МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2023, том 57, № 2, с. 178–184

БИОИНФОРМАТИКА И СИСТЕМНАЯ КОМПЬЮТЕРНАЯ БИОЛОГИЯ

УДК 577.2:599.32

ГЕОМЕТРИЧЕСКИЙ ПОДХОД К ФИЛОГЕОГРАФИЧЕСКОМУ АНАЛИЗУ МОЛЕКУЛЯРНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ: ГЛАВНЫЕ КОМПОНЕНТЫ И ДЕНДРОГРАММЫ

© 2023 г. В. М. Ефимов[a, b, c, d, *], К. В. Ефимов[e], В. Ю. Ковалева[b]

[a] *Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, 630090 Россия*

[b] *Институт систематики и экологии животных Сибирского отделения Российской академии наук, Новосибирск, 630091 Россия*

[c] *Новосибирский государственный университет, Новосибирск, 630090 Россия*

[d] *Томский государственный университет, Томск, 634050 Россия*

[e] *Высшая школа экономики, Москва, 101000 Россия*

**e-mail: efimov@bionet.nsc.ru*

Поступила в редакцию 27.07.2022 г.
После доработки 27.07.2022 г.
Принята к публикации 21.08.2022 г.

Поиски проявлений отбора, вызванного влиянием среды, в молекулярных последовательностях обычно проводят внутри близкородственных видов или на внутривидовом уровне, поскольку считается, что на высоких таксономических уровнях такие поиски бесперспективны из-за филогенетического родства. Аминокислотные последовательности цитохрома b 67 видов грызунов и зайцеобразных с известными географическими координатами оцифрованы с использованием базы данных AAindex. На основе более 200 тыс. признаков получены главные компоненты. Использован ранее не применявшийся для таких задач известный статистический метод, позволяющий ортогонально разложить многомерную изменчивость на внутри- и межтаксонную и анализировать их по отдельности. Выбран уровень подсемейства. Найдена корреляция второй главной компоненты (17.05% межтаксонной изменчивости) с широтой ($r = 0.561$; $n = 67$; $p < E{-}5$). Выявляемое первой главной компонентой (39.48% межтаксонной изменчивости) четкое разделение на две группы, не совпадающее с таксономическим, указывает на возможную физико-химическую подоплеку различий между ними. Это требует дальнейших исследований.

Ключевые слова: Rodentia, цитохром b, евклидовы расстояния, кластерный анализ, ортогональное разложение, географические координаты

DOI: 10.31857/S0026898423020052, **EDN:** EGBWKZ

ВВЕДЕНИЕ

Исторически сложилось так, что положение организмов в системе живой природы определялось их внешним подобием друг другу, прежде всего сходством строения (классический подход). С появлением молекулярной биологии и компьютеров положение организмов стало почти полностью определяться генетическим сходством (молекулярно-биологический подход). Оба подхода оперируют разными типами исходных данных (континуальные морфологические признаки и символьные последовательности соответственно), что предопределило их различное представление при компьютерном анализе, а также характер используемых методов. В классическом подходе – это геометрическое пространство, одномерное или многомерное, осями которого являются признаки, и в которое помещены объекты. Различия между объектами отображаются расстояниями в этом пространстве. На сегодняшний день основными способами анализа считаются линейные многомерные методы типа главных компонент (PCA, MR, DA, CCA и их PLS-аналоги) и нелинейные – типа неметрического шкалирования (NMDS, t-SNE, UMAP). В молекулярно-биологическом подходе сначала превалировал кластерный анализ на основе матрицы расстояний между последовательностями (UPGMA, NJ, ME), а сейчас чаще используется прямое моделирование филогенетического дерева (MP, ML). В результате всегда получаются дендрограммы, т.е. деревья кластеров. По сути, речь идет о противопоставлении непрерывного и дискретного подходов.

Мы полагаем, что оба подхода не противоречат друг другу и их можно и желательно использо-

вать совместно. Разница в типах исходных данных не является препятствием. Поскольку имеется матрица расстояний, последовательности всегда можно представить точками в многомерном пространстве и, следовательно, описать геометрически. Обязательно нужно учитывать дискретность. На любом шаге кластеризации все объекты разделены на несколько непересекающихся кластеров. Для каждого кластера можно вычислить центроид. Для каждого объекта можно вычислить отклонение от центроида своего кластера. В результате общая изменчивость распадется на межвыборочную и объединенную внутривыборочную изменчивость. Все варианты изменчивости можно исследовать с помощью обоих подходов – и классического (непрерывного, геометрического), и молекулярно-биологического (дискретного, кластерного).

