## Part 1: Find and download the hexokinase sequences [3 pts total]

Using NCBI, search the Gene database for the human (*Homo sapiens*) Hexokinase 1 (HK1) RefSeq gene (use the filter "seq refseqgene"). Get the protein sequence (for the first isoform only), and download it to a FASTA file. Next, get the corresponding protein sequences for its **orthologs** in dogs (*Canis lupus familiaris*) and crocodiles (*Crocodylus porosus*). In each case, you only need to download the sequence for the first isoform listed for each protein.

Do the same thing for Hexokinase 2 (HK2). Once you are done, you should have a total of 6 FASTA files, each containing one protein sequence.

#### Question 1A-F: What is the accession number for each of the sequences you downloaded?

This is the unique database ID that NCBI assigns to all of its sequence entries. You can find it on the information page for each protein, or look at the headers in your FASTA files, where each sequence will be named by its accession number *(0.5 pts for each sequence)*.

#### 1A. Human HK 1

NP\_000179.2

#### 1B. Dog HK 1

XP\_038518713.1

#### 1C. Crocodile HK1

XP\_019394065.1

#### 1D. Human HK 2

NP\_000180.2

#### 1E. Dog HK 2

XP\_038547540.1

#### 1F. Crocodile HK 2

XP\_019401522.1

## Part 2: Pairwise Alignments of HK1 with Different Parameters [7 pts total]

Align the Human and Dog HK1 sequences using different options for the Scoring Matrix and the Gap Penalties. Use the web-based tool called EMBOSS Needle from EBI: <https://www.ebi.ac.uk/Tools/psa/emboss_needle/>

[Links to an external site.](https://www.ebi.ac.uk/Tools/psa/emboss_needle/)

You can either copy and paste your sequences into the page, or you can upload them as files. Make sure that you specify that you are running protein sequences! To see and/or change the parameters for the scoring matrices and the gap penalties, click the "More Options" button.

#### 2A. Run the alignment using the default parameters. What scoring matrix is being used? What are the gap open and gap extend penalties? For the resulting alignment, what is the percent identity? What is the percent similarity? How many gaps are there? *(1 pt*)

BLOSUM62  
Gap open: 10  
Gap Extend: 0.5  
Percent Identity: 93.2%  
Percent Similarity: 96%  
Gaps: 8 gaps

#### 2B. Re-do the alignment with the PAM10 matrix. Did this change your results from part A? In what way did the results change? *(1 pt*)

Yes, this changed the results from part A.

Percent Identity: 91.7%

Percent Similarity: 91.7%  
Gaps: 38 gaps

Identity and similarity percentages decreased with an increase in the number of gaps.

#### 2C. Re-do the alignment again, now with the PAM250 matrix. Did this seem to do better or worse than the alignment with the PAM10 matrix? How do these results compare to the results with the default parameters? *(1 pt)*

PAM250 seemed to do better than PAM10.

Percent Identity: 92.4%

Percent Similarity: 96.2%  
Gaps: 26 gaps

There was an increase in similarity and identity, and a decrease in the number of gaps.

#### 2D. Re-do the alignment one more time using the PAM250 matrix. Set the gap open and gap extend penalties to the highest possible values. Now, what does your alignment look like in terms of identity, similarity, and number of gaps? *(1 pt)*

It seemed to do much better than Part C.

Percent Identity: 93.2%

Percent Similarity: 97.0%  
Gaps: 4 gaps

There was a slight increase in similarity and identity, and a decrease in the number of gaps.

#### 2E. Why do you think lowering the number of the PAM matrix had the effect that it did on the alignments? *(3 pts)*

#### 

#### PAM10, when compared to PAM250, has a smaller percentage identity, and percentage similarity. This is due to a shorter time span under observation in PAM10. PAM250 has had more time to analyze the HK1 protein in humans and dogs to genetically diverge.

## Part 3: Use the Cluster to Align HK2 [1 pt total]

Again use EMBOSS Needle to align your sequences, but this time use the cluster to run the alignment rather than the web application.

Start by uploading your 2 FASTA files (for the Dog and Human HK2 sequences) to the cluster.

