Case study in Bioinformatics – Module 4

How well does sequence similarity predict similarity in binding specificity?

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 : Marthe Solleder

Dates: Nov 23, 14-17h; Nov 24, 14-17h; Nov 26, 14-17h

Zoom: https://unil.zoom.us/j/92889642180.

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. Take some time (roughly an hour) to think about how you are going to organize the code to build the PWM (20x10 matrix) of each PDZ domain based on the list of peptide binding to each of these domains, and compute the similarity between PWMs. Write a pseudo code starting from the object 'seq', which is a list of lists of sequences, with seq[[*i*]][*j*] corresponding to the j^{th} peptide binding to the i^{th} PDZ domain (e.g., seq[[1]][1] = "RASSTFGYFC", seq[[1]][2] = "HHVHPFACPR",…).

2. 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').

3. Compute the Position Weight Matrices in R using analyze.R.
   a. Make sure you have the latest version of the packages installed on your machine (especially 'stats' version 3.5.1, you can check with 'packageVersion("stats")'). If not, install them with install.packages, as described in the code.

b. Load the sequences from the PDZLigands/ folder using the script analyze.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"). Make sure you understand what is stored in the different variables.
From this point, the code contains holes (#...) that you need to fill.
c. Compute the frequency of each amino acid at each position, starting with the loop " for(s in seq.all){ ". Make sure Fig=7 and do **not** group amino acids together, neither use codon bias frequencies (see part 7).
d. Compute the similarity between the PWMs using Eq (1). (compPWMSim() function). In this case, make sure Fig=7 and do **not** group amino acids together, neither use codon bias frequencies (this will only be used in part 7).

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

5. Plot PWM similarity vs binding site similarity, as in Figure 7
   a. Color points corresponding to pairs of domains from the same class (column class16 from PDZclass.txt)
   b. Are there qualitative/quantitative differences between Fig 7 and what you get? What could be the 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?

6. Do the clustering of PDZ domains based on their specificity
   a. Open the LOLA software (lola-1.1-beta.jar)
   b. Load the PDZLigands_LOLA/project.txt file in Profile Selection
   c. Load the codon bias file in Codon Bias File (phageLibraryNNKTheoreticalCodonBias.txt).
   d. Click on Open.
   e. Generate the tree in Logo Tree and save it (typically the pdf is saved at the same location as the project.txt file).
      i. Are there differences with the one in Figure 2?
      ii. Do you agree with the authors about the 16 classes?

7. Try to redo the clustering in R (*more advanced*).
   a. Recompute the PWMs this time including codon bias renormalization and grouping similar amino acids. Recompute the PWM similarity based on the new PWMs.
   b. Use the 'hc <- hclust(…)' function in R with distances given as (1-PWMSimilarity) and method="ward.D" to compute a hierarchical clustering.
   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 2 main clusters of Figure 2 and the 16 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.