УДК 57.065

ГЕОМЕТРИЧЕСКИЕ СВОЙСТВА ЭВОЛЮЦИОННЫХ ДИСТАНЦИЙ

© 2013 г. **В.М. Ефимов[1, 2, 3], М.А. Мельчакова[4], В.Ю. Ковалева[2]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: efimov@bionet.nsc.ru;
[2] Федеральное государственное бюджетное учреждение науки Институт систематики и экологии животных Сибирского отделения Российской академии наук, Новосибирск, Россия;
[3] Томский национальный исследовательский государственный университет, Томск, Россия;
[4] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

Поступила в редакцию 15 августа 2013 г. Принята к публикации 5 сентября 2013 г.

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. Рассмотрены метрические свойства некоторых широко известных эволюционных дистанций для нуклеотидных последовательностей. Показано, что расстояния Джукса–Кантора и Кимуры не являются метриками. Введено новое расстояние, аналог расстояния Кимуры, – *PQ*-дистанция. Показано, что *p*-дистанция и *PQ*-дистанция являются квадратами евклидовых метрик, названных в статье E_p-дистанцией и E_{PQ}-дистанцией соответственно. Применимость E_{PQ}-дистанции проиллюстрирована на взятом из GenBank множестве нуклеотидных последовательностей цитохрома *b* 12 видов мышевидных грызунов Западной Сибири и Алтая и сравнена с результатами использования *LogDet*-расстояния.

Ключевые слова: нуклеотидные последовательности, модели эволюции, филогенетические реконструкции, генетические расстояния, геометризация, зоологическая систематика.

ВВЕДЕНИЕ

Эволюционные дистанции (генетические расстояния) – это различия генетической информации двух организмов (например, частот аллелей, нуклеотидных или аминокислотных последовательностей и т. д.), возникшие после их дивергенции от общего предка. Эволюционные дистанции между последовательностями могут быть прямо интерпретированы как филогенетические отношения между формами жизни, от которых эти последовательности получены. Иначе говоря, чем меньше эволюционные дистанции между двумя последовательностями, тем вероятнее, что они имели недавнего общего предка и, соответственно, тем более они родственны друг другу (Лукашов, 2009).

В настоящее время под эволюционной дистанцией понимается исключительно число замен нуклеотидов/аминокислот в пересчете на одну позицию, произошедших за время независимой эволюции двух ДНК-последовательностей после их дивергенции от общего предка, или его оценка различными методами (Ней, Кумар, 2004). Мы ограничимся рассмотрением только нуклеотидных последовательностей. Эволюционные дистанции делятся на истинные, наблюдаемые и расчетные (Лукашов, 2009). Истинные дистанции, как правило, неизвестны, так как для анализа доступны только

ныне живущие или недавно жившие формы, а информация об общих предках отсутствует. Наблюдаемые дистанции, например *p*-дистанция, основаны только на различиях между имеющимися последовательностями. Считается, что они занижают истинные, так как не учитывают ни длину путей от общего предка (ни даже его наличие), ни возможные петли на этих путях (множественные замены, приводящие к той же самой последовательности). Поэтому чаще всего используются расчетные эволюционные дистанции, в которых делается поправка на возможную множественность замен и которые опираются на некоторые модели эволюции нуклеотидных последовательностей. Общим для многих моделей является предположение о существовании матрицы переходных вероятностей между нуклеотидами и постоянстве ее во времени. Примерами являются генетические расстояния Джукса–Кантора, Кимуры и т. п. (Felsenstein, 2003; Ней, Кумар, 2004). В более продвинутых моделях эволюции авторы учитывают непостоянство нуклеотидных замен во времени, например в LogDet-дистанции (Lockhart *et al*., 1994; Zharkikh, 1994).

Эволюционные дистанции исходно разрабатывались для реконструкции эволюционной истории видов и изучения природы и сути селективных сил, формирующих эволюцию генов и видов. На практике это свелось к построению филогенетических деревьев (филодендрограмм) и их применению в биологической систематике. В настоящее время более прогрессивным считается построение филодендрограмм через компьютерное моделирование эволюции самих молекулярных последовательностей (Felsenstein, 2003; Ней, Кумар, 2004).

Полученные на этом пути результаты уже привели к значительной перестройке всего здания биологической систематики (см., например, MSW – Mammal species …, 2005; Млекопитающие …, 2012). Однако «за бортом» остались вся предыдущая работа и весь опыт классических систематиков со времен К. Линнея, несмотря на то что первичное определение видов до сих пор производится на основе морфологии, экологии, географии и прочей доступной информации, и без этого первичного определения молекулярная филогенетика все еще обойтись не может.

