Труды Зоологического института РАН
Приложение № 2, 2013, с. 39–52

УДК 575

ПРИНЦИПЫ РЕКОНСТРУКЦИИ ФИЛОГЕНЕЗОВ: ПРИЗНАКИ, МОДЕЛИ ЭВОЛЮЦИИ И МЕТОДЫ ФИЛОГЕНЕТИЧЕСКОГО АНАЛИЗА

В.А. Лухтанов

Зоологический институт Российской академии наук, Университетская наб. 1, 199034 Санкт-Петербург, Россия; Санкт-Петербургский государственный университет, Университетская наб. 7/9, 199034 Санкт-Петербург, Россия; e-mail: lukhtanov@mail.ru

РЕЗЮМЕ

Предлагаемый вниманию читателя краткий обзор, ни в коей мере не претендуя на полноту и глубину анализа, направлен на то, чтобы дать представление о базовых принципах, на которых основаны современные методы реконструкции филогенезов. В статье рассмотрены исторические связи между такими подходами как интуитивная геккелевская филогенетика, ручная хенниговская кладистика, метод максимальной парсимонии, метод максимального правдоподобия, Байесова филогенетика и методы, основанные на анализе генетических дистанций. Показаны основные преимущества и некоторые принципиальные ограничения каждого из этих методов.

Ключевые слова: Байесова филогенетика, геккелевская филогенетика, генетические дистанции, хенниговская кладистика, метод максимального правдоподобия, метод максимальной парсимонии, модель эволюции, признак, филогенез, филогенетический анализ

PHYLOGENETIC RECONSTRUCTIONS: CHARACTERS, MODELS OF EVOLUTION AND METHODS OF PHYLOGENETIC INFERENCE

V.A. Lukhtanov

Zoological Institute of the Russian Academy of Sciences, Universitetskaya Emb. 1, 199034 Saint Petersburg, Russia; Saint Petersburg State University, Universitetskaya Emb.7/9, 199034 Saint Petersburg, Russia; e-mail: lukhtanov@mail.ru

ABSTRACT

The paper considers some general principles of different methods of phylogeny reconstruction. It demonstrates historical relationships between such approaches as intuitive Haeckel's phylogenetics, Hennig's hand cladistics, method of maximum parsimony, method of maximum likelihood, Bayesian inference and distance methods. The advantages and shortcomings of these methods are briefly discussed.

Key words: Bayesian Inference, Haeckel's phylogenetics, genetic distances, Hennig's cladistics, method of maximum parsimony, method of maximum likelihood, model of evolution, character, phylogenesis, phylogenetic analysis

ВВЕДЕНИЕ

По определению Эрнста Геккеля, которое принимается и многими современными биологами, под филогенетикой следует понимать науку о путях, закономерностях и причинах исторического развития организмов. Нетрудно видеть, что при таком определении филогенетика по существу совпадает с эволюционной биологией (Татаринов 1984). На практике, однако, содержание филогенетики у́же, и она занимается лишь выявлением родственных связей между организмами и рекон-

струкцией путей исторического развития, изображая последние в виде филогенетических схем. Эти схемы могут быть получены посредством филогенетического анализа, в основе которого лежит идея генеалогической передачи признаков. Если распределения признаков, которые наблюдаются у организмов, как рецентных, так и ископаемых, унаследованы от общего предка, то анализируя эти распределения, теоретически можно получить филогенетические траектории отдельных признаков, а затем на этом основании восстановить эволюционные истории таксонов. Для решения этой задачи необходимо, чтобы в наличии были: (1) сами признаки, (2) модели эволюции этих признаков и (3) методы филогенетического анализа, т.е. обоснованные и систематизированные совокупности шагов и действий, которые необходимо предпринять, чтобы на основании изучения признаков и с учетом модели эволюции этих признаков осуществить филогенетическую реконструкцию.

ПРИЗНАКИ

Теоретически для получения филогенетической реконструкции можно использовать изменчивые признаки любой природы, например, экологические и поведенческие. На практике в настоящее время чаще всего используются морфологические, молекулярные и цитогенетические признаки. Главное требование, которое предъявляется к признакам в филогенетике, состоит в том, что признаки должны быть гомологичными. Негомологичные состояния организмов, даже если они похожи, не несут информацию об общем предке, и их нет смысла сравнивать и использовать для филогенетических целей. Конкретные пути и алгоритмы гомологизации признаков разной природы могут быть различными. Принципы выявления гомологии морфологических структур подробно описаны во многих статьях и руководствах (например: Haszprunar 1992, 1998; Клюге 2000; Павлинов 2005; Wiley and Lieberman 2011). Для молекулярных признаков процедура гомологизации включает различение между ортологичными и паралогичными локусами (Рис. 1А–В) с последующим выравниванием нуклеотидных последовательностей (Рис. 1С) (Page and Holmes 1998). Эта процедура сильно отличается от выявления гомологии морфологических признаков. Однако в любом случае гомологизация признаков – это первый и важнейший компонент любого филогенетического анализа. Об этом особенно важно помнить при исследовании молекулярных признаков. Слово гомология, как правило, отсутствует в статьях по молекулярной филогенетике, однако никогда нельзя забывать о том, что нуклеотидное или аминокислотное выравнивание – это не просто матрица признаков, но и гипотеза о гомологии нуклеотидных или аминокислотных сайтов.

ЭВОЛЮЦИОННАЯ ИСТОРИЯ ПРИЗНАКА И ЭВОЛЮЦИОННАЯ ИСТОРИЯ ТАКСОНА

Некоторую сложность при проведении филогенетической реконструкции представляет тот факт, что эволюционные траектории разных признаков, принадлежащие одной и той же группе таксонов, могут различаться. Это можно пояснить следующей схемой (Рис. 2). Допустим, что А и В – разные признаки. Их исходные (плезиоморфные) состояния – A_0 и B_0. Допустим далее, что реальная филогения группы такова, как она показана на схеме слева. A_1 и B_1 – апоморфии (показаны также горизонтальными штрихами). Реконструкции с использованием признаков А и В приводят к разным кладограммам, причем реконструкция по признаку А неправильная.

На рисунке 2В – показано все то же самое, что и на рис. 1А, но не в виде линий, а в виде изменения частот разных генотипов. Вначале мутации в генах, кодирующих признаки A_0 и B_0, приводят к возникновению аллелей A_1 и B_1. До момента физического разделения эволюционных линий идет свободное скрещивание, которое приводит к формированию четырех генотипов. Возникает стадия анцестрального полиморфизма по генам А и В. Затем в ходе сортировки аллелей (в этот процесс может быть вовлечен как генетический дрейф, так и естественный отбор) все три линии приходят к стадии мономорфизма, но генотипы в каждой из линий разные. Реконструкции филогенетической истории этих линий с использованием генов А и В приводят к разным кладограммам.

