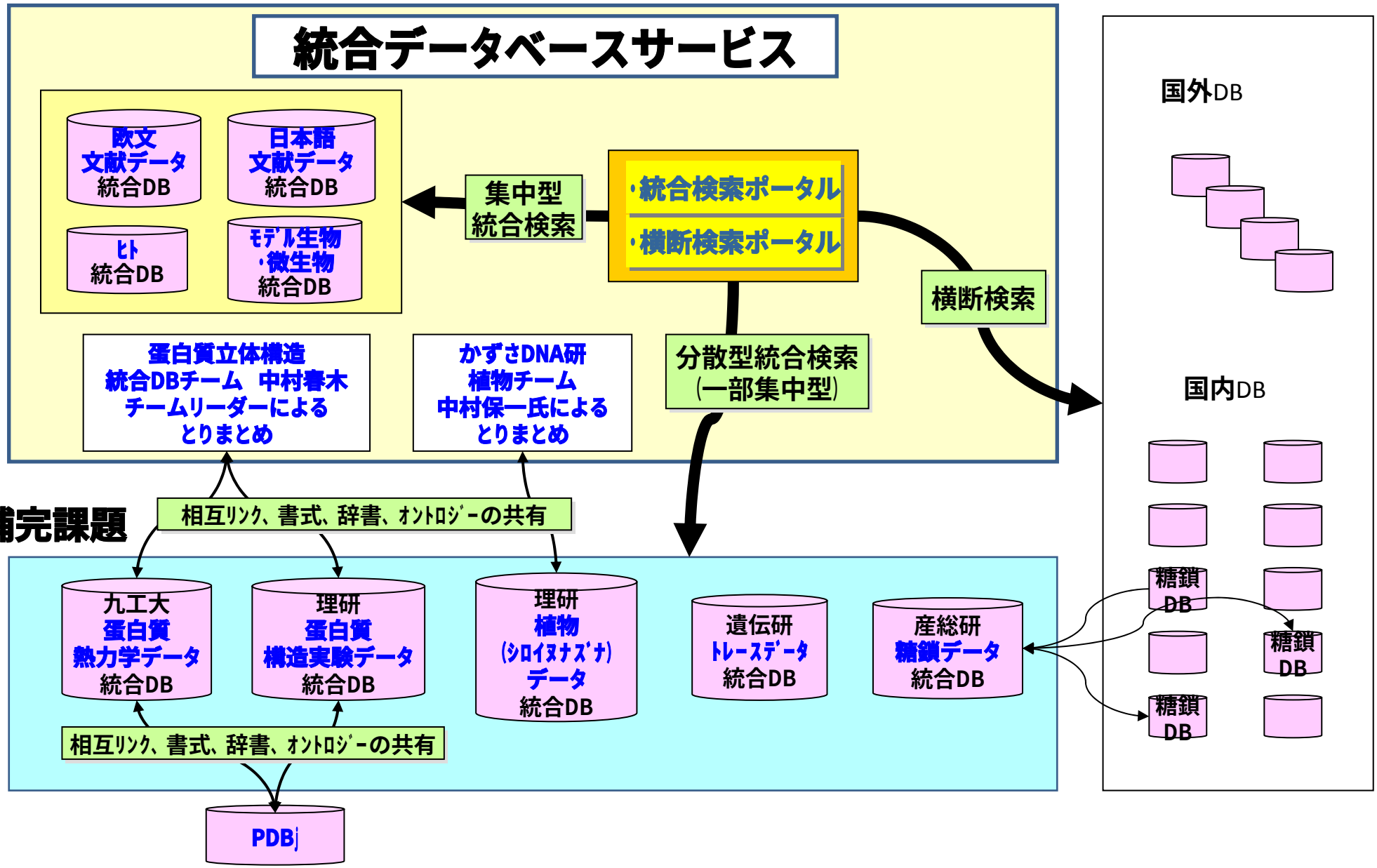
資料4－2

統合データベースプロジェクト
研究運営委員会（第3回）作業部会報告
補完課題の進め方

