Здравоохранение. – Минск, 2007, №3. – С. 14–17

Бутвиловский А.В., Барковский Е.В, Бутвиловский В.Э.

Методы определения времен дивергенции различных таксономических групп организмов

Белорусский государственный медицинский университет

Со времен Ч. Дарвина исследователи пытались реконструировать эволюционную историю организмов на Земле и построить филогенетическое древо жизни. Оптимальное решение этой проблемы возможно по ископаемым останкам, но, к сожалению, они фрагментарны и неполны. Поэтому большинство ученых пользуются методами сравнительной морфологии и физиологии. Используя достижения этих наук, классические эволюционисты смогли предположить большинство аспектов эволюционного развития организмов. Однако эволюционные изменения морфологических и физиологических признаков настолько сложны, что они не позволяют воссоздать точную картину эволюционного процесса, и поэтому детали реконструируемых филогенетических деревьев зачастую противоречивы.

Последние достижения молекулярной биологии решительно изменили ситуацию. Со времени открытия строения нуклеиновых кислот и широкомасштабного секвенирования геномов, стало возможно изучать эволюционные связи между организмами путем сравнения их нуклеотидных последовательностей. Такой подход имеет некоторые преимущества над классическими морфологическими и физиологическими исследованиями, так как:

1. ДНК состоит из 4 типов нуклеотидов (А, Т, Ц, Г), что может быть использовано для сравнения любых групп организмов, включая бактерий, протистов, грибов, растений и животных. При использовании классического подхода такое сравнение невозможно.

2. В процессе эволюции изменения ДНК происходят более или менее регулярно, что позволяет использовать математические модели для определения изме-

нений и сравнения ДНК филогенетически отдаленных организмов. Оценить эволюционные изменения морфологических признаков крайне сложно, особенно за короткий период исторического развития органического мира.

3. Геномы организмов состоят из объемной последовательности нуклеотидов и содержат значительно больше информации, чем морфологические признаки.

Поэтому следует ожидать, что молекулярная филогенетика прояснит ситуацию по установлению многих ветвей филогенетического древа, что невозможно было сделать при использовании классического подхода.

Систематика до сих пор остается одним из наиболее спорных разделов биологии. Определение видов, родов, семейств и других таксономических категорий часто довольно субъективно. В филогенетике гораздо меньше спорных вопросов, так как первоначально устанавливаются филогенетические связи, а лишь затем – таксономическая категория организмов. Однако эти разделы биологии тесно взаимосвязаны, потому что классификация организмов отражает их эволюционную историю.

Следует отметить, что в отличие от палеонтологических доказательств, данные молекулярной эволюции относятся только к группам организмов, существующим в настоящее время.

Определенные времена дивергенции до сих пор остаются достаточно противоречивыми [5, 8, 16, 21]. Например, время дивергенции человека и шимпанзе варьирует от 3,6 [9] до 13 млн. лет назад [5]. Наличие таких противоречий объясняется использованием различных методов определения времен дивергенции, изучением различных частей геномов (ядерных генов, митохондриальных генов, некодирующих участков и др.) и различными калибровочными точками.

Применение «молекулярных часов» при определении времен дивергенции. Теоретической основой, на которой базируется определение времен дивергенции различных организмов в молекулярной эволюции, является гипотеза о

«молекулярных часах», предложенная в 1962-1965 гг. Э. Цукеркэндлом и Л. Полингом [3, 31, 32]. Согласно этой гипотезе число аминокислотных замен в сравниваемых белках организмов двух видов приблизительно пропорционально времени их дивергенции. Отсутствие строгой пропорциональности связано с тем, что ни один ген или белок не эволюционируют со строго постоянной скоростью на протяжении длительного времени, так как через некоторое время может измениться их функции, а также вариация уровня мутаций и репарации у различных групп организмов [7].

