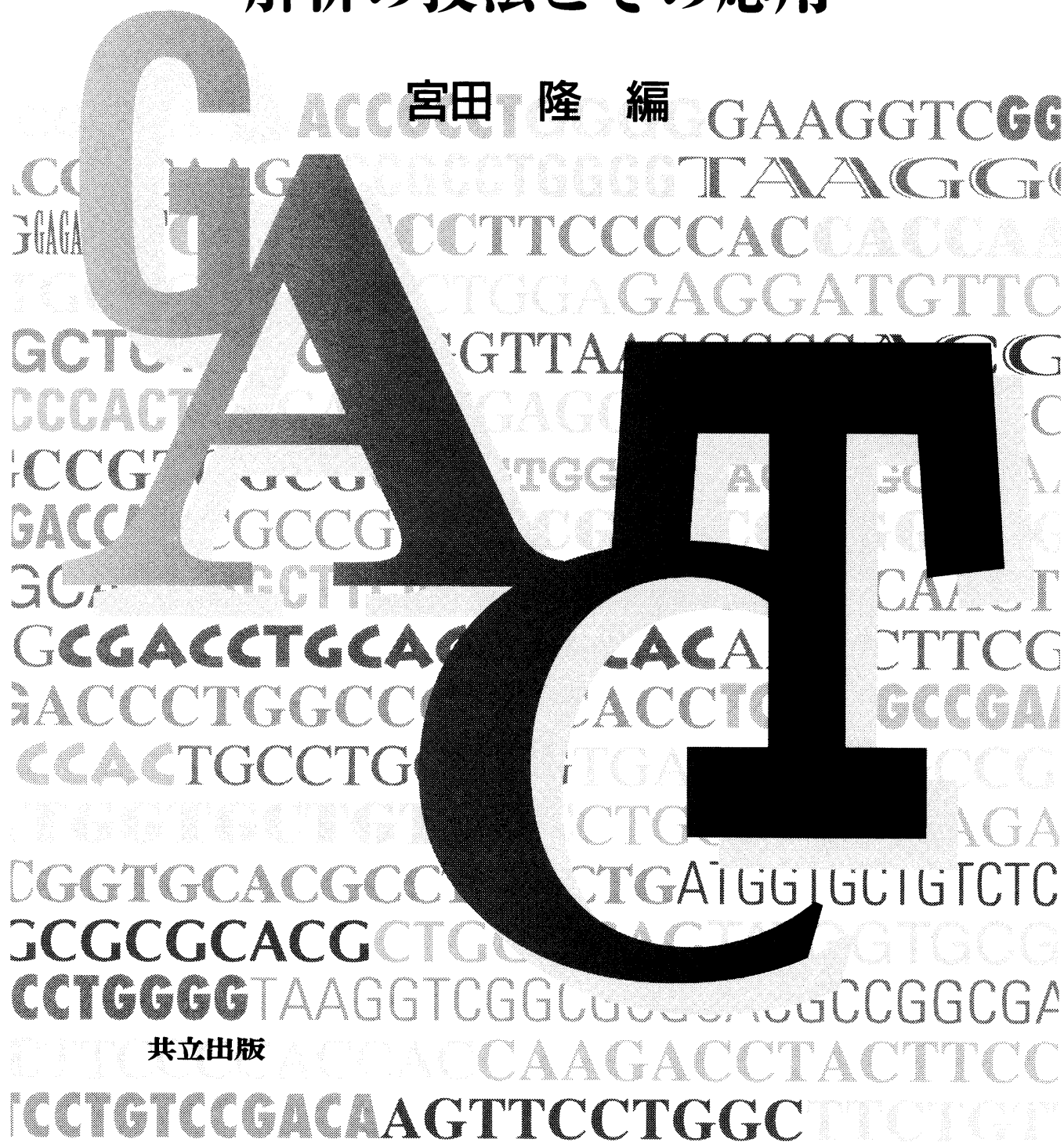


分子進化

—解析の技法とその応用—

宮田 隆 編



共立出版

執筆者一覧

- 宮田 隆 京都大学大学院理学研究科 (第I部, 第II部 2.1 節, 第III部)
- 長谷川 政美 統計数理研究所 (第II部 1.1, 1.3, 2.2, 2.4 節)
- 足立 淳 オックスフォード大学動物学科 (第II部 1.1, 1.3 節)
- 斎藤 成也 国立遺伝学研究所 (第II部 1.2 節)
- 岡田 典弘 東京工業大学生命理工学部 (第II部 1.4 節)
- 館田 英典 九州大学理学部 (第II部 1.4 節)
- 橋本 哲男 統計数理研究所 (第II部 2.2 節)
- 佐藤 矩行 京都大学大学院理学研究科 (第II部 2.3 節)
- 小林 麻里 京都大学大学院理学研究科 (第II部 2.3 節)
- 和田 洋 京都大学大学院理学研究科附属瀬戸臨海実験所 (第II部 2.3 節)
- 寶来 聰 国立遺伝学研究所 (第II部 2.5, 3.1 節)
- 植田 信太郎 東京大学大学院理学系研究科 (第II部 3.2 節)
- 石川 統 東京大学大学院理学系研究科 (第II部 3.3 節)
- 隈 啓一 京都大学大学院理学研究科 (第III部 1.1 節)
- 五條堀 孝 国立遺伝学研究所 (第IV部 1.1~1.5 節)
- 渡邊 日出海 National Center for Biotechnology Information
(第IV部 1.1~1.5 節)
- 宮澤 三造 群馬大学工学部 (第IV部 1.6 節)

括弧内は担当箇所

入れないことと、ある与えられた基準値以上の類似度を示す**部分領域**を含む連続領域のみを比較対象領域とする、という制限を加えていることによっている。この部分領域とは、ある与えられた一定の長さの比較領域である。

