シリーズ企画：研究者のためのコンピュータ環境づくり
研究室 LAN の広域ネットワーク接続とその利用

II. ネットワークを利用した DNA 塩基配列データベースの日々更新と電子メールのためのデータベースサーバーシステム

研究室 LAN の広域ネットワーク接続の現状について述べ，その応用例として USENET ネットワークニュースに日々リリースされる DNA 塩基配列データを取り込み，DNA データベースを日々更新するシステム及び電子メールのためのデータベースサーバーシステムについて報告する。

1. 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 に加入していたため電子メールの送受信は可能であったが送受信に時間がかかったので，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（和光）との間で UUCP 接続をおこなった。また ISDN64 を使用した64kbps の IP link を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 links は DNA 塩基配列データベースの daily update のために多量の DNA 塩基配列データの送信が必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 は当研究室だけでサブネットを構成し，ワークステーションをルーターとして用い群馬大学の campus LAN に接続されている。サブネット化は研究室の LAN を campus LAN から独立させることで，security と完全性を高めるために採用した。また ISDN64を使用した IP 接続をするためにも経路制御の観点からサブネット化が望ましい。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 (Japanese 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) による 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。

広域 IP ネットワーク JAIN への接続に伴い，図1にあるよ

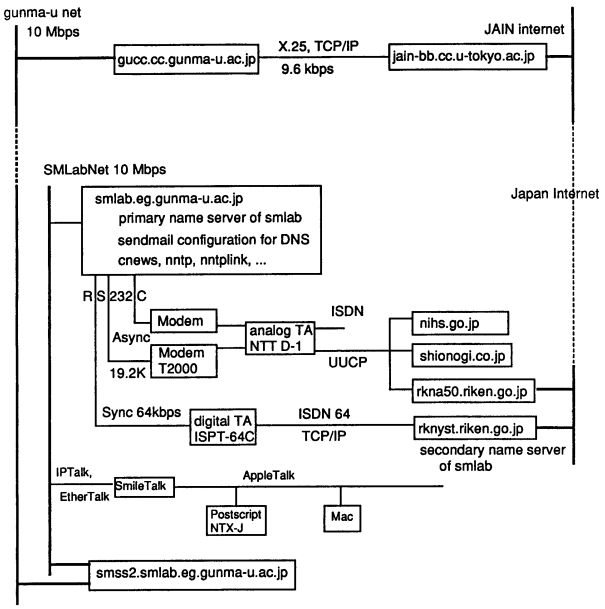

図 I Wide Area Network Links at smlab.eg.gunma-u.ac.jp

DNA/proteins sequence/structure databases		
GenBank	EMBL	eco
GenPept	SwissProt	PIR
PRF	kabat	HGM10
prosite	enzyme	rebase
PDB		
Proprietary softwares		
For networking		
SunLink IR	SunLink DecNet	
Others		
Sun Fortran	Sun Pascal	
IMSL library		
S	Mathematica	
Sybase		
Public Domain Softwares		
For networking		
BIND 4.8.3	sendmail 5.65+1.6W	
in.telnetd	in.ftpd	ftp
traceroute	nslookup	
amd		
wais	irc	
xarchie	xwebster	
X11R5		
X11R5/Wnn/iroha	xgks	
xpic	xkey3	xfig
pbmplus		
GNU		
emacs/nemacs		
gas	bison	gawk
gcc 1.39	g++ 1.39	ansi2knr
ghostscript	graphics	gnuplot
TeX		
TeX3.0	amsTeX2.0	NTT jtex1.01
News		
bnews 2.11	cnews	
nntp	nntplink	
rn/rrn	xrn	
DBMS		
postgres	picaso	
Sequence analysis		
FLAT		
Others		
kcl	akcl	
perl	f2c	p2c
libtiff		
mh		
less	jstevie	nkf
kermit	xmodem	zmodem
...		

表 I Databases and softwares installed at smlab.eg.gunma-u.ac.jp

うにSMLabNetの一計算機(smlab.eg.gunma-u.ac.jp)でsmlab domainのprimary 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とのmailの送受信は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はSMLabNetにおけるソフトウエア環境を示す。

2. ネットワークを利用したDNA塩基配列データベースの日々更新システム

DNA塩基配列データベース(GenBank/EMBL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。GenBank/EMBLデータベースは各々年4回リリースされ，Internetを介してanonymous-ftpにより入手することもできる。しかし，近年の分子生物学の研究においては，年4回リリースはリリースサイクルとして充分ではない。そこでGenBankはUSENET Bulletin Boardにnewsgroupを作成し毎日データをそのニュースグループにリリースしている。同様にEMBLはmailing listにより入力された新データを毎日リリースしている。図2はこのようにして日々リリースされるGenBank及びEMBLのデータ量を示す。日により大きく変動するが，平日は200KB-1MBのデータ量に達する（このデータは既にリリースされたデータの更新データも多数含まれるので，日々のリリースデータを加えたものはかならずしも年4回のリリースに対応しない)。このようにしてリリースされるGenBank及びEMBLデータをデータベースとして取り込むシステムを稼働させているので紹介する。システムはデータ取り込みシステムとデータベース更新システムからなり，smlab.eg.gunma-u.ac.jpではその両方を稼働させ，取り込まれたデータは数カ所に送られている。その内の一つシオノギ製薬研究所では同じデータベース更新システムを稼働させている。また類似のデータ取り込みシステムが理化学研究所 安永照雄研究室でも稼働している。EMBLデータベースの場合は新データは電子メールで配送されるので，データ取り込みシステムは不必要で電子メールによるデータベース更新システムのみが必要である。

