個別タスクの実施事項及び成果 (案)
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　事務局

1．データベース統合戦略立案および評価

業務計画書項目名	個別課題名	成果の概要
●ゲノム注釈とデータベース間の連携における課題	●統合データベース間の連携と課題の整理	ゲノムアノテーション間の詳細な分析比較による質情報およびアノテーションの統合利用の方法に関する報告
●国内外DBの俯瞰と質的量的比較	●国内外DBの俯瞰調査	主に分子に関する国内外の主要なデータベースのリストおよびその維持管理状況やデータサイズなどに関する報告書
●ライフサイエンス分野の研究の俯瞰調査	●ライフサイエンス分野の俯瞰のためのタイトル処理システム	癌学会･生化学会･分子生物学会を含む国内主要学会の過去10年間の要旨に基づく分野研究の俯瞰マップと閲覧システム
	●ライフサイエンス分野の俯瞰のための連携システム開発	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
●検索アルゴリズムを含めた知識情報技術の動向調査	●統合処理技術の動向調査	知識情報処理の動向を文献資料、ネットワーク上での情報検索、専門家からの聞き取り調査、実地見聞などを駆使し調査した調査報告書
	●統合データベースに関する計算機資源の調査	効率的な計算機環境を整えるための、統合データベースに必要とされる計算機資源（CPU数、disk容量など）の予測調査
●臨床情報や医療統計の現状調査	●遺伝統計学分野における解析技術の基礎調査	個体の遺伝子多型と表現型との関連を解析するための有効な手法である遺伝統計学的解析手法の動向調査
	●臨床情報、疾患･健康情報の調査	疾患に関わる各種情報の存在調査と共同利用への課題調査報告書

「統合データベース間の連係と課題の整理」

- **目的**
  - –モデル植物のゲノム統合型データベース群の調査を行い、データベース間の連係状況の現状を知るとともに、あるべき将来像を示す。
- **実施事項**
  1. 「イネ」ならびに「シロイヌナズナ」のデータベースについてコンテンツや相互の連係状況をサイトから調査するとともに、聴き取りとアンケート調査から問題点を抽出。
  2. 調査結果を踏まえ、将来的にライフサイエンスの統合型データベースのあるべき姿を列挙し、必要な機能を実証的に検討するためのプロトタイピングを行う。
- **成果**

  全ゲノム塩基配列が決定されている2植物に関するデータベース群に関するコンテンツ調査と聴き取りならびにアンケート調査を行い、問題点を抽出するとともに、将来の統合型DBへの提言を含む調査報告をとりまとめた

「統合データベース間の連係と課題の整理」実施項目

1. イネならびにシロイヌナズナを対象として、代表的モデル研究植物である両者におけるゲノムアノテーション型のデータベース統合状況に関する詳細な調査を実施した
   1) それぞれの植物のゲノム情報を含む公開データベース群を対象に、それぞれのコンテンツの品質や量に関する網羅的な調査を行った

