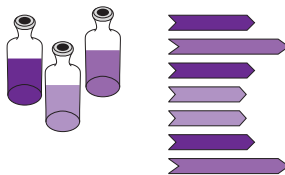


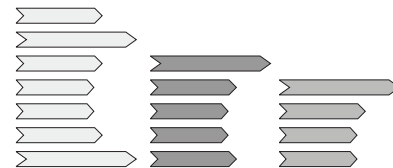
Select marker gene(s) for pathway of interest.



Gather sequences from organisms that have been described to have the pathway.



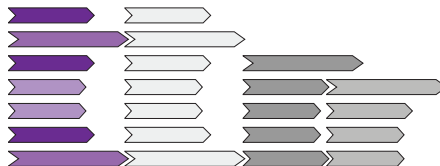
Gather sequences from the same KO or sequences sharing the same PFAM domains.



Phylogenetically define clades of true positives (TPs) and false positives (FPs).



Build profile HMMs for all clades, TPs and FPs

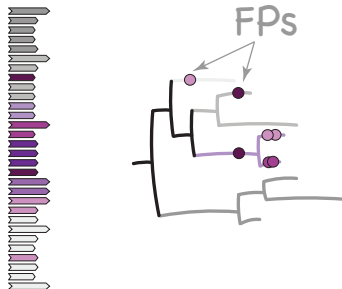


Screen input data using homology sensitive methods against all profiles (e.g. Hidden Markov Models). Consider only the highest scoring hit per sequence.

```
>seq1
MTPEDCVTTWSVATR
>seq2
MTPGDCVSTWSVATD
```



Align all TP hits and reference sequences and validate TP hits through placing them on the tree containing all reference sequences.



Optional: consider operon-like structure or overall pathway completeness as additional quality controls.