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Материалы. Из GenBank (NCBI) [1] взяты аминокислотные последовательности митохондриального белка цитохрома b (Cytb), по одной для каждого вида млекопитающих отряда Rodentia Северного полушария Старого Света, у которых в описании присутствовала опция “lat_lon”, всего 61 вид. Дополнительно в качестве внешней группы взяты шесть последовательностей семейства Ochotona отряда Lagomorpha. Все 67 последовательностей были выравнены с помощью пакета MEGA 11 [2]. Последняя 380-я позиция из-за неполноты данных удалена. Во избежание разночтений использовали таксономическую принадлежность последовательностей, указанную в GenBank. Кроме того, из базы данных AAindex взяты 553 индекса физико-химических свойств 20 протеиногенных аминокислот (далее АА-индексы) [3].

Методы. Все АА-индексы были предварительно стандартизированы (центрированы и нормированы). Для каждого АА-индекса в матрице аминокислотных последовательностей каждая аминокислота заменена на соответствующее значение физико-химического свойства и по полученной числовой матрице рассчитана матрица квадратов межвидовых евклидовых расстояний размера 67 × 67. Все 553 матрицы квадратов расстояний суммированы, и в суммарной матрице из каждого элемента извлечен квадратный корень. Эта процедура эквивалентна вычислению евклидовых расстояний между 67 видами по 209 587 (379 × 553) признакам. Далее по матрице межвидовых евклидовых расстояний методом главных координат вычислены главные компоненты [4]. Для однорангового разбиения на кластеры выбран уровень подсемейств, поскольку даже на родовом уровне молекулярная систематика млекопитающих все еще не устоялась. По матрице главных компонент вычислены центроиды всех подсемейств и для каждого вида вычислено отклонение от центроида своего подсемейства. Этот способ предложил Фишер в методах ANOVA и MANOVA [5, 6]. Идея состоит в том, что из общей изменчивости вычитается межгрупповая и остается внутригрупповая изменчивость, в которой межгрупповая отсутствует. Обе изменчивости по построению отображаются в пространства, ортогональные друг другу. В нашем случае общая матрица главных компонент раскладывается на межтаксонную и объединенную внутритаксонную (таксоны – подсемейства). Для каждой новой матрицы снова вычислена матрица евклидовых расстояний между видами и ее главные компоненты. Сумма квадратов расстояний между двумя видами в новых матрицах равна квадрату расстояния между ними в общей матрице, а главные компоненты новых матриц в совокупности ортогональны друг другу. Таким образом получается ортогональное разложение общей межвидовой матрицы расстояний на две новые, каждая размером 67 × 67. Одна характеризует изменчивость между подсемействами, точнее, между их центроидами, а вторая – изменчивость, которая получится из общей, если совместить центроиды всех 10 подсемейств и собрать вместе их внутривыборочную изменчивость. По всем трем матрицам расстояний методом Варда проведен кластерный анализ, построены дендрограммы и тем самым получено ортогональное разложение общей дендрограммы на две другие, межтаксонную и объединенную внутритаксонную. Точно так же для всех трех матриц главных компонент вычислены корреляции с внешними факторами, в данном случае, широтой и долготой. Расчеты проведены с помощью пакетов MEGA 11 [2], PAST4 [7] и Jacobi4 [8].

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Из табл. 1 видно (столбцы Sum, последняя строка), что доля всей межтаксонной изменчивости в общей составляет 68.3% (71.32/104.48). Видно также (столбцы λ), что первые компоненты общей изменчивости почти полностью соответствуют межтаксонной изменчивости. Суммарное число главных компонент не изменилось (9 + 56 = 65), но произошло перераспределение дисперсий.

Из табл. 2 видно, что высокую и достоверную корреляцию с широтой ($p < 10^{-5}$; $N = 67$) демонстрируют вторые главные компоненты общей и межтаксонной изменчивости.

Очевидно, что это одна и та же корреляция, но в данном случае мы видим, что она целиком относится именно к межтаксонной изменчивости и никак не проявляет себя во внутритаксонной.

