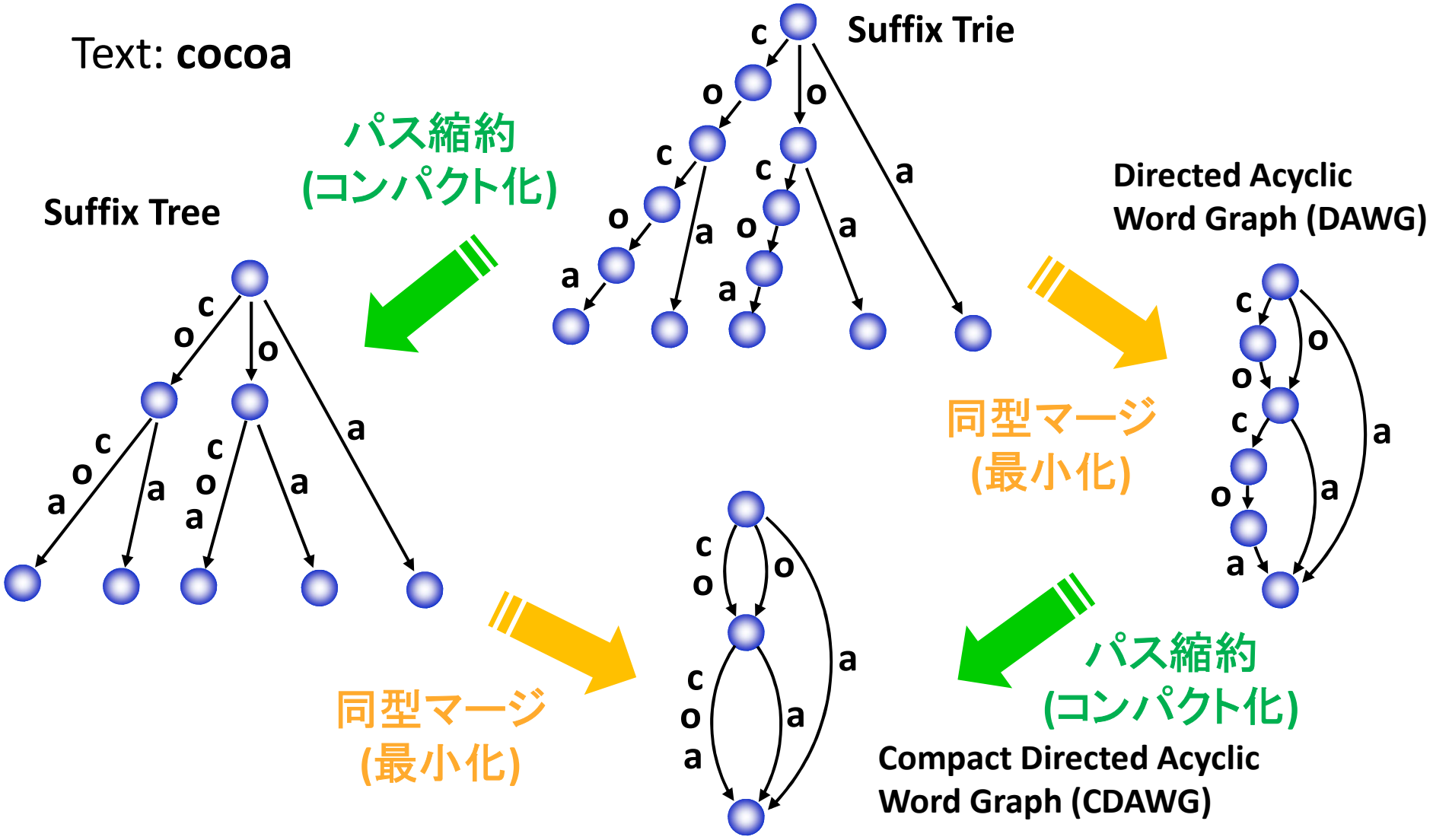
北大有村研セミナー 2023.08.21-23

Symmetric CDAWG

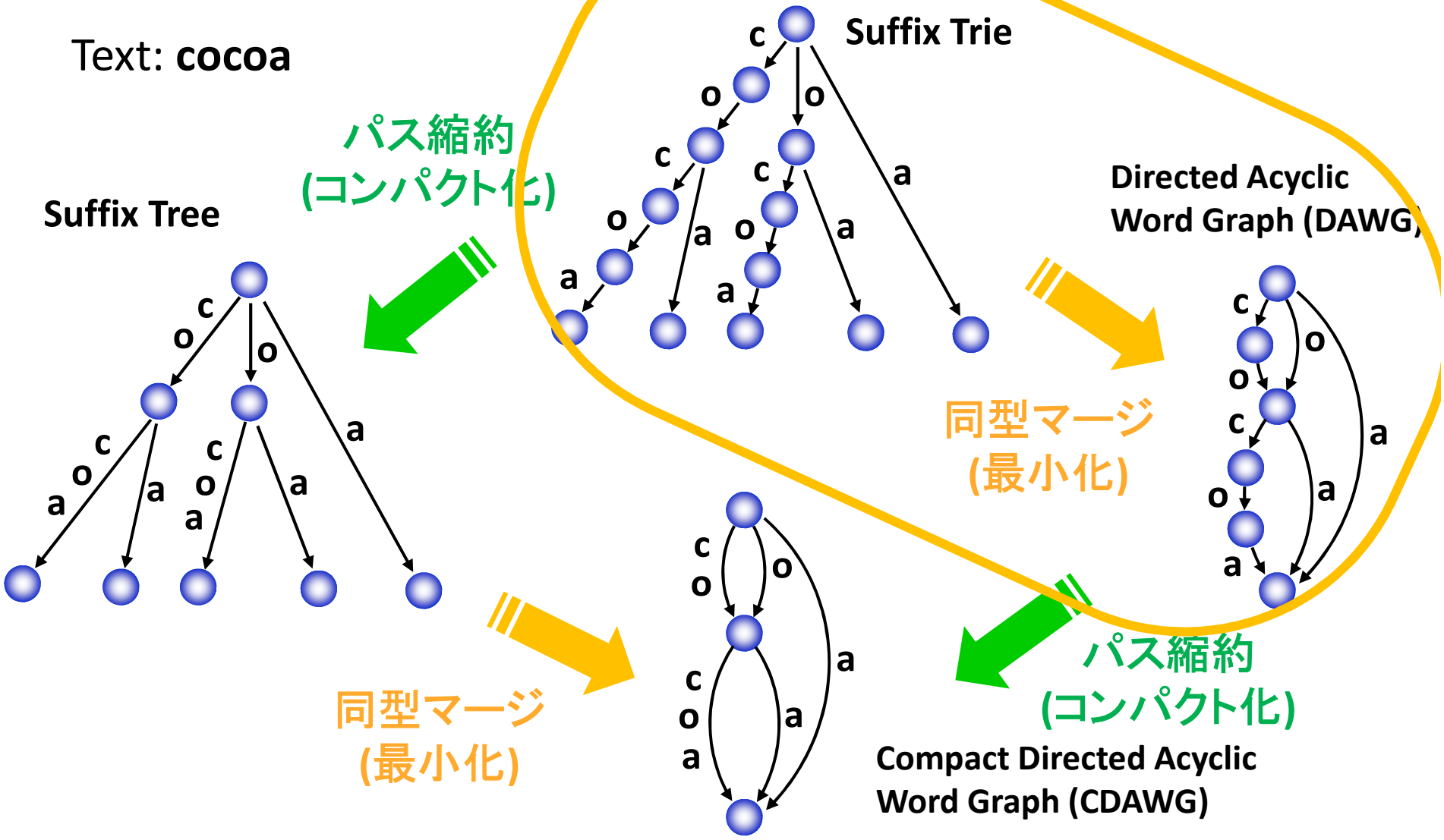
稲永 俊介

基礎的なテキスト索引

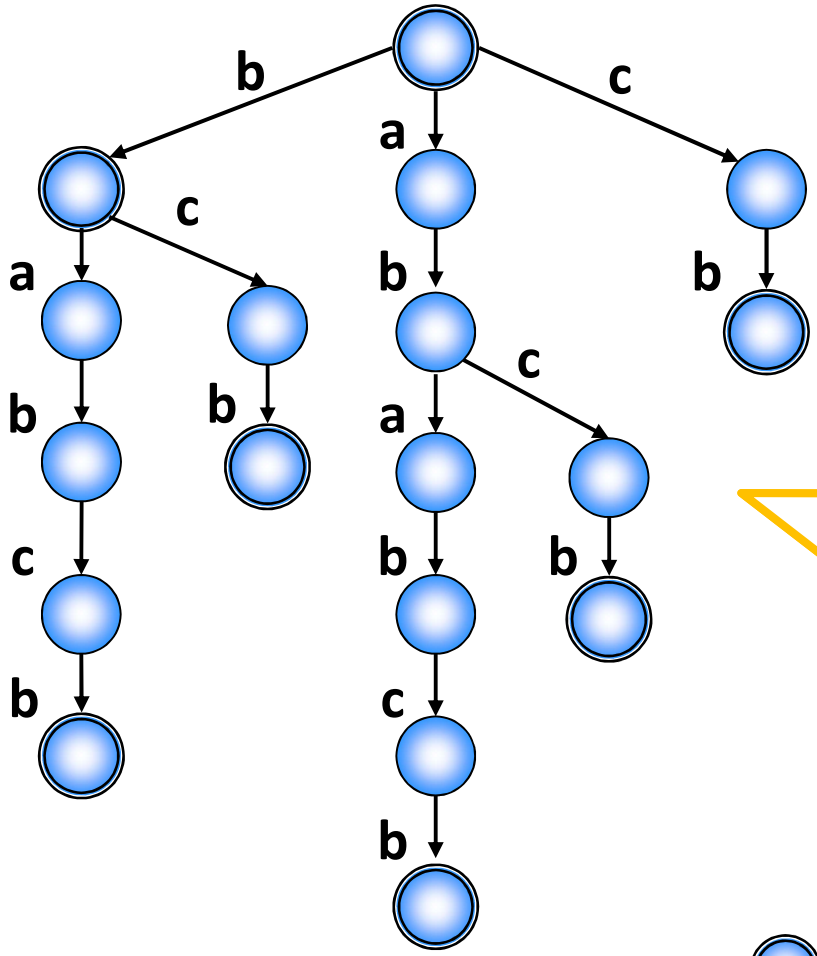

Text: cocoa
Suffix Trie
パス縮約
(コンパクト化)
Suffix Tree
Directed Acyclic
Word Graph (DAWG)
同型マージ
(最小化)
同型マージ
(最小化)
パス縮約
(コンパクト化)
Compact Directed Acyclic
Word Graph (CDAWG)

基礎的なテキスト索引

Text: **cocoa**



From Suffix Trie to DAWG (最小化)

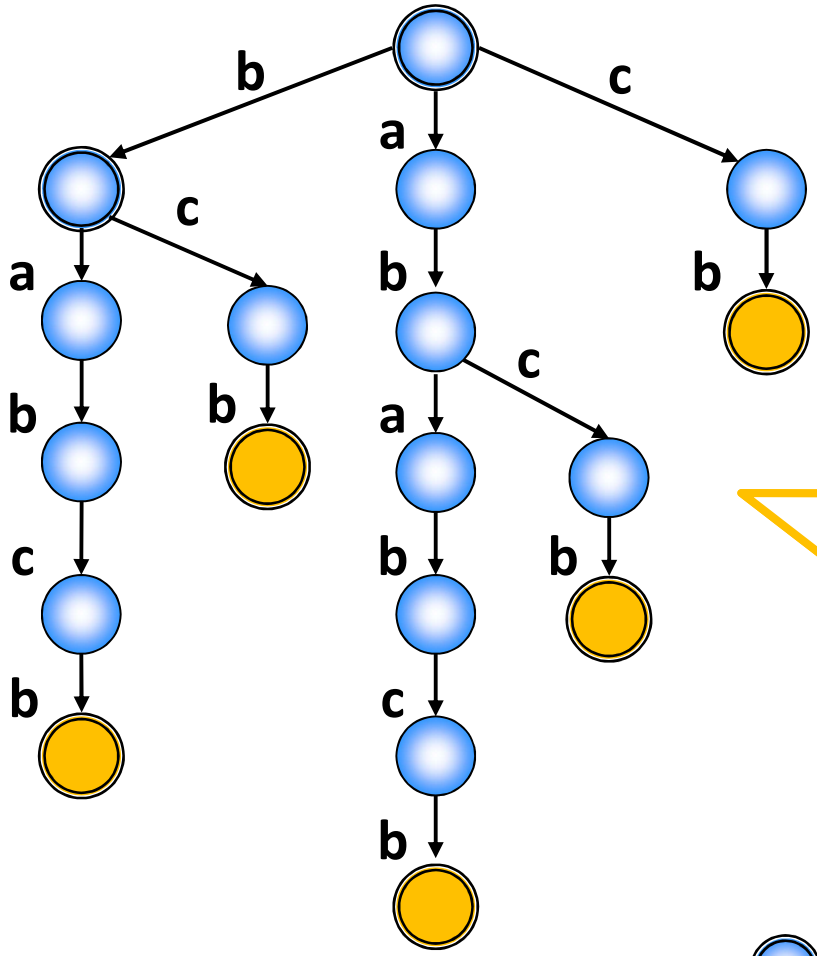


ababcb

同型な木(グラフ)を
ボトムアップに
マージしていく

接尾辞を受理するノード

From Suffix Trie to DAWG (最小化)

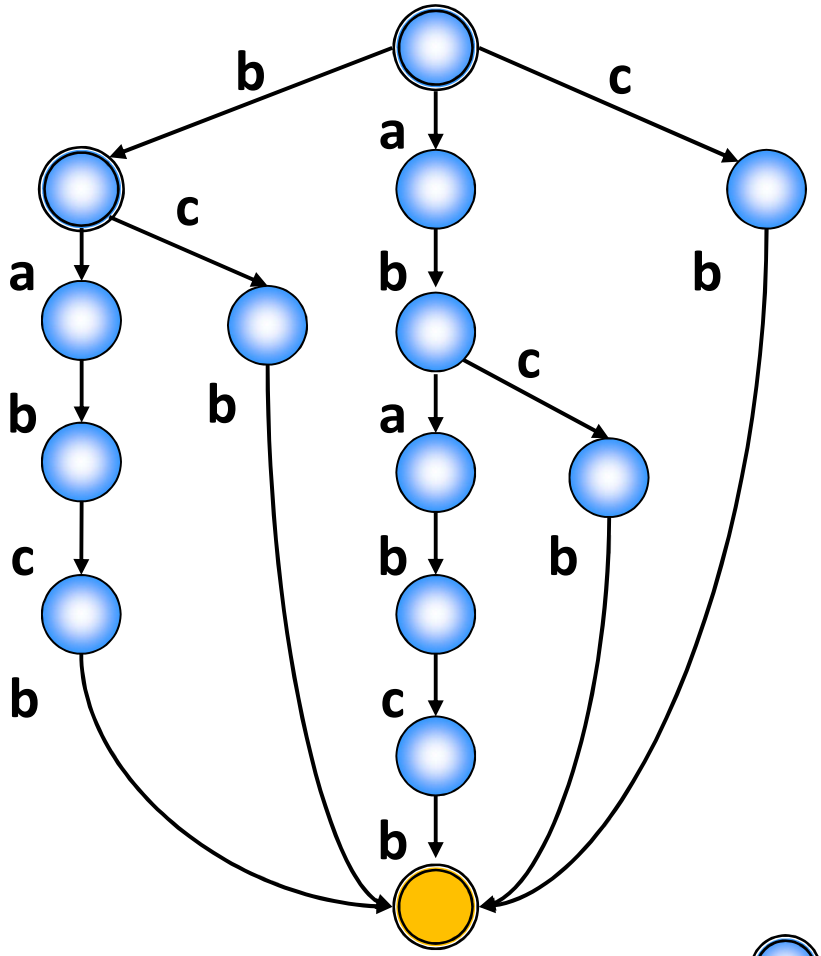


ababcb

同型な木(グラフ)を
ボトムアップに
マージしていく

接尾辞を受理するノード

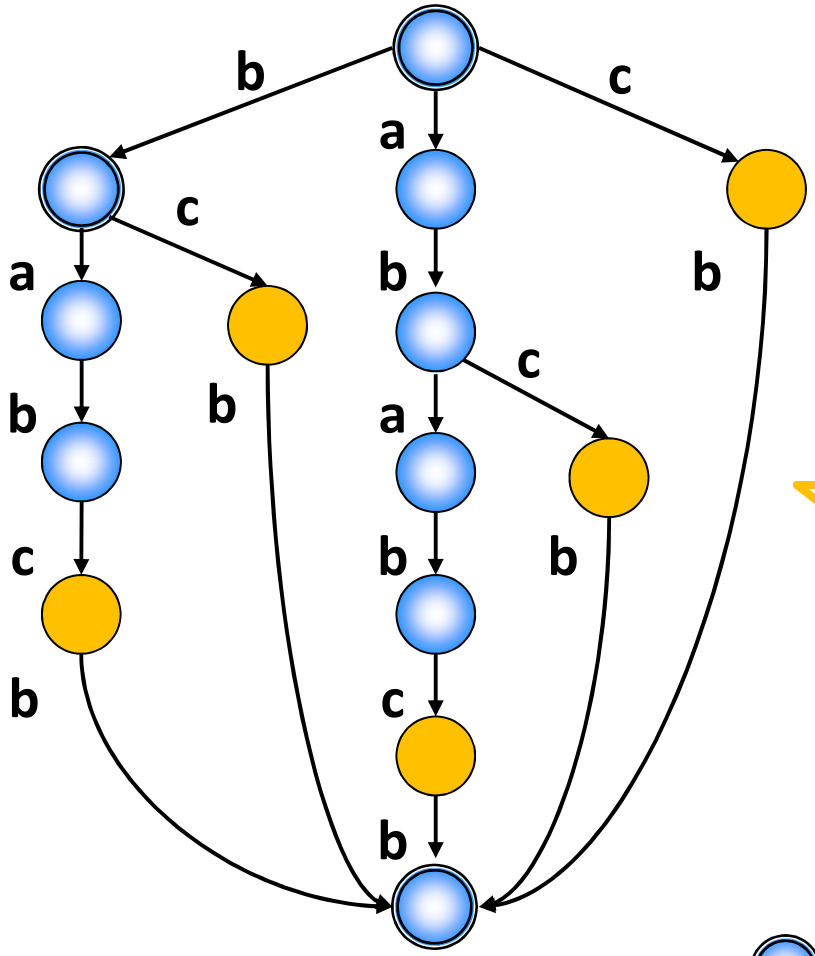
From Suffix Trie to DAWG (最小化)



ababcb

接尾辞を受理するノード

From Suffix Trie to DAWG (最小化)

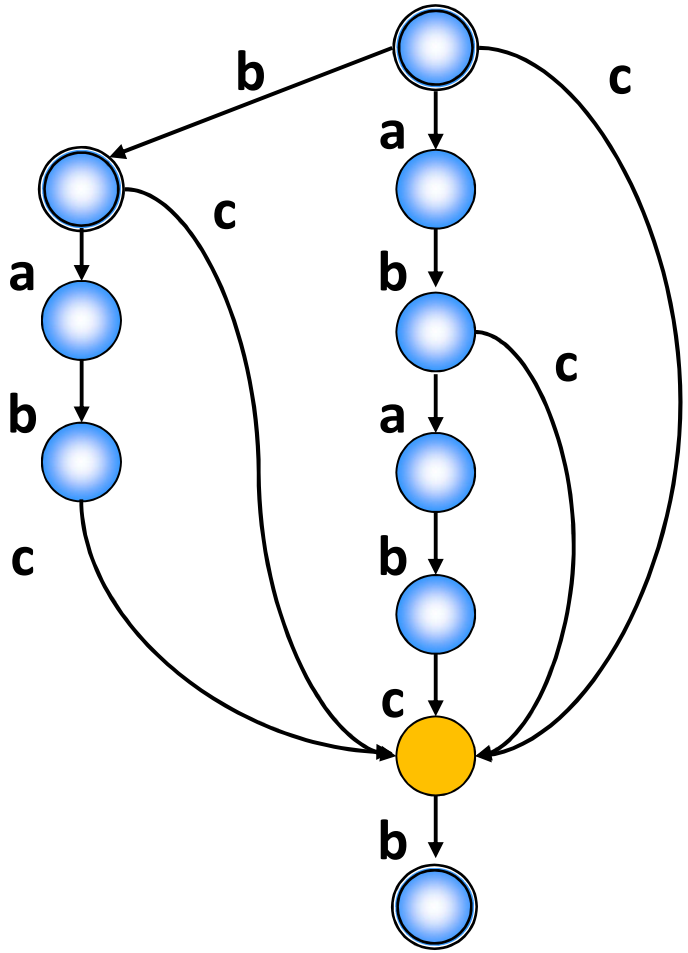


ababcb

同型な木(グラフ)を
ボトムアップに
マージしていく

接尾辞を受理するノード

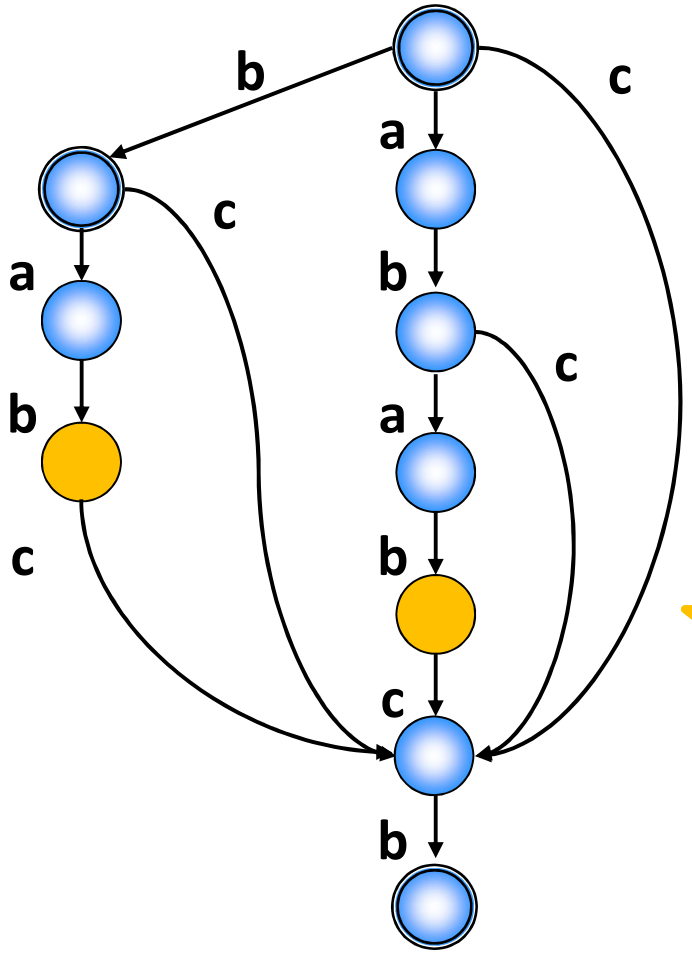
From Suffix Trie to DAWG (最小化)



ababcb

接尾辞を受理するノード

From Suffix Trie to DAWG (最小化)

