研究開発課題別中間評価結果

１．研究開発課題名
ゲノム生物学バックボーンデータベースの構築提供

２．代表研究者名
菅原秀明（大学共同利用機関法人情報・システム研究機構国立遺伝学研究所　教授）

３．研究開発概要
DNA Data Bank of Japan（DDBJ）、EMBL databaseおよびGenBankが共同で構築・提供している国際塩基配列データベース(International Nucleotide Sequence Databases (INSD))は、生物研究にとって必須のデータ資源となっている。一方、個々のエントリーの品質検証、ゲノム単位でのデータ操作への対応、塩基配列に特化したアーカイブからの発想の転換などが求められている。本研究開発では、個々のゲノムに対応してデータの品質を高めた「高品位データベース」として「OASYS(Open Annotation SYStem)」「GTOP(Genome TO Protein structures and functions)」を、遺伝子発現データに対応して空間的な広がりに時間軸を加えた４次元データベースとして「MADB(MicroArray DataBase)」「BSD(BioSimulated Database)」を開発する。

４．中間評価
４－１．研究開発の進捗状況と今後の見込み
・OASYS：研究コミュニティーの要望に沿ったシステム開発を進めている。Web servicesの導入によりデータ処理プロセスの単位の柔軟な組み替えが可能になり、GRID環境で大規模なデータ処理に対処できる見通しがついてきている。

・GTOP：真正細菌(130種)、古細菌(17種)、真核生物(19種)の合計166種のゲノムが解析・格納された。これにより、当初の目的であるゲノム配列が決定した全生物種の収録が達成された。さらに、利用者の要望に応えファージ111種のゲノムが収録された。立体構造の予測に関しては、高精度な計算手法の開発が進められている。GTOPの解析システムは「ヒトcDNAアノテーションプロジェクト、H‐invitational」に提供されており、また、OASYSとDDBJによって解析・確認されたORFセットを解析し、ゲノム情報の高品位化へとフィードバックしている。

・MADB：国際組織Microarray Gene Expression Data（MGED）委員会に参画し、マイクロアレイ実験やデータ表記の標準化、MicroArray Gene Expression‐Markup Language（MAGE-ML）設計などに関わっている。マイクロアレイデータを中心とする遺伝子発現データベースCIBEXを開発している。また、スキーマを簡素化して可用性を向上させている。

・BSD：遺伝子発現データと多様な生物種のデータ収集体制の確立を進めている。種々の遺伝子発現データを共通フォーマットで収納するためのデータベースを作成、テストデータ収集を開始している。ゲノム情報から遺伝子配列やタンパク質構造までを一連のデータとして検索・取得できるように開発を進めている。また、遺伝子発現分布の空間的な把握を容易にするために、三次元可視化のための詳細設計を進めている。

進捗に関しては比較的進んでいるものや、やや遅れているものなど一様ではないが、各課題の位置付けをより明確にしながら、全体として連携していくことが重要であり、今後の展開に期待したい。

４－２．研究開発成果の現状と今後の見込み
Web公開状況

ゲノム生物学バックボーンデータベース　　http://www.jst-bird.nig.ac.jp/

・OASYS：DDBJの微生物ゲノムデータベースGIBのデータを対象として、100万以上の予測ORFの品質を細分類、ランク付けし、その結果INSD未登録のORFを見出している。GRIDにおいて水平移動遺伝子の網羅的解析と大規模クラスタリングを実行している。

・GTOP：比較ゲノム解析研究により、「大腸菌ゲノム偽遺伝子の同定」「好熱菌・好塩菌タンパク質の組成の特徴」「ドメイン構成比較による系統樹の作成」「ゲノム注釈付けの問題点の指摘」等の成果をだしている。

