文科省統合データベースプロジェクト
ライフサイエンス分野の
統合データベース整備事業

データベース整備戦略のための研究俯瞰

I. データベース整備戦略のための研究俯瞰(分野を知りデータを知る)

基礎生命・情報・医学・創薬・育種など異なる領域の専門家および関係省庁の担当者に生命科学DB構築者を加えた運営委員会で戦略は議論されます。同会議に分野俯瞰を継続提供するために下記の開発および調査を行いました。

(1) 学会要旨統合サイト(国内研究俯瞰)

日本の研究俯瞰の重要な情報源として各種学会の過去の抄録を統合し、検索や施設別、テーマ別の再編成が可能なDB化を始めました。平成18年度は分子生物学会8年分の書誌事項に加え、一部要旨を打ち込み、施設名称など基本的な用語の統一を行いました。運営委員会用内部資料目的です。

●研究施設名称辞書 ●生物学名日本語一般名対応辞書 が使われています。

(2) データバンク統合目次(分子研究俯瞰)

分子DBは索引はあるが目次のない本のようです。分子レベルの生物学研究の俯瞰の提供を目的として各種データバンク内容を目次的に表現しようと試みています。またデータの利用機会はバンクの書式について熟知した一部の研究者と一般の研究者で大きな差があるようです。データバンクの内容につきバンクを区別せず自在に一次データを引き出し利用していただくことにも役立つと考えています。平成18年度はINSDCとGEOについて総合データ目次を作成しある程度データ内容について質的な表現を行いました。

1. DNAバンク(INSDC) 目次

DNA配列読み取りをおこなった論文では論文投稿時に国際DNA配列協力(INSDC)へのDNA配列の登録が義務付けられています。従って数十塩基の配列から完全なヒト染色体まで科学論文で新規に報告されたDNA配列は全てINSDCに登録されているはずです。DNAバンク目次は配列登録に至った研究の目次的な表現です。

2. 遺伝子発現バンク(GEO) 目次

NCBIが提供するGEO(Gene Expression Ominbus)はマイクロアレイやSAGEなど(遺伝子、サンプル、値)の三つ組みデータ一般に対するデータバンクです。進展の速い実験領域の1次データバンクは得てして利用者には難解です。少しでもデータが利用しやすいようにデータの整理パイプラインを作成しDNAバンク同様の目次を作りました。まだこれからの部分が多いですがどんな生物のどんな実験データが登録されて利用可能なのか目次でご覧いただけます。

●遺伝子名称シソーラス、●生物学名日本語一般名対応辞書、●動植物解剖学自動分類タガー、●都市名国名自動検出タガー、が使われています。

(3) 戦略立案資料(報告)

統合データベース間の連携と課題の整理

代表的モデル研究植物であり、全ゲノム塩基配列が決定済みである「イネ」ならびに「シロイヌナズナ」のゲノムアノテーション型公開データベース(それぞれ46, 25DB)について　1, データベースの種類と構造、2, 構成、3, webインターフェイス、4, アップデート頻度、5, 管理システム　等の基本項目を調査しました。

同時に、主として実験生物系のデータベースのユーザを対象に「これらのデータベースのなかでよく利用するサイトはどこか」「複数サイトを利用する場合に困っている点はないか」など、聴き取り調査と郵送によるアンケート調査(全188名対象)を実施しました。これによって、主として実験の現場でデータベースを活用している研究者が抱いているゲノムベースのデータベースの連携に関する現状の課題と、将来のデータベース統合にむけた要望を調べ上げました。

調査結果から、データベースのよりよい統合化は、以下のような比較的多数のユーザが抱く不満を解消する方向で行うべきであることを読み取ることができます。

- DB作成の時間差や異なる収集方針による遺伝子名やIDの相違が多く混乱の元になっている。これを吸収したり関連付ける基盤サービスが必要。
- 論文掲載情報やユーザからのフィードバックが直ちに反映されないことへの不満も大きい。
- 誤りの多さを不満とする声がある一方、仮想遺伝子にもなんらかのヒントが欲しいという要望も多い。提供するデータの分類や格付はできないか。
- 植物の分子の研究においては、頻繁に生物種横断的な検索や比較を行う。そのような情報を取得できるサイトがない。

反面、個別に現状のデータベースをみた場合の使いやすさや内容の充実度に関しては約半数が肯定的であり、将来の統合化データベースの作成にあたっては、現在利用頻度の高い個々のデータベースが保持している有用な情報を活かしつつ、齟齬を解消し連携させる形での統合化を考えていくべきであると考えられます。

(4) 検索アルゴリズムを含めた知識情報技術の動向調査

生物情報を扱うデータベースは、いろいろな分野の異なる観点から作成され、それぞれ異なる形式で記述されています。生物情報データベースに含まれるデータ量は計測技術の発展に伴い膨大な量となってきています。また、High Wire Press やPubMed Central をはじめとした文献の電子化・オープン化が進むことで、生物情報として利用可能なテキストや図表の量も増えつつあります。さらに、これら生物情報の利用方法自体、ユーザによって様々です。このような背景をもつ生物情報のデータベースを統合するためには、高度な知識情報技術の利用が不可欠です。そこで、次世代の生物情報データベース統合に必要となる知識情報技術として、検索システム、データマイニング、Web 2.0およびグリッドコンピューティングに焦点を絞り、聞き取り調査や文献調査によって動向を調べました。

