2007－BIO－8 (5)
2007／3／5

化合物の構造式の類似性に基づくパスウェイアライメント

遠里 由佳子 †, 西村 悠 ††

† 立命館大学 情報理工学部 生命情報学科 †† 立命館大学 理工学部 情報学科

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上で重要な情報となる．そこで，本研究では，パスウェイを構成する酵素反応の類似性を，それを構成する化合物の構造式の類似性としてとらえるパスウェイのアライメントアルゴリズムを提案する．化合物の構造式の類似は，構造式の部分構造の定義である MACCS Key と Tanimoto 係数を用いて計算する．実際に，大腸菌の代謝反応ネットワーク全体から，経路をダイクストラのアルゴリズムを用いて抽出し，取り出した全経路に対してアライメントを適用した．その結果，フルクトースとマンノースの生合成パスウェイと，ガラクトースの生合成パスウェイにおいて，類似した構造式変化の系列を見ることができ，その有効性を確かめた．

Metabolic pathway alignment based on similarity between chemical structures

Yukako Tohsato † Yu Nishimura ††

† Department of Bioscience and Bioinformatics,
College of Information Science and Engineering, Ritsumeikan University
†† Department of Computer Science, College of Science and Engineering,
Ritsumeikan University

In many of the chemical reactions in living cells, enzymes act as catalysts in the conversion of certain compounds (substrates) into other comopunds (productes). As the product of a reaction is used as the substrate of other reactions, metabolic pathway is formed. Comparative analyses of the metabolic pathways among species give important information on evolution and on pharmacological targets. In this paper, we propose a method to align the metabolic pathways based on similarity between chemical structures. To measure the chemical similarity, we formalized a scoreing system by using the MACCS keys and the Tanimoto coefficients. The effectiveness of our method is demonstrated by applying the method to pathway analyses of metabolic pathways in *Escherichia coli*. By their result, we have found compound similarity between the fructose and mannose biosynthesis and the galactose biosynthese.

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．これらの化学反応は，ある反応の生成物が別の反応の基質となることで，複雑な反応経路のネットワークを形成する．そして，それらの既存の知見は，KEGG[1] や MetaCyc[2] 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代謝反応パスウェイ（以下，パスウェイと省略）と呼ぶ．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，化合物を合成する方法についての知見を得るためや，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を知る上で重要となる [3].

本研究では特に，代謝反応パスウェイの進化を説明する仮説の 1 つである「パスウェイ重複」(pathway duplication) に着目する．パスウェイ重複は，タンパク質の機能が互いに関連した一連の遺伝子が，まとめてコピーされたのちに，別の機能を持つパスウェイに進化する可能性を示唆する仮説である

[4]．パスウェイ重複を見るためにパスウェイを比較する場合は，酵素間の配列類似性を用いることが多い．しかし，酵素の機能すなわち酵素にラベル付けされたEC(Enzyme Commission)番号[5]が同じでも配列が全く異なる場合（enzyme recruitment[6]）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，文献[7]において，EC番号の機能階層を用いたパスウェイのアライメント手法を提案した．しかし，EC番号は反応の人為的に分類されたものであり，その分類には大きな偏りがある．また，代謝反応ネットワークのデータにはEC番号が付けられていない反応も多い．例えば，大腸菌*Escherichia coli K-12 MG1655*が持つ代謝反応は1115であるが，うちEC番号が付けられていないもの（[1.－.－.－]といったEC番号が一部特定されていないものを含む）は123存在し，全体の約11%を占める．

そこで，本論文では，化合物の構造式に着目する．化合物の構造式の比較は，最大共通部分グラフ問題の一種となるため，NP困難なことが知られており，化合物の構造を計算機上で扱いやすくするために，いくつかの記述子が提案されている[8]．本研究ではその記述子をもとに，化合物間さらには反応間の類似度を定義し，化合物の構造式に基づくパスウェイアライメントアルゴリズムを提案する．そして，実際に提案手法を，大腸菌（*Escherichia coli K-12 MG1655*）の代謝反応パスウェイデータに適用した結果について示す．

