

Homework 1

Due in class, Wednesday, September 21, 2005

A general note about homework: we feel that as much as possible, learning should be a collaborative experience. To that end, we welcome you – and in fact, encourage you – to work together with your peers to make sure you all understand the material thoroughly. However, we do need to enforce some set of guidelines to ensure that your submitted assignment actually reflects what you have learned. Thus, for this and all future homework assignments, be they pencil-paper questions or Web-based questions, please adhere to the following:

1. You may discuss freely with anyone, or refer to any printed or online resource about any aspect of the course **not directly pertaining to the homework questions**. For example, if you want to ask a postdoc in your lab how to do the traceback for a dynamic programming alignment with affine gap penalties, that's fine. Or if you want to Google for sites that talk about when to use PAM versus BLOSUM, that's also fine. However, if we want you to run BLAST to find the orthologs of a particular gene, we don't want you to go looking for a review paper in PubMed that lists them all, or have your PI point them out for you.
2. You **may** discuss homework questions with anybody currently enrolled in the class. However, when it comes to writing up answers, you must do this individually. For pencil-paper questions, you should write answers in your own words, without referring to any notes you may have taken during a group homework session. For Web-based questions, you should do all the mouse clicking etc. on your own.
3. You should cite any help that you give or receive. E.g., if you discussed problem 1 with George Church, Craig Venter, and Sam Karlin (assuming that they are taking GCB 535), write on your homework "I discussed problem 1 with George Church, Craig Venter, and Sam Karlin."
4. The professors and TA are always available for help and clarification. Use your best judgement – remember that the homeworks are meant to give you practice using bioinformatics tools and to make sure you understand the theory behind them. If you are unclear about any particular example or the policy in general, come talk to Prof. Ungar or the TA.

* * * * *

This week, all of the homework questions are pencil-paper questions; this mostly has to do with the material that we've covered so far. Future homeworks will be much more hands-on. For questions that require an explanation, try to be as concise and precise as possible. Wordy answers will generally be looked upon less favorably than ones that are to the point.

0. Algorithms

Write two algorithms that iterate over every index from $(1,1,\dots,1)$ to (n_1, n_2, \dots, n_d) . Make one algorithm iterative and the other recursive. (10 points)

I.e., if $d=2$, $n_1=2$ and $n_2=3$ the algorithms should generate $(1,1), (1,2), (1,3), (2,1), (2,2), (2,3)$
- not necessarily in that order.

1. Doing alignments by hand

a. Given two strings, **TCTCAC** and **TTAC**, perform Needleman-Wunsch global alignment using the following parameters: match = +2, mismatch = -1, gap = -1. Show the score of every square and clearly mark the backpointers. In the event that there exist two (or more) predecessor squares yielding the same score, draw both (all) backpointers. Clearly indicate which square(s) to start from to obtain an optimal traceback, and report the optimal alignment(s). (5 points)

b. Perform Smith-Waterman local alignment using the same strings and parameters as in part (a). Report the two highest-scoring local alignments. (5 points)

[Note: please do appreciate how tedious it is to do this sort of work by hand. Now imagine doing it with ESTs (100s of nucleotides long) or cDNAs (1000s of nucleotides long) or genes (10s of 1000s of nucleotides long) or chromosomes (really really long).]

2. Needleman-Wunsch parameters

You are trying to align a DNA fragment with sequence **TCGTTGCT** to two other DNA fragments, **TCGTAAAAAATGCT** and **TACAGATATAGACAT**. You expect that the alignments will look like this:

```
TCGT      TGCT | T C G T T G C T |
||||      |||| | | | | | | | | |
TCGTAAAAAATGCT | and TACAGATATAGACAT |
```

a. Give a set of Needleman-Wunsch parameters (match, mismatch, gap open, gap extend) such that the two alignments shown above would yield identical, optimal scores, and there exist no other optimal alignments. The actual values of the parameters aren't important as long as the relationship between them is correct. You are welcome to test your parameters by actually performing the alignment, but you shouldn't need to. (3 points)

b. Give a set of Needleman-Wunsch parameters such that the first alignment is optimal and yields a higher score than the second alignment. (3 points)

c. Why might you prefer one alignment over another? Try to give a biological

explanation for why one of the alignments may be “more correct” than the other. (3 points)

