УДК 57 **БИОЛОГИЯ** (BIOLOGY)

Филатов О.В.

консультант по КДП – комбинаторике:

ООО «Физическая исследовательская лаборатория экспериментальной комбинаторики и информатики»

ООО «Прог-рам»

(г. Москва, Россия)

ЭНТРОПИЯ ШЕННОНА: ЕЁ ПРИМЕНЕНИЕ В ФИЛОГЕНЕТИЧЕСКИХ ИССЛЕДОВАНИЯХ И МЕДИЦИНЕ, ВЕКТОРНЫЙ СПОСОБ ВИЗУАЛИЗАЦИИ, ТРИГОНОМЕТРИЧЕСКИЕ ФОРМУЛЫ И ВЫРАЖЕНИЕ ЧЕРЕЗ КОНСТАНТУ ПИ

Аннотация: *приведены примеры применения энтропии Шеннона и теории вероятности для филогенетических исследований, на примере сравнений мтДНК. Показана связь энтропии Шеннона с числом Пи и углом единичного вектора на плоскости. Показано применение основных тригонометрических формул для расчёта энтропии Шеннона мтДНК. Показан способ применения энтропии Шеннона для поиска раковых заболеваний, которые изменяют ДНК и мтДНК.*

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

Сокращения:

ЭШ - информационная энтропия Шеннона, H,

МтДНК - митохондриальная ДНК,

КДП - Комбинаторика длинных последовательностей,

ИСП4 - идеальная случайная последовательность из четырёх равновероятных событий.

Введение

В филогенетике и биологии стали активно применять математические модели. По мере применения моделей энтропийных расчётов для оценки степени хаотичности различных объектов живого и неживого мира стало понятно, что разнообразие предметных областей, в которых измеряется хаотичность, стремительно растёт. По мимо роста предметных областей, для которых рассчитывается энтропия, растёт и ещё и число базовых энтропий. В качестве примера приведём только две базовые энтропии: физическую энтропию (применяется в химии и технических науках) и информационную энтропию Шеннона (о которой пойдёт речь в этой статье). Так Мак-Артур применил информационную энтропию Шеннона в 1955 г. для оценки степени структурированности биоценозов [1]. А в 1957 г. Р. Маргалеф постулировал теоретическую концепцию, согласно которой разнообразие соответствует энтропии при случайном выборе видов из сообщества [2].

В данной статье рассматривается применение энтропии Шеннона (ЭШ) для получения характеристических величин, как для классов и видов животных, так и для отдельных особей. В статье в качестве предметной области, для которой рассчитывается ЭШ, выступает митохондриальная ДНК. В качестве исследовательской базы было использовано 1000 мтДНК разных животных, принадлежащих к основным классам, полученных из крупной базы данных [3]. Для каждого мтДНК рассчитывали следующие пять энтропий Шеннона: H(A) – ЭШ по нуклеотидам A, H(C) – ЭШ по нуклеотидам C, H(G) – ЭШ по нуклеотидам G, H(T) – ЭШ по нуклеотидам T и общую энтропию Шеннона H(A,C,G,T), которая является суммой четырёх частных энтропий: H(A,C,G,T) = H(A) + H(C) + H(G) + H(T).

В результате сравнений пятёрок ЭШ мтДНК: { H(A,C,G,T), H(A), H(C), H(G), H(T) }, которые были рассчитаны для разных видов животных, оказалось, что каждый вид животных обладает своим уникальным сочетанием, образованным множеством из этих пяти чисел – ЭШ мтДНК:

H = { H(A,C,G,T), H(A), H(C), H(G), H(T) }.

Для лучшего восприятия множеств энтропий, чем просто табличное отображение, были построены как обычные графики ЭШ, так и разработаны специальные энтропийные диаграммы. На разработанных энтропийных диаграммах все пять ЭШ мтДНК показаны в качестве углов между вектором, выходящим из центра прямоугольных декартовых координат и положительным направлением оси X. Предельно простое представление особенностей ЭШ мтДНК, в виде единичных векторов, даёт мощный инструмент для филогенетики и филогенетических исследований.

Для демонстрации применения ЭШ для филогенетических исследований были использованы 62 мтДНК человека вида Homo sapiens и 27 мтДНК Homo sapiens neanderthalensis [3]. Уже эти не многочисленные выборки мтДНК рода Homo показали ценность ЭШ для филогенетических сравнений, в частности наглядно видно гораздо большее генетическое разнообразие Неандертальского человека перед сапиенсами нашего вида (шестьдесят два мтДНК которого взято из БД [4]). Но, в то же время, график ЭШ Неандертальского человека демонстрируют разрыв, что означает, что антропологи ещё не обнаружили переходных звеньев между двумя ветками Неандертальского человека, либо это недостающее переходное звено было причислено к роду Homo sapience и отношение родов требует переосмысление, причиной которого может являться фактическое образование единого рода.

Для сравнения, так же приведён график ЭШ мтДНК не человекообразной обезьяны, макаки, построенный из 70 значений ЭШ, мтДНК взято из БД [3]. Этот график не человекообразной обезьяны расположен в другом, не перекрывающимся с человекообразными, диапазоне значений. Таким образом, графики и диаграммы ЭШ очень компактно и наглядно показывают филогенетическое различие видов.

В конце статьи дана математическая формалистика по тригонометрическим формулам, которым подчиняется ЭШ.

Основная часть.

В новом разделе теории вероятности, который носит название «Комбинаторика длинных последовательностей» (КДП), были получены значения энтропии Шеннона (ЭШ) для Идеальной Случайной Последовательности, образованной четырьмя равновероятными случайными величинами (сокращённо - ИСП4). Очень интересно сравнивать ЭШ для: мтДНК и генов различных организмов с уровнями ЭШ случайной математической последовательности - ИСП4. Как мы далее увидим, разные классы животных имеют разную величину удаления уровней ЭШ от уровней природного хаоса (случайной последовательности). Принципы организации членов для формулы энтропии Шеннона подробно описаны в работе [5], в работе 5 эти формулы имеют номера 13 и 14.