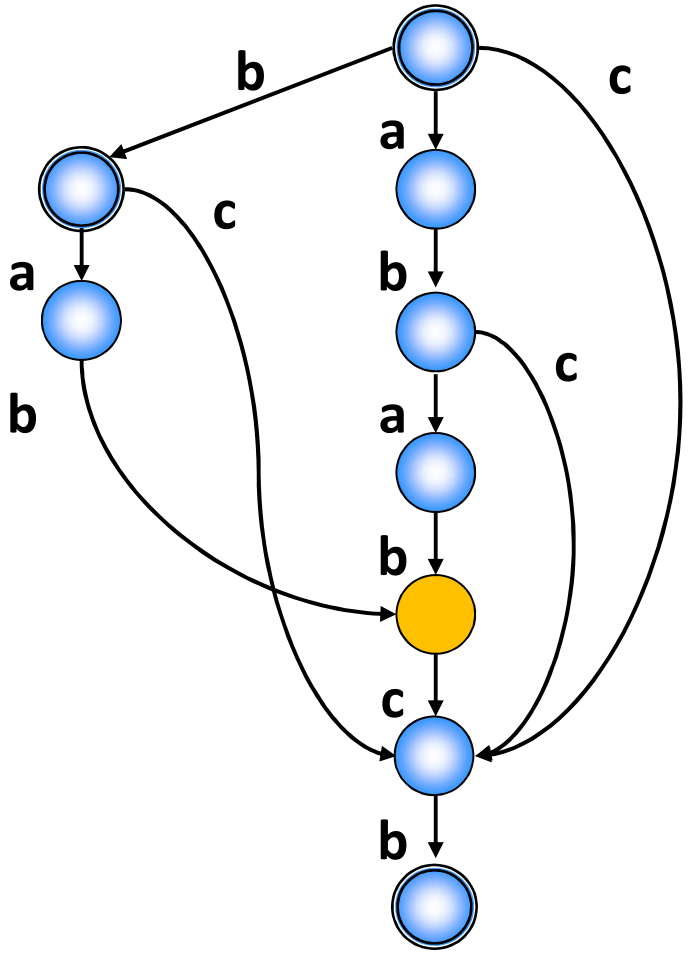


ababcb

同型な木（グラフ）を
ボトムアップに
マージしていく

接尾辞を受理するノード

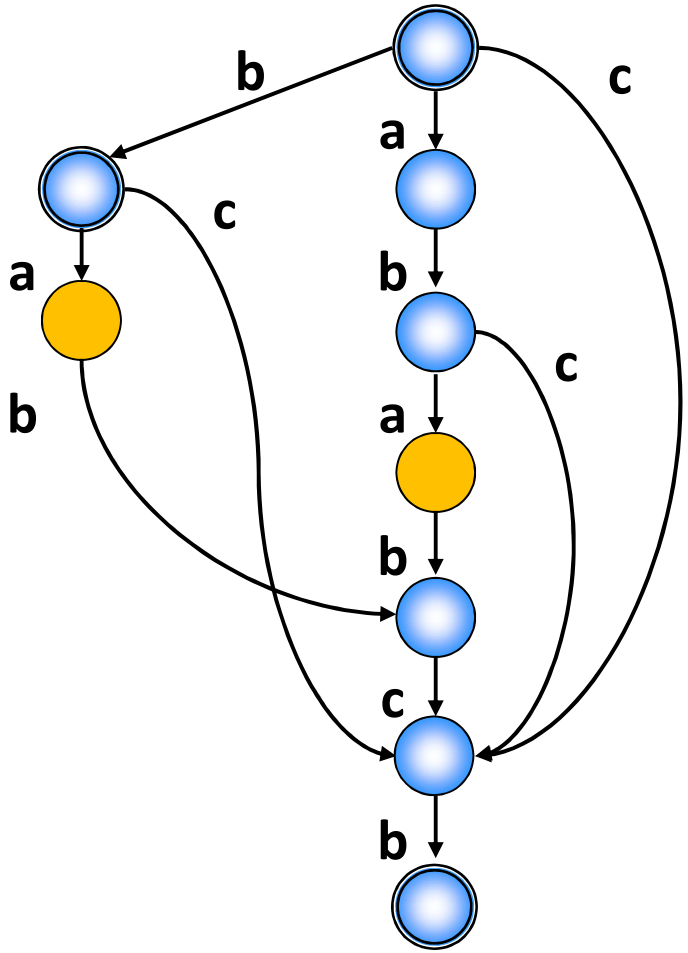
From Suffix Trie to DAWG (最小化)



ababcb

接尾辞を受理するノード

From Suffix Trie to DAWG (最小化)

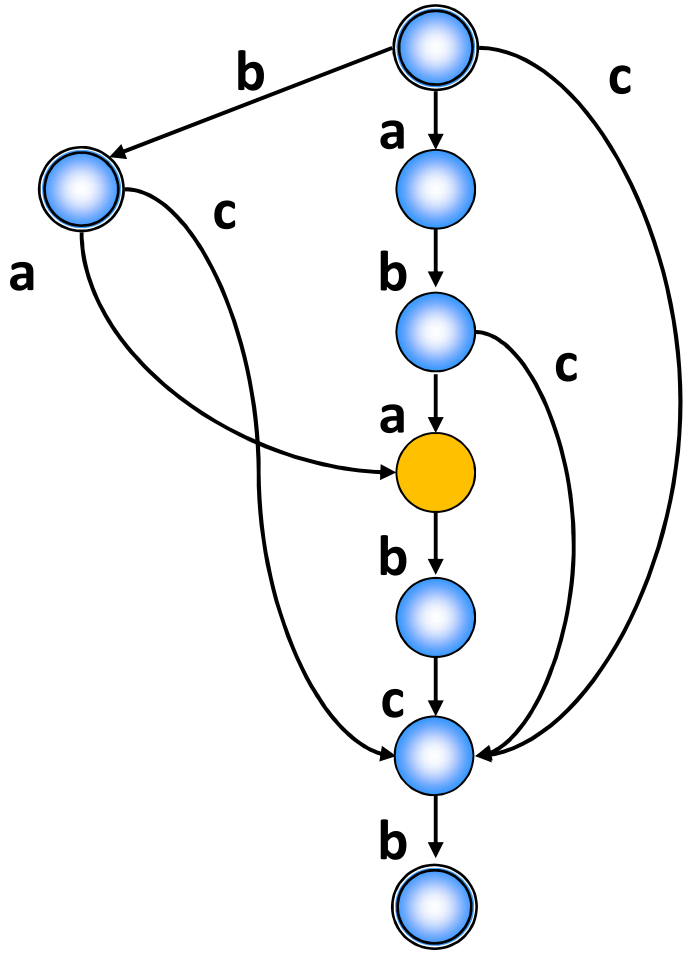


ababcb

同型な木(グラフ)を
ボトムアップに
マージしていく

接尾辞を受理するノード

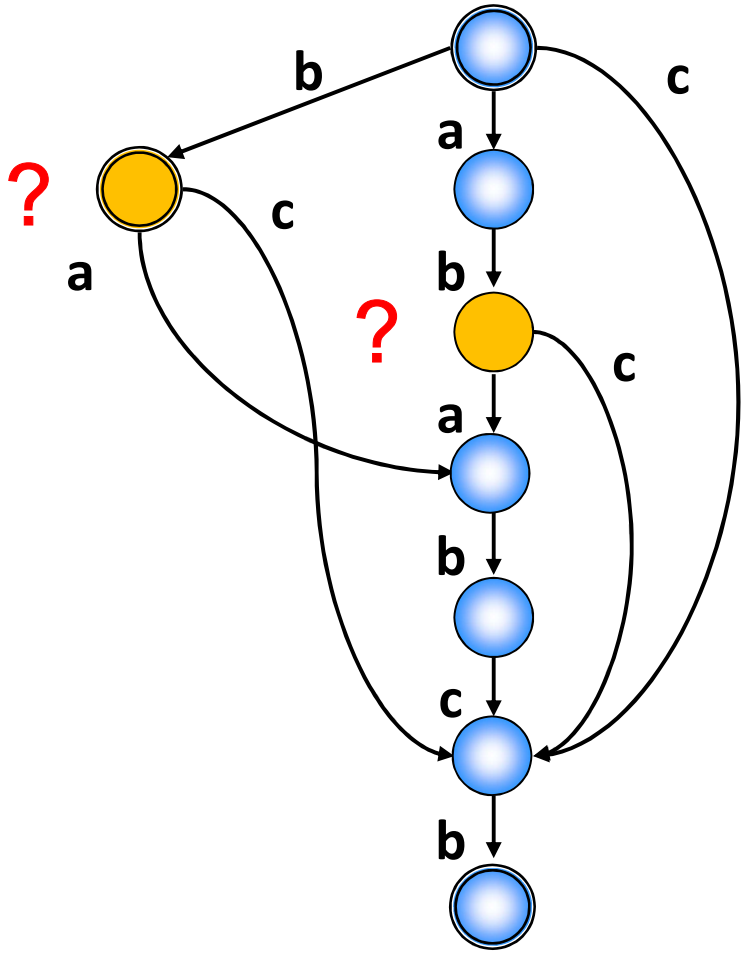
From Suffix Trie to DAWG (最小化)



ababcb

接尾辞を受理するノード

From Suffix Trie to DAWG (最小化)

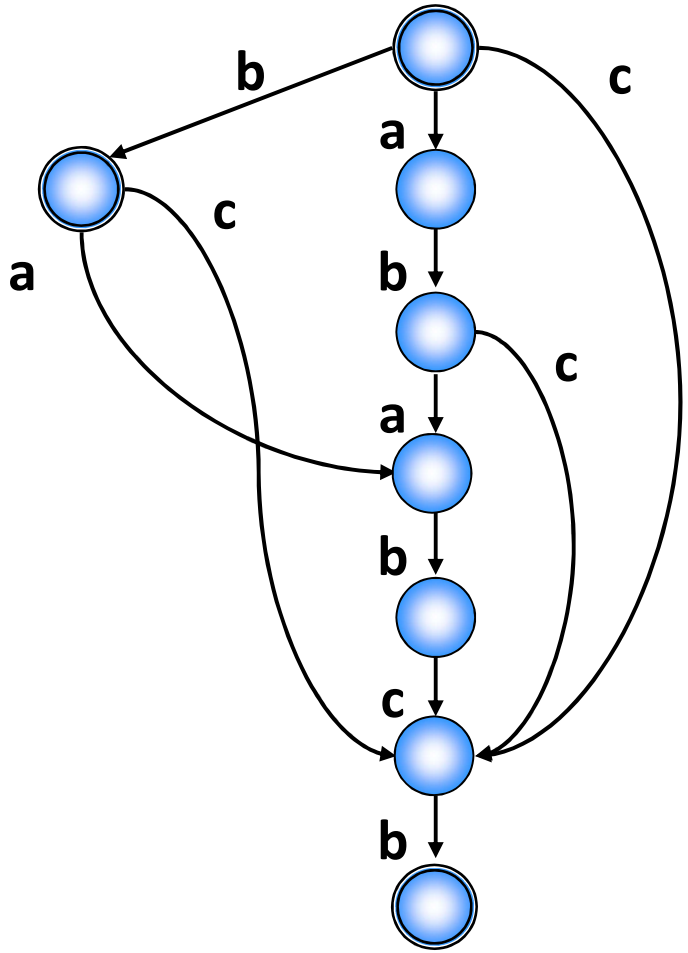


ababcb

同型に見えるが，
一方は受理ノードで
他方は受理ノードでは
ないので，マージしない

接尾辞を受理するノード

From Suffix Trie to DAWG (最小化)



ababcb

DAWG of string
ababcb

接尾辞を受理するノード

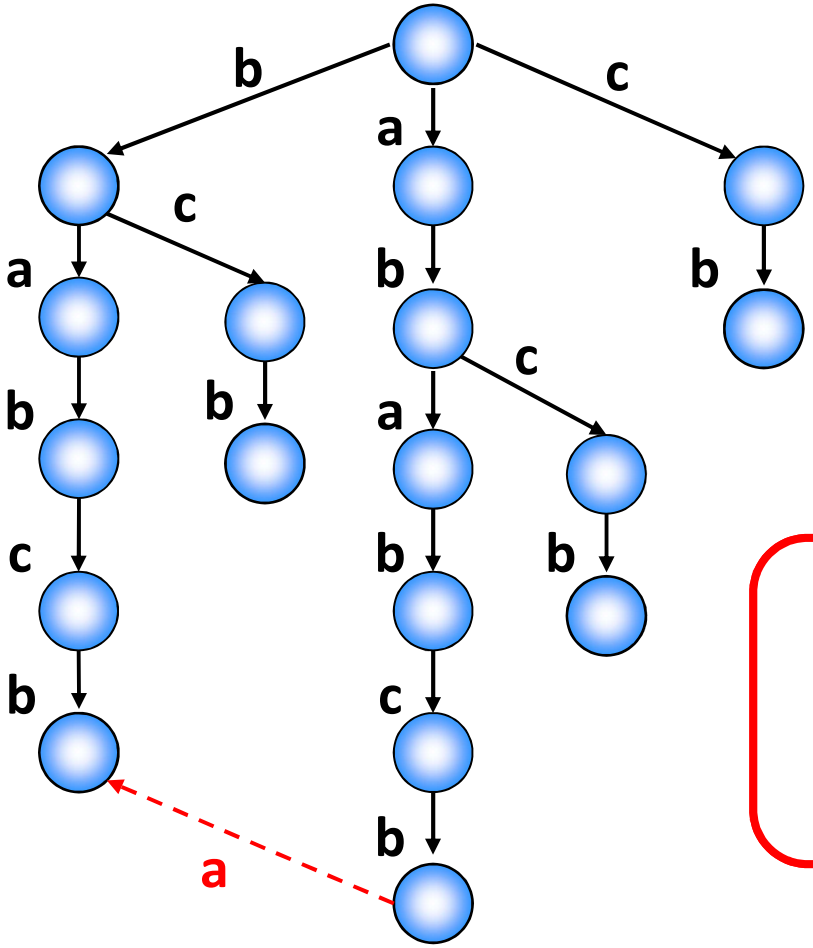
DAWG の最小性

定理 [Blumer et al. 1985]

文字列 w の DAWG は w の接尾辞を
受理する最小のオートマトンである.

- 明らかに suffix trie は w の接尾辞を受理する.
  よって DAWG も接尾辞を受理する.
- マージできるところはすべてマージしたので,
  DAWG が w の接尾辞を受理する最小の
  オートマトンである.

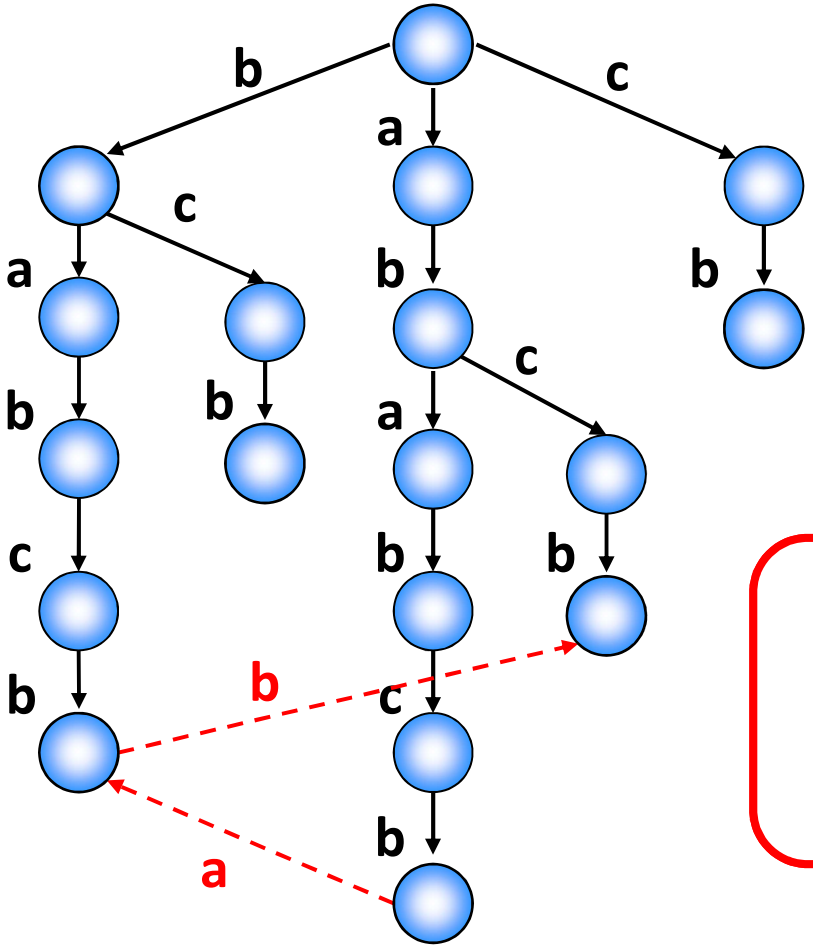
Suffix Links of Suffix Trie



ノード ax の suffix link の
文字ラベルは a で,
その行き先は x である.

$a \in \Sigma,\ x \in \Sigma^*$

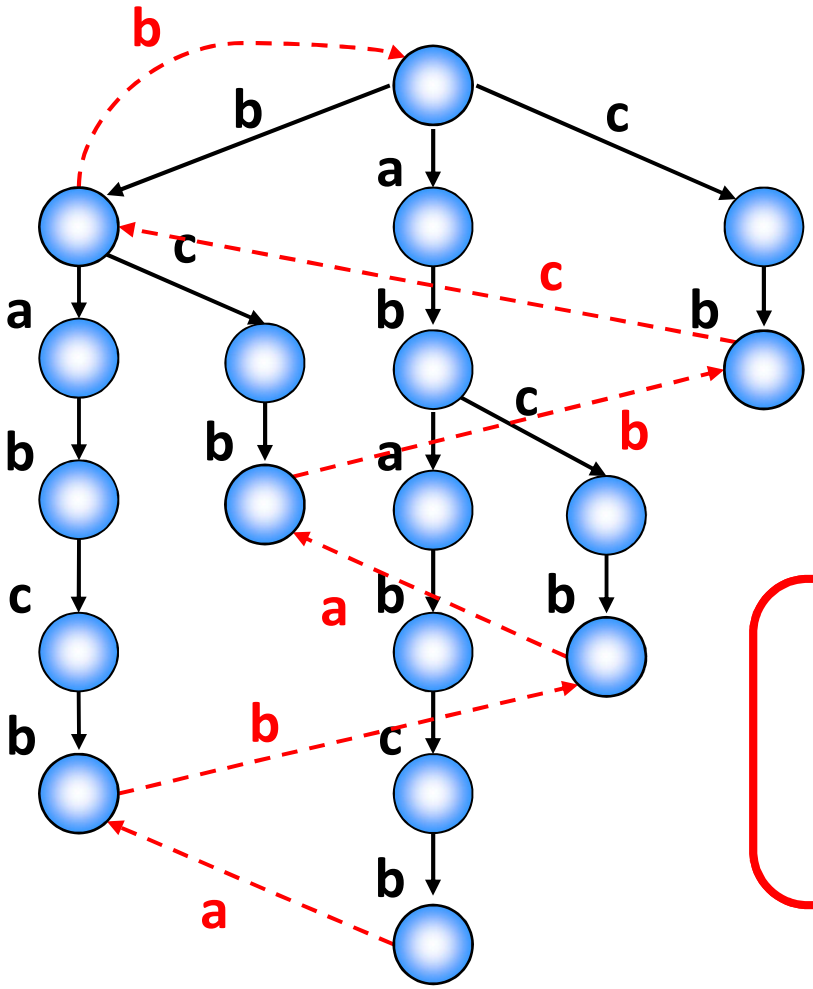
Suffix Links of Suffix Trie



ノード ax の suffix link の
文字ラベルは a で,
その行き先は x である.

$a \in \Sigma,\ x \in \Sigma^*$

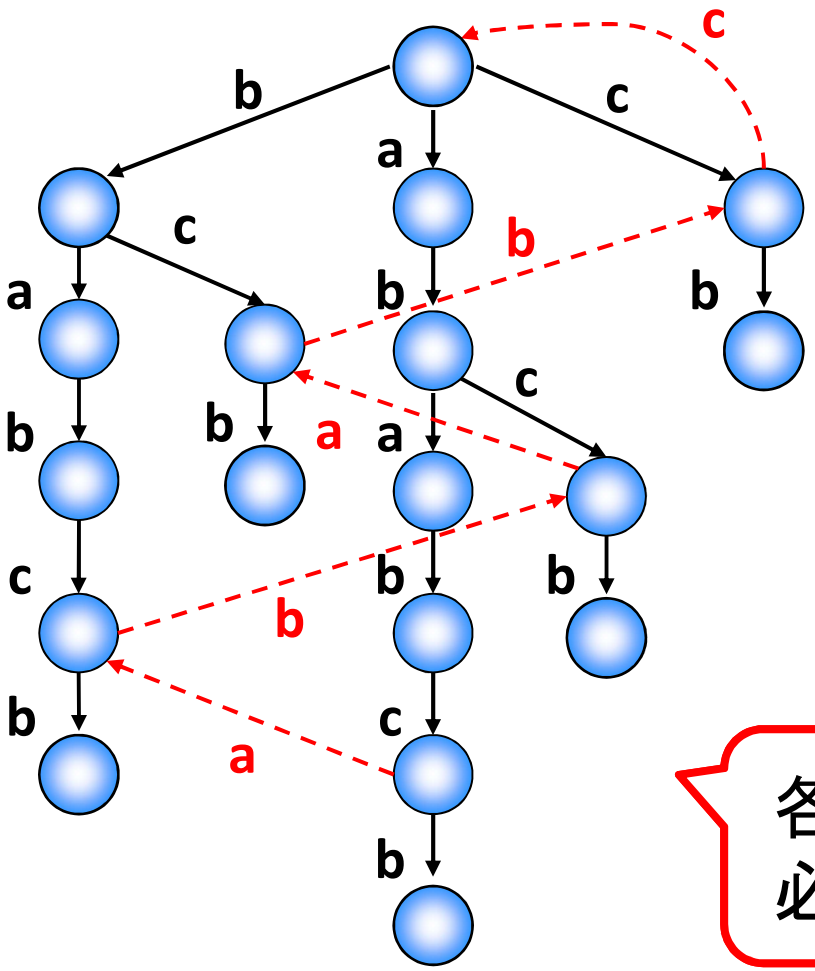
Suffix Links of Suffix Trie



ノード ax の suffix link の
文字ラベルは a で,
その行き先は x である.

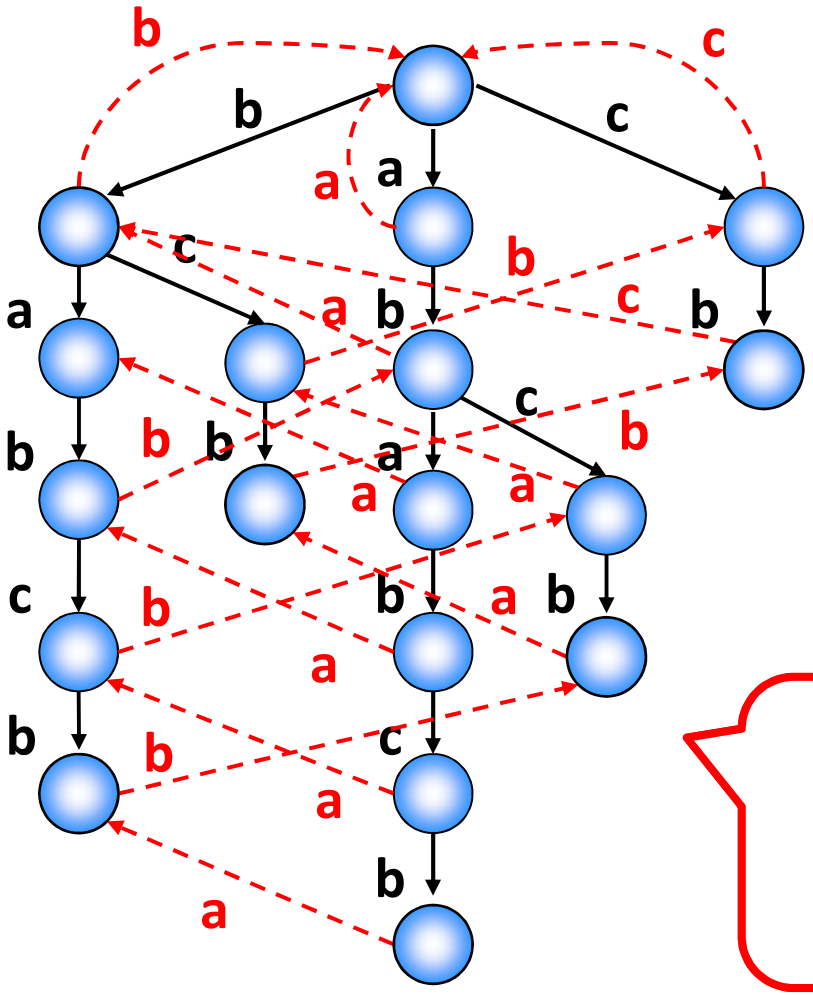
$a \in \Sigma,\ x \in \Sigma^*$

Suffix Links of Suffix Trie



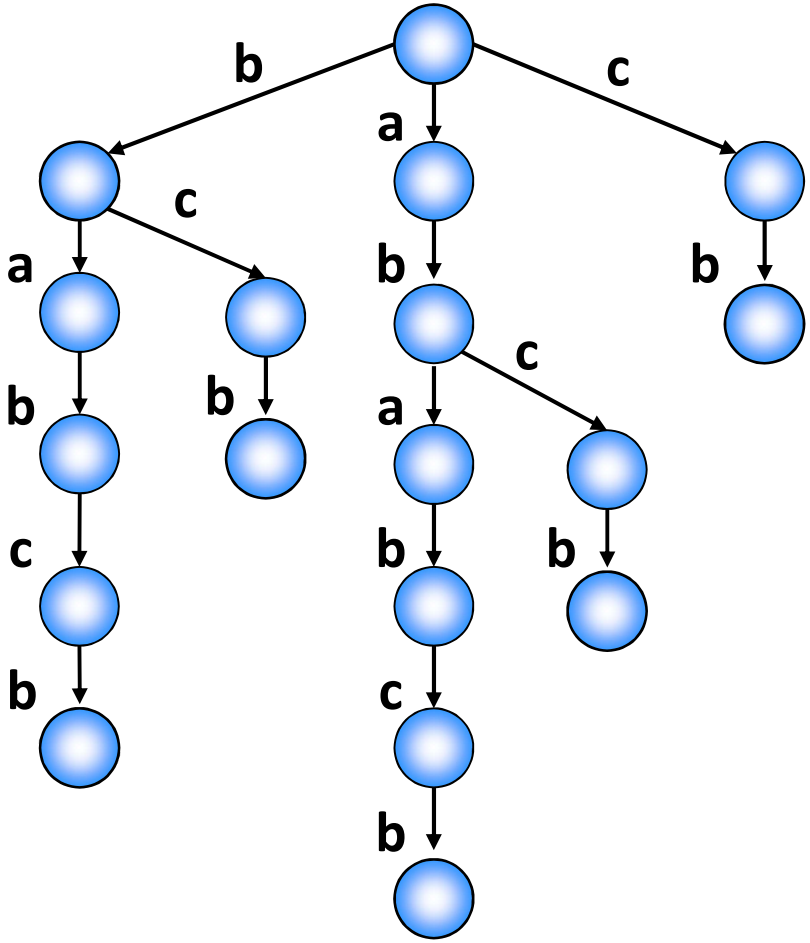
各 suffix link パスは
必ず根に到達する.

