ゲノムからのパスウェイ推定の為の
バイオインフォマティクス研究

守屋 勇樹

概要

生体内の代謝反応やシグナル伝達などの分子間相互作用ネットワーク、いわゆるパスウェイを俯瞰することは、様々な分子生物学的解析の礎となり、生命現象を理解する上で重要な役割を担っている。そのため KEGG PATHWAY や BioCyc、Reactome と言った多くのパスウェイデータベースが構築されており、様々な生物種の持つパスウェイが計算機上で再現されている。KEGG では、ゲノムの決定した生物種の配列データを蓄積し、ゲノムアノテーションを行うことで各生物種においてパスウェイの構築を行っているが、パスウェイの構築には多くの専門的知識と手作業による精査が必要であり、ゲノム配列の決定と、その生物の持つパスウェイの再現を直接結びつけることは容易ではない。一方で、近年の DNA シークエンス技術の発達により、多くの配列データが蓄積されるようになり、配列データからのパスウェイ構築の自動化が求められるようになっている。そのため本研究では３つの研究を行った。

第一の研究は､遺伝子機能の自動アノテーションに基づくパスウェイの構築である。新規に配列の決定された遺伝子の、他の生物種におけるオーソログ遺伝子を予測することにより、その機能の推定を行い、推定された機能を基にリファレンスパスウェイにマッピングを行うことでパスウェイの構築を行った。

第二の研究は、未知代謝パスウェイの予測である。生物界に存在する全ての代謝パスウェイを表現するにはパスウェイデータベースの整備は未だ不十分であり、中間代謝物の判明していない代謝パスウェイが多く存在する。そのため、データベースに登録されている既知の反応パターンを基に、新規に中間代謝物を予測することにより、リファレンスパスウェイには登録されていない代謝パスウェイの構築を行った。

第三の研究は、基質と生成物の構造を用いた代謝酵素の予測である。パスウェイデータベースには代謝活性は確認されているが、触媒する酵素遺伝子が同定されていないオルファン酵素が、また予測されたパスウェイにも遺伝子の同定されていないミッシング酵素が、パスウェイの欠損として存在している。これを補完するため、データベースに登録されている反応パターンを拡張することで代謝反応の類似性をより詳細に計算することを可能とし、また反応を代謝する遺伝子のオーソログ及びパラログを探索することで、代謝反応の酵素遺伝子を予測した。

これらの研究により、パスウェイ推定の大幅な自動化が進み、様々な生物の生命現象の解析に利用できるようになった。また予測されたパスウェイを実験にフィードバックすることで、より正しいパスウェイを求めるための手がかりとして用いることが可能であり、さらなる生命現象の理解への貢献が期待される。

目次

図目次

表目次

第１章　序論

１．１　ゲノム解析とパスウェイデータベース

DNA シークエンス技術の発展に伴い現在では、ゲノム解析は生命現象を理解する上で欠かすことのできない解析手法となっている。1990 年代後半から 2000 年代前半までに、サンガー法 [1] を用いた第一世代シークエンサによりヒトを含む多くのゲノム配列が解読され、ゲノムデータベースに蓄積されている。また、2000 年代後半には第二世代シークエンサの登場により多くの DNA 配列を超並列でシークエンスすることが可能となり、1000 人ゲノムプロジェクト [2,3] などの大型シークエンスプロジェクトが、これまでとは桁の異なる量の配列データが産出している。また、ゲノム解析だけでなく RNA-Seq 技術を用いたトランスクリプトーム解析も可能となり [4]、mRNA 配列データも多数蓄積されるようになった。さらには、コンピュータの計算能力の向上や、DNA 断片をつなぎ合わせるための *de novo* アセンブル法の発展により、生物単体を解析するだけでなく、環境サンプルに含まれる生物集合全体をまとめて解析するメタゲノム解析も盛んになった。その結果、腸内細菌叢メタゲノム解析や海洋メタゲノム解析といった様々な環境におけるメタゲノム解析が行われ [5,6]、単体生物とは規模の異なる量の配列データも蓄積されている。2016 年現在、ゲノム解析、トランスクリプトーム解析、メタゲノム解析等を目的としたプロジェクトを合わせると、97,212 のシークエンスプロジェクトが行われ、日々データが蓄積されている [7]。さらには第三世代シークエンサと呼ばれる、PCR による DNA の増幅を必要としない１分子シークエンシング技術の開発も進んでおり [8]、今後もより高速に、より安価で容易にゲノム配列が蓄積されていくと考えられる。

ゲノム配列を決定しただけでは生物を十分に理解することはできない。そのためゲノム解析ではまず遺伝子領域の予測と生体内で遺伝子が担う機能を推定することが重要である。遺伝子領域予測は確率モデルや配列相同性検索を用いて行われている。また、イントロンやオルタナティブスプライシングが行われる真核生物では予め解析した EST や mRNA 等の転写後の配列情報を用いゲノム配列と比較することで、遺伝子領域の予測が行われている場合もある [9]。予測された遺伝子の機能を推定し、意味付けすることを遺伝子の機能アノテーションと呼び、現在のゲノム解析においては、機能既知の遺伝子との配列相同性によって推定される。また、遺伝子領域の予測や機能アノテーションを生物種（ゲノム）単位で全ての遺伝子に対して行うことをゲノムアノテーションと呼ぶ [10]。機能アノテーションに用いられる機能分類の一つに Gene

Ontology (GO)がある [11]。GO は遺伝子の機能を階層構造で体系的に分類した語彙集で、生物学的プロセス、細胞の構成要素、分子機能の 3 つの視点で分類されている。異なる生物で共通に扱える機能分類を用いてゲノムアノテーションを行うことで、ゲノムの比較解析が可能となる。また、アノテーションされた個々の遺伝子を機能階層のカテゴリに従って分類することで、ゲノム中の遺伝子機能を俯瞰することが可能になり、研究者が生命現象を理解するための手がかりとして利用されている。現在では GO の他に様々な視点からの機能分類のための語彙集が作成されている [12]。現在よく利用されている機能分類の一つに、生物学的パスウェイに基づく分類が挙げられる。

パスウェイデータベース

生物学的パスウェイ（パスウェイ）とは、生体内での遺伝子やタンパク質、その他の化合物等の分子間相互作用ネットワークを"経路"として表現したものであり、経路を俯瞰するために図式化したもの、また計算機で扱うために電子化したものがパスウェイデータベースである。1965 年に初めて、ベーリンガー・マンハイム社（現エフ・ホフマン・ラ・ロシュ社）が代謝経路全体を図式化したパスウェイを出版した [13]。図 1.1 はベーリンガー・マンハイム社が出版したものと同様に代謝経路全体を示したパ

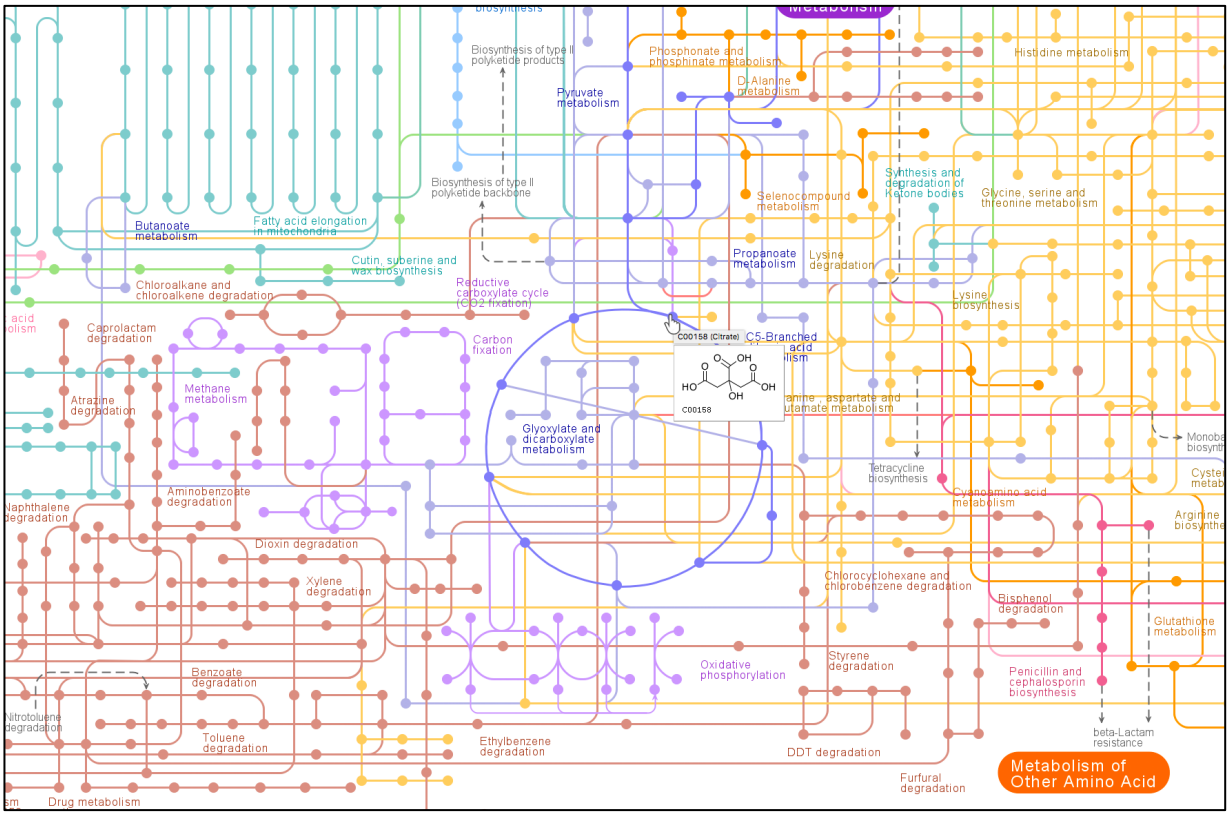


図 1.1　代謝パスウェイの例（TCA 回路とその周辺）

ウェブサイト[14]より引用

スウェイの例で、生体内で連続して起こる代謝反応を、代謝を受ける化合物（ノード、頂点）と反応を触媒する酵素（エッジ、辺）の関係のネットワーク図で表している[14]。これにより代謝経路全体を俯瞰することが可能となり、生命現象の理解への手助けとなっている。

1990 年代の計算機とインターネット技術の発展により、パスウェイを計算機上で表現し、インターネットを通じて可視化したデータを提供することが可能となり、1995 年には KEGG (Kyoto Encyclopedia of Genes and Genomes) が、文献から集めた知識ベースの代謝パスウェイをデータベース化した KEGG PATHWAY データベースの構築を開始した[12,15]。また 1996 年には BioCyc/MetaCyc データベースの前身となる大腸菌代謝パスウェイデータベースである EcoCyc が発表され[16,17]、2004 年にはヒトを中心とした脊椎動物のパスウェイに焦点をあてた Reactome データベースが公開されている[18]。その後もタンパク質の知識ベースである UniProtKB データベースと紐付いた UniPathway[19]、コミュニティベースで開発が進められている WikiPathways など多くのパスウェイデータベースが公開されている[20]。さらに、パスウェイで取り扱う相互作用を代謝化合物と酵素という関係から、多種多様な分子間相互作用へと拡張することによって、パスウェイデータベースは転写制御やシグナル伝達、物質輸送や免疫システム、病気などの分子間相互作用ネットワークを表現する目的でも用いられるようになっている。

多くのパスウェイデータベースでは分子間相互作用ネットワークを機能によって区切ることで分類している。例えば KEGG PATHWAY データベースでは代謝系を炭水化物代謝、エネルギー代謝、脂質代謝など 12 のパスウェイカテゴリに分類しており、さらに炭水化物代謝は解糖系、TCA 回路、ペントースリン酸回路などに細かく分けられている。本来生体内にある一つの大きな分子間相互作用ネットワークを、このような階層構造をもった機能分類に落とし込むことで、個々の生命現象が理解しやすくなり、またゲノムアノテーションにおいても有用な情報となる。

また、いくつかのパスウェイデータベースでは、各遺伝子機能の実証実験の行われていない生物種においても、オーソログ遺伝子の情報を利用することでパスウェイを計算機上で構築している。オーソログ遺伝子とは、祖先生物における同じ遺伝子から分岐した遺伝子であり、一般に生物間で同一の機能を持つ。パスウェイデータベースでは、他の生物において機能の実証された遺伝子に対応するオーソログ遺伝子を推定し、同一の機能を有している同一の相互作用因子としてパスウェイ上で取り扱うことで、様々な生物種のパスウェイを計算機上で表現している。2016 年現在、異株を含め KEGG PATHWAY データベースでは約 4,300 種、BioCyc/MetaCyc では約 7,600 種

のパスウェイデータベースが構築されており、モデル生物に限らず多種多様な生物種の解析においてパスウェイデータベースを利用することが可能となっている。

こうして構築されたパスウェイデータベースは、ゲノムアノテーションに利用されるだけでなく、エンリッチメント解析、シミュレーションなどバイオインフォマティクスの分野で広く利用されている。例えばエンリッチメント解析では、DNA マイクロアレイや次世代シークエンサ技術に基づいた RNA-Seq、または質量分析計を用いたプロテオーム解析など、様々な大規模発現解析で得られた発現変動遺伝子群がどの機能カテゴリで有意に出現しているかを知るために、パスウェイデータベースが用いられている [21,22]。また、より生命現象を理解しやすくするために、発現量の変化をパスウェイのネットワーク画像の上に重ね合わせて可視化する目的でも、パスウェイデータベースは多くの研究で利用されている。

１．２　本論文の目的と概要

上記のように、ゲノム配列の決定とパスウェイデータベースの整備により、これまでの個別の分子間相互作用に焦点をあてた解析だけでなく、ゲノムに内包される相互作用ネットワーク全体を俯瞰した解析を行うことができるようになった。またこれにより、複数のゲノムを用いた比較ゲノム解析も可能となり、ゲノム全体を用いた種の進化と多様性の解明や [23]、共生生物における共生関係の分子生物学的理解 [24]、または病原細菌における病原性遺伝子解析や [25]、モデル生物ゲノムを用いた疾患遺伝子の探索など [26]、これまでとは違った複雑な生物学的知見の獲得や利用を行うことができるようになってきている。しかしながら、パスウェイデータベースの構築は未だ限定的であり、全ての生物種の全ての分子間相互作用を表現するには至っていない。また、手作業によるキュレーションが必要であるため、将来的にもそれら全てを表現することは不可能であると考えられる。そこで本論文では、現在の限定的なパスウェイデータベースを利用して、ゲノムアノテーション及びパスウェイ推定を自動的に高速で行うための手法を開発することにより、可能な限り多くの生物種の分子間相互作用ネットワークを俯瞰することを可能にし、またこれらの生物種の比較ゲノム解析への利用を促進するための基盤となる技術の開発を目的とした。

現在、DNA シークエンス技術の発達により、大量のゲノム配列が産出されるため、より高速なゲノムアノテーション法が必要になっている。ゲノムアノテーションの正確性と高速性を両立するために、ゲノム解析では、機能既知の遺伝子との配列相同性情報を用いて、個々の遺伝子の機能アノテーションを行っている。しかしながら、配列相同性を持った遺伝子（ホモログ遺伝子）同士が常に同じ機能を持っているとは限

らず、異なる機能を持った複数の遺伝子と配列相同性を有している遺伝子も多い[27,28]。そのため、より正確な機能推定を行うためには同一の祖先遺伝子から分岐し、同一の機能を有していると考えられるオーソログ遺伝子を同定することが重要である。KEGG ではオーソロググループを作成し、新規ゲノムの遺伝子をオーソロググループに割り当てることで機能アノテーションを行っているが、これまで KEGG ではオーソログ遺伝子を自動で推定する手法が無く、手作業によるオーソログ遺伝子推定を行っていた。そのため第 2 章において述べるように、生物間の遺伝子の配列相同性を基に、オーソログ遺伝子を統計的に自動で予測することによって、遺伝子の機能アノテーションを行う手法を構築した。これにより新規にシークエンスが行われた生物種においても高速に精度の保たれたゲノムアノテーションを行い、パスウェイを計算機上で可視化できるようになった。

一方で未だパスウェイデータベースに登録されていない未知の分子間相互作用も多く存在する。KEGG PATHWAY データベースでは微生物による環境物質の分解経路や、植物による二次代謝産物の合成経路などの知識の蓄積が遅れているために、データベースに登録されていない未知の代謝経路が存在している。これら未知の代謝経路を推測する手法の一つに、代謝反応における代謝化合物の化学的構造変化の特徴を用いるものがある。代謝パスウェイの中には共通した、または類似した構造変化の特徴を持った代謝反応が多く含まれていることが知られている[29]。そのため、パスウェイデータベース中の代謝反応における構造変化の特徴を反応パターンとして抽出・蓄積し、これら反応パターンを代謝経路の不明な代謝化合物に当てはめることで、中間代謝産物の構造を予測し、ひいては代謝経路を予測することが可能になる。KEGG では代謝パスウェイ中のそれぞれの代謝反応において代謝化合物の生化学的な構造変化のパターンを蓄積したデータベースを構築している。そこで第 3 章で述べるように、代謝経路が同定されていない化合物に対して、データベースに蓄積されている反応パターンを順番に当てはめることによって、パスウェイデータベースには登録されていない新規代謝経路を予測する手法を開発した。

反応パターンを基に代謝反応を予測することが可能となったが、予測された代謝反応には代謝を触媒する酵素が同定されずに残ることになる。また、既存の代謝パスウェイデータベースにおいても代謝酵素が未だ発見されていないオルファン酵素が分子間相互作用ネットワークの穴として存在している。過去 10 年で新たに報告された酵素反応の約 40%では、代謝酵素が未知のまま残っている[30]。これまでに、代謝酵素が未知な反応の EC サブ–サブクラス（EC 番号の前方 3 桁）を予測する手法が開発されている[31]。これは、EC 番号の前方三桁が酵素の代謝反応の種類を表し、四桁目が基

