TM-1144

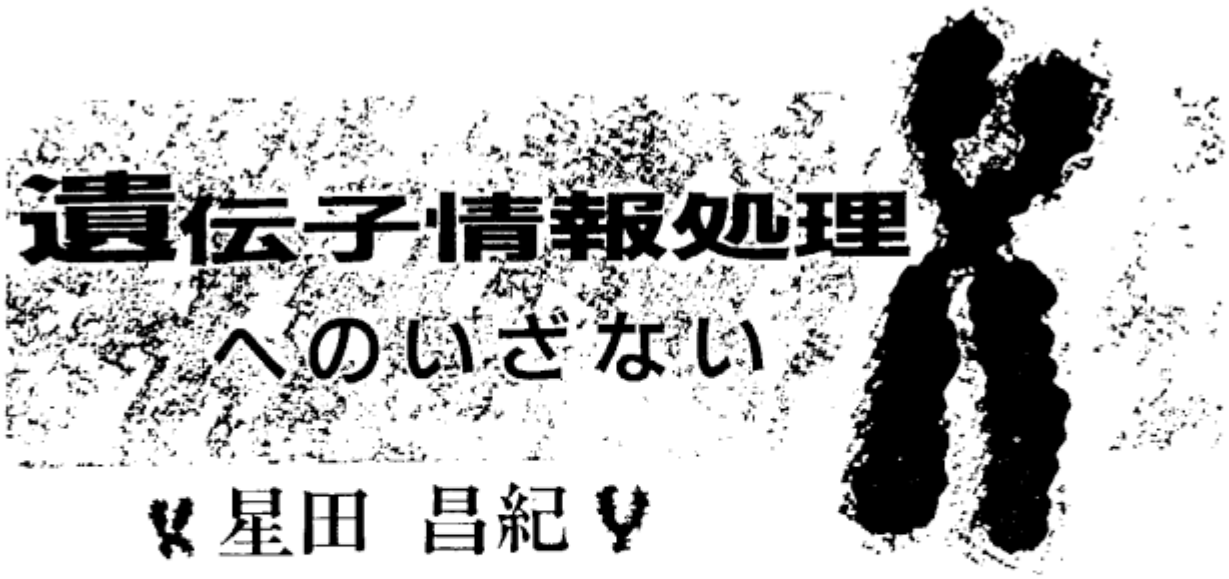
遺伝子情報処理へのいざない

星田　昌紀

January, 1992

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03)3456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology



遺伝子情報処理へのいざない

星田 昌紀

1. はじめに

もうすぐ春である．春になれば，木々は芽吹き，花が咲き，虫たちも動き始める．「生命のいぶき」という言葉があるように，たくさんの生き物が誕生する．ところが，これだけたくさんの生物が生まれてくるにもかかわらず，本当に独力で発生する命は，たったの一つも存在しない．生物には，必ずその親となる生物が存在し，先祖に由来する．他の生命にまったく依存しない「生命の自然発生」は，現在の地球上では皆無である．現存するすべての生命は，太古の昔にたった一度だけ発生した生命に源を発していると考えられている．これはまだよくわからない生命の不思議の一例である．文明が進んでも，このような不思議はまだまだ尽きない．

私も子供のころから生物に興味はあったものの，系統立てて生物学を学ぶチャンスには出会わなかった．時が経ち，大学では計算機科学を学び，職についてからは並列計算機の応用を中心に仕事をしてきた．そこで偶然にも，並列計算機を遺伝子やタンパク質の解析に応用するというプロジェクトに参加することとなった．そして今まで，生物学を学びながら，計算機応用の研究を楽しく行なってくることができた．

ふだん見かけるマクロな生命現象が，遺伝子やタンパク質というミクロな世界で解き明かされていくのは興味深いものである．その研究に私たちの計算機という道具を用いることができるのなら，なお素晴らしい．

最近，遺伝子情報に基づく諸現象を，情報処理の技術を使って研究しようという学際的な研究領域が，少しずつ脚光を浴びつつある．計算機研究者と生物学者の両者が集うコミュニティも徐々に広がっている．この興味深い分野をなるべく大勢の方々に知っていただき，少しでも計算機科学と生物学の交流に役立つことができればと思う．

この学際的な研究領域は，計算機研究者にとって二つの意味でおもしろいと考えられる．一つは，計算機科学の方法論をいろいろな角度から応用できる点であり，もう一つは生命の不思議そのものを楽しめる点である．よい「応用の研究」をするためには，その応用領域に興味をもち，どっぷりつかる必要があると思う．この分野は，情報という概念を明示的に扱うこともあり，おそらく多くの計算機研究者にとって興味深く，取り組みやすいのではないだろうか．

もちろん，おもしろいだけでなく，役に立つという側面も大きい．この分野における計算機の必要性は高い．大量のデータを用いることが多いこと，組合せ的な要素が強くて膨大な CPU パワーを使用すること，より高機能なユーザインタフェースが望まれていること，などがその理由である．

AI，ニューロ，最適化，データベース，並列計算機，ユーザインタフェースなどの研究をされている方で，やりがいのある応用分野を探していらっしゃる方，今まで，方法論の研究を中心にやってきたが，N クイーンはそろそろ脱却したいとお考えの方，その他大勢の方々に興味をもっていただきたい．この領域の研究人

