報告

大容量情報時代の次世代生物学

平成２６年（２０１４年）９月１７日

日　本　学　術　会　議

基礎生物学委員会・統合生物学委員会・農学委員会・

基礎医学委員会・薬学委員会・情報学委員会合同

バイオインフォマティクス分科会

この報告は、日本学術会議 基礎生物学委員会・統合生物学委員会・農学委員会・基礎医学委員会・薬学委員会・情報学委員会合同 バイオインフォマティクス分科会の審議結果を取りまとめ公表するものである。

日本学術会議 基礎生物学委員会・統合生物学委員会・農学委員会・基礎医学委員会・薬学委員会・情報学委員会合同 バイオインフォマティクス分科会

委員長	美宅　成樹	（連携会員）	公益財団法人豊田理化学研究所客員フェロー
副委員長	斎藤　成也	（第二部会員）	情報･システム研究機構国立遺伝学研究所集団遺伝研究部門教授
幹事	諏訪　牧子	（連携会員）	青山学院大学理工学部教授
	岡崎　康司	（連携会員）	埼玉医科大学ゲノム医学研究センターゲノム科学部門教授・所長
	金久　實	（連携会員）	京都大学化学研究所教授
	久原　哲	（連携会員）	九州大学農学研究院教授
	郷　通子	（連携会員）	情報・システム研究機構理事
	五條掘　孝	（連携会員）	情報･システム研究機構国立遺伝学研究所副所長･教授
	高木　利久	（連携会員）	東京大学大学院理学系研究科生物科学専攻教授
	宮野　悟	（連携会員）	東京大学医科学研究所ヒトゲノム解析センター教授

報告および参考資料の作成にあたり以下の方々に御協力いただきました。

有田　正規	情報・システム研究機構国立遺伝学研究所教授
岩崎　渉	東京大学大学院理学系研究科准教授
中村　周吾	東京大学大学院農学生命科学研究科准教授
由良　敬	お茶の水女子大学大学院人間文化創成科学研究科教授

本件の作成にあたっては、以下の職員が事務を担当した。

事務局	中澤　貴生	参事官（審議第一担当）
	渡邉　浩充	参事官（審議第一担当）付参事官補佐
	藤本紀代美	参事官（審議第一担当）付審議専門職

要　　旨

１　作成の背景

ゲノム解析を中心とした大容量生物系ビッグデータにより、生物学は新しいステージに踏み出しつつある。この時代にバイオインフォマティクスという学問分野をどう発展させていくべきかを議論し、バイオインフォマティクスを専門とする研究者に向けてまとめたものが、本報告である。次世代生物学においては、生物情報解析の新しい方向性を打ち出すことによって、バイオインフォマティクスをステップアップさせ、生物についての大きな未解決問題（オープンプロブレム）を解決することになるだろう。本報告では、まず多様な生物系ビッグデータを関係付ける情報解析の方向性について述べ、それによって解決される可能性がある生物学におけるオープンプロブレムを再定義する。さらに次世代生物学における新しい分野開拓、人材養成などについても述べる。

２　生物系ビッグデータの情報解析の方向性

次世代生物学は、生物系ビッグデータの解析によるオープンプロブレムの解明が鍵となる。そのための情報解析は、４つの方向性で特徴付けられる。

(1) ボトムアップ・アプローチとトップダウン・アプローチの統合

生物の理解には、要素還元的により詳細な情報を解析していくアプローチと、ゲノムなどの全情報の中に生物の調和を解析していく全体論的なアプローチの組合せが必要である。

(2) メカニズム解明のアプローチと統計的アプローチの統合

生物のシステムには、メカニズムに関わる側面と統計的な側面が本質的に絡み合っており、生物系ビッグデータはこの２つの側面を統合的に解析していく必要がある。

(3) 配列空間と実体空間の統合的情報処理

生物界は、設計図を書き込んでいる配列空間と生物体の実体空間が緊密に絡み合うことで、大いに繁栄してきた。配列空間と実体空間の統合的な情報処理が必要なのである。

(4) 生物の多様性と複雑さのシミュレーション

生物の非常に大きな多様性と複雑さを、問題のスケールに合わせてシミュレーションできるような情報解析が必要である。

３　生物学における未解決問題（オープンプロブレム）

科学の発展は、各学術分野における大きなオープンプロブレムの解決を軸に成し遂げられてきた。生物学でも従来からいくつかの大きなオープンプロブレムは存在していたが、それらを解決するだけの十分なデータがなく、生物全体を理解するための統一的な考え方も欠けていた。そのために、これまでの生物学では個別の課題に重点を置いた研究開発が行われてきた。大きな未解決問題は、個別課題の積み上げによって

いずれは解決されるかなり遠い目標と考えられてきたのである。しかし、最近生物系ビッグデータが急速に集積されるようになり、その中には生物についてのオープンプロブレムの解決につながるだけの十分な情報が含まれていると考えられる。そこで生物学のオープンプロブレムをビッグデータ時代に合わせて再定義し直した。

(1) 生物の起原、進化、多様性の問題

生物の大進化のメカニズムや多様性を駆動するメカニズムは､生物界におけるゲノムと環境の関係の問題であり､多様な生物のゲノム情報を解析することにより解明されるべきものである。

(2) 生体分子の構造形成と機能相関の問題

すでに多くのアミノ酸配列とタンパク質立体構造のデータが得られているが、配列―構造・機能の相関の問題はまだ解明されておらず、重要なオープンプロブレムとなっている。

(3) 生命システムの問題

セントラルドグマが、アミノ酸配列を設計する原理であるとすれば、生物の部品である多種多様なタンパク質を調和的に組み合わせるシステム形成の問題も重要なオープンプロブレムである。

(4) ゲノムと環境と疾患の関係

疾患のメカニズムの理解には、ヒトという生物種内でのゲノムと環境と疾患の関係を解明しなければならない。それには色々な変異の組合せと生物システムの調和の関係が重要な問題となる。

(5) 意識・精神活動の理解

最近は大量の脳・神経画像データが得られている。脳の仕組みのどこまでがゲノムで規定され、環境によってどのようにコンテンツが形成されるかは、ヒトを理解するための究極の課題である。

