Clint Valentine

BIOL6309 module07

11/15/2016

**Command Line Search with Hmmer**

**Which protein families and clans were identified with this search?**  
The families MH1 and MH2 and the corresponding clans CL0263 and CL0357 were identified from the unknown sequence.  
  
**Give a one-sentence summary of the functions of each family.**  
The MH1 family is separated from the MH2 domain by a non-conserved linker region. The MH1 family is has a conserved residue that is able to bind to the consensus sequence GNCN in DNA. The MH2 family, which does not bind DNA, is responsible for propagating TGF-beta signals by directly associating with the TGF-beta receptor.

**What residues positions have strong conservation in the RHD\_DNA\_bin family of Pfam database?**  
The positions 7, 13, 16, 19, 65, and 97 all have strong conservations as indicated by the large residue logo at those locations.

**After searching PF00554\_seed against the database *pdbaa* how many hits are found?**  
169

**At the end of the HMMsearch output are the matches highly conserved residues still conserved?**  
The *conservation* of sequence in the final results of the *HMMsearch* output are few. For example, the final match has a match score of only 20.8 when compared to the first match at 257.6 and only matches a small fragment of residue (specifically 13 amino acids long). Although there is strong conservation of sequence along the 13 amino acids, the sequence as a whole is barely conserved as much of it is completely missing.

**Find known domains in a query sequence**

**Does the sequence seem to belong to any of the protein families of the HMM library you created?**  
Yes, it appears as if the sequence from *Drosophila melanogaster* matches, in part, much of the seed file for *PF00041* which corresponds to the *fn3* domain. There are a few matches of short size for the *PF00069* seed file which corresponds to the *pkinase* domain. There are no matches for the seed file *PF00076* which corresponds to the *rnn* domain.

**Where can you find such a large library?**  
A large library of HMMs can be found in Pfam which is a curated location for high quality and automatically generated HMMs.

**Compare HMMER to other methods**

After comparing the three methods (*HMMsearch*, *blastp*, and *psiblast*) for searching for *rel* domain homologs it is apparent that all three models return different results with the results from *blastp* and *psiblast* looking more similar in discovery. It appears that *HMMsearch* is looking for protein homology regardless of species as that information is not listed. The two BLAST tools, however, do have a multiple species listed for all protein matches. *Blastp* and *psiblast* each have hundreds of matches yet the output from *HMMsearch* only has169 matches. This could be due to the organization of the reference databases (duplicity of protein sequences due to homology) or it could be due to a more discriminating method for the *HMMsearch*. The readout and results are far easier to read and interpret for the *HMMsearch* tool and I would be inclined to use it again especially if its ability to discriminate true homologous proteins is better than *blastp* and *phiblast*.