［特集］ポストゲノム時代に高まるバイオ自然言語処理への期待：バイオ自然言語処理最新事情

1 語る科学へ向けて
―データ・知識・生命現象をつなぐ―

中村 桂子
keinaka@brh.co.jp / JT生命誌研究館

生命科学は，ヒトゲノムプロジェクトを契機に1つの転換期を迎えている．つまり，生命体を構成する因子を網羅的に解析し，データを蓄積することが技術的・資金的に可能になったのである．ところで，網羅されたデータは，実際の生命現象の理解，特に日常の関心である“生まれる，育つ，老いる，考える”，などの理解につながるはずのものである．しかし現実には，そこをつなげる方法論はない．

そこで，本来生命科学が持っていた，数理で「究める」のではなく論理で「語る」という特徴を活かした方法論の確立が可能ではないか．物理科学を基盤にした従来の科学とは異なる，生命科学独自の知の組み立てへ向けての試みの提案である．

情報科学と生命科学の結びつき

21世紀の生命科学は，ヒトゲノムの全塩基配列解読プロジェクトの終了で始まった．つまり今は，生命科学研究の歴史の中で，エポック・メーキングと呼ぶにふさわしいこのプロジェクトがどのような経緯で何を求めて始まったのかを踏まえ，次の研究の方向を探る議論がなされるべきときにあるということだが，それがなされていない．今後の生命科学にとって，情報という概念，情報科学の方法論などを学び，そこから新しい分野を生み出すことが重要であると思うので，ここで筆者なりに生命科学の歴史と今後とを提示し，情報科学からの議論への参加を求めたい．

ヒトゲノムプロジェクト[1)]

地球上に暮らす多様な生物の観察・分類をする博物学を主体として進展してきた生物学は，19世紀に入って，進化論，遺伝の法則，生化学，細胞説の登場により，すべての生物に普遍的な生命現象の理解を求める学問へと移行した．しかも，20世紀前半には物理学が，“生命とは何か”という問いに関心を持ち始め，生物学が物理学を基盤にした科学の仲間入りをし，分子生物学，さらには生命科学が登場した．この間の歴史は重要だが，ここでは省略し，DNAの二重らせん構造の発見（1953年）に始まるDNAを基本に置いた研究の進展に話を絞る．

多種多様な生物という見方を脇に置き，普遍的生命現象の解明を目指した研究の歴史は，二重らせん構造の発見者の1人，J.D.Watsonの行動に集約されると言ってよい．その象徴として，彼が著した2つの教科書“遺伝子の分子生物学”，“細胞の分子生物学”がある．そして遺伝子と細胞を結ぶものとしてゲノムを位置づけるというところまできたのが生物学の現状であり，彼はこのプロジェクトの推進役をつとめた．細胞は，生物の“構造と機能の単位”であり，その中にあるDNAの全体がゲノムである．ゲノム内にはその細胞，さらにはそれが構

成する個体に必要な遺伝子がすべて含まれていることになり，「細胞内でゲノムがどのように働いているか」を知ることが，生命現象の理解の基本であることが分かったのである．

ここで，研究の歴史として指摘しておくべきことが2つある．1つは，DNAがまず遺伝子として捉えられたことである．DNAは本来ゲノム全体として生命現象のすべてを支えているのに，遺伝子に還元して生命を理解するという考えが広まった．ゲノム解析が終わった今，遺伝子ではなくゲノムが生命を考える単位であるとして，これを理解するところへ視点を移す必要がある．もう1つは，1970年代，米国でDNA研究が医学，特にがん研究と結びつけられたことである．死亡率第1位のがんの原因を知り，予防，診断，治療を進めようというプロジェクトが始まり，その結果，“がん遺伝子”が発見された．ただし，がんという病気の特徴から見て明らかなように，がん遺伝子は1個ではない．今や100個ほどのがん遺伝子が同定されている．そこで，ヒトが持つすべての遺伝子を調べ，細胞増殖とは何かという基本的な問いに向き合わなければがんの理解はできないという認識が生まれた．

