様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月　 ２ 日現在

機関番号：３２６４４
研究種目：基盤研究(B)（一般）
研究期間：2013～2015
課題番号：２５２８０１０８

研究課題名（和文）データ識別子を活用した全生命科学データベースの統合化技術開発

研究課題名（英文）Hyperlink Management System for automated integration of biological databases by use of data IDs

研究代表者
今西　規（IMANISHI, Tadashi）

東海大学・医学部・教授

研究者番号：８０２７０４６１

交付決定額（研究期間全体）：（直接経費）　　14,200,000円

研究成果の概要（和文）：生命科学における世界の主要なデータベースについてデータIDの対応関係を解析し統合化したリンク自動管理システム（http://biodb.jp/）を発展させ、新たにゲノム多型分野の各種データベースを追加し、芋づる式検索が実行できるようにした。これはゲノム医科学分野での知識発見に役立つと期待される。また、データIDやURLを一括取得できるウェブサービスを開発した。今後は関連発見ツール等の新機能の導入を計画している。

研究成果の概要（英文）：Hyperlink Management System (HMS; http://biodb.jp/) is an automatically updating, integrated database of data IDs and their relationships that are used in major biological databases. We newly added several databases concerning genomic polymorphisms into HMS, and realized a comprehensive data search in a recursive manner through a chain of hyperlinks, which can promote discoveries in the field of genome medicine. Also, we developed and released a web service for bulk retrieval of data IDs and URLs. In the near future, we plan to install new functions such as enrichment analysis tools.

研究分野：生命情報学

キーワード：生命情報学　生体分子　統合データベース　論理的データ検索

１．研究開始当初の背景

生物学研究では、事前に膨大な量の文献調査とデータの収集を行い、何らかの仮説を立てた後に実験を開始する。そこで、網羅的な文献調査とデータ収集を効率的に行える情報支援システムがあれば、さらにはデータ解析のための情報基盤が提供されていれば、生物学研究は加速されると予想される。一方で、測定装置の高度化により大量データが容易に取得可能になり、「data-driven science」の研究スタイルが定着した。このような状況で、生物学研究を加速しうるバイオインフォマティクスの情報基盤が求められている。また、昨今は統合データベースプロジェクトも実施されているが、そこでは異なる種類のデータを関連づけるという統合化の作業は遅々として進んでいない。一方、海外ではウェブサービスの統合化や標準化の動きが活発であり、システム生物学研究と合わせて大きな成果を挙げている。このような状況から、生命科学のデータ統合に本格的に取り組むことが必要と考えられる。大規模なデータ統合を行うことによって、新発見がなされる可能性は高まる。また、単に統合化されたデータをユーザに提供するのではなく、「有望な仮説を提示できる統合データベース」が非常に重要である。

２．研究の目的

生命科学におけるさまざまな分子データと知識は、遺伝子配列やタンパク質構造や疾患情報などその種類ごとに分類されて、世界中の 1600 種類を超えるデータベースに分散して格納されている。これらの分子データや知識の間には、例えばある SNP が疾患の原因であるなど、互いに論理的なつながりを持つ場合が多い。そこで本研究では、データ間の論理的なつながりに基づいて生命科学の多様な分子データと知識を連結・再編成し、生命科学データの網羅的な統合化を行う。そして、分子データと知識の巨大ネットワークから直接・間接に関連する全データを芋づる式に検索する機能を実現する。次に、連結されたデータ間の相関構造を考慮しつつ、ユーザが投入したデータの集合（遺伝子名の集合など）の「特徴」を自動探索して発見し、さらに性質の類似する別データを自動探索するシステムを構築して公開する。以上の統合データと発見的情報処理システムは、研究者が有望な作業仮説をみつけることを支援する世界に類のないデータベースとして、研究者に無償で提供することを目的とする。

３．研究の方法

【テーマ１】リンク自動管理システムの拡張による「論理的統合データベース」の開発
【テーマ２】論理ネットワーク上の芋づる式検索の開発
【テーマ３】データ取得用ウェブサービスの開発
【テーマ４】網羅的関連発見ツール（enrichment analysis ツール）の開発
【テーマ５】類似データ自動検索ツール（prioritization analysis ツール）の開発

本研究では上記のテーマ１からテーマ５を設定し、それぞれに年度ごとの達成目標を設け、計画的に研究開発を進めていく。本研究テーマはいずれも計算機を用いたバイオインフォマティクス的研究であるが、一部には理論的研究も必要であり、類似手法の調査も行いつつ世界最高性能の統合データベースと発見的情報処理システムの開発をめざす。体制としては、代表者がシステム全体の設計を行い、代表者のチームのテクニカルスタッフが世界の多数のデータベースの基礎調査、データ統合や検索システム・enrichment analysis のシステム開発を担当する。