Next, create a slurm script that will load the emboss module, run the needle program with default parameters, and save the results to a new file. Note that this will run very quickly, so you do not need much time or memory in your job script. You can use the code below as a guide for your own script; just be sure to change the names of the FASTA files to match your own file names!

#!/bin/bash

#SBATCH --partition=Centaurus

#SBATCH --job-name=needle

#SBATCH --nodes=1

#SBATCH --ntasks-per-node=1

#SBATCH --time=0:30:00

module load emboss

needle -auto \

-asequence dog2.fasta -sformat1 fasta -sprotein1 Y \

-bsequence human2.fasta -sformat2 fasta -sprotein2 Y \

-datafile EBLOSUM62 -gapopen 10 -gapextend 0.5 \

-outfile HK2\_alignment

#### Question 3: What does the alignment of Human and Dog HK2 sequences look like? How does it compare to the HK1 alignment with the same parameters? *(1 pt)*

Percent Identity: 95.6%

Percent Similarity: 97.7%

Gaps: 0

HK2’s pairwise sequence alignment was much higher in identity and similarity percentages with less gaps.

## Part 4: Alignment with a More Distant Ortholog [2 pts total]

Perform alignments of the Human and Crocodile orthologs for both the HK1 and HK2 sequences to see how well conserved they are across a much deeper time scale. Use the default scoring matrix and gap parameters for Emboss Needle. You can do this either on the cluster or on the web.

#### Question 4: How do the identity, similarity, and gap scores compare to the Human/Dog alignments?

#### 4A. For HK1? *(1 pt)*

Percent Identity: 87.8%

Percent Similarity: 94.9%

Gaps: 0

These were lower in identity and similarity percentages, but have a lower number of gaps.

#### 4B. For HK2? *(1 pt)*

Percent Identity: 88.5%

Percent Similarity: 95.5%

Gaps: 0

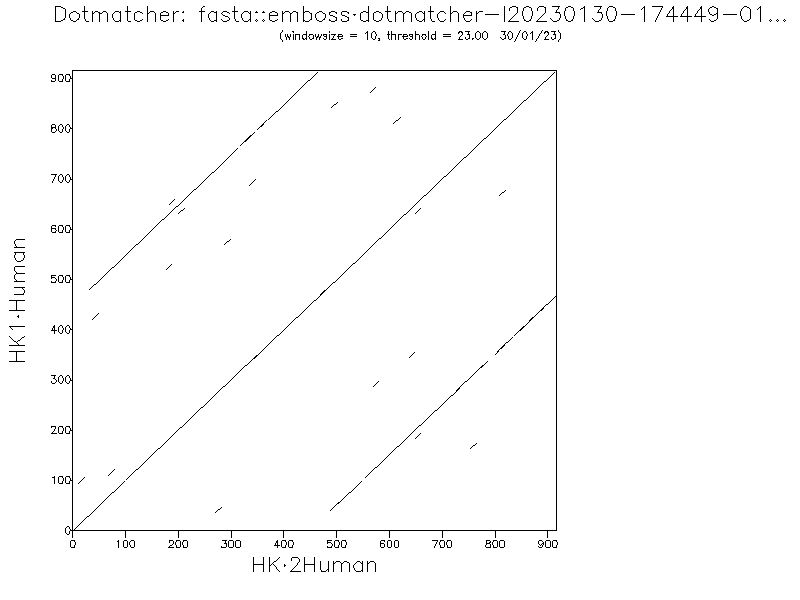
The human crocodile alignments both scored lower in both percent identity and similarity. Both had 0 gaps.

## Part 5: Dot-plot Comparison of HK1 and HK2 [2 pts total]

Use the EMBOSS Dotmatcher tool to create a dot-plot for Human HK1 and Human HK2. You can find the tool online here: <https://www.ebi.ac.uk/Tools/seqstats/emboss_dotmatcher/>

[Links to an external site.](https://www.ebi.ac.uk/Tools/seqstats/emboss_dotmatcher/)

#### 5A. Show your resulting dot-plot. *(1 pt)*



#### 5B. What kind of macro-scale genomic pattern or mutation is indicated by your dot-plot? *(1 pt)*

The pattern displayed by this dot plot can be interpreted as “repeats”.