

反応分類番号によるパスウェイアライメントの提案

新 免 剛 瀬 尾 茂 人
竹 中 要 一 松 田 秀 雄
大阪大学大学院情報科学研究科

代謝ネットワーク中でどの部分の反応経路間で類似が見られるかは、生物が進化の過程でどのように代謝ネットワークを構築してきたかを明らかにする上で重要な情報である。先行研究として、代謝ネットワーク中における反応の類似性を、反応を触媒する酵素に付けられたEC番号を利用して定義し、類似反応列を抽出することが行われてきた。しかしEC番号が全く異なる反応間においても、反応前後の化合物の構造変化が類似している反応の組も多く存在する。そこで本研究では反応間の化合物の構造変化によって定義された反応識別番号を用いて、ネットワーク中の類似性を抽出する手法を提案する。

Pathway Alignment using Reaction Classification Number

GO SHINMEN, SHIGETO SENO, YOICHI TAKENAKA
and HIDEO MATSUDA

Graduate School of Information Science and Technology, Osaka University

Finding a common pattern among metabolic network gives important information on the evolution of the pathway. In previous works, it has been done to define the similarity of the reactions in the metabolic network by using the EC number. There are many reactions whose conformational changes of the compounds are similar, though their EC numbers are quite different. In this research, we propose a method for extracting the similarity in the network by using Reaction Classification Number.

1. はじめに

生物の複雑な生命活動を支える生体代謝ネットワークの1つに代謝反応ネットワークがある。生体内では、酵素を触媒として複数の化合物（基質）を別の複数の化合物（生成物）に合成・分解する多くの化学反応が行われている。このような化学反応を代謝反応とよぶ。個々の代謝反応は、単独の反応ではなく、ある代謝反応の生成物が別の代謝反応の基質となっている。こうしていくつもの代謝反応が連絡しあって生物全体の代謝反応網を形成しており、これを代謝反応パスウェイ（以下、パスウェイと略す）とよぶ。図1に、大腸菌の解糖に関わるパスウェイの一部を示した。現在、パスウェイに関する様々な情報が、生化学的な実験や遺伝学的実験によって得られている。これらパスウェイに関する情報は、KEGG¹⁾や、BioCyc²⁾、WIT³⁾などのデータベースに集められ、WWW上で公開されている。

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、以下のようなものがある。

- パスウェイの再構築^{4)~6)}

基質と生成物の二項関係を利用し、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反応経路を求める。

- パスウェイクラスタリング⁷⁾

パスウェイ中の反応を触媒する酵素間の配列類似性をもとに、反応間の距離を計算しパスウェイを分類する（配列類似性とは、それぞれの酵素として働くタンパク質間の文字列としての類似性をさす）。

- パスウェイに基づくゲノム比較⁸⁾

1つのパスウェイを構成する各酵素と種々のゲノム上の遺伝子を対応づけることにより、パスウェイの酵素の順序と遺伝子の順序を比較し、互いに隣接した関係を求める。

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一つとして、パスウェイ中から類似した反応を抽出する、パスウェイアライメントを考える。パスウェイは約12000種類の化合物と約3500種類の酵素からなる反応のネットワークである。しかし、生物進化の初期段階からパスウェイが現在のようにこれほど複雑な構造をとっていたとは考えにくい。生物が

