

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

フォー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、バイオインフォー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ー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下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ー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

「バイオインフォー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ェ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が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

目 次

1. 文字とコンピューター

1.1 コンピューターの仕組み	1
1.1.1 ハードウェア	1
1.1.2 ソフトウェア	4
1.2 コンピューターにおける文字と符号化	5
1.2.1 ビット列と数値	5
1.2.2 文字と符号化	6
1.2.3 文字フォント	8
1.2.4 制 御 文 字	9
1.3 圧縮とハッシュ	9
1.3.1 符号化を利用した圧縮	9
1.3.2 圧縮の効率と情報量	12
1.3.3 ハッシュ法の利用	14
1.3.4 ハッシュ法の注意点	16
1.4 Python の利用	17
1.4.1 DNA, RNA, およびアミノ酸配列のデータベースとアクセス番号	18
1.4.2 配列データベースへのアクセス	19
1.5 計算量の評価	21
1.5.1 ランダウの記法	21
1.5.2 文字列探索	22
章 末 問 題	24

2. ゲノム配列決定と DNA

2.1 DNA の 性 質	25
2.1.1 核 酸 塩 基	26
2.1.2 糖・リン酸バックボーン	27
2.1.3 ワトソン・クリック塩基対	28
2.1.4 二 本 鎖 DNA	29
2.2 複 製	31
2.2.1 DNA の半保存的複製	32
2.2.2 岡崎フラグメントと DNA 複製の限界	34
2.2.3 PCR	34
2.2.4 PCR 検 査	36
2.3 配 列 決 定	36
2.3.1 電 気 泳 動 法	37
2.3.2 サザンブロッティング	37
2.3.3 制 限 酵 素	37
2.3.4 マイクロアレイ法	38
2.3.5 第一世代シークエンサー	38
2.3.6 第二世代 (次世代) シークエンサー	40
2.3.7 第三世代シークエンサー	41
2.4 k-mer 法	44
2.4.1 k-mer の 例	44
2.4.2 k-mer のメモリー使用量	46
2.4.3 ゲノム中に一度しか現れない単語	47
2.4.4 ゲノムサイズ推定	51
2.5 デノボアセンブリ	51
2.5.1 デブラングラフ	52
2.5.2 順 序	55
2.5.3 橋検出と lowlink	57
2.5.4 フラワーリーのアルゴリズム	57

2.5.5	デノボアセンブリソフトウェアの例	61
2.5.6	オーバーラップレイアウトコンセンサス法	61
2.6	マ ッ ピ ン グ	61
2.6.1	巡回ソートとバロウズ・ウィーラー変換	62
2.6.2	FM インデックスと LF マッピング	65
2.6.3	FM インデックスを用いた逆バロウズ・ウィーラー変換	68
2.6.4	FM インデックスを用いた文字列検索	69
2.6.5	接 尾 辞 配 列	73
2.6.6	接尾辞配列の工夫	74
2.6.7	マッピングソフトウェアの例	75
2.6.8	ヒトゲノム参照配列	75
2.6.9	バロウズ・ウィーラー変換を利用した圧縮	76
2.7	染 色 体	78
2.7.1	セントロメア	78
2.7.2	体細胞分裂と染色体	78
2.7.3	減数分裂・性・遺伝的組換え	79
2.7.4	突 然 変 異	80
2.7.5	シンテニーと遺伝子重複	81
2.7.6	エピゲノムと DNA メチル化	81
2.8	転 写 と RNA	82
2.8.1	RNA を構成する要素	82
2.8.2	転 写	84
2.8.3	スプライシング	85
2.8.4	エクソンとイントロン	85
2.8.5	ノーザンブロッティング	85
2.8.6	テロメアと逆転写酵素	86
2.8.7	ウ イ ル ス	86
2.8.8	RNA-Seq	87
章 末 問 題		88