質特異性やシリアル番号を表しているため、同一の EC サブ–サブクラスには構造変化の特徴、反応パターンが類似した酵素が含まれている傾向があるという観測結果に基づいている[32]。しかし、EC サブ–サブクラスには多くの酵素反応が含まれており、そこから代謝酵素を推定することが難しかった。そこで第 4 章で述べるように、反応において構造変化が起こる部分を局所的に表現したこれまでの反応パターンを、より広い範囲の構造の特徴を含む反応パターンに拡張した。これにより類似した酵素反応をデータベースから探索することが可能となり、また酵素のパラログ遺伝子、オーソログ遺伝子を探索することで、オルファン酵素の候補を予測することが可能となった。

第２章　遺伝子機能の自動アノテーションとパスウェイマッピング

２．１　背景

遺伝情報の解析は生命現象を理解する上で重要である。1977 年のジデオキシ法（サンガー法）[1] の発明により DNA シークエンス技術は分子生物学研究の分野に広く普及し、遺伝情報を解析するために無くてはならないものとなった。1995 年に独立生活を行う生物としては初めて、インフルエンザ菌の 1.8Mb の全ゲノム配列が決定されたのを皮切りに [33]、より大きなゲノム配列の決定が進み、真核生物である出芽酵母のゲノム（12Mb）の解読 [34]、多細胞生物である線虫ゲノム（100Mb）の配列決定が行われた [35]。その後のショットガン・シークエンシング法の普及とゲノムアセンブリ技術の発展により、2004 年にはヒトゲノム（3.1Gb）の解読完了が報告されるまでになった [36]。さらに、近年の次世代シークエンサの登場により、より高速で安価なゲノム配列の決定が行えるようになったことで、膨大な数のゲノム配列が解読、報告がなされている。これらゲノム配列が決定された生物の生命現象を解析するためには、ゲノム上にコードされている遺伝子の機能のアノテーションが不可欠であるが、膨大な数のゲノム配列を高速に処理するためには、計算機による遺伝子の網羅的な機能予測が必要である。それを可能とする最も効果的な手法に、機能既知な遺伝子との配列比較があり、Smith-Waterman アルゴリズム [37] や FASTA[38]、BLAST[39,40] などの配列相同性検索のための手法やツールが開発されている。しかしながら配列相同性を持った遺伝子同士が常に同一の機能を有しているとはいえず、酵素遺伝子における機能予測を行った研究において、90%の精度の機能予測のためには 40～70%の配列相同性が必要であると報告されている [27,28]。また、ゲノム上にコードされている全遺伝子との配列比較を行うことで、種間で比較した際に最も配列相同性が高い遺伝子（ベストヒット）かどうかの情報が遺伝子の機能予測に利用できる。オーソログ遺伝子は共通祖先遺伝子から種分岐によって生じた、生物間で対応する遺伝子で、配列相同性を有し、機能が保存されていると考えられている。そのため、ペアワイズゲノム比較を行うことで、どちらの種から見てもベストヒットとなる遺伝子ペア（双方向ベストヒット）を探索することでオーソログ遺伝子を推定することが可能である [41,42]。このため、新規に解読されたゲノム上にある遺伝子とのオーソログ遺伝子を多くの生物種との間で同定することが、新規ゲノム遺伝子の機能予測への近道である。

機能予測の精度は探索を行うデータベースの質に依存する。KEGG GENES データベースはゲノム配列の決定したほぼ全ての生物種に含まれる全遺伝子が登録されており、KEGG Orthology（KO）と呼ばれるパスウェイに基づいた、全生物種で統一的に利用可能なオーソログ遺伝子の機能分類システムに従って、専門家の手作業による精度の高い機能アノテーションが行われている[12]。KO で定義されている各オーソロググループは KEGG PATHWAY データベースの各要素として紐付けされており、ゲノム上の遺伝子に KO を付与していくことで、生物種の持つパスウェイを計算機上で表現できるようになっている。また、KO システムは KEGG BRITE と呼ばれる階層的に編纂された語彙による機能分類にも拡張されており、KO を介して様々な機能分類へのマッピングが可能となっている（図 2.1）。このため、KEGG はゲノムと生命システムを結びつけるのに適したデータベースとして構築されている。しかしながら KEGG における遺伝子に対する KO の付与には手作業による精査が必要で、多くの時間と労力がかかっていた。

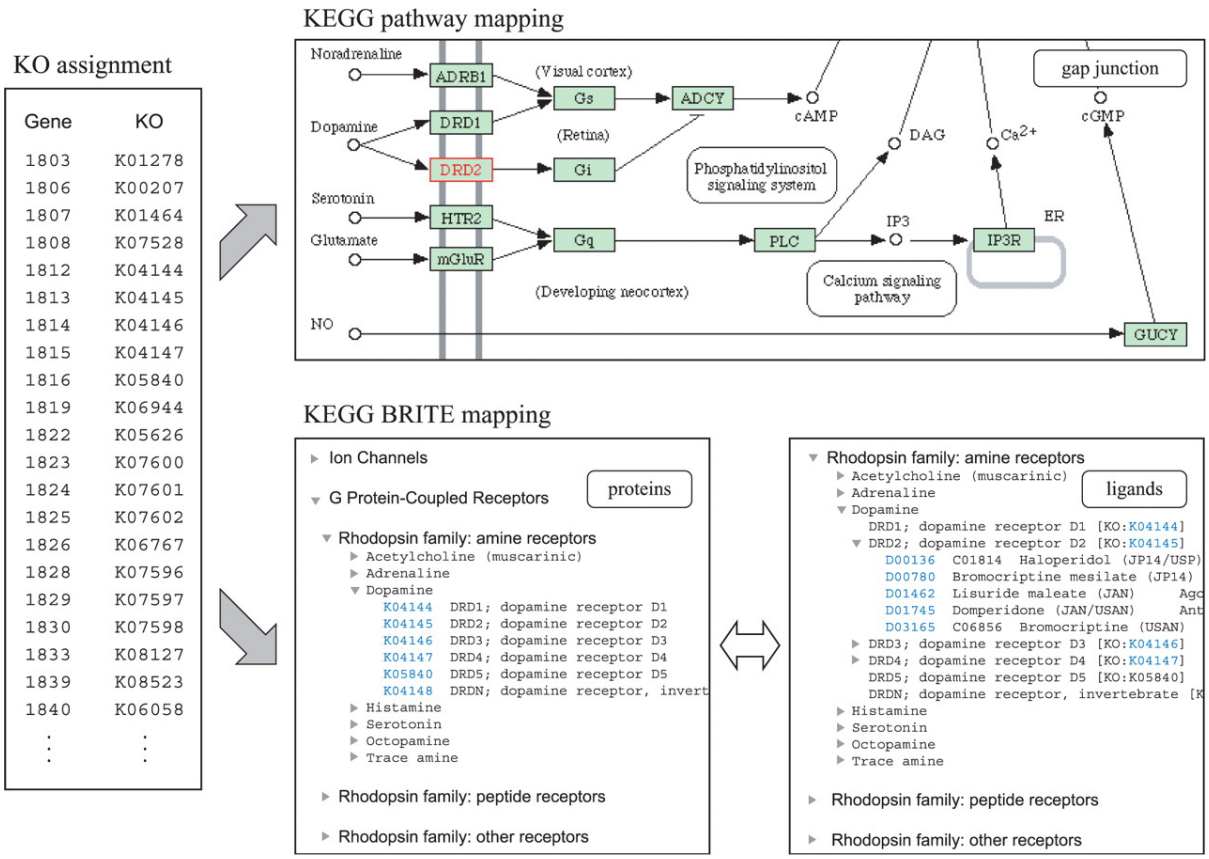


図 2.1　KO を介したゲノムアノテーションの例[44]

そこで本研究では、新規にゲノム配列が決定した生物種において、遺伝子の KO を自動的に予測することで機能アノテーションを行い、KO を介して KEGG PATHWAY にマッピングを行うことで生物種の持つパスウェイを計算機上で表現することを目的とした。さらに、本予測手法を実装したウェブサービスを開発することで、計算資源を持たない研究者にも広く利用できる環境を構築することを目的とした。

２．２　材料と方法

リファレンス配列データベースと代表生物種セット

リファレンス配列データベースとしての KEGG GENES データベースと KO システムは、2006 年 12 月に登録されていた真核生物 36 種、真性細菌 402 種、古細菌 31 種の計 469 生物種のデータを用いた。また、提案する手法の計算コストはリファレンスとなるデータベースのサイズに比例して大きくなる。そのため 2006 年当時、ウェブサービスを展開するにあたって実用的な時間内での機能アノテーションを行うためには、リファレンス配列データベースを制限する必要が生じた。ここでは、KO 予測精度を大きく損なうことなく計算コストを減らすため、KEGG GENES データベースにおいて比較的に、手作業によって機能アノテーションの精査が進んでいる生物種をなるべく系統分類的に広範囲から収集した。また真核生物と原核生物では、機能アノテーションの差異が大きく、有効なリファレンスが異なると考えられるため、真核生物の機能アノテーションを対象とした代表生物種セット 26 種と、原核生物の機能アノテーションを対象とした代表生物種セット 28 種を策定し、ウェブサービスにおけるリファレンス配列データベースの代表例とした（付録 1)。

手法の概要

図 2.2 は遺伝子に KO を割り当てるための手順を示している。まず、問い合わせ遺伝子とリファレンス配列データベースとなる KEGG GENES データベースの遺伝子との間で配列相同性スコア（BLAST ビットスコア）を相同性検索ツール、BLAST を用い算出し相同遺伝子として抽出した。続いて双方向ベストヒットの情報（双方向ヒット率）に基づいた足切りを行い、オーソログ遺伝子候補を選択した。次にオーソログ候補遺伝子を KO 毎に分類し、各 KO グループにおいて確率に基づいた KO 割当スコアを算出し、ランキングを行った。ランキングトップの KO グループを問い合わせ配列の機能アノテーションとして割当を行った。

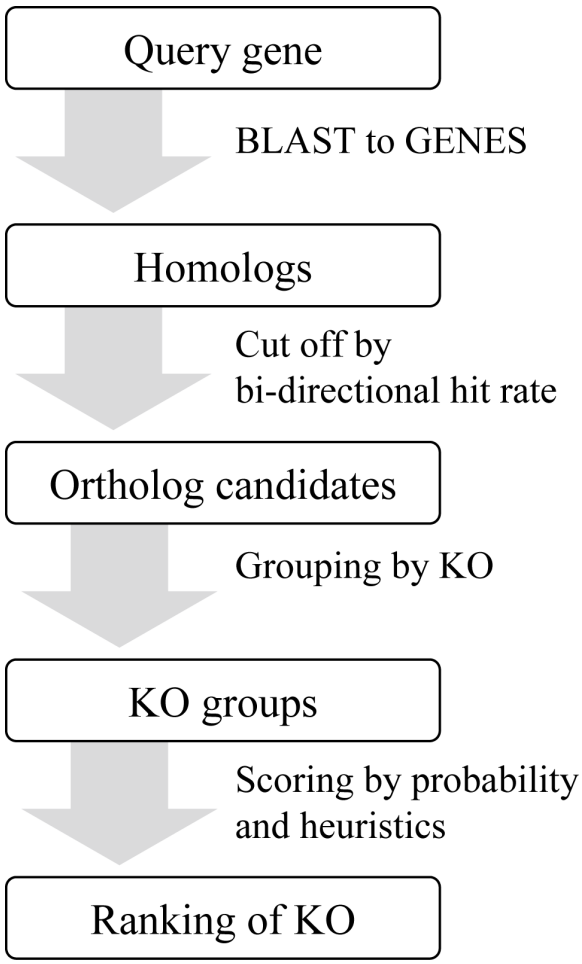


図 2.2　手法の手順 [44]

BLAST 検索と双方向ヒット率

BLAST を用いた相同性検索はベストヒットの情報を取得するために、機能アノテーションを行う種と、リファレンス配列データベースとの間で、種対種毎に計算を行った。その際、BLAST の設定はデフォルト設定を用いた。また BLAST は問い合わせ遺伝子セットとデータベースを入れ替えることにより、ビットスコアが変化する場合があるため、順方向、逆方向の双方向でスコアを計算した。リファレンス配列データにはアミノ酸配列を用い、問い合わせがアミノ酸配列である場合は双方向で BLASTP を、問い合わせが塩基配列である場合には順方向で BLASTX、逆方向で TBLASTN を用いることで、アミノ酸レベルでの配列相同性検索を行った。探索空間の節約を行うため、有意性の低いと思われる、双方向のビットスコアの平均が 60 以下の遺伝子を除外し、残りを相同遺伝子とした。BLAST は相同配列探索が比較的高

速である一方で、Smith-Waterman アルゴリズムのような最適アライメントを保証しないため、誤差を許容した双方向ヒット率（Bi-directional hit rate; BHR）を次のように定義し、相同遺伝子の閾値として用いた。

$$BHR = R_f \times R_r$$

このとき、Rはベストヒット遺伝子のビットスコアに対する、各スコアの割合を示している。例えば、生物種 A の問い合わせ遺伝子 a と生物種 B のターゲット遺伝子 b 間のビットスコアが 250 であり、同時に遺伝子 a と生物種 B のベストヒット遺伝子間のビットスコアが 300 であった場合、遺伝子 a と遺伝子 b のペアの割合は 250/300 になる。また、R_f、R_rはそれぞれ順方向（forward）、逆方向（reverse）での割合を示している。そのため、遺伝子 a と遺伝子 b が双方向ベストヒットであった場合、BHR は 1 になる。ここでは BHR が 0.95 以上の遺伝子を双方向ベストヒットに近い遺伝子とみなし、オーソログ遺伝子候補とした。

KO 割当スコア

最適な KO を問い合わせ遺伝子に割り当てるために、それぞれの KO グループにおいて KO 割当スコア（S_{KO}）を次のように定義した。

$$S_{KO} = S_h - \log_2(mn) - \log_2\left(\sum_{k=N}^{x} {}_xC_k\, p^k (1-p)^{x-k}\right)$$

このとき、S_hはそれぞれの KO グループで最も高いビットスコアを示しており、nはその際のターゲット遺伝子の配列長、mは問い合わせ遺伝子の配列長、Nは各 KO グループに含まれる生物種数、xはリファレンスとなった KO グループに含まれる生物種数、pはリファレンス配列データベースの全遺伝子数に対する KO グループの遺伝子数の割合をそれぞれ示している。第二項は配列長による第一項の補正項となっている。第三項はリファレンスとなった KO に含まれる遺伝子のうち上記の手順によって検出できたオーソログ遺伝子候補の数を考慮した重み付けとなっており、偶然 N 種以上でオーソログが検出できる確率（有意確率）を BLAST のビットスコア同様にビットスコア化することで導出した。

交差検証

手法の評価を行うため、交差検証（Cross-validation）を行った。まずリファレンス配列データベースから 1 生物種抜き出して問い合わせ生物とし、残った生物種を検定におけるリファレンスとした。続いて、提案する手法を用いて問い合わせ生物種の遺伝子の KO を予測することで機能アノテーションを行い、再現性を検証した。検証ではヒト（*Homo sapiens*）、シロイヌナズナ（*Arabidopsis thaliana*）、出芽酵母（*Saccharomyces cerevisiae*）、大腸菌（*Escherichia coli* K-12 MG1655）の 4 種で行った。

２．３　結果と考察

表 2.1 は KEGG GENES データベース全体をリファレンス配列データベースとした場合の交差検証の結果を示している。また表中の Sensitivity（感受性）、Specificity（特異性）、PPV（Positive predictive value、陽性的中率）、Precision（精度）は次のように算出した。

- Sensitivity : KEGG データベースで元来 KO がアノテーションされていた遺伝子において、提案した手法で正しく KO の再割当てが行われた遺伝子の割合
- Specificity : 元来 KO がアノテーションされていなかった遺伝子において、KO 割り当てが行われなかった遺伝子の割合
- PPV : KO の再割当てが行われた全遺伝子における、正しく KO の再割当てが行われた遺伝子の割合
- Precision : 元来 KO がアノテーションされていた遺伝子で、かつ KO の再割当てが行われた遺伝子のうち、正解した遺伝子の割合

表 2.1 KEGG GENES 全体をリファレンスとした場合の KO 再割当ての精度

Species	*H. sapiens*	*A. thaliana*	*S. cerevisiae*	*E. coli*
Sensitivity	83.7%	70.4%	85.2%	97.4%
Specificity	98.6%	91.5%	94.1%	94.3%
PPV	93.6%	47.9%	80.7%	94.9%
Precision	98.0%	85.5%	91.6%	98.5%

大腸菌では全ての指標で 94%を超えており、非常に精度の高い KO 予測を行えた。これはリファレンスとした KEGG GENES データベースに多くの、株違いの同種を含む近縁種が含まれていたためだと考えられる。また、ヒトでも感受性以外の指標で 93%を超える KO 再割当てが行われており、十分に高い精度で予測できたといえる。大腸菌と比較して、ヒトでの感受性が大きく下がった理由は、近縁種が大腸菌ほど多くないことが原因と考えられる。シロイヌナズナにおいては感受性と陽性的中率が顕著に下がっており、これは KEGG GENES データベースに他の植物が登録されておらずリファレンスとなる近縁種が存在しないこと、植物遺伝子の機能同定が進んでおらず多くの遺伝子が機能アノテーションされずに残っていることなどが主な原因と考えられる。

表 2.2 代表生物種セットをリファレンスとした場合の KO 再割当ての精度

Species	*H. sapiens*	*A. thaliana*	*S. cerevisiae*	*E. coli*
Sensitivity	85.4%	62.5%	86.8%	90.1%
Specificity	98.9%	91.3%	96.8%	94.9%
PPV	94.4%	44.3%	87.7%	93.2%
Precision	97.9%	83.8%	94.9%	96.6%

表 2.2 はリファレンス配列データベースに代表生物種セットを用いた場合の交差検証の結果を示しており、4 種とも極度に精度を減じることなく予測が行えている。大腸菌では感受性が下がっており、同種の異株がリファレンスから除外されたことが影響していると考えられる。一方、ヒトと出芽酵母では、予測精度が KEGG GENES データベース全体をリファレンスとした場合とほぼ同程度で、幾つかの指標では数値の向上が見られる。これは KEGG GENES データベースに登録されている生物種は真性細菌が圧倒的に多いため、リファレンス配列データベースに含まれる生物種の系統分類的な偏りが影響したと考えられる。