2 化合物の構造式に基づくアライメントアルゴリズム

2.1 代謝反応パスウェイのグラフ表現

KEGGデータベースの情報に基づき，代謝反応をグラフで表現する．KEGGでは，生物種ごとに，それを構成する代謝反応rは，反応に必要な化合物の集合を，$s_1, s_2, \ldots s_m$，生成される化合物の集合を，$p_1, p_2, \ldots, p_n$とするとき，

$$r : s_1, s_2, \ldots s_m + p_1, p_2, \ldots p_n$$

の形で保存されている．ただし，すべての酵素反応は可逆反応とみなされ，基質および生成物の関係が入れ替えても同じ酵素反応rとする．このようなKEGGデータのもと，入力として，生物種および，出発化合物と最終化合物が与えられた場合に，その間の酵素反応をつなげてできる代謝反応パスウェイを，以下のようなグラフとして用いる．

生物種Xの化合物の集合をC，代謝反応の集合をRとする．このとき，代謝反応（以下，反応と省略する）を化合物間の二項関係としてみると，代謝経路は，頂点集合をC，辺集合をRとする無向グラフ$G = (C, R)$とみなすことができる．反応$r \in R$が化合物$c_1 \in C$から化合物$c_2 \in C$を合成もしくは分解する場合に，$c_1 \overset{r}{\rightarrow} c_2$などと書くことにし，反応$r$を場合によっては化合物のペア$(c_1, c_2)$と同一視することする．

入力として，出発となる化合物c_1と最終生成物である化合物c_mが与えられた場合に，KEGGデータに基づき，パスウェイを，化合物のペアの系列として抽出することを考える．ある反応rに隣接する反応とは，反応rの生成物のうちの少なくとも1つを基質とすることができるr以外の反応とする．

そして，隣接する反応の系列を$r_1 r_2 \ldots r_m$で表し，連接する反応の化合物のペアの系列$(c_1, c_2)(c_2, c_3) \ldots (c_{m-1}, c_m)$と同一視する．このとき，パスウェイの経路長は，化合物のペアの個数する．

2.2 化合物間の構造式とその類似度

反応を化合物の組として考たときの，反応間の類似度を定義するために，まず化合物の構造式の類似度について言及する．

化合物の構造を計算機上で扱いやすくするために，いくつかの記述子が提案されている．広く使われている記述子は，特定の分子構造が分子内に存在するかどうかを示したフィンガープリント（fingerprint）であり，こうした表現は計算機上で拘束で効果的な類似化合物の探索を可能とする．

本研究では，MACCS key[9]という記述子を用いた．MACCS keyは，166個の部分構造を設定している．MACCS keyを用いると，化合物の構造式はその部分構造の有無（該当する部分構造が存在するときは1，存在しないときは0）により166ビットのビット列として表現できる．以後，化合物cの構造式のビット列を$B(c) = (x_1, x_2, \cdots, x_n)$であらわす．

2つのビット列間の類似度にはさまざまな計算方法が提案されている[8]．ここでは，Tanimoto

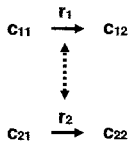

図 1: 反応間の類似性の構成例

係数を用いる．Tanimoto 係数は，ビット列間の類似度を表す指標としてよく用いれており，2 つのビット列 $B(c_1) = (x_{11}, x_{12}, \ldots, x_{1n})$ と $B(c_2) = (x_{21}, x_{22}, \ldots, x_{2n})$ に対して，次の式で与えられる．

$$T(c_1, c_2) = \frac{B(c_1) \cap B(c_2)}{B(c_1) \cup B(c_2)} \quad (1)$$

ここで，$B(c_1) \cup B(c_2)$ は各ビット列でどちらかが 1 となっているビットの数を，$B(c_1) \cap B(c_2)$ は両方のビット列でビットが 1 となっている数を指す．定義により，$T(c_1, c_2)$ は 0 から 1 までの数となり，1 に近づくほど 2 つのビット列間の類似度が高く，0 に近づくほど 2 つのビット列間の類似度が低いことを表す．