Таблица 1. Дисперсии главных компонент для трех матриц межвидовых расстояний

All	λ	λ, %	Sum	Sum, %	Inter	λ	λ, %	Sum	Sum, %	Intra	λ	λ, %	Sum	Sum, %
PC1	28.33	27.11	28.33	27.11	PC1	28.15	39.48	28.15	39.48	PC1	2.66	8.01	2.66	8.01
PC2	12.35	11.82	40.68	38.94	PC2	12.16	17.05	40.31	56.53	PC2	2.17	6.54	4.83	14.56
PC3	9.08	8.69	49.76	47.63	PC3	8.79	12.32	49.10	68.85	PC3	2.05	6.20	6.88	20.76
PC4	7.53	7.20	57.28	54.83	PC4	7.29	10.22	56.39	79.07	PC4	2.00	6.04	8.89	26.80
PC5	5.43	5.19	62.71	60.02	PC5	4.77	6.68	61.16	85.76	PC5	1.74	5.26	10.63	32.06
PC6	4.33	4.15	67.04	64.17	PC6	4.07	5.71	65.23	91.47	PC6	1.65	4.99	12.28	37.05
PC7	2.72	2.60	69.76	66.77	PC7	2.35	3.29	67.58	94.76	PC7	1.49	4.50	13.78	41.55
...	...	...	...	...	PC8	2.05	2.88	69.63	97.64	...	...	...	...	...
PC65	0.01	0.01	104.48	100.00	PC9	1.68	2.36	71.32	100.00	PC56	0.01	0.04	33.16	100.00

Примечание. All– общая, Inter – межтаксонная, Intra – объединенная внутритаксонная.

Таблица 2. Коэффициенты корреляции (×1000) с широтой (lat) и долготой (lon) первых 9 главных компонент трех матриц межвидовых расстояний

PC	Corr	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
All	lat	–69	538*5	–11	–220	82	–12	–61	–38	217
	lon	–122	73	118	53	–178	154	–208	219	66
Inter	lat	76	561*5	–11	–263	42	–61	–30	–26	33
	lon	–139	–63	81	–56	–179	–166	278	63	69
Intra	lat	–165	–110	99	81	–309*	–75	–114	92	217
	lon	–78	128	–61	116	180	–114	–194	305*	83

Примечание: All – общая, Inter – межтаксонная, Intra – объединенная внутритаксонная. *$p < 0.05$; *5$p < 10^{-5}$.

Корреляции на первом уровне значимости можно не принимать во внимание.

На рис. 1*а*–*в* приведено расположение видов на плоскости главных компонент всех трех вариантов изменчивости. Учитывая, что с широтой коррелирует именно вторая компонента, возможно, что на рис. 1*а*, *б* мы наблюдаем последствия давнего отбора под влиянием среды, причем в двух расходящихся направлениях. Ochotona расположилась, хотя и близко, но все же за пределами отряда Rodentia. Что касается первой компоненты, на которую приходится более четверти общей межтаксонной изменчивости, то ее содержательный смысл еще предстоит выяснить.

Объединенная внутритаксонная изменчивость (рис. 1*в*) "размазана" на много компонент с маленькими дисперсиями. Можно предположить, что главное направление изменчивости внутри каждого таксона формируется так, чтобы меньше пересекаться с остальными таксонами, живущими на той же территории. В частности, подсемейство Arvicolinae явно образовало свою собственную компоненту. При этом один вид (*Arvicola amphibius*) вообще отделился от всех в другую сторону.

На рис. 1*г* четко видна ортогональность внутри- и межтаксонной изменчивости. Показаны только первые компоненты, но такая же картина получается на всех рисунках, на которых противопоставлены внутри- и межтаксонные главные компоненты (не приведено).

Ортогональное разложение общей матрицы расстояний позволяет ортогонально разложить не только главные компоненты на блоки, соответствующие разным вариантам изменчивости, но и дендрограммы (рис. 2*а*–*в*). Поскольку по каждой матрице расстояний каким-либо алгоритмом кластерного анализа можно вычислить дендрограмму, мы автоматически получаем ортогональное разложение дендрограммы, отражающей общую изменчивость (1*а*), на две дендрограммы, отражающие межтаксонную и внутритаксонную изменчивость по отдельности (1*б* и 1*в* соответственно).

