60 秒でわかるプレスリリース

2008 年 3 月 19 日
独立行政法人 理化学研究所

約 30 万個の分散データベース群を高速に統合検索する技術を開発

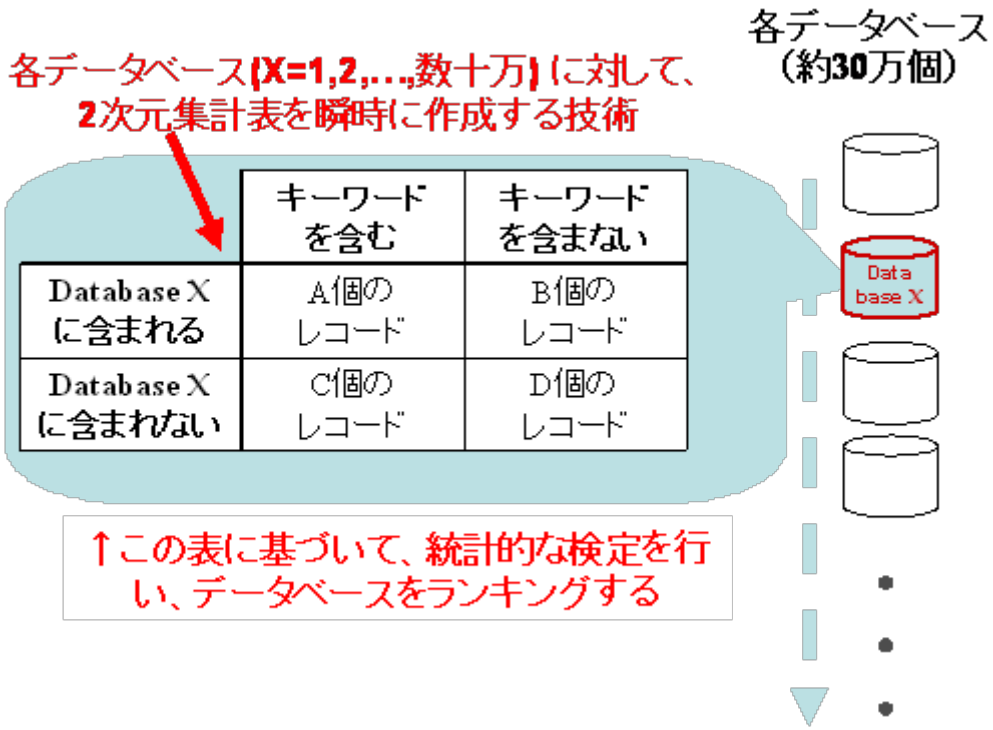
- 生命情報基盤データベース群の統合検索エンジンとして活用 -

生命の設計図であるゲノムを解読し、その機能を見つけることは、生命現象そのものを知るばかりか、がんなどの病気の発症メカニズムの解明や、新たな作物の開発など、さまざまな夢をかなえる基盤となります。わが国では、理研ゲノム科学総合研究センターが、ゲノム解読の中核的な機関としての使命を果たしていますが、世界でも新たなゲノム解読が続き、遺伝子情報のデータ蓄積量はすさまじい勢いで激増しています。

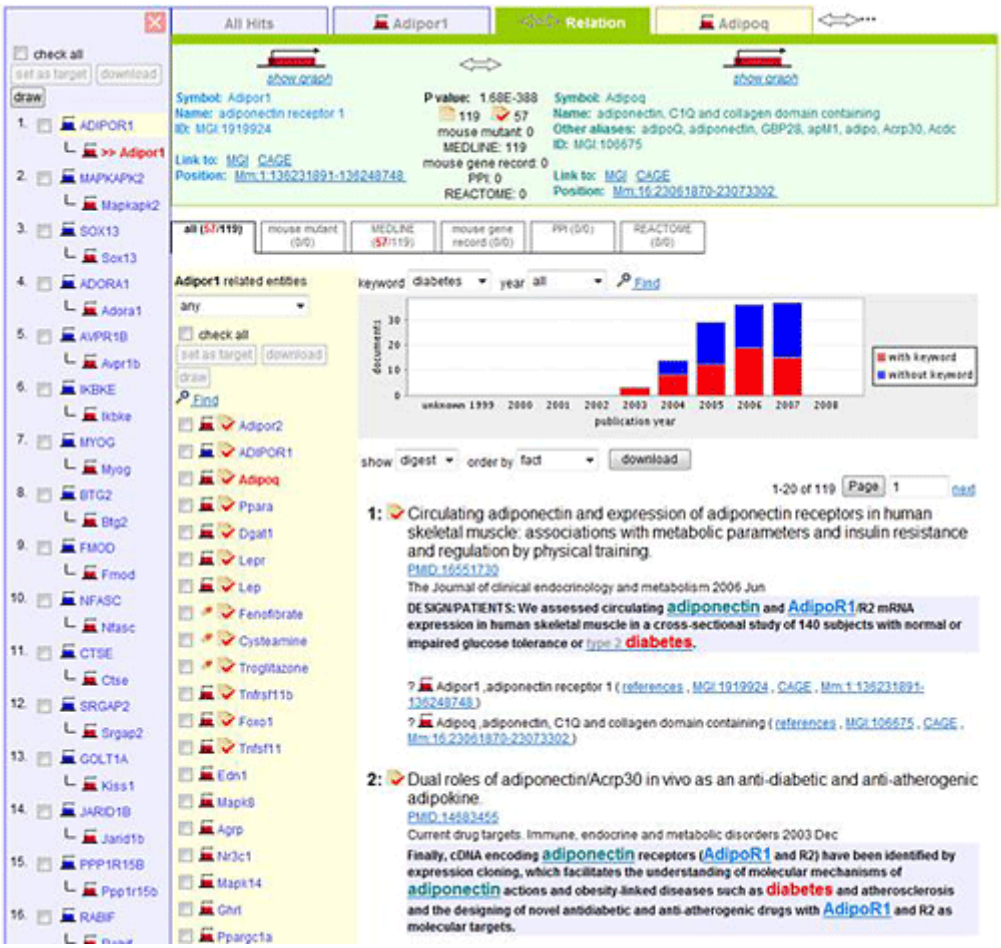
将来的には、この遺伝子情報を活用して人工的に遺伝子をデザインする「ゲノム設計」を行うことで、バイオエタノールなどの植物由来の燃料や、健康を促進する機能性野菜などを作ることができるかもしれません。しかし、この実現のためには、膨大な量の遺伝子データを使いこなすことが必要となります。

