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	ライフサイエンス統合データベース開発運用（統合データベース支援：意見集約システム運用／広報／データベース受入・運用等）
実 施 機 関 名	独立行政法人　科学技術振興機構
代 表 研 究 者 名	白木澤　佳子

１．課題開始時における達成目標

「ライフサイエンス統合データベース開発運用」では、ライフサイエンスやバイオ産業に従事する研究者や技術者がいわゆるゲノムプロジェクト・ポストゲノムプロジェクトの成果や多様な DB や解析ツールをストレスなく利用してより高度な研究開発が効率よくできる環境（統合 DB）を実現することを目的とする。このため、情報・システム研究機構、科学技術振興機構、産業技術総合研究所生命情報工学研究センター、かずさ DNA 研究所、九州大学、奈良先端科学技術大学、長浜バイオ大学、東京大学、お茶の水女子大学は共同して、以下の３つの業務を行う。

（１）戦略立案・実行評価

（２）統合データベース開発

（３）統合データベース支援

このうち、独立行政法人　科学技術振興機構は（３）統合データベース支援の中で、意見集約システムの運用、広報の一部、並びにデータベース受入・運用の一部に関わる研究開発を実施する。

各研究開発の詳細は以下の通り。

①　意見集約システムの運用

個別 DB について、平成 18 年度先行着手分の成果である、DB 構築者および利用者からの意見を集約するためのシステム WINGpro の公開を継続する。

②　広報

文部科学省「ライフサイエンス分野の統合データベース整備事業」等のウェブサイトを運営する。参画各機関の当該事業に関する研究開発動向や活動、ライフサイエンス政策の動向等を広く広報・情報発信をする。

③　データベースの受入・運用

データベースの受入に関して、レポジトリの制度等の調査および国際標準あるいは既存のメタデータについて調査し、結果の要約を公開する。メタデータの受入に必要な国際標準メタデータエレメントサイトの研究開発を行い、運用を実施する。

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

成果概要は以下の通り。

①意見集約システム（WINGpro：http://wingpro.lifesciencedb.jp/）を継続運用している。登録されたデータベース数は平成２２年１０月末時点時点で４３８件。

②以下４つの事業ウェブサイトを運営している。

「ライフサイエンス分野の統合データベース整備事業」(http://lifesciencedb.mext.go.jp/)
「ライフサイエンスの広場」(http://www.lifescience.mext.go.jp/)
「Webリソースポータル for バイオインフォマティクス」(http://tools.lifesciencedb.jp/)
「遺伝子名称シソーラス」(http://nlp.lifesciencedb.jp/)
③平成１９年度テストサイトとして構築した、メタデータとして記述する項目「メタデータ要素（エレメント）」とその記述規則を収録する国際標準メタデータエレメントサイト（MDeR：http://mder.jst.go.jp/）を継続して運用している。

（２）進捗及び成果

進捗及び成果は以下の通り。

① 意見集約システムの運用

当初の予定通り、意見集約システム WINGpro を公開、運用している。【別紙参照】

② 広報

当初の予定通り、４つの事業ウェブサイトを運営している。【別紙参照】

③ データベースの受入・運用

当初の予定通り、レポジトリの制度等の調査および国際標準あるいは既存のメタデータについて調査を実施し、結果の要約を国際標準メタデータエレメントサイトにおいて公開している。【別紙参照】

３．当初目標に対する達成度

当初の目標通り達成した。

４．中間評価に対する対応

相乗的に目標を達成するため、参画機関との間で連携強化を実施した。具体的には、運営委員会等各種委員会等や、シンポジウムへの参加を実施した。

５．他機関との連携

代表機関と緊密な連携をとりながら、活動を実施した。具体的には、運営委員会をはじめとする各種委員会に参加した。また、代表機関である大学共同利用機関法人情報・システム研究機構が主催するシンポジウムにおいて、毎年成果発表を実施した。

６．今後の見通し、計画、展望

意見集約システムの運用、データベースの受入・運用については、本事業成果である WINGpro、MDeR のサービスが利用者にとって有益であると考えるため、継続の可能性について検討を実施する。
広報については、４つの事業ウェブサイトの効果的な公開方法について検討を実施する。ライフサイエンス分野の統合データベース整備事業ページについては更新を停止し、静的な公開を実施することとし、ライフサイエンスの広場については、今後の運用体制についての検討が必要。

７．全体総括

当初の目標について、予定通り達成した。担当した広報活動では、ライフサイエンス分野の統合デー

タベース整備事業を広く一般に対して認知してもらう為にウェブサイトの運営を行い、事業に対する一般の理解を得ることが出来た。また、意見集約システム WINGpro の運用については、平成 21 年度実績で 263610 ページビューを獲得し、多くの研究者等に利用されるサービスを提供することが出来た。

８．特記事項

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）	9,818	0	0	0	0	9,818
人　件　費（千円）	0	0	0	0	0	0
業務実施費（千円）	22,000	75,000	0	0	0	97,000
一般管理費（千円）	3,182	7,500	0	0	0	10,682
合　計　（千円）	35,000	82,500	0	0	0	117,500

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）
	WINGpro http://wingpro.lifesciencedb.org/ http://wingpro.lifesciencedb.jp	公開	ライフサイエンス分野の内外のデータベースの情報を収集、整理、分類し、438DB を収録。ディレクトリからは、データベースの構築法による分類と生物種および対象による分類でデータベースを一覧可能。平成 19 年 3 月 30 日正式公開。平成 19 年 7 月 2 日よりユーザが記事を投稿および編集できる機能を公開。全文検索、分類一覧が可能。入力方式、一括データ取得、tips など記述。ユーザによる登録を促すことが課題。分類一覧の自動作成を開発。
	Web リソースポータルサイト for バイオインフォマティクス http://tools.lifesciencedb.jp/	公開	実験データの解析や公的データの加工に使用する解析ツールや環境を案内するサイト。ベンチメソッドオントロジーの一部を利用。解析ツール総数 456、解析ワークフロー総数 29、著名な論文 18 報のワークフローを収録、一括ダウンロードも可能。平成 19 年 3 月 30 日正式公開。
	Jabion Genome Viewer：HAL http://www.bioportal.jp/ja/GenomeViewer/	公開	HAL（生物種 human）（ヒトゲノムに潜む遺伝子の位置と構造、その機能などの注釈情報を提供。）を既存のゲノムビューアである Jabion Genome Viewer に受け入れたテストサイト。H20.1.7 公開。ヒトゲノムに潜む遺伝子を発見する信頼性の高いコンピュータアルゴリズムにより予測した遺伝子数多数を公開。

（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	メタデータ要素レポジトリ(MDeR) http://mder.jst.go.jp/	公開	メタデータ要素の検索、検索結果の比較、収録メタデータのメタデータ要素の一覧表示が可能。ISO/IEC 11179 Part3(Registry metamodel and basic attributes)に準拠した形で、国際標準５種，データベース３種のメタデータ要素を収録。収録内容の精査と充実によりデータ収集・データベース構築時のデータ形式等の決定に役立つサービスを目指す。

（3）その他の成果物（(1)、(2）に該当しないもの）

通番	名称	公開／未公開	概要
1	事業広報サイト http://lifesciencedb.mext.go.jp/	公開	統合データベースプロジェクトの紹介サイト
2	ライフサイエンスの広場 http://www.lifescience.mext.go.jp/index.html	公開	文科省のライフサイエンス政策等の紹介サイト

①　意見集約システムの運用

意見集約システムWINGproの継続運用を行った。平成22年10月現在438DBを収録。

図1　WINGpro　（http://wingpro.lifesciencedb.jp/）の表示画面例

②　広報

文部科学省「ライフサイエンス分野の統合データベース整備事業」等のウェブサイトを運営し、これらのサイトを登載しているサーバ等の運用を行った。参画各機関の当該事業に関する研究開発動向や活動、ライフサイエンス政策の動向等について広報・情報発信をした。新規記事の追加や修正を行った。

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

図 3 「ライフサイエンスの広場」（ http://www.lifescience.mext.go.jp/ ）

図４「Webリソースポータル for バイオインフォマティクス」（ http://tools.lifesciencedb.jp/ ）

図５「遺伝子名称シソーラス」（ http://nlp.lifesciencedb.jp/ ）

③　データベースの受入・運用

メタデータとして記述する項目「メタデータ要素（エレメント）」とその記述規則を収録する国際標準メタデータエレメントサイト（MDeR : http://mder.jst.go.jp/）のテストサイトを構築し、平成19年から運用している。

図 6　MDeR 画面

（http://mder.jst.go.jp/）

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	統合データベース開発：植物及び植物関連微生物のゲノム情報データベースの統合と高度化
実 施 機 関 名	財団法人　かずさディー・エヌ・エー研究所
代 表 研 究 者 名	中村保一

１．課題開始時における達成目標

植物ゲノムのデータベースを対象とした統合への課題調査により明らかとなった、植物ゲノムデータベースの現状へのユーザの不満を解消することを目標とした業務を行う。ユーザからの情報集積と、それを基にした整理を可能とすることを特徴とするゲノムデータベース環境を構築し、次いでその環境へのデータの集積と高度化を図る。

①高度情報集積データベース開発

分子データ上のポジションを基盤とした、アノテーション・キュレーションの高度化及び統合化を可能とする、基盤データベースを整備し、植物及び植物関連微生物のゲノム解析情報を対象に運用する。同時に、そのようにして集積したアノテーション情報をゲノム上に表示し、優れたインタフェースにより閲覧可能なシステムを開発する。データベース中の遺伝子名や遺伝子IDのトレースが困難であることが利用者にとって大きな障壁であることから、植物遺伝子名や遺伝子IDの整理をおこない、名称の食い違いを解決するためのWWWサービスを行う。

②ゲノムアノテーション情報の蓄積と高度化

①のデータベース環境を整備した後ただちに、その環境を利用したアノテーション情報の蓄積と高度化を実施し、公開する。データベースの維持、アノテーション、キュレーションを行う人材の教育に関わる情報も、将来の継続的なデータベースの充実には不可欠である。そのため、DNA研究所で構築中の人材養成のための教育コンテンツの共有システムを継続して作成し、維持しつつ教育分野の事業と連係し、内容の一層の充実を進める。

（課題実施開始時には定量的目標を設定しておりませんでした）

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

平成 19 年度に、我々が提案した課題とそのロードマップを図１に示す。本事業では当初計画した通りのスケジュールで「高度情報集積データベース」および連携支援するシステムを作成することができた。システムの概要を図２に示す。

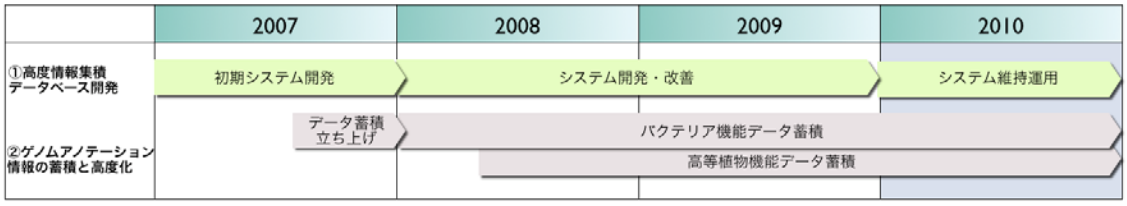


図１　プロジェクトの進捗計画

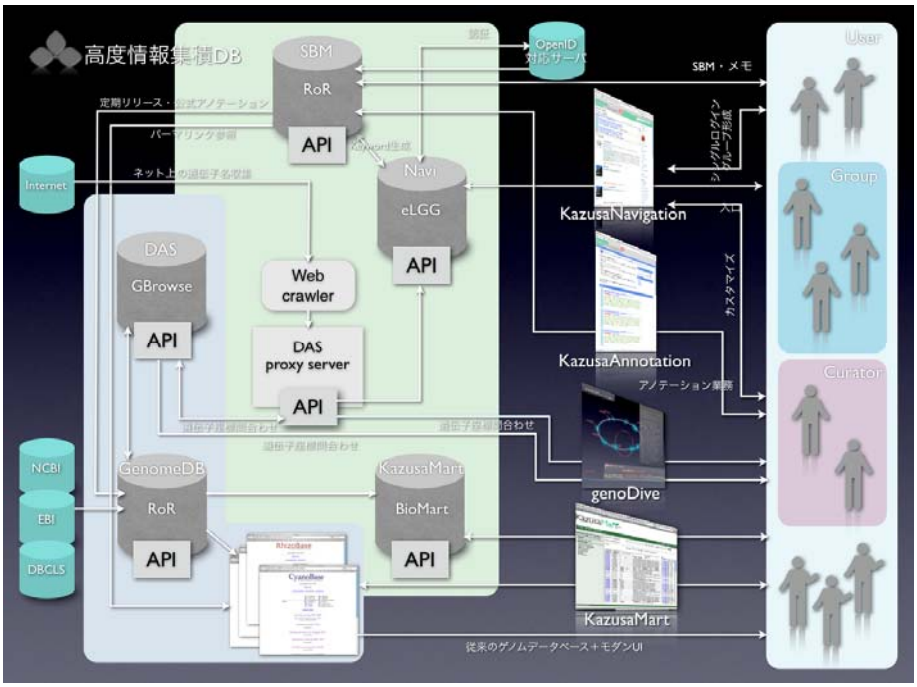


図２　高度情報集積データベースと連携ツール群

ここに示したデータベース群と連携支援するツール群を利用し、実際にゲノムに対するアノテーションとして文献からの情報抽出を行った（Gene Indexing 型アノテーション）。アノテーションの進捗状況を図３に示す。10 月末時点で 7269 報の論文から 180,601 アノテーション情報を抽出し蓄積している。

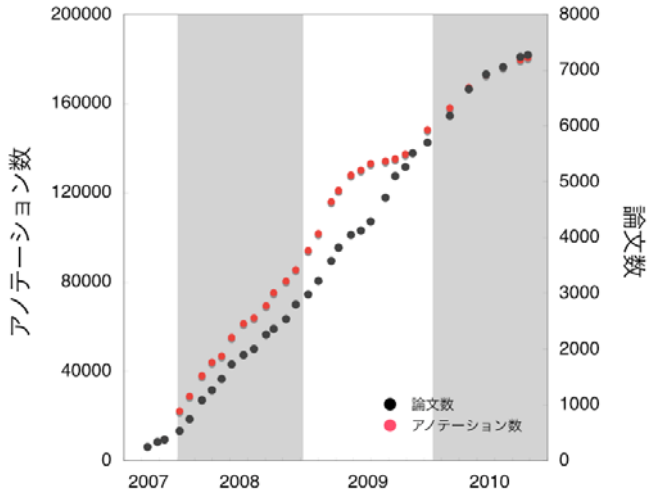


図３　Gene Indexing 型アノテーションの進捗状況

（２）進捗及び成果

１）高度情報集積データベースの運用と改良

KazusaAnnotation（http://a.kazusa.or.jp）は、分子データ上の位置情報を統合基盤情報とし、アノテーション・キュレーションの統合と高度化を可能にする新しいタイプのデータベースである。平成 21 年度までに、当初計画していた基本機能の開発と改善をほぼ完了している。各サービスのアクセス状況を表１にまとめた。

表 1　かずさにおける各サービスのアクセス状況（2009年9月～2010年9月）

サービス名	URL	訪問者数	ユニーク訪問者数	ページビュー
KazusaAnnotation	http://a.kazusa.or.jp	14494	3520	172395
KazusaNavigation	http://navi.kazusa.or.jp	36038	22892	61169
KazusaWiki	http://wiki.kazusa.or.jp	9942	5050	29410
CyanoBase	http://genome.kazusa.or.jp/cyanobase	69082	19190	520095
RhizoBase	http://genome.kazusa.or.jp/rhizobase	15298	5047	113951
genoDive*	http://genodive.org	247	155	764

＊ genoDive は 2010 年 6 月からの集計

今回、KazusaAnnotation の遺伝子単位で文献情報を一覧できる表示（ReferenceView）を、生物別のデータベース（CyanoBase, RhizoBase）から参照可能にした。この仕組みにより、文献マニュアルアノテーションである Gene Indexing の更新結果が、生物別のデータベースに即時反映されるため、エンドユーザーの利便性が向上した（図 4）。

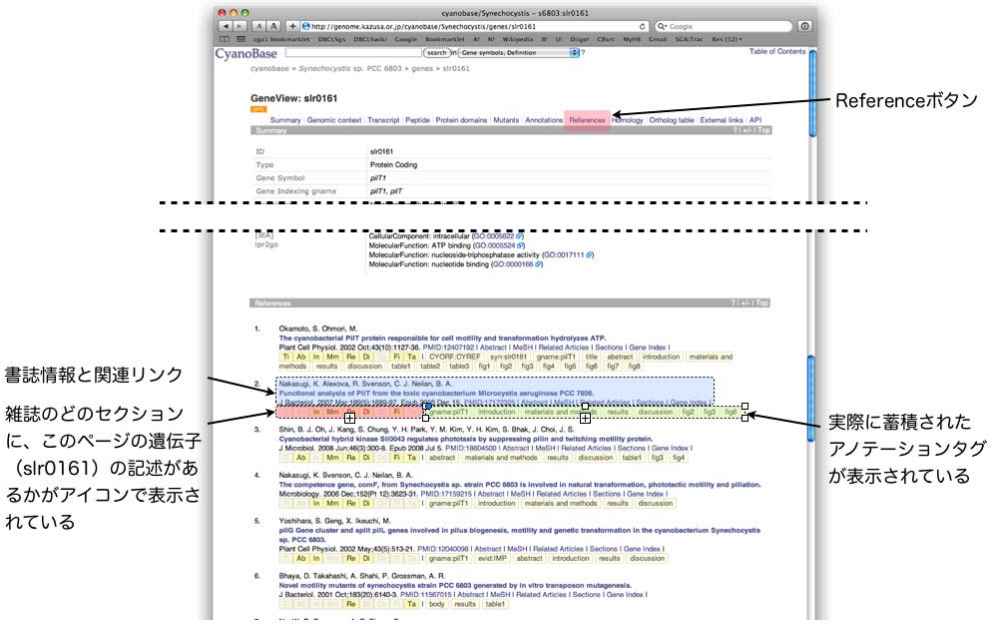


図 4　生物別のデータベースの遺伝子情報ページにある文献情報「Reference」は、KazusaAnnotation に蓄積している Gene Indexing データをもとに自動更新される

平成 22 年度の新しいとりくみとして、KazusaAnnotationを利用して、植物関連土壌細菌のゲノムアノテーションジャンボリーを行った。アノテーションジャンボリーでは、公開前のゲノム情報を扱う。そのため、ジャンボリー参加者のみが情報を共有できる仕組みをKazusaAnnotationに追加した。また、複数人数で一貫性のあるゲノムアノテーションを行うための標準作業手順など協働作業手法を開発しKazusaWiki（http://wiki.kazusa.or.jp）によりマニュアルや資料のかたちで共有した。かずさの提供するシステム組み合わせることにより、研究コミュニティーによるマニュアルアノテーションを成功させた。

genoDive（http://genodive.org）は、ゲノムの位置座標上に様々な情報を統合する可視化システムである。 アプリケーションは、Mac OS 版と Windows 版があり、Web サイトよりダウンロード可能である。真核生物版において、シンテニー領域、カリオタイプ、二項関係の情報可視化表現を開発し追加した（図 5）。統合データベース講習会などで、ゲノム情報可視化に関する講義の教材としても利用している。

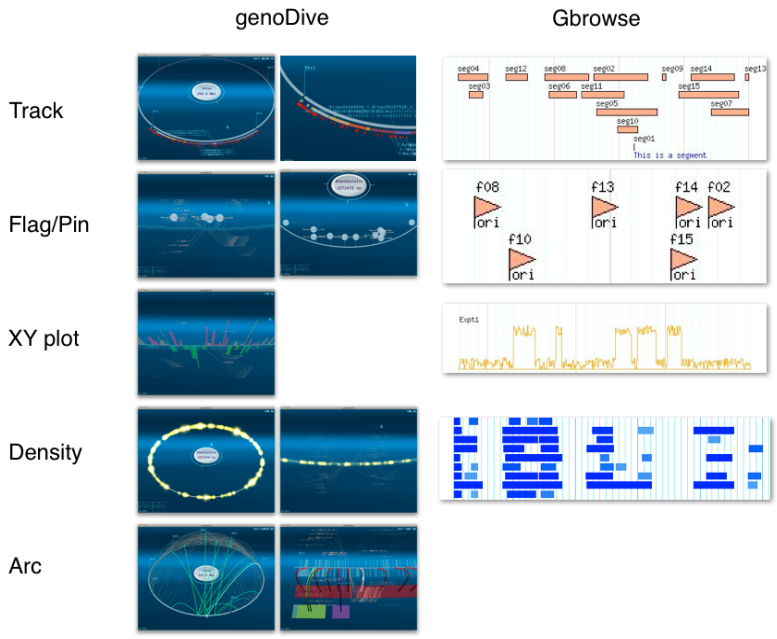


図 5 genoDive と Gbrowse のゲノム情報可視化表現の比較

２）ゲノムアノテーション情報の蓄積と高度化

平成 22 年度も論文中に出現する遺伝子・タンパク質記載のインデクス情報（Gene Indexing 型アノテーション）を KazusaAnnotation を利用し蓄積した。具体的な生物としては、*Synechocystis* sp. PCC 6803 をはじめ 8 種類の光合成細菌に関する合計約 4300 報の論文について、また、*Mesorhizobium loti* MAFF303099 をはじめ 5 種類の窒素固定根粒菌に関する合計約 2600 報の論文について遺伝子の位置情報の登録を完了している（詳細は表）。その他、高等植物としてマメ科のミヤコグサ（*Lotus japonicus*）に関する約 300 報の論文も登録している。合計 14 生物種の約 7200 報の文献から遺伝子とタンパク質の情報を 18 万件抽出蓄積しゲノムの位置情報と論文記載情報の統合を図った（表２）。

表２　Gene Indexing 型アノテーションの対象生物種とアノテーション済み論文数

生物種	論文数
Lotus japonicus	296
Mesorhizobium loti MAFF303099	182
Sinorhizobium meliloti st.1021	176
Bradyrhizobium japonicum USDA110	945
Rhizobium sp. NGR234	128
Rhizobium leguminosarum bv. viciae 3841	1,162
Anabaena sp. PCC 7120	710
Anabaena variabilis ATCC 29413	95
Chlorobium tepidum TLS	177
Nostoc punctiforme ATCC 29133	91
Synechocystis sp. PCC 6803	2,049
Synechocuccus elongatus PCC 7942	720
Synechococcus sp. PCC 7002	257
Thermosynechococcus elongatus BP-1	281
Total	7269

KazusaAnnotation システムを用いたジャンボリー形式のゲノムアノテーションでは、全国から 11 の大学や研究機関に所属する 31 名が参加した。3 日間の合宿と、その後 2 週間の作業で 6948 遺伝子に

対して 2068 の遺伝子名の改善を行った。

これらの情報は上記のデータベースから即時公開し、広く生物学者の利用に供している。また、活用方法を一般の生物学者に周知するため、統合データベース講習会や植物研究関係のワークショップで紹介し利用を促した（平成 22 年度 4 月より 10 月末までの実績。講習会 3 回、ワークショップ 2 回、招待講演 1 回、アノテーションジャンボリー1 回)。

３）植物と関連微生物ゲノム情報の統合

生物種をまたいだ横断的なゲノム情報の取得や、理化学研究所など他機関の植物オミクス情報との連携のため、KazusaMart（http://mart.kazusa.or.jp）を構築した。KazusaMart は、BioMart システム (http://www.biomart.org/)を利用している。このシステムは、グラフィカルユーザインターフェイスを備えており、生物学者がデータベースに複雑な問い合わせをしてデータ取得することができる。植物関連データエントリに付与した種々の属性により高度なデータの絞り込みが出来るように各種データ調査、注釈、整形し KazusaMart に格納した。イネ、シロイヌナズナをはじめとした解析情報の整備された主要植物 8 種、ならびに植物関連微生物 18 種類のゲノム情報を追加し公開した。また、活用方法を一般の生物学者に周知するため、統合データベース講習会や植物研究関係のワークショップで紹介し利用を促した（平成 22 年度 4 月より 10 月末までの実績。講習会 1 回、ワークショップ 1 回)。

３．当初目標に対する達成度

H22年度当初目標に対する現在までの達成度を記する。

１）「高度情報集積データベース」については、Gene Indexing型アノテーション情報を、複合体への注釈を含め18万件以上の蓄積を予定し、現時点でアノテーション180601件（10月末）であり目標を達成ている。

２）植物と関連微生物のコミュニティーへの周知を徹底するとともに基盤データベースと閲覧環境の改善をすすめ、100名規模のアクティブな利用者を確保することを目標した。1土壌細菌コミュニティーでの新規ゲノムアノテーションジャンボリーでの利用し、登録利用者115名達成（10月末）。