Следует отметить, что «молекулярные часы» были предметом дебатов в течение нескольких десятилетий [6]. Однако даже противники этой гипотезы согласны с тем, что если скорость замен не является строго постоянной, то возможно получить приблизительные значения времен дивергенции, которые окажутся полезными при отсутствии палеонтологических данных [8, 12, 30].

Методы определения «молекулярных часов». Методы определения «молекулярных часов» (метод Тадзимы [20, 26] и метод Такезаки [28]) позволяют определить постоянство скорости эволюции в пределах изучаемого набора последовательностей.

Наиболее часто используется метод Тадзимы, также называемый тестом сравнения скоростей эволюции. В этом методе для определения равенства скорости эволюции двух изучаемых последовательностей (или кластеров последовательностей) дополнительно используется третья, заведомо несходная последовательность, называемая внешней группой.

Наблюдаемое число сайтов, в которых последовательности 1, 2 и 3 содержат аминокислоты или нуклеотиды i, j и k, обозначается как n_{ijk}. Согласно гипотезе «молекулярных часов», $E(n_{ijk}) = E(n_{jik})$, независимо от модели замен и различий скорости замен в зависимости от сайта. Изучая нуклеотидные последовательности на наличие «молекулярных часов», можно анализировать транзиции, трансверсии

и различные положения нуклеотида в кодоне в отдельности. Достоверность различий скоростей замен в этом случае определяется по методу χ^2. Значение $p>0,05$ позволяет судить о недостоверных различиях, а следовательно о наличии «молекулярных часов». При обнаружении последовательностей, которые эволюционировали слишком быстро или медленно, возможно их исключение из анализа (хотя оно не всегда необходимо) с последующей оценкой времен дивергенции для остальных видов. Удаление таких последовательностей, безусловно, обеспечит повышение точности вычисляемых в последствии значений времен дивергенции. Еще одним способом повышения точности, который в последнее время чаще всего используется, является определение времени дивергенции для большего числа генов или белков (для нейтрализации различий их скоростей эволюции) [18].

Определение времен дивергенции таксономических групп организмов возможно по одному белку или гену, а также по нескольким белкам или генам.

Определение времени дивергенции по одному белку или гену. Первоначально определение времен дивергенции проводилось по одному белку или гену [18, 23, 24]: после определения наличия «молекулярных часов» строится дендрограмма [1, 4] для одного белка или гена нескольких изучаемых видов организмов. Для создания корня дендрограммы вводится заведомо несходная аминокислотная или нуклеотидная последовательность (внешняя группа) (рис. 1).

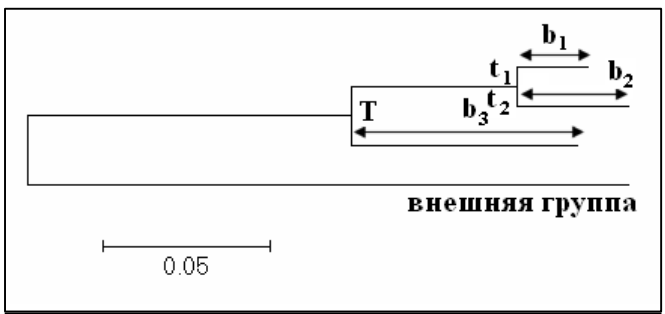

Рис. 1. Ветвь - специфичный вариант метода определения времен дивергенции для одного белка и гена.

Затем вводится еще одна последовательность изучаемого белка или гена, позволяющая ввести калибровочную точку (точно известное время дивергенции двух последовательностей разных видов организмов). Наиболее часто в качестве калибровочных точек используются время дивергенции

птиц и млекопитающих (310 млн. лет), человека и непарнокопытных (90 млн. лет) и др.

Использование первой калибровочной точки основано на том, что установлено появление синапсид и диапсид (предшественников млекопитающих и птиц соответственно) в каменноугольном периоде около 310 млн. лет назад [18].

При введении калибровочной точки производится перерасчет длин ветвей с учетом постоянства скорости эволюции и линеаризация дендрограммы (создание шкалы эволюционных дистанций и временной шкалы). Создание временной шкалы основано на использовании скорости замен (r), вычисленной для калибровочной точки и соответствующей ей длине ветви.

