微生物ゲノムと細胞機能の統合データベースの開発

●金久　實[1,2)]　◆川島　秀一[2)]　◆片山　俊明[2)]

1) 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　2) 東京大学医科学研究所ヒトゲノム解析センター

〈研究の目的と進め方〉

ゲノムから生命システムの理解を目指した新しい研究の流れの中で、データベースの役割も大きく変化しつつある。第1にデータベースの内容として、配列や立体構造といった個々の分子レベルの情報だけでは不十分であり、相互作用ネットワークや細胞プロセスといった高次レベルの情報を提供できなければならない。第2にデータベースの構築法として、著者が配列データや立体構造データをサブミットするやり方では不十分であり、研究コミュニティの知識が継続的にデータベースに反映される形態をとる必要がある。本研究は本領域が重点的に解析を行ってきた枯草菌とシアノバクテリアを主な対象として、遺伝子・分子レベルのデータや知識から細胞・個体レベルでの機能情報を見いだすための統合データベースを開発し、同時に研究コミュニティと連携したコミュニティデータベースの枠組みを構築することを目的としている。

〈研究開始時の研究計画〉

本研究は、これまで研究代表者が開発してきたKEGGの情報技術を発展させ、以下のように枯草菌、大腸菌、シアノバクテリアを中心としたゲノム生物学研究に適用し、特定領域研究「ゲノム生物学」の後期2年間に寄与する。

(1) 遺伝子機能のコミュニティアノテーション

コミュニティアノテーションとは、一人一人の研究者が単にデータベースを利用するだけでなく、データベースを書き換えることも許可するもので、ゲノムが決定された生物種の遺伝子機能に関する最新の情報を文献情報（PubMed ID）と関連づけ、コミュニティ全体の知識を集約する。シアノバクテリアではこれまで我が国の研究コミュニティと連携し、また最近では欧米の研究者とも連携してCYORFデータベースを開発している。枯草菌については奈良先端大と共同でBSORF データベースを開発している。本計画ではこれらを発展させ、さらに大腸菌のデータベースとしてECORFを追加する。

(2) マイクロアレイデータの解析

シアノバクテリア、枯草菌、大腸菌のマイクロアレイデータについては、これまでKEGG/EXPRESSIONデータベースを構築し、解析システムとともに提供してきた。これを上記アノテーションデータベースと統合し、各遺伝子の配列情報やゲノム上の位置情報と関連づけて解析できるシステムを開発する。

(3) タンパク質間相互作用データの解析

本計画が対象とする生物種の2ハイブリッドデータは、マイクロアレイデータほど大量にあるわけではないが、枯草菌を中心に、これもアノテーションデータベースと統合解析できるシステムを開発する。

(4) 転写シグナルの解析

これまで枯草菌の転写因子データベースBSTFを福山大学と共同で開発してきたので、これも枯草菌を中心にマイクロアレイデータと統合して転写シグナル等の解析ができる形に拡張する。

(5) ゲノムと細胞機能の統合データベース

CYORF, BSORF, ECORFはゲノムの情報から細胞システムの理解を目指したデータベースである。細胞システムの「配線図」はKEGGにおいて分子間相互作用ネットワークという形で蓄積されているので、これをシアノバクテリア、枯草菌、大腸菌に対象を絞って取り込み、ゲノムの配列情報、マイクロアレイデータ、2ハイブリッドデータ、さらには転写制御情報と統合したデータベースを構築する。

〈研究期間の成果〉

本研究の成果は以下の5つのシステム開発である。研究計画との対応は括弧内に示した。

(A) 枯草菌ゲノムデータベースBSORFの開発（計画2, 3, 4, 5）

枯草菌ゲノムデータベースBSORFは、我が国の研究コミュニティが大きな役割を果たしたBacillus subtilisゲノムシークエンシングプロジェクトの成果として、奈良先端科学技術大学院大学と東京大学医科学研究所ヒトゲノム解析センターが共同で開発したデータベースである。

