ЭВОЛЮЦИЯ ГЕНОМОВ ЭУКАРИОТ И ПРИНЦИП МАКСИМАЛЬНОЙ ПАРСИМОНИИ

И.Б. Рогозин[1, 2], Ю.И. Вульф[1, 2], В.Н. Бабенко[1, 2], Е.В. Кунин[1]

[1] Национальный центр биотехнологических исследований, Национальные институты здоровья, Бесезда 20894, США; e-mail: rogozin@bionet.nsc.ru;
[2] Институт цитологии и генетики СО РАН, Новосибирск, Россия

Метод Долло-парсимонии, впервые формализованный Джеймсом Фаррисом в 1977 г., основан на предположении о том, что сложная биологическая система, которая была потеряна организмом в ходе эволюции, не может снова появиться в своем исходном виде. Мы применили этот метод для исследования закономерностей эволюции эукариотических геномов. Реконструкция сценария потери и приобретения генов выявила массовые потери генов в группе одноклеточных грибов, в то время как для животных характерно возникновение большого числа новых генов. Исследование экзон/интронной структуры эукариотических генов показало, что общий предок животных, грибов и растений имел большое число интронов. Интересно, что около 30 % интронов в генах малярийного паразита *Plasmodium falciparum* имеют ортологов в других геномах. Были выявлены контрастные отличия в относительных скоростях встройки и потери интронов в разных группах эукариотических организмов. В целом проведенные исследования выявили высокую изменчивость как репертуара генов, так и экзон/интронной структуры ортологичных генов.

Введение

Некоторые результаты этой работы обсуждались первым автором этой статьи с Вадимом Александровичем Ратнером, который делал доклады на SMBE2001 (Ежегодная конференция научного общества по молекулярной биологии и эволюции) в Афинах (штат Джорджия, США) в 2001 г. Поэтому нам представляется интересным осветить тему эволюции геномов в выпуске журнала «Информационный вестник ВОГиС», посвященном памяти этого выдающегося человека.

Вадим Александрович представил два доклада по микроэволюции мобильных элементов в дрозофиле, и эти сообщения вызвали большой интерес у участников конференции. Нами был представлен доклад об исследовании эволюции прокариотических геномов с использованием различных методов филогенетического анализа [краткое описание этих методов приведено в таблице 1, подробное описание методологии дается в книге Вадима Александровича с коллегами (Ратнер и др., 1985а) и в книге Масатоши Нея и Судиа Кумара (Ней, Кумар, 2004)]. После доклада мы много говорили об исследовании глобальных принципов эволюции на геномном уровне и реконструкции ключевых моментов истории про- и эукариотических организмов. Эта тема всегда интересовала Вадима Александровича (Korostyshevsky *et al*., 1974; Ратнер и др., 1985а, б). Мы думали над возможностью совместных работ в этой области, но, к сожалению, этим планам не суждено было осуществиться.

Одним из основных методов, используемых нами для филогенетических исследований эволюции на геномном уровне, является метод Долло-парсимонии. Закон Долло, также известный как «the Law of Irreversible Evolution», был сформулирован бельгийским биологом Луи Долло в 1893 г. (Dollo, 1893). Этот закон утверждает, что сложная биологическая система, которая была потеряна организмом в ходе эволюции, не может снова появиться в своем исходном виде. Другими словами, одна и та же последовательность мутационных событий, которая привела к появлению биологической системы, не может повториться дважды в силу стохастичности эволюционных процессов. Метод Долло-парсимонии как метод филогенетического анализа был впервые формализован Фарри-сом в 1977 г. (Farris, 1977). В простейшем

Таблица 1

Матрица присутствия/отсутствия некоторых ортологичных групп генов (**КОГ**ов) в эукариотических геномах

Организмы	KOG2207, 3'-5' exonuclease	KOG4125, acid trehalase	KOG0006, E3 ubiquitin-protein ligase	KOG0090, Signal recognition particle receptor β-subunit	KOG0050, mRNA splicing factor CDC5
At	1	0	0	1	1
Ec	0	0	0	0	1
Sc	0	1	0	1	1
Sp	0	0	0	1	1
Ce	1	0	1	1	1
Dm	1	1	1	1	1
Hs	0	1	1	1	1

Для каждого **КОГ**а приведен его номер и предсказанная биологическая функция. «1» означает наличие данного гена в организме, «0» – его отсутствие. Сокращения: At – *Arabidopsis thaliana,* Ce – *Caenorhabditis elegans*, Dm – *Drosophila melanogaster*, Ec – *Encephalitozoon cuniculi*, Hs – *Homo sapiens*, Sc – *Saccharomyces cerevisiae*, Sp – *Schizosaccharomyces pombe*.

виде этот метод рассматривает два состояния в каждом исследуемом сайте: примитивное (0) и производное (1). Возникновение производного состояния 1 из примитивного состояния 0 ($0 \rightarrow 1$) разрешается только один раз в рассматриваемом филогенетическом дереве, в то время как число переходов $1 \rightarrow 0$ не ограничено. Метод Долло-парсимонии минимизирует число переходов $1 \rightarrow 0$. В молекулярных исследованиях этот метод иногда применялся для анализа сайтов рестрикции (DeBry, Slade, 1985). В последние годы метод Долло-парсимонии стал актуален для анализа геномных данных (Ней, Кумар, 2004).

