***Доклады Международной конференции "Математическая биология и биоинформатика"**. Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e59. doi: 10.17537/icmbb18.101*

Использование *k*-меров для внутривидового типирования бактерий

Киселев С.С.[1], Озолинь О.Н.[1], Панюков В.В.[2]

[1]*Институт биофизики клетки Российской академии наук*
[2]*ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН*

panyukov@itaec.ru

С середины 70-х годов XX века в качестве стандартного филогенетического маркера в бактериальной таксономии применяют ген 16S рРНК, однако в ряде случаев при его использовании не удаётся отличить различные виды микроорганизмов друг от друга. В связи с этим обстоятельством в последнее время получили широкое распространение методы анализа полных геномов *in silico* – средняя нуклеотидная идентичность, индекс корреляции тетрануклеотидов и другие. Их результаты хорошо соотносятся с данными ДНК-ДНК-гибридизации и мультилокусного типирования, но отдельные подвиды или группы штаммов не всегда удаётся идентифицировать этими методами. В данной работе мы предлагаем использовать для типирования бактериальных штаммов, относящихся к одному виду, сходство между геномами, которое определённым образом выражается через количество *k*-меров длиной 12 нуклеотидных пар, одновременно входящих в состав сравниваемых геномов. В качестве модельных объектов были выбраны геномы 44 штаммов *Escherichia coli*, относящиеся к 6 филогруппам. По результатам расчётов было построено филогенетическое дерево, на котором идентифицируются кластеры, хорошо соотносящиеся с классическими результатами мультилокусного анализа.

Ключевые слова: *k-меры, генотипирование, мультилокусный анализ, филогения.*

Using of *k*-mers for intraspecific bacterial phylotyping

Kiselev S.S.[1], Ozoline O.N.[1], Panyukov V.V.[2]

[1]*Institute of Cell Biophysics of Russian Academy of Sciences*
[2]*IMPB RAS – Branch of KIAM RAS*

Since the mid-1970s, the 16S rRNA gene has been used as a standard phylogenetic marker in bacterial taxonomy, but in some cases, based on their sequences, it is not possible to distinguish different types of microorganisms from each other. In connection with this circumstance, the methods for the analysis of complete genomes *in silico* (average nucleotide identity, tetranucleotide frequency correlation coefficient, and others) have recently become widespread. Although their results are well correlated with DNA-DNA hybridization and multilocus typing data, individual subspecies or groups of strains are not always identifiable. For the typing of bacterial strains belonging to the same species, here we propose to use a similarity between genomes expressed as the number of shared *k*-mers, where $k = 12$. The genomes of 44 strains of *Escherichia coli*, belonging to 6 phylogroups, were chosen as the model objects. Based on the results of the calculations, a phylogenetic tree was inferred. The clusters on this tree are well correlated with the classical results of multilocus sequence analysis.

Key words: k-mers, genotyping, multilocus sequence analysis, phylogeny.

1. Введение

Таксономия прокариот включает в себя идентификацию изолятов известных видов микроорганизмов, классификацию новых штаммов, создание новых таксонов и номенклатуру. В первой половине XX века в основном использовались схемы классификации на основе фенотипических признаков, но затем появился метод ДНК-ДНК-гибридизации. Вид бактерий представляет собой группу штаммов (включая типовой штамм), обладающих, как минимум, 70 % идентичности при ДНК-ДНК-гибридизации, и у которых наблюдается различие в температурах плавления ДНК не более 5 °C [1]. Род представляет собой объединение видов в монофилетическую ветвь филогенетического дерева, построенного с использованием нуклеотидных последовательностей генов 16S рРНК [2, 3]. Однако у данного маркера имеется ряд

недостатков. В их число входит присутствие нескольких оперонов рРНК в одном и том же геноме, частично связанное с этим несоответствие между филогенетическими деревьями, построенными исходя из последовательностей 16S рРНК и деревьями на основе других генов, а также неполнота существующих баз данных. Помимо этого, ряд видов бактерий не отделяется друг от друга из-за высокой гомологии генов 16S рРНК [4]. В связи с данными обстоятельствами при построении филогений наряду с использованием генов 16S рРНК применяют мультилокусное типирование (MLSA/MLST) [5, 6], при котором в качестве маркеров, как правило, выступают «гены домашнего хозяйства» (housekeeping), кодирующие белки, необходимые для поддержания важнейших функций клетки.

