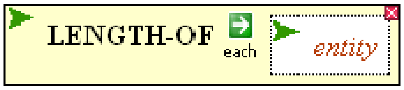
MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics

Patterns and motifs: Do it yourself

I. Familiarization with BioBIKE

I.A. Find the length of *Nostoc punctiforme* (the number of nucleotides in its genome)



1. Use the LENGTH-OF function in the Genome menu. Click the entity box. Find *Nostoc punctiforme* in the Data menu, N-Fixing-Cyanos submenu. Click Npun, the nickname of *Nostoc punctiforme*. Double click the name of the function (or click execute on the function's action menu).
2. What is the length of Npun's genome?

I.B. Find the average length of a gene of *Nostoc punctiforme*

1. Use the LENGTH-OF function in the Genome menu (or the Genes-Proteins menu).
2. Use the GENES-OF function in the Genome menu. Execute it to get a list of the genes of Npun.
3. Find the length of one gene. Copy and paste any gene in the list into the argument of LENGTH-OF. What is its length?
4. Find the lengths of all genes. You should get as many lengths as there are genes. If you get only one number, then the function returned to you the length of the list, not the length of each gene. To get the latter, specify EACH (by clicking the token arrow). This specifies that you want the length of each gene in the list and not the length of the list itself.
5. Run the lengths through the MEAN function (Arithmetic menu, Statistics sub-menu). What is the mean length of a gene in *Nostoc punctiforme*?

I.C. Display the upstream region of NpR3011 and NpF3012

1. Use the SEQUENCE-OF function in the Genome menu to display the chromosome of Npun. Then use the Go to box to go to the region of the chromosome containing NpR3011 (fill in the box with NpR3011 and click Go). What is the start codon of NpR3011? What is the gene upstream of NpR3011?
2. What are the coordinates of the region of the chromosome upstream of NpR3011?
3. Use SEQUENCE-OF and SEQUENCE-UPSTREAM-OF (Genes-Proteins menu Gene-neighborhood submenu) to display the sequence upstream of NpF3012. Compare it to the sequence you displayed in I.C.1. Find it?
4. Replace NpF3012 with NpR3011 to display the upstream sequence of the latter. Compare it to the display of the chromosome sequence. Find it?
5. Get another view of this region by executing the CONTEXT-OF NpR3011 (you can find the function in Genes-Protein, Gene-Neighborhood), using the DRAW option. You'll get both a graphical and a numeric display. Note that the graphical display reverses the orientation of NpR3011, as its policy is to always render the given gene left-to-right. You can mouse over boxes to get a description of each gene. Do the intergenic regions in this neighborhood match what you see in the display of the chromosome?

II. Pattern matching to mine data

Rationale

Suppose you are interested in the interaction of bacteria and their phages. You realize that many phage are quite cozy with bacteria, integrating themselves into a bacterial genome for long periods of time. When a phage is integrated they're called prophages, and it can be difficult to distinguish their DNA from that of the native host, but this is the task you've set for yourself.

To learn how to do identify prophages, it would help to have on hand a set of known prophages on which you can try out your methods. Several prophages in E. coli are known, so your thought is to go there.

A. Gain access to the Genbank file containing the annotation of E. coli

1. You could do this by going to NCBI, finding and downloading the file, and then uploading it into BioBIKE, but I've saved you the trouble. The annotation is available to you in a variable called `ecfile`. You can also scroll through the file by mousing over the black **File** button and clicking **Shared files**. Then find the file called "Escherichia coli…" and click its name. Take a look at it. You'll need it for the rest of this question.
2. Search through the file until you find an annotation of a prophage. Don't pay much attention to gene-specific entries but rather find features that describe the coordinates of the prophage as a whole.
3. Figure out a pattern that will capture the beginning and end coordinates of each prophage along with the description of it (in quotes). Write out the pattern, including the quotation marks.
4. Use MATCHES-OF-PATTERN to extract all prophage information from the E.coli annotation. Be sure to use the CROSS-LINES option, since the information you seek may be on multiple lines.What output did you get?
5. Use DISPLAY-LIST to render in tabular form the information you extracted. What output did you get, and what might you learn from it?

