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

近年のオミックス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重要課題の一つである。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 25 年度も平成 24 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、現在は original（database1

に記述があるリンク)、reverse（database2 に記述があるリンク)、equivalent（データベース間で同じオブジェクトを指していると我々が定義したもの)、indirect（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たもの）の 4 種類を定義している。LinkDB では、このような 3 項関係をプログラムで扱いやすいように RDF 形式でも提供しているが、平成 25 年度は、より高度なデータベース検索へと応用するために、RDF を SPARQL と呼ばれる質問言語で検索できるように Virtuoso を用いてデータベース化した。今後は、Stanza などを用いた使用例の作成とインタフェースの改良を行う。

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 25 年度は、平成 24 年度に構築した KAAS と KEGG MODULE を用いて機能評価する枠組みをウェブツールとして実装し、MAPLE（Metabolic and Physiological Potential Evaluator）（http://www.genome.jp/tools/maple/）システムとして公開した。ユーザーがマルチ FASTA 形式のアミノ酸配列情報を入力すると、システムは KAAS を用いて各アミノ酸配列に対して機能と由来生物種の情報をアノテーションして、その情報を元にしてモジュール充填率を計算する。MAPLE では個別のゲノム配列やメタゲノムサンプルに対してアノテーションするだけでなく、複数のゲノム配列やメタゲノムサンプルを簡単に比較できるようになった。現在、数千遺伝子程度のバクテリアゲノムであれば数時間、数十万遺伝子のメタゲノムであれば数日で計算が終了する。システムのディスク容量の関係で計算結果は 10 日間だけサーバに保存されるため、計算結果をダウンロードする機能もつけている。しかし、現在は計算結果をアップロードする機能がないため、計算結果をウェブ上でグラフィカルに表示して見るためには再度計算する必要がある。今後は、結果をアップロードして再解析できるようにするとともに、計算速度と精度の改善、インタフェースの改良を進める。

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) に関しては、平成 24 年度に引き続き定期的にデータ更新し、最新版は平成 25 年 3 月 13 日版で、2,964 ゲノムの約 1,176 万遺伝子を含む 1,324,921 クラスタ（うち、複数の遺伝子を含むクラスタは 520,731）からなる。データの更新は今後も継続して行う。遺伝子ネットワークを用いた遺伝子機能予測への応用に関しては、現在、生物種分布情報に基づいた OC クラスタの分類を検討しているところである。

化学解析ツールに関して、化合物構造をフィンガープリント表現する新しい方法 KCF-S（KEGG Chemical Function and Substructure）を設計し、2 つの化合物間を触媒する酵素反応が存在するかどうかを判定する問題などに応用した[1,2]。今後は、この方法を改良した上で、新規酵素反応パスウェイを予測する問題や新規酵素反応に対応する遺伝子を予測する問題に応用する。

3) KEGG の拡張

平成 25 年度は、KEGG MODULE や RMODULE とグローバルマップとの関係を整理したパスウェイ表示を実装するとともに、医薬品相互作用データベース検索インタフェースを構築した[3]。モジュールとバクテリアのオペロンなどとの関係から、進化的な制約に関する機能とゲノムとの関連解析は引き続き行う。

ゲノムデータに関しては Human Microbiome Project から約 600 サンプル分のデータを取得し KAAS アノテーションを付けた後、MGENES に登録した。また、サンゴと褐虫藻のゲノムを OIST から取得し、DGENES に登録した。今後は、EGENES を廃止する予定なので、植物ゲノムプロジェクトのデータを積極的に DGENES に追加する予定である。。

4) 医薬品相互作用ネットワークと副作用の解析

FDA AERS に報告された副作用の情報から、医薬品・効能・副作用の関係をバイクラスタリング法を用いて特徴づけし、類似した副作用を引き起こすメカニズムの同定を試みた。また、患者の年齢・性別・体重の情報と関連付けて、それらが副作用に及ぼす影響を解析した[4]。さらに、医薬品の標的タンパク質と副作用から、副作用に関与するタンパク質ドメインを推定する方法を開発した[5]。

参考論文

1. Kotera, M., Tabei, Y., Yamanishi, Y., Moriya, T., Tokimatsu, T., Kanehisa, M. and Goto, S.; KCF-S: KEGG Chemical Function and Substructure for improved interpretability and prediction in chemical bioinformatics. *BMC Syst Biol* 7(S6):S2 (2013).
2. Kotera, M., Tabei, Y., Yamanishi, Y., Tokimatsu, T. and Goto, S.; Supervised de novo reconstruction of metabolic pathways from metabolome-scale compound sets. *Bioinformatics* 29:i135-i144 (2013).
3. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M.; Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 42:D199-D205 (2014).
4. Mizutani, S., Noro, Y., Kotera, M. and Goto, S.; Pharmacoepidemiological characterization of drug-induced adverse reaction clusters towards understanding of their mechanisms. *Comput Biol Chem* in press (2014).
5. Iwata, H., Mizutani, S., Tabei, Y., Kotera, M., Goto, S. and Yamanishi, Y.; Inferring protein domains associated with drug side effects based on drug-target interaction network. *BMC Syst Biol* 7(S6):S18 (2013).