酵素反応式からの基質・生成物ペアの自動推定

Automatic assumption of reactant pairs from enzyme reaction formula

京都大学化学研究所　小寺正明

背景と目的

様々な生物種の体内で触媒される酵素反応は現在 KEGG データベース **[1]** に登録されているだけでもおよそ 8,000 反応に達するが、その中で実際に活性が確認された酵素タンパク質のアミノ酸配列が報告されているものはおよそ 4,300 に過ぎない。また、メタボローム解析技術の進展により、存在の示唆される酵素反応の数は今後大幅に増加することが予想される。これらの酵素反応を触媒する酵素タンパク質やその遺伝子の同定を効率化するためには、実験手法の改良はもちろん、情報学的なアプローチによる支援も欠かせないと考えられる。

代謝化合物の化学情報とそれを代謝する酵素タンパク質の配列情報とをつなぐ中心的な役割を果たしているものとして、国際生化学分子生物学連合（IUBMB）の酵素委員会（EC）により管理されている酵素番号（EC 番号）がある。しかし多くの反応は、存在が確認されているものの、様々な理由で EC 番号が割り当てられないまま残っている。

我々はこれまで、完全または不完全な酵素反応式が与えられた時に、その化学構造から EC 番号を予測する方法を開発してきた **[2]**。今年度 6 月にはそれをウェブプログラム「E-zyme」として実装した論文を発表し **[3]**、スウェーデンのストックホルムで開催された国際学会 ISMB/ECCB 2009 にて口頭発表を行った。この研究では予測性能の検討と改良も行い、既知の反応パターンと EC 番号との関係をただ列記するだけの単純な方法に比べ、予測精度／カバー率ともに大幅に向上することができた。E-zyme はゲノムネットのウェブサービスとして公開されている **[4]**。

E-zyme では、「基質・生成物ペア（reactant pair）」という概念を用いている（**図 1**）。これは「酵素反応式の前後で（水素原子を除く）原子を保存している基質化合物と生成物のペア」という定義で使われており、たとえば **図 1** の右上にある

AB + C = A + BC

という酵素反応式は、化合物 AB から化合物 C へ部分構造 B が転移して化合物 A と化合物 BC を生成する反応を表しており、このときペア AB-A は部分構造 A を、ペア AB-BC は部分構造 B を、ペア C-BC は部分構造 C をそれぞれ保存しているため結線が引かれるが、ペア C-A には引かれない。このことは、酵素反応式中で変化した化学結合を同定するにあたって、化合物 C と化合物 A の化学構造を比較する必要がないことを示している。

現バージョンの E-zyme では、このように酵素反応式を基質・生成物ペアに変換する作業を手作業で行っている。まず、EC 番号が既に与えられている既知の酵素反応を基質・化合物ペアに変換し、各々のペアについて化合物構造比較プログラムである SIMCOMP **[5]** を用いて化合物構造アラインメントを行い、得られた結果をさらに手作業で精査して反応中心を同定し、RPAIR データベースに教師データセットとして蓄えられた **[6]**。ユーザが E-zyme を使用する際もやはり、ユーザ自身が基質・生成物ペアを定義する仕様となっている。

本研究では、このツールの操作性を向上することを目的として、この「基質・生成物ペア」を、酵素反応式から自動的に予測する方法の開発を行った。

方法と結果

まず、KEGG データベースに登録されていた 8,051 反応のうち、左右でバランスの取れていない反応を除去し 6,778 反応を得た。続いて、左右辺どちらか一方でも化合物数が 5 を上回っていればその反応を除去し、6,637 反応を得た。その中で、左右辺どちらか一方

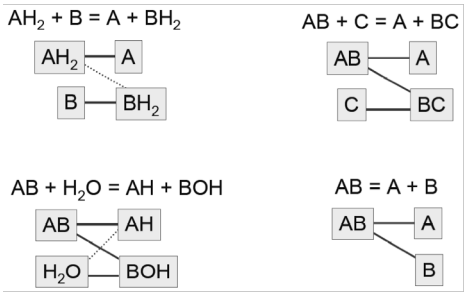

図 1. 基質・生成物ペア

でも化合物数が 1 であれば（**図 1** の右下に相当）、基質・生成物ペアが自明であるため本研究は不要と判断し、除去した。最終的に残った 5,117 反応を本研究に用いた。結果は RPAIR データベースとの一致度で評価した。

