МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ
ОДЕССКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ ИМЕНИ И. И. МЕЧНИКОВА
БИОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

Н. Ю. Васильева

БИОИНФОРМАТИКА

Множественное выравнивание. Филогенетические деревья

Методическое пособие

ОДЕССА
ОНУ
2014

УДК 575.22:004.42(075.8)
ББК 28.042в65я73
В 191

Печатается по решению ученого совета
биологического факультета университета имени И. И. Мечникова.
Протокол № 2 от 10.10.2013 г.

Рецензенты:

Т. В. Гудзенко – доцент, кандидат биологических наук, доцент кафедры микробиологии, вирусологии и биотехнологии Одесского национального университета имени И. И. Мечникова;
С. Л. Мирось – доцент, кандидат биологических наук, доцент кафедры генетики и молекулярной биологии Одесского национального университета имени И. И. Мечникова.

Васильева Н. Ю.

В 191 Биоинформатика. Множественное выравнивание. Филогенетические деревья : методическое пособие / Н. Ю. Васильева. – Одесса : «Одесский национальный университет имени И. И. Мечникова», 2014. – 70 с.

В методическом пособии рассмотрены основные принципы работы с биологическими последовательностями (множественное и парное выравнивание, построение филогенетических деревьев), которые предусмотрены программой курса «Биоинформатика». Большое внимание уделено on-line программам, которые доступны любому пользователю и могут быть хорошим подспорьем студентам биологам на начальных этапах изучения этого курса.

Методическое пособие снабжено большим количеством иллюстраций, позволяющим студентам самостоятельно разобраться с последовательностью действий. Методическое пособие предназначено для студентов биологов и биотехнологов.

.

УДК 575.22:004.42(075.8)
ББК 28.042в65я73

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Биоинформатика – это наука о хранении, извлечении, организации, анализе, интерпретации и использовании биологической информации.

Датой выделения биоинформатики в отдельную научную область можно считать 1980 год, когда началось издание журнала Nucleic Acids Research, целиком посвящённого компьютерным методам анализа последовательностей.

Основополагающий принцип биоинформатики состоит в том, что биополимеры, например, молекулы нуклеиновых кислот и белков, могут быть изображены в виде последовательности цифровых символов. Кроме того, для представления мономеров аминокислотных и нуклеотидных цепей необходимо лишь ограниченное число алфавитных знаков. Подобная гибкость анализа биомолекул с помощью биологических алфавитов привела к успешному становлению биоинформатики [5].

Триединая цель биоинформатики включает в себя:

1) организацию и сохранение биологических данных;
2) разработку программных средств и создание специализированных информационных ресурсов;
3) автоматизацию анализа биологических данных, интерпретацию и использование полученных результатов.

Предметом учебной дисциплин "Биоинформатика" являются компьютерно-ориентированные методы решения информационных задач в биологии и в частности микробиологии.

Для самостоятельной работы выделяется больше половины общего объёма времени, предназначенного для изучения данной дисциплины.

Самостоятельная работа проводится по всем темам, входящим в дисциплину. В процессе самостоятельной работы студент учится самостоятельно приобретать знания, которые затем используются в ходе

выполнения индивидуального задания, практических занятий, при подготовке к выполнению контрольных работ и к тестированию.

ВЫРАВНИВАНИЕ БИОЛОГИЧЕСКИХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

В процессе эволюции все биологические макромолекулы претерпевают множественные мутационные изменения. Это приводит к потере или приобретению протяженных фрагментов последовательностей или отдельных точечных мутаций. В общем справедливым будет высказывание, что если биомакромолекулы имеют общие последовательности мономеров, то они, как правило, обнаруживают подобие в структурах и в биологических функциях. Зачастую для таких биомакромолекул обнаруживается и общий предок. В данном случае, говорят, что если два белка или две нуклеотидные последовательности имеют большое сходство, то они являются **гомологами**, и, как правило, имеют общего предшественника, схожую функцию и похожие структуры [3, 5].

Сравнение двух предположительно гомологичных последовательностей показывает степень их расхождения, то есть силу эволюционных изменений.

Однако не следует забывать, что в биологии из гомологии чаще всего следует подобие функций, тогда как подобие функций может быть следствием как гомологии, так и аналогии.

Гомологичными белками называют белки, чьё происхождение от общего предка доказано. Если же пространственные структуры белков подобны, но первичные последовательности отличны, то такие структуры считаются аналогичными [2].

Для биоинформатики большее значение имеют именно гомологичные последовательности, имеющие общее происхождение (общего предка),

сходную 3D-структуру и в той или иной степени похожую аминокислотную последовательность.

Поэтому при доказательстве гомологичности нескольких последовательностей возникает задача установления **соответствия** друг другу отдельных протяженных участков последовательностей. В этом случае принято говорить о выравнивании последовательностей.

Выравниванием (alignment) последовательностей азотистых оснований в нуклеиновых кислотах или аминокислот в полипептидных цепях белков называют определение взаимного соответствия остатков (нуклеиновых оснований или аминокислотных остатков, соответственно) в двух или нескольких последовательностях, при котором сохраняется исходный порядок остатков в последовательностях [1].

Выравнивание последовательностей – это основной инструмент биоинформатики, его проводят с целью установления структурных, функциональных и эволюционных отношений между последовательностями.

Две последовательности можно «выровнять», написав их гомологичные остатки друг под другом.

При выравнивании двух последовательностей их помещают в две строки друг над другом, записывая их с помощью букв алфавита.

Выравнивание не должно изменять "смысл" последовательностей, поэтому при выравнивании должна сохраняться последовательность символов в строке и не должно быть перестановок символов.

В простейшем случае выравниваются две последовательности (парное выравнивание (pair sequence alignment)), в более сложных случаях выравнивается целый набор последовательностей (множественное выравнивание (multiple sequence alignment)). Как правило, множественное выравнивание осуществляется на основе результатов парного. Множественное выравнивание (на практике) зачастую конструируется повторным слиянием парных выравниваний для всех последовательностей. Последняя строка,

показывающая символы, сохраненные во всех последовательностях выравнивания, называется консенсусом [1, 5].

Результат выравнивания может быть убедительным или сомнительным. Если результат достоверен, то, скорее всего последовательности гомологичны, имеют подобные функции и общего предка.

Для того чтобы найти оптимальное (или наилучшее) выравнивание необходимо определить критерий качества выравнивания как лучший, поскольку в нём получено максимальное число совпадений для нуклеотидов в двух последовательностях и использовано минимальное число вставок.

Оптимальное выравнивание (optimal alignment) – это выравнивание нуклеотидных или аминокислотных последовательностей с самым высоким весом и имеющее биологический смысл. Вес выравнивания рассчитывается исходя из количества замен, с учетом разрывов и т.н. матрицы замен.

Чтобы решить, является ли оно лучшим из всех возможных, необходимо иметь способ систематической проверки всех возможных выравниваний, иметь количественный критерий ("вес" ("weight") или счёт ("score")), по которому возможно сравнивать качество различных выравниваний и определить выравнивание с оптимальным весом (счётом) [1, 5].

При этом от того, какая именно система оценки выбрана для такого сравнения, может зависеть результат сравнения, и даже незначительные изменения в схеме оценки могут изменить рейтинг выравниваний, из-за чего лучшим станет другое выравнивание.

ТИПЫ ВЫРАВНИВАНИЯ

При выполнении процедуры выравнивания идентичные или подобные "буквы" (элементы) этих строк (последовательностей) сдвигают в пределах строки (не меняя исходного порядка следования "знаков") таким образом, чтобы они выстраивались друг под другом в соответствующих столбцах.

Неидентичные, или различные знаки либо помещают в одни и те же столбцы как несовпадения, либо вставляют напротив них во второй последовательности пропуски.

Предположим, что есть две последовательности

Seq 1: ATGTCGTCAAGGTAATCCA
Seq2: ATGCGTCGGTAATGCT

Задача провести выравнивание этих последовательностей так, что бы оно было наиболее результативным.

Варианты написания последовательностей.

1 Неинформативное выравнивание:
```
ATGTCGTCAAGGTAATCCA----------------
------------------ATGCGTCGGTAATGCT---
```

2.Выравнивание без пропусков:
```
ATGTCGTCAAGGTAATCCA
ATGCGTCGGTAATGCT---
```

3. Выравнивание с пропусками:
```
ATGTCGTCAAGGTAATCCA
ATG-CGTC--GGTAATGCT
```

Цветом, выделены замены, которые так же имеют значение при определении веса выравнивания. Как видно, последнее выравнивание является лучшим, поскольку в нём получено максимальное число совпадений для нуклеотидов в двух последовательностях и использовано минимальное число вставок.

Закономерность таких записей так же зависит от типа выравнивания, который используется.

Типы выравнивания последовательностей:

1. Глобальное выравнивание (global alignment) – применимо когда две последовательности гомологичны по всей длине (рис. 1).

Рис.1. Схематическое представление глобального выравнивания.

В данном типе выравнивания символом "|" обозначены соответствия, "пробелы" обозначают несоответствия, "-" обозначает те вставки (инсерции, от англ. insertion) и удаления (делеции, от англ. deletion), которые необходимо сделать в обеих последовательностях, чтобы достичь максимального количества соответствий. Заметьте, что делеция в одной последовательности равносильна вставке в другой. Поэтому делеции и инсерции часто называют в таком типе выравнивания «инделами».

Пример

```
Прасковья. Пет--ровна .                    пела. песни. на. кухне
| |||||||||  |   |||||| ----------------------||||| |||||| |||||||||
Прасковья. Федоровна .пекла . куличи. и. пела .песни. на .кухне
```

2. Локальное выравнивание (local alignment) – применимо для сравнения последовательностей с частичной гомологией (рис. 2).

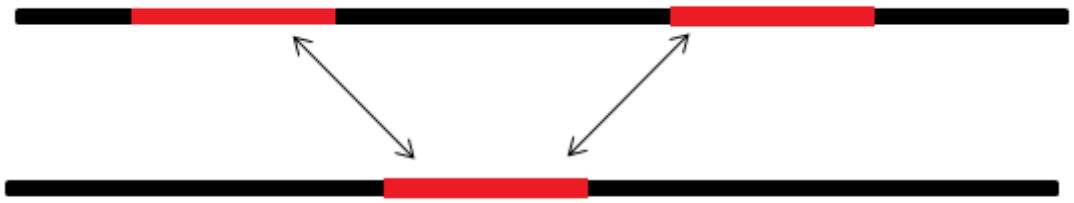


Рис. 2. Схематическое представление локального выравнивания.

Пример

```
                                  Петровна. пела. песни. на.  кухне
                                  ||||| |||||| ||| || |||
Прасковья. Федоровна .пекла . куличи. и. пела. песни. на.  кухне
```

Для локального совпадения выступающие концы не рассматриваются как пропуски (делеции). В дополнение к несовпадениям, возможны также вставки и удаления внутри совпадающей части.

3. Перекрывающееся выравнивание – предназначено для сравнения последовательностей у которых, совпадают только концевые участки (рис. 3). Используется для сборки последовательностей в ходе проектов секвенирования геномов.

Рис. 3. Схематическое представление перекрывающегося выравнивания.

Пример

и. пела .песни. на .кухне. Прасковья. Петровна

Прасковья. Федоровна .пекла . куличи. и. пела .песни. на .кухне.

4. Точечное выравнивание (dot plot) – применяется для общего исследования последовательности с целью обнаружения повторов и выбора фрагмента для множественного выравнивания (рис. 4).

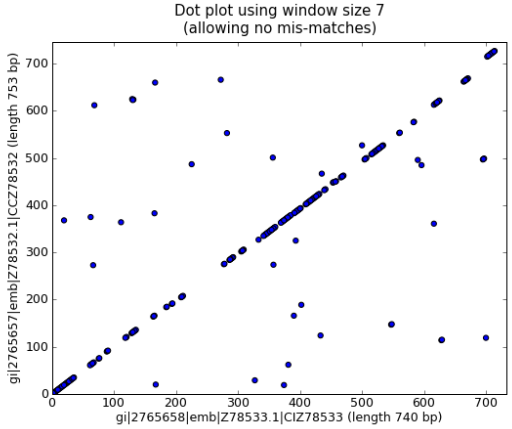


Рис. 4. Схематическое представление точечного выравнивания

Независимо от того проводим ли мы парное или множественное выравнивание, в каждом из вариантов можно использовать как локальное так и глобальное выравнивание (рис. 5).

Точечное и перекрывающееся выравнивания предусматривают использование только попарного сравнения последовательностей.

При выравнивании аминокислотных последовательностей используют специальные матрицы для расчета веса (score) всего выравнивания. Для этого определяют частный вес каждой пары замен при выравнивании аминокислотной последовательностей.

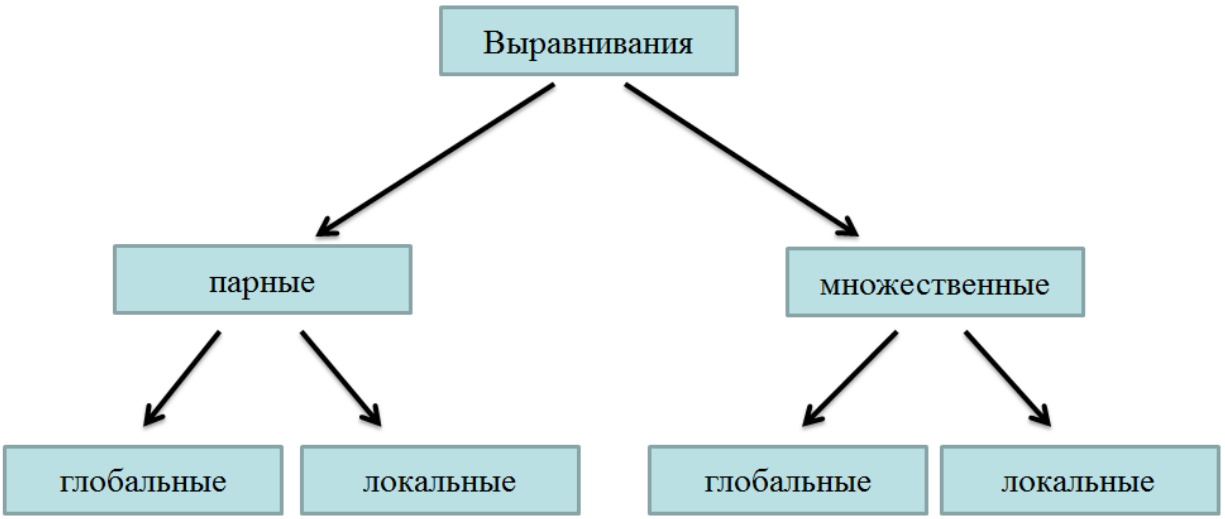


Рис. 5. Типы выравнивания.

МАТРИЦЫ ЗАМЕН

Аминокислоты с близкими биохимическими свойствами, такими как заряд, полярность и т.д. характеризуются большей вероятностью парных замен. Некоторые аминокислоты, например цистеин, глицин, триптофан очень редко заменяются в процессе эволюции. Для того чтобы учесть неравную вероятность замен были разработаны специальные матрицы, которые получили название *матрицы замен*. Эти матрицы содержат оценки частных весов для любой пары замены аминокислоты (или нуклеотида) *i* на аминокислоту (или нуклеотид) *j*. Первыми матрицами были матрицы аминокислотных замен PAM (табл. 1) [1, 2, 3].

Для их создания были использованы эволюционно близкие последовательности различных белков, таких как гемоглобин, цитохром с, фибриноген и т. д. Для оценки весов использовались средние значения частот, вычисленные на большом наборе данных. По этим данным была построена эмпирическая матрица нормированных весов аминокислотных замен.

Вес $S(i, j)$ в ячейке i, j таблицы 1 больше нуля означает, что аминокислота i заменяется на j чаще, чем в среднем по всем заменам. То есть эти аминокислоты сравнительно легко заменяют друг друга, т.к. они

функционально эквивалентны или по другим причинам. Вес меньше нуля указывает на пары аминокислот, которые сравнительно редко заменяют друг друга [2].

Матрицы PAM различаются по числовым индексам. Например, матрице PAM250, соответствует примерно 20% идентичности последовательностей, что считается минимальным уровнем сходства, для которого можно надеяться получить правильное выравнивание, основываясь на анализе самих

Таблица 1

Матрица аминокислотных замен PAM250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	-5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

последовательностей без привлечения дополнительной информации, например, пространственной организации белковой глобулы. Расстояние 250 PAM означает, что при эволюции последовательности длиной 100 аминокислотных остатков произошло 250 мутаций в случайных позициях. Поэтому в некоторых позициях мутаций вообще не было, а в некоторых позициях произошло 3 и более мутационных изменения.

Недостатком матриц PAM является то, что они не очень надежно работают на больших эволюционных расстояниях [3, 5].

Таблица 2

Матрица замен BLOSUM50.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5																			
R	-2	7																		
N	-1	-1	7																	
D	-2	-2	2	8																
C	-1	-4	-2	-4	13															
Q	-1	1	0	0	-3	7														
E	-1	0	0	2	-3	2	6													
G	0	-3	0	-1	-3	-2	-3	8												
H	-2	0	1	-1	-3	1	0	-2	10											
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5										
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5									
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6								
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7							
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8						
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10					
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5			
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15		
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	-1	1	-3	-2	0	-3	1	5

Другим широко используемым семейством матриц весов являются матрицы BLOSUM (табл. 2), предложенные в 1992 г. Они построены на основе выравниваний последовательностей с определенной степенью сходства. В матрицах BLOSUM значение веса S (i, j) для каждой ячейки i, j получено из наблюдений частот замен в частичных выравниваниях близких белков. Каждая матрица соответствует специфическому порогу сходства. Например, при построении матрицы BLOSUM62 были использованы последовательности, имеющие более чем 62% сходства [3, 4, 5].

Матрицы с меньшими пороговыми значениями соответствуют большим временам раздельной эволюции. Поэтому их используют для выравнивания более удаленных друг от друга последовательностей.

Основными отличиями матриц PAM и Blosum являются:

1) использование матрицами PAM простой эволюционной модели (подсчет замен на ветвях *филогенетического древа*);

2) матрицы PAM основаны на учете мутаций по принципу глобального выравнивания (в высококонсервативных и высокомутабельных участках), а матрицы Blosum – локального (только высококонсервативных участков).

При средней степени сходства последовательностей наиболее часто используются матрицы Blosum62 и PAM160. При выравнивании близкородственных последовательностей следует использовать матрицы Blosum с большим порядковым номером и матрицы PAM с меньшим номером.

Матрицы этих двух серий сопоставимы следующим образом PAM 100 – Blosum 90, PAM 120 – Blosum 80, PAM 160 – Blosum 60, PAM 200 – Blosum 52, PAM 250 – Blosum 45. Наиболее часто используются матрицы Blosum 62 и PAM 160 (при среднем сходстве последовательностей).

Так же используются матрицы Gonnet, представляющие собой усовершенствованный вариант матриц Дэйхофф, основанный на большей базе данных.

Зачем необходимо выравнивание?

В первую очередь, как уже отмечали, для подтверждения гомологичости последовательностей. Во-вторых, если открыта новая последовательность с неизвестной функцией, но при этом в базах данных могут быть найдены подобные ей последовательности с ранее установленными структурами и функциями, то результаты выравнивания (сравнения) этой новой последовательности с уже исследованными последовательностями могут стать основанием для предсказания функции или структуры этой новой последовательности.

МНОЖЕСТВЕННОЕ ВЫРАВНИВАНИЕ

Множественное выравнивание (multiple sequence alignment) – это выравнивание набора из трех и более последовательностей одновременно, при котором элементы в одинаковых позициях группируются в колонки.

Какой биологический смысл должно нести множественное выравнивание? С одной стороны, это эволюционная значимость. Правильное выравнивание должно отражать происхождение данных последовательностей из единой предковой последовательности. Если набор последовательностей не имеет единого предка, то и осмысленного выравнивания этих последовательностей не существует. Однако, в этом случае можно обнаруживать участки локального сходства анализируемых макромолекул. Консервативность этих участков свидетельствует об их функциональной важности – они могут являться элементами вторичной структуры, сайтами связывания лигандов, другими функциональными мотивами [2, 3].

С другой стороны, выравнивание последовательностей белков отражает сходство пространственных структур белков. Аминокислотные остатки, стоящие в одном столбце выравнивания, должны занимать довольно близкое пространственное положение. Множественное выравнивание последовательностей использует больше информации, чем парное, поэтому

(теоретически) должно в среднем чаще получаться более биологически осмысленным.

Есть несколько негласных правил при использовании множественного выравнивания.

- Выравнивайте белки, а не ДНК, если есть выбор
- Лучше брать не более 15 последовательностей.

В выборке лучше избегать:

- слишком похожих последовательностей (>90% identically)
- слишком разных последовательностей (<30% identically)
- неполных последовательностей (фрагментов)
- тандемных повторов

Основная **цель** множественного выравнивания – это выявление доменов, содержащихся в изучаемой последовательности. Множественное выравнивание может быть как полным, так и частичным.

Реально не существует полных выравниваний. Даже очень хорошая выборка белков не может быть выровнена по всей длине последовательностей. Значит, мы можем говорить лишь о частичных выравниваниях.

Разработаны различные алгоритмические подходы для построения множественного выравнивания. Наиболее часто используется так называемое *прогрессивное выравнивание*, которое включает следующие этапы:

1) Построить парные выравнивания

2) Построить дерево-подсказку

3) Провести прогрессивное выравнивание по дереву-подсказке

При использовании этого подхода сначала выбираются две наиболее похожие последовательности, которые выравниваются стандартным алгоритмом парного выравнивания. Это выравнивание фиксируется. Далее выбирается третья последовательность, которая «подравнивается» к первому выравниванию, затем 4–я и т.д. до тех пор, пока не будут выровнены все последовательности. При использовании подобного подхода, выравнивание строится в порядке убывания сходства последовательностей (рис. 6).

