Математическое моделирование антигенного сходства штаммов вируса гриппа H1N1 с помощью преобразования вейвлет-пакета

М.А. Форгани[1] М.Ю. Хачай[1,2]

[1]Институт естественных наук и математики
Уральский Федеральный Университет

[2]Институт математики и механики
им. Н.Н.Красовского УрО РАН

X конференция
2018

Содержание доклада

- Измерение антигенного сходства
- Математическое моделирование антигенных вариантов
- Извлечение признаков
- Постановка задачи
- Роль физико-химических свойств аминокислот
- Эффект соседства
- Метод декомпозиции вейвлет-частиц
- Сложности вычислительного алгоритма
- Вычислительные эксперименты
- Результаты и заключение

Измерение антигенного сходства

- Первым шагом к созданию вакцины против гриппа является измерение степени антигенного сходства между штаммами вируса.

- Анализ ингибирования гемагглютинации – это классическая лабораторная процедура для измерения степени сходства между штаммами вируса гриппа.

Image source: https://www.cdc.gov/flu/professionals/laboratory/antigenic.htm

Математическое моделирование антигенных вариантов

Тип модели:

- Модели на основе филогенетического анализа
- Статистические модели
- Модели на основе эпидемиологических данных

Входящие данные:

- Геномные последовательности
- Информация о структуре белка
- Физико-химические свойства аминокислот
- Эпидемиологические данные

Извлечение признаков

- Большинство методов рассматривают корреляцию между заменами аминокислоты в конкретных участках и значением титра (полученным в результате лабораторных анализов).
- Известно, что существует **линейное отношение** между заменами аминокислоты антигенных участках и логарифмом значения титра.
- Это линейное отношение было использовано в **линейной модели смешанных эффектов**, чтобы прогнозировать антигенное сходство для вирусов ящура и гриппа.
- Два фактора могут улучшить качество моделирования:
  - Учет изменения **физико-химических свойств** по ходу эволюции
  - Учет **эффекта соседства**

Извлечение признаков

- Каждая запись анализа ингибирования гемагглютинации состоит из пары идентификаторов штаммов вируса (называем *тестовый* и *референтный* штаммы) плюс результат их измерения, т.е. значение титра.

	A	B	C	D
1	virus	reference	dateOfTest	titre
2	A/BAYERN/7/95	A/BAYERN/7/95	9/5/2001	2560
3	A/JOHANNESBURG/82/96	A/BAYERN/7/95	9/5/2001	2560
4	A/BEIJING/262/95	A/BAYERN/7/95	9/5/2001	40
5	A/NEWCALEDONIA/20/99	A/BAYERN/7/95	9/5/2001	40
6	A/HONGKONG/1252/2000	A/BAYERN/7/95	9/5/2001	<
7	A/MADAGASCAR/57794/20(	A/BAYERN/7/95	9/5/2001	40
8	A/BAYERN/7/95	A/JOHANNESBURG/82/96	9/5/2001	5120

Постановка задачи

- На самом деле, каждая i-я запись это упорядоченная тройка (T_i, R_i, y_i), где T_i и R_i, $i = 1, \ldots, L$ это последовательности тестового и референтного штаммов, соответственно, y_i значение наблюдаемого титра, и L это количество записей.

- Можно сравнивать штаммы в определенной позиции x в последовательности с помощью $T_i(x)$ и $R_i(x)$.

- Для каждой позиции x и векторов $T = [T_1, \ldots, T_L]$, $R = [R_1, \ldots, R_L]$, $y = [y_1, \ldots, y_L]$, необходимо определить такое отображение $Z = \zeta_x \colon \mathbb{R}^2 \to \mathbb{R}$, чтобы величины $\zeta(T(x), R(x))$ и $\log_2(y)$ обладали высоким по абсолютной величине коэффициентом корреляции:

$$AP(\zeta, x \mid T, R, y) = \left| \frac{Cov(\zeta(T(x), R(x)), \log_2(y))}{\sqrt{Var(\zeta(T(x), R(x)))}\sqrt{Var(\log_2(y))}} \right|.$$

Постановка задачи

- Для заданной позиции x в последовательности, метод декомпозиции вейвлет-частиц предоставляет множество функций $\mathcal{F}_x$, обладающих достаточно высоким коэффициентом корреляции с логарифмом y. Необходимо найти $\zeta^* \in \mathcal{F}_x$ (если она существует), такую что:

$$AP(\zeta^*, x \mid T, R, y) = \max\{AP(\zeta, x \mid T, R, y) \colon \zeta \in \mathcal{F}_x\}.$$