図1　トリプシンインヒビターの分子模型

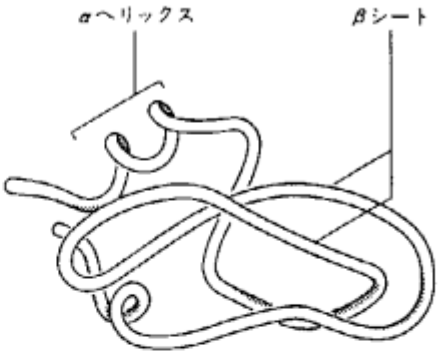


図2　トリプシンインヒビターの模式図

口が増え，領域が活性化されることになればと思う．

とりあえず，新しい領域に関する新しい知識が必要である．まずは，生命の神秘に関する基本的な事がらのダイジェストを楽しんでいただきたい．それから，現在の研究領域の基本的なトピック，続いて計算機サイドから貢献できそうな話題について述べたい．

2．生命の神秘と遺伝子情報

最終的に生体で機能する分子はタンパク質であって，遺伝子はそのタンパク質を作り出すデータの集まりである．したがって遺伝子について語るには，タンパク質についてのイメージを作っておく必要がある．そこで，タンパク質から順に話を始めることにする．

★タンパク質は能動的に機能する

もし私たち人類が植物繊維（セルロース）を分解する酵素をもっていて，カタツムリやシロアリのようにどんな固い草木でも食べることができたら，どうだろう．おなかが減ったとき，新聞紙や落ち葉を食べて空腹をまぎらわすことができるかもしれない．いや，そんなことよりアフリカ地域をはじめとするひどい食糧危機は起こらなかっただろう．残念ながら，この「セルロース分解酵素」はヒトの遺伝子の中に「コード」されていないため，固い草木を食べる練習をどれだけ積んでも，空しい努力に終わることになる．

このような「酵素」は，代表的なタンパク質である．酵素には消化を行なうもののほかにも，DNA を複製するものなどいろいろな種類がある．酵素は普通，細胞内外の溶液中にまるっこい形をして浮かんでいる．それ自身能動的に働き，めったに起こらない化学反応を，非常に起こりやすくする作用をもつ．タンパク質には酵素以外にもさまざまなものがあり，皮膚など生体のかたちを維持するもの，筋肉など運動に使われるもの，などがあるが，一応ここでは，タンパク質というと酵素のことだと考えていただきたい．

このように，タンパク質はすべての生物の重要な構成要素であり「生命のあるところ，必ずタンパク質あり」といってよい．

★タンパク質はアミノ酸の並んだひもである

図1はタンパク質の分子模型の一例である．このタンパク質は，「タンパク質を分解するトリプシンという酵素」にひっついて，その働きを一時的に止める，トリプシンインヒビター（阻害剤）という比較的小さなタンパク質である．

この分子模型を見ると，小さいとはいえ，非常に複雑な形をしていることがわかる．一方，タンパク質はアミノ酸という単位が1次元的に連なってできている．タンパク質の立体構造は複雑なので，「ほんとに1本のひもなの？」と疑ってみたくなるほどである．しかし，たしかにタンパク質はアミノ酸の1本のひもからなっている．両端を引っ張れば一直線にほどけ，結び目はできない．大きいタンパク質では，複数の単位がくっついてできているものもあるが，それぞれの単位は1本のひもからできていて，ひもがツリー状だとか，メッシュ状だとかいうことはない．数十個から数百個，ときには千個におよぶアミノ酸が連なった長い1本のひもが，コンパクトに折れたたまって，タンパク質ができている．図2は，このタンパク質をひもとして表現したものである．タンパク質を構成するアミノ酸は，全部で20種類ある．これらのアミノ酸は，それぞれアルファベット1文字で表記される慣例なので，アミノ酸配列はアルファベットの文字列で表わさ

RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRA
KRNNFKSAEDCMRTCGGA

図3　トリプシンインヒビターのアミノ酸配列

れる．図3は，このタンパク質のアミノ酸配列である．このように立体構造は複雑なタンパク質が，アミノ酸配列としては非常にシンプルに表現される．

★タンパク質のかたちはアミノ酸の配列で決まる

さて，それぞれのタンパク質は，独自の立体構造をもつ．この「かたち」によってタンパク質は物質と結合し，その結果，機能することができる．この構造を決定している要因は，なんだろうか？　実験の結果，タンパク質の構造は，基本的にアミノ酸の配列情報だけで決定されるということが判明した．これは驚くべきことである．タンパク質がとりうる可能な形状の自由度は途方もなく膨大である．にもかかわらず，この膨大な自由度の中の，たった一つの配置に向かって，すみやかにタンパク質が折れたたまっていく過程は，驚異である．多くの研究者が，この過程を計算機でシミュレートしようとして30年近くも努力を続けているが，いまだに成功のめどは立っていない．