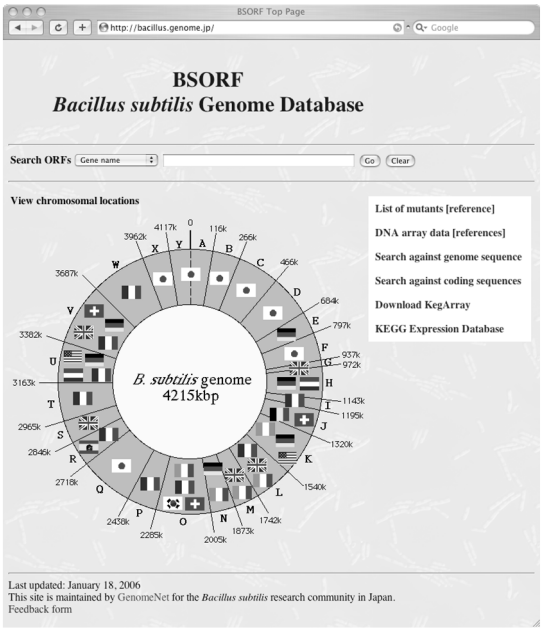

図1. BSORF のトップページ

本研究においては、BSORFをリレーショナルデータベース化し、枯草菌ゲノム、プロテオーム、トランスクリプトームを統合的に解析できるシステムとし、以下のURLで提供した。

http://bacillus.genome.jp /

とくに奈良先端大小笠原研究室のマイクロアレイ実験データと我が国の研究コミュニティによる必須遺伝子の

情報を取り込み、KEGGシステムへのリンクを充実させて、生物種間比較解析やパスウェイ解析を可能とした。また各研究者がデータ入力できるコミュニティデータベースとしての機能も付加したが、実際には利用されていない。

(B) シアノバクテリア遺伝子アノテーションデータベースCYORFの開発（計画1, 2, 5）

CYORFは、研究代表者が我が国のシアノバクテリア研究コミュニティ（らん藻DNAチップコンソーシアム）と共同で行ったゲノムフロンティア研究の副産物として出発したデータベースである。ゲノム配列決定後に明らかにされた新たな遺伝子機能に関する情報をコミュニティ全体で維持していくことを目的とし、各研究者がデータ入力を行うことのできるコミュニティアノテーションデータベースである。CYORFは以下のURLで提供しており、

http://cyano.genome.jp /

本研究において３つの面で大きく発展した。

第１に内容面では、ゲノムを追加して合計8種とし、シアノバクテリア間の比較解析を充実させた。またゲノム比較ツールを改良し、保存されたオペロン構造の解析などを可能とした。本システムは、9番目のゲノムとして、我が国で決定されたSynechococcus sp. PCC6301のコミュニティアノテーションに利用された。

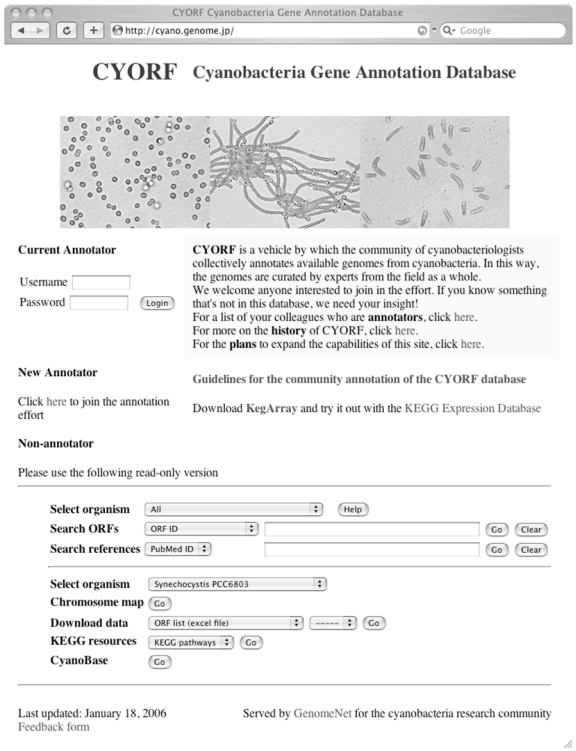

図２．CYORF のトップページ

第２に、CYORFは日本だけでなく国際的な研究コミュニティとの連携に発展した。2006年2月現在、本データベースにアノテーターとして登録しているシアノバクテリア研究者は87名で、その内訳は日本33名、米国26名、フランス7名、インド6名、英国3名、オーストラリア2名、中国2名、その他オーストリア、カナダ、ドイツ、ギリシア、イスラエル、韓国、ロシア、スエーデン各1名である。ただし、実際に活発なアノテーションを行っているのは日本の研究者である。