- Для фиксированного численного отображения последовательности белка h (на основе физико-химических свойств) определяем множество $\mathcal{F}_h^*$, которое содержит ζ^* всех позиций x в последовательности белка.
- Сокращается пространство признаков $\bigcup_{h=1}^{H} \mathcal{F}_h^*$ с помощью методов отбора признаков и новое множество обозначается $\mathcal{F}^*$.
- Признаки, находящиеся в множестве $\mathcal{F}^*$ используются в фиксированной части линейной модели смешанных эффектов.

Постановка задачи

- Для заданной позиции x в последовательности, метод декомпозиции вейвлет-частиц предоставляет множество функций $\mathcal{F}_x$, обладающих достаточно высоким коэффициентом корреляции с логарифмом y. Необходимо найти $\zeta^* \in \mathcal{F}_x$, такую что:

$$\log_2(y_{t,r}) = k_0 + \left[\sum_j k_j \alpha_j(t,r) + \sum_j m_i \delta_i(t,r) + f_t + f_r\right] + \epsilon_D + \epsilon_{res}$$

- Данное исследование сфокусированно на улучшении фиксированной части модели (показана в квадратных скобках).
- Используя признаки ζ^* из множества $\mathcal{F}^*$ можно создать новую модель:

$$\log_2(y_{t,r}) = k_0 + \sum_l \beta_l \zeta_l^* + \epsilon_D + \epsilon_{res}$$

Физико-химическое свойство аминокислоты

- При изучении биологического отношения легко понять, что одного свойства (например, шкалы гидрофобности) недостаточно, чтобы охарактеризовать отношения, а многомерное описание аминокислот лучше подходит для этого.
- Для оцифровки алфавитной последовательности белка используются разные физико-химические свойства с целью изучить отношения с различных точек зрения.
- Аминокислотный индекс, это набор из 20 числовых значений, представляющих собой любые физико-химические и биологические свойства аминокислот.

Физико-химическое свойство аминокислоты

- Существует база данных аминокислотных индексов под названием Aaindex.
- В настоящее время Aaindex содержит более 500 аминокислотных индексов.

```
H ARGP820101
D Hydrophobicity index (Argos et al., 1982)
R PMID:7151796
A Argos, P., Rao, J.K.M. and Hargrave, P.A.
T Structural prediction of membrane-bound proteins
J Eur. J. Biochem. 128, 565-575 (1982)
C JOND750101    1.000  SIMZ760101    0.967  GOLD730101    0.936
  TAKK010101    0.906  MEEJ810101    0.891  ROSM880104    0.872
  CIDH920105    0.867  LEVM760106    0.865  CIDH920102    0.862
  MEEJ800102    0.855  MEEJ810102    0.853  ZHOH040101    0.841
  CIDH920103    0.827  PLIV810101    0.820  CIDH920104    0.819
  LEVM760107    0.806  NOZY710101    0.800  GUYH850103   -0.808
  PARJ860101   -0.835  WOLS870101   -0.838  BULH740101   -0.854
I    A/L     R/K     N/M     D/F     C/P     Q/S     E/T     G/W     H/Y     I/V
    0.61    0.60    0.06    0.46    1.07      0.    0.47    0.07    0.61    2.22
    1.53    1.15    1.18    2.02    1.95    0.05    0.05    2.65    1.88    1.32
//
```

Эффект соседства

- Аминокислоты взаимодействуют друг с другом, в частности с соседними аминокислотами, для образования белковых структур.
- Влияние соседних аминокислот на целевую аминокислоту упоминается как эффект ближайшего соседа или просто эффект соседства.
- Эффект соседства позволяет рассматривать замену аминокислоты не только как изменение в самой аминокислоте, а как локальное изменение.
- Мы рассмотрим этот эффект через линейную комбинацию аминокислот, расположенных в окне вокруг целевого положения.

Декомпозиция вейвлет-частиц

- Преобразование вейвлет-пакет обеспечивает полное дерево декомпозиции как на высокочастотных, так и на низкочастотных компонент.
- Глубина декомпозиции ограничена сверху $log_2 N$, где N это длина исходного сигнала.
- Существуют два вида информации в каждой подполосе: ее коэффициенты и восстановленный сигнал в подполосе из коэффициентов.

