2．特集：バイオインフォマティクスの動向

ライフサイエンス・医療ユニット　　庄司　真理子、茂木　伸一

2.1 はじめに

2001 年 2 月に国際解析チームとセレラ社からそれぞれヒトゲノム配列の概要が報告され、60 以上の生物種のゲノム配列が決定されてきていることから、遺伝子発現解析、タンパク質の構造決定、プロテオーム解析、分子間相互作用解析等をはじめとする、いわゆるポストゲノムシークエンス研究が本格化してきた。これらの研究において、膨大で多種多様な生物情報を効率よく整理・解析し、その生物学的・医学的意味を明らかにすることが必要であり、バイオインフォマティクス(Bioinformatics)が必要不可欠となっている。

本稿では、ヒトゲノム研究を中心とするバイオインフォマティクスの概要を説明し、この分野の課題を述べる。

2.2 バイオインフォマティクスの定義

バイオインフォマティクスは、生命科学に情報科学的な視点や概念を導入した研究分野であるが、近年、この用語の意味する範囲は広がってきている。

東京大学医科学研究所の高木利久教授は、バイオインフォマティクスを以下のように定義づけている。

> 調べるべき、あるいは、実験で確かめるべき膨大な探索空間(例えば、遺伝子やタンパク質の数、あるいは、それらの相互作用の組合せ、など)を狭めてくれる情報技術およびそのための基礎理論

米国では、国立衛生研究所(NIH)の生物医学情報科学技術イニシアチブ(BISTI)コンソーシアムが、バイオインフォマティクスとコンピュテーショナルバイオロジー(Computational Biology)という用語を使い分けて定義している。

> バイオインフォマティクス:
> 生物学、医学、行動学、健康に関するデータの取得、蓄積、体系化、データベース化(archive)、解析及び可視化を含めた展開のためのコンピュータツール及びアプローチの研究、開発または応用
> コンピュテーショナルバイオロジー:
> 生物学、行動学及び社会システムの研究に関するデータ解析手法、理論的手法、数学的モデリング技術及びコンピューターシミュレーション技術の開発及び応用

両者とも本質的な概念は同様であり、NIH では当面の課題を具体的に表しているものと言える。本稿では、NIH の定義するバイオインフォマティクスの事項を中心に述べる。

2.3 バイオインフォマティクスの位置づけと分類

2.3.1 バイオインフォマティクスの位置づけ

バイオインフォマティクスは、主にゲノム解析を活用したライフサイエンス分野の研究を支える基盤技術の一つとして重要な役割を担っている(図表 1)。

2.3.2 ポストゲノム研究におけるバイオインフォマティクスの分類

図表 2 には、ポストゲノム研究の方向性と、それに対応するバイオインフォマティクスのデータベースおよびデータ解析手法を示した。ポストゲノム研究は、「構造の世界」の研究から「関係性の世界」の研究へ移行してきており、さらに、生命のシステム的な理解を目指して、細胞や個体の機能(「機能の世界」)の研究を志向している。

「構造の世界」には、DNA 塩基配列やタンパク質立体構造などの解析が分類される。この解析で一般的に用いられるデータ解析手法の一つに、ホモロジー(相同性)検索がある。これは、配列を比較し、そのホ

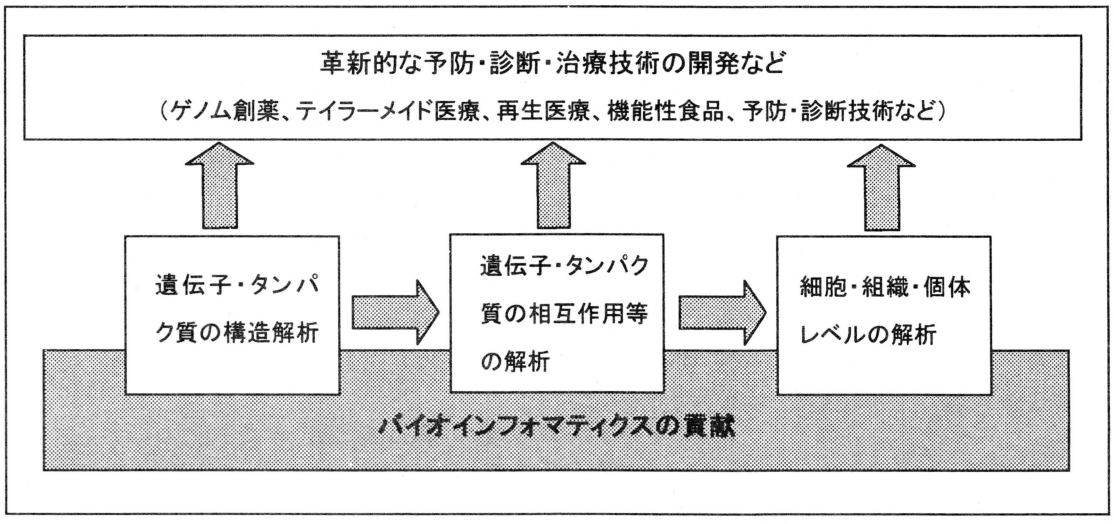
モロジーから遺伝子の構造や機能などの知識を抽出する方法である。この他にも、遺伝子配列に観察される統計的な特徴から遺伝子の位置を推定する手法などが開発されてきている。

