1 バイオデータベースの歴史と展望

平川美夏
京都大学化学研究所
バイオインフォマティクスセンター
hirakawa@kuicr.kyoto-u.ac.jp

分子生物学の台頭とともに成長した核酸，タンパク質のデータベースは，ヒトゲノムプロジェクトを背景として量，質ともに急激に増大した．ゲノムプロジェクトは，さまざまな生物のゲノム塩基配列データを万人が利用できる状況を作り出し，さらに，網羅的にデータを獲得し，系統的に解析するゲノムのアプローチを発展させた「ポストゲノム研究」を生み出している．ここでは，分子生物学の進展とともにデータベースがどう拡大し，変化してきたかに注目してバイオ分野のデータベースのあり方を見直す．データベースを生命科学分野の頭脳として共有するために，専門家による知識編纂の支援や巨大で複雑なデータを統合する方式の新展開が期待されている．

分子生物学とデータベース

本稿では，分子生物学のデータベースをその誕生の背景となった出来事とともに紹介していくことにする．まずデータの発生源に注目して，データの集積を担うレポジトリとしてのデータベースが，研究方法や情報技術の変化とどのようにかかわっていったかについて述べる．さらに，データベースをデータの保存庫から知識の宝庫へと育てるために培われてきたアノテーションを通じて，バイオデータベースの今後を考える．

分子生物学データベースの誕生

1982 年，核酸（DNA，RNA）の塩基配列のデータベース GenBank（米）[1]，EMBL（欧）[2] がデータリリースを開始した．DNA の塩基の並び順を決めるためのシークエンシング技術[3] が開発されてから 5 年後のことである．DNA は，生物の親から子へと受け継がれる遺伝の情報を担い，生物の形を作り，体内で起こるさまざまな現象の中心となるタンパク質の構造を指定する．DNA は，アデニン（A），グアニン（G），シトシン（C），チミン（T）の 4 種類の塩基という化合物が一直線の糸上に繋がった高分子であり（**図 -1**），たとえば，ヒトゲノム 30 億塩基という場合，これが 30 億数珠繋ぎになっていると想像できる☆1．タンパク質を作る遺伝子の DNA の並び方は，アミノ酸の配列を一意に決定する☆2．この変換規則は，コドンと呼ばれ，タンパク質を構成する 20 種類のアミノ酸それぞれに塩基 3 つの並びが対応している．DNA のクローンとは同じ配列の DNA を指すが，DNA にはタンパク質にはないコピーを簡単に増やせる性質がある．そこで研究対象としているタンパク質の遺伝子 DNA のクローンを得ることは，分子レベルの研究を進める上で強力な切り札となった．さらに DNA シークエンシングは，20 種類の分子を振り分ける必要のあるアミノ酸配列の決定よりずっと容易であったため，瞬く間に普及していった．微生物以外の生物の遺伝子 DNA は，タンパク質に変換されるエクソンとタンパク質を作る仲介分子の mRNA を作るときに切り取られてしまうイントロンがある（**図 -2**）．エクソンとイントロンの境目は，配列から簡単には決められない．そこで

☆1 実際のヒトゲノムは，24 本の染色体に分かれており，一続きではない．また DNA は 2 本の分子がより合わさった二重らせんの構造をとるので，その片側 1 列が 30 億塩基という意味である．
☆2 生命科学は，常に変化しており，ここでの説明は，物理学のニュートン力学のような教科書的説明と理解いただきたい．

