

平成 21 年度産業技術研究開発

「統合データベースプロジェクト」 に関する成果報告書

平成22年3月

(委託先) 社団法人 バイオ産業情報化コンソーシアム

目次

1. 総括	1
2. 研究開発成果	2
2.1 経済産業省ライフサイエンスデータベース・ポータルサイトの構築と運営	2
2.1.1 成果物便覧	2
2.1.2 成果物アーカイブの公開	7
2.1.3 横断検索システム	10
2.1.4 ポータルサイト運用	19
2.1.5 成果の追加(PubMedScan:研究ツールの開発)	27
2.2 経済産業省関連機関から産生されるデータベース等の統合	29
2.2.1 ウェブサービスの開発	29
2.2.2 ウェブサービスの拡張	33
2.2.3 データ品質の評価	33
2.2.4 画面の改良	37
2.2.5 分子情報と実験用リソース情報の統合	39
2.3 国内外の最新動向調査及びアンケート調査	42
2.3.1 国内外の最新動向調査と広報・普及活動	43
2.3.2 ユーザーのニーズ調査	47
2.4 他省庁との連携	51
2.4.1 文部科学省ライフサイエンス統合データベースセンターDBCLSとの連携	51
2.4.2 ウェブサービスの統合化	52
2.5 運営委員会	56
3. 成果発表	57
4. 謝辞	59
5. 参加者名簿	60

1. 総括

2008 年 4 月から開始された「経済産業省統合データベースプロジェクト」(2008-2011 予定)は、平成 20 年度に続き平成 21 年度も社団法人バイオ産業情報化コンソーシアム(JBiC)と独立行政法人産業技術総合研究所(産総研)・バイオメディカル情報研究センター(BIRC)の共同チームが事業を実施した。本プロジェクトの目的は第一に、経済産業省関連機関により実施されたライフサイエンス分野の研究開発プロジェクトの成果であるデータベースに関する情報提供サイトを作成することである。第二に、ヒト遺伝子に関連した各種研究成果に関して、ゲノム情報統合プロジェクトにおいて構築したヒト遺伝子の統合データベース H-InvDB を基礎として、経済産業省関連の研究成果を連携して利用できるシステムを構築することである。以上を通じて、これまでに経済産業省関連機関により実施された研究開発プロジェクトの成果等を整備し、ライフサイエンス分野における研究開発の促進に資するデータベースを構築することを目的としている。

平成21年度の主な成果を挙げる。まず、前年度に引き続きポータルサイト MEDALS(<http://medals.jp/>)の公開を継続し、またその内容の大幅な充実を図った。特に、経済産業省関連の各種研究開発プロジェクトで作成されたデータベースやソフトウェアの便覧については、詳細な調査の結果、合計で 100 件以上の成果物をリストに載せることができた。その結果、便覧としての有用性が飛躍的に高まった。同時に、成果物を生み出した各種のプロジェクトに関する情報も、新たに「プロジェクト便覧」として公開した。これにより、経済産業省関連の研究開発動向が理解しやすくなったと考えられる。以上の便覧類は英語版も作成しており、成果物の海外からの利用の増加にもつながると期待される。一方、経済産業省プロジェクトの成果物の利用促進のための各種サービスの提供については、リンク自動管理システムや PubMedScan などの有用な研究支援ツールの機能強化を行った。その結果、これらのツール群は順調に利用者数が増加しており、データや情報のより効率的な流通に貢献することができた。また、他省庁にて実施されている生命科学分野の統合データベースプロジェクトとの連携が前進したことも、今年度の大きな成果のひとつである。特に、文部科学省ライフサイエンス統合データベースプロジェクトとの間では、データベース便覧の共通化やデータベース横断検索の共同開発などの成果を挙げることができた。

さらに、ヒトの分子データの実質的な統合についても多くの進展があった。ヒト全遺伝子のアノテーション統合データベース H-InvDB と産総研・糖鎖医工学研究センターで測定されたタンパク質の糖鎖修飾の情報を集めた糖タンパク質データベース(GlycoProtein DataBase (GPDB) :http://riodb.ibase.aist.go.jp/rcmg/glycodb/Glc_ResultSearch)を連携させるため、双方のシステムに共通仕様のウェブサービスを開発して導入し、ユーザーが双方の情報をひとつの画面の中で閲覧することができるしくみを実現した。また、H-InvDB の新しいビューア Protein view を開発し、GPDB 等の情報を表示する機能を開発した。このほか、独立行政法人製品評価基盤機構(NITE)・生物遺伝資源部門:NITE Biological Resource Center (NBRC) にて公開しているヒト cDNA クローンの情報を、リンク自動管理システムを経由することにより、H-InvDB とリンクさせた。このようにデータベース検索と実験用のリソースの取得をスムーズに連結したことにより、研究開発を加速させることが期待される。以上の成果は利用者の利便性を高め、データベースからの知識の取得を促進・効率化すると期待される。

このほかにも、産業界からのニーズをウェブによるアンケートや面談によるインタビューにより調査したこと、データベース講習会を開催するなどの広報活動を積極的に行いマスメディア等にも取り上げられたこと、論文発表や学会発表を行ったことなど、多くの成果を挙げることができた。今後も利用者となる産業界等の研究者の意見を反映させつつ、本プロジェクトを継続的に実施していくことが、きわめて重要であると考えられる。

経済産業省統合データベースプロジェクト・プロジェクトリーダー 五條 堀 孝
サブリーダー 今西 規
サブリーダー 村上 勝彦

2. 研究開発成果

2.1 経済産業省ライフサイエンスデータベース・ポータルサイトの構築と運営

● 背景

ライフサイエンス分野では、自身の研究成果を既に蓄積されている公知の研究成果情報や他の研究データと対比することにより、自身の研究成果の仮説を考案する手がかりを得たり、その検証を行ったりすることが極めて重要である。それによって例えば効果的な治療薬など新しい実用化の発想が得られる可能性がある。このため、国家プロジェクト等により産生された研究データを基盤データベースとして一括して活用できるデータベースが、産業界や社会から必須の知的基盤として強く要望されている。

● 目的

本事業では、経済産業省関連機関により実施されたライフサイエンス分野の研究開発プロジェクトの成果であるデータベースに関する情報提供サイトを作成し運用する。またヒト遺伝子に関連した各種研究成果に関しては、平成 17～19 年度に実施したゲノム情報統合プロジェクトにおいて構築したヒト全遺伝子のアノテーション統合データベース(H-Invitational Database : <http://www.h-invitational.jp/>)を基礎として、経済産業省関連の研究成果を連携して利用できるシステムを構築する。

以上を通じて、これまでに経済産業省関連機関で実施された研究開発プロジェクトの成果等を整備することにより、ライフサイエンス分野における研究開発の促進に資するデータベースを構築することを目的とする。

● 概要

前年度の事業「平成 20 年度産業技術研究開発(統合データベースプロジェクト)」において、経済産業省関連のライフサイエンスデータベースのうち、独立行政法人産業技術総合研究所(産総研)と独立行政法人新エネルギー・産業技術総合開発機構(NEDO)の一部の成果が成果物便覧に格納された(http://medals.jp/list/list_alldb、http://medals.jp/list/list_alltool)。今年度は、すでに便覧に掲載されている 71 件のデータベース詳細情報の内容の更新、改良を行い、経済産業省ライフサイエンスデータベースポータルサイト(<http://medals.jp/>)から公開した。具体的には、「各データベースの更新により便覧内容が古くなっている」、「表記の統一がとれていない」、「調査が十分でない」などの場合に内容(項目や文言等)を更新・改良した。

また、NEDO の未格納分や独立行政法人製品評価技術基盤機構(NITE)によるライフサイエンス系成果物の便覧を作成して追加公開した。

さらに、文部科学省の統合データベースプロジェクトから公開されている便覧と整合性を図るため、説明項目の統一や同一の説明項目内の文言一致などを協議し、連携を図った。

2.1.1 成果物便覧

(1) 成果物便覧に既掲載分の改良

前年度事業「平成 20 年度産業技術研究開発(統合データベースプロジェクト)」において、経済産業省関連のライフサイエンスデータベースのうち、独立行政法人産業技術総合研究所(産総研)と独立行政法人新エネルギー・産業技術総合開発機構(NEDO)の一部の成果(71 件)が成果物便覧に格納された。平成 21 年度は、すでに便覧に掲載されている上記 71 件のデータベース詳細情報の内容について改良を検討した。その結果、以下の点を改良した。

- 便覧中の解説文、英語について外部委託をして遂行した結果を、チームで再度検討して反映した。
- 便覧中の解説文が、オリジナルサイトに書かれた文章であるか、または本プロジェクトで書かれた文章であるか、翻訳して書かれた文章であるかを「オリジナルサイトから」「～より引用」「～より翻訳」などによって明記した。

- 関連する重要なキーワードを新たに検討して便覧に付与し、Google などの検索でより多くのケースでヒットするようにした。

- データベースの更新頻度情報として何を載せるかを再検討し、その書式を全成果物で統一した。

データベースを調査した時点において、各オリジナルサイトでデータが更新された日付と、我々が調査した日付を便覧に記述してきた。しかし、その後数ヶ月たってしまうと、オリジナルサイトで更新されたとしても、それが便覧に反映されないケースが出てくる可能性がある。さらに、更新頻度の書式は、「年に1回」「毎日」「不定期」等、必ずしも統一的ではなかった。また、更新頻度については、第三者にはわからないものが多く、各データベースの開発者でも毎日自動で予定しているものから予定がたっていないものまで様々であり、頻度が書かれたとしても、それは確定的ではないので、これを今後の頻度であると記述するのは適切でない。これらを解決するため、2つの工夫を行った。第1点は、2. 1. 4節(2)に記述している「情報管理ツール」を開発し、現在進行中の各データベースの更新日を迅速かつ効率的に把握するようにしたことである。第2点は、「更新頻度」でなく「更新実績」を載せ、その情報を各データベースで公開している更新履歴や情報管理ツールによって把握し、「調査日以前の2年間ににおける更新頻度の実績」という内容に統一したことである。

- 文部科学省統合データベースプロジェクト(DBCLS)から公開されている便覧と整合性を図るため、連携を図り、説明項目の統一や同一の説明項目内の文言一致などを協議検討した。

その結果、MEDALS から公開している 30 項目のうち、14 項目を一致させることにした。表 2.1.1-1 に共通とした項目の一部を示す。その他の 11 項目は MEDALS 独自の項目とした。例えば以下のものである。

- ✓ ダウンロード総量(Mbyte)、データ一括取得方法
- ✓ 入力例
- ✓ キーワード
- ✓ 使っている外部リソース
- ✓ データ更新頻度
- ✓ 関連プロジェクト

表 2.1.1-1 共通とした項目の一部

MEDALS	DBCLS (LSDB カタログ)
成果物名→名称	タイトル→名称
成果物の別名→別名	タイトル→別名
成果物に関する説明→説明	説明
運用機関	組織→運用機関
サイト URL→オリジナルサイト	URL→オリジナルサイト
機関所在国	機関所在国
論文等	論文等

(2) 成果物の存在・公開状態の調査

成果物便覧に掲載してあるもの以外で他の成果物がないか、各プロジェクトの成果報告書をもとに、それらの存在・公開状態の調査を行った。主な候補は、新エネルギー・産業技術総合開発機構(NEDO)の未格納分や製品評価技術基盤機構(NITE)によるライフサイエンス系成果物である。成果物の調査から、ポータルサイト MEDALS の成果物便覧に載るまでの状態変化を示したものが、図 2.1.1-1 である。経済産業省の関連機関や Web 上の情報を元に、ライフサイエンスのデータベース(ツールも含む)として適切と判断したものを「選定」した状態とする。調査結果をもとに、各プロジェクトのリーダーや担当の方にコンタクトをとって、どのような具体的な成果物(データベース)があったかを問い合わせる。ここで「回答待ち」の状態になる。回答が明確になった場合、「公開可能(公開している、公開予定である)」または「非公開(製品として商業利用している等)」に分かれる。その他としては、個別の担当者にコンタクトがとれないものや、途中結果にすぎないので公開のレベルでない、担当者が不明ですぐには出てこないものなどがあつた。これらは「非優先(不明)」のカテゴリーとした。「公開可能」となったものは、逐次さらに詳細を調べて便覧に載せていった。なお、「公開可能」とされたものが、「アーカイブ」や「横断検索」という形での活用(2.1.2 項及び 2.1.3 項参照)が検討されることになる。

上記の過程を経て、成果物の件数がどのように変化したかを図 2.1.1-2 に示す。左から 2009 年 2 月、2009 年 9 月、2010 年 2 月時点で、各カテゴリーにいくつの成果物があったかを数値で示している。この各状態は、先の図 2.1.1-1 に示したものと同一(①から⑥)である。今年度の成果としては、まずすべての成果物の関係者にコンタクトが完了した(①の「選定」までしか進んでいないものは 0 件)。報告書では 1 つのシステムにみえたものが、実際には複数あるケースなどのため、成果物の数は 204 件に増加した。公開可能な件数は 146 件となった。そして、前年度の便覧掲載件数は 71 件であったが、2009 年 9 月までに 81 件、さらに 2010 年 2 月には 100 件に達した(図 2.1.1-2、2010 年 2 月時点)。さらに、2010 年 3 月末には、118 件になる予定である。

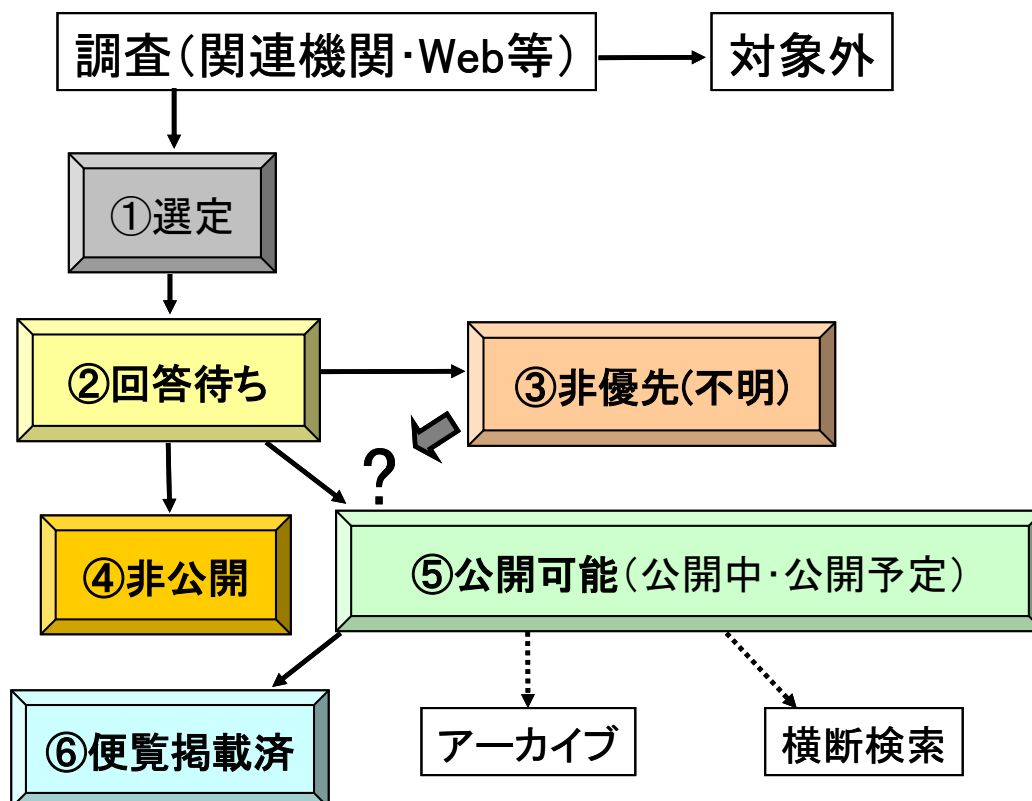


図 2.1.1-1 成果物候補の選定から便覧掲載までの状態遷移

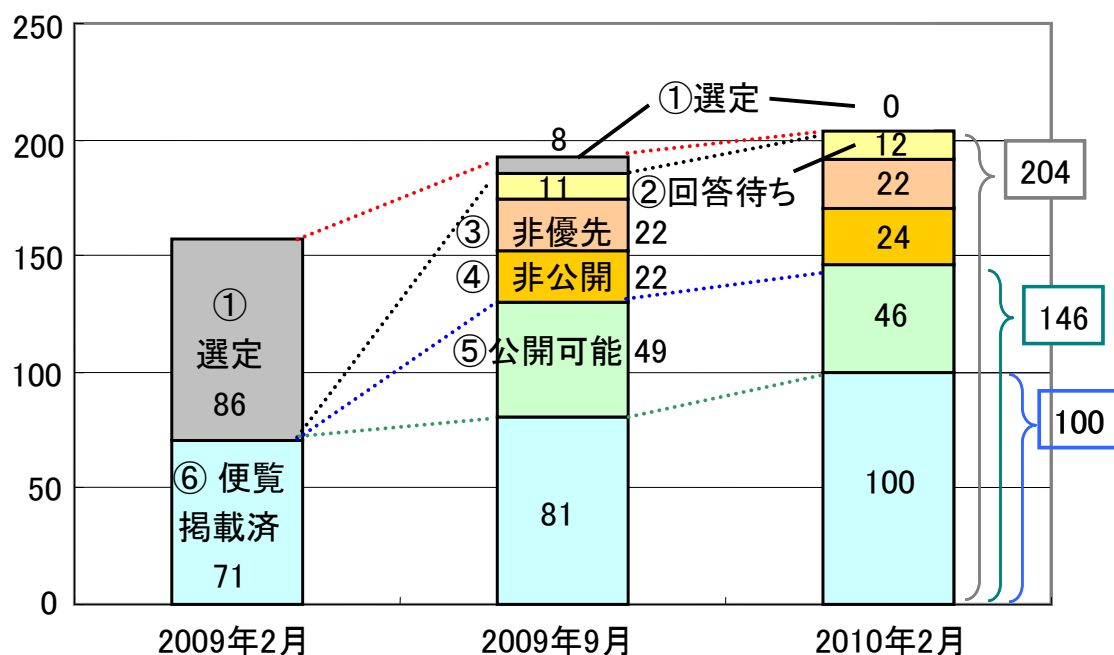


図 2.1.1-2 成果物の発掘件数と便覧掲載数の増加

(3) 成果物の公開状態の調査例

新エネルギー・産業技術総合開発機構 (NEDO) プロジェクトの未格納分や製品評価技術基盤機構 (NITE) によるライフサイエンス系成果物の便覧を作成して追加公開した (表 2.1.1-2)。各便覧作成では内部で調査・作成を行った後に、連絡が可能な開発担当者・現運用担当者へ便覧内容の確認依頼を出し、便覧の精度向上を図った。

表 2.1.1-2 成果物と便覧公開状況

項番	名称	便覧公開状況	運用機関
1	Kappa-view	公開済	かずさ DNA 研究所
2	モデル植物ミヤコグサ完全長 cDNA データベース	公開予定	かずさ DNA 研究所
3	ミヤコグサ EST	公開予定	かずさ DNA 研究所
4	アカシア EST	公開予定	かずさ DNA 研究所
5	カンゾウ EST	公開予定	かずさ DNA 研究所
6	トチュウ EST	公開予定	かずさ DNA 研究所
7	ユーフォルビア EST	公開予定	かずさ DNA 研究所
8	パラゴム EST	公開予定	かずさ DNA 研究所
9	ユーカリ EST	公開予定	かずさ DNA 研究所
10	タカラバイオ DNA アレイ設計配列 データベース (ARTRA)	公開予定	かずさ DNA 研究所
11	遺伝子発現相関計算ツール (SokanProject)	公開予定	かずさ DNA 研究所
12	植物遺伝子発現相関データベース (CoP)	公開済	かずさ DNA 研究所
13	シロイヌナズナ培養細胞過剰発現系統 データベース (M87 Array Viewer)	公開予定	かずさ DNA 研究所
14	シロイヌナズナ遺伝子情報データベース (KATANA)	公開済	かずさ DNA 研究所

15	シロイヌナズナ遺伝子 オントロジービューワー(DAGViz)	公開済	かずさ DNA 研究所
16	シロイヌナズナ培養細胞制御 ネットワークデータベース(RnR)	公開済	かずさ DNA 研究所
17	遺伝子、代謝統合データベース(Komics)	公開済	かずさ DNA 研究所
18	質量分析データベース(MassBase)	公開済	かずさ DNA 研究所
19	質量分析 MS/MS データベース (MS-MS Fragment Viewer)	公開予定	かずさ DNA 研究所
20	GC-MS 質量分析ピーク解析ツール (Metabolometrics)	公開予定	かずさ DNA 研究所
21	CE-MS 質量分析ピーク解析ツール (PeakIdentifier)	公開予定	かずさ DNA 研究所
22	微生物産業利用支援データベース(DIAM)	公開済	(財)バイオインダストリー協会 (JBA)・NITE
23	DOGAN	公開済	NITE
24	NBRC	公開済	NITE
25	化学物質統合検索システム	公開済	NITE
26	PRTR 制度対象物質データベース	公開済	NITE
27	siExplorer	公開済	東京大学
28	GenoBase	公開済	奈良先端科学技術大学院大学
29	MassBank	公開済	慶応義塾大学
30	CADLIVE	公開済	九州工業大学
31	予測プログラム	公開済	広島大学
32	組織別遺伝子発現データベース 癌組織遺伝子発現データベース	公開済	東京大学
33	大腸菌マルチオミクス データベース	公開済	慶応義塾大学
34	Dr.DMASS	公開済	奈良先端科学技術大学院大学
35	Dr.DMASS+	公開済	奈良先端科学技術大学院大学
36	BL-SOM	公開予定	奈良先端科学技術大学院大学
37	KNAPsAck Family	公開済	奈良先端科学技術大学院大学
38	KNAPsAck	公開済	奈良先端科学技術大学院大学
39	DPclus	公開済	奈良先端科学技術大学院大学
40	TREBAX	公開済	奈良先端科学技術大学院大学
41	FLJ Human cDNA Database	公開済	産業技術総合研究所/DBCLS
42	OpenPML	公開済	Object Management Group (OMG)
43	JSNP	公開済	東京大学
44	ATTED II	公開済	東北大学
45	ADMER2	公開済	産業技術総合研究所
46	CGH-DataBase	公開済	東京医科歯科大学

2.1.2 成果物アーカイブの公開

成果物候補の選定フローに準拠し「平成 20 年度産業技術研究開発(統合データベースプロジェクト)」において調査された、アーカイブ対象データベースのリストの中から、電話・メール、また直接面会を行い、外部有識者等の意見を取り入れつつ、権利者・関係者との協議の上、現在 3 件の候補についてアーカイブ化を進めている。また、5 つのプロジェクトに関する成果報告書もダウンロード提供した。本節では、データアーカイブと新たに追加した成果報告書について記載する。

(1) アーカイブ対象の範囲と選定基準

アーカイブを調査するにあたり、アーカイブ対象の範囲と選定基準を定めた。範囲は「オンラインで容易に取得することができないもの(独自なサイトを持たないために公開していない等)」とした。さらに、対象を絞る基準として、「利用価値の高いもの」という趣旨で以下の選定基準を決定した。