   → イネ46データベース・シロイヌナズナ25データベースの調査を完了した

   2) データベースの開発者ならびにユーザ(実験生物学者)への聞き取り調査とアンケート調査を行い、課題を抽出した

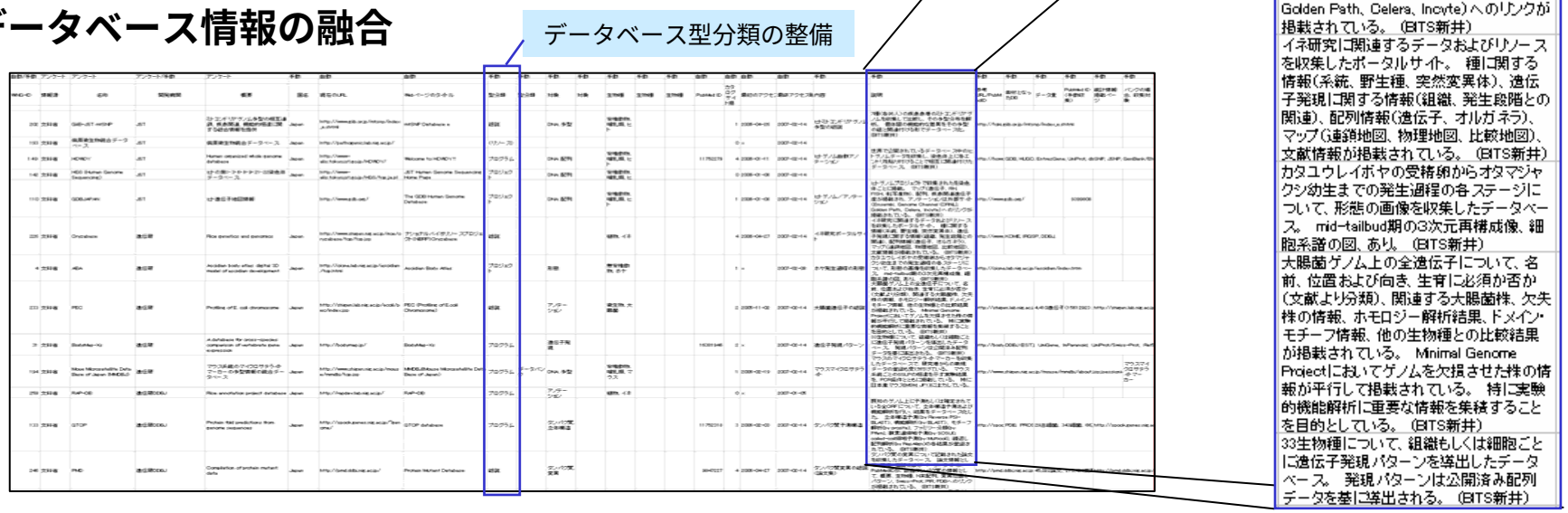
   → 聴き取り調査5名・アンケート調査188名の調査を遂行した
2. 上記の調査結果から、代表的なモデル生物ゲノム解析情報を基盤とした統合型データベース群を対象とした種々の課題を明らかにし、今後のライフサイエンスデータ統合の方策と指針についてとりまとめた

   → 統合の課題とユーザの要望を中心としたあるべきDBの将来像を含む、調査報告書の作成

「国内外DBの俯瞰調査」

- **目的**
  - 内閣府の調査とJST-DBの情報を融合するための調査を行い、データベースカタログを生成する。
- **実施事項**
  1. 内閣府の調査とJST-DBの情報の調査を行い、両者を融合するための調査を行い、融合作業を実施した。
  2. 融合した情報において、内閣府の調査における分類情報を整備し、また、主要なデータベースに関して日本語の解説を追加した。
- **成果**
  1. データベースカタログの生成
     - 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。
  2. データベースカタログ情報の整備
     - 各エントリーに、データベース型分類情報を追加した。
     - 主要なデータベースに関する日本語解説を整備した。

データベース情報の融合



「国内外DBの俯瞰調査」実施項目（成果）

- **内閣府の調査とJSTのDB情報を融合するための調査**
  - [融合対象]
  - 1) 内閣府の調査：ライフサイエンスDBの網羅的収集と多次元的な分類
  - 2) JST：専門家によるライフサイエンスデータベースの解説や分類活動

    →内閣府の調査とJST-DBの内容の調査を実施
- **大型のライフサイエンスデータベースの詳細調査**
  - [調査内容]
  - 1) データの種類 (内容)
  - 2) データ利用のための機能
  - 3) 維持管理状況やデータサイズ