2-1. USENETネットワークニュースからのデータ取り込みシステム

GenBankのデータはUSENET Bulletin Boardのnewsgroup(bionet.molbio.genbank.updates)にリリースされている。genbank.bio.netによってこのnewsgroupに投稿されたデータは図3に示したルートを通ってsmlab.eg.gunma-u.ac.jpに到達する。news articleとして到着したデータは最も容易な方法，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。(図4参照)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。このようにしている理由はGenbankから一遺伝子データが一つのnews articleとして送られてくるので，ある程度まとめて送ったほうがシステムの負荷を軽減できるからである。なおこの際，Bitnetでは最大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。

2-2. 電子メールによるデータベース更新システム

このようにデータが毎日リリースされるので，データベースも日々更新することが要求される。このような更新の頻度を考えると，更新がデータベースの再構築を必要とするようなデータベース管理システムは計算機への負荷が大きく，データベースへの追加が可能な管理システムが望ましい。

smlabにおけるデータベース更新システム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。

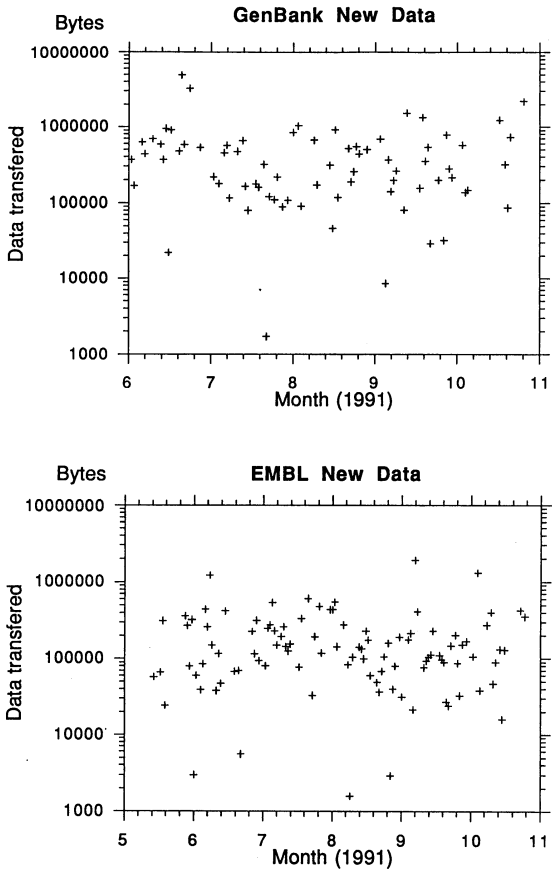

図2 The amount of new data in the daily release of the GenBank and EMBL databases

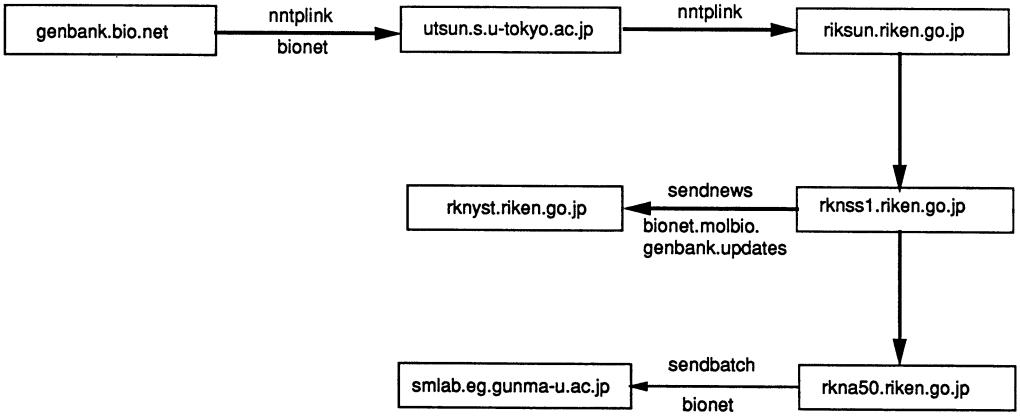

図 3 Route of network news transfer of the bionet newsgroup to smlab.eg.gunma-u.ac.jp

インデックスファイルも，一行にエントリー名とそのエントリーのファイル上の位置を記したフラットファイルである。データは複数のファイルに分割して保持出来る。メールとして到着した更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。これらのファイルは一日分ごとに別ファイルとして管理されている。エントリーの retrieval は，これら複数のインデックスファイルにおいてエントリー名を検索することによりなされる。エントリー名は正規表現を用いて指定する。新データが先に取り出されるよう，日付けの新しい順に検索される。

