ま　え　が　き

本書名に含まれる「ゲノミクス情報処理」という用語は，ゲノム情報学やその周辺で重要となる情報処理技術を意図したものである。ゲノム情報学は，生命現象を情報科学の立場から理解する学問分野であり，生命現象の理解には，コンピュータの計算能力の利用が必要不可欠である。シークエンサーの登場は，病気の診断や治療のみならず，創薬に変革をもたらしていることは，周知のとおりである。シークエンサーが一日に解読可能な塩基数は，1985年の時点で，わずか 10^3 bp 程度であったが，それ以後，シークエンサーの高性能化が加速し，ゲノム DNA 配列や細胞内 RNA 配列を読み取る作業が飛躍的に向上している。今日では，およそ 3×10^9 bp もある個人の全 DNA 配列を短時間で解読できるようになってきており，オーダーメイド医療が盛んになるものと期待されている。このようなことからも，ゲノミクス情報処理の分野においても，ビッグデータ時代を迎えるに至っている。

本書では，ビッグデータ時代における大規模ゲノムデータに対する類似性検索や分析を念頭に置き，国際的なデータバンクで整備されている主要なデータベースの内容とそれらの解析方法を中心に解説している。紙面の都合上，多くの内容を網羅できていないが，本書では，分子進化の重要性に多くの箇所で配慮したつもりである。類似性検索については，分子配列（塩基配列やアミノ酸配列）を対象とする類似検索手法をはじめとして，空間情報を含むタンパク質立体構造を対象にした類似構造検索手法を解説している。分析方法については，生命の進化を解明するために重要な分子進化系統樹の推定法，分子進化を考慮した整列化や多重整列化，曖昧（あいまい）性を持つモチーフの表現法や抽出法，生物の運動機能について解説している。さらに，大規模なゲノムのデータ解析の高速化を可能にする情報処理技術についても解説している。

本書には，大学や高等専門学校等の情報系および生物系の大学生および大学院生が，バイオインフォマティクスをはじめとして，ゲノミクス分野に出現するビッグデータ解析の基本原理を理解することを配慮した内容が含まれている。また，生命科学分野のデータサイエンティストやバイオインフォマティシャンの基礎知識を学習しようとする技術者・研究者にとっても必要不可欠な内容が盛り込まれている。なお，各章の執筆担当については，つぎのとおりである。1 章および 3 章については北上，4 章は斎藤，5 章は太田が執筆を担当した。

また，2章については，斎藤が2.1～2.4節，太田が2.5節および2.6節を担当し，6章については，北上が6.1～6.5節を担当，太田が6.6節を担当した。本書を読んで，将来，一人でも多くのバイオインフォマティシャンあるいはゲノミクス分野のデータサイエンティストが活躍することになれば，著者らにとって，これに勝る喜びはないと考えている。

最後に，本書の出版に際し，株式会社コロナ社の方々に感謝する次第である。

2014年9月

北上　　始

目　　次

1. ゲノム情報のデータベース

2. ゲノム配列の決定と解析

3. モチーフの表現と抽出

4.　分子進化系統樹の推定

5.　新しい運動機能解析

6. 高速ビッグデータマイニングへの展開

1 ゲノム情報のデータベース

本章では，まず，ビッグデータとしてのゲノム情報について触れた後，生命科学と情報科学の関係について紹介する。つぎに，ゲノム情報そのものをデータベース化した**塩基配列データベース**について紹介し，それと深いかかわりのある**モチーフデータベース**，**タンパク質立体構造データベース**，**立体構造分類データベース**，**文献データベース**などについて紹介する。最後に，データベースの統合利用や生物医学の研究には欠かせない**オントロジー**について紹介する。

1.1 ビッグデータとしてのゲノム情報

ビジネス分野や学術分野などでは，古くからデータは市販のデータベース管理システムによって構築されてきたが，インターネットやコンピュータ機器の急速な発達・普及が影響し，2000 年に入ってから従来のデータに比べて性質の異なるデータが急激に増加してきた。このようなデータはビッグデータと呼ばれるようになった[1),2)†]。