По нашему мнению, дальнейшие возможности развития современной компьютерно-молекулярной филогенетики существенно ограничены двумя серьезными обстоятельствами – невозможностью присоединения информации других типов, в частности морфологических данных, и особенно представлением получаемых результатов исключительно в виде филодендрограмм. Вопреки широко распространенному мнению, филодендрограммы не являются единственным способом представления филогенетических взаимоотношений и могут быть дополнены, например, отображением взаимного расположения таксонов в многомерном пространстве (Klingenberg, Ekau, 1996; Revell, 2009; Klingenberg, Gidaszewski, 2010; Ковалева и др., 2012, 2013; Polly *et al*., 2013). Это можно сделать всегда, если отнестись к генетическим расстояниям как к обычным мерам сходства/различия и применить методы неметрического шкалирования (Shepard, 1962; Ковалева и др., 2013). Однако это неизбежно приведет к искажению взаимных расстояний, масштаб которого не всегда можно оценить. Было бы удобнее, если бы генетические расстояния сразу были метрическими расстояниями, т. е. удовлетворяли аксиомам метрики. Еще лучше, если бы они были евклидовыми расстояниями. Однако, несмотря на почти полувековую историю генетических расстояний, их геометрическим свойствам уделялось очень мало внимания.

ЯВЛЯЮТСЯ ЛИ ГЕНЕТИЧЕСКИЕ РАССТОЯНИЯ МЕТРИЧЕСКИМИ РАССТОЯНИЯМИ?

В математике неотрицательная действительная функция $d(x,y)$, определенная на множестве X, называется *метрикой*, если она удовлетворяет следующим условиям:

$d(x,y) = d(y,x)$ (аксиома симметрии)

$d(x,x) = 0 <=> x = y$ (аксиома тождества)

$d(x,y) <= d(x,z) + d(z,y)$ (аксиома треугольника).

Свойство неотрицательности $d(x,y) \geq 0$ вытекает из этих аксиом. Числовое значение функции $d(x,y)$ называется *расстоянием* между элементами x и y (Петровский, 2003). Выполнение аксиом метрики обеспечивает возможность помещения элементов множества X без искаже-

ния взаимных расстояний в некоторое геометрическое пространство и наделения точек этого множества координатами в этом пространстве. Это, в свою очередь, позволяет применять весь арсенал методов многомерного анализа для исследования соотношения внутри- и межвидовой изменчивости, визуализации возможных направлений эволюции, объединения данных различных типов, например молекулярных и морфологических, и оценки их конгруэнтности (Ковалева и др., 2012, 2013).

Простейшими примерами генетических расстояний являются (Felsenstein, 2003; Ней, Кумар, 2004):

***p*-дистанция** – наблюдаемая доля различающихся нуклеотидов для двух последовательностей одинаковой длины. *p*-дистанция является метрикой Хэмминга с точностью до домножения на длину последовательности (Hamming, 1950);

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

$d_{JC} = -\frac{3}{4}\ln(1-\frac{4}{3}p)$ (Jukes, Cantor, 1969);

двупараметрическое расстояние Кимуры – предполагаемое число замен в пересчете на одну позицию, вычисляемое как $d_{K2p} = -\frac{1}{2}\ln(1-2P-Q)-\frac{1}{4}\ln(1-2Q)$, где P – наблюдаемая доля транзиций, Q – наблюдаемая доля трансверсий (Kimura, 1980);

LogDet-расстояние – обобщение расстояния Джукса–Кантора на случай непостоянства вероятностей нуклеотидных замен во времени, вычисляемое как $d_{xy} = -\ln[\det F_{xy}]$, где x, y – последовательности, F_{xy} – матрица 4 × 4 частот совместной встречаемости пар нуклеотидов в каждой позиции для последовательностей x и y (Lockhart *et al.*, 1994).

Рассмотрим последовательности фиксированной длины m. Заметим, что расстояния Джукса–Кантора и Кимуры определены не для всех значений p, P, Q. Для таких значений можем положить значение расстояния, равное ∞. Покажем, что d_{JC} и d_{K2p} не являются метриками, так как для них не выполняется неравенство треугольника.

Запишем неравенство треугольника для расстояния Джукса–Кантора на последовательностях x, y, z:

$$-\frac{3}{4}\ln(1-\frac{4}{3}p(x,y)) \le$$
$$\le -\frac{3}{4}\ln(1-\frac{4}{3}p(x,z)) - \frac{3}{4}\ln(1-\frac{4}{3}p(z,y)).$$

После очевидных преобразований получим:

$$p(x,z) + p(z,y) - p(x,y) \ge \frac{4}{3}\,p(x,z)\,p(z,y).$$

Возьмем такие последовательности длины $m \ge 2$, в которых имеются различия только в позициях 1 и 2: x – AA, y – GG, z – AG. В последнем неравенстве слева будет нуль, а справа – положительное значение. Следовательно, расстояние Джукса–Кантора не является метрикой (Felsenstein, 2003).

Для расстояния Кимуры проводим аналогичные рассуждения для случая, когда доля трансверсий Q равна нулю, т. е. когда

$$d_{K2p} = -\frac{1}{2}\ln(1-2P).$$

Получаем неравенство

$$P(x,z) + P(z,y) - P(x,y) \ge 2\,P(x,z)\,P(z,y),$$

которое нарушается для указанных последовательностей. Следовательно, двупараметрическое расстояние Кимуры тоже не является метрикой (Мельчакова, Ефимов, 2011).

Заметим, что истинная эволюционная дистанция, т. е. число замен нуклеотидов в пересчете на одну позицию, является метрикой (Мельчакова, 2013). Поэтому то обстоятельство, что общеизвестные и широко используемые расчетные оценки этой дистанции, такие как расстояния Джукса–Кантора и Кимуры, теряют ее метрические свойства, вызывает некоторое недоумение и требует дальнейших исследований.

