課題名「ゲノム生物学バックボーンデータベースの構築提供」
代表研究者　国立遺伝学研究所　教授　菅原　秀明

1.代表研究者による成果概要報告

1-1.研究開発のねらい

1987 年 7 月、日本 DNA データバンク(DDBJ)は自ら収集していた塩基配列データの提供を開始した。1992 年 1 月には、米国の National Center for Biotechnology Information (NCBI)ならびに欧州の European Bioinformatics Institute (EBI)とデータ交換した国際塩基配列データベース(International Nucleotide Sequence Databases(INSD))の提供を開始した。それから 2005 年 12 月までに、DDBJ 由来データが INSD に占める割合は、塩基数で 3%から 12%へ、登録件数(エントリー数)で 2%から 17%へと増加して、EBI と肩を並べるところまでに至った。

一方その間 1990 年代後半から徐々に、INSD の遺伝子配列などの既登録データが新規ゲノム配列のアノテーションに利用され、その成果がまた INSD に登録されるというスパイラルが成長し始めた。すなわち INSD はゲノム生物学のバックボーンデータベースとしての色彩を強めてきた。

こうした背景の中で、DDBJ はデータ品質に関する問題点の指摘を受けるようになり、また、INSD のような国際公共データベースが存在していなかった遺伝子発現データへの対応が求められるようにもなった。このため、我々は、データの品質を高めた高品位データベースと、遺伝子発現データの蓄積から利用までに対応する四次元データベースの概念を提唱した。

高品位データベースとして具体には、データバンクやコミュニティーの知見を幅広く集約することを目指した Open Annotation SYStem (OASYS)の開発と、ゲノム配列からタンパク質産物の構造を予測する Genes TO Proteins (GTOP)の拡充によるゲノム配列のアノテーションの再評価を目指した。

四次元データベースとして具体的には、遺伝子発現情報の登録・査定・蓄積・公開さらに空間的・時間的比較解析を可能にする MciroArray expression DataBase (MADB)と Bio-Simulated Database (BSD)の開発と提供を目指した。将来、これらの高品位データベースならびに四次元データベースと、オリジナルの塩基配列、アミノ酸配列、タンパク質立体構造、パスウェイなどを対象とするバイオ・データベースとの連携が、ゲノム生物学を活かし、また、支える情報環境として発展していくものと想定した。

1-2.研究開発の成果

1)高品位データベース

【OASYS によるアノテーション】

OASYS は、計算機のオペレーティングシステムの linux を代表とするオープン・ソフトウエアの開発過程をモデルとして、INSD の公開エントリーに第 3 者によるアノテーションを付与・集積していくという概念でありシステムで

あった。奇しくも、本研究開発課題の初年度にあたる平成 13 年度の INSD 国際実務者会議と国際諮問委員会で、米国 NCBI から Third Party Annotation(TPA)が提案された。TPA は OASYS と異なって独自の実験と論文発表を受付の必要条件としていたが、第 3 者アノテーション導入の一点で、OASYS は国際動向と同調した構想となった。

OASYS の具体システムとして、DDBJ に用意したサーバーと対話型で実行する Web 版と、ネットワークの状況の如何にかかわらず実行可能なポータブル版を開発した。システムの検証のために、DDBJ に登録された微生物ゲノム配列データを再評価するプロジェクトである Gene Trek in Prokaryote Space(GTPS)も立ち上げて、そこでの実証実験の結果をフィードバックしながら開発を進めた。また、システム開発の過程で XML 技術と Web サービス技術を利用したが、DDBJ の既存のデータベースの Web サービス化と相乗効果があった。

GTPS としては、2003 年版 124 株、2004 年版 184 株そして 2005 年版 303 株と微生物ゲノム上の全ての Open Reading Frame（ORF）を再評価して、信頼性が高くかつ新規の ORF 候補も同定することができた。また、GTPS から派生して、比較ゲノム解析のためのパッケージ G-InforBIO も開発した。

【GTOP によるアノテーション】

参照データベースの更新作業を大幅に簡素化し、全ゲノム配列を対象とするタンパク質の立体構造予測、ファミリー分類、機能モチーフ部位予測等を、4 ヶ月ごとに実施することが可能になった。平成 18 年 2 月現在、真核生物 50、古細菌 21、真正細菌 203、ファージ 172 の合計 446 種の GTOP 解析結果をデータベースとして公開している。GTOP 解析の過程で、ORF の産物が立体構造をとりうる可能性を判定することができ、その結果 ORF の再評価を行なうことができた。