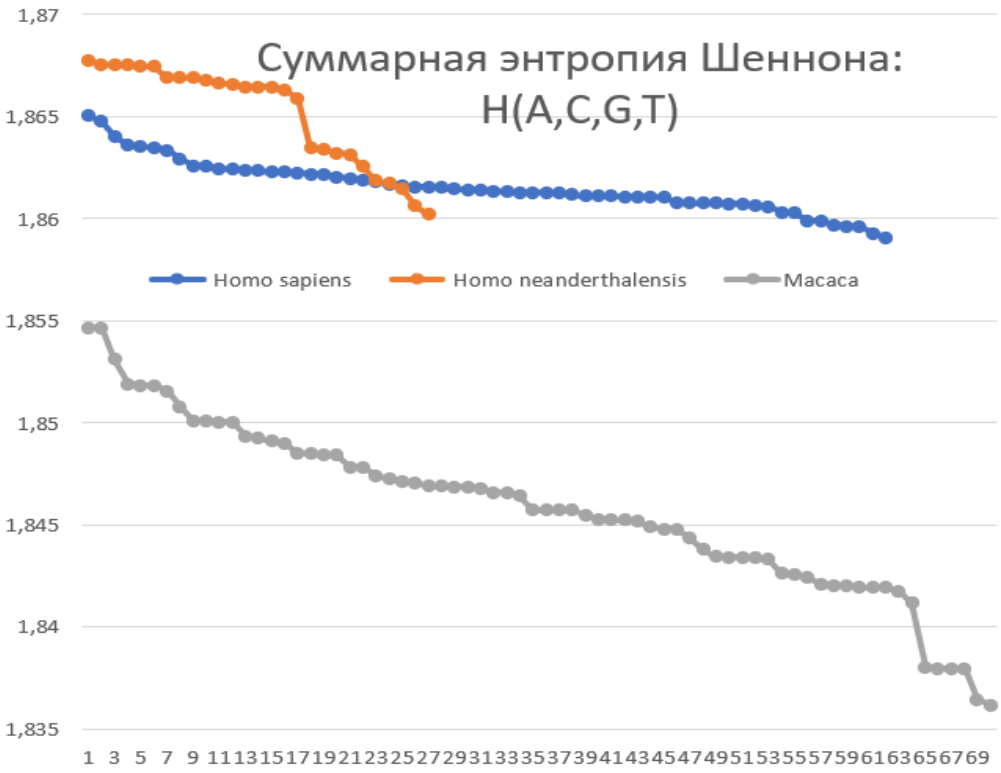

Рисунок 1. «Суммарная энтропия Шеннона H(A,C,G,T) для родов: Homo sapiens, Homo neanderthalensis, Macaca»

Серый график, на рисунке 1, принадлежит роду Макака, он содержит семьдесят значений H(A,C,G,T) мтДНК, которые взяты из БД [3]. Причём в

графике рода макака, на мтДНК Macaca tonkeana приходится две точки и они находятся в самом верху графика Macaca, с наибольшими значениями энтропии H(A,C,G,T). На Macaca arctoides приходится шесть точек, и все эти шесть точек находятся внизу графика Macaca, с наименьшими значениями H(A,C,G,T) энтропии. Причём в области графика Macaca arctoides (шесть нижних точек) виден большой разрыв, который можно объяснить нехваткой образцов мтДНК с промежуточными значениями H(A,C,G,T) энтропии для Macaca arctoides, либо, фактически, необходимостью ввода нового рода.

Филогенетический анализ при рассмотрении пятёрок ЭШ мтДНК.

При помощи рисунка 1 были показаны возможности филогенетического сравнения только при помощи одного параметра мтДНК: H(A,C,G,T) – суммарной энтропии Шеннона. Более глубокий филогенетический анализ возможен при рассмотрении пятёрок ЭШ мтДНК: { H(A,C,G,T), H(A), H(C), H(G), H(T) }. Так на рисунке 2 показана серия общей и частных мтДНК ЭШ: { H(A,C,G,T), H(A), H(C), H(G), H(T) } рода газелей (мтДНК взяты из БД [3]). Вид графика H(A,C,G,T) не вызывает вопросов. Но рассмотрение частных энтропий нуклеотид H(A), H(C), H(G), H(T) выявляет аномальную энтропию у Leptobelus gazella, верхние строчки с аномальными величинами частных ЭШ выделены цветом, рисунок 2.