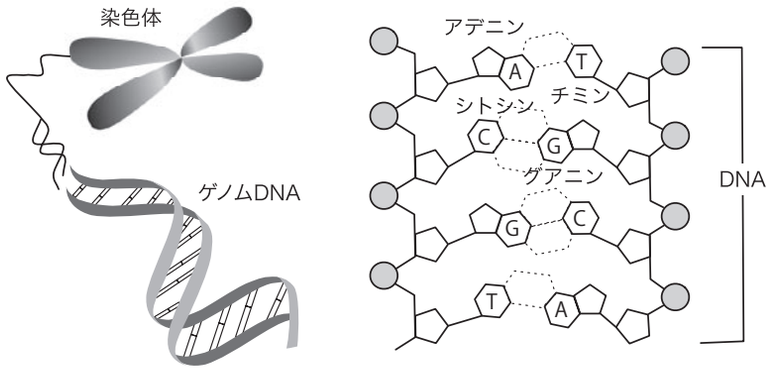

図 -1　ヒト染色体とゲノム DNA（左）　DNA の構造（右）

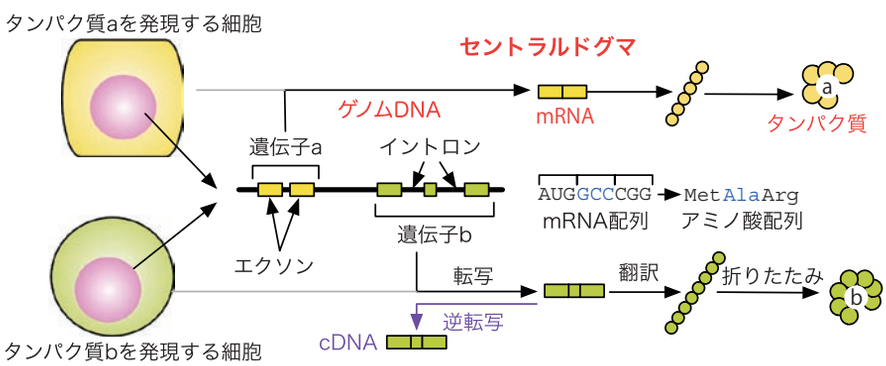

図 -2　遺伝子発現の流れ

イントロンが除かれた mRNA を DNA に再変換（逆転写）した相補的 DNA（cDNA）のセット（cDNA ライブラリ）を作り，その中から目的の遺伝子クローンを選択することで，確実にまた短いシークエンシング作業で遺伝子配列を決められるようになった．このようにして得られた遺伝子の DNA 配列は，AGTC の 4 文字から構成されたテキストとして文献に記述され，それを集めたものが塩基配列データベースの始まりである．

タンパク質立体構造のデータベース PDB（Protein Data Bank）[4] は 1971 年に開設され（2.b）バックボーンデータベースの標準化：PDBj 参照），この当時 200 以上のデータが蓄積されていた．1953 年に Watson と Click が DNA 二重らせん構造を X 線結晶解析データから予見し，次のターゲットとして，タンパク質の立体構造が盛んに解析されたためである．しかし，タンパク質の構造からは，DNA のような法則は見つけられずにいた．

DNA から mRNA に配列情報が転写され，コドンを用いてアミノ酸配列に翻訳されタンパク質に至る，遺伝情報の発現の流れを「セントラルドグマ」という☆3（図 -2）．つまりこの当時，セントラルドグマを記述する情報がすべてデータベースに蓄積され，記号や数値として扱えることになった．その 1982 年に，Oxford Press のジャーナル Nucleic Acids Research（NAR）[5] に，すでに核酸とタンパク質の計算機解析に関する特集が組まれている．生命の本質を担う物質が，単純なテキストとして扱えるのならば，数学や情報処理の専門家が目をつけないはずがなかろう．分子生物学の黎明期も，生命が物質で扱えるならと生命科学に挑戦した物理学者や化学者が中心であった．そして，生命が情報として扱える時代を迎え，生命の原理の探求に向けて，情報科学が名乗りを上げ，それに応えるようにデータベースが成長していったと言える．

☆3「セントラルドグマ」は，Click が提唱し，1 遺伝子から 1 タンパク質が作り出される分子生物学の「原理」とされたが，実際は 1 遺伝子から複数のタンパク質が作られるため，すでに普遍的原理ではない．遺伝子発現の情報伝達の流れの典型と考えればよい．