理研ゲノム科学総合研究センターのオミックス情報統合化研究チームは、数 10 万種類の分散データベース群をキーワードで横断的に検索し、必要とする関連情報を多く含んでいるデーターベースを統計的に選び出して、データーベース単位で高速にランキングする技術「GRASE」を開発しました。さらに、文献やゲノムデータベース群の検索にこの技術を適用して、任意のキーワードで「機能」と「遺伝子」のつながりを推論的に調べることができる強力なウェブ検索サイト「PosMed」を公開しました。ここでは、医学用語に限らずあらゆるキーワードでの検索が可能で、例えば、疾患感受性遺伝子の候補をデータベースの中からランキングづけしたり、変異が見つかった遺伝子の機能解釈を行なうことができます。また、大勢の研究者の中から、必要な専門分野で活躍している人、論文の査読ができる人を探索することに役立つウェブ検索サイト「Researcher Finder」も公開しました。

「GRASE」技術は、バイオ分野だけでなく、あらゆる分野のデータベース統合検索サイトとして活躍することが期待できます。



（図）データベースの高速ランキングを行うための統計(上)と PosMed における文献情報の閲覧画面(下)



報道発表資料

2008 年 3 月 19 日
独立行政法人 理化学研究所

約 30 万個の分散データベース群を高速に統合検索する技術を開発

- 生命情報基盤データベース群の統合検索エンジンとして活用 -

◇ポイント◇
- 多数のデータベース群をキーワードからランキングして利用者にわかりやすく提示
- 疾患関連遺伝子の探索研究に役立つ、情報検索サイトとして成果をあげる
- 専門分野の研究者を探し出すための検索サイトやセマンティックウェブにも応用

独立行政法人理化学研究所（野依良治理事長）は、文献情報やゲノム情報や生体内分子ネットワーク情報など生命科学分野における主要な公開データベース群を国内外から収集し、遺伝子や代謝物や薬物などのトピックごとに細分類化した個別データベース群（約 30 万個）を理研内に分散データベースとして試験的に構築しました。さらに、これらをキーワードで横断的に検索し、関連する情報がより多く含まれているデータベースを統計的に選び出して、データベース単位で高速にランキングし、重要な情報を含むデータベースを利用者が容易に選び出すことができる技術「GRASE（General and Rapid Association Study Engine)」を開発しました。理研ゲノム科学総合研究センター（榊佳之センター長）オミックス情報統合化研究チームの豊田哲郎チームリーダーらによる研究成果です。

この技術を使って「機能」と「遺伝子」のつながりを、既存の知識情報に基づき推論的に検索できる強力なウェブ検索サイト「PosMed（Positional Medline)」を公開しました。PosMed は、医学用語に限らず、あらゆる英語のキーワードから検索することができます（日本語のキーワードは医学用語のみが使用可能)。例えば、ポジショナルクローニング（遺伝学的な手法で狭められた染色体区間内に存在する多数の遺伝子群の中から、原因遺伝子を見つけ出す研究）では、疾患関連遺伝子の候補を、上記データベースを手掛かりにランキングできるほか、変異が見つかった遺伝子の機能解釈や、バイオマーカーの機能解釈を行う目的にも利用することができます。PosMed は、理研ゲノム科学総合研究センターゲノム機能情報研究グループ（城石俊彦プロジェクトディレクター）が推進するENU変異マウス※1開発プロジェクトで、原因変異遺伝子の候補検索システムとして使われてきた実績があり、これまでに 60 以上の遺伝子変異の同定研究に貢献してきました。マウス以外に、ヒト、ラット、シロイヌナズナの遺伝子も検索可能です。

以上のサービスは理研ウェブサイト内のRIKEN Hub Database Project（http://omicspace.riken.jp/）で公開しており、誰でも無料で利用することが可能です。また、今回の成果は、セマンティックウェブ※2（次世代のウェブ技術）における統計的な検索技術が、バイオ分野の情報検索に効果的であることを示し、今後、その他さまざまな分野の情報検索など幅広い応用が期待できます。本研究成果は、英国の科学雑誌『*Bioinformatics*』のオンライン版（3 月 28 日付け）に掲載されます。

1.背　景

理研ゲノム科学総合研究センターでは、これまで、細菌からヒトまで幅広い生物のゲノム解読を推進してきました。ゲノムを解読する装置（シーケンサー）の能力は、現在も急速に進歩しており、世界中で新たな生物から次々と解読される遺伝子情報のデータ蓄積量が指数関数的に急速な増加を続けています。一方で、地球環境保護の観点から化石燃料に代わるバイオエタノールなどの再生可能なバイオ燃料の開発が期待されています。今後、生物由来の素材に依存する経済活動（Bio-based Economy）が拡大すると、遺伝子資源が石油資源にとって代わるようになり、遺伝子資源の世界的な確保がより重要になるだろうとの予測もあります※3。

遺伝子の本質は情報（ATGCの 4 種類の塩基の並び方を表す配列）としてデータベースで保管でき、必要に応じて、DNA合成技術で物質化して生物に導入する研究が、安全に管理された実験施設内で可能です。このため、将来的には、データベースに蓄積された遺伝子情報を組み合わせることで、効率的なエネルギー生産などの新しい機能を付与した有用生物を人工的にデザインする“ゲノム設計学※4”（図 1）が重要技術になると考えられており、データベースを核とした生命情報基盤の整備が求められています。

多様な生物種から見つかる膨大な遺伝子情報の中から、有用な遺伝子セットを人間の都合（例えば、“経済選択”など）によって選び出して、別の生物に導入することによる人為的な生物進化は、“オミックス進化（Omics-driven Evolution）※5”と呼ばれ、従来の“自然選択”による“ダーウィン進化”だけでは起こり得なかった“有用生物の創造”が、データベースを基盤とした生命科学研究によって将来的に可能になっていくと期待されます。

