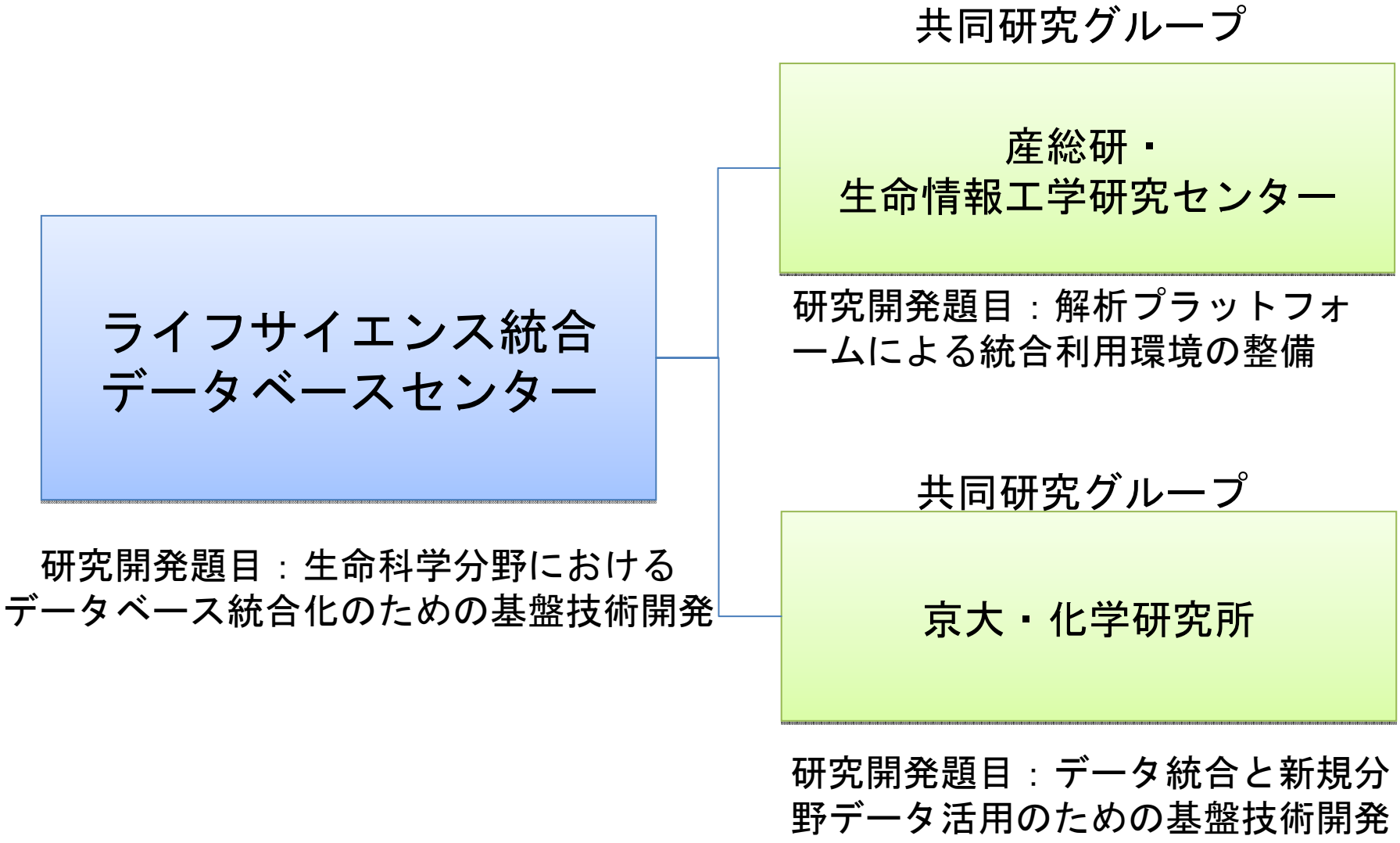
H23年度 基盤技術開発プログラム進捗報告会

ライフサイエンスデータベース統合推進事業
平成23年度進捗報告会

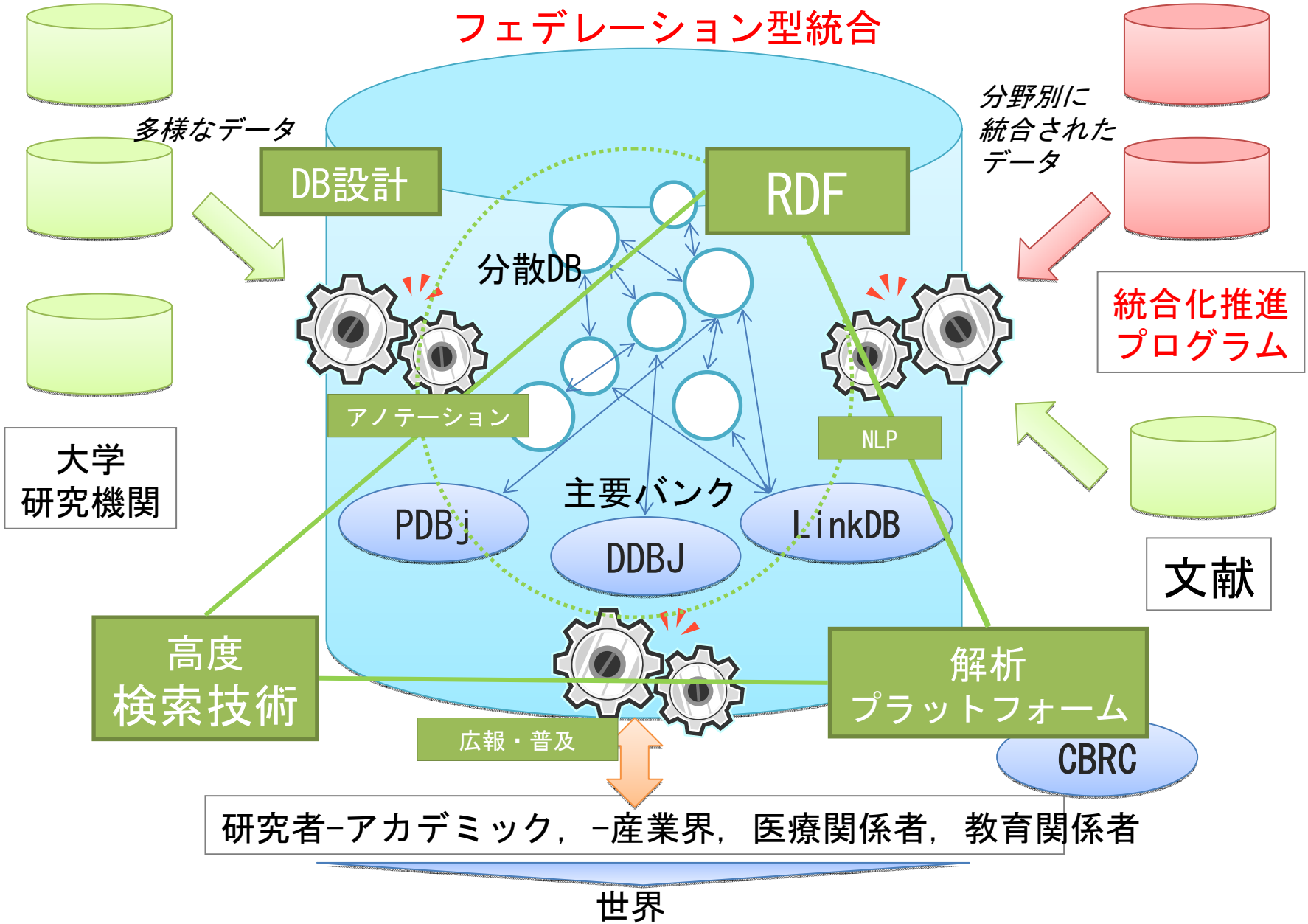
基盤技術開発プログラム
「データベース統合に関わる基盤技術開発」

ライフサイエンス統合データベースセンター

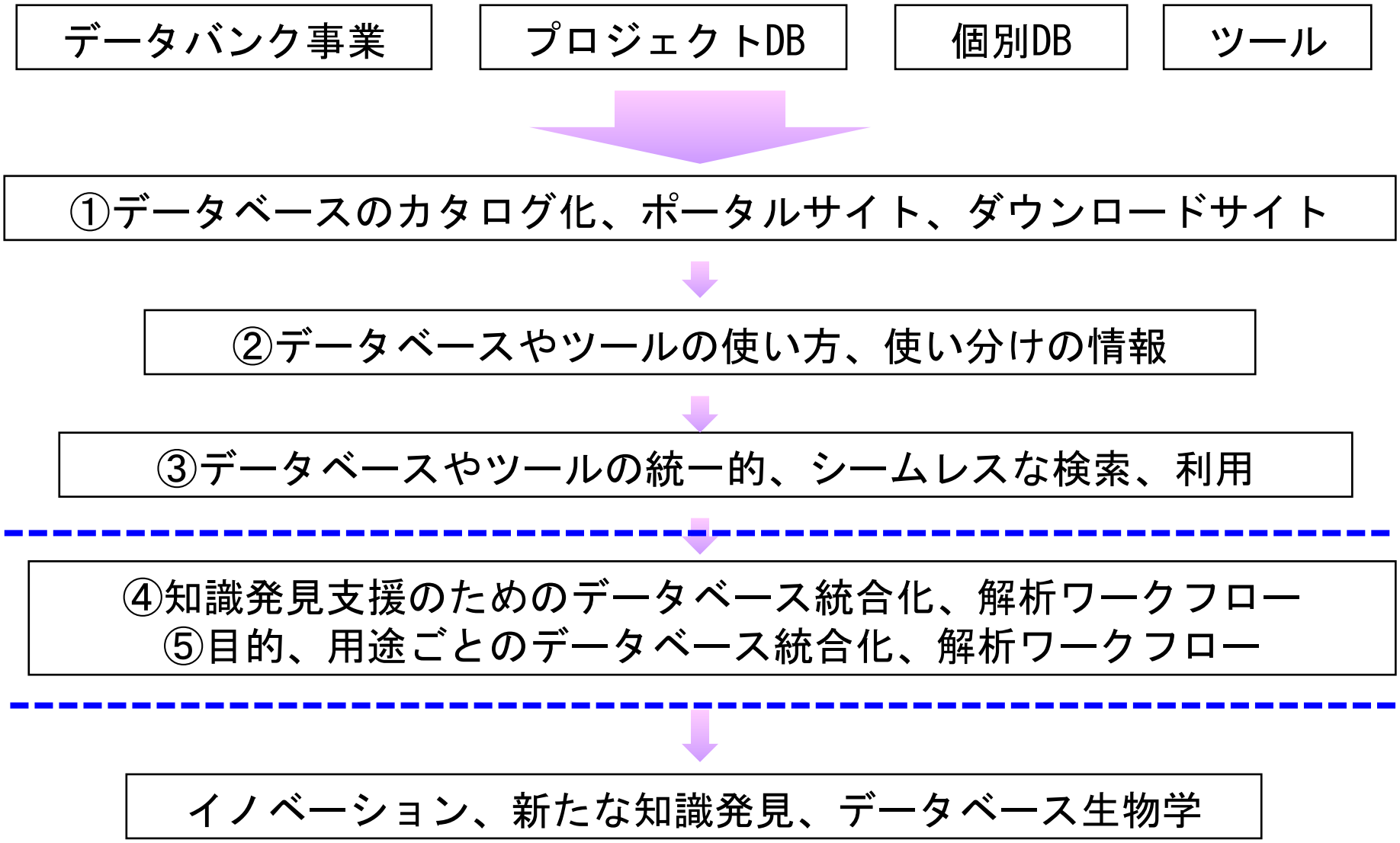
基盤技術開発プログラム実施体制：組織



DBCLSが提案する基盤技術開発の全体像



統合化の全体像とステップ



公募要領の記載(その1)
優先課題

1. 先端的なプログラミング技術によるインターネットを活用した高度検索技術開発を行うこと

2. 国内の基盤的データベースおよび本事業で構築される分野別統合データベースのRDF 化を実現するための，標準フォーマット，オントロジーの提供，RDF 化の支援を行い，RDF コンテンツを公開すること

3. 一連の作業を自動化するための仕組みを構築するとともに，高度なインターネット技術，データベース技術を応用した統合利用環境を整備すること

4. 一貫性のあるポリシーのもとで実用的なオントロジー，辞書，コーパス，標準化技術を開発すること

5. 個人ゲノム等大規模データを活用する技術を開発すること

公募要領の記載(その2)

6. コーパス構築を含んだ文献，画像等コンテンツ活用技術を開発し多様な検索に対応すること
7. 論文解読技術や文献管理システムをベースに論文作成や管理，アノテーション作成を支援する技術を開発し多様な検索に対応すること
8. 医療用画像等に関わる画像データの利用技術（管理・検索・標準化・定量的解析等）を開発すること
9. 統合データベース利用のためのコンテンツ作成（データベースやツールの使い方に係わるコンテンツや，チュートリアルの動画コンテンツ等）とアノテーション支援を行うこと
10. 統合検索システムとアーカイブシステムの有用性の高い高機能化技術を開発すること
11. バイオサイエンスデータベースセンターでの，個人ゲノム等の個人情報に関わるセキュリティ，公開範囲等の検討を踏まえ，適正な内容で公開可能となるよう技術開発を行うこと
12. 脳画像，単一セル計測など，新規分野データの統合化に関わる要素技術を検討し，将来のデータベース化に必要なツール開発等をタイムリーに行うこと

7課題に再編

1. データベースのRDF による統合化 →DBCLS, 京大
2. 解析プラットフォームによる統合利用環境の整備
   → CBRC, DBCLS, 京大
3. インターネットを活用した高度検索技術の開発
   →DBCLS, 京大
4. RDF化に資するオントロジー, 辞書, コーパス整備, 標準化技術開発 →DBCLS, 京大
5. 大規模データの利用技術開発 →DBCLS, 京大
6. 情報統合化・知識発見のためのキュレーション支援 →DBCLS
7. 統合データベースに関わるコンテンツの作成 , 整備 →DBCLS

今年度研究開発内容
(1) データベースのRDFによる統合化

1. RDF化ガイドライン整備に着手
2. RDF化すべきDBの優先順序調査
3. TogoDB RDF化機能プロトタイプ開発
4. DBメタデータRDF蓄積仕様検討
5. 既存のRDFストアの調査・比較検討
6. LinkDB の RDF化着手 (京大化研)

データベースアーカイブの RDF 化

Step1 RDF化するDB絞り込み (50 → 6)

Step2 初期 RDF 化

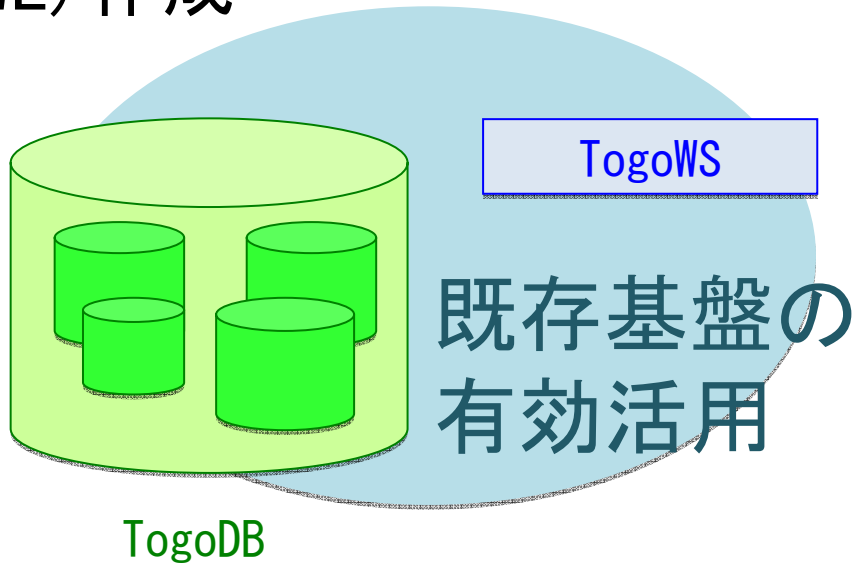
(開発中の TogoDB の RDF化機能を利用)

Step3 メタデータ(データベースのカラム)に対して

初期オントロジー(OWL)作成

Step4 OWL の改良

Step5 SPARQL 検索試行



アーカイブ内の6つのDB
がRDFで利用可能に

RDF化対象のDB

EGTC（マウス：EC細胞のカタログ）
germline（マウス：変異マウスのカタログ）
Yeast Interacting Proteins Database（酵母：Y2HDB）
Building Yeast cDNA sequecing project（酵母：cDNA）
TMPDB（生物全般：膜貫通領域タンパク質の構造DB）
RIKEN SSDB（ヒト・マウス：蛋白3000でのX線構造解析の途中結果の記録と構造）

トリプルストア調査・整備

トリプルストア：RDF(主語-述語-目的語のトリプル)のための
データベース. クエリ言語はSPARQL

トリプルストアの性能はデータの量や問い合わせの複雑さによって大きく異なる

ライフサイエンスデータベースの特徴とトリプルストアの長短の関係を明確にするために，様々なトリプルストアの性能比較を実在するデータベースの RDF を用いて行う

トリプルストアの例:
Virtuoso, 4store, OWLIM, Mulgara, Bigdata, OntoFrame
データベースの例:
Allie, UniProt, DDBJ, PDBj, Cell Cycle Ontology

進捗: テスト用クエリが既に存在するAllie , UniProtから着手.
Allie についてはほぼ終了し，UniProt について進行中

今年度研究開発内容

(2)解析プラットフォームによる統合利用環境の整備

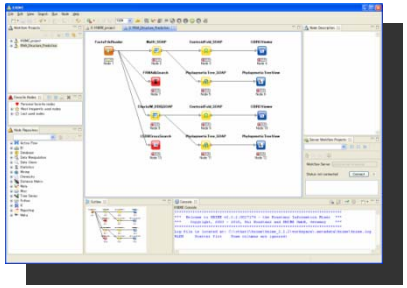
1. テキストマイニングツール群の整備とリポジトリ化
2. 1.を DBCLS Galaxy に反映
3. U-Compare ワークフロー分散処理化
4. ツールの RDF 入出力機能追加(CBRC)

解析ワークフロー

DBCLS Galaxy

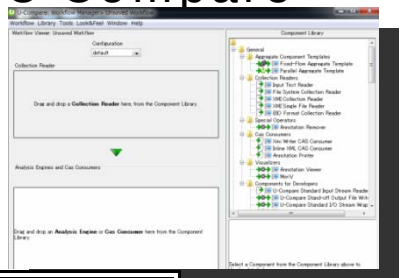
ゲノム解析のためのウェブアプリケーション。ゲノム座標データの演算や主要データベースからのデータインポート, ゲノム配列解析, メール経由のデータ共有, データ中心のヒストリ保存, 解析ワークフローの管理がおこなえる。オリジナルはPSU, EmoryU. DBCLSは文献解析ツールの開発や, TogoWSによる国内DBの利用機能の追加を行っている.

