***Доклады Международной конференции "Математическая биология и биоинформатика"**. Под ред. В.Д. Лахно. Том 8. Пущино: ИМПБ РАН, 2020. Статья № e22. doi: 10.17537/icmbb20.6*

Применение методов глубокого обучения и молекулярного моделирования для идентификации потенциальных ингибиторов проникновения ВИЧ-1

Николаев Г.И.[1], Шульдов Н.А.[2], Босько И.П.[1], Анищенко А.И.[2], Тузиков А.В.[1], Андрианов А.М.[3*]

[1]*Объединенный институт проблем информатики НАН Беларуси*
[2]*Белорусский государственный университет*
[3]*Институт биоорганической химии НАН Беларуси*

*alexande.andriano@yandex.ru

Несмотря на значительный прогресс, достигнутый в разработке новых эффективных ингибиторов проникновения ВИЧ-1, в настоящее время нет лицензированных противовирусных препаратов, основанных на ингибировании критических взаимодействий белка gp120 оболочки ВИЧ-1 с клеточным рецептором CD4. В связи с этим исследования по разработке терапевтических средств, ингибирующих связывание белка gp120 с молекулой CD4, остаются чрезвычайно актуальными. В настоящей работе с помощью технологий глубокого обучения разработана генеративная состязательная нейронная сеть для рационального дизайна низкомолекулярные соединений, способных блокировать CD4-связывающий сайт белка gp120 ВИЧ-1. Проведено тестирование нейронной сети на широком наборе молекул из базы данных ZINC15. Показано, что совместное использование нейронной сети с виртуальным скринингом молекулярных библиотек и методами молекулярного моделирования формирует продуктивную платформу для идентификации базовых структур, перспективных для создания новых противовирусных препаратов, ингибирующих ранние стадии развития ВИЧ-инфекции.

Ключевые слова: *методы глубокого обучения, генеративный состязательный автоэнкодер, ВИЧ-1, белок gp120, ингибиторы проникновения ВИЧ-1, методы молекулярного моделирования, лекарственные препараты против ВИЧ.*

Application of Deep Learning and Molecular Modelling Methods to Identify Potential HIV-1 Entry Inhibitors

Nikolaev G.I.[1], Shuldov N.A.[2], Bosko I.P.[1], Anischenko A.I.[2], Tuzikov A.V.[1], Andrianov A.M.[3*]

[1]*United Institute of Informatics Problems, National Academy of Sciences of Belarus*
[2]*Belarusian State University*
[3]*Institute of Bioorganic Chemistry, National Academy of Sciences of Belarus*
*alexande.andriano@yandex.ru

Despite significant progress in the development of novel potent HIV-1 entry inhibitors, there are currently no licensed antiviral drugs based on inhibiting the critical interactions of the HIV-1 envelope gp120 protein with cellular receptor CD4. In this connection, studies on the design of new small-molecule compounds able to block the gp120-CD4 binding are still of great value. In this work, methods of deep learning have been used to develop a generative adversarial neural network for the rational design of small-molecule compounds able to block CD4-binding site of the HIV-1 envelope gp120 protein. The neural network was validated on a wide range of compounds from the ZINC15 database. The use of the neural network in combination with virtual screening of chemical databases and molecular modeling was shown to form a productive platform to identify basic structures promising for the development of novel antiviral drugs that inhibit the early stages of HIV infection.

Key words: deep learning methods, a generative adversarial neural network, HIV-1, gp120 protein, HIV-1 entry inhibitors, molecular modeling, anti-HIV drugs.

1. Введение

Современные методы компьютерного конструирования лекарств значительно расширяют возможности фармацевтической промышленности, сокращая время и затраты, необходимые для разработки новых терапевтических агентов. На сегодняшний день компьютерное проектирование потенциальных лекарств с использованием методов машинного обучения является одним из наиболее важных и быстро развивающихся направлений хемо- и биоинформатики [1].

