ПРИМЕНЕНИЕ ЭНТРОПИИ ШЕННОНА И КДП КОМБИНАТОРИКИ В ДНК-АНАЛИЗЕ ДЛЯ ВЫЯВЛЕНИЯ БИОЛОГИЧЕСКИХ КЛАССОВ, ЭНТРОПИЙНАЯ ШКАЛА КЛАССОВ

Филатов О.В.

Филатов Олег Владимирович – консультант по КДП - комбинаторике,
ООО «Прог-рам», г. Москва

Аннотация: *в этой статье проверялось предположение, что для родственных групп животных, объединённых в биологический класс, существуют компактные, обнаруживаемые особенности записи ДНК - информации, которые присущи только этому классу животных, и по этим компактным информационным особенностям можно различать классы животных. Действительно, такие уникальные особенности информации для каждого биологического класса были обнаружены в мтДНК. Для расчёта этих информационных меток была применена энтропия Шеннона в сочетании с Комбинаторикой длинных последовательностей. На основании полученных новых научных результатов построена энтропийная шкала классов животных.*

Ключевые слова: *ДНК, МтДНК, КДП, энтропия, энтропия Шеннона, Комбинаторика длинных последовательностей.*

APPLICATION OF SHANNON'S ENTROPY AND CLS OF COMBINATORICS IN DNA ANALYSIS TO IDENTIFY BIOLOGICAL CLASSES, ENTROPY CLASS SCALE

Filatov O.V.

Filatov Oleg Vladimirovich - Consultant for CLS - combinatorics,
LLC "PROG-RAM", MOSCOW

Abstract: *this article tested the assumption that for related groups of animals united in a biological class, there are compact, detectable features of recording DNA information that are inherent only to this class of animals, and classes of animals can be distinguished by these compact informational features. Indeed, such unique features of information for each biological class have been found in mtDNA. To calculate these information marks, Shannon's entropy was used in combination with Combinatorics of long sequences. Based on the new scientific results obtained, an entropy scale of animal classes was constructed.*

Keywords: *DNA, MtDNA, CLS, entropy, Shannon entropy, Combinatorics of long sequences.*

УДК 519.115.8; 575; 57.011; 57.06; 575.2; 575.8

Сокращения: **пос-ть** – последовательность; КДП - Комбинаторика длинных последовательностей.

Введение

Молекула ДНК хранит биологическую информацию в виде генетического кода, состоящего из последовательности нуклеотидов, этот код для восприятия людей отображается в виде последовательности комбинации четырёх букв: A; C; G; T. В ДНК содержатся программы развития и функционирования живых организмов. В настоящий момент компьютерные мощности и базы данных ДНК не могут обеспечить обработку и хранение ДНК сразу всех живых существ, так как объёмы ДНК данных чрезвычайно большие. Поэтому ДНК делят на отдельные функциональные части (гены, хромосомы, мтДНК) и исследуют информацию хранящеюся в них. Поэтому исследование по обнаружению макропризнаков ДНК, которое уникально для каждого отдельного класса существ было проведено на мтДНК - одной из обособленных частей ДНК.

Митохондриальная ДНК (мтДНК) находится (в отличие от ядерной ДНК) в митохондриях, органеллах эукариотических клеток. МтДНК гораздо короче ядерной ДНК, поэтому её информационное исследование не предъявляет высоких требований к вычислительной технике.

В качестве новой научной платформы, которая позволила выполнить поставленную задачу и получить новые научные знания, был выбран новый раздел теории вероятности – Комбинаторика длинных последовательностей (КДП). КДП была соединена с хорошо известным инструментом информатики – энтропией Шеннона. МтДНК последовательности были исследованы при помощи этих двух научных инструментов (КДП и энтропия Шеннона) на предмет существования в них формальных параметров, которые позволяют различать по ним класс организмов, к которым принадлежит исследуемая мтДНК.

Было показано, что эти формальные параметры мтДНК группируются по видам (типам) КДП графиков, по которым определим биологический класс существ. Так же, по величине энтропии Шеннона и КДП, можно определить биологический класс организмов, которому принадлежит мтДНК. Это ожидаемо, так как ДНК содержит информацию об устройстве живых существ. Новым является то, что величинами, которые применяются в информатике и комбинаторике, можно описывать разные классы организмов. При отображении получаемых интегральных величин (энтропия Шеннона, составные события КДП) в виде

графиков и диаграмм, получаемые графические представления для каждого отдельного биологического класса хорошо отличимы от графического представления любого другого биологического класса. Для расчёта энтропии Шеннона взяты смысловые понятия из Комбинаторики длинных последовательностей (КДП является упрощённой, более наглядной версией теории вероятности).

Рассчитываемые величины энтропии Шеннона для мтДНК сравнивались с эталонными значениями, за эталонные значения брались значения энтропии случайной последовательности, по версии КДП (которые можно рассматривать в качестве нулевой отметки на оси координат). Из физики известно, что энтропия в неживой природе не уменьшается (может только возрастать). При расчётах формальных информационных параметров мтДНК ожидалось получить значения энтропии Шеннона меньшими, чем величина энтропии случайной последовательности. Были рассчитаны формальные информационные параметры мтДНК для одиннадцати классов существ (всего 500 мтДНК). Вопреки ожиданиям, значения энтропий для десяти классов оказались не меньше, а наоборот, большими, чем энтропия Шеннона для случайной последовательности. Только один класс показал уменьшение энтропии по отношению к энтропии неживого объекта. Тем не менее, это неожиданное положение формальных информационных параметров на оси хаоса по отношению к положению энтропии случайной последовательности не мешает группировать классы биологических существ по энтропиям и размещать на информационной оси.

