資料 2

「世界最高水準のライフサイエンス基盤整備」

ライフサイエンス分野の統合データベース整備事業

統合データベースプロジェクト

中間評価　報告書

平成２０年７月

「統合データベースプロジェクト」中間評価委員会

目　　次

はじめに

現在、我が国は、第３期「科学技術基本計画」（平成１８年３月２８日閣議決定）の下に、「科学技術創造立国」を目指して諸施策を実施している。

同基本計画においては、「抜本的な科学技術システム改革」が求められており、その中で２０１０年に世界最高水準を目指してデータベースを含む「知的基盤の戦略的な重点整備」を進めることとされている。同基本計画に基づき、総合科学技術会議が策定したライフサイエンス分野の推進戦略では、戦略重点科学技術の一つとして「世界最高水準のライフサイエンス基盤整備」が掲げられている。

生命情報の統合化データベースはライフサイエンス研究を支える基盤であり、その整備を進めるために必要な戦略の検討と技術開発を行なうため、文部科学省では平成１８年度より「ライフサイエンス分野の統合データベース整備事業」（以降、「統合データベースプロジェクト」）を推進している。

平成２０年度は、プロジェクト開始から３年目にあたることから、本プロジェクトに係る中間評価を実施し、本報告書をとりまとめた。

一方、総合科学技術会議「生命科学の基礎・基盤」連携施策群（平成１７年度及び１８年度は「ポストゲノム連携施策群」に相当）においては、同基本計画における分野別推進戦略（ライフサイエンス分野）の戦略重点科学技術の中で、関係府省間の連携を強化して研究体制の構築を行う課題のうち、統合的なデータベース整備に向けた研究開発は、他の研究領域の最も基盤となるものと位置づけて推進している。その一部は、統合的なデータベースの整備に向けた研究を補完的課題として公募・採択（課題名：「生命科学データベース統合に関する調査研究」、研究代表者：大久保公策教授（国立遺伝学研究所）、平成１７年度～平成１９年度）が行われた。本調査報告を受けて、今後具体的な制度設計やロードマップ等について平成２０年度末を目途に取りまとめ、データベース整備を推進して行く予定である。

本統合データベースプロジェクトは、同連携施策群の施策と密接に連携して進めてきたものであり、関係府省と共に引き続き連携をとって進めて行くところである。

Ⅰ　プロジェクトの概要

（目的）

現在、我が国のライフサイエンス分野のデータベースとしてDDBJ[†]（国立遺伝学研究所）、PDBj[‡]（大阪大学蛋白質研究所）、KEGG[§]（京都大学化学研究所）などが国際的に高い評価を受けている一方で、多くのデータベースについて、各機関や各プロジェクトで個々にデータベースが作られ、これらを関連付けて使おうとしたときに使い勝手が悪い、基本的サービスの多くが海外に依存していて、継続的に維持されない等の指摘も寄せられており、国内主要データベースの統合化と継続的な維持方策の必要性が指摘されている。

本プロジェクトは、我が国のライフサイエンス関係のデータベースの利便性の向上を図るため、我が国のライフサイエンス関係データベース整備戦略の立案・評価支援、データベース統合化及び利活用のための基盤技術開発、ポータルサイトの整備等を行い、統合化を推進することを目的としており、文部科学省が委託事業として実施しているものである。

（期待される効果）

本プロジェクトを通じて将来整備される「生命情報の統合化データベース」は、個々の分子生物学研究において蓄積されたデータが戦略的に統合され、付加価値の高いデータベースとして整備されるもので、幅広いライフサイエンス分野の研究者等がこれを活用し、今後の我が国におけるライフサイエンス分野の科学技術の進展に大きく貢献していくことが期待される。

これまでの研究成果の蓄積を網羅的・安定的に利用できるようになり、ライフサイエンス研究の発展に不可欠な基盤となる。

また用法や様式をまたいだ検索機能の開発等による既存データの新たな活用や、産業界・医学関係者等による応用利用を通して新たな知見が得られる。

（実施期間）

平成１８年度～平成２２年度（開始後３年度目に中間評価を実施）

（予算）

平成１８年度　　２．５億円

[†] DDBJ（DNA Data Bank of Japan。国際塩基配列データベースを構築している拠点の一つ。）

[‡] PDBj（Protein Data Bank Japan。生体高分子の立体構造データベースを国際的に統一化されたアーカイブとして運営すると共に、様々な解析ツールを提供。）

[§] KEGG（Kyoto Encyclopedia of Genes and Genomes。遺伝子、タンパク質、また代謝やシグナル伝達などの分子間ネットワークに関する情報を統合したデータベース。）

平成１９年度　　１６億円
平成２０年度　　１１億円

（実施体制）

背景となる国の考え方については、「我が国におけるライフサイエンス分野のデータベース整備戦略のあり方について」（平成１８年５月　科学技術・学術審議会　研究計画・評価分科会　ライフサイエンス委員会　データベース整備戦略作業部会報告書（参考資料（１））。以下「データベース整備戦略作業部会報告書」という。）を参考としている。

平成１８年度は、大学やさまざまな研究機関に蓄積されている生命科学関連の情報を横断的に利用可能とする統合データベースの構築が将来のライフサイエンス研究を支える基盤であるという考えに基づき、フィージビリティ・スタディとしてその整備を進めるために必要な戦略の検討と技術開発を行うため、「戦略立案支援・実行評価支援」、「統合データベース共通基盤技術開発」及び「ポータルサイト整備・広報、普及啓発」の３つの柱について、それら全ての統合化を推進する機関を公募・選定した。

採択された機関名、課題名『カギカッコ内』は以下のとおりである。

《代表機関》
大学共同利用機関法人情報・システム研究機構
『統合戦略立案評価および統合化基盤技術開発』
《参画機関》
独立行政法人科学技術振興機構『ポータルサイト構築』
国立大学法人九州大学『データベース統合化基盤技術開発』

平成１９年度からは、平成１８年度の取り組みの成果を基に、統合データベースの開発・整備に向けて本格的に推進すべく、「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」の３つの柱について実施することとした。

実施機関の選定にあたっては、事業の３つの柱の全てを担う中核機関、及び中核機関の下に「統合データベース開発」の一部を担う分担機関、さらに統合化を一層加速する観点から、中核機関の示す統合化方針に従い、自ら保有するデータ又はデータベースを中核機関に提供する補完課題実施機関を公募・選定した。

採択された機関名、課題名『カギカッコ内』は以下のとおりである。

【中核機関】
《代表機関》
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開発運用
（戦略立案・実行評価／統合ＤＢ開発／統合ＤＢ支援）』
《参画機関》
（１）独立行政法人科学技術振興機構『統合データベース支援：

意見集約システム運用／広報／データベース受入・運用』

（２）独立行政法人産業技術総合研究所『統合データベース開発：
ワークフロー技術を用いた統合ＤＢ環境構築』

（３）財団法人かずさディー・エヌ・エー研究所『統合データベース開発：
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』

（４）国立大学法人東京大学『統合データベース支援：ＤＢ構築者の養成』

（５）学校法人関西文理総合学園　長浜バイオ大学
『統合データベース支援：アノテータ・キュレータの教育』

（６）国立大学法人お茶の水女子大学
『統合データベース支援：ＤＢ高度利用者の養成』

（７）国立大学法人奈良先端科学技術大学院大学『統合データベース開発：
専門用語辞書管理システムと専門用語解析技術の開発』

（８）国立大学法人九州大学『統合データベース開発：多型知識表現技術開発』

《分担機関》

（１）国立大学法人京都大学『ライフサイエンス知識の階層化・統合化事業』

（２）（代表機関）国立大学法人東京医科歯科大学
『統合医科学データベース構築方式の開発』
（参画機関）国立大学法人大阪大学
『統合医科学データベース構築方式の開発』

（３）（代表機関）国立大学法人東京大学
『疾患解析から医療応用を実現するＤＢ開発
（ゲノムワイド関連解析のデータベース開発）』
（参画機関）国立大学法人東京大学医学部附属病院
『疾患解析から医療応用を実現するＤＢ開発
（リシークエンスＤＢの開発）』
学校法人東海大学『疾患解析から医療応用を実現するＤＢ開発
（ゲノムワイドSNPの統計遺伝学的解析手法の開発）』
株式会社日立製作所
『疾患解析から医療応用を実現するＤＢ開発
（ゲノムワイドSNPの疾患関連解析手法の開発）』

《補完課題実施機関》

（１）独立行政法人理化学研究所『植物オミックス情報および蛋白質構造情報』

（２）独立行政法人産業技術総合研究所
『糖鎖修飾情報とその構造解析データの統合
（糖鎖科学統合データベースの構築）』

（３）大学共同利用機関法人情報・システム研究機構　国立遺伝学研究所
『塩基配列アーカイブのデータベース構築と統合への貢献』

（４）国立大学法人九州工業大学『生体分子の熱力学データと構造データの統合』

（公募・選定の経緯）

本プロジェクトにおける実施機関の選定は、研究機関等を公募し、外部有識者により構成される選考委員会において書面および面接ヒアリングによる審査を実施し、選定している。但し、補完課題実施機関の選定においては、書面審査のみにて実施した。

これまでの公募・選定の経緯は以下のとおりである。

・平成１８年度公募・選定

平成１８年７月３日～同７月３１日
一般公募の実施

平成１８年８月２９日　受託実施機関選考委員会　開催
（ヒアリング審査、採択候補の選定）

・平成１９年度公募・選定

（中核機関、分担機関の選定）

平成１８年１２月２７日　　第１回受託実施機関選考委員会　開催
（公募課題、公募・審査方針の決定）

平成１９年１月１１日～同２月８日
一般公募の実施

平成１９年２月２０日　第２回受託実施機関選考委員会　開催
（ヒアリング対象課題の選考、ヒアリング審査方法の決定）

平成１９年３月５日　　第３回受託実施機関選考委員会　開催
（ヒアリング実施、採択課題の選考）

（補完課題実施機関の選定）

平成１９年７月２４日　　第１回補完課題選考委員会　開催
（公募課題、公募・審査方針の決定）

平成１９年８月１日～同８月３０日
一般公募の実施

平成１９年９月２６日　　第２回補完課題選考委員会　開催
（採択課題の選考）

Ⅱ　中間評価の概要

（１）目　的

統合データベースプロジェクトの実施に当たり、各受託機関の実施業務の平成１９年度末段階での中間成果を評価するとともに、プロジェクト全体の今後の方針等についての意見交換を行い、今後の事業展開に資することを目的とする。

（２）方　法

①書面評価および面接評価により行う（総合評価）。

但し、面接評価は中核機関および分担機関を対象に実施する。

②評価対象

・プロジェクト全体の評価（以下の各受託機関の状況を踏まえて総合的に評価）

・中核機関、分担機関、補完課題実施機関のそれぞれについての評価

（平成１９年度末時点の進捗状況）

③評価者　中間評価委員会（平成２０年３月３日設置）

※設置要綱（参考資料（９））、委員一覧（参考資料（１０））を参照。

④評価結果の反映

・今後の事業展開に向けた検討すべき課題の指摘および助言

・次年度の予算配分の重点化等の提案

（３）評価事項

１）プロジェクト全体の評価

・「データベース整備戦略作業部会報告書」（参考資料（１））を実現するために必要な目標設定になっており、ライフサイエンス分野のニーズに合致したものになっているか。

・プロジェクト全体の到達目標の下に、中核機関、分担機関および補完課題の各参画機関の役割に重複等の過不足がなく、全体が組織的に構成されているか。

・プロジェクトが中核機関を中心に円滑に推進され、効率的かつ効果的に進捗しており、プロジェクト実施期間内に期待する成果が達成可能であるか。

２）中核機関の評価

①進捗・達成度（プロジェクト計画の妥当性）

・目標達成のために中核機関として必要かつ十分な施策展開が図られ、適切な計画が立案されているか。

②事業の推進体制（プロジェクトマネジメントの妥当性）

(i) プロジェクトマネジメントに対する評価

・プロジェクトを円滑に遂行するためのマネジメント体制およびプロジェクト進捗管理の仕組みが整理され、日常的に適正に実施されているか。

(ii)連携体制に対する評価

・分担機関および補完課題の参画機関との連携を図り、事業を円滑に実施しているか。

③全体総括（プロジェクトの意義、波及効果）

・プロジェクト全体を推進する上で、中核機関として適切に機能しているか。

・プロジェクト全体の活動や成果が適時に公開されているか。

・ユーザニーズに合致しているかどうかを検証するための評価を適宜実施しているか。

・中長期的観点から、人材育成などのソフト面での整備も含め、ライフサイエンス分野のデータベース基盤整備の実現に向けて着実に進んでいるか。

・オールジャパン体制の意識を持って関係・関連分野のデータを有する機関や研究者等の協力を広く働きかけた提案がなされ、それを実施する体制が整備できているか。

３）分担機関、補完課題実施機関の評価

①進捗・達成度（課題計画の妥当性）

・目標達成のために分担機関又は補完課題実施機関として必要かつ十分な施策展開が図られ、適切な計画が立案されているか。

・課題実施期間内に、担当分野における量的並びに質的に実用に足り得るデータ又はデータベースの整備、統合化が実現できる計画となっているか。

②事業の推進体制（課題マネジメントの妥当性）

（ⅰ）課題マネジメントに対する評価

・プロジェクトを円滑に遂行するためのマネジメント体制および課題進捗管理の仕組みが整理され、日常的に適正に実施されているか。

（ⅱ）連携体制に対する評価

・中核機関と適宜連携をとりながら、業務計画に則って適切に且つ効率的に実施項目が進捗し、活動や成果が適時に公開されているか。

・機関グループ内部の連携を図り、事業を円滑に実施しているか。

③全体総括（課題の意義、波及効果）

・プロジェクト全体を推進する上で、分担機関又は補完課題実施機関として適切な役割を担っているか。

・オールジャパン体制の意識を持って関係・関連分野のデータを有する機関や研究者等の協力を広く働きかけた提案がなされ、それを実施する体制が整備できているか。

以上の観点で評価を以下のとおり中間評価を実施した。

（４）委員会開催実績

平成２０年３月３１日　第１回中間評価委員会　開催
（中間評価の視点、評価の進め方、評価方法の決定）

※中間成果報告（「調査票」）提出期限：平成２０年４月２１日）

※書面評価実施期間：平成２０年４月２３日～同年５月１２日

平成２０年５月２１日　第２回中間評価委員会　開催
（面接ヒアリング、面接評価、総合評価）

Ⅲ　中間評価結果

プロジェクト全体、及び各受託実施機関等について、書面評価及び面接評価による総合評価に基づく中間評価の結果をとりまとめると以下のとおりである。

１　全体評価

（１） プロジェクト全体について

〇進捗・達成度について

本プロジェクトは、平成１８年度から開始されたプロジェクトであり、初年度は中核機関である情報・システム研究機構を中心としてフィージビリティ・スタディを実施した後、平成１９年度より分担機関の３機関が、さらに補完課題実施機関の４機関が、平成１９年度途中より参画し、本格的な実施体制が構築された。

従って、補完課題実施機関に至っては約半年しか経っていないこともあり、全体的に見て目に見えて評価し得る成果が十分に出ているとは言い難いが、計画に対する進捗は順調以上に進んでいると評価される。

特に中核機関においては正味２年間の研究実施期間に満たないものの、国内の300 以上の極めて雑多なライフサイエンス分野のデータベースの横断検索、文献検索、統合ツールなど、見える成果として、短期間で公開できる段階まで到達した点は大いに評価できる。

中核機関としては、プロジェクトの重要性や、その目指すべき方向性、さらにはそのために解決すべき課題については十分把握しており、このままの方向で事業を実施すればライフサイエンス分野のデータベースが現在抱えている問題の多くが解消されるものと期待する。

〇事業推進体制について

体制面において、中核機関、分担機関、補完課題実施機関の相互連携の姿や協力体制が、必ずしも円滑に図られておらず、一部で類似した開発項目や、中核機関からの期待に必ずしも符合しない開発項目が、分担機関等で実施されているといった機関間の意思疎通に係る問題が見受けられた。理想的な役割分担や業務計画の実現に向けて、本評価結果を踏まえたより具体的な議論が必要と思われる。

〇総論

平成１９年度が実質的な立ち上げ段階ということもあり、今後、本格的な連携効果が出てくるものと思われる。しかし、これを加速させるためにも、より中核機関が主導的にプロジェクト全体を管理・運営できるような体制となるよう、後半期は見直すべきである。また、参画機関を含めた中核機関の研究体制を柔軟に見直す裁量を、中核機関に対して与えることが必要である。

さらに研究推進に当たっては、研究運営委員会をより効果的に活用することで、中核機関、分担機関、補完課題実施機関といった現行の組織構造のあり方についても十分に議論し、抜本的な体制の見直しに向けた検討を早急に実施するのが望ましい。

すなわち、中核機関の代表機関である「ライフサイエンス統合データベースセンター」の強力なイニシアチブの下、データベース戦略や方針に沿ってプロジェクト全体が統括・管理され、明確な参画機関各々の役割・機能により事業推進体制を整理することが、世界最高水準の研究基盤整備である統合データベースの実現に向けては不可欠であろう。

（２）全般的に見た今後の課題、助言等

〇本プロジェクトの意義

・本プロジェクトは５年間であるが、単なる「プロジェクト」と見るのではなく、「日本国に必要なライフサイエンスデータベース確立の試行」として事業を進めることが必要である。

〇プロジェクト終了後の今後あり方

・本プロジェクトは、日本のライフサイエンス研究の成果を集約させ、そこから新たなライフサイエンスの知識を生み出す場所の構築のため集中的に整備された。データベースは継続的に整備され、利用されることにより大きな価値が認められる。そのため、プロジェクトが終了した後、整備された「統合データベース」を維持し、発展させることが肝要であることは言うまでもない。
・そのためには、昨年度の総合科学技術会議の優先付けにおける指摘事項（「（統合データベースプロジェクトとバイオインフォマティクス推進センターについては、）一本化を含めた検討を行うことが必要」）**を十分踏まえ、我が国のライフサイエンス基盤データベース（DDBJ、KEGG、PDBj等）を支え、推進してきた科学技術振興機構(JST)のバイオインフォマティクス推進センターに主な経費を一本化し、本プロジェクトの中核機関である情報・システム研究機構による戦略立案機能と、研究開発独立行政法人の事業としての「統合データベース」の維持・運用・高度化等を、有機的に連携させるべきである。
・また、研究室単位のデータベースを研究費交付と連動させて統合化させる等、将来的な統合化に向けた工夫や、完成後の維持を考慮し、少ない経費で維持できるような仕組みの構築に向けた積極的な議論も望まれる。

〇プロジェクト内の事業の整理等

・「統合データベース」として何をどこまでするのかという全体構想を参画者で共有した上で、現在の事業推進体制の中での分担機関のミッションや、個別のデータベース開発の必要性について再度精査し、中止することも必要である。
・特に、汎用的、包括的なデータベースと、特定研究に特化した専門性の高いデータベースとのバランスについては、予め検討を行うことが望ましい。さらに、汎用的

** 総合科学技術会議（第７０回、平成１９年１０月２９日）資料１－２「平成２０年度概算要求における科学技術関係施策の優先度判定等について」(http://www8.cao.go.jp/cstp/siryo/haihu70/siryo1-2-5.pdf、p.11 特記事項：
〇継続性をいかに担保するかが重点課題である。
〇JST-BIRD との連携について、将来的な一本化を含めた検討を行うことが必要である。
〇データベースを作るのみにとどまらず、常に改訂していくことが必要である。）

なものや、利用者が多く出ると予想されるものについては、商業的なベースを活用する等も検討すべきである。
・研究費の投入は新規データベース構築や新しい概念で既存データベースを組み替えるような作業を伴うものに選択的に投入されるべきである。つまり、既存データベースの維持管理については、事業化して民間資金の導入といった点も含めて検討してみてはどうか。
・統合データベースを支えて行くために必要とする人材育成は重要であるものの、本プロジェクトで実施している人材育成施策が「データベース整備戦略作業部会報告書」（参考資料（１））において求められている期待やアウトプットに沿っているのかの観点で再度見直すことが必要であり、限られた事業期間、予算内で対応すべき業務の優先順位を付けた上で、抜本的な整理を敢行すべきではないか。

〇その他の課題、助言等
・中核機関へのバックアップにより、省庁連携統合データベースを促進できる体制を担保すべきである。
・他省庁の主導するライフサイエンスのデータベース、特に医療や健康維持に関する情報とこれからどのように連結していくかと言う課題を念頭に置きつつ、検討を進めて行くべきである。
・メンバーに医学、医療の事情を熟知した研究者を加えると臨床情報との連結がより有効に行えるであろう。分子の統合データベースからライフサイエンス全体の統合データベースへと発展することを期待する。
・市民向けのコンテンツが少し不足しているように感じられる。昨今の市民意識の高まりを考えると、先端研究の内容を噛み砕いて市民向けに開放する作業が必要な時代になっていると強く感じるので、将来的な構想としてこうした観点からのアプローチも期待する。
・本データベースの利用者を増やし、利便性を高めるために、広くモニターを募って定期的に報告を求める制度を始めてはどうか。
・中核機関や分担機関の東京大学グループでは、計算機資源の不足問題が挙がっている。計算機リソースについて、公共の既施設を無償で（廉価に）使えるような仕組みや、支援する仕掛け、あるいは共用できる方策等について検討すべきである。

２　個別評価

（１） 中核機関（代表機関名：情報・システム研究機構、代表者：高木　利久）

〇進捗・達成度について

「辞書の整備」、「知識の整理棚」、文献のオープンアクセス化をにらんだ技術開発、日本語雑誌や学会要旨の検索、教育用教材、ポータルサイト作成、データベース受け入れなど様々な試みがなされ、利用者の利便性向上が緒につき、データベースの統合化が具体的に動き出したことは評価される。

またデータベース構築者のために公開されるデータ全てをダウンロード可能とするといった統合に向けた方針は、より多くの利用者と波及効果を期待できることから望ましい方向である。今後、質の高いデータをより多く蓄積するための取り組みが期待される。

我が国における中核機関として短期間にこれだけの進捗を見せていることを十分に評価するべきと考える。こうした取組を確実に進めていくためにも、中核機関に相応しい予算配分の再構成が必要である。

〇事業推進体制について

一方、中核機関の役割は、自らが成果を挙げると同時に、分担機関あるいは補完課題実施機関と上手く連携して相乗的に目標を達成することにある。しかし、一部の参画機関は必ずしも中核機関の認識どおりの意識にはなっていないようにも見受けられ、連携が不十分である。これはプロジェクト発足当初において、中核機関に指導、連携に関する権限が明確に与えられていないことによるものと思われる。

〇総論

非常に大きなミッションであり、きわめて重要な課題を扱っている。多様な参画機関、分担機関、補完課題実施機関のコントロールが必要であり、現在の中核機関のパワーだけに任せるのは不十分であり、より強いイニシアチブを持たせるべきである。問題と改善策に対しては、研究運営委員会における早急かつ十分な議論を望む。また中核に相応しい予算配分の再構成が必要である。

今後（プロジェクト終了後）統合データベースをどのような体制で進めるかの検討（コンセンサス）が必要である。この予算でできること、できないことを整理しておいた方が良い。

（２）分担機関

①京都大学（代表者：金久　實）

〇進捗・達成度について

化合物に関するソフトウェア整備、医療医薬品、一般用医薬品に関するデータベースの公開、LinkDBの対象データベースの拡張、などについて妥当な進捗が見られる。また、本統合データベースはゲノムネット[††]とKEGGを分離し、前者のサポートということで、明確な切り分けをしており、KEGGとのリンクは有機的に進められている。さらに国内（JAPIC[‡‡]、LipidBank）のみならず、国際データベースやIUPAC[§§]との連携も視野に入っており、継続支援は重要である。

[††] ゲノムネット（ゲノム情報を基盤とした新しい生命科学研究と創薬・医療・環境保全への応用を推進するために、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が提供するインターネットサービス。http://www.genome.jp/）

[‡‡] JAPIC（財団法人日本医薬情報センター。国内外の医薬品に関する臨床的に有用な情報を収集・処理・提供することによって、薬剤の臨床使用の適正化を通じて製薬と医療の間のかけ橋の役目を果たすことを目的に設立された公益法人。）

[§§] IUPAC（International Union of Pure and Applied Chemistry。国際純正・応用化学連合。元素名や化合物名についての国際基準（IUPAC命名法）を制定している国際学術機関。）

〇事業推進体制について

中核機関との連携について、どのようになされているのか、姿が良く見えない。連携は発足当初に考えられていなかったので、そのための話し合いで取り入れた医薬品、化合物データベース構築については、配分比率が高過ぎると思われる。

〇総論

ゲノムネットの医薬品データベースは大学等の研究者または学生にとって、または薬に関心のある一般利用者には有用なデータベースであるかも知れない。しかし、創薬に係る研究者など医薬品企業等の産業界による利用まで考えるならば、JAPIC など廉価で手に入る情報とのリンクではなく、別のあるいは独自の情報とのリンクが必要ではないか。また、当該分野については、便利な有料のデータベースが揃っている点を踏まえつつ、魅力ある検索法を提案するなどの別の方策が必要である。

またプロジェクト全体予算（１１億円）の中で、ゲノムネットの DBGET/LinkDB 解析ツールの開発費用が整合性を持って配分されているのか、また補完課題実施機関の産業技術総合研究所が担っている糖鎖関係のデータベースや中核機関が開発している検索エンジンの開発等において、各々の役割分担が曖昧になって整理されていない開発等が見受けられる点を考慮して、予算配分については適正に査定する必要がある。

さらに、当該機関が持つポテンシャルである KEGG の強固な基盤をどのように役立たせるかについてより深い議論が中核機関との間で必要である。

こういった観点も踏まえ、限られた全体予算の中で効果的な成果を上げるため、利用者のニーズに真に応えるためにどうすべきかを研究運営委員会等で議論すべきである。

②東京医科歯科大学グループ（代表者：田中　博）

〇進捗・達成度について

ライフサイエンスのデータを分子から臨床にいたる多階層の視点からデータベースを構築するという姿勢は評価できる。未だ例数が少ないので評価が難しい。

２つのデータベースをモデル的に構築しようとしている方針は評価できる。しかる後に、そのモデルをどのようにして国内に広めて行くのかについて確たる見通しがないのが不安である。

特定の医師（場合によっては comedical）が限定しているデータベースとしては適当だが、中核機関の姿勢とそぐわないのではないか。

〇事業推進体制について

本データベースの構築に当たっては、疾患記述の標準化など日本全国にインパクトを与える可能性のある重要な事項が散在している他、中核機関との連携の姿が不明確である。

〇総論

限られた予算規模及び時間で一定の成果を出すためには、取組内容について、優先順

位を考えた上で、フォーカスを絞る必要があり、インフォームドコンセントが必要となるなど、その統合に際して大きな困難が予想される医学情報のデータベース作りに向けて「自力で出来ること」を中心に進めるべきである。つまり、データベース開発を目指すのではなく、中核機関が目指すデータベースの統合化に向けて、臨床関係のデータベースはどう統合していくべきか、何をすべきか、何が重要かといった統合に当たっての課題に対する提言や、フィージビリティ・スタディ的なプロトタイプを小規模に作る等のロールモデルの提供に徹するべきである。

③東京大学グループ（代表者：徳永　勝士）

〇進捗・達成度について

GWAS (Genome-Wide Association Study)は世界の潮流であり、理化学研究所以外の GWAS 中核として、国内のデータ保有者と連携の下、多数の施設に亘る共同研究をベースとしたオールジャパン体制で進めており、ナショナルプロジェクトに相応しい。到達目標（アジアのハブ）も明快である。このまま継続されるべきである。

データの品質管理、及び標準化を行うという研究内容は評価できるものであり、限られた予算内で期待される達成度を満足している。

〇事業推進体制について

また、中核機関の方針に馴染んだマネジメント体制が取られていることも評価される。さらに、標準 SNP データベース、GWAS データベースとも平成１９年度計画を十分達成しており、成果は着実に上がっている。

〇総論

今後は、一層のデータを集めるとともにデータを解析する努力が必要であり、特にこのデータベースにどの位のデポジットがなされ、どう発展させるかが重要である。

疾患解析のデータベース構築の際に発生しうる問題点を整理しつつ、小さくても「モデル的データベースの構築」を示して欲しい。

また、計算機資源、データ解析能力の充実に向けては、中核機関との密接な連携が必要であり、両者の一層の協調と協力体制の確立が必要であるが、計算機資源については、公共的なスーパーコンピュータの利用等のコスト便宜にも配慮すべきである。

（３）補完課題実施機関

①理化学研究所（代表者：豊田　哲郎）

理化学研究所の多数あるデータベースを利用しやすいものとするとともに、付加価値を付け統合データベースに組み入れていくテストケースとして、当該取組みの意義は大きい。

主として理化学研究所で行われたシロイヌナズナのオミックス、タンパク 3000 プロジ

ェクトからの高等動植物由来タンパク質の構造データ、微生物由来蛋白質構造データなどを用いて、そのアノテーションなどの構築、およびアノテーションシステムの開発運用の研究について、研究体制は適格に構築できており、時間的な問題も考えると、順調な準備段階であると言える。

今後は、分野毎の国内他機関との連携の構築と、理化学研究所内のあらゆるデータを積極的に公開する方向に進んでいただき、中核機関の期待に沿うよう盛り立てていただきたい。

②産業技術総合研究所（代表者：成松　久）

中核機関との役割分担が明確になって、最もすっきりした関係になっており、糖鎖関連という特定の領域で情報基盤の整備・有効利用が促進されることは関連分野の活性化にとどまらず、データベース統合化のモデルケースとしても評価できる。

利用者コミュニティが比較的まとまっているので、コミュニティとの意見交換を通して、利便性の高いデータベースを作成することが容易な状況にあると言える。一層の努力を期待する。

この種のデータベースにおいては、素材自体の重要性も考えられ、特に利用者を幅狭く想定しないのであれば、「自前」のデータのダウンロードなど様々な利用者のニーズに備えていただきたい。

今後は、プロジェクト終了までになるべく多くの関連データベースの統合を進めていただくとともに、中核機関と連携して、他の分野とも連携したより上位の統合の具体策を検討していただき、中核機関の期待に沿うよう盛り立てていただきたい。

③国立遺伝学研究所（代表者：五條堀　孝）

公開 ftp サイトと WWW サイトに関する開発については、国内主要機関との連携により統合システムの準備段階である。登録処理及び波形表示システムに関する開発についてはシステムの設計、プロトタイプシステムの完成、手法の開発のための情報の調査中の段階である。本プロジェクトが始まって時間が経っていないので、達成度は十分ではないものの概ね妥当である。

ただし、中核機関との連携においては、その目的から鑑みるに、例えば「新しい種類の、あるいは新しい発想に基づくデータベースの開発支援」のような新しい連携のあり方が構築できれば、一層統合データベースへの貢献がなされると考えられる。

今後は、新型 DNA シーケンサの出現により、今後塩基配列データの生産量あるいはその研究対象領域の急速な拡大が予想されるため、単に従来型のトレースアーカイブデータベースを構築するだけでなく、将来を見据えたシステムのあり方を検討していただき、中核機関の期待に沿うよう盛り立てていただきたい。

④九州工業大学（代表者：皿井　明倫）

データベース自体の重要性というよりは、小規模データベースと統合データベースの関係のモデルケースとして重要と思われる。

熱力学データと構造データの統合データベースを構築・提供することによって中核機関に対する保管機能は十分果たされている。XML 化、オントロジー調査はサービスの質の向上に相当するが、データベースそのものの今後の改良とともに、中核機関と連携して中・小規模のデータベース構築者に適用しやすいシステムあるいはツールの開発も期待したい。

理化学研究所、大阪大学（PDBj）と打合わせを持ちながら課題を遂行しており、データベース構築も順調であることから課題マネジメントも問題はないと思われる。

今後は、巨大機関ではなく研究室単位のデータベース構築者としての統合データベース構築参加のモデルケースとしての役割を続けていただき、中核機関の期待に沿うよう盛り立てていただきたい。

おわりに

本プロジェクトは、我が国のライフサイエンス関係のデータベースの利便性の向上を図るため、データベース整備戦略の立案・評価支援、統合化及び利活用のための基盤技術開発、人材育成等を行い、ライフサイエンス関係データベースの統合的活用システムを構築・運用するという、これまでにない取り組みを行うものである。

平成１８年度からの約半年間のフィージビリティ・スタディを経て、平成１９年度より本格的な実施体制が敷かれて間もないため、全体的に見て中間評価として目に見えて評価し得る成果を要求することが厳しい中、中核機関である情報・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）を中心に精力的な活動がなされており、計画に対する進捗は順調以上に進んでいると評価したい。引き続き中核機関を核として分担機関ならびに補完課題実施機関との連携、協力をより密接に図ることにより、データベースの整備が順調に進み、いよいよ本格的に事業が加速、推進していく状況である。参画している各機関、研究者の今後の活動に大いに期待するところである。

報告書のまとめとして、今後、本プロジェクトをより強力に推進していくにあたって、留意すべき点を指摘しておきたい。

○中核機関の代表機関である「ライフサイエンス統合データベースセンター」がプロジェクト全体を統括、目標を管理し、参画機関各々の役割・機能の分担責任を明確にして、データベース戦略や方針に沿って、センター長が組織全体を引っ張るリーダシップ、イニシアチブが思う存分発揮でき、プロジェクト全体をコントロールできるような体制となるよう、後半期は見直すべきである。
また分担機関、補完課題実施機関および中核機関の各参画機関は、「ライフサイエンス統合データベースセンター」との連携をより積極的に働きかけ、本統合データベースプロジェクトを一丸となって盛り立てていただきたい。

○本プロジェクトで築き上げられた統合データベースは、我が国のライフサイエンス上の財産の集積場所として、そこから新たな知識を生み出す場所として、長期的に維持する必要がある。プロジェクトが終了した後の管理運営体制の継続性の担保が必要である。また統合化に向けた仕組みをどのように維持するかという観点から、所在も見えてこない可能性のある研究室単位のデータベースをどのように統合するか等の将来像も必要である。これらの諸課題について研究運営委員会等の場を活用して十分に議論し、検討して解決策を見出していただきたい。

（参考資料）

（１）データベース整備戦略作業部会報告書
（２）プロジェクト実施体制
（３）平成１８年度公募・選定の状況
（４）平成１９年度公募・選定の状況
（５）平成１９年度補完課題の公募・選定の状況
（６）平成１９年度研究運営委員会／戦略作業部会　委員一覧
（７）平成１９年度研究運営委員会／作業部会　委員一覧
（８）平成１８年度研究成果報告書
（９）中間評価委員会設置要綱
（１０）中間評価委員会　委員一覧
（１１）中間成果実績一覧

参考資料（１）

報　告　書

我が国におけるライフサイエンス分野のデータベース整備戦略のあり方について

平成１８年５月１７日

科学技術・学術審議会
研究計画・評価分科会
ライフサイエンス委員会
データベース整備戦略作業部会

目次

（注）フッターにある（ ）付き番号は、
参考資料内のページ番号です。

1. はじめに

（データベース開発の歴史）

ライフサイエンス分野のデータベースは、古くは米国国立医学図書館 NLM (National Library of Medicine) による MEDLARS (Medical Literature Analysis and Retrieval System、1964 年)にさかのぼる。これは現在ライフサイエンス分野で最もよく利用されている文献データベース MEDLINE (MEDLARS On-Line、1971 年)の前身にあたるものである。もっとも紙ベースのものまでデータベースと呼ぶことにすれば、さらに、米国 American Medical Association による CIM (Cumulated Index Medicus、1879 年)や米国 Army Medical Library による CLML (Current List of Medical Literature、1937 年)までさかのぼることができる。これらはすべて医学関連の文献集であるが、1970 年前後からそれ以外の種類のデータベース、すなわち、今日我々がデータベースという言葉から想起する種類のデータベースが次々と作られるようになる。ジョンズホプキンス大のマキューージック博士による遺伝子変異疾患データベース MIM(Mendelian Inheritance in Man、1969 年)、米国エネルギー省のブルックヘブン国立研究所によるタンパク質立体構造データベース PDB (Protein Data Bank、1971 年)などがそれである。

ライフサイエンス分野の最も基盤的なデータベースである核酸配列のデータベースに関しては 1980 年に欧州分子生物学研究所 EMBL(European Molecular Biology Laboratory)で現在の EMBL-Bank の前身が作られ、米国では 1982 年にロスアラモス国立研究所で GenBank が産声をあげた。日本では 1986 年に国立遺伝学研究所で DDBJ (DNA Data Bank of Japan)が始動し、これら三者による現在の日米欧三極体制が出来上がった。

このようにライフサイエンス分野において古くからいろいろなデータベースが作られてきたが、データベースの開発と普及の観点で最も重要な契機は何といっても 1990 年前後に世界的に開始されたヒトゲノム計画であろう。これによりライフサイエンスの大量情報化時代の幕が開いた。冒頭に述べたように、これ以前にも多くのデータベースが作られてはいたが、利用者は一部の研究者に限られていたし、その利用法も限定的なものであった。それが、ヒトをはじめとする種々の生物のゲノム計画が相次いで始まったことにより、また、インターネットの普及やデータベース技術の進歩などによりデータベースの利便性が高まり、利用が一気に拡大した。その後、配列データに加え、遺伝子の発現データ、タンパク質の立体構造や相互作用などのデータが続々とデータベース化され、さらに利用が拡大しつつあることは周知の通りである。データベースは質的にも量的にも種類的にも急成長を遂げ、ライフサイエンス研究に欠かせないものになったのである。

さて、このようにデータベースがライフサイエンスの発展の重要な鍵になるという認識は、昨日今日なされたわけではない。このような時代が到来することはゲノム計画開始当初にすでに一部の研究者には予想されていた。その当時の文献を紐解くとデータの産出よりもその管理や解析のほうが大変で重要だということがいまから 15 年以上も前から指摘されていたことが分かる。この認識は、我が国においても同様で、例えば、1991 年に東京大学医科学研究所にヒトゲノム解析センターが作られたとき、最初にゲノムデータベース分野が設置されたことからもそのことが伺える。このようなデータベースの重要性への理解を背景に、我が国においては、ゲノム関連の科学研究費補助金などにおいて 1991 年頃よりデータベース開発への積極的な取組みが行われてきた。また、これと並行して、国立遺伝学研究所生命情報研究センターや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の設置などデータベース構築を目的の一つに掲げた組織の整備も行われた。これにより
我が国においても多くのデータベースが開発され、一般の利用に供されてきた。そうした中から、関係者の献身的な努力の甲斐もあり、世界的に誇れるデータベースがいくつか生まれてきた。

しかしながら、その一方で、せっかく苦労して作られたにもかかわらず、あるいは、当初は世界でも最先端を行くものであったにもかかわらず、維持更新されずに朽ちていったデータベースも少なくない。この原因の多くは、データベースでは継続的な維持更新が最も重要にもかかわらず、多くは競争的資金で、しかも個々の研究者の個人的な努力により開発が行われていたことに求めることができる。また、データベース構築が高度に知的な活動であることに対する理解や評価の低さもその背景にあったであろう。データベース作りは個々の開発者の創意工夫が必要であるという点に着目すれば研究としてとらえることができるが、一方では組織をあげて長い時間をかけて地道に作って行くものであるという点からは事業としての性格をもつものである。データベース作りはこのような二面性をもつ特殊なものであるにもかかわらず、いわゆる研究と同じ枠組みの中で扱われ評価されてきたことが我が国においてデータベースの健全な発展を阻害してきたと言えよう。