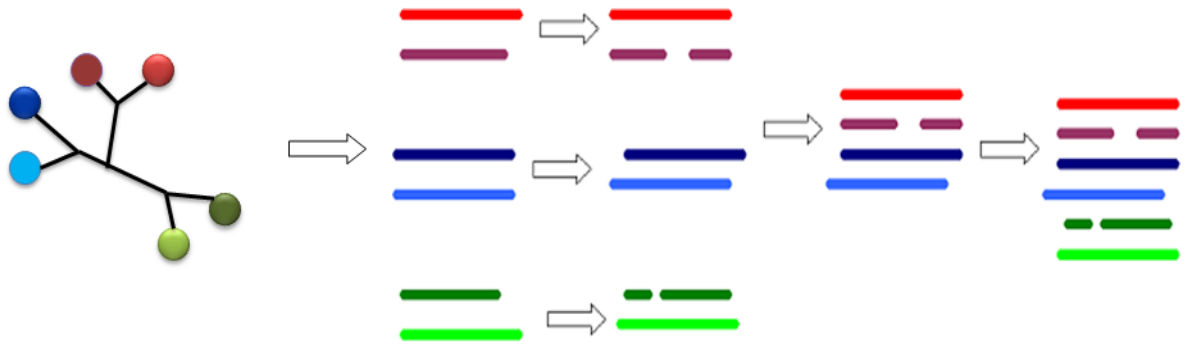


Рис. 6. Схема алгоритма прогрессивного выравнивания.

Пример действия такого алгоритма – результат множественного выравнивания пяти нуклеотидных последовательностей приведен на рисунке 7, а аминокислотных последовательностей на рисунке 8.

G	–	–	C	A	A	C	C	C	A	G
G	C	C	C	T	A	A	C	A	A	G
G	G	T	A	G	A	–	C	A	A	G
G	C	A	C	–	–	A	C	–	A	G
C	C	C	A	G	C	C	C	C	A	G

Рис. 7. Результат множественного выравнивания пяти нуклеотидных последовательностей.

```
FOS_RAT          PEEMSVTS-LDLTGGLPEATTPESEEAFTLPLLNDPEPK-PSLEPVKNISNMELKAEPFD
FOS_MOUSE        PEEMSVAS-LDLTGGLPEASTPESEEAFTLPLLNDPEPK-PSLEPVKSISNVELKAEPFD
FOS_CHICK        SEELAAATALDLG----APSPAAAEEAFALPLMTEAPPAVPPKEPSG--SGLELKAEPFD
FOSB_MOUSE       PGPGPLAEVRDLPG-----STSAKEDGFGWLLPPPPPPP----------------LPFQ
FOSB_HUMAN       PGPGPLAEVRDLPG-----SAPAKEDGFSWLLPPPPPPP----------------LPFQ
                 .   . :   ** .     :..  *:.*   *   . *                  **:
```

Рис. 8. Результат множественного выравнивания пяти аминокислотных последовательностей.

На сегодняшний день множественное выравнивание последовательностей осуществляется несколькими программами доступными в режиме on-line. Одной из наиболее широко используемых реализаций алгоритма прогрессивного множественного выравнивания является программа ClustalW. Это третье поколение программ этой серии, появившейся в 1994 году. Данная версия значительно проще в работе благодаря усовершенствованному алгоритму, основанному на создании множественного выравнивания в результате серий попарных выравниваний, следуя ветвлению направляющего

дерева, построенного методом UPGMA. Кроме этого появилась возможность выбирать матрицы сравнения аминокислот и нуклеотидов, а также устанавливать штрафы за внесение пробелов. Следует отметить, предоставление результатов выравнивания в виде формата FASTA, обеспечивает высокую совместимость программ этого поколения с другими пакетами программ.

Последние программы серии Clustal позволяет создавать наиболее биологически корректные множественные выравнивания биологических последовательностей

Программа доступна на многих серверах (http://npsa-pbil.ibcp.fr, http://www.ebi.ac.uk/services) в двух вариантах – интерактивном и почтовом. Интерактивный вариант предполагает ожидание пользователем получения результатов выравнивания (целесообразно применять при небольшом (<100) количестве последовательностей), а почтовый – по электронной почте (применяется при большом числе последовательностей).

Основным предназначением программы ClustalW является построение множественного выравнивания, вычисление эволюционных дистанций между последовательностями, определение характера и типа аминокислотных замен и т. д.

ЭТАПЫ ВЫПОЛНЕНИЯ МНОЖЕСТВЕННОГО ВЫРАВНИВАНИЯ В ПРОГРАММЕ ClustalW

1. Первоначально необходимо создать файл с последовательностями (нуклеотидными или аминокислотными), которые мы хотим проанализировать. Можно использовать 7 возможных форматов (NBRF/PIR, EMBL/SWISSPROT, Pearson (Fasta), Clustal (*.aln), GCG/MSF (Pileup), GCG9/RSF, GDE). Наиболее часто используется формат FASTA.

В биоинформатике, формат FASTA является текстовым форматом файла для сохранения нуклеотидных или аминокислотных последовательностей, в котором нуклеотиды или аминокислоты передаются с помощью букв. Этот формат также позволяет передавать описание этих последовательностей и

краткий комментарий к ним. Название формата происходит от программного пакета FASTA, но этот формат уже стал независимым стандартом в биоинформатике. Последовательность в этом формате начинается с названия, перед которым ставят символ “>”. Первое слово после этого символа обычно является идентификатором последовательности, таким как номер последовательности в базе данных GenBank. Остальные слова в первой строке могут передавать любую информацию о последовательности. Все слова в первой строке необязательны и могут быть в свободном формате. Однако идентификатор должен следить непосредственно за символом '>', то есть между ">" и идентификатором не должно быть пробелов. Формат рекомендует ограничивать длину строк до 80 символов. Обычно строки последовательности имеют длину в 60 символов. Затем с новой строки вводят саму последовательность.

В FASTA формате используются однобуквенные коды для нуклеотидов и аминокислот, заданные Международным Объединением Биохимии и Международным Объединением Чистой и Прикладной Химии (IUB/IUPAC). Строки могут иметь разную длину – это граница с "рваным" правым краем.

Пример Fasta формата нуклеотидной последовательности

>gi|86197837|emb|AM179887.1| Bacillus sp. C81 partial 16S rRNA gene, isolate C81
TTGCTTCTTCTGATTAGCGGCGGACGGGTGAGTAACACGTGGGCAACCTGCCCTGTAGATTGGGATAACT
CCGGGAAACCGGGGCTAATACCGAATAATCCATTTCTTCACATGAGGAGATGTTAAAAGACGGTTTCGGC
TGTCACTACAGGATGGGCCCGCGGCGCATTAGCTAGTTGGTGAGGTAATGGCTCACCAAGGCGACGATGC
GTAGCCGACCTGAGAGGGTGATCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGC
AGTAGGGAATCTTCCACAATGGACGAAAGTCTGATGGAGCAACGCCGCGTGAGTGAAGAAGGTTTTCGGA
TCGTAAAACTCTGTTGTGAGGGAAGAACAAGTACGAGAGTAACTGCTCGTACCTTGACGGTACCTCATTA
GAAAGCCACGGCTTACTACCTGCCAGCAGCCGCGGTAATACCTAGGTGGCAAGCTGTTGTCCGGAATTAT
TGGGCGTAAAGCGCGCGCAGGCGGTCCTTTAAGTCTGATGTGAAAGCCCACGGCTCAACCGTGGAAGGTC
ATTGGAAACTGGGGGACTTGAGTGCAGAAGAGGAAAGTGGAATTCCAAGTGTAGCGGTGAAATGCGTAGA
GATTTGGAGGAACACCAGTGGCGAAGGCGACTTTCTGGTCTGTAACTGACGCTGAGGCGCGAAAGCGTGG
GGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGTGCTAAGTGTTAGGGGGTTTC
CGCCCCTTAGTGCTGCAGCTAACGCATTAAGCACTCCGCCTGGGGAGTACGGTCGCAAGACTGAAACTCA
AAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAACCTTA
CCAGGTCTTGACATCCCGCTGACCGCTCTAGAGATAGAGTTTTCCCTTCGGGGACAGCGGTGACAGGTGG
TGCATGGTTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGATCT

TAGTTGCCAGCATTCAGTTGGGCACTCTAAGGTGACTGCCGGTGATAAACCGGAGGAAGGTGGGGATGAC
GTCAAATCATCATGCCCCTTATGACCTGGGCTACACACGTGCTACAATGGACGGTACAGAGGGTCGCAAC
CCCGCGAGGGTGAGCTAATCCCATAAAACCGTTCTCAGTTCGGATTGTAGGCTGCAACTCGCCTACATGA
AGCCGGAATCGCTAGTAATCGTGGATCAGCATGCCACGGTGAATACGTTCCCGGGCCTTGTACACACCGC
CCGTCACACCACGAGAGTTTGTAACACCCGAAGTCGGTGGGGTAACCTTACGGGAGCCAGCCGCCGAAGG

Пример Fasta формата аминокислотной последовательности

>gi|228699694|gb|EEL52352.1| Pyridoxine kinase [Bacillus cereus Rock3-44]
MEVIMKKVAVIQDLSSFGKCSLTAAIPVLSVMGVQACPLPTAILSSQTGYPSFFCEDFTSKMKYFEEEWS
KLHVTFDGIYTGFVTGREQIDNIFRFLDTFHTKETILLVDPVMGDIGEAYKLFTEELLVRMRELVKCADV
ITPNVTECCLLTGLSYEKLYSYVNEIDFIKALEEAGKTLQQETDAKVIITGVNPPSANRDKQFIGNMYLD
GNKNFYDQTPYNGKSYSGTGDLFASVIMGSMMRGEDLEKSVQLAEAFLTASIHDTSLEQIPEVEGVNFEK
YLRMLL

Совет: поскольку чаще всего приходится параллельно анализировать нуклеотидные и аминокислотные последовательности одного гена, то лучше использовать ресурсы курируемых баз данных, например KEGG осуществив предварительный поиск и бластование в базе uniProt.

2. Осуществляем вход на страницу браузера веб сервера EMBL-EBI - http://www.ebi.ac.uk/services (рис. 9)

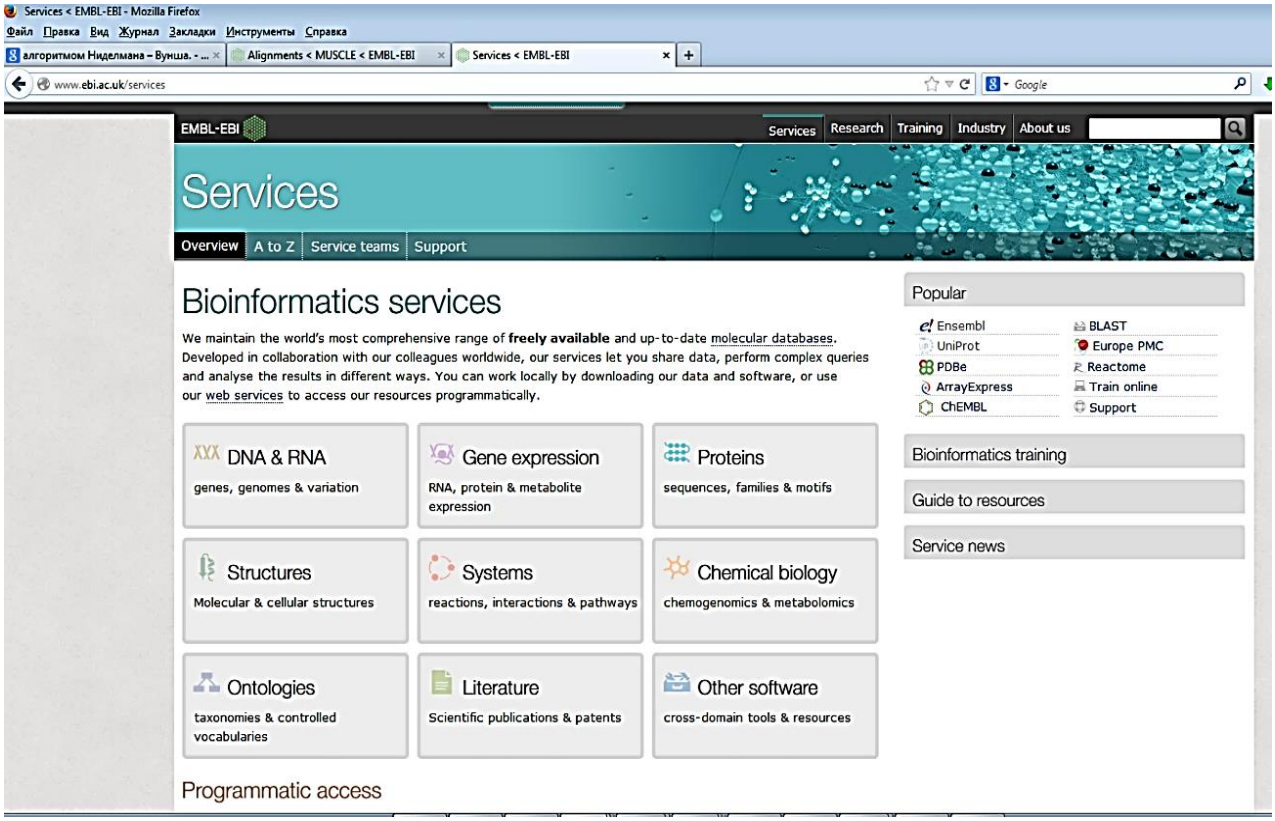


Рис. 9. Страница браузера веб сервера EMBL-EBI.

3. Выбираем блок DNA&RNA (genes, genomes & variation) и переходим на страницу, содержащую программы этого блока (рис. 10).

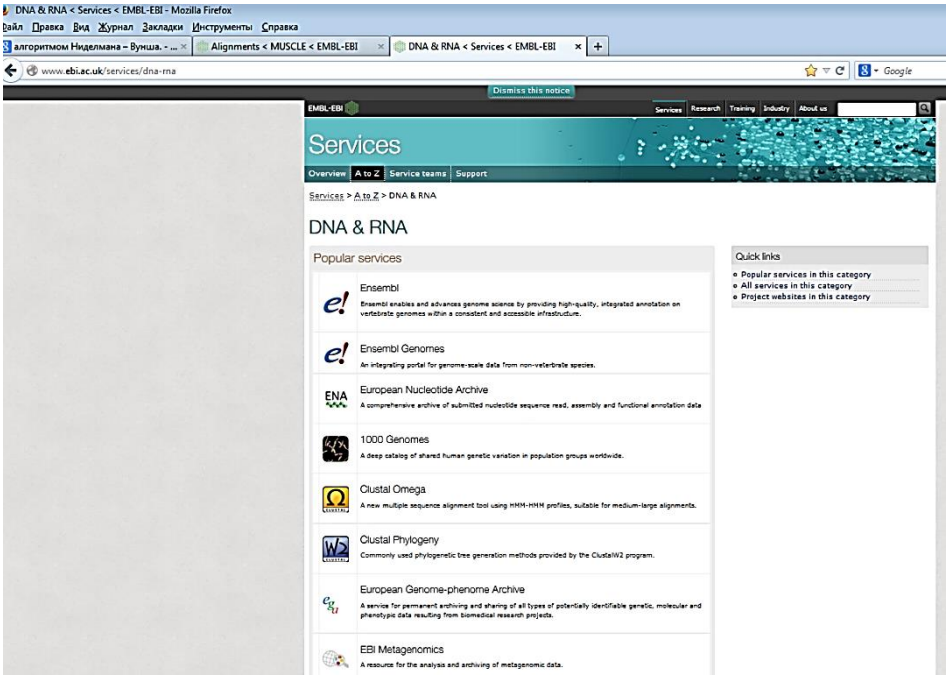


Рис. 10. Страница браузера веб сервера EMBL-EBI с имеющимися программами блока DNA&RNA (genes, genomes & variation).

4. Выбираем ClustalW2 и переходим на страницу с окном программы (рис. 11).

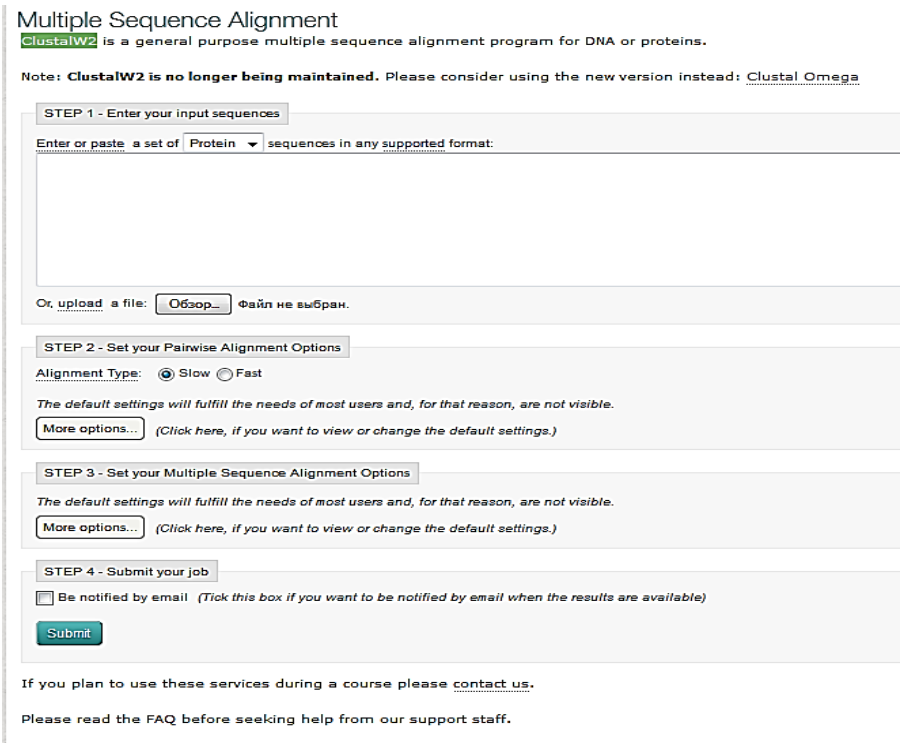


Рис. 11. Окно программы ClustalW2.

5. Перед тем как вставить в окно свой набор последовательностей убедитесь, что у вас стоят верные опции. Для аминокислотных последовательностей – Protein, а для нуклеотидных – DNA. Это необходимо помнить, поскольку выравнивание производится на основании матриц сравнения:

- матрица сравнений нуклеотидов (DNA weight matrix, IUB, Clustal W). В наиболее широко используемой матрице DNA identity совпадение нуклеотидов оценивается в 1 балл, а несовпадение -10000 баллов. Такой высокий штраф за несоответствие облегчает внесение пробелов (табл. 3).

Таблица 3

Матрица DNA identity.

	A	T	G	C
A	1	-10000	-10000	-10000
T	-10000	1	-10000	-10000
G	-10000	-10000	1	-1000
C	-10000	-10000	-10000	1

В последней версии программы ClustalW при выравнивании последовательностей ДНК рекомендует использовать значения "+1" для совпадения, "0" для несовпадения и штрафы d =10 за введение делеции и e = 0,1 за продолжение делеции.

По умолчанию в опциях программы ClustalW стоят матрица IUB для нуклеотидных последовательностей и матрица Gonnet для аминокислотных.

Матрицы сравнения для аминокислотных последовательностей описывали выше.

6. После вставки совокупности нуклеотидных последовательностей в окно, следует выбрать разновидность выравнивания медленное (slow) или быстрое (fast) (показано стрелками на рис. 12). Опция *Alignment* – выбор алгоритма выравнивания.

Медленное выравнивание является более точным, но его не рекомендуется применять в случае большого количества (более 20)

последовательностей значительной длины (более 1000 остатков). Медленное выравнивание характеризуется следующими параметрами:

- *Gap Open Penalty*: штраф на внесение делеции в выравнивание. Смысл этого параметра в следующем. Уменьшение его делает возможным более легко вносить в выравнивание разрывы, при этом качество выравнивания ухудшается. Если этот параметр увеличивать – выравнивание будет представлять собой длинные участки последовательностей почти без вставок или делеций.
- *Gap extension penalty*: штраф на продолжение делеции. Этот параметр контролирует возможность внесения длинных вставок или делеций.
- *Protein weight matrix:* матрица сравнения аминокислот.
- *DNA weight matrix:* матрица сравнения нуклеотидов (рис. 12).

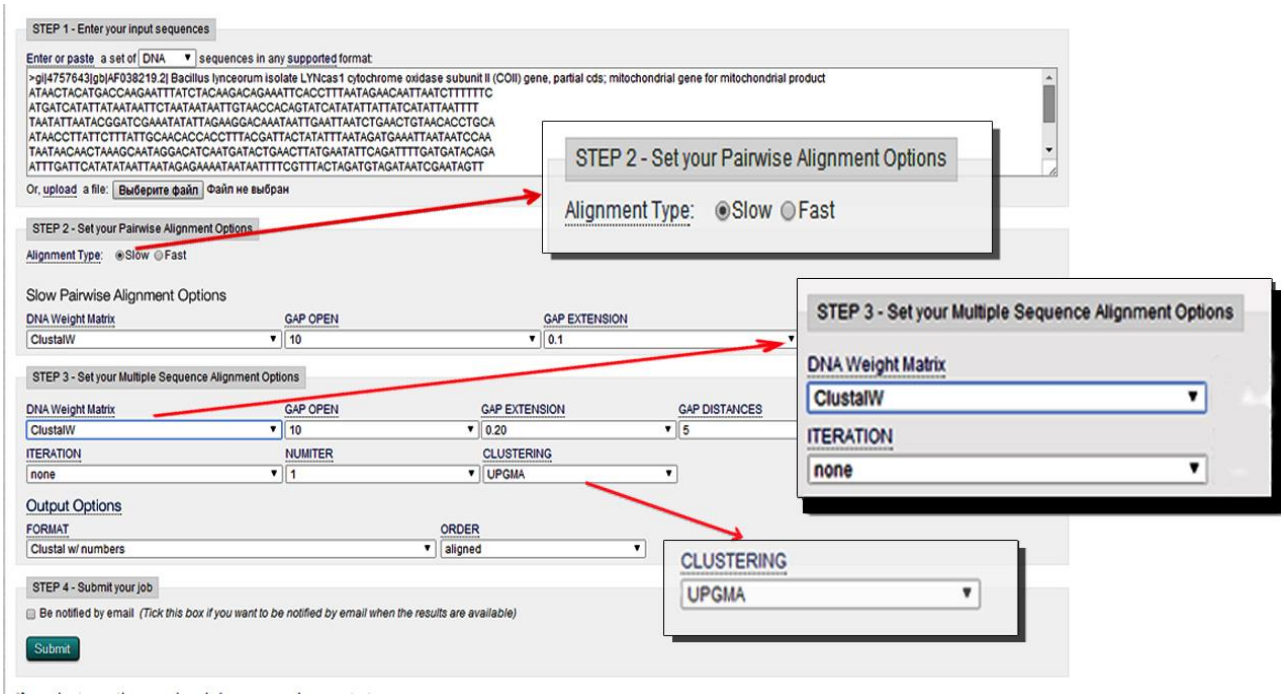


Рис. 12. Окно программы ClustalW с установленными опциями для медленного выравнивания.

