区　　分	□中核機関（□代表機関／□参画機関） □分担機関（□代表機関／□参画機関） ■補完課題実施機関
課　題　名	植物オミックス情報および蛋白質構造情報
実施機関名	独立行政法人理化学研究所
代表研究者名	豊田哲郎

１．課題開始時における達成目標

１．シロイヌナズナオミックス情報の注釈付けと公開

シロイヌナズナの発現、表現型、リソースに関するデータベースを標準的なオントロジーや ID に基づくアノテーションをつけて公開し、XML やテーブル形式でのダウンロードを可能にする。

２．高等動植物等由来蛋白質構造データと実験データの注釈付けと公開

タンパク 3000 プロジェクトで解明された高等動植物由来の蛋白質構造データに付随する実験データに標準的なオントロジーや ID に基づくアノテーションをつけて XML やテーブル形式でのダウンロードを可能にする。

３．微生物由来蛋白質構造データと実験データの注釈付けと公開

理研播磨研究所における微生物由来の蛋白質構造データに付随する実験データに標準的なオントロジーや ID に基づくアノテーションをつけて XML やテーブル形式でのダウンロードを可能にする。

４．理化学研究所のデータベース統合化のためのモデルケース構築

理化学研究所の DB 群を対象とする横断検索を可能にするための理研側のハブサイトを構築し、シロイヌナズナと蛋白質立体構造情報をモデルケースにしながら、他の DB にも横断検索の対象を広げていく（10～100 DB）。

２．平成２２年１０月末時点における事業計画に対する成果

（１）成果概要

１．シロイヌナズナオミックス情報の注釈付けと公開

理研 BASE で開発したデータベース統合システムである理研サイネス上で公開されたシロイヌナズナ変異体データベース（Fox-hunting, *Ds* transposon, Activation tagging line の３データベース）で扱われている表現型変異情報を対象に、2 種類のオントロジー（Plant Ontology, the Ontology of Phenotypic Qualities）を用いた新しい統一的なアノテーションを行なった。オントロジーを用いた表現型の標準化を試み 115 件の表現型に纏めた。これらの表現型と対応するオントロジー、既存の変異体データベースの観察データへのリンクを収めたデータベースを開発し公開した(http://scinets.org/item/cria143u1i)。 本データベースから前述の 3 データベースに対して設定されたリンクの数は、それぞれ 13142 件,221 件,1268 件となる。さらに文献サーベイによるシロイヌナズナフェノーム情報の収集と標準化を進めた。2010 年 10 月までに 266 件の文献をキュレーションし、438 種類の表現型と 21 種類の実験条件を 6 種類のオントロジーを用いて標準化した。これらの文献、表現型、実験条件、関連遺伝子の情報をデータベース化し公開した（http://scinets.org/item/ria224i)。

２．高等動植物等由来タンパク質構造データと実験データの注釈付けと公開

タンパク 3000 プロジェクトで解明された高等動植物等由来のタンパク質構造データ(X 線結晶構造解析)のうち、累計 3 万 1 千件の回折実験データ（画像データ枚数）を公開し、それに関わる実験情報についても関連づけを完了させ、9 月末までに可能な約 50 万件の結晶観察データについても提供した。

３．微生物由来蛋白質構造データと実験データの注釈付けと公開

平成２１年７月の公開分は、微生物由来蛋白質（変異導入蛋白質を含む）に関わる実験データ（試料調製データ（発現プラスミド構築実験 10,000 件、培養実験 5000 件、精製実験 3000 件)、結晶化実験データ（観察 1000 万件)、200 件の回折実験データ（データセット数))、および重原子導入蛋白質に関わるデータ 500 件である。平成２２年７月公開の微生物由来蛋白質オリジナルデータベース Bacpedia は、実験データを有するタンパク質約 9000 種類を含む。平成２２年１０月の追加公開分は、微生物由来蛋白質（変異導入蛋白質を含む）に関わる実験データ（試料調製データ（発現プラスミド構築実験 1800 件、培養実験 2700 件、精製実験 600 件)、結晶化実験データ（観察 250 万件)、回折実験データ 500 件（データセット数))、および重原子導入蛋白質に関わるデータ 2000 件である。

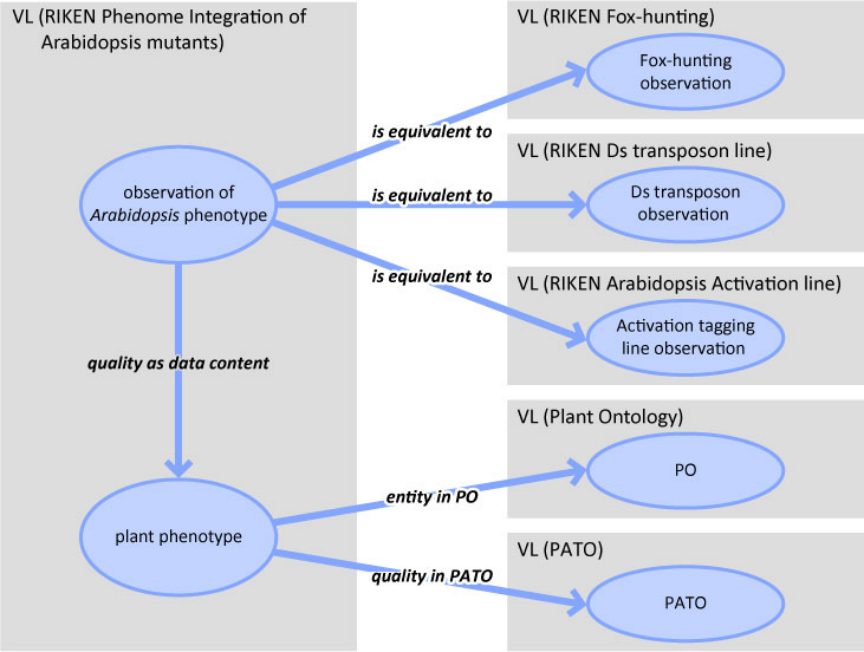
４．理化学研究所のデータベース統合化のためのモデルケース構築

ハブサイトを実現するため、大容量データを安全に保持できるようにデータベースを運用した。特にデータの増大化に伴い、ストレージ系システムのエラーが増加してくるため、これに迅速に対応する体制を整えて運用した。また、公開したデータをクライアントサイドプログラムからウェブ経由で利用できる API である Semantic-JSON の仕様を決定して公開した。植物オミックス統合データベースをサイネス上に構築し（http://scinets.org/db/plant)、この下にシロイヌナズナのゲノム・トランスクリプトーム・プロテオーム・フェノーム関係データベースを含む 29 件の植物関係のデータベース（うち 17 件が理研による研究活動の成果に基づくデータベースである）を統合し、公開した（添付書類１参照)。

（２）進捗及び成果

１．シロイヌナズナオミックス情報の注釈付けと公開

理研サイネス上に統合されたシロイヌナズナ変異体に関する３データベース（Fox-hunting, *Ds* transposon, Activation tagging line）は理研での実験で得られた各変異体の表現型情報を扱っているが、それらはオリジナルデータの作成者によって独自の流儀で表現されていた。本事業では、これらを対象に新しい統一的なアノテーションを行ない公開した。これは、表現型を「該当器官名＋形質状態」として捉え、既存の公開オントロジーである Plant Ontology(PO)および the Ontology of Phenotypic Qualities (PATO)を用いて統一的に再定義しようとするものである（添付書類 2 参照)。この方法によって前述の３データベース中に記述されている表現型情報を整理すると全部で 115 種類となった。この 115 種類を新しいデータベース「シロイヌナズナ変異体情報」に収納し、Fox-hunting, *Ds* transposon, Activation tagging line の各 DB に保存されている既存の観察情報に対して、13142 件,221 件,1268 件のリンクを設定して 2010 年 9 月に公開した（下図参照)。ここまでで、明確な表現型情報が公開されているレコード全てについて新しいアノテーション情報が関連付けられた。



さらに世界中で行われているシロイヌナズナのフェノーム情報を収集し、表現型情報の統合を図った。具体的には、シロイヌナズナ変異体の表現型情報を扱った論文 266 件を対象に、マニュアルキュレーションを実施し、表現型変異・その変異を観察した実験条件・関与が考えられる遺伝子などの情報を抽出した。抽出した表現型情報を 6 種類の公開オントロジー（PATO, PO, Gene Ontology (GO), Plant Trait Ontology (TO), Plant Environmental Ontology (EO), and Chemical Entities of Biological Interest (ChEBI)）に基いて標準化し、438 種類に整理した。また、対象論文に示された実験条件について 2 種類のオントロジー（ChEBI、EO）を用いて 21 種類に標準化した。さらに、それぞれの表現型に関連する遺伝子（AGI コード）との関連付けを行った。以上の情報を Plant phenome データベースとして公開した（http://scinets.org/item/cria224u1i）。

2．高等動植物等由来タンパク質構造データと実験データの注釈付けと公開

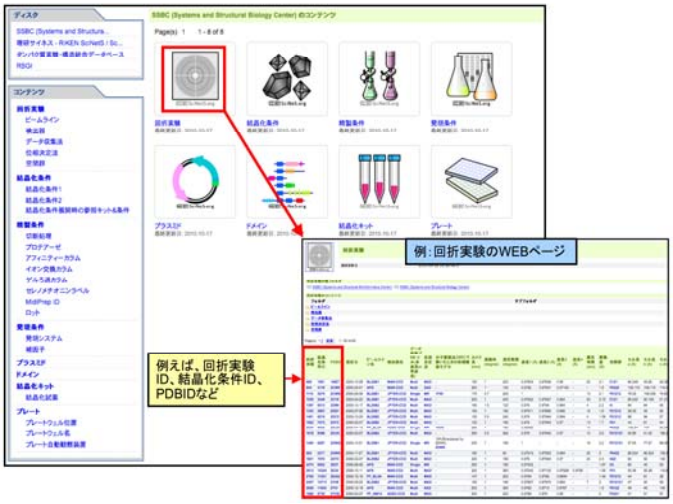
・2009 年 5 月：回折実験データ 2 万件と付随する実験データを理研サイネス上に公開した。

・2010 年 8 月：回折画像データ 1 万 1 千件と付随する実験データを理研サイネス上に公開した。

・2010 年 9 月：約 50 万件の結晶観察データと付随する実験データを提供した。

・2010 年 10 月：上記の回折実験データ、結晶観察データ、付随する実験データをすべて DBCLS へ提供した。

上記の回折実験データおよび付随する実験データを理研サイネス(http://scinets.org/db/ssbc)からダウンロード可能とした。（結晶観察データ約 50 万件をテストサーバで編集中。）



３．微生物由来蛋白質構造データと実験データの注釈付けと公開

平成２１年７月に、微生物由来蛋白質に関わる実験データ、変異導入蛋白質に関わる実験データ、および重原子導入蛋白質に関わるデータについて、アノテーションシステムから第一回の公開を行った。アノテーションデータの公開に合わせてプレス発表を行った。平成２１年９月、国際的な合意を得るため、台湾の３カ所で海外デモンストレーションを実施した。平成２２年７月に微生物由来蛋白質のオリジナルデータベース Bacpedia を公開し、一般の研究者にとって使いやすい環境を整えた。平成２２年１０月に追加データをアノテーションシステムから公開した。

４．理化学研究所のデータベース統合化のためのモデルケース構築

ハブサイト構築のために以下のようなシステム整備を行なった。

まず、平成２１年３月までに、平成１９年度に導入した理研和光研究所をバックアップとする大容量ストレージを安定的に運用する為の体制整備を行った。具体的には１０テラバイトを超える播磨研の X 線結晶構造解析データを順次システムに導入し、ミラーリング等で不具合が生じた部分を修正し、システムの安定的なバックアップ運用を行った。中核機関が設定したメタデータの仕様に従ったデータ変換結果として、アノテーションシステムに登録している成果データベースのメタデータを作成した。横浜研究所と和光研究所にはそれぞれ約 200TB の容量を持つストレージマシン群が備わっている。毎週末に横浜研究所側ストレージに追加されたデータが和光研究所側ストレージに複製する機能を実装した。この機能により遠隔地間で大規模データのバックアップが実現され、データを損失させることなくシステムを連続して安定運用させることが可能となった。

また、公開したデータをクライアントサイドプログラムからウェブ経由で利用できる API として、Semantic-JSON を構築、公開した。この API は、ユーザがデータリクエストを URI で指定し、その結果を JSON 形式データとして返すもので、様々なプログラミング言語で利用が可能となった。これにより、これまで構築したアノテーションシステムに、クライアントサイドのウエブブラウザから理研内外を問わず誰でもプログラムを書いてデータベースを利用できる機能が実現した。

高速かつ精密に表現型情報等をキーとした検索が実行できるようシステム間の連携を組むことで、結果もわかりやすく表示されるよう工夫した。トップページから遺伝子を検索する場合、遺伝子データにキーワードが含まれない場合であっても遺伝子に関連付けられているキーワードを含む表現型データを見出し、当該遺伝子を推論し検索するエンジンにクエリーを投げて表示できるようになった。検索結果の表示画面では、ヒットのあったデータベースそれぞれのデータ総件数とヒット件数を表示し、概略

を簡単に把握できるようにした（添付書類 3 参照）。また検索速度については、データアクセス権のチェックも含め、たいていの場合高々1 秒で結果が表示される。さらには、JSONP を介したキーワードを含むデータアイテム数を返すインターフェイスを公開し、2010 年 9 月までに統合データベースの横断検索機能との連携を可能にした。

2010 年 10 月までに、理研での研究活動に基づく植物オーミックスのデータベース群をサイネス上に順次統合した。現時点で、シロイヌナズナのゲノム・トランスクリプトーム・プロテオーム・フェノーム関係データベースを含む 29 件の植物関係のデータベースが存在し、そのうち 17 件が理研による研究活動の成果に基づくデータベースである（添付書類 1 参照）。それらを束ねる目的で「植物オミックス統合データベース」を構築し公開した（http://scinets.org/db/plant）。植物オミックス統合データベースでは、各種オミックス間の関係を定義すべく新規に開発された「上位オントロジー」（添付書類 4 参照）に準拠して、統合対象データベースを分類し直した。それぞれのトップページに示すメニューにはこの分類体系を反映させている。

植物だけではなく哺乳類を含む他の生物種のフェノーム情報も視野に入れた表現型情報の統合を目指し、表現型・遺伝子型・実験条件・対応する文献などの関連情報を網羅するテンプレート（セマンティックウェブ構築のための枠組み。添付書類 5 参照）を作成した。前述したシロイヌナズナ変異体の表現型データベースもこのテンプレートに沿って作成した。前述の検索システムは、このように定義された意味リンクを考慮して検索を行なうので、対象生物種や実験手法等の違いに関わらず、所定の表現型について関連するオミックス情報をデータベース横断的に検索できるようになった。

３．当初目標に対する達成度

１．シロイヌナズナオミックス情報の注釈付けと公開

サイネス上で公開中のシロイヌナズナ変異体データベースのフェノーム情報について、オントロジーを用いたシロイヌナズナ表現型情報を標準化を行ない、新たな注釈づけを全ての公開データについて完了した。また、文献を対象としたフェノーム情報の標準化も進めた。これらを全てデータベース化して公開し、ダウンロード可能とした。これをもって当初目標は達成できた。

２．高等動植物等由来タンパク質構造データと実験データの注釈付けと公開

当初目標を達成した。

３．微生物由来蛋白質構造データと実験データの注釈付けと公開

公開可能なデータは全て公開した。

４．理化学研究所のデータベース統合化のためのモデルケース構築

基盤システム整備、植物統合データベース・タンパク質統合データベースの開発、検索システムの開発に成功し、データベース間の横断検索を実現した。これを持って当初目標は達成できた。

４．中間評価に対する対応

中間評価においては「分野毎の国内他機関との連携の構築」および「理化学研究所内のあらゆるデータの積極的な公開」をさらに推進するよう求められた。我々は、シロイヌナズナのオミックスデータのかずさ DNA 研究所への全面的提供、哺乳類統合データベース開発チームとの連携によるフェノーム情

報の統合化など、必要な連携を積極的に進めた。また理研から公開されている研究成果のサイネスへの統合化を積極的に進め、植物オミックス統合データベース、タンパク質統合データベース等のコンテンツの充実に努めた。

５．他機関との連携

サイネス上で公開されているシロイヌナズナ関係の全データをＲＤＦ形式に変換して統合データベースセンター（かずさＤＮＡ研究所のグループ）に提供した。

６．今後の見通し、計画、展望

１．シロイヌナズナオミックス情報の注釈付けと公開

フェノーム情報は Multinational Arabidopsis Steering Committee (MASC)の 2010 年の年次報告でもその蓄積が重点課題の一つとして掲げられている通り、現在のシロイヌナズナ研究コミュニティが最も興味を持っている研究対象の一つである。これを踏まえて、フェノーム情報の統合化は今後特に力を入れて推進する必要がある。シロイヌナズナのフェノーム情報を扱った文献は多数あり、今後さらに多くの文献をキュレーションして、データベースの内容の充実を図る。

２．高等動植物等由来タンパク質構造データと実験データの注釈付けと公開

高等動植物等由来タンパク質の回折実験データ、結晶観察データ、および付随する実験データは、一般の生命科学者のタンパク質研究を大いに支援すると期待される。

３．微生物由来蛋白質構造データと実験データの注釈付けと公開

微生物由来タンパク質の結晶構造解析実験データベースは、タンパク質の効率的な構造決定のためのソフトウェア開発などを促進し、また、このデータベースに登録された世界最大規模の類似タンパク質情報により、一般の生命科学者のタンパク質研究を大いに支援すると期待される。変異体タンパク質の結晶構造解析実験データベースは、均一な条件で結晶化されるなど相互比較を行う上で有利な特長があるため、高精度ホモロジーモデリングの開発など、バイオインフォマティクス分野での活用が期待される。重原子標識タンパク質の結晶構造解析実験データベースは、重原子を結合するモチーフ配列を利用して計画的に重原子標識をするなど、タンパク質工学分野への応用が期待される。

４．理化学研究所のデータベース統合化のためのモデルケース構築

サイネスを生命科学研究のためのバーチャルな統合環境の基盤と位置づけ、理研のみならず幅広い分野の科学研究コミュニティと連携が図れるように、プログラミング環境やラボノート機能などを含むシステムの整備を継続していく。

７．全体総括

１．シロイヌナズナオミックス情報の注釈付けと公開

本プロジェクトのもとで植物関連の幅広いオミックス情報が統合でき、特にフェノーム情報の標準化とデータベース化と関連オミックス情報へのリンク付けが実現したことは、遺伝子機能解明などの基礎分野から、バイオ燃料開発や創薬などの応用分野まで、幅広く植物科学研究に資すると考えている。これまで、オントロジーを活用した植物分野でのフェノーム情報の統合化に本格的に取り組んだ例は世界になく、本プロジェクトでの成果は今後の植物フェノーム研究を先導するものになる。

２．高等動植物等由来タンパク質構造データと実験データの注釈付けと公開

今回、タンパク 3000 プロジェクト関連の膨大なタンパク質実験データを公開したことは、大型プロジェクトの社会還元として大きな意味がある。今回のタンパク質実験データベースは、理研生命情報基盤研究部門が開発したアノテーションシステム「理研サイネス」（http://scinets.org）から公開した。このアノテーションシステムでは、各データベースがセマンティックウェブと呼ばれる国際標準形式で再構築されているため、データの再利用や自動処理化が容易である。

３．微生物由来蛋白質構造データと実験データの注釈付けと公開　（担当者：国島直樹）

理研放射光科学総合研究センターは文部科学省「タンパク 3000 プロジェクト（2002 年度～2006 年度)」に参画し、タンパク質結晶構造解析研究グループと放射光システム生物学研究グループが中心となって、主に微生物由来タンパク質の結晶構造を SPring-8 の世界最高輝度を誇る X 線を用いて集中的に決定した。今回、放射光科学総合研究センターに存在するタンパク 3000 プロジェクト関連の膨大なタンパク質実験データ（技術開発関連データを含む公開可能なもの全て）を公開したことは、大型プロジェクトの社会還元として大きな意味がある。今回のタンパク質実験データベースは、理研生命情報基盤研究部門が開発したアノテーションシステム「理研サイネス」（http://scinets.org）から公開した。このアノテーションシステムでは、各データベースがセマンティックウェブと呼ばれる国際標準形式で再構築されているため、データの再利用や自動処理化が容易である。また、今回公開した 3 種類のデータベースは、どれも、未加工の実験データまでさかのぼって利用できるという特長を持っている。従って、データベースを一括ダウンロードし、ほかのデータと組み合わせるなどして新たなデータベースを構築することが可能になり、大規模で予想外の展開が期待できる。

４．理化学研究所のデータベース統合化のためのモデルケース構築

一般にオミックス研究では大量の実験データを対象とした情報処理が必要であり、各種データを一堂に集めた場が不可欠である。しかし、多くの研究者や研究機関にとって、それぞれの研究成果をデータベース化して公開・維持していくことは技術的にも金銭的にも容易ではなく、シロイヌナズナ研究の分野で長年中心的な役割を果たしてきた TAIR のデータベースでさえその存続が危ぶまれる事態となっている。我々はサイネスの構築によって、対象生物種やデータの種類にかかわらず、多様かつ大量のデータを単一のプラットフォームのもとに集積・公開し、低コストで維持する方法を見出した。本事業を通じて、理研において実施されてきた植物オミックスの分野での多くの先進的な研究成果がオントロジーを活用してサイネス上に統合され、データベース横断的な高度な検索が可能になったことは、今後のオミックス研究の発展に広く寄与するものである。

