# パスウェイ解析のための情報量最大化アライメントアルゴリズム

遠里 由佳子† 松田 秀雄† 橋本 昭洋†

生体内の化学反応の多くは，酵素を触媒として，ある化合物（基質）を，別の化合物（生成物）に変換することにより構成される．こういう一連の反応経路を一般にパスウェイと呼ぶ．代謝反応のパスウェイに対する比較分析は，進化の過程で生物がどのようにそのパスウェイを獲得したか，さらには，ある化合物を合成する方法についての知見を得るうえで重要な情報となる．パスウェイを構成する各酵素は，触媒する化学反応のタイプにより4組の数字からなるEC（Enzyme Commission）番号が付けられ，系統的な分類がされている．したがって，共通のパターンをみつける際には，機能階層により条件緩和できることが望ましい．そこで，本研究では，情報量という評価基準を用い，木構造の概念階層を持つ記号に対して拡張されたマルチプルアライメントアルゴリズムを提案する．実際に，解糖やアミノ酸分解，DNAの複製にかかわるパスウェイに対し本アルゴリズムを適用し，その有効性を確かめた．

# An Information Content Maximization Alignment Algorithm for Metabolic Pathway Analysis

Yukako Tohsato,† Hideo Matsuda† and Akihiro Hashimoto†

In many of the chemical reactions in living cells, enzymes act as catalysts in the conversion of certain compounds (substrates) into other compounds (products). Comparative analyses of the metabolic pathways formed by such reactions give important information on their evolution and on pharmacological targets. Each of the enzymes that constitute a pathway is classified according to the EC (Enzyme Commission) numbering system, which consists of 4 sets of numbers that categorize the type of the chemical reaction catalyzed. Therefore, in order to find a common pattern among pathways, it is desirable to be able to use this functional hierarchy to relax the match conditions. In this paper, we propose a multiple alignment algorithm utilizing information content, that is extended to symbols having a hierarchical structure. The effectiveness of our method is demonstrated by applying the method to pathway analyses of sugar, DNA and amino acid metabolisms.

## 1. はじめに

生体内の化学反応の多くは，酵素を触媒として，ある化合物（基質）を，別の化合物（生成物）に変換する化学反応により構成される．その反応経路については近年よく解明されており，WIT☆やKEGG[1]などのデータベースに集められ，WWW上で公開されている．こういう一連の反応経路を一般にパスウェイと呼ぶ．生物の持つパスウェイは糖代謝などの基本的なパスウェイでも多数の分岐をともなう複雑なものとなっている（図1参照）．これらのパスウェイを異なる生物種間や，異なる代謝反応で比較・分析することは，進化の過程で生物がどのようにそのパスウェイを獲得したか，さらには，ある化合物を合成する方法についての知見を得るうえで重要な情報である[2]．

パスウェイを計算機上で解析するこれまでの試みとしては，酵素とそれが触媒する反応での基質と生成物のデータが与えられているときに，パスウェイ再構築[3)~5)]（ある2つの化合物を指定して一方を基質，他方を生成物とするようなパスウェイを求める），パスウェイに基づくゲノム比較[2),6)]（1つのパスウェイを構成する各酵素に種々のゲノム上の遺伝子を割り当てることにより比較），パスウェイのクラスタリング[7)]（パスウェイ中の酵素どうしの配列類似性をもとにパスウェイ間の距離を計算しパスウェイを分類）などがある．

† 大阪大学大学院基礎工学研究科情報数理系専攻
Department of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University

☆ http://wit.mcs.anl.gov/WIT2/

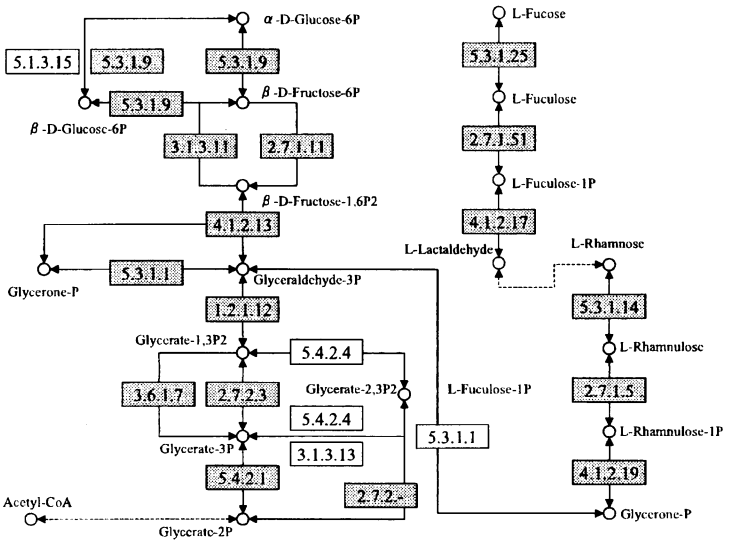


**図 1** 解糖系のパスウェイ．各酵素は EC 番号が書かれた四角で表現されている．色付けされた酵素は，大腸菌ゲノム中の ORF と EC 番号との対応がとられているものを表す．点線で書かれた経路はパスウェイが省略されている．解糖のパスウェイは最終的にアセチル-CoA に至る．

Fig. 1 Sugar degradation pathway. The enzymes are shown in boxes with the EC numbers inside. The shaded boxes represent those enzymes whose genes are identified in *Escherichia coli*. Pathways indicated by dotted lines are not shown. Acetyl-CoA is the final product in the pathway.

