Vol.37 No.10 情 報 処 理 Oct. 1996

解 説

ゲノム情報

1. ゲ ノ ム 情 報 学†

金 久 實††

1. はじめに

分子生物学の進歩は，細胞レベル，個体レベルでの遺伝子や分子の働きを明らかにしつつあり，とくに各生物種におけるゲノムプロジェクトが，生命系システムの基本部品である遺伝子および遺伝子産物(タンパク質およびRNA)のカタログを明らかにしつつある．このカタログを出発点とし，遺伝子相互また分子相互のかかわりに関する膨大なデータから基本部品間の配線図を解明することにより，生命現象を情報の流れという観点で体系化して理解することが可能になっていくものと考えられる．1980年代のおわりに始まったゲノムプロジェクトでは，これまで主に情報処理技術という観点から情報科学と生物学の融合が行われてきた．ゲノムプロジェクトが実際に大量のデータを生産するようになって，今後は情報科学のより基礎的な側面で生物学との融合が始まるものと考えられる．本稿では，1991年度から1995年度まで行われた文部省科学研究費補助金重点領域研究「ゲノム解析に伴う大量知識情報処理の研究」(領域略称名：ゲノム情報)において大きく発展した「ゲノム情報学」の概略を解説する．また1996年度から2000年度までの予定で行われている重点領域研究「ゲノムサイエンス：ヒトゲノム解析に基づくバイオサイエンスの新展開」(領域略称名：ゲノムサイエンス)においても，ゲノム情報学は新たな展開をみせるものと予想され，その展望についても述べる．

2. 統合データベース

重点領域研究「ゲノム情報」の第1の目標は，生物学における大量かつ多様なデータの統合データベースを実現することであった．これは各研究者のデスクトップでデータベースを利用できる環境を作るという情報インフラストラクチャの問題とも関連し，実用的な観点からプロジェクト発足後の最優先課題としてとりあげた．統合化には1つのデータベースに再編成して統合する考え方もあり得るが，さまざまなデータベースが世界各地で作られ日々更新されている状況を考えるとこれは現実的でない．そこで異なるデータはそれぞれ独立に管理し，データ間の相互参照を可能にするメカニズムを導入することにより，弱い統合化をはかるアプローチが一般的である．米国NCBI (National Center for Biotechnology Information) が提供するEntrezは，文献，塩基配列，アミノ酸配列，立体構造を中心とし，各生物ゲノムの染色体地図や生物系統樹なども含めた統合データベースシステムである．NCBIはASN.1形式によるデータ交換を提唱し，独自に作成しているGenBankとMedline以外のデータベースもASN.1形式に変換され，実質的に各データ項目のレベルでの統合を行っている．

ゲノム関連データベースは，いずれもデータの概念的な単位であるエントリーの集合として眺めることができる．たとえば配列データベースであれば1つの配列が，文献データベースであれば1つの文献が1つのエントリーとして，付随情報とともに蓄積されている．そこで，筆者のグループではデータ項目の統合は行わず，エントリーレベルでさらに弱い統合化を行っている．各データベース間では関連するデータを相互に参照することが一般的に行われているので，

データベース名：エントリー名

で1つのデータを参照することにすると，

データベース名1：エントリー名1
→データベース名2：エントリー名2

† Genome Informatics by Minoru KANEHISA (Institute for Chemical Research, Kyoto University).
†† 京都大学化学研究所

の2項関係が統合化の基礎となる．ほとんどのデータベースが単純なフラットファイルとして取得できる形で公開されていること，同じ塩基配列データや同じアミノ酸配列データでもデータベースにより表現の仕方が異なることから，各データベースが提供するフラットファイルそのままを共通のユーザインタフェースで利用できるシステムDBGETを開発した．これはゲノムネットデータベースサービス[1]

http://www.genome.ad.jp/

の中心的なシステムとして公開している．

1996年6月現在，DBGETシステムでは表-1の中のGenBankからLITDBまで16のフラットファイルデータベースがサポートされ，京大化研と東大医科研でデータベースの二重化が行われている．フラットファイルといってもテキストだけでなくイメージや座標データも含まれ，マルチメディアデータベースとなっている．また，上記の2項関係だけを蓄積したデータベースLinkDBを作成しており，これらデータベース間のリンク情報のほかに，さらに表-1のMedlineからTDBまで8つのデータベースに対するリンク情報も含まれている．相互参照表だけを日々更新して維持し，実際のデータは海外にアクセスして取得するわけである．LinkDBはエントリーレベルでの弱い統合化しか表現していないが，実はLinkDBには2項関係から演繹する機能があり，リンク情報を組み合わせたり逆向きにたどったりして，新たな2項関係を作ることができる．すなわち，データベースには直接記載されていない事実データをルールから演繹的に導く能力がある点で，LinkDBは簡単な演繹データベースになっている[2]．ゲノム情報プロジェクト発足時に開始した我々のインターネット(ゲノムネット)活動，およびその後の爆発的なWWWの普及は，DBGET/LinkDBのような弱い統合化の考え方に合致するものであった．