８．特記事項

１．シロイヌナズナオミックス情報の注釈付けと公開

本事業は、PO,PATO 等のオントロジーを活用した植物分野でのフェノーム情報の標準化に初めて本格的に取り組んだ試みである。この成果公開化にいち早く踏み切ったことは、今後の植物フェノーム研究を先導しうるものとして意義が大きい。

２．高等動植物等由来タンパク質構造データと実験データの注釈付けと公開

回折実験データ、結晶観察データ、および付随する実験データは相互に関連付けられているので、 類

似タンパク質の構造機能解析などの参考データとして役立つ。実験データをタンパク質研究の参考情報や、新たな方法論開発に有効活用できる。データの再利用や自動処理化が容易なセマンティックウェブ形式で共有化（当分野で世界初）した。

３．微生物由来蛋白質構造データと実験データの注釈付けと公開

SPring-8 の高輝度 X 線を使ったタンパク質結晶構造解析の膨大な実験データを集積した。

実験データをタンパク質研究の参考情報や、新たな方法論開発に有効活用できる。

データの再利用や自動処理化が容易なセマンティックウェブ形式で共有化（当分野で世界初）した。

４．理化学研究所のデータベース統合化のためのモデルケース構築

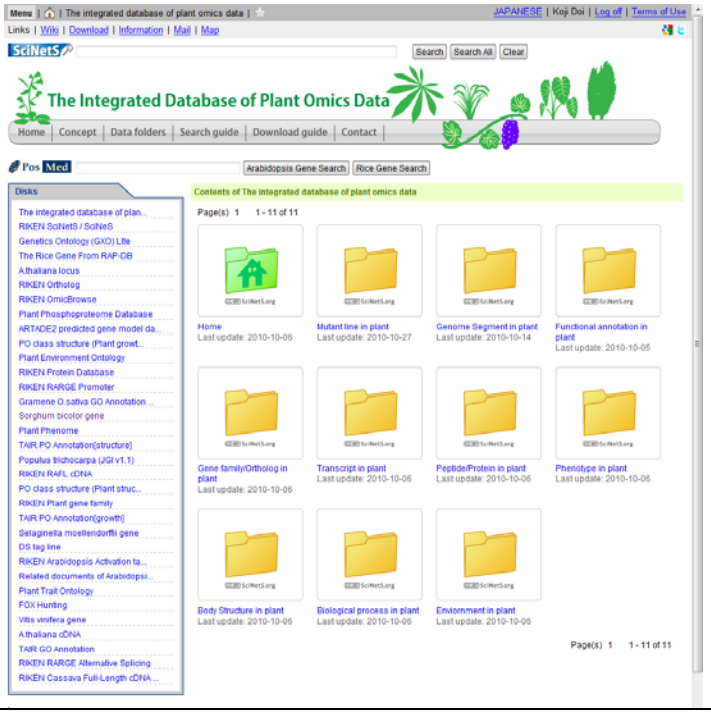
データベース統合化に当たっては、オミックスデータ間の関係を「上位オントロジー」に沿って整理し、フェノーム情報の統合も、植物・動物の分類群の壁を越えて共通のオントロジーでの標準化を進める方針をとった。サイネス上の検索システムはこの特性を生かして検索を実現するので、生物種やデータ種別の違いを超えて、意味的に正しい検索結果が精度よく得られるようになった。生物科学の分野において、セマンティックウェブの特性をこのように生かした統合データベースを構築した前例はなく、先進的な成果が得られたといえる。

９．委託研究費一覧

	１８年度	１９年度	２０年度	２１年度	２２年度	計
設備備品費（千円）		65,088	0	0	0	65,088
人件費（千円）		10,140	44,059	51,100	39,195	144,494
業務実施費（千円）		15,680	19,567	12,535	18,077	65,859
一般管理費（千円）		9,090	6,363	6,363	5,727	27,543
合　計　（千円）		100,000	70,000	70,000	63,000	303,000

整備実績一覧

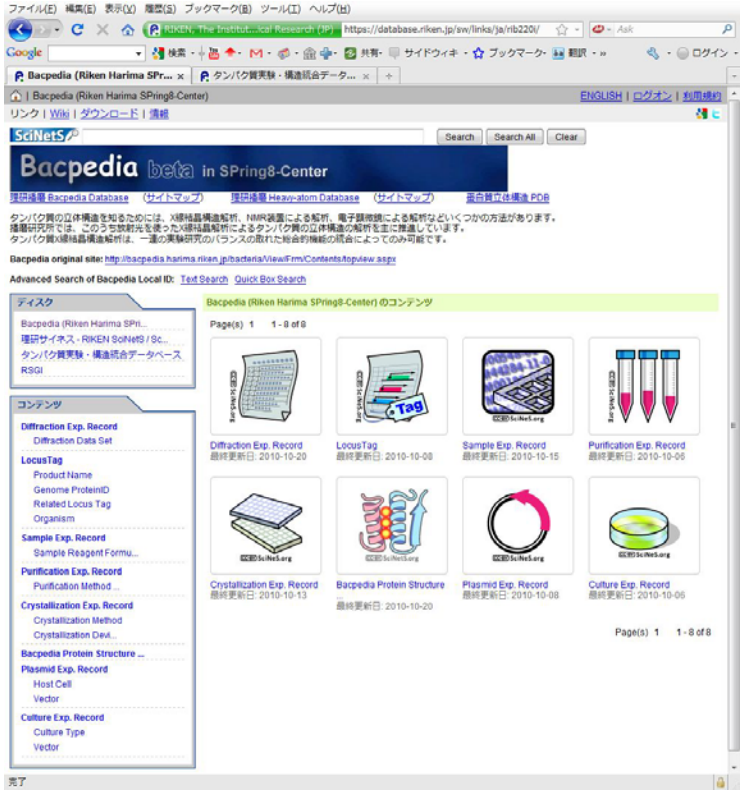
（1）データ（又はＤＢ）の連結、統合化整備

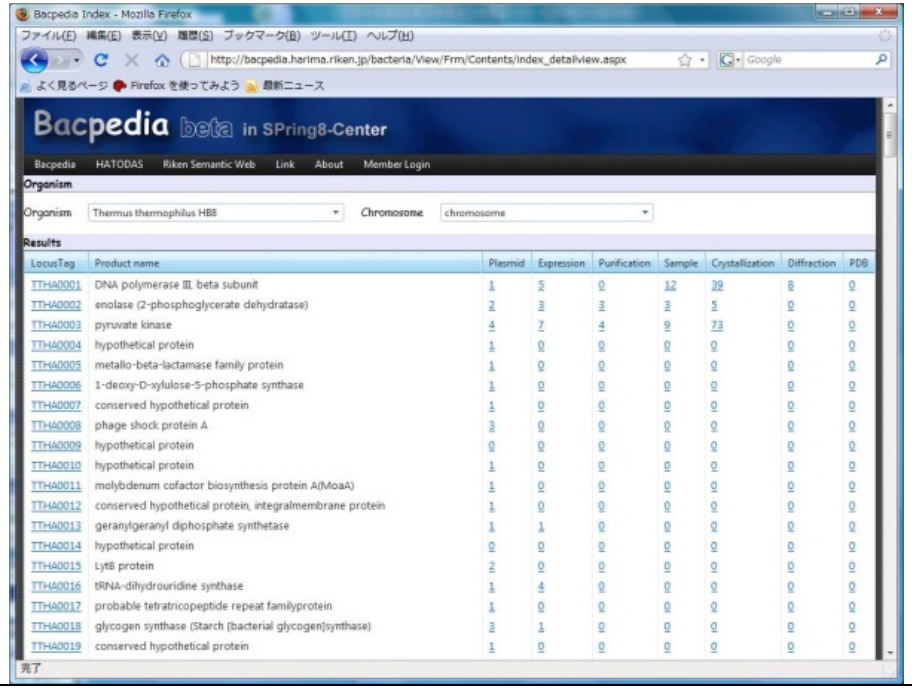
通番	データ（又はＤＢ）の名称	公開／未公開	概要（データの種類（生物種）・数量（kB 等）、本プロジェクトで実施した特徴点、進捗状況、今後の計画・課題などを簡潔にわかりやすく記述）
	植物オミックス統合データベース http://scinets.org/db/plant	公開	2010 年 10 月に、植物関係の 29 件の植物関係のデータベース（うち 17 件が理研による研究活動の成果に基づく）を、各種オミックス間の関係を定義すべく新規に開発された「上位オントロジー」に準拠して分類・統合し公開した。トップページからこの分類体系を反映させた階層式メニューを辿り、各データベースにアクセスできる。 
	RIKEN Activation tagging line http://scinets.org/item/cria37u1i	公開	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 万系統を収納する。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。データは理化学研究所植物科学研究センターのサーバー（http://amber.gsc.riken.jp/act/top.php）で維持されるとともに、理研サイネス上にも統合化されている。理研サイネス上では cDNA 情報・フェノーム情報・TAIR の公共データベースなどとの間にリンクを設定している。公開予定の情報は 2008 年 9 月までに全て公開済みである。
	RIKEN Arabidopsis Fox-hunting line http://nazunafox.psc.database.riken.jp	公開	FOX-Hunting 法によってシロイヌナズナ完全長 cDNA を挿入することによって得られたシロイヌナズナ変異体のデータベース。データ提供者が公開可能とした情報(14070 系統の変異体に関する、遺伝子 ID・挿入方向、観察された表現型情報など)は 2010 年 9 月までに全て公開した。cDNA 情報・フェノーム情報・TAIR の公共データベースなどとの間にリンクを設定している。
	RIKEN *Ds* transposon line http://scinets.org/item/cria278u1i	公開	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を収納している。データは理化学研究所植物科学研究センターのサーバー（http://rarge.gsc.riken.go.jp/）で維持されるとともに、理研サイネス上にも統合化されている。2010 年 8 月にゲノムマッピング情報を TAIR9 を用いて更新した。また、従来は同様の情報が複数の DB に分散して格納されており、利用者にはわかりにくかったため、データベース構造を改善し、変異体情報に容易にアクセスできるようにした。この作業を以て、2010 年 10 月時点で公開可能な Ds トランスポゾンライン情報は全て本データベースを通じて公開されたことになる。
	RIKEN RARGE Promoter http://scinets.org/item/ria12i	公開	シロイヌナズナのプロモーター領域に存在する「シス配列」関連情報をデータベース化した。シス配列は遺伝子発現制御を考える上で重要な数塩基の DNA 配列である。現時点で 391 件のシス配列が登録され、塩基配列・機能・文献情報・関連する完全長 cDNA（RAFL）などの情報を取得できる。データは理化学研究所植物科学研究センターのサーバー（http://rarge.psc.riken.jp/cdna/promoter/）で維持されるとともに、理研サイネス上にも統合化されている。公開予定の情報は 2008 年 9 月までに全て公開済みである。
	RIKEN RARGE Alternative Splicing http://scinets.org/item/cria3u1i	公開	シロイヌナズナ完全長 cDNA をゲノム上にマッピングして得られた選択的スプライシングクローン 1764 種類の情報をデータベース化した。公開予定の情報は全て公開済みである。それぞれのクローンに対応する cDNA の TAIR Locus Identifier の ID が格納されている。また、クローンはスプライシングのパターンをもとに 5 タイプにカテゴライズされており、このカテゴライズ結果も収納されている。データは理化学研究所植物科学研究センターのサーバー（http://rarge.psc.riken.jp/a_splicing/index.pl）で維持されるとともに、理研サイネス上にも統合化されている。公開予定の情報は 2008 年 9 月までに全て公開済みである。
	Omics Gene Models DB	公開	シロイヌナズナを対象に、affymetrix Arabidopsis genome tiling array による RNA 発現

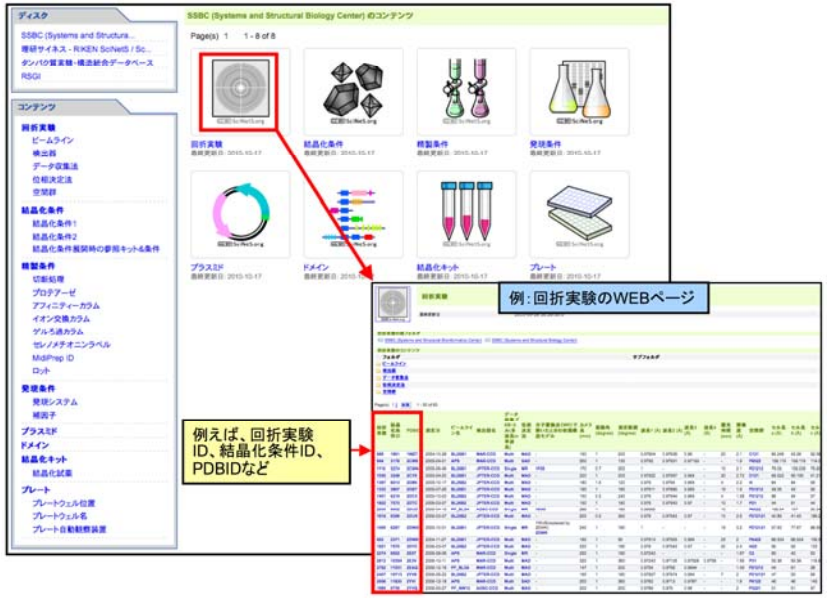
	http://scinets.org/item/cria227s904i		量の測定結果をデータベース化した。9 種の組織別サンプル（dry seed, imbibed seed, root, stem, leaf, flower, early silique, middle silique, late silique）と無処理を含む 9 種類のストレス処理サンプル（control, drought, NaCl, cold, ABA treatment、いずれも処理後 2 時間および 10 時間）を含む。また、遺伝子モデル構築プログラム：ARTADE2 を適応することによって得られた遺伝子モデルと、共発現遺伝子に集積される GO/PO/annotation term の解析結果をデータベース化した。公開を予定していた ARTADE2 遺伝子モデル：17,591 個、GO/PO/annotation term の予測数：1,874,419 個を 2010 年 9 月に全て公開した。
	RIKEN *A.thaliana* gene family http://scinets.org/item/rib123i	公開	*Arabidopsis thaliana, Populus trichocarpa*, *Oryza sativa* subsp. *japonica*, *Physcomitrella patens* の 4 植物種間で、クラスター解析によってオーソローガス遺伝子ファミリーを推定し、その結果認識された遺伝子群（クラスター）のうちシロイヌナズナ遺伝子を含むクラスター8906 個の情報をデータベース化して公表している。2008 年 7 月までに、公開予定の情報は全て公開した。
	RIKEN Ortholog (*A. thaliana* and *O. sativa*) http://scinets.org/item/crib127u1i	公開	シロイヌナズナおよびイネの遺伝子を対象に、配列の類似性に基づいて系統解析し、22631 件のオルソログ遺伝子グループを見出した。このオルソログ遺伝子についてデータベース化した。シロイヌナズナ遺伝子については TAIR locus ID、イネ遺伝子については TIGR の遺伝子 ID が示され、それぞれの遺伝子データベースへのリンクが付されている。2008 年 7 月までに、公開予定の情報は全て公開済みである。
	Cassava full-length cDNA http://scinets.org/item/cria2s1i	公開	キャッサバの完全長 cDNA クローン 8494 件の情報を収容した。それぞれのクローンについて、シロイヌナズナ（TAIR7）およびイネ(RAP-DB)の相同遺伝子へのリンクを付与し、機能アノテーション情報が取得できるようにした。2008 年 9 月までに、公開予定の情報は全て公開済みである。
	RAFL cDNA http://scinets.org/item/cria11s1i	公開	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する完全長 cDNA クローン(RAFL clone)情報 246,203 エントリーを収納する。それぞれの cDNA について、TAIR Locus ID、予測された機能、Entrez Nucleotide ID などの情報が付加されている。本データベースは、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るとともに（http://www.brc.riken.jp/lab/epd/catalog/cdnaclone.html, http://saber.epd.brc.riken.jp/sabre7/SABRE0101.cgi）、理研サイネス上に統合化されている。公開予定の情報は 2009 年 9 月までに全て公開済みである。
	A. thaliana cDNA http://scinets.org/item/cria277u1i	公開	公開されているシロイヌナズナ cDNA59726 件について、TAIR による新しいゲノム情報(TAIR9)に基づいたゲノムマッピングを 2010 年 8 月に実施し、それぞれの cDNA の位置、重複する TAIR locus の ID（AGI コード）、エクソン位置などの情報とともにデータベース化して全て公開した。それぞれのエントリーから、TAIR9 のデータベースの対応するデータアイテムへのリンクが張られており、最新の機能予測などの情報を取得可能とした。

	RIKEN plant phosphoproteome database http://scinets.org/item/cria102s1i	公開	LC-MS/MS-ショットガン法によって得られたリン酸化ペプチド情報を、シロイヌナズナについて 5143 件、イネについて 6919 件収容した。それぞれのリン酸化ペプチド情報には TAIR または TIGR の関連遺伝子情報へのリンクが付与されている。また、これらのペプチドは配列の相同性に基づいてクラスター解析され、オルソログのクラスター（種によらず相同性の高いペプチド同士を集めたグループ）13964 件が識別されている。本データベースではこれらのクラスターそれぞれについて、メンバーであるペプチドの遺伝子 ID と整列済み配列を閲覧できるようにした。この機能は、植物界において保存的なタンパク質リン酸化機構の研究に役立つことが期待される。全てのデータは 2010 年 5 月に公開された。
	RIKEN phenome integration of *Arabidopsis* mutants http://scinets.org/item/cria143u1i	公開	理研サイネス上で公開されている 3 件のシロイヌナズナ変異体に関するデータベース（Fox-hunting, DS transposon, Activation tagging line）に含まれる表現型情報に関して、Plant Ontology（PO), PATO といった公開オントロジーを用いた新しい統一的なアノテーションを設定し、公開した。現在理研サイネス上で公開されているシロイヌナズナ変異体データベースの中で明確な記述のある表現型情報については全て新しいアノテーションで関連付けを実現し、表現型アノテーション（PO・PATO データベースへのリンクを含む）・およびシロイヌナズナ変異体データベースに含まれるフェノーム情報へのリンクを整備し、統合的検索を可能とした。2010 年 9 月までに、表現型アノテーション　115 件、シロイヌナズナ変異体データベース 3 件（Fox-hunting, DS transposon, Activation tagging line）それぞれに含まれるフェノーム情報への表現型アノテーション情報からのリンク（それぞれ 13142 件,221 件,1268 件）づけを行ない、サイネス上で公開されているシロイヌナズナ変異体の表現型情報すべての統合化を完了した。
	Plant phenome http://scinets.org/item/cria224u1i	公開	シロイヌナズナの表現型情報を文献から収集し統合する新しいデータベースを作成した。2010 年 9 月末現在で、266 件の文献から、表現型に関する記述を抽出して慎重にキュレーションし、公開オントロジーを活用して標準化した。標準化した表現型には、形態、種子生産性、環境要因に対する応答性の変異などがある。これらの表現型データに、標準化のベースとなっているオントロジー、表現型変異への関与が考えられる遺伝子（TAIR Locus identifier）、文献データベースである PubMed へのリンクなどを付している。
	Bacpedia (on SciNetS) http://scinets.org/item/rib220i	公開	趣旨：微生物由来蛋白質（変異導入蛋白質を含む）にかかわる試料調製と回折実験データを提供する。 生物種：*Thermus thermophilus* HB8, *Pyrococcus horikoshii* OT3 他、 9 種の微生物 測定方法：X 線結晶構造解析 データの内容：発現プラスミド構築・培養・精製・結晶化・回折データ収集の各段階における実験の条件・結果・生データなど データの量：試料調製データ（発現プラスミド構築実験 11,800 件、培養実験 7700 件、精

			製実験 3600 件)、結晶化実験データ（観察 1250 万件)、700 件の回折実験データ（データセット数) 管理状況：理研サイネス上で統合的に管理 開始年月：平成２１年７月 
	Bacpedia (original) http://bacpedia.harima.riken.jp	公開	趣旨：微生物由来蛋白質（変異導入蛋白質を含む）にかかわる試料調製と回折実験データを一般の研究者にとって使いやすい形で提供する。もともとキュレーションのために構築したが、一般の研究者にとって使いやすい環境を整えるため公開することにした。 生物種：*Thermus thermophilus* HB8, *Pyrococcus horikoshii* OT3 他、９種の微生物 測定方法：X 線結晶構造解析 データの内容：発現プラスミド構築・培養・精製・結晶化・回折データ収集の各段階における実験の条件・結果・生データなど データの量：試料調製データ（発現プラスミド構築実験 11,800 件、培養実験 7700 件、精

			製実験 3600 件)、結晶化実験データ（観察 1250 万件)、700 件の回折実験データ（データセット数) 管理状況：理研放射光科学総合研究センター（RSC）のサーバで維持 開始年月：平成２２年７月 
	HATODAS (on SciNetS) http://scinets.org/item/rib108i	公開	趣旨 : 白金や水銀などの重原子を含む試薬で標識した標識タンパク質のデータを提供する。目的タンパク質の結晶構造を決定するために使用すべき重原子試薬が簡単に検索できるようにすることを目的としたデータベース HATODAS の検索基盤データである。 生物種：多種 測定方法：X 線結晶構造解析 データの内容：重原子・重原子結合サイト・重原子化実験方法・沈殿剤・緩衝液など データの量：重原子導入蛋白質に関わるデータ 2500 件 管理状況：理研サイネス上で統合的に管理 開始年月：平成２１年７月

