Task 1 - Implementation and application of dotplot

Prvá časť – popis dotplotu

DotPlot – je metóda využívaná v bioinformatike na otestovanie podobností v dvoch biologických sekvenciách. Využíva sa nato grafická reprezetácia v podobe obrázku, ktorá obsahuje biele pixely v zhode dvoch sekvencii. Na osiach sa nachádzajú jednotlivé sekvencie. Po vytvorení obrázku je možné sledovať rôzne úkazy medzi sekvenciami (podobnosti, mutácie,...). **Zhody** sa na obrázku prejavia vykreslením bielej diagonálnej čiary. **Mutácie** vykazujú rozdiely medzi sekvenciami a na obrázku sa javia ako medzery medzi diagonálnymi čiarami (rozdeľujú zhody). **Vkladania** sú zas časti sekvencii, ktoré sú v jednej sekvencii, no nie v druhej. Graficky sa javia ako medzery medzi čiarami, ktoré sú zároveň posunuté po nejakej osi. **Vymazania** sú také časti sekvencií, ktoré boli z jednej sekvencie vymazane (podobné s vkladaniami). Pomocou dotplotu môžeme v sekvenciách nájsť redundantne časti takzvané **low-complexity** úseky, ktoré sa graficky javia ako obdĺžnikové oblasti so zhodami. Dotplot-y majú mnoho výhod, ale aj nevýhod.

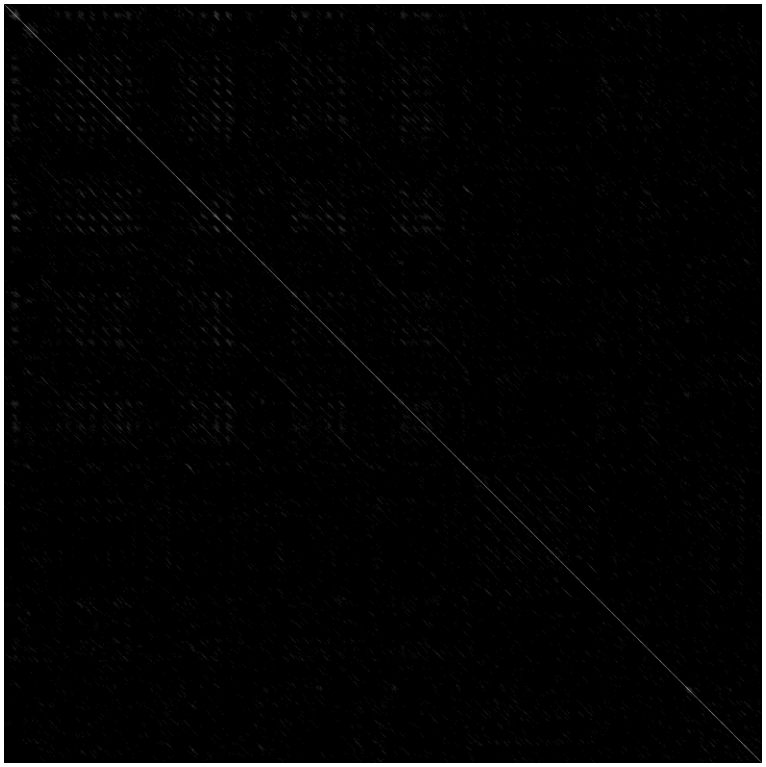
Výhody :

- vytvárajú ľahko interpretovateľný graf.
- možno použiť na porovnanie krátkych, ale aj veľmi dlhých sekvencií (celý kód chromozómu).

Nevýhody :

- nutnosť hľadať najlepšiu veľkosť okna a thresholdu
- porovnávanie najviac dvoch rôznych sekvencií.

Druhá časť – SLIT DROME



Obrázok 1 : Dotplot v ktorom porovnávame sekvenciu SLIT_DROME so sebou. Veľkosť okna je nastavená na 11 s thresholdom rovným 1.

