**О МЕТРИЧЕСКИХ СВОЙСТВАХ ЭВОЛЮЦИОННЫХ РАССТОЯНИЙ**

М.А. Мельчакова

*Новосибирский государственный университет*

mariya.melchakova@gmail.com

В.М. Ефимов

*Институт цитологии и генетики СО РАН*

[efimov@bionet.nsc.ru](mailto:efimov@bionet.nsc.ru)

Одним из способов изучения изменчивости биологических объектов является геометризация задачи: представление объектов точками в многомерном пространстве таким образом, чтобы расстояния между точками как можно лучше соответствовали различиям между объектами. Если различия между объектами являются евклидовыми расстояниями, то эта задача (с точностью до переноса, поворота и отражения) решается методами метрического шкалирования. В некоторых случаях к различиям можно применить монотонное преобразование, переводящее их в евклидовы расстояния. В остальных случаях приходится использовать другие методы, в частности, алгоритмы неметрического шкалирования.

Рассмотрены метрические свойства некоторых эволюционных дистанций для нуклеотидных последовательностей. Показано, что расстояния Джукса-Кантора и Кимуры не являются метриками. Введено новое расстояние – *-*дистанция. Показано, что для -дистанции, *-*дистанции и расстояния Джукса-Кантора существуют монотонные преобразования, приводящие к евклидовым метрикам и позволяющие применять алгоритмы метрического шкалирования.

Определения:

*Расстоянием* (или *различием*) называется функция на , если для всех выполняются условия [1, 2]:

1. (положительная определенность);
2. (симметричность);
3. (рефлексивность).

*Метрикой* называется расстояние, удовлетворяющее неравенству треугольника:

1. *d(x.y)*

*Расширенной метрикой* называется метрика, принимающая значение .

Топологическое пространство *монотонно вложимо* в метрическое пространство , если существует непрерывная строго возрастающая функция на такая, что – метрическое пространство, изометрически вложимое в .

*-дистанция* – наблюдаемая доля различающихся нуклеотидов для двух последовательностей одинаковой длины. *p*-дистанция является метрикой Хэмминга с точностью до домножения на длину последовательности [3].

*Расстояние Джукса-Кантора* – предполагаемое число замен нуклеотидов в двух последовательностях, происшедших от одного неизвестного предка за эволюционное время, в пересчете на одну позицию, вычисляемое как .

*Расстояние Кимуры* – предполагаемое число замен в пересчете на одну позицию, вычисляемое как , где – доля транзиций, – доля трансверсий.

Рассмотрим последовательности фиксированной длины . Заметим, что расстояния Джукса-Кантора и Кимуры определены не для всех значений для таких значений можем положить значение расстояния равным . Покажем, что и не являются метриками, так как для них не выполняется неравенство треугольника.

Запишем неравенство треугольника для расстояния Джукса-Кантора на последовательностях :

После очевидных преобразований получим:

Возьмем такие последовательности длины , что имеются различия только в позициях 1 и 2: —AA, —TT, —AT. В последнем неравенстве справа будет нуль, а слева ненулевое значение, что и доказывает предположение.

Для расстояния Кимуры проводим аналогичные рассуждения для случая, когда количество трансверсий , т.е. . Получаем неравенство , которое не выполнено для указанных последовательностей.

В подобных случаях, чтобы представить исследуемые объекты точками в многомерном пространстве, приходится прибегать к неметрическим методам [4]. Однако расстояние Джукса-Кантора функционально связано с *p*-дистанцией. Покажем, что существует монотонное вложение множества последовательностей с *p* -дистанцией в с евклидовой метрикой в некоторой системе координат. Такое представление даст нам возможность обращаться с данными (после необходимых преобразований), как с матрицей евклидовых расстояний, что расширяет область применимых к ней методов, добавляя, в частности, методы метрического шкалирования [2].

Каждый нуклеотид из последовательности кодируем следующим образом:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A |  | 0 | 0 | 0 |
| G | 0 |  | 0 | 0 |
| T | 0 | 0 |  | 0 |
| C | 0 | 0 | 0 |  |

Обозначим функцию, осуществляющую вложение. Непосредственной проверкой получаем, что , где – метрика Хэмминга. Следовательно, *p*-дистанция является квадратом евклидовой метрики. С другой стороны, .

