生体分子情報データベースの開発

Development of Database for Biomolecular Information

化学研究所バイオインフォマティクスセンター化学生命科学　五斗進

背景と目的

近年のゲノム関連情報解析技術の発展により、ゲノム、メタゲノム、トランスクリプトーム、プロテオーム、メタボロームなどの大量の情報が得られるようになってきた。これらは単に生体分子の情報というだけでなく分子間の関連情報という観点から、新しいタイプの情報でもある。これらを効率よく管理し、そこから新しい生物学的知見を発見するためのツールを備えたデータベースの開発はバイオインフォマティクス分野での課題の一つである。我々は、生体分子情報データベースおよびバイオインフォマティクス技術の開発に取り組み、その成果をゲノムネット（http://www.genome.jp/）で広く公開している。特に、DBGET/LinkDB と KEGG (Kyoto Encyclopedia of Genes and Genomes) はその中核をなすものである。本研究では、ゲノムネットにおけるデータベースおよびシステムの改良を行う。また、データベースを用いた解析として、ネットワークという観点から遺伝子の機能予測や創薬などの応用に結びつけることも目標としている。

検討内容

平成 24 年度も平成 23 年度に引き続き、化合物・反応・遺伝子・ネットワークに関するデータベースの拡張と解析を中心に以下の内容を検討した。

1) DBGET/LinkDB の拡張
2) ゲノムネット計算ツールの拡張
3) KEGG の拡張
4) 医薬品相互作用ネットワークの解析

結果と考察

1) DBGET/LinkDB の拡張

LinkDB はデータベースエントリー間の関係を抽出しデータベース化したものであり、database1 の entry1 と database2 の entry2 に何らかの関係がある場合に以下のような 3 項関係で表現している。

database1:entry1　　database2:entry2　　original

ここで第 3 項は 2 つのエントリー間の関係を表すリンクのタイプであり、例に挙げている original は

database1 の entry1 に database2 の entry2 へのリンクが記載されていたことを示す。平成 24 年度は、従来から提供していた 3 つのタイプ original、reverse（original と逆向きのリンク）、equivalent（異なるデータベース間で同じものを示すリンク）の最新データへの定期的な更新を行うとともに、新しいタイプとして indirect を提供するようにした。indirect リンクは複数のデータベース間にまたがるリンクを、あらかじめ計算で求めておいて検索できるようにしたものである。現在までに、KEGG の遺伝子情報（KEGG GENES）から関連する酵素反応情報（KEGG REACTION）とそれらに関与する化合物の情報（KEGG COMPOUND）を参照できるようにした。

また、LinkDB で提供しているデータベースの一覧を図で分かりやすく表示し、そこから指定した 2 つのデータベース間のリンク情報をダウンロードできるインタフェースを構築して、公開した（図 1）。今後は、LinkDB の情報を高度なデータベース検索へと応用するための仕組みを開発する予定である。

