Annual Report of the Computer Center, Gunma University, **7**, 110-121, 1992

遺伝情報解析分野におけるネットワーク利用の一例：研究室 LAN の広域ネットワーク接続とその利用；ネットワークを利用したデータベースの日々更新とネットワーク上へのその提供

工学部共通　宮澤三造
smiyazaw@smlab.eg.gunma-u.ac.jp

概　　要

遺伝情報解析に欠かせないものに、蛋白質アミノ酸配列、蛋白質/DNA の 3 次元構造、および DNA 塩基配列データベースがある。これらデータベースは、近年解析技術の進歩によりデータの増加が著しく、毎日データが Internet ネットワーク上に電子メール、電子ニュース、 anonymous ftp 等によりリリースされている。私の研究室では UNIX システムを用いてこれらのデータを各々の方法でネットワーク上から取り込みデータベース化して研究に利用している。またこのようにして構築されたデータベースは研究室で利用するだけでなく、データベース検索ソフトウエアの一部を外部から (1) 電子メールと (2)whois protocol により利用できるよう開放している。また wais server も提供している。ここでは研究室 LAN の広域ネットワーク接続の現状について述べこのような活動の概略を報告する。

1　はじめに

近年 (約 10 年ほど前より) 遺伝情報解析における計算機の利用は急速に盛んになり、近頃ではこの分野を Bioinformatics と呼ぶことも多い。従来この分野における計算機の利用者は、計算機を直接研究の道具として使用する比較的少数の研究者に限られていた。しかし近年遺伝子解析にたずさわるほとんど全ての研究者が研究に計算機を使用せざるを得なくなった。このように計算機が急速に利用されるようになった背景には、1970 年代後半にはじまる DNA 塩基配列解析の実験技術の進歩がある。さまざまな生物種において多くの遺伝子が DNA レベルで解析されるようになり DNA 配列データが急速に増大したため、データベースを用いた遺

伝情報解析やデータ交換に計算機の利用が必須になった。最近ではファージの全遺伝子配列のような短いものばかりではなく人間の全塩基配列までも解析しようとするゲノム解析計画が米国，欧州をはじめ日本でも発足した。今後ゲノム解析計画の進行とともに従来以上の多量のデータが予期される。解析にあたって最新のデータが利用できるか否かは研究者にとって死活問題である。

現在、TCP/IP による Internet 広域ネットワークは、データバンクによる研究者からの解析データの収集の際に、また研究者へのデータベースの提供のために必須となった。いまや毎日最新のデータがネットワーク上にリリースされている。研究者にとって、anonymous ftp はもちろん wais, gopher 等はデータの取得および解析ソフトウエアの交換に欠かせないツールである。また電子ニュースも研究者間の情報交換に若手研究者の間で使用されつつある。このような状況にあって研究室 LAN の Internet 広域ネットワークへの接続とその管理、ネットワーク関連ソフトウエアのインストレーションは研究活動を支える重要な土台になりつつある。

私の研究室では UNIX システムを用いて最新のデータを毎日ネットワークから取り込みデータベース化し研究に利用しているのでネットワーク利用の一例として簡単に紹介する。

2　研究室 LAN の広域ネットワーク接続の現状

私がここ群馬大学工学部（桐生）に転任してきた 1991 年 4 月当時，群馬大学工学部は JUNET に加入していたが Internet(IP ネットワーク) には未接続で、電子メールも送受信に時間がかかった。電子メールの高速化の必要からまた多量のファイル転送が必要とされるため 1991 年 6 月に研究室の LAN と理化学研究所 (埼玉県和光市) と間で UUCP 接続をおこなった。また TCP/IP のネットワークアプリケーションを利用する必要から ISDN64 を使用した 64 kpbs の IP link を計画した。当初 1991 年夏頃に設ける計画であったが接続先における ISDN 回線の敷設が遅れ 1992 年 4 月より理化学研究所の一計算機との間でテスト接続を行っている。UUCP link は現在理化学研究所に加え，シオノギ製薬研究所 (大阪)、国立衛生試験所 (東京) との間で結ばれている。シオノギ製薬研究所との UUCP link は、DNA 塩基配列データベースの日々更新のために多量の DNA 塩基配列データの送信が必