また、GTOP 解析とそのグラフィック表示を活用して、選択的スプラインシング、偽遺伝子、好熱性・好塩性タンパク質の特徴、進化系統解析ならびに構造ドメインの構成について新たな知見を得ることができた。

2)四次元データベース

【MADB による遺伝子発現情報データベースの構築】

個々の遺伝子の発現情報をマイクロアレイ技術によって大量に取得できるようになったが、実験条件の影響が大きいため、異なる研究室由来の遺伝子発現情報を比較することが困難であった。この解を目指して発足した Microarray Gene Expression Data（MGED）グループは平成 14 年に MGED Society （http://www.mged.org/）に発展した。MADB グループはこの一連の活動に積極的に参加し、その間に整備されてきたデータ形式の標準（MGED フォーマット）に準拠して Center for Information Biology gene Expression database（CIBEX）を設計・開発し、EBI の ArrayExpress ならびに NCBI

のGEOと共同して遺伝子発現情報の国際共同データバンクを形成しつつある。具体的には、発現情報の信頼度の指標を開発して情報のデータベースの質の向上を図り、欧米の２バンクとのデータ交換のツールもXML技術を応用してすでに用意した。国内では、データバンクへのデータ登録を促進するために、学術誌の協力を獲得しつつある。さらに、マイクロアレイ由来に限らずESTやSAGE由来の遺伝子発現情報にも対応し、高度な解析へと移行できるように次項のBSDとの相互運用性を実現した。

【BSDによる遺伝子発現情報の高度利用】

遺伝子発現情報を空間的な広がりと時間的の広がりの中で解析できるプラットフォームとしてBSDを開発した。対象は、cDNAマイクロアレイ、DNAチップ、SAGE、ESTならびにin situ ハイブリダイゼーション由来の情報である。

このプラットフォームによって、特定の組織で発現する遺伝子や、複数の組織で共通に発現する遺伝子を探索して、その発現情報を3次元で表現した生物の形状データに重ねて表示することができる。また、3次元グラフィック上で発生と分化の段階を追った発現の変化も追うことができる。したがって、BSDによってどの遺伝子がいつ、どこで発現しているのかということを利用者は直感的に理解することが可能であり、BSDは遺伝子間の相互作用とネットワークを理解していくための有力な手段の一つになるであろう。また、BSDにはMGEDフォーマットのデータやCIBEXからデータを読み込ませることができるので、BSDに登録されている発現情報とBSDの利用者独自の発現情報を比較解析することも可能である。

2.事後評価結果

2-1.当初計画の達成度

高品位データベースと四次元データベースという概念を、OASYS、GTOP、CIBEXおよびBSDとして具体化した。OASYSは、DDBJから入手可能なゲノム配列データのアノテーションを再評価することを目的としていたが、新規のゲノム配列に対して比較ゲノム解析を行なってアノテーションを付与するG-InforBIO の開発へと広がった。また、公開されている微生物ゲノム由来のORFの再評価を行い新規ORF候補を発見した。GTOP解析によって、ORFの信頼性評価に加えてORFの構造やタンパク質の立体構造について新たな知見を得ることができた。MADBはCIBEXとしてDDBJ固有の事業から国際公共データベース事業に中に位置づけられた。さらに、4データベースにKEGGとPDBを加えた情報資源の統合検索システムが構築された。本研究開発課題で利用したXML技術とWebサービス技術は、本研究開発課題とDDBJのデータサービスの間の相乗効果をもたらした。

2-2.知的財産権、外部発表(論文等)等研究開発成果の状況

国立遺伝学研究所JST-BIRDポータルサイト：http://www.jst-bird.nig.ac.jp/
GIB (Genome Information Broker)：http://gib.genes.nig.ac.jp/
G-InforBIO：http://www.wdcm.org/inforbio/G-InforBIO/download.html
OASYS：http://althea.ddbj.nig.ac.jp/index.jsp
Gene Trek in Prokaryote Space (GTPS)：http://gtps.ddbj.nig.ac.jp/
DDBJ-XML：http://gtps.ddbj.nig.ac.jp/
GTOP：http://spock.genes.nig.ac.jp/~genome/gtop.html
国際公共遺伝子発現データベース CIBEX：http://cibex.nig.ac.jp/index.jsp
BioSimulated DataBase (BSD)：http://bsd.genes.nig.ac.jp/bsd_web/Top.jsp