３）「BioMart による植物と関連微生物ゲノム情報の統合」については、イネ、シロイヌナズナをはじめとした解析情報の整備された主要植物 8 種ならびに植物関連微生物 16 種類のゲノム情報を追加を目標とした。10 月末時点で、BioMart 植物関連微生物 18 種類、主要植物 8 種類追加を完了している

４．中間評価に対する対応

指摘１：興味深い試みだと思います。が、全体像の見通しがよくないという印象をうけました。対象となる生物種が明示され、かつ、各生物種に対するアノテーションの統計情報等があれば、見通しがよくなるのではないでしょうか。

指摘２：面白い試みと思いますが、評価する段階ではないようです。現状に対する評価でつぶれてしまう可能性があるとすれば、もったいないと思います。

改善：KazusaAnnotationシステムでの見通しを改善するために、トップページにアノテーション数、遺伝子数、タグ数の統計情報を表示しました。さらに、検索用のSearchView、テーブル形式でまとめ

たTableView、文献情報抽出アノテーション (Gene Indexing)用に遺伝子毎、文献毎にまとめた表示、GeneIndexingView, IndexedRefView, ReferenceViewを作成しました。ReferenceViewは、CyanoBase, RhizoBaseなどの生物種別のデータベースからも参照出来るように改善し、ユーザの利便性の向上を図りました。

指摘３：全体的にサイトが重い気がします.
指摘４：リンク先のトップと思われるページが心なしか重い気がします。それと、細かい指摘ですが日本語での解説の文体がそろっていません。
改善：データベースシステムの動作改善、キャッシュ機能の追加による表示高速化を行った。その結果、改善前と比較して検索応答やゲノムマップの表示が数倍から数十倍程度の高速化した。サイト内の解説の文体を出来る限り統一した。

５．他機関との連携

１）DBCLSとの連携
DBCLS内の自然言語解析チームと共同で、手動のアノテーションを支援する固有表現抽出プログラムを作成した。
また、我々のGene Indexingで集積したデータをタンパク質機能俯瞰システムであるGendoo（http://gendoo.dbcls.jp/）で利用した。
２）研究コミュニティーとの連携
根粒菌の研究コミュニティーが新規決定したゲノムのアノテーションにKazusaAnnotationシステムを提供し、31名が参加したゲノムアノテーションジャンボリーを成功させた。
３）研究者への啓蒙と広報活動
主要な学会での発表（**回）、研究コミュニティー別のワークショップ（**回）、統合データベースセンター講習会への参加により、アノテーション手法の啓蒙や、我々の技術の広報活動を行った。
４）理化学研究所との連携
理化学研究所・横浜研究所　生命情報基盤研究部門との理研シロイヌナズナデータのBioMartへの受入による連携（会合4回）

６．今後の見通し、計画、展望

膨大な情報を統合的に扱うオミクス研究では、よく整備されたゲノムデータベースのアノテーション情報が様々な情報への横断的なアクセスの入り口として非常に重要な資源になっている。
しかしながら、人手によるアノテーション作業は依然としてコストの高い困難な仕事であり、アノテーションを支援する技術も学術的に評価されにくい分野であるため、研究開発が活 発とは言いがたい。したがって、JSTライフサイエンスDB統合推進事業（仮称）では、一貫性のあるルールで継続的にアノテーションを付与し続ける、技術的、人材的、制度的な枠組みを整備する事業を行っていただくことを期待する。
課題点としては、以下の二点をあげる。１）アノテーター・キュレーター、システムエンジニア、個別研究者の間の意思疎通を円滑に橋渡しするコーディネータ的人材の養成。２）育児中、介護中でフ

ルタイムに働けない人や、リタイアした専門家など、多様な働き方を望む人材が柔軟に働く事の出来る雇用形態の多様化。質の高い人材の確保と活用が最重要課題であると考える。

7．全体総括

新型シーケンサの普及によってゲノムやトランスクリプトームの配列が従来に比較してさらに高速かつ大量に決定されるようになった現在、リファレンスとなるべきモデル生物種群のゲノムアノテーションの信頼性を増すことは急務であると言える。とくに我が国のライフサイエンスの得意分野である植物や産業的に有用な微生物群のゲノムアノテーションの支援は重要な課題であり、これを目標の中心に据え事業を実施した。アノテーション統合化のため、簡便にデータ集積と整理を支援できるシステムを開発し、そのシステムを用いてゲノム上への論文情報の高速かつ高精度なマッピングと即時公開を実施しその利用価値を実証した。最終年度にはこのシステムを用いたコミュニティゲノムアノテーションの支援をも行い、その汎用性を示すことができた。また、事業を通じて蓄積された高密度な情報によって、植物と植物関連微生物の研究情報資源としてひろく活用される、価値の高いデータベースを形成し公開することができている。

8．特記事項

植物に関連する生物のゲノム研究に於いては世界の中心の一つであるかずさディー・エヌ・エー研究所に於いてこの事業を実施することによって、すでにデファクトスタンダードとして利用されてきた CyanoBase, RhizoBase 等のゲノムデータベースにさらなる価値と信頼性を付与し、のみならずアノテーション基盤として、システムの有用性を示すことができた。生物学に於いて世界をリードする分野での特異なデータベースとして、さらなる優位性を確保することにおおきく寄与したものと考える。また、プロジェクトの推進に際しては、遠隔地にちらばる外部雇用アノテータとの意思疎通と進捗管理をサポートするため、Skype, Ustream 等のインターネット上で提供されているコミュニケーションツールを積極的に活用することで、低コストで高速にアノテーションの充実を実施するシステムづくりに挑戦し、成功をおさめた。また、育児中の主婦（主夫）や、リタイアした研究者をアノテータとして採用することで、限定された人数ながら、こうした人材の支援や再発掘を実現できたことも特筆すべき点である。

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	0	0	0	0	0
人　件　費（千円）	0	11,634	13,178	7,139	7,147	39,098
業務実施費（千円）	0	28,366	14,458	15,043	14,671	72,538
一般管理費（千円）	0	4,000	2,764	2,218	2,182	11,164
合　計　（千円）	0	44,000	30,400	24,400	24,000	122,800

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	KazusaAnnotation http://a.kazusa.or.jp/	公開	高度情報集積型データベースのソーシャルブックマーク型サービスで、独自開発のソーシャルブックマーク によるオープンアノテーションを可能にした遺伝子情報統合データベース。H19.10.3 より公開。基本機能の開発および改善をほぼ完了。植物および関連微生物 14 生物種 7253 文献に関して文献マニュアルアノテーションを実施。約 18 万件の遺伝子およびタンパク質複合体の情報を蓄積、公開。
2	KazusaNavigation http://navi.kazusa.or.jp/	公開	高度情報集積型データベースのポータルサービスで、ソーシャルネットワーク型の Elgg をベースに開発。H19.10.12 より公開開始。登録者数 118 名。研究者コミュニティに利用により情報交換の促進による知識を統合。
3	KazusaWiki http://wiki.kazusa.or.jp/	公開	高度情報集積型データベースのデータとりまとめサイト。Wiki タイプのデータ共有サイトである。MediaWiki をベースに開発し H19.10.23 より公開開始。 研究者による情報とりまとめサイト、ゲノムデータの静的コンテンツ、複合体情報の URL エントリーポイントとして情報の統合化。55 ページ。
4	KazusaMart http://mart.kazusa.or.jp/	公開	植物および関連微生物を対象としたゲノムスケールの遺伝子情報管理システムシステムは BioMart を利用している。国内の植物・関連微生物ゲノム情報の統合基盤の構築に向けたゲノムデータの統合、整備を行っている。 H22.03.15 にシアノバクテリア 34 生物種、公開。 H22.10.01 根粒菌 18 生物種、解析情報の整備された主要植物 8 種公開。
5	CyanoBase http://genome.kazusa.or.jp/cyanobase/	公開	シアノバクテリアゲノムデータベース かずさディー・エヌ・エー研究所で 1995 年より開発・運用が行われているデー

			タベース。遺伝子情報等へのアクセス性と再利用性の向上を目的とした新システムの開発・整備。 H21.2.13 公開。
6	RhizoBase http://genome.kazusa.or.jp/rhizobase/	公開	根粒菌ゲノムデータベース かずさディー・エヌ・エー研究所で 2000 年より開発・運用が行われているデータベース。遺伝子情報等へのアクセス性と再利用性の向上を目的とした新システムの開発・整備。 H21.2.13 公開。

（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	GenoDive Pro/Eu http://genodive.org/	公開	DAS 規格に準拠したゲノム情報を統合的に閲覧することが可能な DAS サーバークライアント。ゲノム情報可視化ツールとしてアプリケーション（MacOSX、Windows 版）を配布している。
2	なんでも DAS	限定公開	任意のウェブページに含まれる遺伝子名を抽出して、GFF 形式に変換するサーバ。このサーバを活用することで、インターネット上のあらゆる遺伝子関連情報をゲノム上にマッピングし、情報の統合的な閲覧を可能とする目的で開発。ゲノムブラウザーや Wiki との連携で限定的に利用・運用中。

外部発表実績一覧

（1）セミナー、研究会等イベント開催

通番	タイトル	発表者（代表者）	年月日	開催場所	イベント名称	概要（対象者（層、参加人数）、出席者の主な反応等）
1	ゲノム情報リテラシー入門	中村　保一	2008 年 5 月 30 日	国内（日本大学藤沢キャンパス・神奈川）	統合データベース・日大共催講習会［口頭］	統合データベース講習会 AJACS にてゲノム解析一般に関する参加形式の実習を行った
2	ゲノム情報リテラシー入門	中村　保一	2008 年 8 月 29 日	国内（北海道大学・札幌）	統合データベース・北大 GCOE 共催講習会（AJACS 蝦夷）［口頭］	統合データベース講習会 AJACS にてゲノム解析一般に関する参加形式の実習を行った

3	ゲノム塩基配列を基盤としたライフサイエンス研究情報統合環境	中村　保一	2008年9月8日	国内（明治大学・川崎）	明治大学農学部ワークショップ「大規模オミックス・データの利用と展望」［口頭］	表題の研究会でKazusaAnnotationの紹介を行った
4	KazusaAnnotation Suiteの活用	中村　保一	2008年10月18日	国内（九州大学・福岡）	統合データベース講習会AJACS博多［口頭］	統合データベース講習会 AJACS にてKazusaAnnotationに関する実習を担当した
5	KazusaAnnotation Suiteの活用	中村　保一	2008年10月31日	国内（長浜バイオ大学・長浜）	統合データベース講習会AJACS近江［口頭］	統合データベース講習会 AJACS にてKazusaAnnotationに関する実習を担当した
6	KazusaAnnotation Suiteの活用	中村　保一	2009年1月23日	国内（東京工業大学・横浜）	統合データベース講習会AJACS長津田［口頭］	統合データベース講習会 AJACS にてKazusaAnnotationに関する実習を担当した
7	Genome AnnotationとKazusaAnnotation	岡本　忍	2009年5月15日	国内（日本大学生物資源学部・神奈川）	統合データベース講習会AJACS湘南2［口頭］	統合データベース講習会 AJACS にてKazusaAnnotationに関する実習を担当した
8	KazusaMart：植物関連生物ゲノムデータの統合と管理	藤澤　貴智	2009.6.12	国内（東大浅野キャンパス内武田ホール ）	統合データベースプロジェクトシンポジウム2009［口頭］	表題の研究会でKazusaMartの紹介を行った
9	ラン藻のゲノム解析とその支援ツールKazusaMart	藤澤　貴智	2009.7.4	国内（東京大学駒場キャンパス）	ラン藻ゲノム研究交流会［口頭］	表題の研究会でKazusaMartの紹介を行った
10	Genome AnnotationとKazusaAnnotation	岡本　忍	2009年7月15日	国内（基礎生物学研究所・愛知）	統合データベース講習会AJACS三河［口頭］	統合データベース講習会 AJACS にてKazusaAnnotationに関する実習を担当した
11	Genome AnnotationとKazusaAnnotation	岡本　忍	2009年9月7-8日	国内（北海道大学・北海道）	統合データベース講習会AJACS蝦夷2［口頭］	統合データベース講習会 AJACS にてKazusaAnnotationに関する実習を担当した
12	KazusaMart：RhizoBaseのデータ統合と管理	藤澤　貴智	2009年9月8-10日	国内（長野県松本市 あがたの森文化会館）	植物微生物研究会第19回研究交流会［ポスター］	表題の研究会でKazusaMartの紹介を行った
13	RhizoBaseの新機能	岡本　忍	2009年9月8-10日	国内（あがたの森・長野）	植物微生物研究会［口頭］	表題の研究会でKazusaAnnotationの紹介を行った
14	Genome Annotationと	岡本　忍	2009年10月	国内（長浜バイ	統合データベース講習	統合データベース講習会 AJACS に

	KazusaAnnotation		16 日	オ大学・滋賀)	会 AJACS 近江 2[口頭]	て KazusaAnnotation に関する実習を担当した
15	Genome Annotation と KazusaAnnotation	岡本　忍	2009 年 11 月 16 日	国内（大阪府立大りんくうキャンパス・大阪)	統合データベース講習会 AJACS りんくう 2［口頭］	統合データベース講習会 AJACS にて KazusaAnnotation に関する実習を担当した
16	Arthrospira platensis NIES-39 のゲノム解析	藤澤　貴智	2009 年 12 月 3-4 日	国内（かずさアカデミアパーク)	ラン藻の分子生物学 2009 (CYANO2009)［口頭］	表題の研究会で KazusaAnnotation の紹介を行った
17	シアノバクテリアのゲノム情報基盤	岡本　忍	2009 年 12 月 4-5 日	国内（かずさアカデミアパーク・千葉)	ラン藻ワークショップ［口頭］	表題の研究会で KazusaAnnotation の紹介を行った
18	ウェット研究者が情報技術的に自立するために：統合データベースプロジェクトからの提案	中村　保一	2009 年 12 月 10 日	国内（パシフィコ横浜･神奈川)	第 32 回日本分子生物学会年会 ワークショップ 2W5［主催］	表題のワークショップを主催し、ひろく統合データベースプロジェクトの紹介を行った
19	ゲノムアノテーション情報の高度化と統合化への取り組み	岡本　忍	2010 年 1 月 27 日	国内（かずさ DNA 研究所・千葉)	かずさ研究交流セミナー［口頭］	表題のセミナーで KazusaAnnotation の紹介を行った
20	アノテーションとは？	岡本　忍	2010 年 3 月 10 日	国内（酒類総合研究所・広島)	統合データーベース講習会 AJACS［口頭］	統合データベース講習会 AJACS にて KazusaAnnotation に関する実習を担当した
21	情報メディアとしてのゲノム	岡本　忍	2010 年 3 月 28-29 日	国内（北海道大学・北海道)	GCOE 若手研究異分野セミナー［口頭］	表題のセミナーで KazusaAnnotation の紹介を行った
22	KazusaMart の使い方	藤澤　貴智	2010.6.4	国内（日本大学生物資源学部・神奈川)	統合データベース講習会: AJACS 湘南 3［口頭］	統合データベース講習会 AJACS にて KazusaMart に関する実習を担当した
23	ゲノム注釈の民主化	岡本　忍	2010 年 6 月 14 日	国内（つくば農業環境技術研究所・茨城)	農業環境技術研究所公開セミナー［口頭］	表題のセミナーで KazusaAnnotation の紹介を行った
24	CyanoBase のゲノムリアノテーション	藤澤　貴智	2010.7.24	国内（東京大学教養学部･東京)	ラン藻ゲノム研究会［口頭］	表題の研究会で KazusaAnnotation の紹介を行った
25	ゲノムアノテーション協働作業を支援する手法と技術	岡本　忍	2010 年 7 月 24 日	国内（東京大学教養学部･東京)	ラン藻ゲノム研究会［口頭］	表題の研究会で KazusaAnnotation の紹介を行った
26	情報の可視化：ゲノムブラウザ、Cytoscape	岡本　忍	2010 年 8 月 4-5 日	国内（東北大学星稜地区･宮城)	統合データベース講習会 AJACS21 みちのく［ロ	統合データベース講習会 AJACS にて KazusaAnnotation に関する実習を

					頭]	担当した
27	かずさアテーションによるゲノム情報注釈とその運用	岡本　忍	2010年9月8日	国内（国立遺伝学研究所DDBJ・静岡）	DDBJ勉強会［口頭］	表題の研究会でKazusaAnnotationの紹介を行った
28	KazusaAnnotationによるゲノムアノテーション協働作業手法の開発	岡本　忍	2010年10月19日	国内（ラフォーレ修善寺･静岡）	統合データベース技術情報交換ワークショップ［口頭］	表題の研究会でKazusaAnnotationの紹介を行った

（2）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	かずさアノテーションを用いた分散型ゲノムアノテーションの実証実験	岡本　忍	第32回分子生物学会年会ワークショップ［口頭］	2009年12月9-12日	国内（パシフィコ横浜・神奈川）
2	KazusaMart:the integration of genomic data in plants and the related microorganisms	藤澤　貴智	第32回日本分子生物学会（MBSJ2009）［ポスター］	2009年12月10日	国内（パシフィコ横浜・神奈川）
3	微生物ゲノム基盤情報の高度化と統合のためのシステム	岡本　忍	第4回日本ゲノム微生物学会年会	2010年3月7-9日	国内（九州大学・九州）
4	A Trial For Genome Re-annotation Using the Literature and Annotation Reference Information	藤澤　貴智	第4回国際バイオキュレーション会議 Biocuration 2010［ポスター］	2010年10月12-14日	国内（東京国際交流会会館・東京）
5	Development of Cooperative Work Tools and Workflow for Multimodal Genome Annotation	岡本　忍，中尾光輝，藤澤貴智，佐藤修正，金子貴一，吉村英尚，中平有香，阿久津智子，鐘ヶ江弘美，笠井誠，山本純子，作田千代子，眞板寛子，鹿島裕子，中村保一	第4回国際バイオキュレーション会議 Biocuration 2010［ポスター］	2010年10月12-14日	国内（東京国際交流会会館・東京）

（3）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	CyanoBase: the cyanobacteria genome database update 2010.	Nakao M, Okamoto S, Kohara M, Fujishiro T, Fujisawa T, Sato S, Tabata S, Kaneko T, Nakamura Y.	Nucleic Acids Research	2010 Jan; **38 (Database issue),** D379-381.	
2	Genomic Structure of an Economically Important Cyanobacterium, *Arthrospira (Spirulina) platensis* NIES-39	Takatomo Fujisawa, Rei Narikawa, Shinobu Okamoto, Shigeki Ehira, Hidehisa Yoshimura, Iwane Suzuki, Tatsuru Masuda, Mari Mochimaru, Shinichi Takaichi, Koichiro Awai, Mitsuo Sekine, Hiroshi Horikawa, Isao Yashiro, Seiha Omata, Hiromi Takarada, Yoko Katano, Hiroki Kosugi, Satoshi Tanikawa, Kazuko Ohmori, Naoki Sato, Masahiko Ikeuchi, Nobuyuki Fujita, and Masayuki Ohmori	DNA Research	2010 **17,** 85-103.	

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	「ライフサイエンス統合データベース開発運用」 （統合データベース開発：ワークフロー技術を用いた統合 DB 環境構築）
実 施 機 関 名	独立行政法人　産業技術総合研究所　生命情報工学研究センター
代 表 研 究 者 名	浅井潔

１．課題開始時における達成目標

利用者が得たい情報（知識）を、パソコンなどの端末から要求すると、必要なデータ、解析手法などを、国内、海外から自動的に選び、データベースと解析ツールのワークフローを作成し、最適な計算資源を使って解析を行なう統合 DB 環境を構築する。

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

センター内に Grid 環境を構築し、以下の Web ワークフロー、アクティブ・ワークフローおよびプラットフォームを開発し、一般公開した。(カッコ内は公開日時)。各ワークフローの詳細は、「（２）進捗および成果」に記す。

Web ワークフローに関しては、DBCLS の TogoProt と連携し、TogoProt で検索された配列で Web ワークフローを利用できるようにした。

１）Web ワークフロー　(http://togo.cbrc.jp/web.html)

・Protein Annotation Workflow （2008 年 12 月） http://togo.cbrc.jp/wf/PAW_EnterSequence.html?Eng
・Comparative Protein Information Workflow（2009 年 3 月）
http://togo.cbrc.jp/wf/comp_prot_info_enter.html?Eng
・Protein Modelling Workflow （2009 年 11 月） http://togo.cbrc.jp/wf/modelling_enter.html?Eng

２）アクティブ・ワークフロー (http://togo.cbrc.jp/active.html)

・ASIAN Active Workflow（2010 年　1 月） http://togo.cbrc.jp/active_local_asian.html
・POODLE Active Workflow（2010 年 5 月） http://togo.cbrc.jp/active_local_poodle.html
・CentroidFold Active Workflow（2010 年 6 月） http://togo.cbrc.jp/active_local_centroid.html
・Modelling Active Workflow（2010 年 7 月） http://togo.cbrc.jp/active_local_modelling.html
・WolfPSORT Active Workflow （2010 年 8 月） http://togo.cbrc.jp/active_local_wolfpsort.html
・LASTActive Workflow （2010 年 8 月） http://togo.cbrc.jp/active_local_last.html
・BLAST Active Workflow（2010 年 8 月） http://togo.cbrc.jp/active_local_blast.html
・MAFFT Active Workflow（2010 年 8 月） http://togo.cbrc.jp/active_local_mafft.html

３）プラットフォーム（ケーススタディ）

・PhylogeneticTree Active Workflow（2010 年 9 月） http://togo.cbrc.jp/active_local_phylo.html
・RNA Structure Prediction Active Workflow（2010 年 10 月）
http://togo.cbrc.jp/active_local_rna_prediction.html
・Protein Structure Prediction Active Workflow（2010 年 10 月）
http://togo.cbrc.jp/active_local_protein_prediction.html

（２）進捗及び成果

１．Web ワークフロー　(http://togo.cbrc.jp/web.html)

Web ワークフローは、利用者が得たい情報を国内、海外から自動的に収集、処理して表示するワークフローである。後述のアクティブ・ワークフロー利用者が処理をアレンジできるのに対して、本ワークフローは、各処理のデータベースやプログラムを変更できるが、処理の流れを変更することはできない。

１－１）タンパク質アノテーションワークフロー（Protein　Annotation　Workflow）

本ワークフローは、立体構造未知のタンパク質に関し、構造及び機能のヒントとなる情報を幅広く実験研究者等に提供することを目的としており、各種プログラム等を Grid により効率的に分散処理し、従来と比し短時間で結果を表示する。本ワークフローは、平成 20 年 12 月 27 日に構築が終了し、一般公開した。下記の図１、図２は、本ワークフローの概念図及び結果画面である。ユーザからアミノ酸配列を受取り、二次構造予測、埋れ残基予測、フォールド認識、ディスオーダー予測、膜タンパク質オールベータ・ベータシート予測、細胞内局在予測を Grid にて分散処理を行う一方、データベース検索及び疎水性予測の実行を他のサーバへ依頼し結果を取得後、全ての結果をユーザが解析し易いよう配置し出力する。

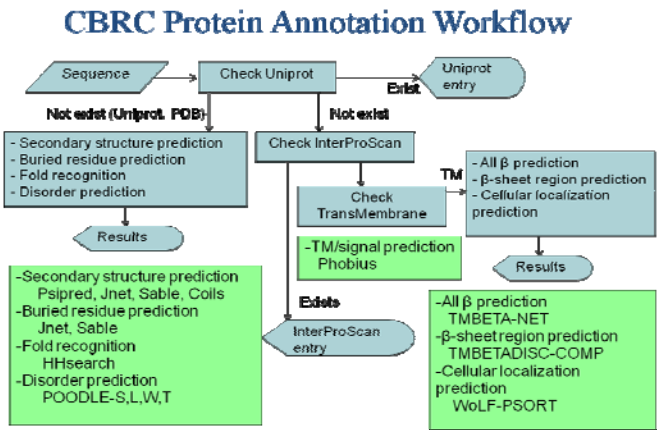


図１　タンパク質アノテーションワークフローの概念図

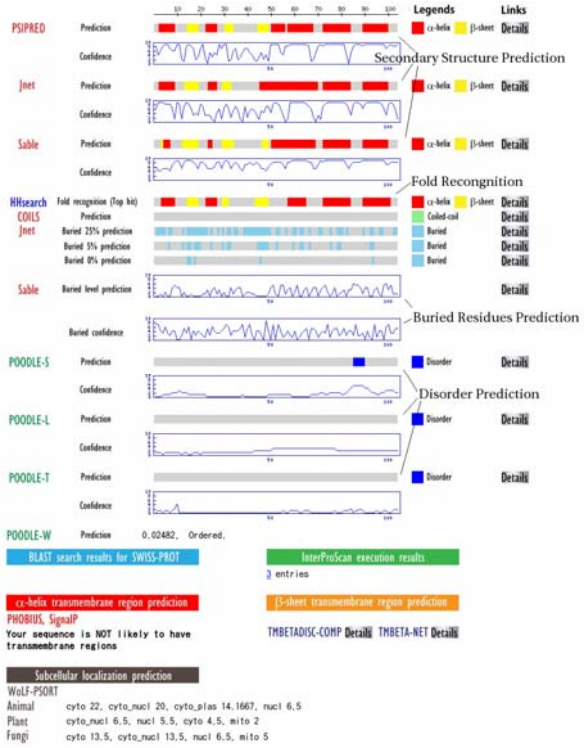


図２　タンパク質アノテーションワークフローの検索画面

１－２）タンパク質比較情報ワークフロー（Protein Comparative Information Workflow）

本ワークフローは、相同なタンパク質を比較することで保存部位等構造上重要な部位を表示し、実験研究者等に提供することを目的としている。本ワークフローは、平成 21 年 3 月 31 日に構築が終了し一般公開した。下記の図３、図４は本ワークフローの概念図及び結果画面である。ユーザからアミノ酸配列を受取り、相同タンパク質を検索、その結果からユーザがいくつかのタンパク質を選択し、マルチプルアラインメントを実行する。その結果より保存性が高い残基を表示する。また、二次構造予測結果も同様にマルチプルアラインメントし、保存性が高い二次構造を表示する。本ワークフローも Grid にて効率的に分散処理を行う。結果は立体構造が存在する場合はその上に表示する。

