National Bioscience Database Center

独立行政法人 科学技術振興機構 Japan Science and Technology Agency

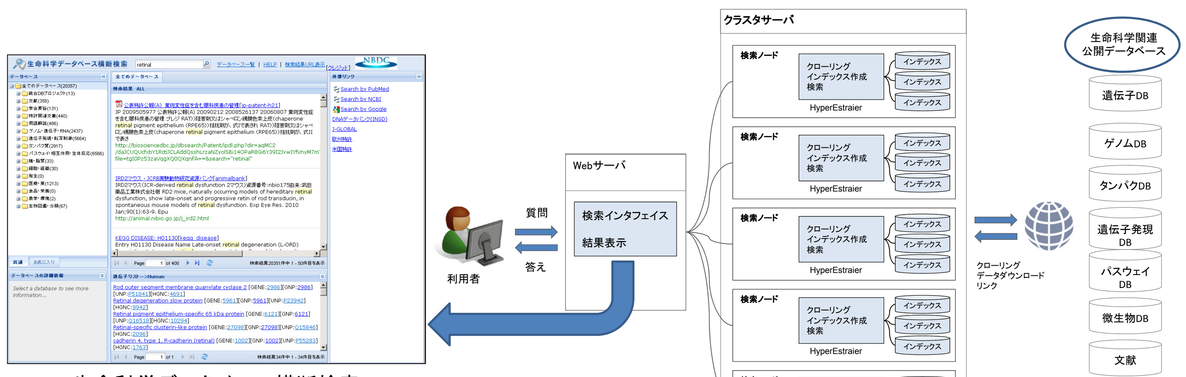
NBDCが公開するデータベース横断検索サービスの現状

○大波純一[1]、杉崎太一朗[2]、青木健一[2]、牧口大旭[2]、奥村利幸[2]、川本祥子[3]、畠中秀樹[1]、三橋信孝[1]

1.独立行政法人科学技術振興機構バイオサイエンスデータベースセンター（NBDC）、2.三井情報株式会社、
3.大学共同利用法人情報・システム研究機構ライフサイエンス統合データベースセンター（DBCLS）

2種類のデータベース横断検索

■多数のDBを分散配置する生命科学データベース横断検索システム



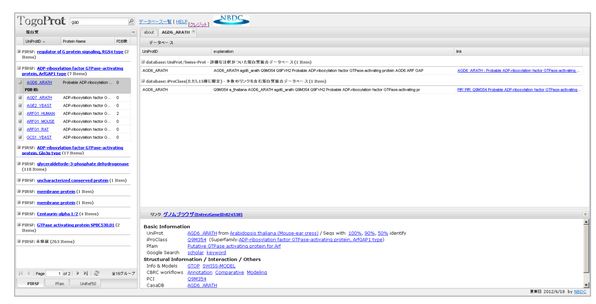
■生命科学データベース横断検索
http://biosciencedbc.jp/dbsearch/

■概要
・生命科学分野のコンテンツを対象にした検索システム
・分子や化合物のDBを文献や特許情報とあわせて一括検索可能
■特徴
・検索対象DB数：327件（右表参照）
・検索エンジンにはHyperEstraierを使用し、サーバやノードを分散配置したP2P連携が可能
・検索キーワードの日英相互翻訳（京都大学ライフサイエンス辞書）を搭載
・検索結果の絞込み機能　（DBのカテゴリ、ヒトの遺伝子名称など・・・）
・日本語コンテンツの充実
　蛋白質核酸酵素のバックナンバー（全文(1985年～2008年)、
　抄録（1985年～2010年））、学会要旨（日本生物物理学会）、
　日本国特許（公開広報、特許実用新案広報の全文（2004年～））
　文科省や経産省「統合データベースプロジェクト」の成果物など
■現在までの状況
・Integbioデータベースカタログとの連携
　→　Integbioデータベースカタログに掲載されるDBを追加し、
　　より網羅的な生命科学横断検索を目指す。
・省庁間システム連携の実現
　経済産業省産業技術総合研究所バイオメディシナル情報研究センター、
　経済産業省産業技術総合研究所糖鎖医工学研究センター、
　農林水産省農業生物資源研究所、厚生労働省医薬基盤研究所における、
　各種データベースとの連携を開始

