Uvod
aky je problem : kecy

DNA sa da predstavit ako retazec (C), guanine (G), adenine (A), or thymine (T) priblizne 10 na 6 az 10 na 10 dlhe.
Tieto retazce urcuju takmer vsetky charakteristiky organizmu.
Preto nas zaujimaju rozdiely v tychto retazcoch. Teky rozdiel moza napriklad znamenat nador.

Problemom je, DNA nevieme merat priamo. Vacsina technologii vie hovorit kratke substringy DNA nazyvane , ready a piared ready. Tie prerobyme na k-mery a tie potom skladaju do deBruijnovych grafov.
Vrcholy su to jednotlyve Kmery a hrany su medzi kmermi ktore maju zhodu na k-1 bazach.
Z debruijnovho grafu je mozne zrekonstruovat povodnu sekvenciu.
Problem je, ze vacsinou tento graf vyzera asi takto. Rekonstrukcia sekvencii je porom taska a nejednoznacna.
Vieme hladat varianty nie v sekvneciach,ale priamo v readoch.

Ak mame dve sekvencie v ktorych je nejaky variant, mozem z nich urobit farebny de bruijnov graf. To je ako klasickz, ale kazdy vrchol ma farbu, podla toho, z ktorej sekvencie je.
Varianty sa nam potom prejavia ako bubliny v tomto grafe.

Moju pracu som si rozdelil takt, a mozme sa pozriet na to, co uz je urobene, pricom som uz napisal TOTO.
Pozrel som sa na sekvenovacie technologie a obznamil som sa s ich specifikami.
Pozrel som sa na skladanie genomov, ktore je relativne blizke hladaniu variantov. Zistil som, ze toto skladanie je vacsinou zalozene na heuristikach, a ze informacia z paired readov sa pouziva vacsinou len pri postprocesingu.

Nasledne som sa pozrel na samotne hladanie variantov. Beznou technikou je hladanie variantov medzi readmi a referencnou sekvenciou. Nejvacsim problemom je tu rychlo a dobre urcit polohu readov v referencnej sekvencii. Na to sa vyuzivaju hlavne specialne hashovacie funkcie.
Tieto pristupy su caste velmi slabo (zvlastne) pravdepodobnostne podlozene.

Co sa tika hladania variantov v colored grafoch, mozme spomenut pracu panov …… . Ty priniesli popis roznych typov variantov a toho, co vytvoria v debruijnovom grafe. Pre tito varianty ale nevedia povedat, ako su pravdepodobne. Tak isto nepouzivaju informaciu o paired readoch.

Dvolezitym krokom v hladani variantov je enumeracia bublin. To sa ukazalo byt NP tasky problem, Sacamoto ale priniesol polynimial delayed algoritmus.

co chceme urobit v najblizsom case
done