

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」

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
小原雄治

基盤技術開発プログラム実施体制

長洲毅志
PO

研究代表者

ライフサイエンス統合データ
ベースセンター
小原雄治

研究開発題目：生命科学分野における
データベース統合化のための基盤技術開発

共同研究グループ

産総研・
生命情報工学研究センター
浅井潔

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

共同研究グループ

京大・化学研究所
五斗進

研究開発題目：データ統合と新規分野
データ活用のための基盤技術開発

基盤技術開発プログラム平成24年度研究参加者

【DBCLS】

小原雄治
大久保公策
岡本忍
川本祥子
金進東
坊農秀雅
箕輪真理
山口敦子
小野浩雅
片山俊明
川島秀一

河野信
内藤雄樹
仲里猛留
山本泰智
飯田啓介
KOCBEK S.
呉紅艷
大田達郎
藤枝香
王悦
鐘ヶ江弘美
RA27名

研究協力者
中村春木
金城玲
小笠原理
岩崎渉
山崎千里

【CBRC】

浅井潔
福井一彦
田代俊行

【京大】

五斗進
時松敏明
小寺正明
中川善一
守屋勇樹

基盤技術開発プログラム背景

1. 生命科学分野の技術革新が進む
 - NGSデータ, オミックスデータ, パーソナルゲノム等データの爆発的増加
 - 測定機器, 解析機械の技術革新
2. 一方でデータベースは散在し, どこに何があるのかわかりづらい, 新規発見のスピードが上がらない, 創薬につながらない, 大量データの解析が困難, 人材不足等の問題点



データ, 文献, 知識を統合化し, 使いやすく提供することで, 研究, 医療, 産業, 研究の発展・深化を支援しイノベーションをもたらすことが急務

統合データベースプロジェクトの課題

- 大型プロジェクトの成果の統合
- 多様なデータを統合する基盤技術の開発
- データ駆動型科学への展開
- 解析ワークフローや統合検索など利用環境の整備
- 大規模データ・新規データの利用技術開発
- データ共有・データ公開の制度設計・場の提供
- 人材育成

統合化の全体像とステップ

データバンク事業

プロジェクトDB

個別DB

ツール



①データベースのカatalog化、ポータルサイト、ダウンロードサイト



②データベースやツールの使い方、使い分けの情報



③データベースやツールの統一的、シームレスな検索、利用



- ④知識発見支援のためのデータベース統合化、解析ワークフロー
⑤目的、用途ごとのデータベース統合化、解析ワークフロー



イノベーション、新たな知識発見、データベース生物学

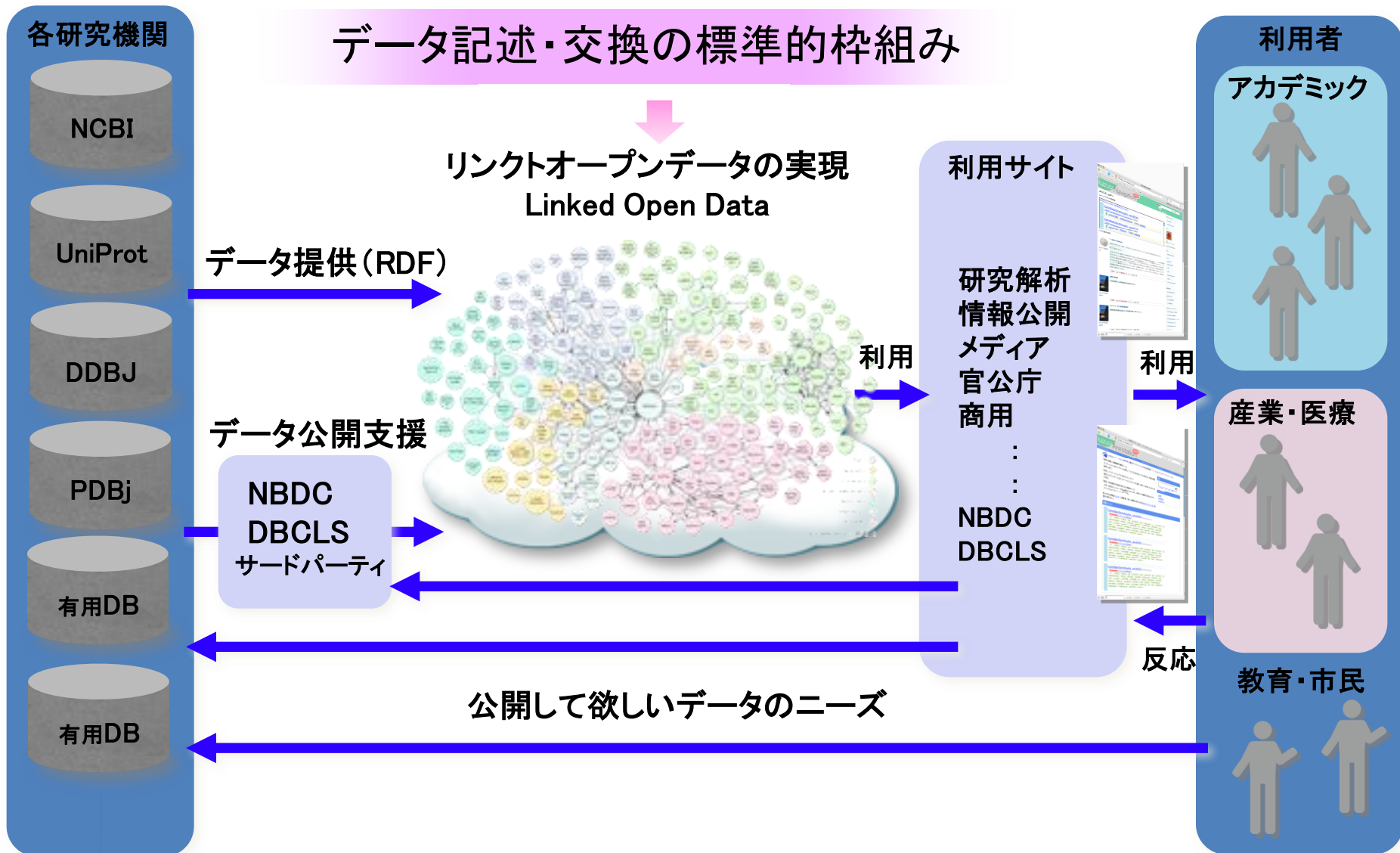
基盤技術プログラム方針(その1)

- ひとつの巨大DBを作るのではなくフェデレーション型の統合を目指す
- セマンティックウェブ技術を基盤に統合化を進める
- DBの連携により個別のDBからは得られない高度な情報や知見が獲得可能となるようなシステムを目指す(高度検索の実現)
- データの性質や特徴などを記述するメタデータの標準化や整理, 可視化技術の開発により, DBを利用しやすい環境を構築する

基盤技術プログラム方針(その2)

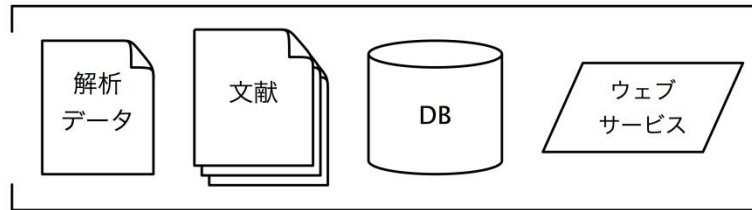
- 解析ツール群を統合し使用しやすい環境を構築する(DBだけでなく, DBを使う道具も整備)
- NGS, メタゲノム, 画像などBigDataにも対応する(NGSデータを中心に公共DBの利用技術を開発)
- NLP, Textmining技術 を生命科学分野に応用し, DB構築に貢献し, アノテーション, キュレーションの高度化を図る
- 開発だけでなく, 講習会, 日本語コンテンツ作成, 人材育成を通じて, 普及につとめる
- BioHackathon等での開発を通じ(国際的な)標準化をリードする

フェデレーション型データ統合による 生命科学分野のオープンデータの実現

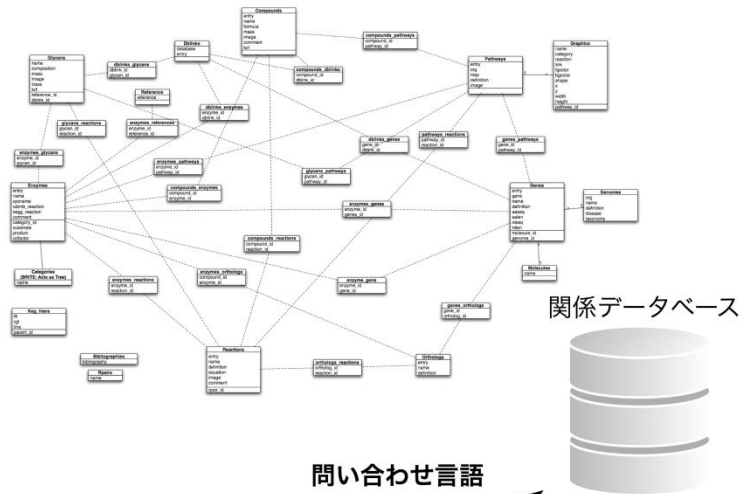


RDFを利用したDB統合・RDBとの比較

関係データベースモデル



スキーマ設計とデータの正規化

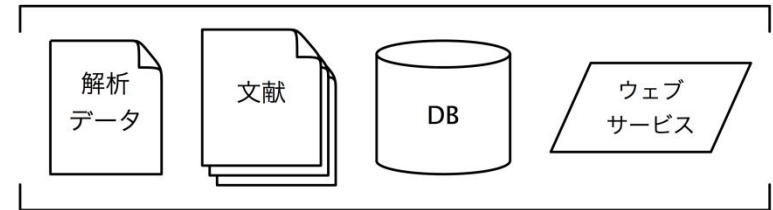


問い合わせ言語
SQL

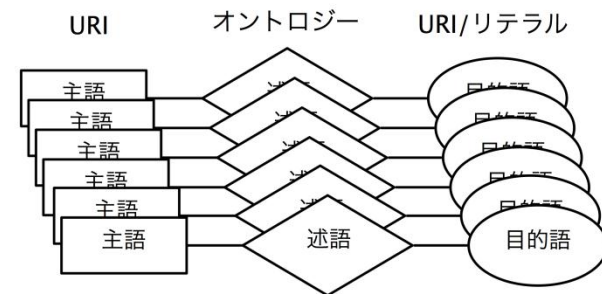
インタフェースと可視化



RDFデータベースモデル



RDFの三つ組に変換



オントロジー (統一用語)

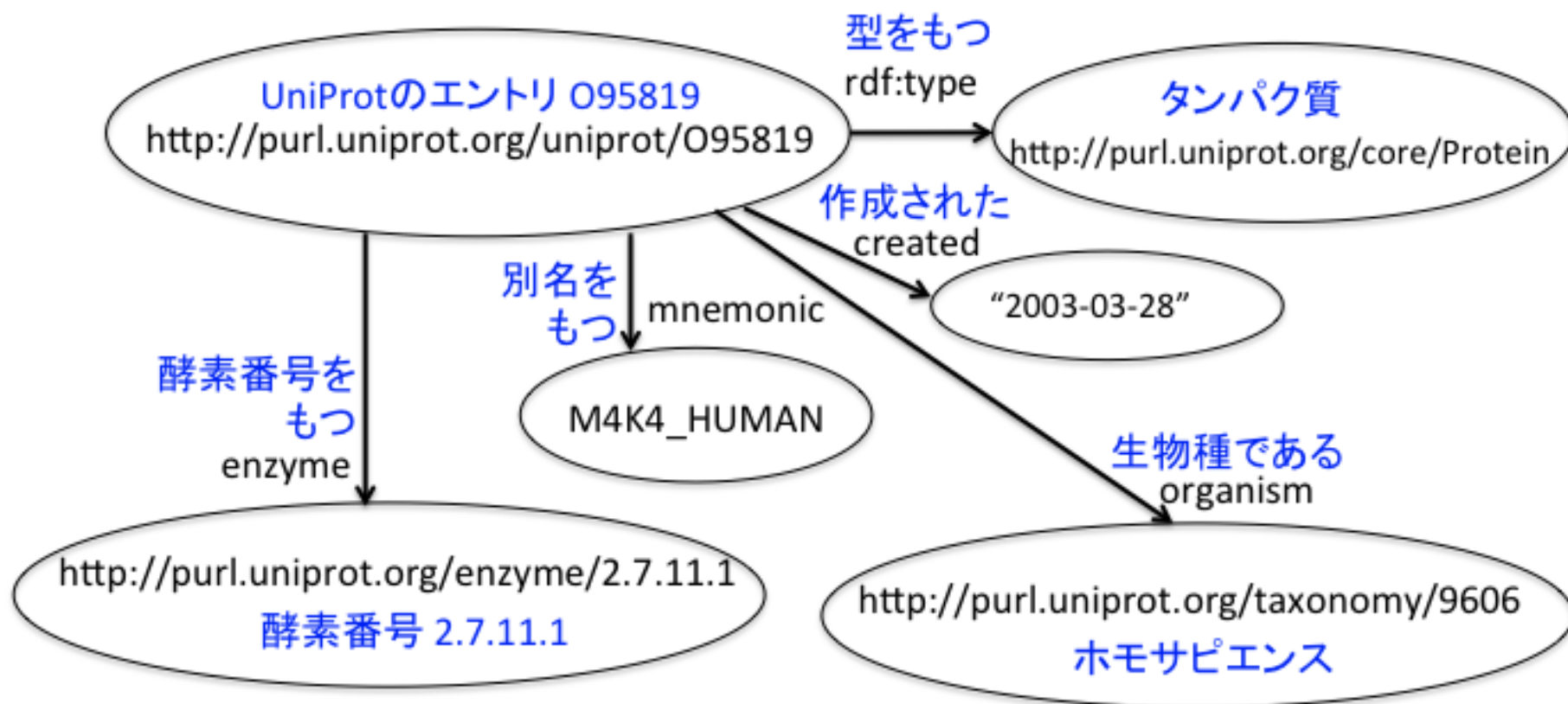
トリプルストア

問い合わせ言語
SPARQL

インタフェースと可視化



UniProtエントリーO95819の情報をRDFで表現



基盤技術開発プログラム研究課題

1. データベースのRDF による統合化(D, 京)
2. 解析プラットフォームによる統合利用環境の整備 (C, D)
3. インターネットを活用した高度検索技術の開発 (D)
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 (D, 京)
5. 大規模データの利用技術開発 (D, 京)
6. 情報統合化・知識発見のためのキュレーション支援 (D)
7. 統合データベースに関わるコンテンツの作成, 整備 (D)

分担 D:DBCLS, 京:京大, C:CBRC

昨年度(H23)の進捗概要

- データベースのRDFによる統合を進めるための各種調査を実施し,標準工程や指針を明らかにした
 - 有用DBの優先度付調査やRDF蓄積DB調査,オントロジ
- 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,統合化推進プログラムと連携し,オントロジ構築やRDF化を実施した

昨年度の課題と今年度の取り組み

- 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)
→トリプルストアやSPARQLについての運用実績を蓄積した
- 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
→自動化のためのツール, サポートツール開発. オントロジー構築や
キュレーションへの応用により問題点からの開発へのフィードバックを
実施した
- 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?)
→微生物やゲノム, プロテオーム, 発現にターゲットを絞って開発
- 必要な計算機の確保が単独では難しい
→遺伝研スパコン, NBDC計算資源の利用で一定の解決

課題別H24年度開発内容一覧

(1) データベースのRDFによる統合化

- 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の継続
- DBメタデータの蓄積及び利用
- LinkDBシステムの統合利用環境への応用（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

- ワークフロー環境整備（CBRC, DBCLS）
- ツールのRDF入出力機能追加開発（CBRC, DBCLS）
- 解析プラットフォームへの機能追加・整備（CBRC）
- ワークフロー環境での結果レポート表示機能開発（CBRC）

(3) インターネットを活用した高度検索技術の開発

- 有用分野でのプロトタイピング,検索結果の情報提示法と可視化開発
- NBDC運用の既開発サービスの機能向上
- NBDC統合化推進プログラムとの連携

課題別H24年度開発項目一覧

(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

- 言語資源リポジトリ開発
- 自然言語処理, テキストマイニング技術の整備
- 反応オントロジーの整備(京大)
- 解析ツールのオントロジー作成(CBRC)

(5) 大規模ゲノム配列データの利用技術開発

- 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
- 遺伝子発現のリファレンスデータセットの整備
- 医療用画像データの利用技術開発
- 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(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

- キュレーション支援システム開発
- 協働キュレーション作業運用技術の整備

(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
- 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

RDFによる統合化 進捗報告

課題(1)(2)(3)(4)(6)

RDF によるデータベース統合 メリットと課題

メリット

一つの検索技術で
様々な DB を縦横
無尽に利用できる
(複雑な検索)

自然文, キーワード,
実験データなど

課題1

検索インターフェイス

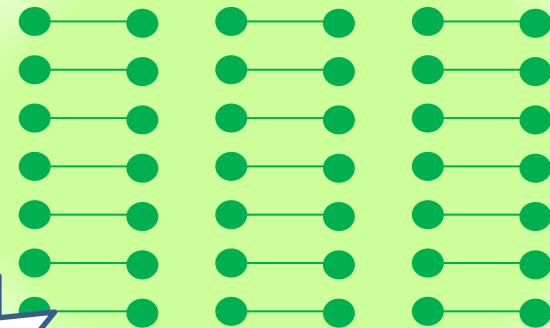
適切な可視化を
用いた結果表示

RDFクエリ結果

RDFクエリ言語

課題2

RDF データ



文献DB

ゲノムDB

タンパク質DB

...

ワークフロー作
成&実行

課題3

解析プラットフォーム

RDF

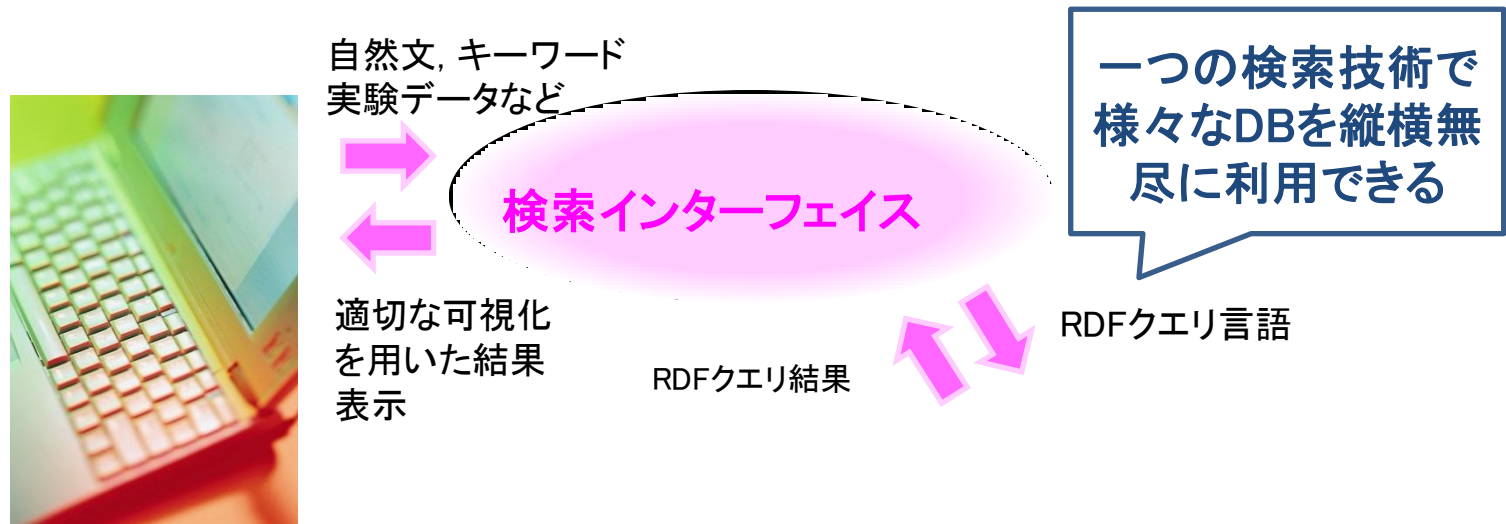
RDF

RDF 化した外部の
データベースも等価
に利用できる
(Web of Data)

メリット



課題1:検索インターフェ이스の課題



- 課題
1. 適切な検索条件・可視化は分野や目的によって異なる
 2. 一般のユーザが RDF クエリ言語 (SPARQL) を書くことは難しい
 3. 検索対象となる RDF DB の応答時間や安定性が不十分

IDの対応表取得の例:

RefSeq(ID)→MEO(ID)一覧取得

```
DEFINE sql:select-option "order"
prefix ncbi_gi:<http://www.ncbi.nlm.nih.gov/nuccore/>
prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#>
prefix meo:<http://purl.jp/bio/11/meo/>
prefix gold:<http://www.genomesonline.org/cgi-bin/GOLD/GOLDCards.cgi?goldstamp=>
SELECT ?seealso ?gi ?gold ?meo
FROM <http://localhost:8890/DAV/id>
WHERE {
  ?s rdfs:seeAlso ?gi
  FILTER regex(?gi, ncbi_gi:CP000768.1, "i" ).
  ?s rdfs:seeAlso ?seealso.
  ?seealso rdfs:label ?typelabel
  FILTER regex(?typelabel, "Genome sequencing", "i" ).
  ?seealso rdfs:seeAlso ?gold.
  ?gold meo:environmentalDescribed ?meo
```

検索インターフェ이스の課題への対策

1. 適切な検索条件・可視化は分野や目的によって異なる

TogoGenome: ゲノムを軸にデータを統合して、データの可視化に必要なテンプレートを揃えたシステム

TogoTable: 一般ユーザによく使われる表形式を入力として、ユーザが欲しいデータを関連する DB から付け加えて表形式で表示するシステム

2. 一般のユーザがRDFクエリ言語 (SPARQL) を書くことは難しい

LODQA: 自然言語(英語等)を SPARQL に自動変換するシステム

3. 検索対象となる RDF DB の応答時間や安定性が不十分

トリプルストア性能調査: 既存の RDF DB システムの性能調査

YummyData: ネットを介したSPARQLによる問合せに対応している外部DBについて、その品質を評価する試み

ゲノム統合環境: TogoGenome

ゲノムを軸とした RDF 統合環境

特長

ゲノムに関連する様々な側面を組み合わせた検索

側面の例:

- 環境 – 地理情報, 培地情報, 寄生/ホスト, ...
- 生物種 – タクソノミー, 進化系統情報, ゲノムサイズ, ...
- 遺伝子 – Gene Ontology, オースログ, 相互作用, モチーフ, アイソフォーム, 位置情報, ...
- 機能 – 疾病, 毒性, 光合成, 代謝パスウェイ, ...
- 文献 – 研究履歴, 参照元文献, 図表, ...

