論文紹介

Weighted Directed Word Graph

Meng Zhang and Yi Zhang
Jilin University, College of Computer Science and Technology, China

To appear in CPM2005.

発表者　稲永俊介
日本学術振興会 特別研究員(PD)

概要

- WDWG (Weighted Directed Word Graph) というテキスト索引構造を提案
  - WDWGの定義
  - WDWGのサイズ
  - WDWGの構築
  - 実験

部分文字列照合問題

- 入力：テキスト文字列T，パターン文字列P
- 出力：PはTの部分文字列か？

パターン： compress

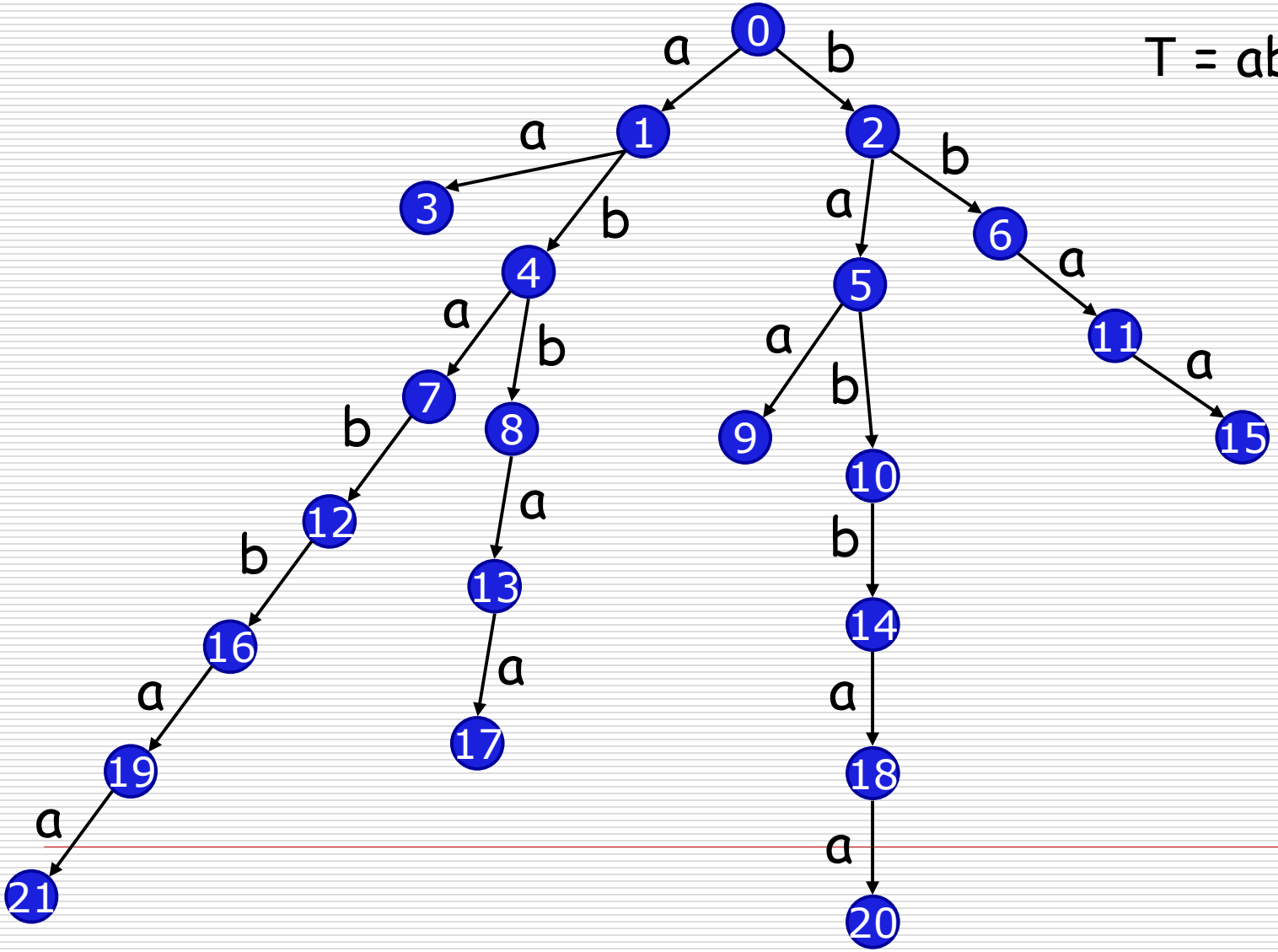
テキスト：

We introduce a general framework which is suitable to capture an essence of compressed pattern matching according to various dictionary based compressions. The goal is to find all occurrences of a pattern in a text without decompression, which is one of the most active topics in string matching. Our framework includes such compression methods as Lempel-Ziv family, (LZ77, LZSS, LZ78, LZW), byte-pair encoding, and the static dictionary based method. Technically, our pattern matching algorithm extremely extends that for LZW compressed text presented by Amir, Benson and Farach [Amir94].