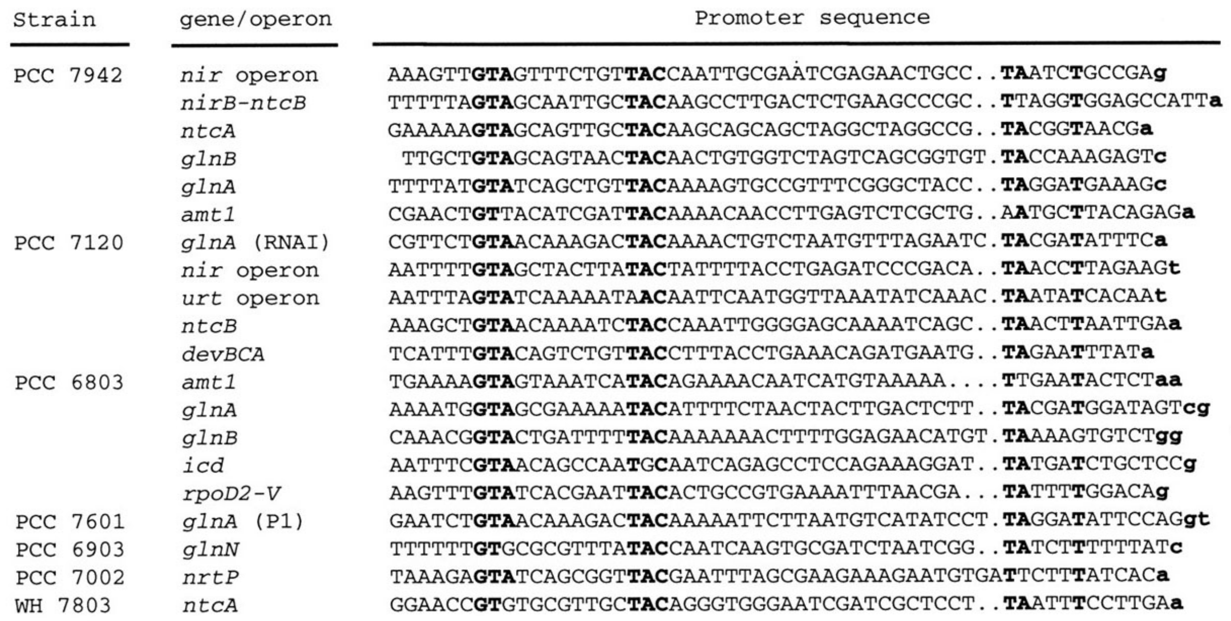
B. Gain access to the Genbank file containing the annotation of E. coli

Many redox proteins have iron-sulfur clusters as cofactors. In some cases, the binding of the cofactor to the protein depend on a precise constellation of cysteine residues, which are spaced Cys-X-X-Cys-X-X-Cys-X-X-X-Cys. Use this observation to identify proteins in Synechocystis PCC 6803 (nickname: S6803) that are candidates to be iron-sulfur proteins. To this end, the function PROTEINS-OF will be useful. In order to access the nominal annotation of the proteins you find, use DESCRIPTION-OF, with the EACH pre-option if you provide the function with a list of proteins, and the DISPLAY option for nice output. What proteins did you find? Which strike you from the annotation as likely iron-sulfur proteins?

III. Position-specific scoring matrices to find proteins with a given motif

Rationale

*NtcA (for **nit**rogen **c**ontrol) is a DNA-binding protein that responds to the availability of nitrogen, e.g. in the form of ammonia, to differentially regulate the expression of many cyanobacterial proteins involved in nitrogen metabolism. In several cases, the DNA sequence to*



Strain	gene/operon	Promoter sequence
PCC 7942	*nir* operon	AAAGTT**GTA**GTTTCTGT**TAC**CAATTGCGAATCGAGAACTGCC..**TA**ATC**T**GCCGA**g**
	nirB-ntcB	TTTTTA**GTA**GCAATTGC**TAC**AAGCCTTGACTCTGAAGCCCGC..**T**TAGG**T**GGAGCCATT**a**
	ntcA	GAAAAA**GTA**GCAGTTGC**TAC**AAGCAGCAGCTAGGCTAGGCCG..**TA**CGG**T**AACG**a**
	glnB	TTGCT**GTA**GCAGTAAC**TAC**AACTGTGGTCTAGTCAGCGGTGT.**TA**CCAAAGAGT**c**
	glnA	TTTTAT**GTA**TCAGCTGT**TAC**AAAAGTGCCGTTTCGGGCTACC..**TA**GGA**T**GAAAG**c**
	amt1	CGAACT**GT**TACATCGAT**TAC**AAAACAACCTTGAGTCTCGCTG..A**A**TGC**T**TACAGAG**a**
PCC 7120	*glnA* (RNAI)	CGTTCT**GTA**ACAAAGAC**TAC**AAAACTGTCTAATGTTTAGAATC.**TA**CGA**T**ATTTC**a**
	nir operon	AATTTT**GTA**GCTACTTA**TAC**TATTTTACCTGAGATCCCGACA..**TA**ACC**T**TAGAAG**t**
	urt operon	AATTTA**GTA**TCAAAAATA**AC**AATTCAATGGTTAAATATCAAAC.**TA**ATA**T**CACAA**t**
	ntcB	AAAGCT**GTA**ACAAAATC**TAC**CAAATTGGGGAGCAAAATCAGC..**TA**ACT**T**AATTGA**a**
	devBCA	TCATTT**GTA**CAGTCTGT**TAC**CTTTACCTGAAACAGATGAATG..**TA**GAA**T**TTAT**a**
PCC 6803	*amt1*	TGAAAA**GTA**GTAAATCA**TAC**AGAAAACAATCATGTAAAAA....**T**TGAA**T**ACTCT**aa**
	glnA	AAAATG**GTA**GCGAAAAA**TAC**ATTTTCTAACTACTTGACTCTT..**TA**CGA**T**GGATAGT**cg**
	glnB	CAAACG**GTA**CTGATTTT**TAC**AAAAAAACTTTTGGAGAACATGT.**TA**AAAGTGTCT**gg**
	icd	AATTTC**GTA**ACAGCCAA**TGC**AATCAGAGCCTCCAGAAAGGAT..**TA**TGA**T**CTGCTCC**g**
	rpoD2-V	AAGTTT**GTA**TCACGAAT**TAC**ACTGCCGTGAAAATTTAACGA...**TA**TTT**T**GGACA**g**
PCC 7601	*glnA* (P1)	GAATCT**GTA**ACAAAGAC**TAC**AAAAATTCTTAATGTCATATCCT.**TA**GGA**T**ATTCCAG**gt**
PCC 6903	*glnN*	TTTTTT**GT**GCGCGTTTA**TAC**CAATCAAGTGCGATCTAATCGG..**TA**TCT**T**TTTTAT**c**
PCC 7002	*nrtP*	TAAAGA**GTA**TCAGCGGT**TAC**GAATTTAGCGAAGAAAGAATGTGA**T**TCTT**T**ATCAC**a**
WH 7803	*ntcA*	GGAACC**GT**GTGCGTTGC**TAC**AGGGTGGAATCGATCGCTCCT..**TA**ATT**T**CCTTGA**a**

