Программа НИС по вычислительной филогенетике

С.А.Спирин, осень 2019

# Комбинаторика деревьев

Терминология (лист, ветвь, узел, клада, корень).

Топология дерева как набор разбиений множества листьев.

Смысл длин ветвей.

Число различных разрешённых деревьев с n листьями (укоренённых и неукоренённых).

Что такое “ультраметричность” (применительно к филогенетическому дереву)

# Принципиальная схема реконструкции филогении

Дистанционные и символьные методы

Прямые и переборные методы

# Оценка эволюционных расстояний между последовательностями

Формула Джукса – Кантора (Jukes-Cantor)

Оценка по принципу максимального правдоподобия

# Поиск в пространстве деревьев

Выращивание (stepwise addition)

NNI

SPR

(\*) TBR

(\*) Генетические алгоритмы

(\*) Markov chain Monte Carlo

# Критерии качества дерева

Максимальная экономия (parsimony)

Максимальное правдоподобие (likelihood)

Минимальная эволюция (minimum evolution)

Наименьшие квадраты (OLS = Ordinary least squares, Fitch – Margoliash)

(\*) Другие (например, квартеты)

# Прямые методы реконструкции филогении

UPGMA

Neighbor-joining

# Укоренение

Укоренение в среднюю точку (midpoint)

Внешняя группа (outgroup)

# Bootstrap и jackknife

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Составные части зачёта

1. Две контрольные 25%
2. Журнальный клуб:
   1. Собственный доклад 30%
   2. Представление о содержании других докладов — в рамках экзамена
3. Практическая работа 25%
4. Экзамен 20%

# Содержание практической работы

## Два варианта:

1. Взять статью, где описывается реконструкция филогении какой-либо группы организмов и повторить часть работы (15–30 организмов) другим способом (те же последовательности, но другой софт, или же другие последовательности из тех же видов или родов и т.п.). Сравнить результаты.  
   Можно не организмы, а мультигенное семейство (ортологи и паралоги из нескольких организмов).
2. Взять набор организмов по своему усмотрению, подобрать ортологичные последовательности, реконструировать несколькими способами.

Технические вопросы задавайте по e-mail [sspirin@hse.ru](mailto:sspirin@hse.ru) . Ответы на наиболее очевидные ниже.

### Где брать последовательности?

* В Supplementary Data к статьям.
* В Uniprot (там хороший Advanced Search).
  + Если у вас есть примерный список организмов, можно попробовать взять белки, закодированные в митохондриях, например цитохром B (Uniprot ID начинается с CYB\_ ) или первую субъединицу цитохромоксидазы (начинается с COX1\_ ). Все их проще всего их найти, вбив в окошко поиска на сайте <https://www.uniprot.org/> строку  
    mnemonic:cox1\_\*  
    или  
    mnemonic:cyb\_\*  
    В крайнем случае можно сначала найти все такие записи, а потом из полученного набора отобрать интересные организмы.
  + Если есть и организмы, и названия белков, разбирайтесь с поиском (ссылка Advanced справа вверху), там всё довольно интуитивно. Варианты названия организма или таксона выпадают как меню по первым буквам этого названия.
* В NCBI (там с поиском по аннотации плохо, но справиться можно, например, используя BLAST). Но часто в статьях приводятся идентификаторы GenBank или RefSeq, тогда проблемы нет.

### Рекомендуемые программы реконструкции филогении

Бонус в случае использования **установленных на своём компьютере** программ из следующего списка:

* FastME
* TREE-PUZZLE
* MrBayes
* PhyloBayes

Но (без бонуса) можно и другими, например MEGA, или на сайте [phylogeny.fr](http://www.phylogeny.fr/) .

### Программы выравнивания

Помните, что невыровненные последовательности нельзя подавать на вход программам реконструкции филогении!

