What do we mean when we describe two objects as being similar? **(1)**

Two objects are similar if they share observable characteristics. TAs the number of shared characteristics increases, the similarity increases.

Exercise 1

Consider the two objects in **Figure 1**. Are these objects similar? In what way(s) would you consider them to be similar? **(2)**

**Yes, the two objects are similar. I consider them similar because they share the characteristics of being vessels that can store materials (they share functional characteristics). They also share overall shape (structural characteristics).**

How could the similarity between the two passages above be quantified? What must be done prior to determining the similarity of these passages? **(3)**

**The similarity between the two passages can be quantified by counting the number of lines in the passage, by counting the number of words in each line of the passages, by counting the total number of words in the passages, by counting the number of rhymes and by comparing the rhyming pattern.**

Considering amino acid residue chemical properties, explain why an alanine substituted with a Serine is assigned a score of 1, while an Alanine substituted with a Tryptophan is assigned a score of -3 in the BLOSUM-62 substitution matrix. **(4)**

**Alanine is a relatively small amino acid with a nonpolar side chain. Serine ia also a relatively small amino acid with a polar side chain. Substitutions between these amino acids give a small positive substitution value due to the similar size, which is unlikely to cause a large disruption in the structure of the protein. The substitution value is not very high because the chemical properties do not match. Tryptophan has a very large bulky side chain that is aromatic and very nonpolar. Substitutions between alanine and tryptophan give negative substitution values because the difference in size will likely cause a large disruption in the structure of the protein. The chemical properties are also different because tryptophan is aromatic.**

What is the total similarity score for these two aligned sequences? **(5)**

Query: MGDVEKGKKIFIMKC

Subject 1: MGEVERGKKLFIMKC

562452655264559

71 compared to 80 when scored to itself

If the query sequence is aligned to a different subject sequence (given below), what is the similarity score? **(6)**

Query: MGDVEKGKKIFIMKC

Subject 2: MCDVWKGKSIFIMKC

536435650464559

58 compared to 80 when scored to itself

Explain why the similarity scores calculated above are different. Consider and refer to information provided in Table 1 as part of your explanation. **(7)**

**In general, the substitutions in the second alignment have greater size and chemical differences than the substitutions in the first alignment. For example, in the second alignment there are substitutions that apply penalties because the differences are so great. This is true for the glycine substitution for cysteine, where the size and chemical differences are extremely large. The glutamate substitution for tryptophan also causes large changes in size and chemistry.**

Did the computationally calculated similarity scores match those that you manually calculated? **(9)**

**Yes**

Pairwise Align Protein results

**Alignment results for 15 residue sequence "Query 1" starting "MGDVEKGKKI" and 15 residue sequence "Subject 1" starting "MGEVERGKKL"**

>Query 1 MGDVEKGKKIFIMKC >Subject 1 MGEVERGKKLFIMKC Alignment score: 71

Pairwise Align Protein results

**Alignment results for 15 residue sequence "Query 1" starting "MGDVEKGKKI" and 15 residue sequence "Subject 2" starting "MCDVWKGKSI"**

>Query 1 MGDVEKGKKIFIMKC >Subject 2 MCDVWKGKSIFIMKC Alignment score: 58

**Exercise 2**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Query Sequence** | | | | | | | | | | | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** |
| S | T | W | G | E | R | G | L | **M** | **P** | **Y** | R | G | L | A | C | E | G | H | I |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Protein Sequence 1** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **21** | **22** | **23** | **24** | **25** | **26** | **27** | **28** | **29** | **30** | **31** | **32** | **33** | **34** | **35** | **36** | **37** | **38** | **39** | **40** | **Total score** |
| M | V | G | V | M | N | S | A | Y | L | N | N | R | L | M | P | Y | G | G | S | G | M | P | E | C | D | Y | D | C | C | W | C | I | M | T | M | N | H | Y | C |
|  |  |  |  |  |  |  |  |  |  |  |  |  | 4 | **5** | **7** | **7** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Score = 23

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Protein Sequence 2** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **21** | **22** | **23** | **24** | **25** | **26** | **27** | **28** | **29** | **30** | **31** | **32** | **33** | **34** | **35** | **36** | **37** | **38** | **39** | **40** | **Total score** |
| R | D | R | N | L | S | S | I | K | L | G | P | Q | V | Q | T | N | Q | M | P | C | G | E | R | D | A | T | N | T | R | C | I | I | N | Y | T | Y | I | S | R |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

