課題３
「文字列処理と動的計画法」

鈴木 潤，吉留 崇

2019年度プログラミング演習A

本課題で学ぶこと

- **編集距離**
  - ◆ 文字列処理 ← 課題 1
  - ◆ 再帰的手続き ← 課題2
  - ◆ **動的計画法**

・問題3-0：編集距離の解説
・問題内にも解説あり

編集距離（Edit Distance）

2つの文字列がどのくらい似ているかを表す指標

（例）「マテオ」、「マンガ」

どちらが「マリオ」に似ている？

	マ リ オ マ テ オ	マ リ オ マ ン ガ
<u>編集距離</u> （後で定義）	**1**	**2**

「マテオ」の方が似ている

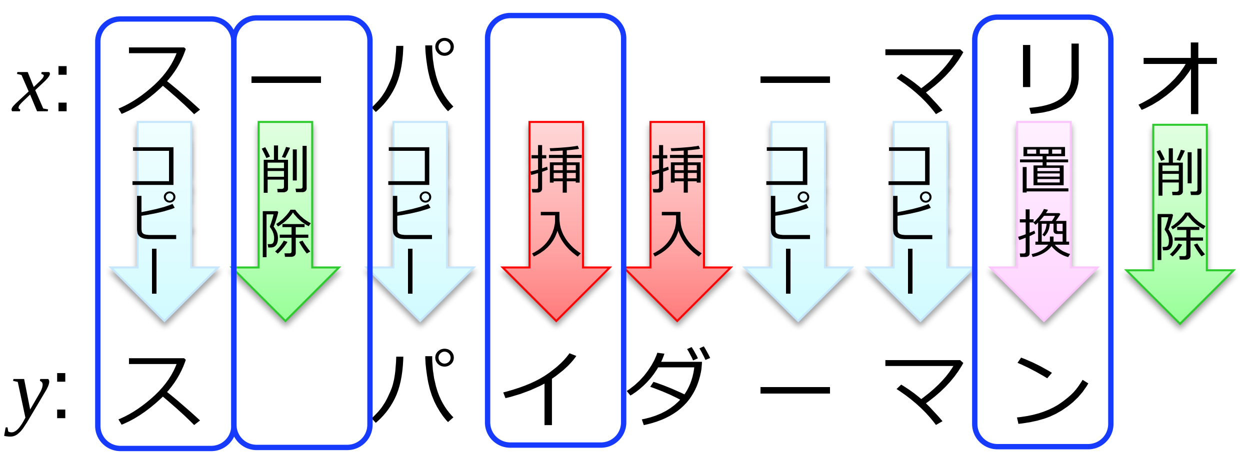
スペルチェッカーやDNAの配列の解析に応用

編集距離（Edit Distance）

文字列x, yの編集距離を求めるために、編集操作を行う

文字列xにコピー、置換、挿入、削除を行い、
文字列yに等しい文字列を得る操作

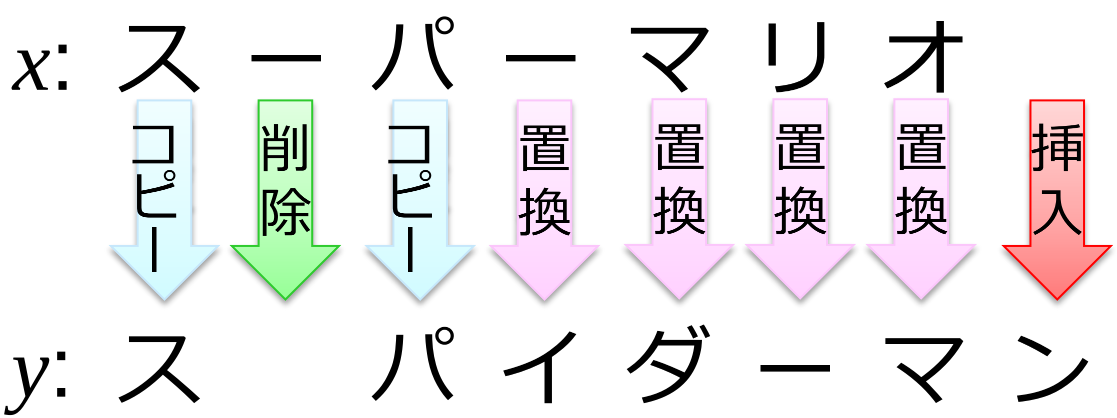
（例）x: スーパーマリオ, y: スパイダーマン



編集距離（Edit Distance）

編集操作は、複数存在

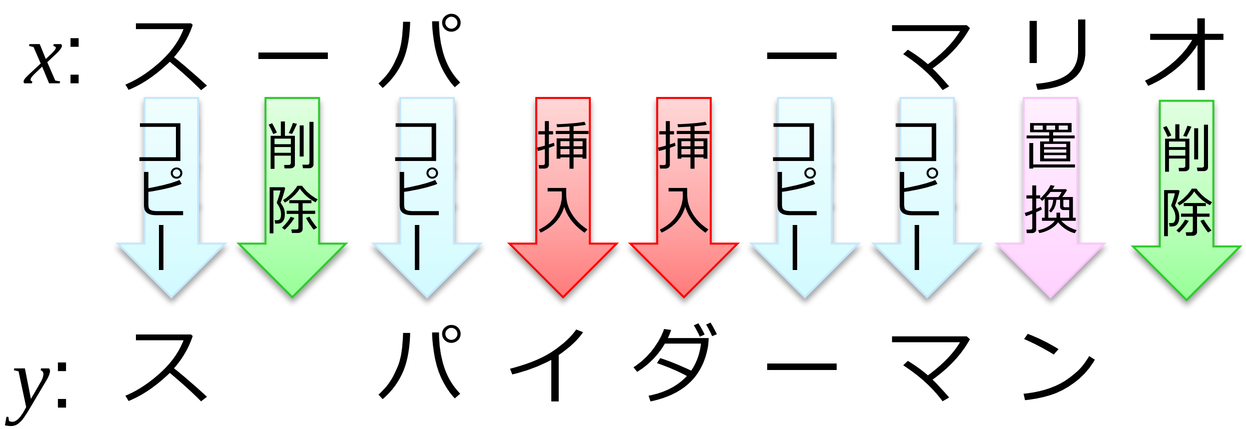
（例）x: スーパーマリオ, y: スパイダーマン



編集距離（Edit Distance）

文字列 x を文字列 y に変換する時の「**削除**」，「**挿入**」，「**置換**」の最小回数（「コピー」はカウントしない）

（例）x: スーパーマリオ, y: スパイダーマン



編集距離 ＝ 5

編集距離の計算方法

文字列に対する数学記号

	1	2	3	4	5	6	7	8	9
x :	a	b	b	a	b	a	a	a	b

文字列 x の i 番目の文字：x_i

例）$x_4 = \mathrm{a}$

文字列 x の先頭からi 番目までの文字列：X_i ⇒ 接頭辞

例）$X_4 = \mathrm{abba}$

長さ 0 の文字列：ε

編集距離の計算方法

文字列 $x=<x_1, x_2,\ldots,x_m>$, $y=<y_1, y_2,\ldots,y_n>$の接頭辞X_i, Y_jの編集距離: $c_{i,j}$

$$c_{i,j} = \begin{cases} \max(i, j) & (i = 0 \text{ または } j = 0 \text{ の時}) \\ \min(c_{i-1,j-1} + d(x_i, y_j), \\ \quad c_{i-1,j} + 1, & (\text{その他}) \\ \quad c_{i,j-1} + 1) \end{cases}$$

$$d(x_i, y_j) = \begin{cases} 1 \ (x_i \neq y_j) \\ 0 \ (x_i = y_j) \end{cases}$$

＊クロネッカーのデルタではない事に注意

編集距離の計算方法

$$c_{i,j} = \boxed{\min}\left(c_{i-1,j-1} + d(x_i, y_j),\ c_{i-1,j} + 1,\ c_{i,j-1} + 1\right),\quad \begin{cases} 1\ (x_i \neq y_j) \\ 0\ (x_i = y_j) \end{cases}$$

置換

コピー

削除

挿入

＊クロネッカーのデルタ
ではない事に注意

（例）x=ab、y=cd

$c_{2,2}$<ab, cd>の編集距離は、以下の距離の最小値

$c_{1,1}$<a, c> +d（b, d）

a b

置換

c d

$c_{1,2}$<a, cd> +1

a b

削除

c d

$c_{2,1}$<ab, c> +1

a b

挿入

c d

編集距離の計算方法

$c_{i,j} = \max(i, j)$ 　　($i = 0$または$j = 0$の時)

$c_{i,0} < X_i, \varepsilon > = i$, 削除

x : a b

削除 削除

y :

$c_{0,j} < \varepsilon, Y_j > = j$ 挿入

x :

挿入 挿入

y : c d

編集距離の計算方法

$$c_{i,j} = \min(c_{i-1,j-1} + d(x_i, y_i),\ c_{i-1,j} + 1,\ c_{i,j-1} + 1), \quad d(x_i,y_j) = \begin{cases} 1 \ (x_i \neq y_j) & \text{置換} \\ 0 \ (x_i = y_j) & \text{コピー} \end{cases}$$

