The DNA database has been managed in a flat-file system at the DDBJ since 1985. The flat-file system is inadequate for building and searching the DNA database which is receiving an explosive increase entries. We carried out a transformation from the flat-file system to the relational database system with GenBank staff. The schema of the relational database was designed as follows:

1. Decomposing the DNA data into both structured and non-structured data
2. Partitioning large tables into small tables without update anomaly
3. Making a flexible relationship among tables to represent complex data

This schema provided the capability for building and searching the DNA database with less memory on the relational database system. However, the schema was implemented as a complex network structure with about 60 tables. It is difficult to use the SQL search language of the relational database system with this schema.

We defined and simplified the schema for easy use of the commands using the view function of the relational database system on the existing schema. The simplified schema implemented in the view function was defined as LOCUS, DEFINITION, ACCESSION, KEYWORDS, SOURCE, REFERENCE, FEATURES, ORIGIN, and SEQUENCE tables which are virtual tables without storing real data. It represents aspects of the traditional DDBJ/EMBL/GenBank data format which are familiar to biologists using the flat-file system.

Users can easily join these virtual tables using attribute storing accession numbers. Since we developed the simplified schema, users are able to use the SQL search command easily and get quick response in DNA data searches.

1. はじめに

DNAデータベースの効率的な構築や検索をはかるために、関係データベース管理システムSYBASEを用いている。データ構築の際は、更新時異常をできるだけ抑えるようにデータベーススキーマを設計しておく事が大切である。このため、この種のスキーマは、テーブル間の関係が複雑なることともに、テーブル数が多くなる傾向にある。特に、生物情報は事務処理用のデータベースとは異なり、それ自身が複雑であるため、データベーススキーマがより複雑になると思われる。実際、DDBJの構築業務に利用されているGenBankスキーマは、約60個のテーブルから構成されており、テーブル間の関係が大変複雑になっている。EMBLでは、さらに多く、約100個のテーブルから構成されている。

著者らは、DDBJ（DNA Data Bank of Japan）で構築されたデータの統計量や信頼性を把握するために、関係データベース管理システムSYBASEのSQL機能を用いてテーブルにアクセスしようとしたが、テーブル数が多い事やテーブル間の関係が複雑なため、なかなかように検索質問文をSQLで書く事ができなかった。

本発表では、図1に示すようなDDBJ/EMBL/GenBankフォーマット（DNA塩基配列に対する属語性記述のための文法）のイメージに近い検索ビューを提案する。また、一般的利用者に、構築されたデータをある程度のスピードで提供するため、簡易な利用者向きデータベーススキーマを作成する方法について提案する。また、このスキーマ上におけるSQLの電子メールサービスの可能性についても述べる。

2. 関係モデルの導入

DNAデータベースは10万件を突破するような勢いで増加しているが、このようなデータをフラットファイルで構築・管理していくには、もはや、効率や機能の面で限界に達しているといえる。著者らは、関係データベースシステムSYBASEを用いて、昨年（1991年）の10月に米国ロスアラモス研究所のGenBankスタッフらの協力により、構
薬用のデータベーススキーマ（GB-Schema）をDDBJにインストールした。このスキーマ上に、構築者がデータを1件ずつ登録するようなインタフェース（Annotator's Workbenchと呼んでいる）を搭載したこともあり、データ構築作業がかなり楽になった。しかし、このスキーマは、図2に示すように、60個ものテーブルが相互に関係づけられた複雑なネットワーク構造を持っている。このため、構築者が登録されたデータ全体を監視したい場合、このようなテーブルを対象として関係データベースの検索質問コマンドSQLを対話入力するには大変な時間を要する。著者らの測定によると、検索質問コマンドを対話入力するのに数十分から数時間も費やしたことがある。検索質問コマンドSQLは、高水準で強力な検索能力を持つが、テーブル数が多くなると、全体の把握が難しくなるため、わずか数行のコマンドを対話入力するのでも大変苦労するのである。

著者らは、これを解決するために、関係データベース管理システムを用いて、このような複雑なテーブル間ネットワークを利用者にとって簡明易いテーブルイメージ（DDBJ/EMBL/GenBankフォーマット）として見せるような検索ビュー（仮想テーブル）を実現した[3]。著者らは、図3に示すように、できるだけ、LOCUS、DEFINITION、ACCESSION、KEYWORDS、SOURCE、REFERENCE、FEATURES、ORIGINといった基本部分（データフォーマットの基本単位）ごとに仮想テーブルを定義するようにし、仮想テーブル間は、必要に応じて登録番号（アクセスション番号）などでジョイントできるように構成した。図中の実線及び矢印は、それぞれ、等号ジョイント及び外ジョイントを示す。以上により、非定型的な検索質問コマンドが直ちに対話入力できるようになったと考えている。また、検索時にフラットファイルに比べて格段に高速化されたので、比較的膨大な時間を費やしていたDNAデータベースの統計報告なども比較的簡単に算出することができた。さらに、構築者が、SQLの集合を扱う能力を習得できるようになれば、DNAデータベースの月間処理などの管理業務などもこの検索ビューで容易に可能になる。

3. 検索ビューの応用

現在、DDBJではこの検索ビューを用いて、オンライン利用者や電子メール利用者を対象にし、SQLサービスの可能性について検討を進めている。構築用スキーマは、構築作業だけを意図して設計されているので、利用者向きに改善する余地がある。そこで、著者らは図4に示すような簡易な検索用スキーマを作成している。検索用スキーマは新たなデータマネージメントを必要とするが、検索ビューのスキーマを用いるよりも高速な検索が期待される。構築用スキーマから検索用スキーマへのデータ変換は、著者らの検索ビューと制御フレームを用いたプログラミングにより実現している。我々は構築用スキーマにおける最新のデータをSQLのトリガー機構により認識し、検索用スキーマに自動的に転送する方法を検討している。これにより、構築用スキーマと検索用スキーマ間のIntegrity Constraintsが保持される。このような2分割型のアーキテクチャは、並列計算機上で運用する時の負荷分散の手段にもなるであろう。以上の詳細は発表時には述べる。

4. まとめ

ここでは、GenBankスキーマを用いて、DDBJ/EMBL/GenBankフォーマットのイメージの検索ビューについて提案した。また、この検索ビューを用いて、検索用スキーマを作成する方法について述べた。今後は、検索用スキーマの評価を行なっていく予定である。また、最近、話題になっているオブジェクト指向データベースとの関係[4]についても検討を加えていく予定である。

[3] 北上礼、山崎由紀子、後藤英弘、嘉藤成也、個野義男。五條模型：関係データベースにおける生物情報の構築と検索、日本微生物保存連盟総会シンポジウム、1992年6月。
[4] 有川正俊、牧之内利文、北上礼、山崎由紀子。五條模型：遺伝子データベースのオブジェクト指向設計、平成4年度電気通信学会九州支部連合大会論文集、1992年10月。
Fig. 1 Tree Structure for File Formats. (DDBJ/EMBL/GenBank)

Fig. 2 A Schema Building DNA Database.
Fig. 3 A Schema Implemented in View

Fig. 4 A Schema Searching DNA Database.