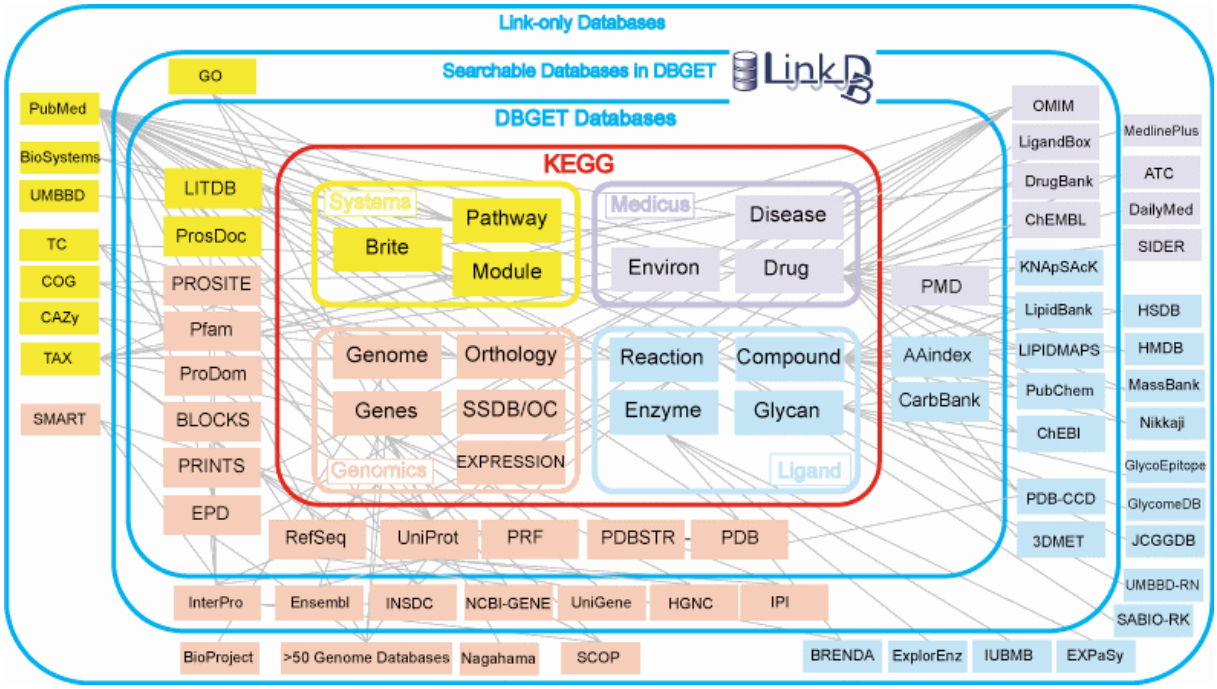

図 1. LinkDB データベースリンク図

2) ゲノムネット計算ツールの拡張

ゲノムネットではゲノムネット計算ツールとして BLAST などの配列解析ツール以外に、遺伝子機能自動アノテーションシステム KAAS などのゲノム解析ツール、化合物の類似構造・部分構造検索システム SIMCOMP/SUBCOMP などの化学解析ツールを開発・提供している。

平成 24 年度はゲノム解析ツールの新規開発と拡張を中心に進めた。近年の次世代シーケンサーによるゲノム・メタゲノム解析では短い（不完全遺伝子長の）塩基配列データから機能予測する必要がある。また、環境メタゲノムデータでは近縁の生物種がデータベースに登録されていない場合もある。そこで、上記 2 点について KAAS で正確に機能予測できるかを評価し、100 アミノ酸配列程度の遺伝子があれば門レベルでの近縁生物種がデータベースになくても 70%程度の精度は得られることを確認

した。この結果を受けて、KAAS と KEGG MODULE を用いて機能評価する枠組みを構築し、8 種の *Bacillus* 属のゲノムデータによる評価とヒト腸内細菌叢メタゲノムによる評価を行った[1]。

KAAS の機能アノテーションではマニュアルで作成されているオーソログデータベースである KEGG ORTHOLOGY（KO）を用いている。KO は精度が高い反面、KEGG に登録されている全遺伝子中の 30%以下しか割り当てられていないためカバー率は低い。そこで、全遺伝子間の配列類似度を元にクラスタリングして、自動的にオーソロググループを同定する方法が提案されてきている。我々も KEGG に登録されているすべての遺伝子（タンパク質をコードする遺伝子）をクラスタリングした KEGG OC (Ortholog Cluster) を開発してきた。OC は、ssearch のアミノ酸配列類似度を使った高速クラスタリングアルゴリズムを用いており、平成 24 年 2 月現在で約 900 万の遺伝子が含まれている。平成 24 年度は、新しいウェブインタフェースとして、キーワード検索の他に生物種分布の表示や類似クラスタとの関係などを調べることができるようにした[2]。今後は、3 ヶ月に 1 回程度の頻度での更新を予定している。また、後述する遺伝子ネットワークを用いた機能予測に応用する予定である。