* Clustal Omega <https://www.ebi.ac.uk/Tools/msa/clustalo/>
* MAFFT <https://www.ebi.ac.uk/Tools/msa/mafft/>
* Muscle <https://www.ebi.ac.uk/Tools/msa/muscle/>

Все три программы можно использовать не только онлайн, но и установив на своём компьютере, а также через редактор выравниваний Jalview.

Кроме того, алгоритмы Muscle и MAFFT встроены в MEGA.

### Проблема переформатирования выравнивания

Многие филогенетические программы не понимают обычных форматов, выдаваемых программами выравнивания. Разберитесь, какие форматы требуются на вход (например, NEXUS/PAUP или PHYLIP), и переформатируйте выравнивание программой seqret:

* Онлайн на сайте <https://www.ebi.ac.uk/Tools/sfc/emboss_seqret/>
* Или поставьте на своём компьютере и освойте пакет EMBOSS (пригодится ещё не раз!), туда входит программа seqret.

### Изображение дерева

MEGA выдаёт довольно качественное изображение по формату Newick. Есть и другие программы для рисования деревьев, см. <https://en.wikipedia.org/wiki/List_of_phylogenetic_tree_visualization_software>

Если выходной формат филогенетической программы — NEXUS, то придумайте что-нибудь (например, найдите или напишите скрипт для переформатирования, или найдите в списке визуализирующую программу, которая понимает этот формат).

## Отчёт

Очень краткий: что делалось, что получилось, выводы.

Но текст должен быть понятен стороннему читателю.

## Несколько статей с описанием работы по реконструкции филогении (*список прошлогодний, будет дополняться*)

1. Erez Shpirer, Arik Diamant, Paulyn Cartwright and Dorothée Hucho. A genome wide survey reveals multiple nematocyst-specific genes in Myxozoa. BMC Evolutionary Biology 2018 18:138 <https://doi.org/10.1186/s12862-018-1253-7>   
   См. Additional Files 3–7. Там есть уже готовые выравнивания в формате NEXUS (под MrBayes, например).
2. Dong Zhang et al. Three new Diplozoidae mitogenomes expose unusual compositional biases within the Monogenea class: implications for phylogenetic studies. BMC Evolutionary Biology 2018 18:133. <https://doi.org/10.1186/s12862-018-1249-3>   
   См. Additional File 13, тем есть идентификаторы в NCBI для полных митохондриальных геномов. Нужно будет выбрать штук 10, открыть каждый в NCBI (вот как здесь: <https://www.ncbi.nlm.nih.gov/nuccore/NC_014591.1/> ) и выбрать в описании какой-нибудь один и тот же белок (например, cytochrome B или COX1), по этим белкам строить дерево.
3. K.E. Bräuer et al. Phylogenetic and genomic analyses of the ribosomal oxygenases Riox1 (No66) and Riox2 (Mina53) provide new insights into their evolution. BMC Evolutionary Biology 2018 18:96. <https://doi.org/10.1186/s12862-018-1215-0>
4. M.F. Medina et al. Molecular phylogeny of Panaspis and Afroablepharus skinks (Squamata: Scincidae) in the savannas of sub-Saharan Africa. Molecular Phylogenetics and Evolution. Volume 100, Pages 409-423. <https://doi.org/10.1016/j.ympev.2016.04.026>  
   Здесь прямо в тексте есть таблица с идентификаторами последовательностей.
5. Sunjoo Joo et al. Common ancestry of heterodimerizing TALE homeobox transcription factors across Metazoa and Archaeplastida. BMC Biology 2018 16:136. <https://doi.org/10.1186/s12915-018-0605-5>   
   Additional File 1 (Table S2) содержит прямо последовательности доменов, но чтобы отобрать нужные для, например, сравнения с рис. 1, потребуются некоторые усилия.
6. A.A. Krinitsina et al. The systematic position of Dryopteris blanfordii subsp. nigrosquamosa (Ching) Fraser-Jenkins within the genus Dryopteris Adans. PhytoKeys. 2017; (90): 89–112. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5784233/>