Систематика, таксономия, филогенетика

Автор: Владимир Шитиков
https://stok1946.blogspot.com/

1. Введение

Определим предварительно общий смысл отдельных базовых терминов, границы между которыми не очень четко обозначены (Christoffersen, 1995).

Систематика - это область биологии, которая классифицирует живые и вымершие организмы в соответствии с некоторым установленным набором правил. При этом организмы распределяются по группам на основе сходства и различия их характеристик, образуя некоторую классификационную иерархию. Таким образом, систематика диктует наиболее общие принципы теории и практики идентификации биологических систем и упорядочения разнообразия организмов.

Таксономия - это техника описания видов и более крупных групп, а также практика распознавания, наименования и упорядочения таксонов в систему категорий, согласующуюся с любыми отношениями между таксонами. Таксономисты используют общие характеристики для создания иерархий, которые субъективно определены, но имеют практическое применение. При этом используется максимально широкий набор признаков, взятых из самых различных областей биологии – от морфологии или анатомии до биохимии. Буквально *таксономия* означает "давать имена вещам", и здесь она идет рука об руку с систематикой, поскольку полезные схемы иерархии называют сущности таким образом, чтобы они наилучшим образом отражали их классификацию.

Кладистика (также известная как филогенетическая систематика) – это систематическая классификация групп организмов на основе эволюционной теории, т.е. наследования общих характеристик, которые, как считается, происходят от общего предка. Иерархическая структура выстраивается в процессе сравнения различных таксонов, чтобы найти объективные феномены сходства и различия между ними; при этом формируются строгие последовательности линий популяций от предков к потомкам и выделяются вложенные наборы этих линий или *клады*, для которых характерна эволюционная близость (то есть общность происхождения). Кладистический метод был впервые описан в 1966 году Хеннигом (Hennig), но в последние десятилетия приобрел популярность из-за наличия мощных компьютеров, дающих возможность проведения многомерного анализа.

Современными графическими моделями для обоснования иерархической классификации биологических систем являются разветвленные асимметричные дендрограммы (филогенетические деревья или кладограммы). В кладистике применяется строгая схема реконструкции родственных отношений между таксонами, включающая требование монофилии и взаимно-однозначного соответствия при выделении групп, которые могут включать всех известных потомков гипотетического ближайшего предка, общего только для членов этой группы и ни для кого другого. И таксономия, и кладистика стремятся дать содержательную классификацию жизни на земле, как живой, так и вымершей. Но методы и цели этих двух методологий различаются. Хотя таксономия и кладистика используют большую часть одних и тех же данных, они не всегда дают одинаковые результаты.

К сожалению, смысл термина *кладистика* оказался запутан тем фактом, что кроме методологии он несет в себе и философию. Эта философия заключается в том, что единственными группами, подлежащими обсуждению, являются клады, то есть группы, состоящие из предка вместе со всеми его потомками. Например, систематики идентифицируют рептилий, птиц и млекопитающих как отдельные и равноправные классы позвоночных. Кладисты, напротив, не признают таксон "рептилии", поскольку

он не включает динозавров и птиц ("пернатых динозавров"), которые являются потомками рептилий в общепринятом понимании. Однако биологи продолжают использовать по отдельности родственные таксоны рептилий и птиц, потому что, хотя это и не настоящие клады, но они полезны и очевидны.

Филогения – это "иерархическая структура древа жизни", посредством которого каждая форма жизни связана с любой другой. Как считают некоторые исследователи (http://www.miketaylor.org.uk/dino/faq/s-class/terms/), термин "филогения" относится к единственному, истинному гипотетическому дереву, не связанному с любыми теориями или домыслами, которыми исследователи руководствуются при его построении. Таким образом, филогения – это не деятельность (совокупность методологий), а некая объективная реальность, которую мы пытаемся обнаружить, но никогда не можем познать до конца. Отметим, однако, что истинность филогении не существует сама по себе, а только применительно к той задаче, которую ставить перед собой исследователь (например, исследование таксономической структуры сообществ или выстраивание эволюционных отношений).

Филогенетика (как отрасль биологии, изучающей филогению) использует унаследованные характеристики для создания групп организмов и занимается, как и кладистика, реконструкцией исторических отношений между этими группами в эволюционном контексте. Филогенетический анализ также использует данные о структуре родственных отношений между видами, представленные в виде филогенетического дерева (Webb et al., 2002; Лукашов, 2009). Дерево состоит из узлов, соединённых ветвями, длина которых связана с постановкой решаемой задачи. Внешние вершины дерева (листья) соответствуют реальным объектам эксперимента или носителям информации. Все остальные узлы считаются внутренними, которые, как правило, упорядочены по иерархии таксонов (роды, семейства и т. п.).

Филогенетическое дерево – это рабочая гипотеза, которая описывает эволюционные отношения между группами организмов, а точки ветвления указывают, когда новые виды расходились от общего предка. Если для деревьев систематики (например, по Линнею) длина каждой ветви является единичной, а расстояние между таксонами определяется числом таксономических категорий L (Clarke, Warwick, 2001), то длина ветвей классического филогенетического дерева может измеряться, в частности, миллионами лет с момента дивергенции. Благодаря развитию молекулярных методов, филогенетические деревья датируют события в масштабе эволюционного времени T, а информация о них становятся всё более доступной для разных систематических групп.

В отличие от кладограммы, филогенетическое дерево не требует строгого соответствия систематики и филогении: в частности, это выражается в признании права на существование в системе парафилетических групп. Изображения деревьев возможны в неукорененной, звездообразной или циркулярной формах, однако и тут объединение потомков, связанных общим узлом-предком, продолжает называться кладами.

Определение последовательности дивергенций в ходе эволюции может быть непростой задачей. Современные генетические методы используют "молекулярные часы", чтобы помочь в построении филогенетических деревьев. Молекулярная систематика и анализ генетических различий между видами позволяет выявить происхождение от предкового гена. Все это привело к серьезной перестройке традиционных линнеевских представлений и наших взглядов на таксономию. По мере обнаружения новых видов таксономические группы перестают быть монофилетическими, а процесс горизонтального переноса генов приводит нас к некой "паутине жизни" вместо простых деревьев.

К настоящему времени разработано большое множество сравнительных филогенетических методов, реализующих различные статистические подходы для

построения деревьев и решения задач эволюционного анализа. В обзоре Brian O'Meara (https://CRAN.R-project.org/view=Phylogenetics) приводится список пакетов R, реализующих всю последовательность этапов сравнительного филогенетического анализа. Сюда включается подготовка данных, визуализация сформированных деревьев и построение моделей, выполняющих поиск неслучайных закономерностей в разных сферах биологии от молекулярной геномики до экологии сообществ.

Формально в дальнейшем изложении речь пойдет не о филогенетических деревьях, а о таксономических дендрограммах, основанных на актуальной биологической систематике. Хотя биологическое разнообразие и возникло в результате филогенеза, многие задачи экологии опираются не на эволюционный контекст, а на традиционную классификацию живых организмов. Более того, многие гипотезы, основанные на современных филогенетических моделях, рассматривают закономерности формирования структур из уже существующих видов и непосредственно не связаны с эволюционной историей (Ives, 2021). Поскольку в классических задачах экологии сообществ смысл использования датированных деревьев не всегда очевиден, то здесь обсуждается применение деревьев, узлами которых служат конкретные таксоны разного ранга, а длина каждой ветви принимается единичной. Легко показать (Chao et al. 2014), что при этом вполне корректно использование многих математических выражений и процедур анализа топологии филогенетических деревьев с датированием эволюционных событий.

2. Подготовка таксономической таблицы

В настоящем сообщении мы проводим анализ таксономической структуры донных сообществ на основе филогенетических представлений и оцениваем степень влияния такого ведущего фактора, как минерализация водной среды, на тесноту родственных связей между видами. Будем опять использовать в качестве примера данные гидробиологической съемки равнинных рек в бассейне Средней и Нижней Волги – см. посты, представленные ранее:

https://stok1946.blogspot.com/2020/09/blog-post.html и
https://stok1946.blogspot.com/2020/11/sdm.html.

С учетом однородности природно-климатических условий из всего массива наблюдений отберем 519 проб, выполненных в 267 точках на 48 малых и средних равнинных реках степной и опустыненной зон Саратовской и Волгоградской областей, где всего было выделено 356 видов или более высших таксонов.

