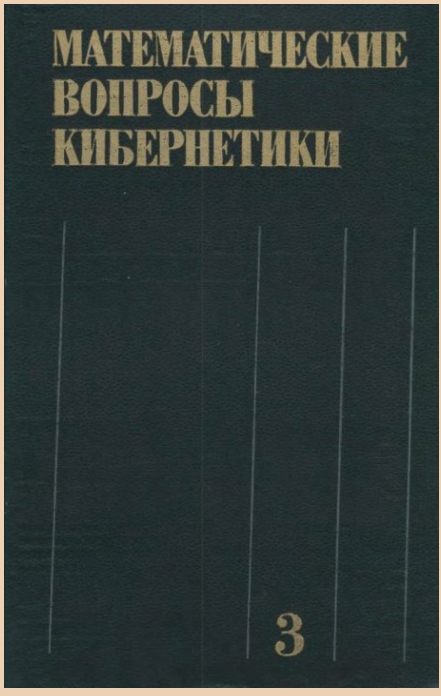
ИПМ им. М.В. Келдыша РАН • Электронная библиотека
Математические вопросы кибернетики • Выпуск 3



С. В. Юшманов

Восстановление биологической эволюции. Методы построения филогенетических деревьев

Рекомендуемая форма библиографической ссылки: Юшманов С. В. Восстановление биологической эволюции. Методы построения филогенетических деревьев // Математические вопросы кибернетики. Вып. 3. — М.: Наука, 1991. — С. 51–76. URL: http://library.keldysh.ru/mvk.asp?id=1991-51

ВОССТАНОВЛЕНИЕ БИОЛОГИЧЕСКОЙ ЭВОЛЮЦИИ. МЕТОДЫ ПОСТРОЕНИЯ ФИЛОГЕНЕТИЧЕСКИХ ДЕРЕВЬЕВ

С. В. ЮШМАНОВ

(МОСКВА)

СОДЕРЖАНИЕ

§ 1. Введение 51
§ 2. Основная модель филогении — модель филогенетического дерева . . . 52
§ 3. Идеальный случай — гипотеза аддитивного дерева 55
§ 4. Минимизация функционалов близости. Общая схема 56
§ 5. Принцип максимальной экономии 58
§ 6. Выбор оптимальной топологии. Принцип максимального топологического подобия 60
§ 7. Метод совместимости и другие 64
§ 8. Вычислительная сложность задачи построения филогенетического дерева 66
§ 9. Согласование филогенетических деревьев 67
§ 10. Модель филогенетического дерева — критический анализ 68
§ 11. Заключение. Проблемы и перспективы 70
Список литературы 72

§ 1. Введение

Данный обзор представляет собой переработанный вариант обзора [25] и предназначен для математиков, желающих ознакомиться с приложениями теории графов к одной из ключевых проблем эволюции — проблеме восстановления эволюционной истории (филогенеза). Эта задача возникает не только в биологии, но и во многих других областях, включая языкознание и классическую филологию, см., например, [34, 88]. Математические методы восстановления филогенеза, применяемые во всех этих областях, однотипны, но в биологии они используются гораздо шире. Поэтому описанные в обзоре методы будут излагаться применительно к задаче восстановления биологической эволюции, хотя и не все из них были разработаны именно с этой целью.

В § 2 формулируется модель филогенетического дерева, лежащая в основе применяемых на практике методов восстановления филогенеза. Она утверждает, что филогенез графически представляется взвешенным корневым деревом, в котором корень соответствует общему предку рассматриваемой группы S, висячие вершины — видам из S, внутренние — их предковым формам, а длины ребер отражают степень эволюционной близости видов, соответствующих инцидентным им вершинам.

В § 3 рассматривается идеальный случай, когда эволюционное расстояние, заданное на S, совпадает с расстоянием между вершинами искомого филогенетического дерева. В этом случае дерево восстанавливается однозначно.

Но в действительности любое используемое нами эволюционное расстояние не вполне адекватно отражает степень эволюционной близости.

Поэтому, как правило, матрица эволюционных расстояний, построенная по реальным данным, не является матрицей расстояний между висячими вершинами какого-либо дерева, и в качестве искомого филогенетического дерева приходится брать дерево, наиболее соответствующее заданному набору данных, причем критерий соответствия может выбираться по-разному.

Один из возможных подходов состоит в минимизации некоторого метрического функционала, измеряющего степень несовпадения матрицы эволюционных расстояний и матрицы расстояний между висячими вершинами искомого дерева. Различные виды таких функционалов и методы построения филогенетических деревьев, оптимальных относительно рассматриваемого функционала, обсуждаются в § 4, 5.

Методы построения филогенетических деревьев, основанные на минимизации неметрических функционалов, описаны в § 6, 7.

В § 8 обсуждается вычислительная сложность алгоритмов построения филогенетических деревьев, а в § 9 — задача согласования различных филогенетических деревьев, построенных по одному и тому же множеству данных.

В основе всех описанных в § 3—7 методов лежит модель филогенетического дерева. В § 10 обсуждаются накопленные за последнее время данные, показывающие, что представление филогенеза филогенетическим деревом «есть грубое, хотя во многих случаях и вынужденное упрощение» [4]. Границы применимости модели филогенетического дерева и математические проблемы, связанные с разработкой более общей модели, включающей модель филогенетического дерева как частный случай, рассматриваются в § 11.

§ 2. Основная модель филогении — модель филогенетического дерева

В основе модели лежат следующие, обычно считающиеся самоочевидными и поэтому, как правило, не формулируемые в явном виде представления о ходе биологической эволюции.

I. Аксиома дивергентности. *Эволюция носит дивергентный характер, т. е. любой вид имеет только одного непосредственного предка.*

II. Аксиома монофилии. *Любой реальный, а не сборный, таксон имеет монофилитическое происхождение, т. е. все виды, в него входящие, происходят от одного общего предка, давшего ему начало.*

III. Аксиома подобия. *Существует количественная мера эволюционной близости видов.*

Пусть нам известны все предковые формы рассматриваемого множества видов S. Тогда филогению S можно представить графически филогенетической схемой — неориентированным графом, вершинами которого являются виды из S и их предковые формы, и в котором любые два вида соединены ребром тогда и только тогда, когда один из них является непосредственным предком другого. Из аксиом дивергентности и монофилии следует, что филогенетическая схема является связным ациклическим графом, т. е. деревом, в котором можно выделить корень — вершину, соответствующую единственному общему предку видов из S. Поэтому вместо термина «филогенетическая схема» обычно используют термин «филогенетическое или эволюционное дерево». По историческим причинам в советской биологической литературе используется также термин «филогенетическое древо». Большая часть рассматриваемых далее методов позволяет восстановить вид филогенетического дерева, но не местоположение корня. Поэтому мы, как правило, будем рассматривать только некорневые деревья.

Аксиома подобия, постулирующая существование количественной меры эволюционной близости — эволюционного расстояния $\delta(x, y)$, поз-

воляет приписать каждому ребру (x, y) филогенетического дерева некоторую длину, характеризующую степень близости видов x, y и равную эволюционному расстоянию $\delta(x, y)$. Именно такие взвешенные филогенетические деревья и будут, как правило, рассматриваться в дальнейшем.

Аксиома III только постулирует существование эволюционного расстояния, но не указывает никаких способов его определения. Более того, прямых методов определения эволюционного расстояния не существует, и все используемые на практике методы вычисляют не сами эволюционные расстояния, а некоторые их аппроксимации. Вопрос о корректности и точности таких аппроксимаций достаточно сложен и лежит за пределами нашего обзора. Мы ограничимся только кратким описанием некоторых используемых в настоящее время аппроксимаций эволюционных расстояний, опуская всюду далее слово «аппроксимация».

Существуют эволюционные расстояния, определяемые непосредственно из эксперимента. Таково, например, иммунологическое расстояние [109], получаемое биохимическим анализом крови. Но большинство эволюционных расстояний вычисляются косвенным образом, на основе сравнения некоторого множества эволюционно значимых признаков.

Пусть для данного множества сравниваемых видов S каким-либо образом выбрано множество эволюционно значимых характеристик (признаков). Эти характеристики могут быть как качественными, так и количественными. Тогда каждому виду из S можно сопоставить характеристическую последовательность $x=(x_1, \ldots, x_m)$, где x_i — значение i-й характеристики для данного вида. Выбирая разные меры близости на множестве таких последовательностей, будем получать разные эволюционные расстояния.

Часто, особенно в генетике [16], используют евклидову метрику

$$\delta(x, y) = \left(\sum_{i=1}^{m} (x_i - y_i)^2 \right)^{1/2},$$

и многие методы построения филогенетических деревьев основаны на ней. Таков, например, метод Эдвардса и Кавалли-Сфорца [49].

Еще более часто используются расстояния вида

$$\delta(x, y) = \sum \alpha_i \delta(x_i, y_i),$$

где α_i — вес i-го признака, а $\delta(x_i, y_i)$ — расстояние между компонентами x_i и y_i. Если x_i и y_i вещественны, то часто берут $\delta(x_i, y_i) = |x_i - y_i|$ (см., например, [38]).

Важным частным случаем такого подхода является определение эволюционных расстояний на основе сравнения синонимичных (выполняющих одни и те же функции) аминокислотных или нуклеотидных последовательностей. Подробное описание механизма кодирования белков нуклеиновыми кислотами см. в [17], а здесь будут даны только самые необходимые сведения. Каждый белок представляет собой линейную последовательность нескольких сотен аминокислот двадцати различных видов, определенным образом уложенных в пространстве, и кодируется в ДНК (или РНК) линейной последовательностью нуклеотидов четырех различных видов. Таким образом, задача вычисления эволюционного расстояния по молекулярным данным сводится к задаче сравнения синонимичных аминокислотных или нуклеотидных последовательностей. В простейшем случае полагают $\alpha_i = 1$ для всех i и $\delta(x_i, y_i) = 0$, если $x_i = y_i$, и $\delta(x_i, y_i) = 1$ в противном случае. Тогда $\delta(x, y)$ равно числу позиций, в которых сравниваемые последовательности различаются. Для нуклеотидных последовательностей так определенное расстояние имеет ясный биологический

смысл — это минимально возможное число единичных мутаций (нуклеотидных замен), переводящих одну последовательность в другую.

Для аминокислотных последовательностей в качестве $\delta(x_i, y_i)$ можно брать также «функциональное расстояние», учитывающее степень физико-химической близости сравниваемых аминокислот (см., например, [104]), и «минимальное мутационное расстояние»— минимальное число нуклеотидных замен, переводящих кодон аминокислоты x_i (тройку нуклеотидов, кодирующих x_i) в кодон аминокислоты x_j. Таблица минимальных мутационных расстояний приведена в [17]. Расстояние между аминокислотными последовательностями, определенное через минимальные мутационные расстояния, тоже имеет ясный биологический смысл — это минимальное число единичных мутаций, переводящих один белок в другой.

На практике дело осложняется тем, что в процессе эволюции кроме замен нуклеотидов могут происходить также вставки и выпадения нуклеотидов, и поэтому сравниваемые последовательности могут иметь разную длину. Для вычисления расстояний в таких случаях применяются различные алгоритмы, в основе которых лежит выравнивание последовательностей, т. е. нахождение их взаимного расположения с максимальным числом совпадающих символов в одних и тех же позициях. При этом возможны два подхода: попарное выравнивание и совместное выравнивание всего множества анализируемых последовательностей. Ввиду вычислительной трудности задачи совместного выравнивания нескольких последовательностей на практике до недавнего времени использовалось только попарное выравнивание, для которого известны достаточно эффективные алгоритмы [31, 61, 67, 89, 99]; наиболее известны из них [89, 99], эквивалентные на достаточно широком классе последовательностей [103]. За последние годы разработаны и стали приобретать популярность и алгоритмы совместного выравнивания [32, 39, 87, 96].

