***Доклады Международной конференции "Математическая биология и биоинформатика"**. Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e56. doi: 10.17537/icmbb18.107*

Локальное влияние замены аминокислоты на антигенность вируса гриппа

Форгани М.А.[1], Хачай М.Ю.[1, 2]

[1]*Уральский Федеральный университет*
[2]*Институт математики и механики им. Н.Н. Красовского УрО РАН*

Majid.Forqani@gmail.com

Первым шагом к созданию вакцины против гриппа является измерение степени антигенного сходства между штаммами вируса. Для этого проводится лабораторная процедура под названием анализ ингибирования гемагглютинации. Известно, что между заменами аминокислоты в последовательности белка гемагглютинина и значением логарифма титра, полученным в результате анализа, существует линейное отношение. Это отношение используется для математического моделирования антигенного сходства. В данной работе, с целью увеличения корреляции этого отношения дополнительно используется информация об изменении физико-химического свойства аминокислоты и эффекта соседства, с помощью метода под названием декомпозиция вейвлет-частиц. Это позволит рассматривать точечную мутацию не как замену аминокислоты в одной позиции, а, как локальное изменение физико-химических свойств в регионе ее происхождения. Наши результаты показывают, что, помимо антигенных участков, существуют другие позиции в последовательности белка гемагглютинина, которые могут косвенно влиять на антигенность вируса.

Ключевые слова: гемагглютинин, вейвлет-преобразование, декомпозиция вейвлет-частиц, антигенность, вирус гриппа.

Local impact of amino acid substitution on the antigenicity of the influenza virus

Forghani M.A.[1], Khachay M.Y.[1, 2]

[1]*Ural Federal University*
[2]*Krasovsky Institute of Mathematics and Mechanics*

The first step towards creating a vaccine against influenza is to measure the degree of antigenic similarity between several strains of the virus. To do this, the laboratory procedure called hemagglutination inhibition assay is carried out. It is known that there is a linear relationship between amino acid substitutions in the hemagglutinin protein sequence and the logarithm of the titer obtained as a result of the hemagglutination inhibition assay. This relationship is used for mathematical modelling of antigenic similarity. In this paper, in order to increase the correlation of mentioned relationship, information about the change in the amino acid physicochemical properties and the neighbour effect is added using a method called the wavelet-particle decomposition. This allows to consider a point mutation, not as a substitution of the amino acid in a site, but a local change in the physicochemical properties to the region where it happens. The results show that in addition to antigenic sites, there are other sites in the hemagglutinin sequence that can indirectly affect on the antigenicity of the virus.

Key words: hemagglutinin, wavelet transform, wavelet particle decomposition, antigenicity, influenza virus.

1. Введение

Антигены являются молекулярными структурами на поверхности вирусов, которые распознаются иммунной системой и способны вызывать иммунный ответ. Белок HA (Гемагглютинин от латинского *Hemagglutinin*) является одним из поверхностных антигенов, вызывающим защитные реакции антител и основным белком для создания вакцины. Этот белок имеет высокую аминокислотную замену и по количеству больше выражен на поверхности вируса гриппа.

***Proceedings of the International Conference "Mathematical Biology and Bioinformatics".** Ed. V.D. Lakhno. Vol. 7. Pushchino: IMPB RAS, 2018. Paper No. e56. doi: 10.17537/icmbb18.107*

Антигенность вируса определяется способностью вируса вызывать иммунный ответ организма на антиген. Первым этапом производства эффективной вакцины против гриппа является измерение степени антигенного сходства между штаммами. Для этого обычно делается лабораторная процедура под названием «анализ ингибирования гемагглютинации» (АИГ). Анализ иногда долгосрочен, поскольку на получение антисыворотки может потребоваться до месяца [1]. Чтобы оперативно и быстро действовать против появляющихся новых штаммов, были созданы разные методы для моделирования и прогнозирования антигенного сходства между штаммами вируса гриппа [2]. Моделирование и прогнозирование антигенности вируса дают лучшее понимание эволюции.

