バイオインフォマティクスが追い求めるもの

由　良　　敬

Ⅰ．はじめに

バイオインフォマティクスという言葉は，比較的最近作られた合成語であるにもかかわらず，聞いたことがない方はほとんどいないほど，広く普及している。バイオインフォマティクスは，1995年の*Haemophilus influenzae*全ゲノム配列決定[1]のころから繰り返し報道されることによって，現在の知名度を得たと考えられる。しかし，改めてバイオインフォマティクスとは何かと問われたときに，その回答を明確に出すのは難しい。専門家とよばれる人々の間でも，バイオインフォマティクスといったときに，指し示している実態が異なっていることが普通である。二人の研究者が，それぞれバイオインフォマティクスを専門にしていると証言しても，両者の間で専門的な会話が成り立たないことすらありうる。このことは，バイオインフォマティクスの分野が非常に広大であることを意味する。

バイオインフォマティクスという学問を定義するのは難しい。しかしあえて以下に，ある定義を抜粋する（日本語訳は筆者による）。「バイオインフォマティクスとは，物理化学的な研究対象である分子に対して，計算機科学や数学によって構築された情報処理技術を適用することで，生体を構成する莫大な数の分子がもつ情報を再構成し理解することを目的とした，概念構築の生物学である。（中略）バイオインフォマティクスには３つの目的が存在する。第１は，研究者に対して，既存情報を検索し新規情報をデータベースに加えることができる環境を設定すること。（中略）第２に，情報解析用の手段と資源を開発すること。（中略）そして第３に，これらの手段を用いてデータを解析し，生物学的に意味のある情報を抽出すること。生物界に普遍的に存在する法則の発見と特殊現象の発見がバイオインフォマティクスによって初めて可能になる。」[2]翻訳文には，筆者の気持ちが入っていることを付記する。

バイオインフォマティクスは情報学であるということをよく聞く。そういう発言を耳にするたびに，筆者は唖然とする。バイオインフォマティクスは生物学である。筆者の周りには，どういうわけだか，思ったことを明快に表現する方が多く，筆者自身はいつもそういう方々に助けてもらっている。あるときバイオインフォマティクスが，どれほど生物学であるかを説明してもらったので，ここに紹介する。「理論物理学において，高度な数学は重要な手段である。しかしこのことを根拠として理論物理学が数学であるという人はいない。現象を説明し予測するという物理学に理論物理学の本質があるからである。バイオインフォマティクスにおいて，情報処理は重要不可欠な技術である。だからといってバイオインフォマティクスは情報学ではない。生命現象を説明し予測するという生物学にバイオインフォマティクスの本質がある。バイオインフォマティク

スは生物学である。」

本解説では，バイオインフォマティクスをデータベース，ソフトウェア，そして生物学の概念構築の順番で概観する。

Ⅱ．データベースの開発

筆者はデータベース開発技術のことに関してはまったくの素人である。どのようにしてデータベースを構築すべきかという技術に関しては，ここで記述することができない。データベース構築方法は情報学の分野で深く研究されていると聞いている。

バイオインフォマティクスにおけるデータベースは，現在莫大な種類が存在する。それらのデータベースはほとんどの場合に，インターネットを介して世界に公開されている[3]。インターネットに公開されているデータベースは大きく分けて３つに分類することができる。第１は，測定結果そのもののデータベースである。この中に含まれるデータベースとして，ゲノム配列決定プロジェクトにより判明した各生物種の塩基配列，タンパク質のアミノ酸配列，タンパク質とDNA／RNAなどの生体高分子立体構造座標，いろいろな細胞におけるmRNAの発現量測定結果，生体中における核酸及びタンパク質の修飾部位，各生物種のいろいろな環境下における表現型の写真，塩基置換と病気との関連などがある。第２は，これらの測定結果から得られる情報を整理したデータベースである。例えば以下のようなデータベースが存在する。ゲノム配列は４種類の核酸が長く連なった高分子であり，データベースの中には４文字の羅列として表現されている。この中のどの部分にタンパク質がコードされているかを示すことで，ゲノム配列からある種の情報を抽出したことになり，その情報が新たなデータベースとして公開されている。あるいはつぎのような情報もデータベースとして存在する。タンパク質のアミノ酸配列にはお互いに類縁関係がある配列と，ない配列が存在するため，配列の類縁性で分類をすることが可能である。分類結果はタンパク質ファミリーとして公開されている。生体高分子の立体構造にもまた類縁性が存在し，その分類結果もまた公開されている。分類は何に注目して分類するかによってその結果が異なってくる。よってこれらの分類データベースは複数種類存在する。第３は，第１と第１，第１と第２，または第２と第２のデータベースを関連づけることでうまれる新たな情報のデータベースである。mRNAの発現量と生物の表現型との関係を調べたデータベース，DNA上近くにコードされているタンパク質には類似のタンパク質があるかを調べたデータベース，タンパク質の配列類縁性と立体構造類縁性の関係のデータベース，これらのデータと文献とを関連づけたデータベースなどである。