В последние годы появилось большое количество исследований, посвященных использованию методов машинного обучения для прогнозирования потенциальных ингибиторов ВИЧ-1 и устойчивости вируса к препаратам против ВИЧ [2]. Однако все эти исследования сосредоточены на вирусных ферментах, а именно на обратной транскриптазе, протеазе и интегразе [3]. Соединения, блокирующие эти ферменты, не могут предотвратить проникновение вируса в клетку-мишень, что привлекает внимание к ингибиторам ВИЧ-1, которые могут вмешиваться в начальные стадии цикла заражения вируса, препятствуя его адсорбции на CD4-клетках и/или слиянию мембран вируса и клетки-мишени [4].

Цель настоящей работы заключалась в разработке и применении генеративного состязательного автоэнкодера для идентификации потенциальных ингибиторов проникновения ВИЧ-1, которые нацелены на гидрофобную полость Phe-43 белка gp120, критически важную для связывания вируса с клеточным рецептором CD4 [5].

Для достижения этой цели были проведены следующие исследования:

1. Построена архитектура генеративного состязательного автоэнкодера;
2. Сформирована виртуальная библиотека низкомолекулярных химических соединений, содержащая потенциальные анти-ВИЧ агенты для обучения нейронной сети;
3. Проведен молекулярный докинг соединений из этой библиотеки с белком gp120 и рассчитаны значения свободной энергии связывания;
4. Созданы молекулярные дескрипторы химических соединений из набора обучающих данных;
5. Осуществлено обучение нейронной сети с последующей проверкой результатов обучения и работы автоэнкодера;
6. Совместное применение разработанной нейронной сети с методами виртуального скрининга и молекулярного моделирования для идентификации базовых структур, перспективных для создания новых противовирусных препаратов, ингибирующих ранние стадии развития ВИЧ-инфекции.

Методы молекулярного докинга, и квантовой химии использовали при тестировании нейронной сети для идентификации потенциальных анти-ВИЧ агентов из множества соединений, обнаруженных в результате скрининга библиотеки молекулярных дескрипторов, сконструированных для химических соединений из базы данных ZINC15. Значения энергии связывания, аппроксимированные оценочными функциями этих методов, сравнивали с величинами, предсказанными с помощью идентичных вычислительных протоколов для ингибиторов ВИЧ-1 NBD-11021 и NBD-14010, представляющих новое поколение антагонистов рецептора CD4 [6, 7].

2. Методы

2.1. Разработка генеративного состязательного автоэнкодера

На рисунке 1 показана архитектура разработанного генеративного состязательного автоэнкодера. Модель состоит из двух нейронных сетей, включающих автоэнкодер и дискриминатор, которые работают во время обучения в соревновательном режиме, что позволяет настроить параметры автоэнкодера в режиме обучения и обеспечить получение выходных данных высокого качества на следующем этапе их генерации. Задача дискриминатора – отличать реальные данные от данных, генерируемых энкодером. Разработанный автоэнкодер представляет собой семислойную нейронную сеть с входным и выходным слоями, скрытым слоем и четырьмя полностью связанными слоями (рис. 1).

Молекулярные дескрипторы поступают на входной слой, данные которого проходят через два полностью связанных слоя (кодировщик) и попадают на скрытый слой, где к полученному результату добавляется численная оценка энергии связывания с молекулярной мишенью. Затем молекулярные дескрипторы проходят через два полностью связанных слоя (декодер) и попадают на выход, который, как и вход, является вектором молекулярного дескриптора. Сеть, работающая в этом режиме, позволяет уменьшить количество нейронов, входящих в латентный слой, содержащий сжатую информацию о векторе, который подается на вход сети, с последующим его расширением на выходе. Скрытый слой состоит из трех нейронов, два из которых получают значения от кодировщика, а третий получает значение свободной энергии связывания. В режиме генерации автоэнкодера случайные числа вводятся на два нейрона латентного слоя, а пороговое значение свободной энергии связывания поступает на третий нейрон, а затем они проходят через декодер, генерируя молекулярные дескрипторы с заданным пороговым значением энергии.

