解説

バイオ，蛋白，化合物の統合情報:ドラッグインフォマティクス

Informatics which Integrate Bio, Protein, and Compounds: Drug Informatics

ライフサイエンス推進室

湯田　浩太郎　　Kohtaro Yuta

ま　え　が　き

DNA配列の変化としてゲノム上に刷り込まれた遺伝情報を私たちの生活に役立たせる応用分野として，遺伝情報を解析し，その情報から新しい医薬品を開発するゲノム創薬が注目されている。ゲノム創薬は疾患関連遺伝子を探索するバイオ分野，遺伝子情報をもとに展開される蛋白分野，蛋白と化合物（薬）の相互作用を研究し，最終的な薬へと導く化合物分野の相互連携に支えられている。このバイオ，蛋白および化合物のそれぞれの分野で産業革命的な研究技術革新が進行しつつある。今後はこれら3分野間の境界領域研究や，3分野を統合し，より高度，かつ複雑な研究を目指した複合領域研究が新たな研究ターゲットとして注目されていくであろう。

本稿では，本特集号が扱うバイオ，蛋白，化合物を分野横断的に概観し，同時に個々の分野の大まかな研究の流れと，代表的な研究項目なども概観する。最後に，バイオ，蛋白，化合物の3分野情報を新薬開発という観点で統一する「ドラッグインフォマティクス」の概念を導入し，その重要性について述べる。

本稿の主たる目的は上記3分野にあまり親しみのない方々にこれらの分野の大まかなイメージをつかんでいただくことである。バイオ，蛋白，化合物の個々の分野は極めて広く，奥が深い。本稿で取り上げた内容はこれら研究項目の一部にしかすぎないことにご理解いただきたい。また，以上の3分野に少しでも興味を持たれたならば，望外の喜びである。

情報的観点から見た3分野

バイオ，蛋白および化合物の3分野の基本となる研究形態はWET（試験管などの器具や溶媒を用いた実際の実験を伴う研究形態）と称されている。このWETに対し，コンピュータ上で行う実験（シミュレーション）は実際の実験器具や溶媒などを用いないのでDRYと称されている。本稿はこのDRY（情報）という観点からバイオ，蛋白，化合物の3分野の研究を概観する。

上記3分野の情報学はそれぞれ，バイオインフォマティクス，プロテオインフォマティクス，およびケムインフォマティクスとして展開されている（**図-1**）。

それぞれのインフォマティクスでは，個々の目的に応じた最適な解を得るべく様々な技術が駆使され，展開されている。しかし現時点では3分野の情報を横断的にまとめ，より高度な目的と統一

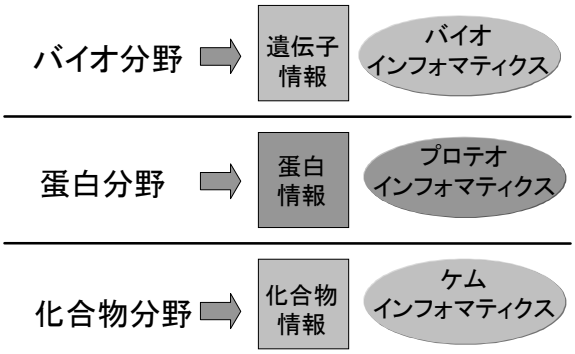

図-1　新薬開発上での3大研究分野と情報
Fig.1-Three major research fields and informatics on new drug development.