KNIME



ドイツのコンスタンツ大学で開発されたeclipseベースのワークフロー型プラットフォーム. 一つの処理をノードという形式で表し, ユーザーはノードを組み合わせることで, ワークフローを構築し, データの読み込み, 計算, 解析, 可視化が可能. CBRCでは核酸や蛋白の他, RNA解析のノードを開発している.

U-Compare



UIMA準拠の統合自然言語処理環境。互換UIMAコンポーネント群とそれらを含むUIMA準拠の統合環境から成ります。開発者向け. H23年度にワークフロー分散処理化を実施した。

Semantic Automated Discovery and Integration

SADI

Find. Integrate. Analyze.

W3Cセマンティックの規格に従い, 分散して存在しているデータと解析技術の相互連携や知識発見へと繋げるフレームワーク. RDFによる入出力

今年度研究開発内容

(3) インターネットを活用した高度検索技術の開発

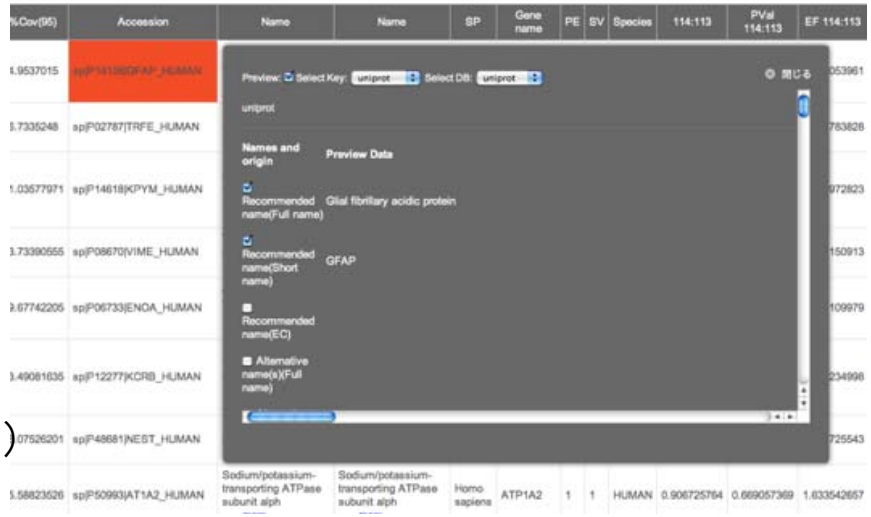
1. 既開発サービスの機能向上
   RDF化による統合を目標に、NBDCと協力し既開発サービス機能向上を図る
2. 有用分野でのプロトタイピング
   微生物、プロテオーム、疾患
3. 検索結果の情報提示法と可視化の開発
   表形式、レポート形式
4. 統合化推進プログラムとの連携
   各グループにヒアリングしてニーズを把握，優先してRDF化するDB整備するオントジー・辞書のフィードバック

有用分野でのプロトタイピング

基盤的技術を活用したユースケースを提示

RDF技術を活用した、表形式データへのアノテーションツールの試作

ユーザが手持ちのデータをアップロード
↓
IDのカラムを指定＆DBを指定
↓
アノテーションを追加するDB名を指定
↓
追加する attributes を指定
↓
（対応するSPARQLを発行し表の右側にカラムを追加）
↓
ユーザが結果をダウンロード



UniProtを対象にRDFモデルを調査し、各attributeを取り出すためのSPARQLを設計して、プロトタイプを作成した。今後は対象DBを増やしていく（RDF化が必要なものは、課題1でRDF化

検索結果の情報提示法と可視化の開発

バクテリアをモデルとしたセマンティックアノテーションDBの試作

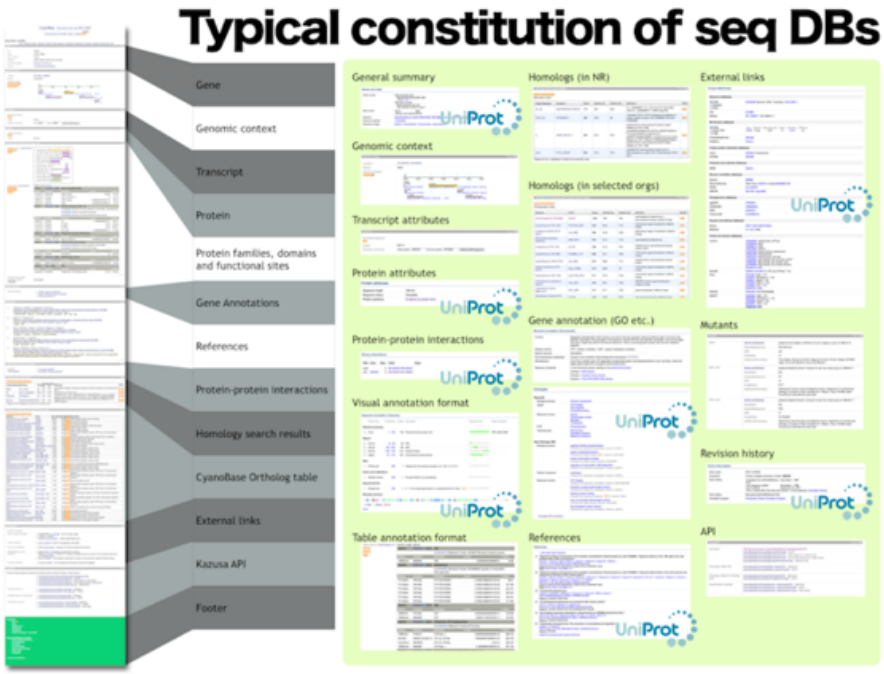
・ゲノムDBで繰り返し利用される
ゲノム情報要素（Stanza）の調査整理

1. General summary of organisms
2. General summary
3. Genomic context
4. Transcript attributes
5. Protein attributes

...

・情報源データベース

GTPS, UniProt, Refseq, MBGD
Pubmed ...

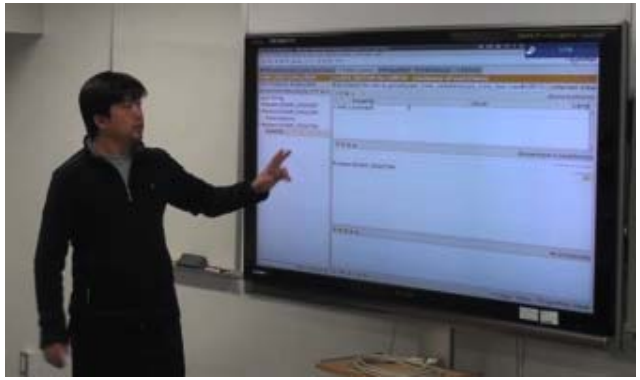


各DBのRDFモデルを調査し、情報を整理して表示するためのSPARQL設計を行った。
プロトタイプ作成中。今後は構成要素（Stanza）を増やしながら、RDFデータの更新系や大量データを格納したときのRDFストアの性能などを検討する。

統合化推進プログラムとの連携

NBDC主催の実務者連絡会に技術協力

9/28　実務者連絡会　準備的会合
10/5　第一回実務者連絡会議
10/21　オントロジーレベルの検討会議
11/11　化合物オントロジーの検討会議
12/6　オントロジー講習会（講師：金）



統合化推進プログラムとの個別連携

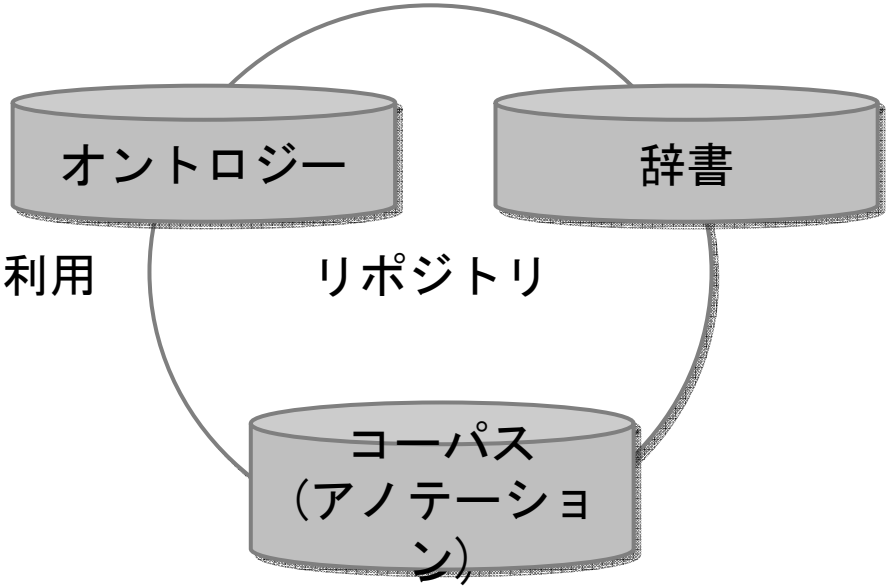
・ゲノム・メタゲノム情報を基盤とした微生物DBの統合（岡本）
　環境ゲノムオントロジー（MEO）の開発
　環境メタゲノムデータへのMEOの適用とRDF化（LODチャレンジ2012投稿）
・糖鎖統合データベースと研究支援ツールの開発（河野）
　糖鎖構造URIの設計
　糖鎖オントロジーの開発
　糖鎖DBのRDF化

今後も引き続き実務者連絡会、個別連携に協力していく

今年度研究開発内容
(4) RDF化に資するオントロジー,辞書,コーパスの整備, 標準化

1. 標準フォーマット開発
2. トップレベルオントロジー開発
3. 自然言語処理、テキストマイニング技術の整備

- Bioportalをプラットホームとして利用
  サポートツール開発
  OntoFinder/OntoFactory
- コーパスアノテーション
  BioNLP Shared Task標準フォーマット利用
  コーパスリポジトリの開発
  アノテーションの為のツール開発



オントロジー構築サポートツール開発

OntoFinder/OntoFactory

	Drop area	IMR	PR	GO	nif
p65 subunit		TF65_ HUMAN [EXACT] (0.0333333)	transcription factor p65 [EXACT] (0.0125)		
intracellular receptor		intracellular ligand receptor [EXACT] (0.0166667)	tumor necrosis factor receptor superfamily member 10C [EXACT] (0.00666667)	intracellular receptor mediated signaling pathway [EXACT] (0.02)	intracellular receptor-mediated signaling pathway [EXACT] (0.00833333)
cytoplasmic inhibitor					
NF-kappa B		NIK [EXACT] (0.00714286)		release of cytoplasmic sequestered NF-kappaB [RELATED] (0.02)	

Ontology URI

OntoFactory

ユーザが用語を入力すると(複数可)
BioPortalから適切なontologyを推薦する

言語解析 アノテーションツール

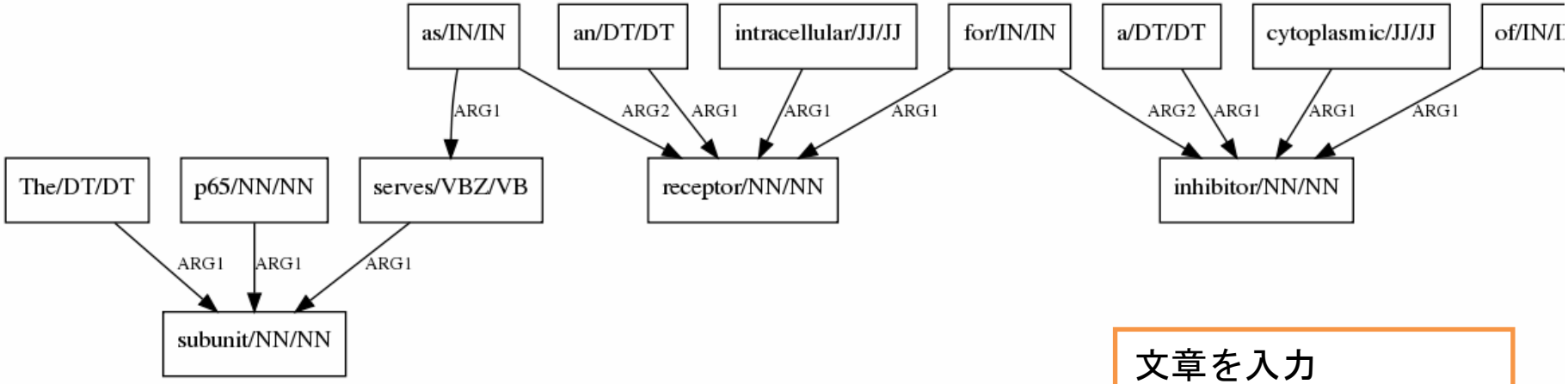
Annotation preprocessing

powered by Enju, a HPSG parser.

Noun chunk extraction

The p65 subunit serves as an intracellular receptor for a cytoplasmic inhibitor of NF-kappa B.

Semantic graph (predicate-argument relations)



Mapping to ontologies

```
p65 subunit
intracellular receptor
cytoplasmic inhibitor
NF-kappa B
```

ontologies to be searched (vids):

submit

文章を入力
↓
構文解析(Enju)
↓
用語を抽出
↓
オントロジにmapping

今年度研究開発内容

(5)大規模ゲノム配列データの利用技術開発

1. メタ情報による大規模ゲノム配列データの整理・再利用促進技術開発
   - ダイジェスト版の構築・開発・維持
   - RNA配列を中心にしたDB利用技術開発
2. 遺伝子発現リファレンスデータセット整備
   - RefExの構築維持管理
   - SRAからのデータ抽出技術開発
3. 医療用画像データの利用技術開発

メタデータを活用した
NGSデータの検索技術開発

（例）論文と対応付けのある信頼度の高いNGSデータを検索する

Search by publications - 文献から探す

→ Publications referring to NGS data

Search by diseases - 疾患から眺める

- Frequency List - 頻度別
- Tree View of Diseases - 疾患カテゴリ別
- All List of the pairs of SRA-Disease - 全リスト

Study Type: Epigenetics Platform: Species:

Total: 98 << first < prev 1 2 3 4 5 6 7 8 9 10 next > last >> 10

PMID	Article Title	Journal	Vol	Issue	Page
21602827	Determinants of nucleosome organization in primary human cells.	Nature	474	7352	516-20
21464311	Substance-specific and shared transcription and epigenetic changes in the human hippocampus chronically exposed to cocaine and alcohol.	Proc Natl Acad Sci U S A	108	16	6626-31
21427231	Lsh, chromatin remodeling family member, modulates genome-wide cytosine methylation patterns at nonrepeat sequences.	Proc Natl Acad Sci U S A	108	14	5626-31
21355038	Integrative genomic analysis of human ribosomal DNA	Nucleic acids research	39	12	4949-60
21355038	Integrative genomic analysis of human ribosomal DNA	Nucleic acids research	39	12	4949-60

Search by statistics - 統計値から探す

The number of projects are indicated in "Study Types" table. The totals in "Platforms" and "Species of Samples" are larger than one of "Study Types" because a project can contain some platforms and sample species.

