

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)

19年度 研究成果報告書

平成20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 科学技術振興機構	黒田雅子
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 科学技術振興機構、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19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。

従って、本報告書の著作権は、文部科学省に帰属しており、本報告書の全部または一部の無断複製等の行為は、法律で認められたときを除き、著作権の侵害にあたるので、これらの利用行為を行うときは、文部科学省の承認手続きが必要です。

目 次

1. プロジェクトの目的	3
2. 戦略立案・実行評価（情報・システム研究機構）	3
2. 1 戦略立案・実行評価の実施計画	3
2. 2 戦略立案・実行評価の実施内容	4
3. 統合データベース開発	8
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）	8
3. 1. 1 共通基盤技術開発の実施計画	8
3. 1. 2 共通基盤技術開発の実施内容	9
(1) 共通基盤技術開発（情報・システム研究機構）	9
(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）	11
(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）	16
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）	19
3. 2. 1 ヒト統合 DB の開発・運用の実施計画	19
3. 2. 2 ヒト統合 DB の開発・運用の実施内容	19
(1) ヒト統合 DB の開発・運用（情報・システム研究機構）	19
(2) 多型知識表現技術の開発（九州大学）	23
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）	25
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	25
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	26
(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）	26
(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）	29
4. 統合データベース支援（情報・システム研究機構、科学技術振興機構）	31
4. 1 ポータル整備・運用、広報、普及啓発	31
4. 1. 1 ポータル整備・運用、広報、普及啓発の実施計画	31
4. 1. 2 ポータル整備・運用、広報、普及啓発の実施内容	32
(1) 統合 DB プロジェクトポータルサイトの整備（情報・システム研究機構）	32
(2) 統合 DB プロジェクトポータルサイトの運用（情報・システム研究機構）	37

(3) 広報・普及啓発 (情報・システム研究機構)	38
(4) ポータルサイト WINGpro の運用 (科学技術振興機構)	38
(5) 事業サイト等の運用と広報 (科学技術振興機構)	39
4. 2 データベース受入	40
4. 2. 1 データベース受入の実施計画	40
4. 2. 2 データベース受入の実施内容	40
(1) データベース受入 (情報・システム研究機構)	40
(2) データベース受入 (科学技術振興機構)	43
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)	45
4. 3. 1 人材育成の実施計画	45
4. 3. 2 人材育成の実施内容	47
(1) DB 構築者の養成 (東京大学)	47
(2) DB 高度利用者の養成 (お茶の水女子大学)	48
(3) アノテータ・キュレータの教育 (長浜バイオ大学)	48
5. プロジェクトの総合的推進 (情報・システム研究機構)	50
5. 1 プロジェクトの総合的推進の実施計画	50
5. 2 プロジェクトの総合的推進の実施内容	50
(1) 研究運営委員会及び研究運営委員会作業部会	50
(2) 分担機関、補完課題実施機関との連絡会	51
(3) 中核機関内の連携のための会合	52
(4) 外部有識者からの意見聴取	53
5. 3 その他セミナー等	53
(1) BioHackathon (情報・システム研究機構、生命情報工学研究センター) ..	53
(2) バイオインフォマティクスへの招待(お茶の水女子大学)	55
(3) アノテータ、キュレータ養成のための運営委員会 (長浜バイオ大学)	56
(4) 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、	57
科学技術振興機構)	
6. 成果のまとめ	58
7. 成果の外部への発表	59
8. 実施体制	59

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

ライフサイエンス、情報処理技術、ライフサイエンスDBの各分野の動向に即した我が国のDB整備戦略を立案するために、上記3分野の専門家、並びに関連府省の専門家、産業界の代表、分担機関の研究代表者を加えた研究運営委員会を組織した。また、研究運営委員会の下に研究運営委員会作業部会を組織し、個別具体的な課題について詳細な議論を行うこととした。また、上記委員会、作業部会に対して、3分野の動向に関する情報の網羅的な収集と分析を行い、戦略立案や実行評価を支援するためのチームを、大学利用機関法人 情報・システム機構（ROIS）内に設立したライフサイエンス統合データベースセンター（DBCLS）内に組織した。

このような体制の中で、本年度は特に、まずDB統合化のスキームの検討とこれに基づくサービスの実現方法を検討し、さらに統合化の年次計画を策定した。次に、これをベースに補完課題の公募方針およびこれに関連したデータベースの受入れ戦略を検討した。さらに、データベース提供を促進するための契約・著作権関係並びに個人情報関係の検討、日本語総説誌を中心とした日本語文献の公開のための方策の検討を行った。年度の後半には、本プロジェクト終了後の我が国のライフサイエンス統合DBの在り方に関する議論も行った。

統合化のスキームについては、主に検索のタイプから、DBカタログ、横断検索及び統合検索の三つのケースについて検討した。DBカタログはライフサイエンス分野のDBのメタデータを集め、これに注釈づけを施したDB情報に網羅的にアクセスできるものである。横断検索は、クローリングなどにより対象となるDBのインデックス情報を集め、これによりライフサイエンス分野のDBや文献の横断的なキーワード検索を実現するものである。また、統合検索は、外部にあるDBからAPIにより入手したデータ、あるいは統合データベースセンター内に受け入れたDBを処理して得たデータに対して検索を行うもので、横断検索より、より高度かつ詳細な情報を獲得できるものである(図2.1参照)。統合化の年次計画としては、DBカタログについては、平成19年度中に、我が国に存在するとされるゲノム、ポストゲノム関係の250のDBのほとんどを網羅することにした。また、横断検索については、平成19年度中に20程度の代表的なDBを掲載し、引き続き平成20年度中に我が国の関連DBの大部分と海外の主要DBを掲載することとした。さらに、統合検索については、国内DBの受入れと併せて、順次対象となるDB数を増やしていくこととした(図2.2参照)。

補完課題については、上記作業部会で公募内容の検討を行い、大型プロジェクトや大量データ産生機関を対象として、センターの標準化仕様に従い統合化するための諸作業を課題化すると結論が得られた。この内容をベースに文部科学省が公募を行い、審査の結果、理化学研究所、産業技術総合研究所糖鎖医学工学センター、国立遺伝学研究所および九州工業大学の提案が採択された。この結果、図2.3に示すように、最終的に統合データベースプロジェクトは、中核機関、三つの分担機関、及び四つの補完課題実施機関で行うこととなった。また、データベース受入については、ターゲットを絞ってパイロット的に受入れることになった(詳細は4. 2項に記載)。

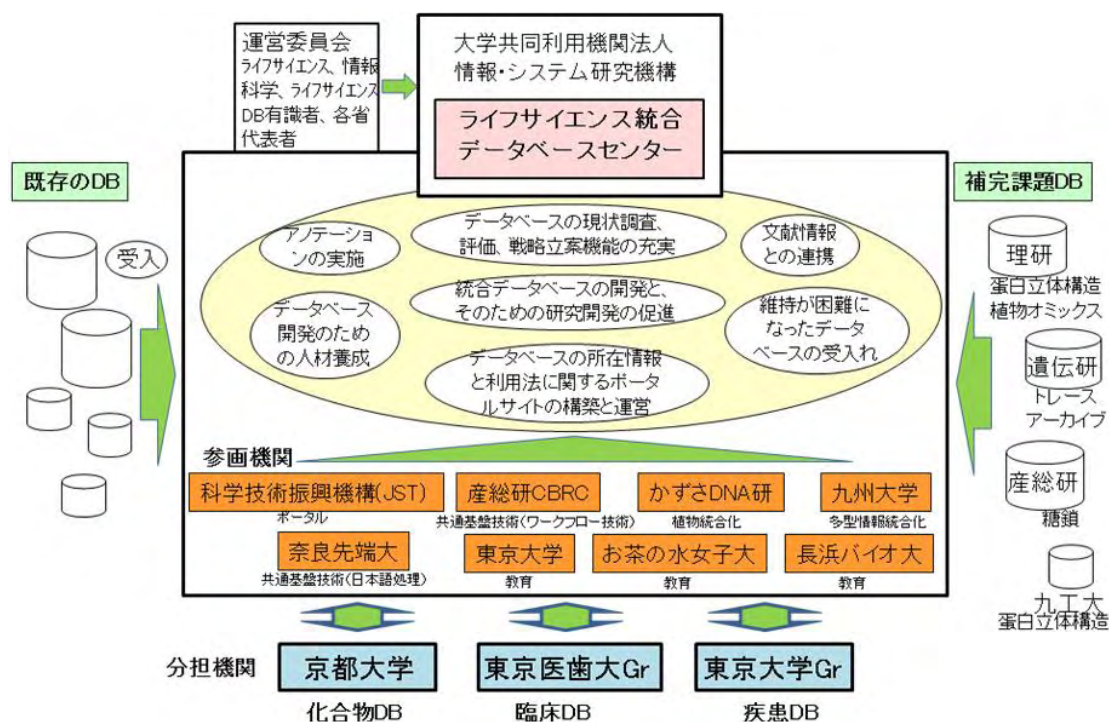
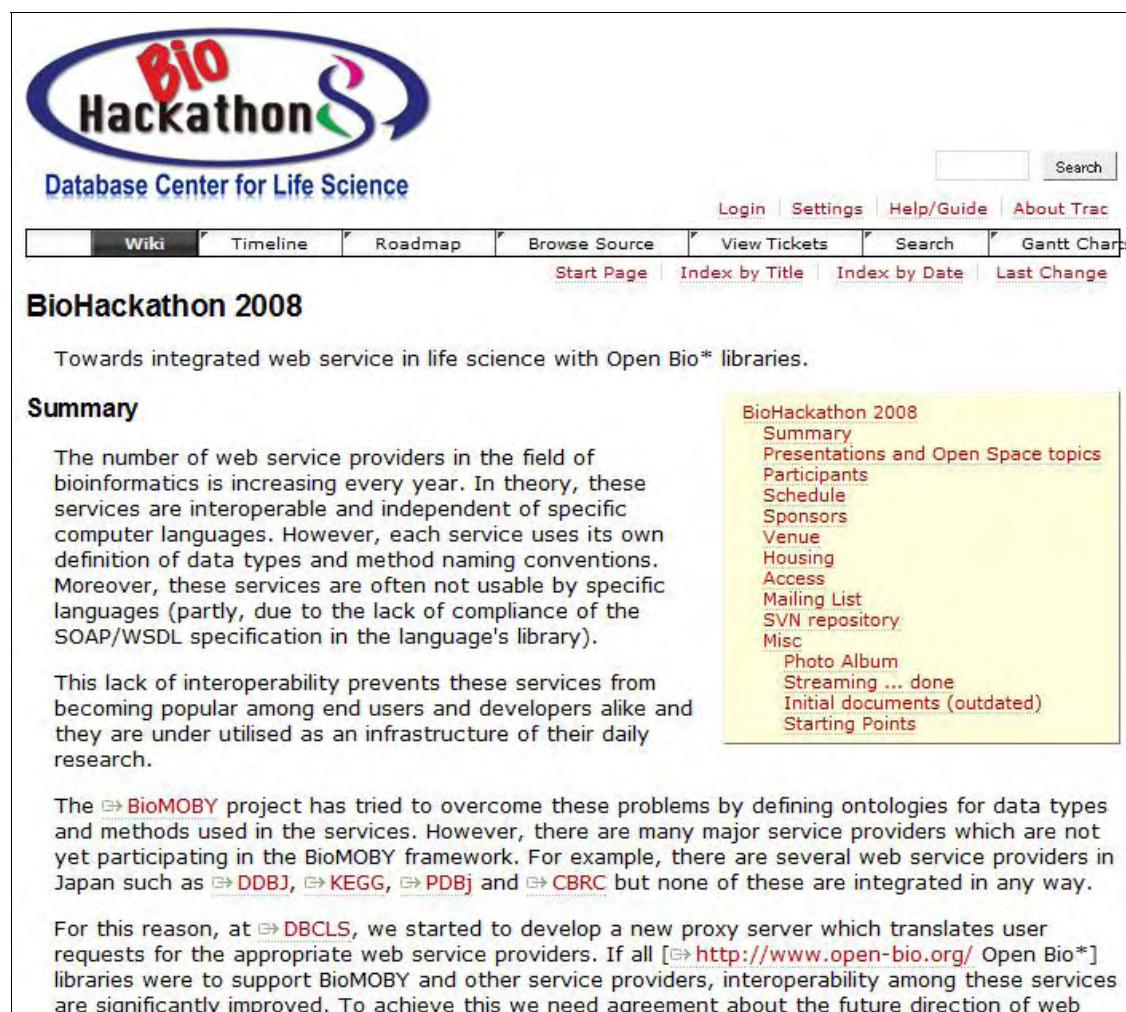


図2.3 統合データベースプロジェクトの体制図

契約・著作権並びに個人情報関係については、まず前者について、国家プロジェクトの契約内容及びその基礎となる法律に関して調査を行った。また後者に関しては、国内の医療情報を取り扱う代表的プロジェクトに関して、個人情報の取り扱いに関する現状調査を行った。こうした検討結果を踏まえ、本年度は特に、今後のライフサイエンス分野の国家プロジェクトからデータおよびデータベースの提供を受けるための方策について検討し、文部科学省との密接な連携の上で、プロジェクトの公募要領にデータおよびデータベース提供条項を盛り込むことができた。また、国内外データベースの利用条件、免責事項、プライバシーポリシー等の整理、分析に基づき、本プロジェクトが運営・管理するサイトのサイトポリシーを作成した。また、日本語文献の公開については、特に日本語総説誌の公開のためのモデルの検討を行い、こうした検討結果をベースに出版社と交渉し、日本語総

説誌公開ための契約に至った。また、日本語資料の公開の一環として総合科学技術会議、文科省等のライフサイエンス関連公開資料の収集も行った。なお、上記の法律、契約に関わる条項については、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。

また、国内外の研究機関等関係者との連絡調整の一環として、平成20年2月11日～15日の5日間、ウェブサービスの統合を目指した現状調査・情報交換および技術開発を目的として、産業技術総合研究所生命情報工学研究センターの協力のもと、国際ワークショップ BioHackathon 2008 を開催した。ここでは、主にライフサイエンスの主要なウェブサービスプロバイダ、クライアント開発者、オープンバイオ開発者などから参加者を募った。その結果、海外から34名、国内から37名の参加者があり、トピックごとに熱心な議論が行われた。主なトピックはオントロジー、データ交換の標準化、非同期通信、大規模データ、セキュリティ、オープンバイオ、RESTvsSOAP、ワークフロー、BioMobyなどであった。この議論の過程および成果はウェブ上で公開している (<http://hackathon.dbcls.jp/> ; 図2.4参照)。



BioHackathon 2008

Database Center for Life Science

Search

Login | Settings | Help/Guide | About Trac

Wiki | Timeline | Roadmap | Browse Source | View Tickets | Search | Gantt Chart

Start Page | Index by Title | Index by Date | Last Change

BioHackathon 2008

Towards integrated web service in life science with Open Bio* libraries.

Summary

The number of web service providers in the field of bioinformatics is increasing every year. In theory, these services are interoperable and independent of specific computer languages. However, each service uses its own definition of data types and method naming conventions. Moreover, these services are often not usable by specific languages (partly, due to the lack of compliance of the SOAP/WSDL specification in the language's library).

This lack of interoperability prevents these services from becoming popular among end users and developers alike and they are under utilised as an infrastructure of their daily research.

The [BioMOBY](#) project has tried to overcome these problems by defining ontologies for data types and methods used in the services. However, there are many major service providers which are not yet participating in the BioMOBY framework. For example, there are several web service providers in Japan such as [DDBJ](#), [KEGG](#), [PDBj](#) and [CBRC](#) but none of these are integrated in any way.

For this reason, at [DBCLS](#), we started to develop a new proxy server which translates user requests for the appropriate web service providers. If all <http://www.open-bio.org/> Open Bio* libraries were to support BioMOBY and other service providers, interoperability among these services are significantly improved. To achieve this we need agreement about the future direction of web

BioHackathon 2008

- Summary
- Presentations and Open Space topics
- Participants
- Schedule
- Sponsors
- Venue
- Housing
- Access
- Mailing List
- SVN repository
- Misc
- Photo Album
- Streaming ... done
- Initial documents (outdated)
- Starting Points

図 2.4 BioHackathon 2008 のホームページ

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）

文献（論文や教科書等）に書かれた高次の生命機能や実験データ（画像や表）、概念図で表現されるような遺伝子・蛋白質ネットワークなどを抽出し、分子データベースや他の知識ベースと連携した効率的な検索を可能にするため、知識表現技術および検索技術を開発する。平成19年度は、近年多くの生命科学系学術論文の全文が電子的に入手し易くなってきたことを踏まえ、当該分野における研究成果を端的に示す要素である、実験結果を示す図表とその脚注から知識を抽出し、索引付けを行う。そして、遺伝子・蛋白質名や、蛋白質間相互作用等の生物学的な現象を検索語として、それに関連する実験結果を出力するシステムのプロトタイプを実装する。また、マイクロアレイにより測定された発現プロファイルからのネットワーク推定プログラムが増えていることを踏まえ、推定されたネットワークから、それに関連する既存ネットワークを検索するための諸技術を開発し、プロトタイプを実装する。

（産業技術総合研究所生命情報工学研究センター）

ワークフロー技術を用いた統合 DB 環境の構築については、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 19 年度は、実施場所内の計算機環境で、統合 DB 環境のプロトタイプを構築する。具体的には、現在構築可能で有益なサービスを調査し、ワークフローを作成することにより、それらのサービスを実現する。また、それらの操作画面は、専門外の利用者でも操作できるよう配慮する。

（奈良先端科学技術大学院大学）

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、専門用語解析技術の開発および専門用語タグ付け手法の設計を行う。専門用語辞書システムの設計については、専門用語間の意味的類似性、同義性を判定する基礎技術を開発し、用語の意味関係を記述することのできる用語辞書システムの基本設計を行う。専門用語解析技術の開発については、専門分野のテキストには複雑な専門用語が数多く出現する。特に、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など、用語辞書の充実だけでは完全な用語抽出を行うことができない。これに対処する解析技術として、専門用語の並列構造解析技術の開発を目指す。専門用語タグ付け手法の設計については、専門用語解析の基礎データの蓄積を行うため、専門分野テキストに対して、用語の出現箇所の特定や用語の種類の分

類を行うため、タグ付けの表現法、および、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)

文献に書かれた高次の生命機能や実験データ、概念図で表現されるような遺伝子・蛋白質ネットワークなどを抽出し、分子データベースや他の知識ベースと連携した効率的な検索を可能にするための知識表現技術及び検索技術として、以下の技術を開発した。まず、近年多くの生命科学系学術論文の全文が電子的に入手し易くなってきたことを踏まえ、文献内の図表の注を索引付けするサブシステムを開発し、学会要旨統合検索システムに組み込んだ。さらに、遺伝子・蛋白質名や、蛋白質間相互作用等の生物学的な現象を検索語として、それに関連する実験結果を出力するシステムのプロトタイプを構築するための基本技術として、遺伝子名やたんぱく質名等の略称とその展開系のペアを MEDLINE から抽出してデータベース化し、検索するシステム Allie を構築した(図3.1参照)。これは、略語として同じ表記であっても、異なった展開系をもつ例が多数存在することを踏まえ、ペア毎に MEDLINE 文献データを分類し、発表順に表示するものである。さらに、オンライン資源検索システム ORe FiLをベースに、ユーザの利便性と検索精度を高めるための資源評価書き込み機能、資源評価閲覧機能などの追加を行った(図3.2参照)。