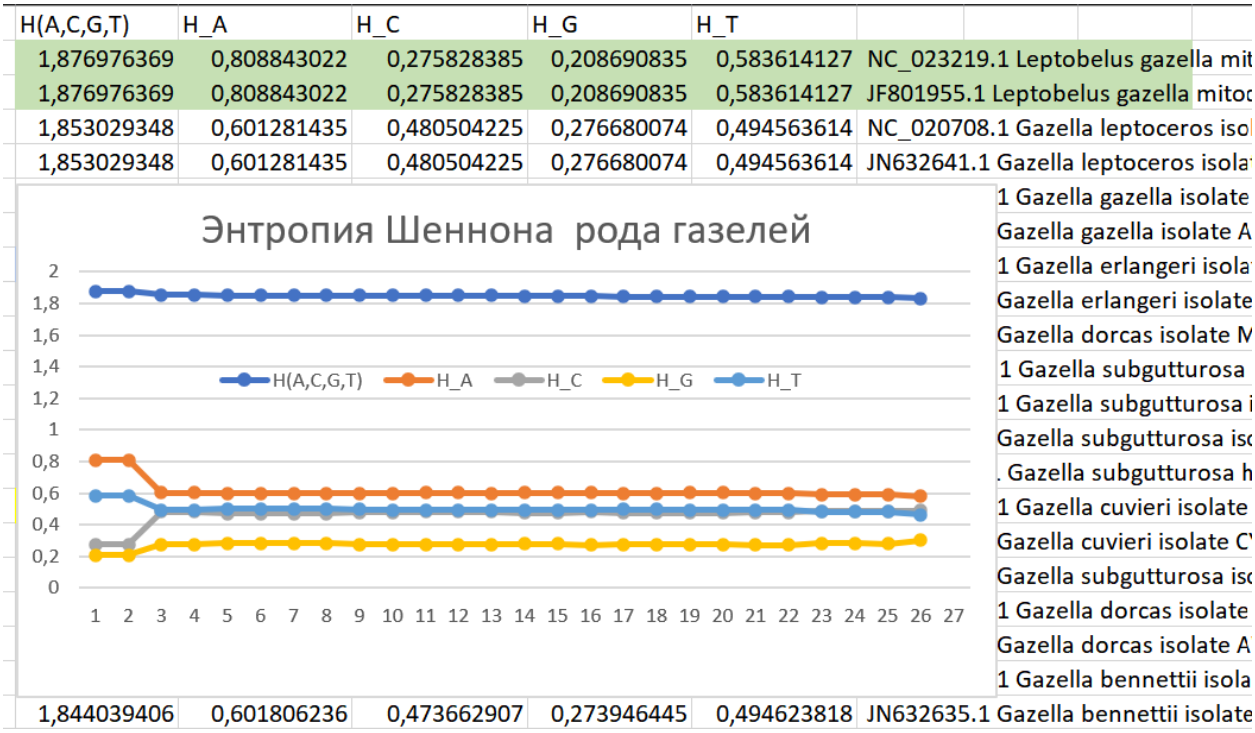

Рисунок 2. «Серия общей и частных ЭШ мтДНК:
{ H(A,C,G,T), H(A), H(C), H(G), H(T) } рода газелей»

Из двадцати шести расстроенных мтДНК, на Leptobelus gazella приходится два. Как видно из графиков H(A), H(C), H(G), H(T), энтропийные значения у Leptobelus gazella сильно отличаются от соответствующих энтропийных значений других газелей в выборке.

Ранее такое аномальное замещение в мтДНК было выявлено только в одном единственном мтДНК человека [4]. Оно одно на 1000 проанализированных мтДНК различных классов животных. Рассмотрим рисунок 3, это слайд из доклада [7].

В таблицах персональные данные двух человек современного вида. Индивид CHM1 обладает типичным мтДНК. У индивида NA24143 мутация, заключающиеся в замене оснований: A<->T; C<->G.

	Homo sapiens CHM1			
	Els(n)_A	Els(n)_C	Els(n)_G	Els(n)_T
1	109,125	-64,875	-933,875	5,125
2	309,0625	327,0625	-568,938	-34,9375
3	256,1484375	298,1484	-301,852	-10,8516
4	126,3828125	242,3828	-109,617	-5,61719
5	69,49462891	184,4946	-35,5054	-10,5054
6	58,34838867	52,34839	-7,65161	4,348389
7	38,01828003	3,01828		3,01828
8	6,862365723			
СУММА:	973,4424133	1042,58	-1957,44	-49,42

	Homo sapiens NA24143 ИНВЕРСИЯ			
	Els(n)_A	Els(n)_C	Els(n)_G	Els(n)_T
1	5,84375	-938,15625	-66,1563	109,8438
2	-35,0781	-571,078125	328,9219	304,9219
3	-13,9043	-295,904297	301,0957	253,0957
4	-5,63477	-113,634766	242,3652	130,3652
5	-5,51086	-35,5108643	179,4891	74,48914
6	4,346741	-7,65325928	52,34674	58,34674
7	3,017799			38,0178
8			6,862228	6,862228
	-46,9198	-1961,93756	1044,925	975,9425

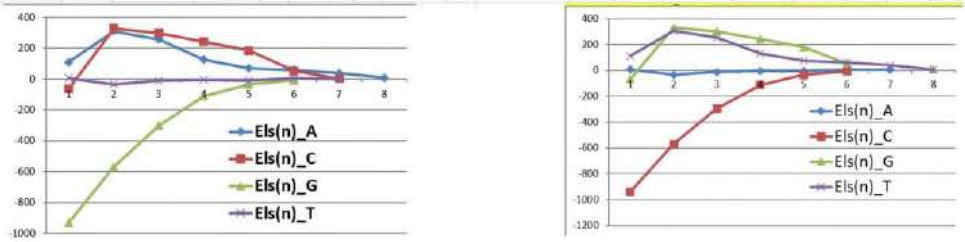

Рисунок 3. «Отклонение численностей нуклеотид от уровня хаоса случайной последовательности»

Практически во всех исследованных мтДНК млекопитающих БД [3] отклонение нуклеотиды «G» имеет вид, как на левом графике рисунка 3, то есть график нуклеотиды «G» начинается с больших отрицательных значений.

На правом графике рисунка 3, файл мтДНК, для которого, скачен из БД [4], нуклеотиды «G» и «C» поменялись друг с другом местами, с больших отрицательных значений начинается график нуклеотиды «C». Так же поменялись местами на левом и правом графиках нуклеотиды «A» и «T». В работе [6] описано получение графиков отклонений численности нуклеотид от случайного уровня (в работе [6] смотри таблицу 1 и формулу 3).

На рисунке 4 дано отклонение численностей нуклеотид от численностей в случайной последовательности для большинства графиков мтДНК газелей (левый график) и аномальный график для Leptobelus gazella (правый график).

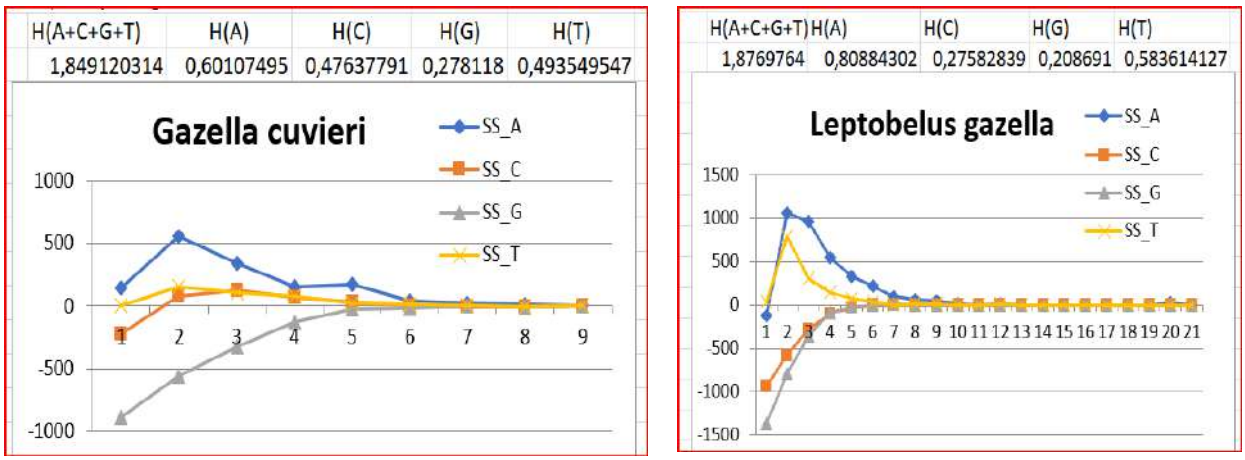

Рисунок 4. «Отклонение численностей нуклеотид от численностей в случайной последовательности, газели»

Слева, на рисунке 4, обычные, типовые отклонения числа нуклеотид от случайного уровня, справа дан график аномальных отклонений у Leptobelus gazella от числа нуклеотид от случайного уровня. Аномальные отклонения у Leptobelus gazella, рисунок 4, имеют максимальные значения, экстремумы, значительно превышающие типовые значения для других газелей (типовой график для газелей приведён слева на рисунке 4).