BLAST アルゴリズムは、類似度を統計的に評価する点で他の類似度のみによる相同性検索アルゴリズムとは異なる有用性を有しているが、ギャップの挿入が許されない等の制限が存在するために、検出能力の点で必ずしも優れているとはいえない。しかし、ギャップの挿入を許す BLAST 版も最近利用可能になっていることを考慮に入れると、今後ますます BLAST の利用価値は高まっていくものと考えられる。

1.6 インターネットの研究利用

TCP/IP プロトコルを利用した広域ネットワークであるインターネットは、近年米国の情報ハイウエー政策にも触発され、インターネットの参加機関およびその利用は指数関数的に増加し、また広域ネットワークにおける情報検索ツールの開発が引金となり、いまや全世界のほとんどの研究者にとって研究に欠かせないツールとなったばかりでなく、一般社会のなかでも確固とした地位を築きつつある。接続されている他の計算機にログインするための telnet、計算機間でファイルを交換するための ftp コマンドは、TCP/IP の典型的アプリケーションとしてよく知られている。ここでは分子進化研究におけるインターネットの利用について概観する。

a. インターネット情報検索ツール

(i) **archie** インターネット上で公開される情報の増加に伴い、利用者がインターネット上で必要な情報を効率良くアクセスするためのプロトコルおよびそのプロトコルにもとづくツールが開発されつつある。そのような情報検索システムはほとんどがクライアントサーバー概念に基づいて設計されている。そのひとつに、一定のリソースを提供している ftp サーバーを検索するための archie とよばれるプロトコル/プログラムがある。この章では、ソフトウェア、データベース等も紹介するが読者がそれらのリソースを入手する際は archie で最新の情報を得てほしい。

(ii) **WAIS, Gopher, そして WWW** 広域ネットワーク情報検索ツールとして最初に出現したのは WAIS (Wide Area Information Servers) である。1992 年ごろにはインターネット上で盛んに使用されるようになった。WAIS サーバーが提