- ・ 生命科学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展開形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、利用者が興味のある略字を検索語として入力すると、その展開形を、おのおの出現している論文の題目と発表年とともに一覧表示する

- ・ さらに、検索された各略語につき、同じ展開形の意味で用いている論文群で共起する他の略字也表示

- ・ 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を利用可能

Allie
http://allie.dbcls.jp/

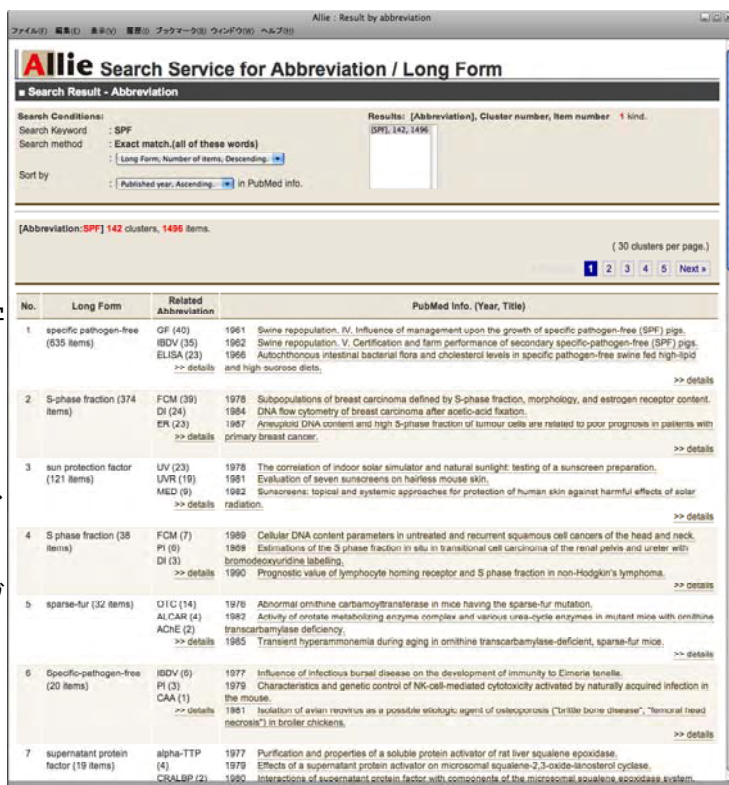


図3.1 Allie: 論文に出現する略字とその展開系を検索

・ 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可能

・ 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

・ 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 を利用可能

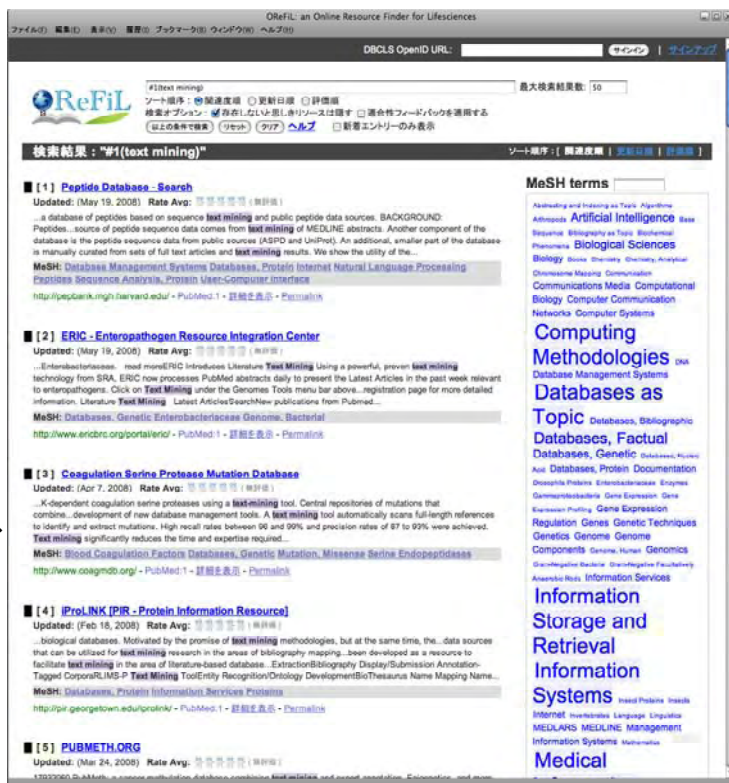


図3.2 OReFiL: オンライン上の資源を、文献情報と共に検索し、提示

また、マイクロアレイにより測定された発現プロファイルからのネットワーク推定プログラムが増えていることを踏まえ、推定されたネットワークから、それに関連する既存ネットワークを検索するための基本技術として、遺伝子の対応関係を視覚的に表示し、対応関係の編集を手助けする技術を開発し、プロトタイプIDEditorを実装した(図3.3参照)。遺伝子対応関係は読み込み時に自動的にクラスタリングされ、また、削除の候補としてクラスターの外れ値となる対応関係が示される。これを用いることで、ユーザは対応関係を視覚的に確認しながら編集を進めることが可能となる。

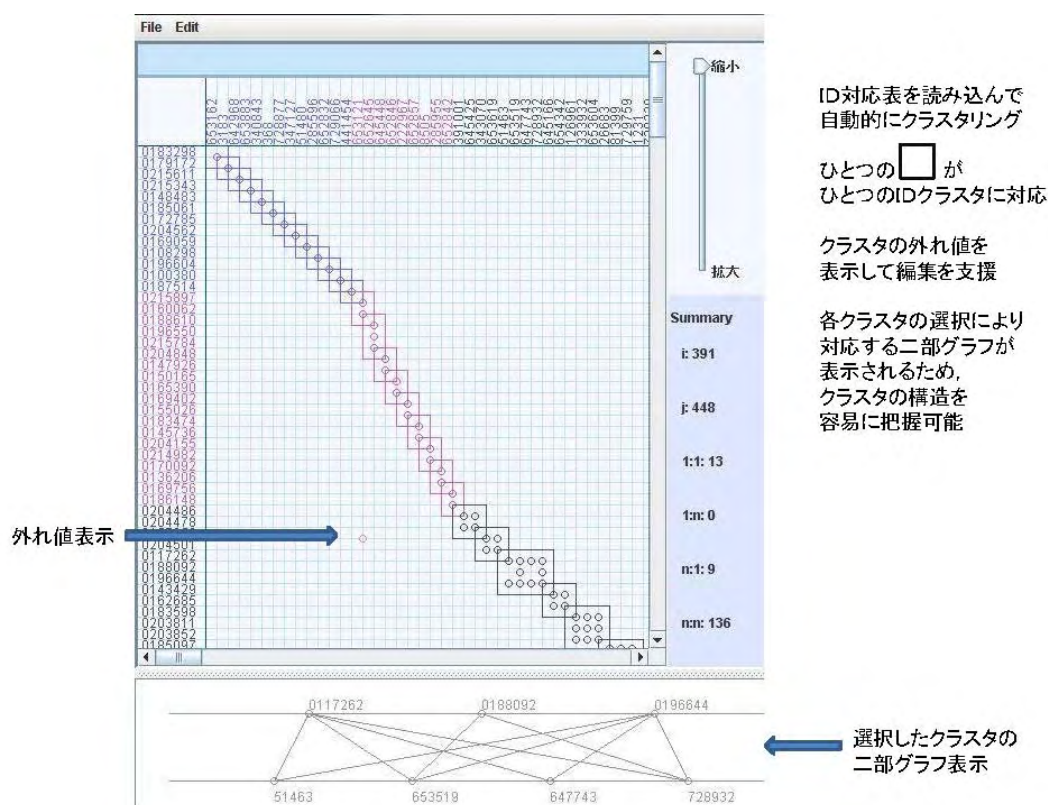


図3.3 IDEditor: ID対応表をクラスタリングし外れ値を表示

(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)

「ワークフロー技術を用いた統合DB環境の構築」のためには、1) システム環境構築、2) ワークフロー設計、3) ソフトウェア、データベース間インターフェース設計、4) マンマシンインターフェースの仕様を決める必要がある。

1) システム環境構築については、グリッドを採用することで、作業を行った。

2) ワークフロー設計については、実用的なワークフロー設計を行うため、配列解析システムをプロトタイプとして開発し、同時にそれを用いて構築したシステム環境のテストを行った。また、ワークフローを固定にするべきか、可変にするべきかの議論も行った。

3) ソフトウェア、データベース間インターフェース設計については、ライフサイエンス統合データベースセンターとの打ち合わせにおいて、SOAPを採用することに決定したので、ソフトウェアやデータベースのSOAP化を進めた。

4) マンマシンインターフェースについては、2)ワークフロー設計の方針と関係するが、ワークフローを可変にした場合には、利用者の操作性も考慮する必要があることから、既存のツール (KNIME, tavernaなど) の調査を行った。

1) システム環境構築

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、必要なデータや解析手法などを自動的に選び、データベースと解析ツールのワークフローを作成し、最適な資源を使って結果を返す統合DB環境を構築するための基盤技術として、下記の観点からグリッド技術を採用することに決定した。

- ① 目標とする統合DBを構築するために必要な技術が、ほぼグリッド環境で実現できている。（例：GEOGrid：地球観測グリッド）
 - －長時間JOBの対応
 - －ソフトウェアやデータベースの利用制限（アカデミック利用のみ等）
 - －他組織の情報・計算資源の利用が容易など
- ② 分散処理を実現する技術として、不特定多数のPCを対象とする技術はあるものの、異なる組織間で計算を可能とする技術は現時点では実用上グリッド以外では存在しない。
- ③ ワークフロー実現のために必要で、かつ現状のグリッド技術にない技術の開発に関しては、産総研グリッド研究センター（平成20年度から情報技術研究部門）の協力が得られる。

グリッド環境を、産業技術総合研究所（産総研）生命情報工学研究センター（CBRC）内に作成し、図3.4に示すシステム概念図で統合DB環境を構築する方針を立てた。

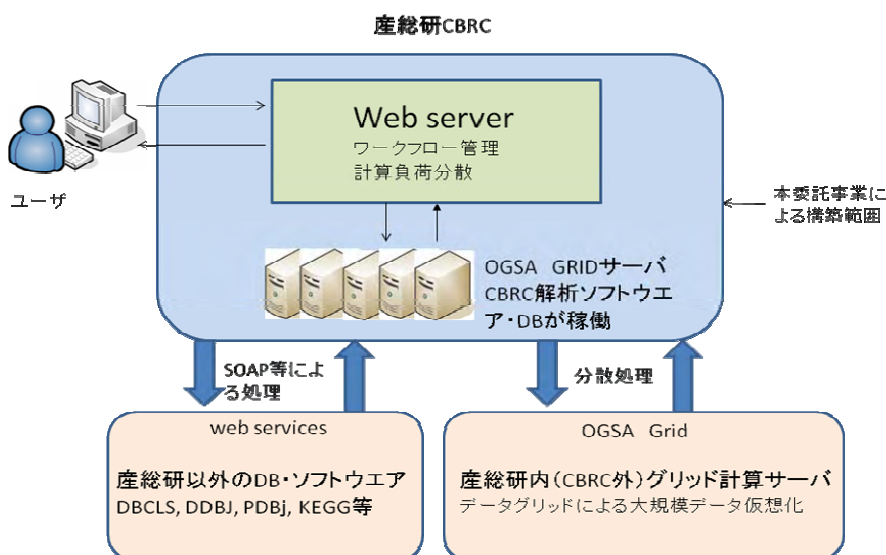


図3.4 システム概念図

ワークフローで実行されるソフトウェアは、OGSAによるグリッドの標準アーキテクチャによりグリッド環境内の計算資源で動作し、データベースは、OGSA-WebDBを用いて大規模データの仮想化され、利用される。一方、グリッド環境に組み込めないソフトウェアやデータベースは、標準的なWebサービスを用いて利用する。

平成19年度は、図3.5に示すグリッド環境を作成し、図3.4における本委託事業による構築範囲部分の開発を行った。具体的には、CBRCで開発されたソフトウェアをまとめ、ライフサイエンス統合データベースセンターの統合ホームページ(LSDBS)で公開している「生命科学データベース横断検索 (<http://lifesciencedb.jp/dbsearch/>) と連携したプロトタイプシステム（次項参照）を開発した。

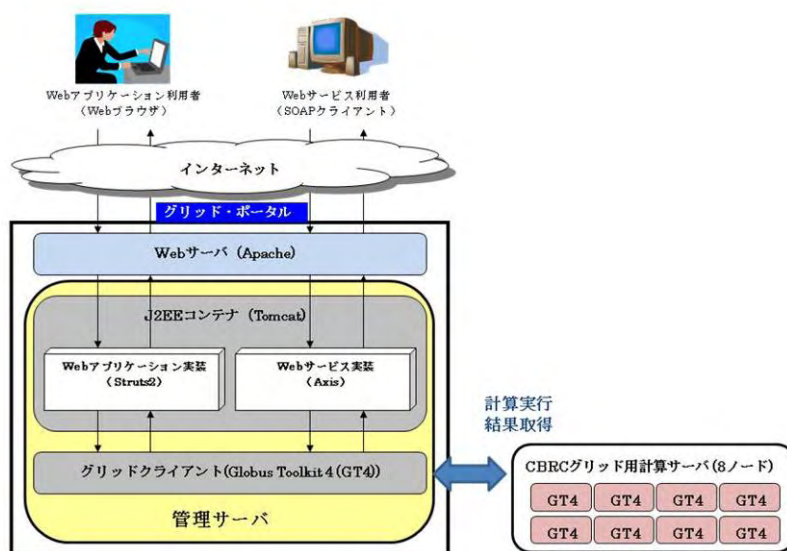


図 3.5 グリッド環境の概要

グリッド環境は、管理サーバとWebサーバからなり、統合DBのサービス専用の計算機資源として、グリッド用計算機サーバを導入することにより、一般公開後の運用に耐えられる環境が整った。

2) ワークフロー設計

統合DBのサービスとして提供するワークフローを開発するためには、

- ① ソフトウェアやデータベースの組み合わせの洗い出し
 - ② 利用するソフトウェアやデータベースのインターフェースの設計
 - ③ 利用者の利便性を考慮したマンマシンインターフェースの設計
- を行う必要がある。

平成19年度は、実用的なワークフロー設計を行うために、上記3点の議論を行った。具体的には、どのような利用者をターゲットにするか、その利用者はどのようなサービスを期待しているか、それを実現するには、どの程度の計算パワーが必要か、どのようなソフトウェアやデータベースが必要かなどを議論したが、結論に至らなかった。他方、構築したグリッド環境を確認する必要があったため、そこで動作するプロトタイプシステム（配列解析）を開発した。図3.6にその流れを示す。プロトタイプシステムでは、CBRCで開発され、WWWで一般公開されている6種類のソフトウェア（3種類のディスオーダー予測

プログラム：POODLE-S,L,W、GPCR-Gタンパク質結合選択性予測システム：GRIFFIN、膜貫通タンパク質のベータストランド予測プログラム：TMBETA-NET、ベータバレル膜タンパク質判定プログラム：TMBETA-DISC）がグリッド環境で動作している。Webの画面上から、配列情報を入力すると各プログラムが各サーバ上で並行に動作し、結果を返す。また、このプロトタイプシステムに、「ポータル整備・運用チーム」が開発した「生命科学データベース横断検索」の機能を追加した。検索用の文字列をプロトタイプシステムの初期画面に入力しておく、この横断検索が動作し、検索結果をCBRCへ転送する。図3.7に全体の概略を示す。各処理結果は、ソフトウェアごとに付けられたタグを選択することにより、結果が表示される形式となっている。

今後は、実用的なワークフローを設計し、グリッド環境に実装するとともに、横断検索の機能を必要に応じて加えていく予定である。



図 3.6 プロトタイプシステム（配列解析）の流れ図

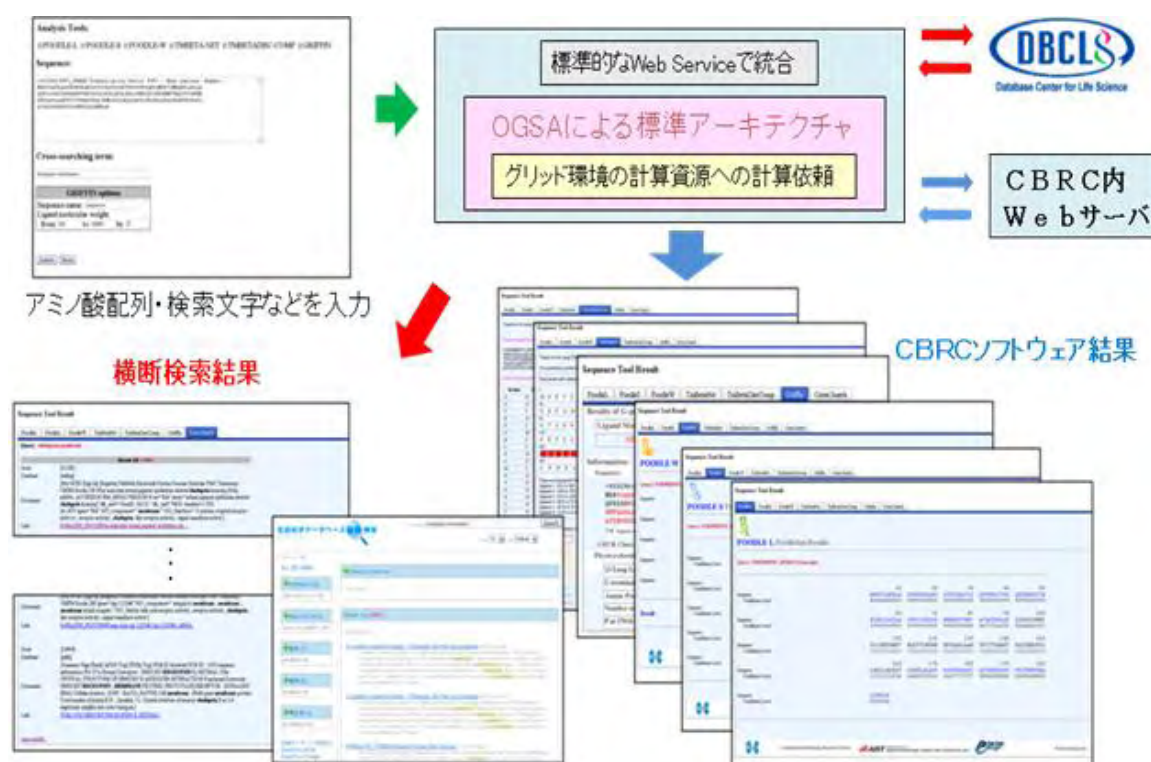


図 3.7 プロトタイプシステム（配列解析）と「生命科学データベース横断検索」の連携概略図

3) ソフトウェア、データベース間インターフェース開発

CBRCで開発されたソフトウェアとデータベースで、当面統合の対象となる7種類のソフトウェア (WoLF-PSORT, POODLE-W, POODLE-S, POODLE-L, TMBETA-NET, TMBETADISC-COMP, GRIFFIN) とその他4種類のソフトウェア (Scarna, RNAmine, Murlet, PHMMTS) のSOAP化が完了した。SOAPは、DBCLSとの打ち合わせにおいて、本委託業務におけるWebインターフェースとして採用を決めたプロトコルである。

4) マンマシンインターフェース調査

本格的なサービスとして提供するワークフローについては、現在議論を進めており、ワークフロー固定と可変のサービスを検討している。可変ワークフローは柔軟性が高い反面、システムが非常に複雑となり、ユーザ教育が必要となる場合が多い。他方、固定ワークフローは、ユーザ教育はほぼ不要で、システムはそれほど複雑にはならないが、柔軟性に欠け、ユーザのニーズに十分対応できない場面も想定される。