No MPY score = 0

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Protein Sequence 3** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **21** | **22** | **23** | **24** | **25** | **26** | **27** | **28** | **29** | **30** | **31** | **32** | **33** | **34** | **35** | **36** | **37** | **38** | **39** | **40** | **Total score** |
| Y | A | W | F | Q | F | R | Q | D | S | F | S | R | E | T | Q | V | T | D | M | P | Y | T | V | W | K | L | I | R |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  | S | T | W | G | E | R | G | -4 | **M** | **P** | **Y** | -1 | G | L | A | C | E | G | H | I |  |  |  |  |  |  |  |  |  |  |

Score = 19

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Protein Sequence 4** | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **21** | **22** | **23** | **24** | **25** | **26** | **27** | **28** | **29** | **30** | **31** | **32** | **33** | **34** | **35** | **36** | **37** | **38** | **39** | **40** | **Total score** |
| M | A | Q | Y | L | A | F | M | P | Y | N | N | S | V | L | H | Y | K | R | A | N | S | I | K | E | Q | H | P | F | R | M | P | Y | E | A | I | A | L | P | Y |
| T | W | G | E | -2 | 0 | 0 | **M** | **P** | **Y** | 0 | 0 | -2 | A | C | E | G | H | I |  |  |  |  |  |  |  | E | R | G | -2 | **M** | **P** | **Y** | 0 | 0 | 2 | 4 | -1 | E | G |  | I |

Score 1 = 19

Score 2 = 25

Which of the proteins from the database is most similar to our query? Which is the least similar? **(1)**

**The most similar is protein 4 because it has the highest similarity score in one of the local alignments identified. The least similar protein is protein 2 because it did not have any match to the query word.**

What problems did you encounter that may have affected your calculated scores? **(2)**

One of the proteins did not have any match to the query word, so it could not be evaluated for similarity. One of the proteins had more than one match to the query word and both matches led to good similarity scores, so it is difficult to know how to characterize the similarity.

How similar do you expect their H4 proteins to be? **(3)**

Given the common role of histone H4 in all organisms and the close evolutionary relationship of humans and chimpanzees, I expect the H4 proteins to be extremely similar.

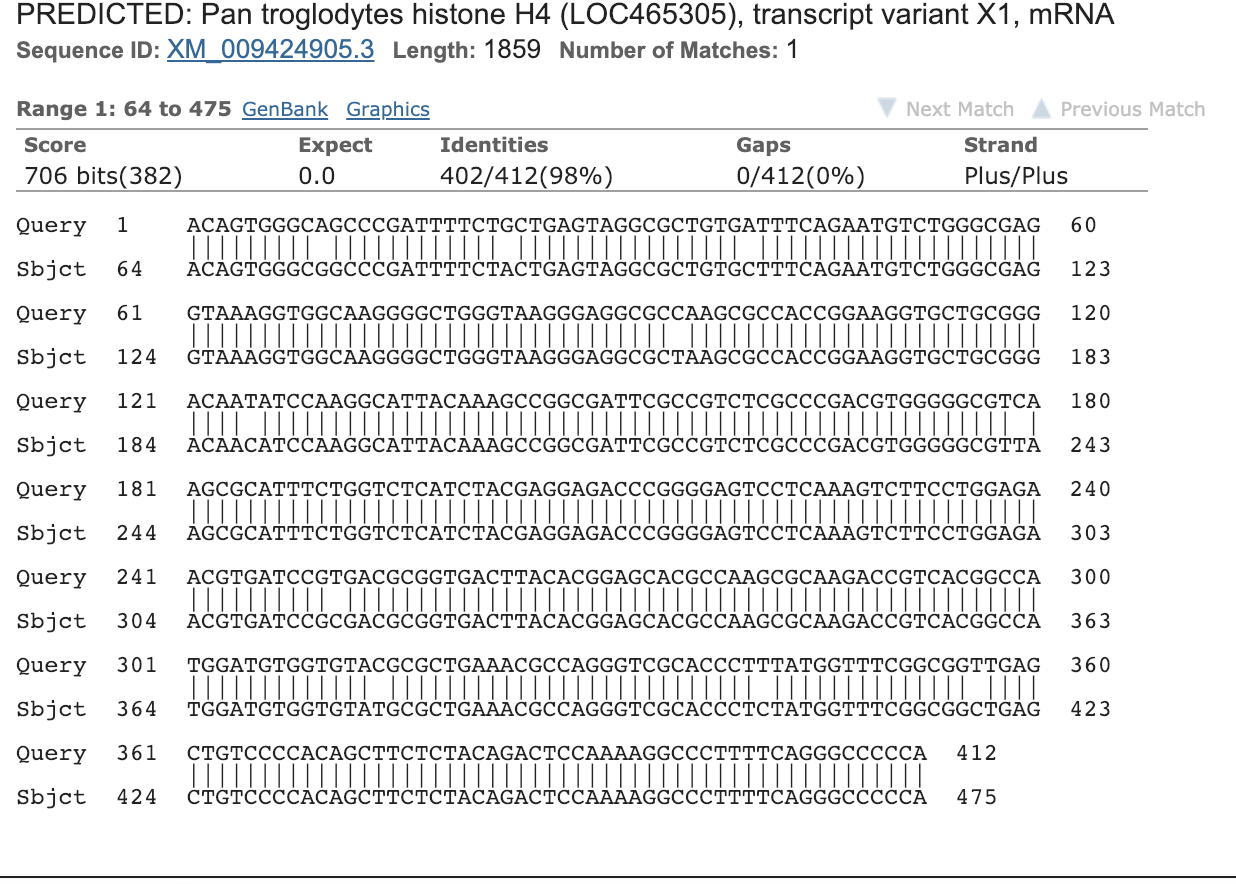
If we were to compare the *nucleotide* sequences for the gene encoding this protein between humans and chimps, do you think they would be identical? Explain. **(4)**