$c_{i-1,j} + 1$：削除

$c_{i,j-1} + 1$：挿入

$c_{i,0} = i$, 削除 　 $c_{0,j} = j$ 挿入

再帰関数による実装

動的計画法による実装

再帰的に計算

$c_{2,2}$<ab, cd>

$$c_{i,j} \langle X_i, Y_j \rangle = \min(c_{i-1,j-1} + d(x_i, y_j),\; c_{i-1,j} + 1,\; c_{i,j-1} + 1)$$

$$c_{i,0} \langle X_i, \varepsilon \rangle = i, \qquad c_{0,j} \langle \varepsilon, Y_j \rangle = j$$

再帰的に計算

$c_{1,1}<a, c>\ +d(b,d)$

$c_{2,2}<ab, cd>$

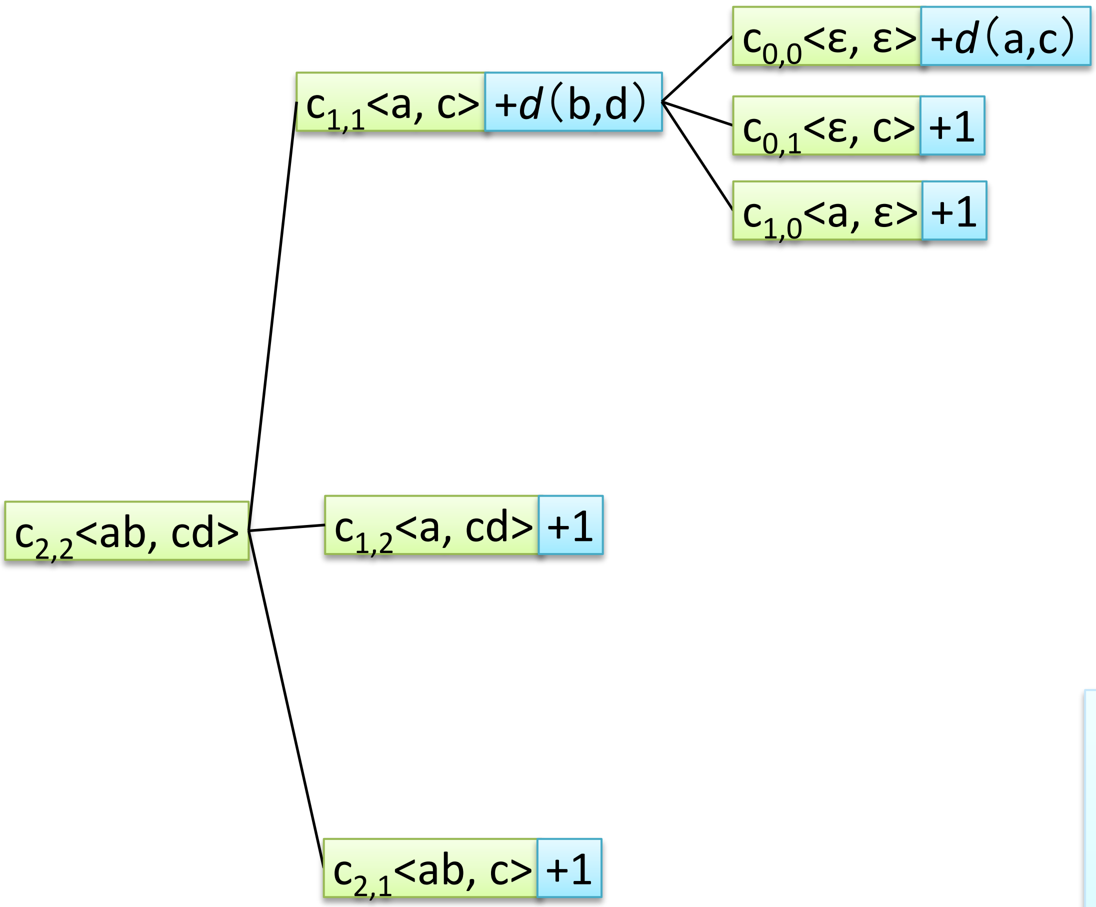
$c_{1,2}<a, cd>\ +1$

$c_{2,1}<ab, c>\ +1$

$$
\begin{aligned}
c_{i,j}<X_i, Y_j> = \min(&c_{i-1,j-1} + d(x_i, y_j),\\
&c_{i-1,j} + 1,\\
&c_{i,j-1} + 1)
\end{aligned}
$$

$$c_{i,0}<X_i, \varepsilon> = i, \qquad c_{0,j}<\varepsilon, Y_j> = j$$

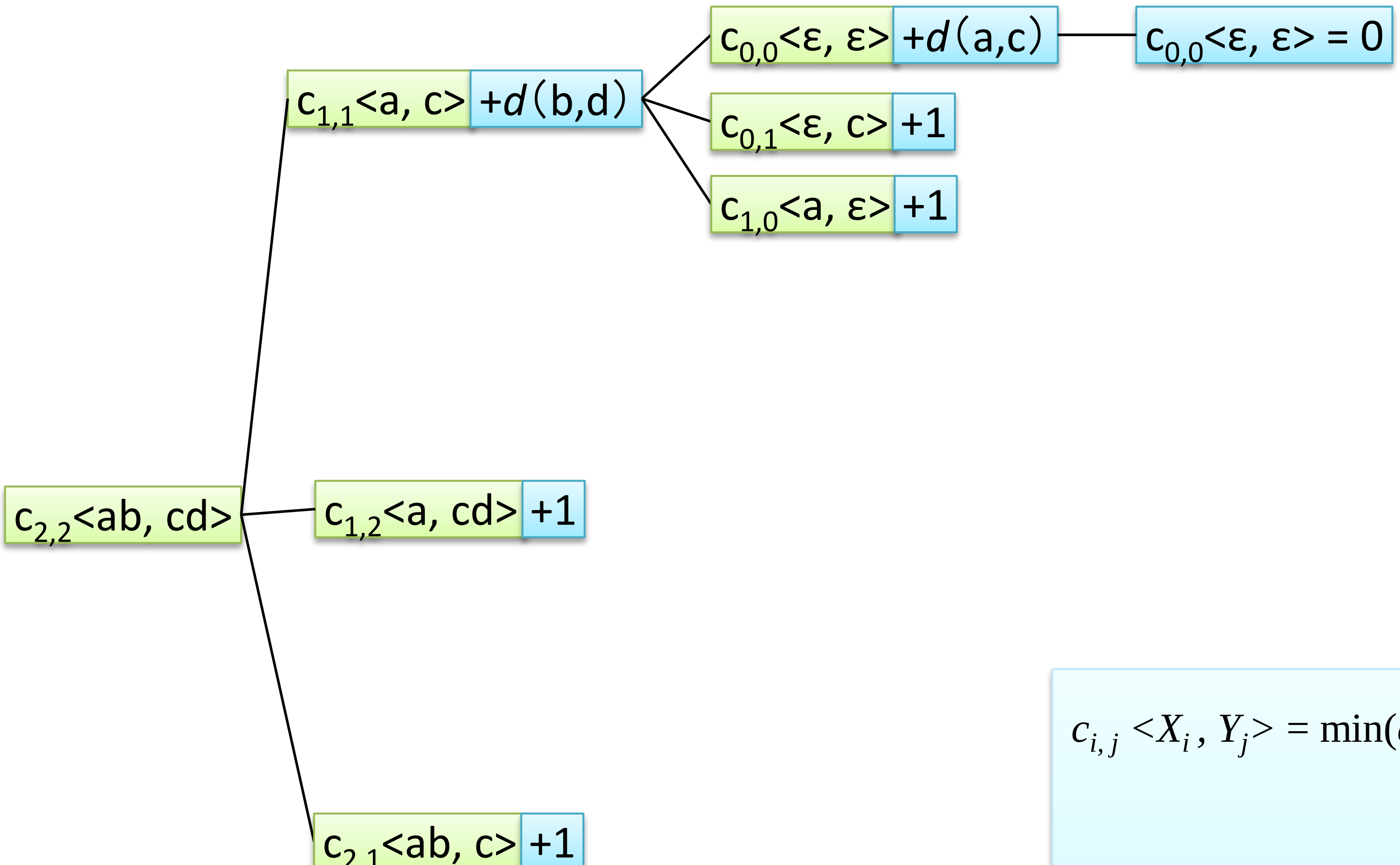
再帰的に計算



$$\begin{aligned} c_{i,j} <X_i, Y_j> = \min(&c_{i-1,j-1} + d(x_i, y_j),\\ &c_{i-1,j} + 1,\\ &c_{i,j-1} + 1) \end{aligned}$$