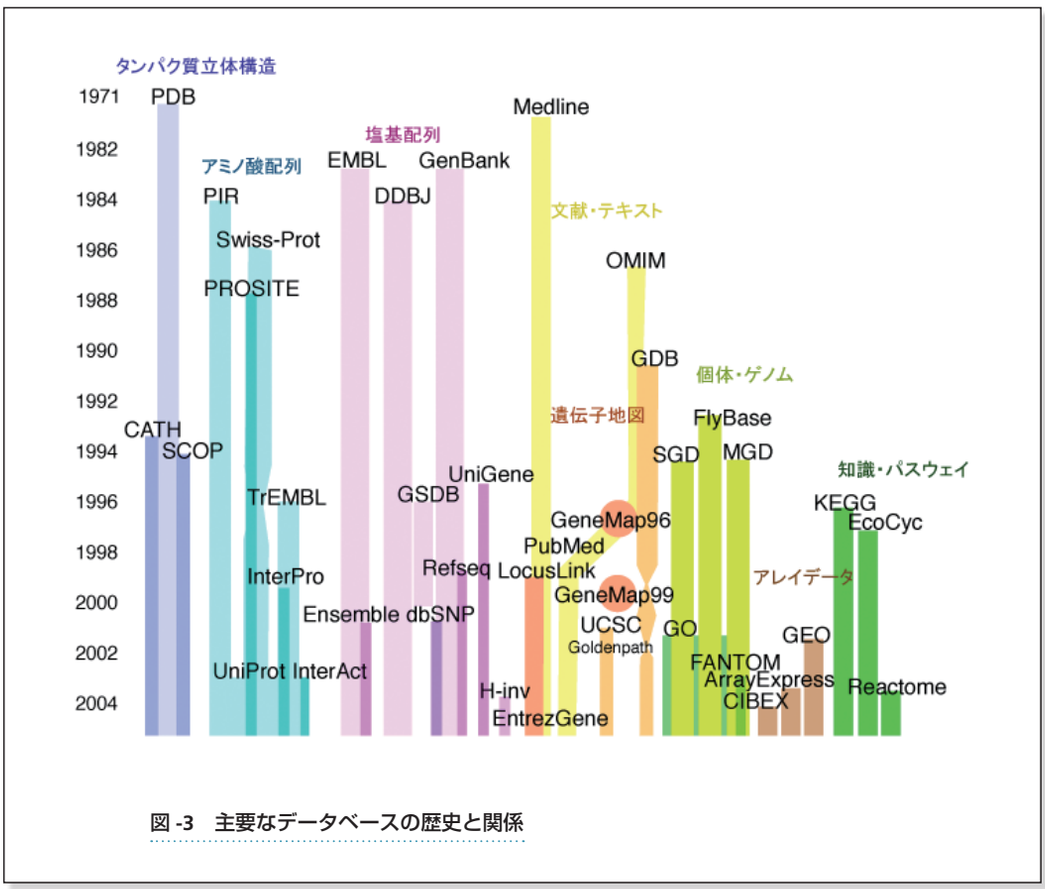

図 -3　主要なデータベースの歴史と関係

図 -3 には，主なデータベースの歴史と関係を示した．本稿では，この中のいくつかの代表的なデータベースにしか触れていないが，インターネットから探索可能であろう．また，前述の NAR 誌は 1 月にデータベース特集号（Database Issue）を発行しており，オンラインで参照することができる．公共データベースの更新情報や，新規に発表されたデータベースの紹介記事やバイオデータベースのリンク集が提供されている．

ゲノムプロジェクトとデータベース

1990 年，ヒトのゲノム DNA，30 億塩基を 15 年かけてシークエンシングする「ヒトゲノムプロジェクト」が開始される．ヒトゲノムプロジェクトは，米国国立衛生研究所（NIH）とエネルギー省（DOE）[6] の下で計画，実行されたが，任意参加の国際プロジェクトの一面も持っていた☆4．プロジェクトは，「ヒトゲノム全塩基配列を医学と生物学の進歩のための人類の共有財産として完成させ，ヒトの全遺伝子のカタログを作る」ことを目的として掲げていた．DNA の塩基配列が共有財産であるという発想には，DDBJ/EMBL/GenBank による国際塩基配列データバンク（INSD）（2.a）バックボーンデータベース：DDBJ 参照）が，誰もが自由に利用できる公共のデータリソースとして生命科学を支えていたことが背景にある．

プロジェクト開始の年にヒトゲノムプロジェクトの公式データベースとして，Johns Hopkins 大学に GDB（Genome Database）が開設された．GDB の前身は，Yale 大学を拠点としていたヒト遺伝学のコミュニティのデータベースであったが，遺伝子やゲノム上の目印である DNA マーカの位置情報を収集していたことから，シークエンシングのための試料の管理機能を期待されて選ばれた．そしてコミュニティの手から開発部隊へと移され，リレーショナルデータベース（RDB）に移植された．RDB 管理システムのベンダが，大量データの管理が必要になるプロジェクトの参加機関に「ゲノム割引」料金を設定したことも手伝い，ゲノム研究での RDB の利用が進んだ．

ヒトゲノムプロジェクト開始当時，1 つの生物の全ゲノム配列というのは，数万塩基のウィルスで決まっていた程度で，一足飛びにヒトというのはいかにも無謀であった．そこでプロジェクトの一部で，テストケースとしてのモデル生物のゲノムプロジェクトも実施された☆5．

☆4 このように言うと分かりにくいが，国際協力や支援の参加の意思は，一般に各国の政府の判断に任されている．米英日仏独中の 6 カ国が参加．

☆5 実際は，シークエンシングのモデルとしての意味はあまりない．ヒトゲノムプロジェクトの一番お手本となったのは，ノーベル賞を受賞した Sulston の線虫研究である．

モデル生物には，大腸菌，酵母菌，線虫，ショウジョウバエ，マウスなど，実験生物としての歴史があり遺伝学などの知識が利用できるものが選ばれ，これまでの研究を支えてきた研究者のコミュニティが中心になって，試料整備やデータベース開発が行われた．そしてマウスのMGD（Mouse Genome Database）[7]，ショウジョウバエのFlybase[8]，酵母菌のSGD（Saccharomyces Genome Database）[9]などのモデル生物のコミュニティデータベースが次々と産声を上げた．この後，WWW（World Wide Web）が急速に広がり，コミュニティやプロジェクト単位でのデータベースの作成，提供に利用されるようになった．特に染色体やゲノムに対するクローンや遺伝子の位置の情報は，画像が大きな伝達能力を果たすため，ゲノムデータの公開には，たいへん有効であった（5. ゲノムデータの視覚化による効果的な理解：参照）．