Следствие*. По заданному расстоянию Джукса-Кантора можно вычислить p-дистанцию* *и, извлекая из нее корень, получить евклидову метрику.*

По аналогии с *p*-дистанцией, введем *-дистанцию*: , где . -дистанция, очевидно, удовлетворяет аксиомам расстояния. Кроме того, так как - дистанция является взвешенной суммой *p*-дистанций, для нее очевидно справедливо неравенство треугольника, следовательно, -дистанция является метрикой.

Покажем, что существует монотонное вложение множества последовательностей с -дистанцией в с евклидовой метрикой . Каждый нуклеотид из последовательности кодируем следующим образом:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| A |  | 0 | 0 | 0 |  | 0 |
| G | 0 |  | 0 | 0 |  | 0 |
| T | 0 | 0 |  | 0 | 0 |  |
| C | 0 | 0 | 0 |  | 0 |  |

Обозначим функцию, осуществляющую вложение. Вычисляя соответствующие значения, приходим к равенству , где – метрика Хэмминга. Следовательно, -дистанция является квадратом евклидовой метрики.

Таким образом, для -дистанции, *-*дистанции и расстояния Джукса-Кантора существуют монотонные преобразования, приводящие к евклидовым метрикам и позволяющие применять алгоритмы метрического шкалирования. Неизвестно, есть ли такое преобразование для расстояния Кимуры. Однако для каждой отдельной матрицы расстояний Кимуры можно эмпирически искать свое преобразование, приводящее к евклидовой метрике. Критерием является неотрицательность собственных чисел матрицы расстояний после двойного центрирования [2].

В частности, на примере данных статьи [5] в качестве такого преобразования было взят корень квадратный. Собственные числа матрицы двойного центрирования, построенной относительно новых значений, удовлетворяют критерию. Это говорит о том, что существует ее представление в евклидовом пространстве некоторой размерности. В частности, можно применять метод главных компонент. После применения алгоритма были выделены три главные оси, берущие на себя 48%, 19%, 9% информации соответственно.

Полученная конфигурация точек в трехмерном пространстве достаточно хорошо описывает филогенетическое дерево (см. рис.1) в смысле явного выделения 4 групп, соответствующих главным группам на дереве (см. рис.2-4). Пространственно эта конфигурация представляет собой двустороннюю вилку по два зубца с каждой стороны, причем зубья двух концов лежат в перпендикулярных плоскостях. Виды, не попавшие в 4 группы, расположились в середине между группами.

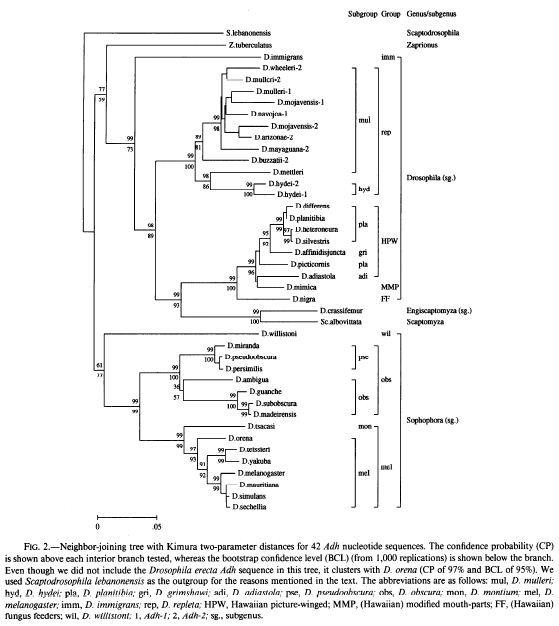


Рисунок Филогенетическое дерево, [5]