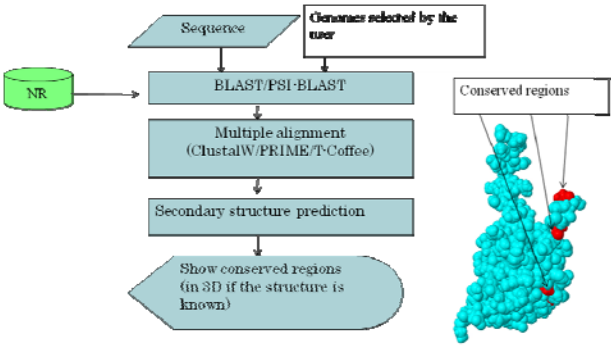


図３　タンパク質比較情報ワークフローの概念図

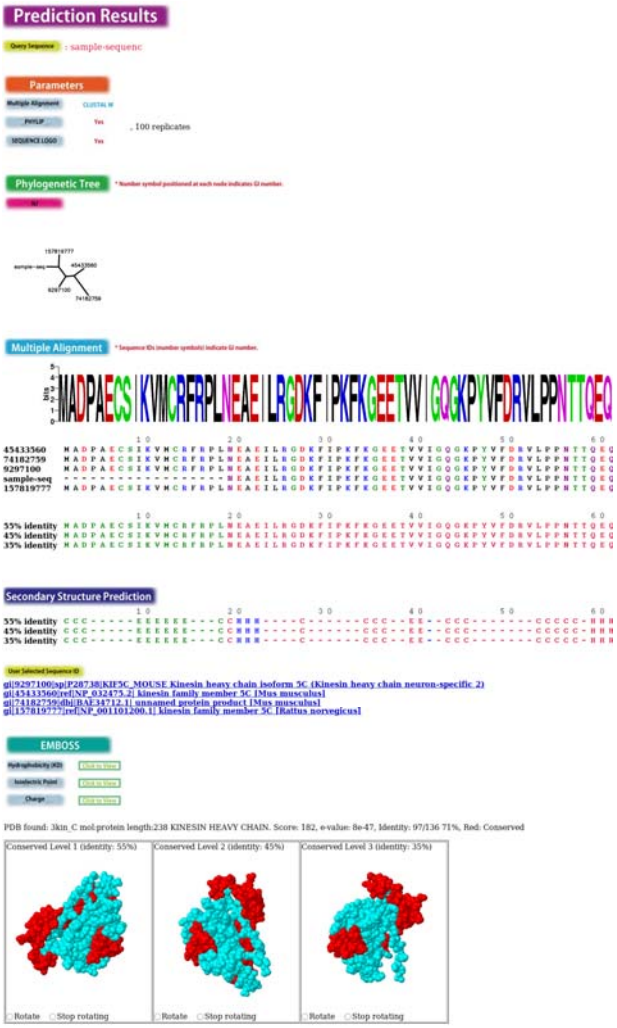


図４　タンパク質比較情報ワークフローの結果画面

１－３）タンパク質立体構造モデリングワークフロー（Protein　Structure　Modelling　Workflow）

本ワークフローは、立体構造未知のタンパク質の立体構造のモデル化を行う。モデル化の前処理として、BLAST/PSI-BLAST及びfold　recognitionにより元になる構造テンプレートを探し出し、そのテンプレートをモデル構築のプログラムに渡す。ユーザからアミノ酸配列を受取り、構造テンプレートを取得後、テンプレート毎に分散処理を行う。下記図５、図６は本ワークフローの概念図及び結果画面である。平成21年11月末に構築が終了し一般公開した。

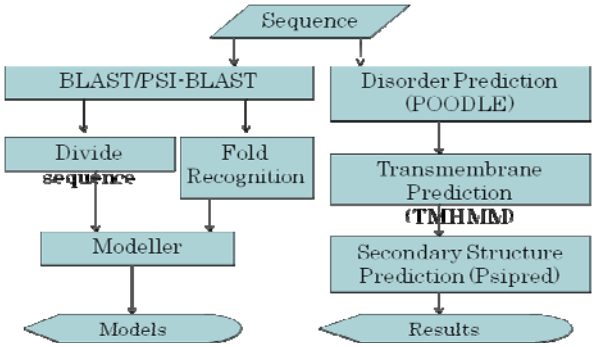


図５　タンパク質立体構造モデリングワークフロー概念図

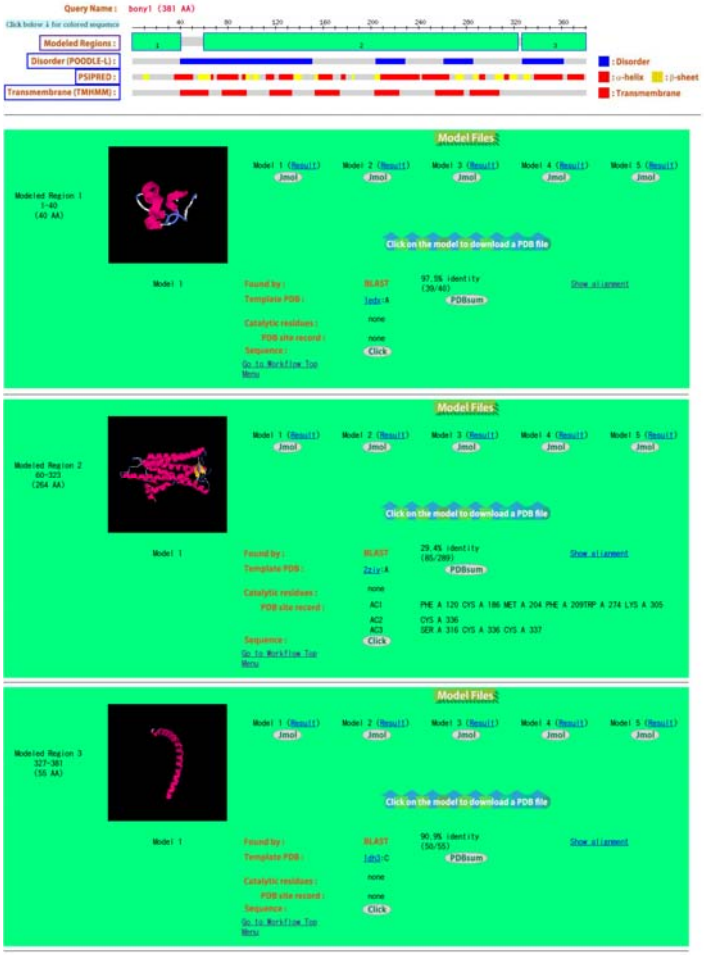


図６　タンパク質立体構造モデリングワークフロー結果画面

２．アクティブ・ワークフロー (http://togo.cbrc.jp/active.html)

アクティブ・ワークフローは操作性・柔軟性に優れたプラットフォーム(KNIME)を使用し、プログラムをインタラクティブに接続し、選択的に組み合わせる事により、独自の解析プラットフォームによるワークフロー解析を可能とするものである。

２－１）ASIANアクティブ・ワークフロー（ASIAN　Active　Workflow）

本ワークフローは、CBRCで開発されたASIAN（Automatic System for Inferring A Network、遺伝子ネットワーク推定システム）をワークフロープラットフォーム（ワークフロー構築ソフトウェア）であるKNIMEを使用してワークフロー化したものである。ASIANは階層クラスタリングとグラフィカル・ガウシアン・モデリング（GGM）を組み合わせたネットワーク推定ツールであるが、これをいくつかのワークフローを構成する部品（ノード）に分け、１つのワークフローとして構築した。実行形態には2種類あり、１つはThin-X、もう１つはLocal PCである。Thin-Xでは、全ての処理をCBRCのサーバ上で行い、画面表示のみユーザ自身のPCで行う。一方、Local　PCは全ての処理をユーザ自身のPC上で行う。下記図７に実行イメージ及び結果表示画面を示す。平成22年1月末に構築が終了し一般公開した。

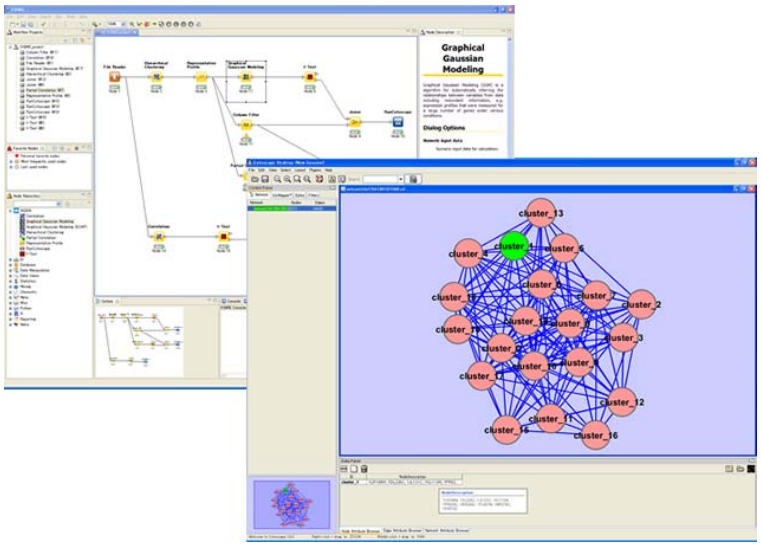


図７　ASIAN アクティブ・ワークフロー(左)、結果表示画面(右)

２－２）POODLEアクティブ・ワークフロー（POODLE Active Workflow）

本ワークフローは、CBRCで開発されたタンパク質ディスオーダー予測POODLE（Prediction Of Order and Disorder by machine LEarning ）をSOAP化し、ワークフロープラットフォーム（ワークフロー構築ソフトウェア）であるKNIMEを使用してワークフロー化したものである。入力されたアミノ酸配列からディスオーダー領域を予測する。平成22年５月に構築が終了し一般公開した。

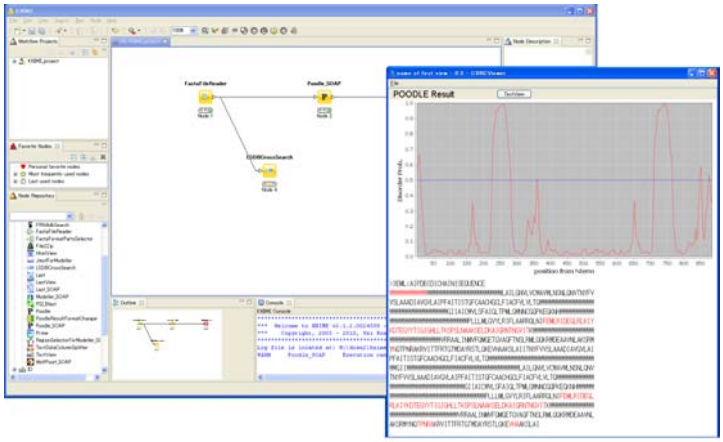


図８　POODLE アクティブ・ワークフロー(左)、結果表示画面(右)

２－３）CentroidFoldアクティブ・ワークフロー（CentroidFold Active Workflow）

本ワークフローは、CBRCで開発されたCentroidFold（RNA2次構造予測）をSOAP化し、平成22年6月に構築が終了し一般公開した。

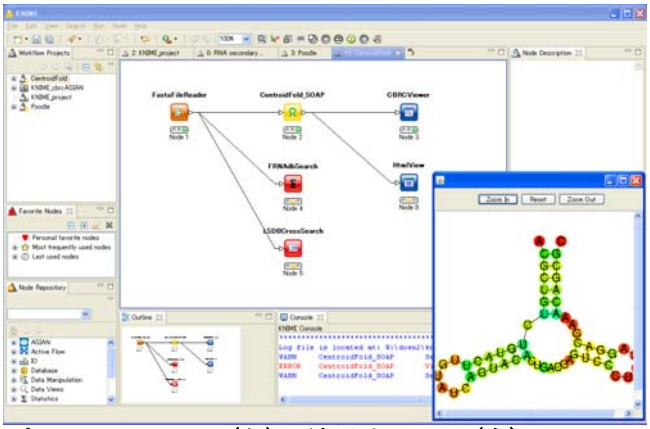


図９　CentroidFold アクティブ・ワークフロー(左)、結果表示画面(右)

２－４）Modellingアクティブ・ワークフロー（Modelling Active Workflwo）

本ワークフローは、Webワークフローのタンパク質立体構造モデリングワークフローをもとにSOAP対応し、各SOAPサービスをKNIMEのコンポーネントとして使用している。平成22年７月に構築が終了し一般公開した。

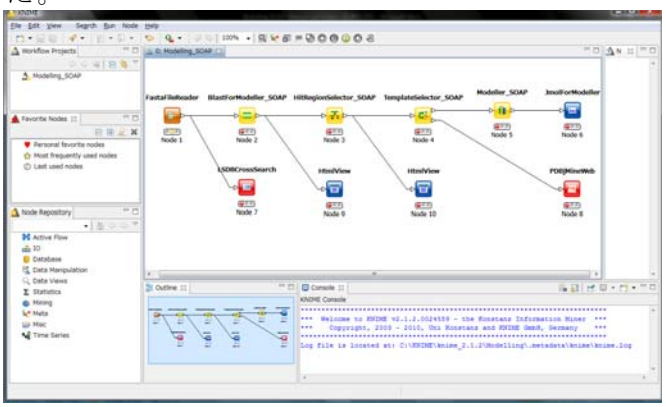


図 10　Modelling アクティブ・ワークフロー

２－５）WolfPSORTアクティブ・ワークフロー（WolfPSORT Active Workflow）

本ワークフローは、CBRCで開発されたタンパク質細胞内局財予測ソフトウェアWolfPSORT（Protein Subcellular Localization Prediction）をワークフロープラットフォーム（ワークフロー構築ソフトウェア）であるKNIMEを使用してワークフロー化したものである。入力されたアミノ酸配列から重み付きkNN法を用いた細胞内局財予測を行う。平成22年８月に構築が終了し一般公開した。

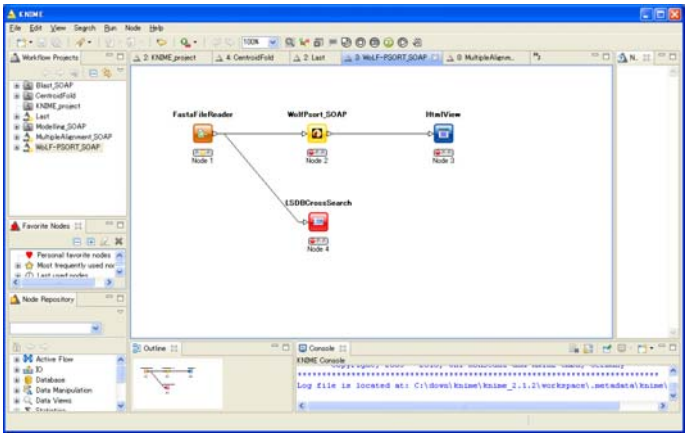


図１１　WolfPSORT ワークフロー画面

２－６）LASTアクティブ・ワークフロー（LAST Active Workflow）

本ワークフローは、CBRCで開発された配列比較プログラムLASTをSOAP化し、平成22年8月に構築が終了し一般公開した。

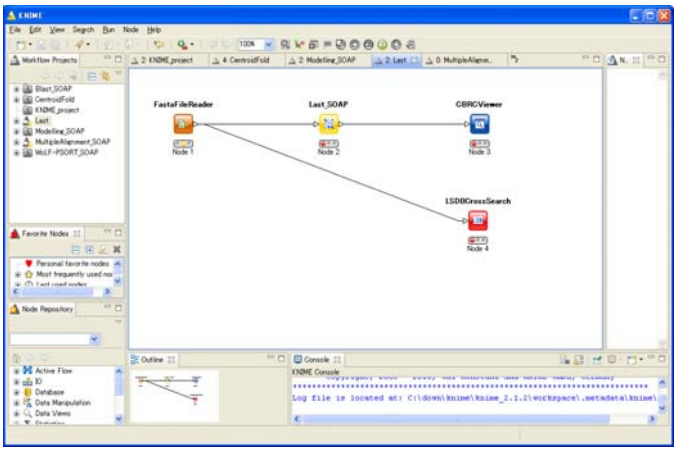


図１２　LAST ワークフロー画面

２－７）Blastアクティブ・ワークフロー（Blast Active Workflow）
本ワークフローは、配列検索プログラムBlastをSOAP化し、平成22年８月に構築が終了し一般公開した。

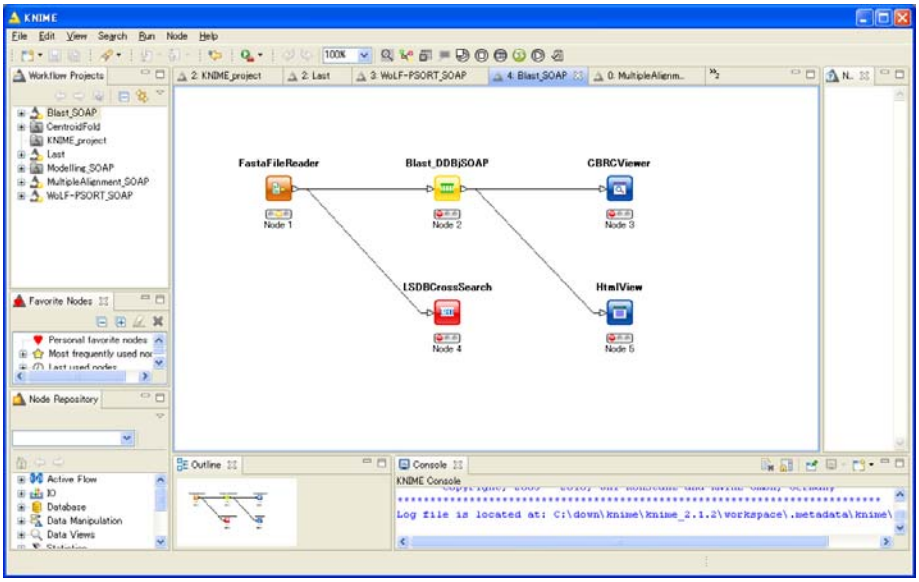


図１３　Blast ワークフロー画面

２－８）MAFFTアクティブ・ワークフロー（MAFFT Active Workflow）
本ワークフローは、マルチプルアライメントプログラムMAFFTをSOAP化し、平成22年8月に構築が終了し一般公開した。

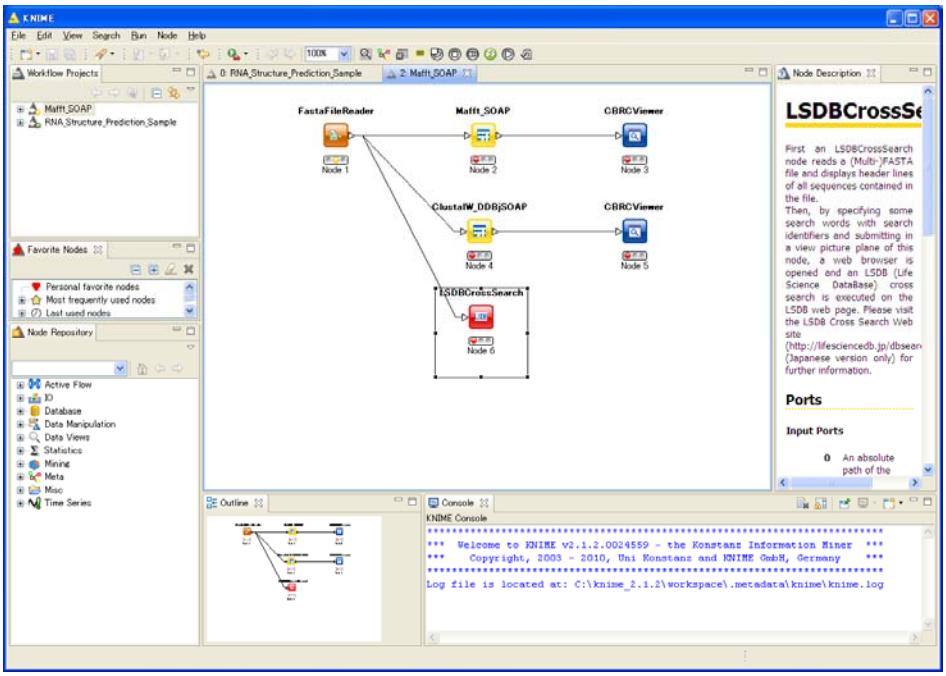


図１４　MAFFT ワークフロー画面

２－９）系統樹作成アクティブ・ワークフロー（PhylogeneticTree Active Workflow）
本ワークフローは、入力された配列(DNA, RNA, アミノ酸)のマルチプルアライメントを行い、系統樹を作成する。このワークフローは、各 SOAP サービスを KNIME のコンポーネントとして使用している。

平成 22 年 9 月に構築が終了し一般公開した。

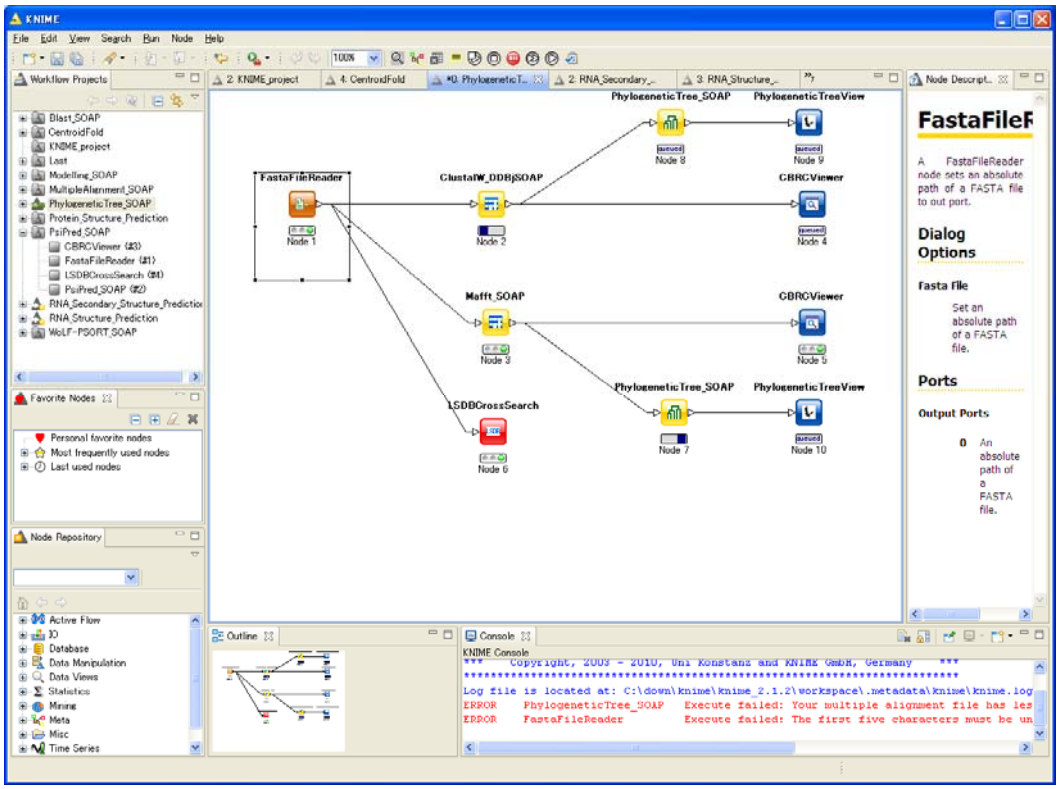


図１５　系統樹作成ワークフロー画面

２－１０）タンパク質立体構造モデリング・アクティブ・ワークフロー（Protein Structure Prediction Active Workflow）

本ワークフローは、入力されたアミノ酸配列からモデリングやディスオーダー予測等を行う。このワークフローは、各SOAPサービスをKNIMEのコンポーネントとして使用している。平成22年10月に構築が終了し一般公開した。

