広域ネットワーク「ゲノムネット」の構築と分散処理環境のもとでの遺伝情報処理

Wide-area Network "GenomeNet" and
Distributed Genetic Information Processing

京都大学化学研究所　荻原淳・金久實
Atsushi Ogiwara, Minoru Kanehisa
Institute for Chemical Research, Kyoto University

1．はじめに（ゲノムネットの紹介）

広域ネットワーク「ゲノムネット(GenomeNet)」とは、文部省科学研究費重点領域研究「ゲノム解析に伴う大量知識情報処理の研究」のプロジェクトの１つとして運営されているものである。これは国内のゲノム研究の主な研究拠点の間を接続するコンピュータネットワークであり、研究者相互の情報伝達のほか、データベースの配布や利用並びに構築、計算機資源の共有による分散処理などを目指したものである。通信に使われているプロトコールはTCP/IPとDECnetであるが、ここでは特にUNIXワークステーションの普及と相俟って広く使われているTCP/IPに関して紹介する。

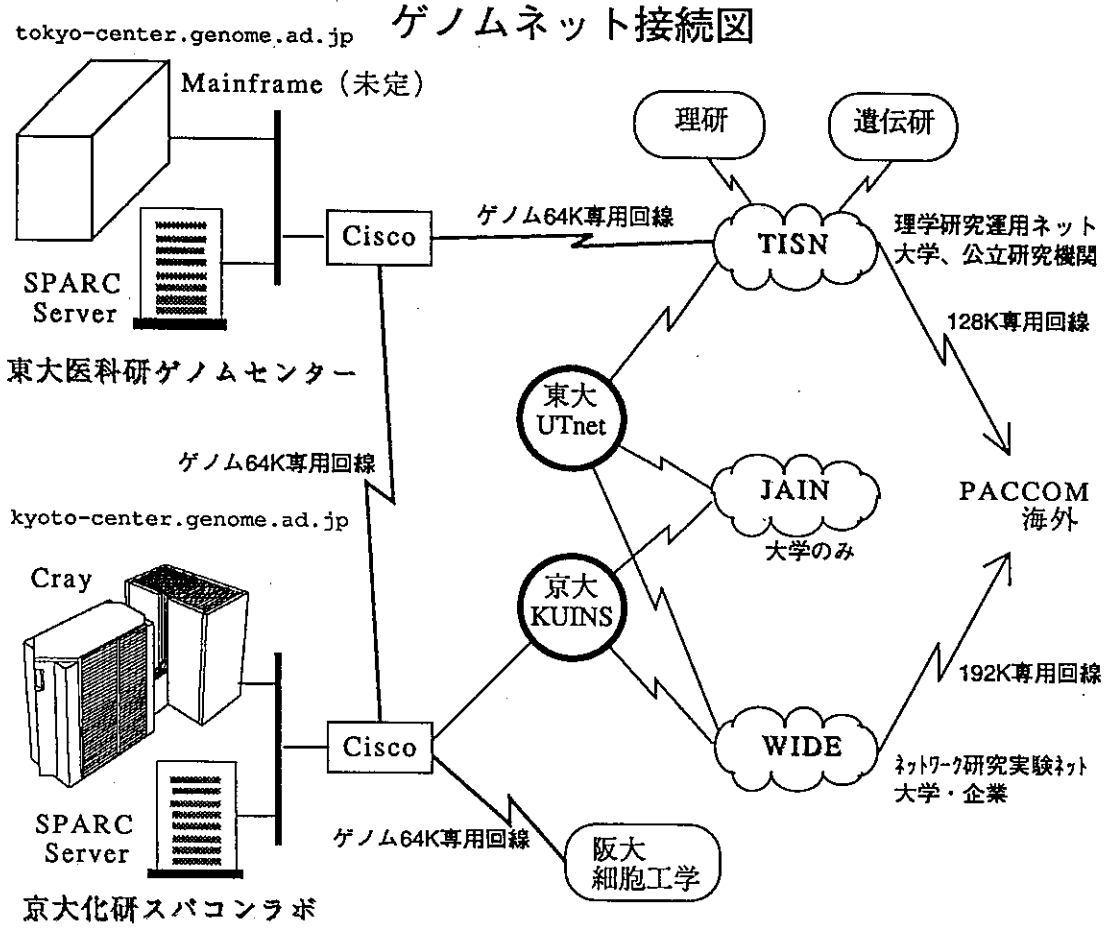
ゲノムネットは、現在日本のアカデミックインターネットを構成しているWIDE, JAIN, TISNのうちのTISNに参加している。このTISN の回線を経由することで海外のインターネットにも到達可能となっており、国内だけでなく海外とも直接情報交換ができるようになっている。

