HW 5

Mathematics 127
Mathematical and Computational Methods in Molecular Biology

Fall 2002
UC Berkeley, CA

Nasser M. Abbasi

Fall 2002 Compiled on August 2, 2022 at 8:11am [public]

Contents

1 Problems

Problem Set 5 (due Tuesday November 26)
MATH 127: Mathematical and Computational Methods in Molecular Biology

Problem 1

a) Consider a tree with n leaves and internal nodes all of which have degree k. How many edges are there in the tree?

b) How many **unlabelled** treees are there on 2,3,4 vertices? Conjecture a formula for the number of unlabelled trees on n vertices.

Problem 2

Go to `http://www.ch.embnet.org/software/ClustalW.html`. Enter the sequences:

GATTACA
AGAGACGATGA
GAGAAGGGAAGGAATTACA
GATATATGCA
GAGAGTG

Align the sequences and look at the output (clustalw aln format). How sensitive is the multiple alignment to the extension and separation gap penalties?

Problem 3^*

Consider the dynamic programming method for aligning k sequences of length n (generalization of Needleman-Wunsch). The divide-and-conquer Needleman-Wunsch algorithm can be generalized to the k sequences problem. What is its running time and space requirement?

Problem 4

a) Find the accession AC129884 at NCBI. What organism is this sequence from? How many pieces is it in?

b) Go to `http://pipeline.lbl.gov/cgi-bin/GenomeVista`. Put in the GENBANK accession above, and find it on the **Mouse Genome**

c) What genes does the sequence contain? (it may help to click on TextBrowser and then look at the Vista picture)

d) Go to

`http://www.nisc.nih.gov/open_page.html?/projects/zooseq/pubmap/PubZooSeq_Target`

Which target region contains the gene of AC129884? Click on that target. How many organisms have sequence available in GENBANK?

Optional Problem Prove the formula conjectured in 1b for the number of unlabelled trees on n vertices.

2 Problem 2

Problem 2
HW 5.
Math 127, UC Berkeley, Fall 2002.
Nasser Abbasi

Went to

http://www.ch.embnet.org/software/ClustalW.html

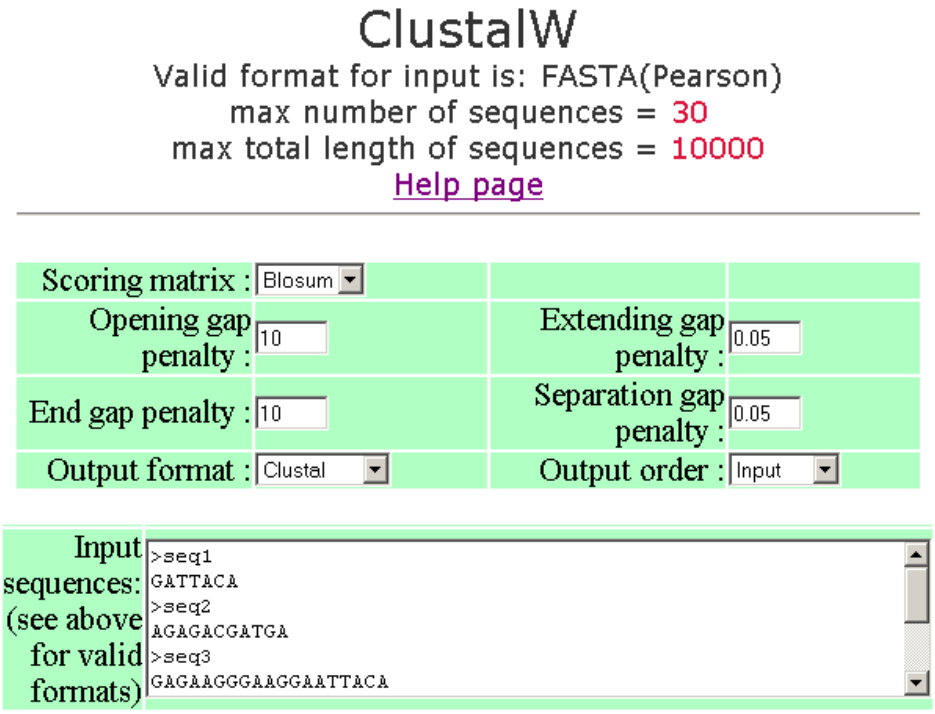
and used the UI to enters the 5 sequences.