そして，図 1 のような反応 $c_{11} \xrightarrow{r_1} c_{12}$ と反応 $c_{21} \xrightarrow{r_2} c_{22}$ $(c_{ij} \in C, r_i \in R)$ の間の類似度 $S(r_1, r_2)$ は，対応する化合物どうしの類似度の平均によって計算する．

$$S(r_1, r_2) = \frac{T(c_{11}, c_{21}) + T(c_{12}, c_{22})}{2} \quad (2)$$

2.3 パスウェイのアライメントアルゴリズム

本研究では，代謝反応パスウェイを反応の系列とみなすことで，反応の系列どうしを比較して類似度を計算するために，動的計画法に基づくローカルアライメント (Smith-Waterman アルゴリズム)[10] を拡張する．動的計画法によるローカルアライメントアルゴリズムでは，x,y を長さ m と n の 2 つの反応の系列とするとき，(i, j) 成分を持つ $(m+1) \times (n+1)$ 行列を用いる．

$$M[i,j] = \max \begin{cases} 0, \\ M[i-1, j-1] + S(r_i, r_j) \\ M[i, j-1] + d, \\ M[i-1, j] + d, \end{cases} \quad (3)$$

ただし，$M[0, j] = 0$, $M[i, 0] = 0$ とする．また，$S(r_i, r_j)$ は，本研究では，反応 r_i と r_j 間の類似度とする．右または下の矢印を通るときギャップのペナルティ d をそれぞれ加えていったものとする．本研究では $d = -1$ で計算した．

```
procedure PathwayAlignment
  input: p1 = r11 r12 ... r1m, p2 = r21 r22 ... r2n;
  output: 最適なアライメント, アライメントのスコア;
  for i := 0 to m do M[i,0] := 0;
  for j := 0 to n do M[0,j] := 0;
  max := 0; i_max := 0; j_max := 0;
  for i := 1 to m do
    for j := 1 to n do
      M[i,j] := max { 0,
                    { M[i-1,j-1] + S(r_i, r_j)
                    { M[i,j-1] + d,
                    { M[i-1,j] + d,
      if max > L[i,j] then
        max := L[i,j]; i_max := i; j_max := i;
  i := i_max; j := j_max;
  align_p1 = ""; align_p2 = "";
  while (M[i,j]! = 0)
    if M[i,j] = M[i-1,j-1] + S(r1i, r2j) then
      align_p1 の先頭に r1i を追加;
      align_p2 の先頭に r2j を追加;
      i := i - 1;
      j := j - 1;
    else if M[i,j] = M[i-1,j] + d then
      align_p1 の先頭に r1i を追加;
      align_p2 の先頭に "-" を追加;
      i := i - 1;
    else M[i,j] = M[i,j-1] + d
      align_p1 の先頭に "-" を追加;
      align_p2 の先頭に r2j を追加;
      j := j - 1;
  endwhile
  return align_p1, align_p2, max;
```

図 2: パスウェイのアライメントアルゴリズム

ただし，$M[0, j] = 0$, $M[i, 0] = 0$ とする．また，$S(r_i, r_j)$ は，本研究では，反応 r_i と r_j 間の類似度とする．右または下の矢印を通るときギャップのペナルティ d をそれぞれ加えていったものとする．本研究では $d = -1$ で計算した．

最適なアライメントは，ここでは最も高いスコアを取るアライメントを指す．ローカルアライメントでは，行列 M でスコアが 0 となる点を出発点として，最もスコアが高い点を目標点とする経路と等価になる．ローカルアライメントを求める場合，行列 M を計算したあと，各成分が最大値としてどの成分から値をとったかをたどることで，対応する最適なアライメントを求めることができる．このアルゴリズムを図 2 に示す．アライメント全体の時間計算量は $0(mn)$ となる．パスウェイにおけるギャップとは，任意の反応を表す．ギャップのペナルティ値 d は -1 とした．

アライメントのスコアは，入力の対象となる系列が長くなるほどスコアが高くなるという性質がある．そこで，本研究ではアライメントのスコアを，取り出されたアライメント長で割ったものを，補正後のアライメントのスコアとする．