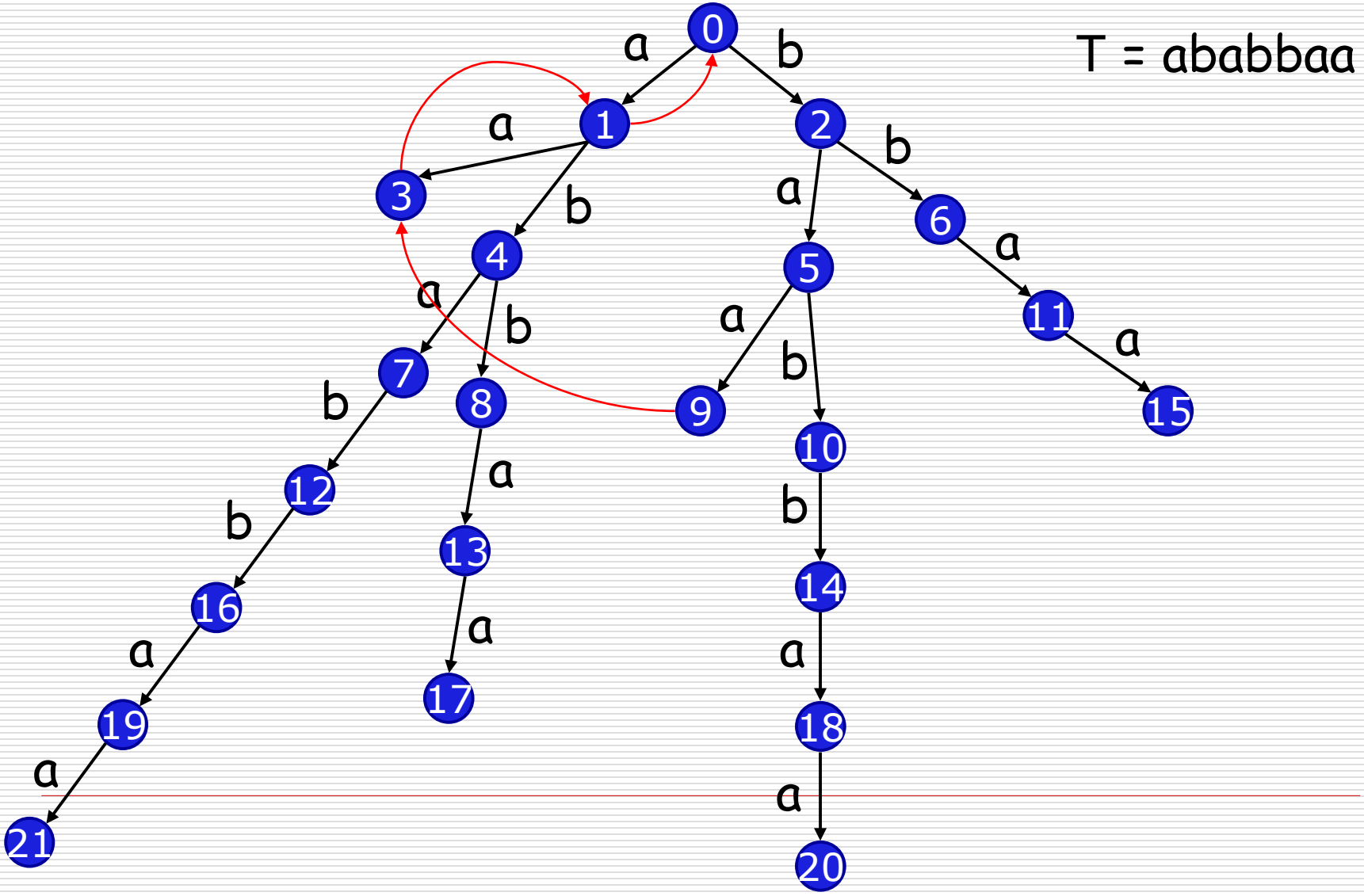
接尾辞トライ

- 文字列Tの接尾辞トライ(Suffix Trie)は，Tのすべての接尾辞を表現する木構造(トライ)である.
  - Tのすべての接尾辞を受理する決定性オートマトンとみなすこともできる.

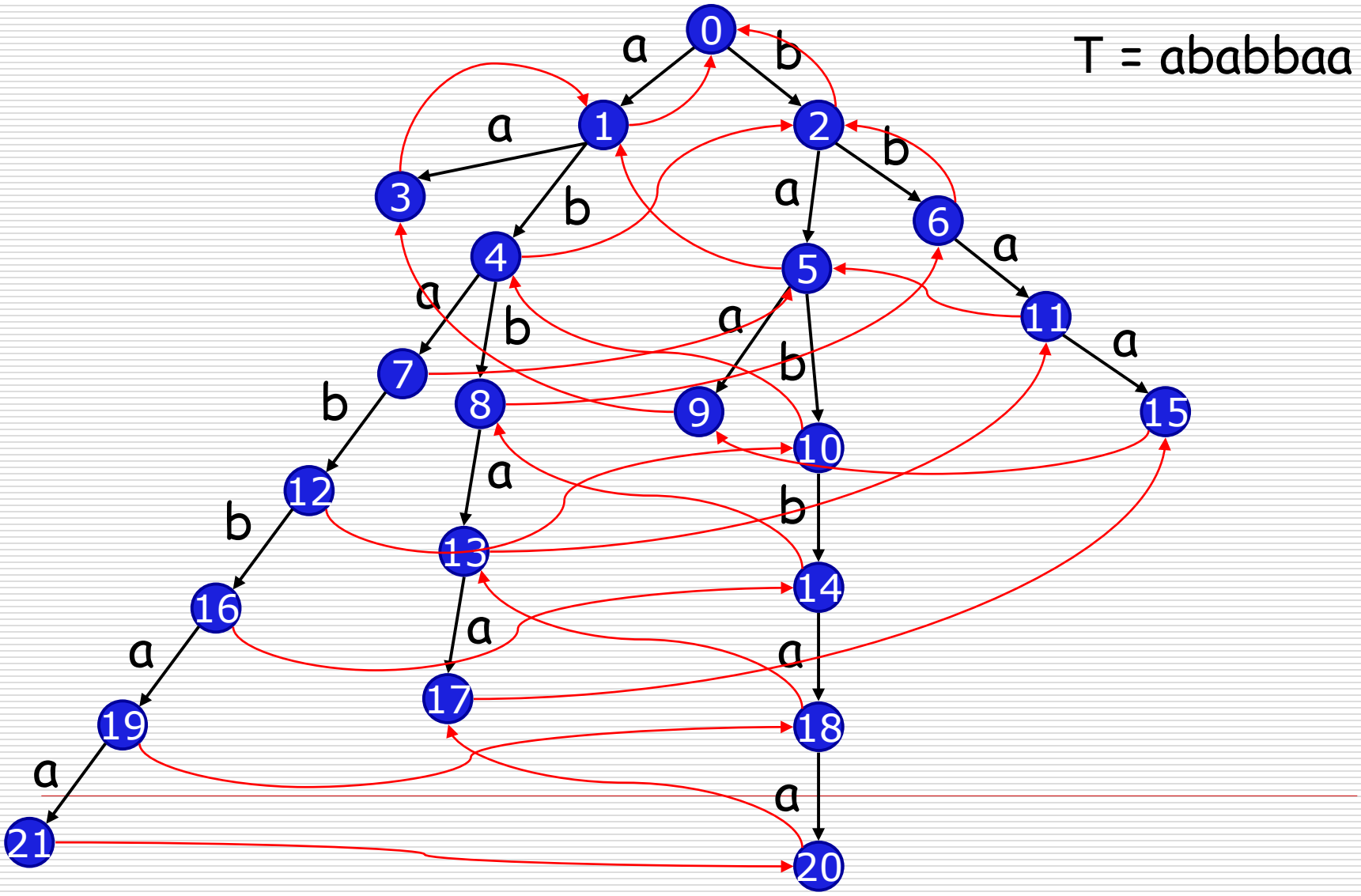
接尾辞トライ(つづき)



