Математическая биология и биоинформатика
2016. Т. 11. № 2. С. 336–350. doi: 10.17537/2016.11.336

========== **БИОИНФОРМАТИКА** ==========

УДК: 004.942

Применение числовой характеристики строя нуклеотидов в геномах прокариот для реклассификации внутри рода *Rickettsia*

Шпынов С.Н.[*1], Гуменюк А.С.[2], Поздниченко Н.Н.[2]**

[1]*ФНИЦЭМ им. Н.Ф. Гамалеи МЗ РФ, Москва*
[2]*ОмГТУ, Омск, Россия*

Аннотация. С целью разработки нового подхода для классификации прокариот геномы представителей семейства Rickettsiaceae были проанализированы с помощью формального анализа строя информационной цепи. Для сравнения геномов использовалась такая числовая характеристика строя, как средняя удалённость. Формальный анализ строя позволяет непосредственно учитывать расположение нуклеотидов в каждой последовательности. Полученные результаты позволили уточнить ранее известную классификацию, выделить внутри рода *Rickettsia* группу *Rickettsia felis* располагающуюся между «предковой» группой и группой клещевой пятнистой лихорадки (КПЛ), и группу *R. akari* на границе между группой КПЛ и родом *Orientia*. Программное обеспечение для анализа нуклеотидных последовательностей с помощью формального анализа строя находится в свободном доступе по адресу: http://foarlab.org.

Ключевые слова: *Rickettsia, классификация, таксономия, геном, формальный анализ строя, средняя удалённость, межнуклеотидное расстояние.*

ВВЕДЕНИЕ

Риккетсии – это строгие внутриклеточные бактерии, передающиеся членистоногими, которые могут вызывать у человека заболевания от лёгкой степени тяжести до летального исхода [1]. Они вызывают эпидемический сыпной тиф (*Rickettsia prowazekii*), пятнистую лихорадку Скалистых гор (*R. rickettsii*), средиземноморскую пятнистую лихорадку (*R. conorii*), клещевой сыпной тиф Северной Азии (*R. sibirica*), крысиный тиф (*R. typhi*), и как минимум 10 других риккетсиозов человека, большинство из которых были описаны за последние 20 лет [1].

На основе фенотипических характеристик в роде *Rickettsia* было выделено три группы: сыпного тифа (СТ), клещевой пятнистой лихорадки (КПЛ) и группа кустарникового тифа (КТ) [2]. Проведённый анализ гена 16S рРНК риккетсий подтвердил наличие групп СТ (*R. prowazekii* и *R. typhi*), КПЛ (*R. rickettsii, R. conorii, R. sibirica* и др.) и установил существование «предковой» группы (*R. bellii* и *R. canadensis*), образовавшейся до дивергенции этих групп в роде *Rickettsia* [3]. При изучении этого гена единственный представитель группы КТ – *R. tsutsugamushi* была помещена в новый род *Orientia* как *Orientia tsutsugamushi* [4]. Предложение о создании «переходной» группы в составе *R. felis* и *R. akari*, сделанное на основании изучения геномов риккетсий и плазмиды *R. felis* [5], не получило широкой поддержки [1].

[*]stanislav63@yahoo.com
[**]nick670@yandex.ru

Создание полифазной таксономии позволило установить взаимоотношения видов внутри рода *Rickettsia*, особенно, в группе КПЛ [6]. Филогенетические отношения видов риккетсий были определены при сравнении последовательностей отдельных генов и фрагментов геномов [1]. Руководящие принципы для классификации риккетсиальных изолятов на таксономических уровнях род, группа и вид были основаны, используя различия последовательностей *rrs* (16S рРНК) и четырёх белок-кодирующих генов [7]. Филогенетические деревья, построенные на основе конкатенации последовательностей хромосомных генов, кодирующих белки 7, 8 и 11 видов риккетсий, позволили чётко определить в роде *Rickettsia* две группы: СТ и КПЛ, и подтвердить обособленное положение видов *R. bellii* и *R. canadensis* [1].

Анализ ортологичных генов в геномах пяти представителей рода *Rickettsia*, основанный на количественных мерах сходства и кладистическом анализе генных порядков позволил обосновать гипотезу о том, что вид *R. felis* обосновался ранее, чем произошла дивергенция групп СТ (*R. typhi*, *R. prowazekii*) и КПЛ (*R. conorii*), и потому не должен включаться в состав последней группы [8].

Существующие биоинформационные инструменты для поиска локальных сходств и филогенетического анализа нуклеотидных последовательностей BLAST [9], MEGA [10], а также новые подходы progressiveMauve [11], CGCPhy [12] и др. основаны на применении математических и статистических методов для текстуального сравнения последовательностей. Мы предложили новый подход, основанный на учете взаимоположения нуклеотидов, который позволяет характеризовать последовательность одним числом [16], что даёт возможность применять его для классификации групп последовательностей, обладающих значительной вариабельностью первичной структуры. Подход был успешно применен к классификации геномов риккетсий, характерной особенностью которых является значительная вариабельность генетического аппарата. Так геном *R. conorii* из 1412 генов имеет только 775 общих с геномами *R. typhi* (877) и *R. prowazekii* (872). В то же время, для риккетсий характерен высокий процент некодирующей ДНК. Так, геном *R. prowazekii* содержит 24% некодирующей ДНК [13]. При сравнении геномов *R. prowazekii* и *R. conorii* межгенные области показали большую вариабельность, чем гены [14].

Цель данного исследования заключается в расширении арсенала средств биоинформатики, применяемых для анализа биологических последовательностей. Характеристика средней удалённости, как средство формального анализа строя, была применена для оценки расположения нуклеотидов в полноразмерных геномах риккетсий и ориентий для изучения их отношений на различных таксономических уровнях.

МАТЕРИАЛЫ И МЕТОДЫ

Последовательности полноразмерных геномов *Rickettsia* и *Orientia*

Сорок один полноразмерный геном представителей родов *Rickettsia* (39) и *Orientia* (2) из семейства Rickettsiaceae был исследован с помощью числовой характеристики строя – средней удалённости нуклеотидов (табл. 1). Все последовательности геномов были загружены из NCBI GenBank (www.ncbi.nlm.nih.gov/genome). Длина геномов составила от 1111445 (*R. prowazekii* BuV67-CWPP) до 1528980 н.п. (*R. bellii* OSU 85-389) в роде *Rickettsia*, и от 2008987 (штамм Ikeda) до 2127051 н.п. (штамм Boryong) среди *O. tsutsugamushi*.

Формальный анализ строя (Formal Order Analysis – FOA)

Сущность предлагаемого подхода состоит в том, что связи между ближайшими одинаковыми нуклеотидами в геноме предлагается представлять интервалами

(межнуклеотидными расстояниями [15, 16]). При этом, в соответствии с теорией информации М. Мазура, нуклеотидные последовательности геномов рассматриваются и моделируются как информационные цепи (в математике такие объекты называются упорядоченными множествами или кортежами) [17]. В работе [18] было дано определение строя информационной цепи и алфавита строя. Там же подробно описан алгоритм построения строя нуклеотидной последовательности, в результате которого последовательность нуклеотидов перекодируется в числовую последовательность. При этом формируется алфавит данной последовательности мощностью m, при помощи которого возможна однозначная обратная перекодировка строя в нуклеотидную последовательность.