可変のサービスを提供する場合のマンマシンインターフェースとして、KNIME (<http://www.knime.org/>) と taverna (<http://taverna.sourceforge.net/>) の調査を行った。図 3.8 に KNIME、図 3.9 に taverna の例を示す。KNIME および taverna はいくつか存在する可変ワークフローの中では実績、ユーザ数、使い勝手、安定性、柔軟性等の観点から上位にランクされるものである。両者共に高い柔軟性があり、ユーザニーズへの対応という点では評

価されるが、ユーザが自由に使いこなすには高水準の教育とかなりの時間が必要との印象を受ける。従って、ユーザの負荷及びシステムの複雑度を最小限に抑えつつ、柔軟性を高めることが重要と思われる。また、現状では、Web サービスで利用するためには、そのための開発を行わなければならない。

以上を踏まえ、今後の方向性の一つとして、まず固定ワークフローを構築、その後一部に可変の考え方を取り入れつつ拡張することを検討している。

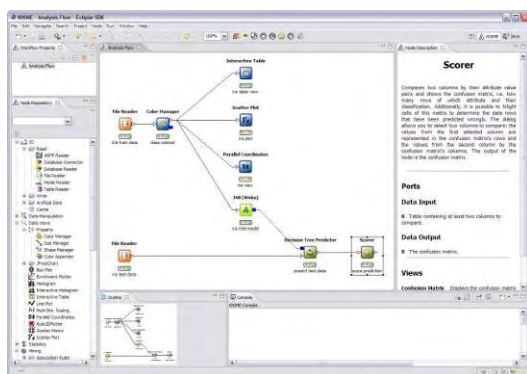


図 3.8 KNIME の例

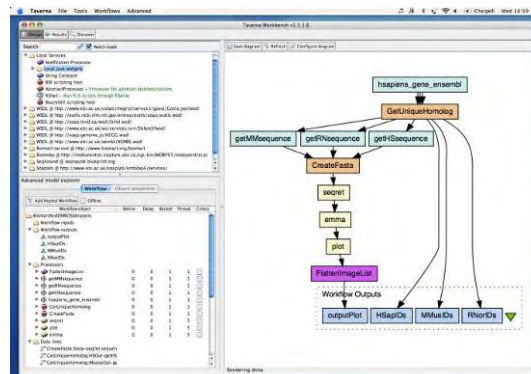


図 3.9 taverna の例

(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)

専門用語獲得技術については、1) 専門用語辞書システムの設計、2) 専門用語解析技術の開発および3) 専門用語タグ付け手法の設計を行った。

1) 専門用語辞書システムの設計

実施計画では、専門用語間の意味的類似性、同義性の判定と、用語の意味関係を記述するための用語辞書システムの基本設計を計画していた。

意味的類似性に関しては、用語が出現する文脈（前後に出現する他の語）によって表現される文脈ベクトルを用いた類似度定義を行い、特に類似語の効率よい検索手法に関する研究を行った。大規模な検索対象が高次元ベクトルによって表現されている場合、類似ベクトルの検索を逐次的に行っていたのでは膨大な検索時間を要する。そこで、ピボット集合を用いた検索処理の高速化が行われている。この方法では、高次元ベクトル空間内に事前に表点としてピボット集合を設定し、すべてのデータ点とピボット間の距離を求めておく。そして、新しいデータとピボット間の距離のみを計算することにより、多くのデータ点との距離の直接計算を省略することを可能とする。ピボット集合は既存のデータ点から選択するのが通常であるが、平成19年度の研究では、その制約を緩和し、かつ、有用なピボット集合の構築法を提案した。（以下の論文として研究発表：Manabu Kimura, 他, “Pivot learning for efficient similarity search,” 11th International Conference, Knowledge-based

Intelligent Information and Engineering Systems (KES2007), September 2007)。今後は、この技術を類義語および同義語の抽出法として利用していく。

また、用語を格納し、用語の表層形や読みあるいはその一部を指定して検索することのできる辞書システムの設計とプロトタイプシステムの実装を行った。次年度以降、ライフサイエンス辞書を組み込み、用語の内部構造のタグ付け機能を実装していく。現状のシステムのスナップショットを図3.10に示す。図3.10では、「～症」という表記をもった用語の一覧を辞書から検索し表示した様子を示している。現在は、同じ語の異表記（漢字・かなによる表記異なるもの）についてのまとめ上げを自動的に行った。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。

順序: Id 昇順 Review
 表示件数: 30 Review
検索条件変更

1 2 | [Next](#)

条件:	単語=～症, 品詞=名詞一般							60 Hits
	ID	単語	読み	発音	品詞	活用型	活用形	BASE
詳細	15234	むちうち症	ムチウチショウ	ムチウチショー	名詞一般			むちうち症
詳細	15236	むち打ち症	ムチウチショウ	ムチウチショー	名詞一般			むち打ち症
詳細	20511	カネミ油症	カネミユショウ	カネミユショー	名詞一般			カネミ油症
詳細	26638	ダウン症	ダウンショウ	ダウンショー	名詞一般			ダウン症
詳細	41003	胃酸過多症	イサンカタショウ	イサンカタショー	名詞一般			胃酸過多症
詳細	48047	炎症	エンショウ	エンショー	名詞一般			炎症
詳細	66592	既往症	キオウショウ	キオーショー	名詞一般			既往症
詳細	71147	急症	キュウショウ	キューショー	名詞一般			急症
詳細	73233	恐怖症	キョウフショウ	キョーフショー	名詞一般			恐怖症
詳細	73635	狭心症	キョウシンショウ	キョーシンショー	名詞一般			狭心症
詳細	75071	禁断症状	キンダンショウジョウ	キンダンショージョー	名詞一般			禁断症状
詳細	80226	軽症	ケイショウ	ケイショー	名詞一般			軽症

図 3.10 辞書管理システムのスナップショット

2) 専門用語解析技術の開発

新しい専門用語が次々に誕生するだけでなく、既存の用語についても、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など、用語辞書への登録を充実するだけでは完全な用語抽出を行うことができない。特に後者の問題に対応するため、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、機械学習を用いることによって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、医療分野の文献を用いた評価を行った。

図3.11では、横と縦に書かれた同一文内の黄色い（色の薄い）単語列が並列構造をなしていることが、三角の行列内の対角線によって表現されている。このように同一文内の類似単語系列を発見することが並列構造の検出につながり、動的計画法などを用いて図3.11のような類似系列の検出が行われるが、単語の類似度の定義や探索にまつわる種々のパラメータの値の決定をアドホックに行うことが多かった。本研究では、系列タグ付けで高い性能を示している条件付確率場(Conditional Random Fields)の考えを拡張し、類似系列の検索のためのパラメータ学習を多数の事例から行う方法を提案した(以下の論文として研究発表：Masashi Shimbo他, “A Discriminative Learning Model for Coordinate Conjunctions,” The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 2007)。これにより、従来手法より高い精度での並列構造の解析が可能になった。今後は、多重に埋め込まれた並列構造や述語並列構造の解析手法の開発を行っていく。

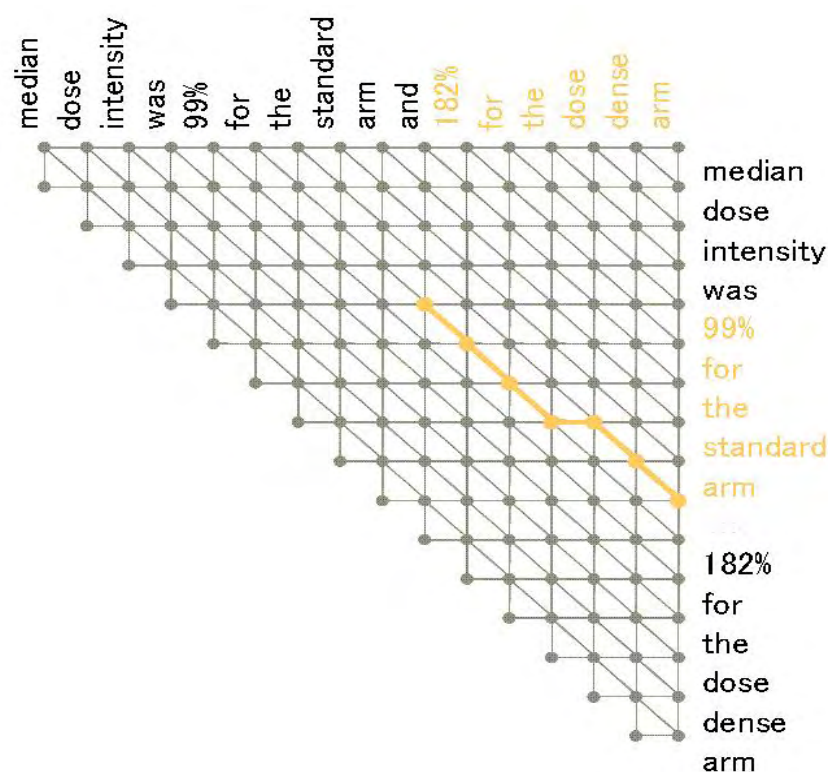


図 3.11 並列構造解析の概念図

3) タグ付け手法の設計

専門用語解析の基礎データの蓄積、特に用語辞書を充実させるため、専門分野テキストに現れる用語の特定や用語の種類の分類を行うことが重要である。機械学習用の学習データを

蓄積するため、テキストへのタグ付け手法の設計を行った。本年は、実際の用語の抽出の対象としてWikipediaを用い、その文書オブジェクト構造を利用した用語獲得手法を設計した。

ほとんどの専門用語は複合語であり、複数の語や形態素からなる。その内部構造を知ることとは、用語の意味分類のみならず、複数の用語が融合して出現する並列構造を含んだ用語の解析の基礎データとしても重要である。専門用語の内部構造（単語間の係り受け関係）を人手によってタグ付けするためのインターフェースの開発を行った。図3.12に具体的な専門用語内の単語の係り受け関係のタグ付けを行っている様子を示す。矢印で表示された係り受け関係をマウスによる簡単なオペレーションで追加・修正することが可能である。

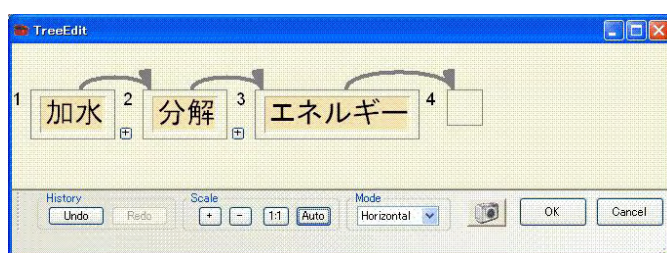


図 3.12 用語の係り受け構造のタグ付けツール

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）

3. 2. 1 ヒト統合 DB の開発・運用の実施計画

（情報・システム研究機構）

ゲノム情報、分子データベース、文献情報等を医学研究や創薬開発研究に統合利用する為に必要な知識の整理を行い、辞書・オントロジーのプロトタイプを開発し、公開する。これらを用いて分子データベース、文献データベースの統合に着手する。

（九州大学）

DNA 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の多型データの共有・公開法を確立する。またこれをもとにゲノム多様性のプライマリーデータベースである多型知識表現 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コミュニティの種々の informatics literacy レベルでの利用者に安定的に供給する。

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）

ゲノム情報、分子データベース、文献情報等を医学研究や創薬開発研究に統合利用するために必要な知識の整理を行うために、文献情報活用技術とそのキュレーションへの適用技術

の開発、アノテーション支援ツールの開発を行った。辞書、オントロジーのプロトタイプとして生物学名の羅英日対応辞書、細胞オントロジーや教科書索引語の専門用語辞書化機能、遺伝子名シソーラスの確認機能の開発、併せて解剖学情報のプラットフォームであるポリゴンマンの高機能化などを行った。また、分子データベース、文献データベースの統合の一環として DNA バンク目次システムの高機能化、及び癌を題材とした遺伝情報と臨床情報の統合による知識の整理法を検討した。

文献情報活用技術とそのキュレーションへの適用技術の開発については、まず、複数の論文の内容を整理してキュレーションに利用するために、出版社が型にはめたものを、型の中で読み取れる情報を型からはずして利用可能にする逆パブリッシングを行うプログラムを開発した。20 の代表的論文の様式を読み解いて、(1) 論文の特定 (PubMed との対応)、(2) リファレンスの抽出 (括弧や肩付き文字として表現されている番号を同定)、(3) 論文の末尾にある参考文献表示を PubMedID などに変換することによるオンライン情報とのリンクなどを行った。さらに、これを数千の入力論文ファイルに対して行い、出版社の区別のない標準化されたタグ付きテキスト形式ファイルで出力するプログラムを開発した。また、論文セットを元にした引用被引用関係を解析し、さらに各論文に付随した MeSH タームをカテゴリ毎に分類し、それらの情報を用いた論文間の類似度の測定を、乳がんに関する論文を例に行った。さらに、統合 DB プロジェクトで内部利用やサービスの為に管理している 10,000 以上の由来や形式の異なる文書を一括管理するためのシステムを構築した。アノテーション支援ツールの開発については、アノテーションやキュレーションの根拠情報を簡単に記録可能にすることを目的に、Web 上の文書の特定の部分に関するメタ情報を簡単な操作で作成するツールを利用が簡便なウェブブラウザ(FireFox)のアドオンとして開発した(図 3.13 参照)。具体的には閲覧中のウインドウで選択した文や図をアドオンが表示するサイドウインドウにドラッグドロップするだけで選択領域のコピーとその URL、およびページ内座標が記録され DB 化される機能を持つ。さらに上記メタ情報を複数のアノテータ間で共有可能にするための出力部分を開発した。具体的にはメタ情報 DB を XML 形式テキストで出力する機能であり、出力ファイルをやり取りすることでそれぞれのツールに他人の情報を移植できる。また XML データから自由形式の表を作成し表示する機能も作成した。



フォルダーを階層的に作成し情報を整理して保存。デザインはスタイルシートでカスタマイズ可能。

web上のドキュメントにアンダーラインをつけるとブックマークとして自動的に保存。



図 3.13 Wired-Marker: インターネットの情報収集共有整理のためのツール

羅英日対応辞書については、平成 18 年度作成したラテン語生物学名和名対応辞書をベースに、和名表記方法の統一、生物学名と英語一般名の対応や見出し語の拡充などを行い、併せて辞書配布システムを構築した。また細胞オントロジーについては、動物細胞をリストアップし細胞の行動に関する情報を DB 化する細胞行動データベースの構築を目的として、収集したデータから細胞名や行動名などを抽出して解析した。教科書索引語の専門用語辞書化機能の開発については、教科書の索引語を専門用語辞書として利用することを目的として、教科書の目次・索引データを管理・表示するシステムを開発した。索引情報を教科書の構成情報と共に格納・参照するデータベース機能に加え、索引情報の専門用語辞書化に必要な同義語処理、曖昧語処理、意味分類の付与等を行う機能を開発した。また、遺伝子名シソーラスの確認機能の開発については、平成 18 年度に集めた遺伝子名称とその関係のうち、各種 DB を情報源とする同義関係情報に根拠情報を付加することを目的に、PubMed アブストラクト中で遺伝子名の同義語関係について陽に表示していると思われる文を機械的に検出し目視で確認分類した。さらにその結果に基づいて名称の同義語関係を表す表記（括弧や also, known as など）をパターンとして集積した。また、以上の開発と併せて、辞書構築システムの基本機能として必要となる、用語管理を行う基本システムの開発を行った。これは用語の概念、字面、セマンティックタイプを管理する基本データベースシステム、および、データベースに用語リソースを挿入するプログラム、文書リソースと用語のインデックスを管理するシステムからなる。ポリゴンマンの高機能化については、ポリゴンマン辞書の要素をグラフィカルユーザインターフェース上で組み合わせて、3次元人体地図情報を生成するアナトモグラフィシステムを開発した。ポリゴンマン辞書の閲覧・検索と解剖学要素の選択、各解剖学要素の表示形式の設定、フォーカスや視点の設定、生成した 3次元人体地図情報の保

存が可能なシステムを開発した。さらに、辞書としての有用性を向上させるために、心臓や腹腔などの詳細化を行った(図 3.14 参照)。

アナトモグラフィー (Anatomography) 実行Webフォーム

1. 左のリストから、表示したい臓器・器官を選択。 2. Organ Informationの"ADD"ボタンでリストに追加。 3. 右下"Draw"ボタンを押下してください。



図 3.14 アナトモグラフィシステムの表示例

DNA バンク目次システムの高機能化については、平成 18 年度構築したシステムをライフサイエンス統合データベースセンターのサーバに移植し、併せて統合ホームページ上で稼働するすべてのサービスと連携して一括検索が可能なシステムを構築した(図 3.15 参照)。さらに文献データベースとの統合の一環として、用語検索でヒットした特許配列に関して米国、欧州、及び日本の特許庁が公開している情報へリンクし、利用者が特許文書を閲覧できる仕組みを開発した。また、DNA バンクの利用検索をアプリケーションプログラムインターフェイスを用いて DDBJ サーバにて行い、その検索結果をプロジェクト分類に基づいて分類し表示する機能を開発しサービスに提供した。さらに、DNA バンクレコードを国別、年度別に分類し、プロジェクト別に自在に序列化しグラフ化することで DNA シーケンスを利用した分子生物学研究のトレンドが一望でき、さらにそのトレンドから特定の研究データを取得できるシステムを開発した。一方、DNA バンク全体に対して類似性検索を行うとヒト遺伝子などでは通常数百から数千件におよぶ類似配列が存在する。そのため、これらの配列の由来を明瞭に表現するために、類似性検索であがってくる多数の配列をプロジェクト単位にまとめプロジェクト分類で展開するシステムも開発した。また、文献を利用した知識整理については、論文の引用部分を解析し PMID を付与するプログラムを作成し引用解析を行った。

対象として過去 5 年間に発表された乳癌に関する臨床論文を利用し、引用・被引用関係をもとに論文を手作業によってクラスター化し、内容との比較を行い、クラスター化部分のプログラム開発に必要な学習データを作製した。



図 3.15 DNA バンク目次システムの表示例

(2) 多型知識表現技術の開発 (九州大学)

1) 多型知識表現技術の検討

信頼度検証可能な形態での多型データの共有・公開法を確立するために、DNA アレイ技術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た。これをもとにゲノム多様性のプライマリーデータベースである多型知識表現 DB を構築してウェブ上に公開するシステムを開発した。これによりヒトゲノム多型知識情報を生命科学コミュニティの種々の情報リテラシーを有する利用者に提示するための基盤が構築できた。

大規模一塩基多型 (SNP) 及びコピー数多型 (CNV) に関する情報は、ヒトゲノムの数十万箇所以上について少なくとも数百人を試料として解析するもので、試料収集、ジェノタイプ解析、解析後データ処理の各段階でデータの誤り及び偏りが不可避免的に生じることが広く知られている。これらの誤差を定量的に検証可能な形態で情報提供を行うことは、データの

有効利用に必須であり、これを実現するためのデータベースの形態について検討した。その結果、ジェノタイプ解析を行う組織（以下データ生産者と）と別個の組織がデータ生産者から一次データ（ジェノタイプ解析実験で得られる生データ）を取得し、データ生産者とは独立してクオリティーチェックを行い、この結果を生データとともに開示（個人遺伝情報を含むので限定開示となる）することが第三者による更なる検証のために必須であると結論した。限定開示の範囲、開示を受ける者の認定方法等には、試料収集組織との協議と同意が必要であり、その具体化は残された課題である。また、SNP 解析の技術は急速に進歩しており、新しい解析法でのデータのクオリティーチェックの方法についても今後遅滞なく対応していく必要がある。

