# Bioinformatics Assignment 1

**S1736880**

## Abstract

The aim of this assignment is to explore core molecular biology and bioinformatics concepts. It follows on directly from the work we did in assignment 1, where we selected KITLG gene which is considered to be responsible for testicular cancer [1, 2, 3]. We use a BLAST query to find potential orthologues of our gene and we consider the phylogenetic implications of our results. Finally we select four orthologues, and create a rooted phylogenetic tree using the UPGMA algorithm.

## 1. Answer to question 1

Firstly we need to define the homology term and the difference between orthologues and paralogues. So, homologue is a gene similar in structure and evolutionary origin to a gene in another species. The homology term may apply to the relationship between genes separated by the event of speciation or to the relationship between genes separated by the event of genetic duplication. Orthologues are genes in different species that evolved from a common ancestral gene by speciation while paralogues are genes related by duplication within a genome. Orthologues usually retain the same function in the course of evolution, whereas paralogues evolve new functions, even if these are related to the original one.

| GENE NAME | E-VALUE | SCORE | ACCESSION NO |
|---|---|---|---|
| SCF_HUMAN | 0.0 | 1435.0 | P21583 |
| SCF_HORSE | 1.29721E-167 | 1202.0 | Q95MD2 |
| SCF_FELCA | 2.44878E-166 | 1193.0 | P79169 |
| SCF_PIG | 2.85171E-164 | 1180.0 | Q29030 |
| SCF_CAPHI | 3.66415E-162 | 1166.0 | Q95M19 |
| SCF_BOVIN | 2.26031E-161 | 1161.0 | Q28132 |
| SCF_NEOVI | 4.97618E-161 | 1159.0 | Q95N18 |
| SCF_CANFA | 6.26409E-161 | 1158.0 | Q06220 |
| SCF_RAT | 1.06517E-157 | 1137.0 | P21581 |
| SCF_SHEEP | 5.65246E-157 | 1131.0 | P79368 |
| SCF_MOUSE | 3.81371E-152 | 1100.0 | P20826 |
| SCF_CHICK | 1.85782E-83 | 649.0 | Q09108 |
| SCF_COTJA | 4.48678E-83 | 647.0 | Q90314 |

*Table 1.* Searching for orthologues in KITLG gene. The code for these results can be found in Appendix A.

Thus, in our case we use BLAST search (blastp) in the 'swissprot' database and we find some potential orthologues as Table 1 indicates. The code for these results is in appendix A. We have to note that the first result in the Table 1 is not an orthologue because it refers to our gene from which we obtain the protein and that is the reason that e-value equals to 0.0.

After that we need to compare our results with Blast search from the ncbi database [4]. We use Protein Blast search tool with protein's accession number NP_000890.1 (protein for our KITLG gene), UniProtKB/Swiss-Prot ('swissprot') database and the algorithm blastp (protein-protein BLAST). In the "Descriptions" section we can see the same results as Figure 1 illustrates.

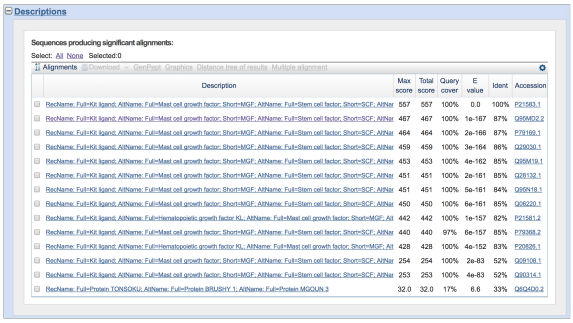
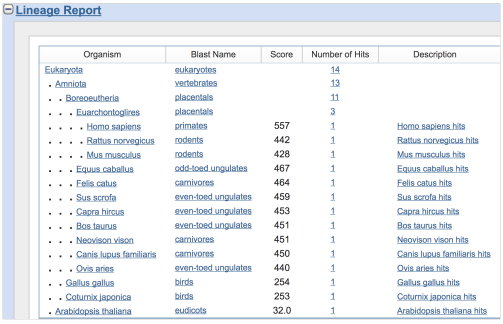