「関係性の世界」には、遺伝子のスイッチのオン・オフを見る遺伝子発現解析や、タンパク質間などの相互作用を見る分子間相互作用解析などが分類される。ここでは、種々の条件下における遺伝子発現情報による遺伝子分類や、アミノ酸配列から予測されるタンパク質の物理化学性などを基に細胞内でのタンパク質の挙動を予測する細胞内局在位置予測といった解析が行われている。

「機能の世界」には、細胞間でのシグナル伝達や個体発生のメカニズムなどの解析が分類される。ここでの有効なデータ解析手法は、ほんの一部を除いてまだ開発されていない。

現在は、「構造の世界」の研究開発に関しては、重点的に公的資金が投入され、整備が進められている。しかし、「関係性の世界」や「機能の世界」、とくに「機能の世界」におけるデータベース整備や、データ解析手法への大規模な取組はまだほとんど手つかずの段階であり、今後の戦略的な推進方策が必要である。

図表 1　ライフサイエンス分野におけるバイオインフォマティクスの位置づけ



（科学技術動向研究センター作成）

図表 2　ポストゲノム研究におけるバイオインフォマティクスの分類

	研究概要	データベース	データ解析手法
構造の世界	配列 立体構造	DNA 塩基配列、遺伝子多型、タンパク質アミノ酸配列、タンパク質立体構造　など	ホモロジー（相同性）検索、遺伝子発見、モチーフ抽出、タンパク質立体構造予測　など
関係性の世界	発現 局在 相互作用	遺伝子発現情報、分子間相互作用、プロテオーム　など	細胞内局在位置予測、分子間相互作用予測、遺伝子発現クラスタリング　など
機能の世界	細胞機能 個体機能	シグナル伝達、発生・生理機能、免疫機能、脳機能　など	パスウェイ比較、計算機シミュレーション　など

研究開発が進行中

今後の戦略的な推進が必要

（東京大学医科学研究所高木利久教授の資料をもとに科学技術動向研究センターで作成）

2.4 バイオインフォマティクスの現状

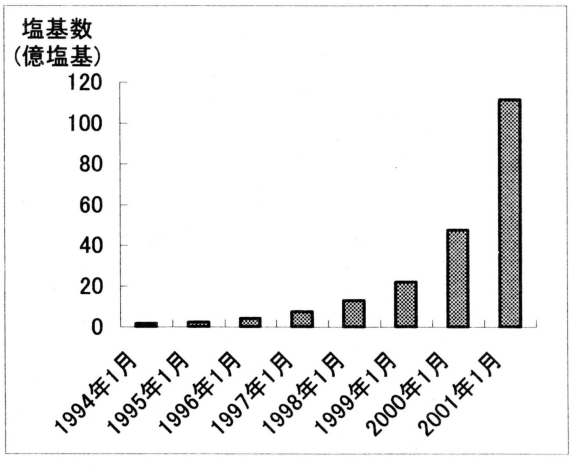
2.4.1 データベース

(1)データ量の増加

解読された DNA 塩基配列などは、公共データベースである GenBank(米)、EMBL(欧)、DDBJ(日本 DNA データバンク:DNA Data Bank of Japan)のいずれかに登録することになっている。三者のデータベースは相互にデータのやり取りを行っているため、データ内容はほぼ同じになっている。近年、mRNA の部分配列(EST)が多数登録されたことや、各種生物の DNA 塩基配列が次々と明らかにされたことを反映して、ここ数年のデータ量の増加は著しく、2001 年 1 月には約 111 億塩基、2001 年 10 月には約 141 億塩基の配列が登録されている(図表 3)。