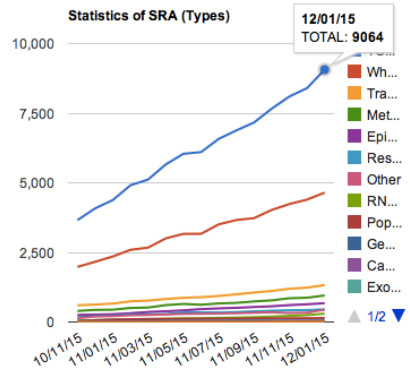
実際のプロジェクト数は"Study Types"に書かれているものです。 "Platforms"や"Species of Samples"のtotalの値は、 複数のプラットフォームで行われた実験が1つのプロジェクトでなされる 場合が多々あり、ダブルカウントしているので、数字が大きくなっています。

2012-02-10 updated.

Study Types

Whole Genome Sequencing	4848
Transcriptome Analysis	1359
Metagenomics	1033
Epigenetics	706
Resequencing	502
Other	464
RNASeq	332
Population Genomics	144
Gene Regulation Study	51
Cancer Genomics	32
Exome Sequencing	25
Pooled Clone Sequencing	13
Synthetic Genomics	5
Forensic or Paleo-genomics	5
Total	9519 (studies)

- Growth of the number of studies (including Total)



- Growth of the number of studies (without Total)

RNA配列DB検索エンジンの開発

統合遺伝子検索
GGRNA

Home | Help | Advanced search

CUGCUGCUGCUGCUGCUGCUGCUGCUGCUGCUGCUG 再検索 Homo sapiens (human)

2012-02-14 18:05:06, GGRNA : RefSeq release 51 (20120113)

Summary:

search term:	hits:	results:
seq:CUGCUGCUGCUGCUGCUGCUGCUGCUGCUGCUGCUG	9	NM_001081560, NM_001081562, NM_001081563, NM_001126054, NM_001126055, NM_001136234, NM_003688, NM_004409, NR_002717
[AND]	9	NM_001081560, NM_001081562, NM_001081563, NM_001126054, NM_001126055, NM_001136234, NM_003688, NM_004409, NR_002717

Results:

検索語に色がつきます。重なると色が濃く表示されます。

Homo sapiens dystrophia myotonica-protein kinase (DMPK), transcript variant 3, mRNA. (2877 bp)
..... ggtccttgtagccgggaatg**ctgctgctgctgctgctgctgctgctgctgctgctgctgctgctgctgctgctgctgctg**gggggatcacagaccatttc .....
position 2304 2307 2310 2313 2316 2319 2322 2325 2328 (CDS: 206 - 2080)
Synonym: DM; DM1; DM1PK; DMK; MDPK; MT-PK
NM_001081560.1 - Homo sapiens (human) - NCBI - UCSC - RefEx発現量

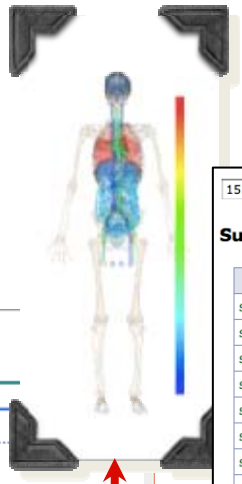
Homo sapiens dystrophia myotonica-protein kinase (DMPK), transcript variant 4, mRNA. (2873 bp)
..... ggtccttgtagccgggaatg**ctgctgctgctgctgctgctgctgctgctgctgctgctgctgctgctgctgctgctgctg**gggggatcacagaccatttc .....
position 2300 2303 2306 2309 2312 2315 2318 2321 2324 (CDS: 206 - 2083)
Synonym: DM; DM1; DM1PK; DMK; MDPK; MT-PK
NM_001081562.1 - Homo sapiens (human) - NCBI - UCSC - RefEx発現量

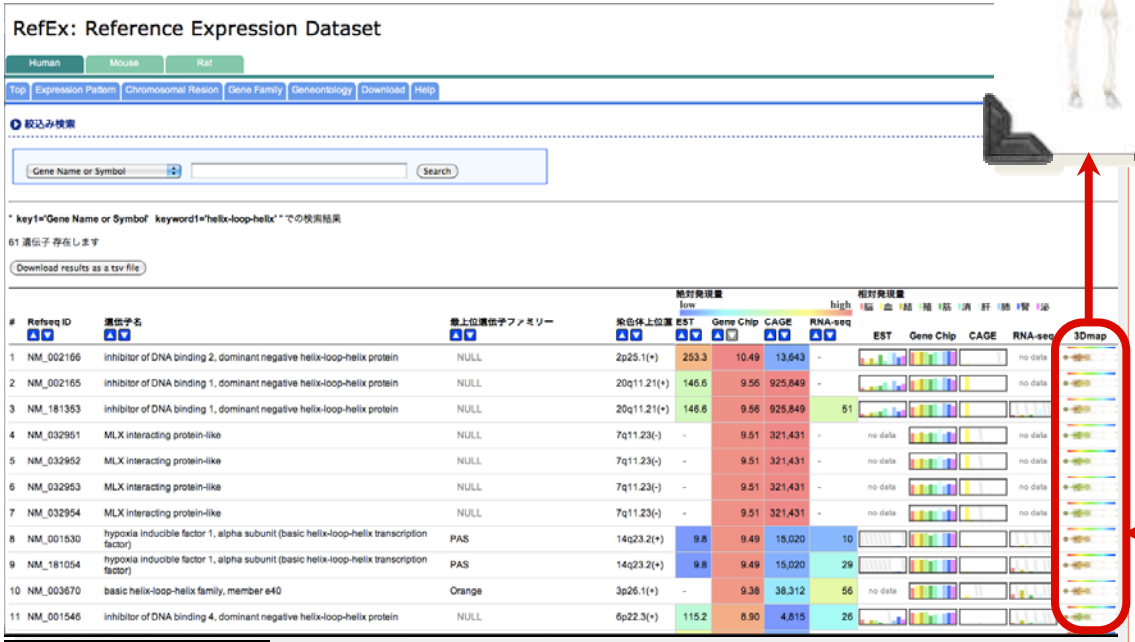
Homo sapiens dystrophia myotonica-protein kinase (DMPK), transcript variant 1, mRNA. (3261 bp)
..... 2691 2694 2697 2700 2703 2706 2709 2712 (CDS: 545 - 2464)

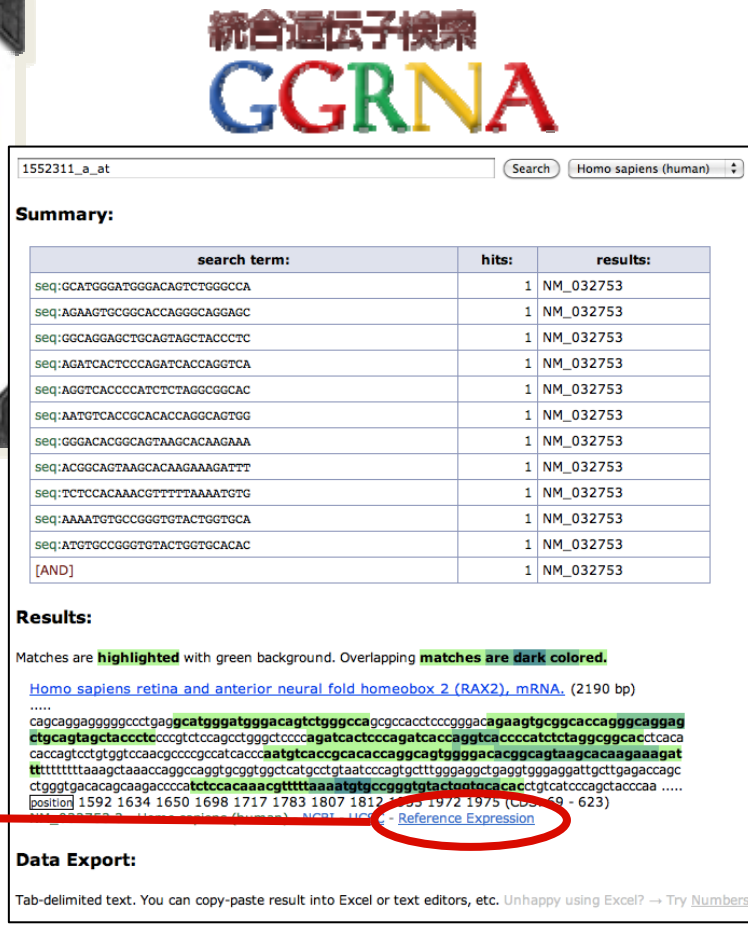
RefEx: Reference Expression dataset

EST, GeneChip, CAGE, RNA-seq の4種類の異なる手法によって得られた体の組織・臓器の遺伝子発現データを並列に表示し遺伝子発現解析を行う上で基準となるリファレンスとして利用することを 目的とした遺伝子発現データベース。ヒト, マウス, ラットが対象

- IDキーワード検索の他, 発現パターン, 染色体領域, 遺伝子ファミリ, 機能で検索ブラウズできる
- GGRNA, BodyMapからのリンク

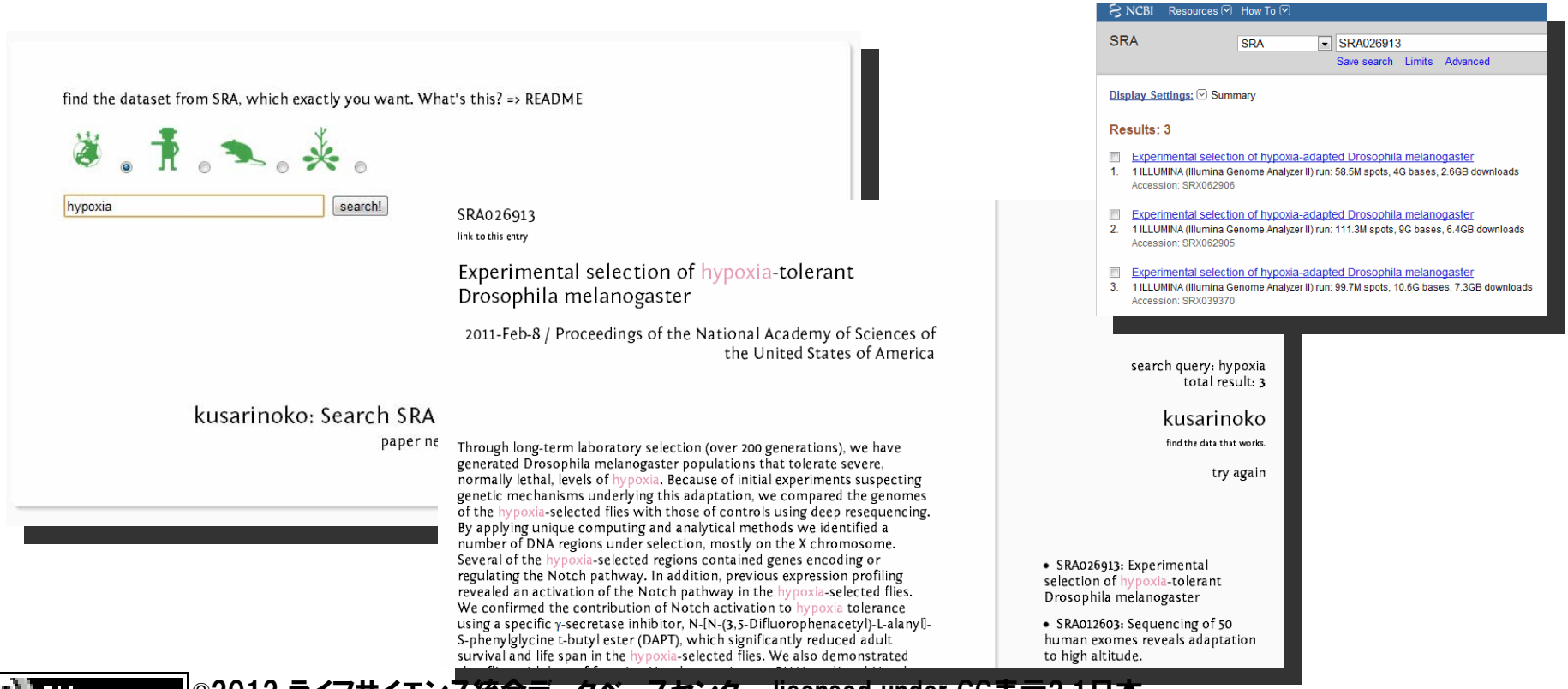






QC値の良いNGSデータの検索：鎖鋸

- 公共データベース （SRA, ENA, DRA）に登録された次世代シーケンサーデータのうち論文で公表され信頼性が確保されたと考えられるデータを収録.
- FastQCを用いて算出した配列データのクオリティーをあらわす統計量付
- メタデータの整理，充実によりユーザーは，キーワード検索を行うとともに，配列データのクオリティー評価を参考に，自分が必要とする信頼性が高いデータを見つけることができる.



今年度研究開発内容
(6)情報統合化・知識発見のためのキュレーション支援

協働キュレーション作業運用技術の整備

1. CSCW技術の調査
   - 協働キュレーション作業に利用できるツールと利用範囲を整理
2. CSCW技術の実際のキュレーション作業への適用
   - 統合化推進プログラムとの連携によるゲノムアノテーション作業へのCSCWシステムの適用
   - DBCLS内部TogoTV作成グループへのTracシステム導入

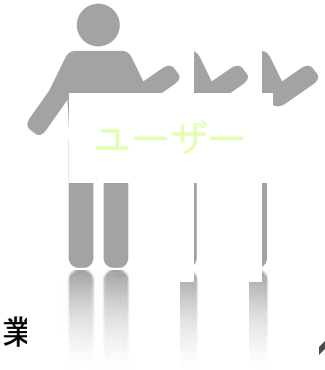
キュレーション支援システムの開発

1. 略語検索システムのRDF化によるDB間連携と再利用性の向上（Allie）
   - RESTインターフェイス，データRDF化完了（SPARQLエンドポイント，FTP，Bioportal，the Data Hubより公開）
2. 科学技術英語表現の添削と推薦技術の開発（inMeXes）
   - nグラムに基づく利用頻度と品詞情報プロトタイピング完了
3. 文献管理推薦システムのマルチプラットホーム化（TogoDoc/Client）
   - iOS版TogoClient開発版ベータテスト中
4. 文献管理推薦システムのユーザカスタマイズ対応とパッケージ化（TogoDoc/Client）
   - ユースケース検討中
5. オルソログクラスターへのNLPによる自動命名手法のための調査（TogoAnnotator）
   - 既存アルゴリズム（LCS+KeyCollision）による命名プログラム試作

