EBO bioinformaatika praktikum
05.06.2021

Juhendajad Reidar Andreson, Age Brauer

Õpilase nimi: _____

Kooli nimi: _____

Tööks vajalike programmide lingid ning analüüsitavad järjestused leiate ülesannete kaupa veebilehelt https://bioinfo.ut.ee/EBO/

Töö tegemiseks on aega 90 minutit ja maksimaalne punktide arv küsimuste peale kokku on 100.

Töö koosneb kolmest osast, mille all on vaja vastata kahele või rohkemale küsimusele (tähistatud tähtedega: a, b, c...). Osade ülesannete puhul on antud lisainformatsiooni bioloogiliste mõistete või programmide kohta. Kõik järjestused ja lingid programmidele on ülesannete kaupa kättesaadavad ülal antud veebiaadressilt (näit. osa_x_ylesanne_y.html).

OSA I Homoloogiaotsingud programmiga BLAST

BLAST (*Basic Local Alignment Search Tool*) on mõeldud kahe järjestuse (DNA ja/või aminohappe) vahelise homoloogia otsinguks. Programm väljastab tulemusena päringjärjestusele kõige sarnasemad järjestused, kus kõige suurema skooriga (*bitscore*) vaste andnud tulemus näidatakse esimesena. Üldjuhul tähendab suurim skoor ka pikimat homoloogset ala järjestuste võrdluses. NCBI BLAST tulemuste veebilehel näidatakse kasutajale kahte skoori väärtust - maksimaalne skoor (*Max Score*) ja täielik skoor (*Total Score*). Esimene neist näitab kõige suuremat skoori joonduse sees, mis on seotud pikima homoloogia andnud alamjärjestusega joonduses. Teine neist tähistab joonduse kõigi homoloogsete lõikude summaarset skoori.

Bioloogilisi järjestusi - nii nukleotiidseid kui ka aminohappelisi - hoitakse enamasti spetsiaalses tekstifailis FASTA formaadis (Joonis 1). Iga uus kirje FASTA formaadis failis algab kirjeldava reaga, mille alguses peab olema alati suurem-kui märk (“>”). Tavaliselt sisaldab see rida järjestuse nime ja unikaalset identifikaatorit. Kirjeldavale reale järgneb bioloogiline järjestus kas ühel või mitmel järjestikusel real. Üks FASTA formaadis fail võib sisaldada rohkem kui ühte järjestust.

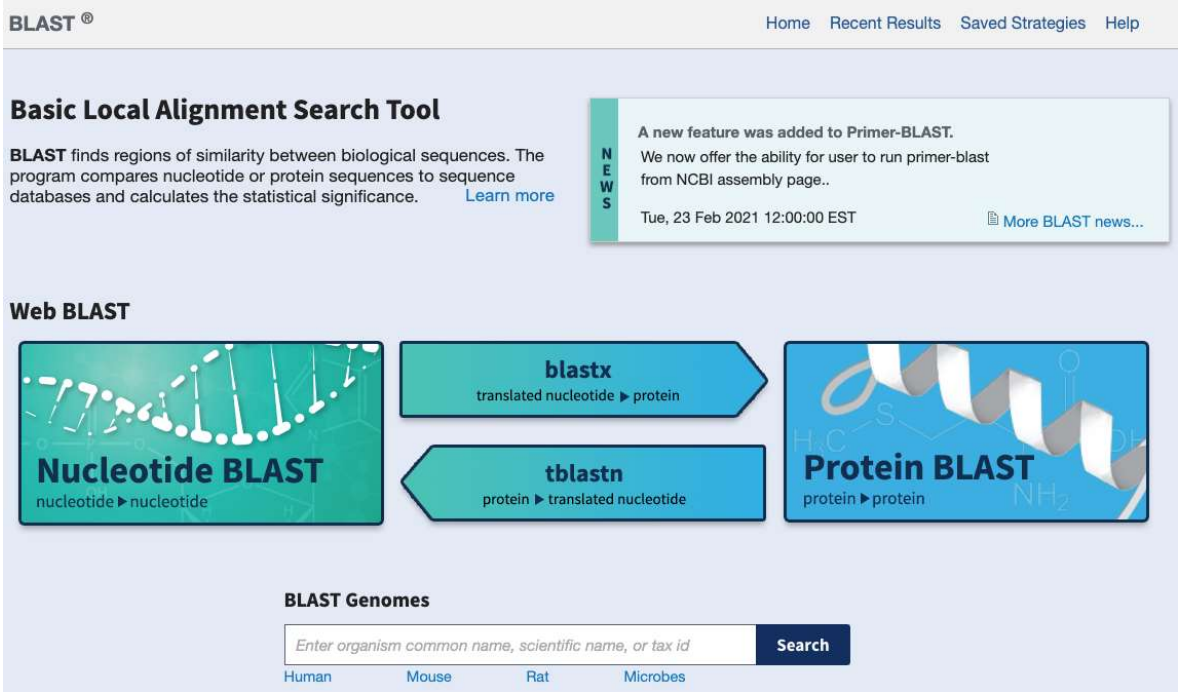
```
>Unknown_DNA_sequence
TTTTTATTCTTTCTGCCAGGTACATCAGATCCATCAGGTCCGAGCTGTGTTGACTACCACTGCTTTTCCC
TTCGTCTCAGTTATGTCTTGGAAGAAGGCTTTGCGGATCCCCGGAGACCTTCGGGTAGCAACTGTCACCT
```