Binding sites of NtcA protein upstream from the promoter of several cyanobacterial genes. The sequences in bold are merely to draw to your attention relatively conserved nucleotides. The left hand portion is the NtcA-binding region and the right hand portion is the RNA polymerase binding region (i.e. the promoter). The table is from Herrero et al (2001) J Bacteriol 183:411-425.

which NtcA binds has been determined in the laboratory. You happen to be interested in a cyanobacterium, Anabaena variabilis (nicknamed Avar) for which there is no laboratory evidence concerning the binding of NtcA. This is an instance where bioinformatics may come to the rescue!

A. <u>Predict binding sites for a regulatory protein in an organism with no pertinent labarotory data</u>

1. Use the sequences from this table to find possible NtcA-binding sites in Anabaena variabilis ATCC 29413 (nickname Avar). Investigate at least the first few matches to determine if the annotation is suggestive of a role in nitrogen metabolism. You'll be interested in the following functions:

   APPLY-PSSM-TO
   DESCRIPTION-OF

   You will be relieved to know that the sequences are available to you (no need to type) by using the variable `ntcA-sites`. What are the top matches and what did you conclude?

2. Determine the information content of the aligned sequences. Based on what you find, consider altering your approach to III.A1. These functions will be helpful:

   INFORMATION-OF
   PLOT

   What does a plot of the information content look like? How did you make use of your findings?

3. The table shown above has variable gaps in the alignment to make both parts of the sequence align. PSSM's don't do well with sequences with variable gaps. Consider ways in which you could make use of the full information in the table. Ideas?

B. Troubleshooting through patterns.

When I first put in the sequences shown in the table, I made some typographical errors. You can verify this by using my original effort (available to you as `ntcA-sites-old`) in III.A. You can stare at the sequences (as I did) by displaying them using DISPLAY-LIST with the EACH pre-option, but you'll probably have better luck if you use pattern matching to detect the problem. What pattern can you use to find the characters in the sequences that are ***not*** legitimate nucleotides? The pattern cheat sheet on the course web site might help. Ideas?

IV. What DNA pattern marks the beginning of protein-encoding genes?

Rationale

*Since the title of today's module is "*Pattern Recognition and Gene Finding", *I feel compelled to end with something related to finding the pattern beginning a gene. How does the cell plow through huge stretches of nucleotide sequences to recognize where a gene should begin? You no doubt have your own ideas, but put them aside for the moment and try the following.*

Define a set of sequences upstream from the protein coding genes of your favorite organism or bacteriophage. Consider only the 20 nucleotides upstream. The following functions will be of use:

SEQUENCE-UPSTREAM-OF (using the LENGTH option)
CODING-GENES-OF
MOTIFS-IN (using the DNA option)

If you choose a bacterium, you'll probably run out of time, hitting the execution time limit of 40 seconds. Not a problem! You don't need thousands of upstream sequences to find a good motif. One solution is to take a random sample of the upstream sequences, using the following function:

CHOOSE-FROM (using the WITHOUT-REPLACEMENT and TIMES options)

200 samples should be plenty and will drastically cut down the execution time. What motifs did you find? What is the significance of the motif(s)?