На примере частных энтропий Шеннона и рисунков 3 и 4 видно, что типовые значения присущи абсолютному большинству исследуемых мтДНК, но встречаются отдельные значения в распределении, которые резко отличаются от типовых. Такие резкие отклонения могут говорить либо об ошибках идентификации мтДНК, либо о действительно интересном случае, который нужно изучить (например, добавить новый род), либо было проведено секвенирование раковой клетки, в которой заболевание вызвало значительные мутации, выходящие за нормальный разброс значений, и диагноз заболевания получен в виде наглядных графиков.

Для детального изучения одного конкретного мтДНК удобно использовать таблицы и графики отклонений численностей нуклеотид (сгруппированных в составные события) от уровней численностей случайной последовательности, рисунки 3, 4. При работе с большими сериями мтДНК для их сравнения и визуализации удобно работать с графиками на которых значения

мтДНК отображены в виде энтропии Шеннона (H(A,C,G,T), H(A), H(C), H(G), H(T)), рисунки 1, 2.

Отображение ЭШ мтДНК в виде угла единичного вектора.

Кроме описанных выше двух способов быстрой оценки на принадлежность и типичность исследуемого мтДНК к энтропийным значениям своего класса, рассмотрим третий способ оценки – векторной. Векторный способ оценки будет востребован в филогенетике при определении на типичность и принадлежность одного неизвестного, исследуемого мтДНК, к некоторому виду животных. То есть, векторный способ поможет определить вид животных, к которому принадлежит исследуемый мтДНК. Иллюстрация различия энтропии для вида «Человек» с энтропиями некоторых других видов дана на рисунке 5.

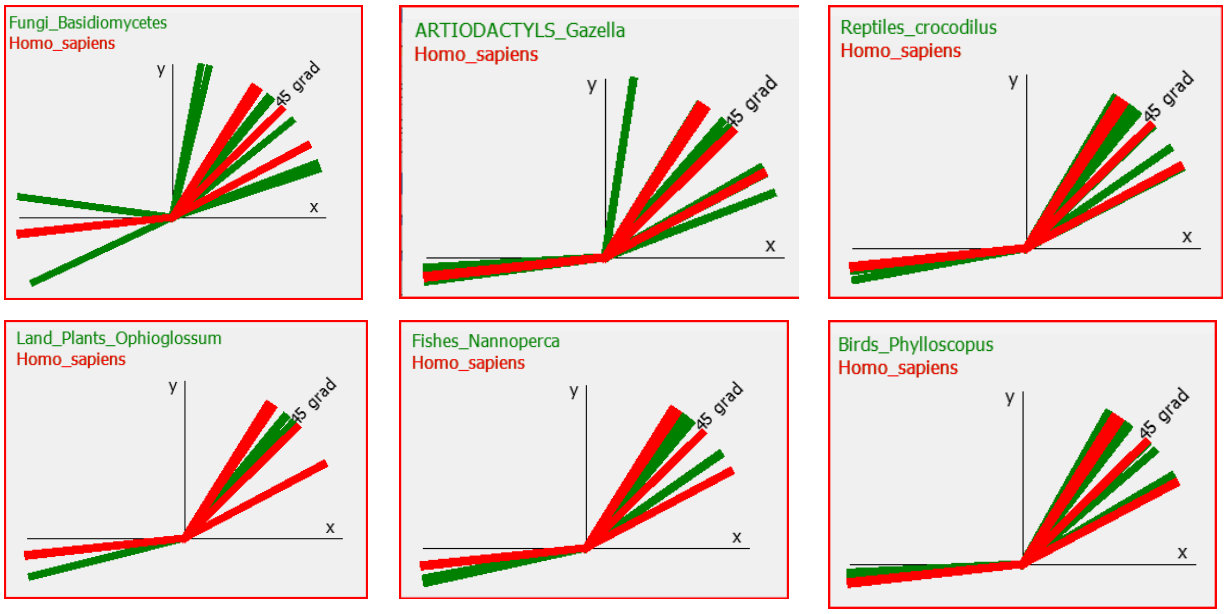

Рисунок 5. «Иллюстрация различия сегментарной энтропии для вида «Человек» с энтропиями некоторых других видов»

На диаграммах рисунка 5 показаны энтропии Шеннона (H(A,C,G,T), H(A), H(C), H(G), H(T)). Энтропии Шеннона мтДНК человека закрашены красным цветом. ЭШ мтДНК других видов животных закрашены зелёным

цветом (названия видов животных написано на диаграммах зелёным, а человека красным).

На диаграммах рисунка 5 выведены не вектора, а сегменты, границы которых образуют вектора с самой большой и самой маленькой энтропией для данного вида. Поэтому на диаграммах видны не тонкие линии, а широкие цветные полосы. Каждая энтропия мтДНК отдельного индивида, особи, вида будет находится внутри этой цветной полосы. Самое интересное в том, что из-за узости таких сегментов вероятность совпадения всех пяти сегментов (которые являются ЭШ) становится маленькой. Достаточно одному сегменту не совпасть с известным распределением на проверяемый вид животного, и сразу можно констатировать, что анализируемое мтДНК не принадлежит животному искомого вида. Надо отметить, что в будущем можно будет организовать контроль по большему числу ЭШ, получаемых из компонентов мтДНК, имеется в виду, что описываемые пять ЭШ получены на основе составных событий «Комбинаторики длинных последовательностей» - КДП, но в КДП существуют и более многочисленные фракции, которые называются «Цуги». Из этих цуг КДП можно создать ещё более точную и чувствительную систему контроля видов.