Определим длину интервала Δ_{ij} (далее интервал) между ближайшими вхождениями j-го компонента алфавита в строе как $\Delta_{ij} = k - l + 1$, где k и l – номера позиций строя, соответствующие i-ому и $(i + 1)$-ому вхождению в строй данного компонента ($\Delta_{1j} = k$, $j = 1,\ldots, m$).

Для нуклеотидной последовательности интервал определяется как расстояние между позициями, соответствующими двум последовательным вхождениям в последовательность одинаковых нуклеотидов. Поскольку геномы бактерий могут иметь как кольцевую так и линейную структуру, то за начало последовательности можно брать точки начала репликации, при этом линейные геномы искусственно «закольцовываются» и «разрезаются» в точке начала репликации. При вычислении интервалов использовалось понятие привязки, которая определяет способ обработки интервалов, лежащих на краях последовательности. В частности, привязка к началу предполагает, что интервал от начала последовательности до первого вхождения элемента будет учитываться, а интервал от последнего вхождения элемента до конца последовательности не учитывается.

Все геномы были проанализированы посредством FOA [18, 19], программное обеспечение которого доступно по адресу: http://foarlab.org. Для сравнения геномов была использована числовая характеристика строя – средняя удалённость g с привязкой к началу, подсчитываемая от точки репликации в нуклеотидной последовательности каждого генома. Показатель g был представлен с точностью до четырнадцати знаков после запятой (например, *R. prowazekii* штамм Breinl – 1.41849525064165). Для целей данного исследования оказалось достаточно применения показателя g с точностью до шестого знака после запятой. Средняя удалённость g всех нуклеотидов в последовательности определялась в виде:

$$g = \log_2 \Delta_g = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log_2 \Delta_{ij}. \quad (1)$$

Таким образом, средняя удаленность определяется посредством логарифмирования среднего геометрического интервала, вычисляемого в виде:

$$\Delta_g = \sqrt[n]{\prod_{j=1}^{m} \prod_{i=1}^{n_j} \Delta_{ij}}, \quad (2)$$

где Δ_{ij} – интервал от i-го до $(i + 1)$-го вхождения j-го символа (интервал между двумя ближайшими гомологичными нуклеотидами); n_j – число вхождений j-го символа (нуклеотида); $m = 4$ – мощность алфавита (число различных нуклеотидов в последовательности: A, T, C, G); n – длина последовательности (генома).

Для дальнейшего анализа при определении g использовалась также формула вида:

$$g = \sum_{j=1}^{m} \frac{n_j}{n} \log_2 \Delta_{g_j}, \quad (3)$$

где Δ_{gj} – средний геометрический интервал между нуклеотидами одного типа:

$$\Delta_{gj} = \sqrt[n_j]{\prod_{i=1}^{n_j} \Delta_{ij}}. \quad (4)$$

В формуле (3) в явном виде представлены частоты нуклеотидов (соотношения мощностей состава нуклеотидной последовательности) $\left\{\frac{n_j}{n}\right\}$, учитывающиеся и в общепринятых оценках (например, GC-составом).

Числовая характеристика строя g обладает высокой чувствительностью и компактно описывает полный геном, представленный одной хромосомой, его отдельные компоненты или фрагменты, имеющие разную длину. Заметим, что все характеристики строя можно вычислять, в том числе, для очень длинных последовательностей (когда $n \sim 10^8$). В данной работе не рассматриваются организмы, геномы которых состоят из более, чем одной хромосомы. Для таких геномов однозначный способ конкатенации хромосом для получения общего/единственного значения характеристики всего генома является отдельной задачей, решение которой выходит за рамки данной работы.

Таблица 1. Характеристики существующих и сгенерированных кортежей нуклеотидов

	Нуклеотидная последовательность	Длина (н.п.)	*g*	*G*	GC-состав
1	$A_{500000}C_{500000}T_{500000}G_{500000}$	2000000	0.000029690	59.38	50.00
2	$A_{250000}C_{250000}T_{250000}G_{250000}$	1000000	0.000056380	56.38	50.00
3	$A_{250}C_{250}T_{250}G_{250}$	1000	0.026492879	26.49	50.00
4	$A_{25}C_{25}T_{25}G_{25}$	100	0.166207926	16.62	50.00
5	*Вторая половина последовательности №13*	*473330*	*1.414624222*	*669584.08*	*28.67*
6	*Два нуклеотида добавлено в конец*	*1111522*	*1.418485522*	*1576677.86*	*29.00*
7	*Удалено 10 нуклеотидов подряд*	*1111510*	*1.418486545*	*1576661.98*	*29.00*
8	*Заменён последний нуклеотид*	*1111520*	*1.418486756*	*1576676.40*	*29.00*
9	*Добавлен один нуклеотид в конец*	*1111521*	*1.418486798*	*1576677.86*	*29.00*
10	*Один нуклеотид добавлен случайно*	*1111521*	*1.418487179*	*1576678.29*	*29.00*
11	*Удалён последний нуклеотид*	*1111519*	*1.418487924*	*1576676.28*	*29.00*
12	*Один нуклеотид заменён случайно*	*1111520*	*1.418488074*	*1576677.86*	*29.00*
13	*Rickettsia prowazekii* str. NMRC Madrid E | *CP004888.1*	*1111520*	*1.418488074*	*1576677.86*	*29.00*
14	*Два нуклеотида переставлены в конце*	*1111520*	*1.418488448*	*1576678.28*	*29.00*
15	*Один нуклеотид удалён случайно*	*1111519*	*1.418488587*	*1576677.02*	*29.00*
16	*Один добавлен случайно в другом месте*	*1111521*	*1.418488656*	*1576679.93*	*29.00*
17	*Один нуклеотид удалён случайно в другом месте*	*1111519*	*1.418488827*	*1576677.28*	*29.00*
18	*Два нуклеотида удалено в конце*	*1111518*	*1.418489201*	*1576676.28*	*29.00*
19	*Один нуклеотид заменён случайно в другом месте*	*1111520*	*1.418489584*	*1576679.54*	*29.00*
20	*переставлены половины последовательности №13*	*1111520*	*1.418491081*	*1576681.21*	*29.00*
21	*первая половина последовательности №13*	*638190*	*1.421346081*	*907088.86*	*29.24*
22	*Rickettsia japonica* YH | NC_016050.1	1283087	1.431179153	1836327.37	32.35
23	*Rickettsia japonica* YH 16s rRNA | NR_074459.1	1508	1.517556143	2288.47	50.86
24	$(ACTG)_{25}$	100	1.965849625	196.58	50.00
25	$(ACTG)_{250}$	1000	1.996584963	1996.58	50.00
26	$(ACTG)_{250000}$	1000000	1.999996585	1999996.58	50.00
27	$(ACTG)_{500000}$	2000000	1.999998292	3999996.58	50.00

В таблице 1 приведена также другая характеристика строя G, называемая глубиной [18], представляющая собой суммарное количество информации в последовательности. Глубина может быть вычислена как произведение средней удалённости g на длину последовательности n: $G = gn$.