Таким образом, используя жаргон филогенетиков, можно сказать, что филогения признака – не обязательно то же самое, что филогения таксона (Nichols 2001). В то же время достаточно очевидно, что генеалогические линии разных признаков, относящихся к одной и той же эволюционной линии организмов, должны быть в среднем более или ме-

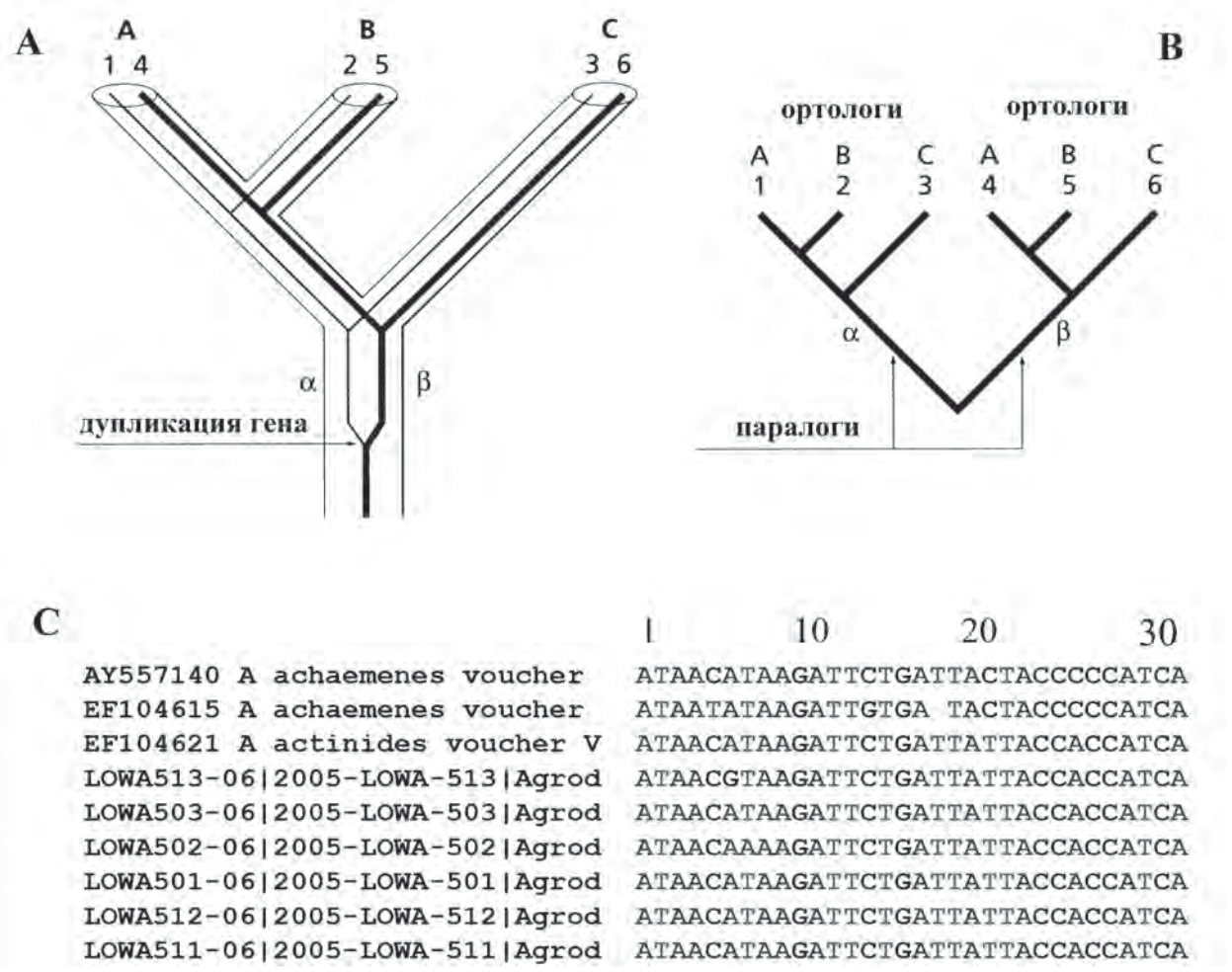

Рис. 1. Гомологизация молекулярных признаков. A – в результате дупликаций возникает пара похожих, но не гомологичных генов ***α*** и β. Они занимают разные локусы в геноме и эволюционируют независимо. B – неразличение настоящих гомологичных (=ортологичных) и структурно похожих негомологичных (=паралогичных) генов ведет к ошибочной реконструкции филогенеза (по: Page and Holmes 1998). C – нуклеотидное выравнивание является примером позиционной гомологии: нуклеотиды в пределах каждой из 32 показанных позиций (32 вертикальных столбцов) гомологичны. A, B и C – таксоны. 1, 2 и 3 – ортологичные варианты гена α. 4, 5 и 6 – ортологичные варианты гена β. α и β – паралоги.

Fig. 1. Homology of molecular characters. A – as a result of a duplication, a pair of similar (but not homologous) genes α and β arise. These genes have different position in genome and evolve independently. B – confusion of true homologous (=orthologous) and non-homological (=paralogous) genes results in false phylogeny reconstruction (after: Page and Holmes 1998). C – nucleotide alignment is an example of position homology: nucleotides are homologous within each of the 32 indicated positions. A, B and C are taxa. 1, 2 and 3 are orthologs of the gene α. 4, 5 and 6 are orthologs of the gene β. α and β are paralogs.

нее конгруэнтны (изоморфны), и реконструкция филогении по их совокупности возможна.

МОДЕЛИ ЭВОЛЮЦИИ ПРИЗНАКОВ

Второй важнейший компонент филогенетического анализа – это выбор модели эволюции признаков. Модели – это или словесные, или имеющие вид математических формул описания закономерностей эволюционных преобразований признаков. На ранних этапах развития филогенетики в качестве моделей часто использовались нечетко сформулированные (а иногда не сформулированные вообще) интуитивные представления о том, как могла идти эволюция изучаемых признаков.

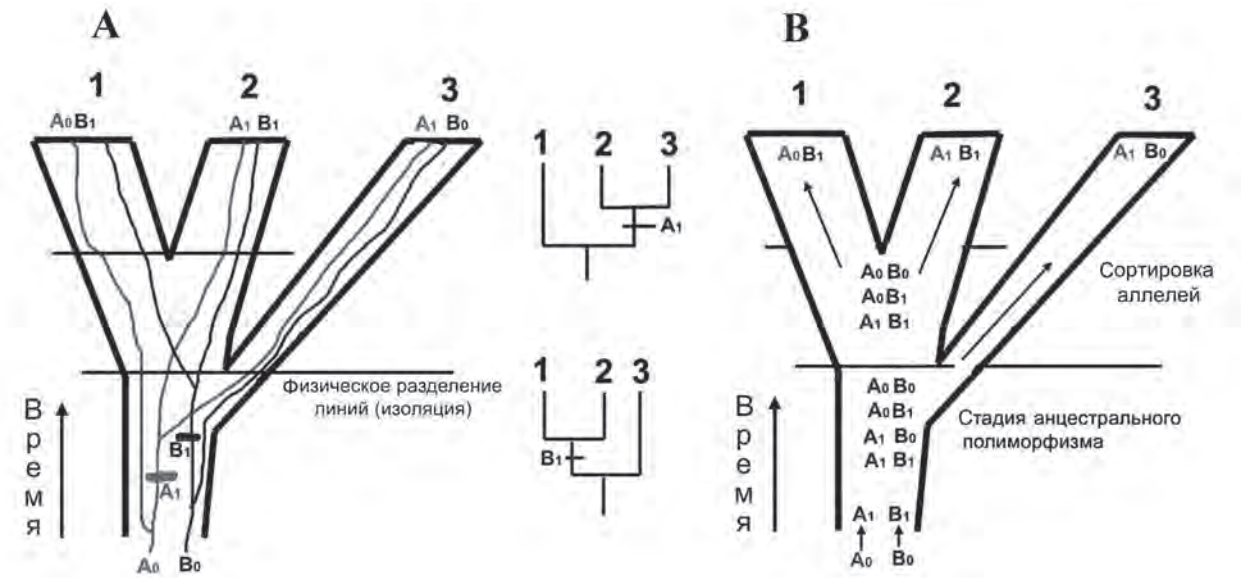

Рис. 2. Филогения признака – совсем не обязательно то же самое, что филогения таксона (по: Лухтанов и Кузнецова 2009) (см. объяснения в тексте).

Fig. 2. Phylogeny of a character and phylogeny of a taxon are not the same (after: Лухтанов и Кузнецова 2009) (see explanations in the text).

