УДК 573.22 + 575.89

О ФУНДАМЕНТАЛЬНОЙ СВЯЗИ ГЕНОМОВ МИТОХОНДРИЙ С ГЕНОМАМИ ОРГАНИЗМОВ-НОСИТЕЛЕЙ

Садовский М.Г.

ФГБУН «Институт вычислительного моделирования» Сибирского отделения Российской академии наук, Красноярск, e-mail: msad@icm.krasn.ru

Представлены предварительные результаты исследования связи митохондриальных геномов с геномами организмов-носителей. 1132 генома митохондрий преобразовывались в частотные словари триплетов, затем в 63-мерном пространстве этих частот строилась классификация методом динамических ядер (с разбиением на два и три класса). Полученные классы сравнивались по своему составу. Установлено, что видовой состав полученных классов не случаен: деление на два класса полностью отделяет геномы позвоночных от геномов беспозвоночных, а при делении на три формирует эволюционно близкие группы геномов. Эта корреляция доказывает факт сильной коэволюции митохондриальных геномов и соматических геномов, поскольку кластеризация в пространстве частот проводилась по митохондриальным геномам, а определение таксономической близости – по морфологическим признакам (т.е. по соматическому геному). Коэволюция проявляется в том факте, что физически эти два генома (геном органеллы и основной геном организма) никак не связаны.

Ключевые слова: порядок, триплет, частота, классификация, корреляция, таксономия, эволюция

ON A FUNDAMENTAL RELATION BETWEEN MITOCHONDRION AND HOST GENOMES

Sadovskiy M.G.

Institute of Computational Modeling of Siberian Branch of Russian Academy of sciences, Krasnoyarsk, e-mail: msad@icm.krasn.ru

Some preliminary results are provided approving the strong co-evolution of mitochondrion genomes and host genomes. To do that, 1132 mitochondrion genomes has been converted into frequency dictionaries of triplets, and unsupervised classification (by *K*-means) has been carried out. Classifications at two and three classes were implemented. It was found the taxonomy composition of those classes is extremely regular. Two class classification stably and reliably separates vertebrates from invertebrates. Three class classification forms the groups of evolutionary related organisms. This correlation directly proves the strong co-evolution of these two genetic systems, since the proximity in frequency space has been determined over mitochondria, while the proximity in taxonomy has been determined morphologically (i.e. over somatic genomes).

Keywords: order, triplet, frequency, classification, correlation, taxonomy, evolution

Изучение связи между структурой нуклеотидных последовательностей в ДНК и той функции, которая в них закодирована, составляет центральную проблему современной клеточной и молекулярной биологии. Поток работ в этом направлении необозрим. Другим важным направлением (популяционная геномика) является изучение связи между структурой этих последовательностей и таксономическим положением их носителей.

В настоящей работе изложены предварительные результаты изучения такого рода связи на примере геномов митохондрий. Митохондрии – клеточные органеллы, ответственные за энергетику клетки, обладают собственным геномом, состоящим из одной хромосомы (характерная длина митохондриального генома 5×10^4 пар нуклеотидов); митохондрии есть только у эукариотических организмов. Важная особенность этих геномов – с точки зрения настоящей работы – в том, что все они кодируют абсолютно одну и ту же функцию и, следовательно, при изучении связи между структурой и таксономией можно ожидать, что на эту связь функциональные различия генетических систем оказывать влияния не будут.

Если с таксономическим положением носителя генома всё более или менее понятно: оно определяется по морфологическим признакам (которые, в свою очередь, определяются соматическим геномом организма), – то определение того, что есть структура, требует специального разъяснения. Заметим, что и в таксономии различных организмов происходят изменения, однако они совсем не радикальны, и логика построения классификации живых организмов в целом ясна и понятна. Структуру нуклеотидной последовательности можно определять многими способами; в рамках настоящей работы мы будем под структурой понимать частотный словарь триплетов. Дадим строгое определение. Пусть имеется символьная последовательность из четырёхбуквенного алфавита $\aleph = \{A, C, G, T\}$. Будем предполагать, что никаких других символов в последовательности нет. «Лишние» символы, присутствующие в некоторых геномах,

игнорировались, а полученный текст объединялся в связную последовательность после их удаления.

