

Определение вида картины замен в нуклеотидных и аминокислотных последовательностях. сообщение 1.

Теоретические аспекты

Белорусский государственный медицинский университет

В статье рассмотрены теоретические аспекты определения вида картины замен в нуклеотидных и аминокислотных последовательностях. Описаны методы вычисления композиционной дистанции, pdf-фактора, индекса несоответствия, метод Монте-Карло для проверки гипотезы о гомогенности. Охарактеризованы альтернативные методы (χ^2 -тест и тест Ржетского-Нея) и указаны возможные области применения теста индекса несоответствия. Ключевые слова: композиционная дистанция, картина замен, гомогенность, гетерогенность, индекс несоответствия, метод Монте-Карло.

Общим предположением при проведении сравнительного анализа нуклеотидных и аминокислотных последовательностей является то, что они эволюционировали с неизменной картиной нуклеотидных замен (гипотеза о гомогенности эволюционного процесса). Однако в ряде случаев картина замен в ходе эволюции может изменяться, то есть становиться гетерогенной. Отсутствие учета гетерогенной картины замен препятствует получению корректных филогенетических выводов и проверке эволюционных гипотез [8].

Для оценки вида картины замен в 2001 г. С. Кумар и С. Гадагкар [6] предложили индекс несоответствия (disparity index, ID), который позволяет определить различия картин замен для пар последовательностей. На основе этого индекса ими разработан метод Монте-Карло, предназначенный для проверки гипотезы о гомогенности. Рассмотрим основные параметры, используемые при определении вида картины замен.

Композиционная дистанция. Композиционная дистанция (дистанция состава) – это мера сходства состава сравниваемых нуклеотидных или аминокислотных последовательностей. Рассмотрим методику вычисления композиционной дистанции.

Пусть X и Y – это 2 последовательности ДНК длиной L каждая. Пусть x_i – это количество нуклеотида типа i (i=A, T, Ц или Г) в последовательности X, а y_i – в последовательности Y. Тогда композиционная дистанция между этими двумя последовательностями определяется как:

$$D_C = \frac{1}{2} \sum_i (x_i - y_i)^2 \quad (1)$$

где i = A, T, Ц или Г. Для получения расчетного значения DC представим последовательности X и Y как (a1a2a3...a4) и (b1b2b3...b4). Для данного нуклеотида типа i, расположенного в сайте k, определим:

+ 1 если $a_k = i$, а $b_k \neq i$

$\neq k_i = -1$ если $a_k \neq i$, а $b_k = i$

0 другие варианты

(2)

$$\delta_{i}^k = \begin{cases} +1 & \text{если } a_k = i, \text{ а } b_k \neq i \\ -1 & \text{если } a_k \neq i, \text{ а } b_k = i \\ 0 & \text{другие варианты} \end{cases}$$

Используя уравнение 2, формулу (1) можно записать как:

$$D_C = \frac{1}{2} \sum_i (\sum_{k=1}^L \delta_i^k)^2 \quad (3)$$

Расчетное значение композиционной дистанции определяется как:

$$E(D_C) = \frac{1}{2} E(\sum_i (\sum_{k=1}^L \delta_i^k)^2) \quad (4)$$

Учитывая независимость сайтов, получаем

$$E(D_C) = \frac{1}{2} E(\sum_i \sum_k (\delta_i^k)^2) + \frac{1}{2} E(\sum_i \sum_k \sum_{k' \neq k} \delta_i^k \delta_i^{k'}) \quad (5)$$

Первое слагаемое – это расчетное количество нуклеотидных различий между двумя по-следовательностями (N_d), которое определяется степенью их дивергенции, картиной эволюци-онных замен и степенью гетерогенности сайтов. То есть:

$$\frac{1}{2} E(\sum_i \sum_k (\delta_i^k)^2) = E(N_d) \quad (6)$$

Поскольку суммирование происходит для независимых сайтов, то второе слагаемое в формуле (5) можно записать следующим образом:

$$\frac{1}{2} E(\sum_i \sum_k \sum_{k' \neq k} \delta_i^k \delta_i^{k'}) = \sum_i (E[\sum_k \delta_i^k] E[\sum_{k' \neq k} \delta_i^{k'}]) \quad (7)$$

