OPEN MIDDLEWARE

3,096万レコードの遺伝子情報データベースを高速検索。ヒトゲノムの応用科学発展を支える並列RDB「HiRDB（ハイアールディービー）」

東京大学 医科学研究所 ヒトゲノム解析センター

東京大学 医科学研究所 ヒトゲノム解析センターでは、2003年1月にスーパーコンピュータシステムを刷新したのを契機として、遺伝子情報データベースを誰でも検索できるインターネット検索サービスをリニューアルした。365億文字、3,096万レコード（2003年12月15日現在）もの膨大なデータを高速に検索するために、日立のスケーラブルデータベース「HiRDB」を採用。HiRDBならではのスケーラブルな並列処理と、柔軟なデータ格納によって高速レスポンスを実現した。HiRDBは、豊かな拡張性で日々増え続ける遺伝子情報を効率よく格納して、今後のゲノム応用科学の発展を支えている。

東京大学 医科学研究所
ヒトゲノム解析センター
機能解析イン・シリコ分野 教授
中井 謙太 氏

東京大学 医科学研究所
ヒトゲノム解析センター
ゲノムデータベース分野 助手
片山 俊明 氏

ヒトゲノム解析プロジェクト完了 オーダーメイド医療など応用の時代へ

「ヒトの全DNA配列を読み取るヒトゲノム解析計画は、世界の科学者の協調のもと、2003年に終了宣言が出ました。いよいよこれからは、読み取った情報を医療などに活かしていく時代になります」と、東京大学 医科学研究所 ヒトゲノム解析センター 機能解析イン・シリコ分野 教授 中井 謙太氏は語る。

ゲノムとは遺伝情報の総体を指す。その本体であるDNAでは、4種類の塩基（文字）で遺伝情報を記述している。世界中で解読されたDNAは延べ365億文字（2003年12月15日現在）ある。しかし、これらの配列のそれぞれの役割や組み合わせたときの機能を解析するのはまだまだこれからだ。

「読み取った情報を活かして、一人ひとりに最も適合する薬の配分を把握するパーソナル医療も本格化します。より学問的に、遺伝情報のデザイン原理を知り、生命の本質に迫ろうという取り組みも活発になるでしょう」と中井氏は目を輝かせる。

東京大学 医科学研究所 ヒトゲノム解析センターは、ヒトゲノム解析計画が立ち上がり始めた1991年、日本のゲノム解析の拠点として設立された。日本のゲノム研究における国際貢献と国際競争の拠点でもある。

生物学者、遺伝学者、オーダーメイド医療やパーソナル医療の実現を目指す基礎医学研究者など、多様な分野の研究者が、膨大な配列情報の中から自分の求める情報を得るため、高速なスーパーコンピュータシステムが重要な役割を果たす。2003年1月、ヒトゲノム解析センターでは、スーパーコンピュータシステムを十数台のUNIX計算機で構成されるシステムへと再構築した。データは毎日大変な勢いで増え続けるため、スーパーコンピュータシステムのデータを格納している日立のディスクアレイサブシステム「SANRISE」は、およそ150TBの容量を用意している。

「単一の膨大な数値計算を行うためなら、従来型の超大型コンピュータを1台設置するのが効率的かもしれません。しかし、ヒトゲノムは数値情報ではなく未知の文法に従って配列された文字情報であるうえに、解析センターのシステムは、データベース検索、ゲノム解析、配列解析、シミュレーション、共同研究など、多様なサービスを提供しなければなりません。負荷を分散することが重要なテーマになっていたのです」と、中井氏は再構築のねらいを説明する。

「読み取った情報が増えるにつれて解析の種類も増え、やりたいことがどんどん多様化しているのです」と、東京大学 医科学研究所 ヒトゲノム解析センター ゲノムデータベース分野 助手 片山 俊明氏は言葉を添えた。