Очевидно, что на рис. 1*а*–*в* отражена та же самая многомерная изменчивость, что и на рис. 2*а*–*в*, только первые показывают взаимное расположение видов в пространстве и направления их изменчивости, а вторые – сходство видов и их объединений между собой. Таким образом, главные компоненты и дендрограммы дополняют друг

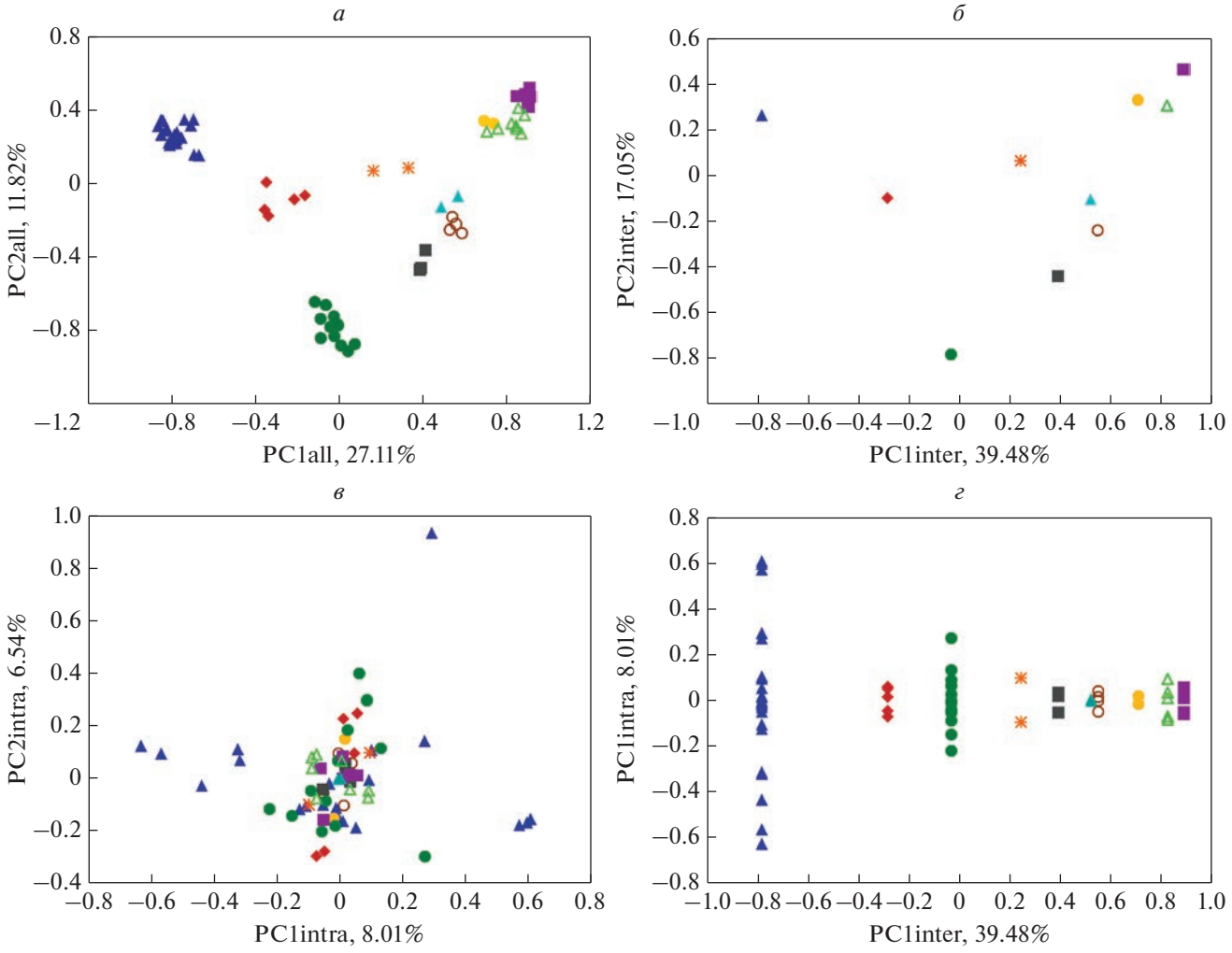

Рис. 1. Расположение видов на плоскости главных компонент. *а* – Первые две компоненты общей изменчивости; *б* – первые две компоненты межтаксонной изменчивости; *в* – первые две компоненты внутритаксонной изменчивости; *г* – первые компоненты межтаксонной и внутритаксонной изменчивости. ○ Allactaginae; ▲ Arvicolinae; ◆ Cricetinae; ■ Dipodinae; ● Leithiinae; ● Murinae; ■ Ochotonidae; ▲ Sicistinae; ✕ Spalacinae; △ Xerinae.