3 実験と結果

3.1 実験データとパスウェイデータの作成

KEGG (2006 年 10 月 Version 40.0) に登録された大腸菌 (*Escherichia coli K-12 MG1655*) のデータを用いて実際に解析を行った．なお,KEGG のデータには，化合物の構造式のビット列データは存在しない．そこで，MESA の Fingerprint Module[11] を用いた．これは化合物の構造式の SMILES 記法 (`http://www.daylight.com`) のデータから MACCS key の 166 ビットのうち 164 ビットを生成するツールである．これを用いて，KEGG の化合物と PubChem (`http://pubchem.ncbi.nlm.nih.gov/`) での SMILES データを対応させ，その SMILES データから 164 ビットのビット列データを作成した．

同じ代謝マップに存在する二つ化合物間のパスウェイは，最短経路のアルゴリズムを用いて抽出することができる [12][13]．本研究では，最短経路を求める古典的なアルゴリズムの一種であるダイクストラ法を用いて，代謝マップごとのすべての化合物間のパスウェイを求めた．1 つの反応はすべて長さ 1 で重み付けして計算した．このとき，抽出された最短経路が，水や ATP などの化合物を経由して生物学的に正しくない経路をとる可能性がある [12][13]．これをできるだけ回避するため，本研究では KEGG 中の reaction_main と呼ばれる，すべての反応データから主要な反応のみを取り出した反応データを用いた．なお，本来同じ代謝マップの分類が異なるが本来は同じ経路どうしが入力の対象となることを防ぐために，表 1 に示す互いに独立な 37 の代謝マップ [14] に限定して経路を抽出した．その結果 15444 のパスウェイが得られた．

3.2 アライメント結果と考察

大腸菌から取り出した 15444 のパスウェイを用いて，代謝マップの分類が異なる任意の 2 つのパスウェイに対してアライメントを行った． なお，本論文で提案するアライメントのスコアはそのアライメント長で割るため，結果的に短いアライメント結果が，スコア上位となる可能が高い．また，パスウェイが重複したパスウェイ同士のアライメントのスコアが高くなりやすい．そこで，代謝マップが異なる 2 つのパスウェイのアライメントの結果で，その補正後のアライメントのスコアが，0.95 以上になるアライメント結果を求め，アライメント長別に分類した．

得られたアライメントの最大長は 5 となった．その中で長さ 5 と長さ 4 のアライメント結果を表 2 に示す．表中のスコアは補正後のアライメントのスコアである．代謝マップ ID は，代謝マップの分類を表し，表 1 の代謝マップ名と対応をとることができる．

得られたアライメント結果のうち，最もアライメント長が長いアライメント結果のスコアは，0.968 であり，フルクトースとマンノースの生合成パスウェイと，ガラクトースの生合成パスウェイのアライメントとなった (図 3 参照)．

図 3 中で，化合物にはその名前と KEGG における化合物番号，反応には，反応名と EC 番号，KEGG における反応番号をラベル付けした．反応の矢印の向きは，実際に起こりえる反応の流れを表現している．アライメントで対応の取れた反応の組み合わせは破線矢印で結び，その横に反応の類似スコアを示した．

図より，これらの化合物の構造変化が互いによく類似していることが確認できる．また，文献 [7] で示した EC 番号に基づくアライメント結果の経路と一部重複しており，その部分に関して，妥当な結果が得られたといえる．対応が取れた反応の中で，EC 番号 [3.1.3.22] が割り当てられた反応と，[2.7.1.69] が割り当てられた反応は，互いにその構造変化が似ているとみることができる．

3.3 提案する反応の類似スコアの評価

本節では，本論文で提案する反応間の類似スコアと，文献 [7] で提案された EC 番号に基づく類似スコアとの関係を調べた．

EC 番号に基づくスコアは，[1.1.1.3] と [1.1.2.4] という EC 番号がラベル付けされた反応 r_1 と r_2 があるとき，その共通の階層は [1.1] と考える (詳細は文献 [7] 参照)．共通の階層に含まれる反応数を n ，すべての反応の数を m とすると，

