9　バイオテクノロジー分野におけるサーチ手法に関する調査研究

近年のバイオテクノロジー分野における遺伝子解析技術の飛躍的な進展は、関連する特許出願数の増大と一出願に含まれる配列データ量の増大をもたらし、特許庁の配列検索システムに影響を与えている。このような状況において、配列検索システムや配列データ蓄積の在り方の見直し、また、配列検索システムの検索応答性の向上が求められている。そこで、本調査研究では、次世代配列検索システムを考慮し、⑴ システムの現状と課題、暫定対策の検討、⑵ データの蓄積の検討、⑶ データ圧縮の有効性の検討、配列データ圧縮の実機検証、⑷ 特許庁への意見・要望など、様々な角度から特許審査に適したバイオテクノロジー分野のサーチの在り方について検討を行った。

Ⅰ　ＤＮＡ配列データベースの現状と課題

1　特許庁内データベース

平成10年に現行の特許庁内ＤＮＡ検索システムが構築された。その後のＤＮＡ配列解析技術の発達と解析対象の多様化により出願数が増加し、特許庁内に蓄積される配列データ量が増大するような状況に、いかに対応すべきか様々な角度からの検討が必要である。

⑴　ＤＮＡ配列関連出願に関する審査の現状

ＤＮＡ配列関連出願の審査が一般的な技術の審査と異なる点は、先行技術文献の調査において、出願された配列データと先行技術文献に記載された配列データとを相同性検索により比較するホモロジー・サーチを用いることである。また、公開前配列データを含むＤＮＡ配列関連出願の審査では、ホモロジー・サーチの検索式に含まれる配列データそのものが発明の本質を表すため、検索式、検索結果などすべてを秘密に扱われなければならない。

⑵　特許庁内ＤＮＡ検索システムの現状

特許庁内ＤＮＡ検索システムは、検索ソフトウエアと、ＣＰＵ・ハードディスクなどのハードウエア、さらに検索対象である配列データに大別される。

検索ソフトウエアは、FASTA、BLAST、Smith & Waterman など一般的なものを利用している。その理由は先行技術文献調査に当たって、特殊な検索ソフトウエアではなく、一般的な検索ソフトウエアによるホモロジー・サーチで発見可能であったかどうかがポイントとなるためである。ハードウエアは出願数、データ量、利用頻度等の増加予測に基づき、定期的な増強が行われているが、近年の配列データ量の予測を上回る増加にハードウエアの増強が追い付かなくなりつつある。配列データは特許出願されたデータのほかに、登録機関に登録され、外部で公開済みのデータが含まれているという状況である。

⑶　特許庁内ＤＮＡ検索システムの課題と暫定対策

（ⅰ）　システム・審査への影響

近年のＤＮＡ配列に関する出願データの急激な増大により特許庁内データベースの容量不足が懸念され、また検索時間の増大を招くなど審査への影響が出始めている。

（ⅱ）　暫定対策

①　公的データベース（GenBank, **DDBJ**, EMBL）のNR（Non-Redundant）化対策

平成12年末からGenBankを基準に公的データベースの重複を排除し、特許庁内データベースの容量の空きの確保と検索時間の短縮化が可能となった。

②　キーワード検索機能の簡略化（インデックス・データの削除）

③　システム稼働時間の延長

上記暫定対策を行っても、近い将来に再びディスク容量不足となることは明らかで、データの圧縮、データベースのアウトソーシング化など、根本的な視点から検討を進める必要がある。

2　公的データベース

⑴　国立遺伝学研究所

（ⅰ）　DNA Data Bank of **Japan**（**DDBJ**）

DDBJ に直接登録されたデータは、評価（データ・フォーマット、データ表記及び基本的な生物学的意味の評価）・蓄積され、EMBL及びGenBankから毎日送付されるデータ、特許庁（**JPO**）からのデータと統合して公開されている。

（ⅱ）　将来の課題

①　公共データベースの在り方

・長期的安定稼働
・私有データベースを公開する機構としての機能
・データの形式と表記の標準化
・異なるデータ源からのデータ間の整合性の検証努力
・医学生物系データベース開発のガイドラインの提示、積極的な技術移転

などが挙げられる。

②　**DDBJ** の課題

・情報環境の整備

塩基配列データベースの増大に対応したコンピュータ・システムの増強が必要である。また、塩基配列データベースのようなバックボーン・データベースとネットワーク上に分散したデータベースを統合したデータベースの需要の増大と、新しい実験技術由来の画像データの流通が見込まれるため、**DDBJ** を含む主要な公共データベースは広帯域のネットワークへの接続が必要である。

・人材の確保

データベース構築という基盤事業の中核には医学生物学分野での博士課程修了相当の能力を持つ人材が必須であり、その確保には待遇・評価に十分配慮していく必要がある。

③　**DDBJ** の具体的課題

・データの受付→評価→公開の長期的な安定稼働
・データ更新の迅速化
・現基本的な評価から本格的なアノテーションへの展開
・塩基配列の1次データベースを加工した2次データベースの開発への取組み
・塩基配列データ以外の新しい型のデータのアーカイブの検討
・医学生物学データの統合利用と解析のツールの展開

などが挙げられる。

④　アウトソーシングへの対応

DDBJ は広く一般に公開されているので、利用者が利用の秘匿性を必要とする場合は、専用 ISDN回線、ＳＳＬ、利用者認証などの対応が可能である。

(2)　東京大学医科学研究所　HGC
［Human Genome Center；ヒトゲノム解析センター］

(ⅰ)　現状と課題

医科学研究所のデータベースには、組織内部及び世界の公的データベースのデータが格納され、外部利用向けに公開されている。データ量の増加によるデータベースの容量の不足、サービス・プログラムが遅くなるなどの問題が発生し、記録ディスクの増設、ネットワーク関連の増強、信頼性の向上を行っている。将来的に、巨大なセンターとそこにアクセスするための強力なネットワークが必要である。

(ⅱ)　アウトソーシングへの対応

アウトソーシングの最大の問題はセキュリティの保護である。ネットワーク監視システムの導入、アクセス制御などで対応している。

(ⅲ)　データベースのトピックス

①　使いやすく、信頼性のあるデータベース

使いやすく、信頼性のあるデータベースとするには、検索に対するレスポンスが速く、最新の情報を提供すること、新しい種類のデータ、生体分子間の相互作用、発現の情報などについて、配列や立体構造データとの関係を作っていくこと、データ・クォリティの面では、データの質・一貫性などをデータベース・サービス側が保証するデータベースを提供することが必要である。