Если T – калибровочная точка, то скорость замен равна длине ветви (b_3), деленной на время дивергенции $r = b_3/T$. Зная r, можно определить t_1:

$$t_1 = b_1/r = b_1/ (b_3/T) = b_1/b_3 \times T \quad (1).$$

Аналогично определяется значение t_2:

$$t_2 = b_2/r = b_2/ (b_3/T) = b_2/b_3 \times T \quad (2).$$

В случае использования нескольких калибровочных точек среднее значение скорости замен определяется с учетом всех полученных значений r.

Когда имеются данные по многим различным белкам или генам, часто вычисляется среднее время дивергенции t_1 для всех изученных белков или генов:

$$t_1 = \sum_{i=1}^{k} t_{1i}/k \quad (3).$$

где t_{1i} – значение t_1, полученное для белка или гена i, а k – число изученных белков или генов.

Однако полученное таким образом значение t_1 теоретически не является корректным, даже если значения длин ветвей для каждого белка корректны. Если b_{1i} и b_{3i} – это корректные значения b_1 и b_3 для белка i, тогда $t_{1i} = (b_{1i} \times T/b_{3i})$. В этом случае корректное значение t_1 предпочтительно равно $t_1 = (\sum_i b_{1i}) \times T/(\sum_i b_{3i})$, а не $t_1 =$

$\sum_i (b_{1i}/b_{3i})\times T/k$. Если предположить, что b_1 и b_3 – случайные величины (индекс отброшен), то математическое ожидание b_1/b_3 приблизительно равно:

$$E\ (b_1/b_3) = (E(b_1)/E(b_3)) - (Cov(b_1, b_3)/(E^2(b_3)) + (E(b_1)V(b_3)/E^3(b_3)) \qquad (4).$$

где $V(b_3)$ – варианса b_3, $Cov(b_1, b_3)$ – ковариансы b_1 и b_3. Оценка второго и третьего слагаемого в формуле (4) свидетельствует, что значения, полученные по формуле (1) будут завышены, особенно если точка калибровки меньше, чем определяемое время дивергенции.

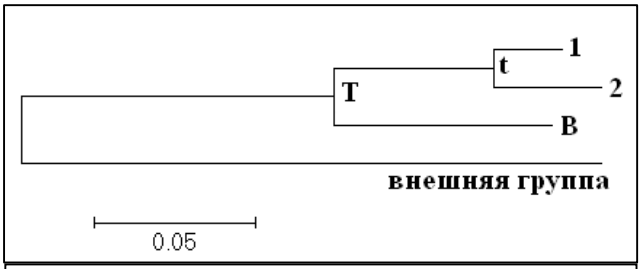

Рис. 2. Неспецифичный вариант метода определения времен дивергенции для одного белка и гена.

В случае, когда известно t_1, а не T, его можно вычислить по формуле:

$$T = \sum_i (b_{1i}/b_{3i})\times t_1/k \qquad (5).$$

Значение $T_i \equiv (b_{1i}/b_{3i})\times t_1$ может стать слишком большим, если b_{3i} близко к нулю или маленьким (но не меньше t_1), если b_{3i} относительно велико. Поэтому T имеет тенденцию к завышению при больших вариациях b_{1i}. Для устранения этого завышения необходимо использовать связанные дистанции. Рассмотренный выше вариант определения значения t_1 и t_2 называется ветвь - специфичным (линия - специфичным).

При использовании неспецифичного варианта [18] (рис. 2) значение t вычисляется по формуле:

$$t = d_{12}/2r = d_{12}\times 8T/2(d_{1B} + d_{2B}) \qquad (6),$$

где d_{12} – эволюционная дистанция между последовательностями, время дивергенции которых необходимо определить, а d_{1B} и d_{2B} – дистанции между последовательностью, введенной для калибровки, и каждой из последовательностей, время дивергенции которых необходимо определить.