2016 年現在までに、提案した手法における配列相同性検索を、より精度の高い配列相同性検索を行える SSEARCH プログラムに置き換えた手法及び、ベストヒットだけでなくタンパク質のドメイン情報など、より多くの情報を用いた自動アノテーションプログラム、KOALA[43] によって KO のアノテーションが進み、KEGG GENES データベースに登録される生物種は 4,200 種を超え、非常に多岐の系統分類に渡って KO アノテーションがなされている。また、リファレンス配列データベースとしての質を

維持するため、本手法のリファレンスとしての対象は手作業による精査を行った生物種に限られるが、それでも 1,500 種を超えており、提案する手法にとって非常に有用なリファレンスデータベースとなっている。

本研究で提案した手法は KAAS: KEGG Automatic Annotation Server (http://www.genome.jp/tools/kaas/)としてウェブサービスが提供されており、FASTA 形式の問い合わせ遺伝子配列を入力し、リファレンス配列データベースを、予め設定した代表生物種セットだけではなく KEGG GENES データベースから自由に選択することにより、自動で KO アノテーションの予測を行い、KEGG PATHWATY や BRITE へマッピングすることが可能となっている (図 2.3) [44]。ウェブサービスでは、リファレンスとなる配列データベース及び機能アノテーションとなる KO システムが KEGG GNENS データベースの更新に伴い逐次更新される体制を取っている。これにより様々な生物において、これまで難しかったパスウェイ再構築が、様々な生物種において容易にできるようになり、ネットワーク解析やエンリッチメント解析に有用なツールとなっている。

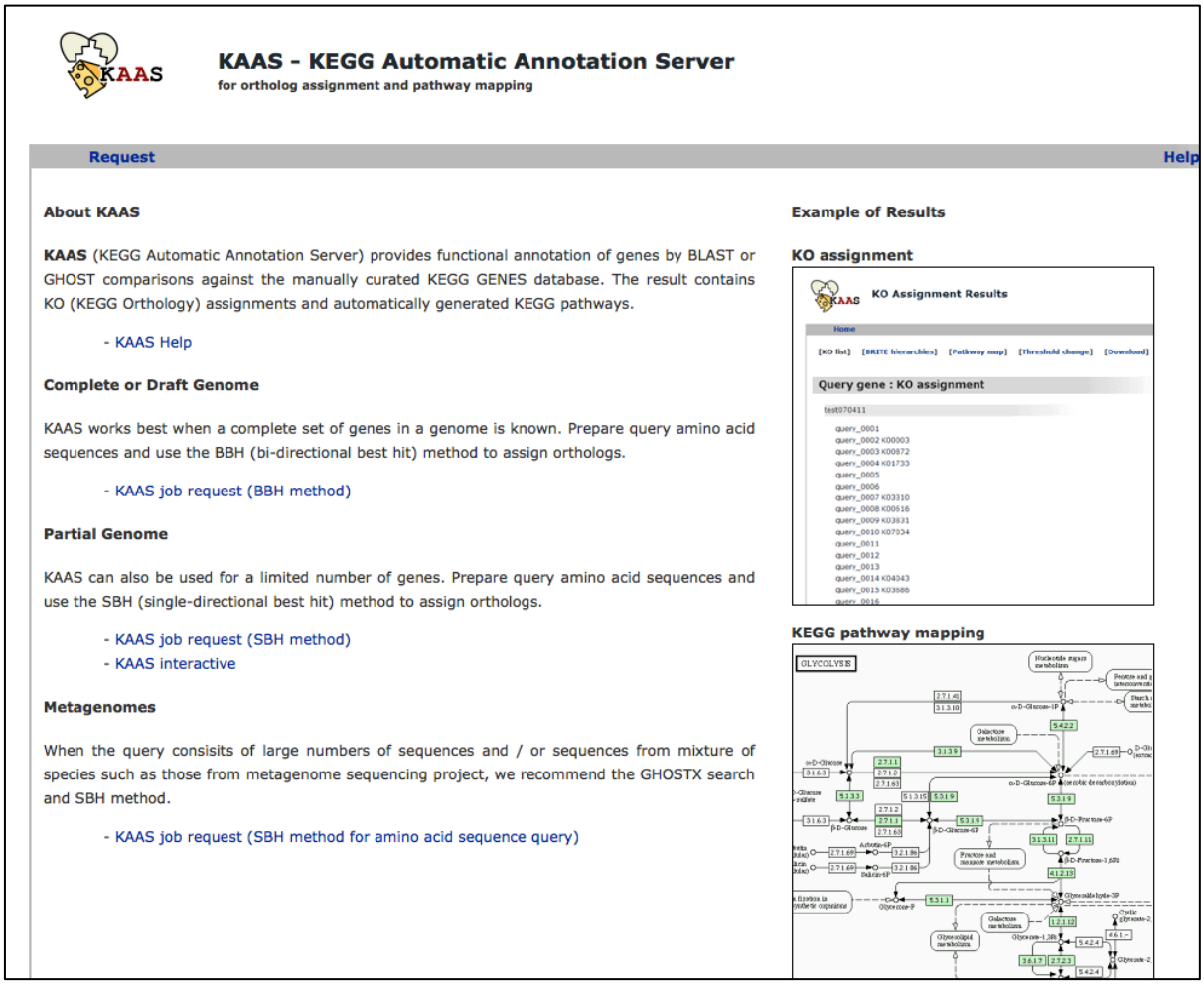


図 2.3 KAAS ウェブサービス

第３章　未知代謝パスウェイの予測

３．１　背景

前章における KAAS の開発により配列相同性を基に KEGG PATHWAY データベースにマッピングを行うことで、新規にゲノム配列の解読された生物の持つパスウェイを推定することが可能になった。しかしながら KEGG を含む既存のパスウェイデータベースでは、生物の持つ全てのパスウェイを表現するには未だ不十分であり、代謝パスウェイに限定したとしても、環境物質の生体内における分解経路や、植物における二次代謝産物の合成経路などで知識の蓄積が遅れている。さらには生物界全体の全ての分子間相互作用の同定を行うことは、物量的に今後とも困難が続くと考えられる。この問題への対処の一つとして、酵素反応における代謝物の化学的構造変化の特徴を蓄積し、それらの特徴を利用して中間代謝化合物を予測することで代謝パスウェイを推定する手法が挙げられる。UM-BBD（University of Minnesota Biocatalysis/Biodegradation Database）は化合物の生体異物の生体内分解経路に主軸を置いた酵素反応データベースで、反応による化合物の構造変化の特徴を抽出し、反応規則として収集している [45]。この規則を基に、UM-PPS（Pathway Prediction System）では基質の構造から反応による構造変化を推定し、代謝経路を予測することが可能になっている [46]。しかしながら 2009 年当時のシステムでは一段階及び二段階の反応経路予測に留まっており、さらに多段階の反応予測を行うには、予測された複数生成物の中から次の基質となるものを選択する必要があるため、網羅的な反応経路を予測するのは困難であった。

一方、KEGG データベースでは IUBMB（International Union of Biochemistry and Molecular Biology）Enzyme Nomenclature[47] 及び KEGG PATHWAY データベースに登録されている酵素反応を蓄積した KEGG REACTION データベースを構築しており、生体内分解経路に限らない多くの種類の酵素反応がデータベース化されている。また、酵素反応における基質と生成物のペア（リアクタントペア）を収集した KEGG RPAIR データベースも構築しており、このデータベースでは基質と生成物の間で起こった生化学的な構造変化の特徴が、反応中心周辺の変化を基に定義した RDM（Reaction center, Difference atom, Matched atom）パターンとして記述されている（図 3.1 A）。このため基質の構造と RDM パターンだけから生成物の構造を、または生成物の構造と RDM パターンから基質の構造を推定することが可能となっている（図 3.1 B）。また反応の未知な化合物に対しても、適用可能な RDM パターンを当て

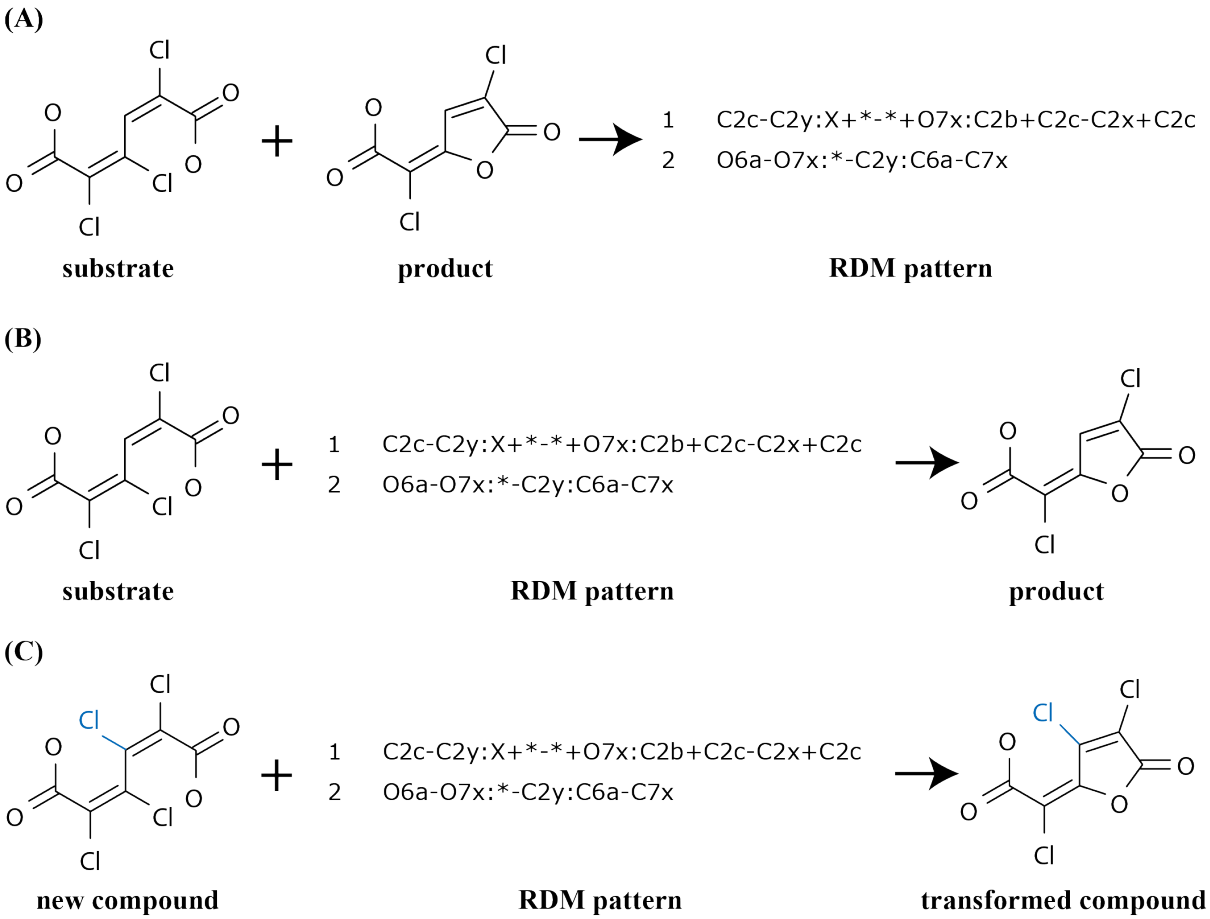


図 3.1 RDM パターンを用いた反応予測

（A）基質と生成物との間の構造変化の特徴は RDM パターンとして抽出される。(B) 基質の構造と RDM パターンから生成物の構造を推定することが可能。(C) 新規化合物に RDM パターンを適用することで、反応後の構造を類推できる。

はめることで、起こりうる反応を推定できる（図 3.1 C)。本研究では、この RDM パターンを反応規則とし、多段階の反応経路を可能な限り多く推定する手法の開発を目的とし、KEGG PATHWAY データベースで特に整備の遅れている、微生物における生体異物の生体内分解経路及び植物における二次代謝産物の合成経路の代謝経路予測に焦点を絞って開発を行った。この手法では、KEGG をリファレンスデータベースとすることで、予測結果と配列情報を結びつけることが可能であり、他の反応予測システムにはない特徴である。さらに、本手法を実装したウェブサービスを構築し、誰でも利用可能な環境の構築を目的とした。

３．２　材料と手法

リファレンスデータベース

2009 年 12 月の段階で、KEGG COMPOUND データベースには 16,110 化合物が蓄積されており、KEGG RPAIR データベースにはリアクタントペアが 12,032 ペア蓄積されている。KEGG では化合物の各原子は、炭素、窒素、酸素、硫黄、リン酸、ハロゲン原子、その他の原子に分類され、さらに周辺の結合情報により最大 3 文字の記号で表した 68 種類の"KEGG atom type"で記述されている（付録 2)。表 3.1 は炭素における atom type を示している。1 文字目は原子の種類を示しており、atom species と呼称している。2 文字目は主に機能分類と多重結合の種類を反映しており、前方から二文字目までの記号を atom class と呼称している。3 文字目は主に結合する原子の数と環状構造に含まれる原子かどうかを示している。

また、リアクタントペアは反応内で担う役割によって次の五つに分類されている。

- main pair：KEGG PATHWAY マップに表示されるような反応における主たる化合物のペア
- cofac pair：酸化還元反応における cofactor 化合物のペア
- trans pair：転移反応に於ける官能基転移を行っている化合物のペア
- ligase pair：リガーゼによる加水分解におけるヌクレオチドのペア
- leave pair：炭素などの無機化合物の分離や結合を表現する化合物のペア

本研究では main pair タイプのリアクタントペアのみに着目し、生体異物の生体内分解経路において 724 化合物からなる 853 のリアクタントペア、植物の二次代謝経路において 993 化合物からなる 1,126 リアクタントペアを抽出し、代謝パスウェイ予測のためのリファレンスデータベースとした。

表 3.1 炭素原子における KEGG atom type の例

Functional group	Atom species	Atom class	Atom type	Description
Alkane	C	C1	C1a	R-CH3
			C1b	R-CH2-R
			C1c	R-CH(-R)-R
			C1d	R-C(-R)2-R
Cyclic alkane			C1x	ring-CH2-ring
			C1y	ring-CH(-R)-ring
			C1z	ring-C(-R)2-ring
Alkene		C2	C2a	R=CH2
			C2b	R=CH-R
			C2c	R=C(-R)2
Cyclic alkene			C2x	ring-CH=ring
			C2y	ring-C(-R)=ring or ring-C(=R)-ring
Alkyne		C3	C3a	R≡CH
			C3b	R≡C-R
Aldehyde		C4	C4a	R-CH=O
Ketone		C5	C5a	R-C(=O)-R
Cyclic ketone			C5x	ring-C(=O)-ring
Carboxylic acid		C6	C6a	R-C(=O)-OH
Carboxylic ester		C7	C7a	R-C(=O)-O-R
			C7x	ring-C(=O)-O-ring
Aromatic ring		C8	C8x	ring-CH=ring
			C8y	ring-C(-R)=ring
Undefined C		C0	C0	

RDM パターン

KEGG RPAIR データベースでは基質と生成物上の原子間のアラインメント情報も蓄積しており、反応前後の生化学的な構造変化の特徴を、RDM パターンとして表現している。RDM パターンは反応中心（Reaction center）となる R 原子、基質と生成物で変化が生じた D 原子（Difference atom）、基質と生成物で変化の生じない M 原子（Matched atom）で構築されている（図 3.2）[29,48]。R 原子は次の 3 つの定義により決定される。1）原子の結合に変化が生じた場合。2）原子の酸化数に変化が生じた場合。3）シス−トランス異性化が生じた場合。また、D 原子、M 原子は R 原子に隣接するものだけを定義している。

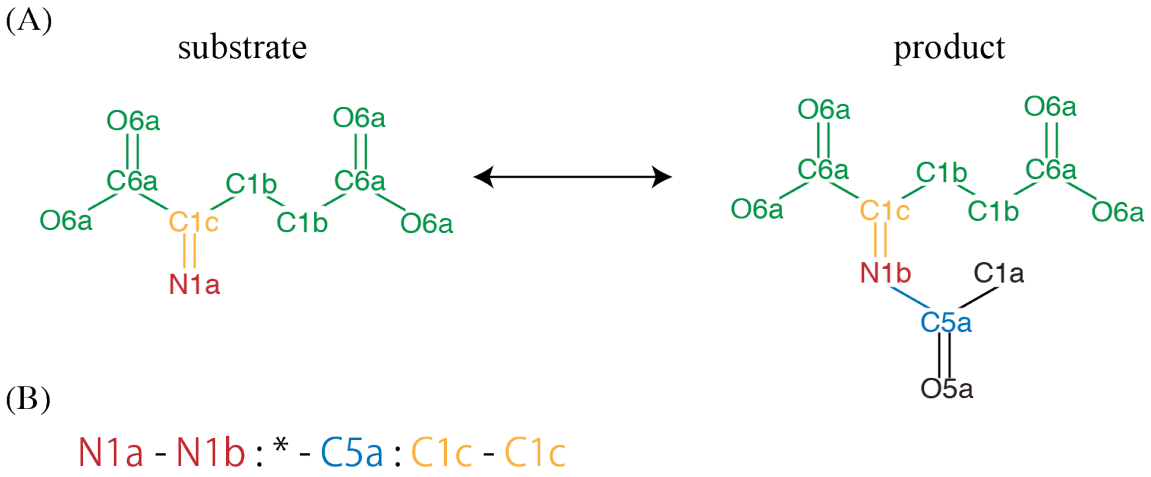


図 3.2 RDM パターン

（A）は KEGG atom type で表記したリアクタントペアの例を示しており、各原子の色は R 原子（赤）、D 原子（青）、M 原子（黄）、その他のリアクタントペア間で異なっている原子（黒）、その他の一致している原子（緑）を示している。（B）は R、D、M 原子から定義される RDM パターンの記述形式を示している。

アルゴリズム

RDM パターンは反応中心の少数の原子で構成されている。そのため、ある化合物に適応可能な RDM パターンは多く存在し、全ての可能性を想定しつつ多段階の反応経路を予測した場合、実際には起こり得ないであろうと思われる経路を含んだ、非常に多くの反応経路が構築されることになる。これを避けるため、全体構造が類似した

