化学反応カーネルを用いた代謝ネットワークにおける未知酵素の予測

津田　宏治
機械学習研究班
班長 主任研究員

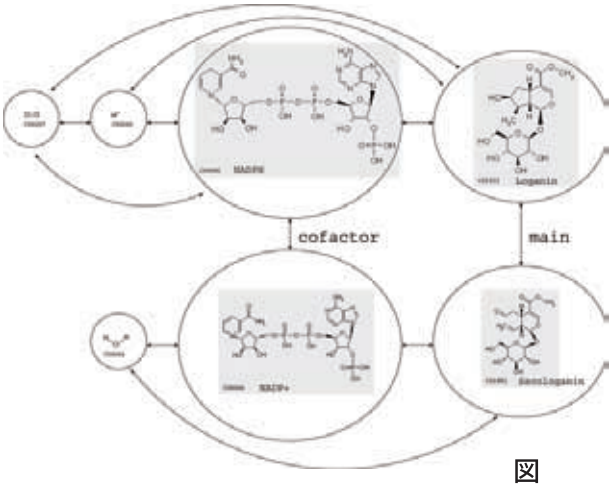
CBRCでは今年度から、機械学習研究班が組織され、私が研究班長を拝命することになりました。これまでのバイオインフォマティクスに加え、東京大学生産技術研究所の喜連川優先生のプロジェクトである「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」にも参画して研究を進めていく予定です。本研究紹介では、最近の研究成果の中から、代謝ネットワークに関連したものを紹介します。

近年、KEGGやARMといった文献をベースとした代謝ネットワークのデータベースが整備されていますが、そこにあるパスウェイは必ずしも完全ではありません。基質や生成物が分かっていても、触媒として働く酵素が未知の場合が多くあり、そのような場合には、データベースの当該部分が空欄として残されています。このような未知酵素は、究極的には、生物学的実験で決定される必要がありますが、すでにデータベース内に類似した反応があり、それに関与している酵素が分かっている場合には、ある程度、計算機によって類推することが可能であると考えられます。酵素を分類する際には、一般にEC番号が用いられます。EC番号は、EC1.3.3.4のような四つ組であり、左から、クラス、副クラス、副々クラスを表します。最後の数字はシリアルナンバーです。副々クラスまで完全に予測することはできなかったとしても、より上位の分類を知ることができれば、代謝経路の理解に貢献することができると考えられます。

生体反応のデータベース（例えばKEGG Reaction）から、今興味を持っている反応と類似した反応を見つけだすには、反応同士の類似度を測る必要があります。本研究では、まず複数の化合物からなる化学反応を、反応グラフとして表現します(図)。反応グラフでは、各化合物がノードとなり、それが、化合物間の関係を表すエッジで繋がれています。エッジには、main, leave, cofactor, transferase, ligaseの５種類があります。反応グラフ間の類似度を表すため、従来のグラフカーネル[2]を拡張した反応グラフカーネル[1]を用います。化合物は、原子のグラフ構造として表されます。反応グラフでは、化合物がノードとなっているため、「グラフのグラフ」という構造を持っています。ここでは、各ノード間の類似度を、あらかじめグラフカーネルで表しておいて、それを統合することによって、最終的な類似度を計算します。

KEGG Reactionを用いた計算実験において、反応グラフカーネルによる類似度検索の精度を、交差検定法によって測定したところ、最上位のクラスで94%程度、最下位の副々クラスでも82%程度の高い精度を得ました。この精度は、KEGG Reactionにおいて用いられている検索システムであるe-zymeよりも優れたものでした。また、植物の二次代謝パスウェイに存在する56個の酵素未知の反応について、EC番号の予測を行い、専門家による予測と照合したところ、36個について一致しました。今後の課題としては、実際の生物学的なタスクにこの手法を適用するということと、複数の酵素が関与する反応も扱える手法を開発することが挙げられます。

参考文献
[1] H. Saigo, M. Hattori, H. Kashima and K. Tsuda, "Reaction Graph Kernels Predict EC Numbers of Unknown Enzymatic Reactions in Plant Secondary Metabolism", *BMC Bioinformatics (APBC 2010)*, (2010).
[2] H.Kashima, K. Tsuda, and A. Inokuchi, "Marginalized kernels between labeled graphs", *Proceedings of the 20th International Conference on Machine Learning*, pp. 321-328, (2003).



図