Информационные процессы, Том 1, № 1, 2001, стр. 64–77.

ТЕОРИЯ И МЕТОДЫ ОБРАБОТКИ ИНФОРМАЦИИ

Об одном способе построения деревьев эволюции видов по множественным генетическим данным

М.С.Гельфанд[*], В.В.Вьюгин[**], В.А.Любецкий[**]

[*] *Государственный научный центр ГосНИИГенетика, Москва, Россия*

[**] *Институт проблем передачи информации, Российская академия наук, Москва, Россия*

Поступила в редколлегию 3.01.2001

Аннотация—Хорошо известно, что филогенетические деревья, построенные по различным семействам белков, часто не совпадают. Причиной этого являются как неточности в построении деревьев генов, так тот принципиальный факт, что деревья генов могут отличаться от дерева видов из-за дупликаций и потерь генов, горизонтальных переносов и т.п. Поэтому возникает задача построения "консенсусного"дерева видов, наилучшим образом согласованного с заданным множеством деревьев генов. Один из методов построения подобного дерева видов рассмотрен в настоящей работе. Он отличается от существующих методов, в частности, тем, что в нем учитывается не только топология исходных деревьев видов, но и надежность отдельных ветвей и, тем самым, явно учитывается возможность ошибок в деревьях генов, вызванная отсутствием надежных моделей эволюции последовательностей, неравномерностью эволюции в разных семействах генов и таксономических группах, а также и другими причинами.

1. ПОСТАНОВКА ЗАДАЧИ

Реконструкция филогенетического дерева видов — одна из основных проблем эволюционной биологии (теории эволюции). Поскольку палеонтологическая летопись часто бывает неполна и противоречива, это невозможно сделать без применения молекулярных методов. При этом современные последовательности биополимеров используются, в рамках той или иной модели молекулярной эволюции, для построения дерева генов. Однако более или менее реалистические эволюционные модели приводят к непреодолимым вычислительным сложностям, и поэтому приходится применять приближенные методы. Обзор методов построения деревьев генов см., например, в [1]–[13].

Для построения филогенетического дерева можно использовать различные белки и соответствующие им последовательности ДНК. Практика показывает, что очень часто эти деревья могут различную структуру. Поэтому возникает задача построения "консенсунсного" дерева видов, наилучшим образом согласованного с данными деревьями генов.

Однако даже в отсутствие ошибок, связанных с выбором неадекватной модели молекулярной эволюции, дерево генов не тождественно дереву видов. Причинами этого являются дупликации генов, потери генов и горизонтальный перенос. Действительно, если в предковом виде происходит дупликация гена, изменения в каждой из копий происходят независимо от другой, и обе копии наследуются потомками. В результате *дупликации гена* в предковом виде два вида могут нести гены a_1 и a_2, которые дивергировали еще до расщепления видов, вследствие чего различия между генами a_1 и a_2 из одного генома превышают различия между ортологичными генами, например типа a_1, из разных геномов. Если теперь в каких-либо потомках произойдет *потеря гена*, точнее, какой-то одной из двух копий гена, скажем a_1, то сравнение единственного оставшегося представителя семейства a_2 из этого генома с выборкой генов типа a_1 из других геномов приведет к искажению топологии дерева "видов".

По-видимому, впервые постановка задачи сравнения филогении генов и видов была рассмотрена в [3]. Построение дерева видов на основе модели дупикаций и потерь рассматривалось в [4], [8], [10].

На первом этапе обычно решается задача сравнительного анализа древовидной генетической и видовой структур — находится наименьшее число (или стоимость) элементарных операций (имеющих разумную биологическую интерпретацию), наиболее экономно переводящих дерево генов в дерево видов. Этот метод основан на естественном предположении, что дерево генов, наименее отличающееся от дерева видов, вместе с набором биологически значащих операций, переводящих одно в другое, наиболее объективным образом отражает эволюцию генов. Таким образом, при выводе филогенетического дерева мы отдаем предпочтение тому дереву, для которого стоимость операций перевода всех деревьев генов минимальна. В данной работе вводится новая более сложная функция стоимости различия деревьев, которая учитывает степень достоверности различных частей дерева генов. Предполагается, что это позволит включить в "консенсусное" дерево правильно установленные узлы, тогда как те узлы частных деревьев по отдельным генам, которые недостоверны и ошибочно отражают филогению видов, будут отброшены. Также будут рассмотрены алгоритмы, нахождения дерева видов, минимально расходящегося с данным набором деревьев генов.

В качестве иллюстрации применяемых алгоритмов на примере множественных генетических данных для конкретных семейств организмов будут построены деревья эволюции видов, для которых функция стоимости различия достигает локального минимума.

2. ТАКСОНОМИЧЕСКИЕ ДЕРЕВЬЯ И ИХ ГОМОМОРФИЗМЫ

Рассмотрим основные события, которые объясняют рассогласованность структуры дерева генов и дерева видов, обычно рассматриваемые в литературе. Это дупликация гена и потеря гена. Еще одна операция, горизонтальный перенос, не характерна для высших эукариот и не будет рассматриваться в данной работе.

Рассматриваются следующие события, происходящие с генами в процессе исторического развития вида. Ген необходимым образом разделяется во время видообразования, т.е. разделения вида на два вида-потомка. При этом каждый из генов-потомков переходит в соответствующий вид-потомок; в последующем, уже независимым образом, происходит накопление изменений в этих генах. Такая пара генов называется *ортологической*.

Событие другого рода заключается в разделении одного гена a на две идентичные копии, скажем, a_1 и a_2 внутри одного вида и последующем независимом развитии этих копий внутри этого вида и его потомков. Такая пара генов называется *паралогической*. С этим событием можно связать событие *дупликации гена*. Различие между деревом генов G и деревом видов S может быть объяснено дупликацией некоторого гена a в корне на две копии a_1 и a_2; после этого при последующем видообразовании происходит разделение потомков каждой из копий. Может так случиться, что потомок копии a_2 не будет обнаружен в виде - потомке. Для объяснения этого явления приходится рассматривать событие другого рода. Ген, присутствующий внутри данного вида в некотором узле, может не иметь потомков в листьях, находящихся под этим узлом. В этом случае говорят, что произошло событие *потери гена*. Это событие произошло в том узле соответствующей ветви, в котором впервые отсутствуют потомки данного гена.

В этом разделе определяется операция отображения дерева генов в дерево видов. Отображения деревьев генов в деревья видов можно использовать для:

- восстановления процесса эволюции отдельных генов во время развития генов;
- выявления степени достоверности отдельных частей видового дерева на основе их согласованности с различными деревьями генов;
- вычисления стоимости различия дерева генов и дерева видов.