また, これらの側面を組み合わせたときの, 統計情報が容易に得られる

結果表示に再利用性の高い可視化テンプレートを利用

開発中(検討中のものを含む)の典型的テンプレート:

表, グラフ(棒グラフ, 円グラフ, 折れ線グラフなど),

木構造(ツリーマップなど), 地図情報,

生物学特有の表示テンプレート(ゲノムブラウザ, 遺伝子構造, モチーフなど)

テーブル型統合環境: TogoTable

(表形式 RDF 統合環境プロトタイプ)

特長: ユーザが持つ表データへ外部の DB からデータを簡単に追加可能

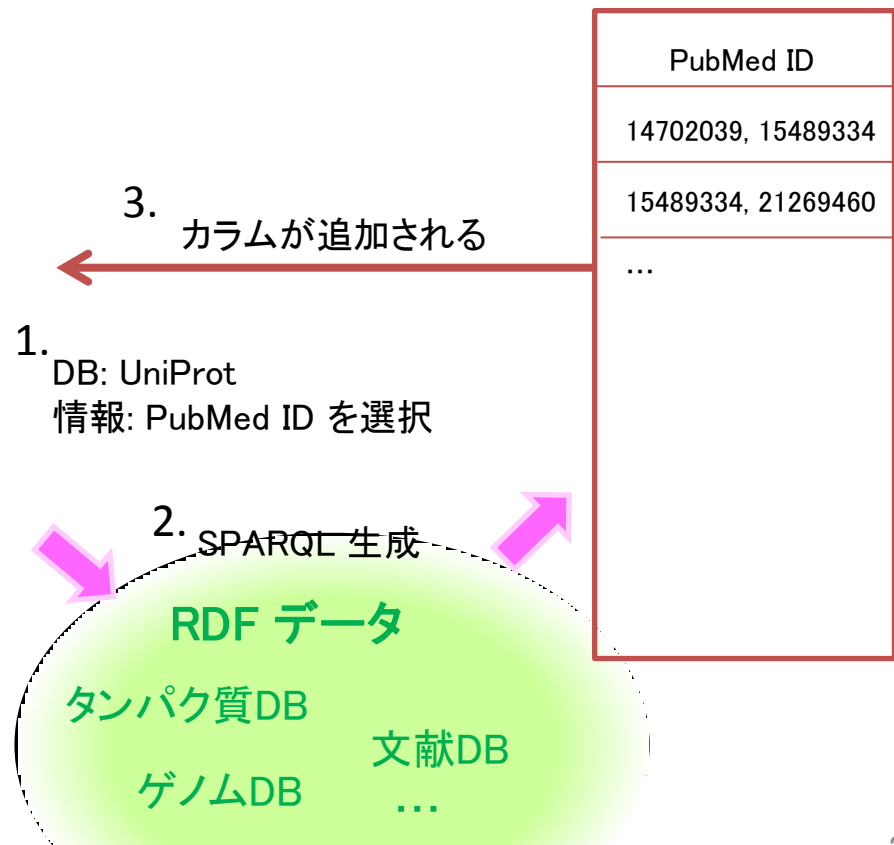
RDF to Table Tool | Upload

filename: 5_MALDI_PRO5d
id:50c90f70703e786658000001
total number:80
status: 0
Merged:0%
[Output csv](#)

N	Unused	Total	%Cov	%Cov(50)	%Cov(95)	Accession	Name
4	120.26	120.31	96.7592597	90.9722209	84.9537015	sp P14136 GFAP_HUMAN	Glial fibrillary acidic protein OS=Homo sapiens G more
7	68.68	68.72	70.48710585	57.73639083	56.7335248	sp P02787 TRFE_HUMAN	Serotransferrin OS=Homo sapiens GN=TF PE=1 SV=2
19	49.16	49.19	81.92090392	61.95856929	51.03577971	sp P146189 GPYM_HUMAN	Pyruvate kinase isozymes M1/M2 OS=Homo sapiens GN more
20	47.63	59.89	84.97853875	65.23604989	63.73390558	sp P08670 VIME_HUMAN	Vimentin OS=Homo sapiens GN=VM PE=1 SV=4
23	46.27	46.33	83.17972422	69.81566548	69.67742205	sp P06733 ENO4_HUMAN	Alpha-enolase OS=Homo sapiens GN=ENO1 PE=1 SV=2
27	42.55	42.55	85.82677245	78.21522355	73.49081635	sp P12277 KCRB_HUMAN	Creatine kinase B-type OS=Homo sapiens GN=CKB PE= more
28	41.19	41.23	45.89759409	21.34484947	18.07526201	sp P48681 NEST_HUMAN	Nestin OS=Homo sapiens GN=NES PE=1 SV=2

列を選択し、欲しい情報を指定すると、
RDF 化された関連する DB から
その情報を検索して表に付け加える

➤ 今年度是对应可能なDBを充実させた



LODQA

Linked Open Data Question Answering

自然言語をSPARQLに変換するシステム

生物学研究者には SPARQL で自分のやりたいことの記述が困難



自然言語を SPARQL へ自動変換するシステム LODQA を開発

- 期待される効果

SPARQL を書けないユーザでも、自然言語を用いて
RDF 化された DB 群の検索が可能に

LODQA

LODQA for SNOMED-CT@BioPortal

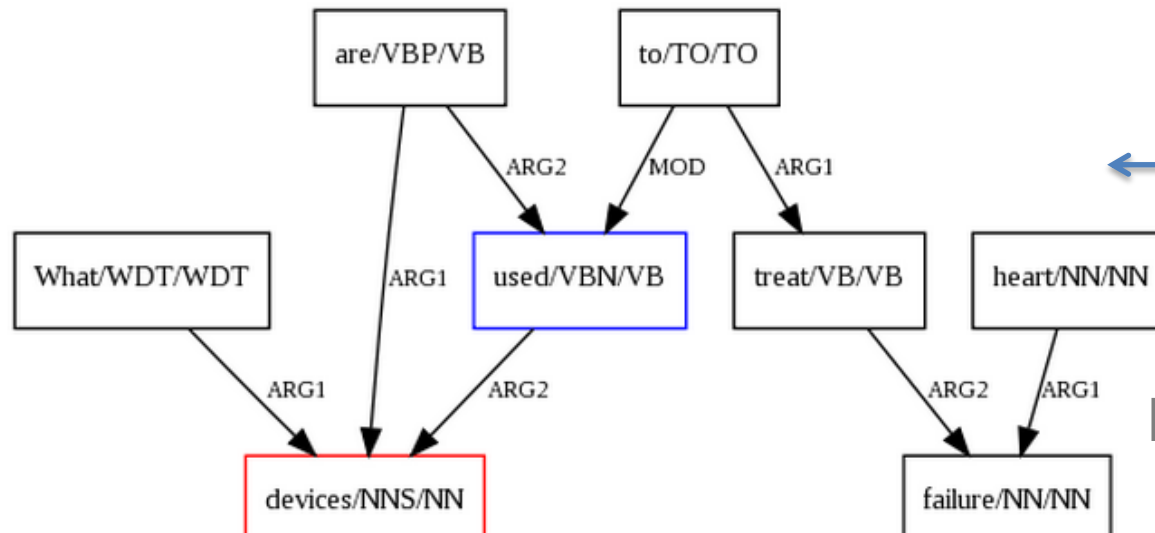
powered by Enju, a HPSG parser.

Query

What **devices** are used to treat **heart failure**?

← 質問文

Semantic analysis (predicate-argument relation graph)



← 質問文の構造解析

自動生成

生成されたSPARQL

SPARQL

```
SELECT ?t1 ?l1
WHERE {
  GRAPH <http://bioportal.bioontology.org/ontologies/SNOMEDCT> {
    ?t1 ?d1 <http://purl.bioontology.org/ontology/SNOMEDCT/49062001> .
    ?t2 ?d2 <http://purl.bioontology.org/ontology/SNOMEDCT/84114007> .
    ?t1 ?r1 ?t2 .
    ?t1 <http://www.w3.org/2004/02/skos/core#prefLabel> ?l1 .
  }
}
```

トリプルストア性能調査

トリプルストア: RDF データを格納し, SPARQL での検索が可能な DB

- SQL と異なり, SPARQL 検索は検索対象がグラフのため, グラフ比較が高速にできる必要がある
 - データ格納時に SPARQL に適したインデックス作成が必要だが, ストアによってインデックス作成方法は様々である
 - 生物学データのような巨大なデータに対する性能調査の例は現時点ではない

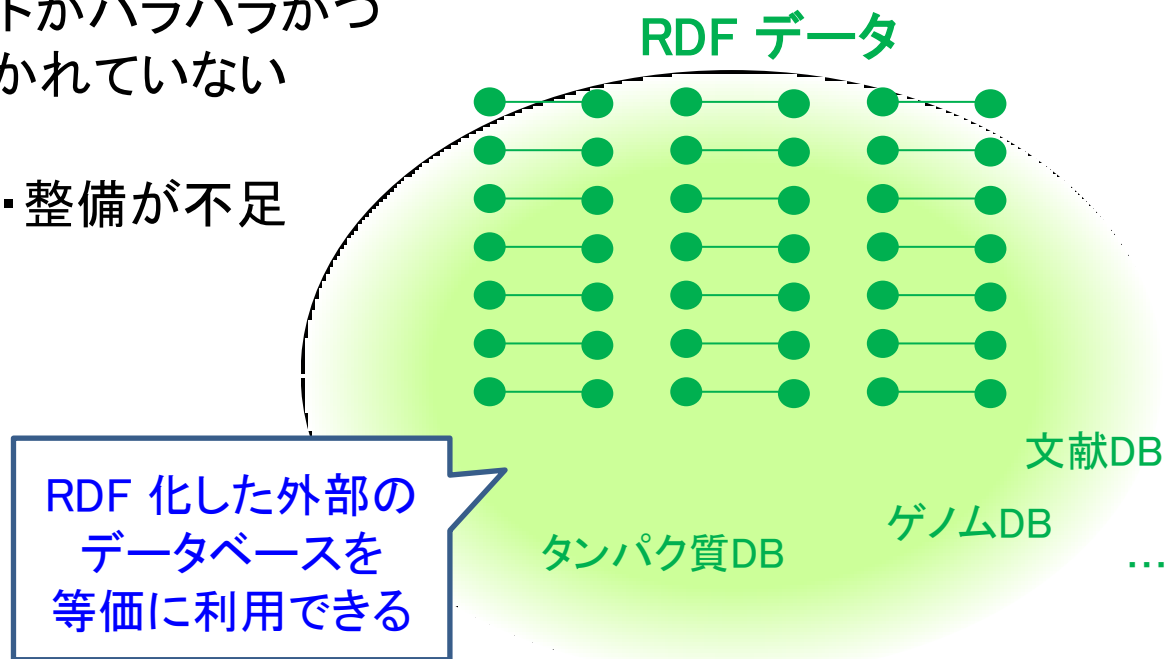


- DDBJ, UniProt など, 生物学 RDF データを利用して性能調査
 - 対象ストア: Virtuoso, 4store, OWLIM-SE, Mulgara, Bigdata
 - ロード時間およびクエリ応答時間等に関する性能調査を行った.

結論: 巨大データに対しては OWLIM-SE, Virtuoso が優れている

課題2: RDFデータの課題

1. 検索したいデータがRDFになっていない
2. RDF化をサードパーティで行おうとすると、
元データのフォーマットがバラバラかつ
セマンティックスが書かれていない
3. オントロジーの統合・整備が不足



RDFデータの課題への対策

1. 検索したいデータが RDF になっていない

DB の RDF 化: 検索に必要なデータ, 既存サービスの RDF 化

TogoDB: 表形式データの RDF 化支援ツール

PubAnnotation: 文献アノテーションの RDF 化支援ツール

2. RDF化をサードパーティで行おうとすると, 元データのフォーマットがバラバラかつセマンティックスが書かれていない

OpenBio*: 各種フォーマットに対するRDF化支援ツール

3. オントロジーの統合・整備が不足

OntoFinder/OntoFactory: オントロジー構築支援ツール

MEO, MCCV, GMO: 微生物・メタゲノム用オントロジーの開発

FALDO: ゲノム配列アノテーションの位置情報を記述するためのオントロジー

DB の RDF 化

1. 検索・解析に必要な DB のRDF 化
 - ゲノム情報の RDF 化
RefSeq 等のDBをRDF化
TogoGenome からの利用のため,
 - Life Science Dictionary の RDF 化
横断検索・Allie などのサービスからの利用のため
2. 既存データの再利用性・流通性を高めるためのRDF化
 - DB アーカイブのRDF化
DB アーカイブ内の60件のDBをRDF化
アーカイブ内のデータと外部のデータを組み合わせて利用可能に

TogoDB の RDF 化機能開発

TogoDB: 表形式のデータをインポートするだけで, ウェブデータベースの公開・管理が簡単にできるシステム

The screenshot displays the TogoDB web interface. At the top, the browser address bar shows 'semantic.togodb.dbcls.jp/togodb/config/antmossdb'. The page header includes 'TogoDB | Public Databases' and navigation links 'my Databases | Create | Release | Logout'. The user is logged in as 'http://openid.dbcls.jp/user/ktym'. The main section is titled 'Database config: antmossdb' and shows '56590 records, 19 columns'. Below this are tabs for 'Operation', 'Release', 'HTML', 'Metadata', and 'User'. The 'Release' tab is active, showing a table with columns: Dataset, File size (bytes) (CSV, JSON, RDF (Turtle), RDF (XML), FASTA), Release date, and Status. The 'default' dataset is listed with file sizes: CSV (13,609,132), JSON (31,952,704), RDF (Turtle) (88,452,734), RDF (XML) (227,096,790), and Release date (2011-11-23 10:11:13 UTC). A 'Reload' button is at the bottom of this section. The 'Column setting' section is below, with tabs for 'General', 'Link', 'RDF', and 'Dataset'. The 'RDF' tab is active, showing a table with columns: Column, Label, Property / Class, and Class. The table lists columns like 'number_registered', 'date_collected', 'latitude', 'longitude', and 'taxonomy_class' with their corresponding labels and RDF property/class settings.

Dataset	CSV	JSON	RDF (Turtle)	RDF (XML)	FASTA	Release date	Status
default	13,609,132	31,952,704	88,452,734	227,096,790		2011-11-23 10:11:13 UTC	Release

Column	Label	Property / Class
number_registered	Number registered	Property: [dropdown], Class: [dropdown]
date_collected	Date Collected	Property: dc, date, http://purl.org/dc/elements/1.1/date; Class: [dropdown]
latitude	Latitude	Property: [dropdown], http://www.w3.org/2003/01/geo/wgs84_pos#lat; Class: [dropdown]
longitude	Longitude	Property: [dropdown], http://www.w3.org/2003/01/geo/wgs84_pos#long; Class: [dropdown]
taxonomy_class	Class	Property: [dropdown], Class: [dropdown]

- TogoDB 内のデータを効率的に RDF 化するための機能を開発
- 今年度: オントロジー管理インターフェイスを開発, インポートされたデータをオントロジーに沿って半自動的に RDF に変換可能

PubAnnotation

文献アノテーション標準化ストレージ

文献アノテーションの例

タンパク質

タンパク質

Runx3 and T-box proteins cooperate to establish the transcriptional program of effector CTLs.

細胞

文献アノテーションの問題点

- アノテーションされた文献が蓄積してきているが
- フォーマットがバラバラ, テキスト, 文字コードも違う
- 異なるグループが作ったアノテーションの比較・交換・統合ができない



- ✓ 文献アノテーション標準フォーマット PubAnnotation Annotation Format を提唱 (RDF に変換可能)
- ✓ そのフォーマット上で文献アノテーションを蓄積するストレージ & システム PubAnnotation を開発

文献アノテーションの例

TextAE (source: <http://localhost:3000/annsets/bionlp-st-ge-2013-sample/pmcdocs/2626671/divs/1/annotations.json>)

Span Types

- ☐ Sentence
- ☐ Anaphora
- ☒ Protein
- ☐ Entity
- ☐ Binding
- ☐ Localization
- ☐ Gene_expression
- ☐ Transcription
- ☐ Phosphorylation
- ☐ Regulation
- ☐ Positive_regulation
- ☐ Negative_regulation

Relation Types

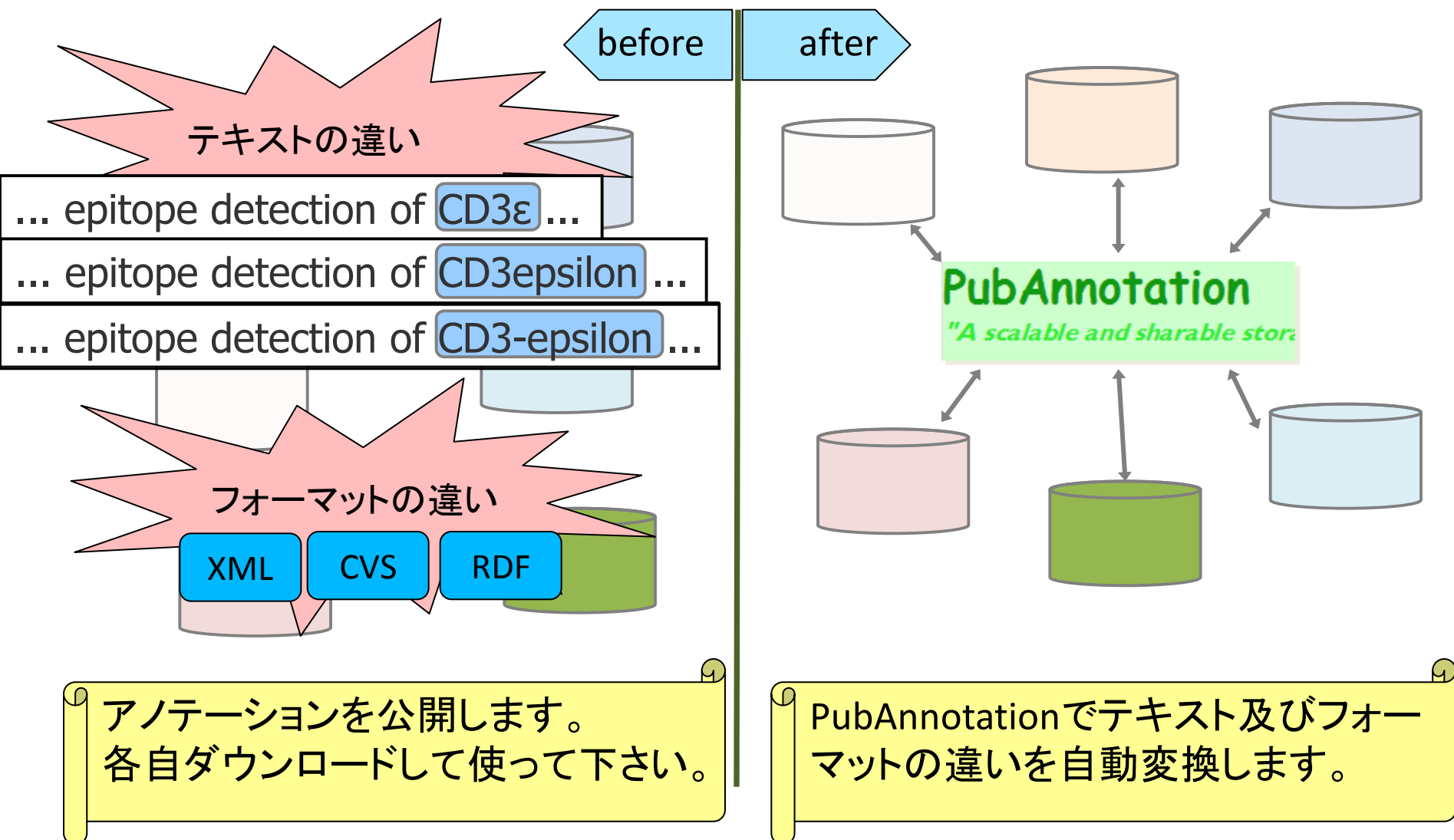
- ☐ themeOf
- ☐ causeOf
- ☐ coreferenceOf
- ☒ equivalentTo
- ☐ partOf

Although Runx3 is not expressed in naive CD4+T cells, its expression is up-regulated during Th1 cell differentiation, and Runx3 influences Th1 cell differentiation and function through direct regulation of the Il4 and Ifng cytokine genes (15,16).

In contrast, all three Runx proteins are expressed in mature CD8+T cells (10,12), and Runx3-deficient CD8+T cells show reduced cytolytic activity (12,13).

Example from BioNLP Shared Task 2011

文献アノテーションの統合



OpenBio*

元データのフォーマットがバラバラかつセマンティックスが書かれていない



各種フォーマットに対する RDF 化支援ツール
(RDF コンバータ)開発

RDF コンバータの例:

- RefSeq
- GFF3
- GTF(Cufflinks)
- Bio-table(表データ)
- ...