2) 多型知識表現データベース SJAG の開発と開示

平成 19 年度は上記視点に合致するデータベースのプロトタイプとして「Shared database of Japanese Applied Genomics」(SJAG)を構築した。SJAG では、現在主流の SNP タイピング法であるアフィメトリックス社の DNA アレイを用いた解析データを Linux 環境下、MySQL で管理している。このシステムでは、以下の 3 段階のインフォーマティックス環境レベルの利用者がウェブを利用してそれぞれ種々のデータ検索を行うことを可能にした。レベル 1 はウェブブラウザが使えるレベル、レベル 2 は解析プラットフォーム専用の解析ソフトウェアが利用可能なレベルで、レベル 3 は Perl スクリプトを用いて大量データの解析が可能なレベルを想定したものである。さらにデータのクオリティコントロール (QC) を第三者が行うことを可能とするために、解析生データである Affymetrix cel file をダウンロードすることを可能とした。実際のデータとしては、ゲノム特定領域研究・基盤ゲノム・ジェノタイプセンタ（東京大学）が Affymetrix 500K アレイを用いて連結不可能匿名化された日本人 200 人についてタイピングしたものを利用した（図 3.16）。さらに、SJAG の評価を目的として、多型解析に実績のある東京大学、東海大学及び九州大学のゲノム研究グループに限定的に開示した。

3) 多型知識表現データベース SJAG の評価

2) 項記載の三つのグループへの限定開示の結果、データベース構造の合理性、利用時の解りやすさ等に関して一定の評価をいただくと同時に、一つのグループから、提供した生データをもとに QC 解析を実施した結果、データ精度に問題点を発見したとの指摘があった。これは、本データベースが QC に有効に機能したことを示すもので、上記の検討結果を裏付ける結果が得られた。本年度は、専門家への限定開示にとどまり、公開というフェーズには至らなかった。その理由の一つは、データ提供元のゲノム特定領域研究グループでデータ公開、共有に関する指針が決まらなかったことであり、もう一つの理由は多型情報の解析プラットフォームが一代古くなってしまったことである。現在は次世代の SNP 解析プラットフォームである Affymetrix SNP 6.0（100 万個の SNP を対象とする）による解析が主流と

なっており、今後は、この一次データの提供を受けて、データベースのバージョンアップを行うことを計画している。また、今後半年程度で検討が終了する予定のゲノム特定領域研究におけるデータ公開、共有指針の決定後、バージョンアップが完了したデータベースの一般への公開と研究コミュニティへの開示を進める予定である。その際、個人情報を含まない統計データは一般公開可能なものの、生データに関しては研究コミュニティでの審査を受けたグループへの限定開示になるものと考ええる。いずれにしても、平成19年度の開発により、データ生産者とは独立した組織としてのデータ公開と共有のためのシステムは構築できたものと考ええる。



図 3.16 SJAG の概要と表示例

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)

ゲノム情報、分子データベース、文献情報等を基礎生物学研究および産業応用に統合利用する為に必要な知識の整理を行い、辞書・オントロジーのプロトタイプを開発し、公開

する。また、ゲノム情報、分子データベース、文献情報等を感染症研究・発酵食品産業等に統合利用する為に必要な知識の整理を行い、辞書・オントロジーのプロトタイプを開発し、公開する。これらを用いて分子データベース、文献データベースの統合に着手する。

(かずさDNA研究所)

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については、植物ゲノムのデータベースを対象とした統合への課題調査により明らかとなった、植物ゲノムデータベースの現状へのユーザの不満を解消することを目標とした業務を行う。具体的には、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構築し、次いでその環境へのデータの集積と高度化を図る。

1) 高度情報集積データベース開発

分子データ上のポジションを基盤とした、アノテーション・キュレーションの高度化及び統合化を可能とする、基盤データベースを整備し、植物及び植物関連微生物のゲノム解析情報を対象に運用する。同時に、そのようにして集積したアノテーション情報をゲノム上に表示し、優れたインターフェースにより閲覧可能なシステムを開発する。

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利用者にとって大きな障壁であることから、植物遺伝子名や遺伝子IDの整理をおこない、名称の食い違いを解決するためのWWWサービスを行う。

2) ゲノムアノテーション情報の蓄積と高度化

上記データベース環境を整備した後ただちに、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実施し、公開する。

データベースの維持、アノテーション、キュレーションを行う人材の教育に関わる情報も、将来の継続的なデータベースの充実には不可欠である。そのため、かずさDNA研究所で H18に構築した人材養成のための教育コンテンツの共有システムを継続して作成し、維持しつつ教育分野の事業と連携し、内容の一層の充実を進める。

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)

ゲノム解析が進んでいる多様な微生物菌株を当面のターゲットとして、統合 DB の開発・運用に着手した。

微生物菌株については、1) 微生物ゲノム情報の統合利用・解析環境の調査、2) 微生物情報資源の機能、ならびに 3) 微生物ゲノム論文からの知識抽出について調査して、微生物に関して、系統保存の情報からゲノム解析の結果の情報までを統合利用可能とする環境構築の要素を整備した。

1) 微生物ゲノム情報の統合利用・解析環境の調査

Web ブラウザーを利用してアノテーション情報を参照するタイプとクライアント側にインストールして使うタイプの双方について調査し、また、第2世代さらには第3世代のシーケンサの開発と普及がもたらす膨大なゲノム情報から有用な生命情報を効率よく抽出するために必須と考えられる自動アノテーションシステムについて調査した。これらの結果をもとに、平成20年度以降、もっとも有用と思われるサービスを設計する。

① ゲノムブラウザーについて

ゲノムブラウザーは、アノテーションされた情報を閲覧するタイプと、アノテーションされた情報を解析し可視化するタイプに大別される。前者のタイプは膨大なデータとセットになることが多く、多種のデータを一度に見たい時に使用し、後者のタイプは利用者独自のデータを含む興味のあるデータのみを利用して多種の解析をする場合に使用する。また、前者はWeb向きであるのに対して、後者はクライアントで実行するのが向いている。したがって用途に応じて開発する方が適切と考える。

Web ブラウザーを使用したオープンソースのゲノムブラウザーシステムでは、表示領域を移動させると表示画面全体が更新され、画面表示に時間がかかる。これを補うために、AJAXの技術を取り入れたシステムも開発されつつあるが、利用者視点から見たときに使い易さ・実用性を兼ね備えたものはまだない。特にこのようなシステムでは動作の軽さを求められるのでまだ開発・改良の余地があると考えられる。また利用者のPCで実行するタイプのクライアントツールも利用者側の解析手法のニーズが変わってくることも想定される。これからは更に両タイプシステムについて実際にゲノム情報を利用する研究者の意見を重視した機能集約が必要になる。

② 微生物ゲノム自動アノテーションシステムについて

微生物ゲノム自動アノテーションシステムは、国際塩基配列データベースへの登録をサポートするタイプと比較ゲノム解析のために大量処理するタイプに大別される。前者は新規ゲノムの解析の機能を備えているものが多い。新規ゲノム解析の大まかな流れ・ツールは共通しており、これをシステム化することで利用者が自前で解析の環境を構築することなく使用できればメリットが大きいと考える。いずれにしても、どちらのタイプについても、参照するデータベースと計算機資源を更新・維持していけるのは大規模な機関に限られていくであろう。

2) 微生物情報資源の機能

171サイトを系統保存(資源センター)、ゲノム/アノテーション、ゲノム解析、細胞バンク、生物遺伝資源部門リンク、データバンク、統合DB及び反応経路DBの観点に分類し、各サイトで提供されている機能のマップを作成した。

調査の対象サイトは、国立遺伝学研究所のGenome Information Broker (GIB)、Gene Trek in Prokaryote Space (GTPS)および微生物世界データセンター(www.wdcm.org)、日本微生物資源学会のオンラインデータベース(www.jscc-home.jp)、財団法人バイオインダストリ

一協会の JBAM、基礎生物研究所の MBGD など国内の主要な微生物情報資源、ならびに、国外の微生物情報資源(例. Genome Online Database (GOLD), RDP11, SILVA, GreenGene, www.StrainInfo.net, www.Mycobank.org) などの微生物データベースの名称および URL 情報についてデータベース一覧 (一次データベース一覧) を作成した。また、これらデータベースが取り扱うリンク集についても調査を実施し、データベース一覧に存在していなかったデータベースも追加した。

情報資源をマップするために設定した項目は、DB ID、名称 (英文、和文)、URL、分類軸(DB 型)、分類軸(対象)、分類軸(生物種)、分類軸(DB グループ)、開発機関、開発機関 (副)、国名、掲載カタログサイト、PubMed ID、参照 PubMed ID、DB の 1 行説明、英文説明、和文説明、データ量、統計情報掲載ページ、独自データおよび独自解析結果である。

微生物ゲノムのアノテーションは、ゲノム配列とその断片配列と表現形質を紐づける作業であるため、今回の調査でとりまとめた微生物情報資源を探索しやすい形で公開することが、従来の微生物学者からゲノム解析グループまで幅広いコミュニティーにとって有用であると考えられる。一方で、表現形質のデジタル化については著名な出版物からのデータ抽出の努力がさらに必要と思われる。

オントロジーについては、ゲノムアノテーションで最も問題になる「参照データベースにおいて実質的には機能が不明なことを示す表現の揺れ」を調査して、その結果にもとづいて有意な記述と判別する Web API を整備した。この機能によって、例えば possible ORF, putative ORF, probable ORF などのさまざまな表現や綴りの誤りを hypothetical protein に統一することができる。この Web API は DDBJ の Web サービスのカテゴリーGTPS の中の checkAnnotation メソッドとして公開し、表現の揺れに関する辞書も公開した (<http://gtps.ddbj.nig.ac.jp/terminology/index.html>)。

3) 微生物ゲノム論文からの知識抽出

微生物アノテーションとして、ゲノムが決定された 500 種以上の微生物の特徴を発表された論文から抽出し、再利用可能な知識として表形式にまとめた。その一部を図 3.17 に示す。各論文に対応する PubMed ID 毎に、第 0 カラムには記載されている微生物の名称に関する抽出情報が記載され、以下第 1 カラムから第 8 カラムにかけて、菌同定学分類に対応する形態構造運動に関する記述、対象微生物の病気や食品、工業応用といった人社会との関係に関する記述、代謝に関わる有機分子との関係に関わる記述、寄生や共生といった生態系との関係に関わる記述、分類、進化といった他の生物種との概念的関係に関わる記述、基礎科学的意義を有する学問上の概念に関わる記述、及び対象微生物に関係しないが上記の各項目に対応する関係先の説明に関わる記述を抽出して記載した。図 3.17 には、対象とした微生物に関わる名称を、関連する PubMed ID や分類とともに示した表の一部も示した。本アノテーションは、論文中の記述を引用し文献の関連部分へのリンクをすべて残す為に情報共有ツール (WiredMarker) を利用し、大学院生のリサーチアシスタントを中心にツールの改良を図りな

から実施した。

PubMed ID	0. 対象微生物	1. 対象の形態・運動・生理・生態・系統・分類	2. 対象と宿主の関係 (感染・共生・寄生)	3. 対象と環境分子の関与 (代謝・信号伝達)	4. 対象と生態系・環境・地理・気候・気象・気候変動 (分布)	5. 対象と学問上の関係 (学際性)	6. 対象と学問上の関係 (学際性)	7. 対象と学問上の関係 (学際性)	8. 対象と学問上の関係 (学際性)
12217542	1. <i>Prochlorococcus</i> 2. and 3. <i>Prochlorococcus</i> 4. <i>Prochlorococcus</i> 5. <i>Prochlorococcus</i> 6. <i>Prochlorococcus</i>	1. The marine unicellular cyanobacterium <i>Prochlorococcus</i> is the smallest known photosynthetic autotroph [Ref.1].	1. and is responsible for a significant fraction of global photosynthesis 2. As an autotroph, <i>Prochlorococcus</i> provides an important pathway for the transfer of carbon from the atmosphere to the ocean [Ref.2].	1. and is responsible for a significant fraction of global photosynthesis 2. As an autotroph, <i>Prochlorococcus</i> provides an important pathway for the transfer of carbon from the atmosphere to the ocean [Ref.2].	1. and is responsible for a significant fraction of global photosynthesis 2. As an autotroph, <i>Prochlorococcus</i> provides an important pathway for the transfer of carbon from the atmosphere to the ocean [Ref.2].	1. and is responsible for a significant fraction of global photosynthesis 2. As an autotroph, <i>Prochlorococcus</i> provides an important pathway for the transfer of carbon from the atmosphere to the ocean [Ref.2].	1. and is responsible for a significant fraction of global photosynthesis 2. As an autotroph, <i>Prochlorococcus</i> provides an important pathway for the transfer of carbon from the atmosphere to the ocean [Ref.2].	1. and is responsible for a significant fraction of global photosynthesis 2. As an autotroph, <i>Prochlorococcus</i> provides an important pathway for the transfer of carbon from the atmosphere to the ocean [Ref.2].	1. and is responsible for a significant fraction of global photosynthesis 2. As an autotroph, <i>Prochlorococcus</i> provides an important pathway for the transfer of carbon from the atmosphere to the ocean [Ref.2].
1222482	1. <i>Escherichia</i> 2. <i>Escherichia</i> 3. <i>Escherichia</i> 4. <i>Escherichia</i> 5. <i>Escherichia</i> 6. <i>Escherichia</i> 7. <i>Escherichia</i> 8. <i>Escherichia</i> 9. <i>Escherichia</i> 10. <i>Escherichia</i> 11. <i>Escherichia</i> 12. <i>Escherichia</i> 13. <i>Escherichia</i> 14. <i>Escherichia</i> 15. <i>Escherichia</i> 16. <i>Escherichia</i> 17. <i>Escherichia</i> 18. <i>Escherichia</i> 19. <i>Escherichia</i> 20. <i>Escherichia</i> 21. <i>Escherichia</i>	1. <i>Escherichia</i> is a member of the family Enterobacteriaceae and is one of the most common bacteria found in the human gut [Ref.1].	1. <i>Escherichia</i> is a member of the family Enterobacteriaceae and is one of the most common bacteria found in the human gut [Ref.1].	1. <i>Escherichia</i> is a member of the family Enterobacteriaceae and is one of the most common bacteria found in the human gut [Ref.1].	1. <i>Escherichia</i> is a member of the family Enterobacteriaceae and is one of the most common bacteria found in the human gut [Ref.1].	1. <i>Escherichia</i> is a member of the family Enterobacteriaceae and is one of the most common bacteria found in the human gut [Ref.1].	1. <i>Escherichia</i> is a member of the family Enterobacteriaceae and is one of the most common bacteria found in the human gut [Ref.1].	1. <i>Escherichia</i> is a member of the family Enterobacteriaceae and is one of the most common bacteria found in the human gut [Ref.1].	1. <i>Escherichia</i> is a member of the family Enterobacteriaceae and is one of the most common bacteria found in the human gut [Ref.1].

微生物リスト

図 3.17 微生物アノテーションと微生物リストの一部

(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)

ユーザからの情報集積を受け入れ、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情報データベース環境を構築し、その環境へのデータの集積と高度化を図った。

1) 高度情報集積データベース開発

分子データ上のポジションを基盤とした、アノテーション・キュレーションの高度化及び統合化を可能とする基盤データベースを整備し、植物及び植物関連微生物のゲノム解析情報を対象として公開運用を開始した(KazusaAnnotation: <http://a.kazusa.or.jp/>(図3.18参照))。

このシステムは独自に開発したソーシャルブックマーク型サービスであり、ゲノム上の任意の遺伝子やロケーションにアノテーションを付与することができ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である。また、同時に登録されたデータやコンテンツの共有や議論を行い、さらにそのとりまとめを行うための場として、ソーシャル・ネットワークサービスとWikiシステムを立ち上げ、その運用を開始した (KazusaNavigation/KazusaWiki: <http://navi.kazusa.or.jp/>)。

同時に、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、直感的に扱うことのできる優れたインターフェースにより閲覧可能なシステムを配布可能なプログラムとして開発した (genoDivePro, genoDiveEu)。これらのプログラムは、これまでのゲノムブラウザと異なり、視点を保持したままで高速なズームイン、ズームアウトを可能とした。また、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を開発した。このサーバの利用により、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。これらはいずれも、ゲノム上の位置に高度に集積した情報を、直感的な形で閲覧することで、生物学実験に必須となるゲノム解析情報の統合的な提供を図るための基盤となるプログラム群である。

また、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利用者にとって大きな障壁であることから、名称の食い違いを解決するための関係性のデータを収集し、植物遺伝子名や遺伝子IDの整理を実施し、提供を開始した。

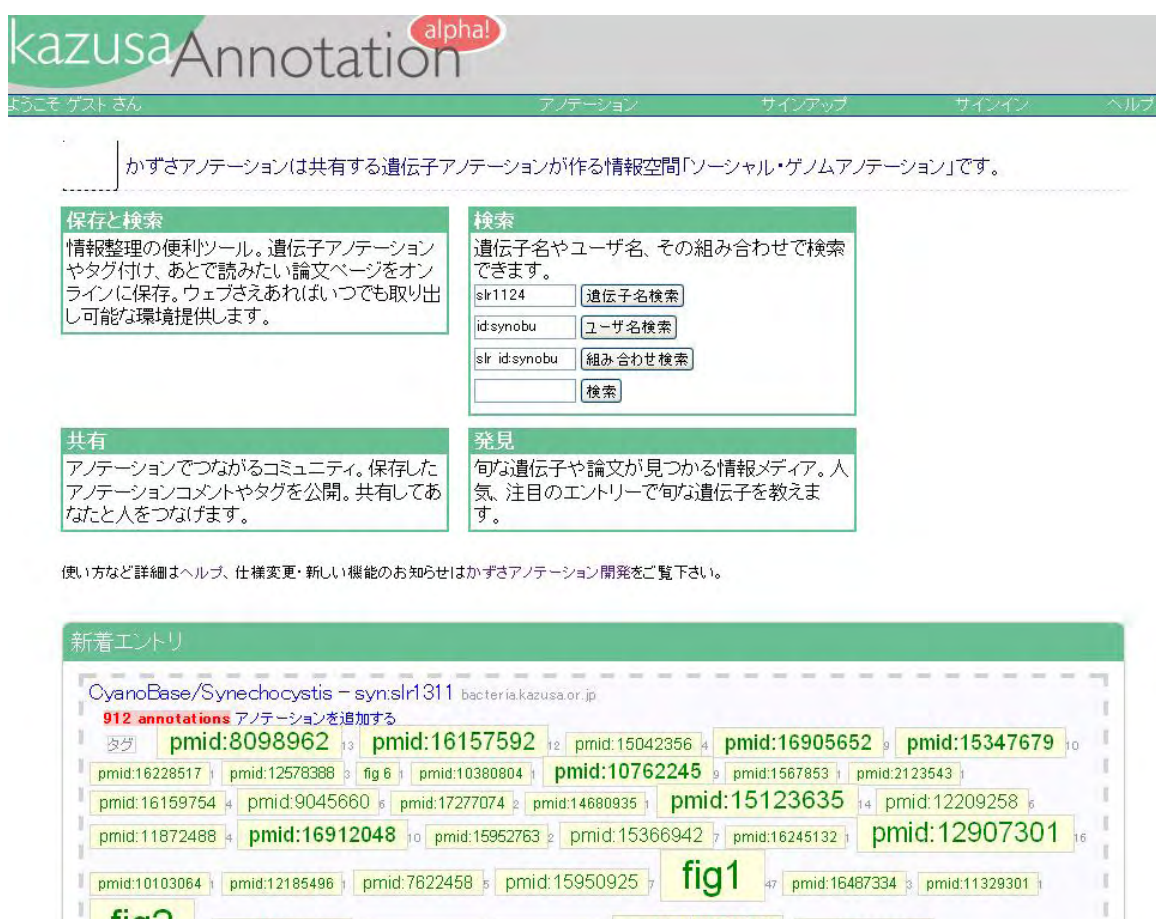


図3.18 かずさアノテーションシステム

2) ゲノムアノテーション情報の蓄積と高度化

上記のデータベース環境を整備し運用しながら、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KazusaAnnotation(<http://a.kazusa.or.jp/>)上で実施し、公開した。具体的には、酸素発生型光合成細菌である *Synechocystis* sp. PCC 6803 の遺伝子について記載のあるすべての研究文献中、実施期間内に入手閲覧が可能であった1,160報を、アノテータの目視により読み取り、その論文に言及されている遺伝子名と本研究所で命名した遺伝子のIDとを、該当論文のID(PubMed ID)と関連づけて収録した。平成19年度は、約12万件 (5,657 エントリー、96,827 アノテーション、21,983 タグ) の情報を蓄積し、広く利用に供している。