Использование каждого из рассмотренных вариантов имеет строгие показания. Так при разной длине ветвей последовательностей, между которыми необ-

ходимо определить время дивергенции (ситуация, изображенная на рис. 1 и 2), следует использовать исключительно неспецифичный вариант. При одинаковой длине ветвей – возможно использование обоих вариантов, но предпочтительно используется более простой ветвь - специфичный вариант.

Эволюционные дистанции, используемые для определения времен дивергенции.

Использование для построения дендрограмм количества наблюдаемых аминокислотных или нуклеотидных замен имеет ряд недостатков [1, 2]. Если все изучаемые последовательности сходны, то достаточно точные значения времен дивергенции можно получить, используя дистанции, корректированные в соответствии с распределением Пуассона. Однако вычисление этой дистанции основано на предположении о постоянстве скорости замен во всех сайтах. Однако это предположение редко выполняется, поскольку скорость эволюции изменяется в соответствии с гамма-распределением [29].

Сложный способ определения дистанции между аминокислотными последовательностями, учитывающий различия скорости замен среди различных сайтов и среди различных пар аминокислот, разработан Н. В. Гришиным [1, 13]. Однако эта дистанция приблизительно равна гамма-дистанции при параметре α равном 0,65, что делает наиболее целесообразным использование гамма-дистанции [22].

С. Кумар и соавт. [19] обнаружили четкую взаимосвязь между дистанцией по четырехкратно-вырожденным сайтам (d_4), определенной методом Тадзимы-Нея [27], с временами дивергенции различных видов млекопитающих. На основании этого можно предположить, что d_4-дистанция также может использоваться для определения времен дивергенции.

Определение времени дивергенции по нескольким белкам или генам.

Методы этой группы позволяют получить корректные значения b_1 и b_3 путем усреднения значений времен дивергенции, полученных при попарных сравнениях различных белков и генов [23]. К данным методам относятся:

1. Метод вычисления простой средней дистанции (d_1), основанный на установлении дистанции для каждого белка с последующим расчетом среднего значения для всех белков. Недостатком дистанции d_1 является то, что она не учитывает различия длины последовательностей.

2. Метод вычисления средней дистанции с учетом длины последовательностей (d_2). Поскольку белки состоят из разного количества аминокислот, то предпочтительно следует вычислять среднюю дистанцию с учетом количества мономеров.

3. Метод вычисления гамма-дистанции для множества белков (d_3). Зная, что параметр α варьирует для различных белков и скорость аминокислотных замен следует гамма-распределению, можно предположить, что при одновременном рассмотрении последовательностей многих белков как единого целого, скорость замен аминокислот в сайтах также изменяется в соответствии с гамма-распределением. Поэтому возможно определение параметра α для всех изучаемых последовательностей [14] с последующим вычислением гамма-дистанции. Определенная таким образом гамма-дистанция, часто называется гамма-дистанцией для множества белков. Стандартные ошибки значений дистанций при этом вычисляются методом бутстрэп [1].

Современные основные и альтернативные направления при определении времен дивергенции.

Согласно данным Г. Глазко и М. Нея [11] основные направления при определении времен дивергенции заключаются в преимущественном использовании:

1) методов определения времен дивергенции по нескольким белкам или генам;

2) дистанций между последовательностями белков (вычисляются при использовании более простых моделей, последовательности белков более консервативны [2, 18, 22]), а не между последовательностями генов;

3) ядерных, а не митохондриальных белков;

4) дендрограмм для групп видов;

5) нескольких достоверных калибровочных точек.

Некоторые из этих направлений являются довольно дискутабельными, так как существуют и альтернативные направления. Некоторые авторы [8, 12] полагают, что некодирующие участки последовательностей ДНК не подвержены действию естественного отбора и могут содержать более надежную информацию. Однако из-за часто наблюдаемых в этих участках ДНК вставок и делеций, получаемые значения менее точны [11].