Определим теперь отображение дерева генов в дерево видов.

Пусть задано множество I операционных таксономических единиц OTE. Без потери общности можно считать, что $I = \{0, 1, 2, \ldots N-1\}$. Даны два дерева: S — дерево видов и G — дерево генов.

Каждое дерево имеет N листьев и каждый лист помечен одноэлементным подмножеством I, причем различные листья помечены попарно непересекающимися подмножествами. Корень дерева помечен множеством I. Другие узлы каждого из деревьев помечены подмножествами I, причем, если узел помечен подмножеством $s \subset I$, а его потомок помечен $q \subset I$, то $q \subset s$. Мы будем отождествлять узел и множество, его помечающее. Из определения следует, что несравнимые узлы имеют пустое пересечение. Таким образом, каждое такое таксономическое дерево состоит из кластеров, разделяющих элементы I.

По дереву генов G и дереву видов S однозначно строится отображение

$$\alpha : G \to S$$

следующим образом: для каждого $g \in G$ значение $\alpha(g)$ определяется как минимальное по теоретико-множественному включению $s \in S$ такое, что $g \subset s$.

Для любого внутреннего узла дерева или его корня g, через cg обозначается его левый потомок, а через $\hat{c}g$ — его правый потомок. Если g не корень, то pg обозначает отца g.

Рассмотрим основные характеристики, которые задают отличие гомоморфизма $\alpha(g)$ от изоморфизма. Это дупликации в области определения, т.е. узлы g и g', такие, что g' является сыном g и $\alpha(g) = \alpha(g')$, а также пропуски в области значений, т.е. наличие узлов s, таких что $\alpha(g) \subset s \subset \alpha(pg)$. Пара (g, s), где $s = \alpha(g)$ называется *односторонней дупликацией*, если $\alpha(g) = \alpha(cg)$ или $\alpha(g) = \alpha(\hat{c}g)$, но не одновременно. Если выполнены два этих условия одновременно, то (g, s) называется *двухсторонней дупликацией*.

Множество односторонних дупликаций обозначается $O(G, S)$.

Узел $s \in S$ называется g-промежуточным, если он находится строго между $\alpha(g)$ и $\alpha(pg)$. Пусть I_g — множество всех g-промежуточных узлов. Общее множество промежуточных узлов определяется

$$M(S, G) = \cup_{g \in G} I_g.$$

В [6] вводится мера различия деревьев G и S - функция потерь

$$c(G, S) = |M(G, S)| + |O(G, S)|,$$

которая является количественной характеристикой степени отличия гомороморфизма $\alpha(g)$ от изоморфизма, а тем самым, степенью отличия деревьев G и S.

Выделение односторонних дупликаций и промежуточных узлов для оценки потерь обосновывается тем, что с каждым из этих явлений можно связать одну потерю гена (см. [7]). Таким образом, мера различия деревьев определяется суммарным числом потерь генов, необходимых для объяснения этого различия.

Большинство алгоритмов построения деревьев по генетическим последовательностям вычисляют также численные характеристики, выражающие время, прошедшее при переходе от одной последовательности к другой. Эта величина выражает вычисленную некоторым способом долю различия между выравненными последовательностями. В этом случае она может интерпретироваться как время при предположении, что скорости эволюции рассматриваемых генов одинаковы. В общем случае предположение одинаковой скорости эволюции чересчур сильное, и степень сходства последовательностей можно скорее интерпретировать как меру достоверности связи между вершинами дерева генов. В качестве дополнительной меры достоверности узла филогенетического дерева используется индекс поддержки этого узла при псевдорепликах ресэмплинга (будстрепа).

Пусть заданы длины ребер $c(a, b)$ дерева генов. Тогда можно ввести в качестве меры отличия деревьев G и S функцию потерь

$$L(G, S) = \sum_{pg \in O(G,S)} c(g, pg) + \sum_{g \in M(G,S)} c(g, pg)|I_g|,$$

в которой первый член характеризует потери от дупликаций, а второй — потери от пропущенных узлов. Таким образом, функция стоимости различия деревьев определяется суммарным числом потерь генов, взятых с разными весами, причем веса представляют собой преобразованный показатель сходства последовательностей или будстрэп поддержки.

3. МЕТОДИКА РАСЧЕТОВ

3.1. Алгоритм построения гомоморфизма деревьев и расчета стоимости их различия.

Легко построить построить алгоритм, позволяющий вычислять функцию стоимости за среднее время $O(N \log N)$, где N — число ОТЕ. Отображение $\alpha(g)$ из G в S строится за среднее время $O(N \log N)$.

Вычисление основывается на следующем замечании: значение $\alpha(g)$ на листьях g дерева генов равно соответствующим листьям дерева видов; на внутренних вершинах дерева значение $\alpha(g)$ равно наименьшему (относительно естественного частично порядка на вершинах дерева видов) общему предку значений $\alpha(cg)$ и $\alpha(\hat{c}g)$, который вычисляется за среднее время $O(\log N)$, где N — число ОТЕ. Применяя более совершенные алгоритмы поиска наименьшего общего предка в двоичном дереве можно уменьшить оценку времени работы этой части и всего алгоритма до $O(N)$ (см. [12]).

Число односторонних дупликаций и пропусков по вершинам дерева видов S рассчитывается одновременно с построением отображения $\alpha(g)$.

Для этого на каждой вершине s дерева видов S вводится счетчик стоимости дупликаций $d(s)$ (начальное значение $d(s) = 0$) и счетчик стоимости пропусков $i(s)$ (начальное значение $i(s) = 0$). Счетчики определяются одновременно с построением отображения $\alpha(g)$ следующим образом: (напомним, что отображение $\alpha(g)$ строится по дереву G от листьев к корню), если $\alpha(g) = \alpha(pg) = s$, а для другого непосредственного потомка g' вершины g это неверно, то полагаем $d(s) := d(s) + c(g, pg)$; если $\alpha(pg)$ не являтся отцом $\alpha(g)$, то для каждого s, такого что $\alpha(g) \subset s \subset \alpha(pg)$, полагаем $i(s) := i(s) + c(g, pg)$.

Среднее время работы такого алгоритма $O(N \log N)$ (в худшем случае $O(N^2)$).

3.2. Алгоритм построения дерева видов.