			
	SSBC(理研サイネス) http://scinets.org/db/ssbc	公開	概要：高等動植物等由来タンパク質構造データと実験データ 公開開始年月、データの内容と量： ・2009 年 5 月：回折実験データ 2 万件と付随する実験データを理研サイネス上に公開した。 ・2010 年 8 月：回折画像データ 1 万 1 千件と付随する実験データを理研サイネス上に公開した。 ・2010 年 9 月：約 50 万件の結晶観察データと付随する実験データを提供した。 ・2010 年 10 月：上記の回折実験データ、結晶観察データ、付随する実験データをすべて DBCLS へ提供した。 上記の回折実験データおよび付随する実験データを理研サイネス(http://scinets.org/db/ssbc)からダウンロード可能とした。（結晶観察データ約 50 万件をテストサーバで編集中。）

			管理状況：理研サイネス上で統合的に管理 

（２）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）
	理研サイネス	公開	「セマンティックウェブ形式」と呼ばれる国際標準規格に準拠したデータベース統合システムで、以下のような特徴を持つシステムとして開発した。 ①一数万個以上の個別データベース構築活動を、大勢の研究者がインターネット経由で並行して実施できる。 ②大規模なデータを介した業務フローを柔軟に設定でき、人的連携や自動処理を容易化できる。 ③各活動群をセキュリティの高い状態で区切り、未公開の状態でデータベース構築ができる。 ④構築したデータベースをその基盤から直接公開できる。 ⑤公開後も、研究者がシステムの維持コストを負担することなく、その基盤でコンテンツを継続的に

			更新することができる。 ⑥複数の世界標準形式に準拠したデータ配信が容易。 同システムは 2009 年 3 月にベータ版が公開され、植物・タンパク質統合 DB 計算機環境として利用されている。2010 年 10 月までに、タンパク質結晶構造解析実験データベースなどに伴う大容量データも安定的に扱えるように、またデータ検索等がスムーズにできるようにシステムの改良が進められた。また、公開したデータをクライアントサイドプログラムからウェブ経由で利用するための API である Semantic-JSON を構築した。また、これまで構築したアノテーションシステムに、クライアントサイドのウエブブラウザから理研内外を問わず誰でもプログラムを書いてデータベースを利用できる機能を実装し、プログラムやその処理結果を交換、公開できるオープンな参加型システムを実現した。今後は、同システムを研究成果公開のためのサイバーインフラストラクチャと位置付け、整備を進めていく。

外部発表実績一覧

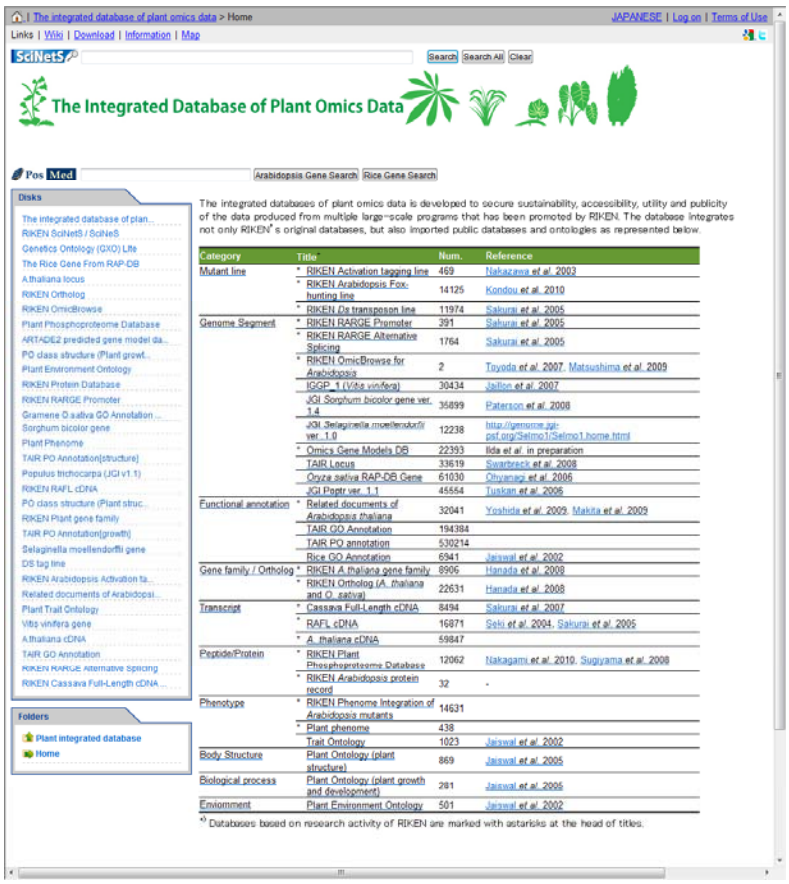
（1）プレス発表、取材対応

通番	タイトル	発表媒体	年月日	特記事項
1	理研のデータベース構築基盤の公開基準をセマンティックウェブに統一	理化学研究所プレスリリース	2009 年 3 月 31 日	
2	世界最大級のタンパク質結晶構造解析実験データベースを公開ー90 万件の結晶化条件などを整備した実験データ群が利用可能にー	理化学研究所プレスリリース	2009 年 7 月 23 日	

（2）学術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	SciNetS: A New Publication Medium For The International Plant Database Integration Consortium Using LaaS-Cloud-Computing And Secure-Semantic-Web Technologies	Koji Doi, Yuko Makita, Norio Kobayashi, Yoshiki Mochizuki, Yuko Yoshida, Shuji Kawaguchi, Kei Iida, Akihiro Matsushima, Manabu Ishii, Koro Nishikata, Yukiko Kanda, Satoshi Takahashi, Erimi Harada, Toshiaki Watanabe, Motoaki Seki, Takeshi Yoshizumi, Kosuke Hanada, Keiichi Mochida, Hirofumi Nakagami, Tetsuya Sakurai, Hiroshi Masuya, Takashi Kuromori, Masatomo Kobayashi, Minami Matsui, Kazuo Shinozaki and Tetsuro Toyoda	Plant Cell and Physiology	投稿中	
2	The RIKEN integrated database of mammals	Hiroshi Masuya, Yuko Makita, Norio Kobayashi, Koro Nishikata, Yuko Yoshida, Yoshiki Mochizuki, Koji Doi, Terue Takatsuki, Kazunori Waki, Nobuhiko Tanaka, Riichiro Mizoguchi, Kouji Kozaki, Tei-ichi Furuichi, Hideya Kawaji, Shigeharu Wakana, Kaoru Fukami-Kobayashi, Osamu Ohara, Yoshihide.Hayashizaki, Yuichi Obata, Tetsuro Toyoda,	Nucleic Acids Research	2010, 1–10	

植物オミックス統合データベース
http://scinets.org/db/plant



理研発の各種オーミックス情報をはじめ、世界の研究機関が公開している遺伝子アノテーション情報などを統合化。

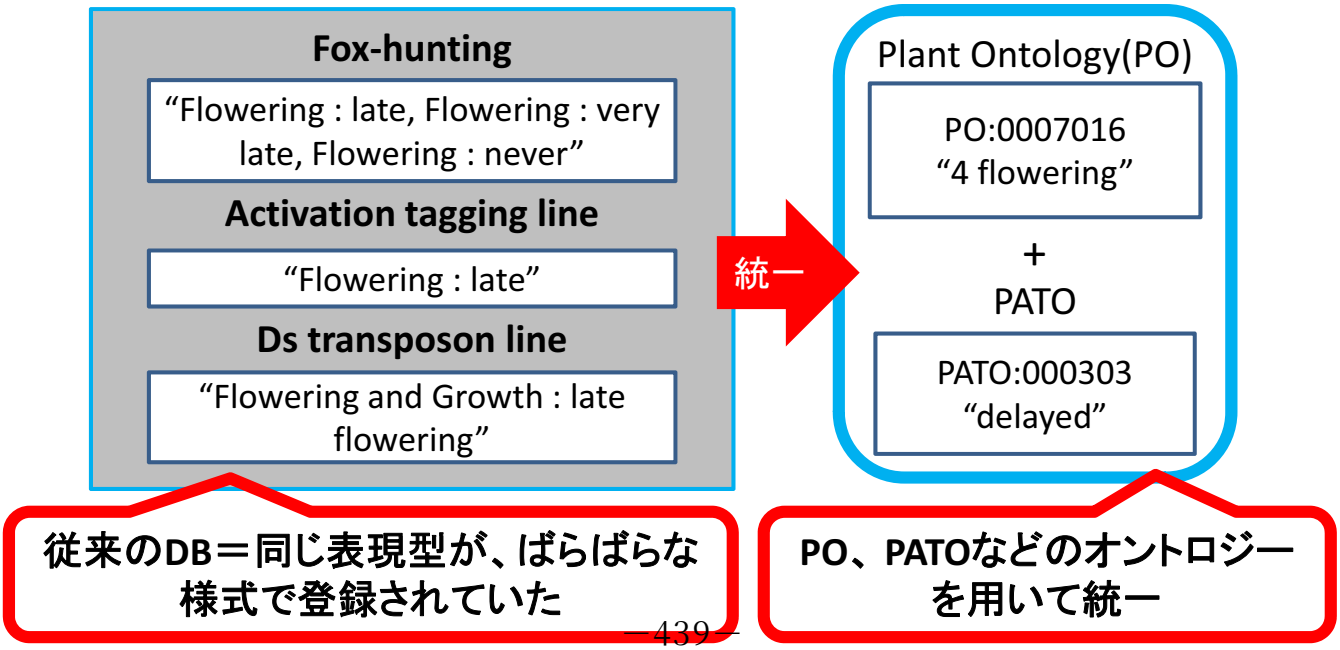
統合化されたDBは、「上位オントロジー」に基づいて分類し、トップページに階層式メニューで示している。

2010年10月時点で29の植物関係のデータベースが統合されている。

うち17件が理研による研究活動の成果に基づくデータベースである。

植物フェノーム情報の統合

理研に複数ある変異体データベース群を統合化し、オントロジーによる表現型情報の統一を図る



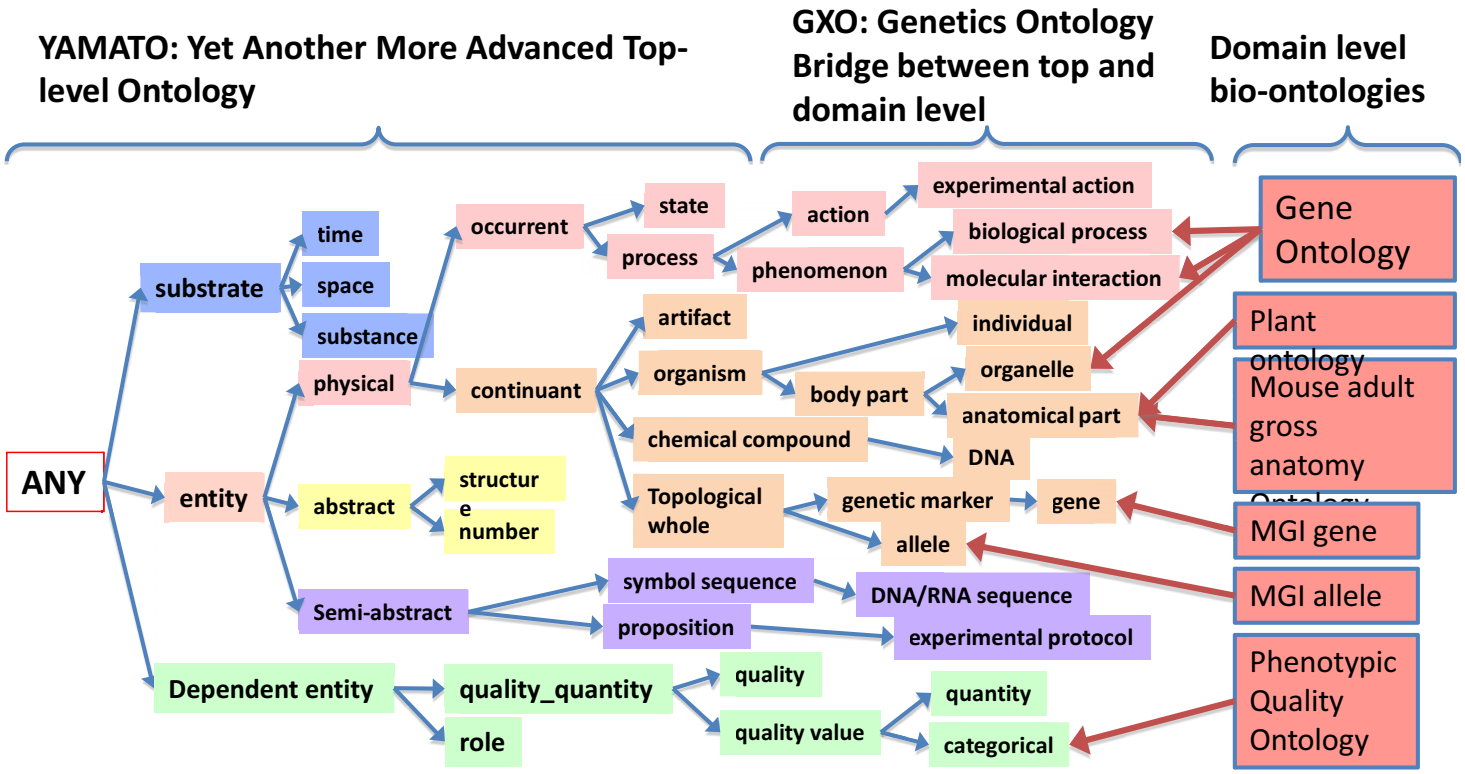
データベース横断型の検索

検索キーワード
（ここでは、drought(乾燥)）

検索結果の概略
（ヒットのあったデータベースではデータ総件数とヒット件数を表示）

アイコンをクリック
↓
各DB内の検索結果を閲覧できる

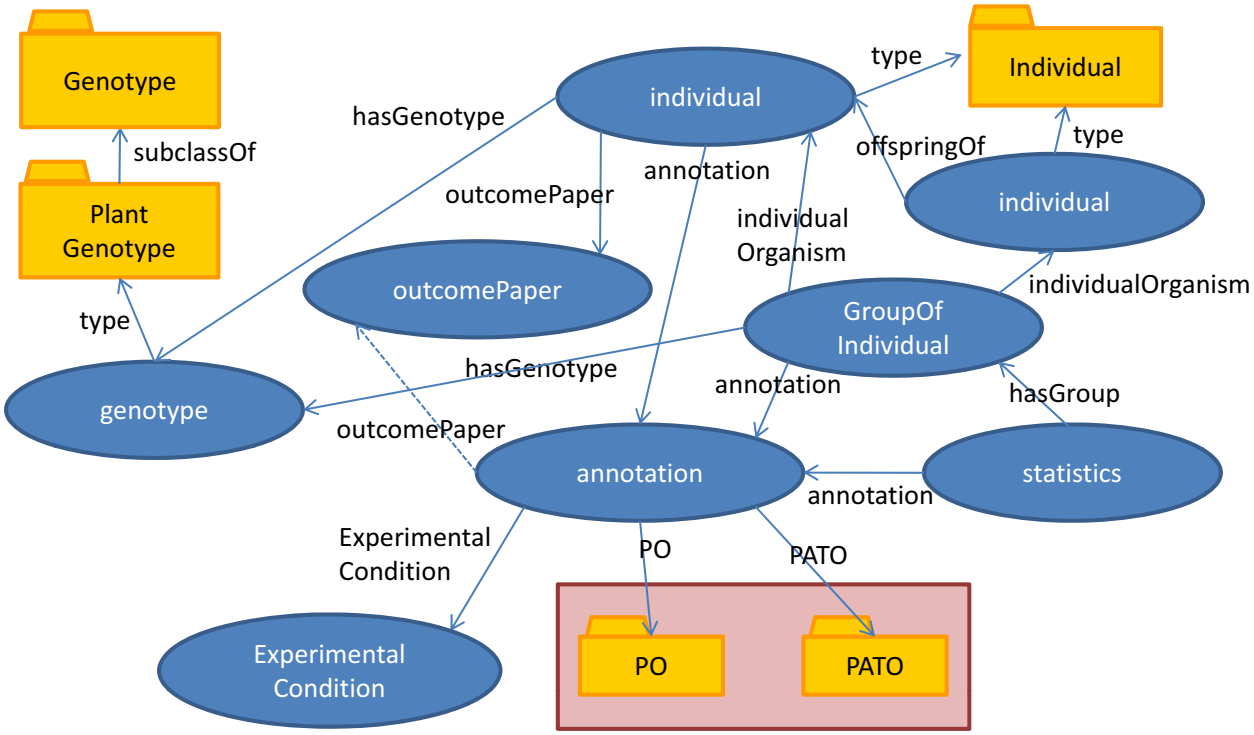
Top-level ontology based integration of broad information



Version 0.5 is available now at: http://www.brc.riken.go.jp/lab/bpmp/Ontologies/

データベース統合化の範囲を拡大

シロイヌナズナで実施してきた統合化をモデルケースとして、
シロイヌナズナ以外の生物種にも適用→マウスフェノタイプへ



生物種を問わず適用できるフェノーム情報表現のための枠組みを設計

区　　分	□中核機関（□代表機関／□参画機関） □分担機関（□代表機関／□参画機関） ■補完課題実施機関
課　　題　　名	「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）
実 施 機 関 名	独立行政法人産業技術総合研究所
代 表 研 究 者 名	成松　久

１．課題開始時における達成目標

糖鎖業界に散在するデータベースを産業技術総合研究所・糖鎖医工学研究センター（以下、糖鎖医工学研究センター）に設置予定のデータベースサーバに集約し糖鎖科学統合データベースを構築する。まずは、糖鎖医工学研究センターが構築した５種類のデータベース（糖鎖関連データベース、糖転移酵素特異性に関するデータベース、レクチンデータベース、糖タンパク質データベース、および質量分析データベース）について、統合を行う。

また、我が国に存在する糖鎖関連データベースを統合するために、糖鎖医工学研究センターは糖鎖関連データベースを保有している研究機関と交渉し、了解を得た上でデータの提供元となるよう促す役割を行う。そのために、各データ提供機関が糖鎖医工学研究センターに円滑にデータを提供できるように糖鎖医工学研究センターは必要な支援を行う。複数の機関から得られたデータを標準化した上でデータベースに格納する。

糖鎖医工学研究センターはデータベースに格納した専門的な情報を直感的に理解できるインターフェースを開発し、糖鎖研究分野以外の研究者等にも理解可能な情報として公開する。

各種様々なデータをいくつかのカテゴリーに分類した上で、統合に必要な情報を中核機関である情報・システム研究機構に受け渡し、情報・システム研究機構の DB ポータル等の検索方法と連携できるように開発を行う。

最終的には情報・システム研究機構等とのデータの統合を目指し、統合検索を共同で開発する業務を行う。

２．平成２２年１０月末時点における事業計画に対する成果

（１）成果概要

横断検索・統合検索インターフェース開発

・Hyper Estraier によるキーワード検索

・キーワードによる統合検索（現在プロトタイプ版）

構造に関する統合検索インターフェースの開発

・糖鎖業界の単糖標準表記（CFG シンボル）による検索

・Chem info のエディターによる検索

産総研の DB 開発（詳細については資料３－２　別紙１を参照のこと）

・糖鎖関連遺伝子データベース（GGDB）の改良

・レクチンフロンティアデータベース(LfDB) の改良

・糖鎖のスペクトルデータベース(GMDB) の改良

・糖タンパク質データベース(GlycoProtDB) の改良

・グリコシダーゼデータベースの開発
・阻害剤データベース(GGIDB) の開発
・単糖のデータベース(JMSDB) の開発
・糖鎖構造のデータベースの開発
・糖鎖関連特許のデータベースの開発
・病原体と宿主の糖鎖との結合情報のデータベースの開発
・腫瘍マーカーリファレンスデータベース(TuMaRDB) の開発
・実験プロトコルオンラインデータベース(GlycoPOD) の開発
・糖鎖関連疾患の遺伝子データベース(GDGDB) の開発
・有機合成による化合物とその合成反応のデータベースの開発
・JCGGDB オンラインリポートの開発
・ノックアウトマウスのデータベースの開発

連携のための API 開発
・立命館大学・GlycoEpitope の API 開発
・理化学研究所・糖鎖コンフォメーション DB の API 開発
・創価大学・FlyGlycoDB の API 開発
・名古屋市立大学・多次元 HPLC データベースの DB と API 開発
・野口研究所・有機合成 DB を JCGGDB の合成 DB へデータを移行する際の API 開発

クロスリンク
・京都大学・KEGG　Glycan
・EuroCarbDB の GlycomeDB

横断検索インデックス化の協力
・生化学工業・水谷財団の GlycoForum

（2）進捗及び成果

円滑にプロジェクトを運営できるように公開サーバやデータ整備等のプログラミングの環境を整え、糖鎖統合データベースのポータルサイトを構築した。そして、糖鎖医工学研究センターが保有しているデータベース（糖鎖関連データベース、糖転移酵素特異性に関するデータベース、レクチンデータベース、糖タンパク質データベース、および質量分析データベース。以下、糖鎖センターの DB と略す。）を公開用にリニューアルし、構築したポータルサイトから横断検索できるように開発を行った。糖鎖医工学研究センターが平成１９年度から構築している横断検索を基盤にし、平成２０年度から国内に散在する糖鎖関連データベースを集約し始めた。　平成２２年度末日まで引き続きデータベースを保有している研究機関との連携を行い、横断検索と統合検索の対象となるように作業を行っている。