Обязательный компонент любой филогенетической модели – это топология, то есть геометрическая, обычно двухмерная схема, показывающая генеалогические связи между единицами филогенетического анализа. Часто топология задается в виде ветвящегося дерева, имеющего корень (Рис. 3А). Такая модель допускает передачу признака только от предка к потомку и не разрешает обмен признаками между разными филогенетическими линиями. Для представления филогении в случаях ретикулярной эволюции или в случаях, когда есть конфликт между признаками (см. Рис. 2), удобно использовать модель укорененной филогенетической сети (Рис. 3В) (Huson et al. 2010). Если направление передачи признака неизвестно, можно использовать модель неукорененного дерева (для случаев строго дивергентной эволюции) или модель неукорененной сети (для эволюции, включающей случаи ретикулогенеза).

Кроме того, у филогенетических моделей могут быть различные качественные и количественные параметры, выраженные словами, числами, соотношениями и вероятностями. Примеры таких параметров: признак, который был потерян организмом в ходе эволюции, не может снова появиться в своем исходном виде (модель Долло) (Wiley et al. 1991; Рогозин и др. 2005); или: эволюционные изменения признака полностью обратимы (модель Фитча-Вагнера) (Wiley et al. 1991; Felsenstein 2004).

МЕТОДЫ ФИЛОГЕНЕТИЧЕСКОГО АНАЛИЗА И ИНТУИТИВНАЯ ГЕККЕЛЕВСКАЯ ФИЛОГЕНЕТИКА

Третий компонент филогенетического анализа – это собственно построение дерева (или сети) с использованием определенного метода. Каждый метод должен иметь теоретическое обоснование возможности его применения, а также включать набор алгоритмов, которые позволяют с учетом выбранной модели трансформировать изученное распределение признаков в филогенетическую реконструкцию. Эти алгоритмы могут быть неявным, интуитивным, или они могут быть различным образом формализованы.

Эти три компонента анализа – признаки, модели и методы – одинаково важны для реконструкции филогении, но осознание этого пришло не сразу. **Классическая геккелевская филогенетика** преуспела лишь в первой из этих трех составляющих филогенетического анализа – в сравнительном изучении признаков. Многие

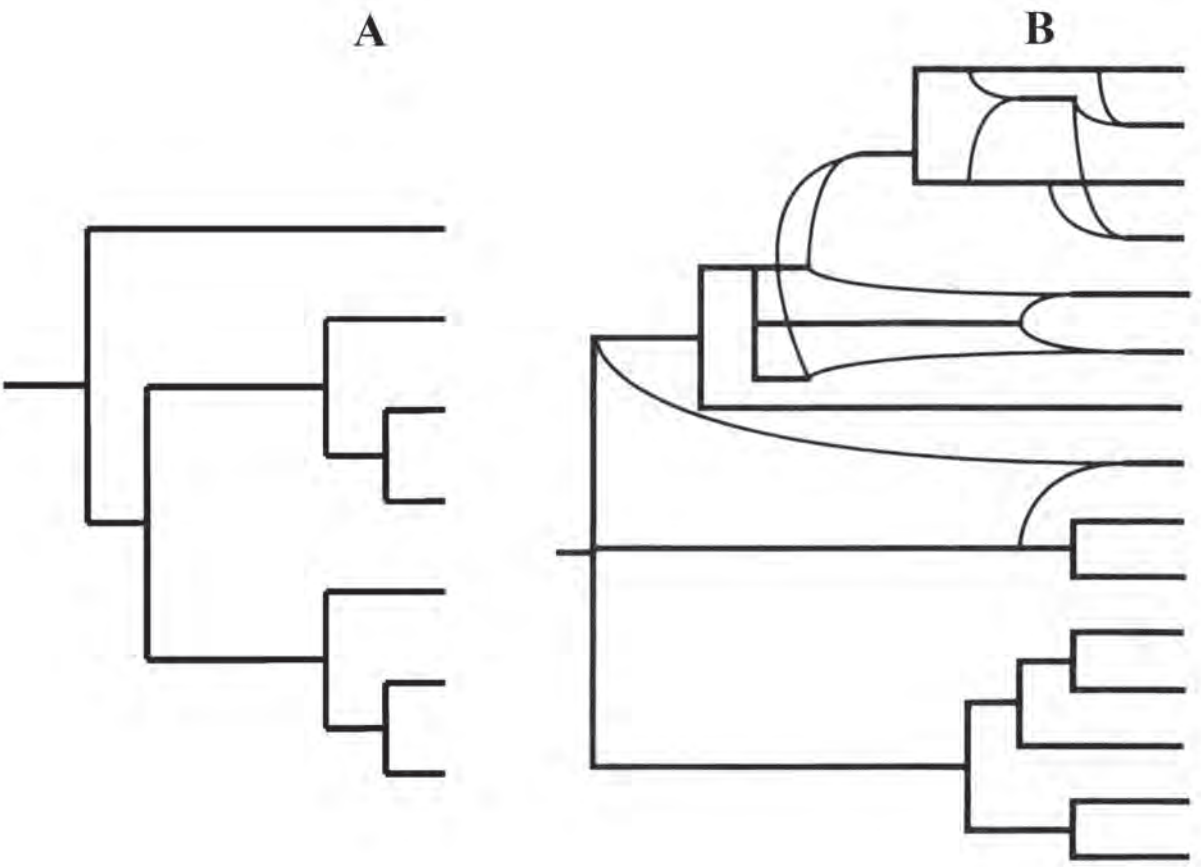

Рис. 3. Примеры топологии в виде укорененного дихотомически ветвящегося дерева (A) и в виде укорененной филогенетической сети (B).

Fig. 3. Examples of topology: A – rooted dichotomous tree; B – rooted phylogenetic network.

зоологи и ботаники второй половины XIX и начала XX века были прекрасными морфологами и оставили великолепные, не утратившие до настоящего времени сравнительно-морфологические исследования, легшие в основу филогенетических построений. Что касается собственно алгоритмов филогенетического анализа, то они были совсем не разработаны, и зачастую обоснование филогений ограничивалось словами: «*я предлагаю принять филогенетические отношения, представленные на рисунках*» (Кузнецов 1915).

ХЕННИГОВСКАЯ, ИЛИ РУЧНАЯ КЛАДИСТИКА

В середине XX века появляется **хенниговская кладистика** (Hennig 1950, 1965, 1966), и ситуация резко меняется. Хенниговская кладистика тщательный анализ признаков дополнила абсолютно четким алгоритмом перехода от признаков к филогениям. Этот алгоритм основан на последовательном выявлении соподчиненных монофилетических линий с использованием анализа синапоморфий (Рис. 4). При этом филогении выводятся на основании анализа относительно небольшого числа «надежных» синапоморфий. Основополагающим является принцип, согласно которому одна истинная синапоморфия может разрешить узел ветвления филогенетического дерева. Если возникает конфликт между потенциальными синапоморфиями, то основной путь его решения – переисследование материала, поиск и изучение дополнительных признаков и таксонов.

Этот подход теоретически и методологически весьма совершенен, что и предопределило его

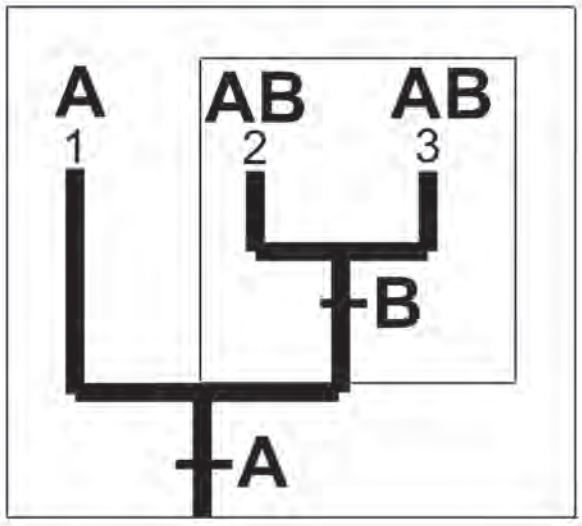

Рис. 4. Построение филогенетического дерева с использованием метода, основанного на анализе синапоморфий. Монофилетический таксон – группа, которая включает предка и всех его потомков. На рисунке показаны два соподчиненных (один вложен в другой) монофилетических таксона. А – это синапоморфия таксона (1+(2+3)), которая однозначно характеризует таксон (1+(2+3)). В – это синапоморфия таксона (2+3), которая однозначно характеризует таксон (2+3). Другие варианты монофилетических таксонов не существуют.

