ゲノムと環境の統合解析による生命システムの機能解読（KEGG）

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
東京大学医科学研究所ヒトゲノム解析センター
金久　實

Deciphering Systemic Biological Functions by Integration of Genomic and Environmental Information (KEGG)

Minoru Kanehisa
Bioinformatics Center, Institute for Chemical Research, Kyoto University
Human Genome Center, Institute of Medical Science, University of Tokyo

KEGG is a database resource that integrates genomic, chemical, and systemic functional information. In particular, gene catalogs in the completely sequenced genomes are linked to molecular networks (pathways, modules, and brite hierarchies) representing higher-level systemic functions of the cell, the organism, and the ecosystem. In addition to expanding the knowledge base for such molecular networks and improving the linking (annotation) procedure, we have incorporated drug and disease information from the point of view of perturbed molecular networks. As the result, KEGG has become an international standard for integration and interpretation of large-scale datasets generated by genome sequencing and other high-throughput experimental technologies.

1. はじめに

日常的な病気は様々な遺伝因子と環境因子が複雑に絡み合った多因子性疾患である。ヒトの健康や病気を理解するには、これらの因子を含む生体システムが安定した状態にあるかゆらいだ状態にあるか、システム的な観点でのアプローチが必要である。本研究では、ゲノム情報とケミカル情報に関する大量データから、生命システムと環境との相互作用を理解し、医療、創薬、環境保全に役立てることを目的として、生命システム情報統合データベース KEGG の高度化を行った [1][2]。KEGG では高次生命システム機能に関する知識を、広い意味での分子ネットワークとしてコンピュータ化することで、分子レベルの大量データとの統合処理を可能にしたリソースである。本研究では高次生命システム機能をパスウェイマップで表現した KEGG PATHWAY と階層リストで表現した KEGG BRITE を大幅に拡張し、新たに単純リストで表現した KEGG MODULE の開発を行った。またゲノムと分子ネットワークをつなぐ KEGG ORTHOLOGY (KO)システムの充実と、これに基づくアノテーション手続きの自動化を進め、ゲノムが解読されたすべての生物種に対応し続けている。日米欧の医薬品情報を化学構造をベースに一元的に集積・管理している KEGG DRUG では、ターゲット分子、薬物代謝酵素をはじめ、分子間相互作用に関する知識を集約した。ヒト

疾患については、ゆらぎ状態の分子ネットワーク（疾患パスウェイマップ）表現とともに、分子ネットワークが未知の大多数の疾患情報をコンピュータ化するため、単一遺伝子疾患・多因子性疾患と病因遺伝子との関連、感染症疾患と病原体ゲノムとの関連を蓄積した KEGG DISEASE の開発を行った。KEGG と NCBI との緊密な連携体制ができたこと（NCBI からのリンクは 1,000 万件）もあり、KEGG ウェブサイトには月間 20 万近くのユニークビジター（UniProt とほぼ同じ）がある。また 2 年ごとに発表している Nucleic Acids Research 誌 Database 特集号での KEGG 論文全体の引用回数は年間 1,000 件（Pfam についで 2 位）に達している。本研究により、KEGG はゲノム、トランスクリプトーム、メタボローム、ケミカルゲノム、メタゲノムをはじめとした大量データを解釈するための国際標準データベースに発展させることができた。

2. 研究開発の成果

2.1 KEGG データベースの内容

KEGG は PATHWAY/BRITE をはじめとしたシステム情報、GENES/GENOME をはじめとしたゲノム情報、COMPOUND/REACTION をはじめとしたケミカル情報からなる統合データベースである。表 1 に現時点でのデータベースの内容を示し、本研究開始年度末とのデータ数の比較を示した。生体システムのゆらぎ状態とゆらぎ物質の観点から、DISEASE/DRUG はシステム情報の一部としてデータベース化を行っている。

