# 生命システム解明の基盤データベース構築

●金久 實[1,2]　◆服部 正泰[1]　◇片山 俊明[2]
1）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　2）東京大学医科学研究所ヒトゲノム解析センター

**＜研究の目的と進め方＞**

本研究は、遺伝子･分子レベルの網羅的な解析から、細胞･個体･生態系レベルでの生命システムの全体像を明らかにすることを目指し、新しい情報技術の開発とともに、新しいタイプの基盤データベースを構築することを目的とする。細胞・個体レベルでの生命システム情報は、これまでのKEGGにおいてすでにデータベース化が行われているので、本研究では生物種間相互作用や環境との相互作用といったより高次レベルの生命システム情報をゲノムの情報と統合し、医療や産業をはじめ、ゲノム情報の有効利用へつなぐ基盤データベースを構築する。同時に支援班との協力の下に、様々な利用・解析ツールを開発して、特定ゲノム４領域との間でフィードバック連携をはかり、これら領域の研究推進及び成果の統合化に寄与する。

**＜2008年度の研究の当初計画＞**

文献データやなまの実験データを蓄積した従来型のデータベースに対し、本研究のデータベースはこれらのデータから得られる「知識」を蓄積する。その知識をもとに新しい研究を推進することが可能となるように、本データベースの様々な利用ツールを開発する。具体的な内容は以下の通りである。

・生命システムを構成する部品の情報として、ゲノム情報から直接的に規定される遺伝子とタンパク質はKEGG GENESデータベースに、それ以外のケミカル情報に関連した分子はKEGG LIGANDデータベースに蓄積する。
・KEGG GENESの機能アノテーションは、KO（KEGG Orthology） システムをさらに拡張することでその割合を高めていく。またKAAS, EGassembler, GENIESといった機能アノテーションツール群を改良し、その活用のための支援を行う。
・KEGG LIGANDの情報はこれまでRDMパターンとしてのみ体系化を行ってきたが、化合物の共通部分構造などに着目して、遺伝子のKOに相当するようなグループ化を行い、生物的な意味づけを行っていく。また化学構造比較プログラムSIMCOMP, SUBCOMP、化学反応予測プログラムe-zyme, PathCompなどの改良、その他のプログラムの開発、これらの活用のための支援を行う。
・これまで糖鎖、PK/NRP、脂肪酸などで行ってきたゲノムからの化学構造予測の方法論を、植物の二次代謝物質に適用する。
・ゲノムネットのWebサービスを通して、KEGGをはじめとしたリソースをプログラムから呼び出し、カスタマイズして利用できるインターフェースKEGG APIの開発を継続して行う。
・基盤ゲノム領域及び他のゲノム領域に対して、ゲノムデータやEST データの機能アノテーションなど、本研究の成果を生かした支援を、本領域の支援班と連携して行う。
・KEGGの利用講習会やKEGG API入門コース等を京都または東京で定期的に開催して、我が国全体での普及をはかる。
・高度の専門知識を効率的に集積するために、研究コミュニティと密接に連携し、その知識を集約する「コミュニティデータベース」の枠組みとして、CYORFデータベース等をさらに発展させる。

**＜2008年度の成果＞**

KEGG GENESデータベースには909生物種の425万遺伝子に関する情報が蓄積され、昨年度と比較して50%の増加である。一昨年度から昨年度への増加率は52% であったので、ほぼ同じような増加傾向にある。今年度の最も大きな成果は、このようなゲノムデータの急増に対応できる新しいアノテーションシステムを開発したことである。

KEGG GENESは全ゲノム配列が決定されたすべての生物種について、その遺伝子セットをRefSeq その他の公共データベースから自動生成し、KEGG独自のアノテーションを行っているデータベースである。アノテーションはKO (KEGG Orthology) システムに基づき行われる。KOとはKEGGパスウェイの各ノードまたは BRITE 機能階層の最下層ノードに対応したオーソロググループを手作業で定義したものである。オーソロググループはK番号で識別されるので、ゲノム中の各遺伝子にK番号を付与することが、KEGGのアノテーションで、これによりゲノムからKEGGパスウェイやBRITE機能階層へのマッピング（エンリッチメント）が可能となる。オーソログ関係を決めるために、SSEARCHプログラムでアミノ酸レベルのゲノム比較を行い、ゲノムペアごとに全遺伝子間の配列類似性スコアとベストヒット関係を保持したKEGG SSDBデータベースが維持されている。これをもとにGFITツールでゲノムごとに手作業のアノテーションを行うのが従来のやり方であった。