こうしたデータベースの維持、アノテーション、キュレーションを行う人材の教育に関わる情報も、将来の継続的なデータベースの充実には不可欠である。そのため、教育分野の事業と連携し、人材養成のための教育用コンテンツを、中核機関が作成した教育コンテンツ用サーバ(<http://MotDB.DBCLS.jp/>)へと移行させつつ、実習の実施とともに改善を加え、公開を続けている。

以上の開発と情報蓄積の周知や、その利用によるデータベース統合の可能性については、合計10回の(国際会議を含む)学会年会や講習会等での機会を得て発表をおこなっており、とくに実験生物学者を対象として、データベース統合の必要性をアピールするとともに、この委託業務による成果物の有用性を示し、利用促進を図った。

4. 統合データベース支援

4. 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、科学技術振興機構)

4. 1. 1 ポータル整備・運用、広報、普及啓発の実施計画

(情報・システム研究機構)

1) 統合DBプロジェクトポータルサイトの整備

統合DBプロジェクトのウェブサイトとしてwww.lifesciencedb.org(仮)等のドメイン名を取得し、統合DBプロジェクトの研究開発成果の公開、各種お知らせ、コンテンツ検索システムの提供を行うとともに、国内外のウェブサイトで配信されるライフサイエンス関連情報の収集システムのプロトタイプを開発し、ポータルサービスに必要な言語情報資源の収集・編纂等を行う。また利用者の意見集約の方法について調査する。

2) 統合DBプロジェクトポータルサイトの運用

サイトの運用に当たっては、参画各機関で作成されるウェブコンテンツ、各種ツールを集約し、利用者が容易に情報を取得できるような目次、索引体系を構築し、更新情報の配信を行う。

3) 広報・普及啓発

統合DBプロジェクトの事業内容や趣旨を解説したパンフレットを発行し学会等で配布するとともに、参画各機関に寄せられる意見や質問に対応する窓口を設ける。教育/人材育成において作成された教育プログラムを利用した講習会の開催をJSTと協力して行う。

(科学技術振興機構)

1) ポータルサイトWINGproの運用

個別DBについて、DB構築者および利用者からの意見を集約するためのシステムWINGproの公開を継続する。その際、内容の追加と利便性の向上を図る。

2) 事業サイト等の運用広報

文部科学省「ライフサイエンス分野の統合データベース整備事業」等のウェブサイトを運営する。参画各機関の当該事業に関する研究開発動向や活動、ライフサイエンス政策の動向等を広く広報・情報発信をする。新規記事の追加や修正を行うほか、検索機能の追加など利便性の改善や表示方法の変更を実施する。

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 (情報・システム研究機構)

統合DBプロジェクトのウェブサイトとして <http://lifesciencedb.jp> のドメイン名を取得し平成 19 年 10 月より公開を開始した(図 4.1 参照)。掲載コンテンツは平成 20 年 3 月現在で、新着情報、人材募集情報、LSDB ブログ、ニュース、生命科学データベースカタログ、DB



図 4.1 統合データベースプロジェクトサービス公開サイト

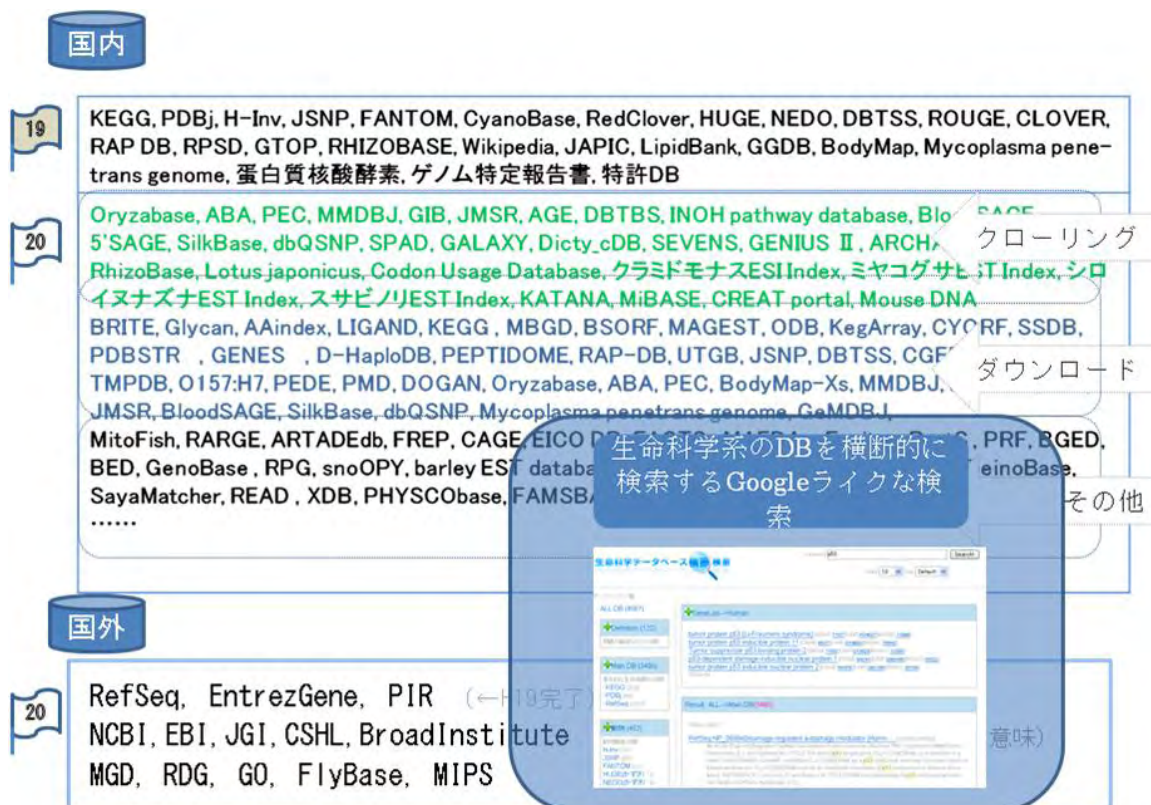


図 4.3 横断検索の掲載データベースと今後の計画

また、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及び著作権処理を行った。学会要旨の収集や学協会情報の収集も行い、さらに新聞記事のうち生命科学に関する記事の収集と解析を行った。日本語キーワードでの検索を可能にするための辞書の収集を行い、京都大学のライフサイエンス辞書、JSTの科学技術対訳辞書、当プロジェクトで開発した遺伝子名称シソーラス辞書並びに生物学名日本語一般名対応辞書を用いた検索システムを検討し、上記日本語文献の検索システムを構築した(図 4.4 参照)。



図 4.4 日本語文献検索サイト

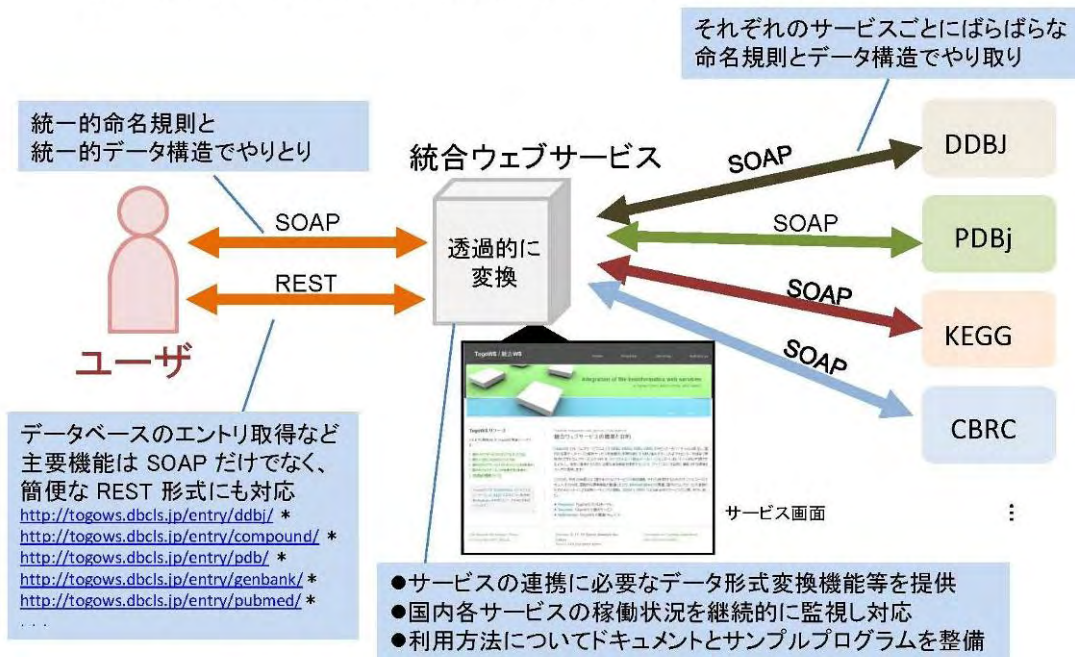
さらに、利用者の意見集約の方法について検討し、データベースカタログ及び統合 TV やブログなどのコメント欄を開設した。統合 TV は、データベースやウェブツールの使い方などを動画で発信するウェブサイトで、平成 19 年 7 月の公開以降、9 ヶ月間に 59 個の動画コンテンツを作成した(図 4.5 参照)。これまでに訪問数 30,423 を数える人気サイトに育っている。また、平成 20 年度からの本格的運用に備えて、外部公開用サーバ、データ収集・解析サーバ、文献・画像解析サーバ等からなる計算機環境を整備した。外部公開用サーバは本事業で開発したサービスを公開するためのサーバで、ハード故障によるサービスの停止を避けるための一部冗長構成を有し、クアッドコア 12CPU、メモリトータル 96GB、ディスク容量 4TB からなるものである。データ収集・解析サーバは横断検索のクロールや公開サービスのコンテンツを作成するための計算を担うものであり、クアッドコア 24CPU、メモリトータル 192GB で、ディスクは文献・画像解析サーバと共用している。文献・画像解析サーバは、文献や画像の解析を行うために同時に大きな容量のメモリを使える共有メモリ型のサーバで、デュアルコア 8CPU、メモリ 128GB、ディスク容量 90TB のものである。



図 4.5 統合 TV：データベースとツールの動画教材

また、国内外の主要データベースでは、ウェブサービスが提供されていることが多い。これらをうまく利用するため、国内外でどのようなウェブサービスが用意されており、また、その稼働状況がどうなっているかを調査した。この調査結果に基づき、現時点のウェブサービスの問題点を洗い出した。そして、各サービスでやりとりされる生物学データは、フォーマットの違いなどにより簡単には相互利用できない状況であること、さらにサービス毎にメソッドの命名規則が異なることやドキュメントが不足しているといった問題を解決するため、ウェブサービスの変換サーバ **TogoWS** を開発した(図4.6参照)。これにより、ユーザは統一命名規則とデータ構造に基づいて、国内外のウェブサービスを一元的に利用することができる。また、**BioHackathon 2008** での議論を反映して、**BioMoby** 対応のための機能を **TogoWS** のプロトタイプ機能として作成した。さらに、産業技術総合研究所生命情報工学研究センター (CBRC) の協力を得て、CBRCの開発した配列解析プログラムについて、SOAP/WSDLによるウェブサービスを提供できるよう開発を行った。対象となる配列解析プログラムは、以下の7個のプログラムである：WoLF-PSORT, POODLE-W, POODLE-S, POODLE-L, TMBETA-NET, TMBETADISC-COMP, GRIFFIN。この開発により、これらのプログラム群をSOAPクライアントから利用可能になった。

国内外のウェブサービスを透過的に統合しワークフロー化を促進



- BioHackathonにより世界標準の規格を検討、国内のサービス連携を開始
- 相互運用を容易にし、主に国内のサービスを統合したワークフローを集積

図 4.6 TogoWS: 統合ウェブサービス

(2) 統合 DB プロジェクトポータルサイトの運用 (情報・システム研究機構)

サイトの運用に当たっては、各参画機関で作成されるウェブコンテンツ、各種ツールの集約のためにヒアリングを行い、情報交換の手法について検討し横断検索を中心にAPIの開発を行った。安定的なサービスの実現のためのサーバ設定業務により停電時以外の常時サービス運用を行った。また開発環境の外部サーバから機能の移転などを行い、パフォーマンスの向上に努めた。また、シングルサインオンサービス **OpenID** サーバを開発し、利用者がプロジェクト内統一IDでサービスを利用するための機能を備えた。**OpenID** とは、一つのIDで複数のサイトの認証が可能なサービスである。認証プロトコル仕様はすべて公開されているため、それを利用してサーバを構築した(図4.7参照)。これにより、統合データベースに関連する認証をこのサーバで一元的に行うことが可能となり、複数のサイト間を行き来するユーザの負担を減らすことができる。さらに、利用者が容易にデータベース情報を取得できるようにカタログ型のユーザーインターフェースのプロトタイプである生命科学分野のデータベース閲覧検索ユーザーインターフェースを開発し、そのインターフェースを用いてデータベースカタログ及び学協会検索を運用した。

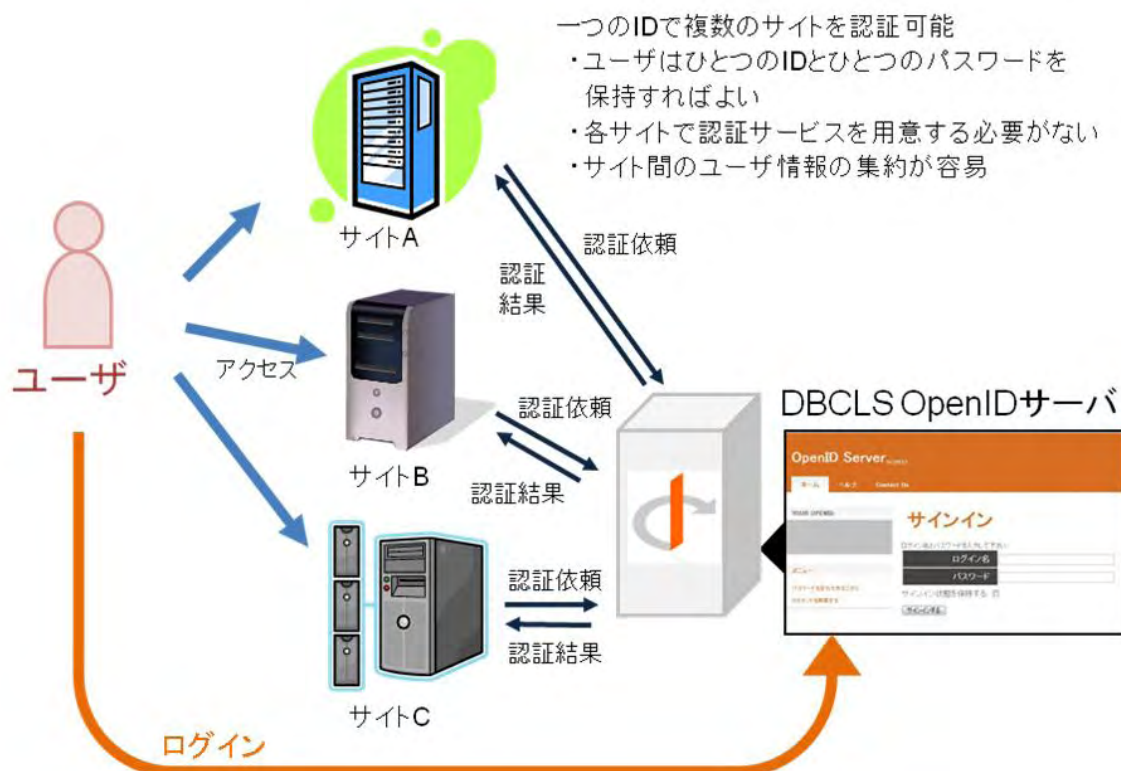


図 4.7 OpenID の概要

(3) 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、学会発表や講習会において統合DBプロジェクトのサービスを紹介した印刷物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。各参画機関に寄せられる意見や質問に対応するために、窓口となるメールアドレスinfo@dbcls.rois.ac.jpを用意しウェブサイトに掲載した。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公開するためのウェブサイトMotDBを公開するとともに、統合データベース講習会（平成20年2月25日於長浜バイオ大学、平成20年3月5日於科学技術振興機構）を開催した。

(4) ポータルサイトWINGproの運用 (科学技術振興機構)

個別DBについて、平成18年度当該事業の成果である、DB構築者および利用者からの意見を集約するためのシステムWINGproの公開を継続した。50余りのDBを追加し、平成20年3月現在420DBを収録している。利便性の向上として、ユーザが記事を投稿あるいは編集できる機能を平成19年7月に公開した。平成19年12月には、第30回日本分子生物学会年会・第80回日本生化学会大会 合同大会でポスター発表および紹介パンフを配布した。また、平成19年日本バイオインフォマティクス学会年会 (JSBi2007) でもパンフを配布した。



図4.8 WINGpro (http://wingpro.lifesciencedb.jp/)

(5) 事業サイト等の運用と広報(科学技術振興機構)

文部科学省「ライフサイエンス分野の統合データベース整備事業」等のウェブサイト運営し、これらのサイトを登載しているサーバ等の運用を行った。参画各機関の当該事業に関する研究開発動向や活動、ライフサイエンス政策の動向等を広く広報・情報発信をした。新規記事の追加や修正を行うほか、サイト内検索機能やRSS機能を追加し利便性を改善した。図を加え、見やすい表示となるようサイトの一部を変更した。また、記事の追加や修正のための編集システムの改良を行った。



図 4.9 ライフサイエンス分野の統合データベース整備事業サイト(その 1)
(http://lifesciencedb.mext.go.jp/)



図4.10 ライフサイエンス分野の統合データベース整備事業サイト(その2)
(<http://lifesciencedb.mext.go.jp/>)

なお、サイト内検索機能には、「ワードプラス」機能を追加した。これは漏れのない検索を支援する機能である（図 4.10）。具体的には、入力されたキーワードの検索結果画面に、「同義語アイコン」が表示され、そのアイコンをクリックするとキーワードの異表記語・同義語の一覧が表示される。一覧から適宜選択することで入力キーワードに同義語を追加した再検索が可能となっている。

