Zarovnání sekvencí

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

March 24, 2020

Contents

1 Dotlet

Seznamte se s programem Dotlet. Zadejte programu různé (např. náhodné) vstupní sekvence (nukleotidů / aminokyselinami) a vyzkoušejte si vliv vstupních parametrů (jako jsou typ skórovací matice nebo velikost posuvného okýnka) na výsledek dotplot grafu.

1.1 Region s nízkou složitostí

Tento příklad ukazuje vliv regionu s nízkou složitostí na výsledek dotplot grafu.

1. Jako vstupní sekvenci použijte antigenu Plasmodium falciparum (parazit způsobující malárii) a zarovnejte ji vůči sobě samé.

2. Zvolte matici Blosum45 a velikost okýnka 15.

3. Spusťte výpočet a upravte si světlost výsledného grafu např. na -40 a +40 (posuvníky pod histogramem).

4. Ve výsledném grafu by jste měli spatřit tmavou oblast odpovídající sekvenci identických znaků. Přemístěním kurzoru do tmavé oblasti se vám objeví ve spodní části okna podrobnější výpis sekvence.

Otázky:

- Jak dlouhá a z jakých znaků se skládá uvedená oblast?
  - 191-225, znaky S

1.2 Opakování

Tento příklad ukazuje zajímavé vlastnosti, kterých si můžete všimnout při zarovnání sekvence oproti sobě samé.

1. Jako vstupní sekvenci použijte SLIT protein Drosophila melanogaster (vinná muška).

2. Zvolte matici Blosum62 a velikost okýnka 15.

3. Spusťte výpočet a upravte si světlost výsledného grafu např. na 0 a +90.

4. Ve výsledném grafu by jste měli spatřit dva shluky, jeden tvořený z delších a druhý z kratších opakujících se podřetězců (zkráceně opakování).

Otázky:

- Jak v dotplot grafu poznáte opakování?
  - Krátké (symetrické?) diagonální pruhy i mimo hlavní diagonálu
- Kolik jste napočítali delších a kolik kratších opakování?
  - Pokud počítám jen dolní trojúhelníkovou matici, tak 12 regionů s několikanásobným opakováním (krátké diagonální čáry)
  - Delších opakování bych z diagramu napočítal čtyři, ale tam záleží kritériích opakování.

2 Dynamické programování

2.1 Míra identity u náhodných biologických sekvencí

1. Vytvořte si 2 náhodné sekvence nukleotidů a aminokyselinami.

2. Dále budeme pracovat s online nástroji pro globální (Needle Nucl a Needle Prot) a lokální (LALIGN Nucl a LALIGN Prot) zarovnání nukleotidových/aminokyselinových sekvencí.

3. Pomocí nástrojů Needle resp. LALIGN vykonajte zarovnání vybraných nukleotidových/aminokyselinových sekvencí. Při zarovnávání aminokyselinových sekvencí zvolte matici BLOSUM50. Jinak ponechte původní nastavení. (LALIGN nabízí i pěkný grafický výstup v záložce *Visual Output*)

4. Sledujte hodnoty vyjadřující míru identity (Needle) a parametr E (pouze u LALIGN).

Otázky:

- Co vyjadřuje parametr E?
  - Očekávaný počet shod v databázi o N sekvencích (děleno N). Hodnota blížící se 0 odpovídá významné shodě, hodnota blížící se 1 znamená buď nevýznamnou (malou) shodu, nebo shodu krátké sekvence s databází dlouhých sekvencí, kde by byla běžná čistě pravděpodobnostně.
- Jaké míry identity a parametru E je obvykle dosahováno u náhodných nukleotidových sekvencí?
  - Identita 0.3 globálně
  - Identita 0.3, E 1 pro náhodné shody v podsekvencích
- Jaké míry identity a parametru E je obvykle dosahováno u náhodných sekvencí aminokyselin?
  - Identita 0.15 globálně
  - Identita 0.3, E 1 pro náhodné shody v podsekvencích

2.2 Objevení podobnosti mezi onkogeny

Russell Doolittle byl průkopníkem v oblasti algoritmů pro analýzu sekvencí v druhé polovině 70 a první polovině 80 let. Doolittle používal tehdejší databáze biologických sekvencí pro své experimenty s geny a hledání jejich funkce na základě podobnosti. V následujícím cvičení si zopakujeme kroky, které Doolittle provedl při objevu funkce **v-mos** onkogenu viru *Moloney Murine Sarcoma*. Ne dlouho poté, co byl nasekvenován v-mos v Salk Institutu, studovala skupina vědců vztah mezi **v-src** onkogen viru *Rous Sarcoma* a v-mos onkogenu. První pokusy o hledání podobnosti však dopadly neúspěšně.

1. Pro tenhle úkol budeme potřebovat nástroj Sixpack pro překlad nukleotidové sekvence na aminokyselinovou.

2. S použitím nástrojů Needle resp. LALIGN pro globální resp. lokální zarovnání analyzujte podobnost obou nukleotidových sekvencí vmos.fasta a src.fasta.