Fig. 4. Pylogeny reconstruction based on analysis of synapomorphies. Two nested monophyletic taxa are shown. A is a synapomorphy of the taxon (1+(2+3)) that characterizes this taxon unambiguously. B is a synapomorphy of the taxon (2+3) that characterizes this taxon unambiguously. Other variants of monophyletic taxa do not exist.

успех. Неудивительно, что он часто и успешно используется в современных исследованиях (см. например: Клюге 2000; Krell 2005). Однако у этой методологии есть серьезные недостатки. В частности Хенниг и его последователи фактически предложили отказаться от использования большой части гомологичных признаков, а именно от плезиоморфных и простых апоморфных (несинапоморфных) признаков. Отбрасывая такие признаки, мы теряем содержащуюся в них филогенетическую информацию. Это приводит к снижению разрешающей способности анализа и исчезновению информации об анагенетической составляющей эволюции. В итоге, реконструкции, получаемые с использованием кладистики по Хеннигу, как правило, чрезвычайно схематичны, что вызывает неудовлетворение у многих биологов.

Проблемой хенниговской кладистики, имеющей непосредственное отношение к систематике, является сложность выявления парафилетических таксонов. Дело в том, что принцип монофилии лежит в самой основе алгоритма построения дерева в хенниговской кладистике. Сипапоморфии однозначно определяют только монофилетические линии, а немонофилетические группы таксонов, например, парафилетические группировки не могут быть определены однозначно (Рис. 5). Таким образом, истинная причина отказа от парафилетических групп, который пропагандируют кладисты, лежит не в том, что парафилетические группы неестественны, а в том, что кладизм в принципе не умеет с ними работать. Фактически это является слабостью подхода, но кладисты нашли блестящий выход из положения, заявив, что парафилетические таксоны не реальны и так преуспели в распространении этого мифа, что многие биологи убеждены в том, что есть какой-то естественный биологический закон, запрещающий парафилетические группировки.

В действительности, парафилетические таксоны широко распространены на видовом уровне. Например, к парафилии приводит распространенное в природе перипатрическое видообразование, когда новые дочерние виды отпочковываются от исходного материнского вида, а последний не вымирает (Coyne and Orr 2004). Так как в этом случае филогенетическая линия, представленная материнским видом, не включает всех потомков, то она по определению является парафилетической. Де факто парафилетическими являются и почти все таксоны высокого ранга в том объеме, в каком мы их знаем, ведь почти все они не включают один или несколько вымерших видов, о существовании которых мы, скорее всего, никогда не узнаем.

МЕТОД МАКСИМАЛЬНОЙ ПАРСИМОНИИ

Решить некоторые из проблем традиционной хенниговской кладистики позволяет увеличение числа анализируемых признаков, и развитие филогенетики в этом направлении привело к появлению **метода максимальной парсимонии**, который вместо того, чтобы сконцентрироваться на изучении немногих ***«надежных»*** синапоморфий, пытается оперировать максимально возможным числом ***потенциальных*** синапоморфий. На этом пути сразу возникает другая сложность: противоречия между предполагаемыми **синапоморфиями**, которые на практике существуют почти

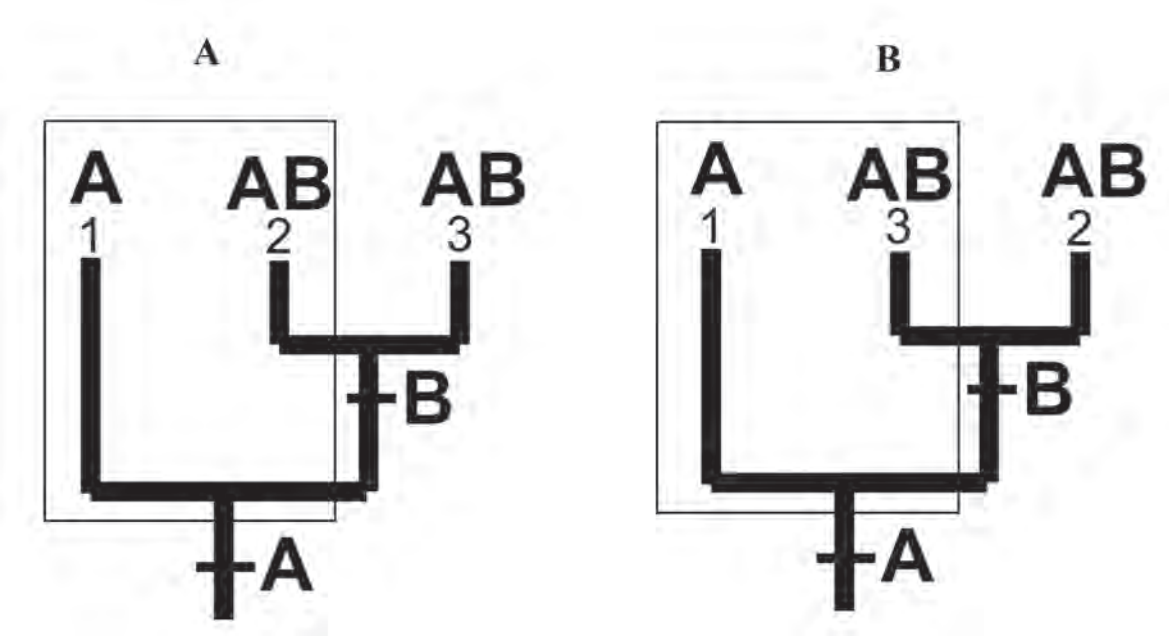

Рис. 5. Парафилетические таксоны, даже если они реальны, не могут быть распознаны с использованием метода синапоморфий. На филогении всегда существует несколько вариантов частично пересекающихся парафилетических группировок, и ни одна из них не имеет уникальной комбинации признаков. Например, для парафилетической группы (1+2) (рис. 5A) признак A не уникален, а признак B характеризует лишь часть таксона (1+2) и тоже не уникален. То же самое можно сказать в отношении парафилетической группы (1+3) (рис. 5B).

Fig. 5. Paraphyletic taxa, even if they are real, cat not be recognized by using the method of synapomorphies. There are two variants of partially overlapping paraphyletic groups, and none of them has a unique combination of characters. For example (fig. 5A), the character A is not unique for the group (1+2), and the character B can be found in only a part of the group (1+2). The same can be said concerning the paraphyletic group (1+3) (fig. 5B).

всегда, свидетельствуя о наличии **гомоплазий**[1], т.е. независимо приобретенных одинаковых состояний признаков.

Метод максимальной парсимонии решает это противоречие, исходя из предположения о том, что эволюция «экономна», и поэтому при выборе филогенетической гипотезы при прочих равных условиях предпочтительнее та, в которой число параллелизмов минимально (Павлинов 2005, с. 51). Этот критерий имеет теоретическое обоснование: если эволюционные события редки, то гипотеза, в которой их число минимизировано, может быть хорошим приближением к действительности (Page and Holmes 1998). Однако в общем виде критерий парсимонии несостоятелен, и было теоретически показано, что при некоторых условиях его использование приводит к ошибочным реконструкциям (Felsenstein 1978, 2004).

Так, к искажению результатов филогенетической реконструкции с использованием метода максимальной парсимонии приводит эффект притяжения длинных ветвей (long branch attraction), который особенно остро дает о себе знать при работе с молекулярными признаками. Если в пределах какой-либо филогении две или большее число ветвей эволюционировали быстрее, то в силу случайных причин они имеют шанс накопить больше гомоплазий. Поскольку одна из презумпций кладистического анализа – рассмотрение одинаковых состояний признаков в качестве синапоморфий, а не гомоплазий (Hennig 1966; Расницын 2005), то формальный анализ приведет к тому, что такие линии могут появиться на реконструкции как сестринские, даже если фактическая филогения была другой.