**The nucleotide sequences are not likely to be identical because the genetic code is degenerate and there can be nucleotide changes that result in codons that code identical amino acids According to nucleotide BLAST it is 98% identical.**

On what chromosome is the gene for human histone H4 located? **(5)**

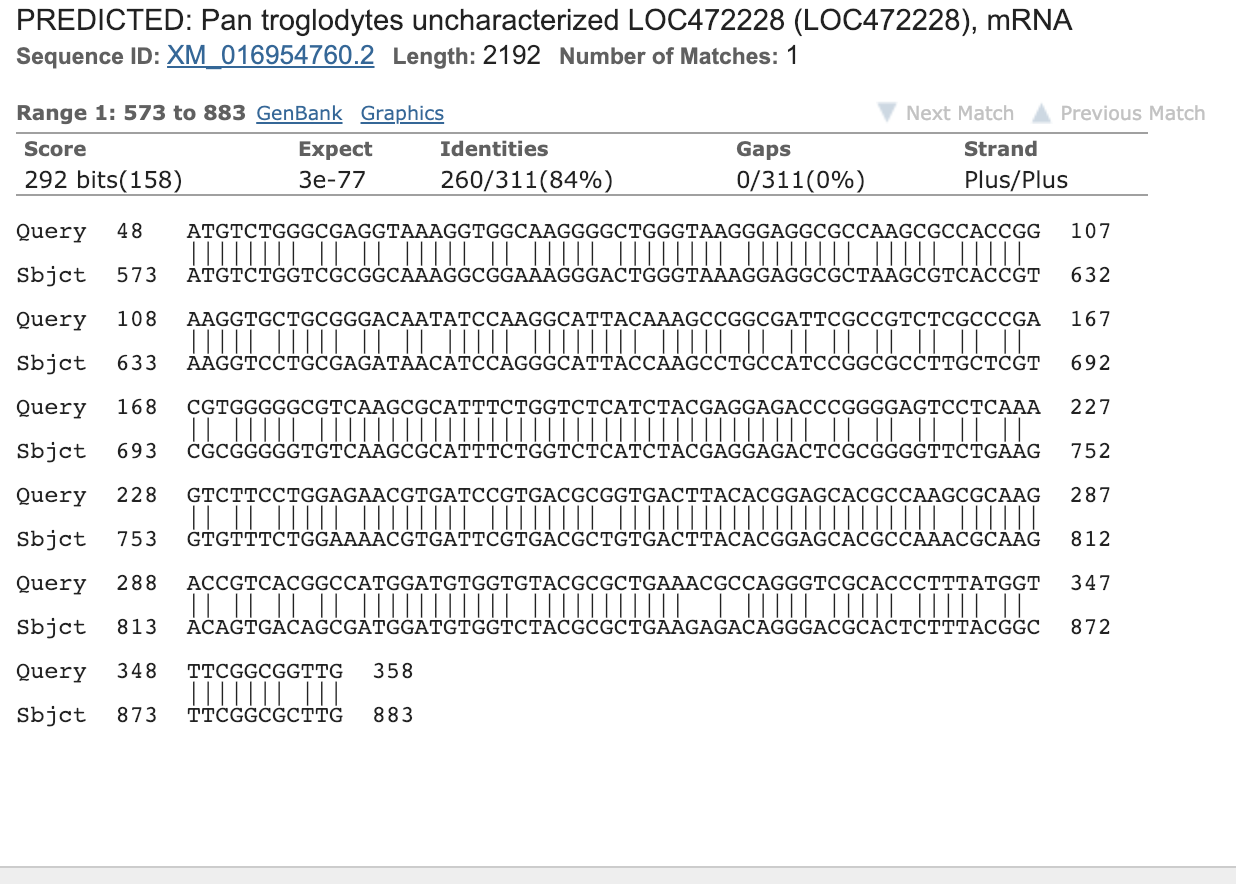
**Human histone H4 is located on chromosome 12.**