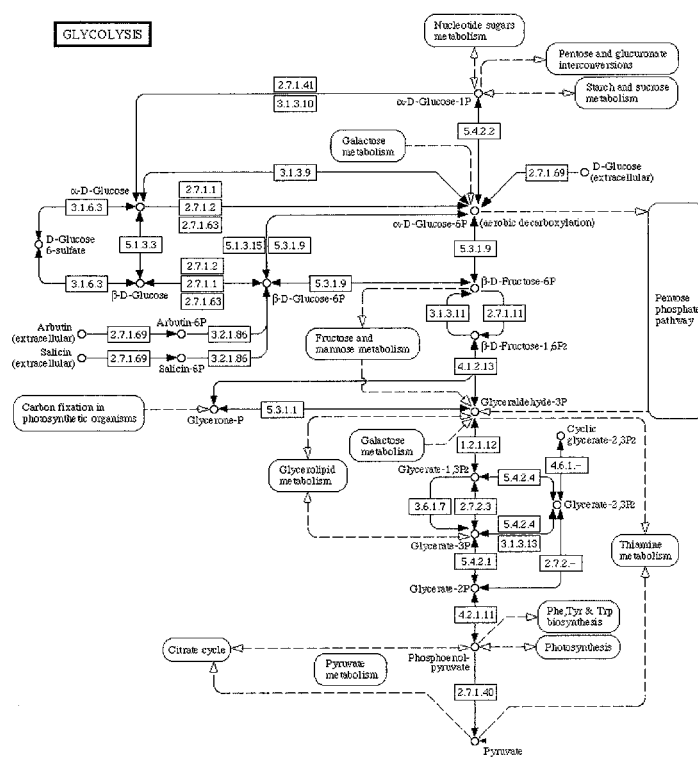


図1 解糖に関わるパスウェイの一部

進化の過程において、ネットワークの一部を複製したり変化させながら、現在の複雑なパスウェイを獲得してきたものと考えられる⁸⁾。このため、パスウェイ中に存在する類似した反応経路を見つけることは、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上で重要な情報をもたらすと考えられている⁸⁾。

パスウェイ中で反応の類似性を考えるとき、反応や反応経路を計算機上で比較するこれまでの試みでは、パスウェイ中の反応に酵素として働くタンパク質の配列類似性をもとに反応間の類似度を定義し、パスウェイを分類することが多い⁷⁾。しかし、同じ反応を触媒する酵素が複数存在するとき、それらの酵素として働くタンパク質の配列間にほとんど類似性が見られない場合 (enzyme recruitment⁹⁾) があることが知られている。そのため、反応の類似性を、酵素の機能すなわち EC 番号を用いて定義しアライメントを行う研究も行われている¹⁰⁾。

EC 番号による反応分類は、反応を触媒する酵素の機能に着目して定義、分類されたものであるため、基質・生成物が異なる反応に対し、同じ EC 番号が割り当てられることもある。また、反応自体は知られているが EC 番号が割り当てられていない反応も存在する。そこで本研究では、反応前後の化合物の構造に対して定義された反応分類 (RC) 番号¹¹⁾ を、反応の類似性として利用したパスウェイアライメント手法を提案する。以下では、EC 番号と RC 番号による類似性指標の差、RC 番号による反応類似性スコアを用いたパスウェイアライメント手法、およびこの手法を適用した結果について示す。

2. パスウェイの反応類似性とアライメント

パスウェイにおける反応の類似性とは、ある反応によっておこる化合物構造の変化が類似していることを指す。図2は図1の中のパスウェイの一部を取り出したものであり、互いに反応が類似したパスウェイの例となっている。図では酵素をその名称 (括弧内は大腸

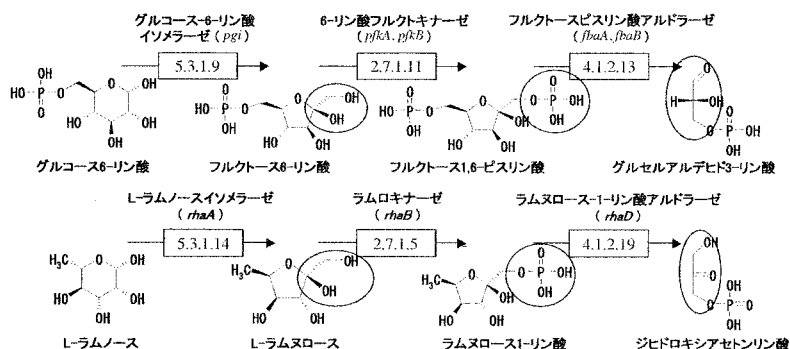


図 2 パスウェイにおける反応の類似部分の構造比較

菌での遺伝子名)と EC 番号 (四角の中) で表し、酵素の基質・生成物を化学構造型と化合物名で表している。化学構造型中で円で囲まれた部分は酵素により変化する部分を表す。円で囲まれた部分に着目すると、これらの 2 つのパスウェイの反応が類似していることが分かる。

