解　説

知識情報処理技術とヒトゲノム計画

Knowledge Processing Technologies and Human Genome Project

金久　實*1　新田　克己*2　小長谷　明彦*3　田中　秀俊*2
Minoru Kanehisa　Katsumi Nitta　Akihiko Konagaya　Hidetoshi Tanaka

*1 京都大学化学研究所
Institute for Chemical Research, Kyoto University.
*2 新世代コンピュータ技術開発機構
Institute for New Generation Computer Technology.
*3 日本電気(株) C&C システム研究所
C&C System Laboratory, NEC Corporation.

1991年5月27日　受理

Keywords: DNA, RNA, protein, genome, motif, deductive and object oriented detabase, alignment, simulated annealing, MDL.

1. はじめに

生物の全遺伝情報を構成するゲノム (genome) を解析する実験的方法論が急速に洗練され，DNA の文字数にして 30 億，遺伝子の数にして 10 万と言われるヒトのゲノムを解読する計画がアメリカを中心に各国で始まった．この計画では，同時に他の生物ゲノムの解読も並行して行い，その比較の中からヒトの遺伝子の働きが解析される．その成果は医学・生物学の基礎研究のみならず，がんや遺伝病，遺伝的要因を持つ日常的な多くの病気の診断と治療に役立つものと考えられる．さらに，生物個体の発生・分化と老化の問題から生命の起源と進化の問題まで，生命の不思議を解明する手掛りを与えることになるだろう．

ゲノム解析がもたらすのは，後述するように大量のマッピングデータと配列データ，およびそれに基づくさまざまな生物的知見であり，生物学はかつて経験したことのない大量情報の時代を迎える．もはや生物学もコンピュータ技術なしに発展はあり得ない．しかしながら，そこでは物理学や化学のような数値計算中心の方法論では不十分である．生物学は基本的に経験的な学問であり，ゲノム解析がもたらす大量の情報を生物学の知識として体系化することと，知識処理という形でのコンピュータ化をすることが必要なのである．

筆者らは，1989 年より新世代コンピュータ技術開発機構 (ICOT) の中で，遺伝子情報処理ワーキンググループ (GIP/WG) の活動を行ってきた．分子生物学の関係者と計算機の関係者がまずお互いの言葉を理解し合うところから始めて，新しいデータベース技術およびデータ解析の技術を実際にゲノムの問題に適用し，ソフトウェアを開発するまでに至っている．生物科学，特に分子生物学と，計算機科学，特に人工知能との間には，研究分野として自然な接点があると我々は考えており，本稿でその簡単な紹介をしたい．

2. ヒトゲノム計画の概要

2・1 歴史的位置づけ

ヒトをはじめあらゆる生物の遺伝情報は，DNA (一部のウイルスでは RNA) の分子情報としてコードされている[1]．DNA は 4 種類のヌクレオチド，A，C，G，T が鎖状につながった生体高分子であり，ヌクレオチドの違いを与える側鎖部分を塩基という (図 1)．遺伝情報は塩基の並び方，すなわち塩基配列としてコードされているので，基本的に 1 次元の文字列情報であると考えてよいが，実際に生体内では立体的な形を持った複雑な分子どうしの相互作用として情報は読み取られていることにも注意を要する．ある生物の持つ遺伝情報全体をゲノムと呼び，生物種によって数千から数十億の塩基数が含まれている．ゲノムの中で特に重要な部分はタンパク質や機能性 RNA に翻訳される遺伝子の部分で，これらの分子の働きで生物の生命活動が維持されている．タンパク質 (protein) は 20 種類のアミノ酸が数百個程度つながった生体高分子で，ヌクレオチド 3 文字のコドンからアミノ酸 1 文字が決定さ

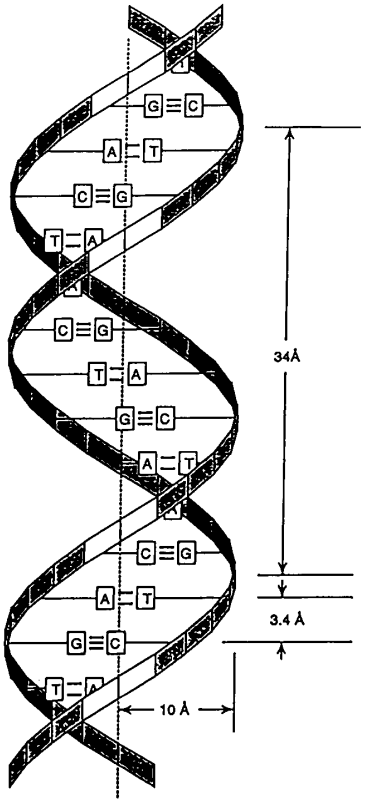
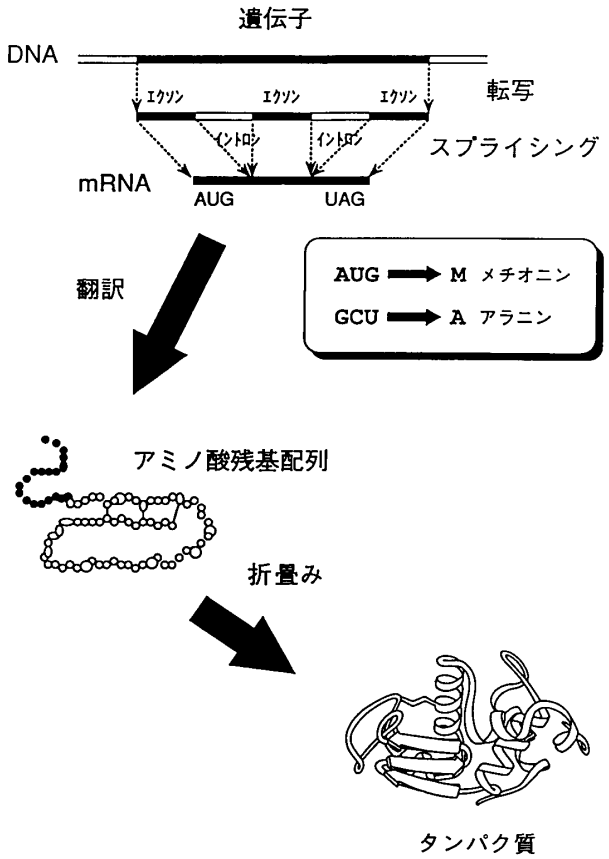