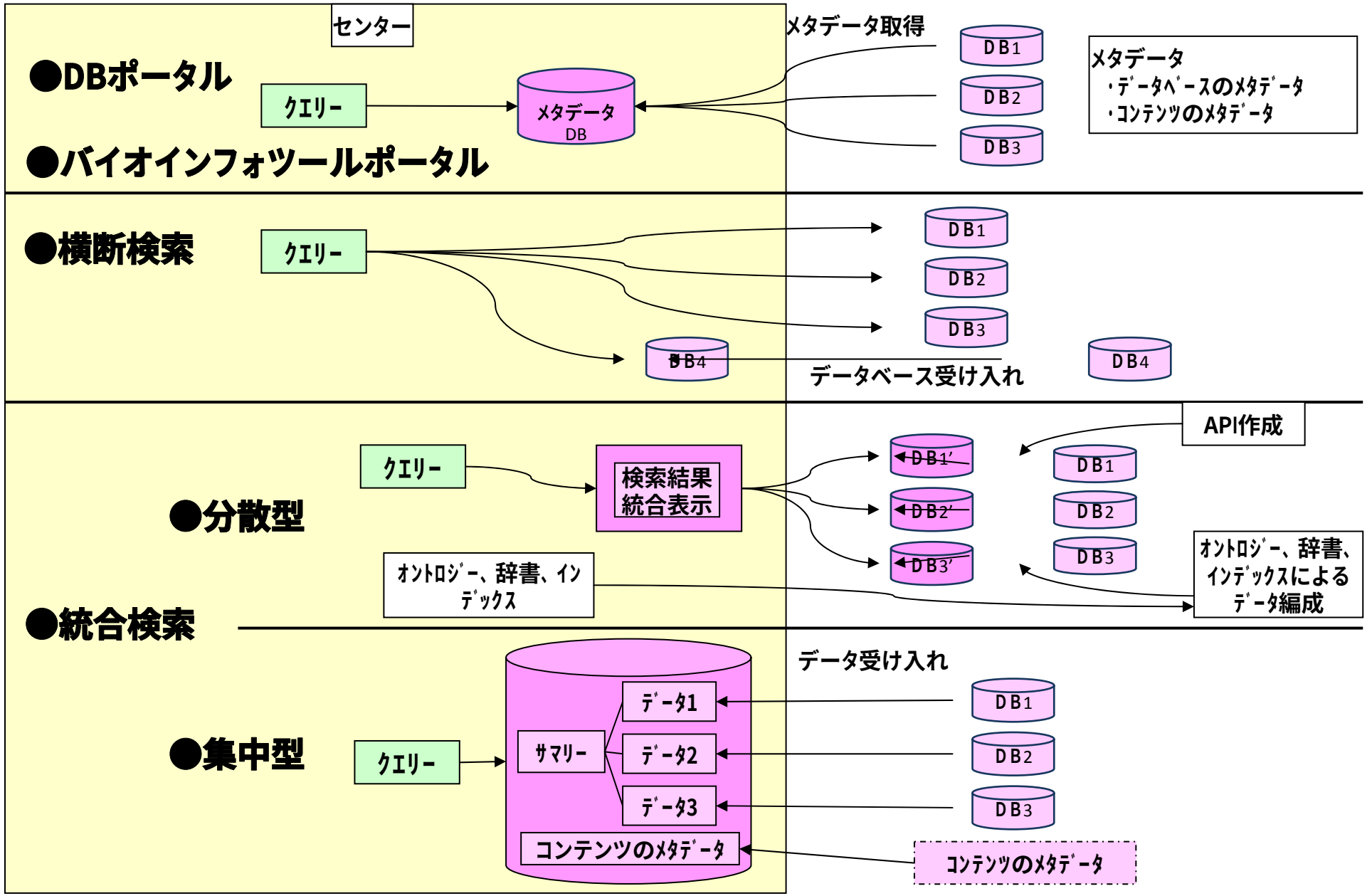
- 補完課題テーマと中核機関の連携の概要
- データベース検索のタイプと受け入れ
- 補完課題の概要
  - 補完課題の各テーマの概要、データ、進め方
  - 各テーマの説明

