

理研：植物オミックスとタンパク構造



『植物を活かしたグリーンイノベーション
のための統合データベースづくりへ』

(独)理化学研究所
生命情報基盤研究部門
豊田哲郎

①シロイヌナズナオミックス情報注釈 (豊田)

– トランスクリプトームデータ

- リポジトリを9月までに公開し、それらの詳細な解析結果に基づいたアノテーションデータを年度末までに公開する。

– フェノタイプデータ

- 理研で観測されたフェノタイプデータを、理研以外で観測されたフェノタイプデータと統合化し、共通のオントロジーで整理し9月までに公開する。

– データベース統合化の範囲を拡大

- シロイヌナズナで実施してきた統合化をモデルケースとして、シロイヌナズナ以外の生物種(マウス等)にも適用することで、データベース統合化の範囲を拡大する。

平成21年度実績(豊田)

- シロイヌナズナオミックス情報の注釈付けと公開
 - 理研PSCで作成されたFOX-Hunting法によるシロイヌナズナ変異体のフェノーム情報をデータベース化し、理研サイネス上で公開した(系統数14125件、表現型情報13292件)。
 - 理研サイネス上で公開されている3件のシロイヌナズナ変異体データベース中の表現型情報に関して、Plant Ontology (PO), PATOといった公開オントロジーを用いた新しい統一的なアノテーションを考えた。現時点で115種類の表現型を定義し、既存3データベースの該当データに対してリンク付け(3DBあわせて14631件)を行った。これにより、公開済みの表現型情報については、関連する変異体の情報が実験手法等の違いに関わらず検索可能となった。

トランスクリプトームデータの統合公開

Semantic Web

SAT gene

SAT ヘア

SAT-PS5-02240

SAT gene

ARTADE2 遺伝子領域

OMAT5P020730

SAT gene

ARTADE2 遺伝子領域

OMAT5S101280

ARTADE2 遺伝子領域

OMAT5S101280

Estimate principle

予測法則

Secondary gene

ARTADE2_exon

ARTADE2 エクソン

OMAT5S101280_exon_2

ARTADE2_exon

ARTADE2 遺伝子領域

OMAT5S101280

ma, region: 174-585, strand: +

位置 (Tair8)

dna, region: 26680830-26681241, strand: -

tair8

Chr5

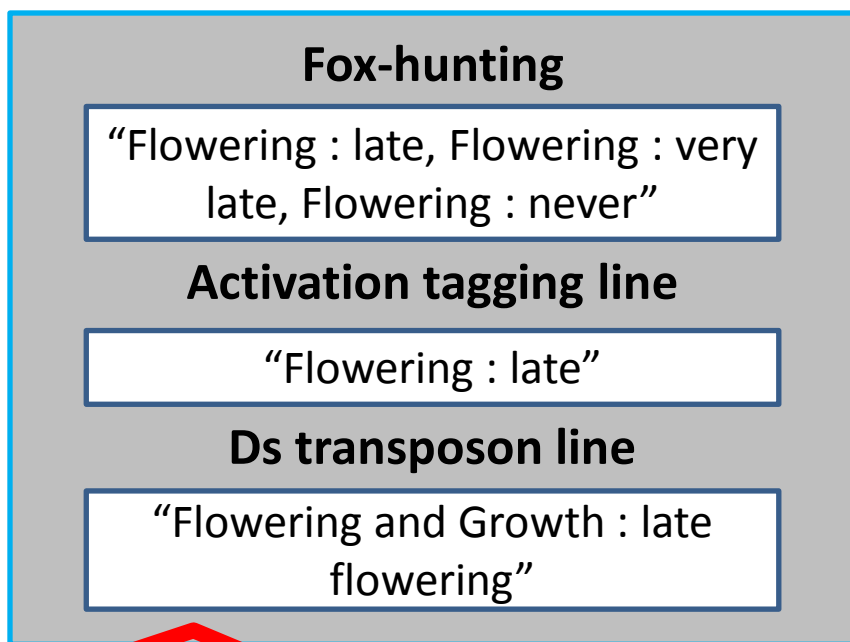
ARTADE2 遺伝子領域	
OMAT5S101280	
外部ID	OMAT5S101280
最終更新日	2009-06-23 20:49:25.0
Estimate principle (予測法則)	Secondary gene
OMAT5S101280 関連事項	
SAT ヘア	SAT-PS5-02240
SAT gene (ARTADE2 遺伝子領域)	OMAT5P020730
SAT gene (ARTADE2 遺伝子領域)	OMAT5S101280
ARTADE2 エクソン	OMAT5S101280_exon_2
ARTADE2_exon (ARTADE2 遺伝子領域)	OMAT5S101280
位置 (Tair8) (tair8)	Chr5 dna, region: 26680830-26681241, strand: -
ARTADE2 エクソン	OMAT5S101280_exon_1
ARTADE2_exon (ARTADE2 遺伝子領域)	OMAT5S101280
位置 (Tair8) (tair8)	Chr5 dna, region: 26681349-26681521, strand: -
発現情報	EXPRESSION-Information-STEM-OMAT5S101280
Expression (ARTADE2 遺伝子領域)	OMAT5S101280
中央値	20.5833
HLE	31.7333
Expression P	0.394831
発現情報	EXPRESSION-Information-SILIQUELATE-OMAT5S101280
Expression (ARTADE2 遺伝子領域)	OMAT5S101280
中央値	326.917
HLE	546.565
Expression P	4.37363e-05
発現情報	EXPRESSION-Information-SILIQUEMIDDLE-OMAT5S101280
Expression (ARTADE2 遺伝子領域)	OMAT5S101280
中央値	82.8967
HLE	122.99
Expression P	0.0328992

- 理研のシロイヌナズナデータと世界のデータを統合化して公開
 - tiling array発現解析
 - シーケンスデータ
 - 遺伝子構造推定プログラムARTADE2で発見されたシロイヌナズナ新規遺伝子(主にアンチセンスRNA遺伝子)

