1　ゲノム解析の現状と将来

宮崎　　智・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　生命情報・DDBJ研究センター
データベース運用開発研究室

Current Status and Future View for the Genome Sequence Data Analysis

Satoru MIYAZAKI and Hideaki SUGAWARA

Laboratory for the Research and Development of Biological Databases,
Center for Information Biology and DDBJ,
National Institute of Genetics,
Research Organization of Information and System

はじめに

微生物での完全長ゲノム配列の決定と公開を皮切りに，この10年間で様々な生物種で完全長ゲノムを決めるプロジェクトが行われてきている．ほとんどの場合で，全ゲノムショットガン法が使われている．全ゲノムショットガン法を用いる場合はよく知られているように，産出される配列は，のべにして最終産物である完全長ゲノムの約5～10倍に及ぶ．また，遺伝子配列情報を得るために（特に，真核生物では）ESTの決定やfull length cDNAの決定も同時進行となってきている．ゲノムプロジェクトから産出されるデータを反映するように，国際塩基配列データベースに蓄積されているデータ量は，1.7～2.0倍/年で伸び続けている．2003年の4月にヒトゲノム計画が一段落した旨の発表があった後もこの傾向は変わっていない．むしろ，「ポストゲノム解析」と名目がかわっただけであり，遺伝子ネットワークの解析となれば，先にきまった配列のすべての組み合わせを考慮して新たな配列決定を行うなど，配列決定の需要は増加する傾向にあると思われる．さて，このような生命現象解析の基本的なデータが公共データベースに登録されて全研究者の共有資産となっているのは非常に重要で良い傾向であると思われるが，データ量の爆発的な増大と，各研究者や研究プロジェクトのアノテーションストラテジーの相違が思わぬ落とし穴を作ろうとしている．この報告では，ゲノムデータの到来に対する国際塩基配列データベースの取り組みを紹介しながら，ゲノムデータに関する現状と将来像を紹介する．

国際塩基配列データベースにおけるゲノムデータへの取り組み

DDBJ/EMBL/GenBank国際塩基配列データベースでは，ヒトゲノム計画から産出される塩基配列の津波が押し寄せはじめた1996年ごろから，ゲノム配列の受け入れを念頭にさまざまな改革を振興させてきた．もろもろの状況から，1997年～1998年の1年間に予想される登録配列数が，それまでの10年間での登録数に等しくなるであろう

Reprint requests to: Satoru MIYAZAKI
Faculty of Pharmaceutical Sciences
Tokyo University of Science
2641 Yamazaki,
Noda 278-8510 Japan

別刷請求先： 〒278-8510　千葉県野田市山崎2641
東京理科大学薬学部薬学科　　宮崎　智

ということが強く示唆されたことを受け，データバンクでは，

1）既存アノテーションソフトウェアの抜本的な改良
2）一括大量登録にも対応可能な登録者ーバンク間でのデータ交換フォーマット
3）ゲノム配列に特化した公開用のカテゴリーの新設

を模索してきた．日本DNAデータバンク(DDBJ)における成果は，以下の通りである．1）について，DDBJのみならず，EMBLやGenbankにおいても，データの管理システムには，リレーショナル型のデータベースシステムが使われていたが，市販のSYBASEやORACLEに付随のツールでは，データベース構造とユーザーインターフェースがタイトに関連する仕組みになっていた．このため，データベース構造が変更されると，それに合わせてユーザーインタフェースをほとんど作り直す必要があった．ゲノムデータの受け入れは，登録の頻度や一度に登録される塩基数の大きさがそれまでに想定されていたものの，数100倍に上ることから，既存のデータベース構造がそのまま流用できるとは考えにくい．元来，生物学的データはその実験手法の改良などにともない，時と場合に応じて常に変遷するものである．従って，データベース構造の柔軟さをいかに効率よく実装できるかが運用上の鍵となる．DDBJでは，オブジェクト指向的な考えを積極的に導入し，データベース管理システム層とユーザーインターフェース層の間に中間層のライブラリーを構築し，データベース構造の変遷があってもユーザーインターフェースの改良が最小限で済むような仕組みを導入した．また，この3層構造化によって，新規にグラフィカルユーザインターフェースを実装しても，データベース側とは独立に作業が行えるようになった．こうして，未知のデータベース構造の要求に対しても効率よく対応できることになる．2）については，100万レコードの一括登録にも耐えうる大量登録フォーマットをデザインした．5列からなるタブ区切りのテキストファイルを形式化したものである．このフォームを構築するのに特別なソフトは必要なく，またプログラム処理にも適しているものである．3）については，国際実務者会議での検討を経て，完全長ゲノムを実現するための，部分配列の「糊しろ」情報を専用に提供するための，Contig（CON）divisionの新設と，ドラフトの部分配列を受け入れるためのHigh Throughput Genome（HTG）が新設されることになった．また，完全長ゲノムとは独立に，full length cDNAを受けいれるためのHTC divisionも新設されている．