OntoFinder/OntoFactory

既存のオントロジーを利用したオントロジー開発支援ツール

今年度:ユーザインターフェイスを改良

OntoFinder powered by NCBO BioPortal

Data

heart
thumb
skin
hair
finger

Get an example

Configuration

キーワード群
入力

自分用オントロジー

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns=""
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xml:base="">
  <owl:Ontology rdf:about="" />
  <owl:Class rdf:about="http://sig.uw.edu/fma#Heart">
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Heart</rdfs:label>
  </owl:Class>
  <owl:Class rdf:about="http://purl.org/obo/owl/UBERON#UBERON_0001463">
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">hand digit 1</rdfs:label>
  </owl:Class>
  <owl:Class rdf:about="http://sig.uw.edu/fma#Skin">
```

自分用オントロジー

OntoFactory Show / Hide columns Show / Hide rows

Input terms	Selected terms clear	EP	MCCL	LNC	UBERON	FMA
heart		Heart	Heart	Heart	circulatory organ	Heart
thumb		Thumb				
skin		Skin	Skin	Skin	integument	Skin
				Skin	zone of skin	
				Skin		
hair		Hair	Hair	Hair	strand of hair	Hair
				Hair		

オントロジー検索結果
自動ソート

ターム選択

OntoFactory

Input terms	Selected terms clear	FMA	UBERON
heart	FMA Heart	Heart	circulatory organ
thumb	UBERON hand digit 1	Thumb	hand digit 1
skin	FMA Skin	Skin	integument
hair	FMA Hair	Hair	strand of hair
finger	UBERON hand digit	Finger	hand digit

Export selected terms to an OWL file

Ontology URI default: blank

Filename Filename not or leave it blank to output to the screen

エクスポート

微生物・メタゲノム用 オントロジー開発

微生物やメタゲノムの採取箇所、生育箇所、菌株、培地など様々な側面から検索するためのオントロジーが存在しなかった



統制語彙集として三つのオントロジーを構築中

MEO : 環境オントロジー (BioPortal で公開済)

MCCV : 菌株オントロジー(構築中, 公開予定)

GMO : 培地オントロジー(構築中, 公開予定)

今後の展開

微生物ゲノム・メタゲノムグループのもつデータに適用予定

FALDO

Feature Annotation Location Description Ontology

ゲノム配列アノテーションの位置情報を記述するためのオントロジー

背景:

生命科学の多くのアノテーション情報は配列上の位置情報に帰属できるが、分野やDBごとに記述方法は統一されていなかった

目的:

塩基配列, アミノ酸配列, 糖鎖などの領域にアノテーションをつける際に利用する位置情報の記述方法を標準化

2012年11月 <http://biohackathon.org/resource/faldo> にて公開

2013年より UniProt, TogoGenome で利用開始予定

NBDC/DBCLS BioHackathon 2012

外部の組織と国際協力のもと取り組むべき課題

検索対象となる外部 DB の応答時間や安定性が不十分

YummyData: ネットを介したSPARQLによる問合せに対応している外部DBについて, その品質を評価する試み

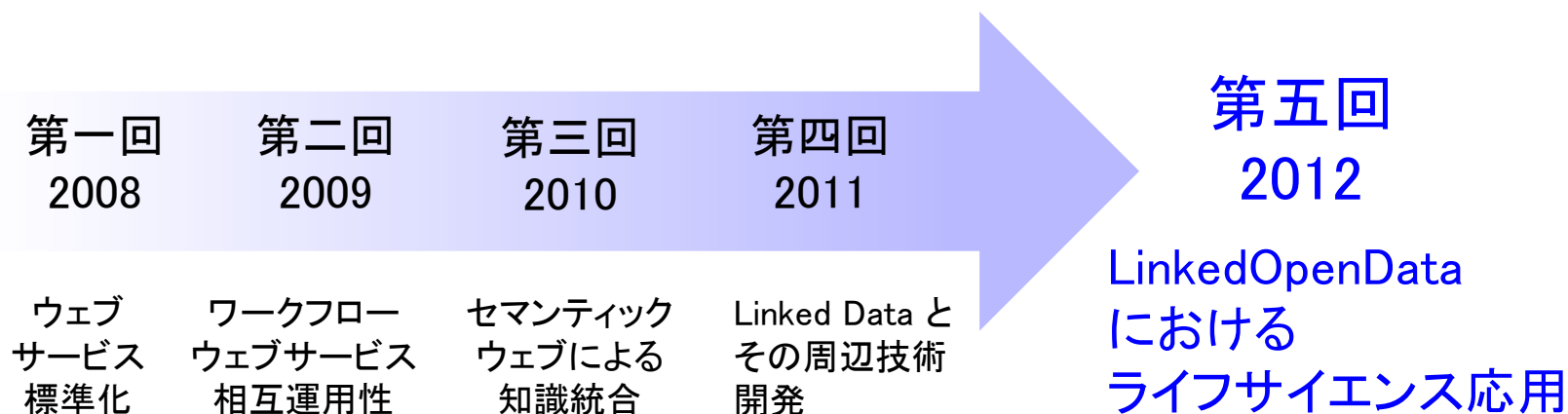
RDF化をサードパーティで行おうとすると, 元データのフォーマットがバラバラかつセマンティックスが書かれていない

OpenBio*: 各フォーマットに対するRDF化コンバータの作成

オントロジーの統合・整備が不足

FALDO: ゲノム配列アノテーションの位置情報を記述するためのオントロジー

NBDC/DBCLS BioHackathon 2012



参加者

海外 24名

参加プロジェクト SADI, UniProt, BioPAX, W3C HCLSIG, NCBO, OepnBio*,
BioGateway, EnvO, Glycoinfo, PhenomeBrowser, OntoFrame, TextMining, ...

国内 54名

参加プロジェクト OpenBio*, 東北メディカル・メガバンク,
KEGG, H-inv, G-language, Glycoinfo, TextMining, ...



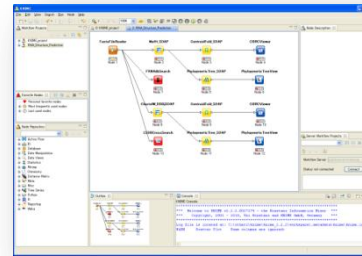
課題3: 解析プラットフォーム



解析プラットフォームとして、

- 創薬向けを中心に使われている **KNIME** --- CBRC 担当
- ゲノム解析を中心に使われている **Galaxy** --- DBCLS 担当

を選択

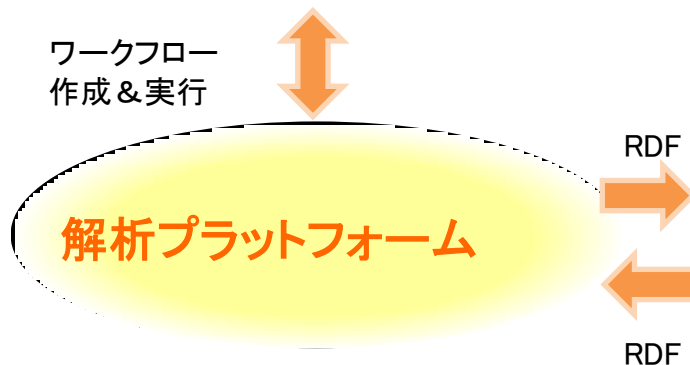


KNIME



Galaxy

RDF 入出力機能を開発



課題: 解析プラットフォームとツール群 / DB の入出力がバラバラ

➡ SADI(RDF 入出力のフレームワーク)に
解析プラットフォーム・ツール群を対応させる

大規模ゲノム配列データの利用技術開発 進捗報告

課題(5)

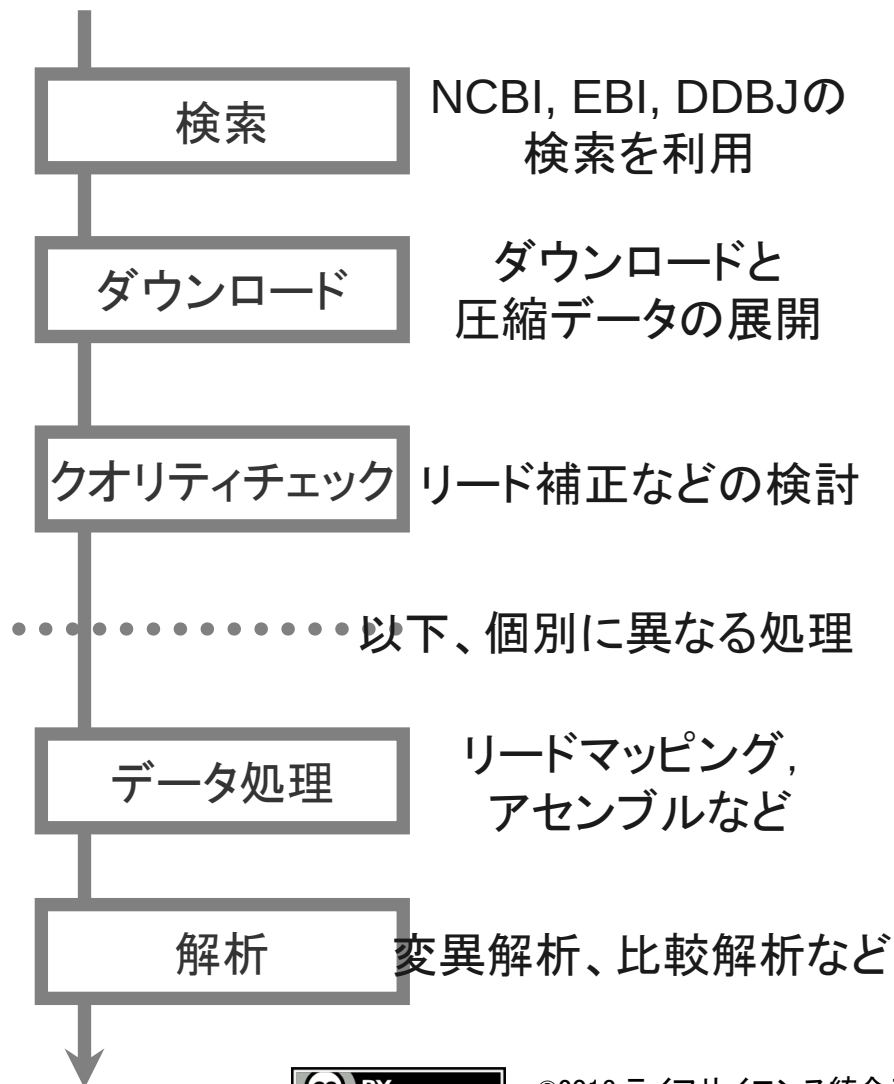
公共DBを使う上での問題点

1. 流行りのNGSデータが使えない
2. 遺伝子を簡単に検索したい
3. 発現データは似たようなものが多数登録されていて、どれを使ったらいいのかわからない

1. NGSデータが使えない

大規模データ利用技術開発: 公共次世代シーケンスDBのための研究開発

公共NGSデータの一般的な利用ワークフロー



国内でのユースケース

- 論文で発表された結果の追試
- 他チームのデータを自身の解析環境において再解析
- 自身のシーケンスデータと組み合わせて利用
- SNP情報のフィルタリング, 比較解析など
- 自身の開発する解析ツールのテストに利用

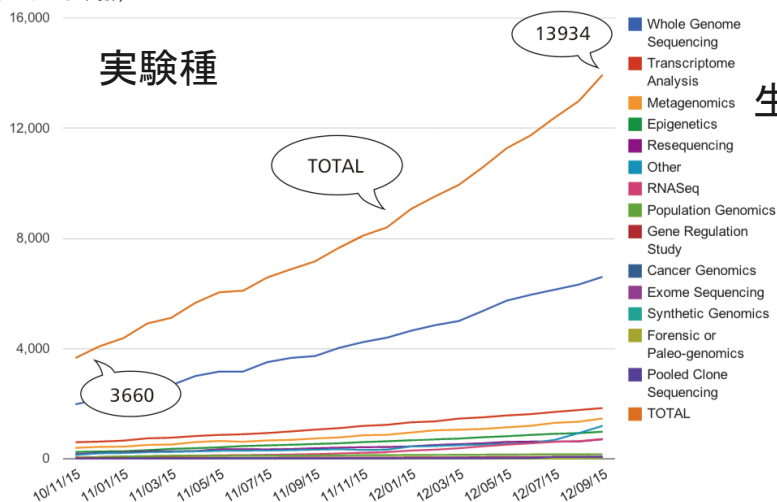
データ利用時における共通の問題点

- 検索におけるコスト
- データ量が多く数百件のヒットがあることも
- 欲しいデータは登録されているのか?
 - データベースの全体像が分からない
- ダウンロードにおけるコスト
- データサイズが大きく長い時間を要する
- 質の低いシーケンスデータが混在している
- より質の高いデータを利用したい

大規模データ利用技術開発: 公共次世代シーケンスDBのための研究開発

DBCLS SRA: メタデータの整理と統計による公共NGSデータのインデックス <http://sra.dbcls.jp/>

(プロジェクト数)



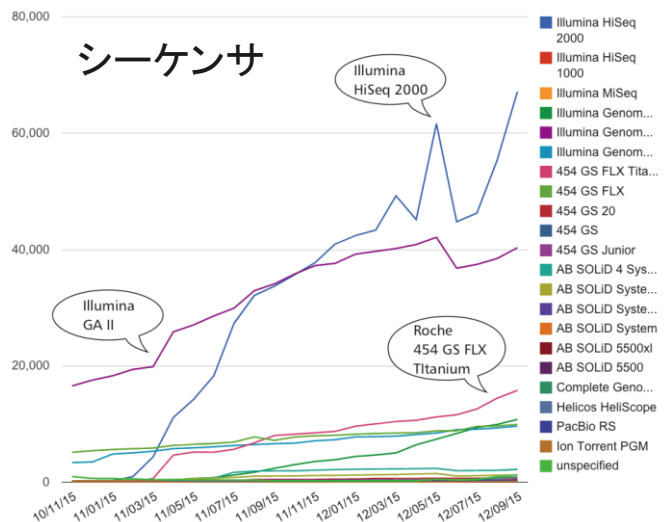
Species of samples (top 15)

生物種

<i>Homo sapiens</i>	1376
unidentified	886
<i>Mus musculus</i>	886
<i>Drosophila melanogaster</i>	305
<i>Mustela putorius furo</i>	233
<i>Caenorhabditis elegans</i>	203
metagenome sequence	188
<i>Arabidopsis thaliana</i>	171
marine metagenome	169
unclassified phages	139
soil metagenome	138
<i>Escherichia coli str. K-12 substr. MG1655</i>	137
<i>Saccharomyces cerevisiae</i>	131
<i>Bacteria</i>	120
<i>Human rhinovirus A</i>	95
Total	17742 (studies)

文献情報とのリンク

PMID	Article Title	Journal	Vol	Issue	Page	Date	SRA_ID_ump	SRA_ID	SRA Title
2319858	Biochemical Characterization of Foreign Gene Expression in Bacterial Hosts	PLoS genetics	8	11	e1003039	2012-Nov	ERR019170	ERR019170	Adenovirus transcriptome
2319874	Whole-genome sequencing for analysis of an outbreak of multidrug-resistant <i>Staphylococcus aureus</i> : a descriptive study	The Lancet infectious diseases	-	-	-	2012-Nov-9	ERR005540	ERR005540	MRSA sequencing, test
23148455	Transcriptome de novo assembly from next-generation sequencing and comparative analyses in the hexaploid salt marsh species <i>Spartina maritima</i> and <i>Spartina alterniflora</i> (Poaceae)	Hereditas	-	-	-	2012-Nov-14	SRA058648	SRA058648	<i>Spartina maritima</i> Transcriptome or Gene expression
23148455	Transcriptome de novo assembly from next-generation sequencing and comparative analyses in the hexaploid salt marsh species <i>Spartina maritima</i> and <i>Spartina alterniflora</i> (Poaceae)	Hereditas	-	-	-	2012-Nov-14	SRA058652	SRA058652	<i>Spartina alterniflora</i> Transcriptome or Gene expression
23118011	Mice Lacking a Myc Enhancer Element That Includes Human BMP-6/8/32/7 Are Resistant to Intestinal Tumors	Science (New York, N.Y.)	-	-	-	2012-Nov-1	ERR189597	ERR189597	Mice Lacking a Myc Enhancer Element That Includes Human BMP-6/8/32/7 Are Resistant to Intestinal Tumors
23110403	Characterization of a transcriptome from a non-model organism, <i>Cladonia rangiferina</i> , the grey reindeer lichen, using high-throughput next-generation sequencing and EST sequence data	BMC genomics	13	1	575	2012-Oct-30	SRA050676	SRA050676	Characterization of a transcriptome from a non-model organism, <i>Cladonia rangiferina</i> , grey reindeer lichen using next-generation sequencing and EST sequence data
23105094	Draft Genome Sequence of <i>Aeromonas hydrophila</i> Strain H-1	Journal of bacteriology	194	22	6395	2012-Nov	SRA055305	SRA055305	<i>Aeromonas hydrophila</i> H-1 Genome sequencing
23108058	Complete Genome Sequence of <i>Bacillus subtilis</i> Strain GB208, a Strain Widely Used in B. subtilis Genetic Studies	Journal of bacteriology	194	22	6308-9	2012-Nov	SRA057093	SRA057093	<i>Bacillus subtilis</i> GB208 Genome sequencing
23099672	Coordinated repression of cell cycle genes by KMSA and E2F4 during differentiation	Proceedings of the National Academy of Sciences of the United States of America	109	45	18495-504	2012-Nov-6	SRA033703	SRA033703	Genome-wide maps of KMSA/LARP2/ARBP2 in mouse ES cells
23087246	Real-time evolution of new genes by innovation, amplification, and divergence	Science (New York, N.Y.)	338	6105	364-7	2012-Oct-19	SRA057534	SRA057534	Real time evolution of new genes by innovation, amplification and divergence



整理されたメタデータを元に公共NGSデータを検索する

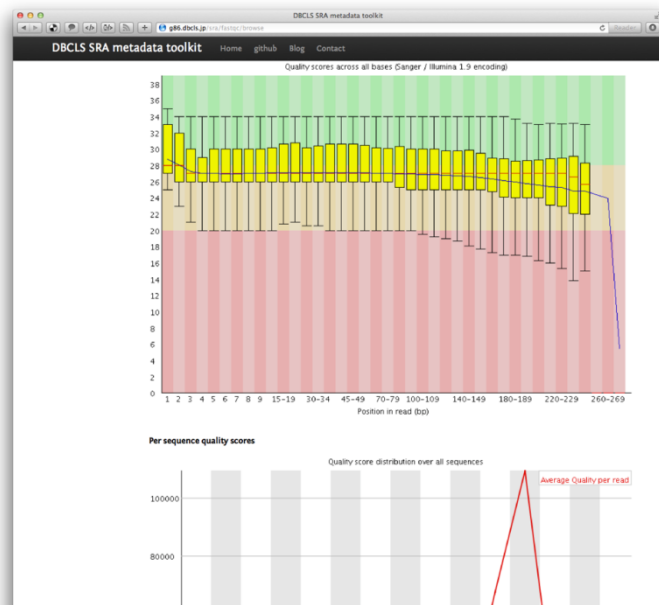
- 絞り込み検索の提供により、検索が容易に
- 統計情報により、データベースの俯瞰が可能に
 - どのようなデータがどの程度登録されているのか
- 関連する文献情報を追加し、情報を補完
 - よりメタデータの充実したデータを選ぶことができる

大規模データ利用技術開発: 公共次世代シーケンスDBのための研究開発

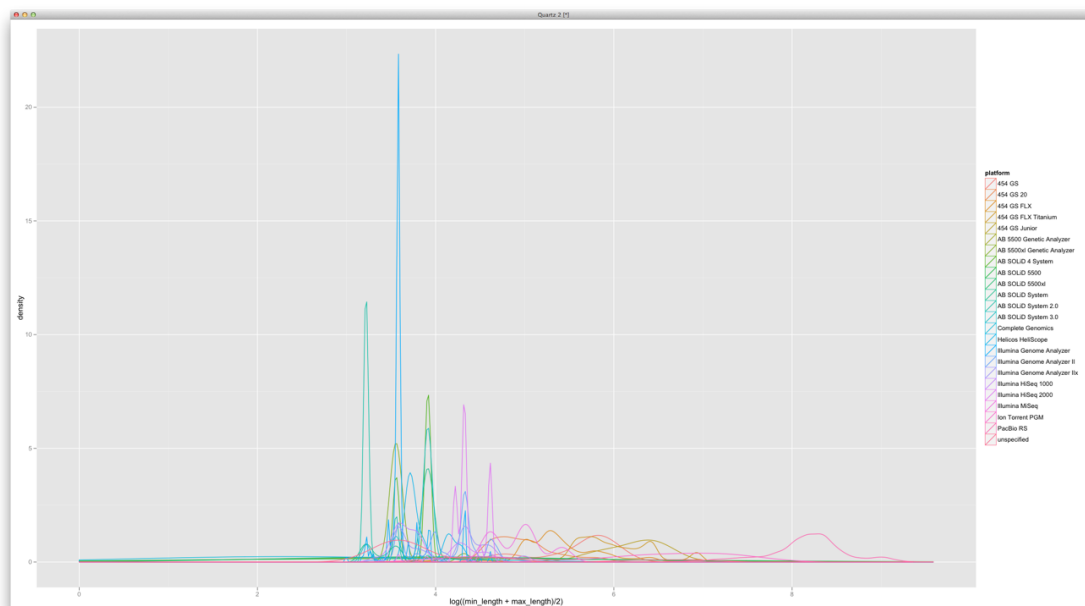
DBCLS SRA: 個別にデータの特徴を計算しメタデータを補完する

<http://sra.dbcls.jp/>

個別データのシーケンスクオリティ



個別に計算したプロフィールを利用した統計の例:
シーケンサ別リード長のばらつき



個別にシーケンスクオリティを確認する

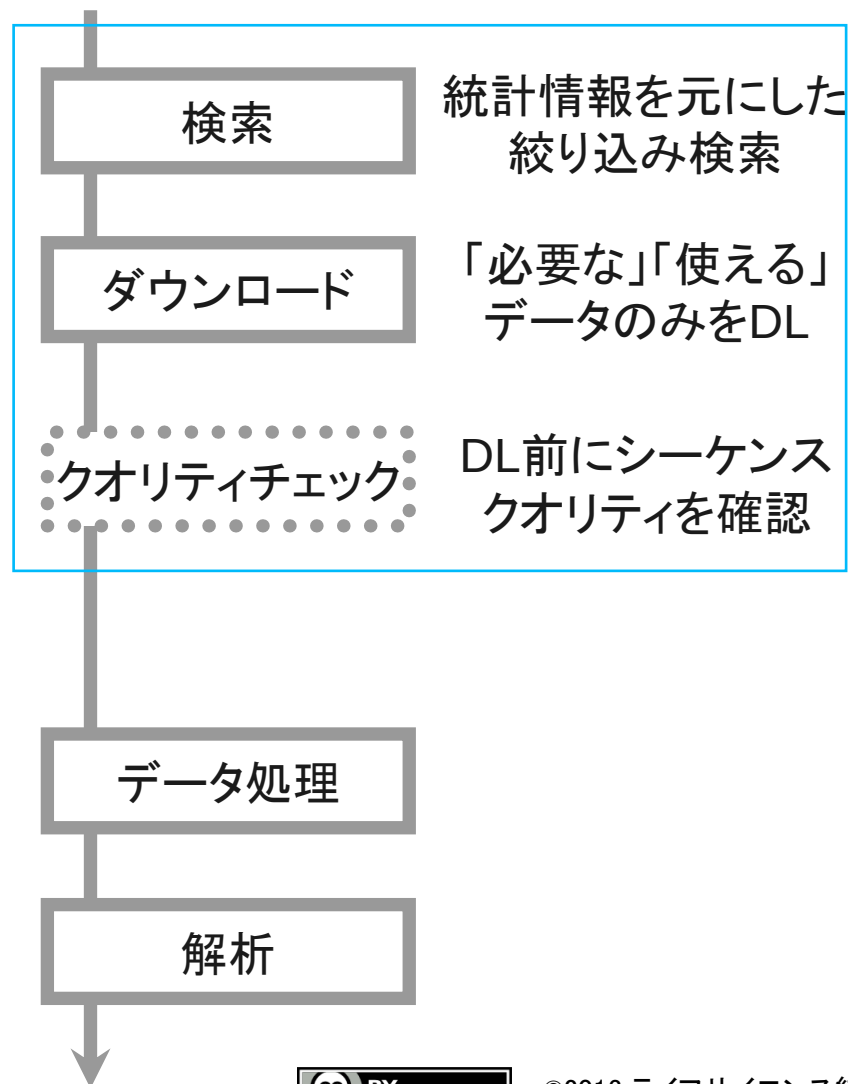
- より質の高いデータを選ぶことが可能
- 失敗したシーケンスのデータを避ける
- 有用なデータを選抜、ダウンロードコストも下がる

計算結果を利用した統計でDB全体を俯瞰する

- 全体のデータの分布の情報で検索を補助
 - どんなデータがあってどんなデータがないのか
- 研究動向などの二次的な情報もユーザに提供

大規模データ利用技術開発: 公共次世代シーケンスDBのための研究開発

DBCLS SRAによる利用ワークフローの改善



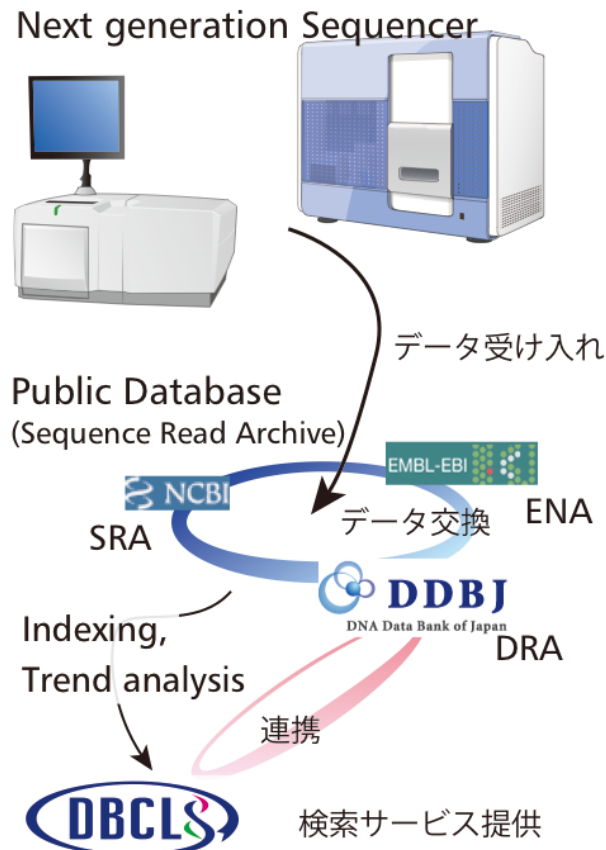
各ユースケースに共通する問題点を解決

- 検索のコストダウン
 - 絞り込み検索によって検索を容易に
 - 論文の情報を加えることで情報を補完
 - 統計情報により「そもそも必要なデータは登録されているか」把握することができる
- 事前にシーケンスクオリティをチェック
 - DL後のチェックのコストを省く
 - 「必要なデータだけ」を選ぶことでダウンロードのコストも下げることができる

DBCLS SRAを通じて公共データを研究に利用した例

- 公共のEpigenome dataを利用して比較解析
 - Waki et al. PLoS Genet, 2011
- カイコのRNA-seqを利用して遺伝子発現解析
 - Ogata et al. PLoS One, 2012

大規模データ利用技術開発: 公共次世代シーケンスDBのための研究開発



Sequence Read Archive(SRA)における他機関との連携

- Bioproject, Biosampleなど新たなデータ整理の枠組み
 - SRAの共同運営を行うNCBI, EBI, DDBJによるプロジェクト
 - ID関係が従来から変更される
 - ユーザにとってはより複雑に
 - 進行中の開発について情報交換のためDDBJと連携
- Sequence Read Archiveに関連する開発について
 - EBI SRAチームと技術情報交換ミーティング (2012.09)
- DDBJ スパコンの利用
 - 利用可能な全てのデータのクオリティチェック計算
 - 約30万のシーケンスファイルをダウンロード

2. 遺伝子を簡単に検索したい

<http://GGRNA.dbcls.jp/>

- ## ■ あらゆるキーワードや塩基配列・アミノ酸配列からすばやく簡単に遺伝子を検索するウェブサーバ