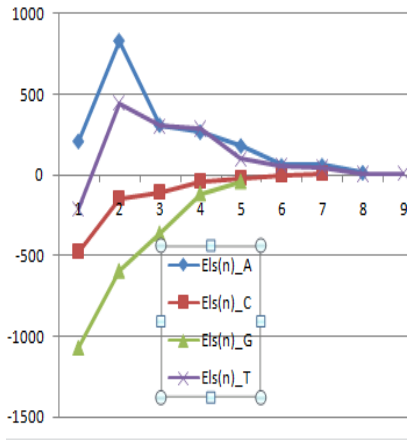
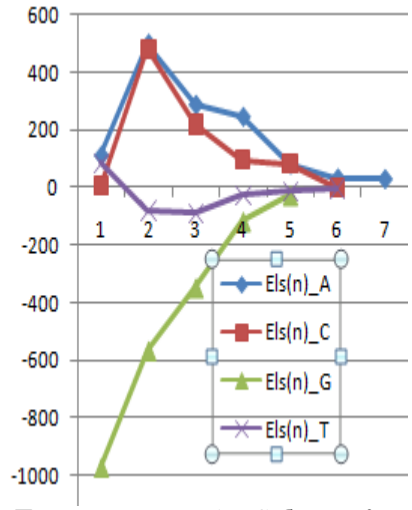
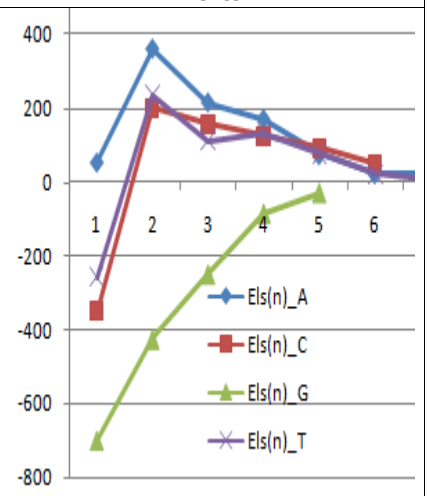
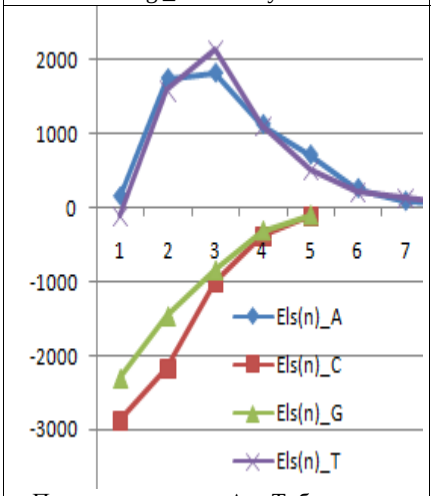
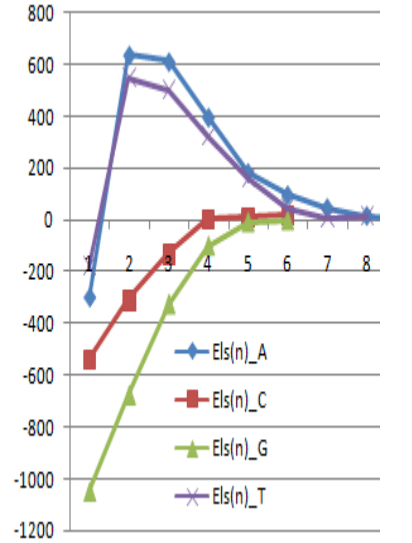
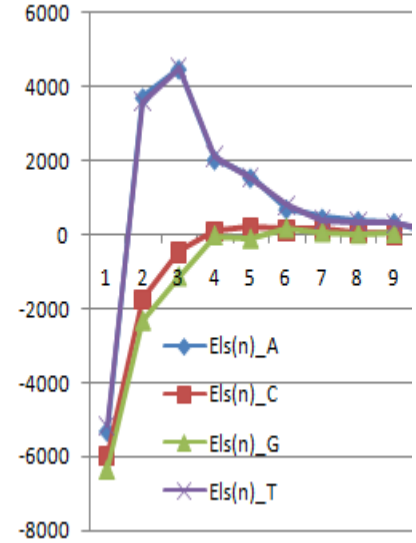
Были рассчитаны информационные параметры для мтДНК следующих биологических классов: Amphibians; Birds; Fishes; Mammals; Reptiles; Fungi Ascomycetes; Fungi Ascomycetes Candida; Fungi Basidiomycetes; Insects; Land Plants; Round worms, в каждом из классов анализировалось примерно 45 случайным образом выбранных мтДНК. Кроме построения принципиально новых графиков, полученных путём учёта объектов Комбинаторики длинных последовательностей в мтДНК, в рамках имеющейся выборки были рассчитаны максимальные, минимальные и средние значения энтропий по каждому классу существ. Так, например, у класса Fungi Basidiomycetes оказалась самая большая энтропия Шеннона из всех рассмотренных классов, она равна: ${}^{\text{КДП}}_{ACGT}H(F\_B)$ = 1,7362. Для сравнения, энтропия случайной пос-ти: ${}^{\text{КДП}}_{ACGT}H$ = 1,5409. Средняя энтропия класса Mammals, к которому относятся люди (${}^{\text{КДП}}_{ACGT}H(M)$ = 1,5749) оказалась очень близка по расположению на энтропийной шкале классов к энтропии случайной пос-ти. Из всех проанализированных классов выделяется один класс Fungi Ascomycetes Candida - единственный класс из рассмотренных, со средней энтропией меньше энтропии случайной пос-ти: ${}^{\text{КДП}}_{ACGT}H(F\_A\_C)$ = 1,4894.