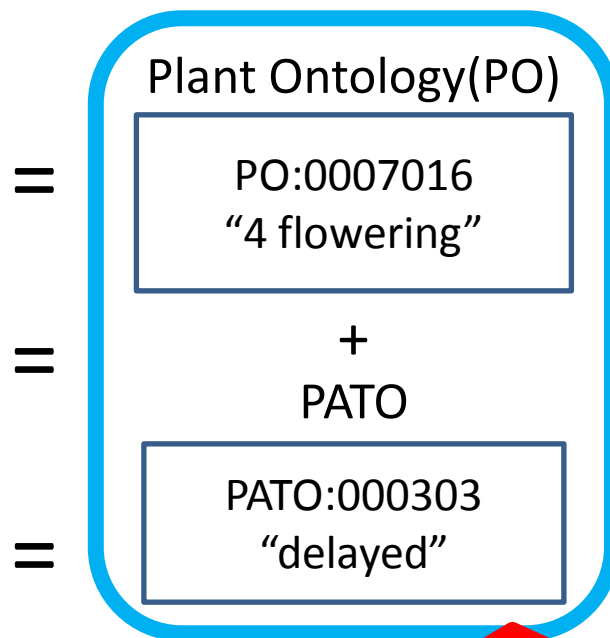
PCPにデータベースの特集号を予定(来年1月号)

フェノタイプ（表現型）情報の統合

理研に複数ある変異体データベース群を統合化し、オントロジーによる表現型情報の統一を図る地道なアノテーション作業。



従来のDB＝同じ表現型が、ばらばらな
様式で登録されていた



PO、PATOなどのオントロジー
を用いて統一

外部フェノタイプ（表現型）情報の統合

- 理研以外で観測されたフェノタイプデータを既存の理研内のデータと統合し、9月までに公開

世界中の研究者が報告しているシロイヌナズナのフェノタイプ情報の文献を収集する。



文献内の記述を注意深く解釈し、オントロジーを利用して表現型情報を標準化し、データベース化する。



理研サイネス上で、各種オミックス情報と関連付ける。

PCPにデータベースの特集号を予定（来年1月号）

理研以外(慶応大)のDBもサイネスから公開

植物リン酸化プロテオームデータベ...

植物リン酸化プロテオームデータベース

ENGLISH | ログイン | 利用規約

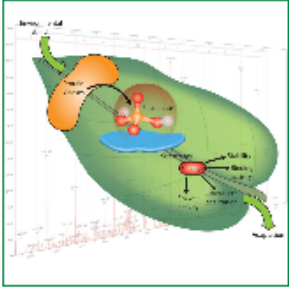
リンク | Wiki | ダウンロード | 情報

検索

ディスクを検索 ☒ サイネス全体から検索

Hash値に変換

Plant Phosphoproteome Database



The Plant Phosphoproteome Database provides information on phosphopeptides which were identified from Arabidopsis and rice cells by LC-MS/MS-based shotgun phosphoproteomics approach. MS/MS spectral information for the identified phosphopeptides at [PepBase \(Keio University\)](#) are available. Alignment views of orthologous proteins in Arabidopsis and rice mapped with the identified phosphopeptides allow users to predict conserved phosphorylation sites in plants. Current database houses information for 5,143 and 6,919 phosphopeptides from Arabidopsis and rice, respectively. Detail information is available from following articles.

Reference to the Plant Phosphoproteome Database:
We kindly ask users to cite the following papers when publishing results derived from the use of the Plant Phosphoproteome Database:
Nakagami, H., Sugiyama, N., Mochida, K., Daudi, A., Yoshida, Y., Toyoda, T., Tomita, M., Ishihama, Y. and Shirasu, K. Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants. (2010) Plant Physiology ([Published on May 13, 2010; 10.1104/pp.110.157347](#))
Sugiyama, N., Nakagami, H., Mochida, K., Daudi, A., Tomita, M., Shirasu, K. and Ishihama, Y. Large-scale phosphorylation mapping reveals the extent of tyrosine phosphorylation in Arabidopsis. (2008) Molecular Systems Biology 4:193 ([MSB](#) | [PubMed](#))

植物リン酸化プロテオームデータベースの関連ディスク

[TIGR Rice Gene Models](#) [植物リン酸化プロテオームデータベース](#) [A.thaliana gene](#)