Для изучения зависимости числовых характеристик G и g от длины, состава и распределения нуклеотидов в последовательности наряду с геномами двух риккетсий были проанализированы искусственно сгенерированные последовательности (1-4 и 24-27 в табл. 1), а также последовательности, полученные путем модификации генома *Rickettsia prowazekii* str. NMRC Madrid E (5-12 в табл. 1, выделены курсивом).

Все последовательности в таблице упорядочены по характеристике g (обозначена серым фоном). Малые изменения фиксируются отличиями характеристики g в шестом разряде после запятой. Большие изменения (строки 5 и 21) – фиксируются во втором-третьем знаке после запятой. GC-состав при модификациях последовательностей в этом эксперименте практически не меняется. В искусственных последовательностях увеличение длин серий приводит к уменьшению величины g ($g \to 0$). Увеличение числа повторов олигонуклеотида ACTG приводит к увеличению g вплоть до $g \to 2$ ($g \to \log_2 m$).

Кластерный анализ

Для верификации полученной схемы классификации представителей семейства *Rickettsiaceae* и критериев формирования таксонов внутри рода *Rickettsia* осуществлялся кластерный анализ с помощью программы Past, доступной по адресу: http://folk.uio.no/ohammer/past/. Анализ показателей средней удалённости осуществлялся с применением алгоритма UPGMA (Unweighted pair-group average). Кластеры были сформированы на основании величины среднего расстояния (Distance) между членами всех групп.

Таблица 2. Характеристики полных геномов представителей семейства Rickettsiaceae

№	Группа / кол-во изолятов	Виды и штаммы *Rickettsia* и *Orientia*	№ референтной последовательности в NCBI / № доступа в GenBank*	Размер генома (н.п.)	G + C (%)	g
1	СТ / 11	*R. prowazekii* str. Katsinyian	NC_017050/CP003392	1111454	29	1.41823242
2		*R. prowazekii* str. BuV67-CWPP	NC_017056/CP003393	1111445	29	1.41824129
3		*R. prowazekii* str. Madrid E	NC_000963/AJ235269	1111523	29	1.41824702
4		*R. prowazekii* Rp22	NC_017560/CP001584	1111612	29	1.41826432
5		*R. prowazekii* str. GvV257	NC_017048/CP003395	1111969	29	1.41831401
6		*R. prowazekii* str. RpGvF24	NC_017057/CP003396	1112101	29	1.41832503
7		*R. prowazekii* str. Chernikova	NC_017049/CP003391	1109804	29	1.41838369
8		*R. prowazekii* str. Breinl	NC_020993/CP004889	1109301	29	1.41849489
9		*R. typhi* str. B9991CWPP	NC_017062/CP003398	1112957	28.9	1.41989726
10		*R. typhi* str. TH1527	NC_017066/CP003397	1112372	28.9	1.41989962
11		*R. typhi* str. Wilmington	NC_006142/AE017197	1111496	28.9	1.41990899
12	ПГ / 4	*R. bellii* OSU 85-389	NC_009883/CP000849	1528980	31.6	1.42461059
13		*R. bellii* RML369-C	NC_007940/CP000087	1522076	31.6	1.42478461
14		*R. canadensis* str. CA410	NC_016929/CP003304	1150228	31	1.42505826
15		*R. canadensis* str. McKiel	NC_009879/CP000409	1159772	31	1.42576173
16	-	*R. monacensis*. str. IrR/Munich	NZ_LN794217/LN794217	1353450	32.39	1.42639158
17	-	*R. felis* URRWXCal2	NC_007109/CP000053	1485148	32.6	1.42911848

18	КПЛ / 21	*R. rhipicephali* str. 3-7-female6-CWPP	NC_017042/CP003342	1290368	32.4	1.43088037
19		*R. japonica* YH	NC_016050/CP003342	1283087	32.4	1.43117915
20		*R. australis* str. Cutlack	NC_017058/CP003338	1296670	32.3	1.43123685
21		*R. montanensis* str. OSU 85-930	NC_017043/CP003340	1279798	32.6	1.43172146
22		*R. slovaca* str. D-CWPP	NC_017065/CP003375	1275720	32.5	1.43199033
23		*R. slovaca* 13-B	NC_016639/CP002428	1275089	32.5	1.43200176
24		*R. parkeri* str. Portsmouth	NC_017044/CP003341	1300386	32.4	1.43209422
25		*R. conorii* str. Malish 7	NC_003103/AE006914	1268755	32.4	1.43213976
26		*R. rickettsii* str. Arizona	NC_016909/CP003307	1267197	32.4	1.43264659
27		*R. rickettsii* str. Iowa	NC_010263/CP000766	1268188	32.4	1.43268083
28		*R. rickettsii* str. Brazil	NC_016913/CP003305	1255681	32.5	1.43272749
29		*R. rickettsii* str. Morgan	NZ_CP006010/CP006010	1269809	32.5	1.43278832
30		*R. rickettsii* str. Hino	NC_016914/CP003309	1269837	32.5	1.43279406
31		*R. rickettsii* str. Colombia	NC_016908/CP003306	1270083	32.5	1.43286685
32		*R. rickettsii* str. Hlp#2	NC_016915/CP003311	1270751	32.5	1.43286997
33		*R. rickettsii* str. R	NZ_CP006009/CP006009	1257005	32.5	1.43288002
34		*R. rickettsii* str. "Sheila Smith"	NC_009882/CP000848	1257710	32.5	1.43291631
35		*R. massiliae* MTU5	NC_009900/CP000683	1360898	32.5	1.43314584
36		*R. philipii* str. 364D	NC_016930/CP003308	1287740	32.5	1.43325070
37		*Candidatus* R. amblyommii GAT-30V	NC_017028/CP003334	1407796	32.5	1.43334598
38		*R. peacockii* str. Rustic	-/CP001227	1288492	32.6	1.43514666
39	-	*R. akari* str. Hartford	NC_009881/CP000847	1231060	32.3	1.43747340
40	*Orientia*/2	*O. tsutsugamushi* str. Boryong	NC_009488/AM494475	2127051	30.5	1.44599461
41		*O. tsutsugamushi* str. Ikeda	NC_010793/AP008981	2008987	30.5	1.44642319

РЕЗУЛЬТАТЫ

Показатель g рассчитывался для каждого из 41 геномов *Rickettsia* и *Orientia* (табл. 2). Как видно из формулы (3) показатель g определяется не только расположением компонентов, но и распределением частот нуклеотидов в последовательности. Поэтому, показатель g коррелирует с общепринятым показателем GC-состава. Применение статистического подхода (при анализе объектов не статистических по природе) позволяет сравнивать геномы, имеющие, в том числе, разную длину. На это указывает формула (1). Однако геном конкретного организма, как целостная сущность, имеет определённую длину, состав и расположение компонентов, и показатель g будет адекватно отображать структуру генома данного организма только при учёте всей последовательности и таким образом – длины данного генома (рис. 1,A и 1,B).