Известно, что существует линейное отношение между заменами аминокислот в конкретных позициях (таких как антигенные участки) в последовательности HA и значением титра, полученного из АИГ. Это отношение было использовано в разных исследованиях для создания математической модели антигенных вариантов [3–5].

Одновременное рассмотрение двух факторов позволит изучать данное линейное отношение более подробно. Первый фактор – это изменение свойства аминокислоты, а второй – это эффект соседства.

Часто обнаруживаются химические и физические сходства между аминокислотами, которые участвуют в процессе замены в одной позиции. Чтобы понять все функциональное разнообразие белков, важно оценить физико-химические свойства различных аминокислот [6]. Это можно сделать с помощью оцифровки последовательности белка с использованием физико-химических индексов.

Существует эффект соседства, который представляет собой влияние соседних аминокислот на аминокислоту в конкретной позиции и наоборот [7–9]. Это влияние изучается на соседних аминокислотах на разных уровнях с помощью математических фильтров. Это позволяет рассматривать замену аминокислоты не только как изменение в самой аминокислоте, а как локальное изменение.

Первый фактор представляет собой алфавитную последовательность белка, представленную в виде численного сигнала, удобного для применения методов обработки сигналов. Метод на основе вейвлет-преобразования использует численную последовательность и извлекает информацию о втором факторе для создания признаков, коррелирующих с антигенностью. Созданные признаки можно использовать в качестве переменных модели для прогнозирования антигенного сходства.

Подробности об этих факторах и их вычислении предоставлены далее.

2. Физико-химические свойства аминокислот

Аминокислоты варьируются по размеру, заряду, форме и химическому составу. Важно понимать роль изменения физико-химических свойств по ходу мутаций, т.е. рассматривать мутацию не как замену аминокислоты, а как изменение ее характеристик.

При изучении биологического отношения легко понять, что одного свойства (например, шкалы гидрофобности) недостаточно, чтобы охарактеризовать отношения, а многомерное описание аминокислот лучше подходит для этого [10]. Поэтому для оцифровки алфавитной последовательности белка используются разные физико-химические свойства с целью изучить отношения с различных точек зрения. Выбор численного представления геномного сигнала влияет на то, как его биологические свойства могут быть отражены в числовой области.

Каждое физико-химическое свойство имеет свой аминокислотный индекс. Аминокислотный индекс, это набор из 20 числовых значений, представляющих собой любые физико-химические и биологические свойства аминокислот. Существует база данных аминокислотных индексов под названием Aaindex [11]. Раздел Aaindex1 этой базы в настоящее время содержит более 500 аминокислотных индексов. Каждая запись состоит из номера доступа, краткого описания индекса, справочной информации и числовых значений свойств 20 аминокислот. Аминокислотные индексы далее используются для оцифровки последовательности, чтобы применить такой метод обработки сигналов как вейвлет-преобразование.

3. Эффект соседства

Аминокислоты взаимодействуют друг с другом, в частности с соседними аминокислотами, для образования белковых структур. Большая часть выбора соседа может быть объяснена склонностью аминокислот к образованию различных структур, в особенности вторичных [7–9]. В данной работе, используется преобразование вейвлет-пакета, чтобы подсчитать этот эффект на разных уровнях.

Преобразование вейвлет-пакет (ПВП) представляет собой богатую коллекцию обширной информации с произвольным временно-частотным разрешением и обеспечивает полное дерево декомпозиции как на высокочастотных, так и на низкочастотных компонент [12, 13]. После декомпозиции, входящий сигнал можно восстанавливать на подполосах декомпозиции.

Пусть длина численной последовательности равна N, ее можно показать в виде дискретного сигнала $f(x)$ где $1 \le x \le N$. Пусть j будет уровень декомпозиции ПВП, после декомпозиции получатся $M = 2^j$ подполосы, в каждой из которых можно восстанавливать сигнал. Восстановленные

сигналы вместе можно представить в виде матрицы таким образом, что каждая строка матрицы является восстановленным сигналом в подполосе с соответствующим индексом строки. Размер полученной матрицы равен $M \times N$.