В наши дни благодаря доступности огромного числа полностью секвенированных геномов всё больше и больше распространяются методы их компьютерного анализа [7]. Расчёт средней нуклеотидной идентичности (ANI) позиционируется в качестве *in silico* замены для традиционной ДНК-ДНК-гибридизации, поскольку при использовании порового уровня в 94–96 % для разделения видов результаты этих методов хорошо соотносятся друг с другом [8]. В программе JSpecies [9] для сравнения геномов наряду с ANI используется индекс корреляции тетрануклеотидов (TETRA), который является параметром, не основанным на выравниваниях. В другом методе (Genome-to-Genome Distance Calculator, GGDC) [10] используется «цифровая» ДНК-ДНК-гибридизация (dDDH) для оценки расстояний между геномами [10, 11]. Показано, что результаты dDDH не только обладают высокой корреляцией с результатами традиционной ДНК-ДНК-гибридизации [10, 12], но также позволяют выделить подвиды [13].

Также разрабатываются методы характеристики геномов, основанные на исследовании олигонуклеотидов длиной k оснований (k-меров). Основным преимуществом таких подходов является использование полного объема данных, т.е. всего генома. Одним из самых простых и эффективных методов сравнения геномов является расчёт попарных сходств или различий между двумя нуклеотидными последовательностями на основании спектра k-меров – нормализованного вектора частот присутствия индивидуального k-мера в геноме. Для решения разных задач разработаны соответствующие алгоритмы, в которых используются различные интервалы значений k. Так, например, при $k = 4–7$ возможны: контроль качества секвенирования [14], идентификация индивидуальных геномов в метагеномных образцах [15–17] или сравнение метагеномов микробных сообществ, выделенных из различных местообитаний [18–20]. Для $k = 15–30$ значительно возрастают вычислительные затраты при обработке полного спектра. В промежуточном диапазоне значений k (7–12) также можно анализировать полный набор k-меров для сравнения геномов с высокой специфичностью и точностью идентификации [17, 21].

2. Материалы и методы

В работе были проанализированы нуклеотидные последовательности геномов 44 штаммов *E. coli*, перечисленных в таблице 1.

Таблица 1. Список штаммов *E. coli*

№ доступа Genbank	Название штамма	Фило-группа
NC_000913.3	K12 MG1655	A
NC_010468.1	ATCC 8739	A
NC_012759.1	BW2952	A
NC_012892.2	BL21(DE3)	A
NC_012947.1	BL21-Gold (DE3) pLysS	A
NC_012967.1	B str. REL606	A
NC_017633.1	O78:H11:K80 str. H10407	A
NC_017638.1	DH1 (ME8569)	A
NC_017663.1	P12b	A
NC_009801.1	O139:H28 str. E24377A	B1
NC_011415.1	SE11	B1
NC_011741.1	IAI1	B1
NC_011748.1	55989	B1
NC_013353.1	O103:H2 str. 12009	B1
NC_013361.1	O26:H11 str. 11368	B1
NC_013364.1	O111:H- str. 11128	B1
NC_016902.1	KO11	B1
NC_017664.1	W	B1
NC_018650.1	O104:H4 str. 2009EL-2050	B1
NC_018658.1	O104:H4 str. 2011C-3493	B1
NC_018661.1	O104:H4 str. 2009EL-2071	B1
NC_022364.1	LY180	B1
NC_002695.1	O157:H7 str. Sakai	E
NC_011353.1	O157:H7 str. EC4115	E
NC_013008.1	O157:H7 str. TW14359	E
NC_013941.1	O55:H7 str. CB9615	E
NC_017656.1	O55:H7 str. RM12579	E
NC_010498.1	SMS-3-5	F
NC_011750.1	IAI39	F
NC_017646.1	O7:K1 str. CE10	F
NC_011751.1	UMN026	D
NC_017626.1	042	D
NC_007946.1	UTI89	B2
NC_008253.1	536	B2
NC_011742.1	S88	B2
NC_011601.1	O127:H6 str. E2348/69	B2
NC_011993.1	LF82	B2
NC_013654.1	SE15	B2
NC_017628.1	IHE3034	B2
NC_017631.1	ABU 83972	B2
NC_017644.1	NA114	B2
NC_017651.1	str. 'clone D i2'	B2
NC_022370.1	PMV-1	B2
NC_022648.1	JJ1886	B2

Значения расстояний между геномами были использованы для построения филогенетического дерева с помощью метода UPGMA (попарная невзвешенная группировка с арифметическим усреднением) [22, 23]. Полученное дерево визуализировано в MEGA 6 [24].