Декомпозиция вейвлет-частиц

- Допустим f(x) это дискретный сигнал с длиной N точек.
- Пусть $j \in \mathbb{Z}^+$ это глубина декомпозиции, ограниченная сверху $log_2(N)$. Существуют $M = 2^j$ подполос на последнем уровне декомпозиции.
- Восстановленные сигналы в подполосах последнего уровня можно переформировать в виде следующей матрицы (матрица восстановления):

$$[f(1), f(2), \ldots, f(N)] \quad \begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,N} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,N} \\ \cdot & & \cdots & \cdot \\ \cdot & & \cdots & \cdot \\ \cdot & & \cdots & \cdot \\ f_{M,1} & f_{M,2} & \cdots & f_{M,N} \end{bmatrix}_{M \times N}$$

Декомпозиция вейвлет-частиц

- Каждая строка матрицы восстановления это восстановленный сигнал подполосы с соответствующим индексом строки.
- Каждый столбец этой матрицы является результатом декомопозиции конкретной точки исходного сигнала по отношению к различным подполосам.
- Процессы декомпозиции и восстановления обеспечивают представление каждой записи сигнала $f(x)$ с $M = 2^j$ точками в различных подполосах. Обозначим это представление $\Omega\,\Omega(f(x)) = [f_{1,x}, f_{2,x}, \ldots,$

Декомпозиция вейвлет-частиц

- Как известно, на основе теории вейвлета значение $f(x)$ можно приблизительно восстанавливать в следующем виде:

$$f(x) \approx \sum([f_{1,x}, f_{2,x}, \ldots, f_{M,x}])$$
$$= \sum(\Omega(f(x)) = \sum_{p=1}^{M} f_{p,x}, \quad (1 \leq x \leq N),$$

где вектор $[f_{1,x}, f_{2,x}, \ldots, f_{M,x}]$ это транспонированный вектор x-ого столбца восстановленной матрицы.

- Порядок представления в Ω (с высокочастотной до низкочастотной подполосы или наоборот) содержит важную информацию, которую предполагается рассмотреть и извлечь.

Декомпозиция вейвлет-частиц

- Рассматривается представление $\Omega(f(x))$ как новый сигнал, к которому снова применяется преобразование вейвлет-пакет (также с глубиной j).

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

- Результат можно переформировать как новую матрицу восстановления:

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

- Аналогично предыдущему процессу можно писать:

$$\Omega(f_{p_1,x}) = [f_{1,p_1,x}, f_{2,p_1,x}, \ldots, f_{M,p_1,x}]$$

для $1 \leq x \leq N$ и $1 \leq p_1 \leq M$.

Декомпозиция вейвлет-частиц

- На основе теории вейвлета для

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

можно писать следующую аппроксимацию:

$$f_{p_1,x} \approx \sum (\Omega(f_{p_1,x})) = \sum_{p_1=1}^{M} f_{p_2,p_1,x},$$

для $1 \leq x \leq N$ и $1 \leq p_1 \leq M$.

- Поскольку

$$f(x) \approx \sum (\Omega(f(x)) = \sum_{p=1}^{M} f_{p,x}, \quad (1 \leq x \leq N),$$

значение точки $f(x)$ исходного сигнала, $1 \leq x \leq N$, можно приблизительно восстанавливать с помощью:

$$f(x) \approx \sum_{p=1}^{M} f_{p,x} = \sum_{p_1=1}^{M} \sum_{p_2=1}^{M} f_{p_2,p_1,x}, \quad (1 \leq x \leq N)$$

Декомпозиция вейвлет-частиц

- Эта новая декомпозиция точки называется **декомпозиция вейвлет-частиц**.
- А **частица** это любое значение, полученное в результате декомпозиции вейвлет-частиц.
- Частица каждой точки $f(x)$ это линейная комбинация точек, находящихся в окне соседства целевой позиции x.

Декомпозиция вейвлет-частиц

Утверждение: *Пусть f будет дискретным сигналом с длиной N точек из $L^2(R)$. Для любого уровня q декомпозиции вейвлет-частиц и для любой позиции x, $1 \leq x \leq N$, справедливо следующее уравнение:*

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

Доказательство можно получить с помощью математической индукции на q.