$$c_{i,0} <X_i, \varepsilon> = i, \qquad c_{0,j} <\varepsilon, Y_j> = j$$

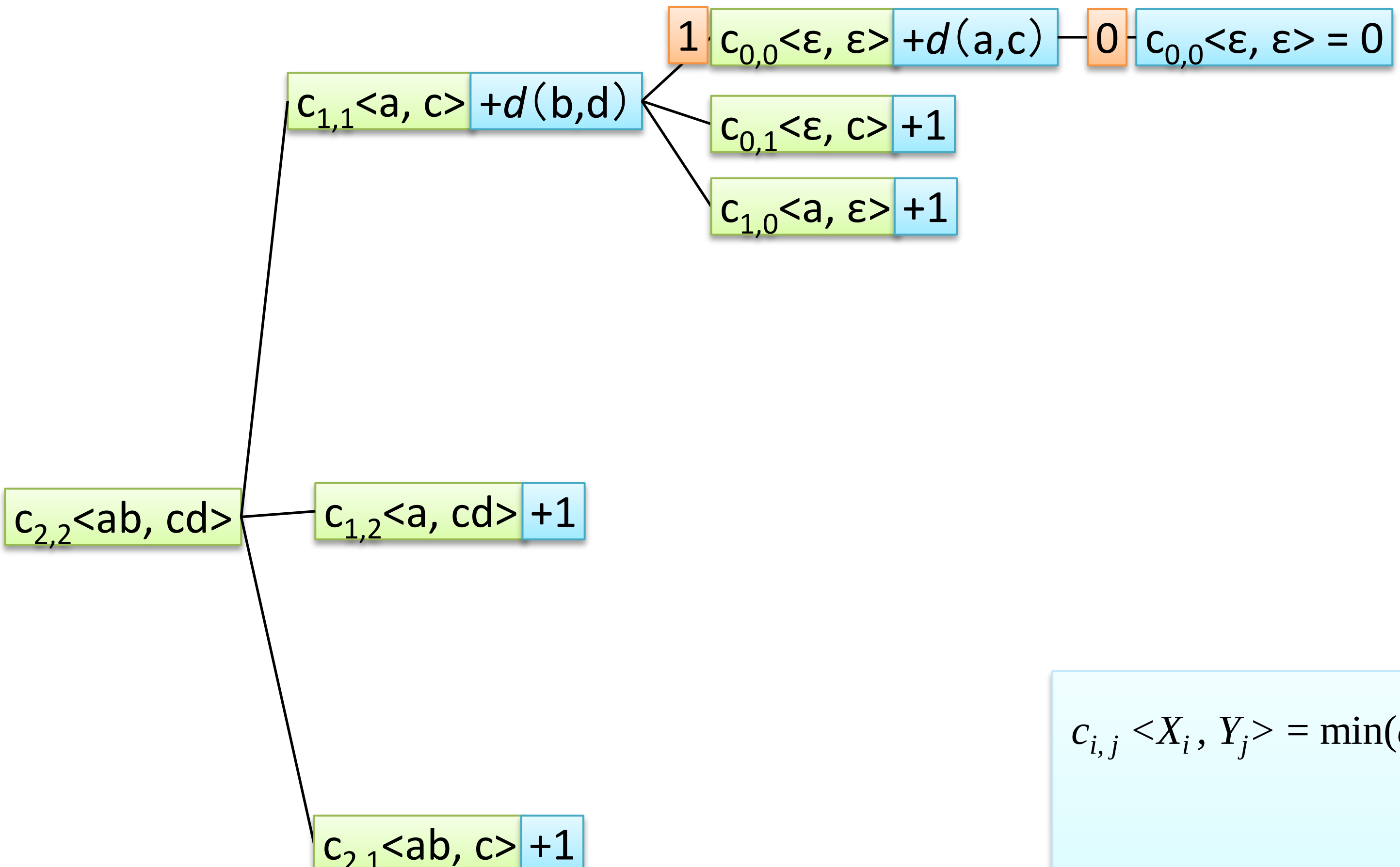
再帰的に計算



$$
\begin{aligned}
c_{i,j} \langle X_i, Y_j \rangle = \min(&c_{i-1,j-1} + d(x_i, y_j),\\
&c_{i-1,j} + 1,\\
&c_{i,j-1} + 1)
\end{aligned}
$$

$c_{i,0} \langle X_i, \varepsilon \rangle = i, \qquad c_{0,j} \langle \varepsilon, Y_j \rangle = j$

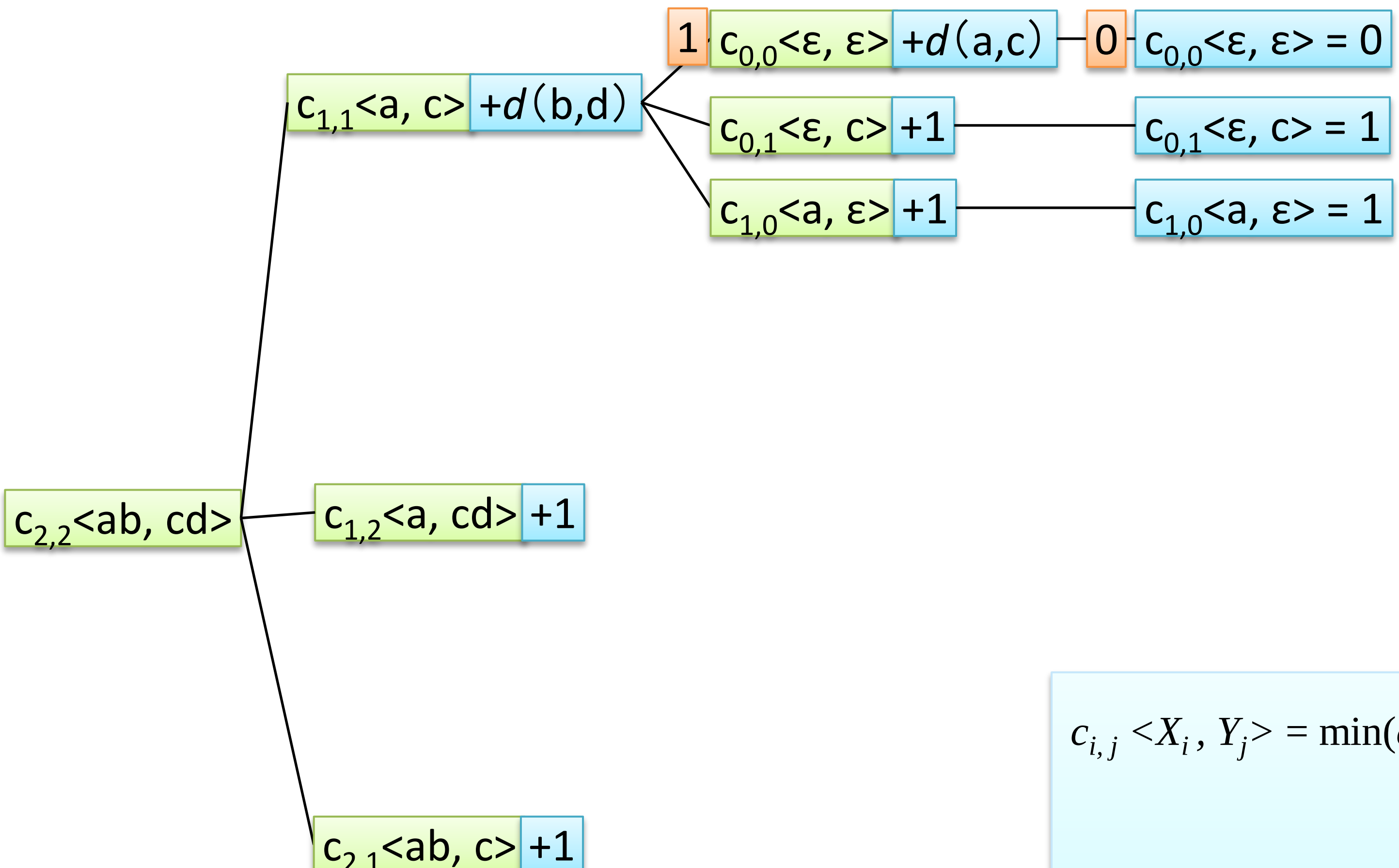
再帰的に計算



$$c_{i,j}\langle X_i, Y_j\rangle = \min(c_{i-1,j-1} + d(x_i, y_j),\; c_{i-1,j} + 1,\; c_{i,j-1} + 1)$$

$$c_{i,0}\langle X_i, \varepsilon\rangle = i, \qquad c_{0,j}\langle \varepsilon, Y_j\rangle = j$$