本研究では，反応の類似性に基づくパスウェイの比較分析手法の 1 つとして複数のパスウェイのマルチプルアライメントを考える．パスウェイの比較では，上述のパスウェイのクラスタリングのように酵素間の配列類似性を用いることが多いが，酵素の機能すなわち EC 番号が同じでも配列がまったく異なる場合（enzyme recruitment[8]）があることが知られており，配列類似性に基づく比較は必ずしも適切ではないと考えられる．

そこで，本研究では，酵素の持つ機能階層（EC 番号）を利用することにより，階層構造を持つ記号に対して拡張されたマルチプルアライメントアルゴリズムを提案する．以下では，パスウェイのアライメントアルゴリズム，およびこのアルゴリズムを，解糖，DNA 複製やアミノ酸分解にかかわるパスウェイ適用した結果について示す．

## 2. 酵素階層と反応の類似性

酵素は，それが触媒する化学反応のタイプにより EC 番号と呼ばれる 4 組の数字により，階層的な分類（以下，これを酵素階層と呼ぶ）がされている．たとえば，酵素階層の最上位（EC 番号の第 1 番目の数字）で，酵素は大きく次の 6 つに分けられる[9]．

(1) 酸化還元酵素（oxidoreductase）
(2) 転移酵素（transferase）
(3) 加水分解酵素（hydrolase）
(4) 除水付加酵素（lyase）
(5) 異性化酵素（isomerase）
(6) 合成酵素（ligase）

第 1 階層より下の階層は，たとえば (1) の酸化還元酵素では，第 2 階層はドナー，第 3 階層はアクセプタ，第 4 階層は基質の種類により分類される．つまり，[1.1] のグループは CH-OH をドナーとする酵素を表し，[1.1.1] はその中でも NAD+もしくは NADP+をアクセプタとするグループを表す．したがって，[1.1.1.1] と [1.1.1.2] などの酵素は，すべて CH-OH をドナーとし，NAD+もしくは NADP+をアクセプタにする．

本研究では，上記の階層に任意の酵素を表す [*] を加えた 5 つの階層を考える（**図 2** 参照）．

**図 3** に互いに反応が類似したパスウェイの例をあげる．これらのパスウェイは互いに類似した反応を触媒する酵素どうしは，EC 番号の 3 番目までが同じ値になっており酵素階層上で近い位置にあることが分かる．そこで，本研究では，酵素階層上での近さと反応

の類似性との間の関連が大きいと考え，パスウェイの反応の類似性を酵素階層に基づいて表すことを考えることにする．

## 3. 情報量最大化アライメントアルゴリズム

### 3.1 酵素間の類似性スコア

前章で述べたように酵素の触媒する反応の類似性と，その酵素階層上での近さは密接に関連すると考えられるが，酵素間の類似性を単純に酵素階層上での近さ（たとえば，酵素階層上である酵素の位置から別の酵素の位置に行くときにたどる階層の段数）だけで表現するのは問題がある．これは，酵素階層にはその分布に大きな偏りがあるためである．たとえば，酵素階層で [1.1.1] の下に位置する EC 番号は 1.1.1.1 から 1.1.1.254 まであるのに対して，[5.3.4] の下に位置する EC 番号は 5.3.4.1 だけしかない．

また，アライメントすべきパスウェイの集合中に含まれる酵素において，それらの多くが酵素階層において近い位置を占める場合，それらの酵素どうしをランダムに組み合わせても，類似した酵素の組と判定されてしまう．

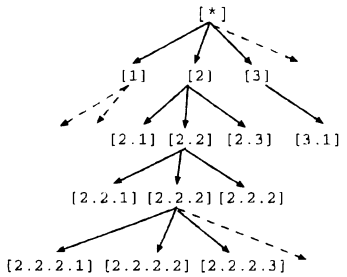


図 2 酵素階層の例
Fig. 2 Example of enzyme hierarchy.