４　新しい分野開拓と人材養成

これまでの生物学では、縦割りの学術的体系の下で研究開発が行われてきた傾向が強い。これに対して、本質的に学問横断的に存在している生物系ビッグデータとそれを解析するバイオインフォマティクスの将来的なあり方について、分野の開拓、人材養成、社会的インパクトを考察した。

(1) 分野の開拓

生物系ビッグデータの時代に入り､細分化された従来の生物学の諸分野ではそれを扱いきれないという状況が生まれてきている。生物系ビッグデータ時代に合わせた学問開拓が求められており､バイオインフォマティクスのステップアップが必要である。それによって、生物学のオープンプロブレムを柱とし、研究拠点の設立や予算配分などを体系的に設計することが可能になるだろう。

(2) 人材養成

バイオインフォマティクス分野の人材不足が言われて久しい。そのことは生物系の

ビッグデータを必要としている学問分野が急速に拡大していることの反映だと考えられる。しかし、本質的には生物系ビッグデータを用いて生物全体にわたる原理的な仕組みを解明していくような人材、あるいはそのような志向を持って新しい学問分野を開こうとする若い生物学人材を養成していくことが必要である。そのとき新たに発生する倫理的問題も同時に意識できる人材であることが求められる。

(3) 社会的インパクト

地球に暮らす多様な生命全体の持続的な存続と、人々の幸福にとって、生命科学は様々な形で重要な関わりを持っている。そして、様々なオープンプロブレムは、生物に関連する課題をほぼ網羅するほど広範囲な分野をカバーしており、次世代生物学が社会に貢献できる内容はきわめて大きい。そして、この産業的・経済的価値や社会的インパクトの反面として、生物科学の最先端分野にあるからこそ人間としての謙虚さを忘れず、生命倫理の問題を常に意識し、十分注意を払っていかねばならない。

目　　次

１　作成の背景

ゲノム解析を中心とした大容量生物系ビッグデータにより、生物学は新しいステージに踏み出しつつある。この時代にバイオインフォマティクスという学問分野をどう発展させていくべきかを議論し、バイオインフォマティクスを専門とする研究者に向けてまとめたものが、本報告である。

バイオインフォマティクス(Bioinformatics)では、生物に関する全ての情報を扱っている。生物には階層性があり、生体高分子、細胞、生物個体、生態系、地球全体の生物界など、それぞれの階層における膨大な情報が存在している。また、生物は DNA 塩基配列を中心とした一次元の配列情報と、三次元の実体としての生物情報が緊密に関係している。そして、それらの情報はダイナミックに変化していて、時間軸でも分子のゆらぎの時間から進化の時間まで多様なダイナミクスがある。まさに多階層、異質、多次元の膨大な情報がバイオインフォマティクスの対象となっているのである。日本学術会議の 2010 年の報告「統合生物学分野の展望」[1]（以下、「2010 年報告」という）が指摘しているように、バイオインフォマティクスは生命科学（ライフサイエンス）における中心的・基盤的分野となり、データ解析・シミュレーション駆動型のプレディクティブ・サイエンスへと脱皮し、生物学は情報マネジメントの科学となると考えられる。実際 21 世紀に入り、ゲノム情報を中心とした大容量生物系ビッグデータが急速に解析されるようになってきた。しかし、生物系ビッグデータは、大規模かつ多階層・異質・多次元の情報であることから、それを統一的にデータベース化すること自体が困難であり、バイオインフォマティクスが 2010 年報告で述べているプレディクティブ・サイエンスへ脱皮するには、より深い考察に基づく高度な研究開発戦略が必要である。

バイオインフォマティクスについての未来的な研究戦略を考えるために、まず簡単に生物学の歴史を振り返ってみると、何回かのパラダイムシフトの波によって生物学が大きく発展してきたことが分かる。第１の波は、1900 年オランダの植物学者ド・フリーズらによる「メンデルの法則の再発見」などをきっかけとした波である、それによって遺伝子の実体を探索する研究が盛んになった。この時期の研究の進め方は、観察や観測によってデータを集め、それらを経験的に判断するようなものであった。第２の波は、1953 年ワトソンとクリックが遺伝子の実体である DNA の二重らせん構造を解明したことをきっかけとした波である[2][3][4]。これによって分子生物学やバイオテクノロジ―技術が生み出され、生物学は大きく進展した。そこでは生物を遺伝子、タンパク質などの分子の集合体としてとらえ、これらの個別分子の機能を見出し、応用してきた。そして第３の波は、2000 年に初めてヒトのゲノムドラフト配列が公開されたこと[5]から始まり、その後のゲノム、プロテオーム、トランスクリプトームおよびそれら全体の関係性であるシステム生物学が急速に発展してきた波である[6]。同時期にビッグデータの産業への応用が進められ、生物学も生物系ビッグデータの時代に入ったのである[7][8][9][10]。

以上のように歴史を振り返ってみると、今後さらに生物系ビッグデータの蓄積をきっかけとした生物学の第４の波が期待される。そして、そこではバイオインフォマティクスの果たすべき役割は非常に大きいことが予想される。そのためには生物系ビッグデータの持つ意味を多角的に考察し、新しい情報解析手法を研究開発していかねばならない。生産されつつあるデータ量の拡大に応じた情報解析のスケールアップは当然であり、これまで常にその努力が払われてきた。しかし、生物を全体として理解するという見地からみると、生物にとって本質的なユニットである生物個体や生態系全体に対する新しい情報解析の手法も求められる[1]。そこで第２節では、ビッグデータ時代のバイオインフォマティクスに必要な生物情報解析の方向性について述べる。さらにその先の大きな目標として、大容量の生物情報を用いて生物学における大きな未解決問題（オープンプロブレム）[11][12]に真正面から取り組み、解決していかねばならない。第３節では、従来はデータ不足や解析手法の不十分さのために、未解決として残されてきたオープンプロブレムをビッグデータ時代に合わせて再定義する。最後に第４節では、この新しい生物学を推進していくためのバイオインフォマティクスを中心とした分野開拓と人材養成、社会的インパクトについて述べる。それと同時に、大きなオープンプロブレムの解明に伴う社会的問題を常に意識し、注意を払っていかねばならないということを指摘する。