Триплетом будем называть три подряд стоящих символа $v_1v_2v_3$. Частотным словарём W_3 будем называть список всех триплетов (их, очевидно, не более 64) с указанием их частот. Все частоты связаны соотношением

$$\sum_{v_1v_2v_3} f_{v_1v_2v_3} = 1. \qquad (1)$$

Частота определяется стандартно: как отношение числа копий данного триплета, обнаруженных в последовательности, к их общему числу (равному, очевидно, длине всей последовательности; для этого мы замыкаем последовательность в кольцо).

Тем самым каждый геном отображается точкой в 63-мерное пространство частот. Собственно, задача выявления структурной близости ставится следующим образом: требуется выделить в этом пространстве группы точек (геномов), которые образуют достаточно плотные и чётко выделяемые кластеры. Если такая кластеризация возможна, будем говорить, что на множестве геномов можно задать некоторый порядок. Связь между структурой и таксономией заключается в том, что видовой (таксономический) состав таких выделяемых кластеров оказывается существенно неслучайным [1, 2].

Забегая вперёд, анонсируем основной результат: на множестве из 1132 митохондриальных геномов был обнаружен весьма сложно устроенный порядок, который обладал высокой корреляцией с таксономией носителей этих геномов – в разных кластерах группировались представители разных таксономических групп и, более того, близкие группы попадали в один кластер, а таксономически более далёкие – в разные.

Материалы и методы исследования

Геномы брались в EMBL-банке (*www.ebi.ac.uk/genomes/organelles*); использовался релиз от октября 2009 года. Всего на тот момент в банке хранилось свыше 3500 геномов митохондрий (в настоящее время – более 7000). Для исследования была собрана база, содержащая 1132 генома. Это связано с тем, что в базу были включены не все геномы, а лишь те, которые представляли таксон уровня семейства не меньше чем пятью видами. Такое ограничение связано с тем, что для базы, содержащей геномы, в которых таксоны высокого уровня представлены единственным видом, никакой классификации построить невозможно: такие «одиночные» геномы вносят сильный шум, полностью перекрывающий «сигнал», но сами при этом не могут эффективно повлиять на распределение точек в пространстве.

Таксономическое описание носителя генома содержится в файле, хранящем собственно геном, и извлекалось оттуда для целей анализа таксономического состава кластеров. Общий состав получившейся базы геномов митохондрий был таков: порядок *Batrachia* содержал 51 геном, порядок *Chondrostei* – 5 геномов, порядок *Crocodylidae* – 7 геномов, порядок *Cryptodira* – 25 геномов, порядок *Dinosauria* – 94 генома, порядок *Eutheria* – 193 генома, порядок *Gymnophiona* – 16 геномов, порядок *Metatheria* – 18 геномов, порядок *Neopterygii* – 500 геномов и порядок *Squamata* – 78 геномов.

Классификация проводилась методом динамических ядер; это линейный метод классификации, минимизирующий суммарное расстояние в классе от его членов до центра (среднего арифметического). Для построения классификации использовалось свободно распространяемое ПО *ViDaExpert* (*http://bioinfo-out.curie.fr/projects/vidaexpert/*). Опишем кратко метод. На первом шаге все объекты (словари в нашем случае) разбиваются случайным образом на K классов. В каждом классе определяется центр:

$$c^{(j)}_{v_1v_2v_3} = \frac{1}{M^{(j)}} \sum_{i=1}^{M^{(j)}} f_{v_1v_2v_3}(i). \qquad (2)$$