そこで，本研究では上記のような点を考慮して，以下のように 2 つの酵素の間の類似度を決めることにした．この類似度を表すのに必要ないくつかの定義を次に示す．なお，以下では，特に断りのない限り，酵素とその EC 番号とを区別せずに，EC 番号だけで酵素を表すものとする．

**定義 3.1** 2 つ以上の酵素が与えられたとき，酵素階層でそれらの酵素の上位にある階層のうち共通するものの中で最も下位に位置する階層を**共通上位階層**と呼ぶ．同じ酵素の間での共通上位階層は自分自身であるとする．

たとえば，2つの酵素 1.2.3.4 と 1.2.3.5 において，これらの共通上位階層は [1.2.3] であり，1.2.3.4 と 1.2.4.1 では [1.2]，1.2.3.4 と 2.1.1.1 では [*] となる．

**定義 3.2** ある酵素階層 $h$ が与えられたとき，その階層以下の階層に含まれるすべての酵素の集合を $E(h)$ で表す．また，$E(h)$ の要素数を $C(h)$ で表す．

**定義 3.3** パスウェイ集合 $S = \{s_1, \ldots, s_n\}$ が与えられたとき，$S$ の要素であるパスウェイ $s_i$ $(1 \leq i \leq n)$ に現れる酵素の数を $N(s_i)$ で表す．また，ある酵素 $e$ が $s_i$ に現れる数を $o(e, s_i)$ で表す．このとき，

$$\sum_{i=1}^{n} o(e, s_i) / \sum_{i=1}^{n} N(s_i) \qquad (1)$$

を $e$ の $S$ における**出現頻度**と呼び，$p(e)$ で表す．

**定義 3.4** パスウェイ集合 $S$ が与えられたとき，ある

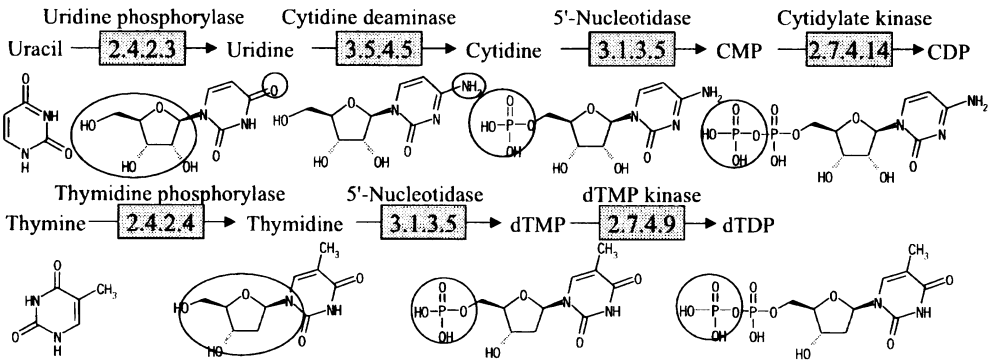


図 3 パスウェイにおける反応の類似部分の構造比較．EC 番号の上に酵素名を，酵素と酵素の間に化合物名を示した．円で囲まれた部分は酵素により変化する点を表している．
Fig. 3 Structural comparison of similar pathways. The enzymes' name are shown above their EC numbers, and the compounds' names are shown between the enzymes. The structural components that are modified by the enzymes are circled.

酵素階層 $h$ において $E(h)$ に属するすべての酵素の出現頻度の足し合わせたものを，$h$ の $S$ における**出現頻度**と呼び，$p(h)$ で表す．

**定義 3.5** 酵素階層 $h$ が与えられたとき，次式で表される $I(h)$ を $h$ の**情報量**と呼ぶ．

$$I(h) = \log_2 \frac{1}{C(h)} - \log_2 p(h) \qquad (2)$$

2 つの酵素 $e_i$，$e_j$ において，それらの共通上位階層が $h_{ij}$ であるとすると，$I(h_{ij})$ は $e_i$ と $e_j$ との類似度を表す（値が大きいほど類似度が高い）．

上記の情報量の値は与えられたパスウェイ集合に依存するが，最上位階層 $[*]$ についてはつねに $p([*]) = 1$ であるため，その情報量は $\log_2(1/3705) - 0$（現在，EC 番号は 3705 種類ある）で，つねに $-11.85$ となる．また，任意の酵素階層 $h$ に対して，それより 1 階層分低い階層である酵素階層が $h_1, h_2, \ldots, h_n$ であるとすると，$C(h)$ は階層 $h$ 以下の階層に含まれるすべての酵素の集合の要素数であるから $C(h) \geq C(h_i)$ $(1 \leq i \leq n)$ であり $\log_2(1/C(h)) \leq \log_2(1/C(h_i))$ が成り立つ．また $p(h)$ は与えられたパスウェイ集合において，階層 $h$ 以下の階層に含まれるすべての酵素の出現頻度の合計であるので，$-\log_2 p(h) \leq -\log_2 p(h_i)$ が成り立つ．以上のことから，上位と下位の階層の間では $I(h) \leq I(h_i)$ が成り立つ．

