
CS/ST 480

HOMEWORK 3 (DUE MARCH 3RD)

1. **FASTA [10 pts].**

In the FASTA algorithm — what are the advantages and disadvantages of using longer k-mers for finding the initial matches?

2. **Progressive multiple sequence alignment [10 pts].**

What is the major problem with progressive alignment schemes?

3. **MJ1477: the Missing aaRS? [30 pts].**

Aminoacyl-tRNA synthetases (aaRSs) are important proteins that catalyze the reaction that attaches amino acids to their corresponding tRNAs (tRNA is transfer RNA, which is used by the ribosome to extend peptide sequences during translation). aaRSs are present in all cellular organisms (with few exceptions) since they are essential for protein synthesis. The bacterium *Methanococcus jannaschii*, does not appear to have an aaRS for cysteine (CysRS). Fabrega et al (2001) devised a novel computational method to search for the missing CysRS and concluded that the MJ1477 protein was responsible for the CysRS activity. Here we will attempt to validate their results. (NOTE: use the default parameters unless otherwise specified.)

- (a) Go to Swiss-Prot website, search for MJ1477, and get its amino acid sequence in FASTA format (you don't need to turn in the sequence with this assignment).
- (b) Go to the NCBI website and access the BLAST page. Enter the amino acid sequence as the query, and use nr as the database. To increase the sensitivity of the search, use PSI-BLAST rather than BLAST. PSI-BLAST uses the initial hits of BLAST to generate a profile which is then used to search for additional hits. Those additional hits are then included in the profile, and the process can be iterated. To execute a PSI-BLAST search check the "Format for PSI-BLAST" option, and then click "BLAST!". On the next page you will need to click "Format!" to submit the query.
- (c) Take a quick look at your results. Note that most of the genes are "predicted" or "hypothetical", and don't have a function associated with them.
- (d) Run PSI-BLAST for enough iterations until you obtain genes that have a known function.
- (e) Based on the score, alignment, and E-value of your hits, would you consider these proteins homologous? If so does the function of those proteins provide evidence that MJ1477 might be the missing CysRS protein? Do you agree with the conclusion of Fabrega et al.? Compare your conclusion with the results from Ruan et al (2004).

References

- Altschul, S.F., Madden, T.L., Scheifer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res.* 25:3389-3402.
- Fabrega, et al. An aminoacyl tRNA synthetase whose sequence fits into neither of the two known classes. *Nature.* 2001 May 3;411(6833):110-4.
- Ruan B, et al. CysteinyI-tRNA(Cys) formation in *Methanocaldococcus jannaschii*: the mechanism is still unknown. *J Bacteriol.* 2004 Jan;186(1):8-14.

4. Motif identification by Gibbs sampling: E. coli ribosome binding sites [50 pts].

E. coli ribosomes recognize start codons by binding to the mRNA at the Shine/Dalgarno consensus just upstream of an initiation AUG. The 3' end of small subunit ribosomal RNA base-pairs to the Shine/Dalgarno motif. This base pairing does not have to be perfect, and it also occurs with variable spacing relative to the initiating AUG. It is therefore not entirely trivial to identify the consensus sequence.

The strongest part of the Shine/Dalgarno motif is **AGGA**. The strongest ribosome binding sites tend to have something close to **UAAGGAGG**, exactly complementary to the 3' end of the small-subunit ribosomal RNA.

The file `rbs.fa`, accessible from the homework page, contains 823 *E. coli* ribosome binding sites in FASTA format. Each sequence starts from position -20 to +2 relative to the A of the initiator AUG. (If you look at the file, notice that not all *E. coli* genes start with AUG; GUG, CUG, AUA are rare but known initiators.) For example:

```
>thrL Escherichia coli K-12 MG1655
ttacagagtacacaacatccatg
```

A second file, `rbs50.fa`, contains upstream regions from the same 823 genes, 50 nt upstream instead of 20 nt upstream. With a bigger search space, this makes for a somewhat tougher motif identification problem.

These upstream regions have been pulled out based on the annotation in the *E. coli* genome. Undoubtedly they contain some bad annotations; they also contain examples where there is no obvious ribosome binding site at all. However, despite these sources of noise, there's still plenty of information in these sequences, sufficient to automatically determine the Shine/Dalgarno consensus.

Implement a Gibbs sampling algorithm for identifying ungapped consensus motifs. Run your program on the files `rbs.fa` and `rbs50.fa`. Since Gibbs sampling is a probabilistic method, different runs will provide varying answers, so run your program several times. Compare the results from those different runs on both `rbs.fa` and `rbs50.fa`. Based on your results, what is your best guess of the Shine/Dalgarno weight matrix?

Attach a printout of your program to the homework, and send your code by email to the TA.

References:

Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignments. C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, and J.C. Wootton. *Science* 262:208-214, 1993.