生命科学分野のデータは、今後も急増することが予想されており、データの種類も大幅な増加傾向にあることから、我が国でもこの情報資源を適切に保全し有効利用することが必要です。そこで、データベースに関する活動を強化する国の方針※6に沿って、理研は生命情報基盤研究部門（豊田哲郎部門長）を 2008 年 4 月から新設し、国内外の他機関とも連携して、共有の情報基盤※7を構築していきます。

生命科学分野のデータベースは、データ量だけでなく個数も種類も大幅な増加傾向にあり、世界的には約 10,000、日本にも約 250 のデータベースがあるといわれています。このため、データベースを有効に活用するためには、多様かつ多数のデータベースに対する検索を瞬時に行う技術の確立が重要な課題です。例えば、数 10 万個のデータベース群を 1 つのキーワードで横断的に検索した場合、それぞれのデータベースから返ってくる検索結果も数 10 万に及ぶため、さらにその中から有用な検索結果を瞬時に絞り込んで、ランキング表示させる技術が必要となっています。

2. 研究手法

理研では、文献情報やゲノム情報や生体内分子ネットワーク情報など、主要な公開データベース群を国内外から収集し、これらをさらに、遺伝子や代謝物や薬物などのトピックごとに細分類化することで、トピックごとの個別データベース群（約 30 万個）を理研内に試験的に構築しました。そして、これらを対象に、キーワードから関係性が高いものを、データベース単位で瞬時にランキング化し、重要な情報を含むデータベースを容易に選び出すことができる技術開発研究を行ってきまし

た（図 2)。個々のデータベースの構築は、各トピック（各遺伝子など）と、それらについて記述している文献とを個々に対応づけることで実現しますが、通常の遺伝子名にはその略称や別名が多数存在するため、単純にテキストマッチングで対応づけ処理すると、誤った対応づけとなってしまいます。このため、マウスゲノムに存在する約 2 万の遺伝子を、作業者が 1 つずつチェックしながら、1,600 万件を超える生命科学分野の文献情報（Medline※8やOMIM※9やミュータントなど）に対して正しい対応づけルールを約 2 年かけて作成し、各遺伝子に対応する個別データベースを構築しました。さらに、遺伝子と代謝物の関係性や研究者情報、代謝パスウェイ情報などのさまざまな要素間のつながりを、次世代のウェブ技術であるセマンティックウェブという形式で知識表現したデータも作成しました。そして、これらのデータを統合的に検索するための検索エンジンとして「GRASE（General and Rapid Association Study Engine)」を開発し、PCクラスタで分散処理する高速検索処理を実現しました。GRASEは、数 10 万個のデータベースのそれぞれに対し、図 3 に示すような 2 次元集計表を高速に算出します。この各集計表を統計検定（フィッシャーの正確検定）することで、各データベースがキーワードに関する情報を有意に含んでいるかどうかを調べあげ、その検定結果に基づいて、関連性の高いデータベース群を瞬時にランキングします。また、GRASEは、文献が発表された年代で区切って上記集計表を作成する機能もあり、時間軸に沿ったトレンドを解析することも可能です。

ランキング対象が遺伝子の場合は、各遺伝子に対応する文献情報としてさまざまな種類の情報ソース（MedlineやOMIMなど）があり、各遺伝子に複数のデータベースが対応しています。そのため、上記の 2 次元集計表をさらに遺伝子ごとに集めることで、各遺伝子に対応する 3 次元集計表を作成し、これを統計検定して全遺伝子の総合的なランキングを計算します。GRASEは、これら一連の処理をわずか 1 秒程度の間にすべて実行できる点で、従来にない技術となっています。

これまでは、セマンティックウェブに対して上記のような統計的なサーチを行うための検索言語は存在しませんでした。今回の技術を応用することで、統計的なサーチが可能な新たな検索言語「GRASQL」を定義する研究が可能になり、生物学分野に典型的な、疾患感受性遺伝子の候補推定など、いくつかのケースでその検索パターンの有効性を検証しました。

3. 研究成果

本研究では、マウスが持っている約 2 万の遺伝子ごとに、その遺伝子に関する文献などの情報をまとめたデータベースを構築し、GRASEを使って、キーワードからそれらのデータベースをランキングできる情報システムを構成しました。これは、キーワードに関連性の高い遺伝子を文献情報に基づいてランキングすることに相当する検索であることから、「PosMed (Positional Medline) 」という名称のサービスでウェブに公開しました。PosMedでは、ユーザが入力した任意のキーワードを使って、関連性の高い遺伝子から順番にランキングリストを作成します。さらに、検索条件として染色体領域を限定した場合は、その中に含まれる遺伝子だけに限定してランキングを行います（つまり、「キーワード → 遺伝子A → 染色体領域」というつながりを探索します)。 さらに、この探索と並行して、PosMedは、遺伝子

ー遺伝子間の関連性情報を使った推論も自動的に行い、ネットワーク的なつながり「キーワード → 遺伝子B → 遺伝子C → 染色体領域」で条件に合うものも探し出して、総合的にランキングします（図 4）。この他にもさまざまな生体内相互作用情報（代謝物ー遺伝子、変異マウスー遺伝子、薬物ー遺伝子、疾病ー遺伝子、タンパク質間相互作用やオルソログ※10情報など）をPosMedに登録しており、任意のキーワードを使って「機能」と「遺伝子」を網羅的に結びつけることができ、その推論過程で使った文献の詳細情報も閲覧できる強力なツールです（図 5）。PosMed は、医学用語に限らず、あらゆる英語のキーワードから検索することができます（日本語のキーワードが入力された場合は内部的に医学用語辞書※11を使って英語のキーワードに自動変換してから検索を実行します。このため、PosMedで使える日本語のキーワードは医学用語のみです）。さらに、このサイトでは、ヒトやマウスの各組織における遺伝子発現量や遺伝子構造情報もゲノムブラウザで視覚的に閲覧できるため、ポジショナルクローニング（遺伝学的な手法で狭められた染色体区間内に存在する多数の遺伝子群の中から、原因遺伝子を見つけ出す研究）において、疾患関連遺伝子の候補を上記データベースを手掛かりに絞り込むために利用できます。このほか、変異が見つかった遺伝子の機能解釈や、バイオマーカーの機能解釈を行う目的にも利用することができます。このシステムは、ポジショナルクローニングの候補遺伝子選びのために世界中から利用され始めており、理研におけるENU変異マウス開発プロジェクトでも、これまでに 60 以上のENU変異マウスで遺伝子変異の同定研究に貢献してきました。