２　生物系ビッグデータの情報解析の方向性

生物学・生命科学は、生物がきわめて複雑かつ多様であるという事実と、学問としての歴史的経緯のために、スモールサイエンスの集合として発展してきた。そして、各個別研究がカバーする範囲が相対的に狭いため、生物系のビッグデータが得られるようになってきた現在も、生命全体の本質をとらえるような根源的で大きな課題が必ずしも提起されていない。そのことについて 2010 年報告では、より大きな観点からのバイオインフォマティクスの役割が大きいとして、次のように言及している[1]。「最近は、地球環境、エネルギー、食料、人口など現代的な諸問題を、生物の情報抜きに議論することができなくなっており、バイオインフォマティクスの対象および方法も大きく拡大されつつある。膨大な生物情報のデータベースを構築し、そこから有用な知識を発見し、生体高分子の構造・機能や生物システムの予測などを行うことが、今後の新しいバイオインフォマティクスの方向性であろう。」この節では、生物の全ての階層に共通の情報解析の考え方や手法について新しい方向性をまとめる。

(1) ボトムアップ・アプローチとトップダウン・アプローチの統合

一般に複雑な現象に対する科学的解析手法には、ボトムアップ（要素還元論的）アプローチとトップダウン（全体論的あるいは構成的）アプローチがある。生物学の場合、歴史的経緯からボトムアップの方向性による研究が最近の主流になっている。実際、タンパク質の機能のメカニズムも詳細な立体構造の情報から多くの議論ができ、研究開発におけるボトムアップの方向性が非常に有効であることは明らかである。しかし、生物のもう一つの側面として、生物は非常に多くのユニットが複雑に組み合わされたものでありながら、高度に調和が取れたシステムであるという事実がある。日本学術会議「日本の展望―生命科学からの提言」が指摘している通り、地球生命系の総合的理解のために、人類は未だ基礎情報の継続的収集とその統合的な理解の深化を必要としており、近代的インフォマティクスによって強化された自然史科学、分類学、生態学にみられるような網羅的・統合的生物学の視点も不可欠である[13]。現在蓄積されつつある生物系のビッグデータは、全体論的な解析に耐えるだけの量と質を含んでおり、トップダウンの方向性の解析が今こそ求められていると考えられる。

次世代の生物学は、生物のユニットをできるだけ詳細に解析していくボトムアップの解析と、生物の全ユニットの調和を理解していくトップダウンの解析[14]を統合するものでなければならないと考える。

(2) メカニズム解明のアプローチと統計的アプローチの統合

生物におけるプロセスでは、必然性と偶然性が緊密に絡み合っており、これが生物の現象を非常に複雑にしている。この問題を解決するために、色々な情報解析手法が用いられてきているが、それぞれ一長一短で生物系ビッグデータを統一的に解析することはまだできていない。例えば、配列の類似性検索は、配列が似ているタンパク質

の立体構造や機能が似ているという経験的ルールに基づいており、生物のプロセスについてのメカニズムに全く触れることなく、偶然性に隠されずに残っている重要な部分を浮き彫りにすることができる[15]。他方、類似性の低い配列領域にも立体構造を維持保存させるという重要な物理的メカニズムが内在している。生物における配列にはまず機能的に重要で保存性の高い部分があり、その解析には偶然性に重みをかけた統計的アプローチが有用であるが、保存性の比較的低い領域には立体構造を維持保存するような物理的特徴があり、その解析には物理的パラメータを用いたアプローチが求められるのである［14]。さらに、物理化学的なアプローチによる解析結果にもランダム性（揺らぎ）が顔を出すので、次世代生物学における生物系ビッグデータの解析ではこの２つのアプローチを組み合わせた新しい方法を開発していかねばならない。

(3) 配列空間と実体空間の統合的情報処理

生物界全体は、配列空間に書き込まれた設計図と、それに基づいて実現された実体空間の中の生物体によって構成されている。この２つの空間の関係は別の言い方をすると、遺伝子型と表現型の関係である。この長年の未解決問題を、生物系ビッグデータに対する統合的情報処理によって解決するには、２種類の空間の中で、現実の生物がどのような位置を占めているかということを理解する必要がある。この２種類の空間は、分子、細胞、個体、生態系など全てのレベルが関係しているが、最も分かりやすい分子レベルについてみてみると、現実のタンパク質のアミノ酸配列集団は巨大な全配列空間の中のきわめて小さなサブ空間に過ぎない。また、タンパク質立体構造のバラエティも、あるアミノ酸配列が取り得る全構造空間の中できわめて小さなサブ空間に過ぎないことが分かっている。

長い進化の歴史にあったと言われる環境の大激変に対して、生物は非常にロバストに生き延び、多様性を増やしてきた。このことは、巨大な配列空間と実体空間の中から生物に対応する非常に小さなサブ空間を効率よく選択する何らかの仕組みがあるということを強く示唆している。そのことが、タンパク質の立体構造形成のシミュレーションや進化上の配列変化のシミュレーションにどう反映するかを考えてみる。アミノ酸配列が折れ畳まれてタンパク質が立体構造を作る過程で、構造が完全にランダムに変化するとすれば、「組み合わせ爆発」の問題が発生し、計算機によるシミュレーションは事実上不可能になってしまう。また、進化における大量の変異が起こる過程を考えた場合、配列空間の中で完全にランダムな変異を仮定すると、配列についての「組み合わせ爆発」の問題がやはり発生する。しかし、この解決不能にみえる難問は、結局生物が利用している配列のサブ空間と、その結果としてできる実体空間中のサブ空間が、どのようなルールで形成されているかという問題に帰着する。このように問題を書き直すと、現在生産されつつある生物系ビッグデータを用いて解答可能な、バイオインフォマティクスの問題となる。具体的には、ゲノム中に実現している DNA 塩基配列がどのくらいランダムから外れているか、逆にどのようなタイプの秩序を含

