

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。シーケンサーの登場は、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。シーケ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、シーケ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、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を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則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ヤ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。

ii ま え が き

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

1.1	ビッグデータとしてのゲノム情報	1
1.2	生命科学と情報科学	3
1.3	塩基配列データベース	3
1.4	モチーフデータベース	8
1.5	タンパク質立体構造データベース	9
1.6	構造分類データベース	12
1.6.1	SCOP	12
1.6.2	CATH	14
1.7	さまざまなデータベース	15
1.8	オントロジー	16
	引用・参考文献	18

2. ゲノム配列の決定と解析

2.1	次世代シーケンサーとは	21
2.1.1	サンガー法	21
2.1.2	次世代シーケンス法	22
2.2	塩基配列決定における情報学的側面	23
2.2.1	ショットガン法	23
2.2.2	リシーケンシング	24
2.3	相同性検索と多重整列化	26
2.3.1	相同性とは	26
2.3.2	相同性検索の原理	26
2.3.3	BLAST	27

2.3.4	2 個の配列を整列化する原理	27
2.3.5	多重整列化の原理	29
2.3.6	MISHIMA	30
2.3.7	長大なゲノム配列間の相同性解析	32
2.4	塩基置換数の推定	32
2.4.1	1 変数法による推定	32
2.4.2	2 変数法による推定	34
2.4.3	同義置換数と非同義置換数の推定	35
2.5	遺伝子予測に基づくゲノムアノテーション	37
2.6	SNP の同定と解析	39
	引用・参考文献	42

3. モチーフの表現と抽出

3.1	類似性検索	44
3.1.1	非類似度に基づく検索と整列化	45
3.1.2	類似度に基づく検索と整列化	50
3.2	多重整列化	61
3.2.1	階層併合的クラスタリング	62
3.2.2	Feng-Doolittle 累進法	63
3.2.3	プロファイル累進法	64
3.3	プロファイルと類似性検索	68
3.3.1	モチーフの表現法	68
3.3.2	正規表現の導出法	75
3.3.3	プロファイルを用いた類似性検索	78
3.4	プロファイル HMM の導出法	86
3.4.1	整列行列からプロファイル HMM の導出	86
3.4.2	Baum-Welch アルゴリズム	88
3.5	頻出な類似部分配列の抽出	92
3.5.1	列挙法	93
3.5.2	ギブスサンプリング法	95
3.5.3	探索問題の効率的解法	99
3.6	ネットワークモチーフの抽出	101
3.6.1	ネットワークモチーフ	102
3.6.2	開近傍と排他的近傍	102

3.6.3 連結部分グラフの列挙	103
3.6.4 グラフの同型性判定	104
3.6.5 ランダム化グラフ	106
引用・参考文献	107

4. 分子進化系統樹の推定

4.1 系統樹と系統ネットワークの数学的性質	111
4.1.1 系統樹の基礎的事項	111
4.1.2 樹形の表現方法	113
4.1.3 樹形と樹形のあいだの関係	115
4.1.4 系統樹で表現できない関係と系統ネットワーク	115
4.2 系統樹の生物学的性質	116
4.2.1 個体の系図と遺伝子の系図	116
4.2.2 遺伝子の系図と種の系統樹	117
4.2.3 遺伝子の系統樹：種分化と遺伝子重複の混合	118
4.2.4 さまざまな系統樹概念	119
4.3 距離行列からの分子系統樹の推定法	120
4.3.1 系統樹作成法の分類	120
4.3.2 進化速度一定を仮定した UPGMA	121
4.3.3 近隣結合法	122
4.3.4 その他の距離行列法	127
4.4 塩基配列やアミノ酸配列の多重整列化からの分子系統樹の推定法	129
4.4.1 最大節約法	129
4.4.2 最尤法	133
4.5 系統ネットワーク	138
4.5.1 塩基配列から系統ネットワークを作成する方法	138
4.5.2 距離行列データから系統ネットワークを推定する方法	140
引用・参考文献	141