Сравнительные исследования геномов меняют наше понимание закономерностей эволюции про- и эукариот. Например, было показано, что массовая потеря и горизонтальный перенос генов являются широко распространенными эволюциоными событиями у прокариот (Doolittle, 1999; Koonin *et al*., 2000). Это яркий пример смены научной парадигмы в эволюционных исследованиях, и теперь массовая потеря и горизонтальный перенос генов из разряда редких событий перешли в разряд основных эволюционных механизмов эволюции прокариот. Темпы потери генов в некоторых прокариотических линиях просто удивительны: у некоторых паразитических микробов больше 80 % генов было потеряно за последние 200–300 млн лет (Moran, 2002). Выявление горизонтального переноса генов требует сложных теоретических и экспериментальных подходов, тем не менее были получены многочисленные доказательства важности этого процесса в эволюции прокариот (Ochman *et al*., 2000; Koonin *et al*., 2001; Gogarten *et al*., 2002; Nakamura *et al*., 2004). Вопрос о переносе генов между эукариотами является открытым. В настоящее время доминирует точка зрения, что этот процесс не является существенным для эволюции эукариот. Однако число потерь генов в некоторых линиях эукариот может быть значительным: сравнение геномов двух дрожжей, *Saccharomyces cerevisiae* и *Schizosaccharomyces pombe*, показало, что более 10 % генов *S. cerevisiae* было потеряно после расхождения этих двух видов (Aravind *et al*., 2000). В эукариотических паразитах, таких, как микроспоридия *Encephalitozoon cuniculi*, потери генов еще более драматичны (Katinka *et al*., 2001). В целом закономерности геномной эволюции в многоклеточных эукариотах, в частности роль процесса потери генов, остаются неясными.

Исследование закономерностей эволюции геномов на основе Долло-парсимонии

Ортологичные и паралогичные гены. Секвенирование геномов из различных таксонов позволяет проводить количественный анализ эволюции на геномном уровне. Необходимым условием таких исследований является подробная классификация генов из исследуемых геномов на основе их функции и сходства (гомологии) на уровне аминокислотной последовательности. Гомологичные гены разделяются на ортологичные и паралогичные гены (Fitch, 1970; Sonnhammer, Koonin, 2002). Ортологи – это гомологичные гены в геномах разных организмов, являющиеся результатом эволюции одного и того же гена, который присутствовал в геноме общего предка этих организмов (вертикальное наследование). Паралогичные гены – это гомологичные гены внутри одного генома, произошедшие путем дупликации предкового гена. Ортологи и паралоги – это две стороны одного явления: если дупликация произошла после разделения сравниваемых организмов, то ортология устанавливается между группами паралогов, а не между индивидуальными генами. Гены, которые принадлежат к таким ортологичным генам, называются кооортологами (Sonnhammer, Koonin, 2002). Дупликации генов до расхождения основных линий про- и эукариот дали начало разным семействам ортологичных генов (например, разные семейства аминотрансфераз) (Cho, Doolittle, 1997).

Надежное выявление ортологов и паралогов важно для реконструкции эволюции геномов, основными механизмами которой являются вертикальное наследование, потери генов и горизонтальные переносы генов (Snel *et al*., 2002). Самым лучшим способом определения ортологических взаимоотношений генов в разных организмах (включая кооортологию) являются построение филогенетических деревьев и сравнение этих деревьев с эволюционной историей исследуемых организмов. В идеальном случае филогенетическое дерево ортологов должно соответствовать эволюционному дереву видов, которое предполагается известным. Однако во многих случаях дерево видов неизвестно. Кроме того, построение надежных филогенетических деревьев для тысяч–десятков тысяч генных семейств технически трудоемко, если вообще возможно. В связи с этими проблемами были разработаны методы определения наборов возможных ортологов без использования филогенетических деревьев. Эти подходы основаны на программах быстрого поиска сходства между двумя геномами, например, на программе BLASTP (http://www.ncbi.nlm.nih.gov/BLAST/).

BLASTP сравнивает две базы данных аминокислотных последовательностей, которые кодируются геномами A и B. Для каждой аминокислотной последовательности X (для краткости мы будем использовать слово «ген») из генома A выявляется ген Y из генома B, который имеет наибольшее значение некоей меры сходства (например, самое большое значение «BLASTP alignment score»). В этом случае ген Y является наилучшим кандидатом на ортолога гена X в геноме B. Такое взаимоотношение ($X \rightarrow Y$) между генами X и Y называется организм-специфичным наилучшим выравниванием (**S**pecies-**S**pecific **B**est **H**it, **SSBH**) (Tatusov *et al*., 1997). Данный подход основан на предположении, что ортологичные гены больше похожи друг на друга, чем на другие гены из-за структурно-функциональных ограничений, которые накладываются на эволюцию этих генов. В некоторых случаях (например, в случае быстро эволюционирующих генов) это условие не выполняется и тогда установление ортологических взаимоотношений между такими генами осложняется. Для более надежного выявления ортологических взаимоотношений при исследовании двух геномов иногда используются пары генов X и Y, которые имеют симметричные **SSBH** ($X \rightarrow Y$ и $Y \rightarrow X$) (Tatusov *et al*., 1996), однако это условие выполняется не для всех ортологичных генов.

При сравнении нескольких геномов пары вероятных ортологов, определенных на основе **SSBH,** могут быть объединены в кластеры ортологичных генов (Tatusov *et al*., 1997). Именно такой подход, объединенный с процедурой автоматического выявления кооортологов и анализа генов, содержащих несколько доменов, был использован при построении базы данных кластеров ортоло-

гичных групп генов (**КОГ**ов) [the database of Clusters of Orthologous Groups (**COG**s) of proteins, http://www.ncbi.nlm.nih.gov/COG/] (Tatusov *et al*., 1997; Koonin *et al*., 2004). Текущая версия базы данных **КОГ**ов включает ~ 70 % белков, которые были обнаружены в прокариотических геномах. Текущая версия базы данных **КОГ**ов также включает 7 эукариотических организмов: человек (Hs), червь *Caenorhabditis elegans* (Ce), муха *Drosophila melanogaster* (Dm), два вида дрожжей [*Saccharomyces cerevisiae* (Sc) и *Schizosaccharomyces pombe* (Sp)], микроспоридия *Encephalitozoon cuniculi* (Ec), а также растение *Arabidopsis thaliana* (At) (Koonin *et al*., 2004). Использование базы данных **КОГ**ов стало одним из стандартных подходов для аннотации новых геномов и выбора новых мишеней для структурных исследований белков (Koonin, Galperin, 2002).

