バイオインフォのよしなしごと

白　井　　剛

Ⅰ．はじめに

なんでもいいという話なので，どうでもいいことを書こうと思う。

たまたまある雑誌の書評欄を見ていたら，心温まる逸話が紹介されていたので，その引用から始めてみよう[1]。“I was once introduced to Sir Sydney Brenner as someone involved in ‘bioinformatics.’ This is not really how I would describe what I do, but I didn’t protest quickly enough: Dr. Brenner laughed heartily and said, “Bioinformatics? The last refuge of scoundrels.”

Brennerは線虫のアポトーシス（プログラム細胞死）の研究で，2000年ノーベル医学生理学賞を受賞した分子生物学者で，遺伝暗号の解読にも多大な功績があった。彼の言葉は実験生物学者のバイオインフォマティクスに対する典型的な見方といえる。つまり，実験屋が苦労して出したデータを使ってキーボードをカチャカチャやっている連中，くらいの意味である。これはどこの世界にもある異分野に対する色眼鏡であるが「ごもっともにございます，サー」と思われる節もなくはない。

バイオインフォマティクスは流行である。筆者もなにげなく「ご専門は？」と聞かれたら，無愛想に「バイオインフォ」と答えるかもしれないが，それはそう答えたほうが面倒くさくないからである。便利な言葉ではある。しかし，いまひとつしっくりこない気もする。何が問題なのか？

Ⅱ．実験系へのサービスとしてのバイオインフォ？

ひとつには，現在のバイオインフォマティクスが，濃厚に実験系生物学へのサービス業として存在していることがある。バイオインフォマティクスの２本柱は，データベース構築とデータ解析法（ソフトウェア）の開発だが，いずれも実験データの入力を得て，実験系へのフィードバックを主な目的としている。というか，研究者自身そう主張する場合が多い。研究領域としての自立性が低いと見なされているのである。

もしバイオインフォマティクスの定義を，生物情報の数値化や電算機処理をともなう研究とすると，全くその要素をもたない生物学はもはや考えにくい。一方，バイオインフォマティクスは，この言葉が発明される以前に存在した研究領域とは全く重複しないものだとすると，そんな領域は未だ存在しない（ちなみに，PubMedがカバーする範囲で最初にbioinformaticsという言葉が使用されるのは1993年）[2－4]。

バイオインフォマティクスがはやり始めるのは，ゲノムプロジェクトの成熟時期と大体一致する。ゲノムプロジェクトの結果，それまでも行われていた計算を大量迅速に，しかも複雑に行う必要が生じたことが，この造語と流行を生み出したのである。ようするに，家内制手工業でやってきた生物学が，好景気に乗って工場を建てたときに，パソコンを何台か買ってSEを雇ったのである。きっとどこかの営業が「この規模でやられるんなら，財務諸表は電子化しないと…勘定奉行つけときますから」とか言ったのだろう。

Ⅲ．なにも新しいことをいわないバイオインフォ？

もうひとつの問題は，バイオインフォマティクスからの逆照射が少ないと思われていることにある。「バイオインフォマティクスの連中は，結局こっちがもう知っていることしか言わない」といったことを言う実験屋さんは多い。

もし，ヒトゲノム30億塩基対を比較して類似した配列を探すような計算の結果を，あらかじめ知っているというのなら，それは単なる勘違いだが，これはそういうことを言っているのではないのだろう。例えば，遺伝子はDNAであるとか，蛋白質の構造は遺伝子に暗号化されているとか，大きな生物学のパラダイムをバイオインフォマティクスが提出できないということである。

これはある程度は事実であって，現状のバイオインフォマティクスは，実験でもできる（むしろ実験したほうが速くて確実）な事を，計算機上で再現するのに苦戦している。例えば，蛋白質の立体構造を理論的に予測するという古くからの課題がある。これは分子量数十万の分子のコンフォメーションを予測する問題なので，扱う変数は膨大である。これまでにさまざまな近似法を導入して試みられてきたが，部分的な成功はあっても，一般的に適用可能な方法論は確立していない。対して，構造が知りたければ実験して求めれば良い，というのが実験構造生物学の立場であり，組織的に大量に構造を決定するための構造ゲノミクス計画が，世界規模で実現している[5]。これに，実験装置の改良，ロボットや量産型ポスドクの大量投入もあって，現状「生産性」においてはバイオインフォマティクスの貢献は少ない。