遺伝子情報データベースのインターネット検索サービスを開始

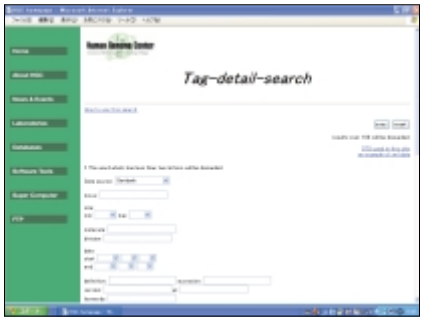
スーパーコンピュータシステムの再構築を契機として、サービスを開始したのが「HI-getシステム」である。これは、インターネット経由で誰でもアクセスして検索ができる、ゲノムデータベースの公開検索サービスである。

通常、論文中で遺伝子の配列情報を参照するにはアクセッション番号を用いる。そこで、たとえば研究者が論文を読んでいるとき、「この

USER PROFILE

東京大学 医科学研究所 ヒトゲノム解析センター

所 在 地：東京都港区白金台4-6-1
設　　立：1991年
U R L：http://www.hgc.jp
概　　要：世界中の科学者が協力して推進するゲノム解析プロジェクトに国際貢献するため、日本のゲノム解析の中心拠点として設立された。現在では、ゲノムデータベース、ゲノム解析、DNA情報解析など8分野での先端的な基礎研究のほか、日本の幅広い研究者に対する研究資材の提供、技術指導、若手研究者の受け入れを行い、国際的な利用を前提としたデータベース構築にも取り組んでいる。



HI-getシステム検索画面 http://www.hgc.jp

アクセッション番号の遺伝子はどのような配列なのか？」と疑問を抱いたらすぐ、インターネット経由でHI-getシステムにアクセスし、アクセッション番号を指定しての検索ができる。あるいは、自分が研究している生物種でみつかった遺伝子にはどのようなものがあるかを知りたいときには、キーワードを設定して条件検索を行い、該当する遺伝子を一覧表示することも可能だ。遺伝子情報を使ってできることや、やりたいことが増え、利用者の層が拡大すればするほど、HI-getシステムの重要性は高まっていく。

「インターネット経由の遺伝子データベース検索サービスということでは必ずしも目新しいものではありませんが、従来提供してきたサービスよりも柔軟に条件を絞り込んで検索できるため、利用者は、目的の情報をより短時間で入手することができるはずです」と片山氏は言う。

HI-getシステムのデータベースには、日立のスケーラブルデータベース「HiRDB」を導入した。スーパーコンピュータシステム全体を取りまとめてきた日立は、HI-getシステムの高速検索を実現するため、CPU、I/Oインタフェースなど、ハードウェアとデータベースやアプリケーションを一体として捉えて全体最適化を行った。インデクス用バッファを大きくとるなど、検索スピードを上げるためのシステム設計にも気を配った。さらに、検索タグの階層化を設計するにあたってはデータベースシステムの研究を行っている日立中央研究所プラットフォームシステム研究部が、ゲノムやDNAについての学問的な知識については、日立ライフサイエンス推進事業部が一体となって取り組んだ。

しかも、データベース構造の詳細まで把握しているSEが迅速にサポートできる国産メーカーならではの「地の利」を活かして、システムの信頼性や可用性も高度に維持している。日立だからこその総合力が、ゲノム研究を支えている重要なデータベースシステムの構築を成功させたのである。

柔軟なデータ格納で膨大なデータの高速検索を実現したHiRDB

HI-getシステムは、世界中のユーザーを対象に24時間サービスを提供する重要なデータベース・サービスであるだけにデータを2系統保持して、片方のデータに対してデータの追加・再構築を行いながら、もう片方のデータで検索サービスを継続することが可能だ。そして、データの追加・再構築が終わったら検索対象データを切り替えるといった運用を可能にしている。また、サーバには1.25GHzのプロセッサを96個搭載したハイエンドUNIX並列サーバを用いて並列RDBMSであるHiRDBの性能を、最大限に引き出せるプラットフォームを構築した。HiRDBは、365億文字、レコード数にして3,096万レコードもの膨大な遺伝子情報を登録し、インターネット経由の問い合わせに快適なレスポンスを返している。

