**Bioinformatika**

1 laboratorinis darbas

Tomas Ukrinas

Programų sistemos 4 kursas 5 grupė

**Suraskite visas baltymų sekas prisegtoje plazmidės sekoje (remiantis start ir stop kodoną - iš kelių galimų baltymų ORF'e rinkitės ilgiausią), atmesti mažesnes nei 100 bp koduojančias sekas.**

Gauti 9 baltymai:

MAQNLSRYGMIAPGRESIQGGECETSNVIRCRRVCRCLLSDRFPRGEPGQPRFCENAGKSGSGDGGAELHSQPRGTTTGGQTVVADWRCHLQSGPARAVANCRGD

MLYPAVNHHQTGFSPAGANQRGPLAATLSGPGGEGQSAVARLTGEKKNHPGAQYANRLSPRVGRFINAAGTTGFPTGKRAVIPDLHRRMLLATLWNTYICINEALALTLSDFSLVPPHPYRQLFTLTTFQ

MRGSHHHHHHGMASMTGGQQMGRSGDDDDKLKIAAFNIRTFGETKMSNATLASYIVRIVRRYDIVLIQEVRDSHLVAVGKLLDYLNQDDPNTYHYVVSEPLGRNSYKERYLFLFRPNKVSVLDTYQYDDGCESCGNDSFSREPAVVKFSSHSTKVKEFAIVALHSAPSDAVAEINSLYDVYLDVQQKWHLNDVMLMGDFNADCSYVTSSQWSSIRLRTSSTFQWLIPDSADTTATSTNCAYDRIVVAGSLLQSSVVPGSAAPFDFQAAYGLSNEMALAISDHYPVEVTLT

MAELNYIPNRVAQQLAGKQSLLIGVATSSLALHAPSQIVAAIKSRADQLGASVVVSMVERSGVEACKAAVHNLLAQRVSGLIINYPLDDQDAIAVEAACTNVPALFLDVSDQTPINSIIFSHEDGTRLGVEHLVALGHQQIALLAGPLSSVSARLRLAGWHKYLTRNQIQPIAEREGDWSAMSGFQQTMQMLNEGIVPTAMLVANDQMALGAMRAITESGLRVGADISVVGYDDTEDSSCYIPPLTTIKQDFRLLGQTSVDRLLQLSQGQAVKGNQLLPVSLVKRKTTLAPNTQTASPRALADSLMQLARQVSRLESGQ

MLRSPATLFGSCVVTTSSSSRRSETATWWLWGSSWTISTRMTQTPTTMWSVSRWAATATRSATSFCSDPTRCPCWTPTSTTTAASPAGTTASAGSPLWSSSHPTPPRSRNLPLLPCTRPHRTQWLRLILSTMSTWMSSRSGT

MGVMIPMKRERMLTIRVTDDEHARLLERCEGKQLAVWMRRDQRKITQGQCQRFVNTDVGVPQGSQQHPAMQIRNYCPLSSRETCRASCINESANARGEAVCVLGARVVFLFTSETGNS

MSIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAEDQLGARVGYIELDLNSGKILESFRPEERFPMMSTFKVLLCGAVLSRVDAGQEQLGRRIHYSQNDLVEYSPVTEKHLTDGMTVRELCSAAITMSDNTAANLLLTTIGGPKELTAFLHNMGDHVTRLDRWEPELNEAIPNDERDTTMPAAMATTLRKLLTGELLTLASRQQLIDWMEADKVAGPLLRSALPAGWFIADKSGAGERGSRGIIAALGPDGKPSRIVVIYTTGSQATMDERNRQIAEIGASLIKHW

MRGSHHHHHHGMASMTGGQQMGRSGDDDDKLKIAAFNIRTFGETKMSNATLASYIVRIVRRYDIVLIQEVRDSHLVAVGKLLDYLNQDDPNTYHYVVSEPLGRNSYKERYLFLFRPNKVSVLDTYQYDDGCESCGNDSFSREPAVVKFSSHSTKVKEFAIVALHSAPSDAVAEINSLYDVYLDVQQKWHLNDVMLMGDFNADCSYVTSSQWSSIRLRTSSTFQWLIPDSADTTATSTNCAYDRIVVAGSLLQSSVVPGSAAPFDFQAAYGLSNEMALAISDHYPVEVTLT

MLRSPATLFGSCVVTTSSSSRRSETATWWLWGSSWTISTRMTQTPTTMWSVSRWAATATRSATSFCSDPTRCPCWTPTSTTTAASPAGTTASAGSPLWSSSHPTPPRSRNLPLLPCTRPHRTQWLRLILSTMSTWMSSRSGT

Rezultatai taip pat pateikti faile *res.txt* (kartu surastais ORF)

**Pabandykite apskaičiuoti, kurio ORF'o kodonai/dikodonai labiausiai skiriasi nuo kitų ORF'ų**

Gauta atstumų matrica, pateikta faile *codon\_distance\_matrix.txt*

Matrica išvesta naudojant Phylip formatą.

Naudojant jūsų rekomenduotą puslapį, suformuotas medis neighbor joining metodu:



Jame galime pamatyti, kad labiausiai išsiskiria ketvirtasis baltymas.

**Pabandykite apskaičiuoti, kurio ORF' G/C simboliu daznis skiriasi nuo ORF'u**

Iš faile res.txt išspausdintų GC simbolių dažnių, matome, jog labiausiai skiriasi 7 ORF, kurio GC dažnis yra ~49%.

7 ORF:

ATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGCAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGTTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAA

GC: 49.24506387921022

Visi kiti GC dažniai yra ~55%