Ⅳ．昔の話

いまこの時点のバイオインフォマティクスは，これまでやられてきた事をちょっと大きな規模でやっているにすぎない。それでは昔はどうだったのか（図１）？　生命情報の数値化ということで言えば，バイオインフォマティクスのルーツの一つは数理遺伝学にあると思う。遺伝学は文字どおり生物の遺伝現象を扱う学問で，生物学のなかで数学との親和性がもっとも高い。数理遺伝学の源流であるメンデルの３法則（分離・独立・優性の法則）に代表されるように，おおむね粒子的にふるまう遺伝子の性質は，数学モデルに載せやすかったのである。特に１次元のデジタル情報であるDNA配列が大量に使えるようになってからは，現在バイオインフォマティクスと呼ばれる領域のいくつかの萌芽が産み出されて来た。よって，バイオインフォマティクスにおける顕著な成功例も遺伝学の中に見つけることができる。例えば木村の中立説である[6]。

当時遺伝子の配列が各種生物から決定されて，異なる生物で同じ役割をもつ遺伝子の配列はよ

く似ていることがわかって来た。もちろん，異なる種の間で配列は全く同一ではなく，種の隔たりが大きいほど違いも大きかった。要するに遺伝子も進化するのだが，一つ謎があった。DNAは３文字のコドンで１つのアミノ酸を暗号化する。塩基の種類は４種類なのでコドンは4^3＝64通りある。対してアミノ酸は20種である。つまり遺伝子が変化しても，蛋白質のアミノ酸配列は変化しない場合があるのだが，種間で遺伝子配列を比較したとき，アミノ酸を変化させない変異の方がより多く見つかったのである。

なぜこれが問題なのか？　遺伝子配列の変異は進化の原因である。そして進化は適者生存の競争によって起こると考えられていた。遺伝子レベルでも，配列が変化するのは，それがより適応した遺伝子に変化したからだと考えたいところだが，アミノ酸が変化しなければ蛋白質の機能も変化しないので，適応は起こらないはずなのである。これでは観察結果と矛盾する。

これに対して，中立説は以下のように答えた。遺伝的浮動という現象があって，ある遺伝子の集団内での頻度は，例えその遺伝子が有利でなくても，たまたま少しだけ増加することができるし，結果として，ある確率で集団内で100%の頻度に達することができる。モデル計算から，この遺伝的浮動により有利でも不利でもない中立的な変異が，集団の大きさにかかわりなく一定速度で蓄積することが示せる。

観察の結果を正しく説明できるにも関わらず，中立説は当初猛烈な反発を受けた。発表当初は，“You are stupid.”と言われたという話もある。それだけその当時の常識（適者選択のダーウィニズム）を覆す理論であったとも言える。

あまり意識されていないが，中立説はバイオインフォマティクスの重要なツールの理論的背景になっている。異なる生物種の遺伝子配列は，その種の間の隔たりが大きいほど，より異なっている。したがって配列の違いを種分岐後の時間経過に換算することが可能で，これにより系統樹を描いて生物進化の過程を推定することができる。系統樹推定法は，生物の歴史を調べるためのもっとも定量的で客観的な道具である。

中立説を持ち出さないでも，この道具はつかえるのだが，配列の変化が確率的な浮動の産物であるというバックグラウンドがないと，結果が系統樹であることが保証されない。分子情報から系統樹を推定する場合は，暗に中立説が正しいと仮定しているのである。

中立説の提唱は1960年代後半だった。もちろん，木村は自身の仕事をバイオインフォマティクスとは呼ばなかったし，今更そのように呼ばれるのも迷惑だろう。しかし，生体分子の構造情報の解析から生物学のそれまでの常識を覆したという意味で，成功したバイオインフォマティクスだと思う。