3. ペアワイズアラインメント

3.1 最適化問題としてのペアワイズアラインメント推定	90
3.1.1 分子進化	90
3.1.2 目的関数と最適化問題	91
3.2 DNA 間のスコア	92
3.3 アミノ酸間のスコア	93
3.3.1 アミノ酸の構造	94
3.3.2 タンパク質を構成するアミノ酸	94
3.3.3 タンパク質を構成するアミノ酸の特性	95
3.3.4 アミノ酸変異の表記	100
3.3.5 アミノ酸ペアにスコアを与える方法	100
3.3.6 BLOSUM スコア	101
3.4 最適化問題とその難しさ	103
3.5 動的計画法	105
3.5.1 動的計画法とニードルマン-ヴァンシュ (1970) のアルゴリズム	105
3.5.2 アフィンギャップスコア	114
3.5.3 後藤 (1982) のアルゴリズム	115
3.6 加藤ら (2002) のアルゴリズムと MAFFT	120
3.6.1 連続化と正規分布	121
3.6.2 ずれと相互相関関数	122
3.6.3 DNA 配列およびアミノ酸配列の 2 次元ベクトル化	123
3.6.4 高速フーリエ変換 (FFT) と相互相関関数	127
3.6.5 相互相関関数のピークと相同性	129
3.6.6 場所の確認	134
3.6.7 動的計画法 1 回目	137
3.6.8 動的計画法 2 回目	138
3.7 Biopython によるアラインメントアプリケーションの呼び出し	142
3.7.1 Biopython からの MAFFT 呼び出し	143

3.7.2 Biopython からの Clustal W 呼び出し	143
------------------------------------	-----

章末問題	144
------	-----

4. 分子系統樹推定と多重配列アラインメント

4.1 進化距離と進化速度	147
---------------	-----

4.1.1 置換	148
4.1.2 確率過程	150
4.1.3 ポアソン距離	152
4.1.4 DNA 配列から求める進化距離	153
4.1.5 DNA における遷移速度行列と塩基置換速度の推定	154
4.1.6 DNA 配列比較による進化距離推定の今後の発展	158
4.1.7 翻訳のプロセスとコドン表	159
4.1.8 アミノ酸配列から求める進化距離	165
4.1.9 アミノ酸配列比較による進化距離推定の今後の発展	168
4.1.10 k-mer 法を利用したアラインメントに頼らない進化距離推定法	169

4.2 進化距離推定をもとにした分子系統樹再構築法	170
---------------------------	-----

4.2.1 距離行列	170
4.2.2 Python による系統樹の表示	171
4.2.3 非加重結合法 (UPGMA)	175
4.2.4 近隣結合法 (NJ 法)	175
4.2.5 非加重結合法と近隣結合法の比較	177
4.2.6 距離行列と genotype value decomposition	177
4.2.7 長枝誘引	180
4.2.8 結合した OTU と残りの OTU の間の距離	182
4.2.9 ブートストラップ検定	184
4.2.10 PartTree 法	186
4.2.11 配列数と計算時間の関係	191
4.2.12 分類群の利用	193
4.2.13 オースロガスとパラロガス	194
4.2.14 真核生物の細胞と細胞内共生説	196
4.2.15 ミトコンドリアのコドン表	197