さて、いまから 6、7 年前の 2000 年前後にデータベースの構築や維持更新に関するこのような問題が顕在化するのとほぼ時を同じくして、それを担うバイオインフォマティクス分野の人材不足も大きな社会問題として浮上してきた。データベースの問題も含め、バイオインフォマティクスのこのような状況を改善するためには、どういう取り組みをすればよいのであろうか？その当時、この件に関して、さまざまなところでさまざまな議論や取り組みが行われたが、その中でも最も具体的かつ実効性があったものの一つが、その当時の科学技術会議ライフサイエンス部会ゲノム科学委員会において行われた議論とそれに基づく提言である。その内容は 2000 年の 11 月に出された報告書「ゲノム情報科学におけるわが国の戦略について」にまとめられている。この中では、バイオインフォマティクスの人材養成、研究開発の振興、データベース整備戦略、の 3 つの課題に関して推進方策が提言されている。そして、この提言をもとに、2001 年度より科学技術振興機構(JST) にバイオインフォマティクス推進センター (BIRD)が設立され、これら３つの課題について精力的な取組みが行われてきた。

この報告書で提言された３つの課題（人材養成、研究開発振興、データベース整備戦略）の推進方策は 5 年経ったいまでも決して古くはなっていないし、BIRD での取組みも大きな成功を収めてきた。この５年間にゲノム、トランスクリプトーム、プロテオーム、メタボロームなどで大規模なデータが次々現れてくるなど、表面的には状況が大きく変化しつつある面もあるが、ここで示された提案内容の多くは基本的にいまでも十分通用するものである。しかしながら、データベース整備戦略に関して言えば、提言されたことがすべて実現されたわけではないこと（例えば、国家レベルの整備戦略を立案する機能や、それを実施する中核的な機能は実現されなかったこと、など）、複数の省庁におけるライフサイエンスに関連するデータベース開発の活動はあったものの文部科学省だけの動きに留まった面があること、BIRD はその守備範囲からして国家的な戦略を考える組織としては不十分であったこと、その後ライフサイエンス分野で続々と行われることになる大型プロジェクトのデータベースの受け皿作りが 2000 年の報告ではあまり考慮されていなかったこと、データベースが予想をはるかに超え多く作られるようになったこと（そのためのポータルサイトや統合化などの必要性が増したこと）、オントロジーや文献の知識などの重要性が

増したこと、など、更なる体制の強化を図ったり、戦略を再度見直したりすることが必要な面が出てきた。また、期せずして産業界からもデータベース整備の強化や見直しを求める声もあがってきた。

（データベース整備戦略作業部会における検討）

そこで、ライフサイエンス委員会の下に、データベース整備戦略に関する作業部会を設けて　2005年年8月12日、11月10日、2006年1月16日、2月28日、3月24日、5月11日の計6回にわたって、ライフサイエンス分野における今後のデータベース戦略をどうすべきか議論を重ねてきた。この作業部会では、ライフサイエンス分野のデータベースとしては、ゲノム、トランスクリプトーム、プロテオーム、メタボローム、などの網羅的かつ基盤的なデータ（いわゆるオームデータと呼ばれるもの）とそれを解釈するためのパウスェイやオントロジーなどの知識のデータベース化、すなわち、ライフサイエンスの基盤となるデータベースに重点をおいて議論を展開してきた。また、利用者もこれらのデータベースを使って研究開発を行う研究者、技術者をおもな対象として議論を進めてきた。当然のことながら、ライフサイエンス分野のデータベースあるいはバイオ分野のデータベースという範疇には、これら以外にもさまざまなデータベースが存在する。これには、生物資源等の研究用材料に関するもの、医療現場で用いられる臨床情報や医薬品情報、化合物の構造や毒性情報、食品の成分や安全性に関するもの、作物や家畜の育種に関するもの、産業上有用な微生物の情報、など多岐にわたるものが該当する。これらのデータの重要性は、オームデータに比べて決して劣るものではないし、本作業部会でおもな議論の対象とした基盤的データとこれらのデータとの連携・統合を図ることが重要であるが、これらの多くは、文部科学省以外の省庁で精力的に取り組みがなされており、また、現在科学技術振興調整費「科学技術連携施策群の効果的・効率的な推進」の一テーマとして調査研究が進められていることもあり、十分な配慮はしたが、検討のおもな対象とはしなかった。

このように本作業部会では、ライフサイエンス全般のデータベースを視野に入れながらも、その基盤となるようなデータベースに重点をおき、ライフサイエンス研究におけるデータベースの意義や重要性、世界的なデータベース開発の動向、我が国のデータベースの現状と問題点などについて綿密な調査を行うとともに、今後のデータベース整備戦略の基本的な考え方や推進方策あるいはその実現に向けた体制作りについて活発な議論を戦わせてきた。本報告書はそれらの結論をとりまとめたものである。ライフサイエンス研究においてデータベースの重要性が今後ますます高まることは万人の認めるところである。本報告書での提言が速やかに実施されることが、我が国におけるデータベース開発の発展、ひいては、それを基盤としたライフサイエンス研究、医療、バイオ産業の発展につながるものと確信している。なお、本報告書では、主たる検討の対象をライフサイエンスの基盤的なデータベースに絞ったが、上に述べたように、ライフサイエンス分野のデータベースは多岐にわたる。また、現在、関係各省を横断的に俯瞰したライフサイエンス分野のデータベース統合に関して、内閣府を中心に活発に議論が進められていることから、本報告書で取り上げた提言の具体的実施に際しては、これらの議論や提言と十分な整合性をもって進める必要がある。

2.ライフサイエンス研究におけるデータベースの意義および整備の必要性

1977 年のマクサムとギルバートおよびサンガーによる DNA 塩基配列決定法の開発を端緒とし、その後 1990 年頃より開始されたヒトゲノム計画の進展により、ライフサイエンスは莫大な量の DNA 塩基配列データ産生の時代を迎えた。現在はこれに加えて、いわゆるポストゲノム時代を迎えて、タンパク質の立体構造データや遺伝子の発現データも爆発的に増えている。このようなデータの洪水の中で、データベースの活用なしにライフサイエンス研究を行うことは事実上困難になってきている。また、実験データだけでなく、文献数も着実に増加しており、米国の文献データベース MEDLINE に登録されている文献の総数は 1500 万件を超え、一人の研究者が関連する分野の論文すべてに目を通すのは不可能な状況になっている。こうした情報の洪水という状況下では、一方で、データベースを活用することにより、従来不可能であったような質と量のデータを個々の研究者が利用できるようになっており、これが現在のライフサイエンス分野の研究開発効率の飛躍的向上を可能としている。今後も情報の増加が加速することが予想されるライフサイエンス研究をさらに進展させていくためには、より網羅的かつ正確で付加価値のついたデータベースを整備し、活用していくことが不可欠である。

データベースの整備は、研究開発の効率化のためばかりではない。ライフサイエンスという多岐にわたる学問体系を生命のシステムとして再統合し、俯瞰することにより、生命の理解がより深まり、ライフサイエンス研究の一層の進歩が見込まれる。データベース整備はその礎となるものであり、国内における整備が不可欠である。ゲノムネットワーク、タンパク 3000、国際Ｈａｐ Ｍａｐプロジェクトなどの近年のライフサイエンスの大規模プロジェクトに期待される成果の一つはデータベース整備にあるといってもあながち間違いではない。すなわち、データベースはライフサイエンス研究に不可欠の基盤であると同時に、次世代の研究への糸口を提供する。あわせて、医学の発展を通しての国民の健康への貢献、食糧問題、環境問題、資源問題への貢献、さらには産業利用など、ライフサイエンス研究の成果の適切な活用という観点からも、データベースの公開・整備は重要である。

一方で、データベースの構築、提供は、単なる既存の情報の提供サービスに止まるものではなく、その知識の蓄積が新たな研究分野を作っていくものであり、その構築時には予想もしていなかった研究の展開が期待できる。例えば、ヒトやマウスの完全長 cDNA 配列データ（巻末の用語解説参照）がデータベースに多数蓄積された結果、タンパク質に翻訳されないが機能をもつ RNA 分子が多数存在すること、多くの遺伝子が読み方を変えることにより複数種類のタンパク質を作りうること、などの発見により多様なＲＮＡに着目した新たな研究分野が拓けたことなどがそれである。また、累積されたデータを整理・統合するということ、すなわち、データベース開発を推進することは、とりもなおさず、生命を担う構造の空間軸、時間軸を俯瞰する能力のある人材を育成することにもつながる。これにより将来新たな視点でのライフサイエンスを拓くことが可能になる。また、データベースは、多額の税金を使って行われるライフサイエンス研究の成果を医療、民間企業、さらには一般社会に還元するための手段としても重要である。データベース化により、その成果が誰の目にも明らかにできるからである。

後述するように、このような意義や重要性をもつデータベースへの理解や整備が我が国では遅れており、学界のみならず産業界からもデータベース整備に対する強い要望が寄せられている。例えば、ゲノム分野の国家プロジェクト等の成果を広い分野で迅速に実用化研究に活用できるよ

うに、一元的に集約・統合され様々な角度からデータを参照できる、無償公開を原則としたデータベースを国が積極的に整備してほしいなどの要望がある。

以上、ライフサイエンス研究の観点からも、産業の観点からも、更なるデータベース整備の強化充実がいままさに求められている。そのためには、今後、公開性・透明性・客観性・科学性を担保しつつ国内でデータベースを戦略的に整備していくことが重要である。その際、データベースを構築する側の立場に立った整備ではなく、それを利用する側の立場に立った整備に努めることが必要である。このためには、情報系、実験系との緊密な共同作業によるデータベース整備がなされなければならない。また、データベースの構築と普及には長い年月を要するため、また、いわゆる研究とは異なる側面をもつため、上記の整備は定常的な経費をもって永続的に続ける必要があることは言うまでもない。

3. 国内外のデータベース開発の現状と動向

3-1 データベース開発の世界的動向

DNA 塩基配列の登録データ量が指数関数的に増加してことはよく言われていることであるが、英国の科学雑誌である NAR (Nucleic Acids Research)が毎年 1 月に発行しているデータベース特集号に登録されているデータベース数からみると、ライフサイエンス分野のデータベースそのものの数も指数関数的に増大していることが伺える。また、タイトルにデータベースという記載のあるライフサイエンス分野の論文数から判断すると、データベースの累計は一万にも及ぶと推定される。また、その種類も多岐にわたってきており、文献から抽出したデータやオントロジー（巻末の用語解説参照）といった知識に関するデータも急激に増加してきている。また、DNA やタンパク質の配列や立体構造といった生体関連物質の構造に関わるデータばかりでなく、遺伝子やタンパク質の発現や相互作用といった生体関連物質間の関係に関わるデータ、およびパスウェイ（用語解説参照）、疾患、表現型という機能に関わるデータも数多くデータベース化されるようになってきた。すなわち、生体関連物質の個々の部品のデータから、それらが構成するシステム全体に関わる情報のデータベース化へと、開発の重点が移りつつある。また、DNA 配列データバンクのようなデータ生産者からの一次データが登録されるデータ登録型のデータベースだけでなく、すでに登録された様々な分野のデータを加工した、あるいは、文献から抜き出したデータを収録した、知識集約型の二次的なデータベースも増加してきている。

以上紹介したように、ライフサイエンス分野のデータベースは量的にも、質的にも増加、拡大してきており、研究面でも、産業応用面でもその重要性はますます高まってきている。すなわち、データベースはライフサイエンス研究や医療、バイオ産業の国家戦略を考える上で欠かせない大きな柱の一つであるとの認識が世界的に広がっている。

3-2 欧米におけるデータベース整備の現状

本報告書の冒頭「1. はじめに」に記載したとおり、米国では、1964 年には、国立衛生研究所 NIH(National Institutes of Health)の下部組織である NLM で現在の文献データベースサービス PubMed につながる活動が始まっている。その後、1969 年に遺伝子変異疾患データベース MIM が、1971 年にタンパク質立体構造データベース PDB が、1982 年に核酸配列データベース GenBank が、というように、現在のライフサイエンス研究になくてはならないデータベースが続々と誕生した。欧州では、1980 年に欧州分子生物学研究所 EMBL において核酸配列データベース EMBL-Bank の前身が形作られた。その後、米国では 1988 年に NLM の配下に国立バイオテクノロジー情報センター NCBI(National Center for Biotechnology Information)が設立され、欧州でも、1992 年に EMBL の下部組織として欧州バイオインフォマティクス研究所 EBI(European Bioinformatics Institute)が設立された。これらの組織は、それぞれ GenBank、EMBL-Bank をはじめとするライフサイエンス分野のデータベースの開発および維持更新を専門に担うとともに、バイオインフォマティクスの研究とデータベースや情報解析のサービスの中核拠点として機能するように設けられたものである。

表 1 に NCBI と EBI の概要を示した。NCBI は、NLM の一部門として設立されたものであり、予算規模は 85 億円、人員規模は約 400 名である。サービスは GenBank を中心とする核酸配列データを中核に、標準配列である RefSeq データベースの提供、Entrez システムによる統合データベース環境の構築、および世界標準的な相同（ホモロジー）解析ソフトウェアである BLAST を中心とする各種解析ソフトウェアにその特徴がある。さらに、国立医学図書館 NLM の下部機関としての特徴を活かした文献データベースサービス(PubMed)の提供は、他にはない大きな特徴である。一方、EMBL の一部門である EBI の予算規模は 32 億円、人員規模は 300 名弱である。サービスの特徴には、UniProt や InterPro といったデータベースに代表されるタンパク質配列を対象とした機能情報の提供と、Ensembl と呼ばれるデータベースにおける真核生物のゲノムを対象とした詳細なアノテーション（データに生物学的医学的な解釈を加えること）情報の提供などがある。資金的には、英国のウエルカム財団や米国 NIH のからの資金も得て活動している。そのため、他機関との共同開発も多い。NCBI、EBI とも人員の二割から三割程度の研究部隊を抱えており、単にデータベースの整備に限定された組織ではなく、データベースの整備やサービスの提供とそれに関連する研究開発とが対になった組織が形成されている。また、これらの組織ではデータベースの開発が明確な目標をもったプロジェクト制で実施されている。

3-3 我が国におけるデータベース整備の現状

日本においても欧米同様にライフサイエンス分野のデータベースは、医学関連の文献情報にその起源を求めることができる。1903 年の医学中央雑誌がそれである。その後だいぶ時は下るが、1958 年に JST において科学技術文献速報が発行され、1976 年にはオンラインデータベースサービスが開始されている。ただし、これらはライフサイエンスだけでなく科学技術分野全般を対象としたものである。

核酸配列データに関しては、前述したように、1986 年に DDBJ が国立遺伝学研究所で産声をあげ、米国の GenBank、欧州の EMBL-Bank と合わせた日米欧の三極体制がこのときに形作られた。その後、NCBI、EBI の設立と同じような時期に、ゲノムネットと呼ぶ国際的なバイオ情報サービスが京都大学化学研究所と東京大学医科学研究との連携にもとに立ち上がった。また、東京大学医科学研究所にヒトゲノム解析センターが、国立遺伝学研究所に生命情報研究センターが、京都大学化学研究所にバイオインフォマティクスセンターが次々と設置され、データベース構築やバイオインフォマティクス研究の下地が整えられた。しかしながら、欧米の NCBI や EBI に匹敵するような中核機関の設置までには至らなかった。ちなみに、我が国におけるセンターとしては最大規模を誇り、また、日米欧との三極体制を担う DDBJ は年間予算 12 億円で、人員は事務員も含めて約 60 名で活動している。実態的には、国立遺伝学研究所の生命情報・DDBJ 研究センターの 5 つの研究室が基盤となっており、タンパク質の構造、機能に関するデータベースである GTOP や遺伝子発現に関するデータベース CIBEX など、独自のデータベースの開発も行っている。

上述のようなデータベース構築やバイオインフォマティクス研究の振興を目的としたセンターの設置の動きに加え、前述の「ゲノム情報科学におけるわが国の戦略について（平成 12 年 11 月科学技術会議ライフサイエンス部会ゲノム科学委員会）」を受けて科学技術振興機構 JST が 2001 年にバイオインフォマティクス推進センターBIRD を設立し、データベース整備やバイオインフォマティクス人材養成に関する競争的資金を拡充した。この枠組みによって、京都大学の KEGG(Kyoto

Encyclopedia of Genes and Genomes)や大阪大学のPDBj(Protein Data Bank Japan)といった世界的に定評のあるデータベース構築や国際協調によるデータベース整備への支援などが進められてきた。KEGGは、細胞レベルでの生命システムの機能に関する知識を分子間相互作用ネットワークの情報としてデータベース化したパスウェイデータベースを中心に、遺伝子カタログ情報(GENES)、生体関連化学物質情報(LIGAND)、機能情報(BRITE)などから構成される一種の統合データベースであり、論文からの引用が多いことでも知られている。また、PDBjは、米国のRCSB(The Research Collaboratory for Structural Bioinformatics)およびEBIのMSD(Molecular Structure Database)との連携のもとに運営されているタンパク質の立体構造データベースである。PDB自体は、前述したようにもともとは米国のブルックヘブン国立研究所により運営、維持されていたものであるが、現在は上記の体制で、国際的な連携のもとに運営されている。PDBjでは、XML(Extensible Markup Language)などの最新情報技術を利用した新しいデータ記述と解析ソフトの開発、およびタンパク質表面形状と物性に関するデータベースeF-siteなどの二次データベースの開発を行っている。

また、文部科学省科学研究費補助金の特定領域研究の中でも、研究成果の一環としてデータベースの構築が行われている。「ゲノム4領域」では、研究成果としてのデータベースリストを公開しており、現在68件のデータベースが登録されている。また、「発生システム」でも、ユウレイボヤなど6種類の個別生物に関するデータベースが開発され、公開されている。その他、一般の科学研究費補助金などでも多くのデータベースが開発されているものの、全体にどの程度のデータベースが構築されているかの把握はなかなか難しい。英国の科学雑誌であるNAR の2006年1月のデータベース特集号、JSTバイオインフォマティクス推進センターBIRDのデータベースディレクトリサイトWING、上記の文部科学省科研費特定領域のホームページ、知的基盤整備委員会による平成16年11月のデータベース見直しリスト、および平成16年度学術情報データベース実態調査などから総合的に判断すると、公開予定のものも含めて文部科学省関連では190件程度のデータベースがあるものと推定される。ただし、これらのうちの約三分の二は一種類の情報にしか登録されていないため、さらに多くのデータベースが水面下にある可能性は否定できない。

さらに、文部科学省以外の他省に目を転じてみると、まず、経済産業省関係では1998年以来進められてきたヒト完全長cDNAの構造解析プロジェクトで得られた成果を基盤に、タンパク質機能解析・活用プロジェクトが実施され、遺伝子の発現頻度解析により得られたデータについてデータベース化され公開されている。また、ヒト完全長cDNA構造解析プロジェクトで得られた完全長cDNAクローンの配列情報は、DDBJに登録されると同時に、世界各機関の全長cDNA配列情報とともにアノテーションがつけられ、H-Invitationalデータベースとして一般に公開されている。その他、産業技術総合研究所あるいは製品評価技術基盤機構からも20件程度のデータベースが公開されている。また、厚生労働省関係では、ミレニアムプロジェクトの一環として進められてきた「疾患データベース」プロジェクトの成果が、GeMDBJ データベースという形で公開されている。解析データは、アルツハイマー病、がん、糖尿病、高血圧、喘息からなる5疾患に関わるゲノムワイドなSNP解析（巻末の用語解説参照のこと）と、ファーマコジェネティックス（用語解説参照)に関するSNP解析、また一部の疾患等に関するチップによる遺伝子発現データを含んでいる。農林水産省関係では、1991年に開始されたイネゲノム解析研究の成果として、イネゲノム配列情報を中心に、遺伝地図情報、物理地図情報、EST 情報も含め幅広い、質の高い情報が蓄積され、それらはINEをはじめとするデータベースで公開されている。また、ブタのcDNA情報や蚕のゲノ

ム情報も解析され、それぞれ公開されている。データベースの件数としては、主要なもので 20 件程度になる。

3-4 日米欧の競争力比較

欧米では 3-2 に記載の通り、それぞれ国家戦略に基づき、データベースに関わる中核機関を設置し、この機関を中心にデータベースの開発、維持を一元管理している。さらに、NCBI、EBI とも独自の研究部隊をもっており、単にデータベースの整備だけでなく、将来のデータベースのサービスを見据えた研究開発も同時に行える体制になっている。また、ゲノム解析の時代からデータ産出（実験系）と強い連携をもってデータベース開発を進めてきた。一方、日本は表 1 に示すように、JST バイオインフォマティクス推進センターBIRD の人員は 61 名（JST 雇用者）、年間予算は 19 億円(平成 17 年度)であり、欧米の中核機関と比較して優位な状況にはない。予算の仕組みや事業内容が異なるため、厳密な比較は困難であるが、予算的には、BIRD に大学共同利用機関（例えば DDBJ）の予算を加えてはじめて欧州と対等の存在になるが、人員的には、その他の中核機関（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や東京大学医科学研究所ヒトゲノム解析センター）を含めても欧州と対等の存在になるかならないかである。米国には、予算的にも、人員的にも水をあけられたままである。ただし、上にも書いたように、何をもってデータベースに関する予算や人員であると定義するかはいろいろ難しい面があり、また、上記の見積もりには、我が国におけるデータベース構築・維持活動が網羅されているわけでもないため（欧米においてもしかり）、予算でも人員でも大幅に足りないと安易に結論付けるわけには行かない。これに関しては、現在内閣府を中心に進められている調査の結果を待つ必要があろう。ここでは、欧米と比較して予算面人員面で決して優位な状況にはないこと、むしろ遅れをとっている可能性が高いこと、米国との比較に関してはそれがより顕著であることを述べるにとどめておく。

さて、欧米との違いは予算面人員面だけではない。データベースの構築支援のあり方、戦略の立て方、統合化のレベル、データ産出側との連携、などにも大きな違いを見出すことができる。我が国では欧米と異なりプロジェクト制ではなく、研究者の創意に基づくデータベースを支援しているケースが多い。また、繰り返し述べているように日本にはデータベース整備に関わる中核機関がないなど、これまで国家戦略がなかったため、JST が資金提供してきた一部のデータベースを除いて、統合化、標準化が遅れている。また、データ産出側との連携の弱さやデータベース利用者のニーズの把握不足も大きな問題であろう。

以上、まとめると、予算面人員面の不足、国家戦略の欠如、その推進を担う体制や省庁を超えた連携の不備により、欧米と比較して、データベースの整備やその標準化・統合化が総じて遅れていると言えよう。そしてこのことが、バイオインフォマティクスの人材不足とあわせて、我が国におけるライフサイエンス研究やバイオ産業の競争力の低下の一因になっていると言えよう。

表１　日米欧の主要中核機関の概要

	日本 （中核的な機能を果たしている機関の例）		米国	欧州
	科学技術振興機構 バイオインフォマティクス推進センター	国立遺伝学研究所 生命情報・ＤＤＢＪ研究センター	国立バイオテクノロジー情報センター （NCBI）	欧州バイオインフォマティクス研究所 （EBI）
組織形態	独立行政法人科学技術振興機構（JST）の組織新しい生物情報の研究開発によるデータベースの整備等の推進と普及のための拠点　統括、副統括、事務局で構成	大学共同利用機関国立遺伝学研究所の付属施設。「生命情報学」の我が国における研究拠点。我が国を代表するDNAデータベースのDDBJを運営	米国ＮＩＨ傘下のＮＬＭの付属機関 分子生物学分野を支援する公共データや解析ソフトの提供と計算機を利用した基礎研究機関	EMBL の傘下の非営利学術機関　バイオインフォマティクスの研究とサービスの中心機関
組織の永続性	JSTの運営費交付金（バイオインフォマティクス推進事業）により運営	国立遺伝学研究所の運営費交付金により運営	根拠法：Public Law 100-607	上部機関EMBLは18カ国からの公的研究資金で運営されている。EBIの資金の半分を負担。やや不安定
予算	19億円	12億円	85億円	32億円
人員	61名（JST雇用者）	62名(事務員含)	約400名	283名
サービスの概要	生物情報データベースの高度化・標準化、バイオインフォマティクスの創造的研究開発、新しい情報生物学の創造のための起業支援センター	国際塩基配列データベースの共同構築と運営	配列情報データの標準配列(RefSeq)の提供やEntrezによる統合データベース、各種解析ソフト提供の世界的な中心 アクセス数：4000万/日	タンパク質配列を基礎とした機能情報(UniProtやInterPro)や真核生物のゲノム情報の統合サービス(Ensembl)
特色	KEGG、PDBjなど国際的データベースの開発支援	タンパク構造(GTOP)や遺伝子発現(CIBEX)など独自データベース開発	文献情報（PubMed）と配列情報、各種解析ソフトの充実、データベース間の連携	タンパク情報の充実
国内プロジェクトとの連携	国内の代表的データベースの構築･高度化を支援	ゲノムネットワークなど他プロジェクトに参画	スタッフは、塩基配列解析、遺伝子同定、遺伝子発現に関する実験的解析において、他のNIHの機関と協力	EU、NIHなどの研究資金を得てプロジェクトに参加している。英国のウエルカム財団、サンガー研究所と連携
自前の研究機能	なし	生命情報･DDBJ研究センターの5研究室で実施	基礎研究グループは、Computational Biology Branchの中にあり、70名のsenior scientist、staff scientist、research fellow、postdoctoral fellowからなる	バイオインフォマティクスの研究グループ１７（新規３）からなる。EBI独自のデータベースや機能サービスを担うグループを含む
教育機関人材養成との連携	バイオインフォマティクス分野の明日を担う人材の育成を目指した講座の開催	総合研究大学院大学傘下の研究機関として博士課程教育を実施。実習付の講習会の実施	"A Field Guide and NCBI Resources"がアメリカ全土で開催。コースは3時間の講義と2時間の実習	学位（PhD）取得を目指す学生から独立した研究者に対するコースを提供
その他特筆すべき事項			NLMには外部への研究資金配布機能がある。NIHの研究資金により実施された研究に由来する論文やデータについて、受け皿をNCBIやNLMが用意する	企業へ最先端の技術を普及することや企業からの寄付を得るための仕組みを整備

4. 我が国におけるデータベースの問題点と今後取り組むべき課題

4-1 データベースの問題点

3-3 で記載したように、我が国のライフサイエンス分野においても、これまで非常に数多くのデータベースが開発されてきている。その中には、国立遺伝学研究所の核酸配列データベース DDBJ、京都大学の統合パスウェイデータベース KEGG、大阪大学のタンパク質立体構造データベース PDBj、東京大学/科学技術振興機構の日本人一塩基多型データベース JSNP など世界に誇れるデータベースも一部にはあるが、多くのデータベースに関しては、各機関や各プロジェクトでバラバラにデータベースが作られ、所在情報が誰にでも分かるようにはなっていない、似たようなものがいくつもありどれを使ってよいか分からないなど、それらを関連付けて使おうとしたときに大変使い勝手が悪いという状況である。また、多くは十分な解析や解釈がなされず生データをただ格納したものになっており、また、臨床情報などの表現型データとの統合が十分でなく、医療や創薬その他の産業への応用が困難になっている。多くのデータベースが日本語化されていないことも広く応用展開していくことを困難にしている要因であろう。さらには、それらを使いこなして有用な生物学的あるいは医学的な知識を発見するための利用技術の開発も十分ではない。また、それ以前に、各機関や各プロジェクトで構築されたデータベースがなかなか公開されない、仮に公開されても永続的に維持更新されないという問題がある。プロジェクト期間が終了すればそのまま放置される場合が多く、せっかくの成果が広く活かされないうちに消えてしまう。また、別の問題として、各機関や各プロジェクトで産出された実験データと文献に書かれた知識とを対応づけることが近年重要になってきているが、それらの連携が我が国では弱いという問題もある。

このような状況の背景には、制度上・予算上の問題、データベース構築への理解不足やそれを担う人材不足などがある。

まず、データベース構築、維持に関わる制度上の問題点としては以下の点が指摘できる。第一に、データベース整備は国家をあげて取り組むべき重要な課題であり、我が国のライフサイエンスの未来を決するものであるにもかかわらず、ライフサイエンス全般を見渡して、我が国のデータベース整備はどうあるべきか、目的や対象をどう設定すべきかを常に考え、戦略を立案する体制が十分には整っていないことである。JST バイオインフォマティクス推進センターBIRD や国立遺伝学研究所 DDBJ の委員会など一部にはそのような機能を期待されるものもなくはないが、それらの守備範囲は限定的であり、また、他に本業を抱える非常勤の委員からなる委員会がその役割を担っており、それらに我が国全般の戦略立案を期待することはできない。データベースの整備戦略の立案は片手間では行えない、また、高度な専門性を有する仕事である。

第二に、仮にデータベースの整備戦略がうまく立案できたとしても、それを速やかに効率よく実行に移す体制が整っていないことである。これに関しても前述の JST BIRD や国立遺伝学研究所 DDBJ その他の組織はあるが、予算の制約や制度上の問題から、多くは期待できない。これは、複数の制度、機関、研究分野を有機的に組み合わせる実施体制が整っていない、すなわち制度的、分野的縦割りを打破できる体制が整っていないことによるものである。

第三に、第二の問題とも関連するが、実験系研究者と情報系研究者間の共同研究を実施する土壌や体制が整っておらず、両分野の研究者による相乗効果が十分発揮できていないという問題である。そのため、データの解析や解釈が十分に施されず、いろいろなデータがバラバラなまま、データベース化されてしまうことにつながっている。

第四に、これが最も大きな問題かもしれないが、予算上の問題である。データベースを継続的に維持していくための予算的枠組みが十分でないこと、および期限付きの予算で作成されたデータベースの予算終了後の受け皿、すなわち、そのための予算確保の制度（例えば、プロジェクト経費の一部を必ずデータベース整備と維持管理に当てるなど）が整っていないことである。データベースは一般に長い間丹念に維持更新してはじめて大きな価値を生むという性格を有している。我が国のデータベースのほとんどが、多くの人が使う非常に基盤的なものでも、個々の研究者が行う、いわゆる研究と同じ扱いをされ、競争的資金によって維持管理されている。これが大きな問題を生み出している。

さて、このような制度上、予算上の問題以外にも、我が国において、データベースを開発することの意義、重要性が少なくとも研究の観点では必ずしも重要視されていないことが指摘されよう。新しいデータベースの開発については論文を出せるようにはなったが、データベースの維持・管理については論文化が難しく、論文出版が研究者に対する主な評価対象となっている我が国では、データベースの開発・維持に対する意欲刺激が高くならない。そのため、これらの開発や作業に従事する研究者およびデータに生物学的医学的な解釈を加える専門職員（アノテータ）やデータベースの編集作業に従事する専門職員（キュレータ）の人材不足を招いている。また、ライフサイエンス分野のことを十分に理解して、データベースシステムの開発や運用にあたるシステムエンジニアやオペレータなども不足している。その結果として、データベース自体、研究の片手間に作ればよいという風潮を招いている。これらのことが我が国で世界的に競争力のある、また、使い勝手のよいデータベースの開発を遅らせている要因になっており、また、構築されたデータベースの価値を正しく評価する目利きや仕組みが我が国に育っていないことにもつながっている。そのため、存在する個々のデータベースの重要度を評価し、支援すべきデータベースを選別することが難しくなっており、整備戦略の立案を困難にしている遠因になっている。

4-2 取り組むべき課題

4-1 で述べた問題を解決し、ライフサイエンス研究や産業応用に十分に役立ち、また、世界的な競争力を備えたデータベースを整備するには、これまでの整備方針や既存の組織やプロジェクトの見直しを図るとともに、それを踏まえて、上で指摘した問題点を解消するための制度作り、予算措置、体制作り、および、人材養成に取り組む必要がある。具体的には、以下の取組みが必要であると考える。

(1)ライフサイエンス、バイオ産業全般を見渡して、また、国家的視点に立って、データベースの整備戦略を立てること。また、そのための体制を整備すること。ライフサイエンス分野は急激に進展しているため、また、データベースは欧米との競争・協調の中で整備を進める必要があるため、戦略そのものよりも、常に我が国のデータベースの現状を質的・量的の両面から調査・評価し、それに基づいて戦略の見直しが柔軟に図れる体制作りが重要である。
(2)データベース整備は研究とは異なる事業的な側面をもつことを十分に認識し、DNA 配列やタンパク質立体構造のような、ライフサイエンスの基盤として不可欠なデータベース、あるいは、有

用性が広く認識され、かつ、世界的に競争力のあるデータベースを安定的に支援するような体制を整備すること。
(3)我が国で開発された種々のデータベースの所在情報や利用法などを漏れなく掲載したポータルサイトを構築し運用すること。
(4)高度に統合化されたデータベースを開発し、大学・研究機関等に加えて産業界における利用者の利便性を一層向上させること。また、そのための技術を研究開発すること。また、これらの推進に必要な体制を整備すること。その際、データベースを構築する側の立場だけでなく、利用する側の立場に十分に配慮すること。また、そのために、実験系の研究者や技術者がデータベース作りに参画する仕組みを確立すること。
(5)プロジェクト終了後にそこで開発されたデータベースを受け入れ、維持管理するための体制を整備すること。そのために必要となる予算をプロジェクト設置と連動させて、また、ライフサイエンスの進展状況に応じて機動的に確保する仕組みを制度化すること。
(6)種々の実験データと文献に書かれた知識とを対応させる仕組みを強化すること。
(7)未解釈、未解析のままのデータをアノテーションして付加価値を高めること。これを継続的に行う仕組みを確立すること。また、そのための、実験系、情報系の連携を図る仕組みを設けること。
(8)ライフサイエンスの展開に対応して、新しい種類の、あるいは、新しい発想に基づくデータベースの開発を支援する体制を整備すること。
(9)データベースの利用技術（バイオインフォマティクス）の研究開発を促進する方策を講じること。
(10)データベースを構築したり、そこから有用な知識を引き出したりできる人材（研究者、キュレータ、アノテータ、システムエンジニア、オペレータ、など）を養成すること。

これらの取組みに際しては、一部繰り返しになるが、データベースに対するニーズを十分に把握すること、すなわち、データベースを利用する側の意見を十分に取り入れる必要がある。そのために、データベースのおもな利用者である実験系の研究者や技術者がデータベース構築に深く関与する仕組みを作ること、また、そのために情報系と実験系が共同でデータベースを構築したり、そのための共同研究を行ったりすることを積極的に支援するための環境整備にも十分な配慮が必要である。また、新たな制度や体制の構築に際しては、既存の体制の限界を認識するとともに、ライフサイエンスの進展に応じて出てくる新しい要望に対して国をあげて柔軟に対応できる体制や制度を構築することが必要である。さらには、データベースはライフサイエンス分野の知的基盤であり、安定的に活用するという観点からの事業の継続性が最も重要である。

5. データベース整備戦略の基本的考え方

前節の後半 4-2 で述べた「取り組むべき課題」は、どれもこれも重要なものであり、一つとしてゆるがせにすることは許されない。また、我が国で作られている数多くのデータベース（3-3 参照のこと）にはそれぞれに存在意義があり、十分な評価を行った上でできるだけ支援することが望ましい。しかしながら、予算の制約もあり、また、データベースの問題は 2 節でも述べたように、我が国のライフサイエンスのあり方にも大きく関わる面があり、具体的な方策を講じるには十分な検討を加える必要がある。また、一方で、日本語化の問題を除けば、データベース開発は欧米との競争にさらされざるを得ず、そのような視点からも整備戦略を練る必要がある。3-4 で述べたように、一部のデータベースを除いて我が国のデータベースの現状は決して優位にあるとは言い難い。

このような状況を踏まえ、今後、日本がとるべきデータベース整備戦略の策定にあたっては、以下の方針で臨むことが大切であると考える。

第一に、データベースはこれまでのライフサイエンス研究の叡智をまとめた宝であると同時に、ライフサイエンスやバイオ産業にとって不可欠の基盤であるとの認識に立ち、国が構築を支援したデータベースは原則的に公開すべきである。

第二に、我が国におけるデータベース整備の見直しにあたっては、日本としての特徴（強み）を出せるように進めることが必要不可欠で、そのための新たな機能や目的の付加が必要である。単に重要なデータベースだから、あるいは、欧米でも作られているから、というだけでは支援するわけには行かない。

第三に、過去の研究資金の投資状況、国内の研究者の裾野の広がり、当該研究分野の国際的な位置づけ、民間資金における代替可能性などを総合的に分析し、データベースとして備えるべき機能や取り込むべき研究分野についてメリハリを付けること、優先順位付けが重要である。そのためには、整備戦略立案は質的・量的両面にわたる調査・評価のしっかりとした裏づけに基づくものでなければならない。

第四に、安全保障、波及効果等の競争以外の観点にも配慮が必要である。上に書いたことと一見矛盾するようだが、仮に欧米ですでに作られていても我が国独自に構築すべきと判断すれば重複や後追いを恐れず構築すべきである。

第五に、知的基盤として長期的な取組みが可能となるような枠組みの構築が必要ということである。これには人材養成、責任体制の明確化も含む。データベース構築およびそのための体制整備は一朝一夕にはできない。5 年後、10 年後を見据えた年次計画を立案すべきである。

第六に、人材養成においては人材の備えるべき能力に対する社会の要望のみならず、養成された人材の将来の処遇（キャリアパス）についても十分に配慮する必要がある。

第七に、国際的データベースを分担開発するような場合でも、単に分担するだけにとどまらず、国内のデータベースとの連携や統合に十分な配慮を払うべきである。

最後に、生物資源等の研究用材料（バイオリソース）や最先端・高性能汎用スーパーコンピュータ等、データベース以外の研究基盤の整備計画と十分な連携をとり、互いにその効果を高め合うような配慮が必要である。

なお、上記の日本としての特徴を出せる例としては、cDNA、イネゲノム、微生物ゲノムなど強みのある分野の機能を付加したデータベース、SNP や生物多様性情報等の地域性のあるデータベ

ース、パスウェイ等国際的に優位にあるデータベース、日本で開発・収集されたバイオリソースとの連携に関わるデータベース、最先端・高性能汎用スーパーコンピュータ等の高速計算機と情報解析技術の活用が期待できるデータベース、我が国の国家プロジェクトの成果を戦略的に活用する方向性のあるデータベースなどが挙げられる。

なお、狭義のデータベースではないが、我が国では文献からの知識抽出技術の開発や遺伝子辞書の構築に関して強みがある。これらを活かした整備戦略も検討に値する。もちろん、日本語での利用環境の提供は日本独自のものであり、この点も十分検討すべき項目である。

最後に、4-2 で述べたことの一部繰り返しになるが、ライフサイエンス研究の進展、欧米のデータベース整備状況の変化等により、日本としての特徴も時々刻々と変わりうるため、状況の変化に国をあげて柔軟に対応できる、すなわち時代に応じて進化可能な体制の整備やそのための制度設計が最も重要な課題の一つである。

6. 推進方策とそれを実現するための体制

6-1 推進方策

前節「5. データベース整備戦略の基本的考え方」を踏まえて、「4-2 取り組むべき課題」を具体的に実現するためには以下の推進方策を遂行する必要がある。推進すべき方策とその留意点は下記の通りである。

(1)データベースの現状調査、評価、戦略立案機能の充実

現在、データベース整備の戦略立案機能はJSTバイオインフォマティクス推進センターBIRDや国立遺伝学研究所DDBJに設けられている委員会あるいは文部科学省のライフサイエンス委員会などによって一部担われているが、それらは非常勤の委員からなる委員会活動であり十分ではない。また各機関の委員会では、その守備範囲もその組織の活動に関するものに限られ、限定的なものとなっている。そこで、専門家による日常的活動（研究者の常勤）を基盤とし、データベースの現状や動向の定常的な調査および既存の戦略や活動の弛まぬ評価に立脚して、省庁の枠を超えて国家的視野に立って、ライフサイエンス研究全般やバイオ産業全般を見渡した戦略立案する機能が是非とも必要である。

