バイオインフォマティクスシリーズ 5

ゲノム配列情報解析

浜田 道昭 監修
三澤 計治 著

コロナ社

シリーズ刊行のことば

現在の生命科学においては，シークエンサーや質量分析器に代表される計測機器の急速な進歩により，ゲノム，トランスクリプトーム，エピゲノム，プロテオーム，インタラクトーム，メタボロームなどの多種多様・大規模な分子レベルの「情報」が蓄積しています。これらの情報は生物ビッグデータ（あるいはオミクスデータ）と呼ばれ，このようなデータからいかにして新しい生命科学の発見をしていくかが非常に重要となっています。

このような状況の中でその重要性を増しているのが，生命科学と情報科学を融合した学際分野である「バイオインフォマティクス」（生命情報科学，生物情報科学）です。バイオインフォマティクスは，DNA やタンパク質の配列などの，生物の配列情報をディジタル情報として捉え，コンピュータにより解析を行うことを目的として誕生しました。このような，生物の配列情報を解析するバイオインフォマティクスの一分野は「配列解析」と呼ばれます（これは本シリーズでも主要なテーマとなっています）。上述の計測機器の進歩とともに，バイオインフォマティクスはここ数十年で飛躍的に発展し，いまや配列解析にとどまらずに，トランスクリプトーム解析，メタボローム解析，プロテオーム解析，生物ネットワーク解析など多岐にわたってきています。また，必要な知識も，統計学，機械学習，物理学，化学，数学などの多くの分野にまたがっています。しかしながら，これらのバイオインフォマティクスの多岐にわたる分野を，教科書的・体系的に学ぶことができる成書シリーズは，国内外を見てもほとんどありません。

そこで，大学生，大学院生，技術者，研究者などに，バイオインフォマティクスの各分野を体系的に学習することを可能とするための教科書を提供することを目的として本シリーズを企画しました。これを実現するために，バイオイン

フォマティクス分野の最前線で活躍をしている，若手・中堅の研究者に執筆を依頼しております。執筆者の方々には，バイオインフォマティクス研究の基盤となる理論やアルゴリズムを中心に，可能な限り厳密かつ自己完結的に解説を行うようにお願いしています。そのため，本シリーズは，大学などにおけるバイオインフォマティクスの講義の教科書として活用可能であるのみならず，読者が独学する場合にも最適な書籍になっていると確信しています。

最後になりますが，本シリーズの企画の段階から辛抱強くサポートしてくださったコロナ社の皆様に御礼を申し上げます。本シリーズが，今後のバイオインフォマティクス研究さらには生命科学研究の一助となることを切に願います。

2021 年 9 月

「バイオインフォマティクスシリーズ」監修者　浜田道昭

ま　え　が　き

ゲノムとは，ある生物の持つすべての遺伝情報を意味する。ゲノム解析学は，ゲノム配列の意味を解読することにより生命の謎に迫る学問である。そしてゲノム配列情報解析は，近年，さまざまな領域で用いられている。ヒトゲノム配列の情報を利用し，個人個人の体質に合わせた治療法を選択する個別化医療も始まっている。また，病原体の研究にも活用され，疾患の原因も次々に明らかになってきている。ウイルスゲノム解析は，変異型の発見や感染ルートの解明に応用されている。さらに，感染予防の現場では，ゲノム配列決定と縁が深いPCR 技術が使われ続けている。生物学の世界では，ゲノム情報を生物の同定に使う。研究室ではなく，フィールドでゲノム配列決定をする研究が始まっている。このような背景から，本書を執筆する運びとなった。

本書の特色は，ゲノム配列情報解析に必要な，生物学の知識とプログラミングの技術を同時に記載しているところにある。プログラミング言語 Python の解説も加えた。本書は五つの章と付録から構成されている。1 章では，コンピューターがデータ処理を行う方法を紹介する。2 章では，DNA 配列決定法について解説し，その際に必要となる DNA・RNA の分子の性質を取り上げる。また，文字列検索のアルゴリズムを紹介し，ゲノム配列を復元する方法を扱う。3 章では，ペアワイズアラインメントについて解説する。その際に必要となる，分子進化学やアミノ酸配列と翻訳についてもこの章で取り上げる。4 章では，解析対象の配列が複数ある場合の，分子系統樹推定方法とマルチプルアラインメントを解説する。5 章では，ゲノム配列情報の意味を解読する方法をいくつか紹介し，特に，ホモロジーサーチ法を解説する。

