

ライフサイエンス分野の統合データベース整備事業

ライフサイエンス統合データベース開発運用

(戦略立案・実行評価／統合データベース開発／統合データベース支援)

20年度 研究成果報告書

平成21年3月

大学共同利用機関法人 情報・システム研究機構	高木利久
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	浅井 潔
財団法人 かずさディー・エヌ・エー研究所	中村保一
国立大学法人 九州大学	林 健志
国立大学法人 奈良先端科学技術大学院大学	松本裕治
国立大学法人 東京大学	森下真一
国立大学法人 お茶の水女子大学	瀬々 潤
学校法人 関西文理総合学園長浜バイオ大学	池村淑道

本報告書は、文部科学省の科学技術試験研究委託事業による委託業務として、大学共同利用機関法人 情報・システム研究機構、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、財団法人 かずさ DNA 研究所、国立大学法人 九州大学、国立大学法人 奈良先端科学技術大学院大学、国立大学法人 東京大学、国立大学法人 お茶の水女子大学及び学校法人 関西文理総合学園長浜バイオ大学が共同で実施した、平成20年度の「ライフサイエンス統合データベース開発運用」の成果を取りまとめたものです。

目 次

1. プロジェクトの目的	4
2. 戦略立案・実行評価（情報・システム研究機構）	4
2. 1 戦略立案・実行評価の実施計画	4
2. 2 戦略立案・実行評価の実施内容	5
3. 統合データベース開発	15
3. 1 共通基盤技術開発（情報・システム研究機構、生命情報工学研究センター、 奈良先端科学技術大学院大学）	15
3. 1. 1 共通基盤技術開発の実施計画	15
3. 1. 2 共通基盤技術開発の実施内容	15
(1) 共通基盤技術開発（情報・システム研究機構）	15
(2) ワークフロー技術を用いた統合 DB 環境構築 （生命情報工学研究センター）	19
(3) 専門用語辞書管理システムと専門用語解析技術の開発 （奈良先端科学技術大学院大学）	24
3. 2 ヒト統合 DB の開発・運用（情報・システム研究機構、九州大学）	27
3. 2. 1 ヒト統合 DB の開発・運用の実施計画	27
3. 2. 2 ヒト統合 DB の開発・運用の実施内容	28
(1) ヒト統合 DB の開発・運用（情報・システム研究機構）	28
(2) 多型知識表現技術の開発（九州大学）	37
3. 3. モデル生物・産業応用生物統合 DB の開発・運用 （情報・システム研究機構、かずさ DNA 研究所）	39
3. 3. 1 モデル生物・産業応用生物統合 DB の開発・運用の実施計画	39
3. 3. 2 モデル生物・産業応用生物統合 DB の開発・運用の実施内容	40
(1) モデル生物・産業応用生物統合 DB の開発・運用 （情報・システム研究機構）	40
(2) 植物及び植物関連微生物のゲノム情報データベース統合と高度化 （かずさ DNA 研究所）	43
4. 統合データベース支援（情報・システム研究機構）	47
4. 1 ポータル整備・運用、広報、普及啓発	47
4. 1. 1 ポータル整備・運用、広報、普及啓発の実施計画	47
4. 1. 2 ポータル整備・運用、広報、普及啓発の実施内容	47
(1) 統合 DB プロジェクトポータルサイトの整備・運用	47
(2) 広報・普及啓発	51
4. 2 データベース受入	54

4. 2. 1	データベース受入の実施計画	54
4. 2. 2	データベース受入の実施内容	54
4. 3	人材育成(東京大学、お茶の水女子大学、長浜バイオ大学)	60
4. 3. 1	人材育成の実施計画	60
4. 3. 2	人材育成の実施内容	61
	(1) DB 構築者の養成(東京大学)	61
	(2) DB 高度利用者の養成(お茶の水女子大学)	63
	(3) アノテータ・キュレータの教育(長浜バイオ大学)	64
5.	プロジェクトの総合的推進(情報・システム研究機構)	66
5. 1	プロジェクトの総合的推進の実施計画	66
5. 2	プロジェクトの総合的推進の実施内容	66
	(1) 研究運営委員会及び研究運営委員会作業部会	66
	(2) プロジェクト参加機関および関連機関との連携	68
	(3) 外部有識者からの意見聴取	70
5. 3	その他シンポジウム、セミナー等	70
	(1) 統合データベース講習会(情報・システム研究機構、長浜バイオ大学、 かずさ DNA 研究所)	70
	(2) BioHackathon2009(情報・システム研究機構)	73
	(3) ゲノムテクノロジー第 164 委員会との共催による研究会 (情報・システム研究機構)	76
	(4) バイオインフォマティクスへの招待(お茶の水女子大学)	76
	(5) アノテータ、キュレータ養成のための運営委員会と公開イベント (長浜バイオ大学)	77
6.	成果の外部への発表	79
7.	実施体制	79
8.	整備実績一覧	79

1. プロジェクトの目的

ライフサイエンスやバイオ産業に従事する研究者や技術者がいわゆるゲノムプロジェクト・ポストゲノムプロジェクトの成果や多様な DB や解析ツールをストレスなく利用してより高度な研究開発が効率よくできる環境（統合 DB）を実現することを目的とする。このため、情報・システム研究機構、科学技術振興機構、産業技術総合研究所生命情報工学研究センター、かずさ DNA 研究所、九州大学、奈良先端科学技術大学、東京大学、お茶の水女子大学、長浜バイオ大学は共同して、以下の3つの業務を行う。

（1）戦略立案・実行評価

ライフサイエンス（基礎生物学および応用生物学（医・薬・農））、情報処理技術、ライフサイエンス DB の3分野の専門家による研究運営委員会を組織し、各々の分野の動向に即した我が国の DB 整備戦略を立案する。また、3分野の動向に関する情報の網羅的収集・分析を行い、戦略立案や実行評価を支援するチームを組織する。

（2）統合データベース開発

ライフサイエンス、バイオ産業にかかわる情報へのアクセスと利用に関する格段の利便性向上とそれによる研究開発の飛躍的な効率化と質的向上を目指して、ヒト統合化 DB およびモデル生物・産業応用生物統合 DB の開発・運用と、それに必要な統合 DB 構築技術、DB の標準化、文献等からの知識獲得技術等の共通基盤技術開発を実施する。

（3）統合データベース支援

個々の公開 DB や利用可能な DB 解析ツールなどを網羅的に調査し、提示するポータルサイトを構築する。運用にあたって専門家及び利用者の意見集約システムを開発し、意見をフィードバックして運用する。また日本語による研究情報の流通促進に注力する。プロジェクトデータベースを受入れ、相互運用可能な DB へ変換した上で公開し、運用する。また、広報・普及啓発の活動を行う。人材育成は実績のある機関と連携し階層別に、アノテータ・キュレータ教育、DB 構築者養成、及び DB 高度利用者養成を行う。

2. 戦略立案・実行評価（情報・システム研究機構）

2. 1 戦略立案・実行評価の実施計画

ライフサイエンス（基礎生物学および応用生物学（医・薬・農））、情報処理技術、ライフサイエンス DB の3分野の専門家による研究運営委員会を組織し、各々の分野の動向に即した我が国の DB 整備戦略を立案する。上記委員会に対し、3分野の動向に関する情報の網羅的収集・分析を行い、戦略立案や実行評価を支援するチームを組織する。このような体制のもとで、戦略立案、実行評価の支援業務とそのための情報収集・分析だけでなく、

関係府省、利用者、産業界、学会・出版社、国内外の研究機関等関係者との連絡調整等の活動もあわせて行う。

2. 2 戦略立案・実行評価の実施内容

1) 戦略立案

戦略立案については、平成20年度に行われた本プロジェクトの中間評価結果や文部科学省ライフサイエンス情報基盤整備作業部会の報告書等の内容に対応すべく、これまで進めてきた統合データベースプロジェクトの目標や年次計画の具体化、国内外の動向調査をベースに、さらには、研究運営委員会の委員の意見も踏まえ、データベース統合化のあるべき姿の検討を進めた。併せてデータ共有化や公開促進の観点から、著作権、個人情報保護に関する調査、検討、並びに今後のデータ産出の中心となることが想定される次世代シーケンサの生み出すデータに関する調査を行い、特にヒトゲノム多型データの共有に関わる指針策定を行った。日本語文献の公開については、コンテンツを充実すべく学会要旨を提供いただく学協会との交渉を行い、コンテンツを充実した。また、国内外の研究機関等関係者との連絡調整という観点から、平成19年度に引き続き、世界的に分散しているデータの統合利用に関する国際ワークショップであるBioHackathonを開催し（参加者は海外30名、国内28名）、国際的な標準化を実現する場、及び国内の関連研究者が海外の第一線の研究者、開発者と共同作業を通じて緊密な関係を築く場としての役割を果たした（3. 1. 2項、5. 3項参照）。その他、日本学術振興会ゲノムテクノロジー164委員会と共催したシンポジウムでのパネルディスカッション等を通じた産業界及び研究者コミュニティの意見収集や、ヨーロッパのEMBL-EBI主催の生物情報基盤構築プロジェクト「ELIXIR」の調査等によるニーズと動向把握を行った。また、法律、契約に関わる事項については、当該分野に詳しい弁護士のコンサルテーションを得ながら進めた。

まず、データベース統合化のあるべき姿の検討については、本プロジェクトの周辺においても、平成20年度には、平成20年7月付で報告書が出された本プロジェクトの中間評価、平成21年度概算要求時の総合科学技術会議のヒアリング、及び文部科学省ライフサイエンス委員会の情報基盤整備作業部会（平成20年10月～平成21年1月）、さらには総合科学技術会議ライフサイエンスPTの統合DBタスクフォース（平成20年12月～平成21年4月）の場において、我が国の関連分野の専門家による議論、検討が行われた。これらにおける共通点は、本プロジェクト終了後も、本プロジェクトの成果である統合データベースを恒久的に維持、発展させる必要があること、また、そのための具体策として、本プロジェクトと科学技術振興機構（JST）のBIRD事業を一本化するための組織をJST内に構築すべきである、との提言であった。これらの提言を受け、本プロジェクトでも松原謙一研究運営委員長のリダーシップの下、運営委員の先生方から意見を収集して、誰でも理解できるライフサイエンス統合データベースのアウトラインを提示することになっ

た。また、これと並行してプロジェクト側でも統合化のビジョンということで、改めて統合化の全体像とステップについて検討した。本検討結果は、まだ研究運営委員会場でオーソライズされているわけではないが、以下で現在までの検討結果の概要を述べる。

運営委員の方々の意見として出てきたのは、まずデータベース統合化の理念として、
1. 公的資金を用いて得られたデータ、データベースは広く共有して、その活用を図るべきであること。
2. データベース統合により、生命科学の新たな展開とバイオ産業の国際的競争力向上に貢献するべきであること。
3. 統合データベースは重要な基盤であり、省庁に囚われない存在として国が継続的支援すべきである、との3点が重要であり、また、これを実現するためには、公的資金で得られたデータをコミュニティで早期に共有するためのルールを作成すること、及び公的資金で得られたデータを保全、管理し育成するための中立的機関を設置する必要があるとの指摘である。次に統合化は、アカデミアの知識を集約し、日本で生産するデータの価値を最大化することにより、日本のユーザ、更には世界のユーザに貢献することを目標にするべきであり、具体的には、制約のないデータの共有、操作感の統一及び知識発見支援を提供することによる生命科学のブレークスルー、イノベーション、新たな研究スタイルの実現を目指すことが究極の目標になる。

統合データベースの持つべき機能として挙げられたなかの代表例は、サービスの側面として、

- ・ 個々のデータベース、解析ツールへの案内機能
- ・ 複数のデータベースの統合化による付加価値の向上
- ・ 標準インタフェース、フォーマット、オントロジーなどの共通基盤の提供
- ・ 維持困難なデータベースの保存によるサービスの継続
- ・ 一次データの登録、編集、標準化、品質管理、閲覧サービス
- ・ 企業アクセスも視野に入れたセキュアな環境の提供

基盤的側面として、

- ・ 使い勝手およびデータの付加価値向上のための研究開発機能
- ・ データベース統合利用促進のためのファンディング機能
- ・ 世界的情報共有化のための国際的な連携機能
- ・ データベースの構築や統合化のための人材育成機能と教材提供

などである。

また、今後の進め方としては、上記の提言の趣旨に従ってプロジェクト終了後の体制の構築を検討していくと同時に、関係4省の連携など全体としての統合を実現するための組織の重要性、及び統合データベースの整備と併せて基盤データベースの維持や個別データベース構築への支援にも留意する必要があるとの意見が寄せられた。その他、データの所有権、公開原則あるいはデータベースへの登録義務といった考え方を、省庁を越えて統一することや、個人情報保護、著作権、営利目的利用に関する考え方を整理して、

関連法令、規則等の整備を推進する必要があるとの指摘があった。

以上のような統合データベースのアウトラインに関して寄せられた意見を受けて、中核機関でも、改めてデータベース統合推進に関する基本的考え方、及びデータベース統合化の全体像とそれに至るステップを整理した。基本的考え方は、

- ・データの共有化、価値の最大化が目的であり、必ずしも一カ所集中型を目指すものではない
- ・データを統合化し、価値を高めるには、知識の集約、デジタル化が必要であり、これに積極的に取り組む
- ・技術的には可能でも、制度的その他の理由により、共有化できていないものは、その障害の解消に積極的に取り組む
- ・DB利用者だけでなく、DB構築提供者にも喜ばれる仕組みを作る
- ・サービスの安定的、永続的な維持更新

である。図2.1にデータベース統合化の全体像とそれに至るステップを示す。

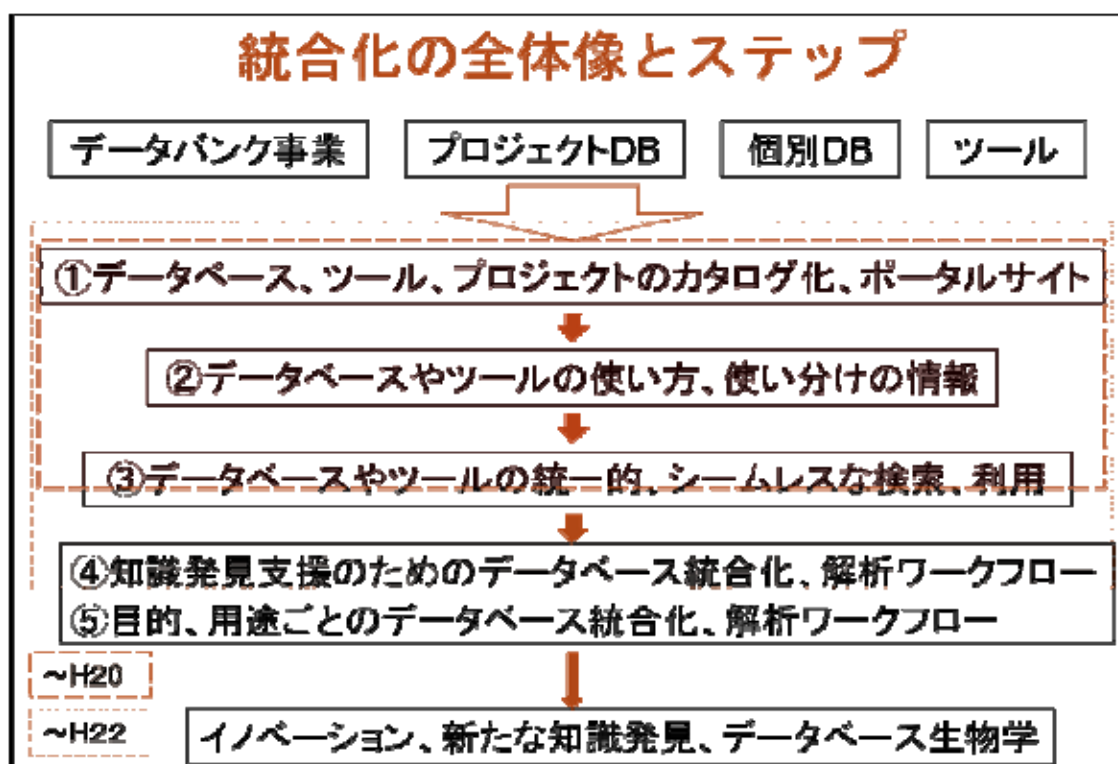


図2.1 統合化の全体像とステップ

図2.1に示した5段階のうち、4番目のステップまでを本プロジェクトで実現することを考えている。5番目のステップは、生命科学全体というよりはその中の特定の分野ごとに必要となるサービスであり、次のステップと考えられる。これ以外にも、中長期的には、個人ゲノム時代への対応、次世代DNAシーケンサへの対応、オープンアクセス化への

対応、医療、健康情報への対応、医薬品、化合物への対応、分子イメージング、画像、動画への対応、環境、生物多様性情報への対応、DBの視点からの国家プロジェクト管理、世界的な連携、役割分担への対応、上記のためのさまざまなガイドライン、法整備といった項目も、統合データベースとして検討、考慮する必要のある課題である。現在、3番目のステップの大部分に開発の手が及んでおり、その中の多くのものがサービスとして公開されている。4番目のステップまでの個別のサービスの進捗状況を図2.2に示す。図中の赤字の部分はすでに開発がほぼ完了しているものであり、青字の部分が進行中のものである。今後は、青字の部分の開発を完了させていくと同時に、統合検索など着手が遅れている部分を早期に立ち上げていく必要がある。

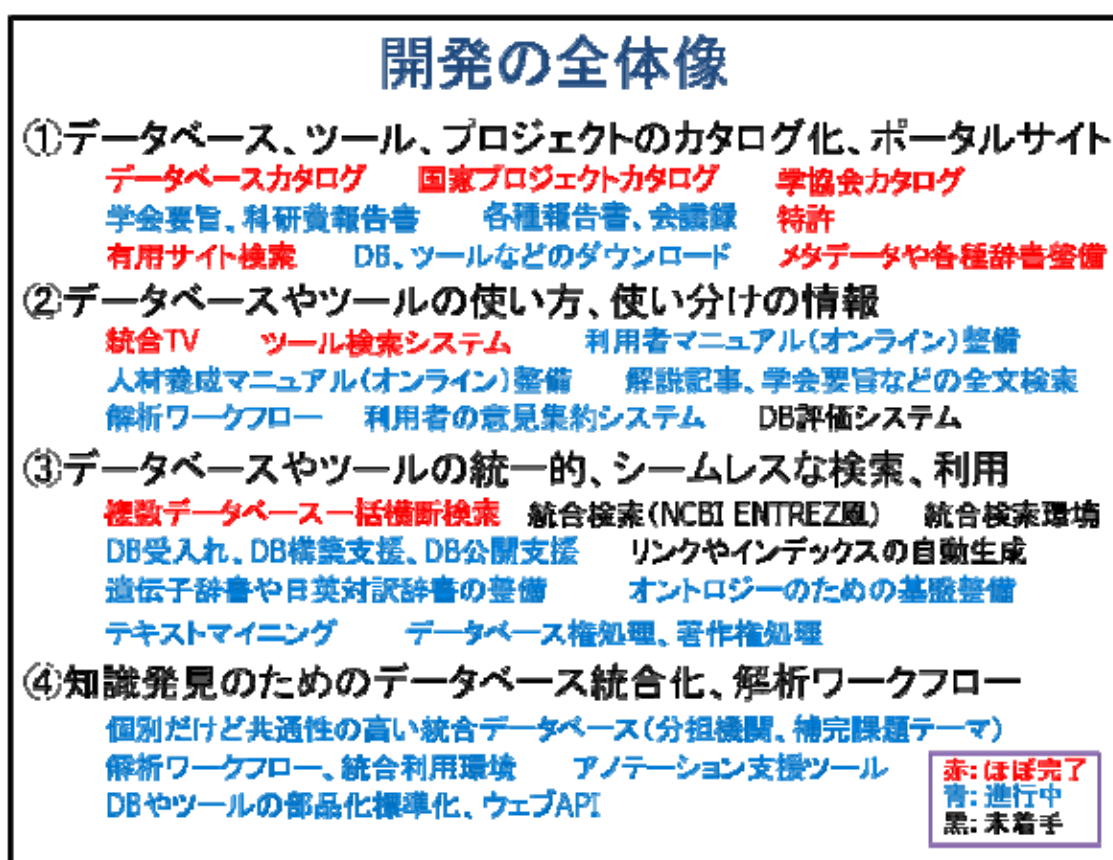


図2.2 開発の全体像と進捗状況

個人情報保護に関する調査、検討については、まず、次世代DNAシーケンサ技術の進展に伴い予想される個人ゲノム情報の蓄積とその医学への応用という局面で想定されるデータ量とデータのセキュリティの確保という観点での調査を行った。次に、こうした調査の結果をベースに、本プロジェクトの分担機関である東京大学の疾患解析データベースグループと連携して、同グループが提供するゲノムワイドな多型情報および関連解析情報および疾患関連遺伝子のリシーケンスデータのデータカテゴリー別の共有方針の

検討を行った。併せて、データ提供に伴うデータアクセス同意書の起案や提供先でのデータの漏えい等の危険に対する安全性を確保するためのセキュリティ手順書の策定などを行った。これにより、疾患解析データベースグループの提供するデータの共有が進むことが期待できる。

2) 実行評価

実行評価については、想定されるユーザドメインから、有用な回答が期待できる有識者を選定し、本プロジェクト成果のユーザ評価を実施した。また出展した学会等の展示会でのアンケート結果も踏まえ、逐次サービス内容の改善を行った。併せて、サービスサイトへのアクセスログを持続的に収集する体制を整え、ユーザからの利用状況の把握に努めた。

ユーザ評価については、プロジェクトに先立ち実行された科学技術連携施策群 生命科学の基礎・基盤 補完的課題「生命科学データベース統合に関する調査研究」において意見の聴取を行った関係者等、想定されるユーザドメインから、有用な回答が期待できる 92 名の有識者を選定し、平成 19 年度に開発した本プロジェクトの公開サービスを中心に、サービス内容の紹介も兼ねたウェブサイト記入方式のユーザ評価を実施した。サイトにおいては、代表的な 17 種類のサービスを 5 つに分類し、それらに対する 5 段階評価とフリーコメントの記入をお願いした（図 2.3）。回答数は項目ごとに異なるが、のべ 58 名（回答率 63%、評価者の背景を図 2.4、5 段階評価の結果を表 1.1 に示した）であり、項目ごとにもらったフリーコメントは 26～58 件（合計 188 件）、全体に対するフリーコメントとして今後提供すべきサービスについての意見は 28 件あった。また、これらのフリーコメントに対して回答可能なものについては、担当者からの回答を併記して、ウェブサイト“H19 年度「統合データベースプロジェクト」事業成果に対する外部評価の結果”

（<http://lifesciencedb.jp/feedback/hyouka19/?result>）から公開した(4. 2. 2 項参照)。これらのコメントのうち、すぐに実現可能なものは実施し、次年度の開発に盛り込めるものは盛り込んだ。例えば、インタフェースの改善点の意見を取り入れてデータベースカタログ等の結果表示の方式を変更したり、英語での紹介が有効と考えられるサービスを中心にウェブサイトの英語化の準備を進めた。また、出展した学会等の展示会においてアンケートを実施し、サービス改善のための情報を得るとともに、ユーザ評価者の募集を行い、合計 184 名の新たな評価者を得た。

ライフサイエンス統合データベース | FrontPage - LSDB 平成... | H18年度外部の評価

LSDB 文部科学省委託研究開発事業 統合データベースプロジェクト

[登録サイトトップ] [総合ホームページ]

ライフサイエンス統合データベースセンターでは、文部科学省委託研究開発事業「統合データベースプロジェクト」において、平成19年度に行った事業の成果を「統合ホームページ」(<http://lifesciencedb.jp/jp/>)にて公開しております。お忙しいところ大変恐縮ではありますが、19年度の事業成果について外部からのご評価をいただきたく協力をお願い申し上げます。昨年廣いたいただいたご意見(※)は[こちら](#)に匿名で転記し、まとめ、ご意見・ご提案への対応も記載しております。今年度も同様に、ご意見を匿名でまとめたサイトの公開をもちまして19年度の評価のご報告とさせていただきますと同時に、今後のサービスの改善に活かしてまいります。

※内閣府調査(代表研究者・大久保公策)の一環として実施

設問は5つの項目(代表的な17のサービスを分類)について選択肢とコメント欄が1つずつございます。それぞれの項目についてコメントをご記入の上、その進め方についてご意見をお聞かせ下さい。検索の行えるサービスについては、「TEST SEARCH」をクリックしますとテストクエリの入ったサービス画面に飛びますので検索(search)ボタンをクリックしてください。ツールなどのサービスについては「SHOW SAMPLE」をクリックしますとサンプル画面等をご覧いただけます。

1. 生命科学研究の情報へ案内する

TEST SEARCH

生命科学系データベースカタログ ([リンク](#))

生命科学のデータベースが検索できます。

TEST SEARCH

生命科学学会カタログ ([リンク](#))

国内の生命科学系学会を網羅したデータベースが検索できるほか、特許も検索できます。

TEST SEARCH

ゲノム・ポストゲノム主要プロジェクト一覧 ([リンク](#))

4省のゲノム・ポストゲノム関連主要プロジェクトの内容や成果を公開しています。

サービス全般について、今後の進め方に関するご意見をお聞かせください。

選択肢	数値	
かなり役に立ちそうなので、どんどん進めてほしい	21	投票
それなりに役に立ちそうなので、できる範囲で進めてほしい	4	投票
責任をもって評価できない	1	投票
あまり役に立たなさそうなので、進めなくてもよい	1	投票
まったく役に立たなさそうなので、やめたほうがよい	0	投票

サービス全般、あるいは個別のサービスについて以下のコメント欄にご意見をお聞かせ下さい。

コメント:

- Googleのような一般的な検索システムが充実する中で、ライフサイエンスの中ではそれに勝る、それより使いやすい便利なものにしていかねばなりません。ポータルとして常にHPIとしてもらえるような工夫(例えばGoogle検索や重要サイトへのリンクをはるとか、RSSを入れるとか、ライフサイエンス関連のBLOGのポータルを入れるとか)が必要ではないでしょうか。また日本語はGlobal Languageにはない特長なのでEnglishは絶対必要だと思います。
- 現在Web Siteの発展は早くてそれに追いつかないとおもわれます。例えばコンテンツメニューのパーソナライズ機能は必要でしょう。試しておませんがIFだけでなく他のブラウザに対応しつつAJAXなどの最新技術を入れていく事は大変だとは思いますが期待しておりますので宜しくお願い申し上げます。8月1日 -- 2008-08-01 (金) 10:13:48

http://lifesciencedb.jp/feedback/hyouka18/hyouka_H18.html

図 2.3 評価記入サイトの一部

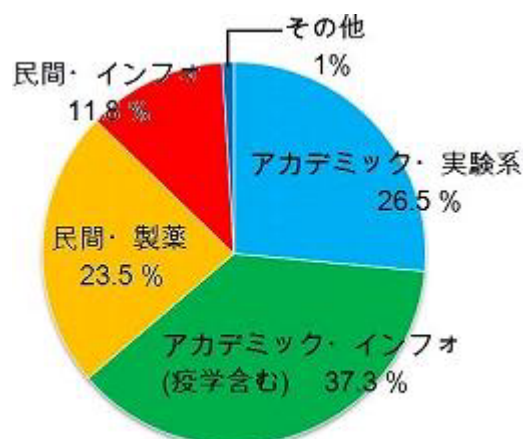


図 2.4 評価者の背景

表 1.1 項目ごとの 5 段階評価の結果

項目タイトル（フリーコメントのあった具体的サービス名）	評価				
	5	4	3	2	1
生命科学研究の情報へ案内する（生命科学系データベースカタログ、学協会カタログ、主要プロジェクト一覧）	79	15	4	2	0
国内外の様々な情報の検索サービスを提供する（横断検索、学会要旨統合検索、遺伝子発現バンク（GEO）目次、蛋白質核酸酵素全文検索、DNADB 総覧と検索）	56	31	4	9	0
情報整理や抽出の道具を提供する（アナトモグラフィー/BodyParts3D、Allie、OReFiL、Wired-Marker）	27	52	17	4	0
DB 統合に必要な新しい基盤技術を創る（TogoDB、TogoWS）	41	27	27	4	0
DB の未来を担う人材を育成する（統合 TV、MotDB）	17	32	26	15	11

(%)

評価 5：かなり役立ちそうなので、どんどん進めてほしい

評価 4：それなりに役立ちそうなので、できる範囲で進めてほしい

評価 3：責任を持って評価できない

評価 2：あまり役に立たなそうなので、進めなくてもよい

評価 1：まったく役に立たなそうなので、やめたほうがよい

また、アクセスログの収集については、公開サービスのアクセスログ取得のシステムを整え、集計処理を行った。集計ソフトは AWStats ver.6.6 (<http://www.bflets.dyndns.org/Tools/AWStatsJpn.html>)を利用した。ウェブサーバ Apache ver. 2.2 が出力する NCSA combined フォーマット形式のアクセスログを AWStats が解析し、日別、月別に以下の集計値等を出力する（図 2.5）。集計は毎日自動更新される。

- 一意な訪問者数（一意な IP アドレス数）

- 訪問回数（IP アドレス数。ただし 1 時間以上の間隔を空けての同一 IP アドレスからのアクセスは別の訪問する）
- ページ数（いわゆるページビュー）

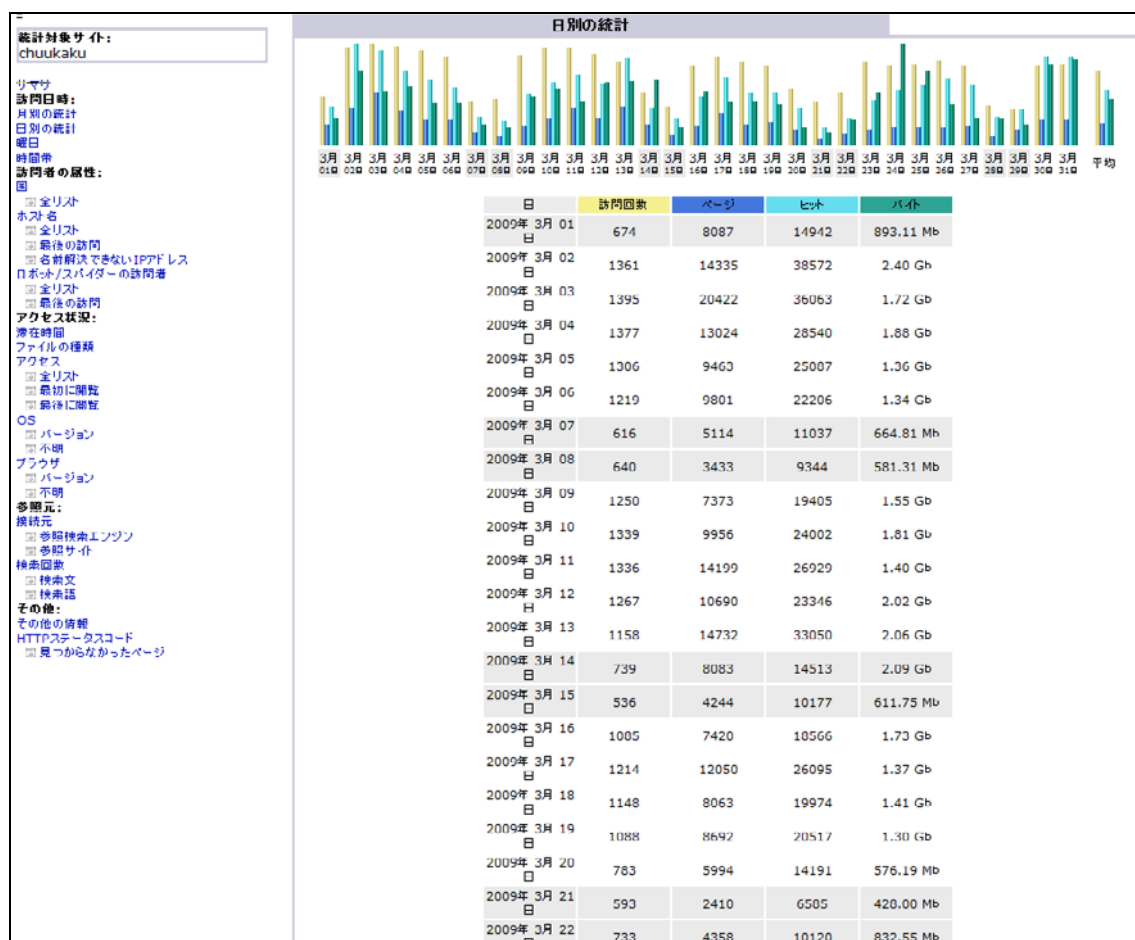


図 2.5 AWStats の出力画面例

集計値について分析すると、統合データベースプロジェクトポータルサイト (<http://lifesciencedb.jp> で始まる URL)全体の月間訪問回数は、2009 年 3 月の集計値は 31,795 であり、1 年前の平成 20 年 3 月の 3,807 と比較すると 10 倍近い増加となっている（図 2.6）。