Если основной процесс замен является гомогенным, то для данной пары нуклеотидов

$$\frac{1}{2} E(\sum_i \sum_k \sum_{k' \neq k} \delta_i^k \delta_i^{k'}) = \sum_i (E[\sum_k \delta_i^k] E[\sum_{k' \neq k} \delta_i^{k'}])$$

– это число сайтов с нуклеотидом i в последова-тельности X и нуклеотидом j в последовательности Y . Таким образом,

$$E(\sum_k \delta_i^k) = E(n_{i.}) - E(n_{.i}) = 0 \quad (8)$$

Поэтому,

$$E(\sum_i \sum_k \sum_{k' \neq k} \delta_i^k \delta_i^{k'}) = 0_{(9)}$$

Подставив формулы (6) и (9) в формулу (5), получаем

$$E(D_C) = E(N_d)_{(10)}$$

где N_d – это количество сайтов в последовательностях X и Y с различными нуклеотидами.

Формула 10 показывает, что расчетное число различий между двумя последовательно-стоящими является половиной суммы всех частот различий соответствующих выровненных оснований (или аминокислот) в сравниваемых последовательностях. В 1977 г. А. Корниш-Боуден [4] первым предложил формулу (1). Однако доказательство формулы (10), представленное в его работе, подразумевало гомогенность эволюционного процесса и независимость значений x_i и y_i . Понятно, что его второе предположение является неверным, потому что эти значения коррелируют в связи с наличием общего предшественника последовательностей X и Y . Рассмотренное выше доказательство С. Кумара и С. Гадагкар не требует этого предположения и основывается на независимости сложности модели замен нуклеотидов (или аминокислот), примененной к наблюдаемой картине, и степени гетерогенности сайтов. Следует отметить, что этот факт был подтвержден компьютерным моделированием.

Однако полученные значения композиционных дистанций оказываются достаточно большими, что делает неудобным их интерпретацию и сравнение между собой. Поэтому эти значения делят на количество сравниваемых нуклеотидных или аминокислотных сайтов, получая композиционную дистанцию в расчете на сайт [7].

Фактор pdf. Когда две сравниваемые последовательности имеют различные картины замен, то полученная с использованием формулы (1) композиционная дистанция будет больше, чем полученная в случае гомогенности. Это связано с тем, что наблюдаемые различия в частоте сходных сравниваемых параметров двух последовательностей ($x_i - y_i$) будут больше. Для подтверждения этого было проведено компьютерное моделирование для аминокислотных последовательностей. Вероятность замены одного аминокислотного остатка на другой в нем была принята одинаковой для всех остатков в эволюции последовательности в обеих линиях (гомогенность). В случае гетерогенности принята намного большая вероятность замены на данный остаток в выбранной линии для выявления больших отклонений в картинах замен (фактор девиации картины, pdf) с равными остальными вероятностями замен.

Фактор pdf-фактор, показывающий отличие вероятности замены на данную аминокислоту (или нуклеотид) при гетерогенности от ожидаемой в случае гомогенности.

Для s возможных состояний ($s=4$ для нуклеотидов и $s=20$ для аминокислот) вероятность замены на данную составляет $1/(s - 1)$, если все замены равновероятны. Если Pdf равен значению f , то это означает, что вероятность

данной замены равна $f/(s - 1)$; $f=1$ соответствует гомо-генному процессу. Вероятности любой другой замены равны и определяются как $(1 - f/[s - 1])/(s - 2)$. Более высокое значение pdf указывает на большую гетерогенность в картинах замен.