Завершая краткий обзор эволюционных расстояний, отметим, что далеко не все они являются расстояниями в точном смысле слова, т. е. удовлетворяют аксиомам метрики. Так, в силу вырожденности генетического кода, минимальное мутационное расстояние не удовлетворяет неравенству треугольника.

Таким образом, в рамках модели филогенетического дерева задача восстановления филогенеза может быть сформулирована следующим образом. Заданы некоторое множество S вершин искомого филогенетического дерева, матрица эволюционных расстояний между видами из S и, возможно, характеристические последовательности видов из S. Требуется по этой информации восстановить искомое дерево, причем если характеристические последовательности видов из S известны, требуется еще восстановить и характеристические последовательности всех внутренних вершин дерева. Обычно, рассматривая эту задачу, вводят еще две аксиомы.

IV. Аксиома бинарности. *Филогенетическое дерево бинарно, т. е. все его вершины, за исключением корня, имеют степень* 1 *или* 3, *а корень, если он есть, имеет степень* 2.

V. Аксиома висячих вершин. *Висячие вершины дерева соответствуют видам из* S, *а внутреннние вершины — их предковым формам.*

Вопрос о биологической обоснованности аксиом, лежащих в основе модели филогенетического дерева, будет обсуждаться позже, а сейчас отметим, что принципиальный характер носят лишь первые три аксиомы, а последние две введены для упрощения расчетов. Более того, с точки зрения биолога они неверны, так как один предковый вид может дать начало одному или нескольким новым видам, и появление нового вида не обязательно приводит к вымиранию вида, от которого он произошел.

Тем не менее, введение этих аксиом не нарушает общности модели, так как из любого небинарного дерева с выделенным подмножеством вершин S, содержащим все висячие вершины, можно добавлением новых вершин и ребер длины 0 получить такое бинарное дерево, что множество его висячих вершин совпадает с S, и расстояния между вершинами из S в обоих деревьях одни и те же. Пример такого преобразования приведен на рис. 1, где вершины из S помечены буквами, и числа, приписанные ребрам, обозначают их длины.

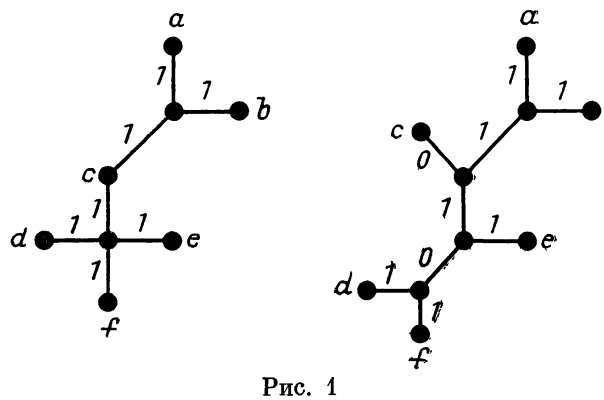


Рис. 1

И, наконец, отметим, что хотя описываемые далее методы построения филогенетических деревьев будут излагаться только для бинарных деревьев, тем не менее они почти все либо непосредственно применимы к небинарным деревьям, либо могут быть соответствующим образом переформулированы.

§ 3. Идеальный случай — гипотеза аддитивного дерева

Согласно определению филогенетического дерева длина любого его ребра (x, y) равна эволюционному расстоянию $\delta(x, y)$. Определим расстояние $d(x, y)$ между вершинами дерева x, y как сумму длин ребер простой цепи, соединяющей x и y. В [40, 49] было выдвинуто предположение, получившее название гипотезы аддитивного дерева и утверждающее, что *для любых двух вершин x, y произвольного филогенетического дерева*

$$\delta(x, y) = d(x, y).$$

В предположении, что гипотеза аддитивного дерева справедлива, задача восстановления филогенетического дерева сводится к задаче восстановления дерева с неотрицательными длинами ребер по его матрице расстояний между висячими вершинами.

Теорема Смоленского — Зарецкого. *Взвешенное дерево с неотрицательными длинами ребер восстанавливается по матрице расстояний между висячими вершинами с точностью до вершин степени* 2 *и ребер длины* 0, *причем произвольная матрица расстояний $D = \|d_{ij}\|$ является матрицей расстояний между висячими вершинами дерева с неотрицательными длинами ребер тогда и только тогда, когда для любых i, j, k, l среди чисел $d_{ij} + d_{kl}$, $d_{ik} + d_{jl}$, $d_{il} + d_{jk}$ два равны и не меньше третьего.*

Эта теорема, дающая полное решение задачи восстановления аддитивного филогенетического дерева, была первоначально сформулирована для невзвешенного случая [7, 8, 19] и позднее была переоткрыта уже применительно к задаче восстановления филогенеза [37, 47]. Сформулированному в теореме условию, которое мы будем называть условием четырех вершин, можно придать следующую форму: *матрица расстояний является матрицей расстояний между висячими вершинами некоторого дерева с неотрицательными длинами ребер тогда и только тогда, когда каждая ее главная подматрица порядка* 4 *является матрицей расстояний между висячими вершинами некоторого дерева* [100]. Алгоритмы восстановления аддитивного филогенетического дерева, основанные на этой теореме, описаны в [6, 13, 42, 97, 108].

Но в действительности реальные филогенетические деревья почти никогда не удовлетворяют гипотезе аддитивного дерева. Рассмотрим дерево на рис. 2, где данные представлены условными последовательностями, и эволюционные расстояния на них равны числу несовпадающих позиций. Легко видеть, что дерево не является аддитивным и матрица эволюционных расстояний между его висячими вершинами не удовлетворяет условию четырех вершин. Как видно из рисунка, одной из причин нарушения аддитивности являются обратные и параллельные замены. Более того, нарушение аддитивности может происходить и при отсутствии обратных и параллельных мутаций, если в одной и той же позиции последовательностей, лежащих на одной простой цепи, встречается более двух различных символов.

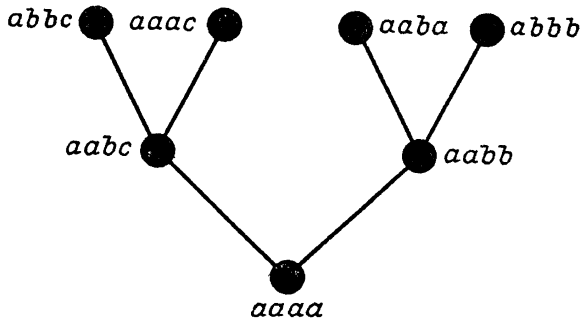


Рис. 2

Поэтому матрица эволюционных расстояний между висячими вершинами реального филогенетического дерева может не удовлетворять условию четырех вершин. Более того, даже если матрица удовлетворяет этому условию, то дерево, построенное по ней, может не отражать реального эволюционного процесса. Чтобы убедиться в этом, достаточно рассмотреть дерево на рис. 3, *а*. Его матрица эволюционных расстояний удовлетворяет условию четырех вершин, но построенное по нему дерево, изображенное на рис. 3, *б*, не совпадает с исходным.

Таким образом, филогенетические деревья не обязаны удовлетворять и, как правило, не удовлетворяют гипотезе аддитивного дерева. Поэтому задача построения филогенетического дерева по набору реальных данных сводится не к задаче восстановления дерева по однозначно его определяющему набору данных, а к задаче построения дерева, наиболее соответствующего заданному набору данных, причем наиболее трудной частью задачи является выбор содержательно обоснованного критерия

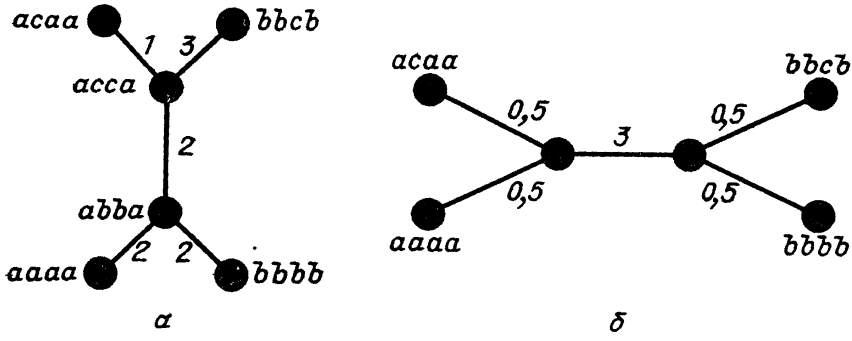


Рис. 3

соответствия, а не само построение дерева согласно этому критерию. Следующие параграфы обзора и будут посвящены обсуждению применяемых в настоящее время критериев соответствия и основанных на них методов.

§ 4. Минимизация функционалов близости. Общая схема

Рассмотрим наиболее общий случай, когда известна только $n \times n$-матрица $\Delta = \|\delta_{ij}\|$ эволюционных расстояний между видами из S. Пусть T — произвольное невзвешенное бинарное дерево с n висячими вершина-

ми, в котором висячие вершины занумерованы числами от 1 до n, а ребра — числами от 1 до $m = 2n - 3$, и пусть $\mathcal{T}_n$ — множество всех таких деревьев. Определим взвешенное бинарное дерево как пару $T_X = (T, X)$, где $T \in \mathcal{T}_n$, а $X = (x_1, \ldots, x_m)$ — вектор длин, приписанных ребрам T, и обозначим через $D = D(T_X) = \|d_{ij}\|$ матрицу расстояний между висячими вершинами T_X.

Естественно в качестве критерия степени соответствия филогенетического дерева исходным данным брать степень близости матриц D и Δ. Тогда задача построения филогенетического дерева по заданной матрице Δ сводится к задаче минимизации некоторого функционала $F(D, \Delta)$, т. е. к задаче нахождения дерева $T_X^* = (T^*, X^*)$ с матрицей расстояний $D^* = D(T_X^*)$, для которого

$$F(D^*, \Delta) = \min_{T \in \mathcal{T}_n} \min_X F(D(T_X), \Delta). \qquad (4.1)$$

Конкретизируя вид функционала и вводя в случае необходимости дополнительные ограничения на вид решения, будем получать разные методы построения филогенетических деревьев, основанные на одной и той же вычислительной схеме (4.1).

Одним из первых на практике был проведен квадратичный функционал Фитча — Марголиаш

$$F(D, \Delta) = \sum \left(\frac{d_{ij} - \delta_{ij}}{d_{ij}} \right)^2.$$

В [65] этим методом построено филогенетическое дерево цитохромов c (цитохром c — один из белков цепи клеточного дыхания) — пожалуй, наиболее известное из филогенетических деревьев, построенных по молекулярным данным. Именно оно, как правило, помещается в учебниках, научно-популярных работах, посвященных проблемам молекулярной эволюции (см., например, [1, 2]). Но несмотря на свою популярность, оно во многом противоречит традиционным филогениям. Например, в этом дереве эволюционная ветвь человека и обезьян ответвляется от общего ствола раньше, чем ветвь кенгуру ответвляется от ветви плацентарных млекопитающих. Кроме того, в нем:

1) есть ребра отрицательной длины;

2) для 75 пар висячих вершин из 190 не выполнено условие $d_{ij} \geqslant \delta_{ij}$ [18];

3) длины ребер принимают нецелые значения.