第３はCYORFの内容が米国NCBIのRefSeqデータベースに取り込まれるようになったことである。配列決定をした著者が登録したGenBankに対し、RefSeqはNCBIが独自にアノテーションを行ったデータベースとされている。しかしながら、急増する多数のバクテリアゲノムに対して、NCBIの専属アノテーターは３名程度しかおらず、ほとんどがコンピュテーショナルなアノテーションに過ぎない。ゲノム配列決定後に続々と明らかにされる個々の遺伝子機能を常にデータベースに反映させていくために、コミュニティの協力が必要なことは、NCBI側でも十分に認識している。CYORFによるシアノバクテリアコミュニティとの連携は、そのテストケースとして始められたもので、CYORFのデータはゲノムネットのFTPサイト

ftp://ftp.genome.jp/pub/db/community/cyorf/

に置かれており、これをNCBI側で取得してもらっている。コミュニティアノテーションの仕組みをNCBIのような公共的なリソースに反映することで、本研究の成果が幅広く提供されることとなった。

(C) マイクロアレイデータ解析ツールKegArrayの開発（計画2, 5）

マイクロアレイ解析によるトランスクリプトーム情報を、ゲノムの情報やパスウェイ・ネットワーク情報と統合して解析することは、様々な生物種で一般的に行われるようになっている。KEGG/EXPRESSIONシステムはKEGGのWebサーバー上でこのような統合解析を可能とするものであった。同様のことを枯草菌ではBSORFを、シアノバクテリアではCYORFを使ってできるようにするため、本研究では単独のJavaアプリケーションであるKegArrayツールを開発した。KegArrayは異なるデータベースに追加するモジュールのようなもので、それぞれの生物種固有データベースにあるゲノム情報とKEGGのパスウェイ情報を統合してマイクロアレイデータ解析を行うことができるようになっている。

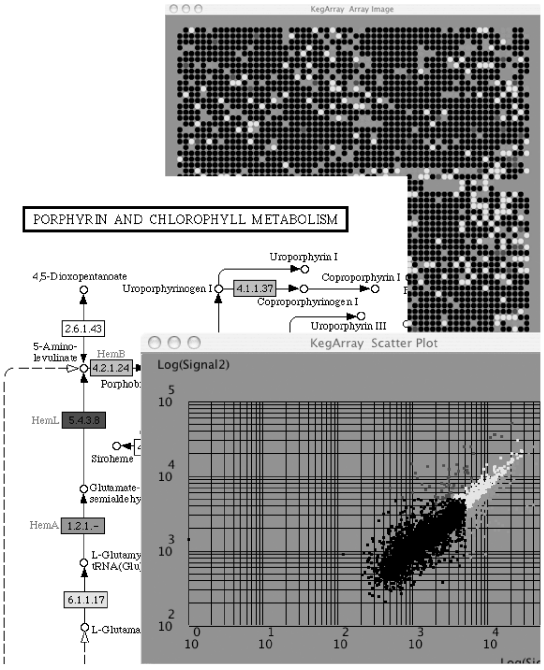

図３．KegArray の画面の例

KegArrayはhttp://www.genome.jp/download/からダウンロードして、誰でも自由に利用することのできるフリーアプリケーションである。Web経由でサーバーにデー

タを送るのではなく、自分のマシンで解析できるので、未発表の個人的なデータを取り扱うのにも適している。

(D) オーソログクラスター自動分類システムKEGG OCの開発（計画5）

特定生物種での遺伝子機能に関する実験データが、他生物種の類似遺伝子にも適用できるかを情報科学的に判定するため、これまでにも様々なオーソログ遺伝子推定法が作られてきた。本研究では、全ゲノム塩基配列が決定されたすべての生物種に含まれるすべての遺伝子をオーソログクラスターに自動分割する方法の開発を行った。

