2項関係に基づくゲノムと生命システムの機能解読

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 biological functions of genomes and biological systems by binary relations

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

BRITE is a collection of hierarchical classifications (controlled vocabularies) representing our knowledge on various aspects of biological systems, categorized into (i) genes and proteins, (ii) chemical compounds and reactions, (iii) drugs and diseases, and (iv) cells and organisms. In contrast to KEGG PATHWAY, which is limited to molecular interactions and reactions, BRITE incorporates many different types of relationships (binary relations). BRITE is now part of the KEGG suite of databases and the mapping of genomic and molecular data to KEGG BRITE supplements the KEGG PATHWAY mapping for inferring higher-order functions. A standalone Java application for browsing BRITE files has also been developed.

1. はじめに

ヒトからバクテリアまで数多くの生物種において全ゲノム配列が続々と決定され、ゲノム情報を基盤に細胞、個体、生態系といった高次生命システムの理解が進み、同時に創薬、医療をはじめとしたゲノム情報の有効利用が活発化している。これまでに決定されたゲノムの配列情報は国際 DNA データベース DDBJ/EMBL/GenBank に登録され、誰でも自由に利用できる形になっている。しかしそれだけでは不十分であり、ゲノムに書かれた生命のはたらきや有用性を見いだすことを可能にするデータベースがなければならない。生物情報データベース高度化の一貫として、本研究では新しいタイプの機能情報データベース BRITE を構築した。

機能情報のデータベース化については、大きく2つのやり方がある。1つは配列データベースのアノテーションのように、言葉として記述することである。機能情報はただ利用者が読んで理解できればよいとするのなら別であるが、一般には生物種間の比較をしたり、コンピュータ処理をするために、語彙の標準化を行わなければならない。また、分子、細胞、個体といった生命システムのどのレベルでの機能なのか、機能情報の階層化を行う必要がある。これは狭い意味でのオントロジーの問題である。GO（Gene Ontology）では、異なる生物種での遺伝子アノテーションの標準化と知識の共有のために、Biological process、Cellular component、Molecular function の3つの観点（オントロジー）で語彙の階層的な定義を行っている。

もう1つは我々が KEGG において提唱し実践しているやり方で、細胞レベルの生命システムの「はたら

き」を分子間相互作用ネットワークの「かたち」として表現する。ゲノムの遺伝子の並びからタンパク質同士のつながり方（かたち）を予測し、すなわち KEGG パスウェイを再構築し、そこから例えばリジンを合成するはたらきがある、浸透圧変化に応答するシグナル伝達のはたらきがあると判定する。主観的な機能情報を客観的な形の情報に置き換えているところが KEGG の特色である。これは逆の言い方をすると、KEGG のネットワーク表現ができる機能情報とは、分子間ネットワークが解明されたものに限られるわけで、例えばこの遺伝子は細胞周期に関与しているらしいといった手がかり程度では、KEGG では表現ができない。

本来、機能とは曖昧なものであり、大雑把な手がかりであっても有用性はあり得る。そこで本研究の BRITE では、上記 GO のやり方と KEGG のやり方を、語彙の2項関係で融合する。2項関係とは2つのオブジェクト間の関係情報であり、KEGG パスウェイにおける分子間の関係、オントロジーの階層における親子関係、さらには「かたち」と「はたらき」の関係の例として、配列や立体構造と分子機能との関係、ネットワークと細胞機能との関係、といった様々な関係を含めて考える。2項関係の集合はグラフであり、語彙で表現された様々なオブジェクトがそのノードとなる一般的なグラフを考えていることになる。BRITE データベースは階層テキストファイルと呼ぶ多数のテキストファイルの集合で、基本的には GO と同じ DAG（Directed Acyclic Graph）で表現されている。しかし階層に属すメンバーは単なるリストではなくサブグラフである点で DAG より一般的な形をしており、このような構造は後述するようにメタグラフと呼んでいる。なお本研究開発の最終年度に、BRITE は KEGG BRITE [1] として KEGG の中に取り込み、