ЧТО ДОЛЖНЫ ОТРАЖАТЬ ЭВОЛЮЦИОННЫЕ ДИСТАНЦИИ?

Все расчетные генетические расстояния оценивают исключительно суммарное число замен нуклеотидов в пересчете на одну позицию, несмотря на то что для многопараметрических расстояний, начиная с расстояния Кимуры, предполагается, что вероятности замен различных типов (например, транзиций и трансверсий) тоже могут быть различными. «Поскольку транзиции в целом более вероятны, более часты, чем

трансверсии, т. е. занимают меньше времени, логично считать, что эволюционная дистанция между последовательностями с одной транзицией меньше, чем между последовательностями с одной трансверсией, а сами последовательности с одной транзицией более родственны друг другу (имеют более недавнего общего предка), чем последовательности с одной трансверсией» (Лукашов, 2009). Но в формулах многопараметрических расстояний это никак не отражено! В расстоянии Кимуры, например, расчетные частоты транзиций и трансверсий оцениваются раздельно по наблюдаемым частотам, а потом просто суммируются. Вообще говоря, логично было бы суммировать их с разными весами, причем вес трансверсий должен быть больше, чем вес транзиций. К сожалению, даже «взвешенное» расстояние Кимуры ни при каких весах не будет являться метрикой.

Заметим, что расстояние Джукса–Кантора получается монотонным преобразованием *p*-дистанции, которая относится к наблюдаемым дистанциям. По аналогии с *p*-дистанцией для «взвешенного» расстояния Кимуры можно предложить его наблюдаемый аналог – *PQ*-дистанцию:

$$d_{PQ} = P + (1 + \alpha)Q,\ \alpha \geq 0,$$

где P – наблюдаемая доля транзиций, Q – наблюдаемая доля трансверсий. Вопрос об их монотонной зависимости рассмотрен в следующем разделе.

Покажем, что и *p*-дистанция, и *PQ*-дистанция являются квадратами евклидовых расстояний. Если закодировать каждый нуклеотид в последовательностях набором чисел в соответствии с табл. 1 и для каждой пары последовательностей вычислить сумму квадратов разностей, т. е. квадрат евклидова расстояния, то, очевидно получим *p*-дистанцию с точностью до постоянного множителя. Если сделать то же самое в соответствии с табл. 2, то очевидно получим *PQ*-дистанцию, также с точностью до постоянного множителя. Отсюда следует, что разумнее всего извлекать из обеих дистанций квадратные корни, что приведет к евклидовым метрикам. Поэтому, по нашему мнению, именно их и следует использовать при применении геометрических методов для филогенетических реконструкций. Назовем их соответственно E_p-дистанция и E_{PQ}-дистанция.

Заметим, что в предположении постоянства нуклеотидных замен во времени *LogDet*-расстояние сводится к *p*-дистанции (Lockhart *et al.*, 1994; Zharkikh, 1994), следовательно, является ее обобщением. Соответственно, само *LogDet*-расстояние не евклидово, а его евклидовым аналогом является E_p-дистанция.

ПРИМЕНЕНИЕ К БИОЛОГИЧЕСКИМ ДАННЫМ

Для иллюстрации применимости предлагаемых нами евклидовых генетических расстояний – E_p-дистанции и E_{PQ}-дистанции – в качестве эмпирических данных были взяты 87 нуклеотидных последовательностей цитохрома *b* митохондриальной ДНК серых полевок (род *Microtus: **M. agrestis**, M. levis = M. rossiaemeridionalis, **M. oeconomus**, **M. gregalis***), лесных полевок (род *Myodes = Clethrionomys: My. Glareolus, My. **rufocanus**, My. **rutilus***), водяной полевки (род *Arvicola: **Arv. amphibious** = **Arv.***

Таблица 1

Кодировка нуклеотидов для *p*-дистанции

	A	G	T	C
A	$\frac{1}{\sqrt{2}}$	0	0	0
G	0	$\frac{1}{\sqrt{2}}$	0	0
T	0	0	$\frac{1}{\sqrt{2}}$	0
C	0	0	0	$\frac{1}{\sqrt{2}}$

Таблица 2

Кодировка нуклеотидов для *PQ*-дистанции

	A	G	T	C	A/G	T/C
A	$\frac{1}{\sqrt{2}}$	0	0	0	$\sqrt{\frac{\alpha}{2}}$	0
G	0	$\frac{1}{\sqrt{2}}$	0	0	$\sqrt{\frac{\alpha}{2}}$	0
T	0	0	$\frac{1}{\sqrt{2}}$	0	0	$\sqrt{\frac{\alpha}{2}}$
C	0	0	0	$\frac{1}{\sqrt{2}}$	0	$\sqrt{\frac{\alpha}{2}}$