Недопустимость пункта 1) очевидна, так как любое эволюционное расстояние, определенное сколь-либо разумным образом, не может принимать отрицательных значений. Что же касается пунктов 2), 3), то в [65] в качестве эволюционного расстояния δ_{ij} бралось минимально возможное число мутаций, переводящих один белок в другой. Поэтому в построенном дереве расстояние d_{ij} соответствует реально затраченным на это мутациям и, следовательно, должно быть целочисленно и не меньше δ_{ij}.

Другой квадратичный функционал $\sum (d_{ij} - \delta_{ij})^2$ рассматривался в [40, 72, 79]. В последних двух работах рассматривался также функционал $\sum |d_{ij} - \delta_{ij}|$. Они обладают теми же недостатками, что и функционал Фитча — Марголиаш. Следует только отметить, что использованные в [40, 72, 79] эволюционные расстояния не требуют целочисленности и выполнения условия $d_{ij} \geqslant \delta_{ij}$. Кроме того, в [40] было предложено рассматривать только деревья с положительной длиной ребер.

Другой путь, применяемый, как правило, в молекулярной биологии, состоит в использовании линейных функционалов. В [36, 108] рассматривались функционалы: 1) $\sum (d_{ij} - \delta_{ij})/\delta_{ij}$; 2) $\sum (d_{ij} - \delta_{ij})/\delta_{ij}^2$; 3) $\sum x_i$ —

т. е. общая длина дерева. При этом на вид решения накладывались следующие линейные ограничения, смысл которых разобран выше:

1) длины всех ребер дерева неотрицательны;
2) для любых i, j $d_{ij} \geqslant \delta_{ij}$.

Иногда к этим ограничениям добавляют еще требование целочисленности длин ребер. При фиксированном T как сформулированная задача является задачей линейного программирования, и для ее решения могут быть использованы стандартные методы линейного программирования [107, 108].

Как правило, из трех указанных выше линейных функционалов употребляется только функционал $\sum x_i$, который более подробно рассматривается в следующем параграфе.

§ 5. Принцип максимальной экономии

Минимизация линейного функционала

$$F(D, \Delta) = \sum x_i$$

при линейных ограничениях:

1) для всех i $x_i \geqslant 0$;
2) для всех i, j $d_{ij} \geqslant \delta_{ij}$

имеет ясный биологический смысл, так как минимизация общей длины дерева при соблюдении указанных выше ограничений соответствует минимизации общего числа мутаций и является реализацией сформулированного в [40, 49] принципа максимальной экономии, утверждающего, что реальный ход эволюционного процесса осуществляется вдоль пути с наименьшим числом эволюционных событий. Филогенетические деревья, отвечающие принципу максимальной экономии, обычно называют минимальными филогенетическими деревьями или деревьями максимальной экономии.

Используемые в молекулярной биологии эволюционные расстояния вычисляются по молекулярным последовательностям, которые информативны и сами по себе. Поэтому желательно восстановление не только минимального дерева, но и приписанных его внутренним вершинам молекулярных последовательностей. Этого можно достичь двумя путями. Первый подход состоит в восстановлении минимального дерева по матрице эволюционных расстояний и последующем нахождении молекулярных последовательностей внутренних вершин по известным последовательностям вершин из S [62, 75]. При втором подходе минимальное дерево и молекулярные последовательности строятся одновременно. Это так называемая задача Штейнера в филогении [68].

Принцип максимальной экономии долгое время был очень популярен в молекулярной биологии, и большинство известных к настоящему моменту филогенетических деревьев различных биополимеров построено алгоритмами, основанными на этом принципе. Во многом это объясняется тем, что функционал $\sum x_i$ — это один из немногих известных на настоящее время функционалов, минимизация которых имеет ясный биологический смысл.

Но в последние годы популярность принципа максимальной экономии значительно упала. Причины этого заключаются в следующем.

Во-первых, несмотря на свою простоту, принцип максимальной экономии вряд ли удовлетворителен с философской точки зрения, так как в неявной форме предполагает наличие у эволюции цели. Поэтому даже в период наибольшей популярности этого принципа раздавались голоса,

утверждавшие, что «широкое практическое построение родословных древес по принципу экономии представляет по существу насилие над природой» [10].

Во-вторых, дёревья максимальной экономии, так же как и деревья, минимизирующие функционал Фитча — Марголиаш, как правило, во многом противоречат здравому смыслу и традиционным филогениям. Например, согласно дереву максимальной экономии цитохромов *с* семи видов млекопитающих [69], изображенному на рис. 4, *а*, кролик происходит от кита.

И наконец, как выяснилось, деревья максимальной экономии неустойчивы по отношению к локальным изменениям данных [94].

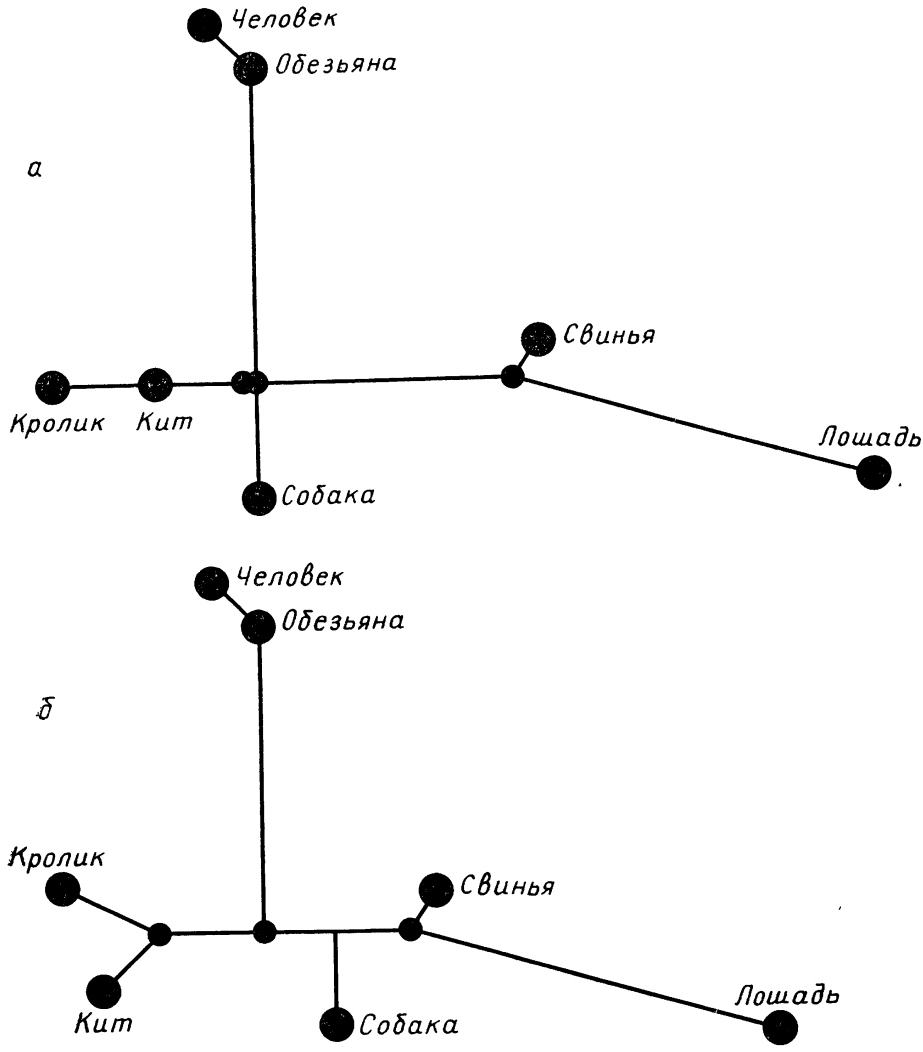
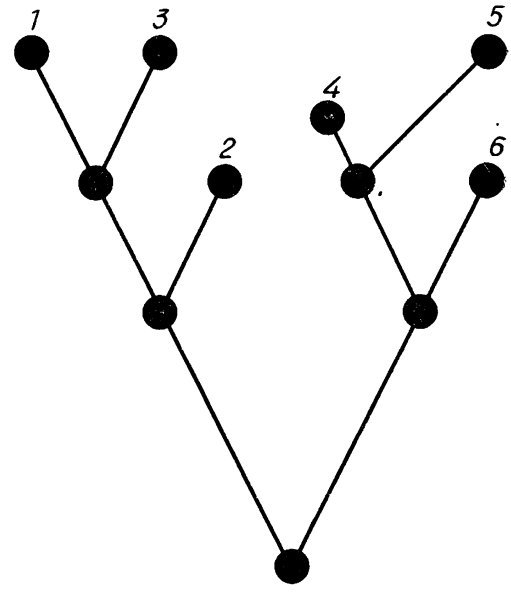


Рис. 4

На рис. 5, взятом из [94], изображены два условных множества данных и построенные по ним минимальные деревья. Эти множества отличаются только значением признака 2 вида 6, а деревья различаются взаимным расположением видов 1—3, лежащих в ветви дерева, не содержащей вида 6, в то время как взаимное расположение видов, лежащих в ветви, его содержащей, не изменилось. Этот пример показывает, что при работе с филогениями, построенными по принципу максимальной экономии, теоретически возможна ситуация, когда, например, незначительное изме-

Вид	Данные
1	01
2	20
3	22
4	70
5	72
6	60



Вид	Данные
1	01
2	20
3	22
4	70
5	72
6	62

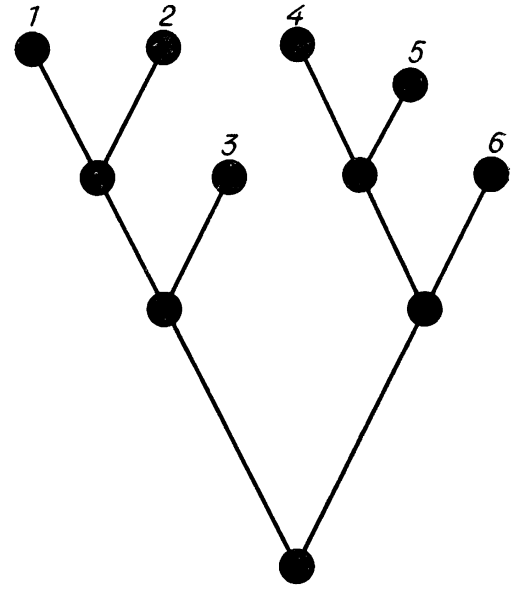


Рис. 5

нение данных об одном виде бабочек, возможно, и не требующее изменения филогении отряда бабочек, повлечет пересмотр филогении отряда двукрылых, или даже более далеких таксономических групп.

§ 6. Выбор оптимальной топологии. Принцип максимального топологического подобия

Задачу построения оптимального филогенетического дерева $T_x = (T, X)$ можно разбить на две подзадачи: выбор оптимальной топологии T и определение оптимальных длин ребер X. Все описанные выше методы решают эти задачи одновременно. Но очевидно, что определение истинной топологии дерева является существенно более важной задачей по сравнению с определением конкретных длин ребер.