2.1 EC 番号によるアライメント

2.1.1 EC 番号

酵素は International Union of Biochemistry and Molecular Biology (IUBMB) により、それが触媒する代謝反応の種類に基づいて EC (Enzyme Commission) 番号と呼ばれる 4 組の数字により階層的に機能分類されている。この EC 番号の階層的な酵素の分類を機能階層と言い、EC 番号の左から順に第 1 階層, ..., 第 4 階層と言う。例えば第 1 階層において、酵素は大きく次の 6 つの機能に分類されている。

- 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)

2.1.2 EC 番号による類似スコア

先行研究¹⁰⁾ においては、反応間の類似性として EC 番号の類似性を用いている。図 2 に示した 2 つの反応経路を見ると、対応する反応間の EC 番号は 3 階層まで一致している。この点に着目し、先行研究では反応間の類似性スコアを、EC 番号における階層構造の類似性を利用して定義することにより、反応経路間の類似部分を抽出している。

2.2 問題点

EC 番号を類似度の指標としている上記の手法では、EC 番号の類似した反応列を抽出することができた。

しかし、EC 番号では反応の類似を表現できない場合もある。

図 3 で示した 2 つの反応にはそれぞれ EC1.7.1.6 と EC1.6.6.- が割り当てられている。図 3 の丸で囲んだ部位は、両反応で共通して見られる反応で変化する構造であるが、2 つの反応の EC 番号は 2 階層目から異なっている。両反応で共通である 1 階層目の "1.-" は、反応が酸化還元反応であることを表すが、2 階層以下の番号は異なっており、両反応の EC 番号の類似度は低いと言える。これは、EC 番号が反応全体の機能的な特徴を表現するものであり、化合物変化の特徴を詳細には表現しきれていないためと思われる。

3. 反応分類番号を用いたパスウェイアライメント

2 章に示したように、EC 番号では反応中の化合物の構造変化についての比較が適切に行えないことがある。パスウェイデータベース KEGG¹⁾ では、パスウェイ中の各反応について、反応前後の化合物構造に注目することで反応を分類し、識別番号 (RC 番号) を付けることが行われている¹¹⁾。本章では、反応の類似度の指標としてこの RC 番号を用いることによるパスウェイアライメント手法を提案する。

3.1 RC 番号

RC 番号の定義ではまず、反応の前後に現れる化合物について、化合物中の原子をラベル付けする。その後、化学構造比較アルゴリズム¹²⁾ を適用し、2 つの化合物の構造アライメントを行い、共通保存領域と反応中心部位を特定する。図 4 は 2 つの化合物の原子をラベル付けし、共通保存領域と反応中心部位を特定した例である。

ラベル付けされた原子のうち、反応中心部位である

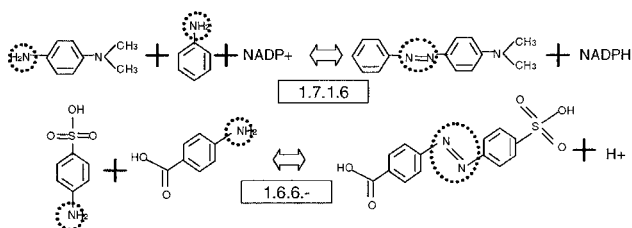


図 3 EC 番号の異なる反応の比較

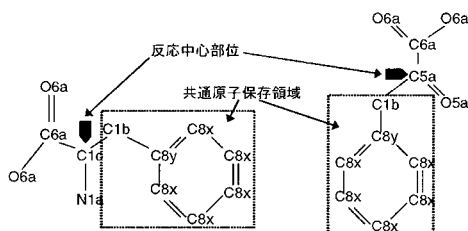


図 4 2つの化合物の構造比較

原子 (R-atom), 共通保存領域のうち反応中心部位に隣接する原子 (M-atom), 保存されない領域のうち反応中心部位に隣接する原子 (D-atom) の, 3つの原子の組み合わせについて, それぞれ識別番号を与え, 3つの識別番号を並べたものが RC 番号である. 図 5 はある反応の, 前後の化合物の組み合わせについて RC 番号を付加した様子を示している.

