市場創造研究会研究報告会

「生物実験」と「情報解析」の関係　～知識の粘着性か？～

第 7 分科会　町田雅之、安達宏

生物の全遺伝子セットを明らかにすることを目的として、1980 年代後半からゲノム解析が始まりました。例えば、ヒトのゲノムの実体は約 30 億個のヌクレオチドからなる DNA ですが、4 種類のヌクレオチド（A、T、G、C）の繋がり（配列）のほぼ全てが読み取られ、これを利用して、研究・開発効率が飛躍的に向上しました。またこれにより、生物学は情報科学的な側面が強くなりました。一方で、情報解析から得られる結果は予測であり、生物実験による確認は欠かせません。そのため、多くの生物学の研究では、生物実験（ウェット）と情報解析（ドライ）の両方が必要です。この 2 つの分野は専門性の違いが大きいことから、それぞれの分野の研究者が協力して進めてきました。しかし、専門用語が全く違うことから始まって、当初からこの連携は容易ではありませんでした。ウェット・ドライの連携は、重要な課題として努力が続けられてきたにも関わらず、ゲノム解析の開始から 30 年近く経った現在でも、満足にできていないと考えている（困っている）人がほとんどです。約 20 年前まで何千万円もかかったゲノム解析も今では数十万円程度になり、遺伝子情報は短時間・低コストで入手できるようになりました。この様な状況により、研究者は、質の高い多量なデータを有効に使いこなすことへの強い要望（未充足ニーズ）を持っています。

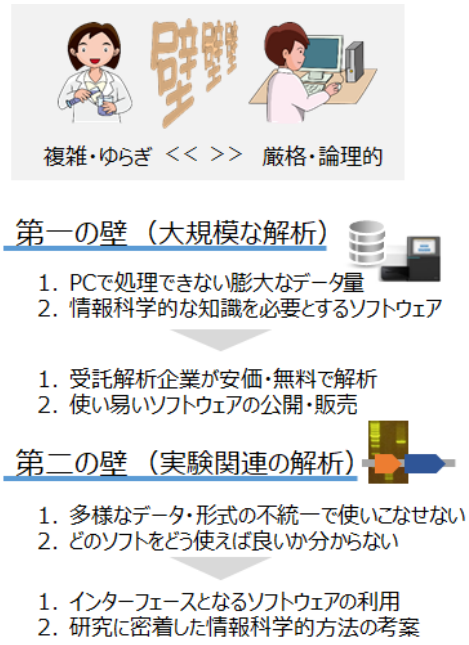

図 1　ウェットとドライの壁

DNAの配列は、基本的に 4 種類の文字を並べたものであり、生物情報の中では最も計算機に馴染みやすいと言えます。しかし、生物は複雑でゆらぎが大きいことなどから、条件を一定にしても測定結果（情報）が必ずしも同じにならないなど、一般的な誤差とは異なる感覚が必要になることがあります。現状では分からないことが多いことがこの理由の本質と思われますが、論理学を基本とする情報科学の分野では共感を得がたい部分と思われます。この様な、人間の感覚を大きく左右するような生物と情報の違いが、それぞれの立場の人の関わり合いを難しくしているのではと想像しています。即ち、ウェットとドライがなかなか納得できる連携を進められない理由は、「知識の粘着性」にあるのではないかと考察しました。

上記より、ウェットとドライの間には高い壁があると考えられます。第一の壁は、大規模なデータを処理できる高性能なハードウエアが必要で、その上で動作するソフトウェアの使い方も含めて、計

算機に関する深い知識と技術が必要なことです。現在では、ゲノム解析などを請け負う企業が誕生し、一般の PC や Excel などを使って解析ができるサイズと形式のデータに処理して提供されています。PC の性能の向上もこの壁を乗り越える一助となっています。

第二の壁は、データが多様で形式が不統一であること、ソフトウェアがたくさんあって使いこなせないことがあります。生物のデータは例外が多く、画一的な形式で表せないことがよくあります。このため、規格化された表形式に備考欄を付けて書き込むことが行われてきました。DNA 配列のデータベースの代表的な表記方法として、GenBank 形式があります。配列とともに、タンパク質として翻訳される領域など、生物学的意味に関する情報が項目毎に階層化されて記述されています。テキストで記述されているため、人が目で見て理解できること、記述に柔軟性があり、Word のようなソフトウェアで書換えることもできます。これにより、当初は想定されなかった情報も記述できますが、情報処理が複雑になり、使用するソフトウェアを選択が必要となることや、使い方が煩雑化する原因にもなると考えられます。

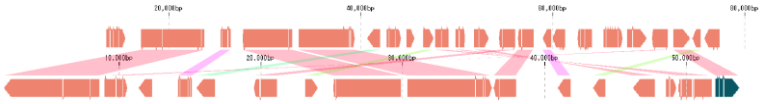

図 2A 生物学者から見て「分かりやすい」と言われる図の例

現在の生物学では、データの種類・大きさの変化が激しいこと、未開拓の研究領域が多く存在するため、興味の対象や実験方法についても変化が激しい状態にあります。そのため、情報科学的にも様々なソフトウェアが開発され、ユーザーは最も適当なソフトウェアを探したり、それを使いこなしていかなければなりません。