3. Результаты

Для заданного k сходство оценивалось на основании количества k-меров, одновременно используемых в сравниваемых геномах g_1 и g_2. По определению, k-мер M, полинуклеотид размера k, используется геномом, если M, по крайней мере, один раз встречается в какой-либо позиции либо на верхней, либо на нижней цепи генома. Таким образом, при вычислении сходства всякий k-мер учитывается только один раз, независимо от его повторяемости и независимо от его местоположения в геноме. Первоначально сходство sim геномов g_1 и g_2 вычислялось по формуле Сёренсена [25]:

$$sim = \frac{2c}{n_1 + n_2},$$

где n_1, n_2 – количество k-меров, используемых в геномах g_1 и g_2 соответственно, а c – количество тех k-меров, которые используются в обоих геномах одновременно.

Пусть X обозначает ядро группы штаммов G_r, то есть все общие k-меры, используемые каждым геномом $g \in G_r$. Тогда при исследовании геномов в группе G_r формула сходства может быть записана следующим образом:

$$sim = \frac{2c}{2x + m_1 + m_2},$$

где x – размер ядра X, и $m_1 = n_1 - x$, $m_2 = n_2 - x$.

Очевидно, что точность типирования зависит от размера ядра и при определенных отношениях между геномами в группе она может падать. В связи с этим в работе использовалась такая формула сходства:

$$sim = \frac{2c}{m_1 + m_2}.$$

Для геномов 44 штаммов *E. coli* (табл. 1) была рассчитана матрица попарного сходства. Средний размер генома 5 млн. пар нуклеотидов, объём материала ≈ 220 Mb. Вычисления проводились на персональном компьютере оригинальной программой. Для сокращения времени счёта использовался отсортированный массив всех k-меров, присутствующих в геномах штаммов исследуемой группы.

Для филогенетического анализа матрица попарного сходства была преобразована в матрицу расстояний **dist** по формуле

$$\mathbf{dist} = 1 - sim.$$

Число всех возможных вариантов k-меров при $k = 12$ составляет 4^{12}, т.е. 16777216. Основными факторами, ограничивающими число реализованных в геномах k-меров является размер генома и его GC-состав.

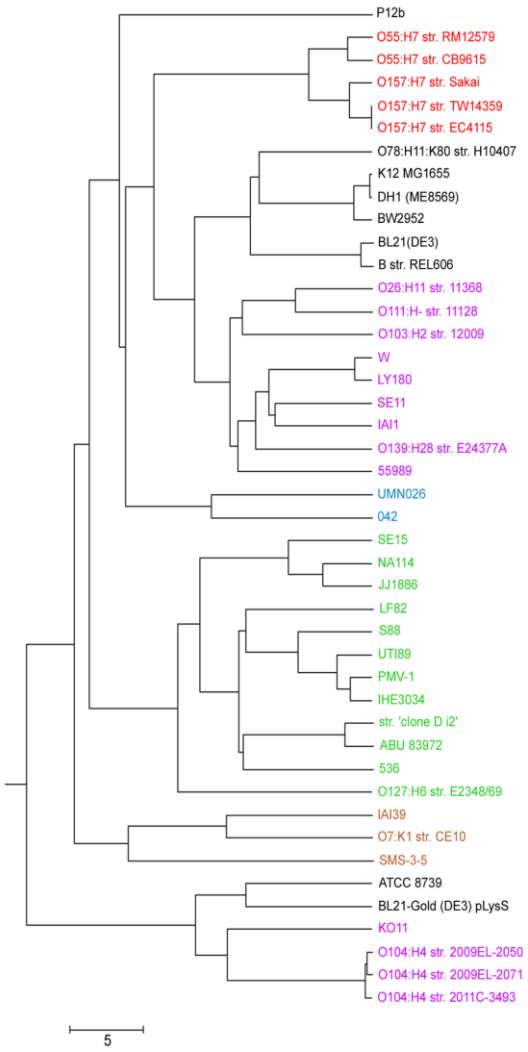

Рис. 1. Филогенетическое дерево, построенное с помощью метода UPGMA [22] на основе матрицы расстояний по использованию 12-меров между полными геномами *E. coli*. Принадлежность отдельных штаммов к 6 различным филогруппам по [26] обозначена соответствующими цветами: чёрным – A, сиреневым – B1, красным – E, коричневым – F, синим – D, зелёным – B2. Масштабная линейка соответствует процентам расстояний между геномами.

Общее количество используемых 12-меров для 44 анализируемых геномов *E. coli* составило 11055581 (т.е. 65.9 % от теоретически возможного), а размер ядра – 1605769.