*Figure 1.* Results from Protein Blast in protein NP_000890.1

Comparing the results between Table 1 and Figure 1 it can be observed that the results are the same (they have the same accession number) and that is reasonable because we obtain them with the same procedure but with different tools. In the first way we acquire them through Python code which allows us access the same database as we use with ncbi protein blast directly in the second approach. We need to point out that in the first approach we have thirteen results and in the second we have fifteen because we use different threshold for e-value. Furthermore, the results match very well in e-value (with ncbi blast they rounded without decimals and with Python code we have more accuracy in decimals). For instance protein Q95MD2 for SCF_HORSE gene has 1.29721e-167 with the first approach and 1e-167 with the first approach but approximately it is the same. In NCBI results, it can be noticed that the score is expressed as bit score and as raw score [5] and we can see that raw scores are the same with these that we calculated from the Python code with the first approach. So, besides e-value, the score for each gene has also the same value.



Lineage Report

| Organism | Blast Name | Score | Number of Hits | Description |
|---|---|---|---|---|
| Eukaryota | eukaryotes | | 14 | |
| . Amniota | vertebrates | | 13 | |
| . . Boreoeutheria | placentals | | 11 | |
| . . . Euarchontoglires | placentals | | 3 | |
| . . . . Homo sapiens | primates | 557 | 1 | Homo sapiens hits |
| . . . . Rattus norvegicus | rodents | 442 | 1 | Rattus norvegicus hits |
| . . . . Mus musculus | rodents | 428 | 1 | Mus musculus hits |
| . . . Equus caballus | odd-toed ungulates | 467 | 1 | Equus caballus hits |
| . . . Felis catus | carnivores | 464 | 1 | Felis catus hits |
| . . . Sus scrofa | even-toed ungulates | 459 | 1 | Sus scrofa hits |
| . . . Capra hircus | even-toed ungulates | 453 | 1 | Capra hircus hits |
| . . . Bos taurus | even-toed ungulates | 451 | 1 | Bos taurus hits |
| . . . Neovison vison | carnivores | 451 | 1 | Neovison vison hits |
| . . . Canis lupus familiaris | carnivores | 450 | 1 | Canis lupus familiaris hits |
| . . . Ovis aries | even-toed ungulates | 440 | 1 | Ovis aries hits |
| . . Gallus gallus | birds | 254 | 1 | Gallus gallus hits |
| . . Coturnix japonica | birds | 253 | 1 | Coturnix japonica hits |
| . Arabidopsis thaliana | eudicots | 32.0 | 1 | Arabidopsis thaliana hits |