Joonis 1. Näide FASTA formaadis nukleotiidsest järjestusest.

Pöörake tähelepanu, milliste bioloogiliste molekulidega on tegemist nii uuritava järjestuse kui andmebaasi puhul. Järgnevas tabelis (Tabel 1) on välja toodud järjestuste tüübid ja programmide nimed, mille vahel NCBI BLAST veebilehel valida (Joonis 2).

Tabel 1. BLAST programmid ja nendega seotud järjestuste tüübid.

Programm	Päringjärjestus	Andmebaas
blastn	nukleotiidid	nukleotiidid
blastp	aminohapped	aminohapped
blastx	nukleotiidid (transleeritakse 6 raami)	aminohapped
tblastn	aminohapped	nukleotiidid (transleeritakse 6 raami)
tblastx	nukleotiidid (transleeritakse 6 raami)	nukleotiidid (transleeritakse 6 raami)



Joonis 2. NCBI BLAST avaleht, kus sobiva programmi pildile klikkides pääseb selle otsingulehele.

NCBI BLAST otsingulehel on spetsiaalne väli, kuhu kopeerida vastav nukleotiidne või aminohapetega järjestus (Joonis 3).

Joonis 3. NCBI BLAST otsingulehel olev väli järjestuste jaoks.

Kui järjestus on sisestatud ja parameetrid seadistatud, tuleb programmi käivitamiseks vajutada BLAST nupule (Joonis 4).

Joonis 4. NCBI BLAST käivitub BLAST nupule vajutades.

Ülesanne 1. Teile on antud FASTA formaadis **DNA** järjestus, mille puhul tuleb kindlaks teha, millise liigi - **inimene** [*Homo sapiens (taxid:9606)*] või **hiir** [*Mus musculus (taxid:10090)*] - **geeni** selle pealt kodeeritakse. Joondage antud järjestus sobiva NCBI BLAST programmiga kõigi teadaolevate geenide (**nukleotiidid**) andmebaasi **"nr/nt"** vastu.
Vihje: Määrake NCBI BLAST otsingulehel **liigiks (*Organism*)** ühel juhul **inimene** ja teisel juhul **hiir** ladinakeelse tähistusega (nagu ülal kirjutatud).

a. Kummale liikidest annab BLAST esimeseks vasteks suurema **täieliku skooriga** (*Total score*) tulemuse? [5 punkti]

- ☐ Inimene
- ☐ ~~Hiir~~

BLAST joonduse skoori statistilist olulisust kirjeldavat arvu nimetatakse **E-väärtuseks** (*expected value* või *E value*). E-väärtus näitab, mitu juhuslikku antud skoori või veel kõrgema skooriga joondust võiksime kasutatud andmebaasist leida, kui andmebaas koosneks ainult juhuslikest järjestustest. Mida **väiksem E-väärtus**, seda **statistiliselt olulisem** on joondus (tulemus) ja **väiksem on tõenäosus**, et leiame andmebaasist juhuslikult mõne teise sarnase skooriga joonduse.

b. Valige vastuses a leitud suurima täieliku skooriga hiire **geeni identifikaator** (tulbas *Accession*) ja avage uus NCBI BLAST otsinguleht, mis võimaldab joondada **nukleotiidseid** järjestusi **aminohappeid** sisaldavate järjestusete vastu. Sisestage geeni identifikaator järjestuse aknasse ja valige andmebaasiks "**RefSeq Select proteins**". Tulemuste lehel filtreerige joondused **e-väärtuse** (*E value*) järgi sellises vahemikus: **0** kuni **1e-05**. Mitmele **mitte-imetaja** liigile BLAST vasted leidis? [5 punkti]
Vihjed: NCBI BLAST tulemuste lehel saab tulemusi filtreerida identsuse protsendi, E-väärtuse või päringjärjestuse katvuse järgi sisestades neisse soovi korral sobilikud vahemikud ja vajutades nuppu "**Filter**". Liikide kokkuvõtliku info saamiseks võib vajutada "**Taxonomy"** lingile.