5. 新しい運動機能解析

5.1 ミクロとマクロをつなぐもの	144
5.2 運動機能解析の歴史	147
5.3 生体力学	149

5.4	神経筋骨格モデル	152
5.5	逆運動学と逆動力学	155
5.6	バーンスタイン問題	157
5.7	順動力学とシミュレーション	159
5.8	体性感覚とホムンクルス	161
5.9	遺伝子型と表現型	164
5.10	ゲノムと進化生体力学	168
	引用・参考文献	173

6. 高速ビッグデータマイニングへの展開

6.1	タンパク質立体構造	180
6.2	類似構造検索	181
6.2.1	平均二乗誤差	181
6.2.2	二重動的計画法	184
6.2.3	CMO 問題	187
6.3	タンパク質の構造や機能の予測	190
6.3.1	アミノ酸配列からの構造予測	191
6.3.2	構造からの機能予測	193
6.3.3	機械学習と予測	194
6.4	分子動力学法	204
6.5	高速化技術	207
6.5.1	サフィックス木の構築と検索	208
6.5.2	座標配列に対するサフィックス木	209
6.5.3	バッファ管理システム	212
6.5.4	並列処理	213
6.6	Mathematica の並列処理	216
	引用・参考文献	218
	索引	223

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。

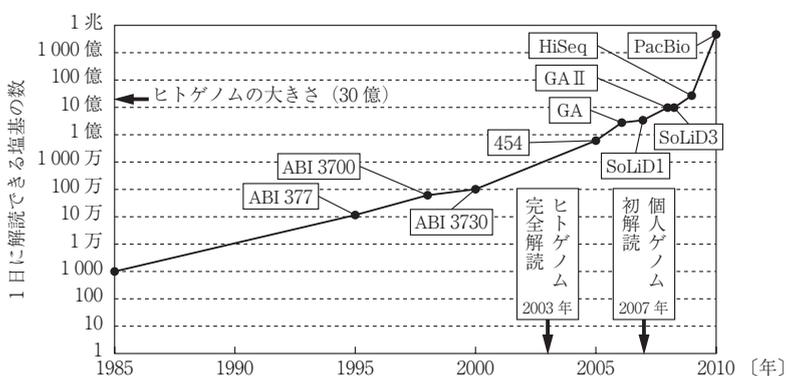
本書で注目しているゲノミクス (genomics, ゲノム科学) は、1980年代に出現した概念である。ゲノミクス分野で情報処理が注目されるようになったのは、1990年代にヒトゲノム解析計画⁴⁾ によるゲノム情報の解読が開始してからである。近年、この解読されたデータ

† 肩付き数字は、各章末の引用・参考文献番号を表す。

2 1. ゲノム情報のデータベース

は、ビッグデータの一つに分類されることからわかるが、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ケンサー (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きつき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1000塩基程度であったが、2000年にはその1000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ケ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：水島一菅野純子，菅野純夫：次世代シーケ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，榮研化学（2011）を転載。

図 1.1 DNA シーケンサーの性能向上

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、^{がん}癌などの難病の解明、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