4.3 多重配列アラインメント 198

 4.3.1 逐次追加法 202

 4.3.2 ガイド樹法 204

 4.3.3 外部からの系統樹読み込み 204

 4.3.4 反復改善法 209

章末問題 209

5. 機能解析と相同性検索

5.1 統計的仮説検定 210

 5.1.1 実験計画法 211

 5.1.2 超幾何分布 214

 5.1.3 帰無仮説と第1種の過誤 215

 5.1.4 フィッシャーの正確確率検定 216

 5.1.5 Z 検定 217

 5.1.6 カイ2乗検定 218

 5.1.7 対立仮説と第2種の過誤 219

5.2 ゲノムワイド関連解析による疾患関連遺伝子の探索 220

 5.2.1 ゲノムワイド関連解析 220

 5.2.2 ファミリーワイズエラー率とボンフェローニの補正 221

 5.2.3 GWAS カタログ 222

5.3 相同性検索 222

 5.3.1 スミス-ウォーターマン (1981) のアルゴリズム 223

 5.3.2 相同性検索とセグメントスコア 225

 5.3.3 相同性検索の統計検定 225

 5.3.4 BLAST 227

 5.3.5 相同性のある配列が見つからない場合 228

5.4 アノテーション 229

5.5 ゲノム配列の変化とタンパク質の機能の関係についての今後の展開 229

章末問題 231

付 録	232
A.1 Python の基本文法	232
A.1.1 Python のデータ型	232
A.1.2 Python とオブジェクト指向プログラミング	233
A.1.3 Python の構文	233
A.2 ファイルフォーマット	235
A.2.1 FASTA フォーマット	235
A.2.2 FASTQ フォーマット	235
A.3 ベクトルと行列	236
A.3.1 ベクトルの成分表示	237
A.3.2 ベクトルの和	237
A.3.3 ベクトルの大きさ	238
A.3.4 ベクトルの内積	238
A.3.5 ベクトルの内積の成分計算	238
A.3.6 直 交	239
A.3.7 行 列	239
A.3.8 行 列 の 積	240
A.3.9 2次元のベクトルと面積と行列式	241
A.4 解 析 関 数	241
A.4.1 微分とベルヌーイの不等式	242
A.4.2 指 数 関 数	245
A.4.3 対数関数と自然対数の底 e	245
A.4.4 三角関数と円周率 π	250
A.5 統計にかかわる関数	256
A.5.1 標 本	256
A.5.2 同時確率と条件つき確率	256
A.5.3 確率密度関数	256
A.5.4 累積分布関数	257
A.5.5 確率変数の期待値	258
A.5.6 分 散	258
A.5.7 k 次のモーメント	258

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 ハードウェア

コンピューターの機械部分をハードウェアという。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

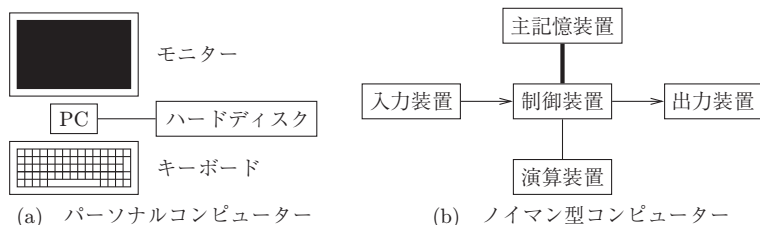


図 1.1 コンピューター

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、 $2^{10} = 1024$ を 1k、 $2^{20} = 4^{10} = 1048576$ を 1M 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。またプロセッサとも呼ばれる。一つのプロセッサ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 ソフトウェア

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ェアという。ソフトウェアを大きく分けると、システムソフトウェア (system software) とアプリケーションソフトウェア (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 システムソフトウェアはコンピューターを管理するソフトウェアの総称である。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、ハードウェアの機能をソフトウェ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 pipeline）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

【あ】	
アクセッション番号	18, 140, 193, 195
アスキーアート	171
アスキーコード	7, 171, 174
アスパラギン	95, 97
アスパラギン酸	95, 97
アダマール行列	155
圧縮	9
アデニル酸キナーゼ	196
アデニン	26, 83
アデノシンーリン酸	83
アデノシン三リン酸	83
アドレス	2
アノテーション	210, 229
アフィンギャップスコア	114
アプリ	4
アプリケーション	
ソフトウエア	4
アミノ基	94, 96
アミノ酸	94
アミノ酸含量	168
アミノ酸配列	94
アラインメント	89
アラニン	95, 96
アルギニン	95, 97
アレル	148
アレル頻度	148
アンチコドン	162
アンプリコン	36
【い】	
イオン半導体シークエンス	41

鋳型 DNA	33
閾値	131, 216, 225
イソロイシン	95, 98
一塩基多型	148
一次構造	94
一分子リアルタイム	
シークエンシング法	42
一般時間反転可能モデル	158
一本鎖 RNA	84
一本鎖 DNA	27
遺伝暗号	159, 197
遺伝子	164
遺伝子重複	81, 165, 194
遺伝的浮動	148
遺伝的連鎖	80
インデント	233
イントロン	85

【う】

ウイルス	86, 100
ウラシル	83

【え】

エクソン	85
枝	90
エバネッセント光	42
エピゲノム	81
絵文字	8
塩基	25, 82
塩基性アミノ酸	97
塩基対	28
円周率	122, 255
エンハンサー	84

【お】

オイラーの等式	254
大文字	7
岡崎フラグメント	34
オス駆動進化仮説	80
オーソログス	194
オーバーライド	54, 233
オーバーラップレイアウト	
コンセンサス法	52, 61, 76, 144
オフサイドルール	233
オブジェクト	233
オブジェクト指向	
プログラミング	53, 233
オペレーティングシステム	4

【か】

改行	9
改行コード	9
改行文字	234
外群	194, 195
開始コドン	161
開始点	70
ガイド樹	204
ガイド樹法	204
カイ 2 乗検定	216, 218, 219, 221
カイ 2 乗分布	218
核酸	25
核酸塩基	25, 84
学名	36
確率過程	150
確率行列	151
確率密度関数	121, 226, 256