ライフサイエンス統合データベースセンター

補完課題テーマと中核機関の連携の概要



データベース検索のタイプと受け入れ



補完課題の各テーマの概要、データ、進め方

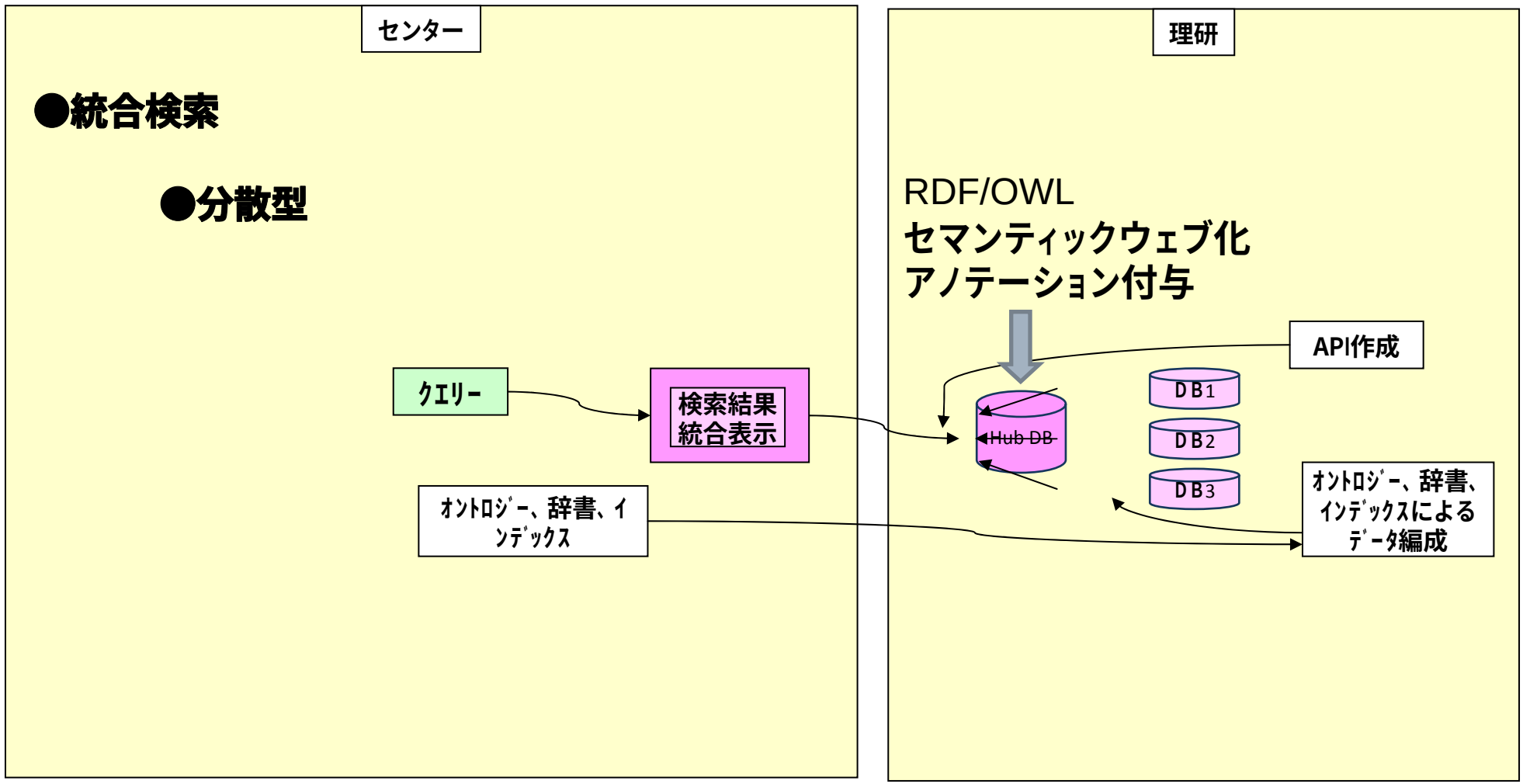
機関 題目 代表	概要	データ	進め方
理研 **「植物オミックス情報および蛋白質構造情報」** **ゲノム科学総合研究センターチームリーダー 豊田哲郎**	●シロイヌナズナにおけるオミックス情報、およびタンパク3000プロジェクトで生産された蛋白質構造とそれに付随する実験情報について、キュレーションやアノテーション作業を行い、それらのデータを中核機関の方針に従って本事業の統合データベースに提供する。 ●これによって、理研の他のデータベースも将来的に統合化していくためのモデルケースとしての流れを作ることを目的とする。	**1) シロイヌナズナにおけるオミックス情報とそのアノテーションデータ** ・トランスクリプトームデータ ・メタボロームデータ ・フェノームデータ ・リソース（完全長cDNAクローン） **2) タンパク3000プロジェクトで生産された蛋白質構造とそれに付随する実験情報とそのアノテーションデータ** ・動物、植物、微生物のNMRや、X線による構造解析データと実験データ ・試料調整・結晶化・回折実験データベース ・変異体構造解析データベース ・重原子データベース	**1) シロイヌナズナデータ（かずさの中村保一氏がとりまとめる）** **●かずさDNA研植物チームとの連携** 辞書、オントロジー、ID、データバージョンやアノテーションフロー等の共通化、ゲノム情報との統合 **●中核との連携** **テキストマイニング** **2) タンパク3000データ（中核の中村春木蛋白構造統合化チームリーダーがとりまとめる）** **●ターゲットタンパクPJ（PDBjも）との連携** オントロジーや辞書、API等の共通仕様 **●中核との連携** オントロジーや辞書、API等の共通仕様、テキストマイニング
産総研 **「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）** **糖鎖医工学研究センター長 成松久**	●糖鎖業界に散在するデータベースを集約し糖鎖科学データベースを構築。 ●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について、統合化。 ●我が国に存在する全ての糖鎖関連データベースを統合化（交渉し、標準化して格納） ● 糖鎖研究分野以外の方にも理解可能なインターフェースの開発 ●メタデータを中核機関に受け渡し、中核機関のDBポータルや横断検索と連携。 ●最終的には中核機関や他の分担機関とのデータの統合を目指す。	**●糖鎖医工学研究センターが構築したデータベース** ・糖鎖関連DB ・糖転移酵素特異性に関するDB ・レクチンDB ・糖タンパク質DB ・ＭＳスペクトルDB **●我が国に存在する全ての糖鎖関連データベース** ・名古屋市立大学加藤晃一先生　GALAXY ・立命館大学川嵜先生　GlycoEpitope ・名古屋大学古川鋼一先生　マウスのフェノタイプ ・北海道大学西村先生　PDBファイルから糖鎖構造 ・その他交渉中	**日本糖鎖科学統合データベース（糖鎖のポータルサイト）JCGGDB** 日本糖鎖科学統合データベース運営事務局 **●産総研・データ提供チーム** ・糖鎖遺伝子機能解析チーム ・糖鎖分子情報解析チーム ・レクチン応用開発チーム **●参画機関・データ提供機関（募集）** ・研究機関 ・大学 ・企業

