Processing of multidimensional data by means of technology TOFI

***Saule Kairollievna Sagnayeva**, cand., Associate Professor, L.N.Gumilyov Eurasian National University, ENU*

***Shynar Erlankyzy**, undergraduate, L.N.Gumilyov Eurasian National University, ENU*

The article deals with the analysis and visualization of multidimensional data Tofi technology means. We describe the technological features of the construction of the Tofi data warehouse, multidimensional cubes and slices of cubes. The main purpose of the Tophi technology is to support analytical activities, ad-hoc query users and analysts.

Keywords: OLAP, data warehousing, online analytical processing, TOFI technology.

УДК 519.95

СВЯЗЬ МЕЖДУ СТРУКТУРОЙ И ТАКСОНОМИЕЙ ГЕНОМОВ ХЛОРОПЛАСТОВ ХВОЙНЫХ

***Михаил Георгиевич Садовский**, доктор физико-математических наук*
Тел.: +79029904597, e-mail: msad@ism.krasn.ru
Институт вычислительного моделирования СО РАН
***Анна Игоревна Чернышова**, студентка 4 курс*
***Тел.:** +79620709211, e-mail: anna2121695@gmail.com*
Сибирский федеральный университет, ИФБиБТ

В работе представлены предварительные результаты по изучению связи между структурой и таксономией геномов хвойных хлоропластов, полученные методом динамических ядер. Данная связь была выявлена. Показаны особенности применяемых методов, а также представлены выводы по полученным результатам.

Ключевые слова. *ДНК, слово, частота, распределение, корреляция, эволюция, порядок.*

Введение


М.Г. Садовский


А.И. Чернышова

В современном мире особый интерес в биологическом мире представляет изучение структуры геномов. Данная работа посвящена изучению связи между структурой и таксономией геномов хвойных хлоропластов. Ранее уже были представлены результаты по выявлению связи между структурой и функцией биологических макромолекул [1]. В работе [2] была выявлена связь между структурой и таксономией геномов хлоропластов. Анонсируя результаты, скажем, что связь у хвойных хлоропластов также выявлена. Под структурой здесь подразумевается частотный словарь толщины три W_3 символьной последовательности, соответствующей ДНК. В свою очередь, под частотным словарём толщины 3 понимается список всех троек $\nu_1\nu_2\nu_3$ подряд идущих символов с указанием их частот. Всякий частотный словарь W_3 отображает геном в 64-мерное метрическое про-

странство. Близость двух геномов задаётся естественным образом. В данной работе использовалась Евклидова метрика.

Близость по таксономическому показателю определяется по морфологическим признакам, которые в свою очередь определяются нуклеарным (основным) геномом. Здесь мы пользовались традиционными схемами филогении и таксономии. Тот факт, что эти два типа близости определяются разными генетическими системами, которые физически независимы друг от друга, является прямым доказательством сильной синхронии в эволюции двух (независимых) генетических систем.

Материалы и методы

Генетические данные были взяты в GenBank банке. В нём было представлено 219 геномов. Релиз представлен от 24 июня 2015 года. Для исследований исходную базу данных сократили, чтобы получить равномерное распределение. В исследуемой базе осталось 97 геномов.

Существует несколько методов кластеризации. И все их можно разделить на две категории: линейные и нелинейные. В данной работе использовался линейный метод — метод динамических ядер (МДЯ) [3].

При построении кластеризации МДЯ одна из 64 частот исключалась. Это связано с тем, что сумма всех частот в словаре W_3 равна 1. Теоретически исключать можно любой триплет. В данной работе исключался тот триплет, для которого стандартное отклонение, наблюдаемое по той выборке, по которой проводилось исследование, являлось минимальным; в данном случае исключался триплет AGC. Построение классификации методом динамических ядер производилось в программе VidaExpert.

Результаты и обсуждение

При построении классификации методом динамических ядер всегда возникают две проблемы. Во-первых, метод динамических ядер не гарантирует единственности построения классификации, так как каждый раз её построение начинается с (нового) случайного разбиения точек на классы. Во-вторых, число классов (минимальное), на которые следует кластеризировать исследуемое множество данных, было установлено эмпирическим путем.

Следует отметить, что устойчивость разбиения на классы при разных значениях числа этих классов не гарантирована. К примеру, можно столкнуться с ситуацией, когда разбиение на четыре класса оказывается устойчивым (попадание большого числа геномов в один и тот же класс), а на пять — не устойчивым, а на шесть опять устойчивым. Один из вариантов решения этой проблемы: определять разделимость классов и объединять неразделимые. В данной работе этого не проводилось.