図1　DNAの構造



RNAはA, U, G, Cの4種類のヌクレオチドからなる配列である．3文字の連続するヌクレオチド（これをコドンという）は一つのアミノ酸に翻訳される．例えば，AUGからM（メチオニン）が生成され，GCUからA（アラニン）が生成される．

図2　DNAとRNAとタンパク質

れる．タンパク質は3次元的に複雑な形をしており，その特異的な立体構造が生物機能と深く関連している（図2）．

1963年にWatson-Crickが提唱したDNAの二重らせん構造モデルにより，AとT，CとGの塩基対が持つ相補的構造から，遺伝という生命の基本問題の分子的理解が可能となった．その後30年近くの生物学の歴史は，生命現象を分子レベルで理解しようという分子生物学の歴史であった．それは同時に，タンパク質と核酸（DNA，RNA）の1次構造および立体構造を解明するために，次々に新しい実験技術が開発されてきた技術革新の歴史でもあった．その中でも1970年代初めの遺伝子組換え技術に端を発したバイオテクノロジーの進歩は，生物のゲノムを直接解読することを可能にしたのである．ヒトゲノム計画は分子生物学の技術革新の必然的な帰結であり，今後15年ほどの間にヒトの全遺伝情報を解読しようとしている．

学問の発展がもたらしたものとはいえ，この計画が生物学で初めてのビッグプロジェクトとして発足するまでには，いくつかの政治的要因があった．最も重要なことはアメリカエネルギー省（DOE）が1986年に最初に計画を立案したことである[2]．ヒトの全遺伝子を解析することの重要性には誰も異論がないが，それをいつどのようなスタイルで実践するかについては，アメリカ国立衛生研究機構（NIH）を中心とした従来の生物学の主流派からはおそらく現在のような計画推進策は生まれなかったであろう．実際，これまで個人的な研究に依存してめざましい発展を遂げてきた生物学に，物理学のような大規模プロジェクト研究を導入することへの反感は今でも根強い．しかしながら，この計画を契機に知識情報処理技術を中心とした新しいコンピュータ技術が導入され，生物学はさらに新たな展開を示すのではないだろうか．

2・2　日米のヒトゲノム計画

アメリカでは1988年にNIHとDOEの間で合意が成立し，両者は協調してヒトゲノム計画が推進されることになった[3]．1990年10月より正式に共同の5カ年計画が始まり，NIHは年間1億ドル程度，DOEはその半分程度の予算を使っている．

塩基配列の決定法にはトップダウン方式と，ボトムアップ方式がある．ボトムアップ方式とは，DNA断片の配列決定をした後，コンピュータで断片間の順序関係を調べてつなぎ合わせていく方式である．現在の塩基配列決定法では一度に読み取れるDNA断片の長さは数百塩基であり，一般にゲノムが大きくなると繰返

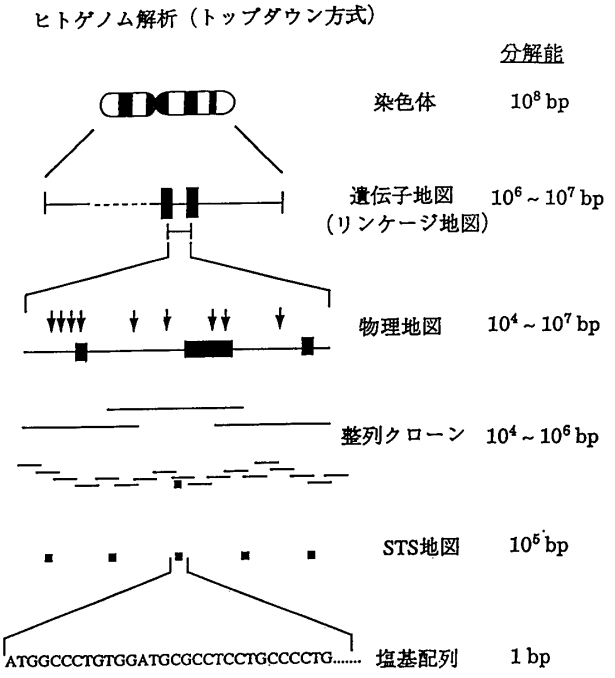


図3 ヒトゲノム解析（トップダウン方式）

し配列が多くなるので，ボトムアップ方式だけで解読できるゲノムの大きさには限界がある．そこで，配列決定をする前に断片どうしがどのように並んでいるかを調べる物理的マッピングを行うのがトップダウン方式である．トップダウン方式を理解するためには，まずマッピングとは何かを理解する必要がある．

マッピングにはさまざまな種類があるが，結局のところは図3に示すように，ゲノムを異なる解像度および異なる方法で眺めたときの地図を作成する作業である．ヒトには46本の染色体があり，22対の常染色体とX，Yの性染色体の24種類から構成される．ある遺伝子が24種類のどの染色体上にあるかを調べることは，ヒトのゲノムの中でどこにあるかを最も低い解像度で調べたことに相当する．逆に最も高い解像度で調べるには，30億の全塩基配列を決定して，その中で遺伝子の塩基配列を同定することになる．地図には大きく分けると遺伝的地図と物理的地図の二つがある．前者は二つの場所の相対的位置関係を遺伝的解析（リンケージ解析）で調べるものであり，後者は絶対的な位置をさまざまな方法で調べるものである．トップダウン方式とは物理地図の解像度をどんどん高めていって，塩基配列決定が可能な大きさまでもっていく方式であるといえる．