- ## ■ RefSeqのmRNA+ncRNAを 高速に全文検索する

- PCRのプライマー、マイクロアレイのプロブ、siRNAの標的配列を即座に確認

- REST API提供により外部のシステムと連携可能

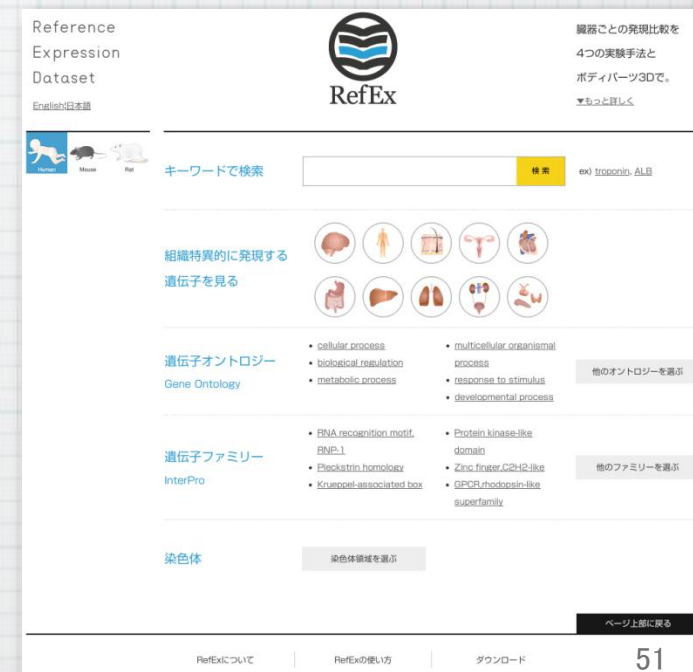
[illegible]

3. 発現データは似たようなものが
多数登録されていてどれを使ったら
いいのかわからない

遺伝子発現リファレンスデータセット『RefEx』

<http://refex.dbcls.jp/>

- 臓器ごとの発現比較を4つの実験手法とボディパーツ3Dで
 - 正常組織・臓器における遺伝子発現のリファレンス
 - 再利用可能で有用なパブリックデータの活用例
- 発現データを素早く比較し、詳細に調べられる
 - インクリメンタルサーチ
 - 「組織特異的遺伝子」検索機能の実装
 - BodyParts3Dによる発現状況の可視化
- 今後の開発方針
 - Illumina RNA-seq データの追加
 - 正常組織 vs ガン組織
 - 細胞・細胞株データ(FANTOM5)版



Reference
Expression
Dataset
[English](#) | [日本語](#)



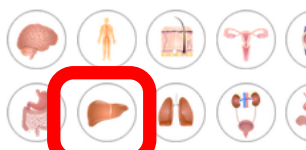
臓器ごとの発現比較を
4つの実験手法と
ボディパーツ3Dで。
[▼もっと詳しく](#)



キーワードで検索

検索 ex) troponin, ALB

組織特異的に発現する遺伝子を見る



遺伝子オントロジー
Gene Ontology

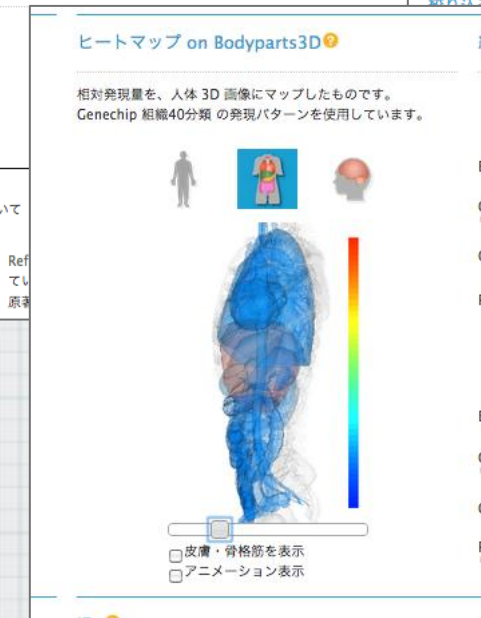
- cellular process
- biological regulation
- metabolic process
- multicellular organismal process
- response to stimulus
- developmental process

遺伝子ファミリー
InterPro

- RNA recognition motif, RNP-1
- Pleckstrin homology
- Krüppel-associated box
- Protein kinase domain
- Zinc finger, C2H2
- GPCR, rhodopsin superfamily

染色体

RefExについて



Reference
Expression
Dataset

[English](#) | [日本語](#)



検索

臓器ごとの発現比較を
4つの実験手法と
ボディパーツ3Dで。
[▼もっと詳しく](#)



一覧

絞り込み検索

結果:463件 検索キーワード:Genechip tissue40 ID=31 (肝臓)

« Previous 1 2 3 4 5 6 7 8 9 10 11 ... 47 Next »

1ページ当たりの表示件数 10

遺伝子名

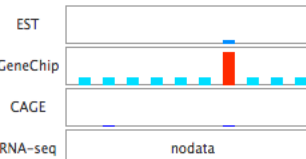
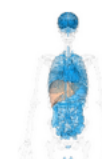
3Dマップ

相対発現量



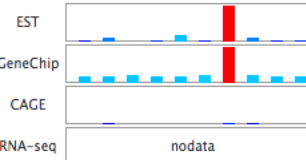
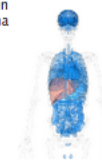
[mannose-binding lectin \(protein C\) 2, soluble \(opsonic defect\)](#)

同義遺伝子名 Mannan-binding protein, MBL, MGC116833, MBL2, MBP-C, Mannose-binding protein C precursor, MGC116832, Mannose-binding lectin, COLEC1, HSMBPC, MBP1
Refseq ID NM_000242
GeneID 4153
Unigene ID Hs.499674
Probe set ID 207256_at, [HG-U133_Plus_2]



[fibrinogen alpha chain](#)

同義遺伝子名 MGC119423, FGA, fibrinogen alpha chain, Fibrinogen alpha chain precursor, Fib2, MGC119422, MGC119425
Refseq ID NM_000508
GeneID 2243
Unigene ID Hs.351593
Probe set ID 205649_s_at, [HG-U133_Plus_2]



[aldo-keto reductase family 1, member D1 \(delta 4-3-ketosteroid-5-beta-reductase\)](#)

同義遺伝子名 3-oxo-5-beta-steroid 4-dehydrogenase, 3o5bred, SRD5B1, AKR1D1, Delta(4)-3-



統合データベースに関わるコンテンツの作成,整備 進捗報告

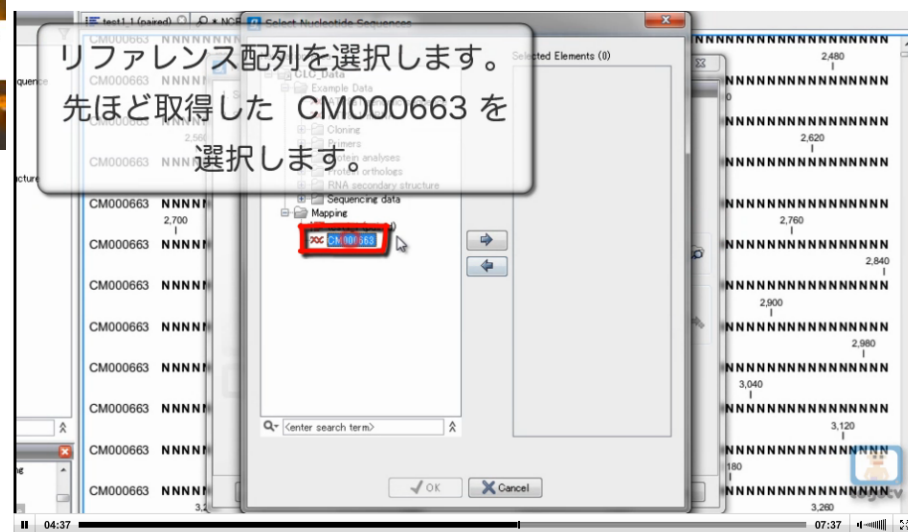
課題(7)

DBやツールの使い方がわからない
日本語のリソースが不足している

統合TV (togoTV)

- 動画によるDBやツールのチュートリアル
 - 各DBやツール名、で検索
- 統合データベース講演会AJACSの動画も
- You からも
- Videocastでも 
- 約650のコンテンツ
- (アップデート含む)

 [NGS][ゲノム][塩基配列][多型情報][配列解析][発現情報] CLC Genomics Workbench でショートリードのマッピングを行う
CLC Genomics Workbench は CLC bio 社によって提供されているゲノム解析ツールの商用パッケージです。
生命科学データの解析に使われるいろいろなツールを統合されたインターフェースで扱うことができます。
今回は NGS データの解析に欠かせないショートリードのマッピングの方法を紹介します。
導入方法、基本操作については過去の動画を参照してください。



ライフサイエンス新着論文レビュー

<http://first.lifesciencedb.jp/>

Nature, Science, Cell などのトップジャーナルに掲載された日本人を著者とする生命科学分野の論文について論文の著者自身の執筆による専門分野の異なる生命科学研究者にむけた日本語によるレビューを,自由に閲覧・利用できるようウェブ上にいち早く無料で公開するサイト



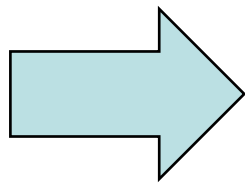
「ライフサイエンス新着論文レビュー FirstAuthor's を使い倒す」
<http://togotv.dbcls.jp/20110301.html>

2012年、新着論文レビューは...

- 162本を公開(計画書「月平均12本年間で150本程度の継続的な公開を維持」)
- 論文の出版(オンライン出版の場合も含めて)からレビューの公開までの平均所要日数 24.43日(目標1か月以内)
- 月間ユニークユーザー数は2万以上
- 公式ツイッターフォロワー数は3000以上

学会と共同で日本語レビューの オンラインジャーナルの立ち上げ

日本分子生物学会	日本蛋白質科学会
日本細胞生物学会	日本植物生理学会



4つの学会の協力のもと、9月に
「ライフサイエンス 領域融合レビュー」
公開スタート

「ライフサイエンス 領域融合レビュー」とは

- 生命科学分野の学問分野・領域を広く総合的にとりあげる日本語による総説(レビュー)を, だれでも自由に閲覧・利用できるよう, Webにて無料かつ継続的に公開するもの
- 学会からは編集委員を推薦, 10名の編集委員により編集委員会を構成し, 執筆者・執筆テーマおよび査読者の選定など, 学術面を担当
- DBCLSは, 編集・制作・公開など, 実務面・制作面を担当
- 2012年は4か月で8本のレビューを公開. 2013年度は30本程度を公開予定



「ライフサイエンス領域融合レビューLeadingAuthor'sを使い倒す」

<http://togotv.dbcls.jp/20121113.html>

LEADING AUTHOR'S

幹細胞の多能性を規定する分子機構 7 users

2012年12月11日

丹羽 仁史

(理化学研究所発生・再生科学総合研究センター 多能性幹細胞研究プロジェクト)

email:丹羽仁史

領域融合レビュー, 1, e008 (2012) DOI: 10.7875/leading.author.1.e008

Hitoshi Niwa: **Molecular mechanism governing pluripotency of stem cells.**

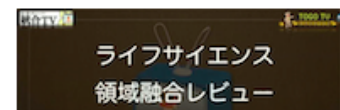
 PDF Download

要約

マウスの着床前の胚から樹立されたマウスES細胞は、サイトカインのひとつLIFに依存的に多能性を維持している。このLIFシグナルは複数の転写因子の制御を介し多能性に関連する転写因子ネットワークに入力している。これとは並列に、液性タンパク質であるWntも多能性の維持に寄与するが、このWntシグナルはTcf3による*Esrrb*遺伝子の発現抑制を解除するというひとつの点において多能性に関連する転写因子ネットワ

このサイトについて

生命科学において注目される分野・学問領域における最新の研究成果について、第一線の研究者の執筆による日本語のレビューを、だれでも自由に閲覧・利用できるよう、無料で公開します。くわしくは、「領域融合レビュー」とはをご覧ください。



統合TVにて解説動画を公開しました。

こちらをご覧ください



FOLLOW ME ON [twitter](#)



MBSJ

ホーム

口演

ポスター発表

夜ゼミ

ようこそ 坊農 秀雅 さん ▼

演題、要旨、名前、メール



日

En

セッション 3W2III : データベースから始まる分子生物学~研究の新しいスタイルを模索する~



3W2III-1 12月13日 15:45 第2会場 (地図)



データベースから始まる分子生物学~トランスクリプトーム解析研究の新しいスタイル~

○坊農 秀雅

情シ機構・ライフサイエンス統合DBセ

@ bono@dbcls.rois.ac.jp

要旨

DNAマイクロアレイや新型シーケンサに代表される大規模解析は、かつては大型プロジェクトに限定されていたが、コストダウンやコモディティの整備がすすみ、一般のウェット系研究室でも無理なく実行可能になった。また我が国ではまだ不十分であるが、今後、解析情報の共有と公開原則の意識が浸透すれば、インターネット上に公開データの蓄積がすすみ、それを利用することにより、解析コストをかけずに大規模データを研究に資することができるようになるだろう。しかしこうした実験のデータ量は膨大であり、そのデータハンドリングは実験生物学者には困難であることが現状の問題と言える。

演者が所属している情報・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)では文部科学省統合データベースプロジェクトの一環として、遺伝子発現バンクNCBI GEOに登録されているデータを測定技術と材料の属性に基づい

て整理するウェブツール「GEO目次」を開発し、維持してきた(<http://lifesciencedb.jp/geo/>)。その結果、特定の生物種や発現解析プラットフォーム、キーワードで遺伝子発現データを検索可能だけでなく、生物学的な興味に基づいて種横断的にデータが俯瞰できるようになってきている。

本講演では、演者自らの個別研究として進めてきた、公共データをフル活用した低酸素刺激前後のトランスクリプトームの生物種横断的比較解析を、GEO目次を駆使するその解析手法に力点を置いて研究の事例を紹介する。GEO目次のRSS機能を活用することで新規に登録されるデータを監視し続けるなど、データベース統合化によって可能となっているちょっとしたデータハンドリングテクニックに関しても触れる予定である。

本講演が実験生物学者の情報技術的な自立を促すためのきっかけとなり、活発に議論されることを期待している。

演題の関連性のネットワーク図



左のネットワーク図は、すべての要旨に対して日本語テキストマイニングの手法を用い、生命科学分野専門用語のキーワード抽出を行ってそれらの関連性に基づいて近い要旨を線で結び可視化したものです。

筆頭著者の名前が各要旨のラベルとなっております。

詳しくは実際に解析してくれたDBCLSの学生アルバイトさんの作業ブログ（tabris2012の不定期日記、鳴いて血を吐く時鳥）を参照して下さい。

サムネイル図をクリックすると拡大できます。



利用技術のフィードバック

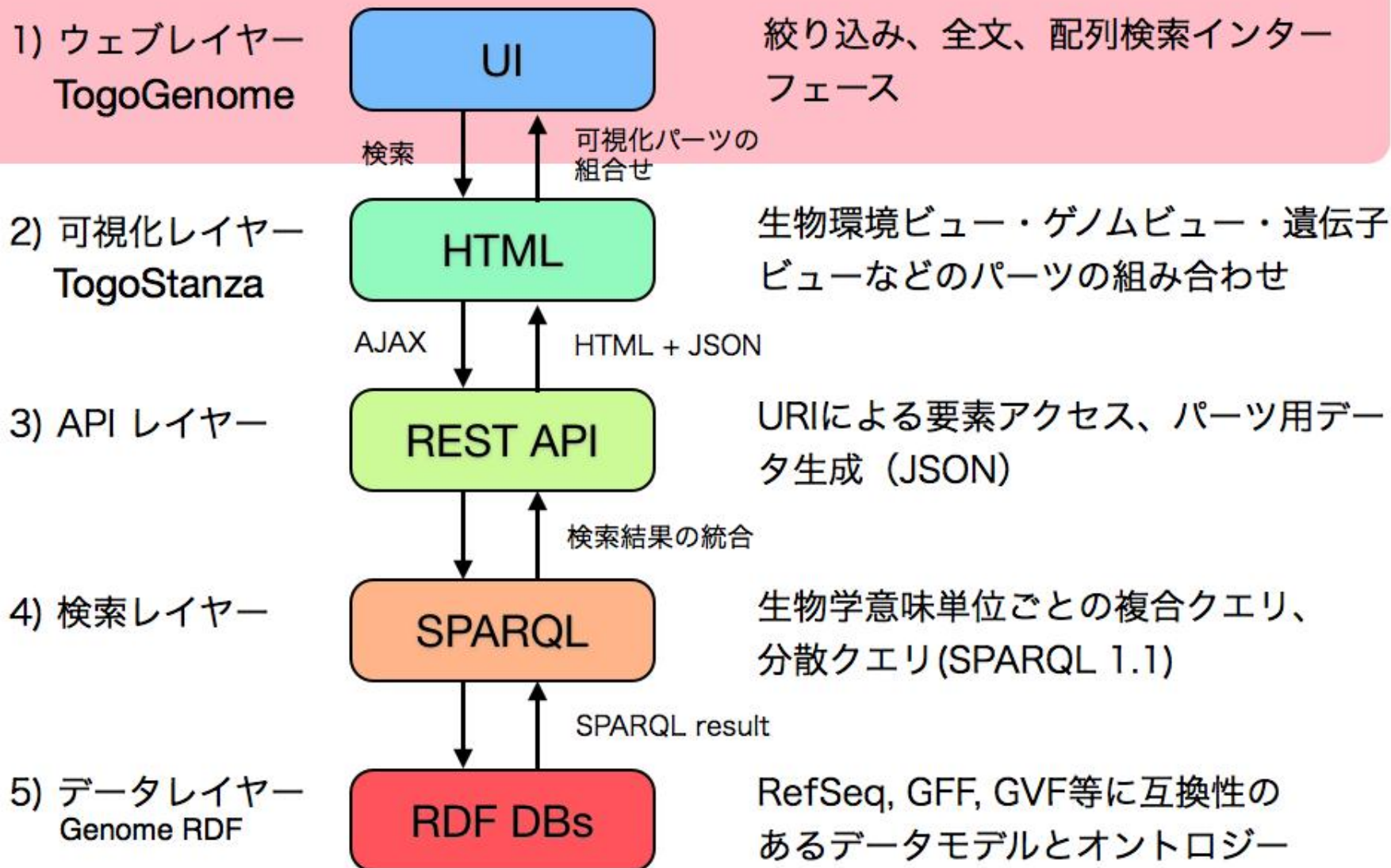
- 利用技術を開発するだけでなく、普及・フィードバックが得られるよう努めました
- 全国各所での講習会、講演活動
 - 統合DB講習会AJACS
 - IPAユーザー会、生命科学夏の学校、同志社大(京田辺)、明治薬科大(東京)、CERI(埼玉)、岡山大(倉敷)、分子生物学会年会(福岡)
- 共同研究
 - 酸化ストレス、低酸素関連研究

TogoGenome: ゲノムを軸とした RDF 統合環境

ゲノムデータベースを RDF化する背景と意義

- 生命科学の情報基盤→リファレンスゲノム情報の整備
 - 多種多様な知識をゲノム配列上に統合して整理可能
- オミクス実験の高速化技術革新→データ統合が低コスト
 - DBセンターや研究機関においてゲノム情報の増加に対応するコスト削減の要求
 - 研究者が自分のデータやアノテーションをゲノム基本情報と併せて閲覧したいという要求
- 幅広いユーザ像が設定できる→データの網羅性, 均質性, 再利用性
 - 生物学者: プログラミングしない
 - 生物学者: 簡単なプログラミングをする
 - 情報生物学者: 網羅的で均質なデータを利用する
 - データベース管理者: データベースの管理運営をする
 - アノテーター: 知識をデータベースに付加する

ゲノムデータベースを構成するための技術階層

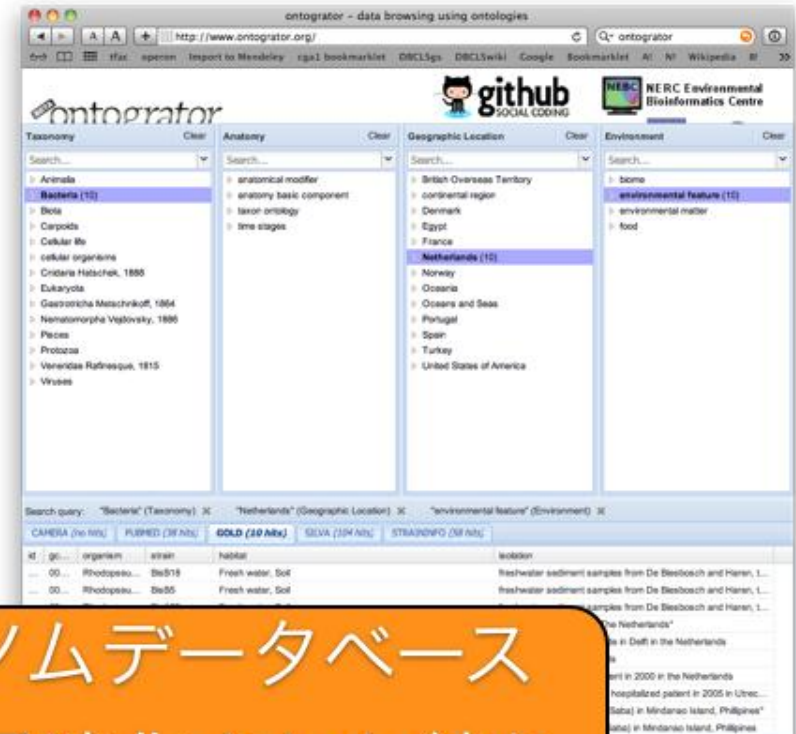
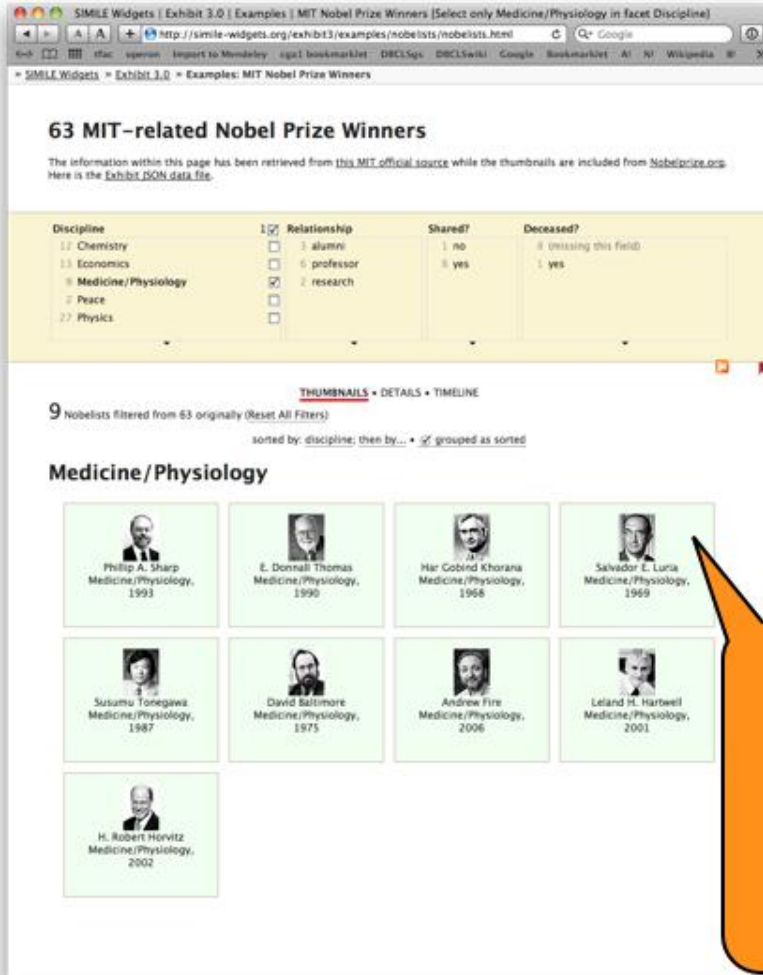


UI

1) ウェブレイヤー ← 検索インターフェース
絞り込み検索、文字検索、配列検索などのインターフェースを開発中

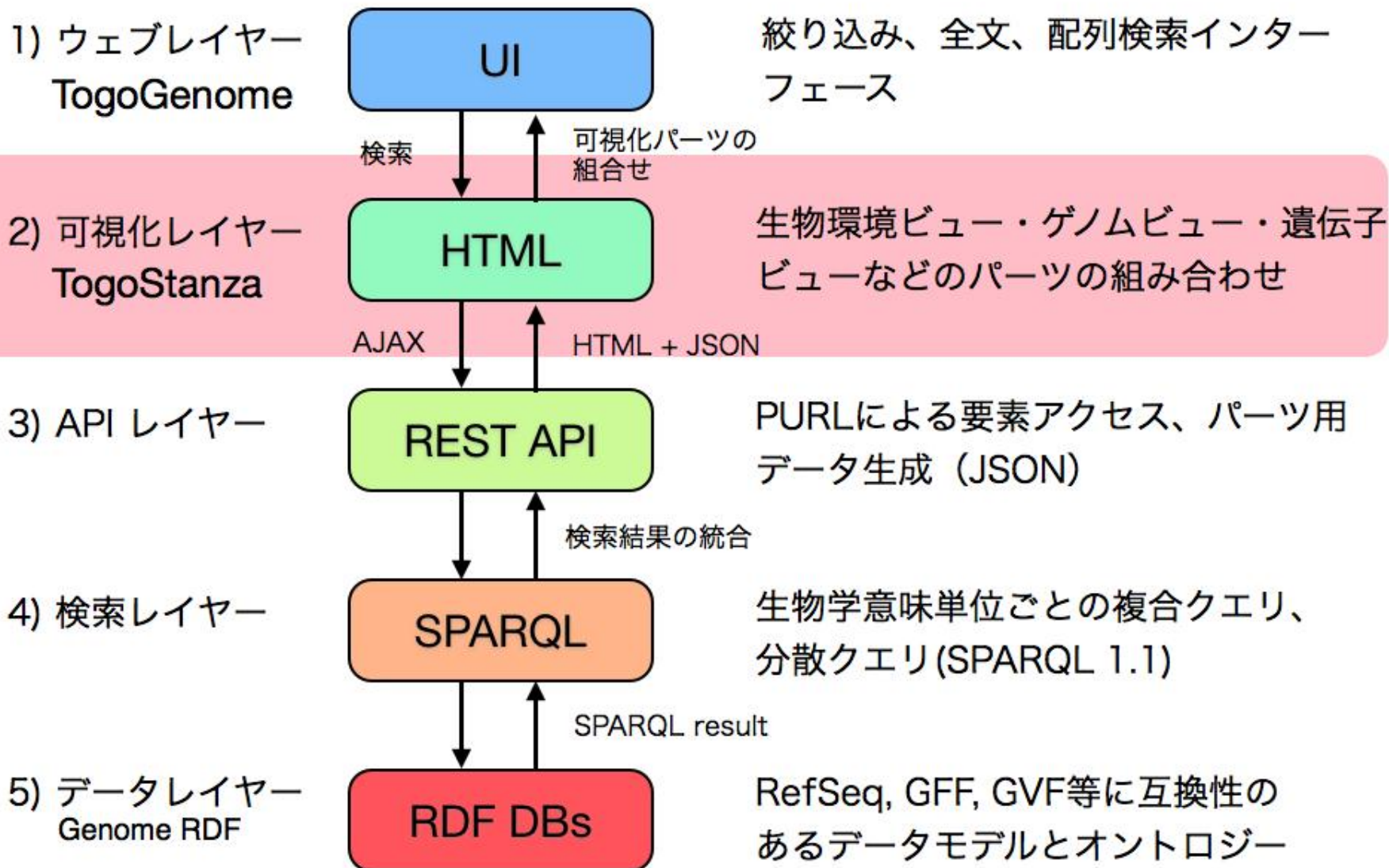
Exhibit 3.0

ontogrator



ゲノムデータベース
に最適化された絞り
込み検索

ゲノムデータベースを構成するための技術階層



2) 可視化レイヤー ← 生物環境ビュー・ゲノムビュー・遺伝子ビューなどのパーツの組み合わせ

典型的なゲノムデータベースを構成するパーツ（Stanza）の調査および整理

典型的なゲノムDBの構造

1. General summary of organisms
2. General summary
3. Genomic context
4. Transcript attributes
5. Protein attributes
6. Protein-protein interactions
7. Visual annotation format
8. Table annotation format
9. Homologs (in NR)
10. Homologs (in selected orgs)
11. Gene annotation (GO etc.)
12. References
13. External links
14. Mutants
15. Revision history



HTML

2) 可視化レイヤー ← 生物環境ビュー・ゲノムビュー・
遺伝子ビューなどのパーツの組み合わせ

既存のゲノムデータベースを参考にしながらパーツ (Stanza)
をデザインする

2. Summary of genes

TogoStanza

Summary		?	+	-	Top
ID	slr0473				
Type	Protein Coding				
Gene Symbol	<i>cph1</i> , <i>hik35</i>				
Gene Indexing gname	<i>cph1</i> , <i>hik35</i> , <i>phy</i>				
Definition	cyanobacterial phytochrome 1, two-component sensor histidine kinase				
Functional category	Regulatory functions				
Genes and Mutants	Gene: by Teruo Ogawa (1998-04-15) Gene: phy by Masahiko Ikeuchi (1999-09-17) Gene: phy by Masahiko Ikeuchi (1999-09-17) Gene: Hik35 by Iwane Suzuki (2000-05-25) Mutant: Δphytochrome by Teruo Ogawa (2003-05-08) Mutant: SK35 by Teruo Ogawa (2003-05-08) Mutant: Δhik35 by Iwane Suzuki (2007-11-01)				

UniProtKB

Names and origin

Protein names	Recommended name: Phytochrome-like protein cph1 EC=2.7.13.3 Alternative name(s): Bacteriophytochrome cph1 Light-regulated histidine kinase 1
Gene names	Name: cph1 Ordered Locus Names:slr0473
Organism	Synechocystis sp. (strain ATCC 27184 / PCC 6803 / N-1) [Complete proteome] [HAMAP]
Taxonomic identifier	1148 [NCBI]
Taxonomic lineage	Bacteria · Cyanobacteria · Chroococcales · Synechocystis



HTML

2) 可視化レイヤー ← 生物環境ビュー・ゲノムビュー・
遺伝子ビューなどのパーツの組み合わせ

パーツ (Stanza) を容易に再利用できるゲノムデータベースの提供

RDF genome database prototype
(<http://rdfgenome.dbcls.jp/>)