んでいるかというルールを明らかにすることは、十分バイオインフォマティクスの範疇の問題である。そういう方針で大量のバクテリアゲノムにおけるDNA塩基配列の空間を解析してみると、実際に配列上の非常に偏ったサブ空間だけが利用され、そのサブ空間はバクテリアの生存環境によらないことがすでに分かっている［14］。

(4) 生物の多様性と複雑さのシミュレーション

生物は、言うまでもなく多様であり、複雑である。そうした多様かつ複雑な現象を再現する方法として、計算機によるシミュレーション手法が用いられる。生物学をプレディクティブ・サイエンスへとステップアップするためには、現在までも行われてきた分子構造のシミュレーションだけではなく、細胞、生物個体、生態系のシミュレーション、さらに進化プロセスのシミュレーションが必要である。このように生物の全ての階層を視野に入れて考えると、それぞれの問題のスケールに合わせたユニットを取り、シミュレーションを行うことは至極自然である。つまり、ユニットの中の微細構造については性質の平均化を行い、ユニット間の相互作用だけを考える粗視化の手法を導入することになる。

この時、全ゲノムという単位で問題を考えると、階層の最も小さな分子レベルにおけるタンパク質の構造や細胞内局在のシミュレーションと、最も大きな進化レベルのゲノム変化のシミュレーションの間に、共通点があることに注目しなければならない。そもそもゲノムは多くの変異が集積してできたものなので、進化レベルのシミュレーションでは導入される大量の変異によってゲノムが時間的にどう変化していくかを調べることになる。他方、タンパク質の構造や細胞内のシミュレーションでは、生物個体のゲノムが対象となり、それから得られるタンパク質の構造や細胞内局在の分布をシミュレーションすることになるので、進化という観点からみると、ある時点での構造や分布を切り出したものということになる。つまり、進化のシミュレーションはゲノム変化の時間ドメインのシミュレーションとなるのに対して、タンパク質の構造や分布のシミュレーションでは同じ対象を空間ドメインで調べることに相当すると考えられる。このように生物に関連するシミュレーションを行う時に、最小と最大の問題が同じ問題の別の側面であるということは、粗視化の仕方に関する非常に重要な情報を与えてくれる。

どのような情報をどのように粗視化するかという具体的な問題は、バイオインフォマティクスの将来に関わってくる。ただ示唆的なのは、第２節の(3)の配列空間と実体空間の統合的情報処理で述べた配列のサブ空間の中で DNA 塩基配列の変異シミュレーションを行い、疎水性などの物性の粗視化を行った膜タンパク質予測システムで評価してみると、膜タンパク質の割合が一定であるという現実のゲノムでもみられる事実がよく再現することである[14]。このことはタンパク質の構造についてのシミュレーションであると同時に、生物ゲノムの進化における大量の変異のシミュレーションともなっているのである。

３　生物学における未解決問題（オープンプロブレム）

生物系ビッグデータのインパクトはきわめて大きい。生物の分子レベルの問題から進化の問題までの基礎科学分野はもちろん、人類の健康に関する医学応用分野にも決定的なインパクトを与えると考えられるのである[16]。本節では、生物系ビッグデータを利用することによって解決される可能性がある未解決問題（オープンプロブレム）をまとめたいと考えているが、その前に、それらを解決するための土台となるビッグデータの産出状況と、それに対する情報解析のアプローチについて簡単に述べておかねばならない。

ビッグデータとは、現実的な時間内でデータ収集、管理などの処理を行う従来からのソフトウェアやツールの能力をはるかに超えたデータ集合体のことである。単に大容量というだけでなく、ボリューム（volume、データ量）、速度（velocity、入出力データの速度）、バラエティ（variety、データタイプとデータ源の範囲）の３ｖで特徴付けられる[7][8][9]。そのデータ量は年々増加の一途にあり、例えば、2013 年には Google の１日のデータトラフィック量は、すでに数ペタ（10^{15}）バイトをはるかに超え、数エクサ（10^{18}）バイト、さらには数ゼタ（10^{21}）バイトといった単位にまで向かっている[10]。そのためこれらビッグデータを許容時間内で処理するための新しい技術やソフトウェアツール群の開発が行われている。例えば、情報解析の観点からは、統計解析（相関性解析、分類、クラスタリング、回帰分析）、機械学習（ニューラルネット、ＳＶＭ，相関ルールの発見、教師あり学習と教師なし学習）、シミュレーション、遺伝的アルゴリズム、 自然言語処理、 パターン認識 、予測モデリング、時系列解析およびデータの可視化などや情報転送技術、クラウドネットワークなどがそれらに相当する[17]。

生物系のビッグデータでは、現在、ゲノム配列データ、パーソナルゲノムデータ、ゲノムコホートデータ、遺伝子発現データ、エピゲノムデータ、タンパク質立体構造データ、バイオイメージングデータなど、大規模（volume）かつ異質・多階層・多次元(variety)の情報が急速(velocity)に蓄積され続けている[18][19][20]。象徴的な例は新型シーケンサーの産出する膨大な配列データである。シーケンサー自体は、日進月歩で進歩しており[21]、ヒトゲノム計画は当初その処理に約 10 年かかったが、今では一週間も経たないうちに達成することができる。新型シーケンサーは、過去 10 年間でシーケンシングのコストを１万分の１に削減したのである。その解析技術の進展速度はコンピュータの性能に関するムーアの法則の 100 倍にも達したとも言われている[22]。

ここではいわゆるビッグデータと生物系ビッグデータを並列させて述べたが、それらは実際のところ全く異なる側面を持っている。生物系ビッグデータの場合、その主体である生物自体が一つの情報処理機械でもあるということである。生物という情報処理機械は、自らの持つ情報に基づいて自分を維持し、環境と相互作用し、次の世代を生み出す。例えば、生物ゲノムを入力とし、生物の構造や形質を出力と考えると、