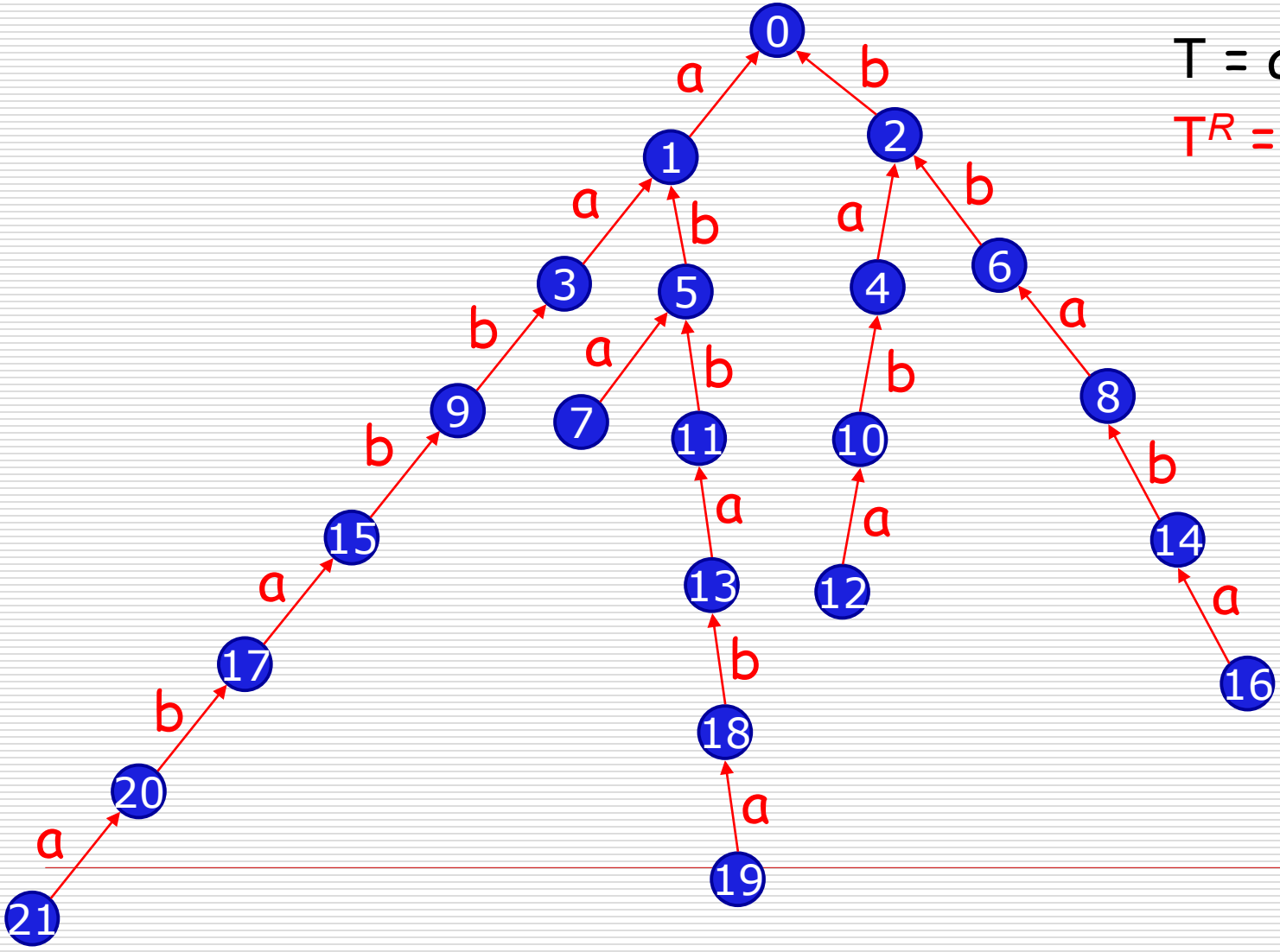
接尾辞リンク



接尾辞リンク



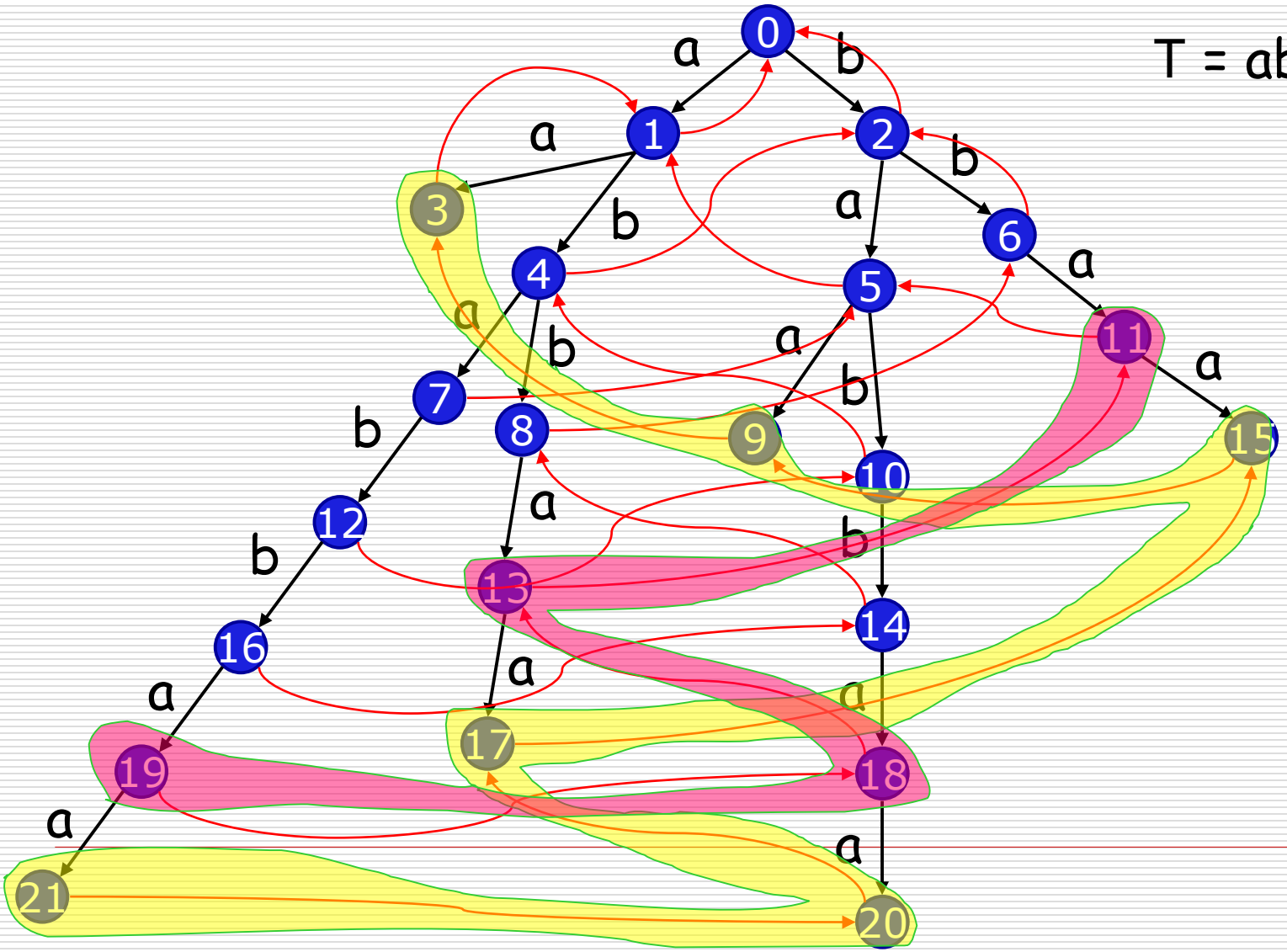
接尾辞リンク(つづき)


0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
a
b
T = ababbaa
T^R = aabbaba

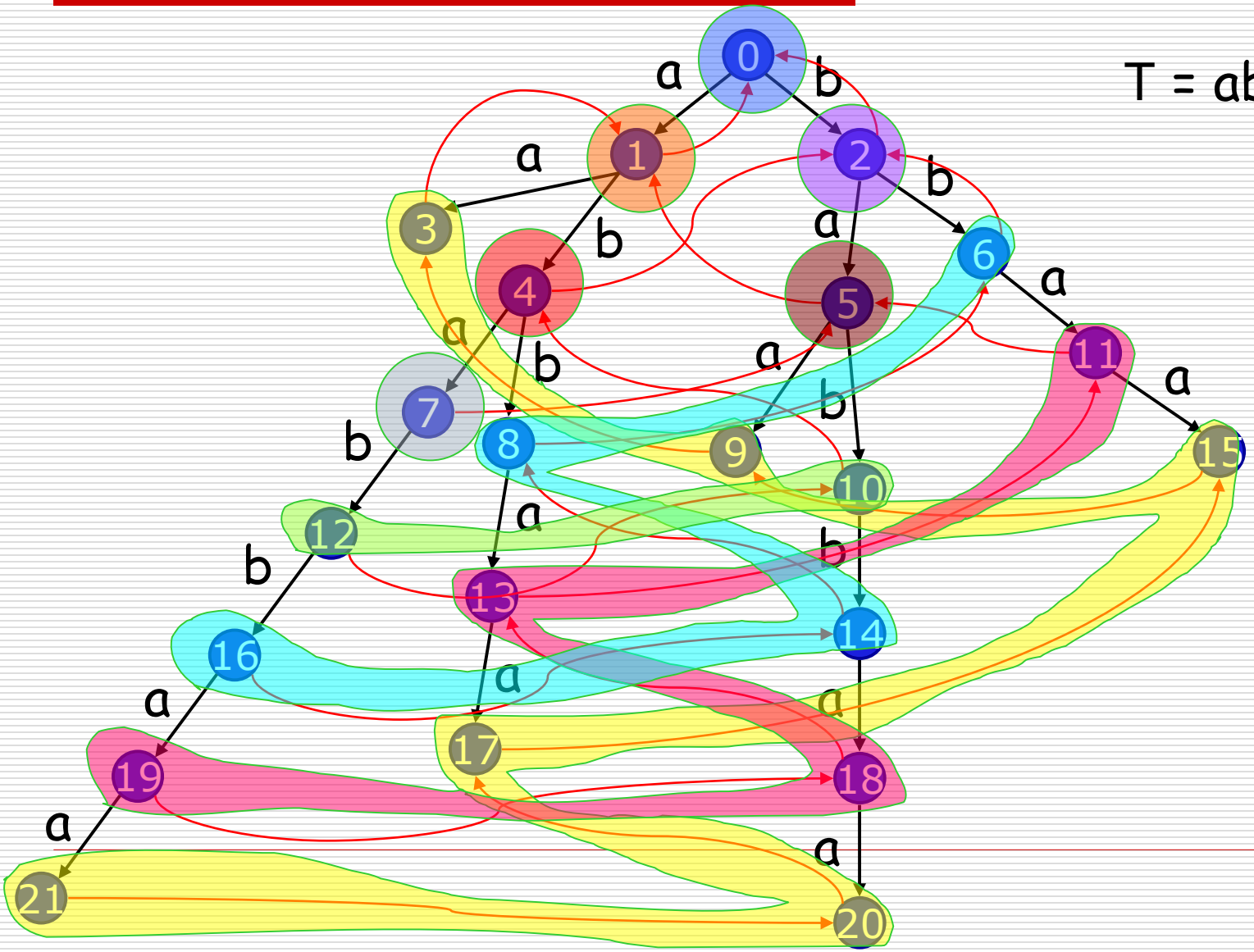
DAWG (Directed Acyclic Word Graph)