Характеристика g с привязкой к началу и циклической привязкой была вычислена для всех геномов в виде, представленном в GenBank, а также в виде, выровненном по точке инициации репликации. Характеристика средней удалённости g с привязкой к началу продемонстрировала стабильность не зависимо от варианта начала обсчёта для всех последовательностей, так как при смене позиции начала обсчёта она меняется только в пятом-шестом знаке после запятой. Характеристика средней удалённости с циклической привязкой была неизменной при любой позиции начала проведения вычислений. Независимо от вида привязки при проведении вычислений изменение величины характеристики средней удалённости не влияло на результат группирования внутри изучаемой группы микроорганизмов.

Геномы гомологичных изолятов имели очень близкие значения показателя *g* включая *R. prowazekii* (8), *R. rickettsii* (9), *R. typhi* (3), *R. bellii* (2), *R. canadensis* (2), *R. slovaca* (2) и *O. tsutsugamushi* (2) (табл. 1). Этот показатель имел наиболее низкие значения от 1.418232 у *R. prowazekii* штамм Katsinyian до 1.419908 у *R. typhi* шт. Wilmington среди риккетсий группы СТ, при этом наблюдалось минимальное различие между размерами геномов, 1111454 и 1111496 н.п., соответственно. Все штаммы *R. prowazekii* (8) образовали целостную группу (рис. 1,A и 1,B; рис. 2) вместе с примыкающими штаммами *R. typhi* (3). Подобные данные были получены в группе КПЛ где показатель *g* ранжировался от 1.430880 у *R. rhipicephali* шт. 3-7-female6-CWPP до 1.433345 у *Candidatus R. amblyommii* шт. GAT-30V. При этом *R. peacockii* шт. Rustic со значением показателя *g* – 1.435146 была значительно отдалена от ядра «классической» группы КПЛ (рис. 1,A и 1,B; рис. 2). На удивление, низкие различия показателя *g* были определены для *R. bellii* (2) и *R. canadensis* (2) в «предковой» группе, несмотря на максимальные отличия в размерах геномов от 1528980 до 1150228 н.п., соответственно (табл. 1, рис. 1A), что поддерживает существование этой группы.

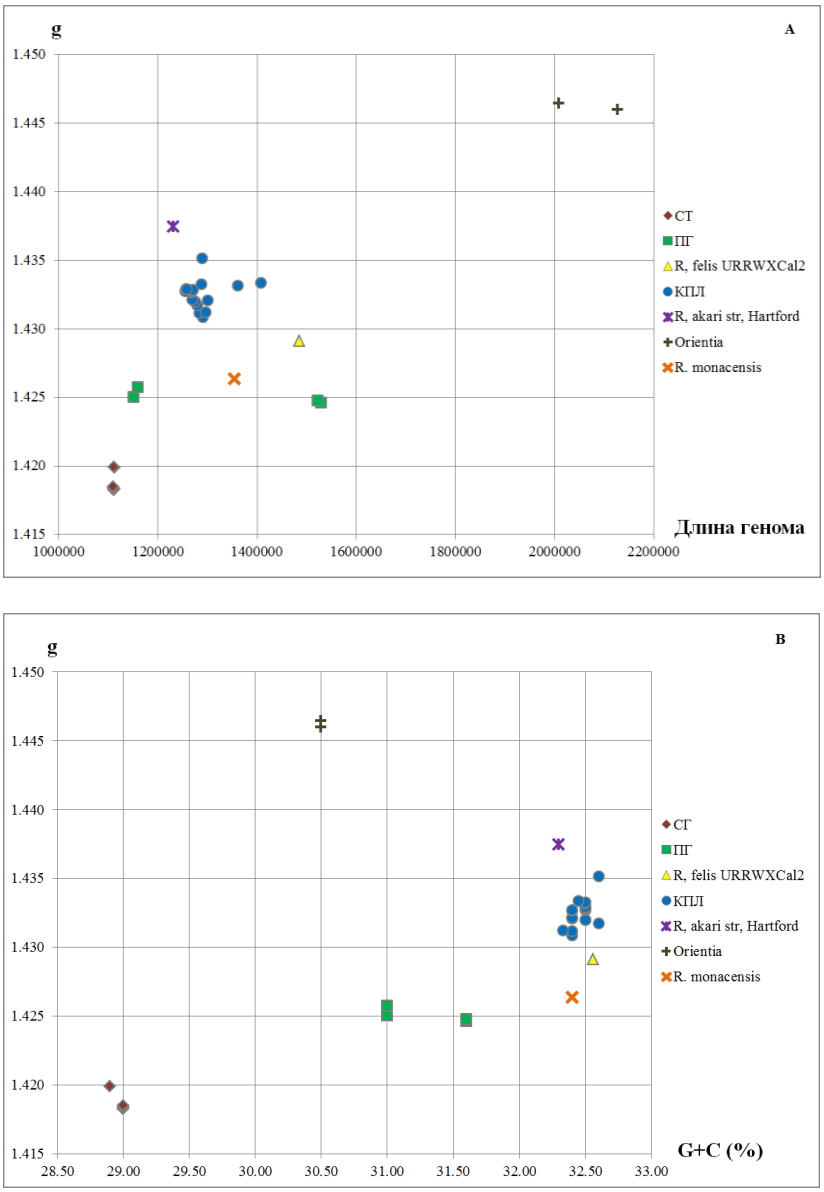

Рис. 1. Распределение исследованных геномов по значениям средней удалённости: **A** – по длине, **B** – по GC-составу.