物理地図作成で必要な作業は，整列クローンライブラリといって，重なりあったDNA断片を作り，並べていくことである．解像度を反映して，大きな断片で並べたり小さな断片で並べたりすることが必要である．1980年代になって染色体および大きなDNA断片を取り扱う技術が大きく進歩し，マッピングを系統的に行うことが可能になってきた．例えば，大きな断片を作るにはゲノムDNAを少数の箇所で切断しなければならないが，普通の制限酵素*1では認識配列が短いためたくさんに切れすぎてしまうので，新しいはさみのデザインが必要である．また，大きな断片を単離して（クローン化して）増幅するための方法として，パルスフィールド電気泳動法とYACベクタが開発された．アメリカの5カ年計画では，このような技術開発と実際のマッピング作業に重点が置かれており，配列決定はまだ行われない．

我が国では主に文部省と科学技術庁がヒトゲノム解析計画を進めている[4]．文部省では1989年に学術審議会の建議により，ヒトゲノム解析の準備研究が開始され，1991年4月より5カ年計画のプログラムとして正式に発足した[5]．科学研究費（創成的基礎研究費および重点領域研究費）が年間6億円で，東京大学医科学研究所に設置されたヒトゲノム解析センターや全国各地の施設等整備の経費を含めても，アメリカNIHの10分の1以下の予算規模である．アメリカではヒトゲノムの全塩基配列決定という明確なゴールを設定し，そこに至る最短方法をとっているが，我が国では予算面でも組織作りの面でも同じことは行えない状況にある．文部省の計画では，アメリカのように全ゲノムの完全なトップダウン方式ではなく，機能的に重要な部分から出発し解析する領域拡大方式がとられる．また，ゲノムDNAだけでなく，遺伝子部分がメッセンジャーRNAとして発現したcDNAライブラリの解析も行われる．また，情報解析の部分は独立の重点領域研究として行われ，生物科学と情報科学の接点に新しい研究分野を形成することを目指している．

一方の科学技術庁では，1981年より科学技術振興調整費でDNA塩基配列決定法の自動化に関する装置開発に着手し，アメリカDOEと同じ時期の1985年頃にはヒト遺伝子解析に関する議論もなされたが，国家プロジェクトには至らなかった．現在では，理化学研究所での自動シーケンシングシステム開発と酵母6番染色体の配列決定，科学技術振興調整費による染色体地図作成，新技術開発事業団によるレーザ光を使った技術開発などが行われており，年間10億円程度の予算が使われている．このほかにヒトゲノム解析計画に関連したものとして，厚生省が病気の遺伝子に関する計画を，農水省がイネゲノム計画を開始したが，通産省はまだ正式プロジェクトを発足させていない．

＊1 DNAの特定の塩基配列部分を認識し，そこを切断する酵素．

3. 情報解析技術に関する課題

3・1 データベース技術

生物科学における情報解析でおそらく最も重要なことは，データをどのように蓄積し，体系化するかであろう．ヒトゲノム計画の中でのデータベースの問題は以下の五つのレベルに分けて考えることができる．

① 生物材料バンク
② ラボノートブック
③ コミュニティデータベース
④ 公共データベース
⑤ 知識ベース

第1に，実験研究の前提となるクローンライブラリなどの生物材料バンクはこれまで実際に生物材料を収集，保存，頒布することがなされていたが，最近STSと呼ばれる情報化が進んでいる．STS(Sequence Tagged Site)とは，例えばクローンを短いユニークな塩基配列で代表させたもので，個々のクローンを保存しなくてもSTSを目印としてPCR法というものにより目的のクローンを作り出すことができるようにしたものである．

第2に，生の実験データを蓄積したのがラボノートブックで，これは研究室あるいは研究所といった単位で整理される．

第3は，例えば，cDNA解析グループ，大腸菌ゲノム解析グループといった特定の実験研究に係わるデータベースで，グループ内でデータを共用するためのものである．

第4は，雑誌等に公表されたデータを蓄積した公共データベースで，一般に生物関係でデータベースというときはこれをさす．

第5は，本稿の中心課題である生物学の知識を体系化した知識ベースで**3・2節**で述べることにする．

上記①，②，③にもそれぞれ興味深い問題があるが，ここでは④の公共データベースに限ってもう少しみてみよう．公共データベースとは，国などの資金援助によりデータベース作成活動が行われているもので，表1に示すように，分子生物学関連ではほぼアメリカが独占している．1970年代初めより急増しはじめた配列データについては，これまで雑誌等に公表されたデータがGenBank/EMBL核酸塩基配列データベース，およびPIRタンパク質アミノ酸配列データベースに蓄積されてきた．学問分野の急激な進展とともに，アメリカではデータベースの再編成が行われており，GenBankとPIRは1992年から異なる形式になる見込みである．文献データベースのMedlineを基礎とした配列データベース作り，塩基配列とアミノ酸配列の一本化，配列のみのバックボーンデータベースと付加価値のあるデータベースとの区別，などが議論されている．ゲノム解析がもたらすのは配列データだけでなく，マッピングデータもまた大量に蓄積される．これまで古典的なリンケージ解析による遺伝子地図のデータベースとして，エール大学のHGML (Human Gene Mapping Library)が知られていたが，これは1990年にジョンズ・ホプキンス大学のGDBに移され，今後は物理地図のデータを中心としたデータベースに変遷していくものと思われる．

ゲノム計画を支えるデータベースのデザインとして今後の重要な研究課題は，異なる種類のデータをいかに統合するかであろう．各種マッピングデータから配列データまで解像度を反映した階層構造があること，ヒトとマウス，大腸菌と枯草菌といった生物種間の関連があること，ゲノムの情報が発現すると分子の立体構造や機能の情報となること，などはどのように表現すればよいのだろうか．また，これらのデータは文字列だけでなく，グラフィクスやイメージなども含まれるので，ユーザインタフェースの研究も重要である．さらに，データが大量でかつ頻繁に追加・更新される状況では，データベースの構築と提供を，ネットワークでの分散処理環境で実現しなければならない．

3・2 データ解析技術

大量データをデータベースとして蓄積し提供する技術と深く関連しているのが，大量データを解析する技術である．実験的なゲノム解析の最終産物は配列データであり，それから生物的な意味を解釈するには別の方法論が必要である．これをコンピュータ解析技術だけで可能にすることが我々の目標である．