Поэтому представляется разумным сначала выбирать топологию и уже потом искать длины ребер, оптимальные именно для этой топологии. Вторая задача решается эффективно для всех функционалов мини-

мизации, описанных выше, и основную трудность представляет выбор топологии. Для решения этой задачи целесообразно в максимальной степени отвлечься от конкретных расстояний и сосредоточить основное внимание на выборе общей структуры дерева, что в свою очередь требует разработки математического аппарата, позволяющего описывать топологию дерева и оценивать степень ее соответствия сведениям о взаимном расположении объектов, запечатленным в исходных данных. Впервые

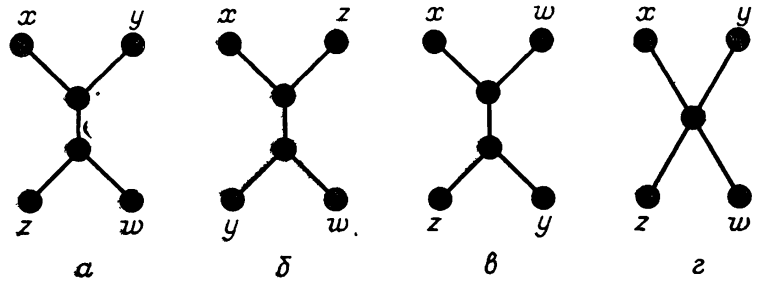


Рис. 6

такой топологический подход был предложен Фитчем в [64]. Для его описания нам понадобится ряд дополнительных определений.

Придадим условию четырех вершин следующую, более удобную для наших целей форму.

Матрица D расстояний между элементами некоторого множества S реализуется деревом тогда и только тогда, когда любые 4 элемента из S можно обозначить x, y, z, w таким образом, что

$$d(x, y)+d(z, w)\leqslant d(x, z)+d(y, w)=d(x, w)+d(y, z). \qquad (6.1)$$

Легко видеть, что матрица расстояний между x, y, z, w, удовлетворяющая условию (6.1), реализуется деревом на рис. 6, *а*, вырождающимся в дерево на рис. 6, *г*, если все три суммы, входящие в (6.1), равны. Назовем вершины x, y в дереве на рис. 6, *а*, *соседями* и обобщим это определение на произвольное дерево T_x следующим образом [97]. *Степень соседства* $\mathrm{dn}(x, y)$ висячих вершин x, y равна числу четверок $\{x, y, z, w\}$ различных висячих вершин, удовлетворяющих условию (6.1). Матрицу степеней соседства висячих вершин T_x обозначим $\mathrm{DN}(T_x)$. В [35] показано, что она однозначно определяет топологию дерева, т. е. задает дерево с точностью до вершин степени 2 и длин ребер.

Пусть на S задана матрица эволюционных расстояний Δ (напомним, что функция δ не обязана удовлетворять аксиомам метрики). Будем говорить, что элементы $x, y, z, w \in S$ удовлетворяют обобщенному условию четырех вершин, если

$$\delta(x, y)+\delta(z, w)\leqslant\delta(x, z)+\delta(y, w)\leqslant\delta(x, w)+\delta(y, z). \qquad (6.2)$$

Легко видеть, что если (6.2) выполнено, то дерево на рис. 6, *а* является лучшей аппроксимацией матрицы эволюционных расстояний между x, y, z, w, чем деревья на рис. 6, *б* и 6, *в*. Поэтому, как и выше, назовем x, y соседями и определим для любых двух элементов $x, y \in S$ *степень соседства* $\mathrm{dn}(x, y)$ следующим образом. Припишем каждой четверке $\{x, y, z, w\}$, удовлетворяющей условию (6.2), две единицы соседства, распределенные между парами, в нее входящими, следующим образом. Если

$$\delta(x, y)+\delta(z, w)<\delta(x, z)+\delta(y, w)\leqslant\delta(x, w)+\delta(y, z),$$

то приписываем каждой из пар $\{x, y\}$, $\{z, w\}$ степень соседства 1. Если же

$$\delta(x, y)+\delta(z, w)=\delta(x, z)+\delta(y, w)<\delta(x, w)+\delta(y, z),$$

приписываем каждой из пар $\{x, y\}$, $\{z, w\}$, $\{x, z\}$, $\{y, w\}$ степень соседства 1/2. И, наконец, если все три суммы равны, припишем каждой из пар $\{x, y\}$, $\{z, w\}$, $\{x, z\}$, $\{y, w\}$, $\{x, w\}$, $\{y, z\}$ степень соседства 1/3. Тогда степень соседства $\mathrm{dn}(x, y)$ любых двух различных элементов из S равна сумме степеней соседства пары $\{x, y\}$, взятой по всем четверкам $\{x, y, z, w\}$ различных элементов из S, удовлетворяющим условию (6.2).

Пусть на S задана матрица эволюционных расстояний Δ, и пусть $DN(\Delta)$ есть матрица степеней соседства элементов из S. Тогда, согласно [64], в качестве дерева, наиболее соответствующего матрице расстояний Δ, берется дерево T, минимизирующее функционал $|DN(T) - DN(\Delta)|$, равный сумме модулей разности между соответствующими элементами матриц степеней соседства. Преимущества такого подхода состоят в следующем.

1. Класс матриц Δ, чья матрица степеней соседства $DN(\Delta)$ есть матрица степеней соседства некоторого дерева, шире класса матриц расстояний, реализуемых деревом (удовлетворяющих условию четырех вершин).

2. Метод нечувствителен к малым погрешностям исходных данных. Пусть матрица Δ удовлетворяет условию четырех вершин, и пусть для какой-либо четверки вершин x, y, z, w из трех сумм $s_1 = d(x, y) + d(z, w)$, $s_2 = d(x, z) + d(y, w)$, $s_3 = d(x, w) + d(y, z)$ две, скажем, s_2, s_3, равны и много больше третьей. Тогда любая малая погрешность в вычислении s_2, s_3 или входящих в них расстояний нарушает реализуемость Δ деревом, но не нарушает реализуемость деревом матрицы степеней соседства.

3. Функционал $|DN(T) - DN(\Delta)|$ дает простой объективный способ оценки степени соответствия построенного дерева T исходной матрице расстояний Δ.

Но на настоящее время не известно отличных от полного перебора методов построения деревьев, чьи матрицы степеней соседства близки к заданным. Поэтому топологический метод Фитча практического применения не нашел.

Другим недостатком метода Фитча является то, что матрица степеней соседства лишь очень грубо описывает топологию дерева.

Гораздо более тонким средством описания топологии дерева являются функции древовидности, используемые в топологическом подходе, носящем название принципа максимального топологического подобия. Этот подход был первоначально предложен в [26, 48] и получил окончательную форму в [24, 28].

Функция древовидности F_T, определенная на множестве четверок висячих вершин дерева T, задается следующим образом: $F_T(x, y, z, w) = 1$, если в (6.1) имеет место строгое неравенство; $F_T(x, y, z, w) = 0$, если все три суммы в (6.1) равны; во всех остальных случаях функция F_T не определена. В [41] показано, что функция F_T однозначно определяет топологию дерева, а в [26, 35, 41] получен ряд критериев того, что произвольная функция F является функцией древовидности некоторого дерева T.

Аналогично определяется функция древовидности F_Δ матрицы расстояний Δ: $F_\Delta(x, y, z, w) = 1$, если в (6.2) имеет место строгое неравенство; $F_\Delta(x, y, z, w) = 0$, если все три суммы в (6.2) равны; во всех остальных случаях функция F_Δ не определена.

Тогда в соответствии с принципом максимального топологического подобия в качестве дерева, наиболее соответствующего матрице Δ, берется дерево максимального топологического подобия T, минимизирующее функционал топологического несовпадения $|F_\Delta - F_T|$, равный числу четверок $\{x, y, z, w\}$, на которых $F_\Delta \neq F_T$.

Принцип максимального топологического подобия обладает всеми достоинствами топологического метода Фитча, а именно:

1. Класс матриц Δ, чья функция древовидности F_Δ есть функция древовидности некоторого дерева, шире класса матриц расстояний, реализуемых деревом.

2. Метод нечувствителен к малым погрешностям исходных данных.

3. Функционал топологического несовпадения $|F_\Delta - F_T|$ дает простой объективный способ оценки степени соответствия построенного дерева T исходной матрице расстояний Δ.

Но кроме этих достоинств, присущих обоим методам, принцип максимального топологического подобия обладает еще и следующими.

4. В отличие от топологического метода Фитча принцип максимального топологического подобия позволяет в явной фрме получить алгоритмы построения деревьев [28].

5. Сравнение функций F_Δ и F_T на всех наборах четверок, включающих интересующий нас вид, позволяет получить количественную оценку надежности его размещения на дереве в виде величины локального топологического несовпадения. Иными словами, принцип максимального топологического подобия позволяет выявлять в филогенетических деревьях места, требующие более осторожной трактовки.

Для иллюстрации сказанного рассмотрим деревья максимальной экономии цитохромов *с* семи видов млекопитающих [69], изображенные на рис. 4. Анализ этих деревьев с филогенетической точки зрения показывает, что дерево 4, *б* более адекватно хотя бы уже потому, что в отличие от дерева 4, *а* не утверждает, что кролик произошел от кита, а лишь предполагает наличие у этих видов общего предка. Этот вывод находится в полном соответствии с принципом максимального топологического подобия, так как а) в дереве 4, *а* величина локального топологического несовпадения для кролика значительно выше величины локального топологического несовпадения для остальных видов [28], и следовательно, возможно, что положение кролика в дереве неверно; б) более приемлемое филогенетически дерево 4, *б* имеет меньшую величину топологического несовпадения, чем дерево 4, *а* [28].

И, наконец, отметим, что если исходные данные представлены последовательностями, то принцип максимального топологического подобия может быть использован для получения приближений к филогенетическим деревьям, основанных на других принципах минимизации, скажем, на принципе максимальной экономии или на любом из принципов, описанных в § 7. Мы опишем построение таких приближений для деревьев максимальной экономии, а приближения к другим типам филогенетических деревьев строятся аналогично. Это делается следующим образом [27, 35]. Для каждой четверки последовательностей строим дерево максимальной экономии и определяем по полученному дереву значение функции древовидности F на этой четверке. После этого в соответствии с принципом максимальной экономии строим дерево, чья функция древовидности так близка к F, как это возможно.

Так полученное дерево, вообще говоря, не является деревом максимальной экономии, так как при его построении принцип максимальной экономии применяется не ко всему множеству данных, а только к его подмножествам из четырех последовательностей. Тем не менее, такие приближения, получившие название деревьев локальной экономии [27], достаточно эффективны. Например, для последовательностей цитохромов *с* семи видов млекопитающих, рассмотренных выше, дерево локальной экономии совпадает с деревом на рис. 4, *б* — наиболее филогенетически приемлемым из двух деревьев максимальной экономии этого множества последовательностей [27].

§ 7. Метод совместимости и другие

Все определенные выше методы, несмотря на свои различия, характеризуются тем, что они используют всю, зачастую противоречивую информацию, данную им, и пытаются на основе этой информации построить наилучшее возможное приближение к истинному филогенетическому дереву, не используя при этом никаких априорных предположений о ходе эволюции, если не считать ограничений, в неявном виде накладываемых выбором функционала $F(T_x, \Delta)$. Для данных, представленных последовательностями, возможны и другие, принципиально отличные от них методы.