ここで定義した情報量は，配列におけるアライメントで使われるスコアリングマトリックスと以下のように対応する．

Durbin らは文献 10) において，任意の 2 つのアミノ酸 $a$ と $b$ の間の類似度が次式のように表せるとしている．

$$S(a,b) = \log \frac{p_{ab}}{q_a q_b} = \log(p_{ab}) - \log(q_a q_b) \quad (3)$$

ここで，分子の $p_{ab}$ は，$a$ と $b$ との間に共通祖先（$a$ か $b$ と同じであってもよいとする）が存在し，それが $a$ と $b$ に置換する確率（すなわち $a$ と $b$ とが進化的に関連している確率）であり，分母の $q_a q_b$ はランダムな置換により $a$ が得られる確率と $b$ が得られる確率の積（すなわち $a$ と $b$ が進化的に無関係にランダムに生じた確率）を表す．つまりこのスコアは，$a$ と $b$ とが進化的に関連する度合いを log-odds 比で表していると考えられる．

本論文で定義した情報量も，2 つの酵素が酵素階層上でより類似している確率（両者の共通上位階層以下の酵素が少ないほど類似度が高いと考える）とパスウェイ集合でランダムにマッチする確率との log-odds 比と考えることができる．

### 3.2 ペアワイズアライメント

本研究では，階層構造を持つ局所最適なペアワイズアライメントからパターンを作成するために，動的計画法に基づく大域アライメントアルゴリズム[11]を拡張する．

動的計画法による大域アライメントは，アライメントの対象となる系列の各要素を 2 次元に配置し，それぞれの要素間に頂点をおき，一番左上の頂点を出発点とし，これらの辺を通って右下の頂点を目標点とする経路を求める．

アライメントのスコアは，**図 4** で表される経路において，右下斜めの矢印を通るときそれに対応する共通上位階層（**図 5** 参照）の情報量，右または下の矢印を通るときギャップの情報量をそれぞれ加えていったものとする．最適なアライメントは，ここでは最も高いスコアをとるアライメントを指す．パスウェイにおけるギャップとは，任意の酵素を表す．ギャップの情報量は $[*]$ より小さいことが望ましいと考え，ここではギャップの情報量を，$-15$ とした．

最適なアライメントがとる経路に現れる酵素階層またはギャップの列を，そのアライメントに対応するパターンと定義する．パターン $p$ の持つ情報量 $I(p)$ はアライメントのスコアと同一とする．

たとえば，パスウェイとして [2.4.2.3] [3.5.4.5] [3.1.3.5] [2.7.4.14] と [2.4.2.4] [3.1.3.5] [2.7.4.9] の酵素の系列があるとき，それらのアライメントに対応するパターン [2.4.2] "-" [3.1.3.5] [2.7.4] とその情報量との関係を**図 6** に示す．ギャップ "-" はこの場合

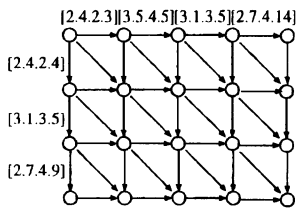


**図 4** 2 つのパスウェイのアライメントの経路
Fig. 4 Paths of alignment between two pathways.

| | [2.4.2.3][3.5.4.5] | [3.1.3.5] | [2.7.4.14] |
|---|---|---|---|
| [2.4.2.4] | [2.4.2] | [*] | [*] | [2] |
| [3.1.3.5] | [*] | [3] | [3.1.3.5] | [*] |
| [2.7.4.9] | [2] | [*] | [*] | [2.7.4] |

**図 5** 共通上位階層参照表．パスウェイ [2.4.2.3] [3.5.4.5] [3.1.3.5] [2.7.4.14] と [2.4.2.4] [3.1.3.5] [2.7.4.9] の場合
Fig. 5 Common higher rank class reference table.

| Pathway1 | Pathway2 | Pattern | MatchScore |
|---|---|---|---|
| [2.4.2.3] | [2.4.2.4] | [2.4.2] | −3.40 |
| [3, 5, 4, 5] | | "-" | −15 |
| [3.1.3.5] | [3.1.3.5] | [3.1.3.5] | 1.80 |
| [2.7.4.14] | [2.7.4.9] | [2.7.4] | −2.51 |

**図 6** 2つのパスウェイのアライメントの例．アライメントされた要素どうしが，図 2 の酵素階層のうちどの階層に分類されたかを 3 列目に，その情報量を 4 列目に示した．アライメントの結果は，酵素 ([3.1.3.5]) と酵素階層 ([2.4.2] や [*])，ギャップ "-" で構成される．

Fig. 6 Example of an alignment between two pathways, showing construction of a covering pattern from the minimally inclusive class covering both aligned elements (see Fig. 2) and the resulting information content between aligned pathway elements. The result of the alignment consists of sequences representing the enzymes (e.g. [3.1.3.5]), enzyme classes (e.g. [2.4.2] and [*]) and gapped positions, "-".

