

Санкт-Петербургский политехнический университет Петра Великого

Институт прикладной математики и механики

Высшая школа прикладной математики и вычислительной физики

Биоинформатика:
Отчёт по практической работе

Выполнил студент:
Курова Анна Николаевна
группа: 3630102/70401

Санкт-Петербург 2020

Содержание

1	Выбор гена	4
2	Нуклеотидная и белковая последовательности	4
3	TBLASTN	5
4	Анализ	7
5	Филогенетические деревья	7
6	Tanglegrams	8
7	Матрица расстояний и Neighbor joining	9
8	Приложения	9

Список иллюстраций

1	RAB20 info	4
2	RAB20 proteins	4
3	Часть файла RAB20 transcripts	5
4	Часть файла выравнивания протеинов	6
5	Часть файла выравнивания нуклеотидов	6
6	Участок с неконсервативными заменами	7
7	Филогенетическое дерево variable1_transcripts	8
8	Tanglegram transcripts_aligned and variable3	8
9	Neighbor joining для последовательностей	9

1 Выбор гена

Исследуемый ген: RAB20, member RAS oncogene family

Электронно-микроскопические исследования показали, что RAB20 расположен в апикальных плотных канальцах, эндоцитарных структурах, лежащих под апикальной плазматической мембраной, что позволяет предположить, что он играет роль в апикальном эндоцитозе/ресайклинге.