ヒトゲノムプロジェクトは，その進行において，INSDにさまざまなデータを発生させ，爆発的なデータの増加をもたらしていた．1990年代半ばには，自動シークエンシングが進歩し，シークエンサーによる大量データ産出が始まった．そしてまず，EST（Express Sequence Tag）というcDNAの断片の配列データの大量登録が始まった．ESTは，ターゲットを絞らず，cDNAライブラリのクローンを端からシークエンシングしたものである．つまり，遺伝子の配列には違いないが，何のタンパク質の遺伝子かは分からないものが含まれる．これは「何々の遺伝子配列」という単位で配列を登録していたINSDに，どこから採取されたかの情報しかない配列が登録される契機となった．

いよいよヒトゲノムシークエンシングが本格化すると，INSDは，これまでのデータ収集とは異なる役目を期待される．重複や競合を避けるため，ゲノムのどの位置をどの国のどこの機関が担当するという分担表が作られ，分担をこなしているかを表明することが参加機関の義務となった．そこで進捗状況の管理とデータの共有を目指して，INSDに日々のシークエンシングの途中経過をすべて登録することになった．GSS（Genome Survey Sequence）カテゴリには，シークエンサーの出力データをすべて登録し，HTG（High Throughput Genome Sequence）には，精度を上げ配列を連結するフィニッシングの作業の過程のデータを更新のたびに登録する取り決めが交わされた．塩基配列と名の付くデータは，すべて登録するレポジトリデータベースとして，INSDは，バージョン管理や大量データ受付システムを開発してこの困難な要求に応えた．この頃には，インターネットは十分に普及しており，それが日々のデータ登録や大量データ転送を可能にした．WWWや電子メールが活用されることで，国境を越えたデータ確認や調整が容易になったことも，プロジェクトの進展を大きく加速させた．

1998年，国際コンソーシアムによるヒトゲノムプロジェクトに転換期が訪れた．Venterが代表を務めるセレラ社が，民間企業として国際コンソーシアムより先にヒトゲノムを決定すると宣言したのである☆6．その結果，国際コンソーシアムは計画の短縮を余儀なくされ，シークエンシングに予算を集中するためにGDBの打ち切りを決めた．GDBは，プロジェクトの顔としてオブジェクト指向型のデータ形式を取り入れるなど，技術的な先進性を追い求めたが，生物系の研究者の支持が得られず，それも中止の一因と言われていた☆7．これを機にGenBankを運営するNCBI（National Center for Biotechnology Information）[1]が，ヒトゲノム情報提供の中心を担うことになった．GDBや遺伝子と疾患の関係をまとめたOMIM（Online Mendelian Inheritance in Man）のデータを取り込み，ヒトゲノムの完成に備えた．国際コンソーシアムのデータは，カリフォルニア大Santa Cruz校[10]のバイオインフォマティクスチームが編集し，Golden pathと呼ばれた．その一方で，ゲノムを完成するのに必要なデータは，すべてデータベースに公開されているため，いくつかの研究チームがゲノムの再構成を試み，ライバルであったVenterもセレラ社のヒトゲノム配列の完成に利用していた．Venterと国際コンソーシアムの戦いは，それぞれが同時にヒトゲノム配列を完成させたとして，同着で握手の記者発表があった．その後民間企業であるセレラ社は，株主の利益を優先するためにデータをINSDに登録しないことを表明したことが発端となり，協力関係には至らなかった☆8．学術発表される塩基配列は，すべてISNDに登録される，という伝統はここで崩れ，以降商業目的の配列は，文献発表されても公開されないことも起きている．現在，ヒトゲノムの完成データは，ゲノムプロジェクトで発生したデータだけでなく，データベースに登録されているヒト配列データ全部を用いてNCBIで編集され公開されている[11]（このゲノムの再構成データはBuildと呼ばれ，最新データはBuild36[12]である）．

ゲノムプロジェクトが築いたエポックは，生物に由来する完全なデータに手が届いたことにほかならない．そしてそれがすべてデータベースに格納されている．そ

☆6 Venterは，細菌の全ゲノムシークエンシングを世界で初めて完了した気鋭の研究者であり，シークエンサーの最大手企業アプライドバイオシステム社が，Venterに最新機器を提供してセレラ社が誕生した．

☆7 GDBは，公的予算を打ち切られた後も，当時の開発者を中心に公開を続け，今は米カリフォルニアの第三セクタRTI（Research Triangle Institute）で継続されている．

☆8 2001年2月国際コンソーシアムがNature Vol.409 No.6822に，セレラ社がScience Vol.291 No.5507に論文を発表し，どちらの号の表紙にも“The Human Genome”とある．

れまでのデータベースのデータは，たとえば発癌とか，AIDS とか，最新の研究ターゲットとして魅力的でインパクトのある分野に関連する遺伝子のデータが目立ち，全体として見た場合偏りがあることが当然であった．しかし，全ゲノムのデータベースは，その対象生物が存在し，生存するためにゲノムとして維持している情報のみ，そのすべてを含むため，データベース全体を意味のあるひとまとめとして解釈することが可能になる．つまり，統計的解析によるデータベースの分析の結果は，その生物を研究して得られた生物学的発見と見なすことができる．個々の遺伝子の配列を探すことから始まった DNA シークエンシングは，ゲノム配列全体に一気に対象を広げ，データ量を増やしただけではなく，木から森へと生物学研究の視点を変えた．