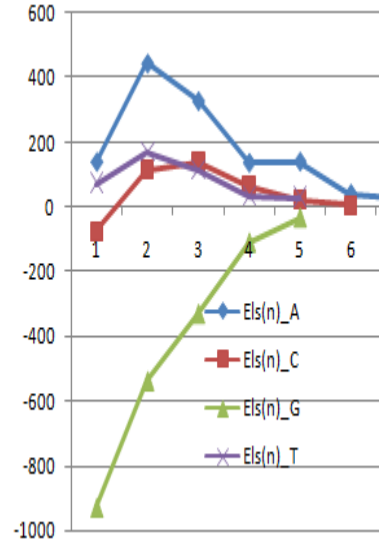
Основная часть.

На рисунках в таблице 1, показаны отклонения нуклеотид (A; C; G; T), по каждой из букв, от чисел нуклеотид случайной пос-ти (подробный пример будет разобран ниже, сейчас дано формальное описание в терминологии КДП). Нулевая отметка на вертикальной оси (вся горизонтальная ось) соответствует числу нуклеотид составных событий каждой буквы (A; C; G; T) в идеальной случайной пос-ти, то есть число нуклеотид в составных событиях идеальной пос-ти бралось за ноль. Число составных событий по каждой букве, для идеальной случайной последовательности, рассчитывается по формуле 1 (формула вводится в работах [1, 2]).

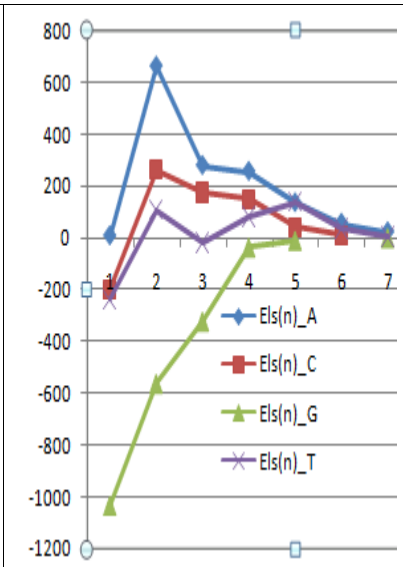
В исследованных мтДНК нуклеотиды отклоняются от расчётных идеальных значений в большую или меньшую сторону, причём видно, что отклонения для разных классов имеют свои характерные особенности, по которым можно определить какому классу принадлежит график, таблица 1. В таблице 1 приводится по одному наиболее общему графику на каждый класс, из примерно 45 имевшихся графиков для каждого класса.

Таблица 1. «Характерные графики разности числа нуклеотид составных событий мтДНК и расчётного числа нуклеотид в ИСП».

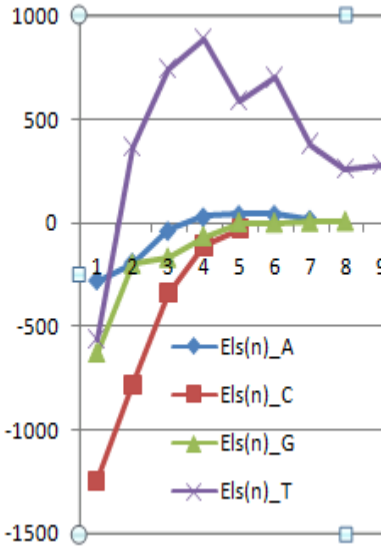
Amphibians	Birds	Fishes
 - Пиковые значения A и T больше 0 и меньше 1000 с номера 2, а C при этом не пиковое и отрицательное;	 - Пиковые значения A и C больше 0 и меньше 1000 с номера 2; - T имеет участок отрицательных значений, когда A и C находятся в положительной полуволне;	 - есть участки с номером больше 1, где A, C, T одновременно положительны; - В большинстве случаев максимумы A, C, T одновременны или близки
Fungi_Ascomycetes Fungi_Basidiomycetes	Insects	Land Plants
 - Пиковые значения A и T больше 1000; - C и G начинаются со значений менее минус 2000 и выходят к нулевой области	 - Пиковые значения A и T меньше 1000 и обычно для обеих букв больше 400; - C и G начинаются со значений меньше 0, но больше минус 1500 и выходят к нулевой области	 - Самые длинные мтДНК (длиннее мтДНК всех других типов существ); - Низкое начало значений A и T меньше -1500
Mammals	Reptiles	Roundworms



Схожи с Reptiles, идентифицируются только при применении цуговых КДП фильтров [5] или цифровых образов.



Схожи с Mammals, идентифицируются только при применении цуговых КДП фильтров [5] или цифровых образов.



