究極のデータ圧縮法

簡潔データ構造による大規模データ処理

発表者:定兼　邦彦　（情報学プリンシプル研究系）

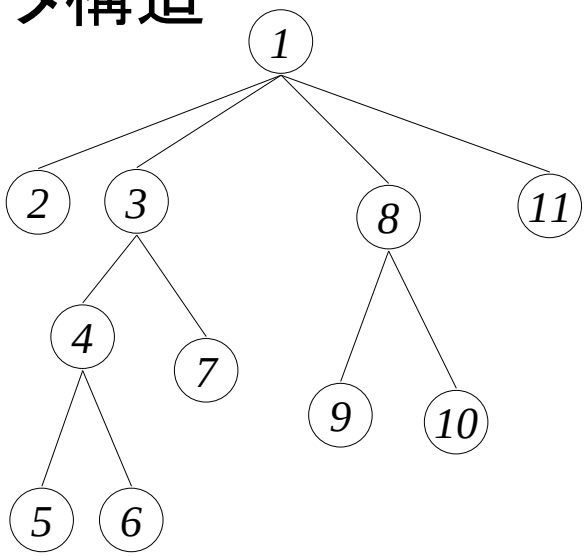
文字列検索の簡潔データ構造

ヒトの全DNA(28億文字)に対し
従来のデータ構造
　接尾辞配列:13GB
　接尾辞木:40GB
簡潔データ構造
　圧縮接尾辞配列:1GB
　圧縮接尾辞木:4GB
サイズが10分の1以下に

BW	*SA*	
g	9	$
$	1	acagcagg$
c	3	agcagg$
c	6	agg$
a	2	cagcagg$
g	5	cagg$
g	8	g$
a	4	gcagg$
a	7	gg$

順序木の簡潔データ構造

従来のデータ構造
→木構造をポインタで表現
簡潔データ構造
→括弧列で表現
　各種操作の時間は同等
サイズが50分の1以下に



1 2 3 4 5 6 7 8 9 10 11

P (( ) ((( ) ( )) ( )) (( ) ( )) ( ))

E **12123434323212323210**