BIOINFORMATIKA – ZADÁNÍ 2. SAMOSTATNÉ PRÁCE

Skládání sekvencí, zarovnávání sekvencí

Podmínky

Pro získání plného počtu bodů (tj. 20) je potřeba si vybrat jeden z dále uvedených úkolů a vypracovat jej

Úkoly budete odevzdávat na 11. cvičení, přičemž pro pozdní odevzdání budou platit podmínky na webu Marka Cravena (tady) - za každý den navíc se vám bude odečítat 1 bod.

Varianta 1: Skládání sekvencí

1. Máme dány následující 3-mery:

{AGT, AAA, ACT, AAC, CTT, GTA, TTT, TAA}

Zkonstruujte graf překryvů („overlap graph“) a nalezněte hamiltonovskou cestu (obsahující 7 hran). Zapište nadsekvenci odpovídající této hamiltonovské cestě.

2. Máme zadáno následující spektrum:

S = {ATG, GGG, GGT, GTA, GTG, TAT, TGG}

Ukažte, jak lze pomocí metody založené na eulerovských tazích nalézt sekvenci **s** takovou, že ***Spectrum(s,3) = S*** (**nepovinně:** jak lze nalézt **všechny** takové sekvence?).

3. Vyberte si jeden z postupů uvedených výše a zautomatizujte jej (tj. napište program pro řešení úloh daného typu). Můžete použít existující kód pro hledání hamiltonovských cest, příp. Eulerovských tahů.

Varianta 2: Zarovnávání sekvencí

- Napište program pro výpočet lokálního (nebo globálního – můžete si vybrat) zarovnání dvou sekvencí s lineární penaltou pro mezery (**ambiciózněji varianta: použijte afinní penaltu pro mezery – pozor, algoritmus pro afinní variantu udržuje tři matice namísto jedné**).
- Vstupem programu by měly být dva soubory se sekvencemi, které chceme zarovnat, a matice penalt substitucí
- Aplikujte váš program na libovolné sekvence a výsledek porovnejte s výsledkem, který dostanete např. pomocí toolkitu BioJava nebo jiného nástroje (**Pozor: nezapomeňte, že může existovat více než jedno optimální zarovnání, tedy nepanikařte, pokud dostanete jiné řešení, než vám dá BioJava**).