Здесь индекс i ($1 \le i \le M^{(j)}$) перечисляет элементы класса; понятно, что среднее арифметическое определяется для каждого триплета $v_1v_2v_3$. На втором шаге для каждого из полученных K центров и для каждой точки из всего множества определяются K расстояний – до каждого из центров:

$$\rho^{(i)} = \sqrt{\sum_{v_1v_2v_3} \left(f_{v_1v_2v_3}(l) - c^{(i)}_{v_1v_2v_3} \right)^2}. \qquad (3)$$

Здесь индекс i теперь перечисляет все полученные классы $1 \le i \le K$, а индекс l перечисляет все точки множества, вне зависимости от того, к какому классу она принадлежит.

На третьем шаге принадлежность каждой точки переопределяется: точка переносится в тот класс, чей центр к ней ближе всего. Такая процедура продолжается до тех пор, пока точки не перестанут менять свою принадлежность к классу; подробности см. в [3–5].

Метод динамических ядер не увеличивает числа классов; строго говоря, после построения классификации следует проверять различимость классов, однако в нашей версии метода мы не делали этой работы. Число классов является важным параметром: заранее не очевидно, каким оно должно быть. Собственно, в рамках настоящей работы мы проводили классификацию с разбиением на два класса и на три класса.

Результаты исследования и их обсуждение

Как уже было сказано выше, в рамках настоящей работы строились классификации с разбиением на два класса и на три класса. Разбиение на два класса было высоко устойчивым: из 500 реализаций лишь в 7 случаях наблюдалось такое разбиение, при котором один из классов состоял из единственного элемента. Во всех остальных случаях наблюдалось разбиение на два класса мощностью 154 и 978 геномов соответственно. Более того, в этом разбиении не наблюдалось подвижных геномов -(т.е. таких, которые бы меняли свою принадлежность к классу).

Из 154 геномов первого класса 142 принадлежали классу беспозвоночных, и лишь два генома беспозвоночных (*Reticulitermes flavipes* и *Gampsocleis gratiosa*; номера

доступа EF206314 и EU527333 соответственно) принадлежали второму классу. Во втором классе 976 геномов принадлежали порядку позвоночных, при этом этот класс включал лишь 12 геномов беспозвоночных (номера доступа приведены в скобках): *Ranodon sibiricus* (AJ419960), *Aneides flavipunctatus* (AY728214), *Ensatina eschscholtzii* (AY728216), *Rhyacotriton variegatus* (AY728219), *Desmognathus fuscus* (AY728227), *Hydromantes brunus* (AY728234), *Geotrypetes seraphini* (AY954505), *Pachyhynobius shangchengensis* (DQ333812), *Onychodactylus fischeri* (DQ333820), *Dermophis mexicanus* (GQ244467), *Dicamptodon aterrimus* (GQ368657) и *Hemiechinus auritus* (AB099481).

Разбиение на три класса также было весьма устойчивым. Из 500 реализаций классификации были получены следующие распределения по классам: 975–147...10–8 реализаций, 510–147...475–474 реализации, 511–146...475–18 реализаций. При этом опять же подвижных геномов не наблюдалось.

В табл. 1 представлены результаты распределения геномов по классам. Отметим, что черепаховые полностью попадают во второй класс. Кроме того, млекопитающие фактически полностью попадают в один класс (третий); этим же свойством обладает и семейство геномов порядка новокрылых (насекомые) – у них лишь 4 генома из 143 попадают в иной класс, чем большинство. Очень близки по этому свойству – попадать в один класс – и геномы ископаемых рептилий (архозавры и лепидозавры).