Разработанный состязательный автоэнкодер основан на модели нейронной сети, которая была создана для генерации химических соединений с противораковыми свойствами [8], но имеет следующие особенности (рис. 1):

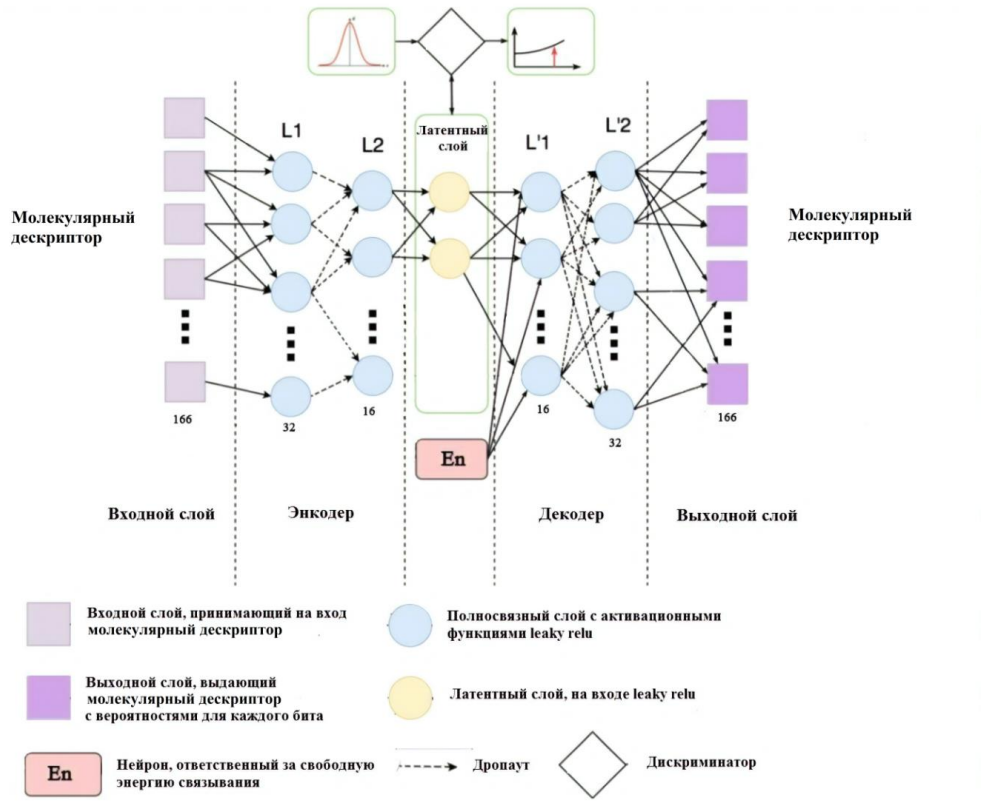

Рис. 1. Архитектура нейронной сети для генерации потенциальных ингибиторов ВИЧ-1, блокирующих CD4-связывающий сайт белка gp120 оболочки вируса

— На скрытом слое используется нейрон, отвечающий за значения свободной энергии связывания. Этот нейрон не взаимодействует с кодировщиком и подается только на вход декодера вместе с данными, полученными кодером. Скрытый слой содержит 3 нейрона;

— Кодировщик состоит из двух последовательных уровней L1 и L2 с 32 и 16 нейронами соответственно. Декодер включает два слоя L'1 и L'2, содержащие 16 и 32 нейрона соответственно;

— Дискриминатор состоит из четырех слоев, включающих 2, 16, 3 и 1 нейрон соответственно;

— На промежуточных уровнях автоэнкодера используется функция активации Leaky ReLU[9];

— Сигмоидные функции активации используются на всех уровнях дискриминатора;

— Для обеспечения дополнительного уровня защиты от перетренированности между двумя полностью подключенными уровнями кодера и декодера используется функция отсева, позволяющая нейронной сети полностью использовать свои параметры (веса) и выборочное случайное отключение во время обучения.