① 運営と開発体制の準備

平成１９年度、糖鎖医工学研究センターに運営チームを設置し、糖鎖科学統合データベースのプロジェクト期間中の中核的な役割を果たせるように運営と開発の体制を整えた。また、開発と業務が円滑に進められるようにプロジェクト専従者にはプログラミング用のコンピュータを用意し、UNIX 環境でデ

ータベースや XML を利用したサービスを JAVA 言語、Perl 言語、Flash 等で開発できるように必要なソフトウェアを購入し環境を整えた。

次に、糖鎖科学統合データベースの運営方針や活動を糖鎖業界の方に理解して頂けるようにポータルサイトを構築した（下図）。

平成２０年度の早い時期に、糖鎖研究のみならず糖鎖との関連が示唆される糖鎖周辺分野の研究者も利用できるようにデータベースの設計を開始した。

平成２０度はまずは糖鎖医工学研究センターの糖鎖のデータベースを統合し、それを基盤として平成２０年度以降に、外部の研究機関のデータベースと連携できるように計画を進めた。

まずは、キーワードによる横断検索のポータルサイトを平成２０年８月にリニューアルして公開した。日本糖質学会で統合データベースの発表を行いその後にアクセスが急増した。また、その他の学会や日本糖鎖科学コンソーシアム（JCGG）年会など糖鎖の研究者が集結する場所で発表を行い統合データベースの活動報告を行った。その際に研究機関の代表者と話合い、平成２１年度から統合プロジェクトに参加への協力の意思が得られた。平成２１年度から協力する機関へはこれまでの統合データベースの活動と今後の計画を説明し、その上で個々の機関のデータの意義やプロジェクト内での役割を理解して頂いた上で、統合プロジェクトの中でどのように統合を進めていくか平成２１年度の業務計画を立てながら具体的に話し合った。平成２１年度も上述と同様の方法で協力機関を取り込むための活動を継続した。

参加募集の活動以外にも、糖鎖に関連するライフサイエンス分野の研究者に、統合検索システムの活用方法を広めていく普及活動を行った。日本糖質学会に於いて、データベース展示ブースを設置し、使い方の説明やプロジェクトに参加している研究機関のデータベースの解説などを行った。糖鎖関連 DB の使い方の勉強会も行った。

②データ提供機関との交渉

平成１９年度国内の数多くの主要な研究機関・大学・企業が集結している最大の団体である「日本糖鎖科学コンソーシアム（JCGG）」に協力を要請し、JCGG 企画委員会の会議に参加して統合データベー

スの活動を説明した。その結果、JCGG のデータベースとしての公認を頂いた。その後、毎年行われている JCGG のシンポジウムで経緯と今までの活動と今後の展開を説明し、統合データベースプロジェクトへの参加を広く呼びかけた。その活動と同時に、既に公開されている糖鎖関連データベースを所有している機関にデータを提供して頂くように個別に交渉し、平成２０年度の計画に Web サービスの通信プロトコル SOAP を利用しデータベース間の連携を行い、糖鎖統合データベースのポータルサイトから横断検索や統合検索ができるように産総研と各機関が連携したシステムを構築した。
平成２０年度は、立命館大学（代表：糖鎖工学研究センター・センター長川嵜敏祐氏）のGlycoEpitopeの横断検索のインデックス化するためのテキスト出力のAPIの開発と、REST方式を使用したキーワードによる検索とその検索結果並びに個々のエントリーの詳細情報をXMLで出力できるAPI（以後、統合検索用ＡＰＩと略す。）を開発して設置した。現在稼働しているキーワードによる横断検索のインデックスにテキスト出力の結果を利用している。RESTを利用したＡＰＩは統合検索に利用することにしている。名古屋市立大学（代表：名古屋市立大学大学院薬学研究科生命分子構造学分野・加藤晃一氏）の多次元ＨＰＬＣデータベース（ http://hplc.glycoanalysis.info/ ）を構築して公開した。API用のデータとしてXMLを出力するシステムを公開した。その他の連携として、データベース化していないデータを保有している３機関と話合いを持つことができた。まずは、財団法人野口研究所（代表：常務理事・白井孝氏）では有機化学合成を用いて糖鎖を合成していることから化合物の未公開のデータベースを持っている。更には合成経路や合成のノウハウの情報も持っている。野口研究所と産業技術総合研究所が文科省の統合データベースプロジェクト内で平成２１年度から２年をかけて糖鎖合成の統合データベースを構築することとなった。JCGGDBのサーバにデータを移行するAPIの仕組みを構築した。糖鎖センターでは論文等から化学構造と反応経路のデータをDBに入力する作業を行っている。DBに登録した化合物を検索するインターフェースが完成しておらず未公開ではあるが、ステレオケミストリーを考慮した検索技術の問題も解決したため、平成２２年度末日までに公開する予定となっている。
次に、創価大学（代表：工学部　専攻長・西原祥子氏）のショウジョウバエの糖転移酵素の遺伝子をノックダウンしてフェノタイプ情報を格納している FlyGlycoDB（平成２２年 10 月時点に於いても未公開）の公開版への作りかえと統合検索用のＡＰＩの設置を平成２１年度行った。FlyGlycoDB が公開後に JCGGDB から検索できるようになっている。
更に、平成 21 年度に理化学研究所（代表：システム糖鎖生物学研究グループ　グループディレクター・谷口直之氏）の糖鎖コンフォメーションデータベース（平成２２年度末日までに公開予定）の公開支援と構造による横断検索・統合検索用ＡＰＩの設置を行った。API が完成しているため公開後に連携を図る予定となっている。

また、糖鎖医工学研究センターが保有しているデータベースのデータ自体は経産省・ＮＥＤＯプロジェクトの成果物であるため、経産省側のプロジェクトにも協力する義務がある。平成２０年度から産業技術総合研究所・バイオメディシナル情報研究センター（旧生物情報解析研究センター）の代表者今西氏と協議を開始し、経産省の統合 DB プロジェクト側での GGDB の API の設置に協力した（GGDB は平成 19 年度の本委託費で構築した。GGDB の API に関しては経産省統合 DB プロジェクトの平成 20 年度の予算で構築した。）。平成 21 年度には経産省統合 DB プロジェクトの H-invDB の API を利用して糖鎖センターの GGDB や LfDB にある遺伝子の詳細画面に組織発現や遺伝子多型（SNP やマイクロサテライトなど）のデータを動的に引き出し画面に表示している。

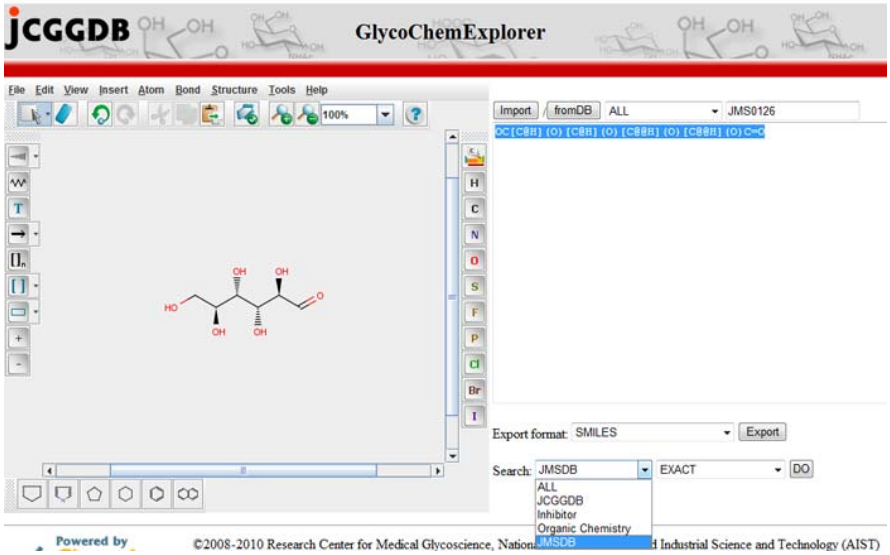
③新規データベース構築と糖鎖関連 DB の統合インターフェースの開発

平成１９年度には、糖鎖医工学研究センターが構築した５種類のデータベース（糖鎖関連遺伝子データベース：GlycoGene Database、糖転移酵素特異性に関するデータベース:KEM-C、レクチンデータベース Lectin frontier Database、糖タンパク質データベース:GlycoProt Database 、および質量分析データベース：GlycoMass Database）について、既存のデータベースのインターフェースを改良し糖鎖科学統合データベースの基盤を整えた。まずは糖鎖関連キーワードや遺伝子名等を利用した横断検索機能を搭載し仕組みを作り、平成 20 年 8 月に公開した。

平成２０年度は、それらの DB を基盤とし、糖鎖科学統合データベースのコンテンツと検索機能を拡充した。まずは、「構造解析と検出に関するカテゴリー」と題して産総研の「質量分析によるスペクトルデータベース」と新規開発した名古屋市立大学の「多次元 HPLC データベース」、日本脂質生化学会脂質データベース構築委員会の「LipidBank（糖脂質のデータのみ)」と立命館大学の「GlycoEpitope」と産総研の「レクチンデータベース」を組み合わせて、質量分析の m/z 値や単糖組成などから想定される糖鎖構造を抽出し、各種様々な方法論で解析されたデータに辿りつけるように統合した。構造による統合検索もできるようになった。各研究機関の糖鎖構造の管理番号やデータの表記の仕方が異なっていたが、糖鎖センターの構造 ID を基に各機関の糖鎖構造との連携が可能になった。構造情報だけではなく質量分析の結果を参照できるように m/z や単糖組成（簡易版・詳細版）から構造検索できるようにした。また、「糖鎖関連遺伝子」のデータベースのカテゴリーの中で、名古屋大学（代表：古川鋼一氏）の協力を得てノックアウトマウスのデータベースを新規に構築した。国内の研究者が保有している糖鎖関連遺伝子のノックアウトマウスの情報を研究者が自ら登録できる仕組みを構築した。現在、名古屋大学のグループにマウスの情報の登録を依頼しており、平成２２年度末日までに全ての情報を公開する予定となっている。

平成２１年度には、糖鎖科学統合データベースのコンテンツと検索機能を拡充した。まずは DB 間の連携をより強めるためにも新規の DB の開発を行った。例えば感染と疾患の関連や病原菌と糖鎖の結合情報をデータベース化して PACDB と名付けた。糖鎖構造と病原体の情報がリンクしている。他にも糖鎖関連遺伝子の変異によって起こる病気の情報を収集（Glyco-Disease Genes Database と命名し GDGDB を略す）し、遺伝子と病名や病態がリンクする。遺伝子名を利用して GGDB、GDGDB、ショウジョウバエの糖鎖関連遺伝子のデータベース（FlyGlycoDB）との連携を行った。遺伝子のホモログ同士を統合・連携できた。このように新しい DB を作ることでより付加価値の高い情報提供を行うことができた。FlyGlycoDB がまだ公開されていないこともあり GGDB と GDGDB のクロスリンクレベルの連携ののみ利用できる。

平成２２年度新規に開発している DB とインターフェースには、糖転移酵素やグリコシダーゼの阻害剤 DB、有機化学合成の化合物 DB、糖鎖構造 DB、単糖 DB をそれぞれ構築し、ステレオケミストリーを重視して全ての DB を同時に構造検索できるようにインターフェースを開発中である。構造検索機能としては、完全一致検索、部分構造検索、類似構造検索、スーパーストラクチャー検索が検索できるようになっている。



③-1 研究領域毎の統合データベース構築

ｉ「構造解析と検出に関するカテゴリー」（構造の横断検索）

平成２０年度に構築した「構造解析と検出に関するカテゴリー」（構造の横断検索）に理化学研究所の「糖鎖コンフォメーションデータベース」が参加することになり API の開発が完了し、構造による横断検索システムで連携できる状況になった。この連携は、糖鎖コンフォメーションデータベースの論文で発表した後にこのデータベースを公開することになっているので、JCGGDB との連携部分の公開が遅れている。

ⅱ オンラインプロトコルの統合検索システム構築

「糖鎖大量合成」研究領域の中で、財団法人野口研究所と共同で有機化学による糖鎖合成とその後末端構造にバラエティを持たせるための酵素合成法の２つを融合させて糖鎖標準品の合成を支援するシステムを２年かけて開発しているところである。データの提供元として、野口研究所のほかに岐阜大学・木曾研究室の協力を得ることができた。糖鎖の標準品の合成法を誰でも検索できるように、外部の研究者が反応経路と合成に関わるノウハウを登録できるシステムを開発した。

立命館大学の川嵜らの協力により、糖鎖研究の初心者でも糖鎖の解析を行えるように各種様々な解析に関するオンライン実験プロトコルを整備し、約５０種類のプロトコルを公開した。平成２２年度末までに合計１００種類以上のプロトコルを公開する予定になっている。

GlycoPOD

Parts | About Database

Glycolipid and related compounds

Extraction of glycolipids

Authors: Nozomu Okino, Makoto Ito (Ver 1.0) Q&A | Rating

Introduction | Protocol | References

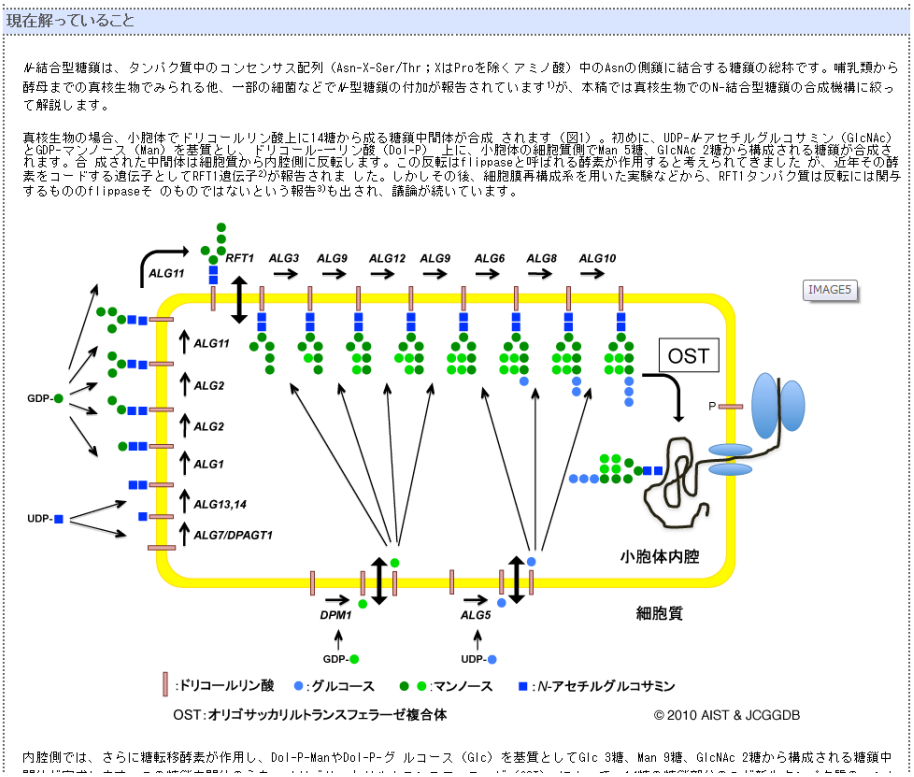
Category	Glycolipid and related compounds
Protocol name	**Extraction of glycolipids**
Authors	*Nozomu Okino** Department of Bioscience and Biotechnology, Graduate School of Bioresource and Bioenvironmental Sciences, Kyushu University *Makoto Ito* Department of Bioscience and Biotechnology, Graduate School of Bioresource and Bioenvironmental Sciences, Kyushu University *To whom correspondence should be addressed.
Keywords	Extraction of glycosphingolipids (GSLs) Neutral GSLs Acidic GSLs (ganagliosides)
Reagents	* chloroform ($CHCl_3$) * methanol (MeOH) * $CHCl_3$-MeOH mixture (C/M, volume/volume) * 4 M NaOH * 4 M acetic acid * distilled water (DW)
Instruments	* OASIS HLB cartridge 1 cc/30 mg (reverse-phase cartridge, Waters) * OASIS MAX cartridge 1 cc/30 mg (anion exchange cartridge, Waters) * sonic bath * Speed Vac concentrator (Savant)
	1. **Extraction of GSLs from tissues**

ⅲ糖鎖関連疾患とその糖鎖関連遺伝子のデータベース

名古屋大学の協力を得て国内にあるノックアウトマウスのデータベースを構築した。国内の研究者が保有しているノックアウトマウスの情報を研究者が自ら登録できる仕組みを構築した。更には登録したリソースの管理（配布方法・問合せ先等の情報も管理）ができるシステムも構築した。

ⅳ糖鎖の機能に関する情報のデータベース

糖鎖の役割の重要性を広く知ってもらうために、一般の方から専門の方まで理解しやすい易しい読み物を作成して公開した。平成２１年度に依頼した原稿を公開し、２２年度分は原稿を収集中である。蓄積した文章に対して、データマイニング技術を駆使して、糖鎖の専門用語と遺伝子名などを結びつける情報を抽出できる技術の確立を目指し開発した。この方法で自動検出した専門用語に隊捨て JCGGDB と関連する各種データベースへリンクは付ける自動リンクシステムを開発した。例として、遺伝子名や糖鎖構造名などが文章中に検出されるとリンクすることになっている。



ｖ経産省側の統合 DB プロジェクトとの連携

糖鎖医工学研究センターのデータベースに遺伝子多型やタンパク質ドメイン構造等の H-invDB の情報を表示できるように、平成２０年度には産業技術総合研究所・メディシナル情報研究センターの代表者と協議を行い API の開発を行った。平成２１年度はお互いの DB の画面に補完してメリットのあるデータを表示させた(経産省側との DB 連携)。産総研・糖鎖センターが保有している GGDB には遺伝子産物のアイソフォームの情報を表示し、産総研・バイオメディシナル情報研究センターの H-InvDB には GGDB へのリンクを表示できるように協力した。また、糖鎖センターが保有する GlycoProtDB に API を設置し、H-InvDB から GlycoPortDB にリンクできるようになった。マウスやヒトの糖タンパク質の情報を公開したときに、API の改良せずに同様にリンクできるように API の設計・設置とデータ提供に協力した。H-invDB に GlycoProtDB の糖鎖修飾の情報を表示できるように協力した。

③-2 統合データベース構築に向けた開発と中核機関との連携

ｉ検索用インデックスの自動更新システムの開発

これまでに構築した外部 DB と連携するための API を活用し、横断検索のインデックスを自動更新するシステムの開発を行った。更新した情報をライフサイエンス統合 DB センターの公開しているポータルサイトの検索用インデックスに反映するようにした。

ⅱ直観的なポータルサイト実現のための技術開発

平成２０～２２年度に作成した API を利用し統合検索システムを構築しているところである。専門用語の共起頻度を計算するマシンの処理速度が遅いためライフサイエンス統合 DB センターに協力をお願いする予定で、検索システムを平成２２年度末までに公開する予定で作業中である。

３．当初目標に対する達成度

当初産総研に与えられた使命である「国内にある糖鎖関連 DB を可能な限り統合すること」に対しては、十分達成できたと思っている。平成２０年８月の公開から平成２２年１１月２３日付、JCGGDB のサービスにアクセスしたユニーク IP アドレス数が 5467 件。

４．中間評価に対する対応

継続して更に協力機関との統合を進めた。

５．他機関との連携

中核機関との連携では、中核機関である、ライフサイエンス統合 DB センターには横断検索のインデックスを提供した。また、ライフサイエンス統合 DB センターからはマシンパワーや情報提供などを受けた。

協力機関との連携については、各機関の希望を伺い産業技術総合研究所だけの成果にならないように気を配った。協力して頂いた機関には、今後 JST を通して著作物の共同保有の契約変更（名称が適切ではないかも）を行う予定としている。当プロジェクトに協力して頂いた機関には、立命館大学、名古屋大学、名古屋市立大学、LipidBank 構築委員会の４機関が平成２０年度から協力し、平成２１年度からは理化学研究所、創価大学、野口研究所が新たに参加した。平成２２年度は、前年度の２ヶ年計画の研究機関と継続して協力して頂いた。

協力機関との連携確保のために行ったことは、今期のプロジェクト体制では共同研究も再委託契約もできないことから協力機関には実務ベースにおける役務費の提供を行った。産総研の一般公募を通して役務契約をしたこともあり、実務を行うまでにタイムラグがあり作業が遅れた。また、協力機関の方々の通常の研究もあり、連携のための作業負担を与えないように企業に外注し、協力者の負担を軽減することができた。

６．今後の見通し、計画、展望

今後も研究側が必要とする情報（実験データ・論文抽出）をデータベース化する。それと同時に、新規に構築されてくる糖鎖関連データベースと既存の DB とを連携する作業を継続する。既存の DB の更新や研究者が自ら登録できるように改良を加える。また、新規 DB のデータの精度を再確認しながら再度修正する必要がある。

