DNAデータベースの構築と検索

1G−8

北上始, 山崎由紀子, 鵜川義弘, 池尾一穂,
斎藤成也, 舘野義男, 五條堀孝

国立遺伝学研究所　日本DNAデータバンク

1. はじめに

DNAは、生物の設計情報であり、4種類のヌクレオチド、A, T, C, Gの並び即ち塩基配列としてコードされている。DNAデータベースは、塩基配列とその配列に関する付属情報（生物学的意味を含む情報）から構成されている。

DNAデータベースの構築は、DDBJ（日本DNAデータバンク)、NCBI（米国バイオテクノロジー情報センター）、EMBL（欧州分子生物学研究所）の三箇所で行われている。我々が所属するDDBJは静岡県三島市、NCBIは米国のワシントンDCにほど近いベセスダ市、EMBLはドイツのハイデルベルグ市にある。これら三箇所のデータバンク間では、毎日、コンピュータネットワークを用いて、最新のDNAデータが交換されている。交換されるデータには、新しく作成されたデータの他に、既に交換済みのデータについての更新情報も含まれている。DNAデータベースの各レコードは、図1に示されるような記述文法[1]（DDBJ/EMBL/GenBankフォーマット）で表現されたテキストデータとしてファイルシステムに保存されている。しかし、DNAデータベースは10万件を突破する勢いで指数関数的に増加しているため、ファイルシステムによる構築や管理は、もはや効率や機能面で限界に達している。

本発表では、UNIX環境下の関係データベースを用いてこのような問題点を解決するための方式やオブジェクト指向データベースとの関係について述べる。

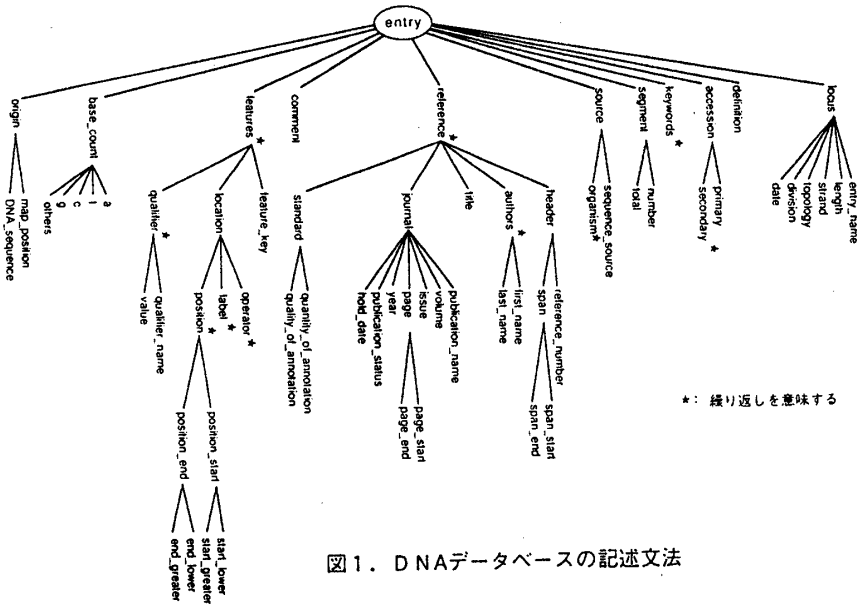

図1．ＤＮＡデータベースの記述文法

2. システム構成

DDBJのデータ構築作業には、(1) 生物学者が解読したDNA配列データに予め取り決められた付属情報を正しく付加する作業、(2) DDBJのデータとNCBI（あるいはGenBank）やEMBLのデータと統合する作業がある。以下では、前者を構築作業#1と呼び、後者を構築作業#2と呼ぶ。三箇所のDNAデータバンクには、同じデータが含まれていたり、異なるデータでも同じような付属情報が含まれていたりするので、統合の際は、統合データベースの冗長性や矛盾が排除される。

図2に、現在開発中のシステム Alife(A large-scale database integration facility with e-mail server)の構成を示す。DDBJでは、関係データベース管理システムとしてSyBaseが用いられている。