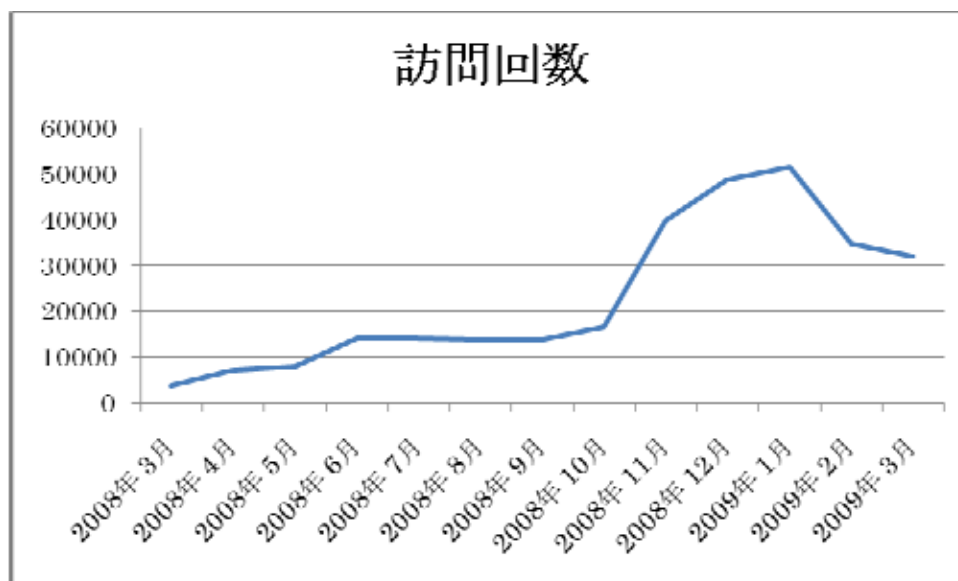


図 2.6 訪問回数の変化

また、サービス別では、生命科学データベース横断検索（2009 年 3 月月間訪問回数 14,346）、統合 TV(同 8,091)、DNA データベース総覧と検索(DDBJ/EMBL/GenBank)(同 4,233)が年間にわたりアクセス上位を占めている。さらに、訪問者のドメイン内訳（サービス開始時 2007 年 10 月からの累積）は図 2.7 が示す通りである。ac ドメイン（教育機関）と co ドメイン（民間企業）からのアクセスがほぼ同割合である。

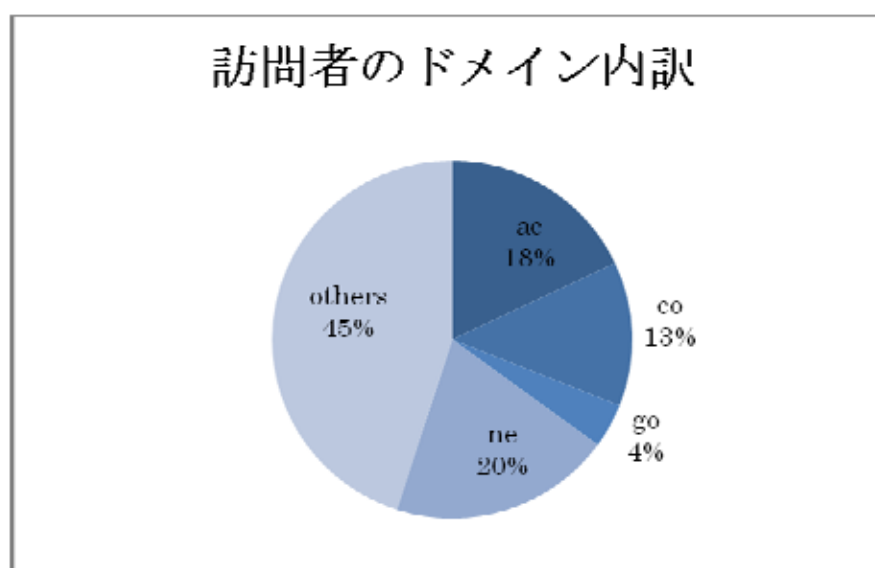


図 2.7 訪問者のドメイン内訳（サービス開始時 2007 年 10 月からの累積）

なお、分担機関についても、統合データベースプロジェクトポータルサイト

(<http://lifesciencedb.jp>)からリンクされているサービスについては、アクセスログの取得を依頼し、分担機関内で中核機関と同じ NCSA combined フォーマット形式で保存している。収集したログの中核機関への送付手順については、各分担機関と協議中である。

3. 統合データベース開発

3. 1 共通基盤技術開発（情報・システム研究機構、生命情報工学研究センター、奈良先端科学技術大学院大学）

3. 1. 1 共通基盤技術開発の実施計画

（情報・システム研究機構）

国内外のライフサイエンスに関するツールやデータベースについて、データの受け渡しの標準化や命名規則の標準化などのために必要な技術開発を行う。さらにこれらの標準化技術をより広くより容易に利用可能にするため、ツールへのアクセスを助けるユーザインタフェースの規格についての調査及びプロトタイプの実装を行う。また、文献を始めとするテキストデータや図表データ等の非構造化データから 各種生命科学情報を、より詳細に、精度良く、幅広い関連資源から抽出する要素技術開発及びそれを用いて生命科学者の調査・研究活動を支援するための応用技術開発、そしてこれらのために必要な調査・研究を行う。

（産業技術総合研究所生命情報工学研究センター）

利用者が得たい情報（知識）を、パソコンなどの端末から要求すると、必要なデータ、解析手法などを、国内、海外から自動的に選び、データベースと解析ツールのワークフローを作成し、最適な計算資源を使って解析を行う統合 DB 環境を構築する。

平成 20 年度は、平成 19 年度に構築した産総研内の計算機環境における統合 DB 環境のプロトタイプを拡張し、国内外の DB と連携した統合 DB システムを構築する。具体的には、プロトタイプおよび新規に開発するワークフローと、情報・システム研究機構の DB との統合を行い、それらのサービスを公開する。

（奈良先端科学技術大学院大学）

専門文書から専門用語を自動抽出するための手法の設計およびそのツール化を目指す。また、抽出された用語を管理するための辞書システムの設計を行う。具体的には、専門用語獲得技術と専門用語抽出ツールの設計を行う。

専門用語獲得技術については、専門用語辞書システムの設計と専門用語解析技術の開発を行う。前者については、専門用語の内部構造や同義性等の意味関係を管理するための辞書システムの設計を行い、後者については、専門分野のテキストに多く出現する並列構造の解析技術を精緻化し、用語の内部構造を解析する技術を開発する。専門用語抽出ツールの設計については、専門用語タグ付け手法の設計として、専門分野のテキストに対して、用語の出現箇所の特定制や用語の意味分類を行う自動タグ付け手法を設計する。

3. 1. 2 共通基盤技術開発の実施内容

（1）共通基盤技術開発（情報・システム研究機構）

国内外のライフサイエンスに関するツールやデータベースについて、データの受け渡しの標準化や命名規則の標準化などのために必要な技術開発として、ウェブサービスの設計指針を作成した(表 3.1 参照)。データ形式、エラー処理などの扱いがサービスごとにばらばらなことや、連携に必要な機能が備わっていないことが、ウェブサービスのスムーズな連携を妨げている。この指針は生物系のウェブサービスの相互連携を容易にするため、各サービスではどのような機能を最低限用意すべきであるか、またエラー処理の扱いをどのように統一すべきかなどの設計指針を示したものである。

表 3.1 ウェブサービスの設計指針

REST サービス 全般	1. データベース系のサービス提供においては REST での検索とエントリの取得を可能とすること、 2. クエリを統一するため、検索キーワードの書式は Common Query Language に準拠すること、 3. 検索キーワードからヒット件数が取得できること、検索キーワードからヒットしたエントリの ID のリストが取得できること、 4. エントリの ID から実際のエントリが取得可能であること、エントリが無かった場合 HTTP レスポンスコード 404 (Not Found) と空のエントリを返すこと、 5. サーバ側でサービスの処理中にエラーが発生した場合、HTTP レスポンスコード 505 (Internal Server Error) を返すこと、 6. REST の URL にはパス変数 (/foo/bar/baz など) を用い、クエリ変数 (param1=value1&param2=value2 など) は使用せず、パス変数には原則として小文字を使用すること、 7. URL が / で終わっている場合、原則としてスラッシュの後に続けることのできる文字列のリストを返すこと、 8. リストを返す場合は、エントリ単位で改行し、エントリ内はタブ区切りとすること。
SOAP サービス 全般	1. メソッド毎にマニュアルを書くこと、パラメータについて、ちゃんと動く値の例を示すこと、少なくとも Java, Perl, Ruby, Python での動作確認を行うこと、サンプルコードが提供されることが望ましい、 2. 非同期サービスはジョブ ID を発行すること、 3. サーバ側で処理中にエラーが発生した場合は SOAP Fault を返すこと、 4. 戻り値の型を可能な限り統一すること(配列 INSDC、発現 MIAMI、糖鎖 GLYDE-II、配列アノテーション GFF、key - value ペア タブ区切り)。
REST にするか SOAP にするか の指針:	データベースエントリの検索や取得については容易に URL にマッピング可能なので REST で提供することが望ましい。それ以外の解析サービスや実行時間のかかるサービス、さまざまな引数を必要とするサービスを提供する場合は SOAP を用いるとよい。

そして、平成 19 年度開発した TogoWS (<http://togows.dbcls.jp/>)の仕様をこの設計指針に準拠させた。

また、これらの標準化技術をより広くより容易に利用可能にするため、ツールへのアクセスを助けるユーザインタフェ이스の規格についての調査及びプロトタイプの実装を行った。具体的には、ペンシルバニア州立大学が開発したプラットフォーム Galaxy を中心にツール組み込みインタフェ이스の調査を行い、その結果を基に TogoWS の一部を組み込んだ DBCLS Galaxy をプロトタイプとして開発した (図 3.1)。この開発により、主にプログラム内部に組み込んで利用されてきた TogoWS を、プログラムを書くことなく、マウス操作で簡単に他のツールとも組み合わせて利用できるようになった。

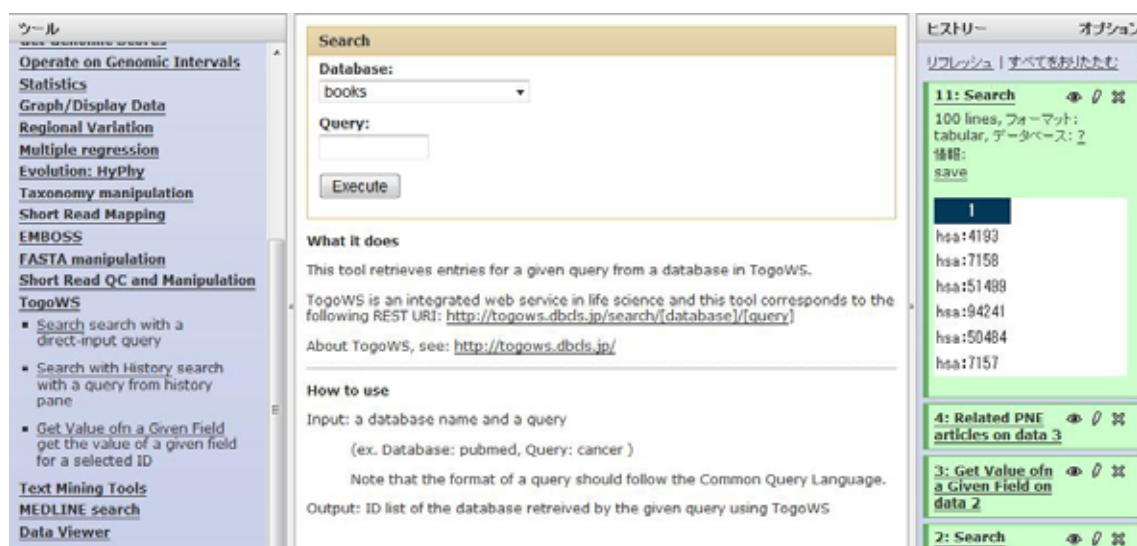
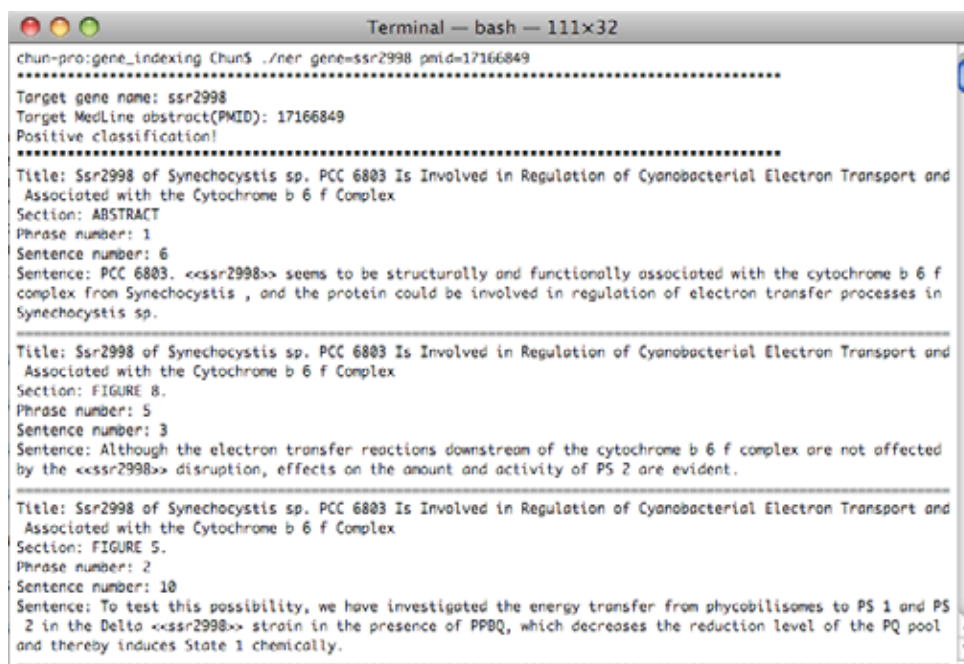


図 3.1 TogoWS を組み込んだ DBCLS Galaxy 画面

さらに、標準化技術および標準化を利用する技術をより安定化する目的で、開発合宿 BioHackathon 2009 を開催した。合宿内ではいくつかのグループに分かれて活発な議論と開発を行った。主なグループとして、上記ウェブサービス設計指針を改良するグループ、Galaxy の国際化や今後の展開について議論・開発するグループ、次世代シーケンサからの大量データをどう扱うかを議論するグループ、大量データの可視化について議論・開発するグループ、マイクロアレイデータ解析について議論・開発するグループ、実際の生物系データを持ち寄ってユースケースについて議論するグループ、セマンティックウェブをどう実装すべきか議論するグループ、文献情報のツール群について議論するグループなどがあった。最終的に多くのソフトウェア開発やコラボレーションにつながる大きな成果を得ることができた。グループや成果についての詳細は本報告書の 5. 3 項および <http://hackathon2.dbcls.jp/> を参照されたい。

また、文献を始めとするテキストデータや図表データ等の非構造化データから各種生命

科学情報を、より詳細に、精度良く、幅広い関連資源から抽出する要素技術開発として、フルペーパー内の藍藻遺伝子名を認識するツールを開発した。図 3.2 は本ツールを 87 本の藍藻遺伝子関連論文に適用した、検索システムプロトタイプの実出力例である。この検索システムでは、藍藻遺伝子名からの文献検索、PubMed ID からその論文に含まれる藍藻遺伝子名頻度情報取得、藍藻遺伝子名と PubMed ID からその遺伝子が論文内のどこに出現するかについての情報を取得することができる。



```
Terminal — bash — 111x32
chun-pro:gene_indexing Chun$ ./ner gene=ssr2998 pmid=17166849
*****
Target gene name: ssr2998
Target MedLine abstract(PMID): 17166849
Positive classification!
*****
Title: Ssr2998 of Synechocystis sp. PCC 6803 Is Involved in Regulation of Cyanobacterial Electron Transport and Associated with the Cytochrome b 6 f Complex
Section: ABSTRACT
Phrase number: 1
Sentence number: 6
Sentence: PCC 6803. <<ssr2998>> seems to be structurally and functionally associated with the cytochrome b 6 f complex from Synechocystis , and the protein could be involved in regulation of electron transfer processes in Synechocystis sp.

Title: Ssr2998 of Synechocystis sp. PCC 6803 Is Involved in Regulation of Cyanobacterial Electron Transport and Associated with the Cytochrome b 6 f Complex
Section: FIGURE 8.
Phrase number: 5
Sentence number: 3
Sentence: Although the electron transfer reactions downstream of the cytochrome b 6 f complex are not affected by the <<ssr2998>> disruption, effects on the amount and activity of PS 2 are evident.

Title: Ssr2998 of Synechocystis sp. PCC 6803 Is Involved in Regulation of Cyanobacterial Electron Transport and Associated with the Cytochrome b 6 f Complex
Section: FIGURE 5.
Phrase number: 2
Sentence number: 10
Sentence: To test this possibility, we have investigated the energy transfer from phycobilisomes to PS 1 and PS 2 in the Delta <<ssr2998>> strain in the presence of PPBQ, which decreases the reduction level of the PQ pool and thereby induces State 1 chemically.
```

図 3.2 藍藻遺伝子名認識ツール出力例

MEDLINE には 1800 万件にも及ぶ生命科学系の書誌情報が格納されているが、研究者がこの膨大な書誌情報から自身に興味のあるものを効率的に検索することは困難になっている。一方、多くの研究者は自身の研究分野における興味のある文献リストを保持している。そこで、この研究者の既に保持している文献情報を基に MEDLINE データ全体を対象として関連度の高いものを高速に検索し、提示するシステム TogoDoc を開発し、内部向けに公開した（図 3.3）。本システムは一般公開に向けて現在開発を継続中であり、文献管理機能やインタフェイスの検討を行うために、文献情報の推薦は NCBI の提供する PubMed Related Articles 機能を用いている。しかしながら、PubMed Related Articles は処理速度の遅さと仕様の不安定性、検索の硬直性などが問題であるため、DBCLS 独自に運用可能な文献推薦システムが必要と考えており、平成 20 年度は、一般的な文献推薦を行う既存システム Hotate（株式会社プリファードインフラストラクチャー社製品）を、MEDLINE の部分集合(2,753,786 件)に適用する試みを行った。その結果、全 MEDLINE データを対象としても動作可能である見通しを得ることができた。

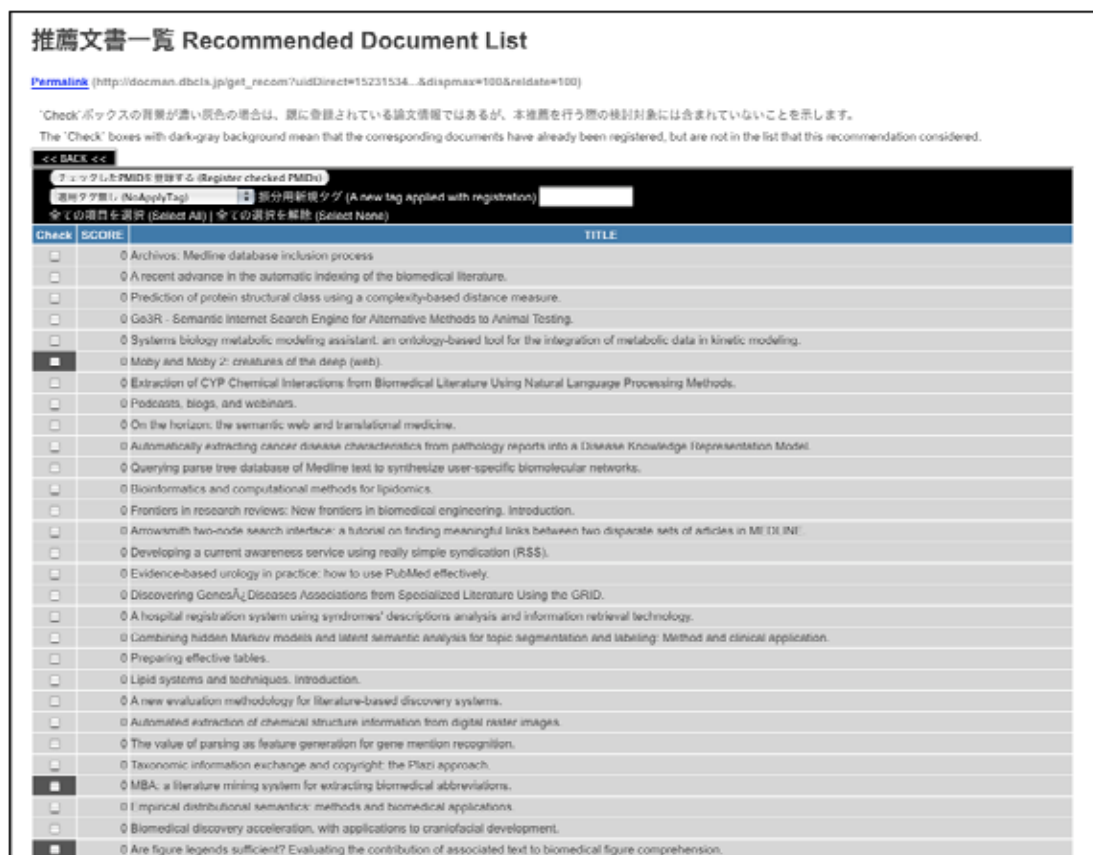


図 3.3 TogoDoc の画面：研究者のもつ文献情報集合を基にした推薦結果が示される。

(2) ワークフロー技術を用いた統合 DB 環境構築 (産業技術総合研究所生命情報工学研究センター)

本開発では、ユーザが利用する DB やソフトウェアの所在を意識することなく、様々なワークフローが動作し、情報を返すシステムを実現するための環境を構築する。平成 20 年度は、平成 19 年度に実施した調査・環境構築・プロトタイプ構築の成果を基に構築したグリッドを用いた統合 DB 環境に、以下の 3 つの一般公開用ワークフロー（ただし、ワークフローは固定）を開発し、実用に耐えうる統合 DB 環境ができているかどうかの確認を行った。

1) タンパク質構造情報ワークフロー (Protein Structure Information Workflow)

本ワークフローは、構造に関する予測プログラム等を Grid により効率的に分散処理し、非分散時の 1/5 以下の処理時間で行うものである。ユーザとしては、立体構造未知のタンパク質に関し、何らかの構造上のヒントとなる情報を必要とする実験研究者等を想定している。本ワークフローは平成 20 年 8 月 29 日に構築が終了し、プロジェクト関係者に限定

し公開した。また、本ワークフローは、次バージョンワークフローへの一部として位置付け構築した。下記の図 3.4、3.5 は、本ワークフローの概念図及び結果画面である。ユーザーからアミノ酸配列を受取り、二次構造予測、埋れ残基予測、フォールド認識、ディスオーダー予測を Grid にて分散処理を行い、結果を出力する。

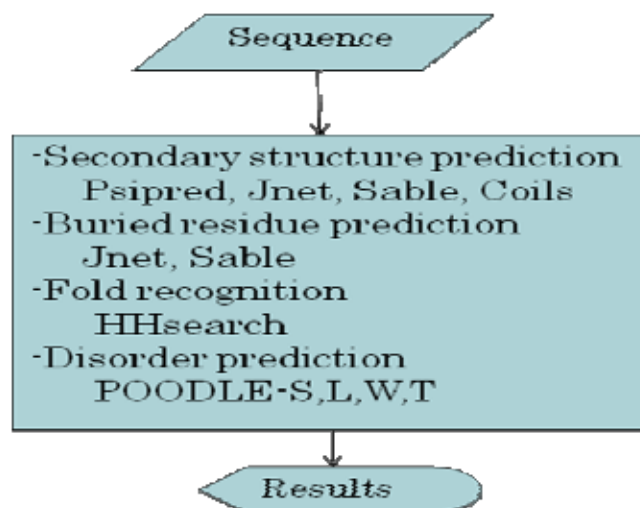


図 3.4 タンパク質構造情報ワークフローの概念図

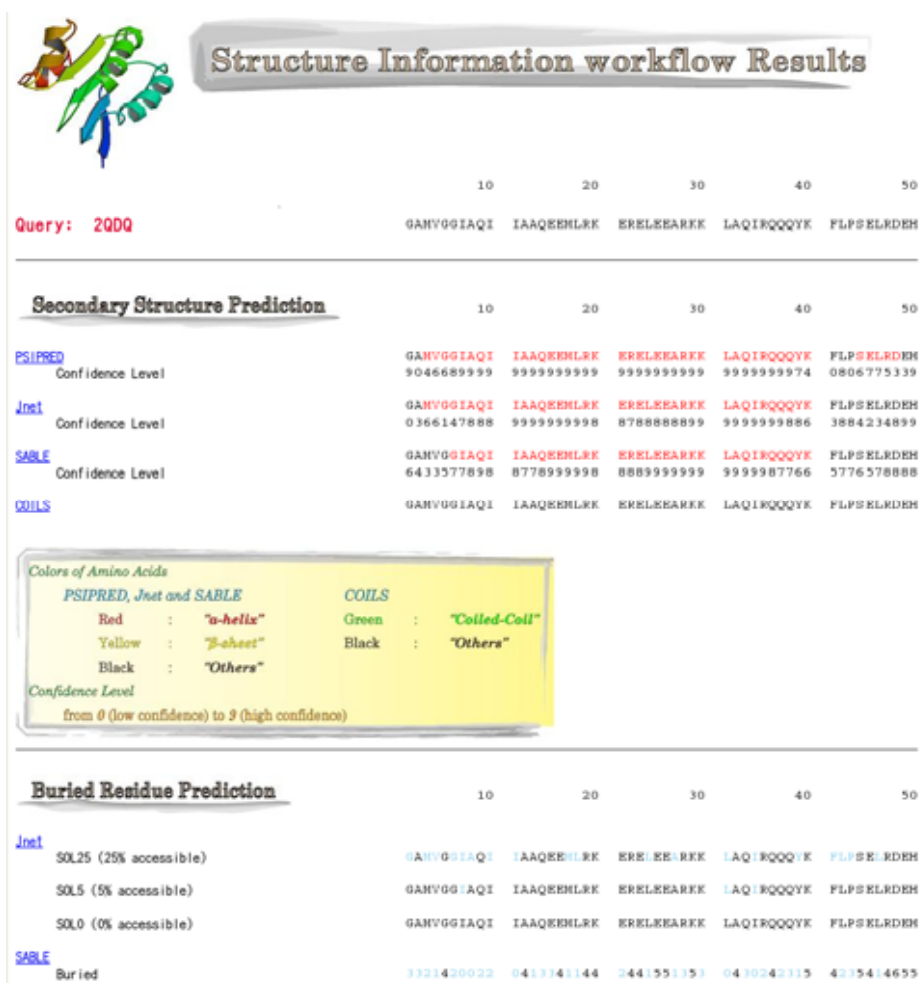


図 3.5 タンパク質構造情報ワークフローの結果画面

2) タンパク質アノテーションワークフロー (Protein Annotation Workflow)

本ワークフローは、立体構造未知のタンパク質に関し、構造及び機能のヒントとなる情報を幅広く実験研究者等に提供することを目的としており、1) のタンパク質構造情報ワークフローを発展・拡張したものである。1) と同様に各種プログラム等を Grid により効率的に分散処理し、従来と比し短時間で結果を表示する。本ワークフローは、平成 20 年 12 月 27 日に構築が終了し一般公開した。下記の図 3.6、3.7 は、本ワークフローの概念図及び結果画面である。ユーザからアミノ酸配列を受取り、二次構造予測、埋れ残基予測、フォールド認識、ディスオーダー予測、膜タンパク質オールベータ・ベータシート予測、細胞内局在予測を Grid にて分散処理を行う一方、データベース検索及び疎水性予測の実行を他のサーバへ依頼し結果を取得後、全ての結果をユーザが解析し易いよう配置し出力する。

CBRC Protein Annotation Workflow

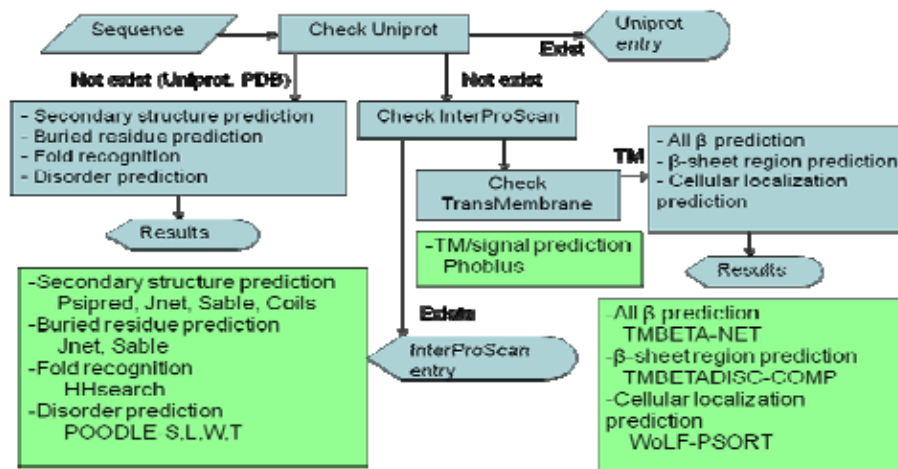


図 3.6 タンパク質アノテーションワークフローの概念図

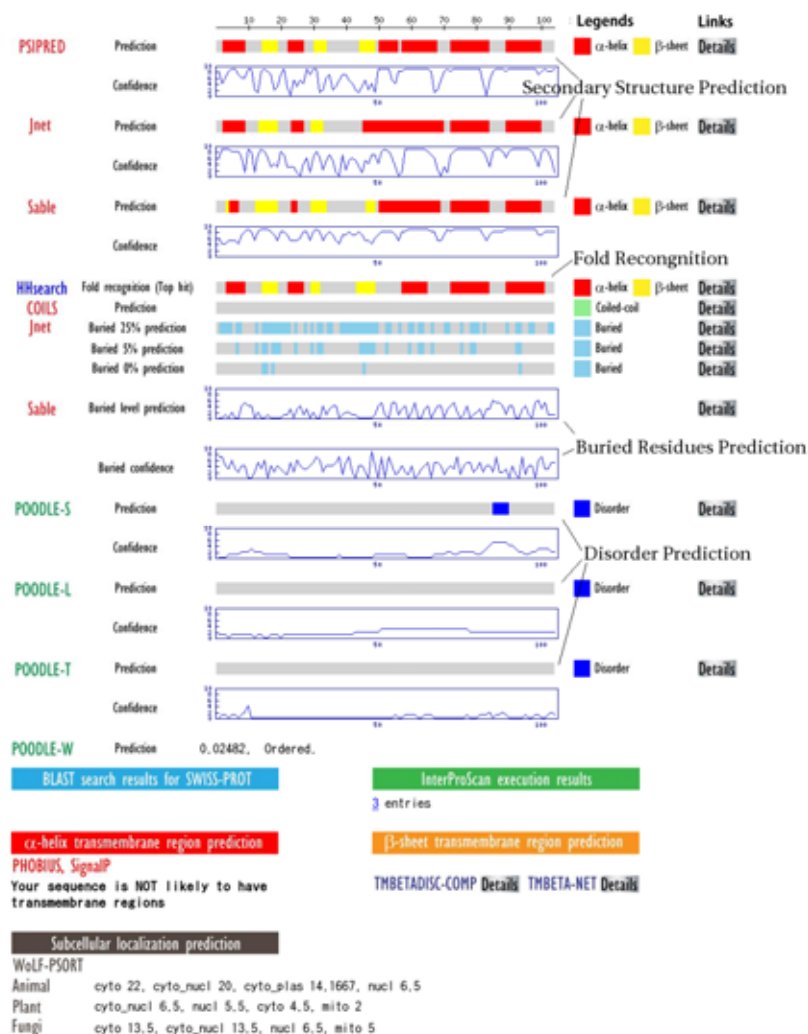


図 3.7 タンパク質アノテーションワークフローの検索画面

3) タンパク質比較情報ワークフロー (Protein Comparative Information Workflow)