- Обычно графики всех букв начинаются в отрицательной области, буква С ниже всех, наибольший максимум у буквы Т

мтДНК взяты из открытой интернет базы данных [3]

Для определения принадлежности графика, таблица 1, к классу животных оцениваются параметры кривых на этом графике. Так, например, для Roundworms характерно начало кривой «С» с отрицательного уровня, ниже начал кривых A; G; T, а значения кривой Т в положительной области больше значений F; C; G. Такое сочетание признаков не встречается в графиках других классов. Выявление характерных особенностей графиков классов и вхождение в одну из групп особенностей графика анализируемого мтДНК, определит принадлежность мтДНК данному классу животных.

Так как графики построены на основании Комбинаторики длинных последовательностей, то рассмотрим два основных объекта КДП – «Элементарное событие» и «Составное событие» на примере коротких фрагментов мтДНК (элементарные и бинарные составные события вводятся в работе [4]). Элементарные события и составные события входят в формулу 1 [1, 2] и в графики таблицы 1. Числами 1; 2; 3; … на горизонтальной оси графиков обозначены n - числа нуклеотид (элементарных событий) в составных событиях.

В следующем фрагменте каждая отдельная буква является в КДП элементарным событием: «..GTGTTTTTCTTTTTGTTG..», в этом фрагменте 18 элементарных событий. Подчёркнуты буквы Т которые образуют Т - составные события ${}_{V=4}^{T}S$. Фрагменты мтДНК составленные из одной нуклеотиды (образованны повторяющейся одной буквой) называются в КДП составными событиями, составное событие обозначается большой буквой S при которой маленькими буквами записаны уточняющие параметры. Для пояснения, заменим в фрагменте: «..GTGTTTTTCTTTTTGTTG..» составные события образованные буквами Т их символьными обозначениями ${}_{V4}^{T}S$: «..G ${}_{V4}^{T1}S$ G ${}_{V4}^{T5}S$ C ${}_{V4}^{T5}S$ G ${}_{V4}^{T2}S$ G..» - уточнение «V4» обозначает, что мтДНК пос-ть образована четырьмя нуклеотидами (A; C; G; T), уточнение «Т1» обозначает, что составное событие имеет вид: «Т»; уточнение «Т5» обозначает, что составное событие имеет вид: «ТТТТТ»; уточнение «Т2» - это: «ТТ».

Так как мы работаем только с составными событиями, которые образуют мтДНК пос-ти, то можно не писать символ составного события S и не писать уточнение числа элементарных событий: V = 4, из которых образованы мтДНК. После исключения символов S и V4 из записи пос-ти составных событий, фрагмент: «..AACACCCGTAGGGATTCACAA..», запишется более компактно: «..$A2\ CAC3\ GTAG3\ AT2\ CACA2$..».

Составные события S нужны для организации сравнения мтДНК и ДНК с длинными идеальными случайными пос-ми, которые рассчитывают по КДП формуле 1, где N – число нуклеотид (длина) мтДНК. В формуле 1, рассчитанные значения ${}_{\text{т}}^{n}S(L)$ соответствуют числу составных событий не в реальных мтДНК, а в идеальной случайной пос-ти, из N нуклеотид. Численность составных событий в идеальной пос-ти по каждой букве $\{A, C, G, T\}$ будет одинакова: ${}_{V4}^{n}S(L) = {}_{V4}^{n}S(A) = {}_{V4}^{n}S(C) = {}_{V4}^{n}S(G) = {}_{V4}^{n}S(T)$. Буква N - длины случайной (идеальной) пос-ти или число нуклеотид в мтДНК. Буква L - обобщённое обозначение одной из четырёх букв: A, C, G, T. Для любой из четырёх нуклеотид случайной пос-ти число составных событий длины n рассчитывается по ф.1 [1, 2]:

$$ {}_{\text{т}}^{n}S(L) = \frac{1}{V} \cdot \frac{(V-1)^2}{V^{n+1}} \cdot N_{\text{мтДНК}}; \quad L = \{A, C, G, T\} \qquad \text{Ф.1} $$

Где: n = 1; 2; 3; … - число одинаковых букв в составном событии; нижняя буква «т» означает, что величина: ${}^{n}_{т}S(L)$ – расчётная (теоретическая); число равновероятных исходов образующих случайную пос-ть: V = 2; 3; 4…, в нашем случае V = 4; A, C, G, T – обозначение нуклеотид (равновероятных исходов, V = 4).

Для определения, насколько реальная мтДНК расходится с идеальной случайной пос-тью рассчитаем составляющие части (составные события) Идеальной Случайной Пос-ти (ИСП) по формуле 2, которую получена из формулы 1. По формуле 2 рассчитывают число составных событий по отдельности по каждой из четырёх равновероятных нуклеотид ($L = \{A, C, G, T\}$) когда они образуют идеальную случайную пос-ть с полным числом всех нуклеотид (длиной): $N_{\text{мтДНК}}$, при равновероятном выпадении четырёх исходов ($L = \{A, C, G, T\}$: V = 4):

$$ {}^{n}_{т}S(L) = \frac{1}{4} \cdot \frac{(4-1)^2}{4^{n+1}} \cdot N_{\text{мтДНК}} = \frac{9 \cdot N_{\text{мтДНК}}}{4^{n+2}} \qquad \text{Ф.2} $$

где: n – число букв в составном событии.