的な視野で情報をまとめるという流れは始まったばかりである。

● バイオインフォマティクス

バイオ分野における大まかな研究の流れと，個々の研究分野において適用される様々な研究項目を**図**-2にまとめた。

なお，個々の研究分野の果たす役割や目的をイメージ的につかめるように文例解析に例えてある。

(1) ゲノム配列決定（役割：原文の文字種/配列の決定）

ゲノム配列決定自体は主としてシーケンサーと呼ばれる実験装置を用いて実施される。これらのWET系実験でもコンピュータの果たす役割は大きく，Celera Genomics社のJ．Craig Venter社長が数百台のコンピュータを駆使してヒトゲノム配列解析を先導し，歴史に名を残したことは記憶に新しい。現在は人以外の動物，植物，微生物などのゲノム配列解析が急速に進展している。これらのゲノム配列のデータは米国のGenBank，日本のDDBJなどの公的機関が運営するデータベースに登録され，インターネットを通じて全世界に公開されている。この情報量は1,700万件以上が登録されている膨大なデータベースである。

(2) 遺伝子探索（役割：文章中における単語部分の特定）

ゲノム配列が解読されても，この段階では単に4種類の核酸を示すA，G，C，Tの文字コードで構成される文字パターン情報にしかすぎない。医薬開発のターゲットである遺伝子をこの文字列から探し出すことが必要である。

遺伝子探索は，純粋に遺伝子のみをゲノム配列データから取り出す一般遺伝子探索と，特定の疾病に関与する疾病遺伝子を選択的に取り出す手法とに分類される。

- 一般遺伝子探索に関するアプローチはホモロジー検索を主体とし，FASTA，BLAST，PSI-BLASTなどの検索手法を用いて展開されている。ホモロジー検索とは，相同性検索とも呼ばれ，類似した配列の遺伝子を上記の膨大なゲノム配列データベースから検索することである。最近では学習的機能を取り入れた隠れマルコフモデル（HMM：Hidden Markov models）の利用が多い。
- 疾患遺伝子探索には，遺伝子多型解析と呼ばれる手法が適用される。この遺伝子多型解析を行うときに根拠とする基本原理は，遺伝子の組み替え時に遺伝子上に存在する遺伝子マーカー（染色体上の遺伝子の位置を特定付ける目印）

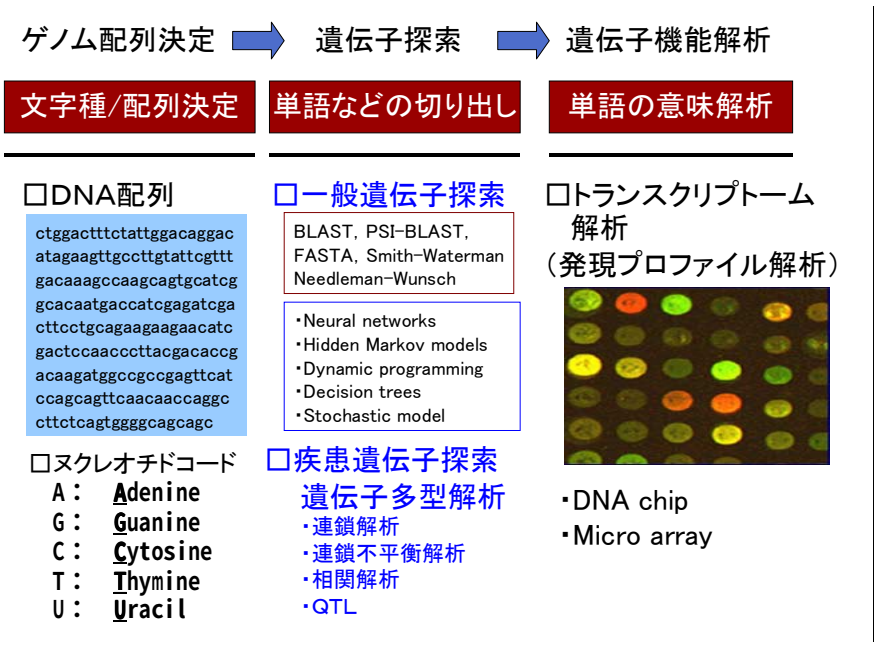

図-2　バイオ分野における研究の流れ
Fig.2-Research flow of bio related fields.