Для записи деревьев будет использоваться скобочный формат. Например, одно из возможных деревьев генов с листьями 1, 2, 3, 4, может быть записано в виде $(((1, 2), 3), 4)$. Кроме указанной информации может быть приписана также информация о стоимости ребра.

Пусть заданы деревья генов $G_1, G_2, \ldots, G_n$. Рассмотрим задачу построения дерева видов S, для которого величина

$$c(S) = c(G_1, S) + c(G_2, S) + \cdots + c(G_n, S) \tag{1}$$

достигает минимума. По-видимому, в общем случае, решение этой задачи требует экспоненциального времени. Мы рассмотрим эвристический алгоритм, поиска локального минимума величины (1).

Начальное дерево видов S выбирается из каких-либо внешних соображений, например, в качестве дерева видов выбирается одно из деревьев генов. Алгоритм заключается в последовательной локальной перестройке дерева S в окрестности каждой вершины с целью поиска дерева S, дающего локальный минимум величине $c(S)$.

Предварительно задается глубина h перестройки. Для каждой вершины p дерева S рассматриваются все вершины, лежащие на глубине h под этой вершиной и производится их всевозможные перестановки вместе с соответствующими поддеревьями. При этом подсчитываются величины $c(S)$. В конечном счете выбирается дерево с наименьшим значением этой величины. После этого можно вновь повторить указанную процедуру. Время одного прохода по всем вершинам оценивается как

$O(nT(h))$, где n — число вершин дерева, а $T(h)$ — число всех двоичных деревьев глубины h. Проходы осуществляются последовательно (можно задавать разные направления проходов дерева) до тех пор пока стоимость $c(S)$ не перестанет уменьшаться.

Например, в случае $h = 2$ все эти деревья можно перечислить следующим образом. Допустим, что вершина A имеет на глубине 2 три вершины 1, 2, 3 (в случае двух вершин перестройка не требуется, так как их перестановка приводит к изоморфным деревьям). Тогда из них можно образовать попарно не изоморфные деревья:

$$((1,2),3),((1,3),2),((2,3),1).$$

Если вершина A имеет на глубине 2 четыре вершины 1, 2, 3, 4, то из них можно образовать попарно не изоморфные деревья:
$(((1,2),3),4),(((1,2),4),3),$
$(((1,3),2),4),(((1,3),4),2),$
$(((1,4),2),3),(((1,4),3),2),$
$(((2,3),1),4),(((2,3),4),1),$
$(((2,4),1),3),(((2,4),3),1),$
$(((3,4),1),2),(((3,4),2),1),$
$((1,2),(3,4)),((1,3),(2,4)),((1,4),(2,3))$.

Может использоваться также более сложный стохастический алгоритм прохода по дереву видов для поиска минимума функции потерь, при котором задается и постоянно пересчитывается распределение вероятностей, с помощью которого выбирается очередная вершина дерева, в которой будет производиться локальная перестройка.

4. БИОЛОГИЧЕСКАЯ ПОСТАНОВКА ЗАДАЧИ И ТЕХНОЛОГИЯ РАСЧЕТОВ.

Практическое применение рассмотренных методов основано на определенной технологии, которая включает в себя биологическую постановку задачи, сбор и обработку исходной информации, расчеты с помощью разработанных и стандартных алгоритмов и программ.

Биологическая постановка заключается в отборе единиц (видов или таксонов) для проведения таксономического анализа. Подобный анализ может заключаться в постановке различных проблем классификации — от вопросов о близости каких-либо таксономических единиц к определенным известным группам, до задачи полного построения деревьев видов. Результатом биологической постановки является набор конкретных видов, отношения между которыми надо изучать. Этот набор будет корректироваться в зависимости от возможностей и наличия генетических данных.

После того как виды определены, производится выгрузка файлов, относящихся к соответствующим организмам из генетических баз данных, например, GeneBank: адрес в Интернете

http://www.ncbi.nlm.nih.gov.

После этого, проводится анализ выгруженных файлов на полноту информации. Требование полноты заключается в необходимости наличия достаточно большой группы белковых семейств или генов, общих всем изучаемым организмам. В дальнейшем, по каждому такому семейству будет проведено построение соответствующего дерева генов. По этим деревьям генов будет построено дерево видов.

Иногда для выполнения требования полноты приходится производить укрупнение таксономических единиц, путем слияния наиболее близких единиц (для того, чтобы увеличить количество генов какого-либо из намеченных белковых семейств).

Ряд вспомогательных компьютерных программ производят предварительную подготовку данных. Конвертор SELECTOR производит выбор последовательностей заданных белковых семейств из геномов организмов и представление их в виде отдельных файлов в FASTA-формате.

Пакет программ CLUSTAL используется для множественного выравнивания последовательностей, относящихся к данным белковым семействам. Пакет программ FORCON является конвертором

по переводу результатов множественного выравнивания из формата CLUSTAL в формат пакета программ PHYLIP. Последний используется для построения таксономических деревьев по каждому из белковых семейств.

После подготовки данных, с помощью программы TIQMAX, производится подбор и оптимизация дерева видов, минимизирующего стоимость различия с заданными деревьями генов. Поскольку эта программа находит лишь локальный минимум, нельзя исключить влияния на конечный результат выбора начального значения дерева видов.

Обычно в качестве начального дерева видов используются какие-либо деревья генов по отдельным белковым семействам, экспертные оценки или ранее существующие, часто упрощенные классификации, например, те которые предоставляет таксономический браузер базы данных GeneBank.

5. ПРИМЕРЫ РАСЧЕТОВ

5.1. Построение дерева видов по 9 белковым семействам.

Указанная технология применялась к задаче классификации 14 групп живых организмов: bone fishes (рыбы – лосось, форель, сом...), amphibians, lizard and snakes (игуана, анолис, геккон, кобра), crocodylidae (крокодил, аллигатор), anseriformes (утка, гусь), galliformes (курица, фазан, индейка), logomorpha (кролик), fissipedia (собака, лиса, кошка), cow (корова), sheep (овца), pig (свинья), rat (крыса), mouse (мышь), primates (мармозет, макака, человек).

Размер групп определялся в зависимости от объема генетических данных, выделенных из организмов группы.

Из организмов были выделены 9 белковых семейств: Adolase, Alphafetoprotein, Lactate dehydrogenase, Prolactin, Rhodopsin, Trypsinogen, Tyrosinase, Vasopressin, Wnt-7.

Каждому данному белковому семейству соответствует файл, содержащий идентификаторы белков, входящих в это семейство, соединенные с последовательностью аминокислот, составляющих эти белки. По данным 9 файлам с помощью пакета программ CLUSTAL были построены 9 деревьев генов.