集約したデータを利用して研究に役に立つグライコインフォマティクスのツール作りを行い研究者に提供する。

７．全体総括

当初産総研に与えられた使命である「国内にある糖鎖関連 DB を可能な限り統合すること」に対しては、十分達成できたと思っている。平成 22 年 10 月末日時点でも、公開できていないものが沢山あるが、公開された時には間違いなく幅広い研究者へ良い影響を与えられると思っている。

プロジェクトの中では、DB の統合や連携だけではなく、実験プロトコルを構築するにあたり、糖鎖業界が一体となって糖鎖とかかわりのある周辺領域の研究者・学生が糖鎖研究をする際のサポート体制が同時に整ったものと確信している。

８．特記事項

今日の NCBI でも、糖鎖のカテゴリーがないこともあり、日本が糖鎖の DB の中でトップに立つためには国内の DB から整備することが急務であった。現在３極化している日本、ヨーロッパ（糖鎖構造・糖鎖構造解析(NMR,MS)）、米国の DB（糖鎖構造・レクチンアレイ・ノックアウトマウス）がある。

今回のプロジェクトで構築したデータベースは日本固有のデータが多いため糖鎖構造解析・構造情報以外では連携できない。構造情報とキーワード（遺伝子名、糖鎖構造名）が連携するための共通のキーワードである。しかし、これまでの日本のナショナルプロジェクトの成果と当プロジェクトで構築したデータベースにより、３極の中ではデータ量とデータの種類では間違いなくトップとなった。また、日本の糖鎖関連の全ての DB を併せることで糖鎖研究領域のカバー率も圧倒的に高くなった。

９．委託研究費一覧

	１８年度	１９年度	２０年度	２１年度	２２年度	計
設備備品費（千円）	0	18,645	0	0	0	18,645
人　件　費（千円）	0	5,224	18,123	19,057	26,753	69,157
業務実施費（千円）	0	21,449	15,051	18,029	20,539	75,068
一般管理費（千円）	0	4,526	3,344	3,799	4,708	16,377
合　計　（千円）	0	49,844	36,518	40,885	52,000	179,247

整備実績一覧

（1）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	キーワードによる横断検索	公開	国内にある糖鎖関連のＤＢを横断検索できるように整備。
2	糖鎖構造による統合検索	公開	国内にある糖鎖関連のＤＢを対象に、糖鎖構造で検索できるエントリーを整備し、構造の冗長を除いたもの。一回の検索で国内の糖鎖構造解析や名称などを検索できる。
3	糖鎖関連遺伝子の阻害剤データベース(GGIDB)	未公開	糖転移酵素やグリコシダーゼの阻害剤を中心に収集。構造検索できるように検索インターフェースを開発中。
4	糖タンパク質統合データベース	未公開	糖鎖修飾位置の情報を網羅的に収集したデータ。
5	オンラインリポート http://jcggdb.jp/doc/	公開	パスウェイの情報や糖鎖の技術を産業に応用している方のリポート
6	グライコサイエンスプロトコルデータベース（GlycoPOD）	公開	開発中。実験プロトコルをオンライン化する。専門分野の研究者に執筆を依頼。平成２１年度４４プロトコルを収録。公開間近。平成２２年度までに１００以上のプロトコルの収録を目指す。
7	糖鎖関連疾患とその原因遺伝子のデータベース(GDGDB) http://jcggdb.jp/doc/	公開	糖鎖関連遺伝子の変異が原因でおこる疾患を収集。遺伝子の情報とあわせて、病態の情報を収録。 病態の情報はＷｅｂサイトなどから収集。著作権者から許可を得て公開している。
8	病原菌と糖鎖のデータベース(PACDB) http://jcggdb.jp/search/PACDB.cgi	公開	病原体と呼ばれる細菌、真菌、ウイルス、トキシンが糖鎖と結合する情報を論文から収集。
9	糖鎖関連遺伝子のデータベース(GGDB) http://riodb.ibase.aist.go.jp/rcmg/ggdb/	公開	糖鎖関連遺伝子と呼ばれている糖転移酵素、糖ヌクレオチドトランスポーター、硫酸転移酵素、糖ヌクレオチド合成酵素などの情報を収録。
10	レクチンと糖鎖の相互作用のデータベース(LfDB) http://riodb.ibase.aist.go.jp/rcmg/glycodb/LectinSearch	公開	レクチンの配列情報、生物種の情報、単糖特異性の情報をベースにし、個々のレクチン毎に１００種類を超える糖鎖との結合情報を計測したプロファイルのデータベース。
11	糖鎖のスペクトルデータベース(GMDB) http://riodb.ibase.aist.go.jp/rcmg/glycodb/Ms_ResultSearch	公開	標準糖鎖を MS/MS/MS/MS まで計測したスペクトル比較できるデータベース。 糖鎖構造によってスペクトルパターンが異なることから比較をすることで構造をある程度まで絞り込むことができる。
12	糖タンパク質データベース（GlycoProtDB）	公開	IGOT 法と呼ばれるＮ結合型糖鎖の付加位置を特定する方法を用いて糖

	http://riodb.ibase.aist.go.jp/rcmg/glycodb/Glc_ResultSearch		ペプチドを検出し、それらのデータをデータベース化したもの。現在は線虫のグライコプロテオーム解析であるが、間もなくマウスの情報が公開される。
13	腫瘍マーカーリファレンス DB (TuMaRDB)	未公開	臨床現場で利用されている腫瘍マーカーの情報をデータベース化した。適応と精度についての情報が掲載。
14	JCGGDB 単糖データベース(JCGG MSDB)	未公開	単糖の構造データベース。約４００種類の単糖を収録。
15	有機化学合成の化合物 DB	未公開	野口研究所で蓄積した情報と産総研の情報を集約する。データ移行の仕組みが完成している。検索システムを開発中。
16	糖鎖関連特許のデータベース	未公開	テキストマイニング中。類似検索ができるように開発中。
17	グリコシダーゼの基質特異性 DB	未公開	阻害剤 DB との連携作業。公開前の準備。
18	統合検索(GlycoExplorer)	未公開	キーワード検索から専門用語を選択するところから始まり、DB・特許・論文に共起した単語が連鎖的に出てくるインターフェースを開発中。

外部発表実績一覧

（1）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	

（2）展示会等出展

通番	タイトル	展示会等の名称	年月日	特記事項
1	JCGGDB の使い方	第 29 回日本糖質学会・飛騨・世界生活文化センター(岐阜県高山市)	2009 年 9 月 8 日～12 日	外部研究機関協力者： 山田　一作 川嵜　敏祐 八杉　悦子

（3）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)	2007 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	
3	「糖鎖の質量分析スペクトルデータベース：何ができるか、何が必要か？ 」 「糖鎖 MSn スペクトルデータベースの現状と展望」	亀山昭彦	第 56 回質量分析総合討論会	2008 年 5 月 14 日～16 日	
4	Publication and integration of glycodatabase. （ポスター）	新間陽一	全国糖生物学会（中国・蘇州）	2008 年 6 月 6 日～9 日	
5	日本糖鎖科学コンソーシアムデータベース （シンポジウム）	鹿内俊秀	日本糖質学会年会	2008 年 8 月 19 日	
6	糖鎖統合データベース講習・体験会(1)「GLYCOGENE DATABASE 入門」	鹿内俊秀、新間陽一	第１回 GLIT 勉強会（JBA）	2008 年 8 月 27 日	
7	糖鎖統合データベース講習・体験会(2)「LfDB とレクチンマイクロアレイ入門」	平林淳、舘野浩章	第３回 GLIT 勉強会（JBA）	2008 年 12 月 16 日	
8	糖鎖データベースの構築と統合 （ポスター）	新間陽一	BMB2008(第 31 回日本分子生物学会年会・第 81 回日本生化学学会大会合同大会	2008 年 12 月 9 日	
9	Integration of Glycoscience-related Database in Japan （ポスター）	鹿内俊秀	Clinical and Translational Research on Cancer:Glycomics Applications(HGPI,HUPO)	2009 年 3 月 24 日～27 日	
10	Japan Consortium for Glycobiology & Glycotechnology Database	木下聖子(依頼出張)、鹿内俊秀	EuroCarbDB meeting(ユトレヒト大学)	2008 年 11 月 27 日～29 日	
11	「糖鎖データベース」と「糖鎖産業技術フォーラムの具体的取り組み」	新間陽一	GLIT(産業技術総合研究所)	2008 年 1 月 23 日	
12	糖鎖データベースの構築と統合	鹿内俊秀	BioJapan2008	2008 年 10 月 15 日～17 日	
13	JCGGDB の概要と利用方法 （データベース講習会）	鹿内俊秀	第 3 回糖鎖産業技術フォーラム(GLIT) 産業技術総合研究所・臨海副都心センター	2009 年 5 月 12 日	
14	日本糖鎖科学統合データベース （ポスター発表）	鹿内俊秀	第 3 回糖鎖産業技術フォーラム(GLIT) 産業技術総合研究所・臨海副都心センター	2009 年 5 月 12 日	
15	日本糖鎖科学コンソーシアムデータベース ～	鹿内俊秀	データベースが拓くこれからのラ	2009 年 6 月 12 日	

	JCGGDB～　（ポスター発表）		イフサイエンス・東京大学・武田ホール（根津）		
16	日本糖鎖科学コンソーシアムのデータベース（ポスター発表）	鹿内俊秀	第29回日本糖質学会・飛騨・世界生活文化センター(岐阜県高山市)	2009年9月8日～12日	
17	日本糖鎖科学コンソーシアムのデータベース（ポスター発表）	鹿内俊秀	第82回日本生化学会大会・(神戸ポートアイランド)	2009年10月24日	
18	糖鎖構造に関連した実験データの横断検索(JCGGDB)　（ポスター発表）	鈴木芳典	第82回日本生化学会大会・(神戸ポートアイランド)	2009年10月24日	
19	日本糖鎖科学コンソーシアムの データベース構築（口頭発表）	鹿内俊秀	バイオジャパン2009	2009年10月7日	
20	ＪＣＧＧＤＢの公開したサービスと最終年度の計画　（口頭発表）	鹿内俊秀	第7回日本糖鎖科学コンソーシアム年会（大阪）	2009年12月8日	
21	Japan Consortium for Glycobiology and Glycotechnology DataBase (JCGGDB)　（口頭発表）	鹿内俊秀	1st Asia communication of glycobiology and glycotechnology	2009年10月31日	
22	Japan Consortium for Glycobiology and Glycotechnology DataBase　（ポスター発表）	鹿内俊秀	IUBMB（International Union of Biochemistry and Molecular Biology）(上海)	2009年8月2日～8日	
23	Japan Consortium for Glycobiology and Glycotechnology DataBase (JCGGDB)　（口頭発表）	鹿内俊秀	20th International Symposium of Glycoconjugates（プエルトリコ）	2009年11月28日～12月6日	
24	糖鎖構造の解析データの横断検索(JCGGDB)（ポスター発表）	鈴木芳典	産総研ライフサイエンス関連分野と産技連ライフサイエンス部会・バイオテクノロジー分科会の合同発表会LS-BT	2010年2月4日～2月5日	
25	病原体が接着する糖鎖と疾患のデータベース（PACDB）（ポスター発表）	鹿内俊秀	産総研ライフサイエンス関連分野と産技連ライフサイエンス部会・バイオテクノロジー分科会の合同発表会LS-BT	2010年2月4日～2月5日	
26	ここ10年の糖鎖研究の進歩とそのデータベース化によるさらなる飛躍（口頭発表）	成松久	学術会議シンポジウム「メタボロミクス研究の最前線とメタボライトデータベースの役割」	2010年1月15日	

区　　分	□中核機関（□代表機関／□参画機関） □分担機関（□代表機関／□参画機関） ■補完課題実施機関
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
実 施 機 関 名	国立遺伝学研究所
代 表 研 究 者 名	平成 19～20 年度 五條堀　孝、平成 21～22 年度　JST BIRD 事業「バイオ情報資源の高準化と共用化（代表研究者　菅原秀明）」の一課題として実施

１．課題開始時における達成目標

わが国における塩基配列決定における Trace データの保存と有効利用を目的として、当機関である情報・システム研究機構国立遺伝学研究所生命情報・DDBJ 研究センターの DDBJ が、Trace データのデータベース構築とデータ提供の業務を実施し、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へのデータ提供や連携を可能にする（当初、定量的目標は設定されていなかった）。

２．平成２２年１０月末時点における事業計画に対する成果

（１）成果概要

旧来のシーケンサー由来データを対象とする DDBJ Trace Archive (DTA)と次世代シーケンサー由来のデータを対象とする DDBJ Sequence Read Archive (DRA)の仕組みを確立し、実データの登録・蓄積・公開を進めた。図１に仕組みの整備進捗状況と DRA データ受付状況をまとめた。

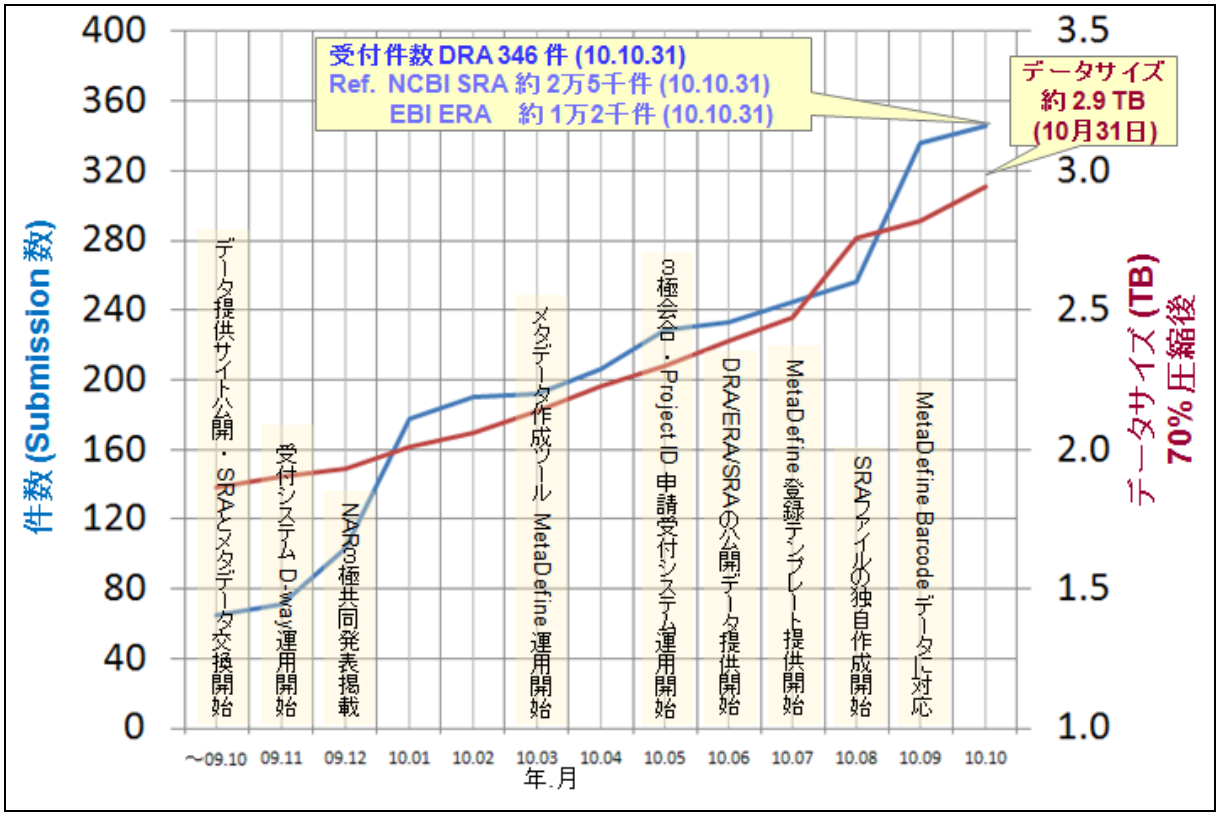


図１　DRA の進捗状況

2009 年 10 月から 2010 年 10 月の期間における DRA が受付けた件数とデータサイズの伸びに、DRA サービスの着実な展開を重ねて表示

（２）　進捗及び成果

1．主要な仕組み

DRAとDTAはhttp://trace.ddbj.nig.ac.jp/dra/index.shtmlから利用可能である（図２）。

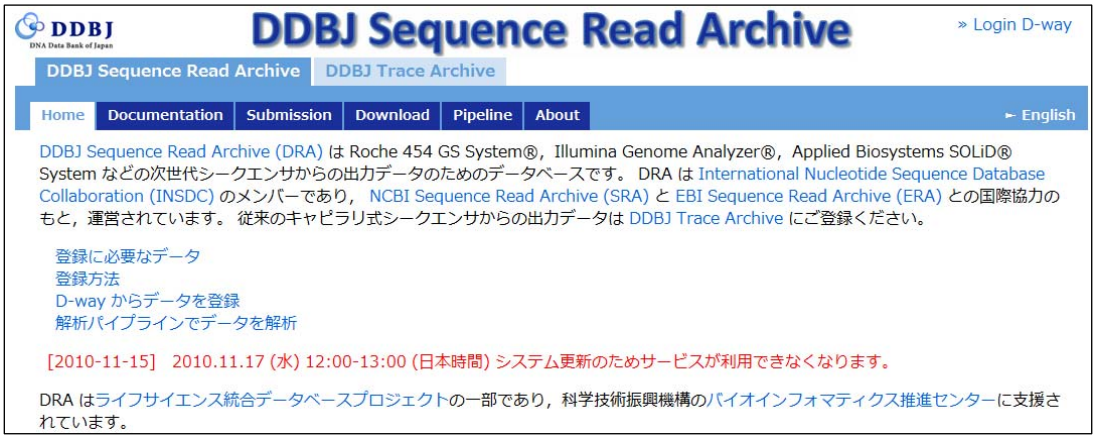


図２　DRA と DTA サイトのトップページ

Documentation のページにはメタデータの解説、データファイルの解説（Roche 454、Illumina Genome Analyzer、Life Technologies SOLiD System、および Helicos Heliscope）、および各種マニュアル（MetaDefine 操作、SRA Barcoding Guide、SRA ファイル形式、および SRA の XML スキーマ）を用意した。また、Submission のページには、概要、D-Way アカウントの申請・取得方法、メタデータの作成、登録ファイルの送付、ならびに登録の完了で構成される DRA 登録マニュアルを用意した。

1）登録受付管理システム D-way

D-way では登録を登録者のアカウントごとに管理している。登録者は D-way にログイン後、ウェブ上で新規登録の作成、進捗状況の確認や公開日の延長などができる。登録者はウェブ上で登録作業を効率よく進めることができる。図３に D-way の登録申請画面と登録状況の確認画面を紹介する。

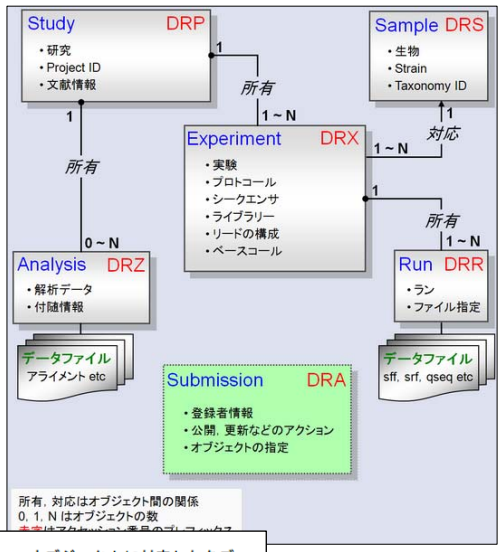
図３　D-way の登録申請画面と閲覧画面例

2）ウェブベースのメタデータ作成ツール MetaDefine

DRA のメタデータは図４に示すような構造をした DDBJ/EBI/NCBI 共通の XML を採用している。イ

ンフォマティクスに不慣れな登録者にとってメタデータを XML で作成することは大きな負担である。登録者は MetaDefine ウェブ画面上のボックスや項目に内容を入力していくだけで、XML を意識せずにメタデータを作成することができる。また、過去の登録内容や準備されているテンプレートを利用することで、必要な箇所の修正だけでメタデータを素早く作成することができる。

図５に示すように、MetaDefine では図４のメタデータがタブに展開されている。続く図６にはテンプレートの再利用を含む操作画面例を示した。



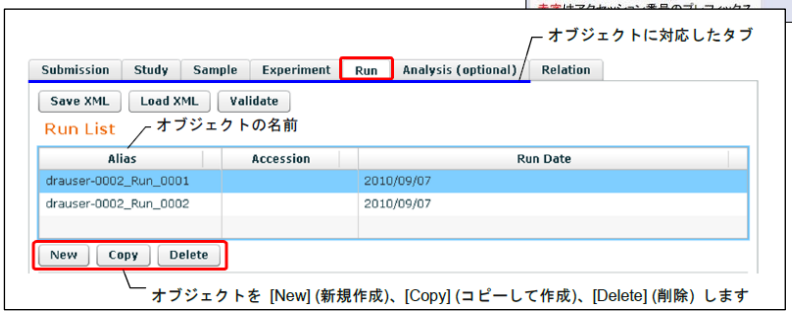


図５　MetaDefine ではメタデータが Submission、Study、Sample、Experiment、Run、および Analysis のタブとして展開される（MetaDefine マニュアルより）