- DAWGは文字列Tのすべての接尾辞を受理する最小の決定性オートマトン.
  - 接尾辞トライを最小化するとDAWGになる.

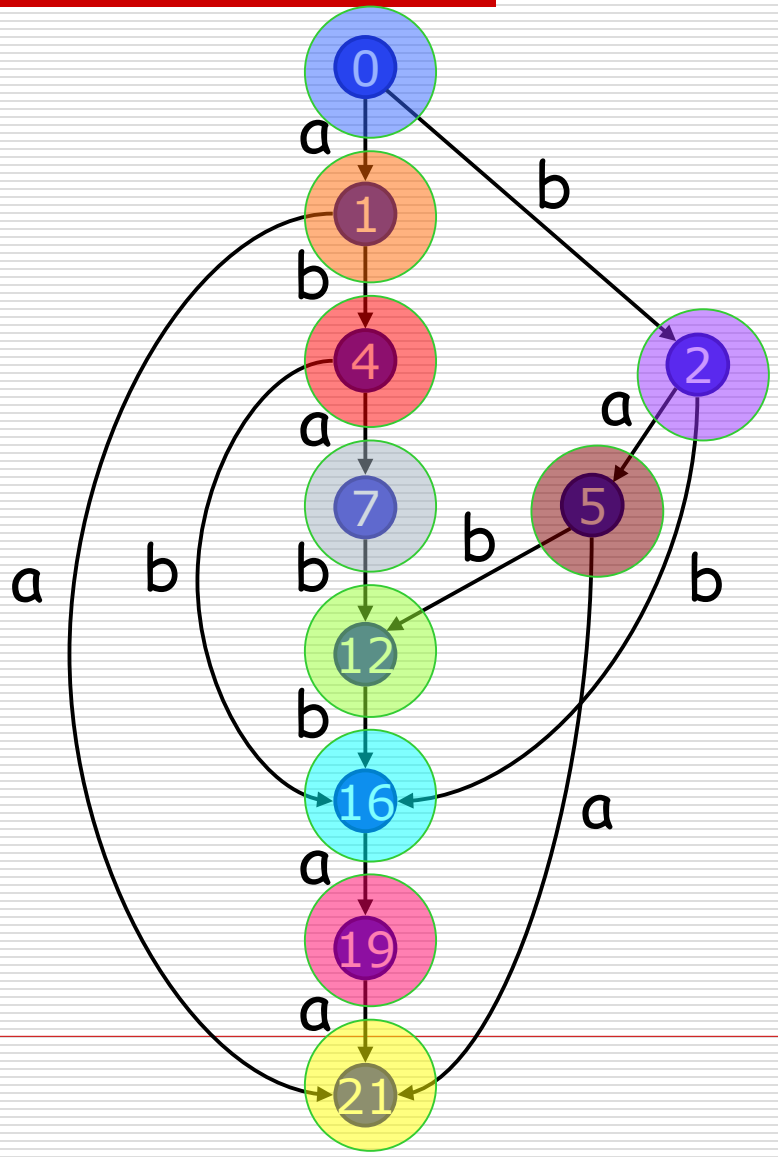
DAWG(つづき)



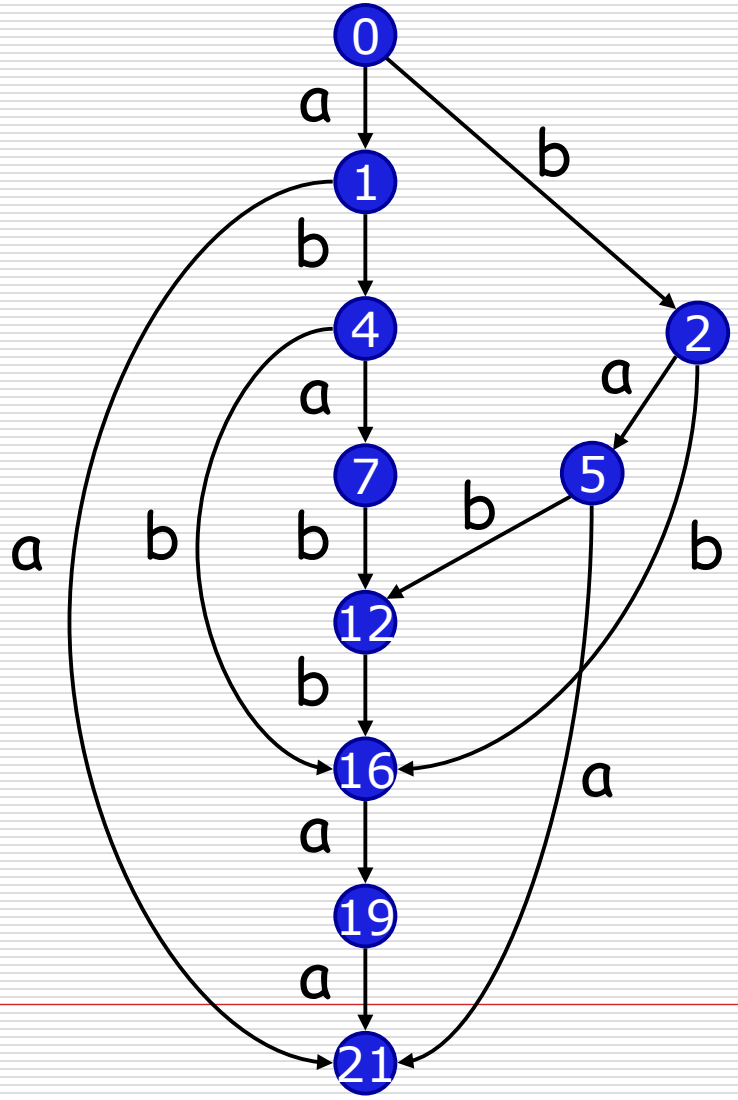
DAWG(つづき)