本研究との比較としてまず、ランダムに予想した場合の的中率を以下のように計算した。全 5,117 反応中、左右の化合物の可能な組み合わせは 32,349 ペアであった。その中で RPAIR データベースにて定義されているのは 16,087 ペアだった。そこで、酵素反応式が与えられた場合に 16,087 / 32,349 = 0.497 の確率でランダムに基質・生成物ペアを割り当ててみたところ、精度は約 36%であった。

続いて、部分構造フラグメントを用いた方法を以下のように行った。まず、与えられた酵素反応式の左右の辺に含まれる化学構造をそれぞれグラフとして表現したあと、複数の部分グラフに切断し部分構造式として表した（**図 2**）。このとき、最適な基質・生成物ペアの選択問題は、左右辺の部分構造式の組み合わせ最適化問題として定式化することができた。すなわち、左右辺のフラグメントの一致を元に**図 3** のようなグラフを形成し、原子の過不足無く結合できるエッジを探索することで、選ばれた最大部分グラフが基質・生成物ペアを、選ばれたフラグメント対が化学構造アラインメントを表すことができた。実際の探索はバックトラック法を用いて実行し、約 85%の精度を得た。

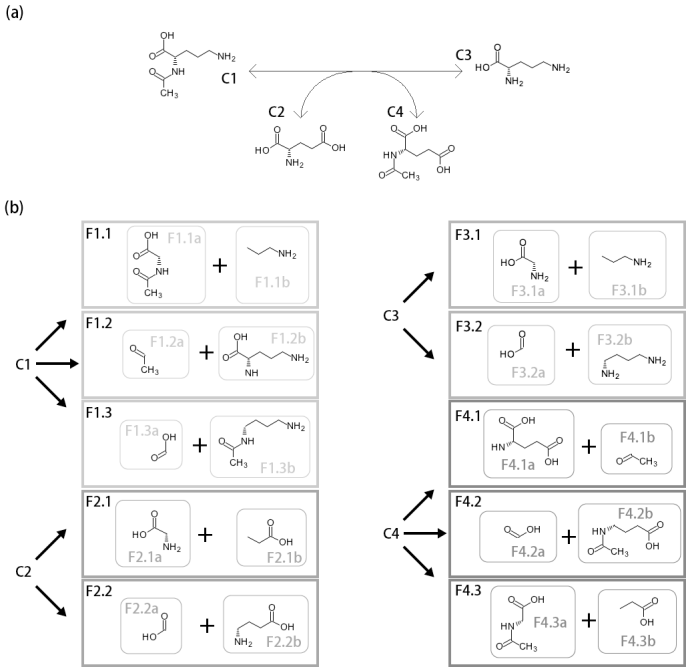

図 2. 部分構造フラグメントへの分解

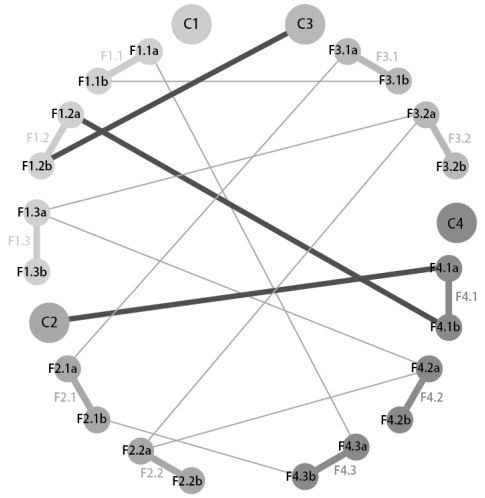

図 3. 基質・生成物ペアを探索するための部分構造組み合わせグラフ

結論

本研究では、酵素反応式から基質・生成物ペアを予測する手法を開発することで、EC 番号予想プログラム E-zyme の操作性の向上を試みた。今後は、RPAIR データベースを解析することで反応中心になりやすい原子に重み付けをするなど、予測精度や計算時間の改善を検討することで、実用化につなげる予定である。

文献・URL

[1] Kanehisa *et al.* "**KEGG for representation and analysis of molecular networks involving diseases and drugs**", *Nucleic Acid Research,* 2010, 38: D355–D360.

[2] Kotera *et al.* "**Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions**". *J. Am. Chem. Soc.* 2004, 126: 16487-16498.

[3] Yamanishi *et al.* "**E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs**". Bioinformatics 2009, 25: i179-i186.

[4] http://www.genome.jp/tools/e-zyme/

[5] Hattori *et al.* "**Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways**". *J. Am. Chem. Soc.* 2003, 125, 11853-11865.

[6] Kotera *et al.* "**RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions**". *Genome Informatics*, 2004, 15, P062.