terrestris), полевых мышей (род *Apodemus: A. agrarius*, *A. peninsulae*), домовой мыши (род *Mus: Mus musculus*), серой крысы (род *Rattus: R. norvegicus*) длиной 1138 п.н. из базы данных GenBank, ранее использованные в работе В.Ю. Ковалева с соавт. (2012). Для всех последовательностей с помощью пакетов MEGA5 (Tamura *et al*., 2011) и Excel вычислены матрицы расстояний Джукса–Кантора и Кимуры, а также *LogDet*-дистанций, E_p-дистанций и E_{PQ}-дистанций (при $\alpha = 1$). В табл. 3 приведены попарные коэффициенты корреляции между этими матрицами, полученные с помощью теста Мантеля (Mantel, 1967; Mantel, Valand, 1970). Видно, что на данном эмпирическом материале все дистанции отражают фактически одно и то же. На рис. 1 показана зависимость расчетных генетических расстояний от их наблюдаемых аналогов. Монотонная зависимость расстояния Джукса–Кантора от E_p-дистанции следует из определяющих их формул и ожидаема для *LogDet*-расстояния. Монотонная зависимость расстояния Кимуры от E_{PQ}-дистанции из формул не следует, но из графика видно, что их эмпирическая зависимость, невзирая на небольшие отклонения, выглядит точно так же. С точки зрения геометрического подхода E_p и E_{PQ}-дистанции наиболее удобны для применения, так как являются евклидовыми метриками.

Далее для матрицы E_{PQ}-дистанций был применен один из методов многомерного шкалирования – метод главных координат (Torgerson, 1952), включенный в пакет Jacobi4 (Ефимов и др., 2011). Из рис. 2 видно, что внутривидовая изменчивость незначительна на фоне межвидовой и ею можно пренебречь. Однако на взаимное расположение видов может влиять разный объем выборок. Поэтому для всех видов по главным координатам были вычислены их выборочные центроиды и между ними – матрица евклидовых расстояний. Соответственно, все виды получили равные веса. (Заметим, что без геометрического подхода это сделать не так просто.) К полученной матрице расстояний снова был применен метод главных координат (рис. 3, 4). Первая главная координата четко отвечает за различия между семействами Cricetidae и Muridae. Хорошо видна родовая структура, совпадающая с принятой на сегодня зоологической классификацией (MSW – Mam-

Таблица 3

Тест Мантеля
для матриц эволюционных дистанций

r	*JC*	*K2p*	E_p	E_{pq}	LogDet
JC	1,000	0,999	0,961	0,978	0,999
K2p	0,999	1,000	0,962	0,978	0,999
E_p	0,961	0,962	1,000	0,995	0,963
E_{pq}	0,978	0,978	0,995	1,000	0,978
LogDet	0,999	0,999	0,963	0,978	1,000

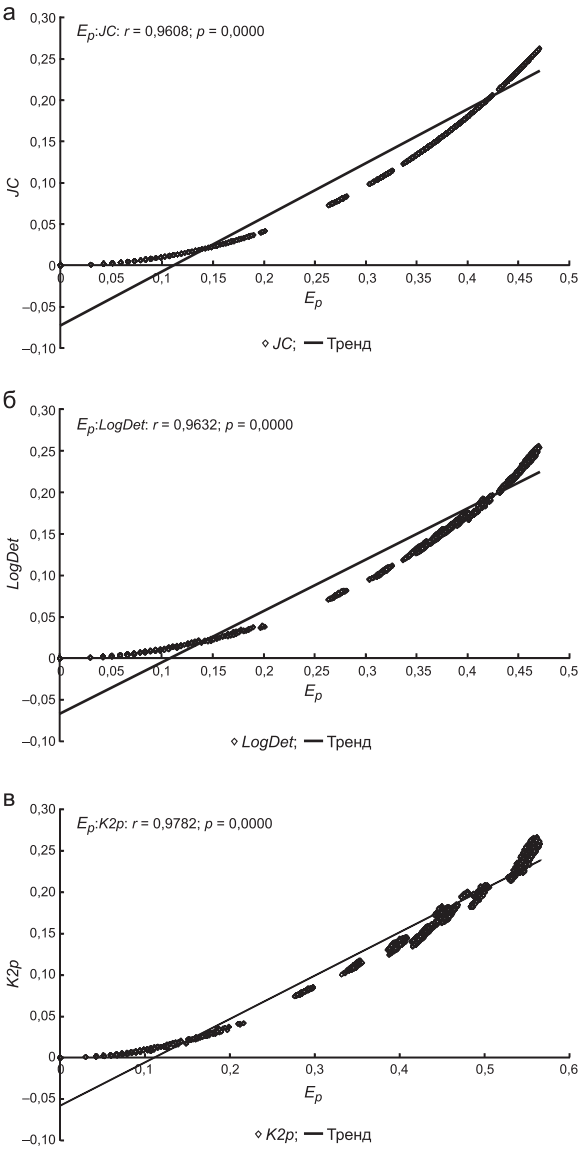

Рис. 1. Зависимость расчетных генетических расстояний от их наблюдаемых аналогов.

а – расстояние Джукса–Кантора от E_p-дистанции; б – *LogDet*-расстояние от E_p-дистанции; в – расстояние Кимуры от E_{PQ}-дистанции.