$$E(r_1, r_2) = -log_2 \frac{n}{m} \tag{4}$$

表 1: 実験に用いた代謝マップ 37 種類のリスト

代謝マップ ID	代謝マップ名	代謝マップ ID	代謝マップ名
MAP00010	Glycolysis / Gluconeogenesis	MAP00400	Phenylalanine, tyrosine and tryptophan biosynthesis
MAP00020	Citrate cycle (TCA cycle)	MAP00450	Selenoamino acid metabolism
MAP00030	Pentose phosphate pathway	MAP00500	Starch and sucrose metabolism
MAP00040	Pentose and glucuronate interconversions	MAP00520	Nucleotide sugars metabolism
MAP00051	Fructose and mannose metabolism	MAP00530	Aminosugars metabolism
MAP00052	Galactose metabolism	MAP00561	Glycerolipid metabolism
MAP00130	Ubiquinone biosynthesis	MAP00620	Pyruvate metabolism
MAP00220	Urea cycle and metabolism of amino groups	MAP00630	Glyoxylate and dicarboxylate metabolism
MAP00230	Purine metabolism	MAP00640	Propanoate metabolism
MAP00240	Pyrimidine metabolism	MAP00650	Butanoate metabolism
MAP00251	Glutamate metabolism	MAP00670	One carbon pool by folate
MAP00252	Alanine and aspartate metabolism	MAP00710	Carbon fixation
MAP00260	Glycine, serine and threonine metabolism	MAP00730	Thiamine metabolism
MAP00271	Methionine metabolism	MAP00760	Nicotinate and nicotinamide metabolism
MAP00280	Valine, leucine and isoleucine degradation	MAP00770	Pantothenate and CoA biosynthesis
MAP00330	Arginine and proline metabolism	MAP00790	Folate biosynthesis
MAP00340	Histidine metabolism	MAP00860	Porphyrin and chlorophyll metabolism
MAP00360	Phenylalanine metabolism	MAP00910	Nitrogen metabolism
MAP00362	Benzoate degradation via hydroxylation		

表 2: アライメント結果

(a) 長さ 5 のアライメント結果

スコア	代謝マップ ID	アライメント結果
0.968	MAP00051	C01131 C00111 C05378 C05345 C00644 C00392
	MAP00052	C01286 C00118 C03785 C01097 C06311 C01697

(b) 長さ 4 のアライメント結果

スコア	代謝マップ ID	アライメント結果
0.992	MAP00051	C00111 C05378 C05345 C00644 C00392
	MAP00052	C00118 C03785 C01097 C06311 C01697
0.992	MAP00051	C00118 C05378 C05345 C00644 C00392
	MAP00052	C00111 C03785 C01097 C06311 C01697
0.977	MAP00052	C00103 C00446 C00124 C00243 C05396
	MAP00500	C05345 C00668 C00267 C00208 C02995
0.960	MAP00051	C01131 C00111 C05378 C05345 C00644
	MAP00052	C01286 C00118 C03785 C01097 C06311
0.957	MAP00051	C01094 C05378 C05345 C00644 C00392
	MAP00052	C00118 C03785 C01097 C06311 C01697
0.957	MAP00051	C01094 C05378 C05345 C00644 C00392
	MAP00052	C00111 C03785 C01097 C06311 C01697
0.955	MAP00052	C00052 C00446 C00124 C05402 C00031
	MAP00500	C00029 C00103 C00089 C00267 C00208
0.955	MAP00052	C00052 C00446 C00124 C05400 C00159
	MAP00500	C00029 C00103 C00089 C00267 C00208