なお、これらの調査、評価、立案に際しては、以下の点に十分な配慮・検討が必要である。

- データベースだけの問題と捉えるのではなく、ライフサイエンス研究の方向性も十分に踏まえた戦略を立案すること。
- データベース構築は、個々の研究者の創意工夫による研究とは異なる事業的な側面をもつことを十分に認識し、その推進および体制の整備に努めること。
- データベースは、ライフサイエンス研究全般、医療、バイオ産業全般の知的基盤、後方支援との明確な位置づけを行い、ニーズを的確かつ継続的に把握すること。
- データベースを構築する側の立場だけでなく、利用する側（例えば、医療や産業界）の意見が十分に取り入れられるように配慮すること。また、そのための仕組みを確立すること。
- 現在、ともすれば別々の戦略をもって収集・管理が行われている医学情報や薬学情報との連携にも十分配慮すること。
- データベース間の連携強化のためのデータベースの形式や構造の標準化や知識の体系化に向けた用語の統一化（辞書作成・標準化）のための戦略もあわせて立案すること。
- また、用語の統一化やデータの記述形式の標準化などをデータベース構築の際に義務づけるための制度設計もあわせて行うこと。
- データベースの開発とそのための技術開発(研究）とを緊密に連携させる仕組みを考案すること。
- 国として支援するデータベースや国として構築するポータルサイトの厳格な評価を行うための仕組みを検討すること。具体的にはモニター制度、利用者評価等を取り入れることを検討すること。
- 文献データベースとの連携のための仕組みを検討すること。
- 既存のデータベースだけでなく、ライフサイエンスの進展に対応した、新しい種類のデータベースあるいは従来にない発想に基づくデータベースの開発の振興にも十分配慮すること。

・データベース構築だけでなく、それを利用する技術開発の促進策も検討すること。
・長期的視点に立って、人材養成の促進を図る教育体制を構築すること。
・国家プロジェクトの成果活用の方向性を検討し、効果的な情報提供に向けた連携のための施策を考案すること。
・海外との連携をさらに進める方策を立案すること。特にアジア諸国のデータ生産者、バイオインフォマティクス研究者およびデータベース運営機関との連携について留意し、積極的な交流を図ること。

(2)基盤データベースの安定的な支援

我が国のライフサイエンス研究の基盤として欠かせないデータベース、世界的競争力の確保に向けて戦略的に重要なデータベースなどについては、安定的、永続的に支援することが必要である。現在この機能の一部は国立遺伝学研究所DDBJで実施されており、その他にもJSTバイオインフォマティクス推進センターによる支援が行われているが、データベースの数も限られており、また現在支援を受けているものについても、予算や期間の制約があり十分とはいえない。今後の更なる拡充が望まれる。なお、基盤データベースの安定的な支援に際しては、以下の点に十分な配慮・検討が必要である。

・我が国が独自に保有することが不可欠のものや世界的に存在が認められる知識基盤に限定して支援すること。その際、存在意義が認められる期間、安定的に維持するための必要額を十分に精査し支援すること。そのための評価基準として、論文への引用件数、アクセス数、一次データ量などによる定量的評価、外部有識者や利用者による定性的評価、およびサービス体制の充実度等を用いること。
・データベースの存在価値を維持するためのデータの収集・精査、サービス向上に直接関連する研究開発に限定して支援すること。新たな研究開発要素などは別予算（別途審査）（下記の(8)や(9)を参照のこと）で対応すること。
・ここで支援するデータベースについては、用語の統一化、データベースの記述形式や構造の標準化などの制約を課して、我が国のデータベースの統合化に寄与することを義務づけること。
・価値の高いデータベース、世界的に競争力のあるデータベースでありつづけるためには、それに関係した研究グループと密接な関係を常に維持していなければならない。そのための配慮を十分に行うこと。

(3)データベースの所在情報と利用法に関するポータルサイトの構築と運営

ライフサイエンス関係のデータベースに関する所在情報や利用法に関するポータルサイトを構築し運営することが必要である。これに関しても、いくつかの機関（JSTバイオインフォマティクス推進センター、国立情報学研究所など）でその試みはあるが、十分とは言えない。その理由は、常勤の専門家による運営が必ずしもなされていないこと、利用者からのフィードバックを常に活かしてサイトを最新のものに更新する仕組みが整っていないことによる。その背後には、このような仕事への評価の低さと予算面の手当てのなさの問題がある（国立情報学研究の活動は予算的な裏づけがあったが、平成17年度末で終了）。3-3節で紹介したように、我が国では数多くのデータベースが日々作られている。これらを十分に活用するためには、常に最新の情報を保持したポ

ータルサイトが不可欠である。このサイトの構築・運用に際しては、以下の点に留意すべきである。

- 何といってもポータルサイトにとって重要なことは、その網羅性である。日々、新しいデータベースが作られているような今日の状況では、個々の利用者が関連するデータベースすべての所在情報や利用法を把握するのは事実上不可能である。ポータルサイトにはデータベース作成者の意向も踏まえた上で、我が国のデータベースを漏れなく収載することが欠かせない。
- 一方、ポータルサイトに掲載されるデータベースが玉石混淆ではかえって混乱を招く。これを避けるため、引用数、アクセス数、データ量等を調査し、利用者側から見て分かりやすいよう、掲載するデータベースの分類をすること。
- 使いやすさによるデータベースの評価や利用法からみた分類などによるガイダンス機能の導入など、利用者の視点に立ったポータルサイトの運用に努めること。
- ポータルサイトの自動構築や評価のための技術開発もあわせて行うこと。
- ライフサイエンス分野の研究者、技術者を主たる対象とするが、一般の医療関係者あるいは育種家といった利用者も想定し、日本語での情報提供にも十分配慮すること。

(4)統合データベースの開発とそのための研究開発の促進

データベースの統合化に関しては、我が国においてもいくつかの機関でそれぞれの取組みが行われている。それらには一長一短あり一概には評価することはできないが、いずれも我が国のデータベース全般を統合化するという視点は弱い。その理由は、そもそもそのような使命を負わされてわけでもないし、権限があるわけでもなく、そのための予算の裏づけがあるわけでもないからである。JSTバイオインフォマティクス推進センターにおいても、データベースの高度化・標準化が謳われているが、統合化は必ずしも視野には入っていない。しかしながら、上述のポータルサイトの構築・運営だけでは、我が国の様々なデータベースの価値を十分に引き出すことはできず、ライフサイエンス研究のみならず産業界からの要請にも応えることはできない。多種多様なデータが生物的医学的に整理された形で統合されなければ、膨大なデータの洪水に流されてしまうだけになってしまい、ライフサイエンスの発展が止まってしまう。逆に、バラバラだったデータベースを統合化することができれば、これまで別々のデータベースに収められていたデータ間の潜在的な関係（例えば、遺伝子と疾患と薬剤との間の新たな関係やゲノムの進化と表現型の進化の間の対応関係）を見出すことが可能になる。ポータルサイトだけでは、このような新たな知識の発見を直接的に支援することはできない。データベース構築の大きな目標の一つはそこから新たな発見をすることにあり、統合化はまさにそのためのものである。一朝一夕には無理でも、我が国のデータベースの統合化に向けた研究開発を強力に、かつ、地道に推し進める必要がある。

ただし、統合化と言っても生命階層のどのレベルの、どのような知識を発見したいのか、どのようなことに統合データベースを使いたいのかによって、その目指すところ、意味するところは異なってくる。仮に目指すところが同じでもいろいろなアプローチがありうる。そのため、我が国としてどのようなアプローチでどのような統合化を目指すべきかに関しては、将来のライフサイエンスの動向や産業界からのニーズも十分踏まえた検討を行い、その議論に基づいて推進を図るべきである。幸い、現在、科学技術振興調費「科学技術連携施策群の効果的・効率的な推進」の一テーマとして調査研究が進められているところでもあり、その結果も踏まえて、前記(1)の戦

略立案機能の中で推進策を練ることが望ましい。

ところで、どのような統合化を目指すにせよ、統合化にあたっては、そのための用語や概念の統一化、データベースの記述形式や構造の標準化が前提となる。これらの中はすでに欧米で開発が進んでいるものもあり、それらを採用することも考えられるが、5節の「データベース整備戦略の基本的な考え方」に述べたように、我が国の特徴や強みが十分に発揮できるように十分な配慮・検討が必要である。

この他にも、データベースの統合化とそのための技術開発に向けては、以下の点に十分な配慮・検討が必要である。

- 国が支援するデータベースの構築者に対し、情報提供や技術指導を行うなど十分な連携をとり、用語の統一や記述形式の標準化を図ること。
- データベースの専門家（特にバイオインフォマティクス研究者）だけでなく、実験研究者や医療やバイオ産業に従事する人でも簡単に使えるような検索ソフトの開発や日本語環境の整備にも努めること。
- 欧米の後追いにならず、次世代の統合化を先取りするためにも、最先端の情報処理技術の活用や開発を行うこと。例えば、画像情報や新しい計測機器の出力結果等、新しい形式のデータに対応した情報処理技術や、新たな情報共有の枠組みのための情報処理技術を開発すること。
- 概念や用語の統一が統合化の鍵を握ることから、また我が国独自の特徴を出す意味からも、分野毎に、実験系の研究者と情報系の研究者の双方からなる専門家集団を形成し、それらの専門家集団の知識の融合に基づく統合データベースを目指すこと。
- 上のことと関連するが、データベース構築には実験研究者も深く関与できるような体制作りが必要である。

(5)維持が困難になったデータベースの受入れ

4-1節「データベースの問題点」に述べたように、各機関や各プロジェクトで開発されたせっかくのデータベースが、予算が切れると維持更新されなくなってしまうという問題がある。これに関しては、現在は研究者あるいは研究室の自発的な努力に頼るしかない状況であり、我が国のライフサイエンスにとって由々しき問題である。当然のことながら、すべてのデータベースを管理し続けるのは意味もないし不可能であるが、存続することが重要と判断されたものに関しては十分な支援が必要である。すなわちプロジェクトや科研費などの研究費が終了するなどして維持が困難になったデータベースの受け皿を、国として用意する必要がある。もちろん、闇雲に受け入れる必要はなく、存続価値を十分に厳正に評価して受入れや支援を判断すべきである。その際、以下の点に留意すべきである。

- ライフサイエンスの進展とともに、支援しなくてもよくなるデータベースも出てくるが、その一方で新たに支援すべきデータベースも出てくる。このような変化に柔軟に対応できるような制度（例えば、データ産出プロジェクトの設置に際しては、そのプロジェクト経費の一部をプロジェクト終了後も一定期間データベースの維持更新が可能なように積んでおくことを義務付けるなど）を導入すべきである。
- 文科省以外の省庁が整備したデータベースについても受け入れを検討すること。その際、内閣府の委員会、調査なども踏まえて検討すること。

・データベースの受け皿機関への移管に関しては、権利関係、事務手続きなどに配慮すること。
・ここで支援するデータベースについても、移管する際に、可能な限り、用語の統一化、データベースの記述形式や構造の標準化などの制約を課して、我が国のデータベースの統合化に寄与することを義務づけること。

(6)文献情報との連携

3-1節「データベース開発の世界的な動向」に述べたように、機能情報のデータベース化が重要な課題になりつつある。機能情報の多くは論文の中にテキストとして記述されていることから、文献中に記述されたデータや知識と、配列や立体構造などの実験データとの連携と統合に今後取り組まなければならない。米国NCBIでは、同じ組織で実験データも文献データも管理されていることから連携は比較的スムーズであるが、我が国ではこれまで別々に扱われてきたことから、今後連携を図っていく方策を講ずる必要がある。具体的には、遺伝子名や塩基配列のアクセッション（用語解説参照）などによる共通識別キーでの統合的検索を可能とするほか、ライフサイエンス分野の知識を計算可能な形へ変換し、概念対概念の関係を自動生成することにより、増大する論文データに対応できる知識の体系化を実現する必要がある。また、これにより提示される知識体系を活用しつつ、各データベースで利用されている各種用語の標準化を図る必要がある。また、このようなことを可能とする技術開発（概念・知識の収集の自動化やデータベースからの知識発見など）を並行して進める必要がある。なお、これらは上記(4)の「統合データベースの開発とそのための研究開発の促進」と重なる部分がいくつかあり、十分な連携のもとに進める必要がある。

(7)アノテーション（情報解読による実験データの注釈付け）の実施

基盤データベースの支援や維持が困難なデータベースの受入れ、さらにデータベース統合化および文献情報との連携といった活動と連動して、我が国で産出されたデータにもかかわらず未解析、未解釈のまま放置されている種々の実験データの意味付け（生物学的、医学的な解釈）を強力に推進すべきである。また、すでにアノテーションされているものでも正確さを欠くものもあり、それらについても再度アノテーションを実施すべきである。これについては、現時点で未解析・未解釈・不正確なデータのアノテーションを実行するだけでは不十分である。今後出てくるデータに対しても常に最新の技術、知識をもった専門家によるアノテーションを施す体制の確立が望まれる。アノテーションされていないデータは統合化しても意味がないし、ポータルサイトで所在が明らかになっても利用価値は低い。アノテーションを実施する際に留意すべき点は以下の通りである。

・アノテーションは独自の基準でバラバラに行うのではなく、上記(4)「統合データベースの開発とそのための研究開発の促進」や(6)「文献情報との連携」で開発された用語やガイドラインに基づいた注釈を行うこと。これによりデータベースの統一化が可能となる。
・実験系と情報系の研究者が協力できる体制を構築して、より正確で意味のある情報解読・注釈付けを実施すること。
・cDNA、イネゲノム、微生物ゲノムなど日本の強みを発揮できるデータについては、統一基準でより信頼性の高い形での再アノテーションを実施し、それを公開データベースに反映することを検討すること。

(8)新たなデータベース構築への投資

上述の基盤的データベースや評価の確立したデータベースの安定的な支援のほかに、ライフサイエンス研究の進展に対応した新たなデータベース、新たな発想に基づくデータベースの構築にも投資すべきである。これは、現在一部科研費特定領域研究「ゲノム４領域」やJSTバイオインフォマティクス推進センターで実施されているが予算や期限が限られており十分ではない。データベースには長期的視点が必要である。今後このような観点にたった支援制度を是非とも設けるべきである。長期的視野に立つといっても、新しく作られるデータベースは玉石混淆である。5年程度の時限を設けた競争的研究資金により実施することが望ましい。そこで評価が確立したものついては、例えば、上記の(2)「基盤データベースの安定的な支援」により支援することが考えられよう。なお、新たなデータベース構築への投資を行う際には、以下の点に十分な配慮・検討が必要である。

- ・ここで支援するデータベースについては、最初から用語の統一化、データベースの記述形式や構造の標準化などの制約を課して、我が国のデータベースの統合化に寄与することを義務づけること。
- ・新たなデータベース構築は機関で行ってもよいし、数人から個人の研究者レベルで行ってもよい。個人で行う場合の支援策に関しては、特に若手研究者が行う場合には、データベース開発に対する研究者の理解が不足している現状を考慮して、任期付きあるいは終身雇用の職をどこかに用意するなどの点に配慮すること。
- ・データベースの構築そのものでなくても、それの基盤となる、分散処理、高速通信、データベースマネジメントシステム等の基盤的技術開発を支援することも必要。

(9)データベースを活用した研究（バイオインフォマティクス）の促進

当然のことながら、データベースはそこから有用な知識を発見してこそ意味がある。逆に言えば、そのことを見越してデータベース開発を進める必要がある。そこで、データベース構築への支援と並行して、それを活用する技術の研究開発、いわゆるバイオインフォマティクスの促進も図る必要がある。バイオインフォマティクスそのものは、科研費特定領域研究「ゲノム４領域」やJSTバイオインフォマティクス推進センターなどで振興が図られているが、データベース構築と一体となった研究開発は必ずしも活発には行われていない。今後この面での支援策を講ずる必要がある。また、従来の施策の更なる拡充を図る必要がある。なお、これに関しては、若手研究者の育成、そのための任期付きあるいは終身雇用の職の確保、競争的な資金制度におけるバイオインフォマティクス分野の研究と連携、などに十分配慮して遂行すべきである。

(10)データベース開発のための人材養成

いくつかの大学において、21世紀COEプログラムや科学技術振興調整費人材養成などの支援を受けながら、バイオインフォマティクス分野の研究者や技術者の養成が行われている。しかしながら、質の高いデータベース構築を行う上で不可欠の人材である、アノテータ（データに生物学的医学的な解釈を加える専門職員）やキュレータ（データベースの編集作業に従事する専門職員）を目的としたものはほとんどない。経済産業省の産業技術総合研究所生命情報科学研究センターや国立遺伝学研究所で一部実施されてはいるものの十分ではない。問題は、教育する側にあるのではなく、受け手の少なさにある。それはアノテータやキュレータの技術を身につけても我が国

にはその職がないからである。また、そのような仕事の重要さへの理解が不足しているからである。我が国で世界的に競争力のある、また、意味付けがきちんとされた、有用なデータベースを開発するには、まずはアノテータやキュレータの安定的な職を数多く確保するとともに、それに相応しい人を養成することが不可欠であり、そのための体制を早急に確立する必要がある。また、そのためにその後の将来の処遇（キャリアパス）につながるような学会の認定資格などの方策も検討する必要がある。高度に専門的な知識や技術をもったアノテータやキュレータを養成するには、振興調整費人材養成プログラムあるいは大学の専門教育との連携がなくてはならない。このことを十分に考慮した人材養成の仕組みを構築する必要がある。上記の観点は、データベースのシステム開発や運用を専門的に担ういわゆるシステムエンジニアやオペレータの育成に関しても言えることである。

6-2 推進体制

6-1 で示した推進方策(1)から(10)を具体的に、かつ、効率よく遂行するために、以下に述べるような体制の整備を提案する。ただし、この提案は、委員会として中長期的な視点から望ましいと考える姿を示したものである。

- ・関係省庁間の連携のための戦略委員会の設置（内閣府総合科学技術会議の議論を踏まえて決定）
- ・関係省庁のデータベース関係機関による連携、調整のための枠組み
- ・関係省庁のデータベース関係機関による連携、調整のための枠組みの中核的機能を担う体制の整備（文部科学省が整備）

それぞれについて、以下にその役割や機能を述べる。

(A) 連携のための戦略委員会

データベースに関する関係省庁間の連携のため、前記 6-1 節「推進方策」の(1)「データベースの現状調査、評価、戦略立案機能の充実」に述べた役割、すなわち司令塔的な役割を担う。具体的にはデータベースの現状や動向、ニーズを定常的に調査し、それに基づき、我が国のデータベース戦略や構築活動を評価し、ライフサイエンスに関わるデータベースの整備戦略を練る。さらに、後述の関係機関による連携、調整のための枠組みの種々の活動を監督・指導する。上記枠組みに属する各機関が有機的に連携しているか、効率的に予算が使われているか、真に役立つデータベースを構築しているか、などを常に監視し、必要に応じてそれらに指導を行うものとする。

なお、統合化データベースを含むライフサイエンス基盤整備は、第３期科学技術基本計画に基づくライフサイエンス分野推進戦略の戦略重点科学技術に位置づけられている。そこで、連携のための戦略委員会のあり方については、総合科学技術会議において、今後詳細を検討することが適当である。総合科学技術会議では、統合化データベースを含むライフサイエンス基盤整備を、第３期科学技術基本計画ライフサインエス分野別推進戦略の戦略重点科学技術に掲げ、今後の推進方針の検討中である。それを受けて文部科学省としては今後の方向性を踏まえて、詳細を検討することとする。

(B) 関係機関による連携、調整のための枠組み

上記の戦略委員会が立案した計画を具体的に実行に移す組織として、各省庁のデータベース関係機関の連携、調整を行う枠組みを設ける。後述するように、統合データベース構築や関係機関による連携、調整及び戦略委員会の計画を実際に遂行するためには、中核的機能を担う体制の整備が不可欠であるが、しかしながらこれだけでは十分ではない。6-1 の(2)の留意点で述べたように、データベースはそれを構築したり利用したりする研究者グループと密接な関係を常に維持することが必要である。そのため、ある種のデータベースは研究者グループのいる機関にそれぞれ分散して保有し、それを連合体として有機的に連携させることが望ましい。関係機関による連携、調整のための枠組みはまさにそのためのものである。

(C)中核的機能を担う体制

上述したように、データベースは研究者グループのいる機関にそれぞれ分散して保有し、それを連合体として有機的に連携させることが望ましい。しかしながら、我が国で作られている多くのデータベースが共通の情報処理を行っていること、また、共通に使うデータを重複して持っていることから、データベースは集中的に構築維持管理したほうが効率がよい面が多々ある。また、標準化や統合化を図るには、連合体では不十分でありそれらに関して強力な指導力を発揮できる能力をもった専門家集団が必要である。すなわち、我が国におけるデータベース構築の効率化、標準化、統合化のためには中核的機能を担う体制の整備が欠かせない。連携のための戦略委員会で立案された計画を漏れなく速やかに遂行するためにも、また、人材養成や国際対応の観点からもこのような中核となるものが不可欠である。

そこで、上記(B)の関係機関による連携、調整のための枠組みの中核的機能を担う体制を置き、6-1 節の「推進方策」の(2)から(10)すべてを担当させることとする。また、当該体制では日常的にデータベース開発の世界的動向や我が国のデータベース構築活動、および様々な分野の利用者のデータベースに対するニーズを調査・評価し、それに基づきデータベース戦略を提案するなどして、上記のデータベース連携のための戦略委員会を補佐する、すなわち、推進方策の(1)にも貢献する。なお、「推進方策」の中の(2)基盤データベースの安定的支援、(7)アノテーション、(10)人材養成、については、必要に応じて外部へその役割を委託する。また、若手研究者の受け皿となる職（任期付きあるいは終身雇用）を用意して、新たなデータベース構築(8)や利用技術(9)の振興に努める。

この中央に設置された体制には、自前でのデータベース開発や外部のデータベース構築を支援するために、一定規模の計算機資源（中央データベース・サーバ）を保有させるものとする。この中央データベース・サーバの利用に際しては、最先端・高性能汎用スーパーコンピュータ（平成 18 年度より整備）との連携も検討する。この他に、データベースの広報、普及啓発活動のために、シンポジウムや講習会・トレーニングコースの開催やホームページの整備を行わせ、あわせて国際的な連携や産業界との連携のための活動も実施させる。なお、中核的な機能を担う体制は、優秀な人員の確保や養成の面からも、データベース整備が時間のかかる作業であるという面からも、5 年から 10 年にわたる年次計画を立てそれに基づいて段階的に整備を行ってゆく必要がある。

なお、上記の(A)(B)(C)という３つの機能・制度の設置、運用に際しては、その役割分担を明確にし、透明性・公平性・客観性を十分に担保した形で進めなければならない

6-3 中核的機能を担うための体制案について

現在、想定している中核的機能を担うための体制案を図 1 に示す。これはあくまでも例示に過ぎないが、統括のもと以下の 5 つのチーム構成とする。

・ポータルサイト構築運用チーム：(3)「データベースの所在情報と利用法に関するポータルサイトの構築と運営」を主に担当。また、日常的にデータベース開発の世界的動向や我が国のデータベース構築活動を調査・評価し、それに基づきデータベース戦略を提案するなどして、前述の連携のための戦略委員会を補佐する。

・データベース運用チーム：(2)「基盤データベースの安定的な支援」、(5)「維持が困難になったデータベースの受入れ」、(7)「アノテーション（情報解読による実験データの注釈付け）の実施」および中央データベース・サーバの運用に関わる業務を主に担当。必要に応じて、外部機関にこれらの業務を委託する。

・統合データベース開発チーム：(4)「統合データベースの開発とそのための研究開発の促進」、(6)「文献情報との連携」に関わる業務を主に担当。

・技術開発チーム：データベースの運用、統合化、その他必要となる様々な技術開発を行う。(8)「新たなデータベース構築」、(9)「データベースを活用した研究（バイオインフォマティクス）」に関する業務も一部担う。

・国際対応、産学連携チーム：国際対応や産学連携の業務を担う。それに加え、(10)「データベース開発のための人材養成」のための各種教育、トレーニング、シンポジウム、ホームページ作成、大学連携、などに携わる。

なお、統括のもとに運営委員会を設けて、適宜運営に関して助言を求めることとする。

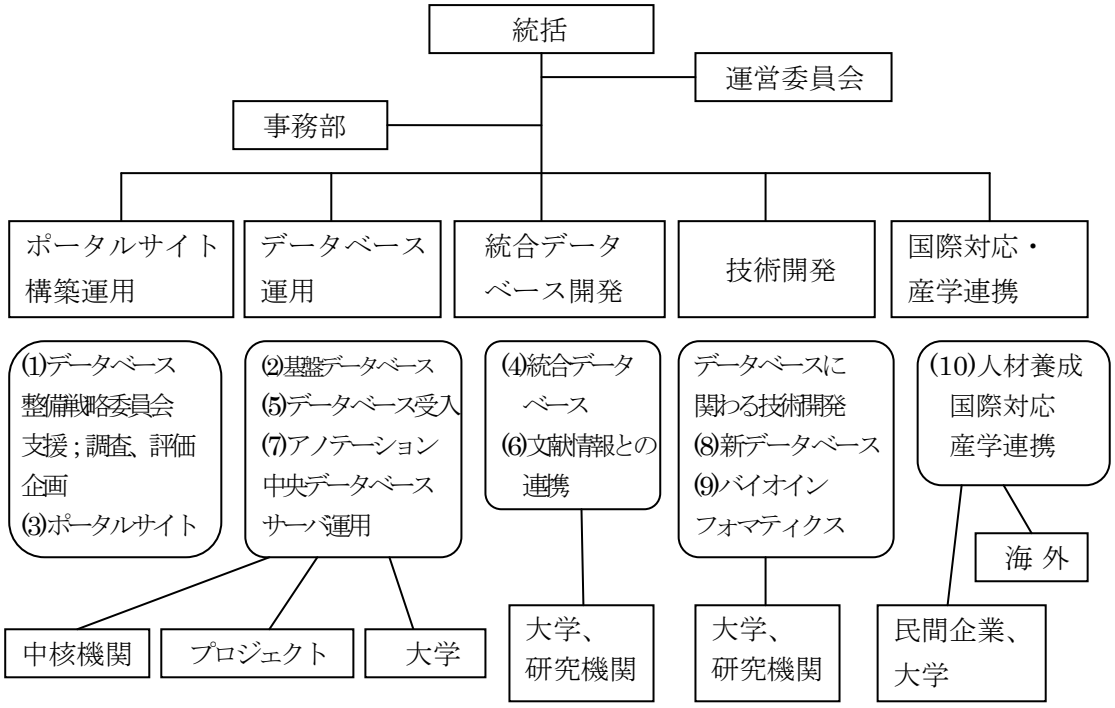


図 1　中核的機能を担うための体制案

7. 緊急に取り組むべき課題

ライフサイエンス研究は進展が速い。その成果であるデータベースも例外ではない。1年、2年の遅れが取り返しのつかない事態を招く可能性もある。そのような事態が起きないようにするには、6節「推進方策とそれを実現するための体制」を一日でも早く実行に移すことが肝要である。しかしながら、予算の確保や体制の整備には、早くても1年ほどの時間を要する。そこで、それほどの予算や体制を必要としないもののので、かつ、緊急性を有するものを6節で示した10個の課題、あるいは(A)から(C)の3つの機能・制度の中から優先度の高いものを選び出し、実行に移すことが望まれる。

また、すでに指摘したとおり我が国において整備されているデータベースの数は把握されているだけでも二百弱を数える。また、それらのデータベースは様々な背景に基づき多様な主体・資金によって作成されたものである。従って今後の整備にあたっては、始めは国家プロジェクトの成果としてのデータベースを対象とし、その後順次対象を広げていくなど、段階的な整備の計画を立案することが望ましい。

まず、1番目に(A)連携のための戦略委員会の設置は、すべての推進方策の基盤となるもので、これを最も急がなければならない。戦略委員会の設置を支援するために十分な透明性や客観性を確保した上で早急に作業グループを整備し、 6-1の(1)「データベースの現状調査、評価、戦略立案機能の充実」を図るべきである。

2番目は、6-1節の(3)「ポータルサイトの構築・運用」に向けたポータルサイトの設計と試作品の作成である。試作品を早急に構築し、利用者に試験的に使ってもらい、その意見を反映させて本運用のサイトを設計することが重要であろう。また、可能なら、我が国で作られた主なデータベースに関してはその試作品に盛り込み、産業界などのからの要望に直ちに応えるべきであろう。各データベースの使いやすさの指標や利用法による分類手法の開発などデータベースの評価法・分類法の開発にも急いで取り組むべきであろう。

3番目は、6-1節の(4)「統合データベース開発」に向けた、用語の統一、記述形式の標準化、データの共有化およびそのための技術開発である。これも、今後開発するデータベースやポータルサイトの基盤になるものであり、これの開発を急がなければならない。

4番目は、何といっても(10)の人材養成である。これにはある程度長い時間がかかるのでできるだけ前倒しで実施することが望ましい。具体的には、アノテータやキュレータの確保、養成に着手すべきである。大学や研究機関や学会と連携して、また、民間企業の力も借りて、集中的な講習会を開くなどして、潜在的なアノテータやキュレータの掘り起こしや教育を実施すべきであろう。

なお、上記以外の(2)、(5)、(6)、(7)、(8)、(9)に関しては、JSTバイオインフォマティクス推進センターや科研費をはじめとした既存の取組みはあるものの、可能な限り急いで取り組むことが重要である。

8. おわりに

ライフサイエンスにおけるデータベースの位置づけは、研究成果の整理・編集だけを目的とする所から、様々な階層に対する様々な種類のデータの統合化により、新たな発見や俯瞰的な理解を与えるだけでなく、個々のデータが持つ生命科学全体における意味を鮮明にし、対応する個々の研究自体の意味とその将来への方向付けを明確にするという極めて重要な役割を担うようになってきた。データベースの整備が覚束なければ、ライフサイエンスやバイオ産業の未来も危ういと言えるほどの存在にまでなってきた。

もちろん、これらデータベースに蓄えられた文献データをはじめ、ゲノム、プロテオーム、生体分子構造、遺伝子やタンパク質の発現や分子間相互作用、ネットワークやパスウェイ、臨床疾病などのデータそのものは、日本はもとより世界中のライフサイエンス研究者の努力と社会からの請託の結晶であり、それゆえ、データベースは、長い歴史における人類の叡智をまとめた宝とも言えよう。この宝物をどのように活かし、さらに発展させられるかが現在、問われているのである。この活用とは、直接的なライフサイエンス研究に対してだけでなく、産業界、一般社会への活用も含まれ、特に多数のデータベースの連携や統合による高度化によって、データベース構築時には考えもしなかった異分野を横断する新たな応用が実現し、生命科学が全く新たな次元での発展を迎えられるものと確信している。

本報告書では、このような現状の認識と将来への展望をもとに、これまで我が国においてあまり議論されることがなかったライフサイエンス分野におけるデータベース、その中でもとくにライフサイエンスの基盤となるデータベースの整備に対して、問題意識を明確にした専門家による議論を通じて様々な問題点を鮮明にし、それを解決するための戦略と具体的な施策のあり方を論じた。本報告書の結論を一言で要約すると、ライフサイエンスのデータベース整備に関して、省庁の枠を超えた権限と責任をもった、国家の司令塔的な役割を担う連携のための戦略委員会の設置、そこで立案された計画を実行する関係機関による連携、調整のための枠組みおよびその中核機能を担う体制の整備が不可欠であるということである。

なお、本報告書の冒頭に述べたように、ライフサイエンス分野には本報告書ではあまり触れなかったデータ（生物資源等の研究用材料に関するもの、医療現場で用いられる臨床情報や医薬品情報、化合物の構造や毒性情報、食品の成分や安全性に関するもの、作物や家畜の育種に関するもの、産業上有用な微生物の情報、など）が数多く存在する。これらに関しては、現在内閣府科学技術連携施策群（ポストゲノム）で進められているような各省庁連携の体制と連携をとって取組を進めることが望ましい。また、緊急に取り組むべき課題については、研究の進展が速いことから、その決定過程における透明性には留意しつつ、遅滞のない施策の立案と実施が望まれる。

データベース整備戦略作業部会委員名簿

（委員）

	秋山　泰	（独）産業技術総合研究所生命情報科学研究センター長
	江口　至洋	三井情報開発（株）フェロー
	宇高　恵子	高知大学医学部教授
	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
	鎌谷　直之	東京女子医科大学附属膠原病リウマチ痛風センター所長
◎	郷　通子	お茶の水女子大学学長
○	五條堀　孝	国立遺伝学研究所生命情報・DDBJ 研究センター長
	佐藤　清	（社）バイオ産業情報化コンソーシアム事務局長
	菅原　秀明	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	東京大学大学院新領域創成科学研究科教授
	田中　博	東京医科歯科大学情報医科学センター長
	中村　春木	大阪大学蛋白質研究所教授
	姫野　龍太郎	（独）理化学研究所情報基盤センター長
	深海　薫	（独）理化学研究所バイオリソースセンター　情報解析技術室長
	細江　孝雄	（独）科学技術振興機構理事
	宮野　悟	東京大学医科学研究所教授
	山本 博一	アステラス製薬㈱研究本部研究企画部日本橋部長

◎主査、○主査代理

（参考人）

大久保　公策	国立遺伝学研究所生命情報・DDBJ 研究センター教授
楠木　正巳	大阪大学蛋白質研究所助教授
藤山　秋佐夫	国立情報学研究所教授

データベース整備戦略作業部会における審議の過程

○第１回　平成１７年８月１２日

（１）データベース整備戦略作業部会について
（２）ライフサイエンス研究に関するデータベースに係る施策について
（３）データベース整備の現況について
（４）課題整理と今後の展開について

○第２回　平成１７年１１月１０日

（１）内閣府総合科学技術会議の科学技術連携施策群における統合データベースの動きについて
（２）科学技術振興機構　バイオインフォマティクス推進事業における「生命情報データベース高度化・標準化」研究開発課題の公募について
（３）データベース整備に係る作業部会委員の意見概要について

○第３回　平成１８年１月１６日

（１）データベース整備に係る作業部会委員の意見概要について
（２）データベース整備戦略として必要な機能について
（３）データベース整備戦略として必要な組織体制について

○平成１８年１月１９日
ライフサイエンス委員会に対してデータベース整備戦略作業部会における議論の状況を報告

○第４回　平成１８年２月２８日

（１）データベース整備戦略作業部会報告書骨子（案）について
（２）平成１８年度「統合データベースプロジェクト」について

○第５回　平成１８年３月２４日

（１）データベース整備戦略作業部会報告書（案）について
（２）平成１８年度「統合データベースプロジェクト」について

○第６回　平成１８年５月１１日

（１）データベース整備戦略作業部会報告書（案）について

付録：用語解説

ヒト完全長cDNA(p.6 下から 8 行目)：cDNAとは、個々の遺伝子が、ゲノムDNAから読み取られてタンパク質を作る際の鋳型となる塩基配列情報を有するDNAであり、その配列が完全な状態で取得できたものを完全長cDNAという。これを得るために日本独自の技術が使われている。

オントロジー(p.8 上から 9 行目)：知識、語彙、概念などと、それらの間の関係を明確にした辞書。生物学では異なる分野で同じ用語を異なる意味に用いたり、異なる用語で同じ意味を表したりすることがあり、これを明確化することを目的に遺伝子オントロジー (GO)プロジェクトなどが進められている。

パスウエイ(p.8 上から 12 行目)：生体分子間の相互作用の一連の流れのことである。生命科学の進展により、個々の生体分子の役割の解析から多様な生体分子の一連の相互作用がおりなすシステムに解析の興味が移り、代謝パスウェイや信号伝達パスウェイといった高次の生命現象に関わるパウスェイの解析が精力的に進められている。

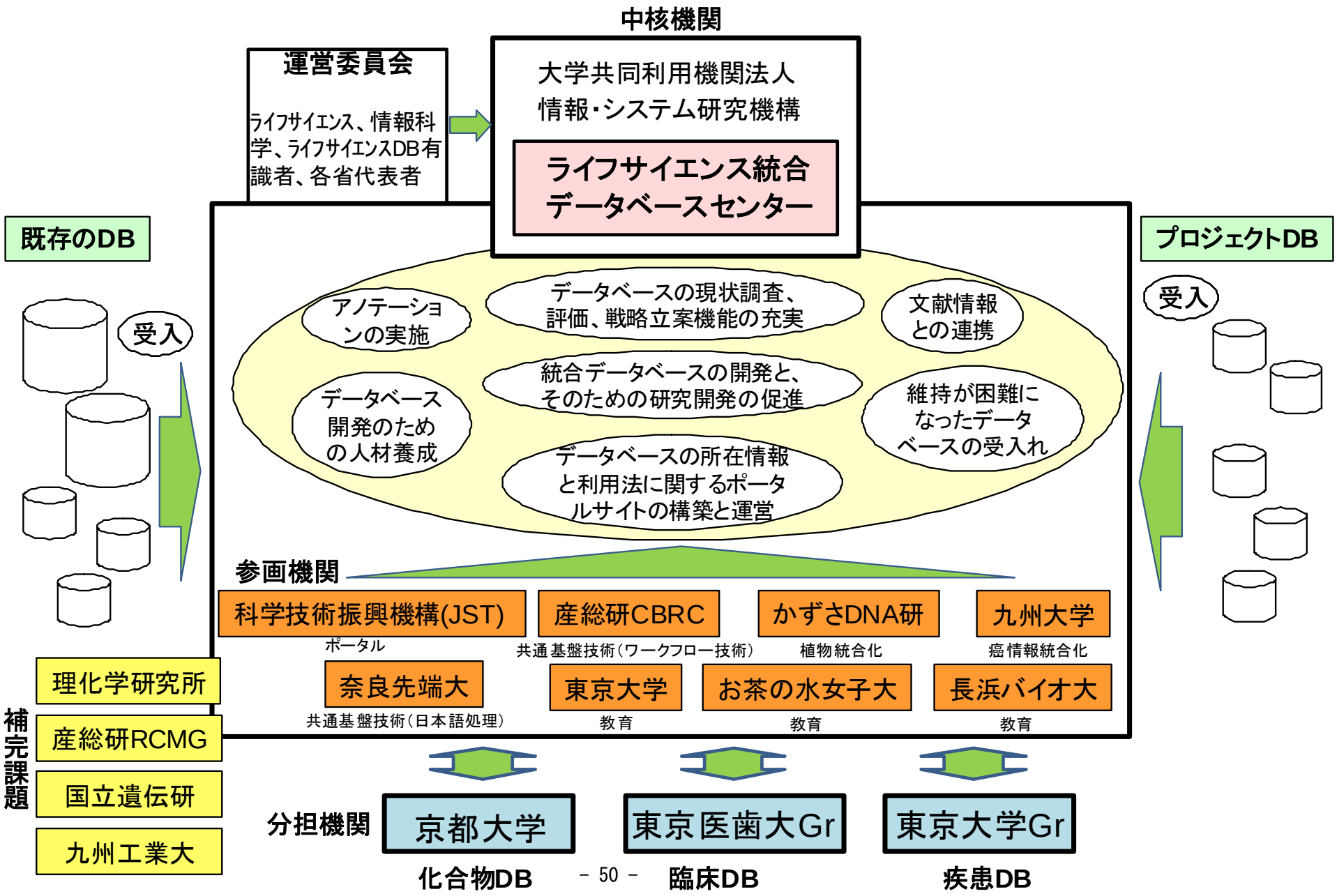
SNP解析(p.11 上から 13 行目)：SNPとは、一塩基多型と呼ばれるヒト一人一人の遺伝子中に見られるDNA配列の違いのことである。これを解析することで、疾患感受性や薬剤応答性の個人差が分かり、疾患関連遺伝子の同定や薬剤による副作用を避ける診断方法の確立が期待できる。

ファーマコジェネティックス(p.11 上から 14 行目)：医薬品の効果や毒性に及ぼす遺伝的特性の影響を調べる学問のことで、特に遺伝的多型が薬理作用に及ぼす影響をゲノムワイドに調べた情報をベースにしたテイラーメイド医療が注目されている。