- (a) 公開による利用を前提として出されたもの
- (b) 統一的条件で収集された網羅的データ

例えば、以下の様なデータセットやデータベースはアーカイブの選定基準を満たさない。

- (a) プロジェクトの進行途中でできた実験のパラメータに偏りのあるデータセット
- (b) 研究担当者が個人の PC に格納して共有、利用できないデータセット
- (c) 遺伝子のリストのみが羅列された紙媒体によるデータセット

(2) アーカイブ化を進めているデータベース 3 候補

- (a) DataBiosafety for Industrial Applications of Microbes (DIAM) (<http://medals.jp/list/detail/113>)

遺伝子組換え技術を利用する事業者の利便性向上と安全・安心に対する社会的要請に応える方策の一助として、本システム(微生物産業利用支援データベース、DataBiosafety for Industrial Applications of Microbes)(略称:DIAM)を構築した。本システムは、産業界におけるバイオセーフティ情報源としての有用性を高めることを目的として、それぞれ独自に開発された二つのデータベース(「組換え微生物等の産業使用促進情報システム」(略称:JBAM)、及び「バイオセーフティデータベース」)を統合したものである。(出典:オリジナルサイトより)(図 2.1.2-1)



図 2.1.2-1 DIAM トップページ(JBA 運用)

JBAM は経済産業省の委託事業「バイオインダストリー安全対策調査」の一環として作成され、財団法人バイオインダストリー協会(JBA)にて一般公開された。また、「バイオセイフティデータベース」は、新エネルギー・産業技術総合開発機構(NEDO)の委託事業「遺伝子組換え体の産業利用におけるリスク管理に関する研究」(平成14年度～平成18年度)の一環として作成され、JBAにて一般公開された。JBAではDIAMの今後の運用更新の予算がなく、2010年4月からの公開が問題となっており、アーカイブ化を現在進めている。

DIAMのアーカイブ化については、2009年11月4日にJBA、文部科学省統合データベースプロジェクト(DBCLS)の3者で打ち合わせを行い、ユーザーの利便性とJBAの意向を加味し、DBCLSが運用している寄託システムにアーカイブとして運用する方針を固め、現在も公開に向けた調整を行っている。

(b) Protein-Ligand INteractions (ProLINT)

新エネルギー・産業技術総合開発機構(NEDO)プロジェクト等の成果を活用し、2003年から社団法人バイオ産業情報化コンソーシアム(JBiC)で公開されていた有償データベースのバイオDBでは、発現プロファイル情報や生物機能情報、疾患遺伝子、文献リガンド情報、プロテイン結合リガンド情報を提供しており、研究機関や企業でも利用されていた。しかし、2005年に本データベースの公開が終わり、有用な情報は公開されることなく産業技術総合研究所(産総研)・バイオメディシナル情報研究センター(BIRC)にて置かれていた。プロジェクト目標である「過去の成果物をダウンロードできる環境を整備」するにあたり、特に外部研究機関からの要望があったリガンド情報(ProLINT(図2.1.2-2))をダウンロード可能にするため、保管されている媒体から該当するデータを抽出し、アーカイブとして公開・閲覧できる作業を行った。作業の前調査では、過去にシステムを利用していた先生や運用を担当していた企業にコンタクトをとり、アーカイブ化の実現可能性を検討した。

2010年2月上旬からの作業で、約4000件のリガンド情報の抽出に成功し、現在3月の公開に向けて、メタ情報の付加とファイルフォーマットの変換などについて作業を行っている。

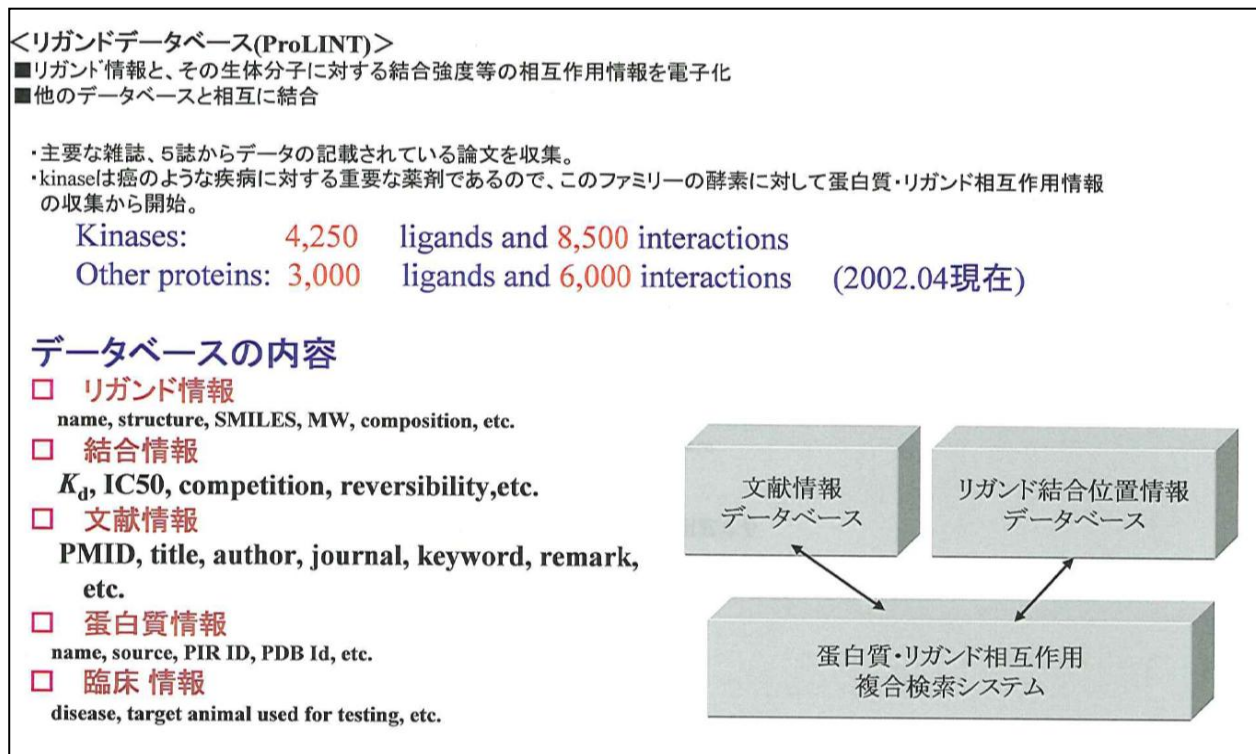


図 2.1.2-2 ProLINT 概要(公開時資料より抜粋)

(c) GenoBase (<http://medals.jp/list/detail/130>)

GenoBase は新エネルギー・産業技術総合開発機構(NEDO)プロジェクト「生物機能を活用した生産プロセスの基盤技術開発」(平成 13 年度～平成 17 年度)の成果を含む、大腸菌 K-12(W3110)の生細胞システムを総合的に理解するためのデータベースである(図 2.1.2-3)。継続的に運用更新されてきた GenoBase であるが、運用資金の問題で最新版 7.0 から運用機関を奈良先端科学技術大学院大学(NAIST)から米国パデュー大学に移っている。2009 年 7 月に管理者である NAIST 森浩禎教授にインタビューを実施し、2010 年 2 月 3 日に ver6.0 のデータの取得作業を行った。取得したファイル 2 種(31Gbyte、85Mbyte)に対するメタ情報の付加作業とファイルフォーマットの変換などについて作業を行っている。

図 2.1.2-3 GenoBase ver.6.0 トップ画面(奈良先端科学技術大学院大学(NAIST)運用)

(3) 成果報告書のダウンロード提供

平成 21 年度では経済産業省系ライフサイエンスプロジェクト MEDALS のうち、社団法人バイオ産業情報化コンソーシアム(JBiC)が実施した5プロジェクトに関する成果報告書(7ファイル)をMEDALSダウンロードページ(<http://medals.jp/contents/report>)より提供した(表 2.1.2-1)。

表 2.1.2-1 平成 21 年度より MEDALS から提供を始めた成果報告書一覧

項番	プロジェクト名	ファイル名	委託元
1	統合データベースプロジェクト	平成 20 年度成果報告書	経済産業省
2	ゲノム情報統合プロジェクト	平成 17～19 年度成果報告書	
3	バイオインフォマティクス関連データベース整備	平成 12～16 年度成果報告書	
4	遺伝子多様性モデル解析技術開発	平成 15～17 年度成果報告書	NEDO
5	機能性 RNA プロジェクト	平成 18 年度中間年報	
6		平成 19 年度中間年報	
7		平成 20 年度中間年報	

2. 1. 3 横断検索システム

経済産業省ライフサイエンスデータベースポータルサイトMEDALSの便覧にあるデータベースの中に格納されているデータの横断検索を可能にした。なお横断検索システムについては、以下の2点において、文部科学省統合データベースプロジェクト(DBCLS)と連携を図った。

- (a) 作成した検索用データを、本ポータルサイトで利用するだけでなく文部科学省サイトへ提供する。
- (b) DBCLS との間で、検索用データを互いに利用できるようにして、どちらのサイトからも同じデータを検索可能にする。

特に、本ポータルサイトで検索した場合には、ヒト遺伝子や転写産物の研究者(H-InvDB を利用するユーザー)にとって利便性のある表示形式で提供した。例えば、疾患に関連する医学情報や、ヒトに関連する遺伝子情報などにテーマを絞った検索結果が得られるように、検索語を疾患と関連のある言葉で補完し、目的の情報に辿り着きやすくした。以下、横断検索の概要について記載し、「文部科学省統合データベースプロジェクト(DBCLS)との連携」と「ユーザビリティ(使いやすさ)」について詳細に記載する。

(1) 横断検索の概要

横断検索システムはインターネットブラウザから利用できるウェブシステムである。横断検索の代表的な例として、Google やYahoo!もインターネット上のコンテンツを特定キーワードで横断的に検索するシステムであり、ワールドワイドに一般的なシステムとなってきた。

今年度の開発では、MEDALS 便覧に掲載されている各データベースに格納されているデータを、データベースの区分に関係なく横断的にキーワードで検索できるシステム(MEDALS 横断検索システム)の構築を行った(図 2.1.3-1)。本システムで横断的に検索を行うことで、ユーザーは興味あるキーワードに関する新たな生物種や分子情報のみならず、遺伝子 ID や糖鎖、特許情報を知ることができ、知見を広げる場合に非常に役に立つシステムである。また、一般的には横断検索ではできる限り多種多様なデータベースに検索を実施したい(情報の拡張)ニーズと、検索結果が大量に得られるため、重複なく目的の情報に絞りたい(情報の収縮)ニーズがある。MEDALS 横断検索システムではこれらの問題を「文部科学省統合データベースプロジェクト(DBCLS)との連携」と「高いユーザビリティ仕様」で解決した(詳細については以降で記載する)。

ライフサイエンス系データベースの横断検索は平成 20 年度の段階では DBCLS にて運用を行っており、日本糖鎖科学統合データベースとの横断検索も実装していた。本プロジェクトでは、平成 20 年度にサイト内の検索を実装し、平成 21 年度から横断検索の開発を始めた。MEDALS 横断検索システムでは、便覧に掲載してい

るデータベースの中から自動巡回で取得できた対象に対して検索で利用するサマリーファイル(Index)を作成した(表 2.1.3-1)。検索エンジンが Index を通じて経済産業省関連のデータベース(METI 関連データベース)を検索できるようになった。本システムは、OS が Linux のサーバーで稼働するウェブシステムであり、Ruby にて構築した。ユーザーが使用する際にも Internet Explorer6/7/8、Firefox3.5 以上、Safari4 以上のウェブブラウザが利用できる。

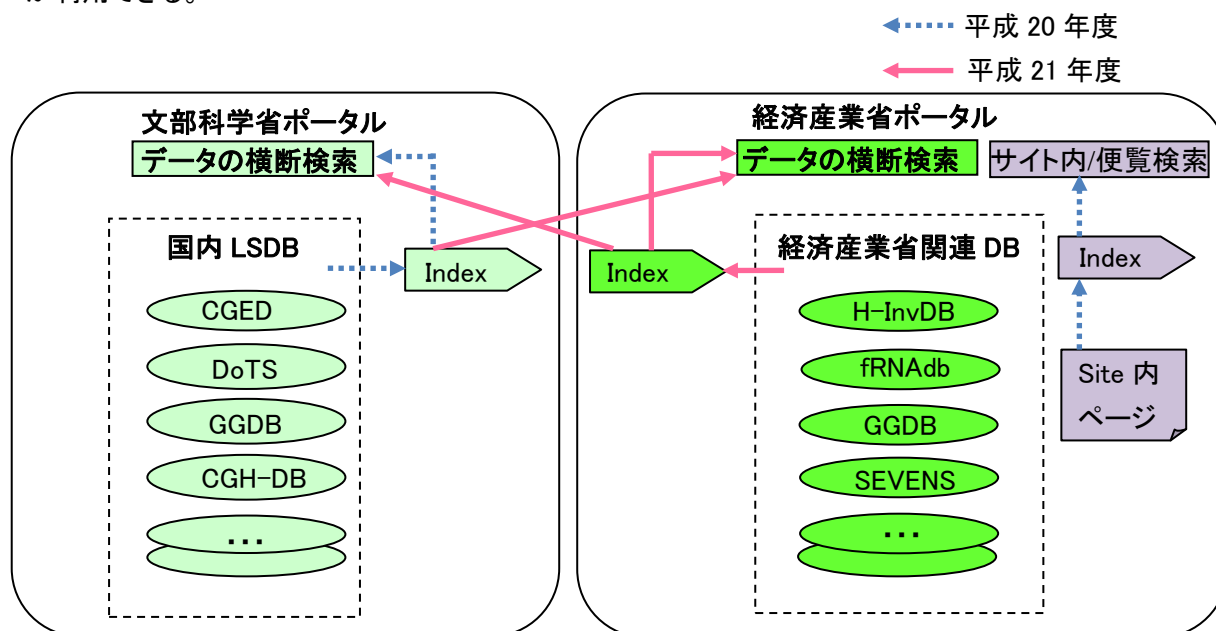


図 2.1.3-1 横断検索での連携イメージ

表 2.1.3-1 横断検索インデックス対象一覧

No.	データベース名称	No.	データベース名称
1	Evola	17	GMDB
2	LEGENDA	18	Arch GeNet
3	遺伝子多様性データベース (H-GOLD)	19	Kappa-View
4	H-DBAS	20	RnR
5	G-compass	21	MassBase
6	統合がんゲノムデータベース	22	DOGAN
7	VarySysDB	23	NBRC
8	機能性 RNA 配列データベース	24	siExplorer
9	CellMontage	25	GenoBase
10	SEVENS	26	組織別遺伝子発現データベース
11	脳画像データベース	27	癌組織遺伝子発現データベース
12	HGPD	28	FLJ Human cDNA Database
13	ProSeg	29	OPEN-PML
14	ASTRA	30	JSNP
15	ラットの脳断面図	31	ATTED II
16	糸状菌ゲノム・転写制御データベース	32	ADMER2
		33	CGH-DataBase

(2) 文部科学省統合データベースプロジェクト(DBCLS)との横断検索システム連携

連携を実現するに当たり DBCLS の関係者と協議を行い、MEDALS 横断検索において5つの点を満足する仕様で開発を行った。

(a) 同種の検索エンジンの採用と Index 構造の統一

検索エンジンは、DBCLS と同種の Hyper Estraier^{*1}を採用し、Index ファイルの形式や配置を統一した。これは、既に横断検索サービスの実績がある DBCLS と仕様を共通化し、効率的な連携を果たしユーザーに迅速にサービスを提供することを目的としたためである。特に(e)で記載する Index の直接参照機能を実現するためにも検索エンジンを統一する必要があった。

^{*1} Hyper Estraier は独立行政法人情報処理推進機構(IPA)による 2004 年度第 2 回未踏ソフトウェア創造事業の支援を受けて開発された国産の全文検索エンジンである。

(b) Index の重複を回避

平成 20 年度から横断検索サービスを運用している文部科学省統合データベースプロジェクト(DBCLS)では MEDALS 便覧に含まれているデータベースの Index を既に作成していた。そのため、DBCLS から Index リストを提供してもらい、既に DBCLS が作成済で重複となるデータベース 10 件に関しては MEDALS では Index 作成対象から除外した。例えば、新エネルギー・産業技術総合開発機構(NEDO)プロジェクトで開発したデータベースを含む日本糖鎖科学統合データベースは、平成 20 年度に DBCLS と横断検索を実装しており、Index から排除した。

(c) 文部科学省統合データベースプロジェクト(DBCLS)側の Index カテゴリーに準拠

検索結果をカテゴリー分けして表示する機能が文部科学省統合データベースプロジェクト(DBCLS)の横断検索システムにて実装されている。MEDALS 横断検索でも双方の Index を利用した場合に、同様の検索結果が表示されるためには、Index の重複を排除するのみでは不十分であり、DBCLS が提供する各 Index のカテゴリーに準拠する必要があった。そこで、DBCLS から提供されているカテゴリーファイル(http://lifesciencedb.jp/dbsearch/lsdb_tree.xml)を調査し、MEDALS 横断検索 Index を DBCLS カテゴリーに準拠させた。

(d) 出力形式の共通化

文部科学省統合データベースプロジェクト(DBCLS)で提供している API(Application Program Interface)は、他のプログラムから横断検索を利用する際に用いるプログラムモジュールである。この API を通じて検索を実行すると検索結果が JSON^{*2}形式で出力される。MEDALS 横断検索においても、一般ユーザーがプログラムを通じて横断検索を利用できるように API を作成した。作成した API は独自仕様を組み入れながら、DBCLS との互換(利用する際の入出力仕様)を重視した。これにより、DBCLS の他のシステムから MEDALS 横断検索を利用する場合でも JSON 形式で情報を取得できるようになった。出力データの構造を表 2.1.3-2 に記載する。

^{*2} JSON (JavaScript Object Notation)は、簡単なデータ交換フォーマットであり、人間にとって読み書きが容易で、マシンにとっても簡単に活用や生成を行なえる形式である。

表 2.1.3-2 API の出力データ構造

要素	要素内の構造		値の内容
node_url	文字列		検索対象ノードの URL
result	結果オブジェクトの配列		検索結果を格納 result[0]は、1 件目のオブジェクトを格納。デフォルトは最大 30 件を格納する。
		doc	各検索結果のスニペット(キーワード*周辺のテキストの抜き出し)
		db	各検索結果のデータベース名
		url	各検索結果の URL
		title	各検索結果のタイトル
		score	各検索結果のヒットスコア
		icon	各検索結果のアイコン表示用 img タグ(pdf ファイルの場合変数を設定して出力)

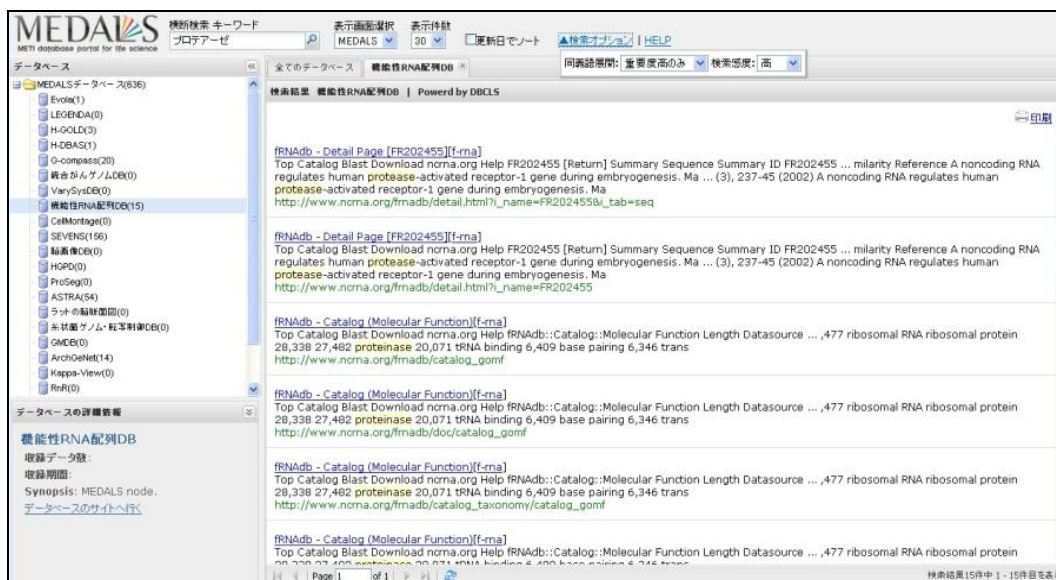


図 2.1.3-2 MEDALS(上)とDBCLS(下)双方の横断検索画面レイアウト

(b) 表示画面選択機能

横断検索を利用する際、MEDALS の独自機能(後述する(d)～(g))を利用したい場合と、DBCLS の独自機能(遺伝子リスト検索など)を利用したい場合がある。そこで、ユーザーが横断検索するサイトを選択する機能を実装した(図 2.1.3-3)。

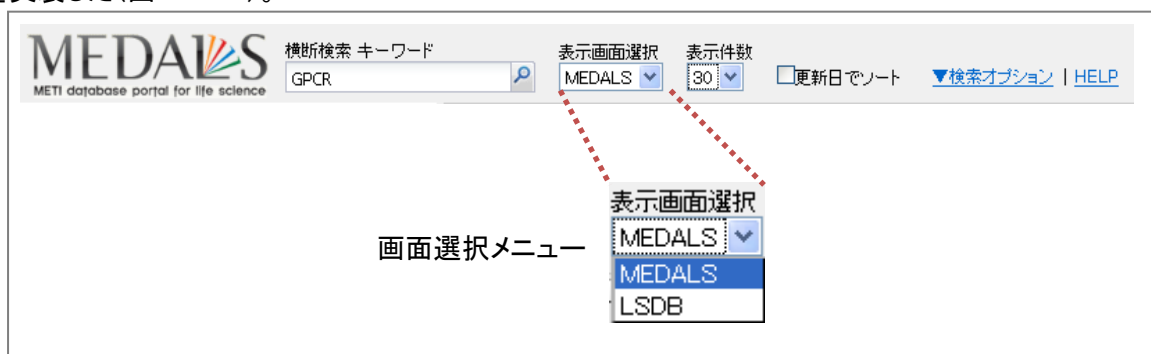


図 2.1.3-3 表示画面選択機能

(c) ライフサイエンス辞書(LSD)利用

ライフサイエンス辞書プロジェクトでは、様々な学問領域の研究者が協力し合い、広範な生命科学の学問領域で使われる専門用語などを電子辞書として開発、配布を行っている。本プロジェクトでは、ライフサイエンスプロジェクト担当者からライフサイエンス辞書(LSD)の利用承諾を受け、横断検索システムに取り込んだ。LSD を取り込むことで、ライフサイエンスの専門的な用語を、かな漢字変換だけではなく、日本語と英語の変換などを行えるようになった。例えば、英語の「Parkinson Disease」は LSD で変換すると「パーキンソン病」であることが簡単にわかる。

(d) 広げる検索(同義語展開と検索感度)

