



DOI: <https://doi.org/10.15688/jvolsu11.2016.4.3>

УДК 577.21

ББК 28.04+22.18

ПРОСТРАНСТВЕННАЯ МОДЕЛЬ ГЕНЕТИЧЕСКОГО КОДА В ВИДЕ ПОЛУПРАВИЛЬНОГО МНОГОГРАННИКА «УСЕЧЕННЫЙ ОКТАЭДР»

Ян Илларионович Корчмарюк

Кандидат технических наук,
главный специалист УНИТ ВЦ,
Волгоградский государственный технический университет
yakor77@yandex.ru
просп. им. Ленина, 28, 400005 г. Волгоград, Российская Федерация

Александр Владимирович Коваленко

Старший преподаватель кафедры биоинженерии и биоинформатики,
Волгоградский государственный университет
biobio@volsu.ru
просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация

Аннотация. В настоящей работе, относящейся к области биоинформатики, математико-статистическим методом факторного анализа анализируются все $4! = 24$ перестановки чисел «1, 2, 3, 4» как 4 фактора. В результате численного анализа получены 4 факторные координаты (x, y, z). Минимальные евклидовы расстояния между всеми 24 вершинами образуют ребра фигуры «усеченный октаэдр». Каждой вершине усеченного октаэдра сопоставлен триплет – кодон генетического кода и соответствующая ему аминокислота. Построены графики проекций фигуры «усеченный октаэдр» и изометрическое (трехмерное) ее изображение. Проанализированы обнаруженные закономерности в физических и химических свойствах аминокислот, распределенных по вершинам построенной фигуры. Произведена интерпретация трех факторных осей координат.

Ключевые слова: усеченный октаэдр, факторный анализ, генетический код, перестановки, биоинформатика.

Методы математического анализа давно и широко используются для изучения генетического кода [4; 12; 20]. Наиболее популярен кластерный анализ [13; 18], а также методы комбинаторики [7] с широкими возможностями перестановок [6; 8; 10; 11]. Но с исходным представлением генетического кода на уровне кибернетики до сих пор существуют различные вариации. Мы обратили внимание, что пространственные модели лучше опи-

сывают взаимосвязь элементов в системе, а усеченный октаэдр встречается только в поле изучения скаффолдов [17]. Хотя он может быть основой для связи нескольких десятков элементов друг с другом, его до настоящего момента не рассматривали как каркас для генетического кода. В связи с вышесказанным цель нашей работы – построить пространственную модель генетического кода на основе усеченного октаэдра.

Материалы и методы

Исходными данными были взяты $4! = 24$ всевозможные перестановки натуральных целых положительных чисел $\{1, 2, 3, 4\}$ [5; 9]. В качестве математико-статистического метода обработки данных был взят классический метод факторного анализа в программном комплексе «SPSS» с использованием методов главных компонент, Варимакс и нормализации Кайзера [1; 3; 16].

С целью построения трехмерной фигуры по данным трех факторных координат были построены евклидовы расстояния $(R_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2})$ между всеми $\{i, j\}$ 24 вершинами усеченного октаэдра. Все расстояния были нормированы на наименьшее из расстояний. Из всех нормированных расстояний была вычтена 1 (единица), для того, чтобы расстояние у вершин было равно 0 (нулю). Затем были выбраны все пары вершин, между которыми было нормированное 1 (единичное) расстояние, которое было взято за ребро усеченного октаэдра.

Для графического представления итоговых данных по усеченному октаэдру использовались графики трех проекций факторных осей.

Для корректности построенной аналогии 4 нуклеотида ДНК/РНК (CUAG) были упорядочены по возрастанию их молекулярной массы [2; 14; 15; 19] и поставлены в соответствие меткам вершин усеченного октаэдра (1234), следующим образом: 1 = C, 2 = U или T, 3 = A, 4 = G.

Результаты

Построенная модель «усеченного октаэдра» обладает достаточной мерой общности, чтобы интерпретировать ее в различных теоретических приложениях. В настоящей статье авторами данная модель интерпретирована как трехмерный аналог известной кодовой таблицы 4 нуклеотидов и кодируемых ими 20 аминокислот. Из кодовой таблицы ДНК были выбраны все аминокислоты, которые соответствовали кодонам с неповторяющимися тройками нуклеотидов. Эти аминокислоты были приписаны соответствующим вершинам усеченного октаэдра по первым трем меткам четырехзначного кода их вершин. Старт-коды и стоп-коды были обозначены везде как «STOP» (рис. 1–4). Данные приведены в таблице.