すべてのデータベースには，２つの重要な側面が存在する。第１は何のデータをデータベースとするのか。そして第２はどのようにしてデータベース化するかである。これらは自転車の両輪であってどちらも重要である。しかしバイオインフォマティクスにとってはどちらがより重要であるかと問われれば，第１番目であると答える。バイオインフォマティクスは生命現象の普遍性と特異性を発見することにその学問の意義があるのだから，どのような現象をデータベース化するかが重要であるのは自ずとあきらかであろう。データベースの構造そのものは，もっと一般性

の高い情報学に位置づけることができる。しかし，生物学固有のデータ構造が存在するのもまた事実である。生物学においては，各データの関連が非常に複雑になっており，データの関連をつけるだけでも莫大な人材と計算機能力を投入する必要がある[4]。さらに問題になるのは，そのようなデータをどのようにして人に提示するかである。データを集める目的は，生命現象を研究者や学生に見せ，解釈し，理解することにあり，集めることそのものが目的ではない。データをどのように視覚化するかもまた，バイオインフォマティクスにおける大きなテーマである[5]。

Ⅲ．ソフトウェアの開発

生命情報解析用プログラムの開発には歴史がある。1950年代半ば頃から，タンパク質のアミノ酸配列を直接決定することが可能になった。このことに伴ない，アミノ酸配列を比較する研究が始まっている[6]。1953年にはX線散乱像からDNAの立体構造が推定されている[7]。タンパク質の立体構造は1958年にX線結晶解析により初めて発表されている[8]。ここにバイオインフォマティクスの始まりがある。これらの研究では，得られたデータから重要な情報をアルゴリズムにしたがって抽出している。当時はコンピュータがまだ普及していなかったので，現在普通に開発されているようなプログラムは存在しなかった。しかし，1950年代のこれらの研究は，本質的な部分で現在のバイオインフォマティクスによる解析と何らかわりはない。

バイオインフォマティクスが始まってから50年たった現在，ソフトウェアの解析では，「大量データ解析」，「高速解析」，「統合解析」が重要な意味を持つようになってきた。現在開発されるソフトウェア[9]は３点の少なくとも２点は満たしているように見える。

ゲノム配列の決定により，我々は膨大な量のDNA配列を手に入れることとなった。日本におけるゲノム情報のデータベースであるDDBJ[10]に保存されている核酸配列の量は現在約290億塩基である。わずか10年前の1993年には１億2000万塩基であり，その５年前の1988年には約20万塩基しか存在しなかった。データ量が幾何級数的に増えていることになる。この大量データの中から情報を抽出することができなければならない。大量のデータに対応するためにはソフトウェアのみならずハードウェアにおいても発展しなければならないことがたくさんある。例えば，複数のゲノム配列を比較して類似部分配列を図示するためには，全ゲノム配列データを，コンピュータメモリ上に読み込むことが一番理想的である。比較的簡単に手に入るパソコンクラスターでは，現在既知の全ゲノム配列をすべてメモリ上に読み込み，比較しN次元行列に格納することができない。CPUあたりのメモリ量に上限が存在するためである。ソフトウェアでの工夫でこの問題を回避すると，解析速度を犠牲にすることになる。

ヒトを含む多くの生物種のゲノム配列の決定に続き，ゲノムにコードされているタンパク質の立体構造決定が大規模に行われようとしている[11]。タンパク質は生体で機能する際には，ほとんどの場合固有の立体構造を形成する。タンパク質の立体構造を知ることによって，そのタンパク質がどのような機能を果たしているかを知るヒントを得ることができる。しかし，アミノ酸配列のみから理論的にタンパク質の立体構造を求めることは，いまでも多くの研究者によって研究されている未到達の領域である。そこで，代表的なタンパク質の立体構造を実験的に決定するプロジ