供するのはおもにテキスト情報だが、高速な検索を可能にするためテキスト情報からキーワード抽出を行い、インデックスファイルとして逆ファイルを作成する。多くのファイル形式 (mail, NetNews, その他) に関してインデックス作成をサポートしている。筆者もインデックス作成プログラムを改良し DNA/タンパク質データベースの検索用としてテストしてみたが、(1) インデックスファイルの作成に時間がかかること、(2) インデックスファイルのサイズがソースと同程度、という点が問題であった。しかし現在商品化されているサーバーではそれらの欠点も改良されているようである。WAIS においては、まず WAIS サーバーのディレクトリーを検索し必要な情報を提供しているサーバーを見つけるといったように階層的に、検索を行う。全世界に分散している情報の検索という観点では、インタフェースはいまひとつである。それゆえ引き続き出現した Gopher に主流の座をとって代わられた。

Gopher は階層的メニュー形式でインターネット上に分散するリソースをブラウズしていくシステムである。メニュー項目としてはそのサーバーで提供しているリソースのみならず、プログラムを起動したり世界中に分散した Gopher サーバーへのポインターを含めることができる。リソースとしては各種のテキストファイル、音声、画像データ等のマルチメディアが可能で、個々のデータに固有なビューアーがクライアント上で実行される。

一方 WAIS, Gopher に 1 年ほど遅れて出現した CERN で開発された WWW (World Wide Web) は、任意の文字列やイメージ (アイコン) からなるいわゆるハイパーテキストを用いインターネット上に分散したリソースにタグをつけ効率よくアクセス可能にすることによって、Gopher より柔軟にリソースを関係づけることが可能である。中心となるのは HTTP (Hyper Text Transfer Protocol) と HTML (Hyper Text Markup Language) である。WWW ではインターネット上のリソースを一意的に指定する方法が URL (Uniform Resource Locator) として定義されている; URL: protocol://hostname/filename. クライアントとしてはインターネットブームをもたらした NCSA Mosaic や Netscape が有名である。いずれも HTTP だけでなく Gopher および WAIS 用の Z 39.50 プロトコルも取り扱えるので、Gopher, WAIS に関してもゲートウエーを経由せず直接サーバーへアクセス可能である。archie の場合はゲートウエーが利用できる。

WAIS, Gopher, WWW のいずれも世界中に広がったインターネット上に分散しているリソースを同一の GUI (Graphic User Interface) でアクセス可能という優

れた特徴をもつ。とくに Mosaic はハイパーテキストという見た目にきれいな GUI で WAIS, Gopher へもアクセス可能なため、わずか 1~2 年で世界中で使用されるようになった。また、WWW サーバーの立ち上げも比較的容易なため世界中のあちこちで WWW サーバーが稼働し、WWW サーバー上のデータを検索する機能を提供するサイトも多数存在する。以下では、その中から分子進化の研究に関連したリソースについて述べる。なおリソースの記述はすべて URL で記す (表 1 参照)。

b. 各種データベース

1994 年末の時点で分子生物学関係のデータベースは 30 をくだらない。それらは (1) DNA 配列 (GenBank, EMBL), (2) タンパク質配列 (PIR, SwissProt), (3) タンパク質配列モチーフ (PROSITE), (4) プロモーター配列, (5) 転写因子, (6) 類似タンパク質配列, (7) 繰返し配列データベース, また特定の生物を対象にした (8) ヒト, (9) マウス, (10) ジョウジョウバエ, (11) *Arabidopsis*, (12) 線虫 (*C. elegans*), (13) 大腸菌, (14) HIV ウイルス, (15) T4 ファージ等のゲノムデータベース, (16) 免疫グロブリンのアミノ酸配列, (17) 制限酵素についての各種情報, (18) 酵素についての各種情報データベース, そして (19) タンパク質構造原子座標 (PDB), (20) タンパク質構造分類データベース (SCOP) 等多種にわたる。今後も研究の発展とともに各種のデータベースが作成され利用されることになろう。これらのデータベースはすべて anonymous ftp によりインターネット上に提供されている。代表的な大規模 ftp サーバーを表 1 に示す。

c. ネットワークを利用したデータベースの日々更新

GenBank, EMBL は修正データを含め入力したデータを毎日ネットワーク上にリリースしている。1998 年現在データは平均約 15 MB/日に達する。PDB も 1 日平均約 5 エントリーがリリースされている。これら新データは 1992 年ごろまで、GenBank は internet/uucp/bitnet 上の電子ニュース (bionet.molbio.genbank.updates ニュースグループ) に投稿され、また EMBL は電子メールにより配布されていた。しかしインターネット接続機関の増加とともに anonymous ftp を利用するようになった。現在では WWW, Gopher, WAIS を通じ、キーワード検索および類似配列検索や種々の解析も可能である。しかしこれらの定型処理だけでは不十分な場合も多々あり、その場合は研究室のワークステーション上で配列解析を行う必要がある。この際、新