このほか、理研の研究者が過去 5 年間に発表した文献要旨についても、研究者ごとにデータベース化し、GRASEで検索できるウェブ検索サイト「Researcher Finder」を開発して公開しました。このサイトでは、理研研究者を中心に 5,585 名の研究者（2008 年 3 月現在）のそれぞれに対応したデータベース群を構築して、検索の対象にしています。これにより、キーワードにヒットする論文要旨をより多く発表している順で各研究者を探し出すことができるため、大勢の研究者の中から、ある専門分野で活動している研究者を探したい場合（論文の査読者探しなど）に利用できます※12。

以上のサービスは、理研ウェブサイト内のRIKEN Hub Database Project（http://omicspace.riken.jp/）で公開しており、誰でも無料で利用が可能となっています。

4. 今後の期待

理研では、セマンティックウェブでのデータベース公開基盤※13を準備しており、今後、さまざまなデータをここから公開していく予定です。多くの人がセマンティックウェブをわかりやすく利用するためには、検索技術が重要であり、今回の成果の応用が期待される分野です。今後、検索対象の個別データベースの数が数千万個以上になっても高速検索が可能なようにGRASEを拡張（スケーラブル化）していく計画です。これにより、セマンティックウェブで表現された膨大な知識データを、さまざまな概念や観点から多面的にまとめた個別データベース群として細分化して検索対象とすることが可能になり、セマンティックウェブの統計的な高速サーチ

エンジンとしての汎用化が期待されます。

現在も、ゲノム解読を行うシーケンサーの解読能力は飛躍的な進歩を続けており、それを使って生物種や個体から膨大な遺伝子情報が世界中で解読され、データベースに蓄積するデータ量も指数関数的に伸びています。遺伝子情報のデータベースは、単にDNAの塩基配列情報だけでなく、他のさまざまな種類の実験から得られた知見と統合化することによって、初めて機能的な観点からの有効活用が可能になるため、遺伝子情報資源の多面的な活用を推進するための情報基盤づくりが重要であり、今回の技術はこうした基盤づくりに貢献することが期待されます。今後、生物由来の材料やバイオエタノールなどの生物由来の燃料に依存する経済活動（Bio-based Economy）が拡大すると、遺伝子資源が、現在注目されている石油資源にとって代わるだろうとする予測もあり、化石資源の少ない我が国でも「遺伝子資源立国」という観点から、遺伝子情報を高度に活用するための情報技術開発に期待が高まっています。

（問い合わせ先）
独立行政法人理化学研究所
ゲノム科学総合研究センター
オミックス情報統合化研究チーム
チームリーダー　豊田　哲郎（とよだ　てつろう）
Tel：045-503-9610 / Fax：045-503-9553

横浜研究推進部　企画課
Tel：045-503-9117 / Fax：045-503-9113

（報道担当）
独立行政法人理化学研究所 広報室 報道担当
Tel：048-467-9272 / Fax：048-462-4715
Mail：koho@riken.jp

<補足説明>

※1 ENU 変異マウス

化学変異原であるエチルニトロソウレア（ENU）を用いて、ゲノム遺伝子上にランダムに１塩基の変異を誘発させたマウス。理研では、ヒト疾患のモデルを含む多数の突然変異マウスを開発し、表現型の解析や原因遺伝子の探索を行っている。

※2 セマンティックウェブ

ワールドワイドウェブ（WWW）の発展形として英国の計算機科学者であるティム・バーナーズ・リーによって提唱されている次世代のウェブ技術。ワールドワイドウェブは、ネットワーク上に置かれた文書などのリソース間をハイパーリンクで

つなぐもので、インターネット上の標準的なインフラとして爆発的な成功を収めた。しかし、ハイパーリンクは、人間がそのリンクをたどりながら読み進めていくのには適しているものの、単純に 2 つのリソースを結びつけているだけなので、そのリンクがどのような関係づけを意味しているかは表現していない。リンクの意味については、テキストに書かれた内容を人間が読んで解釈するしかなく、コンピュータが意味を認識し、高度な知識処理を行うための情報をほとんど含んでいないことが、ワールドワイドウェブの問題点として指摘された。この反省から、ウェブにセマンティクス（意味論）を与えることが求められるようになった。つまり、情報を持つ文書を機械可読な形で提供できるようにすること、および、リンクにその関係を示す値をつけ加えられるようにすることで、我々が自ら読む以上の情報を、コンピュータの力を借りて取り出せるようにする。この実現を目指すのがセマンティックウェブである。

参考資料：神崎正英： セマンティックウェブのための RDF/OWL 入門： 森北出版株式会社、2005 年

セマンティックウェブのライフサイエンス分野への応用について
（http://omicspace.riken.jp/publications/evolution/page7.html）

※3 遺伝子資源が石油資源にとって代わるようになるだろうとの予測

Robert E. Armstrong “From Petro to Agro: Seeds of a New Economy” Defense Horizons, No.20, Oct. 2002.