Масса (kDa): 25.989 kDa

Специфичность ткани: Присутствует в различных тканях, но не в головном мозге

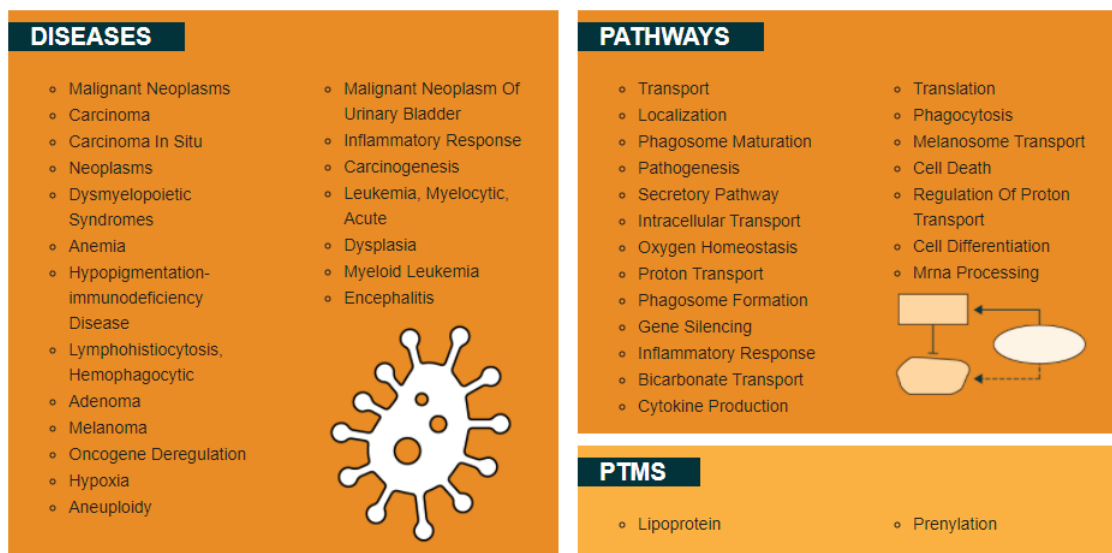


Рис. 1: RAB20 info

2 Нуклеотидная и белковая последовательности

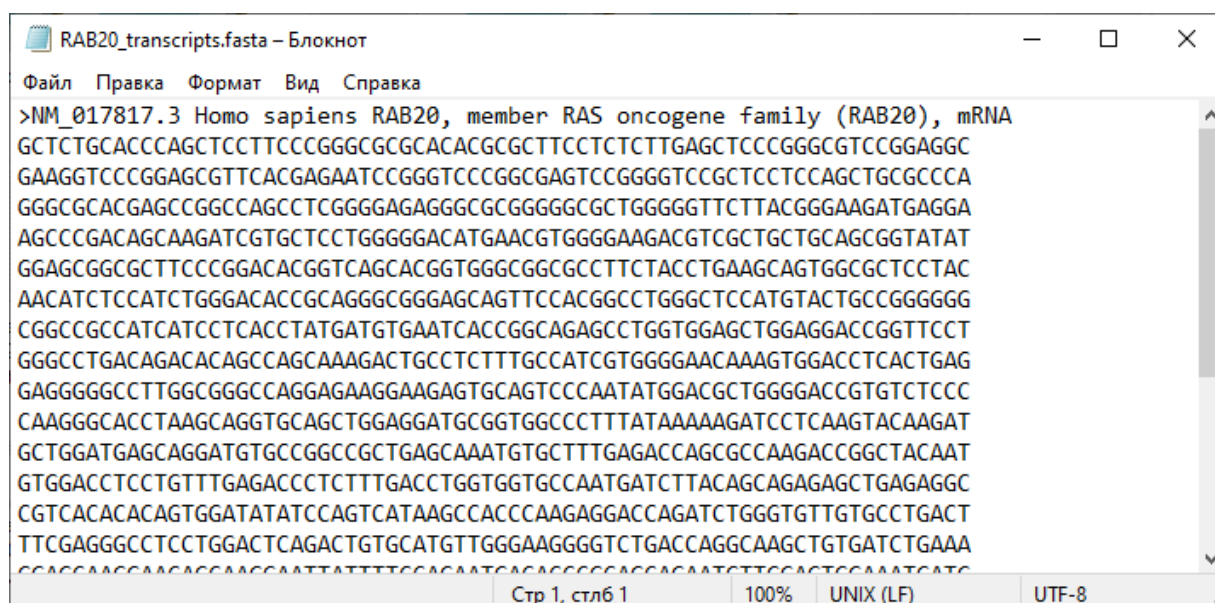
С сайта NCBI были скачены обе последовательности.

The screenshot shows a text editor window titled "RAB20_protein.fasta - Блокнот". The menu bar includes "Файл", "Правка", "Формат", "Вид", and "Справка". The text content is as follows:

```
>NP_060287.1 ras-related protein Rab-20 [Homo sapiens]
MRKPDSKIVLLGDMNVGKTSLLQRYMERRFPDVTSTVGGAFFLKQWRSYNISIWDTAGREQFHGLGSMYC
RGAAAILTYDVNHRQSLVELEDRFLGLTDTASKDCLFAIVGNKVDLTEEGALAGQEKEECSPNMDAGDR
VSPRAPKQVQLEDAVALYKKILKYKMLDEQDVPAAEQMCFETSAKTGYNVDLLFETLFDLVVPMILQQRA
ERPSHTVDISSHKPPKRTRSGCCA
```

The status bar at the bottom indicates "Стр 1, столб 1", "100%", "UNIX (LF)", and "UTF-8".

Рис. 2: RAB20 proteins



```
>NM_017817.3 Homo sapiens RAB20, member RAS oncogene family (RAB20), mRNA
GCTCTGCACCCAGCTCCTCCCGGGCGCGCACACGCGCTTCTCTCTTGAGCTCCCGGGCGTCCGGAGGC
GAAGGTCCCGGAGCGTTACAGAGAATCCGGGTCCCGGCGAGTCCGGGGTCCGCTCCTCCAGCTGCGCCCA
GGGCGCACGAGCCGGCCAGCCTCGGGGAGAGGGCGCGGGGGCGCTGGGGGTTCTTACGGGAAGATGAGGA
AGCCCGACAGCAAGATCGTGCTCCTGGGGGACATGAACGTGGGGAAGACGTCGCTGCTGCAGCGGTATAT
GGAGCGGCGCTTCCCGGACACGGTCAGCACGGTGGGCGGCGCCTTCTACCTGAAGCAGTGGCGCTCCTAC
AACATCTCCATCTGGGACACCGCAGGGCGGGAGCAGTTCACGGCCTGGGCTCCATGTAAGTCCGGGGGG
CGGCCGCCATCATCCTCACCTATGATGTGAATCACCGGCAGAGCCTGGTGGAGCTGGAGGACCGGTTCTT
GGGCTGACAGACACAGCCAGCAAAGACTGCCTCTTTGCCATCGTGGGGAACAAAGTGGACCTCACTGAG
GAGGGGGCCTTGGCGGGCCAGGAGAAGGAAGAGTGCAGTCCCAATATGGACGCTGGGGACCGTGTCTCCC
CAAGGGCACCTAAGCAGGTGCAGCTGGAGGATGCGGTGGCCCTTTATAAAAAGATCCTCAAGTACAAGAT
GCTGGATGAGCAGGATGTGCCGGCCGCTGAGCAAATGTGCTTTGAGACCAGCGCCAAGACCGGCTACAAT
GTGGACCTCCTGTTTGAGACCCTCTTTGACCTGGTGGTGCCAATGATCTTACAGCAGAGAGCTGAGAGGC
CGTCACACACAGTGGATATATCCAGTCATAAGCCACCCAAGAGGACCAGATCTGGGTGTTGTGCCTGACT
TTCGAGGGCCTCCTGGACTCAGACTGTGCATGTTGGGAAGGGGTCTGACCAGGCAAGCTGTGATCTGAAA
CCAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAG
```

