Hidden Markov Model Exploration.

Orfeas Gkourlias
Hanzehogeschool Groningen

Sequences.

De MSA die gebruikt is voor het maken van de HMM staat in de POU5F1.aln file. Hier staan 7 homologen voor hetzelfde, verspreid over verschillende organismen. Het bestand is in fasta formaat gedownload van NCBI.

Daarnast is er nog een fasta bestand. Pou_fam.fasta. Dit bestand is van Pfam gehaald (Dat nu blijkbaar onder interpro valt). Hierbij is er op het Pou domijn geflitered, om vervolgens alle opgeslagen sequenties hierbij op te slaan. Het resultaat is deze fasta file.

HMM.

Om de opdracht uit te voeren heb ik gebruik gemaakt van een zelfgeschreven script. Het programma werkt als volgt: Het .ALN Bestand wordt ingeladen, om hier vervolgens shell commandos mee uit tevoeren. Hierbij is gebruik gemaakt van de subprocess package die standaard in python3. Er zijn vier verschillende functies op de commandline losgelaten, in deze volgorde:

1. Hmmbuild.
2. Hmmpress.
3. Hmmscan.
4. Hmmsearch.

HMMBuild.

HMMER Makat hierbij de HMM profielen voor elke sequence alignment in de .aln file. Dit genereert vervolgens een bestand die niet echt leesbaar is, maar waar wel de verdere functies op uitgeoefend kunnen worden.

HMMPress.

De komende functies moeten een bepaalde bestandsstructuur hebben, die nog niet in de vorige file is voorzien. Deze functie zorgt er voor dat er nieuwe bestanden gemaakt worden die er voor zorgen dat de functies wel goed werken. Er kon gekozen worden om het resultaat hiervaan leesbaar te maken, maar dat is nog niet belangrijk, en wordt in de volgende functie gedaan.

HMMScan.

Dit is de eerste functie die daadwerklijk met een database gaat zoeken. De database die gebruikt wordt zit ook in de directory. Het gaat hier om de .fasta file. Het bestand is een collectie van eiwit sequenties die in hetzelfde domein zitten, het POU domein.

Er komt uit deze functie een output die in een bestand wordt opgeslagen in het data document. Dit bestand toont aan welke eiwit profielen het meest overeenkomen met die van de alignments, door middel van E scores. Op deze manier kan er bepaald worden welke profielen het dichtsbij deze alignments komt.

HMMSearch.

Ten slot word er gebruikt gemaakt van de HMMsearch functie. Dit gaat naar individuele sequenties kijken, en niet aar profielen. Dit komt weer uit dezelfde .fasta database file. Hierbij kan er specifieker gekeken worden welke sequenties nou eigenlijk significant vergelijkbaar zijn. Net zo als de vorige functie komt hier een bestand uit, waarin de vergelijkbaarheid met E scores is aangetoond.