# 生命システム情報統合データベースの構築と ゲノム情報理学の創成

# Biological Systems Database and Genome Information Science

プロジェクトリーダー

金久　實　　京都大学化学研究所・教授

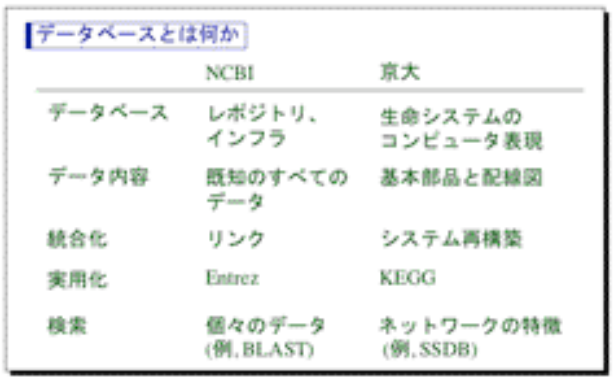
## １．研究目的

ゲノムの情報は、細胞・個体・生物界といった高次レベルの生命システムを原理的に理解するための基盤情報であり、同時に医療や産業をはじめとした応用の可能性をもつ情報でもある。しかしながらこれまでの情報技術では、ゲノムに書かれた個々の遺伝子やタンパク質を解読することはできても、これら基本部品から構成される生命システムとしてのはたらきを直ちに解読することはできなかった。そこで本研究では、ゲノムから細胞レベルでの生命システムのはたらきと有用性を見いだすための知識集約型データベース（生命システム情報統合データベース KEGG）を構築し、生命システムの情報構築原理を理解すると同時に、産業化へつながるバイオインフォマティクス技術を確立する。

## ２．研究成果概要

### ２-１．新しいデータベースの概念

バイオインフォマティクスの究極の目標は、細胞・個体・生物界といった生命システムをコンピュータの中に再現し、高次複雑システムとしての生命現象をゲノムの情報から計算で予測可能にすることである [1]。このような観点から、本研究では「生命システムのコンピュータ表現」というデータベースの新しい概念を提唱し、これを実用化した。



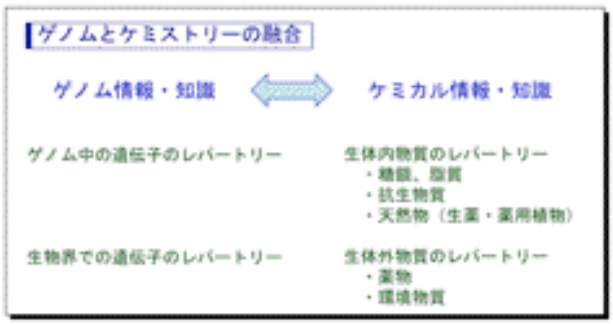

### ２-２．KEGG データベースの構築

KEGG は旧文部省ヒトゲノムプログラム第Ⅰ期の最終年度である 1995 年に特定領域研究「ゲノム情報」の下で開始し、第Ⅱ期に特定領域研究「ゲノムサイエンス」で発展させてきたデータベースである。ミレニアムプロジェクトの一貫として 2000 年に開始された本研究における新しい KEGG は、細胞レベルでの生命システムのコンピュータ表現であり、下図に示したように、ゲノム情報（GENES）とケミカル情報（LIGAND）をネットワーク情報（PATHWAY）で統合した「生命システム情報統合データベース」である [2, 3]。


従来のバイオインフォマティクス技術が、大量データのスクリーニングで有用遺伝子や有用タンパク質など個々の部品を見いだすことに主眼があったのに対し、本研究では部品間の配線図（相互作用ネットワーク）を明らかにすることで、生体システム全体としての機能や有用性を見いだす方法論を開拓した。KEGG に集約された配線図の知識は、代謝再構築をはじめ、ゲノムの情報から細胞レベルでの生命システムの機能解読のため、国際的に幅広く利用されている。