2.2. Формирование обучающего набора данных

Набор обучающих данных был создан с помощью программного пакета AutoClickChem [10]. Эта программа реализует методологию клик-химии *in silico* для генерации большого набора кандидатов в лекарственные препараты, которые включают в себя фрагменты, обеспечивающие лиганд биологической активностью по отношению к целевому белку. На первом этапе с помощью виртуального скрининга подмножества “Drug-Like” базы данных ZINC15 [11] были созданы две библиотеки соединений, содержащие небольшие молекулы с функциональными группами, участвующими в клик-реакции азид-алкинового циклоприсоединения [12]. Для этого использовали программу DataWarrior (URL: http://www.openmolecules.org/help/basics.html). В результате были отобраны ароматические соединения с молекулярной массой < 250 Да, содержащие азидные или алкиновые группы, и помещены в библиотеку 1, а все низкомолекулярные соединения (молекулярная масса < 250 Да) с этими функциональными группами были собраны в библиотеке 2. Выбор ароматических соединений в качестве основы для конструирования потенциальных кандидатов в миметики молекулы CD4 объясняется тем фактом, что их ароматические системы могут имитировать ключевые взаимодействия остатка Phe-43 рецептора CD4 с гидрофобной полостью Phe-43 белка gp120. Как следует из данных рентгеноструктурного анализа [5], ароматическое кольцо $Phe\text{-}43_{CD4}$ погружено в полость $Phe\text{-}43_{gp120}$, что приводит к блокированию остатков белка gp120, критических для адсорбции вируса на клетках CD4+ Т. Кроме того, новые эффективные антагонисты проникновения ВИЧ-1, такие как NBD-11021, NBD-

14010 и NBD-14189, демонстрируют аналогичные способы взаимодействия для связывания с этим гидрофобным “карманом” gp120 [6, 7].

Библиотеки 1 и 2 содержали 1388 и 3769 соединений соответственно. Затем эти соединения были использованы в качестве реагентов для имитации клик-реакции азид-алкинового циклоприсоединения с помощью программы AutoClickChem [10], что привело к генерации 1 655 301 гибридной молекулы. 120 000 соединений из сконструированных молекул, которые полностью удовлетворяли «правилу пяти» Липинского [13], были включены в набор обучающих данных.

Молекулярный докинг отобранных соединений с белком gp120 выполняли с помощью программного пакета QuickVina 2 [14] в приближении жесткого рецептора и гибких лигандов. Структура белка gp120 была выделена из комплекса gp120-CD4-Ab17b в кристалле [5]. К структурам белка gp120 и лигандов были добавлены атомы водорода (программа OpenBabel; (http://openbabel.org/wiki/Main_Page) и проведена их оптимизация в силовом поле UFF. Ячейка для докинга представляла фрагмент структуры белка gp120 [5] с координатами $x \in$ (24 Å; 34 Å), $y \in$ (–15 Å; –5 Å), $z \in$ (78 Å; 88 Å), включающий Phe^{43}–полость гликопротеина; т. е. ее объем составлял $10 \times 10 \times 10 = 1000$ Å^3. Параметр, характеризующий полноту поиска (охват конформационного пространства), был задан равным 50 [14].

Молекулярные дескрипторы соединений из набора обучающих данных – “ключи” MACCS (http://www.dalkescaught.com/writings/NBN/fingerprints.html) были рассчитаны с помощью программного пакета RDKit с открытым исходным кодом (https://www.rdkit.org/). При этом использовали опцию MACCS166KeyBits “ключей” MACCS, представляющую молекулярный дескриптор в виде битовой векторной строки из 166 битов, содержащей нули и единицы (http://www.dalkescientific.com/writings/NBN/fingerprints.html; https://www.rdkit.org/).

2. Оценка результатов обучения и работы автоэнкодера

Для тестирования работы автоэнкодера с помощью программного пакета RDKit была создана библиотека структурных “ключей” MACCS для 21 325 567 соединений из библиотеки “Drug Like” базы данных ZINC15 [11] и рассчитаны 5 молекулярных дескрипторов для сгенерированных автоэнкодером соединений при пороговых значениях энергии связывания с белком gp120, равных –5 ккал/моль и –8 ккал/моль.

В результате виртуального скрининга созданной библиотеки молекулярных дескрипторов в базе данных ZINC15 были найдены три соединения с подобными молекулярными дескрипторами и имеющие низкие значения энергии связывания согласно данным программы QuickVina 2 и полуэмпирического квантово-химического метода PM7 [15] (табл. 1). При этом в качестве меры подобия молекулярных дескрипторов использовали расстояние Хэмминга, определяемое в теории кодирования как число пар несовпадающих компонент сравниваемых векторов [18], и коэффициент Танимото, который вычисляли по формуле [19]:

$$T(a,b) = \frac{N_c}{N_a + N_b - N_c}$$

где T – коэффициент Танимото, принимающий значения от 0 до 1; N_a – количество элементов в первом векторе; N_b – количество элементов во втором векторе; N_c – количество одинаковых элементов в двух векторах. В процессе скрининга библиотеки молекулярных дескрипторов из базы данных ZINC15 отбирали соединения, для которых коэффициент Танимото удовлетворял условию $T > 0.85$ [19].

