THE UNIVERSITY *of* EDINBURGH

Bioinformatics 1

Assignment 2

Henrique Manuel Martins Ferrolho
s1683857 - henrique.ferrolho@gmail.com

November 24, 2016

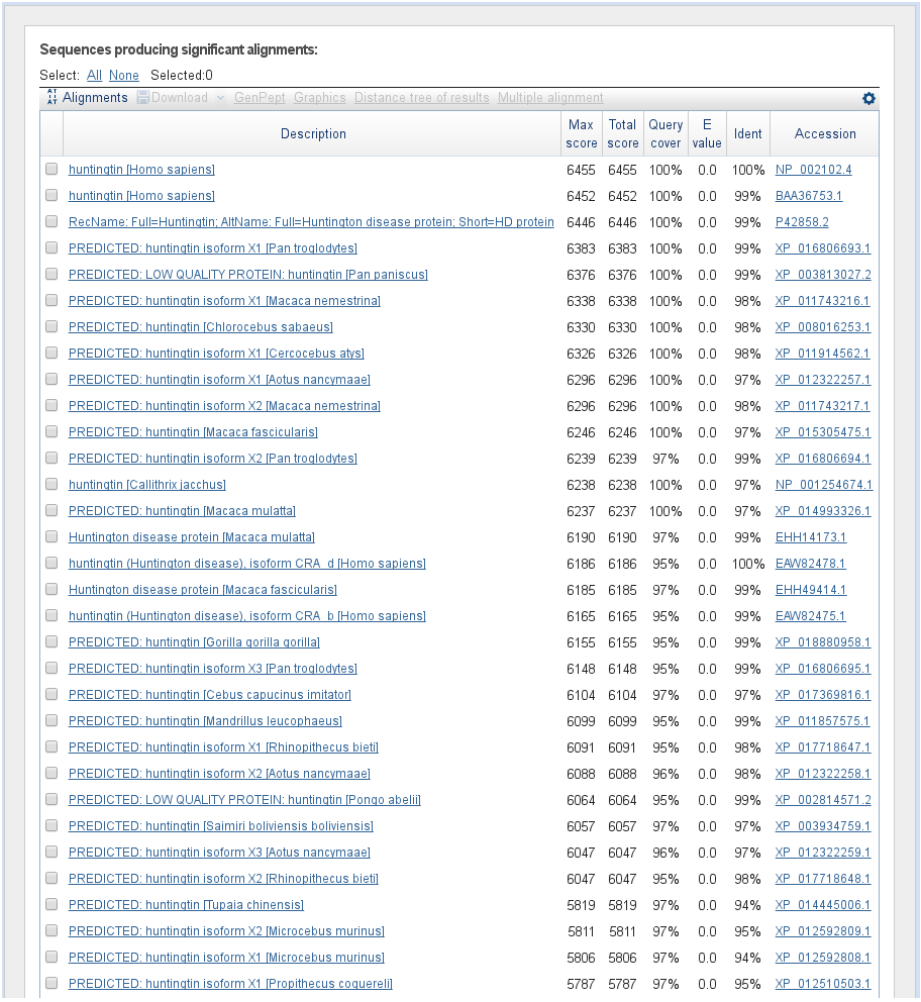
Contents

1 Question

Homologues are two genes from a common ancestor DNA sequence which are related to each other. Furthermore, homologues can be classified as **Orthologues** or as **Paralogues**, whether they have been separated by the event of *speciation* or by the event of *genetic duplication*, respectively.

1.1 Use a BLAST query to find orthologues of your gene. Describe what this query returned, and how you selected potential orthologues.

I have run a Protein BLAST for the *huntingtin* [Homo sapiens] protein (NCBI Reference Sequence: NP_002102.4) - figure 1 is a screenshot of some of the query results.



Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
huntingtin [Homo sapiens]	6455	6455	100%	0.0	100%	NP_002102.4
huntingtin [Homo sapiens]	6452	6452	100%	0.0	99%	BAA36753.1
RecName: Full=Huntingtin; AltName: Full=Huntington disease protein; Short=HD protein	6446	6446	100%	0.0	99%	P42858.2
PREDICTED: huntingtin isoform X1 [Pan troglodytes]	6383	6383	100%	0.0	99%	XP_016806693.1
PREDICTED: LOW QUALITY PROTEIN: huntingtin [Pan paniscus]	6376	6376	100%	0.0	99%	XP_003813027.2
PREDICTED: huntingtin isoform X1 [Macaca nemestrina]	6338	6338	100%	0.0	98%	XP_011743216.1
PREDICTED: huntingtin [Chlorocebus sabaeus]	6330	6330	100%	0.0	98%	XP_008016253.1
PREDICTED: huntingtin isoform X1 [Cercocebus atys]	6326	6326	100%	0.0	98%	XP_011914562.1
PREDICTED: huntingtin isoform X1 [Aotus nancymaae]	6296	6296	100%	0.0	97%	XP_012322257.1
PREDICTED: huntingtin isoform X2 [Macaca nemestrina]	6296	6296	100%	0.0	98%	XP_011743217.1
PREDICTED: huntingtin [Macaca fascicularis]	6246	6246	100%	0.0	97%	XP_015305475.1
PREDICTED: huntingtin isoform X2 [Pan troglodytes]	6239	6239	97%	0.0	99%	XP_016806694.1
huntingtin [Callithrix jacchus]	6238	6238	100%	0.0	97%	NP_001254674.1
PREDICTED: huntingtin [Macaca mulatta]	6237	6237	100%	0.0	97%	XP_014993326.1
Huntington disease protein [Macaca mulatta]	6190	6190	97%	0.0	99%	EHH14173.1
huntingtin (Huntington disease), isoform CRA_d [Homo sapiens]	6186	6186	95%	0.0	100%	EAW82478.1
Huntington disease protein [Macaca fascicularis]	6185	6185	97%	0.0	99%	EHH49414.1
huntingtin (Huntington disease), isoform CRA_b [Homo sapiens]	6165	6165	95%	0.0	99%	EAW82475.1
PREDICTED: huntingtin [Gorilla gorilla gorilla]	6155	6155	95%	0.0	99%	XP_018880958.1
PREDICTED: huntingtin isoform X3 [Pan troglodytes]	6148	6148	95%	0.0	99%	XP_016806695.1
PREDICTED: huntingtin [Cebus capucinus imitator]	6104	6104	97%	0.0	97%	XP_017369816.1
PREDICTED: huntingtin [Mandrillus leucophaeus]	6099	6099	95%	0.0	99%	XP_011857575.1
PREDICTED: huntingtin isoform X1 [Rhinopithecus bieti]	6091	6091	95%	0.0	98%	XP_017718647.1
PREDICTED: huntingtin isoform X2 [Aotus nancymaae]	6088	6088	96%	0.0	98%	XP_012322258.1
PREDICTED: LOW QUALITY PROTEIN: huntingtin [Pongo abelii]	6064	6064	95%	0.0	99%	XP_002814571.2
PREDICTED: huntingtin [Saimiri boliviensis boliviensis]	6057	6057	97%	0.0	97%	XP_003934759.1
PREDICTED: huntingtin isoform X3 [Aotus nancymaae]	6047	6047	96%	0.0	97%	XP_012322259.1
PREDICTED: huntingtin isoform X2 [Rhinopithecus bieti]	6047	6047	95%	0.0	98%	XP_017718648.1
PREDICTED: huntingtin [Tupaia chinensis]	5819	5819	97%	0.0	94%	XP_014445006.1
PREDICTED: huntingtin isoform X2 [Microcebus murinus]	5811	5811	97%	0.0	95%	XP_012592809.1
PREDICTED: huntingtin isoform X1 [Microcebus murinus]	5806	5806	97%	0.0	94%	XP_012592808.1
PREDICTED: huntingtin isoform X1 [Propithecus coquereli]	5787	5787	97%	0.0	95%	XP_012510503.1

Figure 1: Some of the *htt* Protein BLAST results.

The Protein BLAST query returned a list of sequences producing significant alignments, and a graphic summary with their alignment scores.

From this list we can select potential orthologues like: *Pan troglodytes*, *Macaca nemestrina*, *Gorilla gorilla gorilla*, etc.

1.2 Compare your result with a query for homologues. To find these, select 'HomoloGene' in NCBI instead of 'Gene' as search option.

HomoloGene *huntingtin* query: https://www.ncbi.nlm.nih.gov/homologene/1593

The query linked above identifies the list below as being putative homologue genes:

- *H. sapiens*
- *P. troglodytes*
- *C. lupus*
- *B. taurus*
- *M. musculus*
- *R. norvegicus*
- *G. gallus*
- *X. tropicalis*
- *D. rerio*

1.3 How well do the results match when you consider e-value, percentage overlap, or score from your BLAST search?

The following table contains the homologues returned by the HomoloGene query, and their respective *E value*, *query cover*, and *score*, obtained by cross referencing their proteins *accession* with the Protein BLAST results.