Vastus: Peale kahe imetaja (inimene, hiir) leiti vasteid kahele bakteriliigile (*Nocardioides silvaticus*, *Tissierella creatinini*)

Ülesanne 2. Teile on antud FASTA formaadis **valgu** järjestus, millele on vaja leida asukoht inimese genoomis. Valige sobiv NCBI BLAST programm, mille abil oleks võimalik leida uuritava **valgu** järjestusele **"RefSeq Genome Database" DNA** andmebaasist parima skooriga joondus. Määrake NCBI BLAST otsingulehel liigiks **inimene.** Tulemuste lehel sorteerige vasted **päringjärjestuse katvuse** (*Query coverage*) järgi **kahanevalt.**

a. Millise inimese kromosoomi vastu andis uuritav valk suurima päringjärjestuse katvusega homoloogia? Kirjuta välja ka päringjärjestuse katvuse väärtus. [5 punkti]
Vastus: Inimese 11. kromosoom, katvus 88%

b. Kui pikk on andmebaasis eelnimetatud kromosoom nukleotiidides? [5 punkti]
Vastus: 135086622 nukleotiidi

c. Kas suurima täieliku skooriga tulemus vastuses a on ka suurima identsuse protsendiga (*Per. ident*)? [5 punkti]

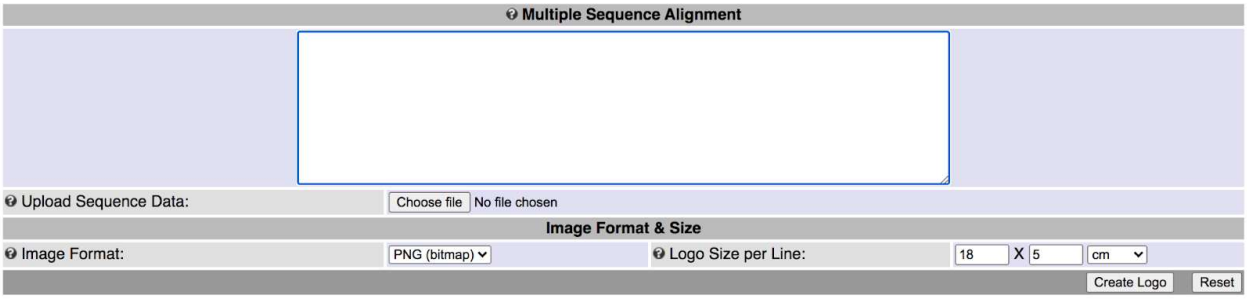
- ☐ ~~Jah~~
- ☐ Ei

d. Vajutage vastuses a **parima tulemuse** andnud järjestuse nime lingile, mis viib **joonduse vaatesse** (*Alignments*). Märkige ära, milline järgnevatest valikutest tähistab uuritava valgu parima tulemuse **algus- ja lõppkoordinaati inimese genoomis**? [5 punkti]

- ☐ ~~64926711 - 64926917~~
- ☐ ~~64926711 - 64933841~~
- ☐ 64925735 - 64933841
- ☐ ~~1 - 497~~
- ☐ ~~67 - 497~~

OSA II Bioloogiliste järjestuste motiivid

Ülesanne 1. Teile on antud DNA järjestuste regioonid, millele seondub **transkriptsioonifaktor nimega CEBPB.** Koostage nende põhjal **CEBPB seondumissaidi järjestuse logo** programmiga **WebLogo** (Joonis 5). Kopeerige järjestused ja sisestage tööriista aknasse täpselt sellisena, nagu need veebilehel on antud.