3. 配列解釈

重点領域研究「ゲノム情報」の第2の目標は，実験によるゲノム解析の最終産物である塩基配列

表-1 DBGETによるデータベースの統合化

データの内容	データベース名	メディア	備考
塩基配列	GenBank(DDBJ含む)，EMBL	テキスト	
アミノ酸配列	SWISS-PROT, PIR, PRF, PDBSTR	テキスト	
立体構造	PDB	テキスト，3次元座標	グラフィックス操作と連動
配列モチーフ	PROSITE, EPD, TRANSFAC	テキスト	
酵素と代謝化合物	LIGAND	テキスト，イメージ，2次元座標	グラフィックス操作と連動
代謝パスウェイ	PATHWAY	テキスト，イメージ，2次元座標	
アミノ酸変異	PMD	テキスト	
アミノ酸指標	AAindex	テキスト	
遺伝病	OMIM	テキスト	
文献(タンパク質)	LITDB	テキスト	
文献(分子生物学)	Medline	テキスト	リンク情報のみ
ゲノムデータ	GDB,MGD,ACeDB,AAtDB,SacchDB,FlyBase,TDB	テキスト	リンク情報のみ
リンク情報	LinkDB	テキスト	

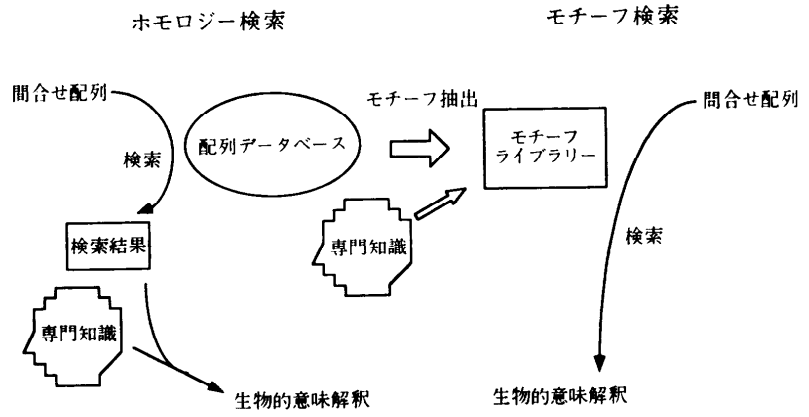

図-1 配列の全体的な類似性(ホモロジー)および機能的に重要な局所保存配列パターン(モチーフ)の検索

データの生物的意味解釈に新しい方法を確立することであった．ゲノムの塩基配列から遺伝子(タンパク質とRNA)のコード領域を予測し，その機能を予測することができなければ，ゲノム解析の生物的意義は生まれない．配列解釈は計算機科学としてもチャレンジングな問題であり，実際，並列処理や知識処理に代表される新しい計算機科学の分野と，分子生物学の技術革新により大量情報の時代を迎えた生物科学の分野の間には，学問としての自然な共通性と発展性があると考えられる[3]．生物学では物理学と異なり，第一原理に基づく ab initio 法はほとんど無力である．それは原理的なことがまだよくわからず，ようやく生命現象の最も基本となるゲノムのデータを観測する技術ができた生物学の状況を反映している．したがって，機能予測の基本は経験的な知識を利用することで，ここに計算機科学における知識処理との接点がある．経験的知識を単に実用面で利用するだけでなく，物理学や化学のような意味で生物学の学問の体系化を行っていくことが，「ゲノム情報学」の長期的な目標である．