★アミノ酸の性質は，構造や機能に寄与する

タンパク質を構成するアミノ酸は20種類だが，これらのアミノ酸はそれぞれ異なった性質をもっている．大きさ，親水-疎水性，酸-塩基性，極性，電荷などの属性がある．これらの性質がタンパク質の構造や機能と密接に結びついていることはまちがいない．しかし，そこにどのようなルールがあるのかはよくわかっていない．たった一つのアミノ酸を他のアミノ酸に置き換えただけで，タンパク質の機能がまったく失われる場合もある．たとえば酵素では，反応に直接関係する場所が，「活性中心」という狭い領域に限られている．この部分のアミノ酸を他のアミノ酸に置き換えると，機能を失う場合が多い．アミノ酸の中にも，性質の近いものと遠いものが存在する．そのため近い性質のアミノ酸に置き換えた場合は，遠い性質のものに置き換えた場合に比べ影響が少ないことが多い．

★アミノ酸はRNAの塩基三つ組によってコードされる

タンパク質を作るアミノ酸配列は，DNA上の遺伝情報に従って決定される．正確には，DNA上の遺伝情報がRNAというやはりひも状の分子にいったん写しとられ(転写という)，このRNAの情報を使って，アミノ酸配列，つまりタンパク質が作られる（翻訳という）．DNAやRNAは，「塩基」と呼ばれる「情報の単位」が並んだ1次元の配列である．RNAにはA，U，G，Cという4種類の塩基が存在し，この塩基配列がアミノ酸配列に翻訳される．4種類の塩基を使って20種類のアミノ酸をコードしているわけである．そうすると，「$4^2<20<4^3$」であるから，一つのアミノ酸を指定するために，塩基が最低三つは必要になる勘定になる．事実，まさに塩基三つ組によってアミノ酸が指定されている．なかなかうまくしたものだ．

どの三つ組がどのアミノ酸に対応するかという対応表が，すでに完成している．たとえば，GCCという三つ組はアラニン（略称A）というアミノ酸に対応している．この対応表のことを「遺伝暗号」と呼ぶ．驚くべきことに，この遺伝暗号は人間から大腸菌に至るまで，すべての生物で基本的に同じであることがわかった．この事実は，地球上の全生物が同一の祖先から進化してできたものであるという説の有力な根拠になっている．

★地球最初の生命は，RNAだけでできていた？

DNAは遺伝子の原簿であり，細胞が生きている限り大切に保存される．一方，RNAは原簿のコピーであり，使い捨てられる．DNAは，A，T，G，Cの四つの塩基からなり，RNAは，A，U，G，Cの四つの塩基からなる．DNAどうしでも，DNAとRNAの間でもAとT（またはU），GとCは相補的に対を作る．この性質を用いてDNAの複製や，DNAからRNAへの転写が行なわれている．さて，タンパク質が「機能の発現」を担当するのに対し，DNAは「情報の貯蔵」を担当している．タンパク質が手続き的であるのに対し，DNAはデータ的であるといってもよいだろう．では，RNAはどうだろうか．実は，RNAはDNAのようなデータでありながら，DNAより複雑な立体構造をとることができる．そのため，RNAの中には，タンパク質に翻訳されることなしに，それ自身が機能をもって働くものもある．つまり，RNAはデータと手続きの一人二役を演じることができるのである．それゆえ，地球最初の生命はRNAだけでできていたという説も登場している．この説に従えば，後になってDNAとタンパク質という専門担当への分業が起こり，RNAは，その仲介役として現在に至ったと考えられる．

★遺伝子はタンパク質の設計図である

一つの遺伝子は，一つのタンパク質の設計図に対応するDNA上の領域である．タンパク質の設計図を1次元的なデータである遺伝子のかたちでもつことにより，「設計図の容易な複製」「必要に応じたタンパク質の増産」「新しいタンパク質の設計」という三つの重要な作業を可能にしている．

さて，遺伝子はタンパク質情報だけを記述し，それ以上のものは，なんら「直接的には」記述していない．「それ以上のもの」とは，ヒトでいえば，顔のかたちや，足の速さや，歌のうまさといったものである．これらの情報はどこからやってくるのだろうか．やはりDNAに陰伏的な形で記述されているのだろうか．このあたりの疑問はまだよく解明されておらず，今後の研究が期待される．

★遺伝子はDNAだが，DNAは必ずしも遺伝子ではない

遺伝子として使用されているのはDNAの一部であり，遺伝子はDNA上に点在している．一般に，DNAの大部分は使われていない領域である．たとえばヒトの場合，遺伝子として実際に使われているのはDNA全体の5パーセントくらいだといわれている．ヒトのDNAは全部で約30億の塩基が連なってできているが，5パーセントというのが正しければ1.5億塩基だけが使われている勘定になる．タンパク質の平均アミノ酸数を300とすると，一つのタンパク質をコードするのに，塩基が約900個必要になる．後で述べる付加的な情報の領域も含めると，1タンパク質当りおよそ1,000塩基と見積もれる．すると1.5億を1,000で割って約15万種のタンパク質があることになる．高々15万種程度のタンパク質で，私たちの身体のかたち，恒常性，免疫機構，五感，運動能力，知能などすべてが実現されていることになる．これはびっくりである．

★DNAはDNAに自己言及する

タンパク質に翻訳されない遺伝子で，とくに重要なのは「ここから遺伝子が始まっていますよ」というようなメタな情報を与える遺伝子である．これらの遺伝子は，普通，タンパク質をコードしている遺伝子と並んで存在している．メタな情報を扱う遺伝子の領域中には，転写を抑制するための領域，つまりタンパク質の生産を抑えるために使われる領域も存在する（図4参照）．この領域はスイッチのような役割をする．この