Быстрое но менее точное выравнивание (последовательности выравниваются с помощью поиска длинных сходных участков «к-плетов», затем эти наиболее сходные участки образуют «блоки» выравнивания):

- *k-tuple size*: Размер участка максимального совпадения (по умолчанию = 1). Для ***увеличения скорости*** надо ***увеличивать*** этот параметр (max= 2 для белков; 4 для ДНК). Для ***увеличения точности*** надо ***уменьшать*** этот параметр.
- *Gap Penalty*: штраф на введение делеции. Практически не влияет на скорость.
- *Top Diagonals:* число непрерывно совпадающих к-плетов на участке парного выравнивания (если к=1, то это просто длина совпадающего сегмента). Для построения выравнивания выбираются только сегменты, превышающие этот порог. Для **увеличения скорости** надо ***уменьшать*** этот параметр, для **увеличения точности** надо ***увеличивать*** этот параметр.
- *Window Size:* длина сегмента, включающего «наилучший выровненный сегмент (см. предыдущий параметр). Для **увеличения скорости** надо ***уменьшать*** этот параметр, для **увеличения точности** надо ***увеличивать*** этот параметр (рис. 13).

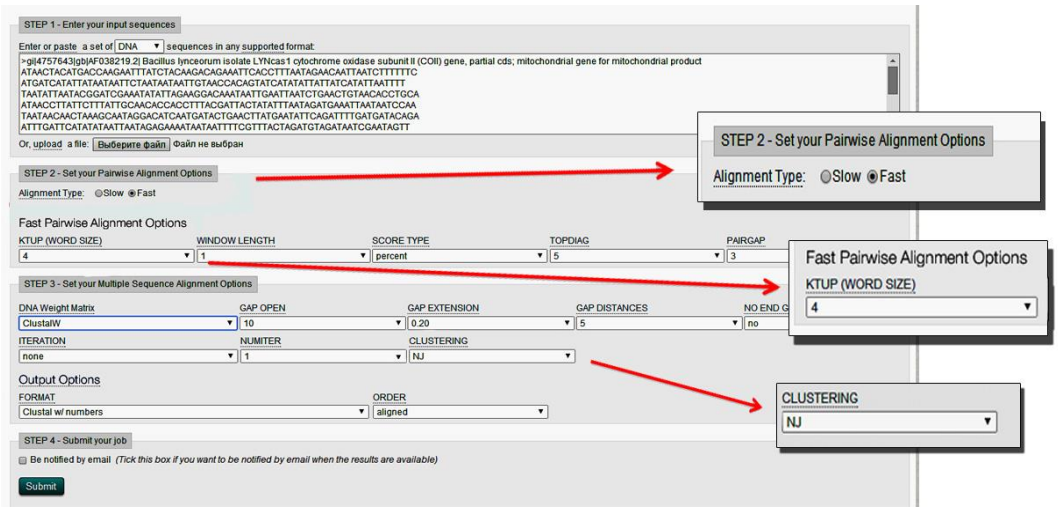


Рис. 13. Окно программы ClustalW с установленными опциями для быстрого выравнивания.

7. Следующим этапом устанавливаем опции собственно для множественного выравнивания.

- *DNA weight Matrix* - выбор матрицы замен, для построения выравнивания;

- *Gap Open* - штраф за начало разрыва;
- *End Gaps* - штраф за окончание разрыва;
- *Gap Extension* - штраф за длину разрыва.
- *CLUSTERING:* алгоритм расчета –NJ (метод связывания ближайших соседей (neighbour - joining или NJ)) или UPGMA (метод невзвешенного попарного среднего – Unweighted Pair-Group Method Using Arithmetic Averages). **Совет**: отнеситесь к данному пункту достаточно внимательно, поскольку, от того какую разновидность выравнивания и какой алгоритм расчета вы выберете, будет зависеть результат. На рисунке 14 приведено сравнение результатов медленного выравнивания нуклеотидных последовательностей, но с разными алгоритмами. На рисунке 15 – результаты по одному алгоритму, но для разных разновидностей выравнивания.

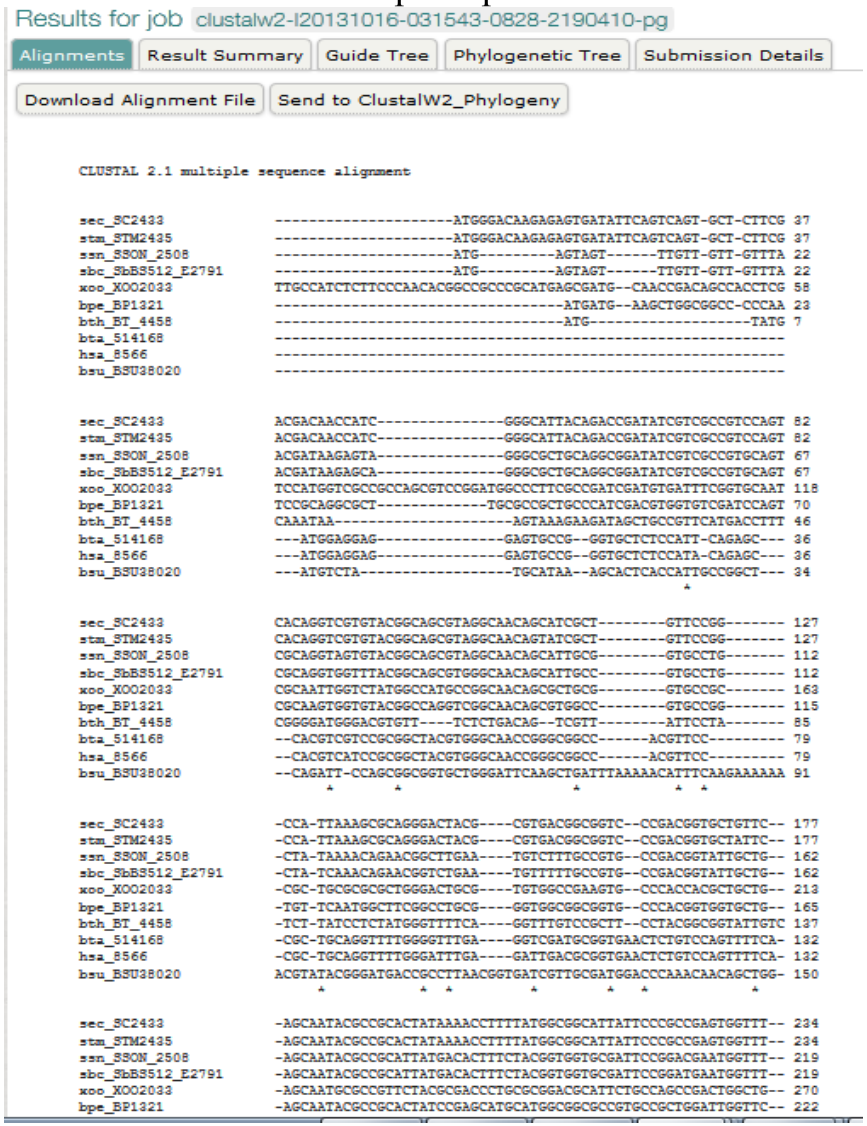
Большинство остальных опций выставлено по умолчанию и не требуют корректировки.

8. После того как все опции установлены нажать Submit.

ПРИМЕР МНОЖЕСТВЕННОГО ВЫРАВНИВАНИЯ

Рассмотрим пример множественного выравнивания нуклеотидных последовательностей. Для примера выбрали последовательности гена pyridoxine kinase следующих организмов: *Salmonella choleraesuis* (strain SC-B67); *Salmonella typhimurium* (strain LT2 / SGSC1412 / ATCC 700720); *Shigella sonnei* (strain Ss046); *Shigella boydii* serotype 18; *Xanthomonas oryzae* pv. oryzae (strain KACC10331 / KXO85); *Bacillus subtilis* (strain 168); *Bordetella pertussis* (strain Tohama I / ATCC BAA-589 / NCTC 13251); *Bacteroides thetaiotaomicron* (strain ATCC 29148 / DSM 2079 / NCTC 10582 / E50 / VPI-5482); *Homo sapiens*. Отбор аминокислотных последовательностей выполняли в базе данных UniProt (http://www.uniprot.org/). Соответствующие нуклеотидные последовательности были получены из базы данных KEGG (http://www.genome.jp/kegg/kegg2.html).

алгоритм расчета NJ



алгоритм расчета UPGMA

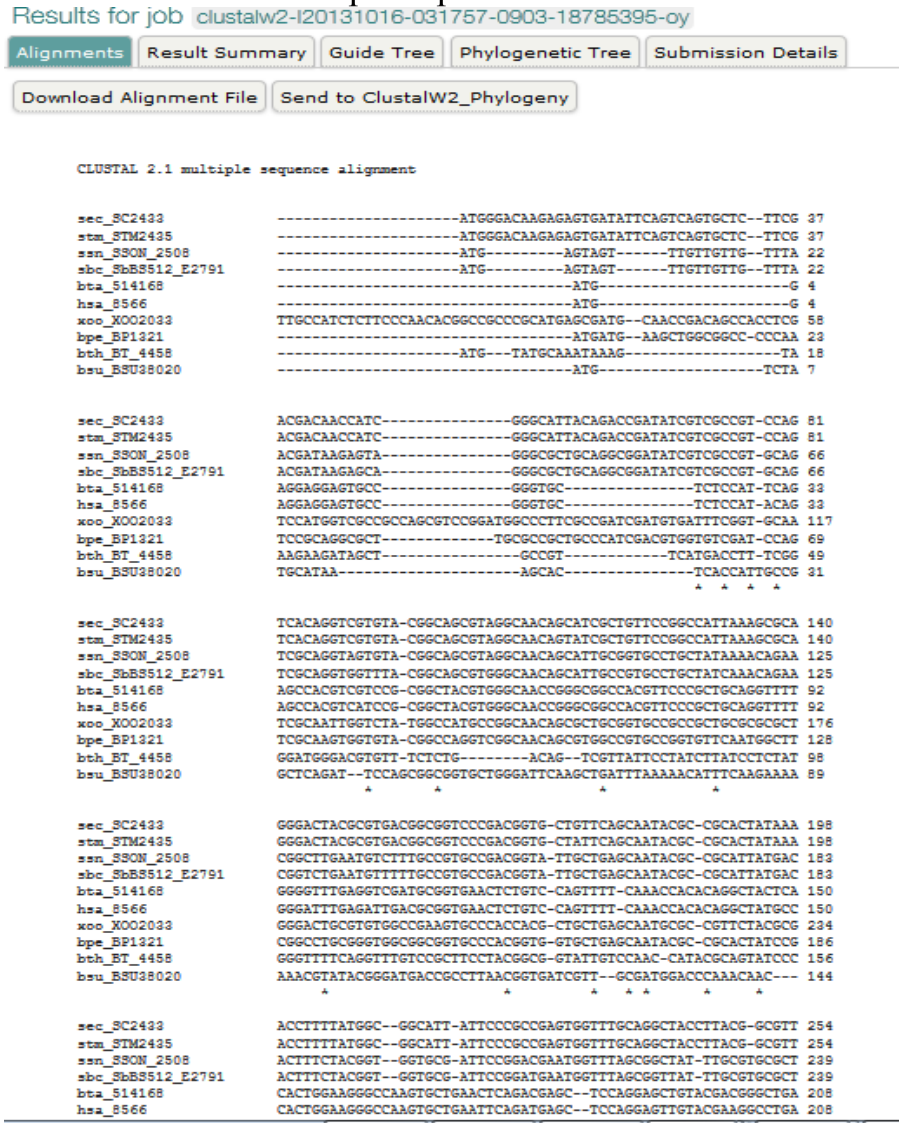
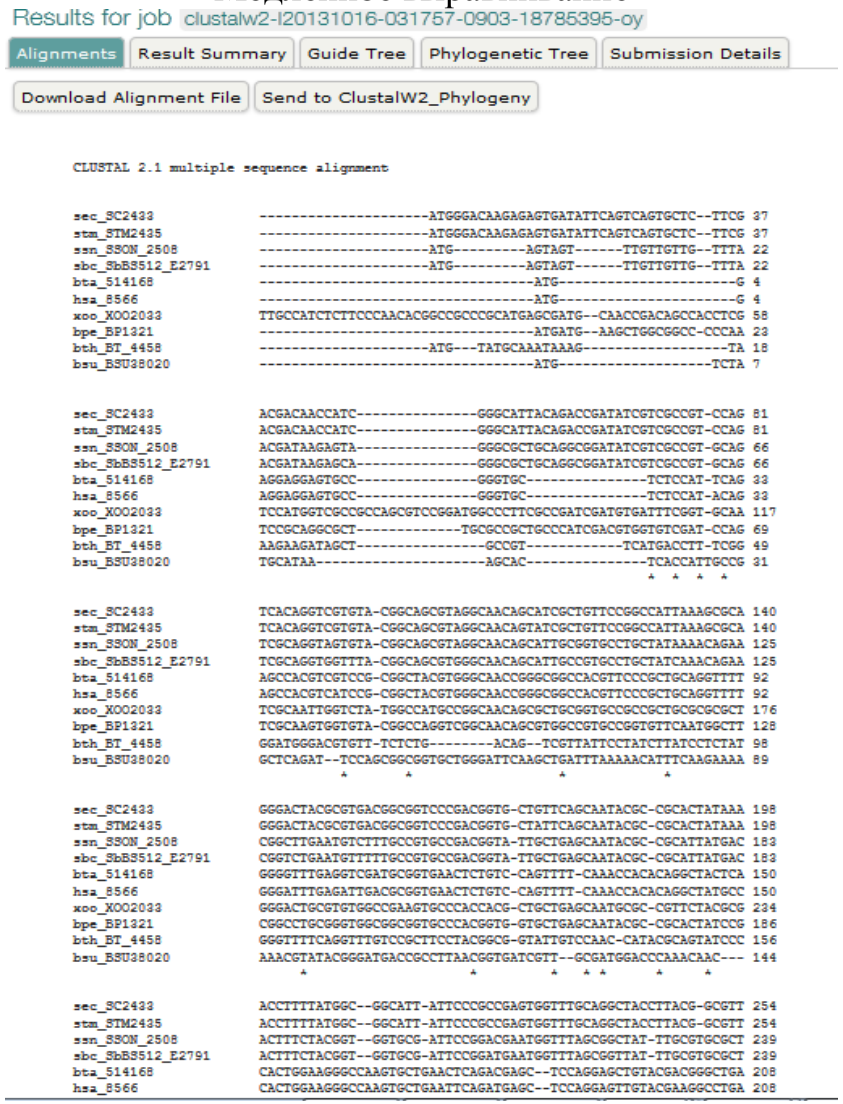


Рис. 14 Сравнение результатов множественного медленного выравнивания с использованием алгоритм расчета NJ и UPGMA.

Медленное выравнивание



Быстрое выравнивание

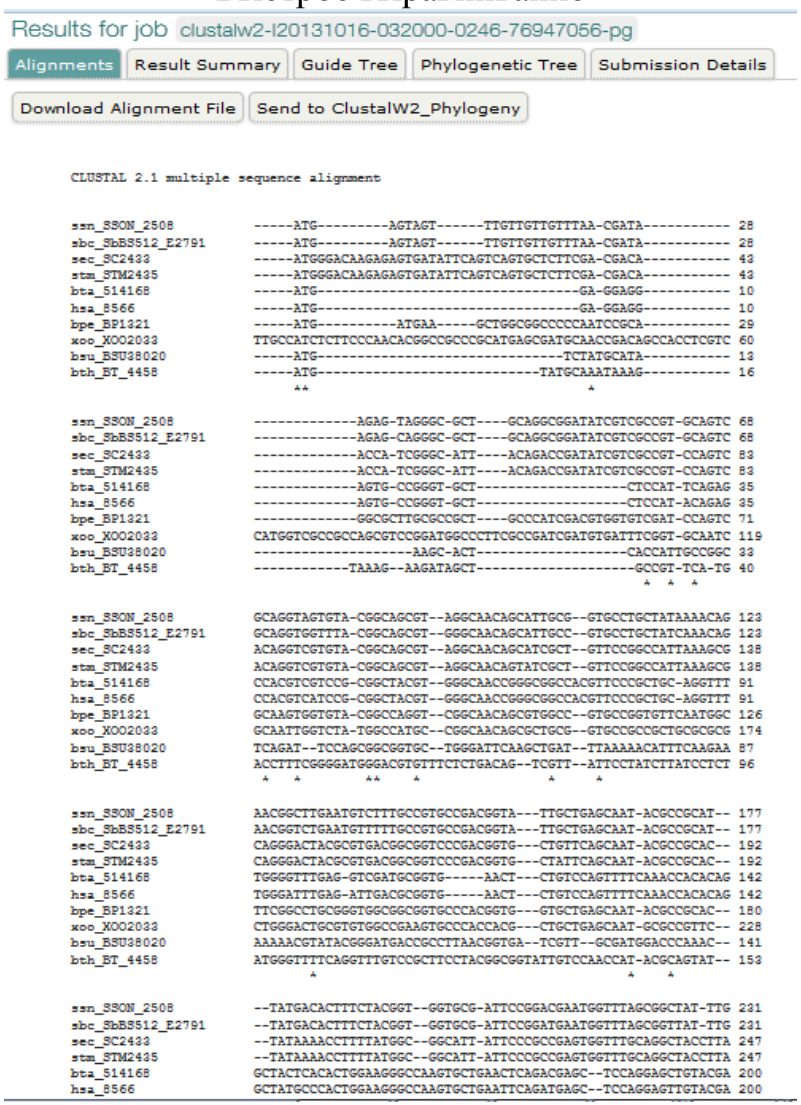


Рис. 15 Сравнение результатов множественного медленного и быстрого выравнивания с использованием алгоритма UPGMA.

Для множества последовательностей сформировали общий файл. Очередность последовательностей в файле, соответствовала очередности в тексте.

1.Выбираем опцию - DNA.

2.В нижележащих опциях установили:

Alignment – slow;

DNA weight matrix – IUB;

Gap Open – 10;

Gap extension – 0.1;

CLUSTERING – NJ.

Остальные установки оставили по умолчанию (рис 16). Ввели последовательности.

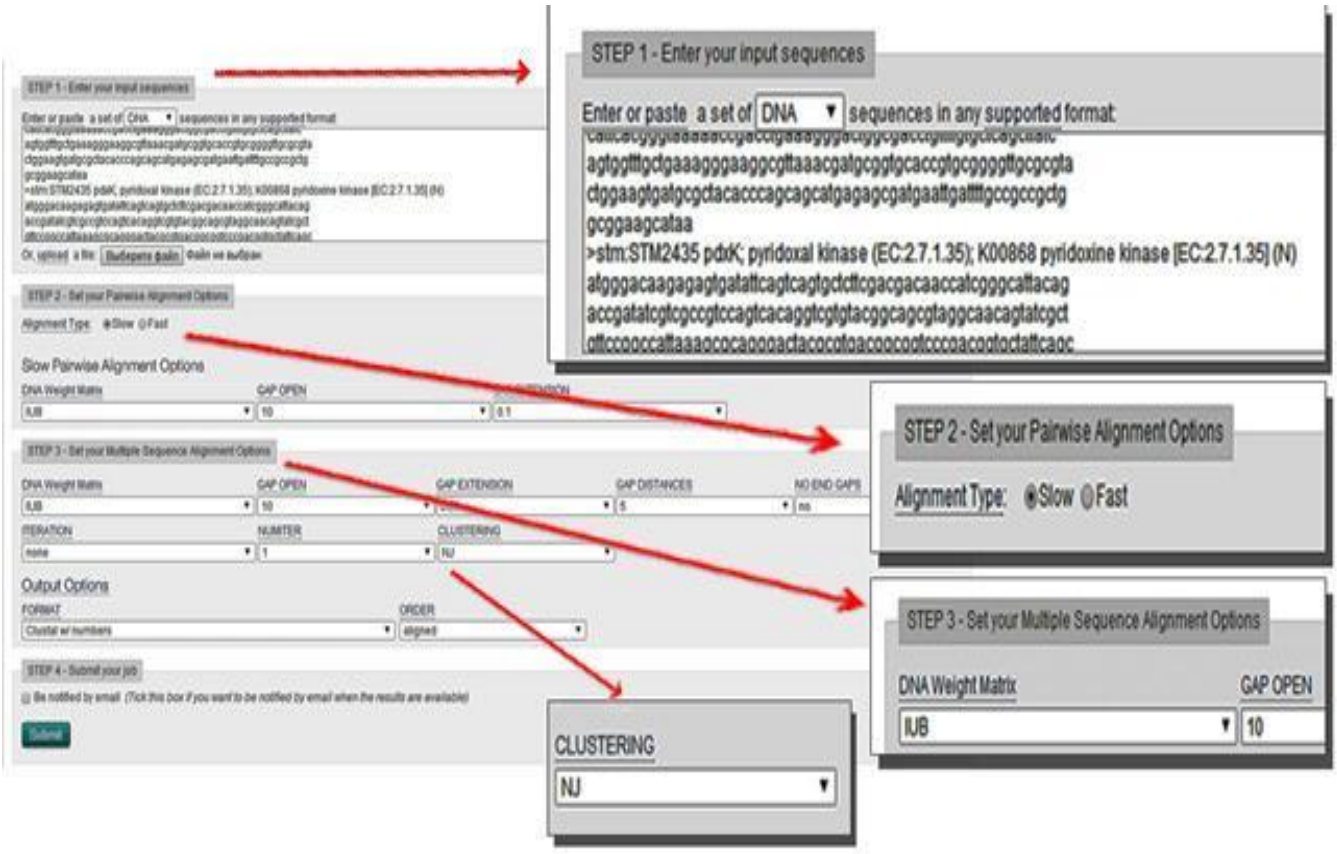


Рис. 16. Окно программы ClustalW с выставленными опциями для выравнивания нуклеотидных последовательностей.