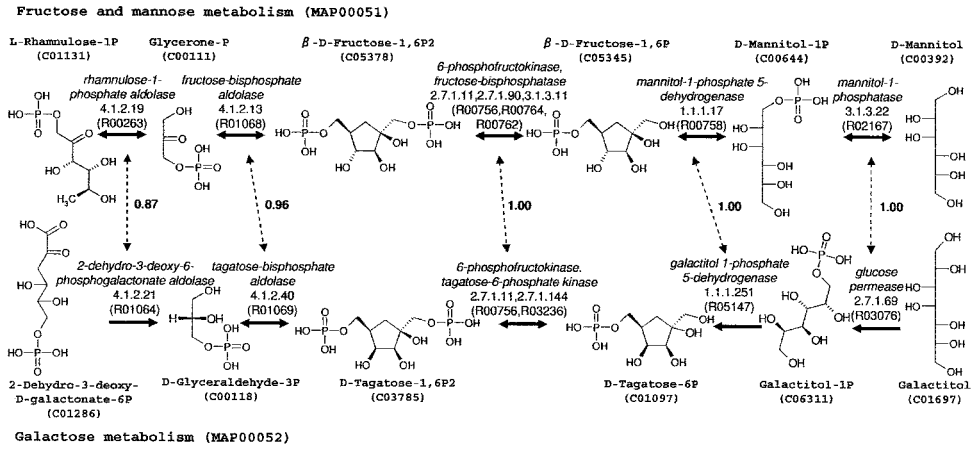

図 3: アライメント長 5 のスコアトップのアライメント結果

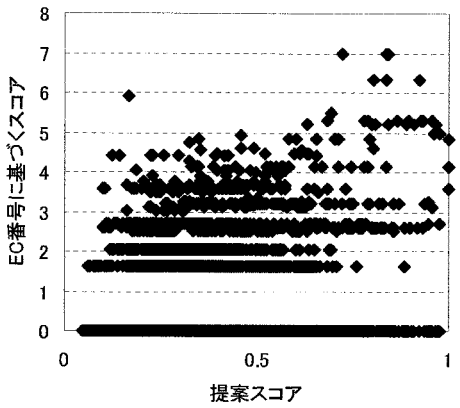

図 4: 提案スコアと EC 番号に基づく類似スコアの相関

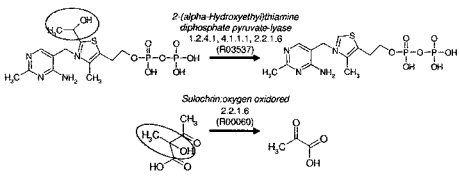

図 5: EC 番号に基づく類似スコアのほうが高くなる例

で計算する．最上位の階層もあっていない場合 (例えば [1.1.1.1] と [6.1.1.1] の共通階層はないとみなす) は，共通の階層に含まれる反応は，すべての反応とし，スコアは 0 となる．このスコアは，共通の階層に含まれる反応の数が少ないほど，その反応は類似しているとみなし，スコアは高くなる．なお，1 種類の反応に何種類もの EC 番号がラベルづけされる場合は，その共通階層に含まれる反応が少なくなる組み合わせ，つまりスコアが高くなる組み合わせを選択した．

反応間の類似スコアは，もととなる反応において，基質となる化合物や生成物となる化合物が複数存在する場合，どの基質と生成物の組み合わせを選ぶかによってスコアが異なる．相関を求める場合には，経路が特定されていないため，どういった組み合わせになるかは定まっていない．そのため，ここでは類似スコアが最も高くなる組み合わせを選択した．

図 4 に 2 種類の類似スコアの相関を取った結果を示す．提案スコアを横軸にとり，縦軸には，同じ反応の組み合わせの場合の EC 番号に基づく類似スコアをとった．図より，EC 番号に基づくスコアでは低いが提案スコアでは高くなることが，EC 番号に基づくスコアでは高いが提案スコアは低くなることよりも，多く見られる．したがって，EC 番号に基づくスコアが低くても，提案スコアが高く

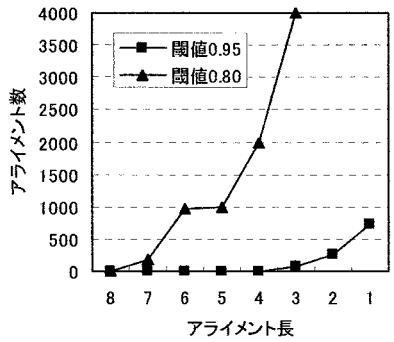

図 6: アライメント長の出力アライメント数

なるケースは多くあり，これは，提案スコアがEC番号に基づくスコアよりも細かい対応付けができていると考えられる．

しかし，EC番号に基づくスコアでは低いが提案スコアでは高くなった反応の組み合わせも存在する．その反応の組み合わせの1つを図5に示す．これは図4中で，EC番号に基づくスコアが5.900となり，提案スコアが0.167となった反応の組み合わせである．

図5で化合物の構造で変化がおきる場所を丸で囲んでいる．ここで最もEC番号に基づくスコアが高くなるEC番号の組み合わせは[2.2.1.6]と[2.2.1.6]であり，EC番号は一致する．しかし，それぞれの反応を構成する化合物の構造式はまったく異なる．このような反応は本論文で提案したスコアでは，類似度が低くなるため，スコアの改良が必要なことがわかる．

