Hey Barbara,

So I’ve run the phylogenetic analysis for the antismashdb v2 A-domains with the 154 aa extracted. I used CD-HIT to redundancy filter at 60% amino acid identity. I then combined with the sp2 training set (colored in the attached tree). I had to use FastTree for treeing because even with the 60% redundancy filtering there were 6,581 antismash db sequences plus the 1,029 sp2 sequences for a total of 7,610 sequences. FastTree provides an “approximate maximum likelihood” which is often fairly good for this type of exploratory analysis. However, if you prefer, I can set up a RAxML run as well– it will take several days to complete, so I’m not sure what your timeline is for getting the sequences to John.

However, what you see here is an overall picture of where we have “gaps” in our knowledge. If I had to guess, I would think some of the more distant sequences from around 8 to 10 o’clock on the tree are fungal sequences, but I don’t know for sure (we’ll need to pare the MIBiG GenBank file to get to that I think?). You can see already there are a couple interesting places where there are clear ‘gaps’ in the circle since these are opportunities where we have large, fairly diverse (<60% identity) A domain clades without experimentally-verified representatives.

Let me know if you have any questions about this! It got too messy to include all tip labels so I just included the MIBiG ones. So for you to access your unknown sequences of interesting what I was thinking is you could draw arrows into the ‘gaps’ you see in this tree that look interesting and number them. And I will pull out the antismashdb clades corresponding to those gaps and send those sequences to you.

An alternative is I can do k-medoids clustering again (based on a ML sequence distance matrix, not the physicochemical properties) and identify “cluster representatives” for those clusters which we have no experimentally-validated MIBiG domains. This would be a bit more algorithmic and reproducible – I’ve done it before for a different project so it would just be repurposing code. Just let me know how you’d like me to proceed or if you want any changes/clarification ☺

Thanks,

Serina