Нажимаем Submit. Результат представлен на рисунке 17.

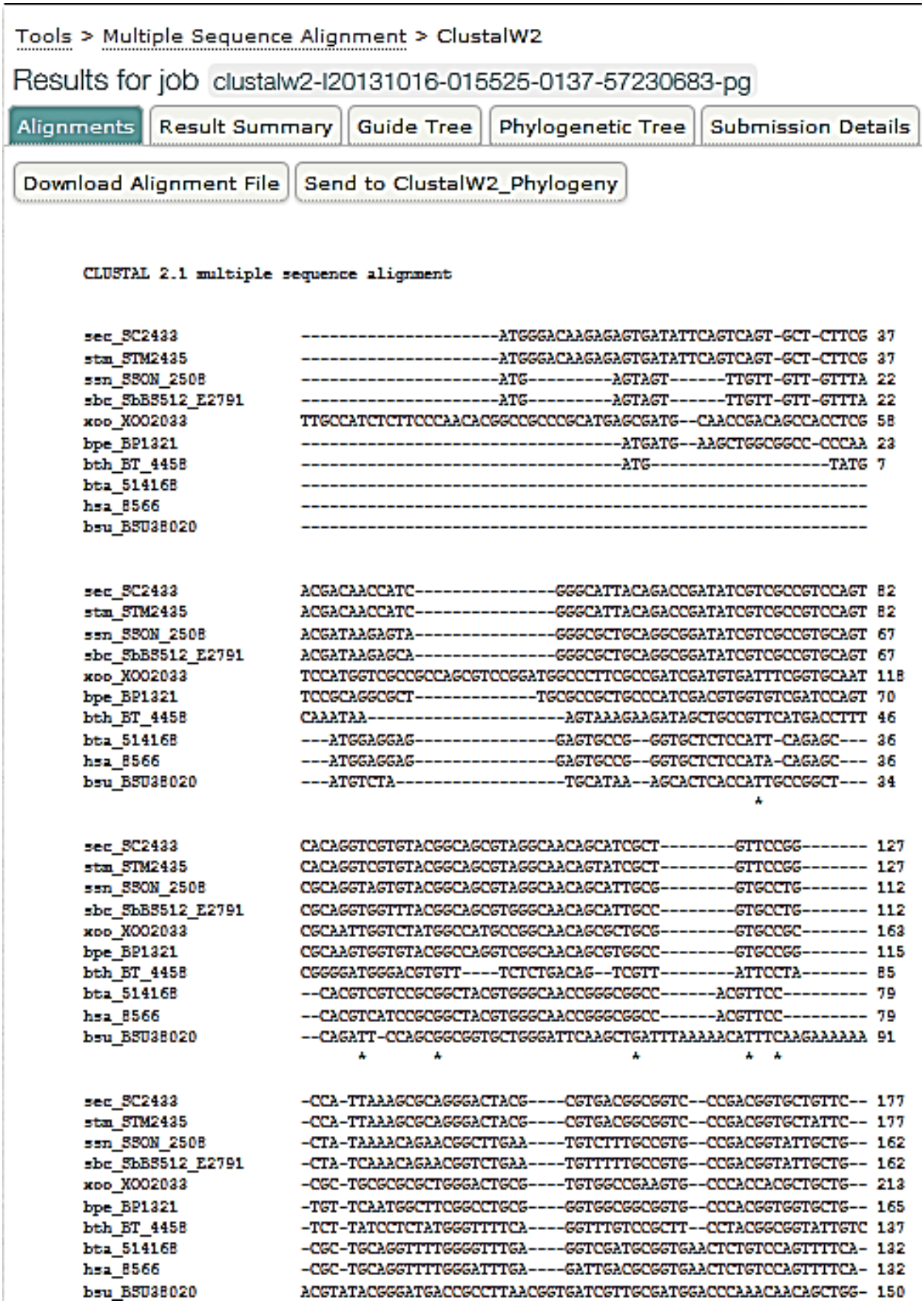


Рис. 17. Результат выравнивания нуклеотидных последовательностей в программе ClustalW2.

Кроме собственно выравнивания результатом работы являются следующие файлы:

Result summary – содержащий таблицу весов (Scores Table);

Phylogenetic Tree – содержащий в скобочной структуре информацию об эволюционном расстоянии между последовательностями;

```
(
(
(
(
sec_SC2433:0.00634373,
stm_STM2435:0.00634373):0.169308,
(
ssn_SSON_2508:0.00938969,
sbc_SbBS512_E2791:0.00938969):0.169308):0.204039,
(
xoo_XOO2033:0.173684,
bpe_BP1321:0.173684):0.204039):0.220587,
(
bta_514168:0.0697551,
hsa_8566:0.0697551):0.220587):0.22854,
(
bsu_BSU38020:0.220588,
bth_BT_4458:0.220588):0.22854):0.22854;
```

Percent Identity Matrix - содержащий матрицу идентичности между последовательностями.

Percent Identity Matrix - created by Clustal2.1

1: sec_SC2433	100.00	98.73	66.27	65.68	49.69	48.81	51.06	55.10	40.63	43.68
2: stm_STM2435	98.73	100.00	66.27	65.68	49.69	48.31	50.71	54.86	40.49	43.94
3: ssn_SSON_2508	66.27	66.27	100.00	98.12	47.87	47.87	49.58	55.66	40.08	43.49
4: sbc_SbBS512_E2791	65.68	65.68	98.12	100.00	47.74	48.50	49.82	56.39	39.95	44.14
5: bta_514168	49.69	49.69	47.87	47.74	100.00	86.05	43.75	47.18	39.13	38.19
6: hsa_8566	48.81	48.31	47.87	48.50	86.05	100.00	43.51	47.55	38.87	38.19
7: xoo_XOO2033	51.06	50.71	49.58	49.82	43.75	43.51	100.00	64.18	39.22	38.08
8: bpe_BP1321	55.10	54.86	55.66	56.39	47.18	47.55	64.18	100.00	37.41	40.13
9: bsu_BSU38020	40.63	40.49	40.08	39.95	39.13	38.87	39.22	37.41	100.00	53.77
10: bth_BT_4458	43.68	43.94	43.49	44.14	38.19	38.19	38.08	40.13	53.77	100.00

Рассмотрим пример множественного выравнивания аминокислотных последовательностей для гена pyridoxine kinase тех же организмов. Отбор аминокислотных последовательностей выполняли в базе данных UniProt (http://www.uniprot.org/).

Выбираем опцию Protein (рис. 18) и вставляем в окно программы последовательности в FASTA формате.

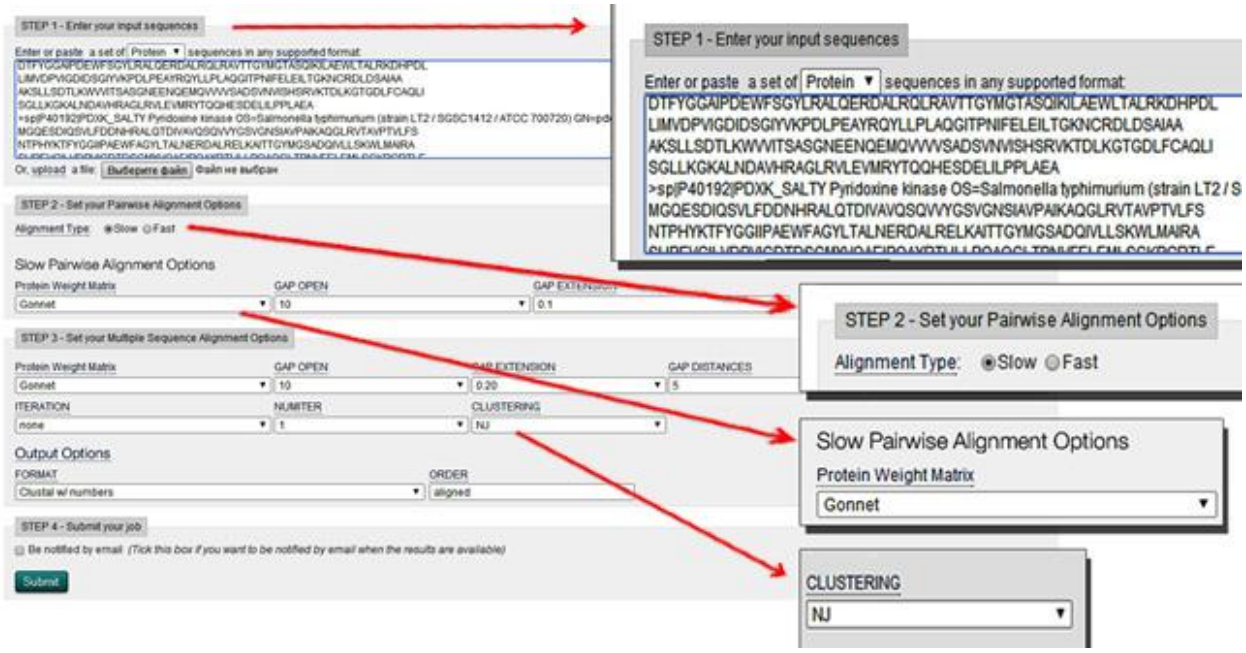


Рис. 18. Окно программы ClustalW с выставленными опциями для выравнивания аминокислотных последовательностей.

1. Выбираем следующие опции:

Alignment – slow;

Protein Weight Matrix –Gonnet;

Gap Open – 10;

Gap Extension -0,1;

CLUSTERING – NJ.

Остальные установки оставляли по умолчанию.

2. Результат представлен на рисунке 19.

Рис. 19. Результат выравнивания аминокислотных последовательностей в программе ClustalW.

Матрица PAM

Матрица GONNET

Рис. 20. Результат множественного выравнивания аминокислотных последовательностей в программе в зависимости от использования различных матриц замен.

В данном случае идентичные аминокислотные остатки отмечаются звездочкой (*), консервативные замены – двоеточием (:), а полуконсервативные – точкой (.).

Полученное выравнивание может быть отображено в черно-белой или цветной гамме, в зависимости от свойств аминокислот. Консервативность и полуконсервативность аминокислотных замен определяются в соответствии с таблицей 4. Если заменяемые аминокислоты расположены в одной группе, то замена считается консервативной.

Таблица 4

Один из возможных способов окраски аминокислотных остатков при визуализации множественного выравнивания белковых последовательностей

Цвет	Тип остатка	Аминокислоты
Желтый	Маленькие неполярные остатки	Gly, Ala, Ser, Thr
Зеленый	Гидрофобные	Cys, Val, He, Leu, Pro, Phe, Tyr, Net, Trp
Фиолетовый	Полярные	Asn, Gin, His
Красный	Отрицательно заряженные	Asp, Glu
Синий	Положительно заряженные	Lys, Arg

Как и в случае с нуклеотидными последовательностями, результатом работы кроме выравнивания, будут являться следующие файлы:

Result summary – содержащий таблицу весов (Scores Table);

Phylogenetic Tree – содержащий в скобочной структуре информацию об эволюционном расстоянии между последовательностями;

```
(
(
(
(
(
sp|Q57LS3|PDXK_SALCH:0.00000,
sp|P40192|PDXK_SALTY:0.00000)
:0.17296,
(
sp|Q3YZC3|PDXK_SHISS:0.00088,
sp|B2TX08|PDXK_SHIB3:0.00266)
:0.15683)
:0.10558,
sp|Q7VYK4|PDXK_BORPE:0.27320)
```

```
:0.02987,
tr|Q5H184|Q5H184_XANOR:0.32665)
:0.04278,
sp|O00764|PDXK_HUMAN:0.38477,
(
tr|Q89ZB9|Q89ZB9_BACTN:0.39935,
sp|P39610|PDXK_BACSU:0.41574)
:0.02317);
```

В данном случае необходимо обращать внимание на то, какую матрицу выбрали в качестве матрицы замен. Этот выбор очень сильно влияет на конечный результат (рис. 20)

ЭТАПЫ ВЫПОЛНЕНИЯ МНОЖЕСТВЕННОГО ВЫРАВНИВАНИЯ В ПРОГРАММЕ MUSCLE

На сегодняшний день наиболее оптимальной и наиболее современной из доступных on-line ресурсов, является программа MUSCLE (http://www.ebi.ac.uk/Tools/msa/muscle/). Данное сокращение произошло от полного названия программы - Multiple Sequence Comparison by Log-Expectation. MUSCLE обеспечивает большую точность выравнивания и большую скорость выполнения работы чем ClustalW и T-Coffee. На рисунке 21 представлено окно программы, находящейся на веб-сервере EMBL-EBI.

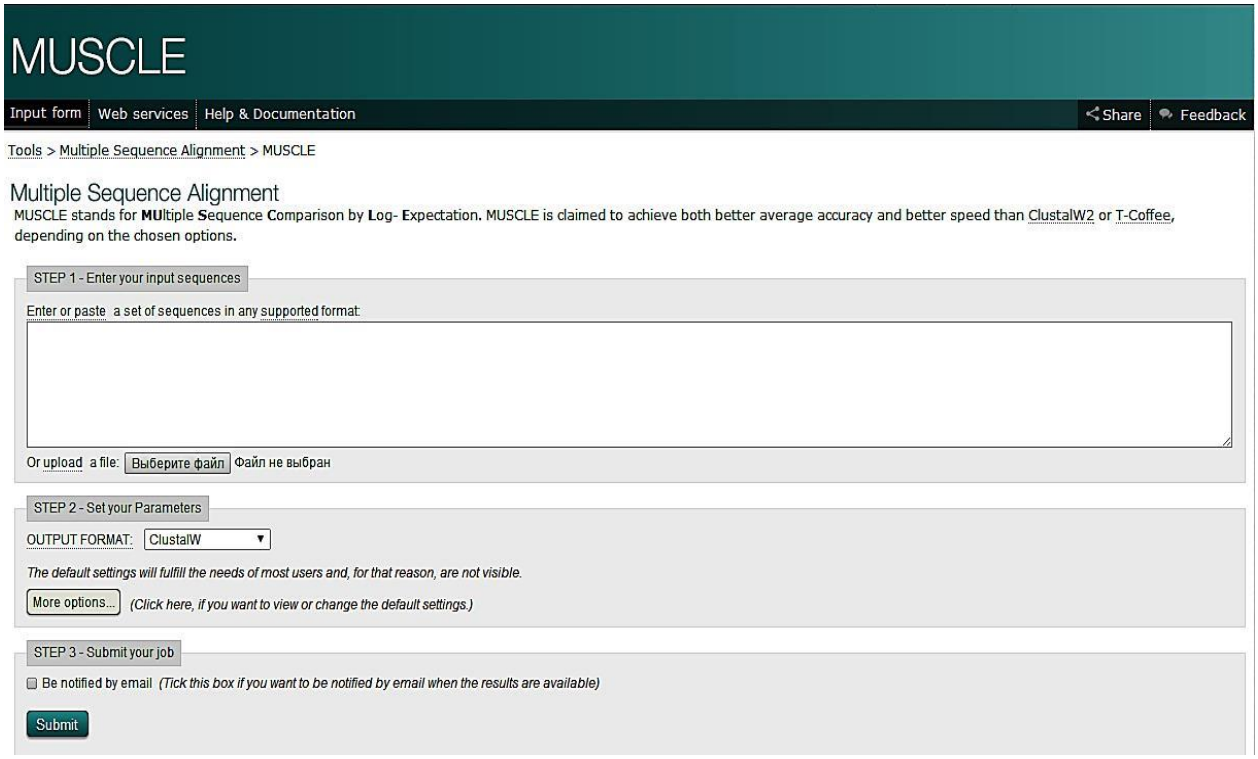


Рис. 21. Окно программы MUSCLE на браузере веб-сервера EMBL-EBI.

2.Для выполнения множественного выравнивания биологических последовательностей в этой программе так же подходит формат FASTA (см. выше). Как и для работы с программой ClustalW, необходимо создать файл с сохраненными в нем FASTA форматами последовательностей. Используем имеющийся у нас набор последовательностей.

3.Опций, которые предлагаются пользователю в этой программе не много:

utput Format – выбор формата для файла содержащего результат множественного выравнивания. Лучше оставить выставленный по умолчанию формат ClustalW, поскольку его воспринимают большинство других биоинформатических ресурсов.

Output Tree – определяет возможность вывода дерева после заданного количества итераций;

Output Order - определяет порядок, в котором последовательности появятся в конечном выравнивание (рис. 22).

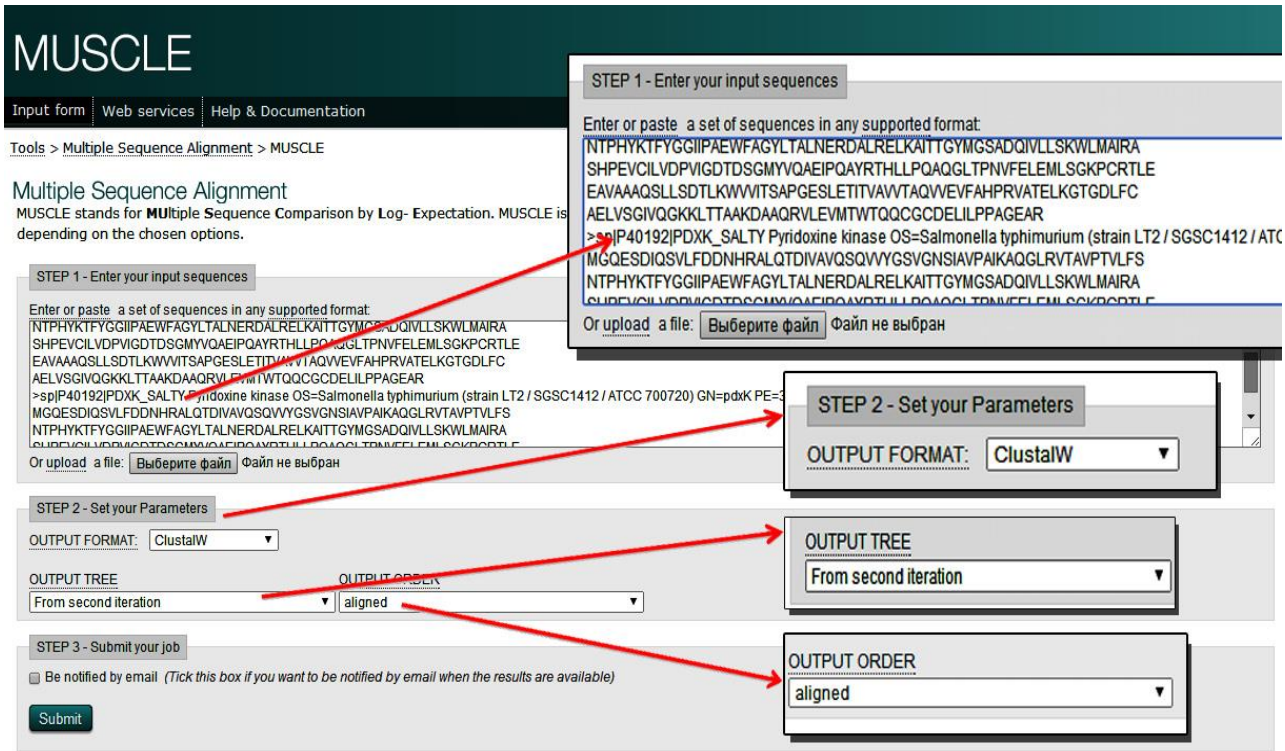


Рис. 22. Окно программы MUSCLE с выставленными опциями и введенными последовательностями.

3. После этого нажимаем Submit. Результат представлен на рисунке 23.

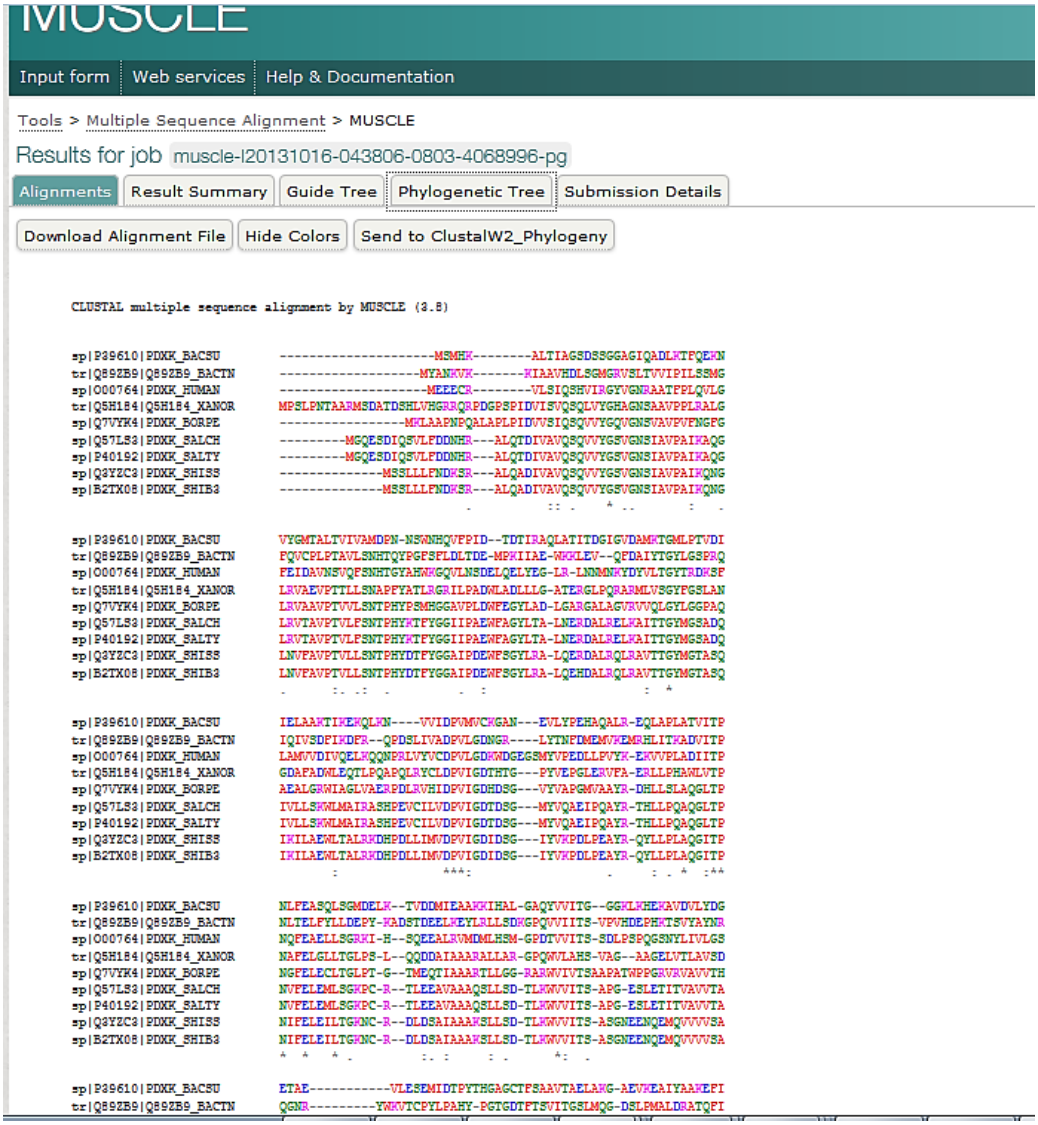


Рис. 23. Результат множественного выравнивания аминокислотных последовательностей с помощью программы MUSCLE.