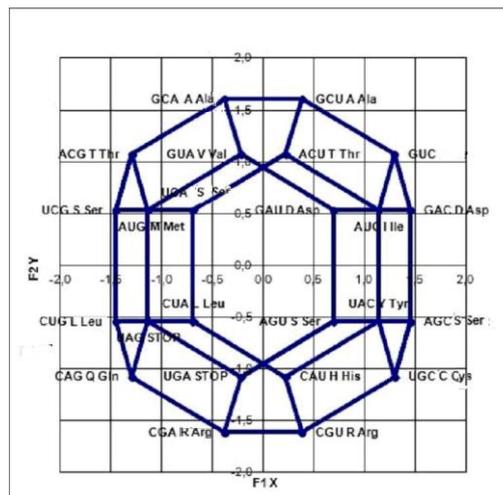


Рис. 1. Генетический код в виде усеченного октаэдра, проекция XY

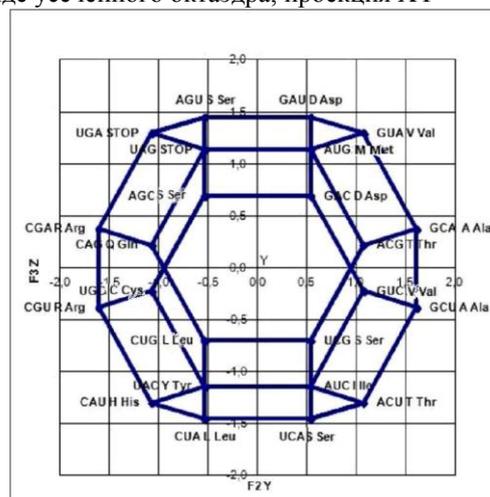


Рис. 2. Генетический код в виде усеченного октаэдра, проекция YZ

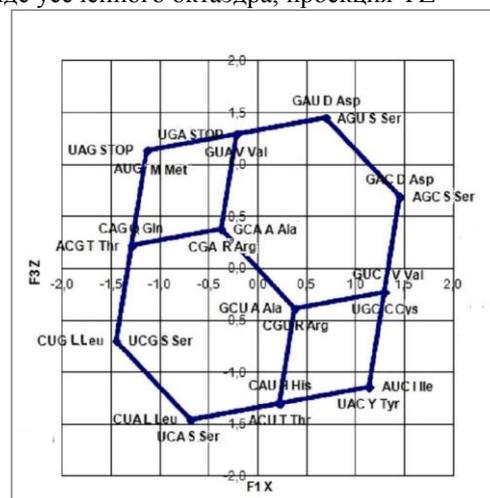


Рис. 3. Генетический код в виде усеченного октаэдра, проекция XZ

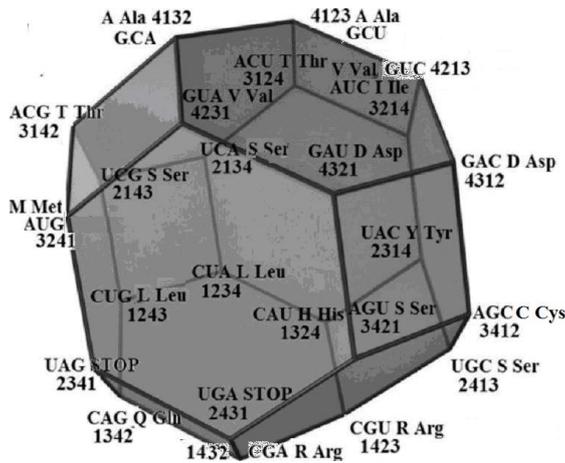


Рис. 4. Генетический код в виде усеченного октаэдра, проекция 3D

Обсуждение

Поскольку все вершины усеченного октаэдра лежат на поверхности описанной около него сферы, а все комбинации перестановок, начинающихся на 1, 2, 3 или 4, – лежат в одной плоскости сечения, то эти перестановки можно интерпретировать как «параллели глобуса», а связывающие их ребра – как «меридианы глобуса». При этом полученная трехмерная модель генетического кода обладает высокой симметрией,

почти сферической, и выделяет подгруппу высокоэнтропийных (невыврожденных) кодонов и их аминокислот из общей группы генетического кода, а также подгруппу «ближлежащие и даль- нележащие коды и аминокислоты». Те, что в вершинах 6-угольников и 4-угольников.