1

2

3

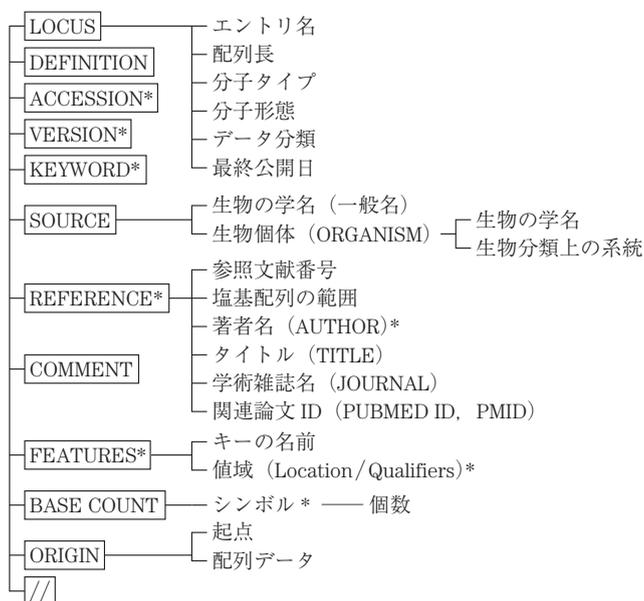
4

5

6

関で設立されているが、2005年以來、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

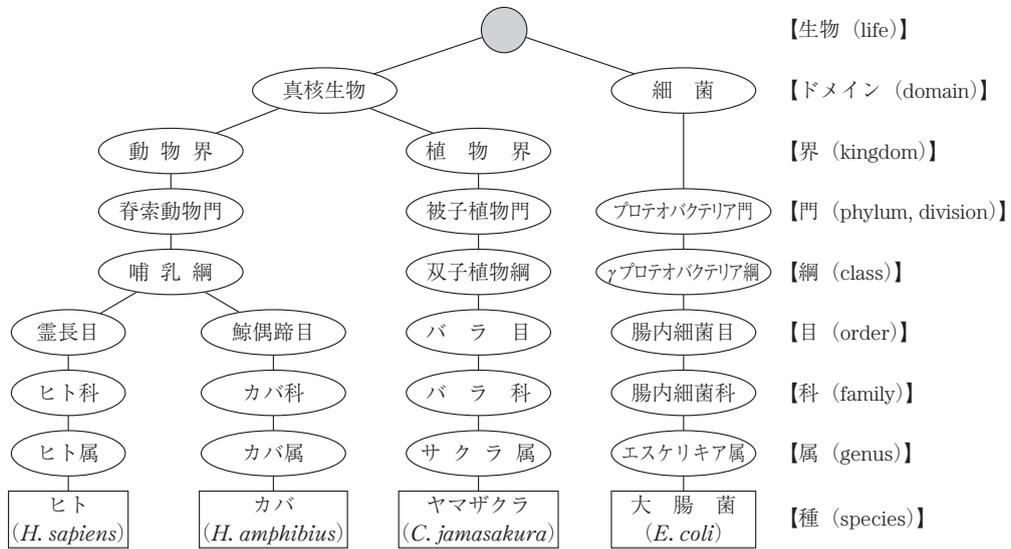
フラットファイル (flat file) とは、1行を1レコードとするプレーンテキスト (plain text) の集まり、あるいは、バイナリ (binary) を保存するファイルを意味する。レコードは、フィールドをデリミタ等の記号で区切った構造になっている。DDBJ フラットファイルフォーマットでは、プレーンテキスト形式のレコード間を二つのスラッシュ (slash) 記号「//」で区切っている。プレーンテキストとは、文字ごとの色や形状、文章に含まれる図などといった情報を含まない文字列形式のコンテンツを意味する。

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を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

