

## 根がなく順序がない木の間の距離の一計算法

六車智一, 田中榮一, 増田澄男

神戸大学工学部

あらまし

これまでの木の間の距離の研究の大部分は、根があり、順序がある木に関するものであった。本文では、木の頂点の置換、挿入、脱落操作を基に、根もなく、順序もない木の間の距離を定義し、その下降形計算法を提案する。2つの木の頂点数を  $m, n$  とするとき、本方法の計算の複雑さは  $O\tau(mn) \sim O\tau(m^3(n-1)(n-2)(n-3) \cdots (n-m+1))$  である。また、このとき、空間複雑さは、 $O_s(mn)$  である。平均原子個数 12.0 個の化合物を計算機（ワークステーション SONY NWS-1860(5 MIPS)）で計算したところ、平均計算時間は 4.3 秒であった。この木の間の距離は化合物の分類に応用できる。

和文キーワード 木の距離、下降形計算法、化合物、分類、検索、編集操作

## A metric between trees with no root and no order and its computing method

Tomokazu MUGURUMA, Eiichi TANAKA, and Sumio MASUDA

Faculty of Engineering, Kobe University

### Abstract

This paper describes a tree metric between unordered trees with no root based on three edit operations on vertices, that is, substitution, insertion and deletion. A top-down computing method is proposed. The computational complexity is  $O\tau(mn) \sim O\tau(m^3(n-1)(n-2)(n-3) \cdots (n-m+1))$ , where  $m$  and  $n$  are the number of vertices of tree  $T_a$  and that of tree  $T_b$ , respectively. The space complexity is  $O_s(mn)$ . The computing time on the calculator(SONY NWS-1860,5 MIPS) is 4.3 seconds on average for chemical compounds which have 12 atoms on average. This metric can be applied to the classification of chemical compounds.

英文 key words tree metric, top-down method, chemical compound, classification, retrieval, edit operation

## 1. まえがき

これまで、木の間の距離の研究はほとんどが根があり、順序がある木に関するものであった。本文では根も無く、順序もない木の間の1つの距離を定義し、その下降形計算法について述べる。

木の間の距離(以後、木の距離という)は一次元の文字列間の距離の研究<sup>1)-6)</sup>の自然な拡張として現れた。Selkow<sup>7)</sup>は根があり、順序がある木について、レベルが同じ頂点の間の対応だけを許した木の距離を考え、Tai<sup>8)</sup>とLu・Fu<sup>9)</sup>はレベルが異なる頂点間の対応をも認めた距離に拡張している。青木<sup>10)</sup>と田中<sup>11)</sup>はTai距離の能率の良い計算法を提案している。また、田中・田中<sup>12)</sup>は構造を保存する写像に基づく木の距離(SPM距離)を定義した。大森<sup>13)</sup>はこれららの距離が同一形式の公式で表されることを示し、同時に弱構造保存写像に基づく木の類似度(WSPM類似度)を定義した。木の変換操作に基づく中林・鎌田<sup>14)</sup>の2分木間の距離は写像の立場から見てどのような距離が明らかでなかったが、大森はWSPM類似度が距離になる特殊な場合であることを明らかにした。Lu<sup>15)</sup>は頂点の分離、融合操作を基に木の距離を定義し、地震波の解析に応用している<sup>16)</sup>。さらに、Wilhelm<sup>17)</sup>は根はあるが順序がない木の距離を定義し、Culik II・Wood<sup>18)</sup>は根もなく、順序もない木の距離を最近接変換操作の必要回数で定義している。グラフの距離についても幾つかの報告がある<sup>19)-21)</sup>。頂点の分離・融合操作に基づく木の距離は計算法が単純ではない<sup>22)</sup>。そこで頂点の変換、挿入、脱落操作に基づく木の距離をグラフの距離に拡張することを考えると、木の距離の研究とグラフ距離の研究の間にはかなり開きがある。すなわち、根があり、順序がある木の距離からグラフの距離へ繋ぐためには、根はないが、順序がある木の距離、根もなく、順序もない木の距離を定義しなければならない。一方、類似化合物を検索するためには、構造式間の距離あるいは類似度を定義し、それらの尺度に基づいた化合物ファイルを作らなければならない。化合物の構造式の類似度を表す尺度はいろいろ提案されている<sup>23)</sup>が、構造式を表すグラフをありのまま取り扱う方法はなかった。構造式を表すグラフは、根もなく、順序もない木で近似できる。本文では、根もなく、順序もない木の距離を頂点の変換、挿入、脱落操作を基に定義し、木を道の集合とみないたときの計算法を提案する。この距離は系統樹作成の立場から考えられたCulik II・Wood<sup>18)</sup>の距離とは異なるものである。

## 2. 諸定義

根がなく、順序のない木を  $T = (V, E)$  で表す。ここで、 $V$  は頂点の集合、 $E$  は辺の集合である。頂点にはラベルがあるものとする。頂点数を  $|V|$  と書く。本文で単に木と呼ぶときはこの木を指すものとする。 $T$  を図 1 (a) の木とするとき、図 1 (b) の木も同じ木である。木の頂点を区別するために頂点に番号を付ける。以後、頂点は番号で表すものとする。木の1つの頂点  $u$  を根として持つ木(*a rooted tree*: 以後 R木という)を  $T(u)$  と書き、 $T(u) = (V(u), E(u))$  と表す。図 1 (c) は図 1 (a) の木から得られた R木  $T(1)$  である。図 1 (d) も  $T(1)$  である。順序のない木であるから、図 1 (c) の木と図 1 (d) の木は同じ木を表している。図 1 (c) の頂点のラベルは図 1 (e) のようである。図 1 (e) で頂点 1 のラベルは  $e$ 、頂点 2 のラベルは  $a$  である。頂点  $v$  の次数を  $\deg(v)$  と書く。 $\deg(v)=1$  のとき  $v$  を葉という。 $L = \{v | v \in V, \deg(v)=1\}$ 、 $L(u) = \{v | v \in V(u), \deg(v)=1, v \neq u\}$  とすると、 $L$  は木

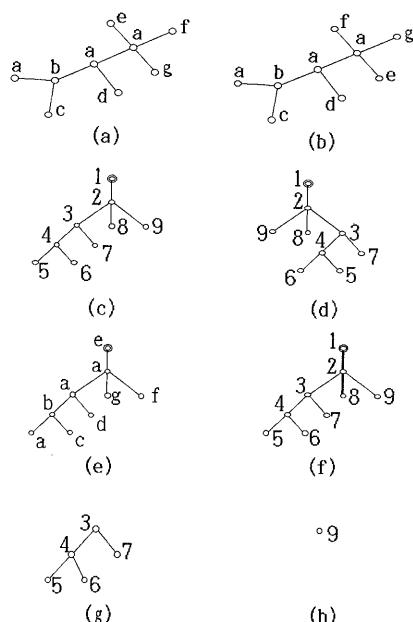


図 1. (a) 木  $T$ 、(b) 木  $T$ 、(c) R木  $T(1)$ 、(d) R木  $T(1)$ 、  
(e) R木  $T(e)$ 、(f) R木  $T(1)$  と幹  $tr(1, 8)$ 、  
(g) 分岐部分木  $subt(2, 3)$ 、(h)  $subt(2, 9)$ 。