アクセッション(p.24 下から 1 行目)：データベースに格納される個々のデータであるエントリー単位に割り振られた固有のＩＤ。DDBJに代表される国際塩基配列データベースなどでは、決まった文字数の英数字が一定の規則のもとに割り当てられる。

参考資料（２）

統合データベースプロジェクトの体制図



参考資料（３）

「ライフサイエンス分野の統合データベース整備事業」の平成１８年度受託実施機関の公募について

平成18年7月3日
文部科学省研究振興局
ライフサイエンス課

文部科学省では、平成18年度から「ライフサイエンス分野の統合データベース整備事業」を推進することとしております。このたび、本事業を実施する機関の公募を行いますのでご案内致します。

１．目的

現在、我が国は、第３期「科学技術基本計画」（平成１８年３月２８日閣議決定）のもとに、「科学技術創造立国」を目指して諸施策を実施しております。同基本計画においては、「抜本的な科学技術システム改革」が求められており、その中で２０１０年に世界最高水準を目指してデータベースを含む「知的基盤の戦略的な重点整備」を進めることとされています。さらに、同基本計画に基づき、総合科学技術会議が策定したライフサイエンス分野の推進戦略では、戦略重点科学技術の１つとして「世界最高水準のライフサイエンス基盤整備」が掲げられています。

生命情報の統合化データベースはライフサイエンス研究を支える基盤であり、その整備を進めるために必要な戦略の検討と技術開発等を行なうため、「ライフサイエンス分野の統合データベース整備事業」の平成１８年度受託実施機関の公募を行ないます。

２．ライフサイエンス分野の統合データベース整備事業」の概要

現在、我が国のライフサイエンス分野の国内主要データベースの統合化と継続的な維持方策の必要性が指摘されています。

そこで本事業では、我が国のライフサイエンス関係のデータベースの利便性の向上を図るため、我が国のライフサイエンス関係データベース整備戦略の立案・評価支援、データベース統合化の基盤技術開発、ポータルサイトの整備等を行い、統合化を推進します。

３．対象

国内の研究機関又は大学、大学共同利用機関法人（以下、「機関等」）を対象とします。（企業にあっては国内に法人格を有するもののみ。）

複数機関で実施体制を組んで申請して頂いても構いませんが、その場合は中核となる機関を定めた上で当該機関は他機関の研究進捗管理、文部科学省との連絡調整などプロジェクトの総合的推進の責任を負う必要があります。

なお、申請は機関の長（学長、理事長等）が行うものとします。

４．公募期間

平成18年7月3日（月曜日）～平成18年7月31日(月曜日)

５．スケジュール

平成18年

7月 3日（月曜日）　公募開始
7月31日（月曜日）　公募〆切
～8月中旬（予定）　審査
～8月下旬（予定）　委託契約、事業開始

※詳細については、文部科学省ライフサイエンスポータルサイト（http://www.lifescience-mext.jp/download/news/life_DBkoubo.html）より、公募要領・提出書類の様式をご参照ください。

（お問い合わせ先）
文部科学省研究振興局
ライフサイエンス課
担当:野田、松永
代表電話:03(5253)4111(内線 4381)
直通電話:03(6734)4104
e-mail:life@mext.go.jp

プレス発表資料

文部科学省
Ministry of Education, Culture, Sports, Science and Technology

平成１８年９月１３日
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」に関する受託実施機関の決定について

文部科学省では、平成 18 年度から「統合データベースプロジェクト」を推進しております。本プロジェクトにおいて取り組むこととしている「ライフサイエンス分野の統合データベース整備事業」について、このたび、受託実施機関を決定しましたので発表します。

１．事業の目的

「統合データベースプロジェクト」は、我が国の生命科学分野のデータベースを戦略的に統合するための戦略立案・評価支援、統合化のための基盤技術開発等を行うことにより、ライフサイエンス関係データベースの統合的活用システムを構築・運用し、幅広いライフサイエンス分野の科学技術の進展に大きく貢献することを目的としています。

２．決定した委託実施機関

外部有識者から構成される受託実施機関選考委員会（別紙１）における審査に基づき、以下の申請機関を受託実施機関として決定し、本プロジェクトを実施することとなりました。（平成１８年度事業の概要などは別紙２参照）

申請機関　　：大学共同利用機関法人情報・システム研究機構
研究代表者　：堀田　凱樹(大学共同利用機関法人情報・システム研究機構長)

(本件照会先)
研究振興局ライフサイエンス課
松永、石塚
TEL:03-6734-4369(直通)

（別紙１）

「ライフサイエンス分野の統合データベース整備事業」
選考委員名簿

主査

郷　通子	お茶の水女子大学　学長
鎌谷　直之	東京女子医科大学 附属膠原病リウマチ痛風センター　所長
榊　佳之	理化学研究所 ゲノム科学総合研究センター・センター長
末松　誠	慶應義塾大学　医学部　教授
山本　博一	アステラス製薬株式会社　研究本部研究企画部　部長

統合データベースプロジェクトの18年度事業概要

(別紙2)

【事業の目的】 我が国のライフサイエンス関係のデータベースの利便性の向上を図るため、データベース整備戦略の立案・評価支援、統合化及び利活用のための基盤技術開発、人材育成等を行い、ライフサイエンス関係データベースの統合的活用システムを構築・運用する。

【想定される成果】 これまでの研究成果の蓄積を網羅的・安定的に利用できるようになり、ライフサイエンス研究の発展に不可欠な基盤となる。また統合化アルゴリズムの開発等による既存データの新たな活用や、産業界・医学関係者などによる応用利用を通して新たな知見が得られる。

【18年度実施内容】

〇データベースの現状調査、評価、戦略立案
〇ポータルサイトの構築、運営
〇統合化技術の研究開発

【実施機関】 責任機関：大学共同利用機関法人情報・システム研究機構
参画機関：独立行政法人科学技術振興機構、国立大学法人九州大学
協力機関：東京大学、埼玉医科大学、(財)かずさDNA研究所、㈱三菱総合研究所、大阪府立成人病センター、(独)産業技術総合研究所　ほか

参考資料（４）

平成 19 年 1 月 10 日
文　部　科　学　省

平成 19 年度「ライフサイエンス分野の統合データベース整備事業」の受託実施機関公募について

第３期「科学技術基本計画」（平成１８年３月２８日閣議決定）に基づき総合科学技術会議が策定したライフサイエンス分野の推進戦略では、戦略重点科学技術の１つとして「世界最高水準のライフサイエンス基盤整備」が掲げられています。生命情報の統合化データベースはライフサイエンス研究を支える基盤であり、その整備を進めるために必要な戦略の検討と技術開発を行なうため、「ライフサイエンス分野の統合データベース整備事業」受託実施機関の公募を行います。

１．公募の受付期間
平成１９年１月１１日（木）～平成１９年２月８日（木）当日必着

２．公募概要
（１）本事業は、我が国のライフサイエンス関係のデータベースの利便性の向上を図るため、データベースの統合化を推進することを目的としています。
事業は、統合データベースの開発・整備に向けて、「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」の３つの柱にて実施します。
提案にあたっては事業の３つの柱のすべてを担う中核機関として、または中核機関の下で「統合データベース開発」の一部を担う分担機関として提案することができます。

（２）採択予定件数は中核機関１課題、分担機関１～３課題程度とします。

（３）平成１９年度の公募要領・申請書等、詳しくは下記ホームページをご覧下さい。

文部科学省ホームページ（http://www.mext.go.jp/）

なお、この公募は、平成１９年度予算の成立を前提に行うものであり、予算の成立状況によっては事業内容や実施予定額を変更する場合がありますので留意して下さい。

＜制度に関するお問い合わせ＞
文部科学省　研究振興局　ライフサイエンス課
担当：松永、石塚
TEL：03-6734-4369（直通）
E-mail：life@mext.go.jp　（注：始めの文字はＬの小文字です）

＜書類作成・提出に関するお問い合わせ＞
科学技術振興機構　キーテクノロジー研究開発業務室
TEL：03-5214-7990（直通）
E-mail：ltogoask@jst.go.jp　（注：始めの文字はＬの小文字です）

プレス発表資料

文部科学省
Ministry of Education, Culture, Sports, Science and Technology

平成１９年４月２日
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」に関する受託実施機関の決定について

文部科学省では、平成 18 年度から実施している「ライフサイエンス分野の統合データベース整備事業」について、このたび、平成 19 年度以降、本プロジェクトを実施する受託実施機関を決定しましたので発表します。

１．事業の概要

「統合データベースプロジェクト」は、我が国のライフサイエンス関係データベースの統合的活用システムを構築･運用し､幅広いライフサイエンス分野の科学技術の進展に大きく貢献することを目的としています（別紙１）。事業は、「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」の３つの柱にて実施し、事業の３つの柱のすべてを担う中核機関、および中核機関の下で「統合データベース開発」の一部を担う分担機関の体制で実施します。

２．決定した委託実施機関

外部有識者から構成される受託実施機関選考委員会（別紙２）において審査を行い、７件の申請機関から以下の機関（中核機関１件、分担機関３件)を受託実施機関として決定しました。

分担機関は連携して、化合物・医薬品、臨床・疾患等の医療に関わるデータベースの統合化を進めます。

【中核機関】
○申請機関　：大学共同利用機関法人情報・システム研究機構
　研究代表者：高木　利久(情報・システム研究機構　特任教授)

【分担機関】（医療に関わるデータベースの統合化）
○申請機関　：国立大学法人京都大学
　研究代表者：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　センター長)

○申請機関　：国立大学法人東京医科歯科大学
　研究代表者：田中　博(東京医科歯科大学情報医科学センター　センター長)

○申請機関　：国立大学法人東京大学
　研究代表者：徳永　勝士(東京大学大学院医学系研究科　教授)

(本件照会先)
研究振興局ライフサイエンス課
松永、石塚
TEL:03-6734-4106(直通)

統合データベースプロジェクト

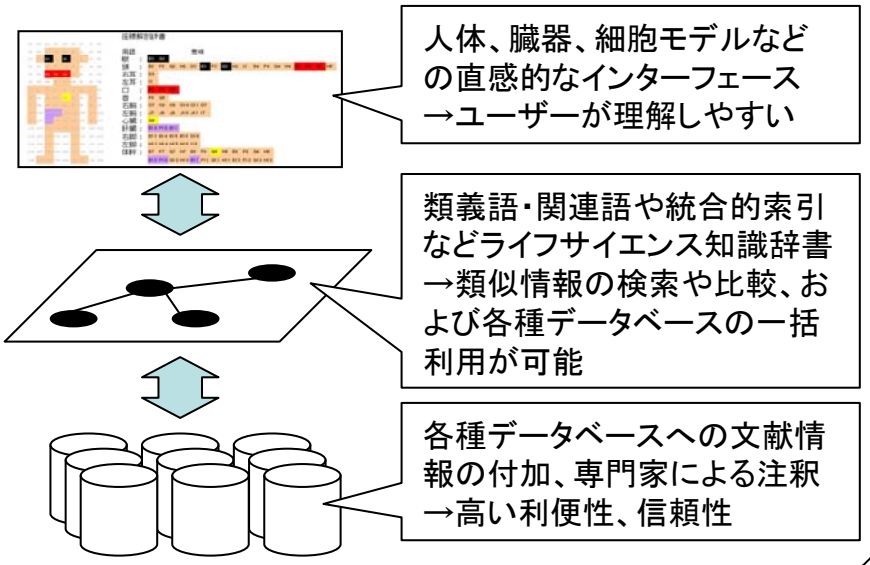
別紙1

【事業の目的】 我が国のライフサイエンス関係のデータベースの利便性の向上を図るため、データベース整備戦略の立案・評価支援、統合化及び利活用のための基盤技術開発、人材育成等を行い、ライフサイエンス関係データベースの統合的活用システムを構築・運用する。

【想定される成果】 これまでの研究成果の蓄積を網羅的・安定的に利用できるようになり、ライフサイエンス研究の発展に不可欠な基盤となる。また統合化アルゴリズムの開発等による既存データの新たな活用や、産業界・医学関係者などによる応用利用を通して新たな知見が得られる。

統合データベースとは？

ライフサイエンス関連データベースの統合的活用システム



【実現するための方策】

※17年度～19年度：
内閣府連携施策群にて実施。
それを受け、文部科学省において、18年度統合DB開始

○18年度先行着手（19年度以降継続）

- データベースの現状調査、評価、整備戦略立案
- ポータルサイトの構築、運営
- 統合化技術の研究開発

○19年度以降本格着手

- 中核的機関整備（公募）による総合的推進
- 統合データベースの開発、運営
- 文献情報との連携やデータへの注釈付加
- 新たなデータベースの構築や活用した研究
- 維持困難となった有用データベースの受入
- データベース開発のための人材育成

別紙２

「ライフサイエンス分野の統合データベース整備事業」
受託実施機関選考委員会委員名簿

鎌谷　直之	東京女子医科大学 附属膠原病リウマチ痛風センター　所長
末松　誠	慶應義塾大学　医学部　教授
中村　春木	大阪大学　蛋白質研究所　教授
深海　薫	理化学研究所　筑波研究所 バイオリソースセンター　情報解析技術室長
宮野　悟	東京大学　医科学研究所 ヒトゲノム解析センター　教授
山本　博一	アステラス製薬株式会社　研究本部研究企画部 日本橋部長

参考資料（５）

平成 19 年 7 月 31 日
文　部　科　学　省

平成 19 年度「統合データベースプロジェクト」補完課題の公募について

第３期「科学技術基本計画」（平成１８年３月２８日閣議決定）に基づき総合科学技術会議が策定したライフサイエンス分野の推進戦略では、戦略重点科学技術の１つとして「世界最高水準のライフサイエンス基盤整備」が掲げられています。生命情報の統合化データベースはライフサイエンス研究を支える基盤であり、その整備を進めるために必要な戦略の検討と技術開発を行なうため、文部科学省では統合データベースプロジェクト」を推進しております。

今回この事業を補完し、さらに充実させるため「統合データベースプロジェクト」補完課題を実施する機関の公募を行います。

１．公募の受付期間

平成１９年８月１日（水）～平成１９年８月３０日（木）当日必着

２．公募概要

（１）平成１９年度より、大学共同利用機関法人　情報･システム研究機構が中核機関として３つの分担機関（東京大学を中心としたグループ、東京医科歯科大学と大阪大学のグループ、京都大学）と連携してライフサイエンス分野のデータベースの統合化を進めています。

統合化を一層加速する観点から、今回、中核機関の示す統合化方針に従い、自ら保有するデータ又はデータベースを文部科学省の統合データベースに提供する事業を行う機関を募集します。

（２）平成１９年度の公募要領・申請書等、詳しくは下記ホームページをご覧下さい。

文部科学省ホームページ（http://www.mext.go.jp/）

＜制度、統合データプロジェクトに関するお問い合わせ＞
文部科学省　研究振興局　ライフサイエンス課
担当：田中、石塚
TEL：03-6734-4104（直通）
E-mail：life@mext.go.jp　（注：始めの文字はＬの小文字です）

＜書類作成・提出に関するお問い合わせ＞
科学技術振興機構　キーテクノロジー研究開発業務室
TEL：03-5214-7990（直通）
E-mail：ltogoask@jst.go.jp　（注：始めの文字はＬの小文字です）

プレス発表資料

文部科学省
Ministry of Education, Culture, Sports, Science and Technology

平成１９年１０月１５日
文　部　科　学　省

「統合データベースプロジェクト」補完課題を実施する機関の決定について

文部科学省では、平成 18 年度から実施している「統合データベースプロジェクト（ライフサイエンス分野の統合データベース整備事業）」について、プロジェクトを補完し、さらに充実させるための補完課題を実施する機関を決定しましたので発表します。

１．背　景

「統合データベースプロジェクト」は、我が国の生命科学分野のデータベースを戦略的に統合するための戦略立案・評価支援、統合化及び利活用のための基盤技術開発等を行うことにより、ライフサイエンス関係データベースの統合的活用システムを構築･運用し、幅広いライフサイエンス分野の科学技術の進展に大きく貢献することを目的としています。

統合化を一層加速する観点から、今回、中核機関（大学共同利用法人情報・システム研究機構）の示す統合化方針に従い、自ら保有するデータ又はデータベースを文部科学省の統合データベースに提供する事業を行う機関を平成１９年８月１日～３０日に一般公募し、この度、決定しました。

２．決定した実施機関

外部有識者から構成される補完課題選考委員会（別紙１）における審査に基づき、以下の実施機関（Ａ型課題３件、Ｂ型課題１件）を決定しました。

［各課題の実施機関に望まれる要件］
Ａ型課題：複数の分野･生物種の実験データを幅広く収集、保有又は整備、維持管理していること
Ｂ型課題：特定の分野･生物種の実験データについて網羅的に保有し、維持管理していること

【Ａ型課題】
○実施機関　：独立行政法人理化学研究所
代　表　者　：豊田　哲郎
課　題　名　：「植物オミックス情報および蛋白質構造情報」
○実施機関　：独立行政法人産業技術総合研究所
代　表　者　：成松　久
課　題　名　：「糖鎖修飾情報とその構造解析データの統合」
○実施機関　：大学共同利用法人情報・システム研究機構　国立遺伝学研究所
代　表　者　：五條堀　孝
課　題　名　：「塩基配列アーカイブのデータベース構築と統合への貢献」
【Ｂ型課題】
○実施機関　：国立大学法人九州工業大学
代　表　者　：皿井　明倫
課　題　名　：「生体分子の熱力学データと構造データの統合」

（本件照会先）
研究振興局ライフサイエンス課
田中、石塚
TEL：03-6734-4104（直通）

別紙１

「統合データベースプロジェクト」補完課題選考委員会選考委員名簿

	宇高　恵子	高知大学医学部　免疫学教室　教授
	金岡　昌治	大日本住友製薬（株）　執行役員　研究本部副本部長 兼　薬理研究所長
	高木　利久	東京大学大学院新領域創成科学研究科　情報生命科学 専攻　教授
	藤　博幸	九州大学生体防御医学研究所　微生物ゲノム情報学 分野　教授
主査	松原　謙一	（株）DNA チップ研究所　代表取締役　社長

参考資料（６）

平成１８年度　研究運営委員会／戦略作業部会　委員一覧

【研究運営委員会】

	氏　名	現　職
	秋　山　　泰	(独)産業技術総合研究所生命情報科学研究センター長
	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　基礎生物学研究所長(JST)
	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
	高　木　利　久	東京大学大学院新領域創成科学研究科情報生命科学専攻教授(JST)
	田　畑　哲　之	(財)かずさDNA研究所植物ゲノム基盤研究部長
	辻　井　潤　一	東京大学大学院情報理工学系研究科教授
	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
	増　保　安　彦	東京理科大学薬学部生命創薬科学科教授
	松　原　謙　一	(株)DNAチップ研究所代表取締役社長
	吉　田　光　昭	東京大学新領域創成科学研究科客員教授(東京大学名誉教授)
	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・システム研究機構理事　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生命情報・DDBJ研究センター長
◎	堀　田　凱　樹	情報・システム研究機構長

◎は責任者を表す。

【統合DB戦略作業部会】

氏　名	現　職
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
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
高　木　利　久	東京大学大学院新領域創成科学研究科情報生命科学専攻教授(JST)
田　畑　哲　之	(財)かずさDNA研究所植物ゲノム基盤研究部長
中　村　桂　子	JT生命誌研究館館長
増　保　安　彦	東京理科大学薬学部生命創薬科学科教授
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生命情報・DDBJ研究センター長
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研究センター教授
高　野　明　彦	情報・システム研究機構国立情報学研究所コンテンツ科学研究系教授
藤　山　秋佐夫	情報・システム研究機構国立情報学研究所情報学プリンシプル研究系研究主幹・教授

参考資料（７）

平成１９年度　研究運営委員会／作業部会　委員一覧

【研究運営委員会】

	氏 名	現　職
	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	大日本住友製薬(株)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・教授
	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・教授
	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礎研究領域ゲノムリソースセンター長
◎	松　原　謙　一	(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	情報・システム研究機構長

◎は委員長を、○は副委員長を表す。

【作業部会】

氏 名	現　職
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	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

参考資料（８）

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基盤整備
１８年度　研究成果報告書

平成１９年３月

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
国立大学法人　九州大学

本報告書は、文部科学省の委託業務として、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、国立大学法人九州大学が共同で実施した、平成１８年度の「ライフサイエンス統合データベース基盤整備」を取りまとめたものです。従って、本報告書の複製、転載、引用等には文部科学省の承認手続きが必要です。

目　次

（注）フッターにある（ ）付き番号は、
参考資料内のページ番号です。

１．プロジェクトの目的

より多くのライフサイエンス研究者等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等をストレスなく利用でき、より高度な研究ができる環境の実現とその持続可能化のために，情報・システム研究機構、九州大学，科学技術振興機構が共同して，以下の４つの業務を行う。

（１）ライフサイエンスおよび知識情報処理の識者にライフサイエンスDBの専門家を加えた研究運営委員会を組織し､十分な情報の収集･分析に基づいて統合化戦略を立案する。

（２）統合化に不可欠な知識表現法、情報共有技術および文献からの知識抽出技術等の開発に着手し、順次戦略立案のための調査やポータルでの利用者誘導に適用する。

（３）DBのカタログおよび解析ツールなどのWEBリソースのカタログを作成し、利用者を目的にかなったDBやWEBリソースに誘導するポータルサイトを構築する。

（４）データベースの構築維持に不可欠な人材の育成に努める。

２．データベース統合化戦略の立案および評価

２．１　データベース統合化戦略の立案および評価の実施計画

DB 統合化の戦略は１）利用者であるライフサイエンス分野の状況、２）素材である個別データベースの状況、３）利用可能な情報技術、の３つの動向を常に考慮しながら継続的にかつ柔軟に立案されねばならない。ここでは3分野からの専門家を集めた研究運営委員会を組織し、同委員会管轄下に調査組織（統合DB整備戦略作業部会）を設ける。

調査組織は、

（１）ゲノム注釈とデータベース間の連携における課題

（２）国内外のDBの俯瞰と質的量的比較

（３）ライフサイエンス分野の研究の俯瞰調査

（４）検索アルゴリズムを含めた知識情報技術の動向調査

（５）臨床情報や医療統計の現状調査

を行い、上記委員会に適宜報告しながら戦略立案を支援する。

なお、本テーマは情報・システム研究機構で実施した。

２．２　データベース統合化戦略の立案および評価の実施内容

ライフサイエンス、知識情報処理、ライフサイエンスデータベース（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DB整備戦略について議論した。上記委員会に対し、3 分野の動向に関する俯瞰データを提示し、適宜与えられる調査課題に対し答えることのできる体制（統合DB整備戦略作業部会）を構築した（研

究運営委員会並びに戦略作業部会の議論の詳細は、６．プロジェクトの総括的推進の項に記載)。また、この体制を利用して戦略立案の基盤となる情報の収集・分析を、以下に示すように行った。

（1）ゲノム注釈とデータベース間の連携における課題

代表的モデル研究植物であり、全ゲノム塩基配列が決定済みである「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の基本項目を調査し、それぞれのデータベース間の連携と課題の整理を実施した。

「イネ」のデータベースについては、Nucleic Acids Research のデータベース特集、国際イネゲノム塩基配列プロジェクト(IRGSP)及び農業生物資源研究所(NIAS)のデータベース検索 Web サイト、PubMed 検索で rice 及び database キーワードにして、AND 検索した結果から重複を除いて得られた、46 種類を対象とした。「シロイヌナズナ」のデータベースについては、同じく Nucleic Acid Research のデータベース特集、国立遺伝学研究所、かずさ DNA 研究所、理化学研究所ゲノム総合科学研究センターのデータベース検索 Web サイト等から重複を除いて得られた、25 種類を対象とした。

調査項目としては、エントリー構成、エントリー数及び、一次データと二次データの区別、関連データベースや一般的公共データベースへのリンクの構成、画面レイアウトや表示ツール、用意されている解析プログラム群といった Web インターフェース構成、定期リリースの頻度やバージョンや過去のバージョン参照可能かどうかといったデータベースアップデートの状況、OS や管理プログラム、マシンスペックや開発言語といったデータベース管理システムに関する事柄をベースに、発表論文や開発者への連絡方法、見易さ、応答性、信頼性などの観点からの検索結果表示の問題といった点まで含めた。

同時に、主として実験生物系のデータベースの利用者を対象に、「これらのデータベースのなかでよく利用するサイトはどこか」、「複数サイトを利用する場合に困っている点はないか」など、聴き取り調査と郵送によるアンケート調査を（全 188 名を対象に）実施した。これによって、主として実験の現場でデータベースを活用している研究者が抱いているゲノムベースのデータベースの連係に関する現状の課題と、将来のデータベース統合にむけた要望を調べ上げた。

調査結果から、データベースのよりよい統合化は、以下のような比較的多数のユーザが抱く不満を解消する方向で行うべきであることを読み取ることができた。

1）DB 作成の時間差や異なる収集方針による遺伝子名や ID の相違が多く混乱の元になっている。これを吸収あるいは関連付ける基盤サービスが必要。

2）論文掲載情報や利用者からのフィードバックが直ちに反映されないことへの不満も大きい。

3）誤りの多さを不満とする声がある一方、仮想遺伝子にもなんらかのヒントが欲しいという要望も多い。提供するデータの分類や格付はできないか。

４）植物の分子の研究に於いては、頻繁に生物種横断的な検索や比較を行う。そのような情報が取得できるサイトがない。

反面、個別に現状のデータベースをみた場合の使いやすさや内容の充実度に関しては約半数が肯定的であり、将来の統合化データベースの作成にあたっては、現在、利用頻度の高い個々のデータベースが保持している有用な情報を活かしつつ、齟齬を解消し連係させる形での統合化を考えていくべきであるとの結論が得られた。

（２）国内外 DB の俯瞰と質的量的比較

科学技術連携施策群の生命科学データベース統合に関する調査研究と JST-DB（WING）の情報の調査を行い、主に分子に関する網羅性の高いデータベースカタログを生成した。データベースカタログの各エントリーには、データベース型分類情報を追加し、主要なデータベースに関する日本語解説を整備した。その結果、分子情報に基づくデータバンクには、索引はあるが目次が欠如していることが多いことが明らかとなった。そこで、一般の研究者でも、分子レベルの生物学研究の現状を容易に俯瞰できるようにすることを目的として、各種データバンクの内容を目次的に表現しようと試みた。これによりデータバンクの内容について、個々のバンクを区別することなく、自在に一次データを引き出し利用することも可能になると考えられる。今年度は INSDC(International Nucleotide Sequence Database Collaboration)が管理する DNA バンクと遺伝子発現バンク（GEO、Gene expression omnibus）について総合データ目次を作成した。これにより、目次項目ごとのデータのダウンロードが可能になった。

１）DNA バンク目次

DNA 配列読み取りをおこなった論文では、論文投稿時に INSDC への DNA 配列の登録が義務付けられている。従って数十塩基の配列から完全なヒト染色体の配列まで、学術論文で新規に報告された DNA 配列は全て INSDC に登録されているはずである。ここでは、INSDC に分類保管されている DNA 登録を登録の背景にある研究（プロジェクト）単位にグループ化し、研究対象や研究目的別に細分類した。これにより、生物種ごとに、さらにその研究対象ごとに、登録レコード数が多い代表的なプロジェクトを把握し、その配列データをまとめてダウンロードすることが可能になった。さらに、タイトルを日本語化することにより、プロジェクトの内容が一目分かりするようになった(図 2.2.1 参照)。

２）遺伝子発現バンク目次

NCBI が提供する GEO はマイクロアレイや SAGE などの遺伝子発現に関する実験結果を集積した、遺伝子、サンプル、値の三つ組みデータを対象としたデータバンクの代表である。進展の早い実験領域の１次データバンクは得てして利用者には難解である。そこで、少しでもデータが利用しやすくなるように、データの整理パイプラインを作成し、DNA

図 2.2.1 DNA バンク目次の表示例

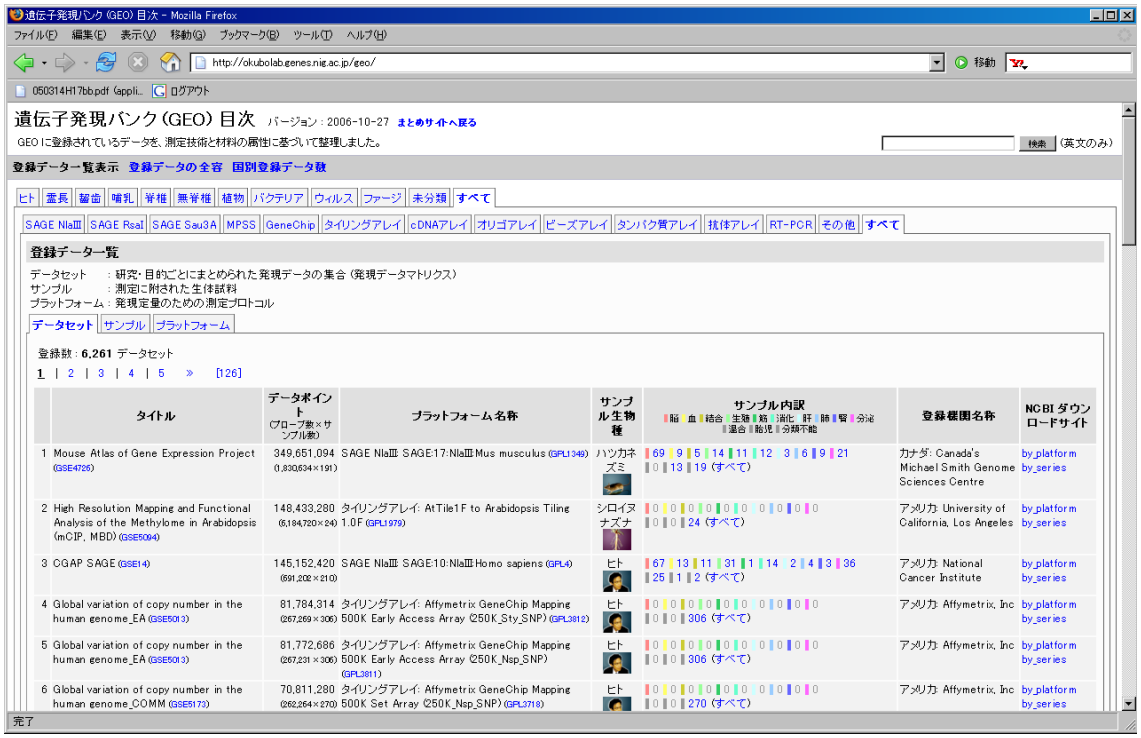


図 2.2.2 遺伝子発現バンク目次の表示例

バンク同様に研究対象や研究手法に基づく目次を作成した。これにより、どんな生物のどんな実験データが登録されて利用可能なのかが、閲覧可能になった(図 2.2.2 参照)。

（3）ライフサイエンス分野の研究の俯瞰調査

国内の研究を俯瞰するための情報源として各種学会の過去の抄録を統合し、検索や、施設別やテーマ別の再編成が可能なDB化を目的として、今年度は分子生物学会８年分の書誌事項に加え一部要旨を電子化し、施設名称など基本的な用語の統一を行い、要旨の検索、ソート、可視化が可能なシステムを開発した。本システムを用いて、要旨中に含まれるキーワードによる要旨の分類やその年次推移を観察することが可能になり、今後の分野の動向を逐次的に調査把握していくことが可能になった。また、同じ目的で日本語総説誌バックナンバー全文電子化作業の一環として、蛋白質核酸酵素のバックナンバー（約10年分）の電子化を行った。以上の検討により、日本語文献のオープン化、データベース化の重要性を認識できた。

さらに､ライフサイエンスの知識を俯瞰するためのデータや知識の整理法を開発する目的で、動物の脳に関する機能的、形態的、分子的、進化的なデータ・知識を集め、教科書的な知識と最新の知見をおりまぜて伝えるシステムの構築を行った。1,000個の脳細胞に対する40個の遺伝子情報のデータベース化と、細胞単位で発現を視覚的に把握することができるビューワー機能の構築により、脳細胞と遺伝子情報の相関性の3次元的な表現が可能になった。これにより、DB統合化の際の表示機能の重要性を把握できた。

（4）検索アルゴリズムを含めた知識情報技術の動向調査

生物情報を扱うデータベースは、いろいろな分野の異なる観点から作成され、また、それぞれ異なる形式で記述されている。また生物情報データベースに含まれるデータ量は計測技術の発展に伴い膨大な量となってきている。また、High Wire Press や PubMed Central をはじめとした文献の電子化・オープン化が進むことで、生物情報として利用可能なテキストや図表の量も増えつつある。さらに、これら生物情報の利用方法自体、ユーザによって様々である。このような背景をもつ生物情報のデータベースを統合するためには、高度な知識情報技術の利用が不可欠である。そこで、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。併せて、統合データベースに必要とされる計算機資源（CPU 数、ディスク容量など）を、知識情報技術の利用の観点から予測し、効率的な計算機環境を整えることを目的に、統合データベースに関する計算機資源の調査を行った。

検索システムについては、単純な項目検索やキーワード検索では、ライフサイエンス分野の、情報の膨大さ多様さゆえに対応が難しく、検索エンジン自体に高度な解析機能、可視化技術が必要になってきている。また、対象がグローバルなWWW上のデータへと広が

ったため、ローカルなデータベース内の構造化されたデータのみならず、WWW上の非構造化データをも扱える必要が出てきた。そこで、膨大かつ多様な情報へ対応する検索技術、および構造化データと非構造化データをともに扱う技術について調査した。

データマイニングとは、大規模なデータやデータベースから隠れた関係性や知識などの情報を帰納的に抽出する技術を指す言葉である。データマイニング手法は出力される情報の方向性と入力されるデータの種類から、おおまかに第一世代と第二世代のものに分けることができるが、第二世代のデータマイニング手法には、ベイジアンネットワーク、隠れマルコフモデルなどの確率モデルや、グラフマイニングなどの構造データからのマイニング手法、さらに、テキストマイニングやストリームマイニングなどの新しいタイプのマイニング手法が含まれる。この調査では、第二世代のデータマイニング、特に構造データからのマイニング手法について調査を行った。

Web 2.0 とは従来の WWW における静的なサービスに対し、次世代にあるべき新しいウェブのあり方に関する総称である。Web2.0 の特徴を持つタームとして、ここでは、web service、ロングテール、集合知、タグ付け、ブログについて調査し、さらに、これらとデータベースの関係について考察した。

グリッドコンピューティングは、元々は遊休計算機資源を有効に活用するために作られた仕組みだったが、現在は、大規模計算を効率よく行うための仕組みとして利用されている。このグリッドコンピューティングの目指す環境を実現するための様々な課題、例えば利用する計算機が別組織に属したり、そのプラットホームがばらばらであっても動的に連携できる仕組みの構築、を解決する必要がある。ここでは、こうした課題の解決策について調査した。

計算機資源の調査については、文献、インターネット、及び、公開されているプログラムの計算速度、及びメモリー使用量、ディスク使用量を計算し、また、今後、新たに利用されると思われる技術については、他分野での同技術を用いているプログラム速度を参考として見積もりを行った。Web クローリング、テキストマイニング、フェノタイプ情報の画像処理・画像検索、配列バンクの項目別検索および多型情報を含む配列解析などについて、必要とされる技術の調査を行い、それぞれで必要となるメモリー量、CPU 数、ディスク容量を概算した。

これらの調査によって、従来の検索技術には情報の膨大さと多様さに基づく限界がすでにきており、データマイニング技術を検索エンジンへうまく組み込む必要があることが分かった。また、グリッドコンピューティングを始めとした分散計算技術は、web service を前提としており、必要な web service をデータベース側で揃えていく事が今後より重要となることが分かった。また、知識情報技術を活用するために必要となる計算機資源の見積り根拠を得ることができた。

（５）臨床情報や医療統計の現状調査

臨床情報の調査に関しては、HL7等の標準規格の役割、データ抽出のためのカルテデータの電子化の状況やインセンティブ等につき、インタビューを含め調査を行った。また、生活習慣病を中心とした我が国のコホート研究の事例を分類・整理した。医療統計の調査に関しては、遺伝子多型解析に関わる遺伝統計学に焦点を絞り、遺伝統計学分野で用いられる解析技術に関して、インタビューを含め調査を行った。

我が国のコホート研究については、循環器疾患を対象とした大迫(おおはさま)研究、生活習慣病その他の種々の疾患を対象とした山形大学の地域特性を生かした分子疫学研究(21 世紀 COE プログラム)、地域住民、大都市検診を対象とした多目的コホートによるがん・循環器疾患の疫学研究、広範な疾患を対象とした久山町研究、癌、循環器疾患を対象とした放射線影響研究所コホート研究、虚血性心疾患を対象とした都市勤労者集団コホート、高血圧を対象とした端野・壮瞥町研究、一般住民の循環器疾患を対象とした NIPPON DATA 80、全国各地で行われている循環器コホート研究の個人データを統計的に統合し、リスク因子を定量的評価することを目的とした JALS (Japan Arteriosclerosis Longitudinal Study) について、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。

遺伝統計学分野で用いられる連鎖解析、連鎖不平衡解析(ハプロタイプ解析）、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、入出力、利用形態、ダウンロード先などを調査し、その評価を行った。また、代表的な商用ソフトの 2 種類の機能、特徴などを調査した。

これらの調査から、医学データ活用における課題として、病名の標準化、前向きコホート研究の推進、人類遺伝学基盤の充実が重要との結果を得た。病名の標準化に関しては、現状は死因統計などの主として保険行政統計用のものか保険診療用のものしかなく、臨床研究向けには使いにくく、抜けもあり、また必ずしも真の病名が記載されない、といった問題がある。前向きコホート研究については、既存のカルテの活用（後向き研究）は、検査値などを除くと難しく、しっかりデザインされた一定規模の前向き研究によって初めて有効なデータが得られることが分かった。また、米国では家系を集めるプロジェクトにも多額の投資がなされているのに対して、日本では家系データが軽視される傾向にあり、これは日本における人類遺伝学基盤の不十分さに起因することが分かった。

２．３　データベース統合化戦略の立案および評価のまとめ

我が国のライフサイエンスＤＢは約２５０あると言われるが、統合化の目的はこれらバラバラに管理運営されている多種多様なＤＢを一つにすることであると一般に考えられている。もし、仮にこのような統合化が実現されるとすると、利用者にとっては、検索、仮説生成、解析が容易になり生産性が向上する。ＤＢ管理者にとっても管理運営がトータル

に効率的されるであろう。

しかしながら、上に述べたような「一つの統合 DB」は現実的に実現困難で有用性も低い。その理由は、分子データの統合化だけでは不十分であること、データの解釈や意味が研究の進展によって変化するため、ある時点で統合化を実現してもそれがずっと有意義かどうかは保障されないこと、ライフサイエンスの最先端はテキスト（論文）で表現されるため、それらとの連携が十分でないと最先端の知識がＤＢに反映されないこと、データをどう眺めたいかは研究者によって異なるため一つの視点で統合化を図っても多くの利用者の満足は得られないこと、などが挙げられる。

そこで、統合ＤＢ構築は完結しないプロセスであると認識し、研究の進展に応じて利用者の求めるものが変化することに柔軟に対応したＤＢを構築することが重要である。一つの統合ＤＢ実現は研究開発の生産性向上のためのあくまでも手段であるので、それを自己目的化することなく、いかにして研究開発の生産性向上を目指すかという原点に立ち返って方針を作成すべきである。

研究運営委員会、統合 DB 整備戦略作業部会での議論、及び上記の調査活動の結果に基づき得られた、統合 DB 構築の際の目指すべき課題、取り組みの際の基本的考えを以下に列挙する。

