MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics

Projects: Computation to Solve Problems

Finding targets for DNA-binding proteins (uses CyanoBIKE)

NtcA (for **ni**trogen **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 which NtcA binds has been determined in the laboratory (see table on the next page).

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. Predict binding sites for a regulatory protein in an organism with no experimental data

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`.* Note that the *with-pssm-from* box of APPLY-PSSM-TO requires a list of sequences (e.g. `ntcA-sites`). You might reasonably think that it takes an actual PSSM, such as that produced by the MAKE-PSSM-FROM function, but it doesn't. What are the top matches and what did you conclude? What code did you use?

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

   Note that, like APPLY-PSSM-TO, INFORMATION-OF requires a list of sequences (e.g. `ntcA-sites`). Again, a PSSM won't work. What does a plot of the information content look like? What code did you use to get the plot? 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?

---

* If you execute a function calling for `ntca-sites` and get the error message "`PROBLEM: I don't understand what you mean by ...`", then see the addendum at the end of this problem.

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..AATGC**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....TTGAA**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**T**G**C**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**AGGGTGGGAATCGATCGCTCCT..**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.

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?

Addendum: What if pre-made variables don't work?

If you attempt to use ecfile, ntca-sites, or ntca-sites-old (properly spelled!) and get the error message "`PROBLEM: I don't understand what you mean by…`*", it could be the variables have disappeared. To get them back, go to the* ***All*** *menu, bring down RUN-FILE, and execute the function as shown to the right.*