Suffix Link Tree of Suffix Trie

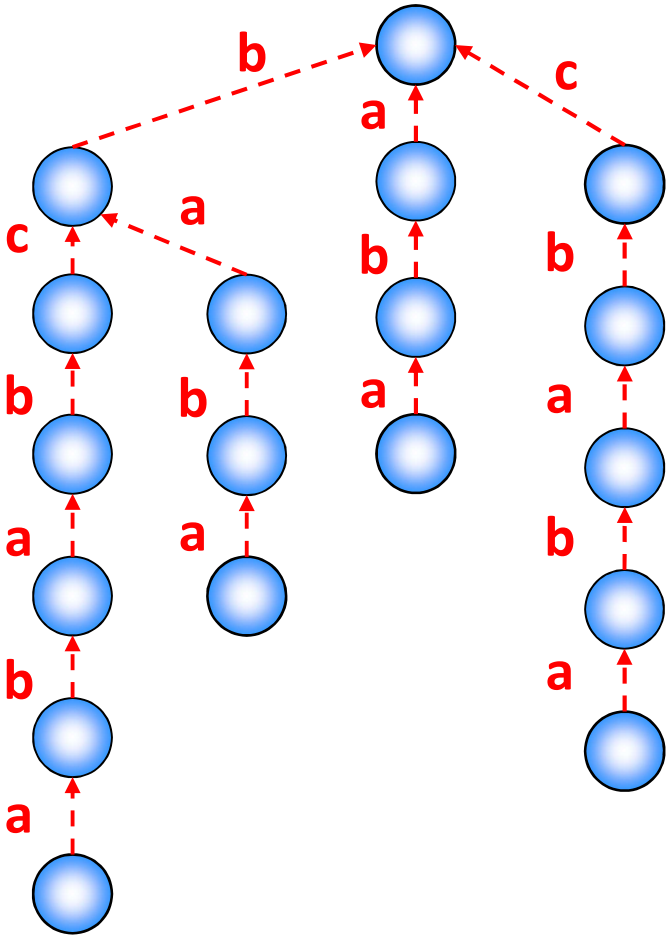


よって suffix link は
(辺が逆向きの)木を成す
(suffix link tree).

Suffix Link Tree of Suffix Trie

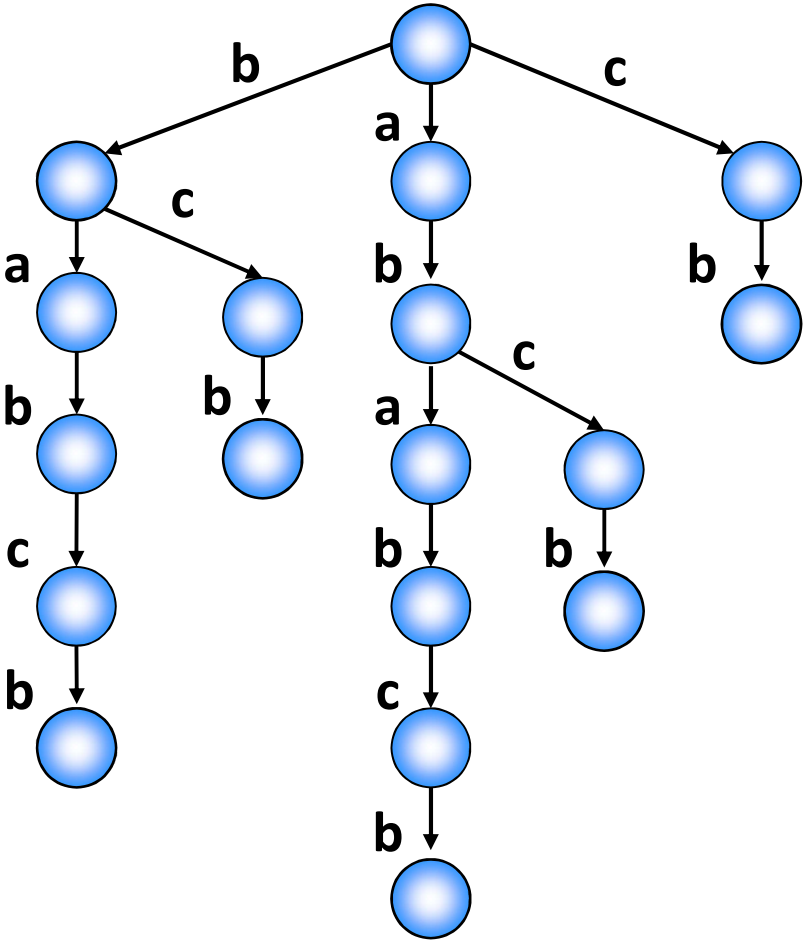


Suffix Trie of **ababcb**

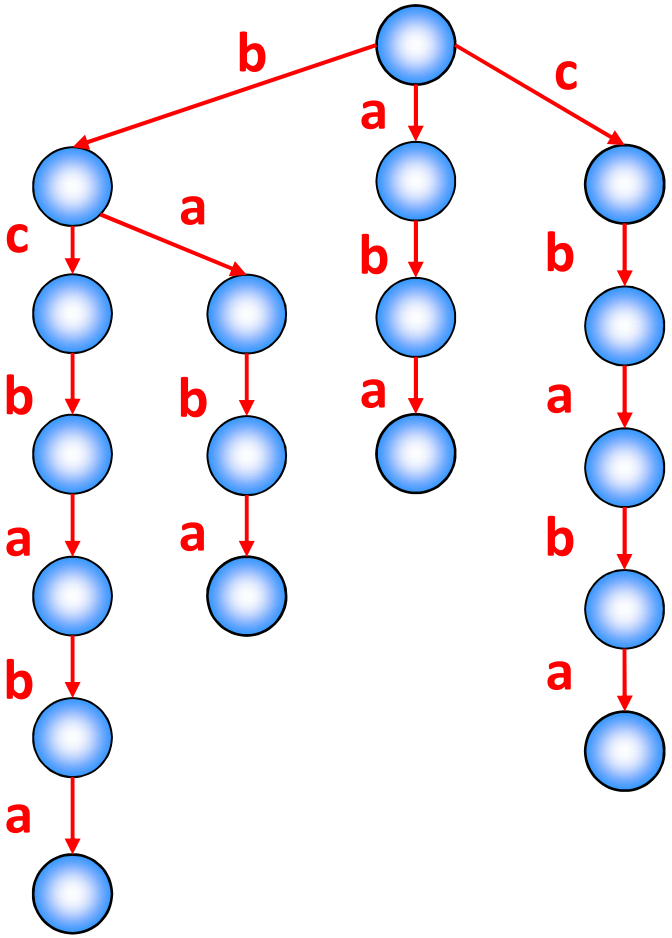


Suffix Link Tree of **ababcb**

Suffix Link Tree = 反転文字列の Suffix Trie



Suffix Trie of **ababcb**



Suffix Trie of **bcbaba**

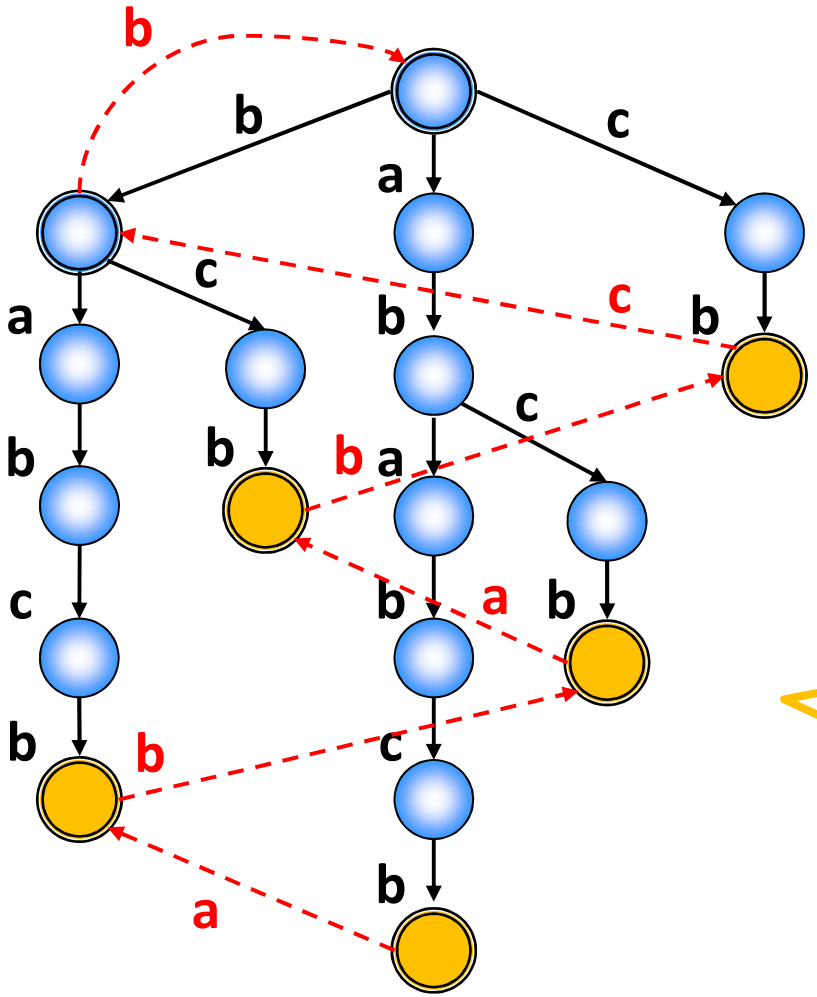
Suffix Link Tree = 反転文字列の Suffix Trie

事実

文字列 w の suffix trie の suffix link tree は
反転文字列 w^R の suffix trie に等しい.

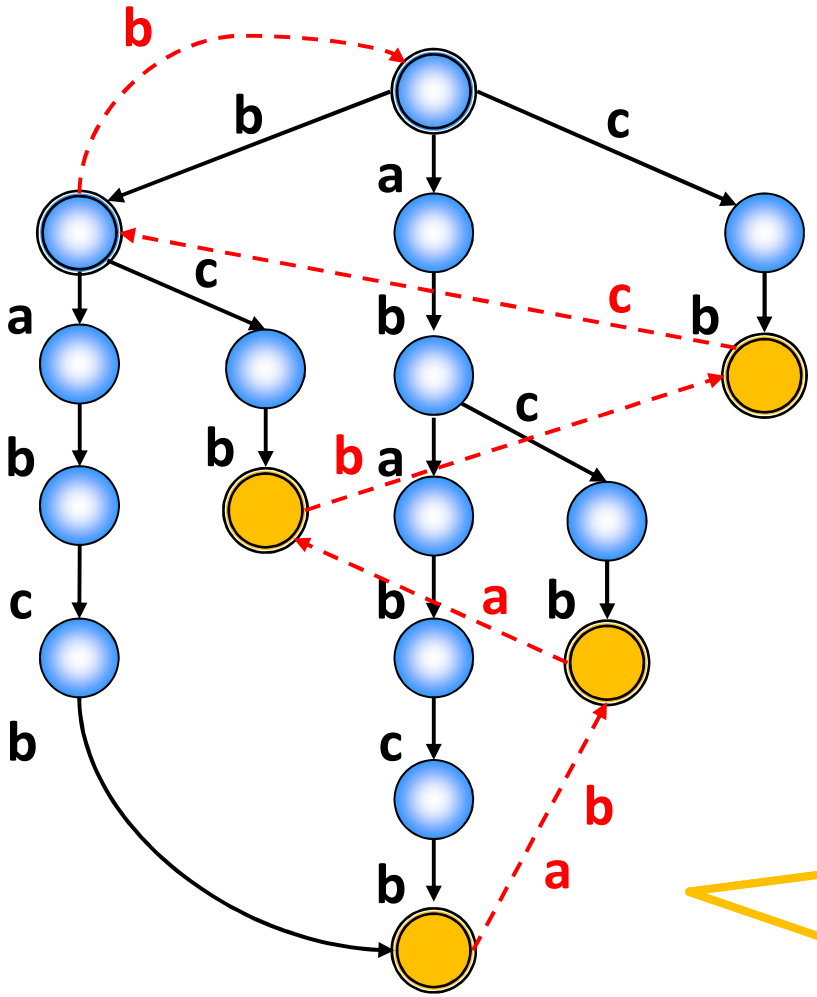
- 文字列 x を表す頂点の suffix link パス は
  x の文字を逆順に読みながら根に到達するから.