Пусть даны четыре вида, представленные бинарными векторами признаков, $A = 1\,100\,000$, $B = 1\,010\,001$, $C = 0\,001\,010$, $D = 0\,001\,101$, и пусть известно, что их истинная филогения представляется филогенетическим деревом, изображенным на рис. 7, где корень R соответствует

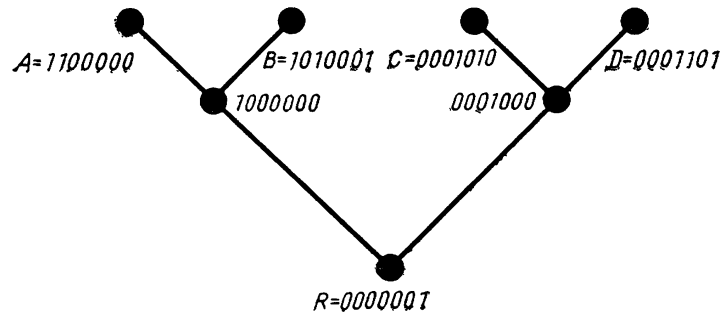


Рис. 7

их ближайшему общему предку. Легко видеть, что это дерево не является аддитивным и не удовлетворяет условию четырех вершин. В данном случае нарушение аддитивности объясняется обратной эволюцией признака, стоящего в седьмой позиции. У предкового вида R этот признак находится в состоянии 1, у видов A и C, от него происшедших,— в состоянии 0, и снова в состоянии 1 — у видов B и D. Если же мы отбросим седьмой признак и станем рассматривать только первые шесть, то в филогенетическом дереве ни по одному из признаков не будет наблюдаться ни параллельной, ни обратной эволюции, дерево будет удовлетворять условию четырех вершин и будет восстанавливаться по матрице эволюционных расстояний с точностью до положения корня.

Рассмотренный пример показывает, что не все признаки одинаково полезны для построения филогении и, более того, в некоторых случаях включение в рассмотрение дополнительных признаков может не только не помочь, но даже помешать восстановлению истинной филогении. Легко видеть, что наиболее полезны для восстановления филогении те признаки, любые состояния которых возникают в ходе эволюции только один раз, и поэтому кажется естественным строить филогенетические деревья только по таким признакам, отбрасывая все остальные. Но, с другой стороны, знание того, у каких признаков любые состояния в ходе эволюции возникают не более одного раза, а у каких нет, уже предполагает знание той самой филогении, которую мы хотим восстановить. Ситуация усугубляется еще тем, что мы не знаем не только самой филогении, но и большей части видов, в нее входящих, так как нам неизвестны вымершие предковые виды.

Эта трудность разрешается переходом к рассмотрению совместимых признаков и построению искомого филогенетического дерева по максимальному множеству взаимно совместимых признаков. Признаки назы-

ваются взаимно совместимыми, если существует филогения, в которой ни одно состояние ни одного из признаков не возникает в ходе эволюции более одного раза. В противном случае признаки называются несовместимыми. Следует подчеркнуть, что филогения, фигурирующая в определении совместимости, не обязана быть истинной, а должна только быть формально возможной.

Такой подход к построению филогенетических деревьев, который будем называть методом совместимости, был впервые предложен в [81, 110], а само понятие совместимости в несколько иной форме было введено в [38]. Математически метод совместимости был обоснован в [51—53], где было дано описание в терминах полуструктур и отображений на полуструктуры.

Построение филогенетического дерева методом совместимости разбивается на три этапа.

I. Нахождение всех пар попарно совместимых признаков.

II. Нахождение максимальной клики, т. е. максимального по мощности множества взаимно совместимых признаков.

III. Построение филогенетического дерева по найденной клике.

Для случая бинарных признаков, т. е. признаков, имеющих два состояния, которые мы будем обозначать 0 и 1, вопрос о попарной совместимости был решен в [81], где было показано, что если неизвестно, какое из двух состояний является начальным, то два признака совместимы тогда и только тогда, когда не более чем три из четырех возможных комбинаций состояний 00, 11, 01, 10 встречаются в исходных данных. Если же известно, какое из двух состояний является начальным, пусть для определенности это 0, то в данных должно встречаться не более двух комбинаций из комбинаций 01, 10, 11. В случае небинарных признаков с известным начальным состоянием проверка совместимости небинарного признака сводится к задаче проверки совместимости всех пар бинарных признаков, его кодирующих [53]. Методы для проверки совместимости небинарных признаков с неизвестным начальным состоянием описаны в [54—56].

Следующий этап основывается на теореме о попарной совместимости, утверждающей, что признаки взаимно совместимы тогда и только тогда, когда они попарно совместимы. Эта теорема верна для бинарных признаков как с известным начальным состоянием так и с неизвестным [52, 53]. Как указано выше, задача проверки совместимости небинарных признаков с известным начальным состоянием сводится к бинарному случаю, и поэтому теорема о попарной совместимости верна и для небинарных признаков с известным начальным состоянием [53]. Если же начальное состояние небинарных признаков неизвестно, то, вообще говоря, теорема о попарной совместимости неверна. Соответствующий контрпример для нуклеотидных последовательностей был построен в [63]. Тем не менее метод совместимости применим и для нуклеотидных последовательностей. В этом случае рассматриваются только те позиции, которые содержат ровно два различных нуклеотида, вычеркивая все остальные [14].

Если максимальная клика уже выбрана, то искомое филогенетическое дерево строится очевидным образом, так как условие, что каждое состояние возникает в ходе эволюции только один раз, накладывает жесткие ограничения как на вид бинарного филогенетического дерева, так и на последовательности состояний, приписываемые его внутренним вершинам, и позволяет однозначно определять пары видов, имеющие непосредственного общего предка, и последовательности состояний, приписываемые этим предковым вершинам. В частности, если признаки бинарны, то филогенетическое дерево, соответствующее любому выбранному множеству взаимно совместимых признаков, аддитивно [14]. Даль-

нейшие подробности можно найти в [86], где подробно описаны все три этапа в удобном для ручных вычислений виде.

Пусть рассматриваемые признаки бинарны и пусть состояние 0 является начальным. Тогда метод совместимости имеет такое подмножество признаков, что существует филогения, в которой каждый признак, перейдя однажды в состояние 1, никогда уже не переходит обратно в состояние 0. Такой подход неизбежно предполагает исключение из рассмотрения части признаков, иногда значительной. Если же мы хотим рассматривать все признаки, нам необходимо рассматривать более широкие классы допустимых филогений. Это можно сделать разными путями.

Камин и Сокал [38] предположили, что переход из состояния 1 в состояние 0 невозможен, но состояние 1 может возникать в ходе эволюции неоднократно. При таком подходе для любого набора признаков можно построить непротиворечивую филогению, причем не одну, и мы выбираем филогению с минимальным числом переходов из 0 в 1. Этот же подход был применен в [38] и к признакам, имеющим более двух состояний.

Другой возможный подход основан на принципе необратимости эволюции, который был выдвинут в конце прошлого века бельгийским палеонтологом Долло и поэтому часто называется принципом Долло. Принцип Долло утверждает, что раз утерянный в ходе эволюции комплекс признаков не может возникнуть вторично. Пусть 1 — это сложное производное состояние, возникшее из простого начального состояния 0. Тогда, согласно методу, основанному на принципе Долло, мы ищем филогению, в которой не более чем один переход из 0 в 1 и так мало переходов из 1 в 0, как это возможно. Этот метод был предложен в [82] и формализован в [59].

Следует только отметить, что если в самом принципе Долло речь идет о комплексах признаков, то в методе, основанном на нем, принцип Долло применяется к отдельным признакам, а согласно современным представлениям (см., например, [10, 22]), принцип Долло к отдельным, пусть даже сложным признакам, неприменим. Например, точечные мутации, обусловленные единичными заменами нуклеотидов, могут быть возвратными. Точно так же дело обстоит и со сложными признаками. Как показано в [15], даже такой сложный признак, как цветное зрение, возникал в ходе эволюции неоднократно, в том числе и в тех эволюционных линиях, в которых цветное зрение было ранее приобретено, а потом утеряно. Анализируя именно этот пример, Красилов [10] подчеркнул, «что многократное возникновение сложных структур, вопреки широко распространенному мнению, не менее вероятно, чем простых». Поэтому истинность филогений, построенных этим методом, очень сильно зависит от способов выделения признаков, к которым метод применяется.

§ 8. Вычислительная сложность задачи построения филогенетического дерева

В предыдущих параграфах подробно рассматривались различные подходы к задаче построения филогенетического дерева, но нигде, за малыми исключениями, не затрагивался вопрос существования и эффективности реализующих эти подходы алгоритмов построения деревьев. Между тем от ответа на этот вопрос зависит практическая применимость описываемых подходов.

Теоретически построение филогенетических деревьев можно осуществить полным перебором. Но такой подход возможен только при малых n, где n — число видов. Уже для $n = 10$ необходимо просмотреть 2 027 025 деревьев. Если же мы ведем перебор не по бинарным, а по

произвольным филогенетическим деревьям, т. е. по произвольным деревьям, в которых помечено некоторое множество вершин, содержащее все вершины степени меньшей 3 и соответствующее рассматриваемому множеству видов, то для того же $n = 10$ необходимо перебрать уже 4 093 236 352 дерева [71].

Поэтому крайне важен вопрос, для каких филогений существуют эффективные полиномиальные алгоритмы их построения. Но к сожалению похоже, что таких алгоритмов не существует, так как, по всей видимости, для всех функционалов близости, не накладывающих слишком жестких ограничений на вид искомого дерева, задача его построения *NP*-полна. Так, *NP*-полны задачи построения деревьев максимальной экономии [43, 68, 73] и деревьев, отвечающих описанным в § 7 критериям совместимости [44], Камина — Сокала [44] и Долло [44].

Выход из этого положения можно искать в двух направлениях — в создании эффективных приближенных методов построения филогенетических деревьев, см., например [6, 26—28, 42, 46, 57, 58, 60, 76, 78, 108] и в разработке альтернативных методов, не минимизирующих функционалы близости [29, 95, 97]. Впрочем, следует отметить, что граница между ними достаточно условна. Так, в топологические алгоритмы [29, 97], использующие степени соседства и отнесенные нами ко второй группе, могут рассматриваться как приближенные алгоритмы для топологического метода Фитча (см. § 6). С другой стороны, различные модификации метода деревьев Вагнера [57, 58, 60, 78], строящие аппроксимации так называемых деревьев Вагнера, задача построения которых *NP*-полна [43], могут быть отнесены и ко второй группе, так как их описания не содержат никаких апелляций к функционалам близости.

§ 9. Согласование филогенетических деревьев

При построении филогенетического дерева мы никогда не используем всей доступной нам информации о видах, филогению которых мы изучаем, а строим дерево по некоторой выборке из нее. Но биологическая эволюция характеризуется неравномерностью темпов преобразования органов и признаков, связанной с тем, что «эволюция частей организма, относящихся к разным координационным цепям, происходит относительно независимо» [4]. Поэтому филогенетические деревья одного и того же множества видов, построенные одним и тем же методом, но по разным выборкам данных, могут несовпадать, да и, как правило, не совпадают ни с истинным, ни друг с другом. Например, среди 39 минимальных и близких к ним деревьев 5 разных белков одних и тех же видов, построенных в [92], совпадают только 2.

Таким образом, построение истинного филогенетического дерева требует использования всей совокупности данных, но необходимый для этого объем вычислений делает такой путь нереальным. Эту трудность можно обойти, строя филогенетические деревья по отдельным группам данных и выбирая в качестве приближения к истинному дереву дерево, наилучшим образом согласованное с уже построенным множеством деревьев. Так же, как и в исходной задаче, основная трудность состоит в выборе содержательно обоснованного правила, а не в построении дерева по выбранному правилу.

Различные правила согласования, в основном для бинарных корневых деревьев, описаны в [30, 33, 84, 85, 90, 93, 105, 106]. Мы рассмотрим только работы [85, 90], показавшие, что задача выбора «разумного» правила согласования в некотором смысле неразрешима.