$[f(x), f(2), \ldots, f(N)]$

→ декомпозиции и восстановления ПВП

$$\rightarrow \begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,N} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,N} \\ . & . & & . \\ . & . & & . \\ & & \cdots & \\ . & . & & . \\ f_{M,1} & f_{M,2} & & f_{M,N} \end{bmatrix}$$

где элемент $f_{q,p}$ матрицы, это p-ая точка в q-ом восстановленном сигнале. Столбец этой матрицы выражает отображение точки сигнала с соответствующим индексом столбца в разных подпространствах ПВП. Как следует из вейвлет-теории, сумма всех элементов столбца равна значению в соответствующей точке во входном сигнале. Можно определить оператор Ω, который отображает точки сигнала на вектор с элементами в разных подполосах.

$$\Omega\left(f(x)\right) = \left[f_{1,x}, f_{2,x}, \ldots, f_{M,x}\right]$$

$$f(x) = \sum_{p_1=1}^{M} f_{p_1,x}$$

Чтобы сохранить порядок отображения точки сигнала в виде вектора $\mathbf{\Omega}(f(x))$, и изучать отношение между подполосами, можно рассматривать данный вектор как новый сигнал для декомпозиции ПВП. Применяя процедуру декомпозиции и восстановления ПВП к вектору $\mathbf{\Omega}(f(x))$, получим новую матрицу размера $M \times M$.

$$\mathbf{\Omega}\left(f(x)\right) = \left[f_{1,x}, f_{2,x}, \ldots, f_{M,x}\right]$$

→ декомпозиции и восстановления ПВП

$$\rightarrow \begin{bmatrix} f_{1,1,x} & f_{1,2,x} & \cdots & f_{1,M,x} \\ f_{2,1,x} & f_{2,2,x} & \cdots & f_{2,M,x} \\ . & . & & . \\ . & . & & . \\ & & \cdots & \\ . & . & & . \\ f_{M,1,x} & f_{M,2,x} & & f_{M,M,x} \end{bmatrix}$$

Сумма всех элементов этой матрицы равна $f(x)$ точке исходного сигнала.

$$f(x) = \sum_{p_1=2}^{M} \sum_{p_2=1}^{M} f_{p_2,p_1,x}$$

Каждая точка, полученная в результате применения оператора Ω называется частицей. Этот оператор Ω будем далее называть оператором декомпозиции вейвлет-частиц (ОДВЧ). Используя математическую индукцию, можно доказать следующее утверждение.

Утверждение: допустим f – это дискретный сигнал в $L^2(R)$. Применяя ОДВЧ до уровня q ($q \geq 1$) на f можно приблизительно восстанавливать значение точки $f(x)$, $1 \leq x \leq N$, с помощью формулы:

$$f(x) \cong \sum_{p_1=1}^{M} \sum_{p_2=1}^{M} \cdots \sum_{p_q=1}^{M} f_{p_q,\ldots,p_2,p_1}.$$

В процессе ДВЧ, на первом уровне производится M частиц, на втором M^2 и на уровне q, M^q. Сумма всех частиц каждого уровня равна значению точки, которая и декомпозируется.

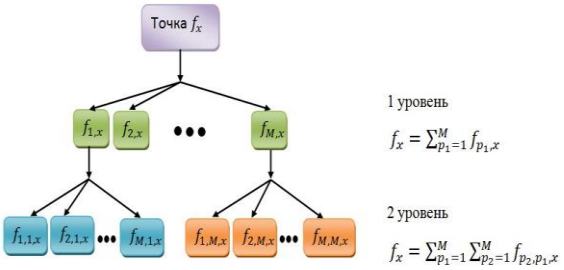

Рис. 1. Декомпозиция значения точки $f(x)$ с помощью ДВЧ.