が，近傍の遺伝子と一緒に組み替えられるという遺伝学上の現象に基づいている。実際の解析では，家系図情報や患者と健常者間のマーカーの出現頻度差データなどを用い，統計手法を適用して解析する。

なお，本手法では疾患遺伝子そのものを特定するのではなく，疾患遺伝子が含まれる近傍領域を特定する。

(3) 遺伝子機能解析（役割：単語の意味，内容解析）

ゲノム上の遺伝子が特定されても，多くの場合はその機能が不明である。遺伝子を医療や新薬開発に結び付けるためには，遺伝子の機能を解明することが必要である。

遺伝子機能解析手法としていろいろあるが，現在注目されている手法はトランスクリプトーム解析（遺伝子発現プロファイル解析）である。この手法はガラスやプラスチック上に目的遺伝子を取り上げるためのプローブ（探索子）と呼ばれるものを貼り付けたもので，DNAチップと呼ばれるものと，スライドガラス上にプローブを貼り付けたマイクロアレイと呼ばれる2種類が存在する。この解析により，様々な外的条件で変化する細胞内部での遺伝子発現（遺伝子の働き状態）に関する情報（高発現遺伝子群，低発現遺伝子群）が得られる。

上記（2）で述べた遺伝子多型解析は，目的疾病と相関する遺伝子を特定するものであり，疾病と遺伝子との「静的」な相関情報を求める手法である。トランスクリプトーム解析で得られる情報は，細胞内の遺伝子群の時間単位での発現（遺伝子の活動状況）情報であり，本手法により疾病と遺伝子との「動的」な相関情報が得られる。

● プロテオインフォマティクス

蛋白関連研究分野とその流れを**図**-3に示す。研究分野の大まかな内容や流れ的にはバイオ分野と大きな差異はなく，解析技術なども一部重複している。

蛋白は遺伝子を構成するA，G，C，Tの核酸配列情報が解読され，これに対応する20種類のアミノ酸に置きかえられて再構築されたものである。ゲノムは生命の設計図そのものであり，種の保存のための情報伝達がその基本であるために，配列情報そのものが重要である。これに対して蛋白は，ゲノムの設計図から実際に生命を維持する目的で構築された機能性物質である。したがって，アミノ酸配列情報そのものよりは，その機能を実現するのに重要な二次構造，とくに三次元構造が研究上での大きなターゲットとなる。

バイオ分野と同様，蛋白研究の流れは大きく3

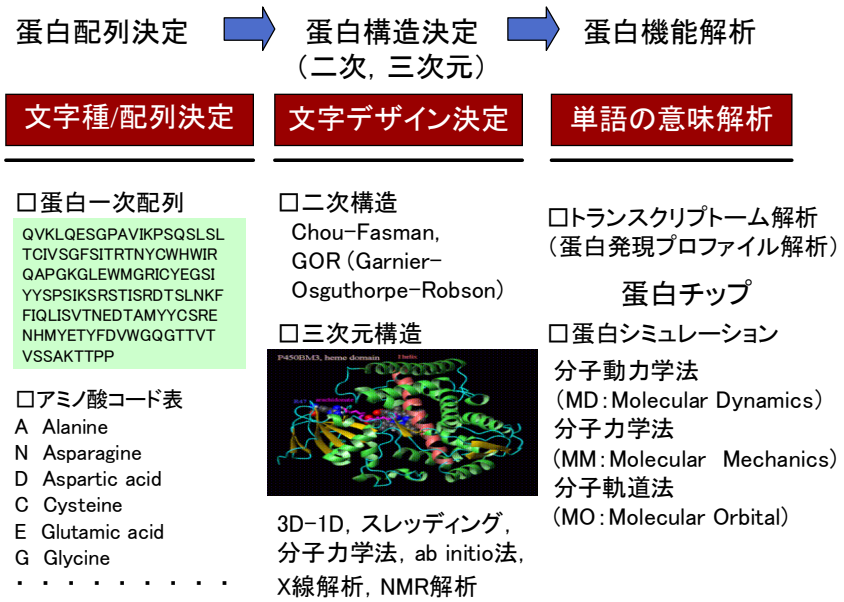

図-3 蛋白分野における研究の流れ
Fig.3-Research flow of protein related fields.