Каждое из этих семейств содержало в среднем 20–30 белковых последовательностей. Число белковых последовательностей, выделенных из организмов одного семейства колебалось от 5 до 27. Вследствие этого, уровень достоверности различных таксономических групп в дереве видов был различным. Фрагмент файла, задающего исходное дерево видов, приведен в ниже:

```
/* Species - genes */
s1 = bone fishes: PRL ANGAN| OPSD ANGAN| SSSERALB...
s2 = amphibians : PWU80581|OPSD AMBTI| PLEWNT7A | AB002267...
s3 = liz snakes : SUU28411|SUU28410| SEOWNT7B...
s4 = crocod: PRL1*CRONO | PRL2*CRONO| PRL1*ALLMI...
s5 = anseriform: NEU1*STRCA| NEU2*STRCA| DUKLDHBCRY...
s6 = galliform: CHKALDB | GGPPALB| DECHLH...
s7 = lagomorpha: RABALDA | OCU85645|OCU18344...
s8 = fissi: S72946| OPSD*CANFA|TRY2*CANFA...
s9 = cow : BOVALBUMIN| BOVPLDH12| PRL*BOVIN...
s10 = sheep: OAMRALDB| OASERALB | PRL*SHEEP...
s11 = pig : PIGALBA| SDU07180| SDU07178...
s12 = rat : PRL*MESAU| OPSD CRIGR | RATALDA...
s13 = mouse : S72537| MMFETO| MUSAFPA...
s14 =primates: OPSL*CALJA| HSALDAR| HSALDCG...
/********* Initial species tree ***************************************/
(s1,(s2,(s4,(s3,(s5,(s6,(s7,(s12,(s9,(s10,(s11,(s8,(s13,s14)))))))))))));
```

Результирующее дерево видов в Phylip-формате:

(bone fishes:2.14832,
(amphibians:0.82307,
((((anseriform:0.158935,
crocod:0.03437):0.193305,
galliform:0.626266):0.586736,
liz snakes:0.36764):0.730343,
((lagomorpha:0.193942,
(((cow:0.210427,
sheep:0.116876):0.250356,
(pig:0.152622,
fissi:0.22021):0.196742):0.142653,
primates:0.566805):0.399384):0.379719,
(rat:0.756555,
mouse:0.610567):0.932419):0.78882):0.770471):1.04102);

Графическое изображение результирующего дерева видов приведено на рис. 3 в следующем разделе. Числа после двоеточий характеризуют относительную степень достоверности данного ребра.

5.2. Построение дерева видов по митохондриальным геномам

Основная цель данного расчета заключается в построении дерева видов основных групп живых организмов по митохондриальным геномам. В данном случае основной метод данной работы и программа TIQMAX проверялись в качестве средства построения консенсусного дерева видов по различным деревьям белковых последовательностей.

Были выбраны геномы митохондрий следующих организмов:

Porphyra порфира, Cyanidoschyzon, Marchantia печеночник, Phytophtora фитофтора, Chrysodidymus, Cafeteria, Allomyces,

Metozoa: Metridium актиния, Platyneris, Lumbricus дождевой червь, Katharina хитон, Drosophila дрозофила, Anopheles комар, Locusta саранча, Ixodes клещ, Artemia креветка, Penaeus, Asterina морская звезда, Florometra морская лилия.

Позвоночные: Branchiostoma ланцетник, Petromyzon минога, Squalus скат, Salmo лосось, Xenopus лягушка, Alligator аллигатор, Gallus курица, Didelphis опоссум, Homo человек.

Геномы митохондрий выли выгружены через Интернет из базы данных NCBI GeneBank, после чего были проведены анализ и редакция файлов. В результате того, из геномов митохондрий указанных организмов был выделен кластер, состоящий из 13 семейств генов, которые встречаются в каждой из митохондрий указанных классов. Приводим их список: NAD1, NAD2, NAD3, NAD4, NAD4L, NAD5, NAD6, COX1, COX2, COX3, ATP6, ATP8, CYTB.

Программа — конвертор SELECTOR выделила из геномов приведенных выше организмов файлы, содержащие идентификаторы генов и соответствующие им белковые последовательности данных 13 белковых семейств. Заметим, что некоторые гены могут иметь различные идентификаторы в разных геномах. Поэтому программа SELECTOR учитывает синонимию в идентификаторах генов.

После этого, для каждого из белковых семейств строилось дерево генов следующим образом. Предварительно производилось выравнивание белковых последовательностей семейства, после чего программами SEQBOOT и PROTPARSE пакета PHYLIP строились варианты деревьев генов данного семейства. Программа CONSENSE строила консенсусное дерево на основе этих вариантов. Внутренние вершины консенсусного дерева имеют численные характеристики, которые отражают степень достоверности, придаваемых будстреп-методом данной связи. Эти численные характеристики используются в качестве длин ветвей при подборе дерева видов с помощью программы TIQMAX.

Для проверки достоверности основного метода подбора видового дерева по митохондриальным геномам, применялись еще 2 метода построения консенсусного дерева видов по деревьям генов.

Первый метод заключался в применении программы CONSENSE к ранее построенным деревьям деревьям генов.

Второй метод основан на соединении всех белковых последовательностей генома каждого организма в одну общую последовательность большой длины и последующем применении программы PROTPARSE для построения дерева по этим последовательностям, которое интерпретировалось как вариант дерева видов.

Все три метода дали практически совпадающий результат: (приводим запись в PHYLIP-формате дерева видов программы TIQMAX; графическое изображение всех трех деревьев видов приведено на рис. 1–5):

(((((CHRY:621.667,
PHYT:645):804.867,
CAFETER:783.333):999.5,
((RECLINOM:950,
MARCH:906.667):804.85,
(PORPH:886.667,
CYAND:831.667):896.783):1013.83):1064.92,
(METRI:600,
ALLOM:566.667):995.917):1488.47,
(((((((GALLUS:687.5,
((SQUAL:1058.33,
SALMO:1083.33):974.242,
XENOP:1083.33):502.317):851.6,
(DIDEL:1233.33,
HOMO:1233.33):842.25):780.117,
ALLIGA:1016.67):729.748,
PETR:1073.33):725.698,
BRANC:825):859.998,
(FLORO:1108.33,
ASTER:1108.33):957.9):875.732,
(((IXODE:873.333,
ARTEM:873.333):812.533,
(PENAEUS:1183.33,

(LOCUST:1058.33,
(ANOPH:1053.33,
DROSOP:1053.33):960.717):928.35):744.467):914.267,
(KATHA:1075,
(PLATY:1153.33,
LUMBR:1153.33):864.533):840.033):964.083):1401.71).

6. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Попытка применить предложенный алгоритм для анализа филогении отрядов позвоночных с использованием данных из [11] оказалась не вполне удачной. При использовании полной выборки было получено множество деревьев существенно различной топологии, но практически одинакового веса. По-видимому, причина заключается в разреженности данных: многие таксоны представлены лишь в одном или двух семействах белков. Аналогичные результаты получаются и при использовании других алгоритмов: так, в [9] было найдено 12 деревьев видов, одинаково хорошо описывающих 53 дерева генов.

Огрубление данных путем слияния родственных таксонов с малым числом представителей и удаления изолированных малочисленных таксонов приводит к относительно приемлемому, но малоинтересному дереву (рис. 1). Дальнейшее продвижение в этом направлении существенно затруднено отсутствием доступных выборок белковых семейств с хорошим представительством в разнообразных таксономических группах.

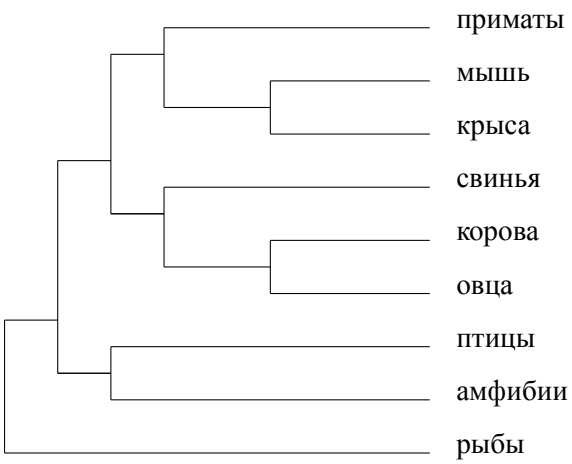

Рис. 1. Дерево видов позвоночных

С другой стороны, можно рассматривать вес дерева видов c(S) как характеристику согласованности деревьев генов и в отсутствие дупликаций. Оказывается, что таким образом можно строить деревья видов по семействам ортологичных белков. В отличие от алгоритмов, в которых используется только топология деревьев генов, наш алгоритм учитывает длины ветвей в деревьях генов, противоречащих дереву видов: короткие (и тем самым плохо подтвержденные) ветви штрафуются слабее, чем надежные.

В качестве примера были рассмотрены полные митохондриальные геномы различных эукариот. Дерево этих видов согласно таксономии GeneBank [14] приведено на рис. 2.

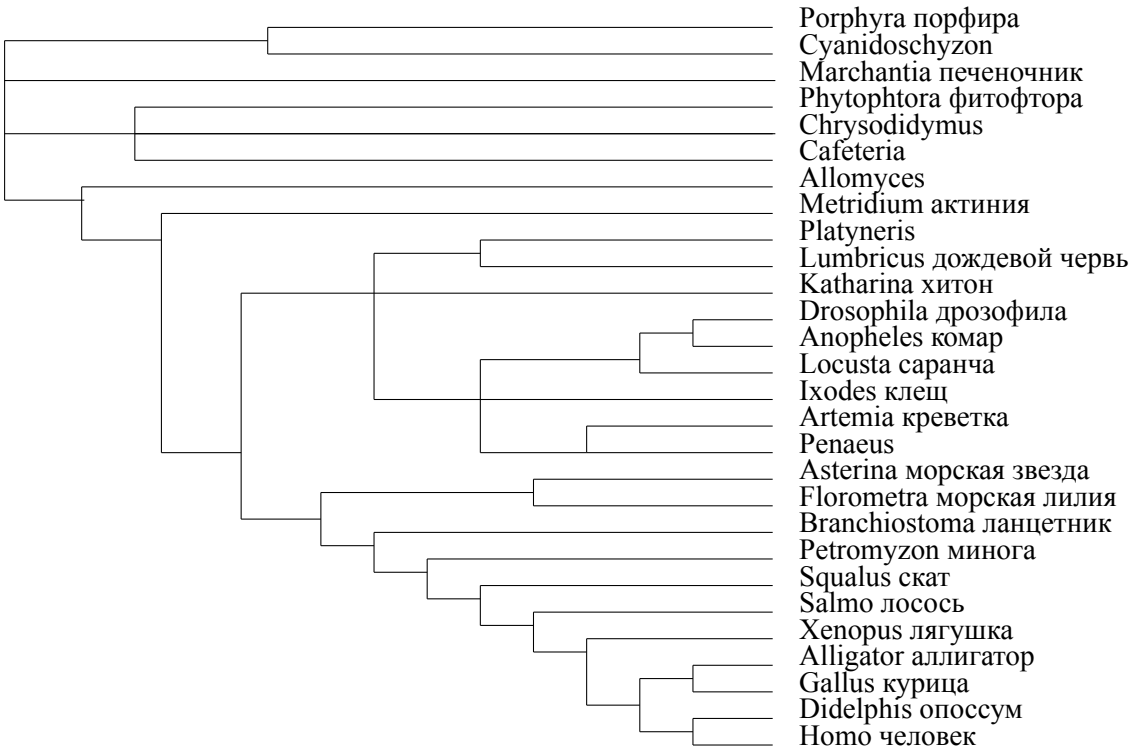

Рис. 2. Стандартная таксономия эукариот согласно http://www.ncbi.nlm.nih.gov/Taxonomy. Reclinomonas не классифицирован

Методом наибольшей экономии (программа PROTPARSE из пакета Phylip [5] были построены деревья следующих генов, имеющихся во всех геномах: COX1, COX2, COX3, ATP6, ATP8, CYTB, ND1, ND2, ND4, ND5, ND6.

Согласованные деревья были построены тремя разными способами. Дерево, построенное описываемой в настоящей статье программой TIQMAX приведено на рис. 3. На рис. 4. приведен результат применения программы PROTPARSE к последовательности, полученной конкатенацией всех рассматриваемых белков (точнее, для каждого вида были сконкатенированы в стандартном порядке все белки данного вида и к полученному набору последовательностей была применена программа). Наконец, на рис. 5 приведены результаты применения программы CONSENSE из пакета PHYLIP к набору деревьев генов.