<p>【あ】</p> <p>曖昧文字 70</p> <p>アーキテクチャ 14</p> <p>アクセション番号 5</p> <p>アブニシオ法 192</p> <p>アフィンギャップスコア 53</p> <p>アミノ酸置換行列 53</p> <p>アミノ酸配列 6</p> <p>網目 116</p> <p>網目構造 138</p> <p>アルツハイマー病患者 70</p> <p>アレイ構造 101</p> <p>案内木 61, 63, 64</p> <p>移住 101</p> <p>移住間隔 101</p> <p>移住率 101</p> <p>異種性 17</p> <p>異種相同 119</p> <p>位置依存スコア行列 8, 68, 78, 96</p> <p>位置依存スコア行列 PSSM 71</p> <p>一塩基多型 39</p> <p>位置スコア関数 192</p> <p>一致状態 73, 83</p> <p>一致性 133</p> <p>一致列 74</p> <p>遺伝暗号 7</p> <p>遺伝コード 7</p> <p>遺伝子</p> <p>——の系図 116</p> <p>——の水平移動 119</p> <p>遺伝子オントロジー</p> <p>——コンソーシアム 17</p> <p>遺伝子型 101, 164</p> <p>遺伝子型空間 164</p> <p>遺伝子系統樹 117</p> <p>遺伝的アルゴリズム 93, 99</p> <p>遺伝的プログラミング 101</p> <p>入れ子構造 138</p> <p>後ろ向きアルゴリズム 91</p> <p>運動協調 155</p> <p>枝 112</p> <p>枝刈り 93</p> <p>エッジ 101</p> <p>エピジェネティクス 165</p> <p>塩基 2</p> <p>塩基置換行列 53</p> <p>塩基配置 130</p> <p>塩基配列 2</p> <p>重み行列 71</p> <p>オントロジー 16</p> <p>オントロジー編集システム 17</p>	<p>【か】</p> <p>開近傍 101, 103</p> <p>外群 40</p> <p>開始状態 73</p> <p>階層的クラスタリング 62, 64</p> <p>階層併合のクラスタリング 62, 64</p> <p>回転行列 182</p> <p>外部節 111</p> <p>下位レベル DP 185</p> <p>核磁気共鳴 9</p> <p>隠れマルコフモデル 73, 79</p> <p>仮想記憶 212</p> <p>活性部位 8</p> <p>カーネル 198</p> <p>カーネルトリック 199</p> <p>カーネル法 194</p> <p>カメレオン配列 190</p> <p>がらくた DNA 41</p> <p>関係データベース管理システム 7</p> <p>関節角度 156</p> <p>完全2分岐樹探索法 121</p> <p>完全一致 59</p> <p>観測列 79</p> <p>癌治療薬 70</p> <p>木 111</p> <p>機械学習 194</p> <p>幾何学的サフィックス木 209</p> <p>擬似度数 71, 75</p> <p>基底条件 46</p> <p>機能ドメイン 12</p> <p>キーバリュ型 7</p> <p>ギブスサンプリング法 76, 92, 95</p> <p>逆運動学 156, 158</p> <p>逆動力学 156</p> <p>ギャップ 47</p> <p>ギャップ開始ペナルティ 53</p> <p>ギャップ伸長ペナルティ 53</p> <p>ギャップ列 53</p> <p>偽陽性 71</p> <p>共通コンタクト 187, 188</p> <p>共有メモリ 215</p> <p>行列解法 194</p> <p>局所環境 184</p> <p>局所構造 68</p> <p>局所最適解 92</p> <p>局所的</p> <p>——な整列化 48, 56</p> <p>——な整列化アルゴリズム 60</p> <p>——な類似配列検索 45, 56</p>	<p>——な累積類似度行列 57</p> <p>許容誤差 45</p> <p>距離閾値 184</p> <p>距離行列 62, 64</p> <p>距離行列法 120</p> <p>距離スコア 62</p> <p>距離節約法 127</p> <p>距離ワグナー法 127</p> <p>筋骨格モデル 152</p> <p>近似文字列検索 44</p> <p>筋電図 155</p> <p>近隣 114, 122</p> <p>近隣結合法 65, 122</p> <p>組合せ爆発 29</p> <p>クラス 13, 14</p> <p>クラスター 114</p> <p>クラスター間距離 63</p> <p>クラスタリング 63</p> <p>クラスタリングアルゴリズム 64</p> <p>グラフ同型性判定問題 105</p> <p>グリッドコンピューティング 215</p> <p>形質状態 121</p> <p>形質状態法 120</p> <p>形態学 12</p> <p>系統関係 5</p> <p>結合エネルギー 205</p> <p>結合角エネルギー 205</p> <p>結合項 205</p> <p>欠失 73</p> <p>血栓 70</p> <p>ゲノミクス 1</p> <p>ゲノミクス情報処理 3</p> <p>ゲノム科学 1</p> <p>ゲノム情報 1</p> <p>ゲノム編集 145</p> <p>原子 11</p> <p>語彙の衝突 17</p> <p>広義の二面角エネルギー 205</p> <p>交叉 100</p> <p>格子状ネットワーク構造 79, 80</p> <p>格子モデル法 192</p> <p>合成によるシークエンシング 22</p> <p>構成論的アプローチ 154</p> <p>構造整列化 181</p> <p>構造ドメイン 12</p> <p>構造比較アルゴリズム 193</p> <p>構造比較プログラム 15</p> <p>酵素触媒残基データベース 194</p> <p>行動テスト 146</p> <p>誤差 69</p> <p>コネクトーム 144</p> <p>コミュニケーション 172</p>