В таблице 1 приведены графики отражающие наиболее типичные распределения ${}^{n}_{Э}\Delta(L)$ - отклонений нуклеотид мтДНК, по классам существ, от числа нуклеотид ${}^{n}_{т}S(L) \cdot n$ в идеальной случайной пос-ти, ф.3. Буква «Т» в КДП обозначает, что результат получен в результате теоретического расчёта. Буква «Э, что результат получен в результате эксперимента.

$$ {}^{n}_{Э}\Delta(L) = \Delta \cdot n = \left({}_{\text{мтДНК}}{}^{n}_{}S - {}^{n}_{т}S(L)\right) \cdot n = \left({}_{\text{мтДНК}}{}^{n}_{}S - \frac{9 \cdot N_{\text{мтДНК}}}{4^{n+2}}\right) \cdot n \qquad \text{Ф.3} $$

В таблице 1 в ячейке Amphibians приведены графики отклонений численности нуклеотид в составных событиях для Dermophis mexicanus (кожистой червяги). Поясним расчёт графиков в ячейке Amphibians при помощи значений таблицы 2. Число нуклеотид в мтДНК Dermophis mexicanus: $N_{\text{мтДНК}}$ = 16162. По формуле 2 рассчитаем число однобуквенных составных событий, ${}^{n=1}_{т}S(L)$: «A»; «C»; «G»; «T», в Идеальной Случайной Пос-ти с Четырьмя равновероятными исходами (ИСП4): ${}^{n=1}_{т}S(L) = \frac{9}{4^{n+2}} \cdot N_{\text{мтДНК}} = \frac{9}{4^{1+2}} \cdot 16162$ =2272.78125, для меньшей трудоёмкости расчётов не производим округлений результатов, смотри таблицу 2, столбец L(теор).

В мтДНК Dermophis mexicanus находится 2482 составных события «A» - единичной длины, обозначаемых: ${}^{n=1}_{Э}S(A)$, смотри столбец «A» таблицы 2. Разница Δ между числом составных событий единичной длины в мт ДНК Dermophis mexicanus ${}^{n=1}_{Э}S(A)$ = 2482, и числом составных событий единичной длины в идеальной случайной пос-ти ${}^{n=1}_{т}S(L)$ =2272 (экспериментальное значение минус теоретическое), равна: Δ = 2482-2272,=209, . По формуле 3, произведение $\Delta \cdot n$ даст разность в нуклеотидах «A» для составных событий единичной длины: ${}^{n=1}_{Э}\Delta(\text{A}) = \Delta \cdot n$ = 209, · 1 = 209,.. , смотри таблицу 2, столбец «A».

Для составных событий образованных двумя нуклеотидами: «AA»; «CC»; «GG»; «TT», число отклонений нуклеотид мтДНК от теоретического значения нуклеотид в идеальной случайной пос-ти, рассчитанно по формуле 2 (при n =2, ${}^{n}_{m}S(L)$ = 568,.. смотри таблицу 2). Так в мтДНК Dermophis mexicanus число нуклеотид (составных событий) типа «AA»: ${}^{n}_{Э}S(A)$ = 983 (таблица 2). В 983 цепочках «AA» содержится 1 966 одинарных нуклеотиды «A». По формуле 3 разница между числом нуклеотид в составных событиях ${}_{мтДНК}{}^{n=2}_{}S$ и числом нуклеотид в ${}^{n=2}_{m}S(L)$ будет: (983 – 568,..) · 2 = 829,.. (смотри таблицы 1 и 2).

Аналогичным образом рассчитаны все значения графиков в таблице 1, по которой видно, что разные классы существ имеют существенно разные особенности графиков ${}^{n}_{Э}\Delta(L)$, а значит, классы существ могут быть определены по значениям, получаемым по формуле 3 и по графикам, которые строятся по значениям формулы 3.

Таблица 2. «Пояснение расчёта значений графиков Amphibians таблицы 1»

	A		**C**		**G**		**T**		**L(теор)**
n	${}^{n}_{Э}S(A)$	${}^{n}_{Э}\Delta A =$ **(Э-Т)** $\cdot n$	${}^{n}_{Э}S(C)$	${}^{n}_{Э}\Delta C =$ **(Э-Т)** $\cdot n$	${}_{т}{}^{n}_{Э}S(L)$	${}^{n}_{Э}\Delta G =$ **(Э-Т)** $\cdot n$	${}^{n}_{Э}S(T)$	${}^{n}_{Э}\Delta T =$ **(Э-Т)** $\cdot n$	${}^{n}_{т}S(L)$
1	2482	209,2188	1798	-474,781	1205	-1067,78	2064	-208,781	2272,
2	983	829,6094	496	-144,391	272	-592,391	790	443,6094	568,
3	245	308,8535	106	-108,146	22	-360,146	243	302,8535	142,
4	102	265,9512	26	-38,0488	7	-114,049	107	285,9512	35,
5	45	180,6097	5	-19,3903	1	-39,3903	29	100,6097	8,
6	12	58,68292	2	-1,31708			12	58,68292	2,
7	9	59,11585	2	10,1155			7	45,11585	0,55
8	2	14,89024					1	6,890244	0,14
9							1	8,687881	0,03
мтДНК Dermophis mexicanus: $N_{мтДНК}$ = 16162 - число нуклеотид мтДНК									

Мы рассмотрели способ определения классов животных по графикам разности параметров мтДНК и идеальной случайной пос-ти (значения которой рассчитываются по формулам КДП). Замечу, что на основе КДП цуг составных событий был разработан другой способ определения принадлежности мтДНК определённому классу [5].

