Take-Home Assignment

Zachary Foster

May 30, 2014

1

In order to group similar proteins into families, I would use BLAST to identify homologs and `mcl` to cluster similar sequences into families. Custom scripts would be used to parse the mcl results into a tables of counts. I would then use R scripts to plot rarefaction curves of the number of pan- and core-genome protein families. The core genome rarefaction curve would decrease with additional genomes and likely level out at some point, but 10 genomes might not be enough. The number of core gene families shared by all genomes once the curve levels out would be the size of the core genome. The pan-genome curve would increase as more genomes are added and likely not level out. It might converge on a positive linear correlation. In theory, the size of the pan genome is infinite, given an infinite number of genomes.

2

I would be interested in the entire microbial community, so I would take water samples at various depths and filter the water through a 0.5mm filter to remove large particulate and organisms. I would use Illumina MiSeq because of its ability to do relatively long paired-end reads affordably. 454 is too expensive given the lower number of sequences obtained and the HiSeq cannot sequences as long of reads as the MiSeq. Long-read sequencers like PacBio would greatly help the assembly of the reads, but would not have enough depth to characterize the rare portion of the community without a prohibitive cost. I would prepare a 300bp paired-end library for the Illumina MiSeq using a variety of insert sizes. I would use DNA-indexes to multiplex the experiment on a single run, using the unique tags to identify both the sample and the insert sizes used.

3

RNA Seq transcriptomics can be conducted without a reference database, but requires one to be useful most of the time. Reference databases are used to identity and infer the function of protein-coding transcripts based off of sequence similarity. This implies that the ability to identify the function of RNA sequences is only as good as the reference database used. If a gene family is not present in a database then sequences of related genes obtained from RNA seq will not be identified. Also, many genes in reference databases only have a predicted function, so many identifications are not very reliable or useful. All the above mentioned limitations also apply to MS-proteomics. In addition, MS-proteomics also requires a reference database to infer protein sequences from MS spectra. In order to translate the raw data obtained from MS, a reference database of proteins is used to simulate the MS pattern each protein would result in. The observed MS patterns are then compared with the reference MS patterns to identify proteins. This means that protein families not present in the reference database will not be sequenced using MS proteomics, making it more reference database dependent than RNA Seq.

4

In bacteria, genome size is typically minimized due to DNA replication time and nitrogen requirements. Since DNA replication is often the limiting factor in the growth rate of fast-growing bacteria, a small genome can provide a selective advantage for bacteria that need to colonize substrate that is only sporadically available. Many pathogens and decomposers have this lifestyle. In low-nutrient environments, a large genome causes bacteria to require more nutrients and energy to replicate and thus is selected against. The presence of transferable plasmides helps to compensate for the loss of adaptability caused by a minimal genome size. A small effective population size in bacteria allows for genome size to shrink due to drift. Drift can reduce the genome size of bacteria by creating pseudogenes which are then selected against due the forces mentioned above.

5

A pathway gap could be detected because another gene takes the place of the missing gene and has not been as well characterized in reference databases. In this case, BLASTp can be use identify proteins in the relevant genome that might be a homolog of the missing protein and perform the same function. If the compensating protein is not a homolog or is too far diverged, then conserved motifs or 3D structure alignments (e.g. iTASSER) can reveal a similar enzymatic function. Potential compensating proteins can be found by comparing ORFs to profile/HMM database. Alternatively, the pathway gap can be real and the organism might not be able to use the pathway. If the end product of the pathway is essential then this can be tested by growing the organism with and without the presence of the end-product of the pathway in the growth medium. If it cant grow without the end product, then it likely is a real gap. Comparing the pathway to that of closely related organisms with a tool such as KEGG could also be informative. If the gap is present in other closely related organisms then it is more likely to be real.