化合物同士において同じ反応が起こる可能性が高いという仮定のもと、手法を構築した。

図 3.3 は手法の全体像を示している。まず、第 1 ステップにおいて、代謝経路を予測する化合物 Q を問い合わせ化合物とし、リファレンスデータベースに対し類似化合物の検索を行った。類似化合物の検索には化合物間の最大共通部分構造探索を行う SIMCOMP プログラムを用いた[49,50]。SIMCOMP では、atom type レベルでの共通部分構造探索の後に、atom species レベルで共通構造を伸長するモードを使用した。次に第 2 ステップとして、第 1 ステップで得られた類似化合物 C_i を基質とするリアクタントペアが持つ RDM パターンに対し、化合物 Q を問い合わせ化合物とした部分構造検索を行った。ここでは化合物 C_i における R 原子、D 原子、M 原子と対応する原子が、それぞれ化合物 Q に存在する場合に検出した。表 3.2 は原子の種類ごとの対応条件を示している。対応例の括弧（ [ ] ）は、括弧の中のいずれかの文字であれば対応することを示しており、アスタリスク（ * ）はどんな文字、数字でも対応することを示している。R 原子の対応条件は"atom class の一致"と"環状構造の一部であるかどう

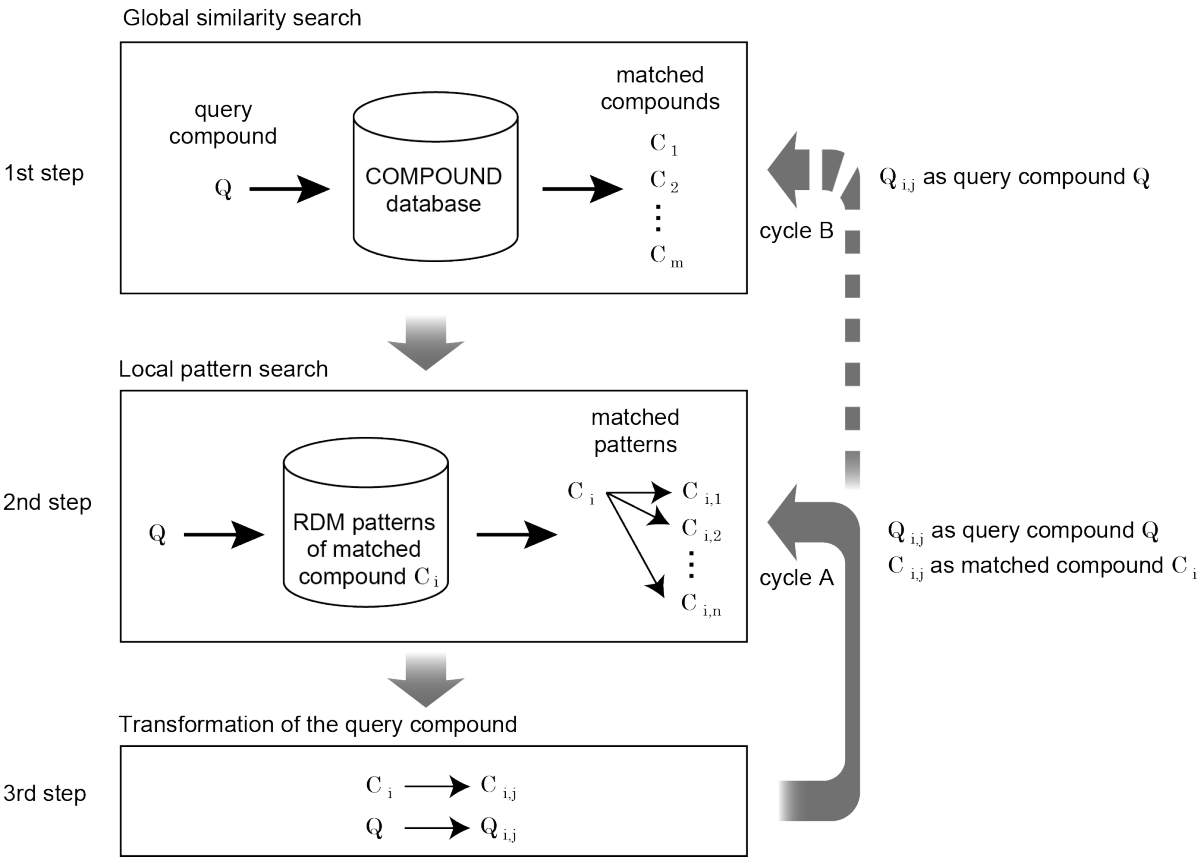


図 3.3 手法の全体像[51]

かの一致"が必要となっており最も厳しく、続いて atom class のみの一致を必要とする D 原子、atom species の一致が必要な M 原子の順に、条件が緩くなるように設定した。また、原子の機能が類似していると考えられる芳香環における炭素と窒素、構造骨格中における炭素と硫黄を、例外として M 原子の対応条件に加えた。

表 3.2 原子の種類による対応条件

原子	対応条件		対応例（C_i の原子: Q の原子）
R 原子	Atom class の一致	直鎖上	C2[abcd] : C2[abcd]
		環上	C8[xyz] : C8[xyz]
D 原子	Atom class の一致		C2* : C2*
M 原子	Atom species の一致		C** : C**
	例外		C8[xy] : N5[xy]
			C1x : S2x
			C1b : S2a

続いて第 3 ステップとして、一致した RDM パターンを持つリアクタントペア C_i →$C_{i,j}$ をリファレンスとして、問い合わせ化合物 Q から新たに中間代謝化合物 $Q_{i,j}$ を構築することで、代謝反応の予測を行った。ここで $Q_{i,j}$ 及び $C_{i,j}$ を第 2 ステップにおける Q 及び C_i とし、第 2 ステップ及び第 3 ステップを繰り返すことで、連続した代謝反応、代謝パスウェイの予測を行った（サイクル A)。また、第 2 ステップで一致する RDM パターンが得られなくなった場合に、$Q_{i,j}$ を第 1 ステップの Q とし、第 1 ステップのリファレンスデータベースに対し類似化合物の検索を行った(サイクル B)。この 2 つのサイクルを繰り返すことで連続した複数の経路からなる代謝パスウェイの予測を行った。反応予測サイクルは、KEGG PATHWAY データベースに出現する既知の化合物または指定した化合物にたどり着く場合に停止する。また、第 1 ステップ及び第 2 ステップで検索結果が得られない場合にも停止する。

スコアリング

算出された複数の経路の妥当性を評価するため次の 2 種類のスコアを定義した。一つは予測反応を評価するスコアで、予測された各反応において、化合物 Q と参照元となった化合物 C_i との間の atom species レベルの原子アラインメントから Jaccard 係数を算出し、反応スコアとして用いた。その際、反応中心の一致がより重要であると

考えられるため、RDM 原子における一致を、atom type での一致、atom class での一致、atom species での一致の 3 つに分けて考えることで、RDM 原子の一致の重みがその他の原子の 3 倍になるように調整した。もう一つは予測経路を評価するスコアで、予測された経路を構成する各反応における反応スコアの平均を、経路スコアとした。スコアの算出は予測の毎サイクルで行い、経路スコアの高いパスウェイを優先して次のサイクルを行うことで、予測経路の妥当性を保ちつつ予測される経路の数の発散を防いだ。

双方向予測

予測を行う代謝経路の始点と終点の化合物が既知である場合、予測サイクルを両端から双方向で行った。両方の予測においてサイクル B に入る前に、双方向それぞれで生成された中間代謝化合物間で SIMCOMP を用いた構造類似性の計測を行い、その結果を考慮することで、双方の予測経路が連結する可能性の高い経路を優先して次のサイクルに回すことで、予測の妥当性の保持と経路の数の発散を防いだ。

３．３　結果と考察

提案する手法の評価を行うため、まず単一の反応予測の Leave-one-out 交差検証を行った。リファレンスデータベースから一つのリアクタントペアを取り除き、基質または生成物を問い合わせ化合物としてペアの対となる化合物の予測を行った。その際、リファレンスデータベース中に一度しか現れない RDM パターンを持ったリアクタントペアは、取り除いた場合の予測が不可能であるため、これらを除いた植物の二次代謝産物合成経路の 872 ペア、微生物の生体異物分解経路での 477 ペアを用いて検証を行った。その結果二次代謝産物合成経路では 71.9%のペアで基質の構造を予測でき、生体異物分解経路では 80.6%のペアで生成物の構造を予測できた。本手法では適用可能な全ての RDM パターンを適用するのではなく、構造類似性の高い化合物に由来する RDM パターンを優先して適用するため、基質や生成物を予測できないものもあった。また、二次代謝産物の多くは生体異物分解経路に含まれる化合物と比較して大きく、反応による構造変化が複雑なものも多いため、正確な RDM 原子のアラインメントが困難になり予測精度が低くなったと考えられる。

次に例として、代謝経路が知られており、KEGG には登録されていない連続した反応の予測を行った。図 3.4 A は 1,2,3,4-テトラクロロベンゼン（Query）を始点としグリコール酸（C00160）を終点とした、生体内分解パスウェイを予測した結果を示している。グリコール酸は炭素鎖の分解産物として比較的構造が単純な化合物であるた

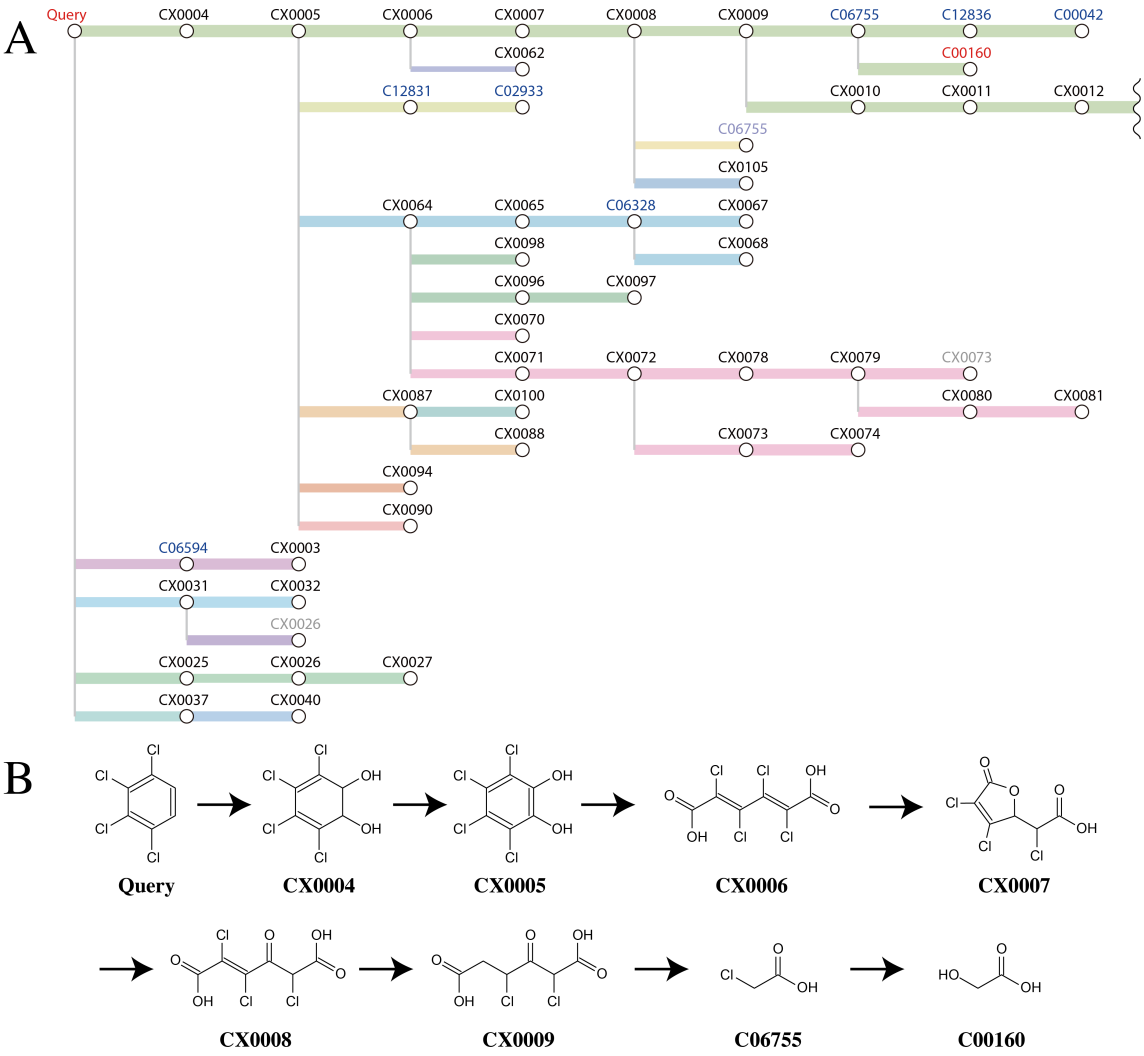


図 3.4 テトラクロロベンゼン分解経路の予測 [51]

A は複数の予測されたパスウェイをツリー状に示したものである。丸印は化合物を示しており、青文字で示された C 番号は KEGG PATHWAY データベースに出現する化合物を、黒文字で示された CX 番号は予測によって生成された中間代謝化合物を示している。また、色の薄くなっている C、CX 番号は、経路の中に同一の化合物が存在することを意味している。線は予測された反応を示しており、線の太さは反応スコアを反映している。また連続した同一の色で示されている反応は、予測においてサイクル B を経由せず、サイクル A の繰り返しによって予測された反応を示しており、KEGG PATHWAY において連続している反応をリファレンスとして予測されたことを意味している。B は始点から終点まで予測された経路における化合物の構造を示している。

め、終点化合物として選択した。1,2,3,4-テトラクロロベンゼンの分解経路は UM-BBD データベースに登録されており、図 3.4 B で示された始点から CX0009 までと一致する。これは図 3.4 A の一番上に見られる緑の線に対応し、このことから、本手法は正しい分解経路を高い妥当性を持って予測できたといえる。その他の経路として、Query から 1,2,4-トリクロロベンゼン（C06594）へ、または CX0005 から 3,4,6-トリクロロカテコール（C12831）へと、芳香環から塩素が取れる経路の可能性も提示している。

2 つ目の例として、図 3.5 A はデルフィニジン（C05908）に 3 個のグルコース（図 3.5 B の青丸）と 2 個のカフェオイル基（同赤丸）が結合した構造を持つ、ゲンチオデルフィン（Query、図 3.5 B）の生合成経路を予測した結果を示している。KEGG において既知の経路を含む、グルコース及び、カフェオイル基の結合の順番の異なる、9 つの経路が提示された。しかしながら、幾つかの経路においてはグルコースとカフェオイル基が同時に結合している。これは本手法が反応中心に重点を置いているため、結合の起こりやすい分子単位を考慮していないためだと考えられる。また、カフェオイル基の結合に関連するリアクタントペアは主に trans pair に分類されるため、main pair のみを用いたリファレンスデータベースでは、予測が正確に行えなかったと考えられる。そのため、より反応予測に影響のあるリアクタントペアを選択する必要がある。

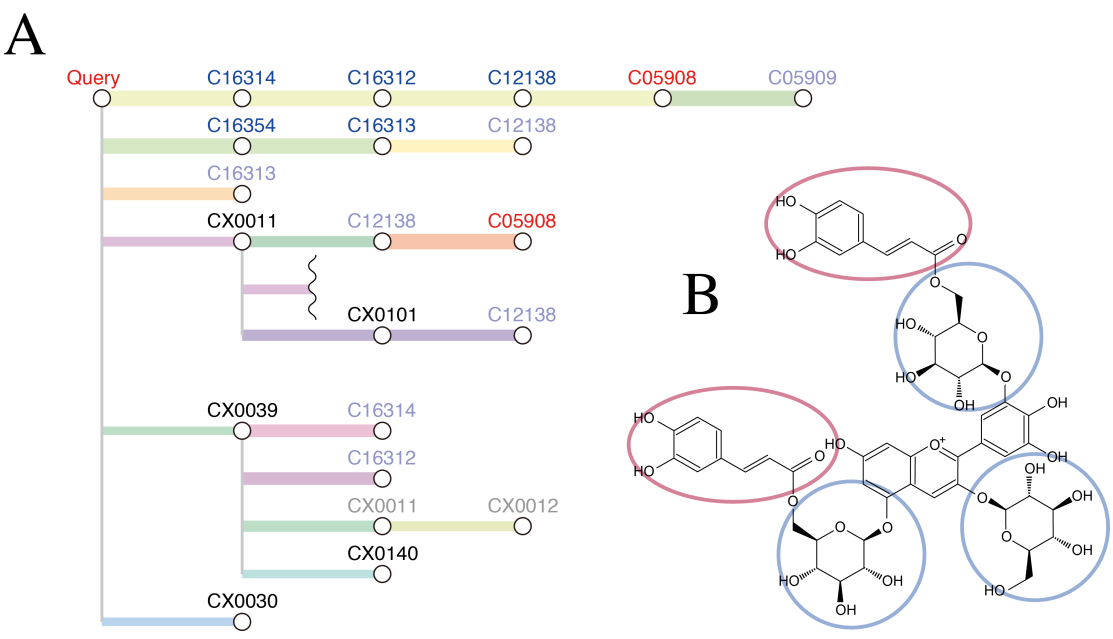


図 3.5 ゲンチオデルフィン生合成経路の予測 [51]

3 つ目の例として、図 3.6 はジベンゾチオフェンの分解経路を予測した結果を示している。赤の矢印で示した経路は UM-BBD データベースに登録されている経路を示しており、本手法で予測された複数の経路で最も正解に近い経路を黒の矢印で示した。前半の反応は正しく予測できているが、後半の反応では誤った経路が予測されている。これは本手法における異性体の扱いに起因すると考えられる。CX0004 と CX0005 は光学異性体であり、CX0003 と CX0005 は互変異性体である。KEGG atom type を用いた化合物の構造表記では光学異性体を識別できず、また互変異性体は全く別の異なる化合物として扱われる。そのため、これら異性体の絡む反応を予測することが困難になっていると考えられる。

