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

統合技術開発・提供

II. 統合技術開発・提供 （統合に足りない技術をつくる）

異種DBレコードの多くは遺伝子(蛋白質)ID、参考文献情報でリンクされ、また配列情報で相互にマップされて"統合"されています。一方でこれらのレコード中の注釈情報などの読むとわかる記述内容は現在信頼できる整理・統合が困難です。研究グループや基本概念での関連付けや整理を可能にするために用語シソーラスの作成および用語の構造化(いわゆるオントロジー)を進めました。これらの辞書は学術論文内容の全文無制限利用(オープンアクセス)時代を睨んだ技術であり、同様の目的で実験手法のリストアップとその構造化を行いました。

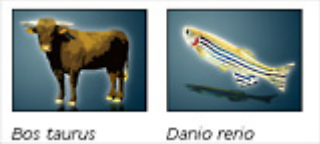
(1) 辞書シソーラス

異なるグループのデータを統合したり、自分なりに情報を整理したいときに必ず必要な辞書やシソーラス(多少階層のついた同義語辞書)を作ります。作成は手作業を基本に地道にバージョン更新を重ねてゆきます。全ての辞書はダウンロード利用可能です。

1. 遺伝子名称シソーラス

検索やDB統合の混乱の主因は同義語の流通です。同義語辞書の使用や配布で統合利用時の混乱を防止せねばなりません。

さまざまなデータベースや論文で利用されている表記を専門的キュレーターが編集して、遺伝子が持つ多様な表記のあいだの関係を明らかにした遺伝子名シソーラスを提供しています。Ver1.0 はヒトをはじめ9種類の生物をカバーしました。現在利用可能な最高の遺伝子名称辞書です。



2. 生物学名日本語一般名対応辞書

DBでは正確を期し、ラテン学名表記を基本としていますが利用者のほとんどは学名になじみがありません。塩基配列登録の多いものから順に学名：日本語一般名 (標準和名が存在しない場合、その生物を説明する一般的な名称)を対応させた日本語生物種辞書を作成しました。登録データ数は合計14028種となっています。主要な73種類についてはさらに認識を容易にするアイコンも付与しました。(2007年3月現在)

3. 施設名称辞書

学会要旨DB化でまず行き当たったのが施設名称研究室名称の表記ゆれです(大阪大、阪大、大阪大学大学院など)。同一の研究室の同一テーマを一塊として把握し国内の研究動向の把握を容易にすることは望まれるサービスの提供に欠かせません。

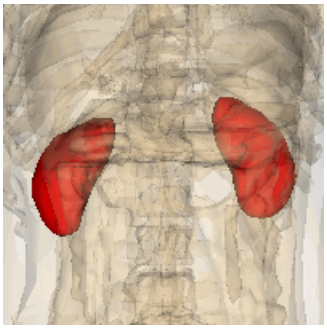
(2) オントロジー、分類機

1. 動植物解剖学自動分類タガー

動植物解剖学自動分類タガーは、解剖学用語、すなわち臓器・器官・部位の名称をカテゴリー分類するプログラムを作成しました。

2. 都市名国名自動検出タガー

論文やDBレコードの国別分類のための辞書を作成しました。かならずしも国名称を記載しなかったり、国名称にみられる表記ゆれを吸収します。

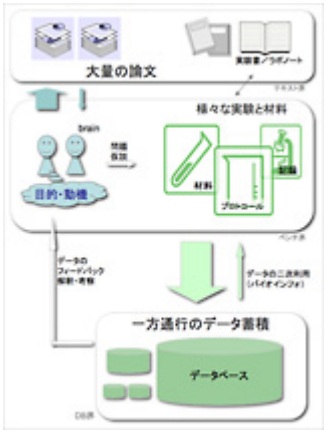


3. 3次元解剖学用語辞書(ポリゴンマン辞書)

解剖学用語、すなわち臓器・器官・部位の名称やそれらをくくる概念をモデル人間中の3次元座標で定義したユニークな辞書を作成しました。ポリゴンマン内の空間関係で用語関係を表現すると、ツリー型表現(いわゆる解剖オントロジー)と違い、多角的に破綻しない表現が可能です。ライフサイエンス辞書(京都大学)の英日対応解剖学用語のうち、PubMedアブストラクトでの出現の多い用語を中心に定義してゆき、2007年2月時点で約130語の定義が終わりました。ポリゴンマンの各臓器・器官の空間座標は、数値人体モデルデータベース(独立行政法人情報通信研究機構が、北里大学、慶應義塾大学及び東京都立大学と共同開発した電磁波影響計算のためのVoxelモデルファントム、分類組織数約50)を基盤に、位置関係や形態を大きく損なうことなく人体解剖模型・図譜等を参考に詳細部分を書き加えました。形態や関係は概念的正確を期していますが定量的正確さは犠牲にしています。また肉付きや顔貌は創作によるものです。