表 1. 分子進化の研究に関連したインターネット上のおもなりソース

URL もしくはメールアドレス	備 考
ftp://ncbi.nlm.nih.gov/	GenBank, SwissProt, ACeDB, Entrez, BLAST, ...
ftp://ftp.ebi.ac.uk/	EMBL, SwissProt, PROSITE, ...
ftp://ftp.bchs.uh.edu/	PIR, ...
ftp://ftp.bio.indiana.edu/	PHYLP, ...
ftp://ftp.pdb.bnl.gov/	PDB, gif and RGB images
ftp://ftp.gdb.org/	Genome Data Base, OMIM
ftp://ftp.genethon.fr/	ヒト遺伝子地図および物理地図
ftp://ftp.cephb.fr/	CEPH ヒトゲノムマップデータ
ftp://ftp.chlc.org/	Cooperative Human Linkage Center
ftp://ftp.gdbnet.ad.jp/	おもな ftp サーバーの日本におけるミラー
http://www.ncbi.nlm.nih.gov/	Entrez, BLAST, キーワード検索, dbEST, dbSTS
http://www.ebi.ac.uk/	BLITZ, FASTA, PROSITE パターン検索, SRS, ...
http://www.gdb.org/	GenQuest (FASTA, BLAST, Smith - Waterman)
http://probe.nalusda.gov/	ACeDB, AAtDB, IGD, MycDB, ...
http://www.embl-heidelberg.de/	データベース検索, SRS Brower, 配列構造 相関
http://expasy.hcuge.ch/	SwissProt, PROSITE, Swiss-2DPAGE, Swiss-3DIMAGE
http://www.gdb.org/	GDB, OMIM, OWL, NRL3D, PIR, EC - Enzyme, REBASE, TBASE, ...
http://www.chlc.org/	ヒト遺伝子地図
http://www.jgi.doe.gov/	Joint Genome Institute
http://www.tigr.org/	TIGR データベース
http://www.genethon.fr/	CEPH ヒトゲノムマップデータ
http://www.jax.org/	マウスゲノム情報
http://www-hgc.lbl.gov/	LBL prosophila Genome Center
http://www.pdb.bnl.gov/	PDB, Gif and RGB images
http://www.nih.gov/modeling/	Molecules R US
http://scop.mrc-lmb.cam.ac.uk/scop/	タンパク質構造分類; SCOP
retrieve@ncbi.nlm.nih.gov.	GenBank メールサーバー
blast@ncbi.nlm.nih.gov	BLAST メールサーバー
blitz@ebi.ac.uk	BLITZ メールサーバー
flat - netserv@smlab.sci.gunma-u.ac.jp	FASTA, PALIGN, ...

データを取り込みデータベースを日々更新することが必要となる。その一例として私の研究室で使用しているシステム^[2]を図4に示す。

このシステムは広域ネットワークを利用したデータ取り込みとして、(1) anonymous ftp によるファイル転送、(2) 電子ニュースからのデータ取り込み、(3) 電子メールによるデータ配布の3つの方法をサポートしている。取り込まれたデータは、cron 機能を利用して一定時間ごとにプログラムで処理し更新データを作成する。この検索システムはUNIXにおけるフィルターとしてはたらくさまざまなツールからなり、正規表現によるキーワード検索、特異な塩基配列をもつ遺伝子の検索が可能である。このようにして構築されたデータベースは計算機に直接ログインしての利用だけでなく、電子メールやUNIXシステムにある whois コマンドにより利用できる。