本ワークフローは、相同なタンパク質を比較することで保存部位等構造上重要な部位を表示し、実験研究者等に提供することを目的としている。本ワークフローは、平成 21 年 3 月 31 日に構築が終了し一般公開した。下記の図 3.8、3.9 は本ワークフローの概念図及び結果画面である。ユーザからアミノ酸配列を受取り、相同タンパク質を検索、その結果からユーザがいくつかのタンパク質を選択し、マルチプルアラインメントを実行する。その結果より保存性が高い残基を表示する。また、二次構造予測結果も同様にマルチプルアラインメントし、保存性が高い二次構造を表示する。本ワークフローも 1)、2) で記載したワークフローと同様に Grid にて効率的に分散処理を行う。結果は立体構造が存在する場合はその上に表示する。

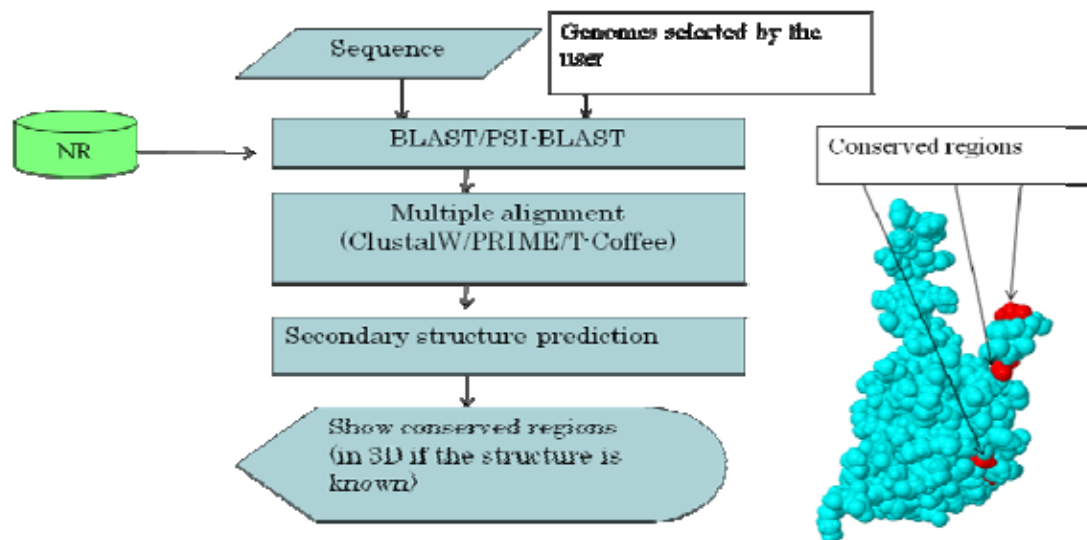


図 3.8 タンパク質比較情報ワークフローの概念図

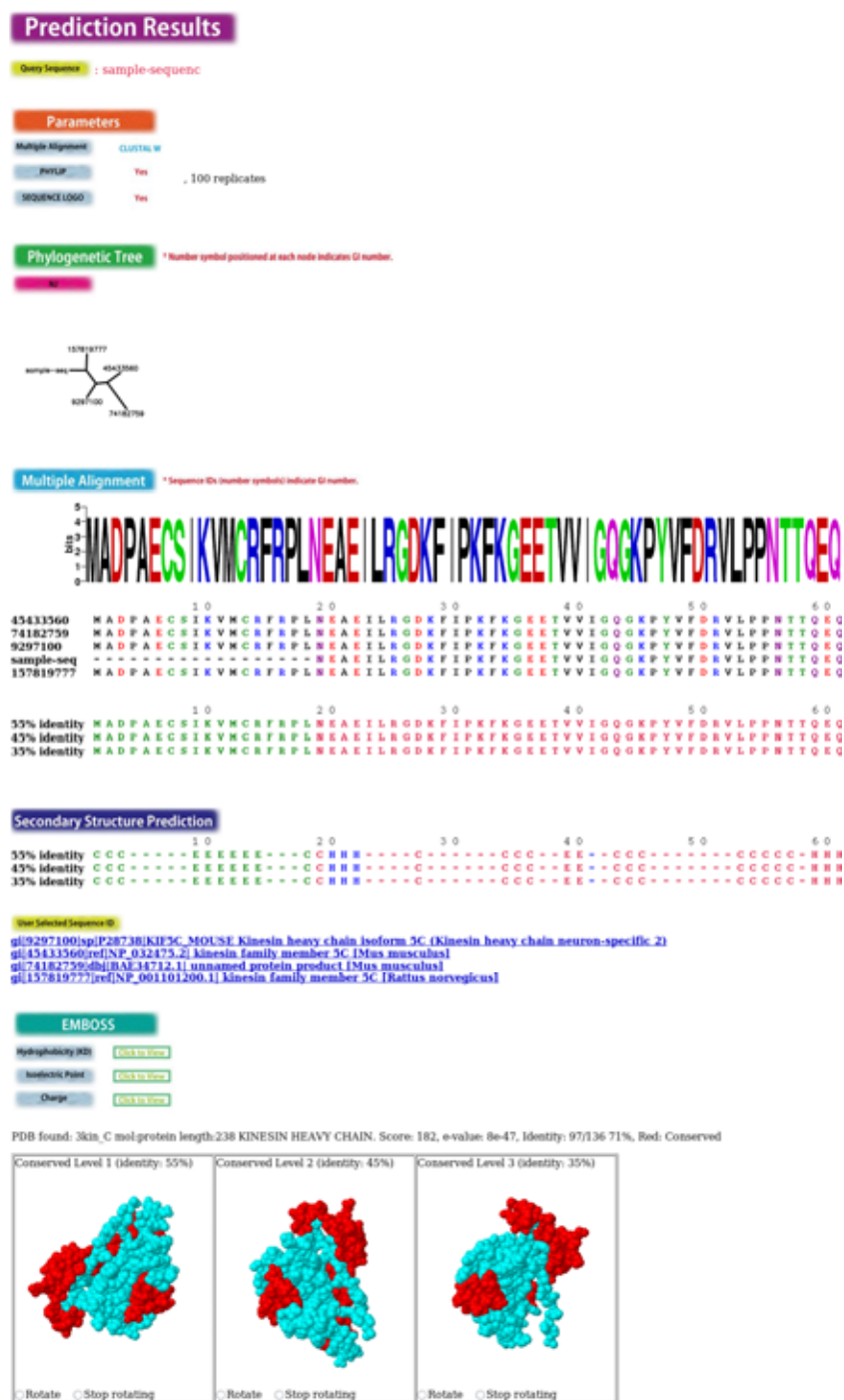


図3.9 タンパク質比較情報ワークフローの結果画面

(3) 専門用語辞書管理システムと専門用語解析技術の開発 (奈良先端科学技術大学院大学)

1) 専門用語辞書システムの設計

医学生物学分野の専門用語の多くは複合語であり、複雑な内部構造を持っている。複合語の内部構造は、それを構成する単語の間の係り受け関係として定義することができるが、係り受けの意味関係にいくつかの違いがある。それらを区別するために4種類の係り受け関係を定義し、内部構造のタグ付けを行うためのインタフェースをもつ辞書システムを設計した。医学生物学関係の用語では、特に、類似した2つ以上の語が並列して現れる場合に一部の漢字を共有して、縮退した形で表現されることがあり（例えば、動脈＋静脈が「動静脈」になる）、そのような並列表現の構造の記述法も考案した。平成19年度に設計した辞書管理システムに、縮退現象を含む複合語の内部構造タグ付け機能を実装し、約800語の専門用語（主に病名）に対して内部構造のタグ付け作業を行った。また、こうしてタグ付けされた複合語を訓練データとして、機械学習による複合語内部構造自動タグ付けプログラムを、文字間の係り受け解析として試作した。

図3.10に、複合専門用語の内部構造情報を管理し、表示するシステムのスナップショットを示す。この例は、「先天性溶血性貧血」の内部構造を表示している例である。また、図3.11には、文字の縮退を伴う複合語の例を示している。一つは、「角膜」と「結膜」が複合する際に末尾の文字が縮退して「角結膜」という語になる例、もう一つは、「大腿骨」と「骨折」が複合する際に前者の末尾と後者の先頭文字が縮退する例である。設計したシステムでは、内部の構成語として縮退する前の単語を指定し、どのような縮退現象が生じたかを記述することによって、このような文字縮退を生じる複合語の記述法を提案した。さらに、文字単位の係り受け解析を用いることにより、このような現象をもつ複合語の内部構造解析が可能であることを示した。

さらに、表記が異なるが同じ概念を指す同義語について、それらを辞書システム内部で同じ識別子を与えることによって同義性を記述する方法を設計した。今後は、表層の表記が異なる同義語に関する情報を辞書に蓄積していく予定である。



図3.10 複合語内部構造表示のスナップショット

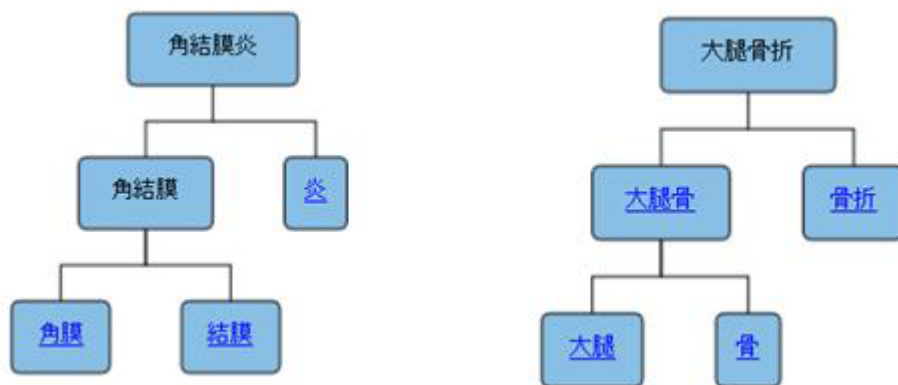


図3.11 単語内の文字縮退を伴う複合語の内部構造記述例

2) 専門用語解析技術の開発

専門分野のテキストには、新しい用語が次々に誕生するだけでなく、既存の用語に関しても、複数の類似の用語が並列表現として出現する場合や部分的に省略が行われる場合には、完全な形で用語が記述されるとは限らず、表現上での並列構造の解析技術が必要となる。そのため、用語辞書への新しい専門用語の登録を充実するだけでは完全な用語解析を行うことができない。特に、並列表現の問題に対応するため、英文および和文の専門文書に現れる並列構造を解析する方法を設計した。

まず、専門用語の並列構造解析技術の開発を計画した。専門分野のテキストには、用語の内部だけでなく、一般的な表現の中にも様々な並列構造が現れる。並列構造は類似の単語列をもつが、どのような単語列が互いに類似しているかを客観的に定義することは容易ではない。並列構造のタグ付けが行われた実データを対象にして、系列アラインメントの部分系列対応のための重みを自動学習する手法を提案し、文中の一般的な並列構造解析を高精度で行う手法の開発を提案した。さらに、複数の並列構造が1文に現れる場合に、それらが重なり合わないという制約を満たすために文法規則と上記手法の融合を行い、医学生物学分野のテキスト解析技術を精緻化した。提案手法では、並列構造を認識する文法規則に基づいて文を解析し、並列構造と認定されようとする部分の系列アラインメントによるスコアを動的に計算し、文法制約を満たしつつ、系列アラインメントによって最大のスコアを与える並列構造を同定する。図3.12に、系列アラインメントと文法制約を融合した並列構造解析の実行例を示す。

(九州大学)

現在の大規模ジェノタイプデータを用いた疾患遺伝子関連解析、集団遺伝学解析研究では、データの信頼度をそのデータと共に客観的且つ検証可能な形態で明確に提示することが求められている。そこで日本人ゲノム多様性に関するプライマリーデータベースとして機能する多型知識表現 DB を構築してウェブ上に公開し、ヒトゲノム多型知識情報を生命科学・バイオインフォマティクスコミュニティの利用者に安定的に供給する方策を確立する。このために DNA アレイ技法及びビーズアレイ技法によって収集された大規模一塩基多型 (SNP)・コピー数多型 (CNV) 情報の一次データの構造を検討し、信頼度検証可能な形態でかつ個人情報保護に配慮した多型データの表現形式を明確化する。またこれをもとに我国の種々の研究組織で決定された大規模ジェノタイプデータに関して、これを利用する立場でクオリティーチェックを独自に行い、即ち収集試料の偏りの程度、ジェノタイプ決定操作におけるエラーのレベル、SNP 及び CNV 判定での確度等を定量的、客観的に検討して、それらの過程及び結果を提示した共有データベースを構築する。

3. 2. 2 ヒト統合 DB の開発・運用の実施内容

(1) ヒト統合 DB の開発・運用 (情報・システム研究機構)

医学研究や創薬開発研究に携わる利用者がゲノム情報、分子データベース、文献情報等の、研究開発に必要な情報を容易に統合利用するための背景知識の整理を行うためには、メタデータの著作やアノテーション付与を行う専門作業者の作業効率を上げ、知的作業の多くの部分を再利用可能な形で記録する必要がある。そのために、雑誌別に様々な html 形式で提供される電子論文の序文、方法、結果、考察、文献などの情報を同定し、引用文献は PubMed の ID まで同定したうえで共通のタグつきテキストとして出力する、主要雑誌を対象とした処理プログラムを開発した。本開発では、主要な学術論文提供プラットフォームである PubMed Central, Blackwell Synergy, Science Direct が HTML 形式で提供している論文データ、および、Nature 各誌、Lancet などの独自フォーマット有する論文データに対応することにより、広範な雑誌に対応するプログラムを開発した。また、引用論文を PubMed に関連付けて NCBI から情報を取得する機能や、論文の本文中から、引用論文に係る文やその周辺の文を抽出する機能など、論文の引用関係や引用内容の解析に利用可能な情報を出力する機能を開発した (図 3.13)。

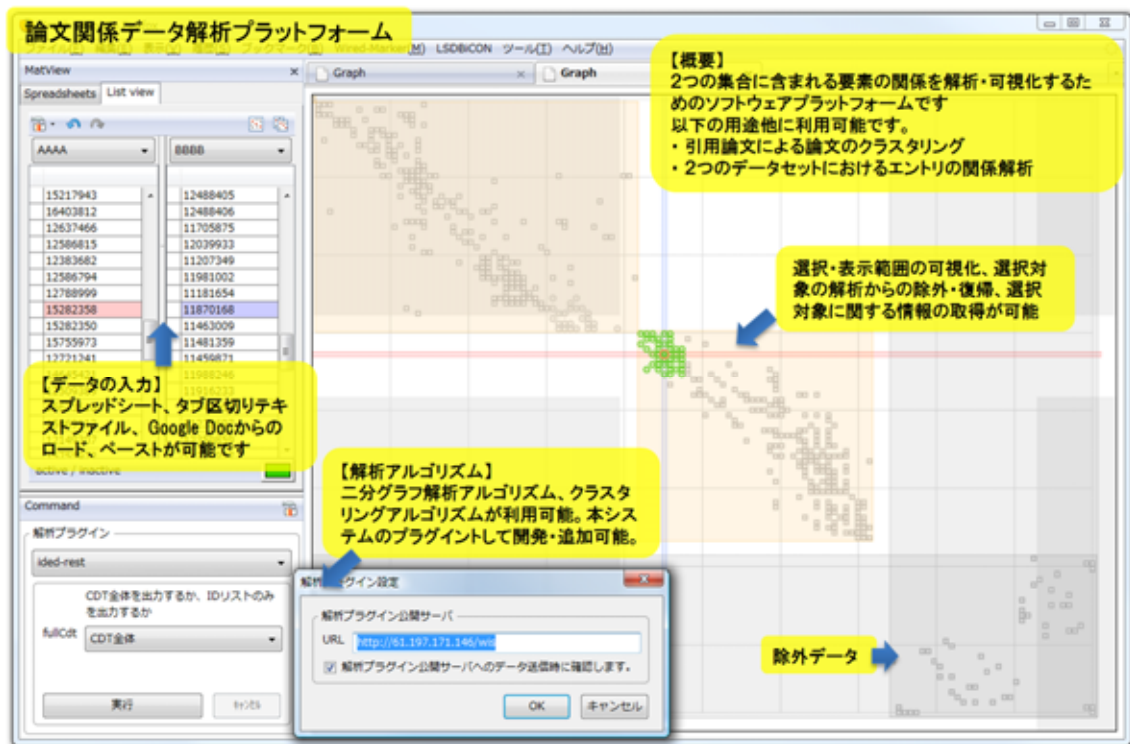


図 3.14 論文関係データ解析プラットフォームの概要

さらに、昨年度までに開発した Wired-Marker の version1.0 に対して要望の多かった、テキストとしてマーキングできないページを画像データとしてインデックスする機能、ディレクトリ管理機能を追加した。マーキングできないページを画像としてインデックスする機能は、Web ブラウザに表示中の情報や、クリップボード中の画像情報の指定領域をインデックスする機能であり、サードパーティーソフト等によって Web ブラウザ上に表示されたコピー＆ペースト操作が不可な情報を Web ページに関連付けて保存することを可能とした。また、情報収集作業において、利用する情報分類フォルダをカレントフォルダとして選択し、配下のフォルダだけを表示することにより、マーキング作業の効率化を可能とする機能を開発した（図 3.15）。

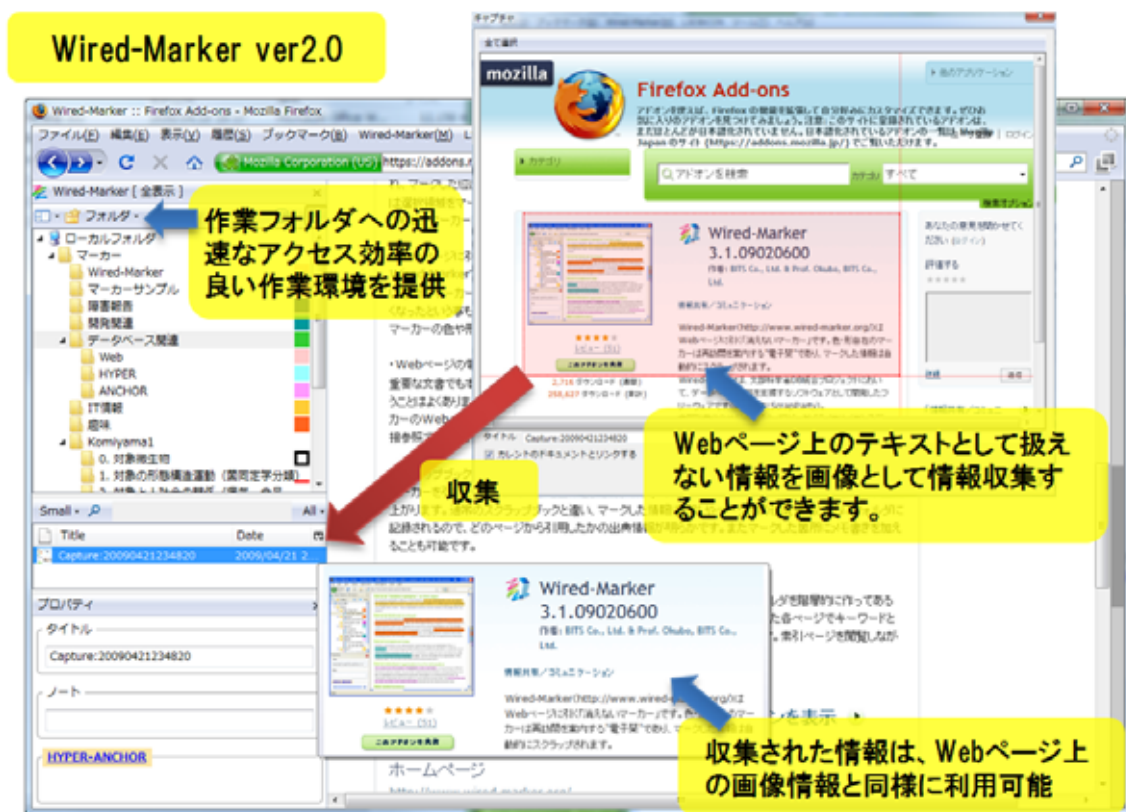


図 3.15 Wired-Marker ver2.0 の概要

また、本プロジェクトで作成した用語辞書による辞書引きをアノテータがフロントエンドで容易に行えるファイアーフォックスのプラグインも開発した。本開発では、外部で作成された辞書を読み込み、Web ブラウザに表示された情報から、辞書に存在するタームを探索し、見つかったタームをインデックスとして登録する機能、インデックスとして集められたタームを辞書に追加する機能、辞書の内容を管理する機能を開発した。本機能を用いて、Web ページとして公開された情報に基づく辞書の作成・メンテナンスが可能である (図 3.16)。

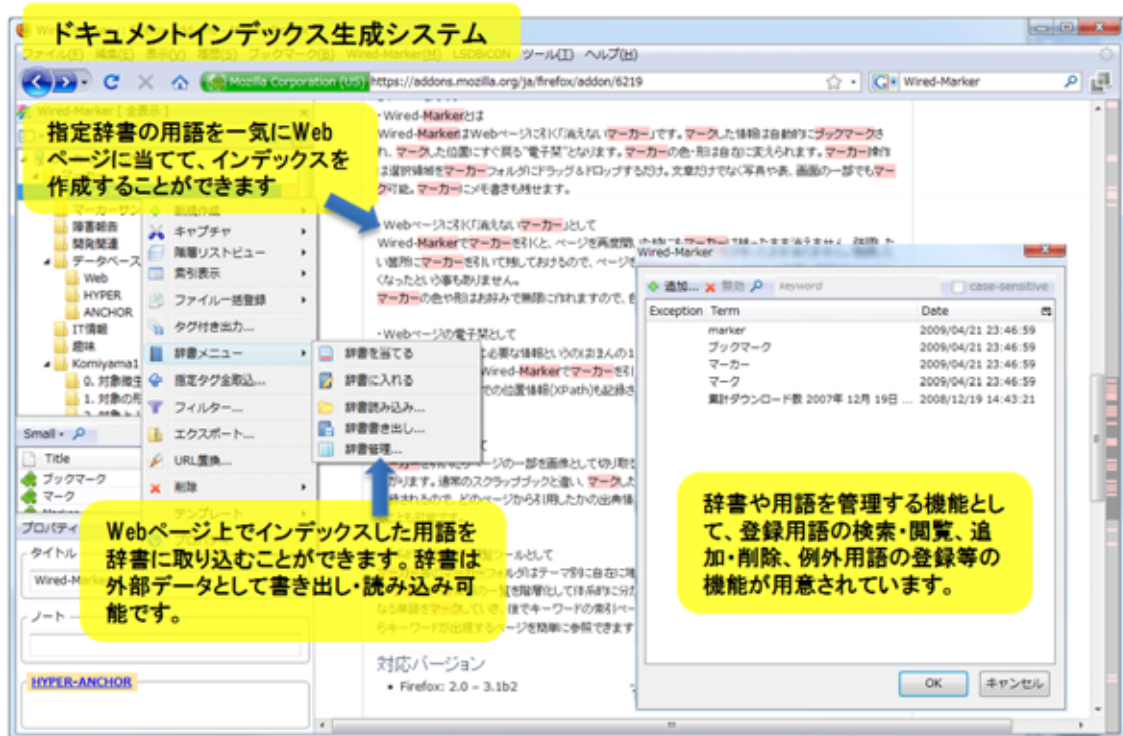


図 3.16 ドキュメントインデックス生成システムの概要

解剖用語は、生命科学の専門用語のうちで生物種名と並んで語数が多いにもかかわらず、体系化がされていないため、医学系の文書の検索や医学系のデータの統合利用の障害となっている。そこで、用語整理とともに用語の定義や索引利用のための構造化を試みてきた。今年度は、BodyParts3D データベースの充実とそれを利用、表示するためのシステムの開発を行った。

前者については、精度の高い教育用模型をスキャンニングし、頭部、腹部、肝臓の詳細デジタル化を行った。複数の段階（フェーズ）に分けてデータベースの構築を進めている。平成 19 年度までに、全身 MRI 断層データから再構築された解像度 2mm 立方の 3 次元（ボクセル）データ（臓器分解数約 50）を 3 次元 CAD (Free Form Modeling Plus、米国 SensAble Technologies 社製品)を用いて編集し、約 230 パーツの全身人体 3 次元モデル（ボディパーツ）をフェーズ 1 データとして作成した。平成 20 年度は、各部位を詳細化したフェーズ 2 データの作成を進めた。まず精度の高い教育局所模型(頭部、腹部、肝臓)を 3 次元スキャナーでスキャンニングし詳細 3 次元デジタルデータを取得した。次に Free Form Modeling Plus を用いて、平成 19 度にスキャンニングした骨格データとともに、フェーズ 1 全身モデルの当該部分と置換した。その際には、解剖学教科書等で定義されている構造上の目印（メルクマール）（例：臓器と肋骨の高さ等）を外さないように、すなわち、個人差を除いた標準的人体モデルとして解剖学的に正しいデータとなるように注意した。図 3.17 は、心臓のエントリ例である。また、モデルの解剖学的正しさの向上だけでなく

く、臓器の部分領域（例えば、心臓の場合は左心房、右心房等）の分割も進め、臓器モデルとして登録した。その結果、2009 年 3 月時点では、フェーズ 1 データも含めた合計が 642 パーツとなり、1 年前の約 230 パーツの約 3 倍に増加した。その大部分が、フェーズ 2 データ(567 パーツ)である。各器官系別の詳細は表 3.2 の通りとなっている。筋肉系も約 300 パーツがほぼ完成しており、平成 21 年度前半には公開する予定にしている。他の解剖学データベースとの関連づけに関しては、University of Washington で開発され、デファクトとなりつつある解剖学オントロジーである FMA(Foundational Model of Anatomy) とのリンクづけを行っている。英語名称も FMA を参照している。日本語名称については、日本解剖学会で定義している解剖学用語（Terminologia Anatomica Japonica）を主に参照した。さらに、フェーズ 2 臓器の 3 次元モデルをポリゴンデータ(VTK フォーマット、STL フォーマット)の形式で公開した。ユーザは自由にダウンロードでき、3 次元モデルを用いた手術シミュレーションや電子カルテなどのシステムに組み込むことができる。ただ、現時点でのモデルの完成度は医療現場で用いるには実用レベルになく、完成度の向上は今後の課題である。

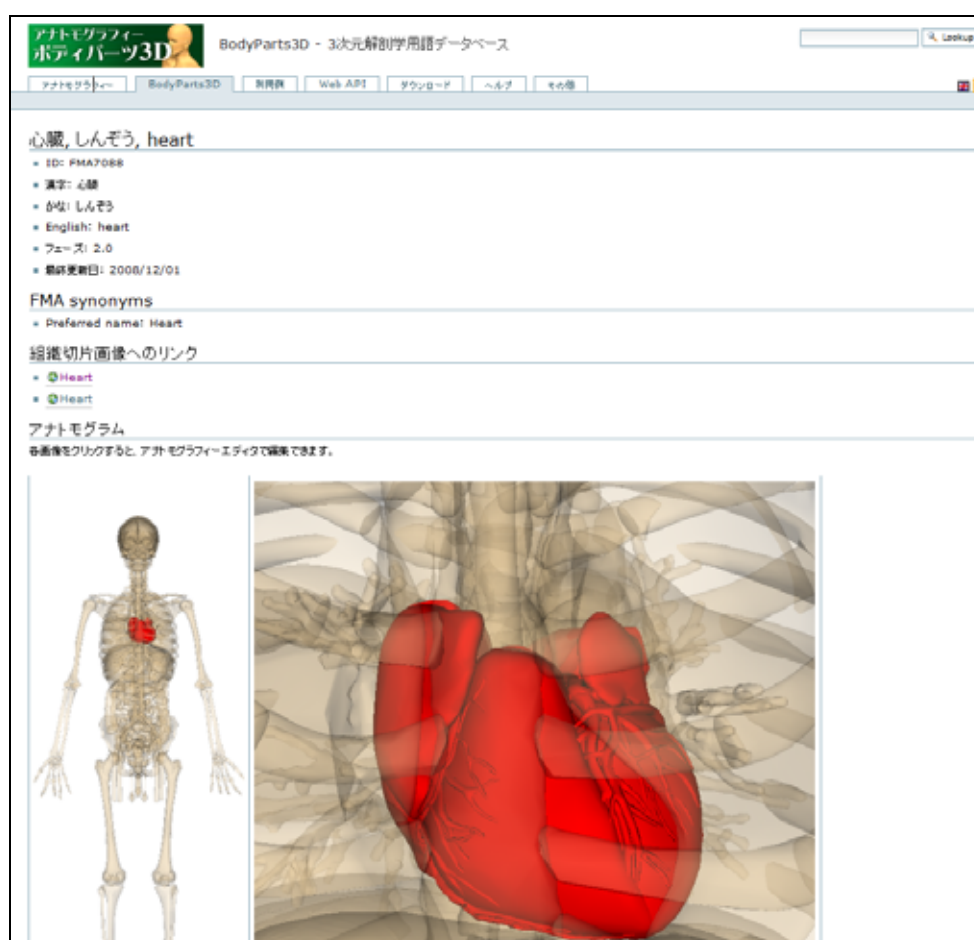


図 3.17 心臓のエントリ例

表 3.2 2009 年 3 月時点での公開済みの 3 次元臓器モデル数

器官系名	フェーズ 1	フェーズ 2	合計
神経系	2	53	55
感覚器系	18	0	18
循環器系	47	97	144
呼吸器系	0	45	45
消化器系	0	32	32
内分泌系	2	13	15
免疫系	0	2	2
泌尿器系	0	9	9
生殖系	0	18	18
骨格系	0	272	272
筋肉系	0	2	2
皮膚/結合組織	3	0	3
その他	3	24	27
合計	75	567	642

後者については、ボディパーツデータを自由に選択して、利用者が必要な身体局所の自由なレンダリング画像（3D 立体を写真的に 2D に表現した画像）を作成し、保存、交信可能にするための 3D データの高速レンダリングサーバの開発及びウェブインタフェイスの開発を行い、公開運用を始めた。レンダリングサーバは、BodyParts3D データベースのポリゴン情報を組み合わせて、自由な座標系においてレンダリング可能な OpenGL ベースの 3D 解剖学 CAD システムとして開発した。それにより、拡大・縮小、平行移動、切断面の生成、イメージ内へのマーカーピンの設置など、自由度が高くかつ高速な図形処理が可能となった。また、レンダリングサーバの機能を利用するためのユーザインタフェイスの開発を行い、BodyParts3D を組み合わせて自由度の高い解剖イメージを生成し、また、イメージ上にマーカーとコメントを付与する機能を開発した（図 3.18）。