検索システムについては、単純な項目検索やキーワード検索では、ライフサイエンス分野の、情報の膨大さ多様さゆえに対応が難しく、検索エンジン自体に高度な解析機能、可視化技術が必要になってきています。また、対象がグローバルなWWW上のデータへと広がったため、ローカルなデータベース内の構造化されたデータのみならず、WWW上の非構造化データをも扱える必要が出てきました。そこで、膨大かつ多様な情報へ対応する検索技術、および構造化データと非構造化データをともに扱う技術について調査しました。

データマイニングとは、大規模なデータやデータベースから隠れた関係性や知識などの情報を帰納的に抽出する技術を指す言葉です。データマイニング手法は出力される情報の方向性と入力されるデータの種類から、おおまかに第一世代と第二世代のものに分けることができますが、第二世代のデータマイニング手法には、ベイジアンネットワーク、隠れマルコフモデルなどの確率モデルや、グラフマイニングなどの構造データからのマイニング手法、さらに、テキストマイニングやストリームマイニングなどの新しいタイプのマイニング手法が含まれます。この調査では、第二世代のデータマイニング、特に構造データからのマイニング手法について調査を行いました。

Web 2.0とは従来のWWWにおける静的なサービスに対し、次世代にあるべき新しいウェブのあり方に関する総称です。Web2.0の特徴を持つタームとして、ここでは、web service、ロングテール、集合知、タグ付け、ブログについて調査し、さらに、これらとデータベースの関係について考察しました。 グリッドコンピューティングは、元々は遊休計算機資源を有効に活用するために作られた仕組みでしたが、現在は、大規模計算を効率よく行うための仕組みとして利用されています。このグリッドコンピューティングの目指す環境を実現するための様々な課題、例えば利用する計算機が別組織に属したり、そのプラットホームがばらばらであっても動的に連携できるようにすること等、を解決する必要があります。ここでは、こうした課題の解決策について調査しました。

これらの調査によって、従来の検索技術には情報の膨大さと多様さに基づく限界がすでにきており、データマイニング技術を検索エンジンへうまく組み込む必要があることが分かりました。また、グリッドコンピューティングを始めとした分散計算技術は、web serviceを前提としており、必要なweb serviceをデータベース側で揃えていく事が今後より重要となることが分かりました。

報告書
検索アルゴリズムを含めた知識情報技術の動向調査.pdf

(5) 臨床情報や医療統計の現状調査

臨床情報の調査に関しては、HL7等の標準規格の役割、データ抽出のためのカルテデータの電子化の状況やインセンティブ等につき、インタビューを含め調査を行いました。また、生活習慣病を中心とした我が国のコホート研究の事例を分類・整理しました。医療統計の調査に関しては、遺伝子多型解析に関わる遺伝統計学に焦点を絞り、遺伝統計学分野で用いられる解析技術に関して、インタビューを含め調査を行いました。

我が国のコホート研究については、循環器疾患を対象とした大迫(おおはさま)研究、生活習慣病その他の種々の疾患を対象とした山形大学の地域特性を生かした分子疫学研究(21世紀COEプログラム)、地域住民、大都市検診を対象とした多目的コホートによるがん・循環器疾患の疫学研究、広範な疾患を対象とした久山町研究、癌、循環器疾患を対象とした放射線影響研究所コホート研究、虚血性心疾患を対象とした都市勤労者集団コホート、高血圧を対象とした端野・壮瞥町研究、一般住民の循環器疾患を対象としたNIPPON DATA 80、全国各地で行われている循環器コホート研究の個人データを統計的に統合し、リスク因子を定量的評価することを目的としたJALS (Japan Arteriosclerosis Longitudinal Study) について、その研究の背景と目的、対象地域、ターゲット疾患、特徴的な検査項目、対象人数、代表研究者、研究開始時期、資金源等について調査しました。

医療統計の調査に関しては、遺伝統計学分野で用いられる連鎖解析、連鎖不平衡解析(ハプロタイプ解析)、QTL解析等の解析手法と、それぞれの手法における代表的なアルゴリズム計8種類の調査を行ない、その特徴および長所・短所を評価しました。併せて、各手法の代表的プログラム計15種類に関して、実装されているアルゴリズム、動作環境、入出力、利用形態、ダウンロード先などを調査し、その評価を行いました。また、代表的な商用ソフトの2種類の機能、特徴などを調査しました。

これらの調査から、医学データ活用における課題として、病名の標準化、前向きコホート研究の推進、人類遺伝学基盤の充実が重要との結果を得ました。病名の標準化に関しては、現状は死因統計などの主として保険行政統計用のものか保険診療用のものしかなく、臨床研究向けには使いにくく、抜けもあり、また必ずしも真の病名が記載されない、といった問題があります。前向きコホート研究については、既存のカルテの活用(後向き研究)は、検査値などを除くと難しく、しっかりデザインされた一定規模の前向き研究によって初めて有効なデータが得られることが分かりました。また、米国では家系を集めるプロジェクトにも多額の投資がなされているのに対して、日本では家系データが軽視される傾向にあり、これは日本における人類遺伝学基盤の不十分さに起因することが分かりました。

報告書
遺伝統計学分野における解析技術の基礎調査.pdf
臨床情報・疾患健康情報の調査.pdf