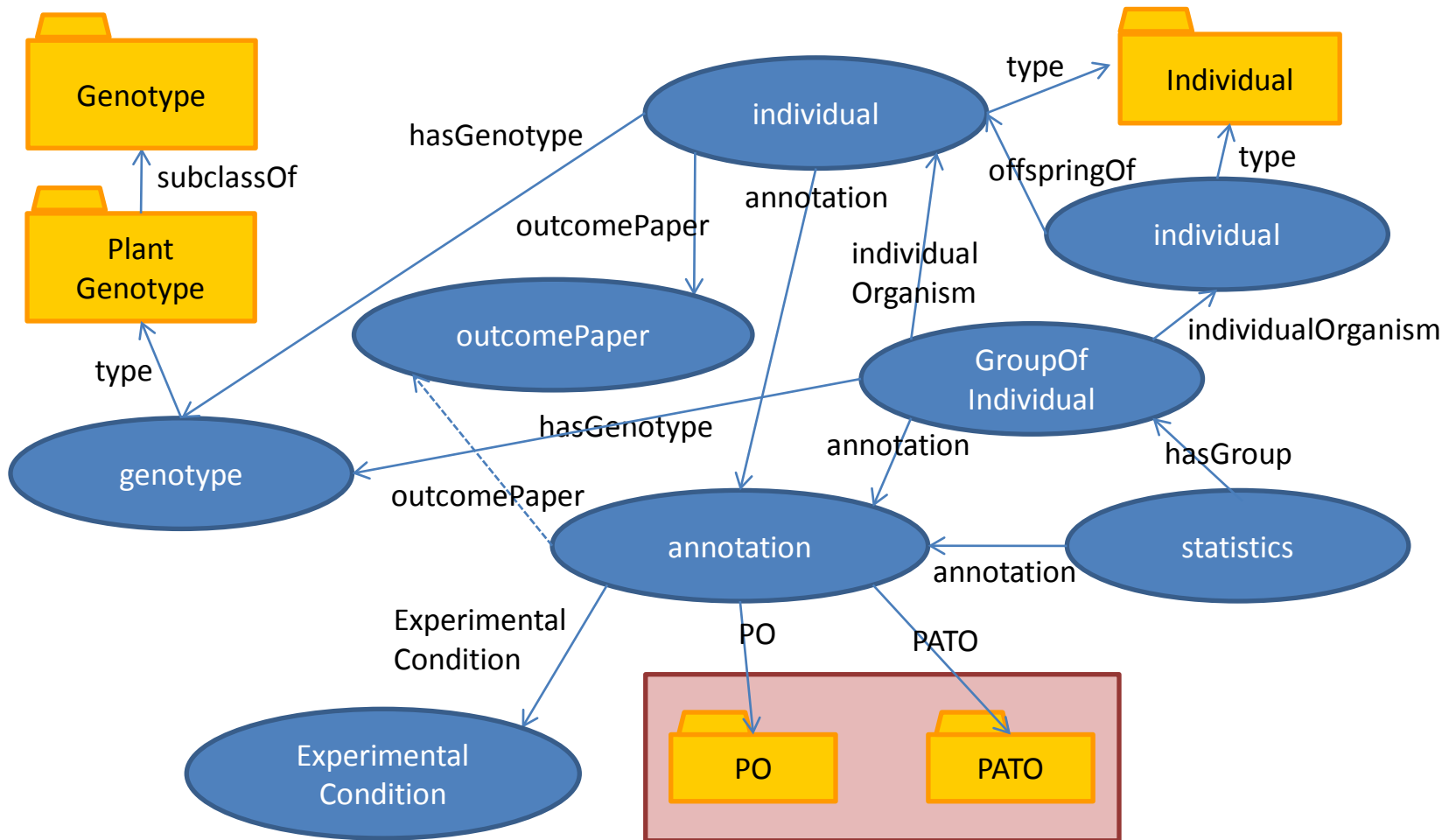
植物リン酸化プロテオームデータベースのコンテンツ

フォルダ	サブフォルダ
リン酸化ペプチド	Atリン酸化ペプチド Osリン酸化ペプチド
ペプチド	Atペプチド Osペプチド
At-Osオルソログクラスター	
ペプチドローカス	Atペプチドローカス Osペプチドローカス

理研外からも植物系DBが集まり統合化される体制を確立

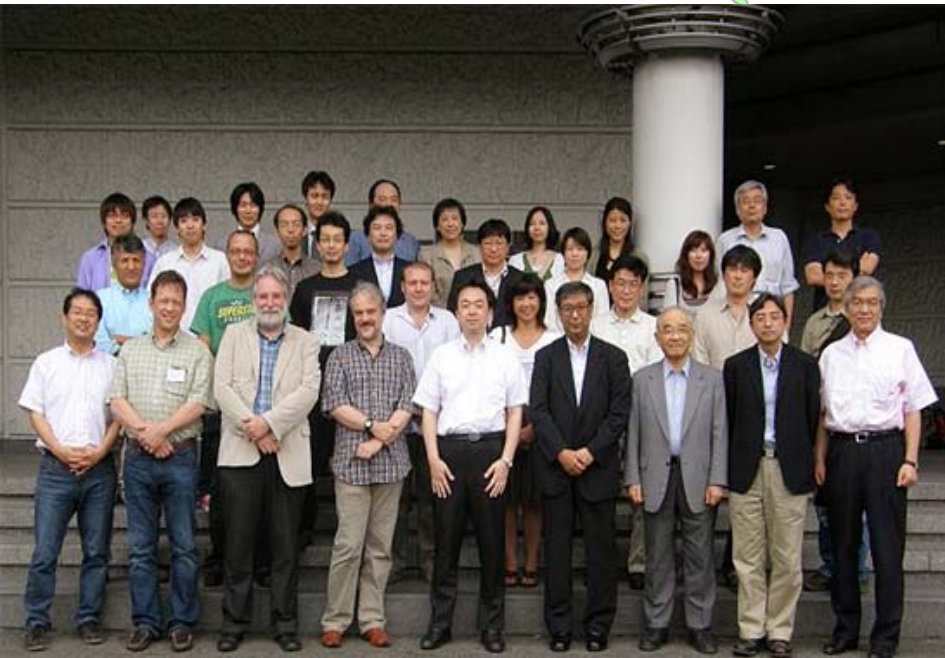
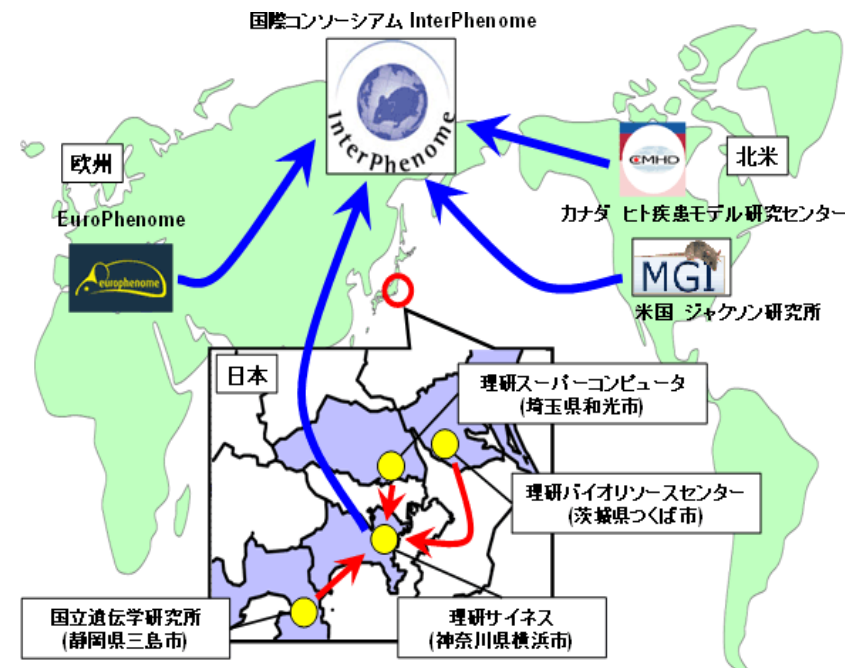
データベース統合化の範囲を拡大