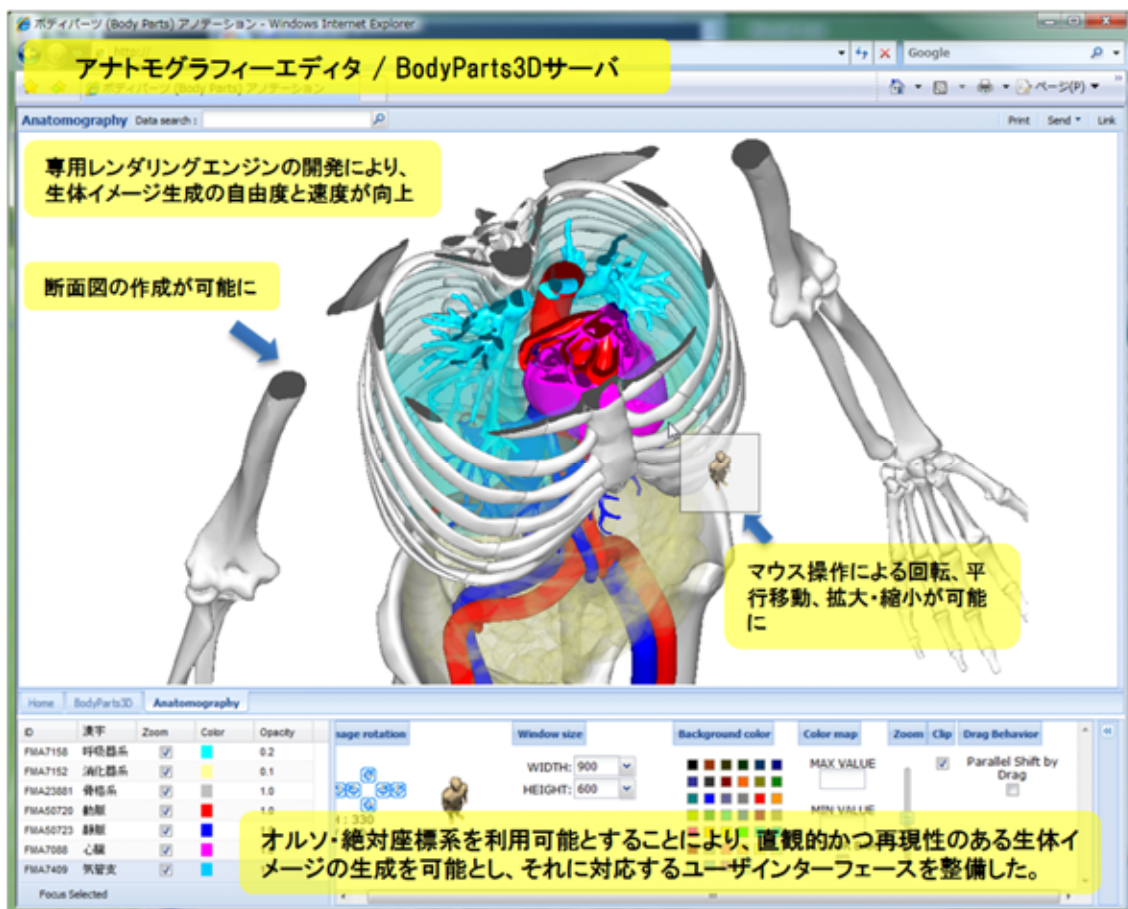


図 3.18 アナトモグラフィーエディタ/BodyParts3D サーバの概要

一方、ボディパーツ 3D データは解剖学知識および画像診断学知識を反映したものでなければならぬ。そのため 10 名程度の臨床医によるデータの精度や利便性に関するレビューを実施した。その結果、本取り組みの将来には期待できるとしながらも、現在の解像度、位置関係等の正確さが、臨床医のレベルからは必ずしも満足のいくものではないとの意見が大半であった。

また、わが国の国立公立研究機関のうちでライフサイエンス分野に関係する研究所、研究室を網羅した機関辞書および機関データベースを充実させ、研究機関情報データベースシステムとして再構築した（図 3.19）。併せてインターネットブラウザのレベルで機関情報が文書中に表現可能なプラグイン辞書化も行った。研究機関情報の収集には、15,000 件以上の研究機関・研究グループについて、各組織がホームページに掲載している情報を中心に収集を行い、Web ページのサムネイルと共に各種情報を利用可能とした。また、学会要旨データベースより収集した機関名別名情報を格納し、機関名の表記ゆれに対応した検索を可能とした。

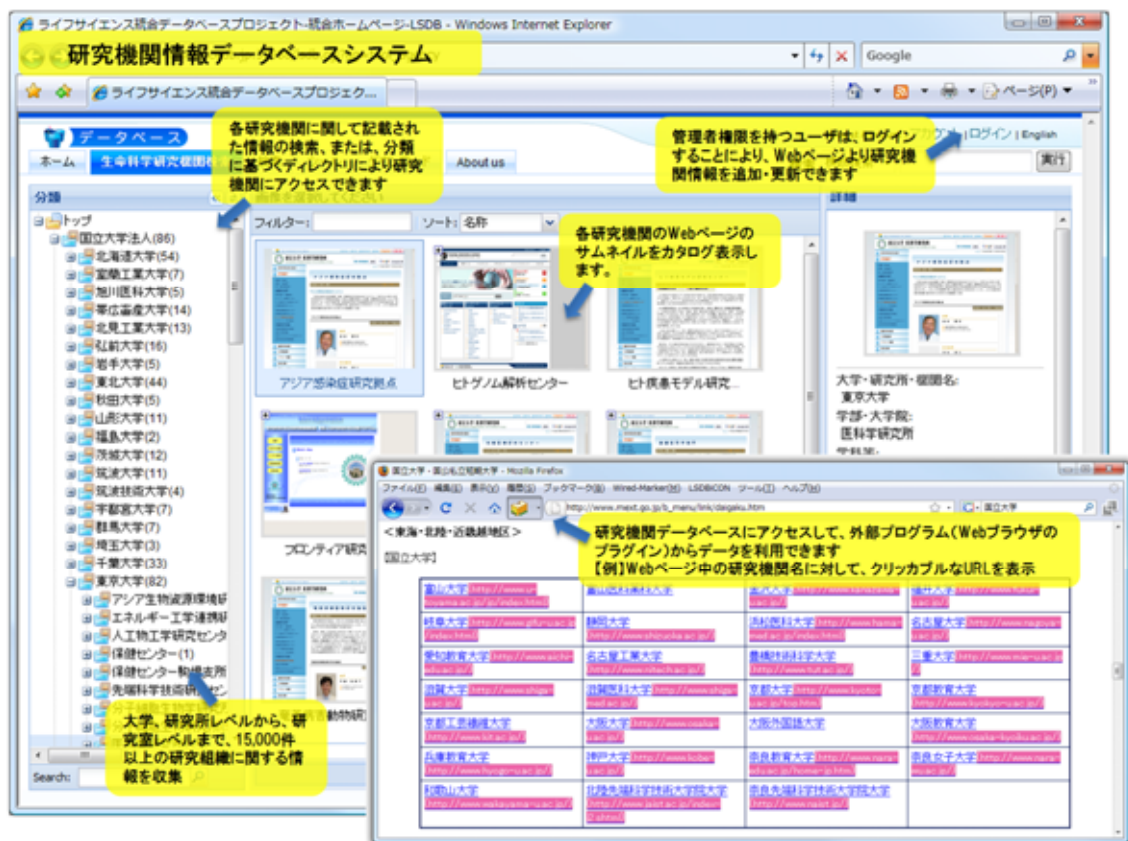


図 3.19 研究機関情報データベースシステムの概要

データベースの整理、統合の一環として進めている国際 DNA バンク目次については、DNA バンク目次および GEO 目次について 3 ヶ月に一回のデータ更新を行い、サービス運用を行うと同時に、サーバの整備を行った。公開サーバの移行に伴い、データ移設に関してサーバ管理者間の協議を行い、併せてパッケージソフトウェアの導入作業を行なって、国際 DNA バンク目次システム、GEO 目次システム、植物ボディーマップ、H18 年度成果公開サイト、生物学名・和名・一般名対応辞書 web 検索システム、サイト内一括検索システムをそれぞれ移行した。遺伝子発現バンク (GEO) 目次、BodyMap-Xs、植物ボディーマップのデータ更新をそれぞれ行った。また BodyMap-Xs に於いて、RSS 1.0 形式に則った更新情報の配信機能を実装した (図 3.20)。分子データのうちでもっとも大量で形式の多い核酸配列バンクについて DDBJ と異なるメタ情報からの整理提示、検索利用を提供するために、全バンクレコードに対する文字列検索システムを開発し、運用に移行した。バンクレコードを纏め上げたプロジェクトを検索の単位とすることで、検索対象を EST ディビジョンを含むように拡大させ、処理時間を約 110 秒だったものを約 8 秒までに約 13 倍ほど向上させた。更に、検索結果情報をキャッシュとして保持しておくことで、同一の検索語に対する応答時間を約 0.7 秒程度にまで約 150 分の 1 程度に短縮させた (図 3.21)。

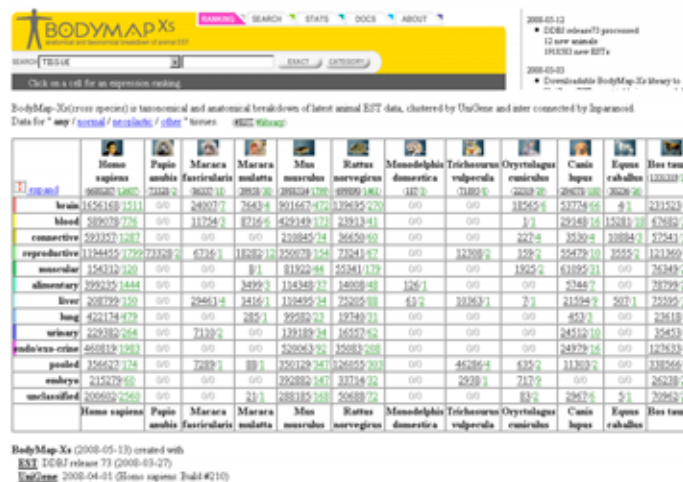


図 3.20 RSS 1.0 形式に則った更新情報の表示



当該データに関わる審査内容に従い承認したデータ開示希望者に、ユーザID、パスワードを与え、閲覧及びダウンロードを可能にすることを意味する。第三者であるためには共同研究者であってはならない。そこで個人ゲノム情報に関する倫理規定に抵触せず遂行可能な業務形態として、QCに作業を限定し、作業終了後は取得データを抹消する、そして解析過程とその結果のみを閲覧及びダウンロード可能とすることを前提とした。

QC作業の内容としては、客観性を確保するために、内容を明示された既定のQCパイプライン（JAGQC pipeline）を用いた（下記に詳述）。またデータ提供を促進するために、解析結果はデータ提供元の同意を得ることを前提とし、これらの条件をデータベース上に明記した。

JAGQC Pipelineでは以下の2段階のQCを行い、それらの解析過程（log file）及び結果をダウンロード可能な形態で提示する。＜一次QC＞受け入れたジェノタイプデータに対してGAINQCプログラム（米国GAIN Projectのために開発されたものである）を用いて sample QC、SNP QC、及びkinship QC sample SNP call rateをもとにデータフィルタリングを行う。＜二次QC＞一次QC／フィルタリングの後のデータセットに対してPLINKプログラム（国際的なデフォルトスタンダードとなっている関連解析プログラムである）を用いて、ジェノタイプ失敗率の偏りの検定（Test of missingness by case/control status）、集団の均一性の検討（multidimensional scaling）、及び同プログラムのデフォルト設定での暫定的関連解析（association）を行う。

現在、JAGQC db では、東京大学人類遺伝学教室から提供された一般日本人 200 人の Affymetrix SNP Array 6.0 を用いた約 100 万 SNP に関するジェノタイプ生データ（.cell ファイル）をモデルとして、上記 JAGQC Pipeline による QC 解析と関連解析を行い、ウェブ上で解析結果を閲覧、ダウンロードできるサイトを構築した

（http://crane.gen.kyushu-u.ac.jp/JAGQC/JAGQC_top.html）。図 3.22 に JAGQC db の画面の例を示す。なお、本サイトの利用には、データ提供元の倫理審査委員会等の承認を得ることが前提になる。ここで提供する情報はいかなる委員会の規定においても開示可能な内容である。今後、順次承認を得た希望者に開示をしていくことになる。

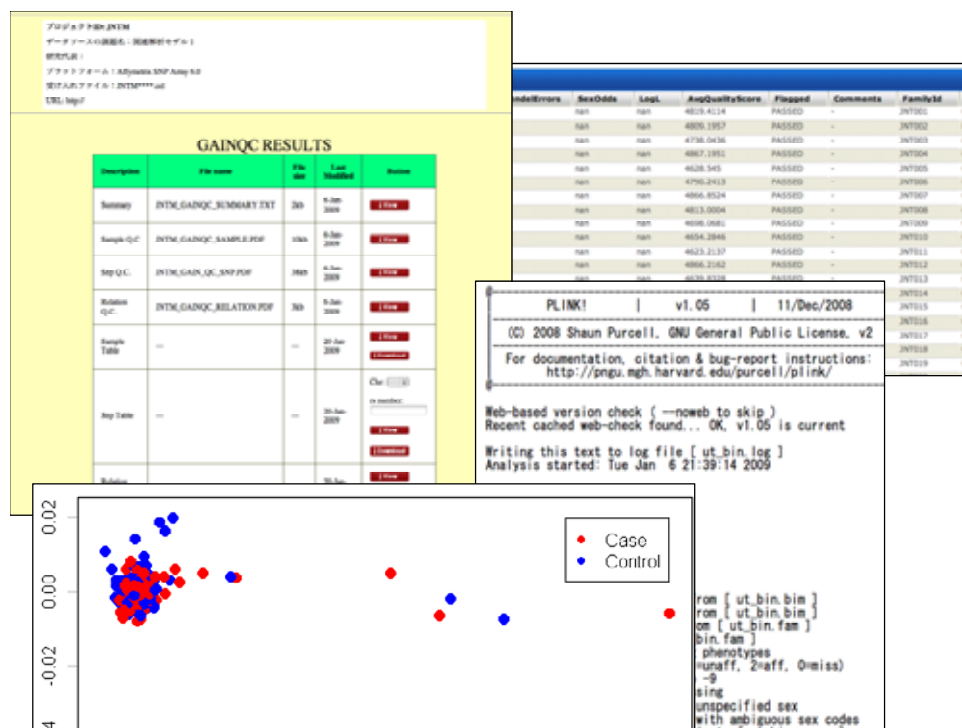


図 3.22 JAGQC db の画面の例

3. 3. モデル生物・産業応用生物統合 DB の開発・運用

(情報・システム研究機構、かずさ DNA 研究所)

3. 3. 1 モデル生物・産業応用生物統合 DB の開発・運用の実施計画

(情報・システム研究機構)

基礎生物学研究および生物の産業応用研究に携わる利用者がゲノム情報、分子データベース、文献情報等の、研究開発に必要な情報を容易に統合利用できるように、背景知識の整理を行い、辞書やオントロジーの形で公開する。 またこれらを用いて データベースの整理、統合を進める。

(かずさDNA研究所)

平成19年度に構築した植物および植物関連微生物のゲノムデータベースを改善し、利便性の向上を目標とした業務を行う。 ユーザからの情報集積とそれを基にした整理を可能とすることを特徴とするゲノムデータベース環境を活用し改善を加えつつ、データの集積によるデータベース統合化を複数の生物種について進める。

1) 高度情報集積データベースの運用と改良

平成19年度に構築した、分子データ上のポジションを統合のための基盤情報とし、アノテーション・キュレーションの統合と高度化を可能とするデータベースを植物および

植物関連微生物のゲノム解析情報を対象として運用しつつ、ユーザの要望を汲みあげて改善を加え、利便性の向上を行う。同時に、集積したアノテーション情報から生物学的な意味を取り出すための情報表現技術の開発と、それを閲覧するためのビューワの開発と提供を行う。また、データベース中の遺伝子名や遺伝子IDの名称の食い違いを解決するサービスを維持し、充実を図る。

2) ゲノムアノテーション情報の蓄積と高度化

1) のデータベース環境を運用しつつ、その環境を利用したアノテーション情報の蓄積と高度化を引き続き進める。平成19年度には光合成細菌であるシアノバクテリア *Synechocystis* sp. PCC 6803 のタンパク質遺伝子の論文記載情報を約1200報分について登録したが、平成20年度は、複数のシアノバクテリアおよび高等植物の遺伝子情報について、引き続き論文記載情報および実験情報の蓄積業務を実施し、ゲノムの位置情報と論文記載情報の統合を図る。登録した情報は速やかに公開していく。

3. 3. 2 モデル生物・産業応用生物統合 DB の開発・運用の実施内容

(1) モデル生物・産業応用生物統合 DB の開発・運用（情報・システム研究機構）

平成 19 年度に調査した微生物ゲノムアノテーションの国内外の状況と急速に普及し始めたいわゆる新世代シーケンサがもたらす研究動向の見通しのもとに、基礎生物学研究および生物の産業応用研究に携わる利用者がゲノム情報、分子データベース、文献情報等の研究開発に必要な情報を容易に統合利用して、データベースの整理、統合を進めることができるような環境構築の第一歩として、微生物ゲノムアノテーションパイプラインを構築して広く一般に公開することにした。すなわち、質の良いデータの生成を支援することによってデータの再利用を促進しようとする試みである。シーケンサの技術開発の進展とともに、大規模なシーケンシングセンターでなくても多くの研究室やグループにおける配列決定が広がっていくと見込まれる。したがって、大規模シーケンシングセンターにおいても専門知識と時間を要していたアノテーションの効率化を進めない限り、塩基の文字列がディスクに蓄積されていくに過ぎない。そこで本課題において、効率の良い微生物ゲノム配列アノテーションを実現する情報環境を目指した。

具体的には、微生物ゲノム解析において定評あるデータベースと定評あるアルゴリズムを組み合わせたパイプライン Microbial Genome Annotation Pipeline（以下、MiGAP）を設計した。MiGAP によって、誰でも、何時でも、どこからでも、微生物ゲノム配列の生物学的意味を入手できるようになる。

MiGAP には、デフォルト解析の初級用のブロンズ (b-MiGAP)、パラメータを設定可能とする中級用シルバー (s-MiGAP)、さらにデータベースやアルゴリズムを付加することができる上級用のゴールド (g-MiGAP) の3つのレベルを設定することにした。今年度はこのうち、b-MiGAP の開発を進め、3月にはライフサイエンス統合データベースセンター内で内部公開した。

b-MiGAP では、MetaGeneAnnotator(MGA) (Noguchi H, Taniguchi T and Itoh T (2008)) によって ORF を予測し、RNA 領域を tRNAscan と高品質の独自 rRNA データベースを利用して同定し、Rfam で IS 領域を同定し、RefSeq と NR で既存アノテーションを参照し、COG、KOG、ならびに eggNOG で機能の手掛かりを得る。

b-MiGAP では、図 3.23 に示すインタフェイスで配列をコピー&ペーストあるいはローカルなコンピュータからアップロードすれば、自動的に解析が進む。

図 3.23 解析対象の配列を投入するユーザインタフェイス

図 3.23 で配列を投入後、解析の進行状況を図 3.24 のように確認することができ、解析が終了後は図 3.24～図 3.25 のように解析結果を確認できる。また、解析結果をダウンロードすることもできる。

Process Status
Start:2009/03/01 14:47:45

Process	Command or DB	Status
Read Parameter File		Done
Read Sequence File		Done
Search ORF Area	Metagene	Done
Search RNA Area	tRNAscan	Done[8/8]
	16sRNA	Done[8/8]
Annotation		Running[1/8][1/2]
Output Result File	Genbank	Before Running
	EMBL	Before Running
	DDBJ	Before Running
	Information	Before Running

図 3.24 投入した配列の解析の進み方の確認例

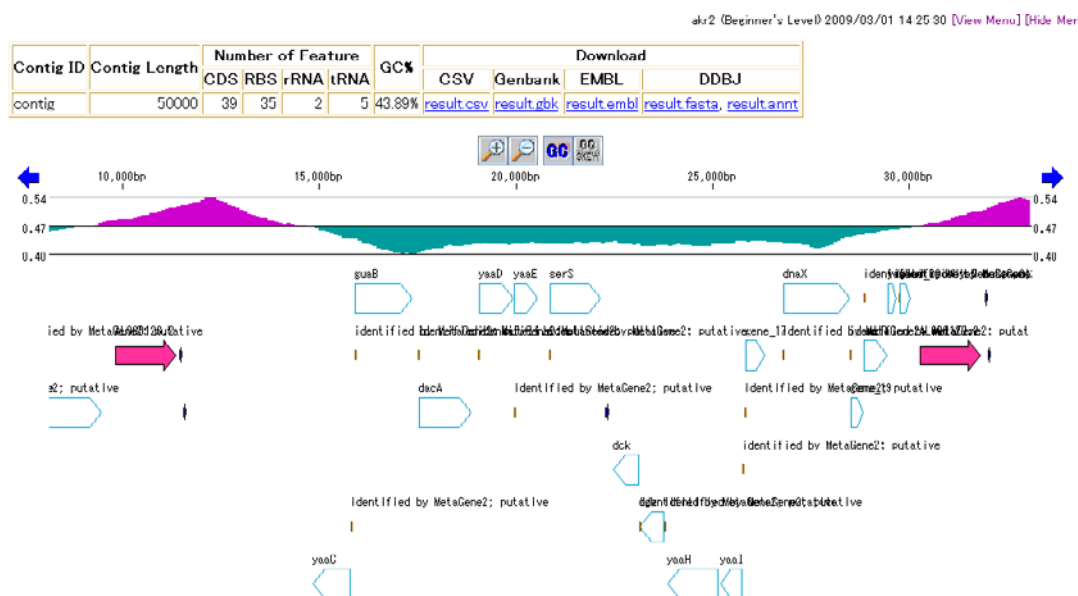


図 3.25 アノテーションの閲覧画面（リニア表示）

また、MiGAP には管理者用のツールも用意しており、図 3.26 に、参照データベースの管理システムの主画面を紹介する。

平成 21 年度には MiGAP を一般に公開し、特に国内微生物ゲノム解析の統合基盤へと斉唱させていきたい。

Logout	Help	Contact Us
User List	DB Updating	
DB Update	Task Name:	
DB Update	Database:	
History	☐ COG ☐ KOG ☐ eggNOG ☐ NR ☐ RefSeq	
DB Version	Run Reset	

図 3.26 参照データベース更新のための入口

辞書、オントロジーについては、生物名称について正式または推奨されている学名と日本語との対応を記したネット上あるいは書籍の形態の情報源を調査し、それぞれに関するメタデータを作成管理した。メタデータは、情報源の名称、配布形式、作成者、編集者、著者、監修者、版、作成日、URL、ISBN、その他 URI、その他書誌情報、章、パス、収集方法、及び個別のフィールドの意味、並びに担当者コメント等である。併せて、昆虫大図鑑 3 分冊を電子化し生物種名日本語名を抽出するとともに、微生物学辞典から日

本語表記が存在する微生物名を全て抜き出し、単離年度単離者名などの付帯情報も抜き出した。また、10万種を超える生物に関するDNAや遺伝子発現に関する情報へ効果的に案内するためには生物種名称の直感的な把握を助ける仕組みが必要であると考え、DNA配列の多い代表的な生物種に関して、直感的な把握が可能なアイコン画像を200種作成し、閲覧取得可能にし、また各画像が一意的URLをもち外部プログラムで表示利用可能なサーバを構築した。アイコン画像は各生物種について、大・小、背景あり・なしの4種類を作成し、系統樹に基づく階層フォルダを通じて閲覧可能とし、また、生物の学名、英名、日本語名、NCBI Taxonomy IDで検索可能とした。生物アイコンの利用方法としては、閲覧インタフェースに表示された生物アイコンをドラッグアンドドロップで他のソフトウェアにコピーする機能、アイコンURLによりネット上のイメージとして利用する機能、全アイコンデータを一括ダウンロードする機能を用意した(図3.27)。



図 3.27 生物アイコンデータベースシステムの概要

(2) 植物及び植物関連微生物のゲノム情報データベース統合と高度化 (かずさ DNA 研究所)

平成 19 年度に構築した、ユーザからの情報集積とそれを基にした整理を可能とするこ

とを特徴とするゲノムデータベース環境を活用し改善を加え、データの集積によるデータベース統合化を複数の生物種について進めた。

1) 高度情報集積データベースの運用と改良

平成 19 年度に構築した、分子データ上のポジションを統合基盤情報とし、アノテーション・キュレーションの統合と高度化を可能とするデータベース KazusaAnnotation (<http://a.kazusa.or.jp/> 図 3.29 参照) を、植物および植物関連微生物のゲノム解析情報を対象として運用しつつ、ユーザの要望を汲みあげて改善を加え、利便性の向上を行った。具体的には、データ登録時に、必要な情報の種類を提示するタグ候補の自動表示により入力タイプ量を減少させると同時にタイプミス回避する機能や、種々の条件指定を可能にし、絞り込みを容易にした検索機能、文献と遺伝子の関連を一覧できる表示機能を追加するなど、操作性の向上を図る改善を盛り込んだ。

また、ゲノム上の遺伝子に関わるアノテーション情報を視覚的に表示し、直感的に扱うことを目標に、優れたインタフェースにより閲覧可能なシステム genoDive について、データ取得ならびに解釈にかかわる基盤ライブラリの共通化をすすめ、ソフトウェアの再利用性を向上させた。

植物ゲノムのデータベース中の遺伝子名や遺伝子 ID のトレースが困難であることが実験生物学者にとりゲノム解析情報を利用する際の大きな障壁であることから、データベースによって異なる ID がついている関係性を検索可能にした遺伝子 ID 相互変換サービス (<http://charles.kazusa.or.jp/gns/>) のデータを更新し、ユーザの利用に供している。

植物のゲノム情報の受入と統合化を進めるために、エントリに付与した種々の属性により高度なデータの絞り込みを可能とするデータマネジメントシステムである BioMart (<http://mart.kazusa.or.jp/> 図 3.28 参照) を導入した。フィージビリティ・スタディとして、このシステムに酸素発生型光合成細菌のゲノム情報を蓄積し供用を開始した。今後、理化学研究所の植物オーミクス実験情報やそのほかの植物データの受け入れと統合的な利用のための枠組みを準備した。

KazusaMart beta

Data management system

Dataset 14 / 3725 Entries
Synechocystis sp. PCC 6803

Filters
GO ID : [ID-list specified]

Attributes
Kazusa Gene ID
Kazusa Peptide ID
Gene Symbol
Description

Export all results to: File TSV Unique results only Go

Email notification to:

View: 20 rows as HTML Unique results only

Kazusa Gene ID	Kazusa Peptide ID	Gene Symbol	Description
slr1404	slr1404		biopolymer transport ExbB protein homolog
slr1017	slr1017	am2	ammonium/methylammonium permease
slr0677	slr0677		biopolymer transport ExbB like protein
slr1314	slr1314		putative C4-dicarboxylase binding protein, periplasmic protein
slr1962	slr1962		probable extracellular solute-binding protein
slr0679	slr0679		periplasmic phosphate-binding protein of ABC transporter
slr0182	slr0182		ABC transporter ATP-binding protein
slr0477	slr0477		putative biopolymer transport ExbB-like protein
slr0108	slr0108	am1	ammonium/methylammonium permease
slr0700	slr0700		probable amino acid permease
slr0537	slr0537	am13	ammonium/methylammonium permease
slr1453	slr1453	cysT	sulfate transport system permease protein
slr1454	slr1454	cysW	sulfate transport system permease protein
slr1455	slr1455	cysA	sulfate transport system ATP-binding protein

biomart version 0.7

図3.28 データマネジメントシステムBioMart

2) ゲノムアノテーション情報の蓄積と高度化

上記の高度情報集積データベース環境 KazusaAnnotation を運用しつつ、その環境を利用したアノテーション情報の蓄積と高度化を平成 19 年度に引き続き進めた。平成 20 年度は、平成 19 年度と同様に論文中に出現する遺伝子・タンパク質記載のインデクス情報を合計約 10 万蓄積した。具体的な生物としては、光合成細菌であるシアノバクテリア *Synechocystis* sp. PCC 6803、*Anabaena* sp. PCC7120、*Synechococcus elongatus* PCC7942 に関する合計約 2100 報の論文について、また、植物と共生し窒素固定を行う根粒菌、*Mesorhizobium loti* MAFF303099、*Sinorhizobium meliloti* strain 1021 に関する合計約 300 報の論文について遺伝子の位置情報の登録を進めた。さらに高等植物であるマメ科のモデル研究植物であるミヤコグサに関する論文についても同様に 300 報の論文でのインデクスを完了し、ゲノムの位置情報と論文記載情報の統合を図った。ゲノムデータベースの変更やデータ更新に伴う関連付けアドレスの変更に対応するなど、蓄積情報の陳腐化を防ぐための更新をすすめた。登録した情報は上記のデータベースから即時公開し、広く生物学者の利用に供している。また、新しいタイプのゲノム情報データベースであるので、その概念と活用方法を一般の生物学者に周知するため、植物研究関係のシンポジウムや講習会で紹介を行い、利用を促した。その結果、月間セッション数、ページビュー数共に昨年当初よりも約 50%、ユニークユーザ数は約 120%増加した（セッション数:1593/1067 (49.3%増), ページビュー数:21742/15303 (48.9%増), ユニークユーザ数:705/318 (121.7%

増)、数値はすべて 2009 年 3 月/2008 年 5 月比較)。

kazusaAnnotation **alpha!**

quest アノテーション 追加 検索 サインアップする サインイン ヘルプ

かずさアノテーションはソーシャル・ゲノム・アノテーションツール。

保存
遺伝子アノテーションの便利ツール。遺伝子アノテーション、タグ、論文

共有
遺伝子アノテーションでつながるコミュニティ。アノテーションのコメントやタグを公開できます。

検索
id:synobu cph1 検索
もっと検索する...

最近のアノテーション

cyanobase/Chlorobium - Chlorobium:CT0609 43 annotations
http://genome.kazusa.or.jp/cyanobase/Chlorobium/genes/CT0609
hkane | pmid:18542834 | discussion ★

cyanobase/Chlorobium - Chlorobium:CT0609 43 annotations
http://genome.kazusa.or.jp/cyanobase/Chlorobium/genes/CT0609
hkane | pmid:18542834 | fg5 ★

cyanobase/Chlorobium - Chlorobium:CT0609 43 annotations
http://genome.kazusa.or.jp/cyanobase/Chlorobium/genes/CT0609
hkane | pmid:18542834 | fg4 ★

Read more ...

最近のページ

cyanobase/Chlorobium - Chlorobium:CT0609 genome.kazusa.or.jp
43 annotations 新しいアノテーションを追加する
タグ materials and methods 2 fg1 3 pmid:16308317 10 fg5 3 table1 3 results 4 fg2 4 fg6 1 fg3 4 pmid:10333495 12 discussion 3 title 4
pmid:16510994 9 experimental procedures 1 table2 1 pmid:18542834 12 fg4 3 abstract 4 table3 1 introduction 4

cyanobase/Chlorobium - Chlorobium:CT1666 genome.kazusa.or.jp
13 annotations 新しいアノテーションを追加する
タグ acknowledgement 1 fg5 1 table1 1 results 1 fg2 1 fg6 1 fg7 1 discussion 1 title 1 experimental procedures 1 pmid:18171624 13
fg4 1 abstract 1 introduction 1

最近のタグ

Read more ...