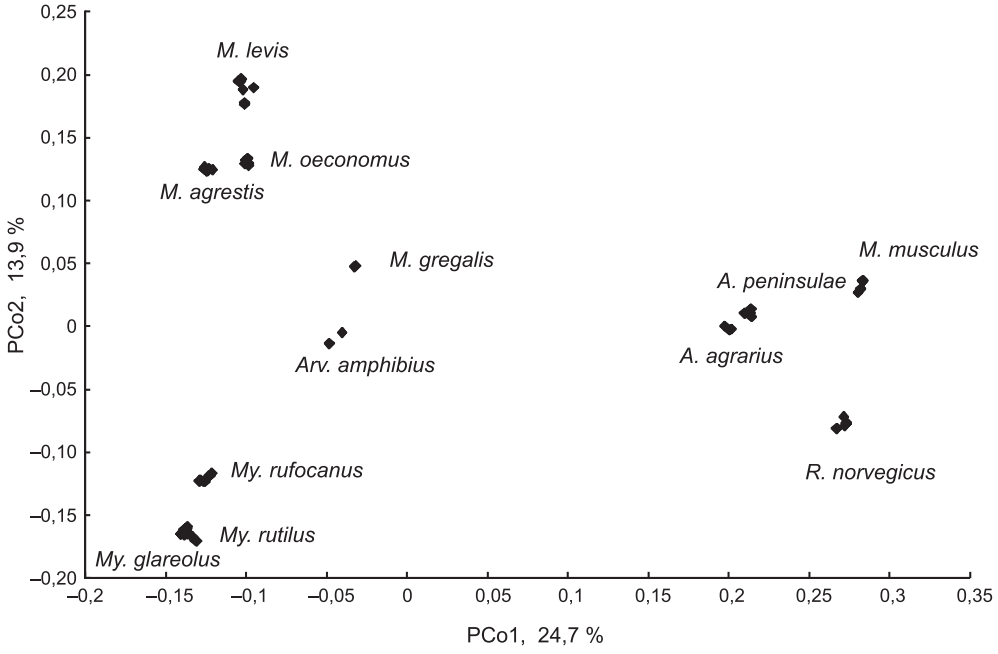

Рис. 2. Расположение нуклеотидных последовательностей на плоскости I–II главных координат.

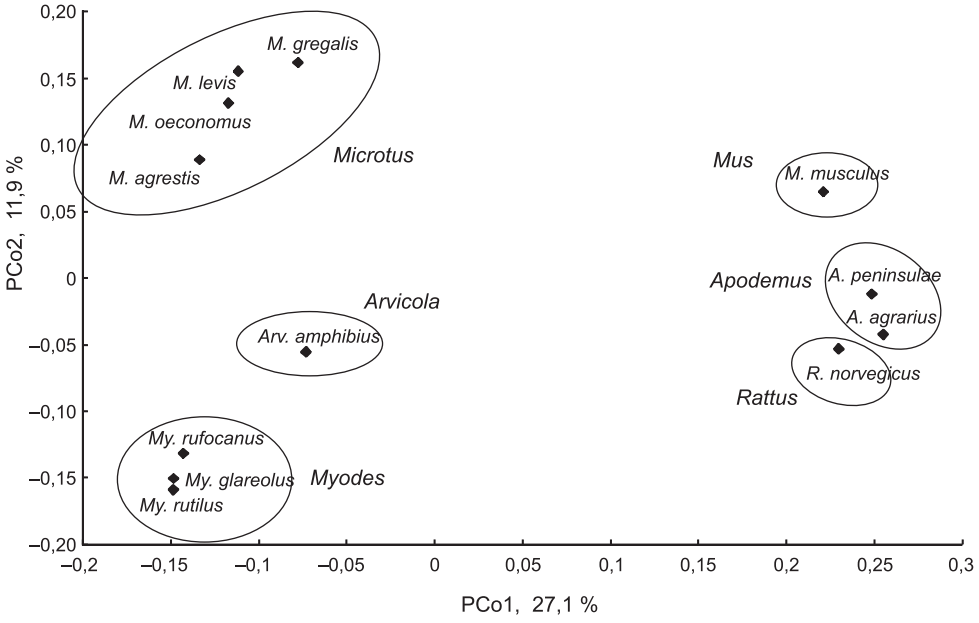

Рис. 3. Расположение видов на плоскости I–II главных координат.

mal species …, 2005), за исключением одного вида – *M. gregalis*, который заметно отличается по третьей главной координате (рис. 4) от других видов рода *Microtus*. Обособленность *M. gregalis* видна и на филодендрограммах, полученных методом UPGMA (невзвешенного попарного среднего) по матрице евклидовых расстояний между видами (рис. 5) и по матрице *LogDet*-расстояний между исходными последовательностями (бутстреп = 1000) (рис. 6). Это совпадает с полученными ранее результатами на основании неметрического шкалирования матрицы расстояний Кимуры (Ковалева и др., 2012) и согласуется с современными тенденциями в зоологической систематике полевок (Абрамсон, Лисовский, 2012).