補完課題の各テーマの概要、データ、進め方

機関名 題目 代表	概要	データ	進め方
遺伝研 **「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」** **国立遺伝学研究所 生命情報・DDBJ研究センター教授 五条堀孝**	●わが国の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、Traceデータのデータベース構築事業とデータ提供の事業を実施する。 ●背景 ・Traceデータは、品質管理や配列決定アルゴリズムの改良、配列断片のアセンブリにおいて大変重要で貴重な情報である。 ・Traceデータは、プロジェクトの完了とともに、消滅してしまう可能性が高い。 ・次世代の超高速の塩基配列決定装置の登場により、Traceデータの量は飛躍的に巨大化し、データハンドリングが困難な状況へ。	**●配列決定の原データ（トレースデータ）** ・サンガー法の波形データ ・次世代型シーケンサーの原データ 454、Solexa、ABI-SOLiD等 **●データ規模** 1プロジェクト500GB、約500エントリーとして、データ保管用ディスクとして240TB以上を用意	**業務フロー** ①登録受付 ②データ査定 ③ID発行 ④アーカイブ処理 ⑤データ提供 **サービス** ・Trace提供 ・ FTP ・キーワード検索 ・相同性検索 ・WWW全般 ・波形表示 ・登録処理
九工大 **「生体分子の熱力学データと構造データの統合」** **情報工学部教授 皿井明倫**	●蛋白質の安定性や相互作用の熱力学データを構造データと統合することを目的とする。 ●そのために、中核機関、PDBjと連携して、相互のデータのリンク、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	**●熱力学情報** ・**ProTherm**:　（約22,000件） 蛋白質熱力学データベース **●相互作用情報** ・**ProNIT**　（約8,000 件） 蛋白質・核酸相互作用データベース ・**ProLINT**　（約24,000件） 蛋白質・リガンド相互作用データベース	**中村春木 蛋白質立体構造統合DBチームリーダーがとりまとめる** **●PDBj、中核との連携** **・統合化技術** ・熱力学情報と構造情報のクロスレファレンス ・熱力学および構造関係の共通のオントロジーの整備 ・データ交換フォーマットの整備 ・検索の統合（APIの共通仕様） **・テキストマイニング技術** ・データを含む論文の自動収集 ・テキストからのデータの自動抽出