微生物ゲノム配列データベースの現状と問題点

比較ゲノム解析ということでもっとも注目されかつ実践的な状況にあるのが微生物ゲノムである．1996年の大腸菌の完全長ゲノム登録から2004年6月現在で173のゲノムが前章で紹介した国際塩基配列データベースに登録されている．国立遺伝学研究所の生命情報・DDBJ研究センターでは，この微生物情報を効率よく公開するシステムとしてGenome Information Broker（GIB）を構築し公開している（**図1**）．GIBでは着目するゲノムに対して，GCプロット解析やそのゲノムのもつ遺伝子群への相同性検索，キーワード検索が行えるほか，複数のゲノムを選択して同時にある問い合わせ実行し，ゲノムを超えた串刺し検索ができるようになっている．こうして簡易の比較ゲノム解析ができる環境を提供してきている．しかし比較ゲノム解析を実行してみると各ゲノムにいくつかの矛盾したアノテーション情報がついていることに気づく．例えば，Open Reading frameという注釈がついている領域内の3'側にそのORFのpromoter領域が登録されている場合がある．また，近縁種あるいは同種でゲノム構造がほとんど同じかつ配列相同性においては90％相同であるにもかかわらず，1つのゲノムではvirulence proteinという記述がありもう片方のゲノムでは，hypothetical proteinと記述されていることがある．後者のケースは，アノテーションを付加した時期に参照したアミノ酸配列データベースの

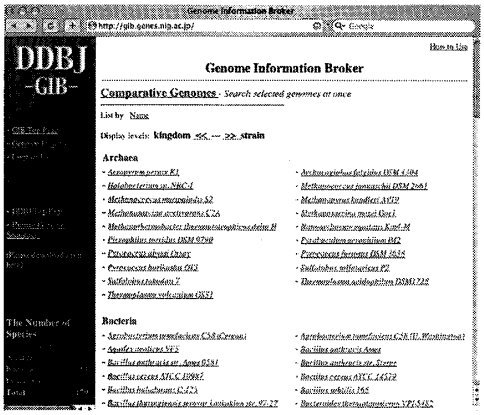

図1-1

図1-2

図1　Genome Information Broker の説明
1）GIB のホームページ．現在公開中のゲノムの種類がまとめられている．
2）GIB で見受けられるゲノムアノテーションの不具合例．

違いから導かれているように思われる．逆にいうと，「hypothetical protein」という注釈がつけられたゲノム配列が，「virulence protein」という注釈を付け得る時を超えて更新されていないとも言える．また，原著論文を参照するとわかるように，各ゲノムプロジェクトにおいて独立のアノテーションルールがあり一般にルールの統一がなされていない．例えば，ORF を認定する上での配列の長さも，あるゲノムプロジェクトでは30残基（90塩基）であるのに対してあるものは，400残基（1200塩基）を基準としているといったことがある（**表1**）．この結果，ゲノムの全長に比較して明らかに多くの ORF をもっているとされるゲノムがある．あるいは，pseudo gene の判定基準がまちまちであるために，pseudo gene の多さから ORF 数が多くなっているものもある．概して，比較ゲノム解析が効率よくかつ高品位に行われるためには，各ゲノムプロジェクトのアノテーションルールを統括し，一定の基準に準拠したアノテーションの再統一が求められているように思われる．

表1　予測遺伝子長の設定の違い

>5aa	1
>20aa	1
>30aa	24
>33aa	1
>33.3aa(100bp)	3
>40aa	1
>50aa	8
>60aa	4
>66.6aa	1
>80aa	2
>100aa	6
>150aa	1
>200aa	1
>300aa	1
>400aa	1

原著論文から各ゲノムプロジェクトのアノテーションの相違を外観した．特に，予測遺伝子の妥当性判定に用いられる遺伝子長が大きく違う場合がある．表中の60AAは60アミノ酸列長であることを示す．右側の数字は，その長さを基準としたプロジェクトの数．