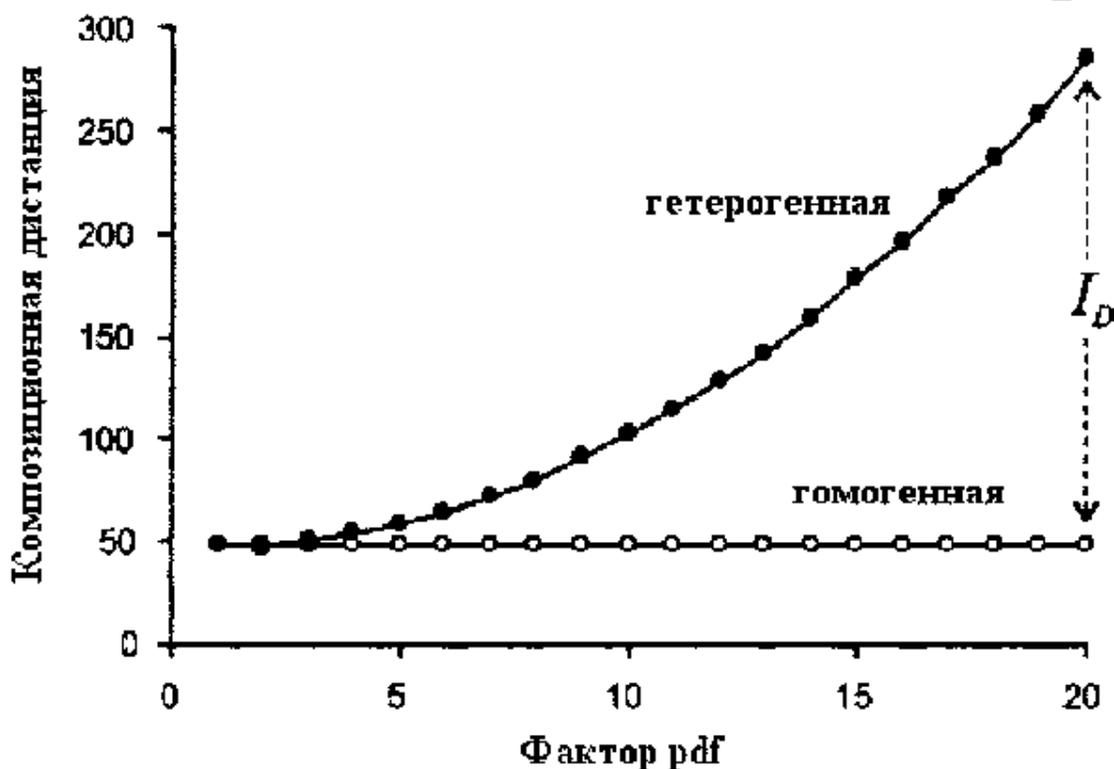


Рис. 1. Зависимость между композиционной дистанцией и pdf-фактором для гомо-и гетероген-ной картин замен, полученная С. Кумар и С. Гадагкар [6]. Индекс несоответствия. Очевидно, что значение DC будет выше при гетерогенном эволюционном процессе (рис. 1). Наблюдаемое между композиционной дистанцией и значением pdf несоответствие, увеличивающееся с ростом гетерогенности, называется индексом несоответствия (index disparity, ID). Значение ID увеличивается при увеличении количества замен и постоянном pdf, а также при увеличении pdf и постоянном количестве замен. При эмпирическом анализе данных ID для данной пары последовательностей вычисляется по формуле:

$$I_D = \frac{1}{2} \sum (x_i - y_i)^2 - N_{d(11)}$$

где x_i и y_i — это частоты нуклеотида (или аминокислоты) типа i в последовательностях X и Y, соответственно, а значение N_d используется для оценки ожидаемой в случае гомогенности DC. Когда предположение о гомогенности подтверждается, то $E(I_D) = 0$, потому что расчетное значение $\sum (x_i - y_i)^2$ равно значению N_d (формула 10). Сходно с композиционной дистанцией индекс несоответствия удобнее вычислять в расчете на сайт.