・MADB：CIBEXを試験公開し、大腸菌やヒトの遺伝子発現データ登録に対応している。
・BSD：ESTを中心とした配列決定に基づく発現プロファイルデータに関して、データ提供者に対して限定的に開放しテストを進めており、将来、一般公開される予定である。
本研究開発の成果はweb上での公開やNucleic Acid Research等の国際雑誌への掲載、様々な学会での発表等、広範に発表されている。
今後は利用者への案内など、さらに個々の成果を見えるようにすることが重要であり、今後の進み方に期待する。

４－３．今後の研究開発にむけて

OASYS、GTOP、MADBおよびBSDの成果に外部の情報資源も加えて、配列データからタンパク質構造そして遺伝子発現データまで一貫検索と解析を可能にすることを目指している。
・OASYS：DDBJでのアノテーションを継続し、大規模なアノテーションジャンボリーの成果の活用も目指している。ゲノム配列の解析からDDBJへの登録までを支援するポータブルOASYSの開発が計画されている。異なる研究者のアノテーション結果を横断的に検索できる機能の開発も考えられている。
・GTOP：ゲノム上にコードされた全タンパク質を分類し、相互の関係性を全体的にとらえるとともに、異種間で比較する方法を考案し、計算プロテオームの手法を開発することが計画されている。また、真の全原子構造を得られるエネルギー計算の方法論確立を試み、立体構造からのタンパク質機能予測システムの開発を目指している。
・MADB：欧米の遺伝子発現データベースとのリアルタイムでのデータ交換を実現させるためにMAGE-MLに対応するデータ交換システムの開発が計画されている。また、CIBEXの公開システム・アプリケーションを最適化スキーマに対応させていくことが考えられている。さらに、独自の統計学的手法を導入したデータ解析ツールの開発が計画されている。
・BSD：遺伝子発現データを比較解析する手法の確立とシステムへの取り込みが今後も続けられる。また、統合システムのための三次元イメージの収集と取り扱いのためのインターフェースを充実させると共に、関連データとの連携およびデータからの意味抽出のためのマイニング機能の充実をあわせて進めていくことが考えられている。最終的には､遺伝子発現プロファイルの比較を定量的におこない､遺伝子発現変化を総合的に捉えるためのシステムとし、公開を目指している。

５．国内外のデータベース高度化・標準化の動向・状況と本課題の位置づけ

・OASYS：国際的協調と競争の双方の観点で完成を急ぐ必要がある。
・GTOP：通常の配列解析では分からない偽遺伝子を感度よく検出できることを実証し、エクソン・イントロン構造の予測機能等によってアノテーションの精度向上に貢献している。
・MADB：国際協力のもとに遺伝子発現データの蓄積を促進し、遺伝子発現データのアーカイブとして機能していくことが期待される。
・BSD：MADBへの対応も含めて、時間的推移も含めた解析を可能とするインターフェースが広く研究社会に普及することが期待される。
本グループは、国外対応の窓口となっているので、広報活動や講習会などを積極的に行い、我が国の研究者にデータの登録をうながすなど、一層の努力と発展を期待したい。

６．研究開発成果の社会への貢献

OASYSとGTOPにより配列データに高品質な評価を加えることを可能とし、MADBとBSDによって配列データから国際標準に基づいた発現データへの展開をもたらし、その成果をフィードバックして、ゲノム情報の高度化をもたらす。品質評価を付加したデータを使いやすいインターフェースで提供することによって、実験研究の計画立案からゲノム情報の診断や処方への応用までに貢献することが期待される

７．総合的評価

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、種々のデータを統合化している点が評価できる。今後は、４つのサブグループが個々に進めてきた成果をまとめ、全体として配列データから蛋白質構造・遺伝子発現まで検索・解析が行えるシステムが構築されることを期待する。

本研究開発には本来事業の支援の面が見られるが、DDBJを中心とするデータバンクシステムの国際的拠点としての立場から、本研究開発の４つの課題の位置付けを、利用者と共に検討することが今後の発展に寄与するのではないか。積極的に公開し、利用者からの反応に答えることによりさらに発展することが重要である。継続して進めていくべき事業であり、今後の発展のためにも研究開発体制の整備が望まれる。