В результате измерений были получены две классификации: устойчивая и неустойчивая. Под устойчивым распределением в данном случае понимается следующее свойство: пусть имеется M реализаций метода динамических ядер с разными случайными начальными разбиениями геномов по классам. Финальная кластеризация будет устойчивой, если некоторые геномы, составляющие чётко выявляемую группу, статистически часто попадают в один и тот же класс при различных начальных распределениях.

В первом случае распределение устойчивое, во втором — неустойчивое.

Анализ результатов показал, что при устойчивом разбиении (рис.1.) вторая группа геномов на протяжении всей кластеризации не меняет свой состав. Кроме этого, ярко выделяется группа геномов, состоящая из *Nageia, Podocarpus, Retrophyllu*, которые при построении распределения на любое число классов всегда состоят в одном классе.

Прежде чем обсуждать неустойчивую часть множества геномов, заметим, что здесь может быть несколько типов неустойчивости. В нашем случае наблюдалась наименее слабая неустойчивость — те геномы, которые регулярно меняли свою принадлежность к классу при последовательной реализации МДЯ, тем не менее всегда принадлежали как целая группа одному и тому же классу. Иными словами, принадлеж-

ность группы к классу могла меняться, но сам по себе состав такой слабо неустойчивой группы оставался стабильным. При неустойчивой кластеризации (рис.2) также наблюдается группа геномов, которая на протяжении всей кластеризации держится вместе: *Abies, Cathaya, Keteleeria, Larix, Picea, Pseudotsuga.* При этом какая-то часть геномов появляется при переходе от двух классов к трём.

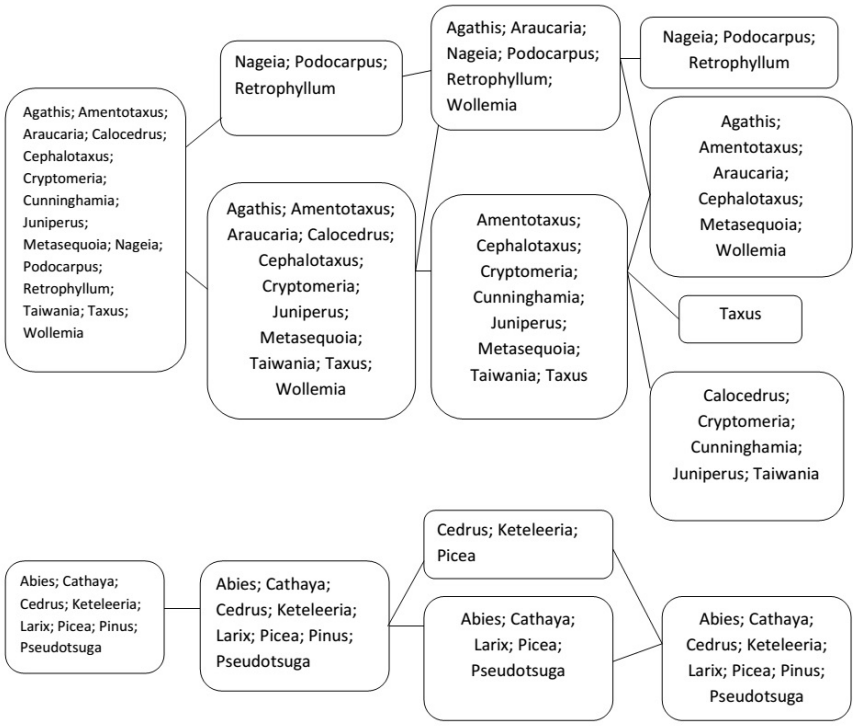

Рис.1. Устойчивое распределение таксонов низкого уровня по классам при построении классификации «снизу вверх» методом динамических ядер