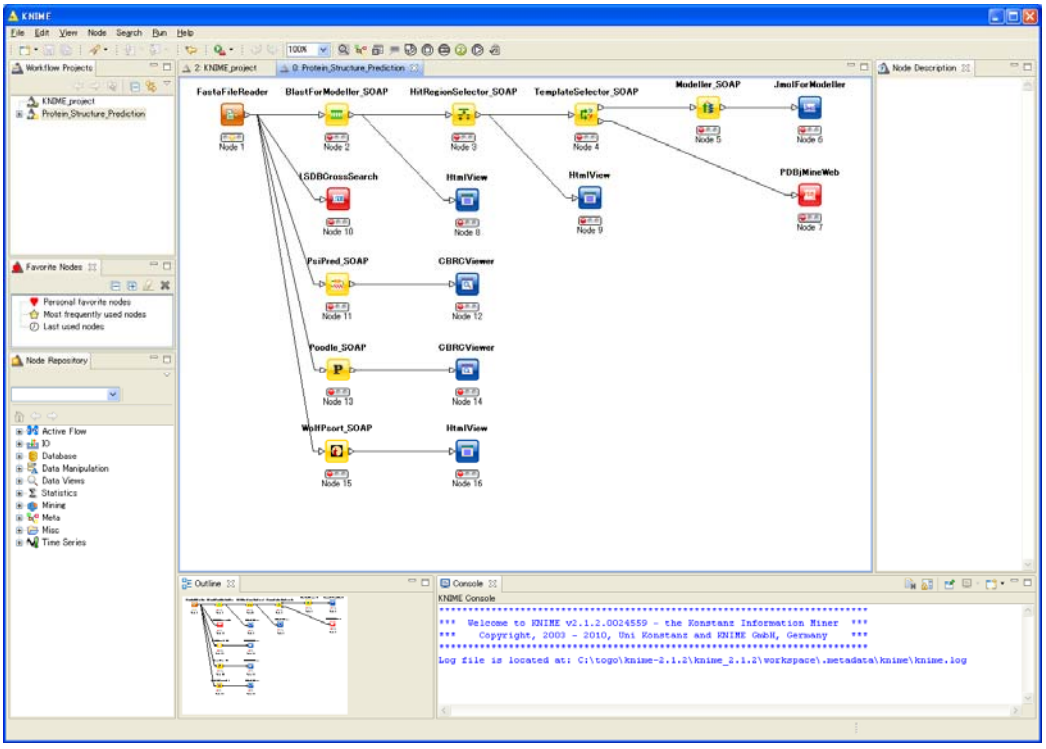


図１６　タンパク質立体構造モデリング・アクティブ・ワークフロー画面

２－１１）RNA構造予測アクティブ・ワークフロー（RNA Structure Prediction Active Workflow）

本ワークフローは、RNA の構造予測等を行う。このワークフローは、各 SOAP サービスを KNIME のコンポーネントとして使用している。平成 22 年 10 月に構築が終了し一般公開した。

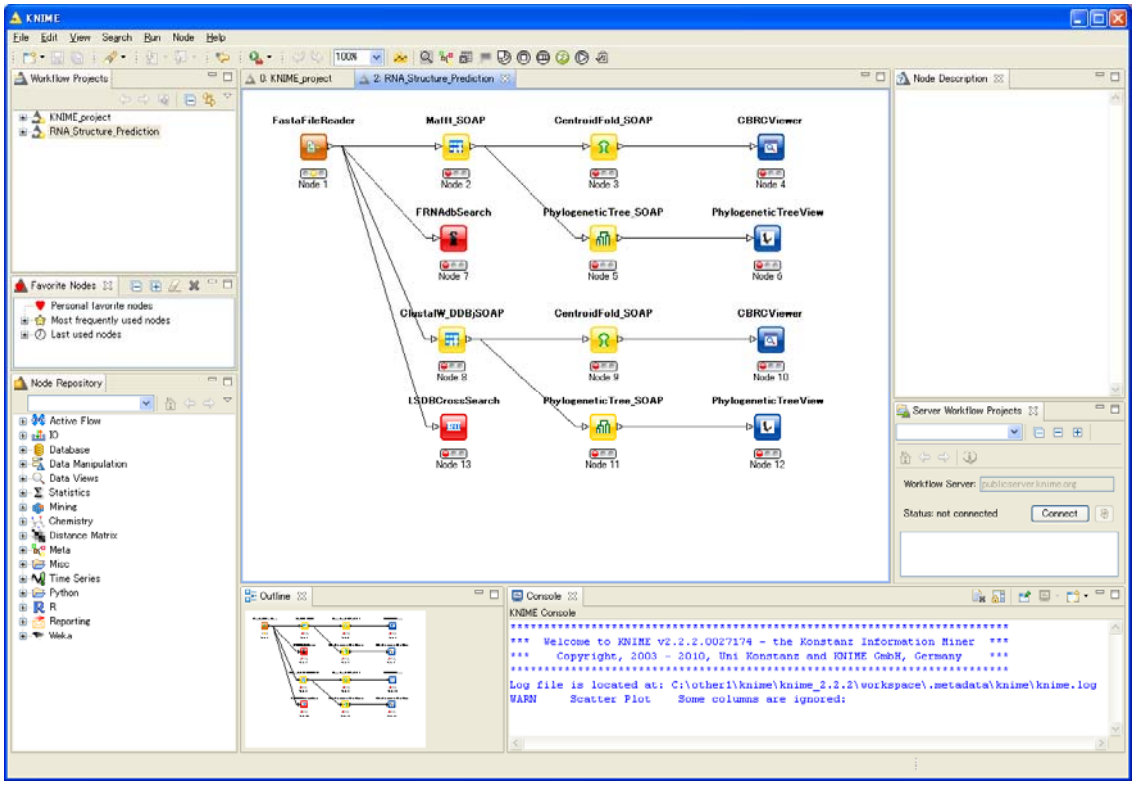


図１６　RNA 構造予測アクティブ・ワークフロー画面

３．当初目標に対する達成度

目標とした Grid を用いた統合 DB 環境を構築し、利用者が得たい情報（知識）を、パソコンなどの端末から要求すると、国内、海外から必要なデータ、解析手法など取得して提供するワークフロー（Web ワークフローとアクティブ・ワークフロー）を開発し、一般公開した。

2010 年 10 月現在におけるアクセス数は以下の通り。

アクセス数：約１６０/月　月間転送量：126MB（7 月）200MB（8, 9 月）760MB（10 月）

４．中間評価に対する対応

産業技術総合研究所　生命情報工学研究センター　統合 DB 情報基盤サイトホームページの情報量が少ないとの指摘を受け、開発状況（News）・研究成果の発信、マニュアルの充実など情報量を増やすとともに、利用者の利便性に注力してホームページの拡充を行った。

５．他機関との連携

Web ワークフローにおいて、DBCLS の TogoProt と連携し、TogoProt で検索された配列で Web ワークフローを利用できるようにした。

６．今後の見通し、計画、展望

・　これまで開発したノードや既存のノードを繋げたケーススタディの公開

・　LINUX や MAC にも対応したノードの開発

７．全体総括

バイオインフォマティクスの各種ソフトウェアとライフサイエンスのデータベースを連携して動作させ、高度な解析を行う Web ワークフローと、ユーザ自らがワークフローを設計できるアクティブ・ワークフロー環境を構築した。本開発で構築された環境は、ライフサイエンス統合データベースの価値を大きく向上させる効果があり、より多くのユーザにライフサイエンス統合データベースの利用する機会を提供するものである。

８．特記事項

ライフサイエンスのデータベースにおいては、単に必要な情報を検索して表示するだけでなく、タンパク質・RNA の構造情報の解析やモデリング、高度な配列情報解析などを組み合わせることによってのみ、高付加価値のある情報を抽出することができる。本開発では、ライフサイエンス統合データベースセンター（DBCLS）との密接な連携の下にワークフロー環境を構築した。個々のバイオインフォマティクス・ツールやデータベース検索の SOAP 化に留まらず、ローカルおよびリモートの計算資源を組み合わせて活用できる柔軟な枠組みは、独自性が高い試みである。

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	1,575	420	0	0	1,995
人　件　費（千円）	0	10,474	7,157	7,168	7,216	32,015
業務実施費（千円）	0	16,951	12,096	12,505	10,420	51,972
一般管理費（千円）	0	2,900	1,967	1,967	1,764	8,598
合　計　（千円）	0	31,900	21,640	21,640	19,400	94,580

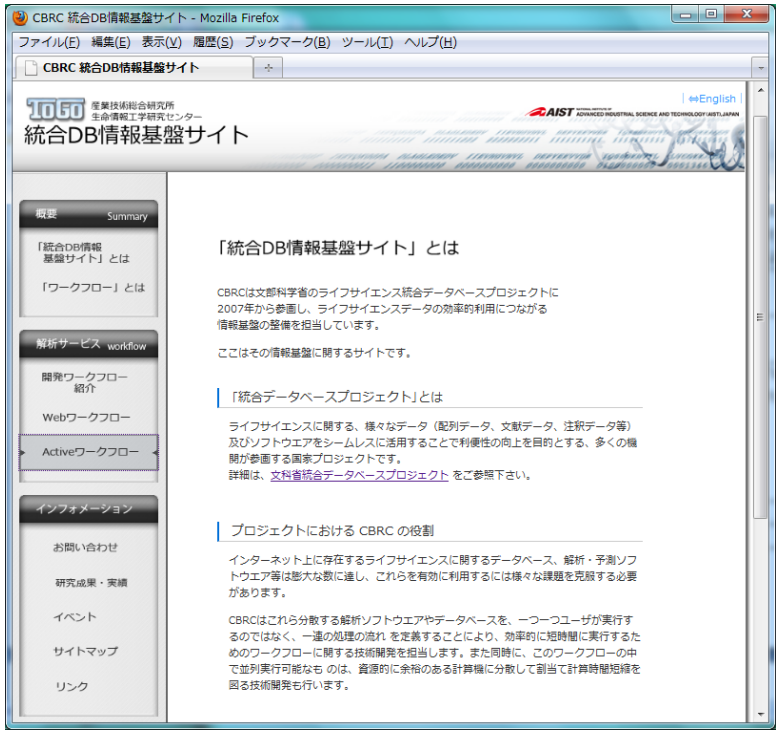
整備実績一覧

（1）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	CBRC　タンパク質構造情報ワークフロー	限定公開	構造に関する予測プログラム等を Grid により効率的に分散処理し、結果を実験研究者等に提供する（プロトタイプ）。平成 20 年 8 月末よりプロジェクト関係者に限定し公開開始。後、下記 2 のタンパク質アノテーションワークフローに統合した。
2	CBRC　タンパク質アノテーションワークフロー http://tgweb01.cbrc.jp/wf/PAW_About.html?Eng	公開	立体構造未知のタンパク質に関し、構造及び機能のヒントとなる情報を幅広く実験研究者等に提供する。二次構造予測、埋れ残基予測、フォールド認識、ディスオーダー予測、膜タンパク質オールベータ・ベータシート予測、細胞内局在予測を Grid にて分散処理を行う一方、データベース検索及び疎水性予測の実行を他のサーバへ依頼し、結果を取得後、全ての結果をユーザが解析し易いよう配置し出力する。平成 20 年 12 月末より一般公開。
3	CBRC　タンパク質比較情報ワークフロー http://togo.cbrc.jp/wf/comp_prot_info_enter.html?Eng	公開	相同なタンパク質を比較することで保存部位等構造上重要な部位を表示し、実験研究者等に提供する。マルチプルアラインメントを用い保存性が高い残基及び二次構造を表示する。平成 21 年 3 月末より一般公開。
4	タンパク質立体構造モデリングワークフロー http://togo.cbrc.jp/wf/modelling_enter.html?Eng	公開	このワークフローはアミノ酸配列を入力とし、二次構造予測、ディスオーダー予測等を行い、タンパク質立体構造のテンプレートとなる構造をデータベースから見つけ出し、それを基に立体構造モデリングを行うことで、タンパク質の配列情報から 3 次元立体構造の構築を可能とするものである。平成 21 年 11 月に一般公開。
5	遺伝子ネットワーク推定システム（ASIAN-K） http://togo.cbrc.jp/active_local_asian.html	公開	ASIAN-K は、遺伝子の発現量（発現プロファイル）を入力とし、クラスタリング技術を用い遺伝子間ネットワーク推定を可能とするアクティブ・ワークフローである。平成 21 年 1 月より一般公開。
6	POODLE Active workflow http://togo.cbrc.jp/active_local_poodle.html	公開	POODLE Active workflow では、CBRC で開発されたディスオーダー予測 POODLE（Prediction Of Order and Disorder by machine Learning）の SOAP サービスを KNIME のノード（コンポーネント）として使用しています。平成 22 年 5 月より一般公開。
7	CentroidFold Active Workflow http://togo.cbrc.jp/active_local_centroid.html	公開	CentroidFold Active workflow では、CBRC で開発された RNA の 2 次構造予測 CentroidFold の SOAP サービスを KNIME のノード（コンポーネント）として使用しています。平成 22 年 6 月より一般公開。
8	Modelling Active Workflow http://togo.cbrc.jp/active_local_modelling.html	公開	Modelling Active Workflow では、入力されたアミノ酸配列からホモロジーモデリングを行います。 WEB 版をもとに SOAP 対応し、各 SOAP サービスを KNIME のノード（コンポーネント）として使用しています。平成 22 年 7 月より一般公開。

9	WOLFPSORT Active workflow http://togo.cbrc.jp/active_local_wolfpsort.html	公開	WOLFPSORT Active workflow では、CBRC で開発されたタンパク質局在予測ツール WOLFPSORT（Protein Subcellular Localization Prediction）の SOAP サービスを KNIME のノード（コンポーネント）として使用しています。平成 22 年 8 月より一般公開。
10	LAST Active Workflow http://togo.cbrc.jp/active_local_last.html	公開	LAST Active workflow では、CBRC で開発された配列比較プログラム LAST の SOAP サービスを KNIME のノード（コンポーネント）として使用しています。平成 22 年 8 月より一般公開。
11	Blast Active Workflow http://togo.cbrc.jp/active_local_blast.html	公開	Blast Active Workflow では、Blast の SOAP サービスを KNIME のノード（コンポーネント）として使用しています。平成 22 年 8 月より一般公開。
12	MAFFT Active Workflow http://togo.cbrc.jp/active_local_mafft.html	公開	MAFFT Active Workflow では、MAFFT の SOAP サービスを KNIME のノード（コンポーネント）として使用しています。平成 22 年 8 月より一般公開。
13	PhylogeneticTree Active Workflow http://togo.cbrc.jp/active_local_phylo.html	公開	PhylogeneticTree Active Workflow では、入力された配列(DNA, RNA, アミノ酸)のマルチプルアライメントを行い、系統樹を作成します。このワークフローは、各 SOAP サービスを KNIME のノード（コンポーネント）として使用しています。平成 22 年 9 月より一般公開。
14	Protein Structure Prediction Active Workflow http://togo.cbrc.jp/active_local_protein_prediction.html	公開	Protein Structure Prediction Active Workflow では、入力されたアミノ酸配列からモデリングやディスオーダー予測等を行います。このワークフローは、各 SOAP サービスを KNIME のノード（コンポーネント）として使用しています。平成 22 年 10 月より一般公開。
15	RNA Structure Prediction Active Workflow http://togo.cbrc.jp/active_local_rna_prediction.html	公開	RNA Structure Prediction Active Workflow は、RNA の構造予測等を行うワークフローです。このワークフローは、各 SOAP サービスを KNIME のノード（コンポーネント）として使用しています。平成 22 年 10 月より一般公開。

（2）その他の成果物（（1）に該当しないもの）

通番	名称	公開／未公開	概要
16	産業技術総合研究所　生命情報工学研究センター　統合 DB 情報基盤サイトホームページ http://togo.cbrc.jp/	公開	産業技術総合研究所　生命情報工学研究センターにおける本プロジェクトの成果を集約したホームページ。本プロジェクトの概要、（2）で開発したワークフローの公開などを本サイトから利用できる。 

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	ワークフローによる配列及び	野口　保、田	2010 年 10 月	ライフサイ	統合データ	CBRC で開発した、Web ワークフローとアクティブワー

	構造解析	代 俊行	4 日	エンス統合データベースセンター	ベース講習会：AJACS 本郷 7	クフローの説明および実習を行った。（対象者（社会人１１名、学生 7 名）、アクティブワークフローの環境作成で戸惑っている講習者が多く見受けられた。）

（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	Life Science Database Integration Project, Workflow 2010	InCob Tokyo2010	2010 年 9 月 26 日～28 日	

（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	統合データベース・ワークフロー、ポスター	田代 俊行	産総研・産技連 LS-BT 合同研究発表会	2009 年 1 月 29 日～30 日	
2	情報統合基盤技術によるワークフロー解析、口頭	福井 一彦、田代 俊行、矢葺 幸光、堀本 勝久、野口 保、浅井 潔	産総研 生命情報工学研究センター2009 年度研究報告シンポジウム	2009 年 12 月 3 日	
3	情報統合基盤技術によるアクティブ・ワークフロー解析、ポスター	福井 一彦、田代 俊行、矢葺 幸光、野口 保、浅井 潔	平成 21 年度 産総研・産技連 LS-BT 合同研究発表会	2010 年 2 月 4 日	
4	Life Science Database Integration Project, Workflows、ポスター	田代 俊行、矢葺 幸光、福井 一彦、野口 保、浅井 潔	産総研 生命情報工学研究センター2010 年度研究報告シンポジウム	2010 年 7 月 28 日～30 日	
5	ライフサイエンス統合 DB プロジェクト ワークフロー ２０１０、ポスター	田代 俊行	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～	平成 2010 年 10 月 5 日	

（4）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	The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows	Katayama, T., …, Noguchi, T., …., Tashiro, T., …., Asai, K., Takagi, T.	Journal of Biomedical Semantics	2010, 1:8	

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	統合データベース開発：専門用語辞書管理システムと専門用語解析技術の開発
実 施 機 関 名	奈良先端科学技術大学院大学
代 表 研 究 者 名	松本裕治

１．課題開始時における達成目標

専門用語獲得技術の開発を目標として，以下の３つの項目を設定した

①┌専門用語辞書システムの設計

専門用語間の意味的類似性，同義性を判定する基礎技術を開発し，用語の意味関係を記述することのできる用語辞書システムの設計と開発を行う．

②専門用語解析技術の開発

専門分野のテキストには複雑な専門用語が数多く出現する．特に，複数の類似の用語が並列表現として出現する場合や部分的に省略が行われる場合など，用語辞書の充実だけでは完全な用語抽出を行うことができない．これに対処する解析技術として，専門用語の並列構造解析技術，および，専門用語の内部構造解析技術の開発を行う．

③専門用語タグ付け手法の設計（その後，「専門用語抽出ツールの設計と開発」に変更）

専門用語解析の基礎データの蓄積を行うため，専門分野テキストに対して，用語の出現箇所の特定や用語の種類の分類を行うため，タグ付けの表現法，および，動タグ付け手法を設計，開発する。

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

②┌専門用語辞書システムの設計

９５００語のライフサイエンス辞書を対象に，新規用語の登録，同義語情報を記述，種々の検索や内部構造のアノテーション機能をもつ専門用語辞書システムの開発を行った．用語の意味的類似性，同義性判定は，項目③の類義用語検索機能として実現した．

②専門用語解析技術の開発

専門用語の並列構造解析技術は，より一般的に，文書内の構文的並列構造解析手法として開発した．専門用語の内部構造解析技術については，文字単位の解析技術を開発し，また，訓練データとして，専門用語辞書システムの機能を利用して，約３５００語の用語の内部構造アノテーションを行った．

③専門用語抽出ツールの設計と開発

文書中の用語の検索を行い，その用語の利用文脈とシソーラス内での位置を表示するシステムを開発した．また，シソーラスにはない新規の用語とその使用文脈を提示することにより，類似の文脈で用いられている用語とそのシソーラス上での位置を示すことにより，新規用語のシソーラス登録を支援するシステムを開発した．

（２）進捗及び成果

項目①の専門用語辞書システムについては，種々の情報とともに専門用語を格納し，情報の付与や検索を行うことができる辞書システムを開発した．ライフサイエンス辞書（約９５０００語）を格納し，

以下のような機能を実装した.

1． 新しい用語の辞書への登録

2． 表層形，読み，シソーラスコードなどの情報を用いた用語検索機能

3． 用語の識別子を共有させることによる用語間の同義情報の登録

4． 複数の用語辞書の同時管理．複数の辞書に含まれる用語は共通部分を共有して保存し，同時に複数の辞書を検索することが可能

5． 用語の内部構造のアノテーション機能：複合語となっている用語に対して，要素間の係り受け関係をアノテーションすることができる．また，要素となる語を指定して検索が可能

6． 利用者の管理機能：３段階の利用者の階層を定義することができ，アクセスできる情報の指定や，内部情報の修正権限などを管理できる.

以上，当初計画のすべての機能をもつ辞書システムを実装した．用語の意味的類似性，同義性の判定については，項目③の類義語検索機能として実現した．本システムの表示画面を別紙参考図の図１～図３に示す.

項目②の専門用語解析技術については，次の２つの解析手法を開発した.

1． 文書中の構文的な並列構造の解析技術：専門文書に頻出する構文上の並列構造は，文書解析における困難な問題の一つである．このような構造を解析するため，並列構造間の単語列としての類似性と文法制約の両方を考慮した解析手法を開発した.

2． 専門用語の内部構造解析技術：専門用語は複合語の形をとることが多く，しかも用語の一部を共有したり，並列の構造をもったりするなど，一般的な複合語に比べて遥かに複雑な構造をもつ．この内部構造解析を実現するため，文字列単位の構造記述法を提案し，また，内部構造アノテーションが施された事例を訓練事例として自動学習を行う用語の内部構造解析手法を開発した．同時に，訓練データとして約 3500 語の用語に対して内部構造のアノテーションを行った．別紙参考図の図３に内部構造アノテーションの結果の例を示す.

項目③の専門用語抽出ツールの設計と開発については，雑誌「蛋白核酸酵素」を対象として，次の２つのシステムを開発した.

1． 用語検索とシソーラス表示：指定した用語を出現文脈とともに表示し，かつ，その用語のシソーラス上での位置を表示するシステムを開発した．シソーラス表示は，ノードをクリックすることにより展開や縮退を行うことができる.

2． 類似用語の検索とシソーラス表示：ある用語の使用例文を提示すると，その用語と類似文脈で使用されている用語を検索し，シソーラスとともに表示する．このシステムは，シソーラスに存在しない新規の用語について，その使用例文をシステムに示すことで用法の類似した専門用語を検索し，そのシソーラス内での位置を示すことで，新規用語のシソーラス登録を支援する用途で用いられる.

これらのシステムのシソーラスは，項目①の専門用語辞書システムとリンクしており，マウスによるクリックによって辞書システムへとジャンプすることができる．前者の用語検索システム

３．当初目標に対する達成度

項目①の専門用語辞書システムについては，当初目標の機能をほぼ達成することができた．用語の同義性判定については，類義性の判定に留まり，同義性判定のための支援を行うに留まった.