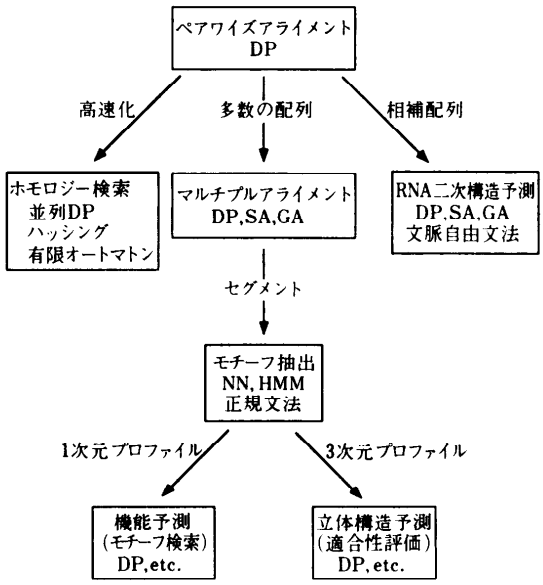

図-2 配列解釈と計算機科学の方法

配列解釈のための経験的知識とは「配列が類似なら機能も類似である」ことである．機能予測でとくに重要なのはアミノ酸配列であるので，ここではタンパク質に限って議論することにすると，その類似性には大きく2つのタイプがある．1つは配列全体の類似性(ホモロジー)で，これは共通の祖先からの分子進化を反映したものと考えられる．もう1つはタンパク質分子が他分子と相互作用する機能部位を反映した配列の局所的な類似性(モチーフ)で，共通の祖先から由来するもの(分岐進化)だけでなく，機能的制約から類似性をもつようになった場合(収束進化)も考えられる．図-1に示したように，前者の類似性を調べるのがホモロジー検索，後者の類似性を調べるのがモチーフ検索である．ホモロジー検索とは問合せ配列とデータベースにあるすべての既知配列とを1つ1つ並べて比較し，もし類似なものがあれば既知配列での例をもとに機能解釈を行うものである．必ずしも類似配列がみつからないことや，みつかっても既知配列で機能が同定されていないことが多いのが難点である．一方のモチーフ検索は，問合せ配列中にモチーフライブラリーに定義された配列パターンが存在するかを検索する．そもそもモチーフとは機能的意味づけがなされた配列パターンであるので，機能解釈は容易であり，ホモロジー検索では検出できない弱いパターンも調べることができる．しかしながら，モチーフを抽出し定義することが必ずしも容易ではないこと，すべての機能部位がモチーフで表現できるわけではないことから，普通はホモロジー検索で明確な答えが出ない場合にモチーフ検索を行う．

図-2はホモロジー，モチーフ，その他配列解釈に関連するさまざまな問題と，それに対する計算機科学の方法をまとめたものである．類似なものを検索することは，あらかじめ定めれらた類似性の尺度で最もよいものを探すことであり，これは計算機科学の立場では最適化問題となる．一方，モチーフについては検索の問題よりも，モチーフをいかに発見するかが重要で，それは機械学習の問題となる．ダイナミックプログラミング(DP)，シミュレーテッドアニーリング(SA)，遺伝的アルゴリズム(GA)，ニューラルネット(NN)，隠れマルコフモデル(HMM)，形式文法を始めとして，さまざまなアルゴリズムが適用されてきたが，

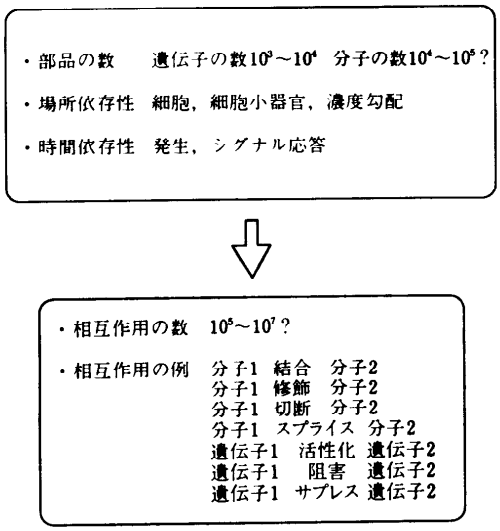

図-3 部品の解析から部品間相互作用の解析へ

これについてはすでに本誌など[4),5)]で解説しているのでここでは省略する．重点領域研究「ゲノム情報」ではマルチプルアライメント，RNA 二次構造予測，モチーフの抽出と検索を中心に，アルゴリズム開発と実用化を行った．ホモロジー検索とモチーフ検索は配列解釈ツールとして，ゲノムネットデータベースサービスで提供されている．なお，重点領域研究「ゲノム情報」には 5 年間で 49 名の研究者が計画研究または公募研究に参加したが，このうち 20 名以上は計算機科学専門の研究者であった．