Рис. 3: Часть файла RAB20 transcripts

3 TBLASTN

Список организмов:

1. Pan troglodytes
2. Pan paniscus
3. Macaca mulatta
4. Mus musculus
5. Rattus norvegicus
6. Canis familiaris
7. Felis catus
8. Bos taurus

Для каждого организма из списка был проведён BLAST, доступный на сайте NCBI:

- Query Sequence: RAB20 protein
- Search Sequence: Species genome

Гомологи с минимальным e value (в соответствии со списком):

1. PREDICTED: Pan troglodytes RAB20, member RAS oncogene family (RAB20), mRNA
2. PREDICTED: Pan paniscus RAB20, member RAS oncogene family (RAB20), mRNA
3. PREDICTED: Macaca mulatta RAB20, member RAS oncogene family (RAB20), mRNA
4. M.musculus rab20 mRNA

5. Rattus norvegicus RAB20, member RAS oncogene family (Rab20), mRNA
6. PREDICTED: Canis lupus familiaris RAB20, member RAS oncogene family (RAB20), mRNA
7. PREDICTED: Felis catus RAB20, member RAS oncogene family (RAB20), mRNA
8. Bos taurus RAB20, member RAS oncogene family (RAB20), mRNA

Для каждого экземпляра были скачены fasta файлы кодирующих аминокислот и убраны стоп кодоны (иначе MEGA-X не сможет выполнить выравнивание). В отдельный файл all_seq_proteins.fasta помещены нуклеотиды 8 видов и RAB20. В MEGA-X были выполнены выравнивания.

В результате получаем 2 файла:

1. transcripts_aligned.fas
2. proteins_aligned.fas

DNA Sequences	Translated Protein Sequences
Species/Abbrv	* * * * *
1. IctNM 017817.3 cds NP 060287.1 1 gene=RAB20 db xref=CCDS:CCDS9512.1 protein=ras-related protein Rab-20 protein id=N	M R K P D S K I V L L G D M N V G K T S L L Q R Y M E R R F P
2. IctNM 001109535.1 cds NP 001103005.1 1 gene=Rab20 db xref=GeneID:689377 RGD:1593487 protein=ras-related protein Ra	M R K P D G K I V L L G D M N V G K T S L L Q R Y M E R R F P
3. IctX80332.1 cds CAA56582.1 1 gene=rab20 db xref=GOA:P35295 InterPro:IPR001806 InterPro:IPR005225 InterPro:IPR013753	M R K P D G K I V L L G D M N V G K T S L L Q R Y M E R R F P
4. IctXM 003832084.2 cds XP 003832132.1 1 gene=RAB20 db xref=GeneID:100977562 protein=ras-related protein Rab-20 prot	M R K P D S K I V L L G D M N V G K T S L L Q R Y M E R R F P
5. IctXM 024348343.1 cds XP 024204111.1 1 gene=RAB20 db xref=GeneID:112205216 protein=ras-related protein Rab-20 prot	M R K P D S K I V L L G D M N V G K T S L L Q R Y M E R R F P
6. IctXM 001082373.4 cds XP 001082373.1 1 gene=RAB20 db xref=GeneID:694578 protein=ras-related protein Rab-20 protein i	M R K P D S K I V L L G D M N V G K T S L L Q R Y M E R R F P
7. IctXM 542668.6 cds XP 542668.2 1 gene=RAB20 db xref=GeneID:485549 protein=ras-related protein Rab-20 protein id=XP 54	M R K P D G K I V L L G D M N V G K T S L L Q R Y M E R R F P
8. IctXM 003980528.5 cds XP 003980577.1 1 gene=RAB20 db xref=GeneID:101089416 protein=ras-related protein Rab-20 prot	M R K P D G K I V L L G D M N V G K T S L L Q R Y M E R R F P
9. IctNM 001193089.1 cds NP 001180018.1 1 gene=RAB20 db xref=BGD:BT12747 GeneID:615760 VGNC:VGNC:33627 protein=M	M R K P D G K I V L L G D M N V G K T S L L Q R Y M E R R F P