На рисунке 1 видно, что на филогенетическом дереве в отдельные кластеры выделились штаммы относящиеся к филогруппам B2 (обозначены зелёным цветом), E (обозначены красным цветом), D (обозначены синим цветом) и F (обозначены коричневым цветом) по современной системе типирования [26]. Разделение филогруппы A на несколько кластеров свидетельствует о её гетерогенности. Аналогичная ситуация наблюдается с филогруппой B1. От неё отделился кластер из четырех штаммов, из которых три являются энтероаггрегативными (O104:H4), а KO11 – нет. Возможно, это обусловлено эффектом «притяжения

дальней группы», но не исключено, что при появлении новых геномов *E. coli* филогруппа B1 будет разделена на более мелкие, как это уже произошло с группой D [26]. По старой системе типирования [27] штаммы, ныне относящиеся к группе F, которые присутствуют на нашем дереве в виде отдельной клады (показаны коричневым цветом), принадлежали к группе D.

Ранее в ряде работ для типирования отдельных видов бактерий были использованы наборы видо- или родоспецифичных уникальных *k*-меров [28, 29]. Однако эти методы являются более вычислительно затратными, поскольку для идентификации наборов уникальных олигонуклеотидов необходимо наличие максимально репрезентативной базы геномных данных. Для поиска общих *k*-меров этого не требуется, поскольку анализируются геномы только конкретного вида. Являясь взаимодополняющими, эти два подхода могут быть особенно эффективны при типировании микроорганизмов на разных таксономических уровнях.

4. Благодарности

Авторы приносят благодарность к.б.н. Зимину А.А. (Институт биохимии и физиологии микроорганизмов РАН) за указание на существование компьютерной программы JSpecies, в которой реализована идентификация видов прокариот на основе нуклеотидных последовательностей геномов.

Исследование выполнено за счёт гранта РФФИ (проект № 18-07-00899) и гранта РНФ (проект № 18-14-00348).

5. Список литературы

1. Wayne L.G., Brenner D.J., Colwell R.R., Grimont P.A.D., Kandler O., Krichevsky M.I., Moore L.H., Moore W.E.C., Murray R.G.E., Stackebrandt E., Starr M.P., Truper H.G. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Bacteriol.* 1987. V. 37. P. 463–464. doi: 10.1099/00207713-37-4-463.
2. Woese C.R., Fox G.E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA.* 1977. V. 74. P. 5088–5090. doi: 10.1073/pnas.74.11.5088.
3. Stackebrandt E., Frederiksen W., Garrity G.M., Grimont P.A., Kampfer P., Maiden M.C., Nesme X., Rossello-Mora R., Swings J., Truper H.G., Vauterin L., Ward A.C., Whitman W.B. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.* 2002. V. 52. P. 1043–1047. doi: 10.1099/00207713-52-3-1043.
4. Janda J.M., Abbott S.L. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol*. 2007. V. 45. P. 2761–2764. doi: 10.1128/JCM.01228-07.
5. Gevers D., Cohan F.M., Lawrence J.G., Spratt B.G., Coenye T., Feil E.J., Stackebrandt E., Van de Peer Y., Vandamme P., Thompson F.L., Swings J. Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* 2005. V. 3. P. 733–739. doi: 10.1038/nrmicro1236.
6. Glaeser S.P., Kampfer P. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Syst. Appl. Microbiol.* 2015. V. 38. P. 237–245. doi: 10.1016/j.syapm.2015.03.007.
7. Chun J., Rainey F.A. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *Int. J. Syst. Evol. Microbiol.* 2014. V. 64. P. 316–324. doi: 10.1099/ijs.0.054171-0.
8. Goris J., Konstantinidis K.T., Klappenbach J.A., Coenye T., Vandamme P., Tiedje J.M. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol*. 2007. V. 57. P. 81–91. doi: 10.1099/ijs.0.64483-0.
9. Richter M., Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. USA.* 2009. V. 106. P. 19126–19131. doi: 10.1073/pnas.0906412106.
10. Meier-Kolthoff J.P., Auch A.F., Klenk H.-P., Goker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics.* 2013. V. 14. Article No. 60. doi: 10.1186/1471-2105-14-60.
11. Meier-Kolthoff J.P., Klenk H.P., Goker M. Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age. *Int. J. Syst. Evol. Microbiol.* 2014. V. 64. P. 352–356. doi: 10.1099/ijs.0.056994-0.
12. Auch A.F., Von Jan M., Klenk H.P., Goker M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.* 2010. V. 2. P. 117–134. doi: 10.4056/sigs.531120.
13. Meier-Kolthoff J.P., Hahnke R.L., Petersen J., Scheuner C., Michael V., Fiebig A., Rohde C., Rohde M., Fartmann B., Goodwin L.A., Chertkov O., Reddy T., Pati A., Ivanova N.N., Markowitz V., Kyrpides N.C., Woyke T., Goker M., Klenk H.P. Complete genome sequence of DSM 30083^{T}, the type strain (U5/41^{T}) of *Escherichia coli*, and a proposal for delineating subspecies in microbial taxonomy. *Stand. Genomic Sci.* 2014. V. 9. Article No. 2. doi: 10.1186/1944-3277-9-2.
14. Plaza Onate F., Batto J.M., Juste C., Fadlallah J., Fougeroux C., Gouas D., Pons N., Kennedy S., Levenez F., Dore J., Ehrlich S.D., Gorochov G., Larsen M. Quality control of microbiota metagenomics by *k*-mer analysis. *BMC Genomics*.