Alignments:



How many differences did you find between the query and subject? What percent identity is there between these sequences? Is this at all surprising to you? Explain. **(6)**

There are 10 nucleotide differences in the alignment which is 98% identity. I do not consider this surprising given the conserved function of the protein and the close evolutionary relationship between humans and chimpanzees.



Imagine that a colleague asks you to align a conserved metabolic protein coding gene sequence from a dog to its human homolog. Which sequence type, DNA or protein, would you expect to exhibit the highest percentage of identities? Why? **(7)**

**I expect the protein sequence will display the highest percentage of identity because the genetic code is degenerate and different nucleotide codons can code for identical amino acids.**

**Exercise 3**

**MSA Calculation Template**

**Example MSA:**

Seq A MGDVEKGKKIFVMKCSQCHTVEKGG

Seq B MVDVEKGMKIFVMKCSQCHTVEAGG

Seq C MVDVEKGMLIFVMKCSQCHTVEAGG

Seq D MGDIEKGKLIFVMSCSQCHTVYAGY

Seq E MIDTEKGYIIFVMTCSQCHTVLMGT

From the aligned sequences above we can individually calculate the fraction of differing amino acid residues between the 25 amino acids that make up Seq A and B (identical residues represented by “|”):

Seq A MGDVEKGKKIFVMKCSQCHTVEKGG

| ||||| |||||||||||||| ||

Seq B MVDVEKGMKIFVMKCSQCHTVEAGG

Alignment value 3/25 = 0.12 this value goes into the **Distance Matrix**

**Directions**: Calculate the following alignment identity values and place them into the Distance Matrix below.

Seq A MGDVEKGKKIFVMKCSQCHTVEKGG

Seq C MVDVEKGMLIFVMKCSQCHTVEAGG

Alignment Value \_\_\_0.16\_\_\_\_\_\_\_

Seq A MGDVEKGKKIFVMKCSQCHTVEKGG

Seq D MGDIEKGKLIFVMSCSQCHTVYAGY

Alignment Value \_\_\_0.24\_\_\_\_\_\_\_

Seq A MGDVEKGKKIFVMKCSQCHTVEKGG

Seq E MIDTEKGYIIFVMTCSQCHTVLMGT

Alignment Value \_\_\_\_0.32\_\_\_\_\_\_

Seq B MVDVEKGMKIFVMKCSQCHTVEAGG

Seq C MVDVEKGMLIFVMKCSQCHTVEAGG

Alignment Value \_\_\_0.04\_\_\_\_\_\_\_

Seq B MVDVEKGMKIFVMKCSQCHTVEAGG

Seq D MGDIEKGKLIFVMSCSQCHTVYAGY

Alignment Value \_\_\_0.28\_\_\_\_\_\_\_

Seq B MVDVEKGMKIFVMKCSQCHTVEAGG

Seq E MIDTEKGYIIFVMTCSQCHTVLMGT

Alignment Value \_\_\_\_0.32\_\_\_\_\_\_

Seq C MVDVEKGMLIFVMKCSQCHTVEAGG

Seq D MGDIEKGKLIFVMSCSQCHTVYAGY

Alignment Value \_\_\_0.24\_\_\_\_\_\_\_

Seq C MVDVEKGMLIFVMKCSQCHTVEAGG

Seq E MIDTEKGYIIFVMTCSQCHTVLMGT

Alignment Value \_\_\_0.32\_\_\_\_\_\_\_

Seq E MIDTEKGYIIFVMTCSQCHTVLMGT

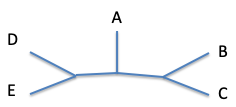
Seq D MGDIEKGKLIFVMSCSQCHTVYAGY

Alignment Value \_\_\_\_0.32\_\_\_\_\_\_

**Distance Matrix**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Seq A | Seq B | Seq C | Seq D | Seq E |
| Seq A | - | - | - | - | - |
| Seq B | 0.12 | - | - | - | - |
| Seq C | 0.16 | 0.04 | - | - | - |
| Seq D | 0.24 | 0.28 | 0.24 | - | - |
| Seq E | 0.32 | 0.32 | 0.32 | 0.32 | - |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Seq A | **Seq B/C** | Seq D | Seq E |
| Seq A | - | - | - | - |
| **Seq B/C** | **0.14** | - | - | - |
| Seq D | 0.24 | **0.26** | - | - |
| Seq E | **0.32** | **.32** | 0.32 | - |