ユーザーが検索で利用するキーワードは、同義語をもつものがあつた場合、キーワードの和英変換・かな漢字変換のみでは検索結果に漏れが発生する。例えば、「プロテアーゼ」で検索する場合には Protease だけではなく、「ペプチターゼ」「タンパク分解酵素」も検索キーワードとして利用すると、より広い範囲で検索結果を手に入れることができる。しかし、これらの同義語をユーザーが全て手作業で指定することは困難である。そこで、MEDALS 横断検索システムでは、LSDを利用して同義語を自動的に展開して検索を行う機能を開発した。しかも展開の範囲もユーザーが選択することができるので、自由度の高い絞り込みが可能となった(図 2.1.3-4)。

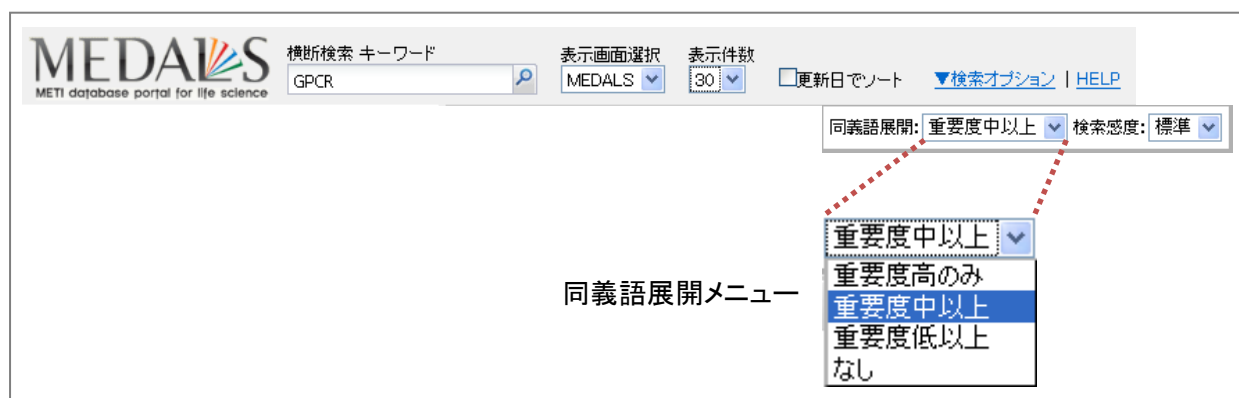


図 2.1.3-4 同義語展開メニュー

この機能は LSD チームで構築しているコーパス^{*3}における同義語の出現頻度に応じて付けられている重要度を利用している。重要度のレベルと出現頻度との関係は表 2.1.3-3 の通りである。

^{*3} コーパスは自然言語の文書を構造化して大量に蓄積したデータのこと。

表 2.1.3-3 重要度のレベルと出現頻度との関係

項番	同義語展開	出現回数
1	重要度高のみ	10000 回以上
2	重要度中以上	100 回以上
3	重要度低以上	1 回以上

実際の横断検索を利用して「プロテアーゼ」を検索した場合における、これらのレベルの違いによる検索結果の違いをまとめた(表 2.1.3-4)。同義語を展開しない場合(14,278 件)と比較して、同義語の重要度を「低以上」に設定すると 128,407 件まで検索結果を広げることができる。また、重要度を「高のみ」に設定すると 79,748 件まで、同義語を展開しつつ絞り込みができた。

表 2.1.3-4 同義語展開による検索結果数の違い

項番	同義語展開	検索結果数
1	重要度高のみ	79,748
2	重要度中以上	128,217
3	重要度低以上	128,407
4	なし	14,278

また、キーワードで検索する際の感度の強度もユーザーが設定する機能も開発した。この機能は Hyper Estraier が持つ検索強度設定の機能を利用して実現した。3 段階(高、標準、低)で設定ができるため、できるだけたくさんの結果を得たい場合は感度を「高」に設定すればよい(図 2.1.3-5)。例えば特許公報に対する検索では、できる限り多くの検索結果を得て漏れを防ぎたい時に有効である。

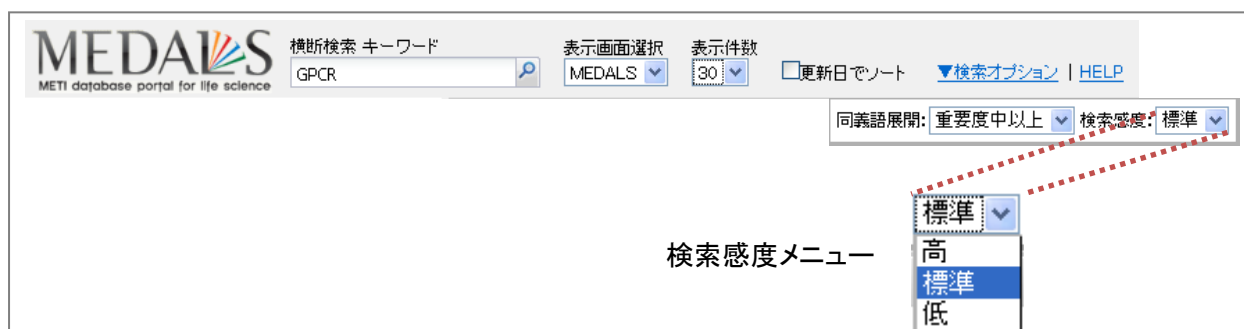


図 2.1.3-5 検索感度メニュー

(e) 絞る検索(関連単語の辞書におけるキーワード入力補完機能)

同義語の展開や検索感度の機能により、関連のある言葉で広げる検索が可能となったが、検索結果が多すぎる場合には人手で全て確認することは非効率であり、広げた検索の結果から目的の情報を抽出する絞る検索が必要となってくる。そこで開発した機能が「関連単語の辞書におけるキーワード入力補完機能」である。関連単語とは薬と疾患や、疾患と遺伝子といった生物学的に関連性を有する単語である。例えば、疾患の「マラリア」との関連する単語の一つは、治療薬の「Mefloquine」である。この様な関連する単語で AND 検索行うことで、目的の情報に簡単に辿り着くことができる。

まず、関連単語の辞書について記載する。関連単語の辞書の元データは産業技術総合研究所(産総研)・バイオメディカル情報研究センター(BIRC) 分子システム情報統合チームで運営しているデータベース LEGENDA である。LEGENDA(<http://hin.jp/legenda/>)は「MEDLINE 文献に書かれた遺伝子、遺伝子機能、疾患、または化合物のうち、2 つが共起した(関連性をもって一緒に出現した)文献を探すシステム」である(詳細は MEDALS 便覧参照 <http://medals.jp/list/detail/3>)。関連単語の辞書におけるキーワード入力補完機能を実現するため、LEGENDA が格納している関連性をもつ単語データセットを辞書として利用した。

次に、本機能を利用したユースケースを記載する。例えばある研究者がパーキンソン病の治療薬である「Levodopa」に関連するデータベースや特許情報を探している場合を想定する。研究者は MEDALS 横断検索システムサービスの検索ボックスに「Levodopa」を入力して、空文字(スペース)を入れるだけで、Levodopa と関連の高い単語(Dopamin、Parkinson Disease、Carbidopa など)の候補を得ることができる(図 2.1.3-6)。

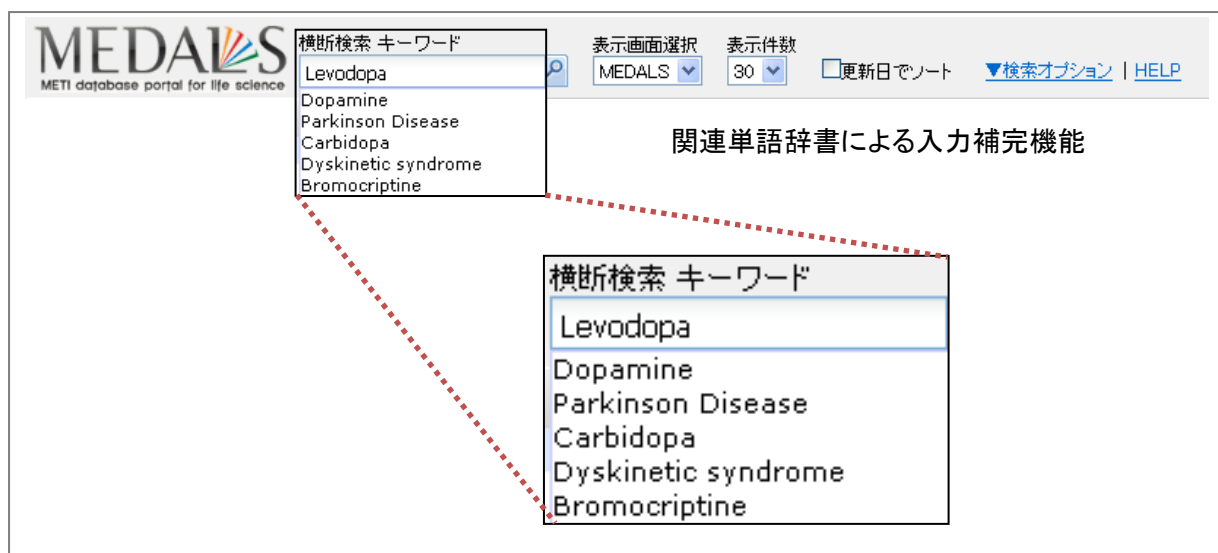


図 2.1.3-6 関連単語の辞書におけるキーワード入力補完機能

候補の中から「Parkinson Disease」を選択して検索を実行した際の結果（一部）として、PharmGKB（薬と疾患と遺伝子の関連性を示すデータベース）や特許公報（日本）のデータがヒットした（図 2.1.3-7）。

PharmGKB(薬と疾患と遺伝子の関連を示すデータベース)

[PharmGKB: Parkinsonian Disorders\[pharmgkb\]](#)
 ? ters Parkinsonian Disorders Curated Relationships Downloads/Cross-references PharmGKB Accessio
 47062 Alternate Names: Autosomal Dominant Juvenile Parkinson Disease ; Autosomal Dominant
 Juvenile Parkinsonism; Autosomal Dominant Parkinsonism; Autosomal Recessive Juvenile Parkinson
 Disease nt Parkinsonism, Autosomal; Experimental Parkinson Disease ; Experimental Parkinson ;
 Experimental Parkinsonism is] Drug Relationship Details catecholamines FA View levodopa FA View
 Common es Medline Plus CTD Sitemap Privacy
<http://www.pharmgkb.org/do/serve?objId=PA447062&objCls=Disease>

特許公報(日本)

[公表特許公報\(A\): パーキンソン病の予防的処置のための置換された2-アミノテトラリンの使用\[japatent\]](#)
 JP 2007518715 公表特許公報(A) 20070712 2006546072 20041223 パーキンソン 病の予防的処置のための置換
 4C086MA56 4C086MA63 4C086NA14 4C086ZA02 4C086ZB21 パーキンソン 病は慢性の進行性神経変性の結果で
 あり、その原因はいま らかである。運動症状を軽減するための医薬として第1に レボドパ、ドーパミンアゴニスト、たとえばロチゴチ
 ン、プラミベはコリン作動性信号カスケードに介入し、かつこの方法で パーキンソン 病タイプの運動障害の症状に影響を与える。
パーキンソン 経皮投与すること、およびこれと関連してUPDRS(Unified Parkinson 's Disease Rating Scale)スコアの改善
 を記載している
<http://lifesciencedb.jp/dbsearch/Patent/ipdl.php?year=2007f&file=xr/7zZ0WKmI2kzoheNMHtw==&search='Levodopa%E3%80%80Parkinson%20Disease'>

図 2.1.3-7 関連単語を使った横断検索結果例

本機能を使うことでどれくらい目的の情報に絞り込めるかをこのユースケースの例をとってまとめた(表 2.1.3-5)。Levodopa のみで検索した場合と比較して、関連語(Parkinson disease や Carbidopa)を併用することで結果総数は 20%に絞ることができた。また、各カテゴリー(医薬データベース、ゲノム・遺伝子データベース、特許データベース)で検索に当たった件数も絞ることができた。本機能で提供する関連単語で多量のデータから目的の情報へ速く簡単に辿り着くことができる。

表 2.1.3-5 Levodopa と関連単語で AND 検索を行った絞り込みの例

検索語	検索結果総数	医薬 データベース	ゲノム・遺伝子 データベース	特許 データベース
Levodopa	1007	611	53	298
+Parkinson disease	200	79	20	84
+Carbidopa	203	55	3	122

(f) 検索結果印刷メニュー

検索結果を会議や発表の印刷資料として利用する際に便利な機能が「検索結果印刷メニュー」である。検索結果が表示されたウェブブラウザを直接印刷しても検索結果のみを全て印刷することはできないため、研究現場での利便性を考慮して本メニューを開発した(図 2.1.3-8)。

The screenshot displays the MEDALS database interface. At the top, there's a search bar with 'Levodopa' entered. Below the search bar, a list of search results is shown. A red box highlights the '検索結果印刷メニュー' (Search Results Print Menu) button. Another red box highlights the '印刷' (Print) button. A third red box highlights the '印刷ページ' (Print Page) button. The print page shows a detailed view of the search results for 'Levodopa' and 'Parkinson disease'. The page includes a table with columns for '検索結果' (Search Results) and '印刷' (Print). The table lists various search results, including 'KEGG OGENES: NC_011087.685', 'BodyMap-Xs Bombyx mori - esophagus', and 'PDB-2zle: DegP12 (タンパク質品質管理因子)・OMP (外膜タンパク質)・EM Navigator'. The page also includes a '検索結果 全てのデータベース - Powerd by DBCLS' section.

図 2.1.3-8 検索結果の印刷メニュー

(g) 表示件数設定メニュー

前述したように横断検索の問題点として結果が多すぎるケースがある。この場合、検索結果が一度にリスト表示されても、ユーザーには閲覧操作が不便だけではなく、視認性にも影響を及ぼす。そこで開発したのが表示件数設定メニューである。ユーザーが1画面中に表示する検索結果件数を3段階(10件、30件、50件)で設定できる機能である(図 2.1.3-9)。特に画面の小さいノートパソコンを利用する研究者にとっては検索結果が閲覧しやすくなり、また、印刷機能を併用することで必要な情報が表示された結果のみ印刷して、資料として利用することができる。



図 2.1.3-9 表示件数設定メニュー

2. 1. 4 ポータルサイト運用

(1) セキュリティ診断

本プロジェクトで開発した MEDALS 及び 4 種のアプリケーション(H-InvDB の Web サービス(ウェブサービス)、機能性 RNA データベース(fRNAdb)のウェブサービス、リンク自動管理システム、H-DBAS)は公共性が強く、インターネットを通じたユーザーに信頼性・可用性を維持する必要がある。しかし、現在のインターネットを取り巻く環境は非常に厳しく、脆弱性を悪用したサイトへの攻撃やウイルス感染によりシステムのサービス停止や情報の漏洩などの被害が発生することがある。そのため、インターネットセキュリティの検査を行うことは、公共的に情報を提供するサイトには必要不可欠である。また、世界的に利用されているセキュリティ診断ソフト(前年度よりプロジェクト内で利用)を利用して検査することで、結果に対する信憑性を確保し、検査漏れを防止し、一般性のあるセキュリティレベルを保持できる。そこで、MEDALS 公開サーバーで動作しているウェブアプリケーション(Web アプリケーション)に対するセキュリティの検査を行い、セキュリティレベルを評価する予定である(期間:2010 年 3 月 20 日~2010 年 3 月 31 日)。

- 診断内容

実施したセキュリティ診断項目を表 2.1.4-1 に示す。

表 2.1.4-1 実施するセキュリティ項目

項番	名称
1	セッション管理
2	クッキー使用法対策
3	クロスサイトスクリプティング対策
4	バッファオーバーフロー対策
5	コマンドインジェクション対策
6	SQL インジェクション対策
7	ファイル改ざん、消去対策
8	エラー画面
9	強制ブラウジング
10	ファイルの誤った公開
11	パス名パラメータの未チェック
12	HTTP レスポンス分割
13	ディレクトリトラバーサル
14	セキュリティ上不適切な画面設計

(2) 情報管理ツール

本プロジェクトではユーザーのニーズ調査としてアンケートと企業インタビューを継続的に実施してきた(詳細は2.3参照)。今までのアンケートやインタビューではデータベースのユーザーは質を重視、特に更新されていることを重視していることが分かった。100 件以上のデータベースやツールを便覧に掲載しているが、いつ、どのサイトが更新されるかわからない状態で、毎日人手で更新を確認することは非効率である。そこで、各サイトの更新情報を簡単に管理するツールを開発した。本ツールは、便覧に掲載されている 100 件以上のオリジナルサイトを 24 時間周期で自動監視し、オリジナルサイトのページ更新が行われた場合に管理者へ通知するシステムである(図 2.1.4-1)。また、監視対象の中から個別サイトの更新部分を簡単に閲覧するための管理 GUI(Graphical User Interface)も開発した。

管理 GUI 画面は機能別に4つの区画に分かれている(図 2.1.4-2)。検索バー(【A】)では、ツールに登録している監視対象のサイトの中から管理者が見たいデータを抽出できる。URL やサイト名、更新された期間で検索できる。検索結果エリア(【B】)では、検索結果を各サイト単位のテーブルリスト形式で表示している。個別のテーブルをクリックすると選択された項目を枠線でフォーカスする。特権モードエリア(【C】)では、特権を与えられた管理者がログインした場合に利用できる機能で、登録しているサイトの追加・削除・変更が行える。変更履歴エリア(【D】)では【B】でフォーカスしたテーブルに記載されたサイトの更新履歴と各更新時のページを参照することができる。

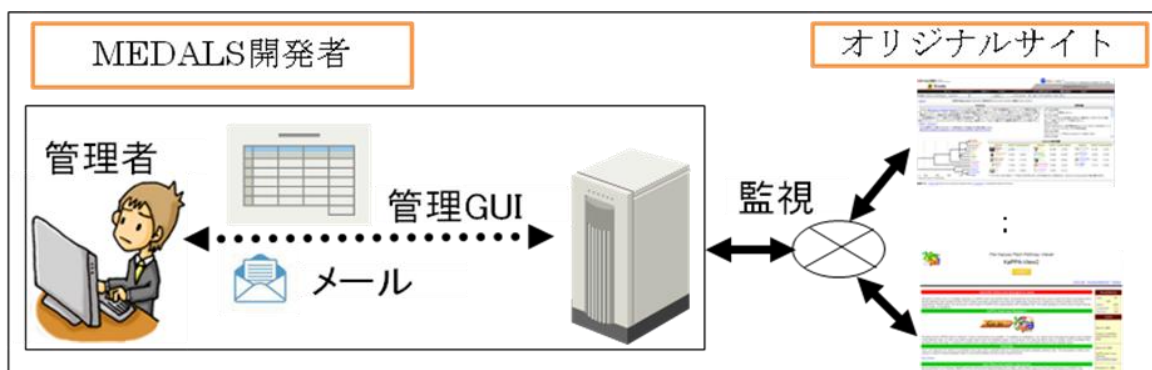


図 2.1.4-1 情報管理ツール概念図

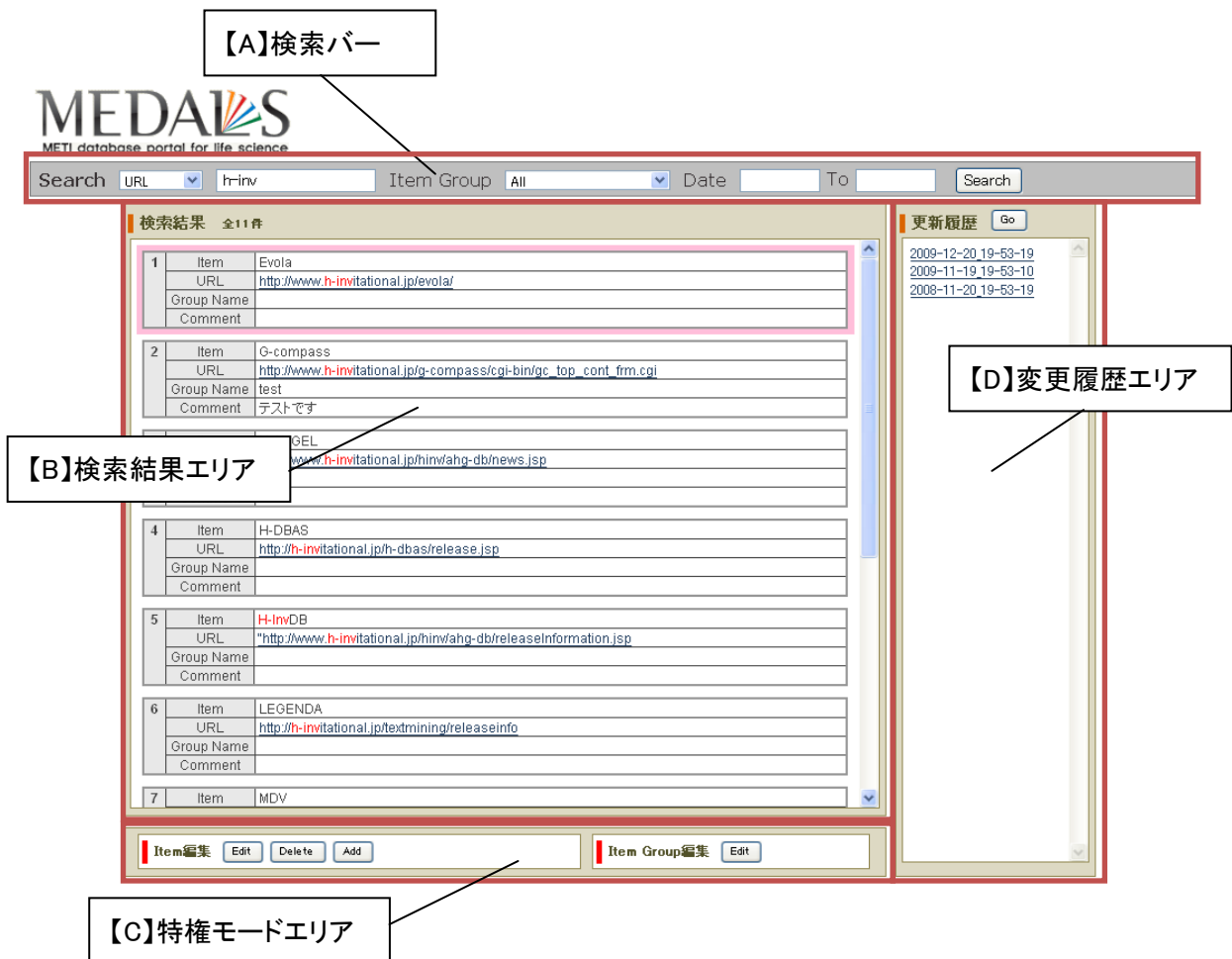


図 2.1.4-2 管理 GUI 画面

(3) プロジェクト便覧の構築

平成 20 年度には、成果物 1 つについてプロジェクト 1 つだけを同じページに示していた。しかし、調査を進めるに従い、同じ成果物が後に別の関連プロジェクトによって機能追加がされたりしているケースが多いなど、1 対 1 では無理が生じることが明らかになってきた。そこで、成果物とプロジェクトが多対多の関係を正しく示せるようデータ管理をし、これらの関連をわかりやすく表示することを検討した。そこで、プロジェクト便覧を新たに構築した。

