Реконструкция деревьев методом максимального правдоподобия

Программа Phyml, которую мы будем использовать для реконструкции филогенетического дерева, является одной из наиболее старых программ по реконструкции деревьев методом максимального правдоподобия.

В отличие от более современных программ (например, raxml, fastml) она позволяет протестировать почти любую модель эволюции нуклеотидных последовательностей. Семь моделей закодированы в готовом виде (JC69, K80, F81, HKY85, F84,TN93, GTR). Кроме того, пользователь может добавить свою пользовательскую модель. Также, есть возможность оценки формы Гамма-распределения (gamma distribution), которое определяет разницу в скорости замен между различными нуклеотидами, а также пропорцию неэволюционирующих позиций (invariable sites).

Алгоритмически поиск оптимальной топологии в программе устроен следующим образом:

- Расчитать дерево дистантным методом neighbor-joining;
- Обновить топологию дерева методом NNI или SPR (или оба), сохраняя параметры модели и длины ветвей;
- Обновить часть длин ветвей дерева, сохраняя топологию и параметры модели эволюции;
- Обновить параметры модели эволюции, сохраняя топологию и длины ветвей;
- Повторять процесс до тех пор, пока значение правдоподобия (log-likelihood) не перестанет улучшаться;

Обратите внимание, программа может читать данные только в формате Phylip. Внимательно прочитайте Practicle1_dataFormats.pdf, чтобы ознакомиться с особенностями разных форматов для филогенетики.

В этой практической работе мы реконструируем филогенетические деревьев для группы рыб-клоунов. Для этого мы используем три разных гена – bmp4, cytB и rag1. Перед тем, как начать работу с данными прочитайте описание группы, которую мы будем изучать.

В ходе этого практического занятия, мы должны найти ответы на следующие вопросы:

1. Какая эволюционная модель для замен нуклеотидов наиболее вероятна для каждого гена?
2. Какова топология деревьев полученных для каждого гена?
3. Какова поддержка бутстрапа для узлов деревьев?

Приступаем к практической части.

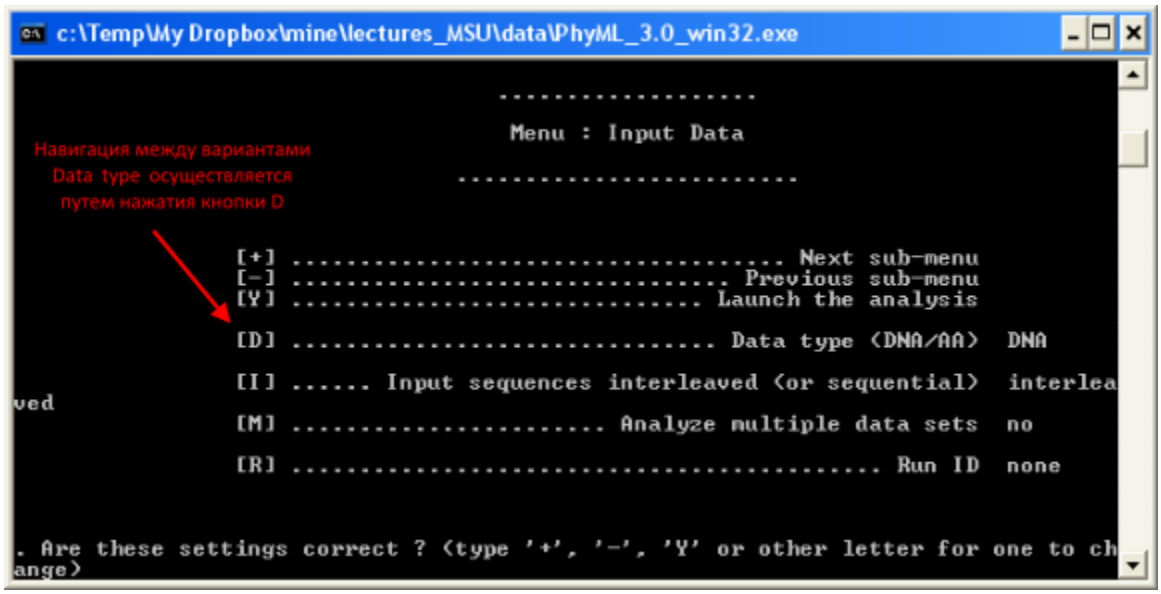
Часть 1. Ген bmp4

1. Определите, где на компьютере установлена програма phyml. Если она не установлена или вы не можете ее найти, скачайте дистрибутив.

2. В ту же папку, куда вы разархивировали дистрибутив программы (или в той же папке где она установлена), скачайте архив с тремя файлами генов и разархивируйте.

3. Для гена bmp4 сконвертируйте данные из формата фаста (Clownfish_bmp4.fst) в формат Phylip. Подсказка – используйте для конвертации программу seaview – если она не установлена, скачайте ее с сервера.

4. Запустите программу Phyml двойным щелчком мыши.
   Навигация по опциям меню программы осуществляется с помощью клавиш + - . Выбор конкретных вариантов из опций меню осуществляется нажатием соответствующей буквы (D, I, M, R, и т.д. – смотреть пример и принтскрин ниже). Запуск анализа – кнопка Y.