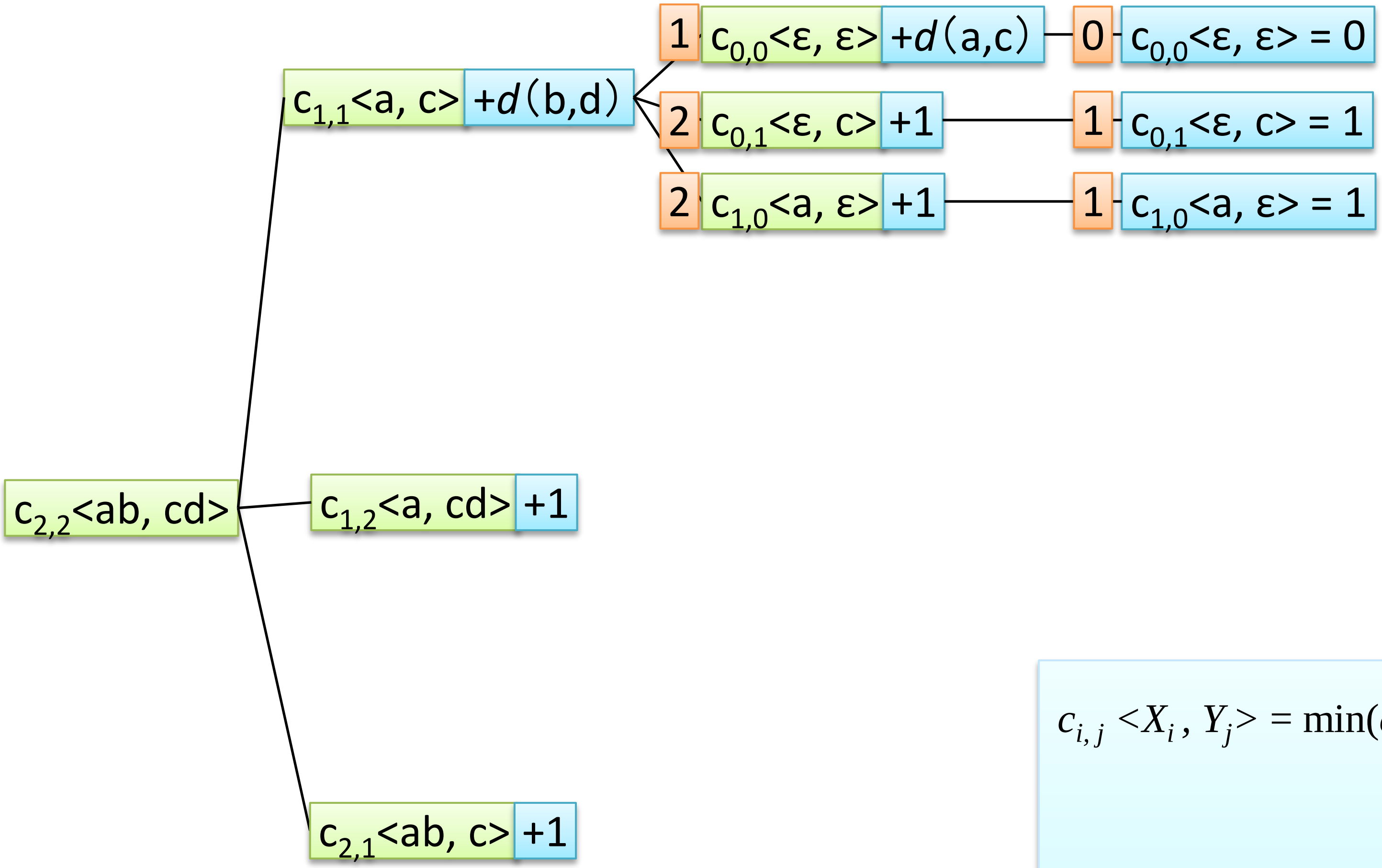
再帰的に計算



$$
c_{i,j} < X_i, Y_j > = \min(c_{i-1,j-1} + d(x_i, y_j), \quad c_{i-1,j} + 1, \quad c_{i,j-1} + 1)
$$

$$
c_{i,0} < X_i, \varepsilon > = i, \qquad c_{0,j} < \varepsilon, Y_j > = j
$$

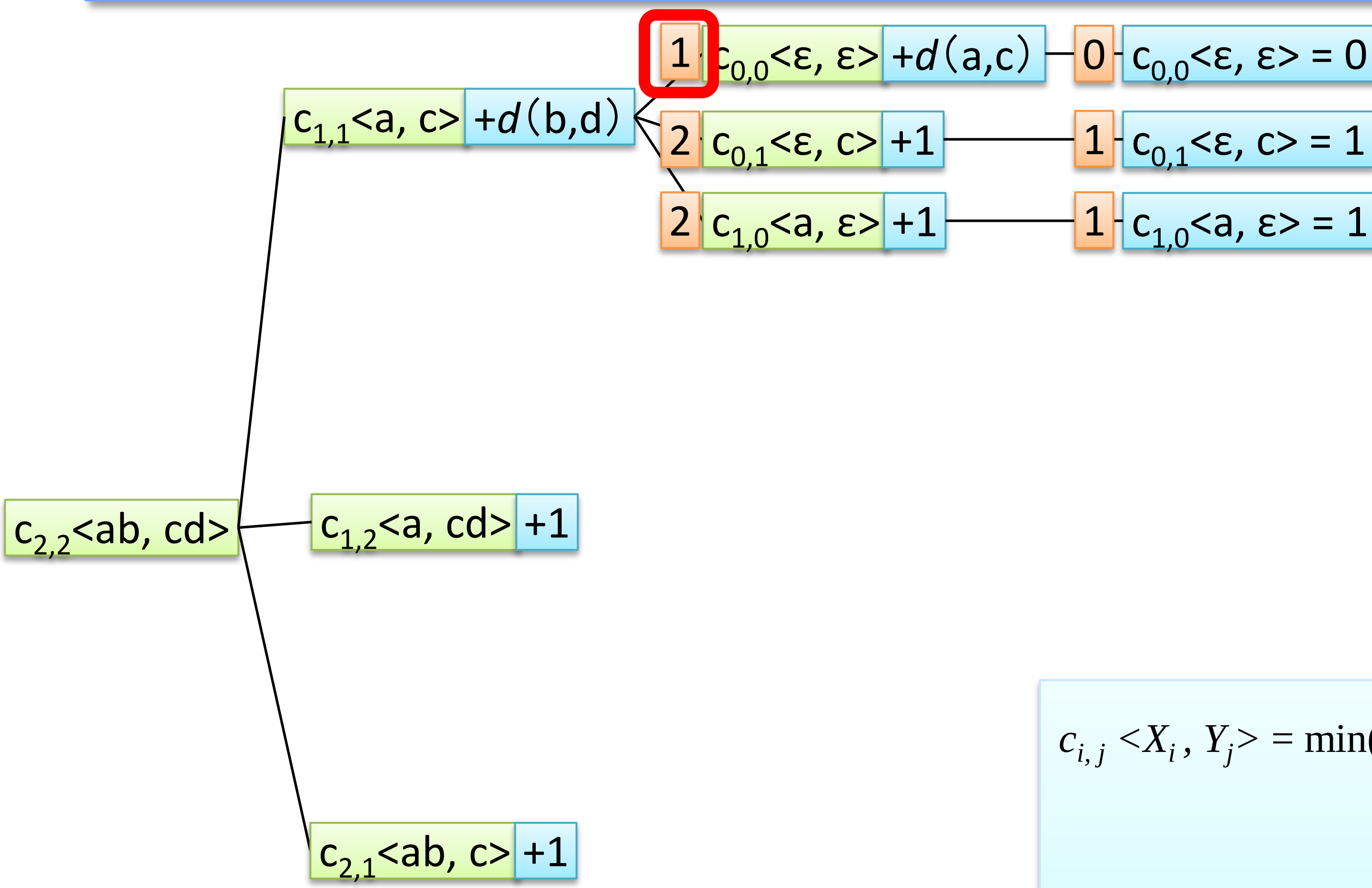
再帰的に計算



$$c_{i,j} <X_i, Y_j> = \min(c_{i-1,j-1} + d(x_i, y_j),\; c_{i-1,j} + 1,\; c_{i,j-1} + 1)$$

