ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FIT

Přibližné vyhledávání nad vlastními indexy

Vedoucí: Jan Holub

Lukáš Hrbek

Problém:
- Přibližné vyhledávání s nejvýše k chybami
- Chyby definované pomocí Hammingovy, či Levenshteinovy vzdálenosti
- Zobrazení všech výskytů a jejich zarovnání
- Vyhledávání v komprimovaných datech

Přibližné vyhledávání:
Dovoluje vyhledat také výskyty, které se od hledaného vzorku mírně liší. Rozdíly mohou být definovány takto:

- Replace CAG → CTG
- Insert CAG → CATG
- Delete CAG → CG

Řešení:
- Zobecněný fitrační algoritmus založený na pigeonhole principu
- Matematický model pro vhodné nastavení algoritmu
- Sublineární složitost vyhledávání při vhodném nastavení
- Použit index FM-Index verze 2

Vlastní index:
Indexování původního textu umožňuje rychleji provádět vyhledávání (např. složitost nebude záviset na délce prohledávaného textu).
Vlastní index navíc umožňuje rekonstrukci původního textu - ten proto nemusí být uložen. Protože indexový soubor je menší, dochází ke kompresi původního textu.

Motivace:
Přibližné vyhledávání ve velkých textech je využíváno při výzkumu DNA, např. pro hledání nových léků.

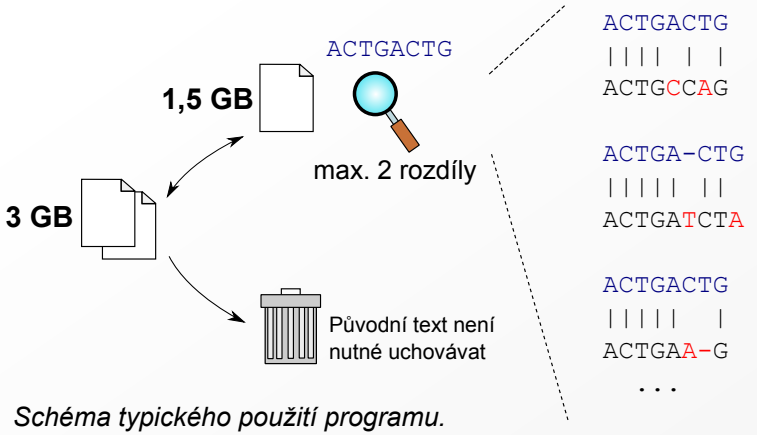
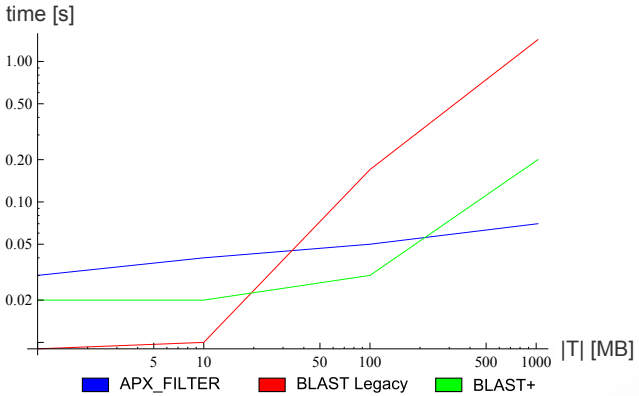
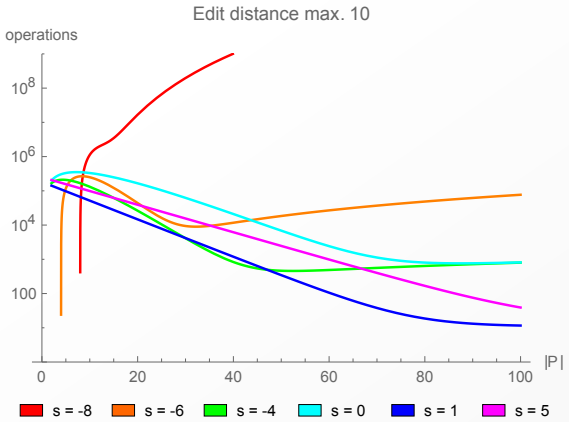


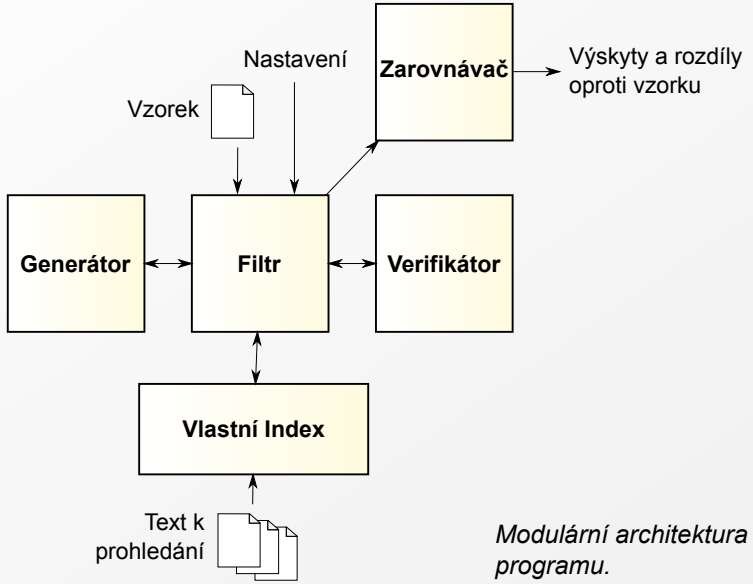
Schéma typického použití programu.



Logaritmický graf vlivu velikosti původního textu (|T|) na čas vyhledávání.



Analytický model odhadující dobu výpočtu v závislosti na délce vzorku (|P|) pro různé nastavení filtračního algoritmu.



Modulární architektura programu.

Srovnání:
Tyto problémy řeší BLAST. Jedná se však jen o heuristickou metodu specializovanou pro DNA.
Nové řešení je pro velké soubory rychlejší než BLAST a pro menší soubory využívá významně méně paměti.