Pathway2 への 1 回の挿入を表す．したがって，図 6 で求められたパターン [2.4.2] "-" [3.1.3.5] [2.7.4] の情報量は $-3.40-15+1.80-2.51$ で $-19.11$ となる．

この手続きにかかる計算量は，パスウェイの最大長を $l$ としたとき，$O(l^2)$ となる．

### 3.3 マルチプルアライメントへの拡張

動的計画法による 2 本のパスウェイに対するペアワイズアライメントアルゴリズムを，3 本以上のパスウェイに拡張することを考える．マルチプルアライメントは，3 本以上のパスウェイの集合が与えられたとき，ペアワイズアライメントを集合中の任意の 2 つの要素に対して行い，そこで得られたパターンを元の集合に加え，さらにペアワイズアライメントを繰り返し，最終的に最適なマルチプルアライメントが得られるまで繰り返す（ペアワイズアライメントを 1 回行うごとに集合の要素が 1 つ減ることに注意）．

ここで最適なマルチプルアライメントとは，次に述べるパターン集合の情報量を最大にするものである．パターン集合の情報量を考える際，単純に集合中のすべてのパターンの情報量の合計とすると，ペアワイズアライメントの繰返しによりパターンの数が減少することにより情報量が減少してしまうので望ましくない．また，ペアワイズアライメントでは，アライメント前のパターン中のそれぞれの酵素階層に対して，アライメントにより得られるパターンでは多くの場合 1 つ上の酵素階層に上がると考えられる．そこで，本論文では，1 階層分酵素階層が異なる場合に，およそ各情報量が $w$ 違うならば，ペアワイズアライメントによりパターン集合の数が 1 つ減った場合に，その減った分に相当する値を加えることを考えた．

具体的には，パターン集合 $P$ の情報量 $I(P)$ は，$P$ 中の各パターン $p$ が $S$ 中の $n_s$ 個のパスウェイのマルチプルアライメントから得られたとき，$k=\#P$, $n=\#S$ とするならば，以下のように計算する（集合 $X$ に対して $\#X$ は $X$ における要素の数を表す）．

$$I(P) = w\bar{l}(n-k) + \sum_{p\in P} \frac{n_s}{n} I(p) \qquad (4)$$

定数 $\bar{l}$ は与えられたパスウェイの平均長を，定数 $w$ は，アライメントの 1 要素あたりに与えられる重みである．$w\bar{l}(n-k)$ は，パターン集合 $P$ の要素数が少ないほど大きな数となる．実際に，1 階層分異なる酵素階層どうしはおよそ 5 だけ情報量が異なることが分かった．そこで，本論文では，$w=5$ と設定することにした．

上記のスコア最大のマルチプルアライメントを可能なすべての組合せについて求めようとすると多次元のアライメントが必要になり計算量が

$$O\left(\sum_{k=1}^{n-1}\binom{n}{n-k+1} l^{n-k+1}\right)$$

となるため，大きな $n$ の値に対しては現実的ではない．配列のマルチプルアライメントでは貪欲算法が多く用いられている（たとえば，文献 12)）ため，本研究でも以下に述べるように貪欲算法に基づくマルチプルアライメントアルゴリズムを実現した（**図 7** 参照）．

$n$ 個のパスウェイが与えられたとき，$k=n-1$ 個

```
procedure IMA
  input: A set of pathways S;
  output: A set of patterns P;
  n := #S; P := S; P_0 := S; max := I(P);
  for k := n − 1 downto 1 do
    foreach p_i ∈ P do
      foreach p_j ∈ P (p_i ≠ p_j) do
        p := G(p_i, p_j);
        P' := (P − {p_i, p_j}) ∪ {p};
        if max < I(P') then
          P_{n−k} := P';
          max := I(P');
      endforeach
    endforeach
    if I(P_{n−k}) < I(P_{n−k−1}) then
      return P;
    P := P_{n−k};
  endfor
  return P;
```

**図 7** 情報量最大化マルチプルアライメントアルゴリズム．$p := G(p_i, p_j)$ はパスウェイ $p_i$ と $p_j$ のペアワイズアライメントを行い 1 つのパターン $p$ を求める手続きを表す．

Fig. 7 Information content maximization alignment algorithm: $p := G(p_i, p_j)$ is a procedure which performs pairwise alignment of two pathways $p_i$ and $p_j$ and calculates a pattern $p$.