・ＤＢ構築者ではなく利用者の思考や意思決定を支援する DB を構築する
・利用者の興味、知識に応じて必要な情報、判断材料をもれなく提示する
・複数 DB をつなぎ異種データ・知識の関係が俯瞰（仮説生成）できるようにする
・いろいろなツールを簡便に組み合わせて解析（知識発見）できるようにする
・できれば、上記のことが日本語で行えるようにする
・ＤＢ化（構造化）されないもの（ 論文（テキスト、ポンチ絵、画像)、特許、教科書、報告書、解説記事、など）もうまく扱えるようにする
・ＤＢ構築に最先端の研究開発は必要だが、あくまでも利用者の利便性向上のためのサービス事業であることを認識する

以上の基本的な考えをベースに得られた、文部科学省が目指すべき統合データベースの概要は以下のとおりである。

まず、想定ユーザとしては、

・ライフサイエンスの研究者および医療、創薬などバイオ産業従事者
・ライフサイエンスプロジェクトの企画立案や評価に関わる人々
・ライフサイエンスデータベースの構築者

とする。

統合化の対象データとしては、ヒト・動植物・微生物の分子データ、文献データ、臨床などの表現型データを考える。

開発すべき提供機能としては、

・DB やツールの所在や利用法を網羅したポータルサイト
・分子データと文献知識（高次生命機能）を統合したデータベース
・分野の俯瞰や仮説生成が容易に行える検索機能
・DB 構築者へのインデックス、辞書、整理棚、書式、DB 構築ツールの提供
・データベース構築者（キュレータ・アノテータなど）の学習用教材
・上記の日本語による検索と表示

が必要との結論を得た。

なお、数年後に上記の統合データベースを実現させるための計画に関しては、図 2.2.3 に示すようなステップをとることが適当と考える。また、年次ごとに想定される具体的実施事項および成果を図 2.2.4 に示す。

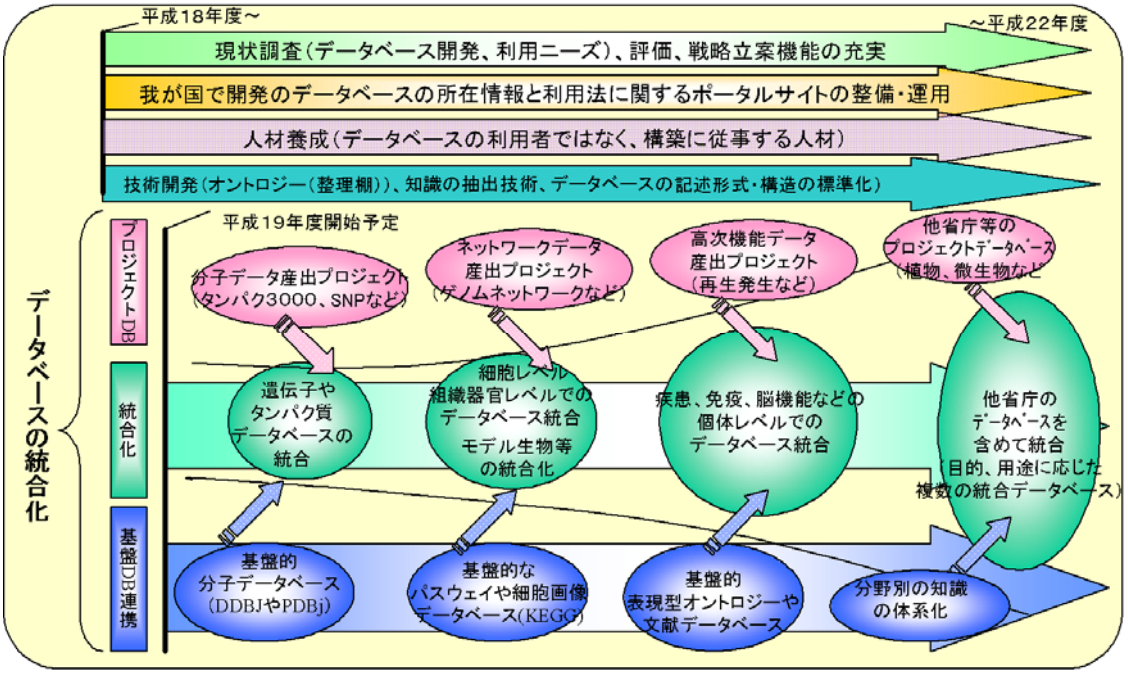


図 2.2.3 統合データベース事業展開の年次計画

図 2.2.4 年次ごとに想定される具体的実施事項および成果

さらに、上記のあるべき姿の統合データベースを具現化する具体的な取り組みとして、以下の提言を行う。

（1）戦略立案・実行評価

統合 DB 構築は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に、より高度な研究開発ができる環境を提供することである。このような環境の構築には一般に長い時間を要することから、また、将来の研究の進展や変化を見越して計画を立てる必要があることから、長期的な視点に立って戦略を立案し、それにそって DB 開発を進めることが不可欠である。そのため、これまでのようにライフサイエンス DB 構築の専門家だけに頼って戦略を立てるのでは不十分である。そこで、引き続きライフサイエンス（基礎生物学および医・薬・農の応用生物学)、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組織を構築し、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必要がある。また、この組織では、その戦略が着実に実行されているか、社会のニーズに適合しているか、などを定常的に評価し、必要に応じて戦略を機動的に変更する必要がある。また上記組織に対し、3 分野の動向に関する情報の網羅的収集・分析を日常的に行い、それら 3 分野の最新の俯瞰マップや動向マップを提示するなどして、戦略立案や実行評価を支援するチームも必要と考える。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、出版社・学会、国

内外の研究機関、等各種利害関係者との連絡調整、以下の各項目の統括、等の活動もあわせて行う必要がある。

統合 DB に関しては、本プロジェクトにより、そのあるべき姿のイメージ作りとそれを実現するための要素技術開発の立上げができたものと考える。しかしながら、統合 DB 構築は統合に必要な技術の開発だけで実現できるものではなく、著作権その他のデータベースにまつわる法律やデータの公開を遅らせる様々な要因といった、検討、解決すべき種々の社会的、制度的課題が存在することも分かった。それぞれが著作権やそのデータ産生の背景に存在する種々の権利関係を有する個別のデータベースを統合して公開する際には、著作権その他の法律をよく把握して、これを十分に遵守する必要がある。また、医療関係のデータベースの公開にあたっては倫理面への配慮が不可避である。一方、国家プロジェクトで産生されたデータに基づくデータベースに関しても、一般への公開が優先か、特許取得や産業振興のための一定期間のデータの非公開を優先するかといった問題が存在する。また、自分で産生したデータを大事にしたいという研究者が一般的にもつ気質もデータの公開を遅らせる大きな要因である。さらに、本年度の取組みでその重要性を認識することができた日本語文献情報に関わる、学会要旨や日本語総説誌のオープン化、データベース化に関しても、個別の学会や出版社との間で相互にメリットがある形の提携を行うための交渉が必要であり、そのためのビジネスモデル作りや著作権などに関する検討も重要な課題である。

以上の課題を解決し、真に役に立つ統合 DB を構築するためには、以下の取り組みが必要と考える。

- ニーズ面、シーズ面から見て今後重要となるデータベースコンテンツの調査
  解析技術の開発動向なども含めたデータ産出側の将来動向、併せてライフサイエンスおよびバイオ産業で今後必要となるデータニーズを調査することにより、ユーザニーズに合致した統合データベースのコンテンツの把握が必要。
- 統合対象の 250 の国内データベースの詳細調査と統合の優先付け
  プロジェクト DB の現状調査に基づく受入れ方法の検討および受入れの優先順位をつける方法を立案がすること必要であり、これに基づいた受入れと運用が望まれる。一方、レポジトリ制度の調査に基づいたレポジトリシステムの構築も重要。
- 国内基幹データベース、大量データ産出機関との連携策の検討
  データ量もユーザニーズも大きな基幹データベースや大量データ産出機関のデータベースとの連携策を検討し、統合データベースの提供する機能に合わせたデータの整形、リンク付けを実現。
- ライフサイエンスプロジェクトの現状調査と将来の統合化の阻害要因の洗い出しと解決策の策定
  各省の代表的なプロジェクトやそれを取り巻くコンソーシアムで産出されるデータの契約・権利関係、データ公開ルールを整理した上で、各プロジェクトや機関も納

得する形でのデータ提供を受けるための枠組みを検討。必要に応じてデータ公開を促進する法律立案も検討要。

・ データに関わる知的財産権や倫理的側面に関する調査と対応策検討

データベースレコード、文献など統合化に際して検討すべき知的財産権の調査と権利保有者との提携方針を策定。また、医療に関わるデータの個人情報の取り扱いなど、データベース公開に関わる倫理面の調査を行い、公開範囲に関する指針を策定。

（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すために、統合 DB とそれに必要な情報技術を開発する必要がある。具体的には、以下の取り組みが必要である。

1）共通基盤技術開発

・ インデックス、専門用語辞書の自動構築技術の開発
・ テキスト情報、画像・ポンチ絵情報ならびに異種 DB からの知識発見技術
・ 情報共有、情報交換のための WEB 技術開発
・ ワークフロー技術等との連携

2）ヒト統合 DB の開発・運用

・ 文献からの知識抽出システム開発とそれによるヒト知識の整理
・ 細胞、組織、器官、個体などの高次レベルの整理棚構築
・ 高精度アノテーションの実施と知識、機能を中心としたヒト統合 DB 構築、運用
・ 医療、医薬品に関するデータとの連携

3）モデル生物・産業応用生物統合 DB の開発・運用

・ 専門家集団からの意見集約による対象生物群の遺伝子名、機能、産物などの辞書構築
・ 環境ゲノム・メタゲノムを含む微生物ゲノム比較解析用 DB の構築と技術開発
・ 高精度アノテーションによるモデル生物・産業応用生物統合 DB の構築、運用
・ 解析ソフト充実と操作性の良いブラウザー開発による統合 DB の高度利用実現
・

（3）統合データベース支援

統合 DB の開発と運用に際しては、個々の DB としてどのようなものが現在開発・公開されているか、どのような DB 解析ツールが利用可能か、などを網羅的に調査し、その情報を一般の利用に供すること、我が国で開発された種々の DB を受け入れ、相互に連携して使えるようにすること、統合 DB の開発・運用やその利用技術開発に従事する人材を育成すること、などの支援業務が欠かせない。これを実現するために、以下の取り組みが必要である。

1）ポータル整備・運用、広報、普及啓発

・ DB サービス、解析サービスサイトに関する網羅的ポータルサイトの整備と運用

・ 検索に必要なインデックスや用語の収集とポータルサイト自動更新技術開発
・ ポータルサイト構築のための専門家ならびに利用者の意見集約システムの開発、運用
・ 日本語情報の収集、整理と日本語による研究情報の流通を促進する仕組みの整備
・ 海外情報日本語化のための技術開発と我が国の研究活動の海外への情報提供
・ ホームページの構築、講習会、シンポジウムの開催、ニュースレター等の発行

2）データベースの受け入れと運用

・ プロジェクト DB の受け入れと相互運用可能 DB への変換と運用、公開
・ 相互運用可能にするための標準化技術と用語の整理
・ 統合 DB 構築のための国内外主要 DB の更新、維持、管理

3）人材育成

・ 実データを用いたキュレータ・アノテータ教育の実践
・ 学部教育と連携した DB 構築者養成カリキュラムの実践
・ 大学院教育と連携した DB 高度利用者の養成と体系的なカリキュラムの作成
・ 教育プログラム・教材の実践・評価

３．データベース統合化基盤技術開発

３．１　データベース統合化基盤技術開発の実施計画

利用者が分子データや文献データを区別なくわかりやすい案内にもとづいて検索利用できる統合化を実現するための基盤技術の開発利用にむけ以下の４項目について開発に着手し利用を試みる。

（１）基盤知識表現技術開発

情報検索の利便性は索引付けの質に大きく依存するが解剖名称や細胞名称などの分野共通の基盤的な概念(用語)に関しても概念(用語)の整理と索引への利用は進んでいない。ここでは応用分野の区別無く索引利用可能な蛋白質名称、動物・植物の解剖用語、細胞名称および実験手法を対象にデータの内容を豊かに表現する索引系を開発しポータルサイト構築や戦略立案に提供する。

（２）癌研究知識表現技術開発

欧米では癌組織のヒトゲノム再配列解析プロジェクトが進行しており、そのサテライトとして統合癌 DB 化が検討されている。一方わが国では、臨床情報については、未だ DB 化に伴う倫理問題が議論されている段階である。そこで本プロジェクトでは大規模再配列解析プロジェクトの成果を念頭に置いて、すべての体細胞レベルでの遺伝情報と臨床情報を統合化した DB の作成において必要な表現法やデータ整理法の開発利用を試みる。

（３）多型知識表現技術開発

これまで多型情報の DB 化として、日本人ゲノムのプロモーター領域に存在する SNP を正確な頻度情報とともに記載したデータベース「dbQSNP」および日本人確定ハプロタイプ情報を記載した「D-HaploDB」の構築に取り組んできた。これらをさらに拡充するとともに、これを統合し日本人試料を用いた関連解析による疾患原因遺伝子の探索に不可欠な、高精度の情報基盤を確立することを目指す。また、XML 化等の標準化にも取り組む。これにより、ヒトゲノム多型データベースの標準化、統合データベースへの組み込みのフィジビリティースタディーを行い、多型情報統合 DB の医学分野におけるあり方を検討する。

（４）キュレーター支援技術開発

多数の文献に書かれた知識をまとめて表やサマリーの形で統合利用可能にする作業は、主に DB キュレーターと呼ばれる研究支援者によって手作業で行われている。論文などを読み解いて関心領域の抽出や記録を行う一連のキュレーターの作業を支援する技術を開発し支援環境の構築を行う。

なお、本テーマの（３）多型知識表現技術開発は九州大学が実施した。それ以外の項目は情報・システム研究機構が実施した。

３．２　データベース統合化基盤技術開発の実施内容

（１）基盤知識表現技術の開発

データベース構築の基盤となる知識表現技術開発の一環として、A．辞書シソーラス、及びB．オントロジー、分類機の開発を行った。さらに、開発したこれらの技術を用いたデータベース統合化の試みとして、C．分子データベース整理統合を行った。

A．辞書シソーラス

ライフサイエンスに関わる種々のカテゴリーのデータを統合し、情報を整理するときに必要となる辞書や、階層のついた同義語辞書であるシソーラスの作成を検討した。本年度は、１）遺伝子名称シソーラス、２）生物学名日本語一般名対応辞書、及び３）施設名称辞書の作成を行った。

1.）遺伝子名称シソーラス

検索やデータベース統合の混乱の主因の一つは、ライフサイエンス分野における同義語、特に遺伝子名称の同義語の氾濫である。そのため、遺伝子名称の同義語辞書を作成し、これを使用することで、種々のデータベースを統合的に利用する際の混乱を防止する必要がある。そこで、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、ここでは、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、遺伝子が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 Ver1.0 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。表 3.2.1 に、今回開発したシソーラスのファイルの構成を示す。また、表 3.2.2 に今回対象とした生物種、およびその遺伝子数ならびに遺伝子名称の数を示した。

表 3.2.1 遺伝子名称シソーラスのファイル構成

SWISS-PROTのID	EntrezGeneのID	その他DBのID	遺伝子名称				
SWISS-PROT:Q9Y6Y9	EntrezGene:23643	HGNC:17156	MD-2 protein	Lymphocyte antige	MD-2	MD2	ESOP-1
SWISS-PROT:Q9Y6Y8	EntrezGene:11196	HGNC:17018	p125	P125	SEC23-interacting	MSTP053	SEC23IP
SWISS-PROT:Q9Y6Y1	EntrezGene:23261	HGNC:18806	Calmodulin-binding	KIAA0833	calmodulin binding	CAMTA1	
SWISS-PROT:Q9Y6X9	EntrezGene:22880	HGNC:23573	MORC family CW-	Zinc finger CW-ty	KIAA0652	MORC family CW-	AC004542.C22.1
SWISS-PROT:Q9Y6X8	EntrezGene:22882	HGNC:18513	KIAA0854	Alpha-fetoprotein	ZHX2	zinc fingers and ho	AFR1
SWISS-PROT:Q9Y6X2	EntrezGene:10401	HGNC:16861	PIAS3	Protein inhibitor of	ZMIZ5	protein inhibitor of	FLJ14651
SWISS-PROT:Q9Y6X0	EntrezGene:26040	HGNC:15573	SETBP1	SET-binding prote	KIAA0437	SET binding protei	SEB
SWISS-PROT:Q9Y6W8	EntrezGene:29851	HGNC:5351	ICOS	CD278	inducible T-cell co	MGC39850	AILIM
SWISS-PROT:Q9Y6W6	EntrezGene:11221	HGNC:3065	MKP-5	Mitogen-activated	MKP5	DUSP10	MAP kinase phosp
SWISS-PROT:Q9Y6W5	EntrezGene:10163	HGNC:12733	WASP2	WASP-family prot	WAS protein famil	Wiskott-Aldrich sy	WAVE2
SWISS-PROT:Q9Y6W3	EntrezGene:23473	HGNC:1484	calpain 7	CAPN7	CALPAIN7	PalB homolog	Calpain-7

表 3.2.2 対象生物種と遺伝子数及び名称数

生物種等	遺伝子数	名称数
遺伝子ファミリ	12,110	27,923
ヒト	38,728	173,630
マウス	60,688	172,260
ラット	38,164	123,726
ゼブラフィッシュ	38,879	83,694
ショウジョウバエ	30,410	95,578
線虫	25,316	97,031
出芽酵母	6,190	33,030
分裂酵母	4,895	9,790
枯草菌	4,106	18,920
合計	259,486	835,582

2）生物学名日本語一般名対応辞書

データベースに記載されている生物の名称は、正確を期してラテン学名表記を基本としているが、利用者のほとんどは学名になじみがないのが実情である。そこで、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)の登録エントリー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。登録データ数は合計 14028 種となっている。主要な 73 種類についてはさらに認識を容易にするアイコン画像を作成した。なお、ここでは農業環境技術研究所の日本野生植物寄生・共生菌類目録および日本産糸状菌類図鑑、日本爬虫両棲類学会の爬虫類のリスト、厚生労働省検疫所の届出対象動物種名リスト、哺乳類頭蓋の画像データベース（第 2 版）や The International Seed Federation (ISF)の植物病害関連生物リストなどを参照し、開発者が最も適切と思われるものを和名として採用した。その他、和名が不明な生物種については論文などから補完した。表 3.2.3 に分類ごとの登録した生物種数を示す。また、表 3.2.4 に辞書の一例を示す。図 3.2.1 には、アイコン画像の例を示す。

表 3.2.3 分類ごとの生物種数

分類	生物種数
霊長類	137
齧歯類	823
その他哺乳類	742
その他脊椎動物	4977
無脊椎動物	647
植物・真菌類など	6578
バクテリア	118
ウイルス	5
バクテリオファージ	0
未分類	1
合計	14028

表 3.2.4 生物学名日本語一般名対応辞書の一例

生物学名	日本語一般名称
Lemur catta	ワオキツネザル
Lepilemur mustelinus	イタチキツネザル
Varecia variegata	エリマキキツネザル
Cynocephalus variegatus	マレーヒヨケザル
Cheirogaleus medius	コビトキツネザル科
Otolemur crassicaudatus	オオガラゴ
Galago senegalensis	ショウガラゴ
Loris tardigradus	ホソロリス
Nycticebus coucang	スローロリス
Perodicticus potto	ポットー
Tarsius syrichta	フィリピンメガネザル

図 3.2.1 アイコン画像の例

3）施設名称辞書

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。これにより、同一の研究室の同一テーマを一塊として把握し国内の研究動向の把握を容易にすることが可能になった。表3.2.5に施設名称辞書の構成の一部を示す。

表 3.2.5 施設名称辞書の構成

標準化名称	標準化英語名称	科研費の機関コー	エリアス(エリアスが無いものは標準化名称を記載)
東大	Univ. Tokyo	172	東京大|東京大学
東大・医	Univ. Tokyo,Fac. Med.	172	東大・医学部
東大・医・一外	Univ. Tokyo,Fac. Med.,1st Dept. Surg.	172	東大・医・一外
東大・医・三内	Univ. Tokyo,Fac. Med.,Third Dept. Int. Med.	172	東大・医・3内|東大・医・第3内科
東大・医・整外	Univ. Tokyo,Fac. Med.,Dept. Orthop. Surg.	172	東大・医・整形外科|東大・整形

B．オントロジー、分類機

オントロジー、分類機として、１）動植物解剖学自動分類タガー、２）都市名国名自動検出タガー、３）解剖学3Dポリゴンマン辞書、４）3DアナトモグラフィーAPI、及び５）メソッドオントロジーとの連携システムを開発した。

１）動植物解剖学自動分類タガー

解剖学用語、すなわち臓器・器官・部位の名称を、専門家が作成したルール（振興調整費DB統合のための調査研究において作成）を用いて自動的にカテゴリー分類するプログラム、動植物解剖学自動分類タガーを開発した。

動物解剖学分類タガーでは、表3.2.6に示すように、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、上記のカテゴリーに分類する。

植物解剖学分類タガーは、表3.2.7に示すように、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。

表3.2.6動物解剖学分類タガーにおける解剖学用語の分類

大分類	小分類								
脳	大脳	小脳	脳幹	脳梁	松果体	末梢神経	脊柱	網膜	目
血	動脈	静脈	リンパ節	末梢血	脾臓	胸腺	骨髄		
結合	脂肪	骨	皮膚						
生殖	胎盤	子宮	前立腺	卵巣	精巣				
筋	心臓	骨格筋							
消化	食道	胃	腸	結腸					
肝	肝臓								
肺	肺								
腎	膀胱	腎臓							
分泌	下垂体	甲状腺	副腎	膵臓	乳腺	唾液腺			

表3.2.7 植物解剖学分類タガーにおける解剖学用語の分類

大分類	小分類		
地上構造	葉	茎	
若い地上構造	若い地上構造		
根	根		
成長点	成長点	カルス	
花・生殖	花粉	子房	花・生殖
種子・果実	胚	種子・果実	

２）都市名国名自動検出タガー

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、都市名国名自動検出タガーを開発した。ここでは、DNAデータバンク(INSDC)のDBレコードの国別分類を行う自動検出タガーのフローを説明する。まず、DBレコード(FlatFile)を構造分解して、国名分類に使用するフレーズを抽出し、抽出した文字列を国名辞書、国名シノニム辞書、国別コードトップレベルドメイン１(ccTLD)辞書、国別研究機関名辞書からなる分類辞書群で順番に検索し、国名分類を行う。

３）解剖学３Ｄポリゴンマン辞書

解剖学用語、すなわち臓器・器官・部位の名称やそれらをくくる概念をモデル人間中の３次元座標（３Ｄポリゴンマン）で定義した辞書である解剖学３Ｄポリゴンマン辞書を開発した。これは、科学技術振興調整費「生命科学データベース統合に関する調査研究」におけるボクセル人体モデルの検討結果を受けたものである。従来のツリー型表現（いわゆる解剖オントロジー）と違い、多角的に破綻しない表現が可能で、ボクセルデータに比べエディットが飛躍的に容易である。ポリゴンマンの各臓器・器官の空間座標は、数値人体モデルデータベース（独立行政法人情報通信研究機構が開発）を基盤に、人体解剖模型・図譜等を参考に詳細化を行った。PubMedアブストラクトで出現の多い用語を中心に定義を行い、約 130 語の定義が完了した。

ポリゴンマンの各臓器・器官の空間座標は、数値人体モデルデータベース（独立行政法人情報通信研究機構の長岡博士らが、北里大学、慶應義塾大学及び東京都立大学と共同開発した電磁波影響計算のための Voxel モデルファントム、分類組織数約 50)を基盤に、位置関係や形態を大きく損なうことなく人体解剖模型・図譜等を参考に詳細部分を書き加えることにより求めた。肉付きや顔貌は創作によるものであり、定量的な正確さはある程度犠牲にしているが、形態や相互の関係は概念的な正確さを期して構築した。今後、用語数を増やし詳細化を進める計画である。また、利用環境が限定されるデータということもあ

り、ダウンロードの形ではなく利用者が自前のデータを貼り付けたり、入力したり、共有したりできるようにアナトモグラフィー（次項）も開発中である。

４）３Ｄアナトモグラフィー ＡＰＩ

解剖学用語が付与された手持ちのデータ（例：器官別の発現解析データ、疾患別症状分布など）を解剖学3Dポリゴンマン辞書にマッピングして俯瞰可能なアナトモグラフィー（新造語）を開発した。マッピング結果は、静止画像もしくは動画で得ることができ、webサービスとして提供予定である。膨大かつ詳細な”体に関する情報”同士の関係を理解する際に有用であると期待される。図3.2.2に３Ｄアナトモグラフィーの出力結果の例を示す。

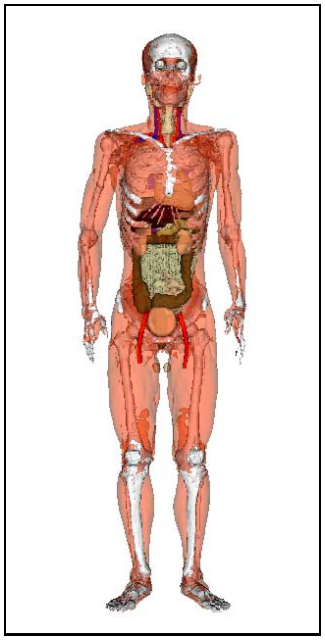


図 3.2.2 ３Ｄアナトモグラフィーの出力結果例

５）メソッドオントロジーとの連携システム

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

今年度は、ドライ系メソッドオントロジーを利用した検索プログラムを作成し、Web リソースポータルに実装することにより、様々な名称がつけられているバイオインフォマテ

ィクスのメソッドを利用者の観点から分類し検索可能にした。表 3.2.8 に、構築したウェット系の実験手法名辞書の一部と図 3.2.3 に Web リソースポータルサイトの検索システムの一部を示す。

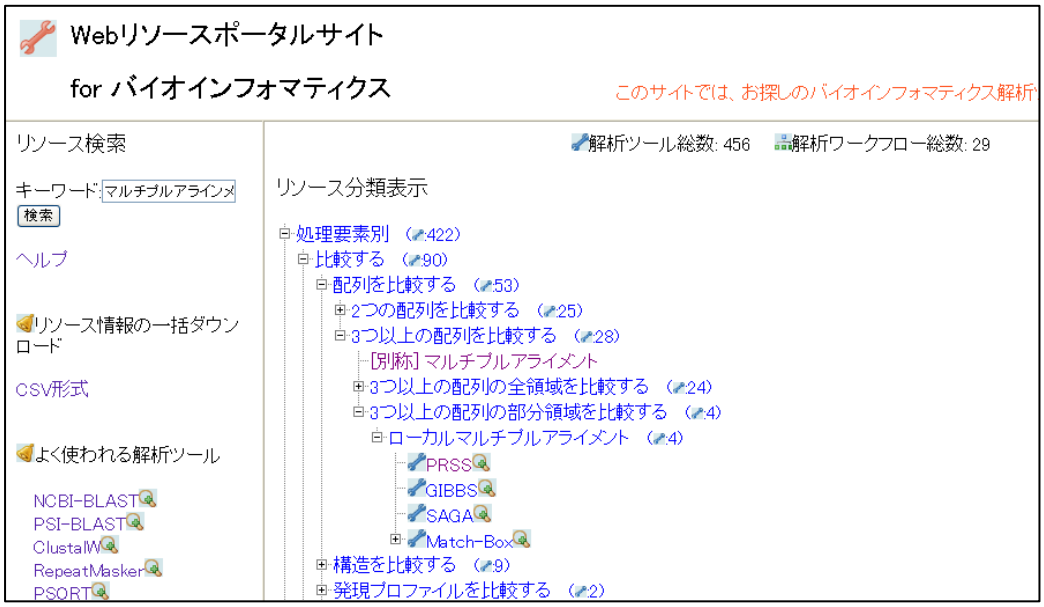


図 3.2.3 ドライ系メソッドオントロジーを利用した検索プログラム

C．分子データベース整理統合

本プロジェクトおよび関連プロジェクトで開発した統合技術およびリソースを用いて未整理のデータを統合した統合データベースとして、１）ヒト遺伝子発現統合の開発を行った。情報が安定した分野についてはこれからもこのような物理統合を進めてゆく考えである。

１）ヒト遺伝子発現統合

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。また、遺伝子の発現パターンを、3種類の生物学的な分類（似た発現パターン、染色体上での隣接、同じ遺伝子ファミリー）に応じて表示することを可能にした。組織情報は、開発した動植物解剖学自動分類タガーで処理し整理分類しており、これにより異なるプロジェクトから得られた発現データ間の比較が可能になった。また、開発した遺伝子名称シソーラスと3Dアナトモグラフィーを、検索部分と表示部分にそれぞれ用いた。

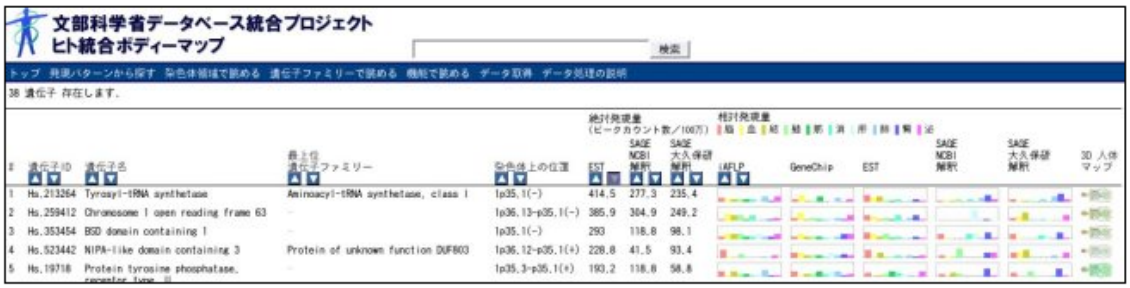


図 3.2.4 ヒト遺伝子発現統合の表示画面の例

（2）癌研究知識表現技術開発

実験的なデータベース作成等を通じて癌の分子データと臨床情報の統合、表現を想定ユーザーにわかりやすい形で実現することを目的として、癌遺伝子発現臨床情報データベースの機能拡張、及び大阪府立成人病センター乳腺内分泌外科の症例を対象に臨床情報の整理を行った。前者については、CGED (Cancer Gene Expression Database)の機能拡張を行い、従来の機能に加えて臨床情報から遺伝子を検索する機能を追加した。図3.2.5に「転移のある癌とない癌で発現の異なる遺伝子の検索」を例とした追加機能を表示する画面の例を示す。また、新規データとして乳癌(抗癌剤耐性研究)、胃癌、甲状腺癌に関するデータのアップロードを行った。臨床情報の整理については、成人病センター内の各種固形癌の臨床情報収集をターゲットとして想定し、乳癌については終了することができた。

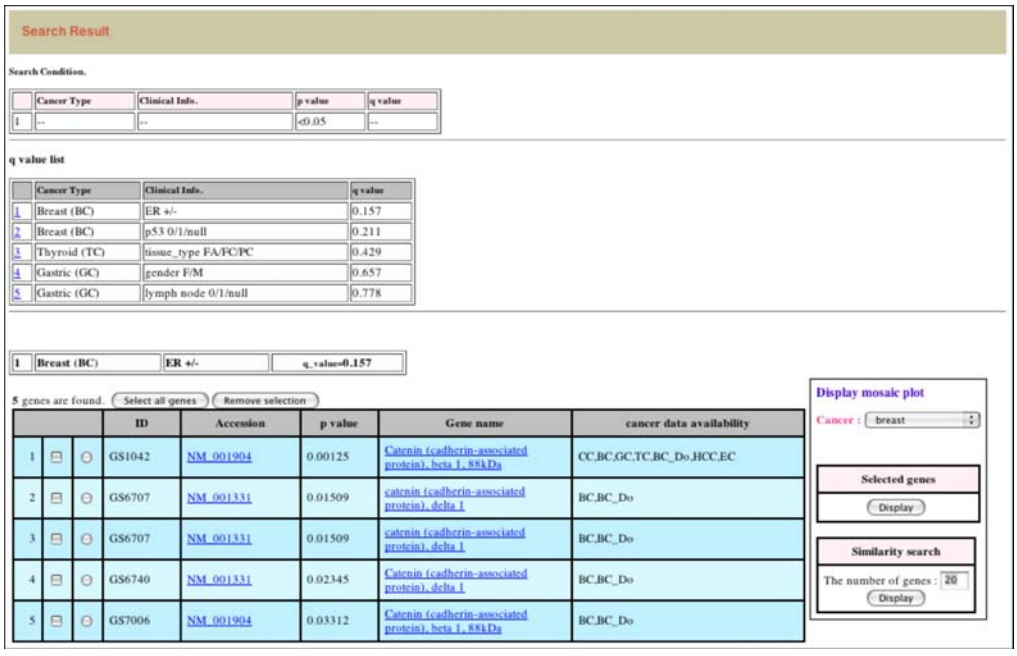


図3.2.5 臨床情報から遺伝子を検索する機能を示す表示例

（３）多型知識表現技術開発

ヒトゲノム多型情報は疾患・体質等の遺伝的背景を解明するための必須の情報基盤である。特にわが国にとって日本人のゲノム多型情報の整備は緊急の課題である。現在最も有効とされる情報源は国際 HapMap 計画での日本人のゲノム多型（SNP）情報とされているが、未だに調べられている SNP のゲノム上の密度が不十分である。また上記の遺伝的背景を明らかにするのに必要なハプロタイプ（SNP の並び方）の情報が必ずしも正確ではない。さらに同計画で調査された個体数が 45 人と少なく、日本人全体のゲノム情報を正確に反映したものとは言い難い状況 にある。そこで新たに日本人ゲノム多型情報に関する一次データを拡大・整備し、医科学等で利用可能な形態として提供することが、ゲノム多型情報を医学 情報と統合するために極めて重要な課題となる。

そこで、日本人ゲノム多型情報を高度化し、医療情報との統合のためのデータポータビリティーを図ることを目的に、疾患の主要な遺伝的要因である遺伝子発現調節領域多型の公開データベース「dbQSNP」の拡充、要因遺伝子探索に必須な全ゲノム確定ハプロタイプ構造の公開データベース「D-HaploDB」の拡充、及び他のデータベースとの多型データ相互利用促進のため上記２つのデータベースの標準言語化を行った。「dbQSNP」の拡充については、自己免疫疾患、がんへの関与が疑われる約 100 個の遺伝子のゲノム領域にある SNP 配列及びその正常日本人でのアレル頻度を直接配列決定及び定量 SSCP 解析により決定した。この結果、約 1.0 X 10^4 個の SNP 情報が記載されることになった。
「D-HaploDB」の拡充については、新たに 100 個の胞状奇胎について Affymetrix 社アレイチップを用いた、既存データの約 2 倍にあたる 5 x 10^5 個の SNP タイピングを行い、ゲノムワイド連鎖不平衡地図を飛躍的に高精度化できた。データベースの標準言語化については、データベース記述標準言語 XML のゲノム多型記述のための機能拡張版である PML を採用し、上記二個のデータベースの PML 版を構築した。これにより、他のデータベースとの多型データ相互利用促進手段を確立できたものと考える。

（４）キュレーター支援技術開発

本事業では、これからの研究に欠かせないのが各分野専門家の手による文献やデータベースからの事実の切り取りと再配置による知識の整理であると考えられる。一方で生命科学の専門家には情報技術を駆使することは容易ではない。そこで、論文から抽出したデータのデータベース構築作業を支援するソフトウェアの開発を目標に、論文情報解析・編集ソフトウェアのベースシステムの開発と論文情報解析・編集用各種解析モジュールの開発を行った。具体的には、論文や Web を渡り歩いて重要箇所だけドラッグドロップすると自動的に URL やページ座標情報が記録され、さらに記事を並べて後からメモ書きを行える環境を FireFox の PlugIn アプリケーション（ScrapParty）として開発した。収集記事は xml 形式で書き出すことができ、他人と共有することも可能である。また、”ScrapParty” の追加モジュールとして、辞書マッピング、ナビゲーション、論文構成認識、引用論文情

報収集を行う基本モジュールを作成した。

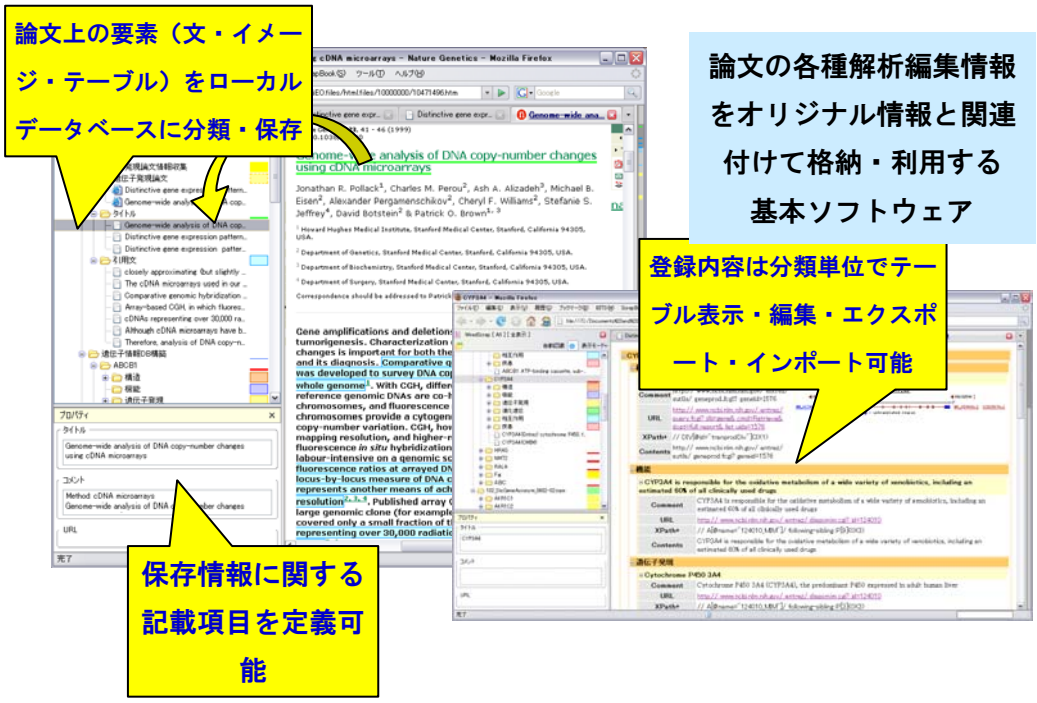


図 3.2.6　ScrapParty の概要

３．３　データベース統合化基盤技術開発のまとめ

① データベースレコードの統合（基盤知識表現技術開発）

検索だけでしか内容をうかがい知ることが出来ない膨大なレコードが与えられたとき、
「要素についてわかるとは全貌のなかでの位置を知ることである」、
「分類整理はＤＢレコード全貌についての最善の表現である」、
「分類整理はＤＢへの入り口であり検索結果の出口を提供する」、及び
「同様な分類整理は統合の一形態である」
という思想に基づいて開発を行った。

整理を行う「整理棚」として、配列や構造とは独立の、
１）研究場所、研究分野、
２）材料の生物種、解剖、
３）研究目的方法
を用いて、それぞれ辞書または分類機を開発し、実際に提供されたサービスに対して開発技術を適用した。

開発技術とサービスの対応は下の表のような関係である。今後も幾つかの対応が追加可能である。

	施設		研究		材料			表示		
1次データ	施設名称辞書	国名辞書	プロジェクトソータ	実験オントロジ	生物アイコン	生物名称辞書	解剖ソータ	ポリゴンマン辞書	アナトモグラフィ:ＡＰＩ (GoogleMap:API)	サービス名称
学会抄録	■									学会俯瞰
INSDC		■	■			■			■	INSDC 目次
GEO		■				■	■	■		GEO 目次
IAFLP		■				■	■	■		発現統合
EST		■				■	■	■		発現統合
SAGE		■				■	■	■		発現統合