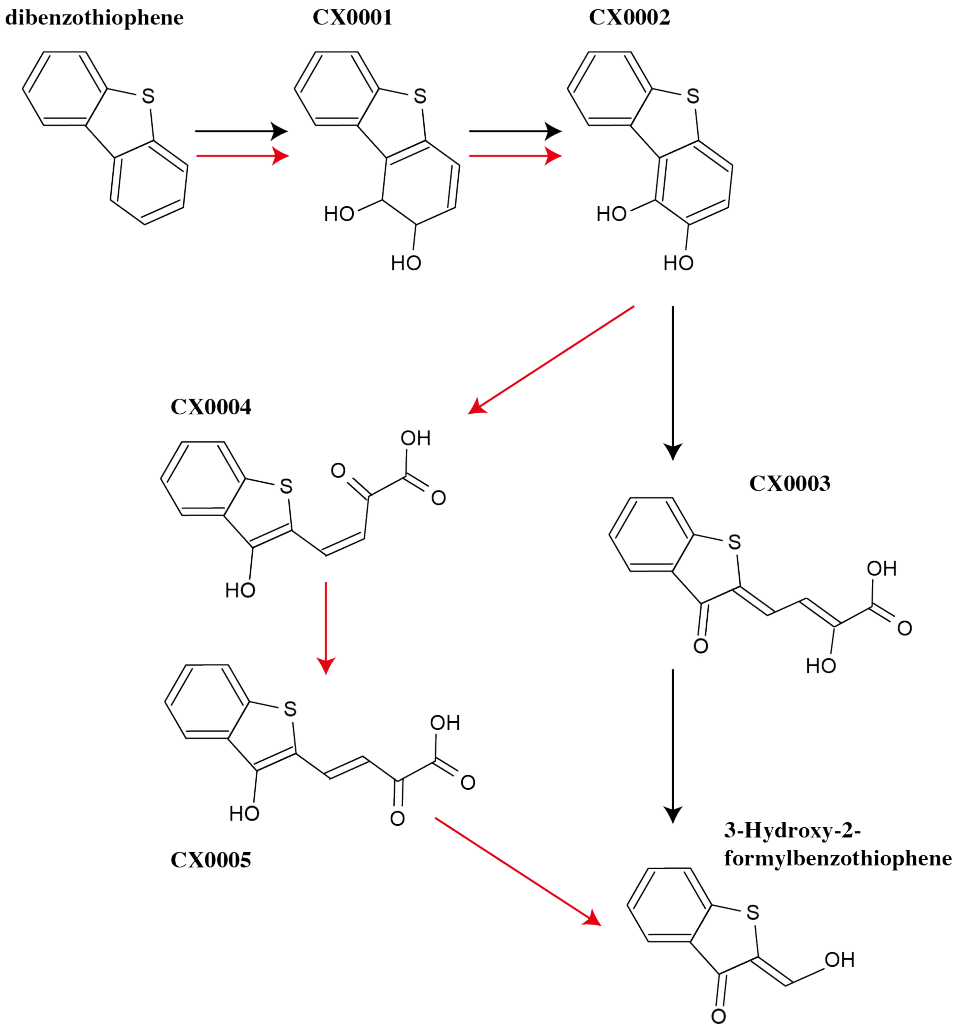


図 3.6 ジベンゾチオフェン分解経路の予測

本研究で提案した手法は PathPred（http://www.genome.jp/tools/pathpred/）として実装、公開されており、誰でも利用可能となっている（図 3.7）[51]。微生物における生体異物の生体内分解経路においては生体異物を、また植物の二次代謝産物の合成経路においては二次代謝産物を起点に、代謝経路の予測を行うことができる。化合物の構造情報は KEGG 化合物 ID（C 番号）及び医薬品 ID（D 番号）MDL MOL 形式または SMILES 形式を受け付けている。

図 3.7 PathPred ウェブサービス

第４章　基質と生成物の分子構造情報を用いた代謝酵素の予測

４．１　背景

2014 年現在、KEGG PATHWAY データベースの代謝パスウェイには約 6,000 の代謝化合物と約 6,200 の代謝反応が登録されている。また酵素反応データベース IUBMB[47] にも同様に約 6,000 の酵素反応が登録されている。しかしながらヒトの体内で見つかった小分子を蓄積したデータベースである HMDB（The Human Metabolome Database）には、すでに 40,000 を超える代謝化合物が登録されており[52]、また二次代謝産物を多く生産する植物界全体における代謝化合物は 100 万を超えると推定されている[53]。そのため自然界にはこれら代謝化合物と同程度のオーダーの酵素反応が存在すると考えられ、パスウェイデータベースはそのほんの一部を蓄積しているに過ぎない。

化合物の代謝反応経路を予測するウェブサービスとして UM-PPS[46] や前章の PathPred[51] などがあり、また機械学習の教師あり学習を用いた予測手法も開発されている[54]。これらの予測手法では、代謝パスウェイに於ける化合物の既知の化学構造変化を基に、自然産物や環境物質、医薬品などの生体内における代謝の予測を行うことができるが、化合物の構造変化を予測することを主目的としており、反応を触媒する酵素遺伝子の予測は行っていない。このような酵素遺伝子配列の同定されていない反応はオルファン酵素反応と呼ばれており[55]、過去 10 年で新たに報告された酵素反応の約 40%は未だ遺伝子配列の決定していないオルファン酵素反応であるという報告がある[30]。そのため、オルファン酵素の遺伝子を推定するための手法が多く開発されており、注目するパスウェイに関わる遺伝子群のゲノムコンテキスト、系統プロファイル、発現プロファイルを用いた予測が行われている[30,56–58]。他にも、統計的アルゴリズムであるベイズモデルや[59]、教師付き学習などを用いた酵素予測法の開発も行われているが[60]、これらの手法にはオルファン酵素反応と関連する遺伝子群の情報や、ゲノムの情報が不可欠である。

一方、反応前後の基質と生成物の分子構造の変化のみに着目し、EC サブ-サブクラス（EC 番号の前方３桁）を予測するツールの一つに E-zyme がある[31]。E-zyme は KEGG RPAIR データベースに蓄積された、反応前後の生化学的な構造変化の特徴を抽出した RDM パターンを基に、構造変化パターンの類似度を計算することでその反

応の EC 番号の予測を行う。これは、同一の EC サブ-サブクラスを持つ酵素反応同士の構造変化パターンは類似している傾向がある、という事実に基づいた手法となっている。しかしながら予測される EC サブ-サブクラスは酵素の基質特異性を表現しておらず、非常に多くの酵素反応が含まれているため、EC サブ-サブクラスからの酵素遺伝子の同定は容易ではない。また、化学構造から類似反応を探索する手法は多く開発されており、結合の変化や基質と生成物間の差異をプロファイル（フィンガープリント）で表現し、類似度を計算するものや[61–64]、機械学習を用いて EC 番号を予測する手法も開発されている[65–68]。しかしながら、これらの手法で酵素遺伝子配列そのものを予測しているものはまだない。

本研究では、E-zyme の手法を拡張し、オルファン酵素反応の遺伝子を予測することを目的とした。基質と生成物のペア（リアクタントペア）の部分情報である RDM パターンのみを用いるのではなく、リアクタントペア全体の構造比較を行うことで、オルファン酵素反応と類似した反応をデータベースから探索し、これを触媒する酵素遺伝子、またそのパラログ及びオーソログをオルファン酵素候補として抽出した。

４．２　材料

リファレンス反応データベース

酵素反応は KEGG REACTION データベース（リリース 67.0+）に登録されている 9,398 反応を用いた。また、リアクタントペアを蓄積した KEGG RPAIR データベースには 14,218 ペアが登録されており、このうち main pair（３．１参照）の 8,846 ペアをリファレンスデータとして用いた。リアクタントペアは 2,831 の RDM パターンと対応している。

オーソログデータベース

KEGG リリース 67.0+に含まれる KEGG Orthology (KO)には、450 万遺伝子から構成される 16,791 グループが登録されており、酵素反応の代謝能と関わりのあるグループは 3,868 を数える。このオーソログデータベースは文献情報等を基に手作業により作成される。また、オーソログデータを拡張するために KEGG Ortholog Cluster (OC)データベース（ver. 2014-04-28）[69] を用いた。OC は KEGG GENES に登録される 3,000 生物種の 1,200 万遺伝子全てを、配列類似性に基づいて自動的にクラスタリングしたデータベースであり、より多くのパラログ及びオーソログがグループ化されている。本研究では KO グループを基にオルファン酵素遺伝子の予測を行うと共に、OC グループを用いた候補遺伝子の探索を行った。

４．３　手法

アルゴリズム

図 4.1 は予測手法の三つのステップを示したフローチャートである。第一ステップでは、酵素遺伝子の予測を行う反応の基質と生成物（問い合わせペア）の間で構造アラインメントを計算した。構造アラインメントには SIMCOMP プログラム [50] を用いた。また、構造アラインメントから反応中心を検出し、RDM パターンを作成した。第二ステップでは、問い合わせペアから作成した RDM パターンと KEGG RPAIR デ

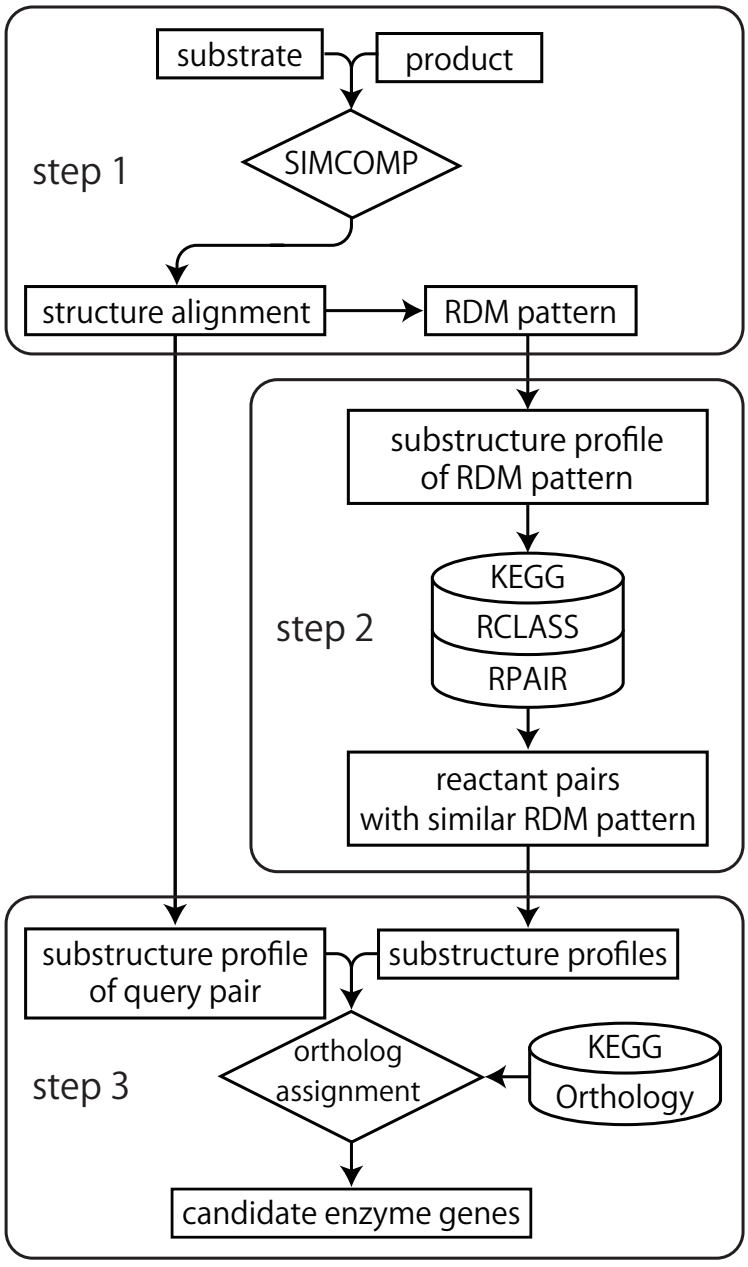


図 4.1 手法のフローチャート [76]

ータベースに蓄積された RDM パターンの間で類似度を計算し、類似度が閾値に満たない RDM パターンの足切りをおこなった（類似度計算の詳細は後述）。残った RDM パターンを持つリアクタントペアを KEGG RPAIR データベースから抽出した。第三ステップでは、問い合わせペアと第二ステップで収集したリアクタントペアの間で類似度を計算し、KO を用いたオーソロググループ推定及び、OC を用いた候補遺伝子探索を行った。第二、第三ステップに於ける類似度計算では、RDM パターン、リアクタントペアをそれぞれ整数ベクトルで表現される構造プロファイルに変換し（図 4.2 D）、プロファイル間の Tanimoto 係数を RDM パターン間及びリアクタントペア間の類似度として用いた。Tanimoto 係数は Jaccard 指数を実数ベクトルでも扱えるように拡張した係数となっており、2 つのプロファイル$X = [x_1, x_2, \ldots, x_i]$、$Y = [y_1, y_2, \ldots, y_i]$の Tanomoto 係数は次の式で表される。

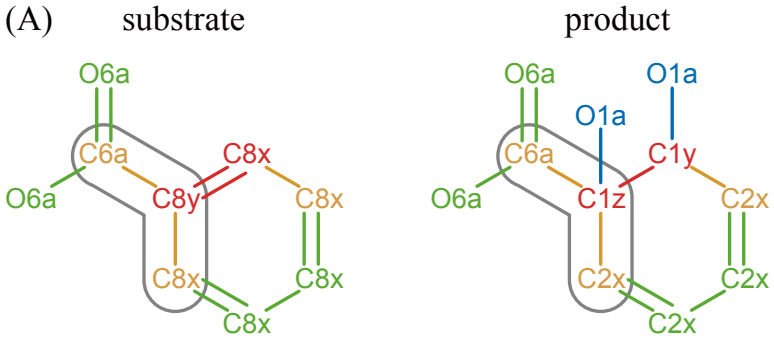
$$\text{Tanimoto 係数} = \frac{\sum x_k y_k}{\sum x_k^2 + \sum y_k^2 - \sum x_k y_k}$$

部分構造プロファイルの構築

第三ステップにおけるリアクタントペア間の類似度を定義するためには、例えば KCF-S[70]や Daylight フィンガープリント[71]のように化合物を部分構造のコレクションで記述する手法が既に発表されている。これと同様に、それぞれのリアクタントペアを、部分構造の出現頻度で表現したプロファイルに変換した。また、それぞれの部分構造は反応の前後で対応する基質と生成物の原子を含むよう設計した。図 4.2 は部分構造プロファイルの例を示している。図 4.2 A の基質と生成物上に灰色の線で囲んだ、連続して結合している 3 原子は、各原子が反応の前後で対応しており、各原子を合わせた 6 原子をアラインメント部分構造として抽出した。各アライメント部分構造は KEGG atom type と KEGG atom class それぞれを用いて表現した。atom type を用いた場合には連続して結合する全ての 2 から 3 原子の鎖を部分構造として抽出し、一方 atom class を用いた場合には R 原子または D 原子を含む、2 から 8 原子の鎖を部分構造として抽出した。また atom class を用いた部分構造では、その原子の生化学的特徴を考慮し、芳香環上の原子を"A"という表記で統一し、それ以外の環上の原子を"R"という表記で統一した。これにより、図 4.2 C に示した例では atom type を用いた場合には"C6a-C6a:C8y-C1z:C8x-C2x"と表現し、atom class を用いた場合には"C6-C6:A-R:A-R"と表現することになった。次に各アラインメント部分構造の出現頻度を、リアクタントペアの部分構造プロファイルと定義し、問い合わせペアの部分構

造プロファイルとリファレンスペアの部分構造プロファイルの間の Tanimoto 係数を、リアクタントペア間の類似スコアとして定義した。また、Tanimono 係数の算出にはプロファイルの要素（アラインメント部分構造）の数を揃える必要があるため、一方のリアクタンペアにしか存在しないアラインメント部分構造は、他方では頻度 0 の要素として、部分構造プロファイルに追加した。問い合わせペアとリアクタントペア間において、類似スコアが最も高いスコアの 3 割に満たないリアクタントペアは、類似性が低いものとして除外した。

(A)



(B)

C8y-C1z:*-O1a:C6a+C8x+C8x-C6a+C1y+C2x
C8x-C1y:*-O1a:C8x+C8y-C2x+C1z

(C)

atom type	C6a-C6a:C8y-C1z:C8x-C2x
atom class	C6-C6:A-R:A-R
atom species	C-C:C-C:C-C

(D)

Aligned substructure	Frequency
C6a-C6a:O6a-O6a	2
C8y-C1z:C6a-C6a:O6a-O6a	2
C6-C6:A-R:A-R	2
C6a-C6a:C8y-C1z:C8x-C1y	1
A-R:A-R	3
⋮	⋮

図 4.2 RDM パターンと部分構造プロファイル [76]

（A）基質と生成物の KEGG atom type での表記（B）RDM パターン（C）A における灰色で示した 3 原子のアラインメントから得られる部分構造(D)各部分構造の頻度を表した部分構造プロファイル。

第二ステップにおける RDM パターンもリアクタントペアと同様に、原子アラインメント情報を保持している。そのため、RDM パターンも同様に部分構造の出現頻度で表現したプロファイルに変換し、パターン間の類似性を定義した。RDM パターンのプロファイルは連続して結合する全ての 2 から 3 原子の鎖を KEGG atom class 及び KEGG atom species で表現した。図に示した例では atom class を用いた場合には"C6-C6:A-R:A-R"と表現し、atom species を用いた場合には"C-C:C-C:C-C"と表現した。プロファイル間の Tanimoto 係数を RDM パターン間の類似スコアと定義し、スコアが 0.3 に満たないものは類似性が低いものとして除外した。

