

# Hledání genů

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

March 19, 2020

## 1 Identifikace otevřeného čtecího rámce

Prostřednictvím nástroje ORF Finder vyhledejte nejdelší otevřený rámec (ORF) na genomové sekvenci bakteriofágu 3A ze souboru bacteriophage<sub>3A</sub>.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 NC<sub>007053.1</sub> (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 svojí 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 následujících úloh využijte sadu nástrojů zvanou EMBOSS toolbox. Experimentování provádějte, není-li uvedeno jinak, na nukleotidové sekvenci proteinu HBA2 ze souboru `proteinHBA2.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 thymín.
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<sub>000006</sub> (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 "Bacillus<sub>subtilis168</sub>" (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 najdete 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

## 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 výstupem GeneMark, tak je to 2081 (následují geny na 3381 a 4268).