Species	E value	Query cover	Score
H. sapiens	0.0	100%	6455
P. troglodytes	0.0	100%	6383
C. lupus	0.0	97%	5700
B. taurus	0.0	97%	5579
M. musculus	0.0	97%	5665
R. norvegicus	0.0	97%	5671
G. gallus	0.0	100%	5295
X. tropicalis	0.0	99%	4949
D. rerio	0.0	99%	4467

E value

From the BLAST Frequently Asked questions: *The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially as the Score (S) of the match increases. Essentially, the E value describes the random background noise.*

Concerning our search results, all the results from the HomoloGene have a E value of 0.0, meaning they are all significant. Nonetheless, this indicator is no good for any comparison between the results.

Percentage overlap

By inspecting the table above, we can see that only three species - *H. sapiens*, *P. troglodytes*, and *G. gallus* - have a total percentage overlap, i.e. query cover of 100%.

Score

H. sapiens is unsurprisingly the Species with the greatest score because we used it to run the HomoloGene query.

P. troglodytes follows it, having a score of **6383**.

The species which has the next greatest score is *R. norvegicus* - **5671**. Even though its query cover is only 97%, its *identity percentage* is greater than that of the other species of the table, which explains the higher score.

1.4 List those results from your initial query that you think are true orthologues, and explain why.

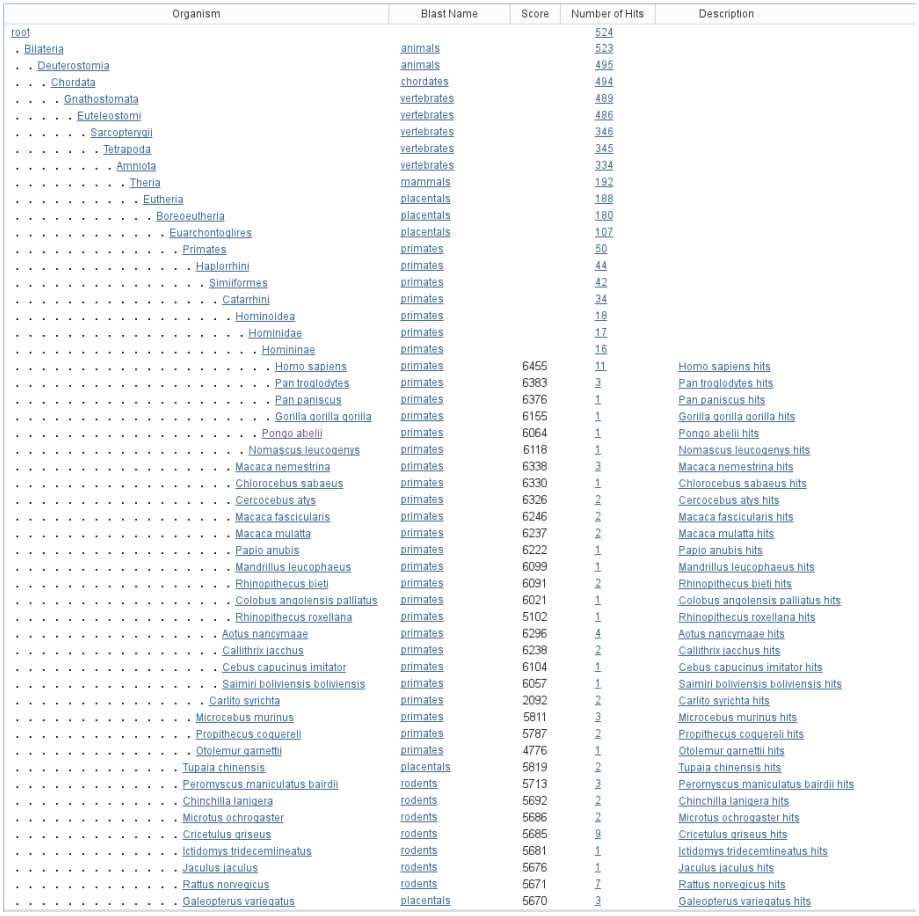
Note that NCBI BLAST has a useful view called 'Taxonomy Reports'. In addition to the NCBI service, you can also try the Ensembl database at http://www.ensembl.org/Homo_sapiens/Tools/Blast?db=core.

Figure 2 shows the lineage tree obtained from the Protein BLAST search.

The following organisms are listed under *Homininae*:

- *Homo sapiens*
- *Pan troglodytes*
- *Pan paniscus*
- *Gorilla gorilla gorilla*

They constitute what I consider to be true orthologues for the *huntingtin* gene.