Хотя способ определения классов животных по графикам, таблица 1, очень нагляден и перспективен, но меня интересовал вопрос об упорядочении классов животных при помощи одного формального параметра. В поисках такого параметра я обратился к энтропии Шеннона, которая широко применяется в информатике, так как энтропия Шеннона является мерой хаотичности, то её применение даст величину хаотичности мтДНК. На рисунке 1 приведено упорядочение классов животных по максимальной хаотичности их мтДНК (энтропии Шеннона по КДП для мтДНК). На рисунке 1 под «ИДЕАЛ ACGT» приведена величина энтропии Шеннона для идеальной случайной пос-ти. Как видно из шкалы хаотичности, наибольшей хаотичностью обладает мтДНК класса Fungi Basidiomycetes. Из одиннадцати классов мтДНК у десяти классов энтропия больше чем энтропия идеальной случайной пос-ти. Поэтому очень интересен класс Fungi Ascomycetes Candida, единственный из рассмотренных классов, обладающий средней энтропией мтДНК меньшей, чем энтропия идеальной случайной пос-ти. То есть, мтДНК Fungi Ascomycetes Candida демонстрирует нарушение закона о не уменьшении физической энтропии (энтропия может только расти, а не может уменьшаться).

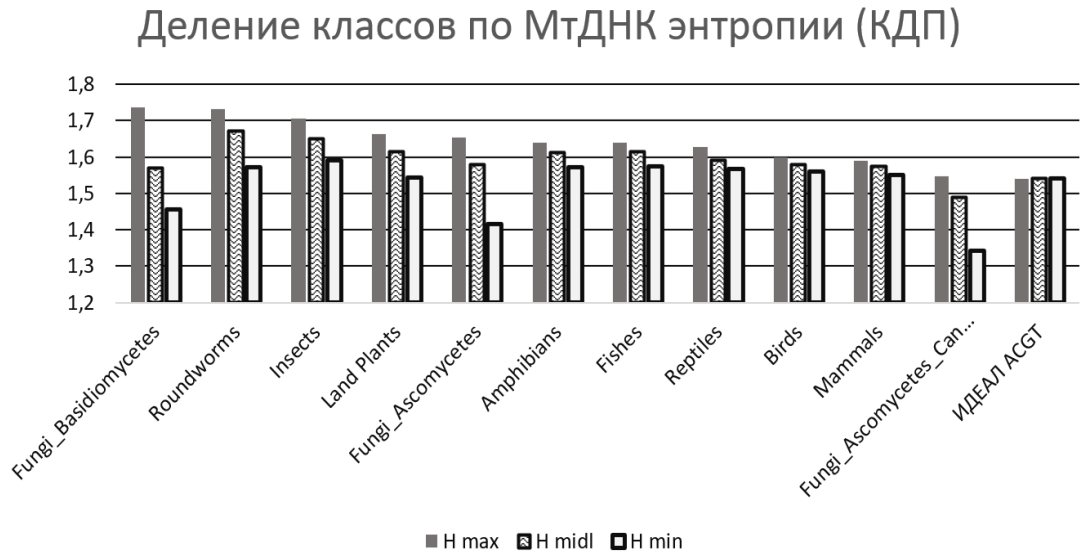


Рис. 1. Последовательность появления классов по величине энтропии Шеннона и КДП

Для построения гистограмм рисунка 1 были случайным образом взяты в среднем по 45 мтДНК каждого класса. В каждом классе была найдена максимальная энтропия, минимальная энтропия и рассчитана средняя энтропия, результаты показаны на рисунке 1. Точные значения рассчитанных энтропий представлены в таблице 3.

Таблица 3. «H max - МтДНК энтропия классов существ как ось времени»

Класс животных	**H max**	**H middle**	**H min**
Fungi Basidiomycetes	1,736168	1,56996	1,456401
Roundworms	1,731469	1,671974	1,572516
Insects	1,704265	1,650837	1,591599
Land Plants	1,66243	1,613937	1,542461
Fungi Ascomycetes	1,653575	1,579919	1,416516
Amphibians	1,639134	1,610871	1,57108
Fishes	1,638777	1,615485	1,574049
Reptiles	1,626587	1,591379	1,566261
Birds	1,598424	1,579015	1,560083
Mammals	1,590039	1,574884	1,551451
Fungi Ascomycetes Candida	1,54712	1,489399	1,342883
ИДЕАЛ ACGT	1,540891	1,540891	1,540891
H max – максимальное значение энтропии Шеннона (формула 5); H middle – среднее арифметическое значение энтропии Шеннона; H min - минимальное значение энтропии Шеннона (формула 5). Button365; ФАЙЛ: «ГРАФИКИ Энтропии Шеннона»			

На рисунке 1 даны распределения энтропий Шеннона, для одиннадцати биологических классов, которые заключают в себе усреднённую информацию по примерно 500 различным мтДНК. Из таблицы 3 видно, что H_max - максимальное значение энтропии убывает от класса к классу и может. H middle – среднее арифметическое значение энтропии Шеннона и H min - минимальное значение энтропии Шеннона, преимущественно убывают от класса к классу.