シロイヌナズナで実施してきた統合化をモデルケースとして、
シロイヌナズナ以外の生物種にも適用→マウスフェノタイプへ



生物種を問わず適用できるフェノーム情報表現のための枠組みを設計

マウス表現型データの国際共有化の合意 (InterPhenome)



理研サイネス - Ano_SHQ36_hom_M - Mozilla Firefox

ファイル(E) 編集(E) 表示(V) 書置(S) ブックマーク(B) LSP(L) ツール(T) ヘルプ(H)

https://database.riken.jp/sw/panel#cria110:ria110:222

よく見るページ Firefox を使ってみよう 最新ニュース

理研サイネス - Ano_SHQ36_hom_M

Annotation of Phenotype Data

独立行政法人 理化学研究所 バイオリソースセンター
日本マウスクリニック
Japan Mouse Clinic

メニュー 「Ano_SHQ36_hom_M」のページ

Ano_SHQ36_hom_M Image Information

作成者: 吉田 孝平
Ano_SHQ36_hom_M < Annotation of Phenotype Data > 日本マウス表現型データベース

Phenotype annotation for Whisker Morphology of B6;D2-Polg male homozygotes in Japan Mouse Clinic.

リンク has_quality value has_quality type Sub Data by Whom 測定対象物 測定対象属性 成果文庫

quality as data content その他

semantic web graph

Download

Annotation of Phenotype Data

Ano_SHQ36_hom_M

最終更新日	2009-11-24 15:45:28.0
外部ID	--
ファイル	--
by Whom (編集)	マウス表現型データベースチーム
上位(母語)	バイオリソースセンター
測定対象属性 (morphology)	whisker morphology of B6;D2-Polg < Rgs625/Rgs625 > male
inheres in (whisker)	whisker of B6;D2-Polg < Rgs625/Rgs625 > male
quality as data content (deviation value)	normal
value of	deviation(from_normal)

Ano_SHQ36_hom_M 関連事項

Statistics

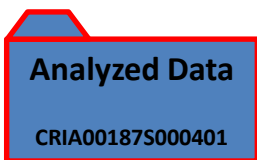
Polg_SHQ36	
by Whom (編集)	マウス表現型データベースチーム
測定対象属性 (morphology)	whisker morphology of B6;D2-Polg < Rgs625/Rgs625 > male
測定対象属性 (morphology)	whisker morphology of B6;D2-Polg < Rgs625/Rgs625 > female
測定対象属性 (morphology)	whisker morphology of B6;D2-Polg < Rgs625/+ > male
測定対象属性 (morphology)	whisker morphology of B6;D2-Polg < Rgs625/+ > female
測定対象属性 (morphology)	whisker morphology of B6;D2-Polg < +/- > male
測定対象属性 (morphology)	whisker morphology of B6;D2-Polg < +/- > female
experimental parameter (experimental Parameter)	Q36: Whisker Morphology
annotation (Annotation of Phenotype Data)	Ano_SHQ36_hom_F
annotation (Annotation of Phenotype Data)	Ano_SHQ36_hom_M
annotation (Annotation of Phenotype Data)	Ano_SHQ36_hom_F
annotation (Annotation of Phenotype Data)	Ano_SHQ36_hom_M
group (Cohort)	B6;D2-Polg_wild_female
group (Cohort)	B6;D2-Polg_wild_male
group (Cohort)	B6;D2-Polg_hetero_female
group (Cohort)	B6;D2-Polg_hetero_male
group (Cohort)	B6;D2-Polg_homo_female
group (Cohort)	B6;D2-Polg_homo_male

Semantic Web

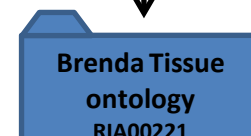
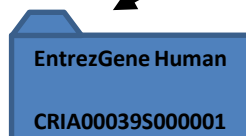
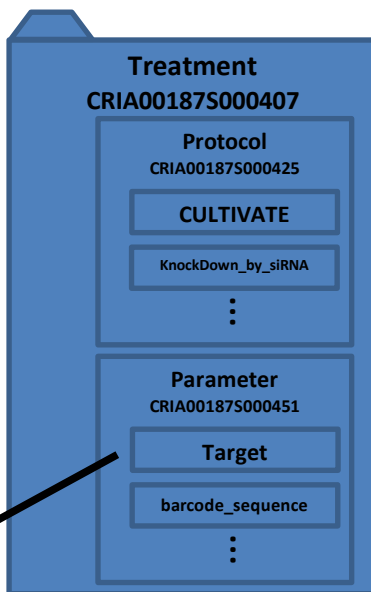
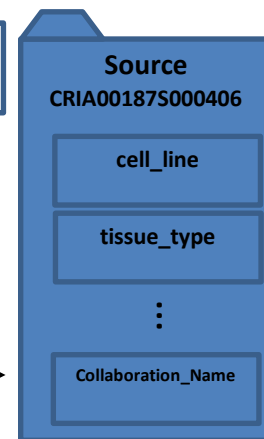
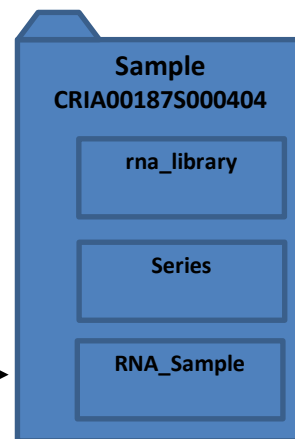
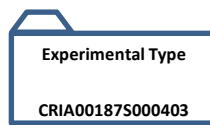
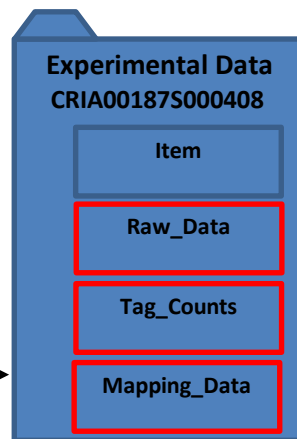
完了