for help on ClustalW, I found this site is useful:
http://www.swbic.org/origin/proc_man/Clustal/search/help.html

Before I show the results, first I needed to better understand the parameters and what they actually mean. From the above help, this the summary:

Open gap penalty: Increasing the gap opening penalty *will make gaps less frequent*.
Extend gap penalty: Increasing the gap extension penalty *will make gaps shorter*.
Separation gape penalty: This is the same as gap distance. From the net, I found this definition:
"GAP SEPARATION DISTANCE tries to decrease the chances of gaps being too close to each other. Gaps that are less than this distance apart are penalised more than other gaps. *This does not prevent close gaps; it makes them less frequent*, promoting a block-like appearance of the alignment"

So, the above tells me that if I increase the gap separation penalty, I should see gaps more far apart.



I started by fixing the value of the extend gap penalty and changing the separation gap penalty. Then fixed the separation gap penalty and changed the extend gap penalty. Then changed both at the same time. These are the result of these trials:

Tries done to see the effect of changing the gap separation penalty:

Extend Gap	Separation Gap	Multiple alignment	observation
0.05	0.0	seq1 ----------GATTACA seq2 AGAGACG-------ATGA-- seq3 -GAGAAGGGAAGGAATTACA seq4 -GATAT--------ATGCA- seq5 -GAGAG---------TG--- *	Original default setting
0.05	0.01	same as above	No change seen
0.05	0.07	same as above	No change seen
0.05	0.08	same as above	No change seen
0.05	0.09	same as above	No change seen
0.05	0.1	same as above	No changes seen
0.05	1.0	same as above	No changes seen
0.05	2.0	same as above	No changes seen
0.05	3.0	same as above	No changes seen
0.05	5.0	same as above	No changes seen
0.05	20	same as above	No changes seen
0.05	200	same as above	No changes seen

Conclusion: In the above sequences, the gap separation penalty have no effect. This shows ClusalW is not sensitive to this penalty, at least in this example.

Tries done to see the effect of changing the gap extension penalty:

Extend Gap	Separation Gap	Multiple alignment	observation
0.05	0.06	seq1 ----------GATTACA seq2 AGAGACG-------ATGA-- seq3 -GAGAAGGGAAGGAATTACA seq4 -GATAT--------ATGCA- seq5 -GAGAG---------TG--- *	Original default setting
0.1	0.06	same as above	No changes seen
0.2	0.06	same as above	No changes seen
0.3	0.06	same as above	No changes seen
0.4	0.06	same as above	No changes seen
0.5	0.06	same as above	No changes seen
0.51	0.06	same as above	No changes seen
0.52	0.06	same as above	No changes seen
0.53	0.06	Same as above	No changes seen
0.530000001	0.06	Same as above	No changes seen
0.530000002	0.06	seq1 ----------GATTACA seq2 ------AGAGACGATGA-- seq3 GAGAAGGGAAGGAATTACA seq4 -------GATAT-ATGCA- seq5 -------GAGAG--TG--- *	A tiny change in the extend gap penealty now shows large effect for first time. First gape on seq3 is gone, and gaps inside seq 2,4,5 are gone. GAPS HAVE BECOME SHORTER AS EXPECTED.
0.54	0.6	Same as above	No changes seen
1.0	0.6	Same as above	No changes seen
2.0	0.6	Same as above	No changes seen
8.99999952	0.6	Same as above	No changes seen
8.99999953	0.6	seq1 ----------GATTACA seq2 ------AGAGACGATGA-- seq3 GAGAAGGGAAGGAATTACA seq4 -------GATATATGCA-- seq5 -----GAGAGTG-------	A tiny change now shows another change. Now all internal gaps are gone. GAPS HAVE BECOME SHORTER AS EXPECTED.
100	0.6	Same as above	No changes seen

Conclusion: ClustalW is more sensitive to gap extension penalty. The larger this penalty, the less gaps are seen inside the sequences as expected. It is very sensitive in that a change from 0.530000001 to 0.530000002 (a change on only 0.000000001) causes such a large effect in the alignment as shown above. As the penalty is increased all the way to 8.99999952 no more change is seen. But a change from 8.99999952 to 8.99999953 caused the final gap inside the last 2 sequences to close.