オープンアノテーションによるデータの高品位化

前章で示したようなアノテーションの不具合は，利用者がGIBなどを利用している際に偶然見つけられることも少なくない．その場合，第3者からDDBJへアノテーションの不具合が報告されてくる．しかし，元のアノテーションについては，登録者が更新権利を持っており，そのような指摘にもかかわらず更新されないことが少なくない．一般にゲノムデータ増えるほど，前章で述べたようなアノテーションの不具合も増える傾向にある．DDBJでは一つの試みとして，ある遺伝子やゲノムの領域に対して複数のアノテーションがあった場合あるいは登録者以外からの指摘がある場合，それらをゲノム配列をベースにして網羅的に表示し，アノテーションの妥当性を利用者が判断できるような情報公開の仕組みを考案している．それが，「オープンアノテーション」である．例えば，よく研究されている大腸菌を例にあげると，完全長ゲノム配列の登録とは独立に，多くの研究者から遺伝子の情報が登録されてきている．これらを網羅的に比較してみると，ゲノムプロジェクトではアノテーションが省かれているが，それ以外の登録では，その箇所に遺伝子の存在が示唆されている領域を発見することがある．こうした視覚化が実現されるとゲノムプロジェクトの情報をだけを見ていた場合には考慮されなかったデータも利用者の判断で用いることができるようになる．

ま と め

これまでに紹介したように，ゲノムデータの爆発的な産出は，登録機関のデータ構造や公開システムに大きな変革をもたらした．研究資材としてのデータ利用の幅が増えたことは非常に大きな成果であるが，2，3章で紹介したように，アノテーションの質に対する危機管理を考慮して使用するようにしたい．もちろん，データのとりまとめをしている国際塩基配列データベースが，トップダウン式にアノテーションルールの標準化を図ることはできるであろうが，「逐次更新」という大きな問題が残っている．現在のゲノムアノテーションの基盤は，既存のアミノ酸配列データベースやモチーフデータベースのエントリーとの比較において抽出された情報である．従って，アミノ酸配列やモチーフ配列データベースが更新されれば，それまで未知であった配列にも機能情報が付加できる可能性があり，逐次的に網羅的な再アノテーションを実行するのが本意であろう．しかしながら，配列決定プロジェクトの多くは，期間限定のものであるため，5年後にそうした操作をできうる計算機資源と人材の確保を保証するものではない．その意味では，多くのデータが個々の研究者からの登録の積み重ねであったように，3章で紹介したオープンアノテーションを拡張して，個々の利用者が少しずつ最新の解釈を与えていける分散システムを構築する必要があろう．

文 献

1）International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. Nature 409: 860-921, 2001.
2）Fumoto M, Miyazaki S and Sugawara H: Genome Information Broker (GIB): data retrieval and comparative analysis system for completed microbial genomes and more. Nucleic Acids Research 30: 66-68, 2002.
3）Goto K, Miyazaki S and Sugawara H: Genome Information Broker for Data Retrieval and Comparative Analysis of Microbial Genomes. Journal of Japan Society of Information and Knowledge 10: 4-13, 2001.
4）Miyazaki S, Hashimoto H, Shimada A, Tateno Y and Sugawara H: A New File Format and Tools for the Large-Scale Data Submission to DNA Data Bank of Japan (DDBJ). In: Miyano S, Shamir R and Takagi T (eds) Currents in Computational Molecular Biology. Universal Academy Press, Inc., Tokyo, pp60-61, 2000.

2 *Helicobacter pylori* の病原性因子と胃十二指腸疾患

山岡 吉生・David Y. Graham
ベイラー医科大学内科

Virulence Factors of *Helicobacter pylori* and Gastroduodenal Disease

Yoshio YAMAOKA and David Y. GRAHAM

Department of Medicine, Baylor College of Medicine

Abstract

Approximately 20 % of *Helicobacter pylori* (*H. pylori*)-infected individuals develop clinically significant diseases such as peptic ulcer, gastric adenocarcinoma or gastric mucosa-associated lymphoid tissue (MALT) lymphoma. It is unknown which bacterial, host, and environmental factors are the critical determinants that predispose to these clinical manifestations of *H. pylori* infection. Host factors that influence acid secretion are potentially good candidates as factors that influence the outcome of an *H. pylori* infection. The proinflammatory cytokine, interleukin (IL)-1β, is a candidate in this regard in that it is expressed in the gastric mucosa infected with *H. pylori*. Here we show that IL-1 genetic polymorphisms influenced *H. pylori*-related gastric mucosal IL-1β levels and were related to gastric inflammation and atrophy, factors thought to be important in gastric carcinogenesis. Experience with other bacterial pathogens suggests that *H. pylori* strain specific factors also influence the pathogenicity of *H. pylori* as well as the risk of developing different *H. pylori*-related diseases. We show that the number of repeats in the 3' region of the

Reprint requests to: Yoshio YAMAOKA
Department of Medicine
Baylor College of Medicine
VAMC (111D), Medicine/GI
2002 Holcombe Blvd.,
Houston, TX 77030 USA