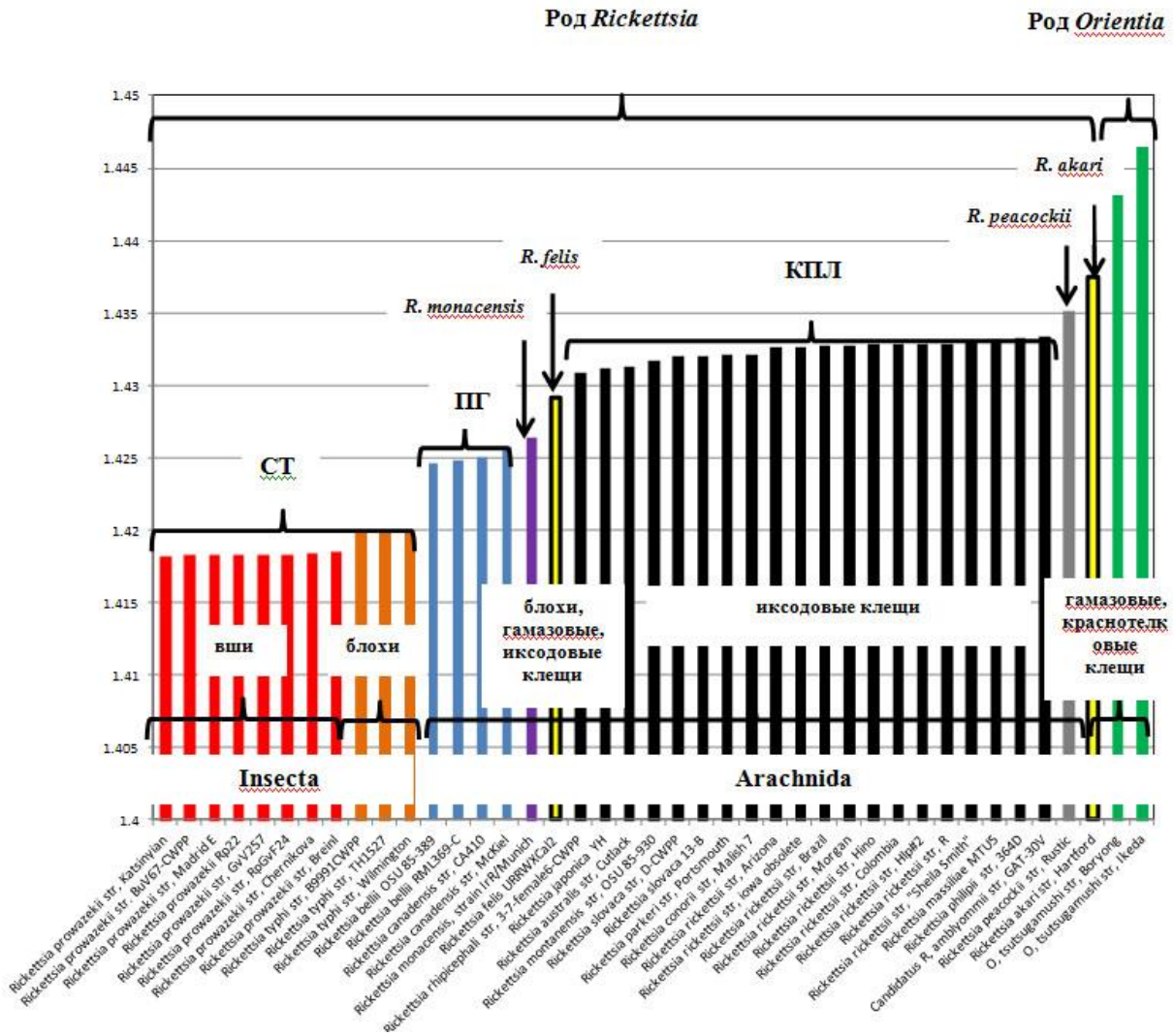

Рис. 2. Классификация, построенная с помощью числовой характеристики средней удалённости нуклеотидов в полноразмерных геномах прокариот из семейства Rickettsiaceae. Обозначения – в тексте.

Полученные в результате исследований данные позволяют рассматривать следующее группирование представителей рода *Rickettsia*. Наш анализ подтверждает существование групп сыпного тифа (СТ), «предковой» группы (ПГ) и группы клещевой пятнистой лихорадки (КПЛ) внутри рода *Rickettsia*. Соответственно, группа СТ включает изоляты *R. prowazekii* и *R. typhi* со средним показателем *g* (1.418746 мат. ожидание – МО + 0.000746 среднеквадратическое отклонение – СКО), «предковая» группа включает *R. bellii* и *R. canadensis* со средним показателем *g* (1.425054 МО + 0.000507 СКО), и группа КПЛ включает все изоляты, связанные с иксодовыми клещами со средним показателем *g* (1.432492 МО + 0.000657 СКО), за исключением *R. peacockii* шт. Rustic (1.435146) (табл. 1, рис. 2), которая вместе с *R. akari* (*g* = 1.437473) расположилась на расстоянии от ядра «классической» группы КПЛ. В соответствии с полученными данными, показатель *g* подсчитанный для *R. felis* шт. URRWXCal2 (*g* = 1.429118) значительно отличался от такового для *R. akari* шт. Hartford. Таким образом, применение показателя *g* как ранжирующего (классифицирующего) параметра позволило поместить *R. felis* между «предковой» группой и КПЛ, в то время как *R. akari* была помещена между группой КПЛ и родом *Orientia*. Таким образом, наш анализ не поддерживает включение *R. felis* и *R. akari* в «переходную» группу [5]. Размер генома *R. felis* является большим, чем у всех других *Rickettsia* spp., включая *R. akari*, за исключением *R. bellii* (табл. 1, рис. 1,А), и видимо *R. felis* ближе к «предковой» группе, а не является эволюционным шагом между группой КПЛ и СТ [1]. *R. monacensis* шт.

IrR/Munich (g = 1.426391) также заняла позицию обособленную от группы КПЛ в пределах «предковой» группы.

Мы сравнили полученную схему с филогенетическими деревьями, построенными при изучении последовательностей гена *rrs* и четырёх белок кодирующих генов (*glt*A, *omp*A, *omp*B и *sca*4) риккетсий [7]. Наш подход, основанный на анализе полноразмерных геномов сравним с 16S рРНК древом Rickettsiaceae за исключением размещения *R. akari*. Этот вид является единственным среди представителей рода *Rickettsia*, который экологически связан с гамазовыми клещами, и в соответствии с нашими данными *R. akari* была помещена между группой КПЛ и видом *O. tsutsugamushi* из рода *Orientia*, который экологически связан с краснотелковыми клещами, которые являются филогенетически близкими гамазовым клещам (рис. 2). Таким образом, полученные нами данные согласуется с таксономией членистоногих (Arthropoda), которые являются хозяевами риккетсий (рис. 2).

При проведении кластерного анализа с помощью показателя средней удалённости множество, состоящее из геномов представителей семейства Rickettsiaceae, было разбито (сгруппировано) на непересекающиеся подмножества (кластеры), состоящие из близких по значению этого показателя геномов и представлено в виде дендрограммы (рис. 3). Номера соответствуют порядковым номерам нуклеотидных последовательностей геномов, указанных в таблице 2, в соответствии со значениями показателя средней удалённости.