段階に分類される。

なお，本節では蛋白を単語に例えてそれぞれの役割について説明している。

(1) 蛋白一次配列（アミノ酸配列）決定（役割：単語の文字種/配列決定）

蛋白は20種類のアミノ酸から構成され，このアミノ酸が数百から数千，数万とつながって構成されている。このアミノ酸のつながりを一次配列といい，個々のアミノ酸を代表する20種類の文字で構成される文字パターン情報となる。

アミノ酸配列の決定はWETによる実験で行われる。最初に二次元電気泳動で蛋白を分離/精製し，続いて質量分析機やシーケンサーを用いてアミノ酸配列を決定するのが一般的である。実験で得られた蛋白とゲノム上の遺伝子の同定はペプチドマスフィンガープリンティング法が一般的に利用されている。

(2) 二次構造，三次元構造決定（役割：文字デザイン決定）

蛋白はゲノムと異なり機能を発現するための二次構造および三次元構造決定が研究上のターゲットとなる。二次構造ではアミノ酸配列上のαヘリックス，βシート，コイルなどの蛋白に特徴的な構造部分が特定される。三次元構造では蛋白全体の立体構造が決定される。

蛋白が純度高く，かつある程度の量が得られればX線やNMR（核磁気共鳴）などを用いて三次元構造を決定する。アミノ酸の一次配列しか分かっていない蛋白は，ホモロジー検索により類似した一次配列を持つ蛋白を検索し，その蛋白の三次元構造を基本にしてターゲットとなる蛋白の三次元構造を再構築する手法（3D-1D法）も用いられる。

(3) 蛋白機能解析（役割：単語の意味/内容解析）

蛋白分野での機能解析には様々な手法が展開されている。現在は様々なWET系中心の実験で特定されているが，ほとんどが蛋白種を限定した状態で研究されている。ゲノム分野のトランスクリプトーム解析と同様，蛋白チップを用いた蛋白発現プロファイル解析も一部では実用化されつつある。本解析により，細胞内部における蛋白/蛋白相互作用，あるいは蛋白/化合物相互作用情報を網羅的に追求することが可能となる。

蛋白シミュレーションによる解析も精力的に展開されている。この分野の基本技術としては分子動力学（MD），分子力学（MM），分子軌道法（MO）などがあり，主として化合物への適用事例が多い。MDおよびMMは三次元構造の最適化とエネルギー計算が可能である。電子関連情報（電子密度，ダイポールモーメントほか）の計算はできないが，高速計算が可能なために蛋白分野でも展開されてきた。電子関連情報はMOで得られるが，計算時間がかかる。このために，MMで蛋白三次元構造を最適化し，この最適化された蛋白を用いてMOの計算により電子関連情報を算出するのが一般的である。

蛋白シミュレーションの今後は，より実際の蛋白環境に近づけた（溶媒効果などの考慮）計算や，シミュレーション時間の延長が求められてくる。このような実環境に近づけた蛋白シミュレーションには膨大なCPU時間を必要とするため，計算アルゴリズムの更なる高速化やグリッドコンピューティングなどのコンピュータ関連高速化技術の導入が加速されつつある。

● ケムインフォマティクス

化合物分野の新薬開発研究の流れはバイオおよび蛋白分野とは大きく異なる。本節では全体を二つの研究分野に分けて議論する（**図**-4）。

本節では新薬開発における個々の研究分野の役割を，植物を育てる場合に例えて説明する。

(1) リード候補化合物探索（役割：「種（タネ）」の発見）

薬の種（タネ）に相当する化合物を「リード化合物」と称する。このリード化合物は，過去においては文献，天然物からの抽出，偶然などにより発見されてきた。現在は，特定の蛋白と反応するリード候補化合物を大量の化合物群から高速に選択（スクリーニングと呼ぶ）する，「コンビナトリアルケミストリー/HTS（**H**igh **T**hroughput **S**creening）」がその役割を担っている。

