

**ゲノム(genome)**

ひとつの生物をつくるのに必要な最小の遺伝子のセット。

**遺伝子**

遺伝情報を持つ最小の機能的単位で、遺伝する形質に対応する。

**DNA(deoxyribo nucleic acid:デオキシリボ核酸)**

すべての生物の細胞に含まれている生命の維持に欠かせない高分子物質を核酸といひ、塩基、糖、リン酸が結合したものを(ヌクレオチドと呼ぶ)が単位となっている。糖の部分がデオキシリボースのものをデオキシリボ核酸(DNA)といひ、この塩基はふつう、アデニン、グアニン、シトシン、チミンである。なお、核酸のなかで糖の部分がリボースのものをRNA(リボ核酸)といひ、この塩基はふつう、アデニン、グアニン、シトシン、ウラシルである。

**ヒトゲノム計画**

ヒトには約10万個の遺伝子が約30億個の塩基対の染色体DNAに記録されていると推定されているが、これらの情報を解読する計画。まず、すべての染色体の塩基対配列を確定する。つぎに、どこかの塩基対配列がどの遺伝子に対応するかを確定する。解析結果のデータは膨大な量になるため、これを生物学的に意味を持った情報にして遺伝子データベースとして蓄積する。

**染色体**

細胞分裂の時に、核内に現われる特殊な色素に染まる棒状の物質で、遺伝子の集合体である。

**バイオリクター**

微生物や動物細胞細胞を用いて有用物質を生産するいろいろな形式の生物反応器をいひ、

**バイオテクノロジー**

生物機能に着目して開発された遺伝子組替え、組織・細胞培養、受精卵移植などの技術の総称。生物学(バイオロジー)と技術を組み合わせた造語。

**mRNA(messenger ribonucleic acid:伝令RNA)**

細胞の核内でDNAのもつ遺伝情報を転写したバク質合成の場であるリボソーム(細胞質内にある直径15~20nmの小顆粒)へ伝える役割をする物質。

**cDNA(complementary DNA)**

mRNAを鋳型として合成された、mRNAと相補的な塩基配列をもつDNA。

**クローン(Clone)**

遺伝子組成が同一の細胞群をいひ、分枝系と訳される。

**SNP(single nucleotide polymorphisms:1塩基多型)**

ゲノムの中にある塩基配列のなかで個人間で異なる塩基を持っている現象を「多型」といひ、塩基1つだけ異なることを1塩基多型といひ、ヒトの遺伝子は基本的には誰でも同じであるが、実際は個人で微妙に差があり、同じ薬が患者毎に効き目が異なるのはこのためである。

**バイオインフォマティクス(bioinformatics)**

生物學とどまらず物理学、数学までを含む生物学的な情報を扱う総合科学で、生物情報学、または生物情報科学と訳される。

## もう一つのIT革命への誘い

今世の中で盛んに言われているIT革命は、ECを中心とするE・ビジネスの分野で進行中である。これとは別に、もう一つのIT革命が、Bio Scienceとその関連ビジネスの分野で起きている。前者のIT革命は、全世界的にビジネスのスタイルを一変させつつあり、その波及速度は、いわゆるdog yearのごとくである。一方、後者のIT革命は、日米欧の研究者間の競争を激化させ、Bioの分野におけるベンチャー企業の勃興を加速させている。また、その変化の波及速度は、前者でのdog yearに対してmouse yearとでもいふべきほどの速さである。例のヒトゲノムの解読にしても、日米欧政府の国際ヒトゲノム計画とそれに対抗する米国Celera Genomics社などのベンチャー企業との競争により、国際ヒトゲノム計画は当初予定の2005年完了を2003年完了に前倒しした。Celera Genomics社にいたっては今年の6月中に解読を完了すると言っているようである。このことを可能にしている主要な要因の一つがBio ScienceとITの連携であり、Bio Scienceの分野に強大なITリソースを持ち込んだIT革命である。米国のCelera Genomics社におけるゲノム解析のITリソースは、1200個のCPUを相互接続したクラスターサーバ、数10TB(数年後に100TBまでを想定した)のデータベースと、公表されているハードウェアだけでも驚くべきリソースである。ハードウェアに加え、これを動かす管理するソフトウェア、さらにはゲノム解析ソフトウェアや解析プロセス管理のソフトウェアなど、想像するだけでも膨大なITリソースが当然存在するであろうことは、IT技術者なら容易に理解できるだろう。