表 1. KEGG データベースの内容

データベース	内容	2011.1.21	2007.3.2
KEGG PATHWAY	パスウェイマップ	382(126,803)	(47,466)
KEGG BRITE	機能階層・オントロジー	96(36,155)	(5,821)
KEGG MODULE	KEGG モジュール	362	-
KEGG DISEASE	ヒトの病気	375	-
KEGG DRUG	医薬品	9,316	4,547
KEGG EDRUG	生薬・天然物	834	-
KEGG ORTHOLOGY	オーソログ (KO) グループ	14,139	9,574
KEGG GENOME	KEGG 生物種	1,516	547
KEGG GENES	高精度ゲノム中の遺伝子	6,154,780	1,980,702
KEGG DGENES	ドラフトゲノム中の遺伝子	372,418	589,268
KEGG EGENES	EST コンティグとしての遺伝子	3,792,883	448,730
KEGG MGENES	メタゲノム中の遺伝子	669,846	-
KEGG COMPOUND	代謝物質、その他の化学物質	16,439	14,397
KEGG GLYCAN	糖鎖	10,976	10,951
KEGG REACTION	化学反応	8,397	6,870
KEGG RPAIR	反応ペアと化学構造変化	12,457	-
KEGG RCLASS	反応クラス	2,306	-
KEGG ENZYME	酵素	5,296	4,673

（注）PATHWAY と BRITE では手作業で作成するレファレンスマップ、レファレンス機能階層の数と、生物種展開をしたものを含めた数を括弧内に示した

本研究で新規に作成したパスウェイマップ（薬の構造マップを含む）の数は現在の約 1/3 の 130、そのうち疾患パスウェイマップが 32、代謝系のグローバルマップが 3 あり、新規 BRITE 機能階層の数は現在の約半分の 50 であった。生物種の数とアノテーションを行っている総遺伝子数は５年間で３倍以上に増えている。詳細な更新履歴は以下の URL を参照していただきたい。

KEGG パスウェイマップ	http://www.genome.jp/kegg/docs/upd_map.html
KEGG DRUG 構造マップ	http://www.genome.jp/kegg/docs/upd_drugmap.html
BRITE 機能階層	http://www.genome.jp/kegg/docs/upd_htext.html
KEGG 生物種	http://www.genome.jp/kegg/docs/updnote.html

ここでは以下に代表的な研究開発の成果を４つ紹介する。

2.2 グローバルマップ

グローバルマップは、代謝の全体像を把握できるように、従来の多数のパスウェイマップを手作業でまとめたものである [2][3]。一次代謝のグローバルマップの他に、植物や微生物による二次代謝物質生合成マップ、微生物による環境物質分解や多様なエネルギー代謝などのマップも提供している。グローバルマップはメタゲノム解析などで広く利用されており、図１はその例で、ヒトと腸内細菌叢の複合代謝能力を、緑はヒト、赤は腸内細菌叢、青は両方がもつパスウェイとして表現している。