生物体ではその間をつなげる素過程は膨大な生化学反応である。そして、バイオインフォマティクスでは同じ問題をコンピュータによる計算という素過程でつなぐことになる。この入出力関係のビッグデータを現在のバイオインフォマティクスは扱い切れていないのだが、生物自体は容易に処理して生きているようにみえる。生物学におけるオープンプロブレムの本質は、生物自身が容易に処理しているビッグデータを、我々は学問としてまだ十分理解できておらず、情報処理の技術も不十分だというところにある。すなわち、それ自体が情報処理機械である生物の原理を、ビッグデータから解明し、さらに応用していくことが、生物学の第４の波につながると考えられる。そこで次世代の生物学に向けて解決すべきオープンプロブレム[11][12]を整理し、今後の研究開発の方向性を示したい。

(1) 生物の起原、進化、多様性の問題

生物が最初にどのように誕生したかという生物の起原の問題、その後生物は大きく複雑化したが、そのプロセスに必然性があったかどうかという生物の大進化の問題、さらに共通のボディプランを持つ生物が何故大きく多様化できたかといういわゆる生物多様性の問題は、生物の進化に関わる大きなオープンプロブレムである[23]。この問題は、第２節（3）において述べた「組み合わせ爆発」という難問を生物がどう解消しているかということと深く関わっており、進化の方向を決めるのは環境変化であるが、その変化スピードや効率などについては生物の内在的な仕組みが関係していると考えられる。そして、環境変化の影響と内在的な仕組みの関係はよく検討しなければならないのである。

生物には、40 億年近い歴史がある。その間には環境の大激変があり、そのために生物は大絶滅の時期も多く経験してきている[24]。しかし、生命は一度も途切れることなく、全体的にみれば複雑化・多様化への道を一貫して進んできている。このことは生物のシステムは、どのような環境であっても複雑化・多様化を可能にできるような、非常にロバストな仕組み（原理）を内蔵していることを示唆している。

これに関連して、生物ゲノムを体系的に解析した ENCODE 計画[25]によれば、生物ゲノムの非コーディング領域（アミノ酸配列をコードしていない領域）も生命現象に重要な役割を果たしていることが分かってきた[26]。非コーディング領域は生物システムの制御領域として重要な役割を果たしているのである。進化の色々な段階にある生物のゲノムを比較してみると、実体空間における生物の複雑化の度合いと、ゲノム中の非コーディング領域の割合に、よい相関がみられる[27]。それとは別に、エピゲノムという概念の下で非コーディング領域の研究が現在盛んに行われている。そこでは生物ゲノムにおける遺伝子発現に対する環境の影響が議論されている[28]。

このオープンプロブレムについては、第２節の４つの情報解析における(1)のボトムアップ・アプローチとトップダウン・アプローチの統合、及び(3)の配列空間と実体空間の統合的情報処理を中心に解析の方向性を検討すれば、解決への道が開かれるだろう。

(2) 生体分子の構造形成と機能相関の問題

セントラルドグマは、DNA 塩基配列とタンパク質のアミノ酸配列をつなぐ生物共通の原理である。しかし、タンパク質のアミノ酸配列と立体構造および機能との相関は未解決の問題である。配列解析技術と立体構造解析技術の急速な発展によって、いずれのデータも大量に得られるようになった。しかし、それらをつなぐ理論的予測手法がまだ確立していないのである。そのため配列情報と立体構造情報が得られた時、配列情報は統計的アプローチによって機能と関係付けられ、立体構造情報は個別のタンパク質のシミュレーションなどによって機能と関係付けられることが多い。しかし、これらを本当に関係付けるためには、配列空間と実体空間の統合的情報処理が必要であり、第２節の(3)、(4)を中心に解析手法の開発を進めていかねばならない。

(3) 生命システムの問題

生物は、分子の膨大な組み合わせによる超複雑かつ階層的なシステムであり、30億余年の進化の過程で数千万種にも及ぶ多様性を獲得している。この生物が示す多様な生命現象を理解するために、個別の要素に還元させて考えていくと、すぐに要素間の組み合わせ爆発の難問に突き当たる。要素の数が増えた時、可能な組合せの数が爆発的に増大するが、そのほとんどは生物の生存とは関係ない無駄な組合せとなり、現実的な時間で妥当な解を得ることができなくなるのである。しかし、実際には生物は、全く無駄な組合せの枝刈りを行う仕組みを持ち、様々な生命現象を至極当たり前に行っていると考えられる[29]。生物システムは、問題を飛躍的に小さくする仕組みを内蔵していると考えられるのである（ここで言う仕組みは、簡単な式で書かれる物理の原理とは異なり、部品・システム・設計図などで表現される機械の「標準化された仕組み」のようなことを意味している)。

このオープンプロブレムでも、第２節の(3)における配列空間と実体空間でのサブ空間の解析、および(1)のトップダウンアプローチと(4)における粗視化解析は重要である。

(4) ゲノムと環境と疾患の関係

ゲノムにおける変異と疾患のリスクの関係については、医科学の分野で精力的に研究が行われてきた。このオープンプロブレムはヒトを対象としており、個々人のゲノムや病気、生活習慣その他色々な情報を基盤として研究が行われる。全ゲノムの解析やその相関解析、環境要因の寄与を考慮した解析、個々人の歴史を追跡するゲノムコホート研究など、ヒトについてのビッグデータのより統合的な研究が行われている[16]。そして、個々人のゲノム情報が得られるようになって、少数の遺伝子が決定的な要因になる病気についての解析は大きく進展している。しかし、寄与の小さな遺伝子変異の集団の抽出という問題（Missing Heritability）は、解析が非常に困難であり、ヒトゲノムにおける変異集団に対する新しい解析法が必要となってきている[30]。

このオープンプロブレムでは、最終的に多くの変異集団に対するシミュレーション（第２節の(4)）が必要となり、その時に(2)のメカニズム解明のアプローチを組合わせることが有効だと考えられる。

(5) 意識・精神活動の理解

ヒトの意識や精神活動は、ヒトを理解する上での最大の未解決問題だろう。例えば意識、知能（抽象化メカニズム）とは何か？創造性・美を感じる感性とは何か？また本能と知識の違いは何か？などの疑問や、これらがどのような生命活動に基づいて生じるのか？そもそも物質および電気的・化学的反応の集合体である脳から、どのようにして主観的な意識体験というものが生まれるのか？という問題[31]などの疑問は未解決のままである。本報告の主眼である生命科学の原理的理解という視点すら、人間が外界を認識する精神活動の一環であり、ヒトを特別視する人にとっては究極の目標とも言えるテーマである。