現在でも，配列データの解釈の際にコンピュータは部分的には大切な役割をしている[6][7]．新しい配列データが決定された場合，それと類似な配列がデータベースの中に存在するかを調べる方法をホモロジーサーチという．配列データが類似であれば生物的な機能が類似であるという経験的事実があるため，類似配列がもしあればそこから機能に関する知見を得ることができる．ここで注目すべきことは，コンピュータはあくまでもデータベースからの情報検索の部分にしか使われていないことで，検索結果から機能についての推論を行うのは専門家の知識に依存していることである．検索に推論機能も含める演繹データベースのアプローチも考えられるが，それ以前の問題としてデータベー

表1　分子生物学関連データベース

データの種類	データベース名	作成地
文献データ	Medline	国立医学図書館(アメリカ)
核酸塩基配列データ	GenBank	ロスアラモス国立研究所(アメリカ)
	EMBL	欧州分子生物学研究所(ドイツ)
タンパク質・アミノ酸配列データ	PIR	ジョージタウン大学(アメリカ)
タンパク質・核酸立体構造データ	PDB	ブルックヘブン国立研究所(アメリカ)
ヒト遺伝子地図データ	GDB	ジョンズ・ホプキンス大学(アメリカ)
ヒト遺伝病データ	OMIM	ジョンズ・ホプキンス大学(アメリカ)

(略称)　EMBI: European Molecular Biology Laboratory
PIR: Protein Identification Resource
PDB: Protein Data Bank
GDB: Genome Data Base
OMIM: Online Mendelian Inheritance in Man

ス自体を見直す必要がある．

塩基配列やアミノ酸配列のデータベースは本来，生物的な分子の情報を蓄積するという立場から出発した．例えば，異なる著者が報告している同じタンパク質のアミノ酸配列を一つのエントリにまとめることは，当然の作業として行われてきた．しかしながら，配列データの急増でこれらの作業が中途半端になっているのが現状で，配列データに生物的解釈を入れず文献データと同じレベルで扱ったデータベース（バックボーンデータベースと呼ばれる）と，さまざまな生物的付加価値を入れたさまざまなデータベースに区別する動きがアメリカで出ている．表1のGenBank/EMBLやPIRは中途半端なデータベースであるため，その検索に推論機能を取り入れることはあまり意味がないだろう．しかし，これらのデータベースに対しても普通の情報検索と同じように，大量のデータをいかに高速に検索するか，特に配列データに置換や挿入・欠失などの曖昧性を許した場合にどうするかという問題がある．これは後述するように，並列処理の応用が最適の問題であると思われる．

生物的付加価値をつけたデータベースの中で特にコンピュータによる推論を目的としたものが，前節で分類した第5のレベルの知識ベースである．ゲノム研究における情報解析で最も重要な課題は，結局，知識ベースの構築と知識処理技術の開発である．ここで，知識の具体例を示すためにモチーフと呼ばれるものを考えてみよう（図4）．

モチーフとは，例えば生物機能が共通の配列データグループで保存されている特徴的な局所配列パターンである．ホモロジーサーチでは，データベース中のデータ一つ一つと照合するが，知識処理の立場では個々のホモロジーグループを特徴づけるモチーフと照合を行い，同時に生物的意味づけに関する推論までを行うことができる．つまり，遺伝子言語で書かれた文章を解釈するのに，以前に出版された文章を一つ一つ比較して類似なものがあるかを探すのではなく，以前の文章から辞書や文法体系を作り上げて言語処理を行うのである．モチーフ辞書は集約されたデータであるのでずっと高速に処理できるだけでなく，ホモロジーサーチでは検出できない弱い類似性でも同じようにモチーフという形に蓄えることができる．しかしながら，実際問題として辞書や文法体系をどのように発見するかについては今後の課題である．現状ではモチーフはそれぞれの分野の専門家により見いだされる場合が多いが，大量のデータをさまざまな見地からグループ化し，グループの特徴抽出をコンピュータで自動的に行うこともある程度は可能である．

図4　モチーフ

情報科学研究者の参入により，コンピュータによる大量データからの知識獲得，生物的な実験事実に対する知識の表現法，配列データの生物的意味解釈といった問題に，新しい展開があることを期待したい．

4. 第五世代コンピュータプロジェクトにおける遺伝子情報処理

4・1 研究目標

前節で述べたように，遺伝子情報の解析には情報処理技術の利用が不可欠である．このような観点からの現在の分子生物学のデータベース技術・データ解析技術の問題点は以下のように集約される．

(1) **データベースにおける機能情報の利用の困難**

データベースやデータ解析で使う生物学的知見の多くは，遺伝子やタンパク質の特徴や機能に関するものと言ってもいいだろう．既存の配列データベースの問題点の一つは，そのような機能記述

表2　データベースの併用例

データベース	補完される情報
PIR＋酵素 DB	マクロな化学反応に関する機能
PIR＋モチーフ DB	特徴的な部分領域とその機能
PIR＋立体構造 DB＋2 次構造 DB	構造
PIR＋遺伝子配列 DB＋遺伝子地図	コーディング領域と塩基配列

が不十分なことである．機能記述は量的にも不十分なうえに，分子生物学上の新たな知見を反映するために頻繁なフォーマットの変更を余儀なくされる．機能記述の充実には酵素 DB，モチーフ DB といった機能データベースとの併用を含めた知識ベース化がまず考えられるが，複数のデータベースを併用するソフトウェアを機能記述の複雑化や詳細化に合わせて改訂していくのは容易なことではない．

したがって，配列データベースの利用は配列の生データと参照文献データに限られているのが現状である．

(2) データ量の爆発

GenBank に含まれる DNA の 1 次構造のデータは，塩基数にして 1982 年に 50 万であったものが，1985 年には 500 万，1990 年には 5 000 万と，爆発的に増加している．配列相互の比較解析にはこの組合せの量を考える必要があり，さらに深刻である．解析ツールの高速化と大容量化の需要は非常に大きいと言る．

(3) 生物学知識の利用技術の不足

配列解析に関しては，配列の特徴記述やタンパク質の機能記述を利用することにより，いっそう質の高い処理を行うことができる．この知識は，教科書にあるような生物の知識だけでなく，実験データから帰納的に得られた知識を含む．