5. После запуска программы, вам необходимо указать название входного файла в формате phylip‘Enter the sequence filename>’
   `'Enter the sequence file name >Clownfish_bmp4.phylip'`

6. Просмотрите все доступные опции с помощью кнопок + -

7. Выберите следующие настройки для анализа:
   - Для модели эволюции нуклеотидной последовательности укажите Jukes-Cantor 60 **`(опция JC69 в меню M [Model of nucleotide substitution])`**

- Отключите оценку формы Гамма-распределения (**опция yes в меню R[One category of substitution rate (yes/no)]**)

8. Запустите анализ нажав кнопку Y.

9. После окончания анализа, ответьте на следующие вопросы (запишите ответы в поля ниже):

   - Какие файлы были созданы программой phyml? Что в них находится?

   - Каково правдоподобие (Log-likelihood) полученного дерева (модель JC69)? Где вы можете найти эту информацию?

   - Какой алгоритм эвристического поиска топологии мы использовали?

   - Сколько свободных параметров оптимизируется в модели JC69 (подсказка – обратитесь к таблице в конце этого документа)?

10. Изучите полученное дерево в программе FigTree. Есть ли в дереве политомии? Где? В чем может быть причина наличия (отсутствия) политомий?

11. Перезапустите анализ, но в этот раз выберите модель HKY+ Гамма (**опции HKY85 в меню M[Model of nucleotide substitution**] и no в меню **R[One category of substitution rate (yes/no)]**.

12. Кроме того, укажите Best of NNI and SPR в меню **S[Tree topology search operations]**.

13. После окончания анализа, ответьте на следующие вопросы (запишите ответы в поля ниже):
    - Каково правдоподобие (Log-likelihood) полученного дерева (модель HKY85)?

    - Какой алгоритм эвристического поиска топологии мы использовали? Чем он отличается от поиска в предыдущем случае?

    - Сколько свободных параметров оптимизируется в модели HKY85 + Gamma?

- Какая модель HKY85 +Gamma или JC69 – лучшим образом апроксимируют наши данные? Каким образом мы можем это определить?

Часть 2. Гены rag1 и cytB.

1. Сконвертируйте файлы Clownfish_rag1.fst и Clownfish_cytb.fst в phylip формат

2. Определите, какая эволюционная модель оптимальна для всех трех генов (rag1, cytb, bmp4). Для этого воспользуйтесь программой jModelTest. Если она не установлена у вас на компьютере, скачайне дистрибутив по ссылке.

3. Загрузите в нее файлы в phylip формате и выберите в главном меню Analysis > Compute Likelihood Scores. Установите настройки как на картинке ниже:

4. Ответьте на следующие вопросы:

Каково правдоподобие (Log-likelihood) и AIC (Analysis > Do AIC calculations) для генов cytB и rag1? Для гена bmp4, который мы анализировали раньше? Объясните разницу между AIC и LogLik значениями.

Заполните таблицу ниже и отметьте лучшую эволюционную модель для каждого гена.

Модель	**LogL bmp4**	**AIC bmp4**	**LogL cytB**	**AIC cytB**	**LogL rag1**	**AIC rag1**
HKY						
TrN						
TPM1uf						
GTR						
SYM						
K80						
TrNef						
TPM1						
F81						
JC						

5. В Phyml реконструируйте дерево для каждого гена с использованием оптимальной модели выбранной выше с помощью jModelTest. Используйте опцию best of NNI and SPR в меню `S [Tree topology search operations]` для поиска оптимальной топологии

6. Сравните топологии деревьев полученных для трех генов в FigTree. Как они отличаются? Опишите отличия и подготовьте pdf файлы для каждого гена

7. С помощью программы sequencematrix объедините все три гена (выровненных). Скачать программу, если она не установлена у вас на компьюетере, можно по этой ссылке. Экспортируйте матрицу объединенных генов в формат “naked” (Garli). См. картинку ниже.

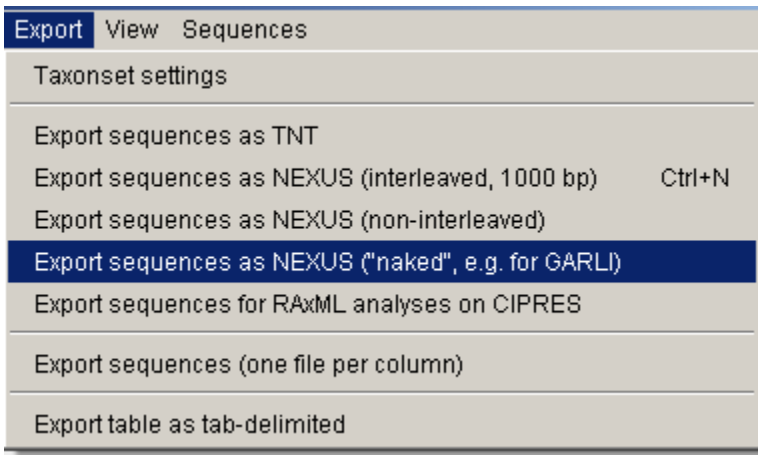

Export
View
Sequences
Taxonset settings
Export sequences as TNT
Export sequences as NEXUS (interleaved, 1000 bp) Ctrl+N
Export sequences as NEXUS (non-interleaved)
Export sequences as NEXUS ("naked", e.g. for GARLI)
Export sequences for RAxML analyses on CIPRES
Export sequences (one file per column)
Export table as tab-delimited