（黒ますは今年度適用試験済み）

例えば、様式や表記が同一でない別年度の分子生物学会の年会抄録を材料に、施設名称辞書を用いて研究ループを同定し人名同定を進めることで、複数年会の要旨が統合されたわけである。複数年の要旨が統合されると、年会のメニューでしかなかった情報が研究室単位の研究内容や年次研究変遷の歴史の情報を持ち、試験利用者からも「研究室ＤＢ」としての価値が指摘されている。また同様な整理棚は学会によらず適用可能であるために、日本の研究全体を表現するＤＢを作ることも可能である。この例では一つの辞書によりＤＢは統合可能で、統合によって新たな価値が生まれることが多いことが示されている。

今後も上記の枡目を全て埋めるような対応をすることで、学会要旨から発現データまでが統合利用され統合表示が可能になると考えられる。

② 複雑応用領域の知識表現（癌研究知識表現技術開発、及び多型知識表現技術開発）

臨床医学では人の共通性ではなく個人の違いが研究される。種内の共通性を研究する生物学とは世界モデルが違う為に、臨床医学領域については領域内で利用可能な整理棚も自明ではないため、領域の専門ＤＢに協力を依頼した。

癌医学については、材料を分類するための癌の臨床分類が整理法としてある程度有効であること判ったが、癌遺伝子遺発現研究は患者背景や治療など研究ごとに多様な特徴を発現と比較する為のコントロール疫学研究であり、データを統合することには殆ど意義がないと思われる。むしろ比較したり和をとったりすることが出来る同種のコントロール疫学研究を集めるようなボトムアップ型の統合が有効であるとの結論に至った。今後は、論文を材料として治療効果やマーカー同定などを含めて、コントロール研究のクラスター化を

行うべきであると結論した。

多型研究知識については、ゲノムワイドな多型同定データ解析時に欠かすことの出来ない日本人のハプロタイプに関するモデルが不在であるために、まずはハプロタイプ情報を集積し信頼できる日本人のモデルつくりを第一歩とするべきであると結論付けた。

③ キュレーター支援技術開発

統合 DB 構築に際して、対象データベースのレコードに注釈づけをおこなっていく際に、いかに効率よく各種文献やデータベースの情報を収集、整理するかが課題である。今回開発した論文情報解析・編集ソフトウェアとその追加モジュールは、各種文献やデータベースに記載されている情報を自由に切り取り、分類、保存するシステムであり、今後のキュレーション作業の効率化の強力な武器になるものと期待している。

４．ポータルサイトの構築

４．１　ポータルサイト構築の実施計画

（1）データベース（DB）等ポータル構築

現存するDBのカタログは遺伝子や蛋白質等のデータ対象によるおおまかな分類が与えられるのみで利用者が目的にかなったDBを選択することは容易でない。ここでは中核機関の情報･システム研究機構が中心となってまとめる DB の俯瞰や戦略立案の目的で作成するDBディレクトリに、科学技術振興機構が維持してきたWINGDB案内の日本語解説を加え、利用者を最適なDBに案内する仕組みを構築する。同時に、DB中のデータを利用するために必要となる解析のためのツールや環境を案内するWEBリソースカタログ作りを増強する。利用者にとってわかりやすいインターフェース作りに配慮し、これらのポータルサイトを構築する。また、我が国に存在するライフサイエンスのいくつかのそれぞれ内容が異なるポータルサイトの機能を生かし、相互に利用すべき部分は利用し、全体としてより高い機能を果たすべく連携する仕組みを考案する。また、本課題を広く周知するためのウェブサイト公開用サーバを準備する。

なお、本テーマは科学技術振興機構が実施した。

（2）文献情報との連携調査

科学技術振興機構が開発運営を続けている文献情報提供事業では国内外の情報が日本語で幅広く提供されている。これを活用し、遺伝子機能に関連する情報について、文献情報との連携を検討調査する。

４．２　ポータルサイト構築の実施内容

（1）データベース（DB）等ポータル構築

中核機関の情報･システム研究機構が中心となってまとめる DB の俯瞰や戦略立案の目的で作成するDBディレクトリに、科学技術振興機構が維持してきたWINGDB案内の日本語解説を加え、利用者を最適なDBに案内する仕組みを構築した。同時に、WEBリソースカタログ作りを増強し、これらのポータルサイトを構築した。また、我が国に存在するライフサイエンスの異なるポータルサイトを調査し、連携について考察した。また、本課題を広く周知するためのウェブサイト公開用サーバを用意し､ウェブサイトを公開した。

１）ＤＢポータル

ライフサイエンス分野のデータベースのカタログサイトで、利用者を最適なDBに案内することを目指し、利用者からの書き込み追加が可能な「Wiki」を利用した。コンテンツの充実と即時性が見込めて、利用方法などの情報交換の場を提供できる仕組みである。公開時、詳細記事は371データベース分が収録されている。関係４省庁の協力による過去の国内調査資料にNucleic Acid Research誌Database Issue2006, Science誌Netwatch/Database等を加えリストアップした。また、データベース一覧（構築型分類）

は、科学技術連携施策群での調査研究の成果を提供した。

図 4.2.1　DB ポータル（WINGpro）のデータベース記事の例

2）WEB リソースポータル

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 WEB リソースカタログサイトを構築した。 本統合技術開発の成果であるベンチメソッドオントロジーの一部をカタログの上に乗せてプロトタイプとして提供した。解析ツール 456 をリストしている。　図 4.2.1 に具体的な検索画面を例示している。

3）ポータルサイト連携のための調査

「Ｊａｂｉｏｎ］日本語バイオポータルサイト、「ライフサイエンスの広場」(文部科学省ライフサイエンスポータルサイト)、「ＵＭＩＮ］大学病院医療情報ネットワーク（一般公開）について調査し、連携の仕組みを考察した。

（2）文献情報との連携調査

独立行政法人科学技術振興機構が実施している文献情報提供事業で提供されている文献情報を活用し、遺伝子情報に遺伝子機能を付加する調査を実施した。100 程度の抄録から遺伝子関連の表現型や疾患名を抽出した。テストサイトを公開した。

４．３　ポータルサイト構築のまとめ

ポータルサイト構築にあたり、多数あると言われるデータベースから、我が国では研究者等の所属機関が把握している DB と海外の DB を加えた 371 について、案内サイトを構築した。今後の方針によるが、収録 DB 数の増加と利用者の意見を反映した内容の充実を図るための仕組みを構築し、利用者を最適な DB に案内し、さらに最適なデータを得ることを可能とするサービスを提供する研究開発を継続すべきである。ポータルサイト連携の仕組みも今後の課題でさらによりよいサービスを検討したい。

５．人材の育成

５．１　人材の育成の実施計画

DB 統合や維持管理のためには①キュレーター（論文の内容を理解して情報抽出整理し定型化した表現に変換して DB 構築を支援する学芸員）、②アノテーター（プログラム処理の結果に総合判断を加え、データに生物学的医学的な解釈を付与する学芸員）、③DB マネージャー（DB について理解しておりデータの参照情報などを自立的に更新できる技術者）の３者が必要である。しかしながらこれらの専門業務の存在は DB 構築運営を行ってきた組織でしか知られておらず、広くこれらの業務内容について衆知し、人材養成のための学習の材料を整備することは将来の DB 統合事業に参加可能な人材の裾野を広げることであり、これらの専門職のキャリアパス作成の第一段階である。ここでは３者の業務内容を区別し、業務に必要な基礎知識や技術について解説した教材の作成に着手し DB の本格的統合化事業に備える。なお、本テーマは情報・システム研究機構が実施した。

５．２　人材の育成の実施内容

キュレーター、アノテーター、DBマネージャーの業務内容を整理し、業務に必要な基礎知識や技術について解説した教材を作成してデータベースの本格的統合化に備える目的で、まずwikiと電子メールによって知識の蓄積が可能な教育資料編纂・閲覧システム「MotDB」を作成した。これを用いて、（１）キュレーターの育成のためのキューレーションDBの調査、（２）アノテーターの育成のための、アノテーターからのノウハウ抽出実施と教育用テキストの作成、（３）DBマネージャーの育成のための実習書の作成を行った。

図 5.2.1 教育資料編纂・閲覧システム「MotDB」

（1）キュレーターの育成

ハイスループット化されたゲノム研究手法により、大量のゲノム配列情報が作り出され、膨大な生物種のゲノム情報が刻々と蓄積し続けている。同時に、従来の小規模で精密な実験やゲノム規模の実験による、発現プロファイル解析、タンパク質相互作用解析や構造ゲノミクス解析、また、それに伴う計算機を用いた大規模予測がなされ、さまざまな角度からタンパク質機能の解析が行われている。それらのデータの統合活用には、自動アノテーションのチェックや文献からの情報抽出、コンピュータ解析によるアノテーションの修正と付加などをタスクとするキュレーション過程の整備と、その担い手であるキュレーターの育成が必要かつ急務であると考えられる。

このような背景から、キュレーターの育成のために、実際にキュレーションが行われている機関での作業内容や流れを調べることを目的として、文献等の資料、及びWeb公開情報をもとに、キュレーション作業の実際を調査した。その結果、著名な生物学データベースの領域でのキュレーションとキュレーターの作業の実際が明らかになった。また、著名なキュレーション型データベースの比較を行い、それぞれのデータベースの人員的な規模や、キュレーションがデータベース全体のなかでどのように関与しているのかを明らかにした(参照)。

表 5.2.1 代表的なキュレーション型 DB のスタッフ規模

DB	組織	キュレーター	コーディネーター	ソフトウエア技術者、プログラマー	参考
RefSeq	NCBI	30	3	38	http://www.ncbi.nlm.nih.gov/RefSeq/staffcredits.html
UniProtKB	EBI	26	5	36	http://www.ebi.ac.uk/Information/Staff/viewgallery_seqdb.php?cid=4
	PIR	12	2	3	http://pir.georgetown.edu/pirwww/about/staff.shtml
	SIB	52*1	5	13	http://au.expasy.org/people/swissprot.html
PATHWAY MAP BRITE KO	KEGG	21*2	-*3	4	http://kanehisa.kuicr.kyoto-u.ac.jp/people.html

*1. キュレーターという肩書きがなくアノテーターとなっていたが、その方々がいわゆるキュレーション作業もすると判断した
*2. KEGGは基本的な生物情報だけでなく、化学物質から疾患情報、薬剤情報など非常に広範囲にわたるDBを作成している。21名という数字はそれら多岐にわたるDBの作成にあたるスタッフの総数であり、実際に生物分野を担当しているスタッフは数名であると考えられる
*3. いわゆるコーディネーターという肩書きはない、前出の21名スタッフが自分の担当分野でおいて状況に応じてコーディネーター的役割もはたしてると考えられる

（2）アノテーターの育成

アノテーターの育成のために、ゲノムアノテーションの実務に携わるアノテーターを支援する教育用システムを作成し、同時に実践的なアノテーター教育テキストを作成することを目的として、①ゲノムアノテーションのノウハウの抽出、②アノテーターによる手動アノテーションの手順を模倣・再現するプログラムの作成、③アノテーションにかかわる知識や注意点を文書化しゲノム解析型統合DB構築に役立つ「アノテーション教育テキスト」のWiki上での作成を行った。

①については、かずさ DNA 研究所で実際に大量ゲノム解析に携わる高度専門技術を有するアノテーターからの聴き取りならびに実務調査を行った。その結果、実際に用いているツールや DB の利用手順、解析結果の解釈法などのノウハウを抽出できた。②については、抽出した情報の解析を元に、アノテーターによる手動アノテーションを模倣するプログラムを作成した。これにより、DB やツールの利用方法やその結果の解釈、さらに実行したアノテーションの根拠を明示することで初心者の学習を支援するプログラムが作成できた。③については、上記項目で抽出されたノウハウをふくめ、具体例をあわせた実践的「アノテーター教育テキスト」コンテンツを作成した。

（3）DB マネージャーの育成

ライフサイエンス関係のデータベースを構築・維持管理の実際を行う DB マネージャーの育成ための教育資料編纂・閲覧システムと実際のコンテンツ構築を行うことを目的に、上記の wiki と電子メールによって知識の蓄積が可能な教育資料編纂・閲覧システムの作成を行い、このシステムを用いて DB マネージャー養成のための実習書を作成した。DB を管理する DB マネージャーに必要不可欠なスキルとして大きく DB 構築とその維持があるので、教科書もそれに合わせて、構築編と維持管理編の二つのカテゴリーからなる構成とした。本年度は既存の DB の維持管理に必要なスキルを解説する維持管理編の構築を主に進めた。二つのカテゴリーには入らない事柄を、基礎編（DB マ

図 5.2.2 DB マネージャー教育用教科書の一部

ネージャーに必要な基礎的な知識を整理）、及びリファレンス編（各編に共通したリファレンスとなりうる事項を収集）としてまとめた。維持管理編は、日々の管理、アップデート、バックアップを中心に作成した。

５．３　人材の育成のまとめ

ライフサイエンス統合データベースを開発、維持していくために不可欠な専門職員であるキュレーター、アノテーター、DBマネージャーの育成に関して、そのベースとなるwikiと電子メールによって知識の蓄積が可能な教育資料編纂・閲覧システムを開発した。さらに、代表的キュレーション型データベース調査などに基づきキュレーションの実際を明らかにし、かずさDNA研究所の実績をベースにしたアノテーター教育システムの構築、およびDBの維持管理をターゲットにしたDBマネージャー教育用教科書の編纂を実施した。これらにより、上記専門職員の人材育成に関する基盤が構築できたものと考える。現在不足している、これら専門職員の育成を図っていくためには、今回開発したシステム、コンテンツをベースにその内容を充実していくことが重要と考える。

６．プロジェクトの総合的推進

６．１　研究運営委員会及び統合 DB 整備戦略作業部会

プロジェクト全体の連携を密としつつ円滑に推進・運営していくため，ライフサイエンス、知識情報処理、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、統合化 DB の整備戦略を議論した。また、研究運営委員会の実働部隊（情報の収集・分析、動向調査、戦略立案支援、など）として統合 DB 整備戦略作業部会を設けた。

（1）運営委員会及び作業部会の構成と活動経過

研究運営委員会のメンバーは下記の通りである。

情報・システム研究機構	堀田凱樹
	小原雄治
	五條堀孝
	大久保公策
自然科学研究機構基礎生物学研究所／JST	勝木元也
東京大学／JST	高木利久
科学技術振興機構（JST）	大倉克美
東京大学	吉田光昭
東京大学	辻井潤一
京都大学	金久　實
大阪大学	中村春木
理化学研究所	榊　佳之
産業技術総合研究所	秋山　泰
東京理科大学	増保安彦
JT 生命誌研究館	中村桂子
DNA チップ研究所	松原謙一
かずさＤＮＡ研究所	田畑哲之
九州大学	久原　哲

統合 DB 整備戦略作業部会のメンバーは下記の通りである。

五條堀孝
田畑哲之
大久保公策
菅原秀明
高木利久
高野明彦
黒田雅子
藤山秋佐夫
久原哲
中村桂子
増保安彦

以下に示す研究運営委員会，統合 DB 整備戦略作業部会の開催を行った。

①第一回研究運営委員会・統合 DB 整備戦略作業部会（合同会議、2006 年 11 月 8 日））

②第二回統合 DB 整備戦略作業部会（2006 年 12 月 18 日）

③第二回研究運営委員会（2006 年 12 月 25 日）
④第三回研究運営委員会・統合 DB 整備戦略作業部会（合同会議、2007 年 2 月 1 日））
⑤第四回研究運営委員会・統合 DB 整備戦略作業部会（合同会議、2007 年 3 月 19 日））

（2）運営委員会及び作業部会の結論

第一回研究運営委員会・統合 DB 整備戦略作業部会、及び第二回統合 DB 整備戦略作業部会と第二回研究運営委員会においては、統合データベース整備事業の全体構成、及び平成 18 年度の実施状況について討議が行われた。第三回研究運営委員会・統合 DB 整備戦略作業部会では、平成 18 年度の実施状況についての討議と共に、今後の具体的進め方についての討議が行われた。第四回研究運営委員会・統合 DB 整備戦略作業部会では、18 年度成果の報告を行った。以上の運営委員会及び作業部会の実施によって「ライフサイエンス分野の統合データベース整備事業」のプロジェクトの総合的推進が達成された。

６．２　教育プロジェクトに関するミーティング

統合 DB の人材の育成プロジェクト（教育プロジェクト）に関して、外部有識者を招聘して意見を聞くための以下の会を開催した。

①統合 DB の教育プロジェクトに関する打合わせ（2006 年 10 月 17 日）
②「シニア世代と学部教育」第一回ミーティング（2007 年 1 月 29 日）

「統合 DB の教育プロジェクトに関する打合わせ」においては、外部有識者として、実際のゲノムのアノテーションを通じて教育を実践されている長浜バイオ大学の池村教授を招いて、ゲノムアノテーションを通じたアノテーション教育のあり方と今後の方針についての討論を実施した。「シニア世代と学部教育」第一回ミーティングにおいては、池村教授が提唱されている「シニア研究者の高度な知識の活用・継承によるアノテーション教育」について、シニア世代 6 名と現役世代 3 名の研究者を招き討論を行った。これらの討論の結果は、「ライフサイエンス分野の統合データベース整備事業」の全体計画策定において反映された。

７．プロジェクトの成果のまとめと評価

あるユーザにとっての情報の Utility（利便性）とは、
Utility = Relevance(質問への関連度)　x Validity(有効度)／Work to Access（入手労力）
として定義される。

１８年度の実施項目のなかでは、分子データに関してあらゆるユーザにとって Relevant な情報への Work to Access を減らすために分類整理を行った。Relevance を表現する手法としては、地理や施設、生物種、解剖など最も多くのユーザに共有されている実世界の整理軸を用いることで、多様なデータ群が統合的に分類可能で Workt to Access を低下させる効果があることが証明された。またこれらの分類は検索後の結果提示にも有効であり、キーワード検索と組み合わせることでより Relevant な情報に簡単に導くことが可能であると期待できる。分類に使用した辞書や分類機、オントロジーは今後も開発を続けることでより感度と精度を増すことが期待され、全ての開発物は次年度からの統合に利用可能である。

一方生命科学領域のデータは配列以外非常に Validity や Reliability, Accuracy が低いと考えられており、Valid なデータを作り出す努力も統合目的にかなう仕事であると考える。発現データについて Reference データを作ることを計画していたが、統合整理し比較表示までの開発は行ったものの相互に矛盾するデータセットが予想外に多く、Reference 作成の方針を立てることが出来なかった。

配列や構造などの説明不要な分子データ以外は多くのデータがコンテキストに大きく依存する観測であり、研究単位で詳細に検討をしない統合は意義が希薄である。すなわちデータは論文のサプリメント情報と扱うべきであると結論した。あわせて大型のデータを伴わない論文も基礎研究、臨床研究論文ともに個々の報告の解釈は説得力を十分に持つものでなく、治療法や診断法を変更させる信頼度は存在しない。従って個別論文は「作業報告」として「操作と結果」にあたる FACT 部分を切り出し相互に強く関連するものをまとめることにより「信頼できる判断」と「証拠」の対を形成してゆくことでおおきな「知識」をくみ上げてゆかねばならない。これは臨床研究における「Evidence Based Medicine」における「Clinical Evidence」と同様である。癌研究や多型研究についてはまさにクリニカルエビデンスを作成することが統合であると結論した。今後のプロジェクトでは「BioMedical Evidence」の生成とＤＢ化によって論文情報を統合してゆく方針である。

８．成果の外部への発表

次ページの添付様式を参照されたい。

９．実施体制

別表１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
06026018	18	和誌	生命科学データベースの現状と課題	大久保 公策	科学	77	4	364-369	2007 年 4 月	
06026020	18	洋誌	D-HaploDB: a database of definitive haplotypes determined by genotyping complete hydatidiform mole samples.	Higasa K, Miyatake K, Kukita Y, Tahira T, Hayashi K	Nucleic Acids Res.	35		D685-689	Jan. 2007	
06026020	18	洋誌	QSNPlite, a software system for quantitative analysis of SNPs based on capillary array SSCP analysis.	Tahira T, Okazaki Y, Miura K, Yoshinaga A, Masumoto K, Higasa K, Kukita Y, Hayashi K	Electrophoresis	27		3869-3878	Oct. 2006	
06026020	18	洋誌	Novel Mutations in Norrie Disease Gene in Japanese Patients with Norrie Disease and Familial Exudative Vitreoretinopathy.	Kondo H, Qin M, Kusaka S, Tahira T, Hasebe H, Hayashi H, Uchino E, Hayashi K	Invest. Ophthalmol. Vis. Sci.	48		1276-1282	Mar. 2007	

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
06026018	18	国内	データベースから見たライフサイエンスプロジェクト	高木 利久	日本分子生物学会 2006 フォーラムシンポジウム[プロジェクト型研究時代の生命科学の課題]	2006 年 12 月 8 日	
06026018	18	国内	知的生産性向上のための情報処理	大久保 公策	日本分子生物学会 2006 フォーラムシンポジウム[プロジェクト型研究時代の生命科学の課題]	2006 年 12 月 8 日	
06026018	18	国内	ライフサイエンスDB その歴史とわが国の現状と課題	高木 利久	日本分子生物学会 2006 フォーラム バイオテクノロジーセミナー	2006 年 12 月 8 日	
06026018	18	国内	オントロジーや辞書は役に立つのか	大久保 公策	日本分子生物学会 2006 フォーラムﾊﾞｲｵﾃｸﾉﾛｼﾞｰｾﾐﾅｰ	2006 年 12 月 8 日	
06026018	18	国内	使い倒し系バイオインフォマティクスによる知のめぐりのよい生物学研究のすすめ	坊農秀雅	お茶の水女子大学「魅力ある大学院教育」第 6 回バイオインフォマティクスへの招待	2007 年 3 月 16 日	
06026018	18	国内	統合 DB の構築に必要な情報技術	高木 利久	情報とシステム 2007	2007 年 3 月 1 日	

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
06026018	18	国内	ライフサイエンスのデータベースの現状と課題	大久保　公策	情報とシステム 2007	2007 年 3 月 1 日	
06026018	18	国内	ライフサイエンスにおけるゲノム情報の高度利用に向けた生命知識の構造化	大久保　公策	知の構造化ワークショップ－知の構造化ツールは、新しいサイエンスを開くのか－	2006 年 12 月 4 日	
06026018	18	国内	知識発見のための癌臨床情報のデータベース化	加藤　菊也	第2回　大阪大学臨床医工学融合研究教育センターシンポジウム	2006 年 10 月 8 日	
06026020	18	国際	Analysis of Genes Affecting Susceptibility to Systemic Lupus Erythematosus (SLE).	OTahira T, Horiuchi T, Sakaguchi D, Yamai M, Miyagawa H, Tsukamoto H, Hayashi K	Annual Meeting of American Society of Human Genetics,	Oct. 9-13,2006	New Orleans, U.S.A
06026020	18	国際	Capillary array SSCP analysis of pooled DNA for association testing.	OHayashi K, Masumoto K, Y. Okazaki, A. Yoshinaga, K. Higasa, Y. Kukita, T. Tahira	Annual Meeting of American Society of Human Genetics,	Oct. 9-13, 2006	New Orleans, U.S.A
06026020	18	国際	D-Haplo: A genome-wide definitive haplotypes determined using complete hydatidiform moles.	OHayashi K	Human Genome Variation 2006	Sep. 14-16, 2006	Hong Kong, China. (Invited)
06026020	18	国際	The power of Definitive Haplotypes in association studies.	OMiyatake K, Kukita Y, Higasa K, Wake N, Hirakawa T, Kato H, Matsuda T, Tahira T, Hayashi K	Human Genome Variation 2006	Sep. 14-16, 2006	Hong Kong, China
06026020	18	国際	Periodicity of SNP distribution around transcription start sites.	OHigasa K, Hayashi K	Human Genome Variation 2006	Sep. 14-16, 2006	Hong Kong, China

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
06026018	18	「ライフサイエンス分野の統合データベース整備事業」の成果公開について－ライフサイエンス研究の発展に向けて－	―	―

別表１　平成１８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.　データベース統合戦略立案および評価	情報・システム研究機構	◎堀田　凱樹
	東京大学大学院新領域創成科学研究科	○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
	かずさDNA研究所植物ゲノム基盤研究部	中村　保一
	JT生命誌研究館	中村　桂子
	情報・システム研究機構 国立情報学研究所	高野　明彦
	情報・システム研究機構 国立情報学研究所	藤山　秋佐夫
	情報・システム研究機構 国立遺伝学研究所	五條堀　孝
	情報・システム研究機構 国立遺伝学研究所	菅原　秀明
2.　データベース統合化基盤技術開発	情報・システム研究機構 国立遺伝学研究所	○大久保　公策
	東京大学大学院新領域創成科学研究科	高木　利久
	情報・システム研究機構 新領域融合研究センター	川本　祥子
	情報・システム研究機構 国立情報学研究所	武田　英明
	情報・システム研究機構 国立遺伝学研究所	西川　　建
	情報・システム研究機構	三橋　信孝
	情報・システム研究機構	水田　洋子
	大阪府立成人病センター	加藤　菊也
	九州大学生体防御医学研究所	林　　健志
3. ポータルサイトの構築	自然科学研究機構基礎生物学研究所	○勝木　元也
	科学技術振興機構研究基盤情報部	大倉　克美
	科学技術振興機構研究基盤情報部	黒田　雅子
	科学技術振興機構研究基盤情報部	小池　俊行
	埼玉医科大学ゲノム医科学研究センター	坊農　秀雅
	東京大学医科学研究所ヒトゲノム解析センター	川島　秀一
	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
	情報・システム研究機構 新領域融合研究センター	川本　祥子
	情報・システム研究機構 国立遺伝学研究所	大久保　公策
4.　人材の育成	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 新領域融合研究センター	川本　祥子
	科学技術振興機構研究基盤情報部	黒田　雅子
	かずさDNA研究所植物ゲノム基盤研究部	中村　保一
	埼玉医科大学ゲノム医科学研究センター	坊農　秀雅
	情報・システム研究機構	岡本　　忍

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。

参考資料（９）

「統合データベースプロジェクト」中間評価委員会設置要綱

１．設置の目的

平成２０年度が統合データベースプロジェクトの開始から３年目にあたることから、本プロジェクト全体にかかる中間評価を公正かつ適正に行い、今後の事業展開に資することを目的とする。

２．組織等

（１）本委員会は、ライフサイエンスに関する分野の研究基盤整備について学識経験のあるプロジェクト実施者以外の者で構成する。（資料１－２参照）

（２）本委員会には主査を置き、文部科学省研究振興局ライフサイエンス課が指名する。

（３）本委員会は主査が招集する。

（４）本委員会は、委員の２分の１以上の者の出席がなければ開会することができない。

（５）本委員会の議事は、出席した委員の過半数の同意を持って決し、可否同数のときは主査の決するところによる。

（６）本委員会に出席できない委員は、主査又は他の委員にその権限を委任することができる。この場合、当該委員は委員会に出席したものとみなす。

（７）委員の委嘱期間は、平成２０年３月３日から平成２１年３月３１日までとする。

３．情報公開

本委員会は参加機関の利害に関わる検討を行うため、会議及び議事については非公開とする。

４．守秘義務

委員は、本委員会において知り得た情報については他に漏らさないものとする。

５．庶務

本委員会に係る庶務は、文部科学省研究振興局ライフサイエンス課において処理する。

６．附則

本要綱は、平成２０年３月３日から適用する。

参考資料（１０）

「統合データベースプロジェクト」
中間評価委員会評価委員名簿

漆原　秀子　筑波大学大学院生命環境科学研究科　教授

岡田　清孝　自然科学研究機構基礎生物学研究所　所長

鎌谷　直之　東京女子医科大学附属膠原病リウマチ痛風センター所長・教授

末松　誠　　慶應義塾大学医学部長

林　哲也　　宮崎大学フロンティア科学実験総合センター　教授

松原　謙一　（株）DNA チップ研究所　代表取締役社長

水澤　博　　（独）医薬基盤研究所　生物資源研究部長

山本　博一　（株）アステラス製薬　研究本部　研究推進部部長

参考資料（１１）

※「調査票」より抜粋

中 間 成 果 実 績 一 覧

中核機関　整備実績

（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およびヒトの多型データを保有
②試料・ライブラリー等の種類、数	シアノバクテリア 12 種、根粒菌 11 種、ミヤコグサ 完全胞状奇胎試料、末梢血由来 DNA など
① 測定方法	ゲノム塩基の配列決定、DNA アレイ法による SNP アレルタイプ決定など
② 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 350、海外 50 の DB 注釈データ 横断検索：検索対象 30 サイトの検索用インデックスデータ ISND 全文検索：国際 DNA 配列バンク ISND の 800 万レコードの構造化データ 統合 TV：動画コンテンツ 59 種類 生命科学学協会検索：国内 594 の学協会情報 学会要旨統合俯瞰システム：3 学会 7400 要旨のテキストデータ 蛋白質核酸酵素全文検索：20 年間 5000 記事のテキストデータ WINGpro：国内外 420 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：05 Mbase 分のゲノム塩基配列に対する、100,680 遺伝子を対象とした遺伝子注釈情報
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
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 10 名
②キュレータ･アノテータ数	中核機関グループ全体で延べ 11 名
③データ構造	関係データベース
④DB 管理ソフト	MySQL、Postgres
⑤サーバの OS	Linux
⑥サーバ規模	PC クラスタ 8 ノード、16CPU など
⑦DB へのアクセス数	本プロジェクトの成果サイトへは、昨年 10 月の公開以来 32,000 件
⑧独立 IP 数	上と同じく、18,000 件
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（データの種類（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	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。今後毎月３０件のペースで掲載データベース数を増やし、今年度中には国内を完全に網羅する。記載内容の充実化を図ることが必要。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。 登録 DB 数：400 件（H20.04 現在）開始年月：平成 20 年 2 月～ 平均アクセス数（月間）：3600
2	DB 横断検索	限定公開	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：30 データベース(H20.04 現在) 開始年月：平成 20 年 3 月
3	ISND 全文検索	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 8000 万レコードを 55 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。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を図る。検索の一層の高速化を図る。EST の臓器分類や発現データとの統合を図る。国外ユーザの獲得を図る。検索対象：8000 万レコード。開始年月：平成 19 年 3 月～ 平均利用者数（月間）：1487, 平均アクセス数(月間):10811,平均ダウンロード数(月間):187
4	統合 TV	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。平成 19 年 7 月公開。年度末の 9 ヶ月間の間に 59 個の動画コンテンツを作成、訪問数で 30,423、総計 180Gbyte のダウンロードがあった。 http://togotv.dbcls.jp/
5	生命科学学協会検索	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ー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。課題は学協会のウェブサイト等に対して検索をかける場合に許諾が必要となる場合もありうるため､各サイトのポリシーなどに配慮して進めなければならない。掲載学協会数：594 学会 開始年月：平成 20 年 4 月～
6	学会要旨統合俯瞰システム	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。課題は、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ことである。検索対象：3 学会 7400 要旨。開始年月：平成 20 年 4 月
7	蛋白質核酸酵素全文検索	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。今回、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った。今後の計画は掲載論文をさらに過去にさかのぼって増やすこと。また、近刊についてはタイトルや要旨検索を可能にすること。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。課題は著作権問題に対処可能な体制が必要ことである。そのため、著作権や知財に詳しい弁護士との協議を重ねている。収録論文：1985 年～2005 年。公開予定年月：平成 20 年 5 月

8	MotDB	公開	DBCLS を担っていく人材であるアノテータ、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 Masterof the database の略）。平成 19 年 3 月公開。平成 20 年 3 月リニューアルして、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とした。リニューアル後、訪問数で 1,830、計 457Mbyte のダウンロードがあった。http://MotDB.dbcls.jp/
9	ヒト遺伝子発現統合	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。 http://okubolab.genes.nig.ac.jp/bodymap_i/
10	ライフサイエンス受入れデータベース	試験的公開	cDNA 関連のデータベースとして、HGPD（Human Gene Protein Database)、FLJ Human cDNA DB（スプライシングバリアント DB）と FLJ-DB（ヒト完全長 cDNA)、DBTSS 、MiBASE（トマト EST DB）と KafTom（トマト全長 cDNA DB)、及び BodyMap、また理研補完課題に関連して KATANA（シロイヌナズナ DB）の受入と連係方法を検討し、受け入れテンプレートを用いてその一部の受け入れの試行を行った。
11	WINGpro	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、420DB 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 197 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。利用状況（平均利用者数 1,300（4-6 月は 300 程度であったが、その後の平均値)、アクセス数　138,208（平成 19 年度)) http://wingpro.lifesciencedb.org/
12	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	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（平均利用者数　100、アクセス数 2,673（平成 19 年度))。 http://tools.lifesciencedb.jp/
13	Jabion Genome Viewer：HAL	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。http://www.bioportal.jp/genome/
14	KazusaAnnotation	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 現在はβ版を運用しながらゲノムデータベースに対する情報蓄積を行い、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、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。公開はH19.10. 3で、一日平均694アクセス。http://a.kazusa.or.jp/
15	KazusaNavigation	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。引き続き改善を加えながら運用する。まだ利用者数、アクセス数ともに少ないが、今後、メーリングリストのよりよい代替として、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。公開はH19.10.12で、登録者数45名、一日平均27アクセス（Wikiと共通） http://navi.kazusa.or.jp/
16	KazusaWiki	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。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。公開年月日はH19.10.23で、登録者数、アクセス数は上と同じ。 http://navi.kazusa.or.jp/wiki/index.php
17	Shared database of Applied Genomics (SJAG)	限定公開	多型情報大規模解析結果（生データを含む）のコミュニティ内での共有を目指して開発したもので、評価を目的とした限定開示を行った。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、my.sql で管理している。

（３）ＤＢ基盤システム、ツール等開発成果物の整備

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	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、遺伝子が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。 http://lifesciencedb.mext.go.jp/result/tech.html#id2-1-1
2	生物学名羅日英対応辞書	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)の登録エントリー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。登録データ数は合計14028 種となっている。 http://lifesciencedb.mext.go.jp/result/tech.html#id2-1-1
3	施設名称辞書	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。 http://lifesciencedb.mext.go.jp/result/tech.html#id2-1-1
4	動物解剖学自動分類タガー	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、上記のカテゴリーに分類する。 http://lifesciencedb.mext.go.jp/result/tech.html#id2-1-1
5	植物解剖学自動分類タガー	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。 http://lifesciencedb.mext.go.jp/result/tech.html#id2-1-1
6	都市名国名自動検出タガー	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。 http://lifesciencedb.mext.go.jp/result/tech.html#id2-1-1
7	ポリゴンマン辞書	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、改変にロバストな表現が可能。今後も用語の追加を続ける。http://lifesciencedb.jp/ag/pgm
8	アナトモグラフィー	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。http://lifesciencedb.jp/ag/
9	受入れテンプレート	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ー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。 http://togodb.dbcls.jp/
10	統合ウェブサービス	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。H20.4.1 公開。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。http://togows.dbcls.jp/
11	DBCLS OpenID サーバ	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。 H20.4.1 公開。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。 http://openid.dbcls.jp/
12	OReFiL	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。H19.8.6 公開。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。http://orefil.dbcls.jp/
13	Allie	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。H20.4.1 公開。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、それ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。http://allie.dbcls.jp/
14	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。
15	WiredMarker	公開	情報共有ツール、アノテーション支援ツール。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。 公開日：H19.12.19　総ダウンロード数：8265 件　http://www.wired-marker.org
16	メタデータ要素レポジトリ(MDeR)	試験的公開	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。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。
17	グリッド環境構築	未公開	グリッドの管理サーバ、Ｗｅｂサーバと計算サーバからなるグリッド環境を構築し、2.のプロトタイプ（配列解析）システムを開発
18	配列解析システム（プロトタイプ）	未公開	ＣＢＲＣで開発された POODLE-S,-L,-W および TMBETA-NET, TMBETADISC-COMP, GRIFFIN を一度に起動し結果を得るシステム。キーワードを入力するとＬＳＤＢの生命科学データベース横断検索を行い結果を得ることも可能。
19	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 サーバークライアント。現在は試験運用中、テスト終了後、ソースコードも含め完全公開し配布を行う予定。
20	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。

21	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。

（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	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。http://lifesciencedb.jp
2	事業広報サイト	公開	統合データベースプロジェクトの紹介サイト　http://lifesciencedb.mext.go.jp/
3	ライフサイエンスの広場	公開	文科省のライフサイエンス政策等の紹介サイト http://www.lifescience.mext.go.jp/index.html
4	統合データベース間の連携調査	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。 http://charles.kazusa.or.jp/lifesciencedb/
5	検索アルゴリズムを含めた知識情報技術の動向調査	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。 http://lifesciencedb.mext.go.jp/result/strategy.html#id1-4
6	遺伝統計学分野における解析技術の基礎調査	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。 http://lifesciencedb.mext.go.jp/result/strategy.html#id1-4
7	臨床情報・疾患健康情報の調査	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。 http://lifesciencedb.mext.go.jp/result/strategy.html#id1-4
8	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。
9	データベース受入に関する	試験的	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

	アンケート調査	公開	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。
10	BioHackathon ホームページ	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 BioHackason2008 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。 http://hackathon.dbcls.jp/
11	データベース、ソフトウェアの SOAP 化	未公開	将来のワークフローシステム構築の準備として、産総研 CBRC で開発した 7 種類のソフトウェアの SOAP 化を行った。

中核機関　外部発表実績一覧

（1）セミナー、研究会等イベント開催

通番	タイトル	発表者（代表者）	年月日	開催場所	イベント名称	概要（対象者（層、参加人数）、出席者の主な反応等）
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	川本祥子	H20.1.25	長浜バイオ大学	統合データベース講演会	学部学生、他機関研究者等 60 名
2	バイオデータベース構築	中谷洋一郎	同上	同上	同上	同上
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	中村保一	同上	同上	同上	同上
4	バイオデータマイニング入門	瀬々　潤	同上	同上	同上	同上
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	坊農　秀雅	同上	同上	同上	同上
6	テキストを対象とした生命科学研究	山本　泰智	同上	同上	同上	同上
7	Objectives of this hackathon and current status of Japanese web services	T. Katayama	H20.2.11	六本木アカデミーヒルズ	BioHackathon2008 Presentation	海外研究者 34 名、国内研究者 37 名
8	From Web API for Biology (WABI) to Semantic Web API for Biology (SABI)	H. Sugawara	同上	同上	同上	同上
9	Current status of the BioMOBY project and vision for the future directions	M. Wilkinson	同上	同上	同上	同上

10	Moby, Legacy Apps and the Semantic Web	P. Gordon	同上	同上	同上	同上
11	Data, Services and Computational Resources Integration at the INB	O. Trelles	同上	同上	同上	同上
12	The EMBRACE project and WS-I standard	J.C. Bryne	同上	同上	同上	同上
13	Soaplab2 project to wrap up command line packages	M. Senger	同上	同上	同上	同上
14	Taverna (part of myGrid project)	T. Oinn	同上	同上	同上	同上
15	Generation Challenge Program effort at building interoperability	R. Bruskiewich	同上	同上	同上	同上
16	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	坊農秀雅、	H20.3.5	JST 東京本部住宅等棟	統合データベース講習会: AJACS 東京	学部生、大学院生 20 名
17	かずさにおける新しいアノテーションの試み	岡本忍	同上	同上	同上	同上
18	統合データベースへの期待	樋口千洋	同上	同上	同上	同上
19	ゲノム情報リテラシー入門	中村保一他	同上	同上	同上	同上
20	テキスト処理して生命科学	山本泰智	H19.6.20	お茶の水女子大学	第 7 回 ﾊﾞｲｵｲﾝﾌｫﾏﾃｨｸｽへの招待	内部・外部学生、その他 22 人
21	KEGG データベースの進化と、統合・標準化	片山俊明	H19.6.20	お茶の水女子大学	第 7 回 ﾊﾞｲｵｲﾝﾌｫﾏﾃｨｸｽへの招待	内部・外部学生、その他 22 人
22	選択的スプライシングと植物の環境適応	飯田慶	H19.9.3	お茶の水女子大学	第 8 回 ﾊﾞｲｵｲﾝﾌｫﾏﾃｨｸｽへの招待	内部・外部学生、その他 37 人
23	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	程久美子	H19.10.29	お茶の水女子大学	第 9 回 ﾊﾞｲｵｲﾝﾌｫﾏﾃｨｸｽへの招待	内部・外部学生、その他 29 人
24	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	深海薫	H19.12.6	お茶の水女子大学	第 10 回 ﾊﾞｲｵｲﾝﾌｫﾏﾃｨｸｽへの招待	内部・外部学生、その他 16 人
25	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	荻島創一	H20.2.20	お茶の水女子大学	第 11 回 ﾊﾞｲｵｲﾝﾌｫﾏﾃｨｸｽへの招待	内部・外部学生、その他 31 人