オーミクスとデータベース

ゲノムを「Gene（遺伝子）+-ome（全体を意味する接尾語）=Genome」として，ある生物の遺伝子の全体という解釈がある☆9．そして個々の遺伝子から（木）ではなく，ゲノム（森）からそこに含まれる遺伝子（木）を見る研究を「ゲノミクス（Genome+ics（学を示す語尾）=Genomics)」という．ここから派生して，網羅的にデータを獲得し系統的に解析する手法に対して，その研究対象に「オーム (全体，-ome)」をつけて造語を作り，その全体を研究対象とする新たな生物学的アプローチ「オーミクス（Omics)」が生まれた[13]．オーミクスとしては，遺伝子発現の状態を mRNA の転写の有無や量の変化を網羅的に測定して調べる「トランスクリプトーム（transcript+ome，転写されたものの全体)」や，特定の細胞にある全タンパク質を測定する「プロテオーム（protein+ome)」などがよく知られており，ゲノムを生命の設計図とすると，トランスクリプトームが生命の注文書，プロテオームが生命による製品群と例えられている．これらのオーミクス研究においては，大規模な処理が可能な実験機器を使って自動測定でデータを獲得し，その結果の大量の測定値はデータベースに保存され，解析も当然ながら計算機処理によって行われる．オーミクスでは，研究の出口が計算機に託され，その結果の成否は情報科学に大きく依存する．

ポストゲノム☆10 のオーミクス研究として，DNA マイクロアレイによるトランスクリプトームの研究が盛んである．たとえば，ある特定の条件（癌と正常組織，投薬前後等）で転写の量が変化する遺伝子を網羅的に検出し，条件の変化が遺伝子発現に及ぼす影響を調べることができるので，医療の診断や創薬ターゲットの探索などの応用が期待されている．網羅的測定によるオーミクスデータは，データ量が多く，論文誌の紙面に全データを載せることはない．したがって，現状では supplementary data として Web 上で提供されることが一般的であるが，塩基配列等の場合と同様に（2. バックボーンデータベースの課題と展望：参照）投稿時に公的データベースに登録することが知的資産の保存のためには有効である．しかし，同じオームから派生していても，オーミクス由来のデータの多くはゲノムのようにスタティックな対象ではなく，経時変化する実験条件下の測定データであるため，標準化が課題である．つまりゲノム配列データは，2 つの試料でシークエンサーの測定値が異なっても，4 種の塩基のいずれかに判定され，配列が一致すれば同じ DNA データになる．それに対して，マイクロアレイの測定値は，時間経過や条件による相対的変化のデータであり，測定のたびに値も基準も異なるので，単純に比較はできない．したがって，データベースにデータを格納することはできても，測定値を同等なデータとして扱えないので，詳細な実験条件だけが検索の手がかりになる．DNA マイクロアレイのデータは，塩基配列，タンパク質立体構造などと同様，日米欧に登録サイトを持つ MGED（Microarray Gene Expression Data）ソサエティが，データベース登録標準形式を提案し，論文投稿時の指定形式としても採用されている[14]．

ある臓器や細胞で遺伝子として発現している mRNA 配列を片端からシークエンシングした EST は，トランスクリプトームの先駆けともいえる．mRNA は，遺伝子の発現量によって何コピーも作られるので，同じ EST 配列が多数得られ，データベースに重複して登録された．塩基配列の場合は，配列比較により選別や連結が可能なので，NCBI では重複した配列から代表配列を抜き出し，良質な配列の最小セットである RefSeq データベースを作り，利用者の便宜を図っている．配列データは，質量ともに多様化しながらも，編集整理していく方法が経験的にも計算機処理としても確立してきた．しかし，すでにオーミックス研究のデータなどさまざまな種類のデータが生まれ，今後も研究の進展によって多様化は一層進むであろう．異なる生き物を異なる研究方法で扱っているデータは，形式としては結びつかないものかもしれない．しかし，バイオ研究の目的は生命現象の理解であり，その探求の手がかりとしては，必ずどこかで結びつけられるべき意味を共有している．このような生物学的な意味の統合が今後のデータベースの課題となっていくだろ

☆9 本来の語源は，遺伝子 gene の gen- と染色体 chromosome の -ome の合成語と伝えられている．

☆10 ポストゲノム，というと，ゲノム研究が終わったかのような誤解を招くと関係者にお叱りをうけるが，ゲノムのアプローチから派生した研究をこう呼んでいる．

う．しかし，ヒトゲノムプロジェクトの完了が宣言されて 3 年を迎えようとしているが，その目標であった「ヒト全遺伝子のカタログ化」もまだ完全には実現していない．次に，そこに至る過程としてのアノテーションに注目して，生物学の知識のデータベース化について考える．

