

国際塩基配列データベース検索システム へのInterstage Shunsakuの適用

Introduction of Interstage Shunsaku to Search Engine of International Nucleotide Sequence Database

あらまし

国立遺伝学研究所生命情報・DDBJ研究センター様は、「国際塩基配列データベース」を構築し、このデータベースの検索システムをインターネットで広く研究者に公開している。近年のバイオ研究の進展により、国際塩基配列データベースは、増加の一途をたどっており、検索システムのレスポンス悪化、システムメンテナンスコストの増大などの問題を抱えていた。この問題を解決するため、国立遺伝学研究所生命情報・DDBJ研究センター様は、富士通が開発した高速XML型データベースエンジン“Interstage Shunsaku Data Manager”を国際塩基配列データベース検索システムの検索エンジンに適用し、検索条件によらず10秒以内の検索レスポンスという良好な結果を得ることができた。これは富士通が提供するIT基盤「TRIOLE」の既存IT資産を活用することによって、柔軟なシステムを実現した事例である。

本稿では、従来の問題点を挙げ、Interstage Shunsakuの適用ポイント、導入効果を紹介し、最後に今後の展望について述べる。

Abstract

The Center for Information Biology and DNA Data Bank of Japan at the National Institute of Genetics has constructed the International Nucleotide Sequence Database (INSD) and made its database search engine accessible to researchers on the Internet. Thanks to the recent progress in biotechnology, this database is continuing to grow. However, this growth has caused a reduction in the search engine's performance and an increase in the system maintenance cost and other costs. To solve these problems, we applied a high-speed XML type database engine called Interstage Shunsaku Data Manager that Fujitsu developed to the search engine of the INSD. As a result, we achieved a retrieval response of within 10 seconds regardless of the search conditions. By using the existing IT assets of Fujitsu's TRIOLE, we have realized a flexible system. This paper describes some common problems with database search engines, some examples of how to implement Interstage Shunsaku, the effects of its introduction, and its future use and development.



西宮直樹（にしのみや なおき）
ライフサイエンスシステム事業部
所属
現在、国立遺伝学研究所のスーパー
コンピュータシステム、DDBJ運用
業務のサポートに従事。

まえがき

国立遺伝学研究所生命情報・DDBJ研究センター様は、欧州のEMBL (European Molecular Biology Laboratory), 米国のGenBank (Genetic Sequence Data Bank) とともに、三大国際DNAデータバンクとして日本DNAデータバンク (DDBJ: DNA Data Bank of Japan) を運営しており⁽¹⁾「国際塩基配列データベース」の構築・提供を行っている。国際塩基配列データベースとは、全世界の研究者が実験によって決定したDNA (またはRNA) の塩基配列および生物学的意味などの情報をDDBJ, EMBL, GenBankが三者間で定めたデータ構築規範に沿って収集・編集し、電子データとして公開しているバイオデータベースである。

2003年4月のヒトゲノム解読完了宣言⁽²⁾に象徴されるように近年のバイオデータは非常に巨大化しており、バイオ分野の研究活動においてコンピュータの利用は必要不可欠となっている。この要望を受けDDBJでは、国際塩基配列データベースを電子データとして提供するだけでなく、研究者が欲しい情報を効率良く得られるようにデータベース検索システムをインターネット上で公開している。しかし、前述のようなバイオデータの急増によって、検索レスポンスの悪化、システムメンテナンスコストの増大などの問題を抱えていた。

本稿では、DDBJが提供しているデータベース検索システムに、富士通が開発した高速XML型データベースエンジン“Interstage Shunsaku Data Manager”(以下、Interstage Shunsaku)を全面適用し、問題を解決した事例を紹介する。これは富士通が提供するIT基盤「TRIOLE」の既存IT資産を活用することによって、柔軟なシステムを実現した事例と言える。以下、従来の検索システムが抱えていた問題、Interstage Shunsaku適用のポイント、導入効果、今後の展望について説明する。

国際塩基配列データベース

DDBJは、日本の国際的な生命科学研究の拠点として位置付けられており、全世界から様々な生物のゲノムプロジェクトのデータや特許庁の公開データなどが集積されている。国際塩基配列データベースは、データの最小単位である「エン트리」の集合と