*Figure 2.* Lineage report from Protein Blast in protein NP_000890.1

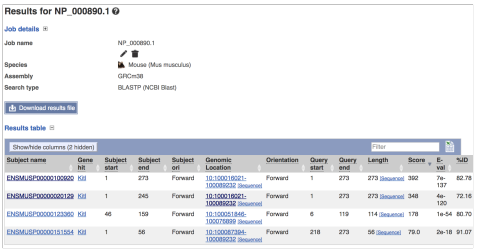
Taxonomy Report

| Taxonomy | Number of hits | Number of Organisms | Description |
|---|---|---|---|
| Eukaryota | 14 | 14 | |
| . Amniota | 13 | 13 | |
| . . Boreoeutheria | 11 | 11 | |
| . . . Euarchontoglires | 3 | 3 | |
| . . . . Homo sapiens | 1 | 1 | Homo sapiens hits |
| . . . . Murinae | 2 | 2 | |
| . . . . . Rattus norvegicus | 1 | 1 | Rattus norvegicus hits |
| . . . . . Mus musculus | 1 | 1 | Mus musculus hits |
| . . . Laurasiatheria | 8 | 8 | |
| . . . . Equus caballus | 1 | 1 | Equus caballus hits |
| . . . . Carnivora | 3 | 3 | |
| . . . . . Felis catus | 1 | 1 | Felis catus hits |
| . . . . . Caniformia | 2 | 2 | |
| . . . . . Neovison vison | 1 | 1 | Neovison vison hits |
| . . . . . Canis lupus familiaris | 1 | 1 | Canis lupus familiaris hits |
| . . . . Cetartiodactyla | 4 | 4 | |
| . . . . . Sus scrofa | 1 | 1 | Sus scrofa hits |
| . . . . . Bovidae | 3 | 3 | |
| . . . . . Caprinae | 2 | 2 | |
| . . . . . Capra hircus | 1 | 1 | Capra hircus hits |
| . . . . . Ovis aries | 1 | 1 | Ovis aries hits |
| . . . . . Bos taurus | 1 | 1 | Bos taurus hits |
| . . . Phasianidae | 2 | 2 | |
| . . . Gallus gallus | 1 | 1 | Gallus gallus hits |
| . . . Coturnix japonica | 1 | 1 | Coturnix japonica hits |
| . Arabidopsis thaliana | 1 | 1 | Arabidopsis thaliana hits |

*Figure 3.* Taxonomy report from Protein Blast in protein NP_000890.1

True orthologues probably are these genes that have e-value less than 1e-100 and the species for these genes have a closer common ancestor with our human gene. So, as we can see from Figures 2, 3 such potential genes are SCF_RAT and SCF_MOUSE for rattus norvegicus hits and mus musculus hits respectively. But all of our results are orthologues because Figure 3 illustrates all the ancestors as we go back in the tree.

In addition to the NCBI service, we try the Ensembl database [6]. In the "sequence data" box we put our protein with accession number NP_000890.1 and we select protein instead of DNA in the radio button. In the "search against" area we try the mouse as species and we select the radio button for Protein database (Proteins GENCODE/Ensembl). Afterward, we select BLASTP for search tool and we let the rest options with their default values and finally we run this job. We have four hits for result as we can see from the figure 4 and we choose the first one because it has the smallest e-value and highest score (we have to note that all of four hits are for the same protein in the mouse). We can see that there are slight differences in e-value (7e-137) and in bit score (392) for this result against the result (4e-152 e-value and 428 bit score) we take from NCBI database.



Results for NP_000890.1

Job details

| | |
|---|---|
| Job name | NP_000890.1 |
| Species | Mouse (Mus musculus) |
| Assembly | GRCm38 |
| Search type | BLASTP (NCBI Blast) |

Download results file

Results table

Show/hide columns (2 hidden) — Filter

| Subject name | Gene hit | Subject start | Subject end | Subject ori | Genomic Location | Orientation | Query start | Query end | Length | Score | E-val | %ID |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ENSMUSP00000100920 | Kit | 1 | 273 | Forward | 10:100018221-100089232 (Sequence) | Forward | 1 | 273 | 273 (Sequence) | 392 | 7e-137 | 82.78 |
| ENSMUSP00000020129 | Kit | 1 | 245 | Forward | 10:100018221-100089232 (Sequence) | Forward | 1 | 273 | 273 (Sequence) | 348 | 4e-120 | 72.16 |
| ENSMUSP00000123360 | Kit | 46 | 159 | Forward | 10:100051846-100078899 (Sequence) | Forward | 6 | 119 | 114 (Sequence) | 178 | 1e-54 | 80.70 |
| ENSMUSP00000151554 | Kit | 1 | 56 | Forward | 10:100087394-100089232 (Sequence) | Forward | 218 | 273 | 56 (Sequence) | 79.0 | 2e-18 | 91.07 |

*Figure 4.* Blastp for protein NP_000890 in the Ensembl database.

## 2. Answer to question 2

Phylogenetics is the study of evolutionary relationships between organisms. It describes genealogical ties among organisms and estimates the time of divergence between them. Comparison of DNA gives clues about mutations that accumulate over time and about gene duplication events that occur occasionally and create paralogues. The result of this analysis is a phylogenetic tree (i.e. Figure 5) about the history of the evolutionary relationships between organisms [7].