知識の記述とデータベース

データに注釈をつけ，データの特徴や意味を記載することを「アノテーション」という．配列データベースにアノテーションによって生物学的な知識を追加することは，文字の並びとなった配列データに生命を吹き込む重要な過程である．

アノテーションは，アミノ酸配列データベースが発達させてきた．シークエンシング結果の公開のために研究者自ら登録する塩基配列データベースでは，提出者が出した通りの記載しか受け入れられない．一方，アミノ酸配列データベースは，塩基配列をアミノ酸配列に変換してデータを作成するので，データベース開発者自らが入力データや記述法を決めることができる．PIR（Protein Information Resource，米ジョージタウン大学）[15] は，配列の類似性からタンパク質のファミリー（構造も機能もほぼ同じグループ），スーパーファミリー（ファミリーより類似性は低いが構造，機能の類似が認められるグループ）などの分類を行っていた．その結果から，生物の進化と同時にタンパク質の配列にも進化にあたる変化が起きることと，そのアミノ酸の変化の傾向を発見した．これを行列表現した PAM（Point Accepted Mutation）は，タンパク質類似性検索に使われている．ジュネーブ大学の Bairoch らは☆11，Swiss-prot[16] に登録された機能未知のタンパク質のアミノ酸配列データに，有用な情報を付加する工夫を行っていた．同じ生化学的性質を持つタンパク質の配列を比較し，共通する配列パターンを探し出して，そのパターンとタンパク質の性質の関係を PROSITE というデータベースに編纂した．そしてタンパク質配列中の PROSITE のパターンの有無から，予想される性質をアノテーションに利用した．アミノ酸配列データベースは，2002 年から PIR，Swiss-prot，EMBL を翻訳した TrEMBL[2] を集約した Uniprot データベースの構築が開始された [17]．それぞれのデータベースの特徴あるアノテーション情報を集約し，知識データベースとしての価値を高めることが目的であり，増加し続けるデータに対応してアノテーションをすることは，組織レベルの連携を要する遠大な事業になっていることを意味している．

EST やゲノム中の遺伝子候補の機能未知の配列に，生物学的な手がかりを与えるアノテーションの第一歩は，よくアノテーションされた他のデータベースを配列の類似性から検索し，結果を参照し取り込んだり，前述の PROSITE のような配列パターンなどを探して情報を付与することである．この判断の経験則を自動化したプログラムによるアノテーションも盛んに行われている．しかしこの方法は，配列が似ていれば機能が似ているという分子レベルでの情報にのみ適用可能な方法にすぎない．また，比較元となった情報を新規の配列のアノテーションに使うため，情報が使われるたびに同じ情報の記述が増えていく．特に，定義名には，参照元の関連配列（XXX related sequence）などという記述が乱用され，キーワードによる検索を困難にする原因となっている（**図 -4**）．

生命現象の記述には，いつどこで，どのような他の分子とかかわるタンパク質なのかといった，データベースの品質管理の専門家であるキュレータが判断して付与するしかない知識データが多い．知識のデータベース化は，すでに文献に書き記された情報や研究者の間で共有されている知識を整理して記載するので，人依存的になりがちである．またキュレータによるアノテーションは，時間もかかり人材も限られるため，データの増加になかなか追いつかないのが現状である．したがって，作業の効率化や人為的なばらつき等の回避のための支援が，良質のデータベース構築に欠かせないものとなる．そこで自動化の精度向上やデータ付与のための条件設定の高度化，またキュレータの作業のためには，既存の知識をもれなく検索するための文献検索支援システムや，適切な表現を見つけるための用語の整備と候補語の提示システムなどの開発が期待されよう．

ゲノムプロジェクトも終盤を迎えた 1999 年頃には，遺伝子のアノテーションのための統制語を整備する GO（Gene Ontology）のプロジェクトが開始された [18]．GO は，遺伝子のアノテーションを実際に手がけている研

