Case study in Bioinformatics – Module 3

Paper: Tonikian et al. A Specificity Map for the PDZ Domain Family, PLoS Biol 6(9):e239, 2008, http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0060239

Teacher: David Gfeller
Assistant : Santiago Carmona
Dates: Nov 20, 9-12, 13-16; Nov 21, 9-12
Room: POL/189.19

Goals:

- Reproduce Fig. 7B and Fig. 2.
- Assess the claim "PDZ Domain Sequence Identity Accurately Predicts Binding Specificity".
- Assess the claim "We find that the PDZ domain family is surprisingly complex and diverse, forming at least 16 unique specificity classes across human and worm".

Instructions:

1. Download the data at: https://www2.unil.ch/cbg/index.php?title=How_well_does_sequence_similarity_predict_similarity_in_binding_specificity%3F
   a. Familiarize yourself with the fasta files in PDZligands/ directory. The 'X' or '-' stand for gaps.
   b. Open the PDZ_SMART_CLUSTAL_sub.fa that lists the sequences of PDZ domains. The first line indicates where the binding site is located ('B').

2. Compute the Position Weight Matrices in R.
   a. Load the sequences from the PDZLigands/ folder using the script analyse.R. Spend some time understanding the code until "#Create PWM matrices" section (do you know what the lapply() function does?). Run this code (until "#Create PWM matrices") and make sure you have all the packages installed. Make sure you understand what is stored in the different variables.
   From this point, the code contains holes (#...) that you need to fill.
   b. Compute the frequency of each amino acid at each position (compCount() and compFreq() functions).
   c. Compute the similarity between the PWMs using Eq (1). (compPWMSim() function). In this case do **not** group amino acids together, neither use codon bias frequencies (see part 6).

3. Compute sequence similarity between PDZ binding sites.
   a. Compute the binding site sequence identity as described in the paper (compSeqSim() function).

4. Plot PWM similarity vs binding site similarity, as in Figure 7
    a. Color points corresponding to pairs of domains from the same class (data from PDZclass.txt)
    b. Are there qualitative/quantitative differences between Fig 7 and what you get? What could be reasons?
    c. Do you agree with the authors about the claim that PDZ domains with high sequence similarity also have similar binding specificity?
    d. Redo the same plot but using another sequence alignment file (PDZ_SMART_MUSCLE_sub.fa, or PDZ_phage_MUSCLE.fa). Are there differences?

5. Do the clustering of PDZ domains based on their specificity
    a. Open the LOLA software
    b. Upload the PDZLigands_LOLA/project.txt file
    c. Upload the codon bias file (phageLibraryNNKTheoreticalCodonBias.txt).
    d. Build the tree (Logo Tree)
        i. Are there differences between the one in Figure 2?
        ii. Do you agree with the authors about the 16 classes?

6. Try to redo the clustering in R (*more advanced*).
    a. Recompute PWM similarity as in Eq (1), this time including amino acid grouping and codon bias.
    b. Use the 'hc <- hclust(…)' function in R with distances given as (1-PWMSimilarity) and method="average".
    c. Plot the tree with plot(as.dendrogram(hc), horiz=T, axes=F)
    d. Is the tree different from the one in Figure 2?
    e. Using the PDZclass.txt file which annotates the two main clusters of Figure 2, seven clusters defined manually, or the sixteen clusters reported in the paper, check if the different clustering obtained with hclust is consistent with the different classes of PDZ domains.

If you are unsure about some R functions/code, do not hesitate to look online or ask us.