Полученные деревья незначительно различаются друг от друга. Обшая таксономия Metazoa реконструирована одинаково, только в дереве TIQMAX объединены Metridium и Allomyces. Bilateria разделяются на Protostomia и Deuterostomia, первые — на Arthropoda и Molluska/Annelida (которые везде образуют один таксон), а вторые — на Echinodermata и Chordata. В Deuterostomia в дереве protparse перепутан порядок ветвей Echinodermata и Branchiostoma.

В то же время, во всех трех деревьях аллигатор, а не скат, является внешним видом по отношению к другим Gnasthomata, скат и лосось образуют таксон, следующим членом которого является лягушка, в то время как курица либо входит в этот таксон (TIQMAX и protparse), либо кластеризуется с млекопитащими, образуя таксон теплокровных (consense). Тем самым, не образуются канонические таксоны Tetrapoda и Archozauria.

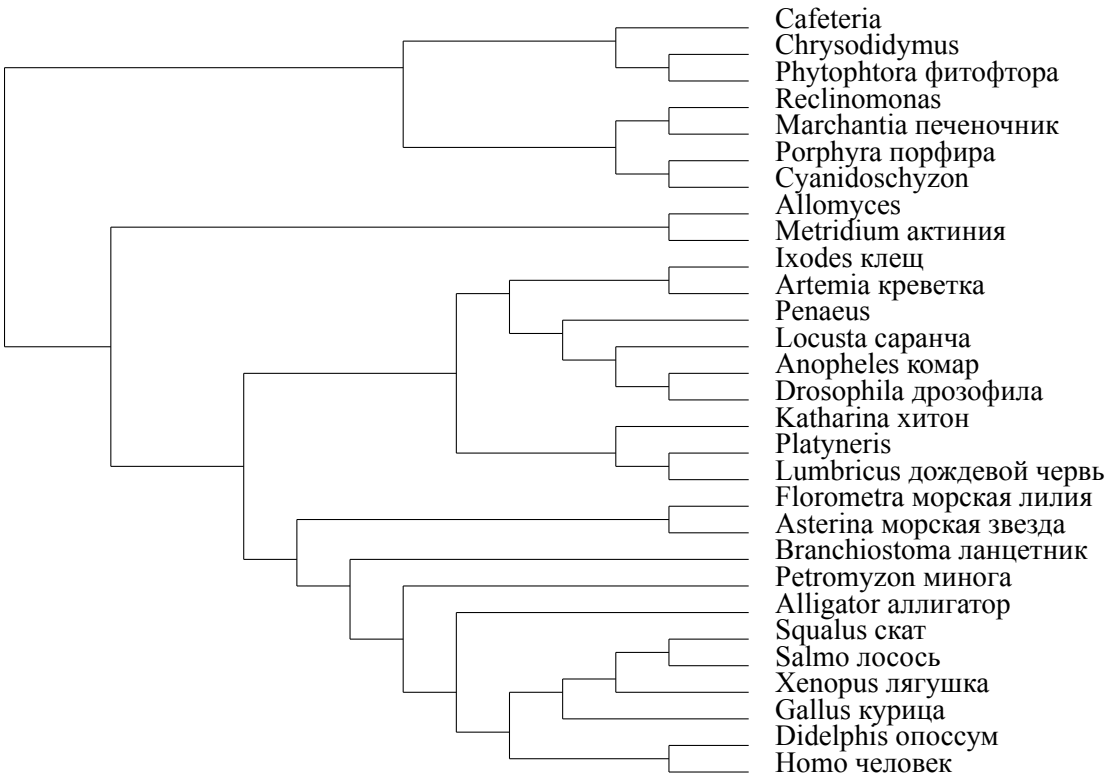

Рис. 3. Таксономия эукариот, построенная алгоритмом согласования деревьев TIQMAX. Положения корня выбрано так, чтобы максимизировать согласование с традиционной классификацией

Странности наблюдаются также в таксоне Arthropoda (членистоногие): во всех трех деревьях он разделяется на Ixodes/Artemia и Penaeus/Hexapoda, в то время как в каноническом дереве происходит трифуркация на таксоны хелицеровых (Ixodes), насекомых (Hexapoda) и ракообразных (Artemia и Penaeus).

Во всех трех деревьев Reclinomonas образует единый таксон с Marchantia, этот таксон далее кластеризуется с красными водорослями (Rhodophyta: Porphyra и Cyanidoschizon). В то же время, таксон Stramenopiles образуется только в дереве TIQMAX.

В таблице (см. следующую страницу) приведены результаты для деревьев видов. Они показывают очень тесную связь рыб и амфибий: в восьми деревьев из одиннадцати образован таксон (скат, лосось), в шести деревьев — таксон ((скат, лосось), лягушка). Минога является внешним членом таксона в пяти деревьях. Млекопитающие почти всегда образуют единый таксон (опоссум, человек). В то же время, положение аллигатора и курицы менее определено, а таксон Archozauria образовался только в трех случаях. Жирным цветом выделены потерянные члены таксонов.

Среди членистоногих в восьми деревьях наблюдается кластеризация Penaeus с насекомыми. Взаимоотношения между членами групп Viridiplantae, Rhodophytes и Stramenopiles довольно запутанны: эти три группы как правило образуют один таксон, однако порядок ветвления в этом таксоне весьма неустойчив. Тем не менее, таксон (Reclinomonas, Marchantia) образован в шести деревьях, а (Reclinomonas, Marchantia, Rhodophyta) — в пяти.

Тем самым, разногласия между согласованными деревьями видов и существующей таксономией в группах позвоночных и членистоногих не являются результатом ошибок алгоритма, а действительно диктуются построенными стандартными деревьями генов. Для объяснения этих разногласий требуется более подробный анализ особенностей эволюции митохондриальных белков в различных таксономических группах, что выходит за рамки настоящего исследования.