Рассмотрим использование возможных рангов таксономии и выполним систематическое описание по 11 уровням: Species → Genus → Tribe → SubFamily → Family → SubOrder → Order → SubClass → Class → SubPhylum → Phylum, осуществляя поиск в справочных испочниках соответствующих идентификаторов уровней для каждого вида из отобранного списка. Выделение трибы в качестве самостоятельного уровня связано с тем, что для многочисленного семейства Chironomidae эта градация имеет важное практическое значение. Выполним затем выравнивание векторов наименований уровней, чтобы они вместе образовали матрицу `Species_Tax`, в которой 11 столбцов описывают упорядоченную таксономическую иерархиею, а строки состоят из последовательности идентификаторов уровней каждого из 356 видов. При этом любому отсутствующему наименованию ранга будет присваивоен псевдо-идентификатор предыдущего уровня.

Файл с выполненным таксономическим описанием представлен на общедоступном ресурсе http://www.ievbras.ru/ecostat/Kiril/R/Blog/Species_Tax.RData, который можно скачать и поместить в рабочий каталог среды R.

```
load(file="Species_Tax.RData")
```

Таблица Species_Tax имеет следующую структуру:

```
str(Species_Tax)
'data.frame':   356 obs. of  14 variables:
 $ Name      : chr  "Dikerogammarus caspius" "Dikerogammarus haemobaphes"
                                "Gammarus lacustris" "Gammarus pulex" ...
 $ Com       : num  4 2 33 1 2 1 1 1 1 11 ...
 $ W         : num  1.398 0.371 8.286 0.48 0.368 ...
 $ Code      : chr  "AmDic.c." "AmDic.h." "AmGam.l" "AmGam.px" ...
 $ Genus     : chr  "Dikerogammarus" "Dikerogammarus" "Gammarus" "Gammarus" ...
 $ Tribe     : chr  "Gammaridae" "Gammaridae" "Gammaridae" "Gammaridae" ...
 $ SubFamily: chr  "Gammaridae" "Gammaridae" "Gammaridae" "Gammaridae" ...
 $ Family    : chr  "Gammaridae" "Gammaridae" "Gammaridae" "Gammaridae" ...
 $ SubOrder : chr  "Gammaridea" "Gammaridea" "Gammaridea" "Gammaridea" ...
 $ Order     : chr  "Amphipoda" "Amphipoda" "Amphipoda" "Amphipoda" ...
 $ SubClass : chr  "Eumalacostraca" "Eumalacostraca" "Eumalacostraca" "Eumalacostraca" ...
 $ Class     : chr  "Malacostraca" "Malacostraca" "Malacostraca" "Malacostraca" ...
 $ SubPhylum: chr  "Crustacea" "Crustacea" "Crustacea" "Crustacea" ...
 $ Phylum    : chr  "Arthropoda" "Arthropoda" "Arthropoda" "Arthropoda" ...
```

Здесь столбцы Name и Code – наименование и внутренний код вида, Com – число проб, в которых он встретился, а W – показатель средней соленосной толерантности (обсуждается далее).

Обычной задачей в биологии является нормализация названий видов, т.е.: а) виды должны иметь самые современные имена, б) они должны быть правильно написаны и в) снабжены научным названием для основного имени. Одним из способов нормализации имен является служба глобального распознавания имен (GNR – Global Names Resolver), предоставляемая Энциклопедией жизни (Encyclopedia of Life).

Обратимся к пакету taxize (https://docs.ropensci.org/taxize/index.html), который интегрирует большое множество источников данных и обеспечивает качественный сервисный интерфейс для разрешения многих актуальных проблем таксономии. Рассмотрим список всемирных источников данных для поиска наименований видов и выделим из них специализированные базы по гидробиологии:

```
library("taxize")
Base_List <- gnr_datasources()
nrow(Base_List)
[1] 109
Base_List[agrep("Water", Base_List$title, ignore.case = TRUE),
                                    c("id", "title","description")]
    id title                                    description
1  144 Freshwater Animal Diversity Assessment - ~ The Freshwater Animal Diversity ~
Base_List[agrep("Marine", Base_List$title, ignore.case = TRUE),
                                    c("id", "title","description")]
    id title                                    description
1    8 The Interim Register of Marine and Nonm~ The Interim Register of Marine and ~
2    9 World Register of Marine Species         An authoritative classification and ~
3  120 Papahanaumokuakea Marine National Monum~ PMNM initial species list test data.
4  181 The Interim Register of Marine and Nonm~ The Interim Register of Marine and ~
```

Нам представлен список из 109 интерактивных баз данных, 5 из которых имеют в своем названии слова "Water" или "Marine". Проверим, имеется ли в этих базах вид *Pontogammarus robustoides* :

```
gnr_resolve(Species_Tax$Name [5]) [,-2]
  user_supplied_name          matched_name                data_source_title       score
1 Pontogammarus robustoides   Pontogammarus robustoides   National Center for Bi~ 0.988
2 Pontogammarus robustoides   Pontogammarus robustoides   Encyclopedia of Life    0.988
3 Pontogammarus robustoides   Pontogammarus robustoides   Index to Organism Names 0.988
4 Pontogammarus robustoides   Pontogammarus robustoides   uBio NameBank           0.988
```

```
5  Pontogammarus robustoides   Pontogammarus robustoides              Arctos                  0.988
6  Pontogammarus robustoides   Pontogammarus robustoides              FishBase Cache          0.988
7  Pontogammarus robustoides   Pontogammarus robustoides              Open Tree of Life Refe~ 0.988
8  Pontogammarus robustoides   Pontogammarus robustoides (Sars, 1894) Catalogue of Life       0.988
9  Pontogammarus robustoides   Pontogammarus robustoides (Sars, 1894) The Interim Register o~ 0.988
10 Pontogammarus robustoides   Pontogammarus robustoides (Sars, 1894) World Register of Mari~ 0.988
11 Pontogammarus robustoides   Pontogammarus robustoides (Sars, 1894) GBIF Backbone Taxonomy  0.988
12 Pontogammarus robustoides   Pontogammarus robustoides (Sars, 1894) EUNIS                  0.988
13 Pontogammarus robustoides   Pontogammarus robustoides (Sars, 1894) BioLib.cz               0.988
14 Pontogammarus robustoides   Pontogammarus robustoides Sars 1894    Index to Organism Names 0.988
15 Pontogammarus robustoides   Pontogammarus robustoides Sars 1894    uBio NameBank           0.988
```

Заданный нами вид встретился в 15 базах данных. Здесь `score` – качество вхождения `user_supplied_name` в `matched_name` (аналог коэффициента корреляции).

Проверим весь список из 356 видов и отсортируем результат по величине `score`:

```
temp <- gnr_resolve(Species_Tax$Name, best_match_only = TRUE)
table(temp$score)
 0.75 0.988 0.995 0.999
   86   227    29     2
head(temp[order(temp$score),-(2, 4)])
  user_supplied_name        matched_name                                score
1 Cingulipisidium fedderseni Cingulipisidium Pirogov & Starobogatov 1974 0.75
2 Euglesa sp.                Euglesa                                     0.75
3 Henslowiana dupuiana       Henslowiana                                 0.75
4 Henslowiana henslowana     Henslowiana henslowiana                     0.75
5 Musculium sp.              Musculium                                   0.75
6 Neopisidium sp.            Neopisidium Odhner 1921                     0.75
```

Отметим в строке 4 ошибочное название `henslowana`. Остальные несовпадения носят непринципиальный характер. Обратим внимание на использование ключа `best_match_only = TRUE`.