機能情報の充実と有効利用には，生物学のデータベースの統合化と，利用ノウハウの知識ベース化とが要求される．データ量爆発への対処には，並列処理などを利用した配列の検索/解析の高速化技術が有効である．生物学知識の有効利用には，生データから知識を抽出する知識獲得技術や，生物学知識ベースを利用した配列の解析技術などが必要である．

第五世代コンピュータプロジェクトとして，逐次推論マシン PSI や並列推論マシン実験機 Multi PSI*2，データベース技術，知識処理技術を開発してきた．それらの技術の応用として，遺伝子情報処理の研究を行い，分子生物学研究の環境を提供することを目的とした活動を行っている．

(1) 分子生物学における統合的な知識ースを目指す研究

(2) タンパク質の構造予測，機能予測システムを目指す研究

- 配列解析の並列プログラムの開発
- タンパク質配列のモチーフを利用した知識獲得技術の開発

以下，それぞれについて簡単に説明する．

4・2 分子生物学の統合的知識ベース

既存のデータベースの問題点は，前述のように機能記述の利用不十分にある．機能記述の充実は，機能のデータベースとの併用で図ることができる．タンパク質配列データベース PIR を例にとると，表 2 のような併用が有効である．

データベースの併用を容易化するには，データベースを何らかの形で統合する必要がある．統合には三つの方法が考えられる．

(1) データベースの標準化

一つはデータベースの標準化を進めることである．国際学術連合会議（ICSU*3）の科学技術データ委員会（CODATA*4）は，記述する属性（生物活性など）の属性名を共通化し，異なるデータベースどうしが属性データを補完しあうという形の相互利用を国際的に可能にするというプロジェクトを推進している(8)．

(2) アクセスソフトウェアの開発

従来から試みられている方法としては，データベースの併用をサポートするアクセスソフトウェアの開発がある．その際データベース全体を DBMS 下に統合管理することは当然有効だが，膨大なデータ量を蓄える環境を用意する必要に加えて，一般的なデータベースモデルである関係モデルには載りにくいデータであること（フォーマットの複雑さ，フォーマットの頻繁な変更など）が問題となる．

現在，PSI 上の非正規関係 DSMS，Kappa に

＊2　第五世代コンピュータプロジェクトで開発された並列計算機．64 台の要素プロセッサが 2 次元メッシュ状に結合されている．

＊3　International Council of Scientific Unions.

＊4　Committee on Data for Science and Technology.

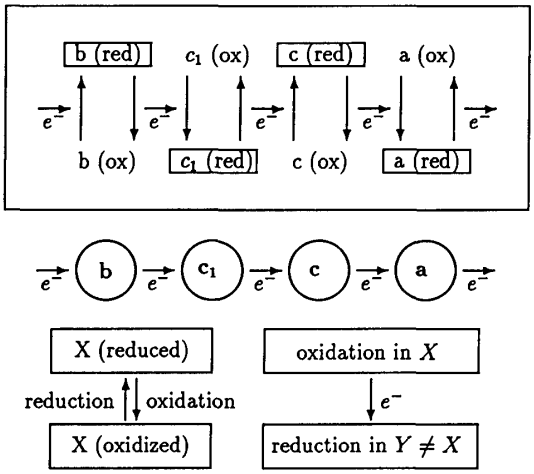


図5 シトクロムの電子輸送・全体図および分解例

GenBank, PIR と酵素 DB が容易に格納できることを確認し，統合環境を試作中である[9]．

(3) 知識ベースアプローチ

機能に関する記述の利用方法を考えると，ただ単に特定のタンパク質の関連する化学反応式を検索したりする程度では利用価値が低い．複数の化学反応式を知識として持ち，あるタンパク質の存在の有無がその反応式で記述される人間の代謝に対してどのような影響を与えそうか，などといったことを，できれば推論できてほしい．

よって，機能記述は推論に利用できるように，意味論のしっかりした言語で記述されていることが望ましい．関係モデルは一つの有力な候補である．しかし，機能記述の典型である化学反応もまた，関係モデルに載りにくいデータ構造をしている．

図5はシトクロムというタンパク質の電子輸送に関する化学反応式と電子の動きを，一挙に表したものと基本的知識に分解したものとで比較した図である．

シトクロムの機能は分解表現のほうがより明確になっている．枠内に示された全体図は，推論できるような枠組みがあれば，下の三つの知識から演繹することが可能である．そして，各知識はフレーム言語のような枠組みがあれば十分表現できる．

ICOTでは現在，Quixote[10][11] というDOOD言語*5[12] を開発し，分子生物学の知識を，データベースを含めて統一的に表現することを考えている．

4・3 タンパク質配列の並列解析プログラム

タンパク質配列データベース PIR には，図6のようなアミノ酸配列が格納されているが，その立体構造はわかっていないものが多い．タンパク質の配列から，その構造や機能を予測することができれば，新しいタンパク質の設計や構造の解明に役に立つ．そのためには，配列から特徴（例えば，モチーフ）を抽出し，その特徴と意味（構造や機能）との関係を求めて，辞書として蓄積することが必要である．特徴抽出や構造予測のプログラムはいくつか開発されているが，その機能は十分でなく，また，実行時間がかかるものも多い．

そのような特徴抽出の一つの方法として，マルチプルアライメント（Multiple Alignment）がある．タンパク質のマルチプルアライメントとは，複数のタンパク質配列が与えられたとき，配列にギャップを入れたり，左右にシフトしたりして，同じアミノ酸が1対1に対応するように並置する技法である．マルチプルアライメントは，モチーフの抽出や系統樹の作成などに用いることができる．例として二つの配列が与えられたときのアライメントをあげる(図7)．どのアミノ酸を対応させるかによって，同じ入力配列から複数のアライメントが行える．アライメントの一つの評価方法に Differential Score を用いる方法がある．これは，アライメントされた配列のミスマッチのアミノ酸の個数を u とし，ギャップの個数を v とするとき，

$$\text{Score}=\text{dif} * u+\text{gap} * v$$

で定義される．ここで，gap はギャップのペナルティ，dif はミスマッチのペナルティである．この Score を最小にするアライメント，すなわちギャップとミスマッチの少ないアライメントを最良のアライメントとする．