Необходимо подчеркнуть, что описания многих геномов содержат значительное количество ошибок различного рода: неправильно предсказанные гены, отсутствие некоторых функциональных генов (например, систематическое недопредсказание генов, кодирующих короткие рибосомальные белки), отсутствие значительных фрагментов геномной ДНК, плохое различение генов и псевдогенов, встройки чужеродной ДНК (например, значительное загрязнение текущей версии генома комара *Anopheles gambiae* последовательностями бактериального происхождения). База данных **КОГ**ов основана на сравнительном анализе геномных последовательностей и, следовательно, содержит определенную долю ошибок, часть из которых устраняется в ходе экспертной проверки (Koonin *et al*., 2004). В целом базы данных геномных последовательностей и ортологичных генов являются полезными источниками информации для биологических наук. Улучшение качества геномных баз данных является непрерывным процессом, для этого используются как экспериментальные (Misra *et al*., 2002), так и теоретические подходы (Natale *et al*., 2000).

Матрицы присутствия/отсутствия генов. Одной из важных концепций базы данных кластеров ортологичных групп генов (**КОГ**ов) является филетический (филогенетический) паттерн, который представляет собой вектор представленности (присутствие/отсутствие) каждого организма в данном **КОГ**е (Tatusov *et al*., 1997; Koonin, Galperin, 2002). **КОГ**и имеют разнообразные филетические паттерны. Интересно, что лишь небольшая часть **КОГ**ов (~ 1 %) встречается во всех исследованных видах, все остальные **КОГ**и имеют представительство только в некоторых геномах. Сравнение филетических паттернов успешно применяется для предсказания функций генов (Koonin, Galperin, 2002). Филетический паттерн **КОГ**а может быть представлен как последовательность единиц (присутствие вида) и нулей (отсутствие вида) (табл. 1). Такие данные являются удобным объектом для методов максимальной парсимонии, которые исходно были разработаны для анализа именно этого типа данных.

Филогенетическое дерево может быть построено с помощью Долло-парсимонии (см. Введение), которая использует матрицу присутствия/отсутствия (табл. 2) в качестве исходных данных. Долло-парсимония может быть использована не только для построения деревьев, но и для восстановления сценария эволюционных событий (потери и приобретения генов) для уже существующего филогенетического дерева путем картирования событий на ветви и корень данного дерева. Такие сценарии, включающие в себя восстановление предковых геномов во всех точках ветвления и в корне филогенетического дерева, имеют большое значение для понимания закономерностей эволюции на геномном уровне. В данной работе мы использовали наши собственные программы, а также программы PAUP* и DOLLOP (пакет PHYLIP) (http://evolution.genetics.washington.edu/phylip/coftware.html).

Эволюция состава генов в эукариотических организмах. Позиция круглых червей, членистоногих и позвоночных в филогении животных остается неопределенной (Hedges, 2002). Традиционно считалось, что членистоногие и позвоночные принадлежат одной группе животных, называемых целоматами. Круглые черви, у которых отсутствует целом, являются внешней группой, которая разошлась с целоматами до расхождения членистоногих и позвоночных (Raff, 1996). Однако в настоящее время доминирует гипотеза о том, что как членистоногие,

Таблица 2

Основные методы филогенетического анализа

Метод	Краткое описание
Методы, основанные на матрице расстояний (distance methods)	Попарные эволюционные расстояния (например, дивергенция) подсчитываются для всех пар анализируемых последовательностей. Филогенетические деревья строятся на основе анализа матриц попарных расстояний. Наиболее популярный метод – метод ближайших соседей (neighbor-joining method).
Методы максимальной парсимонии (maximum parsimony methods)	Каждая позиция исследуемого набора выравненных последовательностей рассматривается отдельно. Метод парсимонии заключается в поиске филогенетического дерева, которое требует минимального числа мутационных событий, объясняющих отличия между исследуемыми последовательностями.
Методы максимального правдоподобия (maximum likelihood methods)	Значения функции правдоподобия подсчитываются для каждой комбинации нуклеотидов (аминокислот) в каждой позиции исследуемых последовательностей для всех возможных топологий филогенетических деревьев, выбирается дерево с максимальным значением функции правдоподобия.

Сводка программ филогенетического анализа доступна в Интернете: http://evolution.genetics.washington.edu/phylip/software.html

так и круглые черви принадлежат к группе линяющих животных экдисозоа (Aguinaldo *et al*., 1997; Giribet *et al*., 2000; Peterson, Eernisse, 2001). Полные геномы нематоды, дрозофилы и человека дают возможность тестирования этих эволюционных гипотез на обширных молекулярных данных (Mushegian *et al*., 1998; Blair *et al*., 2002; Wolf *et al*., 2004). Мы исследовали этот вопрос, используя как традиционные методы филогенетического анализа аминокислотных последовательностей (табл. 2), так и паттерны присутствия/отсутствия ортологичных групп генов (**КОГ**ов) для реконструкции филогенетических деревьев. Для исследования этих паттернов был использован метод Долло-парсимонии. Реконструированное филогенетическое дерево поддерживает «целоматную» гипотезу с большими значениями статистической значимости (рис. 1). «Целоматная» гипотеза поддерживается также другими методами филогенетического анализа полных геномов (Wolf *et al*., 2004). Необходимо отметить, что построенное филогенетическое дерево объединяет животных и растения в одну группу (рис. 1), что противоречит общепринятой точке зрения. Это может быть связано с массовыми потерями генов у одноклеточных грибов (Aravind *et al*., 2000), такие явления часто вызывают ошибки при построении филогенетических деревьев (Soltis *et al*., 2004). Несоответствие построенного филогенетического дерева наиболее общепринятому варианту дерева видов, скорее всего, является результатом такой ошибки.