друга, позволяя составить более точное представление об изучаемой совокупности объектов. Преимущество дендрограмм заключается в том, что они легко отображают сходство объектов в пространстве любой размерности, просто ее игнорируя, а компоненты, из-за особенностей человеческого восприятия, приходится рассматривать попарно, максимум, по три. С другой стороны, дендрограммы не способны адекватно отобразить корреляцию с внешними факторами, из-за этого приходится наносить кластеры прямо на географические карты. Это наглядно, но непригодно для дальнейшей обработки.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Основной результат – это высокая статистически значимая корреляция второй главной компоненты (17.05% межтаксонной изменчивости) с широтой и отсутствие этой корреляции у первой главной компоненты со значительно большей дисперсией (39.48%). Безусловно, нужно учитывать, что исследуемый массив признаков, более 200 тыс., строго говоря, не представляет именно молекулярную изменчивость, хотя и получен оцифровкой набора молекулярных последовательностей. Считается, что физико-химические свойства аминокислот по своей природе ближе не столько к структуре, сколько к функции белков, тем более, что цитохром b участвует, наряду с другими митохондриальными белками, в самой важной биохимической функции: обеспечении организма энергией.

Митохондриальный ген *cytb* в настоящее время является основой молекулярной таксономии организмов. Долгое время он считался селектив-