“今日、炭化水素分子が商業の基本素材となっている。バイオ由来の素材に依存する経済においては、遺伝子が石油にとって代わるだろう。つまり、いま、炭化水素分子（石油）資源の確保に躍起になっているように、近い将来、遺伝子（植物や動物）の広範囲で多様な供給を確保する要求が高まるだろう。この転換は、安全保障面に与える影響もはらんでいる。石油資源に恵まれている国家との関係の重要性が低下し、遺伝子資源の豊富な国家―そのほとんどが赤道付近の生物資源が多様な地域になるだろうが―との関係が、極めて重要になってくるものと思われる。”

※4 ゲノム設計学（Genome Design）

近年の DNA 合成技術の進歩によって、合成エラーが少なく、より長い DNA 鎖の合成が可能になりつつあり、これらの合成技術を使ってゲノムの一部または全部を人工的にデザインした生物を創りだそうとする試みが微生物を中心に世界的に進んでいる（合成ゲノミクス）。これらの合成に関する技術的側面を発展させる一方で、情報的側面からゲノム設計学を発展させる必要がある。データベースの遺伝子情報に基づいてどのようにゲノムを設計すればよいかは、生物学のより深い理解と、情報ツール・インフラの整備が必要であり、今後の課題となっている。（参考：Christopher A. Voigt, “Life from information” *Nature Methods*, 5(1) 27-28,2008）

※5 オミックス進化（Omics-driven Evolution）

オミックスとは、生体内に存在する分子（ゲノムや代謝など）の状態を、網羅的に

計測する技術（シーケンサーやDNAチップなど）が近年急速に発達したことで可能になった研究手法である。オミックスでは、これらの計測装置を使うことで、物質世界に存在する生体内の分子状態を、網羅的なデータとして情報世界に写像化し、次に、その情報を他の情報とも照らし合わせながら統合的に分析することで、生体の状態をより詳細に探るものである。オミックスによって、新たな生物種のゲノム情報が次々と調べられていき、膨大な情報がデータベースに蓄積されつつある。こうした遺伝子情報資源を基盤として、さまざまな生物の遺伝子から有用なものだけを情報的に選び出し、それらを組み合わせて、有用物質を創り出す新たな代謝パスウェイを人為的にデザインできる可能性もあることから、未知の生物から有用な遺伝子を探し出そうとする競争が今後加速する可能性もある。このように、情報世界を介した異生物種間の遺伝子の拡散現象を生物における新たな進化メカニズムという側面から捉える『オミックス進化』についての概要は、図6および下記の資料を参照のこと。

豊田哲郎「コンピュータの中の脳　-情報基盤の進化論-」生体の科学
（http://omicspace.riken.jp/publications/071228_toyoda.pdf）
59(1):20-32, 2008

豊田哲郎「ゲノム解読から生命戦略の解明へ」
（http://omicspace.riken.jp/publications/toyoda1.pdf）
Bionics 26-30, Feb., 2007

※6 データベースに関する活動を強化する国の方針

第3期「科学技術基本計画」（2006年3月28日閣議決定）に基づき総合科学技術会議が策定したライフサイエンス分野の推進戦略では、戦略重点科学技術の1つとして「世界最高水準のライフサイエンス基盤整備」が掲げられており、来年度からの理研の中期目標ではデータベース基盤への貢献を掲げている。

※7 共有の情報基盤

情報社会において共有の情報基盤が担う役割の重要性について
（http://omicspace.riken.jp/publications/evolution/page11.html）

※8 Medline

MEDLINE（Medical Literature Analysis and Retrieval System On-Line）は、米国国立医学図書館（National Library of Medicine; NLM）が提供する生命科学に関連する文献抄録データベース。2008年現在、米国を中心とした80カ国以上の国で出版される学術誌に掲載された1,600万を超える文献抄録が登録されており、誰でも無料で利用できる。

※9 OMIM

OMIM（Online Mendelian Inheritance in Man）は、米国国立医学図書館に属する国立生物情報センター（National Center for Biotechnology Information：NCBI）

が提供するヒト遺伝子変異と遺伝病のカタログデータベース。遺伝子機能の疾患との関連がカタログとしてまとめられてデータベース化されている。

※10 オルソログ

進化の過程で共通祖先から種分化によって生じた、種間で共通に存在する遺伝子のうち、祖先遺伝子の持つ機能をそのまま保存している遺伝子のこと。

※11 医学用語辞書

医学用語辞書として、『ライフサイエンス辞書 2006』を使用した。この辞書は、京都大学大学院薬学研究科生体機能解析学分野 金子 周司教授らによるライフサイエンス辞書プロジェクトの成果の一部で、生命科学を中心に広範な領域の学術用語について 69,311 件の日本語―英語の対訳が含まれている。

※12 Researcher Finder

このサイトは利用者が理研の中から適切な専門家を探すことなどを目的としており、研究者の業績について格付けを行うものではない。このサイトで検索対象となる研究者と論文は、所属が理研になっている論文を最近 5 年間に発表したものに限定しており、Medline が検索対象の範囲である。このため、ここに表示される研究者ごとの論文リストは、その研究者の業績の一部分であり、必ずしもすべてを網羅していない場合がある。また、このサイトはテキストマイニング技術で構築されている。そのため、ここに表示される研究者の論文リストには同姓同名の他の研究者の論文が含まれている場合がある。また、このサイトでは個人情報保護の観点から研究者のメールアドレスは掲載していないため、ヒットした論文に記載されている責任著者からコンタクトすることが必要である。

※13 理研が推進するセマンティックウェブでのデータベース公開基盤

セマンティックウェブ技術を使って構築中の理研総合データベース事業
（http://omicspace.riken.jp/publications/evolution/page9.html）
（将来、公開を予定）