FANTOM4

データの統合作業



level123_clusters
qrtpcr_pma_timecourse
illumina_hg6v2_pma_timecourse
Pol_II/PMA_96h
illumina_hg6v2_siRNAseries



respective data

sample

source

experimental type

sample type

GEO
array design
experimental design

chip
⋮
tf-qrt-pcr

cultivate
knockdown
siRNA
⋮

関連付け

sameAs

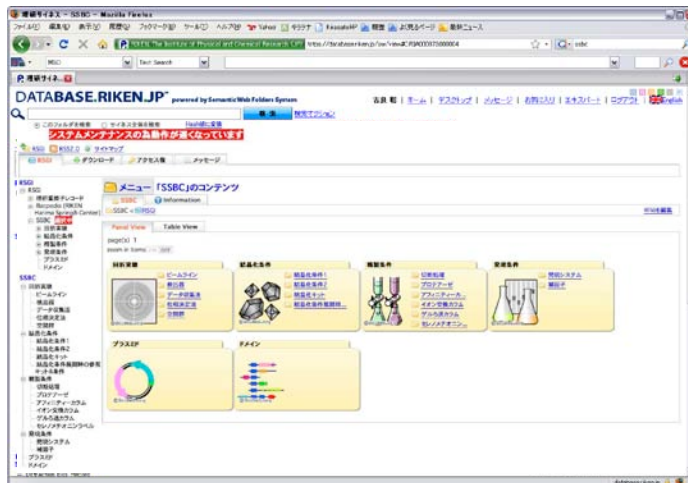
関連付け

②高等動植物等由来蛋白質構造の アノテーション(横山)

- タンパク3000プロジェクト(RSGI)で解明された高等動植物等由来の蛋白質構造データに付随する実験データに標準的なオントロジーやIDに基づくアノテーションをつけてXMLやテーブル形式でのダウンロードを可能にする.
- 9月末までにDBCLSに成果データを直接渡す→畠中さんへ

平成21年度実績（横山）

- 高等動植物等タンパク質の試料調製に関わる発現検討実験 (無細胞タンパク質発現用発現系構築実験4万8千件, スモールスケール発現実験8万件) に基づいてX線結晶構造解析のための大量調製実験 (大量合成および大量発現実験4千件、精製実験5千件) を行い, 結晶化実験 (結晶化条件 138 万件、結晶観察 1517 万件) へと進め, タンパク3000プロジェクトで解明されたタンパク質構造データのうち平成21年度としては1万1千件の回折実験データ (画像データ枚数) を公開可能とした (回折実験画像データの累計枚数は3万1千件となった)。



■試料調製, 結晶化及び構造解析(高等動植物等)実験データ公開用 Web site

- 高等動植物等由来タンパク質の試料調製, 結晶化及びX線結晶構造解析実験情報の閲覧が可能. 回折画像データを含めダウンロードサービスが可能となった.

■URL:

<https://database.riken.jp/item/cria87s4i>

③微生物由来蛋白質構造のアノテーション(国島)

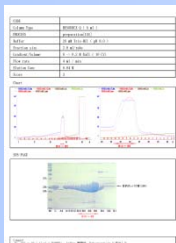
- 平成21年度に引き続き、これまでに公開したデータについてアノテーション作業を行い、アノテーションデータを年度末までに公開する。またそれ以外のデータについてもデータ公開を進める。
- 平成22年度末までに、理研播磨研究所における微生物由来の蛋白質構造データに付随する実験データに関し、統合データベースの一部として生命科学分野の幅広い研究者に役立つ体制の確立を完了する。

MEXT integrated database project - Protein 3000 data in RSC -

Proteins from microorganisms

Bacpedia in SPring8-Center

Purification



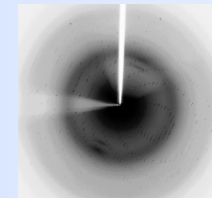
3,000 Purification reports

Crystallization



900,000 Crystallization condition
10,000,000 Crystal observations

Diffraction data

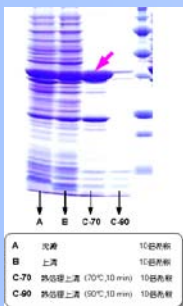


200 Data sets

5,000 Expression reports



Expression



10,000 Plasmid construction reports

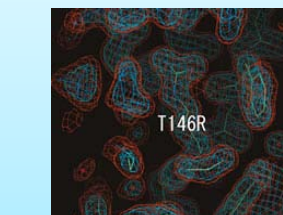
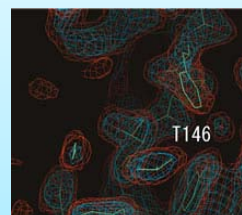
Mutant proteins

Bacpedia in SPring8-Center



Wild type structure at 2.1 Å resolution

150 Systematic mutants



Mutant structure at 1.6 Å resolution

Heavy-atom labeled proteins

Protein	Heavy-atom	Label	Concentration	Volume	Time	Yield	Notes
Protein A	150	100	100	100	100	100	
Protein B	150	100	100	100	100	100	
Protein C	150	100	100	100	100	100	
Protein D	150	100	100	100	100	100	
Protein E	150	100	100	100	100	100	

Heavy-atom database

HATODAS in SPring8-Center

500 Heavy-atom sites



Contents

Data size: 5.0 TB

Publish platform



Bacpedia: <http://scines.org/item/rib220i>
HATODAS: <http://scines.org/item/rib108i>

Data downloading was enabled on 23 July 2009.