[1]Терминологические и эволюционные проблемы, связанные с использованием этого термина, обсуждены в работе Скотланда (Scotland 2011).

Это можно пояснить на следующем примере (Рис. 6). Допустим, филогения таксонов A, B, C и D известна (Рис. 6A). После расхождения линий в точках 0, 0_1 и 0^2, линия A приобрела новое состояние для одного признака (обозначено как 1), остальные признаки сохранили плезиоморфное состояние 0; линия B приобрела новое состояние для 6 признаков (обозначены как 2, 7,8, 9, 10, 12); линия C приобрела новое состояние для 7 признаков (обозначены как 3, 5, 6, 7, 11, 13, 14); линия D приобрела новое состояние для одного признака (обозначено как 4). Таким образом, линии B и C эволюционировали быстрее и приобрели новые состояния для большего числа признаков, чем линии A и D. Чем больше новых состояний признаков появилось в каждой из независимо эволюционирующих линий, тем больше вероятность, что хотя бы некоторые из них будут одинаковы. Так, при сравнении коротких ветвей A и D общих новых состояний признаков не обнаруживается.

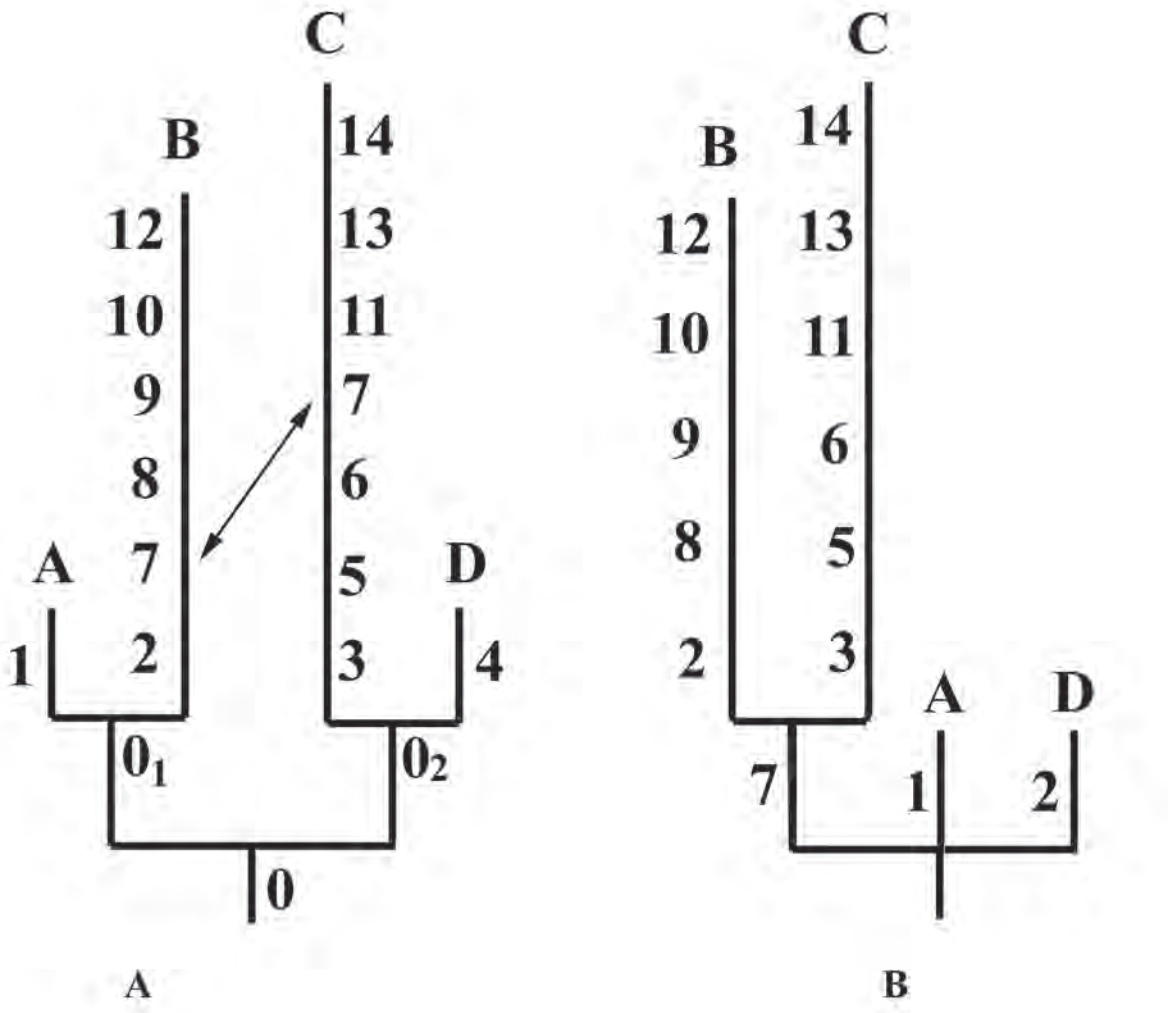

Рис. 6. Влияние эффекта притяжения длинных ветвей на результаты парсимониального филогенетического анализа таксонов A, B, C и D (по: Лухтанов, 2010). 0 – плезиоморфный признак, 1–14 – апоморфные признаки. A – реальная (истинная) филогения и распределение на ней признаков. B – ложная реконструкция филогении A, получаемая при проведении кладистического анализа с использованием метода максимальной парсимонии (см. объяснения в тексте).

Fig. 6. Long branch attraction affects the result of parsimony phylogenetic analysis of the taxa A, B, C and D (after: Лухтанов, 2010). 0 is a plesiomorphy. 1–14 are apomorphies. – A is a true phylogeny. B is a false reconstruction resulted from using the method of maximum parsimony (see explanations in the text).

