Vysoké učení technické
Fakulta informačních technologií

Zarovnání sekvencí

2. cvičení

Alena Tesařová (xtesar36)

21. března 2020

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.

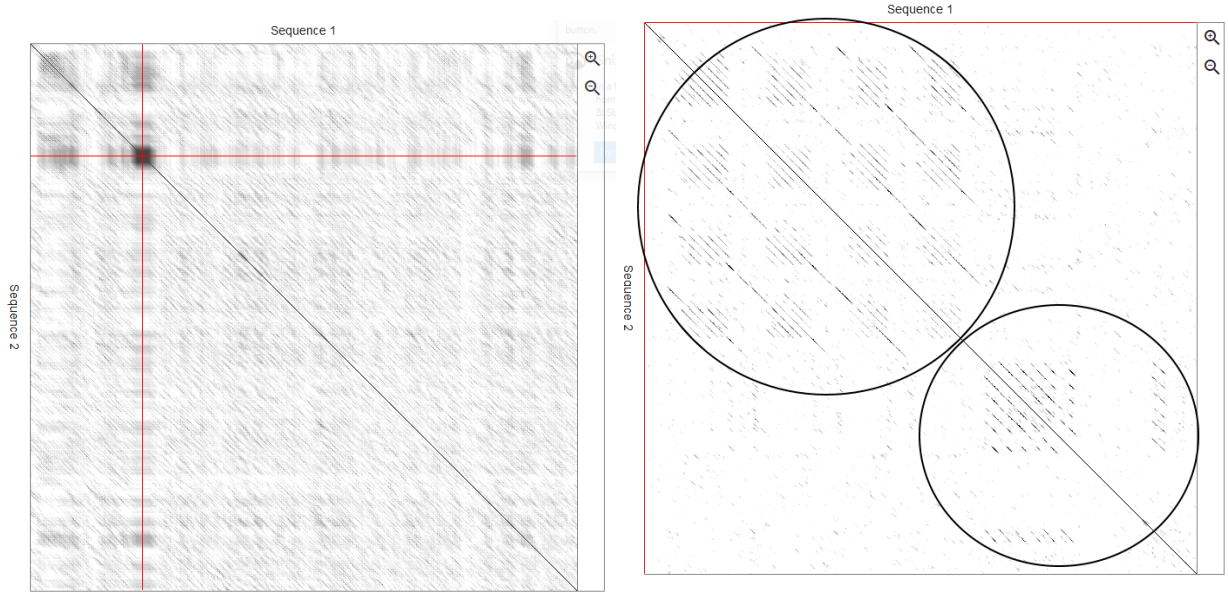
1.1 Region s nízkou složitostí

- *Jak dlouhá a z jakých znaků se skládá uvedená oblast?*
  Uvedená oblast se skládá z 35 znaků S.

1.2 Opakování

- *Jak v dotplot grafu poznáte opakování?*
  Opakování poznáme ze čtverců v grafu, kde každý řádek představuje opakování jedné části kódu na více místech v druhé sekvenci.

- *Kolik jste napočítali delších a kolik kratších opakování?*
  Delších opakování: 4
  Kratších opakování: 6 (viz obrázek 1b)



(a) Plasmodium falciparum. V grafu lze vidět tmavou oblast odpovídající sekvenci identických znaků.

(b) Drosophila melanogaster. Zvýrazněny dva shluky, jeden je tvořený z delších opakujících se podřetězců a druhý z kratších

Obrázek 1: Zobrazení grafů v programu Dotlet

2 Dynamické programování

2.1 Míra identity u náhodných biologických sekvencí

- *Porovnání dvou náhodných sekvencí nukleotidů*
  Identita: 36.0 %
  E(1) $<$ 0.81

- *Porovnání dvou náhodných sekvencí aminokyselin*
  Identita: 24.1 %
  E(1) $<$ 0.25

- *Co vyjadřuje parametr E?*
  Očekávaný počet výskytů zarovnání sekvence délky m v náhodné sekvenci délky N se skórem $\geq$ S

- *Jaké míry identity a parametru E je obvykle dosahováno u náhodných nukleotidových sekvencí?*
  $E = 0.81 - 0.9$ (nejčastější výsledek)
  25 – 30 % identita

- *Jaké míry identity a parametru E je obvykle dosahováno u náhodných sekvencí aminokyselin?*
  $E = 0.5 - 0.9$ (nejčastější výsledek)
  5 – 15 % identita

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.

- *Pro tenhle úkol budeme potřebovat nástroj Sixpack pro překlad nukleotidové sekvence na aminokyselinovou. S použitím nástrojů Needle resp. LALIGN pro globální resp. lokální zarovnání analyzujte podobnost obou nukleotidových sekvencí vmos.fasta a src.fasta. Na základě míry identity a parametru E zhodnoťte výsledky zarovnání.*
  Míra identity: 40.8 %
  Parametr E: 0.016
  Vzhledem k tomu, že parametr E je větší než 10^{-6} a identita je nižší než 70 %, nejedená se o statisticky významný výsledek.

- *Jaké fram-y jste vybrali?*
  Vybrala jsem s fram-y s nejnižší hodnotou čtecího rámce (ORFs) tedy první frame v obou případech.