④アノテーションシステムの開発運用 とデータ変換（豊田）

- 9月までに本事業において公開したデータをクライアントサイドからウェブ経由でプログラマ的に利用できるAPIとツールを構築して公開する。
- 9月までに、これまでに構築したアノテーション作業を行うためのシステムを運用・改善し、アノテーション等のデータ投稿やプログラム投稿を理研内外から実施できるオープンな参加型のシステムとして公開する。
- 9月までに、理研外の研究者から見て、理研のDBが統合されたミニマムモデルの成功事例として、研究者コミュニティから認知されるための統合検索サイトを公開する。

データベース相互連携用インタフェースの公開

Semantic-JSON Specification and Tutorial

Bioinformatics And Systems Engineering division (BASE), RIKEN.

1. [Introduction](#)
2. [Architecture of JSON@SciNeS](#)
3. [SciNeS data structure](#)
4. [Semantic-JSON URI](#)
5. [Sample programs](#)

様々なプログラミング言語からアクセス可能なインタフェース

<http://semantic-json.org>

1. Introduction

On the Semantic Web, each data item is identified by a globally unique identifier, i.e., a URI, and a relationship between two data items is described as a statement including subject, predicate and object. Those data items are written in a standardized format named RDF, and a standardized language named SPARQL is used for querying RDF datasets. Since both RDF and SPARQL require expensive computer resource, it is difficult to effectively retrieve those data items in various computer languages including scripting languages popularly used in bio-informatics.

Semantic-JSON is a light-weight application program interface that allows a user to write a program over Semantic-Web-based data in JSON format. This interface is defined as a collection of URIs and each URI is specified as a simple data access method to Semantic Web data and returns a response in JSON format. Since JSON is supported by most popular programming languages, the functions and semantics for each URI are defined independently on one single programming-language in Semantic-JSON.

As a RDF repository, we employ Scientists' Network System (SciNeS), that is a framework for developing and integrating databases by collaboration among scientists on the Web, and have implemented a Semantic-JSON server interface for SciNeS called JSON@SciNeS. In this document, we will discuss the specification of Semantic-JSON and will illustrate how a user handle Semantic Web data stored in SciNeS via Semantic-JSON with sample programs.

2. Architecture of JSON@SciNeS

JSON@SciNeS is a web-based application of server-client model as shown in Figure 1.

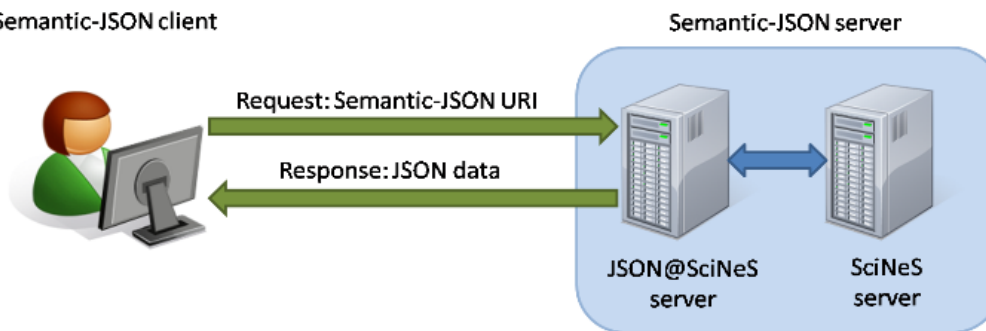


Figure 1. architecture of JSON@SciNeS.

A client is a HTTP or HTTPS client such as a user's program or a web browser.

The JSON@SciNeS server supports the Semantic-JSON interface and provides a service that performs clients' requests by accessing the SciNeS server providing a Semantic Web data repository service.

First, a client sends a URI as a data request to the JSON@SciNeS server. The URI denotes a function that processes Semantic Web data sets and is defined in the Semantic-JSON specification. The JSON@SciNeS server evaluates the client's URI, obtains Semantic Web data properly from the SciNeS server and converts these data into JSON data. Finally, the JSON data is sent back to the Client as the response.

完了

外部DBと連携するためのインタフェース(JSON, RDF他)