上記のオーソログクラスタには機能未知の遺伝子も多く含まれている。このような遺伝子は配列類似性だけでは機能アノテーションは不可能であるが、トランスクリプトームやプロテオームなど様々なオミクス情報を組み合わせることによって機能の手がかりを探すことは可能である。我々は機械学習を用いた遺伝子機能のネットワーク推定システム GENIES (Gene Network Inference Engine based on Supervised Analysis) を開発してきたが、平成 24 年にはウェブインタフェースを整備して、より使いやすくした上で論文として発表した[3]。

3) KEGG の拡張

平成 24 年度は、反応パターンについて平成 23 年度に整備した反応オントロジーをウェブ上にまとめて公開した（http://www.genome.jp/reaction/）。このオントロジーの一つを使ってパスウェイ上で連続した類似反応パターンのパスを抽出する方法を開発し、反応モジュールとして定義した[4]。定義した反応モジュールを詳細に解析したところ、生合成経路・分解経路・中間代謝・二次代謝に関わらずそれぞれに特徴的な反応モジュールの組み合わせがあること、基質の違いによるモジュールの使い分けがあること、ゲノム上のオペロン構造との対応が取れる場合があることなど、パスウェイ進化に関する議論ができることが明らかになった。今後は、これまでに提案されている様々なパスウェイ進化モデルとの関連を明らかにする予定である。

ゲノムデータに関しては犬フィラリアの病原微生物である *Dirofilaria immitis* ゲノムプロジェクトと協力し、ゲノムデータから代謝パスウェイの再構築を行った[5]。この結果は既に KEGG DGENES で公開されている。

4) 医薬品相互作用ネットワークの解析

平成 23 年度までに開発してきた医薬品相互作用のネットワークに対して、標的タンパク質の情報や具体的な副作用の情報を取り込むことによって、医薬品・標的タンパク質・副作用の関係を解析した。

具体的には医薬品と標的タンパク質の関係から副作用を予測するシステムを開発し、標的タンパク質のパスウェイ上での機能類似性が副作用の類似性に関連していることを示した[6]。また、逆に FDA AERS に報告された副作用の情報から標的タンパク質を予測する方法も開発した[7]。

参考論文

1. Takami, H., Taniguchi, T., Moriya, Y., Kuwahara, T., Kanehisa, M. and Goto, S.; Evaluation method for the potential functionome harbored in the genome and metagenome. *BMC Genomics* 13:699 (2012).
2. Nakaya, A., Katayama, T., Itoh, M., Hiranuka, K., Kawashima, S., Moriya, Y., Okuda, S., Tanaka, M., Tokimatsu, T., Yamanishi, Y., Yoshizawa, A. C., Kanehisa, M. and Goto, S.; KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic Acids Res.* 41:D353-D357 (2013).
3. Kotera, M., Yamanishi, Y., Moriya, Y., Kanehisa, M., and Goto, S.; GENIES: gene network inference engine based on supervised analysis. *Nucleic Acids Res.* 40:W162-W167 (2012).
4. Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S. and Kanehisa, M.; Modular Architecture of Metabolic Pathways Revealed by Conserved Sequences of Reactions. *J. Chem. Inf. Model.* in press (2013).
5. Godel, C., Kumar, S., Koutsovoulos, G., Ludin, P., Nilsson, D., Comandatore, F., Wrobel, N., Thompson, M., Schmid, C. D., Goto, S., Bringaud, F., Wolstenholme, A., Bandi, C., Epe, C., Kaminsky, R., Blaxter, M., and Mäser, P.; The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets. *FASEB J.* 26:4650-4661 (2012).
6. Mizutani, S., Pauwels, E., Stoven, V., Goto, S., and Yamanishi, Y.; Relating drug-protein interaction network with drug side-effects. *Bioinformatics* 28:i522-i528 (2012).
7. Takarabe, M., Kotera, M., Nishimura, Y., Goto, S., and Yamanishi, Y.; Drug target prediction using adversed event report systems: a pharmacogenomic approach. *Bioinformatics* 28:i611-i618 (2012).