3. Aligning orthologs

Circadian rhythmicity of biologic processes is a fundamental property of all eukaryotic and some prokaryotic organisms. These rhythms are driven by an internal time-keeping system. Changes in the external environment, particularly in the light-dark cycle, entrain this biologic clock. Under constant environmental conditions devoid of time cues, rhythms driven by the biologic clock show a period near, but usually not equal to, 24 hours. The bilaterally paired suprachiasmatic nuclei (SCN) of the hypothalamus are thought to contain the master circadian clock that regulates most, if not all, circadian rhythms in mammals. The CLOCK gene encodes a basic helix-loop-helix (bHLH)-PAS transcription factor that is essential for circadian rhythm. Polymorphisms within the encoded protein have been associated with circadian rhythm sleep disorders.

A student from Dr. Mimbo’s lab has pulled out the amino-acid sequences of CLOCK for homo sapiens and drosophila meganlastor. He did both Needleman-Wunsch and Smith-Waterman alignment on these two amino-acid sequences using the default parameters. The results are shown below.

EMBOSS-Align Results

Needle Results	
Matrix	Blosum62
Open gap penalty	10.0
Gap extension penalty	0.5
Needle output	needle-20050914-03051425313837.output
<input type="button" value="SUBMIT ANOTHER JOB"/>	

```
#####
# Program: needle
# Rundate: Wed Sep 14 03:03:33 2005
# Align_format: srspair
# Report_file: /ebi/extserv/old-work/needle-20050914-03051425313837.output
#####

#=====
#
# Aligned_sequences: 2
# 1: NP_004889.1
# 2: NP_001014576.1
# Matrix: EBLOSUM62
```

```

# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1146
# Identity:      310/1146 (27.1%)
# Similarity:   433/1146 (37.8%)
# Gaps:         419/1146 (36.6%)
# Score: 1163.5
#
#
#=====