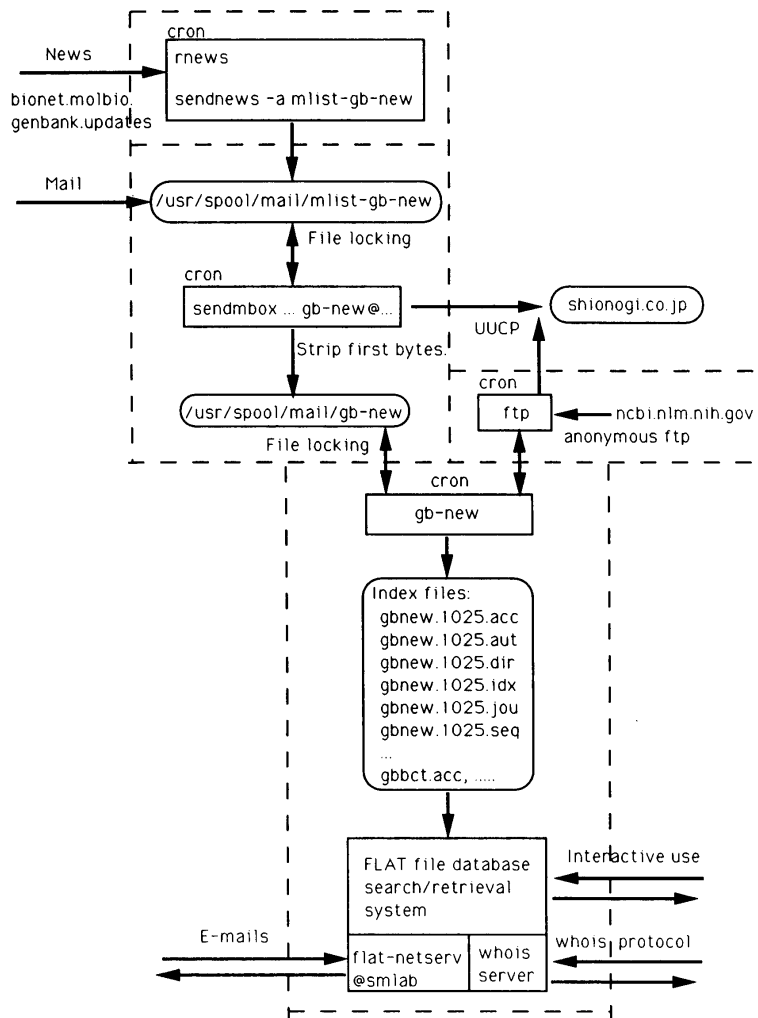


図4. ネットワークを利用したデータベースの日々更新システム

d. データベース検索および配列解析

従来生物学関連のデータベースは、データを含むフラットファイルとインデックスファイルそして検索ソフトウェアおよび各種解析プログラムからなるのが通常であった。このアプローチはシステムを容易に構築できる半面、データ間での相互参照が必要不可欠なゲノムデータベースの構築には適さない。関係データベースやオブジェクトオリエンティッドデータベース管理システムが必要とされるゆえんである。線虫 (*C. elegans*) ゲノムプロジェクトにおいて開発されたゲノムデータベース管理システム ACeDB は生物学者の思考に馴染みやすいデータ構造およびアクセスメソッド、GUI を提供するばかりでなく、新たなゲノムデータベースの構築においてもプログラム作成の必要なく、生物学者が独力でデータベース設計ができるような機能を兼ね備えている。そのため Arabidopsis (AAtdB), mycobacterium (MycDB), 21 番ヒト染色体 (IGD) ゲノムデータベースの構築にも利用された。また ACeDB は DNA 配列の画面表示においても通常の配列解析で必要とされる各種の機能、コーディング領域、類似配列、繰返し配列、プロモータ配列、結合領域等の表示が可能である。また各種解析プログラム (制限酵素地図の作成、フィンガープリントの作成、コドン頻度の計算、スプライス部位におけるコンセンサス配列の表示、エクソン・イントロン領域予測等) も組み込まれている。ACeDB 管理ソフトウェアおよびデータベースは anonymous ftp で入手できる。また ACeDB を用い作成された各種データベースも WWW で利用できる。