可視光 29, 38
 カプセル化 233
 可変長符号 10
 下流 34
 カルボキシ基 94, 96
 完全削除 147
 含硫アミノ酸 98
 関連解析 220, 221

【き】

木 56, 90
 偽遺伝子 164
 棄却 216
 偽常染色体部位 80
 期待値 13, 228, 258
 機能解析 210
 木の深さ 56, 61
 帰無仮説 216, 219
 逆行列 154, 240
 逆転写 80, 165
 逆転写酵素 86
 逆バロウズ・ウィーラー変換 68
 逆フーリエ変換 127, 266
 逆平行 30
 ギャップ 91, 101, 134
 ギャップエクステンション
 ペナルティ 114
 ギャップオープニング
 ペナルティ 114
 ギャップスコア 92, 107
 キャラクターベースユーザー
 インターフェース 4
 狂犬病ウイルス 87
 鏡像異性体 96
 共通祖先 146, 178
 行ベクトル 237
 共有派生形質 179
 極座標系 257
 極性 125, 199
 虚数 250
 距離行列 169, 170
 近隣 175

近隣結合法 170

【く】

グアニン 26, 83
 空間計算量 21
 偶現表 211
 空白文字 7, 233
 クエリ 22, 224, 226
 クラス 53
 グラフィカルユーザー
 インターフェース 4, 173, 229
 グラフィックスプロセッシング
 ユニット 3
 グラフ理論 52
 繰り返し配列 78
 グリシン 95, 96
 グリフ 8
 グルタミン 95, 97
 グルタミン酸 95, 97
 クロック信号 3
 グローバルアラインメント 222
 ゲンベル分布 225

【け】

蛍光 38
 蛍光物質 38, 39
 計算量 21
 継承 55, 233
 系統樹 90
 欠失 80, 81, 91, 148, 163
 ゲノムアセンブリ 51
 ゲノム編集 37
 ゲノムワイド関連解析 210, 221
 減数第一分裂 79
 減数第二分裂 79
 減数分裂 79
 検定統計量 216

【こ】

コア 3

校正 34
 合成によるシークエンシング 41
 高速フーリエ変換 127
 後退辺 56
 紅茶の違いのわかる婦人 211
 構文解析 193
 国際純正・応用化学連合 94
 国際単位系 2
 古細菌 36, 85, 160
 骨格構造式 26
 固定 149
 固定長符号 10
 コドン 160
 コドンの縮重 161
 コドン表 160, 197
 コマンドライン
 インターフェース 4
 小文字 7
 コロナウイルス 87, 159
 コントロール群 220

【さ】

座位 148, 221
 再帰呼び出し 58, 188
 最小値を与える引数 175
 最大値安定性 260
 最短経路問題 103, 104
 最適化手法 174
 最適化問題 89
 サイト 51, 148
 細胞小器官 196
 細胞内共生説 197
 サザンブロッティング 37, 86
 サブクラス 54, 233
 サーマルサイクリング 35
 三角関数 250, 257
 サンガー法 38
 残基 94, 96
 参照配列 61
 参照配列ガイドつき
 アセンブリ 52, 61

