新しいデータ圧縮

簡潔de Bruijnグラフによるゲノムアセンブリ

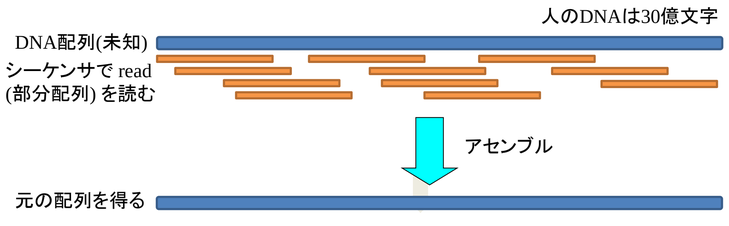
どんな研究？

様々な情報処理を行う際に問題となるのは，データの量と，処理を高速に行うためのデータ構造（索引）のサイズである．本研究ではデータ構造を圧縮してサイズを小さくする手法を開発している．特徴は，圧縮したまま処理を高速に実行できる点である．

何がわかる？

簡潔de Bruijnグラフは，de Bruijnグラフを圧縮したものである．de Bruijnグラフは生物のDNA配列を決定する際に用いられるデータ構造だが，サイズが大きいという問題がある．これを圧縮することで，1台の計算機でDNA配列の決定を行うことができるようになる．

問題設定



- readの長さは 30 ~ 1500 (エラー率1%)
- エラーが多いので複数回読む (40~100回)
- read中の k-mer をノードとする de Bruijn グラフを作る

de Bruijnグラフ

入力: readの集合 R

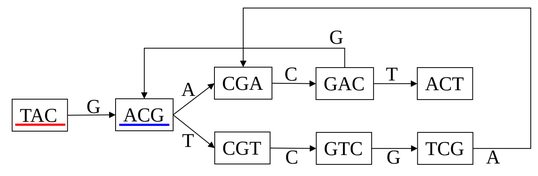
定義: $K_d^k(R) = \left\{ \begin{array}{l} R\text{に現れる長さ}k\text{の部分文字列で} \\ d\text{回以上出現するもの} \end{array} \right\}$

$$G^{k,d} = \left(V^{k,d}, E^{k,d}\right)$$

$$E^{k,d} = K_d^{k+1}(R) \quad (\text{節点と，出る枝のラベル})$$

$$V^{k,d} = \underbrace{\left\{u[1,k] \mid u \in E^{k,d}\right\}}_{V_L\text{: ラベルの長さ}k\text{の接頭辞}} \cup \underbrace{\left\{u[2,k+1] \mid u \in E^{k,d}\right\}}_{V_R\text{: ラベルの長さ}k\text{の接尾辞}}$$

T = TACGACGTCGACT
k = 3



研究内容

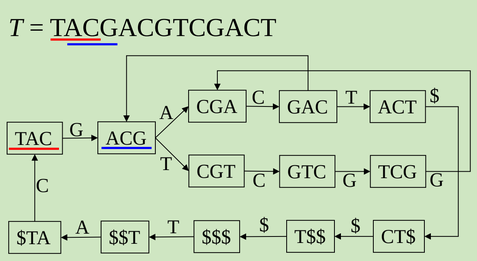
既存手法の問題点

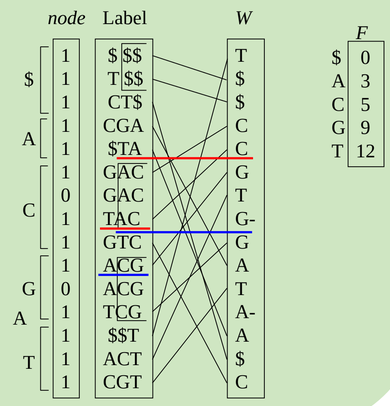
- de Bruijn グラフの使用するメモリ (k = 23~27)
  - ABySS [Simpson et al. 09]
    hash based, ~300 bits / edge
  - Gossamer [Conway, Bromage 11]
    succinct bit vector based, 28.5 bits / edge
  - probabilistic de Bruijn graph [Pell et al. 11]
    Bloom filter, 4~9 bits / edge, ただし false positive あり
  - Minia [Chikhi, Rizk 12]
    probabilistic de Bruijn graph + critical false positives
    13.5 bits / edge
- これらは log k または k に比例して増える
- 本研究: 4+o(1) bits / edge (k に依存しない)

Succint de Bruijn Graph

- *node*, *W*, *F* だけでグラフを表現
- $(m+2k-1)(2+\log \sigma) + o(m)$ bits　(m: 枝数, $R = \{T\}$)
- outdegree(v) , outgoing(v, c), indegree(v): O(1) time
- Label(v), Index(u): O(k) time
- incoming(v, c): O(k log σ) time

T = TACGACGTCGACT





fwd(u)

$c = W[u]$
$r = rank_c(W, u)$
$x = F[c] + r - 1$
$v = select_1(last, x)$

bwd(v)

$x = rank_1(last, v)$
$c = F^{-1}[x]$
$r = x - F[c] + 1$
$u = select_c(W, r)$

O(1) 時間

注: $W[u] \in A^-$ なら対応する A の文字に変換

outgoing(v, c)

節点 v から出る枝ラベルの格納されている範囲で c または $c-$ を探す

$x = select_c(W, rank_c(W, v))$
$y = select_{c-}(W, rank_{c-}(W, v))$
x, y 共に範囲外なら枝は存在しない
$w = fwd(x)$

O(1) 時間

ヒトゲノムでの実験結果

- Readの長さ: 100, #reads 1,408,414,537
- k = 27
- 異なる (k+1)-mer の数: 36,248,317,498
- d = 3 回以上現れる(k+1)-mer の数: 4,998,165,473
- dummy edge の数: 343,459,610
- データ構造のサイズ
  - ABySS: 336GB (ハッシュ)
  - Gossamer: 32GB (簡潔ベクトル)
  - Minia: 5.7GB (Bloom filter)
  - 本研究: 2.5GB

連絡先：定兼　邦彦／ 国立情報学研究所　情報学プリンシプル研究系　准教授　sada@nii.ac.jp
TEL：00-000-0000　FAX：00-000-0000　Email：aaaaaaaa@nii.ac.jp