KO を用いたオーソロググループの推定

残ったリアクタントペアを触媒する酵素遺伝子を含む KO グループから、最も可能性の高い KO グループを酵素遺伝子候補として予測を行うため、それぞれの KO グループ毎に予測スコアを定義し算出した。図 4.3 はスコア算出の例を模式的に示したものである。この例では、ある問い合わせペア(RPQ)に対して、類似性のあるリアクタントペアが 6 ペア(RP1~6)得られている。そのうち RP1 は KO1、KO2 の 2 つの KO グループに入っている酵素遺伝子（KO1 遺伝子）によって代謝され、また RP2~4 は KO1 遺伝子のみ、RP5,6 は KO2 遺伝子のみによって代謝が行われている。図の KO1 のテーブルは、KO1 遺伝子によって触媒されるリアクタントペアの構造プロファイルを並べたものであり、SUB1~5 が部分構造を、数値が出現頻度を表している。問い合わせペアと KO1 遺伝子に触媒される 4 つのリアクタントペアそれぞれの構造プロファイル間で Tanimoto 係数を算出し、最も高いスコアを KO1 グループの予測スコアとした。その際、リアクタントペアの中で出現頻度の"ばらつき"の小さな部分構造を"必須部分構造"と定義し、必須部分構造のみを用いて Tanimoto 係数を算出した。図 4.3 の KO1 では灰色で示した部分構造 SUB3 の出現頻度が RP1-4 で 1,2,0,3 とばらつきが大きくなっている。このため SUB3 を必須部分構造では無いものとしてスコア計算から除外した。一方、SUB5 は全てのリアクタントペアで出現していない。この場合はこの部分構造を持たないことが酵素の基質特異性に関与している可能性を考慮して、スコア計算に用いた。出現頻度のばらつきを基に必須部分構造を定義するため、次の式を満たす部分構造を必須部分構造として定義した。

$$\bar{x} - \sigma \times 2 \geq 0$$

ここで$\bar{x}$は各部分構造の出現頻度の平均を示しており、またσは標準偏差を示している。KO1、SUB3 の例では平均と標準偏差はそれぞれ 1.5 と 1.118 になり式を満たさないため、酵素の特異性に必須でない部分構造として、スコア計算から除外される。

また、KO2 の例ではリアクタントペアのリストが KO1 とは異なるため、必須部分構造も異なっている。そのため異なる KO グループにおける同一のリアクタントペア（RP1）と問い合わせペアの間の Tanimoto 係数も異なる場合がある。そのため、図の例の場合、KO1 グループの予測スコアは 0.61、KO2 グループの予測スコアは 0.93 となる。

図 4.3 スコア算出の模式図 [76]

OC を用いた候補遺伝子探索

KEGG GNESE データベースには KO グループに分類されていない遺伝子も数多く登録されており、酵素遺伝子候補をより広く探索するため、上記の手順により推定された KO グループに含まれる遺伝子を、少なくとも 1 つ含んでいるような OC グループを探索した。OC グループに含まれる全ての遺伝子を候補として抽出した。

４．４　結果と考察

交差検証

提案した予測手法の評価を行うため、少なくとも 2 つのリアクタントペアの触媒を

担うことが分かっている KO グループを用いて、対応する 3,357 リアクタントペアにおいて Leave-one-out 交差検証を行った。それぞれのリアクタントペアを酵素遺伝子が未知である問い合わせペアと仮定し、提案した手法によって予測スコアを算出し、最も高い予測スコアを持つ KO グループを酵素候補とした。図 4.4 は交差検証の結果を示している。横軸は酵素候補の予測率を示しており、左側の縦軸は予測された酵素候補の正答率を示している（青い線と対応）。また右側の縦軸はそれぞれの検証にいて予測された酵素候補の数の平均を示している（赤い線と対応）。ここで、少なくとも 1 つの正解酵素が予測された検証の回数を CA (correct-assign)、酵素候補が予測されたが不正解だった検証の回数を IA (incorrect-assign)、酵素が予測されなかった検証の回数を NA (no-assign)とした時、正答率 (correct assign rate) と予測率 (assign rate) を次のように定義した。

$$\text{正答率} = \frac{CA}{(CA+IA)}$$

$$\text{予測率} = \frac{(CA+IA)}{(CA+IA+NA)}$$

各プロットは、予測スコアを閾値にして予測結果の足切りを行った場合の結果をそれぞれ示している。この結果、予測スコアの閾値を 0.98 に設定した場合に、正答率が最高となり 0.8 を超えている。この場合の予測率は 0.3 程度と低くなっているが、リアクタントペアとして 1,033 ペアにおいて酵素候補を予測できている。さらに、閾値を 0.7 に設定した場合には正答率が 0.7 となり、2,401 のリアクタントペアにおいて酵素候補の予測が行われている。また、赤い線でしめされるように各検証において予測される酵素候補の数は 2 から 3 個であり、化合物の分子構造以外の情報が利用できない場合においても、酵素候補を絞り込むことが十分に可能であることが示された。知られている限り、これは遺伝子発現情報などの他の情報を一切用いず、化学構造の変化のみを用いて酵素遺伝子の予測を行う最初の研究である。

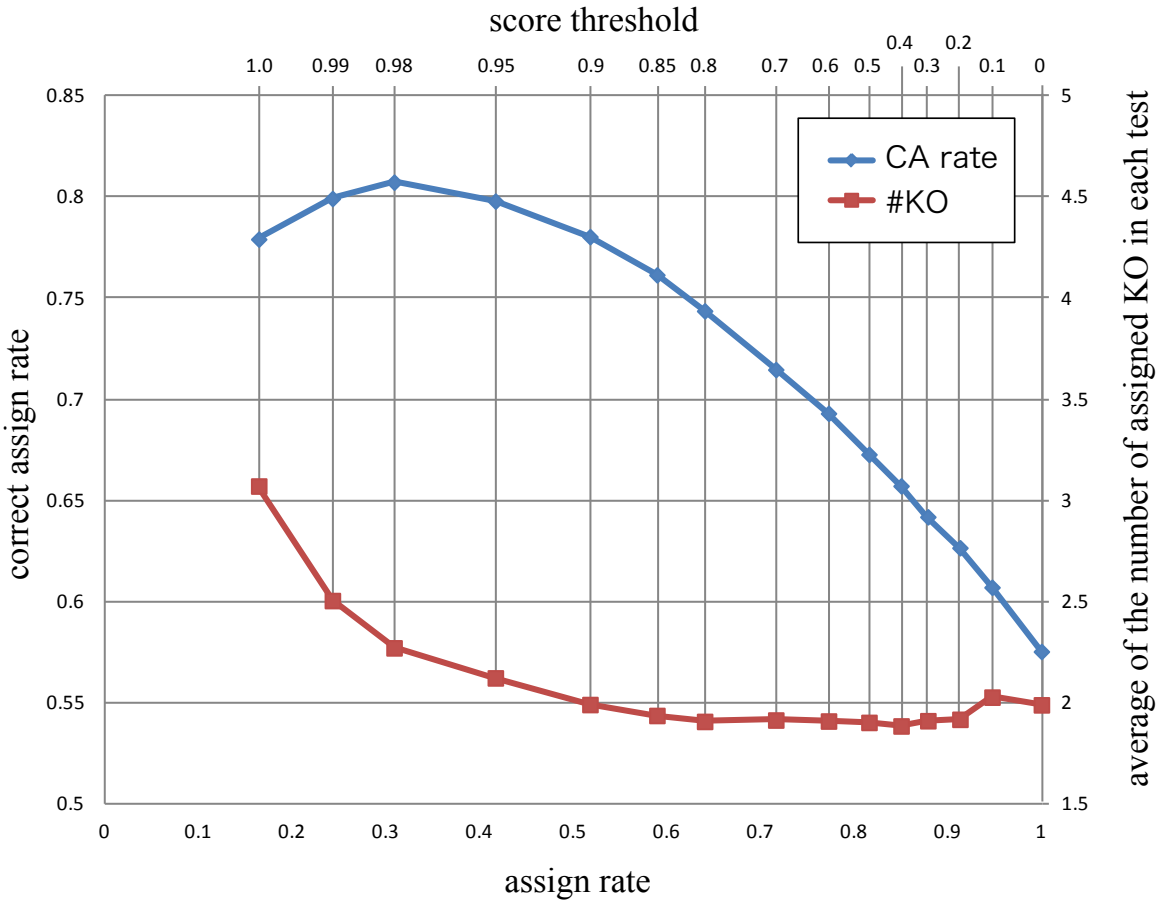


図 4.4 交差検証の結果 [76]

KEGG データベースにおける新規酵素遺伝子の予測

KEGG RPAIR データベースには 8,846 ペアの main pair タイプのリアクタントペアが登録されているが、そのうち 3,865 ペアは未だ KO グループの付与されていないオルファン酵素反応に含まれている。図 4.5 A はこれら酵素の割り当てられていないリアクタントペアで、予測スコアの閾値を 0.7 に設定し、酵素候補の予測を行った結果を示している。その結果、リアクタントペアの 40%以上（1,641 ペア）において酵素候補を初めて予測した。これはゲノムコンテキストを用いて酵素推定を行った先行研究 [56] よりも多くの酵素候補を提示できている。また各リアクタントペアには、平均して 2.08 個の KO グループが酵素候補として予測されている。一方、以前の E-zyme のシステムで予測できる EC サブ-サブクラスを経由して KO を付与した場合、各ペアに平均して 26.89 個の KO グループが関連付けられることになり、提案した手法では酵素候補の大規模な絞り込みが行えている。

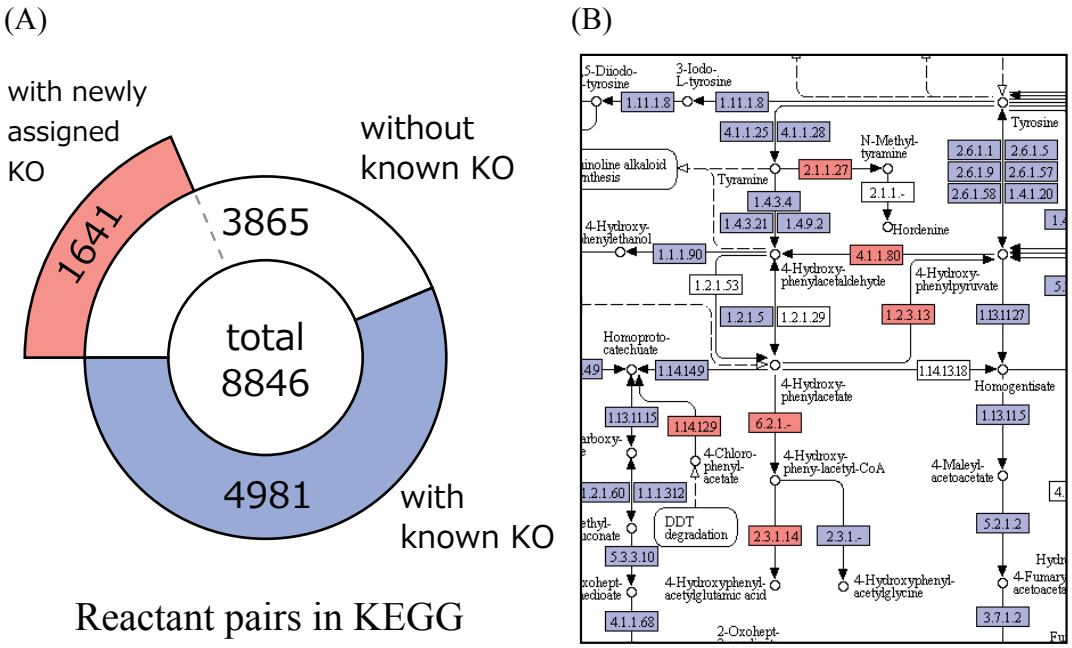


図 4.5 新規酵素遺伝子の予測結果 [76]

図 4.5 B はチロシン代謝パスウェイの一部を示している。酵素が既知である反応を青色で、今回新たに酵素候補が予測された反応を赤色で示しており、これまで繋がっていなかったパスウェイ上のミッシング経路を新たに埋めることができた。

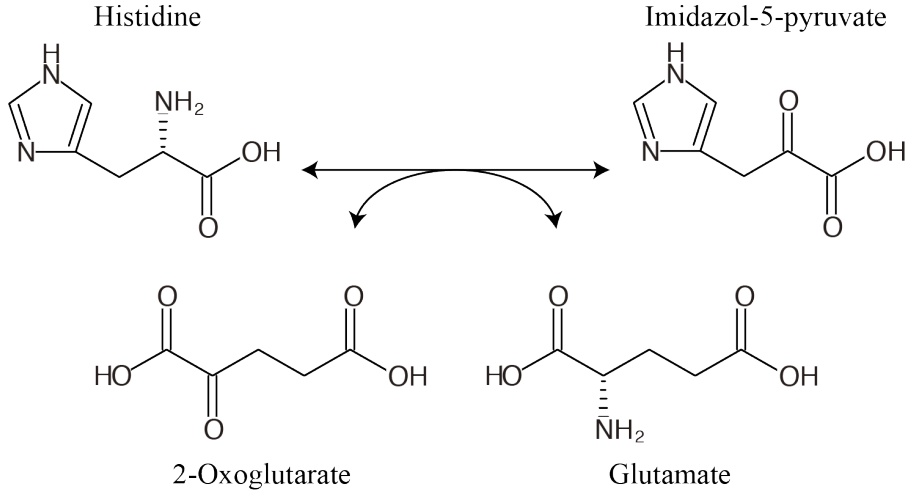
これらの予測結果を評価するため、２つのアミノ基転移反応と脱水素反応、及びメサコン酸パスェイにおいて文献に基づいた検証を行った。一つ目は Histidine transaminase 反応（EC 2.6.1.38）であり、これは Histidine から Glutamate へアミノ基を転移する反応にあたる（図 4.6 A)。Histidine と Histidine のアミノ基がカルボニル基に置き換わった Imidazol-5-pyruvate の問い合わせペアに対して、提案した手法では予測スコア 0.971 という高いスコアで KO グループ K00817 が予測されていた。この KO グループは、Phenylalanine、Tyrosine、Histidinol phosphate のアミノ基をグルタミン酸に転移する機能が知られている。Histidinol phosphate は Histidine がリン酸化された化合物であり、そのため構造もよく似ている。また発現解析によって、Histidine のアミノ基転移能が報告されている *Thermoanaerobacter tengcongensis* の遺伝子 tte:TTE2137 が K00817 に含まれており [30]、実験的に検証された遺伝子を正しく予測できている。

二つ目は Asparagine oxo-acid transaminase 反応（EC 2.6.1.14）で、これは Asparagine から Glutamate へアミノ基を転移する反応である（図 4.6 B)。*Streptococcus mutans* の遺伝子 smu:SMU_1312 がこの反応の触媒能を有していることが実験的に確かめられており [30]、また、ゲノム配列を報告したオリジナル配列リポ

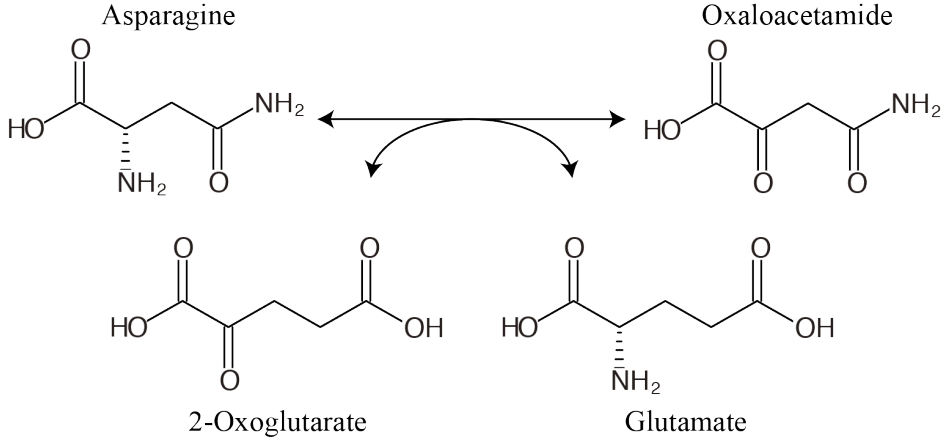
ジトリでは Asparagine transaminase の一つとして登録されている[72]。今回提案した手法では、Asparagine、Oxaloacetamide の問い合わせペアに対して予測スコア 0.581 で 10 個の KO グループが酵素候補として予測された。これら KO グループには *S. mutans* の遺伝子は登録されていなかったため、OC を用いて酵素候補の探索を行い、結果 203 の OC グループと対応が取れた。これらの OC グループにも実験的に触媒能が確かめられた遺伝子 smu:SMU_1312 は含まれて居なかったが、smu:SMU_1312 と最も高い配列類似性を有するパラログ遺伝子 smu:SMU_24 が含まれていた。

三つ目は 3,4-dehydroadipyl-CoA semialdehyde dehydrogenase 反応（EC 1.2.1.77）であり（図 4.6 C）、*Burkholderia xenovoran* の遺伝子 bxe:Bxe_A1420 が反応の触媒を担うことが実験で確かめられている[73]。3,4-Didehydroadipyl-CoA semialdehyde と 3,4-Didehydroadipyl-CoA の問い合わせペアに対して、予測スコア 0.732 で K02618 が酵素候補として予測された。K02618 もまた *B. xenovoran* の遺伝子を含んでいなかったが、この KO グループは実証された遺伝子 bxe:Bxe_A1420 を含む 2 つの OC グループと対応が取れた。

(A) Histidine transaminase



(B) Asparagine oxo-acid transaminase



(C) 3,4-Dehydroadipyl-CoA semialdehyde dehydrogenase

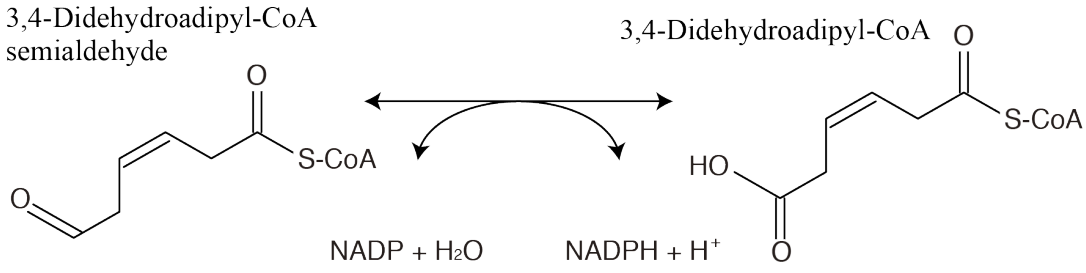


図 4.6 検証に用いた反応

次に、*Clostridium tetanomorphum* で同定されたメサコン酸パスウェイの 2 つの反応 [74] について検証を行った。このパスウェイは Glutamate を起点に Mesaconate を経由して Pyruvate が生じる 4 つの連続した反応からなり（図 4.7 A）、前半の 2 つの反応は酵素が同定されている。また、*C. tetanomorphum* 自体はゲノム配列が決定されていないため酵素遺伝子のゲノム上での配置は分からないが、近縁種である *C. tetani* のゲノム上で隣接して存在しているのが確認できる(図 4.7 B)。一方、第三、第四の反応の酵素は未だ知られていない。ここでは酵素候補の予測に *C. tetanomorphum* の近縁種である *C. tetani* のゲノム情報を組み合わせることで、予測結果の検証を行った。

