Start ir stop kodonų poros:

Visų pirma, suskirstome seką pagal skaitymo rėmelius(3 postūmiai per vieną simbolį). Tą patį padarome ir reverse komplementui. Gauname 6 aibes (pažymėkime jas **a1,a2,..,a6**) su koduotėmis po tris, $\{a1, \dots, a6\} \in$**G** (pvz „AABCDAXAR“ => {„AAB“ „CDA“ „XAR“}, {„ABC“,“DAX“,“AR“},{„BCD“,“AXA“,“R}..... (analogiškai ir reverse komplemetui)). Tuomet surandame dvi aibes {b1,b2}, kuriose išsaugomos start ir stop pozicijos kiekvienai aibei **R**. Iteruojame per abi start ir stop pozicijų aibes ir ieškome start ir stop porų, tarp kurių nebūtų stop. (star < stop). Gauname atsakymą, kuriame pateikiamos start stop koordinatės viename iš 6 gautų skaitymo rėmelių aibių.

Stop kodonui toliausią esantį start kodoną:

Kaip ir start ir stop kodonų porų atveju turime start ir stop pozicijų aibes. Iteruojame konkrečiame skaitymo rėmelyje su konkrečia start ir stop kodonų aibe, taip, kad būtų rastos koordinatės, kad start būtų toliausiai nutolęs nuo stop ir tarp jų nebūtų stop. (stop < start)

Atfiltruoti fragmentus, kurie trumpesni nei 100 fragmentų:

Čia panaudojame poras gautas su pirma funkcija. Turime porą (x,y), kur x start kodono pozicija, y stop kodono pozicija. Kadangi rėmelis sudeda į aibę po tris elementus, norėdami gauti sekos ilgį turime naudoti šią formulę: (y-x+1)*3

Įvertinti kodonų dažnius:

Kodonų bei dikodonų dažnių formulė: dažnis_kodonas = konkretus_kodonas/visi_kodonai, dažnis_dikodonas = konkretus_dikodonas/visi_dikodonai.

Eiga: surandame kiek yra konkrečių kodonų/dikodonų tarp start ir stop kodonų/dikodonų porų visuose rėmeliuose, kurie ilgesni nei 100 fragmentų. Šiuos skaičius padaliname iš šių kodonų/dikodonų sumos. Gauname konkretaus kodono/dikodono dažnį sekoje.

Atstumo funkcija: Atimame kodono/dikodono dažnį tarpusavyje tarp sekų, skirtumui pritaikome absoliutaus dydžio funkciją (modulį), visus dažnius susumuojame ir iš sumos ištraukiame kvadratinę šaknį. Pavyzdžiui:

Seka1: („AAA“, 0.9), („GGT“, 0.2)

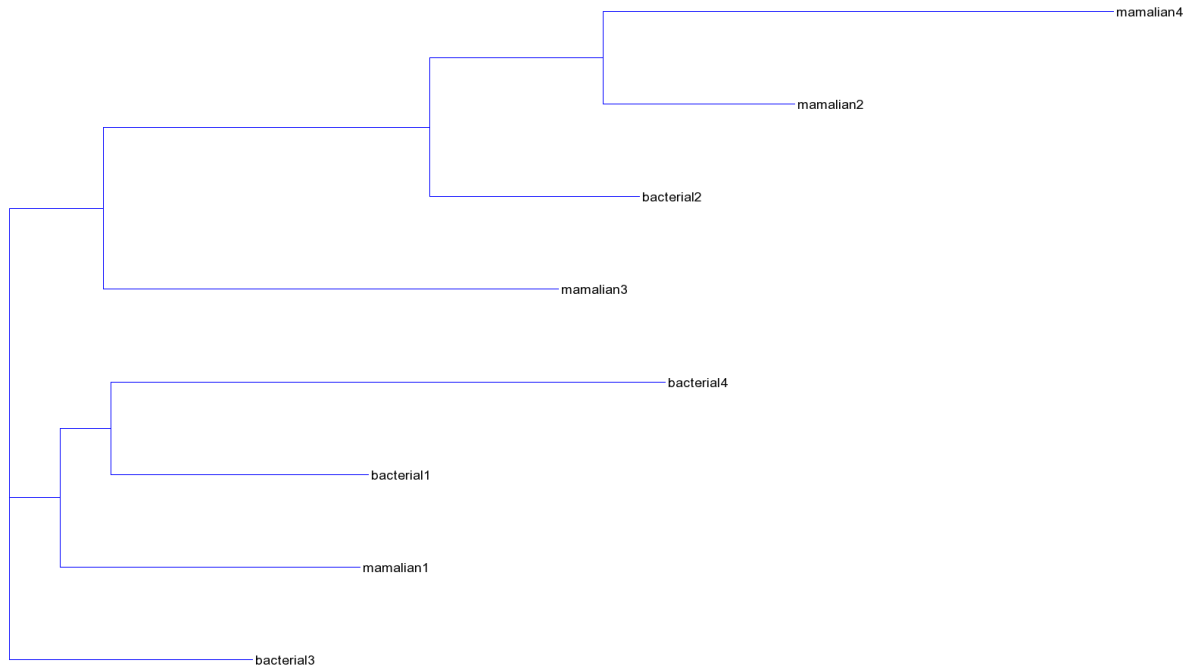
Seka: („AAA“, 0.7),(„GGT“, 0.14)