新しいアノテーションツールはKOALA (KEGG Orthology And Links Annotation）と呼ばれ、従来型のゲノム単位のアノテーション（縦方向アノテーション）だけでなく、K番号単位のアノテーション（横方向アノテーション）をすることができる。すなわちKEGGパスウェイのあるノードにマップされるオーソログ遺伝子群を、すべての生物種に対して一括アノテーションできるのである。また、KOALAにはGFIT作業を自動化した機能があり、間違いの少ない、オーソロググループとしてまとまりのいい安全なK番号は、新規ゲノムに対して自動アサイメントを行う。K番号のグルーピングは常に見直しを行っており、グルーピングをきれいにすることで、最終的には現在約１万あるK番号の大半を自動アサイメントできるようにしていく予定である。さらに、GFITのやり方とは全く独立にKEGG SSDBから計算のみでオーソログクラスターを生成するKEGG OCの情報もKOALAで参照できるようになっており、作業の効率化に役立っている。なお、従来からのKAAS自動アノテーションプログラムは外部向けサービス専用とし、内部でのアノテーションはすべてKOALAに移行した。

KEGG LIGANDデータベースには化合物、糖鎖、医薬品、化

学反応、酵素に関する情報が含まれ、今年度はとくに医薬品関係のデータを重点的に整備した。当初計画に記載した化学情報に関するプログラム群の整備は統合データベースプロジェクトの一環として行われるようになったため、本研究の成果には含めないが、SIMCOMP、SUBCOMP の高速化と高機能化が実現した。

本研究では、ゲノムの情報から、生体内で合成され得る化合物の化学構造を予測する方法論の開拓を行ってきた。今年度はこれを植物二次代謝物質へ拡張することを目指し、新たに KEGG PLANT を開発し公開した。このリソースは植物ゲノムと EST の情報を KEGG GENES に、植物の二次代謝経路を KEGG PATHWAY に、合成経路に基づく植物二次代謝物質の分類を KEGG BRITE に整備し統合したものである。

支援活動については、本特定領域の個別の研究グループに対して cDNA データのアノテーション支援などを行い、また 1 月 29 日と 30 日に東大医科研ヒトゲノム解析センターで KEGG データベース利用講習会を開催した。

**＜国内外での成果の位置づけ＞**

KEGG はその３つの柱のうち、パスウェイ情報とケミカル情報については、国際標準のデータベースとしての地位を獲得している。しかしもう１つのゲノム情報については、アノテーションのカバー率が低いこともあり、国際的評価はまだそれほど高くない。一方、遺伝子・タンパク質の高品質アノテーションを提供していた SwissProt が UniProt となり、ごく一部の手動アノテーション（SwissProt）と大量の自動アノテーション（TrEMBL）が混在する形となってしまい、全体の品質が落ちてしまった印象がある。KEGG では個々のゲノムにアノテーションを行う立場ではなく、生物界すべてのゲノムに一括アノテーションを行う立場をとっている。KEGG SSDB や KOALA はこのような考え方を実用化したものである。一般にはゲノム数の増加がアノテーションの負荷を高め作業を困難にしていると思われているが、KEGG では状況は全く逆で、ゲノム数の増加が生物界の全体像を分かり易くし、アノテーションが容易になる方向へ向かっている。KEGG がゲノム情報においても国際標準となり得る状況になってきたと考えている。

**＜達成できなかったこと、予想外の困難、その理由＞**

前述の通り化学構造比較プログラム SIMCOMP, SUBCOMP、化学反応予測プログラム e-zyme, PathComp などの改良、その他の化学情報関連プログラムの開発は本研究では行わなかった。統合データベースプロジェクトとの重複を避けたことが理由である。また広い意味では、内部で行う KEGG データベース構築のコア部分と、外部で行う KEGG を利用したソフトウェア等の開発部分を切り分け、後者の活動を促進するような方策をとるようにしたためでもある。

**＜今後の課題＞**

本研究はあと１年であり、最終年度までにゲノムアノテーションの部分において KEGG の国際標準化を達成したいと考えている。

**＜成果公表リスト＞**

論文

1. 0901161008
Shimizu, Y., Hattori, M., Goto, S., and Kanehisa, M.; Generalized reaction patterns for prediction of unknown enzymatic reactions. Genome Informatics 20, 149-158 (2008）.

2. 0901161004
Takarabe, M., Okuda, S., Itoh, M., Tokimatsu, T., Goto, S., and Kanehisa, M.; Network analysis of adverse drug interactions. Genome Informatics 20, 252-259 (2008）.

3. 0901131629
Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., and Kanehisa, M.; KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res. 36, W423-W426 (2008）.