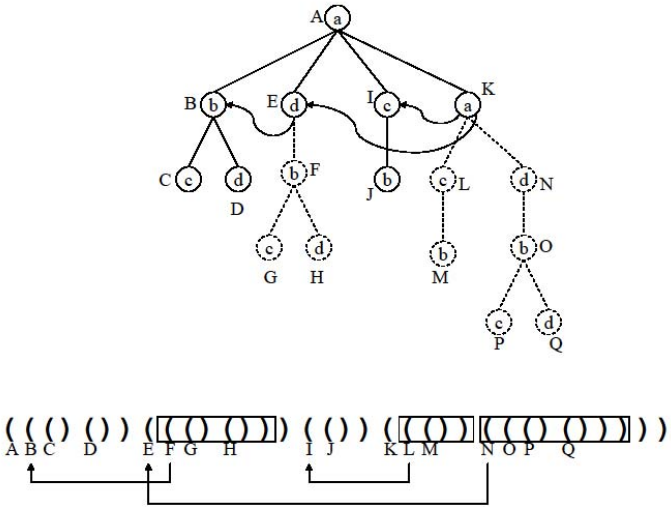
文字列置換と文脈自由文法によるデータ圧縮

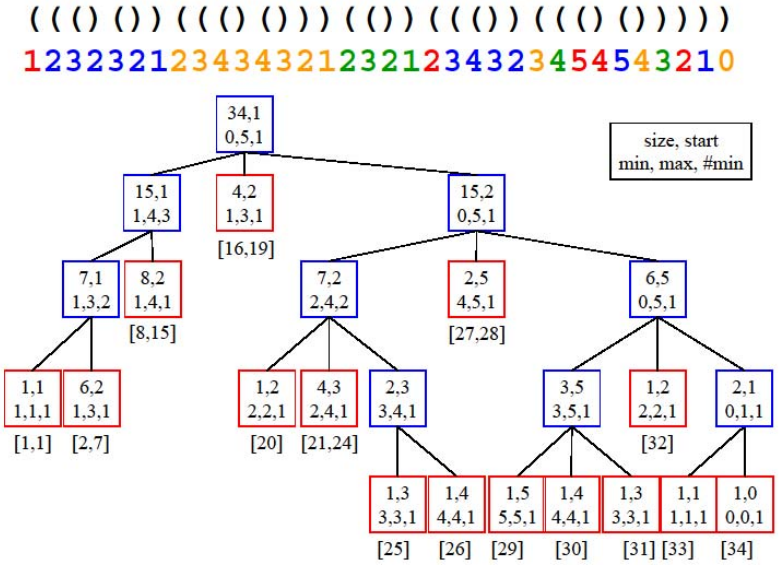
定兼　邦彦

データの圧縮率はその中身に依存する．英語や日本語などの文章は同じ単語を多く含むため良く圧縮できるが，DNA配列は圧縮しにくい．しかし同じ生物のDNA配列はほとんどの部分が同じで，個体による差は少ない．このことを利用すると，大勢の人のDNA配列データをコンパクトに圧縮できる．本研究では文字列置換を用いた類似データの圧縮および検索法と，文脈自由文法を用いた圧縮法を開発している．

文脈自由文法によるデータ圧縮

同じ文字列の繰り返しを文脈自由文法で表現することで圧縮する．既存手法よりも高速に部分文字列を復元できる．また，木構造についても同じ部分木構造を共有することで圧縮できる．圧縮した状態でも従来通りの問い合わせを行える（N節点の木に対しO(log N)時間）．





文字列置換によるDNA配列圧縮・検索

1000人ゲノムプロジェクトのように，多くの個体のDNA配列を格納する場合，それらのほとんどの部分は共通であることを利用して圧縮できる．各配列を，参照配列 R の部分文字列に置換する．R の長さを n, エントロピーを H_k, 参照回数を K とすると，データ構造のサイズは O($n + nH_k + K \log n$) bits, パタン P は O($|P| \log n/\log\log n + \mathrm{occ}(\log n + \log K/\log n)$) 時間で検索できる．

R = ACGTGATAG
S_1 = TGATAGACG = TGATAG, ACG = 8 2
S_2 = GAGTACTA = GA, GT, AC, TA = 5 6 1 7
S_3 = GTACGT = GT, ACGT = 6 3
S_4 = AGGC = AG, GA = 4 5

ID	Segment	Pos. in R
1	AC	1..2
2	ACG	1..3
3	ACGT	1..4
4	AG	8..9
5	GA	5..6
6	GT	3..4
7	TA	7..8
8	TGATAG	4..9

NII　連絡先：定兼　邦彦（Kunihiko Sadakane）／ 国立情報学研究所　情報学プリンシプル研究系　准教授