

Level 4

Выбор гена

выбор пал на [ENSG00000196805](#)

Аминокислотная последовательность

Воспользовавшись [e!Ensembl](#) быстро и непринужденно узнаем что белок выглядит следующим образом

```
MSYQQQCKQPCQPPVCPTRKCPERCPPPKCPERCPPPKCPQPCPPQCCQKYPVTPSPPCQPKYPPKSK*
```

Имя ему [SPRR2B](#), так и назовём файл

Согласно Яндекс нейро

Предполагают, что ген участвует в дифференцировке кератиноцитов

Добавление организмов

После продолжительного повторения итерации ЛКМ -> Ctrl+C -> ЛКМ -> Ctrl+V -> ЛКМ, получаем красивое

Drosophila melanogaster (taxid:7227)
Danio rerio (taxid:7955)
Silurana (Xenopus) tropicalis (taxid:8364)
Crocodylus porosus (taxid:8502)
Ornithorhynchus anatinus (taxid:9258)
Monodelphis domestica (taxid:13616)
Canis lupus familiaris (taxid:9615)
Bos taurus (taxid:9913)
Castor canadensis (taxid:51338)
Pan troglodytes (taxid:9598)
Macaca mulatta (taxid:9544)

BLAST нажат, иду за чаем, как и приказано

P.S. [бобра](#) оценил

Результаты(не результаты)

Пама-пам-пам, с первого раза и не могло не получиться



No significant similarity found. For reasons why, [click here](#)

есть предположение что пробелма в том что при парсинге пептидной последовательности я указал датасет человеческих генов, а надо что-то более общее

нет, если взять датасет мышей, вообще ничего не находит значит нужен точно человеческий

побуем не мелочиться и взять все [гены](#), а дальше уже пробовать на каждом

не знаю, кстати, почему но e!Ensembl разучился загружать fasta файл, и возвращает мне txt при выбраном fasta

Рил - результаты

и вау вау вау, что-то мы да получили

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	PREDICTED: Pan troglodytes histatin 3 (HTN3). mRNA	Pan troglodytes	76.3	76.3	100%	8e-17	78.85%	533	XM_003950334.3
<input checked="" type="checkbox"/>	PREDICTED: Pan troglodytes histatin 1 (HTN1). transcript variant X1. mRNA	Pan troglodytes	60.5	60.5	100%	1e-10	63.79%	597	XM_001143678.6
<input checked="" type="checkbox"/>	PREDICTED: Pan troglodytes histatin 1 (HTN1). transcript variant X2. mRNA	Pan troglodytes	52.4	52.4	100%	1e-07	64.71%	576	XM_054683968.1

но это всё не то, очевидно одного бедного шимпанзе не хватит для построения филогенетического дерева

будем искать и пробовать пока не найдётся что-то поинтереснее



и вот всего ничего, проканал прекрасный **ONECUT3**

и вновь справка от яндекс нейро:

Ген ONECUT3 (One Cut Homeobox 3) участвует в регуляции различных процессов, среди которых:

- Развитие внутрипечёночных желчевыводящих путей у рыбок данио.
- Развитие бета-клеток и нервной системы.
- Метаболизм простаноидов и, возможно, регуляция стемности рака и иммунной эвазии при раке поджелудочной железы.

после фильтрации получаем 154 отборные записи

[файлик](#)

отформатировали файл [скриптом](#)

[результат](#)

Любопытно кстати что в задании сказано

кодирующую последовательность РНК

но потом советуется

выбрать нужно coding DNA

на что ожидаемо в файле куса тиминов

```
> T      Aa ab_* 89 из 19999+
```

и ни одного урацильчика

будем надеяться что *Mr.Bayes* итак схавает

Ура теперь самое интересное

```
mafft --genafpair --maxiterate 3000 ./formed.fasta > maffted.fasta
```

фурычит

```
reallocating...
done.
generating a scoring matrix for nucleotide (dist=200) ... done
All-to-all alignment.
0 / 155
```

мультипроцессинга не хвататет

```
0[|||||]90.1% Tasks: 140, 340 chld, 2 running
1[|]0.7% Load average: 2.01 2.50 2.84
2[||||]9.4% Uptime: 05:01:51
3[|||]5.2%
Mem[|||||]2.07G/5.67G
Swp[|||||]1.29G/5.67G
```

PID	USER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Command
18730	serotonin	26	6	33.4G	747M	2232	R	100.	12.9	0:34.34	/usr/lib/mafft/lib/mafft



ему чет не понравилось

```
/usr/bin/mafft: line 2756: 18730 Killed                  "$prefix/tbfast" _ -u $unalignlevel  
$localparam -C $numthreads $seqtype $model -g $lexp -f $lgop -Q $spfactor -h $laof -O $LGOP  
-E $LEXP -N $usenaivepairscore $focusarg _ -+ $iterate -W $minimumweight -V "$gopdist -s  
$unalignlevel $legacygapopt $mergearg $termgapopt $outnum $addarg $add2ndhalfarg -C $numthre  
adstb $rnaopt $weightopt $treeinopt $treeoutopt $distoutopt $seqtype $model -f "$gop -Q $s  
pfactor -h $aof $param_fft $localparam $algopt $treealg $scoreoutarg $focusarg < infile > /d  
ev/null 2>> "$progressfile"
```