Как и ClustalW программа MUSCLE, в качестве выходных файлов предлагает следующие:

Result summary – содержащий все выходные файлы;

Phylogenetic Tree – содержащий в скобочной структуре информацию об эволюционном расстоянии между последовательностями;

```
(
sp|P39610|PDXK_BACSU:0.41840,
tr|Q89ZB9|Q89ZB9_BACTN:0.38768,
(
sp|O00764|PDXK_HUMAN:0.37346,
(
```

```
tr|Q5H184|Q5H184_XANOR:0.33248,
(
sp|Q7VYK4|PDXK_BORPE:0.26819,
(
(
sp|Q57LS3|PDXK_SALCH:0.00000,
sp|P40192|PDXK_SALTY:0.00000)
:0.17232,
(
sp|Q3YZC3|PDXK_SHISS:0.00087,
sp|B2TX08|PDXK_SHIB3:0.00266)
:0.16102)
:0.09933)
:0.03409)
:0.04343)
:0.01647);
```

Percent Identity Matrix - содержащий матрицу идентичности между последовательностями.

```
1: sp|P39610|PDXK_BACSU    100.00   19.39   19.85   17.29   23.08   21.07   21.07   21.84   21.84
2: tr|Q89ZB9|Q89ZB9_BACTN   19.39  100.00   21.55   22.18   24.34   27.24   27.24   29.10   29.10
3: sp|O00764|PDXK_HUMAN     19.85   21.55  100.00   26.50   25.93   27.31   27.31   27.68   27.68
4: tr|Q5H184|Q5H184_XANOR   17.29   22.18   26.50  100.00   38.21   33.45   33.45   36.65   36.30
5: sp|Q7VYK4|PDXK_BORPE     23.08   24.34   25.93   38.21  100.00   47.12   47.12   46.04   45.68
6: sp|Q57LS3|PDXK_SALCH     21.07   27.24   27.31   33.45   47.12  100.00  100.00   66.67   66.31
7: sp|P40192|PDXK_SALTY     21.07   27.24   27.31   33.45   47.12  100.00  100.00   66.67   66.31
8: sp|Q3YZC3|PDXK_SHISS     21.84   29.10   27.68   36.65   46.04   66.67   66.67  100.00   99.65
9: sp|B2TX08|PDXK_SHIB3     21.84   29.10   27.68   36.30   45.68   66.31   66.31   99.65  100.00
```

ЭТАПЫ ВЫПОЛНЕНИЯ МНОЖЕСТВЕННОГО ВЫРАВНИВАНИЯ В БАЗЕ ДАННЫХ UniProt (http://www.uniprot.org/)

Выполнить множественное выравнивание последовательностей можно непосредственно в базе данных. В данном случае рассмотрим хороший ресурс, представленный в базе UniProt (http://www.uniprot.org/). Ограничения накладываются только на тип анализируемых последовательностей: поскольку это база данных аминокислотных последовательностей, то и алгоритмы, реализующие множественное выравнивание основаны только на матрицах замен, а следовательно, выполнить анализ нуклеотидных последовательностей невозможно.

1. Как и для работы с программой ClustalW, необходимо создать файл с сохраненными в нем FASTA форматами последовательностей. Используем имеющийся у нас набор последовательностей.

2. В верхней части страницы базы данных UniProt, выберем инструмент (toolbar) Align (рис. 24). В поле Sequences, вставить набор последовательностей в FASTA формате из своего файла. Затем нажать кнопку Aling (рис. 24).

3. Результатом выполненых действий будет множественное выравнивание аминокислотных последовательностей (рис. 25). Полученный результат, сравним с результатом множественного выравнивания для аминокислотных последовательностей, проводимого в программе CluastalW с использованием матрицы PAM (рис. 20).

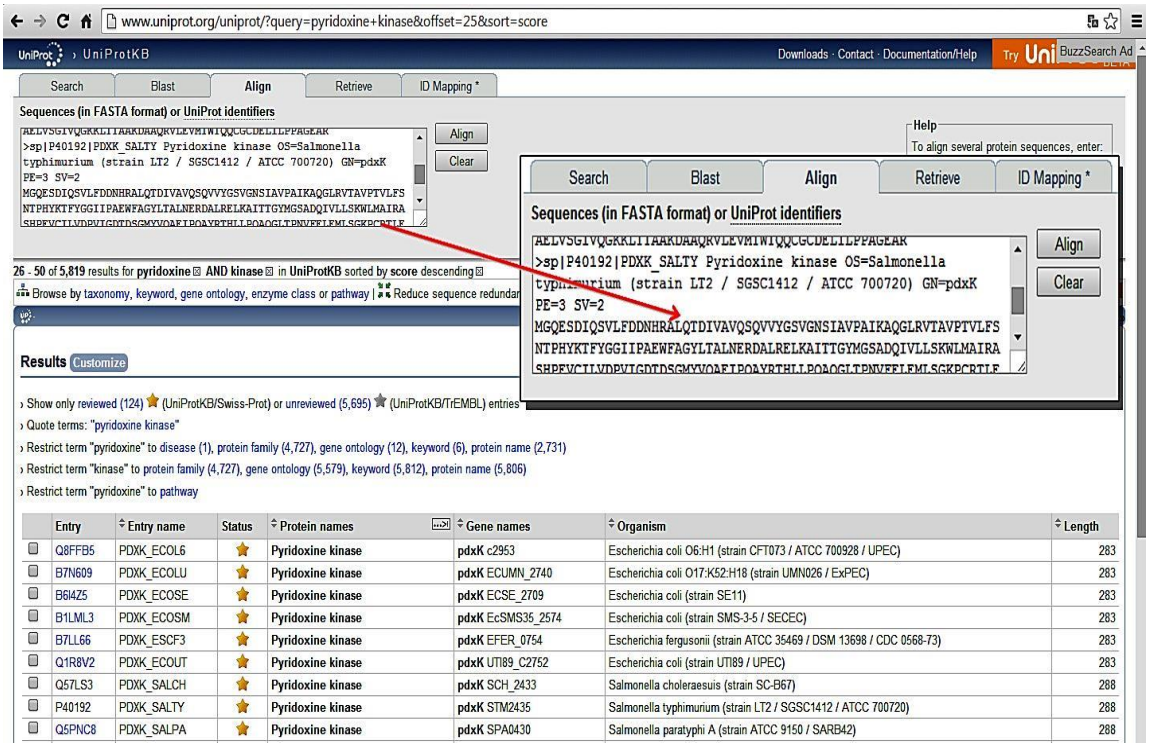


Рис. 24. Страница базы данных UniProt с доступом к toolbar Align.

Как мы видим по результатам множественного выравнивания, если во всех программах выставлены одинаковые опции (например, выбраны матрицы замен только PAM или BLOSUM), то и результат выравнивания будет одинаковым.

Может смутить разница в формировании последовательностей в Ouput file, однако следует помнить, что в алгоритм каждой из программ по множественному выравниванию последовательностей заложено попарное

выравнивание, а формирование результирующего вида бинарного дерева является индивидуальным подходом разработчика.

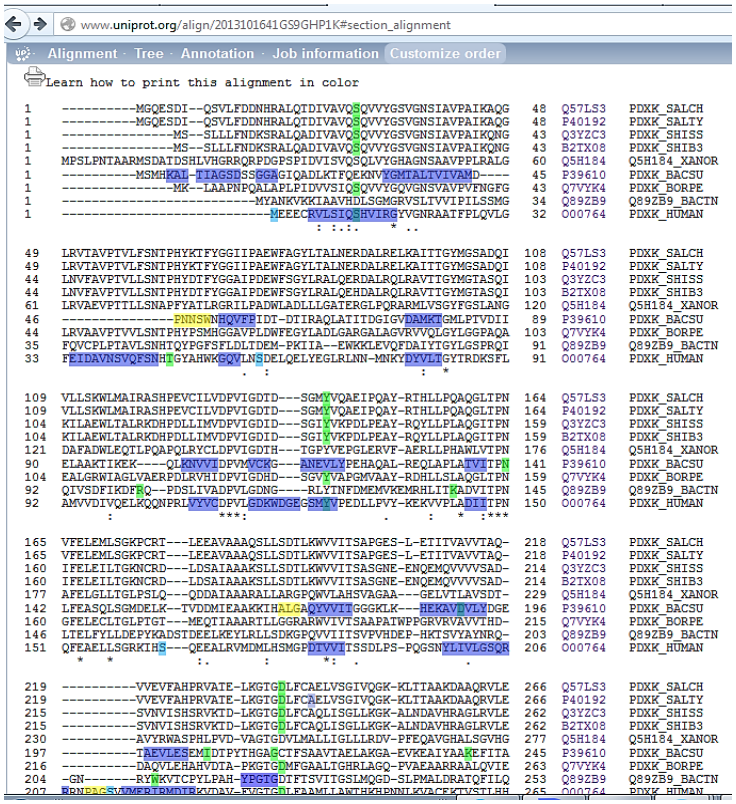


Рис. 25. Результат множественного выравнивания с использование предлагаемого toolbar Align в базе данных UniProt.

Сравнение исходной очередности заданных последовательностей и результирующего Ouput file для каждой использованной программы, приведены в таблице 5.

Таблица 5

Сравнение очередности последовательностей в исходном и Ouput файлах в зависимости от использованной программы.

Исходные последовательности	Последовательности в Ouput file программы Clastalw	Последовательности в Ouput file программы MUSCLE	Последовательности в Ouput file программы toolbar Align в базе данных UniProt
P39610 Bacillus subtilis	sp|Q57LS3|PDXK_SALCH	sp|P39610|PDXK_BACSU	P39610 PDXK_BACSU1
Q89ZB9 Bacteroides thetaiotaomicron	sp|P40192|PDXK_SALTY	tr|Q89ZB9|Q89ZB9_BACTN	Q89ZB9 Q89ZB9_BACTN1
Q5H184 Xanthomonas oryzae	sp|Q3YZC3|PDXK_SHISS	tr|Q5H184|Q5H184_XANOR	Q5H184 Q5H184_XANOR
Q7VYK4 Bordetella pertussis	sp|B2TX08|PDXK_SHIB3	sp|Q7VYK4|PDXK_BORPE	Q7VYK4 PDXK_ BORPE
Q57LS3 Salmonella choleraesuis	tr|Q5H184|Q5H184_XANOR	sp|Q57LS3|PDXK_SALCH	Q57LS3 PDXK_SALCH1
P40192 Salmonella typhimurium	sp|Q7VYK4|PDXK_BORPE	sp|P40192|PDXK_SALTY	P40192 PDXK_SALTY1
Q3YZC3 Shigella sonnei	tr|Q89ZB9|Q89ZB9_BACTN	sp|Q3YZC3|PDXK_SHISS	Q3YZC3 PDXK_SHISS1
B2TX08 Shigella boydii	sp|P39610|PDXK_BACSU	sp|B2TX08|PDXK_SHIB3	B2TX08 PDXK_SHIB3

По результатам выравнивания последовательностей гена pyridoxine kinase у различных организмов можно сделать следующее заключение: наиболее сходными между собой по выбранной последовательности являются *Shigella sonnei* и *Shigella boydii*, а так же *Salmonella choleraesuis* и *Salmonella typhimurium*. Процент идентичности в данных случаях 99.65 и 66.43 %.

Наиболее далекими между собой (по количеству эволюционных событий) являются *Bacillus subtilis* и *Xanthomonas oryzae***.** Процент идентичности **–** 7.01 %.

ВЫПОЛНЕНИЕ МНОЖЕСТВЕННОГО ВЫРАВНИВАНИЯ В ПРОГРАММЕ MATLAB

Для выполнения множественного выравнивания в программе MatLab используется команда multialign.

В общем виде программа в Matlab для построения филогенетического дерева выглядит следующим образом:

seqs = fastaread('pf01.fa')– чтение биологических данных из файла в fasta формате

ma = multialign(seqs,'verbose',true) – выполнение множественного выравнивания с включением информации о последовательностях

showalignment(ma) – вывод множественного выравнивания в отдельное окно

Рассмотрим более подробно эти функции.

1. seqs =fastaread(file) – считывает данные из файла в FASTA формате в структуру MATLAB и создает массив данных seqs, включающий в себя бинарную структуру для каждой последовательности с информацией о тесте последовательности и данные об ней (рис. 26). Перед загрузкой данных в рабочую среду Matlab биологические последовательности, полученные из баз банных или других источников, сохраняют в файле с расширением **fa** в отдельной рабочей директории.

2. multialign – команда, выполняющая прогрессивное множественное выравнивание для набора последовательностей в Matlab

Синтаксис

SeqsMultiAligned = multialign(*Seqs*)

SeqsMultiAligned =multialign(*Seqs*, 'Weights', *WeightsValue*)

SeqsMultiAligned =multialign(*Seqs*, 'ScoringMatrix', *ScoringMatrixValue*)

SeqsMultiAligned =multialign(*Seqs*, 'GapOpen', *GapOpenValue*)

SeqsMultiAligned =multialign(*Seqs*, 'ExtendGap', *ExtendGapValue*)

SeqsMultiAligned =multialign(*Seqs*, 'Verbose', *VerboseValue*)

Описание

SeqsMultiAligned = multialign (Seqs) выполняет прогрессивное множественное выравнивание для набора последовательностей (Seqs). Расстояния между последовательностями вычисляются после проведения предварительного попарного выравнивания на основании матрицы (по умолчанию), с учетом количества различий для каждой последовательности, но с игнорированием пробелов (gap). Конструирование дерева производится на основании алгоритма прогрессивного выравнивания с использованием метода «Ближайших соседей».

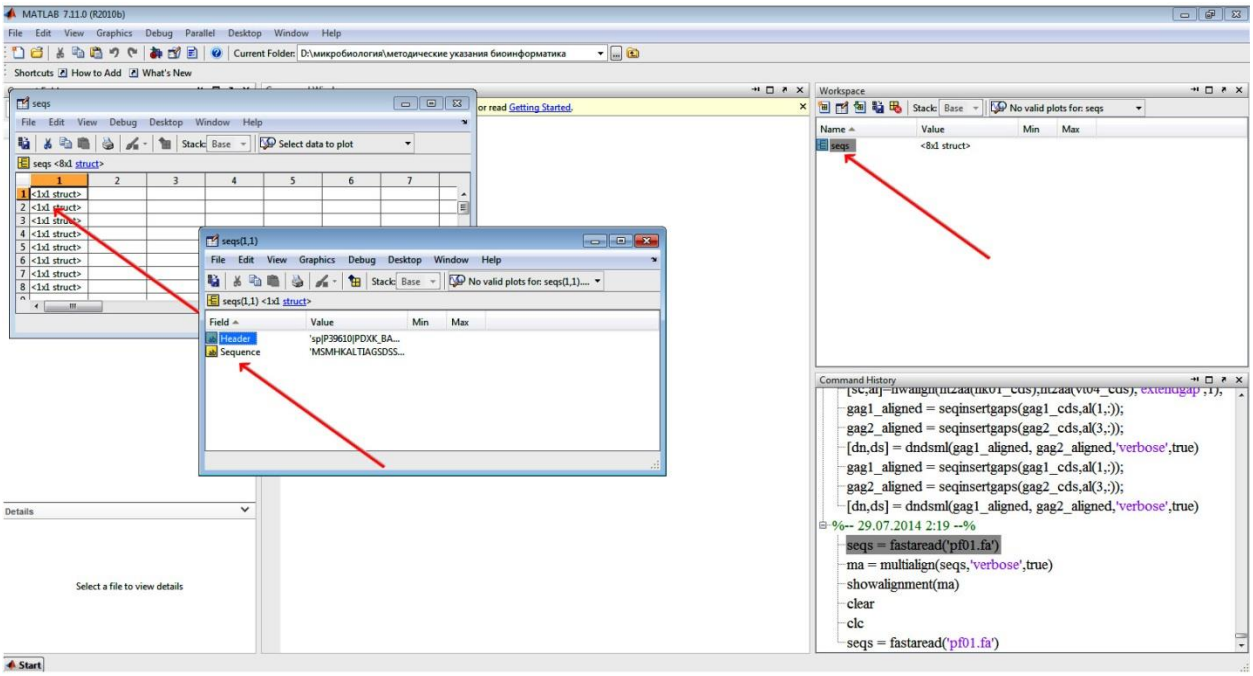


Рис. 26. Массив seqs и его структуры в программе Matlab, хранящие тест последовательности и данные о ней.

SeqsMultiAligned = multialign(..., 'Weights', *WeightsValue*) выбирает метод определения веса взвешивания последовательностей. Вес придает особую значимость несхожим последовательностям, путем масштабирования весовой матрицы и начисления штрафов за пробелы.

Значение свойства 'Weights':

- 'THG ' (по умолчанию) – метод Томпсон -Хиггинса – Гибсона учитывающий разницу веса каждой ветви дерева.
- 'equal' - метод, который присваивает одинаковый вес каждой последовательности.

SeqsMultiAligned = multialign (..., ' ScoringMatrix , ScoringMatrixValue) выбирает матрицу замен (ScoringMatrixValue) для выравнивания. Баллы за совпадения и несовпадения назначаются с учетом расстояния между двумя последовательностями.

Выбор для аминокислотных последовательностей:
'BLOSUM62'
'BLOSUM30'
'BLOSUM100'
'PAM10 '
'DAYHOFF'
'GONNET'
По умолчанию:
'BLOSUM50' – В случае параметра «AlphabetValue» – "AA"
'NUC44' - В случае параметра «AlphabetValue» – 'NT'

SeqsMultiAligned = multialign(..., 'GapOpen, GapOpenValue) определяет начальный штраф за открытие гепа.

SeqsMultiAligned = multialign(..., 'ExtendGap, ExtendGapValue) определяет начальный штраф за открытие продолжение гепа.

SeqsMultiAligned = multialign(...,'Verbose', VerboseValue), управляет включением дополнительной информации о последовательности. По умолчанию false .

3. showalignment(*Alignment*) – команда показывающая выравнивание с использованием цвета (рис. 27).

КОНСТРУИРОВАНИЕ ФИЛОГЕНЕТИЧЕСКИХ ДЕРЕВЬЕВ НА ОСНОВАНИИ МНОЖЕСТВЕННОГО ВЫРАВНИВАНИЯ

Филогения – это описание отношений меду биологическими последовательностями (организмами), обычно изображаемое в виде дерева. Отмеченные подобия и различия между последовательностями (организмами) используют для восстановления филогении.

Филогенетический анализ в систематике описывает взаимоотношения среди таксонов и призван помочь нам понять историю эволюционных отношений между живыми организмами. Эволюционную историю, восстановленную в результате филогенетического анализа, обычно изображают в виде разветвлённых, древовидных диаграмм, которые представляют предполагаемую родословную наследственных отношений между молекулами, организмами или и тем, и другим. Основой для понимания этих процессов являются количественные отношения эволюционных событий произошедших у каждого отдельного организма с

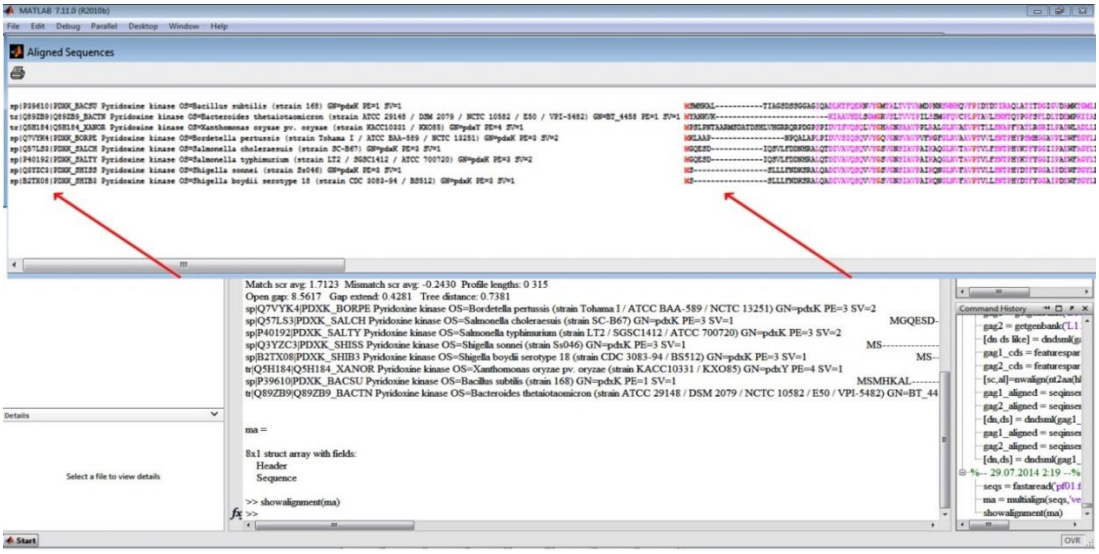


Рис. 27. Результат множественного выравнивания биологических последовательностей в Matlab.