②　ＤＮＡ検索システムの開発状況

ＨＧＣでは、日立中研グループによるＢＡＣＥというWWW検索エンジン技術と連想情報検索技術を組み合わせたシステムの開発をサポートしている。

③　ＤＮＡデータベースの将来像

近未来的には、様々なデータベースを統合的に扱う、例えば比較ゲノム研究を効率的に行えるデータベースが必要となる。分子間相互作用等に関する知識を集めたデータベースを使い、生命現象をシミュレートする研究が盛んになると予想される。

(3)　大阪大学蛋白質研究所

(ⅰ)　蛋白質立体構造データベースＰＤＢ(*1)（Protein Data Bank）の現状と課題

ＰＤＢデータベースはインターネットを通じ全世界に無料で公開している。

①　運営面の課題

ＰＤＢ運営に必要な人員、計算機、ソフトウエアなどの費用が十分ではない。データベースの改良や、ＰＤＢの周辺分野で付加価値を付ける活動が必要とされている。

②　機能面の課題

単に一次データを蓄積するだけでなく、他のデータベースとのリンケージを図るなどの構造化や、キーワード検索以外に他の蛋白との類似性、活性部位の推定などのアルゴリズムの提供等、サービスの高度化を図ることが必要である。また、教育の分野では、より簡便な検索システムが望まれている。近い将来、構造ゲノム・プロジェクトの進展によりＰＤＢのエントリー増加率の増大に備え、エントリーの編集ソフトウエアの効率化、高信頼化が必要である。

(ⅱ)　アウトソーシングへの対応

現在のデータベース保守体制は、アウトソーシングを引き受けられる状況にはなく、大幅な人員増と情報処理の専門家が必要である。また、ネットワークに関する危機管理体制の実施、専用回線の敷設などは、大学では実現が困難であるため、アウトソーシングを引き受けるには大幅な変更が必須であり、課題は非常に大きなものである。

(*1)　大阪大学蛋白質研究所のデータベース（http://pdb.protein.osaka-u.ac.jp/pdb/）は、米国ラトガース大学／米国カリフォルニア大学サン・ディエゴ校のグループがＲＣＳＢ（Research Collaboratory for Structural Bioinformatics）というコンソーシアムで協力して運営しているデータベースのアジア・オセアニア地区での公式アーカイブである。

3 民間データベース

(1) 藤沢薬品工業株式会社

(ⅰ) 社内に遺伝子データベース・システムを必要とする理由

一般公開前に開示される配列情報や共同研究相手又は契約相手などからの配列情報は、漏洩を避けるためにインターネット上で解析することが禁止される場合があるからである。

(ⅱ) 課題と対策

巨大なデータベース・解析システムを社内に整備するには、コストや要員面からも限界があり、アウトソース先としてデータ・センターを利用するなどの検討も始めている。

(2) その他一般的民間データベース

(ⅰ) インハウス・データベースにする理由

民間企業がインハウスでデータベースを構築する理由は、民間企業が公共データベースを検索に利用する際に、インターネットによる社外の公共データベース・サーバからの研究情報の漏洩によるリスクが社内にデータベースを構築するコストを上回ると判断した場合である。しかし、すべてのバイオテクノロジー関連企業がインハウス・データベースを構築しているわけではない。

(ⅱ) 課題と対策

現状ではインターネット上にセキュリティ技術を施した公共のサイトはない。研究情報の漏洩のない状況が確保されるという保証はない。幾つかの民間企業がバイオテクノロジー・データベース・サービスのアウトソーシング・ビジネスに名乗りを上げている。アウトソーシング・サービス提供のコスト・ダウンが図られれば、普及する可能性が高い。

4 諸外国特許庁におけるDNA配列データベースの状況

(1) USPTO (United States Patent Office：米国特許商標庁)

(ⅰ) データベース

USPTOは、提出に関係する情報だけを記録するCRF ISSUED、CRF PENDING及びCRF PUBLISHEDの3種類のデータベースを構築している。他にUSPTO内には、PIR、GENSEQ、SWISS－PROT及びTREMBLのデータを蓄積したアミノ酸配列データベースと、GENBANK、GENBANK－NEW、GENSEQとEMBLのデータを蓄積した塩基配列データベースがある。また化学構造や文献検索のために、CAS、Crystal、Orbit、LEXIS、NEXISなど商用データベースは、電話回線を介してUSPTOと接続可能である。

(ⅱ) 配列検索と解析

USPTOは、Smith & Watermanの利用からBLAST2の採用を検討中である。またSmith＆Watermanのほかに、GCG Word Searchと多重整列プログラムGen Align (GCGに含まれている) を利用している。長大あるいは多数の配列に必要な手動処理の自動化を検討中である。

(ⅲ) ハードウエアとスタッフ

ホモロジー検索には、専用ボードCompugen Bio XL/P4を利用している。

専任サーチャ12名に検索を依頼している。自ら検索する審査官は5名 (USPTO審査官総数約1,600名、内バイオテクノロジー担当500名) である。

(ⅳ) 将来展望

アウトソーシングの検討のために3社（CELERA、COMPUGENおよびBisant [MITのベンチャー]）を試用中である。セキュリティはSSL(*2)またはVPN(*3)を利用し、利用者認証とファイア・ウォールを利用している。また専用のT1回線の利用を検討中である。

ただし、出願された特許のデータベースはUSPTO内で構築維持の方向である。配列検索の管理のために必要なバックアップ・システムは、将来もUSPTO内で維持する方向であり、このため短期的にはハードウェアを増強する予定である。長期的にはMPSRCHソフトウエアに経費節減も含めて期待している。

(2) EPO(European Patent Office：欧州特許庁)

(ⅰ) システム要求

公共データベース、特許データベース及び第三者の利用が可能な「会員制」データベース (例：CELERA, Incyteなど) をすべて網羅した検索システムが必要であると認識している。

(ⅱ) 審査システム

配列データは専用のデータベースに格納している。EPO内で構築したSTRANDシステムからEMBL (European Molecular Biology Laboratory) の核酸塩基配列データベースを利用したEBI－EPOシステムに移行している。EMBLにはGenSeqが設置され、EPO専用のインターフェース開発を委託している。また専用回線で接続しており、相同性検索の結果からPubMedなどを参照する場合は、インターネット接続に切り替わることで、シームレスな操作を可能としている。