$T$  の葉の集合、 $L(u)$  は R木  $T(u)$  の葉の集合である。木  $T_a$  と木  $T_b$  のそれぞれの葉の集合を  $L_a$ 、 $L_b$  と書く。他の記号についても同様に  $a$ 、 $b$  を添字に用いて、それぞれ  $T_a$ 、 $T_b$  に関する記号であることを表す。 $T(u)$  の根  $u$  から葉  $v$

に至る道を $P(u, v)$ と書く。 $P(u, v)$ の上の異なる2つの頂点 $x, y$ について、 $x$ が $y$ より根に近いとき、 $x$ は $y$ の祖先、 $y$ は $x$ の子孫と呼び、 $y$ の祖先の集合を $An(y)$ と書く。 $x$ は $x$ の祖先でも子孫でもないものとする。 $y$ が $x$ に隣接する子孫であるとき、 $y$ を $x$ の子という。 $x \in An(y)$ かつ $y \in An(x)$ のとき、 $x$ と $y$ は分離しているといい、 $Sep(x, y)$ と書く。図1(c)の $T(1)$ で、2は7の祖先、6は3の子孫、 $An(7)=\{1, 2, 3\}$ である。5と6は4の子であり、 $Sep(5, 6), Sep(5, 7)$ である。頂点 $z$ とその子孫が作る木を $z$ を根とする部分木と呼ぶ。 $z$ を根とし、 $x$ を含まない部分木を $T(x, z)$ と書く。根 $u$ から指定された葉 $v$ に至る1つの道を $T(u)$ の幹と呼び、 $tr(u, v)$ と書く。幹は部分木にも定義する。 $T(u)$ の幹 $tr(u, v)$ 上の頂点 $s$ に隣接する子 $t$ が幹上にないとき、 $t$ を根とする $T(u)$ の部分木を $s$ に関する分歧部分木と呼び、 $subt(s, t)$ と書く。図1(f)の $T(1)$ で、幹を $tr(1, 8)$ と選ぶと、頂点2に関する分歧部分木は図1(g)に示す $subt(2, 3)$ と図1(h)の $subt(2, 9)$ である。R木の2つの頂点 $u, v$ の共通の祖先のうち最も近いものを $u$ と $v$ の最も近い共通祖先(the closest common ancestor)と呼び、 $Ccan(u, v)$ と書く。図1(c)の $T(1)$ で $Ccan(5, 7)$ は3である。

2つの文字列 $A=a_1 \dots a_m, B=b_1 \dots b_n$ 間に定義された距離に重み付き $\nu$ - $\pi$ -シグマイン距離(WLD)がある。 $a_i$ が $b_j$ に対応しているとき $(i, j)$ と書き、 $(i, j)$ の集合 $M$ が次の条件を満たしているときAからBへの写像と呼ぶ。任意の $(i_1, j_1), (i_2, j_2) \in M$ について

$$(a_1) i_1 = i_2 \text{ iff } j_1 = j_2.$$

$$(a_2) i_1 < i_2 \text{ iff } j_1 < j_2. \quad (1)$$

$M$ はAからBへの1つの変換を定めている。 $(i, j) \in M$ で $a_i \neq b_j$ のとき、 $a_i$ は $b_j$ に置換したという。 $I_1 = \{i \mid (i, j) \in M\}, J_1 = \{j \mid (i, j) \in M\}$ として、 $k \in I_1 (1 \leq k \leq m)$ のとき、 $ak$ は脱落したといい、 $h \in J_1 (1 \leq h \leq n)$ のとき、 $bh$ は挿入したといい。置換、挿入、脱落の重みを $p, q, r$ とするとき、AをBに変換するのに要する変換(置換、挿入、脱落)操作の重みの和の最小値をAからBへのWLDといい、 $WLD(A, B)$ と表す。 $p, q, r$ は非負で、本文では $p < q + r$ とする。

$WLD(A, B)$ は次式で計算できる。

初期条件

$$d(x, 0) = x \cdot r \quad (0 \leq x \leq m),$$

$$d(0, y) = y \cdot q \quad (0 \leq y \leq n). \quad (2)$$

計算式

$$d(i, j) = \min \{d(i-1, j-1) + p(i, j),$$

$$d(i-1, j) + r, d(i, j-1) + q\}. \quad (3)$$

$$WLD(A, B) = d(m, n). \quad (4)$$

ここで

$$p(i, j) = \begin{cases} 0, & a_i = b_j, \\ p, & a_i \neq b_j. \end{cases} \quad (5)$$

2つのR木を $T_a(u) = (V_a, E_a), T_b(v) = (V_b, E_b)$ とし、 $m=|V_a|, n=|V_b|$ とする。 $V_a$ の頂点には、1から $m$ の番号が、 $V_b$ の頂点には1から $n$ の番号が付けられているものとする。次の条件を満たす $(i, j)$  ( $i \in V_a, j \in V_b$ )の集合 $M_t$ を $T_a(u)$ から $T_b(v)$ への写像という。任意の $(i_1, j_1), (i_2, j_2) \in M_t$ について、

$$(b_1) \quad i_1 = i_2 \text{ iff } j_1 = j_2.$$

$$(b_2) (i) \quad i_1 \in An(i_2) \text{ iff } j_1 \in An(j_2).$$

$$(ii) \quad Sep(i_1, i_2) \text{ iff } Sep(j_1, j_2).$$

$$(iii) \quad Sep(i_1, i_2) \text{ あるいは } Sep(j_1, j_2) \text{ のとき,}$$

$$(i, j) \in M_t. \text{ ここで, } i = Ccan(i_1, i_2), j = Ccan(j_1, j_2). \quad (6)$$

$i, j$ のラベルを $a_i, b_j$ とするとき、 $(i, j) \in M_t$ で、 $a_i \neq b_j$ のとき、 $a_i$ は $b_j$ に置換したという。WLDのときと同様に $I_2 = \{i \mid (i, j) \in M_t\}, J_2 = \{j \mid (i, j) \in M_t\}$ を定義する。