Difference Score を最小にするアライメントを求める手法は，いろいろ開発されている．その中でも DP (Dynamic Programming) の方法がよく知られている．これは二つの配列を

$$A_1, A_2, A_3, \cdots A_m$$
$$B_1, B_2, B_3, \cdots, B_n$$

としたとき，

$$D(i,0)=0$$
$$D(0,j)=0$$
$$D(i,j)=\min(D(i-1,j)+\text{gap},\ D(i,j-1)+\text{gap},\ D(i-1,j-1)+\text{dif}(A_i,B_j))$$

を $0\leq i\leq m, 0\leq j\leq n$ の各点について計算するものである．$D(m,n)$ が最適アライメントの score となる．ここで，$\text{dif}(A_i,B_i)$ はアミノ酸 A_i と B_i の性質の類似度を考慮したミスマッチのペナルティである．各点のデータは隣接する三つの点のデータに依存するので，(0, 0) から (m,n) まで波面状に計算が進行する．この波面

＊5 Deductive and Object-Oriented Database language.

```
///
ENTRY           CCHP        #Type Protein
TITLE           Cytochrome c - Hippopotamus
DATE            19-Feb-1984 #Sequence 19-Feb-1984 #Text 28-May-1986
PLACEMENT          1.0     1.0     1.0     4.0     1.0
SOURCE          Hippopotamus amphibius #Common-name hippopotamus
ACCESSION       A00008
REFERENCE       (Complete sequence with experimental details)
   #Authors     Thompson R.B., Borden D., Tarr G.E., Margoliash E.
   #Journal     J. Biol. Chem. (1978) 253:8957-8961
   #Comment     3-Ile was also found.
COMMENT         The amino end is acetylated.
COMMENT         Cys-14 and Cys-17 covalently bind heme; His-18 and
                  Met-80 are the axial ligands of the heme iron.
SUPERFAMILY     #Name cytochrome c
KEYWORDS        mitochondrion\ electron transport\ respiratory
                  chain\ oxidative phosphorylation\ heme\
                  acetylation
SUMMARY       #Molecular-weight 11530  #Length 104  #Checksum  9264
SEQUENCE
                    5         10        15        20        25        30
      1 G D V E K G K K I F V Q K C A Q C H T V E K G G K H K T G P
     31 N L H G L F G R K T G Q S P G F S Y T D A N K N K G I T W G
     61 E E T L M E Y L E N P K K Y I P G T K M I F A G I K K K G E
     91 R A D L I A Y L K Q A T N E
```

図6　PIRのデータ

に沿って並列計算が可能である（図8）.

DPの手法は二つの配列をアライメントする代表的な手法であるが，三つ以上の配列を同時にアライメントすることはノードの数が指数的に増加するので一般には困難である．そこで通常は，与えられた配列を似ているものから二つずつ選んでアライメントしていく方法が多く用いられている．

並列論理型言語KL 1[*6]で三つの並列のアライメントプログラムを開発したので，その概要を示す[13].

(1) 3次元DPによるアライメント

DPのアライメント手法を3次元に拡張し，三つの配列を同時にアライメントするようにした．すなわち，図8のメッシュ構造を3次元に拡張し，それをMulti PSIの複数のプロセッサに分割して割り当てている．入力された配列（4本以上）は，3本ずつ予備的にアライメントされ，最後にマージされて全体のアライメントが作られる．

並列推論マシンMulti PSIの64台のプロセッサを用いてこのプログラムを実行すると，1台のプロセッサを用いた場合に比べて，30倍以上の高速実行が行われる．

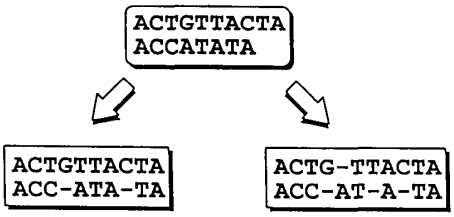


図7　アライメントの例

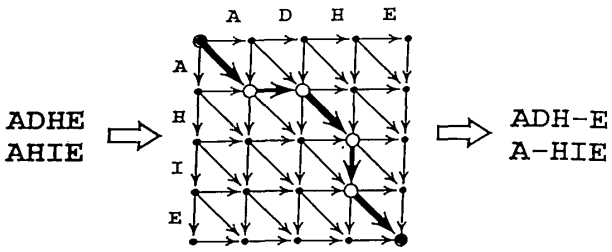


図8　DPによるアライメント

(2) トーナメント方式によるアライメント

まず入力された複数（N本）の配列の中ですべての2本ずつの配列の組を比較する．この組合せは$N*(N-1)/2$通りであるが，これは複数のプロセッサを用いて並列に計算される．その中で最も類似する2本の配列をアライメントして1本の抽象的な配列にまとめる．この配列と残りの$N-2$本の配列との組合せの中で最も類似するものをアライメントする．この操作を繰り返して，入力のすべての配列を最終的に1本の配列にまとめ，その結果を利用して，すべてのアライメントを再構築する．

(3) シミュレーテッドアニーリング[*7]によるアライメント

＊6　述語論理に理論的基礎をおく計算機言語．この言語で書かれたプログラムは，ICOTで開発された並列推論マシン上で並列に実行される．

＊7　組合せ最適化問題を確率的に解く手法．初期状態の組合せに対し，組合せをランダムに微小変形させ，評価値が向上すれば，初期状態を新たな状態に置き換える．評価値が悪化する場合は，特定の確率で新たな状態に置き換える．これを繰り返すことにより，最適な組合せを求める．「温度」というパラメータを変化させて，新たな状態に置き換える確率を制御する．

表3 ミトコンドリアシトクロムCを分別する配列モチーフから計算される記述長

配列モチーフ	ML	PL	UL	総計	$\hat{P}_1$	$\hat{P}_2$
CXXCH	36.2	10.1	225.5	271.7	0.356	0.9993
CXXCH and GPXLXG and PGTKM	80.0	9.4	7.36	163.0	0.932	0.9993
CXXCH and PGTKM	55.8	9.4	80.7	145.9	0.906	0.9993