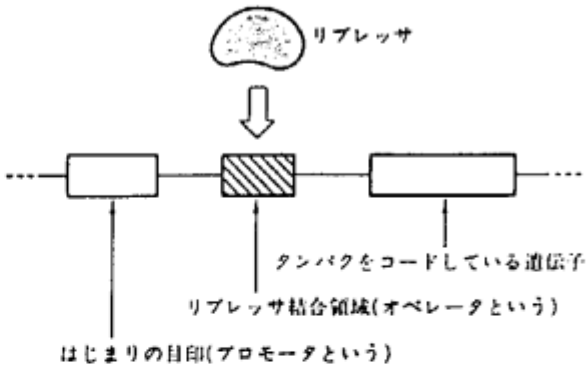


図4　DNA上の遺伝子

領域に，タンパク質生産抑制のためのタンパク質（リプレッサという）が結合すると，「スイッチOFF」になり，この領域と連続する「タンパク質をコードしている遺伝子あるいは遺伝子群」は転写されなくなる．その結果これらのタンパク質の生産がストップする．リプレッサは多種類あり，それぞれ異なるスイッチ領域に対して選択的に働く．このメカニズムを利用して特定のタンパク質の生産が調節できる．一方，リプレッサもやはりタンパク質であるから，同じDNA上にコードされている．そのため，このリプレッサの生産を調節するリプレッサも一般に存在する．きりがないようだが，リプレッサの中には，自分をコードしている遺伝子のスイッチを自分自身でOFFにするものもある．このようにDNAは自分から作り出されるタンパク質を使って自分自身のもつ遺伝子情報の発現を抑制するなど，複雑な自己言及を行なっている．

★受精卵の中ではブートストラップが進行する

DNAの自己言及についてもう少し述べたい．DNAからRNAへの転写を行なう酵素（転写酵素）が存在している．当然，この転写酵素はタンパク質の生産に必須のものである．この酵素自身もタンパク質であり，もちろんDNAの中にコードされている．だから，必要に応じてこの酵素を増産し，増えたこの酵素を使ってさらに他のタンパク質をどんどん作り出すことができる．これも一種のDNAの自己言及である．しかし，もしこの酵素が一つも存在していなかったとしたらどうだろうか．この酵素が一つも細胞内部に存在していなければ，いくらこの酵素がDNAにコードされていても，この酵素を「作り出す」ことができない．いくら完全な遺伝子情報をもったDNAが存在していても，この酵素がなければ，その細胞には何も起こらないということになる．つまり，ブートストラップを起

こして「世界が始まる」ためには，最低一つは転写酵素というプロセッサが必要である．一度ブートが始まってしまえば，後はどんどん「世界」が作られていく．そのため個体における一番最初の細胞，つまり「受精卵」の中には，DNA だけでなく，必ずこの酵素が含まれる必要がある．同じ理由で，翻訳を行なうリボソームという巨大な分子装置も，受精卵の中に必ず存在しなければならない．

★進化は DNA の塩基配列の変化によって起こる

現在のところ，進化は DNA の塩基配列に変化が蓄積することによって起こったと考えられている．この考えに従えば，進化を配列レベルでとらえることができる．この配列レベルで論じられる進化を「分子進化」と呼ぶが，この考えを用いることにより，進化学は急速に進歩した．分子進化学の示す結果は，従来の進化学で論じられていた結果をよく裏付けることがわかった．従来は化石などを使って研究が行なわれていたため，さまざまな生物学的観点から論議を行なうのに時間がかかり，また「似ている」というのに客観性がないという二つの欠点があった．しかし，配列を用いれば，高速に，また客観的に，結果を示すことができる．

また，単に従来の進化学をよく裏付けるだけでなく，今までの説を覆すような新たな事実も見つかるようになった．たとえばコンドルは従来の通説に反して，ワシよりコウノトリに近いということがわかってきた．このように「分子進化学」の意義は大きい．

以上，タンパク質からアミノ酸，RNA，DNA，進化と駆け足で説明してきた．次の章では，実際の研究領域が今どうなっているかということを述べ，計算機が応用されつつある事がらについて解説する．

3. 遺伝子情報処理研究の実際

3.1 配列解析と構造解析

ここでは，アミノ酸配列とタンパク質の構造がどのような関係にあるかという興味深い話題を中心に考えよう．もちろんアミノ酸の配列は，DNA の塩基配列に由来するものであるから，塩基配列を解析する研究も非常に重要であるが，ここではタンパク質の構造を直接規定しているアミノ酸配列に焦点をあてる．

前章で，タンパク質の構造は，基本的にアミノ酸の配列情報で決定されると述べた．つまり，アミノ酸の１次元配列の上に，タンパク質の構造や機能に関する情報がすっかり載っているというわけである．この事実は，われわれのアミノ酸配列に関する興味を喚起する．「配列解析」は，このアミノ酸配列のなかから，なるべく多くの情報を取り出そうとする行為である．

一方，タンパク質の構造を直接解析する研究も非常に重要であり，さかんに行なわれている．この構造解析については後で述べるが，現在，配列解析と構造解析をとりまく以下のような状況が存在している．