Хранение информации о классах животных в виде таблиц с максимальными и минимальными значениями энтропии Шеннона, сгруппированных по родам, рисунки 2 и 5, чрезвычайно резко сокращает объём базы данных, если ставится задача определить род животных, которому принадлежит исследуемое мтДНК.

Поскольку биологический класс образован множеством биологических видов, то можно найти максимальную и минимальную ЭШ не только для вида или рода, но и класса. Носителем максимальной или минимальной ЭШ класса будут виды входящие в этот класс. На рисунке 6, показан разброс суммарной энтропии H(A,C,G,T) для основных классов животных. Максимальные и минимальные наклоны векторов, которые соответствуют только суммарной ЭШ: H(A,C,G,T), варианты диапазонов частных энтропий классов: H(A), H(C), H(G),

H(T) не приводятся (формат статьи не позволяет делать большие информационные выкладки).

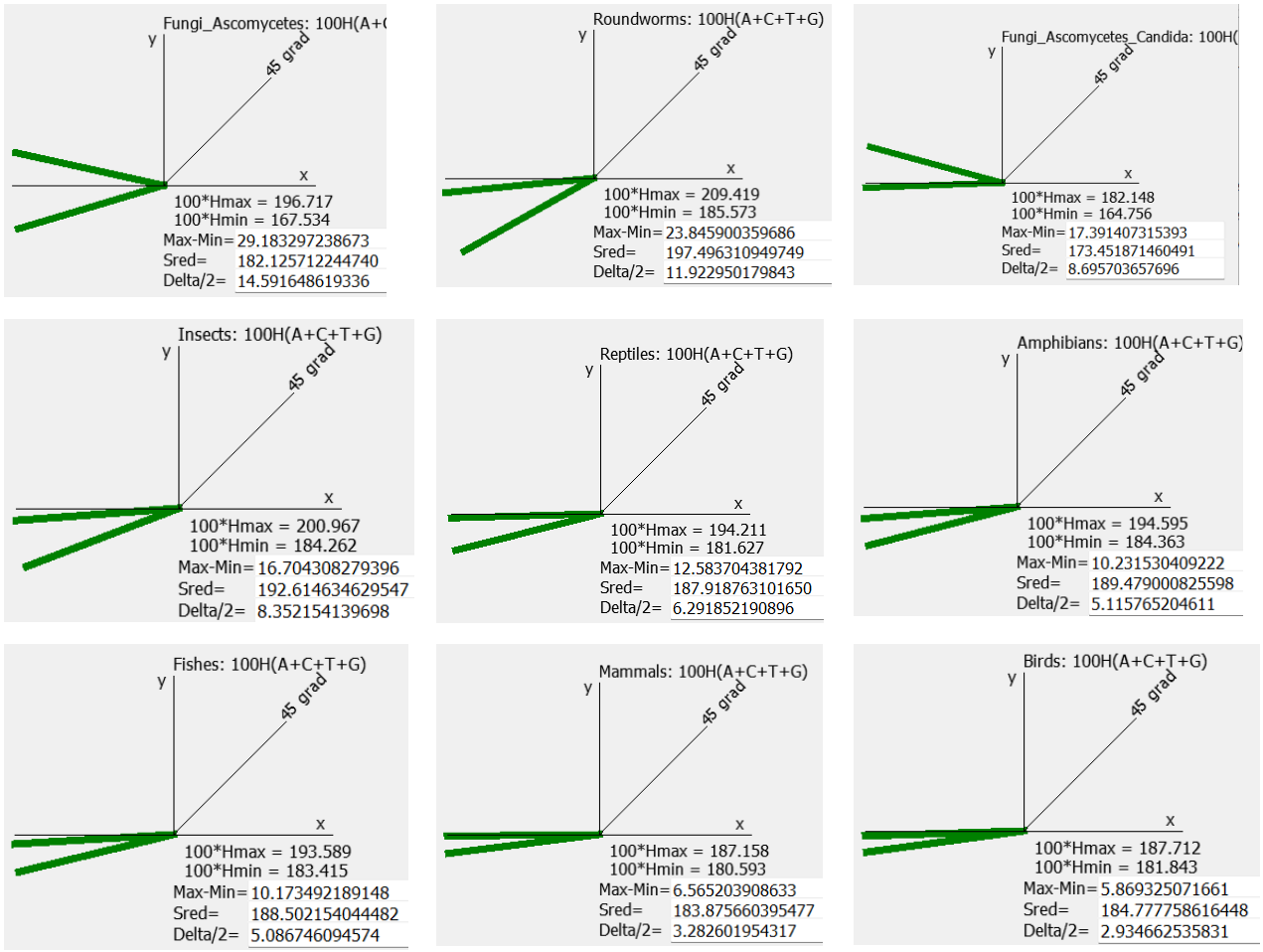

Рисунок 6. «Единичные вектора максимальной и минимальной ЭШ мтДНК, для разных классов животных»

На рисунке 6 видно, что разные классы обладают дельтами суммарной ЭШ: H(A,C,G,T) разной величины, которая находится как разница между максимальной энтропией в данном классе и минимальной. Принято считать, что со временем в ДНК рода накапливаются мутации и энтропия растёт, накапливается. Тогда, по величине дельты энтропии, рисунок 6, самыми молодыми классами, из представленных на рисунке 6, являются птицы и млекопитающие, а самыми древними являются черви и грибы. Но к временной шкале, основанной на рисунке 6, надо относится как к тенденции, так как количество мтДНК, используемое для получения распределений, на рисунке 6,

было крайне невелико по сравнению с численностью животных видов входящих в эти классы.

Обоснование представления ЭШ мтДНК в виде единичного вектора

В работе [6], на формулах 12 – 14, показано, что в идеальной случайной последовательности, образованной четырьмя равновероятными событиями, любая частная энтропия Шеннона равна константе 0,45 (в терминах ДНК это энтропия для одной любой нуклеотиды: A, C, G, T, если бы нуклеотиды были случайны и равновероятны). Обозначив в случайной пос-ти ИСП4, случайные равновероятные события буквами A, C, G, T и рассчитав для них энтропию Шеннона (формулы 12 -14 [6]), получим, что все частные энтропии Шеннона равны (это не для мтДНК, а для ИСП4):

H(A) = H(C) = H(G) = H(T) = 0,45.