本書では実用面も考慮した。コンピューターおよび統計学の用語は，日本産業規格に従った。また，特に断りのないときは，コンピューターのコマンドをタ

イプライター体で，数式をイタリック体で記載し，ローマン体の普通の英単語と区別した。ゲノム解析の現場で使われているソフトウエアの利用法も掲載した。本書で紹介した手法や解析などの一部は，Python を用いることで体験することができ，コードは https://github.com/kazumisawa/genome-infomation-analysis で利用可能である。また，実際の解析の現場で参照するため，ゲノム解析で使われるデータファイルのフォーマットを付録にて解説した。特に断りがない場合，計算には Apple M1 を搭載した MacBook Air を用いた。

ゲノム解析はしばしば地図の作成に例えられる。2 章から 4 章までで紹介する内容は，測量や航空機撮影などで，正確な図面を作ることに似ている。1 章は必要な道具の解説に相当している。そして 5 章の内容は，建物の機能や名称を書き込んでいく作業に対応する。

本書は多くの方の支援と助力に支えられた。大阪大学の加藤和貴博士には，アラインメントの仕組みを教示いただいただけでなく，本書執筆の最初から最後まで継続的な励ましをいただいた。また，北海道大学の長田直樹博士には有益なコメントをいただき，大槻涼博士にはシークエンサーの情報をいただいた。横浜市立大学の松本直通教授と，関西医科大学の日笠幸一郎研究所教授には，執筆のご支援をいただいた。この場を借りて御礼申し上げたい。

著者が Pennsylvania State University にいたときにお世話になった根井正利先生が本書を執筆中に亡くなった。根井正利先生の偉大な業績と研究への情熱から著者が学んだことは計り知れない。根井正利先生に本書を読んでいただく著者の夢は叶わなかったが，この場を借りて御礼を申し上げたい。

最後に，監修の早稲田大学の浜田道昭先生とコロナ社に感謝する。この教科書が，読者の役に立ち，ゲノム解析研究のさらなる発展に貢献できれば幸いである。

2024 年 6 月

三澤計治

本書の URL は，すべて 2024 年 6 月現在である。

目　　次

1. 文字とコンピューター

2. ゲノム配列決定と DNA

3. ペアワイズアラインメント

4. 分子系統樹推定と多重配列アラインメント

5. 機能解析と相同性検索

1 文字とコンピューター

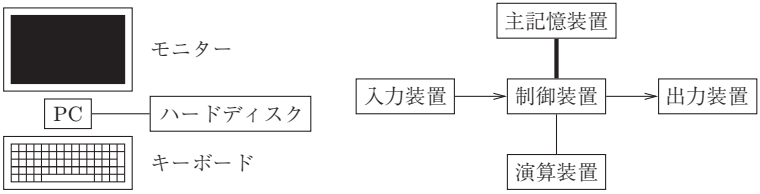
bioinformatics

ゲノム解析では，コンピューターが活用される。読者がゲノム解析を始める際には，まずは手元にあるコンピューターを利用することとなるであろう。大きな大学や研究機関には，スーパーコンピューターが設置されていることも多い。しかし，解析対象によっては，ゲノム解析専用のコンピューターの購入や開発が必要な場合もある。また，多人数で共同利用するコンピューターでは，計算資源を独占しないようにする必要がある。本章では，コンピューターの仕組みを概説し，ゲノム解析に必要な計算機資源の概略を述べる。また，本章では，DNA またはアミノ酸配列をデータベースからダウンロードして，実際に扱う。さらに，圧縮とハッシュや Python の利用方法，計算量の評価方法についても本章で解説する。

1.1 コンピューターの仕組み

1.1.1 ハードウエア

コンピューターの機械部分をハードウエアという。**図 1.1** にコンピューターのイラストを示す。図 1.1(a) に**パーソナルコンピューター**（PC）を図示した。読者の多くの方も PC に触れたことがあるであろう。PC には，多くの場合，出力装置としてモニターが設置され，入力装置としてキーボードやマウスが接続され，さらに外部記憶装置としてハードディスクが接続されている。図 (b) は図 (a) を抽象化したイラストである。PC はもちろんのこと，スーパーコンピューターからスマートフォンやタブレットまで，現在使われているコンピューター



(a)　パーソナルコンピューター　　(b)　ノイマン型コンピューター

図 1.1　コンピューター

の大部分は，**ノイマン型**（von Neumann architecture）コンピューターと呼ばれ，図 (b) に示すように，主記憶装置，制御装置，演算装置，入力装置，出力装置からなる。本書ではまずノイマン型コンピューターの各装置から解説する。

（**1**）　**主記憶装置**　　**主記憶装置**または**メモリー**（memory）は，情報を記録できる装置のうち，後述の制御装置と**バス**（bus）で直接つながっているものをいう。バスは図 1.1(b) では主記憶装置と制御装置の間の太い直線で示されている。主記憶装置には，**アドレス**（番地）がついていて，一つのアドレスに対し，一つの情報を記憶できる。主記憶装置と制御装置の間のバスの情報転送速度が遅いと，そこがボトルネックになってコンピューター全体の処理が遅くなる。これをフォン・ノイマン・ボトルネックという。