```
<!-- nucleotide_attributes -->
<div id="dbcls_nucleotide_attributes" title="syn:slr0473" class="dbcls_genome"></div>
<script type="text/javascript" src="http://rdfgenome.dbcls.jp/javascripts/dbcls_nucleotide_attributes.js"></script>
```

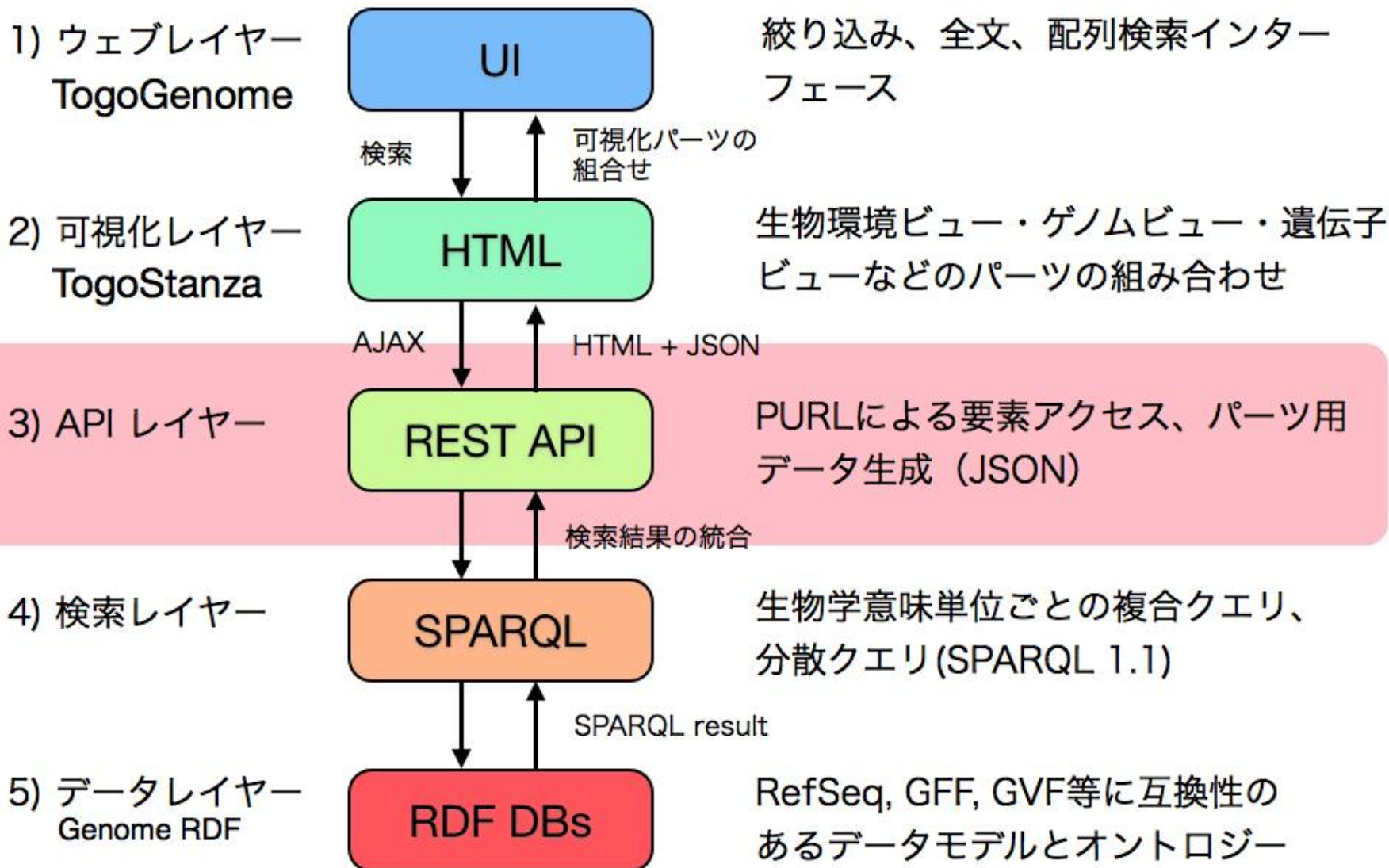
<!-- nucleotide_attributes -->

```
<div id="dbcls_nucleotide_attributes" title="" >  
    <script type="text/javascript" src="http://rd"></script>  
</div>
```

Genome Search	
Genome statistics	
Genome size	3.79
Sequence status	Complete
G+C content	47.4
Total number of chromosomes	5
Total genes	3166
rRNA genes	6
tRNA genes	41
Nucleotide sequences	Synechocystis sp. Synechocystis sp. Synechocystis sp. Synechocystis sp. PCC 6803 plasmid pSYSK DNA, complete sequence 106004 bp AP006555 PCC 6803 plasmid pSYSG DNA, complete genome 44343 bp AP004312 PCC 6803 plasmid pSYSB DNA, complete sequence 103307 bp AP004312 PCC 6803 plasmid pYSM DNA, complete genome 119695 bp AP004310
Sequencing institute	Kazusa DNA Research Institute
References	Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination and assignment of protein-coding regions. Kaneko T., Sato S., Kikuchi H., Tanaka A., Asamizu E., Nakamura Y., Miyajima N., Takeda M., Kohno J., Minamide I., Shimada T., Ohnishi R., Ohta C., Yanai H., Yoshida K., Hirao K., Matsuyama T. Genomics 30(3)(3):109-120 PMID: 8905231 Abstract Medline Related Articles Reference from UniProtKB
Genome names	
Genoma symbol	Recommended symbol: cph1
Genome descriptions	Recommended Phycochrome-like protein cph1 Alternative description: Bacterial phytochrome cph1 Light-regulated histidine kinase 1
Accession numbers	NCBI-GeneID: 16331509 NCBI-Protein ID: Q95260 UniProt: Q95168 Cybase: sinO473 GTPS: ST2165
Nucleotide attributes	
Nucleotide sequences	tcaagcttccatcattgtaacgcgaactttccaaggagtcttcagaaccttacaactctctctctcaactcacggcgaaattcatcaaat tgtagctggaccgaalcccattgatgatacttatgccgttgtggtttgttgcaaagctgatatcaaacctctgctaagataagatc attatagctcttgatfaccataaaaata-agtgcctcaaaaccaacacntgtaggcatactgatatccccctttgaaagatcatcatacda tppgaacctttcagtaacacgcacctbaaaagtaaataagatcgaccacaactgaacttgaagttatcaaataggpgactgaactgatagatag atcgcgtatcataaafgaactgtggcaccagaagaataaactattatgaattatcctggactgaactacatcaggaaggatttgatgcactttc atacgcattatgctattttgtatcttggmttaaactccacgcacgttcggttgaaag
Sequence length	2247
Sequence taking	Initial position: 2607812 Terminal position: 2610058
Protein attributes	
Amino acid sequences	MLDGEKVPVIRNQQGGFATSTFGALRWNIISGLTSDGRDVGWTKTHYVLQQLLAIVPEVVIAAPLAAQSQRCPALGKVITGLSTNSHACPRTL FDMPGTTRVDGDGFWSGALEAGAAVAIVSEKAHQPPNQNGSCVAVPOLVIPFCALAAPPQHQAQTLLQGVDTGNKTHTSHLEYLNQQRSSIAL LGTLTWIKRYPQVKTAHTTTTFATDLQGKLAEALQAGNVAYMVESHALAQGRVLCQQACAFVNRLTGCHLFPGHGMENYNFAAKALKFSYLQGRAVI NQDPPTGGRLDRPLDGQTVYTTSNDSTAOPYTXKDLYGYETVGKVTPTVFQBEPFLSPLVGGINLANVAAAISGSUHLGDPAAMVKOLLDFPVGPQM EQVGRPDQDSVMVDYANTPDFSLNALKAARFPITGRIQCIGCGGDRDKTRLPMGNASAIADLIAYVTSDNPRTDEPSGLADVVSQSSLDBEFWTI GORATAHKAREAKPGDVILAGKHGEDFYQLGETGHPCDDREDAEALLRYIS
Sequence length	748
Sequence status	Complete
References	
Red and far-red light alter the transcript profile in the cyanobacterium Synechocystis sp. PCC 6803: Impact of cyanobacterial phytotoxins Murata N, Börner T, Hubschmann T, Yamamoto H, Gielert Y FEMS Lett. 2005 Mar 14;(259(7)):1613-8.	

[illegible]

ゲノムデータベースを構成するための技術階層



REST API

3) API レイヤー ← URI による要素アクセス、パーツ用
データ生成 (JSON)

アクセスしたいデータセットが直感的に分かるURIの設計

<http://togogenome.dbcls.jp/<スタンザ名>/<ID>.<フォーマット>>

1. General summary of organisms

<http://togogenome.dbcls.jp/organism/1148.json>

2. General summary

http://togogenome.dbcls.jp/gene_attributes/1148.json

3. Genomic context

http://togogenome.dbcls.jp/genomic_region/1148:7229,8311.json

4. Transcript attributes

http://togogenome.dbcls.jp/transcript_attributes/1148:slr1311.json

5. Protein attributes

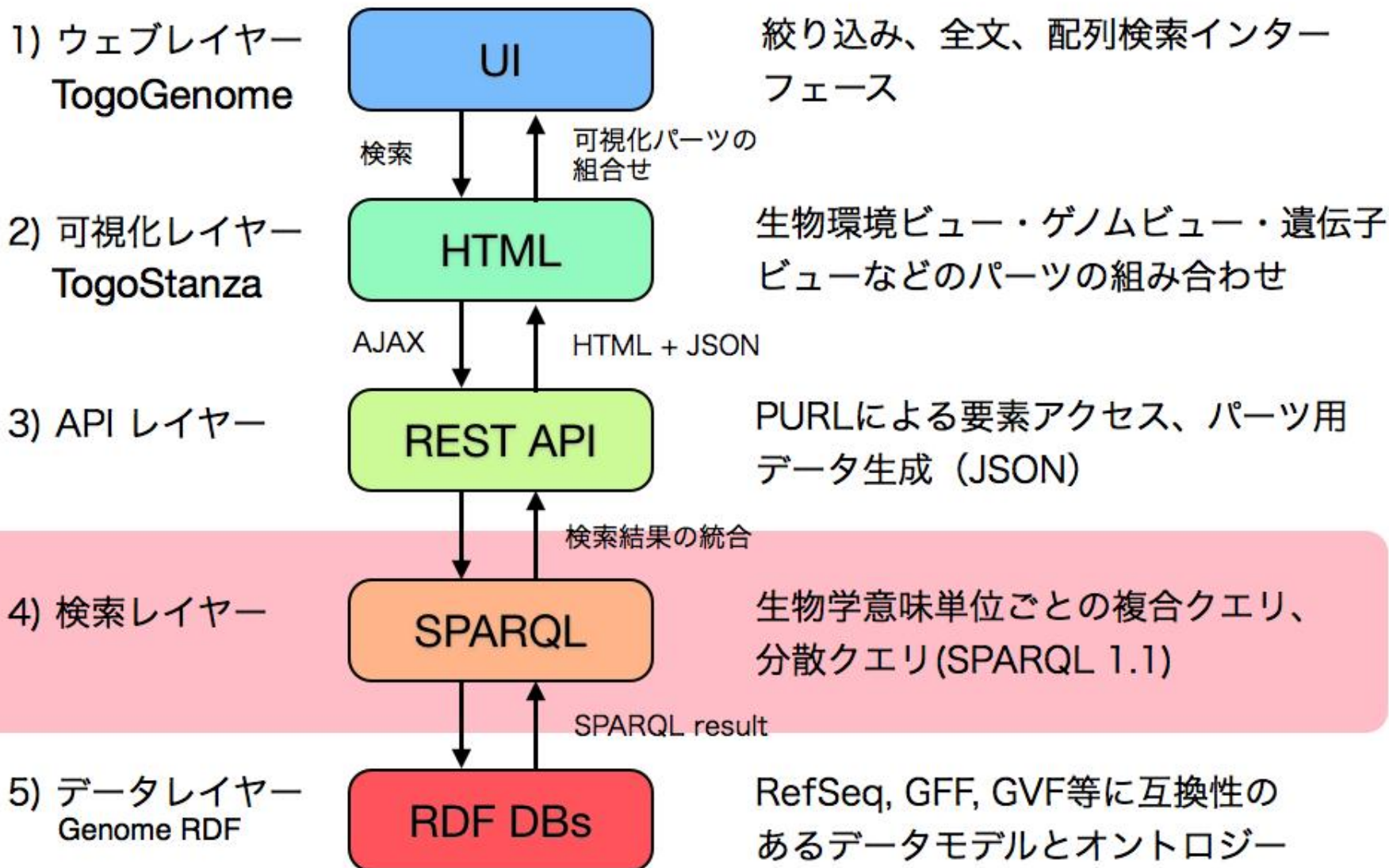
http://togogenome.dbcls.jp/protein_attributes/1148:slr1311.json

6. Protein-protein interactions

<http://togogenome.dbcls.jp/interactions/1148:slr1311.json>



ゲノムデータベースを構成するための技術階層



SPARQL

4) レイヤー ← パーツごとの複合クエリ、分散クエリ (SPARQL 1.1)

各パーツをつくるために必要なデータを分散したRDFデータベースから
取得するための問い合わせ文を設計する

4. Transcript attributes

[illegible]

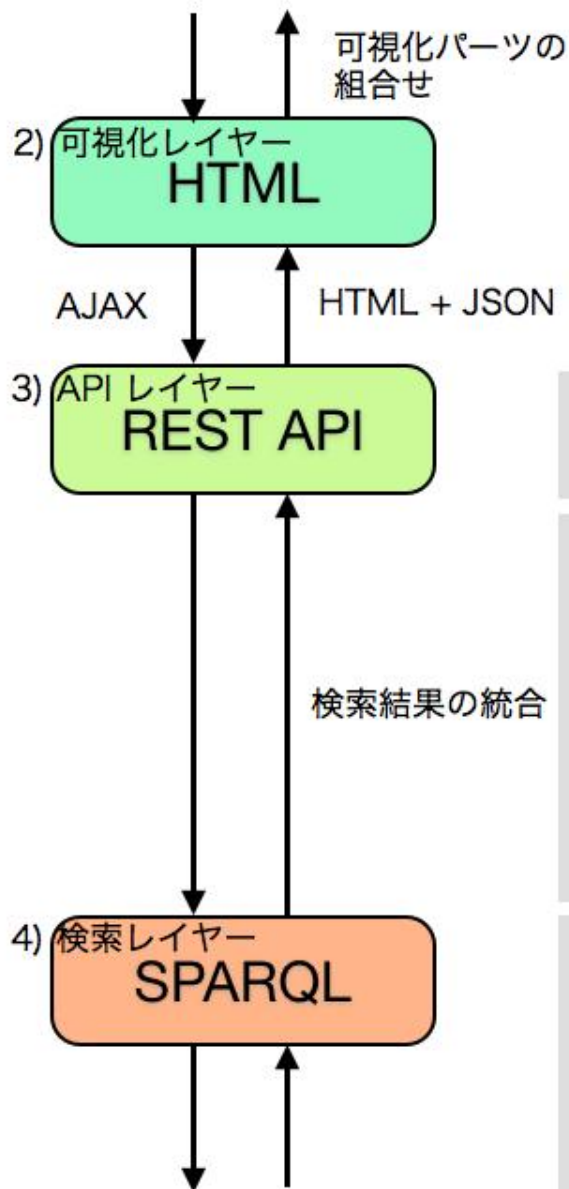
SPARQL