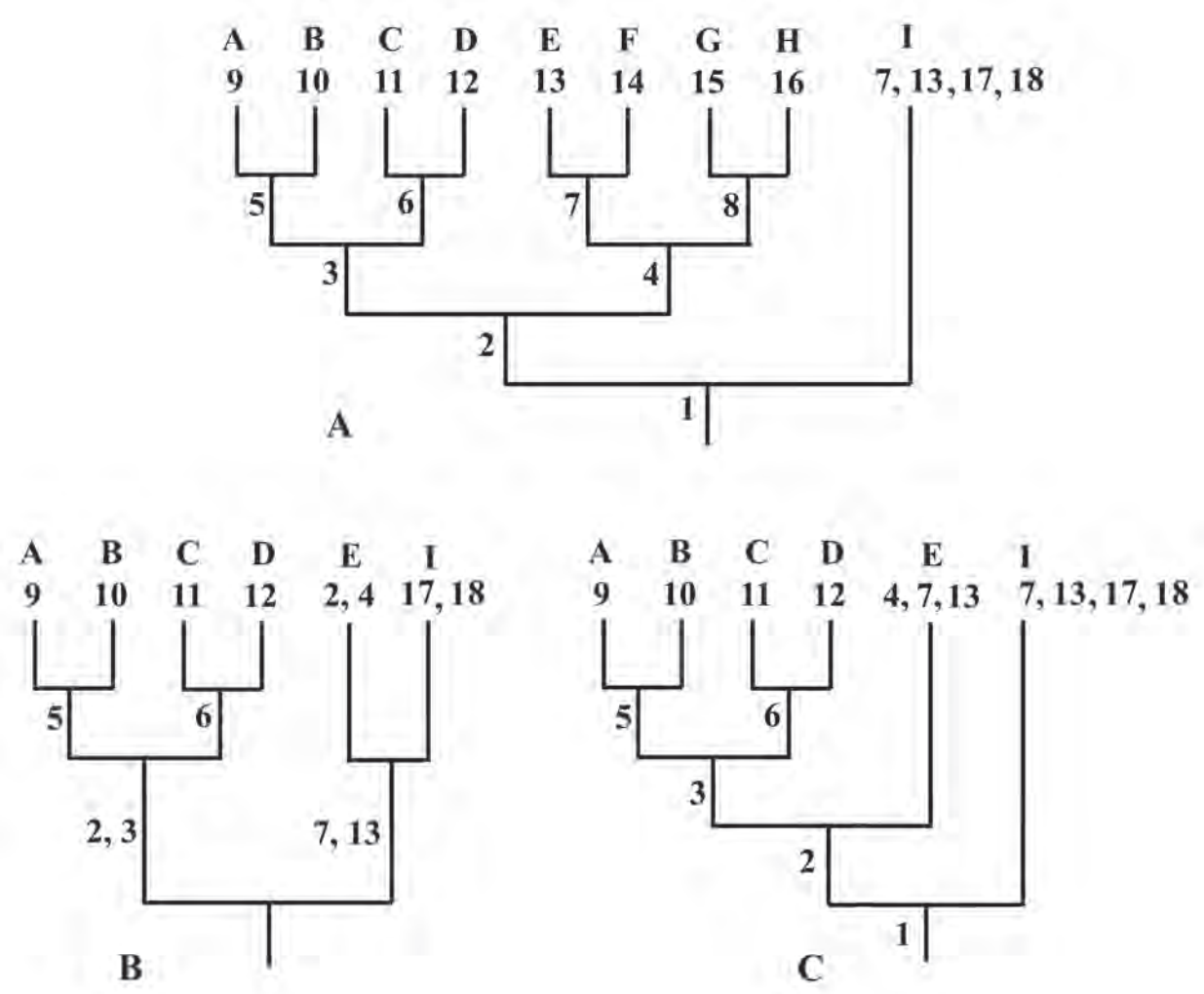

Рис. 7. Влияние неполноты выборки таксонов на результаты парсимониального кладистического анализа (см. объяснения в тексте) (по: Лухтанов, 2010).

Fig. 7. Influence of incomplete taxa sampling on results of parsimony cladistic analysis (after: Лухтанов, 2010) (see explanations in the text).

При сравнении длинных ветвей B и C выявляется новое общее состояние признака (7) (показано стрелкой). Если мы ничего не знаем о филогении и пытаемся ее реконструировать, исходя из признаков, то мы вынуждены рассматривать признак (7) как синапоморфию (хотя в действительности он является гомоплазией). В итоге мы вынуждены остановиться на филогенетической гипотезе, показанной на Рис. 6B и представляющей таксоны B и C как сестринские, хотя она и не верна.

Кроме того, к неправильной реконструкции может приводить неполнота выборки изучаемой группы. Это можно пояснить на следующем примере (Рис. 7). Допустим, филогения таксонов A, B, C, D, E, F, G, H и I известна (Рис. 7A). Скорость эволюции в разных линиях одинакова, поэтому все таксоны накопили равное число изменений состояний признаков, начиная от момента расхождения от общего предка. Арабские цифры у узлов ветвлений показывают синапоморфии соответствующих клад. Арабские цифры у вершин ветвей показывают аутапоморфии. Состояния 7 и 13 являются гомоплазиями для таксонов E и I. Несмотря на это, наличие синапоморфии 4 по-

зволяет в ходе анализа с использованием метода максимальной парсимонии получить правильную реконструкцию. Допустим далее, что таксоны F, G и H не участвуют в реконструкции (или вымерли, не оставив следов в палеонтологической летописи; или сохранились, но не изучены). В этом случае состояние 4 не может быть распознано как синапоморфия, и наоборот, состояния 7 и 13 должны быть интерпретированы как синапоморфии. В итоге использование метода максимальной парсимонии даст неверную реконструкцию, представленную на Рис. 7B, на которой ветви E и I появляются как сестринские таксоны, хотя они таковыми не являются. Признак 2 в линии E интерпретируется на этой реконструкции в качестве единственной гомоплазии. Правильная реконструкция для таксонов A, B, C, D, E и I показана на Рис. 7C, и она не является наиболее парсимониальной, так как содержит две гомоплазии (7 и 13).

Наличие этих недостатков не говорит о том, методом максимальной парсимонии не следует пользоваться. Его неоспоримым достоинством является то, что он основан на минимальных допущениях о закономерностях эволюции используемых признаков, и это важно для тех случаев, когда эти закономерности недостаточно изучены. Этот подход продолжает оставаться рабочим инструментом современной филогенетики, особенно в тех случаях, когда речь идет об анализе морфологических признаков. Его сторонники продолжают разрабатывать теоретические основы метода (Farris 2008) и писать новые программы анализа признаков (Goloboff 2008), однако пользоваться методом максимальной парсимонии (как, впрочем, и другими подходами) следует с осторожностью, имея в виду его ограничения.

МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

Принципиальным прорывом в филогенетике стала разработка во второй половине XX века метода **максимального правдоподобия (**Felsenstein 2004). Суть этого метода может быть сформулирована следующим образом. Если имеется информация о закономерностях эволюционных преобразований признаков (иными словами, если разработана модель эволюции признака) и известно распределение состояний признаков у изучаемых организмов, то можно рассчитать вероятности различных эволюционных траекторий, которые могли привести к современным формам, а затем выбрать наиболее вероятную из них.

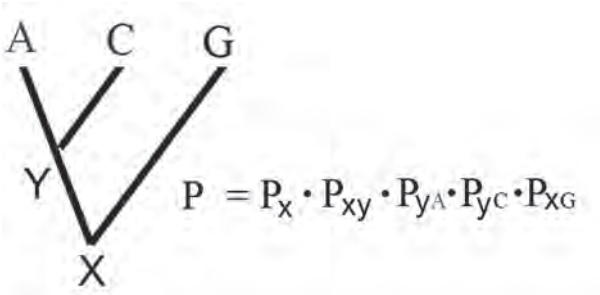

Рис. 8. Принцип реконструкции филогении с использованием метода максимального правдоподобия (см. объяснения в тексте).

Fig. 8. Principle of phylogeny reconstruction by using the method of maximum likelihood (see explanations in the text).

Например, для дерева, состоящего их трех таксонов (Рис. 8) , характеризующихся признаками A, C и G, вероятность наблюдения имеющихся состояний равна вероятности сочетания отдельных эволюционных событий, которые привели к имеющимся данным. Согласно теореме умножения вероятностей эта величина равна произведению вероятности того, что в основании дерева был признак x, умноженной на вероятность перехода признака x в признак y, умноженной на вероятность перехода признака x в признак G, умноженной на вероятность перехода признака y в признак A и умноженной на вероятность перехода признака y в признак C. Эта величина называется правдоподобием эволюционной гипотезы, и метод максимального правдоподобия ищет такое дерево, в контексте которого и в контексте имеющейся эволюционной модели получение наблюдаемого распределения признаков наиболее вероятно.

Таким образом, метод максимального правдоподобия не нуждается в теоретически несостоятельном принципе парсимонии как в критерии истинности филогенетической реконструкции. Совершенно на иных принципах, чем в традиционной кладистике, происходит и построение дерева: вместо анализа синапоморфий производится расчет вероятностей различных эволюционных трансформаций признаков. Преимущество этого подхода состоит в возможности использования для реконструкции филогений любых гомологичных признаков: не только синапоморфий, но и простых апоморфий и плезиоморфий. Так, важная состав-

ляющая этого метода при работе с молекулярными признаками – это учет так называемых инвариантных, т.е. константных сайтов, которые представляют собой не что иное, как плезиоморфии.

Метод максимального правдоподобия дает количественную информацию об анагенетической составляющей эволюции в виде дистанций между узлами ветвлений, что классическая кладистика по Хеннигу даже не пытается делать, а геккелевская филогенетика делает очень приблизительно. Кроме того, он гораздо эффективнее, чем методы, основанные на парсимонии, выявляет ретикулогенез – случаи интрогрессии и гибридизации. Наконец, хотя и не абсолютно иммунен, этот метод на порядок менее чувствителен к эффекту притяжения длинных ветвей.

