Computer Analysis of DNA and Protein Sequences Over the Internet

Part I.

IN CLASS

Download the Lectin sequence output from http://stan.cropsci.uiuc.edu/courses/cpsc265/

Open these in BioEdit (free software).

What is the difference between these outputs?

How can you tell which worked and which did not work so well?

How can you turn these multiple chromatograms into a single sequence of the Lectin gene in A, C, G and T text format?

Use the NIH bioinformatics program web site, http://www.ncbi.nlm.nih.gov/.

Here you can recover sequences for practically any published experiment involving DNA, RNA or protein. You can also search the results of your own experiments against the database of known sequences.

Using BLAST : http://www.ncbi.nlm.nih.gov/BLAST/, search the completed, assembled Lectin sequence you generated above, using nucleotide-nucleotide (blastn), then translated nucleotide-protein (blastx) searches.

What exactly is this gene?

Has anyone sequenced it before?

Who, when, and when was it published?

Does it make a known protein?

Does the protein have a known structure? If so, what features does it have?

Part II.

You are a molecular biologist working for the Centers for Disease Control and Prevention (CDC), a US Govt. body in Atlanta, GA.

A cruise ship docks in Miami, FL, and you get an urgent call to attend the patients aboard the ship. About half of the passengers and crew are sick with an unidentified illness, and it is thought that they might have contracted a new virus. The ship is quarantined at the dock, and the passengers are unable to leave. You go to the ship, where sleepless doctors hand you samples from patients. Back in the lab, you convert RNA from the samples to DNA and get the following sequence:

```
TCAACCAAGTCTGCCTCACCTGACATCGTGGGTACAATCAATGCCCTCCTGGCGAGGATCGCGGCTGCCCGTTCCCTGGTACATCGAGCA
AAGGAAGAACTTTCCACCAGGCCGAGACCCGTTGTCGTGATGATATCAGGAAAACCAGGAATAGGGAAGACCCACCTTGCCAGAGAGCTG
GCTAAGAAGATCGCAACCACCCTTACGGGAGACCAGAGGGTGGGCCTCATCCCACGCAATGGTGTTGACCACTGGGACGCGTACAAGGGA
GAGAGGGTCGTCCTTTGGGATGATTACGGTATGAGTAATCCCATGATCCACGATGTCAGGTTGCAGGAGCTTGCTGACACTTGCCCCCTA
ACACTAAATTGTGACAGGATTGAGAACAAAGGTAAAGTTTTTGACAGTGATGCTATAATTATCACCACTAACTTGGCCAACCCAGCACCA
CTGGTCTATGTCAACTTTGAGGCGTGCTCGAGGCGCAAATGATTTCTTCGTGTACGCCGAAGCCCCTGATGTTGAGAAGGCGAAGCGCGA
CTTCCCAGGCCAACCTGATATGTGGAAGAATGCCTTCAGCCCTGACTTCTCTCACATAAAGCTGATGCTGGCCCCGCAGGGTGGCTTTGA
CAAGAACGGCAATACCCCACATGGGATGGGTGTCATGAAGACCCTCCCCATCGGTTCTCTCATCGCCCGTGCTTCAGGACTCCTCCATGA
AAGGCTGGATGAGTACGAGTTGCAAGGCCCAGCCCTCACAACCTTCAACTTTGACCGAAACAAAGTGCTCGCTTTCAGACAGCTTGCTGC
TGAAAACAAGTACGGGTTGATGGATACGATGAGAGTTGGAGGGCAACTCAAGGGCGTCAGAACCATGCCAGAGCTCAAGCAAGCACTCAA
GAACATCTCAAAGAGGTGCCAGATAGTGTATGGTGGCAGCACCTACACACTTGAATCTGATGGCAAGGGTAGTGTGAGGGTTGACT
```

1. Searching DNA and Protein databases:

Use the NIH bioinformatics program web site, http://www.ncbi.nlm.nih.gov/.

Here you can recover sequences for practically any published experiment involving DNA, RNA or protein. You can also search the results of your own experiments against the database of known sequences.

Using BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi), search the sequence above, using nucleotide-nucleotide (blastn) search.

A. What organism is this sequence from? Genus and species?

B. Is it identical, or similar, to anything that has been previously published?

C. Why are the people on the ship ill? What would you do about it?

2. Open reading frames and protein sequences.

A. Use the sequence above to search the NCBI protein database (GenPept NR) by changing the default search to the "blastx" program that does a 6-frame translation on the protein.

A. Do you think this sequence codes for protein (yes / no).

B. Is this sequence sense or antisense?

C. Does the search give you any idea about the potential function of the protein? This type of protein has a generic name, can you figure it out?

D. Is the sequence above a complete protein coding sequence?

E. Can you find the complete genome sequence of this species? What is its GenBank Accession #?

F. What is the name and accession number of the whole protein from which this sequence is derived?

3. Protein sequences

A. Recover the whole protein sequence related to the DNA above in fasta format using the accession number above. How many amino acids does it contain?

B. What conserved domains does this protein have (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi)?

C. What other conserved domains are present in the other viral proteins in the genome?

D. Do any of these have solved 3D structures?

E. Where is the protein product of the protein in (A) likely to be targeted? (For predicted protein targeting see website at: http://psort.nibb.ac.jp/. Cut and paste the protein sequence that you generated for Part A).

Part III

You are investigating the role of genetic diversity on the flowering time / maturity date of rice. As part of your analysis, you use an Illumina HiSeq 2000 to generate 20,000,000 reads from RNA from five biological replicates at eight time points across the transition from vegetative growth to flowering.

Your analysis reveals that a very small RNA molecule is strongly associated with the transition to flowering.

```
AGAAUCUUGAUGAUGCUGCAU
```

Using NCBI Nucleotide Blast as above, search for similar sequences in the database.

A. Were the parameters you entered in the Blast search altered in any way automatically?

B. Why might it be necessary to use different parameters for a shorter input sequence? Do you notice any difference in the output of the search using this short sequence compared to the longer sequence above?

C. Has this sequence been previously recorded in the database of small, regulatory RNA at miRbase (http://www.mirbase.org/search.shtml)? Note that this database uses a code for different species, such that rice (Oryza sativa) has the code osa, maize (Zea mays) has the code zma, etc.

D. Does this RNA have a known function? What is it?

E. What is the sequence of the full message for this sequence (the pri-miRNA, pre-miRNA, or stem loop sequence)?

F. Use the longer sequence of the pri-miRNA you obtained above to search the rice genome at multiple resources, for example NCBI (select Oryza sativa as target organism), the rice genome project (http://rice.plantbiology.msu.edu/analyses_search_blast.shtml) and phytozome (http://www.phytozome.org) (make sure to select Oryza sativa in the Plant Tree of Life at the left, and no other species).

G. How do these resources differ? Do they all give you the same answer? Why do you think this is?

H. How many hits are there in the rice genome?

I. How many encode this exact pri-miRNA?

J. How many likely encode a version of this miRNA?