Еще одна задача, с которой часто сталкиваются биологи, – получение более высоких таксономических названий для списка таксонов. Ряд источников данных в `taxize` дают возможность получения более высоких таксономических названий, но авторы рекомендуют два наиболее полезных из них: Интегрированная система таксономической информации (ITIS – Integrated Taxonomic Information System) и Национальный центр биотехнологической информации (NCBI – National Center for Biotechnology Information). Получим названия всех возможных таксономических уровней для одного из видов с помощью функции `classification`:

```
classification(Species_Tax$Name [5], db = 'ncbi')
$`Pontogammarus robustoides`
                       name         rank      id
1         cellular organisms      no rank  131567
2                  Eukaryota superkingdom    2759
3                Opisthokonta        clade   33154
4                     Metazoa      kingdom   33208
5                   Eumetazoa        clade    6072
6                   Bilateria        clade   33213
7                 Protostomia        clade   33317
8                   Ecdysozoa        clade 1206794
9               Panarthropoda        clade   88770
10                 Arthropoda       phylum    6656
11                Mandibulata        clade  197563
12                Pancrustacea        clade  197562
13                   Crustacea    subphylum    6657
14              Multicrustacea   superclass 2172821
15                Malacostraca        class    6681
16              Eumalacostraca     subclass   72041
17                  Peracarida   superorder    6820
```

```
18                Amphipoda        order    6821
19              Senticaudata     suborder 1732196
20                 Gammarida   infraorder 1732204
21               Gammaridira    parvorder 1732304
22               Gammaroidea  superfamily   44329
23           Pontogammaridae       family  315623
24             Pontogammarus        genus  225960
25 Pontogammarus robustoides      species  225961
```

Иногда вместо полной классификации может понадобиться только один уровень, например, семейство для выбранного вида. Функция `tax_name` создана именно для этой цели. Как и в случае с функцией `classification`, можно указать источник данных с аргументом `db` (ITIS, либо NCBI).

```
tax_name(Species_Tax$Name [1], get = "family", db = "ncbi")
    db                  query     family
1 ncbi Dikerogammarus caspius Gammaridae
```

Большинство баз данных используют числовой код для ссылки на запрашиваемый вид. Если извлечь эти коды для списка видов, то они могут использоваться для запроса дополнительных данных (Внимание: в каждой базе своя система кодирования). Проверим, сколько видов содержится в номенклатурной части базы NCBI:

```
Uid_List <- get_uid(Species_Tax$Name)
* Total: 356
* Found: 173
str(Uid_List)
 'uid' chr [1:356] "315701" "191520" "52639" "52641" "225961" "2025112" NA NA "143297"...
 - attr(*, "match")= chr [1:356] "found" "found" "found" "found" ...
 - attr(*, "multiple_matches")= logi [1:356] FALSE FALSE FALSE FALSE FALSE FALSE ...
 - attr(*, "pattern_match")= logi [1:356] FALSE FALSE FALSE FALSE FALSE FALSE ...
 - attr(*, "uri")= chr [1:356] "https://www.ncbi.nlm.nih.gov/taxonomy/315701" ...
sum(is.na(as.vector(Uid_List)))
[1] 183
```

К сожалению, полное совпадение по наименованию имеет только примерно половина видов. Поэтому будем ориентироваться на родовые имена:

```
Genus_List <- unique(Species_Tax$Genus)
Uid_List_G <- get_uid(Genus_List)
* Total: 222
* Found: 204
Genus_Fam <- data.frame(Genus=Genus_List,Uid_G=as.vector(Uid_List_G))
head(Genus_Fam[is.na(Genus_Fam$Uid_G),])
             Genus Uid_G
5          Amesoda  <NA>
6  Cingulipisidium  <NA>
9      Henslowiana  <NA>
12     Neopisidium  <NA>
14     Rivicoliana  <NA>
44    Guttipelopia  <NA>
```

В базе NCBI не удалось обнаружить только 18 родов из 222. Обратим внимание, что поисковый запрос может иметь совпадающие строки. В таких случаях поисковые функции возвращают фрейм данных совпадений и просят пользователя ввести номер строки, который необходимо принять. Чтобы блокировать это, нужно использовать параметр `rows=1`, и пользователю выдается, в некотором смысле, случайный вариант.

Выберем из этой же базы названия семейств, соответствующих этим родам:

```
Fam_List  <- tax_name(as.character(Genus_Fam[!is.na(Genus_Fam$Uid_G),1]),
    get = "family", db = "ncbi", rows=1)
str(Fam_List)
'data.frame':   204 obs. of  3 variables:
 $ db    : chr  "ncbi" "ncbi" "ncbi" "ncbi" ...
 $ query : chr  "Dikerogammarus" "Gammarus" "Pontogammarus" "Atherix" ...
 $ family: chr  "Gammaridae" "Gammaridae" "Pontogammaridae" "Athericidae"
...
```

Сравним полученные названия семейств с их наименованиями в столбце `Family` исходной (тестируемой) таблицы:

```
library(tidyverse)
Species_Tax[,c(5,8)] %>%
         inner_join (Fam_List[,-1], by=c("Genus"="query")) %>%
         filter(Family != family) %>% distinct()
           Genus         Family           family
1   Pseudocumidae  Pseudocumidae  Pseudocumatidae
2   Stenogammarus     Gammaridae  Pontogammaridae
3   Lithoglyphus  Lithoglyphidae      Hydrobiidae
4       Ilyocoris      Corixidae       Naucoridae
5       Mesovelia       Veliidae     Mesoveliidae
6      Micronecta      Corixidae    Micronectidae
7     Dicranomyia      Tipulidae       Limoniidae
8       Dicranota      Tipulidae       Pediciidae
9        Hexatoma      Tipulidae       Limoniidae
10   Isochaetides    Tubificidae         Naididae
11 Limnodriloides    Tubificidae         Naididae
12    Limnodrilus    Tubificidae         Naididae
13           Nais       Naididae Halosphaeriaceae
14    Potamothrix    Tubificidae         Naididae
15 Psammoryctides    Tubificidae         Naididae
16        Tubifex    Tubificidae         Naididae
17    Chloroperla       Perlidae   Chloroperlidae
```

Некоторое количество присвоенных имен семейств имеют расхождения с данными базы NCBI, но мы не будем останавливаться на разборе этого обстоятельства.

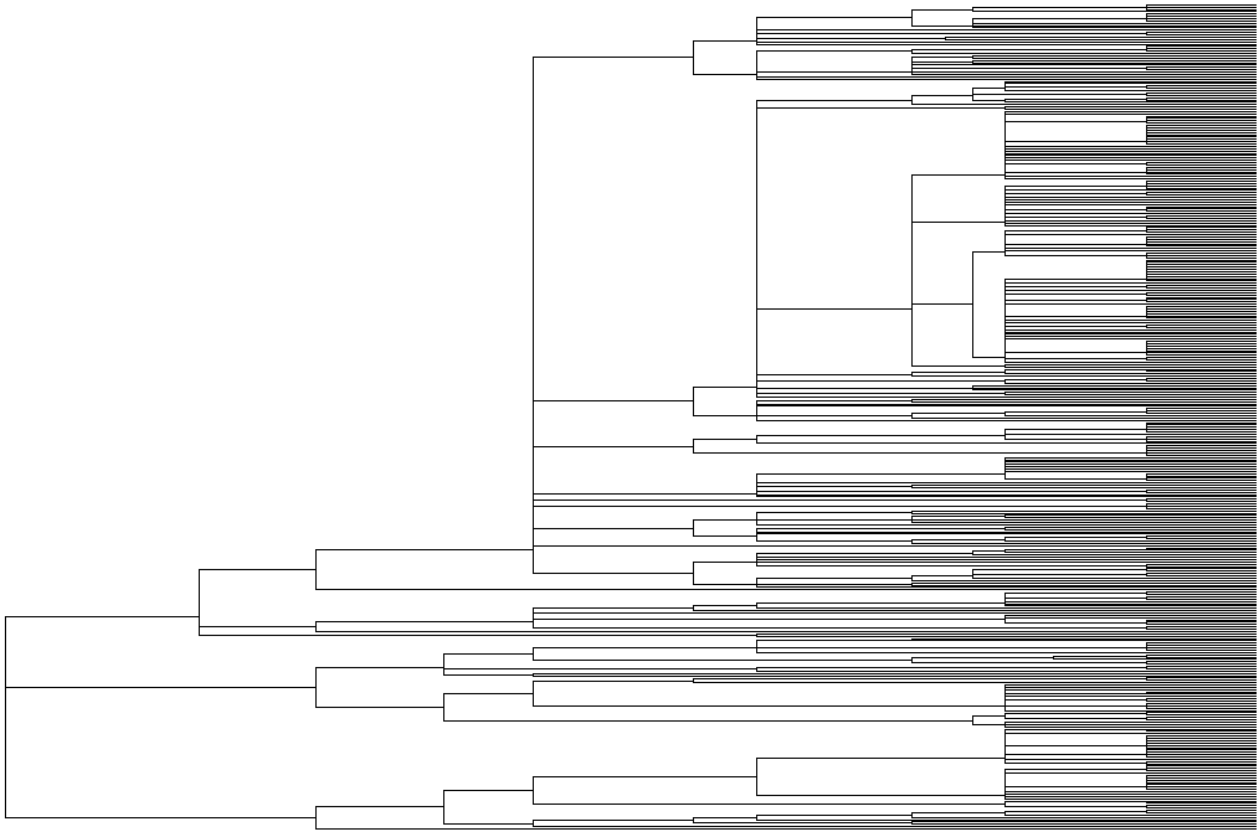
3. Построение филогенического дерева

Деревья R обычно хранятся в объектах `phylo` класса S3, реализованном в пакете `ape` (объект `phylo4` класса S4 реализован в пакете `phylobase`). Существует несколько вариантов создания этих объектов, в частности, `ape` может считывать деревья из внешних файлов в формате Newick (известном также как формат Phylip) или в формате Nexus.