26	生物学者のための生命情報解析ツールの開発	田村浩一郎	H20.3.7	お茶の水女子大学	第12回 ﾊﾞｲｵｲﾝﾌｫﾏﾃｨｸｽへの招待	内部・外部学生、その他 29人
27	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	H19.12.3	かずさアカデミアパーク	ラン藻の分子生物学2007	ラン藻研究者、約100人、KazusaAnnotation, KazusaNavigation, KazusaWiki を限定公開して実際に使っていただき、ツールの有用性
28	第八回オープンバイオ研究会企画運営	中尾光輝	H19.12.18	科学未来館	第八回オープンバイオ研究会	JSBi 年会参加者、約20名、国内のオープンソース関連の開発者とユーザの活発な交流の場となった。

（２）プレス発表、取材対応

通番	タイトル	発表媒体	年月日	特記事項
1	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)	H19.8	
2	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	H19.10.5	
3	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	H19.11.07	
4	「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	H20. 2.12	
5	生命科学DB 統合へ	朝日新聞	H20. 3.31	

（３）展示会等出展（該当なし）

（４）学会等への口頭発表

通番	タイトル	発表者	学会等の名称	年月日	特記事項
1	データベースから見たライフサイエンスプロジェクト	高木利久	日本分子生物学会2006 フォーラムシンポジウム[プロジェクト型研究時代の生命科学の課題]	H18.12.8	

2	知的生産性向上のための情報処理	大久保公策	同上	同上	
3	生物学研究基盤としてのゲノムアノテーション	中村保一	同上	同上	
4	ライフサイエンスＤＢ　その歴史とわが国の現状と課題	高木利久	日本分子生物学会 2006 フォーラム バイオテクノロジーセミナー	同上	
5	オントロジーや辞書は役に立つのか	大久保公策	同上	同上	
6	使い倒し系バイオインフォマティクスによる知のめぐりのよい生物学研究のすすめ	坊農秀雅	お茶の水女子大学「魅力ある大学院教育」第 6 回バイオインフォマティクスへの招待	H19.3.16	
7	統合 DB の構築に必要な情報技術	高木利久	情報とシステム 2007	H19.3.1	
8	ライフサイエンスのデータベースの現状と課題	大久保公策	同上	同上	
9	ライフサイエンスにおけるゲノム情報の高度利用に向けた生命知識の構造化	大久保公策	知の構造化ワークショップ－知の構造化ツールは、新しいサイエンスを開くのか－	H18.12.4	
10	知識発見のための癌臨床情報のデータベース化	加藤菊也	第２回　大阪大学臨床医工学融合研究教育センターシンポジウム	H18.10.8	
11	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ッブとオントロジー研究会	H19.7.23	招待講演
12	An approach to decifer gene regulatory networks from the federation of databases in life science	H. Bono, S. Kawano, S. Kawamoto and T. Takagi	FUNCTIONAL GENOMICS & SYSTEMS BIOLOGY	H19.10.10-13	Hinxton (UK)
13	YUZ: an environment for integration AND ANALYSIS of gene regulatory networks	H. Bono, S. Kawano, S. Kawamoto, and T. Takagi	GENOME INFORMATICS	H19.11.1-5	New York (USA)
14	統合 TV によるがんとハイポキシア研究の推進	坊農秀雅	第 5 回　がんとハイポキシア研究会	H19.12.1	千葉
15	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、坊農秀雅 他 2 名	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)	H19.12.11-15	横浜
16	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　他 11 名	同上	同上	同上
17	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：WINGpro	黒田雅子、小池　俊行　他 10 名	同上	同上	同上
18	ライフサイエンス統合データベースセンターに	河野信、小野	同上	同上	同上

	おけるデータベースポータル構築	浩雅 他7名			
19	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 他3名	同上	同上	同上
20	遺伝子発現統合データベースの開発	有川浩司、飯塚高康 他3名	同上	同上	同上
21	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 他3名	同上	同上	同上
22	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
23	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)	H19.12.17-19	東京
24	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	H20.2.18	
25	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	H20.3.7-8	能美
26	BioHackathon for Web Service の報告	片山俊明	同上	同上	同上
27	DAS workshop 2008 参加報告	中村保一	2008.03.08	同上	同上
28	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』	H20.3.25	三島
29	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
30	OreFiL	Y. Yamamoto	Pacific Symposium on Biocomputing	H20.1.4-8	Hawaii, U.S.A.
31	Funcional analysis of groups of genes with MeSH hierarchy	T. Nakazato, H. Bono et al.	Systems Biology: Global Regukation of Gene Expression	H20.3.27-30	New York, USA
32	Analysis of Genes Affecting Susceptibility to Systemic Lupus Erythematosus (SLE).	T. Tahira, K. Hayashi et al.	Annual Meeting of American Society of Human Genetics	2006.10. 9-13	New Orleans, U.S.A
33	Detection of human copy number variations	Y. Kukita, K.	The Biology of Genomes meeting	2007.5.6-10	Cold Spring

	using a collection of Japanese complete hydatidiform moles	Hayashi, et al.			Harbor Laboratory, U.S.A.
34	Enhanced D-HaploDB: definitive haplotypes and extended haplotype information determined by genotyping complete hydatidiform mole samples	Koichiro Higasa, Kenshi Hayashi, et al.	The Biology of Genomes meeting	同上	同上
35	D-HaploDB: A Database of Genome-Wide Definitive Haplotypes Determined using Complete Hydatidiform Moles	Y. Kukita, K. Hayashi, et al.	第5回国際バイオデータ相互運用性会議	H19.9.26	青海、東京
36	Evaluation of kernel-based link analysis measures on research paper recommendation	新保仁	ACM/IEEE Joint Conference Digital Libraries	H19.6.22	
37	A Discriminative Learning Model for Coordinate Conjunctionsresearch paper recommendation	新保仁	Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning	H19.6.29	
38	A Graph-Based Approach to Named Entity Categorization in Wikipedia Using Conditional Random Fields	渡邉陽太郎	同上	同上	
39	Pivot learning for efficient similarity search	木村学	International Conference, Knowledge-based Intelligent	H19.9.14	
40	ストレスフリーなデータベースにむけて	中村保一	日本植物生理学会	H19.3.28	シンポジウム「植物データベース講習会」への招待講演。
41	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	同上	同上	
42	KazusaAnnotation: ゲノム情報への注釈付け、注釈の利用を支援するシステム	岡本忍	日本ゲノム微生物学会	H19. 3. 6	
43	System update and literature curation at CyanoBase.	岡本忍	International Biocuration meeting 2007	H19.10.25	
44	学部教育としての持続可能型社会への貢献	上原啓史、池	日本遺伝学会第79回大会	H19.9.19	

	遺伝子データベースの構築	村淑道他			
45	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の構築	上原啓史、池村淑道他	第 30 回情報化学討論会	H19.11.16	
46	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、池村淑道他	第２回日本ゲノム微生物学会年会	H20.3.6	

（５）雑誌等への論文寄稿

通番	タイトル	著者名	雑誌等の名称	掲載巻、号、ページ	特記事項
1	生命科学データベースの現状と課題	大久保公策	科学(岩波)	77(4),364-369 (2007)	
2	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),1027-1031 (2007)	連載企画「ライフサイエンス分野の統合データベース」
3	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),1388-1393 (2007)	同上
4	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	同上	53(1),87-93 (2007)	同上
5	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 (2007)	同上
6	利用の立場からのコメント	高木利久	同上	53(5),686-691 (2007)	同上
7	OReFiL: an online resource finder for life sciences	Y. Yamamoto and T. Takagi	BMC Bioinformatics	8:287 (2007)	http://www.biomedcentral.com/1471-2105/8/287/
8	BioCompass: A novel functional inference tool that utilizes MeSH hierarchy to analyze groups of genes	T. Nakazato, H. Bono et al.	*In Silico* Biology	8, 0006 (2007)	http://www.bioinfo.de/isb/2007/08/0006/
9	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）	p175-180 (2007)	
10	Reactome	河野信	同上	p175-180 (2007)	
11	脊椎動物ゲノム進化を推定するロジック	森下　中谷	細胞工学別冊	29 - 40 (2007)	

12	Reconstruction of the Vertebrate Ancestral Genome Reveals Dynamic Genome Reorganization in Early Vertebrates.	Y. Nakatani, H. Takeda, Y. Kohara, and S. Morishita	Genome Research	17(9): 1254-1265	
13	Vertebrate genome evolution examined by comparing the human and fish genomes.	Y. Nakatani and S. Morishita	*Encyclopedia of Life Sciences*, John Wiley & Sons	印刷中	
14	D-HaploDB: a database of definitive haplotypes determined by genotyping complete hydatidiform mole samples.	K. Higasa, K. Hayashi et al.	Nucleic Acids Res.	35, D685-689 (2007)	
15	Periodicity of SNP distribution around transcription start sites.	K. Higasa, K. Hayashi	BMC Genomics	7, 66 (2006)	
16	A Large-scale Protein protein Interaction Analysis in Synechocystis sp. PCC6803	S. Sato, M. Nakamura et al.	DNA Research	14, 207-216 (2007)	本論文の大規模タンパク質間相互作用データを KazusaAnnotation のソーシャルブックマークとして登録し、利用に供している。
17	Complete genomic structure of the bloom-forming toxic cyanobacterium Microcystis aeruginosa NIES 843.	N. Nakajima, S. Okamoto ey al.	同上	14, 247-256 (2007)	新規に決定したシアノバクテリアゲノム論文。本プロジェクトの成果物によるアノテーション高度化の対象である。
18	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 (2008)	
19	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、池村淑道他	同上	26(1) 11-16 (2008)	
20	持続可能型社会への貢献遺伝子データベース	上原啓史、池村淑道他	同上	26(1) 17-19 (2008)	

分担機関（京都大学）　整備実績

（１）保有データ情報（該当なし）

（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（データの種類（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	ゲノムネット医薬品データベース http://www.genome.jp/kusuri	公開	研究の最先端と医療の現場さらには一般社会をつなぐ日本語の医薬品統合データベース。JAPIC 医薬品添付文書情報（医療用医薬品 13,973 件、一般用医薬品 12,658 件、2008 年 4 月現在）を検索可能。KEGG DRUG の構造情報やターゲット情報と統合している。また、文献データベースへのリンクも付加している。医療用医薬品は 2007 年 9 月、一般用医薬品は 2008 年 1 月より公開している。アクセス数等の情報は本文の進捗状況の欄を参照のこと。
2	DBGET/LinkDB: ゲノムネット統合データベース検索システム http://www.genome.jp/ja/gn_dbget_ja.html	公開	2006 年度までに DBGET/LinkDB として開発してきたシステムを、日本語支援環境の整備、LinkDB の拡張、新たな検索システムの開発という観点から改良したもの。全データベース一括検索と外部データベースを含む LinkDB 検索を 2007 年 7 月に、日本語支援環境を 2007 年 10 月に公開した。

（３）ＤＢ基盤システム、ツール等開発成果物の整備

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）
3	SIMPCOMP	公開	類似化合物検索システム。グラフ比較に基づいた精度の高い類似度計算を実現している。検索速度に問題があったため、平成１９年度には高速化についての調査を行い、平成２０年度に高速化を実現する。
4	e-zyme	公開	化学構造変化に基づく反応予測システム。基質と生成物を与えると、その間の反応パターンを予測、EC 番号との対応付けなどを行う。テンプレートとなる反応パターンの充実が課題であったため、平成２０年度に反応パターンデータベースを整備し、化合物データとリンクさせる。また、平成２１年度以降に、複数反応ステップの予測システムを実現する。
5	KCaM	公開	糖鎖類似構造検索システム。糖鎖に特徴的な木構造のための動的計画法を実装したシステムであり、ユーザーインタフェースを他のシステムと統一した。今後は、以下の糖鎖構造予測システムとの連携を計画している。
6	GECS	公開	遺伝子発現データから化合物構造を予測するシステム。ゲノム情報と化合物情報を結ぶためのシステムとして開発している。平成１９年度は糖転移酵素のリストから合成可能な糖鎖構造を予測するシステムを開発し、平成２０年４月に第１版を公開した。今後は、ユーザーインタフェースなどを改良するとともに、脂質など他の化合物のためのシステムを開発し統合する。

分担機関（京都大学）　外部発表実績一覧

（１）セミナー、研究会等イベント開催

通番	タイトル	発表者（代表者）	年月日	開催場所	イベント名称	概要（対象者（層、参加人数)、出席者の主な反応等）
1	ゲノムネットデータベース講習会	五斗進他	2008/1/30,31	東京大学		PC を用いた実習形式での講習会。ホームページ上で一般から 20 名の参加者を募った。大学、公的機関の研究所、企業から幅広く集まった。

（２）プレス発表、取材対応（該当なし）

（３）展示会等出展（該当なし）

（４）学会等への口頭発表

通番	タイトル	発表者	学会等の名称	年月日	特記事項
2	医薬品情報統合データベースの開発	伊藤真純他	BMB2007	2007.12.11-15	ポスター発表

（５）雑誌等への論文寄稿

通番	タイトル	著者名	雑誌等の名称	掲載巻、号、ページ	特記事項
3	医薬品の統合データベース	金久實他	蛋白質核酸酵素	52, 12, 1486-1491	

分担機関（東京医科歯科大学グループ）　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	（１） 東京医科歯科大学：ヒト （２） 大阪大学　　：ヒト
②試料・ライブラリー等の種類、数	（１）東京医科歯科大学：＜試　料＞肝臓癌、大腸癌、口腔癌の手術または生検検体（約３００） ＜データ＞肝臓癌、大腸癌、口腔癌の症例情報（約３００）、 肝臓癌、大腸癌、口腔癌検体の遺伝子発現解析結果（約２００） （２）大阪大学　　：＜データ＞神経難病の症例情報（約４００）
③測定方法	（１） 東京医科歯科大学：診療情報収集、DNA マイクロアレイによる遺伝子発現解析 （２） 大阪大学　　：診療情報収集
④データの内容	（１）東京医科歯科大学：臨床情報（基本情報、病歴・生活歴、臨床検査、画像診断、治療、予後等） 分子情報（マイクロアレイ遺伝子発現情報） （２）大阪大学　　：臨床情報（基本情報、症状、治療等）
⑤その他、特記事項	

（１－２）データソース

①現在のデータ量	（１） 東京医科歯科大学：試験公開可能なデータは、約１００症例分の症例情報。 非公開データを含めると、約３００のがん症例情報を有する。 （２） 大阪大学　　：試験公開可能なデータは、約１００の症例分の統計情報。 非公開データを含めると、約４００の神経難病症例情報を有する。
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
③将来の増加の見込み	倫理審査委員会が承諾する範囲内で、今後も継続的に症例情報を追加していく。
④権利関係	所有者（東京医科歯科大学、大阪大学） 公開（□可　□否　■その他［各機関の倫理審査委員会が承諾する範囲での公開］）
⑤その他、特記事項	個別に試験公開可能なデータの公開をしている。(http://ibmd.tmd.ac.jp) 各データベースの倫理規定に基づいての公開を行うとともに、統合医科学データベースの倫理規定案の策定を推進す

	る。

（１－３）データの管理状況

①更新頻度等の管理状況、体制	症例情報の収集及び検体からの遺伝子発現解析は継続的に実施しており、集積データをクレンジングし、公開用データベースへのデータ更新は年１回～２回を予定している。
②その他、特記事項	

（１－４）データベース関係

①DB 管理者数	（１） 東京医科歯科大学：３名 （２） 大阪大学　　：１名
②キュレータ・アノテータ数	（１） 東京医科歯科大学：３名（ＣＲＣ他） （２） 大阪大学　　：１名
③データ構造	各データベースで独自のデータ構造
④DB 管理ソフト	PostgresSQL
⑤サーバの OS	Linux
⑥サーバ規模	
⑦DB へのアクセス数	
⑧独立 IP 数	1
⑨その他、特記事項	

（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（データの種類（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	網羅的疾患分子病態ﾃﾞｰﾀﾍﾞｰｽ http://ibmd.tmd.ac.jp	試験公開	臨床、病理、分子情報（遺伝子発現情報）を統合化したデータベース。 疾患共通のテンプレートに情報を適応し、100 症例のデータを公開した。 平成 20 年度には 200 症例、最終年度には 300 症例のデータ公開を行う予定。
2	パーキンソンデータベース http://ibmd.tmd.ac.jp	試験公開	神経難病特の、臨床情報データベース。 100 症例の臨床 10 項目に対して、統計情報を取りこれを公開した。 平成 20 年度には 400 症例、最終年度には 500 症例のデータ公開を行う予定。

（３）ＤＢ基盤システム、ツール等開発成果物の整備

通	ＤＢ基盤システム､ツール等の名称	公開／未	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）

番		公開	
1	要件定義書	プロジェクト内部公開	統合医科学データベース構築に関わる統合化技術(情報モデル、オントロジー、セマンティクス)、標準化、公開倫理等に関して、全国の全疾患 DB を対象とした調査に基づき、要件定義を行った。継続的に要件定義を行い、平成 20 年度に要件定義書を完成する。
2	統合検索システムプロトタイプ http://ibmd.tmd.ac.jp/Prototype	未公開	東京医科歯科大学の網羅的疾患分子病態データベース、大阪大学のパーキンソンデータベースに対しセマンティクス検索技術、ユーザインターフェースを検証するためのプロトタイプシステムを構築した。 機能限定版のプロトタイプシステムを評価用に試験公開する予定。

（４）その他の成果物（(２)、（３）に該当しないもの）（該当なし）

分担機関（東京医科歯科大学グループ）　外部発表実績一覧

（１）セミナー、研究会等イベント開催

通番	タイトル	発表者（代表者）	年月日	開催場所	イベント名称	概要（対象者（層、参加人数)、出席者の主な反応等）
1	疾患データベースの統合とオミックス医療への取り組み	田中　博	2007.9.26	日本 IBM	製薬企業向け講演会	製薬企業、10 名
2	医療データベースと医科学研究の変貌	田中　博	2007.10.9	新潟県湯沢町　NASPA ニューオータニ	理化学研究所 理事長ファンドワークショップ	

（２）プレス発表、取材対応（該当なし）

（３）展示会等出展（該当なし）

（４）学会等への口頭発表

通番	タイトル	発表者	学会等の名称	年月日	特記事項
1	文部省ライフサイエンス統合データベースプロジェクトと臨床オントロジー	田中　博	医療知識基盤データベースと用語･言語・知識処理シンポジウム	2008.3.13	
2	Bioinformatics and genomics for opening new perspectives for personalised care'	Hiroshi Tanaka	CeHR 2007	2007.12.3	

3	TMDU Clinical Omics Database Project - Integration of OMICS data and Clinical Information.	Hiroshi Mizushima, Hiroshi Tanaka	The 7th International Workshop on Advanced Genomics	2007.11.27	
4	網羅的臨床情報と網羅的分子情報の統合データベースの構築	水島　洋 田中　博ほか	第 27 回医療情報連合大会	2007.11.24	
5	Omics analysis to predict the aggressive recurrence of hepatocellular carcinoma after curatve hepatectomy.	Tanaka S, Mahmut Y, Mogushi K, Aihara A, Kudo A, Nakamura N, Ito K, Imoto I, Inazawa J, Miki Y, Mizushima H, Tanaka H, Arii S.	Japan Cancer Association Annual Conference	2007.10.3-5	
6	TMDU Clinical Omics Database - Integrating OMICS data and Clinical Information.	Tanaka H, Arii S, Sugihara K, Miki Y, Inazawa J, Mizushima H.	Japan Cancer Association Annual Conference	2007.10.3-5	
7	TMDU Clinical Omics Database Project - Integration of OMICS data and Clinical Information.	Mizushima H, Mogushi K, Ohashi W, Araki E, Nishibori M, Arii S, Sugihara K, Miki Y, Inazawa J, Tanaka H.	ISMB/ECCB07	2007.7.20-26	
8	Development of TMDU Clinical Omics Database.	Fujisaki A, Araki E, Mizushima H, Tanaka H.	ISMB/ECCB07	2007.7.20-26	

（５）雑誌等への論文寄稿

通番	タイトル	著者名	雑誌等の名称	掲載巻、号、ページ	特記事項
1	オミックス情報と医療情報の網羅的統合データベースの構築	水島　洋、田中　博	計測自動制御学会システム情報部門学術講演会 2007	2007 講演論文集 P. 39-40	
2	TMDU Clinical Omics Database System -Integrating OMICS data and Clinical Information.	Mizushima H.Tanaka H.	The7th International Workshop on Advanced Genomics Abstract book	2007, p.82	
3	ライフサイエンス分野の統合データベース整備事業分担機関　「統合医科学データベース構築方式の開発」	田中　博	JSBi ニュースレター	Vol.15	
4	わが国における疾患データベースの統合化について	田中　博	「蛋白質 核酸 酵素」シリーズ『ライフサイエンス分野の統合デー	未定	

			タベース』		
5	網羅的分子病態データベースとシステム病態学	田中　博	医学の歩み	2008 年 5 月号掲載予定	
6	Aurora kinase B is a predictive factor for aggressive recurrence of hepatocellular carcinoma after curative hepatectomy	Tanaka S, Arii S, Yasen M, Mogusi K, Su N-T, Zhao C, Imoto I, Eishi Y, Inazawa J, Miki Y, Tanaka H	British Journal of Surgery	in press	
7	Bioinfomatics and genomics for opening new perspective for personalized care,	Tanaka, H	“eHealth”, (Blobel, B. ed.)	p47-58, IOS Press, 2008	

分担機関（東京大学グループ）　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	Homo sapiens
②試料・ライブラリー等の種類、数	健常者 900 検体、ナルコレプシー220 検体、多系統萎縮症 200 検体、脳動脈瘤 200 検体、パニック障害 200 検体の 50-90 万の遺伝子型データ。理化学研究所ゲノム医学研究センターが実施、保有する GWAS データを除き、国内で実施される GWAS の大部分のデータを収納する予定である。脳神経疾患関連遺伝子のリシークエンスデータ、ALS に関連する遺伝子の mutation 情報及び、臨床情報。
③測定方法	Affymetrix, Illumina の 30-90 万の SNP タイピングセット
④データの内容	検体の性別、疾患情報などの基本情報、genotype データ、genotype calling 前の画像生データ
⑤その他、特記事項	国立がんセンター吉田輝彦部長の協力を得て、同センター研究所で解析された GWAS データも登録予定

（１－２）データソース

①現在のデータ量	健常者のデータ 900 検体、ナルコレプシー220 検体、多系統萎縮症 200 検体、脳動脈瘤 200 検体, パニック障害 200 検体
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
③将来の増加の見込み	アルツハイマー病（2000 検体）、肝炎（B 型、C 型 500 検体）、１型糖尿病（400 検体）、変形性関節症（350 検体）などのデータも登録予定。
④権利関係	所有者（研究代表者に帰属　　）

	公開（□可　□否　■その他 [genotype frequency data や統計解析結果は完全公開する。個々の遺伝子型データ等は、一定の手続きを経て限定された研究者に開示する])
⑤その他、特記事項	

（１－３）データの管理状況

①更新頻度等の管理状況、体制	サーバー管理は日立製作所がおこなっているが、常駐 SE がいるわけではない。更新は新たなデータが提供された都度行う。
②その他、特記事項	

（１－４）データベース関係

①DB 管理者数	1
②キュレータ・アノテータ数	0
③データ構造	Relational DB
④DB 管理ソフト	mySQL
⑤サーバの OS	Redhat enterprise linux ES v4
⑥サーバ規模	Poweredge2900
⑦DB へのアクセス数	未公開
⑧独立 IP 数	1
⑨その他、特記事項	

（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	未公開/未公開	概要（データの種類（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	標準 DB http://133.11.184.23/snpdbnew/snp_top.php	未公開 （テストサイト有）	日本人健常者の 30 万 SNP 約 200 検体、50 万 SNP 約 500 検体、90 万 SNP 約 200 検体の genotype frequency, allele frequency, Hardy-Weinberg 平衡検定値、ハプロタイプ頻度など。 進捗状況：システムは完成しており、論文が accept され次第公開する。 今後の計画：データを随時登録していく。 （画面の snapshot については別紙参照）
2	GWAS-DB http://133.11.184.23/cgi-bin/gwasdbnew/gwas_top.cgi	未公開 （テストサイト有り）	SNP ごとの genotype frequency, allele frequency, call rate、Hardy-Weinberg平衡検定値、genotypic model, allelic model, additive risk model, recessive model, dominant model など各種遺伝統計値を

			登録している。copy number variation, OMIMなどの他の情報と共に上記計算結果をグラフ表示することが可能である。 進捗：システムを構築し、ナルコレプシー、脳動脈瘤、多系統萎縮症などを登録してある。論文がacceptされ次第公開。 今後の計画：ユーザーフレンドリーになるように、インターフェース周りの改良を行うとともに、2次スクリーニング結果を登録できるようにデータベースの拡張を行う。また、臨床情報および、臨床情報による層別化の解析なども登録していく。 　さらに、論文発表、学会発表等により、データのsubmissionを広く呼びかけていく。 （画面のsnapshotについては別紙参照）
3	ALS リシークエンス https://133.11.102.101/resequence/SearchDisease.do?targetId=1	未公開（テストサイト有り）	ALS（筋萎縮性側索硬化症）に関するリシークエンスデータベースであり、東京大医学部附属病院で産出したALS関連遺伝子のリシークエンスデータ及び臨床データのほか、論文から抽出したmutationの位置、頻度、家系情報と共に、発症してから何年で人工呼吸器をつけたか、どのような症状か等の臨床情報、更には、蛋白質の２次構造、３次構造などのデータも登録している。 進捗：システムはほぼ完成している。ユーザのフィードバックで改良を加えた後、8月ごろ公開予定。 （画面のsnapshotについては別紙参照）

（３）ＤＢ基盤システム、ツール等開発成果物の整備

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	Path consistency アルゴリズムによる遺伝子型と表現型のネットワーク解析	未公開	多変量の関連構造をネットワークグラフによって表現型と遺伝子型の関係を表現するソフト

（４）その他の成果物（（２）、（３）に該当しないもの）（該当なし）

分担機関（東京大学グループ）　外部発表実績一覧

（１）セミナー、研究会等イベント開催

通番	タイトル	発表者（代表者）	年月日	開催場所	イベント名称	概要（対象者（層、参加人数）、出席者の主な反応等）
1	Path consistency アルゴリズムを用いた相互作用解析	成田　暁	平成19年11月22日～23日	東京大学 柏キャンパス	第2回インフォマティクス研究者と医学研究者の交流会	
2	Affymetrix Genome-Wide Human SNP Nsp/Sty 6.0 による タイピングプラットフォームの構築	**西田　奈央**	平成19年11月22日～23日	東京大学 柏キャンパス	第2回インフォマティクス研究者と医学研究者の交流会	

（２）プレス発表、取材対応

通番	タイトル	発表媒体	年月日	特記事項
1	統合データベース	日本バイオインフォマティクス学会ニュースレター	平成19年9月1日	

（３）展示会等出展（該当なし）

（４）学会等への口頭発表

通番	タイトル	発表者	学会等の名称	年月日	特記事項
1	A search for genetic variants attributing to the risk of formation of intracranial aneurysms.	安野勝史	米国人類遺伝学会第57回大会	2007年10月23～27日	

（５）雑誌等への論文寄稿

通番	タイトル	著者名	雑誌等の名称	掲載巻、号、ページ	特記事項
1	疾患感受性遺伝子の探査の実際：ゲノムワイド関連解析を中心として	宮川　卓、徳永　勝士	最新医学	62巻、9月増刊号、150-157	
2	第4章 SNPによる連鎖解析	成田　暁	BIOWEB 電子出版	in press	

3	第 6 章 多重検定についての考え方および解決策	成田　暁	BIOWEB 電子出版	in press	
4	第 7 章 遺伝子間相互作用の検出法	成田　暁	BIOWEB 電子出版	in press	
5	ゲノムワイド関連解析データベースの開発	小池　麻子、西田　奈央 徳永　勝士	蛋白質核酸酵素	in press	
6	ゲノムワイド SNP タイピング技術の現状と将来	西田　奈央、徳永　勝士	医学のあゆみ	in press	

補完課題実施機関（理化学研究所）整備実績一覧

（１）保有データ情報

[1 件目]

（１－１）データの種類

① 生物種	シロイヌナズナ
②試料・ライブラリー等の種類、数	シロイヌナズナ完全長（RIKEN Arabidopsis Full-Length, RAFL） cDNA クローン：245,946 個(18,090 グループ、全遺伝子の約 60%に相当。世界最大のシロイヌナズナ完全長 cDNA のリソース)
③測定方法	
④データの内容	全長配列データ:21,130 個 ３’末端側からの端読み配列データ＝151,574 個 ５’末端側からの端読み配列データ＝82,766 個
⑤その他、特記事項	

（１－２）データソース

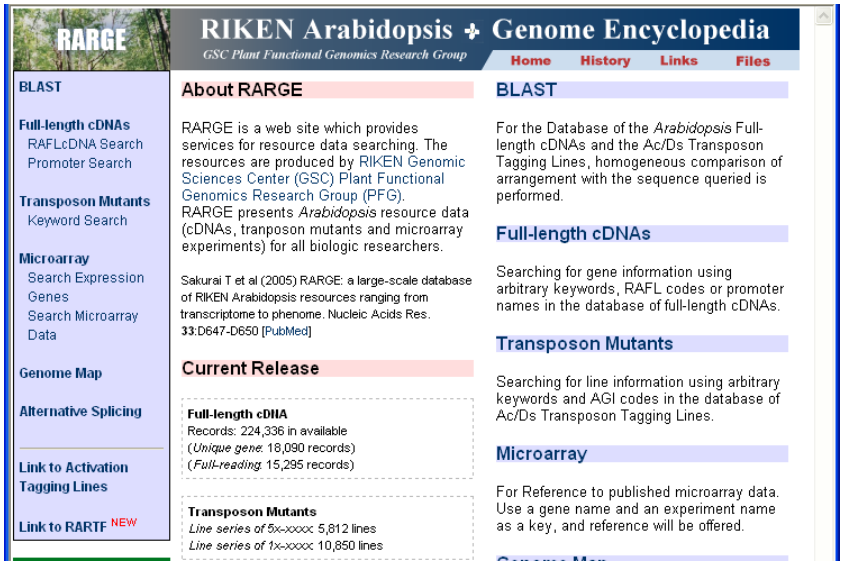
①現在のデータ量	
②データ区分	■自前　□第三者　□文献データ　□計算結果等の二次データ　□その他
③将来の増加の見込み	なし
④権利関係	所有者（　　） 公開（■可　□否　□その他［　　］）
⑤その他、特記事項	

（１－３）データの管理状況

①更新頻度等の管理状況、体制	不定期にアノテーション情報を更新。
②その他、特記事項	

（１－４）データベース関係

(1)DB 管理者数	2 名
(2)キュレータ･アノテータ数	1 名
(3)データ構造	リレーショナルデータベース
(4)DB 管理ソフト	PostgreSQL
(5)サーバの OS	Red Hat Enterprise Linux ES
(6)サーバ規模	1 基
(7)DB へのアクセス数	データベース RARGE 総計　約 5000 ページ/月
(8)独立 IP 数	1
(9)その他、特記事項	画面の様子　RARGE の画面



[2 件目]トランスクリプトーム（タイリングアレイデータ）

（１－１）．データの種類

(1)生物種	シロイヌナズナ
(2)試料・ライブラリー等の種類、数	シロイヌナズナ全ゲノムタイリングアレイを用いた発現プロファイル解析データ　19 種類(各々F および R アレイを用いた計 6 回のハイブリ実験を行う) GEO データベースに登録されている、シロイヌナズナ全ゲノムタイリングアレイを用いた発現プロファイル解析データの内、3 回以上繰り返し実験を行ったものは、わずか 4 種類のみである。
(3)測定方法	
(4)データの内容	シロイヌナズナ全ゲノムタイリングアレイを用いた発現プロファイル解析データ　19 種類(各々F および R アレイを用いた計 6 回のハイブリ実験を行う) 実験内容：播種後 2 週間目の植物体を用いた乾燥、低温、塩ストレス、ABA 処理、再吸水処理による乾燥ストレスからの回復過程など
(5)その他、特記事項	

（１－２）．データソース

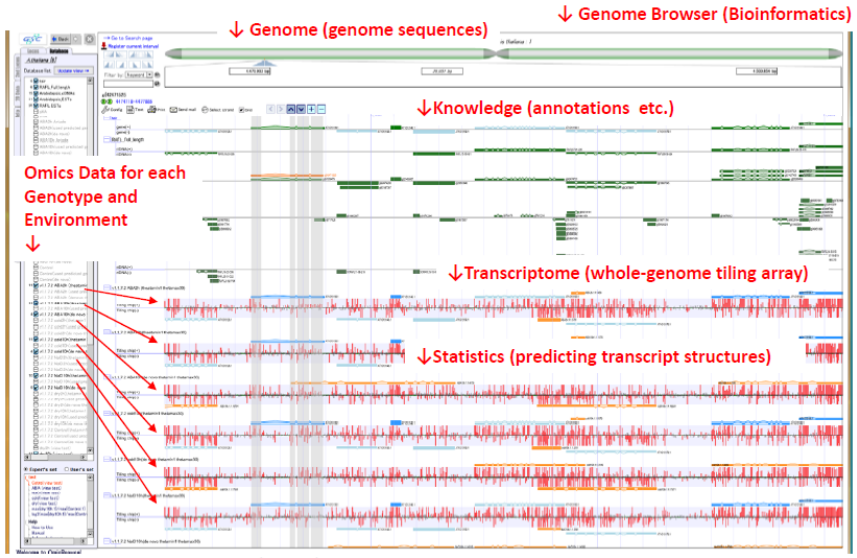
(1)現在のデータ量	シロイヌナズナ全ゲノムタイリングアレイを用いた発現プロファイル解析データ　19 種類
(2)データ区分	■自前　□第三者　□文献データ　■計算結果等の二次データ　□その他
(3)将来の増加の見込み	22年までには、シロイヌナズナ全ゲノムタイリングアレイを用いた発現解析実験を少なくとも100種類以上行う予定である。発現解析だけでなく、今後シロイヌナズナ全ゲノムタイリングアレイを用いてChIP-chipおよびmCIP-chip解析も行い、種々の条件におけるヒストン修飾やメチル化パターンも解析する予定である。また、Solexaシーケンサーを用いたヒストン修飾パターンの解析も行う予定である。
(4)権利関係	所有者（理研 PSC、植物ゲノム機能研究グループ） 公開（□可　□否　■その他［論文発表後データを公開する予定である。）
(5)その他、特記事項	

（１－３）．データの管理状況

(1)更新頻度等の管理状況、体制	実験データの取得に際して随時追加
(2)その他、特記事項	

（１－４）．データベース関係

(1)DB 管理者数	２名

(2)キュレータ・アノテータ数	１名
(3)データ構造	ファイル
(4)DB 管理ソフト	ファイルシステム
(5)サーバの OS	CentOS
(6)サーバ規模	２ノード
(7)DB へのアクセス数	未公開
(8)独立 IP 数	
(9)その他、特記事項	画面の様子　OmicBrowse の画面 

[3 件目]

（１－１）. データの種類

(1)生物種	シロイヌナズナ
(2)試料・ライブラリー等の種類、数	454 シーケンサーを用いた small RNA の大量解析データ 11 種類 454 シーケンサーを用いたシロイヌナズナの small RNA データは、これまでに 6 種類報告されている。
(3)測定方法	
(4)データの内容	454 シーケンサーを用いた small RNA の大量解析データ 11 種類 用いた植物材料：播種後 2 週間目の植物体を用いた乾燥、低温、塩ストレス、ABA 処理、無処理の植物体等
(5)その他、特記事項	

（１－２）. データソース

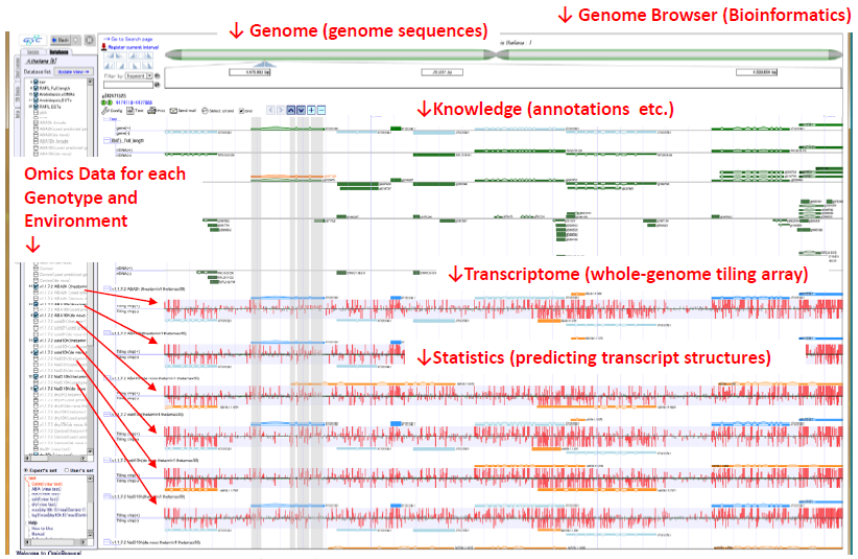
(1)現在のデータ量	454 シーケンサーを用いた small RNA の大量解析データ 11 種類
(2)データ区分	■自前　□第三者　□文献データ　■計算結果等の二次データ　□その他
(3)将来の増加の見込み	今後、454シーケンサーよりも大量の配列解析が可能なSolexaシーケンサーを用いてsmall RNAの大量解析を行う予定である。
(4)権利関係	所有者（理研 PSC、植物ゲノム機能研究グループ） 公開（□可　□否　■その他［論文発表後データを公開する予定である。）
(5)その他、特記事項	

（１－３）. データの管理状況

(1)更新頻度等の管理状況、体制	実験データの取得に際して随時追加
(2)その他、特記事項	

（１－４）. データベース関係

(1)DB 管理者数	１名
(2)キュレータ・アノテータ数	１名
(3)データ構造	ファイル
(4)DB 管理ソフト	ファイルシステム
(5)サーバの OS	CentOS

(6)サーバ規模	１ノード
(7)DB へのアクセス数	未公開
(8)独立 IP 数	
(9)その他、特記事項	画面の様子　OmicBrowse の画面 

[4 件目]メタボローム（代謝物質量プロファイル）

（１－１）. データの種類

(1)生物種	シロイヌナズナ
(2)試料・ライブラリー等の種類、数	✧　野生型および単一遺伝子欠損変異体およそ５０サンプルの網羅的な代謝物質量プロファイル ✧　物質の同定に用いる、標準物質のマススペクトルデータ 10,000 スペクトル（約 1000 物質）
(3)測定方法	ガスクロマトグラフィー質量分析計 液体クロマトグラフィー質量分析計 キャピラリー電気泳動質量分析計
(4)データの内容	質量分析計より出力されるマススペクトル
(5)その他、特記事	計測方法が確立しているため、シロイヌナズナの完全長 cDNA の過剰発現体または遺伝子欠損変異体約数百種類のデータ