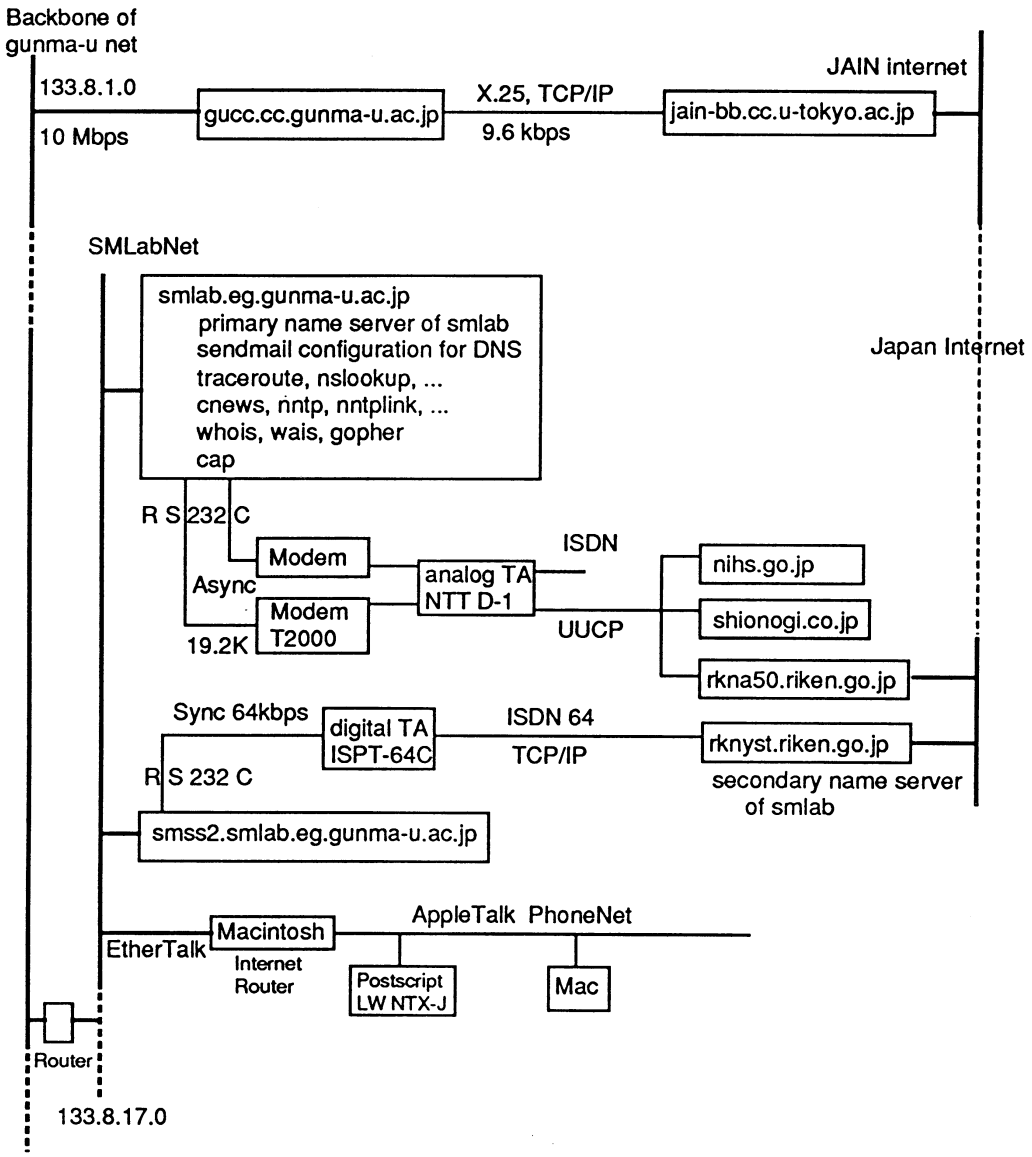


Fig. 1 Wide area network links at smlab.eg.gunma-u.ac.jp

要となるために設けられた。

図 1 は研究室の LAN(SMLabNet) の広域ネットワークへの接続形態を示す。当研究室の LAN は現在工学基礎の LAN の一部を成し群馬大学の campus LAN に接続されている。ISDN64 を使用して IP 接続をするためには経路制御の観点からまた security と完全性を高めるためにサブネット化が望ましい。将来はワークステーションをルーターとして用い当研究室だけでサブネットを構成する計画である。campus LAN の広域 IP ネットワーク (Internet) への接続は 1991 年 12 月に情報処理センターが JAIN (Japanease Academic Inter-University Network) に加入することによってなされた。しかし学術情報センターの X.25 ライン (9.6kbps) を用いての IP link のため低速で電子メールの配送以外 ftp 等には適さない。高速化が望まれるところである。

広域 IP ネットワーク JAIN への接続に伴い，図 1 にあるように SMLabNet の一計算機 (smlab.eg.gunma-u.ac.jp) で smlab domain の parimary name server を立ちあげ，電子メールも Domain Name System の MX record を参照して SMTP により配送している。それ故 Internet 上の全て計算機と mail の送受信が一分以下に高速化された。表 1 にあるようなネットワークソフトウエアを，ネットワーク構築，ネットワーク管理，ネットワークの利用のためにインストールしている。

表 1. ネットワーク構築、管理、利用のために使用している代表的ソフトウエア。

ネットワーク構築	gated	BIND 4.8.3	sendmail 5.65+1.6W	cap	amd
ネットワーク管理	traceroute	nslookup	cmu-snmp	ping[a]	etherfind[a]
	crack[b]				
ネットワーク利用					
Utilities	telnetd[c]	ftpd[c]	ftp[c]		
電子ニュース	cnews	nntp	nntplink	rn	xrn
データベース	whois	wais	gopher	xwebster	

