

## 化合物の部分構造変化に着目した酵素反応の類似性とその評価

遠里由佳子<sup>1</sup>, 西村悠<sup>2</sup>

近年、酵素反応の比較・分析は、生物の進化の解明や創薬に役立つテーマの1つとして注目されている。本研究では、酵素反応を構成する「化合物の構造式の変化」に着目し、酵素反応において基質となる構造式と生成物となる構造式の差を RPAIR データを用いて抽出する。そして、抽出された部分構造間の類似度に基づき酵素反応間を比較する手法を提案し、実際の大腸菌における酵素反応間の比較に適用することでその有効性を評価する。

### Evaluation of Reaction Similarities Focusing Substructure Changes of Chemical Compounds

Yukako Tohsato<sup>1</sup> and Yu Nishimura<sup>2</sup>

Comparative analyses of enzymatic reactions provide important information on both evolution and potential pharmacological targets. We focus on the structural formulae of compounds. The present study perform to extract substructures that change within chemical compounds using the RPAIR data in KEGG. Two approaches were applied to measure the similarity between the extracted substructures: a fingerprint-based approach using the MACCS key and the Tanimoto/Jaccard coefficients; and the Topological Fragment Spectra-based approach that does not require any predefined list of substructures. Whether the similarity measures can detect similarity between enzymatic reactions was evaluated.

#### 1. はじめに

酵素反応の比較・分析は、生物の進化の解明や創薬に役立つテーマの1つとして注目されている。酵素反応の類似性は、一般に反応を触媒する酵素をコードする遺伝子の類似性でとらえることが多い。しかし、配列が異なっても機能に類似が見られる例などの報告もあり、異なった観点からの解析が必要とされている。そこで注目されたものに、Enzyme Commission (EC) 番号 [1] と呼ばれる酵素の機能分類がある。EC 番号は、酵素反応を、基質・生成物の化学構造や補酵素などの化学反応のタイプにより分類する。EC 番号は有用ではあるが、しかし同時に (1) 手作業で番号付が行われていること (2) 酵素が存在すると確認した論文がなければ番号づけをしないという問題がある。

そこで本研究では、酵素反応を構成する「化合物の構造変化」によって酵素の類似性を定義することで、それらを比較する手法を提案する。酵素反応の類似を比較する問題を、単純に反応を構成する化合物の構造式の類似に置き換えた場合、さまざまな問題が生じる [2]。そのため、1つの酵素反応における基質となる構造式と生成物となる構造式の差と

なる部分構造を切り出すことで問題を回避し、比較システムの構築を試みた。

#### 2. 酵素反応の類似性比較

##### 2.1. 部分構造の抽出法

RPAIR は、酵素の反応前後の化合物間で原子アライメントすることにより取り出された、基質と生成物間の変換パターンを記録したデータベースである。これより、反応前後において変化しない部分、付加または脱離によって変化する部分、前記の二つの境界領域にあって反応の中心と考えられる部分の3種類の情報を捉えることができる [3, 4]。そして、各酵素反応は KEGG [5] の代謝反応データに対応づけられている。そのため、RPAIR を用いることで、酵素反応を構成する化合物の構造式から、変化の起こる部分構造を抽出することができる。

本研究では、これら化合物の構造式の部分構造を単に「部分構造」と呼ぶ。例えば、図1の化合物の組 (C00025, C00624) と (C16396, C16420) は KEGG において酵素反応 R00259 と R08036 の主要代謝経路であり、このとき丸で囲まれた原子グループが、部分構造に該当する。

KEGG は RPAIR に基づく酵素反応の階層的な分類 (RC) システムを持っている。しかしながら、RC システムは抽出された部分構造の類似比較を考慮していない。そのため次節で述べる酵素反応間の類似性尺度を提案した。

<sup>1</sup>立命館大学 生命情報学科

Department of Bioscience and Bioinformatics, College of Life Sciences, Ritsumeikan University

<sup>2</sup>立命館大学 理工学研究所 情報理工学専攻

Information Science & Systems Engineering, Graduate school of Science and Engineering, Ritsumeikan University