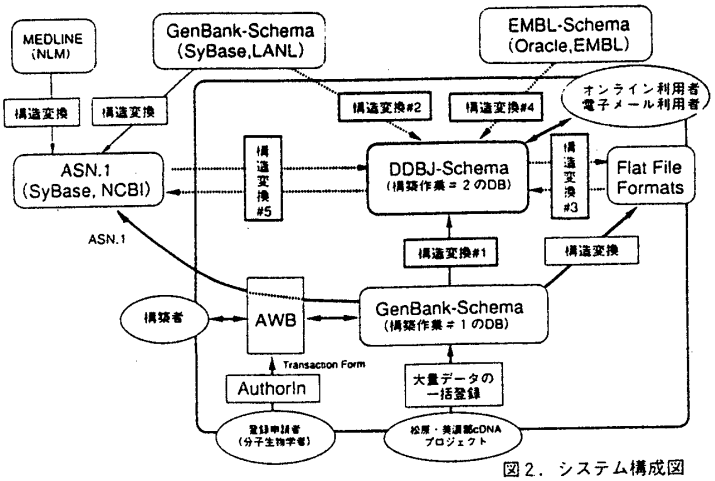

図2．システム構成図

図の構築作業#1用DBで利用されている関係スキーマは、LANL(米国のロスアラモス研究所)が開発したものを利用している。このスキーマは、GenBank-Schemaと呼ばれてる。

図2の構築用インタフェースAWB(Annotators Workbench)は、構築者がこのスキーマに対して1件ずつDNAデータを構築するシステムである。ヒトゲノム解析の松原・美濃部cDNAプロジェクトで代表されるが、大量のデータを一括登録するような場合は、十分な付属情報が付与されているという事情があるため、構築者の手が不要である。このため、著者らは、これを専用に処理するためのシステムを開発した。

図2の構築作業#2のDBにおけるデータベースの統合化は、DDBJ独自の関係スキーマDDBJ-Schemaで行なうことを考えている。また、このスキーマは、図1をよく知っている分子生物学者にとって馴染みやすいスキーマであることから、著者らは、オンラインユーザや電子メールユーザが使えるシステムを開発している。構築作業#1のDBのGenBank-SchemaからDDBJ-Schemaへの構造変換は、図2の構造変換#1で行なわれる。構造変換#1と構造変換#2はほぼ同じ

Building and Searching DNA Database
Hajime Kitakami, Yukiko Yamazaki, Yoshihiro Ugawa, Kazuho Ikeo, Naruya Saito, YoshioTateno and Takashi Gojobori
DNA Data bank of Japan, National Institute of Genetics

機能であり、その中核部分が完成している。構造変換#3,#4,#5は、今後、開発する予定である。図2のASN.1(抽象構文記法)はNCBIが提案しているDNAデータの記述言語を指すが、構造変換#5においてはこの言語により表現されたDNAデータが取り扱われる。

3．関係スキーマ

構築作業#1用DBでは、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になるとともに、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かなり複雑になっている。実際、構築作業#1におけるDBのGenBank-Schema [1] 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-Schemaでは、さらに多く、約100個のテーブルから構成されている。

このような中で、著者らは、構築作業#1のDBで作成されたDNAデータベースの管理業務を簡単に行なうようにするために、図1に示すようなイメージに近い26個の検索ビューの実現に成功している。検索ビューの定義に用いられているSQL言語は、Cなどのプログラミング言語に比べて言語水準が高いため、この様に複雑なGenBank-Schemaを図1のような視点で論理的に整理するのが容易になった。具体的には、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（アクセッション番号）などでジョインできるように構成した。これにより、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。また、検索時間も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情報なども比較的簡単に算出することができた。

さらに、この検索ビューを用いて15個のテーブルから成る構築作業#1のDB向けにDDBJ-Schemaを実現している。図3にDDBJ-Schema [1] を示す。

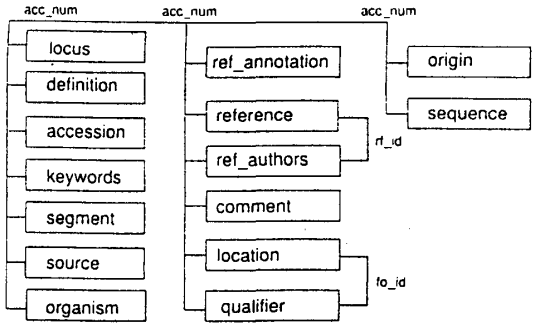

図3．DDBJ-Schema

4．構造変換

構築用スキーマGenBank-SchemaからDDBJ-Schemaへのデータ変換は、著者らの検索ビューと制御フロー言語を用いたプログラミングにより実現している。我々は構築作業#1に於て得られる最新データの登録番号をSQLのトリガー機構により記録し、UNIXのcronデーモンによりDDBJ-Schemaに自動的に転送する方法を実現した。これにより、構築用スキーマと検索用スキーマ間のIntegrity Constraintsが保持されている。このような2分割型のアーキテクチャは、分散型計算機環境や、並列計算機環境などにおける負荷分散の手段にもなるであろう。