http://www.genome.jp/kegg/brite.html

から利用できるようになっている。

2. 研究開発の成果

2.1 メタグラフ表現と抽象化レベル

グラフとはノードとエッジの集合、あるいは2つのノードがエッジでつながった2項関係の集合である。図1(a) の赤いノードのように、ノード自身がグラフであるグラフのことをネステッドグラフ（または複合グラフ）という。ネステッドグラフは抽象化レベルの概念と深い関係がある。タンパク質の立体構造は原子をノードとし原子間結合をエッジとしたグラフとみなすことができ、例えば受容体タンパク質とそのリガンドの相互作用を詳細に解析するには原子レベルでの記述が必要である。一方、異なる生物種間で同一機能をもつオーソログタンパク質を見いだすには、原子レベルでの詳細は抽象化し、各アミノ酸構造を20種類の文字に置き換えて配列解析を行う。すなわち各アミノ酸構造のサブグラフを1つのノードに変換したネステッドグラフとして、上位のグラフ構造（配列）のみ考えるわけである。さらにタンパク質間相互作用ネットワークや酵素の反応ネットワークを考える際には、タンパク質全体を1つのノードとした、さらに上位のネステッドグラフ構造を考える。このように下のレベルでの構造の詳細を抽象化することで、高次レベルでの構造の特徴をよりよく把握することができるわけである。

実際の生物学の問題では、階層化すなわちサブグラフのグルーピングは排他的にならず、しばしば重複が起こる。図1(a) の緑で囲んだサブグラフはノードを共有しており、このような重複を許したネステッドグラフを「メタグラフ」と呼ぶことにする。これは GO などで用いられている DAG と類似の構造である。図2(b) に示したように、単純なツリー構造でなく、複数の親をもつ子が存在する構造が DAG である。