[a] システムに備わっているコマンド；他はすべて public domain software

[b] security 向上のためのパスワードチェッカー

[c] 新バージョンを使用

3　蛋白質アミノ酸配列、3次元構造、およびDNA塩基配列データベース

遺伝情報解析分野における代表的データベースは蛋白質アミノ酸配列、蛋白質/DNAの3次元構造、およびDNA塩基配列データベースである。これらデータベースの重要性をいち早く認識しデータベース構築に貢献した最大の人物は故M. O. Dayhoffである。Dayhoffは蛋白質配列を収集蓄積し、1969年に早くも“Atlas of Protein Sequence and Structure”を出版している。この頃は主に、相同蛋白質の比較から分子レベルで進化を考察すること、またアミノ酸置換の様相から蛋白質の構造と機能が議論された。その後、1970年代後半に始まるDNA配列解析の実験技術の進歩によりさまざまな生物種において多くの遺伝子がDNAのレベルで解析されるようになり、DNA塩基配列のレポートは指数関数的に増大した。それとともにDNA塩基配列データベースの重要性も増し、このような状況の中で、1982年欧州にEMBL Data Library、米国にGenBankがDNAデータバンクとして設立され、現在これらデータバンクは共同でデータの収集提供を行っている。

DNA/蛋白質配列データベースは、配列の類似性検索(homology search)及びSequence alignmentを主な解析手段として、類似性の程度に基づき配列の分子系統樹の作成、類似性を手がかりにDNA蛋白質配列の機能予測、構造予測等、分子進化、DNA塩基配列の遺伝情報解析に欠くことができない。近頃では新しくDNA配列を解析した場合、配列の類似性検索は実験家のルーチンワークとなっている。このようにDNA/蛋白質配列データベースは生物学、医学、農学等の広範囲にわたり研究上必要不可欠となった。一方3次元構造データもPhillips等により1967年にlysozymeの3次元構造がX線解析法によりあきらかにされて以来増加し、Protein Data Bank (PDB)が1971年にケンブリッジの結晶学センターとアメリカのブルックヘブン国立研究所との共同で発足した。PDBは現在蛋白質ばかりでなくDNAの3次元構造データも収集提供している。3次元構造データは蛋白質の酵素反応機構、分子認識機構の解析に必須であるだけでなく一次配列から3次元構造の予測を目指す配列構造相関の解析にも配列データベースとともに欠くことができない。