    →大型のライフサイエンスデータベースを調査に追加
- **データベースカタログの作成**

  内閣府の調査とJSTのDB情報を融合したデータベースカタログの作成

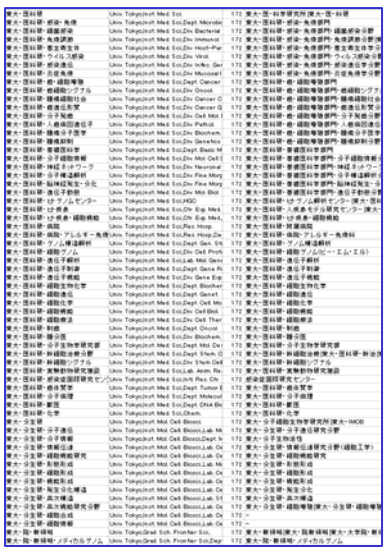
  主に分子に関するデータベースカタログ

  →内閣府の調査とJST-DBの融合と、情報を補完したＤＢカタログを作成

「ライフサイエンス分野の俯瞰のためのタイトル処理システム」

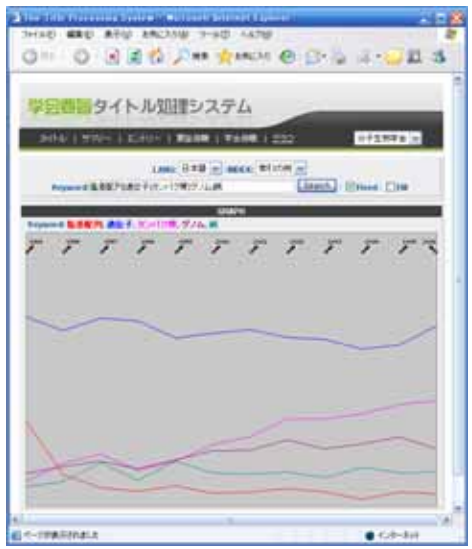
- **目的**
  - 分野研究の俯瞰を可能とする分子生物学界の要旨データベースシステムを開発する。
- **実施事項**
  1. 分子生物学会12年分の学会要旨を入力したデータベースを開発した。
  2. 学会要旨情報利用システムを開発した。
- **成果**
  1. **分子生物学会要旨データベースの作成**
     - 今回、打ち込みにより生成された学会要旨データ（12年分）をデータベース化した。
     - 施設名（所属機関名）表記の揺らぎを吸収する、施設名正規化辞書を作成した。
  2. **分子生物学会要旨情報利用機能**
     - 要旨の検索、ソート、可視化が可能なシステムを開発した。

学会要旨施設名辞書



学会要旨データベース

データ表示機能



「ライフサイエンス分野の俯瞰のためのタイトル処理システム」実施項目（成果）

- **学会要旨データベースシステム**
  - 1) 学会要旨・索引データ登録・検索・閲覧機能
  - 2) 学会要旨クラスタリング機能　→所属機関正規化辞書の作成

    →分子生物学会12年分の要旨を格納した「学会要旨データベース」を開発
- **学会要旨情報可視化システム**
  - 学会要旨エントリーの情報をリスト表示
  - → タイトル・サマリー・エントリー表示
  - → リスト、エントリーイメージ、学会イメージ表示
  - 2) グラフ表示
  - 3) チャート表示

    →学会要旨表示システムを開発
- **学会要旨情報可視化システムのエンハンス**
  - 解析・レポート機能の追加（今後の課題）

「ライフサイエンス分野の俯瞰のための連携システム開発」

- **目的**

1.現在、プラナリアの脳の遺伝子情報を発信するOh!脳サイト（以下URL）を、脳の進化ー発生ー働きに関する内容の充実を図る。

【具体的内容】

a.脳のなりたちと遺伝子に関する情報の充実

b.脳の働き、なりたち、進化の過程に関する情報の充実

c.プラナリアを事例に、細胞と遺伝子情報のデータベース化を図り、関係性を視覚的に把握できるシステムを構築

2.脳の働きについて、プラナリアの細胞と遺伝子を中核に、1,000個の細胞と約40個の遺伝子情報のマトリクスのデータベースを構築するとともに、細胞単位で発現を視覚的に把握することができる（俯瞰できる）ビューワー機能を構築する。

3.当システム開発に伴い、既存サイトと今回拡張する機能も含め、サイトマップを作成する。

- **成果**

1. プラナリアの脳細胞に関するデータベースの構築
   - プラナリアの脳細胞の情報と遺伝子情報の相関性を表現。
2. 3次元表現による遺伝子発現情報の作成
   - WEBブラウザを通じて3次元的表現により、脳細胞と遺伝子情報の相関性を表現。
   - また、インタラクティブなユーザーインターフェイスにより、細部にわたる情報を把握できる仕組みを実現。