図 2.1.4-3 に、プロジェクトと成果物の一覧のページを示す。ユーザーはこのページを見ることにより、あるプロジェクトから出たすべての成果物がわかるようになった。また、個別の成果物便覧詳細画面で、それを産生したプロジェクトの詳細が表示され、各プロジェクト便覧へのリンクをつけた(図 2.1.4-4)。逆に、プロジェクト便覧の詳細画面(図 2.1.4-5)から産生した成果物(図 2.1.4-6)が一覧できる。

プロジェクト	年度	成果物名
ゲノム情報統合プロジェクト	2005-2007	 Evola  G-compass  H-ANGEL  H-DBAS  H-Exp  H-InvDB  HEAT  LEGENDA  MDV  PPI view  TACT  VarySysDB
タンパク質の構造・機能予測法の開発とヒトゲノム配列への適用	2005-2008	 FORTE
機能性RNAプロジェクト	2005-2009	 CentroidFold  Murlet  RNAmine  Rfold  機能性RNA配列データベース  機能性RNAゲノムブラウザ

図 2.1.4-3 プロジェクトと成果物の一覧

 関連プロジェクト ゲノム情報統合プロジェクト バイオインフォマティクス関連データベース整備 統合データベースプロジェクト 関連プロジェクト	成果物名	G-compass
	成果物の別名	G-compass : Comparative genome browser
	成果物に関する説明	G-compass は比較ゲノム解析研究のためのツールとして開発され、進化的に保存されたゲノム領域とオルソログ遺伝子のデータを提供しています。解析対象はヒト+12生物種です(チンパンジー、マカクザル、マウス、ラット、イヌ、ウマ、ウシ、オポッサム、ニワトリ、ゼブラフィッシュ、メダカ、ミドリフグ)。ゲノムの極保存領域(UCE)とコピー数多型領域(CNV)の情報を提供するとともに、ウィンドウ解析やドットプロットを表示するツールを備えています。
	成果物のタイプ	DB
	運用機関	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
	機関所在国	日本
	サイトURL	http://www.hi-invitational.jp/g-compass/
	インターフェイス	検索サイト
	入力例	トップページ上部の“Search”で、“Target”プルダウンメニューからKeywordを選び、テキストボックスに“calmodulin”を入力して“submit”ボタンを押す。

↓ 以下に、プロジェクト便覧が複数続く

図 2.1.4-4 成果物便覧(複数のプロジェクトを表示)

プロジェクト名	バイオインフォマティクス関連データベース整備
分野	バイオインフォマティクス
目的	ゲノム情報とバイオインフォマティクス知的基盤の整備を通じた産業界への貢献を行う。
紹介	本事業では、ヒトゲノム研究成果の共通研究基盤としての統合データベースの開発・提供に取り組み、ヒトゲノムに関する有用なバイオインフォマティクス関連データベースと各種解析用のソフトウェアを統合的に利用できるシステムを整備する。また、ヒトゲノムのアノテーション研究を実施し、ヒト遺伝子とタンパク質に関する生物学的な知識や他の生物の関連情報をデータベース化する。さらに、以上に必要な情報処理技術の研究開発や、比較ゲノム解析等を用いたゲノム配列からの情報抽出に関する基礎研究も推進する。これらの研究開発によって、基礎科学としてのバイオインフォマティクスの発展に寄与するだけでなく、広く産業界等へバイオ情報の高度な利用環境を提供することにより、我が国のバイオ産業における技術開発の促進に貢献する。
キーワード	ヒト完全長cDNA国際アノテーションジャンボリー 統合バイオデータベースシステム ヒト完全長cDNA 比較ゲノム アノテーションシステム
開始-終了年度	2000-2004
代表者	五條堀 孝
代表者所属組織	国立遺伝学研究所 生命情報・DDBJ研究センター 産業技術総合研究所
予算(百万円)	3,026
代表委託機関	(社)バイオ産業情報化コンソーシアム
参加機関	産業技術総合研究所 (株)日立製作所 (株)シーズ・ラボ (株)ダイナコム 富士通(株) 三井情報開発(株) ビッツ(株) CRCソリューションズ (有)アクシオヘリックス 協和発酵工業(株)
論文等(PubMed ID)	15103394 15860772 15608263 15572454 15557284 15208628 15180706 15045488 12799463 12082643
特許(日本、海外)	PGT/JP2004/011624 特開2006-323830 特開2006-072861
成果報告書	 平成12年度-16年度 環境対応技術開発等(バイオインフォマティクス関連データベース整備)に関する研究報告書 [PDF 27.8MB]

図 2.1.4-5 プロジェクト便覧の詳細画面

成果物(データベース、解析ツール)		
G-compass		
 [公式 サイト] [詳細を見る]	要約	G-compassは比較ゲノム解析研究のためのツールとして開発され、進化的に保存されたゲノム領域とオルソログ遺伝子のデータを提供しています。解析対象はヒト+12生物種です(チンパンジー、マカクザル、マウス、ラット、イヌ、ウマ、ウシ、オポッサム、ニワトリ、ゼブラフィッシュ、メダカ、ミドリフグ)。ゲノムの極保存領域(UCE)とコピー数多型領域(CNV)の情報を提供するとともに、ウィンドウ解析やドットプロットを表示するツールを備えています。
	主な対象データ	比較ゲノム
H-ANGEL		
 [公式 サイト] [詳細を見る]	要約	H-ANGELは、ヒトの遺伝子発現についての情報を提供するデータベースです。H-Invitationalプロジェクトの構築した転写産物データに対する遺伝子発現パターンを表示します。遺伝子発現パターンは複数の測定プラットフォームを統合した組織特異的発現データに基づき、実用的な組織分類で表現しています。H-ANGELはまた、遺伝子の発現情報をヒトゲノム上の対応する物理的位置上に表示します。これらの情報は、H-InvDBに蓄積された対応する転写産物あるいは遺伝子座のアノテーションデータとリンクされており、こうした統合を通じて、プラットフォーム横断的に遺伝子発現データを比較して眺めることができます。
	主な対象データ	遺伝子発現
H-InvDB		
 [公式 サイト] [詳細を見る]	要約	H-InvDBはヒトの遺伝子と転写産物を対象とした統合データベースです。ヒトのすべての転写産物の配列をあらゆる手法で解析することにより、ヒト遺伝子の構造、選択的スプライシング、タンパク質としての機能などの精査されたアノテーション(注釈付)情報を提供しています。
	主な対象データ	RNA ヒト完全長cDNA, mRNA

図 2.1.4-6 プロジェクト便覧で表示される複数の成果物

(4) 報告書ダウンロードページの追加

前年度(平成 20 年度)のアーカイブページ(成果物のデータやプログラム自体をダウンロードできるページ)には、次の 2 件が公開されていた。

- myPresto: 医薬品開発支援用分子シミュレーション計算ソフトおよびデータベースの集合体
- PubMedScan: 新規関連文献お知らせツール(local 版)

平成 21 年度には、2.1.2 項(3)に述べたように、「機能性 RNA プロジェクト」の中間年報などを新たに追加し、7 件の報告書を MEDALS に追加、公開した。これらをデータベースやツールと区別して、わかりやすいように、図 2.1.4-7 に示すような「成果報告書ダウンロードページ」を公開した。これらの報告書は 2009 年 4 月～2010 年 1 月までに 153 回のダウンロードがなされた。



図 2.1.4-7 成果報告書ダウンロードページ

(5) 検索エンジン対策

Google, Yahoo!などの検索エンジンで、検索結果に MEDALS のページが出てきた時に、ユーザの誘導を促すための記述の工夫を行った。サーバー側で特に何も設定しない場合は、検索エンジン側で MEDALS 内の一部の文章を適当にピックアップしたものが出ることがあるが、図 2.1.4-8 に示すようなメタタグを設けることで、図 2.1.4-9 に示すような結果が検索エンジンでの検索結果として表示される。このようにして、適切に誘導することを目指した。その他キーワード・タグ程度の対策を行っている。それ以上の対策、言い換えれば、順位の1つや2つを上げるような対策は本質的でないため行っていない。

- <meta name="keywords" content='MEDALS,JBIC,AIST,経済産業省,ライフサイエンス,統合データベースポータルサイト,今西規,五條堀孝' />
- <meta name="description" content='このポータルサイトでは、これまで経済産業省関連機関により実施されたライフサイエンス分野の研究開発プロジェクトについて、1989年以降の成果物（データベース、ツール等）の情報を整理・提供しています。現在の成果物数: 104 (Database:60, Analysis tool:44) 現在のプロジェクト数: 51 今後もこれらの成果物が効果的に利用されることを目的に、内容を更新・追加していきます。（中略）/>

図 2.1.4-8 各ページに埋め込まれたメタタグ

MEDALS 経済産業省ライフサイエンス統合データベースポータルサイト

このポータルサイトでは、これまで経済産業省関連機関により実施されたライフサイエンス分野の研究開発プロジェクトについて、1989年以降の成果物（データベース、ツール等）の情報を整理・提供しています。現在の成果物数: 104 (Database:60, ...

medals.jp/ - [キャッシュ - 類似ページ](#)

図 2.1.4-9 検索エンジンに表示される説明文

(6) アクセスログ解析

MEDALS へのアクセスは、今年度は 22,298 件のページビューがあったので、月平均にすると約 2,200 ページであった。調査期間は、2009 年 4 月 1 日から 2010 年 1 月 31 日までの 10 ヶ月である。どの成果物便覧がよくみられていたかをみるために、成果物のページビュー数トップ 20 を表 2.1.4-2 に示す。調査期間は、上記と同じである。全成果物の数は、1 月末時点で 100 個である。MEDALS 独自のツールである「PubMedScan」や「リンク自動管理システム」など（青色の箇所）が多く、これらを特に紹介することができていることがわかる。2 位の「CentroidFold」（黄色の箇所）は、機能性 RNA プロジェクトの成果物であるが、近年の成果でもあり特に注目されていることがわかる。3 位の「ラットの脳断面図」があるが、実はこれは他とは違って、検索エンジンで固有名詞の検索から訪れるケースよりも、「脳」+「断面図」などのキーワードで来訪するケースが目立っている。そのようなリソースを探している学生などを誘導している可能性があり、そうであれば MEDALS はポータルとしての役割をよく果たしていることになる。10 位の「myPresto」（赤色の箇所）は、「医薬品開発支援用分子シミュレーション計算ソフトおよびデータベースの集合体」であり、創薬分野で有用なソフトが上位にはいっていることがわかる。

便覧へのアクセスは、解説するという MEDALS の役割を果たしているが、さらに言えば実際に成果物のデータベースを訪れて使用してもらうまでステップが進むのが望ましい。我々は独自の工夫により、便覧などから各データベースサイトへ実際にユーザーが訪れた件数の統計をとった。表 2.1.4-3 に、外部サイトへの誘導回数トップ 20 を示す。期間は先のデータと同じで、2009 年 4 月 1 日から 2010 年 1 月 31 日までの 10 ヶ月である。上位の 2 つ、すなわち「ヒト遺伝子統合データベース H-InvDB」と「文部科学省統合データベースプロジェクト」のサイトが、顕著に多くのアクセスをもっていた。これらは、統合データベースに対するユーザーの関心は、どのユーザーも共通して高いことを示唆している。また、我々のサイトもそれに沿って誘導するよう意図して作成しているので、そこへ誘導できていると考えられる。他にリストされている多くの成果物サイトをみると、上記の表 2.1.4-2 に示した「成果物のページビュー数トップ 20」と類似している。これは便覧をみたユーザーがサイトを訪れたと考えられるので、妥当な結果である。

表 2.1.4-2 成果物のページビュー数トップ 20

	成果物(詳細情報ページ)	内容	ページビュー
1	PubMedScan	文献ニュース	450
2	CentroidFold	二次構造予測	323
3	ラットの脳断面図	脳画像	251
4	ID一括変換システム	ID変換	185
5	リンク自動管理システム	IDへのリンク	136
6	WbLF PSORT	細胞内局在予測	135
7	ARCHAIC	古細菌ゲノムDNAの配列とその遺伝子情報	111
8	遺伝子多様性データベース公開システム(H-GOLD)	遺伝子多様性	110
9	G-compass	比較ゲノム	76
10	myPresto	分子シミュレーション計算	70
11	Evola	分子進化	64
12	PPI view	タンパク質間相互作用	64
13	HEAT	遺伝子セットの統計解析	63
14	ALN	配列アラインメント	60
15	HGPD	ヒト遺伝子、タンパク質	57
16	LEGENDA	文献、遺伝子、疾患、化合物	55
17	SNP-system	DNA多型	46
18	KaPPA-View	植物代謝	44
19	H-InvDB	ヒト遺伝子、転写産物	42
20	SEVENS	7本膜貫通ヘリックス型タンパク質のGPCR遺伝子	41

表 2.1.4-3 外部サイトへの誘導回数トップ 20

ランク	タイトル	誘導回数
1	ヒト遺伝子統合データベース(H-InvDB)	443
2	文科省統合データベースプロジェクト(DBCLS)	195
3	myPresto	101
4	ARCHAIC	88
5	ラットの脳断面図	79
6	Evola	74
7	CentroidFold	71
8	WbLF PSORT	71
9	ALN	70
10	ASTRA	67
11	G-compass	54
12	LEGENDA	49
13	ID一括変換システム	40
14	ARMLIPDB	39
15	HGPD	37
16	CellMontage	36
17	PPI view	35
18	リンク自動管理システム	35
19	微生物データベースシステム	32
20	遺伝子多様性データベース公開システム(H-GOLD)	30

2. 1. 5 成果の追加(PubMedScan:研究ツールの開発)

新規関連文献お知らせツール PubMedScan は、NCBI PubMed の文献から、興味のあるトピックスで新規に PubMed に登録された文献を定期的に(毎日)メールでお知らせするツールである。2009 年 3 月に Web 版を開発して公開し、また PMID 抽出ツールの開発・配布を同時に開始した(図 2.1.5-1)。平成 21 年度は、文献表示画面で、文献をキーワードやスコア等で絞れるように機能を追加した(図 2.1.5-2 の赤枠)。その他、以下の更新を行った。

- 提示情報の効率的把握のための改良
 - ✓ 取得論文数をメールの件名に記載
 - ✓ PubMedScan・PMID-Extractor のキャプチャ画像を追加
 - ✓ PMID-Extractor の利用へ誘導
- システム制限の緩和
 - ✓ 多様な入力に対応(全角等も入力可能)
 - ✓ レポート論文数の制限を拡大(100 件→1,000 件)

PubMedScan は便覧閲覧数もトップクラスであり、ユーザーの関心が高い(2.1.4 項(6))。今後もユーザーのニーズを聞きながら、改良を進めていきたい。

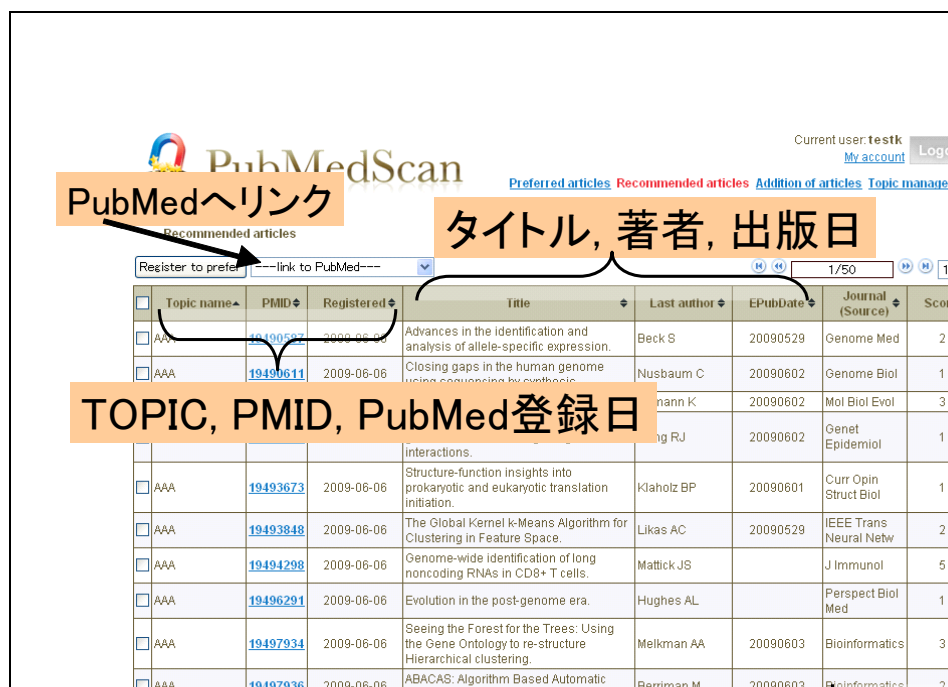


図 2.1.5-1 推薦された文献のリスト画面

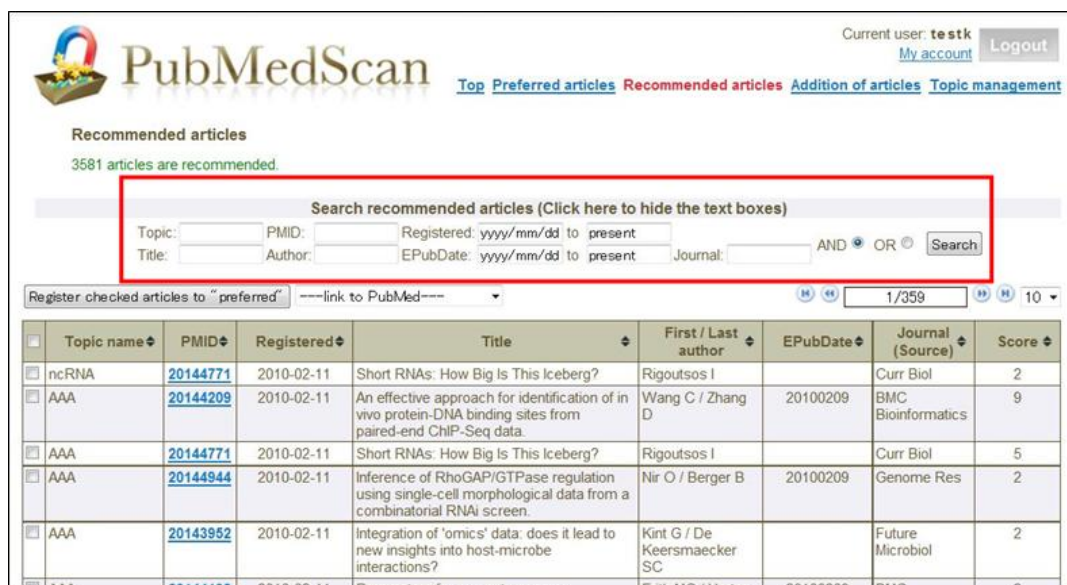


図 2.1.5-2 検索できる機能(赤枠)が付与された文献リスト画面

2.2 経済産業省関連機関から産生されるデータベース等の統合

2.2.1 ウェブサービスの開発

ヒト全遺伝子のアノテーション統合データベース(H-Invitational DB, H-InvDB)を基礎として、糖タンパクやヒト多型情報等の経済産業省関連機関から産生されるデータベース等を統合するために研究開発を行った。具体的には、Web サービス(ウェブサービス)を使ったデータベース連携のしくみを整備するため、連携の基盤となる H-InvDB の Web サービスの改良・機能拡張を行うとともに、連携させるべき 2 種類のデータベース(糖タンパク質およびヒト多型)に対して Web サービスを新規に開発し公開した。

(1) 糖タンパク質データベース、GlycoProtDatabase (GPDB)における Web サービスの開発

GPDB からは、産総研・糖鎖医工学研究センターおよび東京都立大学(現首都大学東京)大学院理工学研究科が液体クロマトグラフィー(LC)と質量分析法(MS)を連結した LC/MS 法を基礎としたプロテオーム解析技術により独自に同定した、糖タンパク質に関する糖鎖モチーフや糖鎖修飾位置などの情報が提供されている。この糖タンパク質の情報はタンパク質の翻訳後修飾(Post Translational Modification, PTM)を正確に理解するために有用であり、H-InvDB に格納されている機能ドメイン、多型、細胞内局在などの情報と連結することが望まれる。そこで、GPDB の糖タンパク質の情報を取得するための Web サービスを開発し、GPDB の公開サーバーに Web サービスのインストールを行った。

具体的には、下記項目の開発を行った。

- XML ファイル設計
- データ検索、データ提供を行う Web サービスプログラムの開発(SOAP および REST)

(データ項目)

- | | |
|-------------------------------|------------------|
| ① GPDB ID | ⑥ signal peptide |
| ② organ | ⑦ taxonomy_id |
| ③ RefSeq ID | ⑧ domain |
| ④ glycosylation site location | ⑨ protein name |
| ⑤ organism | ⑩ length |

- ドキュメント、サンプルソースの提供

日本語および英語のドキュメント、および Perl, Python, Ruby, Java のサンプルソースを提供した(図 2.2.1-1)。なお、GPDB Web サービスの設計は、平成 20 年度統合 Web サービス マニフェスト(20081022 版)に準拠して行った。統合 Web サービス マニフェストについては、後述の「2.4.2 節 Web サービスの統合化」で詳しく報告する。

開発した Web サービスは下記 URL から 2010 年 1 月 29 日に公開した。

- GPDB Web サービス

<http://riodb.ibase.aist.go.jp/rcmg/ws/gpdb/>

(2) VarySysDB の Web サービスの開発

VarySysDB は、ゲノム統合プロジェクトによって開発、公開された H-InvDB のサテライトデータベースで、H-Inv 転写産物に関連付けられる形で注釈付けられたヒト多型情報のデータベースである。転写領域やスプライスサイト上の一塩基多型(Single Nucleotide Polymorphism, SNP)、挿入欠失多型(insertion, deletion)、ショートタンデムリピート(Short Tandem Repeat, STR)、単一アミノ酸多型(Simple Amino acid Repeat, SAR)、コピー数多型(Copy Number Variation, CNV)、連鎖不平衡(Linkage Disequilibrium, LD)について、それぞれ H-InvDB の転写産物と機能ドメインに関連づけて、情報整備して公開している。

GPDB Web Service API マニュアル

1. GPDB Web Serviceとは？
2. どのようなサービスを提供しているのか？
3. どうやって利用するのか？
4. APIマニュアル
 - [searchGpdbId](#)
 - [searchTaxId](#)
 - [searchOrganism](#)
 - [searchGlycoprotein](#)
 - [searchDbXref](#)
 - [getXmlEntry](#)
 - [getTextEntry](#)
 - [getDraftEntry](#)
 - [getFastaEntry](#)
 - [getGlycoprotein](#)
 - [getDbXref](#)
 - [getGlycosylationSite](#)
 - [getSequence](#)
 - [getIdList](#)
5. 共通仕様
 - 5.1. 処理時の結果が0件の場合
 - 5.2. 実行中エラーの処理
 - 5.3. SOAP Fault

1. GPDB Web Serviceとは？

GPDB Web Serviceでは、GPDBのデータを利用することができる様々なAPIを提供しています。

APIを利用することにより外部より簡単にGPDBのデータにアクセスできるので、どなたでもGPDB Web Serviceを利用したアプリケーションの開発が可能になります。

4. APIマニュアル

searchGpdbId	
REST Operation	/search/query+gpdb_id[offset.[limit]][/type]
SOAP Operation	searchGpdbId
Description	GPDB_IDを完全一致で検索して、エントリー一覧を返す