65897 エントリ 216650 アノテーション 195779 タグ

図3.29 高度情報集積データベース環境KazusaAnnotation改善版

4. 統合データベース支援（情報・システム研究機構）

4. 1 ポータル整備・運用、広報、普及啓発

4. 1. 1 ポータル整備・運用、広報、普及啓発の実施計画

平成19年10月に公開したポータルサイト「統合ホームページ」(<http://lifesciencedb.jp/>)の整備運用を継続するとともに、平成19年度の参画機関の成果を集約し統合DBセンターを通じて各サービスを利用できるようにウェブサイトを再構築する。さらに、DB受入にともなうサイトの整備を進める。また平成20年度新規に日本語総説や学会要旨、報告書等、生命科学に関する文献や関連情報を収集し検索を主にしたサービスを公開する。DB横断検索は引き続きクロール対象DBを増やすとともにインタフェースの改良を行う。統合TVについては音声サービス等利用者の利便性をよくする手法並びに、利用者からの登録や意見集約について検討する。データベースやツールに関する講習会を開催する。人材育成担当の参画機関によって作成された教材をとりまとめ公開する。

4. 1. 2 ポータル整備・運用、広報、普及啓発の実施内容

(1) 統合DBプロジェクトポータルサイトの整備・運用

平成19年10月に公開したポータルサイト「統合ホームページ」(<http://lifesciencedb.jp/>)の整備運用について、まず平成20年6月に平成19年度の中核機関、参画機関、補完課題の32の成果を統合HP上に集約し公開した。中核機関として、複数のDBを一括して検索可能な横断検索に加え、蛋白質核酸酵素全文検索、特定領域「ゲノム」報告書検索、学会要旨検索の日本語文献検索サービスを開始した。平成21年3月時点で、新規サービスを含め「統合ホームページ」のトップページからリンクする形で46のサービスを公開した。図4.1にサイトのトップページを示す。デザイン上の大きな変更点として、横断検索サービスにトップページからアクセスが容易になるように、上部の中心に検索窓を配置した。また現在全てのメニューをトップページから一度でたどりつけるようリンクを配置しているが、メニュー数の増加に合わせてカテゴリを整理し、カタログなどのサービスをポータルに、検索機能を中心とするサービスを検索に、参画分担機関によって構築されている各分野の統合データベースをデータベースに、またデータベース受入サービスをアーカイブとして再配置した。



図 4.1 トップページの変遷と変更ポイント

横断検索は統合の第一段階として国内 DB を中心に、各 DB はそれぞれのサイトに置いたまま、キーワード検索を可能にする Google ライクな検索サービスである。平成 19 年度 13 件のデータベースを検索対象にスタートしたが、平成 20 年度は 200 件の DB に対してクロール、インデックス作成を終了しインタフェースの改良を行った。新しいインタフェース(図 4.2)では左側のパネルにカテゴリをツリー表示し、利用者が目的分野の検索結果にたどりつきやすいように改良した。カテゴリの下に各データベースの概要とリンクを表示するパネルを設置した。中央のパネルに検索結果が表示され、カテゴリーツリーをクリックすることによって該当するカテゴリの検索結果が新規タブとして表示されるようにし、利用者が前の絞り込み結果に戻れるように配慮した。補完課題の糖鎖科学統合 DB とは共通の検索エンジンで開発を進め、各機関でインデックスの作成と更新を行いインデックスの共有による検索連携を行った。経済産業省統合データベースプロジェクトとの間でも共通の検索エンジンで開発を行えるよう技術情報を交換した。



図 4.2 横断検索サービスのインタフェース概要

統合 HP 上のデータベースカタログ、学協会カタログ、生物アイコンに関しては、データの更新に加えて、横断検索と共通のインタフェースを採用し使用感の統一を図った。併せて、英語化可能なサービスの英語版を作製し英語ポータルサイトを公開した（図 4.3）。



図 4.3 英語サイトの開設 (<http://lifesciencedb.jp/en/>)

H19 年度まで、ポータルサービスに必要な日本語情報資源として蛋白質核酸酵素のバックナンバーの電子化を行ってきたが、H20 年 6 月より蛋白質核酸酵素(PNE)全文検索サービスを開始した。図 4.4 に示す単独の検索サービスの他、横断検索にも検索結果を含めることにより、利用者が分子の情報を取得すると平行して、総説を参照し理解を深められるようにした。H20 年度 PNE の検索範囲は 1985 年から 2005 年までの 1577 ファイルであり、バックナンバーの入手が不可能な部分については電子化作業を追加で行った。PNE の他に、日本語文書の検索対象として文部科学省特定領域研究「ゲノム」報告書並びに、特許公報の検索サービスも開始した。

データベースやツールの利用方法を動画で解説するサービス統合 TV は今年度 100 以上の動画を作成し 150 件以上計 155 件の作成を達成、既存動画の更新も進めている。特に統合プロジェクト内のサービスについてのヘルプとして主要サービスの利用方法の動画を作成するとともに、広報普及活動の一環として講演会などを音声とスライドの動画として公開した。

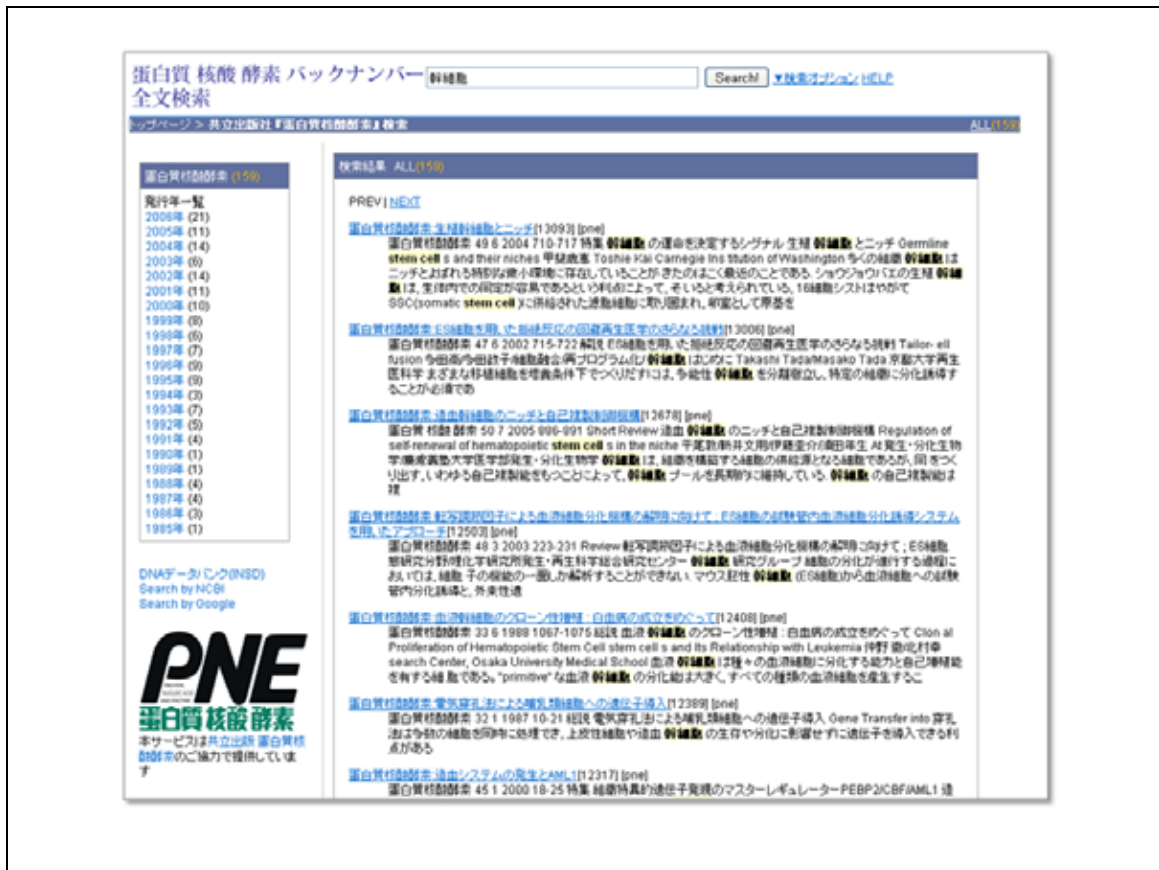


図 4.4 蛋白質核酸酵素全文検索サービス

サービス運用にあたって、新サービス提供サーバにポータルサイト部分の構築作業を行うとともに、200DB を対象とした横断検索サービス開始に備えて増設した計算機に対し、サービスの設計と構築を行った。年間を通じ停電などによる計算機停止の他は安定したサービス運用を実現できた。

(2) 広報、普及啓発（情報・システム研究機構）

統合DBプロジェクトの普及のために、実際にインターネットにつながった端末を前にしてハンズオンで初心者向けに統合データベース講習会を計6回（平成20年5月30日於日本大学生物資源科学部、平成20年7月3日於東京大学本郷キャンパス浅野地区(DBCLS)、平成20年8月29日於北海道大学、平成20年10月18日於九州大学箱崎キャンパス、平成20年10月31日於長浜バイオ大学、平成21年1月23日於東京工業大学すずかけ台キャンパス）、中級者以上向けにNucleic Acids Research誌DB特集号およびウェブサーバ特集号を俯瞰する統合データベース勉強会を計2回（平成20年11月21日と平成21年2月13日於東京大学本郷キャンパス浅野地区(DBCLS)）を開催した（図4.5）。

統合データベース講習会／勉強会で使用したテキストやその様子を録画した動画は、

教育/人材育成の各参画機関において作成された教育プログラムを集約して昨年度より公開してきたウェブサイトMotDBから利用可能となるよう維持管理し、常に最新の情報が提供されるようにした。



図4.5 講習会の開催

開発したデータベースやサービスを学会年会においてポスター発表することのみならず、シンポジウムやワークショップの企画（日本進化学会年会〔8月〕、日本生物物理学会年会〔12月〕、日本分子生物学会・日本生化学会合同年会〔12月〕）、ブースの出展（BioJapan〔10月〕、サイエンスアゴラ〔11月〕、日本分子生物学会・日本生化学会合同年会〔12月〕、日本農芸化学会〔3月〕）やランチョンセミナーの開催（日本農芸化学会〔3月〕）を実施し、その際に統合DBプロジェクトのサービスを紹介した印刷物（総合パンフレット1種〔合計3800部配布〕、個別サービス紹介のためのリーフレット27種）を配布するとともに、製薬企業を中心としたメーリングリスト及び、「ゲノム」特定領域研究の班員のメーリングリストに対し、プロジェクトの案内や人材育成プログラムへの参加者募集を行った。情報発信のために、ウェブサイトの本格公開にあわせて文部科学省を通じたプレスリリースを1回、プロジェクトウェブサイトおよびJSTの「文科省 統合データベース整備事業サイト」を通じたニュース配信を62回行い、そのうちの一部（13

回)については日経のBiotechnology Japanのサイトを利用したPRも実施した。利用者意見を集約するために、平成19年度以降の成果について外部評価アンケートを実施し、評価やフリーコメントの内容を対応状況とともに公開した(図4.6)。また、講習会や展示会等でアンケートを実施し意見を収集し開発の参考とするとともに、次年度の新規ユーザー評価者を獲得した。

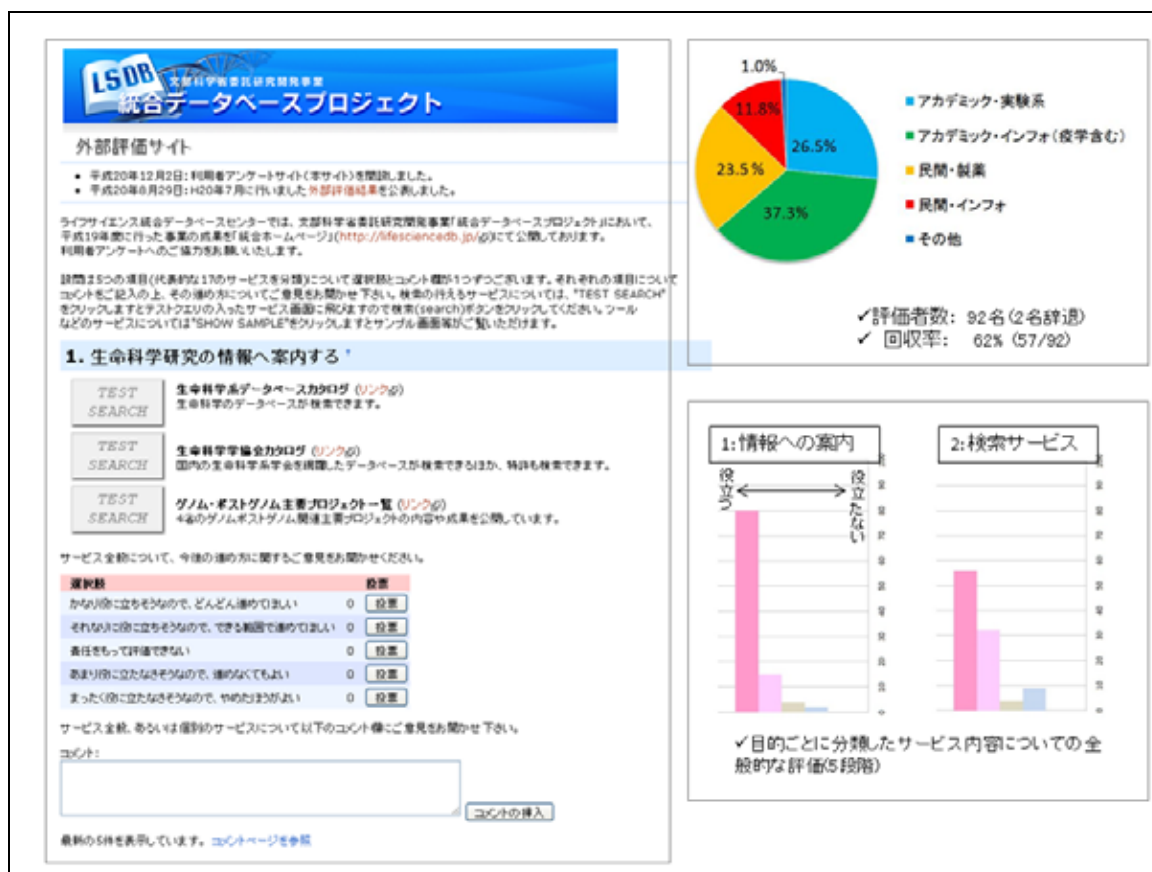


図4.6 外部評価アンケートの実施と結果公開
(<http://lifesciencedb.jp/feedback/hyouka19/?result>)

4. 2 データベース受入

4. 2. 1 データベース受入の実施計画

我が国のプロジェクトデータベースについて、新規なデータベースを含めて現状調査を実施する。データベースのタイプに応じた方法で、データベース受け入れを実施する。受け入れたデータベースの運用システムを、ポータルサイトに構築し、運用を行う。

4. 2. 2 データベース受入の実施内容

1) ライフサイエンスデータベース、及びプロジェクトの調査

主要プロジェクト一覧（全 35 プロジェクト）の英語化を実施し、各項目による検索やソートが可能なインタフェースを平成 19 年度構築した TogoDB を用いて作成し公開した（図 4.7、http://togodb.dbcls.jp/lbdb_project_en）。また、文部科学省のプロジェクトについてのデータ内容の追加(10 件)を行った。さらに、受け入れ対象となるデータベースの把握のため、Nucleic Acids Research 誌（NAR）に収録された国内のライフサイエンスデータベース(77 件)について、利用許諾やダウンロードの可否等の調査を行った。その結果、データダウンロード可能なデータベースは約 6 割、何らかの利用許諾条件が明記されているものは約 3 割であることがわかった。NAR に収録されていないデータベースでは、ダウンロード可能なものの割合はさらに減少すると考えられ、データの再利用がしにくい状況であることがわかった。



Project Name	Implementing Institution	Start Year	End Year	Availability of Research Data	Public Database	Data Download Site	Website
Protein 1000	Ministry of Education, Culture, Sports, Science and Technology (MEXT)	2002	2006	Publicly released	NA	None	-
Genome Network Project	Ministry of Education, Culture, Sports, Science and Technology (MEXT)	2004	2008	Partially shared	-	-	Genome Network Project (http://www-db.jp/genome/english/index.html)
Research for the Future Program (Genome Research)	Ministry of Education, Culture, Sports, Science and Technology (MEXT)	1996	2004	Partially shared, Partially publicly released, Partially unknown	-	-	Research for the Future Program (http://www.rffp.go.jp/~rffp/project/2004.htm)
JST Sequencing Project	Ministry of Education, Culture, Sports, Science and Technology (MEXT)	1995	2001	Shared	http://lifescience.db.jp/ldb.cgi?page=1&emp_tp_dbid=318	Sequencing data download site (http://www-als.tokyo.jst.go.jp/455/top_ja.pl)	JST ALIS Project website (http://www-als.tokyo.jst.go.jp/455/top_ja.pl , Japanese)
SNP Research	Ministry of Education, Culture, Sports, Science and Technology (MEXT)	2000	2003	Unpublished	http://lifescience.db.jp/?page=1&emp_tp_dbid=295	- SNP frequency data, etc. (http://db.hgc.jp/pub/hgc/ldb/emp/) - JPNL mapping data, etc.) (http://exp.mrc-uk.ac.uk/2004/ja.html) - Research tool (http://exp.mrc-uk.ac.uk/2004/emp/)	-
The Hokkaido Japan Project on the implementation of personalized medicine	Ministry of Education, Culture, Sports, Science and Technology (MEXT)	2003	2012	-	-	-	Webpage of Biobank Japan Project for Personalized Medicine (http://biobank.jp/eng/)
RIKEN SNP	Ministry of Education, Culture, Sports, Science and Technology (MEXT)	-	-	-	http://lifescience.db.jp/?page=1&emp_tp_dbid=295	-	RIKEN Center for Genomic Medicine website (http://www.riken.ac.jp/english/index.html)
ARTAGEdb	-	-	-	-	http://lifescience.db.jp/?page=1&emp_tp_dbid=260	-	-

図4.7 英語で記述したゲノム・ポストゲノムプロジェクト一覧の一部

受け入れデータベースの運用システムの構築については、アーカイブデータベース一覧システムを構築した（図 4.8）。

図 4.8 データベース一覧の表示画面

[illegible]

図 4.10 データのメタデータの例

アーカイブデータベース一覧システムでは、各データベースに対して作成した、データベースのメタデータページ（図 4.9）、ダウンロードページよりデータのメタデータページ（図 4.10）、利用許諾のページ（図 4.11）、TogoDB による検索ダウンロードページ（図 4.12）へとリンクが張られている。



図 4.11 利用許諾のページ

[illegible]

図 4.12 TogoDB による検索ダウンロードページ

また、アーカイブデータベース一覧を主要なサービスとする「生命科学系データベースアーカイブサービスサイト」（図 4.13）を立ち上げ、サービスを公開した。「生命科学系データベースアーカイブサービスサイト」では、データベースの寄託を公募するためのページを作成し、その中で「寄託データベース利用許諾作成マニュアル」、「データベース利用許諾用語定義解説」、「データベース公開に関する覚書書式」の作成を行い公開した。

文部科学省委託研究開発事業 統合データベースプロジェクト LSDB@ - アーカイブトップ@ - ヘルプ@

生命科学系データベース アーカイブサービス

- ホーム
- DBCLS標準利用許諾
- サービスの意義と利点
- データベース一覧
- 寄託応募要領

- 更新履歴
- お問い合わせ

生命科学系データベースアーカイブサービス

いくら良質なデータベースでも、説明が十分でない、利用条件が明確でない、ダウンロードできないなどの理由で十分に利用され引用される機会を逃していることがあります。

DBCLSでは、国内の生命科学プロジェクトデータの共有化を推進しています。

データを、利用条件や高一次的な説明とともにダウンロード可能な形で一箇所に集積することで、データへのアクセスの増大とデータマイニング等の再利用による生命科学分野への貢献を期待しています。

ライフサイエンス統合データベースセンター(DBCLS)では、データ共有化の実現を目的として、生命科学系データベースアーカイブサービスを実施します。DBCLSでは、データベースをアーカイブし、さらにダウンロード可能にすることによって、誰もがそのデータを同じように利用できる「データ共有」が実現できると考えています。それにより手間とお金を掛けて得られた貴重なデータが、データマイニングや統合化など、より深く利用されるようになると考えています。さらに、アーカイブ化により、データベース関連プロジェクトの終了後も長期にわたるデータの保全が可能になります。

具体的には、各研究機関から寄託して頂いたデータベースをアーカイブ化し、各データベースの詳細情報(メタデータ)と利用許諾条件(DBCLSでは、[標準利用許諾](#)を用意しています)と共に公開し、ダウンロード可能にする事業を行います。詳細なメタデータの提供と利用許諾条件の開示により、適切にかつ安心してデータをご利用いただけます([リーディスの意義と利点](#))。本サービス向けのデータベース寄託も公募しています。

データベース一覧

現在は、DBCLS、及び「統合データベースプロジェクト」の関連機関において産生されたデータベースを中心に公開しています。また、文部科学省所管のライフサイエンス関連のプロジェクトから今後産生されるデータベースは公開が義務付けられており、それらのデータベースは本サービスで公開される予定です。
※(今後データベースを公開予定のプロジェクトの一覧)(準備中)

データベース寄託公募

本事業で公開するデータベースを寄託していただくための公募を、国内のあらゆる研究機関を対象に行います。データベースを保有する各研究機関におかれましては、寄託の意義と利点(クレジット表示やアクセス統計情報の取得等、[サービスの意義と利点](#))をご理解いただき、ぜひデータベースをDBCLSに寄託していただきますようお願いいたします。尚、具体的な寄託方法につきましては、以下の応募要領を参照願います。

■ [データベース寄託の応募要領](#)

データマイニング
統合化への利用

安心して、
適切に利用可能

ユーザ

データのダウンロード

簡易検索機能

利用許諾条件、メタデータ

データベースの
メンテナンスを
サポート

クレジット表示を
保証

データベース作成者

データベースの寄託

生命科学系データベースアーカイブサービス

データベースアーカイブサービスの意義、利点

図 4.13 生命科学系データベースアーカイブサービスサイト

データベース受け入れについては、補完課題の各機関、及びその他の機関との間で、3つの観点（メタデータ、横断検索、データベースアーカイブサービス）で連携を行うという方針のもとに、それぞれ連携作業を実施し、総計 12 件のデータベースについてメタデータとデータを受け取った。機関名と件数は以下の通り。産総研糖鎖医工学センター（1件）、九工大（2件）、理研（4件）、JST（1件）、かずさDNA研（4件）。これらのメタデータとデータの一部を、アーカイブデータベース一覧システムのテストサイト内に登録(図4.14)し、公開の準備を進めた。

図 4.14 アーカイブデータベース一覧システムのテストサイト

また、理研と九工大のタンパク関連情報の整理の一環として、蛋白質に関する国内外の DB 統合化のための内容検討を行った。蛋白質に関する国内外の DB 情報を収集し、その中で約 40 件のインデックス化が可能なデータベースについて、横断検索システムのプロトタイプを構築した。また、蛋白 ID による統合を目的として、上記の蛋白データベースから ID 情報が抽出できるかどうかの調査を行った。さらに、植物オミックス情報の連携のため理研とかずさ DNA 研究所の連携にも取り組んだ（具体的内容は 3. 3. 2 項に記載）。

その他、経産省統合 DB プロジェクトとの間では、メタデータや横断検索、利用許諾、ウェブサービスなどでの共通化をはかることで合意するなどの連携を図った。メタデータについては、27 件のデータベースについて、双方のデータベースカタログの記述を統一化した。具体的なデータベースの名称は以下の通りである。

1. Brain Atlas Database
2. ARCHAIC
3. Arch GeNet
4. TraP
5. Rat Brain Sections
6. Database of genomes and transcriptional regulations for filamentous fungi (FFDB)
7. fRNA Database
8. ASTRA
9. ConfC
10. TMBETA-GENOME
11. SEVENS
12. GENIUS II
13. fRNAdb
14. EzCatDB
15. DB-SPIRE
16. 機能性 RNA ゲノムブラウザ
17. Evola
18. G-compass
19. H-ANGEL
20. H-DBAS
21. LEGENDA
22. PPI view
23. VarySysDB
24. FLJ Human cDNA Database
25. H-Invitational Database
26. 微生物データベースシステム
27. PDB-REPRDB

4. 3 人材育成（東京大学、お茶の水女子大学、長浜バイオ大学）

4. 3. 1 人材育成の実施計画

（東京大学）

DB 構築者を養成するために以下の3つの演習を実施する。

1) バイオ DB サーバ構築演習

データベースサーバのミラーサイトを構築する。OS, apache, MySQL 等の主要ソフトウェアのインストールおよびネットワークセキュリティに習熟することが目標である。参加者には各自にサーバ構築用ワークステーションを配布する。演習は1年間コースと1ヶ月コースを設け、前者はプログラミング初心者、後者はプログラミング経験者を対象とする。後者は1年間の継続的受講が困難な企業等の受講希望者にたいして1年間コースを要約した演習を提供する。各コースとも受け入れ人数は5名程度、全体で10名のDB 構築者を養成する。

2) プログラミング演習

Java および Perl プログラミングを演習した後に、アルゴリズムの知識を活かした配列処理やデータマイニングの実装を行う。上記①バイオ DB サーバ構築演習では実施がむずかしいプログラミング演習を行うことで、独自にソフトウェア構築ができる能力を身につけることをめざす。演習総時間は90時間で約2ヶ月間を予定している。

3) 独創的サーバ構築演習

バイオDBサーバ構築演習およびプログラミング演習を修了した受講者に対して開講する。大規模計算のためのクラスター利用技術を習得させ、他に類の無いバイオDBサーバを設計、実装、公開することを目標とする。平成20年度の受け入れ人数は5名とする。

演習は英語で書かれたプログラミングの教科書および DB サーバシステムの管理マニュアルの読破も要求し、自ら立ち上げたサーバの英語マニュアルが書けるように指導する。したがって受講者にはある程度の英語能力を想定する。平成 19 年度でほぼ完成し公開中の演習ノートも改善する。

（お茶の水女子大学）

統合DBは日々更新される膨大なデータベースである。このデータベースから最新のデータを取得し、生命理解へと繋げるには、生命科学の知識のみならずDBから必要なデータを抽出、編集、蓄積し、解析や知識発見ができるDB高度利用者が必要である。平成20年度は、平成19年度に作成したDB高度利用者養成カリキュラム案を基に、授業、演習を実施することでDB高度利用者を養成し、カリキュラムの評価と改訂を進める。また、DB高度利用者養成のための自習用教材の作成とそのインターネット上での公開を行う。

(長浜バイオ大学)

1) アノテータ・キュレータの教育

学校法人関西文理総合学園長浜バイオ大学は、DB のアノテータ・キュレータの人材養成を目標の一つとして、実習にゲノムアノテーションを取り入れてきた。本人材育成では、これまでの実績を基礎に、情報・システム研究機構と連絡を密に取り、アノテーションが十分になされていない微生物ゲノム配列を入手し、実データに基づくゲノムアノテーション実習を行い、本統合 DB からの公開データとして提供する。公開を前提としたアノテーションを実践することは、技術の習得に加えて、データベース作成の意義や責任を体験することになり、ミスマッチの少ない人材養成となる。なお、これらの人材育成プログラムには、情報・システム研究機構でアノテータやキュレータとして雇用する人材の育成も含む。アノテータ養成のための自習用教材を作成し、その成果をインターネット上で公開を行う。

2) 自己組織化マップによる混合微生物遺伝子のアノテーション手法を用いた人材養成

バイオデータの特徴である複雑で大量な情報のデータベース化と、その効率的な活用においては、可視化技術の利用が重要となる。大量複雑情報の可視化技術として確立している自己組織化マップのような既存手法を用いて、混合微生物由来の大量遺伝子のアノテーション結果の可視化を行い、得られたアノテーションの成果と可視化データを本統合 DB からの公開データとして、利用しやすい形式で提供する。これらのバイオデータベースの作成と活用に必須な要素技術となる、可視化技術を備えた人材を、環境由来の大量ゲノムデータの解析を実例に養成する。また、その成果をインターネット上で公開を行う。

3) シニア研究者のチームによる分散型の高度なキュレーションモデルの確立

上記の2業務で得られたアノテーション成果の質の向上を目的に、専門知識を持つ複数のシニア世代研究者を中心とする、若手人材も含めたチームによるネットワークを介した分散型の高度なキュレーションモデルの確立をめざす。このキュレーションモデルを本事業のアノテーション作業に活用して、アノテーションデータの品質を高めると同時に、他大学や企業を含む研究機関の人材を対象に、短期滞在型を含む講習会を開催して、大学や研究機関へのキュレーションモデルの普及を図る。

4. 3. 2 人材育成の実施内容

(1) DB 構築者の養成 (東京大学)

公開成果の利用状況は Webalizer または google analytics を使ったアクセス解析により調査した。(Sites: いくつかの IP アドレスからリクエストを受けたかを示す。KBytes: 送信したデータ量。Visits: サイトを訪問したユーザ数。Pages: 送信したページ数。画像などのファイルは含まない。Files: 送信したファイル数。セッション: サイトの訪問数。ページビュー: 送信したページ数。)

1) バイオ DB サーバ構築演習

データベースサーバのミラーサイトを構築する演習を実施した。OS, apache, MySQL 等の主要ソフトウェアのインストールおよびネットワークセキュリティに習熟することを目標とし、参加者には各自にサーバ構築用ワークステーションを配布した。1 名は 1 カ月コースを受講し、残り 9 名は演習を完了するまでには計画通り 1 年間の時間を要した。平成 19 年度より作成している演習ノートの web ページ（下記）を充実させた。

<http://mlab.cb.k.u-tokyo.ac.jp/~mkasa/ensemblmirror/index.php?Introduction>

アクセス数

	Sites	KBytes	Visits	Pages	Files
1 月	130	47283	870	3193	4834
2 月	307	93976	2495	6854	10063
3 月	315	86737	2450	6069	8080

ユーザからの評価

知らない分野についても解説があり、新しい技術も紹介されていて勉強になった。毎週新しいことを勉強するが演習時間内では解決できず時間が足りない。

2) プログラミング演習

Java および Perl プログラミングを演習した後に、アルゴリズムの知識を活かした配列処理やデータマイニングの実装を行う演習を実施した。上記 1) バイオ DB サーバ構築演習では実施がむずかしいプログラミング演習を行うことで、独自にソフトウェア構築ができる能力を身につけさせた。演習総時間は 90 時間で、平成 20 年 7 月下旬～9 月上旬の約 2 ヶ月間で行った。演習ノートの web ページ（下記）を作成した。

<http://mlab.cb.k.u-tokyo.ac.jp/~mkasa/upbsb2008/>

ユーザからの評価

受講者アンケートの結果、テキストや補足資料などの内容については「十分理解できる」「なんとか理解できる」とする受講者が多数だった。また、「積極的に予習復習した」「ときどき予習復習した」という受講者が多く、ウェブで公開された演習資料が役立っている。

3) 独創的サーバ構築演習

バイオ DB サーバ構築演習およびプログラミング演習を修了した受講者に対して独創的サーバ構築演習を実施し、ゲノムブラウザで表示される新規トラックの開発・新規データの表示・塩基配列をゲノムにマッピングし結果をウェブ上に表示するプログラムの実装などを行った。

・塩基配列をメダカゲノムにマッピングし結果をウェブ上に表示するプログラムの開発。

<http://medaka.utgenome.org/>

- ・ cDNA アセンブルプログラムの開発。

<http://musica.gi.k.u-tokyo.ac.jp/>

- ・ ゲノム中の短い配列の出現頻度データベース RepeatScape（下記）を公開した。

<http://repeatscape.utgenome.org>

アクセス数

	セッション	ページビュー
1 月	34	61
2 月	56	89
3 月	41	90

ユーザからの評価

表示が工夫されていてわかりやすい。超高速 DNA シークエンサを用いたゲノム解析に有用である。

（２）DB 高度利用者の養成（お茶の水女子大学）

平成 20 年度の業務においては、平成 19 年度に開発した DB 高度利用者の養成に必要なカリキュラム案を基に、授業・演習を行った。平成 19 年度策定のカリキュラムが今後の生物学データの変化に対応できるか評価のため、世界の先端で活躍する進化学者が集まる国際会議 SMBE08 及びエピゲノムを医療応用する研究を行う世界有数の機関 IFOM-IEO に出張し研究者達とディスカッションを行った。その結果、データ整理の自動化、及び、分子進化に関する講義の量を 19 年度策定カリキュラムより増やす変更を行った。受講生の内、14 人の修士課程、博士課程の学生が所定のコースを終了した。平成 20 年度の受講生の内 2 名は本自習用教材改訂及び公開用サーバ管理において中心的な役割を果たし、データの抽出、編集、蓄積に関する人材として活躍している。他の 2 名は本コースで学んだ内容を基に、統合 TV の教材作成を行っている。一方で「分かりやすかったが、ついていくのが大変だった。」とのアンケート結果もあり、演習中も受講生に進度のばらつきがみられた。より多くの DB 高度利用者養成に向けて、各自の進度に応じて教育できるよう、態勢強化を行う必要がある。特に、編集したデータを知識発見に繋げる所についていくのが難しくなる例が見られたため、基礎統計の演習時間を増やす対策を講じる。

平成 20 年度も平成 19 年度に引き続き、授業・演習と平行して自習用教材の作成とインターネットでの公開を行った。カリキュラム、自習教材は、<http://togodb.sel.is.ocha.ac.jp> で公開中である。平成 20 年 7 月以降、3 ヶ月毎の Google analytics を利用して調査したアクセス数は下記の通りである。

	セッション	ページビュー
平成 20 年 7 月～9 月	62	241
平成 20 年 10 月～12 月	195	1580
平成 21 年 1 月～3 月	265	1260