○構造の決定はむずかしい

タンパク質の分子模型が出回っているくらいだから，タンパク質の立体的な構造はよくわかっているのだろうと思うと，実際はそうではない．タンパク質の世界は電子顕微鏡で見てやっと輪郭が見える程度である．立体構造を決定するのはむずかしい．大きいタンパク質の立体構造を知るには「X 線結晶解析法」という方法が使われるが，これを行なうためにはタンパク質を結晶化しなければならない．しかし，この結晶化はむずかしく，ほとんど運まかせという状況である．タンパク質の立体構造のデータベースで有名なものに PDB（Protein Data Bank）があるが，これに登録されているタンパク質は 200 種類ほどにすぎない．

一方，アミノ酸配列だけがわかっているタンパク質は，構造がわかっているものに比べればずっと多い．タンパク質のアミノ酸配列を決定する技術はすでに確立されている．タンパク質のアミノ酸配列のデータベースで有名なものに PIR (Protein Identification Resource) があるが，これに登録されているタンパク質は数万種類で，しかも増加のスピードも PDB に比べてずっと速い．このため，構造がわからなくても，配列でやれるところまでやろうという動機が生まれる．

○進化などは配列レベルで解析したほうがやりやすい

進化は，DNA 上に起こる変化によって引き起こされている．言い換えれば，進化の痕跡は配列上にまざまざと残っている．この痕跡を調べるには，構造を調べるより DNA やアミノ酸の配列情報を調べるほうがよい．また，進化だけでなく，機能的に重要な部分を推定するにも配列情報を用いたほうが有効なことが多い．

以上のような理由から，タンパク質およびDNAの配列データベースが充実し，大量の配列データを解析する必要性が生じてきた．一方，パソコンやワークステーションが多くの研究室に浸透し，ネットワークも非常に発達してきた．このような状況の中で80年代半ばから，データベースを用いた配列研究が盛んに行なわれるようになった．

3.2 配列解析の実際

アミノ酸配列の中に含まれる情報は，構造・機能に関する情報，進化に関する情報などさまざまである．さて，実際の配列解析はどのように行なわれるのだろうか．ここに1本の配列があると仮定する．この1本の文字列を見て何かを「知る」ことができるだろうか？残念ながら，少なくとも現段階では，この1本のアミノ酸配列だけを見てなんらかの意味のある情報を抽出することは，容易ではない．ところが，うれしいことに，1本ではわからないことが複数本比較することにより明らかになってくる．

3.2.1 相同性検索

最も基本的な配列解析技術は，配列データベースの中からある1本の配列に似ているものを探すもので，「相同性検索」と呼ばれている．相同性検索を行なうと，調べたいタンパク質に似ているタンパク質が，（あれば）引っ掛かってくる．もし幸運にも，この引っ掛かってきたタンパク質の中に機能や構造のよくわかっているものがあると，調べたいタンパク質の機能や構造も似ているだろうと推測することができる．これは，「配列レベルで類似性があるタンパク質は，構造や機能も似ている」という考え方に基づいている．一般に配列の保存性に比べ，立体構造の保存性のほうが高い．つまり，配列が若干変化しただけでは立体構造に及ぶ影響は一般に小さい．もちろん，先に述べたように「活性中心」のような重要な位置に起こったアミノ酸の変化は，1個でも致命的な影響を及ぼすが，そうしたきわめて重要な部分はタンパク質の一部にすぎない．

配列レベルで高い類似性をもったタンパク質は，おそらく共通の祖先タンパク質のアミノ酸配列に（DNAレベルでの変化の反映として）変化が起こって，少しずつ異なったものに進化したと考えられる．

相同性検索には，計算機科学における文字列パターンマッチングの技術が応用できる．なかでもダイナミックプログラミング法がよく用いられるが，計算量が多いため，前処理としてハッシュを使う場合が多い．

3.2.2 マルチプルアライメント

相同な配列が見つかった場合，それらを細部にわたって比較することによりもっと多くのことを知ることができる．似ているとはいっても，配列の中にアミノ酸の置換，挿入，欠失という変化が起こって，随分と異なるものになっている場合がほとんどである．たとえば，図5はそれぞれ異なるウイルスが生産するタンパク質のアミノ酸配列で，DNAを切断するという同じ機能がある（：の左がウイルスの名前，：の右の文字列がアミノ酸配列を表わす）．このように，異なるウイルスでも同じ機能を果たすタンパク質をもっている．たとえば，HTLVというのは，ヒトT細胞白血病ウイルスといってエイズウイルスの兄弟分である．

これらの配列群は，ちょっと見ただけでは，似ているかどうかすらわかりにくい．この複数の配列の比較技術の中で最も典型的な技術が「マルチプルアライメント」である．これは，複数の配列の類似する部分を縦にそろえて並べ合わせる操作である．実例をお見せしたほうがわかりやすいので，図6を見ていただきたい．

図6は，図5の配列のマルチプルアライメント結果である．ところどころにハイフン"-"が挿入されているが，これはギャップと呼ばれる．これらのギャップを挿入することにより，性質の近いアミノ酸をなるべく縦に整列させるわけである．このギャップの挿入は進化的には，あるアミノ酸がDNAレベルで欠失したことを表わしている．マルチプルアライメントでは，同じ列に同じ文字が並んでいるほうがよいが，異なる文字でも，それらが表わすアミノ酸の性質が似ていれば同じ列に置くことを許容する．これは進化的には，あるアミノ酸がDNAレベルで他のアミノ酸に置き換わったことを表わしている．