SOAP	
Operation	searchGpdbId
WSDL	http://riodb.ibase.aist.go.jp/rcmg/ws/gpdb/services/GPDB.wsdl
WSDL Input	gpdb_id (xs:string), type (xs:string), offset (xs:int), limit (xs:int)
Input Description	gpdb_id: 検索したいgpdb_id type: list or count offset: オフセット値 limit: 取得件数
WSDL Output	return (xs:string)
Output Description	エントリーの詳細情報一覧(XML形式)
Sample Code	searchGpdbId.java searchGpdbId.cgi searchGpdbId.php searchGpdbId.rb

REST	
Operation	search
URI Convention	/search/query+gpdb_id[offset.[limit]][/type]
Input	query, type, offset, limit
Input Description	query: 検索したいGPDB_ID type: list or count offset: オフセット値 limit: 取得件数
Output	xml
Output Description	エントリーの詳細情報一覧(XML形式)
Sample URI	http://riodb.ibase.aist.go.jp/rcmg/ws/gpdb/search/GLP-00005e+gpdb_id/ http://riodb.ibase.aist.go.jp/rcmg/ws/gpdb/search/GLP-00005e+gpdb_id/list http://riodb.ibase.aist.go.jp/rcmg/ws/gpdb/search/GLP-00005e+gpdb_id/count http://riodb.ibase.aist.go.jp/rcmg/ws/gpdb/search/GLP-00005e+gpdb_id/1.5 http://riodb.ibase.aist.go.jp/rcmg/ws/gpdb/search/GLP-00005e+gpdb_id/1 http://riodb.ibase.aist.go.jp/rcmg/ws/gpdb/search/GLP-00005e+gpdb_id/1.5/list http://riodb.ibase.aist.go.jp/rcmg/ws/gpdb/search/GLP-00005e+gpdb_id/1.5/count

図 2.2.1-1 糖タンパク質データベース(GPDB)Web サービスドキュメント

http://riodb.ibase.aist.go.jp/rcmg/ws/gpdb/manual/html/GPDB_Manual_jp.htm

本開発では、ヒト多型情報に関する情報のうち、多型アノテーション情報(SNP, STR, SAR)、ゲノム上の多型位置情報、遺伝子機能情報等に関するデータを検索および取得できる Web サービスを開発した。

具体的には、下記項目の開発を行った。

- XML ファイル設計
- データ検索、データ提供を行う Web サービスプログラムの開発(SOAP および REST 方式)
(検索対象データ項目)

- | | |
|--------------------|---------------------|
| ① HIT ID | ⑥ InterProID |
| ② definition | ⑦ chromosome number |
| ③ HIX ID | ⑧ OMIM ID |
| ④ dbSNP ID | ⑨ chromosome band |
| ⑤ accession number | ⑩ OMIM name |

(提供データ項目)

① 多型アノテーション情報

- Polymorphism classification
 - allele (G/T など)
 - feature (SNP, DIP, validation status)
 - region (promoter, 5' -UTR, CDS, 3' -UTR, splice site)
 - type (Nonsynonymous, Synonymous, Nonsense, Unclassified)
- Polymorphism effect on annotation
 - gain/loss of functional domain (InterPro)
 - gain/loss of protein 3D structure (SCOP)
 - OMIM allelic variant
- STR repeat
 - microsatellite (repeat unit, count)
 - single amino acid repeat (repeat unit, count)
- CNV variation
 - division (A, B, C, D, E, F)
 - class (CNV, Inversion)

② ゲノム上の多型位置情報

- Chr. Number, strand, location (start, end)
- Chr. Band (start, end)

③ 遺伝子機能情報

- Representative transcript (yes, no)
- Definition
- Similarity category

- ドキュメント、サンプルソースの提供

日本語および英語のドキュメント、および Perl, Python, Ruby, Java のサンプルソースを提供した(図 2.2.1-2)。なお、VarySysDBWeb サービスの設計は、平成 20 年度統合 Web サービス マニフェスト(20081022 版)に準拠して行った。統合 Web サービス マニフェストについては、後述の「2.4.2 節 Web サービスの統合化」で詳しく報告する。開発した Web サービスは下記 URL から 2009 年 11 月 13 日に公開した。

- VarySysDBWeb サービス

<http://www.h-invitational.jp/hinv/hws/varysysdb/>

[\[Top \]](#) [\[English \]](#)

1. [イントロダクション](#)
2. [RESTサービス](#)
3. [SOAPサービス](#)
4. [REST API一覧](#)
5. [SOAP API一覧](#)
6. [サンプルプログラム](#)

イントロダクション

VarySysDB Web serviceでは、VarySysDBのデータを利用することができる様々なAPIを提供しています。
 どのたでもVarySysDB Web serviceを利用したアプリケーションの開発が可能になります。
 VarySysDB Web serviceでは、REST-API / SOAP-APIを提供しています。

RESTサービス

REST APIを使用するには、HTTPコネクションを使用できる環境が必要になります。
 また、ブラウザや、PerlやJavaのようなプログラミング言語が必要です。
 本サービスにアクセスするには、
<http://h-invitational.jp/hinv/hws/varysysdb/search/hogehoge+definition/>
 のように <http://h-invitational.jp/hinv/hws/varysysdb/> に続けてメソッド名、パラメータを渡します。

SOAPサービス

SOAP APIを使用するには、各プログラミング言語のSOAP対応ライブラリをご使用ください。
 WSDL文書は
<http://h-invitational.jp/hinv/hws/varysysdb/API名?wsdl>
 で参照いただけます。

REST API

Polymorphismデータ取得

No.	概要	URL
1	検索キー(*1)及び検索クエリー(*2)を指定してRS_NUMBERを取得する	http://h-invitational.jp/hinv/hws/varysysdb/search/606953+omim_id/ http://h-invitational.jp/hinv/hws/varysysdb/search/FAM39B+definition/ http://h-invitational.jp/hinv/hws/varysysdb/search/HIT000022181+hit/ http://h-invitational.jp/hinv/hws/varysysdb/search/HIX0000001+hix/
2	検索キー及び検索クエリーを複数指定してRS_NUMBERを取得する	http://h-invitational.jp/hinv/hws/varysysdb/search/HIX0000001.HIX0000002.HIX0000003+hix/
3	位置情報を指定する(6番染色体の1-1000000)	http://h-invitational.jp/hinv/hws/varysysdb/search/6.1-1000000+map/
4	データ取得開始位置、取得件数を指定する(0番目から100個) 開始は0とします	http://h-invitational.jp/hinv/hws/varysysdb/search/6.1-1000000+map/0.100/
5	データ件数を取得する	http://h-invitational.jp/hinv/hws/varysysdb/search/BC053657+accession/count/
6	RS_NUMBERを指定してデータを取得する	http://h-invitational.jp/hinv/hws/varysysdb/entry/rs1578391/
7	RS_NUMBERを指定してマッピングデータを取得する	http://h-invitational.jp/hinv/hws/varysysdb/entry/rs1578391+map/
8	RS_NUMBERを指定して Genomic Polymorphismデータを取得する	http://h-invitational.jp/hinv/hws/varysysdb/entry/rs1578391+genomic_polymorphism/
9	RS_NUMBERを指定して Polymorphism Featureデータを取得する	http://h-invitational.jp/hinv/hws/varysysdb/entry/rs1578391+polymorphism_feature/
10	RS_NUMBERを指定して Polymorphism Effectデータを取得する	http://h-invitational.jp/hinv/hws/varysysdb/entry/rs1578391+polymorphism_effect/
11	RS_NUMBERを指定して Transcript Polymorphismデータを取得する	http://h-invitational.jp/hinv/hws/varysysdb/entry/rs1578391+transcript_polymorphism/

図 2.2.1-2 ヒト多型データベース(VarySysDB)Web サービスドキュメント

http://www.h-invitational.jp/hinv/hws/varysysdb/doc/index_ja.html

2.2.2 ウェブサービスの拡張

データベース連携の基盤となる H-InvDB の Web サービスについて、平成 20 年度に文部科学省と共同で作成した「主要国内 DB に共通な Web サービス開発ガイドライン」(H20 版マニフェスト)に準拠した機能拡張を行った。

具体的には、H-InvDB の Web サービスから公開しているそれぞれ約 30 種の API (REST および SOAP 方式) について、平成 20 年度統合 Web サービス マニフェスト(20081022 版)に準拠したインターフェースの改良を行った。SOAP は Simple Object Access Protocol の略で Web サービスを支える基盤規格、REST は Representational State Transfer の略で HTTP を使って通信を行う手法である。統合 Web サービス マニフェストについては、後述の「2.4.2 節 Web サービスの統合化」で詳しく報告する。

開発した Web サービスは下記 URL から 2010 年 3 月公開予定である。

- H-InvDB Web サービス

<http://h-invitational.jp/hinv/hws/>

2.2.3 データ品質の評価

平成 20 年度経済産業省統合データベースプロジェクトでは、ヒト全遺伝子のアノテーション統合データベース H-InvDB を基礎としてデータ連携のシステムを構築し、データ統合を進めた。具体的には、H-InvDB のヒト遺伝子情報と糖転移酵素 (GGDB) と機能性 RNA (fRNAdb) の精査されたデータを Web サービスをもちいて統合した。この際に「高品質であると評価したデータを公開する」ことを重要な方針とし、データの統合にあたっては H-InvDB のヒト遺伝子情報を基準データとした。

一方、選択的スプライシングとゲノム多型はヒトの転写産物やタンパク質の構造と機能に大きな影響を与えることが知られているが、これらの情報は近年急速に増えている。そのため、ヒト分子情報の統合データの品質を保つためには、H-InvDB と GGDB と fRNAdb の情報を統合する際に選択的スプライシングとゲノム多型の影響を十分に考慮し、最新の情報を用いた品質評価をすることが必要であった。そこで平成 21 年度は、統合化されたデータが正しい情報になっているかどうかを評価すべく、ゲノム多型と選択的スプライシングの観点からヒト分子統合データの品質評価を行った。

(1) VarySysDB の最新データへの更新

ゲノム多型についての影響を評価するために、ゲノム統合プロジェクトによって構築されたヒトゲノム多型に関するデータベース VarySysDB を利用し、そこに登録されたヒトゲノム上の一塩基多型 (Single Nucleotide Polymorphism, SNP)、挿入欠失多型 (insertion, deletion)、ショートタンDEMリピート (Short Tandem Repeat, STR)、単一アミノ酸多型 (Simple Amino acid Repeat, SAR)、コピー数多型 (Copy Number Variation, CNV) の情報を最新に更新して H-InvDB と fRNAdb の統合データの品質評価に利用した。

具体的には、2010 年 2 月にリリースされた H-InvDB リリース 7.0 の構築と並行して、ヒト遺伝子の多型情報を解析し、VarySysDB のデータ更新を行った。データ解析は、独自に開発された高精度自動配列アノテーションツール SuperTACT を利用して PC クラスターにより実施した。

更新した VarySysDB は下記 URL から 2010 年 3 月公開予定である。

- VarySysDB

<http://h-invitational.jp/varygene/home.htm>

(2) H-DBAS の最新データへの更新

選択的スプライシングの影響を評価するために、「ゲノム情報統合プロジェクト」で構築されたヒトの選択的スプライシングのデータベース H-DBAS の情報を最新データに更新して、H-InvDB と GGDB の統合データの品質評価に利用した。

具体的には、2010 年 2 月にリリースされた H-InvDB リリース 7.0 の構築と並行して、ヒト遺伝子の選択的スプライシング情報を解析し、H-DBAS のデータ更新を行った。データ解析は、独自に開発された高精度自動配列アノテーションツール SuperTACT を利用して、PC クラスターにより実施した。

更新した H-DBAS は下記 URL から 2010 年 3 月公開予定である。

- H-DBAS: http://www.h-invitational.jp/h-dbas/simple_search_jp.jsp

(3) H-InvDB と GGDB の統合データ品質評価

平成 20 年度において H-InvDB と GGDB のデータ対応表作成を行い、データベース連携を行った。H-InvDB と GGDB のデータ対応表作成は、H-InvDB の HIT ID と GGDB の Gene symbol を HUGO gene symbol をキーに連携させた結果、GGDB 190 遺伝子中 188 遺伝子が HUGO gene symbol 経由で H-InvDB 1,465 遺伝子座 (HIX) に対応し、残り 2 遺伝子について未対応であった。この H-InvDB と GGDB の統合データ品質評価を行うため、連携したデータ対応表の検証を行った。

具体的には、対応した全遺伝子について連携内容の確認を行い、また未対応の遺伝子について正しいデータ連結を行った。まず、対応した全 1,465 遺伝子座 (HIX) について、下記の項目で連携内容の確認を行い、正しいデータ連結が行われたことを検証した。

(連携検証実施項目)

- ① 遺伝子の機能的 description の整合性
- ② 遺伝子発現情報の整合性
- ③ 外部データベース対応の整合性

次に、未対応の 2 遺伝子について、正しいデータ連結を行うよう下記のようなデータ対応表の修正を行った。

(データ修正内容)

- ① 対応付けの幅を広げ、previous gene symbol, aliases gene symbol 等の情報を元に対応付けを修正。
- ② スプライシング・バリエーションによって異なる糖転移酵素であることが判明している Glucosaminyl (N-acetyl) transferase 2,3 等について手動で対応付けを修正。手動修正の例を図 2.2.3-1 に示す。

例) Glucosaminyl (N-acetyl) transferase 2,3

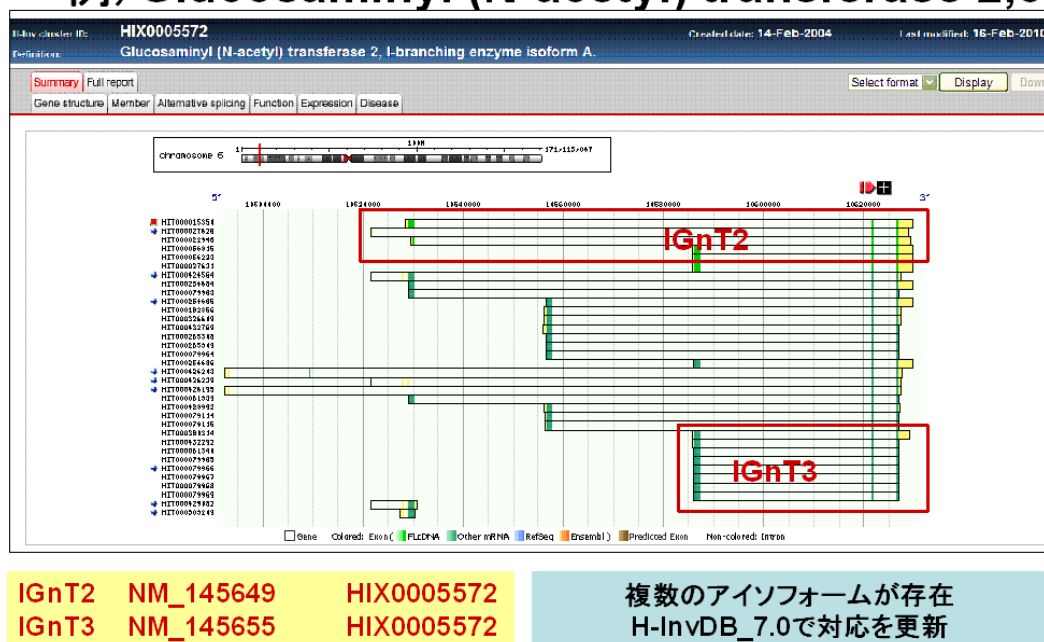


図 2.2.3-1 H-InvDB と GGDB の統合データ品質評価例 (HIX0005572)

http://h-invitational.jp/hinv/spsoup/locus_view?hix_id=HIX0005572

(4) H-InvDB と fRNAdb の統合データ品質評価

平成 20 年度において H-InvDB と fRNAdb のデータ対応表作成を行い、データベース連携を行った。H-InvDB と fRNAdb のデータ対応表作成は、H-InvDB の HIT ID と fRNAdb の FR ID とゲノム上のエクソン位置の 1bp 以上の重なり(センス、アンチセンス)で行った結果、fRNAdb の FR ID 23,931 件と H-InvDB HIT 45,495 件が対応した。この H-InvDB と fRNAdb の統合データ品質評価を行うため、連携したデータ対応表の検証を行った。

具体的には、H20 年度データ対応表で対応した fRNAdb の FR ID 23,931 件と H-InvDB HIT 45,495 件について、下記の項目で連携内容の確認を行い正しいデータ連結が行われたことを検証した。

(対応済遺伝子の特徴抽出)

- ① Overlap (CDS, 5' -UTR, 3' -UTR)
- ② Overlap (exon, intron)
- ③ Strand (sense, anti-sense)
- ④ Sequence ontology (SO)

Sequence ontology (SO)とは、生物学的な配列に対するオントロジーで、ジーンオントロジー(Gene Ontology, GO)に関する機関とゲノムアノテーションセンターの共同でプロジェクトが進められている。

Sequence Ontology Project: <http://www.sequenceontology.org/>

H-InvDB と fRNAdb の対応済遺伝子の特徴抽出の上記 4 パターンの説明を図 2.2.3-2 に、特徴抽出の例を、図 2.2.3-3、図 2.2.3-4 に示す。

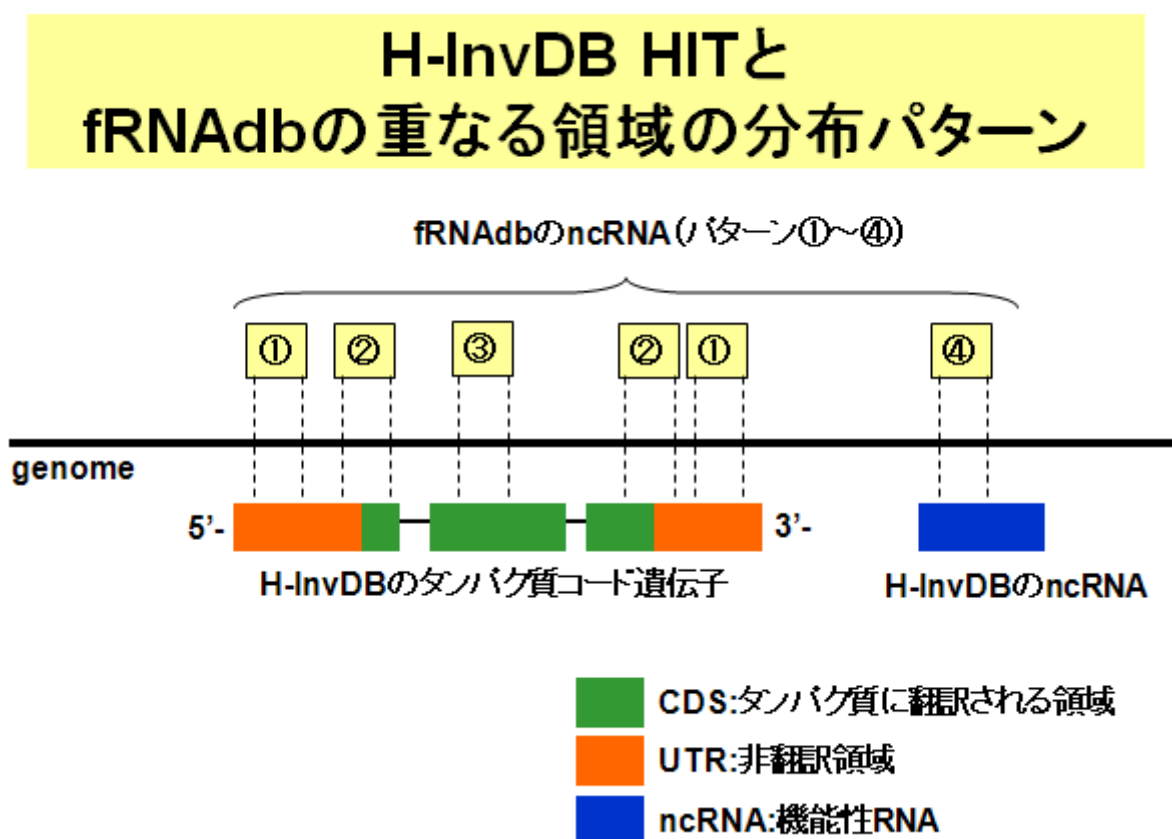


図 2.2.3-2 H-InvDB と fRNAdb の対応遺伝子の重なり領域の分布パターン
fRNAdb ncRNA がアンチセンスでオーバーラップする H-InvDB 転写産物(HIT)の exon 領域