Декомпозиция вейвлет-частиц

- Допустим S_1 и S_2 будут два сигнала (последовательности) с длиной N и , где $1 \leq x \leq N$, это позиция в сигнале. Обозначим следующую величину

$$D(S_1(x), S_2(x)) = \|\Omega(S_1(x)) - \Omega(S_2(x))\|$$

векторным расстоянием между точками $S_1(x)$ и $S_2(x)$.

- Аналогично, такое векторное расстояние можно определить между частицами из $S_1(x)$ и $S_2(x)$ с одинаковым индексом следующим образом:

$$D'(S_1(x), S_2(x)) = \|\Omega(p(S_1(x)) - \Omega(p(S_2(x))\|$$

где $p(f(x))$ это частицы декомпозиции $f(x)$ конкретного уровня.

Декомпозиция вейвлет-частиц

Демонстрация метода ДЧ на две последовательности белка HA. Для этого были взяты 2 штамма вируса гриппа A/NORWAY/1684/2007 и A/NORWAY/1651/2007. Разница между последовательностями находится в позиции 256, где в A/NORWAY/1684/2007 аминокислота треонин, а в A/NORWAY/1651/2007 аминокислота аланин.

(a) (b) (c)

Декомпозиция вейвлет-частиц

В целом, чтобы найти лучший признак (расстояние), необходимо идти и искать по всему дереву (до заданной глубины), предоставляемому декомпозицией вейвлет-частиц. Поскольку сложность алгоритма экспоненциальна, то поиск лучшей частицы (или линейной комбинации) занимает много времени.

уровень	сложность	#частиц
$K = 1$	$N \log N$	NM
$K = 2$	$N(M \log M)$	NM^2
$K = 3$	$NM(MlogM)$	NM^3
$K = 4$	$NM^2(MlogM)$	NM^4
$\dots$	$\dots$	$\dots$

Декомпозиция вейвлет-частиц

- Существуют различные эвристические алгоритмы, способные эффективно обеспечить решения, близкие к глобальным оптимальным решениям.

- Вычислительные эксперименты показали, что алгоритм поиска самого крутого восхождения на вершину хорошо подходит для поиска наилучшего признака.

Результаты

Образец улучшения корреляции между векторным расстоянием D' и $log_2(titr)$ в процессе поиска.

	1	2	3
1	[135,4,8,1]	[0.5211,0.6279,0.6329,0.6432]	*3753x1 double*
2	[137,4,5,3]	[0.5776,0.5874,0.6350,0.6589]	*3753x1 double*
3	138	0.5804	*3753x1 double*
4	139	0.6008	*3753x1 double*
5	140	0.5808	*3753x1 double*
6	[141,7]	[0.6499,0.6750]	*3753x1 double*
7	[142,4,6]	[0.5748,0.6381,0.6770]	*3753x1 double*
8	[143,1]	[0.5451,0.6267]	*3753x1 double*
9	[145,3,5,7]	[0.5558,0.5572,0.5883,0.6148]	*3753x1 double*

Результаты

Было обнаружено три области, которые сильно коррелировали с логарифмом значения титра. Кроме того, в центре последовательности существует слабо коррелированная область

Результаты

Схема процедуры моделирования антигенных вариантов выглядит так:

Результаты

Средняя ошибка модели Харви и др. составляет около 0.9. Аналогично модель данной работы на основе декомпозиции вейвлет частиц была применена для прогнозирования и средняя ошибка получилась равной 0.61.

год	ошибка
2002	0.56
2003	0.60
2004	0.56
2005	0.46
2006	0.50
2007	0.65
2008	0.69
2009	0.67

Заключение

- Представление эффекта соседства как линейной комбинации аминокислот позволяет уменьшить ошибку математического моделирования антигенных вариантов.
- Результаты численных экспериментов показывают, что эффективность декомпозиции вейвлет-частиц чувствительна к порядку соседей.
- Согласно вычислительным результатам алгоритм поиска самого крутого восхождения на вершину способен найти оптимальную функцию с наивысшей корреляцией за меньшее время в сравнении с рассмотренными алгоритмами.
- Следует обратить внимание, что на полученные частицы могут влиять несколько факторов, таких как базис вейвлета и глубина декомпозиции вейвлет-пакетов.
- Обнаруженные области включают результаты других исследований (например, антигенные участки и участок связывания рецептора) по изучению антигенности.

Спасибо за внимание!