3 Problem 4

Problem 4
HW 5
Math 127, UC Berkeley, Fall 2002.
Nasser Abbasi

Part a). from the NCBI web page, AC129884 sequence is from organism Ornithorhynchus anatinus. *Genbank common name:* platypus (to be honest, I do not know what this organism is supposed to be, I just got the name from the NCBI default display for this locus). The length of the sequence is 121,483 bp

For the number of pieces this sequences is made of, I looked down the description, and in the comment section it says that this sequence is a working draft, **and it is made of 7 contigs.**

* 1 4203: contig of 4,203 bp in length
* 4304 15192: contig of 10,889 bp in length
* 15293 24847: contig of 9,555 bp in length
* 24948 34501: contig of 9,554 bp in length
* 34602 51540: contig of 16,939 bp in length
* 51641 76311: contig of 24,671 bp in length
* 76412 121483: contig of 45,072 bp in length.

Part b)

Went to http://pipeline.lbl.gov/cgi-bin/GenomeVista .

First I needed to understand what GenomeVista does. This is below the description from the above web page:

> GenomeVista allows users to perform comparative analysis of their own data sets using the Berkeley Genome Pipeline (Godzilla). The **draft** or **finished** sequences are aligned with the base genome of your choice, and conserved region analysis is performed. The resulting alignments can be browsed via the Vista Genome Browser or the Godzilla Text Browser.

So, GenomeVista locates a sequence on either the human or the mouse genome. The question asks to find this sequence on the mouse genome. So, I set the 'Base Genome' choice to 'Mouse feb 2002' and and entered the above accession number. This is the result:

Chromosome 6
Total Groups: 1 (sorted by alignment size)
6:17409146-17533827 (2 alignments, 121.8Kbp) Text Browser Vista Genome Browser

Godzilla Text Browser

Select Genome Pair: User request - Mouse Feb. 2002
Position in the Base Genome: chr11:113030619-113173035 Go
(Format: chr11:113030619-113173035)

The above result is a little confusing to me. At the top it says that 2 alignments found on Chromosome 6 of the mouse genome. But in the lower part under the text browser, it lists a position in Chromosome 11. I assume this is just to show how the format looks like. I.e. it is an example. (but it should actually say so).

So, I clicked on the 'Text Brower' to see where on Chr 6 these sequences found. And this is the result.

Hits on chr6:17409146-17533827
RefSeq in this region **View in Vista Browser** **View at UCSC** Get conserved regions

user query Contig info	Location on mouse	matches number of matches
AC129884-7 (user scontig) Contig Sequence length = 45072bp **aligned**: between 3401-43318 (39918bp)	**chr6:17409146-17454346** Sequence (softmasked) RefSeq Conserved Regions length=45201bp	12337
AC129884-6 (user scontig) Contig Sequence length = 24671bp **aligned**: between 51566-73595 (22030bp) on the reverse complement	**chr6:17517206-17533827** Sequence (softmasked) RefSeq Conserved Regions length=16622bp	6218

From the above, these are the locations of the sequence on mouse genome:

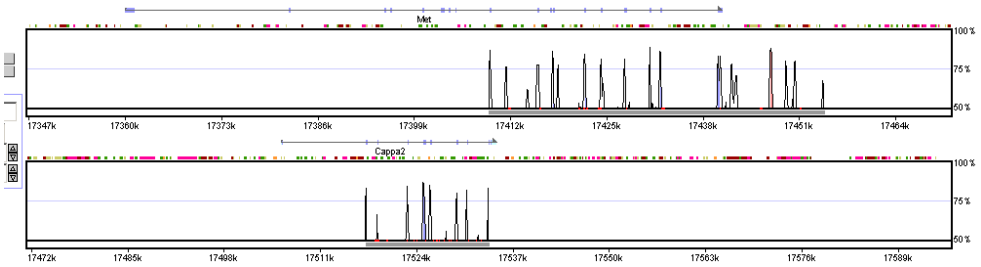
chr6:17409146-17454346 length=45,201bp
chr6:17517206-17533827 length=16,622bp

Looking at the NCBI output, I see that the above alignment seem to have been made on contigs 5 and 7 of the sequence, because those are the lengths closes to result from GenomeVista. (The original sequence had a length of 121,483, but the locations found have smaller lengths to them, so that is why I assume that the alignments was made on the whole sequence but only pieces 5 and 7 were found on the mouse genome).

Part c)
To find what genes in this sequence, I searched in both the mouse genome and the human genome.