現在DNAデータベースには約8万遺伝子 10^8 塩基が収集され、また構造データベースには1000以上の構造データが収集されている(表2参照)。これらのデータベースには、配列や構造データだけでなく文献情報および配列に関する既知の

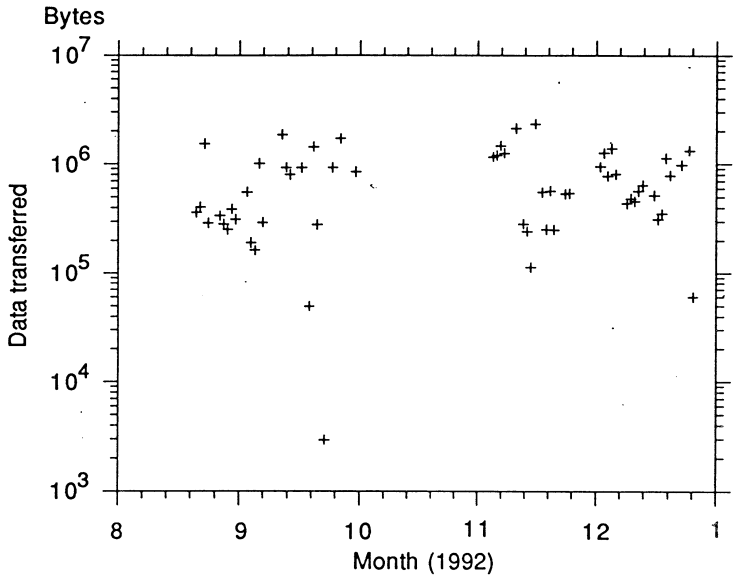
遺伝情報の記述も含まれる。研究者の必要とするデータ検索はいわゆる文献情報データベースのように単純ではなく、関係データベースやオブジェクトオリエンテイドデータベースを構築する試みがあちこちでなされている。しかし利用者への配布は未だ単純なフラットファイルの形でなされている。その容量はDNAデータベースで約300MB、構造データは約250MB、prereleaseのものを含めると約400 MBにも達する。現在ヒトの場合で全ゲノムの約0.6％が解析されている。大腸菌の場合は約75％である。収集されている全塩基数は大腸菌ゲノムの約22倍に相当する約101M Basesである。

表2. DNA, 蛋白質配列データベース及び構造データベースのデータ量

	データベース	バージョン		データ量	
DNA					
	GenBank	73	09/92	300 MB	78 K loci, 101 M bases
	EMBL	32	09/92	260 MB	79 K loci, 101 M bases
蛋白質					
	GenPept	73	09/92	100 MB	
	SwissProt	22	05/92	60 MB	
	PIR	34	09/92	100 MB	45 K proteins, 13 M residues
3次構造					
	PDB		10/92	250 MB	1007 entries
	PDB	prerelease	10/92	140 MB	435 entries

これらデータベースは各々ほぼ3月毎にリリースされ、希望者には磁気テープにより配布されている。しかし広域ネットワークの拡大にともないより便利な配布方法が望まれ、現在ではここで挙げた代表的データベースをふくめ多くのデータベースがanonymous ftpにより入手可能となっている。また近年データはほぼ指数関数的に増加しているため、3月毎のリリースサイクルでは不充分となり、DNAデータベースは毎日新データがネットワーク上にリリースされている。図2はこのようにして日々リリースされるGenBankのデータ量を示す。日により大きく変動するが，平日は平均約500KBのデータ量に達する。PDBも一日平均約一エントリーがリリースされている。(図3参照)。これら新データはanonymous ftpおよびGenBankはinternet/uucp/bitnet上のUSENET電子ニュースのbionet.-

Fig. 2 Data released by GenBank each day from the midle of August to the end of December. No data was trandferred in October.