Метод Монте-Карло для проверки гипотезы о гомогенности. Для проверки гипотезы о гомогенности необходимо вычислить вероятность того, что наблюдаемое значение композиционной дистанции (DCO) больше, чем ожидаемое согласно нулевой гипотезе о гомогенности, то есть $ID > 0$. Поскольку фактическое распределение DC при гомогенности для данных частот оснований и число различий не известно априорно, для его получения следует использовать метод Монте-Карло. В этом методе первоначально используется случайная последовательность длиной L , ожидаемые частоты принимаются равными средним частотам оснований, вычисленным для данной пары последовательностей. Далее производятся случайные замены в двух образовавшихся последовательностях до тех пор, пока количество различий между ними не станет равным Nd для сравниваемой пары последовательностей. Это необходимо для получения DC согласно гипотезе о гомогенности для наблюдаемых данных, учитывая средние частоты оснований для исходной пары последовательностей. Для произведения замен произвольно выбирается участок одной из двух образовавшихся последовательностей. Затем производится произвольная замена нуклеотида в данном сайте (независимо от его исходного вида) на другой на основе полученных ранее средних наблюдаемых частот. Получившиеся последовательности будут иметь те же частоты оснований, поскольку замены происходят со сходным эволюционным процессом в обеих линиях. Эта схема выбрана потому, что нет никакой априорной информации относительно нулевой картины замен и различий скоростей эволюции среди сайтов или линий. Для двух последовательностей, образованных в данном повторе b вычисляем DC, b . Затем этот процесс повторяется необходимое число раз (обычно 1000) и вычисляется доля повторов, в которых DCO выше, чем DC, b ($ID > 0$). Если эта доля больше 95%, то нулевая гипотеза отклоняется на 5%-ом уровне.

Возможности ID-теста. Для оценки эффективности метода Монте-Карло в обнаружении различий эволюционных картин было проведено компьютерное моделирование при различных биологических условиях. Ошибка ID-теста на 5%-ом уровне при неравных скоростях эволюции линий и для аминокислотных последовательностей составляет приблизительно 5%. Возможность ошибки $> 5\%$, особенно при использовании модели Джукса-Кантора [5], обуславливает большую приемлемость 1%-ого уровня значимости.

Была проанализирована эффективность ID-теста в отклонении ложной нулевой гипотезы, когда сравниваемые последовательности эволюционировали с разными эволюционными процессами. Статистические возможности ID-теста в отклонении нулевой гипотезы возрастают с числом замен и длиной последовательности. Для последовательности данной длины и числа замен, эффективность значительно увеличивается даже при малых отклонениях между последовательностями в эволюционных картинах ($pdf = 2$).

Этот тест не требует априорного знания картины замен, степени гетерогенности сайтов и эволюционных взаимоотношений между последовательностями. Компьютерное моделирование показало, что ID-тест применим для множества биологических моделей эволюции последовательностей. При применении этого теста для анализа 3789 пар ортологичных генов человека и мыши получено, что

наблюдаемые картины замен в нейтральных сайтах не являются гомогенными в 41% генов, что обусловлено изменением Г+Ц содержания [6]. Таким образом, предлагаемый тест может использоваться как диагностическое средство для идентификации генов и линий, изменяющимися с существенно отличающимися эволюционными процессами, что отражается в наблюдаемых картинах замен. Идентификация таких генов и линий является важным начальным шагом при проведении сравнительных геномных и молекулярно-филогенетических исследований для анализа эволюционных процессов, в ходе которых формировались геномы организмов.

Эффективность ID-теста по сравнению ранее предложенными тестами. При сходных частотах оснований в последовательностях ранее часто использовался ?2-тест:

$$\chi^2 = \sum (f_{1i} - f_{2i})^2 / (f_{1i} + f_{2i})_{(12)}$$

где f_{1i} и f_{2i} равны количеству основания i в сравниваемых последовательностях. При компьютерном моделировании получено, что ?2 является высоко консервативным, поэтому он является менее эффективным, чем ID-тест [6]. Причина консервативного характера классического ?2-теста заключается в том, что он основан на предположении о независимости сравниваемых показателей. Но это не так, потому что частоты, полученные для гомологичных последовательностей, не являются независимыми из-за общей эволюционной истории. Эта зависимость обуславливает повышение значения знаменателя в формуле ?2-теста, поскольку включает информацию по всем сайтам, включая даже те, в которых не происходили замены. Включение этих инвариантных сайтов в знаменатель сильно снижает значение ?2, в то время как их вклад в числитель автоматически аннулируется. Этот эффект является более жестким для близкородственных последовательностей из-за большей доли идентичных сайтов, что связано с наличием общего предшественника.