(3) アノテータ・キュレータの教育（長浜バイオ大学）

1) アノテータ・キュレータの教育

本学3回生（260名）を対象とした初級アノテーション教育「環境由来メタゲノム配列からの健康に貢献する遺伝子発掘」を平成20年7月に完了、作成した教材を平成20年8月に公開を開始した。新規に発掘した3000件の遺伝子に関する成果については雇用した研究員が精査後にデータベース化し、平成21年3月に公開を開始した。作成した自習用テキストの使用法に関する統合TVコンテンツを含む4件の動画コンテンツを作成し、情報・システム研究機構統合DBセンターから公開した。

本学3回生（50名）を対象とした中級アノテーション教育は、新規微生物由来のドラフトゲノム配列を対象に、平成20年11月より実施した。5グループについて、相互に気付かせずに、同一ゲノムのアノテーションを行った。なお、この新規微生物ゲノムについては、公的機関が別途アノテーションを行っているが、学生が行ったアノテーションは、概略としてこれと同様な結果であった。学生のグループ間で、遺伝子位置や機能で差異の生じた場合について、5グループの結果を多数決的に採用すると、大半について公的機関のものと同一になった。教材を含む教育方法が適切であったと考えられる。

2) 自己組織化マップによる混合微生物遺伝子のアノテーション手法を用いた人材養成

国際DNAデータバンクにおいて遺伝子機能も生物系統もアノテーションがなされていない、メタゲノム解析で得られている微生物遺伝子類を対象として、本学4回生2名・シニア世代1名・雇用した研究員の共同作業として、有用遺伝子の探索とデータベース構築を行った。その成果は平成21年3月から長浜バイオ大学及び情報・システム研究機構統合DBセンターから公開しているデータベースへ追加した。これらの有用遺伝子を主対象に、自己組織化マップによる相同性によらない生物系統の推定を行い、約600件の遺伝子の生物系統を推定した。4回生2名の内の1名は本学の大学院へ進学して、インフルエンザウイルスAの既知の全ゲノム配列を対象にした自己組織化マップによる特徴抽出と、可視化技術を組みこんだデータベース化を行っている。他の1名は、他大学の医学系大学院へ進学し、医療情報処理の分野へ進んだ。シニア世代1名と雇用した研究員は、混合微生物ゲノムからの有用遺伝子探索とそのデータベース化を継続している。

3) シニア研究者のチームによる分散型の高度なキュレーションモデルの確立

3名のシニア研究者チームと学部4回生3名の共同作業により、約14万件のtRNA遺伝子を

データベースに収録し、『tRNADB-CE』として長浜バイオ大学及び情報・システム研究機構統合DBセンターから公開を行い、Nucl.Acids Res. のDatabase Issueにて論文発表を行った。2008年度更新用データとして原核生物416種、41421遺伝子についての精査を完了した。又、「tRNAデータベース『tRNADB-CE』の使い方」の統合TV動画コンテンツを開発し、情報・システム研究機構統合DBセンターからの公開を開始した。

5. プロジェクトの総合的推進（情報・システム研究機構）

5. 1 プロジェクトの総合的推進の実施計画

中核機関である情報・システム研究機構は、分担機関である京都大学、東京医科歯科大学、東京大学大学院医学系研究科に全体戦略を提示し連携して本事業を推進する。このため、分担機関を含めた研究運営委員会を開催する。また補完課題の担当機関である理化学研究所、産業技術総合研究所糖鎖医工学研究センター、情報・システム研究機構国立遺伝学研究所、九州工業大学とも全体戦略を共有し、連携して本事業を推進する。

さらに、中核機関グループにおけるプロジェクトの円滑な運営を推進するために、研究運営委員会の開催等を通して中核機関グループ内の連携・調整、及び進捗の確認を行う。必要に応じて外部有識者を招聘して意見を聞くなど、プロジェクトの推進に資する。

5. 2 プロジェクトの総合的推進の実施内容

（1）研究運営委員会及び研究運営委員会作業部会

研究運営委員会及び研究運営委員会作業部会を組織して、本プロジェクトの円滑な運営に努めた。なお、中間評価結果を踏まえて研究運営委員会と作業部会の在り方を見直し、研究運営委員会ではより広い視点からの議論を行い、具体的なプロジェクトの進捗に関わる議論は作業部会で行うこととした。このため、委員会の構成を見直し、平成 20 年 8 月開催の第五回研究運営委員会で承認を得たのち、以下に示す新たな体制で進めた。

第六回研究運営委員会以降の運営委員は下記の通りである（敬称略、順不同）。

大倉 克美	（独）科学技術振興機構研究基盤情報部長
勝木 元也	自然科学研究機構理事
金岡 昌治	（株）大日本住友製薬研究本部副本部長
金久 實	京都大学化学研究所バイオインフォマティクスセンター長
榊 佳之	豊橋技術科学大学長
豊田 哲郎	（独）理化学研究所横浜研究所生命情報基盤研究部門長
長洲 毅志	エーザイ（株）理事・研究開発担当付部長
中村 桂子	JT生命誌研究館館長
中村 春木	大阪大学蛋白質研究所蛋白質情報科学研究系教授
長村 吉晃	（独）農業生物資源研究所基盤研究領域ゲノムリソースセンター長
松尾 泰樹	（独）理化学研究所横浜研究所研究推進部長
松原 謙一(委員長)	（株）DNAチップ研究所代表取締役社長
湯元 昇	（独）産業技術総合研究所理事
吉田 輝彦	国立がんセンター研究所腫瘍ゲノム解析・情報研究部長
吉田 光昭	東京大学大学院新領域創成科学研究科客員教授
大久保公策	情報・システム研究機構国立遺伝学研究所

生命情報・DDBJ研究センター教授

五條堀 孝 情報・システム研究機構国立遺伝学研究所副所長
高木 利久 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長
堀田 凱樹(副委員長) 情報・システム研究機構長

なお、プロジェクト発足以来、第五回研究運営委員会まで運営委員を務めていただいた方々は以下の通りである（敬称略、順不同）。

秋山 泰 東京工業大学大学院情報理工学研究科計算工学専攻教授
浅井 潔 （独）産業技術総合研究所生命情報工学研究センター長
大倉 克美 （独）科学技術振興機構研究基盤情報部長
勝木 元也 自然科学研究機構理事
金岡 昌治 （株）大日本住友製薬研究本部副本部長
金久 實 京都大学化学研究所バイオインフォマティクスセンター長
久原 哲 九州大学大学院農学研究院遺伝子資源工学部門教授
榊 佳之 （独）理化学研究所横浜研究所ゲノム科学総合研究センター長
田中 博 東京医科歯科大学情報医科学センター長
田畑 哲之 （財）かずさ DNA 研究所副所長
徳永 勝士 東京大学大学院医学系研究科教授
中村 桂子 JT 生命誌研究館館長
中村 春木 大阪大学蛋白質研究所蛋白質情報科学研究系教授
長村 吉晃 （独）農業生物資源研究所基盤研究領域ゲノムリソースセンター長
松原 謙一（委員長） （株）DNA チップ研究所代表取締役社長
吉田 輝彦 国立がんセンター研究所腫瘍ゲノム解析・情報研究部長
吉田 光昭 東京大学大学院新領域創成科学研究科客員教授
大久保公策 情報・システム研究機構国立遺伝学研究所
生命情報・DDBJ 研究センター教授
小原 雄治 情報・システム研究機構国立遺伝学研究所長
五條堀 孝 情報・システム研究機構国立遺伝学研究所副所長
坂内 正夫 情報・システム研究機構国立情報学研究所長
菅原 秀明 情報・システム研究機構国立遺伝学研究所
生命情報・DDBJ 研究センター長
高木 利久 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長/
東京大学教授
堀田 凱樹（副委員長） 情報・システム研究機構長

また、新たな研究運営委員会作業部会のメンバーは下記の通りである（敬称略、順不同）。

浅井 潔	(独) 産業技術総合研究所生命情報工学研究センター長
黒田 雅子	(独) 科学技術振興機構研究基盤情報部バイオインフォマティクス課長
五斗 進	京都大学化学研究所バイオインフォマティクスセンター准教授
田中 博	東京医科歯科大学情報医科学センター長
田畑 哲之	(財) かずさ DNA 研究所副所長
徳永 勝士	東京大学大学院医学系研究科教授
豊田 哲郎	(独) 理化学研究所横浜研究所生命情報基盤研究部門長
成松 久	(独) 産業技術総合研究所糖鎖医工学研究センター長
松原 謙一	(株) DNA チップ研究所代表取締役社長
森下 真一	東京大学新領域創成科学研究科教授
五條堀 孝	情報・システム研究機構国立遺伝学研究所副所長
菅原 秀明	情報・システム研究機構国立遺伝学研究所 生命情報・DDBJ 研究センター特任教授
高木 利久(主査)	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長

上記の体制で、以下に示す研究運営委員会及び研究運営委員会作業部会、同分科会を開催し、プロジェクトの進捗状況、中間評価への対応、プロジェクト終了後を見据えた統合データベースのあるべき姿、体制等に関して討議した。

第四回研究運営委員会（平成20年4月3日）

第五回研究運営委員会（平成20年8月11日）

第五回研究運営委員会作業部会（平成20年10月10日）

第六回研究運営委員会（平成21年1月19日）

研究運営委員会作業部会分科会－医学・化合物・糖鎖関係（平成20年1月21日）

研究運営委員会作業部会分科会

－遺伝研・JST・中核機関の参画機関（基盤技術関係機関）（平成20年1月27日）

研究運営委員会作業部会分科会－立体構造・植物関係機関（平成20年1月29日）

研究運営委員会作業部会分科会－人材育成（平成20年1月30日）

（２）プロジェクト参加機関および関連機関との連携

中核機関とその参画機関、及び分担機関、補完課題実施機関等のプロジェクト参加機関とは、適宜打合せを行うことにより、全体戦略を共有して、連携して本事業を推進するべく努めた。データベースの受け入れという観点で中核機関と連携の可能性のあるプロジェクト外の機関とも、必要に応じて打合せを行った。会合の一覧は以下に示すとおりである。

また、プロジェクトの参加機関が、科学技術振興機構(JST)も含めて、全体で 16 機関と

多岐にわたるため、プロジェクト参加者全員が、全体としての目標を理解して、進捗状況を把握できるようにするために、参加機関ごとの平成 20 年度の具体的目標と進捗状況を示す全体テーブルを作成し、これを定期的にリバイズすることにより、プロジェクト全体での情報共有に努めた。併せて、JST が構築した「参画研究者の情報交換サイト」の活用による情報共有の促進にも努めた。

日付	時間	内容	場所
H20.04.25	13:00-14:40	九工大ミーティング第 1 回	DBCLS
H20.04.30	10:00-12:00	経産省統合 DB とのミーティング	DBCLS
H20.05.09	16:30-17:30	東大グループ DB 移植	DBCLS
H20.05.19	15:00-16:00	SVDB 等移植	DBCLS
H20.05.20	17:00-18:30	経産省統合 DB とのミーティング	DBCLS
H20.05.28	13:30-15:10	九工大ミーティング第 2 回	DBCLS
H20.06.17	11:00-15:30	産総研糖鎖センターとのミーティング第 2 回	糖鎖センター
H20.06.27	15:00-16:00	九工大ミーティング第 3 回	DBCLS
H20.07.02	10:00-12:00	JST との打ち合わせ	DBCLS
H20.07.29	10:00-13:00	微生物グループ打合せ	DBCLS
H20.07.31	14:00-17:00	かずさ DNA 研(植物統合関連)打合せ	DBCLS
H20.08.01	13:00-15:00	理研とのミーティング	理研
H20.08.26	16:00-19:00	経産省統合 DB とのミーティング	DBCLS
H20.08.29	14:00-16:00	東大・磯貝先生とのミーティング(SVDB 関係)	東大薬学部
H20.09.16	15:00-18:30	FAMSBASE 受け入れヒアリング	北里大
H20.09.26	14:00-16:00	DDBJ とのミーティング(トレースアーカイブ関係)	遺伝研
H20.10.22	13:30-16:30	理研とのミーティング	理研
H20.10.30	14:00-18:00	HGPD ミラー化のミーティング	産総研 BIRC
H20.10.31	17:00-18:00	東大グループとの打合せ	東大医学部
H20.11.07	10:30-13:00	トキシコゲノミクスプロジェクトとのミーティング	医薬基盤研
H20.11.07	19:00-20:00	東大グループとの打合せ	DBCLS
H20.11.14	15:00-17:00	人材育成関係打合せ(お茶大、長浜バイオ大)	DBCLS
H20.11.19	15:30-18:30	経産省統合 DB とのミーティング	DBCLS
H20.11.20	13:00-15:00	医科歯科大グループとの打合せ	DBCLS
H20.11.28	14:00-16:30	HGPD ミラー化のミーティング	産総研 BIRC
H20.12.25	13:30-15:30	かずさ DNA 研(植物統合関連)打合せ	かずさ DNA 研
H21.01.08	14:30-17:00	DDBJ とのミーティング(トレースアーカイブ関係)	遺伝研
H21.01.16	14:00-17:00	糖鎖センターとのミーティング	DBCLS

H21.03.24	13:00-14:00	かずさ DNA 研(植物統合関連)打合せ	DBCLS
-----------	-------------	----------------------	-------

(3) 外部有識者からの意見聴取

プロジェクトを推進させるために、以下に示す先生方にセミナーをお願いし、海外における科学技術情報のオープン化の状況や、次世代DNAシーケンサ、情報爆発プロジェクトやテキスト処理などに関する最新動向の把握に努めた。

John Wilbanks (Science Commons)、Alan Ruttenberg (Neuro Commons)
 鳥澤健太郎(情報通信研究機構 グループリーダー)、新里圭司(京都大学 特任助教)
 荒木次郎(三菱総合研究所 主任研究員)、牧野貴樹(東京大学 助教)
 木村宏一(日立製作所 主任研究員)、岩崎渉(東京大学)
 畠中秀樹(九州大学 特任准教授)、坂東明日佳(国立遺伝研)

5. 3 その他シンポジウム、セミナー等

(1) 統合データベース講習会

(情報・システム研究機構、かずさ DNA 研究所、長浜バイオ大学)

4. 1. 2 項で記載した統合データベース講習会の詳細は以下のとおりである。

① 統合データベース講習会: AJACS 湘南

日時：平成 20 年 7 月 3 日 (木) 10:30-17:30

場所：日本大学生物資源科学部

参加者：64 名

対象：ライフサイエンス分野のデータベースを効率よく利用することに興味のある方や統合データベースセンターに（就職する）興味のある方

内容：

10:30-12:00 中村保一 自己紹介 【実習】まずは肩慣らし「検索使い倒し」

13:30-14:40 中村保一 ゲノム塩基配列と類似検索 【実習】配列情報へのアクセス、BLAST
あれこれ

14:50-16:00 中村保一 塩基配列の注釈入門と遺伝子構造予測 【実習】遺伝子発見ツール・
スプライスサイト予測・Martでデータ取得

16:10-17:30 中村保一 遺伝子機能予測とDB高度化 【実習】Interpro・GO・CyanoGenes・
KazusaAnnotation

② 統合データベース講習会：AJACS 本郷 1

日時：平成 20 年 7 月 3 日 (木) 10:30-17:30

場所：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

参加者：21 名

対象：ライフサイエンス分野のデータベースを効率よく利用することに興味のある方や統合データベースセンターに（就職する）興味のある方

内容：

午前：講演 10:30-12:00

10:30-10:50 高木利久「統合データベースプロジェクトとは」

10:50-11:10 岡本忍, 中尾光輝, 藤澤貴智, 中村保一

「Kazusa Annotation Suite によるライフサイエンスデータ統合」

11:10-11:30 上原啓史、阿部貴志、池村淑道

「学部学生による持続可能型社会への貢献遺伝子データベースの構築」

11:30-11:50 大久保公策（国立遺伝学研究所） 「科学データは誰のものか？」

午後：講習 13:30-17:30（統合データベースプロジェクトで提供しているサービスの使い方） 13:30-14:00 坊農秀雅 使いこなすための最低限のコンピュータ使いこなし術

14:00-14:30 川本祥子 生命科学横断検索の利用法

14:30-15:00 休憩

15:00-15:30 三橋信孝 アナトモグラフィ

15:30-16:00 山本泰智 OReFiL と Allie

16:00-16:30 休憩

16:30-17:00 坊農秀雅 DBCLS 独自サービス「目次」、「カタログ」系の使いこなし術

17:00-17:30 坊農秀雅 教育・人材育成のリソース紹介他

③ 統合データベース講習会：AJACS 蝦夷

日時：平成 20 年 8 月 29 日（金）10:00-17:30

場所：北海道大学

参加者：32 名

対象：ライフサイエンス分野のデータベースを効率よく利用することに興味のある方や統合データベースセンターに（就職する）興味のある方

内容：

11:10 - 12:00 中村保一 イン트로ダクション【実習】まずは肩慣らし「検索使い倒し」

13:30- 14:40 中村保一 ゲノム塩基配列と類似検索【実習】配列情報へのアクセス、BLAST
あれこれ

14:50 - 16:00 中尾光輝 ゲノム配列の注釈と構造解析～【実習】ゲノムブラウザの利用、
アノテーションの閲覧

16:10 - 17:30 中尾光輝 遺伝子機能予測と DB 高度化～【実習】機能アノテーション付け、
アノテーションの利用

④ 統合データベース講習会：AJACS 博多

日時：平成 20 年 10 月 18 日（土）10:00-17:30

場所：九州大学情報基盤研究開発センター

参加者：43 名

対象：ライフサイエンス分野のデータベースを効率よく利用することに興味のある方や統合データベースセンターに（就職する）興味のある方

内容：

午前：講演 10:00-12:00

10:00-10:30 高木利久「統合データベースプロジェクトとは」

10:30-10:50 佐藤賢二（金沢大学大学院）「ライフサイエンス用語の意味推定」

10:50-11:00 休憩

11:00-11:20 五斗進（京都大学化学研究所）「KEGG Atlas と KAAS アノテーション ～ KEGG における最近の開発から」

11:20-11:50 大久保公策「科学データは誰のものか」

午後：講習 13:30-17:30（統合データベースプロジェクトで提供しているサービスの使い方）

13:30-14:00 河野信「データベースを使いこなすための最低限のコンピュータ使いこなし術」

14:00-15:00 川本祥子・仲里猛留「生命科学横断検索の利用法」

15:00-15:15 休憩

15:15-15:45 山本泰智「OReFiL と Allie」

15:45-16:15 藤枝香「アナトモグラフィ」

16:15-16:30 休憩

16:30-17:00 中村保一「Kazusa Annotation Suite の活用」

17:00-17:30 坊農秀雅「その他 DBCLS 提供サービスの活用術」

⑤ 統合データベース講習会：AJACS 近江

日時：平成 20 年 10 月 31 日（金）13:30-16:50

場所：長浜バイオ大学

参加者：45 名

対象：長浜バイオ大学生命情報コース配属の 3 回生及び一般参加者

内容：

講演 13:30-14:20

13:30-14:00 坊農秀雅「はじめに」

14:00-14:20 鈴木健介（タカラバイオ株式会社ドラゴンジェノミクスセンター）
「次世代型高速シーケンサによるゲノム解析」

講習 14:30-17:00 統合データベースプロジェクトで提供しているサービスの使い方実習
14:30-15:10 川本祥子「生命科学横断検索の利用法」
15:10-15:30 金子聡子「統合データベースプロジェクトにおける人材育成活動の紹介」
15:30-15:50 小野浩雅「統合 TV による DBCLS 提供サービスなどのバイオデータベース
とウェブツールの活用術」
15:50-16:00 休憩
16:00-16:30 藤枝香「アナトモグラフィ」
16:30-17:00 中村保一「Kazusa Annotation Suite の活用」

⑥ 統合データベース講習会：AJACS 長津田

日時：平成 21 年 1 月 23 日（金）10:30-17:30

場所：東京工業大学すずかけ台キャンパス ものづくり教育研究支援センター

参加者：30 名

対象：ライフサイエンス分野のデータベースを効率よく利用することに興味のある方や統
合データベースセンターに（就職する）興味のある方

内容：

午前：講演＋講習 10:30-12:30

10:30-11:00 坊農秀雅「統合データベースプロジェクトとは？」

11:00-11:30 黒川顕「バイオ大航海時代」

11:30-12:30 河野信「データベースを使いこなすための最低限のコンピュータ使いこなし
術」

午後：講習 14:00-17:30（統合データベースプロジェクトで提供しているサービスの使い
方）

14:00-15:00 川本祥子「生命科学横断検索の利用法」

15:00-15:30 藤枝香「アナトモグラフィ/BodyParts3D の利用法」

15:30-16:00 休憩

16:00-16:30 仲里猛留「自然言語処理技術の活用事例」

16:30-17:00 中村保一「Kazusa Annotation Suite の活用」

17:00-17:30 小野浩雅 「遺伝子発現データの活用事例」

（２） BioHackathon2009（情報・システム研究機構）

1）概要

ライフサイエンスにおける標準化技術および標準を利用する技術をより安定化させるた
め、国内外の生命科学研究機関との情報交換および共同での技術基盤開発を目的として、
平成 21 年 3 月 16 日～20 日の 5 日間、ライフサイエンス統合データベースセンターと沖
縄科学技術研究基盤整備機構の協力の下に BioHackathon 2009 を開催した。

初日の 3 月 16 日に一般向けに東京大学武田ホールにて公開シンポジウムを開催した。17 日から 20 日の 4 日間、沖縄科学技術研究基盤整備機構 Seaside House にて、情報技術開発者や生物系研究者が標準化技術のよりよい利用の議論や開発のための開発合宿を行った。シンポジウム参加者は 80 名、17 日からの開発合宿参加者は 58 名（うち海外参加者 30 名）である。

2) 公開シンポジウム

3 月 16 日東京大学武田ホールにて、BioHackathon 2009 のオープニングとして、公開シンポジウムを行った。シンポジウムのプログラムは以下の通りである。

9:00 - 9:20	Introduction to the BioHackathon 2009 and integration efforts in DBCLS	Toshiaki Katayama
9:20 - 9:40	Integrated reproducible analysis with Galaxy	James Taylor
9:40 - 10:00	Biomolecular Function Discovery with the ANNOTATOR	Tobias Gattermayer
10:00 - 10:20	Accelerating experimentation and biological discovery within RIKEN OSC LifeScienceAccelerator -- eeDB gLyphs	Jessica Severin
10:20 - 10:40	BioMart and third party applications: Working towards better interoperability	Arek Kasprzyk
10:40 - 11:00	The Taverna 2 Workflow Engine API : Extensible, High Performance and Easy to Integrate	Tom Oinn
11:00 - 11:20	Multi-repository architecture for service integration	Oswaldo Trelles
11:20 - 11:40	Integrating literature and annotation resources: the Elsevier Grand Challenge experience	Alberto Labarga
11:40 - 12:00	Looking back on a WABI life for tomorrow	Hideaki Sugawara

3) 開発合宿

3 月 17 日から 20 日の 4 日間、沖縄科学技術研究基盤整備機構の施設である Seaside House にて、開発合宿を開催した。開発合宿では、初日の午前中に参加者全員で共通の問題意識について話し合い、それをもとにいくつかのワーキンググループに分かれ、初日午後以降、グループごとに活発な議論と開発を行った。参加者が多かったグループとして、次世代シーケンサのデータをはじめとした巨大なデータについてどのようにデータを溜めどのように利用するかを議論するグループ、セマンティックウェブ技術をどのように生物系のデータに利用していくか議論するグループ、タンパク質相互作用やゲノムなどの大きなデータをいかに可視化するかについて議論するグループがあった。その他、具体的なユ

ースケースについて議論するグループ、文献データの処理について議論するグループ、ゲノム解析プラットフォーム Galaxy の国際化や今後の展開について議論・開発するグループなど、十数のグループで活発な議論が行われ、開発に反映された。またグループ間の議論の交換も頻繁に行われた。例えば、巨大なデータをどのように蓄積し、どの利用するかに関する話題は、それをどのように可視化すべきかという話題に密接に関わってくる。そのため、関連するグループの参加者間では意見交換や情報交換が行われ、結果として、横断的に開発に生かされることになった。

この合宿の成果物は大きく二つに分けることができる。すなわち、開発による成果物、および、情報共有による成果物である。開発による成果は、セマンティック自動発見統合フレームワーク SADI に対するプログラミング言語 Perl API（アプリケーション・プログラミング・インターフェイス）開発、Galaxy 国際化のための開発、Galaxy とゲノムブラウザ gbrowse の間でのデータ入出力機能の開発、ゲノム解析環境 g-language の新しいデータ交換機能や解析インタフェースの開発、癌ゲノムデータ閲覧ソフトウェア PARP に対する 3D 可視化ツールの開発、可視化ツール Processing でネットワーク解析ツール Cytoscape の結果を表示するための開発、DDBJ-KEGG-PDB ワークフロー連携のための開発、糖鎖情報解析のためのワークフローの開発などが挙げられる。これだけの量のシステム開発が5日間で可能であったのは、関係する開発者同士が顔を合わせて議論をすすめお互いを刺激しながら開発を行う Hackathon スタイルならではのといえる。情報共有による成果は、参加者同士が問題意識の共有を行ったこと、あるいは、顔を合わせて密接に議論したことで、今後の共同研究などにつながっている。実際に、この会議をきっかけに転写制御のグループのメンバーは共同研究を始めており、将来的には潜在的な転写制御を検索するサービスを開始する予定である。また、遺伝子アノテーションパイプライン ANNOTATOR の大規模解析に東京工業大学のスーパーコンピュータ TSUBAME を利用するなどの提案、Galaxy の開発について開発者とライフサイエンス統合データベースセンターの要望についての情報交換、ウェブサービス設計ガイドラインについての議論コミュニティの作成など、BioHackathon 2009 での情報共有からはじまる共同研究の輪は確実に広がっている。

4) 最後に

会議開催後におこなった参加者アンケートでは、他の参加者と議論する十分な時間があつたこと、システムを開発するのに非常に向いた環境であつたことなどの理由で BioHackathon 2009 は高い評価を得ている。参加者からは「セマンティックウェブや巨大データなどのホットなトピックについて議論ができてよかった」「生物系研究者から直接ニーズをきけてよかった」「普段なら数か月かかる開発を5日で終わらせることができた」など、他の機会では得られない成果が得られたという声があがっており、これらの感想はこの会議が主催者の目論見どおりの形で成功裏に終わったことを示唆していると思わ

れる。

(3) ゲノムテクノロジー第 164 委員会との共催による研究会(情報・システム研究機構)

2. 2 項で記載した、日本学術振興会ゲノムテクノロジー第 164 委員会と共催した、「生命科学データの共有に向けて:統合データベースプロジェクトの挑戦」と題した研究会の詳細を以下に示す。

日時:平成 20 年 7 月 22 日(木) 13:30 – 17:30

場所:東京大学山上会館

参加者:約 50 名

プログラム:

13:30-13:45 統合データベースプロジェクト

ライフサイエンス統合データベースセンター 高木利久

13:45-14:15 統合データベースプロジェクトの提供サービスについて

ライフサイエンス統合データベースセンター 川本祥子

14:15-14:45 疾患解析から医療応用を目指す DB 開発:ゲノムワイド関連解析 DB と
リシーケンシング DB

東京大学大学院医学系研究科 徳永勝士

14:45-15:15 日本糖鎖科学統合 DB の取り組みについて

産総研糖鎖医工学センター 成松久

15:15-15:45 休憩(公開サービスのデモ)

15:45-16:15 科学を社会に役立てるには?

国立遺伝学研究所 大久保公策

16:15-17:30 パネルディスカッション「統合 DB に期待するもの」

エーザイ(株) 長洲毅志

協和発酵工業(株) 太田紀夫

理化学研究所 権藤洋一

東京工業大学 秋山 泰

国立遺伝学研究所 大久保公策

(司会) ライフサイエンス統合データベースセンター 高木利久

(4) バイオインフォマティクスへの招待(お茶の水女子大学)

お茶の水女子大学で DB 高度利用者養成の一環として行ったセミナー「バイオインフォマティクスへの招待」の一覧を以下に示す。

第 13 回バイオインフォマティクスへの招待

日時:平成 20 年 5 月 29 日

場所:お茶の水女子大学

参加人数:33 人

タイトル：グライコミクスのためのバイオインフォマティクス

発表者：中原 拓

第 14 回バイオインフォマティクスへの招待

日時：平成 20 年 6 月 26 日

場所：お茶の水女子大学

参加人数：30 人

タイトル：タンパク質の分子進化の制約要因としての相互作用

発表者：藤 博幸

第 15 回バイオインフォマティクスへの招待

日時：平成 20 年 7 月 17 日

場所：お茶の水女子大学

参加人数：25 人

タイトル：SVM 利用に役立つカーネル設計の手法

発表者：申 吉浩

第 16 回バイオインフォマティクスへの招待

日時：平成 20 年 9 月 25 日

場所：お茶の水女子大学

参加人数：17 人

タイトル：イネゲノム計画におけるバイオインフォマティクス

発表者：伊藤 剛

第 17 回バイオインフォマティクスへの招待

日時：平成 20 年 10 月 23 日

場所：お茶の水女子大学

参加人数：19 人

タイトル：比較ゲノム解析から読み解く受容体の機能多様化メカニズム

発表者：諏訪 牧子

タイトル：トランスクリプトーム解析におけるバイオインフォマティクス

発表者：油谷 幸代

第 18 回バイオインフォマティクスへの招待

日時：平成 20 年 11 月 13 日

場所：お茶の水女子大学

参加人数：28 人

タイトル：シグナル伝達パスウェイのバイオインフォマティクス

発表者：高井 貴子

(5) アノテータ、キュレータ養成のための運営委員会と公開イベント（長浜バイオ大学）

長浜バイオ大では、学部レベルでのアノテータ・キュレータの人材養成を効果的に行うために、外部の専門家を含んだ運営委員会を独自に構成している。この運営委員会は、シニア世代の研究者を中心メンバーとしており、学部学生の作成したデータベースの精査をも実施し、品質の高いデータベース構築のための実践的な活動を、学部生との共同作業として実施している。人材育成の実施に積極的に関与しながら、問題点の具体的な洗い出しや評価、取り組むべき課題の方向性を検討した。

本運営委員会は、本年度は以下の日程で開催した。なお、学部学生との共同作業によるデータベース構築の実践例として、「エキスパートのキュレートした tRNA 遺伝子データベース」の精査と構築をおこなった。シニア世代によるデータベースの精査とキュレーションは、日常的にはインターネットを活用して分散型で行い、問題点や取り組むべき課題の洗い出しの検討や評価は長浜バイオ大で開催する会議で行った。シニア世代専門家の精査・キュレーション後のデータベースは、長浜バイオ大学、及び情報・システム研究機構統合 DB センターより公開している。

① 2008.5.29-30

参加者：池村淑道、阿部貴志、上原啓史、井口八郎、武藤あきら、山田優子、金谷重彦、木ノ内誠、学部学生 (1 名)

内容：WGS データ判定と問題点の検討会議

② 2008.6.5-6

参加者：池村淑道、阿部貴志、上原啓史、香川弘昭

内容：統合 DB プロジェクト運営に関する打合せ

③ 2008.11.26-28

参加者：池村淑道、阿部貴志、上原啓史、井口八郎、武藤あきら、山田優子、金谷重彦、金井昭夫、菅原潤一、学部学生 (1 名)

内容：tRNA 遺伝子判定と取り組むべき課題に関する会議

④ 2009.3.18-19

参加者：池村淑道、阿部貴志、上原啓史、井口八郎、武藤あきら、山田優子、金谷重彦、木ノ内誠、学部学生 (1 名)

内容：tRNA 遺伝子判定と DB プロジェクトの長浜バイオ大における次年度計画に関する検討会議

また、長浜バイオ大学のオープンキャンパスとして、平成 20 年 7 月 27 日と 11 月 26 日の 2 回、『『持続可能型社会への貢献遺伝子データベース』の構築にあなたも参加してみ

ませんか？」というテーマで上原啓史が公開イベントを行った。大学生・高校生・一般人が対象で、それぞれ約 50 名および約 40 名の参加を得た。その際、自宅や自分の学校に帰っても使用できる自習用の教材を配布したことが喜ばれた。また、自宅や高校からも遺伝子探索が可能な点が驚かれた。

6. 成果の外部への発表

論文、学会発表等の発表内容については、添付の様式第 2 1（79 ページ以降）を参照されたい。また、新聞、雑誌等における本プロジェクトに関する記事は以下の通りである。