ゲノムデータベースでは遺伝子地図、物理地図を中心に遺伝子情報、クローン、DNA 配列、STS、EST、文献情報等における相互参照が重要である。一方、一般配列データベースにおいては配列相互の関係として配列の類似性に重点がおかれる。そのような観点からデータベースをアクセスするソフトウェアとして、NCBI で開発された Entrez があげられる。配列/構造データベースは配列/構造データ間に配列/構造相互の類似性が定義され、また文献情報にも文献テキストに含まれる字句の頻度の相関に基づいて類似度が導入される。そして配列/構造データと文献は配列/構造データを報告、引用した文献という関係により結ばれる。これらのポインターを用い配列、構造、文献データベースを興味ある情報を求め散策できるようになっている。Network Entrez や WWW を用いて利用できる。

Entrez で利用できる配列は、すでにデータベースに登録されているものであるが、手元にある配列に対する FASTA, BLAST, BLITZ 等による類似配列検索も電子メ

ールや WWW を用いて可能である。BLITZ はタンパク質配列のための高速類似配列検索で並列計算機 MasPar の上で稼働する Smith-Waterman の local similarity algorithm を用いたプログラム MPsrch を用いている。

タンパク質構造データベースにおいては、それぞれのタンパク質に特徴的な構造の静止画像が WWW を用いて利用できる。また分子構造表示ソフトをインストールすれば、CPK 模型の表示や構造の回転も可能である。タンパク質構造分類データベース (SCOP) とともにタンパク質の分子進化を考える際、有用であろう。

e. 類似配列検索

最後に類似配列検索における注意点を述べよう。BLAST, FASTA は高速性を重視するための近似的方法であり、類似配列を見落とす場合もある。また FASTA が出力するアラインメントも近似だから、最終的にはより厳密なアルゴリズムを用いたプログラムでアラインメントを作成することが必要である。2 配列間のアラインメント作成アルゴリズムには 2 種類ある。ひとつは global alignment といい、類似度最大の全領域にわたるアラインメントを計算するもの、もうひとつは local similarity alignment といい統計的に有意と考えられる類似な部分配列を計算する方法である。前者では類似度最大アラインメントを求める Needleman-Wunsch の方法と距離最小のアラインメントを求める Sellers のアルゴリズムがある。この 2 つは双対的な関係で等価であることが証明されている^[4]。後者では Smith-Waterman およびその改良版 Waterman-Eggert のアルゴリズムがある。DNA 配列の場合には通常 local similarity alignment が適している。一方、タンパク質では(マルチドメインタンパク質はドメイン単位で) global alignment が最適であろう。global alignment の場合、アラインメントの信頼性は当然ながら全領域で一様ではないことに注意すべきである。

図 5^[3] はヒト α ヘモグロビンと lupin leghaemoglobin タンパク質のアラインメントの一部 (ヘリックス E から G にかけての部分) である。タンパク質の立体構造の重ね合わせにもとづく構造アラインメント、2 次構造部分に大きなギャップペナルティを用いた類似度最大のアラインメント、至るところ同一のギャップペナルティを用いた類似度最大のアラインメントと確率アラインメント^[3] (残基対のアラインメント確率が 0.5 以上の残基対のみからなるアラインメント) が図示されている。“#”記号は構造アラインメントと一致しない残基対を示す。確率アラインメントの下部

参 考 文 献

1.6 節

- [1] Lesk, A.M., Levitt, M., Chothia, C. : Alignment of the amino acid sequence of distantly related proteins using variable gap penalties. *Protein Engineering*, **1**, 77-78 (1986).
- [2] Miyazawa, S. : DNA Data Bank of Japan : Present Status and Future Plans. *In* Bell, G., Marr, T. eds. : Computers and DNA, Santa Fe Institute Studies in the Sciences of Complexity. vol. VII, pp. 13-19, Addison-Wesley, Reading, MA (1989).
- [3] Miyazawa, S. : A Reliable Sequence Alignment Method Based on Probabilities of Residue Correspondences. *Protein Engineering*, vol. No. 8, 999-1009 (1995).
- [4] Waterman, M. S. ed. : Mathematical Methods for DNA sequences, CRC Press, (1989).