HiRDBの並列処理は、シェアドナッシングアーキテクチャを採用している。この方式は、共有リソースがほとんどないため、CPU数の増加に対してスケーラブルな並列処理能力を発揮できる。遺伝子情報データベースでは、データを32のエリアに分割格納し、HiRDBで並列検索を行うことによって、これまでにない検索スピードを実現しているのである。

HiRDBは、データの持ち方を柔軟に設計できるデータベースである。分割格納/並列検索はそのひとつであり、アクセッション番号による検索では、データベースの登録件数が増えても検索スピードに影響しないという特性を実現できた。

さらに、HiRDB専用の日本語検索エンジン「HiRDB Text Search Plug-In」を搭載し全文検索インデクスをデータベースのフィールドに丸ごと格納。これにより、AND、OR、NOTを含む複雑なキーワード検索条件を与えた場合でも、データベース内部の閉じた論理演算で処理でき、高速レスポンスを得ることができる。

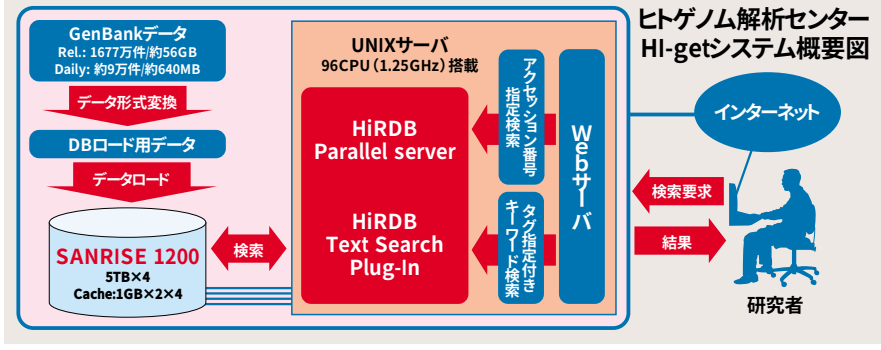
「ユーザーとしては、とにかくレスポンスが速くなって満足しています。アクセッション番号検索でも、条件検索でも、米国の著名なデータベース検索サービスと同等以上の性能です」と中井氏は自信を込めて語る。

並列データベースならではのパフォーマンスは、日々のデータロードでも発揮されている。

米国のパブリックサイトでは、毎日約9万件、2ヵ月ごとに整理した情報として約1,677万件の新しい遺伝子情報が提供されている。HI-getシステムでは現在、365億文字、3,096万レコード（2003年12月15日現在）を扱っているが、これは毎日更新されているのだ。毎日および2ヵ月ごとのデータ取り込みは、HiRDBの並列ローディングを用いて短時間で実行している。

今後、HI-getシステムの利用が増えれば増えるほど、サービスに対する要望も多様化していく。

「遺伝子データベースだけでなく、タンパク質のデータベースとの横断的な検索など、複合的な検索を実現していきます。さらに、新しい形の検索や解析も追加していきますが、常にベースとなるのはHI-getシステムであり、HiRDBとなるでしょう」と中井氏は言う。システムを拡張してもスケーラブルに高い検索性能を発揮できるHiRDBは、これからの日本のゲノム研究の発展をがっちりと支えていく。



お問い合わせ

株式会社 日立製作所
ソフトウェア事業部 販売企画センタ
〒140-8573 東京都品川区南大井6-26-2 大森ベルポートA館
TEL.03-5471-2592 FAX.03-5471-2395
http://www.hitachi.co.jp/soft/hirdb/