Распределение геномов по классам.
N – число геномов в данном порядке;
I, II и III – классы разбиения

Порядок	N	I	II	III
Actinopterygii	510	464	46	0
Amphibia	65	40	17	8
Archosauria u Lepidosauria	177	1	176	0
Mammalia	212	0	1	211
Neoptera	143	0	4	139
Testudines	25	0	25	0

С точки зрения характера распределения по классам выделяются две группы организмов: рыбы и земноводные. И те, и другие в большинстве своём формируют первый класс (более никем не представленный). При этом земноводные являются единственным порядком, представители которого попадают во все три класса. Характер распределения геномов земноводных по трём классам также весьма неслучаен. Всего в порядке земноводных представлено 9 геномов хвостатых амфибий и 13 – бесхвостых. При этом все бесхвостые амфибии попали полностью во второй класс, а 7 из 9 геномов хвостатых – в третий.

Обращает на себя внимание также ещё один необычный факт: если при делении на два класса беспозвоночные полностью выделялись в отдельный класс, то при делении на три класса они фактически полностью объединяются с весьма эволюционно продвинутыми позвоночными (млекопитающими). Кроме того, черепаховые и ископаемые (эволюционно сравнительно близкие друг к другу) попадают в один класс при построении классификации из трёх классов.

Всё сказанное позволяет утверждать, что эволюция митохондриальных геномов очень тесно связана с эволюцией соматических геномов, несмотря на то, что физически эти два генома никак не связаны и обмена генетической информацией между ними не происходит. Данный факт позволяет использовать митохондриальные геномы не только как генетические маркеры (определяющие родство на сравнительно малом числе генераций), но и как эволюционные, позволяющие верифицировать пути формирования видов и иных таксономических разделов.

Список литературы

1. Sadovsky M.G. Comparison of real frequencies of strings vs. the expected ones reveals the information capacity of macromoleculae // Journal of Biological Physics. – 2003. – Vol. 29, № 1. – P. 23–38.

2. Sadovsky M.G. Information capacity of nucleotide sequences and its applications // Bulletin of Mathematical Biology. – 2006. – Vol. 68, № 2. – P. 156–178.

3. Горбань А.Н., Россиев Д.А. Нейронные сети на персональном компьютере. – Новосибирск: Наука. 1996. – 275 с.

4. Fukunaga K. Introduction to statistical pattern recognition. – 2nd ed. – Academic Press: London. 591 p.

5. Горбань А. Н, Попова Т. Г, Садовский М. Г. Классификация нуклеотидных последовательностей по частотным словарям обнаруживает связь между их структурой и таксономическим положением организмов. // Журнал общей биол. 2003. т.64, № 5. С. 16 – 21.

References

1. Sadovsky M.G. Comparison of real frequencies of strings vs. the expected ones reveals the information capacity of macromoleculae // Journal of Biological Physics, 2003, vol. 29, no. 1, pp. 23–38.

2. Sadovsky M.G. Information capacity of nucleotide sequences and its applications // Bulletin of Mathematical Biology, 2006, vol. 68, no. 2, pp. 156–178.

3. Gorban A.N., Rossiev D.A. Nejronnye seti na personal'nom komp'jutere [Neuron networks on PC]. Novosibirsk, Nauka. 1996. 275 p.

4. Fukunaga K. Introduction to statistical pattern recognition. 2nd ed. Academic Press: London. 591 p.

5. Gorban A.N., Popova T.G, Sadovsky M.G. Klassifikatzija nukleotidnykh posledovatel'nostej po chastotnym slovarjam obnaruzhivaet svjaz' mezhdu ikh strukturoj I taksonomicheskim polozheniem [Classification of nucleotide sequences over their frequency dictionaries reveals a relation between the structure of sequences and taxonomy of their bearers] Zhurnal obshchej biologii – Journal of general biology, 2003. v.64, no 5. pp. 16–21.

Рецензенты:

Денисенко В.В., д.ф.-м.н., ведущий научный сотрудник Института вычислительного моделирования СО РАН, г. Красноярск;

Заворуев В.В., д.б.н., профессор, ведущий научный сотрудник Института вычислительного моделирования СО РАН, г. Красноярск.

Работа поступила в редакцию 10.07.2014.