- コンセンサスコア 69
 コンセンサス配列 68, 75
 コンタクトエッジ 187
 コンタクトマップ 187
 昆虫飛行機械 149
 コンピューテーション 172
- 【さ】**
- 再帰ステップ 82
 最急勾配登りアルゴリズム 197
 最近の共通祖先 157
 最小エントロピースコア 61
 最小進化法 127
 最小偏差法 127
 最大節約法 129
 最大ワイルドカード数 95
 最適経路 47
 最適性 65
 最汎パターン 76
 削除状態 83
 座標配列 181, 187
 座標配列長 181
 サフィックス木 48, 208
 サポートベクターマシン 194
 サポートベクトル 195
 作用を与えている点 158
 サンガー法 21
 識別番号 76
 シークエンサー 2
 四元数 182
 支持数 93
 辞書式順 93
 次世代シーケンス法 22
 実現系統樹 120
 シート 10
 シード 59
 指 標 44
 島 101
 島モデル 101
 射影データベース 94
 写真銃 148
 写像エッジ集合 188
 重心 182
 自由度問題 158
 終了状態 73
 種系統樹 117
 主 鎖 181
 出現数 71
 出現頻度 97
 出現頻度行列 68, 69, 96
 出力確率 73, 88
 出力度数 89
 準局的的 48
 —な整列化 48
 —な累積距離行列 49
 順系相同 119
 順系相同遺伝子 119
 順動力学 159
- 順動力学シミュレーション 154
 上位レベル DP 185
 状 態 73
 状態遷移 87
 状態遷移確率 75
 状態遷移行列 80
 状態列 79
 情 報 3
 情報科学 3
 情報を持たないサイト 130
 ショットガン法 24
 進化経路解析 194
 進化生体力学 169
 神経筋骨格モデル 155
 人工知能 151
 真の系統樹 120
 推定系統樹 120
 スキャン開始点 94
 スコア関数 46, 54, 66
 スーパーコンピュータ 213
 スーパーファミリー 12
 スペクトラムカーネル 200, 201
 スラッシュ 4
 スワップ操作 106
 正確さ 1
 正規表現 8, 70
 生体力学 148, 158
 静的負荷分散 214
 静電エネルギー 205
 正のインスタンス 76
 生物医学オントロジー 17
 生物医学論文データベース 6
 生物分類樹 5, 16
 生命科学 3
 整列化アルゴリズム 65
 整列行列 68
 整列集合 188
 セクション 9
 節 111
 遷移確率 73
 全域的 61
 —な整列化 46
 —な類似配列検索 45, 51
 —な累積距離行列 47
 —な累積類似度行列 53
 全域的最適解 92
 遷移度数 89
 線形ギャップスコア 53
 線形判別分析 194
 選 択 99
 全単射 188
 全文検索 7
 相対エントロピー 96
 相同スーパーファミリー 14
 相同性検索 26, 44
 挿 入 73
 挿入状態 83
 挿入列 74
- 側 鎖 181
 速度ベレの方法 206
 疎水性相互作用 192
 ソフトマージン SVM 196
- 【た】**
- 対象空間 198
 対数オッズスコア 83
 対数オッズ比 71
 対数変換 82
 対数尤度 79
 体性感覚 161
 ダイナミックプログラミング 46
 対立遺伝子 164
 多重期待値最大化法 92
 多重整列化 8, 29, 61, 73, 74
 多重配列 65
 多体結合モデル 153, 154
 タブーサーチ 99
 多様性 1
 段階的探索法 121
 タンパク質二次構造予測 192
 タンパク質立体構造データベース 9
 置換行列 32, 60, 65, 77
 逐次改善法 61, 67
 中心間距離 185
 中立論 168
 超曲面 198
 超平面 195, 198
 適応度 100
 適応度比例選択 100
 転写因子結合部 68
 問合せ 78
 問合せ構造 193
 問合せ配列 26, 44, 45, 51
 同型写像 105
 同型性判定 102
 統計的有意性 50
 動的計画法 29, 46, 81
 動的負荷分散 214
 同等に最大節約な系統樹 132
 動力学計算 156
 特異値分解 182, 183
 特徴空間 198
 突然変異 100
 凸 2 次計画問題 196
 トポロジー 14
 ドメイン 9, 12
 トライ構造 204
 トレースバック 47
- 【な】**
- 内部状態 73, 87
 内部状態変数 80
 内部状態列 79, 88
 内部節 112
 二重動的計画法 181, 184