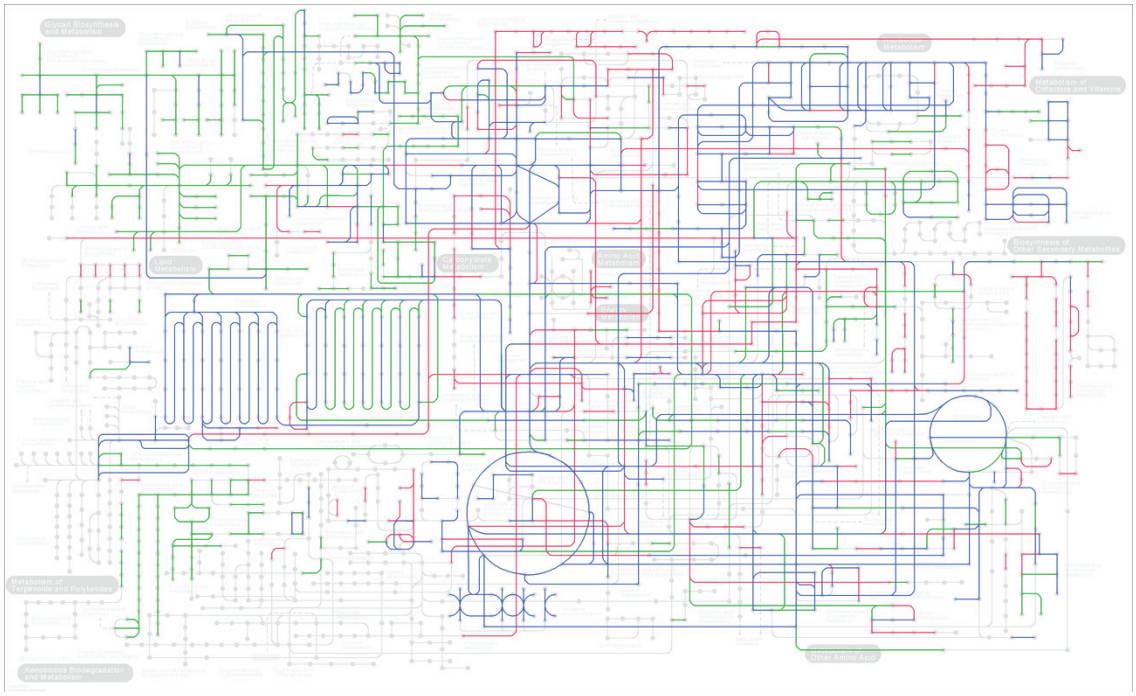

図１．ヒトゲノムと腸内細菌メタゲノムから推定される複合代謝経路

2.3 疾患・医薬品に関する知識集約

医薬品データベース KEGG DRUG は、日米欧の医薬品情報を、化学構造と成分の観点から一元的かつ網羅的に集積・管理したデータベースである。本研究開発では、薬と生体分子ネットワークとの相互作用情報として、ターゲット分子、薬物代謝酵素とトランスポー

ター、ゲノムバイオマーカーなどの遺伝子情報を主に文献から集約し入力した。また JAPIC 添付文書より薬物間相互作用を抽出し、ターゲットの重複、代謝酵素の重複といった形で意味づけを行った[4]。一方、新たに開発を始めた疾患データベース KEGG DISEASE は、疾患パスウェイマップ表現ができない（分子ネットワークが未知の）大多数の疾患を、遺伝子・ゲノムの情報と関連づけ、大量データとの統合処理ができるようにしたデータベースである。単一遺伝子疾患および多因子性疾患では既知の遺伝因子や環境因子などとの関連づけを、感染症疾患では既知の病原体ゲノムおよび病原性因子との関連づけを行っている。

図 2 は薬とターゲットの関係、疾患と病因遺伝子の関係をパスウェイマップに重ね合わせたものである。この例は Alzheimer 病のマップで、既知の病因遺伝子（赤字）以外に、関連した神経変性疾患の病因遺伝子（背景色がピンク）も多数マップ上に存在し、これらの相関関係が推定される。