```
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX faldo: <http://biohackathon.org/resource/faldo#>
PREFIX insdc: <http://rdf.insdc.org/>
SELECT DISTINCT ?description ?genename ?aaseq ?location ?begin ?end ?up
WHERE {
  ?s insdc:feature_locus_tag "slr1311" .
  ?s insdc:feature_product ?description .
  ?s insdc:feature_translation ?aaseq .
  ?s insdc:feature_gene ?genename .
  ?s rdfs:seeAlso ?x .
  ?x rdf:type insdc:Protein .
  ?x rdfs:seeAlso ?up .
  ?s faldo:location ?l .
  ?l insdc:location_string ?location .
  ?l faldo:begin ?b .
  ?b faldo:position ?begin .
  ?l faldo:end ?e .
  ?e faldo:position ?end .
  ?s obo:so_part_of+ ?seq .
  ?seq rdf:type ?seqtype .
  ?seq rdfs:seeAlso ?tax .
  ?tax rdfs:label "taxon:1148" .
  ?tax rdf:type insdc:Taxonomy .
  FILTER (?seqtype IN (obo:SO_0000988, obo:SO_0000155))
}
```

生物学的な意味単位
再利用可能



実際のデータのながれと可視化



Nucleotide attributes	
Nucleotide sequence	tcaagatcttcattatgtattcacgaaatttctcaaaggaatcttcagaaccattccaaatctcttctccaattcagcggctaattttcataat tacgggtggcgccaatgcaattgttttagtatttgacttggttgattggtttgtacaaagatgcaatcttgactatagataggtatc attattagtccttcagtaaccataaaataagtgctccaaaaccacacattgtctaggcatagtgctatatcccttttgaaatgtatcataacta taggaacttctacgtaaccagcacctaaaaagtaataagatcgacccaaaactgaacttgagttatcaaataggggactaagacctgatagatg atccgtgatcataaatgaagctgtggcaccaggaatcaaatctattacagtatatcttcggactcagcatcattgaggaggtatttcgactcttc atacgcatattgtccattgttgatctgttttaaaattcacgagccctggcgtaag
Sequence length	2247
Sequence taking	Initial position 7229 Terminal position 8311

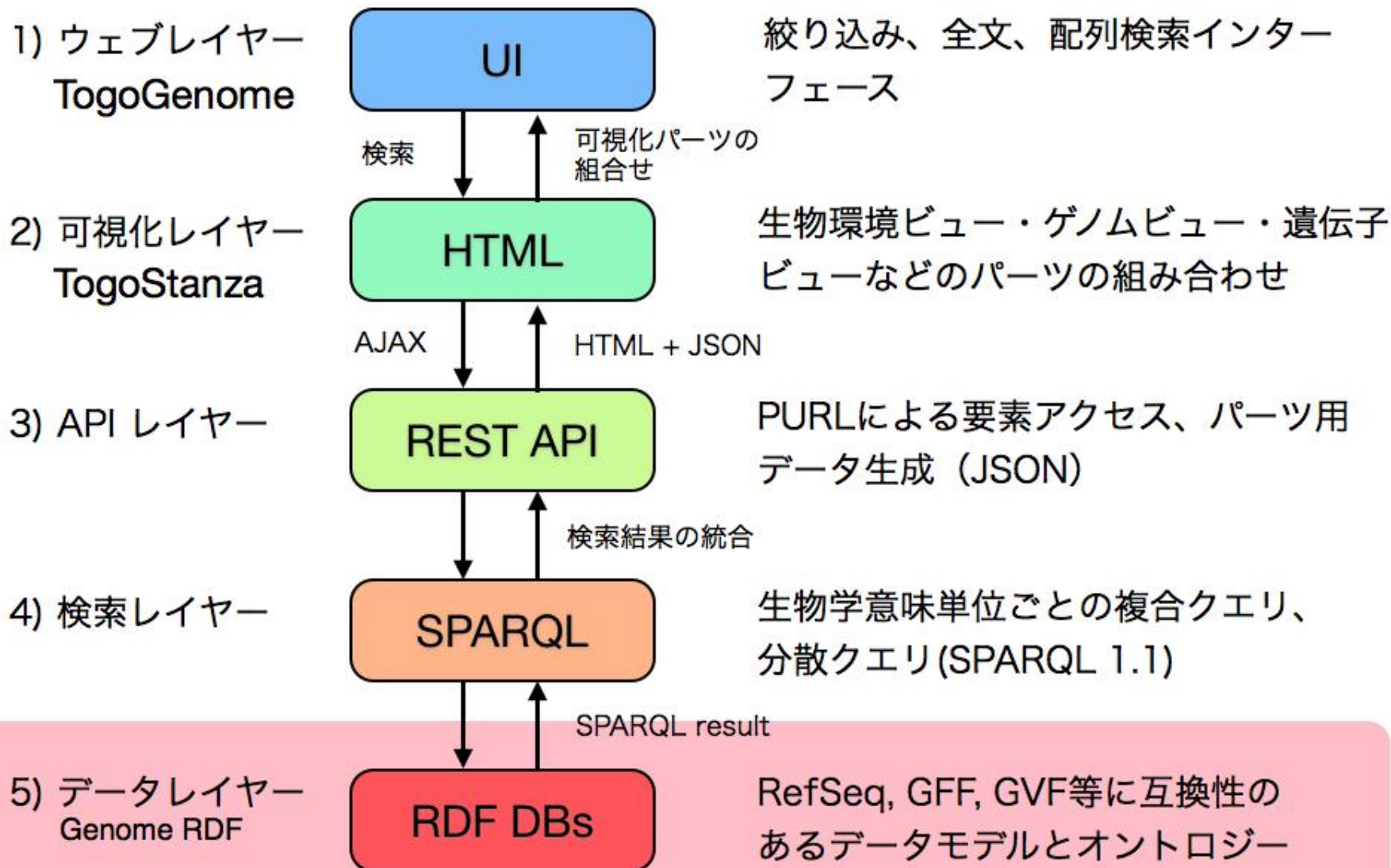
TogoStanza

http://togogenome.dbcls.jp/nucleotide_attributes/1148:slr1311.json

```
{
  "genename": "psbA2",
  "description": "photosystem II D1 protein",
  "nucseq": "tcaagatcttcattatgtattcacgaaatttctcaaagga...",
  "location": "7229..8311",
  "location_start": 7229,
  "location_end": 8311,
  :
}
```

```
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX faldo: <http://biohackathon.org/resource/faldo#>
PREFIX insdc: <http://rdf.insdc.org/>
SELECT DISTINCT ?description ?genename ?nucseq ?location ?begin ?end ?up
WHERE {
  ?s insdc:feature_locus_tag "slr1311" .
  ?s insdc:feature_product ?description .
```


ゲノムデータベースを構成するための技術階層



RDF DBs

5) データレイヤー ← GFF3/DAS, RefSeq 等に互換性のあるモデルとオントロジー

ゲノムの配列情報と多種多様なアノテーションデータを 個別のオントロジー、データ変換プログラムを開発し RDF 形式にして統合する

ゲノム配列

NCBI: BioProject/RefSeq -- 既存のリファレンス配列
DDBJ: Annotation pipeline/GTPS -- 新規ゲノム配列

実験・メタデータ

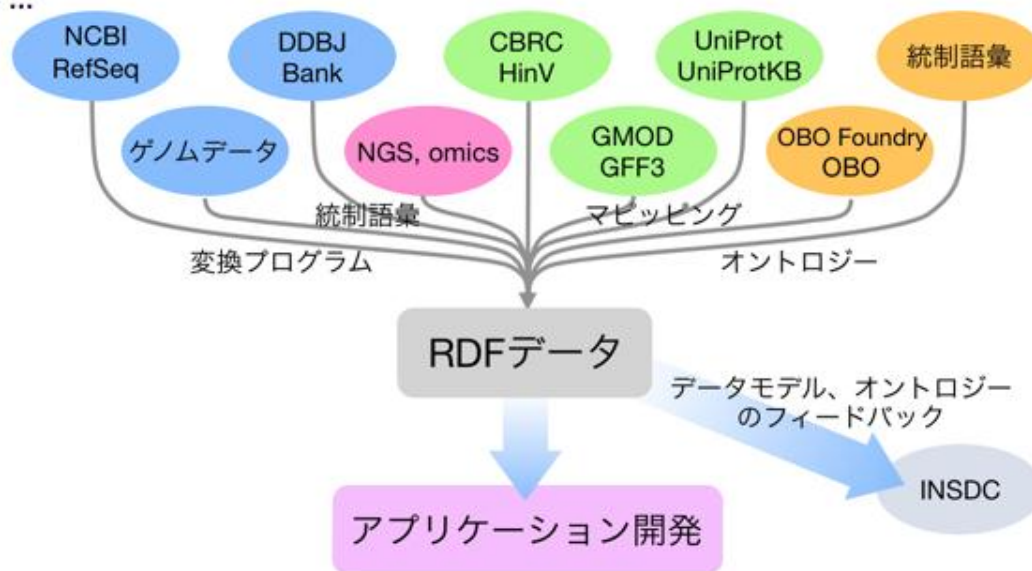
INSDC, NCBI: SRA, GEO
DBCLS: RefEx, Kusarinoko
GOLD, GSC: 環境メタデータ
Bulk data: 文献, 画像 ...

オントロジー

NCBO: BioPortal, OBO (GO, SO ...)
DBCLS: MEO, GMO, MCCV ...

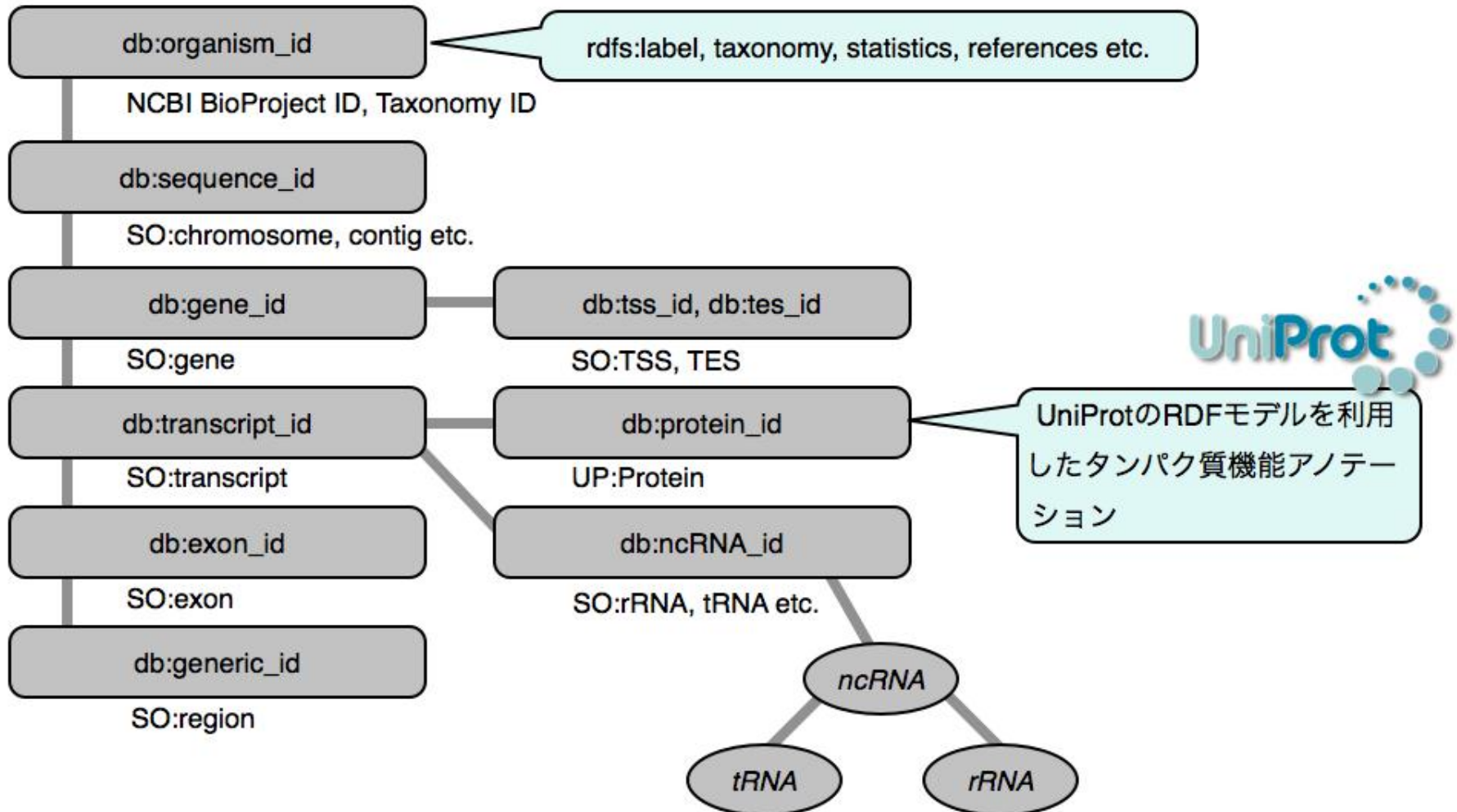
アノテーション

UniProt: Protein functions and links
Formats: GFF3, GTF, GVF, DAS, BED ...
Tools: Cufflinks, BLAST, InterProt ...



ゲノムRDFデータモデルの構造

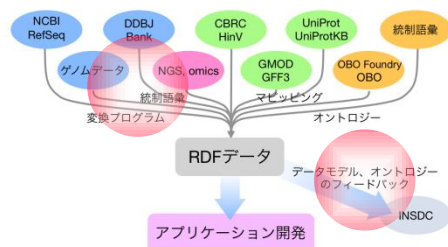
→URI によるゲノム情報要素の識別とオントロジーの利用



RDF DBs

5) データレイヤー ← GFF3/DAS, RefSeq 等に互換性のあるモデルとオントロジー

ゲノムRDFデータモデル開発成果

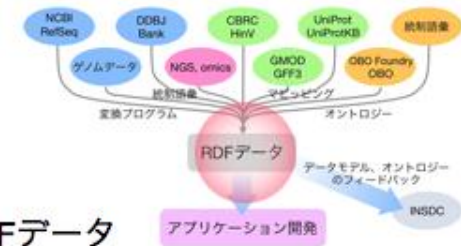


→ Biohackathon 2012, 国内開発会議 BH12.12

1. 全ての要素に URI (URL or URN) で ID (UUID) を付与
2. 配列アノテーションには Sequence オントロジー (SO) を使用
3. FeatureとQualifierには INSDC オントロジーを開発
4. 外部データベースのURIにidentifiers.orgを使用
5. 座標アノテーションには FALDO オントロジーを使用

RDF DBs

5) データレイヤー ← GFF3/DAS, RefSeq 等に互換性のあるモデルとオントロジー



リファレンスゲノムデータ (RefSeq) の RDF 化

リファレンスゲノムデータ

LOCUS NC_000911 3573470 bp DNA circular BCT 21-DEC-
DEFINITION Synechocystis sp. PCC 6803 chromosome, complete genome.
ACCESSION NC_000911
VERSION NC_000911.1 GI:16329170
DBLINK Project: BioProject
KEYWORDS .
SOURCE Synechocystis
ORGANISM Synechocystis
Bacteria

```
FEATURES             Location/Qualifiers
     source            1..3573470
                        /organism="Synechocystis sp. PCC 6803"
                        /mol_type="genomic DNA"
                        /strain="PCC 6803"
                        /db_xref="taxon:114842"

```

```

gene       7229..8311
           /gene="psbA2"
           /locus_tag="psbA2"
           /db="GenBank"
CDS        7229..8311
           /gene="psbA2"
           /locus_tag="psbA2"
           /db="GenBank"
           /codon_start=1
           /transl_table=1
           /product="Photosystem reaction center protein P680"
           /protein_id="P680"
           /db="GenBank"
           /db_xref="GenBank:U00096.1:7229-8311"

```

2. 配列アノテーションには
Sequence オントロジー (SO) を使用

3. FeatureとQualifierには INSDC オントロジーを開発

4. 外部データベースのURIに identifiers.orgを仕様

5. 座標アノテーションには FALDO
オントロジーを使用

1. 全ての要素に URI (URL or URN) で ID (UUID) を付与

```

#Prefix: sdc://rdf.incd.org/.
### S0:chromosome, S0:circular

<urn:uuid:182f171a-7928-4324-8d41-f3e820a872fd>
  rdf:type obo:S0_0000340, obo:S0_0000988 ;
  rdfs:label "Synechocystis sp. PCC 6803 chromosome, complete genome." ;
  insdc:sequence_date "2012-01-19"^^xsd:date ;
  insdc:sequence_length 3573470 ;
  insdc:sequence_version "NC_000911.1" ;
  insdc:source_organism "Synechocystis sp. PCC 6803" ;
  insdc:source_mol_type "genomic DNA" ;
  insdc:source_strain "PCC 6803" ;
  insdc:sequence_fasta <http://togows.dbcls.jp/entry/nucleotide/
NC_000911.1.fasta> ;
  faldo:location <urn:uuid:3b7e843c-bc24-4b63-a1bb-595591ff8e57> ;
  rdfs:seeAlso <http://identifiers.org/bioproject/57659>,
    <http://identifiers.org/ncbigi/16329170>,
    <http://identifiers.org/pubmed/8590279>,
    <http://identifiers.org/pubmed/8905231>,
    <http://identifiers.org/pubmed/9724772>,
    <http://identifiers.org/refseq/NC_000911.1>,
    <http://identifiers.org/taxonomy/1148> .

<urn:uuid:3b7e843c-bc24-4b63-a1bb-595591ff8e57>
  a faldo:Region ;
  faldo:begin <urn:uuid:3c0f2422-67a4-417e-8326-3672b0263e19> ;
  faldo:end <urn:uuid:57d42465-1747-4f64-bb3d-18c43616622c> ;
  insdc:location_string "1..3573470" .

<urn:uuid:3c0f2422-67a4-417e-8326-3672b0263e19>
  a faldo:ExactPosition, faldo:ForwardStrandPosition ;
  faldo:position 1 ;
  faldo:reference <urn:uuid:182f171a-7928-4324-8d41-f3e820a872fd> .

<urn:uuid:57d42465-1747-4f64-bb3d-18c43616622c>
  faldo:position 3573470 ;
  faldo:reference <urn:uuid:182f171a-7928-4324-8d41-f3e820a872fd> ;
  a faldo:ExactPosition, faldo:ForwardStrandPosition .

<http://identifiers.org/bioproject/57659>
  a insdc:BioProject ;
  rdfs:label "BioProject:57659" .

<http://identifiers.org/ncbigi/16329170>
  a insdc:GI ;
  rdfs:label "GI:16329170" .