実際のEBI－EPOシステムの利用は、およそ100人程度の審査官で年間2万回程度である。

機能の向上は、相同性検索結果の保存 最低2年間、理

(*2) Secure Socket Layer
(*3) Virtual Private Network

想的には25年間）を目標とし、また相同性検索プログラムの評価やパラメータの設定知識の必要性を認識している。

その他に審査方法の共通化（例：パラメータの設定）と新技術対応のためのトレーニングを実施している。ＥＢＩへの委託も検討中である。

産業界及び学界との関係では、主要な特許出願人と年1回の会合を開催、またワークショップを開催している。ＥＰＯ審査官と企業の出願人が同一のデータベースにアクセス可能としている。

(iv)　将来の課題

将来のＥＰＯ内システムは、ホモロジー検索の機能だけではなく、文献など関連するあらゆる情報を統合的に参照できるシステムを目指している。また、ＥＰＯへの出願人が利用可能なバリデーション・システムの開発を計画中である。

II　特許庁内ＤＮＡ配列データベースのアウトソーシング

次世代ＤＮＡ検索システムの検討において、緊急課題である増大する配列データの蓄積の在り方について、三つの対策が提案されている。

案1　ハードウエア（ハードディスク）を増強し、特許庁内蓄積を継続する。

案2　データの圧縮によりハードウエアを最大限有効活用し、特許庁内蓄積を継続する。

案3　特許庁内蓄積を最小限にとどめ、その大部分をアウトソーシングする。

ここでは準備段階として案3の実現可能性の検討を行ったが、次世代ＤＮＡ検索システムの構築に案1～案3のいずれを採用すべきかについて検討したものではない。今後、これらは慎重に検討されるべきものである。

1　アウトソーシングの必要性

蓄積能力・検索能力両面で特許庁ＤＮＡ検索システムの増強の負担は大きく、一方、出願公開後のＤＮＡ配列データは公的データベースと特許庁内ＤＮＡ検索システムに重複して蓄積される。その対応策としてデータベースのアウトソーシングがある。

2　アウトソーシングに対する課題

(1)　出願公開の課題

特許出願は、出願から原則として18か月間は公開前情報として扱われる。公開前情報は、先行技術文献調査の対象であり、特にDNA配列データ自身が発明の本質を表すため、秘密情報として厳重に取り扱う必要がある。

(2)　アウトソーシングにおける公開前情報のデータ蓄積の課題

①　データを蓄積するサーバの物理的・ネットワーク的な保全、②　作業者に対する守秘義務、③　秘密漏洩・データ障害等に対するペナルティ、④　公開時に公開前情報から公開情報への移行作業、⑤　検索能力の保証、⑥　検索時の秘密保持、⑦　公共性の高い、特定の者の影響を受けない組織、などが挙げられる。

(3)　通信のセキュリティ

ホモロジー検索は、ＤＮＡ配列情報を含む検索式を用いて、発明の本質そのものが通信回線を流れるため、公開前審査の場合は、その配列データのセキュリティ対策は厳重とする必要がある。

(4)　検索のセキュリティ

ＤＮＡ配列検索では、検索式そのものが発明の本質を表すため、アウトソーシング先で検索式を保存されてはならない。これは契約等によってアウトソーシング先においてログ保存を行わないことを保証する以外はない。

(5)　その他

ＤＮＡ配列データ量の増加対策としてアウトソーシング以外にデータ圧縮技術がある。データ圧縮技術とアウトソーシングを組み合わせることで、より効率良くデータ蓄積を行うだけでなく、圧縮データによる通信、圧縮データに対する検索など、検索効率の向上という効果も期待される。

3　具体的アウトソーシング方法

(1)　検討要素

アウトソーシングの可能性の検討要素として、

①　可能な限りデータ量の大きいもの、②　検索時の利便性がある、③　検索応答性に支障がない、④　データ管理が確実である、⑤　公開された時点で公開前データベースから公開後データベースへデータの移動できる、⑥　検索先のデータが二箇所の場合に検索式を二箇所に送るとともに検索結果を合わせることができる、⑦　データ形式の統一できる、⑧　セキュリティ担保の容易性・確実性である、などが考慮される。

(2)　アウトソーシング方法案

(ⅰ)　策定基本方針

①　現状と同等の機能を維持する。

②　現状と同様のセキュリティを確保する。

③　特許庁内のデータ保有量をなるべく削減する。

(ⅱ)　データベース分類（表1）

表1

データベース分類		データベース内容
公開前データベース		公開前特許の情報
公開後データベース	a	配列情報
	b	キーワード情報
	c	フラット・ファイル (全ての情報を含む)

注：「データベース」は、GenBank、EMBL、DDBJ、SWISS−PROT、PIRを含む。

また公開前情報は、セキュリティ保持の観点からすべて特許庁内保持を前提としている。

(iii) セキュリティ対策

① ネットワーク・セキュリティ…専用回線と暗号化通信を使用する。

② サーバ・セキュリティ

(a) 作業用クエリー情報、検索ログ等の履歴情報を保存せず必ず削除する。

(b) サーバの保護…サーバ設置場所への入室制限。また入退室の際の認証。外部からのサーバへのアクセスの禁止。

(c) 不正アクセス監視機能…インターネットを使用する場合は、サーバに対する不正アクセスを監視する機能が必要である。

③ ユーザ認証…アウトソース先のサーバにアクセスする場合に必須。SSLによるクライアント認証（特定の公開鍵認証書がインストールされたクライアントだけにアクセスを許可する方式）を行う。

(iv) データベースの配置案

アウトソース先に配置する可能性のあるデータは、公開後データベースの中の配列情報a、キーワード情報b、及びフラット・ファイルcである。

① データベース配置案1

公開後データベースa, b, cをすべてアウトソースする。

② データベース配置案2

公開後データベースのうちキーワード情報bとフラット・ファイルcをアウトソースする。

→配列検索自体は特許庁内で行うため、配列検索クエリー式は外部に出ない、検索ログも外部に残らないという特徴があるが、キーワード検索クエリーについてはアウトソース先に出る。