Joonis 5. Pärast järjestuste sisestamist vajutage logo arvutamiseks Create Logo nuppu.

a. Millised **nukleotiidid** ja **mitmendas positsioonis** peavad selle logo põhjal DNA järjestuses **alati olemas olema**, et see toimiks CEBPB seondumissaidina? [5 punkti]
Vastus: Seondumiseks nõutud on nukleotiidid T(3. pos), T(4.), C(6.), A(7.), A(10.).

b. Võttes aluseks loodud logo, otsustage, millisele või millistele allolevatest järjestustest **ei saa** CEBPB seonduda? [5 punkti]

- [x] ~~GGTTTGACAAT~~
- [ ] GATTGCACAAC
- [x] ~~CATTGCATAAC~~
- [x] ~~GGTTACATAAC~~
- [ ] GGTTGCATAAC

Ülesanne 2. Valgudomeenideks nimetatakse alamregioone valgu polüpeptiidahelas, millel on iseseisev teistest valgu osadest sõltumatu stabiilne struktuur ning sellest tulenevalt ka kindel spetsiifiline funktsioon. Tehes kindlaks kõik erinevad domeenid valgujärjestuses, annab see infot valgu võimaliku funktsiooni kohta organismis.

Kasutage vastuste leidmiseks programmi **HMMSCAN**, andmebaasiks (***HMM database***) valige ainult **Pfam**, tulemuste filtreerimise lävendiks (***Cut-Offs***) jätke "**Gathering**".
Pfam on valguperekondade andmebaas, kus on kirjeldatud kõik erinevad elusloodused leitud valgudomeenid neid iseloomustava HMM-mudeli abil. Hmmscan võrdleb meid huvitavat järjestust kõigi mudelitega andmebaasis ning leiab üles need mudelid, mis järjestuse erinevate osadega kõige paremini sobivad - suure tõenäosusega paiknevad uuritavas järjestuses seega ka vastavad domeenid. Hmmscan tulemustes on iga leitud domeeni jaoks antud esimeses tulbas vastava **domeeniperekonna ID Pfam andmebaasis**. 6. ja 7. tulbas on leitud domeeni algus (***Start***) ja lõpp (***End***) uuritavas järjestuses.

a. Mitut **erinevat** valgudomeeni sisaldab etteantud valgujärjestus? [5 punkti]
   Vastus: 5 erinevat domeeni

b. Millistel **koordinaatidel** paikneb selles järjestuses **kõige N-terminaalsem** domeen? [5 punkti]
   Vastus: Koordinaadid 268-416.

c. Millist **domeeni** esineb selles järjestuses **rohkem kui üks kord**? (Kirjutage domeeni nimi) [5 punkti]
   Vastus: SH2 domeen

Domeeni ID-le klikkides jõuate vastava domeeni lehele Pfam'i lehel, kus on lühikirjeldus selle domeeni funktsioonist ja olulisusest. Lisaks informatsioon selle kohta, milliste teiste valgudomeenidega ja millises asetuses vastav domeen erinevates järjestustes esineb (***Architectures***), mitu vastavat domeeni sisaldavat järjestust on Pfam andmebaasis ja millised need järjestused on (***Sequences***), millistes erinevates liikides domeen esineb (***Species***) ning kui palju on ennustatud 3D valgustruktuure, mille üheks osaks vastav domeen on (**Structures**). Erinevates liikides esinemist kirjeldab Pfam lehel nn. ***sunburst*** **graafik**, kus on toodud välja erinevate värvidega suuremad taksonoomilised grupid. Graafik avaneb klikkides domeeni lehel lingil *Species* (kas lehe vasakus servas või paremal ülaservas). Klikkides graafikul, saab lihtsamaks vaatluseks selekteerida ainult osa sellest, näiteks imetajad (Mammals) või primaadid (Primates). Liikudes hiire kursoriga graafikul, kuvatakse vastava taksonoomilise üksuse kohta info, mitu järjestust ning liiki on antud domeeniperekonnaga seotud.

d. Mitmes **erinevas liigis** on vastuses c nimetatud domeeni kirjeldatud? [5 punkti]
   Vastus: 480 liigis.

e. Mitmes **erinevas klass Putukad (*Insecta*) liigis** on vastavat domeeni valgujärjestustes kirjeldatud? [5 punkti]
Vastus: 69 liigis

f. Mitmes **inimese** valgujärjestuses seda domeeni leidub? [5 punkti]
Vastus: 220 järjestuses.