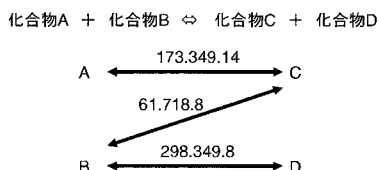


図 5 反応の前後の化合物ペアに, RC 番号を定義する

図 3 で示した, EC 番号の類似度の弱い 2つの反応について, 図 6 左で示した構造比較が図 3 の上側, EC1.7.1.6 の反応, 図 6 右で示した構造比較が図 3 の下側, EC1.6.6.- の反応である. A 1 と B 1, A 2 と B 2 を見ると, R-atom, D-atom, M-atom がそれぞれ一致している.

RC 番号による分類は, EC 番号と同等に酵素の特徴を分類可能で, さらに詳細に反応を分類するものであると報告されている¹¹⁾. このことから, 反応の類似度として RC 番号を用いることで, EC 番号よりも詳細に反応の類似を抽出できると考えられる.

3.2 RC 番号による類似スコア

本手法では, 反応の類似度スコアを, 反応中における化合物変化につけられた RC 番号の一致度を利用して定義する.

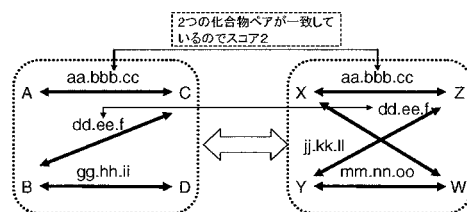


図 7 RC 番号の一致度

2 反応間の類似スコアは, 両反応間に共通して存在する化合物変化のペアの数とする. 反応 A と反応 B 間の類似スコアを $RCs(AB)$ とすると, $RCs(AB) = (A \text{ と } B \text{ に共通して存在する化合物変化の RC 番号の数})$ である. 図 7 の反応間では, 2つの反応間で一致する化合物変化のペアが 2つ存在するためスコアは 2 となる. 両反応中に一致する化合物変化のペアが存在しない場合, スコアは 0 となる.

3.3 反応分類番号によるアライメント

本手法では, 2つの反応列の類似反応パターン抽出に動的計画法に基づくアライメントアルゴリズムを利用する. 動的計画法によるアライメントは, アライメントの対象となる系列の各要素 (反応) を 2次元に配置し, それぞれの要素間に頂点をおき, これらの辺を通して類似スコアが最大となる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 8 で表される経路において, 斜めの矢印を通るときそれに対応する反応間の類似スコア, 右または下の矢印を通るときギャップのスコアをそれぞれ加えていったものとする. 最も高いスコアを取るアライメントを最適なアライメントとする. ギャップスコアは, 化合物ペアが一致しない場合のスコアが 0 であることより, 本研究では -1 とした.