図 2A は、生物学者から見て「分かりやすい」と言われる図の例です。約 8,000 ヌクレオチドからなる 2 つの配列とその意味の概略を表したもので、矢印が遺伝子、遺伝子間の塗りつぶしが対応関係を表しています。この図から、遺伝子の数、大きさ、向き、遺伝子間の間隔、2 つの配列の比較による似た遺伝子の対応に関して、配列の相同性、遺伝子の欠落、移動、転座など、多数の情報を読み取ることができます。生物学者は、この図からイメージ的に情報を読み取り、生物学的な機能や着目する現象の原因を考えます。そのため、考えの方向性によって、情報を整理して分かりやすく表示するための図の形式や情報の内容も変わります。図 2B は、このうちの 1 つの配列を GenBank 形式で記述した場合のもので、DDBJ などのデータバンクでは、計算機でより扱いやすいデータ形式で運用されていると考えられます。

```
LOCUS       X86780               107379 bp    DNA     linear   BCT 12-JUN-2006
DEFINITION  S.hygroscopicus gene cluster for polyketide immunosuppressant
            rapamycin.
ACCESSION   X86780
VERSION     X86780.1  GI:987088
KEYWORDS    ABC-transporter; antibiotic transport complexes regulator;
            cholesterol oxidase regulator; cystathione synthase; Cytochrome
                  ～ 省略 ～
SOURCE      Streptomyces hygroscopicus
  ORGANISM  Streptomyces hygroscopicus
            Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
            Streptomycineae; Streptomycetaceae; Streptomyces.
            ～ 省略 ～
FEATURES             Location/Qualifiers
     source          1..107379
                     /organism="Streptomyces hygroscopicus"
                     /mol_type="genomic DNA"
                     /strain="NRRL 5491"
                     /db_xref="taxon:1912"
     misc_feature    1..107379
                     /product="polyketide immunosuppressant rapamycin"
                     /note="byosynthetic gene cluster"
     gene            complement(<1..871)
                     /gene="orfZZ"
     CDS             complement(<1..871)
                     /gene="orfZZ"
                     /codon_start=1
                     /transl_table=11
                     /product="monosaccharide transporter"
                     /protein_id="CAA60449.1"
                     /db_xref="GI:987089"
                     /db_xref="GOA:Q54286"
                     /db_xref="InterPro:IPR005828"
                     /db_xref="InterPro:IPR016196"
                     /db_xref="InterPro:IPR020846"
                     /db_xref="UniProtKB/TrEMBL:Q54286"
                     /translation="MAPRPSVRSVADVVALIDGLGKITGRAQIIWFLVFGGLFLDAYS
                     NAALSAGLGPMTTQMELTSTQVSILTATAPALILFNPLGGMLATRIGRVPPLLIAKL
                     FAIAGALLAAFAGDFTVVMLGRVLVGVAYGMDFAVAMALLAEYTPAKLGGRLNLMQAV
                     WYIATTSNLVLALVFFNLDVGADIWRWSVGSAAVVAVALLLGQRLMLKESPTWLAGKG
                     RLGEAITNLDKIYGIKAVAGTPDEMTRTPAEAPVIGLRQAGRLFRGEYLPRTILSSVI
                     SLGGSMQYFAVGMY"
ORIGIN      
        1 ggtaccagcc gacggcgaag tactgcatcg actggccgag cgagatcacc gaggagagga
       61 tcgtacgcgg caggtactcg cccctgaaca gccgtccggc ctgccggagc ccgatgaccg
      121 gcgcctcagc cggggtgcgg gtcatctcgt cgggcgtacc ggcgacggcc ttgatcccgt
            ～ 省略 ～
   107341 cggtgatctt cggcggcccg ccgaagaagc ggcggatcc
// 
```

図 2B GenBank 形式の例

第二の壁を乗り越える手段のひとつは、図 2 で

示した様な図を描くことができるソフトウェアの利用であると考えています。これは、生物学でのイメージ的な表現と計算機上でのデータ形式（論理構造）とのインターフェースであると考えられ、多様なデータを感覚的に一望することを助けます。また、表示する結果（データ）を獲得するための計算機能についても、いろいろなものが揃っている必要があります。さらに、取り扱うデータの種類が多いこと、新たなデータの種類の追加があるため、表示方法も柔軟に取捨選択や変更ができる必要があります。また、新規データの解析手段、新たな解析方法など、解析ソフトウェアに関しても、機能の追加、ユーザーによるソフトウェアの使い方の工夫や変更も必要となります。

研究・開発現場では、実験から得られる比較的小規模で多様なデータ、対象とする課題に近い機能の遺伝子の配列情報や関連する生物種のゲノム情報などの中規模なデータを、状況に応じて柔軟に取り扱う使う必要があります。このため、多くの機能と、新たな種類のデータや解析方法への迅速な対応が求められます。図 3 は、このような機能を持ったパッケージ型のソフトウェアの例です。この様なソフトウェアを使ったとしても、使い方の理解は必要であり、機能が豊富で追加が頻繁に行われれば、それだけ多くの時間と労力が必要です。

情報科学的な解析方法には、生物の研究の過程で新たに思いつくものが多くあります。この種な解析は、必要なデータの収集と選択、中小規模のデータの処理がほとんどであり、多くの場合、データ変換、既存のソフトウェアと、簡単な計算処理で実現することができます。これまで、共同研究などで具体的な解析方法を検討してきましたが、迅速な対応や適用範囲を広げるためには、常に実績を積み重ねていくこと、この方法が有効であることを示していくことが必要と考えています。その方策のひとつとして、本年 10 月に(株)ゲノムスケープを設立し、年内には本格的な業務の開始を予定しています。

図 3 遺伝子解析パッケージソフトウェアの例（インシリコバイオロジー(株)の IMC）