Pattern branching jako seed pro Projection algorithm

ZDENĚK TESAŘ

Obsah

Cíl

Pattern branching

- rychlý algoritmus na hledání motifů
- nedosahuje 100% přesnosti

Projection algorithm

- pomalejší algoritmus
- obecně dosahuje 100% přesnosti

Využít rychlosti Pattern branchingu s horší přesností pro urychlení Projection algorithmu s lepší přesností

Pattern branching

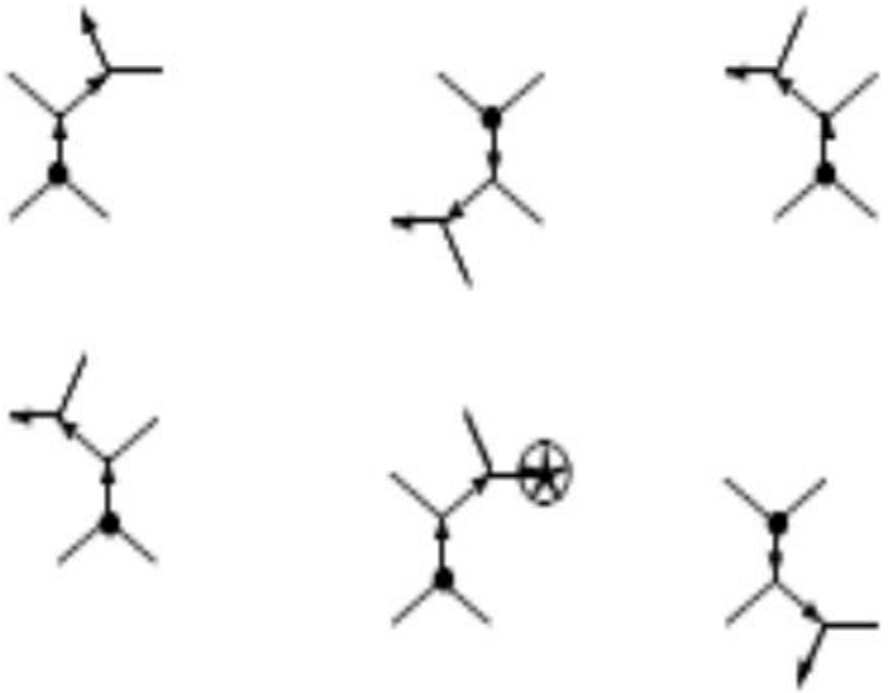
Aproximační algoritmus

Prostor motifů

PatternBranching (DNA, l, k):
$Motif$ = arbitrary motif pattern
For each l -mer A_0 **in** DNA
    For $j = 0$ **to** k
        If $d(A_j, DNA) < d(Motif, DNA)$
            $Motif = A_j$
        $A_{j+1} = BestNeighbor(A_j)$
Output $Motif$

Pattern branching

Skórování



$$d(A, DNA_i) = min\{d(A, P) \mid P \in DNA_i\}$$

$$d(A, DNA) = \sum_{DNA_i \in DNA} d(A, DNA_i)$$

Projection algorithm

Aproximační algoritmus

Prostor startovních pozic

Input: sequences $s_1, \ldots, s_t$, parameters k, s and m

Output: best guess motif

for $i = 1$ to m do

- choose k different positions $I_k \subset \{1, 2, \ldots, l\}$
- for each l-mer $x \in s_1, \ldots, s_t$ do
  - compute hash value $h_{Ik}(x)$
  - Store x in hash bucket
- for each bucket with $\geq s$ elements do
  - refine bucket using EM algorithm

return consensus pattern of the best refined bucket

Projection algorithm

Vhodné k

Vhodné s

Vhodné m

$$k < l - d$$

$$4^k > t(n - l + 1)$$

$$k > \frac{\log(t(n-l+1))}{\log(4)} = \frac{\log(20(600-15+1))}{\log(4)} \approx 6{,}76$$

$$m = \left\lceil \frac{\log(1-q)}{\log\left(B_{t,p'(l,d,k),s}\right)} \right\rceil$$

Možnosti urychlení

Přidat novou sekvenci DNA = nalezenému motifu z Pattern branchingu

Přidat vícekrát (celou sekvenci, vícekrát za sebou, …)

Na základě přidání sekvence upravit parametr m

- Statisticky
- Experimentálně

Problémy

Nepodařilo se mi plně implementovat Projection algorithm

Pokus použít hotové knihovny

- Malý počet knihoven (SeQan C++, Bio.motif, …)
- Většinou používají modernější sofistikovanější úpravy
- Malá či skoro žádná možnost zasáhnout dovnitř
- Použitá knihovna neměla nastavitelný parametr m

Výsledky

Benchmarky

- University of Washington
- Performance evaluation of DNA Motif discovery programs
- Ručně implantované motivy do existujících sekvencí

Pouze pár zkušebních iterací

Při přidání sekvence motifu nalezeného Pattern branchingem urychlení ~ 10-35 %

Děkuji za pozornost

OTÁZKY?