Рис. 4: Часть файла выравнивания протеинов

DNA Sequences	Translated Protein Sequences
Species/Abbrv	* * * * *
1. IctNM 017817.3 cds NP 060287.1 1 gene=RAB20 db xref=CCDS:CCDS9512.1 protein=ras-related protein Rab-20 protein id=N	A T G A G G A A G C C C G A C A G C A A G A T C G T G C T C C T T
2. IctNM 001109535.1 cds NP 001103005.1 1 gene=Rab20 db xref=GeneID:689377 RGD:1593487 protein=ras-related protein Ra	A T G C G G A A G C C C G A T G G G A A G A T C G T G C T G C T T
3. IctX80332.1 cds CAA56582.1 1 gene=rab20 db xref=GOA:P35295 InterPro:IPR001806 InterPro:IPR005225 InterPro:IPR013753	A T G C G G A A G C C C G A T G G G A A G A T C G T G C T G C T T
4. IctXM 003832084.2 cds XP 003832132.1 1 gene=RAB20 db xref=GeneID:100977562 protein=ras-related protein Rab-20 prot	A T G A G G A A G C C C G A C A G C A A G A T C G T G C T C C T T
5. IctXM 024348343.1 cds XP 024204111.1 1 gene=RAB20 db xref=GeneID:112205216 protein=ras-related protein Rab-20 prot	A T G A G G A A G C C C G A C A G C A A G A T C G T G C T C C T T
6. IctXM 001082373.4 cds XP 001082373.1 1 gene=RAB20 db xref=GeneID:694578 protein=ras-related protein Rab-20 protein i	A T G A G G A A G C C C G A C A G C A A G A T T G T G C T G C T T
7. IctXM 542668.6 cds XP 542668.2 1 gene=RAB20 db xref=GeneID:485549 protein=ras-related protein Rab-20 protein id=XP 54	A T G A G G A A G C C C G A C G G G A A G A T C G T G C T C C T T
8. IctXM 003980528.5 cds XP 003980577.1 1 gene=RAB20 db xref=GeneID:101089416 protein=ras-related protein Rab-20 prot	A T G A G G A A G C C C G A C G G G A A G A T C G T G C T T T T T
9. IctNM 001193089.1 cds NP 001180018.1 1 gene=RAB20 db xref=BGD:BT12747 GeneID:615760 VGNC:VGNC:33627 protein=A	A T G A G G A A G C C C G A C G G G A A G A T C G T G C T T T T T

Рис. 5: Часть файла выравнивания нуклеотидов

4 Анализ

- * - консервативные сайты
- : - сайты с консервативными заменами
- . - сайты с полуконсервативными заменами
- () пустая ячейка - сайты с неконсервативными заменами

Рассмотрим полученное выравнивание протеинов.

В основном доминируют именно консервативные сайты. Но если рассмотреть участок на рис. 6 можно заметить, что начинается участок, где присутствуют только сайты с неконсервативными заменами.