$k \in I_2 (1 \leq k \leq m)$ のとき、 $ak$ (あるいは $k$ )は脱落したといい、 $h \in J_2 (1 \leq h \leq n)$ のとき $bh$ (あるいは $h$ )は挿入したといい。頂点のラベルの置換、ラベル付き頂点の挿入、ラベル付き頂点の脱落の重みを、それぞれ $p, q, r$ とする。また、 $p, q, r$ は非負で、 $p < q + r$ とする。 $T_a(u)$ から $T_b(v)$ へ変換するのに必要な変換操作の重みの和の最小値を $T_a(u)$ から $T_b(v)$ への距離と定義し、 $D(T_a(u), T_b(v))$ と書く。距離を与える写像を重み最小の写像という。

$Tai$ 距離は根があり、順序がある木の距離であるが、 $D(T_a(u), T_b(v))$ は根はあるが順序のない木の距離である。根も順序もない木 $T_a, T_b$ の間の距離 $D(T_a, T_b)$ を次のように定義する。

$$D(T_a, T_b) = \min_{u \in L_a, v \in L_b} \{D(T_a(u), T_b(v))\}. \quad (7)$$

### 3. 木の距離の計算法

$D(T_a(u), T_b(v))$ は、 $(b_1), (b_2)$ の条件を満たす写像のうち、最小重みを与える写像の重みである。 $D(T_a(u), T_b(v))$ の計算法を明らかにするために、まず $T_a(u)$ から $T_b(v)$ への写像について調べておく。 $T_a(u)$ から $T_b(v)$ への1つの写像を $M_t$ とする。 $p < q + r$ であるから、 $V_a$ が全て脱落、 $V_b$ が全て挿入する場合は最小重みの写像ではない。したがって最小重みの写像 $M_{min}$ は $M_{min} \neq \emptyset$ ( $\emptyset$ は空集合)である。そこで、一般性を失うことなく $M_t \neq \emptyset$ と仮定す

る。

### [補題 1]

- (1)  $T_a(u)$ に  $I_2$ の頂点を少なくとも 1つ持つ道  $P_a(u, x)$  (図 2) を選ぶ。  $P_a(u, x)$  上の頂点で  $I_2$ に属するものを  $u_1, u_2, \dots, u_k$  とし、  $u_{i-1} \in A_n(u_i)$  ( $1 < i \leq k$ ) とする。  $u_i$  の像を  $v_i$  とすると、  $v_1, v_2, \dots, v_k$  は  $T_b(v)$  上の 1つの道  $P_b(v, y)$  上にあり、  $v_{i-1} \in A_n(v_i)$  ( $1 < i \leq k$ ) である。
  - (2)  $P_a(u, x)$  上の頂点  $c$  に関する分岐部分木  $\text{subta}(c, c')$  の 1つの道  $P_a(c, x_s)$  上の頂点で  $I_2$ に属するものを  $c_1, c_2, \dots, c_h$  とし、  $c_{j-1} \in A_n(c_j)$  ( $1 < j \leq h$ ) とする。  $c_j$  の像を  $d_j$  とすると、  $d_1, d_2, \dots, d_h$  は  $T_b(v)$  の 1つの分岐部分木  $\text{subtb}(d, d')$  の 1つの道  $P_b(d', y_t)$  上にあり、  $d_{j-1} \in d_j$  ( $1 < j \leq h$ ) である。
  - (3) (2) で  $c$  を任意の  $T_a(u)$  の頂点  $x$  としても、 同様に成立する。
- 証明： (1)  $T_a(u)$  の道  $P_a(u, x)$  の頂点のうち  $u_1, u_2, \dots, u_k$  が  $u_{i-1} \in A_n(u_i)$  ( $1 < i \leq k$ ) の関係にあるならば、 写像の条件 (b<sub>2</sub>) (i) から、  $u_1, u_2, \dots, u_k$  の像  $v_1, v_2, \dots, v_k$  の間に  $v_{i-1} \in A_n(v_i)$  ( $1 < i \leq k$ ) の関係があるのは明らかである。
- (2) (1) の  $T_a(u)$  を  $\text{subta}(c, c')$  に、  $T_b(v)$  を  $\text{subtb}(d, d')$  に置き換えれば (2) が成立するのは明らかである。
- (3) についても同様である。

補題 1 は「 $T_a(u)$  の根  $u$  から葉に至る道上、 すなわち幹上の頂点の像是  $T_b(v)$  の幹上にあり、  $T_a(u)$  の分岐部分木  $\text{subta}(c, c')$  の幹上の頂点の像是  $T_b(v)$  の分岐部分木  $\text{subtb}(d, d')$  の幹上にある」(\*1) ことを示している。しかし、 最小重みの写像のとき、  $T_a(u)$  のどの幹が  $T_b(v)$  の

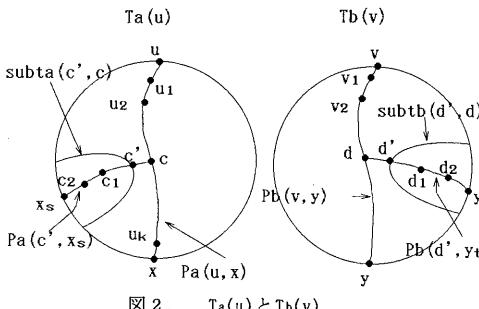


図 2.  $T_a(u)$  と  $T_b(v)$

どの幹に対応し、  $T_a(u)$  の部分木のどの幹が  $T_b(v)$  の部分木のどの幹に対応しているかは明らかではない。それゆえ、  $D(T_a(u), T_b(v))$  の計算をするとき、「全ての可能

な幹の対応を考えねばならない。 $T_a(u)$  の幹が  $m$  本 ( $m = |La(u)|$ )、  $T_b(v)$  の幹が  $n$  本 ( $n = |La(v)|$ ) あるときは、  $m \cdot n$  個の幹の対応を考える」(\*2) ことになる。部分木の幹の間の対応についても同様である。

### [補題 2]

- (1)  $T_a(u)$  の頂点  $c$  が  $\deg(c) > 1$  であるとする。  $c$  の子を  $c_1, c_2, \dots, c_m$  とする。部分木  $T_a(c, c_1), T_a(c, c_2), \dots, T_a(c, c_m)$  のうち少なくとも 2つの部分木の頂点で  $I_2$ に属するものがあるとき、  $c$  は  $I_2$ に属する。
- (2)  $T_b(v)$  の頂点  $d$  が  $\deg(d) > 1$  であるとする。  $d$  の子を  $d_1, d_2, \dots, d_n$  とする。部分木  $T_b(d, d_1), T_b(d, d_2), \dots, T_b(d, d_n)$  のうち少なくとも 2つの部分木の頂点で  $J_2$ に属するものがあるとき、  $d$  は  $J_2$ に属する。

証明： (1)  $I_2$ に属する頂点を持つ部分木を、 一般性を失うことなく、  $T_a(c, c_1), T_a(c, c_2)$  とする。 $I_2$ に属する  $T_a(c, c_1)$  の頂点の 1つを  $x_1$ 、  $I_2$ に属する  $T_a(c, c_2)$  の頂点の 1つを  $x_2$  とする。このとき、  $c = \text{Ccan}(x_1, x_2)$ 。 (b<sub>2</sub>) (iii) から、  $c$  は  $I_2$ に属する。