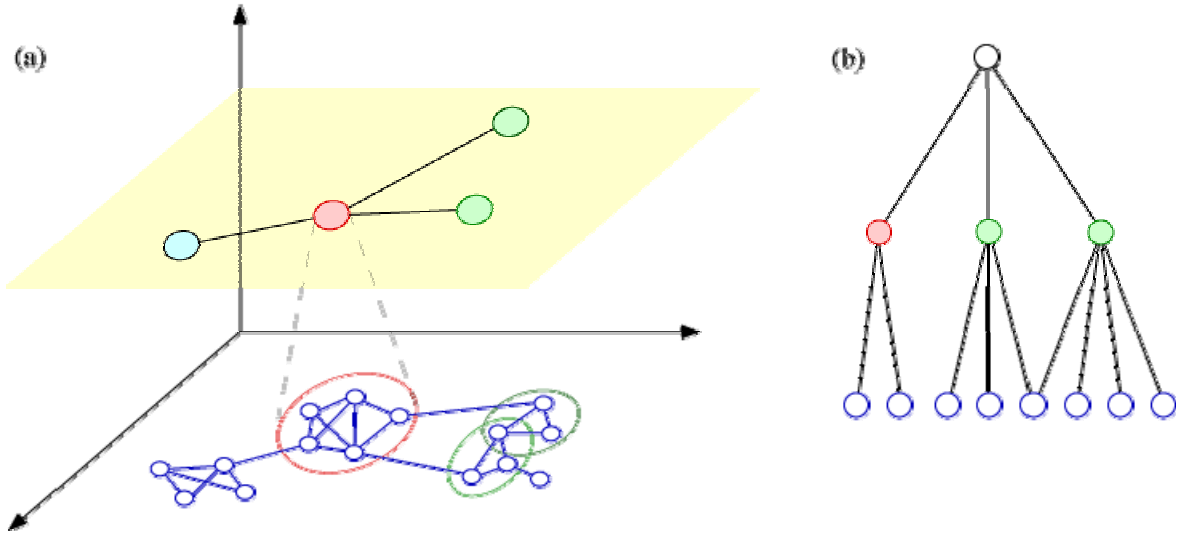

図1．メタグラフと DAG

2.2 KO によるゲノム情報の体系化

KEGG のパスウェイマップは分子間相互作用・反応ネットワークの「はたらき」を表現するものであるが、実際に描かれているのはネットワークの「かたち」である。ネットワークレベルでの構造・機能相関を明確にすることが BRITE の1つの目的であり、そのための機能階層が KO (KEGG Orthology) である。KO は BRITE の Genes and Proteins カテゴリー（表1参照）に定義されている。表1の Network hierarchy は KEGG PATHWAY の分類に従ったネットワーク階層を示しており、GO の Biological process と対比して考えることができる。また Protein families は GO の Molecular function に対応するが、GO のように1つの大きな階層にまとめるのではなく、異なる観点での分類や異なる分野のエキスパートの知識をよりよく反映させるために別々の階層となっている。

表1．BRITE データベースの内容（2006年2月現在）

Genes and Proteins 　Network hierarchy 　　KO 　Protein families 　　Enzymes 　　Cytochrome P450 　　Transcription factors 　　Ribosome 　　Translation factors 　　ABC transporters 　　G-protein coupled receptors 　　GTP-binding proteins 　　Ion channels 　　Cytokines 　　Cytokine receptors 　　Cell adhesion molecules (CAMs) 　　CAM ligands 　　CD molecules 　　Bacterial motility proteins	Compounds and Reactions 　Compounds 　　Compounds with biological roles 　　Lipids 　　Phytochemical compounds 　Compound interactions 　　Ion channel agonists/antagonists 　　Cytochrome P450 substrates Drugs and Diseases 　Drugs 　　Therapeutic category of drugs 　　Drug classification 　Diseases 　　Infectious diseases Cells and Organisms 　Organisms 　　KEGG organisms

これら KO 階層の一番下のレベルは K 番号で識別される KO エントリー、すなわちオーソログ遺伝子グループのエントリーに対応している。KEGG ではアルファベット1文字に続く5桁の数字が識別子となっており、K 番号以外に、化合物の C 番号、薬の D 番号、糖鎖の G 番号、反応の R 番号、化合物ペア（アライメント）の A 番号などがある。これまで、KEGG ではパスウェイ情報に基づき、オーソログの定義がされてきた。K 番号はパスウェイマップのボックス（遺伝子産物）につけられた識別子であると同時に、そこに対応づけられるゲノム中のオーソログ遺伝子の識別子でもある。これが KO の Network hierarchy の部分であるが、パスウェイが既知という大きな制約のため、ゲノム中でカバーされる遺伝子数が限られるという大きな問題があった。そこで、オーソログ遺伝子グループに関する知識をより網羅的に集約するため、2つの観点を取り入れた。

1つは表1にもあるタンパク質ファミリーの知識である。例えば G-protein coupled receptors のように、パスウェイマップにはごく特定のメンバーしか出現しなくても、配列その他の観点で系統的に分類されたタンパク質ファミリーは数多く存在する。それらを本研究チーム内部で再検討し、各ゲノム中の遺伝子との対応づけを行っている。もう1つは NCBI の COG を参考にしたオーソロググループの定義で、機能未知タンパク質も含まれる。機能的に細分化された既知タンパク質ファミリーの分類と比べて、COG は大雑把な分類である。KO システムにはこのように細分化の度合いの異なる分類が混在しており、1つの K 番号が表すオーソロググループに階層が存在する場合も少なくない。例えば、EC 番号の4桁目に対応した基質特異性まで考慮した K 番号と、それを指定しない（4桁目がハイフンになった）K 番号とがある。細分化された K 番号がより有用なのは言うまでもないが、それが分からない場合に大雑把な K 番号はある程度の機能的手がかりを与える。

2.3 パスウェイ再構築と高次機能の推論

GO（Gene Ontology）が遺伝子アノテーションを記述する語彙の体系にオントロジーという言葉を使って以来、図書館情報のような階層分類を何でもオントロジーと呼ぶケースが増えている。本来、オントロジーでは階層に is-a や part-of といった関係を定義し、高度な検索や推論を可能とするためのものである。本研究でも当初はこのようなオントロジーの開発を計画したが、KEGG のメタグラフを利用することへ計画変更を行った。その理由は以下の通りである。