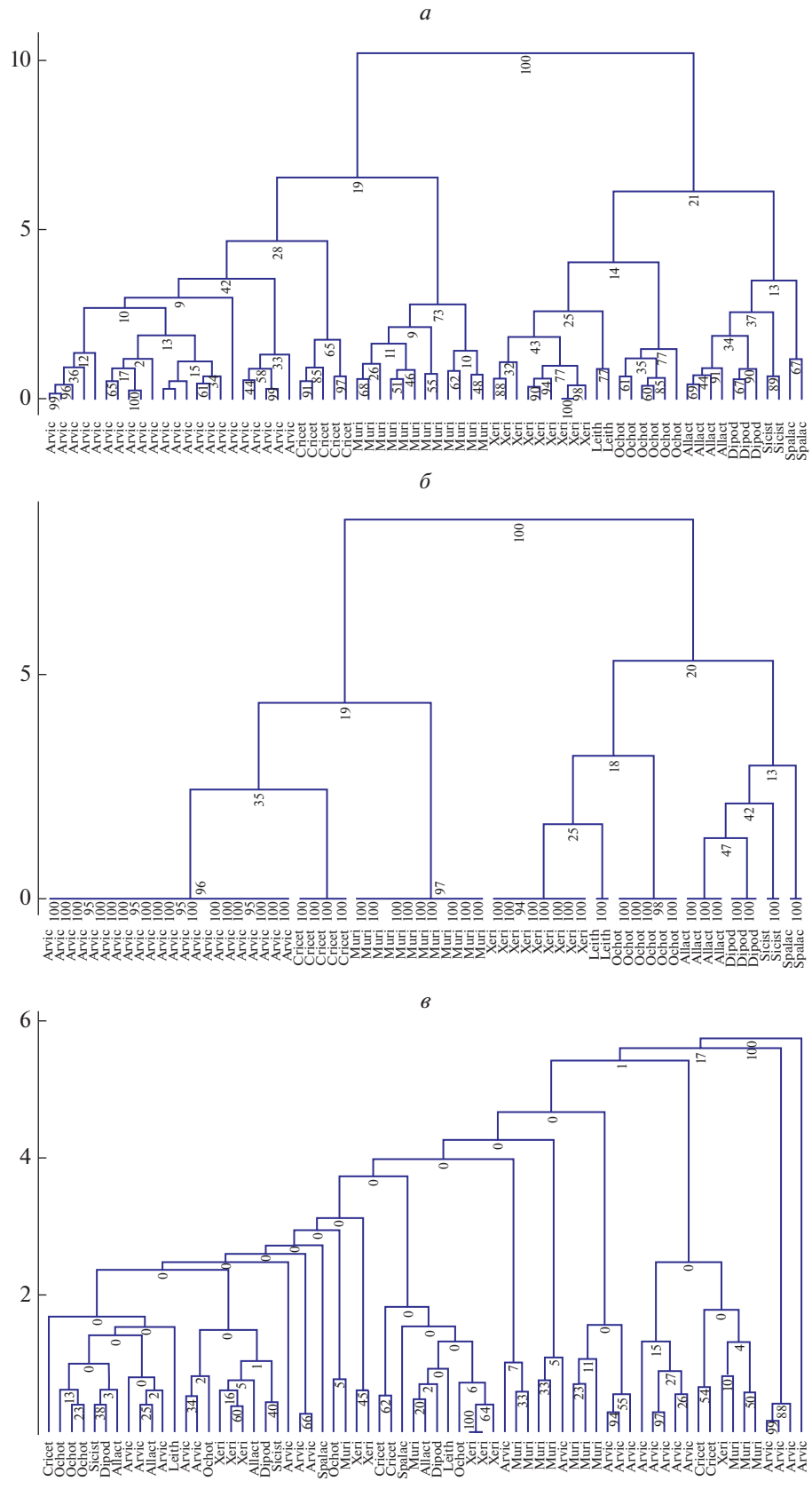

Рис. 2. Ортогональное разложение общей дендрограммы (*а*), характеризующей сходство аминокислотных последовательностей цитохрома b, на межтаксонную (*б*) и объединенную внутритаксонную (*в*) дендрограммы.

но нейтральным. Однако в последнее время это утверждение все чаще подвергается сомнению и выявляются адаптивные компоненты изменчивости кодируемых им аминокислотных последовательностей цитохрома b [9–12].

Несмотря на все разнообразие адаптаций к широкому спектру экологических условий, имеющихся в классе млекопитающих, существуют базовые механизмы функционирования организма, общие для всего класса в целом. Из этого следует, что при адаптации к сходным условиям можно ожидать некоторого сходства эволюционных путей в различных группах, например, перестройки энергетического метаболизма системы клеточного дыхания при адаптации к условиям гипотермии. Это сходство может, в частности, проявляться в одних и тех же мутациях в аминокислотных последовательностях разных видов. В таком случае логично ожидать, что такие мутации должны вызывать и сходные изменения физико-химических свойств самих аминокислотных последовательностей. Обычно вычисляют соотношения синонимичных и несинонимичных различий между нуклеотидными последовательностями или анализируют радикальные аминокислотные замены (т.е. замену на аминокислотный остаток с радикально иными физико-химическими свойствами). Для учета внешних факторов, на которые предположительно идет отбор, обычно подбирают пары контрастных условий, например, высокогорье и равнина, подземный и наземный образ жизни, пресноводные и соленые водоемы и т.д. Для общего представления о близости последовательностей практически всегда вычисляют филогенетические деревья либо на основе матрицы расстояний и алгоритмов кластерного анализа, либо прямым моделированием множественных деревьев и выбирают наиболее подходящее методами максимальной парсимонии или максимального правдоподобия. Считается, что поиски проявлений адаптивной эволюции на высоких таксономических уровнях бесперспективны, так как в этом случае и сходство, и различия между видами очевидно определяются, в первую очередь, степенью их филогенетического родства. Поэтому поиски проявлений адаптивной эволюции в молекулярных последовательностях обычно ведутся внутри близкородственных видов или даже на внутривидовом уровне.

Давно известное в классической математической статистике ортогональное разложение общей изменчивости на межвыборочную и внутривыборочную позволяет искать эти проявления на любых таксономических уровнях. То, что корреляция с широтой (а это, безусловно, следствие корреляции со средой) нашлась на уровне подсемейств, подтверждает перспективность подобных поисков. Более того, выявляемое первой главной компонентой четкое разделение на две группы (рис. 1*а*, 2*а*), не совпадающее с таксономическим, указывает на возможную физико-химическую подоплеку различий между ними. Это требует дальнейших исследований.