ゲノムの解析の必要性は主としてがん研究の側から指摘されたのである．つまり，ゲノム研究は，病気の遺伝子探しとして始まった．科学であれば，できるだけ簡単なモデルを選択するのが通常だが，ここで最も複雑な生物であるヒトを対象にしたのは，がんを知ることが目的だったからである．

こうして1980年代後半，当時としては無謀ともいえる32億塩基が並ぶヒトゲノムの配列解析という方向が出された．1990年代，国際協力と競争の同時進行というかたちで進められたヒトゲノムプロジェクトは，解析技術の開発やベンチャー企業の参加という刺激もあって2003年，DNAの二重らせん構造の発見からちょうど50年目という年に終了した．32億塩基の配列の中には約30,000の遺伝子があること，配列の約50%は，単なるくり返し配列であり機能として遺伝子とは考えにくいことなど，ゲノムの基本構造が見えてきた．さてここで次は何をすべきかという問いが生まれる．

ヒトゲノムプロジェクトを出発点として何をするか[2)]

ゲノムプロジェクトは，もちろんヒトゲノムを対象に始められたものだが，これまで述べてきた経緯からも分かるように，ここには2つの流れが入り込んでいる．1つは，DNA研究を中心にして進めてきた生物学（分子生物学），もう1つは医学・医療の基礎研究である．前者は，普遍的生命現象を知ることが目的なので，すでに，100種ほどのバクテリア，酵母，線虫，ショウジョウバエ，シロイヌナズナなど，生物研究のモデル生物を含む多くのゲノムが解析され，研究は次の段階へ進みつつある．恐らくこのあたりは，専門外の人にはあまり知られていないところだろう．しかし，ここで蓄積されたデータもかなり大量になっており，これらを「細胞内でゲノムがどのように働いているか」という問いへの答にまとめる努力をする必要がある．

一方，医学・医療の基本研究としてのヒトゲノム研究は，国の科学技術政策の主要課題となり，独自の研究機関が設立されて，大型プロジェクトが動いているので，恐らくゲノム研究といえば，これを指すと考えている方が多いだろう．ここでは，遺伝子解析について膨大な量のデータが集積しており，病気の遺伝子探しを目指した米国を主体とする特許取得競争の中で，データの取得は加速している．そこで，網羅的な（ゲノムになぞらえてトランスクリプトーム，プロテオーム，フィジオロームなどというあまり意味のない造語を用いた）生体分子の分析を理解への道とする考えをとっているが，ここには疑問がある．

金子邦彦[3)]は，科学史上網羅的記述の後に理解が現れた例はないという．物理学では，原子や分子のすべてを知ったから熱力学が誕生したのではなく，統計力学も量子力学も熱力学から始まっている．そしていまだに原子や分子から熱力学は導かれてはいないというのである．金子は，理解には細部を捨てて薄目で全体を見る必要があるとも言っているM.Polanyi[4)]も暗黙知という，表現のできない知の存在を指摘し，違う階層の理解にはそれより下の階層の構成員の網羅は無力であると言っている．予算と機械があるので網羅的研究をするというのではなく，何を研究すべきかを考えるには，少なくともここでこれまでに得たデータの解釈に力を入れる必要がある．

生命とは何かという問いに向けて

生命とは何かという大きな問いへの道の1つとして，「細胞内でのゲノムの働き」という具体的課題にどう取り組むかがここでのテーマとなる．

■生命科学の特徴

前述したように生物学は，身辺に存在する多種多様な生物の観察・分類とその記述から始まった．しかし，研究の進展に伴い，現在では，分子の働きで生命現象を理解する研究が主力になり，科学としての性質が強くなった．科学といえば，基本は物理科学であり，生命科学も暗黙のうちに，物理科学を手本にしてきた．それは数理の世界であり，法則に基づいた反証可能性を強く主張できる．無矛盾性がその基本にあり，そのため，実験を伴わない理論物理学が重要な分野として存在する．