さて，マルチプルアライメントを行なうと何がわかるのだろうか．まず進化がどのように行なわれたかを推定できる．アライメント結果から配列間のアミノ酸の差異を数えることができるが，この差異を用いて，タンパク質ごとに何回アミノ酸の置換が起こったかを算出することができる．アミノ酸の置換速度はヒトでもウマでも「タンパク質の種類が同じなら，ほぼ一定」であることが知られている．この事実を用いて，種の

```
copia  :ILDFHEKLLHPGIQKTTKLFGETYYFPNSQLLIQNIINECSICNLAK
M-MULV:LLDFLHQLTHLSFSKMKALLERSHSPYYMLNRDRTLKNITETCKACAQVN
HTLV   :LTDALLITPVLQLSPAELHSFTHCGQTALTLQGATTTEASNILRSCHACRGGN
RSV    :VADSQATFQAYPLREAKDLHTALHIGPRALSKACNISMQQAREVVQTCPHCNSA
MMTV   :ISDPIHEATQAHTLHHLNAHTLRLLYKITREQARDIVKACKQCVVAT
SMRV   :ILTALESAQRSHALHHQNAAALRFQFHITREQAREIVKLCPNCPDWGS
```

図5　ウイルスのタンパク質のアミノ酸配列

```
copia  :----------IL-DF----HEKLLHPGIQKTTK-LF--GET-YY-FPNSQLLIQNIINECSICNL-AK
M-MULV:----------LL-DFL---HQ-LTHLSFSKM-KALLERSHSPYYMLNRDRTL-KNITETCKACAQ-VN
HTLV   :LTDALLITPVLQLSPA-ELHS-FTHCGQTAL-T-LQ--------GATTTEA--SNILRSCHACRG-GN
RSV    :VADSQATFQAYPLREAKDLHT-ALHIGPRAL-S-KA-------CNISMQQA--REVVQTCPHC-N-SA
MMTV   :------ISDPIH-EAT-QAHT-LHHLNAHTL-R-LL-------YKITREQA--RDIVKACKQCVV-AT
SMRV   :------ILTALE-SAQ-ESHA-LHHQNAAAL-R-FQ-------FHITREQA--REIVKLCPNCPDWGS
(パターン)                  H....H                                  C..C
```

図6　マルチプルアライメント結果

間の進化的距離を決定でき，進化系統樹を作成できる．

次にアライメント結果から，タンパク質のどの部分が重要かという推測が可能である．たとえば図6の例では，アミノ酸がばらばらの列やギャップが多い列もあれば，アミノ酸がほとんど同じである列もある．とくに，パターンと書かれた部分は，列がすべて同じアミノ酸で占められている．このようにほとんどのアミノ酸が同じであるような列はアミノ酸が置換しにくく保存性が高いことを示している．ところが一方，突然変異をはじめとする遺伝子上の変化はどこでも等確率で起こる．ではいったい，なぜこのように保存性の高い部分と低い部分が存在するのだろうか．その理由は，次のように考えられる．遺伝子上の変化はどこでも等確率で起こるが，タンパク質には重要な部位とそうでない部位が存在する．もし重要な部位のアミノ酸に変化が起こると，普通そのタンパク質は著しいダメージを受ける．その結果，その変化を受けた個体は生存上不利になり，一般には死んでしまう．そのため，重要な部位に起こった変化は種の中に残らない．一方，あまり重要でないアミノ酸に変化が起こっても，普通そのタンパク質は以前と同様に機能する．その結果，その個体は生き続けて子供を作り，その変化が種の中に固定される可能性が高い．つまり「保存性の高い部分＝重要な部分」と考えられるわけである．この考え方を用いて構造が未知であるタンパク質でも，機能的，構造的に重要な部位を，配列だけから予測できる．

さて，マルチプルアライメントは，これまで生物学者がエディタなどを用いて手作業で行なうことが多かった．配列2本だけを自動的にアライメントする方法は行なわれていたが，複数の配列を扱うマルチプルアライメントは，組合せ的に計算量が増大するため自動化がむずかしく，生物学者の腕にたよる部分が多かった．

「アミノ酸の類似性」と「ギャップの入りにくさ」に対し，生物学的，物理化学的観点から得点を割り当てることにより，アライメント全体の評価値を定義することができる．このような定義を行なうと，マルチプルアライメントは，計算機科学ではおなじみの「組合せ最適化問題」として定式化できることになる．私たちのグループはこのような定式化に従って，並列3次元ダイナミックプログラム法と，並列シミュレーテッドアニーリング法という二つの方法を用いて自動化を試みている．図6の結果は私たちのシステムを使って得られたものである．実際は，アミノ酸の性質に応じて美しく色付けされて表示されるが，誌上ではカラーでお見せできないのが残念である．

3.2.3　モチーフ抽出

配列の保存性の高い部分の中で，特に特徴的なパターンをもっている非常に重要な部分を「モチーフ」と呼ぶ．たとえば図6の例の「H....H　C..C」のパターンはモチーフの一例である．このモチーフはジンクフィンガーと呼ばれる有名なもので，この四つのアミノ酸が亜鉛イオンを挟みこんで指のようなかたちをとり，DNA や RNA にタンパク質が結合する際に使わ