Таблица

таксон ген	Vertebrata	Arthropoda	Viridiplantae, Rhodophytes, Stramenopiles
COX1	(Petromyzon, ((Squalus, Salmo), ((Alligator, Xenopus), (Gallus, (Didelphis, Homo)))))	(Ixodes, (Penaeus, (Artemia, (Locusta, (Drosophila, Anopheles)))))	((Marchantia, (Reclinomonas, (Cyanidoschizon, Porphyra))), (Cafeteria, (Chrysodidymus, Phytophtora)))
COX2	(Alligator, (Petromyzon, (((Squalus, Salmo), Xenopus), (Gallus, (Didelphis, Homo))))	(Artemia, (((Locusta, Ixodes), (Penaeus, (Drosophila, Anopheles))), **Mollusca/Annelida**))	(Cyanidoschizon, ((Chrysodidymus, Reclinomonas), ((Porphyra, Marchantia), (Cafeteria, Phytophtora))))
COX3	(Petromyzon, (((Squalus, Salmo), (Alligator, Gallus)), (Xenopus, (Didelphis, Homo))))	((**Marchantia**, Ixodes), (**Mollusca/Annelida**, (Artemia, (Penaeus, (Locusta, (Drosophila, Anopheles))))))	(((Cyanidoschizon, (Allomyces, (Phytophtora, Porphyra))), (Chrysodidymus, Reclinomonas)), Cafeteria); Marchantia
ATP6	(Petromyzon, (((Squalus, (Salmo, Xenopus)), (Alligator, Gallus)), (Didelphis, Homo)))	(Ixodes, (Artemia, (Locusta, (Penaeus, (Drosophila, Anopheles)))))	((Porphyra, Cyanidoschizon), ((Reclinomonas, Marchantia), (Phytophtora, (Cafeteria, Chrysodidymus))))
ATP8	таксон не образовался	таксон не образовался	таксон не образовался
CYTB	(((Squalus, Salmo), Xenopus), (Gallus, ((Alligator, Didelphis), Homo))); Petromyzon	(Ixodes, (Artemia, (Penaeus, (Anopheles, (Locusta, Drosophila)))))	(((Reclinomonas, Marchantia), (Cyanidoschizon, Porphyra)), (Cafeteria, (Chrysodidymus, Phytophtora)))
ND1	(Petromyzon, (((Squalus, Salmo), Xenopus), ((Alligator, Gallus), (Didelphis, Homo))))	((Ixodes, Artemia), (Penaeus, (Locusta, (Drosophila, Anopheles))))	(Cafeteria, (Chryzodidymus, (Phytophtora, (Cyanidoschizon, (Porphyra, (Reclinomonas, Marchantia))))))
ND2	((Petromyzon, Alligator), (((Squalus, Salmo), Xenopus), (Didelphis, Homo)))	(((Ixodes, Artemia), **Annelida**), (Penaeus, (Locusta, (Drosophila, Anopheles))))	(Chrysodidymus, (Cafeteria, (Phytophtora, ((Reclinomonas, Marchantia), (Porphyra, Cyanidoschizon)))))
ND4	((Petromyzon, **Asterina**), (Xenopus, ((Alligator, (Squalus, (Salmo, Gallus))), (Didelphis, Homo))))	таксон не образовался	(((((Reclinomonas, Marchantia), (Chryzodidymus, Metridium)), (Porphyra, Cyanidoschizon)), Phytophtora), Cafetetria)
ND5	(Petromyzon, (Alligator, ((((Squalus, Salmo), Xenopus), Gallus), (Didelphis, Homo))))	(Artemia, (Ixodes, (**Mollusca/Annelida**, (Penaeus, (Locusta, (Drosophila, Anopheles))))))	((((Reclinomonas, Marchantia), (Porphyra, Cyanidoschizon)), Phytophtora), (Chrysodidymus, Cafeteria))
ND6	(Alligator, (Petromyzon, (Gallus, (((Squalus, Salmo), Xenopus), (Didelphis, Homo)))))	((**Mollusca/Annelida**, (Ixodes, Artemia)), (Penaeus, (Anopheles, (Locusta, Drosophila))))	(Chrysodidymus, (Metridium, (Marchantia, (Reclinomonas, (Phytophtora, (Cyanidoschizon, (Porphyra, Cafeteria)))))))

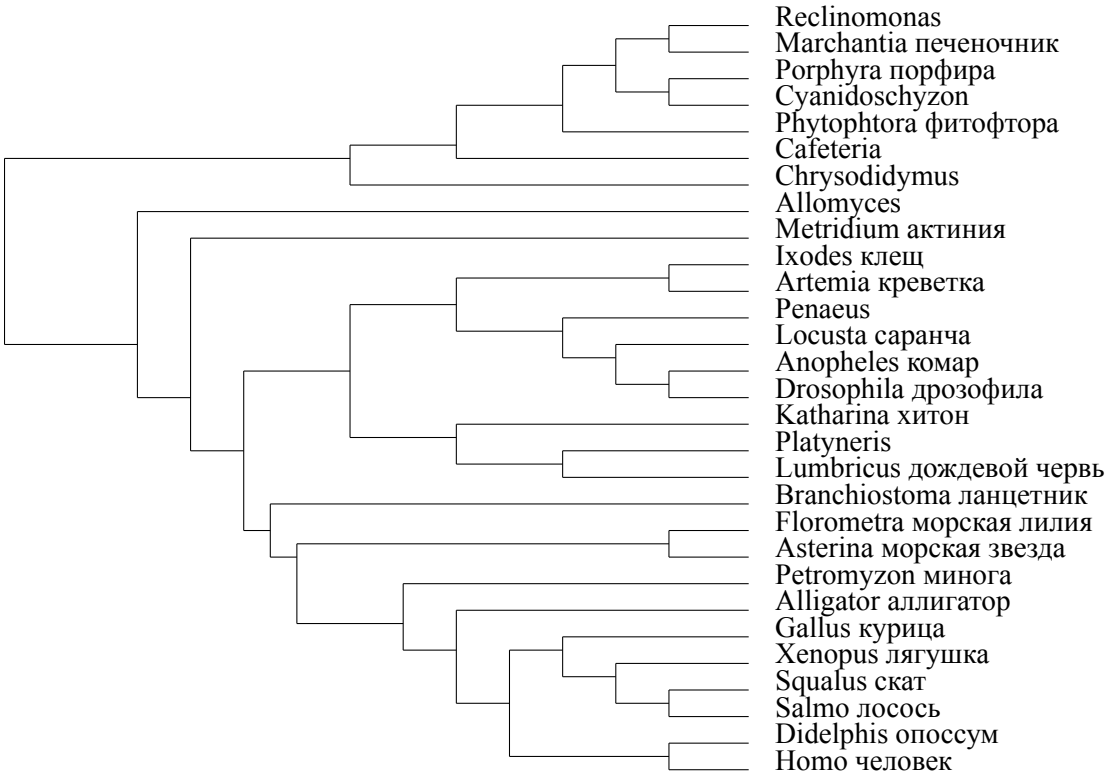

Рис. 4. Таксономия эукариот, построенная программой согласования деревьев protparse из пакета PHYLIP.Положения корня выбрано так, чтобы максимизировать согласование с традиционной классификацией