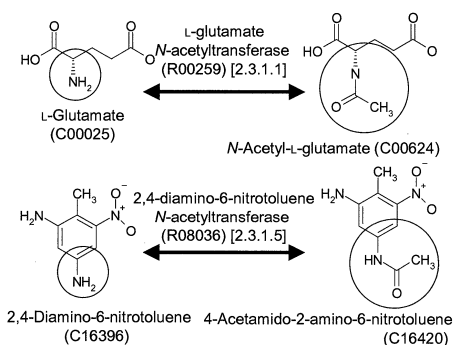


図1 酵素反応の比較の例

## 2.2. 化合物の構造に基づく酵素反応の比較

部分構造間の類似度に基づき酵素反応間の類似度を計算するため、Fingerprint に基づく手法 [6] と Topological fragment spectra (TFS) に基づく手法 [7] を提案する。

### 2.2.1. Fingerprint 法

化合物の構造式をグラフとして扱い、構造式間の類似度を求める問題は、最大共通部分グラフ問題の一種となるため、計算機では困難となる [4]。そこで、類似した構造式の探索には Fingerprint と呼ばれる予め定義された構造式のフラグメントを用いることが多い。MACCS key [6] は、MDL により提案された Fingerprint であり、これにより構造式は 166 種類のフラグメントの有無を 0 もしくは 1 のビットで表したビット列となる。

そこで、部分構造を MACCS key によりビット列で表現し、もっとも一般的な類似性の尺度である Tanimoto (Jaccard) 係数を用いて、化合物  $c_1$  と  $c_2$  の部分構造間の類似度を以下のように定義する。

$$F(c_1, c_2) = c / (a + b - c) \quad (1)$$

$a$  と  $b$  は、部分構造  $c_1$  と  $c_2$  のビット列中の 1 のビットの数、そして、 $c$  は部分構造  $c_1$  と  $c_2$  に共通して 1 となるビットの数を表している。定義より、 $F(c_1, c_2)$  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なる。

次に、反応  $r_1 = (c_{1b}, c_{1d})$  と反応  $r_2 = (c_{2b}, c_{2d})$  間の類似性の尺度  $PF(r_1, r_2)$  を次の式で定義する。

$$PF(r_1, r_2) = (F(c_{1b}, c_{2b}) + F(c_{1d}, c_{2d})) / 2 \quad (2)$$

ただし、1 つの化合物に複数の変化する部分構造が存在する場合や、主要代謝経路が 1 つの酵素反応に複数ある場合は、最も高いスコアとなる組み合わせ

を 1 つ選択する。このとき、図 1 の 2 つの酵素反応の類似度は  $PF = 1$  となる。

### 2.2.2. TFS 法

TFS は化学構造の定量的な類似性評価のために、高橋らによって考案された構造式の記述法の 1 つである [7]。その生成手順は (1) 対象とする構造式から可能なフラグメントをすべて列挙し (2) 列挙したそれぞれのフラグメントに対して数値的な特徴付けを行い、特徴付けの値とその出現頻度のベクトルデータを生成する (図 2)。図 2 は、2-methylbutane からベクトル  $c = (3, 1, 2, 2, 2, 3, 3)$  が作られる過程を示している。

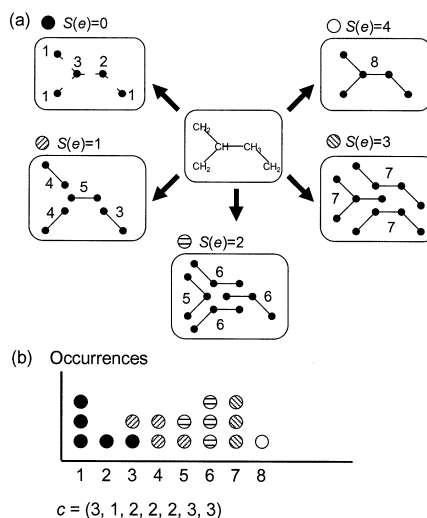


図 2 TFS の手続きイメージ。  $S(e)$  は列挙するフラグメントのサイズ (ノードの数) をさす。

しかし、すべてのフラグメントを列挙するには多くの計算時間を必要とする。そこで、ベクトルの生成に際して、結合 (ノード) の数が 5 までのフラグメントに限定する [7]。また、原子の種類を無視し、骨格構造のみに着目したもっとも単純な手法を採用する。