データ構築の際には、単に塩基の文字列を羅列するだけでなく、遺伝子領域の位置やその構造や機能、またその遺伝子に関連する文献など、解析の結果分かった事項の注釈づけ(アノテーション)を行い、それらの情報も併せてデータベースに収録している。今後は、ここに示した配列データ量が増加することに加えて、配列データのアノテーションを質、量ともに充実させることが重要である。

図表 3 DDBJ に登録されている塩基数の推移



(DDBJ のデータをもとに科学技術動向研究センターで作成)

また、様々な条件の組合せによる遺伝子発現や分子間相互作用のデータなどが加わることにより、データは複雑化してくる。従って、情報システムのハード面では、常に高速化やディスク容量の増加が必要になる。また、研究者によるアノテーションの更新により、高品質なデータを蓄積していくことも重要な課題である。

(2)データベースの例

図表 4 には、ゲノム研究等で利用される代表的なデータベースを示した。一般に、タンパク質に比べて DNA の方が精製しやすく、DNA シークエンサーを用いることにより配列を比較的容易に決定できることから、DNA 塩基配列のデータベースが最も規模の大きいものになっている。DNA 塩基配列の公共データベースには、(1)で述べたデータベースのほか、ヒト遺伝子の DNA 塩基配列や一塩基多型(SNPs)のデータベースなどがある。

また、タンパク質のアミノ酸配列、機能を予測するために有用なモチーフ配列、立体構造のデータベースなどがある。このようなデータベースは世界に約 400 種類あると言われている。なお、Nucleic Acids Research 誌の毎年 1 月 1 日号が、各種データベースの特集号になっている。

また、我が国に関して言えば、パスウェイデータベースなどいくつかのデータベースを除いて、外国からも頻繁にアクセスされるデータベースは数少ないと言われている。ポストゲノム研究において我が国が主導権をとっていくためには、ある程度の規模と独自の内容をもつデータベースを構築して、我が国から世界に情報を発信していくことが必要であろう。

図表 4 データベースの例

データベースの主な内容	データベース名称（運用国）	データベースの主な内容	データベース名称（運用国）
DNA 塩基配列	GenBank（米）、EMBL（欧）、DDBJ（日）	アミノ酸配列	SWISS-PROT（欧）、PIR（米）
ヒト遺伝子の DNA 塩基配列	UniGene（米）	アミノ酸配列ドメイン	Pfam（欧）
一塩基多型	dbSNP（米）、JSNP（日）	アミノ酸配列モチーフ	PROSITE（欧）、BLOCKS（米）
遺伝病	OMIM（米）、Mutation Database（欧）	タンパク質立体構造	PDB（米）、SCOP（欧）、CATH（欧）
総合的なヒトの配列情報	HGREP（日）、Ensembl（欧）	パスウェイ	KEGG/PATHWAY（日）
ヒトの総合情報	LocusLink/Refseq（米）、GDB（カナダ）	文献	MEDLINE（米）

（科学技術動向研究センター作成）

(3)データベースの統合化

バイオインフォマティクスを用いた解析では、一種類のデータベースだけを用いることは少なく、様々なデータベースや検索システ

ムを組み合わせて行われる。そのため、各種データベースや検索ソフトウェアをWeb上でリンク付けさせることによって統合化したシステムが構築されてきている。代表的なデータベース総合検索システムの例を図表5に示した。

データベースの統合化に関する課題としては、データベースのデータ表示方式や検索条件の記述方式などの標準化や統一が図られていないことが挙げられる。そのため、研究対象ごとに特有な概念や用語を使うのではなく、統制された語彙や記述方法による体系的な理論（オントロジー）の構築の重要性も指摘されている。

例えば米国では、INCOGEN や Oracle など 40 以上のバイオ企業および情報系企業等から成る Interoperable Informatics Infrastructure Consortium(I3C)が発足し（2001 年 1 月）、ライフサイエンス分野におけるデータ交換・管理等における標準化を推進する活動に取り組み始めている。I3C では、XML によるデータ記述や、通信プロトコルを標準化することなどを検討している。

図表 5 データベース総合検索システムの例

データベース総合検索システム	サービス提供機関
DBGET（日）	京都大学化学研究所、東京大学医科学研究所
Entrez（米）	米国国立バイオテクノロジー情報センター（NCBI）
SRS（欧）	欧州分子生物学研究所（EMBL）