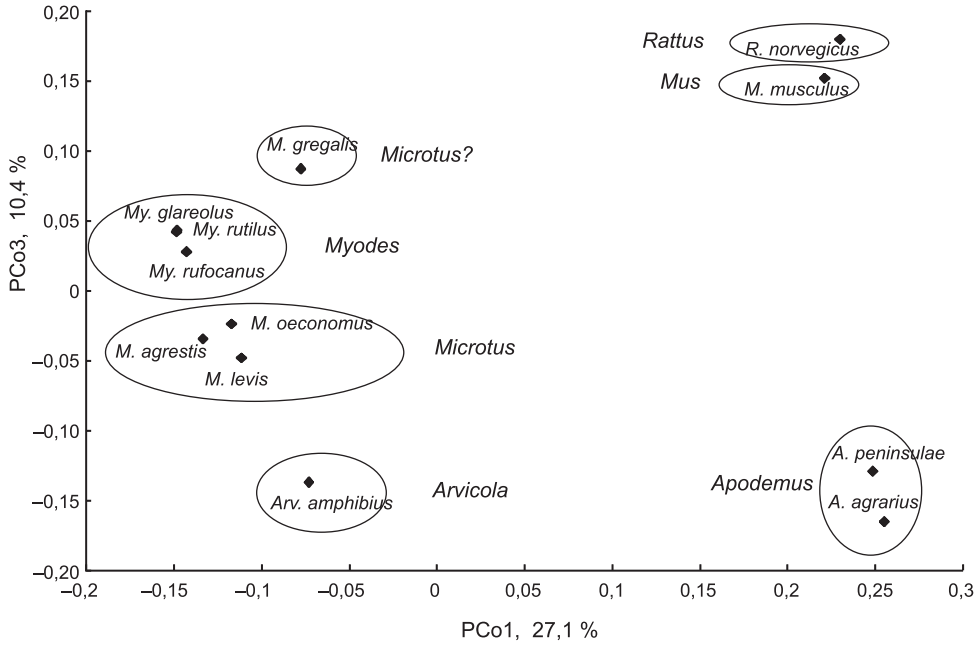

Рис. 4. Расположение видов на плоскости I–III главных координат.

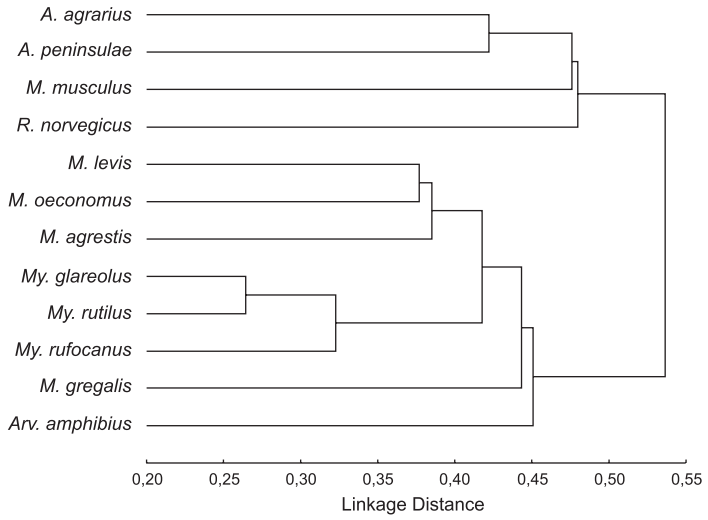

Рис. 5. Филодендрограмма, полученная методом UPGMA (невзвешенного попарного среднего) по матрице евклидовых расстояний между видами.

ЗАКЛЮЧЕНИЕ

Почти все известные генетические расстояния обладают двумя существенными недостатками: даже при разных вероятностях замен нуклеотидов оценивают только суммарное число замен; не являются геометрическими расстояниями. В работе предложены евклидовы аналоги расстояний Джукса–Кантора, *LogDet* и Кимуры – E_p-дистанция и E_{PQ}-дистанция. В E_{PQ}-дистанции числа транзиций и трансверсий взяты с разными весами. Евклидовость предложенных расстояний позволяет дополнительно применять весь арсенал методов многомерного анализа. На реальных данных проиллюстрировано практическое использование E_{PQ}-дистанции. Показана высокая корреляция всех пяти расстояний.

