Amadeus/Allegro – hands-on session

In this exercise we will discover *cis*-regulatory motifs using Amadeus and Allegro. For simplicity, we shall use the sample datasets that are supplied in the Allegro installation. We have already downloaded the relevant sequence files for you (promoter and 3' UTR sequence files for various species are available at http://acgt.cs.tau.ac.il/allegro/download.html).

Launching Allegro: Click on "run_1.3G_mem.bat" in the Allegro installation folder: "C:\ProgramFiles\Allegro_v1.0\".

1. Human G_2+M cell-cycle phase analysis using Amadeus:

In this section, we shall use Amadeus to search for motifs that are over-represented in the promoter sequences of genes whose expression peaks at the G_2 and G_2/M phases of the human cell cycle [Whitfield et al., 2002].

1. **Load the parameters:**
   a. Click on the "Load parameters from file" button at the bottom
   b. In the "params" folder, select the "params_HCC.txt" file
   c. Verify that the parameters were loaded: "Organism" should be "Human", "Target set file" is "targetSets/HCC-G2M.ensid.txt"; we will be analyzing promoter sequences that include the 350 bps upstream the TSS; we shall use the binned enrichment score to handle length and GC-content biases in the target-set sequences.
2. **Set pairs analysis:** Click on the "Analyze pairs" checkbox so that Amadeus will also search for co-occurring motif pairs
3. **Run Amadeus:** Click on the "Run" button at the bottom, and wait for the analysis to complete
4. **Analyze the output:**
   a. The top-scoring motif in the "Results" tab should be the pattern termed **CHR** (TTTGAAA) whose binding factor is yet unknown.
      i. Click the "List" button (the small magnifying glass) to view the list of its putative targets.
   b. The second motif is **NF-Y**, which binds the CCAAT box.
      i. Click on the localization graph to view a histogram of the CCAAT-box locations – the peak is roughly 65 bases upstream the TSS.
      ii. Amadeus reports similar known motifs (e.g. from the Transfac database) in the "Similarities to known motifs" table at the bottom. Click on a motif name to get its full logo.
   c. Click on the "**View TFBSs**" button at the top to open the binding-sites viewer.
      i. Click on the "Sort by" checkbox of the 2nd motif, and then for the 1st motif.
      ii. Notice that many promoters contain both motifs, and that the two motifs are often quite close to one another.
      iii. You can zoom in/out using the buttons at the top left.
   d. Click on the "**Pairs Results**" tab at the top. Notice that CHR and NF-Y receive a statistically significant co-occurrence score.
   e. Click on the "**Output**" tab to view the textual output. Go to the top.
      Amadeus indicates the number of promoters it analyzed (330 in our target set), their average length, and the single-nucleotide frequencies. The target-set promoters in this example are slightly more GC-rich than the rest of the genome (60% vs. 58%).

5. Re-run the analysis using the **HG score**:
    a. Choose the "Hypergeometric" variant for the enrichment score
    b. Click on the "Run" button
    c. Amadeus reports 3 motifs: CHR and NF-Y, as before, and a GC-rich motif, which is probably a false-positive - it is found since the target-set is slightly GC-rich.

2. *Mouse innate immune response analysis using Allegro:*

In this section, we will analyze expression profiles of RAW264.7 monocyte macrophage-like cell line after exposure to several pathogen-mimetic agents (data from http://www.systemsbiology-immunity.org). Each of the agents used is recognized by a different subset of TLRs (Toll-like receptors). Exposure to lipopolysaccharide (LPS) was sampled at several time points.

1. **Load the parameters:**
    a. Click on the "Load parameters from file" button at the bottom
    b. In the "params" folder, select the "params_TLRs.txt" file
    c. Verify that the parameters were loaded: "Data type" (top left) should be set to "Expression", "Organism" should be "Mouse", "Expression file" should be "expression/TLRs_RAW264.7.avg.txt"; we will use a pre-defined cutoff of 1.5 for the expression values (which are log2 change-fold with respect to un-treated cells), i.e., all values above 1.5 are considered "up regulated"; as in the previous example, we will analyze promoter sequences that include the 350 bps upstream the TSS, and we shall use the binned enrichment score.
2. **Run Amadeus:** Click on the "Run" button at the bottom, and wait for the analysis to complete (this may take a couple of minutes)
3. **Analyze the output:**
    a. The top-scoring motif in the "Results" tab should be **ISRE** (Interferon-stimulated response element) (AGTTTC..TT). ISRE is bound by the Isgf3 complex, composed of Stat1, Stat2 and Irf9. Other members of the Irf and Stat families bind the same motif.
        i. Click on the expression profile graph just below the motif logo. The genes whose promoter contains the ISRE element are up-regulated mainly in response to LPS (which activates TLR4) and poly-IC (which activates TLR3). Notice that the genes activated following LPS are up-regulated during late time-points (4 hours and later).
        ii. At the top of the expression pattern window, click twice on the second button from the right ("Switch Pattern"). This should show the expression matrix for the genes in the predicted ISRE transcriptional module. Notice that two members of the Isgf3 complex – Stat2 and Irf9, are found in this module. In other words, we observe a *cis*-regulatory feedback loop, a common theme in transcriptional networks.
    b. The second motif is **NF-kappaB** (GG[AG]..T[CT]CC).
        i. Click on the expression profile graph. The genes in this transcriptional module respond to most treatments, not just LPS and poly-IC. Also, the up-regulation following LPS is much faster than what we observed for the ISRE targets – here, a significant up-regulation occurs within 1 hour.
        ii. As in (a.ii), click twice on the "Switch Pattern" button. Among the members of the NF-kappaB transcriptional module are Nfkb2 and Rel, which are subunits of NF-kappaB itself – another transcriptional feedback loop. Also among the targets of NF-kappaB are Nfkbia and Nfkbiz, inhibitors of NF-kappaB; thus, we also have here a negative *cis*-regulatory feedback loop.

3. Protein binding microarray data for E2F2

4. **Load the parameters:**
   4.1. Click on the “Load parameters from file” button at the bottom
   4.2. In the “params” folder, select the “params_PBM.txt” file
   4.3. Verify that the parameters were loaded: “Data type” (top left) should be set to “PBM”, “Organism” should be disabled (in gray), “PBM file” should be “PBM/E2F2_1022.2_v2_deBruijn.txt” (note that two files are used in this analysis); Score is "Average", k=9 and #k-mers=500.

5. **Run Amadeus:** Click on the “Run” button at the bottom, and wait for the analysis to complete (this will take aprox. 30 seconds)

6. **Analyze the output:**
   6.1. The top-scoring motif in the “Results” tab should be **E2F**.
   6.2. You can play with the score and values of k and #k-mers, and see that the top scoring motif is always E2F. If you change the parameters, remove the previous runs (by clicking on the red x on the left of the line), and press 'Add'.
   6.3. In the output tab you can see the kind of data AmadeusPBM runs on. The target set includes 500 9-mers (average, min, max length are all 9). The nucleotide sequences are uniform, i.e. frequency of 0.25 for each nucleotide (this is due to the design of the sequences).