酸性アミノ酸 97

サンプルサイズ 256

【し】

紫外線 38

時間計算量 21

時間反転可能 158, 159

シークエンシングアダプター 43

次元の呪い 145

辞書型 44, 232

辞書式順序 7

指数関数 35, 155, 245

指数分布 226

システイン 95, 98

システムソフトウエア 4

システムの過誤 228

自然選択 149

自然対数 247

下三角行列 171

疾患関連遺伝子 220

疾患群 220

実験計画法 211

シトシン 26, 82, 83

シフト則 264

姉妹染色分体 78

シャノンのエントロピー 13

シャノンの情報源符号化定理 13

シャルガフの法則 28

ジャンク DNA 165

周期関数 253

重合体 44

終止コドン 161

集団遺伝学 148

周辺度数 211

終了点 70

主記憶装置 2

縮重 162

主鎖 95

出生死亡過程 165

巡回グラフ 56

巡回ソート 62

順序 55

条件つき確率 256

詳細つり合い 158, 197

消失 149

状態 150, 177

状態遷移図 150

情報量 12, 13, 102

常用対数 249

上流 34, 84

ショートリード 36

資料 256

進化距離 147

真核生物 78, 160

新型コロナウイルス 23

新型コロナウイルス感染症 23

親水性アミノ酸 95, 97

真正細菌 36, 85, 160

シntenニー 81, 195, 229

【す】

スクリプト 5

スクリプト実行モード 17

スコア 92

スコア行列 105

スタンダードなコドン表 160, 198

スーパークラス 233

スプライシング 85

スライス 111

スループット 39

ずれ 122

スレオニン 95, 96

【せ】

正規化線形関数 223

正規分布 121, 122, 217

制御文字 9

制限酵素 37

斉時性 151

成分 125

正方行列 239

セグメント 225

セグメントスコア 225

セグメントペア 225

節 52, 90

接尾辞 63

接尾辞配列 64, 73

セリン 95, 96

ゼロモード導波路 42

遷移 150

遷移確率 150

遷移速度行列 154, 155

全エクソーム解析 85, 144

漸化式 104, 116

線形性 263

全ゲノム配列決定 25

染色体 78

先頭移動法 76

セントラルドグマ 84

セントロメア 78

【そ】

相互相関関数 123, 267

操作的分類単位 90

相同 79, 89, 129

相同組換え 79

相同性 79, 102

相同性検索 210

相同染色体 79

相同領域候補 129

挿入 80, 81, 91, 148, 163

挿入・欠失 91

相補的 DNA 30, 87

相補的配列 30

相補累積分布関数 257

側鎖 95

疎水性アミノ酸 95, 100

素数 48

祖先状態 150

【た】

第一世代シークエンサー 38

第1種の過誤 216

対角化 240

対角化可能 156, 240

- 対角行列 154, 239
 第三世代シークエンサー 42, 61
 対象 22, 224, 226
 対称行列 102, 155, 239
 対数 13, 102, 246
 対数オッズスコア 101
 対数関数 245, 247
 体積 125, 199
 第2種の過誤 219
 第二世代シークエンサー 40
 対立仮説 219
 対話モード 17
 ターゲット 47
 ターゲット配列決定 25
 多次元動的計画法 145
 多重配列アラインメント 102, 145, 222
 畳み込み 268
 タブ区切り 9, 124, 171
 タブ文字 9, 11, 233
 タブル 20
 単位行列 155, 240
 単系統群 178
 単語 44
 短枝誘引 182
 単純アミノ酸 95
 単数体 80
 タンパク質 84, 89
 タンパク質を構成するアミノ酸 94
 単量体 44
- 【ち】**
 違いのあるサイトの数 147
 違いのあるサイトの割合 147
 置換 91, 148, 149
 置換行列 153
 置換率 152
 逐次追加法 202
 チミン 26
 中央処理装置 3
 中心極限定理 219, 269
- 中立モデル 149
 超幾何分布 215
 長枝誘引 182
 重複 80, 91
 チロシン 95, 99
- 【て】**
 低エントロピー領域 14
 デイラックのデルタ超関数 122
 デオキシリボ核酸 25
 デオキシリボース 27
 テキストエディタ 9
 データ解析パイプライン 5
 デノボアセンブリ 52
 デブラングラフ 52
 テロメア 86
 テロメレース 86, 141
 転位 93, 155, 157, 163
 転移 RNA 84, 162
 転移因子 80
 展開 9
 転換 93, 155, 157
 電気泳動法 28, 37, 39
 転写 84
 転写因子 84
 転置 237
 転置行列 239
 点突然変異 80, 148
 天然痘ウイルス 87
 伝令 RNA 84
- 【と】**
 同義置換 163, 165
 統計的過誤 228
 統計的仮説検定 210, 216
 統計的検定 216, 219
 凍結された偶然説 163
 動原体 79
 動作周波数 3
 同時確率 256
 到達可能最小 P 値 217
 動的型づけ 232
- 動的計画法 105, 119, 137, 145
 糖・リン酸バックボーン 27
 特殊メソッド 54, 233
 特性関数 264
 突然変異 78, 80, 147
 トーナメント形式 204
 トランスポゾン 80
 トリプトファン 95, 99
 トレースバック 110
- 【な】**
 内積 123, 178
 ナット 13
 ナノボア法 28, 42
 ナンセンス置換 163
- 【に】**
 二価染色体 79
 二重らせん 29
 二倍体 79, 149
 二本鎖 DNA 29
 日本産業規格 7, 219
 尿酸酸化酵素 164
 認識部位 37
- 【ぬ】**
 スクレオシド 31
 スクレオチド 31, 84
- 【ね】**
 根 55, 56, 91
 ネイピア数 246
- 【の】**
 ノイマン型 2
 ノーザンブロッティング 85
- 【は】**
 葉 205
 パーソナルコンピューター 1
 倍数体 80
 排他的 221