このような人の認知メカニズムを解明する端緒として、例えば人の脳神経系のネットワーク構造をパターン化して、他の生物（例えば猿）の脳と比較することが考えられる。最近は、そのような静的なイメージングの解析だけではなく、脳神経系の活性部位の動的変化に対する解析なども可能となっており、このオープンプロブレムに対する研究は確実に進んでいる。しかし、脳神経系の形成と精神活動のメカニズムが、どこまでゲノム情報によって規定されているかなどの問題は、ほとんど分かっていないと言ってよい。そういう意味で、これは脳科学とゲノム科学の重なりのあるオープンプロブレムである。

４　新しい分野開拓と人材養成

日本学術会議「日本の展望―学術からの提言 2010」では、学術分野や人材養成について、一見矛盾する課題の両方を追求しなければならないとして、次のように指摘している[32]。「現在の学術の体系は、……、全ての面において縦割りであり、細分化されている。この中で次世代を担う研究者は、二つの一見矛盾する課題を追求しなければならない。すなわち、その一つは、縦割り的教育・研究体制下において、自分が属する特定の研究分野についての知識と経験を十分に蓄積しながら、そこで研究課題に取り組むことである。もう一つは、学術全体を俯瞰し、学術と社会の関係について深く考察ができるような能力を養うことである。」すでに述べたとおり、生物が非常に複雑な対象であることと学問の歴史的経緯のために、これまでの生物学は縦割りの学術的体系の下で研究開発が行われてきた。これに対して生物系ビッグデータは、縦割りの学術的体系を横断的にデータベース化したものであり、それを対象として解析するバイオインフォマティクスは本来きわめて横断的な学問分野である。従って、生物系ビッグデータは生物に関わる全ての研究者たちに対して、生物を俯瞰的に深く考察をするための材料を提供しており、バイオインフォマティクスはそれを解析するための手法や生物を俯瞰的にみる考え方を提供する学問分野であるはずである。しかし、バイオインフォマティクスはこれまで現在の縦割りの学術分野にできるだけ合わせて研究開発を行ってきたという側面がある。本報告は、本質的に学問横断的な生物系ビッグデータとバイオインフォマティクスの将来的なあり方を述べたものである。以下、分野の開拓、人材養成、および社会的インパクトについて、簡単にまとめる。

(1) 分野の開拓

本報告では、まず生物系ビッグデータに対する４種の情報解析の方向性（(1)ボトムアップ・アプローチとトップダウン・アプローチの統合、(2)メカニズム解明のアプローチと統計的アプローチの統合、(3)配列空間と実体空間の統合的情報処理、(4)生物の多様性と複雑さのシミュレーション）をまとめ、それによって解決できる可能性があるオープンプロブレムを５種に整理した（(1)生物の起原、進化、多様性の問題、(2)生体分子の構造形成と機能相関の問題、(3)生命システムの問題、(4)ゲノムと環境と疾患の関係、(5)意識・精神活動の理解)。物理学などの学問分野では、多様な現象を理解するための原理的考え方があり、その下で個別の課題の研究開発が行われ、その成果は再び学問横断的な原理にフィードバックされるというサイクルがある。これに対して、生物学では細分化された個別課題の集積として研究開発が行われ、物理学の場合のようなサイクルは希薄だった。しかし、生物系ビッグデータの時代に入り、それを細分化された従来の生物学の諸分野では扱いきれない状況が生まれてきている。まさに今、生物系ビッグデータ時代に合わせた学問開拓が求められており、バイオインフォマティクスのステップアップが必要なのである。平たく言えば、「データを積み重ねていくと、いずれ大目標に何がしかの貢献ができる」というスモールサ

イエンスとしてのコミットの仕方ではなく、次世代生物学ではオープンプロブレムに基づくサブテーマを列挙し、慎重に設計した上で、それらを体系的に具体化するというコミットの仕方を考えねばならない。分野に関しては、様々な場で意識的に設計を行っていく必要がある。例えば、国レベルの研究プログラムであれば各々のオープンプロブレムが大きな柱になり、その下のサブテーマの設計に関する議論を醸成していく。各オープンプロブレムの柱が、例えば研究プログラムの予算配分の柱になり、あるいは仮に研究拠点を設立するとすれば、研究グループのテーマの柱になるという設計があるべき姿だと思われる

(2) 人材養成

生物情報分野の人材の必要性は、ヒトゲノム計画がスタートした 1990 年代から再三指摘され、そのための予算も組まれてきた。それにもかかわらず生物情報分野の人材不足が言われて久しい。実際、生物系ビッグデータを解析するための人材は非常に不足している[33]。多様なビッグデータを必要とする生物系の学問分野が急速に拡大し、それぞれに合わせた解析を行うために常に人材不足に陥っていると考えられる。この状況を打開し、より分野横断的な人材を養成するには、従来の人材養成とは異なる仕組みを考える必要がある。例えば、バーチャルな研究教育組織を形成し、それを通して次世代生物学へ向けての新しいバイオインフォマティクスのコアとなる組織と、従来の研究教育組織がつながるような多重構造の人材養成の仕組みが考えられる。本報告で述べた情報解析の方向性を志向する人材の養成には、すでに個別の分野で確立した人たちが、そのマインドを拡大し、要素還元的考え方に対して統合的考え方を加えていけば良いと考えられる。そのためには、従来からの研究教育組織にポストを得ながら、バーチャルな研究教育組織に参加する形が重要な役割を果たす可能性がある。それによって、生物系ビッグデータを用いて、生物全体の原理的な仕組みを解明していくような人材、あるいはそのような志向を持って新しい学問分野を開こうとする若い人材を養成しつつ、従来からの情報解析人材の拡大にもつながるのではないかと考えられる。

人材養成を考えるとき、その量的拡大だけではなく、人間としてのモラルや倫理についても意識していかねばならない。科学の新しい扉が開かれる過程では、想定していなかった社会的問題が顕在化することもある。次世代生物学への研究開発の推進と同時に、生命倫理の問題についても意識し、十分注意を払うことができる人材が求められるのである。