人間の知識を階層的に表現し、下位の概念（例えばゲノムの配列情報）から上位の概念（例えば細胞や個体の機能情報）をコンピュータで推論することは可能だろう。しかしながら、オントロジーの体系に入っていない新たな知識の発見は可能ではない。KEGG においてはすでにメタグラフの概念に基づき、下位構造から上位構造への階層を推定するパスウェイ再構築の方法論が実現されている。これまでにも、配列モチーフや立体構造モチーフといった構造の特徴が分子レベルでの機能につながることがよく知られている。そこで、パスウェイモジュールあるいはネットワークモチーフといった構造の特徴と高次レベルの機能との関係を定義することで、KEGG のパスウェイ再構築に機能的な意味づけが可能となる。また、KEGG のメタグラフはゲノムや生命システムに内在する分子としての関係に基づいており、語彙の関係を定義するオントロジーと比較して主観や曖昧さがはるかに少ない。さらに、未知の構造モチーフを探すことで新たな知識発見も可能である。BRITE ではこのような理由から、オントロジーの体系ではなく、メタグラフの構造体系を用いて推論を行うこととした。

具体的には BRITE におけるゲノムと生命システムの機能解読は以下の手続きをとる。

(1) ゲノム中の遺伝子に K 番号づけを行う。

(2) これをもとに KEGG パスウェイへのマッピングを行う。

(3) 再構築された KEGG パスウェイの機能的意味づけを行う。

ステップ(1)と(2)は KAAS (KEGG Automatic Annotation Server) により自動的に行うことができるようになった。(3)を実現するのが KO の語彙の階層であるが、現時点では機能的な単位に相当するサブグラフ(パスウェイモジュール)の定義が不十分で、高次機能解読の自動化までは達成されていない(人間が色づけされた KEGG パスウェイマップを目で見て解釈することはできる)。本研究においてメタグラフに基づく方法論は完成したので、今後は KO の語彙を精密化することで、推論の自動化という長期目標が達成可能であると考えている。

2.4 ケミカル情報の体系化

ゲノムから高次生命システムの機能を解読することとともに、BRITE のもう1つの大きな目的は、生命システムと環境との相互作用の理解へつながる、ゲノム情報とケミカル情報の関連解析及び経験則の発見である(図2)。BRITE におけるケミカル情報の体系化には大きく3つの側面がある。第1に化合物や糖鎖といった物質の階層分類、第2にレセプタやイオンチャネルといったタンパク質とリガンドとなる化合物等との相互作用に関する階層分類、第3に生体内化学反応の体系分類 RC (Reaction Classification) システムである。ゲノムとケミストリーの融合で最も成果があったのは糖鎖の分野で、糖鎖構造を糖転移酵素の反応の集まりとして表現することで、データ収集と実用的なツール開発を行った [2]。

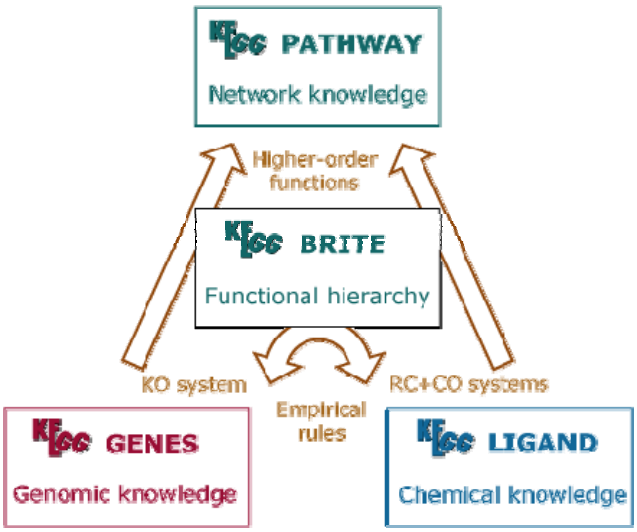

図2. KEGG の一部としての BRITE

RC システムは既知の酵素反応における化学構造変化のパターンを体系化したもので、KEGG LIGAND の ENZYME データベースから以下の手順で作られている。

