Manual annotation of key genes, one gene at a time

José M. González

The first thing to do is to build a database of reliable sequences. I start out with the peptides in the set of complete genomes in the MAR database (https://mmp.sfb.uit.no/databases). Most genomes in the set of complete genomes are from isolates. For a gene described in Rhodobacteraceae it might be ok, but there might not be enough diversity for other abundant groups. For example there are but a few SAR11 genomes. I included then the partial genomes from MAR that has isolates, MAGs and SAGs. However, we need to consider that this set of sequences might be of lower quality, or there might be contaminating contigs for some genomes and the taxonomy is messed up. Anyway, this is a first approach and we get an idea of the diversity of the gene, how many taxa there are, how many organisms have the gene, etc.

As an example, I did this recently as a first approach with the rhodopsin peptides and envision metatranscriptomes. We want to study which taxa are expressing rhodopsins in the envision samples.

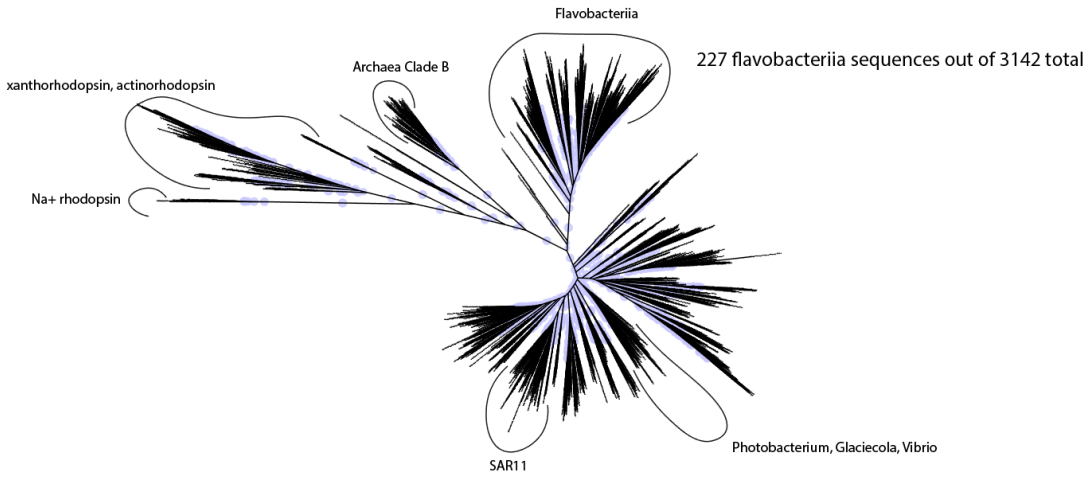
First, I built a reference database of rhodopsins that represented the rhodopsins in the envision metatranscriptomes. We can call this "reference rhodopsin database". To do this I ran the pfam specific for rhodopsins, PF01036, with all peptides in MAR, complete (1270 genomes) and partial (13233 genomes). I included the set in Paoli 2021 (34799 SAG and MAR genomes), with the recommended gathering score for the model. I blasted all envision peptides against this rhodopsin database and retrieved from the reference rhodopsin database only those that had a hit with percent identity above 60 and bitscore above 50. I removed duplicates. Then I made a maximum likelihood tree with IQ-TREE or fasttree, which can handle a large number of sequences. The first rhodopsin tree was too big but all this reduced the size of the reference rhodopsin database tree so I could handle it better and the tree was reliable and made sense.

Now, the envision peptides come from assembled sequences in a metatranscriptome. They are shorter than the complete peptide. There is no way to make a reliable tree with these shorter peptides. The other thing is that I couldn't retrieve many of them with the rhodopsin pfam because most are not long enough to reach the minimum score. From the envision peptides, I also retrieved those sequences that had a hit with percent identity above 60 and bitscore above 50 to any peptide in the reference rhodopsin database. I also included those envision peptides that were retrieved with PF01036. This retrieved some more envision peptides, eukaryotic, that were not represented in my rhodopsin reference database because they only have prokaryotes.

Then comes the double check to confirm that what I retrieved from the envision peptides were actually rhodopsins and not a blast artifact. Some times rhodopsins have additional conserved domains. These additional conserved domains might retrieve the wrong thing. At the same time I double check, I can quantify the peptides in each taxonomic or functional group. I do then a placement of the envision peptides on the reference rhodopsin database. The placement doesn't change the topology of the original reference tree. I do it this way in this next paragraph that I copied from Ben's manuscript:

The package PaPaRa (Berger et al.), EPA-ng (Barbera et al., 2019) and gappa (Czech et al.) were used to confirm the position of the new peptides on the phylogenetic tree inside the rhodopsins after visualization of the tree with iToL (Letunic et al.). The amino acid substitution model was predicted with IQ-TREE using the reference rhodopsin database (Minh et al).

The method removed some envision peptides with long branches on the tree. This is so because a number of envision peptides were retrieved by blast when they were actually something else other than rhodopsins. This below is the reference database tree for the envision samples, although I removed the eukaryotic sequences to simplify.



For now I have only labeled a few of the clusters, the easy ones. The labels are the same as in Pinhassi et al., 2016. You don't see the labels of the sequences because there are hundreds of peptide sequences in the tree. I can quantify now. For example, I have 227 envision rhodopsin peptides that clustered with the flavobacteriia rhodopsins. The envision peptides are also included in the figure. From the few clusters that I labeled, there are functional groups, such as Na+ rhodopsins or xanthorhodopsins/actinorhodopsins (actinorhodopsins are actually xanthorhodopsins, those rhodopsins with two antennae instead of just retinal). Some Sar11 rhodopsins are also expressed in the samples.

Now, the next step is quantifying the number of reads that were need to make the contigs for the flavobacteria rhodopsins. This way we have a number of envision reads that corresponds to flavobacteria rhodopsin sequences. Ben did this part. He retrieved each of the orfs and aligned again the metatranscriptomes to quantify how many reads made up each of orfs for each of the envision samples.

References.

Berger SA, Stamatakis A. Aligning short reads to reference alignments and trees. Bioinformatics. 2011;27:2068–2075.), EPA-ng (Barbera et al. EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. Syst Biol. 2019 Mar; 68(2): 365–369

Czech et al. Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. Bioinformatics, Volume 36, Issue 10, 15 May 2020, Pages 3263–3265

Letunic I and Bork P (2021) Nucleic Acids Res doi: 10.1093/nar/gkab301 Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation.

B.Q. Minh, H.A. Schmidt, O. Chernomor, D. Schrempf, M.D. Woodhams, A. von Haeseler, R. Lanfear (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol., 37:1530-1534.