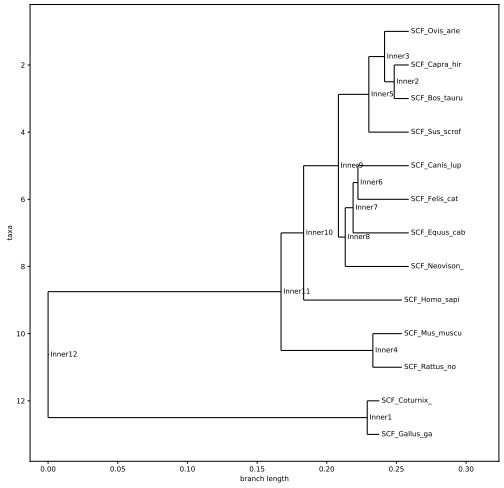



*Figure 5.* Phylogenetics tree for our protein

In the results from the blastp in NCBI database there is a hyperlink 'Distance tree of results' and with this function we create the BLAST tree that it can be seen in Figure 6. We can observe the species and how the evolution was through the tree. Thus, we notice which organisms have closer ancestor and as a result they have more common characteristics.

The phylogenetic analysis can be used to study the same disease among organisms which have recent common ancestor and it can make predictions for this disease. [8] Comparing the organisms from the phylogenetic tree we can find the closer organism to human and study the disease on this. We can also make experiments and we may conclude useful results. Furthermore we may trace a parallelism between human and the other organism or we can make predictions for the disease.

There is known orthologue in mouse [9, 10, 11] and we can verify it using the information under 'General gene information'. We believe that a model such as mouse is more suitable to study testicular cancer than fruit fly or yeast because firstly mouse and human have a closer common ancestor and secondly human and mouse are more similar models (mammals) to study this disease. Another reason is that this gene plays major role in stem cell maintenance besides its other functionalities [12] and so it is more suitable to study it in a mouse model.

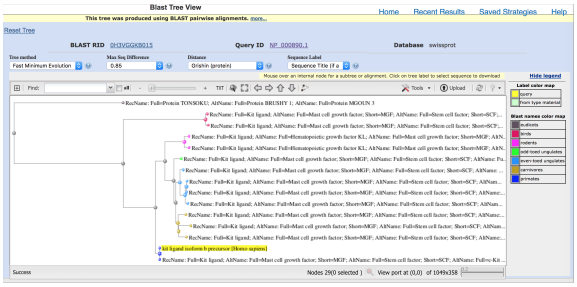



*Figure 6.* Blast tree

## 3. Answer to question 3

We select four orthologues (2nd, 3rd, 9th, 11th with the order that the genes are presented in Table 1) and we create a rooted phylogenetic tree using the UPGMA algorithm as shown in Figure 7. The code we used to construct this tree can be found in appendix B.

We can ran the UPGMA algorithm by hand and as you can see in Figure 8 we have the same result. Firstly, we need the distances for the four orthologues that we have chosen and the code for this calculation is in appendix C. Thus, we assign each sequence (in our case protein) to one cluster and every one of them has leaf with zero height. Afterward, we apply UPGMA algorithm on these four clusters choosing the pair with the smallest distance $d_{ij}$ to merge each time. Finally, we take the result that Figure 8 illustrates and it can be observed that it is the same as that in Figure 7. We need to point out that the results for the distances from the code in appendix C have been rounded to four decimals for the calculations in Figure 8.