Следовательно, полная ЭШ H(A,C,G,T) ИСП4 равна 1,8:

H(A,C,G,T) =H(A)+H(C)+H(G)+H(T) = 1,8.

Замечаем, что при умножении частной энтропии, например H(A), на 100 получаем величину 45. А при умножении полной энтропии на 100 получим величину 180, что соответствует сумме четырёх углов в 45 градусов (45+45+45+45 = 180). Таким образом сумму четырёх ЭШ идеальной случайной пос-ти ИСП4, умноженных на 100, можно связать с величиной в 180 градусов или с числом Пи. Что и было сделано при получении диаграмм энтропий мтДНК на рисунках 5 и 6. На рисунках 5 и 6 ЭШ, умноженная на 100, составляла угол между положительным направлением оси «X» и единичным вектором.

Определение ЭШ как угловой меры:

Полная энтропия Шеннона H(A,C,G,T) и частные энтропии Шеннона H(A), H(C), H(G), H(T), любого мтДНК, могут быть выражены в угловой мере, в виде угла между единичным вектором, берущим начало в точке O на оси «X» и положительном направлением оси «X», с сомножителем равным 100 (умножением ЭШ на 100).

Следствие 1. Полная ЭШ ИСП4 может быть выражена через число Пи.

В развитие темы связи ЭШ с углами единичного вектора (в единичной окружности), замечаем, что 180 градусов — это число Пи в системе счисления в основу которой положены радианы. Таким образом, при переходе в систему, в которой углы измеряются в радианах, полная энтропия Шеннона для случайной пос-ти из четырёх равновероятных событий, после умножения на сто, равна числу Пи.

Так как 180 градусов в идеальной случайной последовательности из 4х равновероятных событий (сокращённо ИСП4) получается путём умножения общей энтропии Шеннона $\mathrm{H}(A,C,G,T)$ на 100, так как $\mathrm{H}(A,C,G,T)$ = 1,8. То есть: $180 = 100 \cdot \mathrm{H}(A,C,G,T)$, то можно ввести коэффициент k, который связывает 180 градусов с числом Пи: $k = \frac{180}{\pi}$. Тогда число Пи с применение этого коэффициента k будет: $\pi = \frac{180}{k}$. Для ИСП4 число Пи через общую ЭШ - $\mathrm{H}(A,C,G,T)$, выражается по формуле 1 (учитывая, что: $\mathrm{H}(A,C,G,T)$ = 1,8):

$$\pi = \frac{100 \cdot H_T(A,C,G,T)}{k} = \frac{180}{k} \qquad \text{Ф. 1}$$

Где: H_T – теоретически полученное значение, помечено символом Т,

так как: $\pi = \frac{180}{k}$, то $k = \frac{180}{\pi}$ и коэффициент k примерно равен: $k \approx$57,296.

С учётом того, что общая энтропия Шеннона, для ИСП4 - это сумма четырёх частных энтропий: H_T (A,C,G,T) = H(A) + H(C) + H(G) + H(T), то формулу 1 перепишем в виде формулы 2 (суммы энтропий):

$$\pi = \frac{100 \cdot (\mathrm{H(A)} + \mathrm{H(C)} + \mathrm{H(G)} + \mathrm{H(T)})}{k} \qquad \text{Ф. 2}$$

Для любой отдельной нуклеотиды в ИСП4, значение угла в радианах будет одной четвёртой от числа π. Если обозначим буквой X любую из четырёх букв (нуклеотид), то есть: $h = H_T(\mathrm{X}) = \mathrm{H(A)} = \mathrm{H(C)} = \mathrm{H(G)} = \mathrm{H(T)}$, то из частной энтропии h, любой нуклеотиды ИСП4, получаем теоретический угол $\varphi_T(h)$, путём деления числа π на четыре, формула 3:

$$\varphi_T(h) = \frac{\pi}{4} = \frac{100 \cdot H_T(X)}{4 \cdot k} = \frac{25}{k} \cdot \mathrm{H}(X) \qquad \text{Ф. 3}$$

Приведённые выше формулы 1 - 3 для ИСП4 выражают угол в радианах через энтропию Шеннона. Не менее интересно переписать формулы 1- 3, наоборот, где уже энтропия Шеннона будет выражена через угол. Так формула 1, связывающая полную энтропию $H_T(A, C, G, T)$ ИСП4 с числом Пи, примет вид формулы 4,

$$H_T(A, C, G, T) = \frac{\pi k}{100} = 1{,}8 \qquad \text{Ф. 4}$$

Переходя от ИСП4 к реальным мтДНК естественно ожидать, что угловая мера будет отличаться от числа Пи. Вместо числа π будем получать угол $\varphi(H)$ – для полной энтропии и углы $\varphi_i(h)$ – для частных энтропий. При обозначении реальной энтропии мтДНК в правом нижнем углу буквы Н уже не будем писать символ теоретического расчёта - «Т». Формула 4 позволяет рассчитать угол общей энтропии $\varphi_i(H)$ для некоторого i-го мтДНК: $\varphi_i(H) = 100 \cdot (H_i(A) + H_i(C) + H_i(G) + H_i(T))$, *г*де: $H_i(X)$ – это одна из частных энтропий исследуемого i-го мтДНК.

Угол для полной энтропия $\varphi_i(H)$ некоторого i-го мтДНК образуется как сумма углов от его частных энтропий, формула 5:

$$\varphi_i(H) = \varphi_i(A) + \varphi_i(C) + \varphi_i(G) + \varphi_i(T) \qquad \text{Ф. 5}$$

Где: $\varphi_i(X) = 100 \cdot H_i(X)$, а $H_i(X)$ – частная энтропия Шеннона в i-ом мтДНК.