ェクトが動き始めている。このプロジェクトによって産出される大量データから，バイオインフォマティクスに期待されている解析の一部として，代表タンパク質を鋳型として全タンパク質の立体構造を推定すること，生体中で存在するタンパク質の超複合体を再構成すること，タンパク質の動き（構造変化）を解明することがあげられる。全タンパク質の立体構造推定には，大量データの高速計算が必要であり，このことに対応するソフトウェアとハードウェアの開発がなされている[12]。生体中に存在するタンパク質の超複合体の再構成問題においては，電子顕微鏡で得られる超分子の像からどのようにして情報を抽出し，X線結晶解析などでわかっている各部品タンパク質の構造と整合性をとりながら再構成していく方法論が問題となっている。高速なデータ読みとり手法と，当てはめ手法の開発が始まっている[13]。タンパク質の動きの解明にも，多くの研究者が手法の開発に力を注いでいる。多粒子系であるタンパク質の生物学的に意味がある動きを見るには，高速計算が不可欠である。いくらすばらしい方法であっても，無限時間かかる方法では意味をなさない。高速化にはソフトウェアでの工夫と次世代のCPUが常に必須である。その好例として構造が判明している超分子の動きを超大型コンピュータによってシミュレーションするソフトウェアのひとつ，NAMDがある。NAMDはその性能のよさゆえに地球シミュレータと並んで，2002年ゴードンベル賞を得ている[14]。タンパク質内部でおこっている化学反応をシミュレーションするソフトウェアの開発もまた進行している。電子がどのように伝達されるのか，光を受けたタンパク質がどのように構造変化をするのかなどは，シミュレーションによって十分あきらかにできることである[15]。これらのシミュレーションには超精度の高速解析が必要である。

生物種のゲノム配列が決定される以前から，遺伝子の配列決定及び比較の解析は行われてきた。その研究によって，いろいろなソフトウェアが蓄積されている（例えば文献16参照）。現在広く行われている解析の大部分は，それらのソフトウェアを大量データに適用すること，および複数のソフトウェアによって複合的な解析を行うことである。大量データに対して既存のソフトウェアを複合的に用いて解析することは，思うほど容易なことではない。どこにデータが存在するのか，どのソフトウェアが解析の目的を達成するのか，どのソフトウェアに対してはどのデータフォーマットを用いなければならないのか，どのソフトウェアはどれだけの計算時間と中間ファイルを必要とするのか，など考えなければならないことが多い。このような解析においては，知りたいことは何かを明確にすることが一番大切であるにもかかわらず，どのように実行するかが重要な問題のようになってしまうことが多い。その結果，本来の問題点を見失ってしまうことすらある。そこで筆者らは，バイオインフォマティクスにおける既存のソフトウェアが一望でき，データフォーマットとコンピュータ資源を気にすることなく目的を達成できるシステムの開発をすすめている。タンパク質高度解析システムBAAQ（Bioinformatics: Ask Any Questions。バークと読む）と命名された本システムは，データベースとソフトウェアをアイコンで示し，アイコンを線で結ぶだけで，データ解析が可能となるシステムである（図１）。コンピュータ資源は現在進められているグリッドコンピューティングを利用することで，ユーザが気にする必要はなくなるようにする。BAAQシステムは，バイオインフォマティクスの専門家のみならず，教育の現場やコンピュータを専門としない生物科学研究者にも利用してもらえるようになることを目指している[17]。

図１　BAAQシステムの概観　ゲノム配列を利用して進化系統樹を描こうとしているところ。データもソフトウェアもすべてがアイコンになっている。

Ⅳ．記述の生物学から予測の生物学へ

多くのデータから抽出すべきであると考えられている生命情報には，ゲノム配列から推定されるタンパク質の機能，ゲノムのタンパク質コード領域，アミノ酸配列情報のみから取得されうるタンパク質立体構造情報，タンパク質の動的性質，タンパク質の細胞内外局在部位，タンパク質間およびタンパク質とDNA／RNAの相互作用，RNA遺伝子の発見などがある。筆者は，ここ数年のあいだ，アミノ酸配列及び立体構造情報から，タンパク質の機能を推定する方法の開発とその適用を行ってきているので，ここではそのことについて概説する。

タンパク質が機能するためには，立体構造を形成しなければならない。例えば，化学反応を触媒する酵素の場合は，触媒となる原子が的確な空間に配置されることが大切である。非常に大きなタンパク質であっても，触媒反応に関与する部分はその一部であり，他の部分は反応を触媒する原子を的確な位置に配置するために存在しているようにも見える。タンパク質の立体構造が類似の場合には，それらのタンパク質の機能も類似の場合がある。タンパク質の立体構造を知れば，そのタンパク質の機能を推定できる可能性があるならば，機能未知のタンパク質の立体構造を決定して，その情報から機能を推定しようという論理が成立する。このような論理構成は，いままでの生物学には存在しなかった。この論理展開によって，機能が推定されたタンパク質の例を挙げよう。