- *Jak hodnotíte výsledky zarovnání nukleotidových sekvencí? Nalezli jste lepší zarovnání v případě přeložených sekvencí?*
  Identita dosáhla hodnoty 28.6%, takže je menší než v případě nukleotidových sekvencí. Nicméně parametr $E(1) = 3.2e^{-20}$, což je výsledek výrazně nižší než u nukleotidů. Zarovnání bych hodnotila celkově jako lepší.

2.3 Hledání původu DinoDNA

- *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.*
  SOURCE: Expression vector pCJH5_pDEST17-*

- *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.*
  Gallus gallus (Kuře)

- *Nalezl Mark lepší sekvenci než Michael? Proč?*
  Nalezl lepší sekvenci, jelikož můžeme považovat, že je kuře blíže k dinosaurovi než nějaký expresní vektor.

- *Mark zabudoval do své sekvence také své jméno MARK. Nalezněte toto jméno v sekvenci.*

```
EFRKRARDKSWHQIQLEIRTDVWQLPQRIHWKCITYPMGAMEFVALGGPDAGSPTPFPDE
AGAFLGLGGGERTEAGGLLASYPPSGRVSLVPWADTGTLGTPQWVPPATQMEPPHYLELL
QPPRGSPPHPSSGPLLPLSSGPPPCEARECVMARKNCGATATPLWRRDGTGHYLCNWASA
CGLYHRLNGQNRPLIRPKKRLLVSKRAGTVCSHERENCQTSTTTLWRRSPMGDPVCNNIH
ACGLYYKLHQVNRPLTMRKDGIQTRNRKVSSKGKKRRPPGGGNPSATAGGGAPMGGGGDP
SMPPPPPPPAAAPPQSDALYALGPVVLSGHFLPFGNSGGFFGGGAGGYTAPPGLSPQI
```

Listing 1: Translace nukleotidové sekvence DinoDNA

2.4 Hledání komplementárních sekvencí

S využitím databáze NCBI GenBank si stáhněte sekvenci nukleotidů libovolného lidského genu napr. KRAS (postačí prvních 1000 znaků genu, Listing 2). S využitím následujícího webového nástroje si ke vstupnímu genu vytvořte

1. reverzní sekvenci (Listing 3),
2. komplementární sekvenci (Listing 4),
3. reverzní+komplementární sekvenci (Listing 5).

Listing 2: Původní sekvence

```
CTAGGCGGCGGCCGCGGCGGCGGAGGCAGCAGCGGCGGCGGCAGTGGCGGCGGCGAAGGTGGCGGCGGCT
CGGCCAGTACTCCCGGCCCCCGCCATTTCGGACTGGGAGCGAGCGCGGCGCAGGCACTGAAGGCGGCGGC
GGGGCCAGAGGCTCAGCGGCTCCCAGGTGCGGGAGAGAGGTACGGAGCGGACCACCCCTCCTGGGCCCCT
GCCCGGGTCCCGACCCTCTTTGCCGGCGCCGGGCGGGGCCGGCGGCGAGTGAATGAATTAGGGGTCCCCG
```

Listing 3: Reverzní sekvence

```
GCCCCTGGGGATTAAGTAAGTGAGCGGCGGCCGGGGCGGGCCGCGGCCGTTTCTCCCAGCCCTGGGCCCG
TCCCCGGGTCCTCCCCACCAGGCGAGGCATGGAGAGAGGGCGTGGACCCTCGGCGACTCGGAGACCGGGG
CGGCGGCGGAAGTCACGGACGCGGCGCGAGCGAGGGTCAGGCTTTACCGCCCCCGGCCCTCATGACCGGC
TCGGCGGCGGTGGAAGCGGCGGCGGTGACGGCGGCGGCGACGACGGAGGCGGCGGCGCCGGCGGCGGATC
```

Listing 4: Komplementární sekvence

```
GATCCGCCGCCGGCGCCGCCGCCTCCGTCGTCGCCGCCGCCGTCACCGCCGCCGCTTCCACCGCCGCCGA
GCCGGTCATGAGGGCCGGGGGCGGTAAAGCCTGACCCTCGCTCGCGCCGCGTCCGTGACTTCCGCCGCCG
```

```
CCCCGGTCTCCGAGTCGCCGAGGGTCCACGCCCTCTCTCCATGCCTCGCCTGGTGGGGAGGACCCGGGGA
CGGGCCCAGGGCTGGGAGAAACGGCCGCGGCCCGCCCCGGCCGCCGCTCACTTACTTAATCCCCAGGGGC
```

Listing 5: Reverzní komplementární sekvence

```
CGGGGACCCCTAATTCATTCACTCGCCGCCGGCCCCGCCCGGCGCCGGCAAAGAGGGTCGGGACCCGGGC
AGGGGCCCAGGAGGGGTGGTCCGCTCCGTACCTCTCTCCCGCACCTGGGAGCCGCTGAGCCTCTGGCCCC
GCCGCCGCCTTCAGTGCCTGCGCCGCGCTCGCTCCCAGTCCGAAATGGCGGGGGCCGGGAGTACTGGCCG
AGCCGCCGCCACCTTCGCCGCCGCCACTGCCGCCGCCGCTGCTGCCTCCGCCGCCGCGGCCGCCGCCTAG
```

- *Shodují se výsledky pro všechny alternativy vstupní sekvence? Zdůvodněte proč.*
  Výsledek byl nalezen pouze u reverzní komplementární sekvence, jelikož se jedná o ten samý úsek DNA. U reverzní a komplementární sekvence nebylo nic nalezeno.