Таблица 1. Значения энергии связывания (ΔG) и констант диссоциации (K_d), рассчитанные для комплексов лиганд/gp120 методами молекулярного докинга и квантовой химии

Код соединения в базе данных ZINC15	ΔG_{DOCK}[1], ккал/моль	K_d[2] (мкмоль)	ΔG_{Kd}[3], ккал/моль	ΔG_{PM7}[4], ккал/моль
ZINC000026430653	–8.8	0.788	–8.6	–39.2
ZINC000191389930	–8.0	1.061	–8.5	–40.8
ZINC000293423658	–7.7	9.591	–7.1	–38.8
NBD11021	–8.4	0.948	–8.5	–29.6
NBD14010	–8.9	0.039	–10.5	–42.8

[1]Значения ΔG в соответствии с оценочной функцией QuickVina 2;
[2]Значения K_d, вычисленные с использованием оценочной функции NNScore 2.0 [16];
[3]Значения ΔG, рассчитанные на основе значений K_d по формуле $\Delta G = R \times T \times \ln(K_d)$ (где ΔG – энергия связывания, R – универсальная газовая постоянная, T – абсолютная температура, равная 310 K) [17];
[4]Значения ΔG, предсказанные на основе расчетов методом PM7 для комплексов лиганд / gp120.

Анализ полученных данных показывает, что совместное использование разработанной нейронной сети с виртуальным скринингом библиотеки молекулярных дескрипторов позволяет идентифицировать лиганды с более низкими значениями свободной энергии связывания по сравнению с заданным пороговым значением. Кроме того, соединения с кодами ZINC000026430653, ZINC000191389930 и ZINC000293423658 (табл. 1) в базе данных ZINC15 демонстрируют значения энергии связывания, сопоставимые (с учетом погрешностей расчета) с величиной –9.5 ± 0.1 ккал/моль, измеренной для комплекса CD4-gp120 методом изотермической титрационной калориметрии [20]. Эти значения также близки к величинам, предсказанным

QuickVina 2 для ингибиторов ВИЧ NBD-11021 и NBD-14010 (табл. 1). Данные о высокоаффинном связывании соединений ZINC000026430653, ZINC000191389930 и ZINC000293423658 с белком gp120 также подтверждаются значениями энергии связывания, рассчитанными для комплексов лиганд/gp120 на основе данных полуэмпирического квантово-химического метода PM7 (табл. 1).

Кроме того, приведенные выше выводы согласуются со значениями энергии связывания анализируемых молекул с белком gp120, рассчитанными из предсказанных значений K_d [17] (табл. 1). Данные таблицы 1 позволяют предполагать, что идентифицированные соединения могут проявлять сильное взаимодействие с гидрофобной полостью Phe-43 CD4-связывающего сайта белка gp120 ВИЧ-1, что согласуется с низкими значениями энергии связывания, сравнимыми с величинами, предсказанными для ингибиторов NBD-11021 и NBD-14010.

Результаты исследования свидетельствуют о том, что разработанная нейронная сеть представляет собой эффективную математическую модель для виртуального скрининга баз данных химических соединений, направленного на поиск малых молекул с высоким сродством к белку gp120 и разработку на их основе новых анти-ВИЧ препаратов широкого спектра действия.

3. Благодарности

Работа поддержана грантом Государственной программы научных исследований «Конвергенция 2020» (проект 3.08) и Белорусским республиканским фондом фундаментальных исследований (проект Х20МС-006).