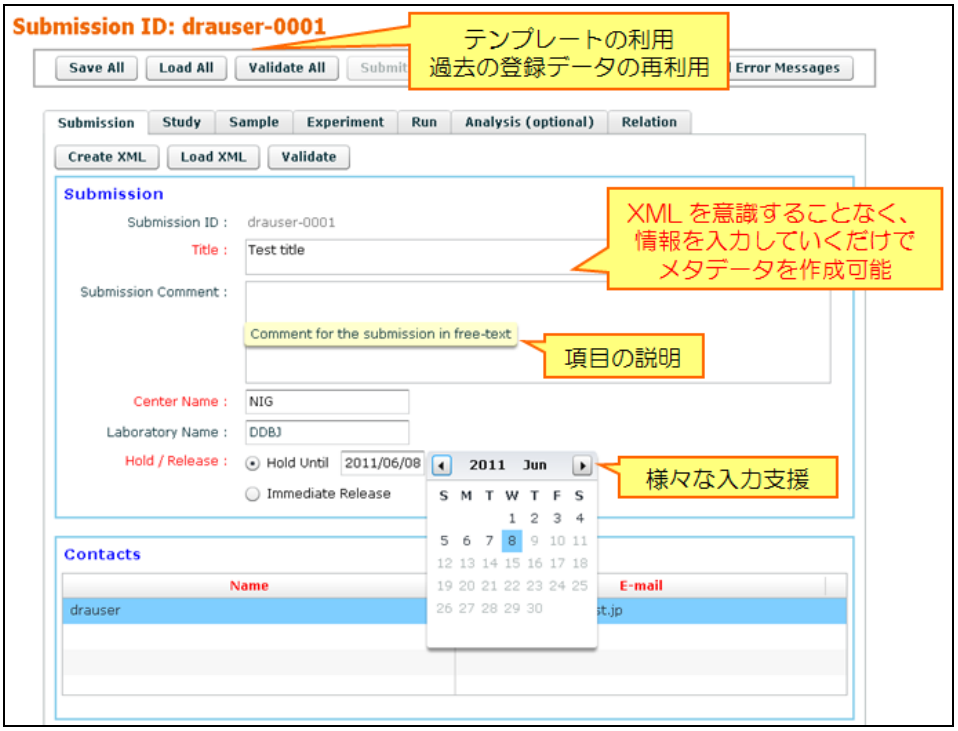


図６　MetaDefine の直感的に理解しやすい入力画面例

３）DRA 検索システム

図７に示すようなパイロット版を開発テスト中である。アクセッション番号、生物名、組織名、研究のカテゴリーとシーケンサーの種類での絞り込みが可能である。現在、キーワード検索機能を実装中で

ある。年度内にリリースし、ユーザのフィードバックを得ながらシステムの改良を進めていく。

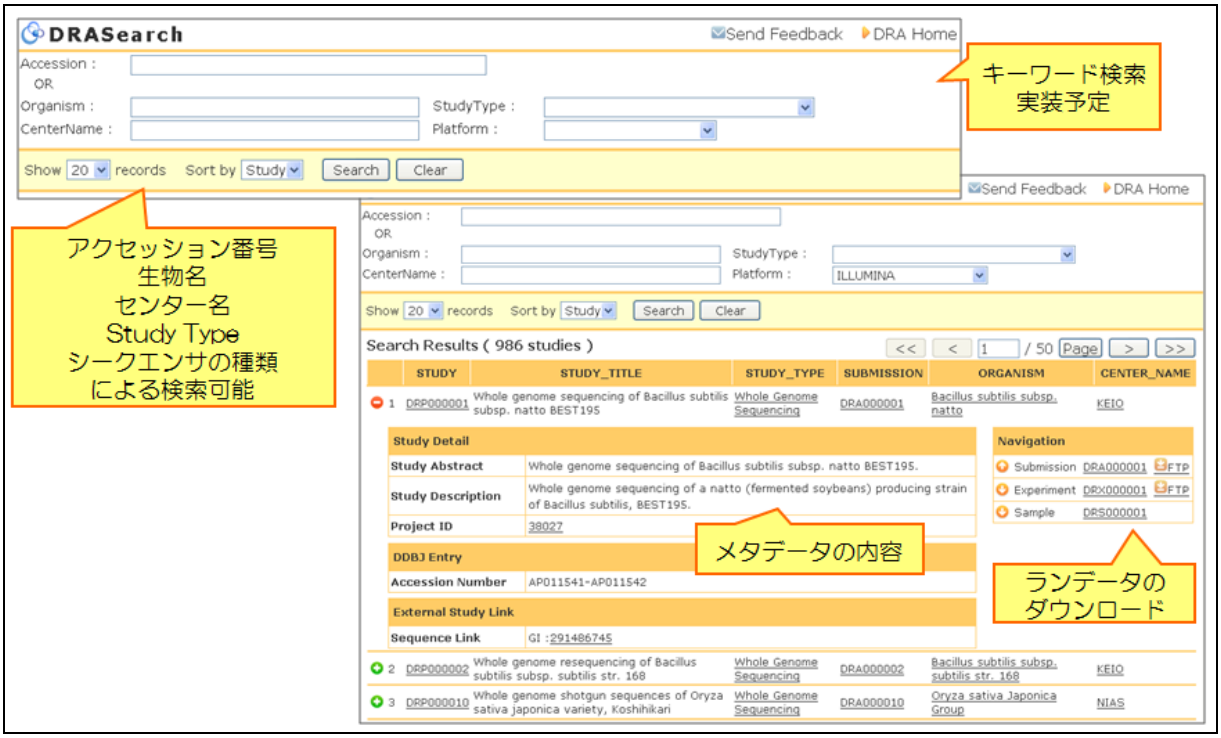


図７　DRA 検索システムパイロット版の画面例

Statistics

Released Entries

Type	Count
Submission	5158
Study	3272
Experiment	29620
Sample	15841
Run	71608

Organism

#	Organism Name	Study
1	Homo sapiens	292
2	metagenome sequence	169
3	Mus musculus	163
4	Drosophila melanogaster	121
5	marine metagenome	79
6	Caenorhabditis elegans	39
7	Arabidopsis thaliana	38
8	synthetic construct	37
9	Saccharomyces cerevisiae	35
10	Panicum virgatum	21

Study Type

#	Study Type	Study
1	Whole Genome Sequencing	1824
2	Transcriptome Analysis	591
3	Metagenomics	414
4	Epigenetics	242
5	Other	114
6	Resequencing	109
7	RNASeq	34
8	Population Genomics	31
9	Gene Regulation Study	24
10	Cancer Genomics	14

図８　DRA 検索システムパイロット版における検索対象データの統計
（2010 年 11 月 17 日現在）

２．仕組みの実装と運用の実績

１）DDBJ Trace Archive (DTA)

・　データ検索、配列、Quality Value、メタデータ、および波形表示システムを公開 (2009.8)。

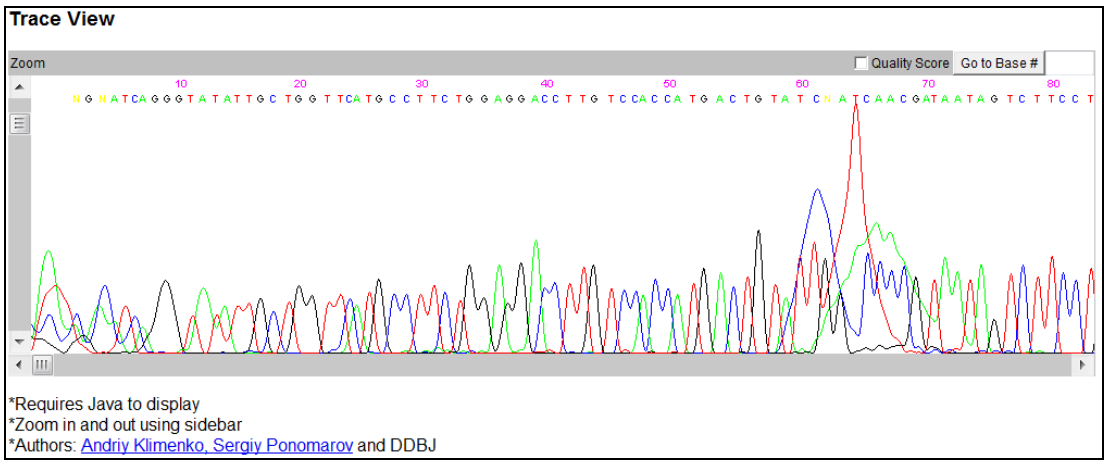


図９　DTA 登録データの波形表示例

・　登録件数 3 件　7,891,318 traces (約 140 GB) (2010.10)。

２）DDBJ Sequence Read Archive (DRA)

・　メタデータ作成支援簡易ツール DRA シートの運用を開始 (2009.9)。
・　データ内部管理システム DRA Submission Manager (DSM) 運用開始　(2009.9)。
・　米国 National Center for Biotechnology Information (NCBI) とデータ交換を開始した (2009.10)。
・　データ登録受付管理システム D-way の運用を開始 (2009.11)。
・　Nucleic Acids Research 誌に NCBI、欧州 European Bioinformatics Institute (EBI) と共同で Sequence Read Archive 国際事業に関する論文を発表した (2009.12)。
・　ウェブベースのメタデータ作成ツール MetaDefine の運用を開始した (2010.3)。
・　EBI で開催された DDBJ/EBI/NCBI 国際実務者会議において、国際データベース事業の運用について議論した。NCBI と EBI の要望に応じ MetaDefine のソースコードを提供した (2010.5)。
・　Project ID の申請システムを D-way に実装した (2010.5)。
・　DDBJ/EBI/NCBI SRA から公開されている全データの FTP 提供を開始した (2010.6)。
・　MetaDefine に登録テンプレートの提供機能を追加した (2010.7)。
・　MetaDefine にバーコードデータの作成機能を追加した (2010.9)。
・　研究の進展に対応すべく NCBI、欧州 EBI と共同でメタデータ XML スキーマ 1.2 を作成した (2010.9-10)。
・　DDBJ Sequence Read Archive で登録データを作成するシステムを整えた (2010.9)。
・　メタデータの検索、及び、それと連動したデータの FTP ダウンロードシステムの試験版を作成した (2010.9)。
・　Nucleic Acids Research 誌に DDBJ/EBI/NCBI と共同で Sequence Read Archive 国際事業の展開に関する論文を投稿し、受理された (2009.10)。
・　登録受付件数 346 件、データサイズ約 2.9 TB に達した (2010.10.31 時点)。
・　ウェブサイト・FTP サイトへのアクセス状況は以下のとおり:

DRA/DTA ウェブサイト

2010 年	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月

アクセス件数	1684	1453	1836	2703	2642	2475	2196	2972	2753	2914
バイト数（MB）	146	133	193	284	345	335	315	364	380	477

DRA/DTA FTP サイト

2010 年	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月
アクセス件数	468	682	333	594	440	1530	65399	10998	564	1629
バイト数（GB）	32	1	27	66	36	315	5155	3292	29	408

（注） 7～8 月は DRA 全件のダウンロードが繰り返し試みられた。

３．当初目標に対する達成度

旧来のシーケンサー由来データを対象とする DDBJ Trace Archive と次世代シーケンサー由来のデータを対象とする DDBJ Sequence Read Archive（DRA）の運用実績を重ねており、当初目標を十二分に達成している。

４．中間評価に対する対応

・中間評価「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」

平成 21 年 10 月に NCBI と Short Read Archive のメタデータ交換を開始して以来、次世代シーケンサー由来データの登録・蓄積・公開サービスを展開して、当該分野の技術・研究の急速な進展に対応している。

・中間評価「…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…」

中核機関におけるDRA統計カタログ (http://mars.dbcls.jp/sra/cgi-bin/studylist.cgi) の展開や 2010 年 12 月に分子生物学会ワークショップ「新型シーケンサーから得られるデータをどう解釈するか：統合データベースからの提案」を中核機関と協力して開催予定など、中核機関との連携を深めている。

５．他機関との連携

１）公的データベースとの連携

米国 NCBI、欧州 EBI と共に Sequence Read Archive と総称される国際共同データベース事業を運営している。他極とメール、ウェブ会議や年一回の国際実務者会議で議論することで連携している。

２）シークエンス拠点との連携

DRA と大型シークエンス拠点との間で大量データの転送手順を確立するため、理化学研究所横浜研究所、東京大学新領域創成科学研究科や国立遺伝学研究所などの拠点と話し合い、転送手順を確立した。

３）次世代シーケンサーベンダーとの連携

次世代シーケンサーの最新状況をフォローするため Roche、Illumina、Life Technologies 各社の日本法人担当者と話し合い、次世代シーケンサーの出力ファイル形式、データ生産能力や今後のアップグレードの予定などについて話し合った。また、DRA の目的や仕組みをベンダーに伝えた。

４）シークエンス受託解析会社との連携

次世代シーケンサーのコストダウンが進み、シークエンスを受託解析会社に依頼する研究者が増えて

いる。受託解析会社から顧客に DRA への登録に必要なデータを納品してもらうため、タカラバイオ社、北海道システムサイエンス社、および理研ジェネシス社と話し合った。

5）研究プロジェクトとの連携

大型の研究プロジェクトから産出される大量データを DRA に登録し、統一的なキーワード付与による検索を可能にするため、理化学研究所 FANTOM プロジェクト、文部科学省科学研究費新学術領域研究「ゲノム科学の総合的推進に向けた大規模ゲノム情報生産・高度情報解析支援」、および文部科学省「革新的細胞解析研究プログラム（セルイノベーション)」担当者と話し合った。

6）研究者への案内、講習

理化学研究所横浜研究所での次世代シーケンサー利用技術講習会 (全 3 回)、主要な学会での発表、DDBJing 講習会、DBCLS 講習会などでデータベースの説明と登録方法の講習を行った。

6．今後の見通し、計画、展望

シーケンシング技術の急速な進歩と普及がもたらすデータ急増とこれまでに存在しなかった新たなタイプのデータに対応していくことが求められている。このため、計算機資源、アプリケーションならびに運用方針までも、柔軟に拡張可能にしておく必要がある。具体的には以下の要件を念頭において、JST ライフサイエンス DB 統合推進事業においても、継続的な開発運用を目指したい：

- 計算機資源の継続的拡充と安定運用によって、研究コミュニティーの信頼に足るアーカイブを実現（特に、１PB～10PB 規模のデータ運用管理の実現）
- DRA/DTA と、米国 NCBI ならびに欧州 EBI のアーカイブとの間、さらに登録者と利用者との間における高速データ転送システムの構築運用
- 第３世代以降のシーケンサー由来データをアーカイブするアプリケーションの開発運用
- 大量リードデータの基礎的な解析サービスの提供
- 米国 NIH や英国 The Wellcome Trust のプロジェクトに比べて遅れている大量の未解析データを効率よく共有するための社会的なルールの整備
- 個人ゲノム情報の取り扱い指針の整備

7．全体総括

圧倒的に生産性が向上した次世代シーケンサーが普及した現在、データ指向（data driven）の研究を支援する情報環境が、研究の命運を握る。そして、研究者コミュニティーがデータの津波（TSUNAMI）を制御して再利用可能とするデータアーカイブこそ、この情報環境の要である。このような認識のもと DDBJ は次世代シーケンサーからの生出力データのための基盤データベース DDBJ Sequence Read Archive (DRA)を EBI ならびに NCBI との共同事業として開始した。

４年間のプロジェクト期間内にデータ登録の効率化、検索・データ提供システムの開発を進め、次世代シーケンサー由来データの受付・提供システムを整備した。登録数、ウェブサイト・FTP サイトへのアクセス数は順調に増えており、流通基盤として有効に機能していることが伺える。また、DDBJ が提供する他のデータベースならびに中核機関との連携も進めており、次世代シーケンサーからの解析前、解析途中、解析後のデータを漏らさずアーカイブしかつ解析を支援する情報環境整備が進んでいる。

8．特記事項

EBI/NCBI の Sequence Read Archive へのデータ登録では、登録者はメタデータを XML ファイルで作成しなければならず大きな負担となっている。DRA への登録では、登録者は MetaDefine を利用することで XML を意識せずに表形式でメタデータを作成することができる。EBI/NCBI からの MetaDefine を評価したいとの申し入れを受け、MetaDefine のソースコードを提供した。MetaDefine には Flex 技術を採用し、開発効率を高めている。MetaDefine 自身は XML をベースにしているが、XML スキーマの更新にも柔軟に対応できるという特徴がある。

DRA/DTA の構築にあたっては、項目５にまとめたように、多様な関係者との意思疎通を最優先し、運用も含めて実践的システムを構築できたと自負している。

９．委託研究費一覧

平成 21～22 年度は JST BIRD 事業「バイオ情報資源の高準化と共用化」における当該課題割当分。
また、当該課題固有の一般管理費は不明。 （千円）

	１８年度	１９年度	２０年度	２１年度	２２年度	計
設備備品費（千円）		34,000	1,000	0	0	35,000
人　件　費（千円）		1,830	19,144	13,385	4,749	39,108
業務実施費（千円）		9,625	7,129	16,615	20,751	54,120
一般管理費（千円）		4,545	2,727	NA	NA	7,272
合　計　（千円）		50,000	30,000	30,000	25,500	135,500

整備実績一覧

（1）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	http://trace.ddbj.nig.ac.jp/dra/index.shtml	公開	データベース名称：DDBJ Sequence Read Archive (DRA) 概要：次世代シーケンサー由来の生出力データのための公共データベース 進捗状況：2008 年 10 月データ受付開始。2009 年 7 月ウェブサイト公開。2009 年 10 月データの FTP 提供開始。2009 年 11 月登録受付管理システム D-way 運用開始。2010 年 3 月 MetaDefine 運用開始。登録受付ウェブシステム D-way: データ登録を登録者のアカウントごとに管理するシステム MetaDefine 運用開始。メタデータ作成支援ウェブツール提供 データの種類：登録された全ての生物種 数量：DRA 受付分 346 件 (約 2.9 TB) EBI/NCBI 由来データを含む総数 約 25,000 件 (約 120 TB)
2	http://trace.ddbj.nig.ac.jp/dta/dta_index.shtml	公開	データベース名称：DDBJ Trace Archive (DTA) 概要：キャピラリシーケンサーからの配列、Quality value と波形データのための公共データベース 進捗状況：2008 年 8 月データ受付公開。2009 年 7 月ウェブサイト公開。2010 年 8 月データの検索、波形データ表示システム公開。 データの種類：登録された全ての生物種 数量：DTA 受付分 3 件 7,891,318 traces (約 140 GB)
3	http://tracedev.ddbj.nig.ac.jp/DRASearch/	未公開	DRA データ検索、提供システム 2010 年 10 月パイロット版完成。アクセッション番号、生物名、登録した組織名、研究とシーケンサーの種類での検索機能を実装。

（2）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	D-way	公開	登録受付管理システム。ログイン後、ユーザは登録の新規作成、進捗状況・アクセッション番号の確認やデータの更新をウェブ上で行える。
2	MetaDefine	公開	メタデータ作成支援ウェブツール。画面に従って入力していくだけで XML を意識することなくメタデータを作成できる。
3	DRA 検索システム	未公開	DDBJ/EBI/NCBI Sequence Read Archive 由来の全てのメタデータを検索し、シークエンシングデータを

			FTP で取得できる。今後、キーワード検索機能を実装し公開する。
4	仕様書	公開	DDBJ/EBI/NCBI 共同でメタデータの XML スキーマを策定している。 http://trace.ddbj.nig.ac.jp/dra/documentation.shtml#schema

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	次世代シーケンサーデータの DRA (DDBJ Read Archive) への登録	児玉悠一	2009 年 12 月 16 日	理化学研究所（横浜研究所）	第 1 回シーケンサー利用技術講習会	
2	次世代シーケンサーデータの DRA (DDBJ Read Archive) への登録	児玉悠一	2010 年 2 月 17 日	理化学研究所（横浜研究所）	第２回シーケンサー利用技術講習会	
3	次世代シーケンサーデータの DRA (DDBJ Read Archive) への登録	児玉悠一	2010 年 7 月 15 日	理化学研究所（横浜研究所）	第３回シーケンサー利用技術講習会	
4	short read archive(SRA)登録紹介と short reads データ解析例の紹介	神沼英里 児玉悠一	2009 年 6 月 18 日	国立遺伝学研究所（三島）	第 21 回 DDBJing 講習会 in 三島	23 名（産 10、学５、官８）(官=産総研、NITE、県の機関) 馴染みのない内容であったが理解が深まった
5	次世代シーケンサー配列の登録・データ解析 ・ 次世代シーケンサーのクラウド型解析パイプライン ・ 次世代シーケンサーアーカイブ DB ・ クラウド型解析パイプライン・実習 assembly/mapping	神沼英里 児玉悠一 望月孝子		ライフサイエンス統合データベースセンター（東京）	AJACS ＆ 第 22 回 DDBJing 講習会 in 東京	28 名（産５、学７、官 16） NCBI の SRA 登録は非常に大変であり DRA は有り難い
6	「次世代シーケンサー由来データの公的アーカイブと利用・解析 - DDBJ Sequence Read Archive & DDBJ Pipeline - 」	児玉悠一	2010 年 9 月 10 日	霞山会館（東京）	ゲノムテクノロジー第 164 委員会 第 34 回研究会	委員会委員を中心に 139 名 今後の超大量データへの対応について質問があった