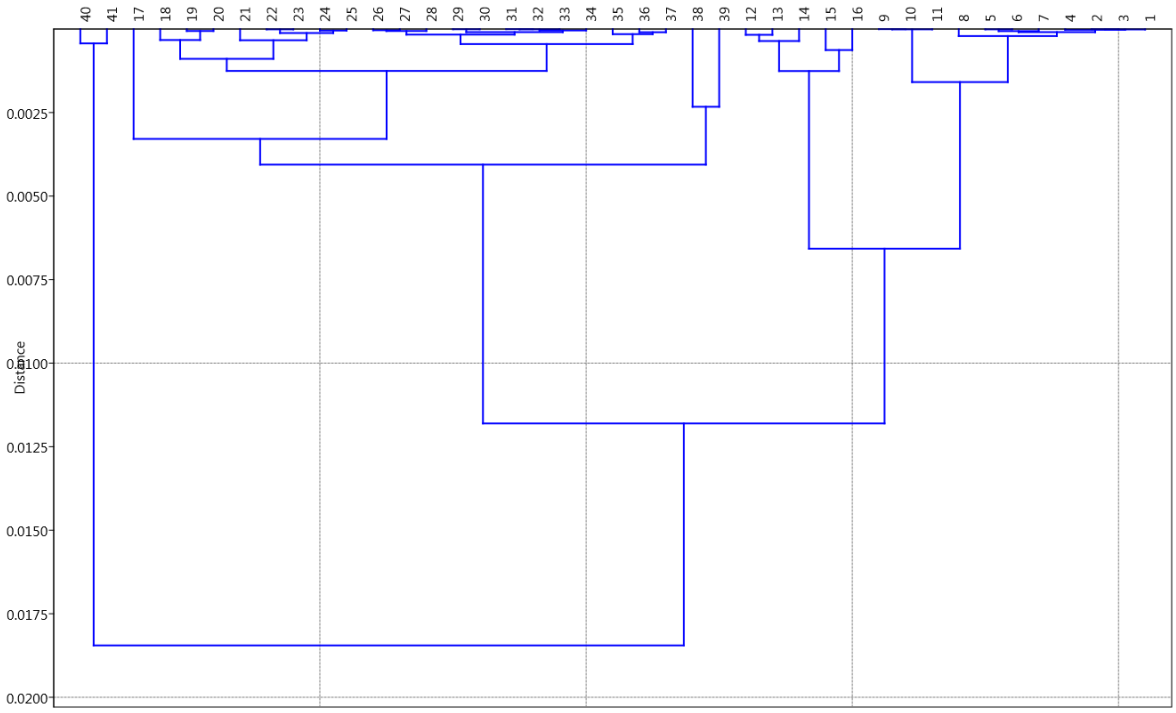

Рис. 3. Кладограмма, построенная при анализе показателя средней удалённости нуклеотидов в полноразмерных геномах представителей семейства Rickettsiaceae.

В соответствии с полученными результатами, при расстоянии по средней удалённости (Distance), в диапазоне от 0.02 до 0.0175 в семействе Rickettsiaceae происходит образование родов *Rickettsia* и *Orientia* (рис. 3). При расстоянии от 0.0125 до 0.01 в роде *Rickettsia* происходит разделение на две мажорные группы. Одна включает в себя группу CT и «предковую», что может быть обосновано с позиции выявления перекрестных реакций с их антигенами [20]. При расстоянии от 0.0075 до 0.0050 происходит разделение на группу CT и «предковою». При расстоянии от 0.005 до 0.0025 происходит разделение в классической группе КПЛ, с выделением трёх

групп: КПЛ, *R. felis*, и *R. akari* с *R. peacockii*. И только в диапазоне ниже 0.0025 происходит формирование видов риккетсий в группе КПЛ.

ОБСУЖДЕНИЕ

Медицинская таксономия имеет решающее значение при диагностике инфекционных заболеваний, так как учитывает эпидемиологические характеристики, клинические проявления и переносчиков, участвующих в передаче этиологических агентов [1].

Применение молекулярно-биологических и филогенетических методов позволило усовершенствовать классификацию и таксономию риккетсий, однако положение некоторых видов остаётся дискуссионным.

Изначально на основании изучения антигенных характеристик *R. canadensis* рассматривалась как член группы СТ [20]. Данные, полученные при изучении генов 16S рРНК и цитратсинтазы риккетсий обозначили позицию *R. canadensis* за пределами группы СТ [21]. Несмотря на создание «предковой» группы [3] таксономическое положение *R. canadensis* и *R. bellii* считается дискуссионным и оба вида можно отнести в зависимости от применяемых критериев (генов) как к группе КПЛ, так и СТ [21]. Положение *R. felis* также является спорным. Первоначально при изучении гена цитратсинтазы этот вид был отнесён к группе СТ, затем по данным филогенетического анализа последовательностей гена 16S рРНК перемещен ближе к группе КПЛ [21]. Это характеризует сложность филогенетической позиции некоторых видов риккетсий и субъективность применения анализа отдельных генов для их классификации и таксономии.

Только комплексный подход, основанный на применении традиционных «классических» методов риккетсиологии и прогрессивных молекулярно-биологических технологий, дополненных подходами, позволяющими учитывать информацию, содержащуюся как в кодирующей, так и некодирующей частях генома позволяет получить объективное представление о взаимоотношениях видов риккетсий.

В данной работе верифицирована достоверность классификации прокариот из семейства Rickettsiaceae, построенной на основании сравнения средней удалённости нуклеотидов в полноразмерных геномах, с учётом таксономии их хозяев – членистоногих (Arthropoda).

Членистоногие из класса насекомых (Insecta) вши и блохи являются хозяевами видов риккетсий из группы сыпного тифа. Членистоногие класса паукообразных (Arachnida) являются хозяевами видов риккетсий из «предковой» группы (иксодовые клещи), группы клещевой пятнистой лихорадки (иксодовые клещи), *R. akari* (гамазовые клещи) и *O. tsutsugamushi* (краснотелковые клещи). Данные, полученные при изучении характеристики средней удалённости ранжируют виды риккетсий группы сыпного тифа (g = 1.418232÷1.419908) (рис. 2), экологически связанные с насекомыми (вши, блохи), отдельно от видов риккетсий «предковой» группы (g = 1.424610÷1.425761) и группы клещевой пятнистой лихорадки (g = 1.431179÷1.435146), связанных с паукообразными (иксодовые клещи). *R. akari* (g = 1.437473), которая по данным других классификаций расположена в группе клещевой пятнистой лихорадки, в полученной классификации разместилась на границе между этой группой и *O. tsutsugamushi* (g = 1.445994÷1.446423) из рода *Orientia*, хозяином которой также являются представители паукообразных (краснотелковые клещи). Учитывая это, можно рекомендовать выделить *R. akari* в отдельную группу внутри рода *Rickettsia* на основании значения показателя средней удалённости и таксономического положения хозяев – гамазовых клещей.

Хозяином *R. felis* (g = 1.429118) считался только представитель класса насекомых – кошачья блоха (*Ctenocephalides felis*). Недавно этот вид риккетсий был генотипирован в различных членистоногих из класса Insecta [1, 22] и Arachnida (иксодовые и аргасовые

клещи) [23–27]. *Candidatus R. senegalensis* была недавно выявлена в *C. felis* [28]; этот вид образовал компактный изолированный клад с видами риккетсий, включающих два официально описанных вида: *R. felis* и *R. hoogstraalii*, что является сестринской группой, соседствующей с *R. australis*/*R. akari* [28]. Candidatus *R. asemboensis* еще один вид, который был недавно определён, как близкий к *R. felis* [29]. Таким образом, можно предположить, что *R. felis* вместе с *Candidatus R. senegalensis* и *Candidatus R. asemboensis* представляют собой отдельную группу и филогенетическую линию внутри рода *Rickettsia,* экологически связанную с кошачьей блохой. Уникальность *R. felis* обусловленная экологической связью с представителями обоих классов членистоногих подтверждает её обособленное положение по отношению к другим видам риккетсий и выделение в классификации в самостоятельную группу.

Таким образом, классификация, полученная на основании анализа числовой характеристики строя – средней удалённости нуклеотидов является обоснованной с позиции таксономического положения хозяев – членистоногих (Arthropoda).

В результате исследования *R. peacockii* заняла позицию рядом с *R. akari* на границе с группой КПЛ и родом *Orientia*. Это может быть обусловлено наличием значительных геномных перестроек, связанных с присутствием недавно описанного семейства транспозонов ISRpe1 близкородственных представителям рода *Wolbachia* [30]. Отсутствие синтении генома *R. peacockii* с геномом близкородственного патогенного вида *R. rickettsii*, объясняется наличием 42 копий транспозонов ISRpe1 в хромосоме, что связано с многочисленными делециями произошедшими в результате рекомбинаций между копиями транспозонов [31]. Наличие многочисленных копий транспозонов ISRpe1 связывают с приобретением в результате горизонтального переноса генов от видов *Cardinium*, это вызвало большие геномные реорганизации и делеции, что в результате привело к потере вирулентности.