С увеличением количества секвенированных митохондриальных белков, их все чаще используют для определения времен дивергенции [5, 16]. Однако полученные таким образом значения времен дивергенции противоречивы, поскольку значительно варьируют скорости эволюции этих белков [10]. В 1990 г. был предложен метод вычисления времен дивергенции, позволяющий учитывать вариации скорости эволюции среди различных групп организмов [15, 17]. Данный метод апробирован именно на митохондриальных белках и достаточно сложен [25].

Общепринятые времена дивергенции различных таксономических групп согласно данным молекулярной эволюции.

Значения времен дивергенции, полученные при анализе генов и белков [11, 18, 19, 23], представлены в табл. 1.

Таблица 1

Времена дивергенции различных таксономических групп, используемые в молекулярной эволюции

Дивергировавшие организмы	Время дивергенции, млн. лет назад

человек/шимпанзе	5,5 [18]
человек/горилла	7 [18]
человек/орангутанг	8 [18], 13 [11]
человек/гиббон	15 [18]
человек/мартышкообразные	23 [18]
мышь/крыса	41 [18]
человек/широконосые обезьяны	48 [18]
человек/тупайи	86 [18]
человек/зайцеобразные	91 [18]
человек/хищные	92 [18], 95 [19]
человек/непарнокопытные	92 [18], 95 [19]
человек/парнокопытные	92 [18]
приматы/грызуны	109-112 [18],120 [23]
человек/неполнозубые	129 [18]
птицы/крокодилы	276 [18]
человек/липидозавры	276 [18]
человек/амфибии	360 [18]
человек/лучеперые рыбы	450 [18]
человек/хрящевые рыбы	528 [18]
человек/бесчелюстные	564 [18]
человек/дрозофилы	833 [23]
человек/нематоды	970 [23]
человек/грибы	1392 [23]
человек/растения	1392 [23]
человек/протисты	1717 [23]
человек/эубактерии	3036 [23]

Примечание. Значения времен дивергенции, вычисленные М. Неем и соавт. [23] вычислены для 104 ортологичных белков по d_3-дистанциям с α =1,24.

Рядом исследователей проведено сопоставление времен дивергенции, полученных по молекулярным и палеонтологическим данным. С. Кумар и соавт. [18] установили, что коэффициент корреляции вычисленных времен дивергенции по

658 генам, относящихся к 207 видам позвоночных животных и времен дивергенции по палеонтологическим данным равен 0,99.

Значения времен дивергенции, вычисленные при использовании методов молекулярной эволюции, как правило, выше, чем таковые по данным палеонтологии. Исследователи, занимающиеся изучением вопросов молекулярной эволюции, объясняют это неполными палеонтологическими данными, а палеонтологи – неточностью данных молекулярной эволюции [23]. Установлению истины, безусловно, будет способствовать увеличение количества секвенированных белков и генов. Еще одной нерешенной проблемой датировки времен дивергенции является отсутствие точных палеонтологических данных для введения калибровочных точек относящихся к ранним этапам эволюционного процесса. Именно по этой причине установленные времена дивергенции человека и эубактерий, человека и растений и т.д., сильно отличаются.

В заключение следует отметить, что создание корректного древа жизни остается одной из актуальных задач современной биологии, решение которой возможно только при взаимодействии палеонтологов, генетиков и исследователей, занимающихся изучением молекулярной эволюции.

Литература.

1. Барковский Е.В., Бутвиловский А.В., Бутвиловский В.Э., Давыдов В.В., Хрусталев В.В., Казюлевич С.Р. Методы молекулярной эволюции и филогенетики: учеб.-метод. пособие. – Мн.: БГМУ, 2005. – 63 с.

2. Бутвиловский А.В., Барковский Е.В., Бутвиловский В.Э., Давыдов В.В. //Здравоохранение. – Минск, 2006, №1 С. 42-44.

3. Кимура М. Молекулярная эволюция: теория нейтральности. - М., 1985. - 398 с.