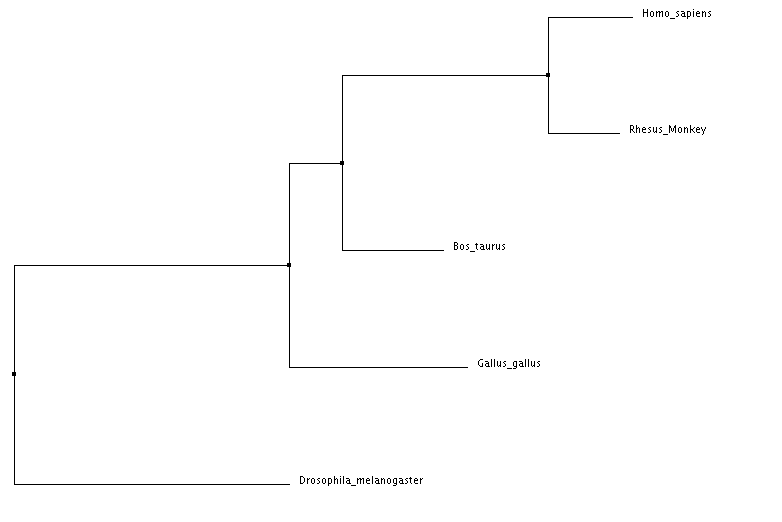
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *H. sapiens* | *M. mulatta* | *B. taurus* | *G. gallus* | *D. melanogaster* |
| *H. sapiens* | - | - | - | - | - |
| *M. mulatta* | 0.038 | - | - | - | - |
| *B. taurus* | 0.095 | 0.095 | - | - | - |
| *G. gallus* | 0.124 | 0.124 | 0.086 | - | - |
| *D. melanogaster* | 0.216 | 0.170 | 0.157 | 0.167 | - |

First Round

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *H. sapiens/ M. mulatta* | *B. taurus* | *G. gallus* | *D. melanogaster* |
| *H. sapiens/ M. mulatta* | - | - | - | - |
| *B. taurus* | 0.095 | - | - | - |
| *G. gallus* | 0.124 | .086 | - | - |
| *D. melanogaster* | 0.193 | .157 | .167 | - |