Для того чтобы подробно изучить вопрос о динамике генов в ходе эволюции, мы реконструировали с помощью Долло-парсимонии сценарий приобретения и потери генов для общепринятого варианта эволюции эукариот (рис. 2). В полученном сценарии каждой ветви могут быть приписаны события возникновения и потери генов за исключением двух ветвей, одна из которых ведет к растениям, другая – к общему предку животных и грибов (рис. 2). Можно видеть, что в линии грибов действительно произошли массовые потери генов. Для животных характерно возникновение большого

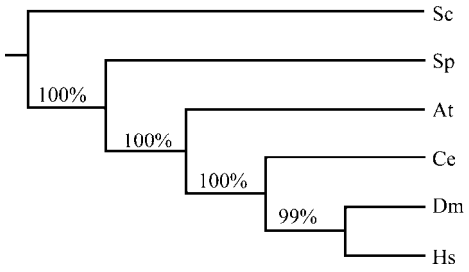

Рис. 1. Филогенетическое дерево эукариот, построенное с помощью Долло-парсимонии на основе матрицы присутствия/отсутствия ортологичных групп генов (**КОГ**ов) в 7 эукариотических организмах.

Сокращения названий организмов см. табл. 1. Значения статистической поддержки, полученные с помощью процедуры бутстрепа (Felsenstein, 1985), приведены для каждой внутренней ветви дерева.

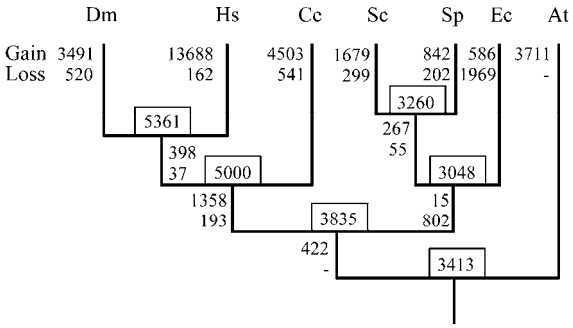

Рис. 2. Сценарий возникновения и потери генов в ходе эволюции эукариот, построенный с помощью Долло-парсимонии.

Числа в рамках соответствуют числу ортологичных групп генов (**КОГ**ов) в предсказанных предках существующих организмов. Числа рядом с ветвями дерева показывают число потерь (loss – нижнее число) и возникновений (gain – верхнее число) генов; дефис означает, что число потерь генов для данной ветви не может быть оценено. Сокращения названий организмов см. табл. 1.

числа новых генов в ветви, ведущей к общему предку всех животных, и в ветви, ведущей к человеку. В ветвях, ведущих к нематоде и дрозофиле, произошли потери значительного числа генов (рис. 2). Описываемый сценарий позволяет также реконструировать предковые геномы в узлах дерева. Долло-парсимония дает консервативные оценки числа генов в предковых геномах, однако даже этот подход оценивает число генов (**КОГ**ов) в общем предке животных, растений и грибов как 3413 генов. Более мягкие оценки (учитывающие, что некоторое число предковых генов было потеряно во всех или почти всех линиях за время последующей эволюции) могут дать еще большее (порядка 4000–6000) число генов в геноме общего предка животных, растений и грибов (Koonin *et al*., 2004).

Эволюция экзон/интронной структуры. Большинство эукариотических белок-кодирующих генов содержат интроны, которые вырезаются из пре-мРНК сплайсеосомой, сложным РНК-белковым комлексом, который консервативен у всех эукариот (Dacks, Doolittle, 2001). Позиции некоторых интронов совпадают в ортологичных генах растений и животных (Marchionni, Gilbert, 1986; Logsdon *et al*., 1995). Однако эукариотические организмы характеризуются большими отличиями по длине экзонов и частоте интронов, и локализация интронов в ортологичных генах не всегда совпадает даже для сравнительно близких видов (Logsdon *et al*., 1998). Основными механизмами эволюции интронов являются вставки и потери интронов (Rzhetsky *et al*., 1997; Logsdon *et al*., 1998). Сдвиг интронов на короткие расстояния (intron sliding) – это достаточно редкий феномен, который не может быть отнесен к разряду основных механизмов эволюции интронов (Rogozin *et al*., 2000). В целом закономерности эволюции интронов остаются неясными. Была высказана гипотеза, что доля совпадающих интронов уменьшается с возрастанием эволюционного расстояния, и поэтому интроны могут быть использованы как филогенетический признак (Ней, Кумар, 2004), однако эта гипотеза не была исследована методами филогенетического анализа.

Мы использовали базу данных **КОГ**ов для эволюционного анализа экзон/интронной структуры эукариотических генов. Для более детального анализа интронов к имеющимся 7 эукариотическим геномам мы добавили геномы комара *Anopheles gambiae* (Ag) и малярийного паразита *Plasmodium falciparum* (Pf). Микроспоридия *Encephalitozoon cuniculi* была исключена из анализа ввиду почти полного отсутствия интронов. Многие **КОГ**и содержат паралогичные гены, для таких **КОГ**ов среди всех