▼細胞情報と遺伝子情報のBDシステム

▼3次元表現による遺伝子発現情報を表現

↑背面焦点（レイヤー15）

↑上部焦点（レイヤー52）

「ライフサイエンス分野の俯瞰のための連携システム開発」の成果

1. プラナリアの脳細胞と遺伝子情報の相関性に関するデータベースを構築

   1) 1,000個の脳細胞に対して、個々の持つ遺伝子情報等をデータベース化し、細胞と遺伝子との相関性を把握できるシステムを開発

      → 表現する細胞個数の検討（今後の課題）

2. 細胞レイヤー別の遺伝子発現情報を把握できる表現方法を確立

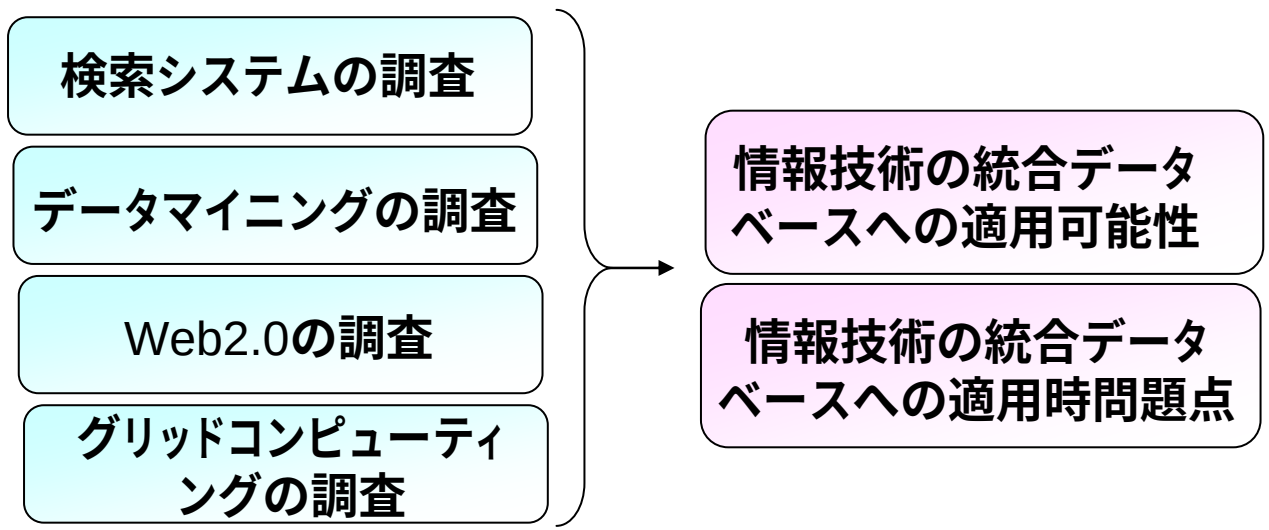
   1)複数層に積み重なる細胞を俯瞰的に把握できる手法と、細胞のレイヤー別で遺伝子の発現情報を把握できる表現方法を確立→XMLによる3次元表現を実現

   2) より鮮明な細胞の構図を把握できるように拡大縮小機能（Ajaxの応用）やユーザビリティを考えたインターフェイスツールの開発（今後の課題）

3. 細胞と遺伝子との相関性をより密接なつながりを表現できる構造化の検討（今後の課題）

「統合処理技術の動向調査」

- **目的**
  - 急速に展開する情報処理技術の動向、中でも知識情報処理の動向を文献資料、ネットワーク上での情報検索、専門家からの聞き取り調査、実地見聞などを駆使し分析を加え、その成果を報告書としてまとめる。
- **実施事項**
  1. 検索システム、データマイニング、Web2.0、グリッドコンピューティングの4項目について調査を行った。
  2. 上記の4項目の調査結果を基に、上記情報技術の統合データベースへの適用可能性について議論し、また、その問題点を考察した。
- **成果**
  1. 検索システム、データマイニング、Web2.0、グリッドコンピューティングの4項目についての調査報告
  2. 上記情報技術の統合データベースへの適用可能性とその問題点についての報告



「統合処理技術の動向調査」実施項目（成果）

1. 検索システム調査
   →従来のキーワード検索では限界がきている
   →検索システム自体に高度な解析機能が必要となってきている
   1)連想検索システムの調査　DualNavi, 想(IMAGINE)
   2)非構造化データ管理技術 UIMAの調査