項	を取得することも可能である。また各植物体について出来るだけ多く（数個体以上）のサンプルを計測することが望ましい。

（１－２）. データソース

(1)現在のデータ量	１サンプルあたり数百メガバイト
(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
(3)将来の増加の見込み	データ数、量を増やすことは容易だが、効率よくアノテーションする作業が困難
(4)権利関係	所有者（理研） 公開（■可　□否　□その他［　　］）
(5)その他、特記事項	全て公開できるが、アノテーション作業を解さない限り、データだけ外に出しても価値が低い

（１－３）. データの管理状況

(1)更新頻度等の管理状況、体制	データは新規代謝物が同定される度にアップデートする必要が生じる。現在は、計測グループが自前で維持管理
(2)その他、特記事項	管理するためのデータベースを現在構築中

[5 件目]フェノーム（トランスポゾン・タグライン）

（１－１）. データの種類

(1)生物種	シロイヌナズナ（*Arabidopsis thaliana*）
(2)試料・ライブラリー等の種類、数	シロイヌナズナのトランスポゾン・タグライン 18,000 系統と、全てのラインに関するトランスポゾン挿入位置情報。 シロイヌナズナ 26,000 遺伝子のうち 5,000 以上の遺伝子に関する変異を含んでいると推測される。
(3)測定方法	トランスポゾン挿入部位近傍の塩基配列の決定
(4)データの内容	変異体番号、トランスポゾン挿入位置情報、近傍遺伝子情報
(5)その他、特記事項	cDNA 情報や他の公共データベースとリンクさせている。

（１－２）. データソース

(1)現在のデータ量	トランスポゾン・タグライン 18,000 系統に関する情報
(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
(3)将来の増加の見込み	現在のところ予定はありません。

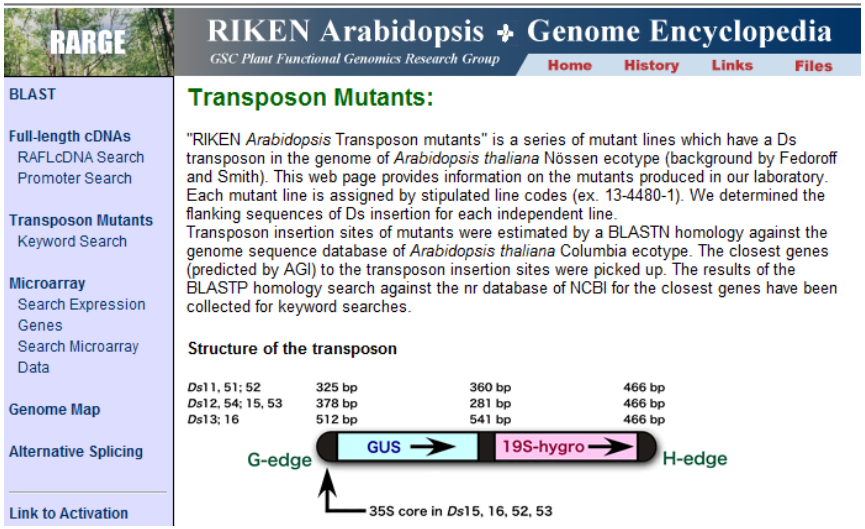
(4)権利関係	所有者（理化学研究所植物科学研究センター） 公開（■可　□否　□その他［　　］）
(5)その他、特記事項	http://rarge.gsc.riken.go.jp/

（１－３）．データの管理状況

(1)更新頻度等の管理状況、体制	理化学研究所植物科学研究センターのサーバーで維持されている。
(2)その他、特記事項	

（１－４）．データベース関係

(1)DB 管理者数	2 名
(2)キュレータ・アノテータ数	1 名
(3)データ構造	リレーショナルデータベース
(4)DB 管理ソフト	PostgreSQL
(5)サーバの OS	Red Hat Enterprise Linux ES
(6)サーバ規模	1 基
(7)DB へのアクセス数	データベース RARGE 総計　約 5000 ページ/月
(8)独立 IP 数	1
(9)その他、特記事項	画面の様子



[6件目]フェノーム（ノックアウト・表現型）

（１－１）. データの種類

(1)生物種	シロイヌナズナ（*Arabidopsis thaliana*）
(2)試料･ライブラリー等の種類、数	シロイヌナズナの4000遺伝子の変異体に関する表現型情報。シロイヌナズナ26,000遺伝子のうち4,000遺伝子に関する変異体を調べている。
(3)測定方法	系統的な表現型解析
(4)データの内容	変異体番号、トランスポゾン挿入位置情報、挿入変異遺伝子情報、表現型の画像データ
(5)その他、特記事項	他の公共データベースとリンクさせている。

（１－２）. データソース

(1)現在のデータ量	トランスポゾン・タグライン4000系統（4000遺伝子）の表現型に関する情報
(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
(3)将来の増加の見込み	現在は地上部形態異常に関して載せているが、他の表現型項目についてもデータを収集中である。
(4)権利関係	所有者（理化学研究所植物科学研究センター） 公開（■可　□否　□その他［　　］）
(5)その他、特記事項	http://rarge.gsc.riken.jp/phenome/

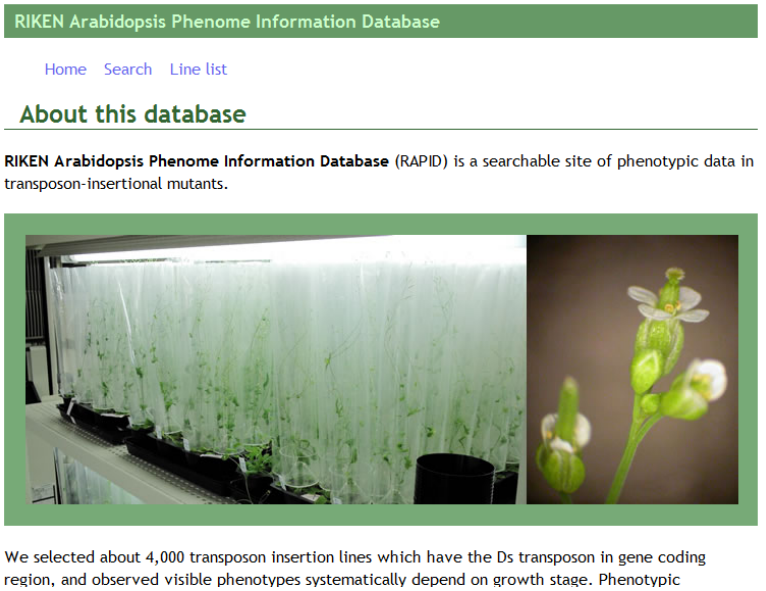
（１－３）. データの管理状況

(1)更新頻度等の管理状況、体制	理化学研究所植物科学研究センターのサーバで維持されている。
(2)その他、特記事項	

（１－４）. データベース関係

(1)DB管理者数	2名
(2)キュレータ・アノテータ数	1名
(3)データ構造	リレーショナルデータベース、ファイル
(4)DB管理ソフト	PostgreSQL、ファイルシステム
(5)サーバのOS	Red Hat Enterprise Linux ES
(6)サーバ規模	1基

(7)DB へのアクセス数	データベース RARGE 総計　約 5000 ページ/月
(8)独立 IP 数	1
(9)その他、特記事項	画面の様子



[7 件目]フェノーム（アクティベーション・タグライン）

（１－１）. データの種類

(1)生物種	シロイヌナズナ
(2)試料・ライブラリー等の種類、数	シロイヌナズナ完全長 cDNA 遺伝子高発現型変異体は、理研オリジナルの変異体であり、約 1 万の遺伝子リソースを網羅する。これは、現在報告されている遺伝子の 40%にあたる。 シロイヌナズナ Activation tagging 変異体系統は、7 万系統あり、シロイヌナズナのほぼすべての遺伝子の活性化をしている数と考えられる。
(3)測定方法	塩基配列決定による遺伝子情報 目視及び計測機器（光合成、色素吸収）による変異形質の情報
(4)データの内容	種子番号 遺伝子番号 遺伝子アノテーション情報 形質情報（光合成、色素、形態） 変異体画像情報
(5)その他、特記事項	完全長 cDNA 高発現変異体系統は、作成した 1 万系統についてすべて遺伝子情報を記載している。

（１－２）．データソース

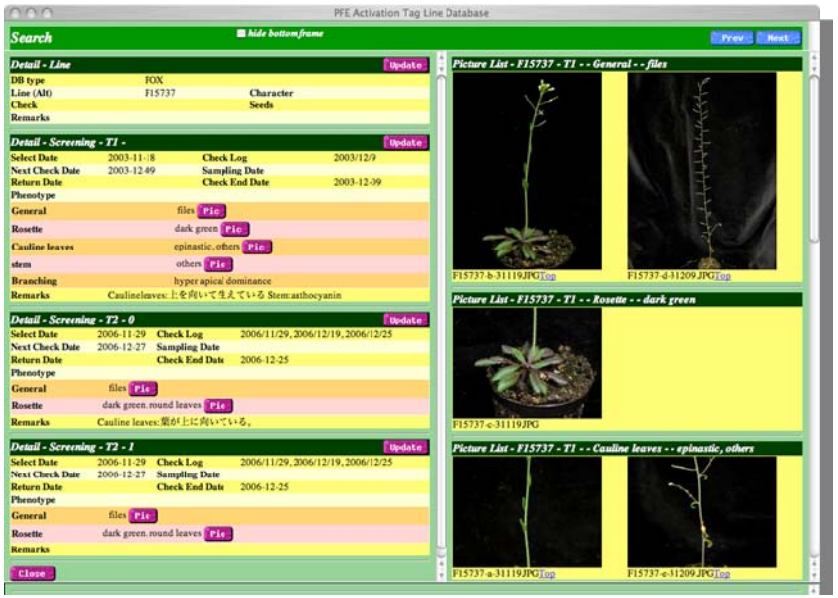
(1)現在のデータ量	シロイヌナズナ完全長 cDNA 高発現型変異体 1 万系統 Activation tagging 変異体 7 万系統
(2)データ区分	■自前　□第三者　□文献データ　■計算結果等の二次データ　□その他
(3)将来の増加の見込み	種々の形質計測により情報増加の可能性がある。
(4)権利関係	所有者（理研・NEC ソフト　　） 公開（■可　□否　□その他［　　］）
(5)その他、特記事項	http://amber.gsc.riken.jp/act/top.php

（１－３）．データの管理状況

(1)更新頻度等の管理状況、体制	1 週間に 1 回アノテーション情報の更新を行っている。
(2)その他、特記事項	

（１－４）．データベース関係

(1)DB 管理者数	1 名
(2)キュレータ・アノテータ数	1 名
(3)データ構造	RDB
(4)DB 管理ソフト	PostgreSQL
(5)サーバの OS	Red Hat Enterprise Linux ES（外部公開用） Red Hat Enterprise Linux AS（内部用）
(6)サーバ規模	2CPU(Pentium Xeon)　（外部公開用） 4CPU(Pentium Xeon)　（内部用）
(7)DB へのアクセス数	登録外部ユーザ数 884（2007 年登録 82、2006 年登録 142、2005 年登録 141、2004 年登録 327、2003 年登録 192）
(8)独立 IP 数	1（外部公開用）、1（内部用）
(9)その他、特記事項	画面の様子

[8 件目]リソース情報

（１－１）. データの種類

(1)生物種	シロイヌナズナ
(2)試料・ライブラリー等の種類、数	完全長 cDNA クローン(RAFL clone) シロイヌナズナ（エコタイプ：Columbia）のほぼ全ての転写領域をカバー
(3)測定方法	cDNA 塩基配列の全長もしくは両端を決定
(4)データの内容	※記録しているデータ項目（例えば、試料番号、遺伝子名、発現データ（画像）等） リソース番号、クローン番号、塩基配列のアクセッション番号、遺伝子コード領域の AGI 番号、塩基配列
(5)その他、特記事項	

（１－２）. データソース

(1)現在のデータ量	246,203 エントリー
(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
(3)将来の増加の見込み	あり
(4)権利関係	所有者（理研ゲノム科学総合研究センター）

	公開（■可　□否　□その他［　　］）
(5)その他、特記事項	http://www.brc.riken.jp/lab/epd/catalog/cdnaclone.html http://saber.epd.brc.riken.jp/sabre7/SABRE0101.cgi

（１－３）．データの管理状況

(1)更新頻度等の管理状況、体制	データの増加に応じて、年に数回までの更新あり 外部業者に運用を業務委託している
(2)その他、特記事項	

（１－４）．データベース関係

(1)DB 管理者数	理研側担当者２名、業者側数名
(2)キュレータ・アノテータ数	なし
(3)データ構造	テキストファイル、リレーショナルデータベース
(4)DB 管理ソフト	PostgreSQL
(5)サーバの OS	Linux
(6)サーバ規模	パーソナルコンピュータ
(7)DB へのアクセス数	計測値なし
(8)独立 IP 数	２
(9)その他、特記事項	画面の様子

[9件目]ストラクチュローム（高等動植物等由来）

（１－１）．データの種類

(1)生物種	動物、植物、微生物
(2)試料・ライブラリー等の種類、数	シロイヌナズナ：40程度、その他：2500程度
(3)測定方法	NMR（核磁気共鳴）や、X線による構造解析等
(4)データの内容	試料番号、タンパク質名、ドメイン名、PDBID、生物種、解析実験装置、発現系、試料の詳細（全長タンパク質の生物的意味、ドメインの機能、構造上の特性、基質結合ポケット・相互作用部位)、PDBに登録した蛋白質立体構造データ、構造決定のもととなった測定データ
(5)その他、特記事項	日本語記載

（１－２）．データソース

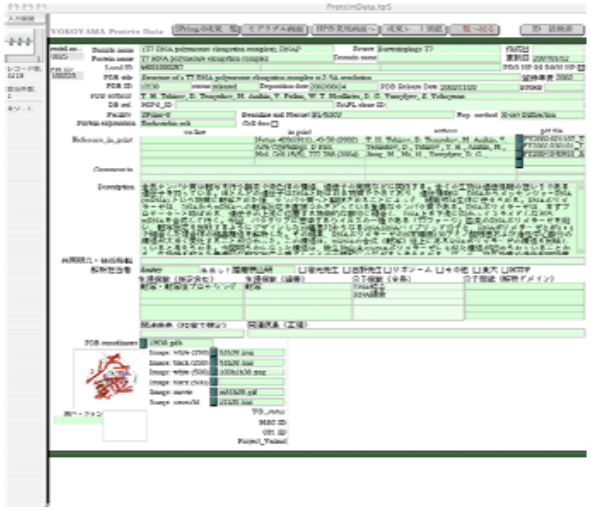
(1)現在のデータ量	シロイヌナズナ：40程度、その他：2500程度
(2)データ区分	■自前　□第三者　□文献データ　□計算結果等の二次データ　□その他
(3)将来の増加の見込み	レコードは随時アップデートで、レコード数は随時追加される。
(4)権利関係	所有者（独立行政法人理化学研究所） 公開（■可　□否　□その他［　　］）
(5)その他、特記事項	

（１－３）．データの管理状況

(1)更新頻度等の管理状況、体制	常時管理。
(2)その他、特記事項	

（１－４）．データベース関係

(1)DB管理者数	
(2)キュレータ・アノテータ数	
(3)データ構造	
(4)DB管理ソフト	ファイルメーカーPro

(5)サーバの OS	
(6)サーバ規模	
(7)DB へのアクセス数	
(8)独立 IP 数	
(9)その他、特記事項	画面の様子 

[10 件目]ストラクチュローム（微生物由来）

（１－１）. データの種類

(1)生物種	（試料調整・結晶化・回折実験データベース）９種類（*Thermus thermophilus* HB8, *Pyrococcus horikoshii* OT3, *Escherichia coli* K-12, *Aeropyrum pernix* K1, *Sulfolobus tokodaii* strain7, *Aquifex aeolicus* VF5, *Geobacillus kaustophilus* HTA426, *Thermotoga maritima* MSB8, *Methanocaldococcus jannaschii* DSM 2661） （変異体構造解析データベース）２種類（*Thermus thermophilus* HB8, *Pyrococcus horikoshii* OT3） （重原子データベース）不明だが多数
(2)試料・ライブラリー等の種類、数	（試料調整・結晶化・回折実験データベース）総レコード数 11190 のうち発現プラスミドがあるもの 11021（９８%）。回折画像数 300（変異体構造解析データベース）変異体総数 241 種類のうち、99 種類を構造決定済み（４１%）。*Pyrococcus horikoshii* OT3 由来 PH0725 蛋白質（265 残基）については、変異体 179 種類をプラスミド構築し 79 種類の結晶構造を決定済み。*Thermus thermophilus* HB8 由来 TTHB049 蛋白質（177 残基）については、変異体 62 種類をプラスミド構築し 20 種類を解析済み。両蛋白質を通じ、セレノメチオニン化のための Leu-Met 変異はほぼ網羅し

	ている。 （重原子データベース）重原子を結合した蛋白質の情報 784 件を収録する。特に水銀に関しては 351 件と豊富なデータを有する。現在のところ 22 種類の重原子をカバーしている。
(3)測定方法	（試料調整・結晶化・回折実験データベース）実験データベースであるため多種多様である。特に発現精製は多様な機器の出力データ等を含む。結晶化は主に結晶化ロボットＴＥＲＡの出力データである。 （変異体構造解析データベース）Ｘ線結晶構造解析により変異体蛋白質の結晶構造を決定。 （重原子データベース）重原子を結合した蛋白質の結晶構造をＸ線結晶構造解析により決定。
(4)データの内容	（試料調整・結晶化・回折実験データベース） （基礎情報） １．試料蛋白質の由来生物種名、遺伝子名、吸光係数、分子量、等電点など、いずれも計算結果等の二次データ。 ２．構築プラスミドデータ（自前のデータ）。ホスト、ベクター、予備発現結果（５段階の発現ランク）。 （発現精製） １．培養情報（自前のデータ）。発現データ（画像と５段階の発現ランク）、発現の諸条件（誘導状態、培養時間、培養温度、培養量溶液量等）。 ２．精製情報（自前のデータ）。各タンパクについて精製の諸条件（懸濁方法、超音波破砕方法、遠心時間、各タンパク質に最適なカラム情報）加えて精製タンパク質の吸収スペクトルの画像、精製蛋白質の SDS および Native ページ画像、精製蛋白質の収量等。精製タンパク質の DLS 測定結果（画像および数値）。選択カラム情報の詳細について：カラム名、緩衝液名、フラクションサイズ、フロー速度、グラジエント方法、溶出濃度、カラムチャート（A280、イオン強度等）の画像、各フラクションの SDS ページ画像。 （結晶化） 結晶化ロボットＴＥＲＡの出力（自前のデータ）。精製タンパク質の結晶化スクリーニング情報（結晶化条件、観察画像、１０段階のスコア）。 （回折実験） 回折実験データ（自前のデータ）。回折画像データとその計算処理のための各種パラメータ。 （変異体構造解析データベース）　二種のタンパク質について網羅的な変異体構造解析データベース。結晶化データ、回折実験データ（画像データ、分解能、測定条件等）、回折画像処理データ、精密化データ、構造座標データ（PDB）。 （重原子データベース）　自前の構造解析から得たデータ。さらに、文献データおよび登録されている PDB からの計算等による二次データ。内容は重原子実験データで、タンパク質名、重原子名、重原子試薬名、実験方法、沈澱剤名、緩衝液名、pH、文献名、重原子結合サイトの二次構造等。インターフェースとして重原子選択予測機能等。
(5)その他、特記事項	

（１－２）．データソース

(1)現在のデータ量	（試料調整・結晶化・回折実験データベース）基礎データ収録数 11190 件。構築プラスミドデータ 11021 件（変異体を含む)。培養情報 5878 件（重複あり)。精製情報 4401 件（重複あり)。結晶化情報 1300752 件、回折画像 300 件。容量は、培養精製情報 120GB、結晶化情報 1.3TB。回折実験情報 1.5TB。 （変異体構造解析データベース）サンプル数 241 件。回折データ 262 件。容量は 1.1TB。 （重原子データベース）784 件。
(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	３データベースとも増加の見込みあり。
(4)権利関係	所有者（独立行政法人理化学研究所） 公開（■可　□否　□その他［　　］）
(5)その他、特記事項	（試料調整・結晶化・回折実験データベース）内部利用で公開はしていない。 （変異体構造解析データベース）内部利用で公開はしていない。 （重原子データベース）HATODAS Ver1 はhttp://hatodas.harima.riken.go.jp/で公開。Ver 2 は内部利用。

（１－３）．データの管理状況

(1)更新頻度等の管理状況、体制	（試料調整・結晶化・回折実験データベース）実験生データのため毎日。 （変異体構造解析データベース）実験生データのため毎日。 （重原子データベース）不定期。
(2)その他、特記事項	

（１－４）．データベース関係

(1)DB 管理者数	１人（淺田）
(2)キュレータ・アノテータ数	キュレータ１人（淺田） アノテータ２人（国島・菅原） その他オペレーターとして実験補助２人程度を採用する予定
(3)データ構造	いずれもリレーショナルデータベース + Web ユーザーインターフェース （試料調整・結晶化・回折実験データベース）サーバ：Apache, Tomcat(JAVA)。バイナリーデータの場合、リンクのみデータベースに格納実体はディスクサーバーに格納 （変異体構造解析データベース）サーバ：Microsoft IIS (ASP.NET)。 （重原子データベース）サーバ：Apache (Ver 1:Perl CGI,Ver 2: Tomcat)。

(4)DB 管理ソフト	（試料調整・結晶化・回折実験データベース）Oracle （変異体構造解析データベース）SQL Server （重原子データベース）Postgres
(5)サーバの OS	（試料調整・結晶化・回折実験データベース）Linux （変異体構造解析データベース）Windows Server （重原子データベース）Linux
(6)サーバ規模	（試料調整・結晶化・回折実験データベース）PC４台 （変異体構造解析データベース）PC１台 （重原子データベース）PC１台
(7)DB へのアクセス数	（試料調整・結晶化・回折実験データベース）内部利用のため実験の進捗に依存 （変異体構造解析データベース）内部利用のため実験の進捗に依存 （重原子データベース）１１００検索/３年
(8)独立 IP 数	（試料調整・結晶化・回折実験データベース）内部利用のためグローバル IP は取得していない。 （変異体構造解析データベース）内部利用のためグローバル IP は取得していない。 （重原子データベース）１
(9)その他、特記事項	

（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（データの種類（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	トランスクリプトーム（タイリングアレイデータ）	未公開	公開整備は次年度につき該当なし
2	454 シーケンサーを用いた small RNA 解析データ	未公開	公開整備は次年度につき該当なし

（３）ＤＢ基盤システム、ツール等開発成果物の整備

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	植物統合 DB 計算機環境	未公開	**主機能・特徴点**＝データデポジトリ用のハードウェア環境の整備・４年間のデータ編纂作業に十分な容量のデータストレージの構築。 **進捗状況**＝アノテーション等のデータを安全に保持できるデータストレージの準備を行い、システム全体の運用体制を整えた。データストレージは４年間のデータ編纂作業に十分な容量を準備した。

			またサービス品質を上げるための工夫として災害時の運用維持対策を考慮し、理研の横浜研究所および和光研究所で同じ内容のストレージを設置し同期をとるようにシステムを構築した。 **今後の計画**＝アノテーションサーバとなる中心の計算機のソフトウェア環境を整備していく予定である。

（４）その他の成果物（（２）、（３）に該当しないもの）（該当なし）

補完課題実施機関（理化学研究所）　外部発表実績一覧

（１）セミナー、研究会等イベント開催

通番	タイトル	発表者（代表者）	年月日	開催場所	イベント名称	概要（対象者（層、参加人数）、出席者の主な反応等）
1	データベースの中に築く生物像 ―生命現象を読み解くためのデータベースと WEB リソース―	豊田哲郎、中村保一、青木考	2008 年 3 月 20 日	札幌コンベンションセンター	第 49 回日本植物生理学会年会シンポジウム	対象層＝植物研究者、 参加人数＝約 80 名 出席者の主な反応＝総合討論にて著作権など実務問題が議題に上がった。

（２）プレス発表、取材対応（該当なし）

（３）展示会等出展（該当なし）

（４）学会等への口頭発表

通番	タイトル	発表者	学会等の名称	年月日	特記事項
1	シロイヌナズナのオミックス統合解析とデータベース	豊田哲郎	理研シンポジウム「植物トランスクリプトーム解析の新展開：シロイヌナズナワークショップ 2007	2007 年 12 月 10 日	
2	シロイヌナズナのオミックス統合解析とデータベース	豊田哲郎	第 30 回日本分子生物学会年会・第 80 回日本生化学会大会（BMB2007）	2007 年 12 月 11-14 日	
3	シロイヌナズナ miRNA の in silico ターゲッ	神沼 英里 松井 章浩,	第 30 回日本分子生物学会年	2007 年 12 月 11-14 日	

	ト探索における特徴パラメータの決定木分析	栗原 志夫, 諸澤 妙子, 関 原明, 豊田 哲郎	会・第 80 回日本生化学会大会 (BMB2007)		
4	シロイヌナズナLucTagラインを用いた遺伝子発現定量解析	神沼 英里、吉積 毅,、栗山 朋子、越 智子、武藤 周、松井 南、豊田哲郎	人工知能学会第 36 回分子生物情報研究会(SIG-MBI)	2008 年 1 月 11 日	
5	セマンティックウェブ技術による理研データベースの統合化	豊田哲郎	「生命をはかる」研究会第 23 回研究会	2008 年 2 月 18 日	
6	デジタル whole mount in situ hybridization に向けたシロイヌナズナLucTagラインの遺伝子発現時系列画像の定量化	神沼 英里、吉積 毅,、栗山 朋子、越 智子、武藤 周、豊田哲郎、松井 南、	第 49 回日本植物生理学会年会(2008	2008 年 3 月 20-22 日	
7	シロイヌナズナのオミックス統合データベース：オミックス進化論とデータベースの役割	豊田哲郎	第 49 回日本植物生理学会年会(2008)	2008 年 3 月 20-22 日	

（５）雑誌等への論文寄稿

通番	タイトル	著者名	雑誌等の名称	掲載巻、号、ページ	特記事項
1	コンピュータの中の脳： 情報基盤の進化論	豊田 哲郎	生体の科学	4 巻 1 号、pp.20-32	

補完課題実施機関（産業技術総合研究所） 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	細菌、ウイルス、植物、線虫、ハエ、マウス、ラット、ヒト
②試料･ライブラリー等の種類、数	・**産総研・糖鎖関連データベース**(ヒト、マウス、ラット、ショウジョウバエの、線虫、植物(シロイネナズナ)、酵母糖鎖合成に関する 163 反応、文献で報告された基質の情報を 260 の反応、ヒトの糖鎖関連遺伝子：189 種類、マウスの糖鎖関連遺伝子：182 種類、ラットの糖鎖関連遺伝子：172 種類、ショウジョウバエの糖鎖関連遺伝子：74 種類、線虫の糖鎖関連遺伝子：63 種類、植物(シロイネナズナ)の糖鎖関連遺伝子：13 種類、酵母の糖鎖関連遺伝子：18 種類) ・**産総研・レクチンデータベース**（哺乳類、両生類、魚類、植物、細菌、ウイルスなど計２００種類ほどの生物、レクチン：266 種、相互作用データ：131 種） ・**産総研・糖タンパク質データベース**（線虫：プロテアーゼで消化したペプチドを 3 種のレクチン（conA、小麦胚芽アグ

	ルチニン（WGA）、線虫ガレクチン 6）で独立に捕集した糖ペプチドを試料とした。同定された糖タンパク質数は 829 種、糖鎖付加部位は 1465 カ所。ゲノム解析から予測される糖タンパク質のポテンシャル数は約 6,000 種なので、カバー率はおよそ 14%。他の研究者による報告は 100 種に満たない。マウス：脳、肝臓、肺、腎臓、精巣から変性条件下で抽出したタンパク質をトリプシン消化して得られた可溶性ペプチドを出発材料とした。3 ないし 5 種（conA、RCA120、AALの 3 種は共通に、肝臓についてはWGAとSSAも含めた）のレクチンカラムで捕集した糖ペプチドを試料とした。同定されたタンパク質は約 2,300 種、糖鎖付加部位は約 4,500 箇所。ポテンシャル糖タンパク質は約 10,000 種なので、カバー率は約 23%。実験データに基づく糖タンパク質としては約 200 種が既報。糖鎖付加部位が判明しているものは 100 種に満たない。） **・産総研・糖鎖スペクトルデータベース**（質量分析スペクトル 2756 スペクトル。糖鎖の種類としては、300 種類程度。）
③測定方法	**・産総研・糖鎖関連データベース** クローニングと酵素活性測定と独自に糖鎖を合成した反応経路（*in vitro*） **・産総研・レクチンデータベース** フロンタル･アフィニティークロマトグラフィーと呼ばれるレクチンをカラムに固定化し糖鎖とレクチンが相互作用すると遅れて溶出される現象を数値化 **・産総研・糖タンパク質データベース** 捕集した糖ペプチドを、安定同位体で標識した水（$H_2{}^{18}O$）中で N-グリカナーゼ処理することによって、糖鎖切除と同時に糖鎖付加位置を特異的に標識（IGOT）した後、LC-MS（質量分析）法で分析、同定した。 **・産総研・糖鎖スペクトルデータベース** リコンビナント酵素を用いて糖鎖を合成し、多段階タンデム質量分析（MALDI-QIT-TOF MS）で計測
④データの内容	**・産総研・糖鎖関連データベース** ドナーとアクセプター構造情報、酵素の情報、アクセプターにしなかった構造情報 **・産総研・レクチンデータベース** レクチン名、由来生物種名、由来器官、タンパク質ファミリー名、一次構造、立体構造、GenBank アクセッション番号、レクチン-糖鎖間相互作用情報 **・産総研・糖タンパク質データベース** 遺伝子名（NCBI など公共データベースの ID）、タンパク質名、糖鎖付加位置、部位ごとの結合したレクチンの種類および検出された組織、タンパク質の特性（分子量、等電点；共に計算値）と構造モチーフ（シグナル配列、膜貫通領域）。遺伝子オントロジー情報。 **・産総研・糖鎖スペクトルデータベース** 糖鎖 ID、糖鎖構造、質量分析スペクトル（画像イメージ）、プリカーサーイオンの *m/z* 値
⑤その他、特記事項	

（１－２）データソース

①現在のデータ量	「１－１）データの種類　②試料・ライブラリー等の種類、数」を参照のこと

②データ区分	**・産総研・糖鎖関連データベース** ■自前　□第三者　■文献データ　□計算結果等の二次データ　□その他 **・産総研・レクチンデータベース** ■自前　□第三者　■文献データ　□計算結果等の二次データ　□その他 **・産総研・糖タンパク質データベース** ■自前　■第三者　□文献データ　□計算結果等の二次データ　□その他 **・産総研・糖鎖スペクトルデータベース** ■自前　□第三者　□文献データ　□計算結果等の二次データ　□その他
③将来の増加の見込み	論文等で公表したデータを優先的に公開する。サプリメンタルデータも公開する。
④権利関係	所有者（　産業技術総合研究所と共同研究先の企業等　　） 公開（■可　□否　■その他［　整備終了後、順次公開　　］） http://jcggdb.jp（糖鎖ポータルサイト） http://riodb.ibase.aist.go.jp/rcmg/ggdb/　（糖鎖関連遺伝子データベース） 他準備が整い次第公開。
⑤その他、特記事項	－

（１－３）データの管理状況

①更新頻度等の管理状況、体制	３～４ヶ月に１度更新する。運営事務局が責任を持って更新する。
②その他、特記事項	－

（１－４）データベース関係

①DB 管理者数	プロジェクト専従雇用３名＋派遣１名、他産総研雇２名
②キュレータ･アノテータ数	10 名
③データ構造	RDBMS、XMLDB
④DB 管理ソフト	Oracle 11, eXist
⑤サーバの OS	Solaris 10
⑥サーバ規模	Web サーバ Fujitsu 社製 T5120
⑦DB へのアクセス数	公開している糖鎖関連遺伝子データベースに関して 2008 年１月～３月、約３０００アクセス
⑧独立 IP 数	不明
⑨その他、特記事項	－

（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（データの種類（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	**産総研・糖鎖関連データベース** http://riodb.ibase.aist.go.jp/rcmg/ggdb/	公開	2007 年 11 月 23 日　GGDB version2 をリニューアル。酵素の基質特異性の情報を配列データベースに統合。定期的に基質特異性の情報を追加。新しく同定された遺伝子情報を追加。2008 年１月～３月、約３０００アクセス。
2	**産総研・レクチンデータベース** http://jcggdb.jp からアクセス可能にする	未公開	構造同定のツールとして平成２０年度外部機関の DB と統合を行う。セキュリティチェック並びに対策を終えた後に公開。
3	**産総研・糖タンパク質データベース** http://jcggdb.jp からアクセス可能にする	未公開	NEDO プロジェクトでヒトのデータを採取中のためプロジェクト終了までに、ヒトのグライコプロテオミクスの糖鎖の付加位置の情報を公開する予定。 セキュリティチェック並びに対策を終えた後に線虫とマウスの情報を公開。
4	**産総研・糖鎖スペクトルデータベース** http://jcggdb.jp からアクセス可能にする	未公開	平成２０年度には構造同定方法論を統合しどの技術であればある糖鎖を同定できるか方法の選択と同定をサポートするシステムを構築する。セキュリティチェック並びに対策を終えた後に公開。

（３）ＤＢ基盤システム、ツール等開発成果物の整備

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	DB のスキーマ定義	未公開	DB を連携するにあたり共通する項目の標準化　　※開発途中
2	構造の定義・描画ツール	未公開	糖鎖構造 ID の共通化（XML,画像）　※開発途中
3	構造の呼び名の統一	未公開	Sialyl Lewis x 等の糖鎖構造に関する有名な名称と構造を辞書化　※開発途中

（３） その他の成果物（（２）、（３）に該当しないもの）（該当なし）

補完課題実施機関（産業技術総合研究所）　外部発表実績一覧

（１）セミナー、研究会等イベント開催（該当なし）

（２）プレス発表、取材対応

通番	タイトル	発表媒体	年月日	特記事項
1	特集レポート「日本の強み糖鎖科学オールジャパン体制へ」	日経 BP 社 BTJ ジャーナル	2008 年 3 月号	

2	タンパク質の機能左右、「糖鎖」総合データベース、産総研、今春ネット公開	日本経済新聞 朝刊、21 ページ	2008 年 1 月 21 日	

（３）展示会等出展（該当なし）

（４）学会等への口頭発表

通番	タイトル	発表者	学会等の名称	年月日	特記事項
1	日本糖鎖科学のポータルサイト	鹿内俊秀	第５回 糖鎖科学コンソーシアムシンポジウム(JCGG)	2008 年 11 月 27 日	
2	「糖鎖データベースの紹介」と「糖鎖産業技術フォーラムの具体的取組」	新間陽一	糖鎖産業技術フォーラム(GLIT)設立総会＆ 第１回糖鎖産業技術フォーラム	2008 年 1 月 23 日	

（５）雑誌等への論文寄稿（該当なし）

補完課題実施機関（国立遺伝学研究所） 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	多数
②試料・ライブラリー等の種類、数	多数
③測定方法	従来型シーケンサーからの出力、一部、第２世代シーケンサーからのもの少数。
④データの内容	塩基配列決定時の波形データ、結果の配列データ、その生物種、目的分類（WGS、ゲノム、等）、他
⑤その他、特記事項	

（１－２）データソース

①現在のデータ量	2 生物種、2,542,476 件
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
③将来の増加の見込み	サービス開始後、急速に増加の見通し。サービス後１年程度で１TB 以上を見込み
④権利関係	所有者（それぞれのデータ登録者に帰属）

	公開（■可　□否　□その他［　　］）
⑤その他、特記事項	

（１－３）データの管理状況

①更新頻度等の管理状況、体制	現在は開発・試験フェーズのため必要に応じて更新。運用開始後は、登録要求に応じて随時更新。
②その他、特記事項	

（１－４）データベース関係（該当なし）

（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（データの種類（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	トレースデータベース	未公開	塩基配列決定の 1 次データとして産生されるトレースデータのデータベース。ウェブサイト、またはウェブ API からの検索サービスを予定。現在、データベース自体は最初の試験実装段階にあるが、ウェブインタフェースは未実装。
2	トレース付随データ	未公開	FTP による公開を予定。トレースデータ自体をこれに含めるかに議論があり検討中。

（３）ＤＢ基盤システム、ツール等開発成果物の整備

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	トレース登録データチェックツール	未公開	トレースアーカイブサービスへのデータ登録に際して、データの文法的・意味的なチェックを行うためのツール。現在は内部試験使用中。将来的には登録希望者への提供を予定。

（４）その他の成果物（（２）、（３）に該当しないもの）（該当なし）

補完課題実施機関（国立遺伝学研究所）　外部発表実績（該当なし）

補完課題実施機関（九州工業大学）　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	複数
②試料・ライブラリー等の種類、数	①蛋白質およびその変異体の構造安定性に関する熱力学データ約 22,000 件 ②蛋白質と核酸の相互作用の定量的な熱力学実験データ約 8,000 件
③測定方法	熱測定、分光測定など
④データの内容	書誌情報、数値情報、画像情報。内容の詳細は別紙参考資料参照。
⑤その他、特記事項	なし。

（１－２）データソース

①現在のデータ量	①蛋白質およびその変異体の構造安定性に関する熱力学データは、22,000 件以上を文献から収集。 ②蛋白質と核酸の相互作用に関する熱力学データは、8,000 件余りを収集。
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	蛋白質の安定性および蛋白質・核酸相互作用の熱力学データともに、年間 2,000～3,000 件程度の新規データが発生。
④権利関係	所有者（データベース作成者） 公開（■可　□否　□その他［　　］）
⑤その他、特記事項	**①ProTherm: http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html** **②ProNIT: http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html**

（１－３）データの管理状況

①更新頻度等の管理状況、体制	月に１回程度更新。データベースサーバーは無休でアクセス可能。
②その他、特記事項	なし。

（１－４）データベース関係

①DB 管理者数	１名
②キュレータ･アノテータ数	２名
③データ構造	リレーショナル

④DB 管理ソフト	SYBASE
⑤サーバの OS	Linux
⑥サーバ規模	ワークステーション
⑦DB へのアクセス数	年間約１０万件
⑧独立 IP 数	約１万個
⑨その他、特記事項	DB の検索メニューの画面コピーは別紙参考資料添付。 オントロジーは今後整備。

（２）データ（又はＤＢ）の連結、統合化整備（該当なし）
（３）ＤＢ基盤システム、ツール等開発成果物の整備（該当なし）
（４）その他の成果物（（２）、（３）に該当しないもの）（該当なし）

補完課題実施機関（九州工業大学）　外部発表実績一覧

（１）セミナー、研究会等イベント開催（該当なし）
（２）プレス発表、取材対応（該当なし）
（３）展示会等出展（該当なし）

（４）学会等への口頭発表

通番	タイトル	発表者	学会等の名称	年月日	特記事項
1	**生体分子間相互作用の熱力学データベースと解析**	皿井明倫	生物物理学会	2007 年 12 月 21 日	

（５）雑誌等への論文寄稿

通番	タイトル	著者名	雑誌等の名称	掲載巻、号、ページ	特記事項
1	**Thermodynamic Database for Proteins: Features and Applications**	M. Michael Gromiha and Akinori Sarai	Methods in Molecular Biology	印刷中	