タイトル	メディア名	掲載日
ライフサイエンス統合 DB センターがポータルサイトで各 DB を公開、「蛋白質核酸酵素」の総説記事は 6 月 2 日	Biotechnology Japan (日経 BP)	H20.5.28
ライフサイエンス分野 データベースを 11 年度から一元化(1 面)・ライフサイエンスの発展 真に必要なデータベースとは	科学新聞	H20.12.5
情報の循環を改善する統合 DB プロジェクト本格化	BTJ ジャーナル	H21.3.25
生命科学分野の DB から文献まで扱うわが国初のサービスを目指す	Medical Tribune	H21.3.26

7. 実施体制

別表 1、2、3（90 ページ以降）を参照されたい。

8. 整備実績一覧

別紙 1（93 ページ以降）を参照されたい。

様式第 2 1

学 会 等 発 表 実 績

委託業務題目: 「ライフサイエンス統合データベース開発運用」(戦略立案・実行評価／統合データベース開発／統合データベース支援)

機 関 名: 大学共同利用機関法人 情報・システム研究機構、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

財団法人 かずさディー・エヌ・エー研究所、国立大学法人 九州大学、国立大学法人 奈良先端科学技術大学院大学

国立大学法人 東京大学、国立大学法人 お茶の水女子大学、学校法人 関西文理総合学園長浜バイオ大学

1. 学会等における口頭・ポスター発表

発表した成果（発表題目、口頭・ポスター発表の別）	発表者氏名	発表した場所（学会等名）	発表した時期	国内・外の別
統合データベースセンターの紹介 (ポスター)	河野 信、坊農 秀雅、全 弘宇、 藤枝 香、川本 祥子、高祖 歩美、 峰崎 雄一、箕輪 真理、三橋 信 孝、永井 啓一、仲里 猛留、西川 哲夫、大野 忠、小野 浩雅、山口 敦子、山本 泰智、八塚 茂、大久 保 公策、高木 利久	第 8 回日本分子生物学会春季シンポジウム	2008 年 5 月 26-27 日	国内(札幌)
Prediction of protein sub-cellular localization using information from texts and sequences. (ポスター)	Hong-Woo Chun, Chisato Yamasaki, Naomi Saichi, Masayuki Tanaka, Teruyoshi Hishiki, Jin-Dong Kim, Jun ' ichi Tsujii, Tadashi Imanishi, Takashi Gojobori and Toshihisa Takagi.	The Annual Meeting of the Association for Computational Linguistics (ACL) with the Human Language Technology Conference (HLT)	2008 年 6 月 15-20 日	国外 (Ohio USA)

ライフサイエンス統合データベースプロジェクト(口頭)	坊農秀雅、全弘宇、藤枝香、川本祥子、河野信、高祖歩美、峰崎雄一、箕輪真理、三橋信孝、永井啓一、仲里猛留、西川哲夫、大野忠、小野浩雅、山口敦子、山本泰智、八塚茂、大久保公策、高木利久	第 28 回日本糖質学会年会	2008 年 8 月 18-20 日	国内(つくば)
KazusaAnnotation Suite: 植物関連生物ゲノム情報統合の試み(口頭)	岡本 忍、中尾 光輝、藤澤 貴智、中村 保一	日本進化学会第 10 回大会	2008 年 8 月 22-24 日	国内(東京)
統合データベースプロジェクトにおける人材育成活動の紹介(口頭)	金子 聡子	日本進化学会第 10 回大会	2008 年 8 月 22-24 日	国内(東京)
統合 TV: 動画によるデータベースの使い方講座(口頭)	河野 信	日本進化学会第 10 回大会	2008 年 8 月 22-24 日	国内(東京)
統合データベースプロジェクトとライフサイエンス統合データベースセンター(口頭)	坊農 秀雅	日本進化学会第 10 回大会	2008 年 8 月 22-24 日	国内(東京)
ライフサイエンス統合データベースセンターにおける言語処理関係の活動(口頭)	山本 泰智、大久保 公策、川本祥子、全 弘宇、高木 利久	第 38 回 人工知能学会分子生物情報研究会	2008 年 8 月 28 日	国内(東京)
Practical Organization of Sequence Data as Gene Expression Data(ポスター)	Hidemasa Bono, Shoko Kawamoto, Kousaku Okubo and Toshihisa Takagi	The 8th Cold Spring Harbor Laboratory/Wellcome Trust conference on Genome Informatics	2008 年 9 月 10-14 日	国外 (Hinxton UK)
Japan started data-sharing center for publicly-funded biomedical science(ポスター)	Kousaku Okubo, Shoko Kawamoto, Hidemasa Bono and Toshihisa Takagi	The 8th Cold Spring Harbor Laboratory/Wellcome Trust conference on Genome Informatics	2008 年 9 月 10-14 日	国外 (Hinxton UK)
データの保護と利用のための情報セキュリティ(口頭)	箕輪真理	科学研究費補助金特定領域研究「ゲノム」4 領域シンポジウム「ゲノム医学のこれから」	2008 年 10 月 14 日	国内(東京)

Text-related Services for Researchers in Life Sciences(ポスター)	Yaunori Yamamoto and Toshihisa Takagi	BIBM2008	2008 年 11 月 3-5 日	国外 (Phila. USA)
生命科学データの共有に向けた「統合データベースプロジェクト」の試み(ポスター)	箕輪真理、坊農秀雅、全弘宇、藤枝香、川本祥子、河野信、高祖歩美、熊谷禎洋、三橋信孝、永井啓一、仲里猛留、西川哲夫、大村蓉子、大野忠、小野浩雅、山口敦子、山口瑤子、山本泰智、八塚茂、横山貴央、大久保公策、高木利久	情報とシステム 2008 「社会のイノベーションを誘発する情報システム 08	2008 年 11 月 5 日	国内(東京)
低酸素刺激における比較トランスクリプトーム解析(口頭及びポスター発表)	坊農秀雅	第 6 回がんとハイポキシア研究会	2008 年 11 月 29-30 日	国内(広島)
A Literature Management System with A Recommendation Function at Database Center for Life Science.(ポスター)	Yaunori Yamamoto and Toshihisa Takagi	The 19th International Conference on Genome Informatics	2008 年 12 月 1-3 日	国外 (Gold Coast)
Life Science Database Portal Site: lifesciencedb.jp.(ポスター)	Shin Kawano, Hidemasa Bono, Shoko Kawamoto, Kousaku Okubo and Toshihisa Takagi	The 19th International Conference on Genome Informatics	2008 年 12 月 1-3 日	国外 (Gold Coast)
Functional organization of transcript sequences as gene expression data.(ポスター)	Hidemasa Bono, Shoko Kawamoto, Kousaku Okubo and Toshihisa Takagi	The 19th International Conference on Genome Informatics	2008 年 12 月 1-3 日	国外 (Gold Coast)
統合データベースプロジェクトは何の役に立つのか?(口頭)	高木利久	第 46 回日本生物物理学会年会 シンポジウム「動き出したライフサイエンス統合データベース」	2008 年 12 月 3-5 日	国内(福岡)
統合 DB プロジェクトのサービスについて(口頭)	川本祥子	第 46 回日本生物物理学会年会 シンポジウム「動き出したライフサイエンス統合データベース」	2008 年 12 月 3-5 日	国内(福岡)

統合 DB の運営と利用にむけた教育活動(口頭)	瀬々 潤	第 46 回日本生物物理学会年会 シンポジウム「動き出したライフサイエンス 統合データベース」	2008 年 12 月 3-5 日	国内(福岡)
DDBJ, PDBj, KEGG, CBRC のウェブサービスによる連携(口頭)	片山俊明	第 46 回日本生物物理学会年会 シンポジウム「動き出したライフサイエンス 統合データベース」	2008 年 12 月 3-5 日	国内(福岡)
ライフサイエンス統合データベースポータル lifesciencedb.jp	河野信、坊農秀雅、全弘宇、藤枝香、川本祥子、高祖歩美、箕輪真理、三橋信孝、永井啓一、仲里猛留、西川哲夫、大野忠、小野浩雅、山口敦子、山本泰智、八塚茂、大久保公策、高木利久	第 31 回日本分子生物学会年会 第 81 回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)
MeSH を用いた生物学的解釈によるマイクロアレイデータの解析の実際(ポスター)	仲里猛留、坊農秀雅、松田秀雄、高木利久	第 31 回日本分子生物学会年会 第 81 回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)
統合 TV - 動画によるバイオデータベースとウェブツールの活用法解説サイト(ポスター)	小野 浩雅、河野 信、大村 蓉子、山口 瑤子、竹若 浩一、川本 祥子、高木 利久、坊農 秀雅	第 31 回日本分子生物学会年会 第 81 回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)
BodyParts3D: 解剖学概念を表現した 3 次元人体構造データベース(ポスター)	三橋 信孝、藤枝 香、田村 卓郎、川本 祥子、高木 利久、大久保 公策	第 31 回日本分子生物学会年会 第 81 回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)
ライフサイエンス分野のメタデータ要素レポジトリ MDeR の構築(ポスター)	黒田 雅子、小池 俊行、安田 絵美、佐藤 早苗、藤田 晶子、西村 佑介、河村 明子	第 31 回日本分子生物学会年会 第 81 回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)

文科省「統合データベースプロジェクト」における生命科学データベース横断検索システムの開発(ポスター)	牧口 大旭、川原 麗子、大野 忠、奥村 利幸、川本 祥子、大久保 公策、高木 利久	第31回日本分子生物学会年会 第81回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)
分子生物学の研究動向の俯瞰を目的としたデータバンク(INSDC, GEO)目次の開発(ポスター)	小笠原 理、渡邊 康司、栗原 英輔、大久保 公策	第31回日本分子生物学会年会 第81回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)
学部教育としての健康への貢献遺伝子データベース構築(ポスター)	上原 啓志、阿部 貴志、池村淑道	第31回日本分子生物学会年会 第81回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)
エキスパートがキュレートした tRNA 遺伝子データベース(ポスター)	阿部 貴志、池村淑道、小原 康雄、上原 啓志、中泉 友紀、平野 晋也、木ノ内 誠、金谷 重彦、山田 優子、武藤 晃、井口 八郎	第31回日本分子生物学会年会 第81回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)
一括学習型自己組織化地図法 (Batch-Learning Self-Organizing Map: BLSOM)による環境生物ゲノム由来断片配列からの知識発見(口頭及びポスター)	阿部貴志、金谷重彦、池村淑道	第31回日本分子生物学会年会 第81回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)
持続可能型社会ならびに健康への貢献遺伝子データベース(ポスター)	上原啓史、阿部貴志、棚橋佳世、陳キン、陸嘉俊、田中秀和、池村淑道	第31回日本分子生物学会年会 第81回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)
KazusaWiki: ゲノムデータベースにおける Wiki の活用(ポスター)	藤澤 貴智、岡本 忍、中尾 光輝、中村 保一	第31回日本分子生物学会年会 第81回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)
Gene Indexing:KazusaAnnotation による遺伝子注釈実証実験(ポスター)	岡本 忍、中尾 光輝、藤澤 貴智、中村 保一	第31回日本分子生物学会年会 第81回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)
かずさアノテーション:持続可能な遺伝子アノテーション整備のための新型システム(ポスター)	中尾 光輝、岡本 忍、藤澤 貴智、中村 保一	第31回日本分子生物学会年会 第81回日本生化学会大会 合同大会 (BMB2008)	2008 年 12 月 9-12 日	国内(神戸)

統合データベースプロジェクトのサービスについて(口頭)	川本祥子	第31回日本分子生物学会年会 第81回日本生化学会大会 合同大会 (BMB2008) シンポジウム「ライフサイエンス統合データベースプロジェクト」	2008 年 12 月 9-12 日	国内(神戸)
日本のライフサイエンス統合データベース構築・維持のための人材育成活動(口頭)	坊農秀雅	第31回日本分子生物学会年会 第81回日本生化学会大会 合同大会 (BMB2008) シンポジウム「ライフサイエンス統合データベースプロジェクト」	2008 年 12 月 9-12 日	国内(神戸)
科学データは誰のものか？(口頭)	大久保公策	第31回日本分子生物学会年会 第81回日本生化学会大会 合同大会 (BMB2008) シンポジウム「ライフサイエンス統合データベースプロジェクト」	2008 年 12 月 9-12 日	国内(神戸)
Funtional profiling of OMIM data using MeSH vocabular(口頭及びポスター)	Takeru Nakazato, Hidemasa Bono, Hideo Matsuda, Toshihisa Takagi	2008 年日本バイオインフォマティクス学会年会 (JSBi2008)	2008 年 12 月 15-16 日	国内(大阪)
Semantic clustering of biomedical abbreviation definitions using bipartite graph(ポスター)	Atsuko Yamaguchi, Yasunori Yamamoto, Hong-Woo Chun and Toshihisa Takagi	The 14th Pacific Symposium on Biocomputing (PSB)	2009 年 1 月 5-9 日	国外 (Hawaii)
ライフサイエンス統合データベースセンサーが提供する微生物ゲノムアノテーションパイプライン(ポスター)	菅原 秀明, 森 宙史, 黒川 顕	第3回日本ゲノム微生物学会年会	2009 年 3 月 5-7 日	国内(東京)
一括学習型自己組織化地図法による全ゲノム・全タンパク質配列を対象にした大規模ポストゲノム解析(ポスター)	阿部貴志、金谷重彦、池村淑道	第3回日本ゲノム微生物学会年会	2009 年 3 月 5-7 日	国内(東京)

エキスパートがキュレートした tRNA 遺伝子データベース(ポスター)	阿部貴志、小原康雄、上原啓史、中泉友紀、平野晋也、木ノ内誠、金谷重彦、山田優子、武藤昱、井口八郎、池村淑道	第3回日本ゲノム微生物学会年会	2009年3月5-7日	国内(東京)
健康への貢献ならびに持続可能型社会への貢献遺伝子データベース(ポスター)	上原啓史、阿部貴志、中泉夕紀、和田千恵子、井口八郎、池村淑道	第3回日本ゲノム微生物学会年会	2009年3月5-7日	国内(東京)
機械学習と系列アラインメントを応用した日本語並列句解析(口頭)	大熊秀治、原一夫、新保仁、松本裕治	人工知能学会全国大会	2008年6月11-13日	国内(旭川)
統合データベース・ワークフロー、ポスター	田代俊行	産総研・産技連 LS-BT 合同研究発表会	2009年1月29-30日	国内(つくば)
GENIA コーパスからのネスト並列句同定(口頭)	原一夫、新保仁、大熊秀治、松本裕治	情報処理学会自然言語処理研究会	2008年9月24-25日	国内(熱海)
Graph-based Analysis of Semantic Drift in Espresso-like Bootstrapping Algorithms.(口頭)	Mamoru Komachi, Taku Kudo, Masashi Shimbo, Yuji Matsumoto	Conference on Empirical Methods in Natural Language Processing	2008年10月25-27日	国外(Hawaii)
Classifying Narrative Patient Records without Any External Resources.	Kazuo Hara	Second i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data	2008年11月7日	国外(Washington D. C.)
グラフを用いたバイオ医療専門用語の類義語推定(口頭)	鈴木郁美、原一夫、新保仁、松本裕治	情報処理学会自然言語処理研究会	2009年1月22-23日	国内(東京)
Evolution of a gene family under human-specific downregulation.(ポスター)	Satoko Kaneko, Naoyuki Takahata, and Yoko Satta	Society for Mol. Biol. and Evol. 2008	2008年6月5-8日	国外(Spain)
Evolution of a gene family under human-specific downregulation.(口頭)	Satoko Kaneko, Naoyuki Takahata, and Yoko Satta	IFOM-IEO	2008年6月10日	国外(Italy)

Evolutionary dynamics of “functional” processed pseudogenes that stabilize Makorin1 mRNAs in mammals.(口頭)	Satoko Kaneko, Naoyuki Takahata, and Yoko Satta	JSPS Japan-Germany Colloquium 2009	2009 年 3 月 8-10 日	国外(Germany)
学部教育における DDBJ の活用について(口頭)	阿部貴志	DDBJing 講習会	2008 年 6 月 19 日	国内(三島)
ヒト特異的に発現量が低下している遺伝子群の進化(口頭)	金子 聡子、高畑 尚之、颯田 葉子	日本遺伝学会第 80 回大会	2008 年 9 月 3-5 日	国内(名古屋)
エキスパートがキュレートした tRNA データベース	阿部貴志、池村淑道、小原康雄、武藤あきら、山田優子、上原啓史、木ノ内誠、金谷重彦、井口八郎	日本遺伝学会第 80 回大会	2008 年 9 月 3-5 日	国内(名古屋)
学部教育としての健康への貢献遺伝子データベース構築	上原啓史、阿部貴志、棚橋佳世、池村淑道	日本遺伝学会第 80 回大会	2008 年 9 月 3-5 日	国内(名古屋)
一括学習型自己組織化地図法 (Batch Learning Self-Organizing Map: SOM) による環境微生物ゲノム群集の解明	阿部貴志、金谷重彦、池村淑道	日本遺伝学会第 80 回大会	2008 年 9 月 3-5 日	国内(名古屋)
ゲノム配列に潜む生物種による特徴をコンピュータで探る	池村淑道	日本遺伝学会第 80 回大会 市民公開講座	2008 年 9 月 3-5 日	国内(名古屋)
A large-scale genome analysis conducted by a small group with reference to “Database of Genes Contributing to Sustainable World” and to Genome Research in Japan.(口頭)	Toshimichi Ikemura	The 13th “Science in Japan” Forum of JSPS	2008 年 6 月 20 日	国外 (Washington D. C.)
Identification of Loci for Systemic Lupus Erythematosus by Pooling-based Genome-wide Association Study. (ポスター)	Tahira T, Masumoto K, Kukita Y, Okazaki Y, Yoshinaga A, Higasa K, Horiuchi T, Hayashi K	The 58th Annual Meeting of The American Society of Human Genetics	2008 年 11 月 11-15 日	国外 (Phila. USA)

Detection of human copy number variations using complete hydatidiform moles carrying haploid genomes. (ポスター)	Kukita Y, Higasa K, Tahira T, Hayashi K	The 58th Annual Meeting of The American Society of Human Genetics	2008 年 11 月 11-15 日	国外 (Phila. USA)
Impact of definitive haplotypes on genetical analysis. (ポスター)	Higasa K, Kukita Y, Tahira T, Hayashi K	The 58th Annual Meeting of The American Society of Human Genetics	2008 年 11 月 11-15 日	国外 (Phila. USA)
Impact of definitive haplotypes for detecting recent positive selection. (ポスター)	Higasa K, Kukita Y, Hayashi K	The Biology of Genome, Cold Spring Harbor Laboratory, New York	2008 年 5 月 6-10 日	国外 (New York)
Detection of human copy number variations using a collection of complete hydatidiform moles: A strategy of CNV-detection using human haploid genomes. (ポスター)	Kukita Y, Higasa K, Tahira T, Hayashi K	The Biology of Genome, Cold Spring Harbor Laboratory, New York	2008 年 5 月 6-10 日	国外 (New York)
Kazusa genome annotation systems (口頭)	Yasukazu Nakamura	The 5th Rice Annotation Project Meeting	2008 年 11 月 15 日	国内(Tokyo)
植物ゲノム情報を基盤とした統合データベース (口頭)	岡本 忍、中尾 光輝、藤澤 貴智、中村 保一	日本育種学会 2008 年秋季大会シンポジウム「植物育種におけるオミックス情報の利用と展望」	2008 年 10 月 12 日	国内(彦根)
ゲノム塩基配列を基盤としたライフサイエンス研究情報統合環境 (口頭)	岡本 忍、中尾 光輝、藤澤 貴智、中村 保一	明治大学農学部ワークショップ「大規模オミックス・データの利用と展望」	2008 年 9 月 22 日	国内(川崎)
小回りが利いて役に立つゲノム情報データベースの提案 (口頭)	岡本 忍、中尾 光輝、藤澤 貴智、中村 保一	第 26 回 日本植物細胞分子生物学会シンポジウム「農作物とモデル植物におけるオミックス情報の利用と展望」	2008 年 9 月 2 日	国内(大阪)

2. 学会誌・雑誌等における論文掲載

掲載した論文（発表題目）	発表者氏名	発表した場所（学会誌・雑誌等名）	発表した時期	国内・外の別
BodyParts3D: 3D structure database for anatomical concepts	Nobutaka Mitsuhashi, Kaori Fujieda, Takuro Tamura, Shoko Kawamoto, Toshihisa Takagi and Kousaku Okubo	Nucleic Acids Research, 2009, Vol. 37, Database issue D782-D785	2008 年 10 月 3 日	国外
tRNADB-CE: tRNA gene database curated manually by experts	T. Abe, T. Ikemura, Y. Ohara, H. Uehara, M. Kinouchi, S. Kanaya, Y. Yamada, A. Muto and H. Inokuchi	Nucleic Acids Research, 2009, Vol. 37, Database issue D163-D168	2008 年 10 月 8 日	国外
Narrowing the regions of allelic loss of chromosome 1p36 in meningioma tissues by an improved SSCP analysis.	Guan Y, Hata N, Kuga D, Yoshimoto K, Mizoguchi M, Shono T, Suzuki SO, Tahira T, Kukita Y, Higasa K, Yokoyama N, Nagata S, Iwaki T, Sasaki T, Hayashi K	Int. J. Cancer 122:1820-1826	2008 年 4 月	国外
Prevalence of copy-number neutral LOH in glioblastomas revealed by genome-wide high-density oligonucleotide array analysis of laser-microdissected tissues.	Kuga D, Mizoguchi M, Guan Y, Hata N, Yoshimoto K, Shono T, Suzuki SO, Kukita Y, Tahira T, Nagata S, Sasaki T, Hayashi K	Neuro-Oncology 10(6):995-1003	2008 年 8 月	国外
Estimation of SNP allele frequencies by SSCP analysis of pooled DNA	Tahira T, Kukita Y, Higasa K, Okazaki Y, Yoshinaga A, Hayashi K	Methods in Molecular Biology: SNP 2nd edition in press	2008 年 9 月 (accepted)	国外

Discovery of unusual haplotype structure in Asian genome by comparing inferred and definitive haplotype datasets.	Higasa K, Kukita Y, Kato K, Wake N, Tahira T and Hayashi K	Plos Genetics in press	2009 年 3 月 (accepted)	国外
蛋白質核酸酵素・連載 ライフサイエンス分野の統合データベース 「『蛋白質 核酸酵素』バックナンバーの全文検索サービス公開にあたって」	川本祥子	蛋白質核酸酵素 Vol.53, No.9, pp.1200-1205	2008 年 7 月	国内
ライフサイエンス分野のデータベース：学部学生とシニア世代の共同作業としての知識発見の可能性をも含めて	池村淑道	日本化学会情報化部会誌 特集 Vol.26, No.1, pp.1-4	2008 年 4 月	国内
ライフサイエンス分野のデータベースの現状と統合データベースプロジェクト	坊農秀雅、高木利久	日本化学会情報化部会誌 特集 Vol.26, No.1, pp.5-7	2008 年 4 月	国内
ゲノムのユーザを育て、研究領域を育てる：ゲノムリテラシー教育活動とオープンゲノムアノテーションの試み	中村保一	日本化学会情報化部会誌 特集 Vol.26, No.1, pp.8-10	2008 年 4 月	国内
エキスパートがキュレートした tRNA データベース	井口八郎、小原康雄、武藤昱、山田優子、木ノ内誠、前野聖、金谷重彦、池村淑道、阿部貴志	日本化学会情報化部会誌 特集 Vol.26, No.1, pp.11-16	2008 年 4 月	国内
持続可能型社会への貢献遺伝子データベース	上原啓史、棚橋佳世、阿部貴志、池村淑道	日本化学会情報化部会誌 特集 Vol.26, No.1, pp.17-19	2008 年 4 月	国内
ゲノム配列情報からの効率的な知識発見のための情報学的手法の確立	阿部貴志、金谷重彦、池村淑道	日本化学会情報化部会誌 特集 Vol.26, No.1, pp.20-22	2008 年 4 月	国内

別表1 平成20年度に於ける実施体制

研究項目	担当機関等	研究担当者
1. 戦略立案・実行評価	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 情報・システム研究機構 国立遺伝学研究所 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 同上	◎○高木 利久 大久保 公策 永井 啓一 箕輪 真理 西川 哲夫
2. 統合データベース開発	情報・システム研究機構 国立遺伝学研究所	○大久保 公策
(1) 共通基盤技術開発	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 同上	高木 利久 山口 敦子
①共通基盤技術開発	同上 同上 同上 同上 同上 東京大学医科学研究所ヒトゲノム解析センター 同上 かずさDNA研究所植物ゲノム基盤研究部 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 同上 同上 同上 同上 同上 同上 同上 同上 同上 同上	山本 泰智 全 弘宇 S. Riedel 大内田 賢太 片山 俊明 川島 秀一 中尾 光輝 浅井 潔 野口 保 諏訪 牧子 光山 統泰 Horton Poul 広川 貴次 福井 一彦 藤渕 航 堀本 勝久 田代 俊行 松本 裕治 新保 仁 浅原 正幸 原 一夫 渡邊 陽太郎 山田 恵美子
②ワークフロー技術を用いた統合DB環境構築		
③専門用語辞書管理システムと専門用語解析技術の開発	奈良先端科学技術大学院大学 情報科学研究科 同上 同上 同上 同上 同上	

注1. ◎:課題代表者、○:サブテーマ代表者

注2. 本業務に携わっている方は、全て記入。

別表2 平成20年度に於ける実施体制

研究項目	担当機関等	研究担当者
(2) ヒト統合DBの開発・運用 ① ヒト統合DBの開発・運用	情報・システム研究機構 国立遺伝学研究所 同上 同上 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 同上 大阪府立成人病センター研究所 九州大学生体防御医学研究所	大久保 公策 同上 小笠原 理 三橋 信孝 藤枝 香 加藤 菊也 林 健志
② 多型知識表現技術の開発	九州大学生体防御医学研究所	菅原 秀明
(3) モデル生物・産業応用生物統合DB ① モデル生物・産業応用生物統合DB	情報・システム研究機構 国立遺伝学研究所 同上 情報・システム研究機構 国立遺伝学研究所 奈良先端科学技術大学院大学 情報科学研究科 自然科学研究機構 基礎生物学研究所 農業生物資源研究所 基礎研究領域	同上 大久保 公策 黒川 顕 内山 郁夫 伊藤 剛
② 植物及び植物関連微生物のゲノム 情報データベース統合と高度化	かずさDNA 研究所植物ゲノム基盤研究部 同上 同上 同上 同上	中村 保一 岡本 忍 中尾 光輝 藤代 維一 小原 光代
3. 統合データベース支援 (1) ポータル整備・運用、広報、普及啓発	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 同上 同上 同上 同上 同上 同上 同上 同上 同上 同上	○川本 祥子 同上 坊農 秀雅 河野 信 仲里 猛留 箕輪 真理 高祖 歩美 小野 浩雅 山口 瑤子 大村 蓉子 千葉 温子 横山 貴央
(2) データベース受入れ	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 同上 同上	西川 哲夫 中村 春木 金城 玲

注1. ◎:課題代表者、○:サブテーマ代表者

注2. 本業務に携わっている方は、全て記入。

別表3 平成20年度に於ける実施体制

研 究 項 目	担当機関等	研究担当者
(3) 人材育成		
①DB 構築者の養成	東京大学大学院 創成科学研究科 同上 同上	森下 真一 中谷 洋一郎 吉村 淳
②DB 高度利用者の養成	お茶の水女子大学大学院 人間文化創成科学 研究科 同上 同上 同上 同上 同上	瀬々 潤 鷹野 景子 小口 正人 加藤 毅 渡辺 知恵美 近藤 るみ
③アナテータ・キュレータの養成	長浜バイオ大学バイオサイエンス学部 同上 同上 同上 同上 同上 同上 同上 同上 同上 同上 同上 同上 同上 奈良先端科学技術大学院大学 山形大学工学部 京都大学大学院 医学研究科 弘前大学 農学生命科学部 株式会社吉田生物研究所 所属機関無し 首都大学東京 岡山大学大学院 奈良先端科学技術大学院大学 慶應義塾大学先端生命科学研究所 慶應義塾大学大学院 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	金子 聡子 池村淑道 白井剛 高橋健一 大島一彦 依田隆夫 塩生真史 阿部貴志 齊藤美保子 上原啓史 井口八郎 米森ゆかり 清水信義 金谷重彦 木ノ内誠 平賀壯太 武藤あきら 和田千恵子 山田優子 駒野照弥 香川弘昭 大橋武久 金井昭夫 菅原潤一 ◎○高木 利久
4. プロジェクトの総合的推進		

注1. ◎:課題代表者、○:サブテーマ代表者

注2. 本業務に携わっている方は、全て記入。

整備実績一覧

(1) 保有データ情報

(1-1) データの種類

①生物種	外部 DB や文献から得たデータは全生物種に及ぶ その他に、グループ内に植物関連のゲノム配列データ、微生物の完全長+ドラフトゲノム(合計 1,353 種)および並びにメタゲノム解析で塩基配列が解読された多数の微生物種データ、ヒトの多型データを保有
②試料・ライブラリ等の種類、数	シアノバクテリア 12 種、根粒菌 11 種、ミヤコグサ 一般日本人 200 人の DNA
③測定方法	ゲノム塩基の配列決定、DNA アレイ法による SNP アレルタイプ決定、tRNA 遺伝子発見プログラムなど
④データの内容	塩基配列、塩基配列上の予測遺伝子の番地、配列類似等の解析情報、遺伝子機能予測結果 ゲノムワイド確定ハプロタイプ、専門家が精査した tRNA データなど
⑤その他、特記事項	

(1-2) データソース

①現在のデータ量	国内 DB カタログ：国内 423、海外 170 の DB 注釈データ 横断検索：検索対象 234 サイトの検索用インデックスデータ ISND 総覧と検索：国際 DNA 配列バンク ISND の 1 億 300 万レコードの構造化データ 統合 TV：動画コンテンツ 160 種類 生命科学学協会検索：国内 600 の学協会情報 学会要旨統合俯瞰システム：4 学会 15,729 要旨のテキストデータ 蛋白質核酸酵素全文検索：21 年間 2,027 記事のテキストデータ WINGpro：国内外 431 の DB データ Web リソースポータル：解析ツール総数 456、解析ワークフロー総数 29、著名な論文 3 報のワークフロー KazusaAnnotation：7 生物種分のゲノム配列に対する、遺伝子注釈情報 115,968 件 JAGQC：2 x 10 ⁸ 個の SNP ジェノタイプ tRNADB-CE：18 万レコード
②データ区分	■自前 ■第三者 ■文献データ ■計算結果等の二次データ ☐ その他（下欄に詳細を記述）

③将来の増加の見込み	開発の進行および環境の整備に伴い、随時増加の見込み
④権利関係	所有者（ 基本的に自前データないし使用許諾を受けたもの ） 公開（ ☒ 可 ☐ 否 ☐ その他 []）
⑤その他、特記事項	公開データについては、下記の（２）、（３）、（４）項を参照願います。

（１－３）データの管理状況

①更新頻度等の管理状況、体制	国内 DB カタログや ISND 全文検索は日々更新、その他は随時更新 DB 毎に担当者がアサインして管理
②その他、特記事項	

（１－４）データベース関係

①DB 管理者数	中核機関グループ全体で延べ 19 名
②キュレータ・アナデータ数	中核機関グループ全体で延べ 15 名
③データ構造	関係データベース
④DB 管理ソフト	MySQL、Postgres、sqlite
⑤サーバの OS	Linux、Solaris、Windos XP、MacOS
⑥サーバ規模	ノード数：68、コア数：478
⑦DB へのアクセス数	本プロジェクトの成果サイトへは、平成 20 年 4 月～平成 21 年 3 月に 293,000 件
⑧独立 IP 数	上と同じく、のべ 161,000 件(月毎集計の総計)
⑨その他、特記事項	保有 DB の詳細は下記[2]の項を参照ください。

（２）データ（又は DB）の連結、統合化整備

通番	データ（又は DB）の名称	公開／未公開	概要（データの種別（生物種）・数量（kB 等）、本プロジェクトで実施した特徴点、進捗状況、今後の計画・課題などを簡潔にわかりやすく記述）
1	生命科学データベースカタログ http://lifesciencedb.jp/lldb.cgi?pg=1	公開	生命科学分野のデータベースカタログ（注釈及びメタデータ）で、データベースサイトのトップページをサムネイルで表示。データベースの稼働状況をモニターして表示。利用者からのコメントを受け付ける。掲載データベース数を増やすことが課題で、海外 DB 数の増加と記載内容の充実化を図ることが必要。その他の課題はデータベースの ID の統一化など、他のデータベースに関する情報との対応を持たせること。平成 21 年 2 月にインタフェースを横断検索、学協会カタログ、生物アイコン間で統一させた。英語版サービスも公開した。

