平成 25 年度　京都大学化学研究所 スーパーコンピュータシステム 利用報告書

基質・生成物ペアの構造に基づく代謝酵素遺伝子の予測
The metabolic enzyme gene prediction based on the structure of the substrate-product pair

化学研究所付属バイオインフォマティクスセンター化学生命科学　守屋 勇樹

背景と目的

生体内における様々な化合物の代謝経路を再構築することは、多様な生体現象を理解する上で非常に重要である。これまで多くの代謝産物が同定され、PubChem、KNApSAcK、KEGG COMPOUND といったデータベースに蓄積されているが、その多くは生合成・分解経路が明らかにされておらず、中間代謝物や代謝酵素などが同定されていない。また、植物界全体では少なくとも 1,000,000 種類の代謝産物を合成するという推定もされており、現在の KEGG PATHWAY や MetaCyc と言った代謝経路データベースはそのごく一部を再構築しているに過ぎない。この状況に対して、我々は以前に、代謝経路の未知な代謝産物における合成・分解反応を推定し中間代謝物を予測するツール、PathPred を作成した[1]。これにより、これまで不明瞭であった代謝経路を、連続反応に置き換えることが可能となった。本研究では代謝経路の再構築に必要な次の段階として、反応の代謝酵素予測を目的とした。

手法

KEGG REACTION データベースでは 6,506 の反応において、その反応を代謝する酵素がオーソロググループ（KEGG Orthology、KO)の形で同定されており、そのうち 4,984 反応は複数の反応を代謝する酵素によっても代謝されている。これは多くの反応が、他の反応と代謝酵素を共有していることを意味している。そこで本研究では同一の酵素によって代謝される反応を推定することで、反応の代謝酵素予測を行った。

同一の酵素によって代謝される反応同士は構造類似性を持つと考えられるが、現在反応間の類似性を検出する手法が存在しない。そこで、化合物を記述子としての部分構造とその数で表現した KCF-S[2]を拡張し、基質・生成物ペアを表現するための aligned KCF-S を定義した（図 1）。これにより反応同士の類似性を基質・生成物ペアのレベルで計算することが可能になった。

まず、代謝酵素の未知な問い合わせ基質・生成物ペアにおいて、E-zyme プログラム[3]を用いて構造アラインメントを行い、問い合わせペアの aligned KCF-S を作成した。次に類似した基質・生成物ペアを KEGG RPAIR データベースから検出した。その際、aligned KCF-S 間の Tanimoto 係数を類似度として用いた。次

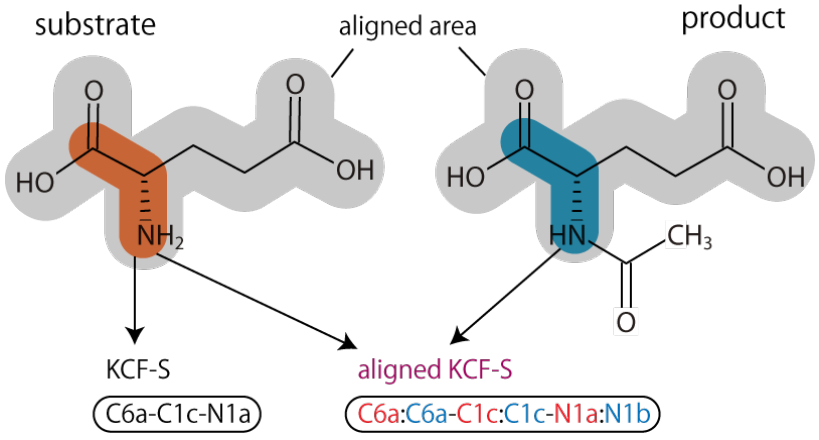

図 1 ：aligned KCF-S と KCF-S の記述子の例

に同一の KO によって代謝される基質・生成物ペアに共通して見られる部分構造を、その KO に特徴的な構造として抜き出すことで、KO を表現する aligned KCF-S を作成した。最後に問い合わせ基質・生成物ペアと KO 間の類似度を計算し、最も類似度の高い KO を問い合わせ基質・生成物ペアの代謝酵素とした。