・コンビナトリアルケミストリー/HTS

本技術で扱う化合物数は年間数百万化合物となる。これにより従来からの化合物合成，薬理スクリーニング（化合物の薬理活性の有無をチェックする），構造-活性相関（化合物構造と薬理活性と

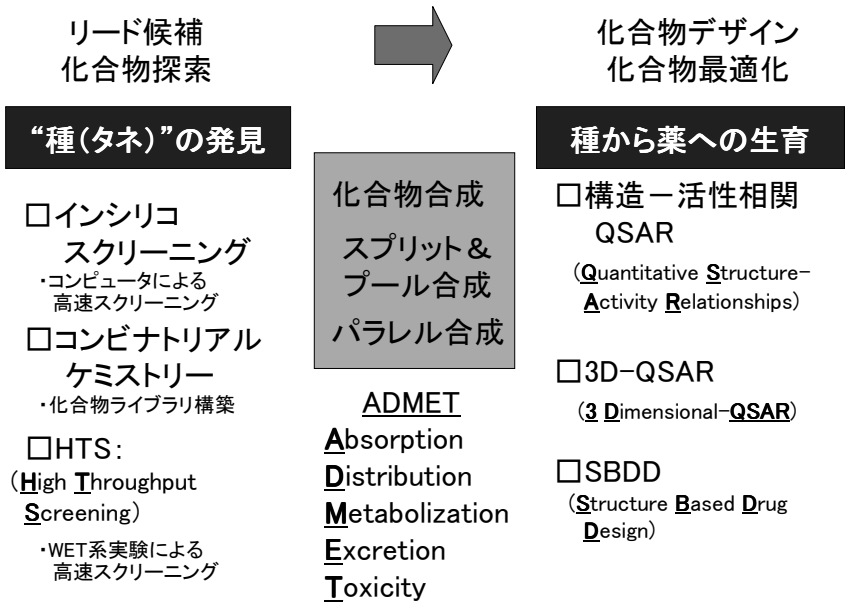

図-4　化合物分野における研究の流れ
Fig.4-Research flow of compound related fields.

の相関を解析する）などの研究の基本的な流れは劇的に変化した。現在の状態は，家内生産的新薬開発現場に大量生産を可能とする新薬開発技術が持ち込まれた状態と言える。

大量の化合物を扱うコンビナトリアルケミストリー/HTSで情報の果たす役割は極めて大きい。現在は情報管理が主体であるが，今後は大量のスクリーニングデータの解析，より効率の高いコンビナトリアルケミストリー/HTS実施のための化合物ライブラリーデザインなどで，高度な情報処理/解析技術が必要となる。

・インシリコスクリーニング（In silico screening）

本技術はコンピュータ上，つまり半導体であるシリコン素子上でスクリーニングを行うものであり，WET系スクリーニング実験の高速化技術であるHTSとは本質的に異なる。化合物の構造式（仮想化合物）だけでスクリーニングを極めて高速に実施できるため，化合物を実際に合成する必要はない。したがって，HTSを実施する前に行うプレスクリーニングとして利用される。

著者が研究している多変量解析/パターン認識によるインシリコスクリーニングは極めて多数の化合物の扱いが可能であり，薬理活性（薬効）のみならず後述するADMETの同時スクリーニングが可能である。このほかの代表的なインシリコスクリーニング手法として，ドッキングシミュレーションがある。これは，蛋白と化合物との結合のし易さをシミュレーションするもので，蛋白とより強く結合する化合物がリード候補化合物となる。このアプローチは詳細な計算を行うと膨大な計算時間がかかるため，さらなる高速計算アルゴリズムの開発が必要である。