生命現象は具体的には化学反応であり，すべてが物理法則に従っているが，法則が働くのは局所的である．検証は経験的な世界で行われることが多く，大量で多様な知識が生まれ，数式や法則による体系化は難しく，言語によってそれを行う以外にない．法則が局所的であるために，生命体全体としては矛盾も見られるが，それはでたらめを意味するものではなく，論理があるのはもちろんである．このようにして見ると，物理科学が数理で“究める”という性質を持つのに対し，生命科学は論理で“語る”ものであるという特徴が見えてくる．

■語る科学としての生命科学

生命科学の現場で，発生，免疫，進化がん，記憶などさまざまな現象をテーマとして実験をしている．DNA解析を中心にした実験が主となるが，いずれにしても得られたデータを単に羅列しても生命現象の理解にはつながらない．研究者はこれまでに分かっている事実をもとに，データを，時間的空間的な制約や文脈の中に位置づけ，生物学的意味を引き出すのである．つまり，データが，常に研究者自身による，もう少し広く言うならそのときの研究者コミュニティによる解釈を通して意味づけされ，結果は“言語”で表現される．ここでいう言語は，数式と対比されるものであり，実際には“図像”での表現も重要である．生命科学の教科書を見れば，すべての現象が言葉と図像で表現されていることは一目瞭然である．

このような生命科学の本質が，これまで注目されずにきたのは，データの解釈と言語化を，研究者個人が行ってきたからである．ところが，ゲノム科学の登場により，そのような文脈，解釈とは無関係の大量のデータが日々産出されることになり事情は変わった．このようにして生産される多様かつ膨大なデータを扱うには，コンピュータの力を借りるほかない．もちろん，そこに，生物学，医学的意味を与えて整理する必要があり，大量データの処理と同時にデータの解釈と言語化が求められる．こうして生命現象の理解のためには，「語りの科学」と呼ぶべき新しい概念と方法が重要であることが分かった．これは，生物，情報科学，言語学などさまざまな分野の概念や手法を必要とするものである．

語りの科学とは

「語り」といっても，これはあくまでも科学であり単なる物語りではない．つまりここで作り上げる体系，最終的には生命体（物理的世界）とつながるものでなければならない．したがってここで行うべきことは，

①遺伝子やタンパク質などの生体分子を見出しとし，その機能を整理した辞書づくり，
②①で作成した辞書を用いて大量で多様なデータを具体的な生命現象と関連づけた整理（知識化），
③整理したものを生命体の理解につなげる，さらなる体系化（知識の体系化），

という作業である．

生命科学研究のデータはまずはデータベースに入り，一方で論文，総説，教科書などに「知識」として整理される．上記の作業に用いるのは，これらの素材であり，それを効率的に行うため，バイオNLPと呼ばれる生命科学の情報検索，自然言語処理，知識処理を対象とした情報技術が重要である．世界的には，この技術への関心は高まっているが，我が国ではまだその重要性への認識不足を実感するので，ここで情報処理の専門家に呼びかける次第である．

ここまで述べてきたことは，言語を用いた科学的知識の体系化であり，世界的にもこのような動きがあるが，ここで，私どもによる新しい試みをつけ加えた．これまでにも生命科学の表現としての言語に，図像を含むと書いてきたが，生物は形に多くの情報があるので教科書などは各ページに図や写真がある．そこで，生体分子の辞書を，生物個体や臓器のはたらきと結びつけて図示する可能性を検討した．時には動きを入れるなどの方法で，データや知識を体系化すると同時に，学問の新しい方向を探れるのではないかと考えている．実際には，脳に関して，その機能，発生，進化などに関する遺伝子やタンパク質の働きを示すという試みをしている．まず，データベース内の遺伝子が具体的にどのように働き，たとえば「記憶する」という日常関心を持つ生命現象とどうつながっていくかを語り，その内容を図と結びつけ，分かりやすく，具体的に示す試みである（**図-1**）．まだ海の