DNA Sequences	Translated Protein Sequences
Species/Abbrev	
1. IctNM 017817.3 cds NP 060287.1 1 gene=RAB20 db xref=CCDS:CCDS9512.1 protein=ras-related protein Rab-20 protein id=NP 060287.1 location=204..908 gbkey=CDS	K V D L T E E G A L A G Q E K E E C S P N M D
2. IctNM 001109535.1 cds NP 001103005.1 1 gene=Rab20 db xref=GeneID:689377 RGD:1593487 protein=ras-related protein Rab-20 protein id=NP 001103005.1 location=	K V D L T T E R G P E G G E K D Q A S G K T G
3. IctX80332.1 cds CAA56582.1 1 gene=rab20 db xref=GOA:P35295 InterPro:IPR001806 InterPro:IPR005225 InterPro:IPR013753 InterPro:IPR020851 MGI:MGI:102789 UniProt	K V D L T S E R D T E G G E K E G P A S G K V
4. IctXM 003832084.2 cds XP 003832132.1 1 gene=RAB20 db xref=GeneID:100977562 protein=ras-related protein Rab-20 protein id=XP 003832132.1 location=237..941	K V D L T E E G A L A G Q E K E E C S P N M D
5. IctXM 024348343.1 cds XP 024204111.1 1 gene=RAB20 db xref=GeneID:112205216 protein=ras-related protein Rab-20 protein id=XP 024204111.1 location=271..975	K V D L T E E G A L A G Q E K E E C S P N M D
6. IctXM 001082373.4 cds XP 001082373.1 1 gene=RAB20 db xref=GeneID:694578 protein=ras-related protein Rab-20 protein id=XP 001082373.1 location=539..1243 gb	K V D L T E E G A L A G Q E K E K C S P D T D
7. IctXM 542668.6 cds XP 542668.2 1 gene=RAB20 db xref=GeneID:485549 protein=ras-related protein Rab-20 protein id=XP 542668.2 location=255..962 gbkey=CDS	K V D L L E D G T A D S G E K E G L G P G V A
8. IctXM 003980528.5 cds XP 003980577.1 1 gene=RAB20 db xref=GeneID:101089416 protein=ras-related protein Rab-20 protein id=XP 003980577.1 location=329..1033	K V D L V E E P A A E D Q R K D G R G P G V A
9. IctNM 001193089.1 cds NP 001180018.1 1 gene=RAB20 db xref=BGD:BT12747 GeneID:615760 VGNC:VGNC.33627 protein=ras-related protein Rab-20 protein id=NP	K V D L S E E A P G E G G Q G G R D P G Q A

Рис. 6: Участок с неконсервативными заменами

В файлах variable* находятся блоки с вариабельными участками. Это такие блоки, где присутствуют сайты с консервативными заменами или сайты с полуконсервативными заменами или сайты с неконсервативными заменами.

5 Филогенетические деревья

Добавим в файл с кодирующими протеинами аутгруппу. Возьмем *Amphiprion* (bony fishes). Построим выравнивание для 10 организмов.

С помощью метода в RAxML построим филогенетические деревья для выравненных последовательностей и для вариабельных участков (всего 8 деревьев).

Следующие параметры были использованы:

- Protein Model: GAMMA BLOSUM62 (для протеинов) и GAMMA (для нуклеотидов)
- Algorithm: Rapid hill-climbing
- Number of starting trees or bootstraps replicates: 1
(когда нет начального дерева, это число независимых запусков на различных начальных деревьях)
- Parsimony random seed: 1
(начальное число случайных чисел для вывода результата)

Деревья сохранены в файлы .newick, чтобы использовать их для построения tanglegrams в дальнейшем

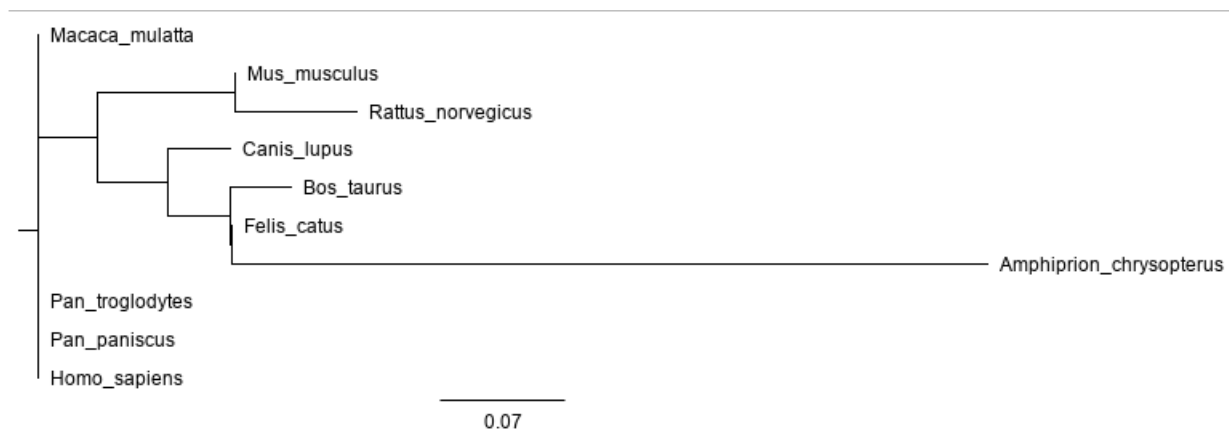


Рис. 7: Филогенетическое дерево variable1_transcripts

6 Tanglegrams

Построим tanglegrams между деревьями для каждого полного выравнивания и деревьями его фрагментов ($2 \times 3 = 6$ tanglegrams).