項目②の専門用語解析技術については，並列構造解析，用語の内部構造解析について，当初目標を達成することができた．後者については，目標としていた解析精度 90%を達成することができた．解析システムの訓練データは，当初 2000 語程度のアノテーションを目標としていたが，約 3500 語に対するアノテーションデータの蓄積を行うことができ，当初予定を上回った．

項目③の専門用語抽出ツールの設計と開発については，新規用語の抽出と，抽出された用語に類義の用語を検索するツールの設計と開発を目標としていた．新規用語の抽出については，Wikipedia に掲載されている用語の意味分類に留まり，任意の用語を抽出するまでには至らなかった．新規用語の類義用語の検索については，当初目標を達成し，また，当初予定していなかったインタフェースの開発を行った．

４．中間評価に対する対応

－

５．他機関との連携

中核機関（ライフサイエンス統合データベースセンター）の方針として示された専門文書の検索の高度化に対して，専門用語への種々の属性情報の追加を行うことのできる専門用語辞書システム，および，新規の用語の抽出と辞書登録の支援として，シソーラス情報の自動推定を研究開発テーマとした．毎年開催された作業分科会での報告とコメント，および，統合データベースシンポジウムでのコメントを参考に，システムの設計と基本機能の実装を行った．

また，代表機関が雑誌「蛋白核酸酵素」から収集した専門用語集を，我々が開発した専門用語辞書システムに格納し，ライフサイエンス辞書や標準病名マスターと統合して検索する機能を提供した．

６．今後の見通し、計画、展望

専門用語辞書システムについては，計画していた機能は実現したが，各辞書項目が持つ属性情報を固定しており，属性情報の追加に対する拡張性に欠けていることがわかった．これは，データベースシステムとして関係データベースを利用していたことに起因するため，現在，非関係データベースシステムの利用へと設計を変更している．今年度中に実装を終わる予定であるが，実利用へ向けたテストを行う必要がある．今後の展望として，本システムを公開し，用語の追加登録や属性情報の追加などを協力して行う環境を構築することが考えられる．

用語の内部構造解析については，90%以上の精度を達成することができた，また，人手によって 3500 語の用語に対して内部構造アノテーションを施すことができた．ただし，ライフサイエンス辞書には 9500 語の用語が登録されており，その多くが複合語である．これらの内部構造解析を行う必要があるが，自動解析を行うには 90%の解析精度はまだ充分ではない．今後は，解析システムの精度向上と，解析誤り検出を精錬化する必要がある．

新規用語の類似用語検索については，既存の手法を上回る手法を提案することができたが，その精度は実用的にはまだ不十分である．手法の精度向上を図るとともに，類似用語の表示やシソーラスへの登録のための支援を効率的に行う手法について検討していく必要がある．

平成２３年度以降の本事業（仮称）に期待する点として，専門用語辞書を自発的に追加発展させていくことができるフォーラムの構築が考えられる．

７．全体総括

専門分野の文書検索の高精度化のための専門用語解析，抽出を目標とし，基盤となる技術とツールの開発を行った．これらの解析技術とツールの機能については，当初予定をほぼ達成することができた．個々のツールや解析手法については，既存のツールや解析手法では実現できなかった機能や性能を達成することができた．

ただし，実際の文書検索への利用に供するところまで至らなかったため，文書検索の精度向上への貢献については評価することができなかった．専門文書検索時に本研究での成果による類義語や内部の部分構造情報を利用することで，どの程度検索性能の向上が得られるかを実利用の立場から検証することが考えられる．

８．特記事項

本開発では，日本語の用語と専門文書を対象としたが，専門用語辞書システムは言語に依存しないシステムとして開発した．また，英語の対訳情報を持つこともできる．用語の内部構造のアノテーションと内部構造を指定した検索機能は，同様の機能をもつシステムは我々が知る限り他には存在しない．

専門用語の内部構造解析については，構成要素となる語の共通部分が縮退（「角膜」＋「結膜」が「角結膜」になる，また，「大腿骨」＋「骨折」が「大腿骨折」になるなどの現象）を考慮した新しい手法を提案した．

類似用語の検索については，用語が出現する文の係り受け解析を行い，係り受け木構造内のウォークの類似性を用いる新しい手法を示し，従来手法を上回る精度を達成した．

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）		6,118	877	0	0	6,995
人　件　費（千円）		4,367	7,678	8,785	8,745	29,575
業務実施費（千円）		3,515	2,118	1,888	800	8,321
一般管理費（千円）		1,400	1,067	1,067	955	4,489
合　計　（千円）		15,400	11,740	11,740	10,500	49,380

整備実績一覧

（1）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	ライフサイエンス辞書検索システム Cradle-LSD	公開	新規用語の登録，同義語情報を記述，種々の検索や内部構造のアノテーション機能をもつ専門用語辞書システム．別紙参考図の図１～図３に使用例を示す．

外部発表実績一覧

（1）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	バイオ医療専門用語の類義語獲得	統合データベースプロジェクトシンポジウム２００９	2009 年 6 月 12 日	
2	専門用語辞書管理システムおよび用語の内部構造アノテーションについて	統合データベースプロジェクトシンポジウム２００９	2009 年 6 月 12 日	
3	コーパスに出現するライフサイエンス分野の専門用語の同定とシソーラスへのマッピング	統合データベースプロジェクトシンポジウム２０１０	2010 年 10 月 5 日	

（2）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	GENIA コーパスからのネスト並列句同定	原一夫，新保仁，大熊秀治，松本裕治	情報処理学会自然言語処理研究会	2008 年 9 月 24-25 日	
2	グラフを用いたバイオ医療専門用語の類義語推定	鈴木郁美，原一夫，新保仁，松本裕治	情報処理学会自然言語処理研究会	2009 年 1 月 22-23 日	
3	Coordinate Structure Analysis with Global Structural Constraints and Alignment-Based Local Features	Kazuo Hara, Masashi Shimbo, Hideharu Okuma, Yuji Matsumoto	Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint	2009 年 8 月 4 日	

			Conference on Natural Language Processing		
4	A Graph-based Approach for Biomedical Thesaurus Expansion	Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, Yuji Matsumoto	ACM Third International Workshop on Data and Text Mining in Bioinformatics	2009年11月6日	
5	Towards automatic biomedical entity annotation by reducing error propagation	Kazuo Hara	The First CALBC (Collaborative Annotation of a Large Biomedical Corpus) Workshop	2010年6月18日	

（3）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	バイパス付きアラインメントグラフを用いた日本語並列句検出と範囲同定	大熊秀治，原一夫，新保仁，松本裕治	人工知能学会論文誌	25 巻、1 号、pp.206-214、2010年1月	
2	文法制約と系列アラインメントによる並列構造の解析	原一夫，新保仁，松本裕治	人工知能学会論文誌	25 巻、5 号、pp.560-569、2010年7月	

図1：専門用語辞書システムの実行画面

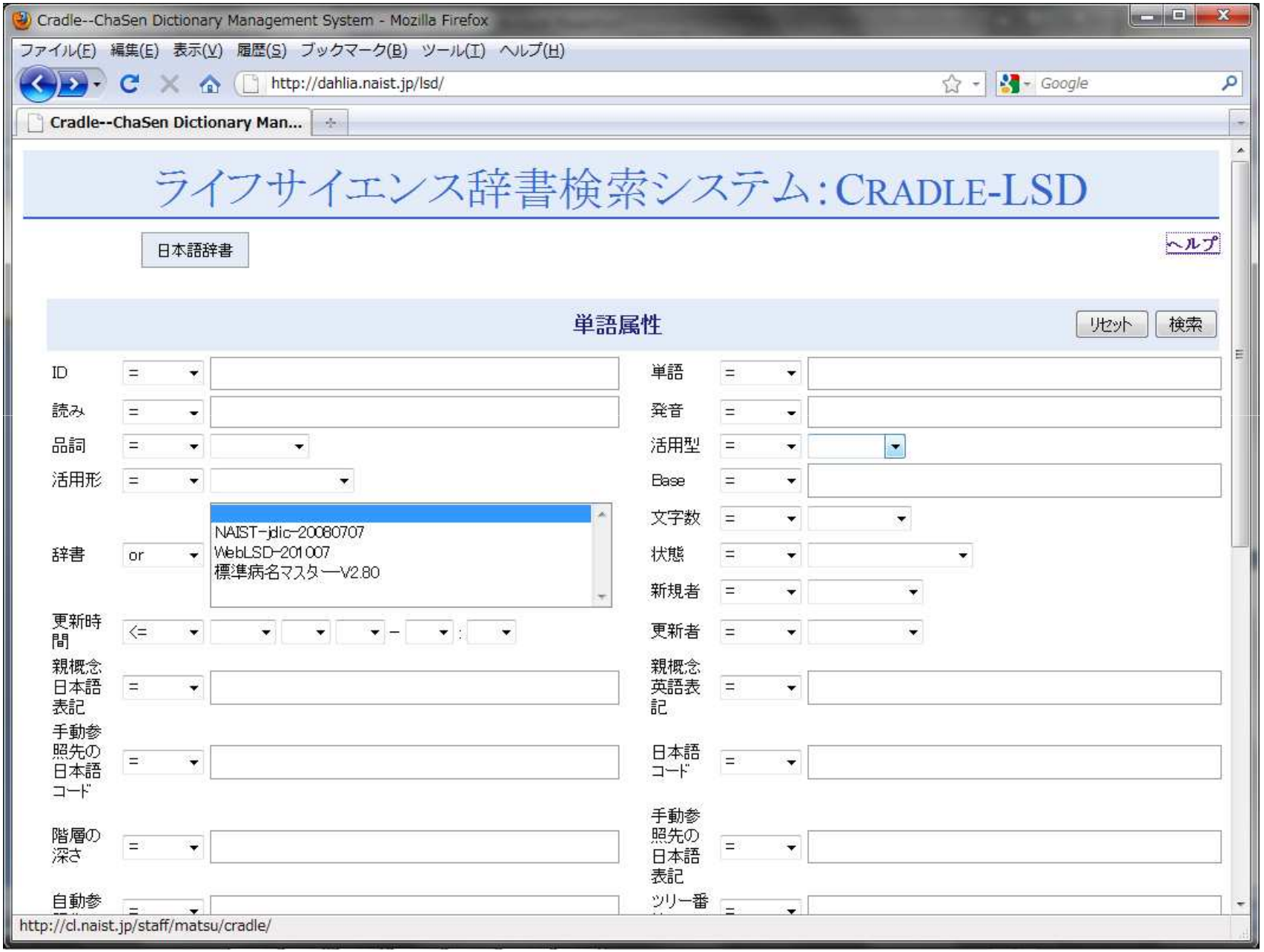


図2：用語の検索結果の例

図3:用語の詳細情報の表示

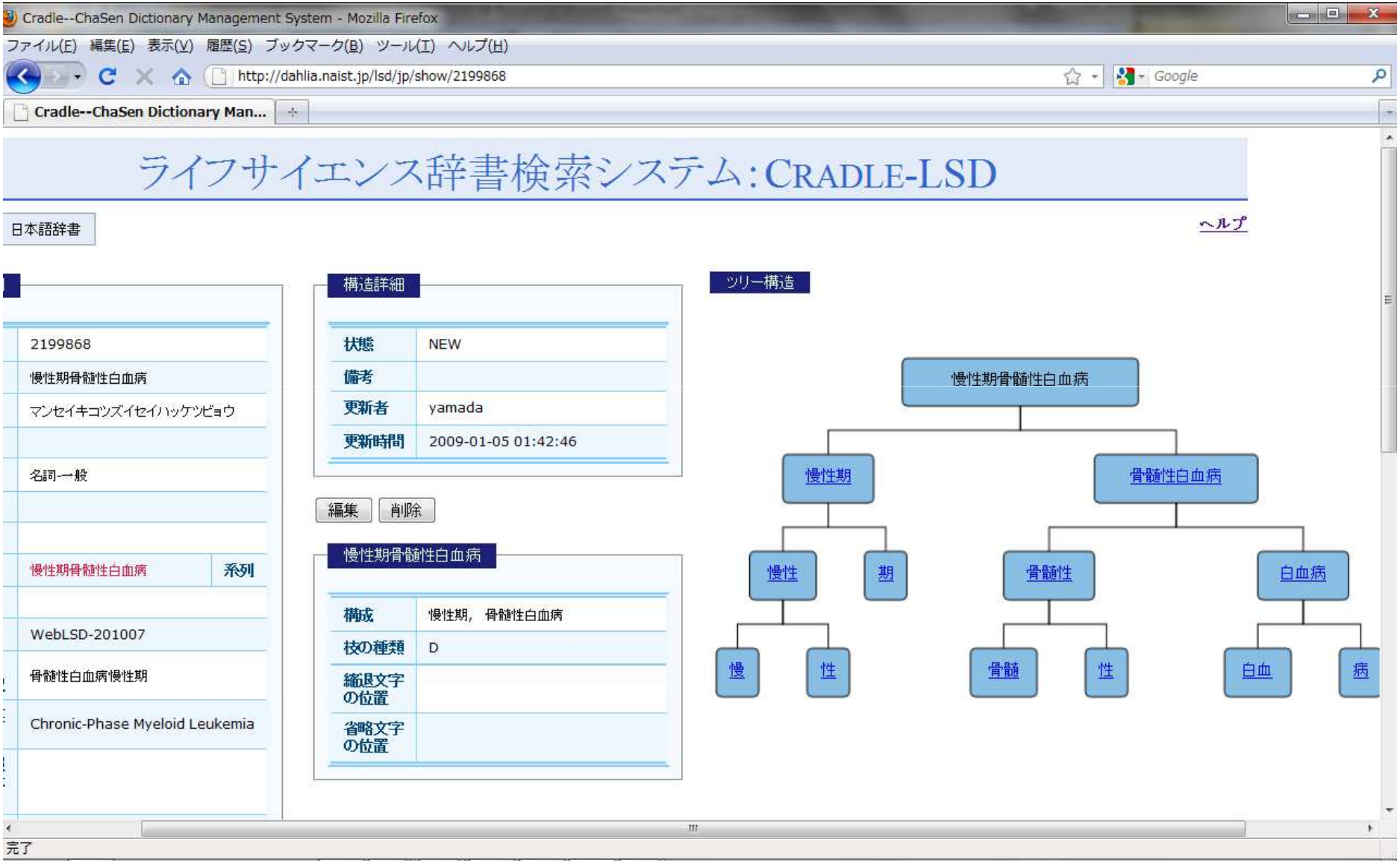


図4：文書中の用語検索とシソーラス表示

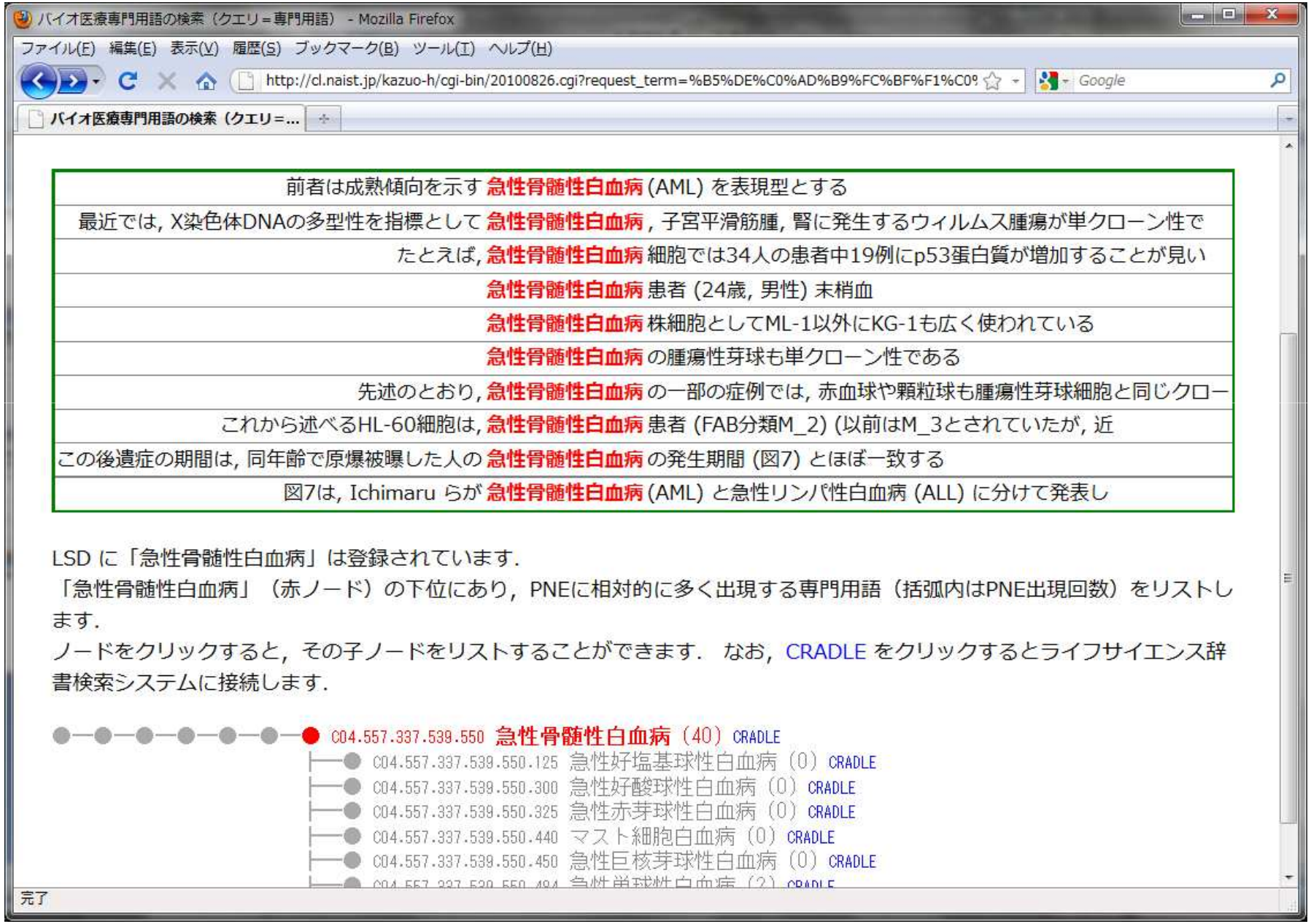


図5：類似用語の検索とシソーラス表示

トップページ　バイオ医療専門用語の検索　類似文脈の検索

類似文脈の検索（クエリ＝文脈＋専門用語）

文脈: 著者らはこのような4症例の患者末梢血リンパ球を起因薬物及びキャリアー蛋白で刺激する際，インターロイキン-2を添加することによつて，リンパ球幼若化反応が増幅されることを観察した．

クエリ: インターロイキン　Submit query　(類似度を約15秒で計算します)

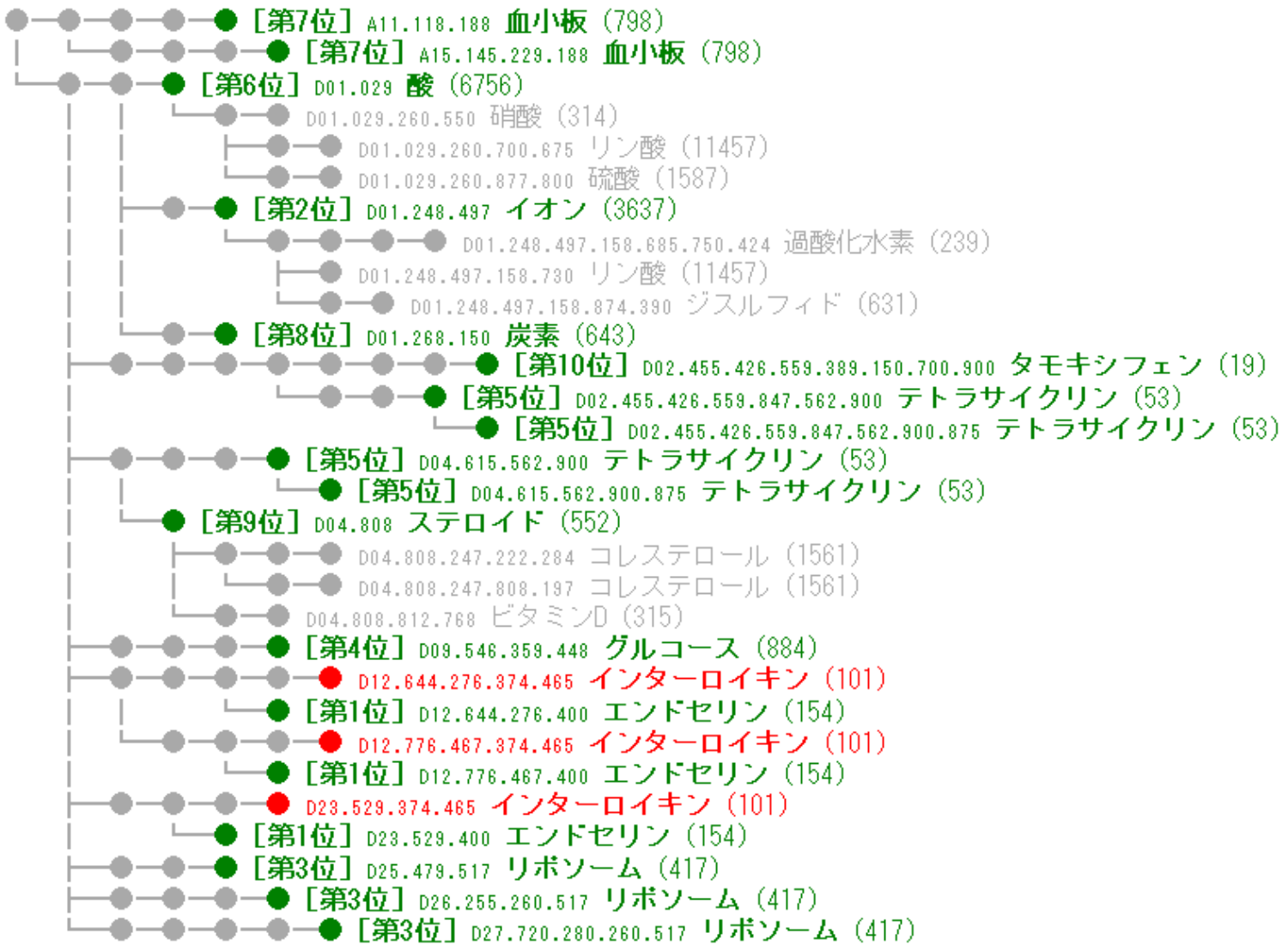
検索結果：

「著者らはこのような4症例の患者末梢血リンパ球を起因薬物及びキャリアー蛋白で刺激する際,インターロイキン-2を添加することによつて,リンパ球幼若化反応が増幅されることを観察した.」
に類似する文脈を，PNEから検索した結果（コサイン類似度による10位までのランキング）を示します．

1	そして外因性の	エンドセリン	-1を添加すると,細胞が収縮した	0.286
2	実際,ATP,Mgイオン,K^+	イオン	を複合体の溶液に添加したところ,すぐさま沈殿が生じ,	0.273
3	PE	リボソーム	を添加した細胞を経時的に観察すると,15分後からブリ	0.266
4		グルコース	を添加することによって,欠損株の表現型は抑制される	0.265
5	しかし,	テトラサイクリン	を添加することで,TetRがTetO配列に結合するこ	0.263
6	そして,この胚様体形成時にレチノイン	酸	を添加することによって神経系の分化がさらに促進される	0.263
7		血小板	にスフィンゴシンキナーゼの阻害剤を添加すると,S1P	0.259
8	ゲナーゼは哺乳動物では5位,12位あるいは15位の	炭素	原子に1分子の酸素を添加するものが知られており,それ	0.255
9	しかし,ステロイドホルモン応答細胞に	ステロイド	ホルモンを添加したり,幼若ラットにテストステロン,エ	0.255
10	この際,	タモキシフェン	を添加しておくとエストロゲン作用は打ち消され,無添加	0.252

図6：類似する専門用語のLSDシソーラスにおける位置

上表に示した類似文脈に関わる専門用語（緑ノード）と「インターロイキン」（赤ノード）に対して，LSDにおいて下位にあり，PNEに相対的に多く出現する専門用語をリストします．



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	多型知識表現技術開発
実 施 機 関 名	九州大学
代 表 研 究 者 名	林　健志

１．課題開始時における達成目標

平成18年度：日本人ゲノム多型情報を高度化し、医療情報との統合の一環として、我々が既に開発している2個の日本人ゲノム多型情報データベースのデータポータビリティーを改善する。
平成19-22年度：我国で公的資金によって行われている疾患のゲノム関連解析（GWAS）研究の透明性確保を意図して、同研究が産出するデータを第三者機関として標準的かつ一貫した方法で品質チェックを行い、その結果を公開する。

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

1）遺伝子発現調節領域多型の公開データベース「dbQSNP」及び全ゲノム確定ハプロタイプ構造の公開データベース「D-HaploDB」にある主要なデータをXML化して公開した。
2）GWAS研究で生産される大規模大量なジェノタイプデータの品質チェックを国際水準に則った一貫した方法で行うための解析パイプラインであるJAGQC Pipelineを開発した。
3）上記JAGQC Pipelineを用いて、文部科学省及び厚生労働省によってサポートされているGWAS研究４件の品質チェック解析を行い、その結果をJAGQC DB（日本語版及び英語版）として公開した。

（２）進捗及び成果

平成 18 年度には、上記「dbQSNP」及び「D-HaploDB」にある主要なデータを PML（ゲノム多型の記述に特化された XML 言語）化し、それぞれ http://qsnp.gen.kyushu-u.ac.jp/ 及び http://orca.gen.kyushu-u.ac.jp/で公開した。

平成 19-22 年度には、品質チェック（Q.C.）解析のための新たな JAGQC Pipeline を開発した。この pipeline は、取得した SNP ジェノタイプデータの形式を先ず PLINK 形式に変換し、常染色体上の SNP について、サンプルのコール率チェック、SNP のコール率チェック、失敗率偏りテスト、Hardy-Weinberg 平衡テスト、血縁関係テスト、集団階層化テストの６個の解析を行うものである。これらの解析には国際的な標準解析プログラムである PLINK 及び EIGENSOFT/smartpca を用いた。また、一部の統計解析、及び結果のグラフ表示には R を用いた。次に、文部科学省が支援する４課題、及び厚生労働省が支援する３件の GWAS 研究について各課題の公式公開データベースに対してジェノタイプデータを請求し、そのうち４件についてデータ提供を受けた。これら４件について JAGQC Pipeline による品質チェック（Q.C.）解析を行い、その結果を JAGQC DB として公開した。JAGQC DB の具体例（日本語版）を添付参考資料に示す。