The BLAST result matches with the findings from the

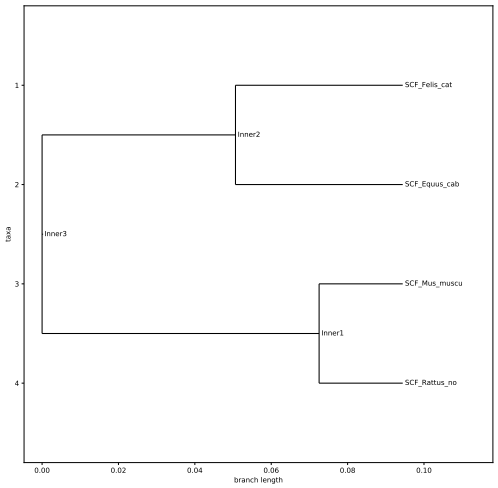



*Figure 7.* Rooted phylogenetic tree

**Algorithm 1** UPGMA algorithm

Initialisation:

1. Assign each sequence $s_i$ to one cluster $C_i$
2. Each sequence has leaf with height 0

Iteration:

3. Find minimum $d_{ij}$ (if multiple, pick random)
4. Define new cluster k: $C_k = C_i \cup C_j$
5. Define node k with daughters i and j, with height $d_{ij}/2$
6. Add k to clusters and remove i and j

Termination:

7. When two clusters i and j remain, place root at $d_{ij}/2$

first two parts. It can be noticed that SCF_HORSE and SCF_FELCA genes has less E-value and higher scores than SCF_RAT and SCF_MOUSE genes (see Table 1) and so the distances from Figure 7 make sense to be smaller for the first two genes than the second ones. In addition, we can compare Figures 6, 7 and we can point out that the categories of rodents and primates from the first figure have similar distances with the horse and cat (primates) and mouse and rat (rodents) from the second figure.

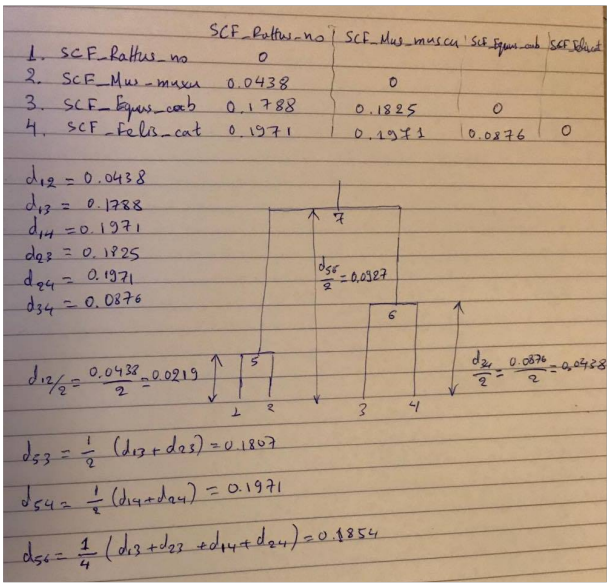



*Figure 8.* Rooted phylogenetic tree

## Appendix A.

```
# import required Biophython functions
from Bio import Entrez
from Bio.Blast import NCBIXML
from Bio.Blast import NCBIWWW
from Bio import SeqIO

Entrez.email = 'A.N.Other@example.com'
my_protein = 'NP_000890.1'

handle = Entrez.efetch(db = "protein", id =
    my_protein, rettype = "gb", retmode =
    "text")
record = SeqIO.read(handle, "genbank")
handle.close()

# here we have chosen the human KITLG gene
result_handle = NCBIWWW.qblast('blastp',
    'swissprot', record.seq)

# parse the results
result_handle.seek(0)
blast_record = NCBIXML.read(result_handle)

print('Gene name\te-value\tscore\taccession
    number')
for a in blast_record.alignments:
   if a.hsps[0].expect > 1e-1:
      break
   print(a.title.split('|')[4].split(' ')[0] +
       '\t' + str(a.hsps[0].expect) + '\t' +
       str( a.hsps[0].score) + '\t' +
       a.accession)
```