(2) (1) と同様に成立する。

### [補題 3]

$T_a(u)$  から  $T_b(v)$  への 1つの写像が与えられたとき、 次の条件を満たす幹が存在する。

(1)  $T_a(u)$  あるいは  $T_a(u)$  の部分木の幹上の頂点  $c$  が脱落したときは、 その分岐部分木は全て脱落する。

(2)  $T_b(v)$  あるいは  $T_b(v)$  の部分木の幹上の頂点  $d$  が挿入したときは、 その分岐部分木は全て挿入した。

証明： (1) 補題 2 から、  $c$  の 2つの部分木に  $I_2$ に属する頂点があるとき、  $c$  は  $I_2$ に属するが、 (a) 1つの部分木 (それを  $T_a(c, c_i)$  とする) に  $I_2$ に属する頂点があるとき、 および (b) 全ての部分木が脱落するときは、  $c$  は  $I_2$ に属してもよいし、 脱落してもよい。(a)の場合、  $c_i$  を通る幹を選ぶことになると、  $c$  が脱落するときは、  $c$  に関する全ての分岐部分木が脱落したことになる。

(2) (1) と同様に証明できる。

補題 1～3 から次の計算法が考えられる。

- (1)  $T_a(u)$  から 1つの幹  $tra(u, x)$  を選び、  $T_b(v)$  から 1つの幹  $trb(v, y)$  を選ぶ。
  - (2)  $tra(u, x)$  から  $trb(v, y)$  の 1次元距離 (WLD) を計算する。このとき、 分岐部分木に関して次の操作をし、 変換の重みを計算する。
- ( $c_1$ )  $tra(u, x)$  の頂点  $c$  が  $trb(v, y)$  の頂点  $d$  に変換される

とき、 $c$ に関する分岐部分木は $d$ に関する分岐部分木に変換される。

(c<sub>2</sub>) $c$ が脱落するとき、 $c$ に関する分岐部分木は全て脱落する。

(c<sub>3</sub>) $d$ が挿入されたとき、 $d$ に関する分岐部分木は全て挿入されたとする。

(3)  $T_a(u)$ および $T_b(v)$ の全ての幹について上記の操作をして、最小の変換重みを求める。

図3の場合を考えよう。 $subta(c, c_2)$ が $subtb(d, d_1)$ に変換する事を $[2, 1]$ と書く。 $p < q + r$ であるから、一般に $D(Ta(x), Tb(y)) < |Ta(x)| \cdot r + |Tb(y)| \cdot q$ である。ここで $|Ta(x)|$ は $Ta(x)$ の頂点数を表す。したがって $c$ が $d$ に写像しているとき、 $c$ の2つの分岐部分木は $d$ の分岐部分木に写像している。 $c$ と $d$ の分岐部分木の間に次の対応を考えられる。

$\{[1, 1], [2, 2]\}, \{[1, 2], [2, 1]\}, \{[1, 1], [3, 2]\}, \{[1, 2], [3, 1]\}, \{[2, 1], [3, 2]\}, \{[2, 2], [3, 1]\}$ 、ここで、 $\{[1, 2], [2, 1]\}$ は $subta(c, c_1)$ が $subtb(d, d_1)$ に、 $subta(c, c_2)$ が $subtb(d, d_1)$ に変換し、 $subta(c, c_3)$ が脱落したことを示す。変換 $\{[1, 2], [2, 1]\}$ の重みは

$$D(subta(c, c_1), subtb(d, d_2)) + D(subta(c, c_2), subtb(d, d_1)) + |subta(c, c_3)| \cdot r$$

である。上記6個の変換の重みの最小値を $d(\{c_1, c_2, c_3\}, \{d_1, d_2\})$ と書く。一般に、 $c$ の分岐部分木の根とな

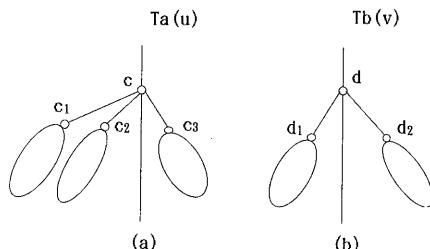


図3. (a) $c$ の分岐部分木、(b) $d$ の分岐部分木

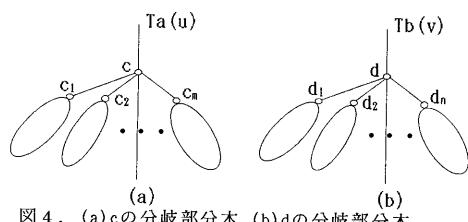


図4. (a) $c$ の分岐部分木、(b) $d$ の分岐部分木

る $c$ の子の数は $m$ 、 $d$ の分岐部分木の根となる $c$ の子の数は $n$ であるとする(図4)。

$(c_1) \sim (c_m)$ に対応して、次の場合がある。  
 $(d_1)c$ が $d$ に変換されるとき、このときの変換の重みを $\triangle 1(c, d)$ とする。

$$\triangle 1(c, d) = p(c, d) + d(\{c_1, \dots, c_m\}, \{d_1, \dots, d_n\}). \quad (8)$$

$(d_2)c$ が脱落するとき、このときの変換の重みを $\triangle 2(c)$ とする。

$$\triangle 2(c) = r + \left( \sum_{i=1}^m |subta(c, c_i)| \right) \cdot r. \quad (9)$$

$(d_3)d$ が挿入するとき、このときの変換の重みを $\triangle 3(d)$ とする。

$$\triangle 3(d) = q + \left( \sum_{i=1}^n |subtb(d, d_i)| \right) \cdot q. \quad (10)$$

一組の幹 $tra(u, x)$ 上の頂点を $u_1, u_2, \dots, u_h$ とし、 $trb(v, y)$ 上を頂点を $v_1, v_2, \dots, v_k$ とする。ここで $u=u_1, x=u_h, v=v_1, y=v_k$ である。これらの幹を指定したときの $Ta(u)$ から $Tb(v)$ への距離を $d((u, x), (v, y))$ と書く。

[ $d((u, x), (v, y))$ を計算するアルゴリズム]

```
procedure D(u, x, v, y)
begin
d(0, 0) := 0;
for i=1 to h do d(i, 0) := d(i-1, 0) + △2(ui);
for j=1 to k do d(0, j) := d(0, j-1) + △3(vj);
for i=1 to h do
for j=1 to k do
begin
d1 := d(i-1, j) + △2(ui);
d2 := d(i, j-1) + △3(vj);
d3 := d(i-1, j-1) + △1(ui, vj);
d(i, j) := min {d1, d2, d3};
end;
d((u, x), (v, y)) := d(h, k);
end.
```

