“分子進化 —- 解析の技法とその応用”

原稿

インターネットの利用

群馬大学工学部共通講座　宮澤三造

smiyazaw@smlab.eg.gunma-u.ac.jp

電話： 0277-40-1027

FAX: 0277-40-1026

インターネットの利用

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。

1　インターネット情報検索ツール

1.1　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基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。その一つに、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、ソフトウエア、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。

1.2　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 主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/蛋白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、インターフェースはいまひとつである。それ故引続き出現した 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 年程遅れて出現した 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 Go-

pher および WAIS 用の Z39.50 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持つ。特に Mosaic はハイパーテキストという見た目に綺麗な 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 サーバーの立ち挙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)。

2 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) 蛋白質配列 (PIR,SwissProt), (3) 蛋白質配列モチーフ (PROSITE), (4) プロモーター配列, (5) 転写因子, (6) 類似蛋白質配列, (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) 蛋白質構造原子座標 (PDB), (20) 蛋白質構造分類データベース (SCOP) 等多種に渡る。 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全て 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。

3　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 年現在でデータは平均約 15MB/日に達する。PDB も一日平均約 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 [1] を図 1 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毎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。

4　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エ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、コーデイング領域、類似配列、繰り返し配列、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 管理ソフトウエ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エアとして、 NCBI で開発された Entrez が挙げられる。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、既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を用い可能である。BLITZは蛋白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。

蛋白質構造データベースにおいては、各々の蛋白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。蛋白質構造分類データベース(SCOP)と共に蛋白質の分子進化を考える際、有用であろう。

5 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の出力するアライ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メントを作成することが必要である。2配列間のアライ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メントを求めるNeedleman・Wunschの方法と距離最小のアライメントを求めるSellersのアルゴリズムがある。この二つは双対的な関係で等価であることが証明されている[2]。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、蛋白質では(マルチドメイン蛋白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2[3]はヒトαヘモグロビンとlupin leghaemoglobin蛋白質のア

ライメントの一部 (ヘリックス E から G にかけての部分) である。蛋白質の立体構造の重ね合わせにもとづく構造アライメント、2 次構造部分に大きなギャップペナルテイを用いた類似度最大のアライメント、至るところ同一のギャップペナルテイを用いた類似度最大のアライメントと確率アライメント [3](残基対のアライメント確率が 0.5 以上の残基対のみからなるアライメント) が図示されている。"#" 記号は 構造アライメントと一致しない残基対を示す。確率アライメントの下部に記された数字は残基対のアライメント確率 $x10$ を示す。類似度最大のアライメントにおける正しくない残基対はほとんどがアライメント確率が 0.5 以下の残基対であることがわかる。アライメント確率に基づく確率アライメントは信頼性の高い部分のみを含むので予測された残基対は正しいことが多い。またヘリックス F の部分のアライメントを見ると、類似性が低い場合、確率アライメントのほうがより正しいアライメントを予測するということが示唆される。もちろん確率アライメントもパラメータの値に依存することはいうまでもない。ギャップペナルティを 置換の起こりにくい 2 次構造部分や蛋白質の内部でより大きな値に設定することにより、より正確なアライメントを得ることができる。しかしいずれにせよ類似性が低い場合に類似度最大のアライメントを使用する際は注意が必要である。

6 まとめ

分子進化研究におけるインターネットの利用について簡単に述べた。今後インターネット上に分散した各種データベースからなる 統合データベース (vertual integrated database) が出現してくるように思える。またデータだけでなく多種多様な解析手段がネットワーク上に提供されるであろう。ネットワークを上手に利用することが肝心である。

7 参考文献

1. Miyazawa, S.: DNA Data Bank of Japan: Present Status and Future Plans. In: Computers and DNA, Santa Fe Institute Studies in the Sciences of Complexity. vol. VII, pp.13-19, Eds. G. Bell and T. Marr, Reading MA: Addison-Wesley (1989).

2. Waterman, M. S. ed: Mathematical Methods for DNA sequences, CRC Press (1989).

3. Miyazawa, S., A Reliable Sequence Alignment Method Based on Probabilities of Residue Correspondences. Protein Engineering, **8**, 999-1009 (1995).

4. Lesk, A.M., Levitt, M. and Chothia, C.: Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. Protein Engineering, **1**, 77-78 (1986).

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リソース

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/	PHYLIP,...
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-2DPHAGE, 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 Drosophila 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/	蛋白質構造分類; 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, ...

図 2　ヒト α ヘモグロビン (H) と lupin leghaemoglobin (L) 蛋白質のアライメントの一部 (ヘリックス E から G にかけての部分)[3] 。構造アライメント及び可変ギャップペナルテイの結果は、Lesk, Levitt & Chothia (1986)[4] より引用。詳しくは本文参照。

```
                 -----   E   ---->         <----  F -----> <--------  G  ---
structural    H: VKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLA
superposition     |  ||      |              |  |   |  |  |    |       | |
              L: LQAHAGKVFKLVYEAAIQLEVTGVVASDATLKNLGSVHVSKG-VADAHFPVVKEAILKTIK

alignment     H: VKGHGKKVADALTNAVAHVD-----DMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLA
with variable     |  ||      |              |  |   |  |  |    |       | |
gap penalty   L: LQAHAGKVFKLVYEAAIQLEVTGVVASDATLKNLGSVHVSKG-VADAHFPVVKEAILKTIK
                              #  #

alignment     H: VKGHGKKVADALTNAVAHVD--DMPNALSALSDLHAHKLR-- VDPVNFKLLSHCLLVTLA
with uniform      |  ||      |              |  |          |    |       | |
gap penalty   L: LQAHAGKVFKLVYEAAIQLEVTGVVASDATLKNLGSVHVSKG VADAHFPVVKEAILKTIK
                              ########################

probability   H: VKGHGKKVA                    LSALSDLHAHK    PVNFKLLSHCLLVTLA
alignment         |  ||                      |  |   |  |       |       | |
              L: LQAHAGKVF                    LKNLGSVHVSK    DAHFPVVKEAILKTIK
                 777766665                    55666666655    5689998888887776
```