入力されたすべての配列をそのまま並べたものを初期状態とする．次に，ランダムに一つの配列の一つのアミノ酸を選び，そこにギャップを挿入して微小変形したアライメントを作り，Differential Scoreを計算する，という操作を繰り返すことにより，アライメントの質を高めていくシミュレーテッドアニーリング（Simulated Annealing）手法を金久は開発した．シミュレーテッドアニーリングの一般的な問題点は，温度を降下させるスケジューリングをいかに決定するかである．ここでは，複数のプロセッサに異なる温度のアニーリングを実行させ，ランダムに二つのプロセッサのデータを交換させることによって温度変化を行わせる並列シミュレーテッドアニーリング[14]を用いたスケジュール不要のアライメントプログラムを実現した．

以上の三つの手法を組み合わせることにより，質の良いアライメントを求めることが可能である．現在は，専門家自身によるアライメントとの比較，アミノ酸の属性を利用したアミノ酸のクラスタ化，疎水性の指標を利用したアライメントの評価などを検討している．

また，マルチプルアライメント以外に，アミノ酸配列が折りたたまれてタンパク質の複雑な構造を形成していく過程をシミュレーションする並列プログラムを開発中である．

4・4 モチーフを用いた知識獲得

生物学知識ベースを構築する際の核となる技術が配列情報からの知識獲得技術である．ここでは，そのような知識獲得技術の一つとして，記述長最小（Minimum Description Length）基準による配列モチーフ抽出技術を紹介する．

（配列）モチーフとは，**3・2節**で紹介したように，共通の先祖を持つタンパク質の配列に固有に見られる配列パターンを意味する．例えば，ミトコンドリアシトクロムCと呼ばれるタンパク質の配列モチーフを抽出することを考える．このアミノ酸配列は“CXXCH”というパターンを持っている．ただし，ここで，文字Cはシステインを，文字Hはヒスチジンを表し，Xは任意のアミノ酸と照合する（名前なし）変数を表す．そこで，配列モチーフの表現法の一つとして，次のようなパターンを分類条件とするような分類規則を利用する方法が考えられる．「もし，配列Sがパターン“CXXCH”を持てば，Sはミトコンドリアシトクロム Cである」．

しかしながら，一般に生物の配列情報は突然変異その他の影響により多くのノイズが混在しているため，上記のような決定的な分類規則を抽出することは極めて困難である．実際，ミトコンドリアシトクロムCの場合，アミノ酸配列データベース（PIR 21.0，全データ6 158例）において，パターン“CXXCH”を持つ配列は189例であるが，そのうち，ミトコンドリアシトクロムCに属する配列は67例である．また，ミトコンドリアシトクロムCに属するが，パターン“CXXCH”を持っていない配列が3例存在する．したがって，配列モチーフの表現法としては以下に示すような確率的な分類規則のほうがより適している．「もし，配列Sがパターン“CXXCH”を持てば，確率67/189でSはミトコンドリアシトクロムCであり，そうでなければ，確率5 966/5 969でその他のカテゴリである．」ただし，このような，確率的な分類規則の抽出を行う際は次の2点において十分な注意が必要である．

- 学習セットに対する分類規則の正確さ（分類精度）よりも，未知のデータに対する正確さ（予測精度）の高い分類規則の抽出．
- 過剰学習の回避．

すなわち，抽出する分類規則の良し悪しは，与えられた学習セットをどれだけ正しく分類したかではなく，未知のデータをどれだけ正しく分類できたかにより評価すべきである．この評価基準を誤ると，学習セットに特化した分類規則の学習，すなわち，「過剰学習」を行ってしまう可能性がある*8．

小長谷らは，この問題を解決するために，記述長最小（Minimum Description Length）基準を配列モチーフの抽出に適用した[15]．このMDL基準の特徴は，

*8 例えば，もともと80%程度の分類精度しか持たない学習セットに対しては，80%程度の正しさで分類できる規則が一番もっともらしい分類規則であり，このような学習セットから100%正しい分類規則を求めることは過剰学習であり，得られた規則の予測精度はかえって悪くなることが多い．

抽出した規則の複雑さと正確さを「記述長」という尺度で定量的に判断することにある．一般に，複雑さと正確さはトレードオフの関係にあるので，両者の記述長の和を最小化するよう分類規則を求めれば，過剰学習の問題を避けることができる．また，MDL基準で求めた分類規則は真の分類規則に早い速度で収束することが知られている[16]．

MDL基準により抽出した配列モチーフの例を表3に示す．表において，MLは抽出した規則の複雑さを，PLは確率パラメータ $\hat{P}_i$ の記述長を，ULは正確さを表す．ミトコンドリアシトクロムCの例では，配列モチーフのパターンとして，"CXXCH" だけでは単純すぎ，"CXXCH" and "GPXLXG" and "PGTKM" は複雑すぎ，"CXXCH" and "PGTKM" がこの三つの中では，MDL基準の意味で一番もっともらしい分類規則であることを示している．MDL基準は必ずしも生物学的に意味のあるモチーフの抽出を保証するものではないが，予測精度の高さを保証しているという意味で，配列モチーフ抽出のための重要な基準の一つといえよう．また，第五世代計算機プロジェクトでは，このような基準に基づいて自動的に配列モチーフの抽出を行うシステムを構築中である[17]．

5. おわりに

知識情報処理技術とヒトゲノム計画の概要について説明した．ここで説明したように，ゲノム情報解析には，データベース技術や知識処理技術が必要である．我々は，昨年，公開ワークショップ「知識情報処理技術とヒトゲノム計画」を開催し，分子生物学者と計算機科学者の交流を図った[18]．この研究は，分子生物の研究者のみでなく，人工知能の研究者にとっても興味あるテーマであり，今後，両者の協力関係はいっそう深まることが予想される．

◇参考文献◇

（1） 分子生物学教科書
Watson, J. D., *et al.* (松原，他 監訳)：遺伝子の分子生物学，上下，トッパン．
Alberts, B., *et al.* (中村，他 監修)：細胞の分子生物学，教育社．
Lewin, B. (松原，他 訳)：遺伝子，上下第3版，東京化学同人．

（2） アメリカ政府機関報告書
Report on the Human Genome Initiative, *U.S. Department of Energy, Office of Health and Environmental Research* (1987).
Mapping and Sequencing the Human Genome, *Commission on Life Sciences, National Research Council* (1988).
Mapping Our Genes, The Genome Projects: How Big, How Fast, *Congress of the United States, Office of Technology Assessment,* Johns Hopkins University Press (1988). (伊藤 訳：Newton Special Issue ヒトゲノム解析計画，教育社)

（3） アメリカ5カ年計画
Understanding Our Genetic Inheritance, The U.S. Human Genome Project: The First Five Years FY 1991-1995, *U.S. Department of Health and Human Services, National Institutes of Health, and U.S. Department of Energy, Office of Health and Environmental Research* (1990).