このアルゴリズムは $\triangle 1(ui, vj)$ で手続き $D(u, x, v, y)$ を呼び出すので再帰的アルゴリズムである。

$D(Ta(u), Tb(v))$ は次のように書ける。

$$D(Ta(u), Tb(v)) = \min_{x \in La(u), y \in Lb(v)} \{d((u, x), (v, y))\}. \quad (11)$$

(7)式を次の(12)式のように変形できないだろうか。ここで、 $u$ は $Ta$ の任意の頂点である。

$$D(Ta, Tb) = \min_{x \in Lb} D(Ta(u), Tb(v)) \quad (12)$$

図5を考えてみよう。いま

$$D(Ta, Tb) = D(Ta(1), Tb(v)) \quad (13)$$

となつたとする。このとき、 $Ta(1)$ の幹として  $tra(1, 9)$  を選んだとき  $D(Ta(1), Tb(v))$  が決定されたとする。即ち、

$$D(Ta(1), Tb(v)) = d((1, 9), (v, y)) \quad (14)$$

ここで、 $(v, y)$  は  $tra(1, 9)$  の頂点の像によって定まる幹  $trb(x, y)$  を表している。いま、 $u=4$  と選んだときは、幹に  $tra(4, 1)$ ,  $tra(4, 7)$ ,  $tra(4, 9)$ ,  $tra(4, 10)$  のいずれかが選ばれる。いずれの場合も頂点 3 は脱落する。したがって、それらの分岐部分木は脱落しなければならない。しかし、 $tra(4, 1)$  の場合は  $subta(3, 5)$  が脱落せず、その他の場合は  $subta(3, 2)$  が脱落しない。ゆえに、任意の葉を根を選んで  $Ta$  から  $Tb$  への最小重みの写像を求ることはできない。すなわち (12) は成立しない。

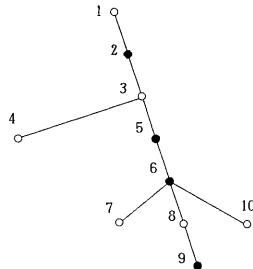


図5. 木  $Ta$ . ●は  $I_2$  に属する頂点. ○は脱落した頂点.

#### 4. 計算量

##### 4.1 計算結果の再利用

手続き  $D(u_1, u_h, v_1, v_k)$  で計算が  $u_1, v_j$  まで進んだときの途中結果を  $d((u_1, u_i), (v_1, v_j))$  と書く。 $Ta(u_1)$  の幹を固定した場合を考える。本アルゴリズムに従い、図6において、まず、 $d((u_1, u_h), (v_1, v_k))$  を計算する。本文の計算法は WLD の計算図式に似て、図7 のような計算図式で表せる。 $tra(u_1, u_h)$  上の頂点  $u_i$ ,  $trb(v_1, v_k)$  上の頂点  $v_j$  に関する図7(a) の格子点  $(u_i, v_j)$  に  $d((u_1, u_i), (v_1, v_j))$  を記憶している。図8 の例で考える。いま、 $Ta(u_1)$  の幹

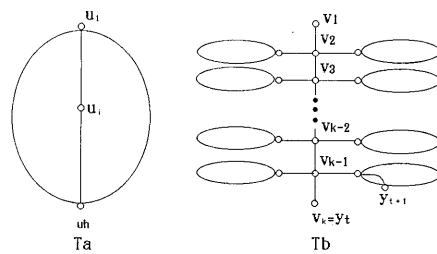
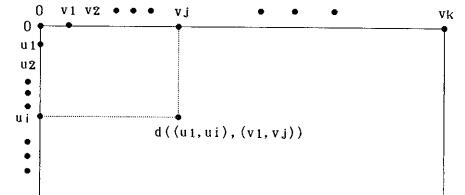


図6. 計算結果の再利用



(a)

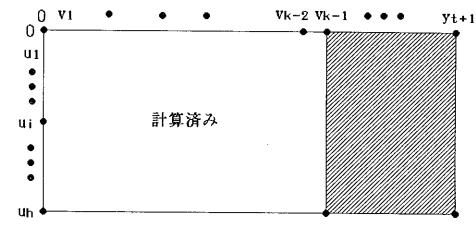


図7.  $d((u_1, u_i), (v_1, v_j))$

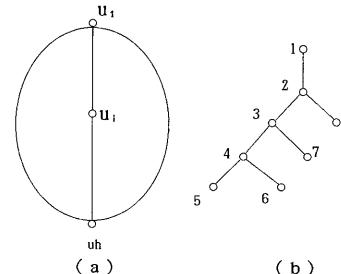


図8. (a)  $Ta(u_1)$ , (b)  $Tb(1)$

$tra(u_1, u_h)$  を固定し、 $Tb(1)$  の幹を順次変化させて  $D(Ta(u_1), Tb(1))$  を計算する場合を考える。まず、 $trb(1, 5)$  を選び、 $d((u_1, u_h), (1, 5))$  を計算する。続いて、 $trb(1, 6)$  を選ぶ。このとき、 $u_1$  を  $tra(u_1, u_h)$  上の任意の頂点とするとき、 $d((u_1, u_i), (1, 3))$  は既に計算されている。したがって新たに計算するのは、 $d((u_1, u_i), (1, 4))$  や  $d((u_1, u_i), (1, 6))$  である。次に  $trb(1, 7)$  を選ぶ。このとき、新たに計算するのは  $d((u_1, u_i), (1, 3))$  や  $d((u_1, u_i), (1, 7))$  である。

一般的に考えよう。 $Tb(v_1)$  の頂点には根を 1 とし、プリオリーダ順に番号がつけられているとする。 $Tb(v_1)$  の葉を  $y_1, y_2, \dots, y_s$  とし、 $y_i < y_{i+1}$  ( $1 \leq i \leq s-1$ ) であるとする。いま  $v_k = y_t$  であるとする。 $d((u_1, u_h), (v_1, v_k))$  の計算後、 $d((u_1, u_h), (v_1, y_{t+1}))$  を計算するとする。このとき、 $d((u_1, u_i), (v_1, v_{k-2}))$  ( $1 \leq i \leq h$ ) は既に計算されている。計算すべき部分は、図7(b) の  $v_{k-1}$  列～ $y_{t+1}$  列（斜線部分）

である。この列の数  $N(y_{t+1})$  は

$$\begin{aligned} N(y_{t+1}) &= |\text{trb}(v_1, y_{t+1})| - |\text{trb}(v_1, v_{k-2})| \\ &= |\text{trb}(v_1, y_{t+1})| - (\text{trb}(v_1, y_t) \text{ と } \text{trb}(v_1, y_{t+1})) \text{ の} \\ &\quad \text{共通部分の頂点数}-1) \end{aligned} \quad (15)$$