（科学技術動向研究センター作成）

2.4.2 ハードウェア

2.4.1.(1)で述べたように、バイオインフォマティクスでは急速なデータ量の増加などから、ハイパフォーマンスコンピュータ（スーパーコンピュータ）を基盤にした高度な情報システム構築の必要性が高まっている。

ゲノム解析に必要とされるコンピュータの性能は、図表 6 のように試算されている。また、生命系のような複雑な系を記述し、シミュレーションする技術などの展開も含めると、少なくとも数千台から1万台程度のプロ

セッサからなる並列機で、数十 Tera flops から数百 Tera flops が必要とされている(「ゲノム情報科学における我が国の戦略について」(2000 年 11 月、科学技術会議ライフサイエンス部会ゲノム科学委員会))。

一方、2001 年 11 月に発表された世界中のハイパフォーマンスコンピュータのランク付け TOP500 において第一位である ASCI white は、ピーク性能 12.3Tera flops である。

またバイオインフォマティクスでは、並列処理アルゴリズムや並列コンピュータの開発が進められているが、その一方で、ハイパフォーマンスコンピュータ上で動作するアプリケーションの並列化が難しいことなど、ソフト面の対応が後れていると言われている。(注:flops は 1 秒間に浮動小数点計算を何回行えるかという、計算機の演算性能指標の一つ。Tera＝10^{12})

図表 6　主なゲノム解析に必要とされるコンピュータ性能

ゲノム解析	性能(flops)
タンパク質ファミリー分類	1　Tera
系統発生図	10 Tera
シークエンスアセンブリ	10^2 Tera
シークエンス比較	＞10^2 Tera
遺伝子モデリング	10^5 Tera

(米国 DOE の科学シミュレーションイニシアチブ(SSI:Scientific Simulation Initiative)レポート「Advanced Computational Structural Genomics」の 2000 年のデータをもとに科学技術動向研究センターで作成)

2.4.3 データ解析手法とソフトウェア

データベースから目的とする知識を発見するデータ解析手法とそれを実行するソフトウェアの例を図表 7 に示した。

図表 7　主なデータ解析手法とソフトウェア

解析手法	解析の概要とソフトウェアの例
ホモロジー検索	配列を比較し、そのホモロジー(相同性)から知識を抽出する方法で、最も一般的に行われている解析方法。BLAST、FASTA、Smith-Waterman などのほか、弱い相同性をも抽出する感度の高いプログラムである PSI-BLAST や SAMT99 などがある。
遺伝子発見	遺伝子配列に観察される統計的な特徴に着目した推定方法。ホモロジー検索では発見できない未知の遺伝子配列を抽出する方法で、GENSCAN や DIGIT などがある。
モチーフ抽出	DNA 結合部位や酵素活性中心などの機能部位がもつ特徴的な短い配列(モチーフ)を見つけだす方法。アミノ酸配列のモチーフを集めた PROSITE などのデータベースに対して配列を比較する手法などがとられている。
細胞内局在位置予測	アミノ酸の疎水性指標や電荷などの物理化学性や局在化シグナルなどの配列を基に、タンパク質が生合成された後、細胞内のどこへ行くかを予測する。シグナル配列の位置を予測する SignalP、局在位置を予測する PSORT、膜貫通領域を予測する SOSUI、TMHMM などがある。
タンパク質立体構造予測	配列の相同性からタンパク質の立体構造を予測するホモロジーモデリングとして、Swiss-Model、MODELLER などがある。また、立体構造の比較から類似性をみる DALI、MODBASE などがある。

(実験医学「ゲノム医科学と基礎からのバイオインフォマティクス(Vol.19 No.11(増刊)P.61～66、P.73～81)」を参照し、科学技術動向研究センターで作成)

2.4.4 バイオインフォマティクスの産業化

近年、先進国を中心にバイオベンチャー企業の活動が盛んであり、2000年には、我が国に約160社、米国に約 1,300 社、欧州に約 700 社のバイオベンチャー企業が存在する。

バイオインフォマティクス分野では、既存のツールやデータベースを利用者向けにカスタマイズしたデータベース、ソフトウェア等を提供するバイオベンチャー企業などが活躍している。ゲノム情報を用いた実際の研究開発では、公表されている公共データベースやソフトウェアだけでは不十分であり、高度なセキュリティ、解析スピードの高速化、研究内容に即したデータ管理システムや検索支援システムなどが要求されるためである。欧米の主なバイオインフォマティクス関連のバイオベンチャー企業を図表 8 に示した。