(6) 情報統合化・知識発見のためのキュレーション支援

CSCWを利用した
協働キュレーション体制

キュレーション情報
公開データベース



キュレーション作業

システム開発者

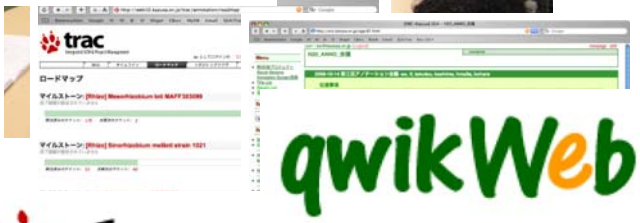
キュレーション情報の統合

協働キュレーション

機能・内容改善要望

遠隔地キュレーター

キュレーション
データベース



キュレーションUI, 情報編集・保存

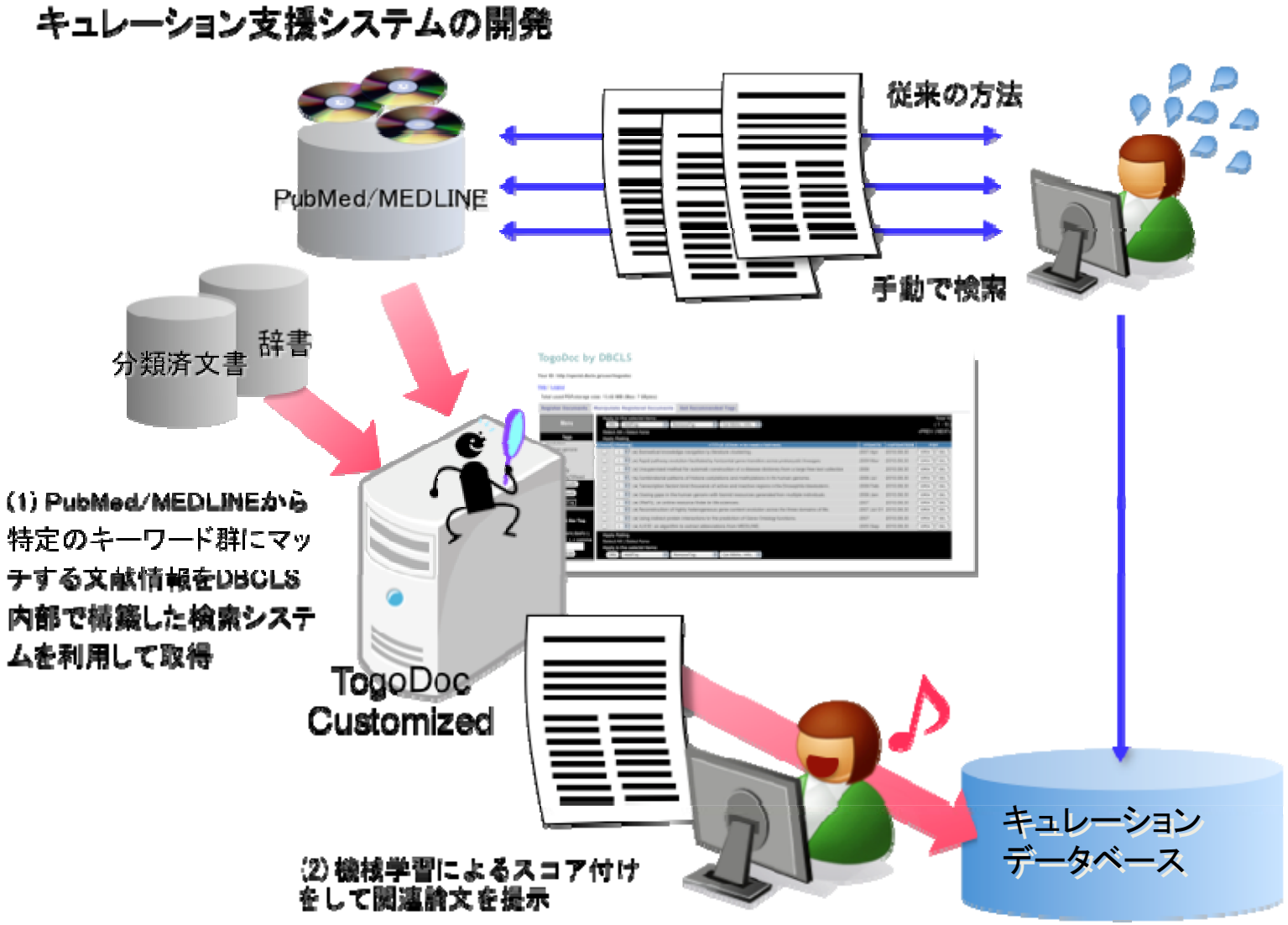
問題発見, 問題共有

qwikWeb

trac
Integrated SCM & Project Management

skype™

(6) 情報統合化・知識発見のためのキュレーション支援

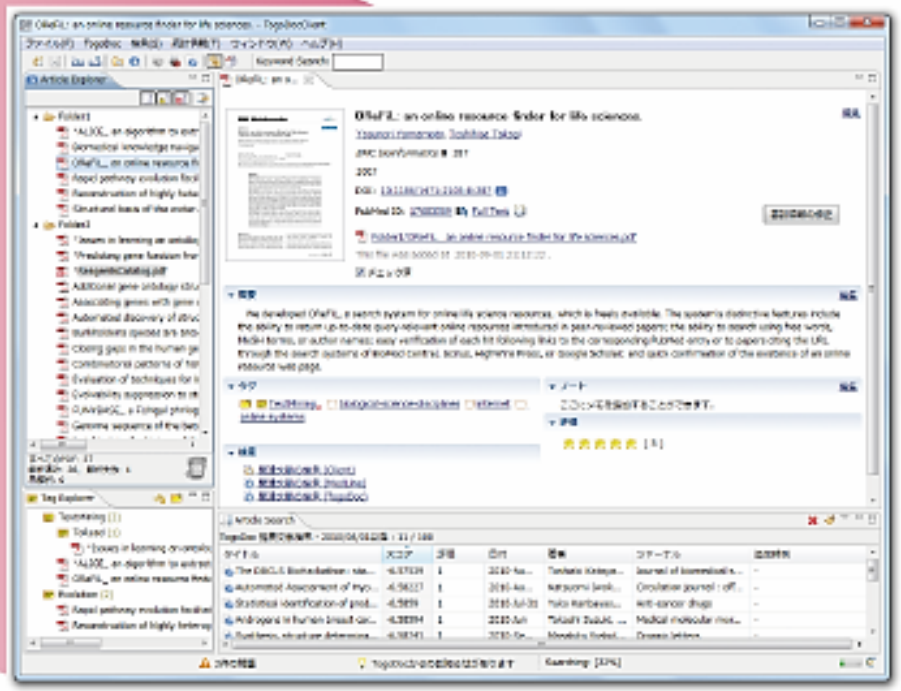


TogoDocClient：ユーザPCにインストール
http://tdc.cb.k.u-tokyo.ac.jp/

論文PDFをダウンロードして
保存するだけの自動文献管理

- 自動解析・書誌情報取得
- 自動ファイル名リネーム
- タグ推薦

解析結果をもとに
最近PubMedに登録された
必読論文を自動推薦



連携

TogoDoc：ウェブブラウザでアクセス
http://docman.dbcls.jp/

どこからでも個人論文ライブラリにアクセス
PC間でのPDFファイルを含む同期

携帯端末からも
必読論文チェック

今年度研究開発内容
(7) 統合データベースに関わるコンテンツの作成，整備

- 動画によるチュートリアル作成 -118本（H23年度4月-2月）
  - 更新動画の維持管理システム整備
  - 音声付与システム開発
- 良質な日本語コンテンツの作成，整備-154本(H23年度4月-2月）
  - 日本語レビューのオンラインジャーナルの立ち上げ準備
- 生物画像コンテンツの作成，整備-BodyParts3D3.0，TogoPictureGallaly公開
  - ヒト3Dマップ（個体レベル）の整備
  - 細胞，分子レベルの画像コンテンツの検討

Fig. 14

DBCLS研究開発サービス一覧

RDF による統合化
TogoDB
TogoDB2
TogoWS

高度検索技術の開発
表形式データへのアノテーションツール
セマンティックアノテーションDB

統合利用環境の整備
DBCLS Galaxy
MiGAP

オントロジー，辞書，コーパス整備
OntoFinder/OntoFactory
アナトモグラフィー/BodyParts3D
生物学名辞書

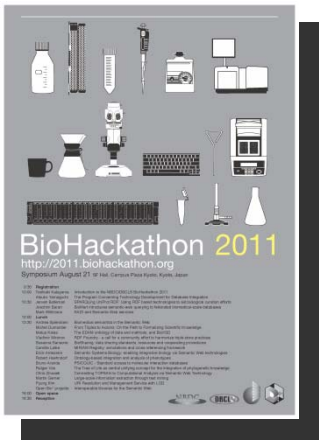
情報統合化・知識発見のための キュレーション支援
Allie
OReFiL
Wired-Marker
inMeXes
TogoDoc

大規模データ利用技術開発
遺伝子発現バンク目次
DNAデータベース総覧と検索
Gendoo
SRAs
Refex
GGRNA
鎖鋸

コンテンツの作成 ， 整備
統合TV
ライフサイエンス 新着論文レビュー
Togo Picture Gallery
医学薬学要旨集

国際会議 NBDC/DBCLS　BioHackathon

BioHackathon = Bio + Hack + Marathon
生物学データをターゲットとした
プログラミングのマラソン
http://2011.biohackathon.org/



第1回 2008
ウェブサービス
標準化

第2回 2009
ワークフロー構築
ウェブサービス
相互運用性

第3回 2010
セマンティック
ウェブによる
知識統合

第4回2011

Linked Data によるデータリソース公開とその周辺技術開発

- 海外　27名(うち24名招聘)
- 国内　68名(ハッカソン参加者 43名)
- NBDC, DBCLS, 京大の共催

（参考）JSBiニュースレター

- 第23号にてBioHackathon2011をレポート

- BioHackathon 2011 開催報告　---　片山俊明
- バイオロジーのデータ格納に適したトリプルストア調査　---　山口敦子
- BioDBCoreにおけるDBメタデータのRDFによる共有　---　山崎千里
- マニュアルキュレーションにおける用語と概念のマッピング　---　岡本忍，山本泰智
- 既存データベースからのLinked Dataの創出　---　川島秀一
- RDFを利用したデータ解析のユースケースの検討　---　荻島創一
- テキストマイニングツールのRDF化と標準化　---　金進東，狩野芳伸
- セマンティックウェブ技術を用いたライフサイエンス系データ利用のための クライアントソフトウェア開発　---　大野圭一朗

http://tinyurl.com/bh11jsbi

統合データベース技術情報交換ワークショップ開催

http://wiki.lifesciencedb.jp/mw/index.php/BH11.11

日程: 2011/11/21 - 25 (5日間)

主催: DBCLS, 共催: DDBJ

場所: ラフォーレ修善寺

概要: 8月に行った BioHackathon の成果を受け, RDF基盤システムおよび統合検索システムの開発を開始した. システムをより有用なものとするために, そこで用いるRDFデータ作成を加速するため, DBのRDF化合宿を行った

NBDC: H23年度「研究加速の取り組み」により実現

NBDC/細胞工学
シリーズ：我が国のデータベース構築・統合戦略

NBDC の広報サイト
バイオサイエンス ×DB＝∞
検索
Web events.biosciencedbc.jp/

Home シンポジウム 講習会 展示会 連載

シリーズ：我が国のデータベース構築・統合戦略

第2回「データベースを統合利用するための基盤としてのセマンティックウェブ技術」

山口敦子
（ライフサイエンス統合データベースセンター）
片山俊明
（東京大学医科学研究所 ヒトゲノム解析センター）

はじめに

ライフサイエンス分野の研究により生み出される多様かつ膨大なデータから必要な情報を効率的に得るためには，ばらばらに構築されているデータベースを統合的に扱うための情報基盤の構築が必要不可欠である．連載第1回「データベースの現状と未来」（http://events.biosciencedbc.jp/article/01）では，データベース統合化のための具体的なステップとして，つぎの3つの段階があげられた．

・第1段階：データベースを網羅的に収集しメタデータを付与すること
・第2段階：それぞれのデータベースにおいてフォーマットと用語の統一を行うこと
・第3段階：複数のデータベースを再構築し使いやすいインターフェイスにまとめあげること

大学共同利用機関法人 情報・システム研究機構 ライフサイエンス統合データベースセンター（DBCLS: Database Center for Life Science, URL: http://dbcls.rois.ac.jp/）では，このうちの第3段階をスムーズに実現することを目標として，現在，セマンティックウェブを利用した第2段階の技術開発を進めている．ここでは，その基幹技術であるRDF（resource description framework）について紹介する．

1. RDFとは

シリーズ：我が国のデータベース構築・統合戦略
- 第3回「植物ゲノムデータベースの統合」
- 第2回「データベースを統合利用するための基盤としてのセマンティックウェブ技術」
- 第1回「データベースの現状と未来」

http://events.biosciencedbc.jp/article/02

まとめ

- データベースのRDFによる統合を進めるための各種調査（有用DBの優先度付調査やRDF蓄積DB調査, オントロジー調査など）を実施し, 標準工程や指針を明らかにした.
- 特定分野のデータベースのRDF化を実施, 必要なオントロジーの作成, RDF化したDBを利用した情報提示法の検討を実施した.
- 大規模ゲノムデータの再利用促進技術開発を目指した検索サービスの開発を進め, 文献情報やQC情報による指標を導入した.
- NLPや協働キュレーション技術により研究者のアノテーション, キュレーション, 文献利用環境の向上を図った.
- BioHackathon等を通じ, NBDC, 統合化推進プロジェクトとの連携し, オントロジ構築やDBのRDF化を実施した.

課題

- RDFを蓄積するトリプルストアの調査並びに整備が難航した.（トリプルストア自身が発展途上）

  →インフラ整備，ストア開発者との共同開発開始

- DBのRDF化やオントロジー構築にかかるコスト（手作業，時間，領域固有の問題）をどう解決するか

  →半自動化のためのツール，サポートツール開発に着手．問題点からの開発へのフィードバック，コストの算出

- 開発規模の決定が難しい（ DBはどこまでRDF化するのか生物種はどこまで増やすのか，対象分野は? ）

  →プロトタイプ開発を通じて今年度の課題とする

- 医療用画像データの利用技術開発の課題設定が困難であった　→分子データ以外のもの、画像等の取り扱いについて検討する

- 節電により計算機の縮退運転を実施し，サービスや開発に影響が出た．今夏の対応，持続的対応が必要

課題別 H24年度 研究計画

(1) データベースのRDFによる統合化

- RDF 化ガイドラインの整備
- 有用 DB のRDF 化作業
- TogoDB の RDF 化機能開発
- 自然文からのSPARQL生成システム開発
- トリプルストアの調査ならびに整備
- LinkDBのRDF化

(2) 解析プラットフォームによる統合利用環境の整備

- ワークフロー環境整備
- ツールのRDF入出力機能追加開発
- 解析プラットフォームへの機能追加・整備

(3) インターネットを活用した高度検索技術の開発

- TogoDB2を利用したDBアーカイブのRDF化
- 有用分野でのプロトタイピング –微生物ゲノム,プロテオーム
- NBDC統合化推進プログラムとの連携-糖鎖との連携

(4) RDF化に資するオントロジー辞書コーパスの整備標準化

- 自然言語処理、テキストマイニング技術の整備
- 言語資源リポジトリ開発
- 反応オントロジーの整備公開

課題別 H24年度 研究計画

(5) 大規模ゲノム配列データの利用技術開

- メタ情報による大規模ゲノム配列データの整理・再利用促進技術開発　SRAからGEO, DDBJに拡大、QC値の利用
- 遺伝子発現のリファレンスデータセット(RefEx) の整備
  癌、細胞株,生物種の拡大
- 医療用画像データの利用技術開発　–公開DBの調査

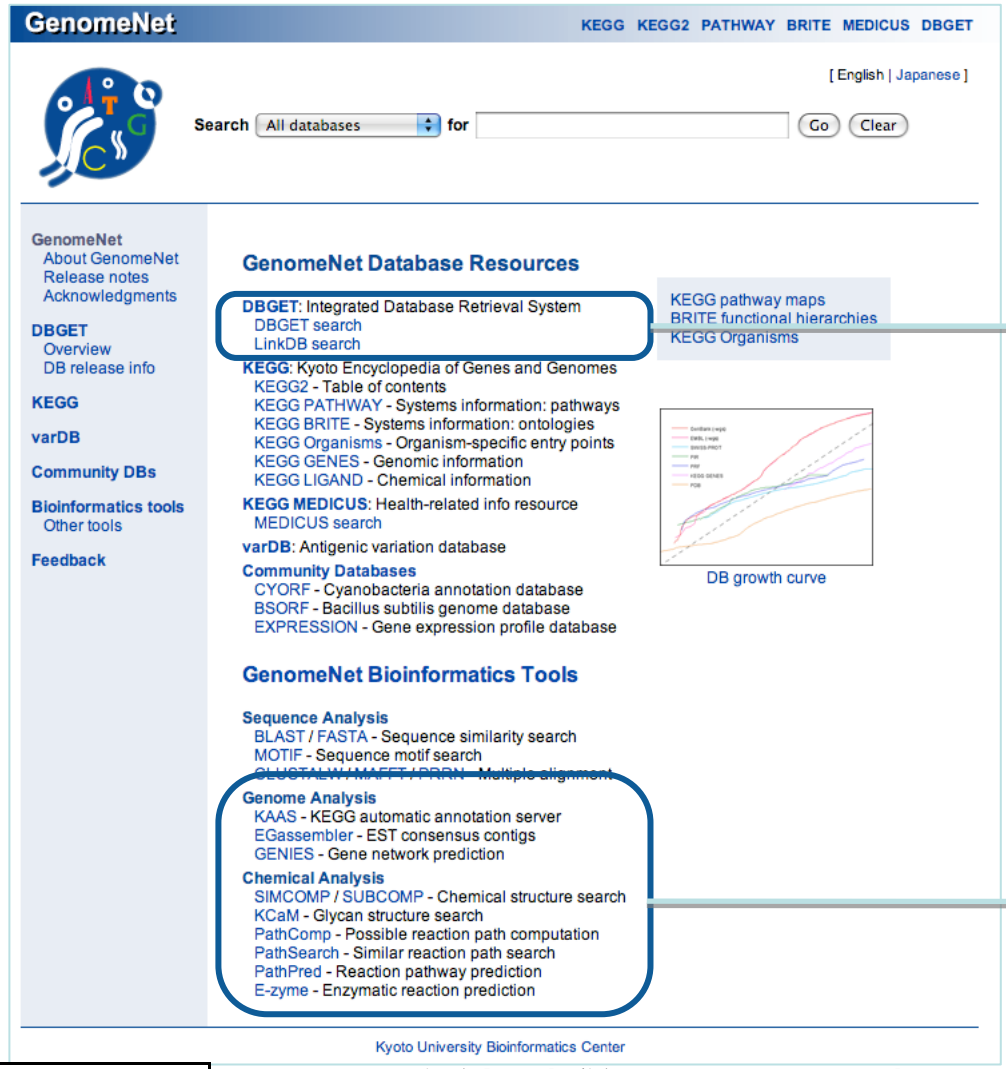
(6) 情報統合化・知識発見のためのキュレーション支援

- キュレーション支援システム開発-他DBとの連携, 出力RDF化
- 協働キュレーション作業運用技術の整備 –統合化推進プログラム連携

(7) 統合データベースに関わるコンテンツの作成, 整備

- チュートリアル動画の作成継続, 音声付与版作成
- 学会オンラインジャーナル立上げ
- 生物並びに臓器3Dデータの作成　男性モデルの完成

京都大学グループ　五斗進（京大化研バイオインフォマティクスセンター）
（1）DBGET/LinkDBシステムの統合利用環境への応用
（2）メタゲノム・メタメタボローム等新規分野データ活用技術の開発
（3）反応オントロジーの整備