（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	遺伝研 DDBJ が次世代シーケンサー配列のクラウド型解析サービスを開始、まずはマッピングとアセンブリーをβ版で公開	日経バイオテクノロジージャパン	2010 年 3 月 25 日	

（3）展示会等出展

通番	タイトル	展示会等の名称	年月日	特記事項
1	日本 DNA データバンクの最新活動	第 32 回日本分子生物学会年会　特別企画「ナショナルバイオリソースプロジェクト」	2009 年 12 月 10 日	

（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	「DDBJ Sequence Read Archive」： 次世代シーケンサーからの出力データのためのアーカイブ	児玉悠一	第 32 回日本分子生物学会年会	2009 年 12 月 10 日	
2	DDBJ Read Archive (DRA) の紹介	児玉悠一	第 309 回 CBI 学会研究講演会	2010 年 6 月 21 日	
3	DDBJ Sequence Read Archive / DDBJ Omics Archive	児玉悠一	Fourth Biocuration Conference	2010 年 10 月 13 日	

（5）学術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	Archiving next generation sequencing data.	Shumway M, Cochrane G, Sugawara H	Nucleic Acids Res	38, D870-871	DDBJ/EBI/NCBI 共同論文
2	DDBJ launches a new archive database with analytical tools for next-generation sequence data.	Kaminuma E, Mashima J, Kodama Y, Gojobori T, Ogasawara O, Okubo K, Takagi T, Nakamura Y	Nucleic Acids Res	38, D33-38	
3	Biological Databases at DNA Data Bank of Japan in the Era of Next-Generation Sequencing Technologies.	Kodama Y, Kaminuma E, Saruhashi S, Ikeo K, Sugawara H, Tateno Y, Nakamura Y	Adv Exp Med Biol	680:125-135	
4	DDBJ progress report.	Kaminuma E, Kosuge T, Kodama Y, Aono H, Mashima J, Gojobori T, Sugawara H, Ogasawara O, Takagi T, Okubo K, Nakamura Y.	Nucleic Acids Res	doi: 10.1093/nar/gkq1041	
5	The sequence read archive.	Leinonen R, Sugawara H, Shumway M.	Nucleic Acids Res	doi: 10.1093/nar/gkq1019	DDBJ/EBI/NCBI 共同論文

区　　分	□中核機関（□代表機関／□参画機関） □分担機関（□代表機関／□参画機関） ■補完課題実施機関
課　　題　　名	生体分子の熱力学データと構造データの統合
実 施 機 関 名	国立大学法人九州工業大学
代 表 研 究 者 名	皿井明倫

１．課題開始時における達成目標

本研究では、情報・システム研究機構がすすめるデータベース統合化を補完するため、蛋白質の安定性や相互作用の網羅的な熱力学データを構造データと統合する。これにより、生体分子の機能に関する研究を促進する。熱力学データは、年間 700～1000 件を文献から収集しており、これらのデータと構造データの間を対応づけるためのクロスレファレンスを作成する。また、構造データベースを構築する PDBj とも連携して、XML などのデータ交換フォーマットの整備、オントロジーなどの統合化技術の開発を行う。さらに、情報・システム研究機構による統合検索との連携を可能にするために、情報・システム研究機構と連携して開発をすすめる。熱力学データは文献から収集しているため、この労力を軽減するため、情報・システム研究機構と協力して、テキストマイニングの技術を用いて文献の自動収集と文献からのデータの自動収集を行うシステムの開発をすすめる。

２．平成２２年１０月末時点における事業計画に対する成果

（１）成果概要

本研究では、以下の４つの項目について開発をすすめ、それぞれ次のような成果を得た。

① 蛋白質と変異体の熱力学データベースの構築と統合

本事業期間に発生した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3,000件と以前に収集したデータを合わせた約25,000件について、熱力学データと構造データを対応づけるクロスレファレンスを作成し両者を統合した。また、情報・システム研究機構による統合検索と連携するため、インデックス作成のための１次データの提供を行った。一方、情報・システム研究機構と連携して、テキストマイニング技術による文献の自動収集やデータの自動抽出法の開発を進め、熱力学データに特化した文献収集ツールを作成した。

② 蛋白質・核酸相互作用の熱力学データベースの構築と統合

本事業期間に発生した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3,000件と以前に収集したデータを合わせた約10,500件のうち、蛋白質・核酸複合体の構造データがあるものについて、熱力学データと構造データを対応づけるクロスレファレンスを作成し両者を統合した。また、統合検索のための１次データの提供を行った。文献の自動収集やデータの自動抽出法の開発では、情報・システム研究機構が開発しているテキストマイニングツール、TogoDocを共同で熱力学データ用にカスタマイズし、文献の自動収集を大幅に向上させることができた。

③ 蛋白質・蛋白質相互作用データの生成と統合

蛋白質･蛋白質相互作用データを格納するデータベースのプロトタイプを作成した。

④ XMLデータフォーマットやオントロジーなどの統合化技術の開発

熱力学データに関するオントロジーを整備するため、熱力学データのControlled Vocabularyの作成

を行った。また、蛋白質・核酸相互作用の熱力学データベースについて、これまではフラットデータのみを公開してきたが、フラットデータをXMLフォーマットのプロトタイプに変換するプログラムを作成し、XMLフォーマットでもデータを公開するようにした。

（2）進捗及び成果

本研究では、以下の４つの項目について開発をすすめ、それぞれ次のような成果を得た。

① 蛋白質と変異体の熱力学データベースの構築と統合

本事業期間に発生した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3,000件と以前に収集したデータを合わせた約 25,000 件について、熱力学データと構造データを対応づけるクロスレファレンステーブルを作成した。すなわち、熱力学データベースに記載された PDB の構造データ（PDBcode）およびそれと 100%同じ配列の PDBcode と対応するすべての熱力学データをリストしたテーブルと、配列の類似（95%以上の類似度）するすべての構造と対応するすべての熱力学データをリストしたテーブルを作成した。

蛋白質と変異体の熱力学データベースの検索画面とデータベースの内容については、それぞれ別紙参考資料１-(1)、１-(3)を参照。クロスレファレンステーブルについては、別紙参考資料１-(4)を参照。蛋白質と変異体の熱力学データベース、ProTherm、へのアクセスは年々増加しており、現在月当たり３千～４千件程度である。利用者層としては、企業からのアクセスが過半数を占めている。国別では、国内、アメリカ、欧州などからのアクセスが多い。本データベースはすでに公開から 12 年がたち、約 25,000 件のデータを保有しているので、我々が開発したデータベースの中では最もアクセスが多い。利用者は専門あるいは関連分野の研究者であるので、アクセスのほとんどは検索を行ってデータのページを参照している。このデータベースは、類似のものが存在しないため、この分野では重要なリソースとして世界中の研究者に利用されている。ProTherm はすでに 150 件以上の論文に引用され、データベースを用いて解析を行った研究論文も数多く出版されている。論文のリストは以下の URL を参照。

http://gibk26.bse.kyutech.ac.jp/jouhou/Protherm/ProTherm_References.htm

本研究が統合の対象としているデータベースの構築にあたっては、熱力学データの含まれている文献を収集し、論文を研究者が読んでデータの抽出を行っている。その後データの入力から照合・チェック、データベースへの登録までをほとんど手動で行っている。特に、研究者が論文からデータを抽出する部分が最も手間がかかる作業となっている。そこで、テキストマイニングの手法などを取り入れて、文献の自動収集や文献からのデータの自動抽出を行い、データベース構築の省力化を計ろうとしている。情報・システム研究機構では、そのような目的のためにテキストマイニングツール、TogoDoc や Wired-Marker を開発している。本研究では、これらのツールが我々の作業に応用できるかどうかの評価を行った。TogoDoc については類似文献検索、Wired-Marker ではテキストからの情報の自動抽出について我々のデータベース構築の有効性の観点から評価を行った。Wired-Marker については、HTML 形式での論文の表の任意の行と列からのデータの抽出、特殊文字の扱い、図からのデータ抽出、XML からの必要データの抽出、PDF の文献からのデータ抽出、などの方法について検討を行った。また TogoDoc については、まず、これまでに我々が収集したデータの記載された文献、蛋白質名、キーワードのリストや、マーキングした論文のサンプルなどをセンター側に提供した。これらの情報をもとに、TogoDoc において PubMed の related articles の機能を用いて類似文献を検索し、データを含む文献がどれだけヒットするかの検証を行った。

② 蛋白質・核酸相互作用の熱力学データベースの構築と統合

本事業期間に発生した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3,000件と以前に収集したデータを合わせた約 10,500 件のうち、蛋白質・核酸複合体の構造データがあるものについて、熱力学データと構造データを対応づけるクロスレファレンステーブルを作成した。すなわち、熱力学データベースに記載された PDB の構造データ（PDBcode）およびそれと 100%同じ配列の PDBcode と対応するすべての熱力学データをリストしたテーブルと、配列の類似（95%以上の類似度）するすべての構造と対応するすべての熱力学データをリストしたテーブルを作成した。

蛋白質・核酸相互作用の熱力学データベースの検索画面とデータベースの内容については、それぞれ別紙参考資料１-(2)、１-(3)を参照。クロスレファレンステーブルについては、別紙参考資料１-(5)を参照。蛋白質・核酸相互作用の熱力学データベース、ProNIT、へのアクセスは年々増加しており、現在月当たり２千～３千件程度である。利用者層としては、やはり企業からのアクセスが過半数を占める。国別では、国内、アメリカ、アジア、欧州などからのアクセスが多い。利用者は専門あるいは関連分野の研究者であるので、アクセスのほとんどは検索を行ってデータのページを参照している。このデータベースを利用してその成果が発表された論文のリストは以下の URL を参照。

http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit_ref.html

テキストマイニングについては、蛋白質と変異体の熱力学データベースの部分と同様、情報・システム研究機構と連携して、テキストマイニングツール、TogoDoc、Wired-Marker の評価を行った。詳細は、前節「蛋白質と変異体の熱力学データベースの構築と統合」の部分を参照。ProNIT に関しては、情報・システム研究機構と共同で独自の学習機能を備えた TogoDoc をカスタマイズし、これまでの正例の文献とともにデータを含まない負例の文献を与えて学習を行った。これにより、PubMed から熱力学データを含む文献の候補を以前より高い精度で予測できるようになった。このツールによりデータ収集の効率が大幅に改善された。

③ 蛋白質・蛋白質相互作用データの生成と統合

蛋白質･蛋白質相互作用データを格納するデータベースのプロトタイプを作成した。データベーススキーマを別紙参考資料２に示す。

④ XML データフォーマットやオントロジーなどの統合化技術の開発

熱力学データと構造データの統合を効率的にすすめるため、熱力学データに関するオントロジーについて調査を行った。生命情報に関してはすでに多くのドメインでオントロジーが整備されつつあるが、熱力学データに関するオントロジーはまだ整備されていないので、まず我々が構築している熱力学データベースについて Controlled Vocabulary の作成を行った（別紙参考資料３を参照）。オントロジー整備にあたっては、構造データベースの代表機関である PDBj や海外の関係機関とも意見交換を行った。

熱力学データベースではこれまではフラットデータのみを公開してきたが、平成２０年度にフラットデータを XML フォーマットのプロトタイプに変換するプログラムを作成し試験的に公開した。平成２１年度は、XML フォーマットを作成するプログラムを完成させ、XML フォーマットのデータを９月に完全公開した。さらに平成２１年度は、XML フォーマットデータをテキストフォーマットに変換するプログラムも作成し、テキストフォーマットデータも XML フォーマットと同時に公開した。XML データのサンプルは別紙参考資料４-(1)を参照。変換のステップは別紙参考資料４-(2)を参照。

なお、本プロジェクトに関する最新情報は、プロジェクト専用の Web ページ（別紙参考資料５を参照）を通して公開している。

３．当初目標に対する達成度

熱力学データは、蛋白質と変異体の熱力学データ及び蛋白質・核酸相互作用の熱力学データについて、これまでにそれぞれ年間 700～1000 件を文献から収集しており、これらのデータと構造データの間を対応づけるためのクロスレファレンスを作成し公開できた。また、蛋白質・核酸相互作用の熱力学データについては、XML フォーマットに変換し公開することができた。オントロジーの整備については、海外の研究機関を含め、他機関と協力しながらすすめている。さらに、情報・システム研究機構と連携し、テキストマイニングの技術を用いて蛋白質・核酸相互作用の熱力学データを含む文献の自動収集するシステムを開発し、文献収集の効率を大幅に改善することができた。熱力学データベースは、類似のものが存在しないため、この分野では重要なリソースとして世界中の研究者に利用されている。すでに 200 件以上の論文に引用され、データベースを用いて解析を行った研究論文も数多く出版されている。アクセスも年々増加しており、今回の統合化により研究への利用効率を改善することができた。

４．中間評価に対する対応

現在、研究の細分化や情報の多様化にともない、専門分野ごとのデータベース構築の需要が高まっており、このような中小規模のデータベースの構築を小グループで行うことができるようにすることが重要であるとの指摘があった。そこで本研究では、データベースキュレータの作業をできるだけ軽減するため、情報・システム研究機構と連携し、テキストマイニングの技術を用いて文献から該当データを含む文献を自動的に抽出する自動収集システムを開発した。これにより、文献収集の効率を大幅に改善することができた。さらに、研究室単位で統合データベースを構築するモデルケースとしての役割を果たすため、データベース構築、データ交換技術、統合化技術などをプロトコル化している。

５．他機関との連携

熱力学データと構造データの統合及びオントロジーの整備については、すでに構造データベースを構築・運営する PDBj とも連携してすすめている。また、データベース統合やオントロジー整備にあたっては、海外研究機関とも連携している。

一方、情報・システム研究機構がすすめている統合データ検索について熱力学データを提供するなどの連携を行った。さらに、文献の自動収集や文献からのデータの自動抽出に関して、情報・システム研究機構と共同で開発を行った。これまでの打ち合わせのリストは、別紙参考資料６を参照。

６．今後の見通し、計画、展望

熱力学データは定常的に生成されているので、今後とも継続してデータ収集を行いデータベース開発をすすめる。また、今後は生命科学の進展を見ながら、構造情報以外に相互作用データ、機能データ、変異データ、疾病データなどとも統合をすすめる。このような中小規模の専門データベースを研究室単位でも推進できるように、文献収集やデータ抽出の自動化技術の開発を情報・システム研究機構と連携してすすめたい。また、データの原著者によるデータ入力システムも開発したい。効率的なデータベー

ス構築に必要となる熱力学データのオントロジーはまだ十分整備されていないので、国内外の研究機関と連携して整備を継続する。一方、データベース利用者が容易にデータを研究に利用できるように、今後ともインターフェイスの改良などによる利便性の向上を計ってゆく。これまでに我々は、インターネットを用いたバイオインフォマティクスのトレーニングシステムを構築しワークショップを開催してきた（http://www.abren.net/workshop/）。今後、このシステムを利用して、データベース開発の人材育成、インターネットを利用したデータベースの共同開発などをすすめてゆきたい。

現在、研究の細分化や情報の多様化にともない、専門分野の知識は専門の研究者しか把握できないようになっている。したがって、専門分野に必要なデータベースの開発は専門研究者に頼らざるを得ない。しかし、専門研究者にとってデータベースの開発はハードルが高い。また、いったんデータベース開発を始めても継続することは極めて困難である。このような状況で、我が国から育って国際的に認知されているデータベースはごく限られている。そこで、新たな JST ライフサイエンス DB 統合推進事業では、大規模なデータベースだけでなく、中小規模の専門データベースの開発に対しても、資金的、人的、技術的支援がなされるような仕組みを構築することを期待する。我々の熱力学データベースに関しては、すでに当該分野では国際的に認知され利用されているので、今後は開発を継続できる仕組みの構築、例えば、データの収集・入力の自動化、データベース開発のプロトコル化、データベース開発の人材育成などを JST ライフサイエンス DB 統合推進事業と協力してすすめ、中小規模の専門データベース開発のモデルケースとしたい。

７．全体総括

本研究の主要目的である、「熱力学データと構造データの統合」に関しては、当初の目標をほぼ達成できたと考えている。一方、本研究が期待されている中小規模の専門データベース開発のモデルとしての役割に関しては、データ収集・入力の省力化について一部達成できたものの今後さらなる努力が必要である。ただ、モデルケースとしての方向性は示すことができた。これまでにこのような専門データベースが１０年以上にわたって開発を継続し国際的にも認知されている例はあまりないので、同様な専門データベース開発を促進する先例となったと考えている。今後ともぜひ継続し日本発の国際的なデータベースとして育てたい。

８．特記事項

我々の熱力学データベースはすでに開発から１０年以上たち、類似のものが存在しないため、この分野では重要なリソースとして世界中の研究者に利用されている。すでに 200 件以上の論文に引用され、データベースを用いて解析を行った研究論文も数多く出版されている。このことは、データベースは開発の継続ということがいかに重要かということを示している。また、国際的優位性を獲得するには、他でしていないことをできるだけ早く軌道に乗せて多くの人に使ってもらうということであろう。熱力学データは地味ではあるが、専門の研究者にとっては研究を進める上で必須である。我々も、最初は自らの研究の必要に迫られてデータベースをスタートさせた経緯がある。ただ、多くのデータベースは資金的・人的な理由から途中で挫折している。個人の力にはやはり限界があるので、データベースのようなインフラの整備と維持には国家的な支援システムの構築が望まれる。

９．委託研究費一覧						
	１８年度	１９年度	２０年度	２１年度	２２年度	計
設備備品費（千円）	0	400,000	0	0	0	400,000
人　件　費（千円）	0	2,353,125	5,649,368	6,215,135	6,347,088	20,564,716
業務実施費（千円）	0	2,696,875	2,206,087	4,628,320	925,640	10,456,922
一般管理費（千円）	0	545,000	785,545	1,084,345	727,272	3,142,162
合　計　（千円）	0	5,995,000	8,641,000	11,927,800	8,000,000	34,563,800

整備実績一覧

（1）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	蛋白質と変異体の熱力学データベース：ProTherm http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html	公開	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を収集し、1998 年より公開。これまでに、25,000 件のデータを収集。この熱力学データをもとに、本プロジェクトでは構造データと統合した。アクセス数は、現在月当たり３千～４千件程度である。今後とも、年間 1,000 件程度を収集する。人材の確保が課題。検索画面とデータ内容の詳細は、別紙参考資料１－(1)、１－(3)を参照。
2	蛋白質・核酸相互作用熱力学データベース：ProNIT http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html	公開	蛋白質と核酸の相互作用の定量的な熱力学データを収集し、2000 年より公開。これまでに、10,500 件のデータを収集。この熱力学データをもとに、本プロジェクトでは構造データと統合した。アクセス数は、現在月当たり２千～３千件程度である。今後とも、年間 1,000 件程度を収集する。人材の確保が課題。検索画面とデータ内容の詳細は、別紙参考資料１－(2)、１－(3)を参照。

（2）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	熱力学オントロジー	公開	我々が構築している熱力学データベースに含まれる熱力学用語について Controlled Vocabulary を作成した。今後、さらに一般的な熱力学オントロジーの整備を行う。Controlled Vocabulary については、別紙参考資料３を参照。

外部発表実績一覧

（1）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	生体分子の熱力学データと構造データの統合	皿井明倫、Shaji Kumar	ライフサイエンスの未来へ～10 年先のデータベースを考える～	2010 年 10 月 5 日	