生物学のオープンプロブレムは、生物情報分野の人材養成だけではなく、理系全体の人材養成にも関係していると考えられる。宇宙科学における「「はやぶさ」の快挙」、物理分野の「ダークマター」や「ヒッグス粒子」の話題、また生命分野の「iPS 細胞」のノーベル賞受賞など、宇宙や生命に関する夢のある話題が続いた後、理学系分野での大学志願者が増加した[34]。生物系ビッグデータの存在と生命の根源的なオープンプロブレムを正しくリンクさせ夢を語ることによって、学生を多く生物情報の分野の

みならず、理系全体に迎え入れることになるのではないかと考える。

(3) 社会的インパクト

地球に暮らす多様な生命全体の持続的な存続と、人々の幸福にとって、生命科学は様々な形で重要な関わりを持っている[13]。そして、ここで述べた様々なオープンプロブレムは、生物に関連する課題をほぼ網羅するほど広範囲な分野をカバーしている。次世代生物学が社会に貢献できる内容はきわめて大きいのである。例えば現在考えられる具体的な問題を見てみても、癌・免疫システムの理解、再生医療への貢献、創薬支援、生命を模した工学研究、生命環境問題などの解決など広範囲に及ぶ。次世代生物学の産業的・経済的価値や社会的インパクトは計り知れないものになるだろう。そして、その反面として、生物科学の最先端分野にあるからこそ人間としての謙虚さを忘れず、生命倫理の問題を常に意識し、十分注意を払っていかねばならない。また、国民の十分な理解を得るように、アウトリーチの努力も行っていかねばならない。

＜参考文献＞

[1] 日本学術会議、『日本の展望―学術からの提言 2010、報告「統合生物学分野の展望」』、2010 年 4 月 5 日.
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-21-h-2-2.pdf

[2] J. D. Watson and F. H. C. Crick: “Molecular structure of nucleic acids - A structure for deoxyribose nucleic acid” *Nature* 171, 737 - 738 (1953).

[3] Crick, F. H.: “On Protein Synthesis”, Symp. *Soc. Exp. Biol.*, XII, 139-163, (1956).

[4] Crick, F. H.: “Central dogma of molecular biology”, *Nature*, 227 (5258): 561-563, (1970)

[5] International Human Genome Sequencing Consortium: “Initial sequencing and analysis of the human genome”, *Nature*, 409, 860 - 921 (2001).

[6] 実験医学「ゲノム医科学・生命科学研究」総集編(2013)vol31,no15(増刊号)羊土社

[7] Hey, T., Tansley, S., Tolle, K.: “The Fourth Paradigm: Data-Intensive Scientific Discovery”, Microsoft Research, (2009).

[8] White, T.: “Hadoop: The Definitive Guide”, O'Reilly, (2012).

[9] 野村総合研究所: “ビッグデータ革命～無数のつぶやきと位置情報から生まれる日本型イノベーションの新潮流～”, アスキー・メディアワークス, (2012).

[10] A special report on managing information: “Data, data everywhere”, The Economist, 2010.
(http://ai.arizona.edu/mis510/other/Data%20Data%20Everywhere%20SAP%20and%20the%20Economist.pdf)

[11] WikiBooks “Unsolved Problem in Biology”
(http://en.wikibooks.org/wiki/Unsolved_problems_in_biology)

[12] “Open questions in Biology” Collection published: 1 February (2013), Last updated: 4, April, (2014) including 16 comments in *BMC Biology*.
http://www.biomedcentral.com/bmcbiol/series/openquestionsinbiology

[13] 日本学術会議、『日本の展望―生命科学からの提言』、2010 年 4 月 5 日.
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-21-tsoukai-2.pdf

[14] 美宅成樹編「ゲノム系計算科学」共立出版 (2013).

[15] Supratim Choudhuri: “Bioinformatics for Beginners: Genes, Genomes, Molecular Evolution, Databases and Analytical Tools”, Academic Press (2014).

[16] 日本学術会議、第二部ゲノムコホート研究体制分科会、提言『100 万人ゲノムコホート研究の実施に向けて』、2013 年 7 月 26 日.
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t176-1.pdf

[17] Manyika, J. et al. “Big data: The next frontier for innovation, competition,

and productivity”, McKinsey Global Institute, (2011).
(http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)
[18] Trifonova, O. P. *et al.*: “Big Data in Biology and Medicine”: Based on material from a joint workshop with representatives of the international Data-Enabled Life Science Alliance, *Acta Naturae*, 5 (3), 13-16, (2013).
[19] Marx, V.: “Biology: The big challenges of big data”, *Nature*, 498 (7453), 255-260, (2013).
[20] “Community cleverness required” *Nature,* 455, 7209-7211. (2008).
[21] Michael,Metzker,:“Sequencing technologies-the next generation” *Nature Reviews genetics*, 457, 32-46, (2010).
[22] Delort P., OECD ICCP Technology Foresight Forum, (2012).
(http://www.oecd.org/sti/ieconomy/Session_3_Delort.pdf#page=6)
[23] スチュアート・カウフマン: ”自己組織化と進化の論理 宇宙を貫く複雑系の法則”日本経済新聞社、ISBN (1999).
[24] 金子隆一 「大量絶滅がもたらす進化」サイエンスアイ新書
[25] The ENCODE Project Consortium: “The ENCODE (ENCyclopedia Of DNA Elements) Project”, Science, 306 (5696), 636-640, (2004).
[26] The ENCODE Project Consortium:“An integrated encyclopedia of DNA elements in the human genome”. *Nature*, 489 57-74 (2012).
[27] Roger P. Alexander *et al.*, “Annotating non-coding regions of the genome”, *Nature Reviews Genetics* 11, 559-571(2010).
[28] The International Human Epigenome Consortium (IHEC),
(http://ihec-epigenomes.org/)
[29] Kitano, H.: “Biological robustness” Nature Reviews Genetics 5, 826-837, (2004).
[30] Joel T. Dudley and Konrad J. Karczewski:“Exploring Personal Genomics” Oxford University Press (2013).
[31] Chalmers, David J.: ″Facing Up to the Problem of Consciousness″. Journal of Consciousness Studies, 2(3), 200-219 (1995).
[32] 日本学術会議、『日本の展望-学術からの提言 2010』、2010 年 4 月 5 日、p28.
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-21-tsoukai.pdf
[33] 日本経済新聞:「ビッグデータ分析に人材の壁、25 万人不足見通し ＩＴ各社、育成へ本腰」2013 年 7 月 17 日記事
[34] 大学探しナビ入学志願者速報
http://www.univpress.co.jp/university_admissions/20140212-2/