３．当初目標に対する達成度

関係する各データベースの利用実績等は以下のとおりである。

dbQSNP アクセス: 5万件／年 ダウンロード: 1600回／年
dHaplo アクセス: 2万件／年 ダウンロード: 160回／年
JAGQC DB アクセス: 76／月 （平成２２年１０月開始）

４．中間評価に対する対応

特になし。

５．他機関との連携

九州大学は、分担機関である東京大学・徳永グループ（文部科学省による GWAS 研究の公式データベースである「GWAS DB」の研究代表）及び国立がん研究センター・吉田グループ （厚生労働省による GWAS 研究の公式データベースである「GEMDBJ」の研究代表）にデータ請求を行い、前者からは一部の、後者からは全ての請求項目についてデータ提供を受けて、これらの品質チェックとその結果の公開を行うことが出来た。従って他機関との連携は完全ではないが、満足すべきレベルであったと考えている。今後、未提供のデータに関しても順次提供を受けることを期待している。

６．今後の見通し、計画、展望

今回我々が行った GWAS 研究データの品質チェック法は必要最低限のものであり、短期的、及び長期的により拡充すべきものと考えている。具体的には、より発生源に近い段階でのデータ取得とその品質チェック、全ゲノムリシークエンシングデータを含む GWAS 研究でのデータ品質チェックシステムの確立である。このために、拡大された規模での品質チェックプロジェクトとして継続されるべきであると考える。

７．全体総括

統合データベースは必然的に極めて広いスペクトラムの利用者を対象としている。しかし、この中の各活動、各データベースはそれぞれの利用者層を明確にして、これに対して検証可能な形体で情報を提供する必要がある。今回我々が公開したデータベースは GWAS 研究当事者、またはそれに密接に関与した分野の研究者を対象としたものであることが明記されている。また本データベースの情報原、得られたデータをどのように処理したか、その結果がどうであったかを利用者が検証可能な形態で開示した。これは統合データベースでの健全な活動であったと考えている。

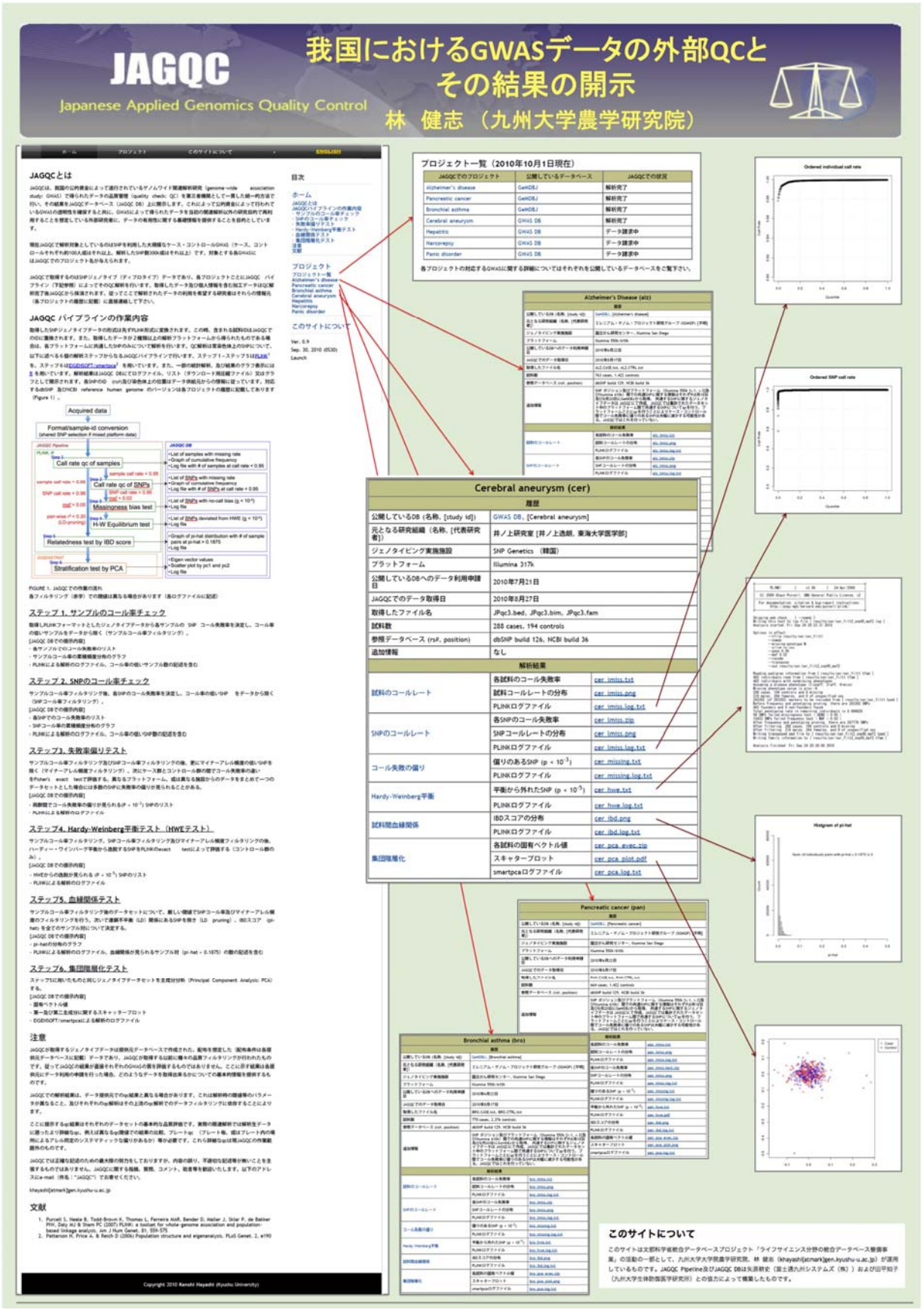
８．特記事項

我国の GWAS 研究を含むゲノム関連研究では、データ生産当事者と、研究成果の公開の場としての GWAS データベースが同一機関で行われていることが多い。このような状況ではデータを第三者が客観的立場で検証し、これを公開することによってはじめて我国の GWAS 研究の信頼性を主張できる。

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）	600	0	0	0	0	600
人　件　費（千円）	0	0	3,992	3,992	3,983	11,967
業務実施費（千円）	13,686	5,238	1,246	1,008	517	21,695
一般管理費（千円）	714	262	262	500	450	2,188
合　計　（千円）	15,000	5,500	5,500	5,500	4,950	36,450

別紙参考資料


JAGQC
Japanese Applied Genomics Quality Control
我国におけるGWASデータの外部QCとその結果の開示
林　健志（九州大学農学研究院）
JAGQCとは
JAGQCパイプラインの作業内容
プロジェクト一覧（2010年10月1日現在）
Cerebral aneurysm (cer)
Alzheimer's Disease (alz)
Pancreatic cancer (pan)
Bronchial asthma (bro)
このサイトについて
Copyright 2010 Kenshi Hayashi (Kyushu University)

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）
	JAGQC DB（http://crane.gen.kyushu-u.ac.jp/JAGQC/JAGQC_top.html）	公開	我国で公的資金によって行われている疾患のゲノムワイド関連解析研究のジェノタイプデータ（対象７件、解析結果公開件数４）の品質チェックを 6 項目について行い、その解析経過のログファイル及び結果を図、表及びダウンロード可能なファイルとして公開している。 平成 22 年１０月開始、アクセス数：76／月

（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）
	JAGQC pipeline	未公開	取得した SNP ジェノタイプデータの形式を PLINK 形式に変換し、常染色体上の SNP について、サンプルのコール率チェック、SNP のコール率チェック、失敗率偏りテスト、Hardy-Weinberg 平衡テスト、血縁関係テスト、集団階層化テストの６個の解析を一貫して行う解析パイプラインである。これらの解析には国際的な標準解析プログラムである PLINK 及び EIGENSOFT/smartpca を用いた。また、一部の統計解析、及び結果のグラフ表示には R を用いた。

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	ヒト遺伝子研究と統合データベース	発表：今西 規　他2名 主催：林　健志	2009年3月17日	九州大学コラボステーションII	講演会及び講習会	ゲノム科学関連の研究者、約50名を対象としたデータベース利用のPC持ち込みで実習を含む講習会。この種の講習会は当地区では始めてであり好評であった。

（2）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	DHaploDB	林　健志	第5回国際バイオデータ相互運用性会議・国際シンポジウム「遺伝多型解析研究の最先端」	2007年9月26日	日本人確定ハプロタイプデータベースDHaploDBに関する講演
2	我国における GWAS データの外部QCとその結果の開示	林　健志	シンポジウム「ライフサイエンスの未来へ」	2010年10月5日	JAGQC DBに関するポスター発表

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	統合データベース支援: DB 構築者の養成
実　施　機　関　名	国立大学法人東京大学
代　表　研　究　者　名	森下真一

１．課題開始時における達成目標

DB 構築者を養成するために以下の３つの演習を実施する。

①バイオ DB サーバー構築演習

データベースサーバーのミラーサイトを構築する。OS, apache, MySQL 等の主要ソフトウエアのインストールおよびネットワークセキュリティに習熟することが目標である。参加者には各自にサーバー構築用ワークステーションを配布する。演習を完了するまでには、受講者の能力と受講可能時間に応じて最短で３ヶ月、最長で１年間の時間を予定している。

②プログラミング演習

Java および Perl プログラミングを演習した後に、アルゴリズムの知識を活かした配列処理やデータマイニングの実装を行う。上記①バイオ DB サーバー構築演習では実施がむずかしいプログラミング演習を行うことで、独自にソフトウエア構築ができる能力を身につけることをめざす。演習総時間は９０時間で約２ヶ月間を予定している。

③　独創的サーバー構築演習

大規模計算のためのクラスター利用技術を習得させ、他に類の無いバイオ DB サーバーを設計、実装、公開することを目標とする。バイオ DB サーバー構築演習およびプログラミング演習を修了した受講者に対して平成 20 年度より開講を予定しており、そのための計算機セットアップを平成 19 年度に準備する。演習は英語で書かれたプログラミングの教科書および DB サーバーシステムの管理マニュアルの読破も要求し、自ら立ち上げたサーバーの英語マニュアルが書けるように指導する。したがって受講者にはある程度の英語能力を想定する。平成 20 年度で作成し公開中の演習ノートも改善する。

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

DB 構築者を養成するために３つの演習を実施し、講義資料・演習ノートのウェブページを作成した。

①┌　バイオ DB サーバー構築演習

データベースサーバーのミラーサイトを構築する演習を実施した。OS, apache, MySQL 等の主要ソフトウエアのインストールおよびネットワークセキュリティに習熟することが目標とし、参加者には各自にサーバー構築用ワークステーションを配布した。演習は平成１９年・２０年・２１年に実施し、演習を完了するまでには、計画通り１年間の時間を要した。演習ノートの web ページ(下記)を作成した。

http://mlab.cb.k.u-tokyo.ac.jp/~nakatani/ensemblmirror/

受講者数は平成１９年度５名、平成２０年度１１名、平成２１年度４名だった。

②　プログラミング演習

Java および Perl プログラミングを演習した後に、アルゴリズムの知識を活かした配列処理やデータマイ

ニングの実装を行う演習を実施した。上記①バイオ DB サーバー構築演習では実施がむずかしいプログラミング演習を行うことで、独自にソフトウエア構築ができる能力を身につけさせる効果があった。演習総時間は９０時間で平成１９年・平成２０年７月下旬～９月上旬の約２ヶ月間ずつであった。演習ノートの web ページ(下記)を作成した。

http://mlab.cb.k.u-tokyo.ac.jp/~mkasa/upbsb2007/

http://mlab.cb.k.u-tokyo.ac.jp/~mkasa/upbsb2008/

http://www.xerial.org/wiki/lecture/2008/Java/Index

③　独創的サーバー構築演習

バイオ DB サーバー構築演習およびプログラミング演習を修了した受講者に対して独創的サーバー構築演習を実施した。ウェブアプリケーション開発・Sun Grid Engine による並列計算・次世代シーケンサーのデータ解析等について演習し、演習ノートの web ページを作成した。

http://mlab.cb.k.u-tokyo.ac.jp/~nakatani/ensemblmirror/

また、ゲノムブラウザーで表示される新規トラックの開発・新規データの表示・塩基配列をゲノムにマッピングし結果をウェブ上で表示するプログラムの実装などを行った。

・塩基配列をメダカゲノムにマッピングし結果をウェブ上で表示するプログラムの開発。

http://medaka.utgenome.org/

・cDNA アセンブルプログラムの開発。

http://musica.gi.k.u-tokyo.ac.jp/

・ゲノム中の短い配列の出現頻度データベース RepeatScape（下記）を公開した。

http://repeatscape.utgenome.org

受講者数は平成２０年度５名、平成２１年度８名、平成２２年度４名だった。

（２）進捗及び成果

資料添付。

３．当初目標に対する達成度

DB 構築者を養成するために３つの演習を実施し、講義資料・演習ノートのウェブページを作成した。サーバー構築用ワークステーションを使用した演習を修了した受講者は、バイオ DB サーバー構築演習で２０名、独創的サーバー構築演習で１７名であった。

演習ノートのウェブページに対する平成２２年度４月～１０月のアクセス統計は以下の通りだった。（**Sites**：いくつの IP アドレスからリクエストを受けたかを示す。**KBytes**：送信したデータ量。**Visits**：サイトを訪問したユーザー数。**Pages**：送信したページ数。画像などのファイルは含まない。**Files**：送信したファイル数。）

Month	Daily Avg			Monthly Totals				
	Files	Pages	Visits	Sites	KBytes	Visits	Pages	Files
Oct	347	321	250	300	5829	3764	4820	5217
Sep	357	337	246	630	16046	7380	10125	10730
Aug	1096	1073	315	978	406870	9778	33287	34006

Jul	1782	1891	486	1906	779913	15087	58623	55244
Jun	580	1108	420	1452	316922	12628	33258	17428
May	1515	1575	483	2230	689661	15002	48825	46988
Apr	1362	1354	344	1636	614311	10320	40644	40863

（10 月のアクセス統計は 10 月 15 日までのアクセスログから作成。）

また、受講者の進路（卒業生の就職先、または現在の所属）は次の通りであった。

DBCLS：２名

研究員：２名

博士課程在学中：５名

修士課程在学中：３名

就職：６名（IBM, JR 東日本, 日立, 野村総研など）

不明、その他：３名

４．中間評価に対する対応

ウェブ上での自習用教材として活用するため、演習ノートの追加・整理を行った。

５．他機関との連携

受講者のうち２名は DBCLS に所属しており、また１名は長浜バイオ大で実施された「アノテータ・キュレーターの教育」を受講し修士課程から東京大学情報生命科学専攻に進学した学生であった。

６．今後の見通し、計画、展望

今後の展開として講義資料・演習ノートのウェブページを自習用教材として活用することが考えられる。残された課題は、DB サーバー構築演習受講者がライフサイエンス分野で就職し、習得した技術を活用できるよう、研究環境を整備することが挙げられる。実際には DB サーバー構築演習の受講者の約半数は卒業後就職しており、養成された人材がライフサイエンス統合データベース分野で十分に活かされているとは言えない状況である。

７．全体総括

当初の目標通り、DB 構築者を養成するために演習を実施し、講義資料・演習ノートのウェブページを作成した。DB サーバー構築演習の受講者が将来的に統合データベース整備事業に貢献することが望まれるが、実際には受講者の約半数は企業に就職している。今後は DB 構築スキルを身に付けた技術者がライフサイエンス分野で継続的に活躍できる環境を整備することが重要である。

８．特記事項

サーバー管理・プログラミングから次世代シーケンサーによる大量配列データの解析まで、最近の研究動向を踏まえて講義・演習を実施した。また、少人数での演習としたことで、各受講者にサーバー構築用ワークステーションを配布し、受講者のレベルに合わせた指導を行うことができた。

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）		12,256	0	0	0	12,256
人　件　費（千円）		3,942	9,965	9,883	9,851	33,641
業務実施費（千円）		2,802	1,435	1,517	1,058	6,812
一般管理費（千円）		1,900	1,140	1,140	1,091	5,271
合　計　（千円）		20,900	12,540	12,540	12,000	57,980

整備実績一覧

（1）その他の成果物

通番	名称	公開／未公開	概要
1	演習ノート	公開	講義資料・演習ノートを整理して公開した。 http://mlab.cb.k.u-tokyo.ac.jp/~nakatani/ensemblmirror/

統合データベース支援：
バイオDBサーバー構築演習
補足資料

目的

- バイオDBを構築できる人材を育てる
  - 膨大なソフト外注費（150～200万円/月）を回避
  - DBの保守・拡張が自前でできること
  - やむをえず外注する場合も、正確な仕様書を書ける力と、納入されたソフトの問題点を見抜く力を養う
- 必要スキルを1年間のカリキュラムで教え込む
- 次の1年で独創的サーバーを構築

計画

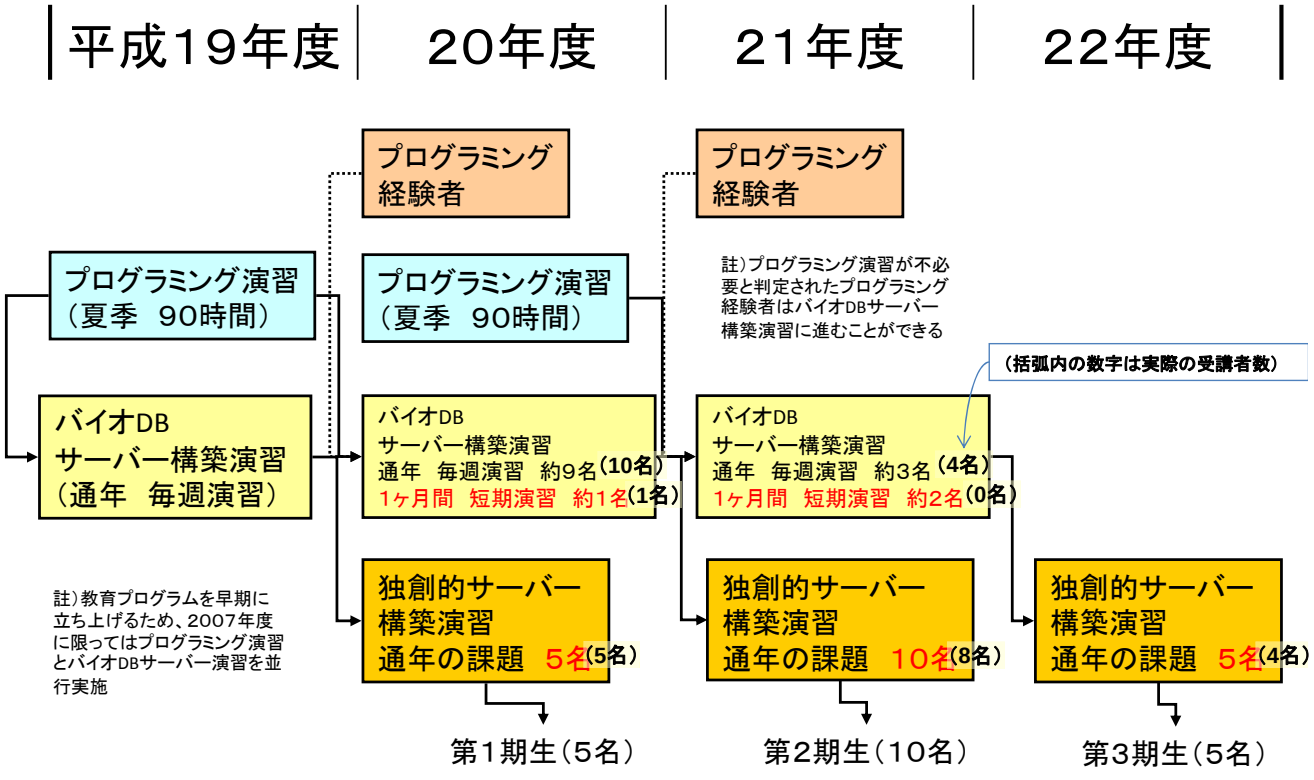
DB 構築者を養成するために以下の3つの演習を実施する。

①バイオ DB サーバー構築演習
データベースサーバーのミラーサイトを構築する。OS, apache, MySQL 等の主要ソフトウエアのインストールおよびネットワークセキュリティに習熟することが目標である。参加者には各自にサーバー構築用ワークステーションを配布する。演習を完了するまでには、受講者の能力と受講可能時間に応じて最短で3ヶ月、最長で1年間の時間を予定している。

②プログラミング演習
Java および Perl プログラミングを演習した後に、アルゴリズムの知識を活かした配列処理やデータマイニングの実装を行う。上記①バイオ DB サーバー構築演習では実施がむずかしいプログラミング演習を行うことで、独自にソフトウエア構築ができる能力を身につけることをめざす。演習総時間は90時間で約2ヶ月間を予定している。

③ 独創的サーバー構築演習
大規模計算のためのクラスター利用技術を習得させ、他に類の無いバイオDBサーバーを設計、実装、公開することを目標とする。バイオDBサーバー構築演習およびプログラミング演習を修了した受講者に対して平成20年度より開講を予定しており、そのための計算機セットアップを平成19年度に準備した。

年次計画



演習用WS15台
（平成19年度予算申請）

註） 1期生と2期生が20年度には重なること（21年度は2，3期生）、WSが15台であること、演習スタッフ1．5名による徒弟制度であるため、各年15名の受け入れが限度である

受講者数

バイオDBサーバー構築演習	独創的サーバー構築演習
• **平成19年度受講者**	**平成20年度受講者**
– 東大情報生命科学専攻から5名	– 東大情報生命科学専攻から5名
• **平成20年度受講者**	**平成21年度受講者**
– 東大情報生命科学専攻から8名 – DBCLSから2名 – 自治医大から1名	– 東大情報生命科学専攻から6名 – DBCLSから2名
• **平成21年度受講者**	**平成22年度受講者**
– 東大情報生命科学専攻から3名 – 東大医学系研究科から1名	– 東大情報生命科学専攻から4名

DBサーバー構築演習の目標設定

1: CentOS を自分のマシンにインストールする
2: ネットワークと接続する
3: セキュリティアップデートを行う
4: Web サーバーを立てる(ファイヤーウォールの設定を行う)
5: CGIを設置する
6: MySQL サーバーを立てる
7: 簡単なデータベースを作成する
8: Ensembl core をインストールしミラーを作成する
9: 複数種の実データをダウンロードして完全ミラーを作る
10: バックアップを作成して即時復旧できる体制を作る

バイオDBサーバー構築演習の概要（平成21年度の例）

- OS (Linux) のインストール
- ネットワーク・ファイアーウォールの設定
- Web サーバーの設置・設定 (apache)
- RDBMSの設置・設定 (MySQL)
- Perl モジュールの設置・設定
- Ensembl の設置・設定
- Perl, Javaプログラミング
- CGIからのデータベース検索
- メンテナンス全般
  - 障害対応
  - ソフトウェアの Security fix やバージョンアップ等

演習日程	テーマ
4/16	イントロダクション 、CentOSのインストール
4/23	セキュリティと定期アップデート、SSHによる外部からの安全な接続
4/30	Webサーバーの設置、シェルスクリプト、Pukiwikiの設置
5/07	Perl演習
5/14	CPANを使いこなす、BioPerlのインストール
5/21	RDBMS、Perlからデータベースを扱う
5/28	PerlによるCGIプログラミング
6/04	Java演習：プログラムの書き方
6/11	Java演習：データ構造とオブジェクト
6/25	Java演習：GUIアプリケーションとデータの入出力
7/02	Java演習：データベースアプリケーション
7/09	CGIでデータベースを検索する
7/16	Ensemblデータベースをミラーする1
7/30	Ensemblデータベースをミラーする2
9/03	Ensemblデータベースをミラーする3
9/17	Ensemblデータベースをミラーする4
10/01	Ensemblデータベースをミラーする5
10/15	サーバーのバックアップ1
10/15	BLATを用いたmRNAのゲノムへのマッピング
10/29	サーバーのバックアップ2
10/29	Ensemblデータの解析、BioMartを使ったデータ取得
11/12	サーバーのバックアップ3
11/26	サーバーのバックアップ4
11/26	OpenCVを使った画像処理演習
12/10	UTGB Toolkitのインストール
12/10	JFreeChartを使用したグラフの描画
12/10	遺伝子発現データベースを使い倒す
12/24	UTGB Toolkitを使ったゲノムブラウザプログラミング
12/24	遺伝子発現データの生物学的な解釈