れることがわかっている．モチーフは先に述べた「活性中心」などに対応していることが多い．

このようにいったんモチーフがわかってしまえば，このパターンを用いてデータベースをサーチするなどして同じ機能のタンパク質を探すことができる．またこれらのモチーフが将来もっと集まって，モチーフどうしの関係や構造が詳しくわかってくれば，タンパク質を記述する「言語」への道が開かれるかもしれない．これについては最後にもう一度触れる．

このモチーフ抽出には，先述のマルチプルアライメントを用いる方法のほかに，あらかじめ，長さとアミノ酸数を固定したパターンを発生させておき，これでデータベースをサーチする方法などが使われている．

3.3 構造予測

ここまで，アミノ酸配列解析について述べてきた．ここからはタンパク質構造の話題を紹介しよう．

3.3.1 フォールディングシミュレーション

タンパク質が長く伸びた状態から，水溶液中でコンパクトに折れたたまった状態に至るまでを，計算によって再現しようとするのがフォールディング（折れたたみ）シミュレーションである．これが完全に成功すればタンパク質の立体構造も予測できることになる．

もっとも素朴なシミュレーション法は，タンパク質を構成する原子と，周辺の水の分子を構成する原子，合わせて数万個を球で表現し，球間に働く力に基づいて，各球を時間的なステップで運動させる方法である．この方法は分子動力学と呼ばれており，厳密解法は，現在のどの計算機をもってしても現実的な時間内に行なえない．なぜなら，原子間に働く力は，静電力のような遠隔力があり，考慮すべき原子数が増えると，計算量が組合せ的に増大するからである．原子の振動や回転がピコ秒以下の現象であるのに対し，フォールディングは秒のオーダの現象であることからも，いかに多くの計算が必要かがわかる．

そこで考えられるのは近似的なシミュレーション法である．アミノ酸自体を一つの仮想的な球にしたり，働く力を制限したり，移動する位置や回転する角度に制限を加えたり，さまざまな近似が考えられている．これはまさに，多くの条件のうちから，フォールディングに不可欠な本質的条件を見極める，知的な作業である．そして，それらの条件を満たす解を高速に求めるアルゴリズムを見いだすのは，計算機科学の課題である．これまでいくつかの試みがなされてきているが，残念ながら，まだ誰も成功していない．私たちのグループは，遺伝暗号を解読したことで有名な，アメリカの国立衛生研究所（NIH）と共同研究を行なっているが，ここの研究者の多くは「これに成功した者には，確実にノーベル賞が与えられる」と話している．

3.3.2 2次構造予測

タンパク質の構造予測には，アミノ酸配列から部分的な構造を予測し，この部分的構造の組合せとして，3次元の立体構造を推定するボトムアップなアプローチもある．その部分的構造を普通2次構造と呼び，図2に示されるようにアミノ酸のひもがらせん状に巻いたαヘリックスと，まっすぐに伸びた部分が複数並んだβシートを指す．2次構造予測とは，アミノ酸配列のどの部分がαヘリックスで，どの部分がβシートかを予測する問題である．この予測には非常に多くの方法が試みられている．古くは，統計的な頻度を利用するものから始まって，最近では，エキスパートシステムやニューラルネット，AIの学習理論も導入されている．しかし，おもしろいことに，どの方法の予測率も60％台に低迷している．これはどうやら，2次構造の決定に，配列の部分的な情報のほかに，タンパク質全体にわたる大局的な情報が必要なことを示しているようである．そうなると2次構造予測の成功の鍵は，大局的な情報をどう認識し，どう予測に盛り込むかであるようだ．それに関して計算機科学の分野に何かノウハウがないだろうか？　音声認識の技術や，制約問題解決の手法が，利用できるかもしれない．

3.3.3 構造解析

2次構造を用いた立体構造予測がうまくいっていないのは，これまでのαヘリックス，βシートという2次構造が，構造予測のための部分構造として不適当だからなのかもしれない．そうした動機から，すでに構造が知られているタンパク質の構造解析がなされている．つまり，αヘリックスやβシートに代わる新たな部分構造の表現を探そうというのである．

典型的なのは，連なった数個のアミノ酸が形成する形を，角度分布や位置分布に注目して，統計的に分類する解析法である．こうした解析により，αヘリックスやβシートよりもさらに精密な構造表現が求まっている．さらに，構造の大局的特徴をおおざっぱに捉

えるにも，計算機科学の知見が使えそうである．その一つはフラクタル解析である．私たちのグループではクラスタ内のアミノ酸分布を主成分分析して，タンパク質構造のフラクタル次元を求める研究を行なっている．