Можно показать, что каждая частица – это линейная комбинация точек входного сигнала, находящихся в определенном окне в окрестности точки $f(x)$. Коэффициенты комбинации вычисляются через коэффициенты фильтров вейвлета. Длина окна определяется длиной фильтра вейвлета и глубиной декомпозиции ПВП. Влияние соседних точек (т.е. соседних аминокислот) регулируется с помощью коэффициентов комбинации.

4. Извлечение признаков антигенности

Как было указано во введении, существует линейное отношение между заменами аминокислот в последовательности белка НА и значением логарифма титра результатов АИГ.

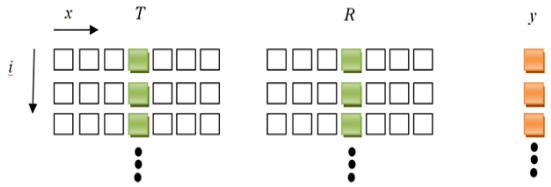

Рис. 2. Демонстрация результатов АИГ.

На рисунке 2 показаны записи результатов АИГ. Каждая запись состоит из идентификатора тестового и референтного штаммов и их значения титра. Каждый идентификатор был заменен на его последовательность НА1 (часть из НА). Индекс i – это индекс записи, а x – это позиция в последовательности. Разница между $T_i(x)$ и $R_i(x)$ коррелирует с $\log(y_i)$, особенно когда x является позицией из антигенных участков. Эта корреляция слабая, но ее можно увеличивать с использованием частиц.

Определяется расстояние между $T_i(x)$ и $R_i(x)$ как векторное расстояние между их частицами. Другими словами, вместо разницы между $T_i(x)$ и

$R_i(x)$ рассматривается расстояние между каждой частицей первого уровня ДВЧ в тестовой и референтной последовательностях. Далее можно рассматривать расстояние между частицами с идентичным индексом в дереве ДВЧ тестовой и референтной последовательности, чтобы найти более коррелирующий признак.

Поскольку число частиц экспоненциально зависит от глубины дерева ДВЧ, поиск самой коррелирующей частицы остается вычислительно сложной задачей. Для ее решения можно использовать эвристические алгоритмы, такие как алгоритм поиска самого крутого восхождения на вершину [14]. Этот алгоритм на каждом уровне выбирает самую лучшую частицу, которая дает корреляцию больше, чем свой отец, и поиск продолжается в его ветви (на следующем уровне). То есть на каждом уровне рассматривается только M частиц (см. рис. 3).

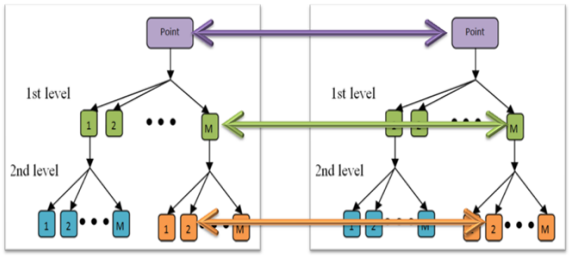

Рис. 3. Поиск наилучшего признака. Каждая двухсторонняя стрелка – это векторное расстояние между частицами тестового и референтного штаммов.

5. Результаты

Применение метода ДВЧ на последовательности HA1 штаммов базы данных [15] с разными свойствами аминокислот из базы данных Aaindex1, дало результаты, показанные на гистограмме рисунка 4.

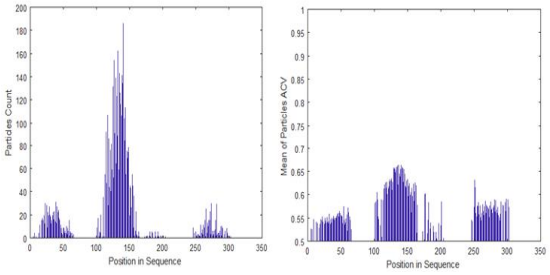

Рис. 4. Слева показано представление общего количества полученных признаков для каждой позиции последовательности белка HA1. Справа изображено среднее значение абсолютной величины корреляции всех полученных признаков каждой позиции в белке.