Определим правило согласования как функцию $C\colon \mathcal{T}_n^k \to \mathcal{T}_n$, сопоставляющему каждому набору $P = \{T_1, \ldots, T_k\}$ деревьев из $\mathcal{T}_n$, построенных по S k различными методами, согласованное с P дерево $C(P) \in$

$\in \mathcal{T}_n$, и рассмотрим, каким ограничениям должна удовлетворять функция C, чтобы ее можно было считать разумно определенной. Для этого нам понадобятся следующие определения.

Пусть $T \in \mathcal{T}_n$ и его множество висячих вершин есть S. Для любого $X \subseteq S$ обозначим через $T|_X$ поддерево T, чье множество висячих вершин совпадает с X. Для любого $V = \{x, y, z, w\} \subset S$ будем писать $(xy, zw) \in T$, если $T|_V$ совпадает с деревом на рис. 4, *а*. И, наконец, для любого $X \subseteq S$ и любых $P = \{T_1, \ldots, T_k\}$, $P' = \{T'_1, \ldots, T'_k\}$ запись $P|_X = P'|_X$ означает, что для всех i $T_i|_X = T'_i|_X$.

Рассмотрим следующее множество возможных ограничений, накладываемых на C.

А1. Аксиома единогласия. *Для любых* $\{x, y, z, w\} \subset S$ *и любого* $P = \{T_1, \ldots, T_k\}$ *из условия* $(xy, zw) \in T_i$, $i = 1, \ldots, k$, *следует, что* $(xy, zw) \in C(P)$.

А2. Аксиома независимости от не относящихся к делу альтернатив. *Для любого* $X \subseteq S$ *и любых* $P = \{T_1, \ldots, T_k\}$, $P' = \{T'_1, \ldots, T'_k\}$ *из* $P|_X = P'|_X$ *следует, что* $C(P)_X = C(P')_X$.

А3. Аксиома отсутствия диктатора. *Для любого* j *найдутся* $P = \{T_1, \ldots, T_k\}$ *и* $\{x, y, z, w\} \subset S$ *такие, что* $(xy, zw) \in T_j$, *но* $(xy, zw) \notin C(P)$.

Содержательный смысл этих аксиом ясен. Так, А1 означает, что если все методы дают одну и ту же филогению некоторой подгруппы $X \subseteq S$, то именно эту филогению для X должно давать и согласованное дерево $C(P)$. А2 означает, что для восстановления филогении по набору частичных филогений не требуется знать филогению видов, не входящих в рассматриваемую группу. Грубо говоря, А2 утверждает, что для восстановления филогении приматов нет необходимости знать филогению парнокопытных. И, наконец, невыполнение А3 означало бы существование такого метода j, что согласованное дерево совпадает с T_j и не зависит от вида деревьев, построенных другими методами, но в таком случае не было бы необходимости в задаче согласования.

Таким образом, каждая из этих аксиом содержательно обоснована. Более того, выражаемые ими свойства кажутся присущими любому разумному правилу согласования. Тем не менее, как показано в [85], для $n \geqslant 5$ не существует функции C: $\mathcal{T}_n^k \to \mathcal{T}_n$, удовлетворяющей всем трем аксиомам одновременно. Аналогичное утверждение справедливо и для корневых деревьев [90].

§ 10. Модель филогенетического дерева — критический анализ

Все описанные в предыдущих параграфах методы восстановления эволюционной истории исходят из справедливости основной модели филогении — модели филогенетического дерева. Напомним первые две аксиомы, лежащие в основе модели.

I. Аксиома дивергентности. *Эволюция носит дивергентный характер, т. е. любой вид имеет только одного непосредственного предка.*

II. Аксиома монофилии. *Любой реальный, а не сборный таксон имеет монофилитическое происхождение, т. е. все виды, в него входящие, происходят от одного общего предка, давшего ему начало.*

Со времен Дарвина справедливость этих утверждений признавалась подавляющим большинством биологов эволюционистов. Более того, в современной форме дарвинизма — синтетической теории эволюции (СТЭ) они по существу считались самоочевидными следствиями из определений вида и таксона. Развитие эволюционной теории привело к пересмотру этих взглядов. Обзор современного состояния СТЭ дан в [4]. Мы же ограничимся только обсуждением истинности аксиом I—II.

Аксиома I основана на предположении, что виды не могут обмениваться генетической информацией, и поэтому единственно возможный путь видообразования состоит в закреплении естественным отбором мутаций, возникающих внутри вида. На самом деле существуют и другие механизмы видообразования, а именно: гибридизация, симбиогенез и вирусная трансдукция. Рассмотрим их по порядку.

I. Видообразование путем гибридизации. Такой способ видообразования широко распространен у растений. Примерами могут служить слива, полученная гибридизацией терна и алычи, и табак, полученный гибридизацией двух видов дикого табака. Другие примеры см. в [3—5, 22]. Основным способом гибридного видообразования у растений является аллополиплоидия — удвоение числа хромосом при межвидовой гибридизации, обеспечивающее плодовитость гибридов даже у видов, различающихся по числу хромосом. «Аллополиплоидия — весьма обычный способ видообразования у покрытосеменных, папоротников и других групп растений. Согласно недавним оценкам, 47 % покрытосеменных и 95 % папоротникообразных (папоротники и близкие к ним виды) — полиплоиды. Большую часть полиплоидов составляют аллополиплоиды» [5]. Роль гибридизации в эволюции покрытосеменных и, в частности, злаков, подчеркивалась также в [9, 23].

Таким образом, уже рассмотрение видообразования путем гибридизации показывает, что помимо дивергентной эволюции существует и широко распространена (по крайней мере у растений) ретикулярная эволюция — филогенетический процесс чередования дивергенции и слияния разных ветвей одного филогенетического дерева.

II. Симбиогенез. Другие примеры ретикулярной эволюции доставляет симбиогенез — возникновение нового вида в результате далеко зашедшего симбиоза, при котором организмы-симбионты теряют способность существовать независимо друг от друга. Классическим примером симбиогенеза являются лишайники — симбиотический комплекс гриба и водоросли. Этот пример долгое время оставался изолированным, но в настоящее время приобрела широкую известность и считается общепринятой теория симбиотического происхождения эукариот из прокариот. Прокариоты — это клеточные организмы с недифференцированным ядром, а эукариоты — это высшие клеточные организмы с дифференцированным ядром. В частности, к прокариотам относятся синезеленые водоросли и большинство бактерий, а к эукариотам — все многоклеточные организмы. Согласно теории симбиогенеза эукариотические клетки возникли в результате кооперации первоначально независимых прокариотических клеток, объединившихся в определенном порядке, и основные клеточные органеллы — митохондрии, фотосинтезирующие пластиды и двигательные органеллы произошли от различных групп свободно живущих прокариот, причем приобретение эукариотами клеточных органелл происходило неоднократно, и в этом участвовали разные виды-партнеры [12]. Например, митохондрии грибов и млекопитающих имеют независимое происхождение [80]. Дальнейшие подробности см. в упомянутом выше обзоре [4] и, в первую очередь, в монографии [12].

III. Вирусная трансдукция. И, наконец, ретикулярная эволюция может происходить за счет вирусной трансдукции — переноса генов от одного вида к другому вирусами. Вирусная трансдукция играет большую роль в эволюции бактерий, но ее роль в эволюции высших организмов пока неясна. Известно только 4 случая трансдукции у эукариот [83], но высказывались мнения о роли трансдукции и в эволюции высших растений [11].

Таким образом, вопреки аксиоме I, в природе помимо дивергентной эволюции существует и широко распространена и ретикулярная эволюция. Уже этого достаточно, чтобы осознать, что реальная филогенетиче-

ская схема может содержать циклы и поэтому не всегда представима деревом. Ситуация осложняется еще тем, что принцип монофилии в той форме, в какой он сформулирован в аксиоме II, также неверен.

Это следует уже из того, что, как указано выше, такие таксоны, как лишайники и эукариоты, возникли в результате симбиогенеза и, следовательно, не являются монофилитическими. Но дело не только в симбиогенезе. Согласно современным представлениям, такие классы, как рыбы, пресмыкающиеся и млекопитающие, не являются монофилитическими в смысле аксиомы II, а представляют собой уровни организации хордовых, прорыв на которые был осуществлен разными эволюционными линиями независимо. Например, разные отряды млекопитающих произошли от разных видов зверозубых пресмыкающихся — териодонтов [21]. Такая же ситуация и с растениями. Так ни покрытосеменные в целом [9], ни такой их отряд, как злаки [23], не являются монофилитическими таксонами.

В общетеоретическом плане осознание этих фактов привело к замене концепции монофилии в узком смысле слова (т. е. в смысле аксиомы II) на концепцию монофилии в широком смысле слова (см. [22]), но для наших целей важен только следующий отсюда вывод, что филогенетическая схема даже тех таксонов, эволюция которых была строго дивергентной, не всегда может быть представлена деревом. Например, филогенетическая схема млекопитающих будет представлять собой лес, компонентами связности которого будут филогенетические деревья отдельных отрядов млекопитающих, и мы сможем достроить ее до связного филогенетического дерева только если объединим ее с филогенетической схемой отряда териодонтов (в предположении, что эта схема сама является деревом).

Таким образом, результаты обсуждения степени обоснованности модели приводят к выводу, что «представление филогенеза в виде дихотомического древа есть грубое, хотя во многих случаях и вынужденное упрощение» [4], и необходима разработка математических методов восстановления филогений, не представимых филогенетическим деревом.

В заключение отметим, что хотя мы анализировали степень содержательной обоснованности модели филогенетического дерева только для биологии, сделанные выводы сохраняют силу и для других областей. Так, рассматриваемые в генетике филогенетические схемы популяций могут иметь циклы уже потому, что в отличие от биологических видов, скрещивание которых затруднено, особи из разных популяций одного и того же вида могут скрещиваться свободно.

§ 11. Заключение. Проблемы и перспективы

Критический анализ модели филогенетического дерева, проведенный в § 10, может создать впечатление о неприменимости модели и ошибочности всех филогений, построенных на ее основе. Но это впечатление обманчиво. Проведенный анализ не отменяет модели, а только указывает границы ее применимости.

Во-первых, ретикулярная эволюция распространена в природе широко, но не повсеместно. Если, как показано выше, ретикулярная эволюция играет большую роль в филогении растений и в начальных этапах филогении эукариот, то эволюция животных преимущественно дивергентна, а эволюция высших животных только дивергентна. Поэтому модель филогенетического дерева неприменима только к тем таксонам, в филогении которых большую роль играет ретикулярная эволюция, и полностью применима к монофилитическим таксонам, эволюция которых носила дивергентный характер. Более того, модель филогенетического дерева применима и к тем таксонам, которые, как млекопитающие, не

являются монофилитическими в узком смысле слова, но возникли в результате дивергентной эволюции. Единственная возникающая здесь проблема состоит в трудности правильной интерпретации нижних ярусов построенного дерева, и в этом случае необходимо привлечение дополнительных биологических и палеонтологических данных, чтобы решить, отражают ли эти ярусы истинную филогению видов, давших начало этому таксону, или нет.