Деревья систематики могут быть построены в двух формах: с использованием только матрицы таксономических уровней, представленной выше, и с учетом таксономических расстояний между каждой парой видов (Clarke, Warwick, 2001). Во втором случае используем функцию `taxa2dist` из пакета `vegan`, которая позволяет сформировать матрицу таксономических дистанций, и на ее основе построим дендрограмму с применением обычных методов иерархической кластеризации. Полученное дерево конвертируется в объектах класса S3 `phylo` функцией `as.phylo`.

```
library(ape)
library(vegan)
# Построение таксономического дерева
    taxdis <- vegan::taxa2dist(Species_Tax[,-(1:3)], varstep = TRUE)
    spe.dend <- hclust(taxdis, method="complete")
    PTree <- as.phylo(spe.dend)
```

```
    str(PTree)
List of 4
 $ edge      : int [1:710, 1:2] 357 358 358 366 371 371 381 410 410 381 ...
 $ edge.length: num [1:710] 12.38 37.62 5.16 3.57 28.9 ...
 $ tip.label  : chr [1:356] "AmDic.c." "AmDic.h." "AmGam.l" "AmGam.px" ...
 $ Nnode      : int 355
 - attr(*, "class")= chr "phylo"
 - attr(*, "order")= chr "cladewise"
plot(PTree, show.tip.label = FALSE)
```



Аналогичным образом работает и функция `class2tree` из пакета `taxize`, на вход которой подается только вектор со списком видов. Функция сама загружает компоненты таксономической структуры из источника данных с аргументом `db` (ITIS, либо NCBI), однако важно, чтобы они там присутствовали.

Степень экологической изменчивости видов оценивается обычно в пространстве их характерных свойств (*trait values*), связанных с морфометрическими признаками, широтой спектра реакций отдельных видов на воздействие факторов среды, значениями продуктивности, стабильности, скорости усвоения питательных веществ, и т.д. (Tilman, 2001; Petchey, Gaston, 2002). Для каждого вида мы рассчитали показатель (*trait value*) *средней соленостной толерантности CCT*, равный средневзвешенному значению минерализации X_i (г/л) для n проб, в которых встретился вид:

$CCT = \sum_n X_i N_i / \sum_n N_i$, где N_i – преобразованное значение численности, экз/м^2.

Для анализа взаимосвязи между родством таксонов и свойствами видов используется концепция *филогенетического сигнала* (Pagel, 1999; Blomberg et al., 2003), которая отражает тенденцию того, что «родственные виды походят друг на друга больше, чем виды, случайно взятые из того же дерева». Если спроецировать филогенетическое дерево на пространство изменения экологических характеристик, то при наличии сильного филогенетического сигнала близкородственные виды будут располагаться рядом, т.е. изменчивость их свойств вдоль листьев дерева будет относительно небольшой.

Для оценки филогенетического сигнала используется алгоритм, который реализует построение так называемых филогенетических обобщенных моделей методом наименьших квадратов. Логика метода заключается в использовании филогенетической информации для подгонки модели эволюции признака, имитирующей броуновское движение, которое не предполагает смену направления, локальных оптимумов, изменений скорости и т. д. Меру филогенетического сигнала можно оценить по двум статистикам, которые равны нулю при отсутствии зависимости и возрастают по мере увеличения корреляционной связи между степенью таксономического родства и тестируемой характеристикой.

Статистика λ Пейджеля (Pagel, 1999) оценивает филогенетический сигнал сравнивая его с корреляцией, ожидаемой при броуновской эволюции. Метод основан на сжатии внутренних ветвей по отношению к верхушке (при λ = 0 сигнал отсутствует и дерево имеет форму полной политомии). Величина λ = 1 означает, что распределение значений признаков по филогении точно такое же, как и ожидаемое по броуновской модели.

```
library(picante)
library(phytools)
traits <- Species_Tax[,c("W","Com")]
CCT <- traits$W
summary(CCT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2700  0.5400  0.8915  3.3582  1.9977 30.5553
names(CCT) <- rownames(traits)
phylosig(PTree, CCT, method = "lambda", test = TRUE)
Phylogenetic signal lambda : 0.803978
logL(lambda) : -1103.83
LR(lambda=0) : 62.3904
P-value (based on LR test) : 2.81695e-15
```

Статистика *K* Блумберга (Blomberg et al., 2003) также основана на броуновской модели случайного дрейфа эволюции, при которой значение *K* = 1. Однако при *K* > 1 считается, что виды более сходны между собой, чем для модели броуновского движения:

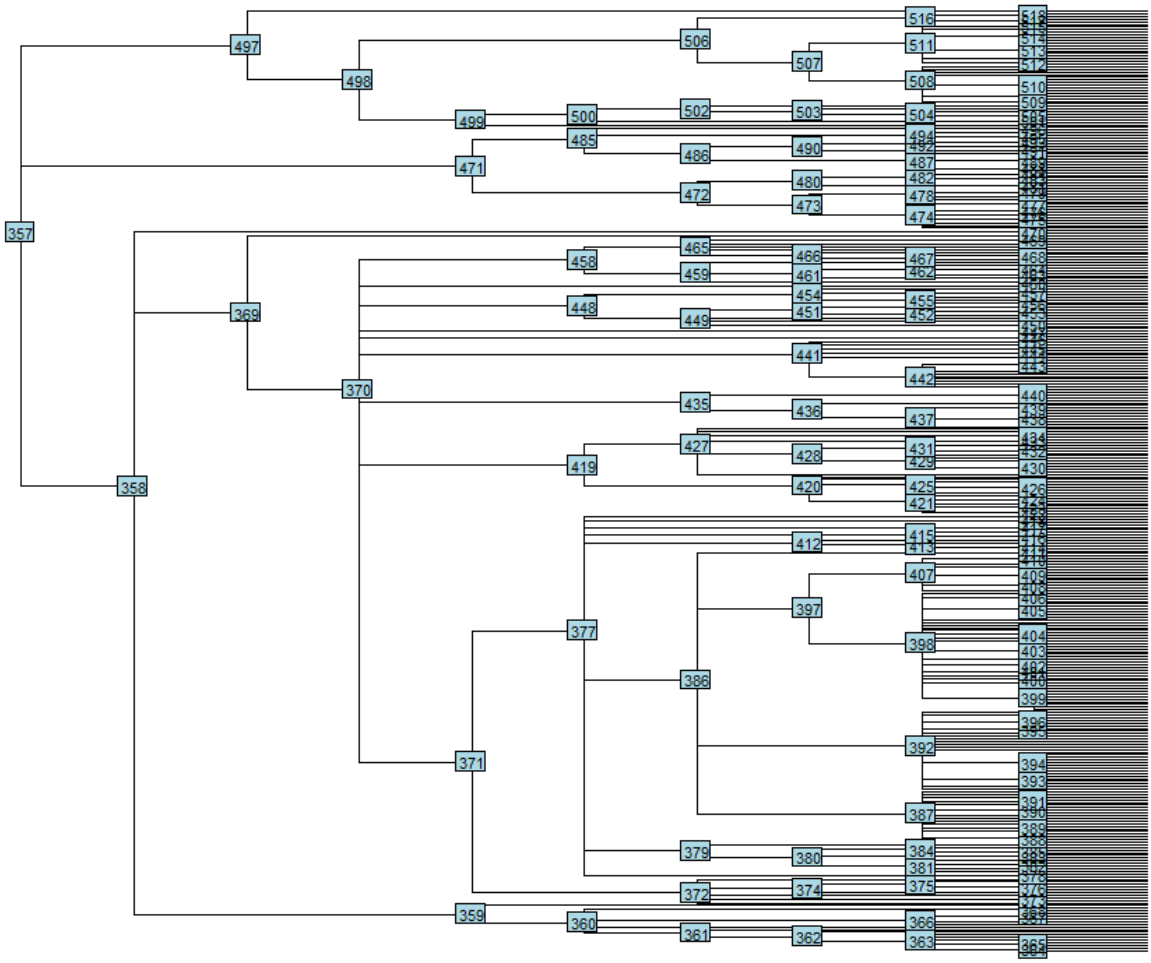
```
apply(traits, 2, Kcalc, PTree)
      CCT       Com
0.3102745 0.1491828
multiPhylosignal(traits, multi2di(PTree))
           K PIC.variance.obs PIC.variance.rnd.mean PIC.variance.P PIC.variance.Z
CCT 0.3102745        2.563954             4.262301         0.001      -5.176719
Com 0.1491828      103.908128            79.997966         0.998       2.721213
```