```

```

NP_004889.1      1 MLFTVSCSKMSSIIVDRDDSSIFDGLVEEDDKDKAK----RVS RNKSEKKR      46
                          .||         |.||||..|         |.||||.|||||
NP_001014576.    1                      MDD-----ESDDKDDTKSFLCRKSRNLSEKKR      27

NP_004889.1     47 RDQFNVLIKELGSMPLPGNARKMDKSTVLQKSIDFLRKHKKEITAQSDASEI      96
|||.|.:.:|.:.:.:|.|.|.|.|.|.|.|.|.:.:|.|.|.|.|.|.:.|.|.|.
NP_001014576.   28 RDQFNSLVNDLSALISTSSRKMDKSTVLKSTIAFLKNHNEATDRSKVFEI      77

NP_004889.1     97 RQDWKPTFLSNEEFTQLMLEALDGGFFLAIMTDGSIIVSESVTSLEHLPL      146
:|||||.|||||:|:|.|||||:|||||.:.:.:|.|.|.||||:|.|.:||
NP_001014576.   78 QQDWKPAFLSNDEYTHLMLESLDGFMMVFSMGSIFYASESITSQLGYLP      127

NP_004889.1    147 SDLVDQSI FNFPIPEGEHSEVYKIL--STHLLLED SLTPEYLKSKNQLEFC      194
.|.:.:|.::.:.:|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
NP_001014576.  128 QDLYNMTIYDLAYEMDHEALLNIFMNPTPVIEPRQTD---ISSNQITFY      174

NP_004889.1    195 CHMLRGTIDPK EPSTY EYVKFIGNFKSLNSV----SSSAHNGFEG-----      235
.:|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
NP_001014576.  175 THLRRGM EKVDANAYELVKFVGYFRNDTNTSTGSSSEVSNGSNGQPAVL      224

NP_004889.1    236 --TIQRTHRPSYEDRVC FVATVRLATPQFIKEMCTVEEPNEEFTSRHSLE      283
..|.:.:.:.:.:|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
NP_001014576.  225 PRIFQQNPNAEVDK KLVFVGTGRVQNPQLIREMSIIDPTSNEFTSKHSME      274

NP_004889.1    284 WKFLFLDHRAPP IIGYLPFEVLG TSGYDYHVDLLENLAKCHEHLMQY GK      333
|||||.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.:
NP_001014576.  275 WKFLFLDHRAPP IIGYMPFEVLG TSGYDYHFDLDSIVACHEELRQTGE      324

NP_004889.1    334 GKSCYRFLTKGQQ WIWLQTHYYI TYHQWNSRPEFIVCTHTVVS YA EVRA      383
|||||.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
NP_001014576.  325 GKSCYRFLTKGQQ WIWLQTDYVVS YHQFNSKPDYV VCTHKVVS YA EVLK      374

NP_004889.1    384 ERRRE---LGIEESL PETAADK-----SQDSGS-----DNRINTVSL      417
.:|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
NP_001014576.  375 DSRKEGQKSGNSNS ITNNGSSKVIAS TGTSSKSASATTTLRDFELSSQNL      424

NP_004889.1    418 KEAL-----ERFDHSPT PSASS---RSSRKSSHTAVSDP      448
...|          ...|.||..|||:  ..|.:.:.:.|.|.
NP_001014576.  425 DSTLLGNSLASLGTETAATSPAVDSSPMWSASAVQPSGSCQINPLKTSRP      474

NP_004889.1    449 SSTPTKIPTDTSTPPRQHLP AHEKMQRRSSFSQSINSQSVGSSLTQP V      498
:|:..| :|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
NP_001014576.  475 ASSYGNU-SSTGISPK-----AKRKC YFYNNRGNDS DSTSMSTDSV      514

NP_004889.1    499 MSQATNLP IPIQGMSQFQFSAQLGAMQHLKDQLEQRTRMIEANIHRQQEEL      548
.:.:.:.:.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
NP_001014576.  515 TSRQSMTHVSSQSQRQRS-----H HREHRENHHN--QSHHHMQQQQQ      556

NP_004889.1    549 RKIQEQLQMVHGQGLQMFLQQS-----NPGLNFGSVQLSSGNSSNIQQ-      591
.:|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.
NP_001014576.  557 HQNQQQQHQQH-QQLQQQLQHTVGT PKMVPLLP IASTQIMAGNACQFPQP      605

```

```

NP_004889.1      592 -----LAPINMQGQ--- 600
                           |:|:|.|.
NP_001014576.   606 AYPLASPQLVAPTFLLEPPQYLTAIPMQPVIAPFPVAPVLSPLPVQSQDTM 655
NP_004889.1      601 -----VVPT-NQIQSGMNTGHIGT'QHMIIQQQTLQSTSTQSQQNVLSGH 643
                           :.|| :|:|...|...:..:| | . . . . | : : | | . :
NP_001014576.   656 LPDVTVMVTPTQSQLQDQLQRKHDELQKLILQQQ--NELRIVSEQLLLSRY 703
NP_004889.1      644 SQQTSLPSQ--TQSTLTAPLYNTMVISQPAAGSMVQIPSSMPQNSTQSAA 691
                           :...:.|. . . . : | | . . . . : . | . . . . : . | . : . . . |
NP_001014576.   704 TYLQPMMSMGFAPGNMTAAAVGNLQASGQRGLNFTGSNAVQPQFNQYGFA 753
NP_004889.1      692 VTT----FTQDRQIRFSQGGQLVTKLVTPVACGAVMPSTMLMGQVVTA 737
                           :.: . . | | : | : . . | . | . | :
NP_001014576.   754 LNSEQMLNQDQQMMMQQQLHTQ----- 778
NP_004889.1      738 YPTFATQQQQSQTLSVTVQQQQQSSQEQQLTQPPQPPQFLQ 787
                           . . . . | | | . | : . | . . | . . | . . | : | | . . . | | . | . | . . | |
NP_001014576.   779 --HQHNLQQQHSHSQLQHTQQQHQQQQQQQQQQQQQQQQQQQQ--- 823
NP_004889.1      788 TSRL LHGNPSTQLILSAAFPLQQSTFPQSHHQHQSQQQQQLSRHRTDSL 837
                           . | | . | | | | | | |
NP_001014576.   824 -----QQQQQQQQQQQL----- 835
NP_004889.1      838 PDPSKVQPQ 846
                           : : | . |
NP_001014576.   836 ---QLQQQNDILLREDIDDIDAFNLNLSPLHSLGSGSTINPFNSSNNNN 881
NP_004889.1      847 846
NP_001014576.   882 QSYNGGSNLNNGNQNNNRSSNPPQNNEDSLLSCMQMATESSPSINFHM 931
NP_004889.1      847 846
NP_001014576.   932 GISDDGSETQSEDNKMHTSGSNLVQQQQQQQQQQQILQQHQQQSNSFFS 981
NP_004889.1      847 846
NP_001014576.   982 SNPFLNSQNQNQNQLPNLEILPYQMSQEQSQNLFNPSHTAPGSSQ 1027