For the mouse genome:

Clicked on the TextBrower, then for each alignment (there are 2 of them as shown above), I click on the 'Vista' link to the right of the screen. To see both alignments, I zoomed out. This is the result



In the above, the first alignment (the first window above) is contig **chr6:17409146-17454346** length=45,201bp While the second alignment (in the lower window) is contig **chr6:17517206-17533827** length=16,622bp

The question asks to find the genes contained in the sequence.

Clicked on the 'VISTA' link, this shows this result (the vista plot shows the gene name on top of the diagram on the arrow line).

Contig	Gene
chr6:17409146-17454346	MEL
chr6:17517206-17533827	CAPPA2

For the human genome, similarly, the following genes found (there were 3 contigs found when search human genome june 2002).

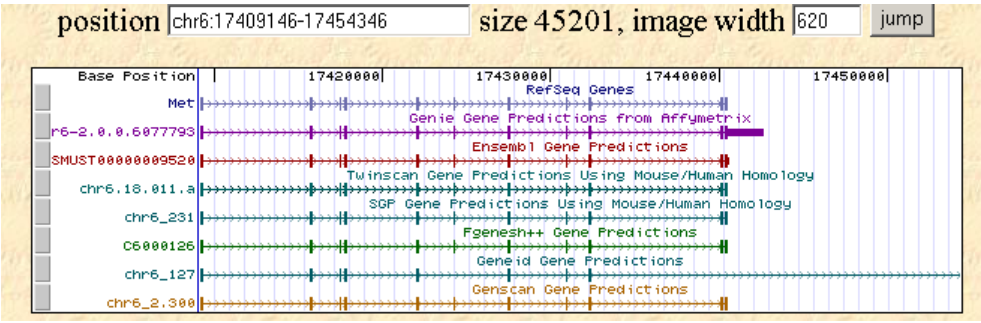
Contig	Gene
chr7:114880956-114924418	MET
chr7:114927608-114928484	NO GENE FOUND
chr7:115006025-115034040	CAPZA2

So, the answer to part C is: Mei, Met, Cappa2, and Capza2. 4 genes, 2 in mouse genome and 2 in human genome.

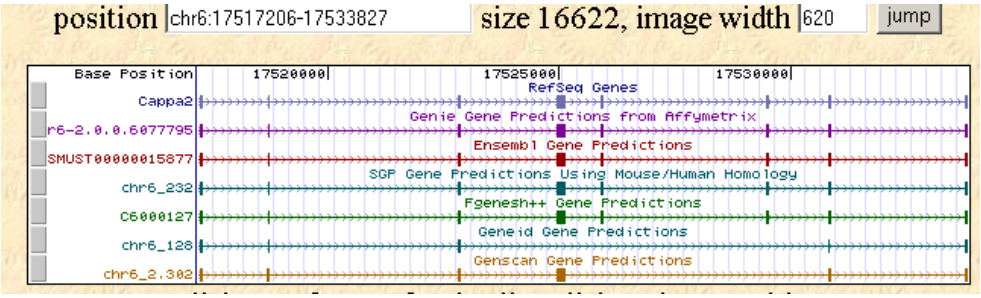
(note: It was hard to read the names of the genes on top of the diagrams in the vista window, but I zoomed in to verify that the names are correct as above).

While working on this part, I show the result of gene prediction using the UCSD software. This below is the output when I hide everything expect the gene predictions from a number of applications.

I clicked on the UCSD brower and in the 'position' window, I typed in **chr6:17409146-17454346** (which is the first contig) which contains gene Mei:



This is for the second contig which contains gene Cappa2:



part d)
went to
http://www.nisc.nih.gov/open_page.html?/projects/zooseq/pubmap/PubZooSeq_Targets

For MET gene, it is contained in target 1
For MEL gene, it is contained in target 1
For CAPPA2, it is contained in target 1
For CAPZA2, it is contained in target 1.

So the answer to part d is target 1.

Target 1 is about 1.5 Mbases. Organisms shown are : Chimp, Orangutan, Baboon, Macaque, Vervet, Lemur, Pig, Horse, Cow, Cat, Dog, Ajbat, Cpbat, Rabbit, Hedgehog, Mouse, Rat, Opossum, Dunnart, Platypus, Chicken, Zebrafish, Fugu, Tetraodon.

24 organisms.