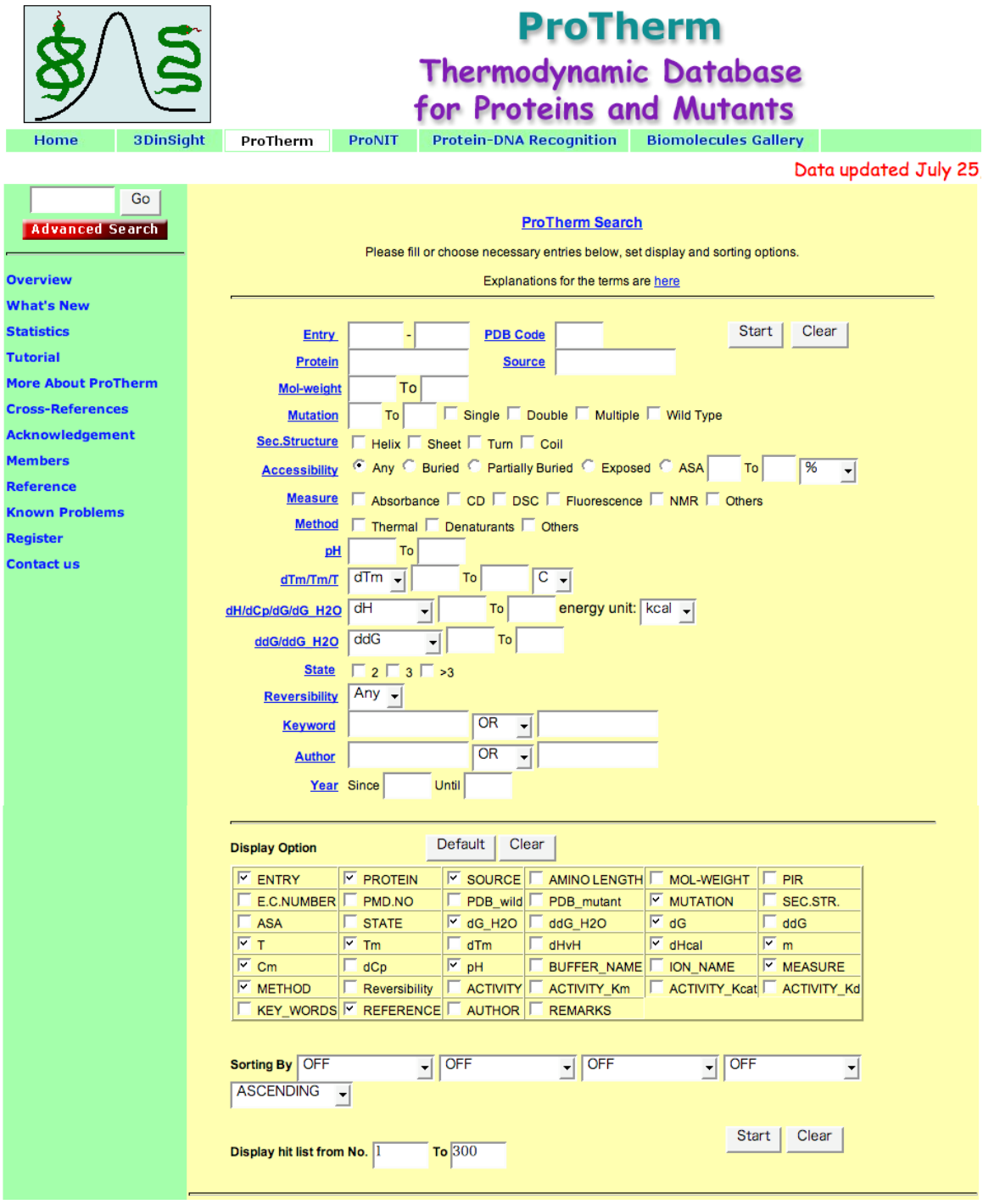
2	生体分子の熱力学データと構造データの統合	皿井明倫、Shaji Kumar	文科省統合データベースプロジェクト シンポジウム「データベースが拓くこれからのライフサイエンス」	2009 年 6 月 12 日	
3	Integration of Thermodynamic and Structural Data	皿井明倫	生物物理学会	2008 年 12 月 4 日	
4	ProNIT: Database Development and Integration	S. Kumar, P. Prabakaran, M. Gromiha, H. Uedaira, K. Kitajima, and A. Sarai	日本バイオインフォマティクス 学会年会	2008 年 12 月 15 日 ～16 日	
5	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）学術雑誌等への論文寄稿

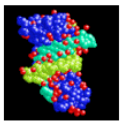
通番	タイトル	著者名	雑誌等の名称	掲載巻、号、ページ	特記事項
1	Thermodynamic Database for Proteins: Features and Applications	M. M. Gromiha and A. Sarai	Methods Mol. Biol.	609, 97-112 (2010)	

別紙参考資料１

（１）蛋白質熱力学データベース ProTherm の検索画面



（２）蛋白質・核酸相互作用熱力学データベース ProNIT の検索画面



ProNIT
Thermodynamic Database for Protein-Nucleic Acid Interactions

Home | 3DinSight | ProTherm | ProNIT | Protein-DNA Recognition | Biomolecules Gallery

Last Update: 28-Aug.-2007, NEW RELEASE: ProNIT 2.0

Quick Search

- ProNIT Home
- What's New
- About ProNIT
- Release Notes
- Statistics
- Tutorial
- Members
- Reference
- Contact us
- Copyright
- Acknowledgement

Welcome to ProNIT Database

Advanced Search

Please fill or choose necessary entries below, set display and sorting options.

ProNIT Entry - Search Reset Entry List
Protein Name Protein List
Protein Source Source List
PDB_Code PDB List
Protein Sequence
Mutation To Single Double Multiple Wild
Sec.Str Helix Sheet Turn Coil
ASA_Free To %
Nucleic Acid Name
Nucleic Acid Source
Nucleic Acid Type dsDNA ssDNA RNA
NDB_Code NDB List
Nucleic Acid Sequence
Method 1.Calorimetry 2.Footprint 3.Filter binding 4.Fluorescence
T To °C
pH To
Kd x 10 To x 10 M
dG To kcal/mol
dH To kcal/mol
dCp To kcal/mol/K
Author AND Author List
Year To Reference List
Keywords AND

Display Options

☑ Protein Name	☑ Protein Source	☐ Biological_unit	☐ Fragment	☐ E.C.Number
☐ PIR_No.	☐ SWISSPROT_NO	☐ PDB_Free	☑ PDB_Complex	☐ NDB_Complex
☑ Mutation protein	☐ ASA_Free	☐ ASA_Complex	☐ ProTherm_No.	☐ Sec_Str
☐ Nucleic Acid Name	☐ Nucleic Acid Source	☐ Nucleic Acid Type	☐ GenBank_No.	☐ Complex_DB_No
☐ Ligand	☑ T	☑ pH	☐ Buffer_Name	☐ Buffer_Conc
☐ Additives	☐ Ion Name	☐ Ion_Conc	☑ Method	☑ Kd_Wild
☐ Kd_Mutant	☐ Ka_Wild	☐ Ka_Mutant	☑ dG_Wild	☐ dG_Mutant
☐ dH_Wild	☐ dH_Mutant	☐ dCp_Wild	☐ dCp_Mutant	☐ Stoichiometry
☐ Activity_Km	☐ Activity_Kcat	☐ Author	☑ Reference	

Sorting Options

Priority 1	Priority 2	Priority 3	Priority 4	Order
OFF	OFF	OFF	OFF	ASCENDING

Entries per page : 50 Search Clear

（３）熱力学データベースの内容

①蛋白質熱力学データベース ProTherm に含まれる主な内容は以下のようである。蛋白質情報:名前、由来種、対応する配列や構造の ID、天然状態における集合数など。変異情報：変異アミノ酸とその位置、２次構造と Accessible Surface Area (ASA)など。実験情報：測定方法や、温度、pH、バッファー、イオン、蛋白質濃度などの実験条件。熱力学データ:熱変性の場合、変性の自由エネルギー変化（ΔG）、エンタルピー変化（ΔH）、熱容量変化（ΔC_p）、変性温度（T_m）、変性の可逆性、変性剤変性の場合、変性剤濃度ゼロに外挿した変性自由エネルギー変化（ΔG^{H2O}）、変性曲線の傾き（m）と変性中点の変性剤濃度（C_m）など。その他の情報:酵素活性値（K_m、k_{cat}）、解離定数（K_d）、転移の状態数。文献情報：ジャーナル名、著者名、出版年、キーワード、リマークなど。

②蛋白質・核酸相互作用熱力学データベース ProNIT に含まれる主な内容は以下のようである。蛋白質情報: 名前、由来種、対応する配列や構造の ID など。アミノ酸変異情報:変異アミノ酸とその位置、２次構造と ASA など。核酸情報：名前、由来種、対応する配列や構造などの ID。塩基変異情報：変異塩基とその位置。複合体情報:複合体構造の ID、複合体形成に伴う構造変化などの記述。実験情報:測定方法や、温度、pH、バッファー、イオン、蛋白質濃度などの実験条件。熱力学データ：解離定数（K_d）、結合の自由エネルギー変化（ΔG）、エンタルピー変化（ΔH）、熱容量変化（ΔC_p）、結合の stoichiometry。その他の情報：酵素活性値（K_m、k_{cat}）。文献情報：ジャーナル名、著者名、出版年、キーワード、リマークなど。

（4）クロスレファレンステーブル。ProTherm データベースに含まれる構造データの PDBcode と 100%同じ配列に対応する蛋白質の熱力学データの対応表の一部。

Cross-Reference to PDB

PDB_ID	ProTherm_EntryNo
1BAL	18200, 18201, 18202, 18203, 18204, 18205, 18206, 18207, 18208, 18209
2TCT	6413, 6414
1IKL	23670, 23671, 23672, 23673, 23674, 23675
1PUC	10219, 10220, 10221, 10222, 10223, 10224, 10225, 10226, 10227, 10228, 10229, 10230, 10231, 10232, 10233, 10234, 10235
1TIO	7485, 7486, 7487, 7488, 7489, 9882, 16833
1IKM	23670, 23671, 23672, 23673, 23674, 23675
1BAV	15105, 15106
1TIU	15280, 15281, 17002, 17003, 17013, 17014, 17628, 17629, 17630, 17631, 17632, 17633, 17634
1F6H	248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 328, 329, 330, 331, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2812, 2813, 2814, 2815, 2816, 2817, 2818, 2819, 2820, 2821, 2822, 2823, 2824, 2825, 2826, 2827, 2828, 2829, 4545, 4546, 5899, 5900, 5901, 5902, 5903, 5904, 5905, 5906, 5907, 5908, 5909, 5910, 5911, 5912, 6215, 6216, 6503, 6504, 6505, 6506, 6507, 6508, 6509, 6510, 6511, 6512, 6513, 6514, 6515, 6516, 6517, 6518, 6519, 6520, 6521, 6522, 6523, 6524, 6525, 6526, 6527, 6528, 6529, 6530, 6531, 6532, 6533, 6534, 6535, 6536, 6537, 6538, 6539, 6540, 6541, 6542, 6543, 6544, 6545, 6546, 7408, 7409, 7410, 7411, 7493, 7494, 7495, 7496, 7497, 7498, 7499, 7500, 7501, 7502, 7503, 7504, 7505, 7506, 7507, 7508, 7509, 7510, 7511, 7512, 7513, 7514, 7847, 10124, 10139, 10140, 10141, 10142, 10143, 10144, 10145, 10146, 10147, 10148, 10149, 10150, 10151, 10152, 10153, 10154, 10155, 10156, 10157, 10158, 10159, 10160, 11133, 11134, 11135, 11136, 11137, 11138, 11139, 11140, 11141, 11142, 11143, 11144, 11145, 11146, 11147, 11148, 11149, 11150, 11151, 11152, 11153, 11154, 11155, 11156, 11157, 11158, 11159, 11160, 11233, 12362, 13271, 13272, 13273, 13274, 13275, 13276, 13277, 13278, 13279, 13280, 13281, 13282, 13283, 13284, 13285, 13286, 13287, 13288, 13289, 13290, 13291, 13292, 13293, 13328, 13329, 13330, 13331, 13332, 13333, 13334, 13335, 13336, 13337, 13338, 13339, 13340, 13341, 13342, 14518, 14519, 14520, 14521, 14522,

（５）クロスレファレンステーブル。ProNIT データベースに含まれる構造データの PDBcode と 100%同じ配列に対応する蛋白質・核酸相互作用の熱力学データの対応表の一部。

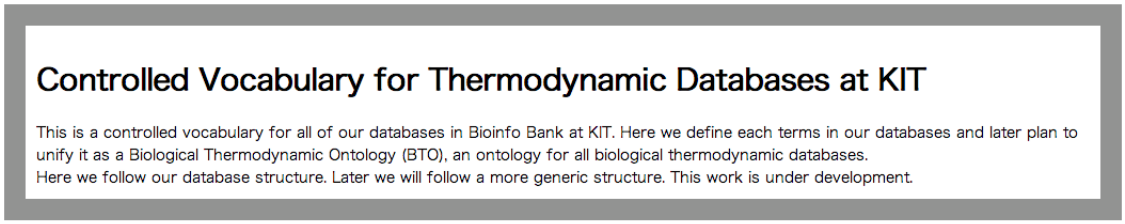
1A1L	2719, 2720, 2721, 2722, 2723, 2724, 2744, 2745, 2746, 2747, 2748, 2749, 2750, 2751, 2752, 2753, 2754, 2755, 2756, 2757, 2758, 2759, 3289, 3290, 3291, 3292, 3293, 3294, 3295, 3296, 3297, 3298, 3299, 6303, 6306
1A28	5959, 5963
1A3C	5592, 5593, 5594, 5595, 5596, 5597, 5598, 5599, 5600, 5601, 5602, 5603, 5604, 5605, 5606, 5607, 5608, 5609, 5610, 5611, 5612, 5613, 5614, 5615, 5616, 5617, 5618, 5619, 5620, 5621, 5622, 5623, 5624, 5625, 5626, 5627, 5628, 5629, 5630, 5631, 5632, 5633, 5634
1A41	6117, 6118, 6119, 6120, 6121, 6122, 6123, 6124, 6125, 6126, 6127, 6128, 6129, 6130, 6131, 6132, 6133, 6134, 6135, 6136, 6137, 6138, 6139, 6140, 6141, 6142, 6969, 6970, 6971, 6972, 6973, 6974, 6975, 6976, 6977, 6978, 6979, 6980, 6981, 6982, 6983, 6984, 6985, 6986, 6987, 6988, 6989, 6990, 6991, 6992, 6993, 6994, 6995, 6996, 6997, 6998, 6999, 7000, 7001, 7002, 7003, 7004
1A43	8505, 8509, 8512, 8515, 8517, 8519, 8521, 8523, 8525
1A4T	7665, 7666, 7667, 7668, 7669, 7670, 7671, 7672, 7673, 7674, 7675, 7676, 7677, 7678, 7699, 7700, 7701, 7702, 7731, 7732, 7795, 7796, 7797, 7798, 7799, 7800, 7809, 7810, 7811, 7812, 7813, 7814, 7815, 7816
1A4X	5592, 5593, 5594, 5595, 5596, 5597, 5598, 5599, 5600, 5601, 5602, 5603, 5604, 5605, 5606, 5607, 5608, 5609, 5610, 5611, 5612, 5613, 5614, 5615, 5616, 5617, 5618, 5619, 5620, 5621, 5622, 5623, 5624, 5625, 5626, 5627, 5628, 5629, 5630, 5631, 5632, 5633, 5634
1A73	1227, 1228, 1229, 1230, 1231, 1232, 1233, 1234, 1235, 1236, 1237, 1238, 1239, 3364, 3365
1A74	1227, 1228, 1229, 1230, 1231, 1232, 1233, 1234, 1235, 1236, 1237, 1238, 1239, 3364, 3365
1AAB	3173, 3174, 3175, 3176, 3177, 3178, 3179, 3180, 3181, 3182, 3183, 3184, 3185, 3186, 3187, 3188, 3189, 3190, 3191, 3192, 3193, 3194, 3195, 3196, 3197, 3198, 3199, 3200, 3201, 3202, 3203, 3204, 3242, 3243, 3244, 3245, 3246, 3247, 3248, 3249, 3250, 3251, 3252, 3434, 3435, 3436, 3437, 3438, 3439, 3440, 3441
1AAY	2719, 2720, 2721, 2722, 2723, 2724, 2744, 2745, 2746, 2747, 2748, 2749, 2750, 2751, 2752, 2753, 2754, 2755, 2756, 2757, 2758, 2759, 3289, 3290, 3291, 3292, 3293, 3294, 3295, 3296, 3297, 3298, 3299, 6303, 6306
1AF5	3366, 3367, 3368, 3369, 3370, 3371, 3372
1AHD	5938, 5939, 5944, 5945, 5950, 5951
1AIE	1919, 1920, 1921, 1922, 1923, 1924, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 750, 751, 752, 753, 754
1AIS	1557, 1558, 1559, 1560, 1561, 1562, 1563, 1564, 1565, 1566, 1567, 1568, 1569, 1570, 1571, 1572, 1573, 1574, 1575, 1576, 1577, 1578, 1579, 1580, 1581, 1582, 1583, 1584, 1585, 1586, 1587, 1588, 1589, 1590, 1591, 1592, 1593, 1594, 1595, 1596, 1597, 1598, 1599, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 4404, 4405, 4406, 4407, 4408, 4409, 4410, 4411, 4412, 4413, 4414, 4415, 4416, 4417, 4418, 4419, 4420, 4421, 4422, 4423, 4424, 4425,

別紙参考資料２
蛋白質･蛋白質相互作用データベースのプロトタイプ

entrynumber
protein information:
protein1 name
protein1 synonyms
protein1 source
protein1 sequence
protein1 biologicalunit
protein1 uniprot
protein1 pdb
protein1 mutation
protein1 asa
protein1 secstr
protein1 prothermnumber:
protein2 name
protein2 synonyms
protein2 source
protein2 sequence
protein2 biologicalunit
protein2 uniprot
protein2 pdb
protein2 mutation
protein2 asa
protein2 secstr
protein2 prothermnumber
Complex information:
pdb complex
ligand
conformation
Experimental Condition:
Temperature
pH
BufferName
BufferConcentration
Additives
protein1 Concentration
protein2 Concentration
Ion_Name
Ion_Concentration
Method
Binding Data:
Kd
Kd Mutant
Ka
Ka_Mutant
dG
dG Mutant
dH
dH Mutant
dCp
dCp_Mutant
Stoichiometry
Literature:
Reference
Author
Keywords
Remarks
RelatedEntries

別紙参考資料３

熱力学データの Controlled Vocabulary



Controlled Vocabulary for Thermodynamic Databases

Expand All　Collapse All

- ProNIT
  - Protein Information
  - Nucleic Acid Information
  - Complex Information
  - Experimental Details
  - Binding Data
  - Literature
  - General
- ProTherm
  - Sequence and Structural Information
  - Experimental Details
  - Thermodynamic Data
  - Literature
  - General

別紙参考資料４

（１）XML フォーマットによる ProNIT データの一部

```
<pronit xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:noNamespaceSchemaLocation="file:pronit.xsd">
 <entry>
  <entryNumber>1</entryNumber>
  <proteinDetails>
   <name>Myb proto-oncogene protein</name>
   <synonyms>c-Myb protein; Transforming protein myb</synonyms>
   <source>Mus musculus (Mouse)</source>
   <fragment>89-193</fragment>
   <sequence>MARRPRHSIYSSDEDDEDIEMCDHDYDGLLPK</sequence>
   <biologicalUnit>1</biologicalUnit>
   <dbReference type="PIR">TVMSMB</dbReference>
   <dbReference type="SWISSPROT">MYB_MOUSE (P06876)</dbReference>
   <dbReference type="PDBFREE">1MBE, 1MBG, 1MBJ</dbReference>
   <dbReference type="PROTHERM">786 787 788 789</dbReference>
   <mutation>wild</mutation>
  </proteinDetails>
  <nucleicAcidDetails>
   <name>MBS-I (22-mer)</name>
   <source>Synthetic</source>
   <type>DDS</type>
   <sequenceTopWild>caccctaactgacacacattct</sequenceTopWild>
   <sequenceBottomWild>agaatgtgtgtcagttagggtg</sequenceBottomWild>
   <mutation>wild</mutation>
   <dbReference type="GENBANK">M33654</dbReference>
  </nucleicAcidDetails>
  <complexDetails>
   <dbReference type="PDBCOMPLEX">1MSE</dbReference>
   <dbReference type="PRONUC">86</dbReference>
   <proteinConformation>R2 and R3</proteinConformation>
   <nucAcidConformation>The base pairs </nucAcidConformation>
  </complexDetails>
  <exptDetails>
   <temperature>20.2 C</temperature>
   <pH>7.5</pH>
   <bufferName>Potassium phosphate</bufferName>
   <bufferConc>100 mM</bufferConc>
   <ionNameOne>Potassium chloride (KCl)</ionNameOne>
   <ionConcOne>20 mM</ionConcOne>
   <method>Isothermal titration calorimetry (ITC)</method>
  </exptDetails>
  <bindingData>
   <Kd_Wild>5.00e-08 M</Kd_Wild>
   <Ka_Wild>2.00e+07 1/M</Ka_Wild>
   <dG_Wild>-1.21e+01 kcal/mol</dG_Wild>
   <dH_Wild>-1.25e+01 kcal/mol</dH_Wild>
   <dCp_Wild>-6.20e-01 kcal/mol/K</dCp_Wild>
   <stoichiometry>1.01</stoichiometry>
  </bindingData>
  <citation>
   <reference>J Mol Biol. 1998; 276(3):571-590 PMID: 9551098</reference>
   <author>Oda M, Furukawa K, Ogata K, Sarai A, Nakamura H</author>
   <keywords>c-Myb; DNA-binding; ITC</keywords>
  </citation>
  <miscellaneous>
   <remarks>R2R3* (C 130 I), a stable mutant</remarks>
   <relatedEntries>2, 3, 4, 5, 6, 7, 8, 9, 10, 11</relatedEntries>
  </miscellaneous>
 </entry>
 <copyright>copyright</copyright>
</pronit>
```

（2）フラットフォーマットから XML フォーマットに変換する手順

１．フラットファイルからデータ構造の情報を取得

２．この情報から XML スキーマを定義

３．XML スキーマの情報をプログラムにインプリメント

４．フラットファイルのすべての項目の読み込みと修正

５．フラットファイルから HTML 情報部分を削除

６．フラットファイルから項目とデータ列を分離

７．各列を読み込み CVS ファイルに書き出し

８．CVS ファイルを読み込みブランク部位は削除

９．CVS ファイルからデータを抽出し XML スキーマに従って XML ファイルを生成

１０．XML 構文をチェックし XML ファイルをバリデート

1. Get the data structure from the flat file
2. Define the XML schema based on this structure
3. Incorporate the XML schema structure information into the program
4. Read the flat file and check all the fields. If there is any error, correct it.
5. Remove the HTML content from the flat file
6. The flat file contains two columns. Split the columns
7. Read the columns and convert the column data into a CSV file
8. Read the CSV file and remove the fields having null data
9. Extract the data from the CSV file and create the XML file as per the XML Schema
10. Verify the XML syntax and validate the XML file.

別紙参考資料５

九工大補完課題プロジェクト専用の Web ページ