して構成されており、2004年10月現在、3,800万件、422億塩基のエン트리 (データサイズ約100 Gバイト) が公開されている。登録件数は年率約140%という非常に速いスピードで増大している。

各エント리는、大きく分けて塩基配列の部分と、注釈の部分から成る (図-1)。塩基配列の部分は、登録するDNA塩基配列そのもの (a, t, g, cなどの塩基の並び) であり、注釈の部分は、登録者、関連文献、生物学的特徴、遺伝子の機能や特性などのDNA塩基配列に付随する情報が記載されている。このようにエント리의データ形式は、複数の意味を持ったフィールドから成る。とくに注釈部分の“FEATURES”は、複数のキーと対応する値が多重になっており、エン트리全体として複雑な多重階層のデータ構造を持つと言える。

データベース検索システムの従来の問題

DDBJでは、このようにデータ量が膨大、かつ、データ形式が複雑である国際塩基配列データベースへの簡易なアクセス方法として、データベース検索システムをインターネットで公開している。世界各国の研究者などから、月間約7万アクセスの利用が

LOCUS	AB000000	450 bp	mRNA	linear	HUM 08-JUL-2002
DEFINITION	Homo sapiens GAPD mRNA for glyceraldehyde-3-phosphate dehydrogenase, partial cds.				
ACCESSION	AB000000				
VERSION	AB000000.1				
KEYWORDS	.				
SOURCE	Homo sapiens				
ORGANISM	Homo sapiens				
	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.				
REFERENCE	1 (bases 1 to 450)				
AUTHORS	Mishima, H. and Shizuoka, T.				
TITLE	Direct Submission				
JOURNAL	Submitted (30-NOV-2000) to the DDBJ/EMBL/GenBank databases. Hanako Mishima, National Institute of Genetics, DNA Data Bank of Japan; Yata 1111, Mishima, Shizuoka 411-8540, Japan (E-mail: mishima@supernig.nig.ac.jp, Tel: 81-55-981-6853, Fax: 81-55-981-6849)				
COMMENT	Human cDNA sequencing project.				
FEATURES	Location/Qualifiers				
source	1..450				
	/chromosome="12"				
	/clone="GT200015"				
	/clone_lib="lambda gt11 human liver cDNA (GeneTech. No. 20)"				
	/map="12p13"				
	/mol_type="mRNA"				
	/organism="Homo sapiens"				
	/tissue_type="liver"				
CDS	86..450				
	/codon_start=1				
	/gene="GAPD"				
	/product="glyceraldehyde-3-phosphate dehydrogenase"				
	/protein_id="BAA12345.1"				
	/transl_table=1				
	/translation="MAKIKIGINGFRIGRLVARVALQSDDELVAINDPFIITDWT YMFKYDTHGQKHHEVVKVDSKTLFGKEVTVFCGRNPKEIPIGETSAEPVIEYTG VFTDKDAVAQLKGAKKV"				
BASE COUNT	102 a 119 c 131 g 98 t				
ORIGIN	1 ccacgcgcgc cggtcgcatc gcacttgtag ctctcgacc ccgcatctca tccctctct 61 cgccttagtc agatcgaat ccgaatgc gaagattag atcggtatca atgggttcg 121 gaggatcgcc agctcgtag ccagggtgc cctgcagac gacgactgc agctctgc 181 cgtcaacac cccttcata ccacgacta catgacata atgtcaagt atgacactg 241 gcacgcgcgc tgaagatc atgagttaa ggtgaagac tcaagacac ttctcttcg 301 tgaagagag gtcacgctg tgcgctgac gacacctaag gagatccat ggggtgagc 361 tagcgttag ttgtgttag agtaccctg ttgtttact gacaaggaca aggcgttcg 421 tcaacttaag ggtgtgtcta agaagctctg				

注釈

塩基配列

//

出典: DDBJホームページ⁽³⁾

図-1 エントリの例

Fig.1-Example of entry (source: DDBJ homepage⁽³⁾).