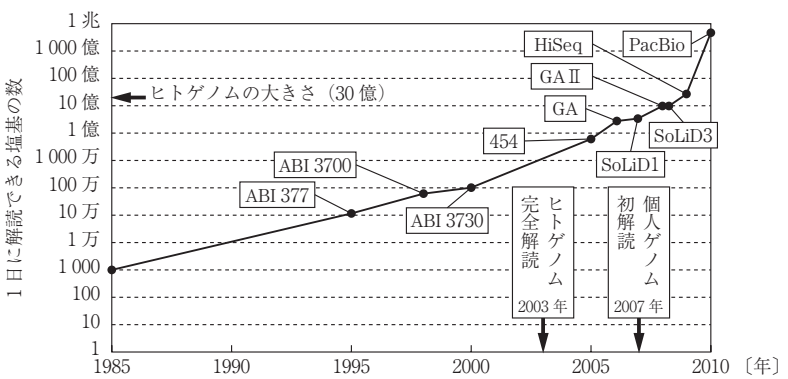
ビッグデータ（big data）とは，① **容量**（volume），② **多様性**（variety），③ **頻度**（velocity）と呼ばれる三つの特徴[3)]の中の二つ以上を持っているデータ集合の集積物を意味する。容量とは，市販のデータベースシステムあるいは標準的な統計処理ソフトウェアの処理能力を超えるぐらいデータが巨大であるという特徴であり，多様性とはデータの種類が多様（非構造な場合が多い）であるという特徴である。また，頻度とは，データが高頻度かつ高速に処理され利用されるという特徴である。このほかに，④ **正確さ**（veracity）という四つ目の特徴が加わっている。正確さとは，データの無矛盾性をはかる指標であり，無矛盾なデータによる信頼できる意思決定を意図している。このような特性を持つビッグデータは，近年，ビジネス分野のみならず学術分野においても扱われる機会が増えている。

本書で注目している**ゲノミクス**（genomics，**ゲノム科学**）は，1980 年代に出現した概念である。ゲノミクス分野で情報処理が注目されるようになったのは，1990 年代に**ヒトゲノム解析計画**[4)]による**ゲノム情報**の解読が開始してからである。近年，この解読されたデータ

† 肩付き数字は，各章末の引用・参考文献番号を表す。

は，ビッグデータの一つに分類されることからもわかるが，巨大化してきている。

ゲノム情報がビッグデータに分類される一つの理由としては，遺伝情報を解読する**シークエンサー**（sequencer）のめざましい性能向上があり，これにより巨大なゲノム情報がつぎつぎと生み出されつつあることが挙げられる。**図 1.1** に表示されているように，1 日に解読できる**塩基**（base）の数を見ると，1985 年ではわずか 1 000 塩基程度であったが，2000 年にはその 1 000 倍の 100 万塩基になっている。また，2010 年には，一日あたり 1 兆塩基近くの解読が可能になっている[5)]。このような高性能なシークエンサーを利用すれば，約 30 億塩基からなる個人ゲノムの短時間かつ安価な解読が可能である。



出典：水島―菅野純子，菅野純夫：次世代シークエンサーの医療への応用と課題，モダンメディア，57 巻，8 号，p. 226 の図 2，栄研化学（2011）を転載。

図 1.1　DNA シークエンサーの性能向上

ゲノム情報がビッグデータに分類され得るもう一つの理由としては，ゲノム情報の解読はされていてもそのゲノム情報の内容がたいへん複雑なため，未知の部分が多いことが挙げられる。このような複雑なゲノム情報を生命科学の知識をもとにコンピュータで分析すれば，

コ ラ ム

DNA

DNA（deoxyribonucleic acid：デオキシリボ核酸）は，1953 年に Watson と Crick がその二重らせん構造を提唱した物質で，生物の親から子に受け継がれる遺伝情報をのせている。DNA は，2 本の相補的な鎖で構成され，2 本の鎖はたがいに絡み合って右巻きのらせん構造をしている。それぞれの鎖は，**ヌクレオチド**（nucleotide）と呼ばれる物質が鎖状に連結された高分子である。ヌクレオチドは，塩基，糖，リン酸から構成される。DNA における塩基の並びは，**塩基配列**（base sequence）と呼ばれ，生物の設計図であり，親から子に受け継がれる遺伝情報の正体である。DNA に含まれる塩基には A（アデニン），G（グアニン），T（チミン），C（シトシン）がある。ただし，RNA の場合は，T（チミン）の代わりに U（ウラシル）となる。

ゲノム情報がタンパク質，細胞，生体系とどのようにかかわっているのかが明らかになるものと期待されている。当然のことではあるが，分析結果に正確さ（ビッグデータの４番目の特性）が要求されることはいうまでもない。

1.2　生命科学と情報科学

ゲノム情報をコンピュータで分析するには，**情報**（information）の本質について明らかにしておくことが重要である。情報とは，本来，「人と人とのコミュニケーションでやりとりされるもの」であったが，近年のインターネットやコンピュータのめざましい発達により，情報やコミュニケーションの概念が拡大解釈され，人とモノとのコミュニケーション，あるいは，モノとモノとのコミュニケーションが注目されるようになってきた[6),7)]。このような背景により，どちらか一方のモノが送信側になり，他方の受信側のモノに変化を与えるとき，その変化を生み出している要因を情報と呼んでいる[8)]。送信側が受信側にメッセージを送信したとしても，受信側に変化を与えない場合は，そのメッセージは情報とはいわない。