生命科学データベース横断検索の対象DBカテゴリとその数

DBカテゴリ	DB数	主なDB
文献	29	蛋白質核酸酵素（共立出版）(1985～)、新着論文レビュー、文科省「ゲノム特定領域」報告書、各種実験プロトコル集
学会要旨	2	日本農芸化学会、日本生物物理学会
特許関連文書	9	日本国特許公報(2004～)
統合DBプロジェクト	4	生命科学DBアーカイブ、統合TV
用語解説	3	Gene Wiki、Proteopedia、Molecule of the Month
ゲノム・遺伝子・RNA	68	EntrezGene、RefSeq、H-Invitational、FANTOM
遺伝子発現・転写制御	29	CGED(がん組織発現)、DBTSS(転写開始部位)、coxpressdb(共発現)
タンパク質	54	UniProt、PIR、PDBj
パスウェイ・相互作用・生体反応	11	KEGG、ゲノムネットワークプロジェクト
糖・脂質	13	LipidBank、産総研糖鎖PJ
細胞・組織	7	SCMD(出芽酵母変異体顕微鏡写真)
発生	3	3D Microscopic Images of Kyoto Collection of Human Embryos
医療・薬	39	OMIM、MedlinePlus、PharmGKB
食品・栄養	9	食品成分DB、トクホなど機能性食品DB
農学・環境	12	イネ品種・特性DB、イネゲノムプロジェクト
生物図鑑・分類	14	水生生物情報ＤＢ，放線菌図鑑、牧野標本館所蔵標本ＤＢ