第三の反応は(S)-2-methylmalate hydrolyase 反応（EC 4.2.1.34）で、Mesaconate を Citramalate に変化させる。予測の結果、*C. tetani* の遺伝子 ctet:BN906_02813 と ctet:BN906_02812 の 2 つの遺伝子が含まれる 2 つの KO グループ K01677 と K01678 が予測された。これらの KO グループの機能は Fumarate hydratase subunits α と β（EC 4.2.1.2）であり、非常に似た反応を触媒している。また、これらの遺伝子

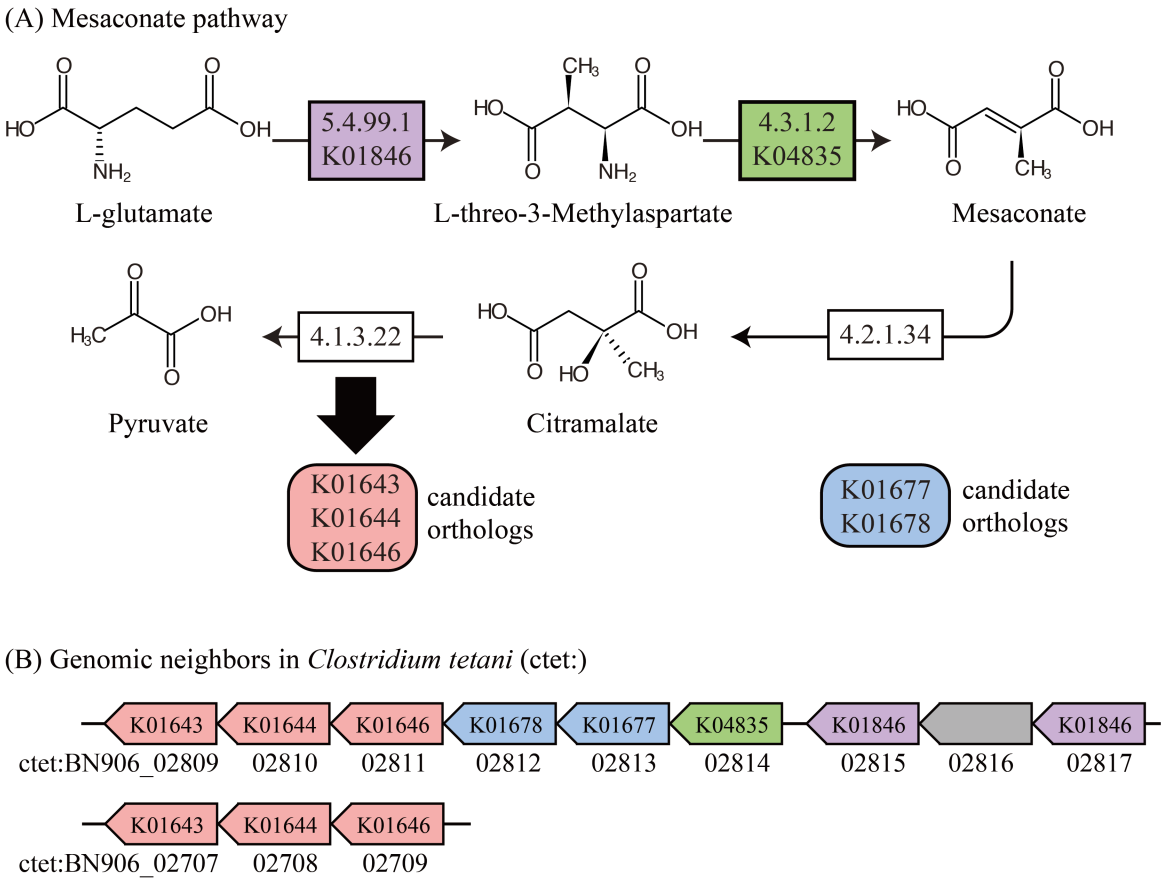


図 4.7 メサコン酸パスウェイと関連遺伝子の並び [76]

はゲノム上で前半 2 つの反応を触媒する遺伝子と隣接している（図 4.7 B）。また、EC 4.2.1.34 の触媒能を持つ酵素は EC 4.2.1.2 の触媒能も有しているという報告もあり [75]、予測された酵素候補は多機能酵素である可能性も高い。

第四の反応は(S3)-citramalate pyruvate-lyase 反応（EC 4.1.3.22）で、Citramalate から Pyruvate と Acetate を生成する反応である。予測の結果、*C. tetani* では Citrate lyase subunits α, β, γに対応する K01643、K01644、K01646 が酵素候補として予測された。これら KO グループが触媒する酵素反応は EC 4.1.3.6 であり、メサコン酸パスェイにおける酵素反応とは異なっているが、一方で 3 つの酵素候補は第一、第二の反応における酵素及び、第三の反応で予測された酵素候補に続いてゲノム上で隣接している(図 4.7 B)。そのため、これらの酵素候補セットも多機能酵素である可能性が示唆される。また、*C. tetani* はゲノム中に 3 つの酵素候補のパラログ遺伝子セットを有している。このため、酵素候補セットとそのパラログ遺伝子セットが、EC 4.1.3.22 と EC 4.1.3.6 の 2 つの酵素反応をそれぞれ分けて触媒している可能性も示唆される。

本研究で提案した手法は E-zyme 2（http://www.genome.jp/tools/e-zyme2/）として実装、公開されており、誰でも利用可能となっている（図 4.8）[76]。基質と生成物の構造情報を入力することで、酵素候補のリスト及び関連する OC グループのリストが得られ、生物種及び分類群による遺伝子の絞り込みが可能となっている。化合物の構造情報は MDL MOL 形式、KEGG 化合物 ID、医薬品 ID のほか、PubChem ID、日化辞 ID、KNApSAcK ID などの入力に対応している。

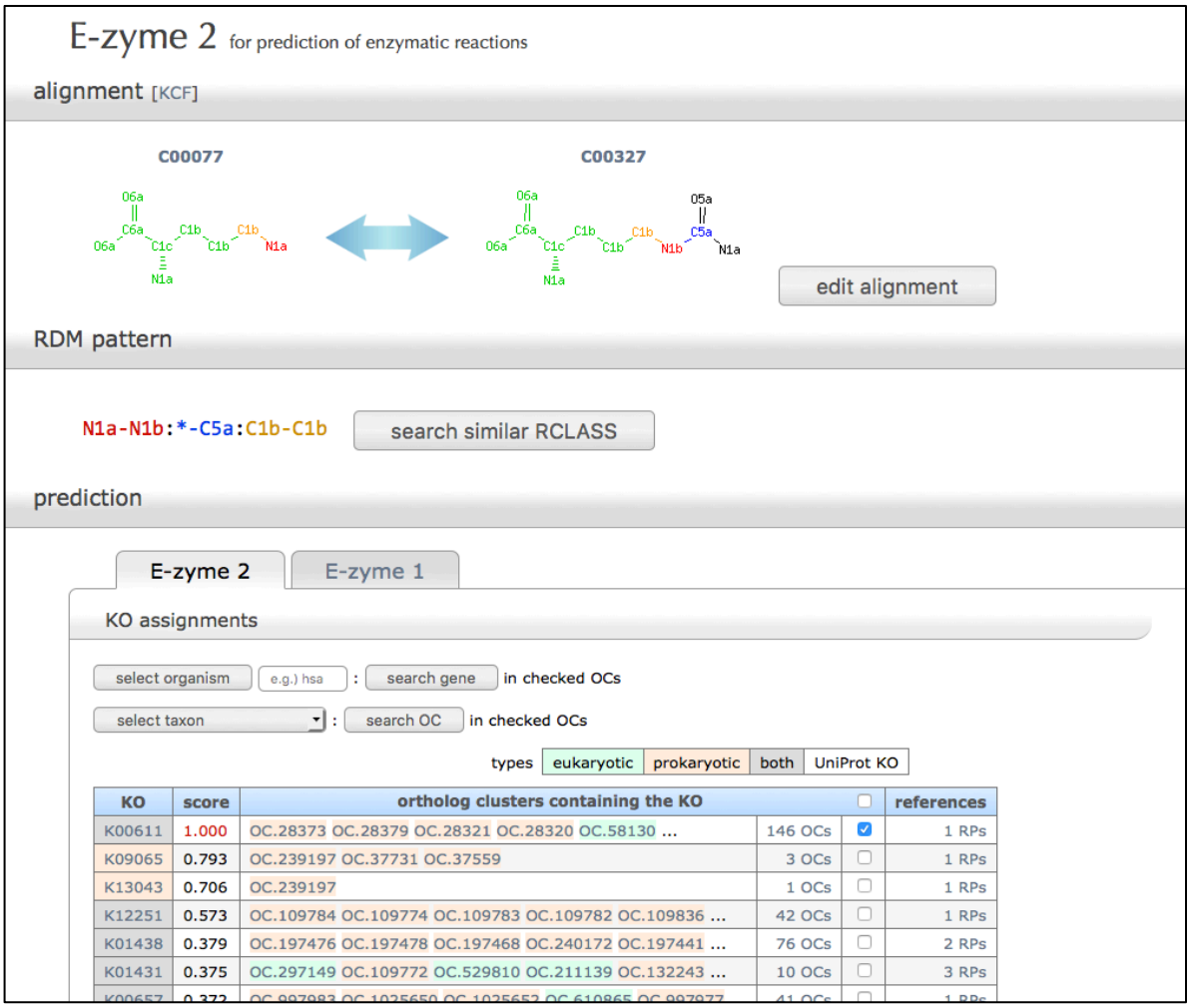


図 4.8 E-zyme 2 ウェブサービス

第５章　まとめと展望

本研究の結果、新規に決定された配列に対してゲノムアノテーションを行うことで、KEGG PATHWAY へのマッピングを介してパスウェイを計算機上で表現することが可能になった。またリファレンスパスウェイで不足している代謝経路の予測や、反応を担う代謝酵素の予測を行うことで、パスウェイデータベースの拡張や補完といった解析が可能となった。また、得られたゲノムアノテーションやパスウェイを利用した比較ゲノム解析などを行うことで、進化解析や医療・創薬などの様々な分野での解析が可能になると考えられる。これら全ての手法を容易に利用できるウェブツールとして公開することで、バイオインフォマティクス分野の研究者のみならず、幅広い分野でのパスウェイデータベースを利用した研究が進むと期待される。

KAAS により、新規に配列の決定した生物種においても即座に KEGG PATHWAY データベースを利用することが可能となった。また、KAAS における配列類似性検索に BLAST プログラムを用いていたが、現在までにサフィックスアレイを利用することで精度をそれほど落とさずに高速に検索を行う GHOSTX[77] および GHOSTZ[78] プログラムが発表されている。2016 年現在、ウェブサービスではこれら高速プログラムを選択できるようになっており、より高速に KO アノテーション、及びゲノム解析が行えるようになっている。また KAAS と同様に、配列類似性に基づきながら、データベースを縮小することで高速に KO アノテーションを行うことを目的とした BlastKOALA、GhostKOALA が KEGG において提供されている [79]。

遺伝子機能の自動アノテーションは生物単位の解析で用いられるだけでなく、現在ではメタゲノム解析にも用いられている。初期のメタゲノミクス研究では主に環境サンプル中の 16S rRNA 遺伝子のみをシークエンスすることによって、生物分類の解析することを目的としていたが、近年のシークエンシングコストの低下に伴い、環境サンプル中の全ての配列をシークエンシングすることが容易となった。そのため、メタゲノムの全配列データに対して、KAAS による遺伝子の機能のアノテーションを行うことで、環境サンプルが有していると考えられる代謝・生理機能の評価を行うことを目的としたシステム、MAPLE が開発されている [80]。

一方で次世代シークエンサの普及に伴い、開発当時とは異なった配列データである、アセンブルのされていない膨大な短いリード配列が蓄積されつつある。配列類似性検索を用いて機能アノテーションを行う KAAS では、アミノ酸配列で 100 から 120 残基が必要であるため [81]、現在普及している Illumina シークエンサの短いリード配列

を直接アノテーションすることは難しかった。しかしながらシークエンスに用いる試薬の開発が進み 300bp までリード長が伸長されため、これらに対しては KAAS によるアノテーションが可能であると考えられる。また鋳型となる DNA 断片の両端（ペアエンド）のリード情報を用いた予測を行うことでより精度の高い機能アノテーションが可能になると考えられる。

PathPred では、植物の二次代謝産物合成経路及び生体異物の分解経路の代謝反応を予測できるようになった。また現在では、３．１で先行研究として挙げた UM-PPS においても多段階の生分解経路予測が可能となっている[82]。さらに、多くの代謝化合物が蓄積されているメタボロームデータから代謝経路を予測する手法も開発されており[83]、未知代謝経路の予測が重要な研究テーマとなっていることが伺える。本研究では計算機資源を考慮して代謝パスウェイの一部を予測の対象としたが、今後の計算能力の発展に伴い、パスウェイカテゴリを考慮しない網羅的な経路予測が可能になることが期待される。また、本研究では光学異性体及び互変異性体などの異性体の扱いに問題点が残った。光学異性体については RDM パターンでは区別することができないが、KEGG COMPOUND データベースを参照することで区別が可能であると考えられる。しかしながら、KEGG COMPOUND データベースでは、多くの化合物の立体構造は MDL MOL 形式同様に、"原子の xy 平面座標"と"結合の z 軸方向の向き（up/down）"で記述されている。そのため立体異性を識別するには座標計算が必要となるため、困難が伴うと考えられる。一方、互変異性体については、共有結合の数を内包した現在の KEGG atom type では表現が難しく、互変異性の変換を反応の一種として記述したリアクタントペアとして、データベースへの蓄積を行う必要性や、そのための互変異性体を推測するツールの開発の必要性があると考えられる。

E-zyme 2 によりオルファン酵素の推定が可能となったが、化合物の構造情報のみを用いた初めての試みであるため、まだ予測できる数は多いとは言えない。しかしながら本手法を用いた網羅的なオルファン酵素の探索において、数例の実証されている遺伝子を正しく予測できることも示された。またメサコン酸パスウェイの例のようにゲノム上における遺伝子の並びの情報を用いることで、酵素予測においてより多くの示唆を得ることができた。そのため、さらに多くの生物学的情報を組み合わせることによって酵素予測の精度が向上すると考えられ、検証実験の足がかりに利用されることが期待される。

現在、東京工業大学においてヒト腸内細菌叢代謝パスウェイデータベースの開発が進んでいる[84,85]。これは KEGG パスウェイデータベースで収集していない代謝経路を含んだ、腸内細菌叢における食物や薬物などの代謝経路を収集したデータベースにな

っている。しかしながら触媒する酵素遺伝子が未だ知られていない反応経路も多く登録されており、このヒト腸内細菌叢代謝パスウェイデータベースでは、本研究で提案した手法で酵素遺伝子の推定を行うことで、データベースにおける遺伝子の情報を蓄積している。このデータベースを KEGG 同様に、本論文で提案する手法のリファレンスデータベースとして利用できる形式に整備することで、KAAS を用いた腸内細菌叢メタゲノムのゲノムアノテーションに有用なりリソースとなり、また PathPred や E-zyme 2 を用いることで、薬物の腸内での代謝経路予測や、それに関わる遺伝子の予測など、創薬や医療の分野における利用も進むことが考えられる。

謝辞

本論文の第２章、第３章において述べた研究は、京都大学化学研究所バイオインフォマティクスセンター生命知識システム領域（現・化学生命科学研究領域）において、金久實教授（現・京都大学特任教授）の指導の下に行われました。同教授にはバイオインフォマティクスという分野で研究する機会と環境を与えていただきました。また多くの助言を頂き、研究に対する姿勢を学ばせていただきました。深く感謝致します。

本論文の第４章において述べた研究は、同研究室の五斗進准教授の指導の下に行われました。研究の機会と環境を与えていただきありがとうございます。五斗准教授には第２章、第３章においても多岐にわたりご指導・助言を頂きました。深く感謝いたします。

また、同研究室に在籍した多くの方に研究の上で大変お世話になりました。服部正泰助教（当時）、小寺正明助教（現・東京工業大学講師）、時松敏明助教（当時）、重水大智博士（現・東京医科歯科大学講師）、中川善一氏には反応データベースを中心とした化学情報データの解析についてご指導頂きました。ありがとうございます。

またバイオインフォマティクスセンターに関わる全ての方に感謝の意を表したいと思います。特に伊藤真純助教（当時）、奥田修二郎博士（現・新潟大学准教授）、山田拓司助教（現・東京工業大学准教授）、吉沢明康博士には公私にわたりお世話になりました。研究室秘書の方々、化学研究所スーパーコンピューターシステムの方々にも感謝致します。

本研究は KEGG データベース無しには行えませんでした。KEGG の維持・更新を行っている KEGG プロジェクトのメンバーにも深く感謝致します。

また、現在所属している情報・システム研究機構データサイエンス共同利用基盤施設ライフサイエンス統合データベースセンターでは、本論文の執筆の場と機会を与えて頂きました。小原雄治センター長とセンターに所属する皆様、及び共同研究機関である科学技術振興機構バイオサイエンスデータベースセンターの高木利久センター長に深く感謝致します。

参考文献

1. Sanger, F., Nicklen, S., and Coulson, A. R. 1977, DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, **74**, 5463–7.
2. The 1000 Genomes Project Consortium. 2012, An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
3. Nagasaki, M., Yasuda, J., Katsuoka, F., et al. 2015, Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.*, **6**, 8018.
4. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. 2008, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–8.
5. Arumugam, M., Raes, J., Pelletier, E., et al. 2011, Enterotypes of the human gut microbiome. *Nature*, **473**, 174–80.
6. Bork, P., Bowler, C., de Vargas, C., Gorsky, G., Karsenti, E., and Wincker, P. 2015, Tara Oceans studies plankton at planetary scale. *Science*, **348**, 873–873.
7. Mukherjee, S., Stamatis, D., Bertsch, J., et al. 2016, Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res.*, Oct 19, 2016, doi: 10.1093/nar/gkw992.
8. Schadt, E. E., Turner, S., and Kasarskis, A. 2010, A window into third-generation sequencing. *Hum. Mol. Genet.*, **19**, R227-40.
9. Brent, M. R. 2005, Genome annotation past, present, and future: How to define an ORF at each locus. *Genome Res.*, **15**, 1777–86.
10. Stein, L. 2001, Genome annotation: from sequenceto biology. *Nat. Rev. Genet.*, **2**, 493–503.
11. The Gene Ontology Consortium. 2008, The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440-4.
12. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. 2016, KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–62.
13. http://www.roche.com/sustainability/what_we_do/for_communities_and_

environment/philanthropy/science_education/pathways.htm (accessed Nov 24, 2016).