である。  $N(y_{t+1})$  は  $\text{trb}(v_1, y_{t+1})$  の頂点の中で「新しく計算に加わった頂点数+1」である。 $L_b(v_1)$  のすべての頂点について調べると、

$$Nb(v_1) = |\text{trb}(v_1, y_1)| + \sum_{t=2}^s N(y_t) \quad (16)$$

が計算に参加する頂点数である。これは、「 $|\text{trb}(v_1, y_1)| + \sum$  (新しく計算に参加した頂点数)」は  $|V_b(v_1)|$  に等しく、

$$\sum_{t=2}^s 1 = |L_b(v_1)| - 1 \quad (17)$$

である。ゆえに、

$$Nb(v_1) = |V_b(v_1)| + |L_b(v_1)| - 1. \quad (18)$$

となる。

#### 4.2 計算量

$T_a$  木  $T_a(u) = (V_a(u), E_a(u))$  と  $T_b(v) = (V_b(v), E_b(v))$  の距離を求める手数を考える。 $|V_a(u)| = m$ ,  $|V_b(v)| = n$  とし、 $T_a, T_b$  の頂点の次数が高々 2 であるとき  $O_T(mn)$  となる。また、図 9 のような  $T_a(u), T_b(v)$  の場合、(8)～(10) 式で距離を計算できる。(9) 式の計算は加算  $m$  回、乗算 1 回、(10) 式では加算  $n$  回、乗算 1 回であり、一般に(8)式の計算に比べて計算量が少ないので無視できる。いま、一般性を失うことなく  $m \leq n$  と仮定する。 $T_a(u)$  の幹の選び方は  $(m-2)$  通り、 $T_b(v)$  では  $(n-2)$  通りあり、 $(m-2)(n-2)$  通りの幹の組み合わせで計算することになる。 $T_a(u)$  の幹  $\text{tra}(u, x)$ ,  $T_b(v)$  の幹  $\text{trb}(v, y)$  が選ばれたとする。このとき  $d((u, x), (v, y))$  は(8)式で計算できる。 $T_a(u)$  の  $x$  以外の葉は  $(m-3)$ ,  $T_b(v)$  の  $y$  以外の葉は  $(n-3)$  あるから、 $(n-3)$  から  $(m-3)$  を選ぶ組み合わせの数は  $n-s C_{m-s}$  である。 $T_a(u)$  の葉が  $T_b(v)$  の選ばれた  $(m-3)$  の葉のどれに対応するかによって、1 つの組み合わせについて  $(m-3)!$  通りの場合がある。ゆえに、(8)式の  $d(\{c_1, \dots, c_m\}, \{d_1, \dots, d_n\})$  を計算するときの分岐部分木の対応の総数  $N$  は

$$N = n-s C_{m-s} \cdot (m-3)! = (n-3)(n-4) \cdots (n-m+1). \quad (19)$$

それぞれの場合について、加減算  $(m-1)$  回、乗算 1 回必要である。また、どの場合が最小重みになるかを知るために  $(N-1)$  回の比較が必要である。ゆえに、1 組の幹についての演算の総数  $\text{Total}$  は

$$\text{Total} = (m+2)N + N-1 = (m+3)N-1. \quad (20)$$

幹の対は、 $(m-2)(n-2)$  通りあるから、時間複雑さは  $O_T(m(m-2)(n-2)(n-3)(n-4) \cdots (n-m+1))$  である。 $D(T_a, T_b)$  の計算では、高々  $(m-1)(n-1)$  通りの根の対について

$D(T_a(u), T_b(v))$  を計算するから、計算量は  $O_T(m^3(n-1)(n-2)(n-3) \cdots (n-m+1))$  である。図 9 の例が最も計算量の多くなる例であるから、計算量は  $O_T(mn) \sim O_T(m^3(n-1)(n-2)(n-3)(n-4) \cdots (n-m+1))$  の間にある。

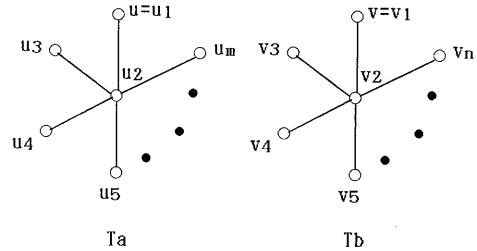


図 9. 頂点の最高次数がそれぞれ  $m-1, n-1$  の木

図 9 の例は異なった方法でより簡単に計算できる。

$T_a, T_b$  の頂点のラベルの集合を  $\Sigma = \{a_1, a_2, \dots, a_s\}$  とする。 $V_a$ においてラベルが  $a_1, \dots, a_s$  である根  $u$  以外の頂点の個数をそれぞれ  $m_1, \dots, m_s$ ,  $V_b$ においてラベルが  $a_1, \dots, a_s$  である根  $v$  以外の頂点の個数をそれぞれ  $n_1, \dots, n_s$  とする。  
 $hs = \sum_{i=1}^s \min\{m_i, n_i\}$  とし、 $m \geq n$  とするととき、頂点の集合  $V_a, V_b$  において、 $hs$  個は同じラベルを持つので、 $T_a(u)$  から  $T_b(v)$  への距離は

$$\begin{aligned} D(T_a(u), T_b(v)) &= (n-2-hs)p + (m-n)r + p(u_1, v_1) \\ &\quad + p(u_2, v_2). \end{aligned} \quad (21)$$

となる。ゆえに、 $O_T(m+n)$  となる。 $m < n$  の場合も同様である。

次に空間複雑さについて考える。(18)式より

$$Nb(v) = |V_b(v)| + |L_b(v)| - 1 \leq 2 \cdot n \quad (22)$$

であり、 $T_a(u)$  についても

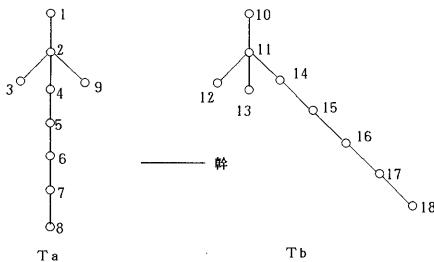
$$Nb(u) = |V_a(u)| + |L_a(u)| - 1 \leq 2 \cdot m \quad (23)$$

である。また、 $T_a(u)$  の幹上の一頂点が  $T_b(v)$  の幹上の 1 つの頂点に写像しているとき、分岐部分木の対応は 4 通りある。ゆえに、空間複雑さは  $O_s(m \cdot n)$  になる。