Различия между более традиционным, основанным на парсимонии, и современными, основанным на вероятностях, подходами состоят не только в принципах, но и в конкретных особенностях работы и даже в терминологии. Для традиционного кладиста поиск синапоморфий, выявление монофилии и боязнь парафилии – существенная часть методологии, часть его работы. Для современных вероятностных методов это не так важно. Эти методы не нуждаются в понятиях «синапоморфия», «монофилия» и «парафилия», хотя термины, обозначающие эти понятия, могут быть полезны при обсуждении уже полученных результатов.

В качестве недостатка метода максимального правдоподобия можно рассматривать зависимость получаемых реконструкций от выбора модели эволюции признака. К счастью, этот подход устойчив даже к существенным отклонениям выбираемых моделей от оптимальных (Felsenstein 2004), однако выбор принципиально неправильной модели эволюции может привести к серьезным ошибкам. Поэтому сравнительный анализ признаков (в том числе молекулярных) с целью выявления их соответствия выбранным моделям является более важной частью анализа, чем формальный процесс получения дерева с помощью той или иной компьютерной программы.

БАЙЕСОВА ФИЛОГЕНЕТИКА (BAYESIAN INFERENCE)

При изучении многих вероятностных процессов часто бывает так, что полученных наблюдений недостаточно для выявления статистически значимых заключений о вероятностях, а в то же время процессы этого рода уже ранее кем-то изучались, и мы что-то о них знаем. Иначе говоря, у нас есть какие-то идеи, некоторые предварительные гипотезы (priors) в отношении вероятностей изучаемых процессов. В работе ученого эта ситуация является чрезвычайно обычной, так как фактов почти никогда не бывает слишком много, а работать с совсем неизученными явлениями приходится редко.

В такой ситуации для расчета вероятностей целесообразно использовать Байесову статистику, которая позволяет комбинировать информацию, извлекаемую из предварительных гипотез, с информацией, получаемую из опытов (наблюдений), для расчета так называемых постериорных вероятностей, т.е. априорных вероятностей, скорректированных с учетом проведенных эмпирических испытаний. Затем эти постериорные вероятности можно и нужно использовать для оценки значимости тестируемых гипотез.

Таким образом, Байесова статистика позволяет более эффективно находить вероятности в условиях недостатка информации. В принципе метод Байеса можно применять и в тех ситуациях, когда конкретные предварительные гипотезы отсутствуют. В этих случаях возможно (и часто это бывает полезно) использование неспецифических априорных вероятностей, которые предполагают, что разные состояния изучаемых параметров одинаково возможны. Однако этот метод становится гораздо более эффективным, если используемые априорные гипотезы специфичны и позволяют отсечь те значения, которые невозможны или маловероятны.

Байесов подход в филогенетике занимается поиском деревьев с наибольшим уровнем правдоподобия (и в этом отношении похож на метод максимального правдоподобия), однако для расчета вероятностей использует Байесову статистику (Ronquist, Huelsenbeck 2003), что дает ряд преимуществ:

1) использование предварительных гипотез в филогенетике фактически означает применение системы запретов и ограничений на определенные топологии дерева и/или определенные переходы признаков из одного состояния в другое. Это потенциально делает этот метод более мощным, чем метод максимального правдоподобия в его классическом виде;

2) в методе Байеса напрямую проверяется, в какой степени получаемая филогенетическая гипотеза соответствует эмпирическим данным, т.е. распределению признаков. В качестве меры соответствия выступает постериорная вероятность параметров получаемых деревьев. В методе максимального правдоподобия соответствие дерева данным напрямую не тестируется, вместо этого выбирается дерево, которое максимизирует степень соответствия наблюдаемых признаков выбранной модели их эволюции;

3) в результате анализа появляется не одно дерево, как в методе максимального правдоподобия, а набор наиболее правдоподобных деревьев, сравнение которых позволяет количественно оценивать уровень неопределенности получаемых топологий и длин ветвей.

Потенциальным недостатком Байесова подхода по сравнению с методом максимального правдоподобия в его классическом виде является то, что использование дополнительной информации в виде предварительных гипотез может быть как плюсом, так и минусом. Так, выбор принципиально неправильных априорных распределений может ухудшать результаты анализа. Однако важно отметить, что правильность этого выбора достаточно легко контролировать, сравнивая значения правдоподобий деревьев, ассоциированных с разными предварительными распределениями. В целом метод Байеса следует признать одним из наиболее эффективных подходов к решению филогенетических задач.

МЕТОДЫ, ОСНОВАННЫЕ НА АНАЛИЗЕ ГЕНЕТИЧЕСКИХ ДИСТАНЦИЙ

Эта группа включает два достаточно разных метода, называемых дистантными и применяемых только в «молекулярной» филогенетике. Один из них основан на ДНК–ДНК гибридизации, позволяющей установить степень тотального сходства сравниваемых геномов (в процентах) и использовать полученные величины для реконструкции филогенезов. Он имеет два неустранимых недостатка. Во-первых, он является фенетическим методом и несет с собой главную проблему фенетики, которая состоит в том, что последняя направлена на выявление сходства, а не родства.

Во-вторых, феномен так называемого С-парадокса (крайне неравномерного накопления некодирующих нуклеотидных повторов даже у близких видов) (Gregory 2001) приводит к тому, что филогенетическая интерпретация значений общего генетического сходства, как правило, затруднительна. Хотя в отдельных случаях этот метод позволяет получить очень красивые результаты (Caccone and Powell 1989), это скорее исключение, чем правило, и с начала 90-х годов XX в. он в филогенетике практически не используется.

Другой метод основан на том, что матрицы дискретных молекулярных признаков преобразуются в генетические дистанции. Построение деревьев затем производится на основании анализа полученных дистанций, а не признаков как таковых. Принципиальным недостатком этого подхода является то, что при преобразовании дискретных признаков в дистанции происходят исчезновение индивидуальности каждого признака и потеря филогенетической информации. Однако в отличие от ДНК–ДНК гибридизации метод не является чисто фенетическим (Page and Holmes 1989). Поскольку при расчете генетических дистанций используют модели молекулярной эволюции (те же самые, что применяются при анализе с помощью метода максимального правдоподобия и метода Байеса), в итоге получают так называемые скорректированные генетические дистанции, которые несут довольно высокий филогенетический сигнал (Felsenstein 2004).

Подходы, основанные на расчете дистанций, были очень популярны в 90-е годы прошлого века, поскольку программы, основанные на алгоритмах этих методов, отличаются быстродействием. Соответственно они не требовательны к мощности компьютеров. В настоящее время их значение сильно уменьшилось, однако они до сих пор достаточно широко применяются, особенно там, где требования к качеству анализа не очень высокие, а матрицы признаков чрезвычайно велики.

ЗАКЛЮЧЕНИЕ: ЕЩЕ РАЗ О МОДЕЛЯХ ЭВОЛЮЦИИ

Сравнивая перечисленные выше методы, нужно отметить, что все они основываются на моделях эволюции признаков, хотя содержание и форма этих моделей сильно различаются. Модели эволюции, которые в неявном виде возникают в головах эволюционных филогенетиков на основании изучения морфологических рядов, могут

быть очень сложными. Как правило, филогенетики-традиционалисты – это зоологи или ботаники, великолепно знающие свои группы, и все нюансы этого знания они учитывают при разработке филогений. Беда только в том, что эти модели так и остаются в голове, а не на бумаге, и в этом нет ничего удивительного: явное выражение моделей эволюции признаков в виде слов и тем более формул – дело очень нетривиальное.