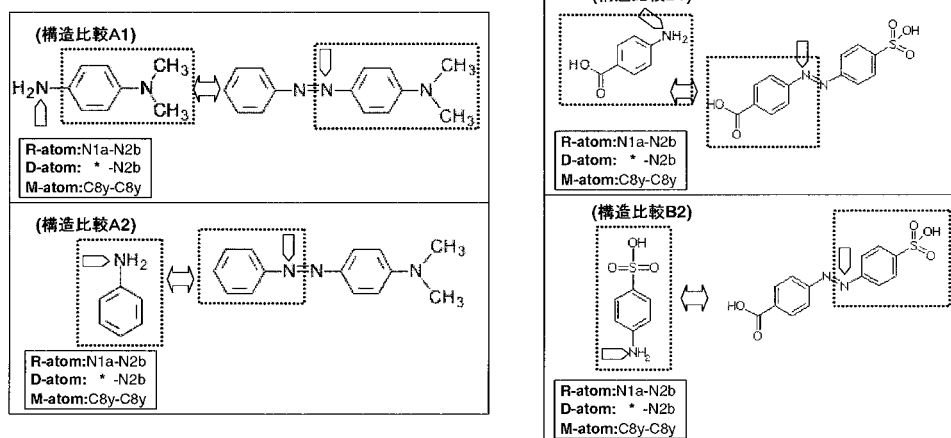


図 6 反応前後の構造比較

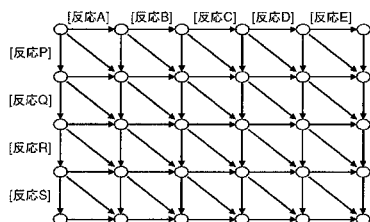


図 8 2つのパスウェイのアライメントの経路

4. 実 験

本手法の有効性を確かめるため、反応経路をKEGGの代謝マップから抽出し、RCスコアによる類似反応列の抽出を行った。

入力データ

実験に用いた反応経路データの抽出法について述べる。KEGGの代謝マップに載っている化合物のうち、その化合物からつながる化合物の数が1つしかない化合物(図9(1))、他の化合物へつながる経路の数が4以上の化合物(図9(2))を特定し、これらの化合物間の全ての経路(閉路は除く)を入力データとして抽出した。(2)の化合物については、経路の分岐点となるという意味で、当初は他の化合物への経路の数が3以上の化合物を対象として抽出を行ったが、抽出される反応経路の多くが反応数1の経路となったため、条件を緩和し4とした。以上のような抽出を行った結果、117の代謝マップから1467の反応経路を得た。

結果

代謝マップから抽出した全ての反応経路の組み合わせ

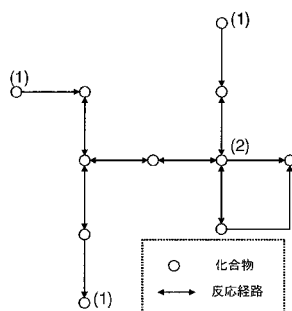


図 9 入力データの抽出法

についてRCスコアに基づくアライメントを行った。表4で示したのは各反応経路の組み合わせの、先行研究におけるEC番号を用いたアライメントのスコア(横)と、本研究で定義したRC番号の一致度を用いたアライメントのスコア(縦)の分布である。各セル内の数値は縦軸と横軸で示されたRCとEC各スコアに該当するアライメントの個数である。

アライメントの結果の中から、反応間でEC番号が類似していないがRC番号の一致があった反応を含む経路の一部分を図10に示す。この結果は、ヒスチジン分解パスウェイから経路長5、リジン分解パスウェイから経路長9の反応経路を入力データとして与えたものであり、RC番号の一致度によるスコアは5、EC番号の類似度によるスコアは-3.38であった。

図10の上下の矢印で示した反応は、RC番号で類似していた反応である。(A)ではEC番号は1桁目から異なるが、上下の反応はそれぞれ化合物の共通した

表 1 反応経路間のアライメントの EC 番号 (横) スコアと RC 番号 (縦) スコアの分布

14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	3	0	8		
8	0	0	0	0	0	0	2	2	1	4	2	5	0	8			
7	0	0	0	0	0	0	0	0	0	6	20	17	7	6			
6	0	0	0	0	0	0	2	12	9	52	141	57	70	71			
5	2	0	0	0	0	1	0	18	79	71	247	102	90	81			
4	60	0	0	21	15	15	232	294	381	824	285	185	176				
3	145	0	6	364	271	369	1159	1227	2002	2747	1001	510	797				
2	1918	0	247	4355	2775	2438	7349	6692	8824	7474	2574	1422	3314				
1	20305	0	1863	24483	23366	8783	28310	24617	26724	18272	5069	1611	20855				
0	185910	0	3983	67272	59663	16842	28885	25591	22843	15595	4354	1169	102637				
Score	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0				