Отображение найденных коррелирующих регионов на белке гемагглютинина предоставлено на рисунке 5. Следует обратить внимание на то, что чем насыщеннее цвет на изображении, тем аминокислота с большим количеством свойств коррелирует с логарифмом титра. Сравнение полученных результатов с другими исследованиями показывает совпадение найденных регионов с участками связывания рецептора, антигенными участками [1, 16, 17] и потенциальными местами [18–20] в связи с эволюцией и антигенностью вируса гриппа подтипа H1N1.

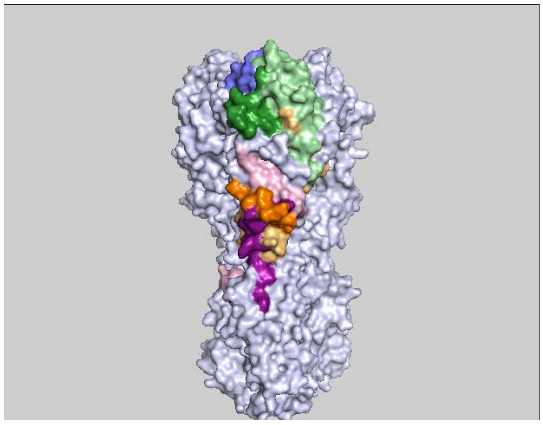

Рис. 5. Визуализация найденных регионов на кристалле белка HA. Изображение создано в программе PyMOL на кристалле 3LZG из [17].

6. Заключение

Решение многих биологических и генетических задач базируется на изучении и моделировании фенотипа на основе генетических последовательностей. Одна из важнейших задач в моделировании эволюции – это прогнозирование антигенного сходства между штаммами вируса, в частности вируса гриппа. Первый этап моделирования – это определение отношения переменных модели и их влияния на целевую функцию.

В данной работе было исследовано влияние замены аминокислоты на антигенность вируса гриппа подтипа H1N1. Результаты показывают, что добавление информации о характеристиках аминокислоты, таких как физико-химические свойства, и информации об эффекте соседства на замену аминокислоты, демонстрирует ее влияние на антигенность более четко. Также совпадение найденных потенциальных мест с результатами других исследований говорит о том, что для изучения антигенности можно рассматривать не только потенциальные иммунодоминантные места, но и позиции, которые косвенно влияют на эти участки. Примером являются аминокислоты, находящиеся по соседству с антигенными участками. Признаки, созданные с помощью метода ДВЧ, также можно использовать как переменные модели для прогнозирования антигенного сходства вируса гриппа.

7. Список литературы

1. Wu A., Peng Y., Du X., Shu Y., Jiang T. Correlation of influenza virus excess mortality with antigenic variation: application to rapid

