公募研究：2000年度

実験手法が異なる情報源からのゲノム比較を可能とする解析システムの研究開発

●菅原秀明　◆宮崎　智

国立遺伝学研究所生物遺伝資源情報総合センター

〈研究の目的と進め方〉

微生物については２６種をこえる種（課題申請時）についてゲノム配列が公開され、ここ１、２年でさらに１００種程度のゲノム配列が公開される見込みであり、比較ゲノム解析が本格的に行われるようになってくる。しかし、公開されているアノテーションは、異なるグループが異なる手法で加えたものであり、単純に比較することはできない。このために、統一プロトコールを策定してそれによって公開ゲノム配列データを解析した結果をデータベース化する。

〈研究開始時の研究計画〉

各研究グループの手法を分析した結果に基づいて予備実験を行い、本研究における統一的な手法を設定し、それを用いて必要に応じてアノテーションを付け替えることが可能な機能を実現する。

またこの手法を一般に公開することによって、信頼性の高い比較ゲノム解析を可能とし、またその結果がフィードバックされる仕組みも用意して、配列データからの生物学的意味の抽出に貢献する。

こうした大量データの迅速な獲得と信頼性の高い解析のために、ネットワークとコンピュータを利用した効率的なデータ獲得と管理、大量情報解析、そして可視化の機能を盛り込む。

このシステムを単なるプロトタイプではなく実用システムとしてDDBJから公開することによって、ゲノム配列決定プロジェクトグループに対して、アノテーションの一つの基準を示すことができる。また、DDBJ大量登録フォーマットでの入出力機能を付加することによって、ゲノムプロジェクトで配列データに対する情報付加作業にも応用できる。大量登録フォーマットに更新があった場合にも迅速に対応することが可能である。これによって、広く一般に利用可能なゲノム解析の情報基盤を整えることにもなる。

〈研究期間の成果〉

本研究課題で開発したシステムはGenome Information Broker(GIB)として広く一般に公開し始めた。

GIBにおいては、国際塩基配列データベースに登録・公開された微生物の完全ゲノムデータをフラットファイルの形式で随時取得し、フォーマットを変換して分散データベースに格納し、複数ゲノムのデータを選択・検索・閲覧可能である。

GIBのシステムの構造として、拡張性を実現するために、データを複数のサーバーに分散して持たせ、それらをCommon Object Request Broker (CORBA)で統合した。また、システム内のデータのやりとりにeXtensible Markup Language (XML)を採用した。

GIBの機能としては、収録されている単独あるいは複数のゲノムデータを対象とするOpen Reading Frame (ORF)検索、相同性検索（blastとfasta）ならびにキーワード検索（遺伝子産物などを対象）とその結果のリスト表示、グラフィカル表示およびダウンロードを実現した。すなわち、異なる研究グループが決定した複数のゲノムデータを比較解析する枠組みを実現した。

〈国内外での成果の位置づけ〉

研究期間終了後もGIBを継続して拡張し、GIBにおいてアーケア25ゲノム、バクテリア275ゲノムおよび真核微生物6ゲノムを登録するに至った（2006年1月時点）。すなわち、課題申請当時の10倍を超えるゲノムを当時の設計に基づいたシステムで問題なく処理できている。またGIBの手法を、ウイルスゲノムに適用したGIB-Vも現在では公開している。

さらに、このWebで公開しているGIB開発経験をもとに、国際塩基配列データベースからダウンロードしたゲノムデータ複数と利用者独自のゲノムデータを比較解析可能なポータブル（手元のコンピュータで利用可能な）なシステムG-InforBIOも開発し広く一般に提供している。

一方で、研究目的の一つであった「統一的な手法で再評価」する機能は研究期間中には実現できなかったが検討した内容を、その後展開することができ、我々が検討・設定したプロとコールにしたがって、公開されている微生物完全ゲノムのORFを再評価した結果をデータベースとして公開する準備が進んでいる。

加えて、本研究課題においてXML技術の使用経験を積んでいる期間に、DDBJ独自の塩基配列データベース用のXMLの開発へと展開していたため、国際塩基配列データベース(International Nucleotide Sequence Databases (INSD))において統一的に使用する予定のINSD-XMLの近年の仕様決定の議論においてDDBJの存在感を示すことができた。

GIBに対する国内外からのアクセス件数は2000年～2001年は年間40万件強であったが、2002年から延び始め、2004年からは年に210万件を超えており、本研究課題は研究期間中および終了後において、大きな波及効果をもたらした。

さらに、遺伝子発現や代謝経路のデータベースとの連携システムも2005年に試作した。

〈達成できなかったこと、予想外の困難、その理由〉

当初計画は3年程度の期間を想定したものであったが、そのなかで初年度に実現すべきものは実現できた。

〈今後の課題〉

今後も毎年1回の網羅的解析を計画しているが、微生物ゲノムの公開数は伸び続けているため、こうした網羅的解析のためには大きな計算機資源が必要になってきている。また、データ処理の殆どの過程を自動化したが、データ処理の最終段階では専門家による評価が必要である。このため、専門家の判断のノウハウを機械化する必要が高まっている。

〈研究期間の全成果発表リスト〉

2）データベース

GIB(http://gib.genes.nig.ac.jp/)