4. アナトモグラフィー

アナトモグラフィー(anatomography)は、anatomy(解剖学)と接尾語-graphy(画法)を組み合わせた新造語です。ユーザは解剖学用語が付与された手持ちのデータ(例:器官別の発現解析データ、疾患別症状分布など)を解剖学3Dポリゴンマン辞書にマッピングして俯瞰できます。膨大かつ詳細な「体に関する情報」どうしの関係を理解する際にgoogle earthのように真価を発揮すると期待し開発しました。マッピング結果は、静止画像もしくは動画で得られます。



5. ベンチメソッドオントロジー

日々増加するデータベースに登録される情報のほとんどはウェットの実験によって生み出されるものです。実験結果そのものでもあるデータベースを理解するためには、実験の目的、材料、手法や条件を理解せねばなりません。しかし最近では手法そのものが高度化、結果も膨大で解釈が難解になりつつあります。そこで本プロジェクトではベンチとデータベースを結ぶために、実験手法の整理を辞書とオントロジーによって行い、データベースや論文を実験手法と目的から分類することに着手しました。平成18年度は核酸に関する実験手法を収集しその分類を行いました。

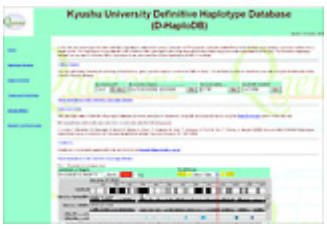
(3)キュレーター支援技術開発

オンラインで文献を読みながら重要な記述や画像、図表を集め、コメントとともに簡単にDB化でき、できたDBは自由にやり取りできるプラグインアプリケーション "ScrapParty"を開発しました。ダウンロード可能なフリーウエアです。

(4)癌研究知識表現技術開発

同じ臓器由来の癌でも進展の早さや薬剤への反応性など細胞としての特徴は様々です。ゲノム医学の分子解析ではこれら 1)臨床上の特徴を性別や年齢など 2)患者さんの背景および遺伝子発現やゲノム変化などの 3. 分子特徴と 関連付けて診断や治療の手がかりを得ようとしています。ところが癌サンプルに関する1)、2)の特徴はあらかじめどの項目が重要なのか知りえないことや カルテの読み取りや追跡調査が必要で簡単ではないことなど未決の課題が残されています。皮肉なことに現在試薬や器具さえあればすぐに集まる分子特徴は十分に集まり始めていますが、地道で体系的な努力の必要な1)、2)の情報を持つサンプルに由来するものはほとんど見られません。

癌標本の臨床記載を長く慎重におこなってきた施設のひとつ大阪府立先人病センターのグループが大阪大学及び京都大学の外科系研究室と共同で行った各種固形癌の遺伝子発現プロファイル解析の成果データベースCGEDに臨床上の特徴によっても解析可能になる部分を開発し有効な1)、2)の記載法について検討しました。



(5)多型知識表現技術開発

ヒトゲノム多型情報は疾患・体質等の遺伝的背景を解明するための必須の情報基盤です。特にわが国にとって日本人のゲノム多型情報の整備は緊急の課題です。現在最も有効とされる情報源は国際HapMap計画での日本人のゲノム多型(SNP)情報とされていますが、未だに調べられているSNP のゲノム上の密度が不十分です。また上記の遺伝的背景を明らかにするのに必要なハプロ

タイプ(SNP の並び方)の情報が正確ではありません。さらに同計画で調査された個体数が45人と少なく、日本人全体のゲノム情報を正確に反映したものとは言い難い状況にあります。そこで新たに日本人ゲノム多型情報に関する一次データを拡大・整備し、医科学等で利用可能な形態として提供することが、ゲノム多型情報を医学 情報と統合するために極めて重要な課題です。

1.「dbQSNP」の拡充

疾患の多くは遺伝子発現の量的異常に起因すると想定されています。我々はゲノムの遺伝子発現量制御に重要なゲノム領域の詳細な多型情報を定量的に記載し、遺伝子等の種々のゲノム情報と統合してグラフィカルに表示し、種々の検索が可能なデータベース「dbQSNP」を確立・公開しており、また拡充を続けています。現在、該当するゲノム領域に記載されているSNPの「dbQSNP」での密度は、国際HapMap計画データベースでの密度の2倍以上となっています。

2.「D-HaploDB」の拡充

我々は既にハプロタイプを誤り無く決定できる材料である胞状奇胎を用いて約3 x 10^5個 のSNPに関する日本人の確定ハプロタイプを決定し、種々の疾患要因遺伝子探索に必要な解析を行い、その結果、及び種々のゲノム情報を統合してグラフィカルに表示し、且つ種々の検索が可能なデータベース「D-HaploDB」を確立・公開しています。このデータベースに、新たに5 x 10^5個のSNPを用いた解析結果を追加しました。これによって高い確率で疾患要因遺伝子を検出可能な日本人の多型情報を整備しました。

3.「dbQSNP」及び「D-HaploDB」のXML化

上記2個のデータベースの情報を種々の外部データベースが取り込んで疾患・体質等の遺伝的背景を解析することに役立てるために、データベース記述標準言語XML(具体的にはゲノム多型記述のための機能拡張版であるPML)により記述したものを構築してウェブページからダウンロード可能としました。