また、今回提案した基質・生成物ペアのレベルでの類似度を用いた手法と比較するため、化合物レベルの類似度を用いた手法を次のように定義した。基質同士の類似度、及び生成物同士の類似度の平均を基質・生成物ペア間の類似度とし、RPAIR データベース中で最も類似度の高くなるペアを代謝する KO を、問い合わせペアの代謝酵素とした。

結果と考察

手法を評価するため、交差検定の一つである leave-one-out cross-validation を行い、次の３つの指標によって評価した。assign rate: KO をアサインしたテストの割合、correct rate: アサインした KO の中に正解が少なくとも一つ含まれていたテストの割合、TP rate: それぞれのテストにおける正解 KO の割合の平均。図２は様々な類似度の閾値で代謝酵素の予測を行った結果を示している。緑で示された線は今回提案する基質・生成物ペアのレベルでの類似度（aligned KCF-S）を用いた結果で、青で示された線は化合物レベルの類似度（KCF-S）を用いた結果を表している。その結果、提案する手法において correct rate 及び TP rate が全体に渡って高くなっており、aligned KCF-S は化合物レベルの類似度に基づいた手法よりも、反応と代謝酵素との関連性を検出できている。また、類似度 0.9 を閾値とした場合 assign rate は 50%を超え、その際の correct rate 及び TP rate はそれぞれ 78%と 69%となり、反応の代謝酵素予測の第一段階として有用な手法と言える。

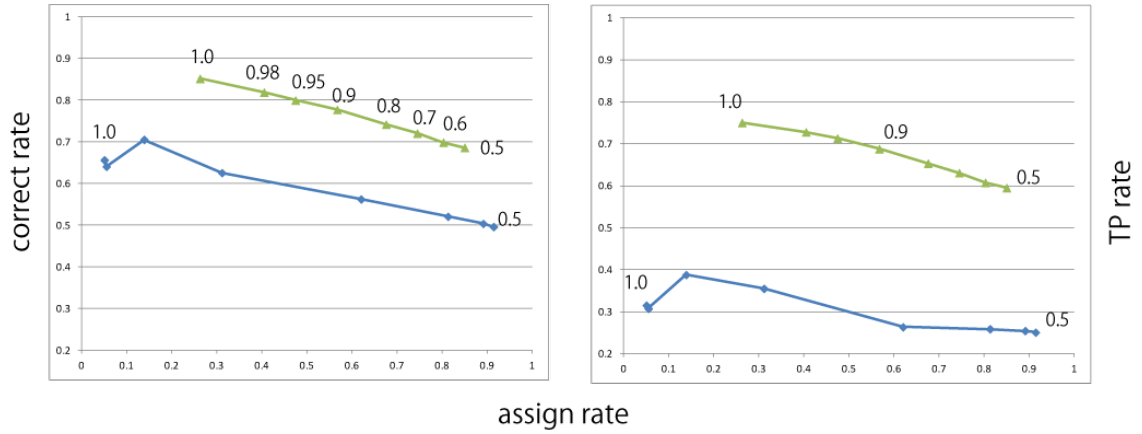

図２：手法の性能比較

参考論文

[1] Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S., and Kanehisa, M.; PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Res. 38, Web Server issue (2010).

[2] Kotera, M., Tabei, Y., Yamanishi, Y., Moriya, Y., Tokimatsu, T., Kanehisa, M., and Goto, S.; KCF-S: KEGG Chemical Function and Substructure for improved interpretability and prediction in chemical bioinformatics. BMC Syst Biol 7 (2013).

[3] Yamanishi, Y., Hattori, M., Kotera, M., Goto, S., and Kanehisa, M.; E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. Bioinformatics 25 (2009).