4. 部品の解析からシステムの解析へ

既存の DNA ・タンパク質データベースは，基本的に分子(または配列)を単位として情報が蓄積されており，その属性の 1 つとして機能に関する注釈が与えられている．たとえば，表-1 の塩基配列データベースやアミノ酸配列データベースではフィーチャテーブルに配列の機能的意味が書かれており，配列モチーフライブラリーでは当然ながら配列パターンと機能の関連が示されている．しかしながら，個々の分子の働きがわかっても，細胞あるいは生物個体の働きがわかるわけではない．個々の分子とはいわば生命系を構成する部品であり，部品間のつながり，つまり結線図のようなものが解明されなければ，生命系システムを理解することはできない．このような認識から，1996 年度より始まった新しい 5 カ年計画である重点領域研究「ゲノムサイエンス」では，遺伝子と分子がいつどこでどのように相互作用しているかを，情報の流れのパスウェイ(経路)としてコンピュータ化することを大きな目標として掲げている．

コンピュータ化の単位となるのは，たとえばシグナル伝達における分子同士の相互作用，あるいは発生における遺伝子同士の制御関係のように，分子と分子あるいは遺伝子と遺伝子の相互作用データである．そして，相互作用の最も単純な形はペア間の相互作用，すなわち 2 項関係である．生命系を構成する部品の数として，バクテリアの遺伝子数は数千，ヒトの遺伝子数は十万近くで，低分子や代謝中間物および生成物を加えても生体内の分子数は 10^5 のオーダではないだろうか．図-3 に示したように，これら分子は時間および空間に依存して相互作用をし，膨大な数の相互作用データが生命系を構成しているのだろう．

さて，ある生物がもつ遺伝子のカタログをデータベースのようなものだと思うと，前述の DBGET を拡張して

生物種名：遺伝子名

で個々の遺伝子(または遺伝子産物)が指定できる．そこで遺伝子間または分子間の関係は，前述の LinkDB と同様に

生物種名：遺伝子名 1
→生物種名：遺伝子名 2

といった 2 項関係で記述できる．生物学における既知のパスウェイとは専門家がこのようなデータの集合から作りあげたものであり，そのプロセスをコンピュータ化することが大量データに対処する唯一の手段であると考えている．

筆者のグループは KEGG(Kyoto Encyclopedia of Genes and Genomes)と呼ぶプロジェクトにおいて，パスウェイデータをゲノムの機能予測に用いることを目指している．現在は代謝系を中心に，パスウェイデータのコンピュータ化と，各生物種の遺伝子産物とパスウェイ上の位置との対応をコンピュータ化しており，すでにゲノムネットデータベースサービスの一部として公開している．これまでのようなホモロジー検索による部品ごとの機能予測では，実験で検証されない限り，予測がどの程度正しいものであるか客観的な基準が存在

しなかった．パスウェイデータがコンピュータ化されると，部品の機能予測の結果をパスウェイ上にマップすることにより，たとえばトリプトファン合成系が再構成できるかといった観点で，すなわち生命系の機能的な単位が存在するかどうかで予測結果を判定することができる．もし，トリプトファン合成系の1カ所だけ部品が足りない場合には，2つの可能性がある．1つは本当はゲノムに遺伝子があるのに見落としている場合で，遺伝子探索をやり直す必要がある．もう1つは部品が欠けていることに生物的な意味があって，これまでに知られていない反応経路が使われている場合で，これはその生物種がもつすべての酵素の基質－生成物の2項関係を演繹することにより探索することができる．その際，化合物の階層分類により質問緩和を行うことも必要である．表-2にはパスウェイに関連した生物学の問題を掲げておいた．これらに対する新しい情報処理技術を開発することも，重点領域研究「ゲノムサイエンス」の大きな目標である．

表-2 パスウェイに関連した問題

パス計算(2項関係の演繹)
始点と終点を与えた場合，与えない場合
さまざまな階層分類に基づく質問緩和
パス計算結果の処理と表示
パスウェイマップの自動生成
パスウェイ比較
パスウェイの全体的・局所的アライメント
複数パスウェイのコンセンサス
パスウェイ検索
パスウェイ解析
遺伝子重複とパスウェイ形成の関連解析
遺伝子発現制御とパスウェイの関連解析
生物種，環境条件とパスウェイの関連解析