*Methanococcus jannaschii*は嫌気性の細菌であり，深海の熱水吹き出し口に生育する。この生物の全ゲノムが決定され，そのゲノムに存在する遺伝子のひとつであるMJ1247は，どのような機能のタンパク質がコードされているのかが不明であった。KimらはMJ1247産物の立体構造をＸ線結晶解析により決定し，その立体構造が大腸菌のグルコサミン６リン酸合成酵素と酷似していることを見いだした（図２）。立体構造の比較から，機能不明であったタンパク質の活性部位が同定され，また他の証拠とともにして，MJ1247にコードされているタンパク質が３ヘキスロース６リン酸イソメラーゼであることがわかった[18]。

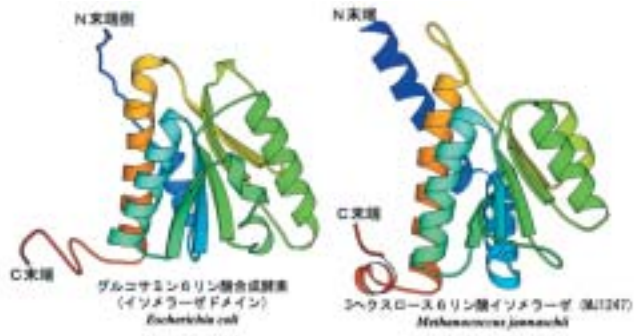

図２　機能不明のタンパク質（右）の立体構造を決定したところ，すでに立体構造が判明しているタンパク質と類似の立体構造であった。タンパク質は方向（Ｎ末端からＣ末端）をもつ長いポリペプチド鎖であり，図では鎖のトレースを抽象的にあらわしている。βストランド（中央に存在する矢印状に抽象化された構造）とαヘリックス（外側にある螺旋状に抽象化された構造）の鎖に沿った順番と空間配置を見比べてほしい。立体構造の類似性から両タンパク質は類似の機能を持つと推定され，実験的に確かめられた。

タンパク質の全体構造ではなく，部分構造のみに類似性が見いだされた場合でも，機能に類似性がみられる場合がある。郷通子らはモジュールとよばれるタンパク質の部分構造に注目して機能同定に取り組んでいる。例えば，異なるタンパク質において，DNAのリン酸基に結合する機能をもつ類似立体構造のモジュールが見つかっている[19]。モジュール立体構造の情報をアミノ酸配列に還元して，ゲノム配列から予測されたアミノ酸配列を調べたところ，このモジュールが，機能未知タンパク質にも見いだされた。このことから，そのタンパク質がDNAに結合する機能をもつこと，及び他の証拠から，そのタンパク質が環境に存在する遊離DNAをゲノムの中に取り込む機能をもつことを予測し[20]，実験によりその機能を確かめることに成功している[21]。

バイオインフォマティクスにおいて，筆者が一番重要と考えているのは，生物学的に意味のある情報を抽出することであり，生物界に普遍的に存在する法則の発見と特殊現象の発見である。このことをとおして生命とは何かを理解することがバイオインフォマティクスの使命だと思う。データベースの構築もソフトウェアの開発も，これらの目的を達成する手段である。そして，普遍性を理解することができれば，予測が可能となるはずである。現在のバイオインフォマティクスが目指している具体的な目標は，データからの普遍性抽出と予測，そして実証実験である。

Ⅴ．これから

生物における普遍的法則は，進化の中に存在すると筆者は思っている。DNA二重螺旋構造を見いだしたクリックは，自著の中でこんなことをいっている（翻訳は筆者による）。「自然選択の存在が，生物学を他の科学と異なるものにしている。生物学は物理とは大きく異なる学問である。物理学の基本法則は，数学によって厳密に表現することができ，その法則は宇宙のどこにおいてもたぶん成立するであろう。これに反して，生物学の法則は，一般則であろう。この法則がしるすものは，何十億年もの間に自然選択によって選ばれた精巧な化学反応である[22]。」現存の生物がもつ情報と，その情報の産物を比較することで，生物がどのように進化し，これからどうなっていくかを推定することができるであろう。これには生物の持つ情報のみならず，地球に刻み込まれた情報をも統合する必要がある[23]。