#### 4.3 枝刈り

本アルゴリズムの枝刈りについて述べる。まず、(i) 木  $T_a, T_b$  の指定した幹  $\text{tra}, \text{trb}$  の頂点数の差  $|\text{tra}| - |\text{trb}|$  を求める。 $|\text{tra}| > |\text{trb}|$  のときは  $\text{tra}$  の頂点の脱落重み  $(|\text{tra}| - |\text{trb}|) \cdot r$ ,  $|\text{tra}| < |\text{trb}|$  のときは  $\text{trb}$  の頂点の挿入重み  $(|\text{trb}| - |\text{tra}|) \cdot q$  を求め、その値を  $\alpha$  とする。実際の幹上の頂点の変換重みは必ず、 $\alpha$  以上になる。次に、(ii) 木  $T_a, T_b$  のそれぞれの分岐部分木の総頂点数の差から、分岐部分木の挿入、脱落重みの  $\beta$  を計算す

る。すなわち、 $K = (|V_a| - |tr_a|) - (|V_b| - |tr_b|)$  とし、 $K > 0$  なら  $\beta = K \cdot r$ 、 $K < 0$  なら  $\beta = -K \cdot q$  とする。もし (i), (ii) の



重みの和  $\gamma = \alpha + \beta$  があらかじめ定めたしきい値  $\eta$  を越えるならば、その指定した幹に対する木の距離の計算

図 10. 枝刈り

を行わない。図 10においてしきい値を  $\eta$  とするとき  $\alpha = r \cdot (|tr(1, 8)| - |tr(10, 13)|) = (7-3) \cdot r = 4 \cdot r$ ,  $\beta = ((|V_b| - |tr(10, 13)|) - (|V_a| - |tr(1, 8)|)) \cdot q = ((9-3) - (9-7)) \cdot q = (6-2) \cdot q = 4 \cdot q$  であるから、 $\gamma = \alpha + \beta = 4 \cdot r + 4 \cdot q$  となり、もし  $\gamma > \eta$  ならば、この幹の組み合わせにおける計算を行わない。

## 5. 計算例及び化合物の分類

### 5.1 計算例

図 11 の 2 つの化合物間の距離の計算過程の一部を示す。枝に関する操作は行っていない。ここでは、 $p = q = r = 1$ とした。

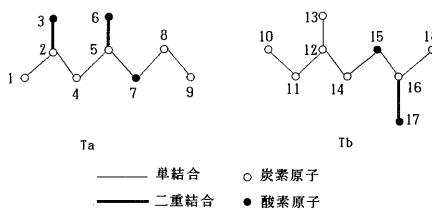


図 11. 化合物 Ta, Tb

$$Ta = (V_a, E_a), Tb = (V_b, E_b)$$

$$V_a = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

$$V_b = \{10, 11, 12, 13, 14, 15, 16, 17, 18\}$$

$$L_a = \{1, 3, 6, 9\} \quad L_b = \{10, 13, 17, 18\}$$

根に選べるのは次数 1 の頂点であるから、根の対は次の 16 通り存在する。

$$(1, 10), (1, 13), (1, 17), (1, 18), (3, 10), (3, 13), (3, 17), (3, 18), (6, 10), (6, 13), (6, 17), (6, 18),$$

$$(9, 10), (9, 13), (9, 17), (9, 18)$$

### (1) 根の対 (1, 10)

根の対 (1, 10) を選ぶ。すなわち、 $T_a(1)$  と  $T_b(10)$  の距離の計算をする。それぞれの幹には次のものがある。

$$T_a(1) \text{ の幹: } tr(1, 3), tr(1, 6), tr(1, 9)$$

$$T_b(10) \text{ の幹: } tr(10, 13), tr(10, 17), tr(10, 18)$$

### (i) 幹の対 ( $tr(1, 3), tr(10, 13)$ )

幹の対  $tr(1, 3), tr(10, 13)$  を選び、 $d((1, 3), (10, 13))$  を計算する。手続き  $D(u, x, v, y)$  より、

各  $d(i, j)$  ( $0 \leq i \leq 3, 0 \leq j \leq 4$ ) は以下のようになる。

$$d(0, 0) = 0, d(0, 1) = 1, d(0, 2) = 2, d(0, 3) = 8,$$

$$d(0, 4) = 9, d(1, 0) = 1, d(1, 1) = 0, d(1, 2) = 1,$$

$$d(1, 3) = 7, d(1, 4) = 8, d(2, 0) = 8, d(2, 1) = 7,$$

$$d(2, 2) = 6, d(2, 3) = 4, d(2, 4) = 5, d(3, 0) = 9,$$

$$d(3, 1) = 8, d(3, 2) = 7, d(3, 3) = 5, d(3, 4) = 5.$$

ゆえに、

$$d((1, 3), (10, 13)) = d(3, 4) = 5.$$

### (ii) 幹の対 ( $tr(1, 3), tr(10, 17)$ )

$$d((1, 3), (10, 17)) = 10.$$

⋮

### (ix) 幹の対 ( $tr(1, 9), tr(10, 18)$ )

$$d((1, 9), (10, 18)) = 4.$$

### (i)～(ix) より、

$$D(T_a(1), T_b(10))$$

$$= \min \{d((1, 3), (10, 13)), \dots, d((1, 9), (10, 18))\} = 4.$$

### (2) 根の対 (1, 13)

$$D(T_a(1), T_b(13))$$

$$= \min \{d((1, 3), (13, 10)), \dots, d((1, 9), (13, 18))\} = 5.$$

⋮

### (4) 根の対 (1, 18)

$$D(T_a(1), T_b(18))$$

$$= \min \{d((1, 3), (18, 10)), \dots, d((1, 9), (18, 17))\}$$

$$= d((1, 3), (18, 17))$$

$$= 3$$

⋮

### (16) 根の対 (9, 18)

$$D(T_a(9), T_b(18))$$

$$= \min \{d((9, 1), (18, 10)), \dots, d((9, 6), (18, 17))\} = 4.$$

(1)～(16) より、 $T_a, T_b$  の距離は、

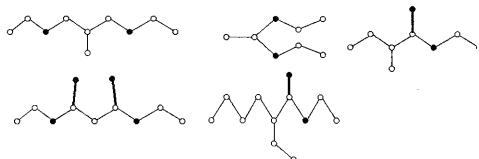
$D(T_a, T_b)$

$$= \min\{D(T_a(1), T_b(10)), \dots, D(T_a(9), T_b(18))\} = 3.$$

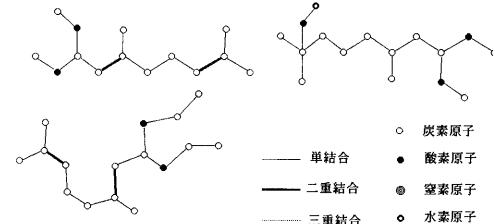
## 5.2 化合物の分類

鎖構造の化合物 111 個について分類を試みた。化合物を表す木の頂点数は 3 ~ 19 個で平均 12.0 個であった。炭素原子に結合している水素原子は原則として省略している。類内の化合物間の距離は 3 以内、異類の化合物間距離は 4 以上としたとき、17 類に分類