$$c_{i,0} <X_i, \varepsilon> = i, \qquad c_{0,j} <\varepsilon, Y_j> = j$$

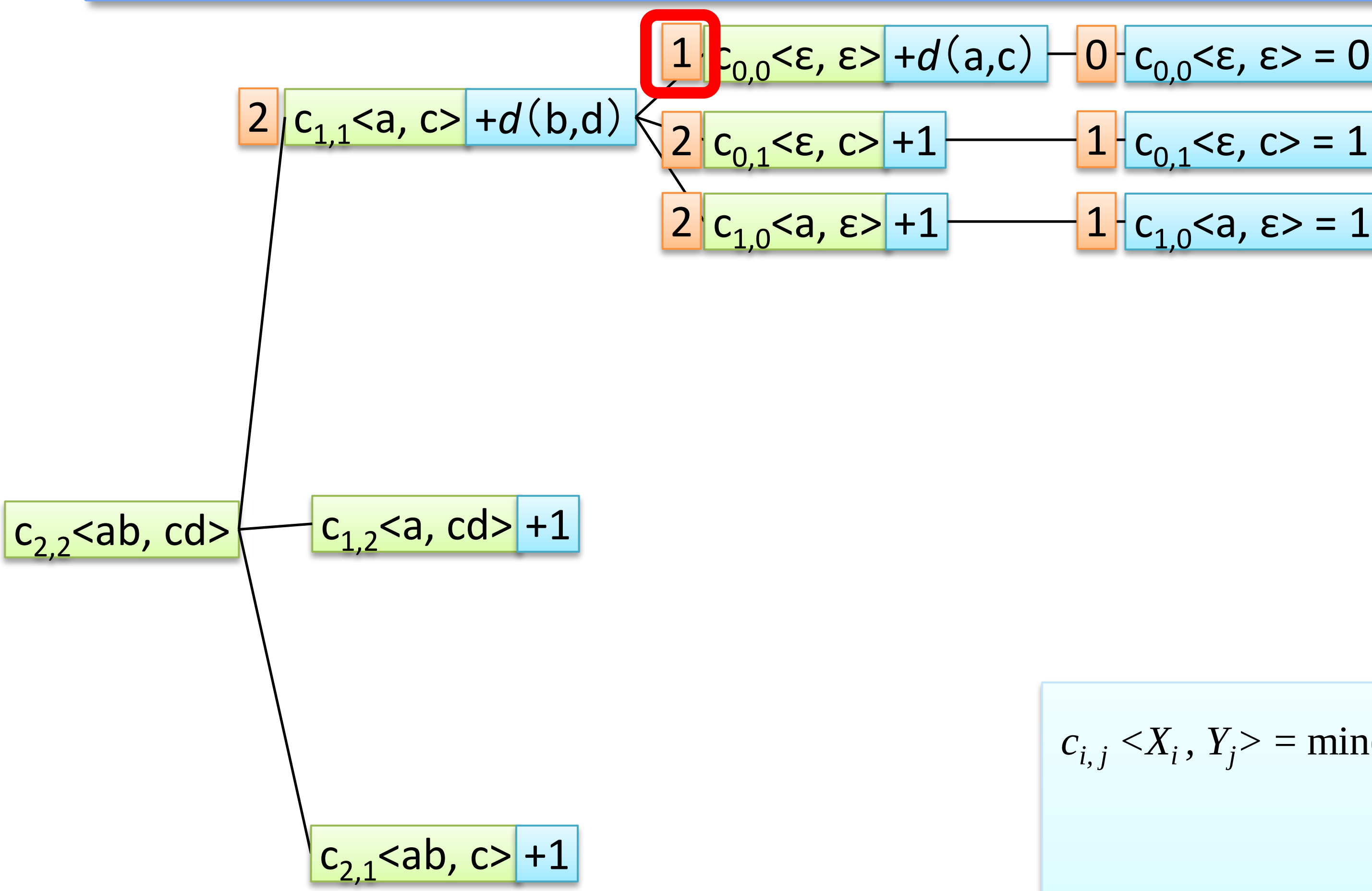
再帰的に計算



$$
\begin{aligned}
c_{i,j} \langle X_i, Y_j \rangle = \min(&c_{i-1,j-1} + d(x_i, y_j),\\
&c_{i-1,j} + 1,\\
&c_{i,j-1} + 1)
\end{aligned}
$$

$c_{i,0} \langle X_i, \varepsilon \rangle = i,$ $\quad c_{0,j} \langle \varepsilon, Y_j \rangle = j$

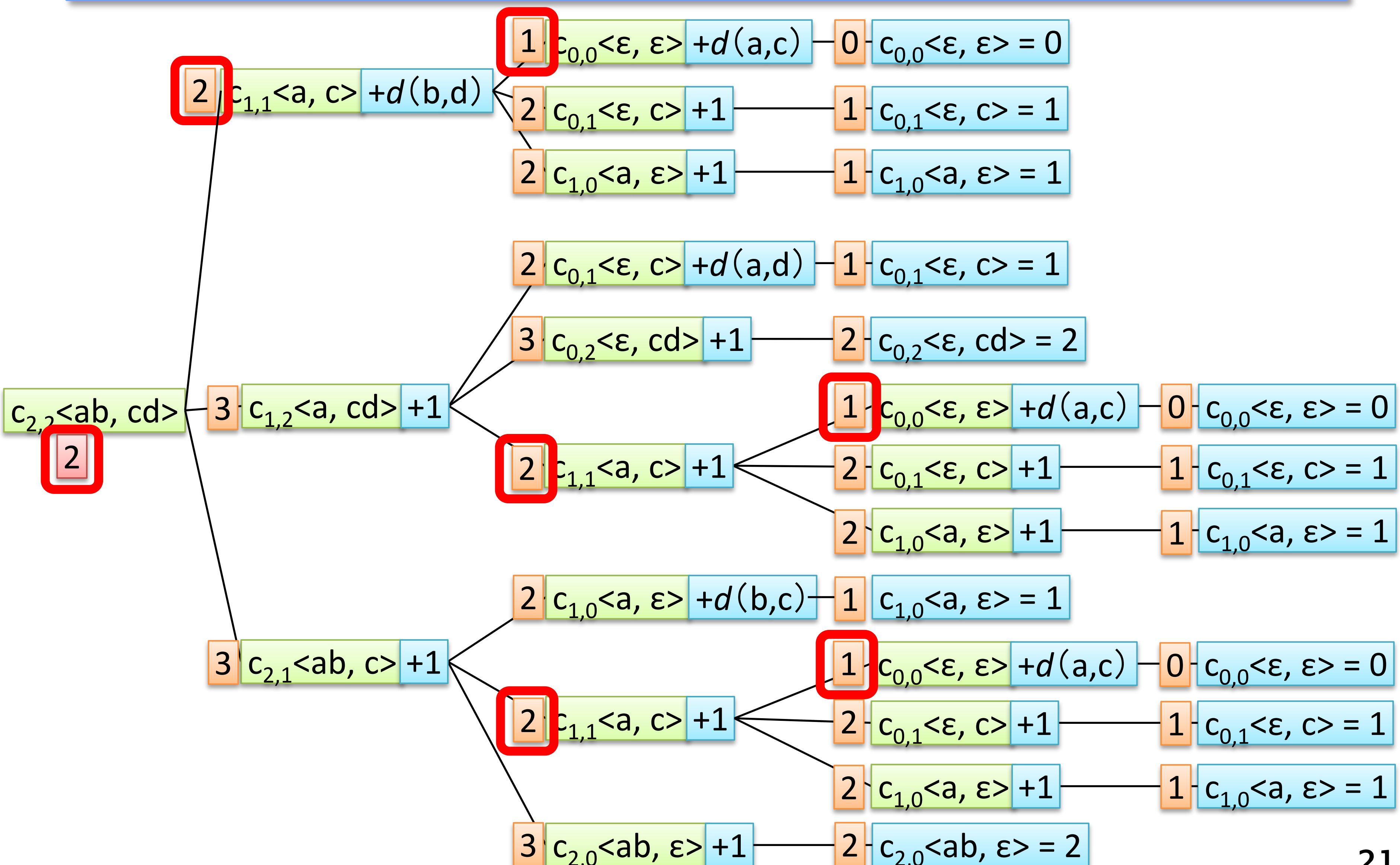
再帰的に計算



$$c_{i,j} <X_i, Y_j> = \min(c_{i-1,j-1} + d(x_i, y_j),\ c_{i-1,j} + 1,\ c_{i,j-1} + 1)$$