二面角エネルギー 205
 スクレオチド 2
 根 112
 ネットワーク 115
 ネットワークモチーフ 68, 101
 脳梗塞 70
 ノード 101

【は】

バイオロボティクス 167
 背景の出現頻度 71, 98
 背景配列集合 71
 排他的近傍 102, 103
 バイナリ 4
 配列データベース 44
 配列ファミリー 14
 配列モチーフ 8, 68
 パイロシークエンシング法 22
 バウム・ウエルチアルゴリズム 75
 パス 80
 パスウェイデータベース 16
 バックボーンモデル 181
 ハッシュ表 59
 バッファ 212
 バッファ管理 212
 ハードマージン SVM 196, 199
 ハミング距離 44
 バーンスタイン問題 158
 反復計算アルゴリズム 194, 197
 比較ベア 182
 非系統樹ネットワーク 115
 非結合項 205
 ビッグデータ 1
 ヒット 60
 ヒトゲノム解析計画 1
 ヒューリスティクス 86
 ヒューリスティック 61
 表現型空間 164
 非類似度 44, 45
 非類似度スコア 46
 ヒルの筋肉モデル 153
 頻出部分配列 93
 類 度 1
 ファミリー 12
 ファンデルワールス・エネルギー 205
 フィジオーム 164
 フィールドフォワードループ 102
 フォールド 13
 フォールド認識法 191
 物理ページ 212
 負のインスタンス 76

部分配列 59
 部分配列パターン 93
 フラグメントアセンブリ法 192
 フラットファイル 4, 9
 プレーンテキスト 4
 プロファイル 61, 68
 プロファイル HMM 62, 68, 73, 79
 —の長さ 73
 プロファイル行列 61, 65, 67, 68
 プロファイル対プロファイルの
 整列化 67, 191
 プロファイル累進法 61
 分岐分類学 129
 分散型ワーカモデル 215
 分子系統学 118
 分枝限定法 99
 分子進化学 168
 分子進化系統樹 61
 分子動力学法 192, 204
 分類階級 5
 平均距離法 62, 64
 平均二乗誤差 181
 並行移動 183
 並行移動ベクトル 182
 ベイズ統計解析 97
 ベイズ法 121
 並列コンピュータ 213
 並列性能比 214
 ページアウト 212
 ページイン 212
 ペプチド結合 181
 ヘリックス 10
 ベレの方法 205
 辺 111
 変異のないサイト 130
 変換距離法 128
 編集距離 44
 傍系相同 119
 傍系相同遺伝子 119
 ポケット形状 194
 ホムンクルス 161
 ホモロジー検索 44
 ホモロジーモデリング法 192
 ポリペプチド鎖 11

