Domača nalog #3:Najdi eksone

Karmen Bezlaj (27132012)

21. november 2013

1 Uvod

Naša naloga je bila v DNA sekvenci, čim bolj natančno napovedati podane gen. Da lahko primerjamo gen, ki je bil podan z aminokislinami, ter DNA, podana z nukliotidi, moramo uskladi zapis. Odločimo se, da DNA v vseh oknih prevedemo v aminokisline. Uporabili bomo alogitma Needleman–Wunsch za globalno poravnavo in Smith–Waterman za lokalno poravnavo. Na koncu moramo za vsak od podanih genov vrniti, kje se nahajajo, odstaniti moramo introne.

2 Podatki

Za nalogo smo potrebovali naslednje tekstovne datoteke:

- dnaC00C39.txt
- dnaC40C79.txt
- aminoacids.txt

Vse dostopne na spletni strani http://193.2.72.57/uvb2013dn3/. Prva datoteka je testna, na njej smo poitzkušali naš algoritem, druga pa je takorekoč odločilna, saj je rezultat naše naloge določen na podlagi rezultata, ki ga pridobimo na njej. Tretja datoteka je datoteka, v katri so shranjeni znani geni.

3 Metode

Od zgoraj omenjenih algorimov bomo potreboali samo Smith–Watermanov algoritem za lokalno poravnavo. Oba smo dobili, kot predlogo v datoteki alignment.py na spletni strani https://ucilnica.fri.uni-lj.si/mod/assign/view.php?id=25573.

Moja ideja je bila za vsak gen s pomočjo Smith–Watermanovega algoritma poiskati potencialno lokacijo za gen v vseh treh oknih v obeh smereh z "filtrom" *blosom50* in "kazijo-10. Z pomočjo funkvcije *izpisi(s,t,mat,pr,loc_score,risi)* pridobim lokacijo v posamazni sekvenci (eda od smeri v enem od oken), kako daleč od začetka gena se poravnava nahaja in dolživo najbolše kosa poravnave.

S pomočje teh inforamaciji zaženem rekurzivno funkcijo *rekurzija(z, k, th, n, s, t)*, ki deluje na podsekvenci. Podsekvenca se začne v *z*, ki začetek poravnave pomanjša za oddaljenost poravanve od začetka gena in konča v *k*, ki je vsota začetka poravnave golžine poravnave ter dložine gena. Delovanje rekurzije je sledeče na vsakem koraku izračunamo začetek in konec podsekvence, ki se najbolje ujema z genom, zmanšamo kazen za algoritem za -5 (v algoritem smo ustopili z kaznijo -30) in ponovimo postopek na dveh manših podsekvencah, ki ne vsebujejo tega dobrega dela, ki ga vrenemo in predstavlja delj eksona. Rekurzija ustavimo, ko so delčki približno 20% velikosti gena.

Na koncu še pregledamo kose eksonov in iščemo vse koščke, ki se razlikujejo za 10 amino kislin in jih zavržemo. Vse skupaj vrnemo v ustezi obliki za oddajo.

4 Rezultati

V metodi je opisana moja ideja za algoritem, najdemo ga v datoteki *3DN_adv.py* vendar po štirih urah algoritem še ni zaključil svojega dela, zato sem proces prekinila, ter podala rešitev, ki mi jo je podal program *3DN.py*, ki deluje kot zgoraj opisano, le brez rekurzije. Za testne podatke sem dobila oceno: 23178

Tabela 1: Napoved genov.

oznaka gena	začetek gena	konec gena	oznaka gena	začetek gena	konec gena
C57	22443	22698	C56	12540	12804
C55	534	792	C54	12504	12603
C53	2931	3081	C52	9504	9780
C51	20433	20679	C50	15087	15291
C71	14151	14460	C70	11487	11991
C73	7626	8400	C72	9066	9777
C75	4989	5718	C74	3717	3894
C59	19971	20583	C58	12132	12471
C68	10614	11244	C69	20184	20403
C67	13398	13512	C44	279	882
C45	19038	19383	C46	15468	16155
C47	15921	16554	C40	24300	24996
C41	17496	18012	C42	14205	14544
C43	25743	25836	C66	23955	24420
C62	4077	4293	C64	468	603
C65	9447	9975	C48	13176	13470
C49	8535	9384	C60	22617	22926
C61	21069	21594	C77	19407	19545
C76	18666	19344	C79	25005	25926
C63	10605	10812	C78	22698	23199

5 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.