主記憶装置とは対称的に，制御装置と直接つながっていない記憶装置を外部記憶装置といい，ハードディスクや USB メモリーなどが該当する。ストレージ（storage）ともいう。ゲノム解析は大量のデータを扱うので，外部記憶装置も欠かせない装置である。外部記憶装置に記録されるデータは，多くの場合**ファイル**（file）という塊で扱われることが多い。後述するように，テキストファイルとバイナリファイルという種類がある。

現代のコンピューターでは，データの大きさを**バイト**（byte）で表す。バイトについては 1.2.1 項にて解説を行う。また，**国際単位系**（SI）で定義されている接頭辞を用いる。**表 1.1** に，大きさを表す接頭辞を解説した。コンピューターの分野では，国際単位系とは別に，$2^{10} = 1\,024$ を 1 k，$2^{20} = 4^{10} = 1\,048\,576$ を 1 M と呼ぶことがあるが，推奨されていない。

表 1.1　1 より大きな数を表す SI 接頭辞

10	10^2	10^3	10^6	10^9	10^{12}	10^{15}	10^{18}	10^{21}	10^{24}	10^{27}	10^{30}
da	h	k	M	G	T	P	E	Z	Y	R	Q
deca	hecto	kilo	mega	giga	tera	peta	exa	zetta	yotta	ronna	quetta
デカ	ヘクト	キロ	メガ	ギガ	テラ	ペタ	エクサ	ゼタ	ヨタ	ロナ	クエタ

（**2**）**制御装置**　ノイマン型コンピューターでは，命令群は主記憶装置に置かれている。ジャンプ命令がない限りは，命令群がアドレス順に読み取られ，順次実行されていく。命令の実行のタイミングを合わせるため，**クロック信号**（clock signal）が使われる。1 秒間で発生するクロック信号の数を，**動作周波数**（clock rate），またはクロック数という。動作周波数の単位は国際単位**ヘルツ**（Hz）が用いられる。動作周波数が高いほど，処理も速い。ノイマン型コンピューターのもう一つの特徴は，データと命令が同じ記憶装置に載っていることである。そのため，ノイマン型コンピューターを，**プログラム内蔵式コンピューター**（stored-program computer）と呼ぶこともある。

（**3**）**演算装置**　演算装置では計算が行われる。制御装置と演算装置を合わせて，**中央処理装置**（central processing unit；CPU）と呼ぶ。またプロセッサーとも呼ばれる。一つのプロセッサーに複数の演算装置を載せる場合，それぞれを**コア**（core）という。コンピューターの処理能力は動作周波数だけではなく，コアの数にも依存する。画像処理用に開発された**グラフィックスプロセッシングユニット**（graphics processing units；GPU）に演算を行わせる**GPGPU**（general-purpose computing on graphics processing units）**手法**も広く用いられている。

（**4**）**入出力装置**　入力装置には，図 1.1(a) に示したキーボードや，マウスなどがある。出力装置には，モニターやスピーカー，プリンターなどがある。

（**5**）**冷却装置ならびに電源**　図 1.1(b) に描かれていないものとして，冷却装置と電源がある。現代のコンピューターは半導体を通る電気で計算を行っており，電気の消費に伴って熱が発生する。そこでコンピューターは冷却される必要がある。冷却装置には大きく分けて空冷式，水冷式，油冷式がある。空

冷式冷却装置ではファンを回して風を起こし，コンピューターを冷却する。大きな騒音が出る特徴がある。水冷式は管に水を通すことで冷却を行う。油冷式は絶縁性の液体を冷却に用いる。消費電力が少ないコンピューターは，その分発熱が小さくなる。

1.1.2　ソフトウエア

コンピューターを動かすプログラムや，それにかかわるデータなどをまとめてソフトウエアという。ソフトウエアを大きく分けると，**システムソフトウエア**（system software）と**アプリケーションソフトウエア**（application software），略して**アプリ**（app）がある。

（**1**）　**オペレーティングシステム**　　システムソフトウエアはコンピューターを管理するソフトウエアの総称である。特に重要なものが**オペレーティングシステム**（operating system；OS）である。OS は，メモリー管理や計算資源の割り当て，ネット接続，ユーザーインターフェースなどを管理している。

ゲノム解析では，Windows，Linux，Mac OS を OS とするコンピューターが多く使われている。近年，仮想化技術が進み，OS 間の垣根は低くなっている。仮想化技術とは，ハードウエアの機能をソフトウエアによって実現する技術である。Windows 10 では，Windows Subsystem for Linux を用いると，Windows から Linux 環境を利用できる。その他の例としては Docker がある。