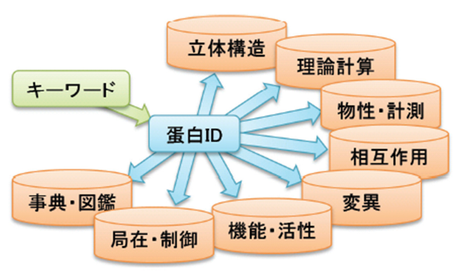
■蛋白IDが各種DBをつなぐTogoProtシステム



TogoProtの対象DBとその詳細

TogoProt対象DB	オリジナルサイトURL	DBの説明
3DinSight	http://gibk26.bse.kyutech.ac.jp/jouhou/3dinsight/3DinSight.html	生体分子の立体構造・性質・機能に関する統合データベース
Biological Magnetic Resonance Data Bank (BMRB)	http://bmrb.protein.osaka-u.ac.jp/	生体高分子のNMRデータのデータベース
ConfC	http://mbs.cbrc.jp/ConfC/	蛋白質構造変化部位データベース
DB-SPIRE	http://mbs.cbrc.jp/DB-SPIRE/	PDBエントリに含まれる配列モチーフのデータベース
eF-site	http://ef-site.hgc.jp/eF-site/	静電ポテンシャルなど分子表面物性のデータベース
EM Navigater	http://emnavi.protein.osaka-u.ac.jp/	生体分子や生体組織の3次元電子顕微鏡データ
eProtS	http://eprots.pdbj.org/jp.cgi	蛋白質構造百科事典
eSOL	http://tp-esol.genes.nig.ac.jp/	大腸菌の全蛋白質を無細胞系で合成した際の可溶率と合成量のデータベース
EzCatDB	http://mbs.cbrc.jp/EzCatDB/	酵素の立体構造等に基づいた触媒機構分類データベース
Gallery of Biomolecules	http://gibk26.bse.kyutech.ac.jp/jouhou/image/gallery.html	分子構造の画像データベース
Gene Wiki	http://en.wikipedia.org/wiki/Portal:Gene_Wiki	Wikipedia中の遺伝子・蛋白質ページ
Genome Network Project	http://genomenetwork.nig.ac.jp/	発現プロファイルや蛋白質相互作用などをまとめたヒト遺伝子統合データベース
GLIDA	http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/	GPCR-リガンド相互作用データベース
Glycoepitope Antibody	http://www.glyco.is.ritsumei.ac.jp/epitope/	糖鎖抗原データベース
HATODAS2	http://hatodas.harima.riken.go.jp/hatodas_V2/query/queryheavy.jsp	蛋白質の重原子化実験データベース
Hepatitis Virus Database (HVDB)	http://s2as02.genes.nig.ac.jp/	肝炎ウイルスの配列データベース
Het-PDB Navi.	http://hetpdbnavi.nagahama-i-bio.ac.jp/index.php	PDB中の低分子-高分子相互作用データベース
Hintdb	http://hintdb.hgc.jp/hint/	実験的に得られた蛋白間相互作用のデータベース
HitPredict	http://hintdb.hgc.jp/htp/	ハイスループットデータから予測された蛋白間相互作用のデータベース
KEGG GENES	http://www.genome.ad.jp/dbget/	ゲノムが解読された生物の遺伝子カタログ
Molecule of the Month	http://www.pdbj.org/mom/	Goodsell博士による毎月のPDB中の立体構造の解説
OLIGAMI	http://protein.t.soka.ac.jp/oligami/	蛋白質のオリゴマー構造のデータベース
PDBj	http://www.pdbj.org/index_j.html	日本蛋白質構造データバンク
PGE	http://www.bioa.jst.go.jp/	マウス肛の遺伝子発現パターンデータベース
PIR	http://pir.georgetown.edu/	多数のリンクを含む蛋白質統合データベース
ProMode	http://cube.socs.waseda.ac.jp/pages/jsp/ja/index.jsp	蛋白質の基準振動解析データベース
ProNIT	http://dna01.bse.kyutech.ac.jp/jouhou/pronit/pronit.html	蛋白質-核酸相互作用の熱力学データベース
ProNuc	http://gibk26.bse.kyutech.ac.jp/jouhou/pronuc/pronuc.html	蛋白質-核酸複合体の構造データベース
Protein Globe	http://www.pdbj.org/globe/	球面上の距離で構造類似性を表現した蛋白質の世界
Protein Mutant Database	http://spock.genes.nig.ac.jp/~pmd/pmd.html	蛋白質変異データベース
Protein Quaternary Structure	http://www.ebi.ac.uk/pdbe/pqs/	蛋白質の4次構造データベース
Proteopedia	http://www.proteopedia.org/	蛋白質などの分子の3次元百科事典
ProTherm	http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html	蛋白質と変異体の熱力学データベース
Rice Proteome Database (RiceProDB)	http://gene64.dna.affrc.go.jp/RPD/	イネの組織やオルガネラの2次元電気泳動
イネタンパク質構造データベース（RPSD）	http://structure.rice.dna.affrc.go.jp/	イネ蛋白質構造データベース
Two-Dimensional Electrophoresis Database of HSG cells proteins	http://hitech-d.iwate-med.ac.jp/biochem/biochem/kamo/index.html	ヒト唾液腺癌細胞蛋白質の2次元電気泳動
UniProt	http://www.uniprot.org/	詳細な注釈がついた蛋白質統合データベース
VIPERdb	http://viperdb.scripps.edu/	正二十面体型ウイルスの構造データベース
万見(yorodumi)	http://pdbj.protein.osaka-u.ac.jp/emnavi/viewtop.php	PDB中電顕の3次元構造を気軽に楽しむウェブページ
PDBSTR	http://www.genome.jp/dbget-bin/www_bfind?pdbstr	PDB中の蛋白質・核酸を鎖ごとに編成しなおした配列・立体構造データベース
PDBWiki	http://pdbwiki.org/wiki/Main_Page	コミュニティによるPDBエントリへの注釈の知識ベース
TPプレスリリース(tppress)	http://www.tanpaku.org/news_release/	ターゲット・タンパク研究で発表されたプレスリリース一覧
TP構造ギャラリー(tpgallery)	http://www.tanpaku.org/tp_gallery/	ターゲット・タンパク研究で取り組まれている現たんぱく質など307の構造データベース
TP Atlas	http://net.genes.nig.ac.jp/tpatlas/	ターゲット・タンパク研究プログラムの各課題の成果を、Cell Illustratorを使用したネットワーク図をベースにしてとりまとめたデータベース
PREIMS	https://preims.tanpaku.org/preims/index.jsp	ターゲット・タンパク研究で開発された実験プロトコルを整理したデータベース
AT Atlas	http://p4d-www.genes.nig.ac.jp/at_atlas/index.html	実験方法をCell Illustratorなどによる概要図と技術の概要や活用事例のテキストを組み合わせて、分かりやすく表現することを目指した枠組み
Autophagy Database	http://tp-apg.genes.nig.ac.jp/autophagy/	オートファジー関連蛋白やそのホモログの構造や機能、文献情報