ゲノムネット
http://www.genome.jp/

（1）

（2）（3）

京都大学グループ

（1）DBGET/LinkDBシステムの統合利用環境への応用

（2）メタゲノム・メタメタボローム等新規分野データ活用技術の開発

（3）反応オントロジーの整備

DBGET/LinkDB統合データベース検索システム（170 DBを統合）

カテゴリ	bget	bfind	blink	DB数
1. KEGGデータベース(DBGET版)	yes	yes	yes	22
2. その他のDBGETデータベース	yes	yes	yes	19
3. Web上の検索可能データベース	no	yes	yes	18
4. Web上のリンクのみのデータベース	no	no	yes	110
5. PubMedデータベース	yes	no	yes	1

LinkDB：8億以上のリンク情報
- 順引きリンク：データベース中に記述されているリンク
- 逆引きリンク：他のデータベースから参照されているリンク
- 等価リンク：データベース間で同じ化合物・遺伝子の関係を定義したリンク

blink: データベース間のリンク情報

bget: エントリ取得、bfind: キーワード検索

- LinkDBのRDF化
  - 化合物データベースの等価リンクから始めて他のデータベースに拡張する
- RDF化したLinkDBの等価リンク情報を横断検索に応用する
  - ヒットしたエントリのうち同じ化合物・遺伝子の情報をまとめるなど
- 共通で使われているIDなどを利用した等価リンクの自動更新化を検討
- ゲノムネット計算サービスのAPI化
  - ホモロジー検索、モチーフ検索との統合

LinkDBのRDF化

- 化合物データベースの等価リンクから始めて他のデータベースに拡張する
- Turtle での表現

```
@prefix 3dmet: <http://www.3dmet.dna.affrc.go.jp/bin2/show_data.e?acc=> .
@prefix chebi: <http://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:> .
@prefix cpd: <http://kegg.jp/entry/> .
@prefix hmdb: <http://www.hmdb.ca/metabolites/> .
@prefix knapsack: <http://kanaya.naist.jp/knapsack_jsp/information.jsp?sname=C_ID&word=> .
@prefix linkdb: <http://www.genome.jp/linkdb/> .
@prefix nikkaji: <http://nikkajiweb.jst.go.jp/nikkaji_web/pages/top_e.jsp?CONTENT=syosai&SN=> .
@prefix pubchem: <http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?sid=> .
cpd:C00002            linkdb:equivalent     3dmet:B01125          .
cpd:C00002            linkdb:equivalent     chebi:15422           .
cpd:C00002            linkdb:equivalent     hmdb:HMDB00538        .
cpd:C00002            linkdb:equivalent     knapsack:C00001491    .
cpd:C00002            linkdb:equivalent     nikkaji:J10.680A      .
cpd:C00002            linkdb:equivalent     pdb-ccd:ATP           .
cpd:C00002            linkdb:equivalent     pubchem:3304          .
```

From DB entry　　リンクの種類　　To DB entry

equivalent, original, reverse

@prefix に URL を記述：LinkDB でサポートする 170 のデータベースについて定義

実際には From と To に URL を組み込んだ形式

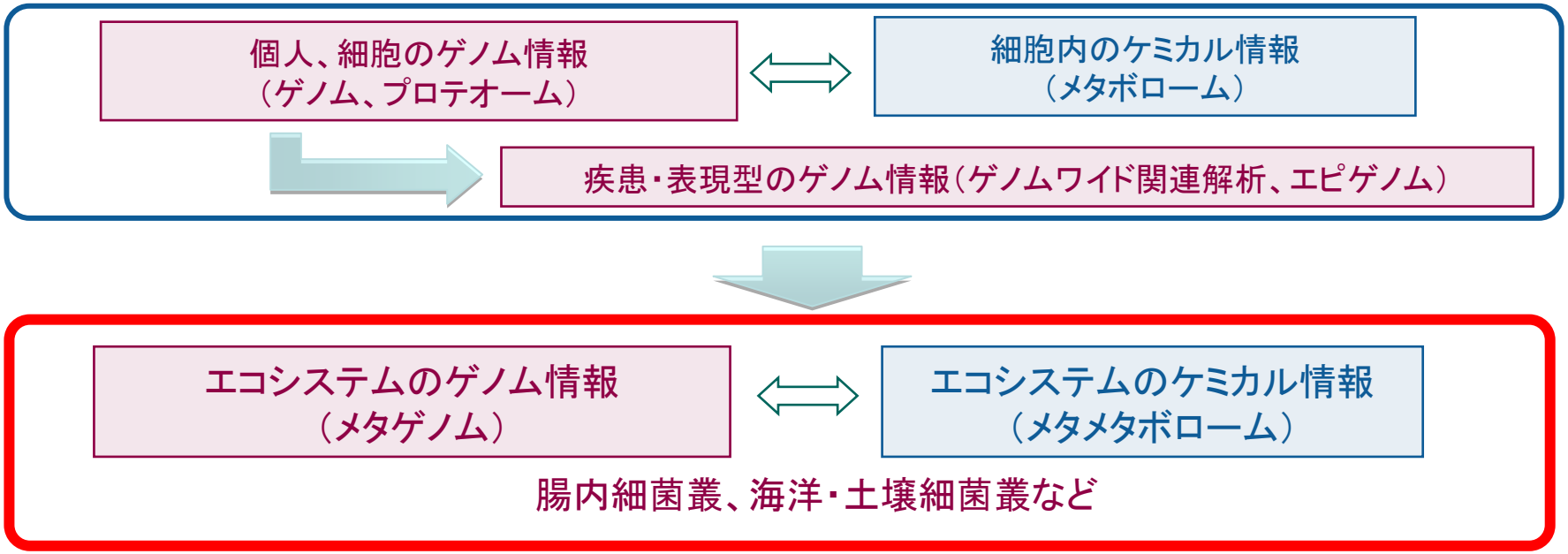
Web 上の LinkDB 検索結果からダウンロード可能にする予定

京都大学グループ

（1）DBGET/LinkDBシステムの統合利用環境への応用

（2）メタゲノム・メタメタボローム等新規分野データ活用技術の開発

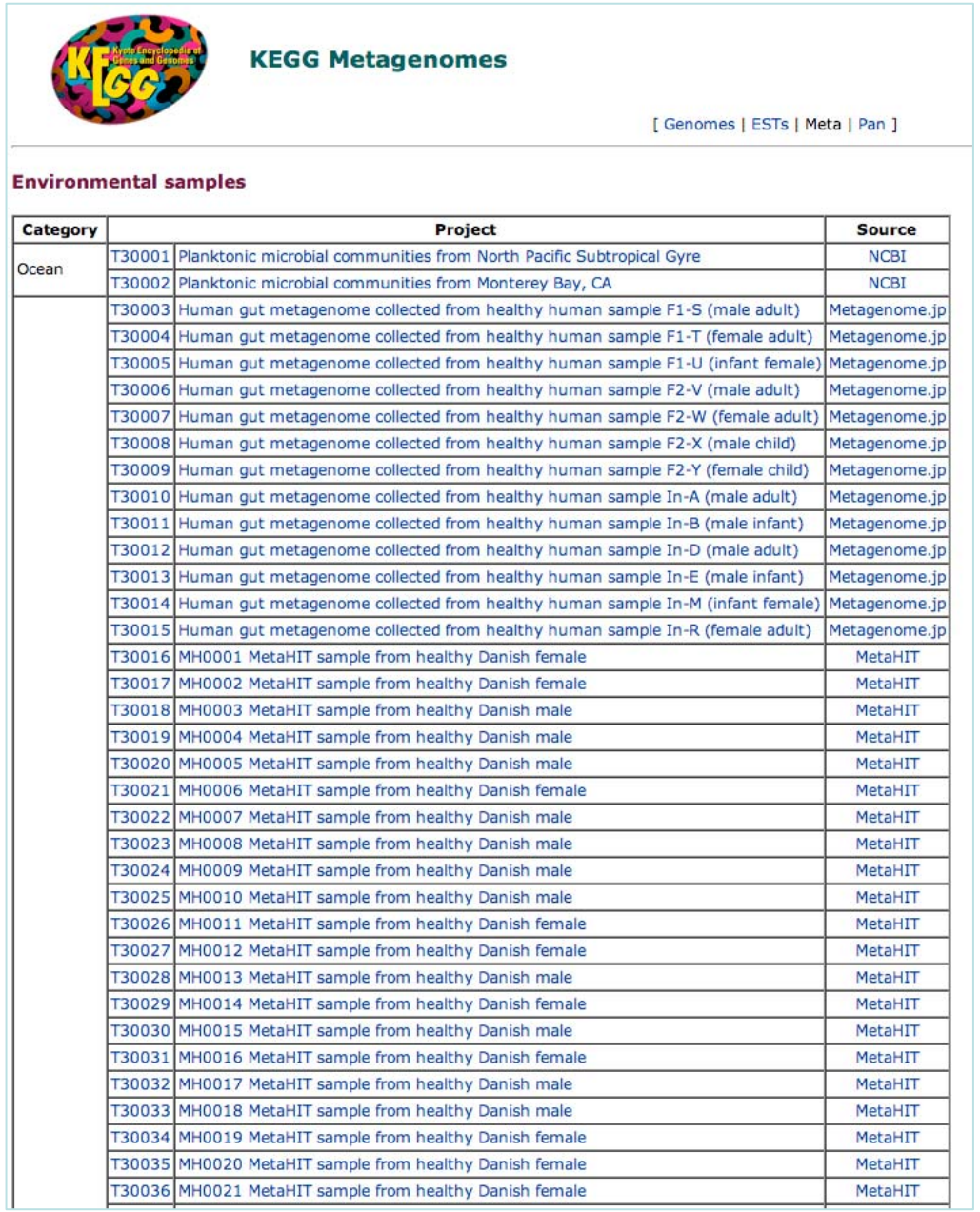
（3）反応オントロジーの整備



- メタゲノムデータ、メタメタボロームデータ利用技術開発
  - データベース化支援技術：機能アノテーション、パスウェイ再構築
- メタゲノムとメタメタボロームデータの統合化に関わる要素技術開発
  - 反応データによるゲノムとメタボロームの関連付け技術
  - 新規パスウェイ予測のための技術
- 反応オントロジーの整備
  - 遺伝子と反応タイプの関連付けによる反応分類システムの開発

メタゲノムデータ、メタメタボロームデータ利用技術開発

- データベース化支援技術：機能アノテーション、パスウェイ再構築
- KAAS: KEGG Automatic Annotation Server を応用
- 海洋メタゲノム
- ヒト腸内細菌メタゲノム
- アミノ酸配列レベルでのアノテーション



KEGG Metagenomes

[Genomes | ESTs | Meta | Pan]

Environmental samples

Category	Project		Source
Ocean	T30001	Planktonic microbial communities from North Pacific Subtropical Gyre	NCBI
	T30002	Planktonic microbial communities from Monterey Bay, CA	NCBI
	T30003	Human gut metagenome collected from healthy human sample F1-S (male adult)	Metagenome.jp
	T30004	Human gut metagenome collected from healthy human sample F1-T (female adult)	Metagenome.jp
	T30005	Human gut metagenome collected from healthy human sample F1-U (infant female)	Metagenome.jp
	T30006	Human gut metagenome collected from healthy human sample F2-V (male adult)	Metagenome.jp
	T30007	Human gut metagenome collected from healthy human sample F2-W (female adult)	Metagenome.jp
	T30008	Human gut metagenome collected from healthy human sample F2-X (male child)	Metagenome.jp
	T30009	Human gut metagenome collected from healthy human sample F2-Y (female child)	Metagenome.jp
	T30010	Human gut metagenome collected from healthy human sample In-A (male adult)	Metagenome.jp
	T30011	Human gut metagenome collected from healthy human sample In-B (male infant)	Metagenome.jp
	T30012	Human gut metagenome collected from healthy human sample In-D (male adult)	Metagenome.jp
	T30013	Human gut metagenome collected from healthy human sample In-E (male infant)	Metagenome.jp
	T30014	Human gut metagenome collected from healthy human sample In-M (infant female)	Metagenome.jp
	T30015	Human gut metagenome collected from healthy human sample In-R (female adult)	Metagenome.jp
	T30016	MH0001 MetaHIT sample from healthy Danish female	MetaHIT
	T30017	MH0002 MetaHIT sample from healthy Danish female	MetaHIT
	T30018	MH0003 MetaHIT sample from healthy Danish male	MetaHIT
	T30019	MH0004 MetaHIT sample from healthy Danish male	MetaHIT
	T30020	MH0005 MetaHIT sample from healthy Danish male	MetaHIT
	T30021	MH0006 MetaHIT sample from healthy Danish female	MetaHIT
	T30022	MH0007 MetaHIT sample from healthy Danish male	MetaHIT
	T30023	MH0008 MetaHIT sample from healthy Danish male	MetaHIT
	T30024	MH0009 MetaHIT sample from healthy Danish male	MetaHIT
	T30025	MH0010 MetaHIT sample from healthy Danish male	MetaHIT
	T30026	MH0011 MetaHIT sample from healthy Danish female	MetaHIT
	T30027	MH0012 MetaHIT sample from healthy Danish female	MetaHIT
	T30028	MH0013 MetaHIT sample from healthy Danish male	MetaHIT
	T30029	MH0014 MetaHIT sample from healthy Danish female	MetaHIT
	T30030	MH0015 MetaHIT sample from healthy Danish male	MetaHIT
	T30031	MH0016 MetaHIT sample from healthy Danish female	MetaHIT
	T30032	MH0017 MetaHIT sample from healthy Danish male	MetaHIT
	T30033	MH0018 MetaHIT sample from healthy Danish male	MetaHIT
	T30034	MH0019 MetaHIT sample from healthy Danish female	MetaHIT
	T30035	MH0020 MetaHIT sample from healthy Danish female	MetaHIT
	T30036	MH0021 MetaHIT sample from healthy Danish female	MetaHIT

メタゲノムデータ、メタメタボロームデータ利用技術開発

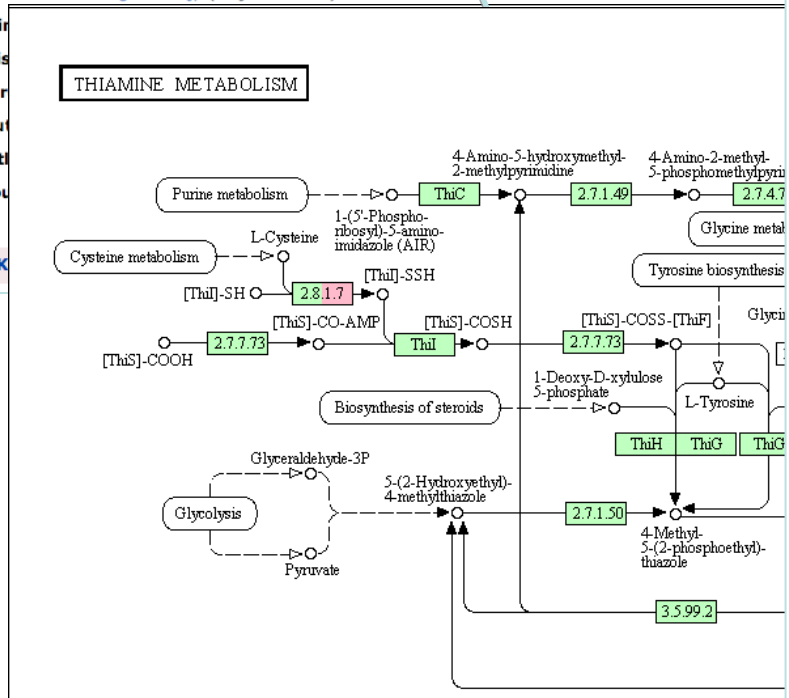
KEGG **Metagenome: T30003**

Genome info | Pathway map | Brite hierarchy

Search genes: Go Clear

Genome information

T number	T30003
Definition	Human gut metagenome collected from healthy human sample F1-S (male adult)
Annotation	KAAS
Taxonomy	TAX: 408170
Lineage	unclassified sequences; metagenomes; organismal metagenomes
Data source	Metagenome.jp (Project:27855)
Origi	
Statis	
Refer	
Aut	
Titl	
Jou	



KEGG **KEGG Orthology (KO) + T30003**

[Brite menu | Organism menu | Download htext]