Статистическую значимость филогенетического сигнала здесь проверялась путем сравнения дисперсии наблюдаемых филогенетически независимых контрастов (*PIC*) признака с нулевой моделью перемешивания меток таксонов по вершинам дерева филогении. Отметим, что филогенетический сигнал для соленосной толерантности `CCT` является статистически значимым (хотя и значительно меньше, чем для модели броуновского движения). В противоположность этому, частота встречаемости видов в пробах `Com` не зависит от филогении, т.е. обнаружение видов из разных групп равновероятно.

4. Выделение клад (фрагментов дерева) и их визуализация

Функция `as.phylo` пакета `ape`, может создать кладограмму прямо на основе таблицы систематики. Длина ветвей такого дерева уже не определена:

```
# Конвертация строк в переменные-факторы
ST  <- lapply(Species_Tax[,-(1:3)],as.factor)
frm <- ~Phylum/SubPhylum/Class/SubClass/Order/SubOrder/Family/SubFamily/Tribe/Genus/Code
ATree <- as.phylo(frm, data = ST)
str(ATree)
List of 3
 $ edge      : int [1:517, 1:2] 357 358 359 360 361 362 363 364 364 363 ...
 $ Nnode     : int 162
 $ tip.label: chr [1:356] "AmDic.c." "AmDic.h." "AmGam.1" "AmGam.px" ...
 - attr(*, "class")= chr "phylo"
 - attr(*, "order")= chr "cladewise"
plot(ATree, show.tip.label = FALSE, node.depth = 2,
                use.edge.length = FALSE)
nodelabels(cex=0.7)
```



Общее число узлов дерева существенно уменьшилось, по сравнению с предыдущим случаем: 356 листьев + 162 внутренних узла = 517 ветвей.

Часто для практической работы необходимо разбиение дерева на фрагменты (клады). Функция `subtrees` создает список всех возможных поддеревьев, количество которых равно числу внутренних узлов:

```
stree <- subtrees(ATree)
# Поддерево, сформированное на основе узла 386
stree [[41]]
Phylogenetic tree with 59 tips and 14 internal nodes.
```

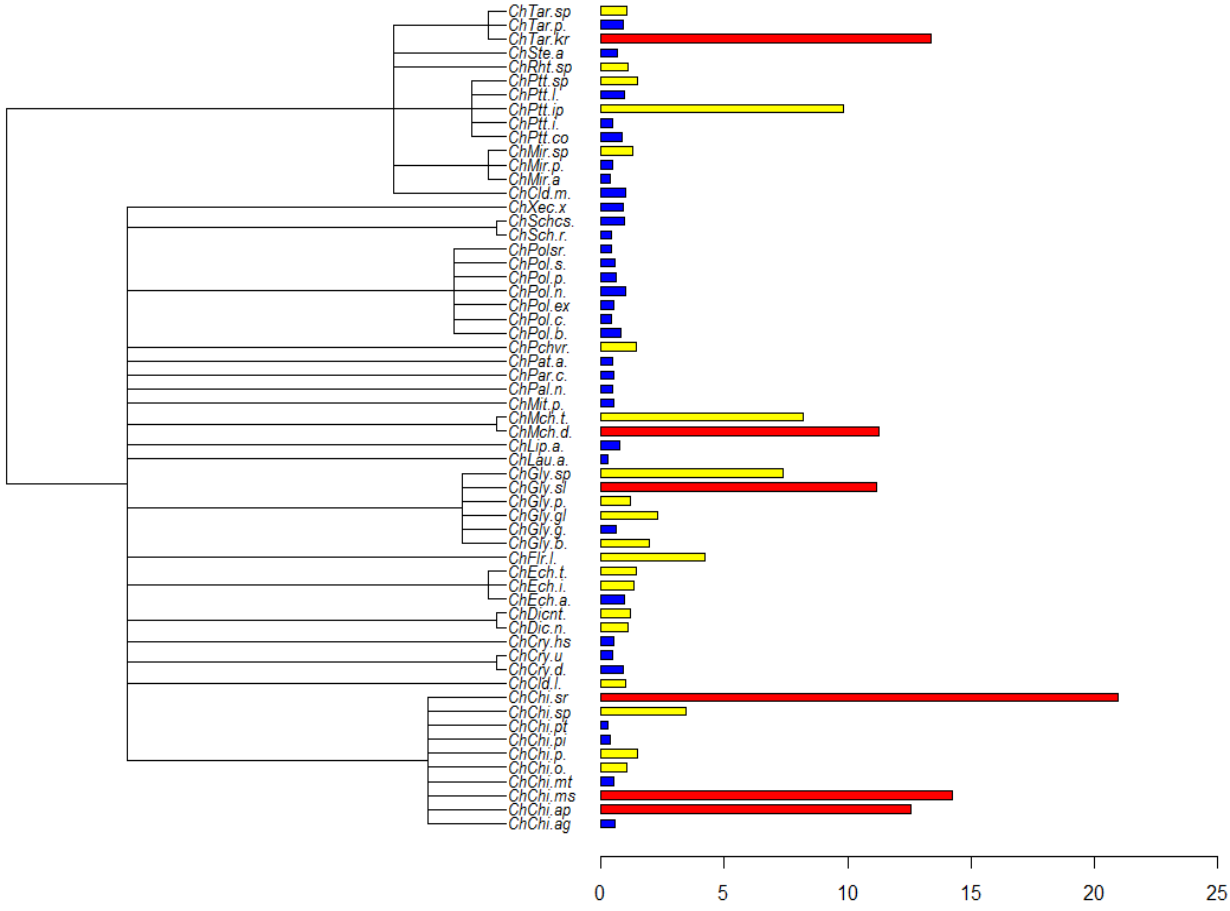
```
Tip labels:
  ChChi.ag, ChChi.ap, ChChi.ms, ChChi.mt, ChChi.o., ChChi.p., ...
Node labels:
  397, 398, 407, 399, 400, 401, ...
Rooted; no branch lengths.
Stree_S <-  stree[[41]]
```

Определим, из каких таксономических уровней состоит выбранная клада:

```
Species_Tax %>% filter(Code %in% Stree_S$tip.label) %>%
   group_by(SubFamily, Tribe) %>% summarize(Ns = n())
# A tibble: 2 x 3
  SubFamily    Tribe          Ns
1 Chironominae Chironomini    45
2 Chironominae Tanytarsini    14
```

Разработано несколько пакетов визуализации и аннотирования, представляющих деревья, как в форме отдельных объектов, так и в сочетании с сопутствующими данными. Приведем пример с использованием пакета `phytools`, в котором индекс средней соленосной толерантности каждого вида показан столбчатой диаграммой

```
target <- tibble(name = Stree_S$tip.label) %>%
  left_join(Species_Tax[,2:4], by = c("name" = "Code")) %>%
  mutate(col = "yellow") %>%  as.data.frame()
target[target$W  > 10, 4] <- "red"
target[target$W < 1, 4] <- "blue"
x <- target$W
names(x) <-  target$name
plotTree.barplot(Stree_S, x,
   args.plotTree=list(fsize=0.8),
   args.barplot=list(col=target$col, xlim=c(0,25)))
```