TogoProt ■TogoProt（蛋白質関連DB統合検索）システム
http://biosciencedbc.jp/togoprot/index.php

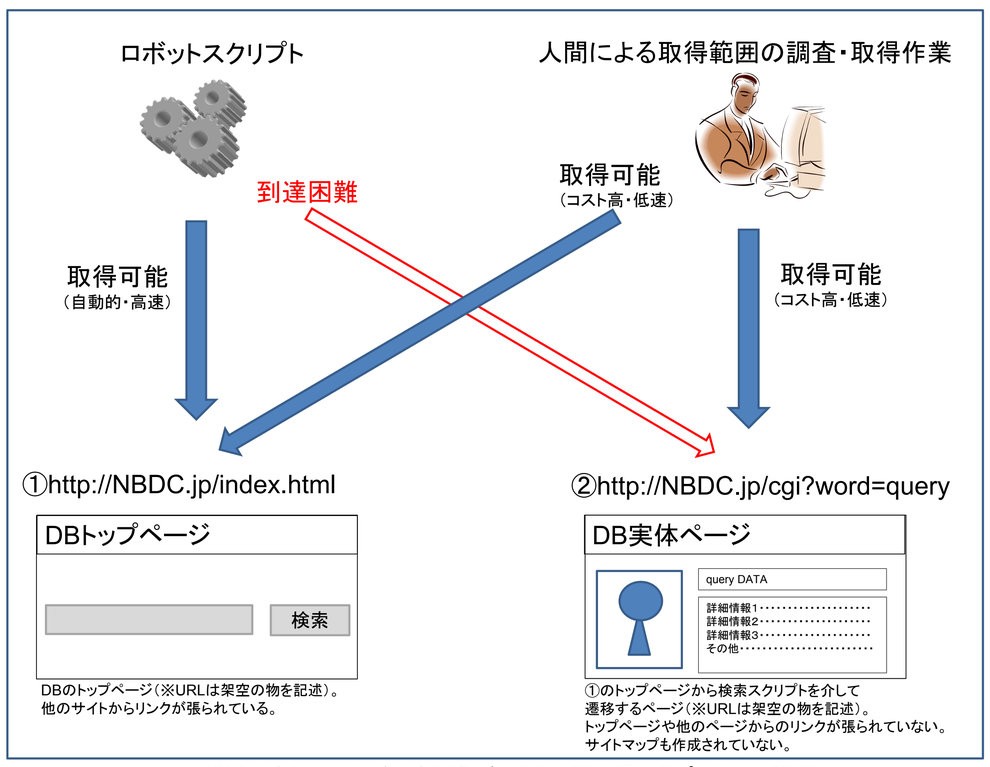


TogoProtの検索概念図

■概要
・散在する多数の蛋白質関連データベースの情報を蛋白質ごとにまとめて一覧できるシステム
■特徴
・検索対象DB数：47件（右上表参照）
・検索エンジンにはHyperEstraierを使用し、サーバやノードを分散配置したP2P連携が可能
・各データベースの蛋白質のIDを検索インデックスの属性として持ち、ウェブブラウザの
　検索インターフェイス上に蛋白質IDによる一括した表示を可能にしている。
・各種蛋白質ファミリーのIDやPDBIDもインデックスの属性として持つので、
　ファミリーで一括した表示、特定の立体構造に限定した表示も可能である。
■現在までの状況
・DBの追加
・データ更新自動化