Bio ScienceとITの連携は大きく分けて二つある。一つは、遺伝子解析のプロセス管理にコンピュータを利用すること、もう一つは遺伝子解析そのものにコンピュータを利用することである。遺伝子解析のプロセス管理にコンピュータを利用する例としては、DNAシーケンス解析のプロセス管理にコンピュータシステムを導入し、DNAシーケンス解析の品質や生産性を大幅に向上させている例がある。DNAシーケンス解析では、解析対象サンプルの作成、サンプル断片のDNAシーケンス解読(これはシーケンサと呼ばれるDNAシーケンス解読機で解読される)、サンプル断片群の解読結果をつなぎ合わせてサンプルの全長シーケンスを決定(これはコンピュータ上で処理される)等、多数のプロセスにおいて、膨大な数のサンプル毎にプロセスの進捗状況や品質、さらには生産性までも管理する。従来のように、DNAシーケンス解析が研究室レベルで行われていた状況では、解析対象サンプル数も少なかったが、現状では一日に数千、数万の単位のサンプルを解析するような状況になってきている。このような膨大なデータを生産し、管理するためのシステムの実態は、工場の生産ラインの生産管理システムを想定するとよい。解析対象サンプルを作成する装置やDNAシーケンス解読機が数十、数百と設置されており、それぞれがPCで制御されている。これらの制御PCが、全体を管理するコンピュータとLANで接続され、データベースに各サンプルのDNAシーケンス解析の進捗状況や、解析結果とその品質、生産量が集

められる。ここでの生産物はDNAシーケンスといひ情報であるが、システムの枠組みは工場の生産管理システムと同じものである。

一方、遺伝子解析そのものにコンピュータを利用する例としては、解析されたDNAシーケンスから、その中に存在する遺伝子領域を予測するアルゴリズムが研究され、いくつかのプログラムが新しい遺伝子の発見のために利用されている。これらのプログラムでは、遺伝子解析の研究の中で得られた一種のヒューリスティックな知識を用い、DNAシーケンスの中から遺伝子領域を予測する。例えば、「遺伝子領域の開始部分と終了部分には特定の塩基配列が存在する」、「遺伝子領域の開始部分の前方には特定の塩基配列が存在する」、さらには予測した遺伝子領域の長さ、A、T、G、Cの含有量の比率、等々。コンピュータ上では単にA、T、G、Cの文字列で表現されているDNAシーケンスの中からこれらの知識を使って遺伝子領域の候補を発見する処理は、多量のコンピュータパワーを必要とし、またその予測精度も問題になる。このため、処理を最適化するアルゴリズムの研究や、予測精度を上げる研究が取り組まれている。このような課題は、AI研究の分野での課題と類似したものを想定してもらってもよい。

新しく発見した遺伝子候補が、既に発見され登録されているものと似たものかどうかを調べるために、homology searchのプログラムがよく使われる(homologyとは日本語で相同性と訳される)。これは、新しく発見した遺伝子候補のA、T、G、Cの配列データをキーにして、データベースとして登録されている既存の配列データの中から類似したものを探すプログラムである。このようなデータベースとしては世界中の研究機関においてWebで公開されており、その数は4、5百ともいわれている。また、登録されているデータ量も毎年指数関数的に増加しており、最も代表的な米国NIHのGENEBANKでは登録数がこの4月で600万を越えている。homology searchでは経路探索の手法であるDP(dynamic programming)が使われている。DPでは、データ量が増加すると処理量が爆発的に増大するため、実際のhomology searchのプログラムでは、処理量の増大を抑えるため、種々の改良が加えられている。このような課題にしてもcomputer scienceにおける課題そのものである。もちろん、遺伝子領域の予測、さらには遺伝子の機能、構造の解析にITが不可欠ではあるが、当然Bio Scienceに関する知見も必須である。

Bio Scienceとその関連ビジネスの世界で起きているもう一つのIT革命では、ITの研究者や技術者が取り組むことのできる課題がいたるところにある。米国では、この分野に多数のITの研究者や技術者が参加し、Bioの研究者と共同で成果をあげているようである。日本でも、この動きは進みつつあるがまだまだである。ITの研究者や技術者がこの分野に関心を持ち、参加する人が増えることを期待したい。



理化学研究所  
ゲノム科学総合研究センター  
遺伝子構造・  
機能研究グループ

吉田清  
(元NTTソフトウェア  
技術開発部長)