3. DNA，蛋白質配列フラットデータベースのためのデータ検索システム（FLAT）[1]と電子メールのためのデータベースサーバー

このようにして日々更新される最新のデータベースは，研究室で使用されるだけでなく，電子メールを用いて誰でも検索できるように，開放されている。

検索システム(Flat)は簡単な機能を果たす様々なツールからなる。基本ツールの例は，

－指定された著者名，論文名，accession numbers を検索しエントリー名を出力する
－エントリー定義ファイルにおいて文字列を検索しエントリー名を出力する
－指定されたタイプのレコードを出力する
－指定された文字列を含むエントリーのエントリー名を出力する
－エントリー名からなるセットに関する and，or，xor。
－指定されたエントリーを出力する

等である。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

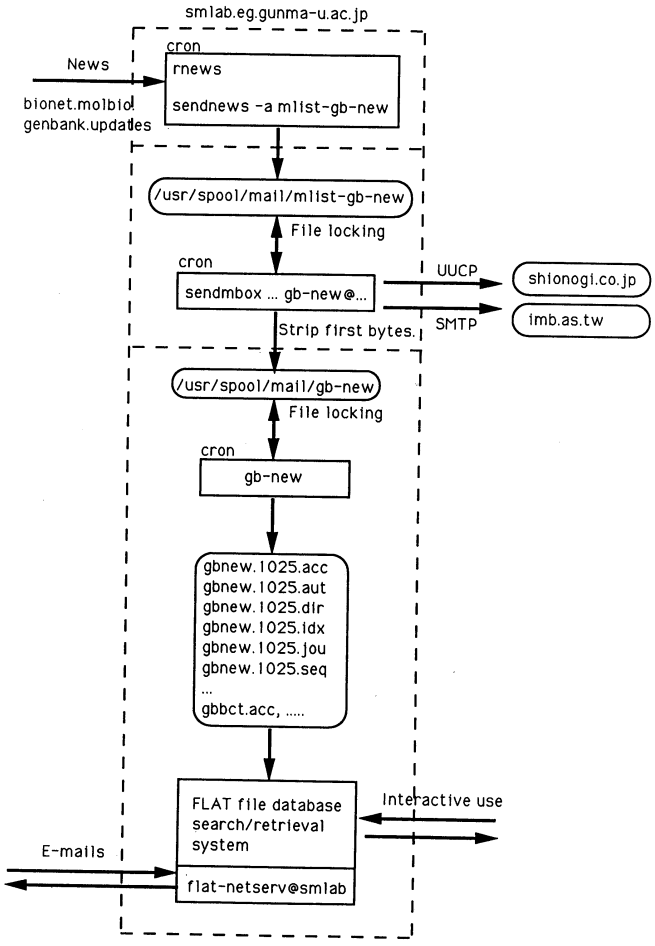

図 4 Data flow at smlab.eg.gunma-u.ac.jp

塩基配列を正規表現で表現できる。この検索システムがサポートするフォーマットは現在 GenBank, EMBL フォーマット及び SwissProt, PIR, PRF フォーマット等である。このソフトウエアは, UNIX システムなら移植可能であると言う利点を持つ。

この検索システムのコマンドの一部（著者名, 論文名, キーワード等による検索, 及びエントリーの retrieval）及び配列の類似性検索プログラムは電子メールによっても利用可能である。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 に送ると検索結果が送り返される。このようなシステムでは, system security を守ることが重要となる。まずメールは mail header が除去された後, PATH 変数, directory を適当に設定し, restricted shell(sh-r) にデータとして渡され処理される。restricted shell では PATH 変数のリセット, 入出力の redirection 及び cd コマンドは使用出来ないことに注意願いたい。その機能によりメール内では, PATH 変数に設定された directory 内にあるコマンドしか使用できない。一方, 配列の類似性検索コマンドはファイルからのみ入力データを読み取るため, 標準入力からデータを読み取りファイルとして書き出すコマンドが追加されている。もちろんファイルはその directory にしか作成されないようになっている。このようにして処理された結果（restricted shell からの出力）はメールの From line から得られた e-mail address に送り返される。

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ネットワークニュースからの DNA データ取り込み, 電子メールによるデータベース更新システム, 及び電子メールのためのデータベースサーバーシステムを簡単に報告した。データベースの管理は time consuming である。すべてのユーザーが個々に管理するのは不可能であろう。今後はネットワークの進展とともに, 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有用であろう。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整備が肝要である。(1992年 4 月)

(追補) その後、ニュースシステムへの GenBank データのリリースの停止(1992/10)に伴い, anonymous-ftp による日々更新に変更された。現在、データベースの提供は電子メールによる利用に加え, wais による提供, "whois"コマンドによる提供も公開されている。

参考文献

1) Miyazawa, S. :DNA Databank of Japan:Present Status and Future Plans. in "Computers and DNA", Santa Fe Institute Studies in the Sciences of Complexity, Eds. G. Bell and T. Marr (Reading, MA: Addison-Wesley), Vol. VII (1989) 47-61.

宮澤三造（群馬大学工学部）