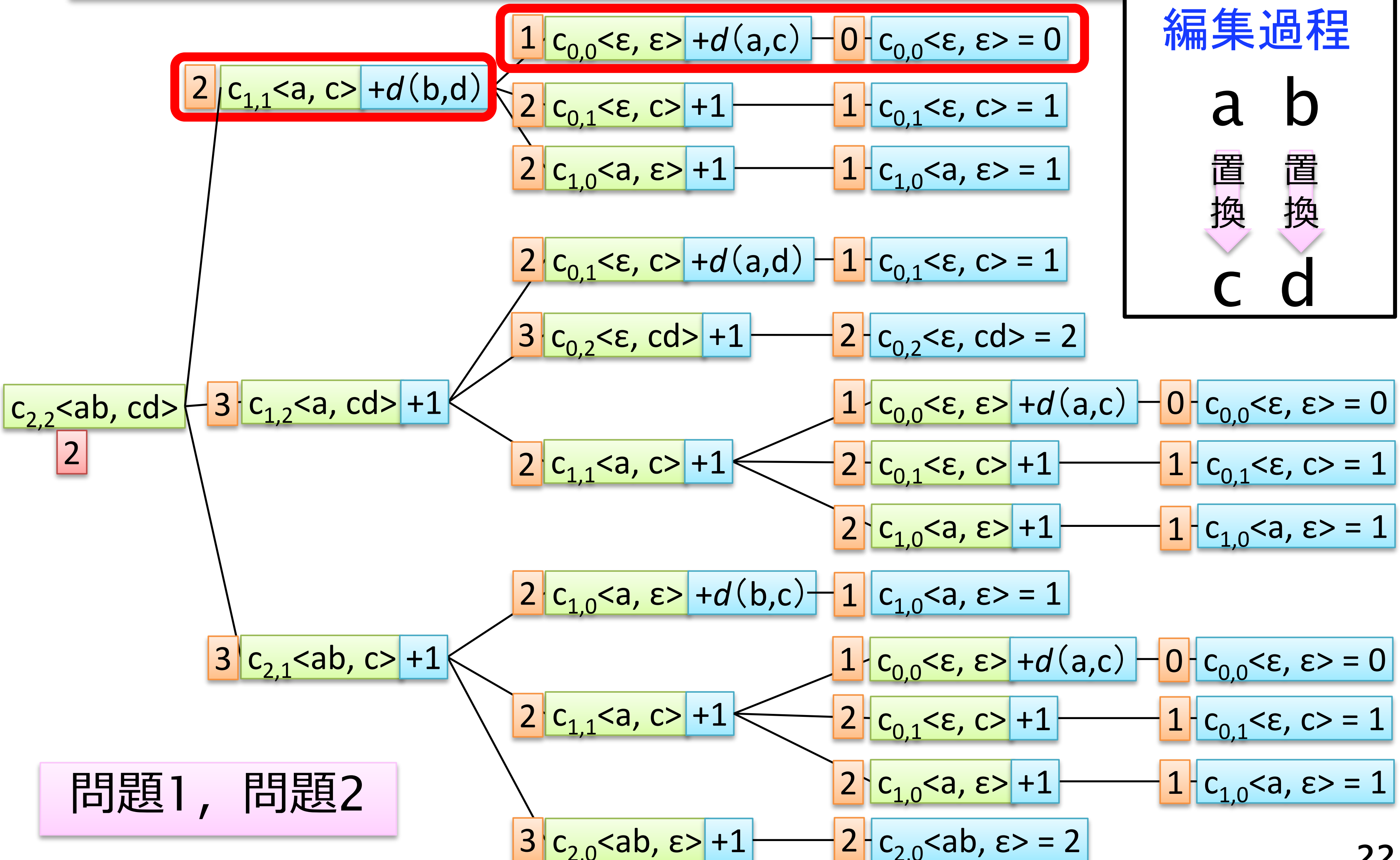
$$c_{i,0} <X_i, \varepsilon> = i, \qquad c_{0,j} <\varepsilon, Y_j> = j$$

再帰的に計算



$c_{2,2}$<ab, cd>
2
2 $c_{1,1}$<a, c> +*d*(b,d)
1 $c_{0,0}$<ε, ε> +*d*(a,c)
0 $c_{0,0}$<ε, ε> = 0
2 $c_{0,1}$<ε, c> +1
1 $c_{0,1}$<ε, c> = 1
2 $c_{1,0}$<a, ε> +1
1 $c_{1,0}$<a, ε> = 1
3 $c_{1,2}$<a, cd> +1
2 $c_{0,1}$<ε, c> +*d*(a,d)
1 $c_{0,1}$<ε, c> = 1
3 $c_{0,2}$<ε, cd> +1
2 $c_{0,2}$<ε, cd> = 2
2 $c_{1,1}$<a, c> +1
1 $c_{0,0}$<ε, ε> +*d*(a,c)
0 $c_{0,0}$<ε, ε> = 0
2 $c_{0,1}$<ε, c> +1
1 $c_{0,1}$<ε, c> = 1
2 $c_{1,0}$<a, ε> +1
1 $c_{1,0}$<a, ε> = 1
3 $c_{2,1}$<ab, c> +1
2 $c_{1,0}$<a, ε> +*d*(b,c)
1 $c_{1,0}$<a, ε> = 1
2 $c_{1,1}$<a, c> +1
1 $c_{0,0}$<ε, ε> +*d*(a,c)
0 $c_{0,0}$<ε, ε> = 0
2 $c_{0,1}$<ε, c> +1
1 $c_{0,1}$<ε, c> = 1
2 $c_{1,0}$<a, ε> +1
1 $c_{1,0}$<a, ε> = 1
3 $c_{2,0}$<ab, ε> +1
2 $c_{2,0}$<ab, ε> = 2

再帰的に計算



編集距離を
与える
編集過程

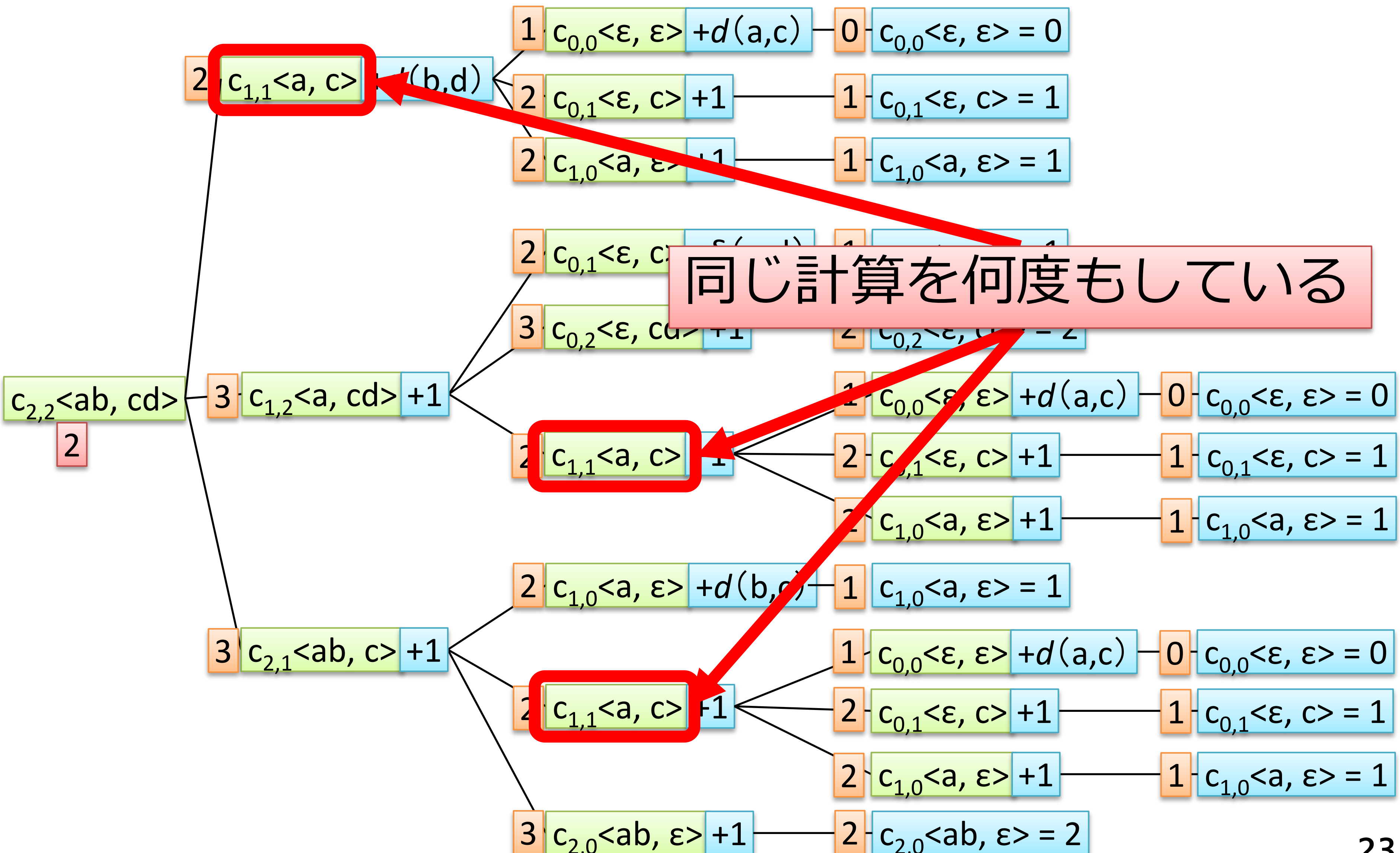
a b

置換 置換

c d

問題1，問題2

再帰的計算の問題点



c2,2<ab, cd>
2
3 c1,2<a, cd> +1
3 c2,1<ab, c> +1
1 c0,0<ε, ε> +d(a,c)
0 c0,0<ε, ε> = 0
2 c0,1<ε, c> +1
1 c0,1<ε, c> = 1
1 c1,0<a, ε> = 1
2 c1,1<a, c>
同じ計算を何度もしている
2 c1,0<a, ε> +1
1 c1,0<a, ε> = 1
3 c2,0<ab, ε> +1
2 c2,0<ab, ε> = 2

再帰的計算の問題点

- **同じ計算を何度も行うため計算量大**

⇒ メモ化（memoization）による対策

一度計算したものをメモとして記憶，次に計算するときはそのメモを呼び出す

問題5

動的計画法(Dynamic Programing, DP)

$c_{i,j}$<ab, cd>

	ε	c	d
ε	$C_{0,0}$	$C_{0,1}$	$C_{0,2}$
a	$C_{1,0}$	$C_{1,1}$	$C_{1,2}$
b	$C_{2,0}$	$C_{2,1}$	$C_{2,2}$

$$\begin{aligned} c_{i,j} <X_i, Y_j> = \min(&c_{i-1,j-1} + d(x_i, y_j), \\ &c_{i-1,j} + 1, \\ &c_{i,j-1} + 1) \end{aligned}$$

$$c_{i,0} <X_i, \varepsilon> = i,\ c_{0,j} <\varepsilon, Y_j> = j$$

$C_{0,0}$、$C_{0,1}$、$C_{1,0}$が計算済みならば、$C_{1,1}$の値分かる

動的計画法(Dynamic Programing, DP)