③ データベース配置案3

公開後データベースのうちフラット・ファイルcのみをアウトソースする。

→この配置案は配列検索、キーワード検索自体は特許庁内で行うため、検索クエリーが外部に出ない、検索ログも外部に残らない。

(3) アウトソーシングの課題の検討

アウトソーシングに係る各課題について検討を行った結果を以下に示す。

① 出願公開・データ蓄積の課題

公開前情報の取扱者の守秘義務のために高レベルなセキュリティの導入の負担が大きい。公開時の公開前情報から公開情報への移行作業は、各データベース共に指定日まで非公開とするなどの対応により、大きな問題ではない。公開前情報が特許庁内に、公開情報が特許庁外に、分割された場合に、それぞれの検索をスムーズに行えるシステムが要求される。

② 通信のセキュリティ

専用回線と暗号化通信を併用する。

③ 検索のセキュリティ

アウトソーシング先での検索式のログ保存を行わないことを契約等で保証する。

④ その他

データ圧縮技術の適用は、特許庁内システムに残されるデータのデータ量、データ内容を考慮し、検討されるものである。

⑤ 各データ種別のセキュリティ・レベルについて

・フラット・データ

公開前データの蓄積には高度なセキュリティが必要である。公開前のフラット・データは、特許庁内の蓄積が望ましく、公開後のフラット・データはアウトソーシングし、相当量のデータが削減できる。

・インデックス・データ

発明のキーとなる情報であり、公開前データはセキュリティが必要であるが、公開後データはフラット・データと同様のセキュリティ・レベルで良く、データ量も大きいことから、公開後データはアウトソーシングが望ましい。

・配列データ

公開前配列データは、最大級のセキュリティが求められるため、特許庁内蓄積である。公開前検索で公開前配列データの検索式をアウトソーシング先に通信することは望ましくない。配列データを公開前後で分割して蓄積することは、新たなシステム構築など負荷が大きい。

以上から配列データはすべて特許庁内蓄積とし、配列検索は特許庁内でのみ行うことで、信頼性・検索性を向上できる。よって、データベース配置案2「公開後データベースのうちキーワード情報bとフラット・ファイルcをアウトソース」が、相当量のデータを削減でき、データ量増大対策として十分に有効であり、かつセキュリティも十分に担保されていると思われる。

4　アウトソーシングに関する意見と要望

(1)　セキュリティについて

(ⅰ)　システムに従事する内部の「人間」を介しての情報漏洩の可能性があり、アウトソーシング先の厳重な情報管理が望まれる。確実性を期すには公開前データはアウトソースから除外することも検討してもらいたい。

(ⅱ)　特許庁内のシステムの大幅な経費削減も期待されるメリットがある。また、審査官に最新のユーザ環境やカスタマイズが提供され、精度が高く迅速な審査につながり、特許出願人の利益にもなると考えられる。アウトソースは十分な保安対策を講じるならば積極的に行う価値があると評価できる。

(ⅲ)　セキュリティ確保について

①　アウトソーシング先と特許庁間のデータ転送時のセキュリティ、②アウトソーシング先のデータ・セキュリティ、の二つがあり、ある条件の下で一定のセキュリティを確保する技術は現時点で十分である。それらの技術をバイオ関連で適用しているのは、現時点では**ＪＢＩＣ**（社団法人バイオ産業情報化コンソーシアム）のシステム以外にない。**ＪＢＩＣ**システムのセキュリティの考え方は、

(a)　インターネット上はＳＳＬ暗号化により保護する、

(b)　サーバ内で実配列データが必要な場所はすべてプロセス内メモリ上で処理する（プロセス内メモリとは、必要時にダイナミックに確保するメモリで他からはアクセスは不可能な性質を持つ）、

(c)　サーバ内でシーケンスを含む一時ファイルが必要な場合は暗号化する、

である。

(2)　アウトソーシング先について

アウトソーシング先としては公的機関が望ましい。競合企業あるいは競合企業と関連する企業がアウトソーシング先となった場合には、公的機関よりも情報漏洩の可能性が懸念される。

(3)　その他

①　アウトソーシングを受ける側は、ある一定のセキュリティを保証するが、確実に安全であるとは言い切れないため、かなりの免責事項を契約に盛り込むことになると思われる。任せる側はこれらの免責事項が特許出願人にとって問題にならないか十分納得がいく説明が求められるであろう。

②　アウトソーシングの検討には、特許庁内システムを維持、増強する場合の負担と、アウトソーシングを行った場合の負担について、人、ハードウェア、予算などの各側面から比較検討を行う必要がある。

Ⅲ　配列データの圧縮

1　圧縮とデータ種別

現在利用可能な技術で、検索の高速化を前提としたデータ圧縮を検討した。

(1)　配列データ（無効）

文字の種類をａ，ｔ，ｃ，ｇの4種類に限定すると、1文字を2ビットで表現できる。これ以上に大幅に圧縮率を上げるのは難しい。また、「ゲノムの有限性に基づく圧縮法」は、配列データが出揃った場合には有効な手法となるが、現時点での有効性は不明確である。独自のファイル形式を採用している複数の既存の検索プログラムを使用する必要のある場合は、現状では配列データに独自の圧縮を行うメリットはほとんどない。

(2)　インデックス・データ（無効）

インデックス・データは、フラット・データＴの中からパターン文字列Ｐの出現位置を探す際の索引構造であり、検索の大幅な高速化が達成できる。目的の検索法に合わせて索引構造を選択し効率よく実装する必要がある。検索目的に合わせて選択され、実装されたプログラムによって構築されたものの場合は、そのインデックス・データを外部から別の方法で圧縮しようという試みは無意味であり、必要とする検索の仕様を決め、それに適した索引構造を選んで実装するべきである。

(3)　フラット・データ（有効）

ファイルを圧縮して保持しておき、結果の表示などの要求に応じて展開するという方法は、ディスク容量の節約という観点からは有効であると考える。展開が高速な圧縮方法を用いれば、十分に実用的なパフォーマンスが得られるであろう。ただし、部分的な展開を要する場合には、それに適した圧縮法を採用する必要がある。適切な圧縮によってファイル・サイズが小さくなれば、検索速度が向上することが期待できる。

(4)　まとめ

圧縮と検索を融合させてとらえようとする研究動向は比較的最近のものであり、状況は急速に変化している。したがって、現時点では有効性が認められないと判断した配列データの圧縮等についても、今後も引き続き慎重な検討を重ねていく必要があろう。

2　実機検証

ゲノム塩基配列の解読後に追加される塩基配列データのほとんどは重複データであり、その部分は高度に圧縮できると予想される。本検証で、ゲノムサイズを越える塩基配列データの圧縮効率がどう変化するか調べたところ、ゲノム解読後の追加データは20分の1（塩基あたり0.4ビット）以下のサイズに圧縮できるという傾向がみられた。