(2) 化合物デザイン，最適化（役割：種から薬への育生）

前項（1）で特定された「種（タネ）」は薬の出発点であり極めて重要であるが，これだけでは薬とならない。薬効を高め，毒性や副作用のチェック，投与形態などをも考慮した化合物デザイン，構造-活性相関の適用が必要となる。すなわち，拾い上げた「種（化合物）」を「薬」として機能するように大きく育てるのがこの研究過程である。最近になり新薬開発分野ではADME（薬物動態）の問題に関する研究が特に注目されつつある。このADMEの文字は **A**bsorption；吸収，**D**istribution；分布，**M**etabolization；代謝，**E**xcretion；排泄，の頭文字を取ったものである。

このほかに薬物開発上失敗の許されない問題として，毒性および副作用がある。医薬品の毒性としては発癌性，染色体毒性，突然変異性などの

様々な毒性評価が必要である。これら種々毒性の予測は主として化合物分野で実施されているが，最近はトキシコゲノミクスとして遺伝子発現プロファイル解析データを用いて予測しようとする試みが始まっている。

なお，前記ADMEと毒性（Toxicity）を一つにして，「ADMET」という言葉で表現することが多い。

この研究分野では以前よりドラグデザインや構造-活性相関などの研究が実施されてきた。この分野にも様々な解析手法が存在するが，QSAR（Quantitative Structure-Activity Relationships），3D-QSAR，およびSBDD（Structure Based Drug Design）の3種類に分けるのが手法的にも原理的にも妥当と考える。

「ドラッグインフォマティクス」の必要性

効率的な新薬開発実現には，バイオ分野で遺伝子とその機能が特定されたならば，その情報は蛋白にフィードバックされることが必要である。蛋白とその機能が特定されたならばその情報は化合物にフィードバックされなければならない。膨大な数の化合物中よりリードとなる化合物を探し出し，続いて薬理活性を最適化しADMETなどを検討して最終的な薬とする。その後，臨床試験に入り，審査と医薬品登録を経て最終製品となる。

以上，新薬開発全体を川の流れとして例えるならば，バイオは最上流に位置し，蛋白はその下流（すなわち，中流）に位置する。化合物は蛋白の下流に位置し，河口（薬）へとつながる（**図**-5）。

薬は，上記3分野の技術が相互連携して作り上げた結晶である。したがって，新薬開発ではバイオ，蛋白，化合物と情報が連続して流れる情報ハイウェイが重要となる。

現状では，個々の分野におけるインフォマティクスは盛んに研究されている。しかし，薬を成果物へと導く最短道となる3分野間の情報連携技術は世界的にも進んでいない。すなわち，3分野内での地方道の整備は進んだが，3分野を貫き，「薬」へと導く縦貫道（情報ハイウェイ）が未整備である。今後はこの縦貫道の早期構築が医薬品開発の効率向上，開発期間の短縮などの観点で極めて重要となる。

バイオ，蛋白および化合物分野の情報を，新薬開発という観点で統一して議論し，情報をバイオ，蛋白そして化合物と流れるようにする情報研究分野を「ドラッグインフォマティクス」と呼ぶ。今後はこの新規研究分野の強力な推進が，世界をリードする新薬開発という観点で極めて重要となる。

最後に，新薬開発を目指すバイオ，蛋白，化合物の研究を情報やデータ解析の効率的運用，あるいは正しい解析を実施するという観点上留意すべき問題点を簡単にまとめる（**図**-6）。

(1) 情報と専門研究分野のコミュニケーション

研究分野のシステムサポートで常に問題となることがある。これは情報関連研究者と専門分野研

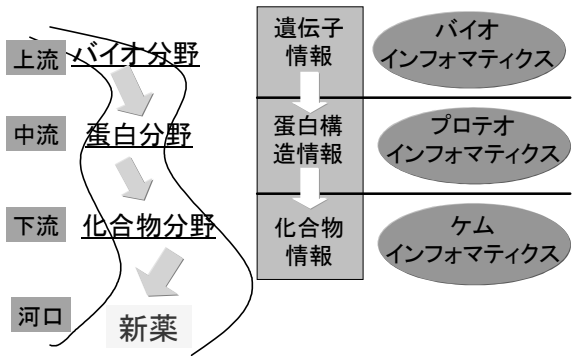

図-5　バイオ，蛋白，化合物の情報の流れを作る情報学：ドラグインフォマティクス
Fig.5-Informatics which make flows of information from bio to protein and compounds: Drug informatics.