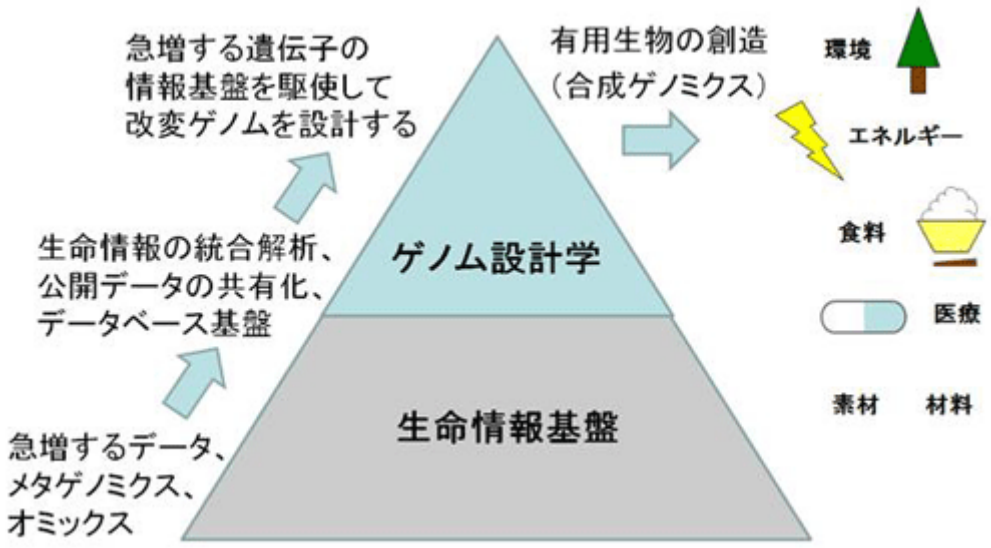


図１　ゲノム設計学

今日、世界中で新たな生物から新しい遺伝子が次々に発見され、蓄積される遺伝子情報のデータ量は、指数関数的に増加している。また、ゲノムに限らずトランスクリプトームやメタボロームなど様々な種類の網羅的な分子情報（オミックス情報）も急速に増加していることから、これらの大量のデータを管理しつつ統合的に解析するための生命情報基盤研究が重要になっている。将来的には、効率的なエネルギー生産などの新しい機能を付与した有用生物を、データベースに蓄積した遺伝子情報を組み合わせることでゲノムの一部または全部を人工的にデザインする“ゲノム設計学”が重要技術になると予想され、データベースを核とした生命情報基盤の整備は将来の科学技術の発展にとって不可欠なものになっている。（実験的な手法については「合成ゲノミクス」という用語が用いられることが多いが、ここでは情報的な手法に焦点を当てるために「ゲノム設計学」という用語を使っている。）

図２　分散データベース群の統合検索による検索結果表示画面

キーワード”diabetes”（糖尿病）から統合検索を実行し、検索結果を表示した画面。青字で書かれたヒットリストは、それぞれの個別データベースに対応しており、赤字で各データベースにおけるヒットレコード（キーワードを含むコンテンツ）の数が表示され、統計的な検定で関連性がより強いと判定されたものから順番にリスト表示されている。さらにわかりやすくするために、各データベースは Drug や Disease など

のカテゴリごとに束ねられており、各カテゴリ内では上位 10 件の個別データベースがリスト表示されている。ここからリンクをたどっていくことでさらに詳細画面（図４や図５）を閲覧していくことができる。この統合検索のサイト（http://omicspace.riken.jp/db/）は誰でも無料で利用できる。

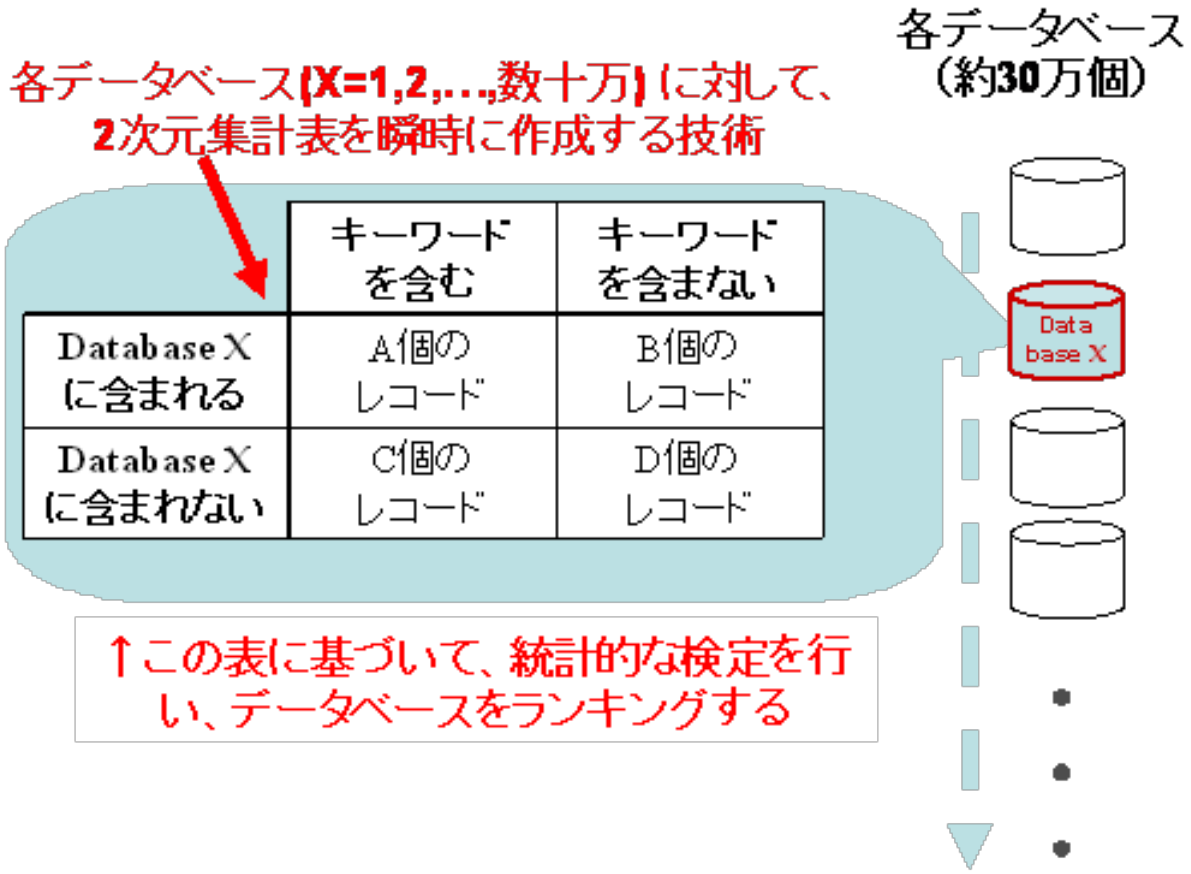


図 3　データベースの高速ランキングを行うための統計処理

GRASE は、ユーザが任意に指定したキーワードあるいはそれらの論理式からなる検索クエリ（問い合わせ）に対して、データベースごとに上記の 2 次元集計表を作成する。図ではデータベース X について、データベース X に含まれかつキーワードを含むレコード数 A、データベース X に含まれかつキーワードを含まないレコード数 B、データベース X に含まれないがキーワードを含むレコード数 C、データベース X に含まれずキーワードも含まないレコード数 D から構成される 2 次元集計表が示されている。さらに得られた 2 次元集計表を用いて統計的な検定を行い、全データベースをランキングするためのスコアを計算する。これらの動作を約 30 万件のすべてのデータベースについて 1 秒程度のわずかな時間で実行する。