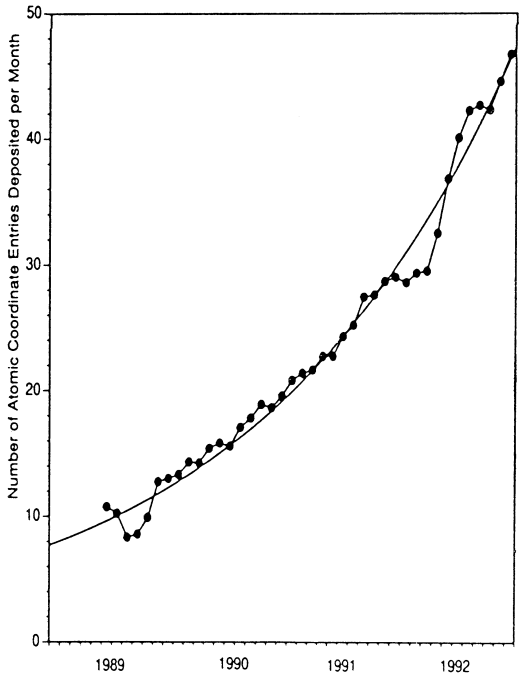


Fig. 3 Running 12-month average number of atomic coordinate entries deposited per month since 1989. The curve shows an exponential fit to the experimental data points.

Taken from Protein Data Bank Newsletter No. 60, 1992.

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る。私の研究室ではこのようにして毎日リリースされるデータを取り込み研究に利用しているのでそのシステムの一端を紹介する。

4 ネットワークを利用したデータベースの日々更新とネットワーク上へのその提供

4.1 ネットワークを利用したデータ取り込み

現在、DNA/蛋白質配列データベースに関しては、広域ネットワークを利用したデータ取り込みとして、

- anonymous ftp によるファイル転送
- 電子ニュースからのデータ取り込み
- 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。図 4 は (GenBank) 配列データベースにおけるデータフローを示す。電子ニュースからのデータ取り込みでは、news article として到着したデータは最も容易な方法, Cnews もしくは Bnews software package に含まれている、ニュースをメールを用い配送するためのプログラム sendnews を用いてメールとして取り出される。このプログラムは CRC チェック用の情報も出力するのでデータをメールとして受け取った際不完全なデータを除去することも容易である。このようにして取り出されたデータはシステムの負荷を軽減するため一旦システムのメールボックスに蓄えられた後、cron 機能を利用して一定時間毎にプログラムで処理し mailing list により必要なサイトに転送している。なおこの際、Bitnet では 300KB/mail 以下のメールしか許されないので、メールのサイズを調整している。またこのプログラムは、データを受ける相手先によっては、sendnews の付加する先頭一バイトの除去も行っている。このようにして取り出されたデータは UUCP または SMTP で配布している。その一つに Bitnet 接続のため news が入手できない台湾のサイトがある.

データが電子メールにより入手できる時はもちろん sendnews を用いてメールとして取り出す部分が不必要となる。また anonymous ftp が利用できる時のために、新データを認識し転送するためのプログラムが用意されている。このプログラム