Так как все расстояния в этой модели укладываются на логарифм (значения F1–F3 таблицы), то появляется возможность анализировать корреляции этих 10 типовых расстояний с параметрами кодонов и аминокислот.

Заключение

В силу большой общности построенной модели ее можно использовать в любых кодовых системах, которые имеют базовый алфавит в 24 (или менее) знаков. Например, в математической лингвистике, для букв и звуков алфавитов различных языков мира. Также построенную модель можно использовать в области искусственного интеллекта, если каждую вершину считать «формальным нейроном» в рамках перцептронной парадигмы Розенблатта – Мак-Каллоха – Питса, а весовые коэффициенты межнейронных связей приписывать ребрам усеченного октаэдра.

Соотнесение генетического кода с усеченным октаэдром

Генетический код CUAG				Метка перестановок	Факторные оси		
C	U	A	L Leu		F1 = X	F2 = Y	F3 = Z
C	U	A	L Leu	1234	-0,69324	-0,53619	-1,45152
C	U	G	L Leu	1243	-1,45152	-0,53619	-0,69324
C	A	G	Q Gln	1342	-1,29448	-1,07238	0,22210
C	G	A	R Arg	1432	-0,37914	-1,60857	0,37914
C	G	U	R Arg	1423	0,37914	-1,60857	-0,37914
C	A	U	H His	1324	0,22210	-1,07238	-1,29448
U	C	A	S Ser	2134	-0,69324	0,53619	-1,45152
U	C	G	S Ser	2143	-1,45152	0,53619	-0,69324
U	A	G	STOP	2341	-1,13743	-0,53619	1,13743
U	G	A	STOP	2431	-0,22210	-1,07238	1,29448
U	G	C	C Cys	2413	1,29448	-1,07238	-0,22210
U	A	C	Y Tyr	2314	1,13743	-0,53619	-1,13743
A	C	U	T Thr	3124	0,22210	1,07238	-1,29448
A	C	G	T Thr	3142	-1,29448	1,07238	0,22210
A	U	G	M Met	3241	-1,13743	0,53619	1,13743
A	G	U	S Ser	3421	0,69324	-0,53619	1,45152
A	G	C	S Ser	3412	1,45152	-0,53619	0,69324
A	U	C	I Ile	3214	1,13743	0,53619	-1,13743
G	C	U	A Ala	4123	0,37914	1,60857	-0,37914
G	C	A	A Ala	4132	-0,37914	1,60857	0,37914
G	U	A	V Val	4231	-0,22210	1,07238	1,29448
G	A	U	D Asp	4321	0,69324	0,53619	1,45152
G	A	C	D Asp	4312	1,45152	0,53619	0,69324
G	U	C	V Val	4213	1,29448	1,07238	-0,22210