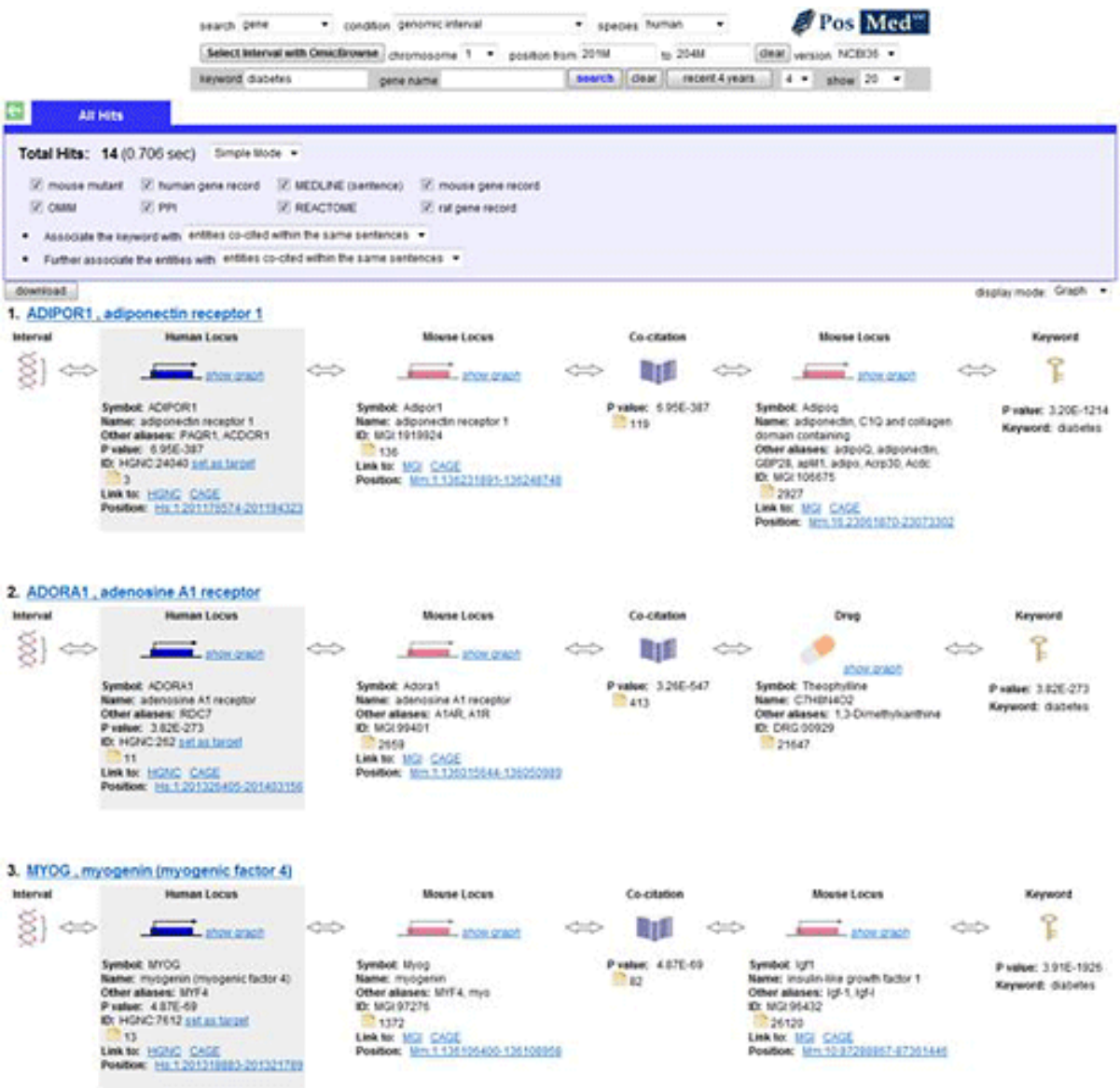


図 4　PosMed における検索画面

ヒト 1 番染色体の 201Mbp～204Mbp の間に存在する遺伝子群の中から、diabetes（糖尿病）に関係する遺伝子をランキングした様子。ADIPOR1（adiponectin receptor 1）が 1 位にランキングされている。これは、キーワードから、まず、マウスのタンパク質 adiponectin に関連づけられ、それを経由してマウスの遺伝子 Adipor1 に、さらにそのオルソログであるヒト ADIPOR1 に結びづけられて、上記の染色体領域に存在する条件を満たすものとして出力されている。PosMed はこの検索を 1 秒足らずで実行するが、この間にすべての Medline のテキスト検索処理や、その結果に基づく推論処理が完了している。