Расположение *R. monacensis* рядом с *R. canadensis* str. McKiel из «предковой» группы за пределами группы КПЛ может быть обосновано с позиции изучения антигенных характеристик и данных филогенетического анализа двух генов (*ompA* и *gltA*), которые показали что она и другие близкородственные виды (штаммы) риккетсий, выделенные из клещей *I. ricinus* занимают обособленную позицию по отношению к группе КПЛ [32].

В проведённом исследовании, характеристика (g) показала высокую чувствительность к расположению нуклеотидов в геномах риккетсий. Мы проанализировали ценность применения этого параметра при определении таксонов: вид, группа, род и семейство для классификации бактерий из семейства Rickettsiaceae.

Текущая версия FOA не включает в себя анализ нуклеотидных последовательностей плазмид являющихся внехромосомными генетическими элементами, которые могут дополнительно способствовать нашему пониманию эволюционных связей различных групп риккетсий.

Таким образом, показано, что этот метод может быть применён в таксогеномике риккетсий. Показатель средней удалённости с привязкой к началу (g) является чувствительным и характеризует геном как информационную цепь.

Построение классификации при этом подходе информативнее и доступнее существующих филогенетических методов, основанных на применении специализированных программных продуктов (MEGA, BLAST), так как позволяет непосредственно учитывать расположение нуклеотидов в полноразмерных геномах, а не сравнивать только гомологичные фрагменты геномов.

Определение термина геном было дано немецким учёным-ботаником Г. Винклером почти сто лет назад в 1920 году [33] и было переведено на английский язык: «Геном это гаплоидный набор хромосом, который, вместе с протоплазмой, определяет материальную основу видов» [34]. «Геномы развиваются через многие события, происходящие с генами: делеции, дупликации, инсерции, перестановки в геномах, а не

через последовательные адаптационные процессы. Геномы – это динамические и химерные сущности с генными репертуарами, которые происходят из вертикальных и горизонтальных приобретений, а также создания новых генов (*de novo* gene)» [35]. Различная организация геномов близкородственных видов риккетсий, связанная с потерей генов в результате редуктивной эволюцией, горизонтальными приобретениями генов, связанные с плазмидами могут повлиять на положение и распределение отдельных нуклеотидов и/или их групп расположенных вдоль хромосомы. Эти накопленные изменения могут быть обнаружены с помощью описываемой здесь числовой характеристики строя. Предлагаемый метод дополняет существующие биоинформационные подходы, используемые для классификации прокариот, и учитывает расположение нуклеотидов в полноразмерных геномах. Применение FOA подтверждает наличие трёх групп в роде *Rickettsia*, сыпного тифа, «предковой» и клещевой пятнистой лихорадки. На основании анализа показателя средней удалённости нуклеотидов в полноразмерных геномах и экологических связей риккетсий с членистоногими можно выделить группу *R. felis,* вместе с *Ca. Rickettsia senegalensis* и *Ca. R. asemboensis*, располагающуюся между «предковой» группой и группой КПЛ, и группу *R. akari* на границе между группой КПЛ и родом *Orientia*. Следует отметить обособленную позицию *R. peacockii* на периферии группы КПЛ и позицию *R. monacensis* в пределах «предковой» группы.

Результаты исследований нуклеотидных последовательностей с использованием аппарата формального анализа строя позволяют сформулировать определение генома организмов с одной хромосомой с позиции теории информации. «Геном – это информационная цепь (упорядоченное множество, кортеж), сформированная четырьмя разными компонентами (мощность алфавита $m = 4$) – азотистыми основаниями (A – аденин, G – гуанин, C – цитозин и T – тимин), содержащая информацию, в том числе, необходимую для создания, развития и функционирования организма, а также для передачи и копирования генетической информации».

ЗАКЛЮЧЕНИЕ

В работе определён формализм строя нуклеотидной последовательности. Сформулированы ограничения, накладываемые на вектор строя. Описана процедура декомпозиции символьной последовательности (кортежа) на строй и алфавит.

Проведено исследование и анализ геномов риккетсий с помощью средней удалённости и установлено, что g является характеристикой их генома и более информативна, чем общепринятые характеристики, не учитывающие расположение компонентов, в том числе – длина генома и GC-состав.

Показано что средняя удалённость и другие характеристики строя чувствительны даже к малым изменениям расположения нуклеотидов в последовательности.

На основании средней удалённости проведена и обоснована реклассификация внутри семейства Rickettsiaceae.

Дано определение генома с позиции теории информации, биологии и биоинформатики как информационной цепи нуклеотидов.

Авторы благодарят сотрудников лаборатории анализа геномов ФГБУ «ФНИЦЭМ им. Н.Ф. Гамалеи» МЗ РФ Андрея Семенова и Марину Кунда за идентификацию точек начала репликации во всех последовательностях, включённых в это исследование.

СПИСОК ЛИТЕРАТУРЫ

1. Merhej V., Raoult D. Rickettsial evolution in the light of comparative genomics. *Biol. Rev. Camb. Philos. Soc.* 2011. V. 86. № 2. P. 379–405.