DAWG(つづき)

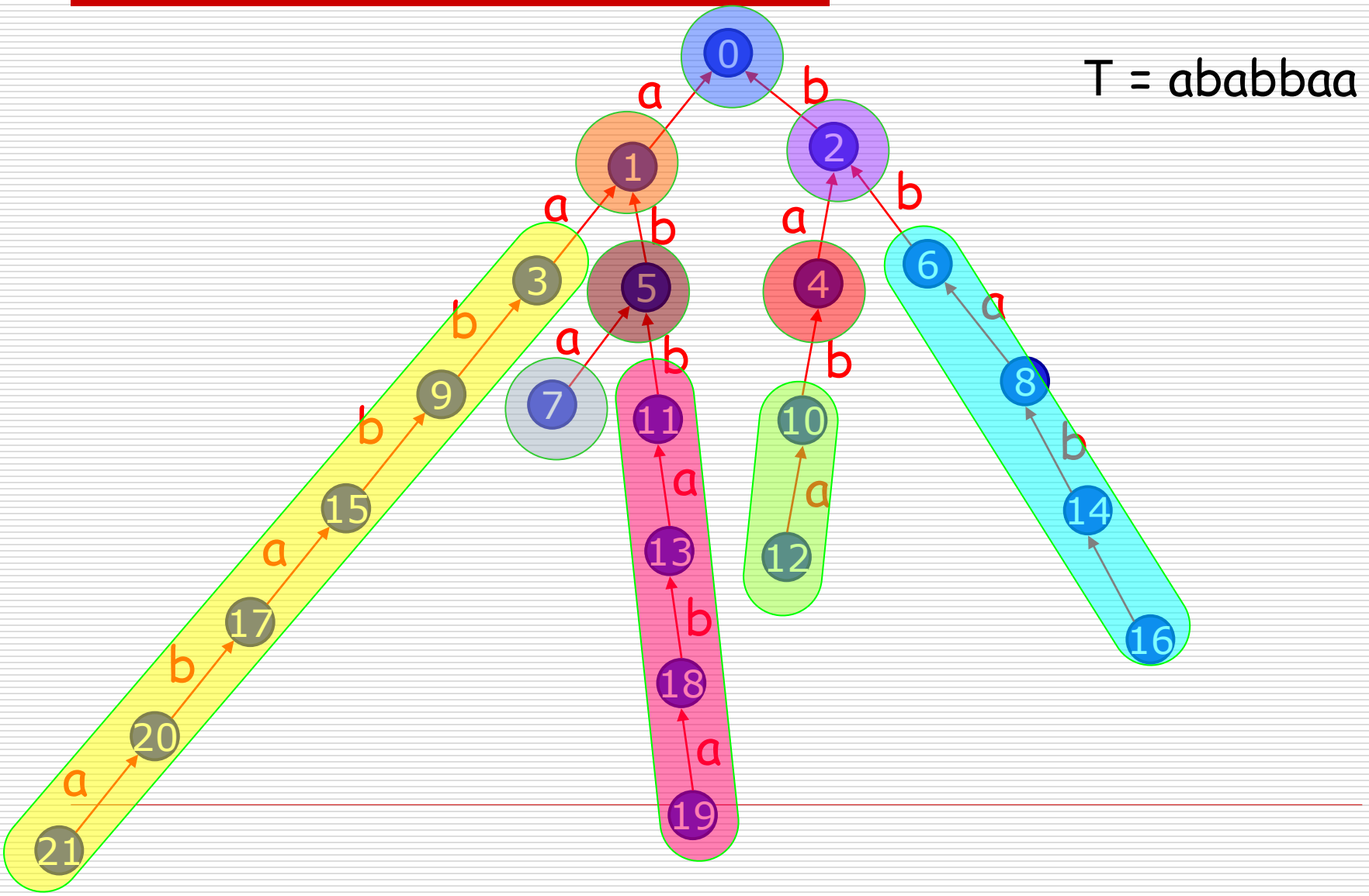


DAWG(つづき)



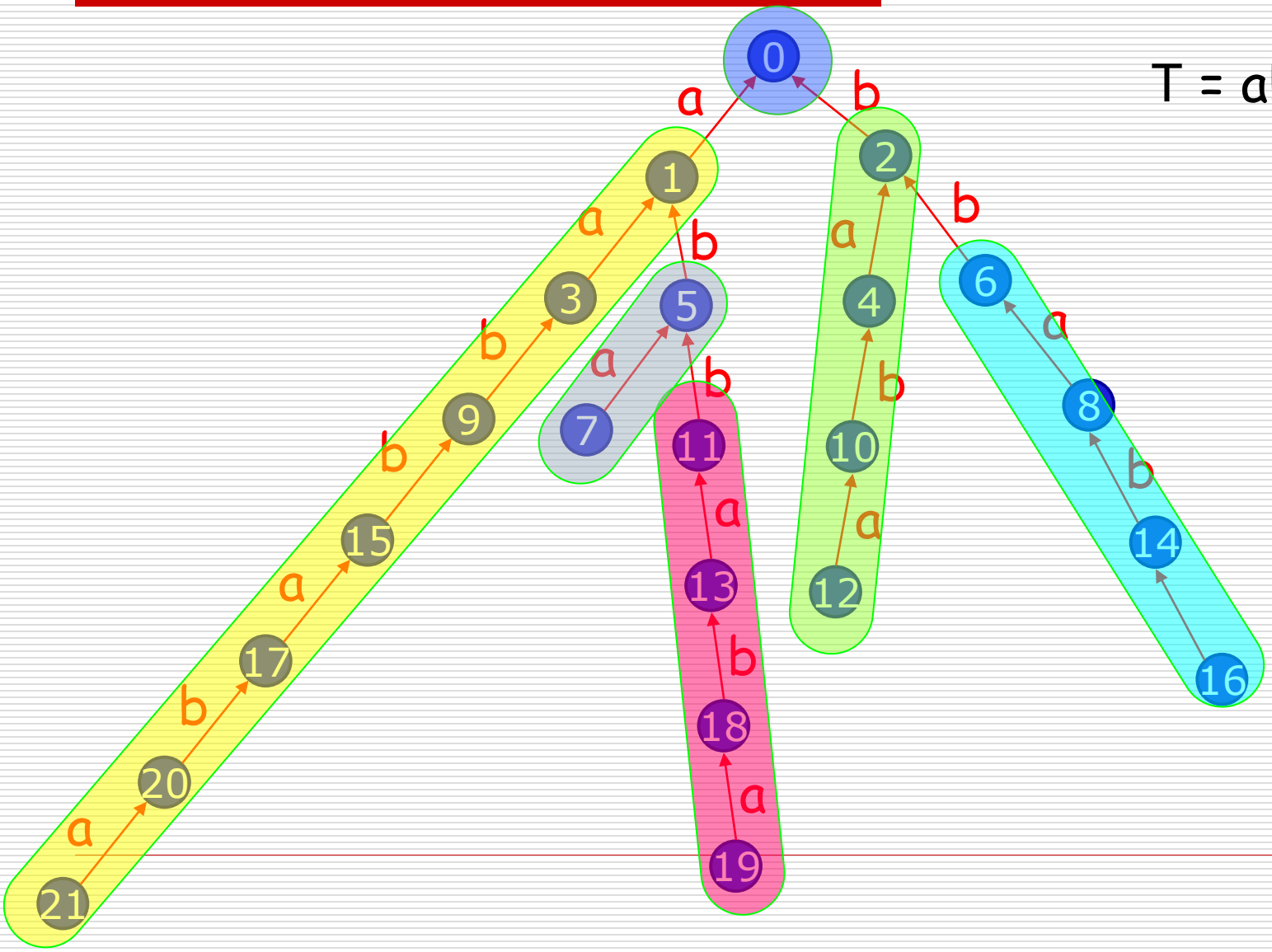
T = ababbaa

DAWG(つづき)