ある。

研究者は、検索システムを利用して、自分の研究に関連したデータを参照したり、ファイルとして取得したりする。検索対象のデータが大きいので検索条件を詳細に指定し、絞り込まないと余計なノイズデータもヒットしてしまう。そのため、検索システムでは、データ形式に沿ったフィールドごとに検索キーワードの指定が可能になっており、複数指定や論理演算子による結合ができるようになっている。

このようにDDBJでは、データベースの利便性向上のため、利用者の要望に応えた機能性の高い検索システムを提供しているが、近年のデータ量の急増に伴って、つぎのような問題に直面していた。

(1) 検索レスポンスの悪化

単純な「IDによる検索」や「単一キーワードの検索」などは、数秒で結果が得られる。しかし、特定の項目に絞ってキーワードを指定したり、複数の条件を組み合わせたり、検索条件を複雑にしていくなにつれて検索時間が増大してしまう。また、検索時間は指定した条件に依存し、数秒から数十分とばらつきが大きく利用者にとっては非常に不便である。

(2) メンテナンスコストの増大

データ量の増加に伴って、検索インデックスの作成時間も増大している。全データを入れ替える際には、インデックス作成に約3日かかっており、その分、マシンコストおよび作業コストが発生している。

(3) システム拡張が困難

今後のデータ増加は避けることができないため、システム拡張によって検索性能を向上させる必要がある。しかし、現状のシステムでは、システム拡張による性能向上があまり望めないことや、拡張による作業コストが大きくなることが予想されることから、今後の対応に不安を抱えている。

これらの問題を解決するために富士通が開発した、高速XML型データベースエンジンInterstage Shunsakuをシステム基盤に適用した大規模データベース検索システムの構築を検討した。

Interstage Shunsaku適用のポイント

Interstage Shunsakuの適用を検討するに当たり、以下の特長が問題解決に有用であると考えられた。

(1) XMLテキストデータの高速検索性能

DDBJのデータ形式は複雑（非整形、多重構造）

であるが、Interstage ShunsakuではXMLデータ形式を採用しているので、DDBJのデータ形式をXML形式へ適切に変換することにより、特定のフィールドに絞った検索などの高度な検索が実現可能と考えられる。また、独自の高速検索技術⁽⁴⁾を採用しているため、検索条件が複数であっても一定のスピードで検索でき、利用者に対して高速かつ一定の検索レスポンスを提供することができると考えられる。

(2) 検索インデックス不要

データを先頭から最後まで読み込んでいくシンプルな検索方式なので、インデックスを作成する必要がない。そのため、データ更新の際に、現状と比べて時間と手間がかからずシステムのメンテナンスコストの削減が期待できる。

(3) シンプルな拡張性

検索時間は、データ量とCPU数のみに依存するので、データ量の増加に対して、サーバ数をそれに合わせて増やすだけで、チューニングすることなしに検索性能を保証できる。また、分散サーバ環境に対応しているため、サーバ増設の作業も簡易である。

Interstage Shunsakuの導入効果

Interstage Shunsakuを検索エンジン部分に適用し、データベース検索システムの構築を行った（図-2）。今回構築したデータベース検索システムをDDBJの研究員に試験公開し評価を実施していただいた。その結果、現状の検索システムとの比較で、以下のような効果が認められた。

(1) 検索レスポンスの飛躍的な向上

3,800万件、100 Gバイトのデータ検索に対して、10秒以内の検索レスポンスが得られ、検索時間の

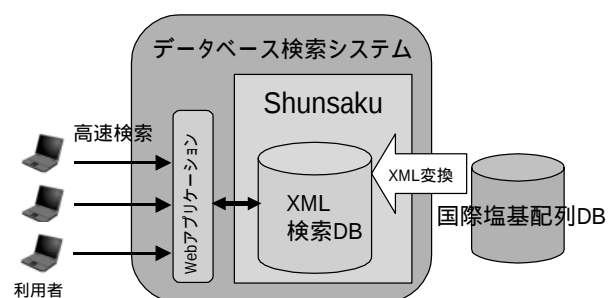


図-2 システム構成
Fig.2-System configuration.