ノードのマージと Suffix Link の関係



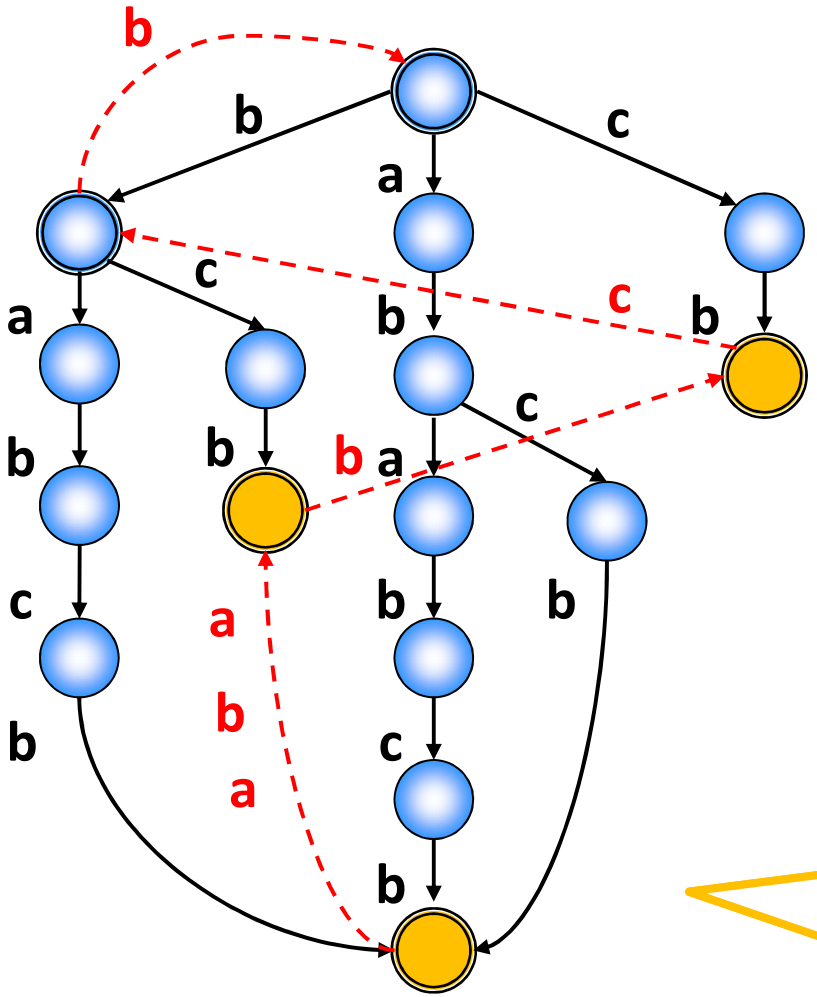
マージされるノードは
suffix link パスで
繋がっている

ノードのマージと Suffix Link の関係



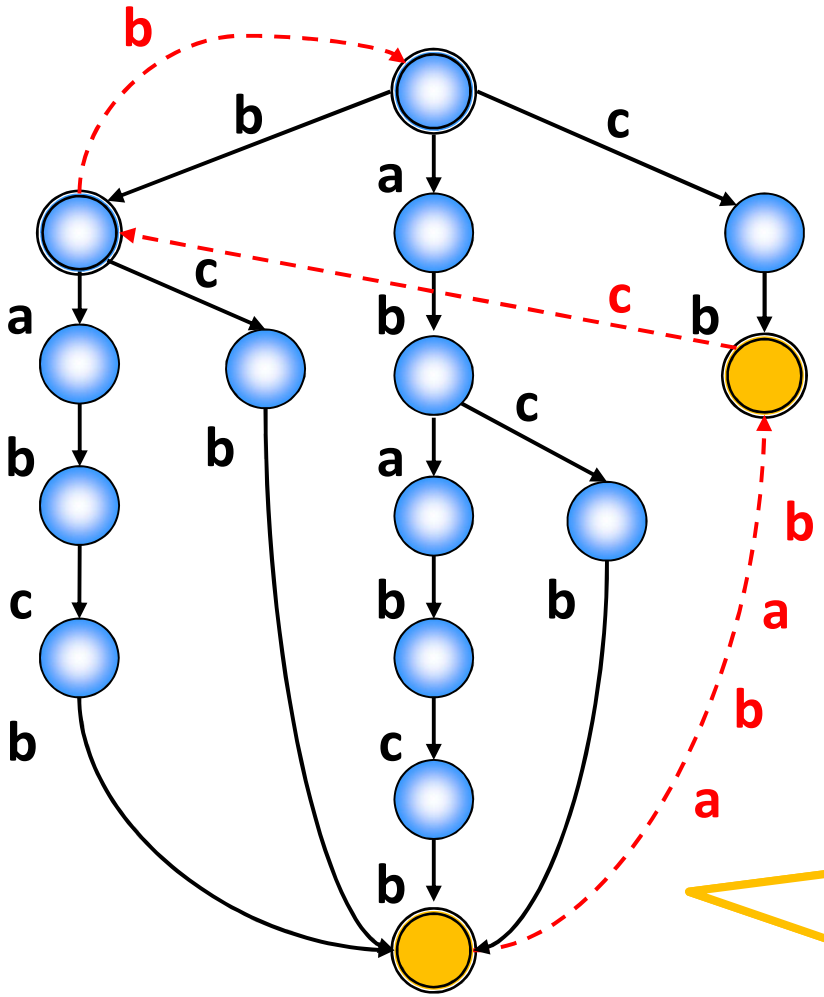
ノードをマージするとき，
suffix link を同時に
パス縮約する

ノードのマージと Suffix Link の関係



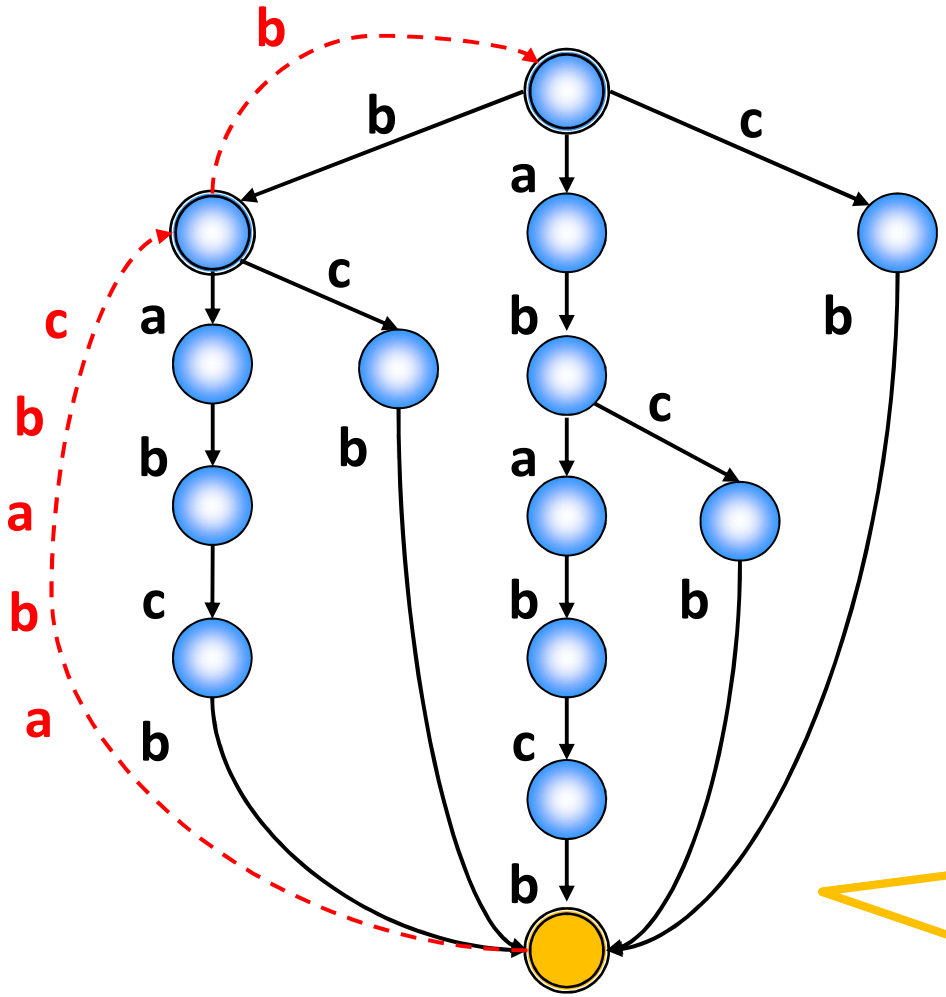
ノードをマージするとき，
suffix link を同時に
パス縮約する

ノードのマージと Suffix Link の関係



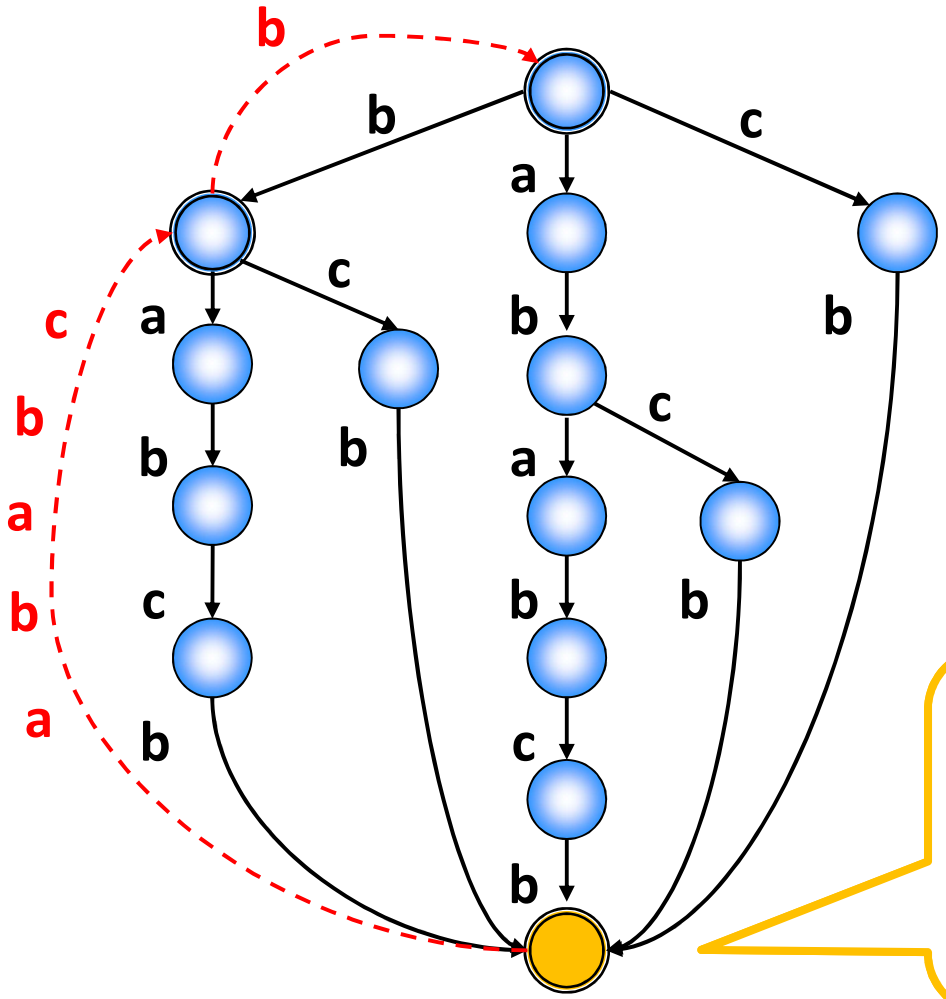
ノードをマージするとき，
suffix link を同時に
パス縮約する

ノードのマージと Suffix Link の関係



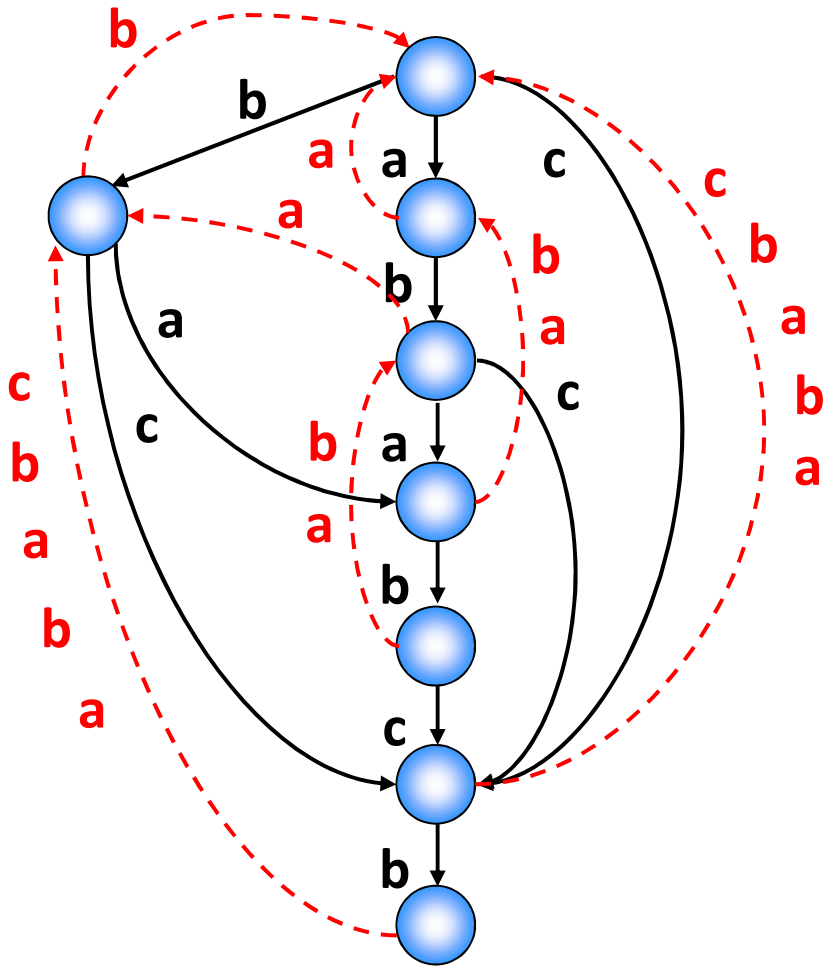
ノードをマージするとき，
suffix link を同時に
パス縮約する

ノードのマージと Suffix Link の関係



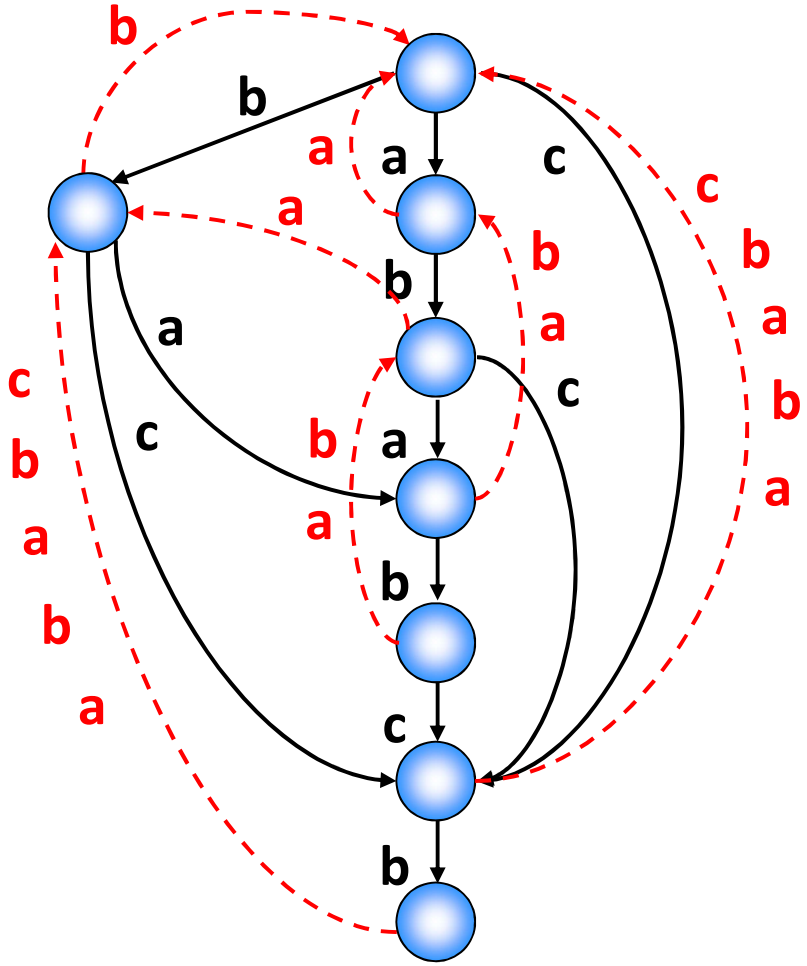
縮約した suffix link パスが
DAWG のこの頂点の
suffix link である.

Suffix Links of DAWG

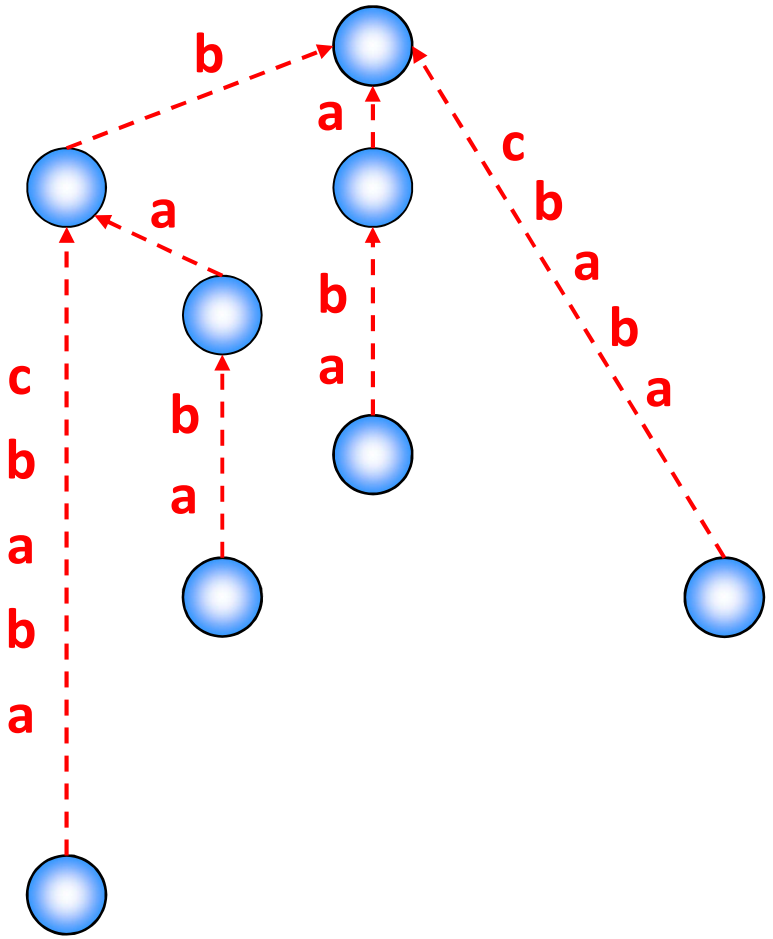


DAWG の suffix link も
また木をなす

Suffix Links of DAWG

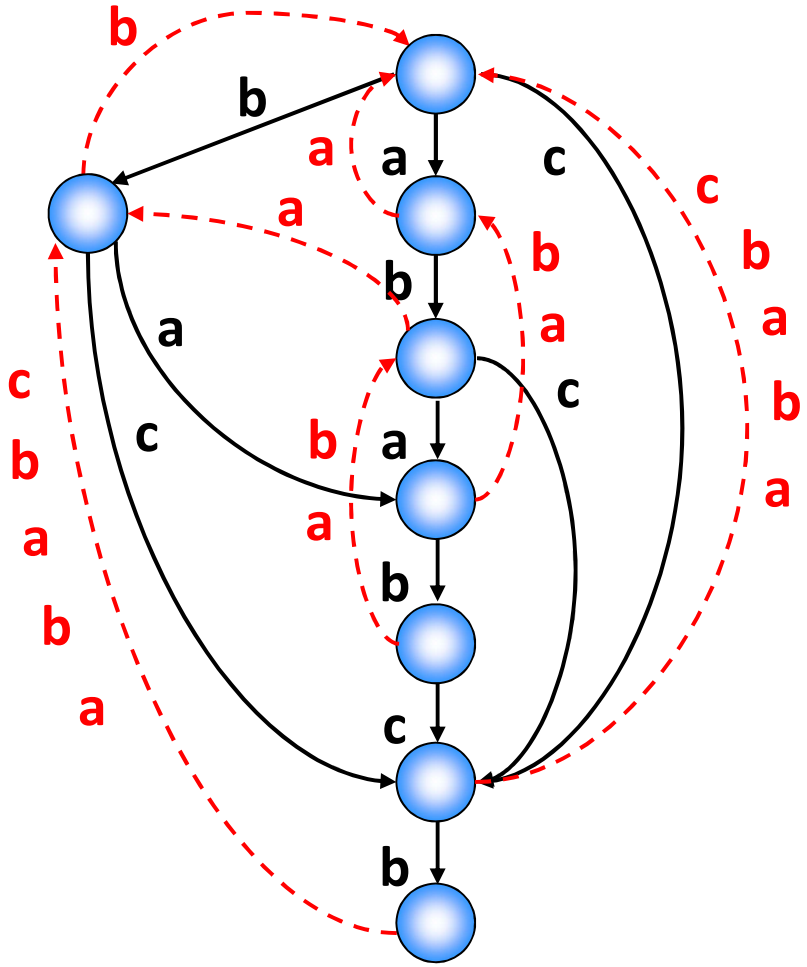


DAWG of **ababcb**

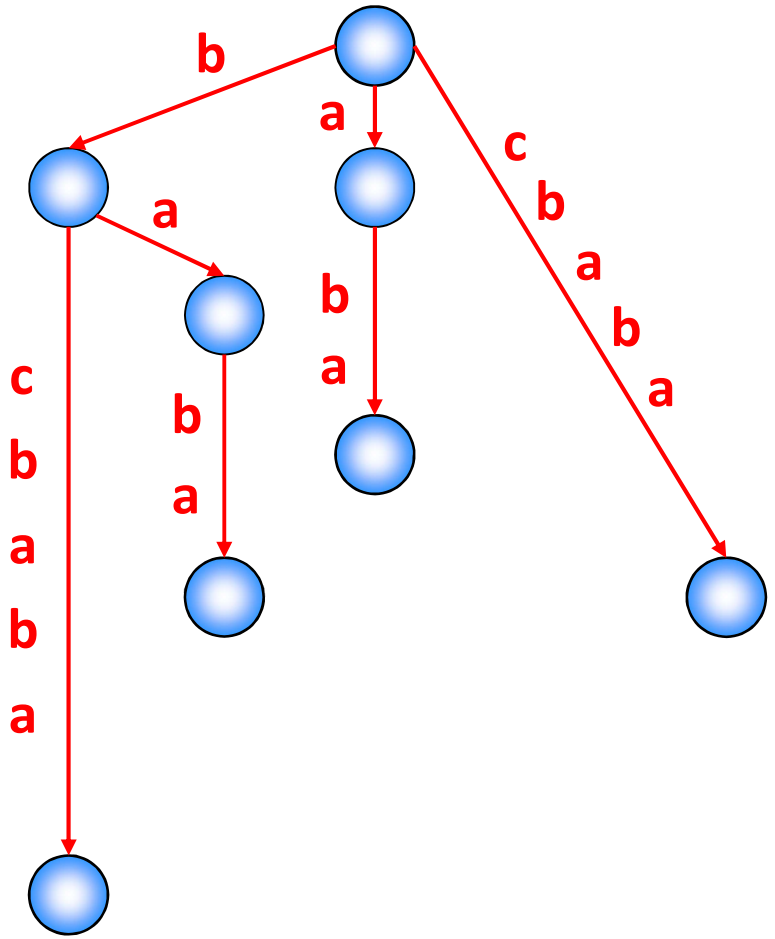


Contracted SLT of **ababcb**

SLT of DAWG = 反転文字列の Suffix Tree

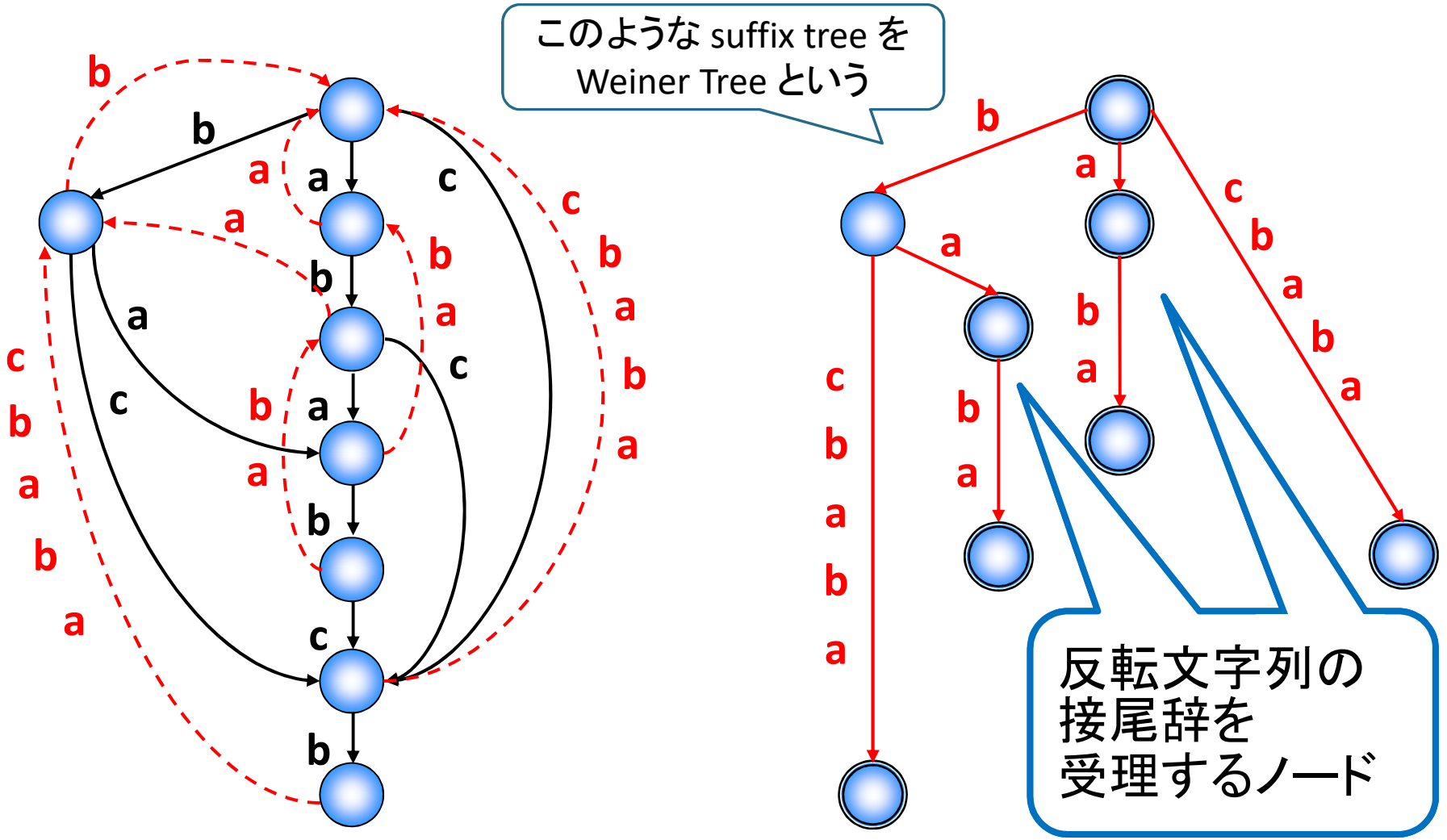


DAWG of **ababcb**



Suffix Tree of **bcbaba**

SLT of DAWG = 反転文字列の Weiner Tree



DAWG of **ababcb**

Suffix Tree of **bcbaba**

SLT of DAWG = 反転文字列の Suffix Tree

定理 [Blumer et al. 1985]

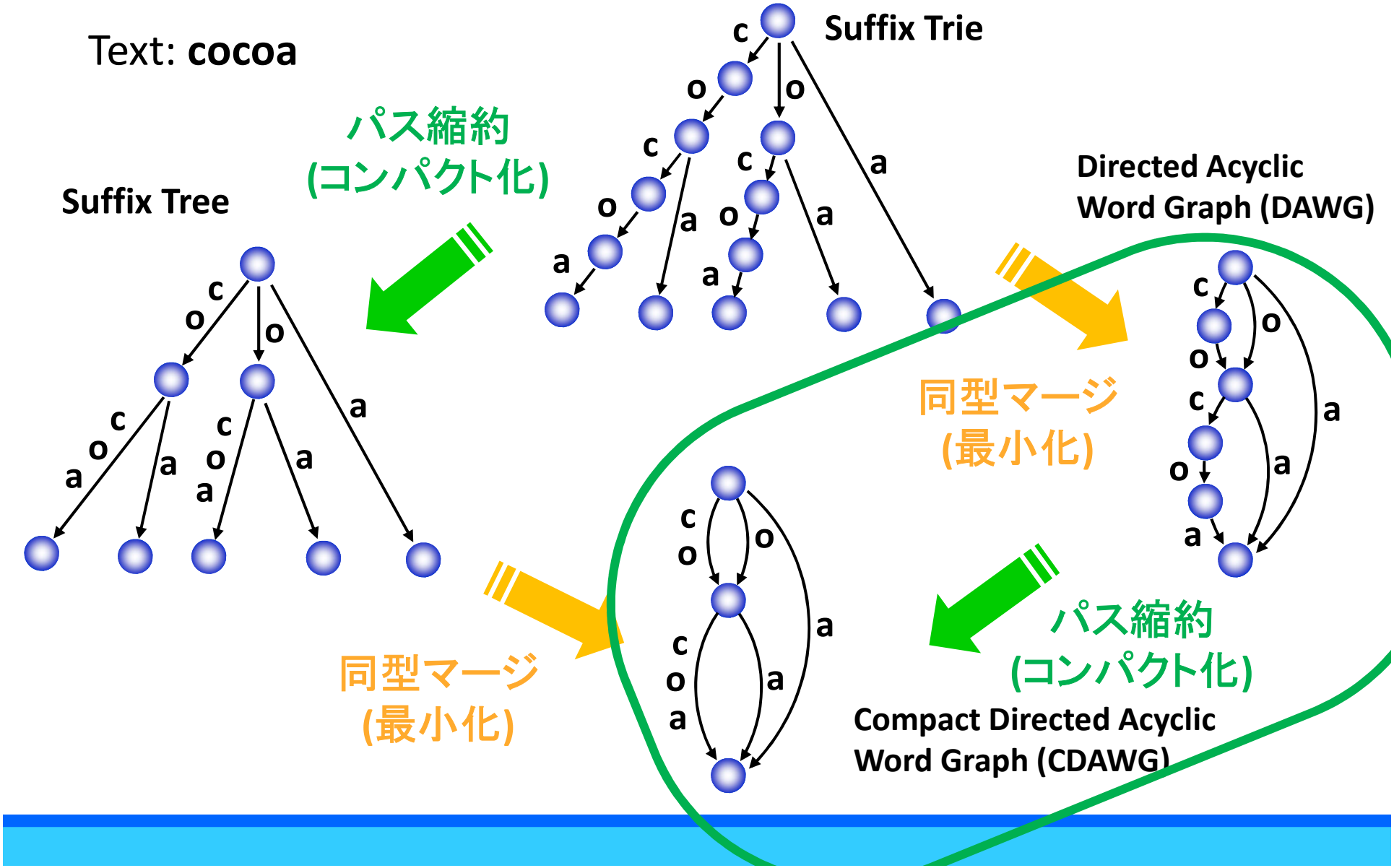
文字列 w の DAWG の suffix link は
反転文字列 w^R の Suffix Tree に等しい.

(1) w の suffix trie のノードをマージしてオートマトンを
最小化しながら，同時に suffix link をパス縮約

(2) w^R の suffix trie の辺をパス縮約

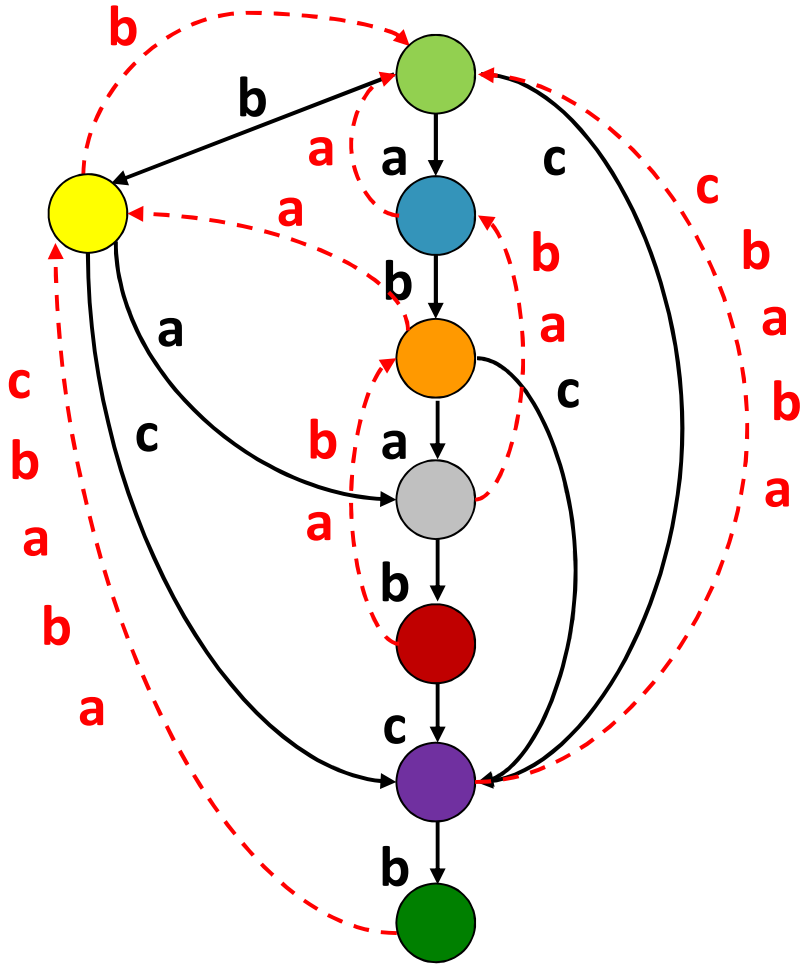
(1) は表の世界（順向き文字列 w）を主体にした操作.
(2) は裏の世界（逆向き文字列 w^R）を主体にした操作.
これらの操作は表裏一体の「同じ作業」である.