$c_{i,j}$<ab, cd>

	ε	c	d
ε	$C_{0,0}$	$C_{0,1}$	$C_{0,2}$
a	$C_{1,0}$	$C_{1,1}$	$C_{1,2}$
b	$C_{2,0}$	$C_{2,1}$	$C_{2,2}$

$$c_{i,j} <X_i, Y_j> = \min(c_{i-1,j-1} + d(x_i, y_j),\; c_{i-1,j} + 1,\; c_{i,j-1} + 1)$$

$$c_{i,0} <X_i, \varepsilon> = i,\ c_{0,j} <\varepsilon, Y_j> = j$$

$C_{1,1}$、$C_{1,2}$、$C_{2,1}$が計算済みならば、$C_{2,2}$の値分かる

動的計画法(Dynamic Programing, DP)

$c_{i,j}$<ab, cd>

	ε	c	d
ε	$C_{0,0}$	$C_{0,1}$	$C_{0,2}$
a	$C_{1,0}$	$C_{1,1}$	$C_{1,2}$
b	$C_{2,0}$	$C_{2,1}$	$C_{2,2}$

$$c_{i,j} <X_i, Y_j> = \min(c_{i-1,j-1} + d(x_i, y_j), \; c_{i-1,j} + 1, \; c_{i,j-1} + 1)$$

$$c_{i,0} <X_i, \varepsilon> = i, \; c_{0,j} <\varepsilon, Y_j> = j$$

ボトムアップに$C_{i,j}$を求める

↓

編集距離$C_{2,2}$得られる

同じ計算は1度のみ ⇒ 計算量減

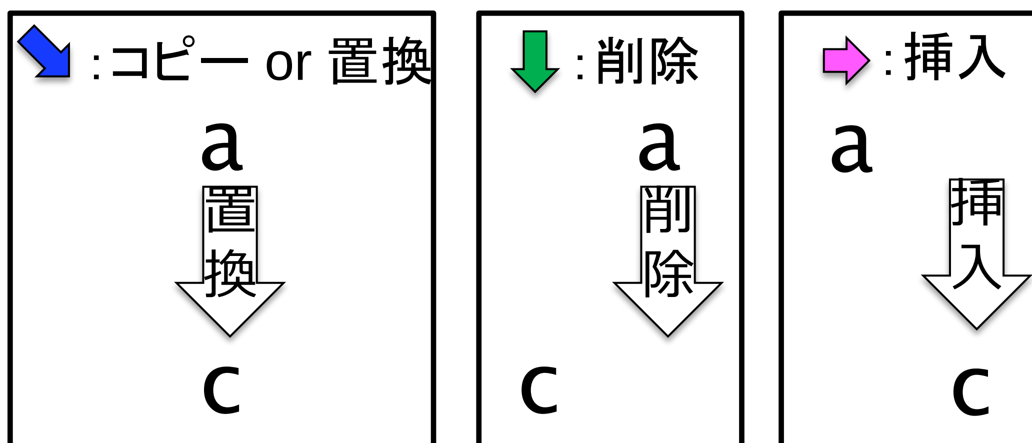
動的計画法(Dynamic Programing, DP)

$c_{i,j}$<ab, cd>

	ε	c	d
ε	$c_{0,0}$	$c_{0,1}$	$c_{0,2}$
a	$c_{1,0}$	$c_{1,1}$	$c_{1,2}$
b	$c_{2,0}$	$c_{2,1}$	$c_{2,2}$

$$c_{i,j} <X_i, Y_j> = \min(c_{i-1,j-1} + d(x_i, y_j),\; c_{i-1,j} + 1,\; c_{i,j-1} + 1)$$

$$c_{i,0} <X_i, \varepsilon> = i,\; c_{0,j} <\varepsilon, Y_j> = j$$



動的計画法による計算

$c_{i,j}$<ab, cd>

	ε	c	d
ε			
a			
b			

$$c_{i,j} \langle X_i, Y_j \rangle = \min(c_{i-1,j-1} + \delta(x_i, y_j), \; c_{i-1,j} + 1, \; c_{i,j-1} + 1)$$

$$c_{i,0} \langle X_i, \varepsilon \rangle = i, \quad c_{0,j} \langle \varepsilon, Y_j \rangle = j$$

動的計画法による計算

$c_{i,j}$<ab, cd>

	ε	c	d
ε	0 →	1 →	2
a			
b			

$$c_{i,j} \langle X_i, Y_j \rangle = \min(c_{i-1,j-1} + \delta(x_i, y_j), \; c_{i-1,j} + 1, \; c_{i,j-1} + 1)$$

$c_{i,0} \langle X_i, \varepsilon \rangle = i$, $c_{0,j} \langle \varepsilon, Y_j \rangle = j$

$c_{0,0}\langle \varepsilon, \varepsilon \rangle = 0$, $c_{0,1}\langle \varepsilon, c \rangle = 1$, $c_{0,2}\langle \varepsilon, cd \rangle = 2$

挿入

動的計画法による計算

$c_{i,j}$<ab, cd>

	ε	c	d
ε	0	1	2
a	1		
b	2		

$$c_{i,j} \langle X_i, Y_j \rangle = \min(c_{i-1,j-1} + \delta(x_i, y_j), \; c_{i-1,j} + 1, \; c_{i,j-1} + 1)$$

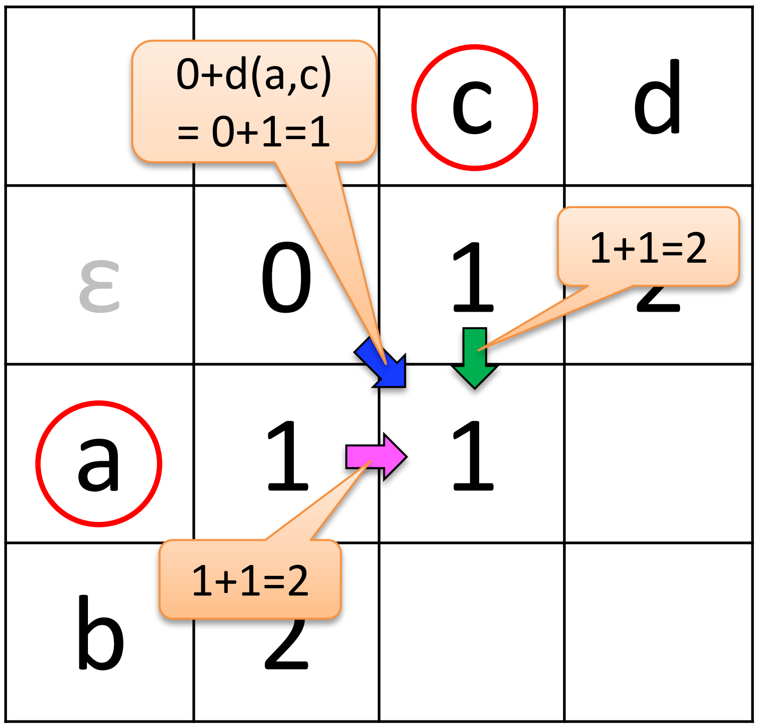
$$c_{i,0} \langle X_i, \varepsilon \rangle = i, \; c_{0,j} \langle \varepsilon, Y_j \rangle = j$$

$c_{0,0}\langle \varepsilon, \varepsilon \rangle = 0$, $c_{1,0}\langle a, \varepsilon \rangle = 1$, $c_{2,0}\langle ab, \varepsilon \rangle = 2$

削除

動的計画法による計算

$c_{i,j}$<ab, cd>



$$
\begin{aligned}
c_{i,j} <X_i\,,\,Y_j> = \min(&c_{i-1,j-1} + d(x_i, y_j),\\
&c_{i-1,j} + 1,\\
&c_{i,j-1} + 1)
\end{aligned}
$$

$$c_{i,0} <X_i, \varepsilon> = i,\ c_{0,j} <\varepsilon, Y_j> = j$$

$$
\begin{aligned}
c_{1,1}<\mathrm{a}, \mathrm{c}> = \min(&c_{0,0}<\varepsilon, \varepsilon> + d(\mathrm{a}, \mathrm{c}), && \text{置換}\\
&c_{0,1}<\varepsilon, \mathrm{c}> + 1, && \text{削除}\\
&c_{1,0}<\mathrm{a}, \varepsilon> + 1) && \text{挿入}
\end{aligned}
$$

動的計画法による計算

$c_{i,j}$<ab, cd>

	ε		d
ε	0	1	2
a	1	1	2
b	2		

1+d(a,d) = 1+1=2

2+1=3

1+1=2

$$c_{i,j}\langle X_i, Y_j\rangle = \min(c_{i-1,j-1} + d(x_i, y_j),\ c_{i-1,j} + 1,\ c_{i,j-1} + 1)$$

$$\langle X_i, \varepsilon\rangle = i,\ c_{0,j}\langle \varepsilon, Y_j\rangle = j$$