バイト 2, 6
 バイト並び順 6
 パイプ 5
 ハイブリダイズ 30, 86
 ハイブリダイゼーション 30
 パイロシークエンシング 40, 41
 バウムクーヘン積分 258
 橋 57
 バス 2
 派生状態 150
 パッケージ 17
 ハッシュ 14
 ハッシュ関数 14
 ハッシュ衝突 16, 48
 ハッシュ値 14
 ハッシュテーブル 47
 ハッシュ法 14, 47
 鳩の巣原理 48
 ハフマン符号化 10, 77
 パーミュテーション 228
 パラメーター 122
 パラログス 194
 バリエント 148
 バリエントコール 149, 229
 バリン 95, 98
 パロウズ・ウィーラー変換 62
 反復改善法 146, 209
 番兵 62, 107, 109
 半保存的複製 33

【ひ】

非加重結合法 170, 183, 186
 非巡回グラフ 56
 ヒスチジン 95, 97
 左確率行列 151
 ビッグエンディアン 6
 ビット 5, 13, 212
 非同義置換 163, 165
 ヒト T リンパ好性ウイルス 87
 ヒト免疫不全ウイルス 87

非復元抽出 212
 微分可能 242
 評価関数 89
 標準化 127
 標準出力 5
 標準正規分布 217
 標準入力 5
 標準偏差 259
 標本 256
 標本空間 212, 258
 標本の大きさ 256
 標本分散 259
 標本平均 259
 ビリミジン塩基 26, 83, 93
 ピロリン酸 32, 40, 41, 83

【ふ】

ファイル 2
 ファミリーワイズエラー率 221
 フィッシャーの正確確率検定 216, 221
 フェニルアラニン 95, 99
 不確定性関係 264
 深さ 51
 深さ優先探索 55
 復元抽出 184
 複製 32, 90
 複製フォーク 32
 複素共役 267
 符号位置 7, 236
 符号化 6, 10, 108
 復帰 9
 復帰置換 150
 ブートストラップ検定 184
 ブートストラップ値 184, 195
 プライマー 35, 47
 フラリーのアルゴリズム 57
 フーリエ反転公式 266
 フーリエ変換 127, 263
 プリン塩基 26, 83, 93

ブルームフィルター 51, 61
 フレームシフト置換 163
 プログラム内蔵式
 コンピューター 3
 プロセス型偽遺伝子 164
 フローセル 40
 プロモーター 84
 プロリン 95, 96
 分割 178
 分割統治法 138, 139, 212
 分割表 211
 分岐点 90
 分散 258
 分子系統樹 90
 分子系統樹再構築 146
 分枝鎖アミノ酸 98
 分子進化 90
 分子進化速度 147
 分子ふるい 37

【へ】

ベアワイズアラインメント 89, 198, 222
 ベアワイズ削除 147
 ハイフリック限界 34
 ペプチド結合 94
 ヘルツ 3
 ベルヌーイの不等式 244
 辺 52, 90
 変異シグニチャー解析 159
 変異スペクトラム解析 159
 変更可能 232
 変更不可能 232

【ほ】

ポアソン距離 153
 ポアソン分布 152
 芳香族アミノ酸 99
 ポリペプチド 94
 ポリメラーゼ連鎖反応 34
 ボンフェローニの補正 221
 翻訳 84

【ま】	
マイクロアレイ法	38, 86
マイナーアレル	148
マイナーアレル頻度	148
マキサム・ギルバート法	40
マッピング	62
マルコフ過程	151
マルコフ連鎖	151
マルチバイト	6
【み】	
右確率行列	151
三澤式コドン表	161
ミスセンス置換	163
ミトコンドリア	14, 196
ミドリムシ	85
【む】	
無限サイトモデル	147, 153, 178
無根系統樹	172
【め】	
メソッド	172
メチオニン	95, 98
メチル基	82
メチルシトシン	82
メモリー	2
メンデルの法則	148