estimation of influenza mortality burden. *PLoS computational biology*. 2010. V. 6. № 8. P. e1000882.
2. Klingen T.R., Reimering S., Guzmán C.A., McHardy A. C. In Silico Vaccine Strain Prediction for Human Influenza Viruses. *Trends in microbiology*. 2017.
3. Reeve R., Blignaut B., Esterhuysen J.J., Opperman P., Matthews L., Fry E., O'Neill H.G. Sequence-based prediction for vaccine strain selection and identification of antigenic variability in foot-and-mouth disease virus. *PLoS computational biology*. 2010. V. 6. № 12. P. e1001027.
4. Harvey W.T., Benton D.J., Gregory V., Hall J.P., Daniels R.S., Bedford T., et al. Identification of low-and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza A (H1N1) viruses. *PLoS pathogens.* 2016. V. 12. № 4. P. e1005526.
5. Harvey W.T. *Quantifying the genetic basis of antigenic variation among human influenza A viruses*: doct. diss. University of Glasgow, 2016.
6. Cozzone A.J. Proteins: Fundamental chemical properties. *ELS*. 2010.
7. Colubri A., Jha A.K., Shen M.Y., Sali A., Berry R.S., Sosnick T.R., Freed K.F. Minimalist representations and the importance of nearest neighbor effects in protein folding simulations. *Journal of molecular biology*. 2006. V. 363. № 4. P. 835–857.
8. Ghadimi M., Khalifeh K., Heshmati E. Neighbor effect and local conformation in protein structures. *Amino acids*. 2017. V. 49. № 9. P. 1641–1646.
9. Toal S., Schweitzer-Stenner R., Rybka K., Schwalbe H. How do Nearest-Neighbor Interactions Effect the Conformational Distributions in Peptides? *Biophysical Journal*. 2013. V. 104. № 2. P. 55a.
10. Sjöström M., Wold S. A multivariate study of the relationship between the genetic code and the physical-chemical properties of amino acids. *Journal of molecular evolution.* 1985. V. 22. № 3. P. 272–277.
11. Kawashima S., Pokarowski P., Pokarowska M., Kolinski A., Katayama T., Kanehisa M. AA index: amino acid index database, progress report 2008. *Nucleic acids research*. 2007. V. 36. P. D202–D205.
12. Li S., Zhou Y., Kou Z., Yan B. A Wavelet Packet Based Approach for the Research of the Avian Influenza virus cross-species infection. In: *2010 First International Conference on Networking and Distributed Computing*. 2010. P. 301–304.
13. Graps A. An introduction to wavelets. *IEEE computational science and engineering*. 1995. V. 2. № 2. P. 50–61.
14. Harman M. The current state and future of search based software engineering. In: *Proceeding FOSE '07 2007 Future of Software Engineering*. Washington: IEEE Computer Society, 2007. P. 342–357.
15. Gregory V., Harvey W.T., Daniels R.S., Reeve R., Whittaker L., Halai C., Douglas A., Gonsalves R., Skehel J.J., Hay A.J., McCauley J.W. *Human former seasonal influenza A(H1N1) haemagglutination inhibition data 1977–2009 from the who collaborating centre for reference and research on influenza*. London: Technical report, University of Glasgow, 2016.
16. Brownlee G.G., Fodor E. The predicted antigenicity of the haemagglutinin of the 1918 Spanish influenza pandemic suggests an avian origin. *Philosophical Transactions of the Royal Society of London. Series B.* 2001. V. 356. № 1416. P. 1871.
17. Xu R., Ekiert D.C., Krause J.C., Hai R., Crowe J.E., Wilson I.A. Pdb id: 3lzg structural basis of preexisting immunity to the 2009 h1n1 pandemic influenza virus. *Science*. 2010. P. 1186430.
18. Zhang W., Qi J., Shi Y., Li Q., Gao F., Sun Y., Lu X., Lu Q., Vavricka Ch.J., Liu D., et al. Crystal structure of the swine-origin a (h1n1)-2009 influenza a virus hemagglutinin (ha) reveals similar antigenicity to that of the 1918 pandemic virus. *Protein & cell*. 2010. V. 1. № 5. P. 459–467.
19. Zhao R., Cui S., Guo L., Wu C., Gonzalez R., Paranhos-Baccal`a Gl., Vernet G., Wang J., Hung T. Identification of a highly conserved h1 subtype-specific epitope with diagnostic potential in the hemagglutinin protein of influenza a virus. *PloS one*. 2011. V. 6. № 8. P. 23374.
20. Tsibane T., Ekiert D.C., Krause J.C., Martinez O., Crowe Jr.J.E., Wilson I.A., Basler C.F. Influenza human monoclonal antibody 1f1 interacts with three major antigenic sites and residues mediating human receptor specificity in h1n1 viruses. *PLoS pathogens.* 2012. V. 8. № 12. P. e1003067.