Работа выполнена в рамках бюджетного проекта Института цитологии и генетики СО РАН FWNR-2022-0019 Министерства науки и высшего образования Российской Федерации.

Выражаем искреннюю признательность рецензенту и редактору выпуска за полезные замечания.

Процедур с участием людей и животных не было.

Авторы заявляют об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. NCBI Resource Coordinators. (2015) Database resources of the national center for biotechnology information. *Nucl. Acids Res.* **43**, D6–D17.
2. Tamura K., Stecher G., Kumar S. (2021) MEGA11: molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **38**, 3022–3027.
3. Kawashima S., Pokarowski P., Pokarowska M., Kolinski A., Katayama T., Kanehisa M. (2008) AAindex: amino acid index database progress report 2008. *Nucl. Acids Res.* **36**, D202–D205.
4. Gower J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*. **53**, 325–338.
5. Fisher R.A. (1919) XV. – The correlation between relatives on the supposition of Mendelian inheritance. *Earth Env. Sci. Transactions Royal Soc. Edinburgh*. **52**, 399–433.
6. Fisher R.A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics*. **7**, 179–188.
7. Hammer Ø., Harper D.A.T., Ryan P.D. (2001) PAST: paleontological statistics software package for education and data analysis. *Palaeontologia Electronica*. **4**, 1–9.
8. Polunin D., Shtaiger I., Efimov V. (2019) JACOBI4 software for multivariate analysis of biological data. *bioRxiv*. 803684.
9. Da Fonseca R.R., Johnson W.E., O'Brien S.J., Ramos M.J., Antunes A. (2008) The adaptive evolution of the mammalian mitochondrial genome. *BMC Genomics*. **9**, 1–22.
10. Abramson N.I., Bodrov S.Y., Bondareva O.V., Genelt-Yanovskiy E.A., Petrova T.V. (2021) A mitochondrial genome phylogeny of voles and lemmings (Rodentia: Arvicolinae): evolutionary and taxonomic implications. *PLoS One*. **16**, e0248198.
11. Bondareva O., Genelt-Yanovskiy E., Petrova T., Bodrov S., Smorkatcheva A., Abramson N. (2021) Signatures of adaptation in mitochondrial genomes of Palearctic subterranean voles (Arvicolinae Rodentia). *Genes*. **12**, 1945.
12. Mori S., Matsunami M. (2018) Signature of positive selection in mitochondrial DNA in Cetartiodactyla. *Genes Genet. Systems*. 17-00015.

Geometric Approach to Phylogeographic Analysis Molecular Genetic Sequences: Principal Components and Dendrograms

V. M. Efimov[1, 2, 3, 4, *], K. V. Efimov[5], and V. Yu. Kovaleva[2]

[1]*Institute of Cytology and Genetics, Siberian Branch, Russian Academy of Sciences, Novosibirsk, 630090 Russia*
[2]*Institute of Animal Systematics and Ecology, Siberian Branch, Russian Academy of Sciences, Novosibirsk, 630091 Russia*
[3]*Novosibirsk State University, Novosibirsk, 630090 Russia*
[4]*Tomsk State University, Tomsk, 634050 Russia*
[5]*Higher School of Economics, Moscow, 101000 Russia*
**e-mail: efimov@bionet.nsc.ru*

Currently, the search for manifestations of selection under the influence of the environment in molecular sequences is usually carried out within closely related species or at the intraspecific level. It is believed that at high taxonomic levels this is unpromising due to phylogenetic relationship. Cytochrome b amino acid sequences of 67 rodent and lagomorph species with known geographic coordinates were digitized using the AAindex database. Based on more than 200 thousand features, the main components were obtained. A well-known statistical method, which has not previously been used for such problems, was used, which makes it possible to orthogonally decompose multidimensional variability into intra- and intertaxon variability and analyze them separately. Subfamily level selected. For the second principal component (17.05% of intertaxon variability), a correlation with latitude was found ($r = 0.561$; $n = 67$; $p < E{-}5$). The clear division into two groups revealed by the first principal component (39.48% of intertaxon variability), which does not coincide with the taxonomic one, indicates a possible physicochemical underlying reason for the differences between them. This requires further research.

Keywords: Rodentia, cytochrome b, Euclidean distances, cluster analysis, orthogonal decomposition, geographic coordinates