LS 2020 Jméno:

BIN – Bioinformatika – 10.9.2020

O1	O2	O3	O4	O5	Zkouška (50)	Cvičení (50)	Celkem (100)

Instrukce: Na vypracování máte 120 min. K odpovědi využijte volná místa pod otázkami, poté strany 1, poté i volné listy. Odpovídejte co nejprodrobněji, ale stručně a strukturovaně.

Otázka 1 *(10 bodů) Zarovnávání sekvencí.*

O zarovnávání sekvencí metodou BLAST lze prohlásit (v každé podotázce je vždy právě jedna odpověď správná, kromě zaškrtnutí odpovědi slovně zdůvodněte svoji volbu).

(a) (2 body) Z pohledu uživatele je po zadání vstupní sekvence q časově nejnáročnějším krokem metody:

i) najít s pomocí substituční matice v q všechna slova délky w (obvykle $w = 3$), která překračují prahovou hodnotu skóre slova T ((obvykle $T = 9$)),

ii) detekovat v databázi všech relevantních sekvencí D (např. UniProtKB) tzv. hity, tj. najít všechny výskyty výše uvedených slov délky w v D,

iii) rozšířit nalezené hity v obou směrech a rozhodnout, zda jde o možná zarovnání (nejprve bez mezer, hledáme segmenty překračující práh S, posléze pro nadějné kandidáty včetně mezer),

iv) rozhodnout o statistické signifikanci reportovaných zarovnání, tj. spočítat pro všechna zarovnání jejich E-hodnoty (náhrada častějších p-values).

(b) (2 body) Kromě rychlosti je důležitým parametrem metody BLAST její citlivost (sensitivita). Citlivost lze ovlivnit volbou parametru T. Definujeme ji jako:

i) poměr mezi počtem nalezených zarovnání a počtem skutečně existujících zarovnání, jak by je v D nalezlo například dynamické programovaní,

ii) procento nalezených zarovnání, které odpovídá skutečné evoluční homologii,

iii) procento hitů, které se následně podaří rozšířit (tj. procento slov délky w s nadprahovou hodnotou skóre slova, která následně vytvoří segment překračující prahovou hodnotu S),

iv) průměrný absolutní počet nalezených zarovnání s podprahovou E-hodnotou, E-hodnotu obvykle volíme 0.05 a průměrujeme přes celý genom daného organismu, citlivost se váže nejenm k metodě, ale i konkrétnímu genomu.

(c) (2 body) Jednotlivá zarovnání hodnotíme pomocí E-hodnoty. O ní lze prohlásit:

i) E-hodnota vyjadřuje průměrný počet zarovnání, které bychom v databázi D našli beze změny parametrů metody pro náhodně vygenerovanou sekvenci stejné délky jako má q a měla by skóre nejméně takové jako je skóre právě hodnoceného zarovnání,

ii) E-hodnota vyjadřuje pravděpodobnost, že dané zarovnání je náhodné a nalezená sekvence není homologická k sekvenci q,

iii) rozhodnout o statistické signifikanci reportovaných zarovnání, tj. spočítat pro všechna zarovnání jejich E-hodnoty (náhrada častějších p-values).

iv) pokud E-hodnota zarovnání roste, pak jeho p-hodnota nutně klesá,

v) E-hodnota zarovnání nemá s jeho p-hodnotou žádný vztah.

(d) (2 body) Součástí metody BLAST je algoritmus:

i) hledání nejkratší cesty v grafu bez záporných cyklů pro urychlení zarovnání,

ii) metoda větví a mezí pro prořezání prohledávacího prostoru možných zarovnání,

iii) aloritmus tvorby fylogenetického stromu sloužící ke shlukování sekvencí v databázi D, shluky se využijí v pre-indexaci sekvencí,

iv) dynamického programování pro zarovnání páru sekvencí, použije se jen pro vybrané kandidáty zarovnání.

(e) (2 body) Pro metodu BLAST **neplatí** tvrzení:

i) jde o heuristickou metodu,

ii) má lineární asymptotickou složitost, díky indexaci hraje roli pouze délka vstupní sekvence q,

iii) její variantu lze použít i pro zarovnávání DNA sekvence vůči proteinové databázi,

iv) může být použita i k vyhledání evolučně vzdálených párů homologických sekvencí.

Otázka 2 *(10 bodů) Sestavování sekvencí.*

Analyzujete sekvenci "i_má_máma_má_mámu" pomocí de Bruijnových grafů.

(a) (3 body) Nejprve sekvenci modelově rozdělte na 5-mery tak, aby byly splněny všechny předpoklady pro znovusestavení sekvence. Diskutujte slovně význam těchto předpokladů.