（4） 日本政府機関報告書
ヒト遺伝子解析に関する総合的な研究開発の推進方策について，航空・電子等技術審議会 (1988).
大学等におけるヒト・ゲノムプログラムの推進について(建議)，学術審議会 (1989).

（5） 文部省5カ年計画
我国におけるヒトゲノム解析の推進に関する調査研究，文部省科学研究費補助金総合研究A「ヒトゲノム・プログラムの推進に関する研究」報告書 (1990).

（6） 五條・堀，他：大量DNAデータを対象とした遺伝情報のコンピュータ，情報処理，Vol. 31, No. 7 (1990).

（7） 西川：タンパク質の2次構造予測，情報処理，Vol. 31, No. 7 (1990).

（8） 沖林，他：蛋白質の属性データベース，情報学シンポジウム 1990, pp. 73-77 (Jan. 1990).

（9） Yokota, K. and Tanaka, H.: GenBank in Nested Relation, *Joint Japanese-American Workshop on Future Trends in Logic Programming 1989* (Oct. 1989).

（10） 安川，横田：ラベルつきグラフに基づくオブジェクトの意味論，情報処理学会研究会報告 人工知能 73-13 (1990).

（11） 森田，羽生田：Quixoteのオブジェクト識別性，情報処理学会研究会報告，人工知能 73-12 (1990).

（12） 横田，西尾：演繹・オブジェクト指向データベース，情報処理，Vol. 31, No. 2 (Feb 1990).

（13） 石川，星田，広沢，戸谷，鬼塚，新田，金久：並列推論マシンを用いたタンパク質の配列解析システム，情報処理学会研究会報告 情報学基礎 23-2 (1990).

（14） 木村，瀧：時間的一様な並列アニーリングアルゴリズム，電子情報通信学会ニューロコンピューティング研究会，NC90-1, (1990).

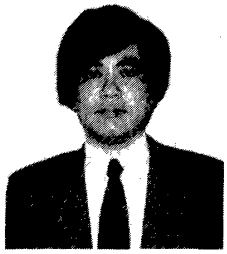
（15） 小長谷，山西：記述長最小基準の遺伝子情報処理への適用について，日本ソフトウェア科学会第7回大会論文集，pp. 101-104 (1990).

（16） Yamanishi, K: A learning criterion for stochastic rules, in *Proc. of the 3rd Annual Workshop on Computational Learning Theory,* pp. 67-81 (1990).

（17） 小長谷，新田，山西：遺伝子情報知識ベースシステムの構想について，情報処理学会研究会報告 人工知能 73-9, pp. 79-88 (1990).

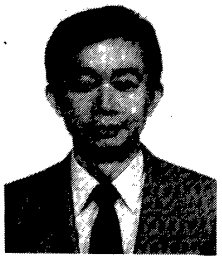
（18） 公開ワークショップ「知識情報処理技術とヒトゲノム計画」プログラム・講演要旨集 (1990).

——著 者 紹 介——



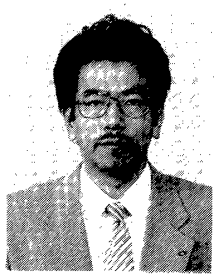
金久　實（正会員）

1970年3月東京大学理学部物理学科卒業．1970～76年同大学院修士・博士課程および日本学術振興会奨励研究員．理学博士．1976～79年アメリカジョンズ・ホプキンス大学医学部博士研究員．1979～81年アメリカロスアラモス国立研究所博士研究員．1981～84年同正研究員（この間，1982～84年にNIHへ出向）．1984～85年アメリカ国立衛生研究所（NIH）客員研究員．1985～87年京都大学化学研究所助教授．1987年同教授．1991年東京大学医学研究所教授（兼任）．理論分子生物学，特に遺伝情報の知識処理とタンパク質の機能設計の研究に従事．物理学会，化学会，生物物理学会，生化学会（評議員），分子生物学会，蛋白工学会，情報処理学会各会員．



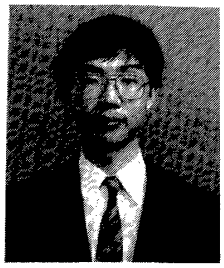
新田　克己

1975年3月東京工業大学電子工学科卒業．1980年3月同大学院博士課程修了．工学博士．同年，電子技術総合研究所入所．1989年7月より（財）新世代コンピュータ技術開発機構に出向．現在，同研究所第7研究室長．並列推論マシンの応用プログラムの開発に従事．情報処理学会，日本ソフトウェア科学会各会員．



小長谷　明彦

1985年東京工業大学大学院理工学研究科情報科学専攻修士課程修了．同年，日本電気（株）入社．現在，C&Cシステム研究所コンピュータシステム研究部主任．以来，先端的プログラミング言語および推論マシンの研究に従事．知識処理，並列処理を利用した遺伝子情報処理に興味を持つ．情報処理学会，電子情報通信学会，日本ソフトウェア科学会会員．



田中　秀俊

1986年3月東京大学工学部計数工学科卒業．同年，三菱電機（株）入社．1989年3月より（財）新世代コンピュータ技術開発機構に出向．現在同研究所第三研究室所属．知識ベースシステムの応用として分子生物データベースの研究に従事．情報処理学会会員．