2. データマイニング調査
   1) 知識発見についてのサマリー データマイニングと知識発見、知識発見プロセス
   2) グラフマイニングの調査　ラベル付順序木マイニングとその応用例

3. Web 2.0 調査
   ・Web 2.0 についてのサマリー ロングテール、集合知、forksonomy、ブログなど
   →Web 2.0 技術に共通するものとして、いかにデータを幅広くかつ効率よく集めるか、また、それらをいかにうまく管理するかという側面がある

4. グリッドコンピューティング調査
   ・歴史的経緯、概念、Web service との関係など

統合データベースに関する計算機資源の調査

- **目的：統合データベースに必要とされる計算機資源（CPU数、disk容量など）を予測し、効率的な計算機環境を整える。**

- **実施事項：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、disk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。**

- **成果：　テキストマイニング、Webクローラなどの8項目において重点的に調査した。**

統合データベースに関する計算機資源の調査　（成果）

- **統合データベースに関する調査**
  - –Web　crawling (bio**版**google)
  - –Text mining (**日本語、英語版**)
  - –Phenotype**の画像処理、画像検索**
  - –DDBJ/EMBL/GenBank**の項目別検索**
  - –**配列情報 (多型情報を含む)**

などについて調査を行い、項目別、段階別 (どこまで対象とするか) に必要とされるメモリー量、CPU**数、**disk**容量 (5年後のデータ量の伸びを考慮) を概算すると共に、必要とされる技術の概要を提示した。**

「遺伝統計学分野における解析技術の基礎調査」

- **目的**
  - 個体の遺伝子多型と表現型との関連を解析するための有効な手法である遺伝統計学的解析手法の動向を調査する。

- **実施事項**
  - 文献･インターネットによる調査を行ない，遺伝統計学分野の解析手法･解析アルゴリズムの特徴を整理して報告書を作成した。

- **成果**

1. 解析手法･解析アルゴリズムの調査
   - 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，それぞれの解析手法における代表的なアルゴリズムの調査を行ない，特徴および長所･短所を評価した。
2. ソフトウェアの調査
   - 連鎖解析，連鎖不平衡解析（ハプロタイプ解析），QTL解析を行なうためのソフトウェアに関して，実装されているアルゴリズム，動作環境，入出力などを調査し，特徴および長所･短所を評価した。

「遺伝統計学分野における解析技術の基礎調査」実施項目（成果）

1. **遺伝統計の現状調査**

   1) **目的**　**遺伝統計に関わるデータ産生の現状と解析手法の調査を行い、本プロジェクトにおける取組み指針を得る。**

   2) **内容**　**遺伝子多型解析に関わる研究手法の分類と解析アルゴリズムの調査。**
   遺伝子発現解析に関わる解析手法及び解析アルゴリズムの調査。
   現行アルゴリズムの問題点と将来のニーズ、方向性の把握。
   →文献、インターネットを利用して、遺伝統計学分野の解析手法・解析アルゴリズムの調査および評価
   遺伝統計学分野のソフトウェアの調査および評価
   調査報告書の作成を行った。

「医療情報、疾患・健康情報の調査」

- **目的**
  - 臨床データの活用を視野に、本プロジェクトにおける医療情報への取組みに関する方向性を探る。
  - 医療統計に関わるデータ産生の現状と解析手法の調査を行い、本プロジェクトにおける取組み指針を得る

- **実施事項**

1. 電子カルテ等における臨床データの規格化とデータ共有化の現状と動向の調査
2. 医学統制用語等の現状調査
3. 臨床データ活用上の課題の論点整理
4. 我が国における代表的なコホート研究の統計分析手法の調査
5. 統計分析手法の課題と方向性の整理

- **成果**

1. 地域臨床データ共有において、現在できていることと課題意識を整理
   - 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モデル化の意味、インセンティブ等
2. 生活習慣病を中心としたコホート研究の幅広い事例調査、統計分析手法、成果等の整理
   - 生活習慣病を中心としたコホート研究の事例を分類・整理
   - 統計分析手法を可能な範囲で論文等から抽出
   - データオリジンの課題の抽出
   - 地域、全国、世界でのデータ連携の動向の整理　　等