Hledání genů

Jakub Zárybnický (xzaryb00@stud.fit.vutbr.cz)

March 24, 2020

1 Identifikace otevřeného čtecího rámce

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

1. Určete nejdelší ORF (nejdelší ORF obvykle bývá ten správný).
   (a) 99% shoda s ORF001 Staphylococcus aureus
2. Je sekvence genu odpovídající nejdelšímu ORF kompletní (odhadněte na základě analýzy blastp - lze spustit přímo z nástroje ORF Finder)?
   (a) E = 0 pro alignment s $\text{NC}_{\text{007053.1}}$ (9430-10989)

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.
   (a) nt=429, aa=142

2. Stáhněte nukleotidovou sekvenci varianty hemoglobinu Constant Spring. Použijte nástroj ORF Finder ke zjištění délky ORF.

   (a) nt=522, aa=173

3 Predikce genů založená na analýze sekvence a sekvenčních signálů

Sekvenční analýza může poskytnout relevantní informace využitelné pro predikci genů. Pro řešení následujích úloh využijte sadu nástrojů zvanou EMBOSS toolbox. Experimentování provádějte, není-li uvedeno jinak, na nukleotidové sekvenci proteinu HBA2 ze souboru $\text{protein}_{\text{HBA2.fasta}}$. 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.

   (a) frekvence 3.91% je pouze 62.5% z očekávané 6.25%, CpG se (dle wiki) u obratlovců vyskytuje méně z důvodu možné degradace cytosinu na thymin.

2. CpgPlot: Identifikujte oblasti CpG ostrůvků a vysvětlete, jak lze znalost o těchto oblastech využít pro hledání genů.

   (a) V HBA2 se vyskytují tři regiony s CpG ostrůvky, na začátku, uprostřed a cca v 75%. Dle wiki jsou CpG ostrůvky většinou následované začátkem genu

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?

   (a) Sekvence AATAAA celkem 39x
   (b) Sekvence ATTAAA celkem 13x

4 Identifikace strukturních genů pomocí aplikace GeneMark

V části bakteriální sekvence Heliobacillus mobilis proveďte prostřednictvím aplikace GeneMark vyhledání strukturních genů. Používejte výchozí nas-

tavení vstupního formuláře, ve kterém změňte druh na "$\text{Bacillus}_{\text{subtilis168}}$" (položka "Select Species").

1. Kolik ORF bylo detekováno na přímém vlákně?

   (a) 15 ORF na přímém vlákně, 2 na komplementárním

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 sestavte tak, že:

   - na první pozici RBS může být A, C nebo G
   - na druhé až páté pozici RBS může být pouze sekvence AGGA
   - na šesté pozici RBS může být A nebo G
   - mezera mezi RBS a start kodónem může být 4-12 nukleotidů

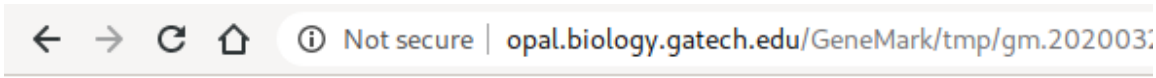
   Jak vypadá Vámi použitý regulární výraz? Kolik jste našli odpovídajících výskytů? Kolik z nich je relevantních (tj. nacházejících se v blízkosti ORF predikovaného pomocí GeneMark)?

   (a) [ACG]AGGA[AG].{4,12}ATG

   (b) 10 výskytů

   (c) pozice RBS na 582, 6188, 7126, 8821, 12869

(GeneMark na přiložené adrese `http://exon.biology.gatech.edu/` nefunguje, používám `http://opal.biology.gatech.edu/GeneMark/gm.cgi`)



GENEMARK PREDICTIONS

Sequence: Hm_dna.fasta
Sequence file: seq.fna
Sequence length: 16361
GC Content: 50.44%
Window length: 96
Window step: 12
Threshold value: 0.500

Matrix: Bacillus_subtilis_168
Matrix author: -
Matrix order: 4

List of Open reading frames predicted as CDSs, shown with alternate starts
(regions from start to stop codon w/ coding function >0.50)

Left end	Right end	DNA Strand	Coding Frame	Avg Prob	Start Prob
595	828	direct	fr 1	0.56	0.44
926	1888	direct	fr 2	0.80	0.47
971	1888	direct	fr 2	0.81	0.77
1139	1888	direct	fr 2	0.83	0.79
1145	1888	direct	fr 2	0.83	0.28
2081	3358	direct	fr 2	0.66	0.53
2213	3358	direct	fr 2	0.65	0.06
2258	3358	direct	fr 2	0.65	0.01
2354	3358	direct	fr 2	0.69	0.85
3381	4229	direct	fr 3	0.81	0.42
3390	4229	direct	fr 3	0.82	0.38
3471	4229	direct	fr 3	0.83	0.53
3717	4229	direct	fr 3	0.78	0.02
3723	4229	direct	fr 3	0.77	0.02
4268	4900	direct	fr 2	0.73	0.01
4286	4900	direct	fr 2	0.76	0.21
4334	4900	direct	fr 2	0.79	0.54
4379	4900	direct	fr 2	0.77	0.25
4385	4900	direct	fr 2	0.77	0.53
4882	6189	direct	fr 1	0.71	0.72
5086	6189	direct	fr 1	0.79	0.00
5161	6189	direct	fr 1	0.85	0.01
5164	6189	direct	fr 1	0.85	0.03
5209	6189	direct	fr 1	0.89	0.63
5212	6189	direct	fr 1	0.89	0.64

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.

   (a) Pokud uvažuju pouze s výstupen GeneMark, tak je to 2081 (následují geny na 3381 a 4268).