Vysoké učení technické
Fakulta informačních technologií

Hledání genů

5. cvičení

Alena Tesařová (xtesar36)

9. dubna 2020

1 Identifikace otevřeného čtecího rámce

Prostřednictvím nástroje ORF Finder vyhledejte nejdelší otevřený rámec (ORF) na genomové sekvenci bakteriofágu 3A ze souboru bacteriophage_3A.txt. Protein kódovaný daným ORF porovnejte prostřednictvím blastp s proteiny dostupnými v databázi nr.

1. *Určete nejdelší ORF (nejdelší ORF obvykle bývá ten správný).*
   Délka: 1584 nukleotidů.

```
   >lcl|ORF23
MKTFNKEQMIAQSHFGKLASQADVMSKKFSSIGDKMTSLGRTMTMGVSTP
ITLGLGAALKTSADFEGQMSRVGAIAQASSKDLKSMSNQAVDLGAKTSKS
ANEVAKGMEELAALGFNAKQTMEAMPGVISAAEASGAEMATTATVMASAI
NSFGLKGSDANHVADLLARSANDSAADIQYMGDALKYAGTPAKALGVSIE
DTSAAIEVLSNSGLEGSQAGTALRASFIRLANPSKSTAKEMKKLGIHLSD
AKGQFVGMGELIRQFQDNMKGMTREQKLATVATIVGTEAASGFLALIEAG
PDKINSYSKSLKNSNGESKKAADLMKDNLKGALEQLGGAFESLAIEVGKD
LTPMIRAGAEGLTKLVDGFTHLPGWFRKASVGLAIFGASIGPAVLAGGLL
IRAVGSAAKGYASLNRRIAENTILSNTNSKAMKSLGLQTLFLGSTTGKTS
KGFKGLAGAMLFNLKPINVLKNSAKLAILPFKLLKNGLGLAAKSLFAVSG
GARFAGVALKFLTGPIGATITCYYNCI
```

2. *Je sekvence genu odpovídající nejdelšímu ORF kompletní (odhadněte na základě analýzy blastp - lze spustit přímo z nástroje ORF Finder)?*
   Sekvenci odpovídá z 99% tento protein: **phage tail tape measure protein, partial [Staphylococcus aureus]** (skóre 1039)

2 Změna otevřeného čtecího rámce vlivem mutace - Single nucleotide polymorphism (SNP)

Mutace protein-kódující sekvence může změnit otevřený čtecí rámec (vznik / poškození na start / stop kodónu). Jedním z mnoha příkladů může být varianta hemoglobinu nazývaná Constant Spring. Tato varianta byla poprvé objevena na Jamaice a od standardní varianty se liší svoji délkou. Více podrobností ohledně této mutace můžete prostudovat v databázi OMIM pod identifikátorem 141850.

1. *Stáhněte z databáze GenBank standardní variantu nukleotidové sekvence proteinu HBA2 homo sapiens - mRNA (stahujte celý záznam ve formátu FASTA). Použijte nástroj ORF Finder ke zjištění délky ORF.*
   Délka: 429 nukleoditů
   Start: 67
   Stop: 495

```
>lcl|ORF1
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS
HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFK
LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
```

2. *Stáhněte nukleotidovou sekvenci varianty hemoglobinu Constant Spring. Použijte nástroj ORF Finder ke zjištění délky ORF.*
   Délka: 522 nukleoditů
   Start: 38
   Stop: 559

```
>lcl|ORF1
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS
HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFK
LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYRQAGASVAV
PPARWASQRALLPSLHRPFLVFE
```

3 Predikce genů založená na analýze sekvence a sekvenčních signálů

Sekvenční analýza může poskytnout relevantní informace využitelné pro predikci genů. Pro řešení následujících úloh využijte sadu nástrojů zvanou EMBOSS toolbox. Experimentování provádějte, není-li uvedeno jinak, na nukleotidové sekvenci proteinu HBA2 ze souboru protein_HBA2.fasta. Pro lehčí hledání odpovědí na níže uvedené otázky si přečtěte něco o methylaci DNA a CPG ostrůvcích.