Reference hierarchy (KO) Go

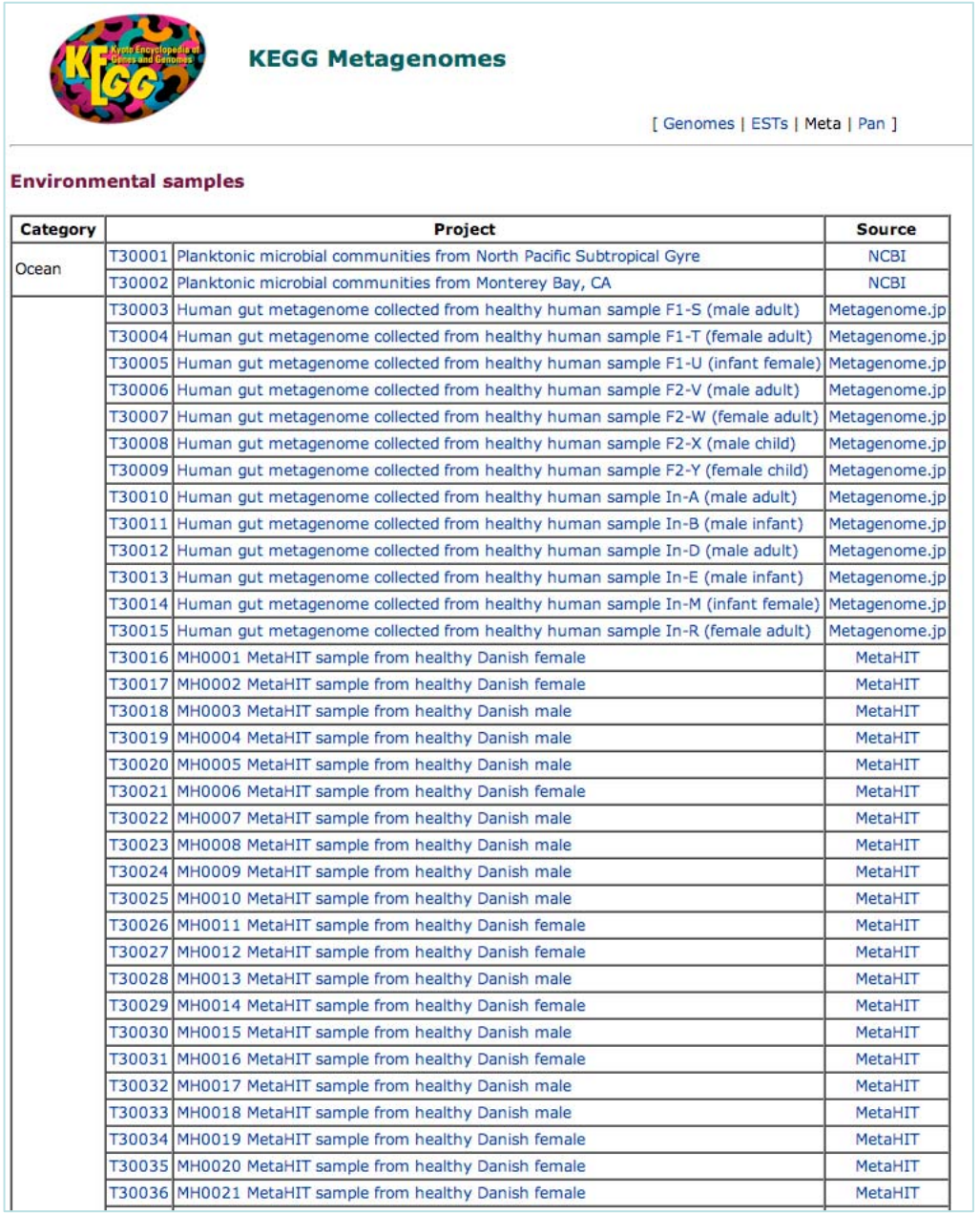
One-click mode

- **Metabolism**
  - **Carbohydrate Metabolism**
  - **Energy Metabolism**
  - **Lipid Metabolism**
  - **Nucleotide Metabolism**
  - **Amino Acid Metabolism**
  - **Metabolism of Other Amino Acids**
  - **Glycan Biosynthesis and Metabolism**
  - **Metabolism of Cofactors and Vitamins**
    - 00730 Thiamine metabolism [PATH:00730]
      - K03147 thiC; thiamine biosynthesis protein ThiC
      - K00877 THI20; hydroxymethylpyrimidine/phosphomethylpyrimidine kinase [EC:2.7.1.49 2.7.4.7]
      - K00941 thiD; hydroxymethylpyrimidine/phosphomethylpyrimidine kinase [EC:2.7.1.49 2.7.4.7]
      - K14153 thiDE; hydroxymethylpyrimidine kinase / phosphomethylpyrimidine kinase / thiamine-phosph
      - K03148 thiF; sulfur carrier protein ThiS adenylyltransferase [EC:2.7.7.73]
      - K04487 iscS, NFS1; cysteine desulfurase [EC:2.8.1.7]
      - K11717 sufS; cysteine desulfurase / selenocysteine lyase [EC:2.8.1.7 4.4.1.16]
      - K03151 thiI; thiamine biosynthesis protein ThiI
      - K03149 thiG; thiamine biosynthesis ThiG
      - K03150 thiH; thiamine biosynthesis ThiH
      - K03153 thiO; glycine oxidase [EC:1.4.3.19]
      - K00878 thiM; hydroxyethylthiazole kinase [EC:2.7.1.50]
      - K00788 thiE; thiamine-phosphate pyrophosphorylase [EC:2.5.1.3]
      - K14154 THI6; thiamine-phosphate diphosphorylase / hydroxyethylthiazole kinase [EC:2.5.1.3 2.7.1
      - K00946 thiL; thiamine-monophosphate kinase [EC:2.7.4.16]
      - K06928 K06928; nucleoside-triphosphatase THEP1 [EC:3.6.1.15]
      - K07251 thiK; thiamine kinase [EC:2.7.1.89]
      - K00949 E2.7.6.2, THI80; thiamine pyrophosphokinase [EC:2.7.6.2]
      - K05307 THTPA; thiamine-triphosphatase [EC:3.6.1.28]
      - K03707 tenA; thiaminase (transcriptional activator TenA) [EC:3.5.99.2]
      - K10811 E2.5.1.2; thiamine pyridinylase [EC:2.5.1.2]
    - 00740 Riboflavin metabolism [PATH:00740]
    - 00750 Vitamin B6 metabolism [PATH:00750]
    - 00760 Nicotinate and nicotinamide metabolism [PATH:00760]
    - 00770 Pantothenate and CoA biosynthesis [PATH:00770]
    - 00780 Biotin metabolism [PATH:00780]
    - 00785 Lipoic acid metabolism [PATH:00785]
    - 00790 Folate biosynthesis [PATH:00790]
    - 00670 One carbon pool by folate [PATH:00670]
    - 00830 Retinol metabolism [PATH:00830]
    - 00860 Porphyrin and chlorophyll metabolism [PATH:00860]
    - 00130 Ubiquinone and other terpenoid-quinone biosynthesis [PATH:00130]

メタゲノムデータ、メタメタボロームデータ利用技術開発

- データベース化支援技術：機能アノテーション、パスウェイ再構築
- KAAS: KEGG Automatic Annotation Server を応用
- 海洋メタゲノム
- ヒト腸内細菌メタゲノム
- アミノ酸配列レベルでのアノテーション

- 検討中＆来年度
  - Short read などへの対応
  - MODULE を用いたアノテーションの評価



KEGG Metagenomes

[Genomes | ESTs | Meta | Pan]

Environmental samples

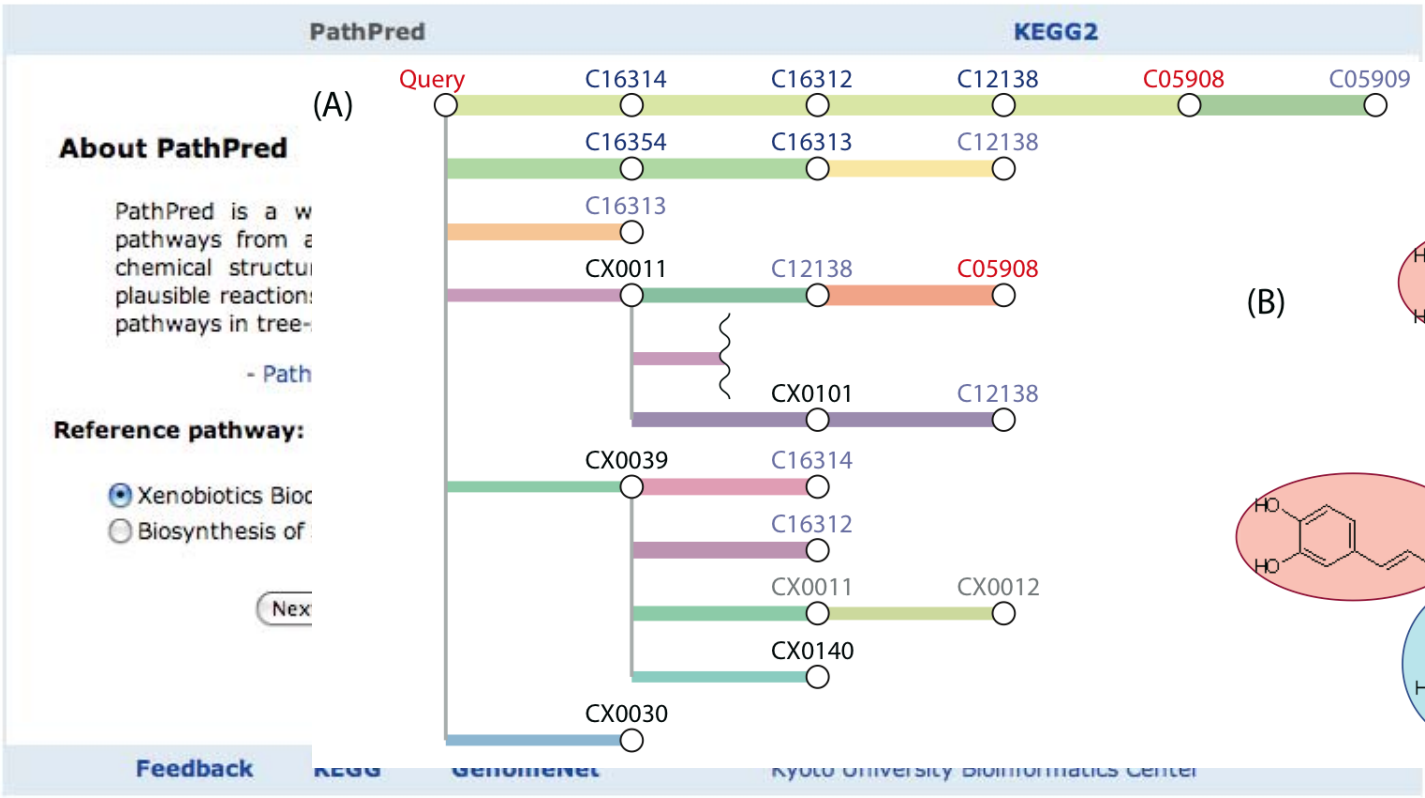
Category	Project		Source
Ocean	T30001	Planktonic microbial communities from North Pacific Subtropical Gyre	NCBI
	T30002	Planktonic microbial communities from Monterey Bay, CA	NCBI
	T30003	Human gut metagenome collected from healthy human sample F1-S (male adult)	Metagenome.jp
	T30004	Human gut metagenome collected from healthy human sample F1-T (female adult)	Metagenome.jp
	T30005	Human gut metagenome collected from healthy human sample F1-U (infant female)	Metagenome.jp
	T30006	Human gut metagenome collected from healthy human sample F2-V (male adult)	Metagenome.jp
	T30007	Human gut metagenome collected from healthy human sample F2-W (female adult)	Metagenome.jp
	T30008	Human gut metagenome collected from healthy human sample F2-X (male child)	Metagenome.jp
	T30009	Human gut metagenome collected from healthy human sample F2-Y (female child)	Metagenome.jp
	T30010	Human gut metagenome collected from healthy human sample In-A (male adult)	Metagenome.jp
	T30011	Human gut metagenome collected from healthy human sample In-B (male infant)	Metagenome.jp
	T30012	Human gut metagenome collected from healthy human sample In-D (male adult)	Metagenome.jp
	T30013	Human gut metagenome collected from healthy human sample In-E (male infant)	Metagenome.jp
	T30014	Human gut metagenome collected from healthy human sample In-M (infant female)	Metagenome.jp
	T30015	Human gut metagenome collected from healthy human sample In-R (female adult)	Metagenome.jp
	T30016	MH0001 MetaHIT sample from healthy Danish female	MetaHIT
	T30017	MH0002 MetaHIT sample from healthy Danish female	MetaHIT
	T30018	MH0003 MetaHIT sample from healthy Danish male	MetaHIT
	T30019	MH0004 MetaHIT sample from healthy Danish male	MetaHIT
	T30020	MH0005 MetaHIT sample from healthy Danish male	MetaHIT
	T30021	MH0006 MetaHIT sample from healthy Danish female	MetaHIT
	T30022	MH0007 MetaHIT sample from healthy Danish male	MetaHIT
	T30023	MH0008 MetaHIT sample from healthy Danish male	MetaHIT
	T30024	MH0009 MetaHIT sample from healthy Danish male	MetaHIT
	T30025	MH0010 MetaHIT sample from healthy Danish male	MetaHIT
	T30026	MH0011 MetaHIT sample from healthy Danish female	MetaHIT
	T30027	MH0012 MetaHIT sample from healthy Danish female	MetaHIT
	T30028	MH0013 MetaHIT sample from healthy Danish male	MetaHIT
	T30029	MH0014 MetaHIT sample from healthy Danish female	MetaHIT
	T30030	MH0015 MetaHIT sample from healthy Danish male	MetaHIT
	T30031	MH0016 MetaHIT sample from healthy Danish female	MetaHIT
	T30032	MH0017 MetaHIT sample from healthy Danish male	MetaHIT
	T30033	MH0018 MetaHIT sample from healthy Danish male	MetaHIT
	T30034	MH0019 MetaHIT sample from healthy Danish female	MetaHIT
	T30035	MH0020 MetaHIT sample from healthy Danish female	MetaHIT
	T30036	MH0021 MetaHIT sample from healthy Danish female	MetaHIT

メタゲノムとメタメタボロームデータの統合化に関わる要素技術開発

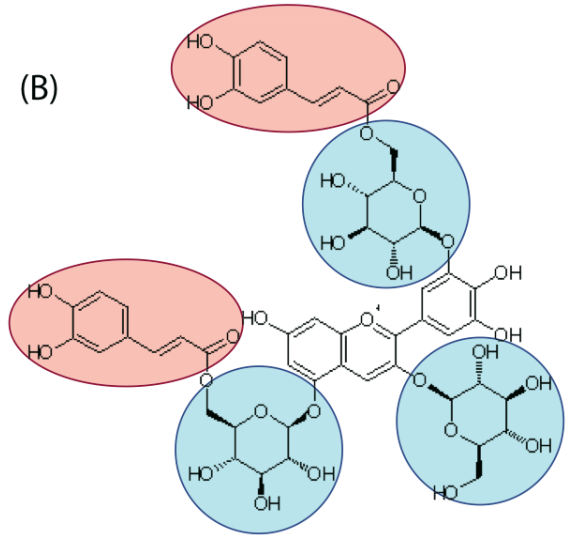
- 反応データによるゲノムとメタボロームの関連付け技術
- 新規パスウェイ予測のための技術

PathPred: Pathway Prediction server

http://www.genome.jp/tools/pathpred/



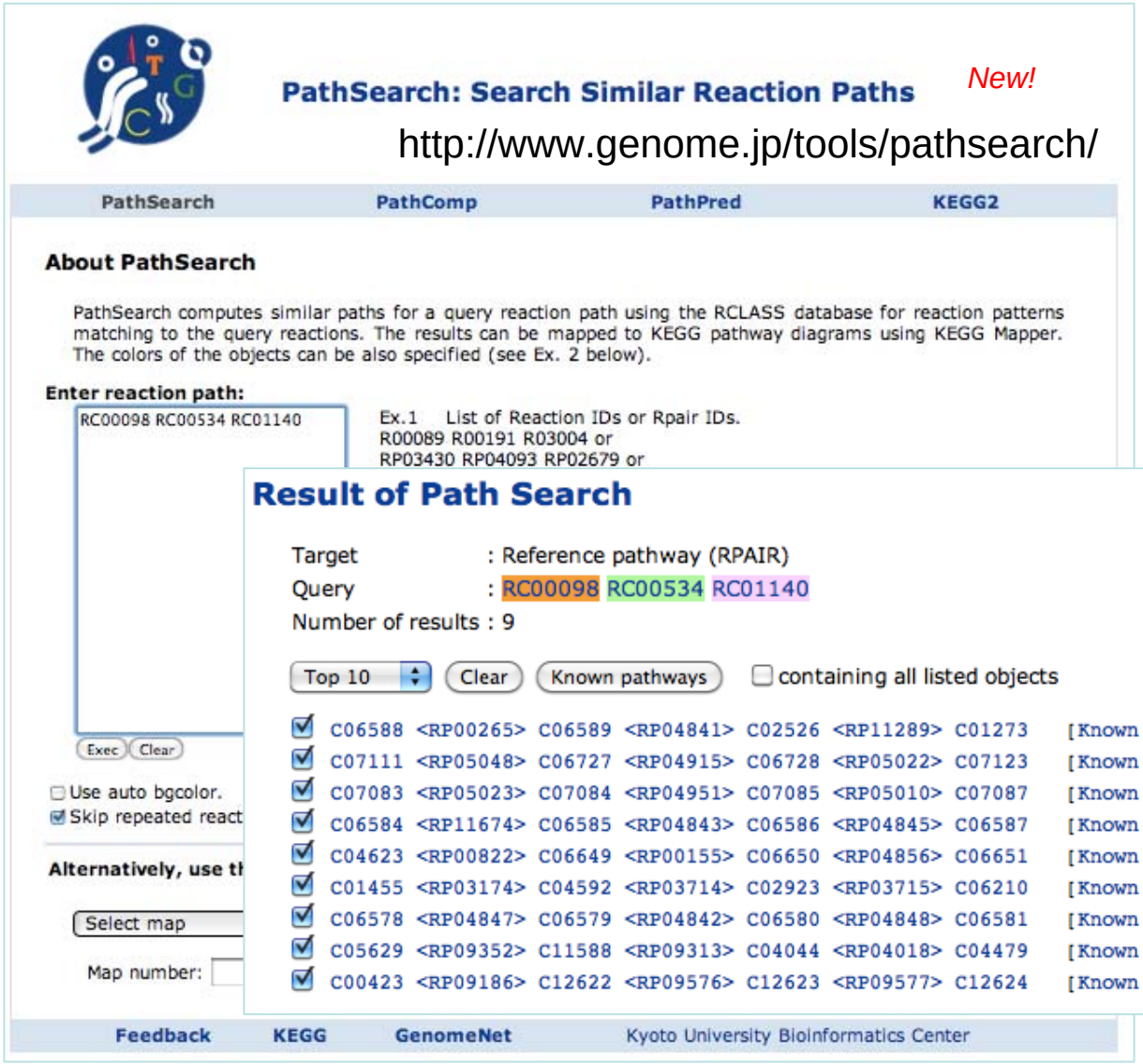
(B)