3. Na základě míry identity a parametru E zhodnoťte výsledky zarovnání.

4. Dále použijte nástroj Sixpack který dokáže přeložit sekvenci nukleotidů na aminokyseliny pro překlad sekvencí *v-mos* a *v-src*. Ponechte původní nastavení. Všimněte si, že nástroj Sixpack provede celkem 6 překladů s ohledem na různou počáteční pozici překladu fram-u. Tři překlady provede s přímou sekvencí a tři překlady provede s komplementární a reverzovanou sekvencí (simulace komplementárního vlákna DNA).

5. Z šesti překladů obou sekvencí vyberte takové, které by mohli vést na co největší míru podobnosti (říďte se tabulkou na konci výstupu, výsledné přeložené sekvence najdete v záložce *Tool Output*).

6. Proveďte globální a lokální zarovnání přeložených sekvencí pomocí nástrojů Needle/LALIGN.

7. Na základě míry identity a parametru E zhodnoťte výsledky zarovnání.

8. Pokud si nejste jisti, zda jste vybrali správné fram-y v bodě 5, vyberte jiné a opakujte experiment. Nápověda: hledáme takové posuny, které vedou na překlad jednoho genu v protein (jeden start/stop kodon)

Otázky:

- Jaké fram-y jste vybrali?

  - Frame 1 v obou případech - jediné souvislé fram-y bez více ORF

- Jak hodnotíte výsledky zarovnání nukleotidových sekvencí?
  - Nepříliš přesvědčivé
- Nalezli jste lepší zarovnání v případě přeložených sekvencí?
  - Lokální zarovnání o délce 582 s $E(1) = 4.6e - 07$
  - Globální zarovnání se skórem 241

3 BLAST

3.1 Hledání původu DinoDNA

1. Film Michael Crichtona o klonování dinosaurů, Jurský park, ukazuje domnělou DNA sekvenci dinosaura. Identifikujte skutečný zdroj této DNA sekvence s využitím programu BLAST a NCBI databáze všech nukleotidů **nr**.
2. Vědec NCBI Mark Boguski však upozornil na to, že jeho sekvence byla určitě kontaminovaná a zásobil Michaela Crichtona lepší sekvencí, pro pokračování tohoto filmu z názvem The Lost World. Identifikujte zdroj této sekvence.

Otázky:

- Nalezl Mark lepší sekvenci než Michael? Proč?
  1. Michael - 99% se shoduje s "Escherichia coli strain Mach1 plasmid pSS1129, complete sequence" s $E(1) = 3e - 117$
  2. Mark - 66% se shoduje s "Gallus gallus GATA binding protein 1" s $E(1) = 0$ (úplná shoda)
  3. Na takovou otázku se špatně odpovídá - ani jedna není dobrá pro klonování dinosaurů. Michaelova sekvence je DNA bakterie E. coli a tudíž naprosto irelevantní, Markova sekvence je asi relevantnější, obsahuje jeden konkrétní gen, který se mohl vyskytovat i v DNA dinosaura, pokud uvažujeme příbuznost *Gallus gallus*, ale ani jedna z nich není sekvence, která by pomohla rekonstruovat DNA dinosaura. Musím tedy asi odpovědět, že Markova sekvence je lepší, GATA1 protein, ač ve verzi Gallus gallus, je v přítomnosti dalších fragmentů DNA užitečnější.

- Mark zabudoval do své sekvence také své jméno MARK. Nalezněte toto jméno v sekvenci.
  - Je součástí "DinoDNA$_{1ORF1}$ Translation of DinoDNA in frame 1, ORF 1, threshold 1, 358aa", polovina třetího řádku:

```
>Translation of DinoDNA in frame 1, ORF 1, threshold 1, 358aa
EFRKRARDKSWHQIQLEIRTDVWQLPQRIHWKCITYPMGAMEFVALGGPDAGSPTPFPDE
AGAFLGLGGGERTEAGGLLASYPPSGRVSLVPWADTGTLGTPQWVPPATQMEPPHYLELL
QPPRGSPPHPSSGPLLPLSSGPPPCEARECVMARKNCGATATPLWRRDGTGHYLCNWASA
CGLYHRLNGQNRPLIRPKKRLLVSKRAGTVCSHERENCQTSTTTLWRRSPMGDPVCNNIH
ACGLYYKLHQVNRPLTMRKDGIQTRNRKVSSKGKKRRPPGGGNPSATAGGGAPMGGGGDP
SMPPPPPPPAAAPPQSDALYALGPVVLSGHFLPFGNSGGFFGGGAGGYTAPPGLSPQI
```

3.2 Hledání komplementárních sekvencí

1. S využitím databáze NCBI GenBank si stáhněte sekvenci nukleotidů libovolného lidského genu napr. KRAS (postačí prvních 1000 znaků genu)
2. S využitím následujícího webového nástroje si ke vstupnímu genu vytvořte:
   - reverzní sekvenci,
   - komplementární sekvenci,
   - reverzní+komplementární sekvenci.
3. S využitím BLASTu vyhledejte v **nr** databázi všechny výskyty vytvořených sekvencí

Otázky:

- Shodují se výsledky pro všechny alternativy vstupní sekvence? Zdůvodněte proč.
  - Výsledky BLAST jsou shodné pro původní a reverse-complement DNA (GTPase), stejně tak pro reverse a complement DNA (OATP-B promotor)
  - DNA čteme v pořadí 5-3 a reverse nebo complement verze jsou obě v 3-5 pořadí, BLAST tedy zřejmě hledá i normální i reverse-complement verze, aspoň pro blastn.