情報科学の分野では，ヒューマンコンピュータインタラクションは，人と情報機器（モノ）とのコミュニケーションに強い関心があり，情報ネットワークは，情報機器（モノ）と情報機器（モノ）とのコミュニケーションに強い関心がある。

生命科学の分野では，細胞や生命体などがモノであり，モノとモノとの間でやりとりされる情報の意味は拡大解釈されている。遺伝子発現や分子間相互作用などにかかわるゲノム，ホルモンや神経伝達物質などにより細胞間に伝達される信号，環境（外界）からの生体システムへの刺激などは，ある種の情報である。また，これらの情報が基本となって引き起こされるタンパク質間の相互作用，遺伝子間の相互作用，タンパク質と遺伝子との相互作用，酵素反応サイクルなどでは，モノとモノとの間で情報がやりとりされているとみなされる。

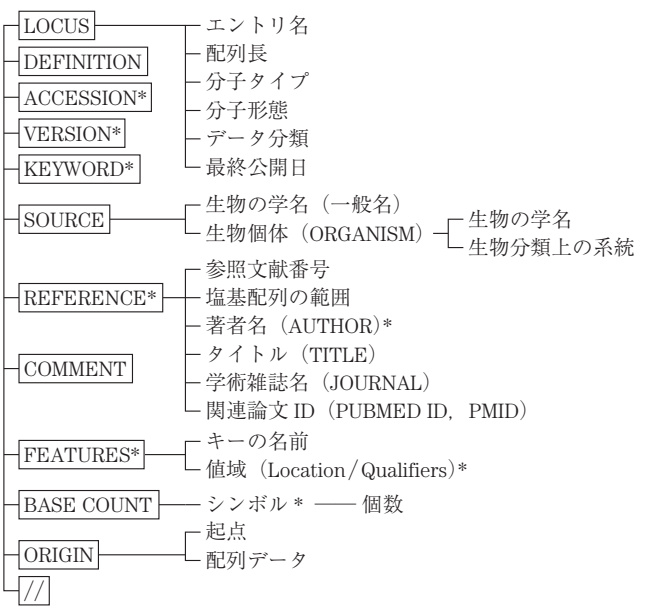
ゲノム情報がタンパク質，細胞，生体システムとどのようにかかわっているのかについては未知の部分が多い。ゲノミクスの研究では，ゲノムと遺伝子について研究し，ゲノム創薬，癌（がん）などの難病の解明，ゲノム比較に基づく生物の進化の解明などが進められている。本書では，これらの研究において，情報科学の知識を用いたコンピュータ分析を**ゲノミクス情報処理**と呼んでいる。なお，情報科学で利用可能な知識としては，データベース技術，機械学習と統計学，自然言語処理，人工知能，コンピュータグラフィックス，画像処理技術などがある。

1.3　塩基配列データベース

塩基配列データベースを構築・維持する組織は，1980 年代から欧州・米国・日本の各機

関で設立されているが，2005 年以来，この塩基配列データベースは，**INSD**（International Nucleotide Sequence Database：国際塩基配列データベース）と呼ばれている[9～11]。日本の機関については，国立遺伝学研究所内の **DDBJ**（DNA Data Bank of Japan：日本 DNA データバンク），欧州については，**ENA**/**EBI**（European Nucleotide Archive/European Bioinformatics Institute），米国については **GenBank**/**NCBI**（National Center for Biotechnology Information）として知られている。国際塩基配列データベース（INSD）には，ゲノム関連の研究者によって直接送付されてきたデータのみならず日本・韓国・欧州・米国の特許庁で処理されたデータも含まれている。

図 1.2 は，塩基配列データベースに登録されている各データの公開形式の概略を図示したものである。各データは，この公開形式でファイルに蓄積される。この公開形式は，フラッ



＊は，繰り返しを表す項目であることを意味する。

図 1.2　塩基配列データベースの公開形式

コ ラ ム

フラットファイル

フラットファイル（flat file）とは，1 行を 1 レコードとする**プレーンテキスト**（plain text）の集まり，あるいは，**バイナリ**（binary）を保存するファイルを意味する。レコードは，フィールドをデリミタ等の記号で区切った構造になっている。DDBJ フラットファイルフォーマットでは，プレーンテキスト形式のレコード間を二つの**スラッシュ**（slash）記号「//」で区切っている。プレーンテキストとは，文字ごとの色や形状，文章に含まれる図などといった情報を含まない文字列形式のコンテンツを意味する。