図表 8 欧米の主なバイオインフォマティクス関連ベンチャー企業

会社名	主な事業内容
Incyte Genomics(米)	cDNA データベース、遺伝子発現(DNA チップ)など
Human Genome Sciences(米)	分泌タンパク質、膜タンパク質データベース、遺伝子薬開発など
Celera Genomics(米)	ゲノムデータベース(ヒト、マウス、ショウジョウバエ)、プロテオーム、SNP など
Gene Logic(米)	遺伝子発現データベースなど
CuraGen(米)	遺伝子発現データベース、SNP データベースなど
Genset(仏)	SNP データベースを用いた疾患解析など
deCODE genetics (アイスランド)	臨床データベース、家系データベース、多型性データベースなど

(「ゲノム医学の最先端と世界のバイオベンチャー『ゲノム情報系ベンチャー(p.96～102)』羊土社(2001 年)」をもとに科学技術動向研究センターで作成)

2.5 バイオインフォマティクス推進への取組

2.5.1 政策的な取組

近年、バイオインフォマティクスは、ゲノム研究の推進を図るために必要不可欠なものとして、世界各国で政策の重点化が図られている。我が国及び米国におけるバイオインフォマティクスの主な推進拠点を図表 9 に示した。そのほかにも、大学を拠点とした取組が活発に行われてきている。

我が国では、第 2 期科学技術基本計画(2001 年 3 月、閣議決定)において、ライフサイエンス分野で重点的・戦略的に取り組む課題の一つとしてバイオインフォマティクスが挙げられている。「平成 13 年度科学技術の振興に関する重点指針」(2000 年 6 月、科学技術会議政策委員会)でも、バイオインフォマティクスは重点化項目の一つとして挙げられ、2001 年度予算では 104 億円が計上された。2001 年度より、文部科学省では科学技術振興事業団にバイオインフォマティクス推進センター(東京都千代田区)を、経済産業省では生物情報解析研究センター(東京都江東区)及び生命情報科学研究センター(東京都江東区)を新たに発足させている。また、87 社の民間企業が参画しているバイオ産業情報化コンソーシアム(JBiC)では、産学官連携のもと、バイオインフォマティクスの研究開発が進められている。

図表 9　日米におけるバイオインフォマティクス推進のための主な拠点

日 本	➢ 文部科学省 ┬ 国立遺伝学研究所生命情報・DDBJ 研究センター ├ 東京大学医科学研究所ヒトゲノム解析センター ├ 京都大学化学研究所バイオインフォマティクスセンター ├ 理化学研究所ゲノム科学総合研究センター └ 科学技術振興事業団バイオインフォマティクス推進センター ➢ 経済産業省 ┬ 産業技術総合研究所・生命情報科学研究センター(CBRC) └ 産業技術総合研究所・生物情報解析研究センター(JBIRC) バイオ産業情報化コンソーシアム (JBiC)　etc.
米 国	➢ 健康福祉省(DHHS) 国立衛生研究所(NIH) ├ 国立医学図書館(NLM) └ 国立バイオテクノロジー情報センター(NCBI) ├ 国立バイオメディカルイメージング・バイオエンジニアリング研究所(NIBIB) ├ 国立一般医学研究所(NIGMS) └ バイオインフォマティクス・コンピュテーショナルバイオロジーセンター(CBCB) ├ 国立ヒトゲノム研究センター(NHGRI) └ 生物医学情報科学技術イニシアチブ(BISTI)コンソーシアム ➢ エネルギー省(DOE) 生物・環境研究局(BER) ➢ 全米科学財団(NSF)　etc.