各テーマの説明 (理研)

データベース検索のタイプ



シロイヌナズナオミックスデータ1

(1)生物種	シロイヌナズナ（トランスクリプトーム）
(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

(1)生物種	シロイヌナズナ（トランスクリプトーム）
(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、無処理の植物体など

(1)生物種	シロイヌナズナ（メタボローム）
(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、特記事項	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。

シロイヌナズナオミックスデータ2

(1)生物種	シロイヌナズナ（フェノーム）
(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

(1)生物種	シロイヌナズナ（フェノーム）
(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

(1)生物種	シロイヌナズナ（フェノーム）
(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

(1)生物種	シロイヌナズナ（リソース）
(2)試料・ライブラリー等の種類、数	※申請者の保有データが、特定の分野・生物種においてどの程度カバーしているかの自己評価も記述 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

蛋白構造関連データ

(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)生物種	（試料調整・結晶化・回折実験データベース）９種類 （変異体構造解析データベース）２種類 （重原子データベース）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	
(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）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。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）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）。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。

シロイヌナズナオミックス注釈化作業

- 理研は、シロイヌナズナを題材に、トランスクリプトーム、メタボローム、フェノーム、リソースデータのアノテーションを、新規の実験データを含めて実施し、データベース化する。その際、ブラウザーでの可視化と共に、ダウンロードできる形でDB化する。
- かずさ、理研、中核との間で、遺伝子名称辞書、オントロジー、ID、データバージョンやアノテーションフロー等の共通化を実施する。
- かずさと理研の課題は互いに相補的になっているので、かずさのゲノムと理研のトランスクリプトームの間を統合化することによって、ゲノムから、トランスクリプトーム、メタボローム、フェノーム、リソースデータに至る統合化を図る。
- 統合化方式の枠組みが、本プロジェクトにおいて確立されれば、これをその後、別の植物や植物以外にも適用可能になることを意識して開発を行う。

タンパク質構造注釈化作業

- ターゲットタンパクPJ（PDBjも）と理研補完課題間の連携を、オントロジーや辞書、API等の共通仕様の観点で行う。

- 両PJ間の形式上の切り分けが必要。例えば、ターゲットタンパクPJでは、実験情報のオントロジーと辞書構築といった共通仕様に関することを実施し、理研では、具体的なタンパク3000データの実験情報のDB構築を行う。今後検討する。

- 中核と各機関との連携

  オントロジーや辞書、API等の共通仕様の観点で連携する。

  テキストマイニングツールの構築と利用に関して連携する。

  上記、オントロジーや辞書、API等の共通仕様の観点での連携では、逐次、情報や仕様の交換を実施する。

- 今後の連携の具体的方法

  具体的な連携を行うための各機関担当者によるチームを編成する。

  理研（豊田、横山、国島）、九工大、阪大、中核機関それぞれ1名

各テーマの説明（産総研）


糖鎖大量合成
・ケミカル
・バクテリア系
・ライブラリー化
糖鎖構造の分析・検出
・質量分析計
・2-D/3-D 糖鎖マッピング法
・糖鎖抗体
・レクチン
糖鎖関連遺伝子
・糖転移酵素・糖分解酵素
・糖ヌクレオチド輸送体
・硫酸転移酵素
・レクチン
・合成と分解パスウェイ
・マテリアルリソースのライブラリー化
日本糖鎖科学
統合データベース
(糖鎖のポータルサイト)
jCGGDB
糖タンパク質
・生物種別、組織別のタンパク質
およびその糖鎖構造
糖鎖欠損動物・糖鎖改変動物
・フェノタイプ解析
・マテリアルリソースのライブラリー化
糖脂質
プロテオグリカン
糖鎖関連疾患
・先天代謝異常
・癌転移
・感染
・アレルギー
糖鎖関連分化マーカー
・癌、免疫、再生医療、受精
糖鎖機能
・タンパク質の機能調節
・細胞間コミュニケーション