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

(情報・システム研究機構)

我が国のプロジェクトDBの現状を調査し、それらをそれぞれの形態に従って分類し、それに基づいた受け入れ方法を検討する。また、どのデータベースから優先的に受け入れるべきかその優先順位付けの方法を立案し、それに基づきデータベースの受け入れと運用を試行する。

(科学技術振興機構)

データベースの受入に関して、レポジトリの制度等の調査および国際標準あるいは既存のメタデータについて調査する。調査結果の要約を公開する。受入に必要な研究開発の一部（設計段階）に着手する。

4. 2. 2 データベース受入の実施内容

(1) データベース受入（情報・システム研究機構）

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

国内のライフサイエンスデータベースに関する情報を漏れなく収集することを目的として、データベースカタログデータの補充を行った。内閣府連携施策群成果のリストに理研、ゲノム特定領域（新旧）、JST SNPポータル、JBIRC、CBRC、かずさDNA研究所、農業生物資源研究所等のリストを追加した。また、多くの予算が投入されてきたこの約10年間のライフサイエンス国家プロジェクトの事業や成果を明らかにする目的で調査を行い、プロジェクトの期間、概要、予算、代表研究者、公開の状況、文献等に関する情報を収集した(図4.11参照、<http://lifesciencedb.jp:8888/lldb.cgi?gg=project>)。

プロジェクト名	実施期間	プロジェクトの概要	データベース公開状況の分類	データベース	データダウンロード
ゲノムネットワークプロジェクト	2004-2008	遺伝子の発現調節機能に関わる網羅的な解析	一部共有 コンパニ内部共有	データベース	ダウンロード
タンパク3000	2002-2006終了	主要タンパク質約3000種の基本構造及び機能解明	公開		
未来開拓事業(ゲノム研究)	1996～1999(終了)ヒトゲノム 2000-2004(終了)ゲノム研究	ヒトゲノム解読と、その応用研究	一部共有 一部公開		
遺伝子多型研究	2000-2003終了	ヒトゲノム遺伝子中のSNP関連情報の取得と解析	共有	データベース	ダウンロード
オーダーメイド医療実現化	2003-2007終了	約30万人のSNPと薬剤の効果、副作用との関係解明			
理研多型	理研運営中期計画等に基づく予算措置のため、限定年限ではない	理研遺伝子多型センター(2007年度まで)→ゲノム医科学研究センター(2008年度以降)の運営			
理研ゲノム	理研運営中期計画等に基づく予算措置のため、限定年限ではない	理研ゲノム科学総合研究センター(2007年度までの)運営	一部共有 公開	データベース	ダウンロード
理研植物	理研運営中期計画等に基づく予算措置のため、限定年限ではない	理研植物科学研究センターの運営	一部共有 公開	データベース	ダウンロード
バイオインフォマティクス研究(バイオインフォマティクス推進事業)	2001-	生命科学分野の基幹データベースの構築・高度化	一部共有 公開	データベース	ダウンロード
統合データベース	2006-	生命科学分野DB戦略立案支援、ポータルサイト整備	一部共有 一部公開	データベース	ダウンロード

図4.11 ゲノム・ポストゲノムプロジェクト一覧の一部

2) ライフサイエンスデータベース受入方法の検討

データベースの分類に基づく受入優先順位の検討と受入方法の検討を行った。まず、収集したデータベースリストに基づき、データベースの対象を分類し、さらにDDBJなどの公開バンクに登録されていないオリジナルデータがあるかないかに基づいて、受入の優先順位の検討を行った。その結果、DNA配列や発現のデータが最もオリジナルデータの割合が多いことがわかった。ここでは、DNA配列の中でも日本が優位性を持つcDNAに着目し受入の試行を行うこととした。

次に、DB作成者との間の様々な状況に応じた受入方法の検討を行った。受入や連係の方法として、一括してデータを集める場合とデータベースが分散している場合に分けて検討した。データを集める場合は、1)データだけをFTPでダウンロード可能にする場合とデー

データベース自体を移設する場合、2)データベースの型を揃え、管理はそれぞれのデータベース構築者が運用する場合（データベース受入テンプレートの利用）3)データベースの型を揃え、統合した形で当センターが構築管理を行う場合、いわゆる集中型統合の場合を検討した。また、データベースが分散している場合の連係方法としては、1)横断検索、及び2)API等を用いる分散型の統合検索を検討した。

3) データベース受入システムの構築

受入を容易に行うためにテンプレートデータベースを作成した。これを用いることで、ユーザはウェブブラウザを用いて、エクセルなどで作成したCSVファイルを読み込み、関係データベースを簡単に構築することができる。さらに、検索インターフェースなど、ウェブ公開に必要なインターフェースが自動生成されるため、データ読み込み後すぐウェブ公開することができる。また、ウェブサービス(SOAP/WSDL)による検索APIも自動生成される(図4.12参照)。

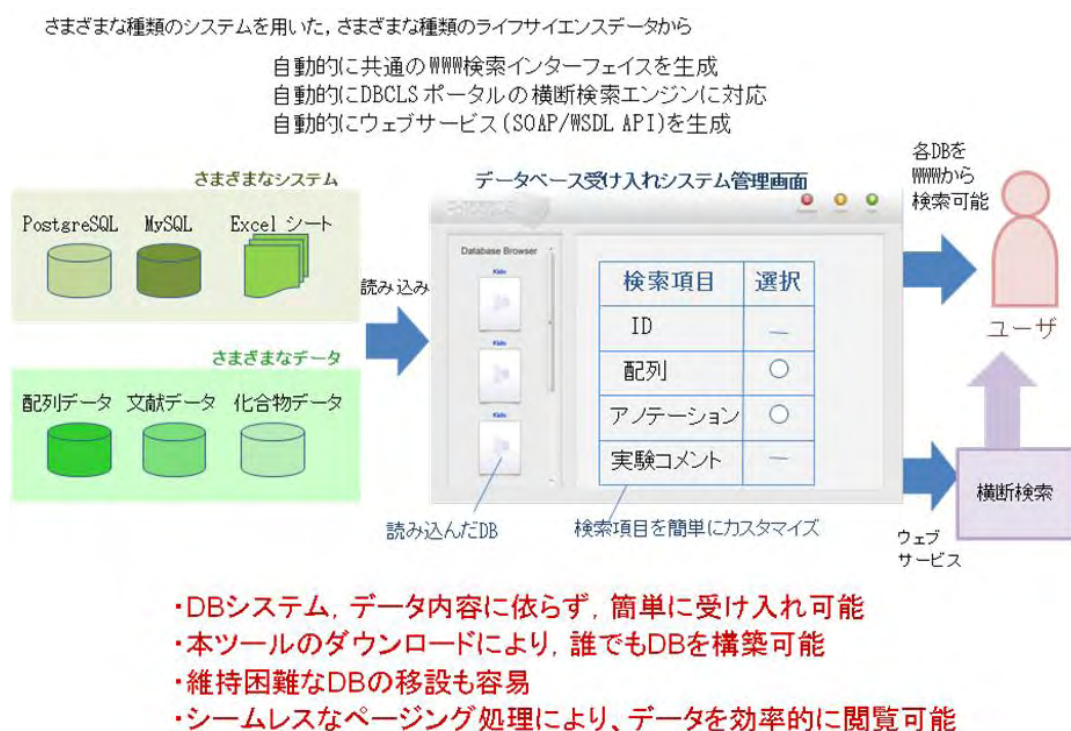


図4.12 データベース受入システムの概要

4) ライフサイエンスデータベース受入の試行

cDNA関連のデータベースとして、BodyMap、HGPD (Human Gene Protein Database)、FLJ Human cDNA DB (スプライシングバリエーションDB) とFLJ-DB (ヒト完全長cDNA)、DBTSS (Database of Transcriptional Start Sites)、及びMiBASE (トマトEST DB) とKafTom (トマト全長cDNA DB) の受入と連係方法を検討した。また、理

研補完課題のテーマ「植物オミックス情報および蛋白質構造情報」に関連したデータベースとして、KATANA（シロイヌナズナDB）の受入と連係方法を検討した。なお、BodyMapについては、データベース受入テンプレートシステムを用いた受入の試行を行った。試行内容は以下のとおりである。まず、BodyMapのEST配列情報を表形式に整理して、受入テンプレートシステムに導入した。BodyMapは、EST配列とそのクラスティング情報からなり、EST配列の表だけではクラスターの情報が見にくいため、クラスターの表を作成して受入テンプレートシステムに導入した。さらに、各クラスターの行からそのクラスターに含まれるEST配列のリストを表示する機能を設けた。これによって、二つの表データ間の関連を表現することが可能になった。一般的に、データベースは単純なリストとは限らず、今回のような他の表とのリンクや同じ表の中でのリンク、外部とのリンクなど様々なリンクが存在する。そこで、ここで実現した表間のリンク機能をさらに拡張していくことが次の課題であると考えられる。

5) ライフサイエンスデータベース受入に関するアンケート調査

本事業の分担機関や補完課題実施機関以外で整備・構築された国内データベースを対象として、データベースの内容、維持・更新状況、提供可能性及び提供時の問題点などを調査し、受入による統合化の方針策定に役立てるとともに、受入対象の候補の抽出を目的としてアンケート調査を実施した。データベースコンテンツと検索方法、データベースの経歴・管理状況、データベース維持の将来計画、提供可能性及び条件、提供方法、受入時の連絡先について質問し、65/104 (62.5%) の回収数（率）を得て、以下のようなアンケート結果を得た。

DBの利用条件では、大半は1次利用に対して制限を設けていないが、2次利用や再配布に対しては半数近くが制限を設けている。今後については、大半が現状維持・データの増強を検討しているが、平成20年度以降もプロジェクトが継続しているのは全体の30%程度にすぎない。また、半数以上が、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ている。提携条件としては、データ更新権限や著作権などの留保、利用条件の継承、統合DB事業による維持・管理・提携費用の負担を挙げる回答が多かった。

(2) データベース受入（科学技術振興機構）

データベースの受入に関して、「学術情報やデータの機関レポジトリ制度の調査」および「国際標準（論文やデータセットの所在情報発信）およびメタデータレポジトリシステム」、「ゲノムビューアをハブとしたデータベース受入」について調査した。「学術情報やデータの機関レポジトリ制度の調査」では、学術論文やデータを受入、公開している機関レポジトリ制度の構築、レポジトリライセンスの要件、我が国の大学での運用方針を調査した。例えば、レポジトリ運用指針では誰がデータを登録できるか、登録対象、登録されたコンテンツの利用、レポジトリ機関としての遵守事項、著作権に関する事項、免責、登録されたコンテンツの削除が取り上げられている。「国

国際標準およびメタデータレポジトリシステム」では、既に構築が試みられているメタデータレポジトリで使用されているメタデータ記述規則、標準通信規約、これらを利用したツールを調査した。国際標準としてネットワーク上の情報資源の所在情報を扱うメタデータ記述規則である Dublin Core、生物の標本、観察データのデータ交換形式である Darwin Core、生物学コレクションデータバンクの情報を最大限に共通化する目的で策定されている ABCD Schema および米国国立癌研究所(NCI)の癌研究ネットワーク情報基盤システムの一部である caCORE を対象とした。この統合データベースプロジェクトで利用できるか（汎用性）、使用する際のコスト面（再利用性、運用）、ライフサイエンス分野での実例の多さや認知度、セキュリティなどを比較した。この調査結果から、メタデータとして記述する項目「メタデータ要素（エレメント）」とその記述規則を収録するメタデータ要素レポジトリをメタデータレポジトリシステムの研究開発の一部（設計段階）として着手し、テストサイトを公開した(図4.13、図4.14)。

「ゲノムビューアをハブとしたデータベース受入」では、テキスト検索や配列の相同性検索とは異なり、諸外国の多くのサービスで行われている特定配列のゲノム上の位置やその位置の相互の関係から検索できるゲノムビューアについて調査した。ゲノムビューアを「ハブ」とする案内機能、データの更新方法、生物学研究者に提供するデータや機能について新規遺伝子関連情報、扱うべき生物種、検索機能を調査し、具体化する方法をまとめた。また、平成18年度の調査「ライフサイエンス分野のポータル連携のための調査」で扱った Jabion のゲノムビューア

(<http://www.bioportal.jp/genome/>) に科学技術振興機構のバイオインフォマティクス推進事業の研究開発課題の成果である HAL を受け入れる試みを実施し、HAL の追加

(<http://www.bioportal.jp/genome/cgi-bin/index.cgi?org=hal>) を実現した。

The screenshot shows the MDeR website interface. At the top, there is a logo and the text "The metadata element repository in life sciences" and "ライフサイエンス分野のメタデータ要素レポジトリ". Below this is a navigation bar with links: "ホーム・検索", "概要", "リンク", "ヘルプ", and "利用条件". On the left, there is a sidebar with a "収録一覧" (List of collections) section, which includes "国際標準" (International standards) and "その他" (Others). Under "国際標準", there are links for "DublinCore", "DarwinCore", and "ABCDSchema". Under "その他", there are links for "MorningGlory", "LipidBank", and "UniProtKB". The main content area is titled "ホーム・検索" (Home Search) and contains a "検索" (Search) section. It has a search bar with a placeholder "入力例 Author" and a dropdown menu set to "ALL". There are radio buttons for "完全一致" (Exact match) and "部分一致" (Partial match). Below the search bar are buttons for "検索" (Search) and "クリア" (Clear). At the bottom, there is a "お知らせ" (Notice) section with a date "2008年3月26日" and the text "ライフサイエンス分野のメタデータ要素レポジトリ(MDeR)を公開しました。"

図 4.13 ライフサイエンス分野のメタデータ要素レポジトリ(その 1)
(<http://mdcr.jst.go.jp/>)



比較結果表示

● 比較表示

No.	1	2	3
□ Data Element			
Data Element Public ID	6460	5479	204
Data Element Long Name	LipidData REFERENCES AUTHOR string	reference citation authorList person name nameListType	Resource ScientificNameAuthor string
Data Element Version	1.0	1.0	1.0
Data Element Definition	author of journal article		The author of a scientific name. Author string as applied to the accepted name. Can be more than one author (concatenated string). Should be formatted according to the conventions of the applicable taxonomic discipline.
Data Element Begin Date			
Data Element Example	Pike, J.E., Licoin, F.H., and Schneider, W.P.	Kitada T.	
Data Element Comment			
Data Element Context Name	Lipid Bank	UniProtKB	DarwinCore2
Data Element Context Version	1.0	1.0	1.2
□ Object Class			
Object Class Public ID	6463	5481	207
Object Class Long Name	entry lipid reference	reference citation authorList person	Resource
Object Class Version	1.0	1.0	1.0
Object Class Definition	lipid data class	reference citation class	
Object Class Context	Lipid Bank	UniProtKB	DarwinCore2

図4.14 ライフサイエンス分野のメタデータ要素レポジトリ(その2)
(<http://mder.jst.go.jp/>)

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2つの演習および次年度以降実施するための環境整備を実施する。

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。演習を完了するまでには、受講者の能力と受講可能時間に応じて最短で3ヶ月、最長で1年間の時間を予定している。

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のための環境構築

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平成 20 年度より独創的サーバ構築演習の開講を予定しており、そのための計算機セットアップを平成 19 年度に準備する。なお、独創的サーバ構築演習では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。

(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 19 年度は、このような DB 高度利用者に必要なスキルを詳細に検討し、DB 高度利用者の養成に必要なカリキュラムの開発を行う。平行して、この養成カリキュラムに則した授業、演習を実施し、人材の養成を行うとともに、カリキュラムの評価を行う。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、ライフサイエンス統合データベースセンターと連絡を密に取り、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、ライフサイエンス統合データベースセンター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。ゲノムアノテーションの実習授業で使用したテキストを基に、アノテータ養成のための自習用教材を作成し、インターネット上での公開を行う。

アノテータ・キュレータの具体的な教育プログラムとして、2 年次の学生に関しては約 200 名を対象に、「ライフサイエンス分野のデータベースの統合利用」に関する 180 分の計算機実習授業を 4 回程度行う。3 年次の学生に関しては約 45 名対象に、「ゲノムアノテーションとそれに必要な計算機基礎技術」に関する 180 分の実習授業を 15 回程度行う。

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

ライフサイエンス分野の大量情報を対象に、多様なアノテーション情報付加と情報可視化に必要となる技術の実践的な教育を行い、作成したデータベースを本統合 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。この教育で使用したテキストを基に自習用教材

を作成し、インターネット上での公開を行う。

多様なアノテーション情報の付加に関する具体的な教育プログラムとして、3年次の約200名の学生を対象に、混合微生物由来の大量な遺伝子候補からの遺伝子探索とアノテーション付加を実践例として「持続可能型社会への貢献遺伝子データベース」の作成を行い、データベースを公開する。4年次の学生の3名程度について、1年間の卒業研究教育として、自己組織化マップを含む既存の可視化技術による大量遺伝子のアノテーション情報の可視化を行う。

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

上記の二つの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ハブとする、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をめざす。本事業のアノテーション作業に活用すると同時に、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3回程度開催して、大学や研究機関への成果の普及を図る。

4. 3. 2 人材育成の実施内容

(1) DB構築者の養成（東京大学）

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。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。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/ensemblmirror/index.php?Introduction>

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る効果があった。演習総時間は90時間で平成19年7月下旬～9月上旬の約2ヶ月間であった。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/upbsb2007/>

3) 独創的サーバ構築演習のための環境整備

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平成20年度より独創的サーバ構築演習の開講を予定しており、そのためのそのために計算機と二次記憶装置を増設した。それにともなって電源を増強する必要性が生じ、さらに増設された計算

機資源を動作させるソフトウェアも必要となった。これらのセットアップを平成 19 年度に準備した。

(2) DB 高度利用者の養成 (お茶の水女子大学)

平成 19 年度の業務において、DB 高度利用者に必要なスキルを詳細に検討した結果、DB 高度利用においては情報科学、生物学に立脚した高度な知識が必要であり、教育対象者の背景知識に応じて異なった基礎科目を用意し、その後 DB 高度利用者に必要な専門教育を行う事が必要だと判明した。そこで、基礎科目として、生命系出身者には DB 高度利用者に必要となるデータベース、ネットワークの知識を補う科目を、情報系出身者には、遺伝学の基礎知識を補うカリキュラムを策定した。これら知識を補完した上で、DB 高度利用者として活躍するには、近年のライフサイエンスにおける大規模データを解析する基礎として「データ解析の基礎」、現代はインターネット上にデータがあり、これらデータを活用する手法を学ぶ「ネットワークを通じたライフサイエンスデータの利用」、大規模データの実践的な解析手法として「データマイニング技術」のスキルが DB 高度利用者に必要であることが分かった。平成 19 年度は、これらの項目に沿って、カリキュラム開発を行った。開発したカリキュラムは、<http://togodb.sel.is.ocha.ac.jp/> を通じ公開している。また、カリキュラム開発に平行して、「ネットワークを通じたライフサイエンスデータの利用」に関し、18 人の修士課程、博士課程の受講生に対し、授業及び演習を行った。受講生へのアンケートにより、カリキュラムの評価を行ったところ、「プログラムをもっと早く書けるようになれば、データベースを利用して自分の研究専用に検索や比較ができるようになる、と思った。」との回答を得ている。また、受講生の過半数から既存のプログラムを改良するかプログラミングを自分で行って、データベースの利用に役立てたいとの回答を得ており、本カリキュラムにより DB 高度利用者となる際、大きなハードルとなるであろうプログラミングスキルを習得するモチベーションが得られていると評価できる。