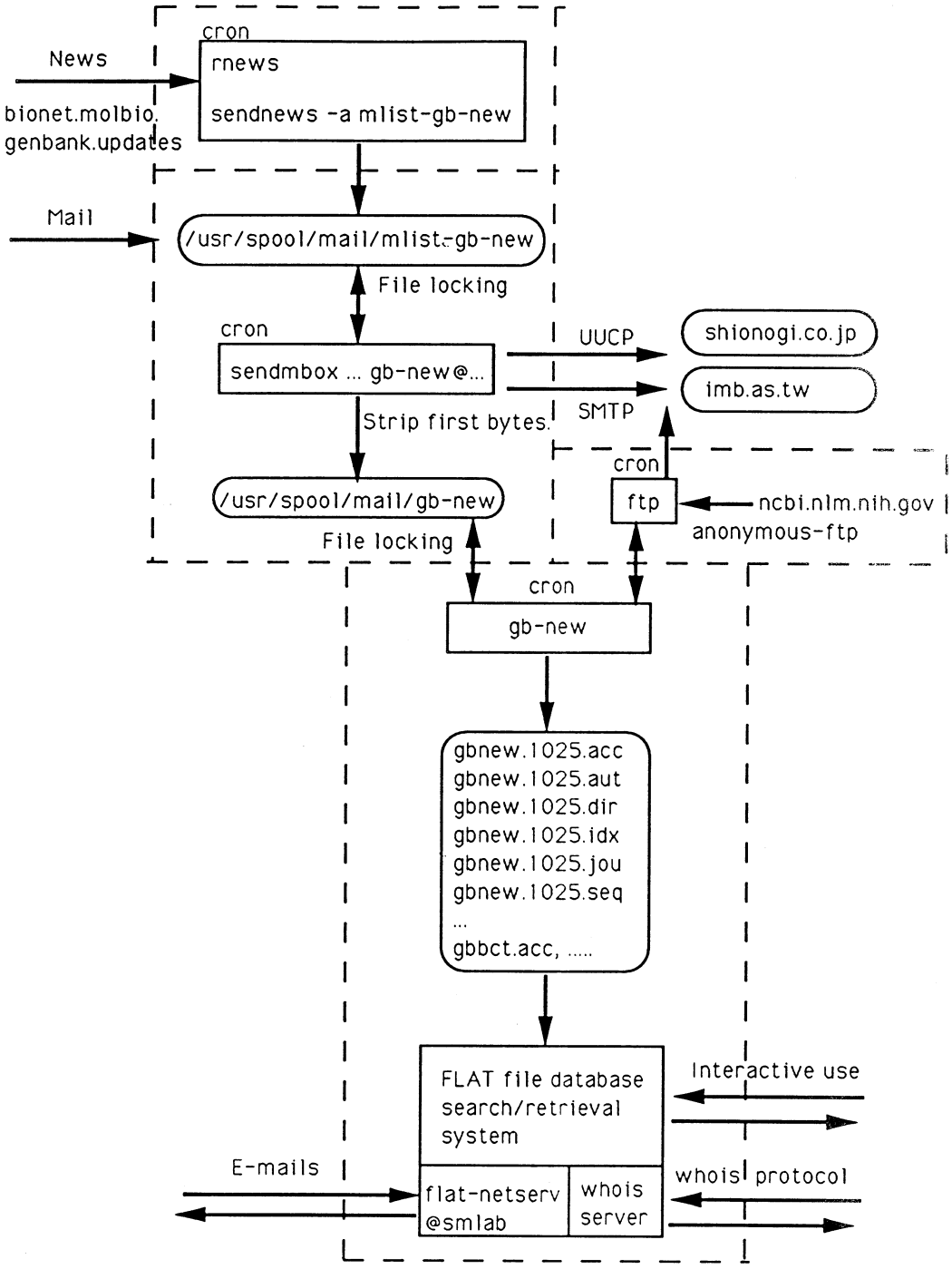


Fig. 4. Data flow in sequence database at smlab.eg.gunma-u.ac.jp

も cron により定期的に起動される。EMBL データは電子メールで送られて来る。GenBank はこれまで電子ニュースからのデータ取り込みを行っていたが anonymous ftp が 1992 年 10 月より利用できるようになったので現在は anonymous ftp を用いている。PDB 構造データベースは、東大吉田氏作成の ftpsync と呼ばれるプログラムを用い anonymous ftp によりデータベースを常に最新のものに保っている。

4.2 データベース構築

このようにして取り込んだ新データは一日数回 cron により起動されるプログラムにより処理され、インデックスファイルやデータベース検索のための各種ファイルが作成される。このデータベース検索システムでは作成及びメインテナンスの容易さのためフラットファイルを用いている。データも複数のファイルに分割して保持出来る。それゆえ更新データは一日分ごとに別ファイルとして管理されている。検索システムは簡単な機能を果たす様々なツールからなる。ほとんどのツールは UNIX におけるフィルターとして働く。このようなツールを UNIX のパイプで組み合わせることにより、著者名、論文名、生物種、遺伝子名、キーワード等による検索が可能である。文字列は通常 UNIX の正規表現で指定する。よってあいまいな文字列による検索が可能である。また特異な塩基配列をもつ遺伝子の検索においても塩基配列を正規表現で表現できる。