2015. V. 16. Article No. 183. doi: 10.1186/s12864-015-1406-7.
15. Zhou F., Olman V., Xu Y. Barcodes for genomes and applications. *BMC Bioinformatics.* 2008. V. 9. Article No 546. doi: 10.1186/1471-2105-9-546.
16. Pride D.T., Meinersmann R.J., Wassenaar T.M., Blaser M.J. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.* 2003. V. 13. P. 145–158. doi: 10.1101/gr.335003.
17. Alsop E.B., Raymond J. Resolving prokaryotic taxonomy without rRNA: longer oligonucleotide word lengths improve genome and metagenome taxonomic classification. *PLoS One.* 2013. V. 8. Article No. 67337. doi: 10.1371/journal.pone.0067337.
18. Silva G.G.Z., Cuevas D.A., Dutilh B.E., Edwards R.A. FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ.* 2014. V. 2. Article No. 425. doi: 10.7717/peerj.425.
19. Langenkamper D., Goesmann A., Nattkemper T.W. AKE - the Accelerated *k*-mer Exploration web-tool for rapid taxonomic classification and visualization. *BMC Bioinformatics.* 2014. V. 15. Article No. 384. doi: 10.1186/s12859-014-0384-0.
20. Liao R., Zhang R., Guan J., Zhou S. A new unsupervised binning approach for metagenomic sequences based on n-grams and automatic feature weighting. *IEEE/ACM Trans. Comput. Biol. Bioinformatics.* 2014. V. 11. P. 42–54. doi: 10.1109/TCBB.2013.137.
21. Jiang B., Song K., Ren J., Deng M., Sun F., Zhang X. Comparison of metagenomic samples using sequence signatures. *BMC Genomics.* 2012. V. 13. Article No. 730. doi: 10.1186/1471-2164-13-730.
22. Sokal R.R., Michener C.D. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* 1958. V. 38. P. 1409–1437.
23. *DendroUPGMA: dendrogram construction using the UPGMA algorithm.* URL: http://genomes.urv.cat/UPGMA/index.php (дата обращения: 12.09.2018).
24. Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* 2013. V. 30. P. 2725–2729. doi: 10.1093/molbev/mst197.
25. Sorensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Kongelige Danske Videnskabernes Selskab. Biol. krifter.* 1948. V. 4. P. 1–34. URL: http://www.royalacademy.dk/Publications/High/295_S%C3%B8rensen,%20Thorvald.pdf (дата обращения: 12.09.2018).
26. Clermont O., Christenson J.K., Denamur E., Gordon D.M. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ. Microbiol. Rep.* 2013. V. 5. P. 58–65. doi: 10.1111/1758-2229.12019.
27. Clermont O., Bonacorsi S., Bingen E. Rapid and simple determination of *Escherichia coli* phylogenetic group. *Appl. Environ. Microbiol.* 2000. V. 66. P. 4555–4558. doi: 10.1128/AEM.66.10.4555-4558.2000.
28. Petit R.A.III, Hogan J.M, Ezewudo M.N., Joseph S.J., Read T.D. Fine-scale differentiation between *Bacillus anthracis* and *Bacillus cereus* group signatures in metagenome shotgun data. *PeerJ.* 2018. V. 6. Article No. e5515. doi: 10.7717/peerj.5515.
29. Panyukov V.V., Kiselev S.S., Alikina O.V., Nazipova N.N., Ozoline O.N. Short unique sequences in bacterial genomes as strain- and species-specific signatures. *Mathematical Biology and Bioinformatics.* 2017. V. 12. P. 547–558. doi: 10.17537/2017.12.547.