момента его отделения от общего предка. В данном случае мы рассматриваем только косвенные показатели фактических событий. В филогенетике наиболее удобный путь визуального представления эволюционных взаимоотношений

среди групп организмов осуществляется посредством графиков, которые называются филогенетическими деревьями. Для выполнения подобных филогенетических реконструкций удобно использовать (в качестве черновых макетов) on-line программы. Наиболее удобной для студентов является пакет программ «Robust Phylogenetic Analysis For The Non-Specialist», представленный на сайте http://www.phylogeny.fr/.

Phylogeny.fr это бесплатный, простой в использовании веб-сервис посвященный реконструкции и анализу филогенетических связей между биологическими последовательностями. В основе работы сервера Phylogeny.fr лежит объединение различных биоинформационных программ с целью реконструкции надежного филогенетического дерева из набора последовательностей.

ЭТАПЫ ВЫПОЛНЕНИЯ ФИЛОГЕНЕТИЧЕСКОГО АНАЛИЗА В ПАКЕТЕ Phylogeny.fr

1. Первоначально нам необходимо иметь файл с нуклеотидными или аминокислотными последовательностями. Используем уже имеющийся у нас файл с последовательностями, который мы использовали для множественного выравнивания с FASTA форматами.

Заходим на нужную нам страничку (http://www.phylogeny.fr/). **Совет**: используйте веб-браузер Mozilla Firefox. Этот веб-браузер быстрее работает с подобного типа сайтами и программами чем Internet Explorer или Google Chrome (рис. 28).

2. Для того чтобы выполнить реконструкцию филогенетического дерева необходимо совершить несколько действий. В первую очередь провести множественное выравнивание в предложенных вариантах на сайте. Принцип и программы идентичны тем, что мы рассматривали выше. Это известные нам MUSCLE, CluastalW. Из неизвестных нам программ – T-Coffee и ProbCons (рис. 29). Выбираем (к примеру) программу MUSCLE и

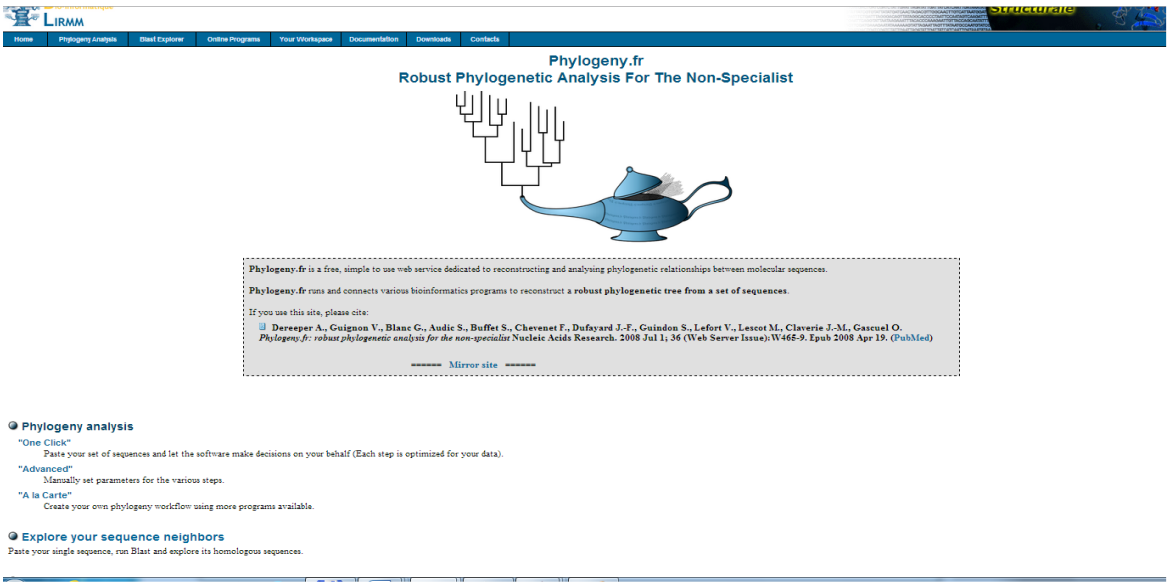


Рис. 28. Начальная страница сервера Phylogeny.fr.

нажимаем на линкованную ссылку. Переходим на страничку с окном программы (рис. 30) и вставляем наши последовательности из файла. В данном случае не надо самим выставлять опции для типа последовательности. Затем жмем Submit и ждем результат работы (рис. 30)

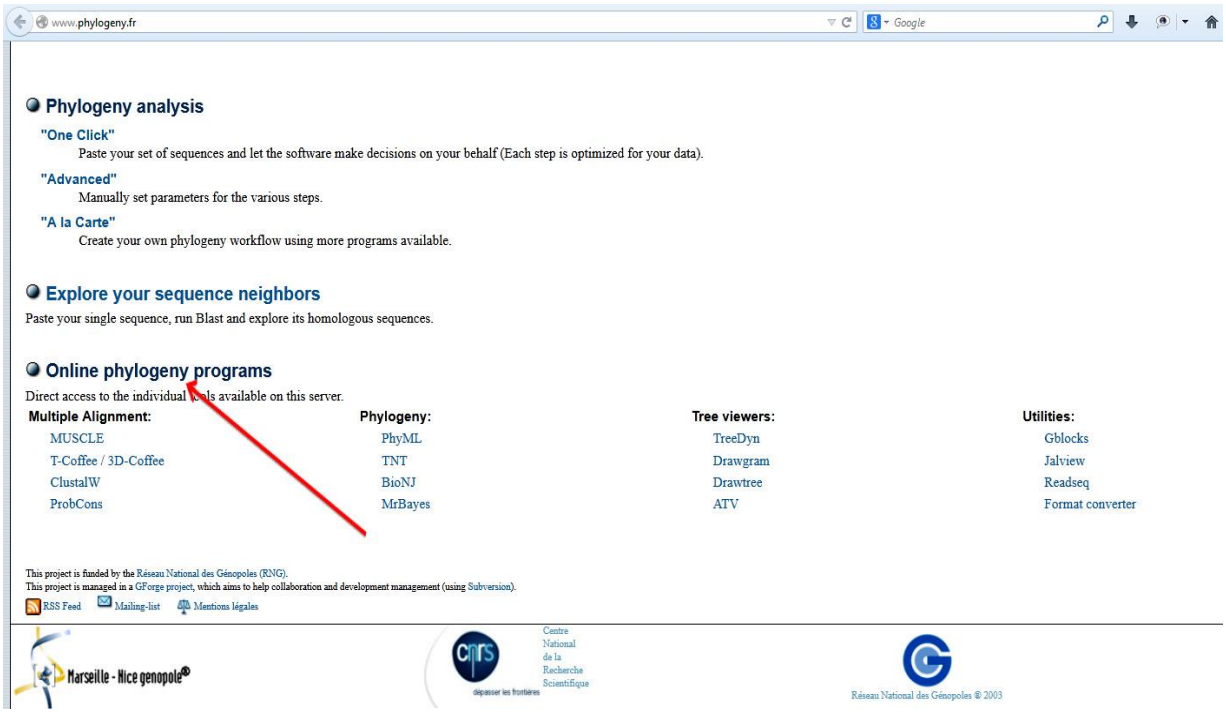


Рис. 29. Начальная страница сервера Phylogeny.fr с блоками программ.

Для дальнейшей работы нам необходимо получить выравненные последовательности в одном из предложенных программой форматов. Чтобы

было понятнее выберем FASTA формат (рис. 31). Результат выравнивания откроется в соседнем окне и его можно будет скопировать или сохранить.

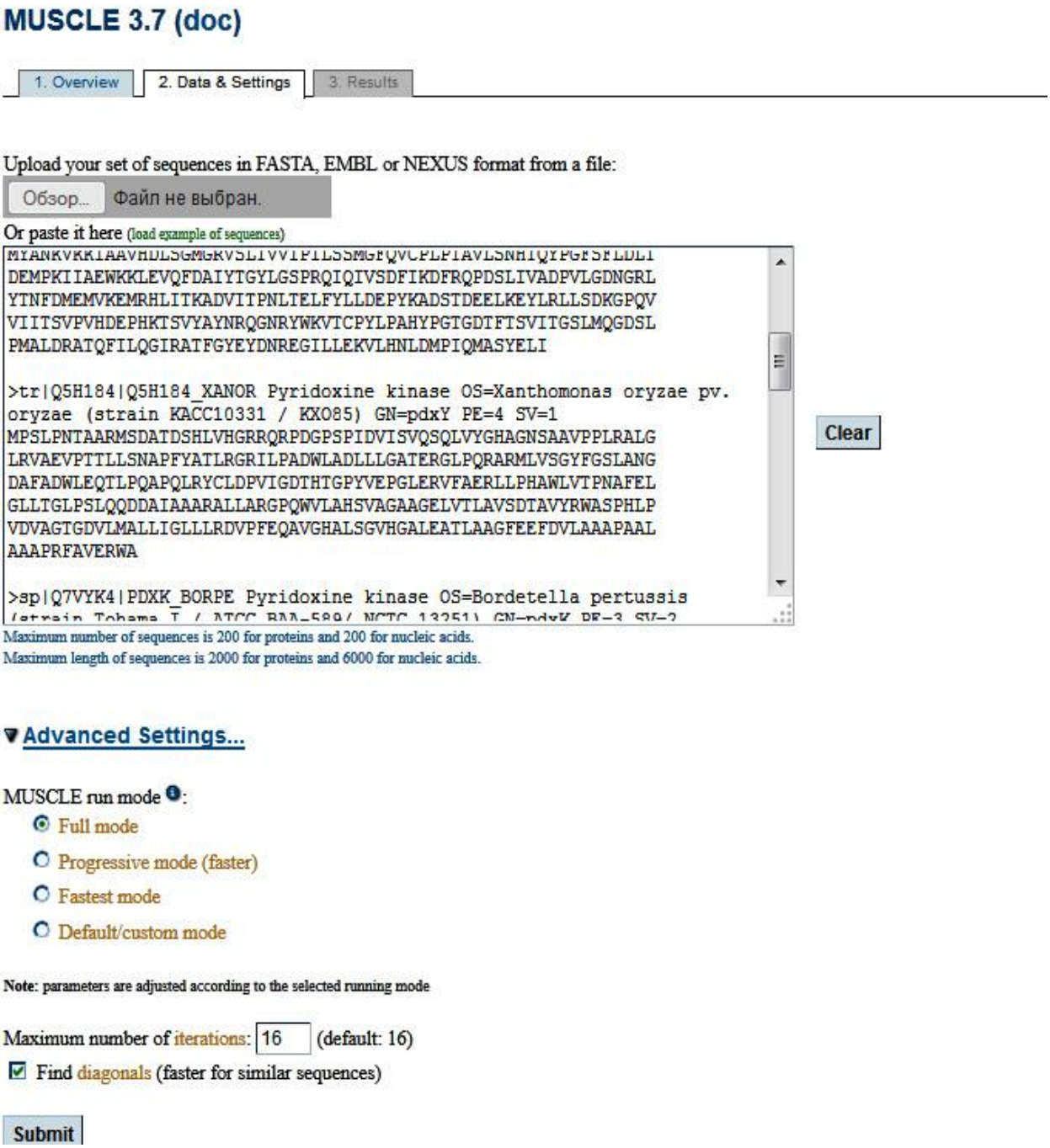


Рис. 30. Окно программы MUSCLE на сервере Phylogeny.fr.

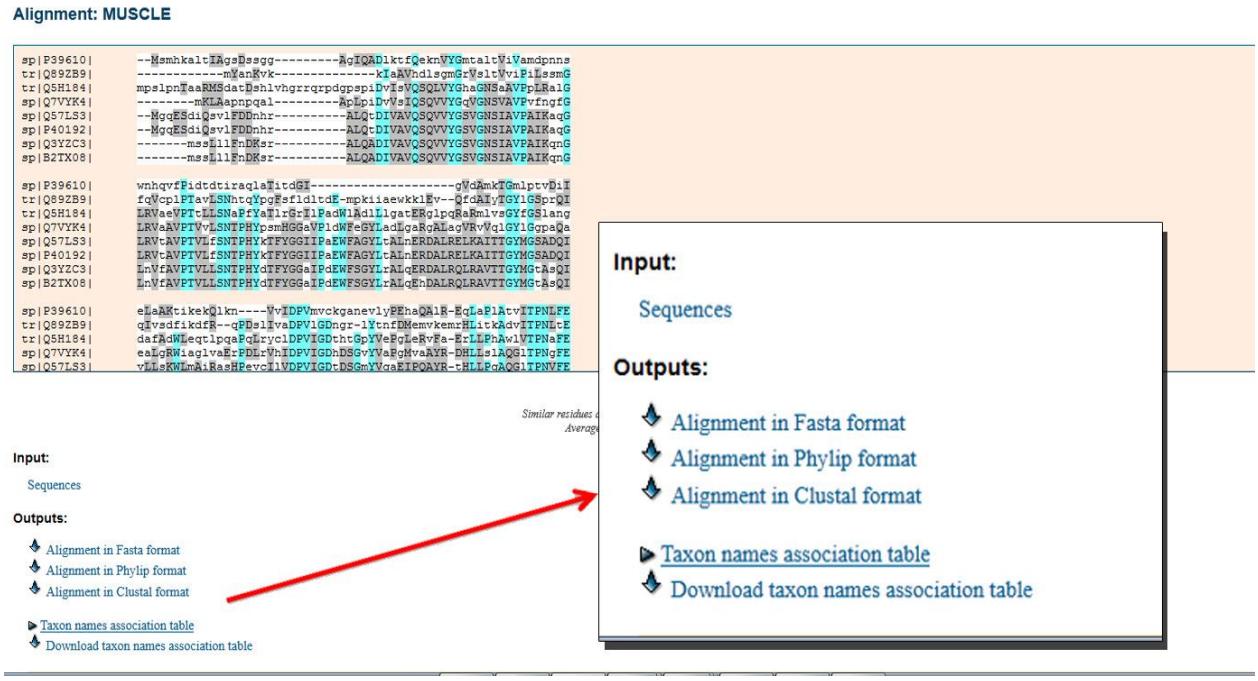


Рис. 31. Результат выравнивания биологических последовательностей в программе MUSCLE на сервере **Phylogeny.fr.**

2. Следующим этапом будет работа с блоком программ связанных с филогенией. Разница между ними заключается в заложенном алгоритме, для каждой программы. На рисунке 32 стрелкой указан блок программ, в который включены программы по реконструкции филогенетических отношений между заданными последовательностями. Это программы PhyML, TNT, BioNJ, Mr Bayes. Рассмотрим их поподробнее.

3. Программы блока Phylogeny:

PhyML – использует алгоритм, основанный на методе максимального правдоподобия. Метод максимального правдоподобия или метод наибольшего правдоподобия (MLE — Maximum Likelihood Estimation) в математической статистике — это метод оценивания неизвестного параметра путём максимизации функции правдоподобия. Этим методом обычно используются, чтобы найти эволюционное дерево или деревья, которые наилучшим образом объясняют наблюдаемые изменения в группе последовательностей. Метод максимального правдоподобия требует понимания вероятностных методов в эволюционной модели.

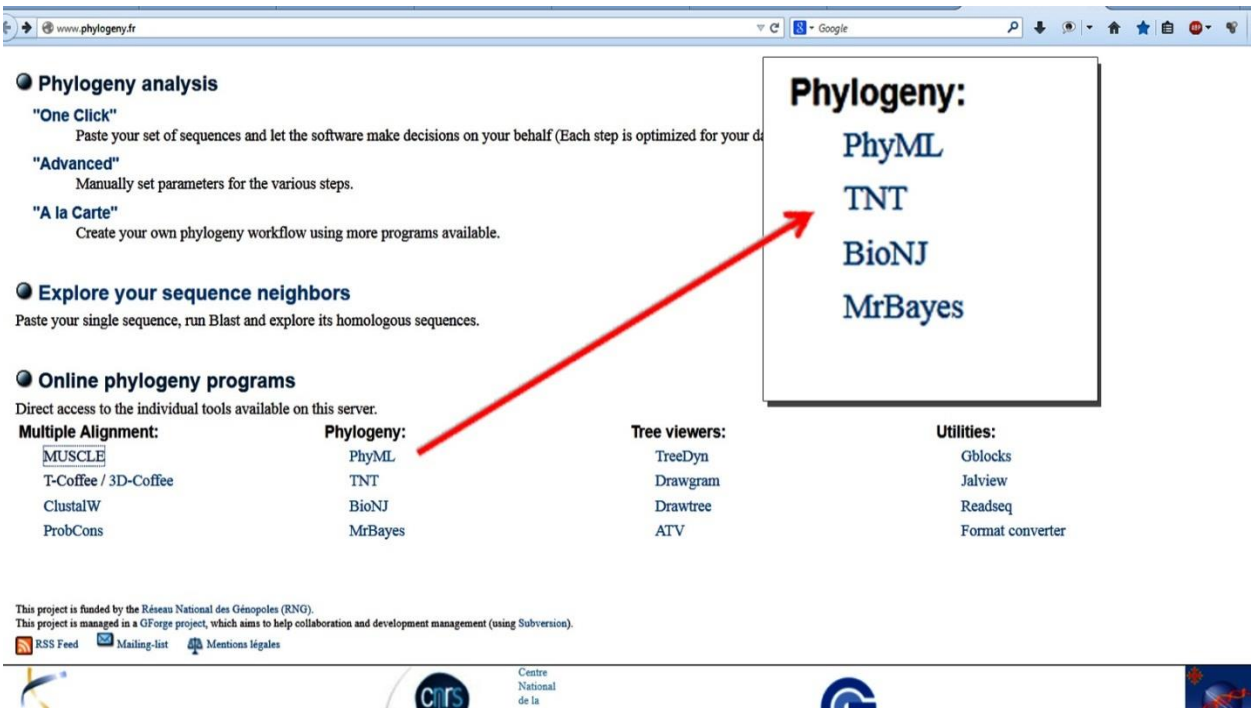


Рис. 32. Страница сервера Phylogeny.fr **с указанным блоком программ, выполняющих филогенетическую реконструкцию.**

Недостатком данного алгоритма является то, что на обычных компьютерах нельзя качественно и быстро обработать большое количество длинных последовательностей, поскольку это многошаговые алгоритмы, основанные на многократном переборе последовательностей и множестве вычислительных шагов. При использовании метода максимального правдоподобия, мы получим значения параметров модели, которые делают данные «более близкими» к реальным. **Совет**: не используйте последовательности длиннее 60 букв.

TNT – принцип максимальной экономии, реализованный в данном алгоритме, сформулирован в методологическом принципе именуемом «Бритва Оккама». В кратком виде он гласит: *«Не следует множить сущее без необходимости».* В филогении данный принцип выражается в том, что наиболее предпочтительным филогенетическим деревом является то, которое предполагает минимальное количество эволюционных изменений, но максимально объясняет эволюционный процесс, т.е. наиболее консервативным является дерево, которое требует наименьшее число мутаций для всех последовательностей. Принцип максимальной экономии популярный алгоритм

реконструкции эволюционных отношений. Однако как и алгоритм максимального правдоподобия это многошаговые алгоритмы, основанные на многократном переборе последовательностей и множестве вычислительных шагов. По мере увеличения количества анализируемых последовательностей, число возможных деревьев растет, что усложняет обработку данных

BIoNJ – алгоритм объединения ближайших соседей (Neighbor-joining) представляет собой метод, который связан с кластерным методом, но не требует, чтобы данные были ультраметрические. Этот метод особенно подходит для набора данных содержащих последовательности, в которых эволюционные события происходили с различной скоростью. Алгоритм требует расчета расстояния между каждой парой таксонов перед объединением их в единое дерево. Для этого строится модифицированная матрица расстояний Данный алгоритм является модификацией метода UPGMA (Unweighted Pair Group Method using arithmetic Averages).

MrBayes – данный алгоритм, называемый принципом Байеса, использует байесовский анализ и цепи Маркова в методе Монте-Карло (MCMC) для оценки апостериорного распределения параметров модели. В данном случае параметр распределения состоит из филогенетического дерева и модели эволюции, основанной на предварительном параметре и вероятности, полученной по результатам множественного выравнивания. Байесовский вывод все чаще применяется в молекулярной филогенетике, для оценки вида филогении и времени дивергенции последовательностей. **Уточнение**: считает очень долго, поскольку требует большого количества интераций.

4. Результатом работы каждой из программ будет структура филогенетического дерева, основанная на использованном алгоритме (рис. 33). Однако, для дальнейшей работы, нас будет интересовать не графическое представление результатов, а один из output файлов. Нам необходимо получить

информации о сгенерированном дереве виде Newick format (рис. 33).