(1) 既知の酵素が触媒する各反応を REACTION データベースに登録する。

(2) 各反応を基質と生成物のペア(一般に複数ある)に分解し RPAIR データベースに登録する。

(3) 各基質・生成物ペアでの構造変化パターンをアライメントを作って抽出する [3]。

(4) 構造変化パターンを RDM パターンと呼ぶ KEGG atom type の変化で分類する [4]。

現在、RDM パターンの類似性を階層的に表現し、RC として集大成する作業を行っている。これを酵素のオーソログ・パラログ遺伝子の階層(KO の階層)と比較することで、既知反応と既知遺伝子の間の経験則を明らかにし、未知反応や未知遺伝子の予測へつなぐことが可能になると考えている。

2.5 薬と病気の情報及びその日本語化

本研究開発の大きな成果は薬と病気に関する知識の体系化が進んだことである。日本及び米国で認可されている薬はこれまでにも KEGG COMPOUND に一部登録されていたが、これを KEGG DRUG として独立させ、化学構造だけでなく薬効、ターゲット情報なども付加して蓄積している。また薬の開発の歴史や基本骨格に関する知識を図3のような KEGG DRUG structure map として表現している。この図の背後にある化学構造変化のパターンを RC のような形で体系化することで、新たな創薬につながる経験則を探索する予定である。