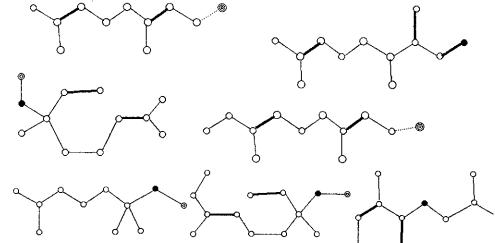
類 1



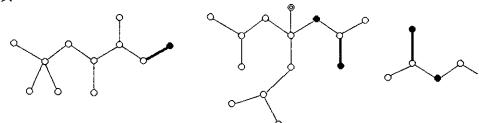
類 2



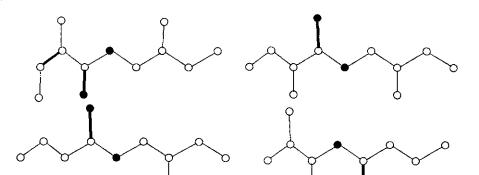
類 3



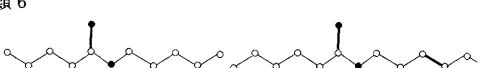
類 4



類 5



類 6



類 7

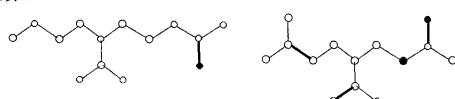


図 12. 分類例（一部）

できた。その内的一部を図 12 に示す。計算機はワークステーション NWS-1860 (SONY)、言語は C を使用した。距離の計算時間は平均 4.3 秒であった。

## 6. あとがき

化合物の構造式の類似度を表す尺度はいろいろ提案されている<sup>23)</sup>が、構造式を表すグラフをありのまま取り扱う方法はなかった。すなわち、一般には構造式をグラフと見て、グラフの距離を定義しなければならないが、この問題に応用できる本格的なグラフの距離の研究はまだない。本文では、この問題の対する第一歩として、根がなく、順序がない木の距離を定義し、その計算法を提案した。計算複雑さは、 $O_T(mn) \sim O_T(m^3(n-1)(n-2)(n-3)\dots(n-m+1))$  の間にある。また、このとき、空間複雑さは  $O_S(m \cdot n)$  となる。平均原子数 12.0 個の化合物を計算機（ワークステーション SONY NWS-1860 (5 MIPS)）で計算したところ、平均計算時間は 4.3 秒であった。さらに、この尺度が、化合物の分類に応用できることを示した。これは、類似化合物を検索するためのファイルの構成等に有用である。辺の変換を考慮した木の距離を考えること、また、頂点数が多くなると計算時間が急速に増大するので、それを改善することが今後の問題である。

## 7. 参考文献

- 1) Sankoff, D. and Kruskal, J. B. : Time warps, string edit, and macromolecules : The theory and practice of sequence comparison, Addison-Wesley Publishing Company (1983).
- 2) Levenshtein, V. I. : Binary codes with correction of deletions, insertions and substitutions of symbols, Dokl. Akad. Nauk SSSR, Vol. 163, pp. 845-848 (1965).
- 3) 奥田, 田中, 笠井 : 距離概念の拡張による語の誤り訂正法, 昭47 電気関係学会東海支部連合大会, 18A-B-6 (1972).  
Okuda, T., Tanaka, E. and Kasai, T. : A method

- for the correction of garbled words based on the Levenshtein metric , IEEE Trans. Comput. , Vol.C-25 , pp.172-178 (1976).
- 4) Wagner, R.A. and Fischer, M.J. : The string-to-string correction problem , JACM , Vol.21 , pp.168-173 (1974).
  - 5) Sellers, P. : An algorithm for the distance between two finite sequences , Journal of Combinatorial Theory , Vol.16 , pp.253-258 (1974).
  - 6) Sankoff, D. : matching sequences under deletion-insertion constraints , Proceedings of the National Academy of Sciences of the U.S.A. , Vol.69 , pp.4-6 (1972).
  - 7) Selkow, S.M. : Tree-to-tree editing problem , Information Processing Letters , Vol.6 , pp.184-186 (1977).
  - 8) Tai, K.C. : The tree-to-tree correcting problem , JACM , Vol.26 , pp.422-433 (1979).
  - 9) Lu, S.Y. and Fu, K.S. : A tree-to-tree distance and its application to cluster analysis , IEEE Trans. PAMI , Vol.PAMI-1 , pp.219-224 (1979)
  - 10) 青木 : 木と木の距離を求める下降型アルゴリズム , 電子通信学会論文誌 , Vol.J66-D , pp.49-56 (1983).
  - 11) 田中 : Taiの定義した木の間の距離の上昇型計算法 , 電子通信学会論文誌 Vol.J66-D , pp.660-667 (1983).
  - 12) 田中, 田中 : 木の間の距離とその計算法 , 電子通信学会論文誌 Vol.J65-D , pp.511-518 (1982)
  - 13) 大森 : A Unified View on Tree Metrics , 宇都宮大学大学院工学研究科情報工学専攻 , 修士論文 (1987)
  - 14) 中林, 鎌田 : 2分木間の距離とその計算アルゴリズム , 電子通信学会論文誌 , Vol.J66-D , pp.455-462 (1983).
  - 15) Lu, S.Y. : A tree-matching algorithm based on node splitting and merging , IEEE Trans. PAMI , Vol.PAMI-6 , pp.249-256 (1984)
  - 16) Cheng, Y.C. and Lu, S.Y. : Waveform correlation by tree matching , IEEE Trans. PAMI , Vol. PAMI-7 , pp.299-305 (1985).
  - 17) Wilhelm, R. : A modified tree-to-tree correction problem , Information Processing Letters , Vol.12 , pp.127-132 (1981).
  - 18) Culik II, K. and Wood, D. : A note on some tree similarity measures , Information Processing Letters , Vol.15 , pp.39-42 (1982).
  - 19) Tsai, W.H. and Fu, K.S. : Error-correcting isomorphisms of attributed relational graphs for pattern analysis , IEEE Trans. SMC , Vol. SMC-9 , pp.747-768 (1979)
  - 20) Sanfeliu, A. and Fu, K.S. : A distance measure for image analysis , IEEE Trans. SMC , Vol. SMC-14 , pp.398-408 (1984).
  - 21) Bunke, H. and Allermann, G. : Inexact graph matching for structural pattern recognition , Pattern Recognition Letters , Vol.1 , pp.245-253 (1983).
  - 22) 田口, 池田, 田中 : 節点の分離・融合操作に基づく木の距離について , 電子情報通信学会論文誌 D-I , Vol.J75-D-I , pp.191-195 (1992).
  - 23) Johnson, M.A. : A review and examination of the mathematical spaces underlying molecular similarity analysis , J. Math. Chem. , Vol.3 , pp.117-145 (1989).