OSのインストール、セキュリティ設定、ネットワーク設定

- 演習受講者にサーバーを1台ずつ割り当て、OSインストールから演習する。
- OSインストール
  - システム・ネットワーク・ウェブ・データベース等に関する基礎的な用語の解説。
  - 各自のサーバーにLinuxをインストール。
    - CDイメージをダウンロードしCentOS最新版をインストールする。
- セキュリティ設定
  - 脆弱性について基礎的な事項を学習（buffer overflow等）。
  - Yum-cronによる定期的なセキュリティアップデート。
- ネットワーク設定
  - 公開鍵認証方式によるssh接続の設定。
  - ファイアウォールの設定。

Webサーバーの設置、Pukiwikiの設置、シェルスクリプトプログラミング

- ウェブサーバーの設置。
  - Apacheのインストール。
  - 設定ファイルの編集。
  - Firewallの設定。
- Pukiwikiの設置し、ウェブ上で情報の共有と整理を多人数で行う。また演習ノートをPukiwikiで作成する。
  - Pukiwikiをダウンロードし、サーバーにインストールする。
  - Pukiwikiの基本操作、文法の解説。
- シェルスクリプトプログラミング
  - 文法の解説。
  - シェルスクリプトを使ったCGIプログラムの作成。

Perlプログラミング、CGIプログラミング

- EnsemblのコードはPerlで書かれておりミラーサイト構築に必要となるため、Perlは時間をかけて詳しく解説した。
  - Perlのインストール。
  - 基本的なPerl文法の解説。
    - File I/O, 正規表現, サブルーチン, ソート等。
  - ゲノム配列データをダウンロードし、Perlを使用して簡単なデータ処理を行う。
  - CPAN (Comprehensive Perl Archive Network)によるソフトウェア・モジュールのインストールを解説。
    - BioPerlのインストール。
    - Makeによるインストールも解説。
- CGIプログラミング
  - HTTP, CGIの解説。
  - Perlでアクセスカウンターを作成。
  - GET方式とPOST方式によるユーザーからの入力の処理。
  - Cookieの解説。

データベースの設置、Perlを使ったデータベース検索

- データベースの設置。
  - MySQLのインストール。
  - MySQLの基本的なコマンドの解説。
  - 遺伝子データ（TreeFam, Ensemblのデータを使用）をサーバーにダウンロードしMySQLで検索する。
- Perlを使ったデータベース検索CGIの作成。（ウェブページからユーザー入力を受け取りデータベースを検索するCGIの作成。）
  - PerlのDBIモジュールを使ってデータベースにアクセスし遺伝子データを検索。
  - BioPerlを使用した遺伝子系統樹解析。
  - CGI作成用のPerlモジュール HTML::Template, HTML::FillInformを使用し、データベースを検索し検索結果を画像で表示するCGIを作成。

Java プログラミング演習

- プログラムの書き方。
  - Javaの仕組みと文法、Eclipse（Javaの開発環境）の使い方
- データ構造とオブジェクト。
  - 配列、オブジェクト指向プログラミング、データ構造。
- GUIアプリケーションとデータの入出力。
  - 文字列、オブジェクト・クラス、入出力、GUIアプリケーションの作成。
- データベースアプリケーション。
  - リレーショナルデータベースとSQL, SQLite JDBCを使ってJavaからデータベースを扱う。

Ensemblミラーサイトの構築
サーバーのバックアップ

- Ensembl ミラーサイト構築。
  - yum, CPANを使って必要なモジュール・外部プログラムをインストール。
  - Ensemblデータのダウンロードとインストール。
  - Ensemblウェブサイトの設定、起動。
- TeraStationへのバックアップ。
  - TeraStationをサーバーにマウントする。
  - “mysqldump”によるデータベースのバックアップとデータ復旧。
  - ”rsync”コマンドによるバックアップとデータ復旧。

バイオDBサーバー構築演習の進捗

平成19年度	平成20年度	平成21年度	
済	済	済	1. CentOS を自分のマシンにインストールする
済	済	済	2. ネットワークと接続する
済	済	済	3. セキュリティアップデートを行う
済	済	済	4. web サーバーを立てる
済	済	済	5. CGIを設置する
済	済	済	6. MySQL サーバーを立てる
済	済	済	7. 簡単なデータベース作成をする
済	済	済	8. Ensembl core をインストールしミラーを作成
			9. Ensemblの完全ミラーを作る
		済	10. バックアップを作成する

独創的サーバー構築演習

- 大規模計算のためのクラスター利用技術を習得させ、他に類の無いバイオDBサーバーを設計、実装、公開することを目標とする。
  - バイオデータの解析演習。
  - 大規模並列計算演習。
    - Sun Grid Engineを用いた並列計算。
  - ウェブアプリケーション開発演習。
    - ゲノムブラウザーによる解析データの視覚化。
  - 受講者の研究結果をデータベースとして実装・公開することが目標。

バイオデータの解析演習

受講者が実際の研究で使用するツール・データの解説と簡単な演習を行った。

- BLATを用いたmRNAのゲノムへのアラインメントと出力結果の処理。
- Ensemblの比較ゲノムデータを使った解析とBioMartによるデータ取得方法。
- OpenCV（画像処理ライブラリー）を用いた画像解析。
- JFreeChartを使用したグラフの描画。
- 遺伝子データベースを使い倒す（BioGPS, NCBI Gene Expression Omnibus, Mouse Genome Informatics データベースを使った発現解析）。
- 遺伝子発現データの生物学的な解釈（DAVID, Reactomeデータベースを使った発現解析）。

Sun Grid Engineを用いた並列計算

次世代シーケンサーの大量データを解析するために、SGEを使った並列計算を習得することが目標。

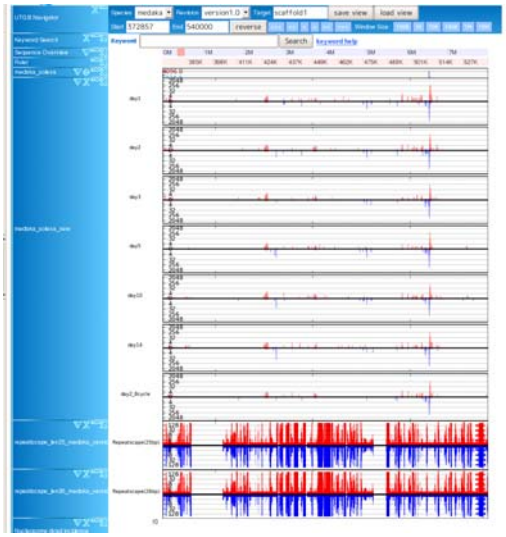
- SGEの使用法を解説
  - 並列ジョブの投入。
  - 実行/エラー状態の確認。
  - ジョブのキャンセル。
  - ジョブ間の依存関係の指定。
  - 必要なリソースを指定して実行。
  - キューの作成、エラー処理。
- SGEを使って、次世代シーケンサーで読んだリード配列をゲノムにマッピング。

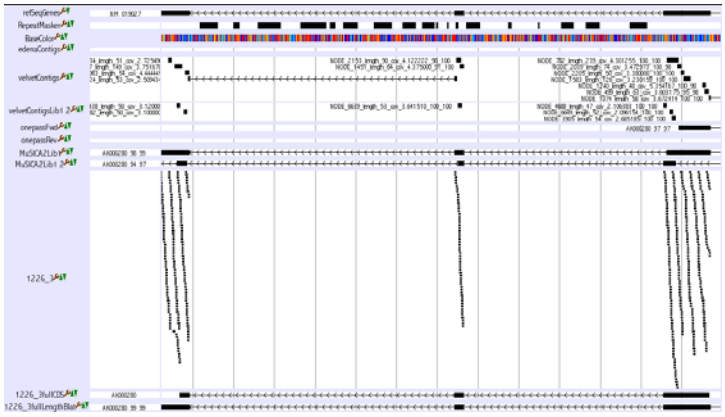
ウェブアプリケーション開発演習

- UTGB toolkitを用いて新しいタイプのゲノムデータをトラックに表示する技術を習得することが目標。
  - GWT(Google Web Toolkit)によるwebアプリケーション開発。
  - ユーザーインターフェイスの作成。
  - AJAXによるデータ通信。
  - JDBCによるデータベース検索。
  - UTGBの紹介、UTGB toolkitのインストール。
  - UTGB toolkitを用いてデータを表示する。

データベースの実装・公開

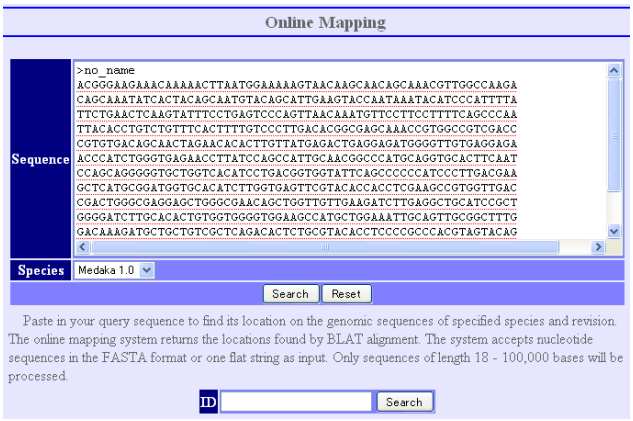
- 受講者が研究で使用する新規データをゲノムブラウザーに表示する。
  - 発現量データを表示するトラックの開発。
  - 配列特異性を視覚化するトラックの開発。
    - “RepeatScape”として公開。
  - Fosmid-end解析, 完全長cDNAアセンブリーの解析をブラウザーに表示。
- データ解析・論文作成に活用されている。





UTGB Medaka Online Mapping

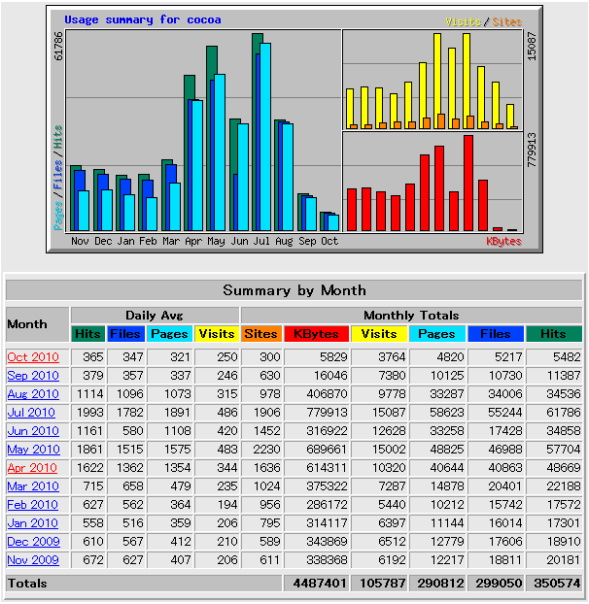
- クラスターでアラインメントの計算。
- ウェブブラウザーでマッピング結果を表示。



演習ノートをウェブ上で公開

過去の演習ノートの整理・改善を行った。

演習ノートのアクセス統計



Summary by Month

Month	Daily Avg				Monthly Totals					
	Hits	Files	Pages	Visits	Sites	KBytes	Visits	Pages	Files	Hits
Oct 2010	365	347	321	250	300	5829	3764	4820	5217	5482
Sep 2010	379	357	337	246	630	16046	7380	10125	10730	11387
Aug 2010	1114	1096	1073	315	978	406870	9778	33287	34006	34536
Jul 2010	1993	1782	1891	486	1906	779913	15087	58623	55244	61786
Jun 2010	1161	580	1108	420	1452	316922	12628	33258	17428	34858
May 2010	1861	1515	1575	483	2230	689661	15002	48825	46988	57704
Apr 2010	1622	1362	1354	344	1636	614311	10320	40644	40863	48669
Mar 2010	715	658	479	235	1024	375322	7287	14878	20401	22188
Feb 2010	627	562	364	194	956	286172	5440	10212	15742	17572
Jan 2010	558	516	359	206	795	314117	6397	11144	16014	17301
Dec 2009	610	567	412	210	589	343869	6512	12779	17606	18910
Nov 2009	672	627	407	206	611	338368	6192	12217	18811	20181
Totals						4487401	105787	290812	299050	350574

受講者の進路

（卒業生の就職先、または現在の所属）

- DBCLS：2名
- 研究員：2名
- 博士課程在学中：5名
- 修士課程在学中：3名
- 就職：6名
  - IBM, JR東日本, 日立, 野村総研など
- 不明、その他：3名

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	統合データベース支援：アノテータ・キュレータの教育
実 施 機 関 名	長浜バイオ大学
代 表 研 究 者 名	池村　淑道

１．課題開始時における達成目標

① アノテータ・キュレータの教育

学校法人関西文理総合学園長浜バイオ大学は、DB のアノテータ・キュレータの人材養成を目標の一つとして、実習にゲノムアノテーションを取り入れてきた。本人材育成では、これまでの実績を基礎に、情報・システム研究機構と連絡を密に取り、アノテーションが十分になされていない微生物ゲノム配列を入手し、実データに基づくゲノムアノテーション実習を行い、本統合 DB からの公開データとして提供する。公開を前提としたアノテーションを実践することは、技術の習得に加えて、データベース作成の意義や責任を体験することになり、ミスマッチの少ない人材養成となる。なお、これらの人材育成プログラムには、情報・システム研究機構でアノテータやキュレータとして雇用する人材の育成も含む。ゲノムアノテーションの実習授業で使用した教材を基に、アノテータ養成のための自習用教材を作成し、インターネット上での公開を行う。

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

② 自己組織化マップによる混合微生物遺伝子のアノテーション手法を用いた人材養成

ライフサイエンス分野の大量情報を対象に、多様なアノテーション情報付加と情報可視化に必要となる技術の実践的な教育を行い、作成したデータベースを本統合 DB からの公開データとして、利用しやすい形式で提供する。この教育で使用した教材を基に自習用教材を作成し、インターネット上での公開を行う。

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

③ シニア研究者のチームによる分散型の高度なキュレーションモデルの確立

上記の２業務で得られたアノテーション成果の質の向上を目的に、専門知識を持つ複数のシニア世代研究者をハブとする、チームによる分散型の高度なキュレーションモデルの確立をめざす。本事業のアノテーション作業に活用すると同時に、他大学や企業を含む研究機関の人材を対象に、短期滞在型を含む講習会を 3 回程度開催して、大学や研究機関への成果の普及を図る。

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

成果物の名称:「健康への貢献遺伝子データベース」ならびに「持続型可能社会への貢献遺伝子データベース」

特徴:学生実習で作成したデータベースであり、実習・自習用の教材もダウンロード可能にしている。

URL: http://dbcls.nagahama-i-bio.ac.jp/

上記以外に、既に完了した実習の経験を基礎に作成した教材として、ダウンロード可能な実習・自習用教材としては、「バイオ技術の統合的データベース作成のための情報収集」、「バイオ分野のデータベースの統合的な利用」、「病気に関係する遺伝子の探索」、「健康への貢献遺伝子探索」の４教材がある。

次世代シーケンサーの産出する大量ゲノム情報処理の教材は、平成 21 年度の本学実習用には作成を完了したが、他大学の実習や自習用の教材としては、平成 22 年度の本学の後期実習の経験を基に、更新を加えた後に公開の予定である。

成果物の名称:　tRNADB-CE：エキスパートがキュレートした tRNA データベース

特徴:世界最大で高品質の tRNA 遺伝子のデータベース

URL: http://trna.nagahama-i-bio.ac.jp/cgi-bin/trnadb/index.cgi

統合 TV の 13 コンテンツを作成し、DBCLS から公開した。

（2）進捗及び成果

①　アノテータ・キュレータの教育

本人材育成の事業においては、ライフサイエンス分野の DB 作成について、学部生を対象にした実習授業を行い、そこで得られる経験を基に、教材の更新を進めて、他大学の実習や自習用として利用可能な教材として完成し、公開を行った。

実習自体については、１回生、２回生、３回生の実習のいずれも、カリキュラムとして設定しているので、計画時間数どおりに完全に実施している。

既に完了した実習の内容と作成教材。

1. 2 回生 320 名を対象とした「バイオ分野のデータベースの統合的な利用」の実習を行い、教材を公開した。DBCLS や JST が作成したコンテンツを主に用いて、DB やポータルサイト利用の重要さを教育した。加えて、作成・公開した教材の説明用の統合 TV コンテンツを作成し DBCLS より公開した。

2. 3 回生 250 名の初級アノテーション教育として、「持続型可能社会への貢献遺伝子データベース」作成の実習を行い、その DB ならびに実習教材の公開を行った。

   成果物の DB の名称は「持続型可能社会への貢献遺伝子データベース」である。特徴としては、学生実習で作成した DB であり、学生の氏名入りでの公開している。教材は、他大学での実習に使用されることを目標にしており、自習用の教材をダウンロード可能にしている。既に、他大学での実習に利用されている。

   URL: http://dbcls.nagahama-i-bio.ac.jp/

   作成・公開した教材と DB の説明用の統合 TV コンテンツを DBCLS より公開している。

3. 3回生250名の初級アノテーション教育として、ヒトゲノム配列を対象にして、「病気に関係する遺伝子の探索」の実習を行い、その実習教材の公開を行った。各学生が自身で興味を持つ病気について、関係する遺伝子やその変異を探索した。実習後のアンケートを見ると、自身で興味を持つ病気を設定したことで、意欲的に実習に取り組む姿勢が感じられた。他大学の実習や自習用、更には一般市民を対象にした講習会等に適した教材と考えている。自習用の教材をダウンローロ可能にしている。

4. 本学3回生（50名）を対象とした平成20年度の中級アノテーション教育では、新規微生物由来のドラフトゲノム配列を対象に実施した。5グループについて、相互に気付かせずに、同一ゲノムのアノテーションを行った。結果を配列の提供を受けた公的機関へ報告した。この新規微生物ゲノムについては、当該の公的機関でも別途アノテーションを行っていたが、学生が行ったアノテーションは、概略としてこれと同様な結果であった。学生のグループ間で、遺伝子位置や機能で差異の生じた場合について、5グループの結果を多数決的に採用すると、大半について公的機関のものと同一になった。教材を含む教育方法が適切であったと考えられる。

5. 3回生（50名）を対象とした平成21年度中級アノテーション教育では、ある国立大学法人の研究グループが次世代シーケンサーで決定した1000万件に近い微生物ゲノム由来の大量断片配列データを対象に実施し、結果を当該研究グループへ報告した。平成20年度とゲノムアノテーションの対象配列の種類を変更した理由は、次世代シーケンサーの登場で大量な断片配列が産出され、その情報処理が重要になったとの状況の変化による。各学生に大量な断片配列データを与えることは、UNIX やLINUXを用いたプログラム処理の必要性や便利さを実感させるのに良い素材となることが明らかになった。この大量ゲノム情報処理の実習用教材は、平成21年度の本学実習用には作成を完了したが、世界的に定評のあるプログラム類も次々に登場しており、他大学の実習や自習用の教材としては、平成22年度の後期実習の経験を基に更新を加えた後に公開を予定している。

6. 1回生50名を対象にしたDB実習において、ライフサイエンス分野の「実験技術に関する統合的データベース作成」のための実習を行い、実験技術の原理やマニュアルや機器画像や動画等に関する情報収集を行った。これは、当初計画には入っていなかったが、DBCLSでの議論や中間評価での議論等で、情報処理と実験の両方に習熟した人材育成の必要性が指摘されたことを受けて、平成21年度から実習を開始した。ここで収集したコンテンツは、当該1回生が2回生ならびに3回生になった際のMySQLやOracleのDB実習において利用する。

② 自己組織化マップによる混合微生物遺伝子のアノテーション手法を用いた人材養成

国際DNAデータバンクには、遺伝子機能も生物系統もアノテーションされていない、メタゲノム解析で得られている大量なゲノム断片配列が集積している。この環境微生物由来のメタゲノム配列を対象として、4回生4名・シニア世代1名・雇用した研究員の共同作業として、有用遺伝子の探索とデータベース構築を行った。これらの有用遺伝子を主対象に、自己組織化マップによる相同性によらない生物系統の推定を行い、約600件の遺伝子の生物系統を推定した。それらの成果は平成21年3月から長浜バイオ大学及びDBCLSから公開しているデータベースへ追加した。

URL: http://dbcls.nagahama-i-bio.ac.jp/

4回生4名の内の進路については、1名は本学の大学院へ進学して、インフルエンザウイルスAの既知の全ゲノム配列を対象にした自己組織化マップによる特徴抽出と、可視化技術を組みこんだデータベース化を行っている。他の1名は、他大学の医学系大学院へ進学し医療情報処理の分野へ進み、他の2名は来年4月に、情報と環境関係の専攻へ進学予定である。

④ シニア研究者のチームによる分散型の高度なキュレーションモデルの確立

シニア研究者3名・学部4回生3名・助教1名・雇用した研究員の共同作業により、tRNAデータベース（tRNADB-CE）のデータと機能の拡充を行った。具体的には、シニア研究者1名と助教1名のグループがハブとなり、地域分散型で高度なキュレーションを行った。原核生物2228種、真核生物種12種、葉緑体121種、ウイルス171種、次世代シーケンサー産出メタゲノム配列を含む2.3億本のメタゲノム配列より、tRNA遺伝子を網羅的に探索し、3種類の計算機プログラム間で結果の異なる遺伝子をエキスパートがキュレートした後に公開している。

成果物の名称:　tRNADB-CE：エキスパートがキュレートしたtRNAデータベース

URL: http://trna.nagahama-i-bio.ac.jp/cgi-bin/trnadb/index.cgi

特徴:世界最大の高品質なtRNA遺伝子のデータベースで、入力レコード数が287,102レコード、データ量は39Mbである。Nucleic Acids ResearchのDB特集号(2009)での発表が受理され、アクセス数は年間で180万件程度に達している。NARのDB特集号(2011)での発表も受理されており、1月に出版予定であるが、既にOn lineで公開されている。

このDBの内容と使用法に関しての統合TVも作成しており、DBCLSから公開している。

講習会や出前実習としては、「環境微生物ゲノム配列からのお宝遺伝子発掘など」について、長浜バイオ大を会場にした3回の実習講習会以外に、他大学・機関（松本大学、統合DB講習会等）での出前授業をも実施した。高校生（滋賀県立虎姫高校等）を対象とした実習講習会としては、「バイオ分野のデータベースの利用」を行った。

データベース高度利用者養成担当のお茶の水女子大学との相互連携を目的に、お茶の水女子大学での講義の実施、ならびに長浜バイオ大学での瀬々准教授による実習を実施した。

成果の学会や研究集会での発表については、参考添付資料に示したように45回行った。DB分野での発表に加えて、教育分野での発表を積極的に行ったところ、教材の使用等に係わる問い合わせ多く受けている。この分野の教材が不足していたことによると考えられる。

統合TVのコンテンツは、学部4回生が総計で13作品を作成し、DBCLSから公開している。

３．当初目標に対する達成度

「tRNADB-CE 2011: tRNA gene database curated manually by experts」の場合、入力レコード数が287,102レコード、データ量は39Mbで、アクセス数は年間で180万件程度である。NARのDB特集号(2009 & 2011)での発表受理を目標に設定したので、収録データ数が当初の目標を超えただけでなく、DB作成過程でのシニア世代の具体的な提言に基づき、検索機能の向上が実現した。アクセス数も当初の予想を超えている。

学生実習で作成した5編の教材、「バイオデータベースの統合的な利用」、「病気に関係する遺伝

子の探索」、「健康への貢献遺伝子探索」、「バイオ分野の実験技術に関する統合的なデータベース作成」、「持続可能型社会への貢献遺伝子探索」を長浜バイオ大だけでなく、統合ホームページからも公開している。件数は当初計画を満たしている。長浜バイオ大での実習時での、学生の結果発表等で理解度や到達度を把握した後に、自習用の教材として更新を加えて公開を行っているので、他大学での利用が容易と考えている。これらの実習に関する統合 TV を作成し、DBCLS から公開していることも、他大学での利用や自習を容易にすると考えている。これらも、当初目標を達成している。

４．中間評価に対する対応

【非常に重要な取り組みだと思うが、成果が外に明らかになっていない。具体的にどのような人材が、どのように活躍しているのだろう。シニア＝池村先生のことなのか？】→【京都大学（２名）、弘前大学（１名）、自治医科大（１名）を退職した、計４名の実験系の研究者に、tRNA 遺伝子 DB の精査をお願いしています。具体的には、これらシニア世代研究者と学部生や若手研究者との共同作業で、世界で最大規模・最高品質と考えられる tRNA 遺伝子 DB を作成し、公開しています。シニア世代から若手への知識継承の場にもなっています。Nucleic Acid Res の DB 特集号や、遺伝学会、分子生物学会、日本 RNA 学会で発表をしております。池村はこれらの活動の統括的な役割をしています。】