OSA III Nukleotiidsete järjestuste GC-sisaldus

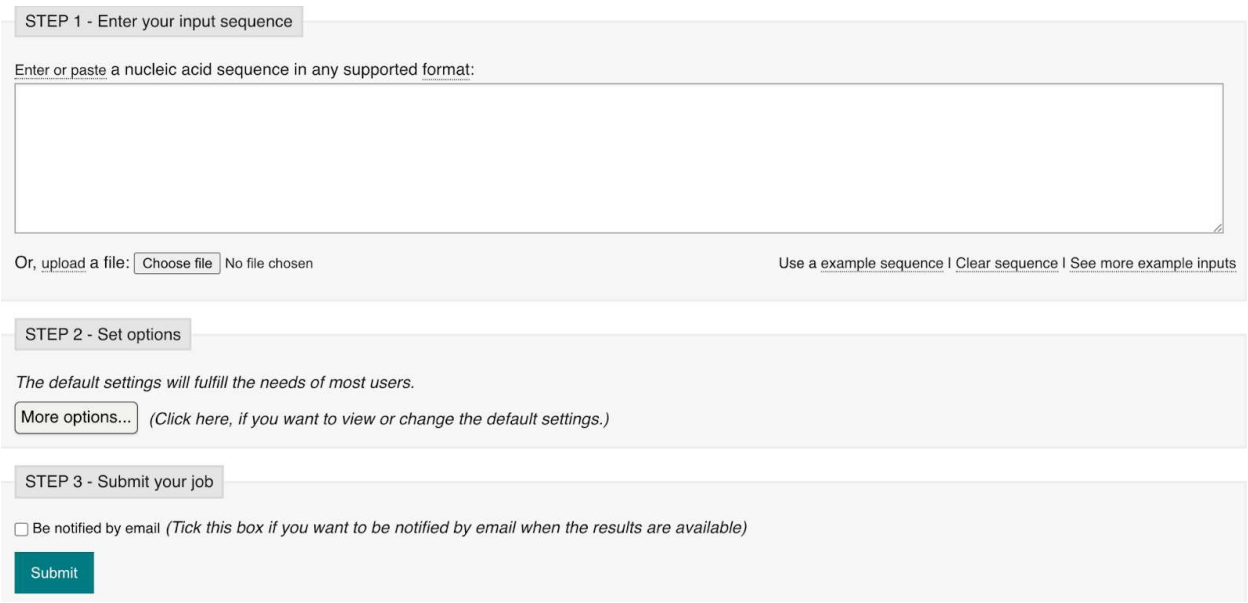
Ülesanne 1. Teile on antud nukleotiidne järjestus. Kasutage ülesande lahendamiseks tööriista https://www.novoprolabs.com/tools/gc-content **Akna suurus 10.**

a. Millisest positsioonist antud järjestuses algab **kõige madalama** GC-sisaldusega aken? [5 punkti]
Vastus: 22. positsioonist

b. Kui suur on sealt algavas aknas GC%? [5 punkti]
Vastus: 20%

c. Kui suur on selles järjestuses **maksimaalne GC%** 10 nukleotiidi pikkuses aknas? [5 punkti]
Vastus: 90%

Ülesanne 2. CpG saared on alad genoomis, kus esineb palju CG-dinukleotiide. (C nukleotiid, mille järel on G nukleotiid, “p” tähistab fosfodiester sidet kahe kõrvuti asuva nukleotiidi vahel) Tihtipeale asuvad need geenide avaldumise regulatsioonis oluliste promooteralade ümbruses ning on seetõttu abiks nii promooteralade kui ka geenide asukohtade ennustamisel. CpG saar on üldjuhul defineeritud kui vähemalt 200 bp pikkune regioon, mille GC% on >50% ning tegeliku ja oodatava CG dinukleotiidide arvu suhe on >60%.
Leidke CpG saared inimese genoomist pärit järjestuse lõigus. Kasutage selleks **Cpgplot** tööriista (Joonis 6) akna suurusega **200** (***window size***), saare minimaalne pikkus **200 nt** (***minimum length***), tegelik/oodatud CG **>0.6** (***minimum observed***) ning minimaalne GC-sisaldus **50%** (***minimum percentage***).



Joonis 6. Cpgplot avalehe selgitus. Klikkides “STEP 2 - Set options” all olevale “More options” nupule, avaneb valik, kus saab täpsemalt määrata otsingu parameetrid (akna suurus, minimaalne saare pikkus jne).

a. Mitu CpG saart selles järjestuses leidub? [5 punkti]
   Vastus: 7 saart

b. Millistel koordinaatidel asub **kõige kõrgema** GC-sisaldusega CpG saar selles järjestuses? [5 punkti]
   Vastus: 2266..2677

c. Millistel koordinaatidel asub ning kui pikk on **kõige pikem** CpG saar selles järjestuses? [5 punkti]
   Vastus: 6325..6991, pikkus 667 nt.