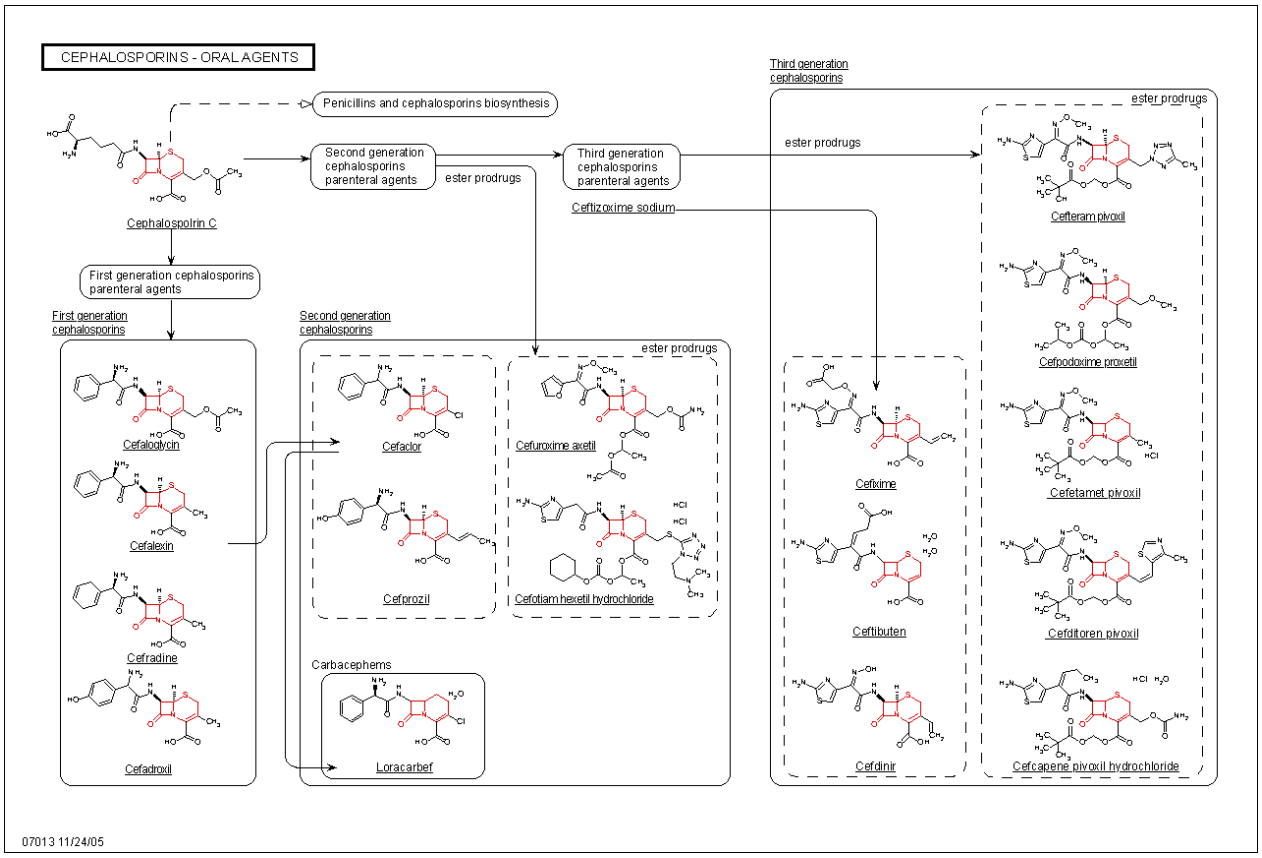

図3. セフェム系抗生物質に関する KEGG DRUG 構造マップの例

日本で認可された薬には階層的な薬効番号がつけられているので、これを薬効分類の BRITE ファイル(階層テキストファイル)として集約している。階層テキストファイルは従来から KEGG の Web サービスの中で get_htext と呼ぶプログラムで表示・検索できるようになっていたが、本研究で新たに KegHier と呼ぶ Java アプリケーションを開発した。これは Mac や Windows など自分のパソコンにダウンロードして利用する形態をとり、図4に示したように iTunes のようなインターフェースをもっている。iTunes でミュー

ジックストアから曲を取得するのと同じように、KegHier では KEGG の Web サイトからファイルを取得して表示したり、自分で作成したファイルを KEGG その他のインターネットリソースと統合して利用したりできるようになっている。