まず各遺伝子をノードとし、遺伝子間の配列類似関係をエッジとしたグラフを考える。エッジには配列類似度のスコアで重みをつけ、ゲノムペアごとの比較でのベストヒット情報（問合せ配列とゲノム中の全配列との比較で最もスコアが高いものをベストヒットと呼ぶ）で方向（矢印）をつけることができる。この膨大な重み付き方向付きグラフに、準クリーク探索というグラフ解析の方法を適用して、KEGGに含まれる全遺伝子を自動的にオーソログクラスター分類したのがKEGG OC (Ortholog Cluster) である。計算結果は以下のURLで公開している。

http://www.genome.jp/kegg-bin/OC/count/lookup

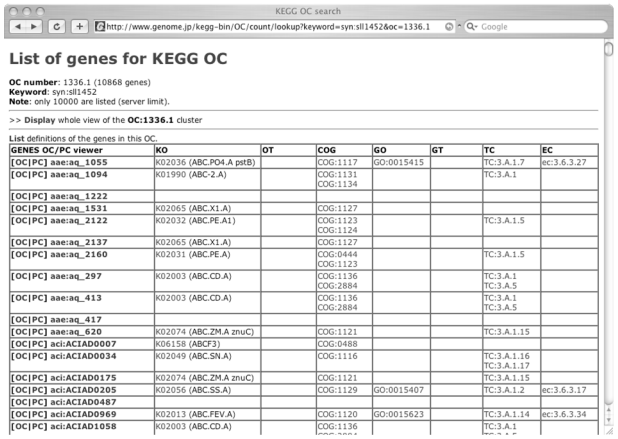

図 4. KEGG OC 検索結果の例

なお、KEGG本体でのオーソログ遺伝子の分類はこれとは別に、手作業を中心とした方法でKO (KEGG Orthology) と呼ぶグループ化が行われている。

(E) KEGG SSDBを用いた遺伝子クラスター解析システムの開発（計画4, 5）

KEGG SSDB (Sequence Similarity Data Base) は上記KEGG OCやKO作成でも利用している全ゲノム全遺伝子間の配列類似情報とベストヒット情報を蓄積したデータベースである。本研究では、ゲノム比較において双方向ベストヒットとなる遺伝子ペア（オーソログ遺伝子である可能性が非常に高い）の情報と、そのペアがそれぞれのゲノム上で隣接しているかの情報を利用し、複数ゲノムで保存された遺伝子クラスター（オーソログ遺伝子群の並び）を検出する方法を開発した。この解析システムはSSDBの 1 つの機能として以下のKEGG GENESのページの中で提供している。

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

異なる生物種間での遺伝子発現制御シグナルの比較解析にも有用なシステムである。

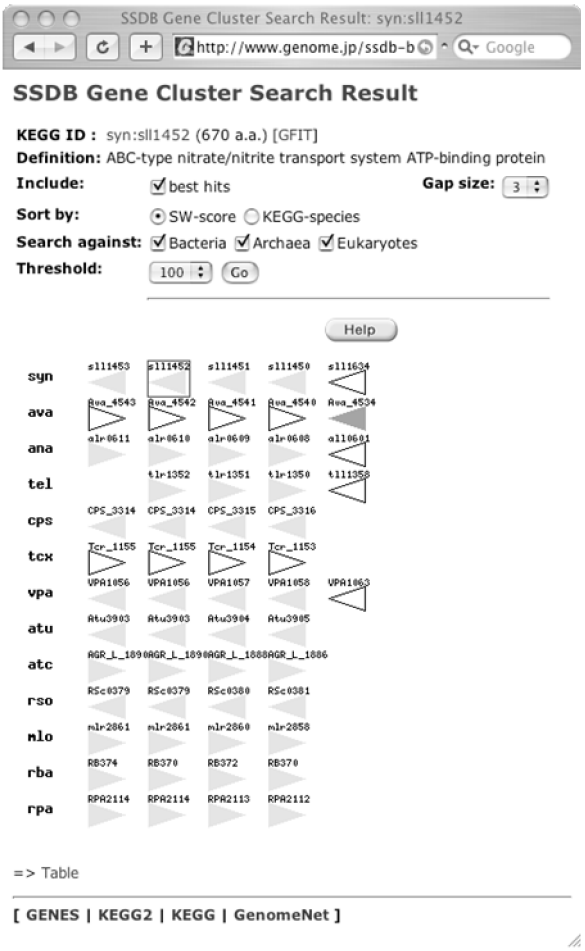

図 5. KEGG SSDB による遺伝子クラスター検索結果の例

〈国内外での成果の位置づけ〉

本研究で開発したデータベースの位置づけは以下の通りである。

枯草菌のゲノム配列はわが国と欧州の研究者が共同で決定したものである。欧州のSubtiListデータベースが2001年4月より更新されていないのに対し、本研究のBSORFは奈良先端大学の小笠原研究室をはじめ国内の研究室との連携により、配列アノテーションがアップデートされ、また配列以外のデータも非常に充実した。このことによりBSORFが枯草菌の国際標準データベースとなった。