5．オブジェクト指向データベースとの関係

現在までのDNAデータベースの開発経験に基づいて、関係データベースの問題点について触れ、オブジェクト指向データベースとの関係について述べてみたい。

（1）SQLの記述能力

検索ビューは、SQLを用いて定義されるが、1つのSQL文で2つ以上の問合せ結果を足し合わせる記述能力がない。このため、ACCESSION, REFERENCE, FEATURESなどを表現する検索ビューが、各々、2個以上定義されている。例えば、ACCESSIONの検索ビューとしては、プライマリの登録番号とセカンダリの登録番号を別々の検索ビューとして定義しておかなければならない。また、SQLによる問合せ文に複数の等号ジョインがあるときに、外ジョインが同時に使用できないという制限がある。外ジョインは、論文の著者名や論文を掲載する雑誌名がまだ定まっていないデータを検索するときに利用される。このため、検索ビューを用いて完全に実テーブルperson, publicationを見せなくするようにはできなかった。さらに、Integrity Constraintsの定義では、SQL以外を用いる事ができないが、制御フロー言語が利用可能になれば、インクリメンタルなデータ変換に、UNIXのcronデーモンを利用しなくてもよくなる。以上は、いずれもSQLコマンドが柔軟に拡張可能な構造になっていれば、問題がないと思われる。このような柔軟性は、オブジェクト指向データベースにおけるメソッドの柔軟な定義に期待する事ができる。

（2）データの表現

関係データベースは2次元のテーブルの中にデータが格納されるようなデータ構造である。このため、図1で表現されるようなデータを格納するには、これを分割し複数のテーブルに分配する必要がある。従って、もとのDNAデータを得るためには、利用者は、それらのテーブルを結合しなければならない。この様に、関係データベースはテーブル間を結合するような情報をデータベースとして管理するパラダイムをもたない。このため、そのような結合に関係した情報は利用者か応用プログラムにより直接管理されている。オブジェクト指向データベース [2] では図1で表現されるような複雑な情報を直接扱えるので、DNAデータはクラスのあるインスタンスとして格納できる。また、データを分割したとしても、分割されたデータ間をポインタで結合するような機能が用意されているので、そのような問題が生じない。

さて、セカンダリの登録番号を複数もつ場合や、論文の著者が複数の場合は、1件のDNAデータに対して複数のタプルがテーブル中に存在する。これは、DDBJ-Schemaを図1の形式に変換するようなプログラムを作成するときに、テーブルスキャン操作がかなりの煩わしくする。この様な複数個のレコードをリスト構造のようなもので表現できれば、これらを1件のタプルとして表現できる。そのような仕組みも、オブジェクト指向データベースのパラダイムで容易に解決可能である。

（3）応用プログラムとの親和性

構造変換#3, #5を関係データベースの応用プログラムとして実現しようとする場合に、予想される事だが、一度定義したスキーマと同じスキーマを応用プログラム内で定義しなければならないという煩わしさがある。これにより、データベース側のスキーマに修正があると、これに対応してプログラム側も修正しなければならない。オブジェクト指向データベースで、応用プログラムをオブジェクトとして扱うと、応用プログラムは既存オブジェクトとメッセージ通信だけで情報のやり取りができる。このため、この様なスキーマ再定義のような煩わしさがない。

6．おわりに

ここでは、関係データベースを用いて、DNAデータベースの構築と検索を行なうシステムAlifeについて述べた。今後は、構造変換#3,#4,#5を具体化する方法について検討を行なっていく予定である。また、5の（2）で述べた複数タプルの件についての問題が生じないようなDDBJ-Schemaの作成についても検討する予定である。さらに、オブジェクト指向データベースによりAlifeを実現する方法についても検討をさらに加えていく予定である。

［参考文献］

［1］北上始, 山崎由紀子, 鵜川義弘, 池尾一穂, 斎藤成也, 舘野義男, 五條堀孝：関係データベースによるDNAデータベースの構築と検索, 第3回ゲノム情報ワークショップ, 1992年12月.

［2］有川正俊, 牧之内顕文, 北上始, 山崎由紀子, 五條堀孝：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　電気関連学会九州支部連合大会論文集, 1992年10月.