## Appendix B.

```
E_VALUE_THRESH = 1e-6

# the following will write all results into a
    FASTA file for the # MSA
def get_seqrecs(alignments, threshold):
  for i,aln in enumerate(alignments):
    # random 4 orthologues
    if i in [1, 2, 8, 10]:
      for hsp in aln.hsps:
        if hsp.expect < threshold:
            id = aln.title.split('|')[4].split('
                ')[0].split('_')[0] + '_' +
                species[i].replace(' ', '_')[:9]
            print(id)
            yield SeqRecord(Seq(hsp.sbjct), id =
                id)
            break

best_seqs = get_seqrecs(blast_record.alignments,
    E_VALUE_THRESH)
# write out to a fasta file
SeqIO.write(best_seqs, 'family_alignment.fasta',
    'fasta')

run Muscle MSA
cmdline = MuscleCommandline(
    './muscle3.8.31_i86linux64', input =
    'family_alignment.fasta', out =
    'family_alignment.aln', clw = True)
cmdline()

from Bio import AlignIO
alignment = AlignIO.read('family_alignment.aln',
    'clustal')
print(alignment)

from Bio.Phylo.TreeConstruction import
    DistanceCalculator
calculator = DistanceCalculator('identity')
dm = calculator.get_distance(alignment)
print(dm)

from Bio.Phylo.TreeConstruction import
    DistanceTreeConstructor
constructor =
    DistanceTreeConstructor(calculator, 'upgma')
tree = constructor.build_tree(alignment)
print(tree)

from Bio import Phylo
import matplotlib.pyplot as plt
fig = plt.figure(figsize = (12, 12))
ax = plt.subplot(111)
Phylo.draw(tree, axes = ax)
fig.savefig('tree.pdf')
```

## Appendix C.

```
from Bio import AlignIO
alignment = AlignIO.read('family_alignment.aln',
    'clustal')

from Bio.Phylo.TreeConstruction import
    DistanceCalculator
calculator = DistanceCalculator('identity')
dm = calculator.get_distance(alignment)
```

```
print(dm)
```

## References


[1] Phuong L Mai Christine Mueller June A Peters Gennady Bratslavsky Alex Ling Peter M Choyke Ahalya Premkumar Janet Bracci Rissah J Watkins Mary Lou McMaster Larissa A Korde Mark H Greene, Christian P Kratz. Familial testicular germ cell tumors in adults: 2010 summary of genetic risk factors and clinical phenotype. *Endocrine-Related Cancer*, 2010. doi: 10.1677/ERC-09-0254.

[2] Jimenez-Trejo F Chavez-Saldana M Arechaga-Ocampo E, Rojas-Castaneda JC. Epigenetic and risk factors of testicular germ cell tumors: a brief review. *Frontiers in Bioscience*, 2017.

[3] Robert A. Huddart-Janet Shipley & Clare Turnbull Kevin Litchfield, Max Levy. The genomic landscape of testicular germ cell tumours: from susceptibility to treatment. *Nature Reviews Urology*, 2016. doi:10.1038/nrurol.2016.107.

[4] https://blast.ncbi.nlm.nih.gov/Blast.cgi. Accessed: 2017-11-09.

[5] https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html. Accessed: 2017-11-12.

[6] http://www.ensembl.org/Homo_sapiens/Tools/Blast?db=core. Accessed: 2017-11-09.

[7] https://en.wikipedia.org/wiki/Phylogenetics. Accessed: 2017-11-12.

[8] Mohamed Chaouchi Mones Abu-Asab and Hakima Amri. Evolutionary medicine: A meaningful connection between omics, disease, and treatment. *Proteomics Clinical Applications*, pages 122–134, 11 January 2008. doi: 10.1002/prca.200780047.

[9] https://www.ncbi.nlm.nih.gov/gene?cmd=retrieve&list_uids=4254#homology. Accessed: 2017-11-13.

[10] https://www.ncbi.nlm.nih.gov/gene/?Term=ortholog_gene_4254[group]. Accessed: 2017-11-13.

[11] http://www.orthodb.org/?ncbi=1&query=4254. Accessed: 2017-11-13.

[12] http://www.genecards.org/cgi-bin/carddisp.pl?gene=KITLG. Accessed: 2017-11-17.