14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	2	1	0
12	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0
11	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
10	0	0	0	0	2	4	0	1	4	5	4	0	0	0	1	0	0	0	0
9	1	6	1	2	6	4	4	7	8	7	3	2	0	0	0	0	0	0	0
8	7	13	3	2	13	7	1	12	14	7	4	0	0	24	0	0	0	0	0
7	9	34	20	4	16	13	5	14	19	1	1	2	2	0	0	0	0	0	0
6	27	77	31	17	35	33	5	7	5	0	0	0	0	0	0	0	0	0	0
5	60	162	80	33	32	37	8	6	9	0	0	0	0	0	0	0	0	0	0
4	92	324	73	37	52	27	32	18	1	0	0	0	0	0	0	0	0	0	0
3	190	835	175	67	43	20	36	14	2	0	0	0	0	0	0	0	0	0	0
2	413	1396	187	10	3	4	0	0	0	0	0	0	0	0	0	0	0	0	0
1	362	1454	169	1	7	8	2	1	0	0	0	0	0	0	0	0	0	0	0
0	234	771	57	8	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Score	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

構造部分のアルデヒド基に作用する反応である。(B)の反応はアミノ基をアルデヒド基に変化させる反応であり、(C)の反応はアルデヒド基を酸化しカルボキシル基に置き換える反応である。(C)では EC 番号は 3 桁目まで一致しており、EC 番号の類似度も高いといえるが、(B)では EC 番号は 1 桁目から異なっており、類似度は低い。

考察

図 11 は図 10 中の (B) の両反応の詳細である。EC 番号 EC2.6.1 はアミノ基を転移する反応を定義するものであり、EC 番号 EC1.4 は部分構造 $CH-NH_2$ に対し酸化還元を起こす反応を定義している。EC 番号は反応全体の機能的な特徴をもとに割り当てられるため、化合物の反応経路としてみた場合の、両反応は共にアミノ基の部分アルデヒド基に変化させる反応であるという類似点を表現できていないことがわかる。このように、本手法により EC 番号は異なるが化合物

の構造変化が互いに類似した反応経路を抽出することができた。

5. ま と め

化合物の構造比較に基づいて定義された反応分類番号を用いてパスウェイのアライメントを行う手法を提案し、実際にパスウェイデータに適用することで、EC 番号によるスコアでは抽出できないアライメントを見つげられることを確認した。

今後の改善点としては、反応分類番号の類似スコアの定義において、RC 番号が完全に一致した場合のみスコアを加算すると定めているが、RC 番号の類似性を考慮することで、例えば反応中心原子のみが一致した場合にも一定のスコアを加算するなどの工夫をすることで、より詳細に反応の類似を抽出できるようになる可能性が考えられる。