データベースを利用するためのプログラムと結果ファイルを保存できる機能

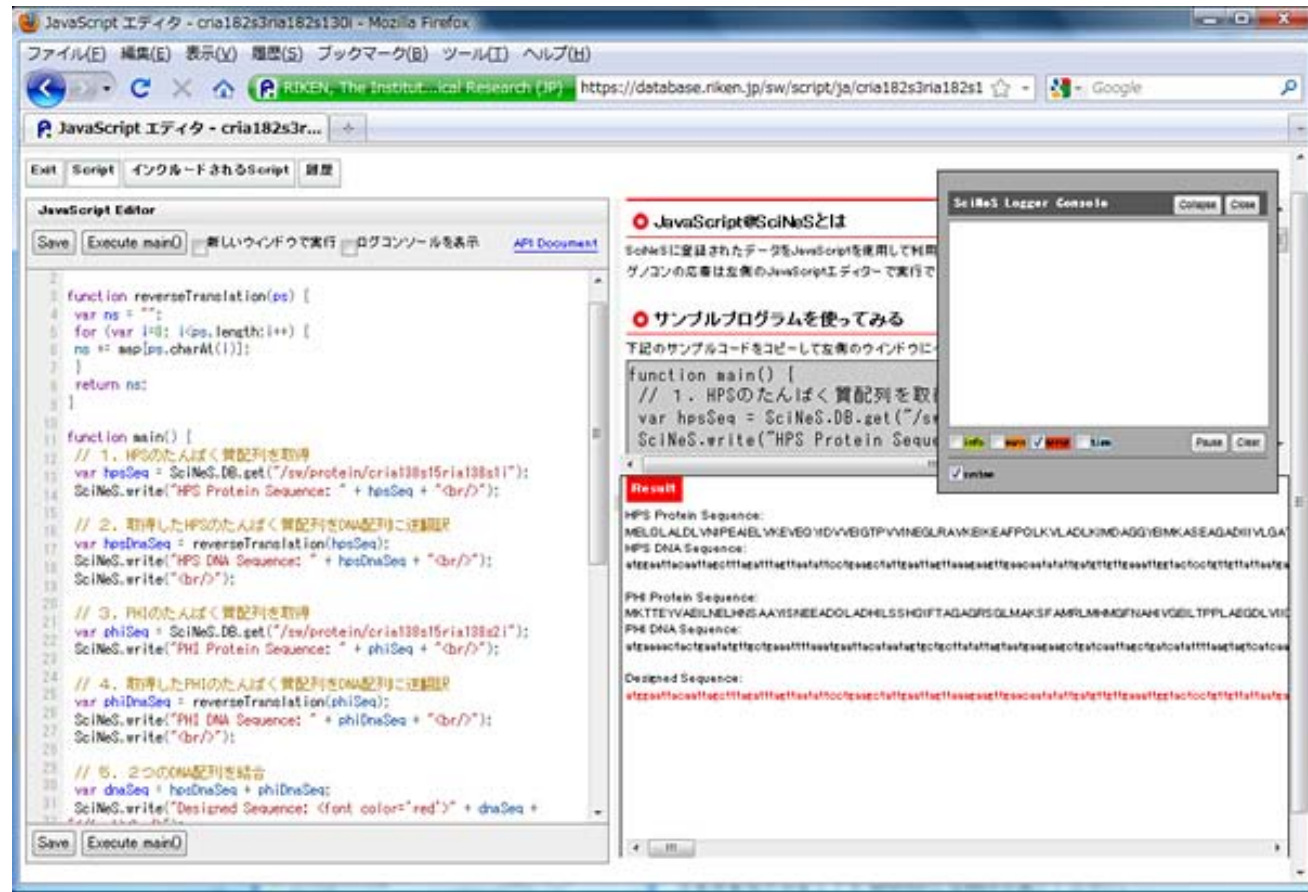
利用者が自分のワークスペースをサイネスにもてる



ウェブブラウザ上で、サイネス上の情報を利用したプログラムを作成



プログラムや処理結果をサイネスから公開



プログラムの蓄積と公開化と再利用が可能に(CC-BY-SA)

オープンな参加型のDBシステム

- ウェブブラウザさえあれば、誰でもプログラムを書いてデータベースを利用できる環境を提供
→ 理研外の人も参加可能
- 応用例の一つとして、合理的ゲノム設計コンテスト (GenoCon) を開催中 (開催費用は理研負担)
 - グリーンイノベーションの課題「環境浄化作用高める植物のゲノム設計」を募集
 - 世界中の研究者からゲノム設計理論をコンテスト形式で蓄積し集合知として共有化
 - 研究者に限らず高校生も参加できる、バーチャルラボでの人材育成型コンテスト



The screenshot shows the GenoCon website interface. At the top, there's a navigation bar with links like TOP, 開催概要, 応募要項, スケジュール, FAQ, and 主催. Below this is a large banner with the text 'GENOCON 合理的ゲノム設計コンテスト'. The main content area is divided into two columns. The left column contains a sidebar with links to 'サブメニュー' (Submenu) and 'リンク' (Links). The right column contains the main text, which includes a description of the contest and a section titled '課題' (Topic) with a sample DNA sequence: ATGGAGGATCAAGTTGGGT TTGGGTTCCGCGAAGCA CGAGGAGCTCGTTGGTAC TATCTCGTAAACAAATCG AAGGAACACTAGCCCGA CGTTGAAGTAGCCATCAGC GAGGTCAACATCTGTAGT ACATCTCTGGAACTTGGG CTCCAGTCAAGTACAAA TCGAGAGATGCTATGTGT ACTCTCTCTCGTAGAGAA. The bottom of the page features a green leaf icon and a disclaimer about safety and participation.

※別添のプレスリリースをご覧ください(参考資料1)

統合検索サイトの実現

統合されたデータベース群から、関連する情報を容易に引き出す。

右図は、一つのシロイヌナズナ遺伝子IDをキーとして関連するサイネス上の情報を検索し、列挙した例である。

本例では、理研の観察データ、オルソログデータなどの他、AGRIS等の公開DBの関連情報へのリンクも提示されている。

異種情報間の高精度検索（たとえばゲノム情報を出発点としたフェノーム情報検索）が容易化し、広く研究コミュニティに貢献できる。

SciNeS Scientists' Networking System

Koji Doi (<http://logon.jp/kdoi/>) | [Log off](#) | [Japanese](#)