のパターンを得るマルチプルアライメントを，与えられたパスウェイ集合から任意の2個の要素を選んでアライメントを行ったときの最も情報量の大きいパターンを選ぶ操作とし，次の $k = n - 2$ 個のパターンを得るマルチプルアライメントでは，前の操作で得られた結果（$n$ 個の中から2個を除いて，新たに作った1個のパターンを加えたもの）を新たな集合として同じ操作を行う．以下これを繰り返し，$k = 1$ となれば停止する（解として単一のパターンが得られる）．あるいは，繰返しの過程で，直前のパターン集合（図7の $P_{n-k-1}$）の情報量と比べて小さい情報量のパターン集合しか得られなくなったときは，解として直前のパターン集合を返して停止する．

この手続きの計算量は，正例の最大長を $l_{max}$ とするとき $O(n^3 l_{max}^2)$ となる．

## 4. 実　　験

本アルゴリズムの有効性を確かめるため，解糖や，DNA または RNA の複製，アミノ酸の分解にかかわる経路を KEGG の代謝マップから抽出し，それらのパスウェイに対して本アルゴリズムを適用した．図7のアルゴリズムの実装および上記のパスウェイ実験は DEC AlphaStation 600 5/333（CPU は Alpha 21164 333 MHz，OS は Tru64 UNIX で gcc を使用）上で行った．

抽出したパスウェイを**図8**，**図9**，**図10**に示す．これらの図では，パスウェイを表形式で表しており，表の各行がそれぞれ1つのパスウェイを酵素の EC 番号の系列として表したものである．

図8の各パスウェイは，図1のような解糖系のパスウェイの中で，α-D-Glucose-6P や L-Fucose，L-Rhamnose，D-Mannose-6P を基質としてとる酵素から始まるものであり，パスウェイ中の各酵素は KEGG によりその EC 番号が大腸菌ゲノム中の ORF との対応がとられているものだけを選んでいる．

このデータに対して図7のアルゴリズムを適用することにより，1つのパターン [5.3.1] [2.7.1] [4.1.2] が得られた．2章で述べたように，酵素の EC 番号の階層は反応の類似性を表していると考えられる．このパターンはいずれも図8のパスウェイ集合中で EC 番号の1段階上の階層で表現できており，妥当であると考えられる．なお，得られた解の情報量は 32.56，アライメントの計算時間は 0.005 秒であった．

図9の各パスウェイは，プリン塩基もしくはピリミジン塩基を基質として，DNA もしくは RNA を複製するパスウェイ（最後の 2.7.7.7 と 2.7.7.6 はそれぞれ

| | | |
|---|---|---|
| 5.3.1.9 | 2.7.1.11 | 4.1.2.13 |
| 5.3.1.25 | 2.7.1.51 | 4.1.2.17 |
| 5.3.1.14 | 2.7.1.5 | 4.1.2.19 |
| 5.3.1.8 | 2.7.1.11 | 4.1.2.13 |

**図8** 大腸菌のグルコース，フコース，ラムノース，マンノースの解糖系パスウェイ
Fig. 8 Glucose, fucose, rhamnose and mannose degradation pathways in *Escherichia coli*.

| | | | | | | |
|---|---|---|---|---|---|---|
| Purine base | 2.4.2.1 | 3.1.3.5 | 2.7.4.3 | 2.7.4.6 | 2.7.7.7 | | DNA |
| | 2.4.2.1 | 3.1.3.5 | 2.7.4.3 | 2.7.1.40 | 2.7.7.7 | | |
| | 2.4.2.1 | 3.1.3.5 | 2.7.4.8 | 2.7.4.6 | 2.7.7.7 | | |
| | 2.4.2.1 | 3.1.3.5 | 2.7.4.8 | 2.7.1.40 | 2.7.7.7 | | |
| | 2.4.2.1 | 3.1.3.5 | 2.7.4.3 | 2.7.4.6 | 2.7.7.6 | | RNA |
| | 2.4.2.1 | 3.1.3.5 | 2.7.4.3 | 2.7.1.40 | 2.7.7.6 | | |
| | 2.4.2.1 | 3.1.3.5 | 2.7.4.8 | 2.7.4.6 | 2.7.7.6 | | |
| | 2.4.2.1 | 3.1.3.5 | 2.7.4.8 | 2.7.1.40 | 2.7.7.6 | | |
| | 2.4.2.4 | 3.1.3.5 | 2.7.4.9 | 2.7.4.6 | 2.7.7.7 | | |
| Pyrimidine base | 2.4.2.1 | 3.5.4.5 | 3.1.3.5 | 2.7.4.14 | 2.7.4.6 | 2.7.7.7 | DNA |
| | 2.4.2.3 | 3.5.4.5 | 3.1.3.5 | 2.7.4.14 | 2.7.4.6 | 2.7.7.6 | |
| | 2.4.2.4 | 3.1.3.5 | 2.7.4.14 | 2.7.4.6 | 2.7.7.6 | | RNA |