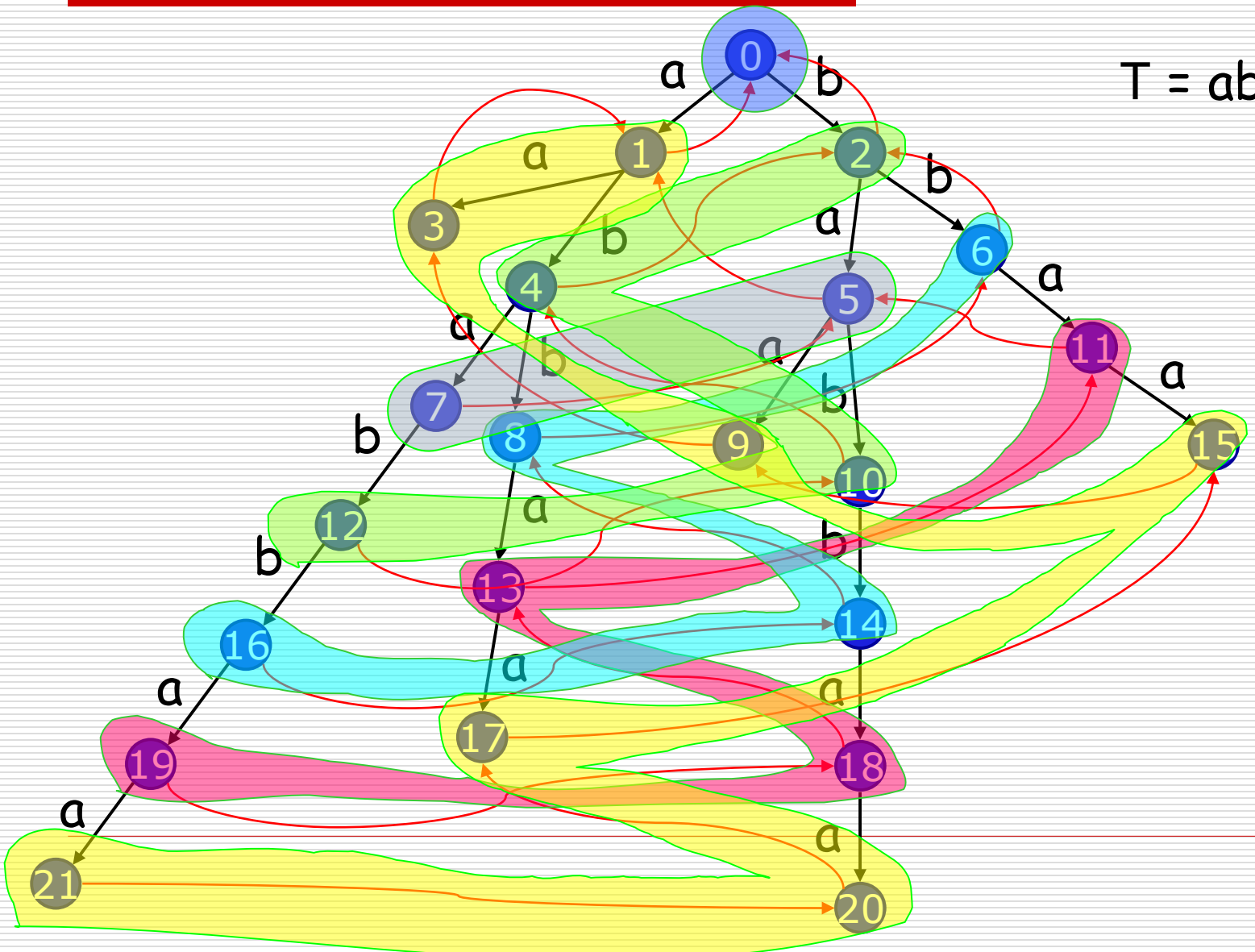


もっと小さく

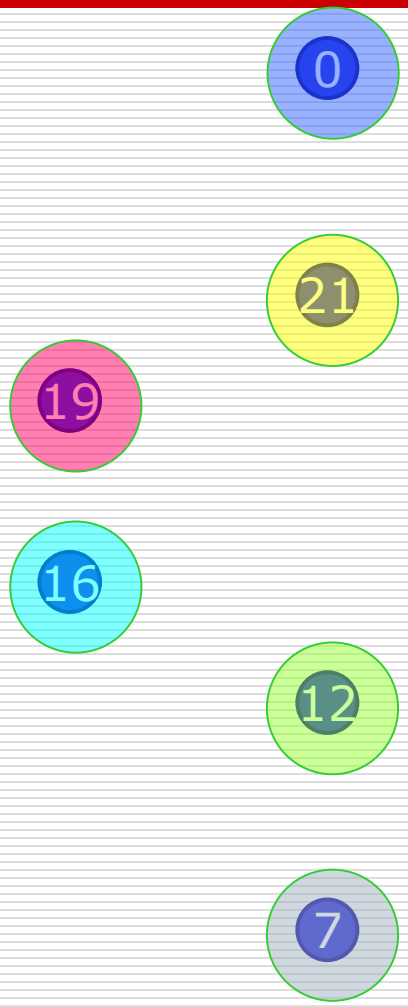
T = abababaa



もっと小さく(つづき)