Проблему сходных оснований в сайте, обусловленную общим происхождением также изучали А. Ржетский и М. Ней [9], разработавшие строгий статистический тест равенства частот нуклеотидов (или аминокислот) для множества последовательностей. Однако этот тест слишком либерален, что обусловлено нарушением некоторых сделанных в тесте предположений.

Применение ID-теста. Определение индекса несоответствия и проведение ID-теста является эффективным способом идентификации пар последовательностей, которые эволюционировали с сильно отличающимися картинами замен. Идентификация генов и видов с нетипичными картинами замен также может быть полезной для объяснения эволюционных механизмов наблюдаемых различий.

В молекулярной филогенетике возможность определения таких пар последовательностей с помощью ID-теста полезна для дифференциального выбора метода реконструкции эволюционного дерева, учитывающего или не учитывающего гипотезу о гомогенности [1, 3]. Однако стоит отметить, что большое количество параметров в этих сложных методах может препятствовать

получению высоко достоверных результатов. Альтернативно исследователи могут удалять последовательности, которые не удовлетворяют гипотезе о гомогенности, предварительно используя ID-тест, или использовать более простые модели для корректных филогенетических оценок.

Помимо этого применение ID-теста будет обеспечивать получение корректных эволюционных дистанций [1, 2, 10], вычисление на их основании точных скоростей эволюции, а также правильность результатов селекционных тестов.

Литература

1. Барковский Е.В., Бутвиловский А.В., Бутвиловский В.Э., Давыдов В.В., Хрусталеv В.В., Козюлевич С.Р. Методы молекулярной эволюции и филогенетики: учеб.-метод. пособие. – Мн.: БГМУ, 2005. – 63 с.
2. Барковский Е.В., Козюлевич С.Р., Бутвиловский А. В., Хрусталеv В.В., Ачинович О.В. Характеристика методов определения эволюционных расстояний между нуклеотидными последовательностями генетических макромолекул. // Здравоохранение. – Минск, 2005, №5. – С. 37-43.
3. Хрусталеv В.В., Барковский Е.В., Бутвиловский А.В., Козюлевич С.Р., Ачинович О.В. Основные методы анализа эволюционных отношений между последовательностями генетических макромолекул. //Здравоохранение. – Минск, 2005, № 8. – С. 11-13.
4. Cornish-Bowden A. Assessment of protein sequence identity from amino acid composition data //J. Theor. Biol. – 1977. – Vol. 65. – P. 735-742.
5. Jukes T.H., Cantor C.R. Evolution of protein molecules//In H.N.Munro, ed., Mammalian protein Metabolism. – 1969. – P.21 – 132. Academic Press, New York.
6. Kumar S., Gadagkar S.R. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences //Genetics – 2001. – Vol. 158. – P. 1321-1327.
7. Kumar S., Tamura K., Nei M. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment //Brief. Bioinform. – 2004. – Vol.5. – P.150-163.
8. Nei M., Kumar S. Molecular Evolution and Phylogenetics.-Oxford University Press, New York, 2000.
9. Rzhetsky A., Nei M. Tests of applicability of several substitution models for DNA sequence data //J. Mol. Evol. – 1995. – Vol. 12. – P. 131-151.
10. Tamura K., Kumar S. Evolutionary distance estimation under heterogeneous substitution pattern among lineages//Mol.Biol.Evol. – 2002. – Vol.19. – P.1727 – 1736.