Опишем расчёт энтропии Шеннона для мтДНК (рисунок 1, таблица 3). В отличие от бинарного алфавита, который состоит из двух символов (0; 1), алфавит ДНК образован алфавитом из четырёх символов (A, C, G, T), поэтому в формуле 4 – формуле энтропии Шеннона расчёт ${}^{\text{КДП}}_{L}H$ ведётся по каждой букве ($L = \{A, C, G, T\}$), а логарифм по основанию четыре ($\log_4$). Полная энтропия Шеннона ${}^{\text{КДП}}_{ACGT}H$ получается при суммировании всех четырёх частных энтропий, формула 5. В файловой записи каждой мтДНК с помощью компьютерной программы подсчитывалось число составных событий по каждой из четырёх нуклеотид $\{A, C, G, T\}$. В результате все нуклеотиды были поделены по соответствующим составным событиям ${}^{n}_{L}S$. Сумма всех составных событий в мтДНК: $S = \sum_{n=1}^{n}({}^{n}_{A}S + {}^{n}_{C}S + {}^{n}_{G}S + {}^{n}_{T}S)$. Энтропия ${}^{\text{КДП}}_{L}H$ определяется для образующих любую последовательность КДП составных событий, это подчёркивают буквы «КДП» в левом верхнем углу символьного обозначения энтропии, формула 4:

$$ {}^{\text{КДП}}_{L}H = -\sum_{n=1}^{n} \frac{{}^{n}_{L}S}{S} \cdot \log_4 \frac{{}^{n}_{L}S}{S}, \qquad \text{Ф.4} $$

где: ${}^{\text{КДП}}_{L}H$ – энтропия Шеннона для одной из четырёх нуклеотид.

${}^{n}_{L}S$ – составные события образованные одной из четырёх букв, подробно описаны выше, при разборе таблицы 1, [2. 4];

S – число всех составных событий в мтДНК, $S = \sum_{n=1}^{n}({}^{n}_{A}S + {}^{n}_{C}S + {}^{n}_{G}S + {}^{n}_{T}S)$.

В идеальной случайной пос-ти, с V равновероятными исходами, ${}_{V}S$ - полная сумма составных событий всех длин $n = 1; 2; 3; \dots$, рассчитывается по формуле 5, [1]. Формула 5 адаптирована под число нуклеотидов идеальной случайной пос-ти:

$$ {}_{V=4}S = \sum_{n=1}^{\infty} {}_{V=4}{}^{n}_{}S = \frac{V-1}{V}N = \frac{3}{4}N \qquad \text{Ф. 5} $$

Полная энтропия Шеннона мтДНК равна сумме частных энтропий, формула 6. Из формул 4 и 5, полная энтропия Шеннона идеальной случайной пос-ти (из 4х равновероятных событий) равна: ${}^{\text{КДП}}_{ACGT}H = 4 \cdot {}^{\text{КДП}}_{L}H =$ 1,54085 .

$$ {}^{\text{КДП}}_{ACGT}H = {}^{\text{КДП}}_{A}H + {}^{\text{КДП}}_{C}H + {}^{\text{КДП}}_{G}H + {}^{\text{КДП}}_{T}H \qquad \text{Ф.6} $$

На рисунке 1 и в таблице 3 приведены максимальные и минимальные значения энтропии, рассчитанные по формуле 6. В выборке по каждому классу содержится примерно 45 мтДНК, среднее значение – это средние значение энтропии по выборке каждого класса.

Обсуждение

Сейчас обще принято, что жизнь развивается от простого к сложному, что жизнь появилась из хаоса и эволюционирует к порядку, при этом в эволюционирующих организмах происходит уменьшение энтропии. Понятие энтропии сейчас пытаются применить к самым разным системам (объектам): физическим, химическим, социологическим, энергетическим и информационным, определяя энтропию как степень неупорядоченности рассматриваемых микросостояний в этих системах (объектах). Комбинаторика длинных последовательностей (КДП) создала свои объекты (составные события), на которые раскладываются любые пос-ти, как случайные, так и не случайные. КДП описала формулами распределения составных событий в случайных пос-тях и рассчитала для случайных пос-тей уровни энтропии Шеннона. Поскольку случайные результаты, как и любые результаты, относятся к области информации, то степень неупорядоченности определялась по отношению информационных объектов – составных событий. Поскольку ДНК идеально подходит под изучение при помощи комбинаторики длинных пос-тей, то были рассчитаны значения энтропии для полутысячи мтДНК различных 11 классов организмов. Оказалось, что для абсолютного большинства различных классов, энтропия их мтДНК не меньше, а наоборот больше, чем энтропия случайной последовательности из четырёх букв. Как это можно объяснить? Есть несколько объяснений.

Сейчас доминирует теория, что в процессе биологической эволюции ДНК накапливает повреждения. А поскольку энтропия только увеличивается, то и чем дольше существует биологический класс, тем большие значения энтропии для его мтДНК. Возникнув из неживой природы, мтДНК имело значение энтропии близкое, практически равное значению энтропии для случайной пос-ти, но по мере накопления

генетических повреждений в процессе эволюции, величина энтропии росла. Таким образом, длительность существования биологического класса можно определять по величине энтропии его мтДНК, смотри рисунок 1 и таблицу 3. Из этого предположения вытекает, что раньше всех из рассмотренных одиннадцати классов возник класс Fungi Basidiomycetes. Классы Mammals и Fungi Ascomycetes Candida возникли самыми последними. Но класс Fungi Ascomycetes Candida обладает меньшей энтропий, чем энтропия случайной пос-ти. Поэтому, либо этот класс эволюционирует совершенно обособлено от всех остальных классов Земли и для него действительно наблюдается эволюция с уменьшением энтропии, либо начальный уровень энтропии мтДНК, с которого начинается жизнь, отличается от уровня энтропии случайной последовательности.