【幅広い層の組み合わせで新しい知識を発見する取り組みは面白いと思います。シニア研究者には、今後増えるであろう退職した研究者なども含まれると面白くなると思います。】→【現在は tRNA 遺伝子を中心に、DB の精査や DB 機能向上の活動を、退職したシニア世代の研究者にお願いをしていますが、ライフサイエンスの他の分野へも少しずつ活動を広めようと計画しています。】

【このような取り組みは重要だとは思いますが、単発の取り組みではなく、継続・発展させていけるような仕掛けがあるとよいと思います。】→【色々な大学等を退職したシニア世代が、学部生を含む若手人材との共同作業として、高品質な DB を作成する試みは、知識継承の場ともなっており、シニア世代が積極的・意欲的に取り組んでいます。着実に成果を挙げているので、このような具体的な取り組みをプロトタイプとして、より継続的に発展させるための方針を考える時期に来ていると感じています。統合 DB センターの事業の全体像における意味付けもあり得るように思えます。】

【高度な専門性に裏付けられた知識と経験のあるシニアが活躍し、研究・教育活動へ貢献することはとても重要です。応援しています。こういった取り組 みのできるシニアがどの程度いるかは不安がありますが、少しずつ裾野を広げていければよいのだと思います。高度な専門知識あるシニアは体系だてた物事の考え方を身につけているはずで、内容は最先端でない場合もあるかもしれませんが、考え方を学ぶという観点から、協同作業としての研究機会をもてる学生は幸せ と思います。】→【退職後のシニア世代研究者が若手人材と共同して高品質の DB を作成することは、シニア世代にとっても自分が携わって来た分野の最新の知識に触れることになり、高品質 DB の公開等を通じて、その分野における存在感をある程度は明示・維持できることになります。この様な活動に意欲を持つシニア世代が、我が国にはかなりの数おられると思えるので、オーガナイザー的な人材がいれば、他大学でも実践可能に思えます。DB の精査のような地味に見える作業は、現役世代の研究者よりもシニア世代に適しているように思えます。この点も、統合 DB センター事業の全体的視点における意味付けもあり得るように思います。】

【取り組み１はシニア研究者の人材活用法として画期的な取り組みであり、興味深い。今後のモデルケースとなりうる。】→【今後のモデルケースとなるために、tRNA 遺伝子 DB 以外に、もう一つの別分野の高品質 DB の作成と公開に挑戦する計画です。】

【tRNA データベースのキュレーションモデルがどのように確立されたのか、上記リンク先からだけでは不明でした。また「学部教育としてのデータ ベース構築」についてもその目的と意義がよくわかりません。教材が成果なのでしょうか？データベースが成果なのでしょうか？データベースだとすると、学部生のアノテーションについて、どのようにクオリティコントロールするのでしょうか？】→【ご指摘の通り、DB の説明文や Nucl Acid Res の論文では、キュレーションモデルをどのように確立し、どのように運用しているのかが不明です。日本語で書いた紹介論文がありますので、DB の説明文の Reference の項で、この日本語論文へのリンクをはりました。長浜バイオ大の学部実習で DB を作成し、その実践を基に他大学でも使用可能な実習教材を作成しています。統合 DB センターの Web ページからもダウンロード可能です。「持続可能型社会や健康への貢献遺伝子データベース」は、現時点では実習で作成したものが中心であり、クオリティコントロールが不十分です。この学生実習で作成した DB を素材に、tRNA 遺伝子 DB のシニア世代チームとは別個なシニア世代チームを形成して、特定分野の高品質 DB の作成を計画しています。】

【取り組み 1 のアウトプットは何でしょうか？キュレーションされたデータベースでしょうか？ キュレーションモデルでしょうか？ 両方とも有用であり，続ける価値はあると思いますが，これらと「アノテータ・キュレータの『教育』」との関係性についての説明が欲しいところです.】→【世界に通用する高品質 DB の作成を実践しながら、キュレーションモデルを構築しようと計画しました。tRNA 遺伝子 DB は、年間で 180 万件程度のアクセスのある DB になりましたので、初期の計画は達成したと考えています。シニア世代と若手人材が共同作業として、高品質 DB を作成することの意義は、DB 作成における責任感の習得を含めて、単なる知識習得以上の教育効果があると考えています。本年度は、次世代シーケンサーが産出する、超大量なゲノム塩基配列のアノテーションを実習に組み入れてました。我が国のライフサイエンス分野が必要としている人材の養成であり、そこで作成する教材も重要になると考えています。】

５．他機関との連携

中核機関より、本事業は 1 大学での人材育成が目標ではなく、その活動の実践から得た経験を基礎に、他大学等での教育にも役立つ教材を作成するようにとの指示を受けていた。従って、自習も可能なように、丁寧な教材の作成を目指した。他大学で実際に使用されることを目標にしたので、学会や研究集会で積極的に発表を行った（合計 45 回）。発表においては、DB 分野での発表に加えて、教育分野での発表を行ったところ、教材の使用等に係わる問い合わせを多く受けた。

次世代シーケンサーの登場で大量な断片配列が産出され、その情報処理を行える人材が不足していることを考慮して、情報・システム研究機構の国立遺伝学研究所 DDBJ のメンバーと相談して、次世代シーケンサー配列の情報処理をゲノムアノテーションの実習に加えた。
外部専門家を入れた運営のための委員会として、「ライフサイエンス分野での実験技術の産業利用の知

識付加キュレータ教育に関わる委員会」を 5 回開催し、「ライフサイエンス分野での実験技術に関する統合データベース作成」に関する意見の提供を受けた。この委員会は、バイオ産業分野のシニア世代の２名で構成されており、実験技術の産業利用ならびにその技術に関するデータベースについて、精査やキュレーションを行える人材に関しての意見と知識提供を受けた。なお、研究分野でのこれらの意見と知識提供は、tRNA 遺伝子データベースのキュレーションを担当している、シニア世代の研究者から随時受けた。

６．今後の見通し、計画、展望

ライフサイエンス分野の情報量の爆発的な増加に伴い、データベースの品質の低下が大きな問題になっている。高品質なデータベースの構築には、各分野の専門知識に基づくキュレーションが重要になるが、少数の個人レベルでの活動には限界がある。シニア世代の研究者のような時間的に余裕のある複数の人材を組織して、現役研究者や学生・院生との共同作業として、分野別に高品質なデータベースを構築することは、知識の継承としても意義深い。大学や研究機関の枠を超えて取り組むのが望ましい活動であり、JST ライフサイエンス DB 統合推進事業において取り組むべき課題に思える。

学部 1 回生や 2 回生の実習では、DBCLS や JST が作成しているデータベースやポータルサイトやコンテンツ類を中心に、それらの有用性や便利さを解説する教材を作成した。他大学での授業等でもこの教材の一部を紹介しているが、授業後の学生アンケートでも良い評価を受けている。DBCLS や JST の活動が、我が国の学生や院生には余り知られていないのが現状であるので、その広報活動の一端を担っていると考えている。この種の授業等を通じての広報活動やその目的のための教材作成は、今後も必要に思える。

７．全体総括

ライフサイエンス分野の情報量の爆発的な増加に伴い、ライフサイエンスの実験と情報解析の両方についての、技術と知識を備えた人材の育成が益々重要になってきた。長浜バイオ大は、コンピュータバイオサイエンス学科を昨年度から開設してこの分野の人材育成を行っているが、既存の教材が乏しく、教材を新規に作成し、学生の習熟度や理解度や反応等を見ながら、更新を続けて来た。各学生が課題を独自に設定し、公開の DB 中ではアノテートされていない新知識の発見を、アノテーション実習に組み入れた。学生の氏名入りで公開するための DB も作成したので、学生が積極的に取り組む姿勢が見られた。他大学でも利用する価値のある教材が作成できたと考えている。

tRNA 遺伝子の世界最大で高品質の DB が作成できたことも、有意義と考えている。

８．特記事項

複数のシニア世代研究者と若手研究者ならびに学生との共同作業として、地域分散型で、DB の高度なキュレーションを行った。tRNA 遺伝子の世界最大で高品質の DB を作成したが、この過程でシニア世代がキュレータとして優れた人材であることが示された。日常的には地域分散型で作業し、年間に 3 回程度会合を持って相互にキュレーションや精査の結果を発表し、集中的に討議する形式の有用性も示された。DB ならびにその紹介論文、学会発表等において、参加したシニア世代の氏名を明記することも、高品質の DB 作成において重要になる。この種の活動のプロトタイプを構築したと考えている。

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）		1,606	0	0	0	1,606
人　件　費（千円）		3,960	4,012	5,315	5,740	19,027
業務実施費（千円）		13,434	8,161	6,867	5,169	33,631
一般管理費（千円）		1,900	1,217	1,218	1,091	5,426
合　計　（千円）		20,900	13,390	13,400	12,000	59,690

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	「tRNADB-CE：エキスパートがキュレートした tRNA データベース」 http://trna.nagahama-i-bio.ac.jp/cgi-bin/trnadb/index.cgi	公開	**データの種類（生物種）：** 微生物の完全長、ドラフトゲノム（合計 1353 種）、メタゲノム解析由来塩基配列（3.5 億件）を対象に、3 種の予測プログラムを用いて網羅的な tRNA 遺伝子探索を行い、287,102 レコード件の tRNA 遺伝子情報（生物種・アミノ酸タイプ・アンチコドン・位置・塩基配列）を 2008 年 7 月より提供している。 開始年月：2008 年 7 月 利用状況：年平均アクセス数　180 万件程度
2	「健康への貢献遺伝子データベース」「持続型可能社会への貢献遺伝子データベース」 http://dbcls.nagahama-i-bio.ac.jp/	公開	人類の健康に役立つ遺伝子を学生が各自で想定して、世界で進行している環境生物試料に関する大規模シークエンシング（メタゲノム解析）で得られた大量な塩基配列を対象に、公的データベースにおいて未だ機能が特定されていない 1,700 万件以上の遺伝子候補に着目して、想定した遺伝子の探索を学生実習として行った。他大学からの探索結果も、相互に情報交換を持ちながら追加し、集合知の形式を目標に学生等の個人名入りで、2007 年 6 月から公開している。自習可能な教材を作成し、公開も行っている。月平均アクセス数は、24,000 件である。登録データの精査を行うために、シニア世代の専門家のキュレーション体制の拡充を図りつつ、高品質なデータの収集を目指す。 開始年月：2007 年 6 月 利用状況：年平均アクセス数　20 万件程度

（2）その他の成果物（（1）、（2）に該当しないもの）

通番	名称	公開／未公開	概要
1	「持続可能型社会への貢献遺伝子探索」の教材と解説用の統合 TV	公開	実習用に作成した教材を、実習での学生の理解度や習熟度を踏まえて、自習用に更新後にダウンロード可能にした。
2	「健康への貢献遺伝子探索」の教材と解説用の統合 TV	公開	実習用に作成した教材を、実習での学生の理解度や習熟度を踏まえて、自習用に更新後にダウンロード可能にした。
3	「バイオ技術の統合的データベース作成のための情報収集」の教材と解説用の統合 TV	公開	実習用に作成した教材を、実習での学生の理解度や習熟度を踏まえて、自習用に更新後にダウンロード可能にした。
4	「バイオ分野のデータベースの統合的な利用」の教材と解説用の統合 TV	公開	実習用に作成した教材を、実習での学生の理解度や習熟度を踏まえて、自習用に更新後にダウンロード可能にした。
5	「病気に関係する遺伝子の探索」の教材	公開	実習用に作成した教材を、実習での学生の理解度や習熟度を踏まえて、自習用に更新後

			にダウンロード可能にした。
6	「tRNADB-CE：エキスパートがキュレートした tRNA データベース」の解説用の統合 TV	公開	tRNADB-CE の DB には多様な検索機能が用意してあるので、それらの機能を含めて DB の内容を日本語で解説した統合 TV のコンテンツである。

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	『持続可能型社会への貢献遺伝子データベース』の構築にあなたも参加してみませんか？	上原啓史	平成 20 年 7 月 27 日、同 11 月 26 日	長浜バイオ大学	オープンキャンパス	大学生・高校生・一般人を対象とし、7 月・11 月それぞれ約 50 名、及び約 40 名の参加者
2	″学部教育における DDBJ の活用について″(講師)	阿部貴志	平成 20 年 6 月	国立遺伝学研究所(三島)	DDBJing 講習会	
3	″ゲノム配列に潜む生物種による特徴をコンピュータで探る″(講演)	池村淑道	平成 20 年 9 月	名古屋大学(名古屋)	日本遺伝学会第 80 回大会市民公開講座	遺伝学会の主催する、大学生・高校生・一般人を対象とした公開講座。

（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	「持続可能型社会への貢献遺伝子データベース」	ゲノムひろば 2007：大阪	2007 年 10 月	一部の参加者に簡単な実習も行った。
2	「持続可能型社会への貢献遺伝子データベース」	ミニゲノムひろば 2007：福岡	2007 年 12 月	一部の参加者に簡単な実習も行った。

（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	"Alignment-free method for phylogenetic classification of sequences from environmental mixed genomes: Self-Organizing Map (SOM)."	T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya and T. Ikemura,	Symposium on Evolutionary Genomics (Costa Rica)	January 7-12, 2007.	Oral
2	"A novel bioinformatics strategy for prediction of functions of a massive amount of poorly-characterized	T. Abe, S. Kanaya, H. Uehara, T. and Ikemura,	XX International Congress of Genetics (Berlin,	July 17, 2008.	Oral

	protein genes obtained by metagenome studies"		Germany)		
3	"Unveiling microbial diversity of uncultured environmental microbe mixtures and functions of metagenome sequences using Batch Learning Self-Organizing Map (BLSOM)"	T. Abe, S. Kanaya, and T. Ikemura,	International Symposium on Marine Genomics 2009 (Okinawa, Japan)	Dec 15-18, 2009.	Oral
4	"tRNADB-CE: tRNA gene database curated manually by experts"	T. Abe, T. Ikemura, Y. Ohara, H. Uehara, M. Kinouchi, J. Sugahara, A .Kanai, S. Kanaya, Y. Yamada, A. Muto, and H. Inokuchi,	23rd tRNA workshop (Aveiro, Portugal)	Jan 28 - Feb. 2, 2010	Oral

（4）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	tRNADB-CE 2011: tRNA gene database curated manually by experts.	T. Abe, T. Ikemura, J. Sugahara, A. Kanai, Y. Ohara, H. Uehara, M. Kinouchi, S. Kanaya, Y. Yamada, A .Muto and H. Inokuchi.	Nucleic Acids Research	November 11, 2010 (Advance Access)	論文としては、2011 年 1 月に出版予定
2	"Novel bioinformatics for inter- and intraspecies comparison of genome signatures in plant genomes"	T. Abe, K. Wada, Y. Iwasaki, and T. Ikemura,	Plant Biotechnology	26, 469-477, 2009.	
3	"A novel bioinformatics strategy for function prediction of poorly-characterized protein genes obtained from metagenome analyses"	T. Abe, S. Kanaya, H. Uehara, and T. Ikemura,	DNA Research,	16, 287-298, 2009.	
4	"tRNADB-CE: tRNA gene database curated manually by experts"	T. Abe, T. Ikemura, Y. Ohara, H. Uehara, M. Kinouchi, S. Kanaya, Y. Yamada, A. Muto and H. Inokuchi,	Nucleic Acids Research,	37, D163-D168, 2009, doi:10.1093/nar/gkn692.	
5	"Batch-Learning Self-Organizing Map for Predicting Functions of Poorly-Characterized Proteins Massively Accumulated."	T. Abe S. Kanaya, T. Ikemura	J.C. Príncipe and R. Miikkulainen (Eds.): WSOM 2009,	LNCS 5629, pp. 1–9, 2009.	

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	データベース高度利用者養成
実 施 機 関 名	お茶の水女子大学
代 表 研 究 者 名	瀬々　潤

１．課題開始時における達成目標

生命科学のデータベース（ＤＢ）は、膨大であり日々更新されている。このＤＢから最新のデータを取得し、生命理解へと繋げるには、生命科学の知識のみならずＤＢから必要なデータを適宜抽出、編集、蓄積し、解析や知識発見ができるＤＢ高度利用者が必要である。ＤＢ高度利用者の育成では、生命科学ＤＢの利用に必要な生物学および情報科学の知識を学んだ上で、インターネット・プログラムを通じたデータの利用法、データ処理、基礎統計、データマイニング技術を学ぶための 90 時間程度のカリキュラムを構築し、実践する。また、社会人、主婦からも受講生を受け入れる。

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

データベース高度利用者養成」HP（http://togodb.sel.is.ocha.ac.jp/)上にてカリキュラム概要、講義資料および自習用教材を公開。平成 19 年度より 4 年に渡って講義を行い、計 89 名が受講した。受講生には、18 名の主婦および社会人も含まれる。受講者から平成 20 年度より統合データベースプロジェクト関連の事業に関わる人材を計 10 名輩出した。

（2）進捗及び成果

データベース高度利用者養成　受講者人数

	total	内訳(学生	社会人)	一部受講	total+一部受講
H19	4	(4	0)	13	17
H20	15	(15	0)	4	19
H21	27	(22	5)	8	35
H22	12	(8	4)	6	18

講義資料公開について

-データベース高度利用者基礎

ネットワーク基礎　平成 20 年度よりホームページ（HP）にて資料公開、H22 年度より講義録(pdf 形式)公開。

データベース基礎　平成 20 年度より HP にて資料公開、講義録(pdf 形式)公開準備中。

現代遺伝学基礎　平成 21 年度より HP にて資料公開。

-データベース高度利用者専門

ネットワークを通じたライフサイエンスデータの利用　平成 19 年度より HP 及び Slideshare を通じて公開。

データ解析の基礎　平成 19 年度より HP 及び Slideshare を通じて公開。

データマイニング技術　平成 21 年度より HP 及び Slideshare を通じて公開。

HP(http://togodb.sel.is.ocha.ac.jp/)のページビュー

	訪問者数	一意な訪問者数	ページビュー
H20 7~9	62	39	241
H20 10~12	195	106	1580
H21 1~3	265	90	1260
H21 4~6	429	141	2601
H21 7~9	1232	818	3017
H21 10~12	1326	873	3057
H22 1~3	1474	952	5397
H22 4~6	2504	1821	5524
H22 7~9	3179	2116	7068

受講生は、以下に示すように、データの抽出、編集、蓄積に関する人材として活躍している。

平成 20 年度　自習用教材改訂　2 名、自習用教材公開用サーバ管理　2 名
　　ライフサイエンス統合データベースセンター(DBCLS)にて　統合 TV の教材作成　2 名

平成 21 年度　自習用教材改訂　3 名、自習用教材公開用サーバ管理　2 名
　　DBCLS にて　文書検索システム開発　1 名、　統合 TV の教材作成　2 名

平成 22 年度　自習用教材公開用サーバ管理　5 名
　　DBCLS にて　テキスト処理　2 名、　統合 TV の教材作成　2 名

３．当初目標に対する達成度

カリキュラムは、基礎(５時間 ＊ ３種類)、応用（２５時間 ＊ ３種類）の９０時間分であり、当初予定した時間数を構築した。また、同カリキュラムを用い、社会人、主婦からも合計１８名の参加者を受け入れて、演習を行った。

４．中間評価に対する対応

人材育成を行いながら、同時に中核機関のデータベース開発や個別機関でのデータベース開発の戦力となることが求められていたため、平成 19 年度に策定したカリキュラムの評価と改訂を行い、DBCLS での統合 TV の教材作成および文書検索システム開発に継続的に人材を輩出した。

５．他機関との連携

中核機関開発のサービス利用：

　カリキュラムにおいて、統合 Web Service を利用してデータ取得を行う講義を実施。利用者の拡大に努めた。

　カリキュラム中の資料において、統合データベースプロジェクト構築の横断検索等のページを入り口として講習資料を作成することで、利用者の誘導を行った。

他機関との連携：

平成 20 年度：

＊ 蛋白質核酸酵素に「ライフサイエンスにおけるデータベース構築のための人材育成」を寄稿。教育に関する分担機関である長浜バイオ大学の池村先生と共著。また、同分担機関である東京大学新領域創成科学科の森下真一先生からは、現状をお聞きして執筆。

＊ 日本進化学会第 10 回大会においてワークショップ「統合データベースの活用法:ゲノム情報などを使いこなした効率的な研究のために」を企画、金子に加えてライフサイエンス統合データベースセンターの坊農先生、河野先生、かずさ DNA 研究所の岡本先生に講演をお願いし、統合データベースプロジェクトの取り組みについて宣伝を行った。

＊ 統合データベース講習会: AJACS 近江において、金子がお茶の水女子大学での講習内容の一部を演習も交えて紹介した。

平成 21 年度：

＊ 統合データベース講習会 AJACS 蝦夷 2 にて、瀬々が講演を行った。

＊ 理研 BASE の豊田哲郎部門長にお茶の水女子大学にお越し頂き、講演をして頂いた。

平成 22 年度：

＊ 長浜バイオ大学の阿部先生にお越し頂き、講義をして頂いた。

＊ 瀬々が、長浜バイオ大学に伺い、講義、演習を行った。

＊ ＤＤＢＪの中村保一教授にお越し頂き、講演会を開く予定。（平成 23 年 1 月）

６．今後の見通し、計画、展望

本事業に於いて、計 89 名の受講生を輩出した。残された課題としては、この受講生の中からライフサイエンス統合データベースセンターへの就職、および民間企業や研究所へ受講内容を活かした就職ができるようにすることである。本事業開始当初に予想した以上に、次世代シークエンサーの運用が本格化してきており、膨大なゲノム配列を処理する人材を欲する声が聞かれる。データベース高度利用者の卒業生には、このような状況に対応できる知識を身につけた人材が多数いる。その一方で、本事業に携わる研究所を含め、研究所や企業の雇用拡大は進んでおらず、新しい人材を雇用するのではなく、既に雇っている人材の再教育を求める傾向が強く感じられる。このため、修了生の就職先には IT 系の大手企業が少なくない。再教育を目的とした講習では、本事業で育成した人材が、統合データベース関連事業に参画できる可能性は薄いため、本事業で行うべきか疑問が残る。再教育の要望に応えるには、長期間の授業・演習より、統合データベース講習会の様にサービスの使い方に関する講習会を継続し、統合データベース事業の対外的なアピールと利用者拡大に努める策もあるのではないかと考える。

長期的な視野に立った場合、講習会だけ行って、JST ライフサイエンス DB 統合推進事業内のメンバーの育成や入れ替えを行わないのでは、データベース技術を担える人材の枯渇や、最新の研究にキャッチアップできないメンバーが増える事になるだろう。本教育事業を継続的に行い、ライフサイエンスデータベースを担う人材の育成と発掘に努める必要がある。JST ライフサイエンス DB 統合推進事業(仮称)においては、いずれのスタンスを取るかで教育事業の内容が変化していくと思うが、長期的に当該分野が継続する為には、統合データベースプロジェクト同様の基礎教育（異分野や暫く研究開発から離れている人材への教育）と専門教育の両方を行い、現在我々が策定したカリキュラムをベースに継続する事が良いと考える。

７．全体総括

教育活動を行う事で、統合データベースセンターにリサーチアシスタントとして雇用される人材を輩出してきたこと、カリキュラムを策定し、講習資料を公開したこと、及び、講習において統合データベースセンター開発のサービスを利用し、利用方法を普及したことは、我々の教育成果とみなせ、十分目標

を達成したと考えられる。

８．特記事項

本事業雇用の特任助教が、2 度の国際会議に出席したことにより、海外でもデータベース高度利用については、必要とされているが網羅的に取り扱う講習会はなく、本事業は国際的にも特異であることが分かった。

社会人は時間的な制約により受講が困難な場合があるため、自習できるよう、手順などについても詳しい説明を加えた教材を作成した。

社会人や主婦の受講生を募るため、メーリングリストでの告知や学会での宣伝活動を行った。

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）		6,843	773	766	0	8,383
人　件　費（千円）		6,162	9,436	9,389	9,391	34,380
業務実施費（千円）		5,993	2,099	2,152	1,063	11,308
一般管理費（千円）		1,900	1,230	1,230	1,045	5,407
合　計　（千円）		20,900	13,540	13,540	11,500	59,480

整備実績一覧

（1）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	データベース高度利用者養成　HP	公開	http://togodb.sel.is.ocha.ac.jp/
2	ネットワーク基礎　講義録	公開	pdf 形式で上記 HP よりダウンロード可能
3	データベース基礎　講義録	未公開	11 月 2 日より公開

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	統合データベースプロジェクトにおける人材育成活動の紹介	金子　聡子	日本進化学会　第 10 回大会	2008 年 8 月 23 日	ワークショップを企画
2	情報量が増え続けるデータベースを使いこなすために：データベースからの知識発見	金子　聡子	第 82 回日本生化学会大会	2009 年 10 月 23 日	
3	情報量が増え続けるデータベースを使いこなすために：データベースからの知識発見	金子　聡子	第 32 回日本生化学会大会	2009 年 12 月 10 日	