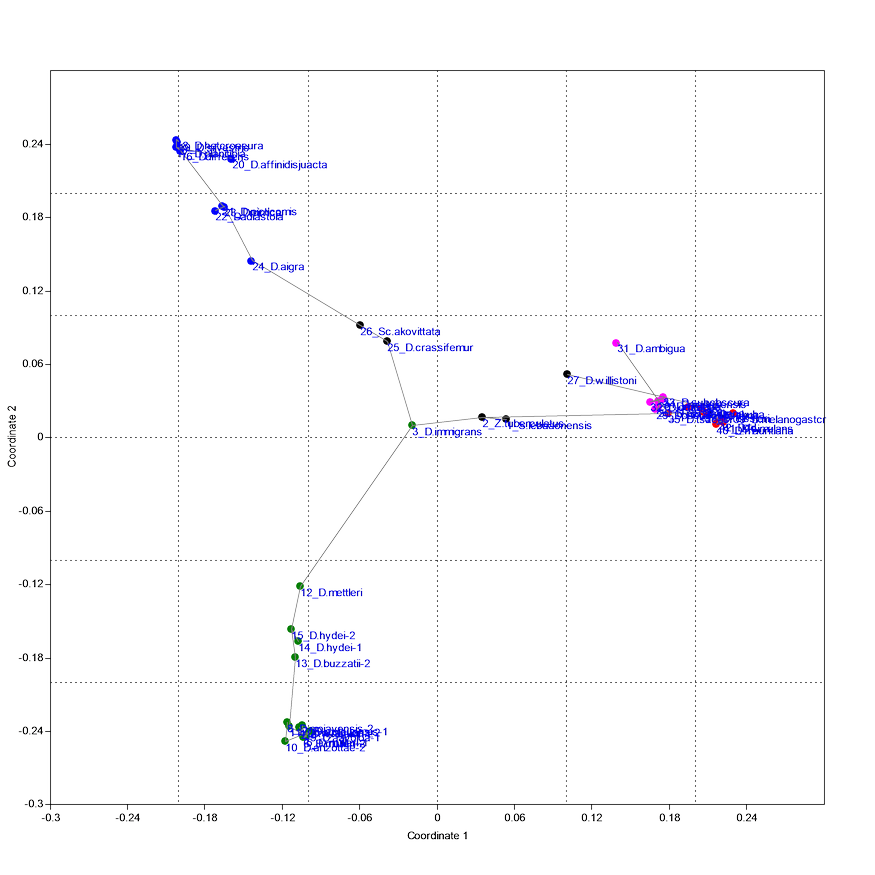


Рисунок Пространственное расположение видов дрозофилы, оси шкалирования 1,2

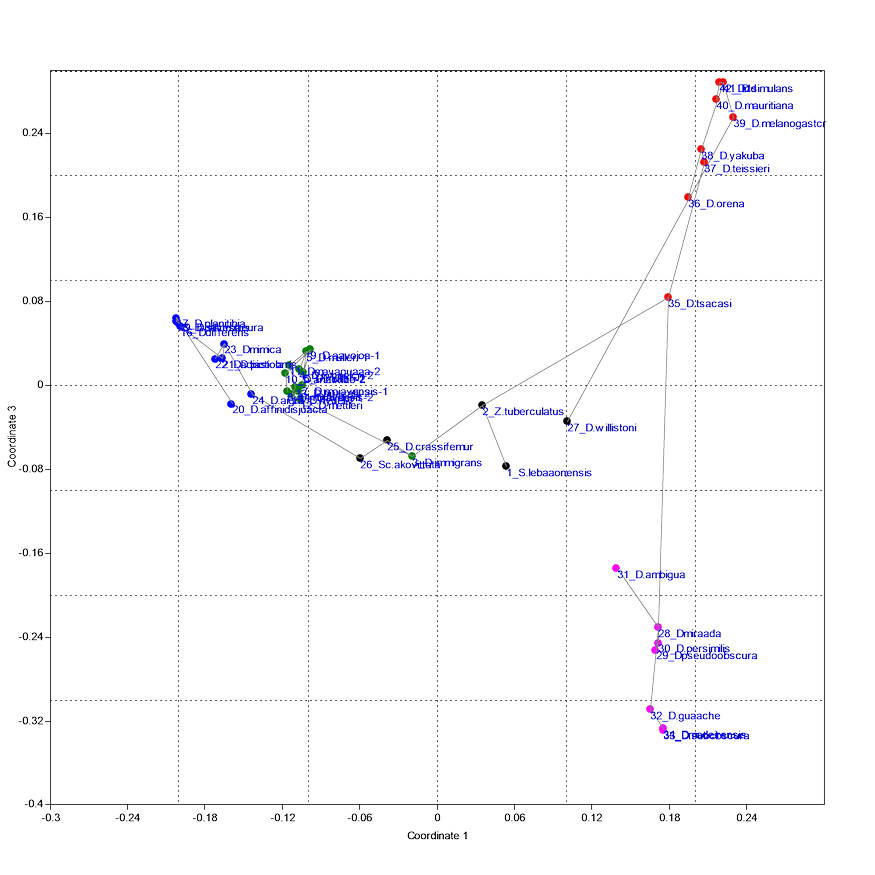


Рисунок Пространственное расположение видов дрозофилы, оси шкалирования 1,3

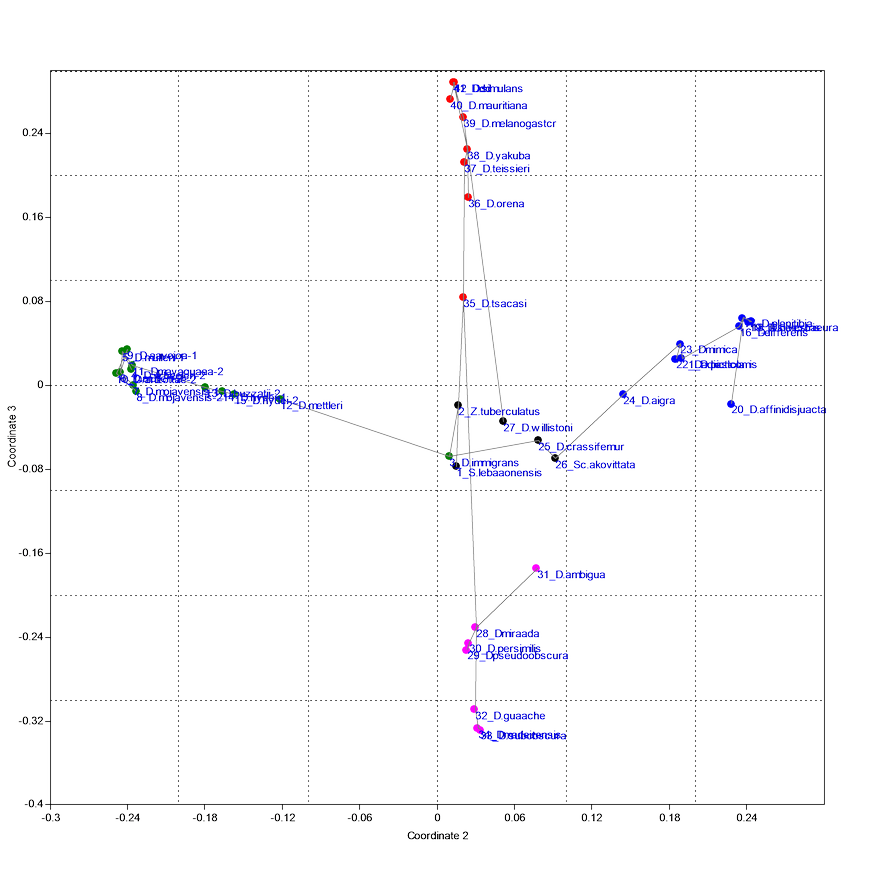


Рисунок Пространственное расположение видов дрозофилы, оси 2,3

ЛИТЕРАТУРА

[1]. Деза Е.И., Деза М.М. Энциклопедический словарь расстояний. М.: Наука. 2008. 444 с.

[2]. Courrieu P. Straight monotonic embedding of data sets in Euclidean spaces // Neural networks. 2002. V. 15. P. 1185–1196.

[3]. Лукашов В.В. Молекулярная эволюция и филогенетический анализ. М: БИНОМ. 2009. 256 с.

[4]. Shepard R.N. The analysis of proximities: multidimentional scaling with an unknown distance function.1. Psyhometrika. 1962. V. 27, №2, P. 125–140.

[5]. Russo C.A.M., Takezaki N., Nei M. Molecular phylogeny and divergence times of drosophilid species. Mol. Biol. Evol. 1995. V. 12. №3. P. 391–404.