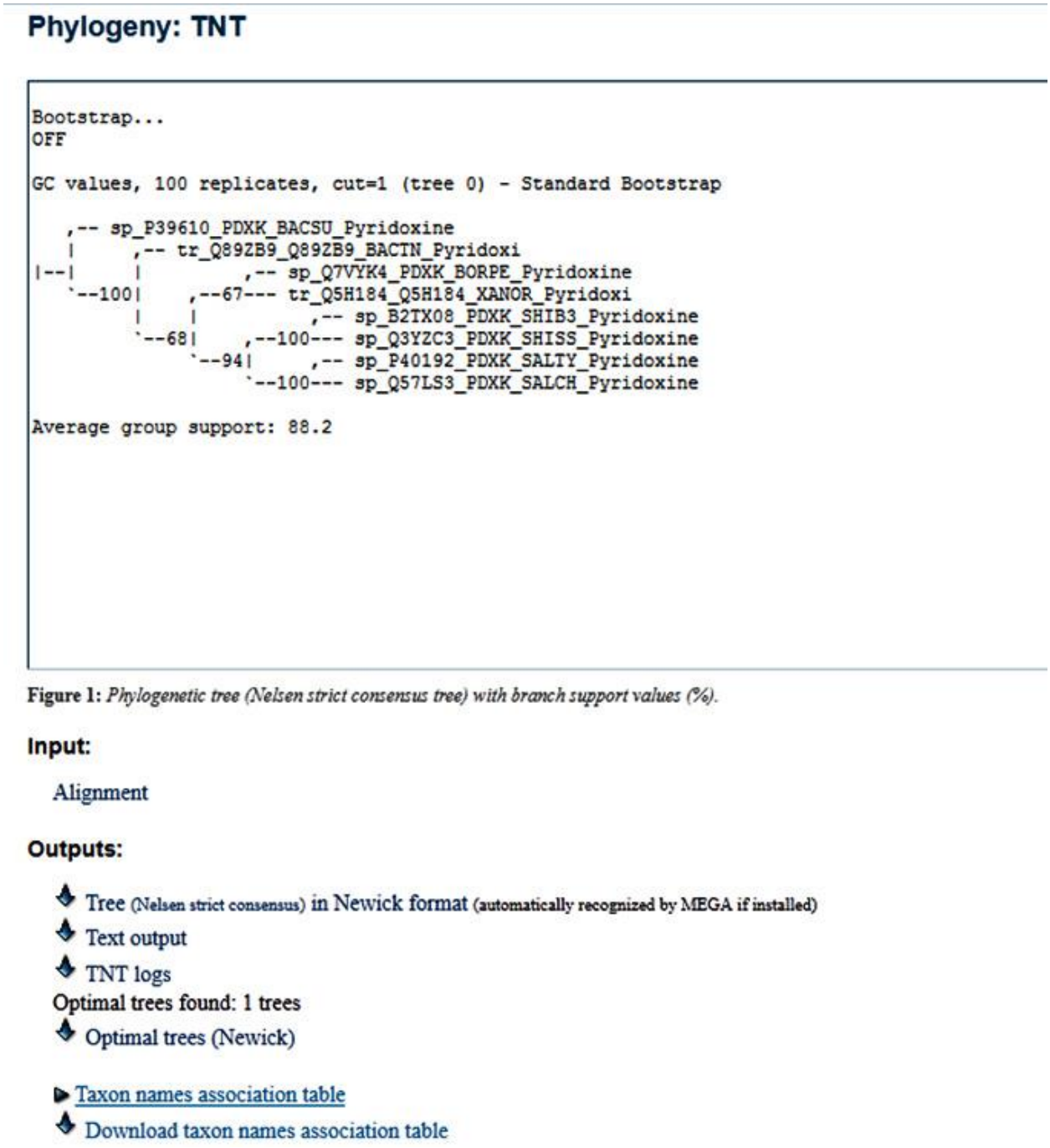


Рис. 33. Результат реконструкции филогенетических отношений между последовательностями (программа TNT).

Для выбранного примера (программа TNT) результат работы в виде Newick формата выглядит следующим образом:

(sp_P39610_PDXK_BACSU_Pyridoxine,(tr_Q89ZB9_Q89ZB9_BACTN_Pyridoxi,((tr_Q5H184_Q5H184_XANOR_Pyridoxi,sp_Q7VYK4_PDXK_BORPE_Pyridoxine),((sp_Q57LS3_PDXK_SALCH_Pyridoxine,sp_P40192_PDXK_SALTY_Pyridoxi

ne),(sp_Q3YZC3_PDXK_SHISS_Pyridoxine,sp_B2TX08_PDXK_SHIB3_Pyridoxine)))));

5. Последним этапом работы является собственно создание филогенетического дерева с использованием скобочной структуры, записанной в Newick формате. Для этого необходимо выбрать блок с соответствующими программами, представленными на сайте **Phylogeny.fr.** На рисунке 34 стрелкой указан необходимый блок программ – **Tree viewers**.

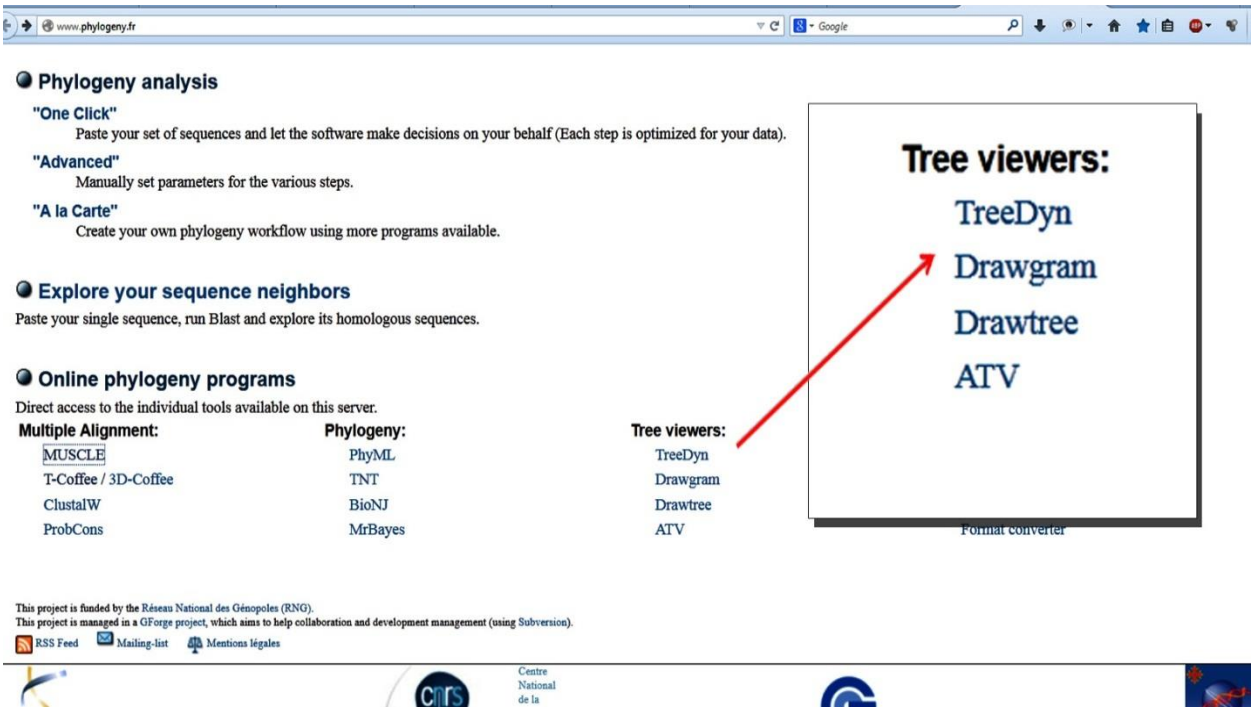


Рис. 34. Страница сервера Phylogeny.fr **с указанным блоком программ, выполняющих построение филогенетических деревьев.**

Отличия в окончательном результате будут заключаться в типе дерева, которое будет построено. Так, например программа **TreeDyn** хороший инструмент для визуализации дерева, использующий дополнительно информацию, связанную с последовательностью: таксономическое описание, идентификатор в базе данных, описание функции и т.д. Данная программа строит корневое дерево и учитывает эволюционное расстояние между последовательностями. Результат работы программы TreeDyn представлен на рисунке 35.

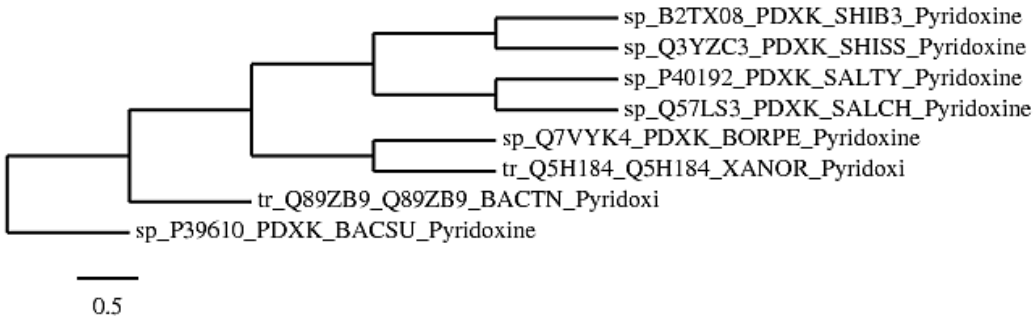
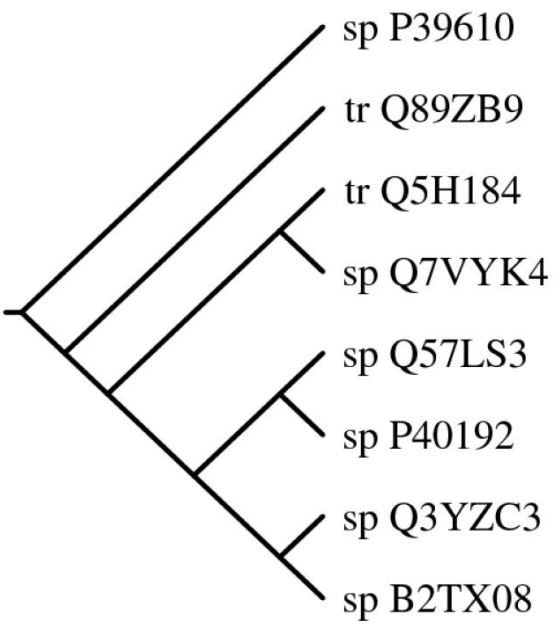


Рис. 35. Филогенетическое дерево построенное с помощью программы TreeDyn на основании филогенетической реконструкции, выполненной в программе TNT.

Программа **Drawgram** не учитывает эволюционные расстояния. Способ построения филогенетических деревьев в этой программе основан на кладистическом подходе. Данная программа строит корневое дерево и снабжает основной информацией о последовательностях. Результат работы программы Drawgram представлен на рисунке 36.

А)



Б)

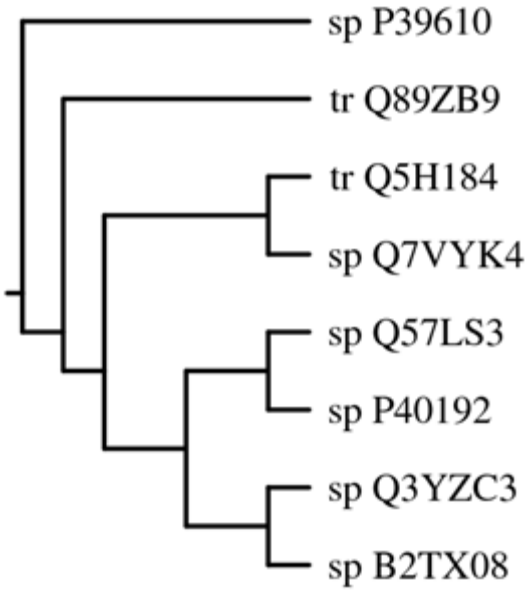


Рис. 36. Филогенетические деревья, построенные с помощью программы Drawgram (А- кладограмма, Б) – фенограмма) на основании филогенетической реконструкции, выполненной в программе TNT.

Программа **DrawTree** строит неукорененное филогенетическое дерево, что особенно важно, если мы не знаем истинной предковой последовательности. Результат работы программы DrawTree представлен на рисунке 37.

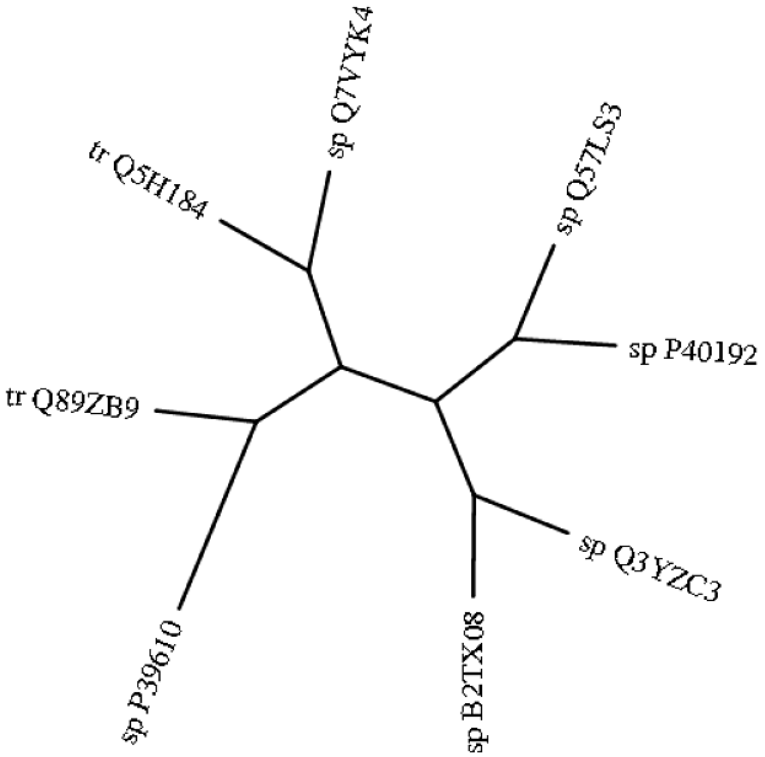


Рис. 37. Филогенетическое дерево построенное с помощью программы DrawTree на основании филогенетической реконструкции, выполненной в программе TNT.

КОНСТРУИРОВАНИЕ ФИЛОГЕНЕТИЧЕСКИХ ДЕРЕВЬЕВ В ПРОГРАММЕ MATLAB

Поскольку работа в программе Matlab основывается на матрицах, то и алгоритмы, заложенные в ней для построения филогенетических деревьев, относятся к фенетическому подходу, не имеющему никакого отношения к исторической модели родства между последовательностями. В этом случае

начинают с измерения расстояний между последовательностями и строят дерево с помощью процедуры иерархический кластеризации.

Кластеризация - это задача разбиения исходного множества данных на подмножества, называемые кластерами, при котором каждый объект может быть отнесен к одному или нескольким заранее неизвестным классам.

Иерархическая кластеризация – это многоступенчатое группирование кластеров из кластеров.

В общем виде программа в Matlab для построения филогенетического дерева выглядит следующим образом:

s= fastaread(' Accession.fasta') – чтение биологических данных из файла в fasta формате

data(1).Sequence=s.Sequence(1:end)

data(1).Header='Accession'; – формирование массива данных для дальнейшей обработки

distances = **seqpdist**(data, ...'PropertyName', PropertyValue, ...) – расчет попарного расстояния между дистанциями

tree= seqlinkage(distances,'single',vir) – конструирование филогенетического дерева на основании попарной дистанции между последовательностями

h = plot(tree,'Type', 'radial','orient','left') – визуализация сконструированного объекта в виде филогенетического дерева

Для использования различных методов построения филогенетических деревьев изменяются входящие аргументы функции. Рассмотрим каждую из функций более подробно.

1. Чтение программой Matlab данных из файла в FASTA формате – функция fastaread

Синтаксис

[Header, Sequence] = fastaread(File)

Описание

Функция fastaread считывает данные из файла в FASTA формате в структуру MATLAB со следующими полями.

Header - Информация о последовательности (заголовок)

Sequence – Однобуквенная запись нуклеотидной последовательности, сохраняющейся как строка букв.

Пример использования функции чтения данных из файла в FASTA формате

1.1. Перед загрузкой данных в рабочую среду Matlab биологические последовательности, полученные из баз банных или других источников, сохраняют в файле с расширением **fasta** в отдельной рабочей директории. **Совет.** Для выполнения этого действия откройте программу «Блокнот» и вставьте выбранную последовательность в FASTA формате. Описание этого формата было приведено выше. Дайте имя файлу в соответствии с ACCESSION number вашей последовательности. Когда даете расширение при сохранении в «Типе файла» выберите «все файлы».

Используя функцию fastaread можно прочитать и загрузить эти данные для дальнейшего использования в программе Matlab .

```
s= fastaread('NC_010658.fasta');
```

В рабочей области видим формирование бинарного массива данных

```
s =

    Header: [1x127 char]
    Sequence: [1x1741 char]
```

1.2. Если у вас не очень много файлов и они не большие по размеру, то лучше сохранить их всех в одном файле с расширением **.fa.** Для создания такого исходного файла откройте программу «Блокнот» и вставьте последовательности в FASTA формате. Используя функцию fastaread можно прочитать и загрузить эти данные для дальнейшего использования в программе Matlab .

seqs = fastaread('pf01.fa');

В рабочей области видим формирование массива данных

seqs =

8x1 struct array with fields:
Header
Sequence

1.3. При наличии версии Matalb старше 8-ой, возможна загрузка данных непосредственно из базы данных, при наличии подключенного Интернета. Например, получение набора данных из GenBank. В данном примере использованы внутренние номера базы данных, соответствующие нуклеотидным последовательностям D-петли, изолированной из различных видов.

```
%       Species Description      GenBank Accession
data = {'German_Neanderthal'     'AF011222';
        'Russian_Neanderthal'    'AF254446';
        'European_Human'         'X90314'  ;
        'Mountain_Gorilla_Rwanda' 'AF089820';
        'Chimp_Troglodytes'      'AF176766';
        'Puti_Orangutan'         'AF451972';
        'Jari_Orangutan'         'AF451964';
        'Western_Lowland_Gorilla' 'AY079510';
        'Eastern_Lowland_Gorilla' 'AF050738';
        'Chimp_Schweinfurthii'   'AF176722';
        'Chimp_Vellerosus'       'AF315498';
        'Chimp_Verus'            'AF176731';
       };
for ind = 1:length(data)
   primates(ind).Header   = data{ind,1};
   primates(ind).Sequence = getgenbank(data{ind,2},'sequenceonly','true');
end
```

2. Расчет попарного расстояния между последовательностями - функция seqpdist

Синтаксис

D = seqpdist(*Seqs*)
D = seqpdist (Seqs, ...'Method', MethodValue, ...)

Описание

D = seqpdist (Seqs) возвращает вектор D, содержащий биологические расстояния между каждой парой последовательностей, которые хранились в массиве данных Seqs.

D состоит из 1-by-(M*(M-1)/2) строк, соответствующих M*(M-1)/2 парам последовательностей в массиве Seqs. Выходной параметр D ранжирован по следующему порядку ((2,1), (3,1),..., (M, 1), (3,2),...(M, 2),...(M, M-1)). Это левый нижний треугольник полной дистанционной матрицы. Чтобы получить расстояние между I-ой и J-ой последовательностями для I> J, необходимо использовать формулу расчёта расстояний D((J-1)*(M-J/2)+I-J).

D = seqpdist (Seqs, ...'Method', MethodValue, ...) указывается метод вычисления расстояний между каждой парой последовательностей. Варианты методов представлены в следующих таблицах 6 – 9.

Таблица 6

Опции, используемые для расчета попарного расстояния между последовательностями

Аргумент	Описание
MethodValue	Строка, которая определяет способ вычисления попарных расстояний. По умолчанию используется ‘Jukes-Cantor’.
IndelsValue	Строка, которая определяет, как рассматривать сайты с пробелами (gap). По умолчанию ‘score’.
PairwiseAlignmentValue	Управление глобальным попарным выравниванием входящих последовательностей, с игнорированием множественного выравнивания входящих последовательностей (если оно есть). Выбор определяется как истинный или ложный. По умолчанию: true - когда все входные последовательности не имеют одинаковую длину. false - Когда все входные последовательности имеют одинаковую длину.
AlphabetValue	Строка, описывающая тип последовательности (нуклеотидная или аминокислотная). Выбор – "NT" или "AA" (по умолчанию).
ScoringMatrixValue	Строка спецификации матрицы замен, для используемого выравнивания. Выбор для аминокислотных последовательностей: 'BLOSUM62' 'BLOSUM30' 'BLOSUM100' 'PAM10 '

	'DAYHOFF' 'GONNET' По умолчанию: 'BLOSUM50' – В случае параметра «AlphabetValue» – "AA" 'NUC44' - В случае параметра «AlphabetValue» – 'NT'

Таблица 7

Методы, используемые для нуклеотидных и аминокислотных последовательностей

Метод	Описание
p-distance	Количественное соотношение различающихся сайтов в двух последовательностях.
Jukes-Cantor (default)	Оценка числа замен в двух последовательностях методом максимального правдоподобия. Для нуклеотидных последовательностей: d = -3/4 log(1-p * 4/3) Для аминокислотных последовательностей: d = -19/20 log(1-p * 20/19)
alignment-score	Расстояние (D) между двумя последовательности (1, 2) вычисляется исходя из веса попарного выравнивания двух последовательностей (score12), и оценки попарного соответствия каждой последовательности самой себе (score11, score22) следующим образом: D = (1-score12/score11) * (1-score12/score22) Это расстояние метод не соответствует ультраметрическим условиям. В редких случаях, когда оценка веса выравнивания между последовательностями больше, чем оценка при выравнивании последовательности с самой собой, то D = 0

Таблица 8

Методы, используемые только для нуклеотидных последовательностей

Метод	Описание
Kimura	Отдельно рассчитывается количество транзиций и трансверсий
Tamura	Отдельно рассчитывается количество транзиций и трансверсий , а также содержание GC пар.

Таблица 9

Методы, используемые только для аминокислотных последовательностей

Метод	Описание
Poisson	Расчёт ведется, основываясь на предположении, что число аминокислотных замен на каждый сайт соответствует распределению Пуассона
Gamma	Расчёт ведется, основываясь на предположении, что число аминокислотных замен на каждый сайт соответствует Гамма распределению с заданным параметром α.