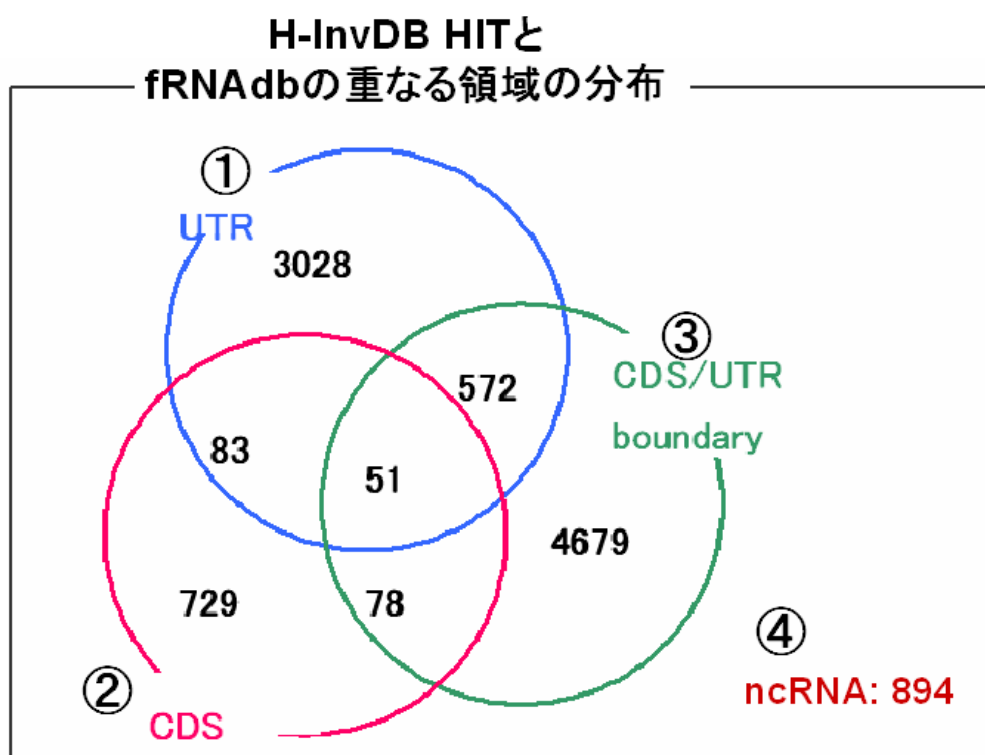


図 2.2.3-3 H-InvDB と fRNAdb の対応遺伝子の特徴抽出例 1

fRNAdb ncRNA がアンチセンスでオーバーラップする H-InvDB 転写産物(HIT)の exon 領域の分布

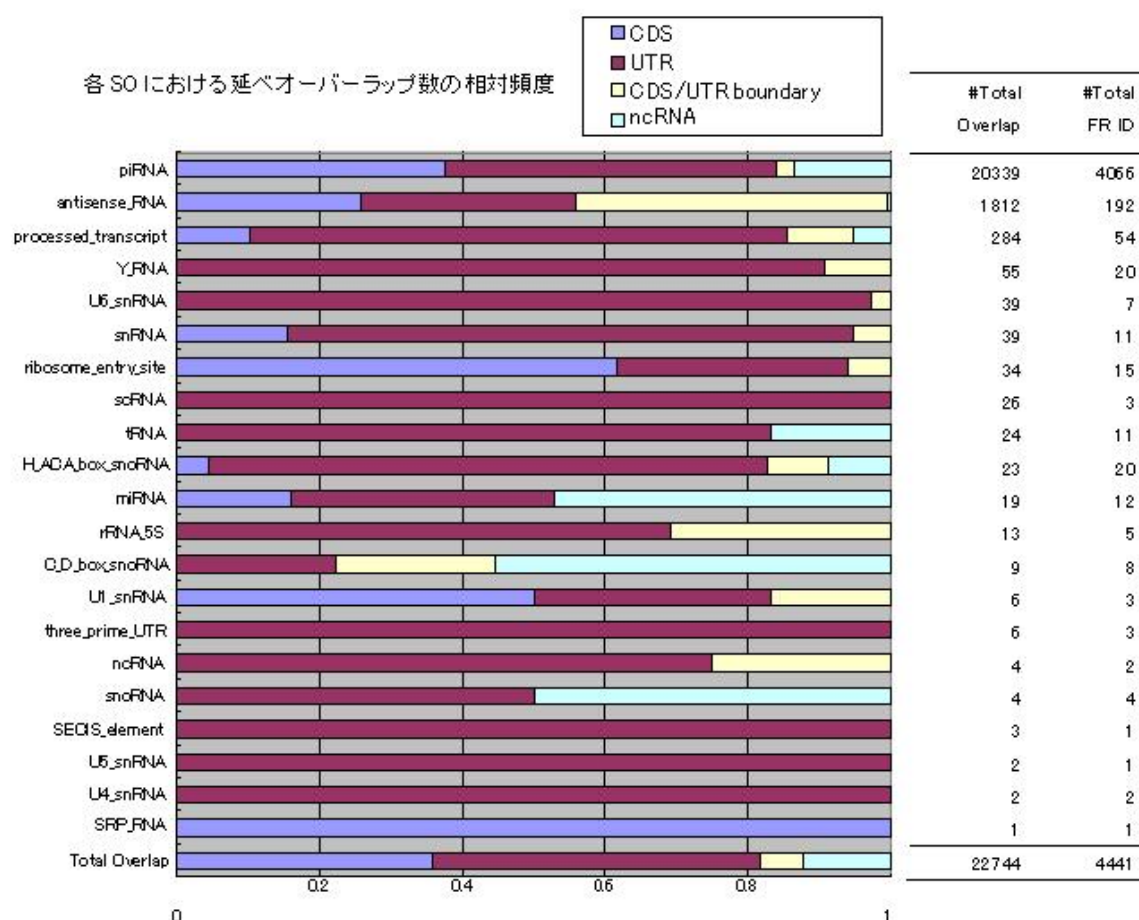


図 2.2.3-4 H-InvDB と fRNAdb の対応遺伝子の特徴抽出例 2

fRNAdb ncRNA と H-InvDB 転写産物(HIT)がオーバーラップする exon 領域と Sequence ontology (SO)の分布

2.2.4 画面の改良

H-InvDB、GPDB、VarySysDB において開発・公開した Web サービスによって、これらのデータベース間でリアルタイムの相互のデータ提供が可能となった。そこで、Web サービスの仕組みを通して、H-InvDB、GPDB、VarySysDB のデータをリアルタイムで表示し、相互にリンクするための画面の改良を H-InvDB に対して実施した。H-InvDB の画面改良について、具体的には下記項目の開発を行った。

- GPDB データの表示とリンク設置

下記データ項目を H-InvDB に新規に開発したタンパク質アノテーション情報表示画面 (Protein view) に Web サービスを介して表示しリンクを設置するようシステム開発を行った。(図 2.2.4-1 および図 2.2.4-2) (データ表示項目)

- | | |
|-------------------------------|----------------|
| ① GPDB ID | ⑥ domain |
| ② glycosylation site location | ⑦ protein name |
| ③ organism | ⑧ length |
| ④ signal peptide | ⑨ organ |
| ⑤ taxonomy_id | |

開発した画面の改良は、2010 年 2 月 16 日に H-InvDB_7.0 より公開した。

- VarySysDB データの表示とリンク設置

下記データ項目を H-InvDB 画面 (Protein view) に Web サービスを介して表示した。(図 2.2.4-2)

- ① HIP ID
- ② Synonymous SNP
- ③ Nonsynonymous SNP
- ④ Simple amino acid repeat (STR)

開発した画面の改良は、2010 年 2 月 16 日に H-InvDB_7.0 より公開した。

H-InvDB Associated Human Gene Database

5th Anniversary !! H-InvDB_7.0 released on February 16, 2010.

Search by Keyword for GO Advanced Search

ホーム クイックガイド 検索ナビ BLAST サイトマップ データダウンロード 問い合わせ ヘルプ

Transcript view Locus view G-integra DiseaseInfo Viewer H-ANGEL Evola PPI view Gene Family/Group

H-Inv protein ID: **HIP000199646** Last modified: 16-Feb-2010

Definition: **Endothelin converting enzyme 2 isoform B.**

Summary Full report

Protein Info. Member Motif Function PTM Subcellular loc. Protein structure Evolution Polymorphism

Protein information	
HIP ID	HIP000199646
Length	736
Codon Adaptation Index (CAI)	0.824
Database links	RefSeq: NM_002808 ; NM_014693 ; NM_032331 ; NM_001037324 ; NM_001100120 ; NM_001100121
	UniProt: A5PLK8
	CCDS: NA

Original transcript information	
Representative H-Inv transcript ID	HIT000435444
H-Inv cluster ID	HIX0003914
Predicted CDS	144..2354; 736[aa]; Orientation: +3;
Genomic location	Chromosome 3 Location 3q27.1 CDS position 183993799-184010819 Strand +
Accession number	BC142950.1

HIP000199646 GLP-00014a *Caenorhabditis elegans*

1 110 210 310 410 510 610 710 737

10 kaa

● Glucosylation site
● Signal peptide
■ Transmembrane (SOSUI)
| Single amino acid repeat (SAR)
| Synonymous SNP
| Nonsynonymous SNP

ドメイン、細胞内局在、立体構造などヒトタンパク質に対するアノテーション情報を提供

図 2.2.4-1 H-InvDB 画面の新規メイン画面 Protein view

ウェブサービスを介して、多型、糖タンパク質データ表示とリンク設置

The screenshot displays the H-InvDB Protein view for the protein HIP000199646, Endothelin converting enzyme 2 isoform 0. The interface includes a search bar, navigation tabs (Summary, Full report, PTM, Subcellular loc., Protein structure, Evolution, Polymorphism), and a detailed Glycosylation section. The Glycosylation section shows a link to GPDB (GlycoProtDB) with the ID GLP-00014a, protein name protease: T05A8.4, organism mixed, and length 736. A blue arrow points to the '糖タンパク質情報' (Glycoprotein information) section. Below this, a diagram of the protein sequence (1-736) shows a glycosylation site (red dot) at position 74, labeled '糖鎖修飾部位' (Glycosylation site). A legend indicates that the red dot represents a glycosylation site, the blue dot represents a signal peptide, and the green dot represents a transmembrane domain (TMD). A yellow callout box points to the 'PTM' tab, stating '翻訳後修飾情報セクション PTMタブ' (Post-translational modification information section PTM tab). Another yellow callout box points to the GlycoProtDB link, stating 'GlycoProtDBへのリンクと詳細 (ウェブサービスで取得)' (Link to GlycoProtDB and details (obtained via web service)). A third yellow callout box points to the amino acid polymorphism section, stating 'アミノ酸配列の多型情報 (VarySysDBウェブサービスで取得)' (Amino acid sequence polymorphism information (obtained via VarySysDB web service)).

図 2.2.4-2 H-InvDB 画面への GPDB 及び VarySysDB データ表示

2. 2. 5 分子情報と実験用リソース情報の統合

分子情報と実験用リソースの情報を結合することにより、双方の利用者に大きな利便性をもたらすことが期待される。そこで、本テーマでは、ヒトの分子情報を対象として両者の情報の統合をすすめた。

独立行政法人製品評価基盤機構(NITE)生物遺伝資源部門の NITE Biological Resource Center (NBRC)では、実験等で用いるリソースの分譲業務を行っており、その対象にはヒトの cDNA クローンも含まれる。NBRC にて配布中のヒト関連リソースとしては、完全長ヒト cDNA クローンおよび Splice-Variant ヒト cDNA クローンがある。これらはいずれも、NEDO「完全長 cDNA 構造解析プロジェクト」および「タンパク質機能解析・活用 プロジェクト:スプライシング・バリエーションの取得技術の開発」による成果である。

本テーマでは、まず、NBRC が配布しているヒト cDNA クローンの情報をヒト遺伝子統合データベース H-InvDB に統合することにより、さまざまな分子情報を確認しながら必要な実験用リソースを発注できるしくみを整備した。そのために、以前より開発を進めているリンク自動管理システムを機能拡張し、NBRC から配布クローンの一覧表を毎日自動でダウンロードして、クローン ID を他のヒト分子に関する ID と対応づけるようにした。これを用いて、H-InvDB から NBRC のクローン詳細ページへ(図 2.2.5-1)、また、NBRC のクローン詳細ページから H-InvDB へ(図 2.2.5-2)のリンクを設置した。同時に、ヒト cDNA クローンの詳細情報を提供する HGPD (<http://www.hgpd.jp/>)にもリンクを張った。2010 年 3 月 12 日時点で、H-InvDB と NBRC の間では 55,405 件のデータがリンクされている。

以上をまとめると、ヒト遺伝子の統合データベース H-InvDB と、NBRC が管理・配布するヒト cDNA クローンの情報を、リンク自動管理システムにより連結した。分子に関する情報と実験リソースの情報を統合したことにより、研究者にとって便利になるだけでなく、情報検索から実験開始までのタイムラグの短縮が期待される。このようなリソース情報の統合化は、cDNA クローンだけでなく、抗体や化合物などに対して実施することが有効と思われる。

H-InvDB 5th Anniversary!! H-InvDB 7.0 released on February 16, 2010.

Transcript view

Search by Keyword for GO Advanced Search

Home Quick guide Navi BLAST Site map Download Contact us Help

Locus view Protein view G-Integra DiseaseInfo Viewer H-ANGEL Evola PPI view Gene Family/Group

H-InvDB ID: HIT000425669 Accession number: AK292872 Created date: 18-Dec-2008 Last modified: 16-Feb-2010

Definition: CMP-N-acetylneuraminase-poly-alpha-2,8-sialyltransferase; EC=2.4.99.-; Alpha-2,8-sialyltransferase 8D; ST8Sia IV; Polysialyltransferase-1;

Summary Full report

Transcript info Gene structure CDS Motif Function GO Subcellular localization Evolution Polymorphism

Select format Display Download

Transcript original information

Accession number	AK292872.1
CAGE tag ID	NA
EST ID	NA
Clone Number	TRACH3013141
Experimental resources	NBRC; HGPD; Antibody
Sequence data provider	NA
Annotation project	NA
Length of cDNA	1575(bp) (No. of exons: 4:485 T:398 G:368 C:324)
Division	HUM
Molecular type	mRNA
Cell type	NA
Tissue type	trachea
Developmental stage	NA
Library origin	NA

Sequence quality information

CDS feature	Complete CDS
Kozak sequence	NA

CAAGGTGCTCATCTGAAAAGAAACCTGAGCCCCAGGGAGGCGGCGCGG
GCGACCCCTGGCAGAGCTGGCGCAACAGGGCGAGAGGTGCTGGGCAGC
GTTCGAGGACCAGGGAGCTCGGCCACAGAAGACCCAGTGATCTGATC

図 2.2.5-1 H-InvDB リリース 7.0 の Transcript view の Transcript info というタブの画面

ここには新たに NBRC ヒト完全長 cDNA クローン検索ページへのリンクが設置され、cDNA クローンの注文がスムーズに行えるようになった。このリンクの設置にはリンク自動管理システムを利用している。

nite Incorporated Administrative Agency National Institute of Technology and Evaluation

[Site Map] [Links] [Japanese]

NBRC Biotechnology Field
Exploring the power of Microorganisms for the Development of Bioindustries

Microorganisms

- Online Catalogue
- Alphabetical Index
- ISQJIS/Pharmacopoeia
- Test Strains
- Ordering of Cultures

DNA Resources

- Microbial Genomic DNAs
- Microbial Genomic DNA Clones
- Human cDNA Clones
- Human Gateway® Entry Clones

Deposit

- Deposit
- Others
- Price
- Request for Certificate
- List of Media
- Contact Us
- Download

Biological Resource Center

Distribution of Full-length Human cDNA Clones

[Clone Search] [MTA Search] [MTA & Ordering Form Download]

Sequence ID Search Result

Back Query: FEKID2000146 Found 1 entries. How to order

Accession No. FLJ No. or PSEC No. Clone ID

AK290561 FLJ76020 FEKID2000146 (H-inv: HIT000423358)

Prior to ordering cDNA clones of your interest, please click the button below and check kinds of MTA

MTA Search

図 2.2.5-2 NBRC におけるヒト完全長 cDNA クローン検索ページ

リンク自動管理システムを利用することにより、H-InvDB の Transcript view へリンクが設置された。

ここで、リンクの設置に用いたリンク自動管理システムの機能拡張について、補足説明をする。リンク自動管理システムでは、世界の主要なデータベースの間のリンク情報を自動で管理するため、データ ID の対応関係を調べ、データ間の対応表を作成している(図 2.2.5-3)。このシステムに NBRC のヒト cDNA クローン情報を追加し、H-InvDB はもちろん、世界の主要なデータベースとの対応関係を解析して、その情報をホームページで提供するようにした。また、データ ID を変換する ID 一括変換システムや、データの対応表をダウンロードする機能も公開した。

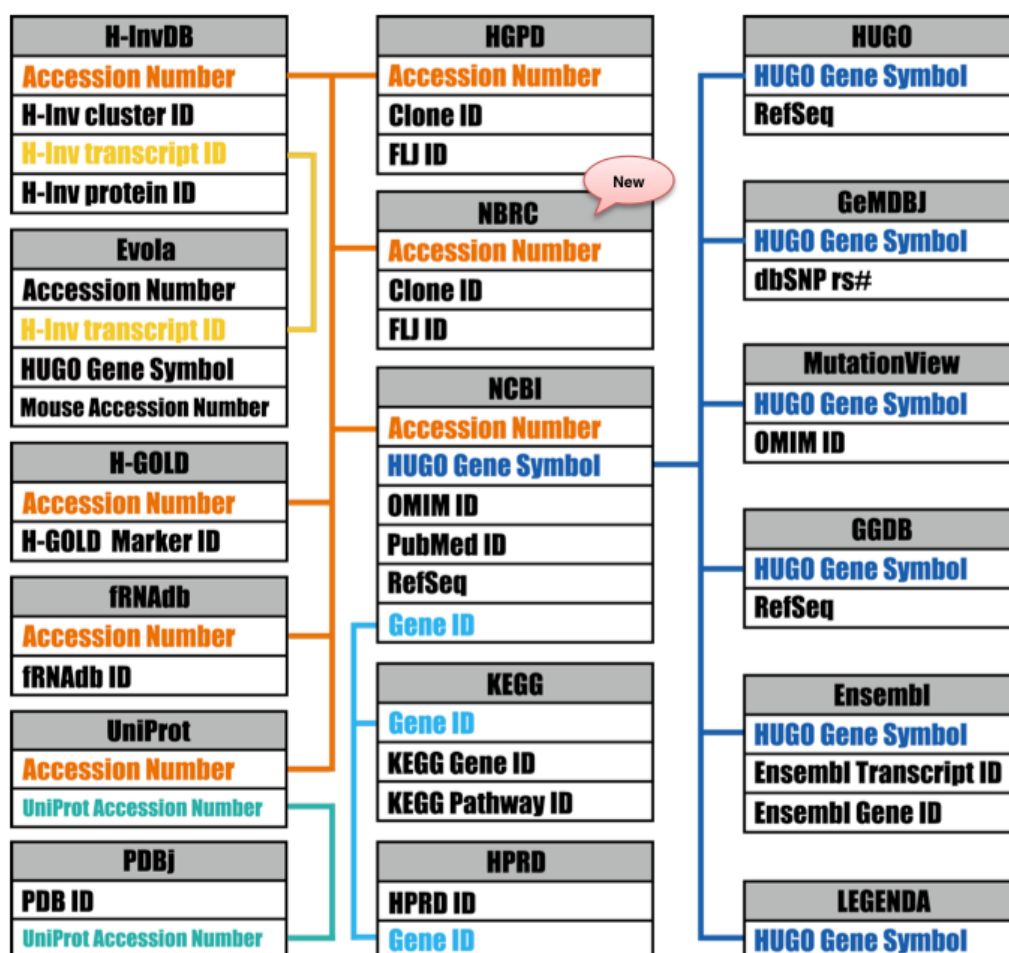


図 2.2.5-3 リンク自動管理システムの拡張

本システムは、色をつけたデータ識別子(ID)を共通 ID と位置づけ、データベース間のデータ対応関係を自動で解析することにより、世界の主要なデータベースの間のリンク情報を管理している。NBRC は、2010 年 1 月の更新により本システムの対象データベースに追加された。

さらに、All Database 検索の機能を新たに作成して公開した(図 2.2.5-4)。これは、ヒト分子に関するいずれかのデータIDをクエリとして、世界の 24 種類のデータベースが持つ対応するデータへのリンクを一括して表示する検索サービスである。24 種類のデータベースは、遺伝子、転写産物、タンパク質、機能、疾患・多型、文献の6つの分野に分けて結果が表示される。この機能はデータIDによる世界の主要なデータベースへの横断検索であり、統合データベースの新たなサービスとして役立つ可能性がある。

Hyperlink management system

Accession Number [All Databases]

遺伝子カタログ	転写産物カタログ	タンパク質カタログ
H-invDB (Locus view) BC053657 HX00005084 H-invDB (G-Integra) BC053657 HX00005084 NCBI (Nucleotide) BC053657 NM_055686 BC053657 NM_175052 NCBI (Entrez Gene) BC053657 7903 HUGO BC053657 GT891A4 Ensembl Human GeneView BC053657 ENSG00000113532 KEGG Gene BC053657 7903	H-invDB (Transcript view) BC053657 HT000053061 H-invDB (Evoia) BC053657 HT000053061 NCBI (GenBank) BC053657 BC053657 Ensembl Human TransView BC053657 ENST00000231481 BC053657 ENST00000451528 NBRC BC053657 No link HGPD BC053657 FLJ82330	H-invDB (PPI view) BC053657 HBP000084442 UniProt BC053657 Q82167 POBJ BC053657 No link HPRD BC053657 23870
遺伝子機能 KEGG Pathway BC053657 hsa0604+7903 GGDB BC053657 GT891A4 trnAdb BC053657 No link	疾患・多型 NCBI (OMIM) BC053657 602547 GenMDBJ BC053657 RMS-J0T080484 BC053657 RMS-J0T109485 BC053657 RMS-J0T109486 BC053657 RMS-J0T131844 BC053657 RMS-J0T131846 BC053657 RMS-J0T131847 BC053657 RMS-J0T131850 BC053657 RMS-J0T131854 BC053657 RMS-J0T131855 BC053657 RMS-J0T131860 BC053657 RMS-J0T131861 BC053657 RMS-J0T131863 BC053657 RMS-J0T131864 BC053657 RMS-J0T131865 BC053657 RMS-J0T131867 BC053657 RMS-J0T131868 BC053657 RMS-J0T131871 BC053657 RMS-J0T131873 MutationView BC053657 No link H-GOLD BC053657 No link	文 献 LEGENDA BC053657 7903 NCBI (PubMed) BC053657 16660622 BC053657 10766705 BC053657 11279095 BC053657 11744834 BC053657 12138100 BC053657 12227054 BC053657 12477932 BC053657 15488034 BC053657 15710344 BC053657 16027151 BC053657 16229622 BC053657 16344560 BC053657 17178603 BC053657 18384787 BC053657 18648501 BC053657 19165232 BC053657 19338400 BC053657 19578676 BC053657 7624364 BC053657 7854457 BC053657 9254414 BC053657 9774403 BC053657 9852130

図 2.2.5-4 リンク自動管理システムにおける「All Database 検索」の実行例

Accession Number の BC053657 を指定することにより、世界の 24 種類のデータベースが持つ対応するデータへのリンクを一度に表示している。この例でわかるように、All Database 検索はIDによる主要なデータベースの横断検索を実現した。

2.3 国内外の最新動向調査及びアンケート調査

2.3.1 国内外の最新動向調査と広報・普及活動

(1) 国内外の動向調査

経済産業省ライフサイエンスデータベース・ポータルサイト MEDALS に関する学会発表または聴講を目的として、国際学会に 6 人回、国内学会等に 47 人回参加した(表 2.3.1-1)。

表 2.3.1-1 参加した国際および国内学会等

日付	会議名	開催場所
2009 年 5 月 12 日	第 3 回 糖鎖産業技術フォーラム[GLIT] 「再生医療における糖鎖の利用と重要性 － 評価技術と政策 －」	お台場 (産業技術総合研究所 臨海副都心センター)
2009 年 6 月 12 日	統合データベースシンポジウム 2009 「データベースが拓くこれからのライフサイエ ンス」	東京 (東京大学 浅野キャンパス)
2009 年 6 月 27 日-7 月 2 日	17 th Annual International Conference on Intelligent System for Molecular Biology [ISMB 2009]	ストックホルム (Stockholmsmässan/Stockholm International Fairs)
2009 年 7 月 1-3 日	第 8 回 国際バイオ EXPO	お台場(東京ビッグサイト)
2009 年 9 月 2-4 日	第 11 回 日本進化学会大会	札幌(北海道大学)
2009 年 10 月 7-9 日	Bio Japan 2009	横浜(パシフィコ横浜)
2009 年 10 月 30 日	JBIC2009 プロジェクト研究成果報告会	品川 (東京コンファレンスセンター)
2009 年 12 月 9-12 日	第 32 回日本分子生物学会年会	横浜(パシフィコ横浜)
2010 年 2 月 4-5 日	第 9 回 産総研・産技連 LS-BT 合同研究発 表会	産業技術総合研究所 つくばセンター共用講堂
2010 年 2 月 8 日	国際開発会議 DBCLS バイオハッカソン 2010	東京大学 浅野キャンパス・ お台場(産業技術総合研究所 臨海副都心センター)
2010 年 3 月 15-19 日	Xgen Congress	サンディエゴ(HILTON SAN DIEGO RESORT)
2010 年 3 月 15-18 日	Cloud Connect	サンタクララ(Santa Clara Convention Center)