(b) (2 body) Z jakých pojmů teorie grafů de Bruijnovy grafy vychází? Jak algoritmus sestavení sekvence pracuje? Jaká je jeho asymptotická složitost a proč?

(c) (2 body) Demonstrujte korektní sestavení původní sekvence z rozkladu z bodu ad a.

(d) (3 body) Porušte postupně tři z předpokladů metody a na příkladech vysvětlete, jaká budou mít tato porušení důsledky pro sestavování sekvence.

Otázka 3 *(10 bodů) Markovské modely biologických sekvencí*

Máte k dispozici množinu sekvencí {ACGGAGA, CGTTGACA, ACTGAA, CCGTTCAC}. Vaším úkolem je vytvořit profilový skrytý markovský model (HMM) pro tuto množinu.

(a) (2 body) Vysvětlete k čemu profilový HMM slouží. Zapište matematicky, demonstrujte na příkladě. Uveďte alespoň 2 způsoby jeho použití.

(b) (1 bod) Formálně definujte úlohu učení profilového HMM (jaké rozdělení se učíme, jaké je kritérium).

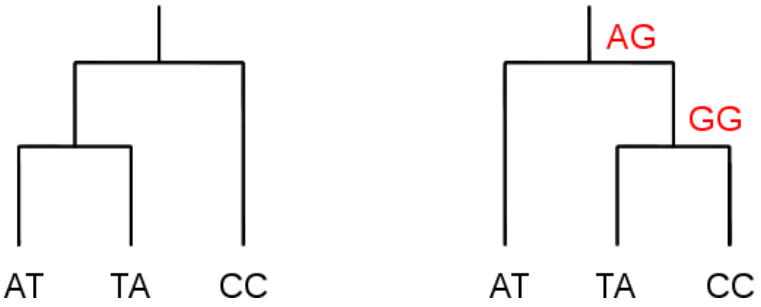
(c) (2 body) Podrobně srovnejte jak by vypadalo učení profilového HMM v situaci, kdy prvním krokem je a není zarovnání vstupních sekvencí. Pojmenujte výhody a nevýhody obou postupů. Který postup je obvyklejší?

(d) (3 body) Stručně popište metodu progresivního stromového zarovnání množiny sekvencí (algoritmus CLUSTALW). Naznačte jeho aplikaci na uvedenou čtveřici sekvencí. Uvažujte triviální skóre, kdy za shodu v páru symbolů počítáme +1, za neshodu/vložení mezery -1. Popište princip, zarovnání neprovádějte.

(e) (2 body) Nakreslete profilový HMM jehož struktura odpovídá zarovnání vzniklému v minulém kroku. To nechť je {ACG–GAGA, -CGTTGACA, AC-T-GA-A, CCGTTCAC-}.

Otázka 4 *(10 bodů) Fylogenetické stromy*

Na základě parsimony metody rozhodněte, zda je pro danou trojici sekvencí vhodnější fylogenetický strom vlevo nebo vpravo. Součástí řešení je nalezení optimálních sekvencí pro vnitřní uzly stromu vlevo, u stromu vpravo již byly tyto sekvence nalezeny. Předpokládejte, že sekvence jsou zarovnány a uvažujte nezávislost mezi rezidui, tj. pozicemi v sekvencích. Záměny mezi nukleotidy hodnoťte dle cenové matice níže.



	A	C	G	T
A	0	0.8	0.2	0.9
C	0.8	0	0.7	0.5
G	0.2	0.7	0	0.1
T	0.9	0.5	0.1	0

(a) (4 body) Použití váženého parsimony algoritmu.

(b) (4 body) Doplnění sekvencí do vnitřních uzlů stromu vlevo.

(c) (2 body) Ohodnocení obou stromů a rozhodnutí o tom, který je lepší.

Otázka 5 *(10 bodů) Sekundární struktura RNA*

Máte za úkol modelovat sekundární strukturu ribonukleové kyseliny. Diskutujte následující otázky.

(a) (2 body) Porovnejte stavbu řetězce RNA a DNA. Pojmenujte rozdíly a zkuste vysvětlit z čeho plynou.

(b) (2 body) Co to je sekundární struktura RNA? Proč je u RNA důležitá a k čemu se používá? Jaké jiné úrovně popisu RNA znáte?

(c) (2 body) Jakým typem gramatiky lze sekundární strukturu RNA popsat a za jakých podmínek? Vysvětlete, uveďte příklad gramatiky.

(d) (2 body) Popište jak se využívá minimalizace volné energie k predikci sekundární struktury RNA.

(e) (2 body) Postup popsaný výše srovnejte s Nussinovovým algoritmem založeným na dynamickém programování (předpoklady, složitost, úspěšnost).