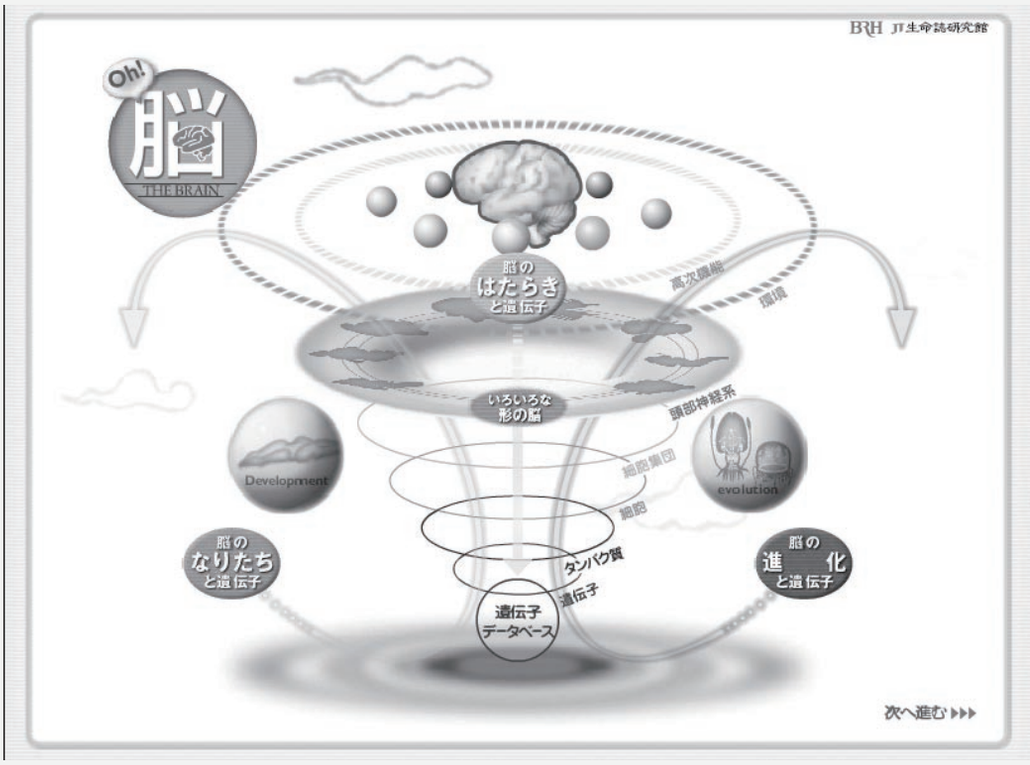

図 -1　脳の遺伝子データベースを生命現象の理解につなげるためのプラットフォーム

脳の機能・発生・進化にかかわる研究成果を遺伝子データベースと結んだ形で知識として整理し，そこから脳における生命現象についての物語りを構成していく試み．専門家にとっては知識整理になり，市民とのコミュニケーションにも使える．また関心のある人は誰もが活用できるものにもなっている．

ものとも山のものとも分からない状態だが，生物学の特性を活かした表現になることを狙っている．

語りの科学の持つ意味の再確認

■実用的側面

ゲノム科学が成立して以来，ゲノム配列はもちろん，遺伝子発現，分子間相互作用など大量に生産されている実験データを効率よく扱い，生物学・医学の知識にしないとデータが無意味に蓄積されるだけのことになり，研究の進展につながらない．情報技術の開発が重要な所以である．

■生命研究の本質にかかわる側面

生命現象は複雑であり，研究が急速に進展したとはいえ，まだその本質のつかみ方さえ分からないというのが現状である．このような複雑さの陰には，冗長性や曖昧性などがあり，なかでも生命体に特徴的なのは階層性である．これらを物理科学のシステムとして捉える試みも重要だが，これらを認識するのは人間（研究者）であり，表現された言語の構造としての認識の構造を解析することが興味深い．テキストからの生命現象の解析は，物理科学を基盤にした生命科学研究とは異なる形で本質に迫り，新しい視点を示す可能性がある．

語りの科学からの展開

テキストを読みとるということで前章までは言語（画像）という表現をしてきたが，これはもう少し広く情報という意味を持たせてもよい．語りの科学にかかわる言語（情報）には，3種類ある．