Moriya, Y., et al. *Nucleic Acids Res*, 38, W138-W143 (2010)

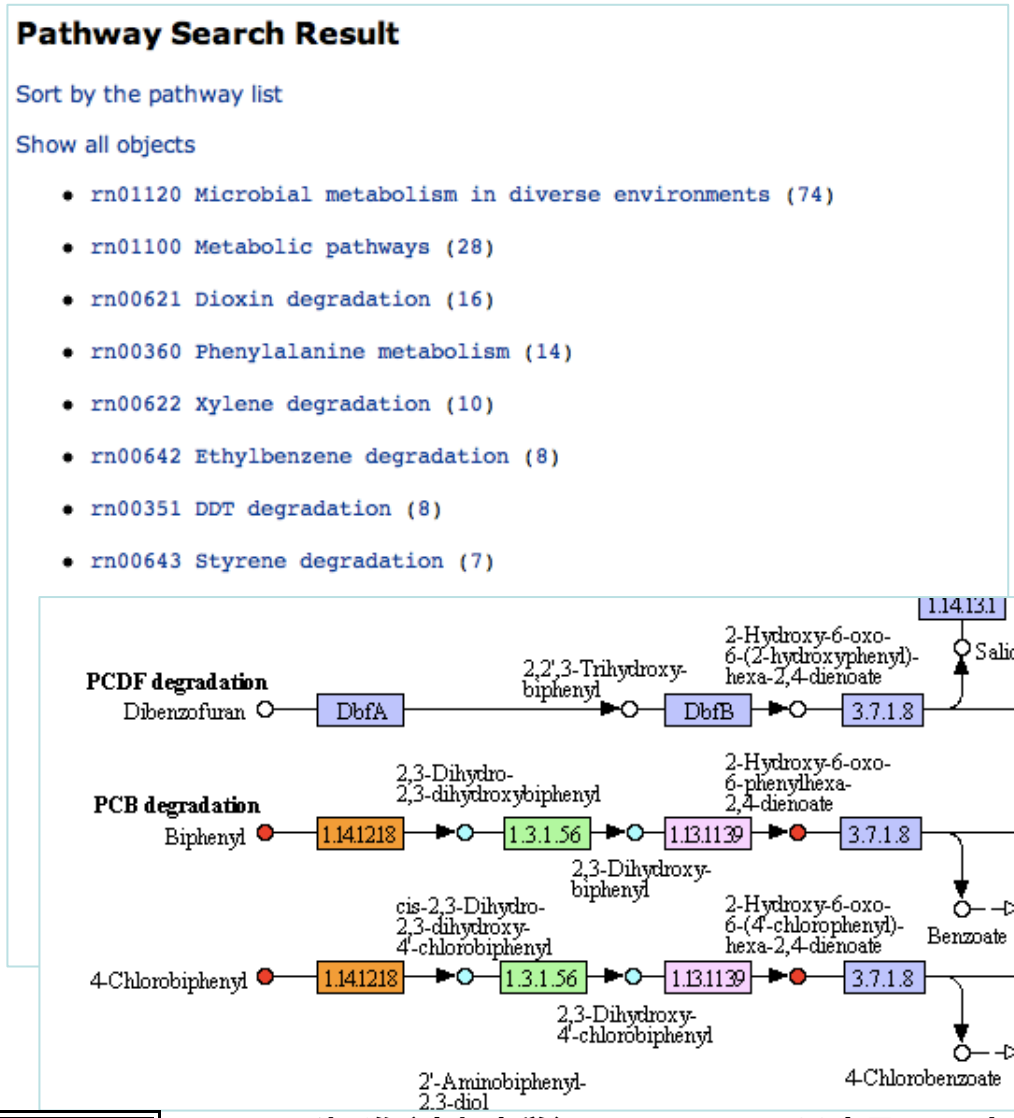
メタゲノムとメタメタボロームデータの統合化に関わる要素技術開発

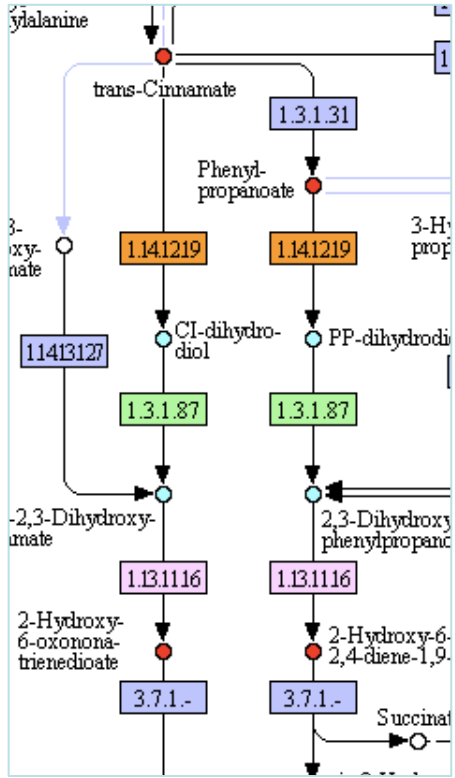
- 反応データによるゲノムとメタボロームの関連付け技術
- 新規パスウェイ予測のための技術



メタゲノムとメタメタボロームデータの統合化に関わる要素技術開発

- 反応データによるゲノムとメタボロームの関連付け技術
- 新規パスウェイ予測のための技術





京都大学グループ

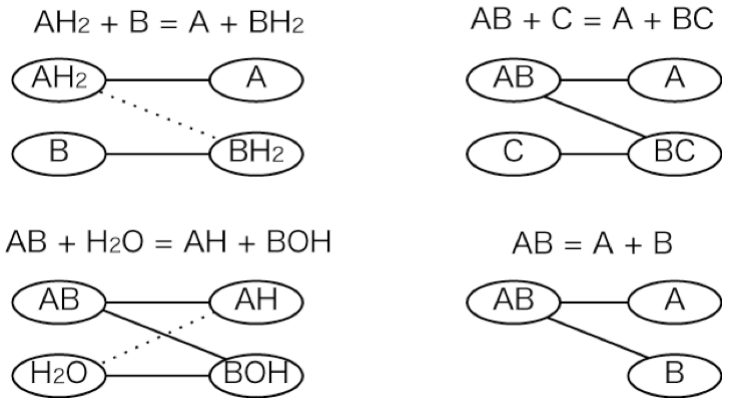
（1）DBGET/LinkDBシステムの統合利用環境への応用

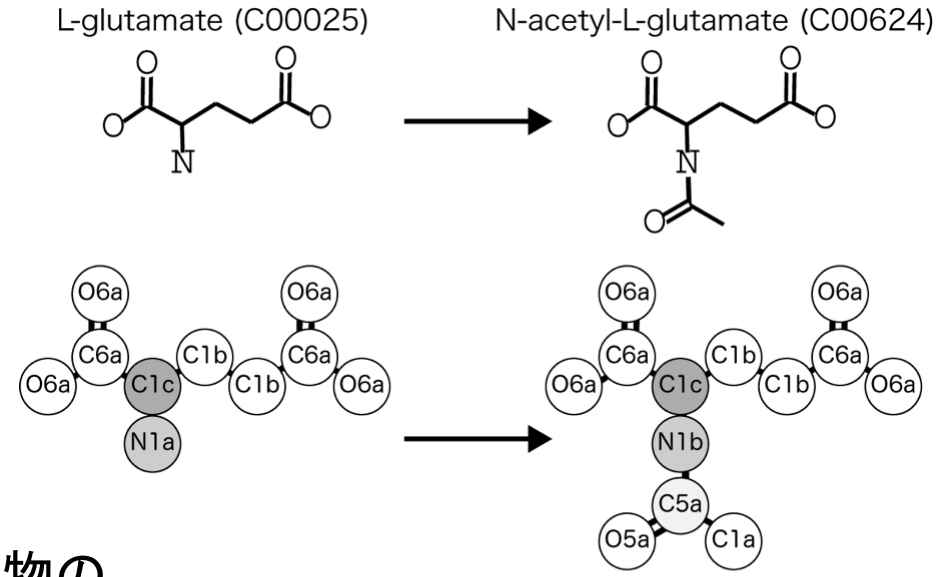
（2）メタゲノム・メタメタボローム等新規分野データ活用技術の開発

（3）反応オントロジーの整備

遺伝子と反応タイプの関連付けによる反応分類システムの開発

反応クラスデータベース RCLASS の構築と整備





反応のタイプごとに基質と生成物のペアを定義して、構造アライメントによって対応する原子を同定

R Reaction Center: N1a -> N1b

D Difference Atom: (H) -> C5a

M Matched Atom: C1c -> C1c

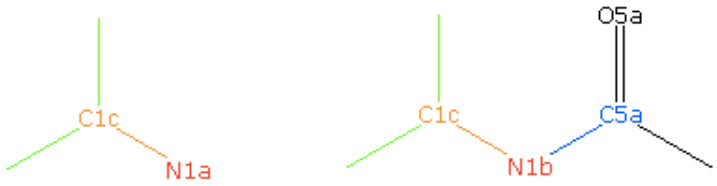
RDM(C00025, C00624) = N1a-N1b : *-C5a : C1c-C1c

反応クラスデータベース RCLASS の構築と整備

KEGG

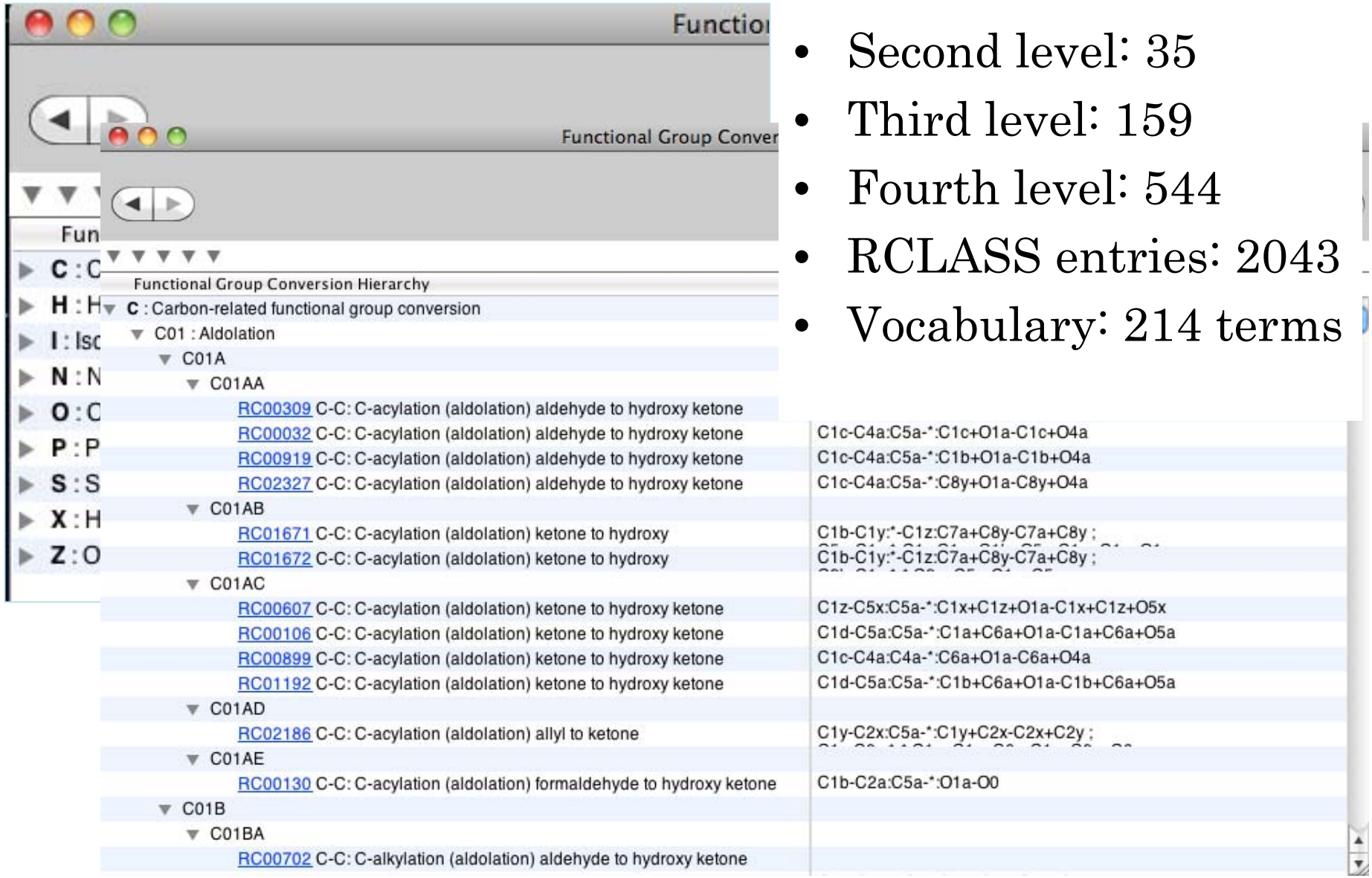
RCLASS: RC00064

Help

Entry	RC00064 RClass
Definition	N1a-N1b:*-C5a:C1c-C1c
	
RPair	RP00167 RP00245 RP00258 RP00269 RP00279 RP00335 RP00423 RP00425 RP00428 RP00431 RP00508 RP00611 RP00724 RP00824 RP00863 RP00864 RP00881 RP00975 RP00991 RP01042 RP01057 RP01071 RP01130 RP01172 RP01186 RP01300 RP01330 RP01331 RP01338 RP01391 RP01607 RP01668 RP01931 RP02027 RP02036 RP02037 RP02039 RP02060 RP02199 RP02248 RP02249 RP02291 RP02319 RP02446 RP02448 RP02494 RP02498 RP02500 RP02828 RP03435 RP03478 RP03482 RP03520 RP03533 RP03556 RP03576 RP03587 RP03738 RP03793 RP03875 RP03951 RP03994 RP04010 RP04219 RP04240 RP04262 RP04266 RP04330 RP04409 RP04457 RP04458 RP04539 RP04559 RP04795 RP04802 RP05279 RP05285 RP05371 RP08259 RP09150 RP09214 RP11706 RP12519 RP12572 RP12945 RP13290 RP13773 RP13910 RP13929 RP14043 RP14066 RP14069 RP14076 RP14349 RP14706 RP14708 RP14710 RP14903 RP15014 RP15016 RP15018 RP15020 RP15022 RP15061 RP15223 RP15337 RP15346 RP15387 DB search
Related class	DB search
Reaction	R00234 R00254 R00259 R00411 R00484 R00487 R00488 R00494 R00546 R00577 R00669 R00693 R00735 R00832 R00894 R00910 R00911 R00912 R00942 R01089 R01149 R01150 R01160 R01164 R01166 R01262 R01263 R01397 R01493 R01494 R01496 R01581

反応オントロジーの構築と整備

- First level: 9
- Second level: 35
- Third level: 159
- Fourth level: 544
- RCLASS entries: 2043
- Vocabulary: 214 terms



京都大学グループの平成24年度の計画

（1）DBGET/LinkDBシステムの統合利用環境への応用

間接的なリンクを含め新たなリンク情報を扱えるようにLinkDBを拡張するとともにRDF化する。LinkDBの自動更新への対応は化合物の構造を用いた対応付けを検討する。