СПИСОК ЛИТЕРАТУРЫ

1. Козлов, Н. Математический анализ генетических кодов / Н. Козлов // Математическая биология и биоинформатика. – 2006. – Т. 1, № 1. – С. 70–96.
2. Козлов, Н. Один способ хранения генетической информации / Н. Козлов // Математическое моделирование. – 2002. – Т. 14, № 8. – С. 72–78.
3. Козлов, Н. О расчете генетического кода / Н. Козлов // Математическое моделирование. – 2011. – Т. 23, № 6. – С. 3–17.
4. Козлов, Н. Тайны генетического кода: взгляд математика / Н. Козлов // Математическое моделирование. – 2015. – Т. 27, № 4. – С. 64–80.
5. Корчмарюк, Я. Усеченный октаэдр как результат факторного анализа полной группы перестановок из чисел {1, 2, 3, 4} / Я. Корчмарюк // Сборник ВорГТУ «Современные проблемы информатизации в технике и технологиях». – 2002. – № 7. – С. 58–60.
6. Макаров, М. Антимонотонные перестановки / М. Макаров // Сибирские электронные математические известия. – 2012. – Т. 9. – С. 346–359.
7. A universal combinatorial design of antibody framework to graft distinct CDR sequences: a bioinformatics approach / J. Haidar, Q. Yuan, L. Zeng [et al.] // *Proteins*. – 2012. – Vol. 80, № 3. – P. 896–912.
8. Avgustinovich, S. Infinite permutations of lowest maximal pattern complexity / S. Avgustinovich, A. Frid, T. Kamae [et al.] // *Theoretical Computer Science*. – 2011. – Vol. 412. – P. 2911–2921.
9. Bona, M. Exact enumeration of 1342-avoiding permutations: a close link with labeled trees and planar maps / M. Bona // *Journal of Combinatorial Theory*. – 1997. – Vol. 80, № 2. – P. 257–272.
10. Chen, W. Minimal permutations and 2-regular skew tableaux / W. Chen, C. Gu, K. Ma // *Advances in Applied Mathematics*. – 2011. – Vol. 47, № 4. – P. 795–812.
11. Fon-Der-Flaass, D. On periodicity and low complexity of infinite permutations / D. Fon-Der-Flaass, A. Frid // *European Journal of Combinatorics*. – 2007. – Vol. 28, № 8. – P. 2106–2114.
12. Ma, H. iFad: an integrative factor analysis model for drug-pathway association inference / H. Ma, H. Zhao // *Bioinformatics*. – 2012. – Vol. 28, № 14. – P. 1911–1918.
13. Madeira, S. Biclustering algorithms for biological data analysis: A survey / S. Madeira, A. Oliveira // *IEEE ACM Transactions on Computational Biology and Bioinformatics*. – 2004. – Vol. 1, № 1. – P. 24–45.
14. Maraia, R. Different types of secondary information in the genetic code / R. Maraia, J. Iben // *RNA*. – 2014. – Vol. 20, № 7. – P. 977–984.
15. Overlapping of Genes in the Human Genome / T. Nakayama, S. Asai, Y. Takahashi [et al.] // *Int. J. Biomed. Sci.* – 2007. – Vol. 3, № 1. – P. 14–19.

16. Pournara, I. Factor analysis for generegulatory networks and transcription factor activity profiles / I. Pournara, L. Wernish // *BMC Bioinformatics*. – 2007. – Vol. 8. – P. 61.
17. Scaffold library for tissue engineering: a geometric evaluation / N. Chan tarapan ich, P. Puttawibul, S. Sucharitpawatskul [et al.] // *Computational and Mathematical Methods in Medicine*. – 2012. – Vol. 2012. – ArticleID 407805. – P. 1–14.
18. Sparse group factor analysis for biclustering of multiple data sources / L. Bunte, E. Leppaaho, I. Saarinen [et al.] // *Bioinformatics*. – 2016. – Vol. 32, № 16. – P. 2457–2463.
19. Whelan, S. The genetic code can cause systematic bias in simple phylogenetic models / S. Whelan // *Phil. Trans. R. Soc. B*. – 2008. – Vol. 363. – P. 4003–4011.
20. Zur, H. Exploiting hidden information interleaved in the redundancy of the genetic code without prior knowledge / H. Zur, T. Tuller // *Bioinformatics*. – 2015. – Vol. 31, № 8. – P. 1161–1168.

REFERENCES

1. Kozlov N. Matematich eskiy an aliz geneticheskikh kodov [Mathematical Analysis of Genetic Codes]. *Matematicheskaya biologiya i bioinformatika*, 2006, vol. 1, no. 1, pp. 70-96.
2. Kozlov N. Odin sposob kh r an en iya geneticheskoy informatsii [One Way to Store Genetic Information]. *Matematicheskoe modelirovanie*, 2002, vol. 14, no. 8, pp. 72-78.
3. Kozlov N. O raschete geneticheskogo koda [On Calculation of the Genetic Code]. *Matematicheskoe modelirovanie*, 2011, vol. 23, no. 6, pp. 3-17.
4. Kozlov N. Tayny geneticheskogo koda: vzglyad matematika [Secrets of the Genetic Code: a View of Mathematician]. *Matematicheskoe modelirovanie*, 2015, vol. 27, no. 4, pp. 64-80.
5. Korchmaryuk Ya. Usechennyy oktaedr kak rezultat faktornogo analiza polnoy gruppy perestanolov iz chisel {1,2,3,4} [The Truncated Octahedron as a Result of Factor Analysis of th e Complete Group of Permutations of the Numbers {1,2,3,4}]. *Sbornik VorGTU «Sovremennye problemy informatizatsii v tekhnike i tekhnologiyakh»* [Collection of VorGTU “Modern Problems of Informatization in Engineering and Technology”], 2002, no. 7, pp. 58-60.
6. Makarov M. Antimonotonnye perestanolvki [Antimonotone Reshuffle]. *Sibirskie elektronnye matematicheskie izvestiya*, 2012, vol. 9, pp. 346-359.
7. Haidar J., Yuan Q., Zeng L., et al. A universal combinatorial design of antibodyframework to graft distinct CDR sequences: a bioinformatics approach. *Proteins*, 2012, vol. 80, no.3, pp. 896-912. DOI:10.1002/prot.23246.