一部公開されたばかりのものもあるが、各データベースとも着実にアクセスされている。原著論文発表、招待・口頭講演は国内外ともに数多く行われている。また、講習会も定期的に行っており、普及活動にも努めている。

2-3.研究開発成果の公開による波及効果

OASYS は国際アノテーションワークショップにて、ORF の再評価に貢献し、GTPS が予測した新規 ORF が採用された。GTOP による ORF の再評価と産物の構造の観点からの分析は独自性が高い。MADB は、遺伝子発現情報の国際公共データベース形成に貢献している。BSD は遺伝子発現情報解析に新たなプラットフォームを提供した。主要なバイオ情報資源で Web サービスが急速に広がってきた。

2-4.成果の実用化の可能性及び成果から予想される波及効果

データバンクの大規模解析と専門家グループの融合により、遺伝子配列とゲノム配列の詳細な解析が広がると考えられる。高品位データベースは、ゲノム生物学に限らず塩基配列データベースを利用するあらゆるバイオ研究に依って足るべき軸を与え、本データベースの品質評価付データは、緻密な研究計画を効率よく立てることに貢献すると期待する。

CIBEX は遺伝子発現情報の国際公共データベースの確立への貢献が期待される。BSD は、ゲノムネットワークの研究に対して高性能な情報環境を提供する。

また、これからの展開により、異種分散したバイオ情報資源の統合利用が広がることを期待する。

2-5.総合評価

三大国際 DNA データバンクのひとつである DDBJ の高度化として、4 つのデータベースを構築し、データの品質の問題への取り組み、遺伝子発現データの基盤への取り組み、また、その利用について新しいインターフェイスを作り上げた点が評価できる。高品位データベースの成果から、一旦公開されたデータについても、定期的再評価が必要なことが明らかになったことは、科学研究において重要である。今後、DDBJ における大規模計算機解析によるアノテー

ションと研究コミュニティー群による専門的アノテーションを融合する枠組みにより、遺伝子配列データとゲノム配列データのアノテーションが豊かになっていくこと、四次元データベースによって、個々の遺伝子の振舞いや遺伝子のネットワークの振舞いの解析が促進されていくことも期待する。本研究開発の成果であるデータベースとツールを活かしていくためには、研究者コミュニティーとの連携を強める必要があるであろう。本研究開発は DDBJ の活動とリンクして国際的な対応も行われているが、今後の国際社会への浸透を期待したい。通常業務である DNA データバンク構築を含め、更なる研究開発には人員、資金の両面から拡大の必要性があると思われる。国際的拠点として研究開発体制の整備を図り、継続的に DDBJ の活動と研究開発に取り組んでいくことを期待する。

3.主な論文発表

1)Sugawara, H., Miyazaki, S., Abe, T., and Shigemoto, Y. (2005). Biological Data Analysisi using DDBJ Web services. Proceedings of BIOINFO2005. 379-382.

2)Riley, M., Abe, T., Arnaud, B.M., Berlyn, M., Blattner, R.F., Chaudhuri, R.R., Glasner, D.J., Horiuchi, T., Keseler, M.I., Kosuge, T., Mori, H., Perna, T.N., Plunkett, G., Rudd, E.K., Serres, H.M., Thomas, H.G., Thomson, R.H., Wishart, D., and Wanner, L.B. (2006). Escherichia coli K-12: a cooperatively developed annotation snapshot—2005, Nucleic Acids Research, 34, 1-9.

3)Homma, K., Fukuchi, S., Kawabata, T., Ota, M. and Nishikawa, K.: A systematic investigation identifies a significant number of probable pseudogenes in the Escherichia coli genome. Gene, 294, 25-33, 2002.

4)Fukuchi, S. and Nishikawa, K. (2004). Estimation of the number of authentic orphan genes in bacterial genomes. DNA Res., 11, 219-231.

5)Matsumura, Y., Shimokawa, K., Ikeo, K., Tateno, Y., Hayashizaki, Y. and Kawai, J. Development of the reliability index for the measurement value of each spot in a DNA microarray (RIESM) and addition of RIESM to READ of CIBEX. Gene (accepted)

参考：

論文発表　国内 9 件、海外 49 件　（原著論文のみ）

口頭発表　国内 45 件、海外 32 件

ポスター発表　国内 34 件、海外 10 件

特許出願　なし

出版　６冊