部分構造  $c_1$  のベクトル  $(c_{1b}, c_{1d}, \dots, c_{1m})$  と部分構造  $c_2$  のベクトル  $(c_{2b}, c_{2d}, \dots, c_{2m})$  間の非類似度をユークリッド距離に基づき次の式で計算する。

$$T(c_1, c_2) = \sqrt{\sum (c_{1k} - c_{2k})^2} \quad (3)$$

ユークリッド距離では値が大きいかほど 2 つ部分構造間の類似度が低い。

反応  $r_1 = (c_{1b}, c_{1d})$  と反応  $r_2 = (c_{2b}, c_{2d})$  間の類似性の尺度  $PT(r_1, r_2)$  を次の式で定義する。

$$PT(r_1, r_2) = (T(c_{11}, c_{21}) + T(c_{12}, c_{22})) / 2 \quad (4)$$

図1の2つの酵素反応間では $PT=0$ となる。

### 3. 実験と結果

KEGG (Ver. 46.0) の大腸菌 K-12 MG1655 の代謝データから, RPAIR に基づき 19,805 の部分構造を MOL ファイル形式で取り出した。そして, 取り出したデータを Open Babel (<http://openbabel.org/>) と MESA の Fingerprint Module (<http://www.mesaac.com/>) を用いてビット列データに変換した。EC 番号が完全に割り当てられていない (例えば [1. -. -.]) 反応のデータは実験対象から除いた。そして, 4,567,734 の反応の組に対し, PF 法および PT 法, WF 法の3種類の反応の類似度を計算した。

WF 法は文献 [8] の酵素反応間の類似度の計算法であり, 提案法と比較するため導入する。WF 法では, 式 (1) の2つの部分構造を2つの化合物の構造式に置き換えたものとなる。そして, 2つの化合物  $c_1$  と  $c_2$  の構造式のビット列間の類似度  $F'(c_1, c_2)$  を, Tanimoto 係数に基づき計算する。酵素反応  $r_1=(c_{11}, c_{12})$  と  $r_2=(c_{21}, c_{22})$  間の類似度は, 以下の式で定義する (詳細は文献 [8] 参照)。

$$WF(r_1, r_2) = (F'(c_{11}, c_{21}) + F'(c_{12}, c_{22})) / 2 \quad (5)$$

解析の基準として EC 番号を採用した。酵素に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 番号が割り当てられ, 階層的に分類されている。以下, これを単純に階層と呼ぶ。2つの酵素反応間で, 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の位置を「レベル」と呼ぶ。例えば [1. 1. 1. 3] と [1. 1. 2. 4] 間で共通する最も上位の階層は [1. 1] であり, そのレベルは2となる。

図3に PF 法および PT 法, WF 法の手法別スコア分布を示した。X 軸にスコアを Y 軸にそのスコアとなった酵素反応の組の数をプロットしている。次に図3の各分布で上位5%に該当する組を選び, EC 番号の階層のレベル別にまとめた (表1)。EC 番号の

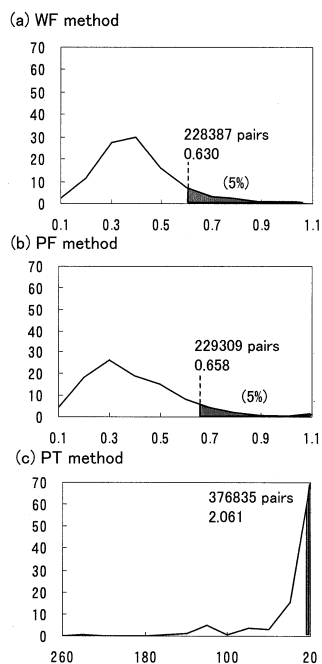


図3 反応間の類似スコアの分布

階層の第1レベルは酵素反応における十分な情報を含んでいない。しかし, 第2, 第3レベルは, より詳細な情報を含んでいる。そこで, 第3レベルまで EC 番号が一致した酵素反応の組み合わせを正解とみなし, 表1において, 第3レベルまで一致する EC 番号を持つ反応の組み合わせが, 上位5%に含まれる割合に着目した。このとき, PF 法が最も高い割合 (44.0%) となった。そこで, PF 法で上位5%の酵素反応の組 (表2) と, それらの中で EC 番号の一致が見られない組み合わせを構成する5つの酵素反応の詳細を確認した (図4)。反応 R00010 と R00803 はその EC 番号が [3. 2. 1. 28] と [2. 4. 1. 7] であり, EC 番号のレベルは一致しない。しかし, その酵素反応のメカニズムは図に示すように大変類似し

表1 一致する EC 番号のレベルごとの分布。EC 番号のレベルは一致する階層の深さをさす。図3の分布から上位5%に一ずる2つの反応中で, 各階層別に該当する反応の組の数と割合を求めた。