(2) 学会発表およびその機会を利用した広報活動

学会等において MEDALS およびヒト遺伝子統合データベース H-InvDB に関連するポスター発表・口頭発表を行い、その機会を利用してデモンストレーションやパンフレット配布等の広報活動を実施した。また、JBIC 成果報告会においてプロジェクトの成果発表を行った。

○2009 年 5 月 12 日 第 3 回 糖鎖産業技術フォーラム [GLIT] 「再生医療における糖鎖の利用と重要性 - 評価技術と政策 -」(お台場)

国内の学術動向と産業化を踏まえた情報収集を目的に、糖鎖研究動向について調査し、MEDALS のポスター発表を行った。糖鎖はほぼ全ての生命現象にキーファクターとして関わることが明らかになってきた。臓器移植の際の拒絶反応には糖鎖の関連が疑われているため、糖鎖が免疫の関連を解明することが重要にな

る。日本の糖鎖研究は世界的にもトップを走っており、世界的にも糖鎖研究を促進する動きがあるため、優位を保っていく必要がある。

○2009 年 6 月 12 日 統合データベースシンポジウム 2009「データベースが拓くこれからのライフサイエンス」(東京大学浅野キャンパス)

MEDALS および H-InvDB について、1 件の口頭発表と 4 件のポスター発表を行った。口頭発表会場はほぼ満席で、データベース開発に関わる多くの人に MEDALS のアピールができた。また、ポスター発表会場では活発な議論が行われた。また、会場では朝から夕方までセッション発表が行われた。その中で、ライフサイエンスデータベースの現状の問題点について「恒久的なナショナルセンターが無く、公的資金で取得したデータの公開の取り扱いが不明確である。」という意見や、「データベースの散在・成果の公開が不十分・維持、アノテーションを行える人材の不足」という意見が出された。その他に恒久的な統合データベースセンターが満たすべき必要条件として①継続性、②情報の創出と利用を反映できる運営、③業務を通しての人材育成、④関係省庁の支援が挙げられていた。

○2009 年 6 月 27 日-7 月 2 日 17th Annual International Conference on Intelligent System for Molecular Biology [ISMB 2009](ストックホルム)

MEDALS や H-InvDB の改良や応用利用への可能性を探るため、バイオ系データベースと最近のバイオデータベース研究の世界的動向について調査した。

統合データベースに関する情報では、欧州のバイオデータベース統合のための基盤整備を行おうとしている ELIXIR プロジェクトでは欧州データベース開発者へのアンケート調査を報告していた。今後のバイオインフォマティクス基盤についての課題を次の 3 つにまとめていた。①データやツールの統合にむけた開発、②データベースの機能の改良(インターフェースやそれに関連したツールなど)、③データの品質。この調査結果から、これらの課題の解決を目指す欧州のバイオインフォマティクス基盤について、今後の戦略を練る方針が得られたということであった。

○2009 年 7 月 1-3 日 国際バイオ EXPO(お台場)

MEDALS および H-InvDB について 2 件の口頭発表と 2 件のポスター発表を行った。ポスター展示ブースではデモンストレーションを行い、H-InvDB オリジナルデザインのエコバッグとパンフレットを合わせて配布した。オリジナルデザインのエコバッグは大変好評であった。

【配布物】

- MEDALS パンフレット: 241 部(日本語版 232 部、英語版 9 部)
- H-InvDB パンフレット: 243 部(日本語版 233 部、英語版 10 部)
- エコバッグ: 185 枚

○2009 年 9 月 2-4 日 第 11 回 日本進化学会大会(札幌)

MEDALS および H-InvDB についてブース出展と、関係するポスター発表を 4 件行った。ブース出展では MEDALS と H-InvDB のポスターを展示し、新たに作成した MEDALS と H-InvDB のオリジナルデザインのエコバッグをパンフレットとともに配布した。出展ブースはセミナー会場のそばにあったため、多くの関係者に配布・アピールすることができた。2009 年 8 月 4-18 日の期間で実施した Web アンケートでは日本進化学会の会員にも通知していたため、重ねてアピールするのに良いタイミングであった。また、要旨集には MEDALS の広告を掲載し、多くの方にアピールすることができた。

【配布物】

- MEDALS パンフレット: 日本語版 100 部

- H-InvDB パンフレット: 日本語版 100 部
- エコバッグ: 100 枚

○2009 年 10 月 7-9 日 Bio Japan 2009(横浜)

出展ブースでは MEDALS と H-InvDB のデモをそれぞれ二回ずつ行う機会を得て、多くの方にアピールすることができた。また、サイトの更新内容に合わせて刷新した MEDALS と H-InvDB のパンフレットをエコバッグとともに配布し、多くの質問や意見をもらうことができた。

【配布物】

- MEDALS パンフレット: 471 部(日本語版 453 部、英語版 18 部)
- H-InvDB パンフレット: 425 部(日本語版 384 部、英語版 41 部)
- エコバッグ: 370 枚

【意見・質問など】

- MEDALS に入ったことがあるが、全体に硬い感じである。
- 理想を言えば、ユーザー参加型でプログラム等を作らせるようなものが望ましい。

○2009 年 10 月 30 日 JBiC2009 プロジェクト研究成果報告会(品川)

MEDALS と H-InvDB について 1 件の口頭発表と 4 件のポスター発表を行った。口頭発表会場はほぼ満席で、MEDALS の紹介からこれまでの進捗成果について発表し、活発な質疑応答を行うことができた。また、ポスター発表では有意義な意見交換を行うことができた。

○2009 年 12 月 9-12 日 第 32 回日本分子生物学会年会(横浜)

MEDALS と H-InvDB についてブース展示、1 件のワークショップ発表、8 件のポスター発表を行った。展示ブースではエコバッグとパンフレットを配布し、MEDALS・H-InvDB の操作のデモンストレーションを行った。展示場所が会場の出入口に近かったことと、ブース位置をライフサイエンス統合データベースセンター(DBCLS)の隣にしたことで PR を効率的に進めることができた。ポスター発表では、事前に行ったワークショップ発表の影響もあり、MEDALS ツールへの関心が高く、多くの方からの質問・意見を頂くことができた。

【配布数】

- MEDALS パンフレット: 515 部(日本語版 508 部、英語版 7 部)
- H-InvDB パンフレット: 526 部(日本語版 516 部、英語版 10 部)
- エコバッグ: 500 枚

○2010 年 2 月 4-5 日 第 9 回 産総研・産技連 LS-BT 合同研究発表会

MEDALS と H-InvDB についてそれぞれ 1 件ずつ、ゲノムとバイオインフォマティクスの区分でポスター発表を行った。また、広報活動として、MEDALS と H-InvDB のパンフレット、および第 3 回講習会(つくば)と第 4 回講習会(大阪)の案内を配布した。

(3) リンクの設置

MEDALS へのアクセスルートの拡大を目的として、関連するサイト(学術的ポータルサイト、大学や研究所のページ、解析サービスへのサイト、または医学生物学関連の商業ウェブサイト等)から MEDALS へのリンクの設置を実現した(表 2.3.1-2)。また、今年度は農林水産生物ゲノム情報統合データベース「AgriTOGO」のページが公開され、相互リンクを実現することができた。

(4) 講習会開催

MEDALS および H-InvDB に関する講習会・セミナーを全国各地で計 5 回開催し、内容の解説とデモンストレーションを実施した(表 2.3.1-3)。文部科学省ライフサイエンス統合データベースセンター(DBCLS)や産総研生命情報工学研究センター(CBRC)、産総研糖鎖医工学研究センターなどの複数の機関と合同で講習会を開催することで、多様かつ密度のある講習会を開催することができた(表 2.3.1-4)。

表 2.3.1-2 MEDALS へのリンク設置一覧

サイト名/リンク元 URL
産総研公式サイト「主な研究成果」 http://www.aist.go.jp/aist_j/new_research/nr20090206/nr20090206.html
JBIC 公式サイト「研究開発成果」 http://www.JBiC.or.jp/activity/i_db_pj/medals.html
H-InvDB オフィシャルサイト「トップページ」 http://www.h-invitational.jp/index_jp.html
H-InvDB オフィシャルサイト「リンク集/H-InvDB 関連データベース&ツール」 http://www.h-invitational.jp/links_jp.html
文部科学省統合データベースプロジェクト http://lifesciencedb.jp/
文部科学省ライフサイエンス統合データベースセンター「リンク」 http://dbcls.rois.ac.jp/link/index.html
財団法人バイオインダストリー協会(JBA)「リンク集」 http://www.jba.or.jp/link/index.html
BIOWEB「バイオリンク: 情報収集サイト」 http://www.bioweb.ne.jp/mid_biolink/index.html
農林水産生物ゲノム情報統合データベース AgriTOGO「トップページ」 http://togo.dna.affrc.go.jp/

表 2.3.1-3 今年度開催した講習会

日付	講習会名	開催場所	参加人数
2009 年 8 月 25 日	第 1 回講習会	お台場(産総研・臨海副都心センター)	35 名
2009 年 11 月 11 日	データベースセミナー	品川(東京大学医科学研究所)	18 名
2009 年 11 月 20 日	第 2 回講習会	香川大学(医学部キャンパス)	15 名
2010 年 3 月 3 日	第 3 回講習会	つくば(産総研・つくば中央 共用講堂)	14 名
2010 年 3 月 5 日	第 4 回講習会	大阪(千里ライフサイエンスセンター)	21 名

表 2.3.1-4 今年度合同で講習会を開催した機関

講習会名	講習会合同開催機関
第 1 回講習会	産総研生命情報工学研究センター(CBRC) 文部科学省ライフサイエンス統合データベースセンター(DBCLS)
第 2 回講習会	香川大学研究推進機構 総合生命科学研究センター 徳島大学疾患ゲノム研究センター
第 3 回講習会	文部科学省ライフサイエンス統合データベースセンター(DBCLS) 糖鎖医工学研究センター

【参加者へのアンケート調査結果】

- 開催情報の入手は「メールマガジンから」が最も多く(46%)、次いで「関係者から」(31%)であった。
- 参加の目的は「情報収集」が 56%で、「業務に活用する」が 37%であった。
- 講習会の内容の満足度については 9 割以上の方が「期待以上・期待通り」と回答した。

【意見等】

- 総合的に説明を受けたのは初めてで、知らない機能があった。(H-InvDB)
- 見やすい、わかりやすい画面を作って欲しい。(H-InvDB)
- とても有用なサイトだと思いました。(MEDALS)
- PubMedScan は有用であろうと思う。(MEDALS)

(5) その他会合等

文部科学省統合データベースプロジェクトでは、データベースの利用許諾をクリエイティブ・コモンズ・ライセンス日本版「表示・継承」に準拠した標準利用許諾と、データベースの寄託者が独自に作成する追加利用許諾から構成している。経済産業省統合データベースプロジェクトではアーカイブの寄託におけるライセンス契約について文部科学省統合データベースプロジェクトの実績を参考にすべく、8 月 6 日 JBiC にて「REVIEW の会」を開催し、17 名が参加した。

「REVIEW の会」には経済産業省・文部科学省統合データベースプロジェクト・弁護士(オブザーバー)・経済産業省統合データベースプロジェクトのメンバーが参加し、データベース等の寄託におけるライセンスのあり方について有意義な意見交換を行うことができた。

2.3.2 ユーザーのニーズ調査

(1) アンケート調査

MEDALS・H-InvDB の潜在的ユーザーのニーズを調査することを主な目的として、専門会社に委託したウェブアンケート調査を 2009 年 8 月 4-18 日の期間で実施し、201 名分の回答を集計した。さらに H-InvDB の連携データベースである機能性 RNA データベース(fRNAdb)と糖鎖関連遺伝子データベース(GGDB)についても質問項目を作成した。質問項目は自由記述回答も含む全 51 問であった。

① 調査対象について

今年度のアンケート回答者の専門分野は、理学系が 28%、医学系が 26%、農学系が 17%、薬学系が 13%という順になった(図 2.3.2-1)。昨年度のアンケート結果では医学系が 41%、理学系が 20%であった。そのため、昨年度よりも特定の分野に大きく偏ることなく、幅広い分野からの回答を集めることができた。

② ライフサイエンス分野のデータベースについて

ライフサイエンス分野のデータベースで情報収集する際に利用している情報源をその利用頻度とともに調べた結果、昨年度と同様に、PubMedScan や DDBJ/EMBL/GenBank、Entrez Gene は月に 1 回以上利用する割合が高かった(図 2.3.2-2)。また、ライフサイエンス分野のデータベースで充実が望まれている情報としては、解析ツールの情報を選ぶ回答者が多く、各種アノテーション情報だけでなく、解析ツールの情報も必要とされていることがわかった(図 2.3.2-3)。ライフサイエンス分野のデータベースで不満に感じる点についての自由記述回答の一部を以下に紹介する。

【利用しているデータベースへの不満点等】

- 目的に合ったものがどこにあるか、どのように使えばよいかわかりづらい。
- 使われている遺伝子コードなどがデータベースによってバラバラでわかりづらい。
- 情報の更新や保守が不完全。
- Web ブラウザでの検索だけでなく、API が充実すると便利だと思う。
- 同じようなデータを含むデータベースが複数あり、違いを比較するのが難しい。

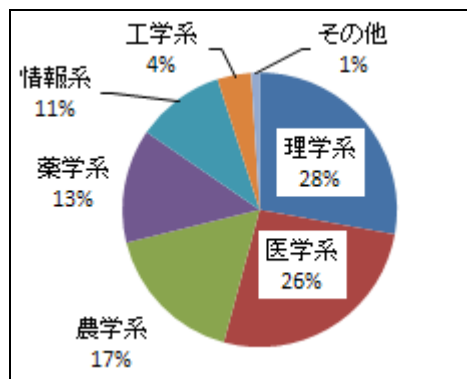


図 2.3.2-1 回答者の専門分野

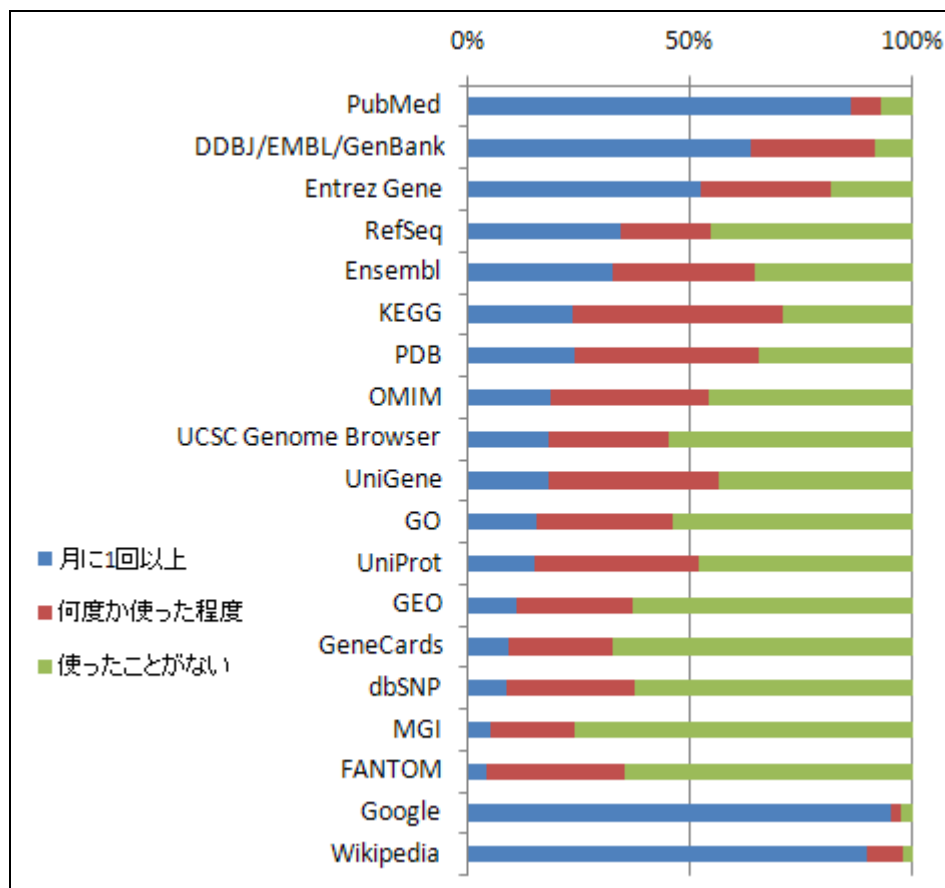


図 2.3.2-2 ライフサイエンス分野の情報収集で利用する情報源

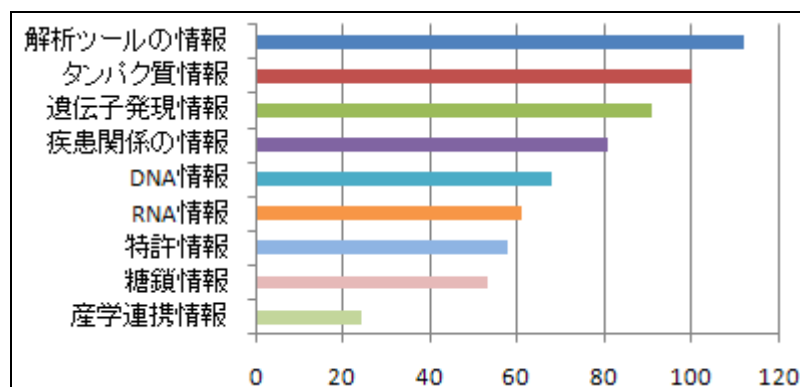


図 2.3.2-3 ライフサイエンス分野のデータベースで充実が望まれる情報

③ MEDALS について

MEDALS の利用状況・要望を把握するために、利用頻度や興味のあるコンテンツについて調査した。回答者の約 20%は MEDALS を使ったことがあり、その中でも薬学系・情報系では定期的に MEDALS を利用しているユーザーが他分野に比べて多いことがわかった(図 2.3.2-4)。MEDALS コンテンツの中で利用したいもの・興味のあるものについては「データベース相関図」、次いで「統合データベース(H-InvDB)」の順に選ばれた(図 2.3.2-5)。「H-InvDB」や「データベース便覧」、「ツール便覧」等、現在提供している多くのコンテンツに興味を持っていることがわかった。また、「データベースの相関図」や「プロジェクト相関図」等のユーザーの新たなニーズを見つけることができた(図 2.3.2-5)。

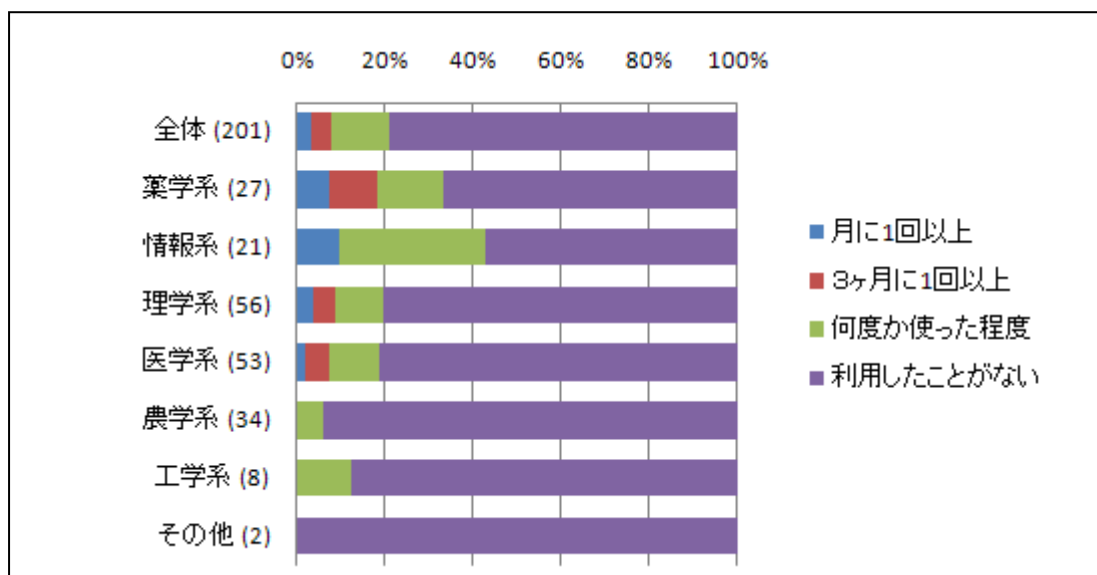


図 2.3.2-4 MEDALS の利用頻度

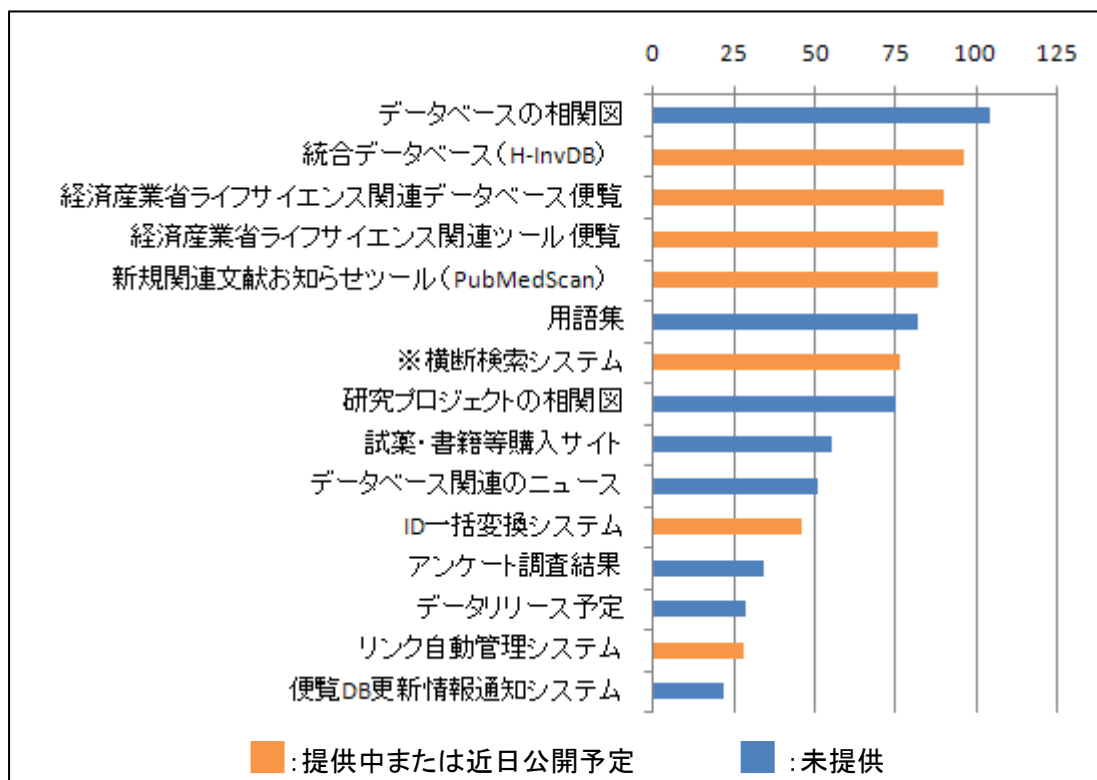


図 2.3.2-5 興味のあるまたは利用した MEDALS コンテンツ

※横断検索システム…文部科学省統合データベースプロジェクト・MEDALS データベース便覧との連携

以下に MEDALS への希望・要望についての自由記述回答の一部を紹介する。

【MEDALS への希望・要望】

- データベース、ツールなどの説明が日本語で提供されているので、情報探索の時間が短縮できて非常にありがたい。
- 予算の打ち切りとともに、データベース存続も打ち切られるような形式でないようにしていただけると嬉しい。
- 詳細情報のメタデータ構造はわりとしっかり考えてあると思いました。トップページでそれぞれの件数が表示されているのもわかりやすいです。
- それぞれの特色(強み、弱みなど)を教えてください。