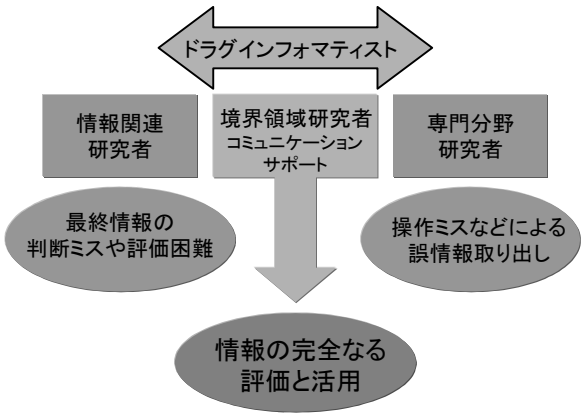

図-6　ドラッグインフォマティストの役割
Fig.6-The role of the drug informatist.

究者間でのコミュニケーションが極めて困難なことである。

専門分野が異なることは，背景となる生活環境も含めて言語が異なる世界と考えるべきである。このような言語間コミュニケーションには言語通訳者が必要なように，異分野研究者間にも研究通訳者が必要である。

(2) 共通基本分野と専門（応用）研究分野

新薬開発は応用研究である。様々な基本技術を駆使して，目的とする結果を導き出すことが必要である。現在，様々なプログラムが販売され，比較的簡単な操作で解析結果が得られる。しかし，データ解析や理論計算は，基本理論や技術を理解しつつ正しい操作をしなければ，結果の保証はない。このような問題を回避するためにも，プログラムの基本となる理論や技術をある程度理解し，臨機応変に信頼性の高い結果を導き出し，専門研究者と議論しつつ解析フィードバックをかける研究者が必要である。すなわち，正しい情報と偽りの情報を正しく判別し，研究自体の方向性を誤らないようにすることが大事である。

文献検索，データ整理/検索，データ解析，理論計算，構造-活性相関など，新薬開発研究に占める情報の役割は日々増大している。また，これらの研究を実施するための基本技術分野も多彩となっており，研究内容や情報自体がバイオ，蛋白，化合物といった様々な分野を横断的にカバーすることが必要となりつつある。

新薬開発における様々な情報や基本技術を統合的に扱い，情報と専門分野研究者間のコミュニケーション確立と融合化を目指し，最高のパフォーマンスを実現する。このような働きをする研究者を「ドラッグインフォマティスト」と呼ぶことにする。情報化投資を無駄にすることなく，その効率的な運用と最大限の成果達成を実現するにはこのような人材の育成が重要である。

むすび

本特集号に関与するバイオ，蛋白，化合物の3分野を，新薬開発という観点で概観し，今後必要となる情報学としてバイオ，蛋白，化合物を統一して議論する「ドラッグインフォマティクス」を提案した。また，情報研究者と専門分野研究者間のコミュニケーションを取り，基本技術の正しい適用を行い，信頼性の高い情報を専門研究者に提供し，議論する「ドラッグインフォマティスト」の必要性についても言及した。

最初に述べたように，本稿で扱った個々の分野は極めて広く，かつ深い。ここで記載，説明した内容は極めて狭く，浅いものであることを改めてことわらせていただく。より詳細な個々の内容に関しては本特集号の論文を参照していただきたい。本特集号を通じて，以上の3分野に関する情報の果たす役割などについて少しでも興味を持っていただければ幸いである。