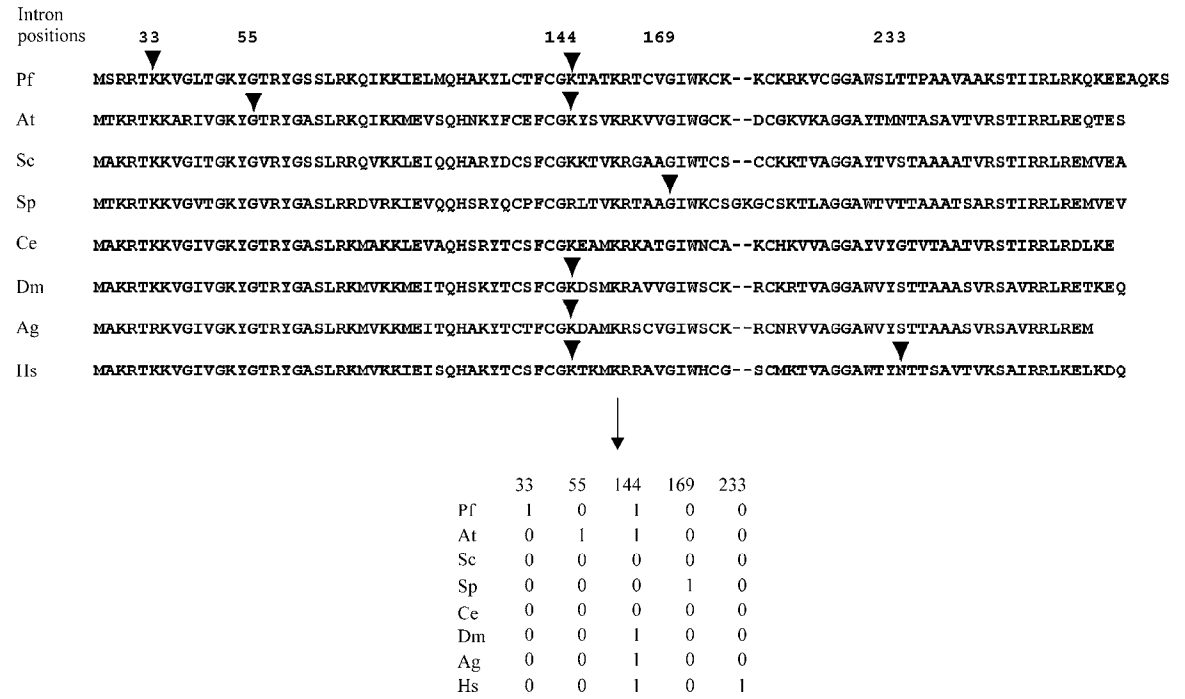

Рис. 3. Примеры консервативности/вариабельности позиций интронов в ортологичных генах (**КОГ**ах) и построения матрицы присутствия/отсутствия интронов.

Приведены данные для KOG0473 – рибосомальный белок L37. Позиции интронов показаны стрелками над последовательностями. «1» означает наличие данного интрона, «0» – его отсутствие. Сокращения: At – *Arabidopsis thaliana,* Ce – *Caenorhabditis elegans*, Dm – *Drosophila melanogaster*, Hs – *Homo sapiens*, Ag – *Anopheles gambiae*, Pf – *Plasmodium falciparum*, Sc – *Saccharomyces cerevisiae*, Sp – *Schizosaccharomyces pombe*.

паралогов из каждого организма был выбран один ген, имеющий наибольшее сходство с генами из других организмов (индексный ортолог) (Rogozin *et al*., 2003). Нами исследовались 684 **КОГ**а, которые были представлены во всех исследуемых геномах. Каждый **КОГ** был выравнен с помощью программы MAP (Huang, 1994), позиции интронов были картированы на полученные выравнивания (рис. 3). Для того чтобы интроны считались ортологичными, требовалось точное совпадение позиций интронов в разных организмах (Rogozin *et al*., 2003). Исследуемые ортологичные гены содержали 21434 интрона в 16577 позициях, 8028 интронов были консервативны в двух и более геномах. Большинство консервативных интронов было найдено в двух геномах, однако значительное число консервативных интронов было найдено в трех и более геномах (табл. 3). Моделирование эволюции интронов в анализируемом наборе выравниваний путем случайного «разбрасывания» позиций интронов отдельно в каждом выравнивании показало, что только ~ 1 % интронов, которые найдены в двух и более геномах, ожидаются по случайным причинам (табл. 3). Этот результат указывает на то, что подавляющая часть интронов, консервативных в двух и более видах, была унаследована от соответствующего общего предка этих видов. Анализ матрицы попарного сравнения присутствия/отсутствия интронов в ортологичных генах (табл. 4) выявил неожиданный эффект: число общих интронов не уменьшается монотонно с возрастанием эволюционного расстояния между видами. Например, человек имеет большее число общих интронов с растением *Arabidopsis thaliana* по сравнению с тремя другими животными, включенными в данный анализ. В консервативных участках выравнивания (этот подход дает более устойчивые оценки числа ортологичных интронов), 24 % интронов человека имели ортологов в растении (~ 27 % интронов растения имели ортологов в геноме человека)

Таблица 3

Консервативность позиций интронов в ортологичных генах (**КОГ**ах) из эукариотических геномов

Число геномов		1	2	3	4	5	6	7	8
Число интронов	Наблюдаемое*	13406	2047	719	275	104	25	1	0
	Ожидаемое	21368	33	0	0	0	0	0	0
	Ожидаемое, 10 %	20083	662	8	0	0	0	0	0

* Вероятность того, что наблюдаемая фракция интронов, которые консервативны в двух и более геномах, может возникнуть по случайным причинам << 0,0001 (как для случая всех позиций выравнивания, так и случая, когда только 10 % позиций выравнивания были разрешены для встройки интрона).

Таблица 4

Число консервативных позиций интронов в наборе ортологичных генов (**КОГ**ов): матрица попарных сравнений

	Pf*	Sc	Sp	At	Ce	Dm	Ag	Hs
Pf	**971**	2	48	137	50	46	54	145
Sc		**46**	7	3	3	3	4	6
Sp			**839**	209	98	114	111	308
At				**5589**	353	255	254	1148
Ce					**3465**	315	312	948
Dm						**1826**	787	802
Ag							**1768**	771
Hs								**6930**

На диагонали показано общее число исследуемых интронов в 684 **КОГ**ах из 8 геномов. * Сокращения названий организмов см. рис. 3.