**図9** 大腸菌の DNA または RNA を複製するパスウェイ．プリン塩基を基質とするパスウェイは8本，ピリミジン塩基を基質とするパスウェイは4本．
Fig. 9 DNA and RNA replication pathways in *Escherichia coli*: 6 purine pathways; 4 pyrimidine pathways.

| | | | | | | |
|---|---|---|---|---|---|---|
| Eco | + | 4.2.1.17 | 1.1.1.35 | 2.3.1.16 | | |
| | + | 4.2.1.17 | 1.1.1.57 | 2.3.1.9 | | |
| Afu | + | 4.2.1.17 | 1.1.1.35 | 2.3.1.9 | | |
| | + | 1.3.99.3 | 4.2.1.17 | 1.1.1.35 | 2.3.1.16 | |
| Cel | | 1.2.4.2 | 2.3.1.61 | 1.3.99.7 | 4.2.1.17 | 1.1.1.35 | 2.3.1.9 |
| | | 1.2.4.2 | 1.3.99.7 | 4.2.1.17 | 1.1.1.35 | 2.3.1.9 | |
| | + | 2.3.1.2 | 1.3.99.3 | 4.2.1.17 | 1.1.1.35 | | |
| | + | 2.3.1.2 | 1.3.99.6 | 4.2.1.17 | 1.1.1.35 | | |
| | + | 1.3.99.2 | 4.2.1.17 | 1.1.1.35 | | | |

**図10** 異なる生物種でのイソロイシン，リシン，トリプトファンなどのアミノ酸を分解するパスウェイ．Eco は大腸菌，Afu は好熱性硫酸塩還元古細菌，Cel は線虫を表す．"+" と書かれた行のパスウェイには基質に到達する経路が存在しない．
Fig. 10 Isoleucine, lysine, tryptophan and other degradation pathways in different organisms: Eco, *Escherichia coli*; Afu, *Archaeoglobus fulgidus*; Cel, *Caenorhabditis elegans*. The pathway marked by "+" does not have a complete pathway from its substrate.

DNA と RNA の複製酵素）であり，パスウェイ中の各酵素は KEGG によりその EC 番号が大腸菌ゲノム中の ORF との対応がとられているものだけを選んでいる．このデータに対して図7のアルゴリズムを適用することにより，1つのパターン [2.4.2] "-" [3.1.3.5] [2.7.4] [2.7] [2.7.7] が得られた．

図7のアルゴリズムの for ループで最終的にこのパターンが得られる直前で，パターン [2.4.2] [3.1.3.5] [2.7.4] [2.7] [2.7.7] とパターン [2.4.2] [3.5.4.5] [3.1.3.5] [2.7.4.14] [2.7.4.6] [2.7.7] のアライメントが行われており，このとき情報量は 239.67 から 245.94 に増加している．この結果は，ギャップを許したパスウェイの

アライメントによりDNAまたはRNAの複製反応のパスウェイの一部に[3.5.4.5]で触媒される反応が含まれていることをうまく表現できていると考えられる．なお，得られた解の情報量は245.94，アライメントの計算時間は0.15秒であった．

図10のパスウェイは，イソロイシン，リシン，トリプトファンなどのアミノ酸を分解するパスウェイから抽出したものであり，KEGGにより大腸菌，好熱性硫酸塩還元古細菌，線虫のそれぞれのORFと対応がとられているEC番号を選んでいる．

このデータに対して図7のアルゴリズムを適用することにより，3つのパターン

{ [4.2.1.17] [1.1.1.] [2.3.1],
"-" [1.3.99] [4.2.1.17] [1.1.1.35] "-",
[1.2.4.2] "-" [1.3.99.7] [4.2.1.17] [1.1.1.35] [2.3.1.9]
}

が得られた．

これらのうちで，最初のパターンは，大腸菌の3つのパスウェイと対応している．図7のアルゴリズムの実行過程では，これら3つを組み合わせて，2つのパターン "-" "-" [1.3.99] [4.2.1.17] [1.1.1.35] "-" と [4.1.2.17] "-" [1.1.1] [2.3.1] を得ようとしたが，情報量が56.70から52.46に減少するため，この段階で繰返しを終えている．なお，得られた解の情報量は56.70，アライメントの計算時間は0.04秒であった．

## 5. おわりに

木構造を持つ記号に対して拡張されたマルチプルアライメントアルゴリズムを提案し，実際に解糖やアミノ酸分解，DNAの複製に関するパスウェイを用いて実験することで，その妥当性を確かめた．今回の実験では，基質と生成物を基準に一部，手作業でパスウェイの実験データを集めた．今後，この部分を自動化する必要がある．また，パスウェイ解析への応用において，今回提案したパターンの情報量という評価基準の有効性をより詳細に検証することが今後の課題としてあげられる．