4. 遺伝子情報処理の将来と夢

さて以上，遺伝子情報処理という研究分野についていくつかのトピックをお話ししてきた．ここでは，この分野への期待を，私個人の夢を含めて述べたい．

配列解析の究極の目標あるいはロマンは，配列上に我々にも理解し得る「言語」が存在しているのかということをつきとめることであろう．この言語をここでは仮に「DNA 言語」と呼ぶことにしよう．DNA 言語は，もし存在していたとしても，人類の使用する自然言語や，プログラミング言語とは，かなり異なったものであるに違いない．なぜなら，その「言語」としての目的がずいぶん異なるからである．自然言語の目的は，主として人が人に対して意思伝達を効率的に行なうことであり，プログラミング言語の目的は，主として人が計算機に対して意思伝達を効率的に行なうことだと考えられる．つまり，自然言語もプログラミング言語も「人」の存在なくして，その目的を語れない．ところが，DNA 言語には「人」の存在は，とりあえず関係がない．「DNA 言語の高級性」に関する議論もなされているが，このような考えがなんとなく捉えにくいのは，「高級性」という概念が普通は「人」の存在を前提にしているためなのかもしれない．このように，ひとくちに「言語」といっても，ずいぶん目的が違うわけである．では，DNA 言語の目的は，あるとしたら何だろうか？　より効率的に増殖が行なえるということだろうか．「効率的に」とは「速く」と同じ意味だろうか．

たしかに「速く」という方向性も存在する．大腸菌をはじめとする「高速な増殖」に力点を置く生物グループは，徹底的に遺伝子をチューンし，驚くほどコンパクトな遺伝子体系をもつ．その結果「彼ら」は非常に高速に増殖することができる．チューンアップの極めつけは，ある一つの DNA の塩基配列を，3 塩基コードの読み枠をずらすことによって，異なる遺伝子として利用していることである．このオーバラップした遺伝子は，驚異というほかはない．現在，生物の自然発生が皆無であるのは「彼ら」があまりに高度に「進化」しているため，自然発生で新しく微生物が生まれたとしても生存競争にかなわないためかもしれない．

一方，ヒトをはじめとする高等な生物は，「高速な増殖」というキーワードでは説明できない．「高等」という言葉を使ってしまったが，だいたい私たち人類は，あの高度にチューンされた大腸菌より高等であると言い切れるのだろうか．この疑問に答えるには，「高等」の意味をきちんと定義する必要があるだろう．

ヒトをはじめとする「高速な増殖グループ以外の生物」は，いったい何を目標にしているのだろうか．そして DNA 言語は，その目標にふさわしいスペックをもっているのだろうか．このような疑問に対して，現在のところ誰も生物学的に説得力のある説明を与えることができない．しかし，いずれは，このような「問いかけ」を私たち自身が行なうべき時が来るだろう．

さらに，DNA 言語の重要な点は，その言語だけを考えていても不十分だということである．言語がどのようなスペックをもっているかということだけでなく，この言語上に築かれた処理系統を解明しなければ遺伝子情報の解明にはならない．先にも述べたように，一つの遺伝子が「直接」記述しているのは，一つのタンパク質にすぎない．それゆえ，それらのタンパク質がどのように相互作用するかということまで含めて理解しなければ，「遺伝子情報に基づく諸現象」を包括的に理解したことにはならない．「ヒト遺伝子解読プロジェクト」という言葉を最近よく耳にする．しかし，単にヒトの DNA の配列を全部読みとっただけでは不十分で，その「意味」まで解釈しなければならない野心的な課題である．私たちは，私たち自身の秘密をどこまで解き明かすことができるだろうか．

私は，遺伝子情報処理の研究が，最終的には，がんをはじめとする遺伝子にまつわる難病で苦しんでいる人たちを救っていくための研究に，間接的につながっていくものだと理解している．その意味で，この研究を推進していく力は，学術的な興味によるものだけではないように思われる．たとえば，タンパク質の立体構造が簡単に予測できるようになれば，単に科学的好奇心を満たすだけでなく，人工的な新しいタンパク質を容易に設計し，薬品や医療に応用する道が開ける可能性がある．また，遺伝子のメカニズムの解明がもっ

と進めば，がんの根本的治療法も見つかるに違いない．

一方，計算機科学にとっても，この研究分野は価値があるように思われる．「遺伝子の研究から計算機科学は何を学べるのか」という問いかけがそれである．最近流行の「遺伝アルゴリズム」などはその先駆けかもしれない．遺伝子の研究から計算機科学に新しいパラダイムがもたらされる日も夢ではない．

5. さいごに

遺伝子情報処理に限らず，これらさまざまな分野で学際研究が盛んになっていくだろう．しかし，学際研究を掛け声だけに終わらせないためには，本当は，並々ならぬ努力が必要である．私の経験では，その努力とは「サービス精神」と「恥を恐れぬ心」だと思う．

二つの異なる立場の人が，サービス精神不在で専門用語を乱発したりしているかぎり，コミュニティは広がってはいかないし，恥をかくのを気にして質問しなければ，いつまでたってもわかり合うことはできない．お互いに興味をもてるように配慮し合い，また理解するために歩み寄ることが不可欠である．

しかし恐れることはない．まずは，おもしろいと思えるところから始めよう．人間が本当に一所懸命やれることは，つまるところ，その人にとっておもしろいことなのだから．学際研究には，うまくやれば，2倍楽しめるチャンスがある．ときにはちょっぴり勇気を出して新しいドアをノックしてみるのもいい．扉を開けば，新鮮な世界が広がるかもしれない．

謝辞

本稿の執筆にあたり，ICOT の石川幹人氏，京都大学の中井謙太氏，松下電器産業の中崎義己氏，小山雅庸氏，大西達也氏には大変貴重なコメントをいただきました．深く感謝いたします．

(ほしだ まさき　㈶新世代コンピュータ技術開発機構)