Atstumas = $\sqrt{abs(0.9 - 0.7) + abs(0.2 - 0.14)}$

Kodonai atstumų matrica

8
bacterial1 0.0000000 0.2440857 0.1333937 0.2022245 0.1535006 0.2797698 0.2314605 0.3804349
bacterial2 0.2440857 0.0000000 0.1981244 0.3225986 0.2416587 0.1455292 0.2680976 0.2199708
bacterial3 0.1333937 0.1981244 0.0000000 0.2319793 0.1559255 0.2475716 0.2343881 0.3256150
bacterial4 0.2022245 0.3225986 0.2319793 0.0000000 0.2230188 0.3780934 0.2650636 0.4445399
mamalian1 0.1535006 0.2416587 0.1559255 0.2230188 0.0000000 0.2771983 0.2158819 0.3698648
mamalian2 0.2797698 0.1455292 0.2475716 0.3780934 0.2771983 0.0000000 0.2823126 0.1746668
mamalian3 0.2314605 0.2680976 0.2343881 0.2650636 0.2158819 0.2823126 0.0000000 0.3464100
mamalian4 0.3804349 0.2199708 0.3256150 0.4445399 0.3698648 0.1746668 0.3464100 0.0000000

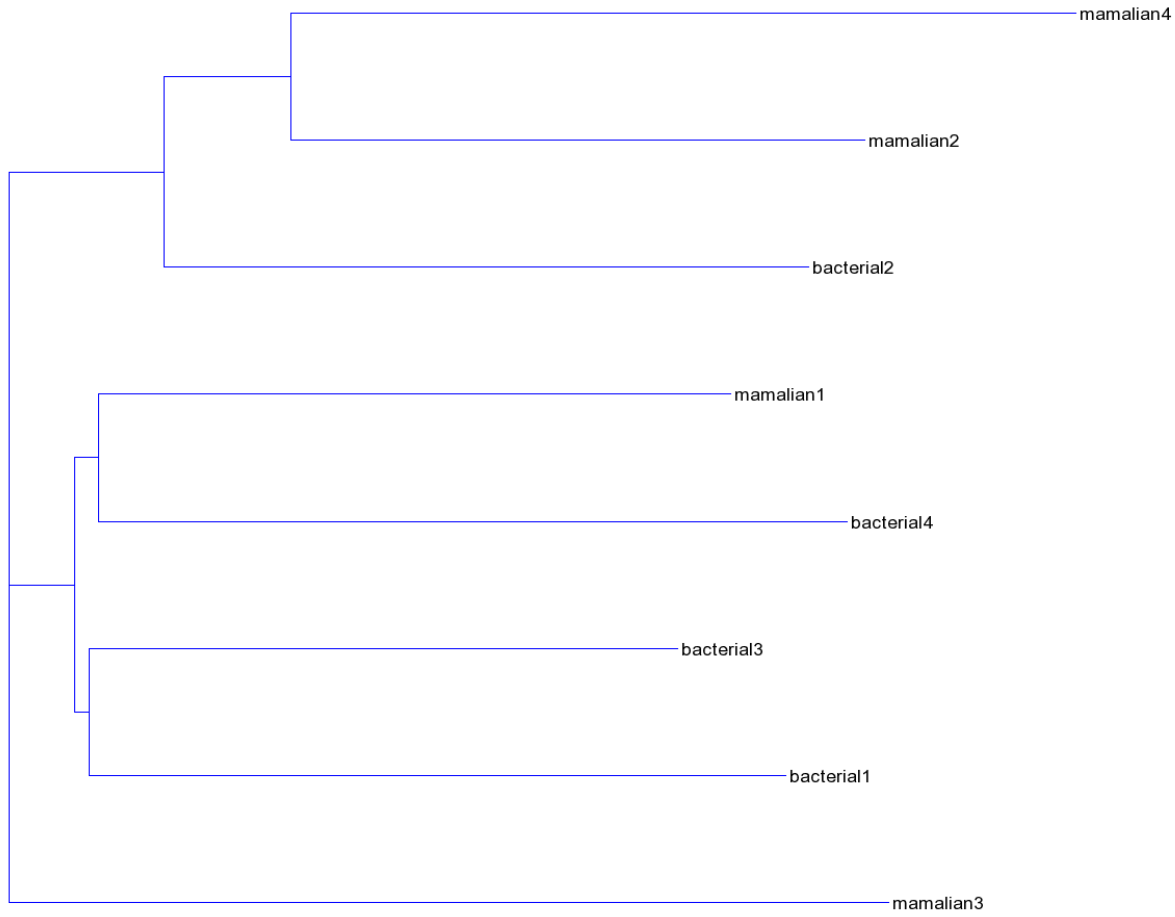
Kodonų medis



Dikodonai atstumų matrica

8
bacterial1 0.0000000 0.5637819 0.4672461 0.5210289 0.5082506 0.5926989 0.6060408 0.6800986
bacterial2 0.5637819 0.0000000 0.5218947 0.6023827 0.5545670 0.4925065 0.6207678 0.5623780
bacterial3 0.4672461 0.5218947 0.0000000 0.5084752 0.4542483 0.5336070 0.5980108 0.6215342