Ⅴ．また昔の話

蛋白質の立体構造が初めて決定されたのは1957年のことで，Kendrewらによるマッコウクジラのミオグロビン（組織中の酸素運搬蛋白質）が第一号である[7]。後にPerutzによりヘモグロビン（血液中の酸素運搬蛋白質）の構造が決定されたとき[8]，彼らは２つの構造を比較して驚いた。アミノ酸配列では20%程度しか一致しない２つの蛋白質の立体構造が，非常によく似ていたのであ

る。蛋白質の立体構造は配列よりも保存性が高いという重要な経験則は，ここから得られた。

時は流れて1980年代後半，あることが流行った。一見関係のなさそうな蛋白質の立体構造を，比較して類似したものを探し出すことである。進化的に関係のある蛋白質は，アミノ酸配列は大きく隔たっていても立体構造は保存されていることは，すでに知られていた。しかし当時の構造解析屋さんたちは，自分の解いた構造を，他人の解いたあまり関係のなさそうな構造と比較してみる習慣を持っていなかった。蛋白質の意外な類似性を見つけだしてきたのは，多くの場合「蛋白質の立体構造や配列情報を計算機をつかって研究」していた人々だった[9]。

いまでは，まとめてバイオインフォマティクスに編入されているが，その頃はこの分野に適当な名称はなかったと思う。

ちょうどこの時期から，立体構造解析手法の改良によって，蛋白質の解析例が飛躍的に増加し始めた。関係のなさそうな蛋白質の構造がお互いに似ていることは珍しくなくなり，次第にコンセンサスが形成されていく。蛋白質の立体構造は，進化的に関係があるなしに関わらず，そんなに種類がないということである。

1992年にChothiaは短いコメントを発表した[10]。年々加速度的に増加する配列や立体構造の情報を比較すると，新しく決定された蛋白質が，すでに知られているものと類似している割合がわかる。その割合から全体のサイズを推定することができるが，それによると蛋白質の構造は，せいぜい千数百種類しかないというのである。

この「蛋白質構造1000個説」はちょっとしたブームを起こした。ちょうどヒトゲノム計画が本格的になったころであり，生物の網羅的な解析へ向かって生物学全体がシフトし始めた時期である。全体は恐れるほど大きくはない，とする説は魅力的だった。

蛋白質構造の推定個数はその後増加して，いまでは最大10,000くらいと考えられているのだが，10,000ならば現在の技術で計画的に資源を投入すれば，数年で網羅的に構造決定できる，というのが構造ゲノミクスのスローガンである。Chothiaの説がこの雰囲気を後押しして，現在のポストゲノムとしての構造ゲノミクスの，重要な精神的なバックグラウンドを果たしたことは確かである。帰結が生物学にとって幸福だったかどうかは置くとして，バイオインフォマティクスがパラダイムを作り出した例といえるだろう。

Ⅵ．いまとこの先の話

ここで，比較的最近のバイオインフォマティクスの２つの分野，細胞シミュレーションとオントロジーについて紹介しようと思う。理由は，これらがいまの生物学におけるバイオインフォマティクスの位置づけを代表していると思うからである。つまり，実験系生物学へのサービスとしての装いを持っていて，しかも実験屋さんが誉めているのを見たことがない，という意味である。

１．細胞シミュレーション

細胞シミュレーションとは，細胞内の代謝過程を計算機上で再現する試みである。もともと生化学などにはあった分野であるが，現在はゲノム解析の結果を受けて，すべての反応を計算機上

で網羅的に扱うことを目指している。何千と連立した反応速度微分方程式を非解析的に解くシミュレーションである。

この分野は，生物を一つの系としてまるごと解析するという意味で，システムバイオロジーとも呼ばれている。日本ではE-Cellというプロジェクトがあり[11]，パーソナルコンピュータ上で細胞シミュレーションを実行できるパッケージを配布している（Linux 環 境 http://bioinformatics.org/project/?group_id=45,Windows98/Me/2000/XP 環 境 http://bioinformatics.org/project/?group_id=49)。パスウェイエディターで反応経路を作製し（代謝中間体を配置し，反応経路で繋ぎ，濃度など既知のパラメータを設定する)，シミュレータで反応を計算する仕組みである。

見たところ，まだシステムの開発と実験データによる検証が平行して行われている段階だが，大方の細胞生物研究者の意見は，「そんなもので細胞が理解できるわけがない」といったところである。細胞のドロドロした実像と，小ぎれいな細胞シミュレータの違和感は大きいのである。