【も】	
目的関数	89, 91, 174
文字コード	7, 124, 173
文字列	20
モーメント	258, 265
【ゆ】	
有意	216
有意水準	216
有害	149
有根系統樹	91, 194
有糸分裂	79
有利	149
ユークリッドノルム	238
【よ】	
葉緑体	85
【ら】	
ラギング鎖	34
ランダウの記法	21
ランダム化試験	211
【り】	
リジン	95, 97
リスト	20
リーディング鎖	34
リード	36, 62
リトルエンディアン	6

リピート	78
リファレンス配列	202
リボ核酸	82
リボソーム	23, 159
リボソーム RNA	84
隣接語	227
隣接コドン	162, 165
【る】	
類似性	263, 269
累積分布関数	226, 257
ルシフェラーゼ	41
【れ】	
列ベクトル	237
レトロウイルス	87
レトロトランスポゾン	81
連結グラフ	55
連鎖	80

【ろ】	
ロイシン	95, 98
ローカルアラインメント	222
ロングリード	36

【わ】	
ワトソン・クリック塩基対	28, 158

【A】	
API	19
ATP	41, 83, 196
ATP 合成酵素	196
【B】	
Biopython	18
BLOSUM	101

【C】	
C 末端	94
Cas9 mediated enrichment 法	43
Catch-22 situation	146
CCS 法	43
CLI	4, 172
CpG hypermutability	158

【D】	
D-アミノ酸	96
dAMP	31
dATP	31
dbSNP	148
dCMP	31
dCTP	32
DDBJ	18
dGMP	31
dGTP	31

DNA	25	INSDC	18	PCR	34, 47, 195
DNA ウイルス	86			PCR 検査	36
DNA 合成酵素	32	[J]		Phred クオリティスコア	
DNA 複製酵素	32	JC20	166		235, 249
DNA 複製装置	34	JTT 法	166	Python	17, 232
DNA メチル基転移酵素	82	Jukes-Cantor モデル	157	[R]	
dNMP	31			RNA ウイルス	86
dNTP	31	[K]		RNA エディティング	
dTMP	31	k-mer	44, 169		85, 159
dTTP	31	k-mer 法	44	RNA 配列決定	87
[E]		[L]		RNA ポリメラーゼ	84
EBI	18	L-アミノ酸	96	RNA-Seq	87
ENA	18	LF マッピング	65	rRNA	84, 87, 159
[F]		[M]		[S]	
FASTA	172	MAFFT	120	SARS-CoV-2	159
FASTQ	235	mRNA 前駆体	85	[T]	
FM インデックス	65	MSP	225	T2T consortium	11, 50, 86
[G]		MSS	225	[U]	
GATK	149	[N]		UTF-8	8, 172, 173
GenBank	18	N 末端	94	[Z]	
GPGPU 手法	3	NCBI	18, 227	Z 検定	216, 217
gRNA	37	newick standard format			
GUI	4		188, 190, 204, 208	[ギリシャ文字]	
[H]		[P]		ϵ -N 論法	21
HSP	225	p 距離	147, 171	ϵ - δ 論法	242
[I]		PAA	94, 126		
indel	91	PAM	38		
		PartTree 法	186, 188		

—— 監修者・著者略歴 ——

浜田 道昭 (はまだ みちあき)	三澤 計治 (みさわ かずはる)
2000年 東北大学理学部数学科卒業	1995年 京都大学理学部理学科卒業
2002年 東北大学大学院理学研究科修士課程修了 (数学専攻)	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)
2002年 株式会社富士総合研究所研究員	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)
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)	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
2010年 東京大学特任准教授	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
2014年 早稲田大学准教授	2013年 理化学研究所情報基盤センター研究員
2018年 早稲田大学教授	2015年 東北大学東北メディカル・メガバンク 機構助教
現在に至る	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

© Kazuharu Misawa 2024

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)



＜出版者著作権管理機構 委託出版物＞

本書の無断複製は著作権法上での例外を除き禁じられています。複製される場合は、そのつど事前に、出版者著作権管理機構（電話 03-5244-5088, FAX 03-5244-5089, e-mail: info@jcopy.or.jp）の許諾を得てください。

本書のコピー、スキャン、デジタル化等の無断複製・転載は著作権法上での例外を除き禁じられています。購入者以外の第三者による本書の電子データ化及び電子書籍化は、いかなる場合も認めていません。落丁・乱丁はお取替えいたしません。