bacterial4 0.5210289 0.6023827 0.5084752 0.0000000 0.5017573 0.6417640 0.5852128 0.7053364
mamalian1 0.5082506 0.5545670 0.4542483 0.5017573 0.0000000 0.5558031 0.5808843 0.6578100
mamalian2 0.5926989 0.4925065 0.5336070 0.6417640 0.5558031 0.0000000 0.6392920 0.4934961
mamalian3 0.6060408 0.6207678 0.5980108 0.5852128 0.5808843 0.6392920 0.0000000 0.6876600
mamalian4 0.6800986 0.5623780 0.6215342 0.7053364 0.6578100 0.4934961 0.6876600 0.0000000

Dikodonų medis



Labiausiai paplitę kodonai ir dikodonai sekose:
bacterial1: 'AAA', 'TTCTTT'
bacterial2: 'GAT', 'TAAATG'
bacterial3: 'GCT', 'GCTGGT'
bacterial4: 'AAA', 'GATGAT'
mamalian1: 'GTT','GGTGTT'
mamalian2: 'ATG', 'TAAATG'
mamalian3: 'ATA', 'TAAATG'
mamalian4: 'ATG', 'TGAATG'

Klasterizacija:

Ir Kodonų, ir dikodonų atveju mamalian4, mamalian2 bei bacterial2 klasterijuoasi atskiroje šakoje. Dikodonų medyje struktūra nežymiai pasikeičia:mamalian1 klasterizuojasi su bacterial4, bacterial3 klasterizuojasi su bacterial1. Visi jie klasterizuojasi atskiroje šakoje (mamalian1,bacterial4,bacterial3,bacterial1). Tuo tarpu kodonų medyje atskiroje šakoje klasterizuojasi bacterial4, bacterial1, mamalian1. Kodonų medyje atskirą šaką turi bacterial3, dikodonų mamalian3.