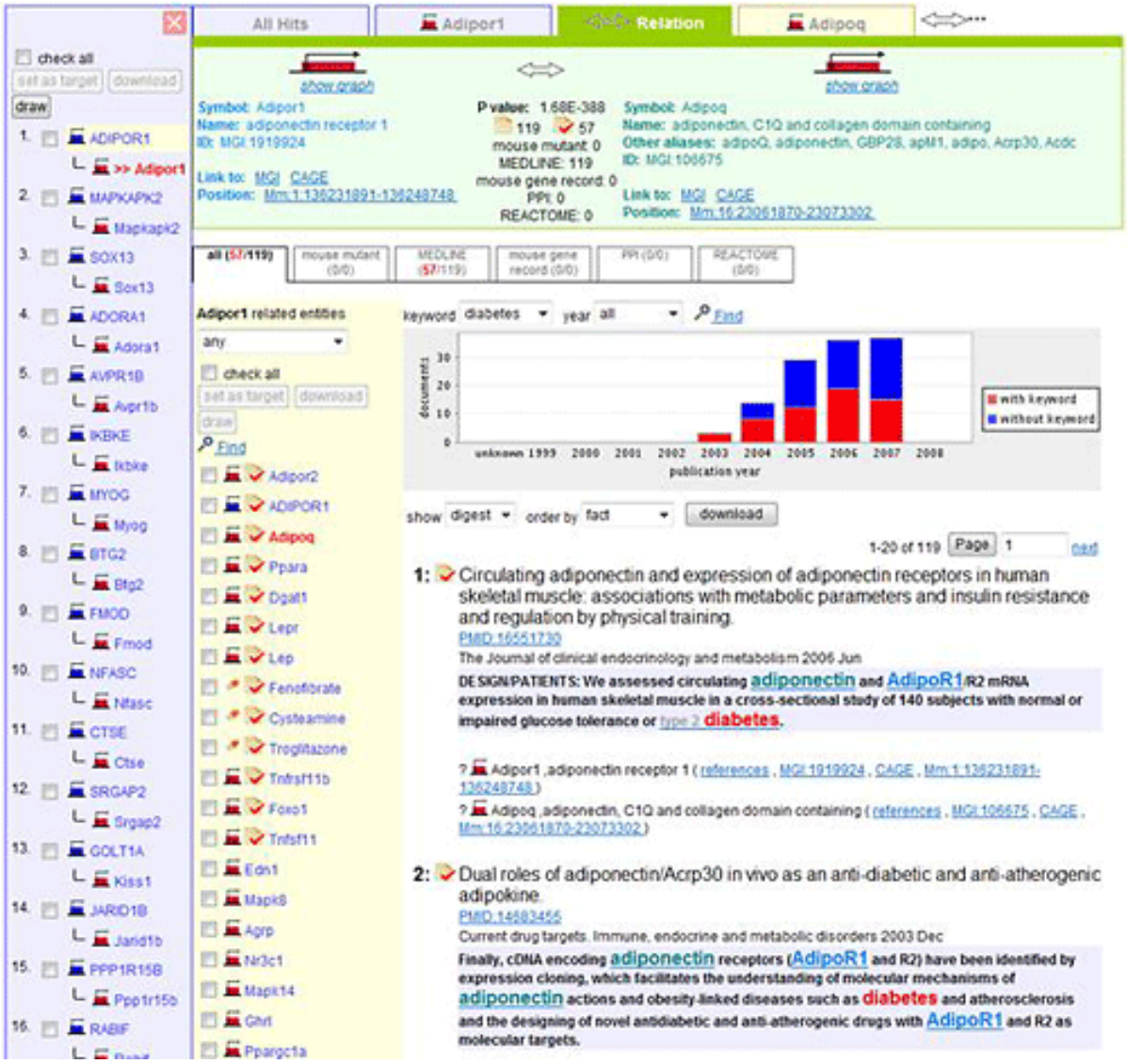


図 5　PosMed における文献情報の閲覧画面

図 4 に示した画面の検索で行われた推論過程についての詳細情報を表示させた様子。遺伝子 Adiopr1 と adiponectin の関係性についての文献情報を表示させた画面で、キーワードである diabetes（糖尿病）が文中で赤くハイライトされている。また、これらの関係性を論じている文献数の年次推移も棒グラフで表示されており、文献数は 2003 年ごろから急速に増加しており、その大半が糖尿病関係で論じられていることが一目でわかる（棒グラフの赤い部分は、diabetes というワードを文中に含む論文数である）。

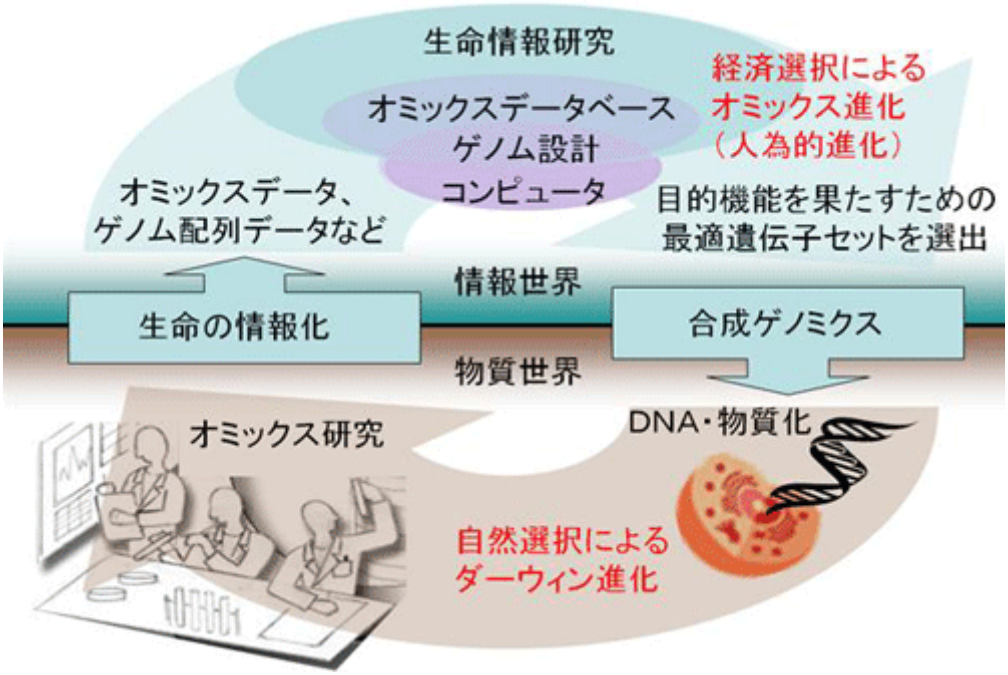


図 6　情報世界にまで拡張された生物進化のサイクル（Omics-driven Evolution）

オミックスデータ、ゲノム配列データなどの生命情報のデータベース化が進むことで、人間が意図した機能を最大限に発揮する最適な遺伝子セットを幅広い生物種の遺伝子から選出することが情報的に試みられるようになりつつある。このように、情報世界を介した異生物種間の遺伝子の拡散現象を生物における新たな進化メカニズムという側面から捉える“オミックス進化”ではデータベースが重要な役割を担う。