```

```

#-----
#-----

```

EMBOSS-Align Results

Water Results	
Matrix	Blosum62
Open gap penalty	10.0

Gap extension penalty	0.5
Water output	water-20050914-03065703113546.output
SUBMIT ANOTHER JOB	

```
#####
# Program: water
# Rundate: Wed Sep 14 03:05:19 2005
# Align_format: srspair
# Report_file: /ebi/extserv/old-work/water-20050914-03065703113546.output
#####

#=====
#
# Aligned_sequences: 2
# 1: NP_004889.1
# 2: NP_001014576.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 924
# Identity: 306/924 (33.1%)
# Similarity: 430/924 (46.5%)
# Gaps: 197/924 (21.3%)
# Score: 1171.5
#
#
#=====

NP_004889.1      27 EEDDKDKAK----RVSRNKSEKKRRDQFNVLIKELGSMPLPGNARKMDKST      72
  .|.|.|.|..|   |.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|
NP_001014576.   4 ESDDKDDTKSFLCRKSRNLSEKKRRDQFNSLVNDLSALISTSSRKMDKST      53

NP_004889.1      73 VLQKSIDFLRKHKEITAQSDASEIRQDWKPTFLSNEEFTQLMLEALDGGFF      122
  |.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|
NP_001014576.   54 VLKSTIAFLKNHNEATDRSKVFEIQQDWKPAFLSNDEYTHLMLESLDGMFM      103

NP_004889.1      123 LAIMTDGSI IYVSEVTSLLEHLPSDLVDQSIFNF IPEGEHSEVYKIL--      170
  :...:|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|
NP_001014576.   104 MVFSSMGSI FYASESITSQLGYLPQDLYNMTIYDLAYEMDHEALLNIFMN      153

NP_004889.1      171 STHLLESDSLTP EYLKSKNQLEFCCHMLRGTIDPKEPSTY EYVKFIGNFK      220
  .|.:|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.:
NP_001014576.   154 PTPVIEPRQTD--ISSNQITFYTHLRGGMEKVDANAYELVKFVGYFR      200

NP_004889.1      221 SLNSV----SSAHNGFEG-----TIQRTHRPSYEDRVCFVATVRLAT      259
  :...:  ||...||..|   ..|:.....:..:|.|.|.:..
NP_001014576.   201 NDTNTSTGSSSEVSNGSNGQPAVLPRIFQQNPNAEVDKKL VFGTGRVQN      250

NP_004889.1      260 PQFIKEMCTVE EPNEEFTSRHSLEWKFLFLDHRAPPIIGYLPFEVLGTSG      309
  |.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|.|
NP_001014576.   251 PQLIREMSI IDPTSNEFTSKHSMEWKFLFLDHRAPPIIGYMPFEVLGTSG      300
```


- b) Based on the above alignments, can we claim that the two aa sequences are orthologs? If yes, give your explanation. If not, what extra information do you need to say these two genes are orthologs? (5 points)
- c) Either global or local alignment can yield multiple optimal solutions. Mathematically speaking, all of the solutions are equally good under the parameters chosen by the users. However, to a biologist, not all of them are good. What kind of biology information, do you think, will be useful to help us eliminate the alignments that don't make sense in biologist's view? (5 points)