Для этого была использована программа Dendroscope. Результаты сохранены в файлы .newick с названиями деревьев для которых был проведён алгоритм.

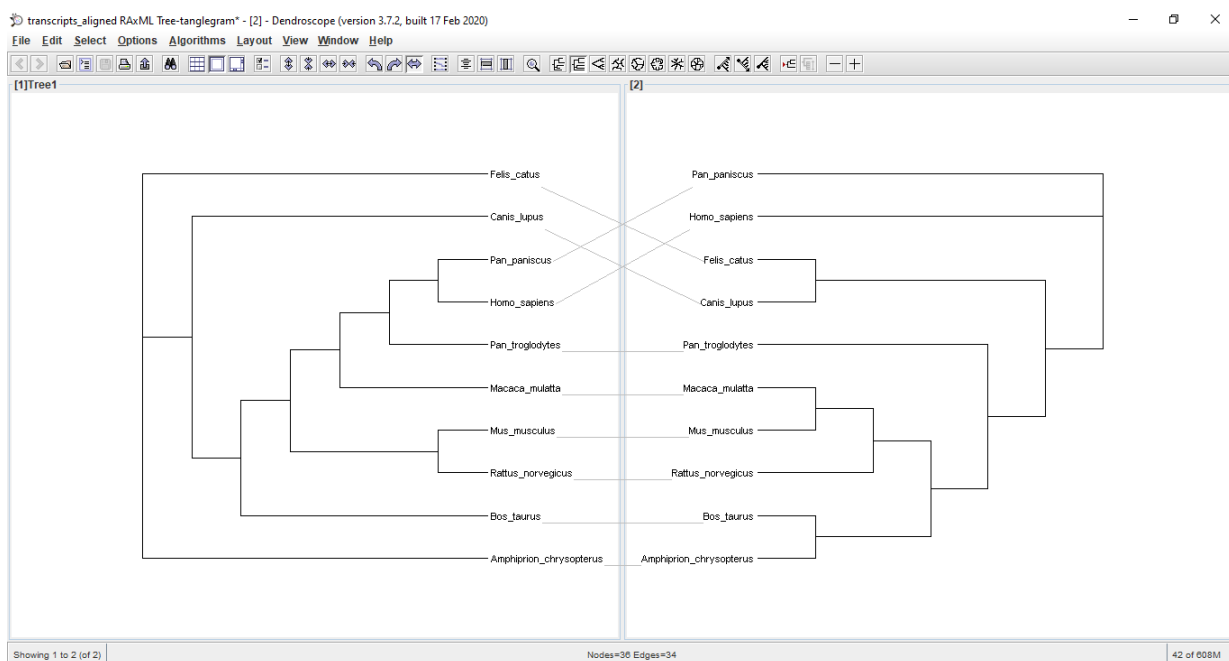


Рис. 8: Tanglegram transcripts_aligned and variable3

7 Матрица расстояний и Neighbor joining

Строить матрицу и neighbor joining будем в R студии с помощью библиотеки `ape`. Нам понадобится функции `dist.toro` и `nj`. Код и html программы в приложениях.

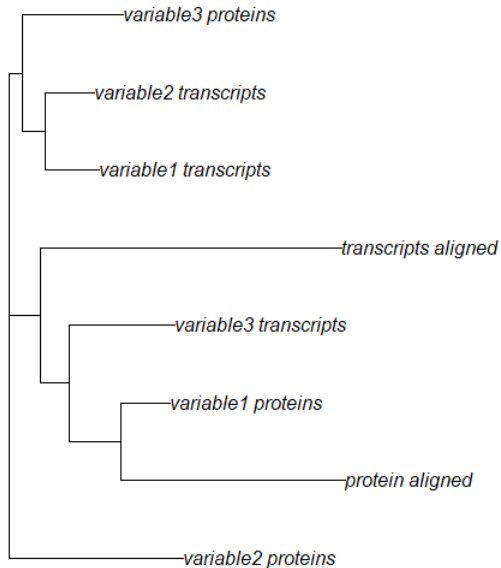


Рис. 9: Neighbor joining для последовательностей

8 Приложения

Код программы и все полученные файлы:
<https://github.com/katonapng/Bioinformatics>

Список литературы

- [1] Ras-Related Protein Rab-20 (Rab20) <https://www.bosterbio.com/bosterbio-gene-info-cards/RAB20>