8. Avgustinovich S., Frid A., Kamae T., et al. Infinite permutations of lowest maximal pattern complexity. *Theoretical Computer Science*, 2011, vol. 412, pp. 2911-2921.
9. Bona M. Exact Enumeration of 1342-Avoiding Permutations: A Close Link with Labeled Trees and Planar Maps. *Journal of Combinatorial Theory*, 1997, vol. 80, no. 2, pp. 257-272.
10. Chen W., Gu C., Ma K. Minimal permutations and 2-regular skew tableaux. *Advances in Applied Mathematics*, 2011, vol. 47, no. 4, pp. 795-812.
11. Fon-Der-Flaass D., Frid A. On periodicity and low complexity of infinite permutations. *European Journal of Combinatorics*, 2007, vol. 28, no. 8, pp. 2106-2114.
12. Ma H., Zhao H. iFad: an integrative factor analysis model for drug-path way association inference. *Bioinformatics*, 2012, vol. 28, no. 14, pp. 1911-1918. DOI:10.1093/bioinformatics/bts285.
13. Madeira S., Oliveira A. Bicluster in g algorithms for biological data analysis: A survey. *IEEE ACM Transactions on Computational Biology and Bioinformatics*, 2004, vol. 1, no 1, pp. 24-45.
14. Marais R., Iben J. Different types of secondary information in the genetic code. *RNA*, 2014, vol. 20, no. 7, pp. 977-984.
15. Nakayama T., Asai S., Takahashi Y., et al. Overlapping of Genes in the Human. Genome. *Int. J. Biomed. Sci.*, 2007, vol. 3, no. 1, pp. 14-19.
16. Pournara I., Wernisch L. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, 2007, vol. 8, pp. 61. DOI: 10.1186/1471-2105-8-61.
17. Chantapanich N., Puttawibul P., Sucharitpwatskul S., et al. Scaffold library for tissue engineering: a geometric evaluation. *Computational and Mathematical Methods in Medicine*, 2012, vol. 2012, pp.1-14. DOI:10.1155/2012/407805.
18. Bunte L., Leppaaho E., Saarinen I., et al. Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics*, 2016, vol. 32, no. 16, pp. 2457-2463. DOI:10.1093/bioinformatics/btw207.
19. Whelan S. The genetic code can cause systematic bias in simple phylogenetic models. *Phil. Trans. R. Soc. B*, 2008, vol. 363, pp. 4003-4011. DOI:10.1098/rstb.2008.0171.
20. Zur H., Tuller T. Exploiting hidden information interleaved in the redundancy of the genetic code without prior knowledge. *Bioinformatics*, 2015, vol. 31, no. 8, pp. 1161-1168. DOI: 10.1093/bioinformatics/btu797.

THE SPATIAL MODEL OF GENETIC CODE IN THE FORM OF A SEMIREGULAR POLYHEDRON “TRUNCATED OCTAHEDRON”

Yan Illarionovich Korchmaryuk

Candidate of Technical Sciences,
Chief Specialist of the Computer Center,
Volgograd State Technical University
yakor77@yandex.ru
Prosp. Lenina, 28, 400005 Volgograd, Russian Federation

Aleksandr Vladimirovich Kovalenko

Assistant Professor, Department of Bioengineering and Bioinformatics,
Volgograd State University
biobio@volsu.ru
Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation

Abstract. This work applies to the field of bioinformatics and analyzes all $4! = 24$ permutations of numbers 1, 2, 3, 4 as 4 factors by means of factor analysis. Computational analysis helped us to get 4 factor coordinates – x, y, z. Minimum Euclidean distances between all 24 vertices create edges of truncated octahedron. Every vertex of truncated octahedron matches with a triplet (it is the genetic code’s codon) and appropriate amino acid. We have created the graph of truncated octahedron’s projections and its isometric (three-dimensional) image. We have also analyzed detected sequences in physical and chemical characteristics of amines which are located in vertices of our figure. We have interpreted three factor axis of coordinate system.

Key words: truncated octahedron, factor analysis, genetic code, permutations, bioinformatics.