На рисунке 6 энтропия Шеннона девяти классов животных представлена в прямоугольной системе координат в виде углов единичного вектора и рассчитаны дельты: $\Delta\varphi_i(H)$. Формула 6 задаёт расчёт дельты $\Delta\varphi_i(H)$ между максимальным ${}^{i}_{Max}\varphi(H)$ и минимальным углом ${}^{i}_{Min}\varphi(H)$ для мтДНК исследуемого класса животных.

$$\Delta\varphi_i(H) = {}^{i}_{Max}\varphi(H) - {}^{i}_{Min}\varphi(H) \qquad \text{Ф.6}$$

Для описания одним числом целого класса животных, введём средний энтропийных угол $\bar{\varphi}_i(H)$, формула 7:

$$\bar{\varphi}_i(H) = \frac{{}^{i}_{Max}\varphi(H) + {}^{i}_{Min}\varphi(H)}{2} \qquad \text{Ф.7}$$

Для совмещения характеризующих возможностей формул 6 и 7, введём формулу 8, которая характеризует общую энтропию класса животных через средний угол энтропии$\bar{\varphi}_i(H)$ и величину поправки $\frac{\Delta\varphi_i(H)}{2}$ (которая показывает ширину зоны $\Delta\varphi_i(H)$ между максимальным и минимальным энтропийными углами ${}^{i}_{Max}\varphi(H) - {}^{i}_{Min}\varphi(H)$).

$$\bar{\varphi}_i(H) \pm \frac{\Delta\varphi_i(H)}{2} = \frac{{}^{i}_{Max}\varphi(H) + {}^{i}_{Min}\varphi(H)}{2} \pm \frac{{}^{i}_{Max}\varphi(H) - {}^{i}_{Min}\varphi(H)}{2} \qquad \text{Ф.8}$$

Аналогичным образом, по формуле 8, рассчитываются средние значения и их отклонения для частных энтропийных углов Шеннона: $\bar{\varphi}_i(X) \pm \frac{\Delta\varphi_i(X)}{2}$.

Обсуждение

Сейчас предполагают, что по мере существования любого класса животных, накапливающиеся флуктуации в мтДНК приводят к росту разнообразия: видов и подвидов, входящих в этот класс. То есть, чем дольше существует класс, тем больше у него дельта $\Delta\varphi_i(H)$, формула 6.

Из рисунка 6 и формулы 6 видно, что наибольшим энтропийным разбросом, дельтой, обладают Fungi Ascomycetes: $\Delta\varphi_{\text{Fungi\_Ascomycetes}}(H) = 29.183$. Они интересны ещё и тем, что их средний угол ближе всех средних углов других классов совпадает со 180 градусами: $\bar{\varphi}_{\text{Fungi Ascomycetes}}(H) = 182.126 \pm 14.592$, то есть они почти сбалансированы, как и природный хаос, вокруг 180 градусов (вокруг Пи).

Интересны так же и Fungi Ascomycetes Candida, они единственные, из рассмотренных классов, у кого угол средней суммарной энтропии меньше 180

градусов:

$\bar{\varphi}_{\text{Fungi Ascomycetes Candidas}}(H) = 173.45 \pm 8.695.$

Если гипотеза о связи ширины зоны энтропийных углов $\pm\frac{\Delta\varphi_i(H)}{2}$ с длительностью существования данного класса верна, то млекопитающие образовались раньше птиц, так у птиц, по полученным данным, наименьшая ширина зоны энтропийных углов: $\bar{\varphi}_{Birds}(H) = 184.778 \pm 2.935.$

При рассмотрении максимальных и минимальных углов в векторном задании границ классов животных, указанные в статье значения являются предварительными, так как число мтДНК в тестовой базе данных было примерно 1000 на все классы животных. В то же время, в настоящий момент только у млекопитающих насчитывается 1258 родов. Поэтому, по мере роста выборки, максимальные значения углов векторов энтропии Шеннона будут увеличиваться для каждого вида, а минимальные значения углов уменьшаться.

В статье были рассмотрены несколько способов анализ данных мтДНК при помощи энтропии Шеннона.

На рисунках 1 и 2 показаны точечные энтропийные графики. Их анализ показывает, как среди общих, обычных для данного класса распределений можно выявить необычное, аномальные распределение, которое говорит либо о выявлении нового вида (пода), либо это выявлена ошибка секвенирования и её необходимо выкинуть из рассматриваемых данных.

Так как энтропия Шеннона рассчитывается на основе составных событий «Комбинаторики длинных последовательностей» [8, 9], то для детального анализа заинтересовавших данных необходимо перейти с уровня энтропии на уровень КДП - составных событий, рисунки 3 и 4. На рисунках 3 и 4 горизонтальная ось обозначает природный уровень хаоса, а линии графиков – это отклонении образований мтДНК (составных событий) от уровня природного хаоса.

Если на рисунках 3 и 4 действительно зафиксированы новые, не типичные мтДНК известных классов животных, то возникает вопрос о этих нетипичных, «неземных», мтДНК, которые явно нельзя включать в общий ряд

мтДНК рассматриваемого класса животного. То есть, для классификации таких мтДНК с мутациями нужно создавать собственные подклассы, списки.

Применения раздела теории вероятности – комбинаторики длинных последовательностей и рассчитанной на основе её составных событий энтропии Шеннона, для мтДНК, произвело фундаментальную подвижку в филогенетике и биологии, так как был перекинут мост между фундаментальными законами физики и математики, с одной стороны, и биологией, и филогенетикой, с другой. Автор КДП надеется, что ему удалось создать принципиально новый исследовательский научный инструмент.

Продолжая математизацию биологии, автор обнаружил интересную взаимосвязь между значениями энтропии Шеннона, для ИСП4, и единичным вектором в декартовой системе координат. На основе этой связи был предложен принципиально новый, простой и наглядный способ отображать энтропию Шеннона для ДНК в виде углов поворота единичных векторов. Автор произвёл поиск в своей тестовой базе данных, на 1000 мтДНК, на предмет выявления совпадений пяти векторов характеризующих мтДНК человека с пятёрками векторов всех других животных в базе. Компьютерный поиск не выявил совпадений. То есть, применение пятёрок углов единичных векторов: $\varphi_i(H), \varphi_i(A), \varphi_i(C), \varphi_i(G), \varphi_i(T)$, формула 5, позволяет предварительно определять класс животных анализируемых мтДНК, когда он неизвестен.