基礎的なテキスト索引

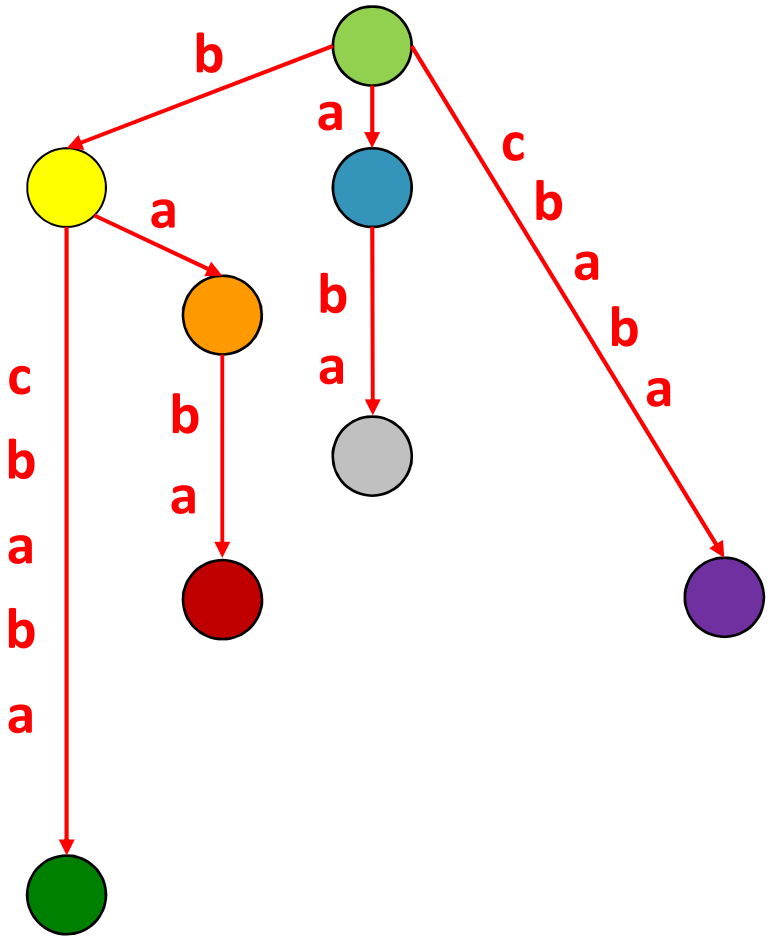

Text: cocoa
Suffix Trie
パス縮約
(コンパクト化)
Suffix Tree
Directed Acyclic
Word Graph (DAWG)
同型マージ
(最小化)
同型マージ
(最小化)
パス縮約
(コンパクト化)
Compact Directed Acyclic
Word Graph (CDAWG)

From DAWG to CDAWG

わかりやすさのために表と裏の世界の
対応する頂点を同じ色で塗っている

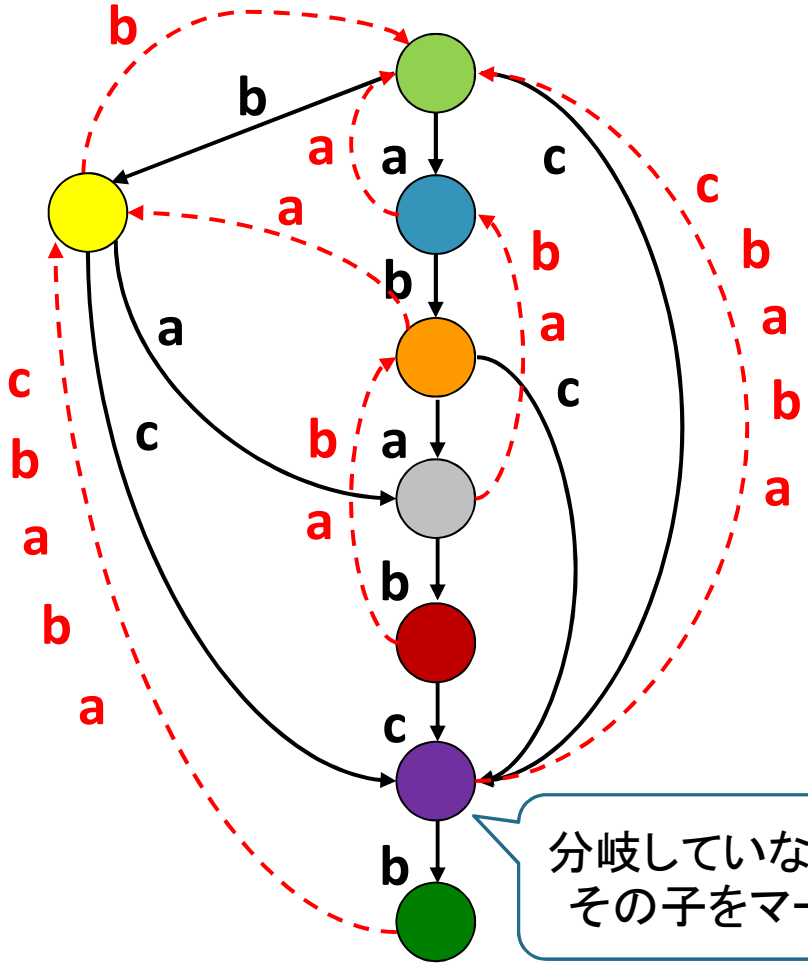


DAWG of **ababcb**

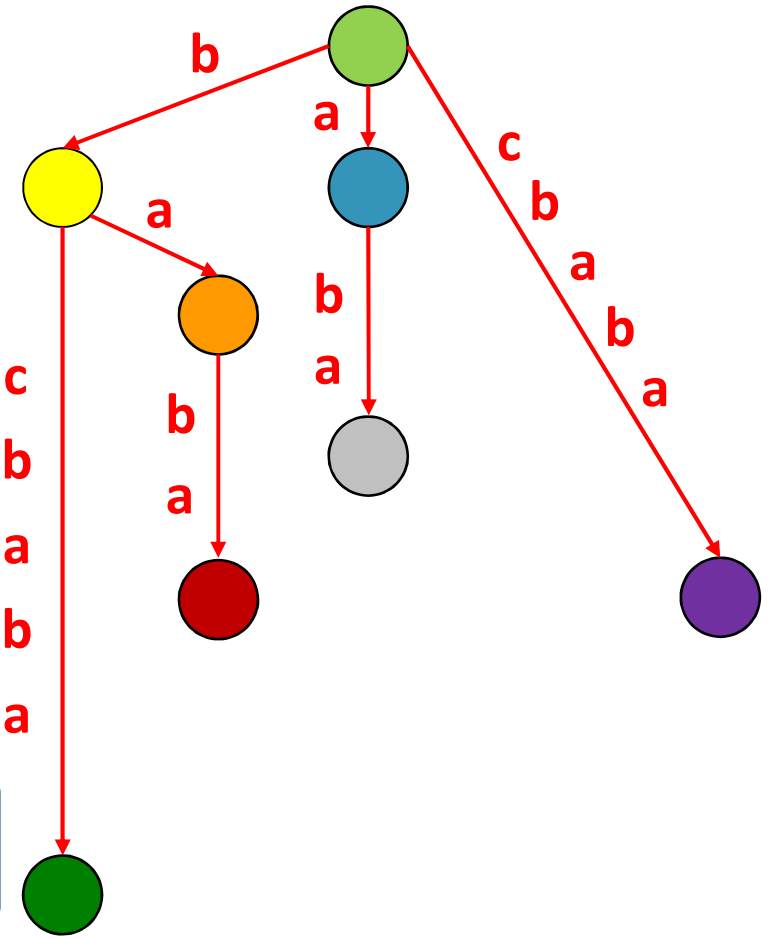


Suffix Tree of **bcbaba**

From DAWG to CDAWG (パス縮約)

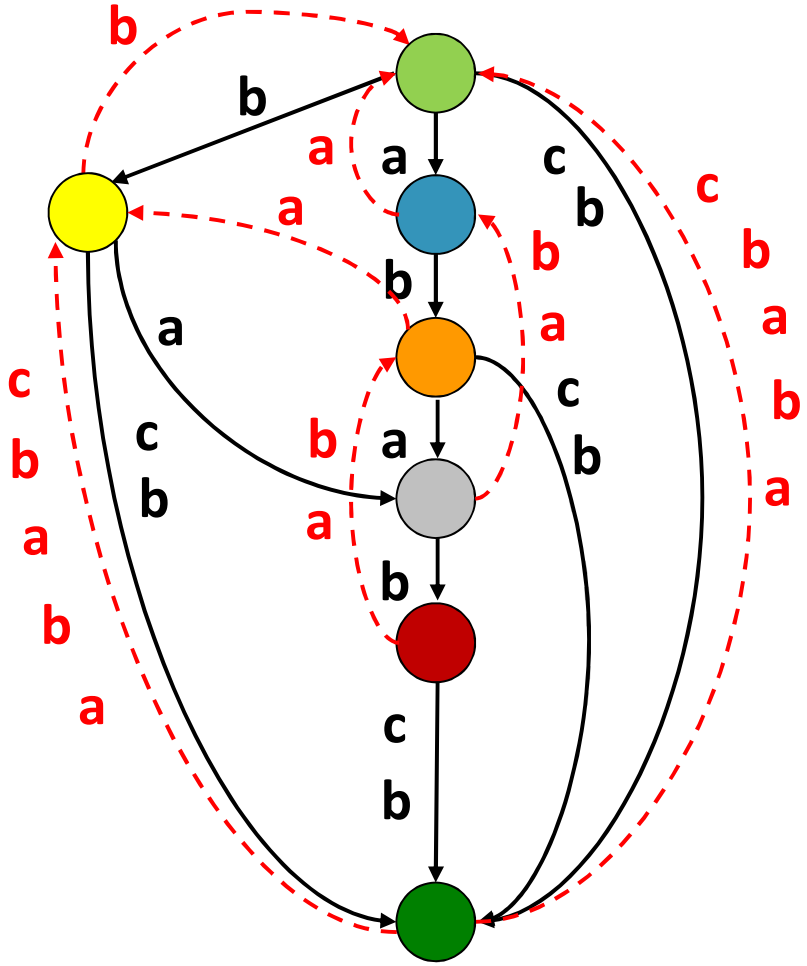


DAWG of **ababcb**

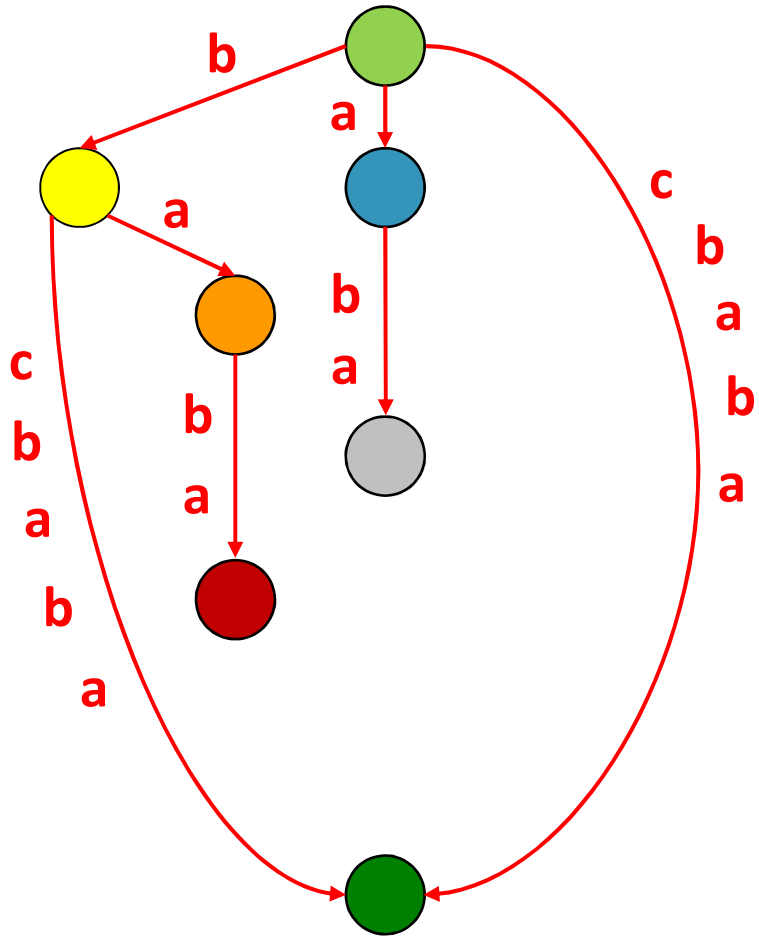


Suffix Tree of **bcbaba**

From DAWG to CDAWG (パス縮約)

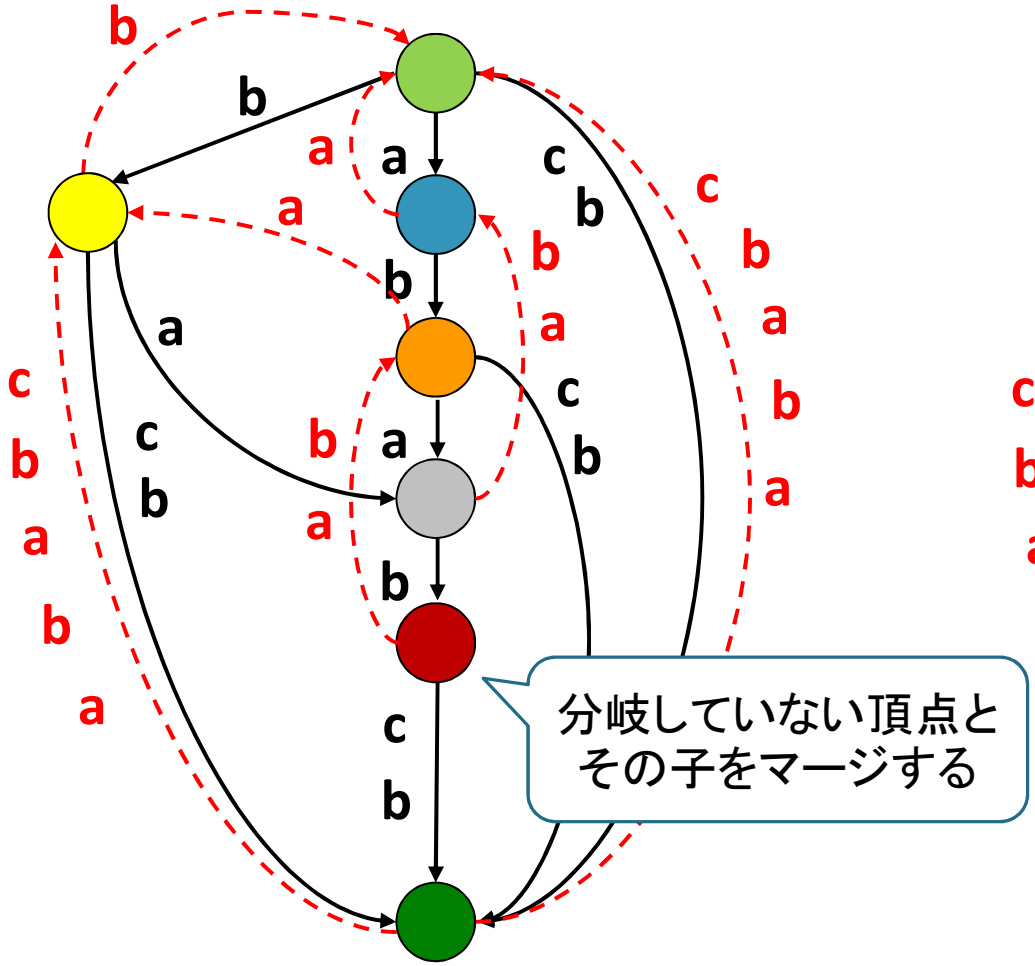


DAWG of **ababcb**

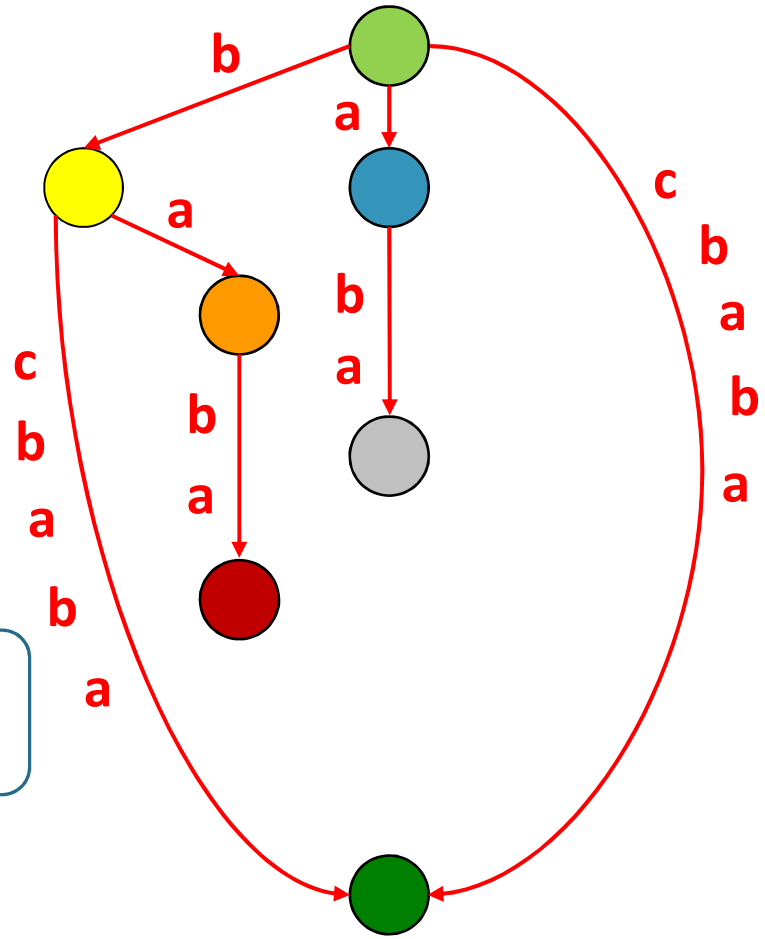


Suffix Tree of **bcbaba**

From DAWG to CDAWG (パス縮約)

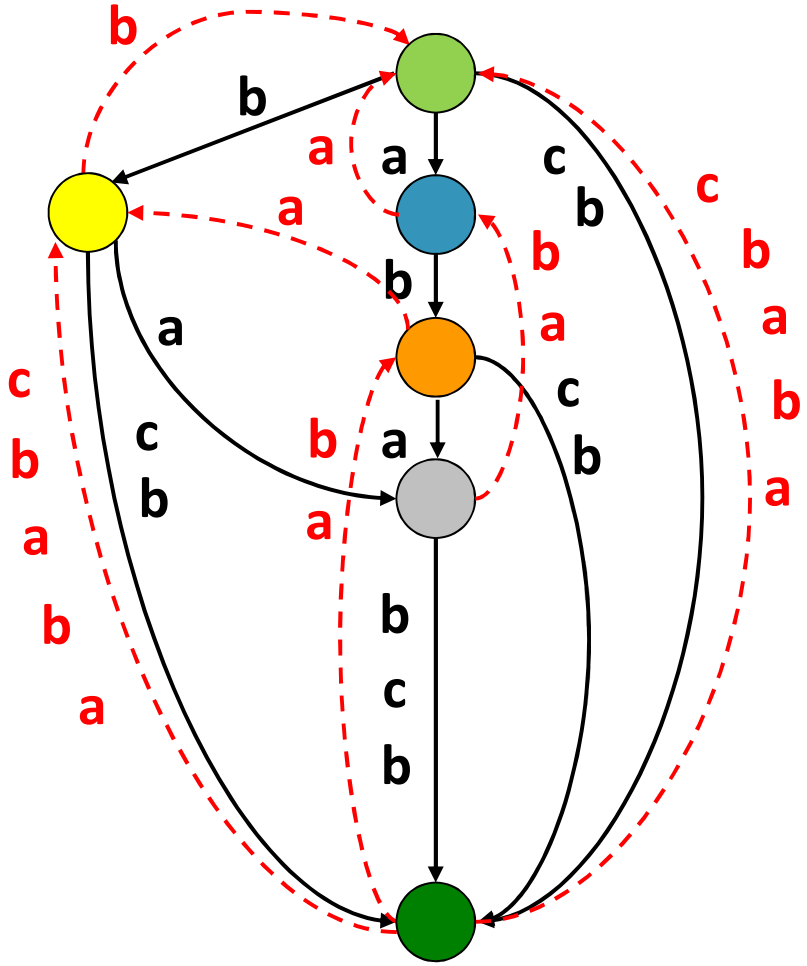


DAWG of **ababcb**

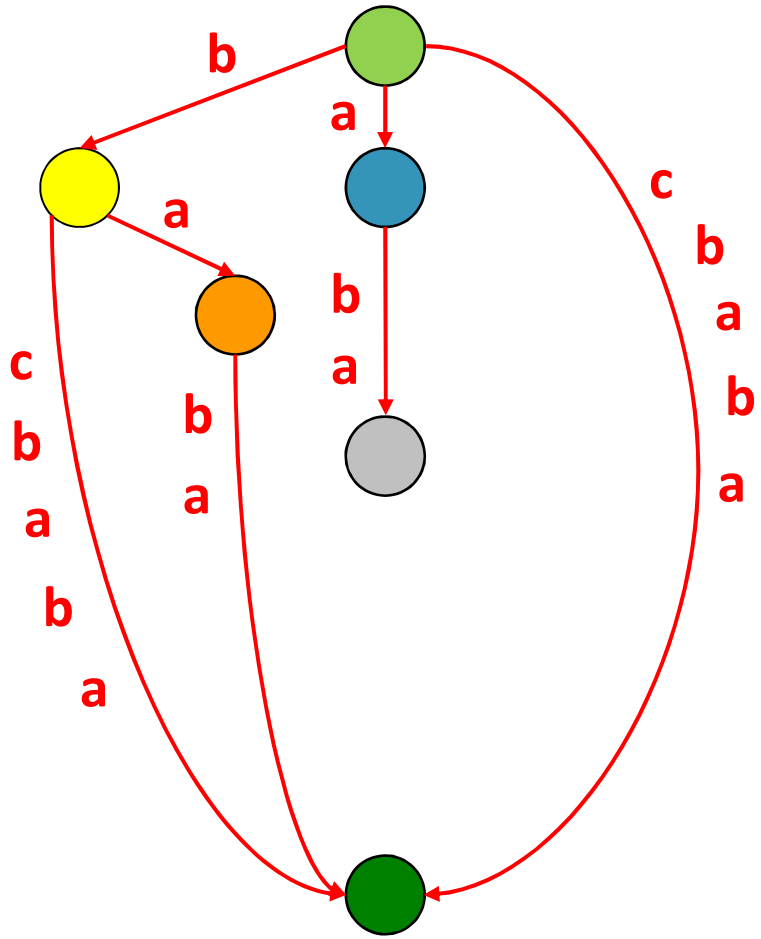


Suffix Tree of **bcbaba**

From DAWG to CDAWG (パス縮約)