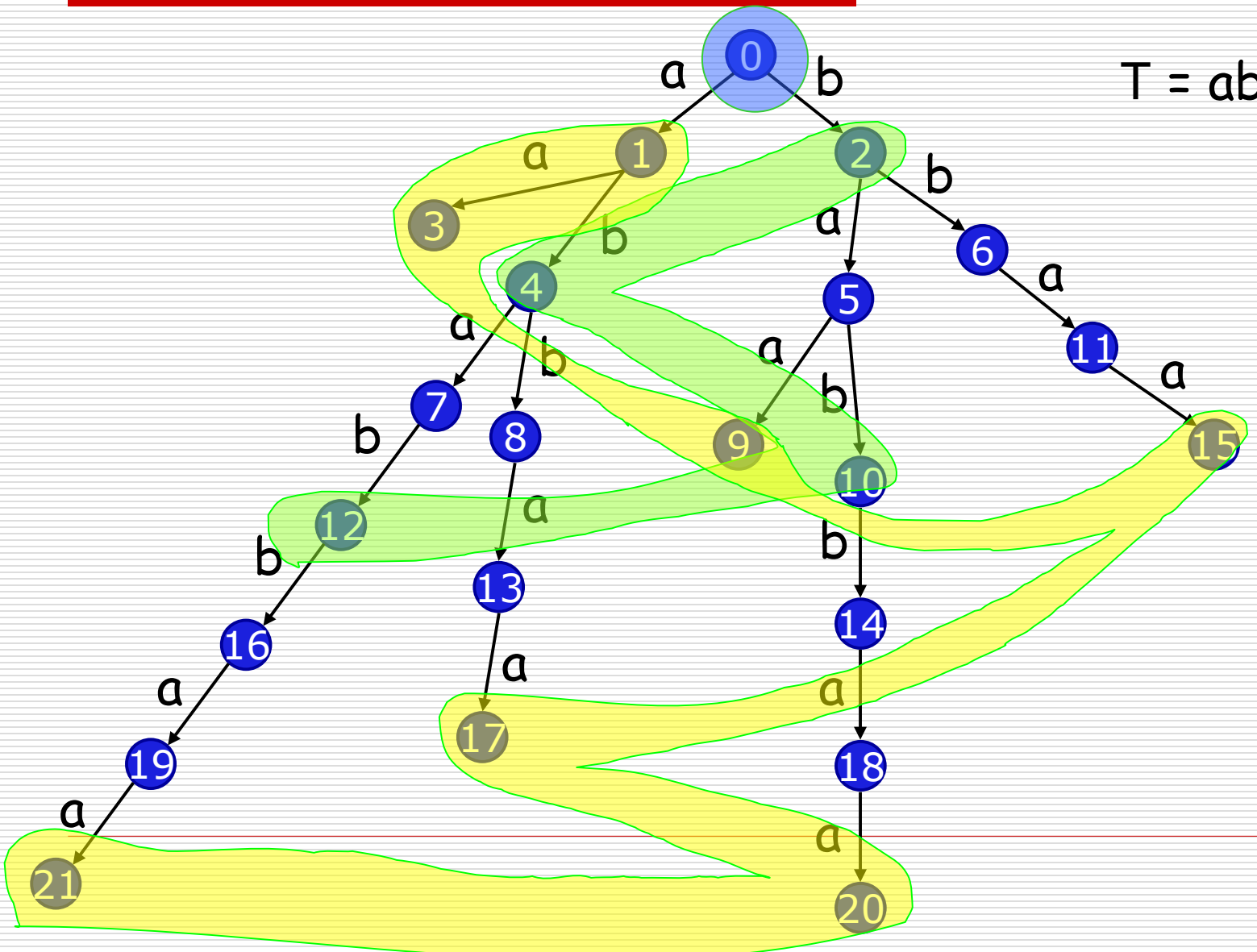


もっと小さく(つづき)

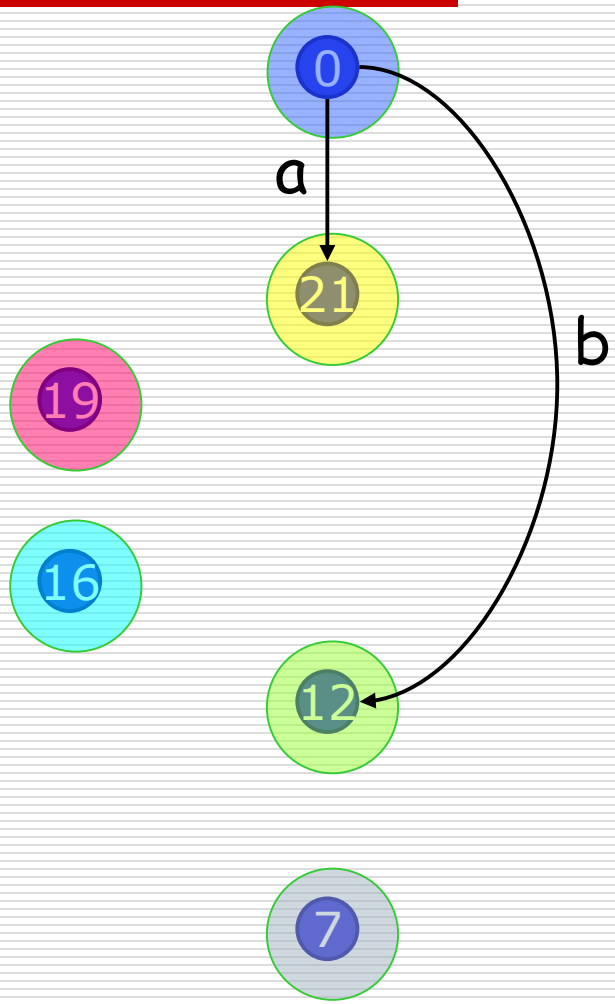


T = ababbaa

もっと小さく(つづき)

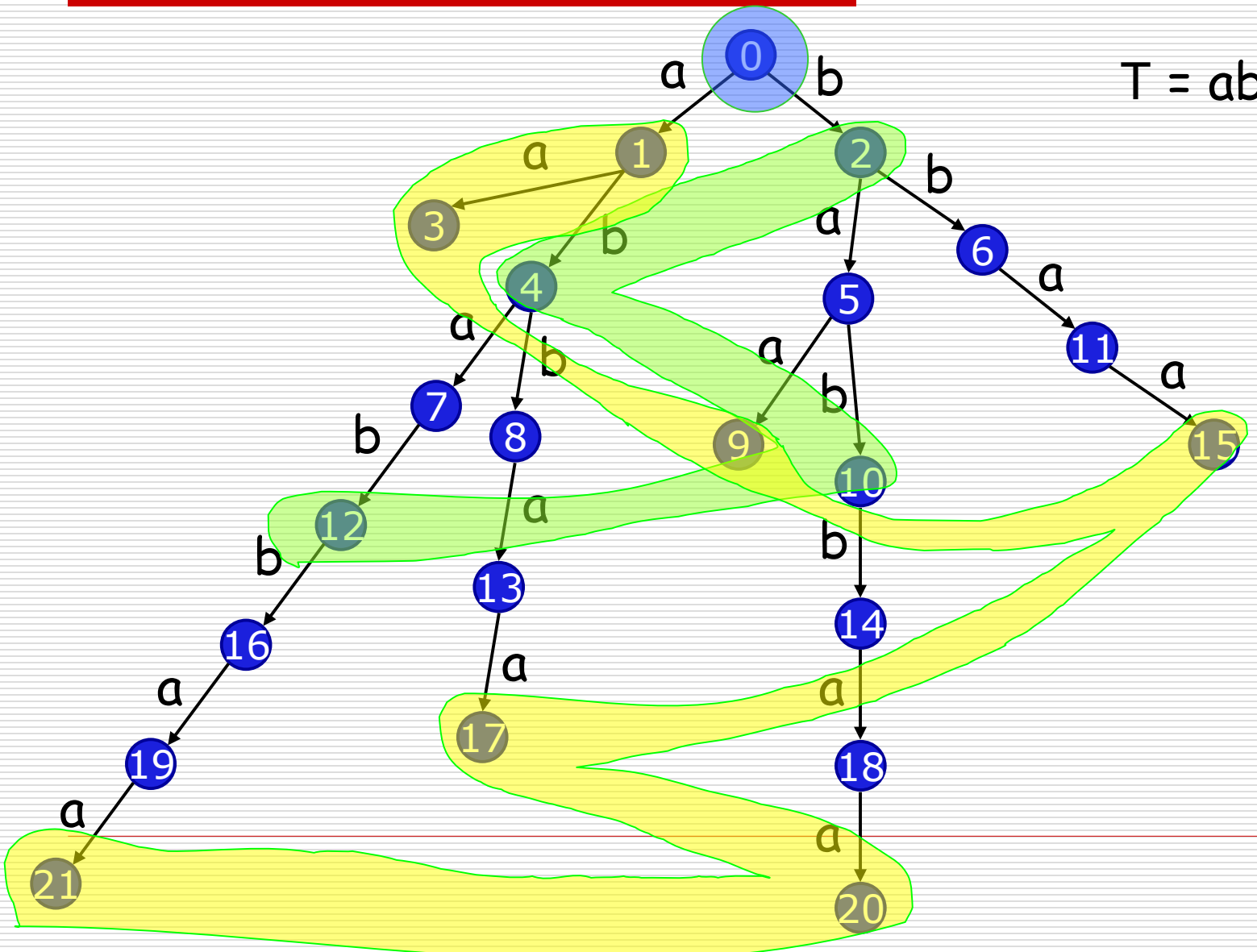


もっと小さく(つづき)