トファイル形式の **DDBJ フォーマット**（DDBJ format）と呼ばれる。DDBJ から定期的にリリースされている塩基配列データベースは，2014 年 7 月の時点の Release 97.0（2014 年 6 月公開）では，172,402,324 件（総塩基数は 161,078,598,329 個）もある[11]。どの 1 件のデータ（エントリ）も，この図の形式で表現されている。

以下に，この公開形式で使用されている各予約語について，その予約語で表記される行の説明を簡単に行う。

（1）“LOCUS”行には，Locus 名，配列長，塩基配列の分子タイプ，塩基配列の分子形態，Division（21 種類に分類），データの最終公開日が記録されている。Locus 名はデータベース中でそのエントリのみが持つユニークな名前であり，かつてはそのエントリーにふさわしいものが使われていたが，データが爆発的に増加したので，現在はアクセッション番号と同一になっている。

（2）“DEFINITION”行には，データの定義や遺伝子などに関する簡略情報が記録されている。

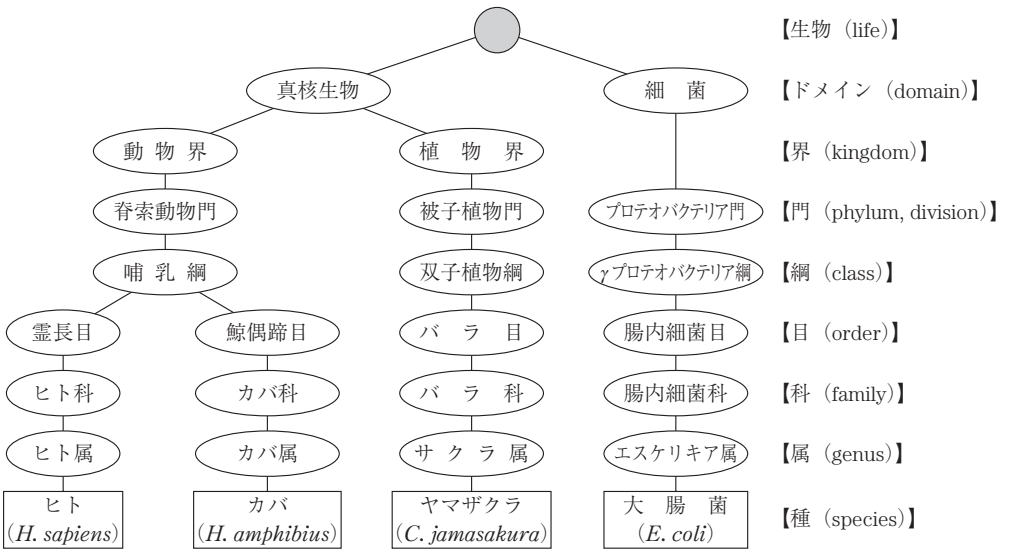
（3）“ACCESSION”行には，**アクセッション番号**が記録されている。ただし，アクセッション番号は，**INSD** が発行する登録番号であり，アルファベット 1 文字＋5 桁の数字，または，アルファベット 2 文字＋6 桁の数字（例 AB123456）で構成されている。

（4）“VERSION”行には，アクセッション番号とバージョン番号が記録されている。ただし，初めて公開されたデータのバージョン番号は“1”で表記されている。

（5）“KEYWORD”行には，データの詳細種別（EST，TSA，HTG，WGS，TPA など），配列の特性，実験手法，ゲノム配列の完成度などが記録されている。

（6）“SOURCE”行には，配列データが由来する生物の学名（一般名が存在する場合はその名前）が記録されている。また，その中の ORGANISM 行には，由来生物の生物名と**系統関係**（lineage）が記入されている。**図 1.3** に生物の**分類階級**とヒト・カバ・ヤマザクラ・大腸菌に対する系統関係を表現した**生物分類樹**の例を図示する。この木構造では，子ノードと親ノードの間に *is-a* 関係の性質を満たす。たとえば，「サクラ属は，バラ科である」という性質は，“*is-a*（サクラ属，バラ科）”と表記される。