☆11 現在は，スイスバイオインフォマティクス研究所．機関名の変更や運用主体の変更は，頻繁に起こるので，文中では話題の当時に合わせており，全部列挙はしていない．

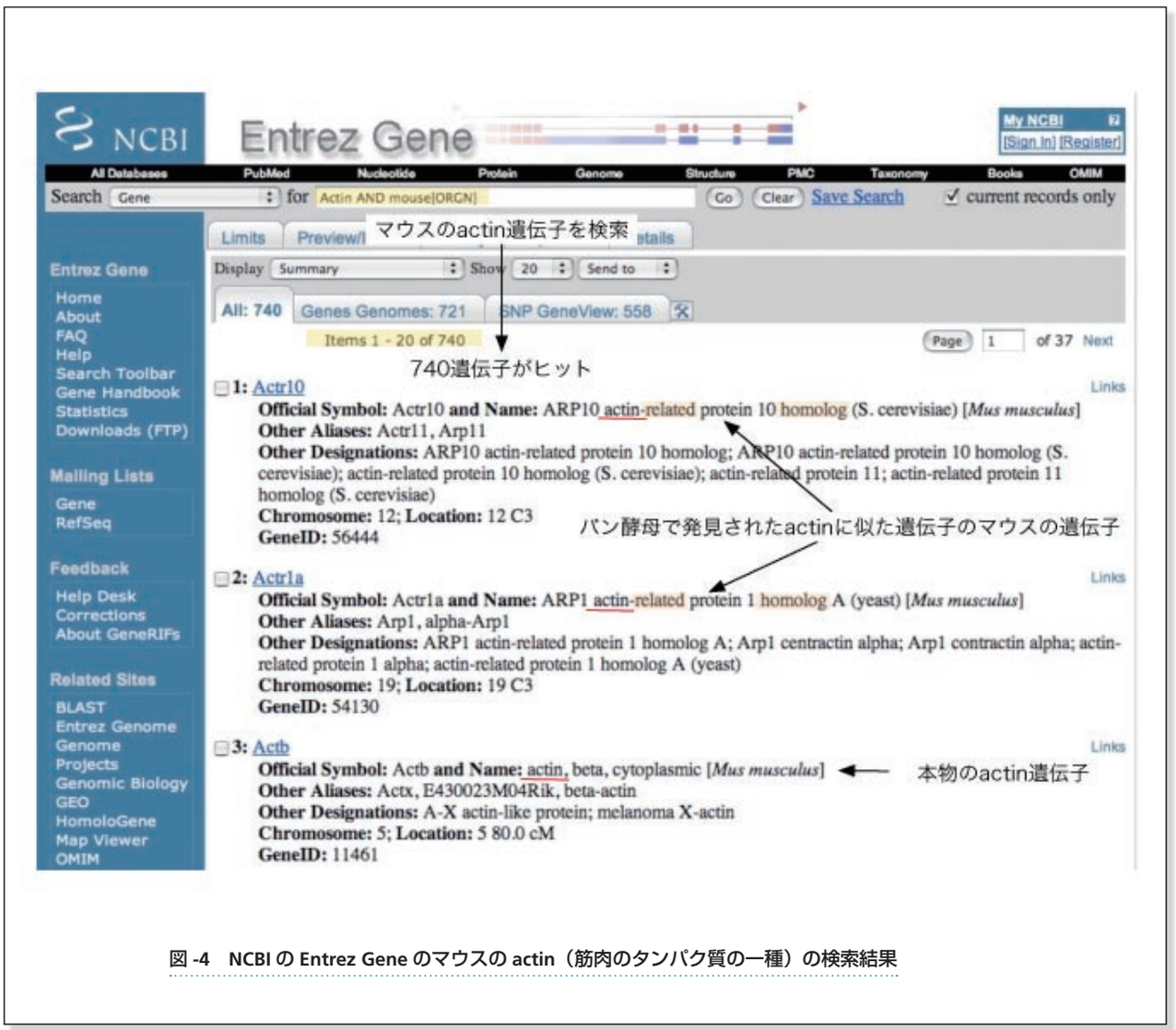

図 -4　NCBI の Entrez Gene のマウスの actin（筋肉のタンパク質の一種）の検索結果

究者のコミュニティが集まって作成している専門用語の体系であり，自身はデータベースではないとしているが，これを付与した遺伝子データベース間は，共通の知識体系で結ばれることになる．20 種以上のデータベースが GO の設けた基準によって，アノテーションを実施しており，これらのデータをリンク，引用することで，さらに多くのデータベースで用語が共有されている．GO は，遺伝子アノテーションというポストゲノム時代における，ゲノムデータの効果的なデータベース化と有効活用の要求から生まれた．生物の全ゲノムシークエンシングは，もはや当たり前の生物学的手法であり，全ゲノムが決定した生物も 300 種に届く日も近い．しかしこれは，シークエンサーの機能の向上やコストの低下，その他の関連技術が安定し，普及してきたことによるところが大きく，その一方でシークエンサーを稼働させる何倍もの時間が，人手による配列の解釈に費やされている．地道な実験に裏付けられた知識をいかに取り込み，ゲノムデータを解釈して生物学的理解へ結びつけるかは，まだまだ模索が続くであろう．その中で語彙と知識（オントロジー，4．バイオ知識の形成と表現：参照）の共有を基本としたコミュニティによるアノテーションは，確実な基盤となるはずである．

バイオデータベースのこれから

技術進歩と飽くなき探求により増加し続けるデータをその費用と労力以上に利用する活路を見いだすには，既存のデータベースを維持し，発展させることと，十分な知識に裏付けられた表現を用い，的確な編纂に基づくデータベースの開発が必要である．レポジトリデータベースは，アーカイブとしての安定した運営を大前提としながらも，データにおいては，解析技術の多様化や新規の知識の増大により追加修正を必要とされ，管理，運用においては，データ量の増大や，利用環境の発展により常に対応を迫られる．たとえば，シートにデータを記入し郵送していた著者登録は，いつの間にか Web ブラウザ