（**2**）　**ユーザーインターフェース**　　コンピューターを利用する際に，ユーザーとコンピューターの間を取り持つものがユーザーインターフェースである。ユーザーインターフェースは，**グラフィカルユーザーインターフェース**（graphical user interface；GUI）と**コマンドラインインターフェース**（command line interface；CLI）がある†。GUI は，画面に文字だけではなくアイコンやウィンドウを表示し，キーボードの他にマウスやタッチパネルで操作できるなど，操作が容易という特徴がある。また，グラフや画像表示なども得意である。

† CLI は**キャラクタベースユーザーインターフェース**（character-based user interface；CUI）とも呼ばれるが，GUI と紛らわしいため，本書では CLI とする。

Windows，Linux，Mac OS などはすべて GUI を備えている。CLI では操作はおもにキーボードで行われ，出力も文字で行われる。文字列の行（ライン）がコマンド（命令）として解釈されるため，コマンドラインと呼ばれる。CLI で重要な概念に，**標準入力**（standard input）と**標準出力**（standard output）がある。アプリが標準出力に出力する結果を，別のアプリ標準入力への入力とすることができる。この仕組みを**パイプ**（pipe）という。Linux，Windows，Mac OS のコマンドラインはすべてパイプを利用できる。

CLI は，**スクリプト**（script）を利用することで，同じ作業を何度も繰り返すときなどに活用できる。また，**データ解析パイプライン**（data analysis pipe line）とは，データの入力から解析結果の出力まで一つの流れで行われる解析の仕組みである。パイプ以外の仕組みを使うものもパイプラインと呼ばれる。

1.2 コンピューターにおける文字と符号化

多くの場合，ゲノム配列は文字列として表現される。本書では，まず，コンピューターにおける文字の扱いについて解説する。現代のほとんどのコンピューターは，電気を通す（on）通さない（off）の二つの状態を用いて，情報を処理している。一つの on/off を表す単位を 1 **ビット**（bit）という。

1.2.1 ビット列と数値

4 ビットのビット列と数値とを対応させる例を**表 1.2** に示す。4 ビットを二つ並べて 8 ビットにすることで，0 から 255 まで表現できる。10 進数はわれわれが普段使っている数値の表現方法であり，10 になるときに桁が一つ上がる。2 進数は 2 になるときに桁が一つ上がる。つまり 10 進数の 2 は 2 進数では 0b10 となる（10 進数と区別するため，先頭に 0b がつけられる）。8 進数は 8 になるときに桁が一つ上がり，先頭に 0o がつけられる。16 進数は 16 になるときに桁が一つ上がり，先頭に 0x がつけられる。表 1.2 に示すように，4 ビットのビット列は 8 進数では 2 桁，16 進数では 1 桁の数字で表される。コンピューターの

索　　引

【ひ】

【ふ】

【へ】

【ほ】

—— 監修者・著者略歴 ——

浜田　道昭（はまだ　みちあき）

2000年　東北大学理学部数学科卒業
2002年　東北大学大学院理学研究科修士課程修了（数学専攻）
2002年　株式会社富士総合研究所研究員
2009年　東京工業大学大学院総合理工学研究科博士後期課程（社会人博士）修了（知能システム科学専攻）
博士（理学）
2010年　東京大学特任准教授
2014年　早稲田大学准教授
2018年　早稲田大学教授
現在に至る

三澤　計治（みさわ　かずはる）

1995年　京都大学理学部理学科卒業
1997年　東京大学大学院理学系研究科修士課程修了（生物科学専攻）
2000年　東京大学大学院理学系研究科博士課程修了（生物科学専攻），博士（理学）
2000年　ペンシルバニア州立大学分子進化遺伝学研究所博士研究員
2003年　千葉県産業振興センター研究員・かずさ DNA 研究所共同研究員
2007年　理化学研究所次世代計算科学研究開発プログラム研究員
2013年　理化学研究所情報基盤センター研究員
2015年　東北大学東北メディカル・メガバンク機構助教
2019年　関西医科大学附属生命医学研究所講師
2021年　横浜市立大学大学院医学系研究科准教授
現在に至る

ゲノム配列情報解析

Genome Information Analysis

2024 年 8 月 15 日　初版第 1 刷発行

検印省略

監修者　浜田　道昭
著　者　三澤　計治
発行者　株式会社　コロナ社
代表者　牛来真也
印刷所　三美印刷株式会社
製本所　株式会社　グリーン

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　https://www.coronasha.co.jp

ISBN 978–4–339–02735–8　C3355　Printed in Japan　（西村）