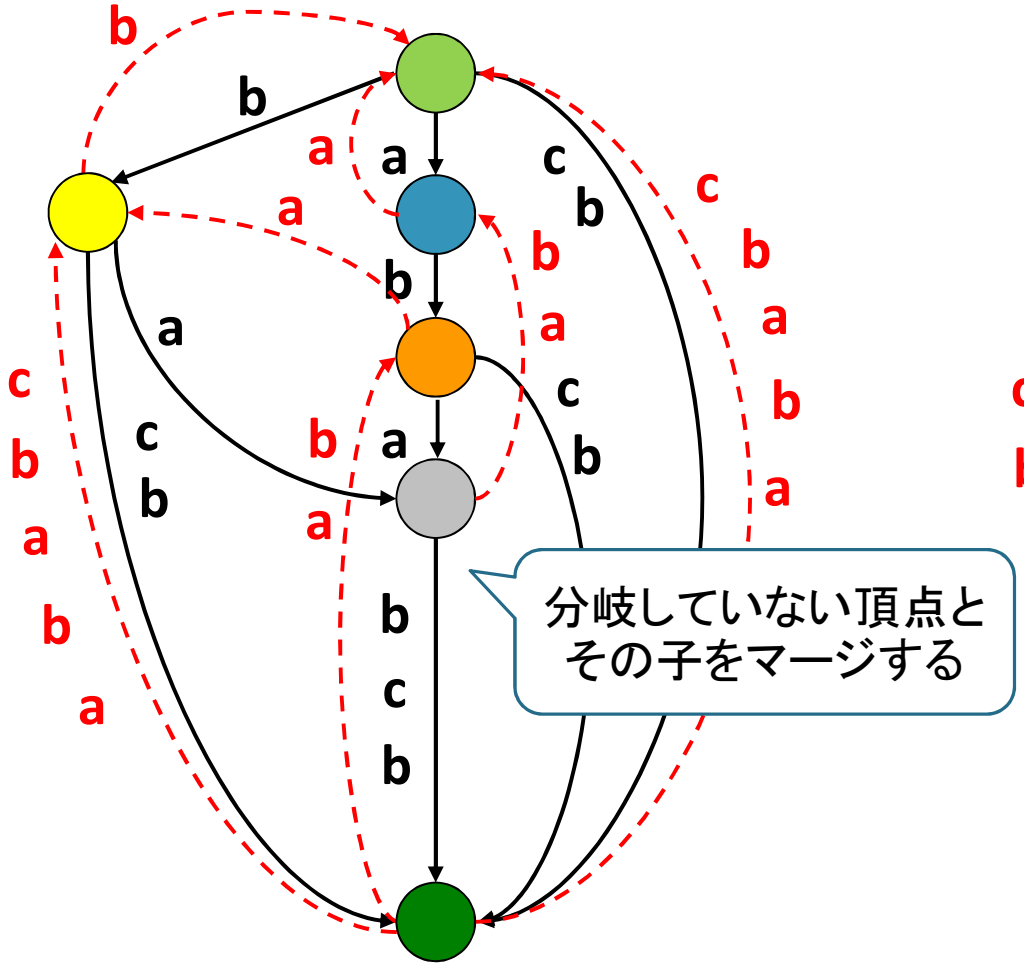


DAWG of **ababcb**

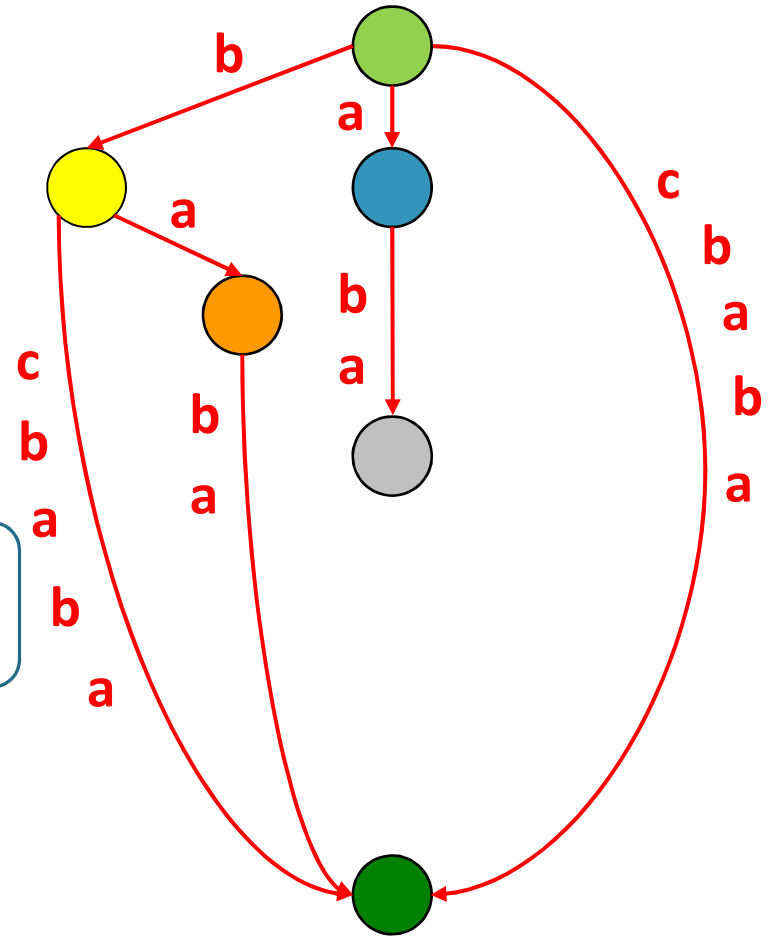


Suffix Tree of **bcbaba**

From DAWG to CDAWG (パス縮約)

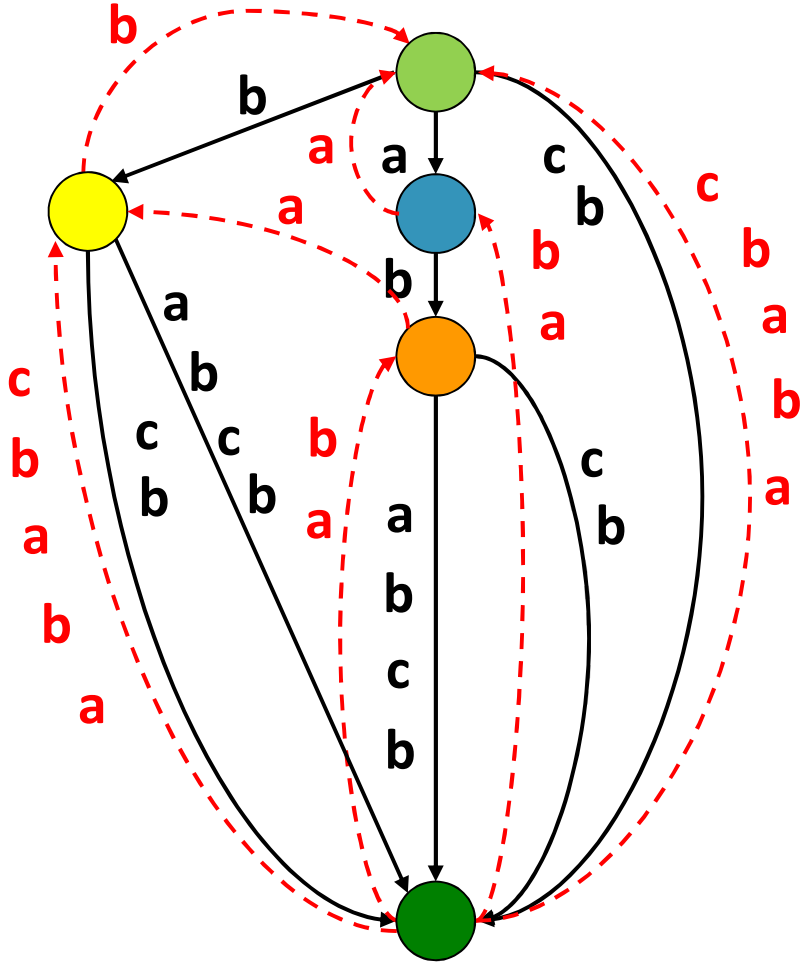


DAWG of **ababcb**

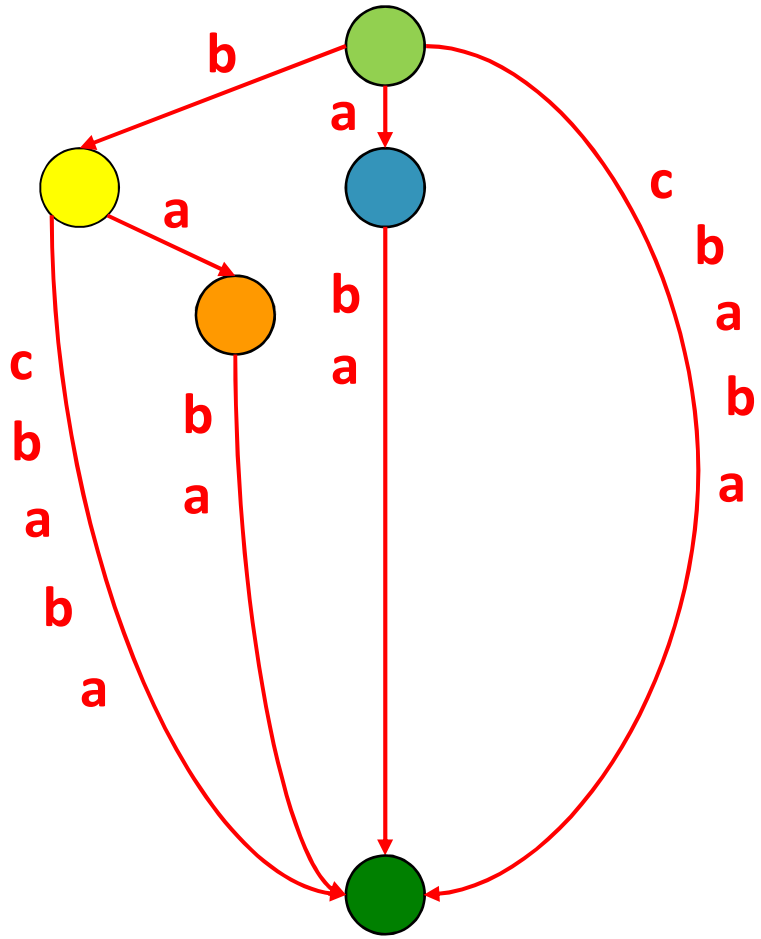


Suffix Tree of **bcbaba**

From DAWG to CDAWG (パス縮約)

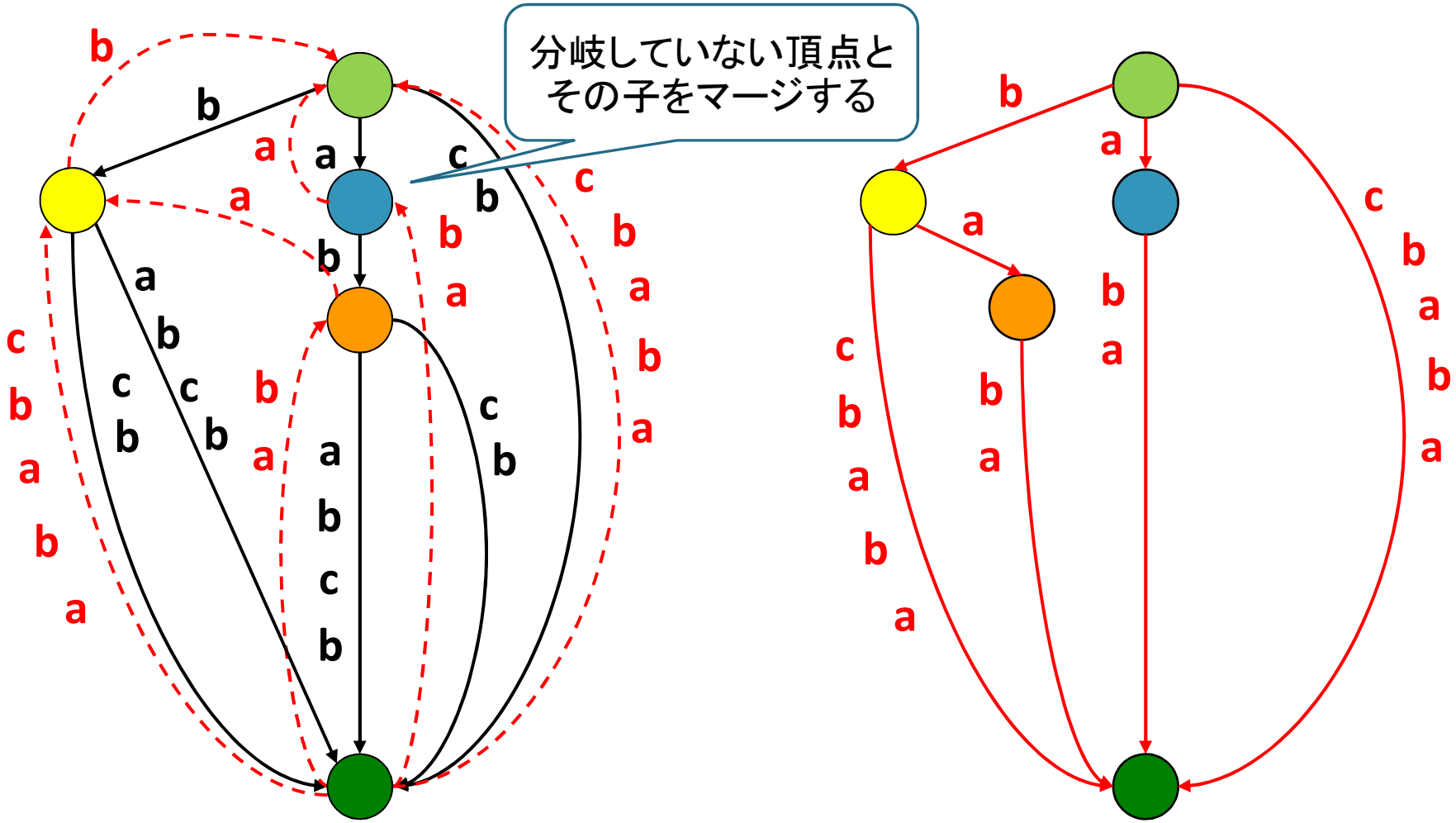


DAWG of **ababcb**



Suffix Tree of **bcbaba**

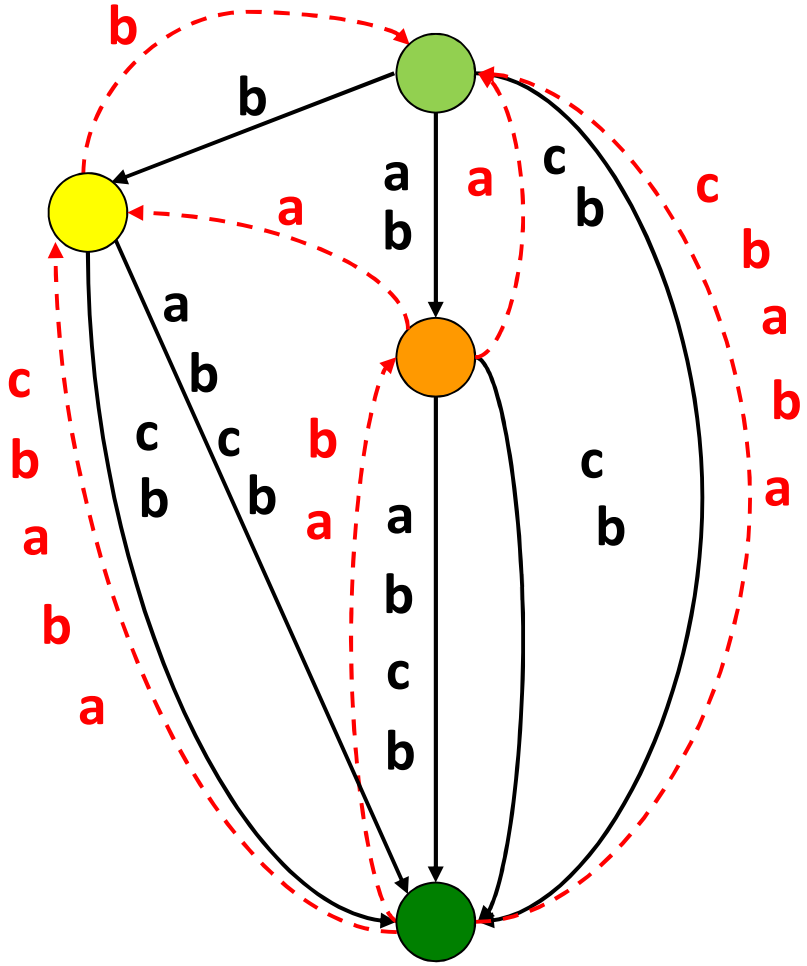
From DAWG to CDAWG (パス縮約)



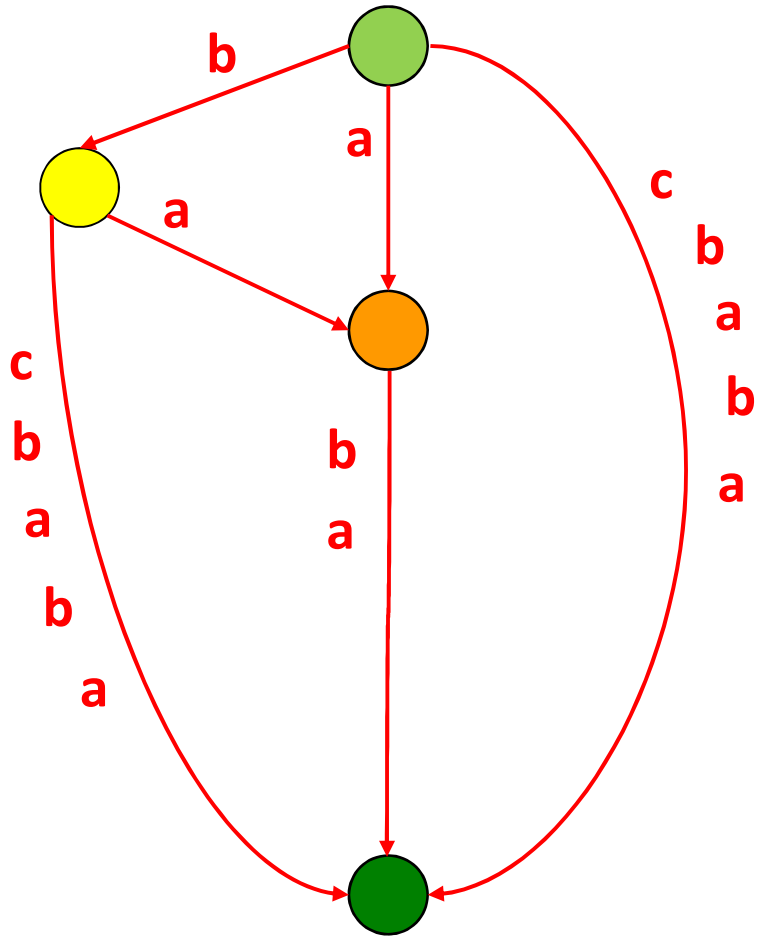
DAWG of **ababcb**

Suffix Tree of **bcbaba**

From DAWG to CDAWG (パス縮約)



CDAWG of **ababcb**



CDAWG of **bcbaba**

From DAWG to (S)CDAWG

定理 [Blumer et al. 1987]

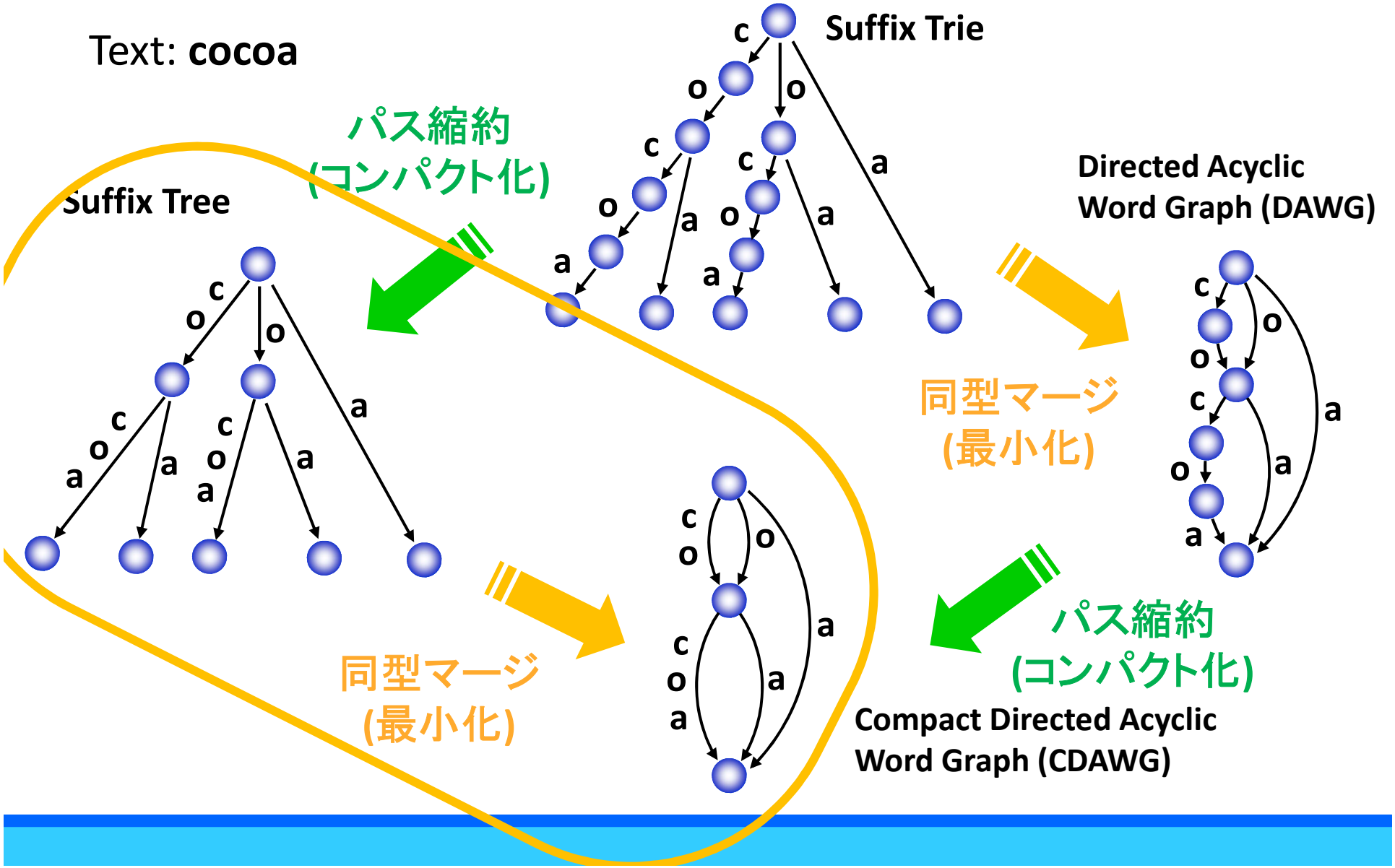
DAWG から CDAWG を $O(n)$ 時間で得られる.

定理 [Blumer et al. 1987]

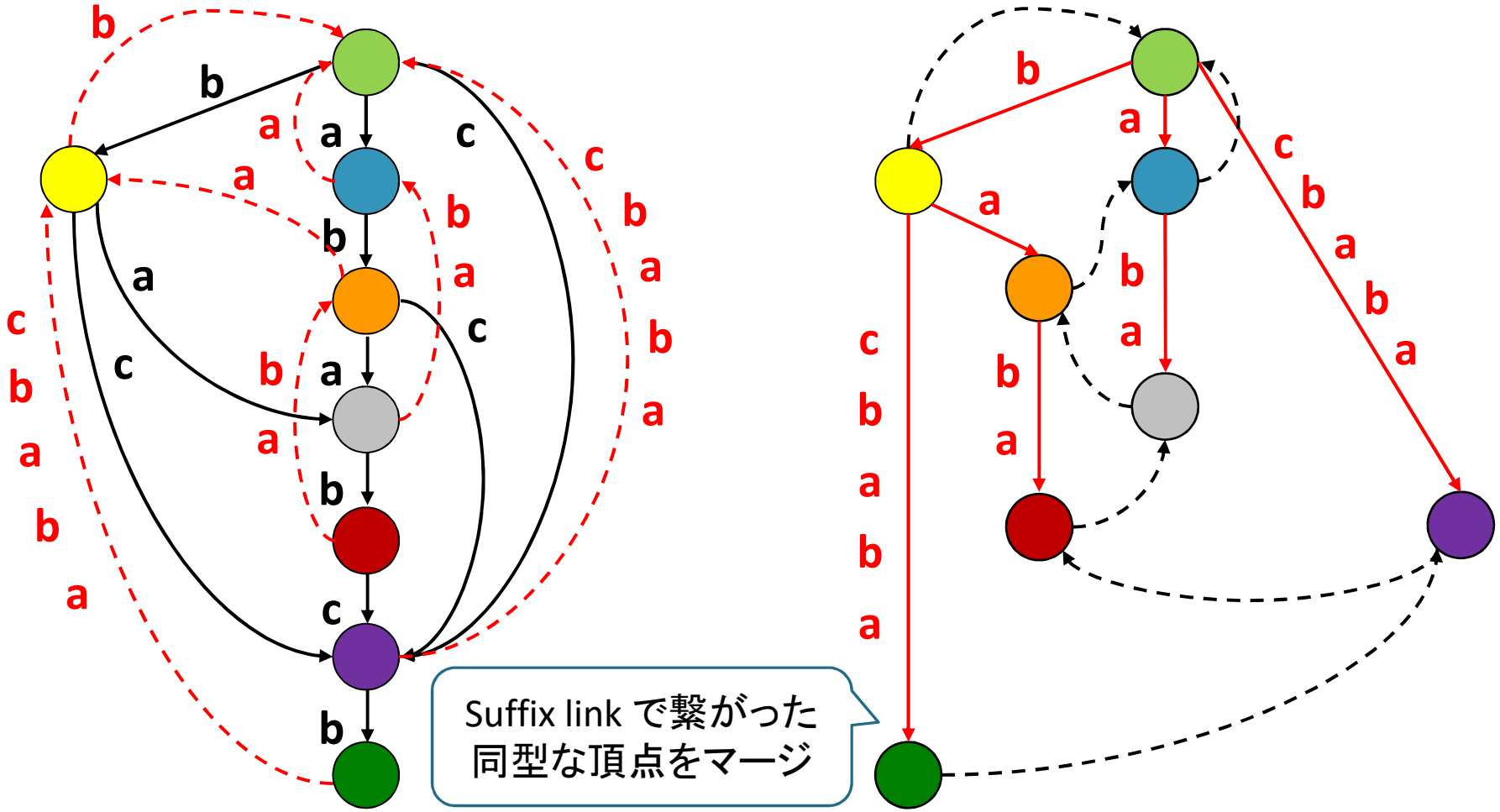
Suffix link 付き DAWG から Symmetric CDAWG を $O(n)$ 時間で得られる.