(2) ヒアリング調査

より使い易いデータベースにするための指針を得ることを目的として、ライフサイエンス分野の企業の現場で働くバイオフィーマティシャン等にヒアリング調査を行った(表 2.3.2-1)。直接対話による意見交換・情報収集は業界の最新動向についても知ることができ、大変有意義であった。以下に頂いたコメントの一部を紹介する。

表 2.3.2-1 ヒアリング実施日と訪問先

日付	訪問先
2009 年 9 月 1 日	大手製薬企業研究所
2009 年 10 月 6 日	大学
2010 年 3 月 11 日	大手製薬企業研究所

【データベース利用の実態】

- プロジェクトの内容により使うデータベースは変わってくるが、データベースを選定する際には、そのデータベースが更新されているかどうかについても重要視する。
- 情報の網羅性という観点から、より多くの情報を持つデータベースを選ぶことが多い。
- コピー＆ペーストですぐに資料が作成できるような見せ方をしているデータベースが使い易い。

【情報の入手】

- 有償の情報配信サービス(アラートサービス)を利用していたため、自分自身でキーボードを打って検索する機会は減ってきている。
- 情報配信サービスでは、ただ単に情報を抽出するだけでなく、一度整理された情報を提供するサービスを望む。
- 企業の技術者は自身で集めた情報を外部のサーバーに預けるサービスは使わない(セキュリティを重視)。
- 新しいツールやデータベースについての情報は仲間内での情報交換で得られることが多い。

【プロジェクト全般について】

- 公的資金によるプロジェクトについてはベンチャー企業にいろいろな案やシステムを競わせる形で行わせるのが良いと考えている。
- 1～3 年の期間で企業を育成、または競争で絞り込んでいく形式も良いのではないかと。

2.4 他省庁との連携

本プロジェクトは、2005 年に総合科学技術会議の唱えた日本のライフサイエンス統合データベースを目指す取り組みの一つである。文部科学省でもこれを受けて、2006 年秋に「統合データベースプロジェクト」を開始し、4 省庁（経済産業省、文部科学省、厚生労働省、農林水産省）のデータベース統合のための中核団体として、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）を設立した。そこで本プロジェクトでは、DBCLS と様々な点で連携をしてプロジェクトを実行した。まず、DBCLS との連携内容について述べる。

2.4.1 文部科学省ライフサイエンス統合データベースセンターDBCLSとの連携

DBCLS にて連携のための会議を開催し、連携内容の決定や進め方の議論を行った。原則的に月一回のペースで開催することとし、4 月、5 月、6 月、7 月、10 月、12 月、1 月の計 7 回開催した。9 月、2 月は運営委員会、11 月は展示会等のため開催を見合わせた。

4 月の第一回会議では、「全体の連携方針」について議論した。両者の便覧の項目を比較し、共通の内容のものについては共有できるよう方策を検討した。DBCLS より、ダウンロード用データの寄託を受けるときに交わす手続きについて説明いただいた。横断検索について連携して進めていくことを確認し、検索エンジン等具体的な内容についても議論した。

5 月の第二回会議では、便覧の項目の共通化について具体的に議論した。また、便覧用のデータベースの調査状況と今後の調査方針について報告した。

6 月の第三回会議では、かずさ DNA 研究所から、モデル植物、実用化植物に関するデータ、質量分析等に関する解析ツールなど、10 個の公開データベース、ツールがすでに公開されていること、さらに多くのデータベースが本年度中に公開されることを報告した。また、ダウンロード用データの寄託を受けるときなどに必要となる著作権等の権利の表示方法等について、DBCLS 側から説明していただくことを提案し了承された。この会（REVIEW の会）は 8 月 6 日、お台場にて開催された。

7 月の第四回会議では、便覧の項目リストに基づき、各々のデータベースが公開可能か、管理者の連絡先が判明したか等調査経過を報告した。また便覧において、成果物だけでなく、プロジェクトについても共通化を提案した。

10 月の第五回会議では、便覧の掲載の対象に選定したものは 193 件であり、そのうち 81 件については既に掲載しており、49 件については公開可能であることを報告した。また横断検索の進捗状況、プロジェクト便覧の例等についても議論した。

12 月の第六回会議では、文献情報取得モジュール、PDF 自動取得機能、更新検知ツールなど開発したツールを紹介した。横断検索システムの特徴を報告するとともに、連携における懸案事項についても議論した。また、成果物便覧調査の進捗状況についても報告した。

1 月の第七回会議では、微生物産業利用支援データベース（DIAM）、大腸菌関連データベース（GenoBase）についてデータを寄託したいとの依頼がありこれに進めていること、DBCLS では既にデータの寄託の実績があり、両者を統一的に進めることについて議論した。また、ポータルサイト MEDALS の更新状況、成果物調査の進捗状況についても報告した。さらに、来年度の連携の進め方についても議論した。

以下に、データベース便覧、ポータルサイトでの検索機能、ウェブサービスの 3 つにおける詳細を述べる。

（a）データベース便覧

● 両サイトの便覧（項目と文言）の統一

文部科学省の統合データベースプロジェクトから公開されている便覧と整合性を図るため、連携を図り、説明項目の統一や同一の説明項目内の文言一致などを協議検討した。その結果、MEDALS から公開している 30 項目のうち、14 項目を一致させることにした。本報告書の 2. 1. 1 項に詳細を記述している。

- 著作権表示の方法

クリエイティブコモンズ(CC)を利用してわかりやすい表示を行うことにした。

(b) データの寄託

- ダウンロード用データ寄託の作業

データベース開発者から、データをダウンロードできる形で MEDALS ポータルサイトから配布できる旨が了承された場合は、寄託の手続きを行う。まず、ダウンロード用データ寄託の作業について、連携して進めることが決定した。

- データ等を寄託する際の書面

文部科学省側では、寄託に際して書面の整備が既に進められており、これをもとに覚え書きレベルのものを準備して説明を行って寄託をとりつけることにした。

(c) データの横断検索機能

- 横断検索用データ交換

横断検索に関しては、両者がそれぞれの担当データベースについてその INDEX を作成して、その INDEX をオンラインで相互利用することを前提として、その仕組みを MEDALS 側で開発するための情報交換を行うことになった。

2. 4. 2 ウェブサービスの統合化

Web サービス(ウェブサービス)統合化については、下記のように実施した。

(1) バイオ Web サービス開発ガイドライン(英語版マニフェスト)の作成

Web サービス共通ガイドラインは、標準的なメソッド、設計、ドキュメント整備等を提案するもので、文部科学省統合 DB Web サービス標準化コミュニティへ参加し平成 20 年度に平成 20 年度統合 Web サービス マニフェスト(20081022 版)を作成した。2009 年 3 月に BioHackathon2009 国際会議に参加し、Web サービス共通ガイドライン作成に協力しバイオ Web サービス開発ガイドライン(英語版マニフェスト(20090320 版))の作成を行った。

以下にバイオ Web サービス開発ガイドライン英語版マニフェスト(20090320 版)を示す。

Guideline for Web Services (Manifesto ver.2.0)

This document is a guideline to design web services toward integration.

- REST

- For a database service, REST services to call entries and search by a keyword should be provided.
- To unify query formats, the query languages should conform to Common Query Language.
- A method to get the number of hits for a keyword should be provided.
- A method to get the list of hit entries for a keyword should be provided.
- A method to get the entry from an ID should be provided.
 - If the entry does not exist, the method should return code 404 and an empty entry.
- Path parameters (/foo/bar/baz) should be used for a REST URL and query parameters (param1=value1¶m2=value2) should not be used.
 - Lower-case letters should be used for path parameters in principle.

- If a URL ends “/”, a list of possible strings after the URL should be returned in principle.
 - When a method returns a list, each line should correspond to an entry and each line should be divided by tab into fields.
 - As REST web services can be called from AJAX, an alternative could be [JSON](#).
 - [HTTP codes](#) should be used for error signaling. It is allowed to also send in the answer some text.
- SOAP
 - A manual for each method should be provided.
 - The manual should include examples of parameters such that the method works well by them.
 - Each method should be verified to use Java (for instance, [JAX-WS](#) or [JAXB](#)), Perl ([SOAP::Lite](#)), Ruby and Python ([SOAP.py](#), [ZSI](#), [TGWebServices](#) or [soaplib](#)).
 - Providing sample codes for the common SOAP toolkits in such languages is recommended.
 - Providing test case projects for free testing packages like [soapUI](#)
 - For asynchronous services, a job ID should be provided.
 - When any unrecoverable error occurs during processing at a server, a method should return “SOAP Fault”.
- The following data types should be used:
 - [INSDC INSDSeq](#) for sequences,
 - [MIAME](#) for expression data,
 - [GLYDE-II](#) for glycans,
 - Tab format for key-value.
- REST or SOAP
 - For retrieval methods (to get an entry or search entries by keywords), REST is recommended. For methods of analysis service, SOAP is recommended.

TogoWS Guidelines

REST

TogoWS stated the rulers to design the services to call and search entries as below:

- <http://togows.dbcls.jp/site/services.html>
- <http://togows.dbcls.jp/site/rest.html>

How to call the entries

`http://togows.dbcls.jp/entry/database/entry_id[,entry_id2,...]/field[.format]`

If the name space of the databases are in URL :

- <http://frnadb.example.org/entry/id> instead of <http://frnadb.example.org/entry/frnadb/id>

How to add the name spaces for the institution, release_number of the database:

- <http://togows.dbcls.jp/entry/birc-hinv-5.0/HIT0012345>

NOTE: The default format for the entries is “.txt”, and can be in the specified format by adding the suffix, such as “.xml”.

Calling a part of the entry: The method to call the part of the entries by field names is the extension of TogoWS using BioRuby/BioPerl?

How to search the entries

```
http://togows.dbcls.jp/search/database/query+string[.format][offset,limit[.format]]
```

```
http://togows.dbcls.jp/search/database/query+string/count
```

The format of “query string” to be based on “Common Query Language” [Common Query Language](#)

The conversion of data format and ID

```
http://togows.dbcls.jp/convert/data_source.format
```

Metadata

- The list of the databases

```
http://togows.dbcls.jp/entry/
```

```
http://togows.dbcls.jp/search/
```

- The list of the fields

```
http://togows.dbcls.jp/entry/DBname/
```

- The list of the formats

```
http://togows.dbcls.jp/entry/DBname/formats
```

SOAP

- Required test environment
 - Perl: Soap::Lite 0.69
 - Ruby: SOAP4R 1.5.5 (Ruby 1.8.6 bundle)

- Python: SOAPpy 0.12.0
- Java: Axis 1.4
- SOAP version 1.2, at least
- Calling a method with asynchronous
 - The name for the method with asynchronous end with “Async” and return job ID
 - Return job ID to “CheckAsyncStatus??” method and give either RUNNING, COMPLETED, or ERROR
 - Method name for the method with asynchronous must be “BeginInvoke? methods + Result” and state job ID as a parameter

以上

(2) 平成 20 年度統合 Web サービス マニフェスト(20081022 版)に準拠した Web サービスの開発

2. 2. 1 及び 2. 2. 2 で詳細を報告したとおり、多型データベース(VarySysDB)、糖タンパク質データベース(GlycoProtDB)およびヒト遺伝子統合データベース(H-InvDB)の Web サービスの開発を平成 20 年度統合 Web サービス マニフェスト(20081022 版)に準拠して行った。具体的には、データ検索およびデータ取得 Web サービスプログラム設計、ドキュメントやサンプルコード提供の点についてマニフェストに準拠した。

開発した Web サービスの URL:

- VarySysDB Web サービス
<http://h-invitational.jp/hinv/hws/varysysdb/>
- GPDB Web サービス
<http://riodb.ibase.aist.go.jp/rcmg/ws/gpdb/>
- H-InvDB Web サービス
<http://h-invitational.jp/hinv/hws/>

2.5 運営委員会

経済産業省統合データベースプロジェクトの進捗状況を明らかにし、議論を通じて外部有識者（アドバイザー）らの意見を成果物に反映させることを目的として、年度内に 2 回、運営委員会を開催した。運営委員会では、まず実施テーマごとにプロジェクトの進捗報告がなされ、それを受けてアドバイザー等からさまざまな意見が述べられ、最後に総合討論が行われた。ポータルサイト MEDALS の便覧に公開する形で成果物の調査が順調になされていること、自動更新などツール開発に工夫がなされていることが確認された。また、国家プロジェクトにおけるデータベースの公開のあり方については、活発な議論がなされた。文部科学省とは既に緊密な連携をとりながら進んでいるが、農林水産省、厚生労働省等他の省庁とも連携を図っていく旨確認された。

開催日時

第一回 2009 年 9 月 10 日(木)9 時 30 分－12 時 10 分

第二回 2010 年 2 月 18 日(木)9 時 30 分－12 時 5 分

アドバイザー名

伊藤隆司(東京大学大学院理学系研究科)

長洲毅志(エーザイ株式会社 CEO オフィス)

3. 成果発表

(1) 論文

1. Imanishi T, Nakaoka H (2009) Hyperlink Management System and ID Converter System: enabling maintenance-free hyperlinks among major biological databases. Nucleic Acids Research37 (Web Server issue): W17-22
2. Takeda J, Suzuki Y, Sakate R, Sato Y, Gojobori T, Imanishi T, and Sugano S (2009) H-DBAS: Human-transcriptome DataBase for Alternative Splicing, update 2010. Nucleic Acids Research38 (Database issue):D86-90
3. Yamasaki C, Murakami K, Takeda J, Sato Y, Noda A, Sakate R, Habara T, Nakaoka H, Todokoro F, Matsuya A, Imanishi T, Gojobori T (2009) H-InvDB in 2009, extended database and data mining resources for human genes and transcripts. Nucleic Acids Research38 (Database issue):D626-32

(2) 発表

1. 村上 勝彦、松矢 明宏、富所 布紗乃、小川 誠、中岡 源、五十嵐 由美子、岸本 晃彦、今西 規、五條堀 孝(2009)「MEDALS:経済産業省ライフサイエンスデータベースのポータルサイト」第3回GLITフォーラム「再生医療における糖鎖の利用と重要性ー評価技術と政策ー」
2. 村上 勝彦(2009)「経済産業省ライフサイエンス統合 DB プロジェクトのポータルサイト MEDALS」文科省統合データベースプロジェクトシンポジウム「データベースが拓くこれからのライフサイエンス」
3. 松矢 明宏、村上 勝彦、富所 布紗乃、小川 誠、中岡 源、五十嵐 由美子、岸本 晃彦、今西 規、五條堀 孝(2009)「経済産業省ライフサイエンス統合 DB プロジェクトのポータルサイト MEDALS と MEDALS ツール」文科省統合データベースプロジェクトシンポジウム「データベースが拓くこれからのライフサイエンス」
4. Murakami K, Todokoro F, Sato Y, Gojobori T, and Imanishi T (2009) PubMedScan: an automatic paper recommendation system for newly published PubMed articles. ISMB/ECCB 2009
5. 村上 勝彦、松矢 明宏(2009)「経済産業省ライフサイエンス統合データベースポータルサイト MEDALS」第8回国際バイオフィオーラム 併催国際バイオ EXPO
6. 村上 勝彦(2009)「経済産業省統合データベースポータルサイト MEDALS」平成21年度 第1回データベース講習会「ヒト遺伝子研究と統合データベース」
7. 村上 勝彦(2009)「統合データベースの発展」JBIC2009 プロジェクト研究成果報告会
8. 松矢 明宏、村上 勝彦、小川 誠、間宮 健太郎、岸本 晃彦、今西 規、五條堀 孝(2009)「経済産業省ライフサイエンス統合データベースポータルサイト MEDALS の拡張」JBIC2009 プロジェクト研究成果報告会
9. 村上 勝彦、小川 誠、今西 規、五條堀 孝(2009)「PubMedScan: Web で使える新規 PubMed 文献をお知らせするツール」JBIC2009 プロジェクト研究成果報告会
10. 村上 勝彦(2009)「経産省統合データベースポータルサイト MEDALS の紹介」平成21年度 第2回データベース講習会「疾患研究のためのゲノム情報の活用」
11. Murakami K, Yamasaki C, Matsuya A, Mamiya K, Ogawa M, Kishimoto A, Gojobori T, and Imanishi T (2009) METI Life science integrated database project. MBSJ2009 (第32回日本分子生物学会年会)
12. Murakami K, Imanishi T, Matsuya A, Mamiya K, Ogawa M, and Gojobori T, (2009) MEDALS: METI Database portal for Life Science. MBSJ2009 (第32回日本分子生物学会年会)
13. Murakami K, Ogawa M, Gojobori T, and Imanishi T (2009) PubMedScan: a web-based recommender of PubMed articles. MBSJ2009 (第32回日本分子生物学会年会)
14. 村上 勝彦、松矢 明宏、小川 誠、間宮 健太郎、岸本 晃彦、今西 規、五條堀 孝(2010)「経済産業省

ライフサイエンス統合データベースプロジェクト:ポータルサイト MEDALS」平成 21 年度 産総研・産技連 LS-BT 合同研究発表会

15. 松矢 明宏(2010)「経産省統合データベースポータルサイト MEDALS の紹介」平成 21 年度第 3 回データベース講習会「ヒト遺伝子研究と統合データベース」
16. 村上勝彦(2010)「経済産業省統合データベースプロジェクトのポータルサイト MEDALS」平成 21 年度第 4 回データベース講習会「ヒト遺伝子研究と統合データベース」第二部 講習会 ヒト遺伝子研究と統合データベース
17. 今西 規(2010)「トランスクリプトームからみたヒト遺伝子の全体像と多様性」平成 21 年度第 4 回データベース講習会「ヒト遺伝子研究と統合データベース」第一部 講演会 ヒト遺伝子研究と統合データベース
18. 武田 淳一(2010)「完全長 cDNA によるヒトトランスクリプトーム解析と次世代シーケンサーの可能性」平成 21 年度第 4 回データベース講習会「ヒト遺伝子研究と統合データベース」第一部 講演会 ヒト遺伝子研究と統合データベース
19. 村上 勝彦(2010)「プロモーターモチーフによるヒト遺伝子発現解析」平成 21 年度第 4 回データベース講習会「ヒト遺伝子研究と統合データベース」第一部 講演会 ヒト遺伝子研究と統合データベース

(3) データベース公開・更新

2009 年

- | | |
|-----------|---|
| 5 月 20 日 | 経済産業省ライフサイエンス関連プロジェクトのポータルサイト MEDALS
データ更新 成果物追加 2 件(MDV, HEAT)
プロジェクト成果報告書ダウンロードページ公開 |
| 6 月 11 日 | MEDALS データ更新 成果物追加(Kappa-view) |
| 8 月 24 日 | MEDALS データ更新 成果物追加 8 件(CoP, KAGIANA, DAGViz, RnR, MassBase, DOAN,
NBRC オンラインカタログ検索, CADLIVE) |
| 9 月 8 日 | MEDALS データ更新 プロジェクト便覧公開、成果報告書 3 件追加 |
| 10 月 29 日 | MEDALS データ更新 成果物追加 5 件(KATANA, DIAM, 予測プログラム, KNApSAcK, Dr
DMASS) |
| 11 月 13 日 | ヒト多型データベース VarySysDB ウェブサービス公開 |
| 11 月 19 日 | MEDALS データ更新 成果物追加(PMID-Extractor) |
| 11 月 30 日 | MEDALS データ更新 成果物 6 件追加(JSNP DATABASE, KOMICS, MassBank,大腸菌マル
チオミクス データベース, OpenPML,FLJ Human cDNA Database)
プロジェクト 13 件追加 |
| 12 月 24 日 | MEDALS データ更新 成果物 3 件追加(ATTED II, ADMER2, CGH Data Base)
プロジェクト 12 件追加
成果報告書(機能性 RNA プロジェクト)追加 |

2010 年

- | | |
|----------|---|
| 1 月 28 日 | MEDALS データ更新 成果物 4 件追加
(GenoBase, 正常臓器 遺伝子発現データベース,RNAi 活性予測アルゴリズムによる siRNA
設計ツール"siExplorer",化学物質統合検索システム)
プロジェクト 8 件追加 |
| 1 月 29 日 | 糖タンパク質データベース GlycoProtDB ウェブサービス公開 |
| 2 月 25 日 | MEDALS データ更新 成果物 4 件追加(KNApSAcK Family, DPCLus, THE MICROARRAY
ANALYSIS SYSTEM - TREBAX -, Dr DMAS+) |

3 月 ヒト遺伝子統合データベース H-InvDB ウェブサービス拡張版公開予定
 ヒト多型データベース VarySysDB 更新公開予定
 選択的スプライシング・バリエーションデータベース H-DBAS 更新公開予定

4. 謝辞

本プロジェクトの実施にあたり、多数の方のご協力、ご貢献を頂いた。ここに深く感謝申し上げます。(敬称略)

- 文部科学省統合データベースとの連携
 高木 利久 文部科学省 ライフサイエンス統合データベースセンター
 西川 哲夫 文部科学省 ライフサイエンス統合データベースセンター
- 便覧作成協力
 各データベース・解析ツール開発者の方々
- 民間企業でのデータベースニーズ調査
 社団法人 バイオ産業情報化コンソーシアム会員企業(3 社)

5. 参加者名簿

役職	氏名	所属機関
プロジェクトリーダー	五條堀 孝	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(併任)
サブリーダー	今西 規	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
	村上 勝彦	社団法人 バイオ産業情報化コンソーシアム
研究員	松矢 明宏	社団法人 バイオ産業情報化コンソーシアム(出向)
	小川 誠	社団法人 バイオ産業情報化コンソーシアム(出向)
	間宮 健太郎	社団法人 バイオ産業情報化コンソーシアム(出向)
	山崎 千里	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
	武田 淳一	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
	坂手 龍一	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
	野田 彰子	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
技術員	羽原 拓哉	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
	遠藤 智宏	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)
	岸本 晃彦	社団法人 バイオ産業情報化コンソーシアム
研究協力機関	独立行政法人 産業技術総合研究所 生命情報工学研究センター 特に経済産業省関連機関から産生されるデータベース等の統合において 浅井 潔 光山 統泰	
	独立行政法人 産業技術総合研究所 糖鎖医工学研究センター 特に経済産業省関連機関から産生されるデータベース等の統合において 成松 久 鹿内 俊秀	
	国立大学法人北海道大学大学院 情報科学研究科	
	学校法人 慶應義塾大学 医学部	
	国立大学法人東京大学大学院情報理工学研究科	
	大学共同利用機関法人 国立遺伝学研究所生命情報・DDBJ 研究センター	

運営委員(アドバイザー)

伊藤 隆司

東京大学 大学院理学系研究科

長洲 毅志

エーザイ株式会社

＜参加企業＞

日立公共システムエンジニアリング株式会社

株式会社 ダイナコム

日立ソフトウェアエンジニアリング株式会社