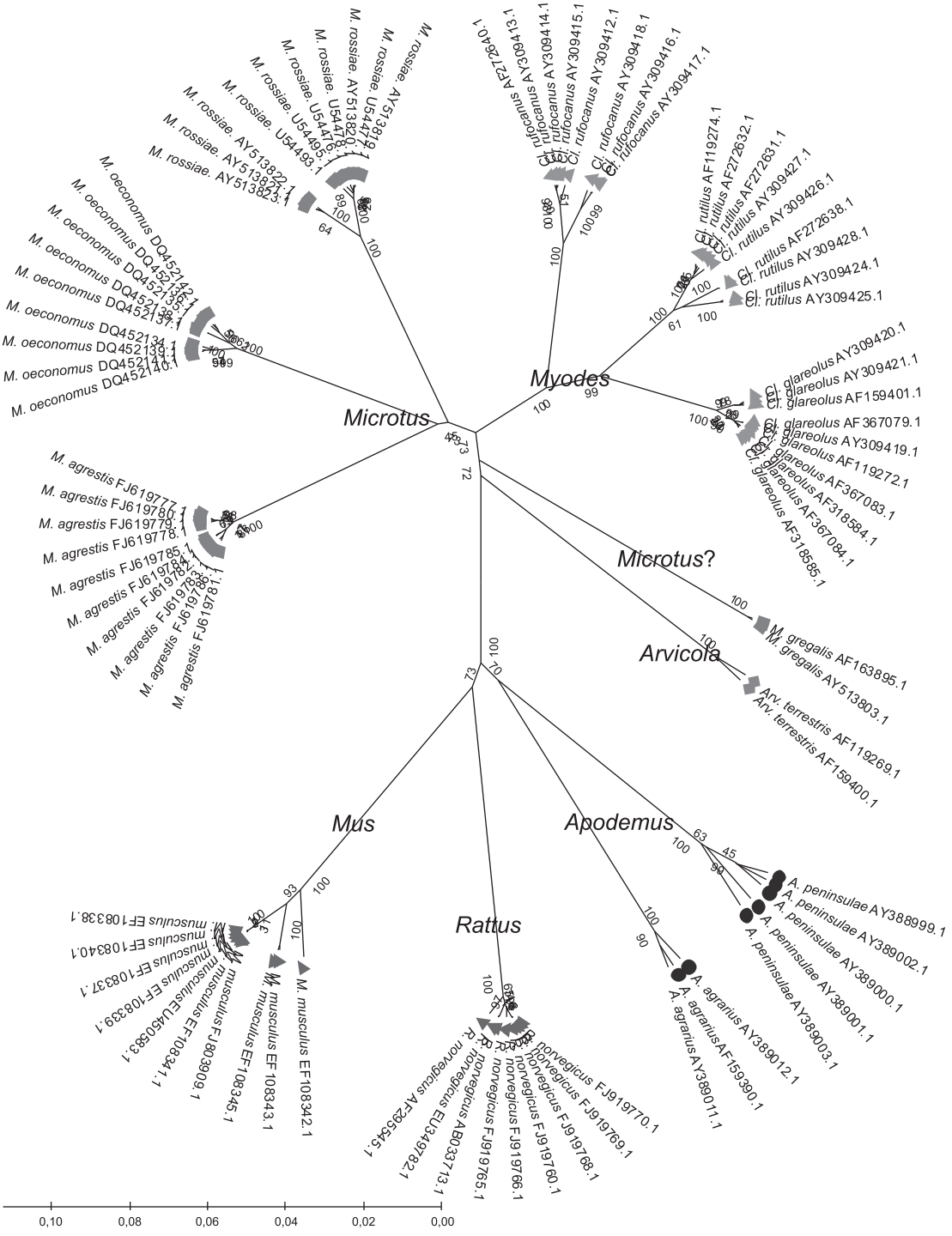

Рис. 6. Филодендрограмма, полученная методом UPGMA (невзвешенного попарного среднего) по матрице *LogDet*-расстояний между исходными последовательностями.

БЛАГОДАРНОСТИ

Работа выполнена при финансовой поддержке Программы Президиума РАН – Интеграция РАН (6.8 и 28), Президента РФ (НШ-5278 2012.4), Интеграционного проекта СО РАН (18.13), Проекта фундаментальных исследований СО РАН и УрО РАН (70) и РФФИ (11-04-00141а, 13-07-00315а).

ВИДЫ И НОМЕРА ПОСЛЕДОВАТЕЛЬНОСТЕЙ ЦИТОХРОМА *b* мтДНК В БАЗЕ ДАННЫХ GENBANK

M. agrestis – FJ619777–FJ619786; *M. levis* – AY513819–AY513823, U54476–U54478, U54493, U54495; *M. oeconomus* – DQ452134–DQ452142; *M. gregalis* – AF163895, AY513803; *Arv. terrestris* – AF119269, AF159400; *My. glareolus* – AF119272, AF159401, AF318584, AF318585, AF367079, AF367083, AF367084, AY309419–AY309421; *My. rufocanus* – AF272640, AY309412–AY309418; *My. rutilus* – AF119274, AF272631, AF272632, AF272638, AY309424–AY309428; *A. agrarius* – AF159390, AY389011, AY389012; *A. peninsulae* – AY388999, AY389000–AY389003; *Mus musculus* – EF108337 – EF108343, EF108345, EU450583, FJ803909; *R. norvegicus* – AB033713, AF295545, EU349782, FJ919760, FJ919765, FJ919766, FJ919768– FJ919770.

ЛИТЕРАТУРА

Абрамсон Н.И., Лисовский А.А. Полевочьи // Млекопитающие России: систематико-географический справочник / Ред. И.Я. Павлинов, А.А. Лисовский. М.: КМК, 2012. С. 220–276.

Ефимов В.М., Штайгер И.А., Полунин Д.А. и др. Программно-алгоритмический комплекс для многомерного анализа микрочиповых данных // II Междунар. науч.-практ. конф. «Постгеномные методы анализа в биологии, лабораторной и клинической медицине: геномика, протеомика, биоинформатика». Новосибирск, Россия, 14–17 ноября, 2011. С. 120.

Ковалева В.Ю., Абрамов С.А., Дупал Т.А. и др. Анализ соответствия и комбинирование молекулярно-генетических и морфологических данных в зоологической систематике // Изв. РАН. Сер. биол. 2012. Вып. 4. С. 404–414.

Ковалева В.Ю., Литвинов Ю.Н., Ефимов В.М. Землеройки (Soricidae, Eulipotyphla) Сибири и Дальнего Востока: комбинирование и поиск конгруэнтности молекулярно-генетических и морфологических данных // Зоол. журнал. 2013. Т. 92. Вып. 11. С. 1–15.