Примеры использования функции расчета попарного расстояния между последовательностями:

2.1. Используем массив данных seqs, содержащий выбранные для анализа биологические последовательности. Произведем расчет попарного расстояния, между последовательностями используя метод 'Jukes-Cantor', учитываем сайты с пробелами и игнорируя сайты с гепами (gap) 'pairwise-delete'. Возможно преобразование входящих данных в квадратную матрицу 'squareform',true.

```
dist = seqpdist (seqs,'method','jukes-cantor','indels','pairwise-
delete','squareform',true)
```

2.2. Используем массив данных seqs, содержащий выбранные для анализа биологические последовательности. Произведем расчет попарного расстояния, между последовательностями используя метод Jukes-Cantor и используя строку указания типа последовательностями 'Alpha','DNA'

```
distances = seqpdist(seqs,'Method','Jukes-Cantor','Alpha','DNA');
```

2.3. Используем массив данных seqs, содержащий выбранные для анализа биологические последовательности. Произведем расчет попарного расстояния, между последовательностями используя метод Jukes-Cantor с учетом точечных мутаций во множественном выравнивании

```
dist = seqpdist(seqs,'Method','jukes-cantor',….
 'Indels',  'score','PairwiseAlignment', true );
```

3. Конструирование филогенетического дерева на основании попарной дистанции между последовательностями – функция seqlinkage

Синтаксис

Tree = seqlinkage(Dist)

Tree = seqlinkage(Dist, Method)

Описание

Tree = seqlinkage(Dist**)** возвращает филогенетическое дерево объектов из попарных расстояний (Dist) между видами. Dist является матрицей или вектором попарных расстояний, рассчитанных с помощью функции seqpdist.

Tree = seqlinkage(Dist, Method) создает филогенетическое дерево объекта, используя указанный метод расчета расстояния.

Таблица 10

Доступные методы

Метод	Описание
'single'	Метод ближайшего расстояния (одиночная связь)
'complete'	Метод полной связи
'average' (default)	Unweighted Pair Group Method Average (UPGMA)
'weighted'	Weighted Pair Group Method Average (WPGMA)
'centroid'	Unweighted Pair Group Method Centroid (UPGMC)
'median'	Weighted Pair Group Method Centroid (WPGMC)

Пример использования функции конструирование филогенетического дерева на основании попарной дистанции

Загрузим данные и создадим массив data для дальнейшей обработки. Используем метод 'Jukes-Cantor' для оценки попарного расстояния между последовательностями. Сконструируем на основе полученных данных филогенетическое дерево. Визуализируем его с помощью функции plot.

Очередность написания команд:

distances = seqpdist(data,'Method','Jukes-Cantor','Alpha','DNA');

UPGMAtree = seqlinkage(distances,'UPGMA',data)

```
h = plot(UPGMAtree,'orient','left');
title('UPGMA Distance Tree of data using Jukes-Cantor model');
xlabel('Evolutionary distance')
```

При наличии версии Matalb старше 8-ой, возможно построение филогенетического дерева методом «ближайших соседей» (Neighbor-Joining), с использованием функции seqneighjoin. Этот метод рассчитывает попарные расстояния с использованием метода максимальной экономии. (maximal parsimony).

```
NJtree = seqneighjoin(distances,'equivar',data)
h = plot(NJtree,'orient','left');
title('Neighbor-Joining Distance Tree of data using Jukes-Cantor model');
xlabel('Evolutionary distance')
```

4. Визуализация сконструированного объекта в виде филогенетического дерева с помощью функции plot

Синтаксис

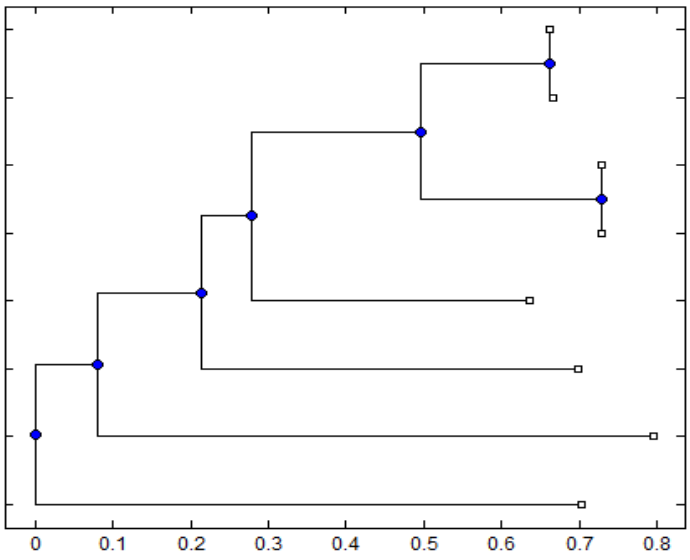
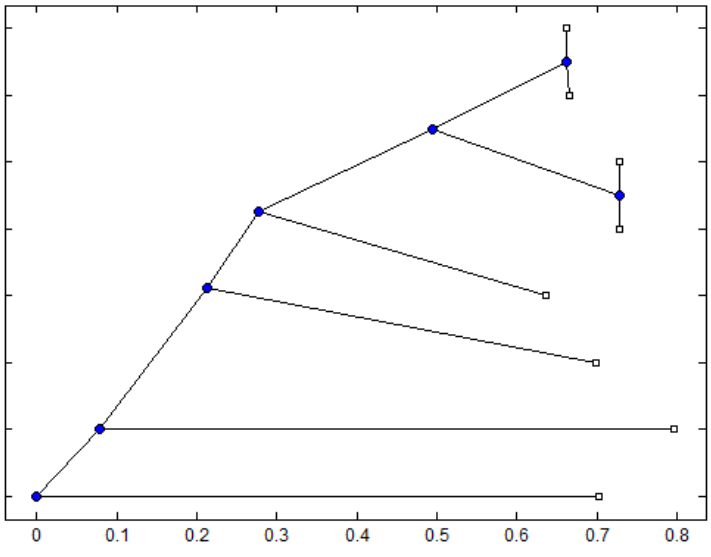
```
plot(Tree)
H = plot(...)
plot(..., 'Orientation', OrientationValue, ...)
```

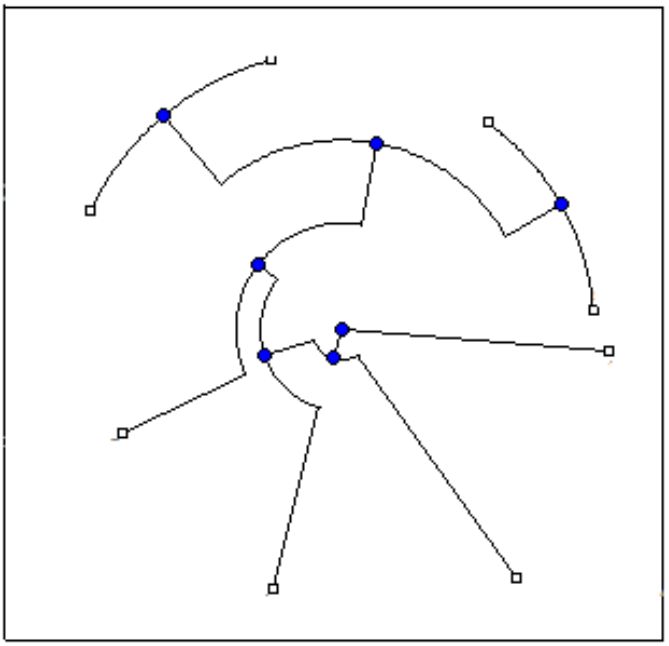
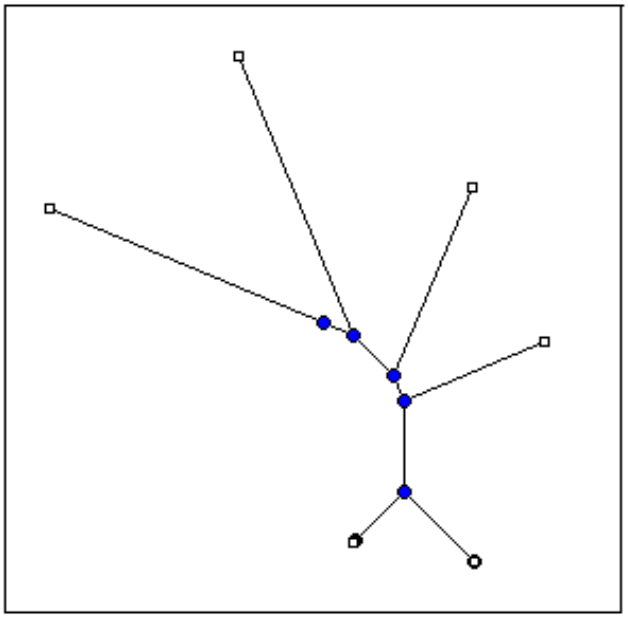
Описание

plot(Tree) преобразует объект в фигуру типа филограмма или кладограмма. Значимые значения узлов филогенетического дерева, отображены в горизонтальном направлении. Вертикальные расстояния являются произвольными и не имеют никакого значения.

plot(..., 'Type', TypeValue, ...) определяет метод для визуализации филогенетического дерева. TypeValue - строка, определяющая способ для рисования филогенетического дерева. Варианты топологии филогенетических деревьев представлены на рисунке 38.

Доступные методы:

Метод	Внешний вид
'square' (default)	
'angular'	

'radial'	
'equalangle'	

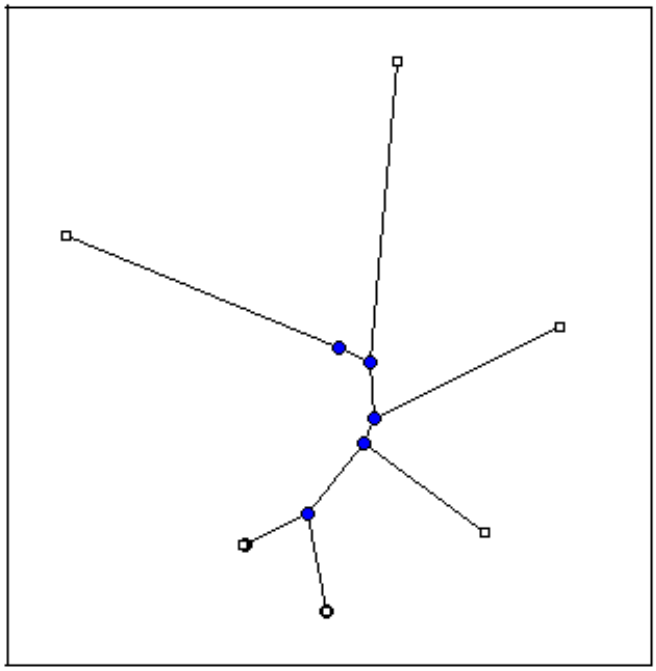
'equaldaylight’	

Рис. 38. Топология филогенетических деревьев, выполненных в программе Matlab с использованием опции TypeValue.

plot(..., 'Orientation', OrientationValue, ...) определяет ориентацию корневого узла, и, следовательно, ориентация филограммы или кладограммы на рисунке, когда выбран тип построение 'square' или 'angular'. Доступные атрибуты:

'left' (default) - ориентация слева на право;

'right' – ориентация справа налево;

'top' - ориентация сверху вниз;

'bottom' – ориентация снизу вверх.

Варианты написания полной программы

Программа 1.

```
seqs = fastaread('pf01.fa')
distances = seqpdist(seqs,'Method','Jukes-Cantor','Alpha','DNA');
tree = seqlinkage(distances,'UPGMA',seqs);
h = plot(tree,'orient','left');
title('Neighbor-Joining Distance Tree of pyridoxine kinase using Jukes-Cantor
model');
xlabel('Evolutionary distance')
```

Результат представлен на рисунке 39

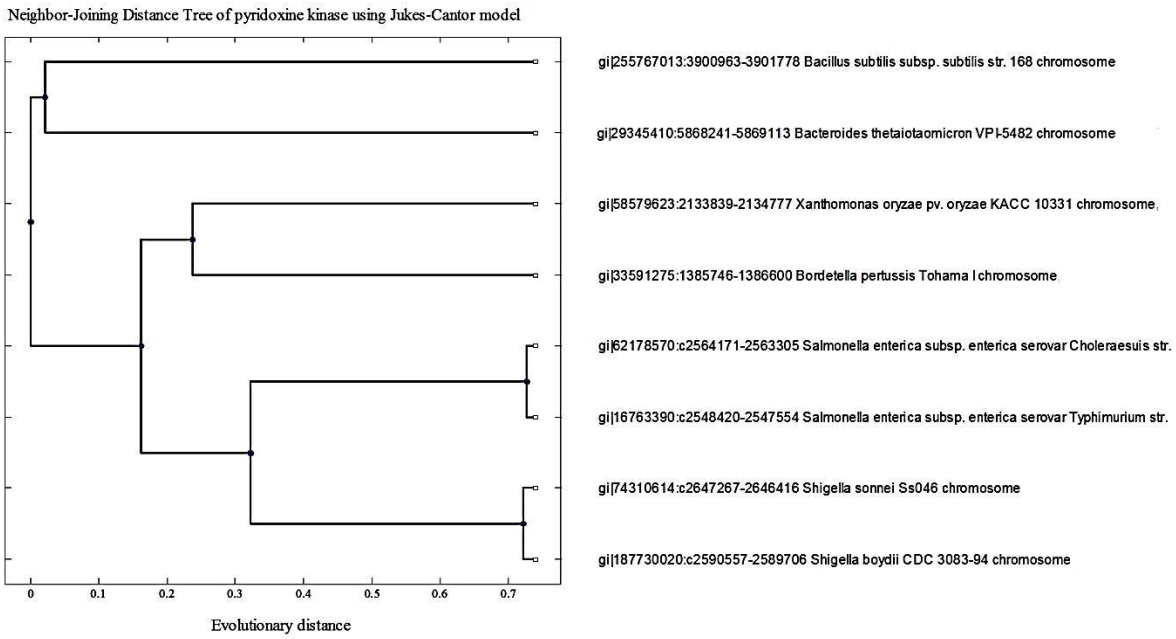


Рис. 39. Филогенетическое дерево изученных последовательностей, по результатам выполнения программы 1.

Программа 2

```
seqs = fastaread('pf01.fa')
distances = seqpdist(seqs,'method','jukes-cantor','indels','pairwise-
delete','squareform',true);
NJtree = seqneighjoin(distances,'equivar',seqs);
h = plot(NJtree,'orient','left');
title('Neighbor-Joining Distance Tree of pyridoxine kinase using Jukes-Cantor
model');
xlabel('Evolutionary distance')
```

Результат представлен на рисунке 40

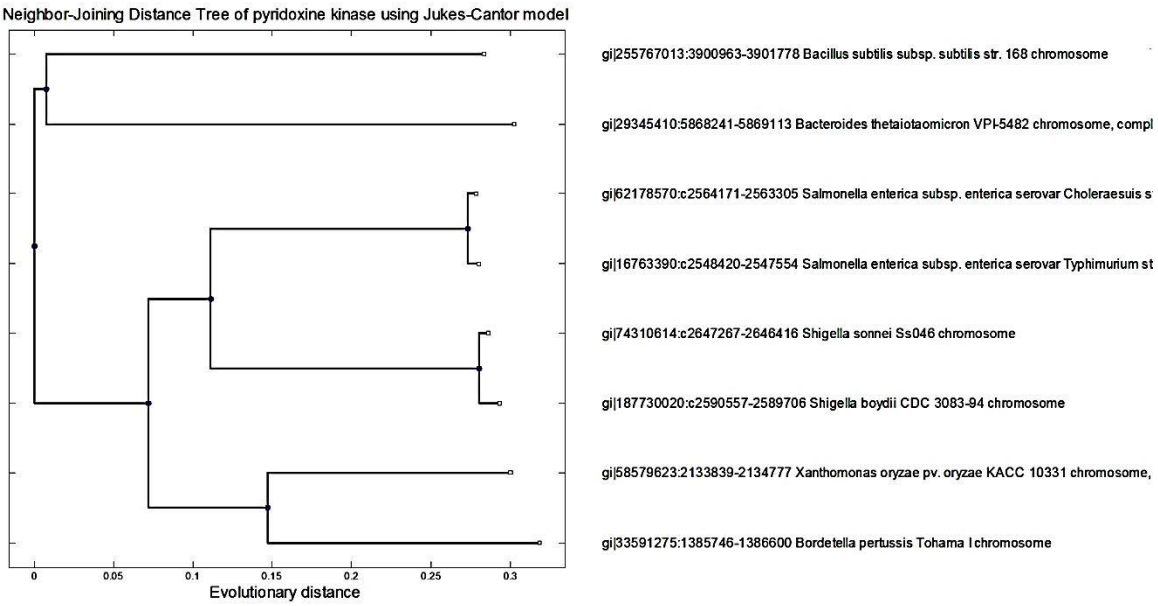


Рис. 40. Филогенетическое дерево изученных последовательностей, по результатам выполнения программы 2

Программа 3

```
seqs = fastaread('pf01.fa')
dist = seqpdist(seqs,'Method','jukes-cantor', 'Indels', 'score','PairwiseAlignment', true
);
tree = seqlinkage(dist,'UPGMA',seqs);
h = plot(tree,'orient','left');
title('Neighbor-Joining Distance Tree of pyridoxine kinase using Jukes-Cantor
model');
xlabel('Evolutionary distance')
```

Результат представлен на рисунке 41

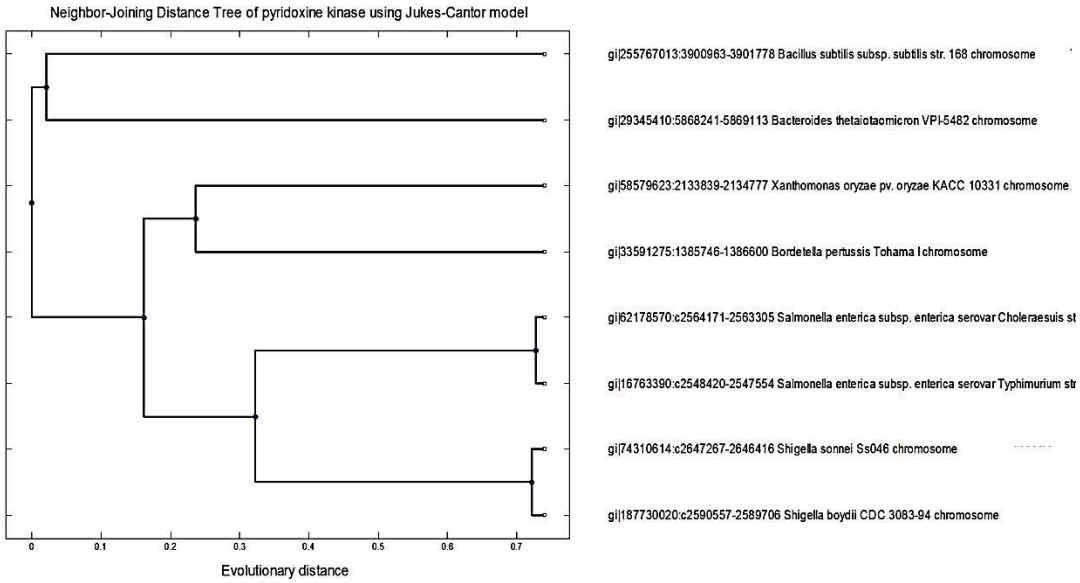


Рис. 4.1 Филогенетическое дерево изученных последовательностей, по результатам выполнения программы 3

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое биоинформатика?
2. Какую дату можно считать датой выделения биоинформатики в отдельную научную область?
3. Где хранятся биоинформационные данные?
4. Каковы цели биоинформатики?
5. Какие задачи стоят перед биоинформатикой?
6. Какие биологические последовательности называются гомологичными?
7. Какую роль играет анализ гомологических последовательностей в расшифровке биологической информации?
8. Какие типы выравнивания существуют?
9. Что называется выравниванием биологических последовательностей?
10. Что такое глобальное выравнивание последовательностей?
11. Что такое локальное выравнивание последовательностей?
12. Что такое множественное выравнивание?
13. Что такое оптимальное выравнивание?
14. Для какого типа биологических последовательностей используются матрицы PAM, BLOSUM и Gonnet?
15. С какой целью проводится выравнивание?
16. Какие программы используются для множественного выравнивания?
17. Дайте характеристику формату FASTA? Для чего он используется?
18. Какая команда Matlab используется для множественного выравнивания?
19. Какая команда Matlab используется для чтения FASTA формата?
20. На основании какой процедуры строится множественное выравнивание?
21. Почему с точки зрения биолога локальное выравнивание может дать более значимые и точные результаты, чем глобальное выравнивания?
22. Какими символами в окне результатов программы ClustalW2 обозначаются: одинаковая аминокислота; сходные аминокислоты; вставки; отсутствие сходства в последовательностях?
23. Что такое матрица PAM?
24. Что такое филогения?
25. Что называется кластеризацией?
26. В чём суть метода наибольшей экономии?
27. В чём суть метода наибольшего правдоподобия?
28. Какая команда Matlab расчитывает попарное расстояние между последовательностями?
29. Какая команда Matlab конструирует филогенетическое дерево?
30. На основе какого принципа конструируется филогенетическое дерево в Matlab?

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Бородовский М. Задачи и решения по анализу биологических последовательностей / М. Бородовский, С. Екишева. – М.-Ижевск: РХД, 2008. – 440 с.
2. Дурбин Р. Анализ биологических последовательностей / Р. Дурбин, Ш. Эдди, А. Крог, Г. Митчисон. – М.-Ижевск : РХД, 2006. – 480 с.
3. Каменская М.А. Информационная биология / М. А. Каменская. – М.: Академия, 2006. – 368 с.
4. Огурцов А.Н. Методы биоинформационного анализа / А.Н. Огурцов. – Х.: НТУ "ХПИ", 2011. – 114 с.
5. Основы биоинформатики : учеб. пособие / А. Н. Огурцов. – Х.: НТУ «ХПИ», 2013. – 400 с.

Навчальне видання

Васильева Наталія Юріївна

Біоінформатика

Множинне вирівнювання. Філогенетичні дерева

Методичний посібник

(російською мовою)

За редакцією автора

Дизайн обкладинки – М. Чабан, Н. Коротаєва

Підп. до друку 27.10.2014. Формат 60х84/8.
Ум.-друк. арк. 8,14. Тираж 50.
Зам. № 955.

Видавець і виготовлювач
Одеський національний університет імені І. І. Мечникова
Свідоцтво суб'єкта видавничої справи ДК № 4215 від 22.11.2011 р.

Україна, 65082, м. Одеса, вул. Єлісаветинська, 12
Тел. (048) 723-28-39. E-mail: druk@onu.edu.ua