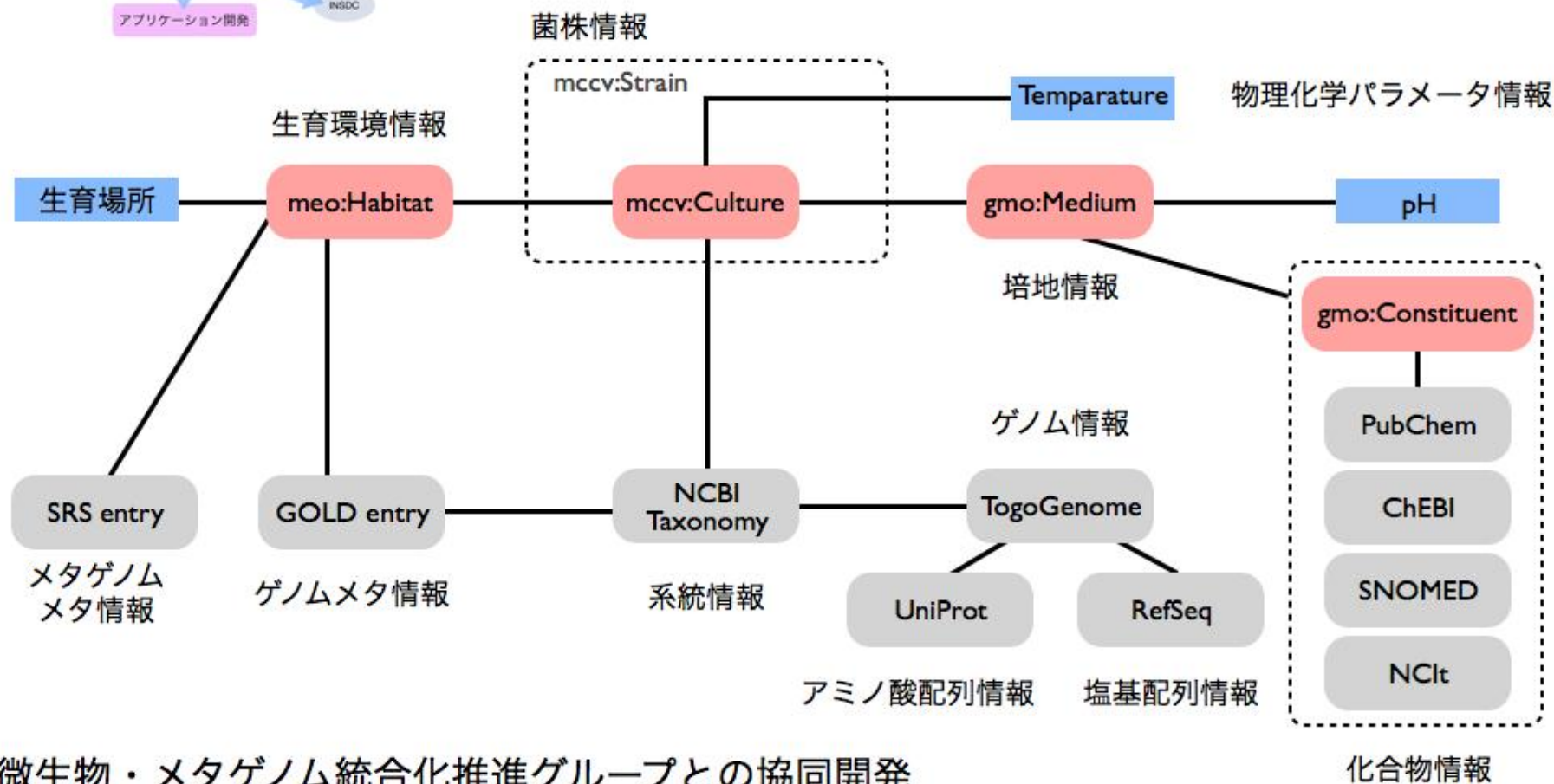
```

RDF DBs

5) データレイヤー ← GFF3/DAS, RefSeq 等に互換性のあるモデルとオントロジー



新規オントロジーを開発してRDFデータをつなげる



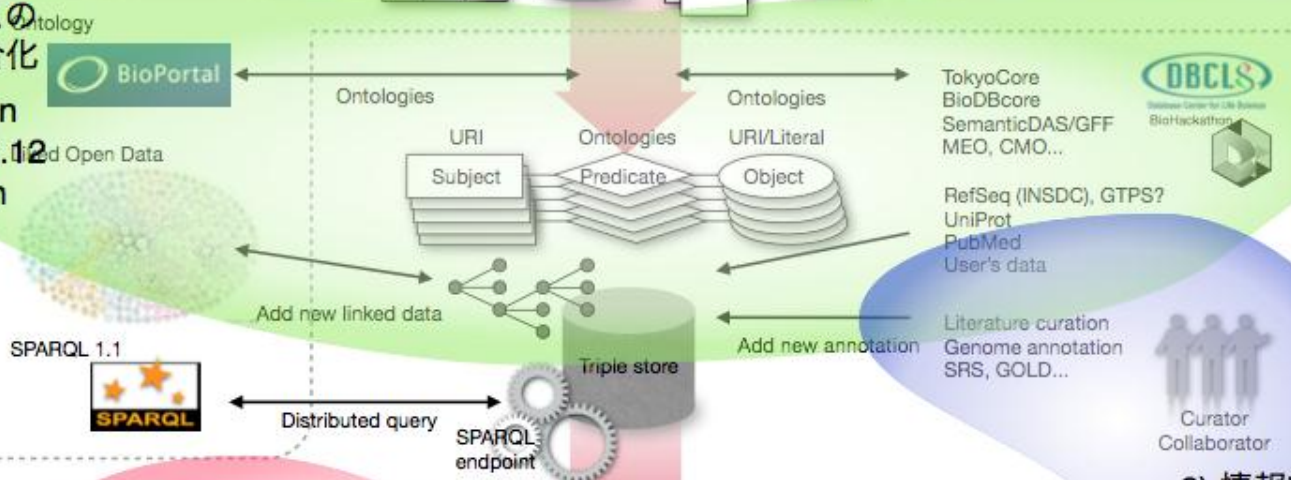
微生物・メタゲノム統合化推進グループとの協同開発

ゲノムデータベースのRDF化によるデータ統合

Legacy data

1) データベースのRDFによる統合化

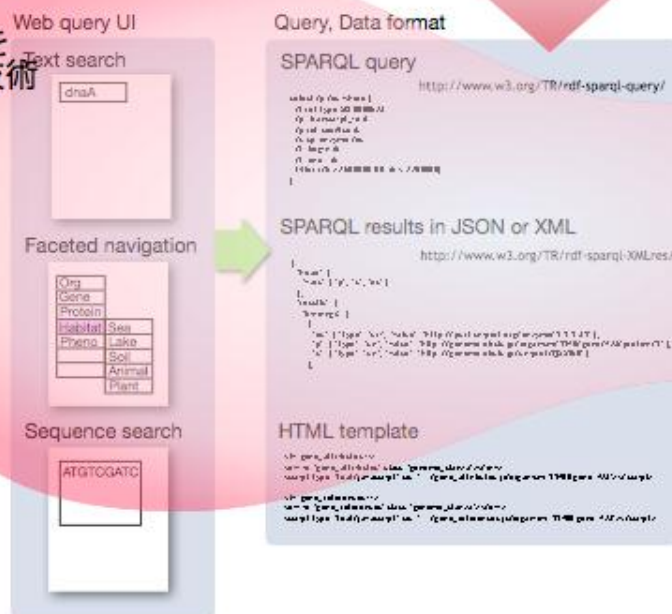
- BioHackathon
- BH.11.11, 12.12
- SPARQLthon



- RDF化
- オントロジー作製

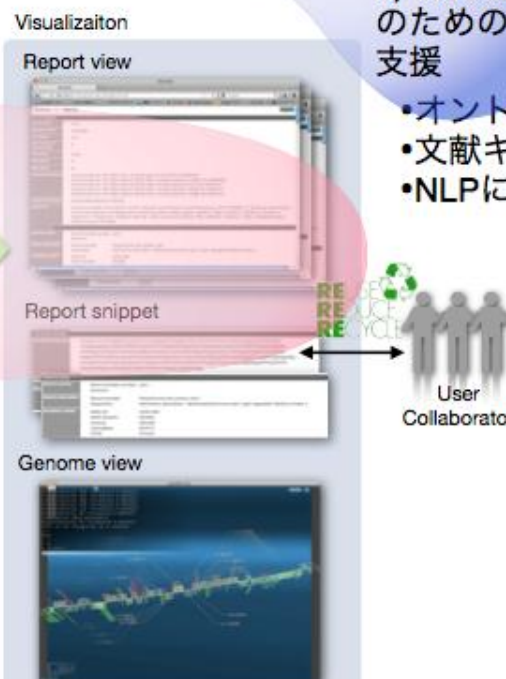
3) インターネットを活用した高度検索技術の開発

- 標準化
- 可視化



6) 情報統合化・知識発見のためのキュレーション支援

- オントロジーへのマッピング
- 文献キュレーション
- NLPによる作業支援



TogoGenomeの現状と方向性のまとめ

- GTPS と UniProt の RDF データを利用したプロトタイプ開発
- 国際会議 BioHackathon で座標情報のオントロジー FALDO を策定
- 微生物ゲノムDB 用にゲノムの RefSeq エントリ RDF 変換を完了 (42GB, 4億トリプル)
- スパコンを活用したトリプルストアでのスケーラブルな運用と更新体制 (DDBJ 連携)
- 他のデータソースやツール出力の RDF 化とインポート (BioHackathon, Open Bio* 国際連携)
- SPARQL 開発、API 開発、スタンザ開発 (DBCLS, DDBJ, 東工大, ベンダー連携)
- ユースケースに応じたこれまでにない柔軟な UI とファセットブラウザの提供 (国内各研究機関連携)

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でき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）を行う。

福井一彦、田代俊行、浅井潔

2013年1月22日

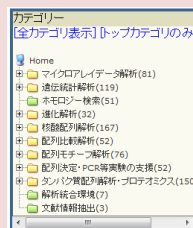
H24年度計画



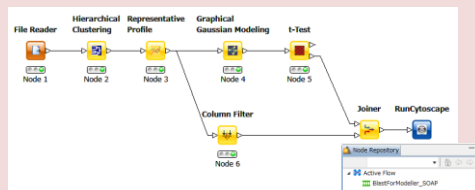
基盤技術開発

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



セマンティックWEB
RDF/OWL/SPARQL

統合DB

連携

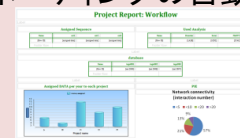
③ 解析・検索結果の可視化

解析結果及び解析ツール群の可視化



レポート作成機能

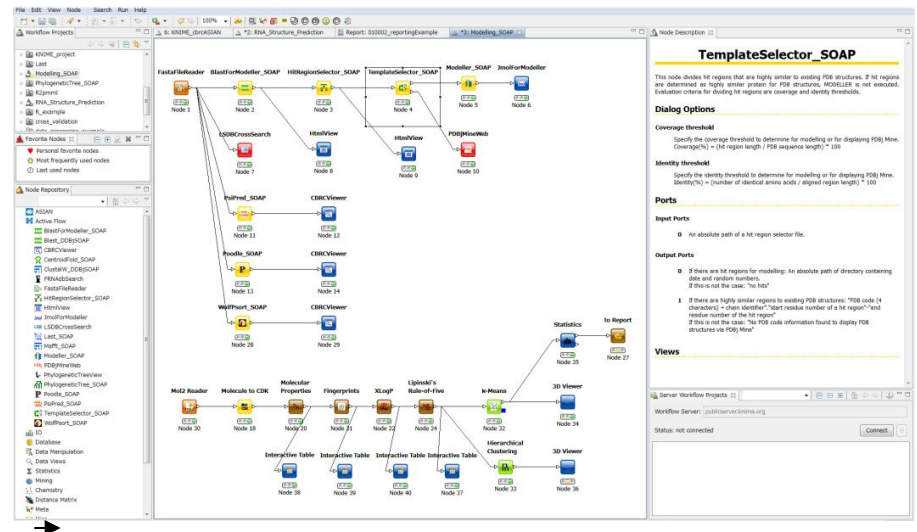
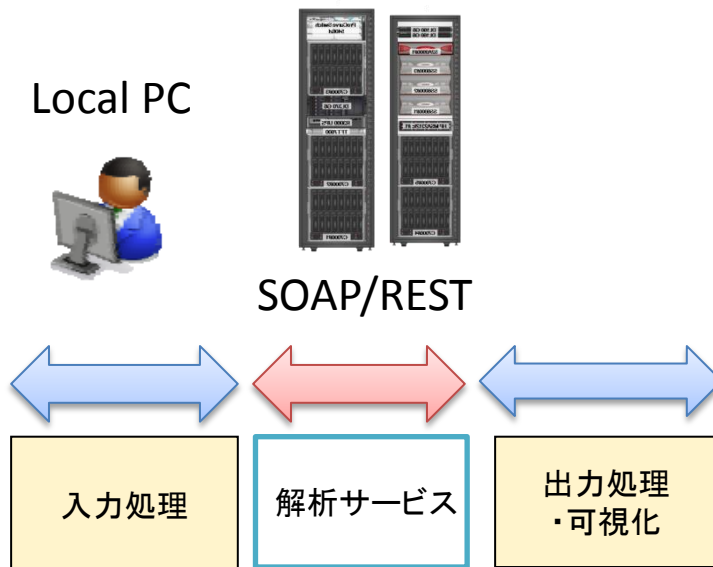
レポートの自動化



H24年度 進捗状況

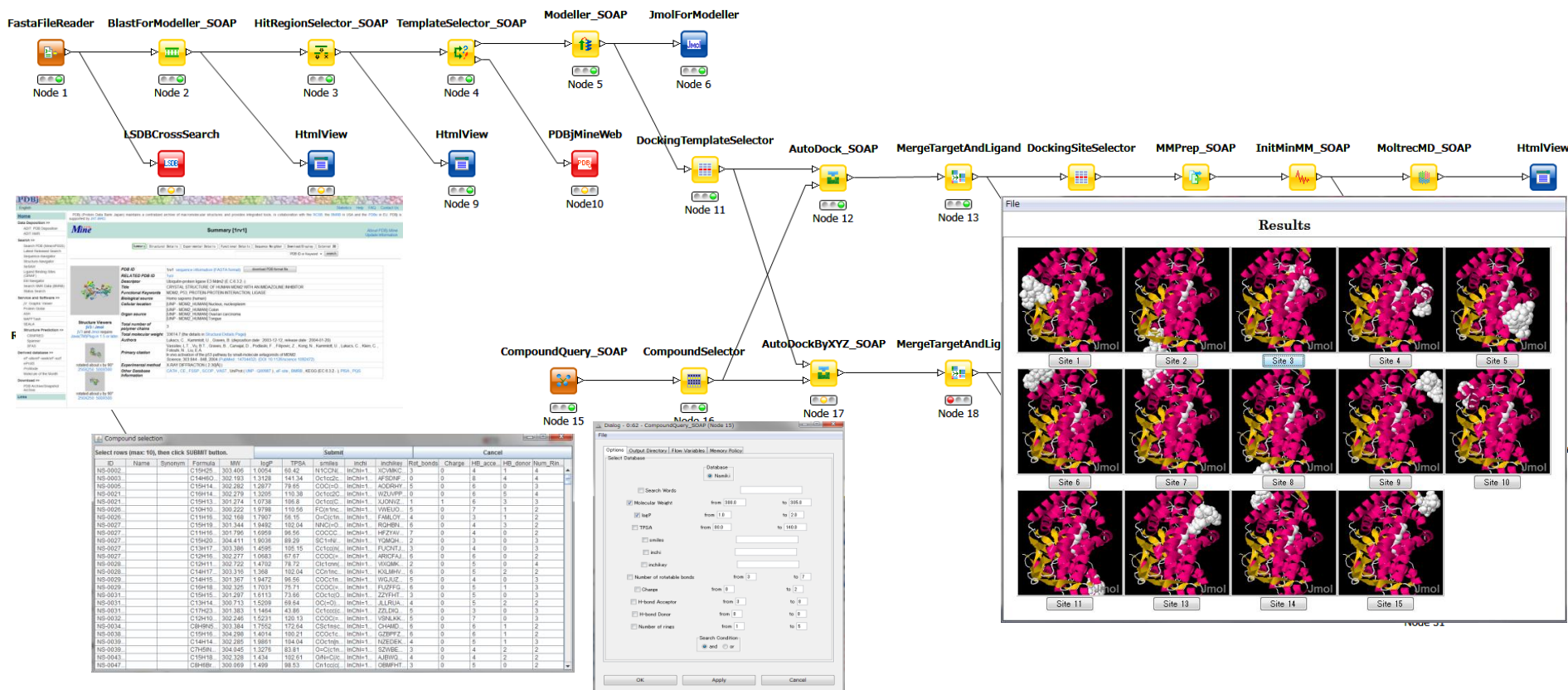
① ワークフロー統合・利用環境の構築・整備

- 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17個のKNIMEノードを公開
 - RNA解析の拡張
 - Molecular Simulation 解析の拡張
- 公開サイトのアップデート
- 利用者マニュアルの英語版作成



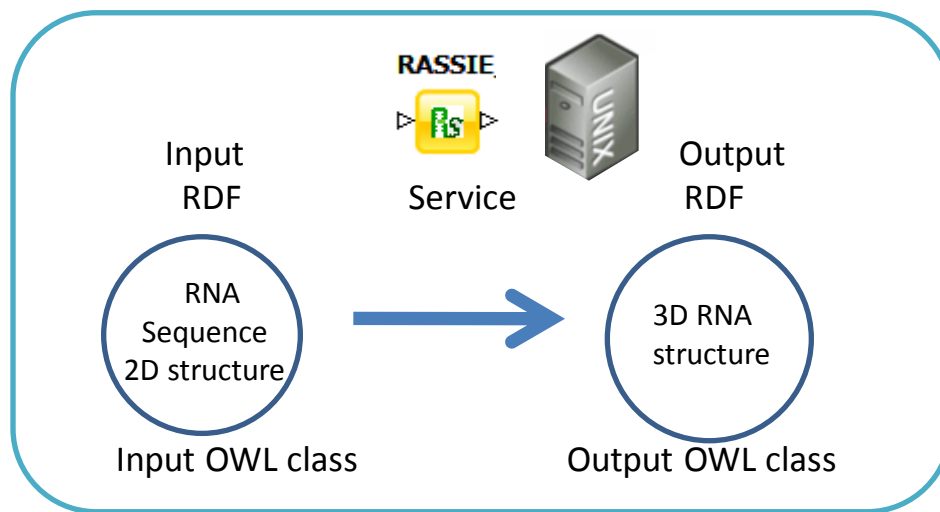
② プラットフォームとDB連携

- 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



<http://sadiframework.org/>



② プラットフォームとDB連携

- SADIサービスの公開
- 利用者マニュアルの日本語・英語版公開

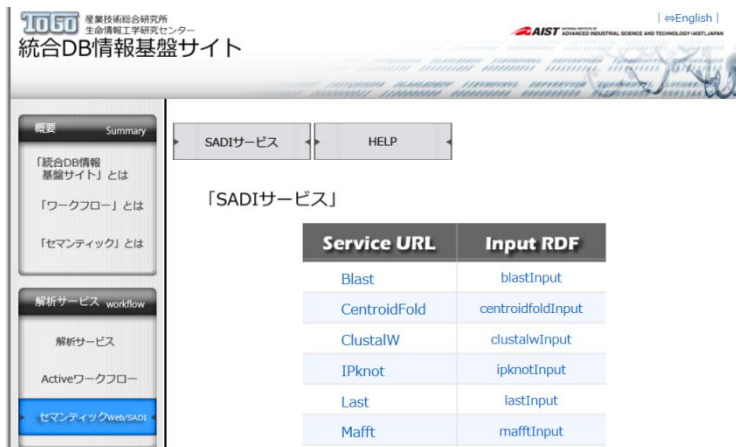
同期通信
(Sync service)

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

非同期通信
(Async service)

② プラットフォームとDB連携

SADI Services



```
% curl --data-binary @input-RDF http://semantic.cbrc.jp/sadi-services/ToolName -o output-RDF
```

```
<rdf:RDF
```

```
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:cbrc="http://togo.cbrc.jp/ontologies/cbrcswo.owl#">
```

```
<cbrc:WolfPsortInput rdf:about="http://togo.cbrc.jp/ontologies/wolfpsort.rdf#1">
```

```
<cbrc:requiresKingdomInformationOf>animal</cbrc:requiresKingdomInformationOf>
```

```
<cbrc:requiresQueryProteinSequence>
```

```
MANLGCWMLVLVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSP.....
```

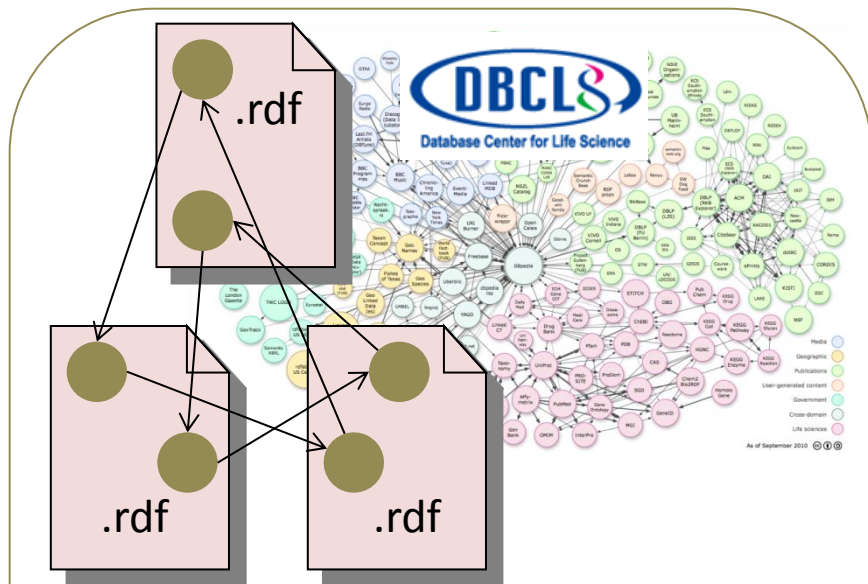
```
</cbrc:requiresQueryProteinSequence>
```