(1)　塩基配列データの圧縮

(ⅰ)　従来使用されている汎用圧縮方法

①　compress：LZW (*4)辞書ベース圧縮手法が忠実に実装されている。

②　gzip：LZ77 (*5)に基づき、一般に①よりもよい圧縮率を実現できる。

③　gzip -9 'gzipに-9オプションをつけたことを意味し、最良の圧縮を行うモードである。

(ⅱ)　従来使用されている塩基配列専用の圧縮方法

①　参照の利用

既に登場した塩基配列と同じ塩基配列部分が後で見つかった場合、前者への参照関係を表現するデータで後者の配列を表現する方法である。ここでは、参照先として使える配列データの全体を「辞書」とする。

(a)　一致した部分の再利用

完全一致した部分のみを参照する方法（complete match)、おおむね一致した部分であれば、それに相違点情報を付加して参照する方法（approximate match)がある。

(b)　相補的部分の再利用（Reverse Complement)

例えば、「AATTGCGC」と「GCGCAATT」は互いに相補配列の関係にあり、一方の配列が既に辞書内にある場合は、それを相補配列として参照して、他方の配列を表する。

②　符号化の利用

出現頻度の高いA，T，G，Cを短いビット列で表し、Nなどまれに使用される記号は長いビット列で表す。Huffman符号化 (*6)は有名である。また、符号化方法を固定せず、塩基配列のコンテキストに応じて次の出現文字を予測し、符号を最適化する「Context Tree Weighting」は塩基配列の圧縮に有効である。

③　formatdbは、ＢＬＡＳＴに付属のプログラムであり、直接サーチ可能な圧縮データベースを生成する。

④　現在の最良圧縮法でも1塩基当たり、平均1.74ビット程度。

「Context Tree Weightingに基づく符号化」と参照の利用」を組み合わせた方法が最も高い圧縮率で塩基配列を圧縮できることが報告されている (*7)。圧縮前データを1バイト／塩基とすると、圧縮後のサイズ比は圧縮前の21～22%程度にまで削減される。

(ⅲ)　本検証作業において、塩基配列データ圧縮方法に求めた条件

①　圧縮後もホモロジー・サーチ可能なデータ構造に近いこと。

②　配列データの増加に伴い、圧縮能力が上昇すること。

③　アノテーション付加による配列数増加要因に対しても有効であること。

ORIGIN（公開塩基配列のデータ・レコードにおいては、1レコード内に含まれている全体配列が記載されている項目）内に含まれる部分配列を抜き出して、独立配列としてデータベースに含める必要があり得ることをも考慮に入れる。

④　公知塩基配列を漏れなく保全してもデータ爆発を起こしにくいこと。

(ⅳ)　本検証作業で加えた改良点

生物学的な特性であるゲノムの有限性を圧縮原理とするように改良を加える。

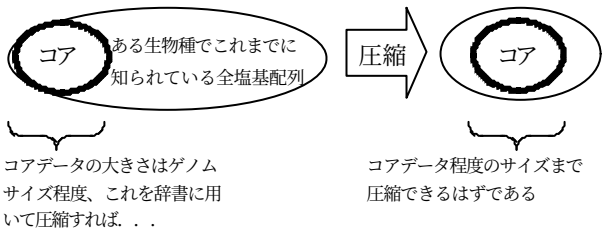
①　ゲノムの有限性（ＤＮＡ配列に特有の性質から）

人工的にランダムな塩基配列を作り出す場合を除き、通常取り扱われる塩基配列は、ゲノム上に存在する塩基配列の一部をコピーした断片である。シークエンシングの実行に比例して、塩基配列の総データ量は増加するが、得られた配列データ間では互いに重複する部分が多くなり、ユニークな断片の総量は頭打ちになってくる。これはゲノムの長さが有限であり、それ以上のデータ量を採取しても、部分的に重複したデータが増すだけであるからである。近年、この原理を応用したショットガン法により、ヒトゲノム全体の解読が成功している。この原理をデータ圧縮法に応用した。

②　辞書のサイズをゲノムワイドに広げること

ゲノムの有限性によれば、その生物から得られた塩基配列は、ゲノムに存在する塩基配列の一部をコピーした断片の情報であり、必要最小限の情報はゲノムサイズ程度である。これを辞書として利用すれば、塩基配列のほとんどの部分は辞書の参照により表現できるため、高度の圧縮率が期待できる（図1参照）。本圧縮方法は、ゲノムサイズを基準にして、少なくともそれ以上に辞書のサイズを広げる点に特徴がある。

図1



(*4)　Welch, T, " A Technique for High-Performance Data Compression.",IEEEComputer,vol.17,No.6,pp.8-19 Jun. 1984.

(*5)　Ziv,J.andLempel,A."Auniversalalgorithm for sequential data compression", IEEE Transactions on InformationTheory, Vol.23,No.3, pp.337-343,May 1977.

(*6)　Huffman,D.A., "Amethodforthe construction of minimum-redundancy codes", Proc. of theIRE,Vol.40, No.9, pp.1098-1101, Sep. 1952.

(*7)　Matsumoto,T., Sadakane,K.and Imai,H., "BiologicalSequenceCompression Algorithums", Genome Informatics, 11:43-52, Dec.2000.

③　「formatdbで作られるサーチ専用ファイル」のみを対象にした圧縮法

本検証で開発した圧縮プログラムは、formatdbにより生成されるＢＬＡＳＴサーチ用の塩基配列データを、さらに圧縮した。

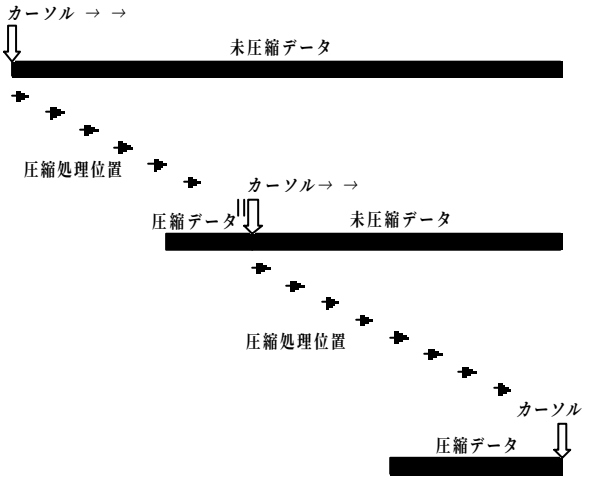
塩基配列データ（1塩基＝1バイト）
↓formatdb
ＢＬＡＳＴ用データ（1塩基＝2ビット）
↓本研究の圧縮方法
高度圧縮データ
↓解凍とＢＬＡＳＴサーチの同時実行
ＢＬＡＳＴ用データ