バイオインフォマティクスにおいて，解くべき問題は山のように存在する。そのため，この分野は常に人材不足で苦しんでいる。この状況は日本だけのことでなく，例えば北米でも同じであることが報告されている[24]。この状況を生んでいる原因の一つには，本分野の多彩性があると考えられる。生物学を理解した上で，数学，物理学，化学，情報学などのいくつかの分野にたけている必要がある。残念ながら現在の大学教育は，このような素養をもつ研究者を生み出すようにはなっていない。これから爆発的に増大する生命情報の解析のために，大学教育に大きな変革が起こり，バイオインフォマティクスを力強く進める人材が登場することを切望する。

参考文献

（1）Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512 (1995)

（2）Luscombe, N. M., Greenbaum, D. & Gerstein M. What is bioinformatics? An introduction and overview. *Yearbook of Medical Informatics* 83-100 (2001)

（3）http://us.expasy.org/alinks.htmlやhttp://www.rcsb.org/pdb/links.htmlを参照されたい

（4）Chicurel M. Bioinformatics: Bringing it all together. *Nature* **419**, 751-757 (2002)

（5）Campbell, A.M. & Heyer, L. J. Discovering genomics, proteomics, & bioinformatics. Cold Spring Harbor Press (2003)

（6）Harris, J.I., Sanger, F. & Naughton, M.A. Species differences in insulin. *Arch. Biochem. Biophys.* **65**, 427-438 (1956)

（7）Watson, J.D. & Crick, F.H.C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953)

（8）Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, H., Wyckoff, H. & Phillips, D.C. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, 662-666 (1958)

（9）菅原秀明．編集 あなたにも役立つバイオインフォマティクス．共立出版（2002）

（10）Miyazaki, S., Sugawara, H., Gojobori, T. & Tateno, Y. DNA Data Bank of Japan (DDBJ) in XML. *Nucl. Acids. Res.* **31** 13-16 (2003)

（11）Burley, S.K. & Bonanno, J.B. Structuring the universe of proteins. *Annu. Rev. Genom. Hum. Genet.* **3** 243-262 (2002)

（12）Yamaguchi, A., Iwadate, M., Suzuki, E., Yura,, K. Kawakita, S. Umeyama, H. & Go, M. Enlarged FAMSBASE: protein 3D structure models of genome sequences for 41 species. *Nucleic Acids Res* **31** 463-468 (2003)

（13）Swedlow, J.R., Goldberg, I., Brauner, E. & Sorger, P.K. Informatics and Quantitative Analysis in Biological Imaging. *Science* **300**, 100-102 (2003)

（14）http://www.ks.uiuc.edu/Research/namd/

（15）Yamada, A., Kakitani, T., Yamamoto, S. & Yamato, T. A computational study on the stability of the protonated Schiff base of retinal in rhodopsin. *Chem. Phys. Lett.* **366**, 670-675 (2002)

（16）由良　敬，篠田和紀，井本剛史，郷　通子．DNA塩基配列を用いたfastDNAmlによる進化系統樹の推定．名古屋大学大型計算機センターニュース 28, 127-135 (1997)

（17）科学新聞「原研の成果」10面　平成15年３月21日

（18）Martinez-Cruz, L.A., Dreyer, M.K., Boisvert, D.C., Yokota, H., Martinez-Chantar, M.L., Kim, R. & Kim, S.H. Crystal structure of MJ1247 protein from *M. jannaschii* at 2.0 A resolution infers a molecular function of 3-hexulose-6-phosphate isomerase. *Structure* **10**, 195-204 (2002)

（19）Yura, K., Shionyu, M., Kawatani, K. & Go, M. Repetitive use of a phosphate-binding module in DNA polymerase beta, Oct-1 POU domain and phage repressors. *Cell Mol Life Sci* **55**, 472-486 (1999)

（20）由良敬，郷通子．モジュールに基づくゲノム機能予測．*蛋白質核酸酵素*47, 1090-1096 (2002)

（21）Yoshihara, S., Geng, X., Okamoto, S., Yura, K., Murata, T., Go, M., Ohmori, M. & Ikeuchi, M. Mutational analysis of genes involved in pilus structure, motility and transformation competency in the unicellular motile cyanobacterium *Synechocystis* sp. PCC 6803. *Plant Cell Physiol* **42**, 63-73 (2001)

（22）Crick, F. What Mad Pursuit. (1990)

（23）Benner, S.A., Caraco, M.D., THomson, J.M. & Gaucher E.A. Planetary Biology-Paleontological, geological, and molecular histories of life. Science 296, 864-868 (2002)

（24）Powell, K. Universities urged to get with IT for biology. *Nature* **419**, 102 (2002)

（ゆら　けい：日本原子力研究所 計算科学技術推進センター 量子生命情報解析グループ）
（http://wwwitblpg.apr.jaeri.go.jp/qbg/yura/index.html）