```
</cbrc:WolfPsortInput>
```

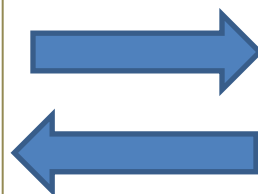
```
</rdf:RDF>
```

② プラットフォームとDB連携

- DBレポジトリ構築
- SPARQL検索との連携



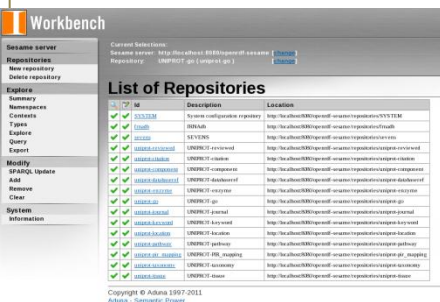
Service Input
RDF/OWL



Service Output
RDF/OWL

CBRC DBレポジトリ

- fncRNA
- Sevens
- UniProt
- KEGG from Bio2RDF
- PDBj

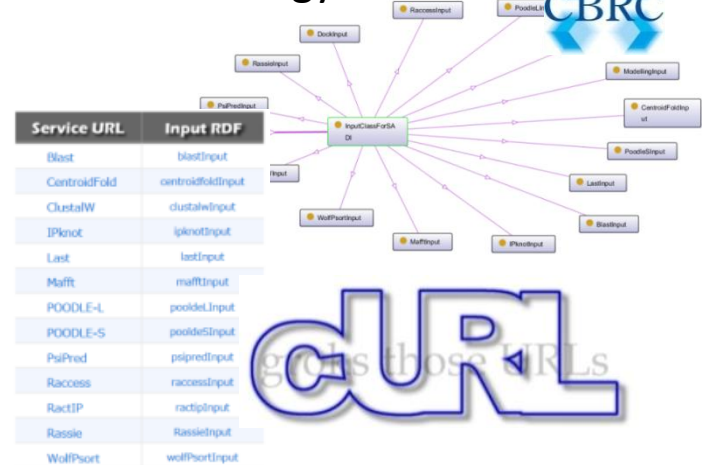


DBのレポジトリ (CBRC,DBCLS,Bio2RDF)



©2013 産業技術総合研究所 生命情報工学研究センター licensed under CC表示2.1日本

SADI用Ontology



Sparql

SequenceSelector



Node 1

Node 2

開発中

SADI Services

③ 解析・検索結果の可視化

解析ツールのオントロジー作成

ゲノム解析ツール リンク集 - Category

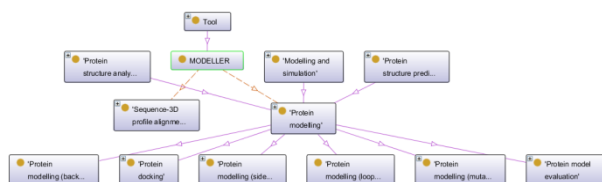
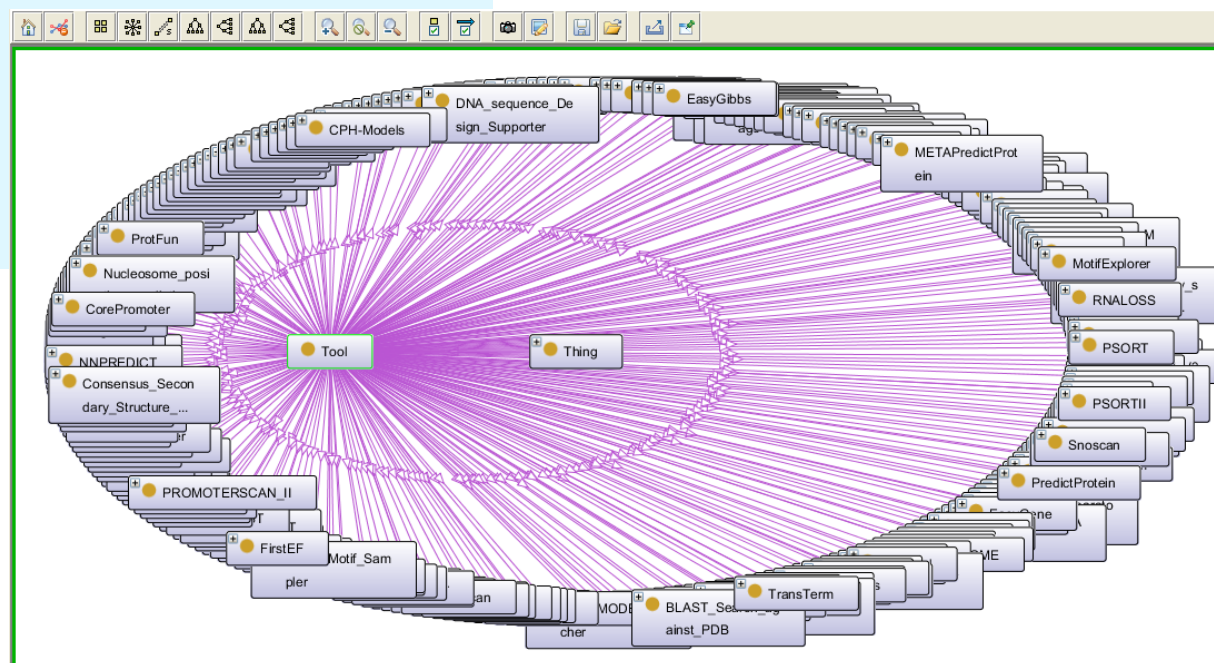
[データベースの説明](#) | [データ項目の説明](#) | [ダウンロード](#) | [利用許諾](#) | [ヘルプ](#)

[Category](#) || [Tool](#) | [How-to](#) | [Review](#) || [Keyword](#)

分子生物学に関わるデータ解析ツールの提供サイトへのリンクと簡単な解説を提供します。
(以前のバージョンをご覧になりたい方は、こちらのリンクをクリックしてください。)

カテゴリ [すべて開く](#) [すべて閉じる](#)

- マイクローレイデータ解析 [\[Tool \(80\)\]](#) [\[How-to \(1\)\]](#)
- 遺伝統計解析 [\[Tool \(117\)\]](#)
- ホモロジー検索 [\[Tool \(50\)\]](#)
- 進化解析 [\[Tool \(31\)\]](#) [\[How-to \(1\)\]](#)
- 核酸配列解析 [\[Tool \(158\)\]](#) [\[How-to \(2\)\]](#) [\[Review \(4\)\]](#)
- 配列比較解析 [\[Tool \(51\)\]](#)
- 配列モチーフ解析 [\[Tool \(74\)\]](#)
- 配列決定・PCR等実験の支援 [\[Tool \(51\)\]](#) [\[How-to \(2\)\]](#)
- タンパク質配列解析・プロテオミクス [\[Tool \(147\)\]](#) [\[Review \(3\)\]](#)
- 解析統合環境 [\[Tool \(6\)\]](#)
- 文献情報抽出 [\[Tool \(2\)\]](#)



H25年度実施計画

1) ワークフローの整備

H24年度に引き続き、有用な既存ツールや本研究グループが独自に開発したツールをKNIMEのプラットフォーム上で動作するようにノード化する。またこれまで公開したワークフローや開発ノードの整備を行う。

2) ワークフロー環境での結果レポート表示機能開発

これまで開発したKNIMEのプラットフォーム上で動作する各種解析結果を、自動でレポート作成可能とする機能開発を行う。

3) 解析プラットフォームへのRDF入出力機能追加

セマンティック技術に対応した解析プラットフォーム(KNIMEやGALAXY)とするため、SADIフレームワークを利用し、RDF化されたデータレポジトリ(エンドポイント)と連携して動作する機能開発を行う。

4) 解析ツールのオントロジー作成

KNIMEやSADIサービスとして開発した解析ツールや既存の解析ツール・ソフトウェアに関するオントロジーの整備を行う。

H24年度成果発表

口頭発表:

- K. Fukui, T. Tashiro, Y. Yabuki, K. Asai; "Development of Bioinformatics Workflows using Platform and Semantic Technologies", ISCB-Asia/SCCG 2012, Shenzhen, China, Dec. 17-19 2012.
- 福井 一彦、浅井 潔; “解析ツール・データベースの統合化と秘密検索”、日本製薬工業協会、トリイ日本橋ビル、11月21日 2012.
- 福井 一彦、田代 俊行、矢葺 幸光、浅井 潔; “プラットフォームを用いた解析ツールの統合・連携”、生命医薬情報学連合大会、JSBi/CBI、10月15日 2012.

ポスター発表:

- 福井 一彦、田代 俊行、矢葺 幸光、浅井 潔; “解析プラットフォームによる統合利用環境の構築”、トーゴの日シンポジウム2012 ～ライフサイエンスデータベース統合の医学への応用を探る～ 10月5日 2012.
- K. Fukui, T. Tashiro, Y. Yabuki, K. Asai; "A Composition of Bioinformatics Workflows using Platform and Semantic Technologies", Bio-IT World Conference & Expo 2012, Boston, April 24-26, 2012.

デモ:

- K. Fukui, T. Tashiro, Y. Yabuki, K. Asai; "Development of a workflow system using integrated information platform technology", BiWo2012, CBRC, Oct.31-Nov.2, 2012.

データベース統合に関わる基盤技術開発

データ統合と新規分野データ活用のための基盤技術開発 (京大グループ)

- (1) データベースの RDF による統合化
 - DBGET/LinkDBシステムの統合利用環境への応用
- (2) 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
 - 反応オントロジーの整備
- (3) 大規模データの利用技術開発
 - 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



(1) DBGET/LinkDB システムの統合利用環境への応用 (データベースの RDF による統合化)

本プロジェクトで予定している内容

- 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

赤字: 実施または着手済み



(1) DBGET/LinkDB システムの統合利用環境への応用 (データベースの RDF による統合化)

今年度を実施した内容

等価リンクの更新

- 化合物データベースの InChI 構造(と CAS)を用いた等価リンク作成
 - KEGG COMPOUND/DRUG と PDB-CCD, KNApSAcK, ChEMBL
- 反応データベースのリンク
 - KEGG REACTION と SABIO-RK
 - KEGG REACTION と UM-BBD reaction(手動)
 - KEGG RCLASS と UM-BBD btrule(手動)

間接的なリンクの定義と更新

- 遺伝子データから反応や化合物へ

データベースのメタ情報の整備

- DB の分類と URI のリスト作成、リンクされている DB 間の関係



URI prefixes of the DBGET and link-only databases

DBGET

LinkDB

KEGG2

1. KEGG Databases in DBGET

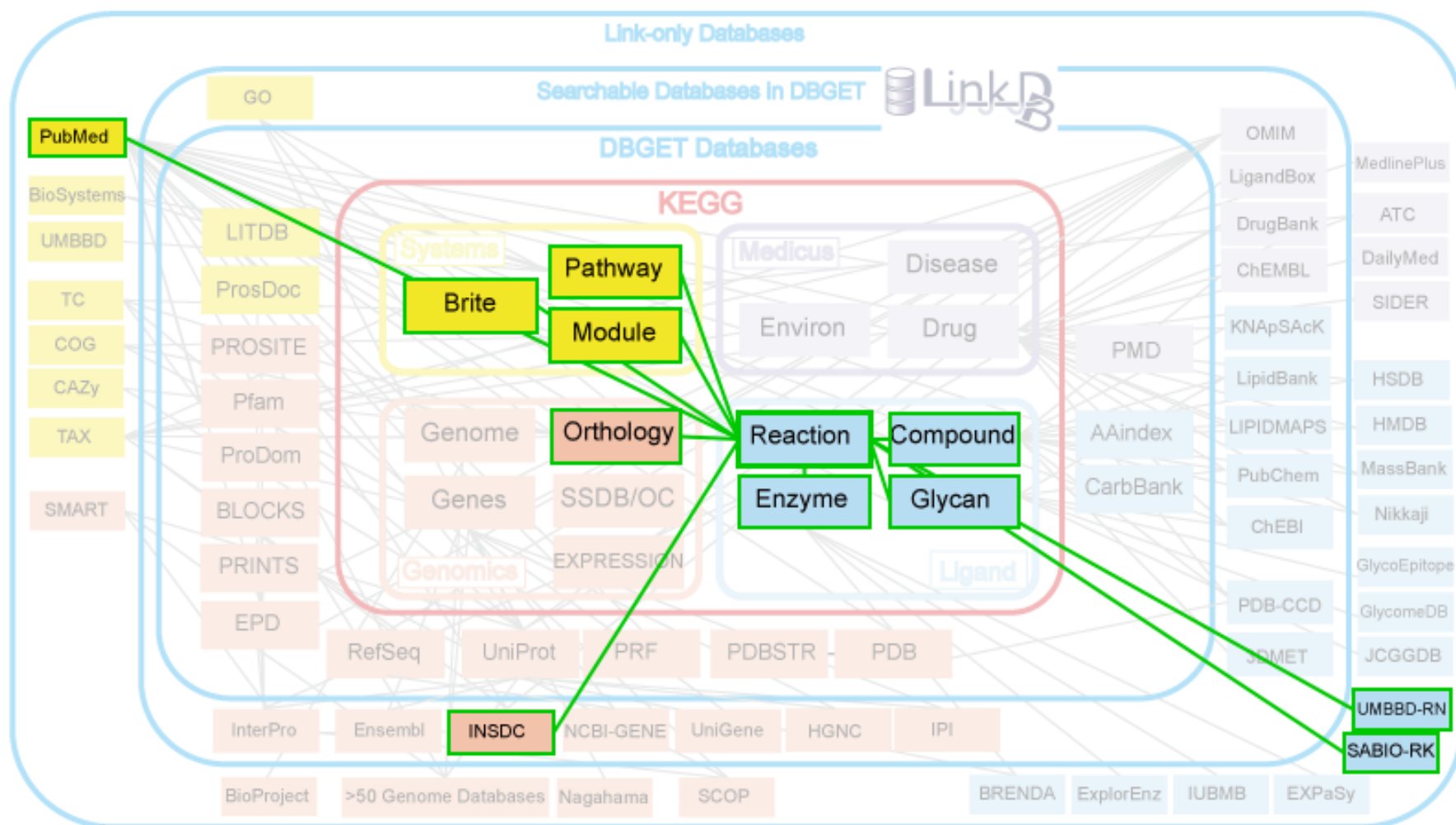
All KEGG databases: <http://www.kegg.jp/entry/>

2. Other DBGET Databases

Database name	Prefix
RefSeq	http://www.ncbi.nlm.nih.gov/nuccore/
UniProt	http://www.uniprot.org/uniprot/
PRF	http://www.example.org/
PDB	http://pdj.org/rdf/
EPD	http://epd.vital-it.ch/cgi-bin/query_result.pl?out_format=NICE&from=-499&to=100&Entry_0=
PROSITE	http://prosite.expasy.org/
Blocks	http://blocks.fhcrc.org/blocks-bin/getblock.pl?
ProDom	http://prodrom.prabi.fr/prodrom/current/cgi-bin/request.pl?question=DBEN&query=
PRINTS	http://www.example.org/
Pfam	http://pfam.sanger.ac.uk/family/
PMD	http://spock.genes.nig.ac.jp/~pmd/cgi-bin/PMD/PMDentry.pl.cgi?
AAindex	http://www.genome.jp/dbget-bin/www_bget?aaindex:
PDBSTR	http://www.genome.jp/dbget-bin/www_bget?pdbstr:
CarbBank	http://www.glycome-db.org/database/showStructure.action?glycomeId=
LITDB	http://www.genome.jp/dbget-bin/www_bget?litdb:
PROSITE (DOC)	http://prosite.expasy.org/

3. Searchable Databases on the Web

Database name	Prefix
GenBank	http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=
EMBL	http://www.ebi.ac.uk/cgi-bin/emblfetch?style=html&id=
NCBI-Gene	http://www.ncbi.nlm.nih.gov/gene/
UniGene	http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=
HGNC	http://www.genenames.org/data/hgnc_data.php?hgnc_id=
GO	http://amigo.geneontology.org/cgi-bin/amigo/go.cgi?view=details&search_constraint=terms&depth=0&query=



Download link information from to in format



(1) DBGET/LinkDB システムの統合利用環境への応用 (データベースの RDF による統合化)

今後実施予定の内容

ゲノムネット計算サービスのAPI化

- LinkDB の REST 検索
- 化合物構造検索の API: curl などで利用できるようにする
- ホモロジー検索、モチーフ検索との統合

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

継続的なデータの更新



(2) 反応オントロジーの整備 (RDF 化に資するオントロジー整備、標準化技術開発)

今年度を実施した内容

RCLASS: Reaction Classification and Ontology

PathSearch PathPred E-zyme KEGG REACTION

Search Chemical reaction for

RCLASS database

RCLASS is a database of reaction classification based on the chemical structure transformation patterns of main reactant pairs. The transformation pattern is described in RDM format defined in the RPAIR database. RCLASS entries are classified in hierarchical ways and used in the pathway analysis tools such as [PathSearch](#).

Reaction Classifications

- [Functional group conversion hierarchy](#)
- [Classification based on atom classes](#)
- [Classification based on EC numbers](#)

Tools using RCLASS database

- **PathSearch:** for searching similar reactions
- **PathPred:** for predicting biosynthesis and
- **E-zyme:** for assigning EC numbers

Feedback KEGG Genome

反応クラスの分類

1. 官能基の変化に基づいた手動分類
2. KEGG Atom Type の分類によるグループ化

RCLASS, RPAIR の RDF 化

FTP でのダウンロード



(2) 反応オントロジーの整備 (RDF 化に資するオントロジー整備、標準化技術開発)

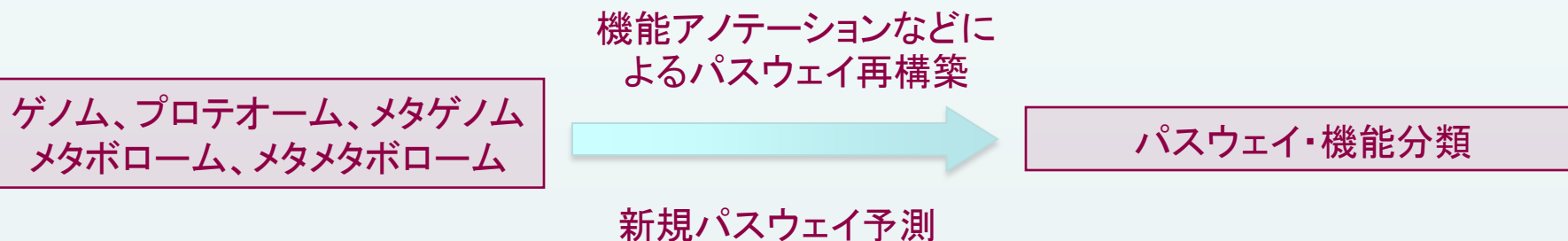
今後実施予定の内容

反応クラス分類

- 官能基などの用語の整備
- 分類間の比較
- 遺伝子 (KEGG Orthology) との対応付け
- 反応経路計算ツールへの応用

RCLASS, RPAIR の継続的なデータ更新

(3) 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（大規模データの利用技術開発）



- 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
- 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

(3) 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（大規模データの利用技術開発）



メタゲノムデータ利用技術開発

 KEGG Metagenomes		
[Genomes Draft 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
Microbial mat	T30140 Metagenome isolated from subsurface mine microbial mat	NCBI
	T30141 Anaerobic oxidation of methane in the Black Sea	NCBI
Organismal metagenomes		
Category	Project	Source
	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 (female infant)	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 (female infant)	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

- データベース化支援技術：
 - KAAS によるアミノ酸配列レベルでの機能アノテーションとパスウェイ再構築
 - Short read 入力時の KAAS アノテーションの評価
 - GenBank env division
- 検討中 & 今年度
 - HMP データの追加
 - MODULE を用いたKAASアノテーションの拡張

(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



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

PathSearch: Search Similar Reaction Paths

About PathSearch

PathSearch computes similar paths for a query reaction path using the RCLASS database for reaction patterns matching to the query reactions. The results can be mapped to KEGG pathway diagrams using KEGG Mapper. The colors of the objects can be also specified.

Enter reaction path:

RC00098 RC00534 RC01140

Ex.1
R00089
RP0343
RC0005

Ex.2
R02740
R04779
R01070

Ex.3
C00345

☐ Use auto bgcolor.
☒ Skip repeated reaction.

Alternatively, use the pathway map into

Select map

Map number:

[Feedback](#) [KEGG](#) [Geno](#)

類似反応検索ツールの開発

<http://www.genome.jp/tools/pathsearch/>

Result of Path Search

Target : Reference pathway (RPAIR)

Query : RC00098 RC00534 RC01140

Number of results : 9

Top 10 ☐ containing all listed objects

☒	C06588	<RP00265>	C06589	<RP04841>	C02526	<RP11289>	C01273	[Known pathways]	[Show compound structures]
☒	C07111	<RP05048>	C06727	<RP04915>	C06728	<RP05022>	C07123	[Known pathways]	[Show compound structures]
☒	C07083	<RP05023>	C07084	<RP04951>	C07085	<RP05010>	C07087	[Known pathways]	[Show compound structures]
☒	C06584	<RP11674>	C06585	<RP04843>	C06586	<RP04845>	C06587	[Known pathways]	[Show compound structures]
☒	C04623	<RP00822>	C06649	<RP00155>	C06650	<RP04856>	C06651	[Known pathways]	[Show compound structures]
☒	C01455	<RP03174>	C04592	<RP03714>	C02923	<RP03715>	C06210	[Known pathways]	[Show compound structures]
☒	C06578	<RP04847>	C06579	<RP04842>	C06580	<RP04848>	C06581	[Known pathways]	[Show compound structures]
☒	C05629	<RP09352>	C11588	<RP09313>	C04044	<RP04018>	C04479	[Known pathways]	[Show compound structures]
☒	C00423	<RP09186>	C12622	<RP09576>	C12623	<RP09577>	C12624	[Known pathways]	[Show compound structures]



(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

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

類似反応検索ツールの開発

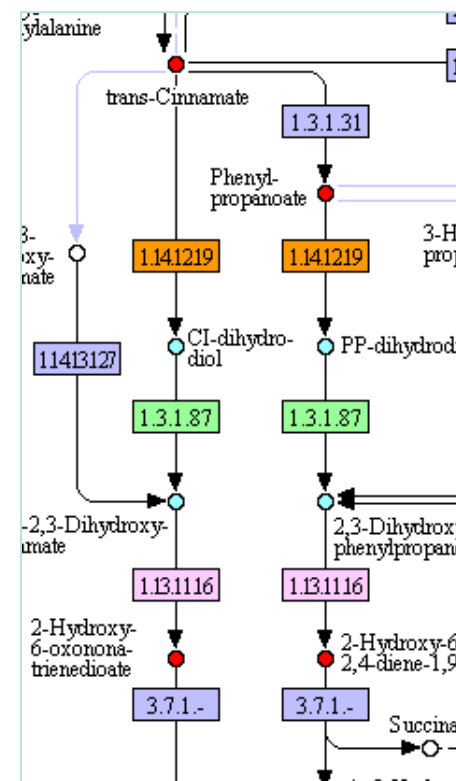
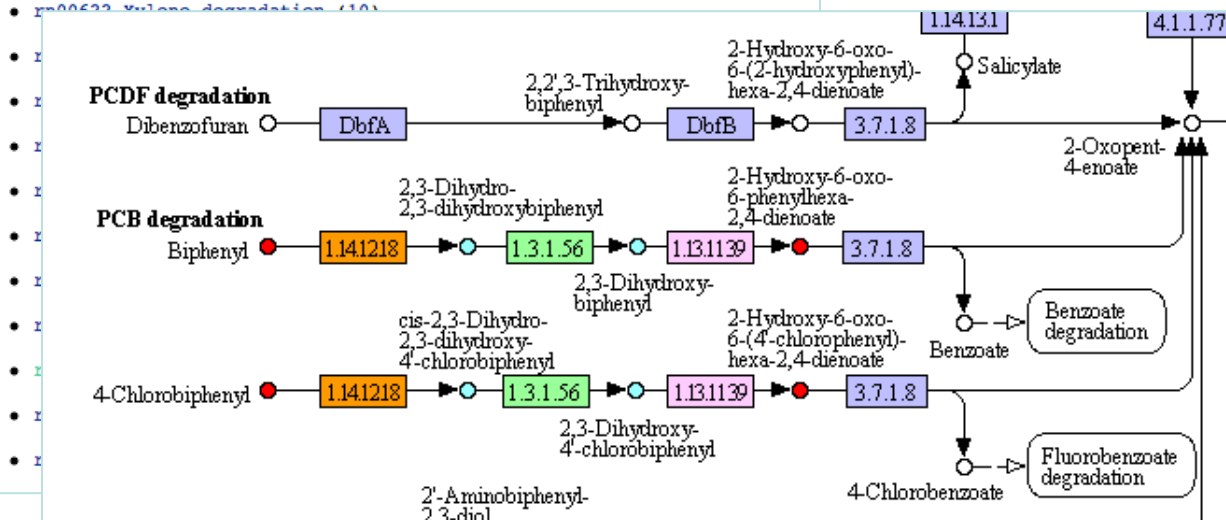
<http://www.genome.jp/tools/pathsearch/>

Pathway Search Result

Sort by the pathway list

Show all objects

- rn01120 Microbial metabolism in diverse environments (74)
- rn01100 Metabolic pathways (28)
- rn00621 Dioxin degradation (16)
- rn00360 Phenylalanine metabolism (14)
- rn00622 Xylene degradation (10)



(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



今年度に実施したその他の内容

遺伝子機能予測システム GENIES (GEne Network Inference Engine based on Supervised analysis) 整備: Kotera, M., *et al.* NAR Web issue 2012

オーソログクラスタ KEGG OC 整備: Nakaya, A., *et al.* NAR DB issue 2013

今後実施予定の内容

新規メタゲノム、メタメタボロームデータへの対応:
得られた化合物間の関連を解析するツール

ゲノム、プロテオーム、メタゲノム
メタボローム、メタメタボローム

機能アノテーションなどによる
パスウェイ再構築

KAAS

パスウェイ・機能分類

新規パスウェイ予測

PathPred, PathSearch, GENIES

データ統合と新規分野データ活用のための基盤技術開発 (京大グループ)

- (1) データベースの RDF による統合化
→ DBGET/LinkDBシステムの統合利用環境への応用
- (2) 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
→ 反応オントロジーの整備
- (3) 大規模データの利用技術開発
→ 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

メンバー

代表: 五斗進
特定助教: 時松敏明、小寺正明
特定研究員: 守屋勇樹、中川善一

協力メンバー

川島秀一、山口敦子(DBCLS)
大久保宏一、上原英也、小澤健太郎(SGI)
山本留美子(KEGG)
高見英人(JAMSTEC)