④　ゲノムワイドな辞書の参照による圧縮効果を検証

「ゲノムの有限性」を圧縮原理による圧縮効果を的確に検証するために、「ゲノムワイドな辞書の参照」という方法のみを実装したプログラムを作成し、その圧縮効果を測定した。プログラムの概略について以下に説明する。

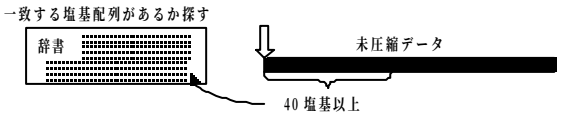
(a)　図2に示すように、圧縮中の処理位置をカーソルとよぶ。本研究の圧縮法では、ファイルの先頭から2ビット（1塩基）単位で圧縮していくため、カーソルは2ビットずつ進む。

図2



(b)　図3に示すように、カーソルから40塩基以上の塩基配列断片と一致するものが辞書内にあるかどうか探索（相補配列も考慮）。

図3

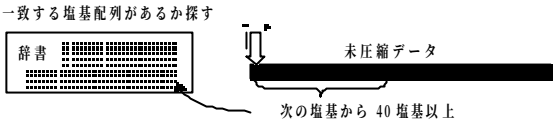


(c)　図4に示すように、見つかった場合は参照情報を該塩基配列データと置き換える。カーソルを次の塩基に移動させて、また(b)から繰り返す。

図4

(d)　図5に示すように、見つからなかった場合は、1塩基分だけカーソルを次にスキップさせて、また(b)から繰り返す。スキップされた塩基は圧縮後データにそのまま付け加えられる。

図5

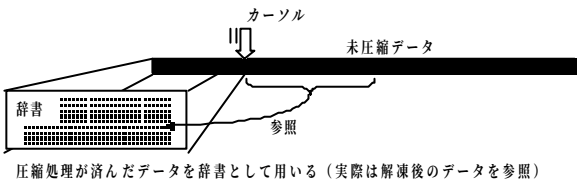


(e)　辞書を固定するかどうかで二とおりの方法を試みた。

方法1：既に圧縮済みのすべての塩基配列データを動的辞書として用いる方法

図6に示すように、複数塩基配列を連続して圧縮していく際に、既に圧縮処理の済んだ塩基データを動的な辞書として利用する。長所としては、最適な圧縮効果を得ることができる点にある。一方、短所としては、辞書のサイズが次第に増大していくため、それに伴い圧縮速度は低下してしまう。

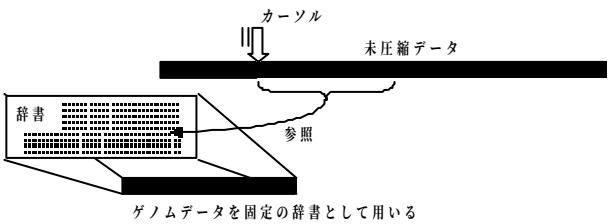
図6



方法2：ゲノムデータを固定の辞書として用いる方法

図7に示すように、将来もっとも基本的に利用されるであろうゲノムデータを固定の辞書として活用する方法。辞書のサイズが固定なので、圧縮速度は一定であるが、動的な辞書を用いる方法1のように圧縮効果を最大限に引き出すことができない。

図7



(v) 実データを用いた検証の方法

E. coli（大腸菌）とS. cerevisiae（酵母）の全公開塩基配列データに対し、「固定辞書（ゲノム塩基配列）を用いて圧縮する方法」と、「動的辞書（既に圧縮済みの塩基配列）を用いて圧縮する方法」を適用し、それぞれの条件での「圧縮能力」と「圧縮速度」（ただし、ここでは省略）を測定した。

(vi) 圧縮処理の高速化

圧縮する塩基配列を幾つかのセクションに分割し、セクションごとにCPUを割り当て、圧縮処理を並列的に行うことも試みた。各セクションを圧縮した後、それらを連結させて、圧縮処理を完了させた。並列化の問題点は、辞書との一致率が低いセクションは圧縮速度が低下し、これが律速となる。

(vii) 圧縮法の検証結果

① E. coli公開塩基配列データの圧縮

DDBJ リリース39（1999年10月）のファイル"ddbjbct.seq"からE. coli由来の塩基配列をすべて抜き出し、この圧縮を試みた。汎用的な各種圧縮方法と本検証で開発した圧縮方法を適用した場合の圧縮後データサイズを表2に示す。

表2 圧縮法の検証結果

圧縮方法	圧縮後サイズ	サイズ比率	塩基当たりサイズ
未圧縮	15,668kb	100.0%	8.00 bits
Compress	4,234kb	27.0%	2.16 bits
Gzip	4,000kb	25.5%	2.04 bits
gzip -9	3,865kb	24.7%	1.97 bits
Formatdb	3,921kb	25.0%	2.00 bits
formatdb後gzip-9	3,647kb	23.3%	1.86 bits
本研究の方法1	1,501kb	9.6%	0.77 bits
本研究の方法2	527kb	3.4%	0.27 bits

圧縮前のE. coli全公開塩基配列データは15,668kbで、E. coliゲノムのコンプリート・データ(Escherichia coli K-12 MG1655) 4,663kbが含まれていた。

(a) 動的辞書を用いる圧縮方法1

E.coli全公開塩基配列データの圧縮後サイズは1,501kbで、圧縮前の9.6%にまで小さくすることができた。1塩基当たりのサイズは0.77ビットであった。

圧縮データの内訳は、ゲノムのコンプリート・データに限って計算したサイズ比は24.5%で、1塩基当たりのサイズは1.96ビットであった。この値はformatdbのそれと同程度であり、圧縮効果は余り高くない。ゲノムのコンプリート・データ以外の部分は360kbにまで圧縮された。この部分のサイズ比は2.3%で、1塩基当たりのサイズは0.18ビットと、非常に高い圧縮効果が得られた。更に塩基配列を追加したとしても、追加データに対してはこれと同程度の圧縮効果が期待できる。