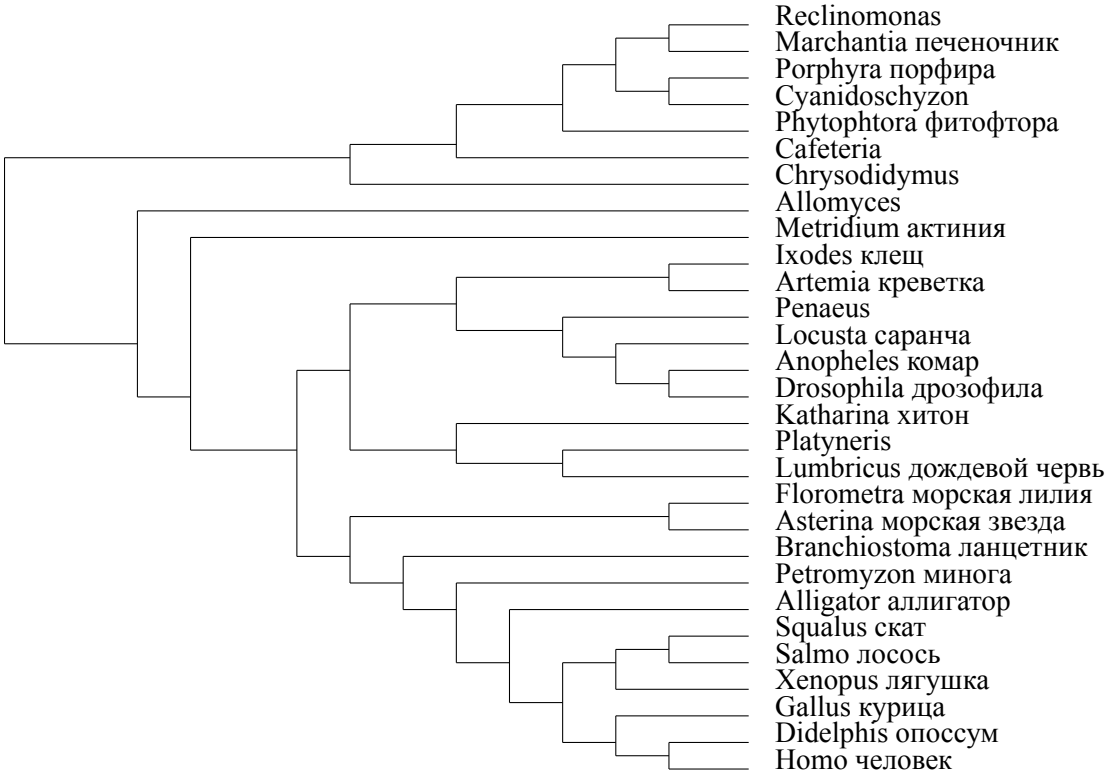

Рис. 5. Таксономия эукариот, построеннная программой согласования деревьев consensus из пакета PHYLIP. Положения корня выбрано так, чтобы максимизировать согласование с традиционной классификацией

В то же время, проведенный анализ показал практическую применимость метода в ситуации, когда выборка содержит представителей семейств генов во всех рассматриваемых таксонах. Кроме того, результаты позволяют предположить родство Reclinomonas, ранее не классифицированного представителя эукариот, с зелеными растениями.

СПИСОК ЛИТЕРАТУРЫ

1. Вейр Б. *Анализ генетических данных.* М.: Мир, 1995.
2. Cavalli-Storza L.L., Edwards A.W.F. Phylogenetic Analysis: Models and Estimation Procedures. *Evolution*, 1967, vol. 21, pp. 550–570.
3. Goodman M., Czelusniak J., Moore G.W., Romero-Herrera, A.E., Matsuda, G. Fitting the Gene Lineage into Its Species Lineage. A Parsimony Strategy Illustrated by Cladograms Constructed from Globulin Sequences. *Syst. Zool*, 1979, vol. 28, pp. 132–163.
4. Guigo R., Muchnik I., Smith T. Reconstruction of Ancient Molecular Filogeny. *Molecul. Phylogenetics Evol.,* 1996, vol. 6, no. 2, 189–213.
5. Felsenstein J., PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics*, 1989, vol. 5, pp. 164-166, [http://evolution.genetics.washington.edu/phylip.html].
6. Eulenstein O., Vingron M. On the Equivalence of Two Tree Mapping Measures. *Arbeitspapiere der GMD*, 1995, no. 936, Bonn, Germany.
7. Eulenstein O., Mirkin B., Vingron M. Duplication-based Measures of Difference Between Gene and Species. 1998, *J. Computational Biology.* 1998, vol. 5, no. 1, pp. 135–148.
8. Page, R.D.M., Charlstone M.A. From Gene to Organismal Phylogeny: Reconciled Trees and Gene Tree/Species Tree Problem. *Mol. Phylogenet. Evol.,* 1997, vol. 7, pp. 231–240.
9. Page R.D.M., Charleston M.A. Reconciled Trees and Incongruent Gene and Species Trees. *DIMACS Series in Mathematics and Computer Sciences*, 1997, vol. 37, *"Mathematical Hierarchies in Biology"*.
10. Page R.D.M. GeneTree: Comparing Gene and Species Phylogenies using Reconciled Trees. *Bioinformatics Appl. Notes*, 1998, vol. 14, no. 9, pp. 819–820.
11. Page R.D.M., Extracting Species Trees from Complex Gene Trees: Reconciled Trees and Vertebrate Phylogeny. *Molecujlar Phylogenetics and Evolution*, 2000, vol. 14, pp. 89-106, [http://taxonomy.zoology.gla.ac.uk/rod/data/vertebrates].
12. Schieber, B. and Vishkin U. On Finding Lowest Common Ancestors: Simplification and Parallelization. *SIAM J. Comput.,* 1988, vol. 17, no. 6, pp. 1253–1262.
13. Waterman M.S. *Introduction to Computational Biology.* Chapman and Hall, 1995.
14. Wheeler D.L., Chappey C., Lash A.E., Leipe D.D., Madden T.L., Schuler G.D., Tatusova T.A., Rapp B.A. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 2000, vol. 28, pp. 10-14, [http://www.ncbi.nlm.nih.gov/Taxonomy].
15. Zhang L. On a Mirkin–Muchnik–Smith Conjecture for Comparing Molecular Phylogenies. *J. Computat. Biol.,* 1997, vol. 4, no. 2, pp. 177–187.

Статью представил к публикации член редколлегии Н.А. Кузнецов