＜用語の説明＞

1)セントラルドグマ

遺伝情報は、DNA 塩基配列に保存され、mRNA 塩基配列を経て、タンパク質のアミノ酸配列に伝達されるという概念である。バクテリアからヒトまですべての生物に共通する基本原理となっている。

2)プレディクティブ・サイエンス

初期条件を与えると現象あるいは現象の基本的特徴を予測できる物理学のような科学分野をプレディクティブ・サイエンスと呼ぶ。バイオインフォマティクスでは各種の予測を試みているが、まだプレディクティブ・サイエンスとしては未熟な段階にある。

3)粗視化

一般に、あるユニット中の物性量の平均を取り、その平均値でユニットの挙動を議論する方法を粗視化と言う。生物は階層的にできているので、粗視化は本来生物の解析には適切な手法である。ゲノムの解析には、配列の物性分布の粗視化が有力だと考えられる。

4)ムーアの法則

コンピュータの性能の長期的傾向に関する法則である。実際には、２年に２倍程度のスピードでコンピュータの性能は向上している。ゲノムの配列データの増大スピードは、ムーアの法則をはるかに超えており、ゲノム情報解析手法の革新的方法が求められている。

5)ロバスト

パラメータを変化させたときに、系の状態が変化しにくいことをロバストと表現している。分子レベルから進化レベルまで生物の様々な状態は、環境に対してロバスト(安定)であることが多い。

6)素過程

化学反応の場合、一般に複雑な経路をたどるので、その要素となる反応過程を素過程と呼ぶ。ここでは、それをコンピュータ上で表現する計算の要素も素過程と呼ぶことにする。

7)ボディプラン

脊椎動物は、頭部の中枢神経や分節化した骨格など共通の基本的な構造でできている。これをボディプランと言う。基本ボディプランを持った生物が非常に多様化し

ているのは脊椎動物だけではない。多くの多細胞生物はボディプランを多様化しつつ、共通のボディプランの生物がさらに多様化している。

8)配列空間と実体空間

可能な配列の全組合せの集合を配列空間、また３次元空間における可能な構造、形態の集合を実体空間とここでは呼んでいる。配列空間の場合は、要素の配列は原理的には数え上げることができるが非常に膨大である（DNA 塩基配列では、N 個の塩基で要素の数は 4^N となる)。また、実体空間では要素の数は数え上げができない膨大なものとなる。タンパク質の立体構造では、可能なすべての構造の集合を、全構造空間と表現する。

9)サブ空間

全ての要素を含む配列空間や実体空間の中で、特定の性質を持った要素の部分集合をサブ空間と呼んでいる。例えば、全生物個体が持つゲノム塩基配列の集合は、配列空間のサブ空間となっている。

10)時間ドメインと空間ドメイン

多くの要素が相互作用している複雑な系が時間変化しているとき、１つの要素の時間変化と、ある時点での系内の要素の分布には関係がある。生物進化の時間変化を現在の生物集団の分布から推論できるのは、生物界でも時間ドメインでの変化と空間ドメインでの分布に強い関係があるからである。

＜参考資料１＞分科会審議経過

平成２４年

３月２８日　分科会（第１回）
○年間活動計画、国立研究所設立の議論とシンポジウム開催について

１０月　２日　分科会（第２回）
○シンポジウム開催準備について

１１月１２日　分科会（第３回）
○国立研究所設立に関連するシンポジウムに関して

平成２５年

１月２５日　分科会（第４回）
○シンポジウム開催について：＜参考資料２＞参照

８月　１日　分科会（第５回）
○シンポジウムについて報告

平成２６年

１月２０日　分科会（第６回）
○提言内容の作成について

４月２５日　分科会（第７回）
○報告書（案）についての審議

６月２３日　分科会（第８回）
○報告書（案）についての審議

８月２８日　日本学術会議幹事会（第 199 回）
基礎生物学委員会・統合生物学委員会・農学委員会・基礎医学委員会・薬学委員会・情報学委員会合同 バイオインフォマティクス分科会報告「大容量情報時代の次世代生物学」について承認

＜参考資料２＞公開シンポジウム

「バイオインフォマティクスのパラダイムシフト 30 年後の生命科学の姿を描いて」
平成 25 年 1 月 25 日（金）13:00～17:30　名古屋大学 ES ホール

13:00～13:30　生物の学問体系構築に向けて
　　久原　哲（九州大学農学研究院教授）
14:00～14:30　ビッグデータとしてのゲノム
　　宮野　悟（東京大学医科学研究所ヒトゲノム解析センター教授）
14:30～15:30　医療立場からバイオインフォマティクスに期待するもの
　　尾崎紀夫（名古屋大学大学院医科学研究科教授）
15:00～15:30　生命動態におけるバイオイメージング
　　上田昌弘（理化学研究所生命システム研究センターグループディレクター）
15:30～15:50　休憩
15:50～16:20　国立研究所構想
　　美宅成樹（名古屋大学大学院工学研究科教授）
16:30～17:30　パネルディスカッション「30 年後の生命科学の姿を描いて」
　　司会：斎藤成也（国立遺伝学研究所教授）
　　五條掘孝（国立遺伝学研究所教授）
　　郷　通子（情報・システム研究機構理事）
　　岡崎康司（埼玉医科大学ゲノム医科学研究センターゲノム科学部門教授・所長）
　　久原　哲（九州大学農学研究院教授）
　　美宅成樹（名古屋大学大学院工学研究科教授）

（所属はシンポジウム開催当時）