(b) ゲノムを固定辞書に用いる圧縮方法2

E.coli全公開塩基配列データの圧縮後サイズは527kbで、圧縮前の3.4%にまで小さくすることができた。1塩基当たりのサイズは0.27ビットであった。ゲノム配列との類似性に基づいて、圧縮したい塩基配列を事前に分類しておくとよい。

② S. cerevisiae公開塩基配列データの圧縮

スタンフォード大学のSGD(Saccharomyces Genome Database)[*8]で公開されている酵母（Saccharomyces cerevisiae）の全塩基配列、yeastGenBank（2001年2月14日現在）を用いて圧縮の効果を試みた。汎用的な各種圧縮方法と本検証で開発した圧縮方法を適用した場合の圧縮後データサイズを表3に示す。

表3 圧縮法の検証結果

圧縮方法	圧縮後サイズ	サイズ比率	塩基当たりサイズ
未圧縮	27,602kb	100.0%	8.00 bits
compress	7,369kb	26.7%	2.14 bits
gzip	7,286kb	26.4%	2.11 bits
gzip -9	6,983kb	25.3%	2.02 bits
formatdb	7,011kb	25.4%	2.03 bits
formatdb後gzip-9	6,481kb	23.5%	1.88 bits
本研究の方法1	3,537kb	12.8%	1.02 bits
本研究の方法2	1,043kb	3.8%	0.30 bits

(a) 動的辞書を用いる圧縮方法1

酵母の公開全塩基配列データの圧縮後サイズは3,537kbで、圧縮前の12.8%にまで圧縮された。1塩基当たりのサイズは1.02ビットであった。

圧縮データの内訳は、ゲノムのコンプリート・データに限って計算したサイズ比は23.9%で、1塩基当たり1.92ビットである。つまり、酵母ゲノム自体は余り重複がなく、圧縮されにくいデータであるといえる。ゲノムのコンプリート・データ以外の部分は679kbにまで圧縮された。この部分のサイズ比は4.3%で、1塩基当たりのサイズは0.35ビットと、非常に高い圧縮効果が得られた。更に塩基配列を追加したとしても、追加データに対してはこれと同程度の圧縮効果が期待できる。

(b) ゲノムを固定辞書に用いる圧縮方法2

S. cerevisiaeゲノムのコンプリート・データを固定の辞書にした場合の圧縮後サイズは1043kbであり、圧縮前の3.8%に

(*8) SGD: http://genome-www.stanford.edu/Saccharomyces/

まで圧縮された。1塩基当たりのサイズは0.30ビットであった。この圧縮率には辞書であるゲノムのデータ・サイズが含まれていない。

更なる塩基配列の解読が進み、S.cerevisiaeの塩基配列データが追加された場合、追加分のデータはゲノムを辞書とすることで上記の圧縮率で圧縮できると推定される。

(2) 検証結果の評価

本評価は今回の実機検証結果のみに基づき、実用化には更なる検討が必要である。

(ⅰ) 辞書のサイズがゲノム・サイズを越えると、圧縮効果が劇的に上昇する。

(ⅱ) 部分配列（全体配列から抜き出された配列）は重複配列であるため、全体配列とともに圧縮すると効果が大きい。

(ⅲ) 辞書がゲノムワイドになり圧縮に掛かる処理が非常に遅いが、圧縮領域を分割して計算機を並列化することにより高速化が可能である。

(ⅳ) ゲノム解読後に追加されるデータは、ゲノム辞書による圧縮で5%以下に圧縮可能であるという傾向がみられた。

(ⅴ) 辞書がゲノムワイドになり、メモリを要する（ヒトで1GB弱）。

(ⅵ) 生物ごとの分類がなされてないと、ゲノム辞書の圧縮効果が得にくい。

Ⅳ 次世代ＤＮＡ検索システム

1 システム的要求

次世代ＤＮＡ検索システムに望まれる機能の例を以下に示す。

① 検索結果表示

・配列比較時における配列間の相違点の表示

・領域ごとの一致度スコアを表示

・検索式である配列内の発明に本質的な部分をマークアップ可能とし、当該マークアップが常に反映される表示形態とする

② 検索結果集合を相同性順だけではなく、様々なキーでソート可能とする

③ 同じ出願について、拒絶理由通知後の補正に対する審査を行う場合など、前回のサーチの状況等、審査の経歴を保存／再現可能とすることで、重複したサーチを回避する

④ アウトソーシング・データの効率的検索

公開前後のデータを庁内外で分割して蓄積する場合などは、同時に庁内外に検索式を投げた場合、データ量の違いから応答にはかなりの時間的ズレが予想され、そのための対策が必要となる。また、どちらかを先に検索対象として指定する場合、システム的には、どのような順番で検索を行うことが望ましいかの検討が必要となる。

さらに、明らかに公開前検索が不必要な場合には、特許庁外データベースのみを指定して検索することになるが、公開前データの一部だけ（公開直前のものだけ）検索が必要な場合、すべての公開前データを検索することは検索効率の低下につながるものである。

現在のＤＮＡ検索システムでは、日付による制限ができないために、すべてのデータが検索対象となって、応答時間の長大化の一因となっている。例えば制限条件として、特定の日付以前のデータのみを検索対象とする機能が望まれる。

⑤ アウトソーシング・データの効率的検索

公開前データを特許庁に、公開後データをアウトソーシング先で蓄積する場合に、両方に同時に検索を行った際の応答時間のズレ対策や、どちらかを先に検索対象とする場合に、どちらを先に検索を行うことが良いか検討が必要である。

さらに、現在のＤＮＡ検索システムでは、日付による制限ができないために、すべてのデータが検索対象であり、検索応答時間の長大化の一因となっている。すべてのデータを検索する必要がない場合や特定の期間のみを検索するような場合に、特定の日付以前又は以後のデータのみを検索対象とする機能が必要である。

2 特許庁に求めること

(1) アウトソーシングの遺伝子配列解析システムの一般ユーザーへの開放

特許審査と同じ条件での遺伝子配列解析が可能となり、さらに実際の審査での評価方法（使用データベースや解析ツールとパラメータなど）の指針が示されれば、出願人が事前に出願の可否を判断できるため、無効な出願とその審査が減少することが期待できるので、アウトソーシングの遺伝子配列解析システムの一般ユーザーへの開放を要望する。