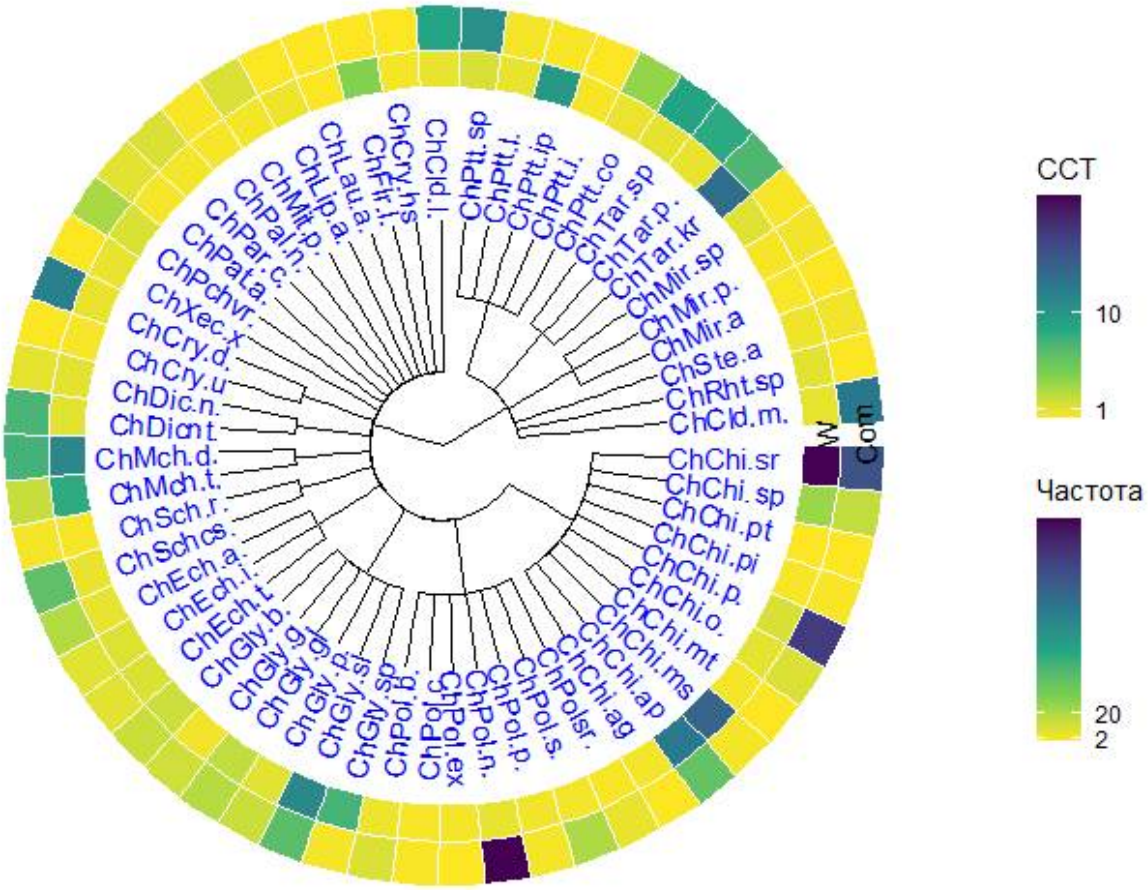


Надежной и программируемой платформой для работы с филогенетическими деревьями является пакет `ggtree` (Yu et al. 2017, см. также детальное описание на https://yulab-smu.top/treedata-book), который позволяет на высоком уровне не только интегрировать и визуализировать многоаспектные данные, но и служит для выявления их ассоциаций и паттернов. `ggtree` создан для работы и отображения древовидных структур на основе грамматики графического пакета `ggplot2` и полностью с ним интегрирован. Пакет выпущен в рамках проекта Bioconductor по создания открытой системы программного обеспечения для вычислительной биологии и биоинформатики. (http://bioconductor.org), поэтому его установка требует предварительной инсталляции функции `BiocManager`, управляющей загрузкой компонентов Bioconductor.

Представим дерево `Stree_S`, свернутое в виде "цирка" (`"circular"`), и листьям его будут соответствовать раскрашенные сегменты колец, связанных с соленосной толерантностью `W` и частотой встречаемости видов в пробах `Com`.

```
BiocManager::install("ggtree")  # Инсталляция пакета
library(ggtree)
library(ggnewscale)
df2 <- target[,2:3]
rownames(df2) <- # Поддерево, сформированное $tip.label

circ <- ggtree(Stree_S, layout = "circular") +
geom_tiplab(aes(angle=angle), color='blue')
pgp1 <- gheatmap(circ, df2["W", drop=F], offset=1.5, width=.2,
         colnames_angle=95, colnames_offset_y = .25) +
     scale_fill_viridis_c( direction= -1, breaks=c(0, 1, 10), name = "CCT")
pgp2 <- pgp1 + new_scale_fill()
gheatmap(pgp2, df2["Com", drop=F], offset=2, width=.2,
         colnames_angle=95, colnames_offset_y = .25) +
  scale_fill_viridis_c(direction= -1, breaks=c(0, 2, 20),name="Частота")
```



5. Использование меток узлов и ветвей дерева

В приведенных выше функциях создания объектов `phylo` отсутствует возможность сохранения имен узлов, которые заменяются сквозной нумерацией. В ряде случаев или в соответствии с характером задачи необходимо нанести на кладограмму систематическе наименования узлов и ветвей, что при построении больших деревьев представляет нетривиальную проблему.

Поставим задачу рассчитать величины соленосной толерантности для всех узлов дерева как средние *CCT* видов, составляющих каждую группу. Ограничим такие расчеты уровнями от Phylum до Family (для упрощения исключив SubOrder) и отобразим полученное дерево в аннотированной форме.

```
library(ape)
library(tidyverse)
df_all <-  Species_Tax %>%
    group_by(Phylum, SubPhylum, Class, SubClass, Order, Family) %>%
    summarize(n=n()) %>%
    select(Phylum, SubPhylum, Class, SubClass, Order, Family) %>%
    as.data.frame()
```

Предварительно приведем скрипты двух весьма полезных функций, осуществляющих конвертацию таксономической таблицы в объект Newick и вывод информации о дереве во вспомогательную таблицу;

```
#++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
#  https://stackoverflow.com/questions/15343338/how-to-convert-a-data-
frame-to-tree-structure-object-such-as-dendrogram
 ## рекурсивная функция
    traverse <- function(a,i,innerl, df){
        if(i < (ncol(df))){
            alevelinner <- as.character(unique(
                          df[which(as.character(df[,i])==a),i+1]))
            desc <- NULL
          if(length(alevelinner) == 1)
             (newickout <- traverse(alevelinner,i+1,innerl, df))
          else {
             for(b in alevelinner)
                desc <- c(desc,traverse(b,i+1,innerl, df))
                il <- NULL; if(innerl==TRUE) il <- a
             (newickout <-
                    paste("(",paste(desc,collapse=","),")",il,sep=""))
            }
        }
        else { (newickout <- a) }
    }
   ## функция конвертации data.frame в newick формат
    df2newick <- function(df, innerlabel=FALSE){
        alevel <- as.character(unique(df[,1]))
        newick <- NULL
        for(x in alevel) newick <- c(newick,traverse(x,1,innerlabel, df))
        (newick <- paste("(",paste(newick,collapse=","),");",sep=""))
    }
#
# Создание таблицы связей узлов дерева
# https://stackoverflow.com/questions/51696837/r-phylo-object-how-to-
connect-node-label-and-node-number
edge_table <- function(my_tree) {
    select.tip.or.node <- function(element, tree) {
        ifelse(element < Ntip(tree)+1,
         tree$tip.label[element], tree$node.label[element-Ntip(tree)])
    }
```

```
    node_labels_in_edge <- my_tree$node.label
                           [my_tree$edge[,1]-Ntip(my_tree)]
    tips_nodes <- my_tree$edge[,2]
    data.frame(
         "is_tip" = my_tree$edge[,2] > length(my_tree$tip.label),
         "parent" = my_tree$edge[,1],
         "par.name" = sapply(my_tree$edge[,1],
                      select.tip.or.node, tree = my_tree),
         "child" = my_tree$edge[,2],
         "chi.name" = sapply(my_tree$edge[,2],
                      select.tip.or.node, tree = my_tree)
               )
}
# ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```

С использованием этих двух функций выполним построение дерева, которое содержит 80 листьев, соответствующих списку семейств, и 29 внутренних узлов, отражающих их таксономическую иерархию. Каждая из 108 ветвей дерева связывает родительский узел (`parent`) с дочерним узлом (`child`). При `is_tip = FALSE` дочерние узлы являются листьями:

```
   myNewick <- df2newick(df_all, TRUE)
# Создание дерева phylo из объекта newick
   All_Tree  <- read.tree(text = myNewick)
   str(All_Tree)
List of 4
 $ edge      : int [1:108, 1:2] 81 82 83 84 84 85 85 85 83 86 ...
 $ Nnode     : int 29
 $ node.label: chr [1:29] "" "Annelida" "Clitellata" "Lumbriculata" ...
 $ tip.label : chr [1:80] "Gnathobdellidae" "Erpobdellidae"
"Glossiphoniidae" "Piscicolidae" ...
 - attr(*, "class")= chr "phylo"
 - attr(*, "order")= chr "cladewise"
# Создание таблицы связей узлов дерева
df_apAll <- edge_table(All_Tree)
# Связи узлов дерева с листьями
head(df_apAll[!df_apAll$is_tip,])
   is_tip parent     par.name child        chi.name
4   FALSE     84 Lumbriculata     1 Gnathobdellidae
6   FALSE     85   Hirudinida     2   Erpobdellidae
7   FALSE     85   Hirudinida     3 Glossiphoniidae
8   FALSE     85   Hirudinida     4    Piscicolidae
10  FALSE     86  Oligochaeta     5   Enchytraeidae
12  FALSE     87   Tubificida     6        Naididae
# Связи внутренних узлов дерева между собой
head(df_apAll[df_apAll$is_tip,])
   is_tip parent     par.name child     chi.name
1    TRUE     81                82     Annelida
2    TRUE     82     Annelida    83   Clitellata
3    TRUE     83   Clitellata    84 Lumbriculata
5    TRUE     84 Lumbriculata    85   Hirudinida
9    TRUE     83   Clitellata    86  Oligochaeta
11   TRUE     86  Oligochaeta    87   Tubificida
```

Добавим в эту таблицу справа столбцы, необходимые для вывода содержательной информации для аннотирования дерева – среднюю соленосную толерантность групп, ее стандартное отклонение и число видов в кладе:

```
# Подсчет средней соленосной толерантности для идентификаторов
# каждого таксономического уровня
Fam_stat <-Species_Tax %>% group_by(Family) %>%
     summarise(n=n(), m=mean(W), sd = sd(W)) %>% as.data.frame()
```

```
Order_stat <- Species_Tax %>% group_by(Order) %>%
     summarise(n=n(), m=mean(W), sd = sd(W)) %>% as.data.frame()
SubClass_stat <- Species_Tax %>% group_by(SubClass) %>%
     summarise(n=n(), m=mean(W), sd = sd(W)) %>% as.data.frame()
Class_stat <- Species_Tax %>% group_by(Class) %>%
     summarise(n=n(), m=mean(W), sd = sd(W)) %>% as.data.frame()
SubPhylum_stat <- Species_Tax %>% group_by(SubPhylum) %>%
     summarise(n=n(), m=mean(W), sd = sd(W)) %>% as.data.frame()
Phylum_stat <- Species_Tax %>% group_by(Phylum) %>%
     summarise(n=n(), m=mean(W), sd = sd(W)) %>% as.data.frame()
# Заполнение таблицы df_apAll для листьев
df_apAll  <-  df_apAll %>%
             left_join(Fam_stat, by = c("chi.name" = "Family"))
head(df_apAll[!df_apAll$is_tip,])
   is_tip parent     par.name child        chi.name  n         m        sd
4   FALSE     84 Lumbriculata     1 Gnathobdellidae  1 0.3668533        NA
6   FALSE     85   Hirudinida     2   Erpobdellidae  2 1.0093906 0.4251258
7   FALSE     85   Hirudinida     3 Glossiphoniidae  4 0.8872398 0.3898640
8   FALSE     85   Hirudinida     4    Piscicolidae  1 1.2404873        NA
10  FALSE     86  Oligochaeta     5   Enchytraeidae  4 8.5527878 9.4576076
12  FALSE     87   Tubificida     6        Naididae 15 3.5347776 4.2008553
# Заполнение таблицы df_apAll для внутренних узлов
df_apAll[is.na(df_apAll$n),] <-  df_apAll[is.na(df_apAll$n),] %>%
    select(1:5) %>%  left_join(Order_stat, by = c("chi.name" = "Order"))
df_apAll[is.na(df_apAll$n),] <-  df_apAll[is.na(df_apAll$n),] %>%
 select(1:5) %>%  left_join(SubClass_stat, by = c("chi.name" = "SubClass"))
df_apAll[is.na(df_apAll$n),] <-  df_apAll[is.na(df_apAll$n),] %>%
      select(1:5) %>%  left_join(Class_stat, by = c("chi.name" = "Class"))
df_apAll[is.na(df_apAll$n),] <-  df_apAll[is.na(df_apAll$n),] %>%
      select(1:5) %>%
      left_join(SubPhylum_stat, by = c("chi.name" = "SubPhylum"))
df_apAll[is.na(df_apAll$n),] <-  df_apAll[is.na(df_apAll$n),] %>%
    select(1:5) %>%  left_join(Phylum_stat, by = c("chi.name" = "Phylum"))
```

Создадим строковые переменные в формате, принятом для отображения

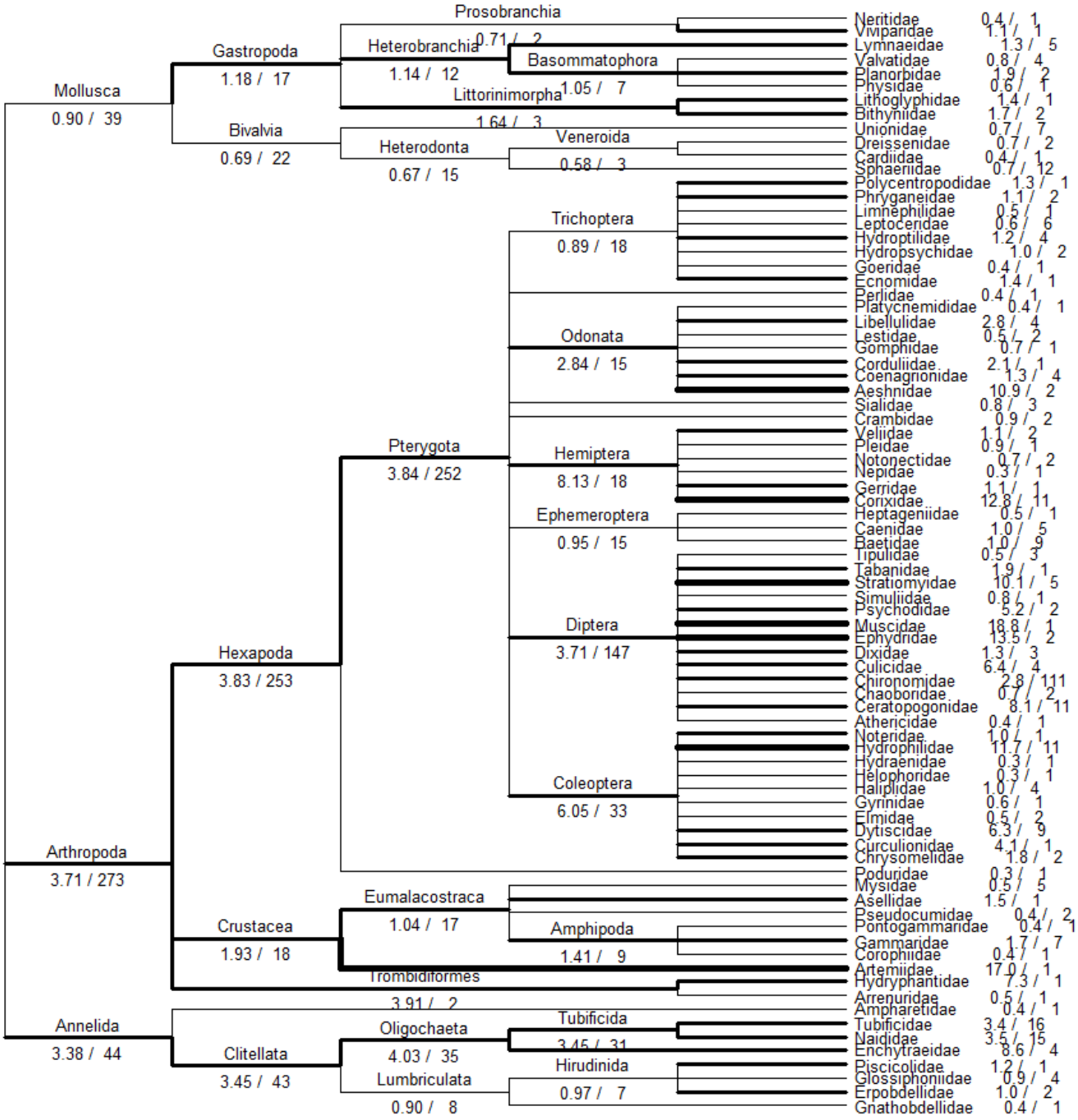
```
df_apAll$node.label <- NA
df_apAll[!df_apAll$is_tip,9] <- with(df_apAll[!df_apAll$is_tip,],
        paste(format(chi.name, width = 19, justify = "left"),
        paste(format(m, digits = 1),format(n), sep=" / ")))
df_apAll[df_apAll$is_tip,9] <- with(df_apAll[df_apAll$is_tip,],
            paste(format(m, digits = 2),format(n), sep=" / "))
```

Создадим переменную, определяющую толщину линий ветвей:

```
df_apAll$edge_with <- 3
df_apAll[df_apAll$m  > 10, 10] <- 5
df_apAll[df_apAll$m < 1, 10] <- 1
```

Выведем кладограмму и всю сопутствующую ей информацию:

```
All_Tree$tip.label <- df_apAll[!df_apAll$is_tip,9]
N_ed <- as.integer(rownames(df_apAll[df_apAll$is_tip,]))
Name_ed <- as.character(df_apAll[N_ed,5])
m_ed <- df_apAll[N_ed,9]
par(mar=c(0,0,0,0))
plot(All_Tree, use.edge.length = FALSE, font = 1,
     edge.width = df_apAll$edge_with , node.depth = 2,
     cex = 0.8, label.offset = .05, family = "mono")
edgelabels(text=Name_ed, edge= N_ed, adj = c(0.5, -0.5),
     cex = 0.8, frame = "none")
edgelabels(text=m_ed, edge= N_ed, adj = c(0.5, 1.5),
      cex = 0.8, frame = "none")
```


Mollusca
0.90 / 39
Gastropoda
1.18 / 17
Bivalvia
0.69 / 22
Prosobranchia
0.71 / 2
Heterobranchia
1.14 / 12
Basommatophora
1.05 / 7
Littorinimorpha
1.64 / 3
Heterodonta
0.67 / 15
Veneroida
0.58 / 3
Arthropoda
3.71 / 273
Hexapoda
3.83 / 253
Pterygota
3.84 / 252
Trichoptera
0.89 / 18
Odonata
2.84 / 15
Hemiptera
8.13 / 18
Ephemeroptera
0.95 / 15
Diptera
3.71 / 147
Coleoptera
6.05 / 33
Crustacea
1.93 / 18
Eumalacostraca
1.04 / 17
Amphipoda
1.41 / 9
Trombidiformes
3.91 / 2
Annelida
3.38 / 44
Clitellata
3.45 / 43
Oligochaeta
4.03 / 35
Tubificida
3.45 / 31
Lumbriculata
0.90 / 8
Hirudinida
0.97 / 7
Neritidae 0.4 / 1
Viviparidae 1.1 / 1
Lymnaeidae 1.3 / 5
Valvatidae 0.8 / 4
Planorbidae 1.9 / 2
Physidae 0.6 / 1
Lithoglyphidae 1.4 / 1
Bithyniidae 1.7 / 2
Unionidae 0.7 / 7
Dreissenidae 0.7 / 2
Cardiidae 0.4 / 1
Sphaeriidae 0.7 / 12
Polycentropodidae 1.3 / 1
Phryganeidae 1.1 / 2
Limnephilidae 0.5 / 1
Leptoceridae 0.6 / 6
Hydroptilidae 1.2 / 4
Hydropsychidae 1.0 / 2
Goeridae 0.4 / 1
Ecnomidae 1.4 / 1
Perlidae 0.4 / 1
Platycnemididae 0.4 / 1
Libellulidae 2.8 / 4
Lestidae 0.5 / 2
Gomphidae 0.7 / 1
Corduliidae 2.1 / 1
Coenagrionidae 1.3 / 4
Aeshnidae 10.9 / 2
Sialidae 0.8 / 3
Crambidae 0.9 / 2
Veliidae 1.1 / 2
Pleidae 0.9 / 1
Notonectidae 0.7 / 2
Nepidae 0.3 / 1
Gerridae 1.1 / 1
Corixidae 12.8 / 11
Heptageniidae 0.5 / 1
Caenidae 1.0 / 5
Baetidae 1.0 / 9
Tipulidae 0.5 / 3
Tabanidae 1.9 / 1
Stratiomyidae 10.1 / 5
Simuliidae 0.8 / 1
Psychodidae 5.2 / 2
Muscidae 18.8 / 1
Ephydridae 13.5 / 2
Dixidae 1.3 / 3
Culicidae 6.4 / 4
Chironomidae 2.8 / 111
Chaoboridae 0.7 / 2
Ceratopogonidae 8.1 / 11
Athericidae 0.4 / 1
Noteridae 1.0 / 1
Hydrophilidae 11.7 / 11
Hydraenidae 0.3 / 1
Helophoridae 0.3 / 1
Haliplidae 1.0 / 4
Gyrinidae 0.6 / 1
Elmidae 0.5 / 2
Dytiscidae 6.3 / 9
Curculionidae 4.1 / 1
Chrysomelidae 1.8 / 2
Podiceridae 0.3 / 1
Mysidae 0.5 / 5
Asellidae 1.5 / 1
Pseudocumidae 0.4 / 2
Pontogammaridae 0.4 / 1
Gammaridae 1.7 / 7
Corophiidae 0.4 / 1
Artemiidae 17.0 / 1
Hydryphantidae 7.3 / 1
Arrenuridae 0.5 / 1
Ampharetidae 0.4 / 1
Tubificidae 3.4 / 16
Naididae 3.5 / 15
Enchytraeidae 8.6 / 4
Piscicolidae 1.2 / 1
Glossiphoniidae 0.9 / 4
Erpobdellidae 1.0 / 2
Gnathobdellidae 0.4 / 1

ЛИТЕРАТУРА

Лукашов В.В. Молекулярная эволюция и филогенетический анализ. М.: БИНОМ. Лаборатория знаний, 2009. 256 с.

Шитиков В.К., Зинченко Т.Д. Использование чисел Хилла для оценки видового и таксономического разнообразия в группах местообитаний // Принципы экологии. 2013б. № 3. С. 23 - 36.

Blomberg S. P., Garland T., Ives A.R. Testing for phylogenetic signal in comparative data: behavioral traits are more labile // Evolution. 2003. V. 57. P. 717-745.

Chao A., Chiu C.-H., Jost L. Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers // Annual Review of Ecology, Evolution, and Systematics. 2014. V. 45, No 1. P. 297–324.

Christoffersen M. Cladistic Taxonomy, Phylogenetic Systematics, and Evolutionary Ranking // Systematic Biology. 1995. V. 44, N 3. P. 440-454.

Clarke K.R., Warwick R.M. A further biodiversity index applicable to species lists: variation in taxonomic distinctness // Mar. Ecol. Prog. 2001. Ser. 216. P. 265–278.

Ives A.R. Mixed and Phylogenetic Models: A Conceptual Introduction to Correlated Data. Leanpub book. 2018. URL: https://leanpub.com/correlateddata.

Pagel M. Inferring the historical patterns of biological evolution // Nature. 1999. V. 401. P. 877–884.

Petchey O.L., Gaston K.J. Functional diversity (FD), species richness and community composition // Ecology Letters. 2002. V. 5, No 3. P. 402–411.

Swenson N.G. Functional and phylogenetic ecology in R. N.Y.:Springer, 2014. 212 p.

Tilman D. Functional diversity / Encyclopedia of Biodiversity (ed. Levin, S.A.). San Diego:Academic Press, 2001. P. 109-120.

Tilman D. Functional diversity / Encyclopedia of Biodiversity (ed. Levin, S.A.). San Diego:Academic Press, 2001. P. 109-120.

Webb C.O., Ackerly D.D., McPeek M.A., Donoghue M.J. Phylogenies and community ecology // Annual Review of Ecology and Systematics. 2002. V. 33, No 1. P. 475–505.

Yu G., Smith D.K., Zhu H., Guan Y., Lam T. `Ggtree`: An r Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data // Methods in Ecology and Evolution. 2017. V. 8, N 1. P. 28–36.