			登録 DB 数：593 件 (H21.04 現在) 開始年月：平成 20 年 2 月。平成 21 年 3 月の訪問者数：251/訪問回数：314/ページビュー：603。
2	生命科学データベース横断検索 http://lifesciencedb.jp/dbsearch/	公開	生命科学分野のデータベースや文献の横断検索サービスで、生命科学分野の国内外のデータベースと文献を網羅的に検索。検索対象とするデータベースに優先度を設け、H19 年度中に掲載するデータベース、インデックス作成等掲載の準備をするデータベース、掲載が困難と予想されるデータベースに分類し計画に従って実施した。今後は、検索対象データベースをさらに広げつつ、主要な DB とそうでないアーカイブ的な DB に対する検索を分離する、またはユーザの対象とする生物種の検索を分離するなどしてユーザの利便性を図る。検索速度並びに検索精度の向上を図る。課題は、対象データベースが増加するに従って、維持管理のコストが高くなることが懸念される。検索対象 DB 数：234 データベース (H21.04 現在) 開始年月：平成 20 年 3 月。平成 21 年 3 月の訪問者数：8485/訪問回数：14346/ページビュー：23520。平成 21 年 4 月に対象データベースを増やし、新しいインタフェイスを公開。
3	DNA データベース総覧と検索 http://lifesciencedb.jp/ddbj/	公開	巨大な国際 DNA バンクを高速に検索し生物種、分子種、プロジェクトに分類して表示するシステム。従来巨大な計算機を必要としていたが、バンク内部を詳細に検討し 1 億 300 万レコードを 58 万プロジェクトに縮退させ実質上、計算機一台での検索を可能にした。さらに、ある研究機関から発表されたある生物のゲノムデータを一括してダウンロードするというような研究単位でのデータ取得を実現した。BLAST による配列検索、核酸関連の特許公報へのダイレクトリンク、レコードの時系列展開も可能である。英語版サービスサイトの整備も完了。今後は、データベース受入れ、横断検索など他のサービスとの統合、検索の一層の高速化、EST の臓器分類や発現データとの統合を図る。検索対象：1 億 300 万レコード。開始年月：平成 19 年 3 月。平成 21 年 3 月の訪問者数：3546/訪問回数：4233/ページビュー：7834/ダウンロード：2.43Gbyte。
4	統合 TV http://togotv.dbcls.jp/	公開	データベースやウェブツールの使い方などを動画で発信するウェブサイト。動画コンテンツ数は 150 に達した (平成 21 年 4 月現在)。開始年月：平成 19 年 7 月。平成 21 年 3 月の訪問者数：2419/訪問回数：8091/ページビュー：22720/ダウンロード：95.86Gbyte。
5	生命科学学協会カタログ http://lifesciencedb.jp/lbdb.cgi?gg=tool_academy	公開	生命科学系の学協会のデータベースと検索サービス。国内の生命科学系学協会を網羅したデータベース。各学会の詳細情報やウェブサイト、学会誌などの公開情報を掲載。H19 年度は日本学術会議の資料等をもとに生命科学系の学協会をリストアップしそれぞれの情報を調査、DB カタログと同じユーザインタフェイスを利用して公開した。今後は、内容の充実化と、各学協会のウェブサイトに対するメタ検索の実施を行う。可能であれば公開している文献情報の検索も一括して行えるようにする。掲載学協会数：600 学会 開始年月：平成 20 年 4 月。平成 21 年 3 月の訪問者数：45/訪問回数：52/ページビュー：64。
6	学会要旨統合検索 http://lifesciencedb.jp/lbdb	公開	学会要旨集の検索サービス。著者名や施設名の表記揺れを吸収する辞書を搭載し、研究者の研究履歴をたどることが可能で、キーワードのトレンドを表示させることもできる。要旨集

	.cgi?gg=tool_tproc		の提供を受けた遺伝学会と進化学会については全文検索全文表示を行っている。非公開の分子生物学会については、索引に対する検索のみ行っており要旨本文は表示しない。今年度は生物物理学会から要旨データの提供を受けた。さらに農芸化学会からも要旨データの提供を受けることが決まっており、次年度追加予定。学会要旨は研究において大変有用な情報資源であり、JSTAGE など公的な仕組みを利用して電子化されているものの、会員以外非公開の場合も多く、オープン化に向けて学会への働きかけが必要であるという課題については根本的な解決はできていない。検索対象：4 学会 15729 要旨。開始年月：平成 20 年 4 月。平成 21 年 3 月の訪問者数：62/訪問回数：68/ページビュー：84。
7	蛋白質核酸酵素全文検索 http://lifesciencedb.jp/pne/	公開 (更新分一部未公開)	蛋白質核酸酵素バックナンバーの全文検索サービス。国内で出版社により発行されている総説誌、研究者にとって重要な研究発表の場であるとともに、無くてはならない良質な情報源でもある。共立出版の協力により蛋白質核酸酵素のバックナンバー電子化と公開が可能となり、すでに 1985～2005 年分を公開中。毎年、1 年分の全文 PDF 公開について追加することにしており、新たに 2006 年分のデータを追加(平成 21 年 4 月公開)、2008 年分までのデータも全文検索可能(ただし、表示は要旨のみ) (同 5 月公開予定) となっている。今後の計画は関連論文、書籍などの検索を可能にすることが必要である。著作権や知財に詳しい弁護士との協議により、共立出版との契約等も円滑に行われている。収録論文：1985 年～2006 年(全文表示分)～2008 年(要旨表示分)。開始年月：平成 20 年 6 月。平成 21 年 3 月の訪問者数：1248/訪問回数：1807/ページビュー：4320。
8	MotDB http://MotDB.dbcls.jp/	公開	DBCLS を担っていく人材であるアノデータ、キュレータ、システムデータベース管理者向けの教材を提供するサイト (MotDB は Master of the database の略)。中核機関で独自に作成した教材 PDF ファイルもダウンロード可能。講習会開催のたびに、情報が追加されている。開始年月：平成 19 年 3 月。平成 21 年 3 月の訪問者数：636/訪問回数：1168/ページビュー：3930/ダウンロード：380.53Mbyte。
9	遺伝子発現バンク(GEO)目次 http://okubolab.genes.nig.ac.jp/bodymap_i/	公開	ヒト遺伝子の解剖学的な発現パターンデータの統合サイトを構築した。発現パターンは、測定法毎に異なる場合があることが知られており、ここではできるだけ客観的な発現パターンの解釈を可能にするために、5 種類の発現データ、即ち iAFLP、GeneChip、EST、NCBI の SAGEmap によるタグマップ、SAGE データの独自タグマップに基づく発現データを表示可能にした。平成 21 年 3 月の訪問者数：539/訪問回数：703/ページビュー：2480/ダウンロード：230.66Mbyte。
10	生命科学系データベースアーカイブサービス http://dbarchive.lifesciencedb.jp/	公開	アーカイブデータベース一覧を提供するサービス。各登録データベースに対して作成したデータベースとデータのメタデータ、利用許諾と共に、アーカイブデータのダウンロードが可能である。データの一括ダウンロードと共に、検索して絞り込んだデータのダウンロード機能も備えている。また、データベースの寄託を公募するためのページを作成し、その中で「寄託データベース利用許諾作成マニュアル」、「データベース利用許諾用語定義解説」、「データベース公開に関する覚書書式」の作成を行い公開している。アーカイブデータベース一覧に

			は、現在 DBCLS で作成されたコンテンツが 4 件登録されている。開始年月：平成 21 年 3 月。
11	ゲノム・ポストゲノム主要プロジェクト一覧 http://togodb.dbcls.jp/lsdb_project	公開	日本国内のゲノム・ポストゲノム関連の主要なプロジェクトに関して詳細な情報を収集しまとめたもの。1990 年代以降の 30 以上にもものぼる 4 省（文科省、経産省、厚労省、農林水産省）のゲノム・ポストゲノムに関連する主要プロジェクトについて、プロジェクトの期間、概要、予算、代表研究者、公開の状況、成果、文献等に関する情報が閲覧できる。開始年月：平成 20 年 3 月。平成 21 年 3 月に、キーワード検索等が可能な新しいインタフェースに更新した。
12	日本の生命科学データベース政策 http://lifesciencedb.jp/sciencenpolicy/	限定公開	これまでにライフサイエンスのデータベースに関わる検討がいろいろなところで行われており、一元的に俯瞰できる状態ではなかったため、本プロジェクトに関与しそうな報告書、会議資料（ウェブページ情報等含む）等を 1 箇所に集約し、時系列に整理した。 開始年月：平成 21 年 3 月。4 月からは外部公開予定。
13	WINGpro http://wingpro.lifesciencedb.org/ http://wingpro.lifesciencedb.jp	公開	ライフサイエンス分野の内外のデータベースの情報を収集、整理、分類、431DB を収録。ディレクトリからは、データベースの構築法による分類と生物種および対象による分類でデータベースを一覧可能。平成 19 年 3 月 30 日正式公開。平成 19 年 7 月 2 日よりユーザが記事を投稿および編集できる機能を公開。全文検索、分類一覧が可能。入力方式、一括データ取得、tips など記述。ユーザによる登録を促すことが課題。分類一覧の自動作成を開発。利用状況（平均訪問者数 2,217 6 月-12 月の平均値）、ページビュー：165,850（平成 20 年度）。
14	Web リソースポータルサイト for バイオインフォマティクス http://tools.lifesciencedb.jp/	公開	実験データの解析や公的データの加工に使用する解析ツールや環境を案内するサイト。ベンチメソッドオントロジーの一部を利用。解析ツール総数 456、解析ワークフロー総数 29、著名な論文 3 報のワークフローを収録、一括ダウンロードも可能。平成 19 年 3 月 30 日正式公開。教科書に近いサービスであるので、多くの研究者が行う実験フローをいち早く提供するなど工夫が必要。利用状況（平均訪問者数 122、ページビュー：3,720（平成 20 年度））。
15	Jabion Genome Viewer: HAL http://www.biportal.jp/genome/	公開	HAL（生物種 human）（ヒトゲノムに潜む遺伝子の位置と構造、その機能などの注釈情報を提供。）を既存のゲノムビューアである Jabion Genome Viewer に受け入れたテストサイト。H20.1.7 公開。ヒトゲノムに潜む遺伝子を発見する信頼性の高いコンピュータアルゴリズムにより予測した遺伝子数多数を公開。
16	KazusaAnnotation http://a.kazusa.or.jp/	公開	高度情報集積型データベースのソーシャルブックマーク型サービスで、独自開発のソーシャルブックマークによるオープンアノテーションを可能にした遺伝子情報統合データベースゲノムデータベースに対するマニュアルキュレーションによる情報抽出を行い、115,968 件の情報を蓄積し、公開している。今後は、改善を加えながら運用し、情報の集積によるゲノムを基盤とした生物学研究情報の統合を図る。アクセス数 209,248（平成 20 年度）。
17	KazusaNavigation http://navi.kazusa.or.jp/	公開	高度情報集積型データベースのポータルサービスで、ソーシャルネットワーク型の Elgg をベースに開発。現在は β 版を運用しながら情報蓄積を続け、公開している。まだ利用者数、ア

			クセス数ともに少ないが、学会、研究会情報などを定期的に更新している。また、かずさ内部のキュレーション作業の情報共有などで有効に活用している。今後、さらに、研究者コミュニティに利用をよびかけ、情報交換の促進による知識の統合を図る。登録者数72名、アクセス数 143,738 (平成20年度、Wikiと共通)
18	KazusaWiki http://wiki.kazusa.or.jp/	公開	高度情報集積型データベースのデータとりまとめサイトで、Wiki タイプのデータ共有サイトである MediaWiki をベースに開発。現在はβ版を運用しながら情報蓄積を続け、公開している。ワークショップの告知、既存データベースの受入などにも活用している。今後は改善を加えながら運用し、研究者による情報とりまとめサイトとして発展させていき、研究情報の統合を図る。提供者側からの情報入力もすすめていく予定である。登録者数、アクセス数は上と同じ。
19	JAGQC	限定 開示	我国で行われている大規模なゲノムワイド関連解析研究でのデータを第三者としてクオリティーコントロール (QC) を行い、開示することにより、これらの研究の透明性、客観性を確保することを目標としたデータベースである。現在は標準日本人 200 人の 100 万 SNP のジェノタイプデータをもとに模擬関連解析を行い、この QC データ及び解析過程を完全に透明な形で提示している。個人情報間接的に含むことから、アクセスに関する委員会の承認のもとに限定開示するように運用されている。今後実際の大規模関連解析の外部 QC データベースとして拡張予定である。
20	「健康への貢献遺伝子データベース」 「持続可能型社会への貢献 遺伝子データベース」 http://dbcls.nagahama-i-bio.ac.jp/	公開	人類の健康に役立つ遺伝子を学生が各自で想定して、世界で進行している環境生物試料に関する大規模シーケンシング (メタゲノム解析) で得られた大量な塩基配列を対象に、公的データベースにおいて未だ機能が特定されていない 1,700 万件以上の遺伝子候補に着目して、想定した遺伝子の探索を学生実習として行った。他大学からの探索結果も、相互に情報交換を持ちながら追加し、集合知の形成を目標に 学生等の個人名前入りで、2007 年 6 月から公開している。 自習可能なテキストを作成し、公開も行っている。月平均アクセス数は、24,000 件である。登録データの精査を行うために、シニア世代の専門家のキュレーション体制の拡充を図りつつ、高品質なデータの収集を目指す。
21	「エキスパートがキュレートした tRNA データベース」 http://trna.nagahama-i-bio.ac.jp/cgi-bin/trnadb/index.cgi	公開	微生物の完全長、ドラフトゲノム(合計 1,353 種)、メタゲノム解析由来塩基配列(1,700 万件)を対象に、3 種の予測プログラムを用いて網羅的な tRNA 遺伝子探索を行い、結果に差異が生じた場合には専門家による精査並びにキュレーションを行い、173,365 件の tRNA 遺伝子情報 (生物種・アミノ酸タイプ・アンチコドン・位置・塩基配列)を 2008 年 7 月より提供している。月平均アクセス数は、63,500 件である。今後は、ウイルス・オルガネラ・真核生物を対象にデータ登録を行ってゆく予定である。

(3) DB基盤システム、ツール等開発成果物の整備 (※試験的、限定的公開済みのものも含む。)

通番	DB基盤システム、ツール等の名称	公開／未公開	概要 (主な機能・特徴点、進捗状況、今後の計画などを簡潔にわかりやすく記述)
1	遺伝子名称シソーラス http://lifesciencedb.mext.go.jp/result/tech.html#id2-1-1	公開	分子情報から医学文献までの遺伝子や蛋白質名称の正確な同定のための遺伝子固有名称 (「遺伝子名」と一般名称 (「ファミリー名」) の辞書データの構築を目的として、様々なデータベースで利用されている名称の収集と専門的キュレータによる編集を行い、遺伝子を持つ多様な名称の関係を明示した遺伝子名称シソーラスを開発した。本シソーラスは、ヒトをはじめ9種類の生物をカバーしている。
2	生物学名羅日英対応辞書 http://lifesciencedb.mext.go.jp/result/tech.html#id2-1-1	公開	研究分野でよく使われる生物種の基準として、学名に日本語一般名を対応させた生物学名日本語一般名対応辞書を開発した。対応付けは、塩基配列データベース (DDBJ) の登録エントリ数が多い生物種から順番に行った。また、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用いた。今年度の追加作業の結果、登録データ数は昨年度末の 14,028 種から 123,057 種へと増加した。
3	施設名称辞書 http://lifesciencedb.mext.go.jp/result/tech.html#id2-1-1 (http://lifesciencedb.jp:32888/agency/agency.cgi)	公開 (内部公開)	日本のライフサイエンス研究を俯瞰するための重要な情報源として、各種関連学会の抄録などの報告文書があるが、これらをデータベース化する際に問題となるのが、例えば、大阪大、阪大、大阪大学大学院などといった、施設名称研究室名称の表記ゆれである。これに対応するために、施設名称辞書を開発した。すでに公開しているものは 1627 の名称についてまとめたものであるが、今年度はその範囲を 15,000 件以上の研究機関・研究グループに広げ、ホームページ掲載情報や学会要旨データベースより収集した機関名別名情報を格納し、機関名の表記ゆれに対応した検索を可能とした。辞書そのものについては内部公開で試験運用中である。
4	動物解剖学自動分類タガー http://lifesciencedb.mext.go.jp/result/tech.html#id2-1-1	公開	動物の臓器、組織を、大きく 10 のグループに分類 (大分類) し、さらにそれぞれのグループを細かく分類、合計 40 の小分類グループに分類する。基本的には、与えられた解剖用語に対して、解剖用語辞書、病理関連語彙の分類辞書、形容詞の解剖用語辞書、一般的な臓器名称の分類辞書、の 4 種類の辞書を順番に検索し、上記のカテゴリに分類する。
5	植物解剖学自動分類タガー http://lifesciencedb.mext.go.jp/result/tech.html#id2-1-1	公開	植物 (維管束植物) の部位、組織を大きく 6 のグループに分類し、さらにそれぞれのグループを細かく分類、合計 11 の小分類グループに分類する。分類における検索対象としては、生物種に合わせて、種子を持たない維管束植物の解剖用語辞書、イネ科の解剖用語辞書、その他被子植物の解剖用語辞書、裸子植物の解剖用語辞書、トウモロコシ属の解剖用語辞書、フウチョウソウ目の解剖用語辞書を用いる。
6	都市名国名自動検出タガー http://lifesciencedb.mext.go.jp/result/tech.html#id2-1-1	公開	論文やデータベースレコードに見られる国の名称の未記載や国の名称にみられる表記ゆれを吸収することを目的に、国別に分類するための辞書を開発した。

7	生物アイコン http://lifesciencedb.jp/lbdb.cgi?gg=resource_icon	公開	ユーザが自身の HP など自由に利用できるような提供されている生物のアイコン集。目的とするイラストをキーワード検索でき、生物に関する外部情報のリンクも付いている。ユーザはイラストに関するコメントを書き込むことができ、さらには同じ条件で配布可能なイラスト等を投稿することもできる。アイコン数：201。開始年月：平成 20 年 6 月。平成 20 年 12 月にアイコン数を増やし、インタフェイスを更新。英語版も作製した。平成 21 年 3 月の訪問者数：210/訪問回数：226/ページビュー：286/ダウンロード：8.8Mbyte。
8	BodyParts3D http://lifesciencedb.jp/ag/pgm	公開	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辞書。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、改変にロバスタな表現が可能。今後も用語の追加を続ける。今年度は昨年度のフェーズ 1 データより精密なフェーズ 2 データの作成を実施し、現在、フェーズ 1(75 種)、フェーズ 2(567 種、今後も増加予定) のデータが格納されている。アクセス情報はアナトモグラフィと一緒にカウントしているため下記参照。
9	アナトモグラフィ http://lifesciencedb.jp/ag/ (http://61.197.171.146/project/ag_annotation/ag_annotation.cgi 、現在 DBCLS 内サーバへの移行準備中)	公開 (内部公開)	ポリゴンマン辞書から部品を選択して自由に人体の部分のモデル図が書ける。視点固定の図譜や 3D アトラスと異なり透過度や着色、視点は自由に選択可能。今後は、解剖学教育や医学用語辞書の助けになる人体版グーグルマップを目指す。今年度はすでに公開したバージョンに加え、拡大・縮小、平行移動、切断面の生成、イメージ内へのマーカーピンの設置など、自由度が高くかつ高速な図形処理が可能なバージョンを試験的に内部公開し、検証中である。平成 21 年 3 月の訪問者数：900/訪問回数：1275/ページビュー：27093/ダウンロード：12.94XGbyte。
10	TogoDB http://togodb.dbcls.jp/	試験的公開	データベース受け入れシステム。さまざまな形式のさまざまな種類のデータを簡単に読み込み、自動的に共通のウェブ検索インタフェイスを生成する。また、自動的にウェブサービス(SOAP/WSDL API)を生成する。これらの機能により、容易に維持困難な DB を移設できる。また、本ツールのダウンロードにより、誰でも DB を構築可能である。H20 年度新たに検索結果のダウンロード機能を追加した。平成 21 年 3 月の訪問者数：525/訪問回数：993/ページビュー：3258/ダウンロード：375.34Mbyte。
11	TogoWS http://togows.dbcls.jp/	公開	国内外のウェブサービスを透過的に統合するシステム。国内各サービスの稼働状況を継続的に監視し、サービスの連携に必要なデータ形式変換機能等を提供する。これにより、ユーザは統一的命名規則と統一的データ構造で国内外のウェブサービスにアクセスすることができるため、相互運用が容易となり、ワークフローの作成・蓄積を促進する。H20 年度キーワード検索の書式を統一した。開始年月：平成 20 年 4 月。平成 21 年 3 月の訪問者数：263/訪問回数：1625/ページビュー：21017/ダウンロード：743.08Mbyte。
12	OpenID 認証システム http://openid.dbcls.jp/	公開	一つの ID で複数のサイトを認証できるアカウントシステム。このシステムを利用することで、ユーザはひとつの ID とひとつのパスワードを保持すればよい。これにより、各サイトで認証サービスを用意する必要がなく、サイト間のユーザ情報の集約が容易になる。開始年

			月：平成 20 年 4 月。平成 21 年 3 月の訪問者数：115/訪問回数：306/ページビュー：792。
13	OReFiL http://orefil.dbcls.jp/	公開	オンライン上に存在する多数の生命科学系の資源(データベースやソフトウェアなど)を効率的に見つけるための検索システム。信頼性や概要を容易に取得するために、資源の所在だけでなく、その資源について書かれている論文情報や、他の利用者の評価を閲覧できる。アカウントを取得することで、自ら評価を行うことも可能である。また、検索に関連する機能を全てウェブサービスとして提供しているため、利用者のプログラムから OReFiL を利用できる。英語版ヘルプなどを整え、環境を整備した。開始年月：平成 19 年 8 月。平成 21 年 3 月の訪問者数：366/訪問回数：569/ページビュー：1891。
14	Allie http://allie.dbcls.jp/	公開	医学生物系論文書誌情報データベース MEDLINE を対象とし、出現する略字とその正規系のペアを検索するシステム。生命科学系の論文では非常に多くの略字が使われており、同じ表記でも全く違う意味を示していることが少なくない。そこで、利用者の興味のある略字を検索語として入力することで、その使われ方を一覧表示すると共に、論文の発表年を提示する。また、検索された各略字について、その意味で使われている論文中で共起する他の略字も提示する。検索に関連する機能を全てウェブサービスとして提供しているため、利用者のプログラムから Allie を利用できる。英語版ヘルプなどを整え、環境を整備した。開始年月：平成 20 年 4 月。平成 21 年 3 月の訪問者数：207/訪問回数：256/ページビュー：683。
15	DBCLS Galaxy http://galaxy.dbcls.jp/	限定 公開	生命科学データに特化した、ウェブベースの対話的ツール組み合わせインタフェース。マウス操作と簡単な入力で、プログラムを書くことなく、途中結果を確認しながら、用意されたツールを対話的に組み合わせて使うことができる。DBCLS Galaxy では、TogoWS やテキストマイニングツールなど、DBCLS で開発されたツール群を組み込み、利用できるようにした。ログインは DBCLS OpenID で可能である。また、メニューの国際化対応へ協力し、メニューの日本語化を行い、DBCLS Galaxy のみならず、元の Galaxy にも反映させた。
16	ScrapParty	公開	ブラウザのプラグインによるキュレータ支援用アプリケーション。オンラインで文献やウェブサイトを閲覧しながら、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 化するツール。H18 年度アノテーションの調査、試験に利用した。下記 Wired-Marker が後継アプリケーションであり、現在 ScrapParty 名での開発は終了。
17	Wired-Marker http://www.wired-marker.org	公開	情報共有ツール、アノテーション支援ツール。Firefox ブラウザのプラグインとして働く。データベース構築、アノテーションを支援するブラウザ用のプラグイン。オンラインで閲覧したページ、論文のテキストや図をマークしブックマークすることが可能。問の際にも同じ箇所がハイライトされアノテーションのエビデンスとして詳細な情報が残せる。またフォルダを自由に階層化して情報を整理することが可能。新たにキャッシュ機能が加わり、ウェブページの書き換えが起きても過去のデータを参照できるようになった。 公開日：H19.12.19 総ダウンロード数：258,627 件。
18	メタデータ要素レポジトリ	公開	メタデータ要素の検索、検索結果の比較、収録メタデータのメタデータ要素の一覧表示が可

	(MDeR) http://mdcr.jst.go.jp/		能。ISO/IEC 11179 Part3(Registry metamodel and basic attributes)に準拠した形で、国際標準3種、データベース3種のメタデータ要素を収録。国際標準2種の調査終了し、搭載予定。収録内容の精査と充実によりデータ収集・データベース構築時のデータ形式等の決定に役立つサービスを目指す。公開日：平成20年3月26日 平成21年3月の訪問者数：47 ページビュー：318。
19	CBRC タンパク質構造情報ワークフロー	限定公開	構造に関する予測プログラム等を Grid により効率的に分散処理し、結果を実験研究者等に提供する(プロトタイプ)。平成20年8月末よりプロジェクト関係者に限定し公開開始。後、下記20のタンパク質アノテーションワークフローに統合された。
20	CBRC タンパク質アノテーションワークフロー http://tgweb01.cbrc.jp/wf/PAW_About.html?Eng	公開	立体構造未知のタンパク質に関し、構造及び機能のヒントとなる情報を幅広く実験研究者等に提供する。二次構造予測、埋れ残基予測、フォールド認識、ディスオーダー予測、膜タンパク質オールベータ・ベータシート予測、細胞内局在予測を Grid にて分散処理を行う一方、データベース検索及び疎水性予測の実行を他のサーバへ依頼し、結果を取得後、全ての結果をユーザが解析し易いよう配置し出力する。平成20年8月末より一般公開開始。
21	CBRC タンパク質比較情報ワークフロー http://tgweb01.cbrc.jp/wf/comp_prot_info_about.html	公開	相同なタンパク質を比較することで保存部位等構造上重要な部位を表示し、実験研究者等に提供する。マルチプルアラインメントを用い保存性が高い残基及び二次構造を表示する。平成21年3月末より一般公開開始。
22	genoDive Pro / Eu	限定公開	DAS 規格に準拠したゲノム情報を統合的に閲覧することが可能な DAS サーバクライアント。現在は試験配布中、ソースコードも含め完全公開し配布を行う予定。
23	遺伝子名称変換プログラム	限定公開	植物遺伝子名や遺伝子 ID の名称の食い違いを解決するための名前変換サービス。現在は試験運用中、テスト終了後、完全公開の予定。
24	なんでも DAS	限定公開	任意のウェブページに含まれる遺伝子名を抽出して、GFF 形式に変換するサーバ。このサーバを活用することで、インターネット上のあらゆる遺伝子関連情報をゲノム上にマッピングし、情報の統合的な閲覧を可能とする目的で開発した。現在は試験運用中、テスト終了後、完全公開の予定。
25	MiGAP	限定公開	技術革新により大量に生産されるようになった微生物ゲノムの利用価値を高めるため、アノテーションを自動的に行うパイプラインを開発し、試験的にプロジェクト内部に公開、評価を受けた。評価の意見を反映した修正作業を進めており、平成21年5月には、一般公開の予定。利用のためには、現状も公開後も OpenID の取得が必要。開始年月：平成21年3月。

(4) その他の成果物 ((2)、(3) に該当しないもの)

通番	名称	公開／未公開	概要
1	統合 PJ 成果公開サイト http://lifesciencedb.jp	公開	統合 DB プロジェクトに関する中核機関の成果公開サイト。生命科学データベースカタログ、DB 横断検索、INSD 全文検索、統合 TV、OReFiL、生命科学学協会検索、学会要旨統合俯瞰システム、MotDB、LSDB ブログ等を掲載している。開始年月：平成 20 年 6 月。
2	統合 PJ 成果公開サイト(英語版) http://lifesciencedb.jp/?lng=en	公開	国外で本プロジェクトの紹介をする機会も増えてきたことから、英語での利用が可能なサービスを集めて、英語版のポータルサイトを公開した。現在紹介されているデータ・サービスは、ゲノム・ポストゲノム主要プロジェクト一覧、生命科学データベースカタログ、生物アイコン、DNA データベース総覧と検索、遺伝子発現バンク(GEO)目次、アナトモグラフィ/BodyParts3D、Wired-Marker、TogoWS, togoTV。そのほか、プロジェクト紹介ページ、参画機関の成果ページ(英語版があるもの)へのリンクをはった。開始年月：平成 21 年 3 月。
3	事業広報サイト http://lifesciencedb.mext.go.jp/	公開	統合データベースプロジェクトの紹介サイト
4	ライフサイエンスの広場 http://www.lifescience.mext.go.jp/index.html	公開	文科省のライフサイエンス政策等の紹介サイト
5	統合データベース間の連携調査 http://charles.kazusa.or.jp/lifesciencedb/	公開	代表的モデル研究植物である「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各 46 サイト、25 サイトの基本項目の調査と実験生物研究者 188 名のアンケートにより調査し、それぞれのデータベース間の連携のための現状の課題と将来の DB 統合にむけた課題の整理を行った。
6	検索アルゴリズムを含めた知識情報技術の動向調査 http://lifesciencedb.mext.go.jp/result/strategy.html#id1-4	公開	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た。
7	遺伝統計学分野における解析技術の基礎調査 http://lifesciencedb.mext.go.jp/result/strategy.html#id1-4	公開	遺伝統計学分野で用いられる連鎖解析、連鎖不平衡解析)、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、動作環境などを調査し、その評価を行った。
8	臨床情報・疾患健康情報の調査 http://lifesciencedb.mext.go	公開	わが国の代表的コホート研究に関し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た。

	.jp/result/strategy.html#id1-4		
9	ライフサイエンスプロジェクトの調査結果	試験的公開	この約 10 年間のライフサイエンス国家プロジェクトの事業内容や成果を明らかにする目的で、4 省庁のゲノム、ポストゲノム関連の主要プロジェクトの調査を行った。プロジェクト名、期間、予算、主要研究者、主要事業内容、公開状況、公開 DB、データダウンロードサイト、公開 HP、事業化関連、報告書、その他文献等に関する情報を情報ソースへのリンクをできるだけつけた形で整理した。
10	データベース受入に関するアンケート調査 http://lifesciencedb.jp/lbdb.cgi?pg=0&gg=ukeire_survey	試験的公開	受入による統合化の方針策定と受入対象の候補抽出を目的として、国内データベースを対象として、データベース内容と検索方法、経歴、維持・管理状況、提供可能性及び提供時の問題点などをアンケートにより調査した。送付した 104 件中、65 件の回答が得られ、半数以上の機関が、統合 DB 事業に対して全データもしくは一部データを開示できると回答した。結果の概要および詳細報告 (PDF) をウェブ上で公開した。
11	BioHackathon2008 ホームページ http://hackathon.dbcls.jp/	公開	分散環境のままで、各種 DB をアルタイムかつ統合的に利用できる環境を構築すること目的に開催した、国内外のライフサイエンス DB のウェブサービスにおけるデータ構造や命名規則を標準化するための BioHackathon2008 の活動内容と経過を紹介するサイト。2008 年 2 月 11 日から 15 日まで都内で開催。海外参加者 34 名、国内参加者 37 名。
12	BioHackathon2009 ホームページ http://hackathon2.dbcls.jp/wiki	公開	標準化技術および標準化を利用する技術をより安定的に使用可能にすることを目的として開催した BioHackathon2009 の活動内容を紹介するサイト。巨大なデータについてどうすべきか、セマンティックウェブ技術をどう利用すべきかなどの議論の報告が掲載されている。2009 年 3 月 16 日から 20 日まで、都内および沖縄で開催、海外参加者 30 名、国内参加者 28 名。
13	データベース高度利用者養成カリキュラム http://togodb.sel.is.ocha.ac.jp/	公開	データベース高度利用者養成において利用したカリキュラム、資料の公開。