我々はゲノムネットを国内のゲノム研究の発展を支えるインフラストラクチャと位置づけ、１年以上前からTISN等関係機関と交渉、アドレス取得などの準備作業を行なってきた。本年９月２０日に東京－京都両センター間のバックボーンが開通したばかりである。接続がこれからのところも多いし、まだアプリケーション運用は試験段階のものもあるが、運用が順調に軌道に乗ってきたところでその経過と今後のプランを紹介するものである。

本稿ではまずネットワークの接続の現状と接続にあたっての技術上の問題点などを述べたうえで、提供されるアプリケーションサービス、データベースサービスなどを説明し、最後に現在試験段階にあるネットワークを利用した分散処理による遺伝情報処理システムについて紹介する。

2．接続経路と相互接続

ゲノムネットの拠点は東京大学医科学研究所のヒトゲノム解析センターと京都大学化学研究所に設置されており、２つのセンター間を基幹回線が通り、これらのセンターを中心に各研究機関まで回線がpoint-to-pointで延びているという構成になっている。１９９１年１２月現在の構成は図のようになっている。TISNのセンターがある東京大学理学部情報科学教室から出た回線は１つは京都大学化学研究所につながり、もう１つは東京



大学医科学研究所に届いている。京都センターからはさらに大阪大学細胞工学センターまで延びている。この間はすべて 64KBPS のディジタル専用回線が使われている。

現在国内にはアカデミックな利用を目的とした広域ネットワークがいくつかあるが、そのうち WIDE, JAIN, TISN の 3 者は参加組織がもっとも広く、また相互接続をもつことで事実上日本のインターネットを形成している。ゲノムネットはこのうち TISN に属しているために当然TISNとの接続をもっているが、このほかに東京センターで東大を介して WIDE および JAIN にもつながっている。また京都センターでは京大学内 LAN を通って WIDE, JAIN に到達可能となる。IP の経路制御では、出発点から到達点まで複数のネットワークを経由して行くときには、基本的には各ゲートウェイでローカルに経路が選択され全体の経路は不定である。このように複数のネットワークとの接続がある場合も理想的にはその時の回線の状況に応じて最適経路を通ることが望ましいが、現在国内のインターネットで採用されている RIP (Routing Information Protocol) ではそこまでの制御が難しい。またネットワークプロジェクトの性格によって例えばJAIN の上は非アカデミックサイト同士のトラフィックの通過を禁止しているなど、現実の経路制御では様々な要素を考慮しなくてはならない。他のプロジェクトネットワークなどでは外部との通信の乗り入れを排除する方向で管理しているものもあるが、我々はむしろ既存のネットワークに対してオープンな姿勢でゲノムネットを構築した。相互乗り入れ地点で

の経路制御の在り方は、基本的には経路情報をどのように内外に流すかに依っているが、さらに経路情報のメトリックを操作するなどして外部ネットワークにわたるトラフィックはなるべく押さえるよう調整されている。

3．提供されるサービス

TCP/IP上で提供されるサービスには多様なものがあるが、このうち電子メールはもっとも代表的なものである。これについては改めて説明するまでもないが、内外のインターネット接続サイト、JUNETなどのUUCPサイトと相互にメールのやり取りが可能である。特にインターネット接続サイトへのメールはドメインネームサーバーのメール交換レコードを使用することで直接に配送されるようになっている。またメールシステムは利用者間の情報交換の他に、検索プログラムなどを通して直接問い合わせを行なえるようなサービスもデータベース作成機関などで行なわれている。特にホモロジーサーチプログラムFASTA, BLASTなどを用いて利用者から送られた配列をサーチして結果をメールで返すメールサーバーサービスが生物学者の間で便利に使われるようになってきた。このようなサービスもゲノムセンターなどの設備が整い次第、始められる予定である。またデータベースの更新にもメールは利用されているが、これは次項で述べる。

このほかに利用者間での情報交換に良く利用されているものとして、ニュースシステムがある。これにはアメリカのbionetのように生物学関係の情報交換のためのグループなども作られており、JUNET でもつい最近 fj.sci.bio というニュースグループが作られた。ゲノムネットとしてはまだニュースシステムを稼働させていないが、近いうちに利用可能となる予定である。ただ、ニュースシステムに関しては運用上の問題点などが議論されてきており、よりアカデミックインターネットの実情に合わせたシステムに再編しようという動きもある。