シアノバクテリアのCYORFに関しては、各研究者がデータベースの内容を書き換えて、コミュニティ全体で最新の知識を共有するという「コミュニティアノテーション」の概念が広まり、わが国だけでなく米国を中心とした海外の研究コミュニティとの連携が進んだ。さらにはKEGGとNCBIとの連携の一貫として行われていることであるが、CYORFとRefSeqとのつながりができたことは、今後コミュニティの知識をいかに公共データベースに反映させていくか、その仕組み作りを考える上で重要な役割を果たした。

〈達成できなかったこと、予想外の困難、その理由〉

当初計画に記載した大腸菌データベースECORFは実際に新たに構築し、一旦はWebで公開した。しかしながら、大腸菌のコミュニティとの連携体制を作ることはできなかった。大腸菌に関してはすでに国内外に多数のゲノムデータベースが存在していること、内容的にECORFはKEGGと変わらず、単に大腸菌専用のインターフェースをもったKEGGシステムであったことがその原因である。

そこで大腸菌については本研究としての開発は断念し、米国Purdue大学のBarry WannerらをE. coli K-12 Information ResourcesにKEGGも参加することとした。幸いこのプロポーザルはNIHの支援を受けることとなったので、今後は国際的なネットワークの一員として、実質的に本研究で目指したことを実現したいと考えている。

〈今後の課題〉

ゲノムから細胞・個体に関する高次機能を解読するには、配列情報だけでは不十分であり、トランスクリプトーム、プロテオーム、パスウェイ等の情報を統合して解析することが必要である。本研究はそのための基礎技術開発と、枯草菌、シアノバクテリアにおいて実用的なデータベース開発を行ってきた。その中で2つの重要な考え方がうまれ、これを今後さらに発展させていく予定である。

1つはコミュニティアノテーションデータベースの内容を公共データベースに反映させ、これによりコミュニティの知識を外へ広げ、他生物種へも広げる考え方である。現時点では米国NCBIを公共データベースの相手としているが、今後は我が国の中で同様のことができることが望ましい。

もう1つは個々のアプリケーションをモジュールとして開発する考え方である。本研究で開発したKegArrayでは、アプリケーションに含まれるデータのリンク先を変更することで、KEGG、CYORF、BSORFといった具合に、それぞれのデータベースに組み込んで利用することができるようになっている。今後ともこのようなモジュラーアプリケーションを開発し、ゲノムから高次機能を解読する統合解析環境を構築していく予定である。

〈研究期間の全成果公表リスト〉

1）論文

1. 0602081746
   Itoh, M., Akutsu, T., and Kanehisa, M.; Clustering of database sequences for fast homology search using upper bounds on alignment score. Genome Informatics 15(1), 93-104 (2004).
2. 0602081739
   Yamanishi, Y., Vert, J.-P., and Kanehisa, M.; Protein network inference from multiple genomic data: a supervised approach. Bioinformatics 20, i363-i370 (2004)
3. 0602081735
   Igarashi, Y., Aoki, K.F., Mamitsuka, H., Kuma, K., and Kanehisa, M.; The evolutionary repertoires of the eukaryotic-type ABC transporters in terms of the phylogeny of ATP-binding domains in eukaryotes and prokaryotes. Mol. Biol. Evol. 21, 2149-2160 (2004).
4. 0602081729
   Itoh, M., Goto, S., Akutsu, T., and Kanehisa, M.; Fast and accurate database homology search using upper bounds of local alignment scores. Bioinformatics 21, 912-921 (2005).
5. 0602081724
   Yamanishi, Y., Vert, J.-P., and Kanehisa, M.; Supervised enzyme network inference from the integration of genomic data and chemical information. Bioinformatics 21, i468-i477 (2005).

2）データベース／ソフトウェア

1. シアノバクテリア遺伝子アノテーションデータベース
   CYORF (http://cyano.genome.jp/)
2. 枯草菌ゲノムデータベース
   BSORF (http://bacillus.genome.jp/)
3. マイクロアレイデータ解析ツール
   KegArray (http://www.genome.jp/download/)

3）特許

なし