5. ゲノム情報学の体系化に向けて

タンパク質の立体構造予測の問題は，自然界に存在するすべての立体構造パターンのグループが明らかになれば，どのグループに帰属するか判定するだけのことになる．同様に機能予測の問題も，生命系を構成するすべての遺伝子と分子のパスウェイが明らかになれば，パスウェイ上のどこに帰属するかを判定するだけの問題となる．立体構造予測も機能予測も，少なくとも実用的な観点からは近い将来に可能となるであろう．しかしながら，経験的知識で機能予測を可能にすることと，その背後にある自然界の法則を解明することとは，必ずしも一致しない．

図-4はゲノム情報に関連した情報科学の分野を主観的にまとめてみたものである．これまで情報科学の基礎は数学や物理学と関連して発展してきた．新しい生物学の展開を担う若手研究者を育成していくためには，情報処理技術の応用を学ぶだけでなく，その背景にある基礎的な考え方を学ぶ教育が必要である．図にあるデータベース基礎，ソフトウェア基礎，知識処理基礎は，現状でのゲノム情報学の理論的基礎として重要である．一方，図の右の方は情報科学自体が生物学と関連して発展していく可能性を秘めた部分である．生命現象のシステム的な側面を解明していく中で，新しい

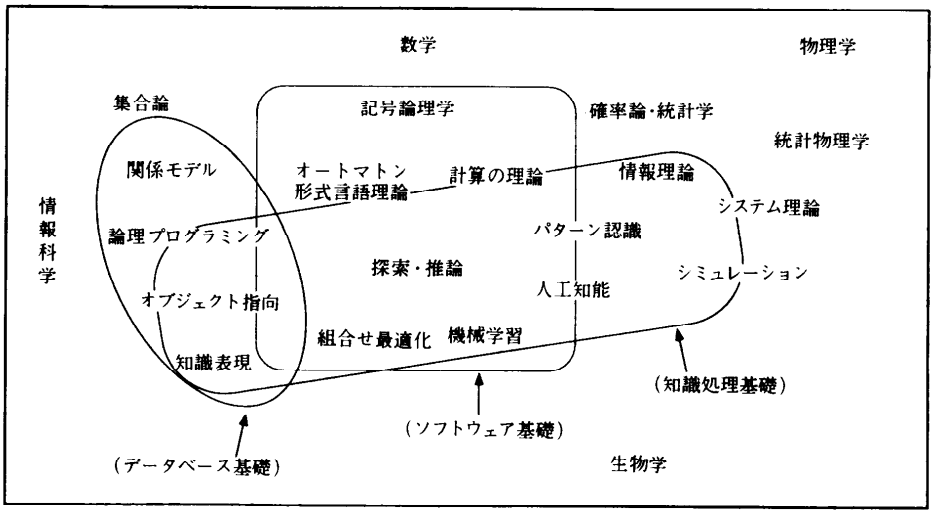

図-4 ゲノム情報学と他の学問領域の関連

情報科学の発展があるのではないだろうか.

参考文献

1)高木利久，金久 實：ゲノムネットのデータベース利用法，200 p，共立出版，東京 (1996).
2)高木利久：演繹データベースのゲノム情報処理への応用，人工知能学会誌, Vol. 10, pp. 17-23 (1995).
3)金久 實，新田克己，小長谷明彦，田中秀俊：知識情報処理技術とヒトゲノム計画，人工知能学会誌, Vol. 6, pp. 630-640 (1991).
4)金久 實：遺伝子とゲノムの情報処理，情報処理 Vol. 35, No. 11, pp. 983-990 (1994).
5)金久 實：ゲノム情報への招待，160 p，共立出版，東京 (1996).

(平成8年7月17日受付)

金久 實(正会員)

1970年東京大学理学部物理学科卒業，1975年同大学院博士課程修了，理学博士．1976 - 79年米国ジョンズ・ホプキンス大学研究員，1979 - 1982年米国ロスアラモス国立研究所研究員，1982 - 85年米国国立衛生研究所(NIH)研究員．1985年京都大学化学研究所助教授，1987年同教授，1991 - 1995年東京大学医科学研究所ヒトゲノム解析センター教授を併任．専門は理論分子生物学，とくに，タンパク質の機能予測と，遺伝子・分子・細胞の情報伝達パスウェイの解析．物理学会，化学会，生物物理学会，生化学会，分子生物学会，人工知能学会各会員．