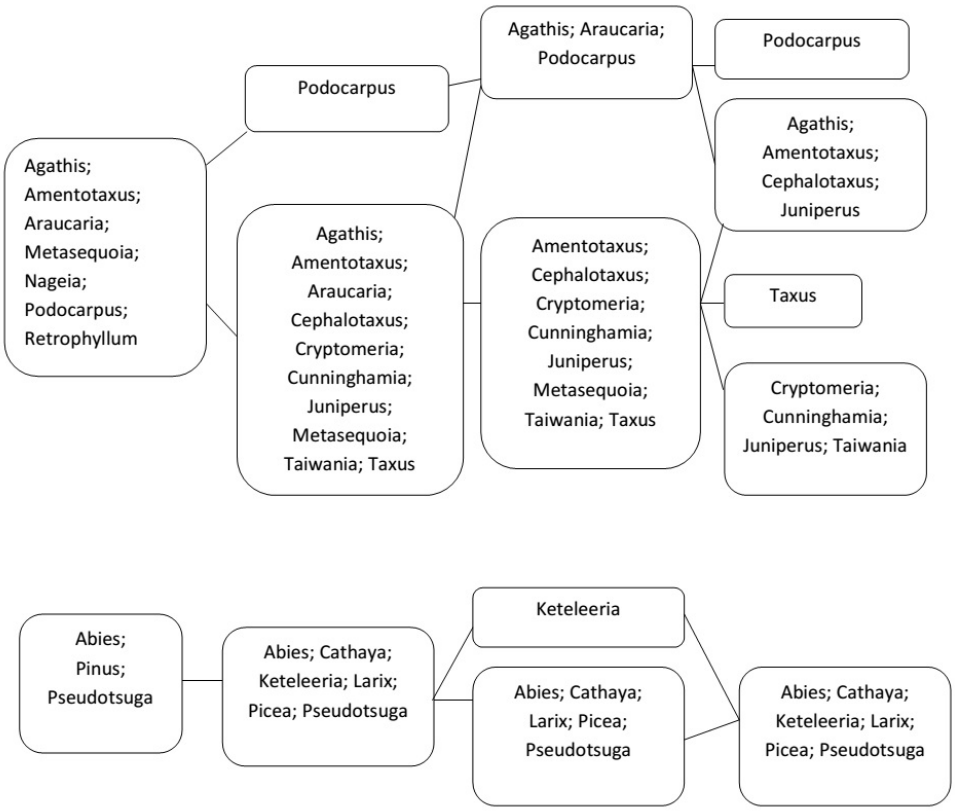

Рис.2. Неустойчивое распределение таксонов низкого уровня по классам при построении классификации «снизу вверх» методом динамических ядер

Заключение

Упорядоченное распределение видов и родов по классам, которые определялись лишь частотами триплетов в эволюции двух генетических систем — соматической и геномов хлоропластов. Показано существование высокого уровня синхронизации геномов хлоропластов и соматических геномов растений, несущих эти хлоропласты. Физически они никак друг с другом не связаны. Полученные результаты будут проверены при построении классификаций на множестве геномов более низкого таксономического уровня, например, на множестве геномов сосновых.

Литература

1. *Зайцева Н.А., Путинцева Ю.А., Садовский М.Г.* К проблеме связи структуры и систематики на примере митохондриальных геномов: труды X международной ФАМ-конференции. 2011. С. 152-156,

2. *Садовский М.Г., Чернышова А.И.* О связи структуры и таксономии геномов хлоропластов: материалы XIII Международнойя конференции по ФАМ и эвентологии многомерной статистики. 2014.

3. *Gorban A.N., Zinovyev A.Yu.* Visualization of data by method of elastic maps and its application in genomics, economics and sociology. *IHES Preprint,* 2001.

Towards the correspondence between structure of pine chloroplast genomes and their phylogeny

***Michael Sadovsky**, ICM SB RAS*

***Anna Chernyshova**, SFU*

Some results are presented exploring the problem of the relation between the phylogeny of various species and taxa, and the structure of corresponding DNA sequences. The features of the methods used are shown in this work. And also presents conclusions on the results obtained.

Keywords: DNA, string, frequency, distribution, correlation, evolution, order.

УДК 332.85; 519.233.5

ОЦЕНКА ВЛИЯНИЯ ВНЕШНИХ ФАКТОРОВ НА РОССИЙСКИЙ РЫНОК НЕДВИЖИМОСТИ

***Елена Валентиновна Смирнова,** д.ф.-м.н., проф.*
Тел.: 8 963 190 88 07, e-mail: seleval2008@yandex.ru
***Андрей Сергеевич Лем**, ассистент*
Тел.: 8 905 086 21 12, e-mail: alem@sfu-kras.ru
Сибирский федеральный университет
http://www.sfu-kras.ru

В статье описаны результаты анализа российской строительной отрасли с помощью оценки корреляции и дисперсии данных. Данное исследование доказывает, что метод корреляционной адаптометрии эффективен при исследовании рынка недвижимости и прогнозировании кризисов.

Ключевые слова: кризис, строительная отрасль, рынок недвижимости, корреляции, адаптация.