14. http://www.kegg.jp/kegg-bin/show_pathway?map01100 (accessed Jan 25, 2017).
15. Kanehisa, M. 1996, Toward pathway engineering: A new database of genetic and molecular pathways. *Sci. Technol. Japan*, **59**, 34–8.
16. Karp, P. D., Riley, M., Paley, S. M., and Pelligrini-Toole, A. 1996, EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.*, **24**, 32–9.
17. Caspi, R., Altman, T., Billington, R., et al. 2014, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **42**, 471–80.
18. Croft, D., Mundo, A. F., Haw, R., et al. 2014, The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **42**, 481–7.
19. Morgat, A., Coissac, E., Coudert, E., et al. 2012, UniPathway: A resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.*, **40**, 761–9.
20. Kutmon, M., Riutta, A., Nunes, N., et al. 2015, WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res.*, **44**, D488–94.
21. Huang, D. W., Lempicki, R. a, and Sherman, B. T. 2009, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
22. Tripathi, S., Pohl, M. O., Zhou, Y., et al. 2015, Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. *Cell Host Microbe*, **18**, 723–35.
23. Zhang, G., Li, C., Li, Q., et al. 2014, Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, **346**, 1311–20.
24. The International Aphid Genomics Consortium. 2010, Genome Sequence of the Pea Aphid *Acyrthosiphon pisum* Eisen, J. A., (ed.), . *PLoS Biol.*, **8**, e1000313.
25. Perna, N. T., Plunkett, G., Burland, V., et al. 2001, Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–33.
26. Rubin, G. M., Yandell, M. D., Wortman, J. R., et al. 2000, Comparative

genomics of the eukaryotes. *Science*, **287**, 2204–15.

27. Rost, B. 2002, Enzyme Function Less Conserved than Anticipated. *J. Mol. Biol.*, **318**, 595–608.
28. Tian, W., and Skolnick, J. 2003, How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity? *J. Mol. Biol.*, **333**, 863–82.
29. Oh, M., Yamada, T., Hattori, M., Goto, S., and Kanehisa, M. 2007, Systematic Analysis of Enzyme-Catalyzed Reaction Patterns and Prediction of Microbial Biodegradation Pathways. *J. Chem. Inf. Model.*, **47**, 1702–12.
30. Yamada, T., Waller, A. S., Raes, J., et al. 2012, Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours. *Mol. Syst. Biol.*, **8**, 581.
31. Yamanishi, Y., Hattori, M., Kotera, M., Goto, S., and Kanehisa, M. 2009, E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. *Bioinformatics*, **25**, i179-86.
32. Kotera, M., Okuno, Y., Hattori, M., Goto, S., and Kanehisa, M. 2004, Computational Assignment of the EC Numbers for Genomic-Scale Analysis of Enzymatic Reactions. *J. Am. Chem. Soc.*, **126**, 16487–98.
33. Fleischmann, R. D., Adams, M. D., White, O., et al. 1995, Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
34. Goffeau, A., Barrell, B. G., Bussey, H., et al. 1996, Life with 6000 genes. *Science*, **274**, 546–67.
35. The C. elegans Sequencing Consortium. 1998, Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, **282**, 2012–8.
36. International Human Genome Sequencing Consortium. 2004, Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–45.
37. Smith, T. F., and Waterman, M. S. 1981, Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–7.
38. Lipman, D. J., and Pearson, W. R. 1985, Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–41.
39. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990, Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–10.

40. Altschul, S., Madden, T. L., Schäffer, A. A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–402.
41. Tatusov, R. L., Koonin, E. V, Lipman, D. J., et al. 1997, A genomic perspective on protein families. *Science*, **278**, 631–7.
42. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., et al. 2001, The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–8.
43. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. 2014, Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199-205.
44. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. 2007, KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, **35**, W182–5.
45. Gao, J., Ellis, L. B. M., and Wackett, L. P. 2010, The University of Minnesota Biocatalysis/Biodegradation Database: improving public access. *Nucleic Acids Res.*, **38**, D488-91.
46. Fenner, K., Gao, J., Kramer, S., Ellis, L., and Wackett, L. 2008, Data-driven extraction of relative reasoning rules to limit combinatorial explosion in biodegradation pathway prediction. *Bioinformatics*, **24**, 2079–85.
47. McDonald, A. G., Boyce, S., and Tipton, K. F. 2009, ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.*, **37**, D593-7.
48. Kotera, M., Okuno, Y., Hattori, M., Goto, S., and Kanehisa, M. 2004, Computational Assignment of the EC Numbers for Genomic-Scale Analysis of Enzymatic Reactions. *J. Am. Chem. Soc.*, **126**, 16487–98.
49. Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M. 2003, Development of a Chemical Structure Comparison Method for Integrated Analysis of Chemical and Genomic Information in the Metabolic Pathways. *J. Am. Chem. Soc.*, **125**, 11853–65.
50. Hattori, M., Tanaka, N., Kanehisa, M., and Goto, S. 2010, SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res.*, **38**, W652-6.
51. Moriya, Y., Shigemizu, D., Hattori, M., et al. 2010, PathPred: An

enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res.*, **38**, W138–43.

52. Wishart, D. S., Jewison, T., Guo, A. C., et al. 2013, HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res.*, **41**, D801-7.
53. Afendi, F. M., Okada, T., Yamazaki, M., et al. 2012, KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol.*, **53**, e1.
54. Kotera, M., Tabei, Y., Yamanishi, Y., Tokimatsu, T., and Goto, S. 2013, Supervised de novo reconstruction of metabolic pathways from metabolome-scale compound sets. *Bioinformatics*, **29**, i135-44.
55. Pouliot, Y., and Karp, P. D. 2007, A survey of orphan enzyme activities. *BMC Bioinformatics*, **8**, 244.
56. Yamanishi, Y., Mihara, H., Osaki, M., et al. 2007, Prediction of missing enzyme genes in a bacterial metabolic network. Reconstruction of the lysine-degradation pathway of *Pseudomonas aeruginosa*. *FEBS J.*, **274**, 2262–73.
57. Kharchenko, P., Vitkup, D., and Church, G. M. 2004, Filling gaps in a metabolic network using expression information. *Bioinformatics*, **20 Suppl 1**, i178-85.
58. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. 1999, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **96**, 4285–8.
59. Green, M. L., and Karp, P. D. 2004, A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, **5**, 76.
60. Kotera, M., Yamanishi, Y., Moriya, Y., Kanehisa, M., and Goto, S. 2012, GENIES: Gene network inference engine based on supervised analysis. *Nucleic Acids Res.*, **40**, W162–7.
61. O'Boyle, N. M., Holliday, G. L., Almonacid, D. E., and Mitchell, J. B. O. 2007, Using Reaction Mechanism to Measure Enzyme Similarity. *J. Mol. Biol.*, **368**, 1484–99.
62. Egelhofer, V., Schomburg, I., and Schomburg, D. 2010, Automatic assignment of EC numbers. *PLoS Comput. Biol.*, **6**, e1000661.

63. Hu, Q.-N., Zhu, H., Li, X., et al. 2012, Assignment of EC numbers to enzymatic reactions with reaction difference fingerprints. *PLoS One*, **7**, e52901.
64. Rahman, S. A., Cuesta, S. M., Furnham, N., Holliday, G. L., and Thornton, J. M. 2014, EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat. Methods*, **11**, 171–4.
65. Latino, D. A. R. S., and Aires-de-Sousa, J. 2009, Assignment of EC numbers to enzymatic reactions with MOLMAP reaction descriptors and random forests. *J. Chem. Inf. Model.*, **49**, 1839–46.
66. Hu, X., Yan, A., Tan, T., Sacher, O., and Gasteiger, J. 2010, Similarity perception of reactions catalyzed by oxidoreductases and hydrolases using different classification methods. *J. Chem. Inf. Model.*, **50**, 1089–100.
67. Nath, N., and Mitchell, J. B. O. 2012, Is EC class predictable from reaction mechanism? *BMC Bioinformatics*, **13**, 60.
68. Matsuta, Y., Ito, M., and Tohsato, Y. 2013, ECOH: an enzyme commission number predictor using mutual information and a support vector machine. *Bioinformatics*, **29**, 365–72.
69. Nakaya, A., Katayama, T., Itoh, M., et al. 2013, KEGG OC: A large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic Acids Res.*, **41**, D353–7.
70. Kotera, M., Tabei, Y., Yamanishi, Y., et al. 2013, KCF-S: KEGG Chemical Function and Substructure for improved interpretability and prediction in chemical bioinformatics. *BMC Syst. Biol.*, **7 Suppl 6**, S2.
71. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed Nov 24, 2016).
72. Ajdić, D., McShan, W. M., McLaughlin, R. E., et al. 2002, Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 14434–9.
73. Bains, J., and Boulanger, M. J. 2008, Structural and Biochemical Characterization of a Novel Aldehyde Dehydrogenase Encoded by the Benzoate Oxidation Pathway in *Burkholderia xenovorans* LB400. *J. Mol. Biol.*, **379**, 597–608.
74. Buckel, W., and Barker, H. A. 1974, Two pathways of glutamate

fermentation by anaerobic bacteria. *J. Bacteriol.*, **117**, 1248–60.

75. Wang, C. C., and Barker, H. A. 1969, Purification and properties of L-citramalate hydrolyase. *J. Biol. Chem.*, **244**, 2516–26.
76. Moriya, Y., Yamada, T., Okuda, S., et al. 2016, Identification of Enzyme Genes Using Chemical Structure Alignments of Substrate-Product Pairs. *J. Chem. Inf. Model.*, **56**, 510–6.
77. Suzuki, S., Kakuta, M., Ishida, T., et al. 2014, GHOSTX: An Improved Sequence Homology Search Algorithm Using a Query Suffix Array and a Database Suffix Array. *PLoS One*, **9**, e103833.
78. Suzuki, S., Kakuta, M., Ishida, T., and Akiyama, Y. 2015, Faster sequence homology searches by clustering subsequences. *Bioinformatics*, **31**, 1183–90.
79. Kanehisa, M., Sato, Y., and Morishima, K. 2016, BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.*, **428**, 726–31.
80. Takami, H., Taniguchi, T., Arai, W., Takemoto, K., Moriya, Y., and Goto, S. 2016, An automated system for evaluation of the potential functionome: MAPLE version 2.1.0. *DNA Res.*, **23**, 467–75.
81. Takami, H., Taniguchi, T., Moriya, Y., Kuwahara, T., Kanehisa, M., and Goto, S. 2012, Evaluation method for the potential functionome harbored in the genome and metagenome. *BMC Genomics*, **13**, 699.
82. Gao, J., Ellis, L. B. M., and Wackett, L. P. 2011, The University of Minnesota Pathway Prediction System: multi-level prediction and visualization. *Nucleic Acids Res.*, **39**, W406–11.
83. Kotera, M., Tabei, Y., Yamanishi, Y., et al. 2014, Metabolome-scale prediction of intermediate compounds in multistep metabolic pathways with a recursive supervised approach. *Bioinformatics*, **30**, i165-74.
84. 奥田修二郎，佃直紀，山本希，et al. 2014，ヒト腸内細菌叢解析のためのパスウェイデータベース構築. *第37回日本分子生物学会年会*.
85. http://www.enteropathway.org/ (accessed Nov 24, 2016).

付録 1

代表生物種セット

古くからゲノム配列が決定され、また研究が進み遺伝子の機能アノテーションが進んでいると考えられる生物種を、真核生物と原核生物を跨いで選択した。このリストに、真核生物、原核生物において系統的に漏れていると思われる生物種を追加することで、真核セット及び原核セットを作成した。

分類群	生物種名	真核セット(26)	原核セット(28)
Eukaryotes	Homo sapiens (human)	○	○
	Mus musculus (mouse)	○	
	Rattus norvegicus (rat)	○	
	Danio rerio (zebrafish)	○	
	Drosophila melanogaster (fruit fly)	○	○
	Caenorhabditis elegans (nematode)	○	
	Arabidopsis thaliana (thale cress)	○	○
	Saccharomyces cerevisiae (budding yeast)	○	○
	Ashbya gossypii (Eremothecium gossypii)	○	
	Candida albicans	○	
	Schizosaccharomyces pombe (fission yeast)	○	
	Encephalitozoon cuniculi	○	
	Entamoeba histolytica	○	
	Plasmodium falciparum 3D7	○	○
	Cryptosporidium hominis	○	
Bacteria	Escherichia coli K-12 MG1655	○	○
	Salmonella enterica subsp. enterica serovar Typhi CT18		○
	Haemophilus influenzae Rd KW20 (serotype d)		○
	Pseudomonas aeruginosa PAO1		○
	Neisseria meningitidis MC58 (serogroup B)	○	○
	Helicobacter pylori 26695	○	○
	Rickettsia prowazekii Madrid E		○
	Mesorhizobium loti		○

	Bacillus subtilis subsp. subtilis 168	○	○
	Staphylococcus aureus subsp. aureus N315 (MRSA/VSSA)		○
	Lactococcus lactis subsp. lactis Il1403	○	○
	Streptococcus pneumoniae TIGR4 (virulent serotype 4)		○
	Clostridium acetobutylicum ATCC 824		○
	Mycoplasma genitalium G37	○	○
	Mycobacterium tuberculosis H37Rv	○	○
	Synechocystis sp. PCC 6803	○	○
	Chlamydia trachomatis D/UW-3/CX		○
	Borrelia burgdorferi B31		○
	Aquifex aeolicus	○	○
Archaea	Methanocaldococcus jannaschii	○	○
	Archaeoglobus fulgidus DSM 4304		○
	Pyrococcus horikoshii		○
	Aeropyrum pernix	○	○

付録２

KEGG atom type リスト

Frequency は KEGG データベースにおける出現回数を示している。

（http://www.genome.jp/kegg/reaction/KCF.html accessed Nov 24, 2016）

Atom	Functional group	Atom type	Description	Frequency
C	Alkane	C1a	R-CH3	16473
		C1b	R-CH2-R	20193
		C1c	R-CH(-R)-R	4964
		C1d	R-C(-R)2-R	698
	Cyclic alkane	C1x	ring-CH2-ring	14010
		C1y	ring-CH(-R)-ring	27376
		C1z	ring-C(-R)2-ring	4463
	Alkene	C2a	R=CH2	634
		C2b	R=CH-R	3965
		C2c	R=C(-R)2	1914
	Cyclic alkene	C2x	ring-CH=ring	2964
		C2y	ring-C(-R)=ring or ring-C(=R)-ring	3722
	Alkyne	C3a	R≡CH	43
		C3b	R≡C-R	282
	Aldehyde	C4a	R-CH=O	350
	Ketone	C5a	R-C(=O)-R	3595
	Cyclic ketone	C5x	ring-C(=O)-ring	2257
	Carboxylic acid	C6a	R-C(=O)-OH	3190
	Carboxylic ester	C7a	R-C(=O)-O-R	1691
		C7x	ring-C(=O)-O-ring	869
	Aromatic ring	C8x	ring-CH=ring	19905
		C8y	ring-C(-R)=ring	20511
	Undefined C	C0		8
N	Amine	N1a	R-NH2	2440
		N1b	R-NH-R	3003

		N1c	R-N(-R)2	374
		N1d	R-N(-R)3+	105
	Cyclic amine	N1x	ring-NH-ring	806
		N1y	ring-N(-R)-ring	1464
	Imine	N2a	R=N-H	230
		N2b	R=N-R	163
	Cyclic imine	N2x	ring-N=ring	357
		N2y	ring-N(-R)+=ring	14
	Cyan	N3a	R≡N	119
	Aromatic ring	N4x	ring-NH-ring	785
		N4y	ring-N(-R)-ring	840
		N5x	ring-N=ring	2131
		N5y	ring-N(-R)+=ring	59
	Undefined N	N0		194
O	Hydroxy	O1a	R-OH	18369
		O1b	N-OH	198
		O1c	P-OH	3111
		O1d	S-OH	332
	Ether	O2a	R-O-R	4199
		O2b	P-O-R	2481
		O2c	P-O-P	502
		O2x	ring-O-ring	5853
	Oxo	O3a	N=O	134
		O3b	P=O	2248
		O3c	S=O	941
	Aldehyde	O4a	R-CH=O	350
	Ketone	O5a	R-C(=O)-R	3595
		O5x	ring-C(=O)-ring	2862
	Carboxylic acid	O6a	R-C(=0)-OH	6384
	Ester	O7a	R-C(=O)-O-R	3382
		O7x	ring-C(=O)-O-ring	1738
	Undefined O	O0		127

S	Thiol	S1a	R-SH	100
	Thioether	S2a	R-S-R	420
		S2x	ring-S-ring	261
	Disulfide	S3a	R-S-S-R	45
		S3x	ring-S-S-ring	48
	Sulfate	S4a	R-SO3	267
	Undefined S	S0		223
P	Attatched to other elements	P1a	P-R	112
	Attatched to oxygen	P1b	P-O	2158
Other	Halogens	X	F, Cl, Br, I	1419
	Others	Z		261