平成 19 年度は、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 19 年度の授業・演習受講生の内 2 名は本自習用教材の作成と公開において中心的な役割を果たし、本人材育成により統合 DB に関わる人材を育成できつつあることが分かる。平成 20 年度以降、授業・演習において明らかとなったカリキュラムの過不足部分を反映し、本自習教材の充実を計画している。

(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 2 回生 220 名を対象としたアノテータ・キュレータ養成を目的に、ライフサイエンス分野のデータベースの実体や意義を習得する目的で、10~11 月 (毎週月曜 180 分 × 6 週) に「バイオ分野のデータベースの統合的な利用」との課題で、JST や統合 DB の関係機関が

作成した、日本語説明のあるバイオデータベース用のポータルサイトを利用して実習を行った。なお、この実習授業用に作成したテキストを、自習用教材としてインターネットを介して公開しており、希望によっては印刷物としても配布している。

本学 3 回生 40 名を対象として「ゲノムアノテーションとそれに必要な計算機基礎技術」の習得を目的に、10～1 月（毎週 木・金曜 各 180 分×16 週）に、機能アノテーションがされていない高度好熱古細菌のゲノム配列を、統合 DB の参画機関より入手し、微生物ゲノムアノテーションの実習を行った。作成したテキストは統合 DB へ提供した。

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

多様なアノテーション情報の付加に関する実習として、6 月～7 月（毎週水曜 180 分 × 8 週）に、本学 3 回生 220 名に対して、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（混合微生物ゲノム解析）で得られている環境微生物ゲノム配列を探索対象として、持続可能型社会への貢献遺伝子の探索とデータベース構築を行った。自然環境の浄化や保全に役立つ、広義には持続可能型社会の実現に貢献できる可能性を持つ遺伝子を各学生が独自にテーマ別に設定し、その遺伝子がどの程度の種類の新規生物種にも存在しているのかを、メタゲノム配列を対象に網羅的に探索し、機能と生物系統のアノテーション情報を付加して、データベース化を行った。主として、環境ホルモン分解、重金属除去、PCB 分解、ダイオキシン分解、石油分解、土壌浄化、バイオエタノール、農薬分解、硝酸性窒素除去、プラスチック分解、有機リン除去等に能力を発揮できる遺伝子類やタンパク質類を探索した結果、公的 DNA データベースでは機能が未同定として残されていた遺伝子を約 2000 件発掘した。さらに、4 回生 2 名が、7 月～2 月までの卒業研究としてアノテーションデータを精査し、新たに発見できた約 5000 遺伝子についても機能と生物系統のアノテーションデータを加えて、合計で 7000 遺伝子類を「持続可能型社会への貢献遺伝子データベース」へ登録し、学生の名前入りでインターネット上での公開を行った。この実習授業用に作成したテキストを、インターネットを介して公開し、希望によっては印刷物としても配布をおこなった。並行して、4 回生の 3 名が 4 月～2 月の卒業研究として、3 月は大学院進学前の自主的な研究として、自己組織化マップを用いた生物系統と機能アノテーションとその可視化を行った。自己組織化マップで得られた結果を、インターネットを介して公開している「持続可能型社会への貢献遺伝子データベース」へ追加した。

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

学部レベルの人材養成プログラムで得られたアノテーション成果の質の向上を目的として、専門知識を有する 3 名のシニア世代研究者を中心に、作業は作成途中のデータベースについてアクセス権を付与してオンラインでの書き込みを可能とし、在宅で行った作業についての件数に基づく謝金の支払いを行うシステムを整備し、分散型の高度なキュレーションモデルの確立を行った。平成 19 年度は、シニア世代研究者チームに 4 回生の 1 名が参加して、高

度なキュレーションのモデルとして、「エキスパートがキュレートした tRNA 遺伝子データベース」を構築し、そのプロトタイプ版を公開した (<http://trna.nagahama-i-bio.ac.jp>)。

長浜バイオ大でのアノテータ・キュレータ養成用に作成したテキストを用いて、シニア世代研究者が協力して、学外者を対象にした以下の講習会を開いた。1) 平成 19 年 10 月 13、14 日のゲノム広場(大阪)、2) 平成 19 年 12 月 23 日のミニゲノム広場(福岡)において、「持続可能型社会への貢献遺伝子データベースの構築」のテキストを用いて、合計で 30 名の参加者に講習を行なった。これらのゲノム広場では、合計で 120 部のテキストの配布も行なった。3) 平成 20 年 1 月 25、26 日に他大学や企業を含む研究機関の計 21 名を対象者として、遺伝子機能と生物系統のアノテーション手法に関する講習会を開催した。参加者である大阪大学微生物病研究所の研究者は、マラリア原虫のゲノムアノテーションにおいて活用することを予定している。

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、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。研究運営委員会のメンバーは下記の通りである (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、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
金岡 昌治	大日本住友製薬（株）研究本部副本部長
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
田畑 哲之	（財）かずさ DNA 研究所副所長
松原 謙一	（株）DNAチップ研究所代表取締役社長
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
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/ 東京大学教授
永井 啓一	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 特任教授

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を開催し、平成19年度の計画、補完課題の公募内容、実施状況と進め方、契約・著作権問題、個人情報問題、海外連携、3年後の在り方などを討議した。

第一回研究運営委員会（平成19年4月27日）

第一回研究運営委員会作業部会（平成19年6月15日）

第二回研究運営委員会作業部会（平成19年10月1日）

第二回研究運営委員会・DBCLSお披露目会（平成19年10月5日）

第三回研究運営委員会作業部会（平成19年12月6日）

第三回研究運営委員会（平成20年1月8日）

第四回研究運営委員会作業部会（平成20年3月27日）

（2）分担機関、補完課題実施機関との連絡会

中核機関と分担機関の実施内容と連携の進め方、実施状況のフォローなどを中心に討議した。全体会議を3回と、個別会議を各1回開催した。また、補完課題の事業に関しては、各

機関と密接な打合せを行いながら進めた。また、特に蛋白立体構造ならびに植物関係の進め方に関する議論は、それぞれ中村春木教授、中村保一室長を中心に中核機関を含めたグループを構成して進めた。会合の一覧は以下に示すとおりである。

日付	時間	内容	場所
H19.06.22	10:00-12:00	第1回中核機関・分担機関研究連絡会	ROIS 会議室
H19.08.09	10:00-12:00	第2回中核機関・分担機関研究連絡会	ROIS 会議室
H19.08.29	15:00-16:00	東大グループとの打合せ	DBCLS
H19.09.28	15:30-17:30	第3回中核機関・分担機関研究連絡会	東大理会議室
H19.10.11	11:00-12:30	産総研糖鎖医工学センターとの打合せ	DBCLS
H19.10.18	18:00-19:00	分担機関との個別打合せ(東大)	東大医学部
H19.10.19	12:30-14:00	九工大との打合せ	DBCLS
H19.10.19	14:00-15:30	理研との打合せ	DBCLS
H19.11.07	10:00-12:00	補完課題(蛋白構造関係)打合せ	DBCLS
H19.11.15	10:00-11:00	補完課題(植物関係)打合せ	DBCLS
H19.11.30	10:00-12:00	分担機関との個別打合せ(京大)	DBCLS
H20.01.07	10:00-11:00	分担機関との個別打合せ(東京医科歯科大)	東京医科歯科大
H20.01.31	13:00-15:00	理研との打合せ	理研
H20.02.06	13:00-15:00	理研(蛋白構造関係)打合せ	播磨理研
H20.02.18	18:00-21:00	補完課題(蛋白構造関係)打合せ	東京三菱ビル
H20.02.22	13:00-18:00	第一回日本糖鎖統合 DB 会議	東京コンファレンス センター品川
H20.03.06	13:00-15:00	遺伝研との打合せ	遺伝研
H20.03.13	14:00-17:00	産総研糖鎖医工学センターとの打合せ	産総研
H20.03.14	15:00-16:00	九工大との打合せ	DBCLS
H20.03.14	17:00-19:00	東大グループとの打合せ	DBCLS

(3) 中核機関内の連携のための会合

各機関の具体的課題の解決と全体的な調整、意識合わせを図るために、各機関と年数回の打合せを行った。会合の一覧は以下に示すとおりである。

日付	時間	内容	場所
H19.05.24	15:30-17:00	講習会計画立案、人材育成計画に関する打合せ	DBCLS
H19.06.06	17:00-19:00	JST との打合せ (DB 受入れ関係)	JST
H19.06.07	15:30-18:00	CBRC、奈良先端大との打合せ(共通基盤技術)	DBCLS
H19.07.04	15:30-17:00	モデル生物有用微生物チーム打合せ	DBCLS
H19.07.10	15:00-18:00	統合データベース支援拡大会議	DBCLS

H19.07.23	16:00-18:00	バイオポータル合同会議	DBCLS
H19.08.21	14:00-16:00	JST との打合せ (DB 受入れ関係)	DBCLS
H19.08.29	18:00-20:00	CBRC との打合せ(共通基盤技術)	DBCLS
H19.09.14	15:00-17:00	奈良先端大との打合せ	奈良先端大
H19.11.05	18:30-20:30	奈良先端大との打合せ	DBCLS
H19.11.15	16:00-18:00	モデル生物有用微生物チーム打合せ	DBCLS
H19.11.27	15:00-17:30	CBRC、かずさ DNA 研との打合せ	DBCLS
H19.12.03	17:00-18:30	JST との打合せ (文献検索)	JST
H19.12.07	10:00-11:00	JST との打合せ (文献検索)	DBCLS
H19.12.13	13:00-15:00	九州大学との打合せ	DBCLS
H19.12.18	14:00-16:00	かずさ DNA 研との打合せ	DBCLS
H19.12.27	13:30-16:30	人材育成関係合同打合せ	DBCLS
H20.01.09	18:40-21:00	かずさ DNA 研との打合せ	DBCLS
H20.01.10	13:00-15:00	奈良先端大浅原先生講演と打合せ	DBCLS
H20.01.22	14:00-15:00	CBRC との打合せ	CBRC
H20.02.04	15:30-18:00	モデル生物有用微生物チーム打合せ	DBCLS
H20.02.25	16:00-17:00	かずさ DNA 研との打合せ	DBCLS
H20.02.26	15:00-17:00	農業生物資源研との打合せ	農資研
H20.03.21	13:00-17:00	大阪府立成人病センター研究所との打合せ	成人病センター

なお、共通基盤技術に関わる東大医科研、かずさ DNA 研との打合せは週一回定例で実施

(4) 外部有識者からの意見聴取

プロジェクトの推進に資するため、以下に示す先生方との打合せやセミナーを開催した。

Martin Yuille先生(Manchester大 教授)、 隅蔵康一先生 (政策研究大 准教授)
金子周司先生(京大薬学部 教授)、 簗島伸生先生(浜松医大 教授)
浅原正幸先生(奈良先端大 助教)、 増井徹先生(医薬基盤研 主任研究員)
加藤和人先生(京大生命科学 准教授)、 神嶋敏弘先生(産総研脳神経情報研究部門)
市瀬龍太郎先生(国立情報研 准教授)、 向井秀幸先生(神戸大 准教授)
権藤洋一先生(理研GSC PJ副ディレクター)、 西村邦裕先生(東大情報理工 助教)
星山大介先生(東大新領域 助教)、 田中英夫氏(トミーディジタルバイオロジー)
野口 祐子氏(弁護士)、 末吉互氏(弁護士)

5. 3 その他セミナー等

(1) BioHackathon (情報・システム研究機構、生命情報工学研究センター)

2. 2 で記載した通り、国内外の研究機関等関係者との連絡調整の一環として、平成20年2月11日～15日の5日間、ウェブサービスの統合を目指した現状調査・情報交換および技術開発を目的として、BioHackathon2008を開催した。本会議は、情報・システム研究機構と生命情報工学研究センターの協力のもとで開催した。

以下に、初日に開催されたの講演会の内容と各分科会の議論の内容を示す。

1) 講演会

- Objectives of this hackathon and current status of Japanese web services - Toshiaki Katayama
- From Web API for Biology (WABI) to Semantic Web API for Biology (SABI) - Hideaki Sugawara
- Current status of the BioMOBY project and vision for the future directions - Mark Wilkinson
- Moby, Legacy Apps and the Semantic Web - Paul Gordon
- Data, Services and Computational Resources Integration at the INB - Oswaldo Trelles
- The EMBRACE project and WS-I standard - Jan Christian Bryne
- Soaplab2 project to wrap up command line packages - Martin Senger
- Taverna (part of myGrid project) - Tom Oinn
- Generation Challenge Program effort at building interoperability - Richard Bruskiewich

2) 分科会の構成と議論の内容

- OpenBio、Web サービスのための Semantics と Ontology
語義、プロトコル、ドメインモデル、データタイプ、サービスタイプ (含 BioMoby 等)、API 等。特に語義の重要性が再確認され、異なるサービス間での同一語義の使用に関して議論された。
- 標準化
Glycoinformatics、相互作用ネットワーク、テキストマイニング、Phyloinformatics の各分野で取り扱うデータが異なるため、それぞれの分野にて標準化のための議論がなされた。
- 非同期 Web サービス
同期 Web サービスのみならず、非同期のサービスも必要であるため、そのためのアプリケーション、環境、方法に関して議論された。
- ラージデータ

バイオでは頻繁に巨大データを扱うが、SOAP では扱えない等の問題がある。その取り扱い方法に関して議論された。

- セキュリティ

Web サービス提供では当然セキュリティを確保する必要がある。認証方法、セキュリティレベル、計算機資源保護等に関して議論された。

- サービス記述

Web サービスに対する適切な説明や、ユーザがそのサービスを迅速に探し当てるためにはどうすればよいか等に関して議論された。

- OpenBio

バイオデータを RDBMS に格納し検索するための標準モデル (BioSQL)、既存のバイオプロジェクトに関して議論された。

(2) 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 7 回 バイオインフォマティクスへの招待

日時：平成 19 年 6 月 20 日

場所：お茶の水女子大学

参加人数：22 人

タイトル：テキスト処理して生命科学

発表者：山本 泰智

タイトル：KEGG データベースの進化と、統合・標準化

発表者：片山 俊明

第 8 回 バイオインフォマティクスへの招待

日時：平成 19 年 9 月 3 日

場所：お茶の水女子大学

参加人数：37 人

タイトル：選択的スプライシングと植物の環境適応

発表者：飯田 慶

第 9 回 バイオインフォマティクスへの招待

日時：平成 19 年 10 月 19 日

場所：お茶の水女子大学

参加人数：29 人

タイトル：バイオインフォマティクスを用いた RNA 研究による生命現象の理解

発表者：程 久美子

第 10 回 バイオインフォマティクスへの招待

日時：平成 19 年 12 月 6 日

場所：お茶の水女子大学

参加人数：16 人

タイトル：バイオリソース事業におけるインフォマティクスの役割

発表者：深海 薫

第 11 回 バイオインフォマティクスへの招待

日時：平成 20 年 2 月 20 日

場所：お茶の水女子大学

参加人数：31 人

タイトル：システム進化生物学：

生命システムのグランドデザインの解明を目指して

発表者：荻島 創一

第 12 回 バイオインフォマティクスへの招待

日時：平成 20 年 3 月 7 日

場所：お茶の水女子大学

参加人数：29 人

タイトル：生物学者のための生命情報解析ツールの開発

発表者：田村 浩一郎

(3) アノテータ、キュレータ養成のための運営委員会（長浜バイオ大学）

長浜バイオ大学では、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た。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。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」を構築し、公開を予定している。

① 2007.6.15

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、木ノ内誠、
和田千恵子、平賀壮太、香川弘昭、駒野照弥、学部学生（8 名）

内容：統合データベースに関わる長浜バイオ大の全体活動のミーティング

② 2007.8.20-21

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（2名）

内容：第1回 tRNA 遺伝子データベース構築会議

③ 2007.10.5-6

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（2名）

内容：第2回 tRNA 遺伝子データベース構築会議

④ 2008.1.25-26

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、和田千恵

子、香川弘昭に加えて、外部研究機関の研究者と学部学生（21名）

内容：ライフサイエンス統合データベース講演会

⑤ 2008.1.27

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、

学部学生（2名）

内容：第3回 tRNA 遺伝子データベース構築会議

⑥ 2008.3.14-15

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（2名）

内容：第4回 tRNA 遺伝子データベース構築会議

**（4）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、
科学技術振興機構)**

4. 1の（3）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演会

日時：平成20年1月25日 午後1時半～午後5時半

場所：長浜バイオ大学

参加者：外部研究機関の研究者と学部学生（21名）

対象：長浜バイオ大学学生並びに興味のある外部研究機関の研究者

内容：

1. ライフサイエンス分野の統合データベースプロジェクト（講師 川本祥子）
2. バイオデータベース構築（講師 中谷洋一郎）
3. ゲノムアノテーションとは何か：その付けられかた／活用方法の紹介（講師 中村保一）

4. バイオデータマイニング入門（講師 瀬々 潤）
5. 使い倒し系バイオインフォマティクスによる遺伝子発現情報解析（講師 坊農秀雅）
6. テキストを対象とした生命科学研究（講師 山本泰智）

②統合データベース講習会：AJACS 東京

日時：平成 20 年 3 月 5 日 午前 10 時～午後 5 時半

場所：科学技術振興機構 東京本部

参加者：20 名

対象：学部 3 年生や修士 1 年生、博士後期課程 2 年生を始めとした統合データベースセンターに就職する興味のある人。統合データベースが実際にどのように作成、維持管理されているかに興味のある人

内容：

1. 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
(講師 坊農秀雅)
2. かずさにおける新しいアノテーションの試み（講師 岡本忍）
3. ゲノムアノテーションとは何か：その付けられかた／活用方法の紹介（講師 中村保一）
4. 統合データベースへの期待（講師 樋口千洋）
5. 講習：ゲノム情報リテラシー入門（講師 中村保一、その他ライフサイエンス統合データベースセンター員）

6. 成果のまとめ

分子生物学以降ライフサイエンスの成果の主体は、遺伝子や蛋白質に関するデータや言葉を使った生命現象の記述である。この膨大な成果はいまやデータベースの助けなしに整理することも利用することも不可能である。従って DB 統合の効果はゲノム研究やポストゲノム研究に限定されず、ライフサイエンス全体に波及する。中核機関では、統合 DB の目標として、網羅性の向上、利便性の向上、持続性の担保を挙げている。網羅性については、生物やデータの種類に関する網羅性と併せて、実験データから文献データまでという学術情報の整理の階層に関する網羅性の向上も目指している。特に分子データの統合にとどまらず、分子データ・文献データの統合も目指している。これについては、特にこれまで利活用の遅れていた日本語文献に関して、総説誌から学会要旨、学協会情報、日本語特許明細書に至るまで横断的に検索可能にした点は意義があると考え。特に、一定の制限はあるものの商業誌の全文検索を可能とした点は、将来への大きな第一歩と考える。利便性の向上に関しては、DB の所在が分からない、どう使ったらいいか分からないというユーザからの二つの大きな要望に対して、生命科学データベースカタログと統合 TV を開発、公開した。前者については、現状把握可能な国内 DB のほとんどを網羅するという平成 19 年度の目標をクリアするものである。また、後者についてもすでに 50 以上のサイトに関する動画コンテンツを公開済みで、本