また、データベースからデータを直接取ったり、公開されている解析プログラムなどを自分のところにもってくるのには、ftpが欠かせない。特に、利用資格がなくてもアクセスできるanonymous ftpがいくつかのサーバーサイトで公開されるようになってきた。

代表的なanonymous ftpサービスサイトの一例

hostname	services
embl-heidelberg.de	EMBL, SwissProt, PDB, DROMAP, ECD, ENZYME, etc
genbank.bio.net	GenBank, GenPept, EMBL, SwissProt, ALU, Prosite, etc
ncbi.nlm.nhi.gov	NCBI software, ENZYME, EPD, LiMB, BLAST, FASTA, etc
irisc2.chm.bnl.gov	PDB

ゲノムのセンターでもサーバーの設備が整い次第GenBank, EMBL, DDBJ, PIR, SwissProt, Prositeなどの代表的なデータベースの他に、我々の行なっているMotif Dictionary, Enzyme DataBase, AAindexや蛋白質奨励会のデータベースなども公開するため準備中である。ただ、ゲノムネットを含めた日本のインターネットのように比較的低速回線をバックボーンとしてもつところではftpは重負荷である。国内の利用者が直接海外にアクセスを繰り返すと現在128KBPS + 192KBPSしかない海外リンクをかなり圧迫するので、データベースサーバーを国内にも複数分散させることで転送負荷の軽減を計る必要がある。

このほか、遠隔のホストのアカウントをもっている場合には telnet 等の遠隔端末機能を利用し直接そのホストで仕事を行なうことも、専用回線によるスピードアップでスムーズに行なえるようになった。

4．分散データベース

生物分野のデータベースは大容量のテキスト形式のものが多い。大部分のサイトでは通常のファイルの形で扱われており、定期的に一括更新されている。このようなものはネットワーク上でも ftp やメールといった通常のアプリケーションで扱うことは可能である。例えば京大化研では現在 GenBank の最新エントリーの配付を遺伝研を経由して電子メールで受けている。このデータは受けるとすぐににフラットテキストのデータベースに追加され、同時にインデックスも更新されるようになっているため、常に最新のデータが検索可能となっている。ただデータベースの整合性のチェックまで行なわれていないためメールの不着などがあってもこれではわからないのが難点である。データベースの整合性を保証するのにもっとも簡単にはリリースがフィックスしたものを一括して配布する方法があり、現に多くは磁気テープを使ってこのように更新している。これをそのまま ftp などでネットワーク上で行なおうとするとデータ量が余りにも膨大なため回線がパンクしてしまう。

一方、GenBank などのデータベースを作成しているところでは、内部的にはリレーショナルデータベース管理システム (RDBMS) を導入してトランザクション管理している。この RDBMS としては米国のゲノム分野では Sybase を利用しているところが主流であるため東大ゲノムセンターでも Sybase を導入している。ここでも内部的なデータベースの維持に使われるほか、Sybase の 1 つの特徴である分散型データベースの構築にも利用される見込である。Sybase は東京センターのほかに京都センター（京大化研）にも導入されるため、少なくとも 2 箇所のサーバーでサービスが可能となる。Sybase を利用したデータベースとして、GDB, GenBank, GenInfo などのサービスを検討中である。

5．分散処理

ゲノムネット上では東京、京都の 2 つのセンターを始め、各地でデータベースサーバーおよび計算資源サーバーのサービスが行なわれることを期待している。データベースサーバーについては前述したが、計算資源サーバーとしてはメールによるホモロジー検索サービスなどがある。このようなジョブは京都センターで来年 1 月から稼働するスーパーコンピュータラボラトリの Cray Y-MP2E でも空時間を利用して提供する予定である。またメールによるバッチ形式の処理のほかに、クライアントーサーバーモデルに基づくより対話性のよい検索システムを現在構築中である。これはデータベース検索・解析統合システム IDEAS の新リリースという形で進められており、現在 SEQMAN の GET などに相当するエントリー抽出プログラムのネットワーク対応の試作版が試用されている。このクライアント部分は X ウィンドウ対応版なども設計されており、よりユーザーフレンドリーなものがネットワーク上で利用可能となるであろう。