4.3 広域ネットワーク上へのデータベースの提供

このようにして構築されたデータベースは研究室で利用するだけでなく、広くネットワーク上に開放している。広域ネットワーク下でのデータベース利用は

- 電子メールによるデータベース検索
- UNIX システムにある whois コマンドを利用してのデータベース検索
- wais protocol

が可能である。電子メールによるデータベース検索では、検索システムのコマンドのサブセット (著者名、論文名、キーワード等による検索および配列の類似性検索) のみがメールの本文のなかで利用できる。利用者はオンライン使用時と全く同じように検索コマンドを書いた電子メールを flat-netserv@smlab.eg.gunma-u.ac.jp

に送ることにより検索結果をメールとして得ることができる。キーワード等による検索はまたwhoisコマンドを利用しておこなうこともできる。そのためにinetdを利用し簡単なwhoisサーバーも稼働させている。また試験的にwaisサーバー稼働させている。 waisはテキストデータのための分散データベースをInternet上に構築する手順で、並列計算機メーカーであるThinking Machinesで開発されたものである。 waisを用い現在さまざまなテキストデータベースが米国をはじめ世界上で提供されている。しかし研究目的のデータベースのまれでDNA/蛋白質データベースはその珍しい例である。しかしwaisデータベースの構築は通常CPU時間とデイスクスペースを非常に食うinverted fileの作成をふくむ。DNA/蛋白質データベースは日々更新することが要求されるので、更新がデータベースの再構築を必要とするようなシステムでは計算機への負荷が高過ぎるように思われる。

ちなみに私が提供しているシステムの利用の程度を調べてみると、電子メールによる利用が450-500 queries/ month程ある。利用の目的は配列の類似性検索がほとんどである。表3は1992年9-12月の4ヵ月間における電子メールでデータベース検索を行った利用者の国毎の分布である。 USA, Finland, Germanyがベスト3である。残念なことに日本ではそれほど利用されていない。というのも、日本では生物関連分野の実験家にとってまだ電子メールはなじみが薄いからであろう。しかし学生、研究者に対する適切な教育が欠けているようにも思われる。

5 まとめ

研究室LANの広域ネットワーク接続の現状を述べ，広域ネットワークを利用したデータベースの日々更新とネットワーク上でのデータベースの提供に関し簡単に報告した。データベースの管理は手間がかかる。今後ゲノムプロジェクトの進展とともにデータベースは増大する一方である。データ量だけでなくその利用もより高度な形が要求されよう。全てのユーザーが個々に管理するのは不可能である。ネットワーク上へのデータベースの提供が必要とされる結縁である。今後はネットワークのの進展とともに，サーバー，クライアントタイプのデータベース検索システムがあちこちで利用できるようになると思われる。いずれにせよ，広域ネットワークは研究に必須でありその高速化と整備が必要とされている。

残念ながら日本では広域ネットワークの構築および整備を研究者が片手間に行っ

ている現状である。米国では、NSF、DOE(エネルギー省)、NASA 等がネットワークを構築し、backbone は T1 回線 (1.5Mbps) から T3 回線 (45Mbps) に既に移行した。1Gbps も計画されている。一方日本では backbone で 192kbps 程度であり、群馬大学の場合は広域ネットワークへの接続はなんと 9.6kbps である。インフラストラクチャーとしてはあまりに貧弱である。研究にとって情報交換はなくてはならないものである。情報交換に果たすネットワークの役割ははかりしれない。米国では副大統領の掲げる重要な政策の一つにもなっている。日本でもせめて T1 回線程度の高速化を望みたい。

表 3. 1992 年 9 月から 12 月の 4ヵ月間に電子メールでデータベース検索を行った利用者の発信アドレスの分布

#mails	domain	country #mails	domain	country	
9	at	Austria	28	gov	USA governmental
5	au	Australia	1	gr	Greece
36	be	Belgium	1	ie	Ireland
42	bitnet	Bitnet	80	jp	Japan
26	br	Brazil	1	kr	Korea
111	ca	Canada	42	nl	Netherlands
172	de	Germany	1	no	Norway
493	edu	USA educational	5	nz	New Zealand
2	es	Spain	101	org	USA organization
577	fi	Finland	22	sg	Singapore
91	fr	France	16	uk	United Kingdom