補完課題　糖鎖･平成19年度の進捗

		11月	12月	1月	2月	3月
運営と開発体制の準備	一般入札（公開用サーバ）/納品スケジュール		公告	入札･選定		履行期限
	開発体制					
	･人材確保		テクニカルスタッフ採用･人材派遣会社から採用			
	･開発機器購入		個人の開発用PC等購入			
	･会議用機材購入		提供機関との打ち合わせのための環境を整える			
データ提供機関との交渉	交渉					
	･活動を糖鎖業界に報告		日本糖鎖科学コンソーシアムで発表･コンソーシアムの活動の一環として認定			
	･活動に賛同を頂いた機関のリストアップ		上記発表後交渉開始			
	･H20年度の計画･立案			･1月後半から具体的に打ち合せを行う ･参加機関にヒアリング		
統合に向けたデータベース構築	糖鎖統合DBの活動を示すためのサイト		立ち上げ			
	産総研のDBを統合DB用に改良					
	産総研･糖鎖関連遺伝子DB		公開	更新･改良		
	産総研･糖タンパク質DB		外注		納品予定	
	産総研･レクチンDB		外注		納品予定	
	産総研･MSのDB		外注		納品予定	

各テーマの説明（遺伝研）

補完課題名：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

代表：国立遺伝学研究所・生命情報DDBJ研究センター、五條堀孝

事業の目標・概要

各種生物の遺伝子やゲノムの塩基配列を決定するいわゆるシーケンシングセンターにおいてはその配列決定の原データになる波形データなどのいわゆるアーカイブ(Trace archive; 以下Traceデータという)を有している。このTraceデータは、品質管理やそれを基にして行う配列決定アルゴリズムの改良ならびに配列断片を連結するアセンブリにおいて大変重要で貴重な情報である。そして、これらのTraceデータは、シーケンシングセンターの活動がその支持母体のプロジェクトの完了とともに終了するとすると、原則的には完全に消滅してしまう可能性が極めて高い状況にある。また、454やSolexaあるいはABI-SOLiDといった次世代の超高速の塩基配列決定装置の登場により、そのTraceデータの量は飛躍的に巨大化してきており、シーケンシングセンター自身でもそのデータハンドリングを含めて保存はもちろんのこと対処が非常に困難な状況になっている。

この状況の理解の下、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、当機関である大学共同利用機関法人　情報・システム研究機構の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事業とデータ提供の事業を実施する。

表１　予測されるデータ規模の例アカデミアDNAシーケンシングセンターにおける解析中データ（遺伝研・小原所長）

					波形データ総容量 (Gbytes)			（2007年８月現在）
		リード数	プレート換算	HQリード数	非圧縮	圧縮(1)	圧縮(2)	
原始紅藻	GSZW	530,607	693	427,705	175	80	16	ゲノム（完成配列）登録済み
カタユウレイボヤ	GCiW	3,198,499	4,176	2,694,372	1,056	480	96	NCBI TraceArchive登録済み
メダカ	GOLW	15,675,095	20,464	12,925,557	5,173	2,351	470	ゲノム（ドラフト）登録済み
メダカ	GOLN	4,522,752	5,904	3,781,645	1,493	678	136	
マウス	GMS	10,249,629	13,381	9,173,139	3,382	1,537	307	
立襟鞭毛虫	GMOW	1,452,431	1,896	1,181,375	479	218	44	
細胞性粘菌	GASW	482,780	630	387,328	159	72	14	
Diploscapter	GNDW	1,089,826	1,423	809,113	360	163	33	
2006年度計		5,603,961	7,316	4,499,702	1,849	841	168	