по сравнению с ~ 12–17 % интронов человека, имеющих ортологов в дрозофиле, моските и нематоде (Rogozin *et al*., 2003). Интересно, что около 30 % интронов в генах малярийного паразита *Plasmodium falciparum* имеют ортологов в других геномах. Это означает также, что интроны могли возникнуть до расхождения малярийного паразита (простейшее) и остальных эукариот (~ 1,5 млрд лет) (Hedges, 2002; Rogozin *et al*., 2003).

Мы исследовали эволюционную динамику интронов, используя филогенетический анализ. Позиции интронов были представлены в виде матрицы присутствия/отсутствия интронов (рис. 3). Мы применили Долло-парсимонию для реконструкции сценария встройки и потери интронов для общепринятого варианта эволюции эукариот (рис. 4). В полученном сценарии каждой ветви могут быть приписаны события возникновения и потери генов за исключением двух ветвей, одна из которых ведет к малярийному паразиту, другая – к общему предку животных, грибов и растений (рис. 4). В

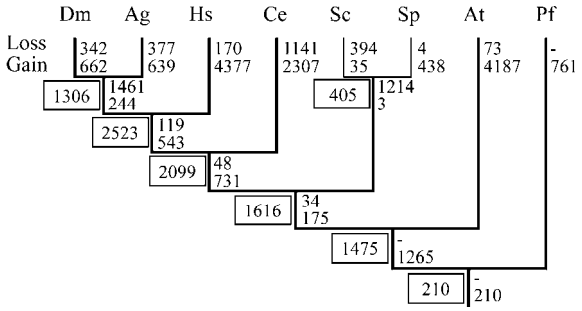

Рис. 4. Сценарий возникновения и потери интронов в ходе эволюции эукариот, построенный с помощью Долло-парсимонии.

Числа в рамках соответствуют числу интронов в предсказанных предках существующих организмов. Числа рядом с ветвями дерева показывают число потерь (loss – верхнее число) и вставок (gain – нижнее число) интронов; дефис означает, что число потерь интронов для данной ветви не может быть оценено. Сокращения названий организмов см. рис. 3.

полученном сценарии общий предок животных, грибов и растений имеет большое число интронов, а в ветви, ведущей к общему предку дрожжей, наблюдаются массовые потери интронов (рис. 4). Выявляются контрастные отличия в относительных скоростях встройки и потери интронов в разных линиях животных. В частности, в ветви, ведущей к человеку, наблюдается явное смещение в сторону встроек интронов, в то время как в ветви, ведущей к членистоногим, наблюдается противоположный эффект (рис. 4). Удивительный факт, что человек имеет большее число общих интронов с растением *Arabidopsis thaliana* по сравнению с тремя другими животными, включенными в данный анализ, объясняется массовыми потерями интронов в ветвях, ведущих к членистоногим и нематодам (рис. 4). Таким образом, эволюция интронов неравномерна, и наш недавний анализ интронов в паралогичных генах (Babenko *et al*., 2004) подтверждает, что эпохи активной встройки и потери интронов сменяются периодами относительной стабильности экзон/интронной структуры генов. Наличие большой фракции интронов, общих у малярийного паразита и других эукариот (~ 30 %), совместимо с гипотезой возникновения большого числа интронов на ранних стадиях эволюции эукариот.

Условия применимости Долло-парсимонии для изучения эволюции геномов

Долло-парсимония является полезным инструментом для изучения эволюции на геномном уровне. Однако необходимо помнить, что адекватное применение данного подхода требует соблюдения основного принципа Долло-парсимонии: возникновение производного состояния 1 из примитивного состояния 0 ($0 \rightarrow 1$) разрешается только один раз. В прокариотах горизонтальный перенос генов между различными линиями – это один из важных механизмов эволюции, поэтому повторное появление гена после его потери может происходить достаточно часто. В такой ситуации не следует применять Долло-парсимонию.

Не вызывает сомнений, что недавние горизотальные переносы генов из прокариот в эукариоты или между разными линиями эукариот не очень часты или отсутствуют совсем (по крайней мере, это характерно для позвоночных). В этом случае повторное появление гена в дереве после того, как произошла его потеря, достаточно маловероятно, поэтому нет сомнений, что Долло-парсимония является адекватным методом анализа паттернов присутствия/отсутствия генов. Ошибки классификации эукариотических генов могут создавать определенные проблемы для такого рода филогенетического анализа, однако доля таких ошибок не столь велика, так как мы не пытаемся учесть количество членов паралогичных семейств в каждом геноме (среди членов мультигенных семейств могут часто встречаться псевдогены), а оперируем с более простой и устойчивой характеристикой – присутствием/отсутствием **КОГ**ов в исследуемых геномах.

Применимость Долло-парсимонии к эволюции интронов более проблематична. Известно, что вставки интронов (или фиксация интронов после вставки) преимущественно происходят в нуклеотидном контексте (A/C)AG|G, называемом протосплайс-сайтом (Dibb, Newman, 1989; Sverdlov *et al*., 2004a). Мы попытались учесть возможный эффект ограничения на общее число позиций, разрешенных для вставки интронов. Для этого мы повторили моделирование эволюции интро-

нов, разрешая встройку интронов только в 10 % позиций выравнивания (табл. 3). Это привело к увеличению ожидаемого числа интронов, консервативных в двух и более видах (табл. 3). Однако нет никаких оснований полагать, что протосплайс-сайты консервативны в ходе эволюции, поэтому ограничение на 10 % позиций может быть явно завышенным. Это подтверждается исследованием экзон/интронной структуры генов, которые произошли в результате древних дупликаций (Cho, Doolittle, 1997). Был выявлен только один случай независимой встройки интронов в одну и ту же позицию древних паралогов, что составляет < 1 % интронов (всего было исследовано 237 вставок интронов) (Cho, Doolittle, 1997). Этот результат подтверждает, что независимая повторная встройка интронов в ортологичные позиции генов – это чрезвычайно редкое явление и, следовательно, Долло-парсимония является адекватным методом анализа эволюции интронов.