しかし以下のように考えることもできる。「その研究の目的は？」という問いに対して「何々を理解すること」と答えたことはないだろうか。結局のところ，理解するとはどういう事かという定義に，この問題は関わっている。細胞の要素（蛋白質とか代謝中間体）がすべて見つかって，細胞現象の記述が完備すれば，細胞が理解できたといえるだろうか。我々のサイエンスの枠組みでは，対象を数学モデル化し，そのモデルがちゃんと動くことが，対象を理解したことの最終的な証明になるのではないだろうか。細胞レベルまでいくと，もはやすべての要素を頭の中に収めることは不可能である。しかもそれらは共時的に変動するのである。「細胞についての理解」は計算機上に実装される以外に道はなく，最終的には細胞シミュレータに行き着くしかないと思う。

２．オントロジー

オントロジーとは「対象領域についての概念を網羅的に集積して，それぞれの概念に明確な定義を与え，各概念の間の関係を定義する」ことである。これをゲノムサイエンスとその周辺領域に対して行うのがGene ontologyである[12]。現在行われているのは，生物学の記述に使われる用語を樹状グラフに整理してゆく作業で，下位の（よりspecificな）用語は，上位のより抽象度の高い用語に接続される。

Gene ontologyにはhttp://www.geneontology.org/からアクセスできる。例えば，ヘモグロビンという言葉を検索すると，IDがGO:0005833の“ hemoglobin complex ”という用語が見つかって，その定義は，“ An iron-containing, oxygen carrying complex. In vertebrates it is made up of two pairs of associated globin polypeptide chains, each chain carrying a noncovalently bound heme prosthetic group. ”である。そしてこの用語は次頁のように，上位の概念に接続される（より左側からはじまるものが上位)。

GO:0003673 : Gene_Ontology (80972)
GO:0003674 : molecular_function (66225)
GO:0008150 : biological_process (56741)
GO:0005575 : cellular_component (38547)
GO:0005623 : cell (28087)
GO:0005622 : intracellular (20002)
GO:0005737 : cytoplasm (12607)
GO:0005829 : cytosol (2318)
GO:0005833 : hemoglobin complex (17)

Gene ontologyは，計算機環境に適した生物学辞書づくりを目指しているように見える。これがいかに受けが悪いか，想像に難くない。「それで研究といえるのか」というのが標準的な反応である。

だが，このように考えたことはないだろうか。研究して何か新しいことを発見したいとき，まずしなければならないことは，過去にその分野でどんな研究がなされたかを整理して理解することである。実際にはこの過程はすっとばされていることも多いのだが，まあ建前的にはそうである。ところが，このスタートラインに立つための労力は年々増加する一方で，理屈から言えばいずれかの時点で，準備に費やす時間が個人の持つ資源（＝寿命か知的活動の限界）に追いつくだろう。

一つの解決策は外部記憶装置を有効に使うことである。もちろん文献情報はこれまで有効に機能してきた外部記憶なのだが，やがてデータ転送速度（＝読んで理解するための時間）が要求に合わなくなる。Gene ontologyは，文献情報から必要な知識を自動抽出して整理するという，より上位の構想の一環でもある。誰かがやらなくてはならない作業で，いま始めても遅すぎるくらいだと思う。

以上の２つの分野の共通点は，これまでヒトの脳で行ってきた活動を計算機上に移動する作業という点である。現在ほとんどの研究者は，自分の仕事に必要な情報はだいたい頭の中に入っているし，例え忘れても，どの文献を見ればよいか知ってると自負していると思う。それはすでに間違っているという気もするし，これからますます自信がなくなっていくはずである。

今後バイオインフォマティクスから新しいパラダイムが生みだされるとしたら，ヒトの脳内で処理しきれない情報量の中から，何らかの有意なパターンを発見するという形になると思う。単なるサービス業ではないという点では，バイオインフォマティクス研究者にとって理想の時代なのかもしれない。だがその時代は，実験研究者もバイオインフォマティクス研究者も，実態は「研究プログラムさま」に末節情報をアップロードするためのドローンにすぎないという，二流のSFみたいな世界かもしれない。その時まで生きていたいとはあまり思わない。