注意：

・プレート換算は、リード数/766 で計算　　(384ウェル - マーカ) x 表裏

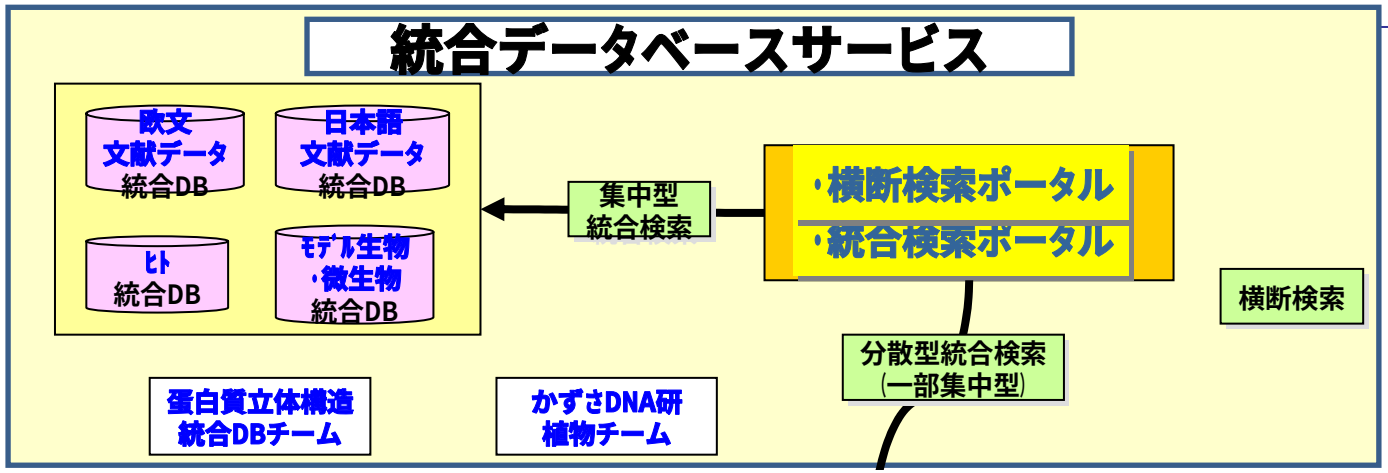
・HQリード数は、HQ長>=300を閾値に算出

・波形データ容量は、ABI generic形式の非圧縮ファイル1件330Kbytesとして算出

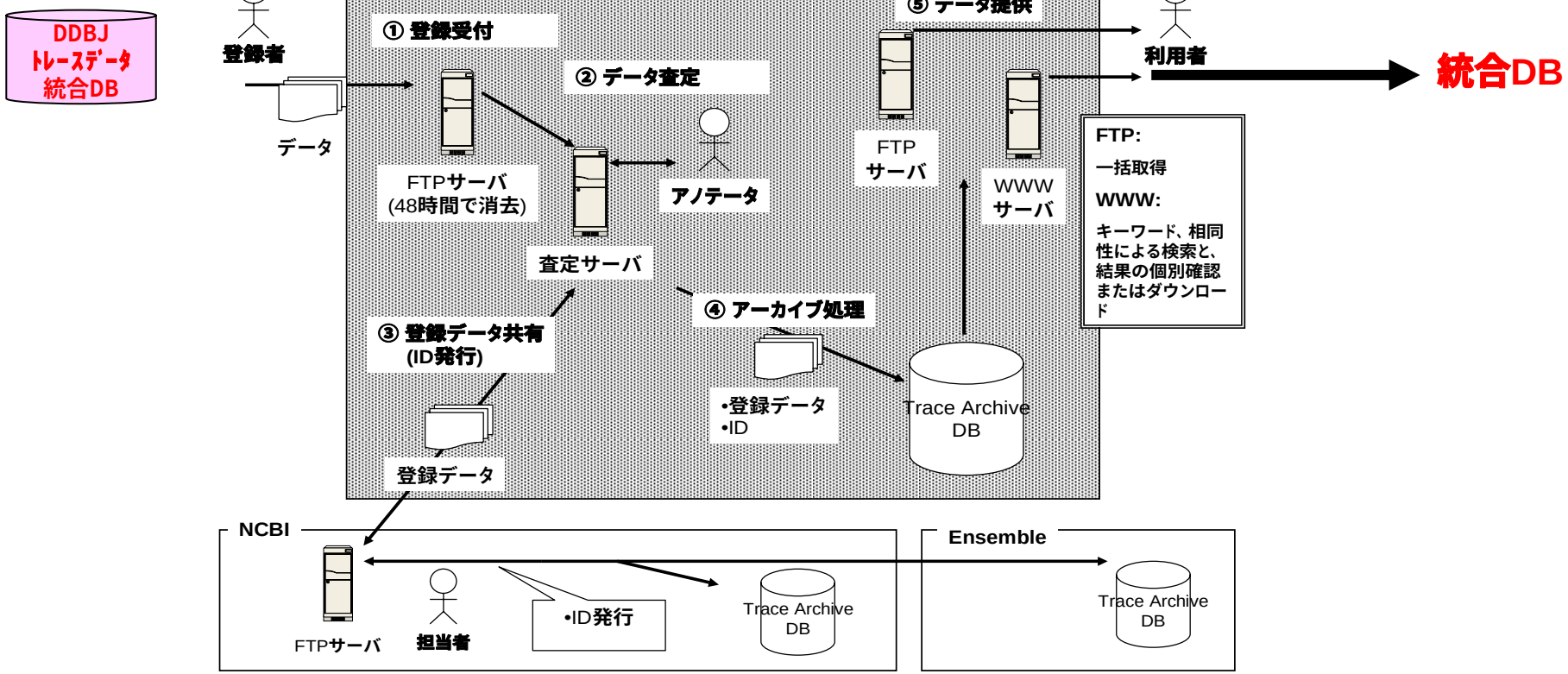
・圧縮(1)は、gzip圧縮した結果を基に算出（1ファイル150Kbytes）