EC level	#of pairs	Hit number of reaction pairs (%)		
		WF	PF	PT
>=0 (all)	4567734	228387 (5.0)	229309 (5.0)	376835 (8.2)
>=1	1270206	73853 (5.8)	128403 (10.1)	205968 (16.2)
>=2	232100	29100 (12.5)	72871 (31.3)	66509 (28.6)
>=3	120184	19245 (16.0)	52896 (44.0)	45108 (37.5)
>=4	3681	2178 (59.1)	2349 (63.8)	2370 (64.3)

ていることが分かった。これらの結果は PF 法が EC 番号に依存することなく酵素反応の類似を見つける可能性を示唆した。

表 2 上位 5 つの反応の組み合わせ

Rank	Pair	Score	EC Pair
1	R00006 R01841	1	[2.2.1.6] [4.1.2.30]
2	R00006 R03403	1	[2.2.1.6] [4.1.2.30]
3	R00010 R00802	1	[3.2.1.28] [3.2.1.48]
4	R00010 R00803	1	[3.2.1.28] [2.4.1.7]
5	R00010 R00837	1	[3.2.1.28] [3.2.1.93]

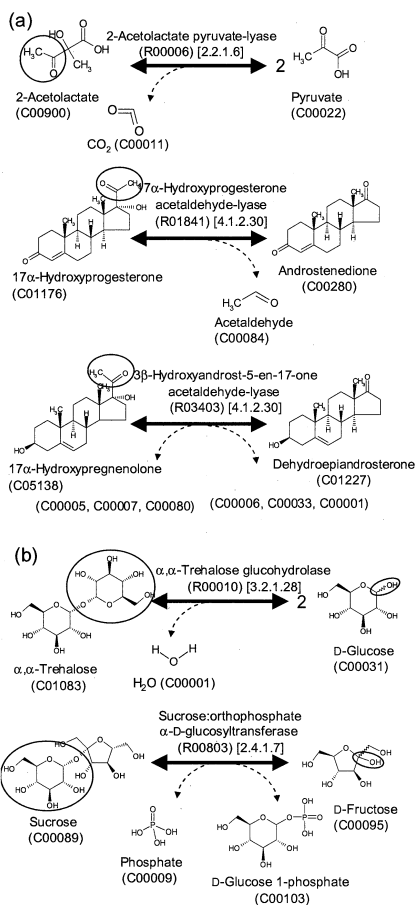


図 4 EC 番号のレベルが異なるが PF 法で高スコアとなった酵素反応の詳細。

#### 4. おわりに

RPAIR を用いることで酵素反応を構成する化合物の構造式から変化する部分構造を抽出した。そして、抽出した部分構造に基づく酵素反応の類似性尺度を提案し、実際の大腸菌の酵素反応データに適用し、

EC 番号と比較することで評価した。その結果、提案手法の中で PF 法を用いた場合がもっとも EC 番号の分類に一致するという結果が得られた。一方で、EC 番号における番号づけのあいまい性の問題から、異なる分類番号を振られた酵素反応間においてもよく似た部分構造変化が存在することを確かめた。このような解析結果を踏まえ、今後、代謝反応パスウェイ中の類似した反応の系列を見つける問題へ応用を試みたい。

#### 参考文献

- [1] Webb, E.C., editor, Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes, Academic Press (1993).
- [2] Tohsato, Y., Matsuda, H. and Hashimoto, A.: An application of a pathways alignment method to the analysis of metabolic pathways, *Res. Comm. In Biochem, Cell & Mol. Biol.*, Vol.5, pp.179-191 (2003).
- [3] Hattori, M., Okuno, Y., Goto, S. and Kanehisa, M.: Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways, *J. Am. Chem. Soc.*, Vol.125, No.39, pp.11853-11865 (2003).
- [4] Kotera, M., Okuno, Y., Hattori, M., Goto, S. and Kanehisa, M.: Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions, *J. Am. Chem. Soc.*, Vol.126, No.50, pp. 16487-16498 (2004).
- [5] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y.: KEGG for linking genomes to life and the environment, *Nucleic Acids Res.*, Vol.36, pp. D480-D484 (2008).
- [6] MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.
- [7] Takahashi, Y., Ohoka, H. and Ishiyama, Y.: Structural similarity analysis based on topological fragment spectra, *Advances in Molecular Similarity*, Vol.2, pp.93-104 (1998).
- [8] Tohsato, Y. and Nishimura, Y.: Metabolic Pathway alignment based on similarity between chemical structures, *IPSJ Transactions on Bioinformatics*, Vol.48, No.SIG17(TBIO3), pp.736-745 (2007).