（7）“REFERENCE”行には，先頭に付けられた番号でデータベース登録者と掲載ジャーナルの情報を区別している。AUTHOR 行で，番号 1 は原則としてそのエントリの登録者（Submitter（s）），2 以降の番号は論文の著者名を記録している。TITLE 行で，番号 1 は“Direct Submission”がつねに表示，2 以降の番号は論文のタイトル，あるいは，まだ出版されていない場合は予定されるタイトルを記録している。JOURNAL 行で，番号 1 は 1 行目にはそのエントリの受付日（Accept Date），2 行目以降には，コンタクトパーソンの氏名，所属等の情報を記入している。2 以降の番号については，論文が出版



動物界では「門」を英語で phylum と呼び，植物界では「門」を division と呼んでいる。植物界では，「属」と「種」の間に「節（section）」がある。また，この例にはないが，ウイルスのみ「界」と「門」の間に「群（section）」がある。一部の細菌のみ，「門」と「綱」の間に「群」がある。これらの例外を除き，一般に，分類階級間に階級が必要になる場合は，基準となる分類階級から見て上位か下位かによって，階級名の先頭に「上（super）」「亜（sub）」などが付与される。

図 1.3　ヒト・カバ・ヤマザクラ・大腸菌に対する系統関係を表現した生物分類樹の例

された場合，あるいは印刷中の場合には，論文の雑誌名等が記録されている。PUBMED 行には，**生物医学論文データベース** PubMed に登録されている関連論文 ID（PMID）のリストが記録されている。

（8）“COMMENT”行では，つぎの（9）で記述できないその他の情報やコメントを記録している。

（9）“FEATURES/Location/Qualifiers”行では，塩基配列の生物学的な特徴について，Feature key（特徴を表す項目）ごとに，Location（配列上の位置情報）および Qualifier（特徴をさらに特定する項目）で記録している。Feature key は，source，CDS，rRNA，variation，conflict などの多くのキーワードを用いて，特徴を記録している。その中の source 行では，由来生物の特徴が記録されている。CDS や rRNA などの行では，配列の中の一定の領域が持つ生物学的機能が記録されている。塩基配列の翻訳により得られた**アミノ酸配列**は，CDS 行の /translation の箇所に記録されている。variation，conflict などの行には，配列の差違や変更が記録されている。

（10）“BASE COUNT”行には，塩基配列に含まれる各塩基の出現数が記録されている。

（11）“ORIGIN”行では，塩基配列をすべて小文字で記録している。10 塩基ごとにスペースで区切られ，60 塩基ごとに改行している。

索　　引

【あ】

【か】

—— 著者略歴 ——

北上　始（きたかみ　はじめ）
1976 年　東北大学大学院工学研究科博士前期課程修了（電子工学専攻）
1976 年　富士通株式会社入社
1978 年　株式会社富士通研究所入所
1982 年　財団法人新世代コンピュータ技術開発機構入所
1991 年　国立遺伝学研究所客員助教授
1992 年　博士（工学）（九州大学）
1994 年　広島市立大学教授
現在に至る

太田　聡史（おおた　さとし）
1995 年　北陸先端科学技術大学院大学修士課程修了（知識工学専攻）
1998 年　総合研究大学院大学博士課程修了（遺伝学専攻）博士（理学）
1998 年　財団法人遺伝学普及会情報資源研究センター研究員
1999 年　シカゴ大学(Dept. of Ecology and Evolution) ポストドクター
2001 年　科学技術振興事業団研究員（於国立遺伝学研究所）
2004 年　独立行政法人理化学研究所専任研究員
現在に至る

斎藤　成也（さいとう　なるや）
1979 年　東京大学理学部生物学科卒業
1981 年　東京大学大学院理学系研究科修士課程修了
1986 年　テキサス大学ヒューストン校生物医科学大学院修了，Ph.D.
1987 年　日本学術振興会特別研究員
1989 年　東京大学助手
1991 年　国立遺伝学研究所助教授
1992 年　総合研究大学院大学助教授
2002 年　国立遺伝学研究所教授
2002 年　総合研究大学院大学教授（兼任）
2006 年　東京大学大学院教授（兼任）
現在に至る

ビッグデータ時代の　ゲノミクス情報処理
Genomics Information Processing in the Era of Big Data

2014 年 10 月 30 日　初版第 1 刷発行　★

検印省略

著　者　北上　始
斎藤　成也
太田　聡史
発行者　株式会社　コロナ社
代表者　牛来真也
印刷所　新日本印刷株式会社

112-0011　東京都文京区千石 4-46-10

発行所　株式会社　**コロナ社**
CORONA PUBLISHING CO., LTD.
Tokyo　Japan
振替 00140-8-14844・電話(03)3941-3131(代)
ホームページ　http://www.coronasha.co.jp

ISBN 978-4-339-02485-2　（中原）　（製本：愛千製本所）
Printed in Japan