В настоящее время мы исследуем более сложные модели эволюции эукариотических геномов, в частности, баесовские модели (Huelsenbeck *et al*., 2001). Однако сложные модели эволюции могут создавать дополнительные проблемы для филогенетического анализа как в случае ошибок с выбором модели потери/приобретения генов, так и при оценке параметров этих моделей на ограниченных объемах данных в условиях явной неравномерности эволюции генного репертуара и экзон/интронной структуры (Suzuki *et al*., 2002). Применимость этих методов требует дополнительных исследований.

Заключение

Проведенное исследование выявило высокую степень изменчивости генного репертуара в геномах эукариот. Массовые потери генов в некоторых линиях являются характерной чертой эволюции эукариот. Тем не менее выявляется большое число генов, которые встречаются во всех исследованных геномах. Сходная картина наблюдается для эволюции экзон/интронной структуры эукариотических генов. Полученные результаты позволяют предположить, что некоторые интроны существуют длительное время. Вопрос функционального значения консервативности таких древних интронов представляет большой интерес. К настоящему времени существует целый ряд гипотез о функциях интронов (Соловьев, Колчанов, 1985; Fedorova, Fedorov, 2003). Возможная функциональность некоторых интронов совместима с наблюдениями о неравномерности вставок и потерь интронов по длине генов, хотя мы не можем исключить, что эта неравномерность связана с механизмами встройки и потерь интронов (Sverdlov *et al*., 2004b). Сравнительное исследование эволюционной консервативности древних интронов с теми интронами, которые были потеряны/приобретены относительно недавно (< 100 млн лет), может быть полезно для понимания функций интронов. К сожалению, данное исследование сдерживается отсутствием пары близких видов, в которых, тем не менее, наблюдается достаточно высокая изменчивость экзон/интронной структуры ортологичных генов.

В целом проведенные исследования выявили высокую изменчивость как репертуара генов, так и экзон/интронной структуры ортологичных генов. Некоторые организмы находятся, по всей видимости, под давлением отбора на уменьшение размера генома, это относится к дрожжам и в меньшей степени к насекомым и нематодам. Другие группы эукариот, такие, как позвоночные и высшие растения, не испытывают такого давления отбора и сохраняют большой генный репертуар и более сложную организацию экзон/интронной структуры генов.

Литература

Ней М., Кумар С. Молекулярная эволюция и филогенетика. Киев: КВІЦ, 2004.

Ратнер В.А., Жарких А.А., Колчанов Н.А., Родин С.Н., Соловьев В.В., Шамин В.В. Проблемы теории молекулярной эволюции. Новосибирск: Наука, 1985а.

Ратнер В.А., Омельянчук Л.В., Жарких А.А., Колчанов Н.А. Теоретический анализ структурных характеристик и эволюции транспортных РНК // Журн. общ. биологии. 1985б. Т. 46. С. 732–742.

Соловьев В.В., Колчанов Н.А. Экзон-интронная структура эукариотических генов может быть

связана с нуклеосомной организацией хроматина и регуляцией экспрессии генов // Докл. АН СССР. 1985. Т. 248. С. 232–237.

Aguinaldo A.M., Turbeville J.M., Linford L.S., Rivera M.C., Garey J.R. *et al*. Evidence for a clade of nematodes, arthropods and other moulting animals // Nature. 1997. V. 387. P. 489–493.

Aravind L., Watanabe H., Lipman D.J., Koonin E.V. Lineage-specific loss and divergence of functionally linked genes in eukaryotes // Proc. Natl. Acad. Sci. USA. 2000. V. 97. P. 11319–11324.

Babenko V.N., Rogozin I.B., Mekhedov S.L., Koonin E.V. Prevalence of intron gain over intron loss in the evolution of paralogous gene families // Nucl. Acids Res. 2004. V. 32. P. 3724–3733.

Blair J.E., Ikeo K., Gojobori T., Hedges S.B. The evolutionary position of nematodes // BMC Evol. Biol. 2002. V. 2. P. 7.

Cho G., Doolittle R.F. Intron distribution in ancient paralogs supports random insertion and not random loss // J. Mol. Evol. 1997. V. 44. P. 573–584.

Dacks J.B., Doolittle W.F. Reconstructing/ deconstructing the earliest eukaryotes: how comparative genomics can help // Cell. 2001. V. 107. P. 419–425.

DeBry R.W., Slade N.A. Cladistic analysis of restriction endonuclease cleavage maps within a maximum-likelihood framework // Syst. Zool. 1985. V. 34. P. 21–34.

Dibb N.J., Newman A.J. Evidence that introns arose at proto-splice sites // EMBO J. 1989. V. 8. P. 2015–2021.

Dollo L. Le lois de l'evolution // Bulletin de la Societe Belge de Geologie de Paleontologie et d'Hydrologie. 1893. V. 7. P. 164–167.

Doolittle W.F. Lateral genomics // Trends Cell. Biol. 1999. V. 9. P. M5–M8.

Farris J.S. Phylogenetic analysis under Dollo's Law // Syst. Zool. 1977. V. 26. P. 77–88.

Fedorova L., Fedorov A. Introns in gene evolution // Genetica. 2003. V. 118. P. 123–131.

Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrep // Evolution. 1985. V. 39. P. 783–791.

Fitch W.M. Distinguishing homologous from analogous proteins // Syst. Zool. 1970. V. 19. P. 99–106.

Giribet G., Distel D.L., Polz M., Sterrer W., Wheeler W.C. Triploblastic relationships with emphasis on the acoelomates and the position of *Gnathostomulida*, *Cycliophora*, *Plathelminthes*, and *Chaetognatha*: a combined approach of 18S rDNA sequences and morphology // Syst. Biol. 2000. V. 49. P. 539–562.