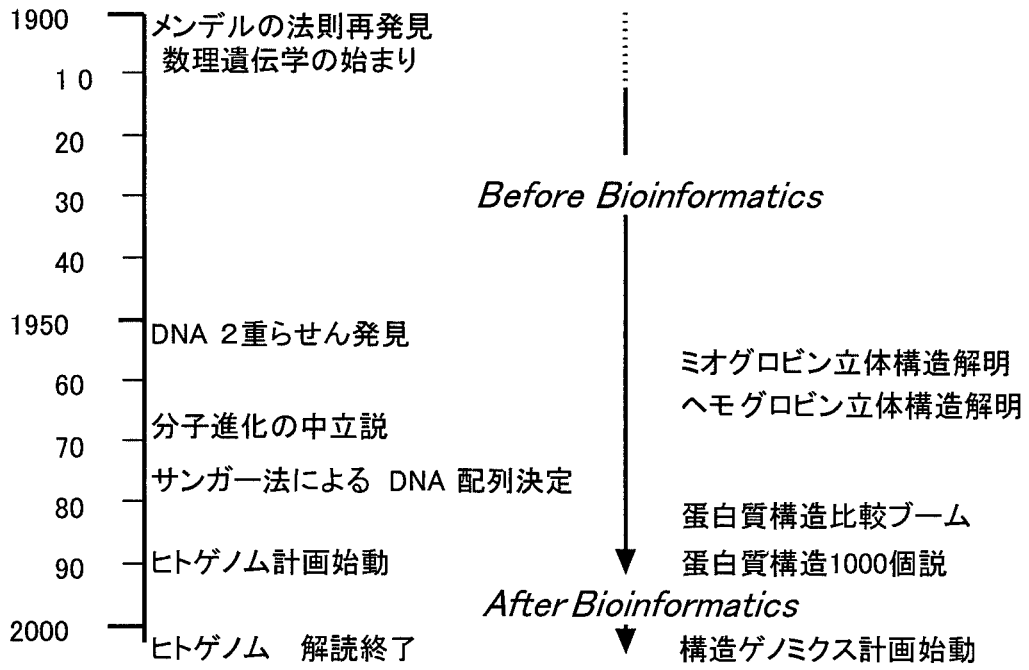

図1　超簡略バイオインフォ100年史。この文章中で触れた出来事の年表である。

参考文献

（1）Dunbrack, R.L., A scoundrel's refuge. *Nat. Struct. Biol.*, **10**, 590 (2003)

（2）Franklin, J. Bioinformatics changing the face of information. *Ann N Y Acad Sci.*, **700**, 145-152 (1993)

（3）Bains, W. Bioinformatics in Europe-the federation strikes back. *Trends Biotechnol.*, **11**, 217-218 (1993)

（4）Beltrame, F., Tagliasco, V. Confocal microscopy and cellular bioinformatics. *Cytotechnology*, **Suppl 1**:S72-74 (1993)

（5）Burley, S.K. An overview of structural genomics. *Nat. Struct. Biol.*, **suppl 7**: S932-934 (2000)

（6）Kimura, M. *The Newtral Theory of Molecular Evolution* (Cambridge Univ. Press), pp. 149-193 (1983)

（7）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)

（8）Perutz, M.F., Rossmann, M.G., Cullis, A.F., Muirhead, G., Will, G., North, A.T. Structure of haemoglobin: a three-dimansional Fourier synthesis at 5.5 A resolution, obtained by X-ray analysis. *Nature*, **185**: 416-422 (1960)

（9）Swindells M.B. Structural similarity between transforming growth factor-beta 2 and nerve growth factor. *Science*, **13**:1160-1161 (1992)
（10）Chothia, C. One thousand families for the molecular biologist. *Nature*, **18**:543-544 (1992)
（11）Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J.C., Hutchison, C. E-CELL: Software environment for whole cell simulation. *Bioinformatics*, **15**: 72-84 (1999)
（12）The Gene Ontology Consortium, Gene ontology: tool for the unification of biology. *Nature Genet*., **25**: 25-29 (2000)

（しらい　つよし ： 生物分子工学研究所・生命情報解析部門）