データベース横断検索サービス公開の意義と、Web検索サービスとの比較

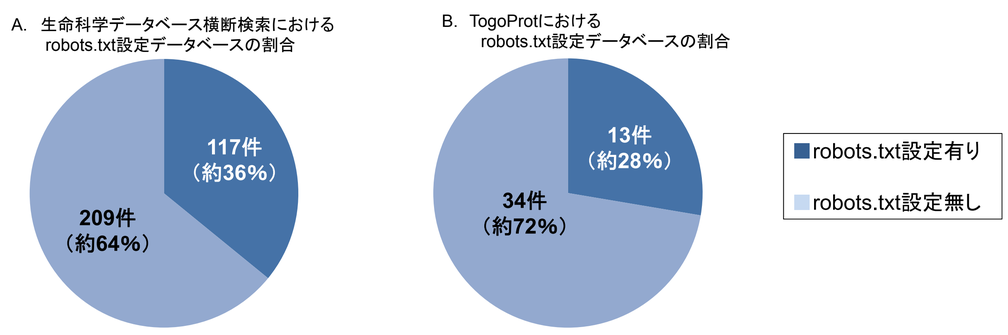
■検索対象データの自動取得に関する問題と対策



横断検索システムが検索対象データを自動取得するプロセスの種類

ネットワーク上をロボットスクリプトがリンクを辿って自動的に巡回し、検索対象のデータをインデックス化する方式は、低コストで高速だが、深層Webのデータ（他のページから直接リンクされていない）は取得が困難である（上図参照）。
公開されている学術情報データベースはこのような深層Webに配置している物も多いため、DBごとにWebのどこに設置しているかを見分ける必要がある。また、ページタイトルやスニペット表示もDBごとに人間が設定することで、利用者が検索結果から目的の結果を取得しやすくなる。
このため、NBDCの横断検索ではデータベースインデックスをオリジナルサイトから取得する際には、Web検索サービスやDB横断検索サイトで一般的に利用されている自動クローリングロボットの他に、人間による取得範囲の調査・取得作業を併用し、検索結果の網羅性と品質を保っている。

■データベースがロボットスクリプト向けに公開する情報



各横断検索サービスの中で、robots.txtを規定し、検索ロボットスクリプトへ検索対象を指示している割合を示す。robots.txtファイルはデータベースサイトごとに設置され、サイト内のどこにデータが存在するかを記述するサイトマップ、データをロボットで自動取得してもよいか指定する「allow」および「disallow」パラメータから構成されている。横断検索とTogoProtでrobots.txtを設置しているサイトはそれぞれ36%、28%程度であり、ロボットスクリプトへ積極的に情報を開示しているDBは多くはない。
今後、DBサイト管理者ごとにrobots.txtやサイトマップ情報を設置することで、ロボットスクリプトでのデータ取得が容易になれば、横断検索サービス管理の効率化につながると考えられる。

■特定分野用語の検索方法について

生命科学分野と一般分野で複数の意味がある単語の例

検索語句	生命科学分野における意味	一般的なWeb検索エンジンで優先的に表示される意味
マウス	生物名	情報端末の入力デバイス
プロテイン	タンパク質	タンパク質を主成分とするサプリメント
翻訳	mRNAからタンパク質が生合成されること	ある言語で表された文章を他の言語に置き換えて表すこと
ボトルネック	生物集団における遺伝的多様性の減少	電子情報や各種トラフィックの流れにくい部分

検索語が複数の意味を持つ場合、一般的なWeb検索エンジンでは片方の意味の検索結果が優先されて表示され、利用者の要求が満たされないことがある（表参照）。 このような場合、利用者自身が検索語を複数入力し再検索を行うか、検索結果一覧から必要な結果を取捨選択する必要がある。
このような利用者の想像と検索結果の不一致をなくすため、検索対象DBの大部分を、Integbioデータベースカタログに記載された生命科学系、タンパク質系DBに限定した。このために、生命科学関連に絞った信頼度の高いDBの検索結果を迅速に入手できる。
またクエリ変換にはライフサイエンス辞書を採用し、精度の高い語句変換（上表の各単語の変換や、「遺伝子」→「gene」「偽遺伝子」→「pseudogene」等）を利用できる。

今後の展望

NBDCより公開している生命科学系データベース横断検索およびTogoProtは、一般的なWeb検索サービスと比較し、高い網羅性と生命科学系情報に特化した検索精度を備えている。
今後は利便性と運用効率向上のため、対象データベースの範囲を広げ、検索対象データベース管理者ともrobots.txt（サイトマップ）の情報公開など相互に協力し、さらに利用者からの意見も積極的に取り入れていく方針である。

トーゴーの日シンポジウム 2012　ライフサイエンスデータベース統合の医学への応用を探る

会期：2012 年 10 月 5 日（金）　於：時事通信ホール

Licensed under a Creative Commons 表示2.1 日本 license (c)2012 バイオサイエンスデータベースセンター