から入力するのが当たり前になった．このように，ゲノムプロジェクトを経て大きく躍進したバイオデータベースは，同時にインターネット技術や計算機の高速化などの情報分野での発展にも大きく影響を受けている．

バイオデータベースの最先端を担う NCBI は，この 20 年近いバイオデータの変革期に，どのようなデータ種をどう追加するかということに対しても，常に時代や研究者の要請に応えてきた．この 1 つの理由に，ASN.1（Abstract Syntax Notation One）フォーマットによるデータの管理がある．複雑な階層を持つデータ構造を記述でき，計算機親和性の高いフォーマットを一貫して使うことで，DNA，タンパク質などの階層性を持つ配列情報から，疾患，生物分類などあらゆるデータを統合し，さらに 50 年の索引作業の資産を持つ医学文献データベース Medline を PubMed としてすべてのデータに配している．複雑で種類の多い生命科学のデータを扱う場合には，信頼性の高い構造化されたデータや知識が，いかに強力か，そして重要かを教えてくれる．XML 化を中心としてさまざまなデータ交換形式が提案され試されているが，データベース間の取り決めにすぎず，生命情報全体への適応にはほど遠い．生命システムの構造に立ち返り，その再構成を実現するためのデータフォーマットの考案も情報科学と生命科学の協力を要する課題である．

新しいデータを追加するだけでなく，既存のデータを最新の学問事情に応じて，改訂していくための指針や方策も問われている．ヒトゲノム完成を目前にした頃から，有志によるアノテーション（Third Party Annotation）が更新に有効として試みられているが，これが一番機能しているのは，関心を同じくしている研究者によるコミュニティデータベースである．その基盤を支えるオントロジーやデータ交換形式を，生物種，データベースを超えて共有していくことが真の知識の統合への発展の鍵となるであろう．生物学の知識と実験技術の進歩を背景に，データベースも成長していく必要がある．生命の設計図をテキスト処理して進んできた生命情報の情報科学は，より高次の生命現象をいかに標準化して，どう取り扱うことを目指すのか，まだようやくそのスタート地点に立ったところである．

参考 URL

1）http://www.ncbi.nlm.nih.gov/
NCBI の Web サイト．GenBank，Refseq，PubMed，OMIM など，統合検索環境 Entrez ではすべてのデータベースを検索できる．

2）http://www.ebi.ac.uk/Databases/index.html
European Bioinformatics Institute (EBI) の提供するデータベースサイト．EMBL 核酸データベース，Uniprot データベースや，Ensemble ゲノムブラウザなどがある．

3）http://www.dnalc.org/ddnalc/resources/sangerseq.html
コールドスプリングハーバー研究所の DNA 学習センターが提供する DNA 解析技術の学習ページ．シークエンシングの方法をアニメーションで再現している．最近の手法は，メニューの Cycle Sequencing である．後述のモデル生物，DNA アレイもある．

4）http://www.rcsb.org/pdb/
PDB の米国，RCSB（Research Collaboratory for Structural Bioinformatics）サイト

5）http://www.oxfordjournals.org/nar/database/c/
Nucleic Acids Research のオンラインサイト．データベース特集号 Database Issue は，無料で見ることができる．

6）http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
米国エネルギー省のヒトゲノムプロジェクト情報サイト．

7）http://www.informatics.jax.org/
マウス研究の中心である Jackson 研究所のマウスデータベースサイト．

8）http://flybase.bio.indiana.edu/
インディアナ大学にあるショウジョウバエデータベースのサイト．

9）http://www.yeastgenome.org/
スタンフォード大学のパン酵母（Saccharomyces cerevisiae）データベースのサイト．

10）http://genome.ucsc.edu/
カリフォルニア大学 Santa Cruz 校のゲノムブラウザのサイト．

11）http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.figgrp.1442
NCBI の Build データの編集プロセス．

12）http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=9606&build=36
NCBI の Build36 データの統計情報．使ったデータ，編集結果のデータを提供している．

13）http://www.lsbm.org/omics/index.html
東京大学 RCAST システム医生物学ラボラトリの Omics の解説サイト．

14）http://www.mged.org/
Microarray Gene Expression Data (MGED) Society のサイト．標準形式の検討過程やオントロジーの整備の情報が提供されている．

15）http://pir.georgetown.edu/
米国ジョージタウン大学の Protein Information Resource (PIR) サイト．

16）http://www.expasy.org/
Swiss Institute of Bioinformatics (SIB) のプロテオミクスサーバ．Swiss-prot や PROSITE などのデータベースがある．

17）http://www.uniprot.org/
Uniprot データベースの提供サイト．

18）http://www.geneontology.org/
Gene Ontology プロジェクトのサイト．アノテーションへの適用基準なども提供している．

（平成 18 年 2 月 1 日受付）