４．研究成果

生命科学における世界の主要なデータベースについてデータ ID の対応関係を解析し統合化したリンク自動管理システム（http://biodb.jp/）を基盤として、以下の研究開発を行った。

【テーマ１】リンク自動管理システムの拡張による「論理的統合データベース」の開発、に関しては、世界の主要な公開データベースの追加を行った。特に、ヒトのプロテオーム分野のデータベースについては、Human Protein ATLAS、Peptide ATLAS、Ensembl Protein、neXtProt をリンク自動管理システムに追加することにより、ヒトのタンパク質に関する新しい実験データに容易にアクセスできるようになった。ヒトゲノムの single nucleotide polymorphisms（SNPs）に関するデータベースについては、dbSNP、VaDE、iJGVD、HGVD、JSNP を追加した結果、さまざまな人類集団での SNP 頻度のデータや疾患との関連などの情報に容易にアクセスできるようになった。さらに、NCBI の作成するヒトの健康に関係するゲノム変異のデータベース ClinVar を追加し、ゲノム医科学分野の研究者への利便性を高めた。以上により、ゲノム多型の公共データベース dbSNP が提供する rs 番号を使った検索により、6 種類のゲノム多型データベースへの一括検索が可能になった。

また、リンク自動管理システムで大量データの高速な処理を実現するため、新たにデータ作成用のサーバ環境を構築し、さらにプログラムの大幅な改良・並列処理化を行った。これにより、以前のおよそ半分の時間でのデータ作成を実現した。さらに、公開サーバを産業技術総合研究所から東海大学へ移設した。

【テーマ２】論理ネットワーク上の芋づる式検索の開発、に関しては、各データベースの分子情報を Gene、Transcript、Protein、Structure、Disease などのカテゴリーに細分類するとともに、ID 間の論理的な関係の整理を行い、因果関係、対応関係、相関関係、結合関係（例えば薬剤と標的タンパク質の関係）、属性関係（例えば SNP は遺伝子の属性である）の 5 種類に分類した。これを基礎として、データ量の増加に対応しつつ、データ間の論理的な関係を利用したデータの芋づる式検索ソフトウエアの設計を行った。

【テーマ３】データ取得用ウェブサービスの開発、については、データ ID や URL を一括取得できるウェブサービスを開発し、公開した。REST による API を使用しており、その詳細はリンク自動管理システムのヘルプページで解説している（http://biodb.jp/help/ws_jp.html）。

【テーマ４】網羅的関連発見ツール（enrichment analysis ツール）の開発および【テーマ５】類似データ自動検索ツール（prioritization analysis ツール）の開発、については、システム設計は完了したものの、最終年度に発生した計算機故障の影響によりソフトウエア作成が大幅に遅れている。今後、これらのツール群の開発を進め、成果を順次公開していく計画である。

以上の研究成果については、学会発表等で精力的に発表を行い、普及に努めた。リンク自動管理システムは公開データベースとして継続的に運用しており、利用者の便宜を図っている。また、リンク自動管理システムはヒト遺伝子の統合データベース H-InvDB や、学術文献から抽出したヒトの遺伝的形質に関わるゲノム多型情報のデータベース VaDE においても、リンク作成のために利用している。本システムは、今後もできる限り継続して運用を行っていく計画である。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計５件）

①高橋定子、世良実穂、間宮健太郎、羽原拓哉、今西規、ヒト遺伝子多型の疾患リスク情報データベース VaDE の開発、第 38 回日本分子生物学会年会、2015/12/1-4、神戸国際会議場（兵庫県神戸市）

②Nagai Y, Takahashi Y, and Imanishi T. VaDE: a manually-curated database of reproducible associations between various traits and human genomic polymorphisms. ISMB/ECCB 2015, 2015/7/10-14, The Convention Center Dublin (Dublin, Ireland)

③今西規、小尾信男、リンク自動管理システムによるデータベース統合化とデータマイニング基盤の構築、第 37 回日本分子生物学会年会、2014/11/25-27、パシフィコ横浜（神奈川県横浜市）

④今西規、小尾信男、バイオデータベースの統合化とデータマイニングを推進するためのリンク自動管理システム、第 36 回日本分子生物学会年会、2013/12/3-6、神戸国際会議場（兵庫県神戸市）

⑤Imanishi T. Integration and data-mining of human transcriptome and proteome databases in H-InvDB. HUPO 2013, 2013/9/14-18, パシフィコ横浜（神奈川県横浜市）

〔その他〕
ホームページ等

リンク自動管理システム
http://biodb.jp/

６．研究組織
(1)研究代表者

今西　規（IMANISHI, Tadashi）
東海大学・医学部・教授
研究者番号：80270461