**謝辞** 本研究は一部，文部省科学研究費補助金特定領域研究「ゲノムサイエンス」(課題番号08283103)，科学技術振興事業団戦略基礎研究推進事業および計算科学技術活用型特定研究開発推進事業によっている．本アルゴリズムの応用として重要なヒントをいただいた京都大学の西岡先生に感謝いたします．また，階層を持つパターンの定義と妥当性についてアドバイスしていただいた，九州工業大学の篠原武先生，下園真一先生，九州大学の有村博紀先生に感謝いたします．

## 参考文献

1) Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M.: KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Research*, Vol.27, No.1, pp.29–34 (1999). Available at: http://www.genome.ad.jp/kegg/
2) Dandekar, T., Schuster, S., Snel, B., Huynen, M. and Bork, P.: Pathway Alignment: Application to the Comparative Analysis of Glycolytic Enzymes, *Biochemical J.*, Vol.343, No.1, pp.115–124 (1999).
3) Mavrovouniotis, M.L.: Identification of Qualitatively Feasible Metabolic Pathways, *Artificial Intelligence and Molecular Biology*, Hunter, L. (Ed.), pp.325–364, AAAI Press/MIT Press, Menlo Park (1993).
4) Gaasterland, T. and Selkov, E.: Reconstruction of Metabolic Networks Using Incomplete Information, *Proc. Intl. Conf. on Intelligent Systems for Molecular Biology*, pp.127–135 (1995).
5) Goto, S., Bono, H., Ogata, H., Fujibuchi, W., Nishioka, T., Sato, K. and Kanehisa, M.: Organizing and Computing Metabolic Pathway Data in terms of Binary Relations, *Pacific Symp. Biocomputing 97*, pp.175–186 (1997).
6) Bono, H., Ogata, H., Goto, S. and Kanehisa, M.: Reconstruction of Amino Acid Biosynthesis Pathways from the Complete Genome Sequence, *Genome Research*, Vol.8, No.3, pp.203–210 (1998).
7) Forst, V.C. and Schulten, K.: Evolution of Metabolisms: A New Method for the Comparison of Metabolic Pathways Using Genomics Information, *J. Computational Biology*, Vol.6, No.3, pp.343–360 (1999).
8) Galperin, M.Y., Walker, D.R. and Koonin, E.V.: Analogous Enzymes: Independent Inventions in Enzyme Evolution, *Genome Research*, Vol.8, No.8, pp.779–790 (1998).
9) 今堀和友，山川民夫（監修）：生化学辞典，第2版，付録1-6 酵素の分類と命名法，pp.1494–1495，東京化学同人 (1990).
10) Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G.: *Biological Sequence Analyses: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge (1998).
11) Needleman, S.B. and Wunsch, C.D.: A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins, *J. Mol. Biol.*, Vol.48, pp.444–453 (1970).
12) Feng, D. and Doolittle, R.F.: Progressive Se-

quence Alignment as a Prerequisite to Correct Phylogenetic Trees. *J. Mol. Evol.*, Vol.25. pp.351–360 (1987).

(平成11年 8 月14日受付)
(平成11年12月20日再受付)
(平成12年 1 月14日採録)

**遠里由佳子**（学生会員）

昭和47年生．平成7年九州工業大学情報工学知能情報工学科卒業．平成9年同大学院情報工学研究科情報科学専攻（修士課程）修了．同年三菱電機伊丹製作所に入社，平成11年3月退職．現在大阪大学大学院基礎工学研究科情報数理系専攻（博士課程）在学中．

**松田 秀雄**（正会員）

昭和34年生．昭和57年神戸大学理学部物理学科卒業．昭和59年同大学院工学研究科システム工学専攻（修士課程）修了．昭和62年同大学院自然科学研究科（博士課程）修了．同年同大学工学部助手となり，同大学講師，助教授を経て，平成6年10月より大阪大学基礎工学部情報工学科助教授，現在に至る．この間，平成3年4月より10カ月間米国アルゴンヌ国立研究所客員研究員．学術博士．論理型言語による並列処理，遺伝子情報処理の研究に従事．電子情報通信学会，IEEE CS，ACM各会員．

**橋本 昭洋**（正会員）

昭和36年大阪大学工学部通信工学科卒業．昭和41年同大学院工学研究科博士課程修了．工学博士．同年NTT電気通信研究所に勤務．昭和44～46年イリノイ大学計算機科学科客員助教授．昭和60年NTTデータ処理研究部長，昭和62年情報科学研究部長．この間計算機の故障診断，自動設計，大型計算機DIPSの開発等に従事．平成1年大阪大学基礎工学部情報工学科教授，平成6年同大学情報処理教育センター長を併任，現在に至る．最近は分子生物学関連の情報処理技術の研究に従事．著書「計算機アーキテクチャ」（平成7年，昭晃堂）．電子情報通信学会，IEEE，ACM各会員．