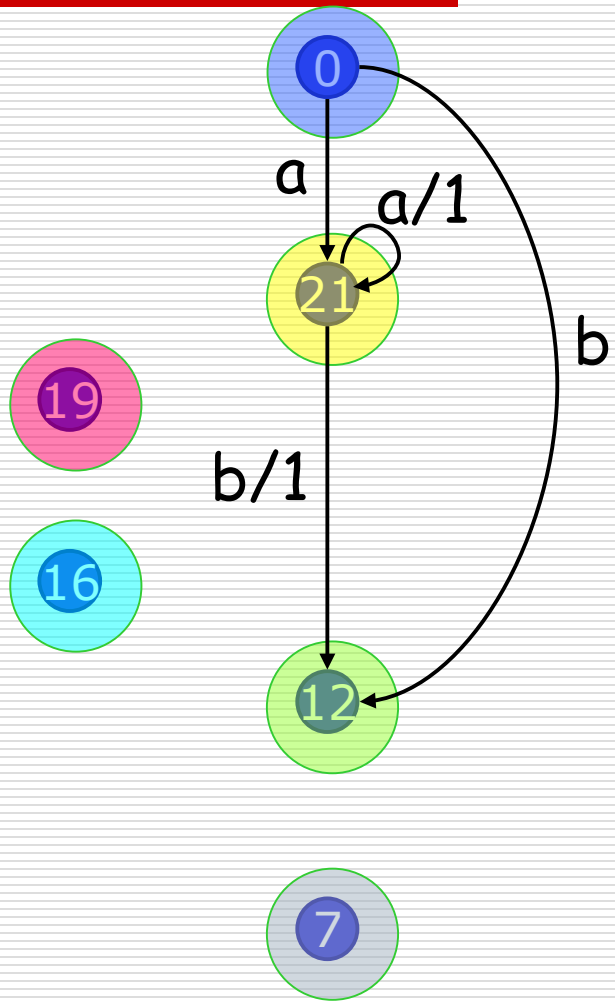


T = ababbaa

もっと小さく(つづき)

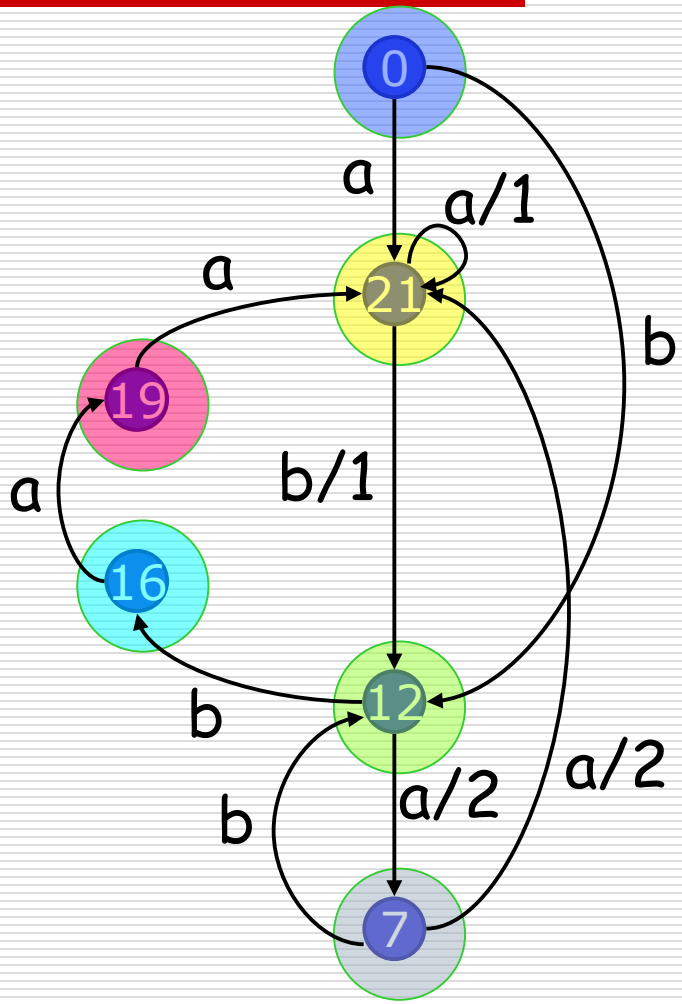


WDWG (Weighted Directed Word Graph)



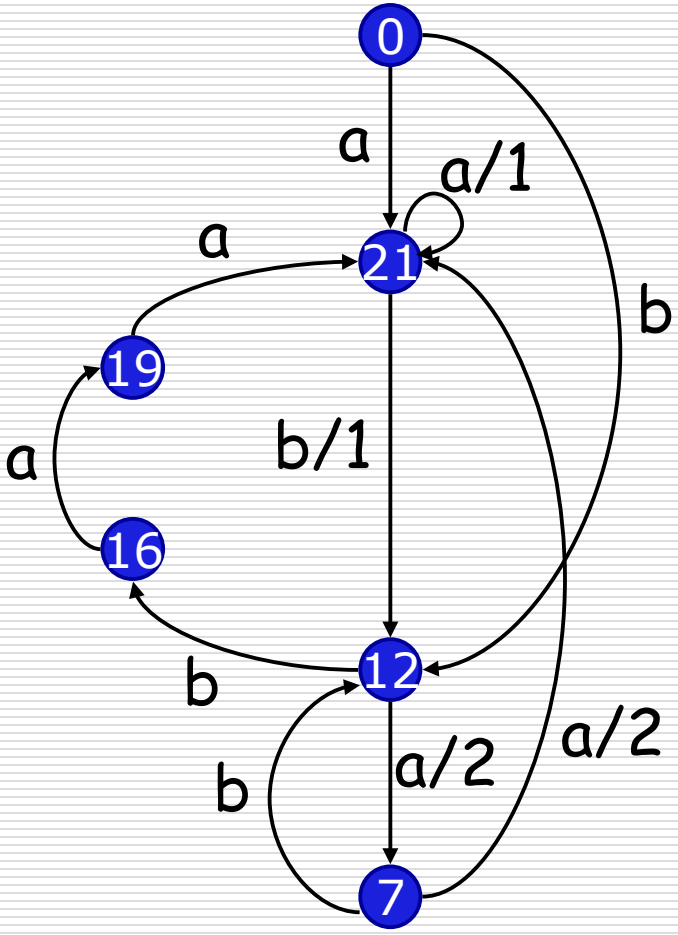
T = ababbaa

WDWG (Weighted Directed Word Graph)



T = ababbaa

WDWG (Weighted Directed Word Graph)



T = ababbaa

WDWGのサイズ

- DAWG
  - 状態数：2n-1
  - 遷移数：3n-3
- WDWG
  - 状態数：n+1
  - 遷移数：2n-1

n：入力文字列Tの長さ

WDWGの構築

- WDWGは線形時間・領域でオンライン構築可能である.

実験

Table 1. Statistic on the size of real DAWGs and WDWGs

source x	$\|\Sigma\|$	$\|x\|$	DAWG			WDWG		
			Number of states	Number of transitions	(Number of states) /$\|x\|$	Number of states	Number of transitions	Bytes per character of x
DNA	4	500000	844244	1235805	1.688488	499978	792996	6.52
DNA	4	500000	826941	1262603	1.653882	499989	808128	6.79
DNA	4	500000	829619	1259255	1.659238	499993	797433	6.60
Random	4	500000	881696	1181151	1.763392	499910	729621	5.38
English	71	100000	153044	214086	1.530440	99996	155982	15.13
English	71	100000	152753	215485	1.527530	99995	157529	15.27