Organism	Blast Name	Score	Number of Hits	Description
root			524	
. Bilateria	animals		523	
. . Deuterostomia	animals		495	
. . . Chordata	chordates		494	
. . . . Gnathostomata	vertebrates		489	
. . . . . Euteleostomi	vertebrates		486	
. . . . . Sarcopterygii	vertebrates		346	
. . . . . Tetrapoda	vertebrates		345	
. . . . . Amniota	vertebrates		334	
. . . . . Theria	mammals		192	
. . . . . Eutheria	placentals		188	
. . . . . Boreoeutheria	placentals		180	
. . . . . Euarchontoglires	placentals		107	
. . . . . Primates	primates		50	
. . . . . Haplorrhini	primates		44	
. . . . . Simiiformes	primates		42	
. . . . . Catarrhini	primates		34	
. . . . . Hominoidea	primates		18	
. . . . . Hominidae	primates		17	
. . . . . Homininae	primates		16	
. . . . . Homo sapiens	primates	6455	11	Homo sapiens hits
. . . . . Pan troglodytes	primates	6383	3	Pan troglodytes hits
. . . . . Pan paniscus	primates	6376	1	Pan paniscus hits
. . . . . Gorilla gorilla gorilla	primates	6155	1	Gorilla gorilla gorilla hits
. . . . . Pongo abelii	primates	6064	1	Pongo abelii hits
. . . . . Nomascus leucogenys	primates	6118	1	Nomascus leucogenys hits
. . . . . Macaca nemestrina	primates	6338	3	Macaca nemestrina hits
. . . . . Chlorocebus sabaeus	primates	6330	1	Chlorocebus sabaeus hits
. . . . . Cercocebus atys	primates	6326	2	Cercocebus atys hits
. . . . . Macaca fascicularis	primates	6246	2	Macaca fascicularis hits
. . . . . Macaca mulatta	primates	6237	2	Macaca mulatta hits
. . . . . Papio anubis	primates	6222	1	Papio anubis hits
. . . . . Mandrillus leucophaeus	primates	6099	1	Mandrillus leucophaeus hits
. . . . . Rhinopithecus bieti	primates	6091	2	Rhinopithecus bieti hits
. . . . . Colobus angolensis palliatus	primates	6021	1	Colobus angolensis palliatus hits
. . . . . Rhinopithecus roxellana	primates	5102	1	Rhinopithecus roxellana hits
. . . . . Aotus nancymaae	primates	6296	4	Aotus nancymaae hits
. . . . . Callithrix jacchus	primates	6238	2	Callithrix jacchus hits
. . . . . Cebus capucinus imitator	primates	6104	1	Cebus capucinus imitator hits
. . . . . Saimiri boliviensis boliviensis	primates	6057	1	Saimiri boliviensis boliviensis hits
. . . . . Carlito syrichta	primates	2092	2	Carlito syrichta hits
. . . . . Microcebus murinus	primates	5811	3	Microcebus murinus hits
. . . . . Propithecus coquereli	primates	5787	2	Propithecus coquereli hits
. . . . . Otolemur garnettii	primates	4776	1	Otolemur garnettii hits
. . . . . Tupaia chinensis	placentals	5819	2	Tupaia chinensis hits
. . . . . Peromyscus maniculatus bairdii	rodents	5713	3	Peromyscus maniculatus bairdii hits
. . . . . Chinchilla lanigera	rodents	5692	2	Chinchilla lanigera hits
. . . . . Microtus ochrogaster	rodents	5686	2	Microtus ochrogaster hits
. . . . . Cricetulus griseus	rodents	5685	9	Cricetulus griseus hits
. . . . . Ictidomys tridecemlineatus	rodents	5681	1	Ictidomys tridecemlineatus hits
. . . . . Jaculus jaculus	rodents	5676	1	Jaculus jaculus hits
. . . . . Rattus norvegicus	rodents	5671	7	Rattus norvegicus hits
. . . . . Galeopterus variegatus	placentals	5670	3	Galeopterus variegatus hits

Figure 2: Screenshot of the taxonomy reports view from the *htt* Protein BLAST results.

2 Question

2.1 Now consider the phylogenetic implications of your results. First, create a BLAST tree (using the function 'Distance tree of results').

Figure 3: Distance tree generated from the BLAST query.

2.2 Describe the result and its relevance.

The resulting distance tree is a visual representation of the relationship between organisms related to the *huntingtin* gene and their ancestors.

It is relevant to analyse and study the proximity of organisms which express the same gene one is researching - *huntingtin* in this case.

Figure 4: Sub-tree of the distance tree in figure 3.