(2) データベース整備と公開

引用特許・文献、被引用特許、米国や欧州諸国など他国の特許、遺伝子・蛋白データベースとのリンクなどの整備、ＢＬＡＳＴやＰＦＡＭ（Protein families database of alignments and HMMs）などの配列解析結果や、様々なカテゴリー分類、出願者などの記載項目からの検索機能、分かりやすい「審査状況・米国や欧州諸国などの他国出願状況・有効期限」などを含む有用性の高い特許データベース・解析システムを整備・構築して、これらを一般に公開することを要望する。

更に高度なキーワード検索ができる付加価値の高いデータベースを民間で容易に作製できるように、特許配列データをまとめてダウンロード可能とすることを要望する。

(3) 特許配列情報の公共データベースへの公開

公開された特許配列の公共データベースへの迅速な収

載を要望する。また、特許出願の公開時期と配列の公開時期が一致し、公開前に出願を取り下げた場合には、配列自体が単独で公開されないようなシステムの構築が望まれる。

3　その他

①　配列の「寄託制度」をシステムに導入することが望ましい。具体的には、出願番号（出願前は何らかのID）と配列番号により特定される配列を特定機関に「寄託」し、配列表の提出を省略するというものである。寄託した本人は「寄託」配列を出願に引用でき、出願公開後は第三者も配列を引用できることとする。

世界的にこの制度を採用すれば、日本だけでなく、ＰＣＴ受理官庁や各国特許庁の事務負担が大幅に削減されることは確実であり、特許庁のみならず出願人の労力軽減という観点から魅力的な制度である。

②　ＰＣＴ出願の翻訳文の提出に際して配列データの提出を省略するシステム構築を望む。

Ⅴ　まとめ

1　ＤＮＡ配列データベースの現状と課題

まず、本調査研究では、特許庁内データベースを使用したＤＮＡ配列出願検索の現状を踏まえ、ＤＮＡ配列データベースの現状と課題、及び特許庁内ＤＮＡ配列検索システムについて議論を行った。

実際にＤＮＡ配列に関する特許出願の審査を行う審査官のニーズを満たすデータベースを構築するためには、ＤＮＡ配列データベースのデータ量の問題のみならず、データベースの在り方、質的問題（信頼性）等についての総合的な検討を行う必要があることが明らかである。その上で、データベースのアウトソーシングの利害得失について議論を行うことが不可欠であるものと思われる。しかしながら、本調査研究委員会においては、委員会の目的や時間的制約からデータベースの質的問題についてはほとんど議論がなされなかったため、データベースの質的問題は今後議論すべき課題である。

次に、公的データベースとしては、委員会委員が所属する国立遺伝学研究所、東京大学医科学研究所及び大阪大学蛋白質研究所において、実際に構築し、使用しているデータベースの現状について、詳細に紹介いただき、また今後の課題についても、技術的問題だけでなく、管理面の問題や質的問題等の多くの問題点に基づき、具体的かつ貴重な提言をいただいた。公的データベースといっても、取り扱うデータの種類が相違するだけでなく、そもそもデータベース構築の目的が相違するものであることから、現状や課題についても必ずしも共通するものでないが、データ量の増大や、データ更新・送信、アウトソーシングを行う際の問題点等の技術的な問題、データベースの管理問題という組織的な問題も存在していることや、データベースの質的問題にしても、長期的な安定稼働が求められていること、私有データも含めたデータの公開という機能を有すべきこと、データ形式・表記の標準化や異なるデータベースのデータ間の整合性が求められることなどの多くの課題が指摘され、データベースのネットワーク化を念頭に置きつつ、使いやすく、信頼性のあるデータベースを構築することの重要性が指摘された。これらのことは、すべて特許庁内のＤＮＡ配列検索システムにも共通する課題であると考えられることから、今後更なる具体的かつ詳細な検討が行われることが期待される。

さらに、利用者の立場から、民間データベースの現状と課題の多くは公的データベースと共通するものであるが、データの流出やシステムのセキュリティの問題等も存在するものの、データ量の増大や管理コストの観点から、アウトソーシング化も不可避であることが報告された。

また、ＵＳＰＴＯ及びＥＰＯにおけるＤＮＡデータベースの状況についての報告と議論がなされ、両特許庁と我が国の特許庁とは、日米欧三極特許庁会合でデータベースの共通化やデータの交換等のプロジェクトを推進しており、三極特許庁の保有しているデータは互いに重複するものが多く存在するものと考えられる。我が国の特許庁がＤＮＡデータベースの構築を行う際には、両特許庁の状況にも留意しつつ、重複するデータの取扱いやデータベースの形式・表記等の統一化に向けての議論を充分に行い、その上で、データベースのアウトソーシングの是非についても検討を行うべきである。

2　特許庁内ＤＮＡ配列データベースのアウトソーシング

データの蓄積対策及びデータベースの検索応答性向上の観点から、アウトソーシングの必要性について、アウトソーシングをすることの必要性及び有効性について基本的に疑問とする意見があった。これはアウトソーシングを行う上での基本的方針の欠落に対して疑問を呈するものであったことから、まず、アウトソーシングについての基本的方針を明確にすることが必要であろう。

実際に、アウトソーシングを行うのであれば、アウトソーシング自体についての評価に関する議論をより詳細かつ具体的に行った上で、アウトソーシングを行う際の具体的課題についての議論を行うべきであろう。

3　配列データの圧縮

配列データベース検索の高速化の観点から、データの圧

縮の評価及び実機検証の結果は、フラット・データのみがデータ圧縮が有効であることが確認された。また、実機検証の重複データの圧縮は、特定の条件の下である程度の効果が得られることが確認された。データ圧縮の分野の今後の研究の進展次第によっては、これらの評価が変更される可能性もあるが、特許庁におけるＤＮＡ配列データベースを構築する際にも参考となるものであろう。

4　次世代検索システム

次世代検索システムについては、検索結果表示やアウトソーシング・データの効率的検索等のシステム的要求、また、システムの一般ユーザーへの開放、データベース整備・公開、特許配列情報の公共データベースにおける公開等の特許庁に対する要望等を含め、特許庁の次世代検索システム全体のなかでの位置付け、すなわち、特許庁の情報政策全体におけるＤＮＡ配列データベースの位置付けを明確にした上で議論をすべきものである。

以上、時間的制約とデータベース分野の著しい技術進展等から、完全な結論を得るには至らなかったが、一定の方向性を得ることはできたと思われる。なお、本調査研究報告書紀要において採り上げた意見は、飽くまで個人的見解に基づくものであり、必ずしも委員会の総意ではない。

(担当：研究員　池上　敬)