・圧縮(2)は、NCBI TraceArchiveにおける圧縮後SCFファイルのサイズより算出（1ファイル30Kbytes）

補完課題における連携の概要



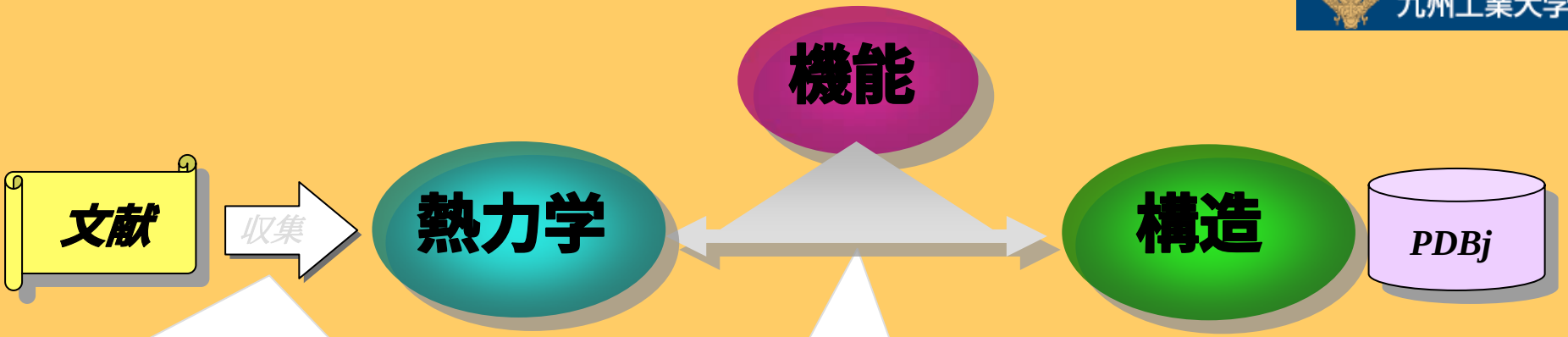
補完課題



各テーマの説明（九工大）

生体分子の熱力学データと構造データの統合

九州工業大学



テキストマイニング技術

- データを含む論文の自動収集
- テキストからのデータの自動抽出

統合化技術

- 熱力学情報と構造情報のクロスレファレンス
- オントロジーの整備
- データ交換フォーマットの整備
- 検索の統合 (APIの共通仕様)