Verslag

Big data wordt steeds belangrijker in de bio-informatica wereld. Om goed met dit soort data om te kunnen gaan, zijn er verschillende omgevingen die gebruikt kunnen worden. Een van die omgevingen is Hadoop. Hadoop heeft een hadoop distributed filesystem (hdfs) waar gemakkelijk een grote hoeveelheid data op kan staan. Hadoop heeft een framework genaamd mapreduce die grote hoeveelheden data kan analyseren in parallel op grote clusters (“MapReduce Tutorial”, z.d.). Mapreduce heeft een map gedeelte en een reduce deel. Het map gedeelte filtert de data die nuttig is voor een bepaald probleem en het reduce deel maakt de berekening op de gefilterde data.

In dit onderzoek hebben we op een ‘grote’ eiwitsequentie (in fasta format) gedownload van Uniprot een berekening uitgevoerd. We hebben een functie geschreven die alle aminozuren telt en vervolgens het aantal voorkomen van een aminozuur vermenigvuldigt met een specifieke waarde die is meegegeven in het script. Hierbij is een waarde -1 apolair, 0 neutraal en +1 polair. Als voorbeeld komt het Aminozuur R (Arginine), is een polair aminozuur, 4 keer voor in de sequentie, dan is de berekening voor dat aminozuur 4 \* +1 is 4. Als alle aminozuren zijn geteld en vermenigvuldigd, worden de berekende getallen bij elkaar opgeteld. Vervolgens wordt er gekeken of dit getal positief, neutraal of negatief is, of te wel polair, neutraal en apolair.

Om te kijken of het script wel functioneert, is er eerst een testsequentie gemaakt met een verkorte sequentie “ACGAPPEL”. Hier hebben we zelf de aminozuren geteld en vermenigvuldigd met de polariteitwaarde. Deze waarde hebben we vergeleken met de waarde die uit het script komt. De zelf berekende waarde en de waarde uit het script kwamen overeen met elkaar.

De eiwitsequentie die we voor dit onderzoek gekozen hebben, is haptoglobine. Dit is een eiwit, wat gevormd wordt in de lever, die helpt bij de transport van hemoglobine uit rode bloedcellen. Wanneer een patiënt een te kort heeft aan haptoglobine, kan dit leiden tot bloedarmoede, omdat de rode bloedcellen uit elkaar vallen en niet worden gerecycled doordat hemoglobine niet wordt gebonden aan haptoglobine (Uniprot Q61646). De waarde die hieruit komt is +55, dit betekent dat de sequentie van haptoglobine polair is.

Het script kan redelijk simpel worden uitgebreid door bijvoorbeeld meerdere eigenschappen van aminozuren toe te voegen. Ook kunnen er meerdere eiwitsequenties worden toegevoegd aan je hdfs systeem, waardoor je, in theorie, meerdere berekeningen kan doen in parallel.

Literatuur

*MapReduce Tutorial*. (z.d.). Geraadpleegd op 18 juni 2020, van https://hadoop.apache.org/docs/r1.2.1/mapred\_tutorial.html