Связывание ЭШ с углами единичного вектора, рисунок 5 и 6, формулы 3 и 4, формально позволяет применить основные формулы тригонометрии к ЭШ, включая получение проекций ЭШ на оси координат. Так проекции $\varphi_x(H)$ - максимальных и минимальных суммарных энтропий, рисунки 5 и 6, на ось X будут косинусами энтропийного угла единичного вектора, формула 9:

$$\varphi_x(H) = \cos(\varphi_i(H)) \qquad \text{Ф. 9}$$

Также по формуле 9 и 10 находят проекции углов частных энтропий H(A), H(C), H(G), H(T) на ось икс, где: $H_i(X)$ – это одна из частных энтропий исследуемого i-го мтДНК, формула 10:

$$\varphi_x(H_i(X)) = \cos(\varphi_i(H_i(X))) \quad \text{Ф. 10}$$

По формуле 11 находят проекции единичных энтропийных векторов на ось Y (включая вектор полной энтропии):

$$\varphi_y(H_i) = \sin(\varphi_i(H_i)) \quad \text{Ф. 11}$$

А также можно применить формулу Пифагора к проекциям энтропийных векторов Шеннона на оси координат, формула 12:

$$(\varphi_x(H_i))^2 + \left(\varphi_y(H_i)\right)^2 = \sin^2(\varphi_i(H_i)) + \cos^2(\varphi_i(H_i)) = 1 \quad \text{Ф. 12}$$

Выводы

- Энтропия Шеннона H(A,C,G,T) рассчитанная для мтДНК родов Homo sapiens и Macaca не имеют общих областей значений, что наглядно показывает принадлежность к разным родам.
- Наличие больших разрывов на графических представлениях энтропии Шеннона H(A,C,G,T), для Homo sapiens neanderthalensis и Macaca, наглядно демонстрирует нехватку исследуемых образцов мтДНК, что можно трактовать, что эти образцы с переходными свойствами ещё не найдены или, что необходимо произвести разделение рода на два (ввести новый род).
- Разделение Энтропии Шеннона на пять энтропий: общую энтропию и четыре частных энтропии, позволяет обнаруживать не обнаруживаемые только одной общей энтропий Шеннона особенности исследуемого мтДНК, так энтропийные свойства мтДНК Leptobelus gazella резко отличаются в частных энтропиях от основного массива мтДНК представителей рода gazella.
- Построение графиков отклонения от уровня природного хаоса составных событий мтДНК позволяет обнаружить: не корректно проведённое секвенирование, либо новый биологический род, либо обнаружить раковые клетки, что показано на графиках составных событий для рода Homo sapiens и gazella.

- Для увеличения наглядности при визуальном сравнении различных мтДНК было разработано представление мтДНК в виде единичного вектора (-ов), в которых угол к оси X связан через константу с Энтропией Шеннона.
- Расчёт разности углов у векторов с максимальной и минимальной суммарной энтропией Шеннона связано с временем существования исследуемого класса животных (вектора которого использовались), чем больше разность, тем дольше существует род.
- Представление энтропии Шеннона в виде единичного вектора позволило связать число Пи в формулах с энтропией Шеннона.
- При представлении энтропии Шеннона в виде единичного вектора применимы все основные тригонометрические формулы для его преобразования и его разложения на проекции.
- Предельно простое представление особенностей ЭШ мтДНК, в виде единичных векторов, даёт мощный инструмент для филогенетики и филогенетических исследований.
- Если ставится задача определить род животных, которому принадлежит исследуемое мтДНК, то хранение информации о классах животных в виде таблиц с максимальными и минимальными значениями энтропии Шеннона, сгруппированных по родам, вместо файлов формата FAST, резко сокращает объём базы данных.

СПИСОК ЛИТЕРАТУРЫ:

1. MacArthur R.M. Fluctuation of animal populations and measure of community stabiliry// Ecology. – 1955. – V. 36, №3. – P. 533–536. 13. Margalef R. Information theory in ecology // Gen. Syst. – 1958. – V. 3. – P. 36–71.
2. Margalef R. Information theory in ecology // Gen. Syst. – 1958. – V. 3. – P. 36–71.

3. АДРЕСС БД ДНК: https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles , дата обращения 03.2020 г.

4. АДРЕСС БД ДНК: https://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Homo_sapiens/ARCHIVE/BUILD.35.1/ , дата обращения 2019 г.

5. Филатов О. В., статья «Применение энтропии Шеннона и числа Эйлера «е» для описания случайных последовательностей и мтДНК, получение числа «е» через энтропию Шеннона», «Вестник науки и образования», №7 (127), 2022 г., с.29-40, DOI: 10.24411/2312-8089-2022-10706.

6. Филатов О. В., статья «Применение энтропии Шеннона и КДП комбинаторики в ДНК анализе для выявления биологических классов, энтропийная шкала классов», «Вестник науки и образования», №7(127), 2022 г., с. 18-29, DOI: 10.24411/2312-8089-2022-10703.

7. Филатов Л.О., доклад: «МЕТОДИКА ВЫЯВЛЕНИЯ КЛАССА ЖИВОТНЫХ ПО ВЕРОЯТНОСТНЫМ ПАРАМЕТРАМ мтДНК», XXIII Колмогоровские чтения, МГУ, 2023 г. (доклад был удостоен бронзовой медали МГУ).

8. Филатов О. В., Филатов И.О. «Закономерность в выпадении монет – закон потоковой последовательности». Германия, Издательский Дом: LAPLAMBERT Academic Publishing, 2015, с. 268, ISBN 978-3-659-71144-2.

9. Филатов О. В., Филатов И.О., Макеева Л.Л. и др. «Потоковая теория: из сайта в книгу». Москва, «Век информации», 2014, с.200, ISBN 978-5-906511-06-5.

Filatov O.V.

consultant on KDP - combinatorics:

LLC "Physical Research Laboratory of Experimental

Combinatorics and Informatics",

LLC "Prog-ram"

(Moscow, Russia)

SHANNON ENTROPY: ITS APPLICATION IN PHYLOGENETIC RESEARCH & MEDICINE, VECTOR VISUALIZATION METHOD, TRIGONOMETRIC FORMULAS & EXPRESSION THROUGH THE CONSTANT PI

***Abstract**: examples of the application of Shannon entropy and probability theory for phylogenetic studies are given, using mtDNA comparisons as an example. The relationship of the Shannon entropy with the number Pi and the angle of the unit vector on the plane is shown. The application of basic trigonometric formulas for calculating the Shannon entropy of mtDNA is shown. A method of using Shannon entropy to search for cancers that alter DNA and mtDNA is shown.*

***Keywords**: phylogenetics, phylogenetic studies, DNA, mtDNA, entropy, Shannon entropy, CLS, Combinatorics of long sequences.*