Кладисты, использующие методологию парсимонии, впадают в другую крайность: схема эволюции, которую они применяют, основана на предположении о строго дихотомической дивергенции таксонов и модели экономной эволюции признаков. Очевидно, что эта схема слишком проста для того, чтобы быть универсальной. Модель эволюции, которую используют филогенетики, исповедующие ручной кладизм по Хеннигу, по необходимости также чрезмерно упрощена и описывается схемой, согласно которой строго дихотомическая дивергенция сопровождается накоплением синапоморфий, при этом предполагается, что синапоморфии возникают чаще, чем гомоплазии.

Современные подходы, основанных на методах максимального правдоподобия и Байеса, в какомто отношении являются золотой серединой между интуитивной геккелевской филогенетикой и традиционным парсимониальным кладизмом. Они используют такие модели эволюции, которые (1) достаточно адекватно описывают закономерности эволюционных трансформаций признаков, (2) могут быть эксплицитно описаны и формализованы в виде алгоритмов и формул, и (3) по возможности просты (но не чрезмерно упрощены), что позволяет их практическое использование в филогенетике. К настоящему моменту времени в деле разработки моделей эволюции признаков молекулярные биологи продвинулись значительно дальше, чем морфологи, по той причине, что закономерности молекулярной эволюции проще изучать. Кроме того, модели молекулярной эволюции легко формализуются, они хорошо изучены теоретически, а их соответствие реальности проверено на огромном фактическом материале (Page and Holmes 1998; Felsenstein 2004). Конкретные параметры этих моделей, необходимые для построения филогений, легко рассчитываются на основании сравнения частот нуклеотидных замен разного типа у изучаемых организмов. По этой причине метод максимального правдоподобия и метод Байеса пока широко применяются только для анализа молекулярных признаков. Однако эта ситуация, вероятно, временная, и первые попытки использования признаков морфологии в рамках современных параметрических методов филогенетического анализа уже сделаны (Lewis 2001; Ronquist and Huelsenbeck 2003; Nylander et al. 2004; Ronquist 2004; Müller and Reisz 2006).

В заключение следует еще раз отметить, что ни один из методов филогенетического анализа не является идеальным, однако преимущества и недостатки разных подходов не совпадают. Неудивительно поэтому, что многие филогенетики, особенно исследователи, работающие с молекулярными признаками, используют для получения реконструкций современную «классическую триаду» методов – парсимонию, правдоподобие и Байесов анализ. Как правило, эти три метода дают очень похожие филогенетические реконструкции. В то же время получение поддержанных статистическими тестами, но несовпадающих результатов свидетельствует о наличии систематических ошибок при проведении анализов, и это дает импульс для дальнейших исследований признаков и моделей их эволюции.

БЛАГОДАРНОСТИ

Автор благодарен организаторам конференции «Современные проблемы биологической систематики» за предоставленную возможность обмена мнениями с коллегами, а также А.В. Бочкову за неформальное обсуждение излагаемых вопросов и высказанные критические замечания. Работа выполнена при финансовой поддержке программ президиума РАН «Динамика и сохранение генофондов» и «Происхождение биосферы и эволюция гео-биологических систем» и Российского фонда фундаментальных исследований (гранты 11-04-01119, 11-04-00734, 11-04-00076 и 12-04-00490).

ЛИТЕРАТУРА

Клюге Н. Ю. 2000. Современная систематика насекомых. Принципы систематики живых организмов и общая систематика насекомых с классификацией первичнобескрылых и древнекрылых. Лань, Санкт-Петербург, 336 с.

Кузнецов Н. Я. 1915. Насекомые чешуекрылые (Insecta, Lepidoptera). Том 1. Введение. Danaidae (Pieridae + Leptalidae auct.). Выпуск 1. Петроград: Зоологический музей Императорской Академии Наук, 337 с.

Лухтанов В.А. 2010. От геккелевской филогенетики и генниговской кладистики к методу максимального правдоподобия: возможности и ограничения современных и традиционных подходов к реконструкции филогенезов. *Энтомологическое обозрение*, **89**(1): 133–149.

Лухтанов В. А. и Кузнецова В. Г. 2009. Молекулярно-генетические и цитогенетические подходы к проблемам видовой диагностики, систематики и филогенетики. *Журнал общей биологии*, **70**(5): 415–437.

Павлинов И. Я. 2005. Введение в современную филогенетику (кладогенетический аспект). Товарищество научных изданий КМК, Москва, 391 с.

Расницын А.П. 2005. Избранные труды по эволюционной биологии. Товарищество научных изданий КМК, Москва, 347 с.

Рогозин И.Б., Вульф Ю.И., Бабенко В.Н. и Кунин Е.В. 2005. Эволюция геномов эукариот и принцип максимальной парсимонии. *Вестник ВОГиС*, **9**(2): 141–152.

Татаринов Л.П. 1984. Кладистический анализ и филогенетика. *Палеонтологический журнал,* **1984**(3): 3–16.

Caccone A. and Powell J. R 1989. DNA divergence among hominoids. *Evolution*, **43**: 926–942.

Coyne J.A. and Orr H.A. 2004. Speciation. Sinauer Associates, Sunderland, Massachusetts, 545 p.

Farris J.S. 2008. Parsimony and explanatory power. *Cladistics*, **24**: 825–847.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, **27**: 401–410.

Felsenstein J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts, 664 p.

Goloboff P.A., Farris J.S. and Nixon K.C. 2008. TNT, a free program for phylogenetic analysis. *Cladistics*, **24**: 774–786.

Gregory T.R. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological Reviews of the Cambridge Philosophical Society*, **76**(1): 65–101.

Haszprunar G.1992. The types of homology and their significance for evolutionary biology and phylogenetics. *Journal of Evolutionary Biology*, **5**: 13–24.

Haszprunar G.1998. Parsimony analysis as specific kind of homology estimation and the implications for character weighting. *Molecular Phylogenetics and Evolution*, **9**: 333–339.

Hennig W. 1950. Grundzüge einer Theorie der phylogenetischen Systematik. Deutscher Zentralverlag, Berlin, 370 p.

Hennig W. 1965. Phylogenetic Systematics. *Annual Review of Entomology*, **10**: 97–116.

Hennig W. 1966. Phylogenetic Systematics. University of Illinois Press, Urbana, 284 p.

Huson D.H., Rupp R. and Scornavacca C. 2010. Phylogenetic networks: concepts, algorithms and applications. Cambridge University Press, New York, 362 p.

Krell F.-T. 2005. A Hennigian monument on vertebrate phylogeny. *Systematics and Biodiversity*, **3**(3): 339–341.

Lewis P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, **50**: 913–925.

Müller J. and Reisz R. 2006. The phylogeny of early Eureptiles: comparing parsimony and Bayesian approaches in the investigation of a basal fossil clade. *Systematic Biology*, **55**: 503–511.

Nichols R. 2001. Gene trees and species trees are not the same. *Trends in Ecology and Evolution*, **16**: 358–364.

Nylander J.A.A., Ronquist F., Huelsenbeck J.P. and Nieves-Aldrey J.L. 2004. Bayesian phylogenetic analysis of combined data. *Systematic Biology*, **53**(1): 47–67.

Page R.D.M. and Holmes E.C. 1998. Molecular evolution: a phylogenetic approach. Blackwell Publishing, Oxford, 346 p.

Ronquist F. 2004. Bayesian inference of character evolution. *Trends in Ecology and Evolution*, **19**(9): 475–481.

Ronquist F. and Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**(12): 1572–1574.

Scotland R.W. 2011. What is parallelism? *Evolution and development*, **13**(2): 214–227.

Wiley E.O., Siegel-Causey D., Brooks D.R. and Funk V.A. 1991. The compleat cladist: a primer of phylogenetic procedures. Lawrence, Kansas: The University of Kansas. 158 pp.

Wiley E.O. and Lieberman B.C. 2011. Phylogenetics. Theory and practice of phylogenetic systematics. Second Edition. Wiley-Blackwell Publishing, New Jersey, 406 p.