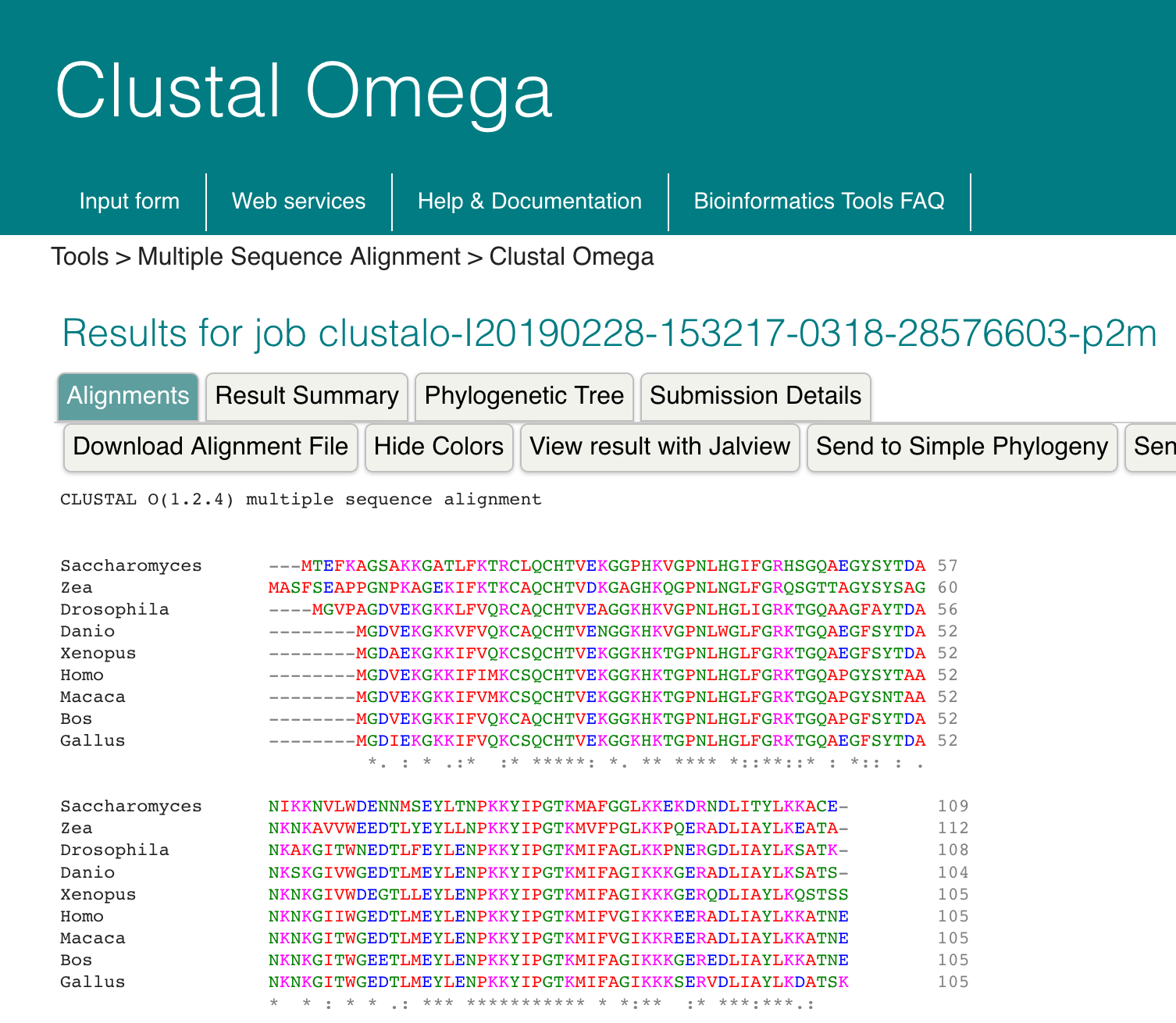
|  |  |  |  |
| --- | --- | --- | --- |
|  | *H. sapiens/ M. mulatta* | *B. Taurus/ G. gallus* | *D. melanogaster* |
| *H. sapiens/ M. mulatta* | - | - | - |
| *B. Taurus/ G. gallus* | 0.109 | - | - |
| *D. melanogaster* | 0.193 | 0.162 | - |

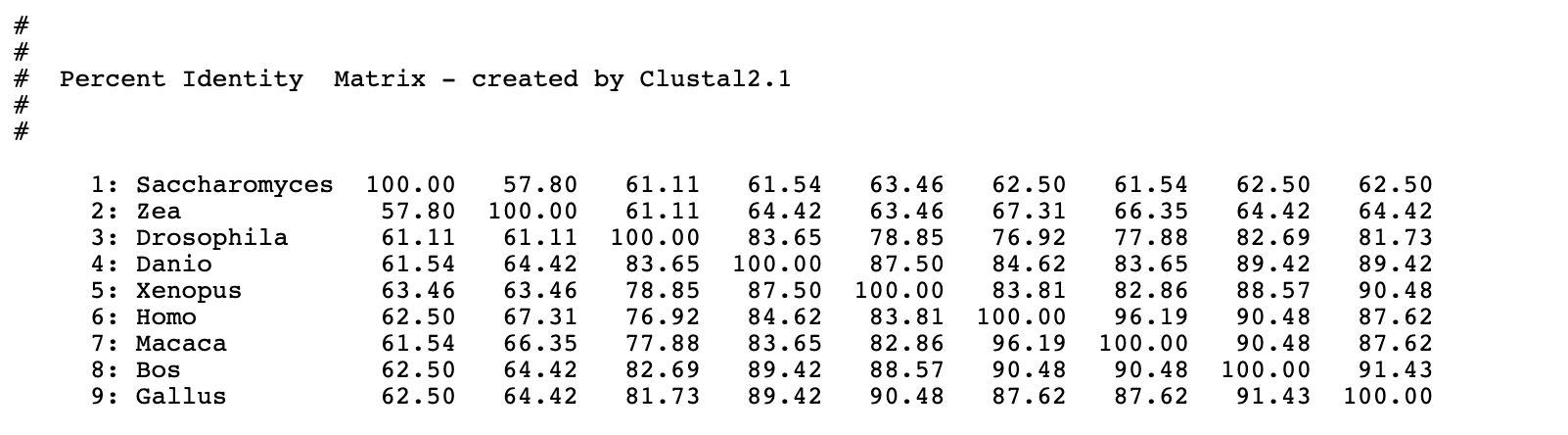
|  |  |  |
| --- | --- | --- |
|  | *H. sapiens/ M. mulatta/ B. Taurus/ G. gallus* | *D. melanogaster* |
| *H. sapiens/ M. mulatta/ B. Taurus/ G. gallus* | - | - |
| *D. melanogaster* | 0.177 | - |

