Introduction to Bioinformatics

Problem Set 7: Motif Discovery Through PSSM's

Motif discovery in previously recognized proteins through MOTIFS-IN

1. In the example of a protein-centered project (see Research Project web page; Advice and Examples), you saw a list of protein motifs characteristic of cytosine methyltransferases. Get the full list by downloading Bujnicki & Radlinska (1999), as described in the tour. How can you make a similar list yourself? To test your skills, recreate their list (here's how).

1a. First, you need a collection of plausible cytosine methyltransferases. The tour describes the process of getting such a list from proteins proven in the laboratory. You'll take a short cut by developing a collection based on sequence similarity rather than experimental results. Go into CyanoBIKE, and ask for SEQUENCES-SIMILAR-TO p-NpR6310 (one of the proteins found in the tour). Use the BYPASS-LOOKUP option.[*] Describe the target proteins that are found. How many are they? How similar are they to the query p-NpR6310? (Look at the Expect values) What fraction of the protein is similar, just a small portion of it or the whole thing?

1b. Define a set of putative cytosine methyltransferases (perhaps `C-MT-subset`). Note that SEQUENCES-SIMILAR-TO ***displays*** a table of results, but it ***returns*** in the Results Pane two things: (1) a list of the proteins found, in order from best hit to worst, and (2) the table that was popped up, in computer-readable form. DEFINE a variable that consists of the FIRST 15 hits. Note that the FIRST function allows you to specify a number, indicating how many *entities* you want. The larger the set, the more accurate the motifs you will determine but the longer the process will take. A set size of 15 is a reasonable compromise.

1c. Use `C-MT-subset` as input to MOTIFS-IN.[†] Specify the **Protein** option (because the function is currently too stupid to figure out what you're giving it by itself) and ask it to **Return** 10 motifs. Then execute the function. Note that MOTIFS-IN both displays an extensive analysis of the motifs it finds and also returns a list of sequences. Actually, if you look carefully, you'll see that it is really a ***list of lists***. How can you tell that it is a list of lists rather than a simple list? Connect the result and the display by finding in the display the sequences listed in the result. DEFINE a variable (perhaps `all-motifs`) as the result of MOTIFS-IN. To do this, use an asterisk (*) as the *value*, indicating the previous result in the Result Pane, or simply drag the result into the *value* box. Don't waste tens of seconds by re-executing the MOTIFS-IN function.

1d. DEFINE a variable (perhaps `motif-1`) as the FIRST list of sequences in `all-motifs`. Verify that the result shown indeed is the first list and only the first list.

---

[*] This is currently necessary in CyanoBIKE (though not in PhAnToMe/BioBIKE), because the super-fast lookup table used by CyanoBIKE gives results that are not easily interpreted. This is one of many things that need to be fixed!

[†] See the example of a protein-centered research project, Part II, for an explanation of how MOTIFS-IN works.

2. Now you'll analyze the first motif. Scroll the display to Motif 1. You'll see several sections:

Section 1: **Statistics**
Section 2: **Simplified position-specific probability matrix**
Section 3: **Information content**
Section 4: **Consensus sequence**
Section 5: **Sites**
Section 6: **Block diagram**
Section 7: **Blocks**
Section 8: **Position-specific scoring matrix**
Section 9: **Position-specific probability matrix**

2a. Consider Section 1 (**Statistics**). What do **width** and **sites** mean? Be sure to account for the specific numbers given. In plain English, what does E-value mean (by analogy with the similarly named quantity used in Blast statistics)?

2b. Consider Section 5 (**Sites**). What do you make of the column labeled **Start**? Test your hypothesis by displaying the sequence of the first protein. Where do you find the sequence shown in the first line under **Sites**? How do the sequences under **Sites** relate to the sequences in the **Blocks** section? How do they relate to the sequences returned in the Results Pane?

2c. How would you construct the **Consensus sequence** section from the **Sites** section? What rules would you follow?

2d. Qualitatively, how do you relate the height of the bars in the **Information content** section to the **Sites** section? To get quantitation, use the INFORMATION-OF function in BioBIKE, giving it `motif-1`. How do you relate the numbers returned by this function to the **Information content** section?

(As for the rest of the sections,… wait until BNFO601 Integrated Bioinformatics)

3. Compare the motifs found by MOTIFS-IN to those found by Bujnicki and Radlinska.

3a. To the extent possible, find a motif amongst the motifs presented by MOTIFS-IN that corresponds to those shown by Bujnicki and Radlinska. Note the equivalences, e.g. Motif 1 (mine) = Motif XII (B&R).

3b. Scroll down to near the bottom of the display given by MOTIFS-IN to the **Summary of Motifs** section. Interpret what you see. Are the all motifs found in all proteins in the same order?

4. Align your set of cytosine methyltransfereases, `C-MT-subset`, in two ways:

4a. Use the ALIGNMENT-OF function to produce an alignment of the 15 sequences in `C-MT-subset`. Use the **Colored** option to get pretty output. What does each section of the displayed output mean?

4b. Run the same ALIGNMENT-OF function but without the **Colored** option. This gives you pure text output. Copy this into your favorite word processor. Highlight the region of the sequence corresponding to Motif I as defined by Bujnicki and Radlinska.

Motif discovery in unrecognized proteins through APPLY-PSSM-TO

5. Does *Anabaena* PCC 7120 (A7120) possess cytosine methyltransferases? To address this question, search through all proteins of A7120 for those with motifs similar to those discovered by MOTIFS-IN. To do this, use the APPLY-PSSM-TO function, giving A7120 as the *sequence-source* and `motif-1` as the value of *with-pssm-from*. The function will use the block of sequences you provide in `motif-1` and construct from it a PSSM.[‡] The result of executing this function is a list of lists. Each sublist consists of a protein, coordinate, direction, and score. The higher the score, the closer the match between the region of the given protein and the motif described by the list of sequences found in `motif-1`.

5a. How many proteins of A7120 were identified by APPLY-PSSM-TO?

5b. What is the current annotation of these proteins? To find out, use the DESCRIPTION-OF function. For the *entity* field, provide the first element of the result of APPLY-PSSM-TO. You can do this readily by using the FIRST function in IN-EACH mode (mouse over the IN icon to change it to IN-EACH), and provide the result as the argument to FIRST. Are the annotations consistent with the proteins being cytosine methyltransferases?

5c. Display the sequence of the first protein found. Is there a plausible motif at the coordinate given in the result?

6. One motif is not sufficient to identify a cytosine methyltransferase. Determine whether the proteins identified in Question **5** have the full complement of motifs by adding them to the set of proteins considered by MOTIFS-IN. To do this, use the JOIN function to join `C-MT-subset` with the protein names you extracted in Question **5b**. Execute MOTIFS-IN with this new set of proteins, and scroll down to the **Summary of Motifs** section. Do the *Anabaena* proteins contain the proper motifs in the proper order?

[‡] See *What is a Gene*, Part III, for an explanation of what a PSSM is.