Действительно, например, в природе существуют кристаллы с очень высокой степенью упорядоченности, почему бы и ещё не живым биологическим цепочкам, из которых возникла жизнь не иметь уровень энтропии ощутимо отличающийся от значения уровня случайной последовательности? Допустим, стартовый уровень энтропии образования жизни был больше, чем у Fungi Basidiomycetes, просто мы не имеем мтДНК этих первых организмов, которые доэволюционировали до класса Fungi Basidiomycetes, и уменьшили значения энтропии мтДНК, который и за стабилизировался у Fungi Basidiomycetes. Дальнейшая эволюция приводила к дальнейшему уменьшению величины энтропии. И, вот, наконец Fungi Ascomycetes Candida – это венец Земной (? если они не результат панспермии) эволюции, этот класс появился самым последним, и так уж получилось, что именно он преодолел барьер энтропии случайной пос-ти в процессе развития жизни.

Подтверждением, что всё же энтропия мтДНК уменьшается в процессе эволюции, могут служить четыре энтропии человеческих мтДНК [3]: 1,58413 - денисовский человек, 1,58137 - неандертальский человек; 1,58087 - человек современного типа, 1,58078 - человек современного типа.

Длительность существования на Земле класса животных можно оценить по разности максимальной и минимальной энтропии: H_max - H_min, таблица 4. Обоснование этого предположения так же базируется на том, что скорость изменения энтропии мтДНК одинакова для всех классов, и, класс с максимальной разностью энтропий существует дольше всех классов.

Таблица 4. «Длительность существования класса: H max - H min»

Класс животных	**H max - H min**	**H max**	**H min**
Fungi_Basidiomycetes	0,279766412	1,736168	1,456401
Fungi_Ascomycetes	0,237058363	1,653575	1,416516
Fungi_Ascomycetes_Candida	0,204237053	1,54712	1,342883
Roundworms	0,158952541	1,731469	1,572516
Land Plants	0,119969279	1,66243	1,542461
Insects	0,112665726	1,704265	1,591599
Amphibians	0,06805487	1,639134	1,57108
Fishes	0,064728638	1,638777	1,574049
Reptiles	0,060325849	1,626587	1,566261
Mammals	0,03858747	1,590039	1,551451
Birds	0,038340929	1,598424	1,560083

Из таблицы 4 видно, что наиболее долго из рассмотренных классов на Земле существуют Fungi (грибы). И позже всех появились (существует меньше всех) классы Mammals и Birds.

Выводы

1. Проводившиеся ранее расчёты энтропии Шеннона для мтДНК были мало информативны, так как не использовали знания о структуре случайных последовательностей (из которых образовалась мтДНК в процессе эволюции) и описывающие эту структуру формулы. Эти структуры и формулы были раскрыты в «Комбинаторике длинных последовательностей» и применены в этой статье.
2. При применении нового направления теории вероятности - Комбинаторики длинных последовательностей к анализу мтДНК, стали доступны КДП графики, на которых отображены особенности отклонений чисел нуклеотид мтДНК от идеальной случайной пос-ти.
3. КДП графики отклонений чисел нуклеотид мтДНК, от идеальной случайной пос-ти, имеют признаки, по которым определяется их принадлежность к одному из классов животных (к таким признакам относятся: формы и расположения линий на графике, величины значений, комбинации нуклеотидных линий отклонений).
4. Для каждого класса животных можно рассчитать энтропию Шеннона, для КДП составных событий, диапазоны энтропий большинства классов частично перекрываются.
5. Рассчитанные энтропии классов были упорядочены по величине максимальной энтропии, их упорядоченное расположение отражает длительность существования класса, так как сейчас предполагается, что скорость изменения энтропии ДНК одинакова для всех живых существ.

6. Из рассмотренных одиннадцати классов наибольший интерес представляет класс Fungi Ascomycetes Candida, так как он единственный из классов обладает средней энтропий меньшей, чем энтропия случайной последовательности.

7. Относительная длительность существования класса (по сравнению с другими классами животных) определяется путём набора животных по определённым признакам в члены этого класса (образование класса) и определением разности между максимальной и минимальной энтропии мтДНК, у членов этого класса. Класс, у которого разница энтропий больше дольше существует.

Список литературы / References

1. *Филатов О.В.* «Описание структур любых последовательностей образованных равновероятными случайными событиями» «Проблемы современной науки и образования». № 5 (138), 2019. С. 9-15, DOI: 10.24411/2404-2338-2019-10501.
2. *Филатов О.В.* «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. С. 27-36.
3. АДРЕС БД ДНК: [Электронный ресурс]. Режим доступа: https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/ (дата обращения: 30.03.2020).
4. *Филатов О.В., Филатов И.О.* «О закономерностях структуры бинарной последовательности». «Журнал научных публикаций аспирантов и докторантов», 2014. № 5 (95), С. 226–233.
5. *Филатов О.В.* «Применение структур случайных последовательностей для описания свойств мтДНК и определения принадлежности отдельных мтДНК к их хозяйской группе животных». «Проблемы современной науки и образования». № 5 (150), 2020. С. 6-12.