4. Хрусталев В.В., Барковский Е.В., Бутвиловский А.В., Казюлевич С.Р., Ачинович О.В. //Здравоохранение. – Минск, 2005, №8. – С. 11-13.

5. Arnason U., Gullberg A., Janke A. //J. Mol. Evol. – 1998. – Vol. 47. – P. 718-727.

6. Britten R.J. //Science. – 1986. – Vol. 231. – P. 1393-1398.

7. Bromham L**.** //Mol. Biol. Evol. – 2002. – Vol. 19. – P. 302-309.

8. Chen F.C., Li W.-H. //Am. J. Hum. Genet. – 2001. – Vol. 68. – P. 444-456.

9. Easteal S., Herbert G. //J. Mol. Evol. – 1997. – Vol. 44. – P. 121-132.

10. Gissi C., Reyers A., Pesole G., Saccone C. //Mol. Biol. Evol. – 2000. – Vol. 17. – P. 1022-1031.

11. Glazko G., Nei M. //Mol. Biol. Evol. – 2003. – Vol. – 20(3). – P. 424-434.

12. Goodman M., Porter C.A., Czelusniak J., Page S.L., Schneider H., Shoshani J., Gunnell G., Groves C.P. //Mol. Phylogenet. Evol. – 1998. –Vol. 9. – P. 585-598.

13. Grishin N. V. //J. Mol. Evol. – 1995. – Vol. 41. – P. 675-679.

14. Gu X., Zhang J. //Mol. Biol. Evol. – 1997. – Vol. 15. – P. 1106-1113.

15. Hasegawa M., Thorne J.L., Kishino H. //Genes Genet. Syst. – 2003. – Vol. 78. – P. 267-283.

16. Horai S., Hayasaka R., Kondo R., Tsugane K., Takahata N. //Proc. Natl. Acad. Sci. USA. – 1995. –Vol. 92. – P. 532-536.

17. Kishino H., Hasegawa M. //Meth. Enzymol. – 1990. – Vol. 183. – P. 550-570.

18. Kumar S., Hedges S.B. //Nature. – 1998. – Vol. 392. – P. 917-920.

19. Kumar S., Subramanian S. //Proc. Natl. Acad. Sci. USA. – 2002. – Vol. 99(2). – P. 803-808.

20. Kumar S., Tamura K., Nei M. //Brief. Bioinformatics. - 2004. - Vol. 5. – P. 150-160.

21. Lee M.S. //J. Mol. Evol. – 1999. – Vol. 49. – P. 385-391.

22. Nei M., Kumar S. //Molecular Evolution and Phylogenetics. - Oxford University Press, New York, 2000.

23. Nei M., Xu P., Glazko G. //Proc. Natl. Acad. Sci. USA. – 2001. – Vol. 98. – P. 2497-2502.

24. O´hUigin C., Li W.-H. //J. Mol. Evol. – 1992. – Vol. 35. – P. 377-384.

25. Sanderson M.J., //Mol. Biol. Evol. – 1997. – Vol. 14. – P. 1218-1231.

26. Tajima F. //Genetics. – 1993. – Vol. 135. – P. 599-607.

27. Tajima F., Nei M. //Mol. Biol. Evol. –1984. – Vol. 1. – P.269–285.

28. Takezaki N., Rzhetsky A., Nei M. //Mol. Biol. Evol. – 1995. – Vol. 12. – P. 823-833.

29. Uzzell T., Corbin K. //Science. – 1971. – Vol. 172. – P. 1089-1096.

30. Wilson A.C., Carlson S.S., White T.J. //Annu. Rev. Biochem. – 1977. – Vol. 46. – P. 573-639.

31. Zuckerkandl E., Pauling L. //In: Evolving Genes and Proteins. – 1965. – Acad. Press., NY. – P. 97-166.

32. Zuckerkandl E., Pauling L. //In: Horizons in Biochemistry. – 1962. – Acad. Press., NY. – P. 189-225.