Во-вторых, модель филогенетического дерева может использоваться как тест для проверки утверждений о роли ретикулярной эволюции в филогении данного таксона. В этих случаях именно противоречие между филогенетическим деревом, построенным в предположении, что эволюция таксона носила дивергентный характер, и данными о его филогении, вытекающими из общебиологических соображений, и может рассматриваться как доказательство роли ретикулярной эволюции в филогении данного таксона. Например, именно тот факт, что в филогенетическом дереве 5*S*-РНК ([98], см. также [12, с. 61]) 5*S*-РНК пластид ряски гораздо ближе к РНК коккоидной цианобактерии, чем к РНК цитоплазмы ржи и хлореллы, рассматривается как сильный аргумент в пользу теории симбиогенеза. Более того, анализ филогенетических деревьев, построенных по молекулярным данным, позволил сделать вывод не только о симбиотическом происхождении эукариот, но и о путях симбиогенеза, например, о том, что эукариоты получили генетический материал в результате симбиоза по крайней мере трех отдельных бактериальных линий [45] и что митохондрии грибов и млекопитающих имеют независимое бактериальное происхождение [80].

Таким образом, пересмотр старых филогенетических концепций не перечеркивает математических методов восстановления филогенеза, разработанных в рамках модели филогенетического дерева, а только указывает границы их применимости, ставя тем самым задачу разработки новых более общих методов, применимых к филогениям, не представимым деревьями.

Не касаясь всего комплекса связанных с этим проблем, закончим обзор обсуждением математической стороны вопроса.

Проблема восстановления филогении, не представимой деревом, может быть сформулирована следующим образом. Задано некоторое множество S вершин искомой филогенетической схемы и матрица эволюционных расстояний видов из S. Требуется восстановить искомую филогенетическую схему, т. е. построить взвешенный граф G, содержащий все вершины из S и минимизирующий некоторый функционал $F(D, \Delta)$ — меру близости матрицы эволюционных расстояний Δ к матрице D расстояний между вершинами из S в G.

Рассмотрим сначала идеальный случай, когда Δ является метрикой. Тогда эта задача сводится к задаче построения взвешенного графа G, содержащего все вершины из S, и такого, что матрица расстояний между вершинами из S совпадает с Δ. Такой граф будем называть реализацией Δ. Как известно, любая матрица расстояний реализуется взвешенным графом [74], и, более того, любая целочисленная матрица расстояний реализуется невзвешенным графом [20]. Но одна и та же матрица может иметь несколько реализаций. В частности, любая матрица расстояний $\Delta = \|\delta_{ij}\|$ может быть реализована полным графом, в котором любые две вершины i, j соединены ребром длины δ_{ij}. Ясно, что такая реализация не может иметь биологического смысла. Поэтому необходимо сузить класс допустимых реализаций. Один из возможных подходов состоит в рассмотрении только оптимальных реализаций, т. е. реализаций, удовлетворяющих следующим условиям:

1. длины всех ребер G положительны;
2. степень всех вершин G, не входящих в S, больше двух;

3. сумма длин ребер любой реализации, отличной от G, не меньше суммы длин ребер G.

Такой выбор допустимых реализаций находится в соответствии с принципом максимальной экономии, минимизирующим общее число эволюционных событий. Другое достоинство этого подхода состоит в том, что если Δ реализуется деревом, то эта реализация оптимальна [74]. Поэтому для матриц, реализуемых деревом, рассматриваемый подход приводит к тем же результатам, что и методы, основанные на модели филогенетического дерева.

Практическая реализация этой идеи требует знания свойств оптимальных реализаций и алгоритмов их построения. На настоящий момент об этих вопросах известно не слишком много. Доказано, что любая конечная метрика имеет оптимальную реализацию [77]. Это утверждение менее очевидно, чем кажется, так как для квазиметрик, т. е. для функций расстояния, удовлетворяющих всем аксиомам метрики, кроме условия симметричности, это уже не так [101]. Некоторые свойства оптимальных реализаций установлены в [74, 77, 91, 102]. В частности, известно, что оптимальные реализации не могут содержать треугольников [74], и известны необходимые и достаточные условия единственности для оптимальных реализаций, множество вершин которых совпадает с S [77].

Обобщая этот подход на произвольные матрицы эволюционных расстояний, возможно, не имеющие реализаций, сведем задачу построения филогенетической схемы по матрице Δ эволюционных расстояний некоторого подмножества S ее вершин к задаче построения связного взвешенного графа G, удовлетворяющего следующим условиям:

1. все вершины S принадлежат G;
2. длины всех ребер G положительны;
3. степени всех вершин G, не входящих в S, больше двух;
4. для любых двух вершин i, j из S $d_{ij} \geqslant \delta_{ij}$, т. е. для любых двух вершин из S длина кратчайшей простой цепи, их соединяющей, не меньше эволюционного расстояния между ними;
5. матрица расстояний между вершинами из S в G так близка к Δ, как только это возможно;
6. сумма длин ребер G не больше суммы длин ребер любого связного взвешенного графа, удовлетворяющего условиям 1—5.

Неконструктивное условие 5 существенно, так как отказ от него превращает сформулированную задачу в задачу построения минимального филогенетического дерева, а конкретизировать его можно разными способами. Например, можно потребовать, чтобы сумма $\sum(d_{ij} - \delta_{ij})^2$ была минимальна по всем G, удовлетворяющим условиям 1—4. Возможно использование и любых других мер близости матриц. Вопрос же о том, какая из возможных мер близости наиболее полно отражает специфику задачи, требует, как и вся проблема в целом, дополнительной разработки и может быть решен только на основе анализа конкретных филогенетических схем, построенных по реальным данным.

СПИСОК ЛИТЕРАТУРЫ

1. Айала Ф. Дж. Механизмы эволюции // Ж. Всесоюз. хим. о-ва.— 1980.— Т. 25.— С. 227—294.
2. Айала Ф. Дж. Популяционная генетика.— М.: Мир, 1986.
3. Бреславец Л. П. Полиплоидия в природе и опыте.— М.: Изд-во АН СССР, 1963.
4. Воронцов Н. Н. Синтетическая теория эволюции: ее источники, основные постулаты и нерешенные проблемы // Ж. Всесоюз. хим. о-ва— 1980.— Т. 25.— С. 295—315.
5. Грант В. Эволюция организмов.— М.: Мир, 1980.

6. Жарких А. А. Алгоритм построения филогенетических древ по аминокислотным последовательностям // Математические модели эволюции и селекции.— Новосибирск: ИЦиГ СО АН СССР, 1977.— С. 7—52.
7. Зарецкий А. К. Построение дерева по набору расстояний между висячими вершинами // УМН.— 1965.— Т. 20, № 6.— С. 94—96.
8. Зыков А. А. Теория конечных графов.— Новосибирск: Наука, 1969.
9. Красилов В. А. Предки покрытосеменных // Проблемы эволюции. Т. 4. Современные проблемы эволюции.— Новосибирск: Наука, 1975.— С. 76—102.
10. Красилов В. А. Современные проблемы соотношения филогении и систематики // Зоология позвочных. Проблемы теории эволюции.— М.: ВИНИТИ, 1975.— С. 118—147.
11. Красилов В. А. Эволюция и биостратиграфия.— М.: Наука, 1977.
12. Маргелис Л. Роль симбиоза в эволюции клетки.— М.: Мир, 1983.
13. Миркин Б. Г., Родин С. Н. Графы и гены.— Новосибирск: Наука, 1977.
14. Омельянчук Л. В., Колчанов Н. А. Алгоритм построения аддитивных деревьев по набору гомологичных последовательностей. Достоверность восстановления филогении // Вычисл. системы.— 1985.— № 112.— С. 46—55.
15. Орлов О. Ю. Об эволюции цветного зрения у позвочных // Проблемы эволюции. Т. 2.— Новосибирск: Наука, 1975.— С. 69—94.
16. Пасеков В. П. Генетические расстояния // Общая генетика. Т. 8.— М.: ВИНИТИ, 1981.— С. 3—75.
17. Ратнер В. А. Молекулярно-генетические системы управления.— Новосибирск: Наука, 1975.
18. Ратнер В. А. Математические модели в теории молекулярной эволюции // Математическая биология и медицина. Т. 1.— М.: ВИНИТИ, 1978.— С. 210—257.
19. Смоленский Е. Ф. Об одном способе линейной записи графов // ЖВМиМФ.— 1962.— Т. 2.— С. 371—372.
20. Стоцкий Э. Д. О вложении конечных метрик в графы // Сиб. мат. ж.— 1964.— Т. 5.— С. 1203—1206.
21. Татаринов Л. П. Морфологическая эволюция териодонтов и общие вопросы филогенетики.— М.: Наука, 1977.
22. Тимофеев-Ресовский Н. В., Воронцов Н. Н., Яблоков А. В. Краткий очерк теории эволюции.— М.: Наука, 1977.
23. Цвелев Н. Н. О происхождении и основных направлениях эволюции злаков // Проблемы эволюции. Т. 4.: Современные проблемы эволюции.— Новосибирск: Наука, 1975.— С. 107—117.
24. Чумаков К. М., Юшманов С. В. Принцип максимального топологического подобия в молекулярной систематике // Мол. генетика, микробиол. вирус.— 1988.— Т. 3.— С. 3—8.
25. Юшманов С. В. Методы теории графов в эволюции. Построение филогенетических схем // Математическая кибернетика и ее приложения к биологии.— М.: Изд-во МГУ, 1987.— С. 101—140.
26. Юшманов С. В. Восстановление филогенетического дерева по поддеревьям, порожденным четверками его висячих вершин // Математическая кибернетика и ее приложения к биологии. М.: Изд-во МГУ, 1987.— С. 141—147.
27. Юшманов С. В., Чумаков К. М. Деревья локальной экономики: топологический подход к задаче построения деревьев максимальной экономии // Ж. эвол. биохим. и физиол.— 1989.— Т. 25, № 4.— С. 532—535.
28. Юшманов С. В., Чумаков К. М. Алгоритмы построения деревьев максимального топологического подобия // Мол. генетика, микробиол. вирус.— 1988.— Т. 3.— С. 9—15.
29. Abdi H., Barthelemy J.-P., Luong N. X. Thee representations of associative structures in semantic and episodic memory research // Trends in mathematical psychology.— Amsterdam: Elsilver, 1984.— P. 3—33.
30. Adams E. N. Consensus techniques and comparison of taxonomic trees // Syst. Zool.— 1972.— V. 21.— P. 390—397.
31. Altschul S. F., Erickson B. W. Optimal sequence alignment using affine gap costs // Bull. Math. Biol.— 1986.— V. 48.— P. 603—616.
32. Altschul S. F., Lipman D. J. Trees, stars, and multiple biological sequence alignment // SIAM J. Appl. Math.— 1989.— V. 49.— P. 197—209.
33. Barthelemy J.-P. Tresholded consensus for n-trees // J. Cpassif.— 1988.— V. 5.— P. 229—236.
34. Barthelemy J-P., Luong N. X. Sur la topologie d un arbe phylogenetique: aspects theoretigues, algorithmes et applications a l-analyse de donnes textuelles // Math. et Sci. Hum.— 1987.— V. 25, № 100.— P. 57—80.
35. Bandelt H-J., Dress A. Reconstracting the shape of a tree from observed dissimiarlity data // Advances in Appl. Math.— 1986.— V. 7.— P. 309—343.
36. Beyer W. A., Stein M. L., Smith T. E., et al. A molecular sequence metric and evolutionary trees // Math. Biosci.— 1974.— V. 19.— P. 9—25.