短縮を実現することができた。また、従来であれば検索条件が複雑になった場合、分オダの時間がかかるケースがあったが、新システムでは検索条件の複雑さに依存せず一定の検索レスポンスを保証することができた。公開システムでは、単純な検索時間の面だけではなく安定した検索レスポンスの提供がより重要視されるため、「どんな検索条件でも10秒以内で結果が得られる」という信頼性は非常に価値が高い。

(2) 検索機能の向上

Interstage Shunsakuでは、XMLデータの全タグに対する検索が可能である。そのため、従来は検索対象のフィールドとして指定できなかった箇所に対してもXMLタグとして明示化することにより条件指定が可能になり、柔軟な検索機能を実現することができた。

(3) メンテナンスコストの削減

Interstage Shunsakuの導入に際して、リレーショナルデータベース(RDB)のようなDBスキーマの設計は必要なく、導入コストを抑えることができた。DBスキーマの役割は、XMLデータ側が担っているが、仮に、XML形式の仕様に変更が発生した場合でも、Interstage Shunsakuの設定変更は必要なくアプリケーションの変更だけで済む。そのため、今後のシステム変更で発生するメンテナンスコストを従来よりも抑えることができると考えられる。

また、インデックスを使用しないため、従来のインデックスの設計や作成をする必要がなくなった。データ更新の際にはデータ取込み処理が必要になったが、全データの入替えでも約2時間と速いため、従来(約3日)と比較してデータのメンテナンスコストを大きく削減することができた。

(4) システム拡張が柔軟

今回のシステム導入に当たって、データ量(3,800万件, 100 Gバイト)と検索時間(10秒以内)の関係からシステム規模を決定した。その結果、期待どおりの検索性能を得ることができた。事前に検索性能を見積もることができる点は非常に有用だと考える。例えば、RDBであればテーブルやインデックスの設計によって検索性能が影響されるため、考慮すべき対象が増え見積もりがより困難になると考えられる。一方、Interstage Shunsakuは、事前に検索性能を高精度で見積もることができ、かつ、分散

サーバ環境に対応しているため、今後のデータ量の増加に対応して柔軟に対応可能である。

今後の展望

バイオデータは、今回のDNA配列情報のほかに、文献情報、タンパク質情報(立体構造など)、遺伝子発現情報など多岐にわたる。今後は、これらのデータ形式にも対応した検索システムの構築を進めていく予定である。具体的には、それぞれのバイオデータについての詳細な検索を実行可能にする。さらには多様なバイオデータ間を串刺しで一括検索できるような統合バイオデータベースへの拡張を目指す。また、データの検索機能だけではなく、得られた結果を利用者の望む形式でダウンロードできるなどのデータ提供の観点での機能増強も考えていきたい。

国際塩基配列データベースだけではなく多様なバイオデータについて、研究者のポータルサイトとしてDDBJのデータベース検索システムを実現することができれば、DDBJを通して大きな付加価値を研究者に提供することが可能になると考える。現在、バイオ分野において、データ量の増大およびデータ形式の多様化が進んでおり、巨大データベースの検索システムおよび管理システムは重要性が増していくことが必定である。今後もInterstage Shunsakuをコア技術として、大規模データに対する複雑なデータ処理の観点から、バイオ研究開発において世界をリードすべくDDBJを支援していきたい。

む す び

大規模なバイオデータベース検索において、レスポンスの悪化、システムメンテナンスコストの増加、という問題に対して、検索エンジンにInterstage Shunsakuを適用することにより問題を解決できた。とくに、国際塩基配列データベースの検索で、「検索条件によらず10秒以内の検索レスポンスの実現」という大きな成果をあげることができた。バイオデータベースの検索システムは、全世界に多く存在するが、対象とするデータ量、検索性能の面から今回構築したシステムはほかに類を見ないものではないだろうか。

今後は、Interstage Shunsakuをコア技術として、DDBJとともに、他データベースへの適用や大規模

データの複雑処理の観点から付加価値の高いサービスを研究者に提供していきたい。

参 考 文 献

- (1) DDBJホームページ .
<http://www.ddbj.nig.ac.jp/>
- (2) International Human Genome Sequencing Consortium : Initial sequencing and analysis of the human genome . *Nature* , Vol.409 , p.860-921 (15 February 2001).
- (3) DDBJホームページ : エントリの例 .
<http://www.ddbj.nig.ac.jp/sub/ref10-j.html>
- (4) 日経BP企画 : 革新のXML型データベースエンジン「Shunsaku」 . 日経BP企画 , 2004/03 .