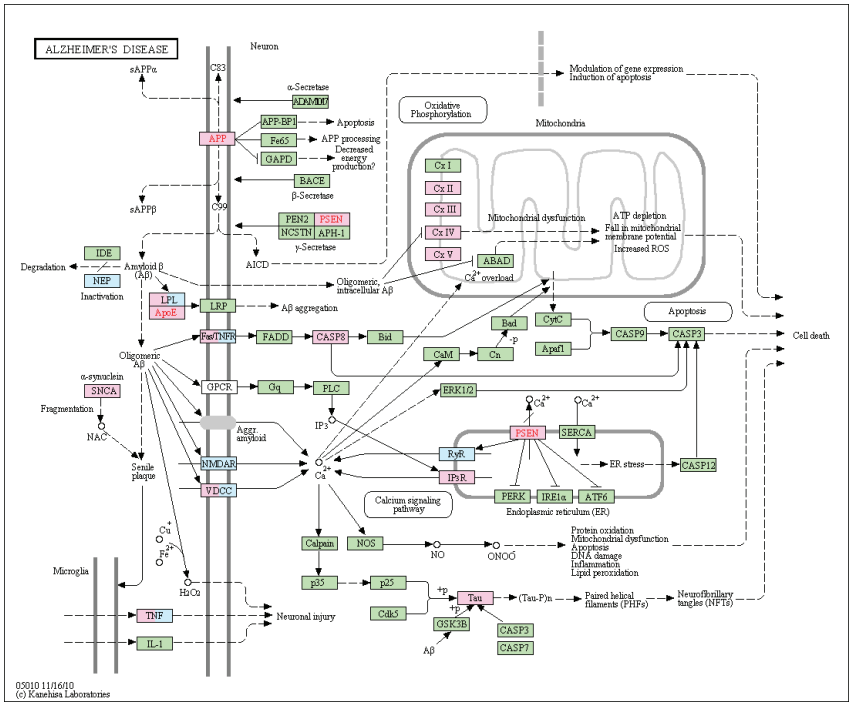

図 2． KEGG disease/drug マッピングの例

2.4 環境物質に関する知識集約

本研究開発では、環境保全やエネルギー資源の観点で、微生物がもつ代謝能力とゲノムとを関連づける知識集約も行った。微生物による環境物質の分解経路は、当初 UMBBD データベースをもとに作成し、統一性に欠けていたので、これを全体的に見直して 12 ほどのパスウェイマップにまとめ直した。さらにこれらを様々な環境に特化したエネルギー代謝とともにグローバルマップ化した。エネルギー代謝の部分は国際的なメタゲノムコミュニティからの要望に応える形で開始したものである。環境物質そのものについては、内分泌攪乱物質の BRITE 機能階層といった形で、化学構造と関連づけて集約している。また、環境物質の分解経路や植物二次代謝物質の合成経路を予測するソフトウェア開発も行った[5]。

2.5 KEGG アノテーションパイプライン

KEGG GENES には高精度ゲノム配列が決定された 1400 生物種の合計 600 万を越える遺伝子（non-coding RNA も含む）が蓄積され、その約 4 割（ヒトは 46%、大腸菌は 69%）にアノテーションがつけられている。これはパスウェイマップ、モジュール、BRITE 機能階

層といった高次レベルの情報に基づく KEGG ORTHOLOGY (KO)システムを用いた独自のアノテーションで、ゲノム単位でなく、種間のオーソログ単位に行っている。本研究ではKEGG アノテータのノウハウをコンピュータ化した KOALA を開発し[1]、KO システムの改良とともに、KOALA による自動アノテーションの比率（現状では原核生物で約 7 割）も順次向上している。自動アノテーションとそれを補う手動アノテーションをパイプライン化することで、作業効率は飛躍的に向上した。

3. まとめ

次世代シークエンサーに代表されるハイスループット実験技術の進歩に伴い、多数の生物種ゲノム、個人ゲノム、メタゲノムなどの配列データを迅速かつ大量に決定することが可能となり、そこに含まれる遺伝子の機能、さらにはその総体としての高次生命システム機能を自動的に解読する情報技術の必要性が高まっている。KEGG はこのようなニーズに対応し、ゲノム機能解読の国際標準として広く利用されるようになった。その最大の要因は人手による知識集約作業である。これは多くの関連論文にある知識をまとめる作業、すなわち総説を書くような作業であるが、文章で表現するのではなく、分子レベルでコンピュータ処理できる表現（パスウェイマップ、階層リスト、単純リスト）にしている所に特色がある。その際、実験データは取捨選択が必要で、例えば KEGG DRUG にあるターゲット情報は薬効との関連がある生理機能分子に限定し、単に相互作用するものをすべてターゲットとするのではない。散在する多様なデータをコンピュータ処理で統合し提供しているデータベースは多数存在するが、品質の評価無しに集めれば利用者に誤った情報を与えかねない。表 1 にある多様なデータについて高品質の知識集約と統合化を行ったことで、KEGG は国際的な信頼を得ることができたと考えている。

4. 研究開発実施体制

代表研究者　金久　實（京都大学化学研究所）

(1)　KEGG 生命・環境グループ
　　グループリーダー　金久　實（京都大学化学研究所・教授）

(2)　KEGG 創薬・医療グループ
　　グループリーダー　金久　實（東京大学医科学研究所・教授）

5. 参考文献

[1] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M.; KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38, D355-D360 (2010).

[2] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y.; KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, D480-D484 (2008).

[3] 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).

[4] Takarabe, M., Shigemizu, D., Kotera, M., Goto, S., and Kanehisa, M.; Characterization and classification of adverse drug interactions. Genome Informatics 22, 167-175 (2009).

[5] 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, W138-W143 (2010).