3.4 アライメントスコア補正の問題点

最後に，アライメントのスコアの補正の影響を確認する．閾値により求められるアライメントの数が変化する．図6に，閾値を0.95としたときと，0.80にしたときのアライメントの数の変化をアライメント長別に示した．この図より，閾値の値を小さくすると，得られる最大のアライメント長が長くなり，得られるアライメントの数が増加することが分かる．したがって，アライメント長を考慮に入れたスコアの提案が必要であるといえる．

4 おわりに

化合物の構造式に基づくアライメントアルゴリズムを提案した．そして，実際に大腸菌の代謝反応データに適用した結果，最もスコアが高くなったのは，フルクトースとマンノースの生合成経路と，ガラクトースの生合成経路のアライメントとなった．

今後の課題としては，アライメントのスコアの補正方法や，反応間の類似スコアの改良，経路抽出アルゴリズムの検討などがあげられる．

謝辞

本研究の一部は，文部科学省ハイテク・リサーチ・センター整備事業および，2006年度科学研究補助金（若手研究（B）課題番号17700297）による．

参考文献

[1] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M.: The KEGG resource for deciphering the genome, *Nucleic Acids Research*, Vol. 32, pp. D277–280 (2004).

[2] Caspi, R., Foerster, H., Fulcher, C. A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S. Y., Tissier, C., Zhang, P. and Karp, P. D.,: MetaCyc: a multiorganism database of metabolic pathways and enzymes, *Nucleic Acids Research*, Vol. 34, pp. D511–516 (2006).

[3] Dandekar, T., Schuster, S., Snel, B., Huynen, M. and Bork, P.: Pathway Alignment: Appliation to the Comparative Analysis of Glycolytic Enzymes, *Biochemical J.*, Vol. 343, No. 1, pp. 115–124 (1999).

[4] Schmidt, S., Sunyaev, S., Bork, P. and Dandekar, T.:Metabolites: a helping hand for pathway evolution?, *Trends in Biochemical Sciences*, Vol. 28, No. 6, pp. 336–341 (2003).

[5] Webb, E. C., editor, *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classication of Enzymes*, Academic Press, (1993).

[6] 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).

[7] Tohsato, Y., Matsuda, H. and Hashimoto, A.: An application of a pathways alignment method to the analysis of metabolic pathways, *Research Communications in Biochemistry, Cell and Molecular Biology*, Vol. 5, pp. 179–191 (2003).

[8] Xue, L., Godden, J. W., Stahura, F. L. and Bajorath, J.: Profile Scaling Increases the Similarity Search Performance of Molecular Fingerprints Containing Numerical Descriptors and Structural Keys, *Journal of Chemical Information and Computer Sciences*, Vol. 43, No. 4, pp. 1218–1225 (2003).

[9] MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.

[10] Smith, T. F. and Waterman, M. S.: Identification of Common Molecular Subsequences, *Journal of Molecular Biology*, Vol. 147, pp. 195–197 (1981).

[11] MacCuish, N.E. and MacCuish, J.D.: Clustering Compound Data: Asymmetric Clustering of Chemical Datasets, Chemometrics and Cheminfomatics, ACS Symposium Series, Vol. 894, ed. B. K. Lavine, Oxford University Press, (2005) `http://www.mesaac.com/`.

[12] Rahman, S.A., Advani, P., Schunk, R., Schrader R. and Schomburg, D.: Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC), *Bioinformatics*, Vol. 21, No. 7, pp. 1189–1193 (2005).

[13] Arita, M.: The metabolic world of Escherichia coli is not small, *Proceedings of the National Academy of Sciences USA*, Vol. 101, No. 6, pp. 1543-1547 (2004).

[14] Zhang, Y, Li, S., Skogerbo, G., Zhang Z., Zhu X., Zhang Z., Sun S., Lu H., Shi B. and Chen R.: Phylophenetic properties of metabolic pathway topologies as revealed by global analysis, *BMC Bioinformatics*, Vol. 7, pp. 252-264 (2006).