Gogarten J.P., Doolittle W.F., Lawrence J.G. Prokaryotic evolution in light of gene transfer // Mol. Biol. Evol. 2002. V. 19. P. 2226–2238.

Hedges S.B. The origin and evolution of model organisms // Nat. Rev. Genet. 2002. V. 3. P. 838–849.

Huang X. On global sequence alignment // Comput. Appl. Biosci. 1994. V.10. P. 227–235.

Huelsenbeck J.P., Ronquist F., Nielsen R., Bollback J.P. Bayesian inference of phylogeny and its impact on evolutionary biology // Science. 2001. V. 294. P. 2310–2314.

Katinka M.D., Duprat S., Cornillot E., Metenier G., Thomarat F. *et al*. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi* // Nature. 2001. V. 414. P. 450–453.

Koonin E.V., Aravind L., Kondrashov A.S. The impact of comparative genomics on our understanding of evolution // Cell. 2000. V. 101. P. 573–576.

Koonin E.V., Makarova K.S., Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification // Annu. Rev. Microbiol. 2001. V. 55. P. 709–742.

Koonin E.V., Galperin M.Y. Sequence – Evolution – Function. Computational approaches in comparative genomics. N.Y.: Kluwer Acad. Publ., 2002. 313 p.

Koonin E.V., Fedorova N.D., Jackson J.D., Jacobs A.R., Krylov D.M. *et al*. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes // Genome Biol. 2004. V. 5. P. R7.

Korostyshevsky M.A., Schtabnoy M.R., Ratner V.A. On some principles of evolution viewed as a stochastic process // J. Theor. Biol. 1974. V. 48. P. 85–103.

Logsdon J.M., Tyshenko M.G., Dixon C., D-Jafari J., Walker V.K. *et al*. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory // Proc. Natl Acad. Sci. USA. 1995. V. 92. P. 8507–8511.

Logsdon J.M., Stoltzfus A., Doolittle W.F. Molecular evolution: recent cases of spliceosomal intron gain? // Curr. Biol. 1998. V. 8. P. R560–R563.

Marchionni M., Gilbert W. The triosephosphate isomerase gene from maize: introns antedate the plant-animal divergence // Cell. 1986. V. 46. P. 133–141.

Misra S., Crosby M.A., Mungall C.J., Matthews B.B., Campbell K.S. *et al*. Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review // Genome Biol. 2002. V. 3. P. R0083.

Moran N.A. Microbial minimalism: genome reduction in bacterial pathogens // Cell. 2002. V. 108. P. 583–586.

Mushegian A.R., Garey J.R., Martin J., Liu L.X. Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes // Genome Res. 1998. V. 8. P. 590–598.

Nakamura Y., Itoh T., Matsuda H., Gojobori T. Biased biological functions of horizontally transferred genes in prokaryotic genomes // Nat. Genet. 2004. V. 36. P. 760–766.

Natale D.A., Galperin M.Y., Tatusov R.L., Koonin E.V. Using the COG database to improve gene recognition in complete genomes // Genetica. 2000. V. 108. P. 9–17.

Ochman H., Lawrence J.G., Groismn E.A. Lateral gene transfer and the nature of bacterial innovation // Nature. 2000. V. 405. P. 299–304.

Peterson K.J., Eernisse D.J. Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences // Evol. Dev. 2001. V. 3. P. 170–205.

Raff R.A. The shape of life: genes, development, and the evolution of animal form. Chicago: University of Chicago Press, 1996.

Rogozin I.B., Lyons-Weiler J., Koonin E.V. Intron sliding in conserved gene families // Trends Genet. 2000. V.16. P. 430–432.

Rogozin I.B., Wolf Y.I., Sorokin A.V., Mirkin B.G., Koonin E.V. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution // Curr. Biol. 2003. V. 13. P. 1512–1517.

Rzhetsky A., Ayala F.J., Hsu L.C., Chang C., Yoshida A. Exon/intron structure of aldehyde dehydrogenase genes supports the «introns-late» theory // Proc. Natl Acad. Sci. USA. 1997. V. 94. P. 6820–6825.

Snel B., Bork P., Huynen M.A. Genomes in flux: the evolution of archaeal and proteobacterial gene content // Genome Res. 2002. V. 12. P. 17–25.

Soltis D.E., Albert V.A., Savolainen V., Hilu K., Qiu Y.L. *et al*. Genome-scale data, angiosperm relationships, and «ending incongruence»: a cautionary tale in phylogenetics // Trends Plant. Sci. 2004. V. 9. P. 477–483.

Sonnhammer E.L., Koonin E.V. Orthology, paralogy and proposed classification for paralog subtypes // Trends Genet. 2002. V. 18. P. 619–620.

Sverdlov A.V., Rogozin I.B., Babenko V.N., Koonin E.V. Reconstruction of ancestral protosplice sites // Curr. Biol. 2004a. V. 14. P. 1505–1508.

Sverdlov A.V., Babenko V.N., Rogozin I.B., Koonin E.V. Preferential loss and gain of introns in 3' portions of genes suggests a reverse-transcription mechanism of intron insertion // Gene. 2004b. V. 338. P. 85–91.

Suzuki Y., Glazko G.V., Nei M. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics // Proc. Natl Acad. Sci. USA. 2002. V. 99. P. 16138–16143.

Tatusov R.L., Koonin E.V., Lipman D.J. A genomic perspective on protein families // Science. 1997. V. 278. P. 631–637.

Tatusov R.L., Mushegian A.R., Bork P., Brown N.P., Hayes W.S. *et al*. Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with *Escherichia coli* // Curr. Biol. 1996. V. 6. P. 279–291.

Wolf Y.I., Rogozin I.B., Koonin E.V. Coelomata and not ecdysozoa: evidence from genome-wide phylogenetic analysis // Genome Res. 2004. V. 14. P. 29–36.