n は入力文字列長

基礎的なテキスト索引


Text: cocoa
Suffix Trie
パス縮約
(コンパクト化)
Suffix Tree
Directed Acyclic
Word Graph (DAWG)
同型マージ
(最小化)
同型マージ
(最小化)
パス縮約
(コンパクト化)
Compact Directed Acyclic
Word Graph (CDAWG)

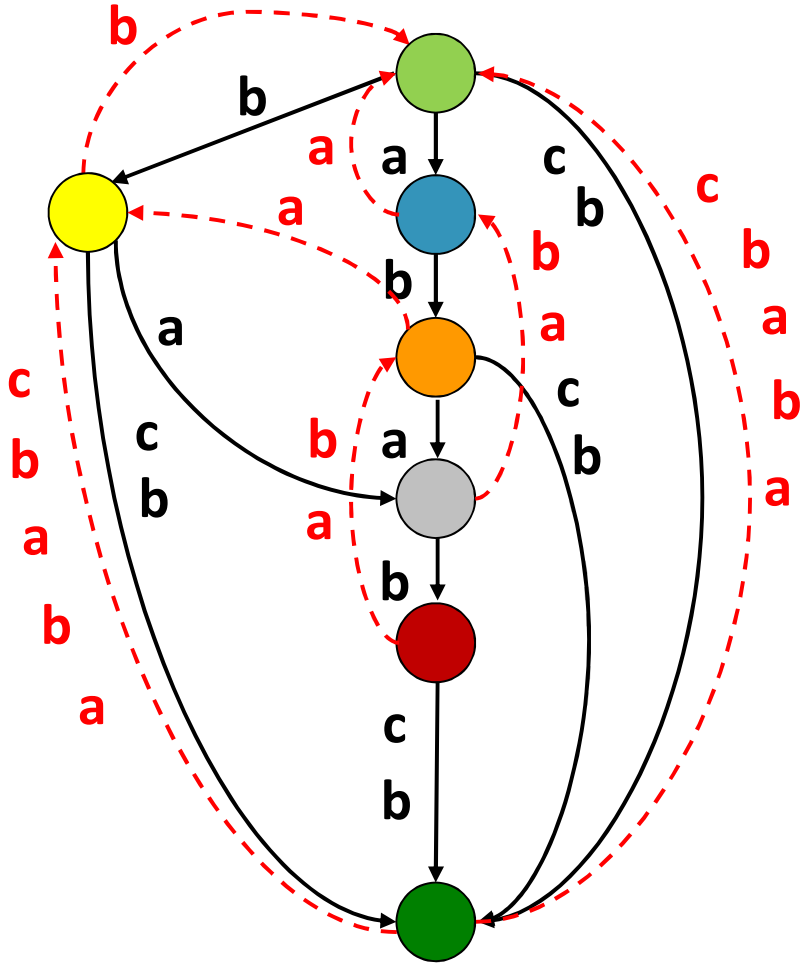
From Suffix Tree to CDAWG (最小化)



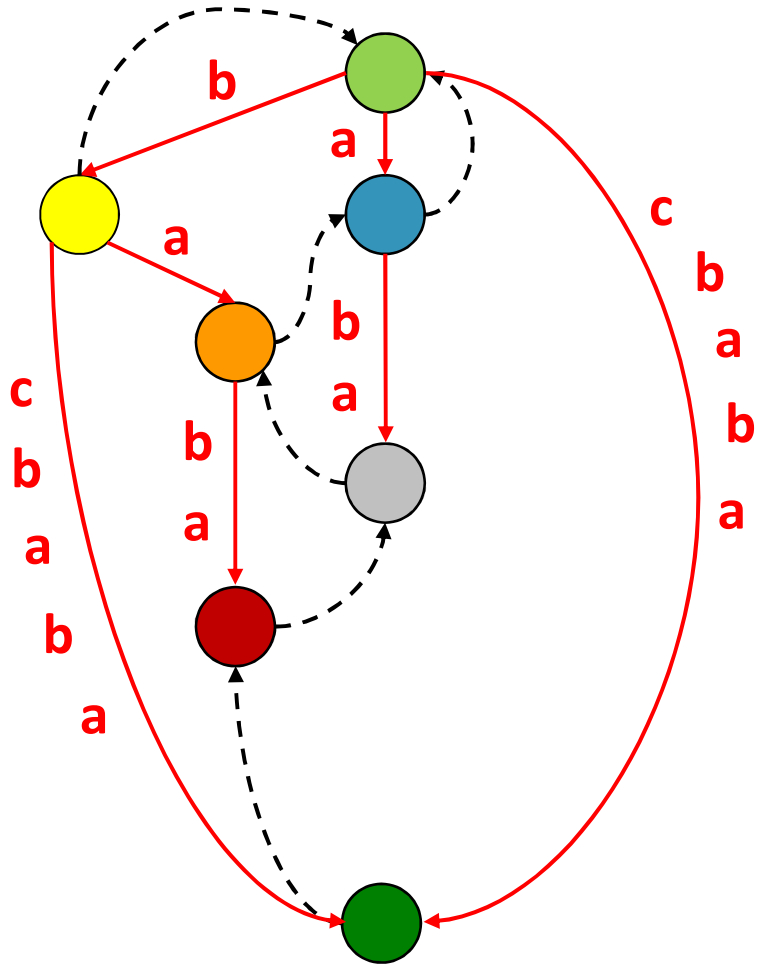
DAWG of **ababcb**

Suffix Tree of **bcbaba**

From Suffix Tree to CDAWG (最小化)

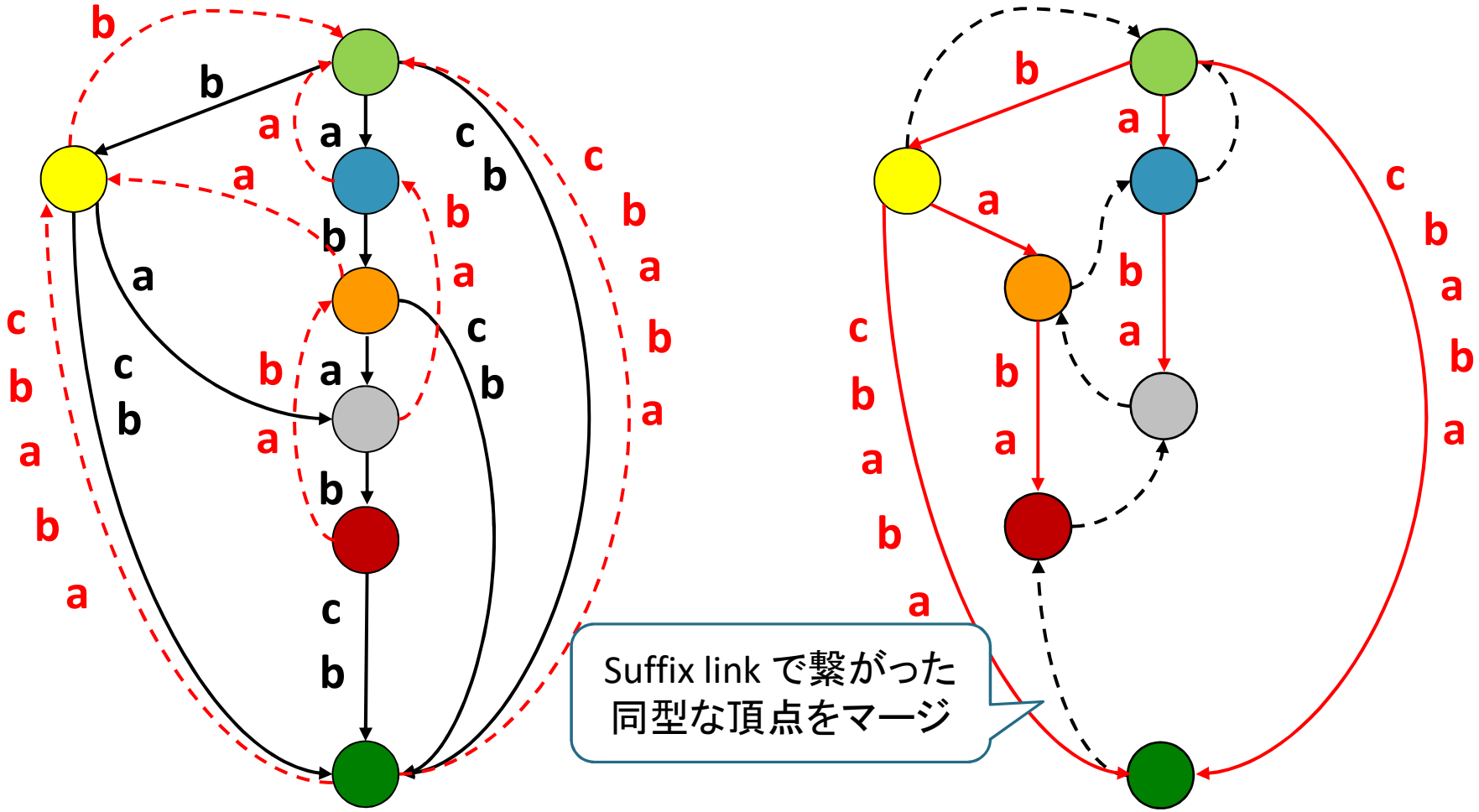


DAWG of **ababcb**



Suffix Tree of **bcbaba**

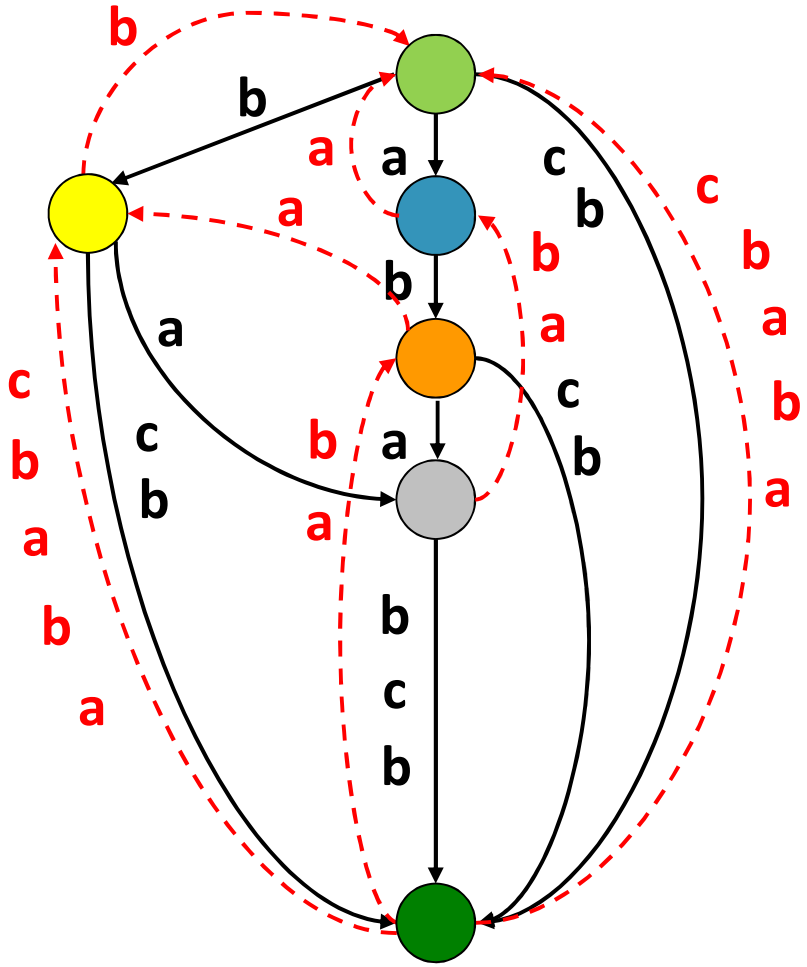
From Suffix Tree to CDAWG (最小化)



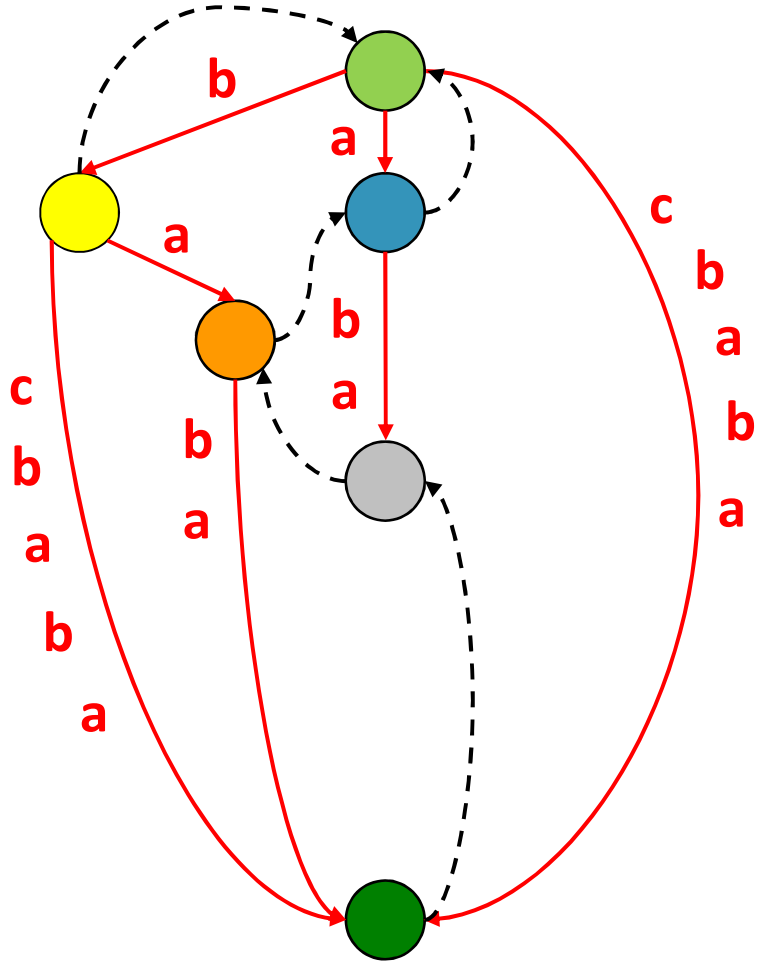
DAWG of **ababcb**

Suffix Tree of **bcbaba**

From Suffix Tree to CDAWG (最小化)

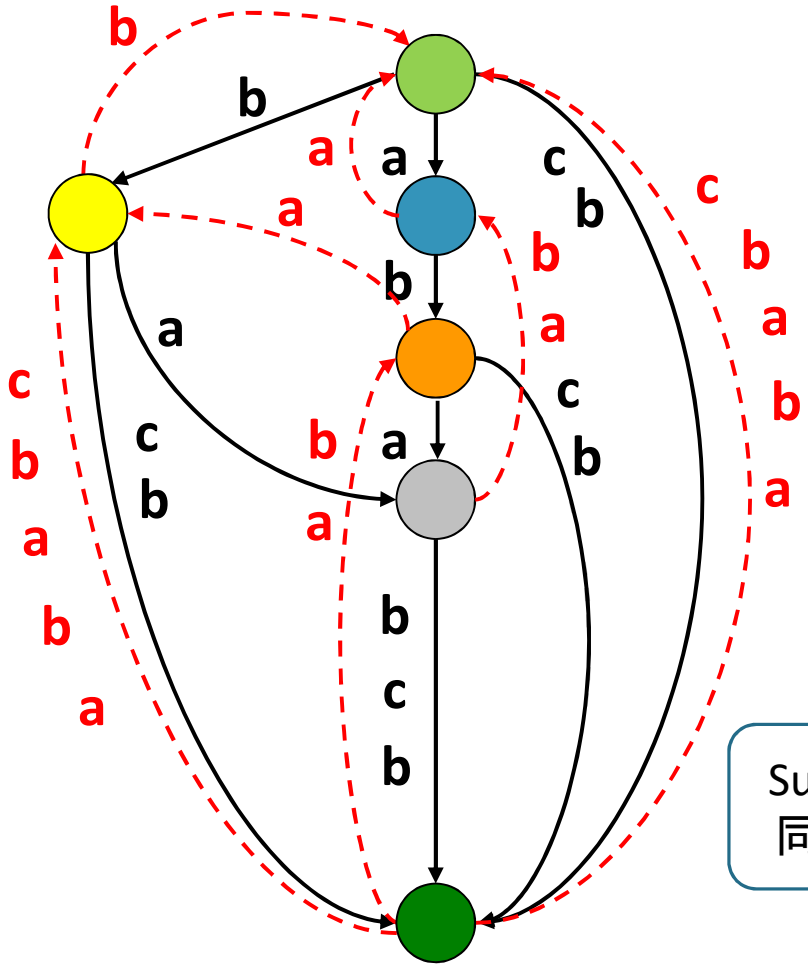


DAWG of **ababcb**

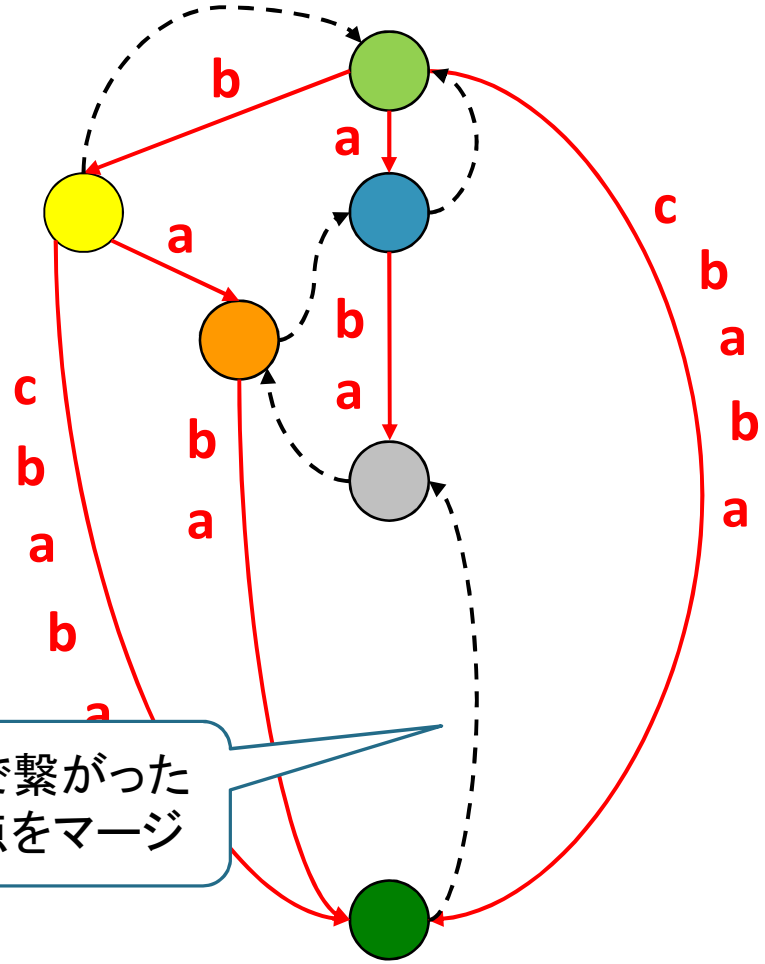


Suffix Tree of **bcbaba**

From Suffix Tree to CDAWG (最小化)

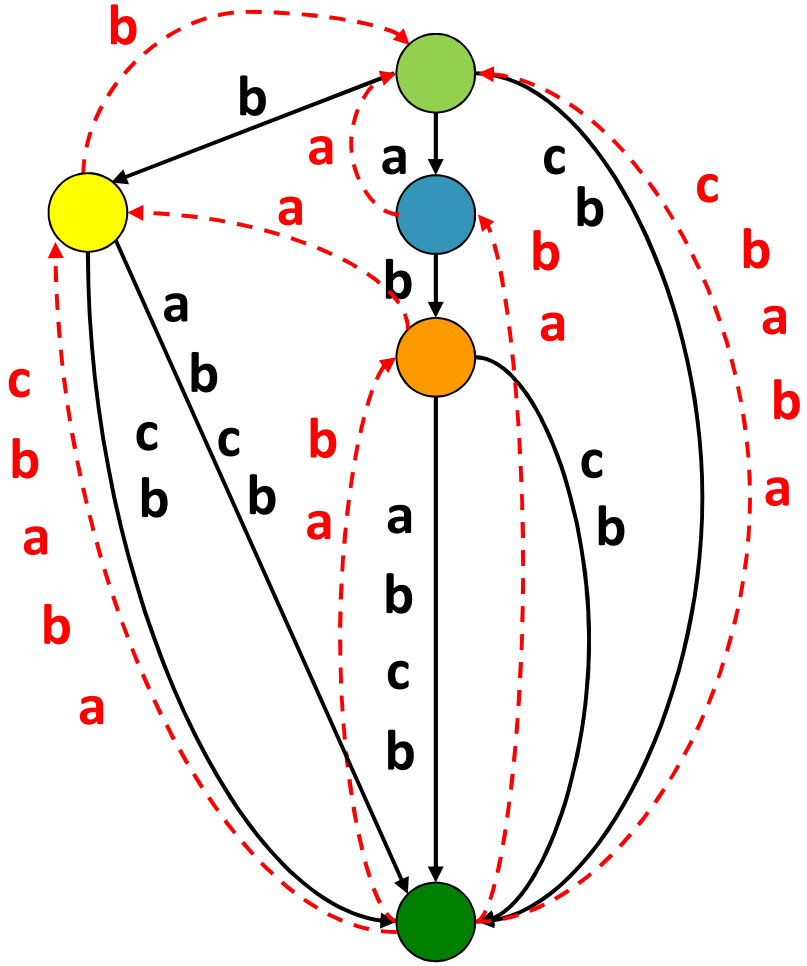


DAWG of **ababcb**

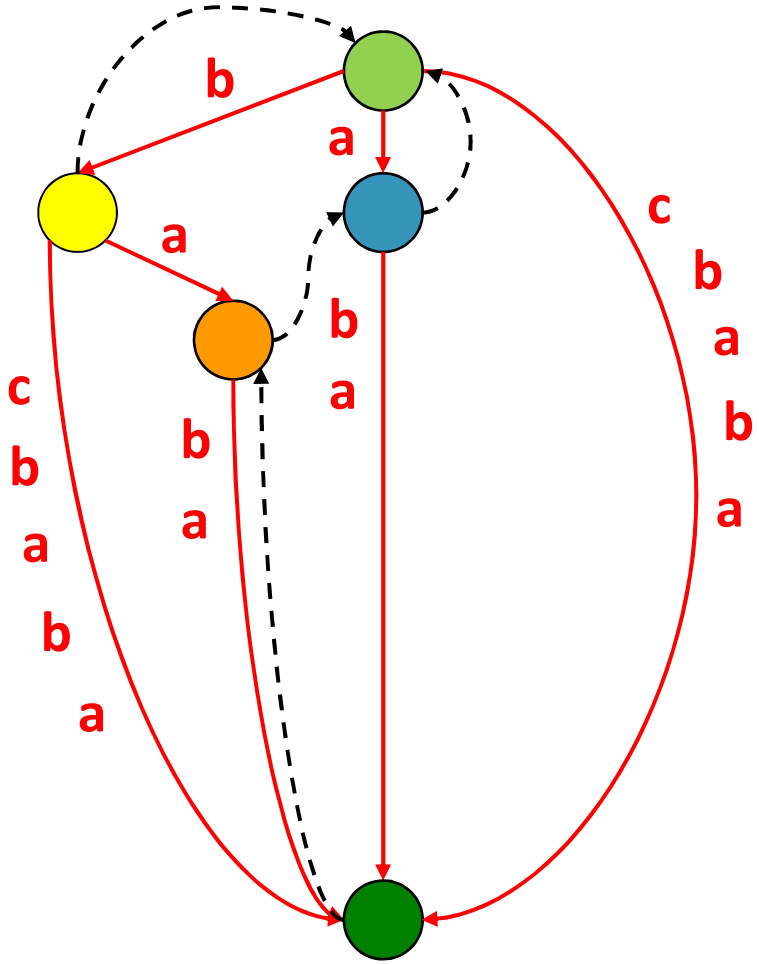


Suffix Tree of **bcbaba**

From Suffix Tree to CDAWG (最小化)