****

What species in your list of sequences derived from HomoloGene could you utilize as an “outgroup” if you wanted to build a rooted tree? **(2)**

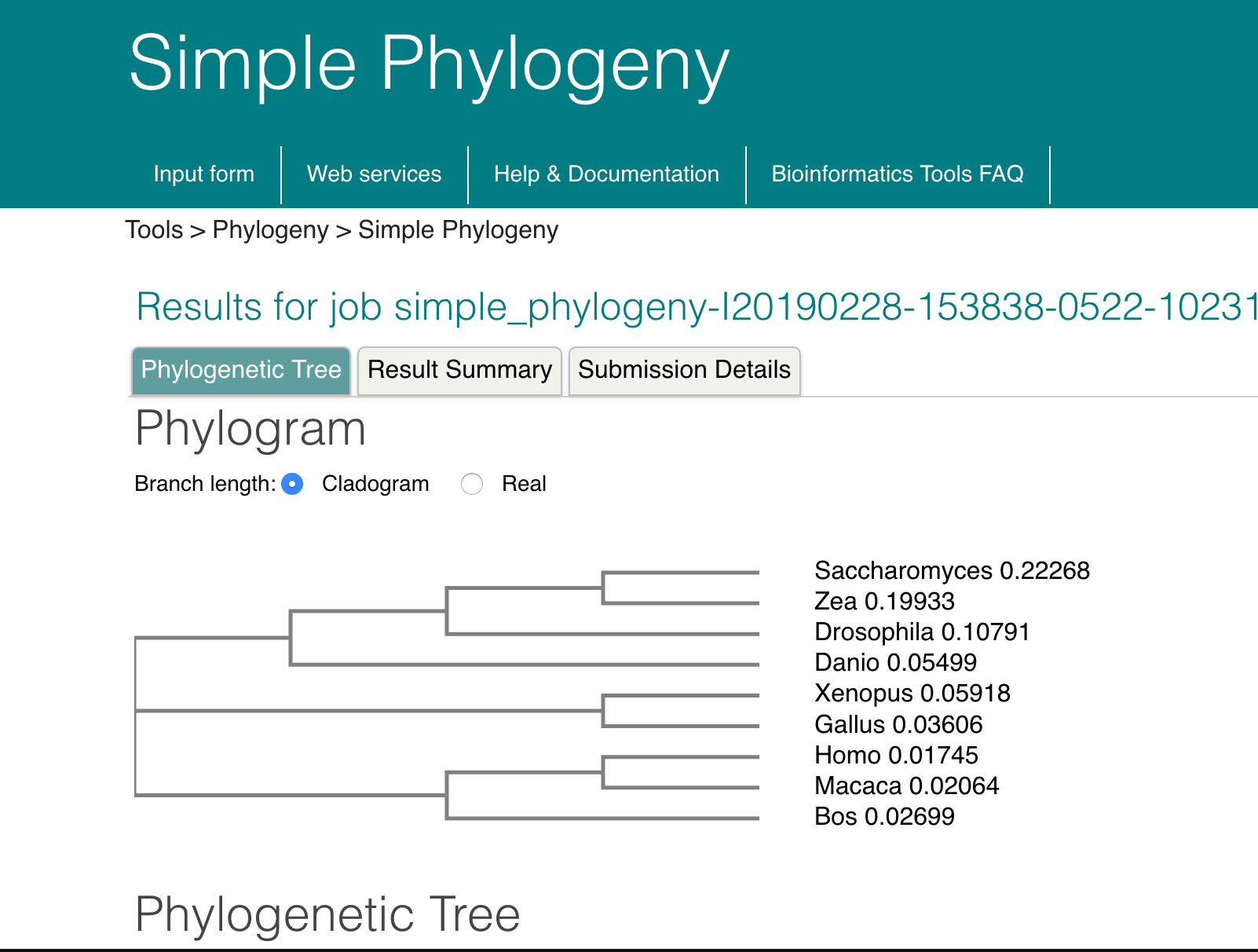
Use *Saccharomyces cerevisiae*

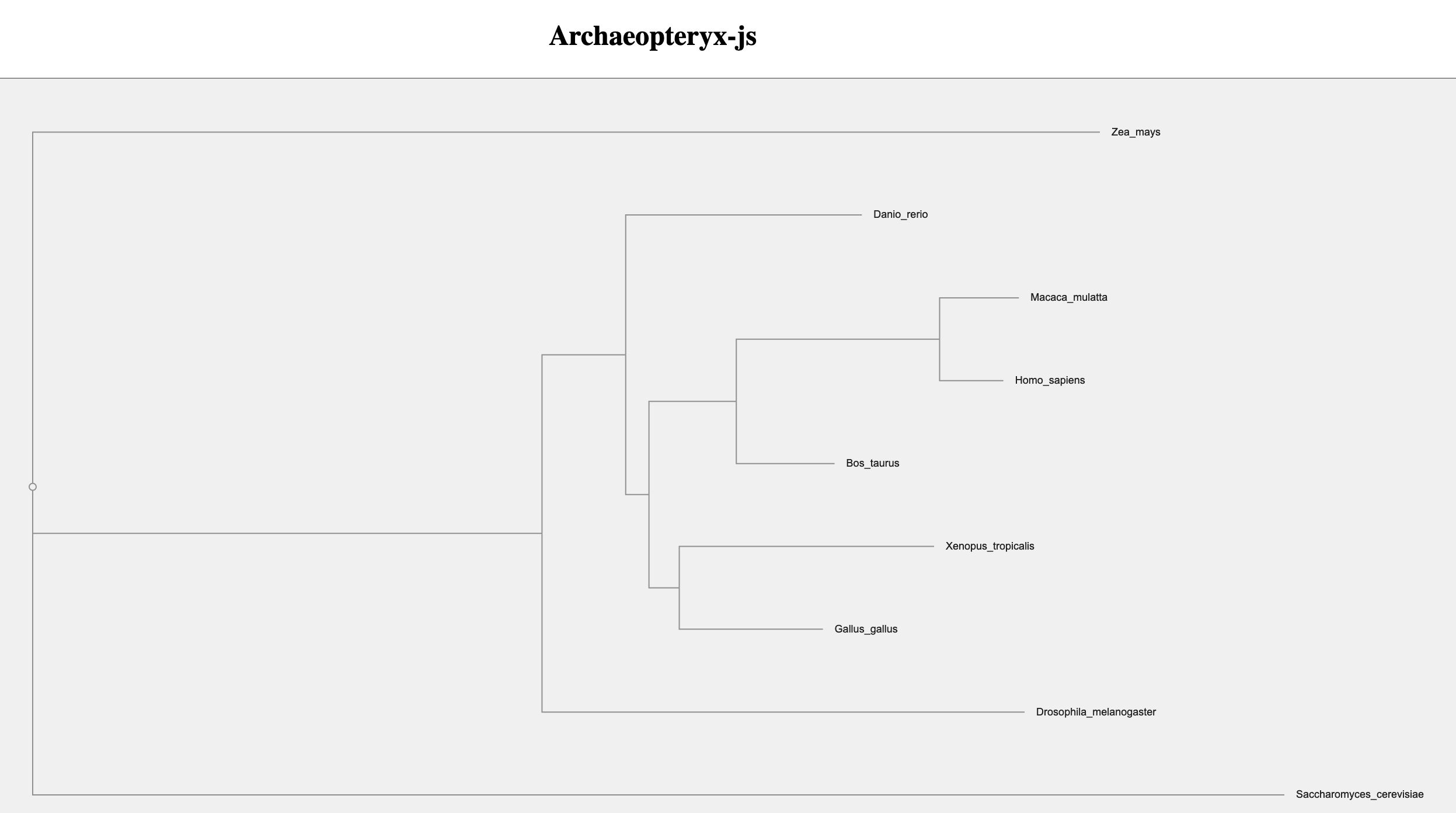
**

**

Using the percent identity matrix, which two organisms have the least evolutionary time separating them based on the molecular data analyzed? **(5)**

Homo and Macaca (human and rhesus macaque) with 96% identity have the least evolutionary time separating them.





Based on your knowledge of eukaryotic organisms, does your phylogram represent how you would predict the relative evolutionary relationships between the organisms to be? Explain. **(7)**

**Not completely, I was expecting humans to be most closely related to macaque, but I would expect humans to be more closely related to cows.**

**I am also surprised to see the close relationship between chickens and frogs.**

What is meant by homology when applied to organisms or genetic sequences?

If two genetic sequences are similar, are they homologous?

If two sequences are homologous, will they be similar?

Thus, if a query sequence returns a long list of small E value hits that correspond to a described gene in several different species, it is likely that you have identified a homologous sequence.