（2）メタゲノム・メタメタボローム等新規分野データ活用技術の開発

平成23年度に引き続きデータ収集を進めるとともに、機能アノテーション支援の拡張としてKEGG MODULEを用いた方法を開発し、ウェブサービスとして公開する。さらに、配列以外の情報を用いた機能アノテーション・予測ツールの整備を行い、上記のサービスと連携して利用できるようにする。

反応オントロジーを利用できるように反応経路探索サービスを改良するとともに、上記のゲノムアノテーションサービスとを連携して利用できるようにする。また、本研究課題で開発したサービスのAPI化を検討する。

（3）反応オントロジーの整備

平成２３年度に整備した反応クラスを、反応パターンの種類と反応において変化する官能基の種類に基づき分類し、反応オントロジーの第1版を公開する。

ライフサイエンスデータベース統合推進事業
基盤技術開発プログラム進捗状況報告

産業技術総合研究所　生命情報工学研究センター　（CBRC）

研究開発題目：解析プラットフォームによる統合利用環境の整備

目標:
本グループでは、ライフサイエンス分野においてDBに蓄積された多種多様なデータと解析ツールを連携させ、必用な情報を効率よく入手、活用することを目的とし、高度なインターネット技術を用いた解析プラットフォームの基盤技術開発（ノード化）、利用環境の整備（RDF化）及び解析結果の統合化されたレポート表示機能（可視化）の開発を行います。

福井一彦、田代俊行、浅井潔

2012年2月24日

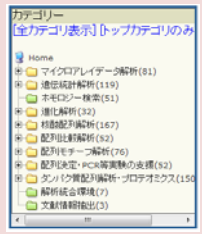
H23年度計画

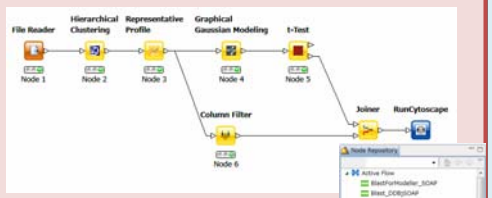
基盤技術開発

① ワークフロー統合 利用環境の構築・整備

分散している既存解析ツールの統合化

CBRC独自解析ツールのノード化





効率性・柔軟性を考慮したプラットフォーム

コンテンツ作成

利用者のためのマニュアル作成

② プラットフォームとDB連携

RDF化

解析プラットフォームとDBに格納されたデータ連携による高度な知的基盤

Semantic Web Serviceとの相互運用性

・解析ツールのRDF化

連携

統合DB

DBCLS Database Center for Life Science

セマンティックWEB
RDF/OWL/SPARQL

③ 解析・検索結果の可視化

解析ナビゲーション

解析結果の可視化及びDB連携ノードやその検索結果を統合した可視化ツール開発

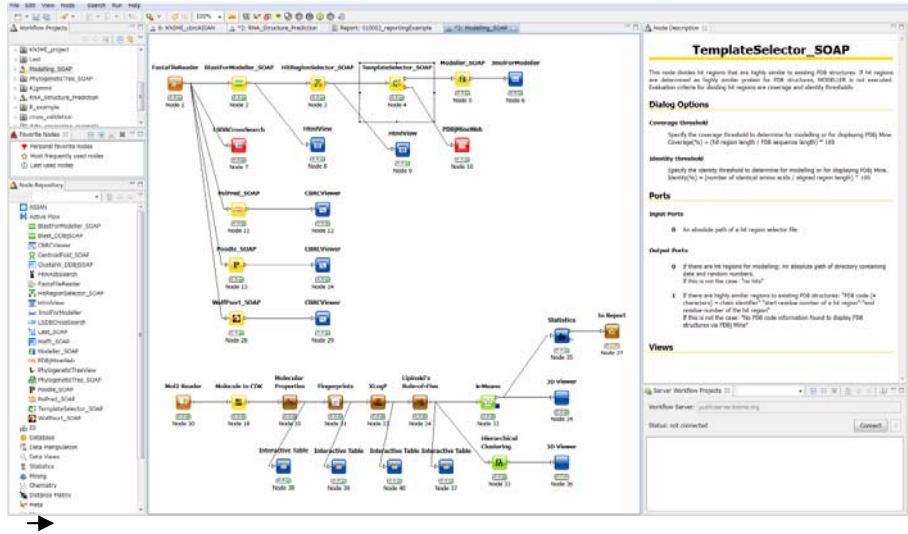
レポーティング作成機能

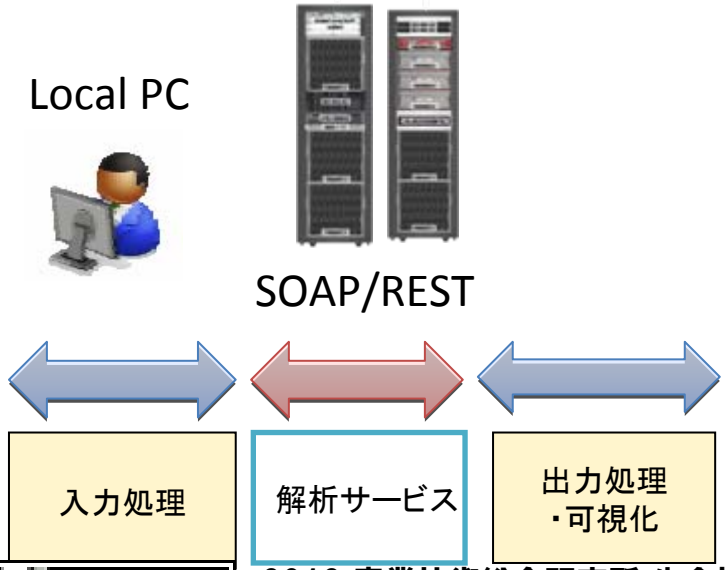
レポーティングの自動化

H23年度　進捗状況

①　ワークフロー統合・利用環境の構築・整備

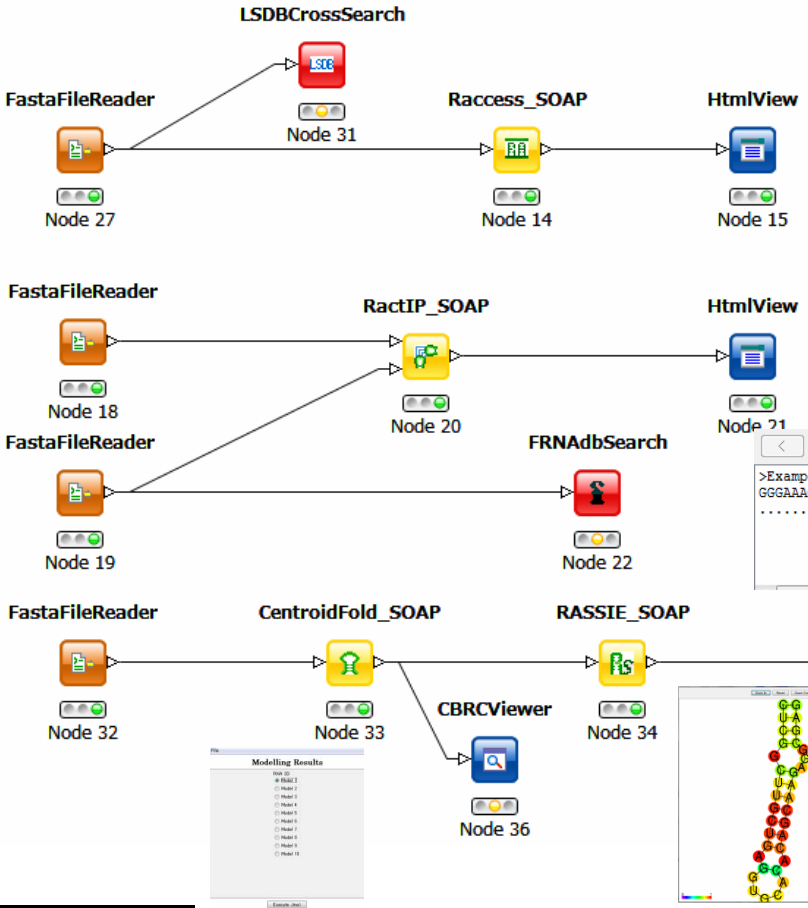
- 本研究グループが独自に開発したツールや有用な既存ツールをKNIMEのプラットフォーム上で動作するように新規にノード化しワークフローとして公開。

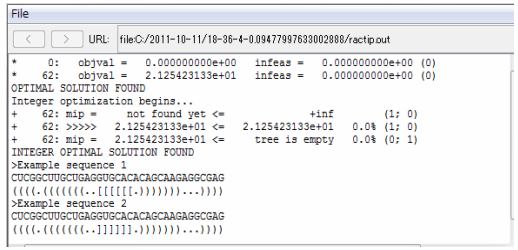


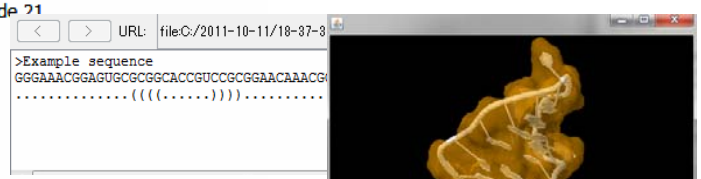


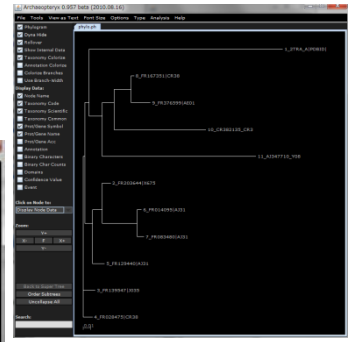
① ワークフロー統合・利用環境の構築・整備

- 新規5つの解析ツールを追加し公開 （SOAP／REST）
- 可視化ノード
- 公開サイトのリニューアル
- 利用者マニュアルの統一化









公開サイト
http://togo.cbrc.jp/active.html#1

H23年度進捗状況

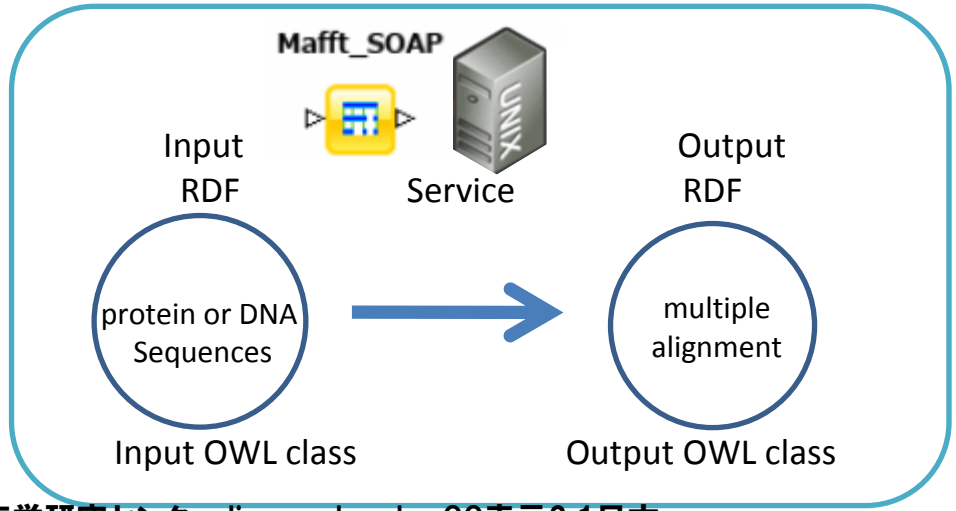
② プラットフォームとDB連携

- DBCLSが進める主要・有用DBのRDF化と連携し、高度な解析ツール群を広く利用可能とするために、各ツールにRDF入出力機能の追加。

SADI (Semantic Automated Discovery and Integration) フレームワークの利用

SADI
Find. Integrate. Analyze.

http://sadiframework.org/

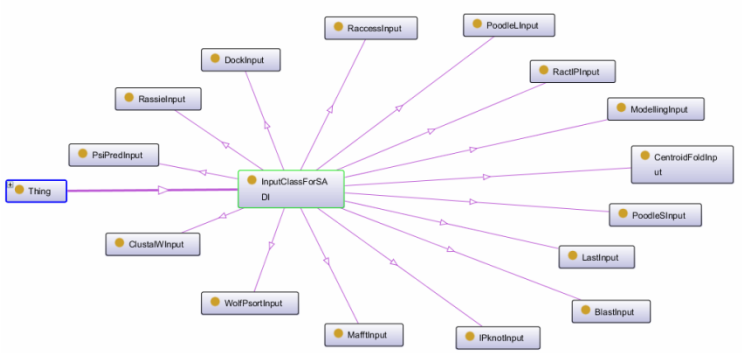


② プラットフォームとDB連携

同期通信
(Sync service)

非同期通信
(Async service)

解析ツール名	入力RDFを定義したOWLクラス
WoLF-PSORT	http://togo.cbrc.jp/ontologies/cbrcswo.owl#WolfPsortInput
CentroidFold	http://togo.cbrc.jp/ontologies/cbrcswo.owl#CentroidFoldInput
IPknot	http://togo.cbrc.jp/ontologies/cbrcswo.owl#IPknotInput
Raccess	http://togo.cbrc.jp/ontologies/cbrcswo.owl#RaccessInput
RactIP	http://togo.cbrc.jp/ontologies/cbrcswo.owl#RactIPInput
PSIPRED	http://togo.cbrc.jp/ontologies/cbrcswo.owl#PsiPredInput
MAFFT	http://togo.cbrc.jp/ontologies/cbrcswo.owl#MafftInput
POODLE-L	http://togo.cbrc.jp/ontologies/cbrcswo.owl#PoodleLInput
POODLE-S	http://togo.cbrc.jp/ontologies/cbrcswo.owl#PoodleSInput



SADI用　OWL作成
http://togo.cbrc.jp/ontologies/cbrcswo.owl

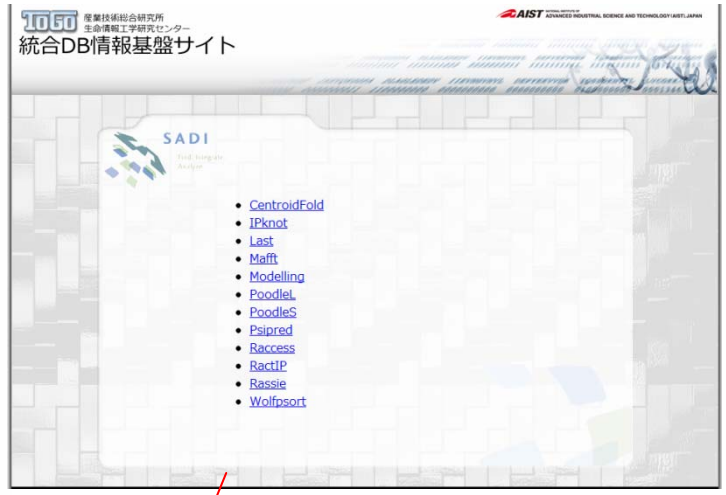
② プラットフォームとDB連携

http://togo.cbrc.jp/semantic.html#2

http://semantic.cbrc.jp

「SADIサービス」

Service URL	Input RDF
Blast	blastInput
CentroidFold	centroidfoldInput
ClustalW	clustalwInput
IPknot	ipknotInput
Last	lastInput
Mafft	mafftInput
POODLE-L	pooldeLInput
POODLE-S	pooldeSInput
PsiPred	psipredInput
Raccess	raccessInput
RactIP	ractipInput
Rassie	RassieInput
WolfPsort	wolfPsortInput



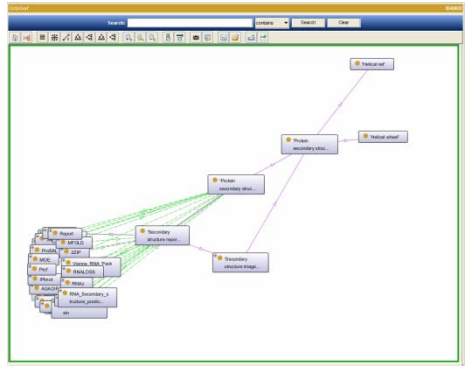
サーブレット

RDF Output file

サンプル RDF Input file

% curl --data-binary @mafftInput.rdf http://semantic.cbrc.jp/sadi-services/Mafft -o mafftOutput.rdf

H24年度実施計画



1)ワークフロー環境への新規ノード追加

H23年度に引き続き、本研究グループが独自に開発したツールや有用な既存ツールをノード化し公開する。

2)ツールのRDF入出力機能開発

解析ツールのSADIサービスによるRDF入出力機能を開発し公開する。また解析ツール・ソフトウェアに関するオントロジーの整備を行う。

3) ワークフロー環境での結果レポート表示機能開発

KNIMEのプラットフォーム上で解析された各種結果を、自動でレポート作成可能とする機能開発を検討する。

4) 解析プラットフォームへのRDF入出力機能追加

解析プラットフォームからSADIサービスを利用し、DBCLSが進める主要・有用DBのRDF化と連携してデータをRDF入出力可能とする機能開発を行う。