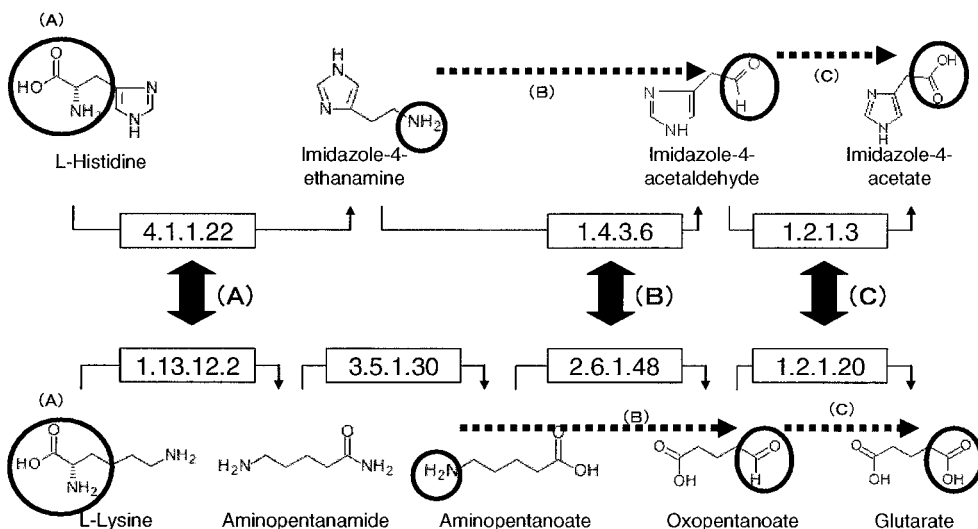


図 10 RC 番号の一致度でアライメントした結果の例：上列はヒスチジン分解経路の一部，下
列はリジン分解経路の一部

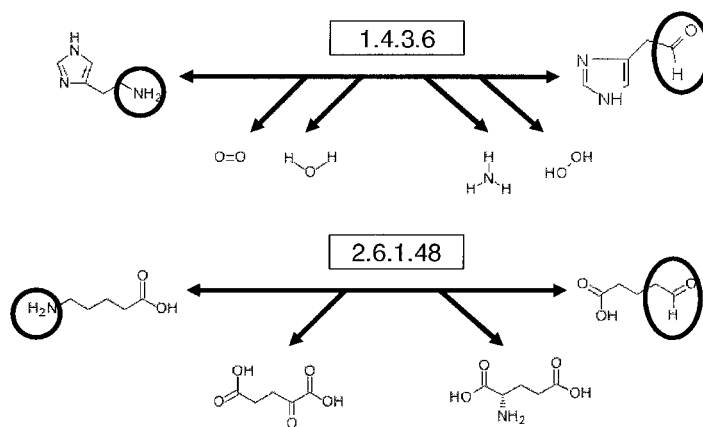


図 11 EC 番号と構造変化

参考文献

- 1) Kanehisa, M., Goto, S., Kawashima, S. and Hattori, M. "The KEGG resource for deciphering the genome," *Nucleic Acids Research*, Vol.32, pp.D277-D280(2004).
- 2) Keseler, I.M., Collado-Vides, J.C., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M., and Karp, P.D. "EcoCyc: a comprehensive database resource for *Escherichia coli*," *Nucleic Acids Research*, Vol.33, pp.D334-D337(2005).
- 3) Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Selkov, E. Jr., Kyrpides, N., Fontstein, M., Maltsev, N., Selkov, E. "WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction," *Nucleic Acids Research*, Vol. 28, pp. 123-125 (2000).
- 4) Mavrouniotis, M.L. "AI and Molecular Biology," *Artificial Intelligence and Molecular Biology*, AAAI Press/MIT Press, Menlo Park, pp.

- 325–364(1993).
- 5) Paley, S.M., Karp, P.D. "Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*," *Bioinformatics*, Vol. 18, No 5, pp.715–724 (2002).
 - 6) Gaasterland, T., Selkov, E. "Reconstruction of metabolic networks using incomplete information," *Proceeding of the International Conference on Intelligent Systems for Molecular Biology*, pp.127–135 (1995).
 - 7) Forst, C.V. and Schulten, K. "Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information," *Journal of Computational Biology*, Vol.6, pp.343–360(1999)
 - 8) Dandekar, T., Schuster, S., Snel, B., Huynen, M., Bork, P. "A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy," *Biochemical Society*, Vol.343, pp.115–124(1999)
 - 9) 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)
 - 10) Tohsato, Y., Matsuda, H., Hashimoto, A. "A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy," *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, pp.376–383(2000)
 - 11) Kotera, M., Okuno, Y., Hattori, M., Goto, S., Kanehisa, M. "Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions," *Bioinformatics Center, Institute for Chemical Research*, Vol.125, pp.16487–16498(2004)
 - 12) Hattori, M., Okuno, Y., Goto, S., Kanehisa, M. "Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways," *Bioinformatics Center, Institute for Chemical Research*, Vol.126, pp.11853–11865(2003)