Лукашов В.В. Молекулярная эволюция и филогенетический анализ. М.: БИНОМ, Лаборатория знаний, 2009. 256 с.

Мельчакова М.А. Геометрические аналоги генетических расстояний: Магистерская диссертация. Новосибирск: НГУ, 2013. 33 с.

Мельчакова М.А., Ефимов В.М. О метрических свойствах эволюционных расстояний // Тез. докл. конф. «Соврем. пробл. математики, информатики и биоинформатики», посвящ. 100-летию А.А. Ляпунова, 11–14 окт. 2011 г. Новосибирск, 2011. С. 88.

Млекопитающие России: систематико-географический справочник / Ред. И.Я. Павлинов, А.А. Лисовский. М.: КМК, 2012. 604 с.

Ней М., Кумар С. Молекулярная эволюция и филогенетика. Киев: КВЩ, 2004. 418 с.

Петровский А.Б. Пространства множеств и мультимножеств. М.: Едиториал УРСС, 2003. 248 с.

Felsenstein J. Inferring phylogenies. Sunderland: Sinauer Associates, 2003. 664 p.

Hamming R.W. Error detecting and error correcting codes // Bell Syst. Tech. J. 1950. V. 29. No. 2. P. 147–160.

Jukes T.H., Cantor C.R. Evolution of protein molecules // Mammalian Protein Metabolism / Ed. H.N. Munro. N.Y.: Acad. Press, 1969. P. 21–132.

Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences // J. Mol. Evol. 1980. V. 16. No. 2. P. 111–120.

Klingenberg C.P., Ekau W. A combined morphometric and phylogenetic analysis of an ecomorphological trend: pelagization in Antarctic fishes (Perciformes: Nototheniidae) // Biol. J. Linn. Soc. 1996. V. 59. No. 2. P. 143–177.

Klingenberg C.P., Gidaszewski N.A. Testing and quantifying phylogenetic signals and homoplasy in morphometric data // Syst. Biol. 2010. V. 59. No. 3. P. 245–261.

Lockhart P.J., Steel M.A., Hendy M.D., Penny D. Recovering evolutionary trees under a more realistic model of sequence evolution // Mol. Biol. Evol. 1994. V. 11. No. 4. P. 605–612.

Mammal Species of the World: a Taxonomic and Geographic Reference / Eds D.E. Wilson, D.M. Reeder. 3rd ed. Baltimore: J. Hopkins Univ. Press, 2005. 2142 p. Available at http://www.departments.bucknell.edu/biology/resources/msw3/ browse.asp

Mantel N. The detection of disease clustering and a generalized regression approach // Cancer Res. 1967. V. 27. P. 209–220.

Mantel N., Valand R.S. A technique of nonparametric multivariate analysis // Biometrics. 1970. V. 26. P. 547–558.

Polly P.D., Lawing A.M., Fabre A.C., Goswami A. Phylogenetic principal components analysis and geometric morphometrics // Hystrix, the Italian J. Mammalogy. 2013. V. 24. No. 1. P. 1–9.

Revell L.J. Size-correction and principal components for interspecific comparative studies // Evolution. 2009. V. 63. P. 3258–3268.

Shepard R.N. The analysis of proximities: multidimensional scaling with an unknown distance function. 1 // Psychometrika. 1962. V. 27. No. 2. P. 125–140.

Tamura K., Peterson D., Peterson N. *et al*. MEGA5: molecular evolutionary genetics analysis using maximum likelihood; evolutionary distance; and maximum parsimony methods // Mol. Biol. Evol. 2011. V. 28. P. 2731–2739.

Torgerson W.S. Multidimensional scaling: I. Theory and method // Psychometrika. 1952. V. 17. No. 4. P. 401–419.

Zharkikh A. Estimation of evolutionary distances between nucleotide sequences // J. Mol. Evol. 1994. V. 39. P. 315–329.

GEOMETRIC PROPERTIES OF EVOLUTIONARY DISTANCES

V.M. Efimov[1, 2, 3], M.A. Melchakova[4], V.Yu. Kovaleva[2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, e-mail: efimov@bionet.nsc.ru;
[2] Institute of Systematics and Ecology of Animals SB RAS, Novosibirsk, Russia;
[3] Tomsk National Research State University, Tomsk, Russia;
[4] Novosibirsk National Research State University, Novosibirsk, Russia

Summary

One way to study the variability of biologic objects is their geometrization: the objects are presented by points in a multidimensional space in such a way that the distances between the points would be best consistent with the dissimilarities between objects. If the dissimilarities between the objects are Euclidean distances, this task (up to translation, rotation and reflection) is solved by metric scaling. We consider the metric properties of some well-known evolutionary distances of nucleotide sequences. It is shown that the Jukes-Cantor and Kimura distances are not metrics. We introduce a new Kimura distance analog, the *PQ*-distance. It is shown that the p and *PQ* distances are the squares of Euclidean metrics named E_p-distance and E_{PQ}-distance, respectively. The applicability of the E_{PQ} distance is illustrated by the example of a cytochrome b sequence set of 12 rodent species from West Siberia and Altai, taken from the GenBank, and compared with the results of the use of the *LogDet*-distance.

Key words: nucleotide sequences, evolution models, phylogenetic reconstructions, genetic distances, geometrization, zoological systematics..