1つは，生命情報．ゲノム配列がどのような構成で働

	生命科学 （物理科学）	生命誌
DNA	遺伝子	ゲノム（生命子）
方法	分析 還元 数理 （無矛盾性）	分節 統合 論理 （矛盾許容）
理解の方法	構造・機能	関係・変化（進化）
生命体の捉え方	機械	生命体（時間）
科学として	究める	語る

表 -1　生命科学 と 生命誌の比較

生命科学は生命体を機械と見なし，その構造と機能を解明すればすべて理解できると考えている．生命体を機械論と物理科学の中に置いているのである．DNA も遺伝子を単位と考える．生命誌（Biohistory）は，生命体は構造と機能に加えて歴史と関係を見なければ理解できないと考える．具体的には DNA はゲノムを単位とする（これを生命子と名づけ，生命体が存在するための単位とする）．もちろん生命誌にとって生命科学の知識は不可欠であるがそれを基盤に生命体を生命体として捉える知を作っていく．これが語る科学となるのである．

いているか，より言語に近い表現をするならどのような文法で書かれているのかというゲノム言語の解読はゲノム研究そのものである．たとえば，遺伝子の中でもタンパク質指令領域と発現調節領域☆1を規定している文法を明らかにすることができたら面白い．

第2は人間の言語であり，社会情報である．まさにこれは人間による理解であり，その内容の人間の間での相互伝達，相互理解である．これをコンピュータがいかに助けるか興味深い．

そして第3，機械言語である．理解の相互伝達，それによるさらなる理解の展開に役立つために，生命現象や生体機能を形式的記述に変換し，大量の情報を基にした理解を助けたり補完したりすることが求められる．

これまでのバイオインフォマティクス研究は，データの解析に主眼が置かれ，その解釈は研究者が論文を読んできたが，論文数が急速に増加し，研究の全体像をつかむことは不可能に近くなってきた．そこで，知識のデータベース化の動きが世界的に起こりバイオ NLP（Natural Language Processing）と呼ばれたり，Bibliomics（文献に書かれた知識のすべてを扱う学問）という言葉が提案されたりしている．しかし，これらの動きは，今回ここで紹介したような“語る科学”という概念には至っていない．“語る科学”の中には，単なるテキスト処理を超えた新しい考え方があり，しかもそれは生命科学の次の展開に重要な役割を果たすことが期待されるものである．Conceptual Biology（概念生物学）という用語も考えられるが，当面は“語りの科学”として提案する．筆者は，ゲノムに注目して生命体を見たときには，進化の歴史が重要になることから，生命科学でなく生命誌（Biohistory）の方が生命理解の知としては適切であると考えており，それを語る科学の具体的な姿として提唱している（**表-1**）．

“語りの科学”はまだ生まれたばかりで，どう展開するか見えない部分もあるので，“騙りの科学”ではないかと眉に唾をつける方もあるかもしれない．しかし，テキストの処理によって膨大な知識を共有できる形にする必要性は誰もが感じていることであり，そこから生命現象の何かが見えてくるかもしれないという予測は，それほど見当違いではないはずである．情報の専門家から面白いアイディアや，具体的な方法の提案，さらには実際に仕事をしてみようという申し出が出てくることを期待して筆をおく．

参考文献
1）榊　佳之: ヒトゲノム ― 解読から応用・人間理解へ，岩波新書（2001）．
2）松原謙一: 遺伝子とゲノム― 何が見えてくるか，岩波新書（2002）．
3）金子邦彦: 生命とは何か―複雑系生命論序説，東大出版会（2004）．
4）Polanyi, M.: 暗黙知の次元―言語から非言語へ，紀伊国屋書店（1980）．

（平成16年12月29日受付）

☆1　ゲノム DNA の中，遺伝子として働いている中にも，実際にタンパク質のアミノ酸配列を指令する部分とそれがいつ，どこで，どれだけ合成されるかを調節する部分がある．この調節が生きものらしさを産み出している．