At1g78600 [Hash this value](#) Search for all scines

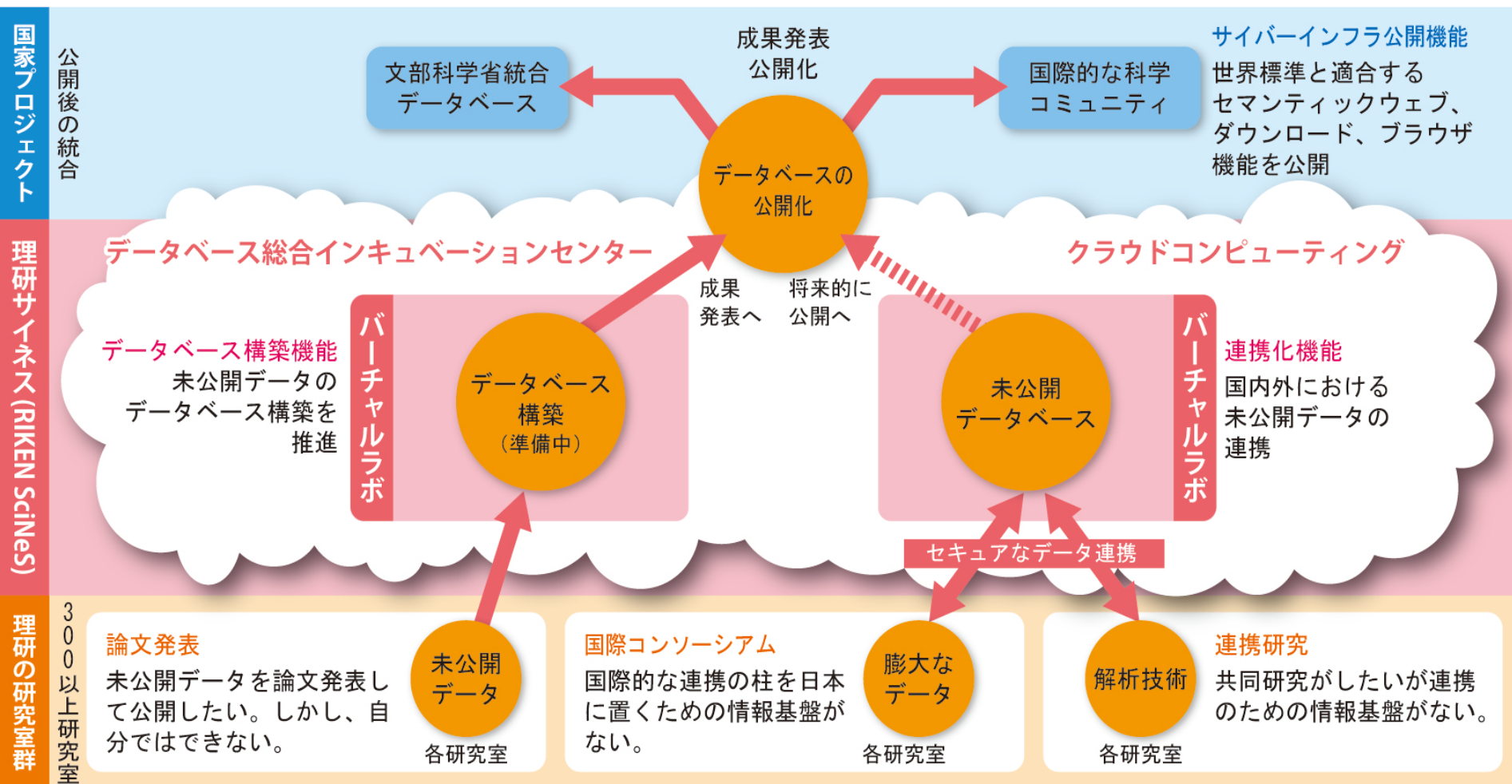
Search depth 2 Reverse search depth 1

Results 1 - 10 of 23 for find. (10.05 seconds)

- AT1G78600**
[AT_TAIR locus](#)
Last Update Date 2008-05-29 19:48:52
<https://blade2gw2.base.riken.jp/sw/item/crib102u0rib102u10000078600>
- ATH-SMO_Ortholog03810**
[Ortholog \(A.thaliana and Selaginella moellendorffii\)](#)
linked page in 1 steps
[AT1G78600](#)
Name: **AT1G78600**
Last Update Date 2009-09-02 03:55:55
<https://blade2gw2.base.riken.jp/sw/item/cria54s6ria54u3810i>
- At-Os_Ortholog9151**
[RIKEN Ortholog\(A.thaliana and Rice\)](#)
linked page in 1 steps
[AT1G78600](#)
Name: **AT1G78600**
Last Update Date 2008-06-20 18:53:33
<https://blade2gw2.base.riken.jp/sw/item/crib127u1rib127u9151i>
- RAFL09-27-A12**
[RAFL cDNA](#)
linked page in 1 steps
[AT1G78600](#)
Name: **AT1G78600**
Last Update Date 2008-09-01 20:44:22
<https://blade2gw2.base.riken.jp/sw/item/cria11s1ria11u8059i>
- ATH-SBI_Ortholog02614**
[Ortholog\(A.thaliana and Sorghum bicolor\)](#)
linked page in 1 steps
[AT1G78600](#)
Name: **AT1G78600**
Last Update Date 2009-09-02 07:30:20
<https://blade2gw2.base.riken.jp/sw/item/cria54s4ria54u2614i>
- ATH-VVI_Ortholog05317**
[Ortholog\(A.thaliana and Vitis vinifera\)](#)
linked page in 1 steps
[AT1G78600](#)
Name: **AT1G78600**

DBCLSの横断検索システムとの連携化も予定

DBCLSへのデータ提供体制



理研サイネスの公開基準としてクリエイティブコモンズを採用

理研外コミュニティへの貢献体制

- Plant and Cell Physiology誌においてDatabase Issue 刊行予定(2011年)
- TAIR(シロイヌナズナ)の国際分業化の一翼に参加
- マウス関連データの国際的な交換体制づくり
- データベース上で国際コンテストを開催(GenoCon等)※
- 理研データの公開化・共有化に向けたコンセンサス※

※別添の理研ニュースをご覧ください(参考資料2)