37. Buneman P. The recovery of trees from measures of dissimilarity // Mathematics in archaeolodgical and historical sciences. Edinburg: University Press, 1971.— P. 387—395.
38. Camin J. H., Sokal R. R. A method for deducting branching sequences in phylogeny // Evolution.— 1965.— V. 19.— P. 311—326.
39. Carrillo H., Lipman D. J. The multiple sequence alignment problem in biology // SIAM J. Appl. Math.— 1988.— V. 48.— P. 1073—1082.
40. Cavalli-Sforza L. L., Edwards W. F. Phylogenetic analysis: models and estimation procedures // Evolution.— 1967.— V. 32.— P. 550—570.
41. Colonius H., Schulze H. H. Tree structures for proximity data // Briitish J. Math. Statist. Psych.— 1981.— V. 34.— P. 167—180.
42. Canningham J. P. Free trees and bidirectional trees as representations of psycholodgical distance // J. Math. Psychol.— 1978.— V. 17.— P. 165—188.
43. Day W. H. E. Computationally difficult parsinomy problems in phylogenetic systematics // J. Theor. Biol.— 1983.— V. 103.— P. 429—438.
44. Day W. H. E., Johnson D. S., Sankoff D. The comutational complexity of inferring rooted phylogenies by parsinomy // Math. Biosci.— 1986.— V. 81.— P. 33—42.
45. Dayhoff M. O., Schwartz R. M. Evolution of the rhodospirillaceal and mitochondria: a view based on sequence data // Origin life. Dordrecht: e. a., 1981.— P. 559—566.
46. De Soete G. A leas squares algorithms for fitting additive trees to proximity data // Psychometrica.— 1983.— V. 48.— P. 621—626.
47. Dobson A. J. Unrooted trees for numerical taxonomy // J. Appl. Probab.— 1974.— V. 11.— P. 32—42.
48. Dress A., Haeseler A., Krueger M. Reconstructing phylogenetic trees using variants of the «four-point-condition» // Studlen zur Klassif.— 1986.— V. 17.— P. 299—305.
49. Edwards W. F., Cavalli-Sforza L. L. The reconstruction of evolution // Am. Hum. Genet.— 1963.— V. 27.— P. 105.
50. Edwards W. F., Cavalli-Sforza L. L. A method for cluster analysis // Biometrics.— 1965.— V. 21.— P. 362—375.
51. Estabrook G. F., Johnson C. S. Jr., McMorris F. R. An idealized concept of the true cladistic character // Math. Biosci.— 1975.— V. 23.— P. 263—272.
52. Estabrook G. F., Johnson C. S. Jr., McMorris F. R. An algebraic analysis of cladistic characters // Discrete Math.— 1976.— V. 16.— P. 141—147.
53. Estabrook G. F., Johnson C. S. Jr., McMorris F. R. A mathematical foundation for the analysis of cladistic character compatibility // Math. Biosci.— 1976.— V. 29.— P. 181—187.
54. Estabrook G. F., Landrum L. A simple test for the possible simultaneous evolutionary divergence of two aminoacid positions // Taxon.— 1975.— V. 24.— P. 609—613.
55. Estabrook G. E., McMorris F. R. When are two qualitative taxonomic characters compatible // J. Math. Biol.— 1977.— V. 4.— P. 195—200.
56. Estabrook G. F., Meacham C. A. How to determine the compatibility of undirected character state trees // Math. Biosci.— 1979.— V. 46.— P. 251—256.
57. Farris J. S. Methods for computing Wagner trees // Syst. Zool.— 1970.— V. 19.— P. 83—92.
58. Farris J. S. Estimating phylogenetic trees from distance matrices // Am. Nat.— 1972.— V. 106, № 951.— P. 645—648.
59. Farris J. S. Phylogenetic analysis under Dollo's law // Syst. Zool.— 1977.— V. 26.— P. 77—88.
60. Farris J. S., Kluge A. G., Eckardt M. J. A numerical approach to phylogenetic systematics // Syst. Zool.— 1970.— V. 19.— P. 172—189.
61. Fickett J. W. Fast optimal alignment // Nucl. Acids Res.— 1984.— V. 12.— P. 175—180.
62. Fitch W. M. Toward defining the course of evolution: minimum chandge for a specified tree topology // Syst. Zool.— 1971.— V. 20.— P. 406—416.
63. Fitch W. M. Toward finding the tree of maximum parsinomy // Proc. Eighth. International Conference on numerical taxonomy. San francisco: Freeman, 1975.— P. 189—230.
64. Fitch W. M. A non-sequential method for constructing trees and hierarchical classifications // J. Mol. Evol.— 1981.— V. 18.— P. 60—67.
65. Fitch W. M., Margoliash E. Construction of phylogenetic trees // Science.— 1967.— V. 155, № 3760.— P. 279—284.
66. Fitch W. M., Smith T. F. Implications of mimimal lenght trees // Syst. Zool.— 1982.— V. 31.— P. 68—75.
67. Fitch W. M., Smith T. F. Optimal sequence alignments // Proc. Nat. Acad. Sci. USA.— 1983.— V. 80.— P. 1382—1386.
68. Foulds L. R., Graham R. L. The Steiner problem in phylogeny is NP-complete // Advances in Appl. Math.— 1982.— V. 3.— P. 43—49.

69. Foulds L. R., Penny D., Hendy M. D. A graph theoretic approach to the development of minimal phylogenetic trees // J. Mol. Evol.— 1979.— V. 13.— P. 127—149.
70. Foulds L. R., Penny D., Hendy M. D. A general approach to proving the minimality of phylogenetic trees illustrated by a set of 23 vertebrates // J. Mol. Evol.— 1979.— V. 13.— P. 151—166.
71. Foulds L. R., Robinson R. W. Determing the asymptotic of phylogenetic trees // Combinatorial Mathematics. VII.— Berlin: Springer, 1980.— P. 110—126.
72. Friedlaender J. S. et al. Biological divergences in south-central Bougainville // Amer. J. Human. Genet.— 1971.— V. 23.— P. 253—270.
73. Graham R. I., Foulds L. R. Unlikelihood that minimal phylogenies for realistic biological study can be constructed in reasonable computational time // Math. Biosci.— 1982.— V. 60.— P. 133—142.
74. Hakimi S. L., Yau S. S. Distance matrix of a graph and its realizability // Quart. Appl. Math.— 1965.— V. 22.— P. 305—317.
75. Hartigan J. A. Minimum mutation fits to a given tree // Biometrics.— 1973.— V. 29.— P. 53—65.
76. Hendy M. D., Penny D. Branch and bound algorithms to determine minimal evolutionary trees // Math. Biosci.— 1982.— V. 59.— P. 277—290.
77. Imrich W., Simoes-Pereira J. M. S. On optimal embeddings of metrics in graphs // J. Combin. Theory.— 1984.— V. B36.— P. 1—15.
78. Jensen R. J. Wagner networks and Wagner thees: A presentation of methods for estimating most parsimonious solutions // Taxon.— 1981.— V. 30.— P. 576—590
79. Kidd K. K., Sgaramella-Zonta L. Phylogenetic analysis: concepts and methods// Am. J. Hum. Genet.— 1971.— V. 23.— P. 235—252.
80. Kuntzel H., Kochel H. G. Evolution of rRNA and origin mitochondria // Nature.— 1981.— V. 293, № 5835.— P. 751—753.
81. Le Quesne W. J. A method of selection of characters in numerical taxonomy // Syst. Zool.— 1969.— V. 18.— P. 201—205.
82. Le Quesne W. J. The uniqely evolved character concept and its cladistic application // Syst. Zool.— 1974.— V. 23.— P. 513—517.
83. Lewin R. Can genes jump between eukaryotis species // Science.— 1982.— **V. 217,** № 4554.— P. 42—43.
84. Margush T., McMorris F. R. Consensus n-trees // Bull. Math. Biol.— 1981.— V. 43.— P. 239—244.
85. McMorris F. R. Axioms for consensus functions on undirected phylogenetic trees // Math. Biosci.— 1985.— V. 74.— P. 17—21.
86. Meacham C. A. A manual method for character compatibility analysis // Taxon.— 1981.— V. 30.— P. 591—600.
87. Murata M., Richardson J. S., Sussmann J. L. Simultaneous comparision of three protein sequences // Proc. Nat. Acad. Sci. USA.— 1985.— V. 82.— P. 7657—7661.
88. Najock D. Principles and modifications of local genealogical algorithms in textual history // Computers and humanities.— 1980.— V. 14.— P. 171—179.
89. Needleman S. B., Wunsh C. P. A general method applicatible to the search for similarities in aminoacid sequences of two ptoteins // J. Mol. Biol.— 1970.— V. 48.— P. 443—453.
90. Neumann D. A. Faithful consensus methods for n-trees // Math. Biosci.— 1983.— V. 63.— P. 271—287.
91. Patrinos A. N., Hakimi S. L. The distance matrix of a graph and its realization // Quart. Appl. Math.— 1972.— V. 30.— P. 255—269.
92. Penny D., Foulds L. R., Hendy M. D. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences // Nature.— 1982.— V. 297, № 5863.— P. 197—200.
93. Rohlf F. J. Consensus indices for comparing classifications // Math. Biosci.— 1982.— V. 59.— P. 113—144.
94. Rohlf F. J. A note of minimal length trees // Syst. Zool.— 1984.— V. 33.— P. 341—343.
95. Saitou N., Nei M. The neighbor — joining method: a new method for reconstructing phylogenetic trees // Mol. Biol. Evol.— 1987.— V. 4.— P. 406—426.
96. Sankoff D., Gedergren R. J. Simultaneous comparision of three or more sequences related by a tree // Strings and macromolecules: The theory and practice of sequence comparison. Reading, MA, Addison — Wesley, 1983.— P. 253—263.
97. Sattath S., Tversky A. Additive similarity trees // Psychometrica.— 1977.— V. 42.— P. 319—345.
98. Schwartz R. M., Dayhoff M. O. Origins of prokaryotes, eukariotes, mitochondria and chloroplasts // Science.— 1978.— V. 199, № 4327.— P. 395—403.
99. Sellers P. H. On the theory and computation on evolutionary distances // SIAM J. Appl. Math.— 1974.— V. 26.— P. 787—793.
100. Simoes-Pereira J. M. S. A note on the tree realizability of a distance matrix // J. Combin. Theory.— 1969.— V. 6.— P. 303—310.

101. Simoes-Pereira J. M. S. A note on optimal and suboptimal digraph realizations of quasidistance matrices // SIAM J. Algebraic Discrete Methods.— 1984.— V. 5.— P. 117—132.
102. Simoes-Pereira J. M. S., Zamfirescu C. Submatrices of non-treerealizable distance matrices // Linear Algebra Appl.— 1982.— V. 44.— P. 1—17.
103. Smith T. F., Waterman M. S., Fitch W. M. Comparative biosequence metrics // J. Mol. Evol.— 1981.— V. 18.— P. 38—46.
104. Soto M. A., Toha J. Dissimilar rates in molecular evolution // Orig. Life.— 1984.— V. 14.— P. 637—642.
105. Stinebrickner R. *s*-consensus trees and indices // Bull. Math. Biol.— 1984.— V. 46.— P. 923—925.
106. Stinebrickner R. J. An extension of intersection methods from trees to dendrograms // Syst. Zool.— 1984.— V. 33.— P. 381—386.
107. Wagner H. J. The minimum number of mutations in evolutionary network // J. Theor. Biol.— 1981.— V. 91.— P. 621—636.
108. Waterman M. S., Smith T. F., Sing M. et al. Additive evolutionary trees // J. Theor. Biol.— 1977.— V. 64.— P. 199—213.
109. Wilson A. C., Carlson S. S., White T. J. Biochemical evolution // Annu. Rev. Biochem.— 1977.— V. 46.— P. 573—639.
110. Wilson E. O. A consistency test for phylogenies based on contemporaneous species // Syst. Zool.— 1965.— V. 14.— P. 214—220.