2. Weiss E., Moulder J.W. Order I. *Rickettsiales* Gieszczkiewicz 1939. In: *Bergey's manual of systematic bacteriology*. V. 1. Eds N.R. Krieg and J.G. Holt. Baltimore, Md.: The Williams & Wilkins Co., 1984. P. 687–703.
3. Stothard D.R., Clark J.B., Fuerst P.A. Ancestral divergence of *Rickettsia bellii* from the spotted fever and typhus groups of Rickettsia and antiquity of the genus *Rickettsia*. *Int. J. Syst. Bacteriol*. 1994. V. 44. P. 798–804.
4. Tamura A., Ohashi N., Urakami H., Miyamura S. Classification of *Rickettsia tsutsugamushi* in a new genus, *Orientia* gen. nov., as *Orientia tsutsugamushi* comb. nov. *Int. J. Syst. Bacteriol*. 1995. V. 45. P. 589–591.
5. Gillespie J.J., Beier M.S., Rahman M.S., Ammerman N.C., Shallom J.M., Purkayastha A., Sobral B.S., Azad A.F. Plasmids and rickettsial evolution: insight from *Rickettsia felis*. *PLoS One*. 2007. V. 2. Article No e266.
6. Roux V., Raoult D. Phylogenetic analysis and taxonomic relationships among the genus Rickettsia. In: *Rickettsiae and Rickettsial diseases at the turn of the third millennium*. Marseille: Elsevier production, 1999. P. 52–66.
7. Fournier P.E., Dumler J.S., Greub G., Zhang J., Wu Y., Raoult D. Gene sequence-based criteria for identification of new rickettsia isolates and description of Rickettsia heilongjiangensis sp. nov. *J. Clin. Microbiol*. 2003. V. 41. P. 5456–5465.
8. Марков А.В., Захаров И.А. Использование количественных мер сходства генных порядков для построения филогенетических реконструкций на примере бактерий рода *Rickettsia*. *Генетика*. 2008. Т. 44. № 4. С. 456–466.
9. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J. Mol. Biol*. 1990. V. 215. № 3. P. 403–410.
10. Kumar S., Stecher G., Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*. 2016. V. 33. P. 1870–1874.
11. Darling A.E., Mau B., Perna N.T. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*. 2010. V. 25. № 5(6). Article No e11147.
12. Du W., Cao Z., Wang Y., Sun Y., Blanzieri E., Liang Y. Prokaryotic phylogenies inferred from whole-genome sequence and annotation data. *Biomed. Res. Int*. 2013. Article No. 409062.
13. McLeod M.P., Qin X., Karpathy S.E., Gioia J., Highlander S.K., Fox G.E., McNeill T.Z., Jiang H., Muzny D., Jacob L.S., Hawes A.C., Sodergren E., Gill R., Hume J., Morgan M., Fan G., Amin A.G., Gibbs R.A., Hong C., Yu X.J., Walker D.H., Weinstock G.M. Complete genome sequence of *Rickettsia typhi* and comparison with sequences of other rickettsiae. *J. Bacteriol*. 2004. V. 186. P. 5842–5855.
14. Ogata H., Audic S., Renesto-Audiffren P., Fournier P. E., Barbe V., Samson D., Roux V., Cossart P., Weissenbach J., Claverie J.M., Raoult D. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science*. 2001. V. 293. P. 2093–2098.
15. Nair A.S.S., Mahalakshmi T. Visualization of genomic data using inter-nucleotide distance signals. B: *Proceedings of IEEE Genomic Signal Processing*. Bucharest, 2005.
16. Afreixo V., Bastos C.A.C., Pinho A.J., Garcia S.P., Ferreira P.J.S.G. Genome analysis with inter-nucleotide distances. *Bioinformatics*. 2009. V. 25. № 23. P. 3064–3070.
17. Мазур М. *Качественная теория информации.* Москва: Мир, 1974. 240 с.
18. Гуменюк А. С., Поздниченко Н. Н., Родионов И. Н., Шпынов С.Н. О средствах формального анализа строя нуклеотидных цепей. *Математическая биология и биоинформатика*. 2013. Т. 8. № 1. С. 373–397. doi: 10.17537/2013.8.373
19. Shpynov S., Pozdnichenko N., Gumenuk A. Approach for classification and taxonomy within family Rickettsiaceae based on the Formal Order Analysis. *Microbes and Infection.* V. 17. № 11. P. 839–844.
20. Игнатович В.Ф. Антигенные связи риккетсий Провачека и риккетсий Канада, установленные при изучении сывороток больных болезнью Брилля. *Журнал*

гигиены, эпидемиологии, микробиологии и иммунологии. 1977. Т. 21. № 1. С. 48–52.

21. Raoult D., Roux V. Rickettsioses as paradigms of new or emerging infectious diseases. *Clin. Microbiol. Rev*. 1997. V. 10. № 4. P. 694–719.
22. Socolovschi C., Pages F., Ndiath M.O., Ratmanov P., Raoult D. Rickettsia species in African Anopheles mosquitoes. *PLoS One*. 2012. V. 7. № 10. Article No. e48254.
23. Ishikura M., Ando S., Shinagawa Y., Matsuura K., Hasegawa S., Nakayama T., Fujita H., Watanabe M. Phylogenetic analysis of spotted fever group rickettsiae based on gltA, 17-kDa, and rOmpA genes amplified by nested PCR from ticks in Japan. *Microbiol. Immunol*. 2003. V. 47. P. 823–832.
24. Choi Y.J., Lee E.M., Park J.M., Lee K.M., Han S.H., Kim J.K., Lee S.H., Song H.J., Choi M.S., Kim I.S., Park K.H., Jang W.J. Molecular detection of various rickettsiae in mites (acari: trombiculidae) in southern Jeolla Province, Korea. *Microbiol. Immunol*. 2007. V. 51. P. 307–312.
25. Oliveira K.A., Oliveira L.S., Dias C.C., Silva A.Jr., Almeida M.R., Almada G., Bouyer D.H., Galvao M.A., Mafra C. Molecular identification of *Rickettsia felis* in ticks and fleas from an endemic area for Brazilian Spotted Fever. *Mem. Inst. Oswaldo. Cruz*. 2008. V. 103. № 2. P. 191–194.
26. Abarca K., López J., Acosta-Jamett G., Martínez-Valdebenito C. *Rickettsia felis* in *Rhipicephalus sanguineus* from two distant Chilean cities. *Vector Borne Zoonotic Dis*. 2013. V. 13. № 8. P. 607–609.
27. Soares H.S., Barbieri A.R., Martins T.F., Minervino A.H., de Lima J.T., Marcili A., Gennari S.M., Labruna M.B. Ticks and rickettsial infection in the wildlife of two regions of the Brazilian Amazon. *Exp. Appl. Acarol*. 2015. V. 65. № 1. P. 125–140.
28. Mediannikov O., Aubadie-Ladrix M., Raoult D. *Candidatus* ‘Rickettsia senegalensis’ in cat fleas in Senegal. *New Microbe and New Infect*. 2015. V. 3. P. 24–28.
29. Jiang J., Maina A.N., Knobel D.L., Cleaveland S., Laudisoit A., Wamburu K. Molecular detection of *Rickettsia felis* and *Candidatus* Rickettsia asemboensis in fleas from human habitats, Asembo, Kenya. *Vector Borne Zoonotic Dis*. 2013. V. 13. № 8. Article No. 550e8.
30. Simser J.A., Rahman M.S., Dreher-Lesnick S.M., Azad A.F. A novel and naturally occurring transposon, ISRpe1 in the *Rickettsia peacockii* genome disrupting the rickA gene involved in actin-based motility. *Mol. Microbiol*. 2005. V. 58. № 1. P. 71–79.
31. Felsheim R.F., Kurtti T.J., Munderloh U.G. Genome sequence of the endosymbiont *Rickettsia peacockii* and comparison with virulent *Rickettsia rickettsii*: identification of virulence factors. *PLoS One*. 2009. V. 21. № 4(12). Article No. e8361.
32. Jado I., Oteo J.A., Aldámiz M., Gil H., Escudero R., Ibarra V., Portu J., Portillo A., Lezaun M.J., García-Amil C., Rodríguez-Moreno I., Anda P. *Rickettsia monacensis* and human disease, Spain. *Emerg. Infect. Dis*. 2007. V. 13. Article No. 9. P. 1405–1407.
33. Winkler H.L: Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche. Jena. Verlag Fischer. 1920.
34. Lederberg J., McCraw A.T: 'Ome sweet 'omics– a genealogical treasury of words. *Scientist*. 2001. V. 15. № 7. P. 8.
35. Merhej V., Raoult D. Rhizome of life, catastrophes, sequence exchanges, gene creations, and giant viruses: how microbial genomics challenges Darwin. *Front. Cell. Infect. Microbiol*. 2012. V. 28. № 2. P. 113.

Рукопись поступила в редакцию 03.08.2016.
Дата опубликования 05.12.2016.