4. Список литературы

1. Cherkasov A., Muratov E.N., Fourches D., Varnek A., Baskin I.I., Cronin M., Dearden J., Gramatica P., Martin Y.C., Todeschini R. QSAR modeling: where have you been? Where are you going to? *Journal of Medicinal Chemistry*. 2014. V. 201457. P. 4977–5010. doi: 10.1021/jm4004285.
2. Dubey A. Machine learning approaches in drug development of HIV/AIDS. *International Journal of Molecular Biology: Open Access*. 2018. V. 3. № 1. P. 23–25.
3. Li W., Lu L., Li W., Jiang S. Small-molecule HIV-1 entry inhibitors targeting gp120 and gp41: a patent review (2010–2015). *Expert Opinion on Therapeutic Patents*. 2017. V. 27. P. 707–719.
4. Andrianov A.M., Nikolaev G.I., Kornoushenko Y.V., Xu W., Jiang S., Tuzikov A.V. *In silico* identification of novel aromatic compounds as potential HIV-1 entry inhibitors mimicking cellular receptor CD4. *Viruses*. 2019. V. 11. P. E746.
5. Kwong P.D., Wyatt R., Robinson J., Sweet R.W., Sodroski J., Hendrickson W.A. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*. 1998. V. 393. P. 648–659.
6. Curreli F., Kwon Y.D., Zhang H., Scacalossi D., Belov D.S., Tikhonov A.A., Andreev I.A., Altieri A., Kurkin A.V., Kwong P.D., Debnath A.K. Structure-based design of a small molecule CD4-antagonist with broad spectrum anti-HIV-1 activity. *Journal of Medicinal Chemistry*. 2015. V. 58. P. 6909–6927. doi: 10.1021/acs.jmedchem.5b00709.
7. Curreli F., Kwon Y.D., Belov D.S., Ramesh R.R., Kurkin A.V., Altieri A., Kwong P.D., Debnath A.K. Synthesis, antiviral potency, *in vitro* ADMET, and X-ray structure of potent CD4 mimics as entry inhibitors that target the Phe43 cavity of HIV-1 gp120. *Journal of Medicinal Chemistry*. 2017. V. 60. P. 3124–3153. doi: 10.1021/acs.jmedchem.7b00179.
8. Kadurin A., Aliper A., Kazennov A., Mamoshina P., Vanhaelen Q., Khrabrov K., Zhavoronkov A. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget*. 2017. V. 8. P. 10883–10890.
9. Xu B., Wang N., Chen T., Li M. Empirical Evaluation of Rectified Activations in Convolutional Network. *arXiv*: 1505.00853 [cs.LG]. 2015.
10. Durrant J.D., McCammon J.A. AutoClickChem: click chemistry in silico. *PLOS Computational Biology*. 2012. V. 8. № 3. P. e1002397.
11. Sterling T., Irwin, J.J. ZINC 15–ligand discovery for everyone. *Journal of Chemical Information & Modeling*. 2015. V. 55. № 11. P. 2324–2337.
12. Kolb H.C., Finn M.G., Sharpless K.B. Click chemistry: Diverse chemical function from a few good reactions. *Angewandte Chemie International Edition*. 2001. V. 40. № 11. P. 2004–2021.
13. Lipinski C.A., Lombardo F., Dominy B.W., Feeney P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*. 2001. V. 46. № 1–3. P. 3–26.
14. Alhossary A., Handoko S.D., Mu Y., Kwoh C.K. Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics*. 2015. V. 31. № 13. P. 2214–2216.
15. Stewart J.J.P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *Journal of Molecular Modeling*. 2013. V. 19. P. 1–32.
16. Durrant J.D., McCammon J.A. NNScore 2.0: A neural-network receptor–ligand scoring function. *Journal of Chemical Information & Modeling*. 2011. V. 51. № 11. P. 2897–2903.
17. Sharma G., First E.A. Thermodynamic analysis reveals a temperature-dependent change in the catalytic mechanism of bacillus stearothermophilus tyrosyl-tRNA synthetase.

Journal of Biological Chemistry. 2009. V. 284. № 7. P. 4179–4190.
18. Blahut R.E. *Theory and Practice of Error Control Codes*. Addison-Wesley, 1983. 500 p.
19. Tanimoto T.T. *IBM Internal Report 17th*. Armonk, New York: IBM Corp., 1957.
20. Myszka D.G., Sweet R.W., Hensley P., Brigham-Burke M., Kwong P.D., Hendrickson W.A., Wyatt R., Sodroski J., Doyle M.L. *Proc. Natl. Acad. Sci. USA*. 2000. V. 97. P. 9026–9031.