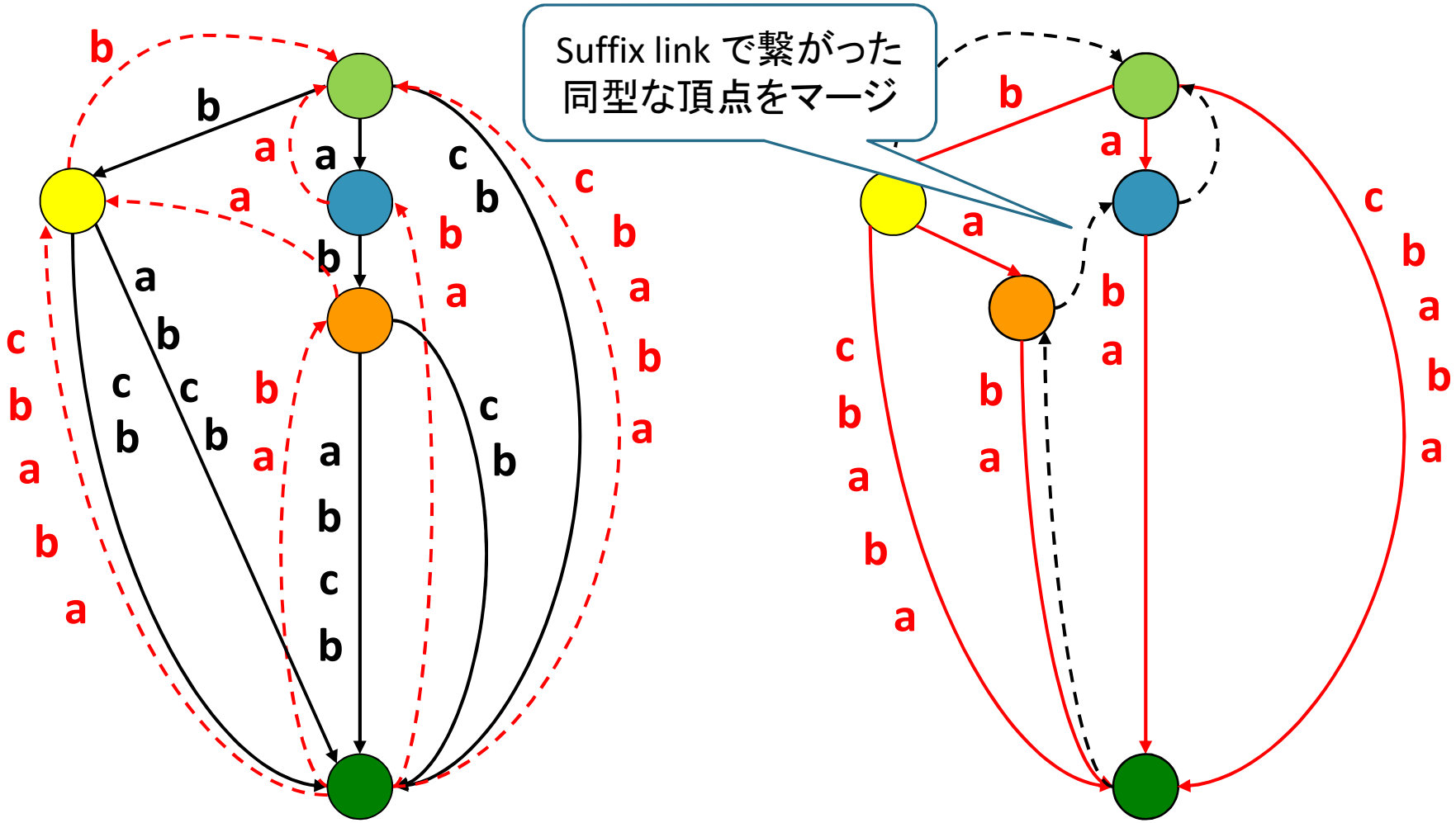


DAWG of **ababcb**



Suffix Tree of **bcbaba**

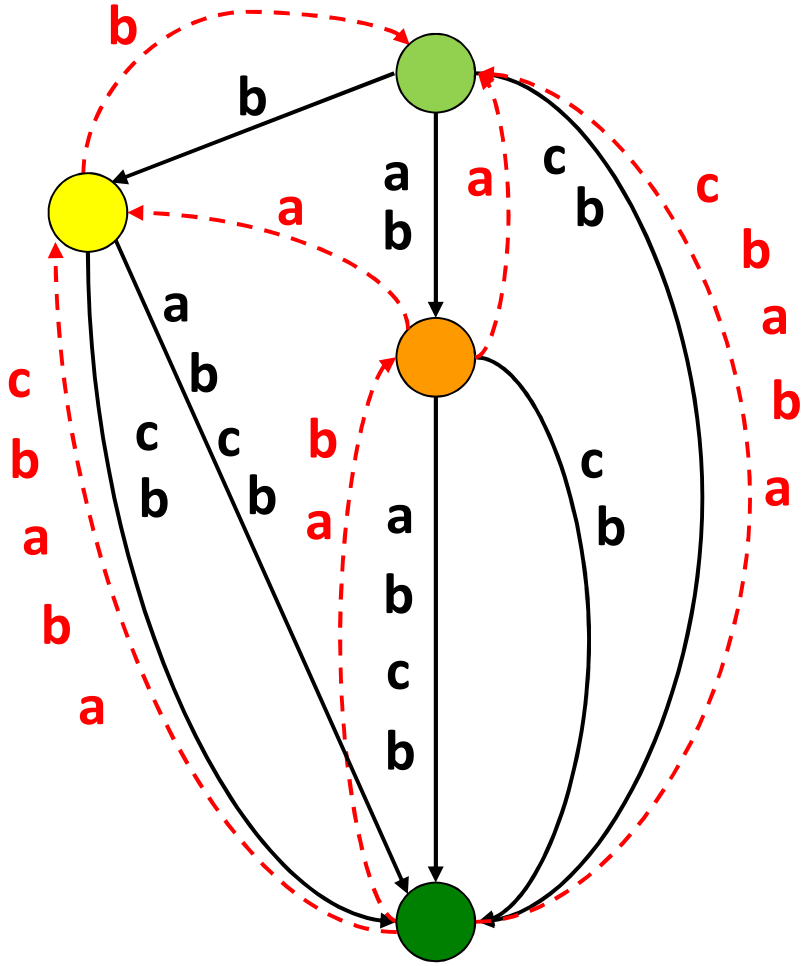
From Suffix Tree to CDAWG (最小化)



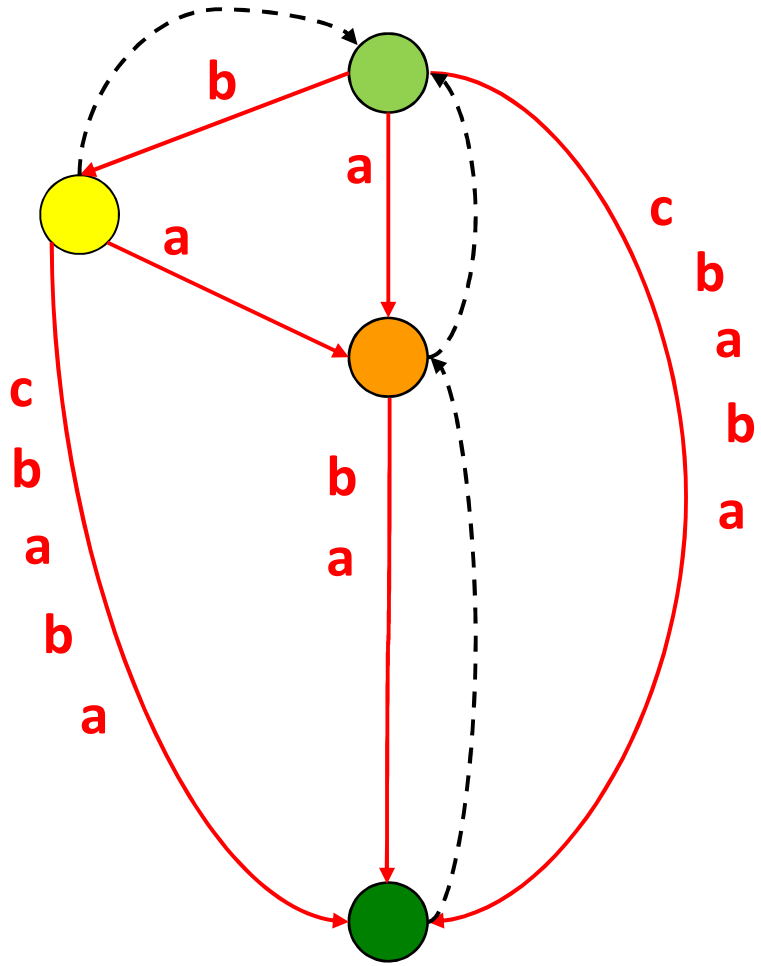
DAWG of **ababcb**

Suffix Tree of **bcbaba**

From Suffix Tree to CDAWG (最小化)

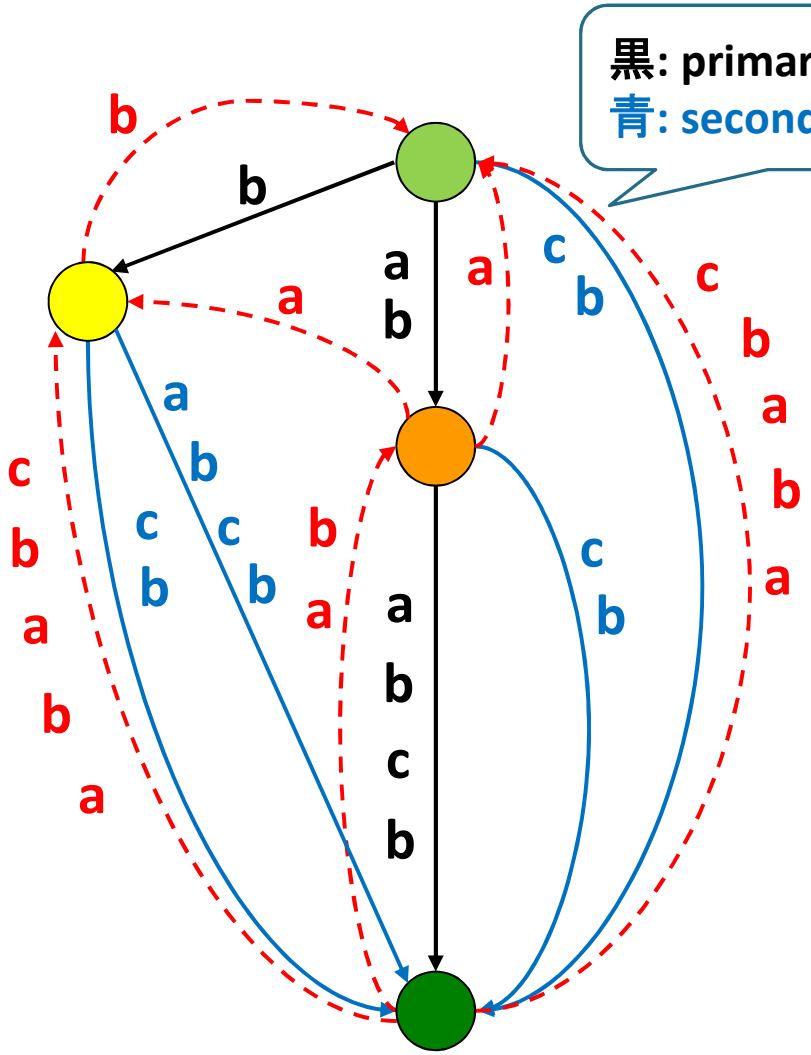


CDAWG of **ababcb**

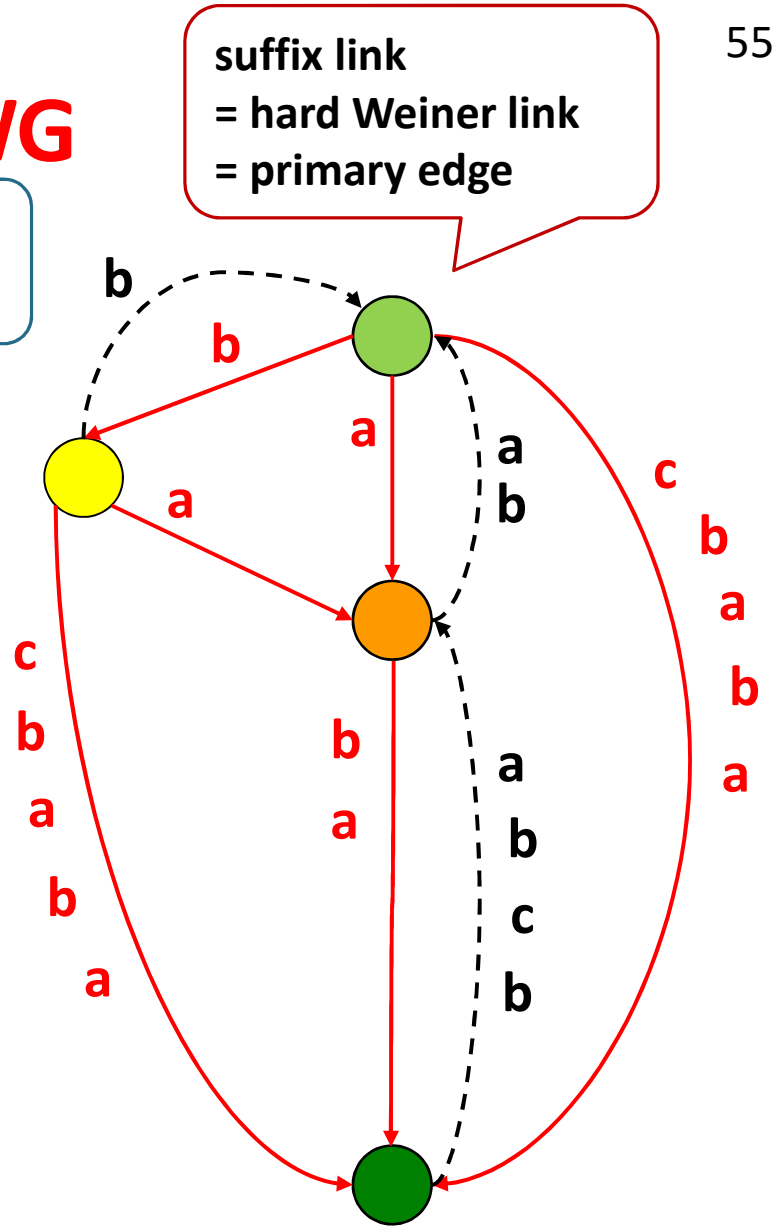


CDAWG of **bcbaba**

From Suffix Tree to CDAWG



CDAWG of **ababcb**



CDAWG of **bcbaba**

From Suffix Tree to (S)CDAWG

定理 [Blumer et al. 1987]

Suffix Tree から CDAWG を $O(n)$ 時間で得られる.

定理 [Blumer et al. 1987]

Soft Weiner link 付き Suffix Tree から
Symmetric CDAWG を $O(n)$ 時間で得られる.

- Soft Weiner link = DAWG の secondary edge なので,
  結局 DAWG ➔ CDAWG の変換とまったく同じ

n は入力文字列長

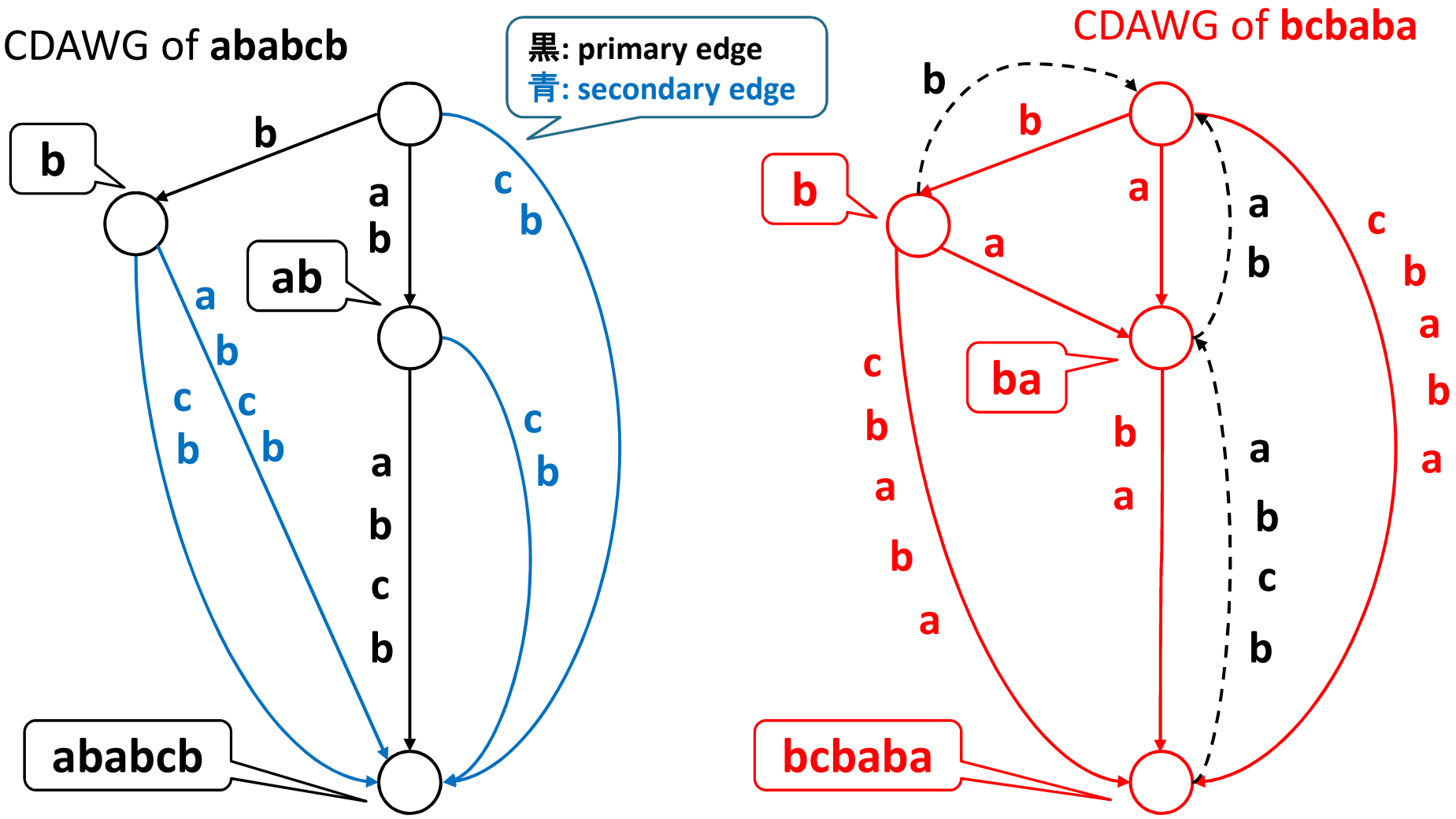
SCDAWG の出力線形領域構築

定理 [Inenaga 2023?]

長さ n の入力文字列 w の SCDAWG を
$O(n \log \sigma)$ 時間, $O(e_l + e_r)$ 領域で構築できる.
e_r と e_l はそれぞれ w と w^R の CDAWG の辺数.

- CDAWG(w\$) と CDAWG(w^R\$) をそれぞれ $O(n \log \sigma)$ 時間, $O(e_r)$ 領域・$O(e_l)$ 領域で構築 [Inenaga et al. 2005] し, \$辺を削除.
- CDAWG(w) と CDAWG(w^R) の頂点の対応を求めたい.
- 順向き文字列 w 側の代表元を lex order でソートする.
  → (すでにソートされている) primary edge を巡回すればよい.
- 逆向き文字列 w^R 側の代表元を co-lex order でソートする.
  → CDAWG(w^R) の suffix link を辞書式順序でソートする.
  $O(\min(e_r, e_l) \log \sigma)$ 時間, $O(\min(e_r, e_l))$ 領域
  → CDAWG(w^R) の suffix link = CDAWG(w) の primary edge
  なので, ソートされた suffix link を巡回すればよい.

Symmetric CDAWG



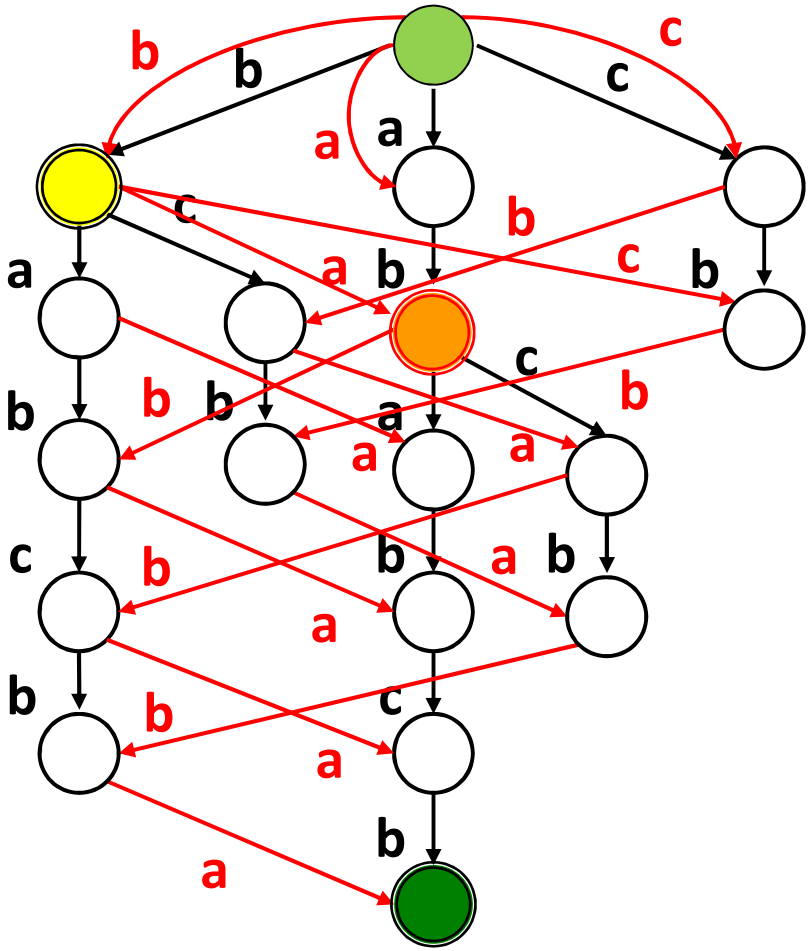
代表元の lex order のソート順:
ab, ababcb, b

代表元の co-lex order のソート順
ba, bcbaba, b

Suffix Trie と SCDAWG の関係

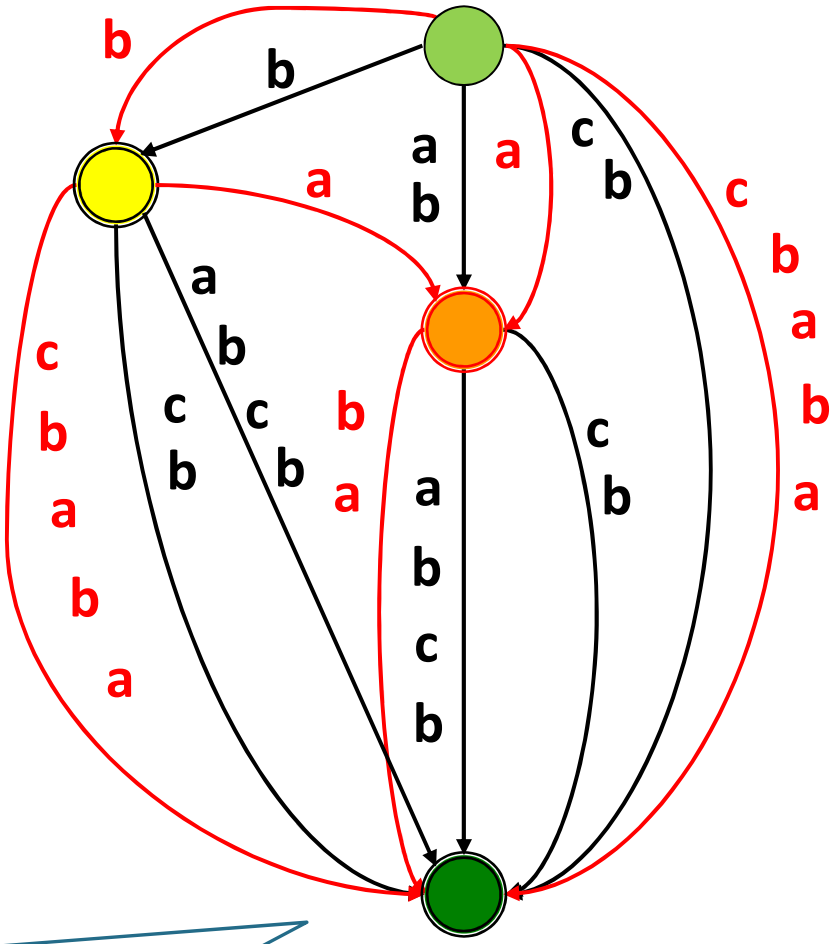
Suffix Trie of ababcb

Suffix Trie of bcbaba



CDAWG of ababcb

CDAWG of bcbaba



Symmetric Suffix Trie の頂点のうち，
順向き文字列 w と逆向き文字列 w^R の<u>両方で</u>
枝別れ or suffix に対応している頂点だけを残したのが SCDAWG