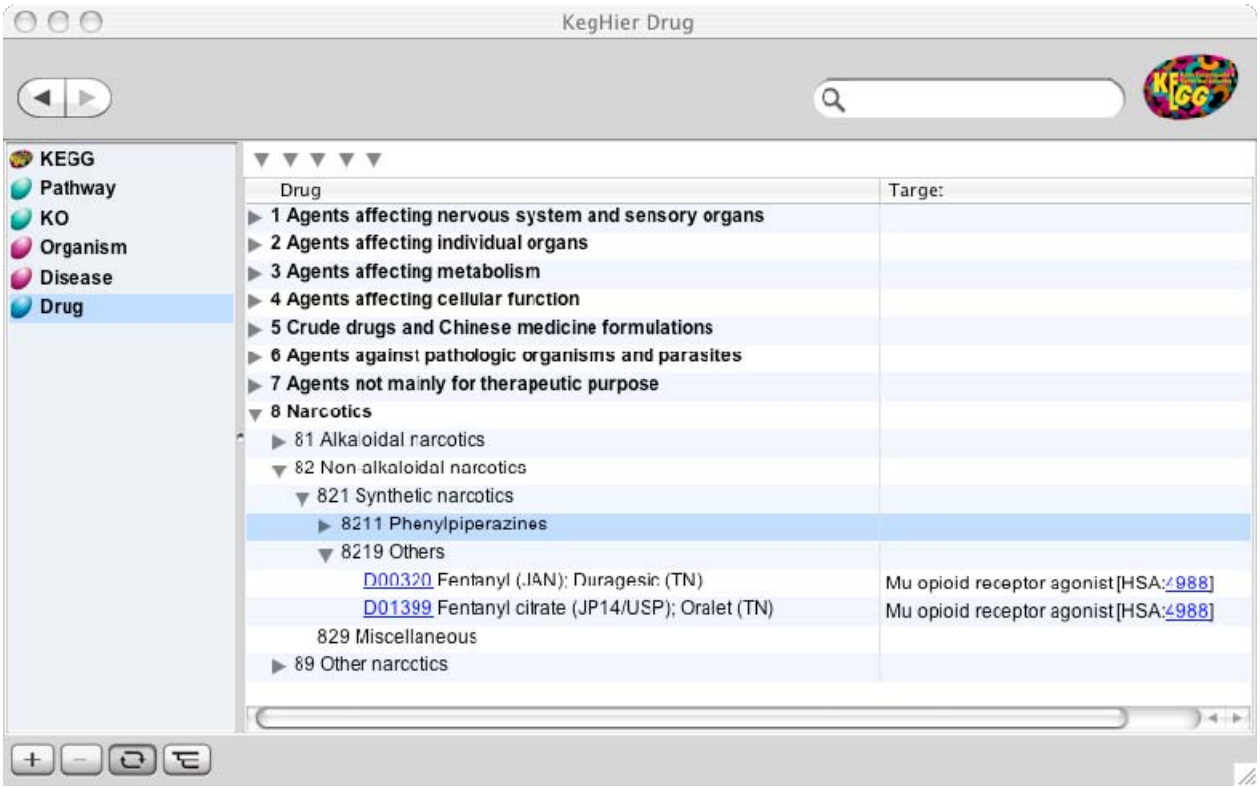

図4. KegHier による薬効分類とターゲット情報の表示

病気に関しては主に感染症を対象とし、続々と全ゲノム配列が決定されている病原微生物と関連づけ、また ICD など既存の病気分類と対応づけ、複数の BRITE ファイルとして知識集約を行っている。また、コレラ菌及びピロリ菌については感染経路のパスウェイマップも作成した。その他、代謝疾患についても、Ⅰ型糖尿病、Ⅱ型糖尿病、若年者糖尿病の成熟発現（MODY）などのパスウェイマップを作成した。薬、病気、生物種に関する BRITE ファイルは日本語化も行っており、KEGG システムへの日本語インターフェースの役割を果たしている。

KegHier が機能的に get_htext より優れている点は、階層テキストファイルにタブ区切りでフィールド分割をして、すなわち Excel ファイルに階層を付け加えた感じで、利用できる点である。図4にあるように BRITE の Drug ファイルには薬とそのターゲットの関係が含まれている。また感染症の Disease ファイルでは病原生物とそのベクターの関係も含まれている。BRITE の階層は個別にブラウズするためだけでなく、複数の階層を重ねあわせて、上記以外にも例えば RC と KO、レセプタの階層とリガンドの階層などを重ね合わせて解析し、経験的な法則を見いだすことができるようにしたいと考えている。BRITE が GO のような DAG としてではなく、メタグラフとして構築されていることは以上のことからも明らかであろう。

3. まとめ

BRITE データベースの最初の構想は1994年12月に分子生物学会年会シンポジウムで発表した (http://www.genome.jp/Japanese/docs/mbs94.html) もので、これは1995年5月に開始した KEGG よりも古い (http://www.genome.jp/Japanese/docs/mbs95.html)。その背景には1990年代の初めに研究代表者が関与した第5世代コンピュータプロジェクトでの推論マシンの考え方がある。人間の知識を関係という形でコンピュータに蓄積し、関係を組み合わせて推論を行うといった演繹データベースの考え方である。しかしながら、当初の構想の多くは KEGG に取り込んで実現されていき、BRITE の位置づけも変遷をとげてきた。本研究開発事業ではじめて BRITE に本格的に取り組むことが可能となり、すでに国際的なデータベースとなっている KEGG の高度化・標準化の観点から、すなわち KEGG のネットワーク構造の階層に基づく高次機能の推論と、KEGG パスウェイでは表現できない知識のコンピュータ化を、階層テキストファイルという語彙の体系で実現した。

4. 研究開発実施体制

代表研究者　　金久　實

(1) BRITE 分子グループ

グループリーダー　金久　實（京都大学化学研究所・教授）

(2) BRITE 疾患グループ

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., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M.; From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354-357 (2006).

[2] Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K.F., Ueda, N., Hamajima, M., Kawasaki, T., and Kanehisa, M.; KEGG as a glycome informatics resource. Glycobiology, in press (2006).

[3] Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M.; 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 (2003).

[4] Kotera, M., Okuno, Y., Hattori, M., Goto, S., and Kanehisa, M.; Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. J. Am. Chem. Soc. 126, 16487-16498 (2004).