2.3 How can this be used to study the disease in another model organism?

As it has been previously said, this kind of representation - distance tree obtained with blast - can be used to list in a tree structure organisms which are closely related to the disease.

This is very useful for situations where one needs to look for a model organism to study a disease.

For example, in our case we are interested in *HD*. If we wanted to use a model organism to study the disease gene expressed in humans, we would like to choose the model organism as close to humans as possible. The tree representation makes it easier to answer this question: one can look for the leaf where the gene is expressed, and from there look for the nearest leaf which corresponds to a model organism.

2.4 Are there known orthologues in mouse, fruit fly or yeast?

There are for the mouse and for the fruit fly, but not for yeast. The *NCBI* database contains orthologs of numerous species, including *Mus musculus* and *Drosophila melanogaster*.

2.5 Using the information under 'General gene information', discuss whether mouse, fruit fly or yeast (or any of these for which an orthologue exists) are suitable models to study the disease.

Note: To give an example, if the gene is implicated in cell cycle, yeast may well be a good model because it is easier to study than mice. But if the gene is relevant for brain function, yeast may of course be less relevant even if an orthologue exists.

From the *General gene information* - `https://www.ncbi.nlm.nih.gov/gene/3064#general-gene-info` - one can read: *The HTT gene is conserved in chimpanzee, dog, cow, mouse, rat, chicken, zebrafish, and frog.*

Considering the disease symptoms and taking into account other details discussed on Assignment 1, the mouse is the most suitable model to study the disease. Yeast is not an ortholog of the gene, and the mouse should be chosen over the fruit fly because its proximity to *H. Sapiens* is much greater.

3 Question

3.1 Select four orthologues, and create a rooted phylogenetic tree using the UPGMA algorithm as shown in class.

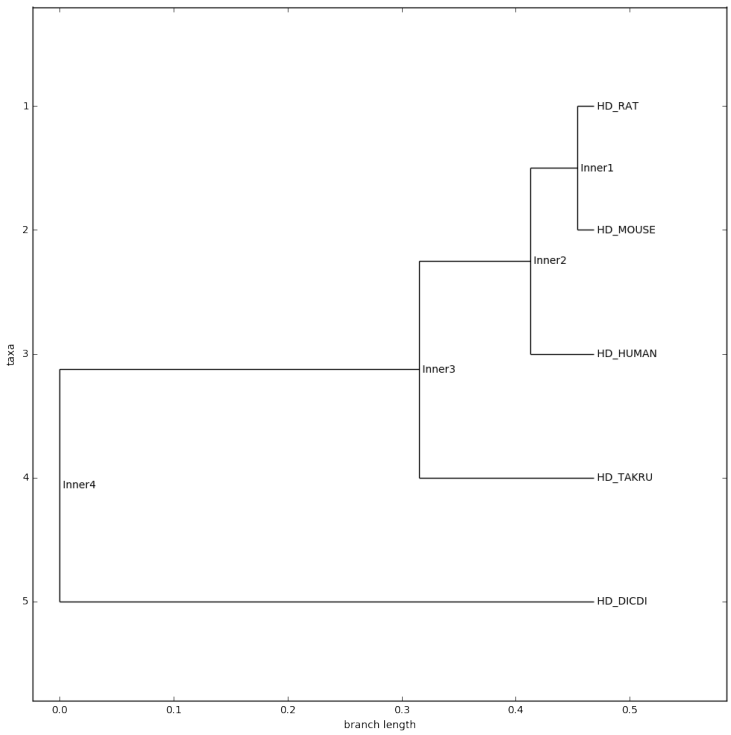


Figure 5: Rooted phylogenetic tree obtained using the UPGMA algorithm with Jupyter Notebook for protein NP_002102.4.

3.2 Does it match the BLAST result from Q1/2?

The orthologues in the obtained phylogenetic tree using UPGMA are:

Rattus norvegicus, *Mus musculus*, *Homo sapiens*, *Takifugu rubripes*, and *Dictyostelium discoideum*.

Although I did not focus on them answering the homework, they do match the outputs of my queries. For example, *Rattus norvegicus*, *Mus musculus*, and *Homo sapiens* are listed on section 1.2. And they also show up on the distance tree from question 2. Therefore, yes they do match the results from questions 1 and 2.

4 Sources

Ensembl genome browser
http://www.ensembl.org/Homo_sapiens/Tools/Blast?db=core

Genetics Home Reference
https://ghr.nlm.nih.gov/condition/huntington-disease

National Center for Biotechnology Information
https://www.ncbi.nlm.nih.gov/

Wikipedia
https://en.wikipedia.org/wiki/Huntington%27s_disease