プロジェクトの公開サービスの中でもアクセスの多いものとなっている。また、この他に検索の利便性、精度向上を支える種々の辞書やオントロジーも開発、公開している。なお、本プロジェクトでは公開するデータをすべてダウンロード可能とする方針をとっており、研究面での利用者だけでなく、DB 構築者の役に立つことも狙っている。持続性の担保については、維持困難な DB を中心とした DB 受入れについて、DB 保有者アンケートや受入れスキームの検討を行う中から、当初の予定である 5 個を上回る受入れ候補 DB を選定した。これらについては順次受入れ中である。また、DB の受入れ、公開に伴う権利関係などの制度面の問題についても検討を開始し、少なくとも将来のプロジェクトに関してはデータ提供に関する一步を踏み出すことができた。しかしながら中核機関の努力だけでは限界があり、知的財産権や個人情報に関する制度面の環境整備も望まれる。

引き続き、生命科学データベースカタログ、横断検索および統合 TV などのメニューの充実を図り、また、受入れ DB の拡充やこれをベースに高度な検索機能を付加し、さらには統合利用へむけた種々の基盤的な技術課題を克服することで、より多様な背景知識をもった研究者や技術者がストレスなくプロジェクト研究やバイオインフォマティクスの成果を有効活用できる情報基盤の本格的実現に向けた環境が整うものとする。そうした観点からは、ウェブサービスやワークフローに関する基本技術、及び専門用語を獲得するための自然言語処理技術などの共通基盤技術の開発についても一定の成果があがったものとする。また、アノテーション、キュレーションという点では、キュレーションを支援するためのツールやアノテーション情報の集積を可能とするシステムを開発した。またシニア研究者によるキュレーションモデルの構築といった新しい試みも行った。これらは、今後統合されていくデータベースに質の高いアノテーション、キュレーション情報を蓄積していくための重要な成果と考える。また、持続性の担保という点で重要な人材育成についても、DB 構築者、DB 高度利用者、アノテータ、キュレータ教育という複数の観点からの教材の開発と人材育成の実践を行った。育成された人材を継続的に積極的に活用していくことが今後の課題と考える。

最後に、統合 DB を広く活用してもらうためには広報活動も重要な要素である。本年度の成果として、プロジェクトの紹介や成果の公開サイトも整えることができた。これらをベースに、さらに多面的な広報活動を行うことにより、DB 利用者層が格段に拡大され、また利用機会が増大することでライフサイエンス分野全体にプロジェクト研究の成果還元が進むことが期待できる。

7. 成果の外部への発表

次ページ以降の添付様式を参照されたい。

8. 実施体制

別表 1、2、3 を参照されたい。

添付様式

論文寄稿

業務コード	実施 年度	和誌/ 洋誌	論文タイトル	発表者名	発表誌名	巻	号	ページ	掲載年月	メモ
07046015	19	和誌	生命科学系データベース統合化の背景	大久保公策	蛋白質核酸酵素(共立出版)	52	9	1027-103 1	平成 19 年	連載企画「ライフサイエンス分野の統合データベース」
07046015	19	和誌	統合データベースセンターがめざすもの	高木利久	同上	52	11	1388-139 3	平成 19 年	同上
07046015	19	和誌	統合データベースプロジェクトのサービスとその利用法	川本祥子、坊農秀雅	同上	53	3	281-287	平成 20 年	同上
07046015	19	和誌	利用の立場からのコメント	高木利久	同上	53	5	686-691	平成 20 年	同上
07046015	19	洋誌	OReFiL: an online resource finder for life sciences	Y. Yamamoto and T. Takagi	BMC Bioinformatics	8		287	平成 19 年	
07046015	19	洋誌	BioCompass: A novel functional inference tool that utilizes MeSH hierarchy to analyze groups of genes	T. Nakazato, T. Takinaka, H. Mizuguchi, H. Matsuda, H. Bono and M. Asogawa	<i>In Silico Biology</i>	8		6	平成 19 年	
07046015	19	和誌	Ensembl (Ensembl Genome Browser)	坊農秀雅	バイオデータベースとウェブツールの手とり足とり活用法(羊土社)			76-83	平成 19 年	
07046015	19	和誌	Reactome	河野信	同上			175-180	平成 19 年	
07046041	19	洋誌	A Large-scale Protein protein Interaction Analysis in <i>Synechocystis</i> sp. PCC6803.	Sato, S., Shimoda, Y., Muraki, A., Kohara, M., Nakamura, Y. and Tabata, S.	<i>DNA Research</i>	14	5	207-216	平成 19 年 11 月	
07046041	19	洋誌	Complete genomic structure of the bloom-forming toxic cyanobacterium <i>Microcystis aeruginosa</i> NIES 843.	Kaneko, T., Nakajima, N., Okamoto, S., Suzuki, I., Tanabe, Y., Tamaoki, M., Nakamura, Y., Kasai, F., Watanabe, A., Kawashima, K., Kishida, Y., Ono, A., Shimizu, Y., Takahashi, C., Minami, C., Fujishiro, T., Kohara, M., Katoh, M., Nakazaki,	<i>DNA Research</i>	14	6	247-256	平成 19 年 12 月	

				N., Nakayama, S., Yamada, M., Tabata, S. and Watanabe, M.M.						
07046089	19	洋誌	Reconstruction of the Vertebrate Ancestral Genome Reveals Dynamic Genome Reorganization in Early Vertebrates.	Yoichiro Nakatani, Hiroyuki Takeda, Yuji Kohara, and Shinichi Morishita	<i>Genome Research</i>	17	9	1254-1265	平成 19 年	
07046089	19	洋誌	Vertebrate genome evolution examined by comparing the human and fish genomes.	Yoichiro Nakatani and Shinichi Morishita	<i>Encyclopedia of Life Sciences</i> , John Wiley & Sons					in press
07046089	19	和誌	比較ゲノムを支える情報学:脊椎動物ゲノム進化を推定するロジック	森下真一, 中谷洋一郎	細胞工学別冊「比較ゲノム学から読み解く生命システム」		9		平成 19 年	
07046089	19	和誌	ゲノム進化のデータベース	中谷 洋一郎	『実験医学』増刊号 生命研究への応用と開発が進むバイオデータベースとソフトウェア最前線		4		平成 20 年	
07046091	19	和誌	ライフサイエンスにおけるデータベース構築のための人材育成	瀬々 潤, 池村淑道	蛋白質核酸酵素	53	1	87-93	平成 20 年 1 月	
07046077	19	和誌	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も含めて	池村 淑道	日本化学会情報化学部会誌	26	1	1-4	平成 20 年 4 月	
07046077	19	和誌	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 あきら, 山田 優子, 木ノ内 誠, 前野 聖, 金谷 重彦, 池村 淑道, 阿部 貴志	日本化学会情報化学部会誌	26	1	11-16	平成 20 年 4 月	
07046077	19	和誌	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	日本化学会情報化学部会誌	26	1	17-19	平成 20 年 4 月	
07046077	19	和誌	ライフサイエンスにおけるデータベース構築のための人材育成	瀬々潤, 池村淑道	蛋白質核酸酵素	53	1	87-93	平成 20 年 1 月	

講演

業務コード	実施年度	国内/国際	講演タイトル	発表者名	講演会名	発表年月日	メモ
07046015	19	国内	ゲノム研究のためのオントロジーとテキストマイニング	○高木利久	第 16 回セマンティックウエブとオントロジー研究会	平成 19 年 7 月 23 日	招待講演
07046015	19	国際	An approach to decipher gene regulatory networks from the federation of databases in life science	○ H. Bono, S. Kawano, S. Kawamoto and T. Takagi	FUNCTIONAL GENOMICS & SYSTEMS BIOLOGY	平成 19 年 10 月 10-13 日	Hinxton (UK)

07046015	19	国際	YUZ: an environment for integration AND ANALYSIS of gene regulatory networks	○ H. Bono, S. Kawano, S. Kawamoto, and T. Takagi	GENOME INFORMATICS	平成 19 年 11 月 1-5	New York (USA)
07046015	19	国内	統合 TV によるがんとハイボキシア研究の推進	○坊農秀雅	第 5 回 がんとハイボキシア研究会	平成 19 年 12 月 1 日	千葉
07046015	19	国内	MeSH terms を用いた生物学的機能付与による遺伝子群の解析手法の開発	○仲里猛留、滝中徹、坊農秀雅、麻生川稔、松田秀雄	第 30 回日本分子生物学会年会第 80 回日本生化学会大会合同大会 (BMB2007)	平成 19 年 12 月 11-15 日	横浜
07046015	19	国内	ライフサイエンス統合データベースプロジェクトの課題と成果	○川本祥子、坊農秀雅、池田佳代子、河野信、三橋信孝、峰崎雄一、永井啓一、西川哲夫、小野浩雅、山口敦子、山本泰智、大久保公策、高木利久	同上	同上	同上
07046015	19	国内	ライフサイエンス統合データベースセンターにおけるデータベースポータル構築	○河野信、小野浩雅、仲里猛留、坊農秀雅、岡本忍、中村保一、川本祥子、大久保公策、高木利久	同上	同上	同上
07046015	19	国内	3 次元人体モデルのセグメンテーションによる解剖学用語の形式表現	○三橋信孝、藤枝香、川本祥子、高木利久、大久保公策	同上	同上	同上
07046015	19	国内	遺伝子発現統合データベースの開発	○有川浩司、飯塚高康、渡邊康司、小笠原理、大久保公策	同上	同上	同上
07046015	19	国内	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	○小笠原理、飯塚高康、渡邊康司、有川浩司、大久保公策	同上	同上	同上
07046015	19	国内	ゲノムセントラルな遺伝子発現情報融通システム YUZ (柚子)	○坊農秀雅	同上	同上	同上
07046015	19	国内	TogoTV - a broadcast station of tutorial movies about bioinformatics resources	○S. Kawano, H. Ono, H. Bono, S. Kawamoto, T. Takagi	日本バイオインフォマティクス学会年会 (JSBi2007)	平成 19 年 12 月 17-19 日	東京
07046015	19	国内	ライフサイエンス統合データベースプロジェクト	○高木利久	「生命をはかる」研究会	平成 20 年 2 月 18 日	
07046015	19	国内	統合データベースセンターの紹介	○河野信	第 37 回人工知能学会分子生物情報研究会 (SIG-MBI)・第 9 回オープンバイオ研究会	平成 20 年 3 月 7-8 日	能美
07046015	19	国内	BioHackathon for Web Service の報告	○片山俊明	同上	同上	同上
07046015	19	国内	ウェブサービスの統合に向けた BioHackathon の成果	○片山俊明	国立遺伝学研究所研究会『生物情報資源の相互運用性』	平成 20 年 3 月 25 日	三島
07046015	19	国内	ライフサイエンス統合データベースプロジェクトの目標と成果	○川本祥子	同上	同上	同上
07046015	19	国際	OReFiL	○Y. Yamamoto	Pacific Symposium on Biocomputing	平成 20 年 1 月 4-8 日	Hawaii, U.S.A.

07046015	19	国際	Funcional analysis of groups of genes with MeSH hierarchy	○ T. Nakazato, H. Bono, H. Matsuda and T. Takagi	Systems Biology: Global Regulation of Gene Expression	平成 20 年 3 月 27-30 日	New York, USA
07046027	19	国内	ライフサイエンス統合データベース基盤整備のためのデータベースポータル構築:WIINGpro	○黒田 雅子, 小池 俊行, 安田 絵美, 佐藤 早苗, 藤田 晶子, 河村 明子, 川越 千晴, 坊農 秀雅, 川島 秀一, 谷口 丈晃, 川本 祥子, 高木利久	第 30 回日本分子生物学会年会・第 80 回日本生化学会大会 合同大会	平成 19 年 12 月 11 日	国内学会でのポスター発表, 横浜
07046041	19	国際	System update and literature curation at CyanoBase.	○岡本忍	International Biocuration meeting 2007	平成 19 年 10 月 25 日	San Jose, USA
07046041	19	国内	オーミクス研究情報の統合基盤としてのゲノムアノテーション	○中村保一	東京大学 21 世紀 COE 公開シンポジウム「言語から読み解くゲノムと生命システム」	平成 19 年 4 月 15 日	招待講演. 東京
07046041	19	国内	CyanoBase とソーシャルゲノムアノテーション	○岡本忍	ラン藻ゲノム研究会	平成 19 年 7 月 7 日	東京
07046041	19	国内	ソーシャル・ゲノム・アノテーション	○中尾光輝	CBRC セミナー	平成 19 年 8 月 31 日	東京
07046041	19	国内	KazusaAnnotation システムによるゲノムアノテーション高度化の試み	○岡本忍	ラン藻の分子生物学 2007	平成 19 年 12 月 3 日	木更津
07046041	19	国内	植物ゲノムのアノテーション	○中村保一	ライフサイエンス分野の統合データベース整備事業・遺伝学談話会 共同開催講習会	平成 20 年 1 月 25 日	長浜
07046041	19	国内	かずさにおける新しいアノテーションの試み	○岡本忍	統合データベース講習会	平成 20 年 3 月 5 日	東京
07046041	19	国内	ゲノム情報リテラシー入門	○中村保一	統合データベース講習会	同上	東京
07046041	19	国内	DAS workshop 2008 参加報告	○中尾光輝	37 回 人工知能学会 分子生物情報研究会 (SIG-MBI) (第 9 回オープンバイオ研究会と共催)	平成 20 年 3 月 8 日	能美
07046041	19	国内	KazusaAnnotation: ゲノム情報へのアノテーションを支援するシステム	○岡本忍	日本植物生理学会シンポジウム	平成 20 年 3 月 20 日	札幌
07046053	19	国際	Enhanced D-HaploDB: definitive haplotypes and extended haplotype information determined by genotyping complete hydatidiform mole samples	Higasa K, Kukita Y, Miyatake K, Tahira T, ○Hayashi K	57th Annual Meeting of the American Society of Human Genetics	平成 19 年 10 月 24 日	San Diego, USA
07046053	19	国際	Detection of human copy number variations using a collection of Japanese complete hydatidiform moles	Y. Kukita, K. Higasa, T. Tahira, ○ K. Hayashi	57th Annual Meeting of the American Society of Human Genetics	同上	San Diego, USA.
07046053	19	国際	Enhanced D-HaploDB: definitive haplotypes and extended haplotype information determined by genotyping complete hydatidiform mole samples.	Higasa K, Kukita Y, Miyatake K, Tahira T, ○Hayashi K	The 7th international workshop on advanced genomics	平成 19 年 11 月 27-28 日	Tokyo, Japan

07046053	19	国際	Detection of human copy number variations using a collection of Japanese complete hydatidiform moles	Kukita Y, Higasa K, Ishikawa S, Tahira T, ○Hayashi K	The 7th international workshop on advanced genomics,	同上	Tokyo, Japan
07046053	19	国際	D-HaploDB: A Database of Genome-Wide Definitive Haplotypes Determined using Complete Hydatidiform Moles	Yoji Kukita, ○Kenshi Hayashi	第5回国際バイオデータ相互運用性会議	平成19年9月26日	青海、東京
07046065	19	国際	Evaluation of kernel-based link analysis measures on research paper recommendation	○Masashi Shimbo, Takahiko Ito, Yuji Matsumoto	ACM/IEEE Joint Conference Digital Libraries	平成19年6月	
07046065	19	国際	A Discriminative Learning Model for Coordinate Conjunctions	○Masashi Shimbo, Kazuo Hara	ACL Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning	平成19年6月	
07046065	19	国際	Pivot learning for efficient similarity search	○Manabu Kimura, Kazumi Saito, Naonori Ueda	International Conference, Knowledge-based Intelligent Information and Engineering Systems	平成19年9月	
07046065	19	国内	文内共起パターンと格要素共有情報による事象間関係知識の獲得	○阿部修也, 乾健太郎, 松本裕治	言語処理学会第14回年次大会	平成20年3月	
07046091	19	国内	バイオデータマイニング入門	○瀬々 潤	ライフサイエンス分野の統合データベース整備事業・遺伝学談話会 共同開催講習会	平成20年1月25日	
07046077	19	国内	“学部教育としての持続可能型社会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, 中原拓, 大島一彦, 池村淑道	日本遺伝学会第79回大会	平成19年9月19日	
07046077	19	国内	“学部教育としての持続可能型社会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	第30回情報化学討論会	平成19年11月16日	
07046077	19	国内	“持続可能型社会への貢献遺伝子データベース～膨大な環境由来メタゲノム配列からの有用遺伝子探索～”	○上原啓史, 棚橋佳世, 阿部貴志, 池村 淑道	第2回日本ゲノム微生物学会年会	平成20年3月6日	

プレス発表

業務コード	実施年度	発表タイトル	掲載新聞名	掲載日
07046015	19	統合DBプロジェクトが本格始動 仮説の枯渇と蛸壺化を改善	BTJ ジャーナル (日経 BP)	平成19年8月

07046015	19	大学共同利用機関法人情報・システム研究機構、ライフサイエンス統合データベース・ポータルサイト開設	日経バイオテク	平成 19 年 10 月 5 日
07046015	19	生命科学の情報を網羅 今年度から統合データベース作成	読売新聞	平成 19 年 11 月 7 日
07046015	19	「BioHackathon 2008」が六本木で開幕、13 カ国から 80 人参加、12 日から 4 日間お台場で「Hack、hack、hack」	日経バイオテク	平成 20 年 2 月 12 日
07046015	19	生命科学 DB 統合へ	朝日新聞	平成 20 年 3 月 31 日

別表1 平成19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	◎○高木 利久
	情報・システム研究機構 国立遺伝学研究所	大久保 公策
	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	永井 啓一
		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	高木 利久
① 共通基盤技術開発		同上
	同上	山口 敦子
	同上	山本 泰智
	同上	全 弘宇
	東京大学医科学研究所ヒトゲノム解析センター	片山 俊明
	同上	川島 秀一
	かずさDNA 研究所植物ゲノム基盤研究部	中尾 光輝
② ワークフロー技術を用いた統合 DB 環境構築	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
	同上	野口 保
	同上	諏訪 牧子
	同上	秋山 泰
	同上	金 大真
	同上	Horton Poul
	同上	広川 貴次
	同上	福井 一彦
	同上	藤渕 航
	同上	堀本 勝久
	同上	原田 肇
	同上	廣瀬 修一
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	松本 裕治
	同上	新保 仁
	同上	浅原 正幸
	同上	木村 学
	同上	呂 嘉
	同上	小町 守
	同上	阿部 修也

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19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	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	○川本 祥子 同上
(1)ポータル整備・運用、広報、普及啓発 ①ポータル整備・運用、広報、普及啓発	同上 同上 同上	坊農 秀雅 河野 信 仲里 猛留
②意見集約システム運用、広報	科学技術振興機構研究基盤情報部 同上 同上	黒田 雅子 小池 俊行 安田 絵美

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19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) 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	西川 哲夫
① データベース受入れ	データベースセンター	同上
	同上	中村 春木
	同上	金城 玲
② データベース受入・運用	科学技術振興機構研究基盤情報部	黒田 雅子
	同上	河村 明子
	同上	佐藤 早苗
	同上	藤田 晶子
	同上	大倉 克美
(3) 人材育成		
① DB 構築者の養成	東京大学大学院 創成科学研究科	森下 真一
	同上	中谷 洋一郎
	同上	吉村 淳
② DB 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	瀬々 潤
	研究科	鷹野 景子
	同上	小口 正人
	同上	渡辺 知恵美
	同上	近藤 るみ
	同上	金子 聡子
③ アノテータ・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	池村 淑道
	同上	白井 剛
	同上	高橋 健一
	同上	大島 一彦
	同上	依田 隆夫
	同上	塩生 真史
	同上	中原 拓
	同上	阿部 貴志
	同上	上原 啓史
	同上	井口 八郎
	同上	米森 ゆかり
	同上	清水 信義
4. プロジェクトの総合的推進	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	◎高木 利久

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。