$c_{1,2}\langle a, cd\rangle = \min(c_{0,1}\langle \varepsilon, c\rangle + d(a, d),$ 置換
$c_{0,2}\langle \varepsilon, cd\rangle + 1,$ 削除
$c_{1,1}\langle a, c\rangle + 1)$ 挿入

動的計画法による計算

$c_{i,j}\langle ab, cd\rangle$

	ε	c	d
ε		1	2
a	1	1	2
b	2	2	

1+d(b,c)
= 1+1=2

1+1=2

2+1=3

$$c_{i,j}\langle X_i, Y_j\rangle = \min(c_{i-1,j-1} + d(x_i, y_j),\ c_{i-1,j} + 1,\ c_{i,j-1} + 1)$$

$$c_{i,0}\langle X_i, \varepsilon\rangle = i,\ c_{0,j}\langle \varepsilon, Y_j\rangle = j$$

$$c_{2,1}\langle ab, c\rangle = \min(c_{1,0}\langle a, \varepsilon\rangle + d(b, c),\ \text{置換}\quad c_{1,1}\langle a, c\rangle + 1,\ \text{削除}\quad c_{2,0}\langle ab, \varepsilon\rangle + 1)\ \text{挿入}$$

動的計画法による計算

$c_{i,j}$<ab, cd>

	ε	c	d
ε	0		2
a	1	1	2
b	2	2	2

1+d(b,d) = 1+1=2

2+1=3

2+1=3

$$c_{i,j} <X_i, Y_j> = \min(c_{i-1,j-1} + d(x_i, y_j),\\ c_{i-1,j} + 1,\\ c_{i,j-1} + 1)$$

$$c_{i,0} <X_i, \varepsilon> = i, \; c_{0,j} <\varepsilon, Y_j> = j$$

$c_{2,2}<ab, cd> = \min(c_{1,1}<a, c> + d(b, d),$置換
$c_{1,2}<a, cd> + 1,$削除
$c_{2,1}<ab, c> + 1)$挿入

動的計画法による計算

$c_{i,j}$<ab, cd>

	ε	c	d
ε	0	1	2
a	1	1	2
b	2	2	2

ここが編集距離

$$c_{i,j} \langle X_i, Y_j \rangle = \min(c_{i-1,j-1} + d(x_i, y_j), \; c_{i-1,j} + 1, \; c_{i,j-1} + 1)$$

$$c_{i,0} \langle X_i, \varepsilon \rangle = i, \; c_{0,j} \langle \varepsilon, Y_j \rangle = j$$

問題3，問題4

動的計画法による計算

$c_{i,j}<ab, cd>$

	ε	c	d
ε	0	1	2
a	1	1	2
b	2	2	2

編集操作の求め方

値を埋める際に使ったデータの流れの矢印を表示

動的計画法による計算

$c_{i,j}$<ab, cd>

	ε	c	d
ε	0	1	2
a	1	1	2
b	2	2	2

編集操作の求め方

矢印を180°回転

動的計画法による計算

$c_{i,j}$<ab, cd>

	ε	c	d
ε	0	1	2
a	1	1	2
b	2	2	2

編集操作の求め方

右下から０の部分までたどっていったときの矢印が編集操作（ただし，編集操作が１つに決まるとは限らない）

a　b
置換　置換
c　d

R　R

問題6

課題3：全部で7問

- 問題1（5点）　[記述問題]再帰による編集距離の計算（手書き可）
- 問題2（10点）　再帰関数による実装
- 問題3（10点）　[記述問題]動的計画法による編集距離の計算（手書き可）

2週目の面接のみ受付

- 問題4（10点）　動的計画法による実装
- 問題5（15点）　[プログラム＋記述問題]再帰呼び出しにおけるメモ化
- 問題6（10点）　編集操作の表示
- 問題7（10点）　スペルチェッカーの作成

4週目の面接のみ受付

問題1

X=“on”, Y=“so”に対して, $c_{i,j}$の漸化式を手作業で再帰的に適用することにより$c_{2,2}$の値を求めよ.

◎ $c_{2,2}$を求める途中経過を省略することなく, すべてレポートに記すこと(課題説明スライドと同様な木構造を使って記すこと).

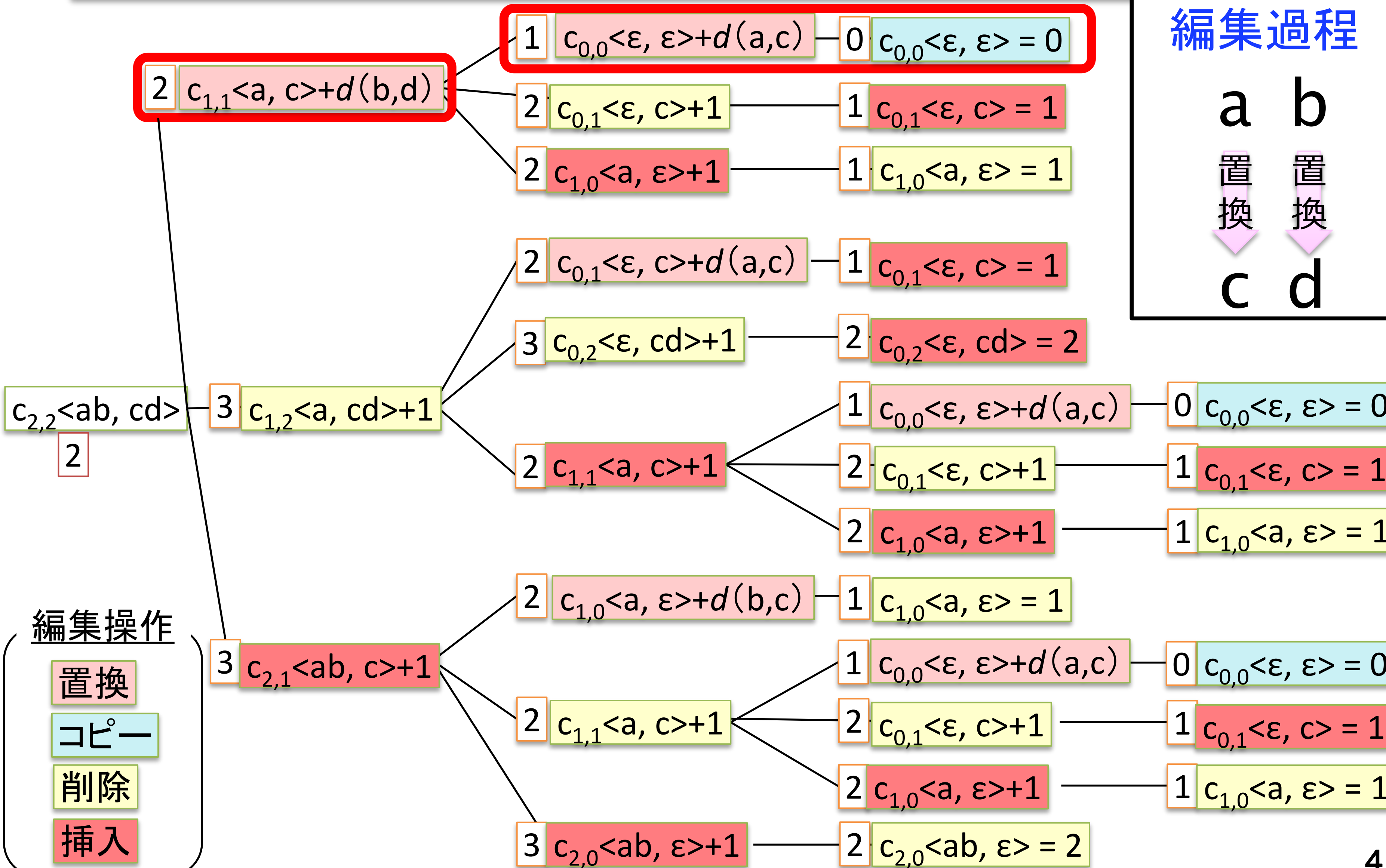
◎ 注意 : 以下の**3つすべて**を記せ

- $c_{i,j}$の値を求めるのに使った接頭辞の編集距離の値とその添字.
- 編集操作が分かるように明示せよ.
- 編集距離を与える編集過程をすべて示せ.

問題1

編集距離を
与える
編集過程

a b
置換 置換
c d



面接を受けるときの注意事項

- **C言語の用語や関数の動作をあらかじめ調べてくること**
  プログラムのソースコードに説明を書いてかまわないので,面接時に説明できるようにすること
- **課題に関する問題文をきちんと読むこと**
  問題文を読むことが解答への近道になることもあります
- **2週目の面接時に,途中だったとしても,取り組んでい る問題を持ってきましょう**
  困っていることやわからないことがあれば，面接の時にアドバイスをするので，途中のプログラムでも 遠慮なく持ってきましょう