KEGG データベース構築に関して、本研究の具体的な成果は以下の通りである。

(1) ネットワーク（パスウェイ）情報に関しては、代謝系中心の KEGG からシグナル伝達をはじめとした様々な制御系や病気のパスウェイを含む KEGG へと発展させた。また KEGG パスウェイマップの XML 化を実現し、パスウェイデータベースの国際標準となった。
(2) ゲノム情報に関しては、KEGG パスウェイに対応づけてオーソログ遺伝子グループ KO を定義し、ゲノム中の遺伝子に KO づけを行うことで、パスウェイ再構築とそれに伴う高次機能解釈を可能とした。KO づけの自動化システムを広く提供することで、遺伝子アノテーションにおいても KEGG は国際標準になると考えている。
(3) ケミカル情報に関しては、化合物・化学反応に加えて糖鎖の構造情報をデータベース化した [4]。また化学構造比較アルゴリズムを開発して、化合物や糖鎖の類似構造検索を可能とし、さらに生体内化学反応の分類体系 RC を開発して、EC 番号づけの自動化を実現した [5, 6]。このような先駆的研究により、米国 NCBI、欧州 EBI、米国糖鎖コンソーシアムなどとの国際連携が進んでいる。
(4) 幅広い基盤的なデータベースである KEGG を個々のニーズに応じて利用できるよう、標準的なプログラミングインターフェース KEGG API を開発し提供した。




## 2-3．ゲノム情報理学の創成

本研究では、生命システムのコンピュータ表現（オントロジー）について、ネステッドグラフ（階層グラフ）とライングラフの概念を導入した。ネステッドグラフは KEGG パスウェイの階層表現に使われ、ゲノムからパスウェイ再構築と高次機能解釈を行う方法論として実用化した。

もう１つの概念であるライングラフとはノードとエッジを入れ替えたグラフのことで、代謝系において酵素（遺伝子）ネットワークと化合物ネットワークの相補性に関する概念である。これをもとにゲノム中の遺伝子レパートリーから生物が生産し得る物質を予測したり、逆に天然物の構造からゲノム中の遺伝子や合成経路を予測したり、ゲノムとケミストリーを融合した研究を開拓した [4-6]。これは近年のケミカルゲノミクス研究とともに、ゲノムと環境との相互作用を理解するゲノム情報理学の研究領域としてさらに発展しつつある。

## ３．結論

KEGG を中心としたゲノムネットのウェブサイト（http://www.genome.jp/）へのアクセス件数は、本研究開始時に月間 200 万件であったのが終了時には月間 800 万件に達し、５年間で KEGG が飛躍的に発展したことを物語っている。アクセス件数の６ー８割は海外からであり、国際的な知的情報基盤としての地位を確立した。これは下に示した Google のリンク検索（他のサイトからどの程度リンクされているかの目安）でも明らかであり、KEGG は世界で最もよく利用されているデータベースの１つとなった。

Googleによるデータベース引用回数

| データベース | アドレス | 回数 |
|---|---|---|
| NCBI | www.ncbi.nlm.nih.gov | 29,800 |
| ExPASy (SwissProt) | www.expasy.org | 18,300 |
| EBI | www.ebi.ac.uk | 13,200 |
| GenomeNet (KEGG) | www.genome.jp | 9,430 |
| DDBJ | www.ddbj.nig.ac.jp | 620 |
| JSNP | snp.ims.u-tokyo.ac.jp | 55 |
| PDBj | www.pdbj.org | 23 |
| H-invitational | www.h-invitational.jp | 19 |

（2005年7月16日のリンク数検索結果）

## ４．主な発表論文


1. Kanehisa, M. and Bork, P. (2003) Bioinformatics in the post-sequence era. *Nat. Genet.* **33**, 305-310.
2. Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42-46.
3. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277-D280.
4. Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K.F., Ueda, N., Hamajima, M., Kawasaki, T., and Kanehisa, M. (2005) KEGG as a glycome informatics resource. *Glycobiology*, in press.
5. 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**, 11853-11865.
6. 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**, 16487-16498.