

IV107 Bioinformatika I

Přednáška 10

Katedra informačních technologií
Masarykova Univerzita Brno

Jaro 2009

Řetězce a algoritmy na
řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu
hledaného motivu

Algoritmus využívající analýzu
prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Předchozí týden

- ▶ **Systemová biologie zkoumá**
 - ▶ život na více úrovních
 - ▶ vztahy mezi součásti živého systému
 - ▶ matematické modely biologických procesů

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Outline

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

DP - Needleman-Wunsch

Vylepšení pro maximálně k chyb

video HHMI

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Základní pojmy

abeceda	$\{\epsilon, a, c, g, t\}$
podřetězec	aag gtacg cgt
prefix	gtacg cgtggt
suffix	cgtat gtacg

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Řetězce a algoritmy na
řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu
hledaného motivuAlgoritmus využívající analýzu
prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Základní operace

konkatenace

 $x = \text{cggat}$ $y = \text{att}$ $x.y = \text{cggatatt}$

průnik

 $x = \text{cggat}$ $y = \text{att}$ $\text{Over}(x, y) = \text{at}$

sjednocení

 $x = \text{cggat}$ $y = \text{att}$ $\langle x, y \rangle = \text{cggatt}$

Výskyt sekvenčních motivů v databázích

Cílem je zjistit všechny pozice delšího řetězce, na kterých se vyskytuje kratší řetězec

- ▶ přesný výskyt
- ▶ přibližný výskyt

řetězec t dlouhý (n), např genomová sekvence
motiv p krátký (m), např `cgcgggctgggtggctcg`

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu
hledaného motivu

Algoritmus využívající analýzu
prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Naivní algoritmus

a c t g t g t a t g a a a t c g c
1..n → t g t c a
1..m →

Složitost: $O(mn)$

Boyer-Moore

a c t g t g t a t g a a a t c g c
→ g a t c a t
 x ↑ ↑ ←

máme v motivu další t?

a c t g t g t a t g a a a t c g c
+1 → g a **t** c a t

kde máme v motivu další výskyt suffixu at?

a c t g t g t a t g a a a t c g c
+3 → g **a t** c a t

Realizujeme krok, který je větší

Složitost

konstrukce: $O(\|abeceda\|.m)$

hledání: $O(mn)$ (v praxi ale blíže k $O(n)$)

Řetězce a algoritmy na
řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu
hledaného motivu

Algoritmus využívající analýzu
prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

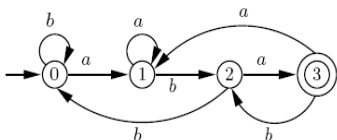
Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Automat pro hledání řetězce aba



Automat vytvořen z motivu p postupně čte symboly z řetězce m . Koncový stav automatu dosáhneme po načtení celého hledaného motivu.

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

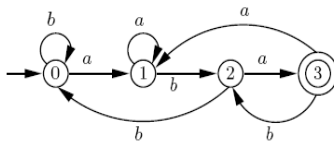
Vylepšení pro maximálně k chyb

Příště

video HHMI

t=bababaa p=aba

ε	0
b	0
ba	1
bab	2
baba	3
babab	2
bab aba	3
bababaa	1



Složitost

konstrukce: naivní $O(m^3)$; optimální $O(\|abeceda\|.m)$

hledání: $O(n)$

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

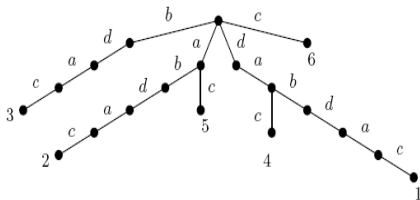
Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Suffixový strom pro řetězec dabcdac



Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

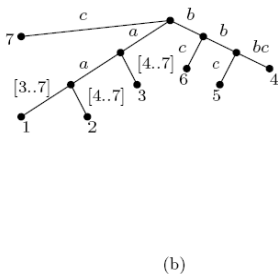
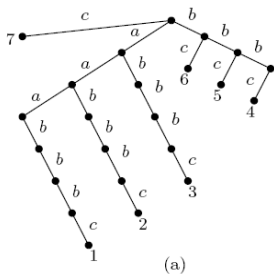
Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Kompaktní suffixový strom pro řetězec aaabbbbc



Konstrukce: $O(n \cdot \log n)$
 Hledání: $O(m \cdot \|abeceda\| + k)$

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Sufixové pole - ukazovatele na polohy suffixů seřazené lexikograficky

Dlouho bylo považováno za méně kvalitní datovou strukturu, protože neobsahuje přímo informace o společných prefixech. Ty lze však spočítat do lcp pole (least common prefix) tak, že konstrukce pole i stromu má stejnou složitost.

$t = \text{dabdac}$

$\text{sa}(t) = 6, 1, 4, 2, 5, 0, 3$

$\text{rank}(t) = 5, 1, 3, 6, 2, 4, 0$

$\text{lcp}(t) = 0, 0, 1, 0, 0, 0, 2$

```

6 0
1 0 abdac
4 1 ac
2 0 bdac
5 0 c
0 0 dabdac
3 2 dac

```

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Outline

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

DP - Needleman-Wunsch

Vylepšení pro maximálně k chyb

video HHMI

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Tandemová a palindromická opakování nesou biologický i praktický význam

palindrom možná sekundární struktura DNA nebo RNA

tandem regulace genů, telomery, identifikace jedinců z DNA

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Řetězce a algoritmy na
řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu
hledaného motivu

Algoritmus využívající analýzu
prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

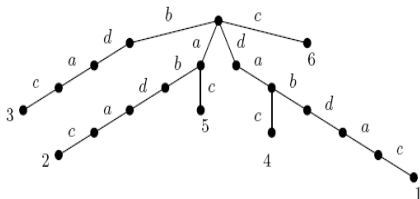
video HHMI

Nejdelší společný prefix dvou pozic

t g c a g a a g c a g a t c c t g a c g
↑ ↑

Složitost naivního algoritmu $O(n^3)$

Posuzování tandemových opakování pomocí suffixových stromů, příp. polí ($t=dabdac$)



$\text{lcp}(1,4)=?$

Nalezneme větve označené 1 a 4

$\text{lcp}(1,4)=da$

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Řetězce a algoritmy na
řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu
hledaného motivu

Algoritmus využívající analýzu
prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Hledání tandemových opakování

- ▶ konstrukce stromu: $O(n \cdot \log n)$
- ▶ hledání lcp pro dvě konkrétní pozice $O(n \cdot \log n)$
- ▶ Prohledávání sekvence

Složitost: $O(n \cdot (\log n)^2 + p)$

Outline

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

DP - Needleman-Wunsch

Vylepšení pro maximálně k chyb

video HHMI

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Řetězce a algoritmy na
řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu
hledaného motivu

Algoritmus využívající analýzu
prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Výpočet omezeného počtu buněk v tabulce DP

Stačí počítat $2k+1$ diagonál bez ohledu na délku sekvencí

Složitost: $O(kn)$ (naproti $O(mn)$)

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu
hledaného motivu

Algoritmus využívající analýzu
prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Využití SA a LCP k rychlému postupu po diagonále

Složitost: $O(k^2)$

Příště Video HHMI

Řetězce a algoritmy na řetězcích

Základní pojmy

Základní algoritmy

Algoritmus využívající analýzu hledaného motivu

Algoritmus využívající analýzu prohledávaného řetězce

Hledání opakování

Tandemové opakování

Palindromy

Srovnávání dvou sekvencí

Vylepšení pro maximálně k chyb

Příště

video HHMI

Outline

Dodatek

Dodatek

For Further Reading

Dodatek

For Further Reading

For Further Reading
X