Таблица для определения количества свободных параметров для разных моделей

Model	Base frequencies	Substitution rates	Number of free parameters
JC	$\pi_A = \pi_C = \pi_G = \pi_T$	$\phi_{A\text{-}C} = \phi_{A\text{-}G} = \phi_{A\text{-}T} = \phi_{C\text{-}G} = \phi_{C\text{-}T} = \phi_{G\text{-}T}$	0
K80	$\pi_A = \pi_C = \pi_G = \pi_T$	$\phi_{A\text{-}C} = \phi_{A\text{-}T} = \phi_{C\text{-}G} = \phi_{G\text{-}T} \neq \phi_{A\text{-}G} = \phi_{C\text{-}T}$	1
SYM	$\pi_A = \pi_C = \pi_G = \pi_T$	$\phi_{A\text{-}C} \neq \phi_{A\text{-}G} \neq \phi_{A\text{-}T} \neq \phi_{C\text{-}G} \neq \phi_{C\text{-}T} \neq \phi_{G\text{-}T}$	5
F81	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$	$\phi_{A\text{-}C} = \phi_{A\text{-}G} = \phi_{A\text{-}T} = \phi_{C\text{-}G} = \phi_{C\text{-}T} = \phi_{G\text{-}T}$	3
HKY	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$	$\phi_{A\text{-}C} = \phi_{A\text{-}T} = \phi_{C\text{-}G} = \phi_{G\text{-}T} \neq \phi_{A\text{-}G} = \phi_{C\text{-}T}$	4
GTR	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$	$\phi_{A\text{-}C} \neq \phi_{A\text{-}G} \neq \phi_{A\text{-}T} \neq \phi_{C\text{-}G} \neq \phi_{C\text{-}T} \neq \phi_{G\text{-}T}$	8

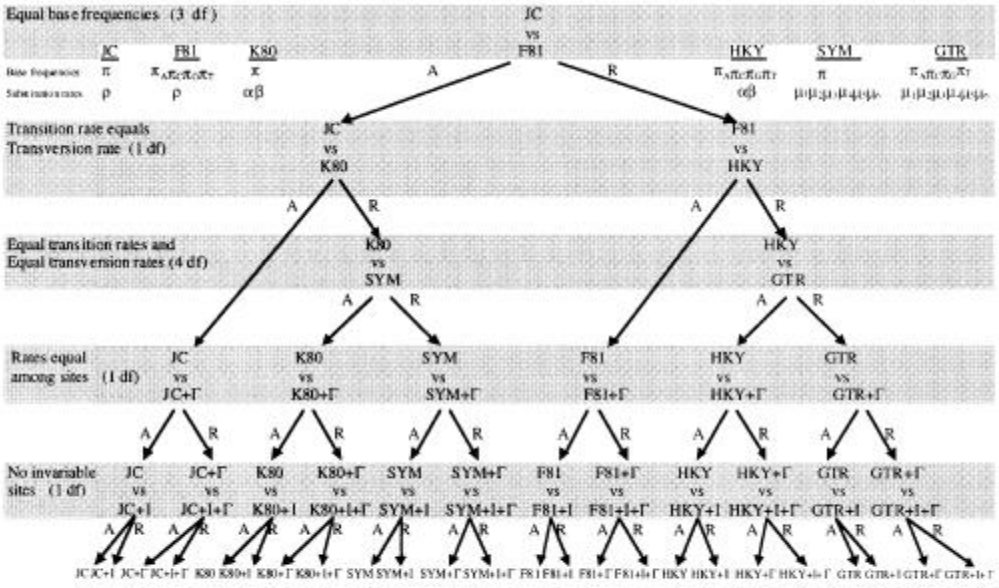


Fig. 1. Hierarchical hypothesis testing in MODELTEST. At each level the null hypothesis (upper model) is either accepted (A) or rejected (R). The models of DNA substitution are: JC (Jukes and Cantor, 1969), K80 (Kimura, 1980), SYM (Zharkikh, 1994), F81 (Felsenstein, 1981), HKY (Hasegawa *et al.*, 1985), and GTR (Rodriguez *et al.*, 1990). Γ: shape parameter of the gamma distribution; I: proportion of invariable sites. df: degrees of freedom. 1: equal base frequencies (0.25), π_A: frequency of adenine, π_C: frequency of cytosine, π_G: frequency of guanine, π_T: frequency of thymine. ρ: equal substitution rate, α: transition rate, β: transversion rate; μ_1: A⇒C rate, μ_2: A⇒G rate, μ_3: A⇒T rate, μ_4: C⇒G rate, μ_5: C⇒T rate, μ_6: G⇒T rate.