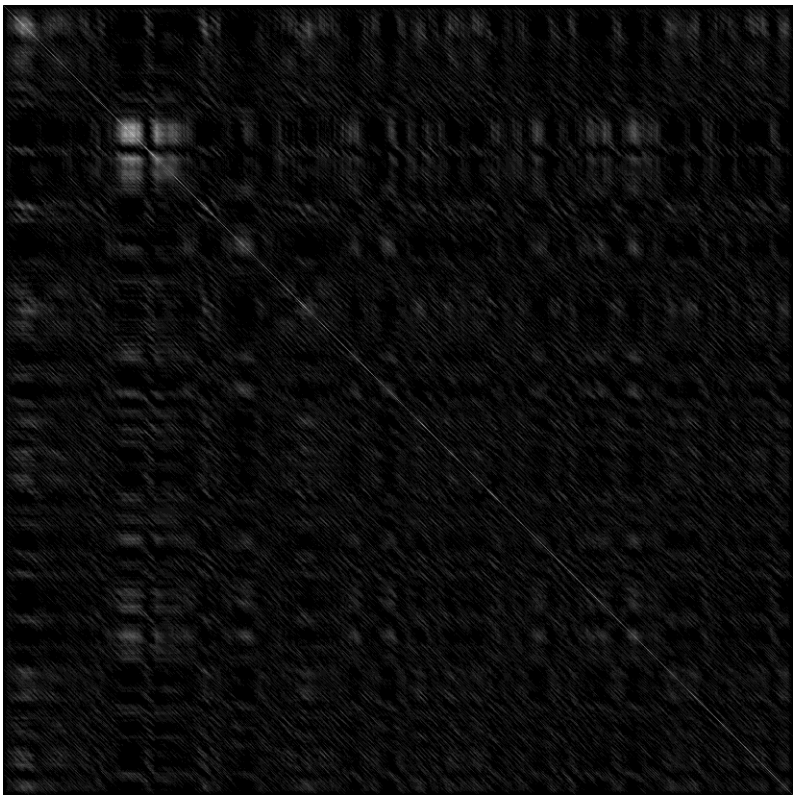
Ako je vidieť v dotplote sa nachádza viacero oblastí, ktoré sa v grafe javia ako prerušované štvorce/obdĺžniky, čo znamená, že sa v týchto oblastiach opakujú určité časti sekvencií viacero razy (tzv. minisatellite patterns). Po nahliadnutí na obrázok sa dajú vymedziť okraje týchto 5 štvorcov pomocou pozicií :

- [101,257]
- [330,457]
- [561,671]
- [755,880]
- [938,1178]

Po zhliadnutí tohto proteínu v databázi EBI na stránke http://www.uniprot.org/uniprot/P24014 sú v prvých 4 regiónoch 19 LRR (leucine-rich) úsekov. Posledný štvorec obsahuje 7 EGF domén.

Tretia časť – NM_000044

Podľa databáz EBI a NCBI je DNA sekvencia s označením NM_000044 tiež nazývaná Homo sapiens androgen receptor AR. Sekvencia je má ako zdroj uvedeného človeka. Dotplot sekvencie NM_000044 s veľkosťou okna 25 kvôli veľkosti obrázku do dokumentu nepridám, namiesto neho vložím obrázok dotplotu stredu sekvencie => 4000-6000 pozicia v sekvencii.

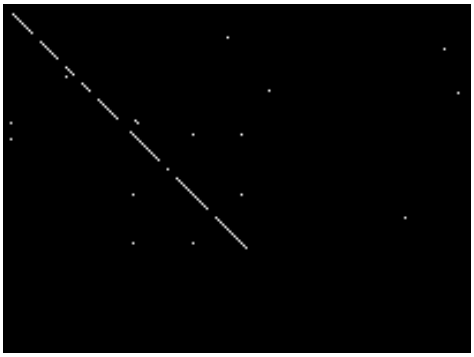


Obrázok 2 : DotPlot kde porovnávame stred sekvencie NM_000044 so sebou.

V tomto dotplote je vidno 2 opakujúce sekvencie na pozíci [4276,4352] a [4372,4450]

Štvrtá časť – hemo vs. Pásovec

Na zistenie najdlhšej identickej sekvencie využijem dotplot program s tým, že nastavím okno na 3 a threshold na 2 (okno veľkosti 3 musí obsahovať všetky aminokyseliny rovnaké, aby sme vykreslili pixel), čím dostaneme graf, v ktorom je vidieť iba spoločné sekvencie. Teraz nám stačí vybrať tú najdlhšiu z nich a ku každej strane pridať ešte jednu aminokyselinu kvôli zvoleniu veľkosti okna rovnej 3.



Obrázok 3 : Porovnanie sekvencií hemoglobin, pásovec. Threshold je 2 veľkosť okna je 3.

Jedna z najdlhších sekvencií je na pozícií [74,89] čiže má dĺžku 16. Sekvencia obsahuje

AVTNVEDLSSLEEYLA.
Pre zistenie najdlhšej moc podobnej sekvencie môžme takisto využiť dotplot s veľkosťou okna 3 a thresholdom rovným 1 (okno veľkosti 3 musí obsahovať aspoň 2 podobné aminokyseliny, aby sme vykreslili pixel).

Obrázok 4

V takomto prípade má najdlhšia podobná sekvencia veľkosť 53 s tým, že najviac 16 aminokyselín je v tomto úseku rovnakých. Tieto dve dané sekvencie očividne majú spoločného predka kvôli prerušovanej čiare.

Piata časť – HIV

Na DNA sekvenciu vírusu Human immunodeficiency virus type 1, isolate PV22, complete genome získaného zo stránky http://www.ebi.ac.uk/ena/data/view/K02083 po vykreslení dotplotu dostávamé obrázok, na ktorom sú v ľavom spodnom a v pravom hornom rohu rovnobežné čiary s diagonálou. Takéto druhy čiar indikujú opakujúce sa oblasti v smere čítania na rôznych častiach sekvencie (duplikáty cca [0,662] úsek sekvencie). Obrázok kvôli príliš veľkej veľkosti nepridávam.