CAP5510 Introduction to Bioinformatics

Fall 2009

Programming Assignment #1

(Assigned Sept.16. Due: Oct.5, 2009)

1. Sequence Alignment Algorithms

A) BLAST on Web

* Try the following sample sequences for BLAST (NCBI) (http://www.ncbi.nlm.nih.gov/BLAST/) on the web against "**Swiss-Prot Database**".

1) protein sequence (proteinSeq1.txt) (use protein blast)

>protein1

```
MRVLKFGGTSVANAERFLRVADILESNARQGQVATVLSAPAKIT
NHLVAMIEKTISGQDALPNISDAERIFAELLTGLAAAQPGFPLAQLKTFVDQEFAQIK
HVLHGISLLGQCPDSINAALICRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHY
LESTVDIAESTRRIAASRIPADHMVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACL
RADCCEIWTDVDGVYTCDPRQVPDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQF
QIPCLIKNTGNPQAPGTLIGASRDEDELPVKGISNLNNMAMFSVSGPGMKGMVGMAAR
VFAAMSRARISVVLITQSSSEYSISFCVPQSDCVRAERAMQEEFYLELKEGLLEPLAV
TERLAIISVVGDGMRTLRGISAKFFAALARANINIVAIAQGSSERSISVVVNNDDATT
GVRVTHQMLFNTDQVIEVFVIGVGGVGGALLEQLKRQQSWLKNKHIDLRVCGVANSKA
LLTNVHGLNLENWQEELAQAKEPFNLGRLIRLVKEYHLLNPVIVDCTSSQAVADQYAD
FLREGFHVVTPNKKANTSSMDYYHQLRYAAEKSRRKFLYDTNVGAGLPVIENLQNLLN
AGDELMKFSGILSGSLSYIFGKLDEGMSFSEATTLAREMGYTEPDPRDDLSGMDVARK
LLILARETGRELELADIEIEPVLPAEFNAEGDVAAFMANLSQLDDLFAARVAKARDEG
KVLRYVGNIDEDGVCRVKIAEVDGNDPLFKVKNGENALAFYSHYYQPLPLVLRGYGAG
NDVTAAGVFADLLRTLSWKLGV
```

2) DNA sequence (use DNA blast) against "**nr database**".

```
>337..2799
ATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTT
GCCGATATTCTGGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCC
GCCAAAATCACCAACCACCTGGTGGCGATGATTGAAAAAACCATTAGCGGCCAGGATGCT
TTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTTGACGGGACTCGCCGCC
GCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTTGCCCAA
ATAAAACATGTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCT
GCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCG
CGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTAC
CTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCG
GCTGATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTG
GTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCC
GATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGTCAGGTG
CCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTC
GGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGC
CTGATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGAT
GAAGACGAATTACCGGTCAAGGGCATTTCCAATCTGAATAACATGGCAATGTTCAGCGTT
```

TCTGGTCCGGGGATGAAAGGGATGGTCGGCATGGCGGCGCGCGTCTTTGCAGCGATGTCA
CGCGCCCGTATTTCCGTGGTGCTGATTACGCAATCATCTTCCGAATACAGCATCAGTTTC

* Did you find out where the two sequences originated from?

*You should pay attention to **E-value** when using BLAST.

* Find more about BLAST at NCBI Education
* BLAST Statistical background

B) Standalone version is available from NCBI FTP site: (ftp.ncbi.nih.gov/blast/executables/). In this part, you are running the search engine in your computer or laptop. You have to download the executables.

If you are working under the windows operating system, then:

1) Click the ftp link
2) Click "LATEST" directory
3) Download "blast-2.2.18-ia32-win32.exe" to "bin" directory

[If you have a Lynex account download "blast-2.2.18-ia32-linux.tar.gz" to "bin" directory

Check whether BLAST is in your path
> which blastall]

2. **Target sequences should be formatted** before it's searched against.
   a. Copy E.Coli protein sequences (NC_00913.faa) from the following directory:

www.cs.ucf.edu/~shzhang/CAP5510/NC_000913.faa

   b. Now perform 'formatdb' in the BLAST directory
   >formatdb -i NC_000913.faa -n EColi -p T
   c. You will see these files created in the same directory.
   EColi.pin, EColi.psq, EColi.phr, formatdb.log

3. Let's perform a simple BLAST of "**proteinSeq1.txt**"
   a. Copy the "proteinSeq1.txt" into the BLAST directory.

   b. >blastall -p blastp -d EColi -i proteinSeq1.txt -o proteinSeq1.out

```
blastall -p blastp -d EColi -i proteinSeq1.txt
```

4. Change the following options
   A. -e : expectation value (Default: 10)
   B. -m : alignment view option (Default: 0)
   C. -b : Number of databse sequences to show alignments (Default: 250)
   D. -v : Number of database sequences to show one-line descriptor (Default: 500)
   E. -g : Perform gapped alignment (Default: T)
   F. -M : Scoring Matrix (Default: BLOSUM62)

5. There are many options you can adjust. Simply run blastall without any option.

6. Try to make BLAST print out result in html (with -T T)
   >blastall -p blastp -d EColi -i proteinSeq1.txt -o //index.html -T T

C) Read the BLAST Tutorial and Statistical Background