（科学技術動向研究センター作成）

米国では、国立衛生研究所(NIH)を中心に早期よりバイオインフォマティクスの推進が活発に行われている。米国のバイオインフォマティクスの中核を担うNCBIをもつNLMの2001年度予算は約2億4,600万ドル(約290億円)である。また、2001年にはバイオインフォマティクス研究を助成する機関として、バイオインフォマティクス・コンピュテーショナルバイオロジーセンター(CBCB)が新設され、初年度は約1,000万ドル(約12億円)を大学などに助成することとしている。そのほか、エネルギー省(DOE)、全米科学財団(NSF)、国防総省高等研究計画局(DARPA)などにおいてもバイオインフォマティクス分野の研究推進が取り組まれている。

欧州では、欧州分子生物研究所(EMBL)及びその中の機関の一つである欧州バイオインフォマティクス研究所(EBI)、さらにサンガーセンター(英国)などを中心に推進が図られている。2000年の欧州における公的なバイオインフォマティクス予算は、1億ユーロ(約100億円)であり、その内EBIには1,000万ユーロ(約10億円)が投資されている。

2.5.2 人材育成への取組

近年のライフサイエンス分野の研究では、保有するバイオインフォマティクスツールの品質・性能や、それを使いこなすテクニックが、研究開発の進展に大きく関係する。そのため、ツールを使いこなせる人材や、より優れたアルゴリズムやソフトウェアを開発できる人材等の需要が急速に高まっている。しかし、この分野は新興分野であることから、十分な人材が確保できておらず、各国とも人材育成の対策が重要課題となっている。

「ゲノム情報科学における我が国の戦略について」(2000年11月、科学技術会議ライフサイエンス部会ゲノム科学委員会)では、当面の人材育成について必要とされる事項を以下のようにまとめている。

① 即戦力の養成:既存の人材を活用するための研修・訓練プログラムの開発、訓練の機会及びインセンティブの提供
② 中長期的な人材の育成:大学院の専攻や学部の学科の整備等、研究交流・試行錯誤の「場」としての拠点形成

①に関しては、例えば、2001 年度より科学技術振興事業団のバイオインフォマティクス推進センターでは、東京大学医科学研究所と共同で、実践的な研修プログラムである「ゲノムリテラシー講座」に取り組み始めた。ここでは、研究者を対象に、データベースや解析ソフトウェアの利用法を習得するためのプログラムを開催している。

また②に関しては、2001 年度の科学技術振興調整費により、バイオインフォマティクスの人材育成事業が始められ、4 つのテーマが採択された。これにより、例えば慶應義塾大学では、理工学部に生命情報学科を新設する「システム生物学者育成プログラム」を開始した。本プログラムでは、学部教育で、化学、物理・情報、数学を基礎におき、生物を理解する方法としての実験と計算機科学を修得させる。そして、製薬会社やコンピュータ会社、コンサルティング会社などの産業界へ、年間 40 名程度の学部卒業生、25 名程度の修士学位取得者を供給することなどを目標としている。

米国においても人材育成は重要課題として取り組まれている。NIH の研究機関において、人材育成への取組を行っている主な機関には、フォガティ国際研究所(FII)、国立がん研究所(NCI)、国立老化研究所(NIA)、国立一般医学研究所(NIGMI)、国立医学図書館(NLM)、国立ヒトゲノム研究センター(NHGRI)などがある。その中で、例えば NLM では、エール大学やコロンビア大学など 12 の人材育成プログラムに助成金を出資している。

2.6 おわりに

バイオインフォマティクスは、ライフサイエンス分野の研究を支える基盤技術であると同時に、異分野融合型の生命科学を担う新しい領域として捉えられる。

現在、DNA 塩基配列やタンパク質立体構造を解析するためのデータベースやデータ解析手法の整備が進行中である。今後は、細胞機能や個体機能の研究に求められる、データベースやデータ解析手法の研究開発を戦略的に推進していく必要がある。

データベースは全ての解析のもととなるため、データには常に高い品質が求められる。アノテーションなどのデータ更新等、適切なデータベース管理を維持していくことが重要である。また、効率的なデータベース検索やデータ解析には、データベースの統合化などの取組も必要である。

ハード面については、現在の最高性能のハイパフォーマンスコンピュータでも、種々のゲノム解析に必要とされる性能には足りない状況であり、さらに性能を向上させることが望まれる。また、ハイパフォーマンスコンピュータ上で動作するアプリケーションの並列化などソフト面での対応も望まれる。

我が国では、バイオインフォマティクス推進のための拠点形成や人材育成のためのプログラムなどに関する取組はまだ始められたばかりである。今後の生命科学におけるバイオインフォマティクスの重要性を考慮すると、継続的な施策がより一層望まれる。

【謝辞】

本稿をまとめるにあたり、東京大学医科学研究所の高木利久教授には、ご指導いただくとともに、関連資料を提供していただきました。文末にはなりますが、ここに深甚な感謝の意を表します。