【ま】

マウスンクルス 163
 前向きアルゴリズム 79, 91
 マージン 195
 マスタワーカモデル 215
 マルコフ過程 73

マルチコア 213
 マルチプルアラインメント 61
 ミスマッチカーネル 200, 202
 ミスマッチ木 203
 ミスマッチクラスタ 75
 無向グラフ 102, 187
 無根系統樹 112
 矛盾 17
 文字出現数行列 69, 97
 文字出力確率行列 80
 モーションキャプチャ 148
 文字列カーネル 200
 文字列探索アルゴリズム 7
 モーターコマンド 144, 155
 モチーフ 8, 68
 モチーフカーネル 200, 204
 モチーフデータベース 8
 モチーフライブラリー 8
 モデルパラメータ 73

【や】

焼き鈍し法 93, 99, 161
 山登り法 99
 有向グラフ 102, 187
 有根系統樹 112
 尤度関数 136
 尤度面 135
 ユークリッド空間 199
 容 量 1

【ら】

ラグランジュの未定乗数法 196
 ランダム化グラフ 102
 ランダム配列 51
 ランダムプロジェクト法 92
 ランダムモデル 88
 リシークエンシング 25
 立体構造モチーフ 68
 リモデリング 166
 類似性検索 44
 類似度 44, 45, 51
 類似部分構造 181
 累進法 61
 累積非類似度行列 47
 ルーレット選択 100
 列挙木 76
 列挙法 92, 93
 連結部分グラフ 101
 論理ページ 212

【わ】

ワイルドカード文字 93
 ワード 60

[A]		is-a	16	PrefixSpan 法	93
Aho-Corasick アルゴリズム	7	[K]		PSI-BLAST	61
ASH	186	KEGG	16	PSIST	209
[B]		KKT 条件	196	[R]	
Baum-Welch アルゴリズム	89	KMP 法	48	RCSB-PDB	9
BLAST アルゴリズム	59	Kringle ドメイン	70	RDBMS	7
BMRB	9	Kringle モチーフ	70	RMSD	181, 209
BM 法	48	ktup	60	[S]	
[C]		Kunitz モチーフ	70	SA	93
CLUSTAL W	61	[L]		SCCS	14
CMO	181	LRU	213	SIB	16
CMO 問題	187	[M]		SIGMA アルゴリズム	7
CSA	194	Mathematica	216	Smith-Waterman アルゴリズム	56
C 末端	181	MD	204	SMO 法	198
[D]		MEDLINE	16	SNP	39
DDBJ	4	MISHIMA	30	SPSP	71
DDBJ フォーマット	5	MSSD	210	SP スコア	61, 65
DDP	181, 184	[N]		SSAP	184
DNA	2	NBRF	15	SV	195
[E]		NCBO	17	SVD	182
Ecocyc	16	Needleman-Wunsch アルゴリズム	52	SVM	194
EMG	155	NER	186	SwissProt	15, 191
EM アルゴリズム	89	NIH	16	[T]	
ENA/EBI	4	NMR	9	TOP-Q	213
[F]		NoSQL	7	[U]	
FA	192	NR	191	UniProt	16, 191
FASTA アルゴリズム	59	N 末端	181	UniProtKB	16
Feng-Doolittle 累進法	61	[O]		UPGMA	64, 121
FFL	102	OBO-Edit	17	[V]	
FR	191	opt スコア	60	Viterbi アルゴリズム	79, 80
[G]		OS	212	[W]	
GA	93	OTU	111	what if	170
GenBank/NCBI	4	[P]		WIT	16
[H]		PAM 行列	53	wwPDB	9
Henikoff の BLOSUM 行列	53	part-of	16	[Z]	
HSP	60	PDB	9	Z-スコア	102, 106
[I]		PDBe	9	~~~~~	
IK	156	PDBj	9	3D-1D 整列化	191
INSD	4, 5	PDB データベース	9		
		PDB フォーマット	9		
		PIR	15		

— 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

© Kitakami, Saitou, Oota 2014

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



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
き禁じられております。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おりません。

落丁・乱丁本はお取替えいたします