1. ***CompSeq**: spočítejte frekvenci výskytu jednotlivých dinukleotidů v sekvenci. Má dinukleotid CG jinou než očekávanou frekvenci výskytu? Pokud ano, zdůvodněte proč.*
   Ano, má. Očekávaná frekvence je větší než pozorovaná jelikož čekávaná frekvence se vypočítá jako suma počtu nukleotidů C krát suma počtu nukleotidů G děleno délkou frekvence. Tím pádem se v sekvenci nachází více nukleotidů C a G samostatně (nebo je sekvence příliš velká).

Word	Obs Count	Obs Frequency	Exp Frequency	Obs/Exp Frequency
CG	502	0.0390783	0.0625000	0.6252530

2. *CpgPlot: Identifikujte oblasti CpG ostrůvků a vysvětlete, jak lze znalost o těchto oblastech využít pro hledání genů.*
   CpG ostrůvky jsou části sekvence, kde je vysoká frekvence párů GC. Pro hledání genů se dá využít tato znalost například k hledání promotoru, jelikož ten typicky obsahuje oblasti CpG ostrůvků.

   Listing 1: Oblasti CpG ostrůvků z `http://emboss.bioinformatics.nl/emboss-explorer/output/292568/`

```
Length 257 (65..321)
Length 770 (6113..6882)
Length 286 (6955..7240)
Length 770 (9924..10693)
Length 283 (10766..11048)
```

3. *Dreg: Identifikujte polyadeninové signály v sekvenci NG_000006 (stahujte celý záznam ve formátu FASTA). Nejčastějšími polyadeninovými signály jsou AATAAA a ATTAAA. Jak často se v sekvenci vyskytují?*
   AATAAA – výskyt 40x
   ATTAAA – výskyt 13x

4 Identifikace strukturních genů pomocí aplikace GeneMark

V části bakteriální sekvence Heliobacillus mobilis proveďte prostřednictvím aplikace GeneMark vyhledání strukturních genů. Používejte výchozí nastavení vstupního formuláře, ve kterém změňte druh na "Bacillus_subtilis_168"(položka "Select Species").

1. *Kolik ORF bylo detekováno na přímém vlákně?*
   75 (*total*) − 6 (*complement*) = 69 na přímém vlákně.

2. *Lokalizujte ribozomální vazebná místa (RBS). Za konsensuální model pro E.Coli je považována sekvence AAGGAG, která je umístěna typicky 4-12 nukleotidů před start kodónem. Tato RBS najděte pomocí utility Dreg z balíku EMBOSS.*
   Regulární výraz:

   ```
   (A|C|G)AGGA(A|G).{4,12}ATG
   ```

   Celkem nalezeno 10 odpovídajících výsledků. Z toho 8 jich je relevantních, jak lze vidět v seznamu níže.

Start	End	Sequence
582	597	GAGGAGGAGCATCATG
4463	4475	CAGGAAGGAGATG
6188	6204	AAGGAGTCACCGTAATG
6682	6700	GAGGAACAAAACCTCGATG
7126	7145	CAGGAGACTGAGTTGCAATG
8821	8837	GAGGAAGGATCACCATG
11474	11494	AAGGAAAACAACTGCTCGATG
12869	12886	AAGGAATGAGACGGTATG
15088	15107	GAGGAAATTATGTCTTTATG
15535	15552	GAGGAATGCAACTTTATG

5 Predikce operonů

Operony jsou sekvencí nukleotidů, resp. řadou po sobě jdoucích genů v bakteriálním chromozomu, které mají společný promotor a jsou regulovány společným operátorem a exprimovány najednou. Tyto geny kódují většinou enzymy zapojené v jedné metabolické dráze. Predikujte operony nad bakteriální sekvencí Heliobacillus mobilis pomocí 40bp pravidla: Pokud je intergenová vzdálenost dvojice nepřímo transkribovaných genů menší než 40 párů bází, potom je tato dvojice nazývaná operon.

1. *S využitím výstupu genové predikce GeneMarku z předchozí úlohy určete první operon na přímém vlákně.*
   Na základě pravidla *40bp* jsem určila, že první operon se bude nacházet na pozici **6189**, jelikož intergenová vzdálenost je $6202 - 6189 = 13$ a to je méně než 40.

Left end	Right end	DNA Strand	Coding Frame	Avg Prob	Start Prob
5242	6189	direct	fr 1	0.89	0.63
6202	7146	direct	fr 1	0.77	0.48

Listing 2: `http://exon.gatech.edu//tmp/gm.20200407.185017.30576.gm.out`