H23年度 統合化推進プログラム進捗報告会

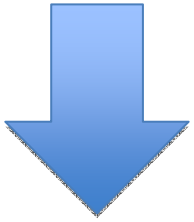
ライフサイエンスデータベース統合推進事業
統合化推進プログラム
平成23年度　進捗報告会

ゲノム・メタゲノム情報を基盤とした微生物DBの統合

東京工業大学大学院生命理工学研究科
黒川　顕

研究開発の目標・ねらい

ゲノム情報を核として様々な微生物学上の知識を統合し、幅広い分野での微生物学の発展に資することのできる「微生物エンサイクロペディア」の構築を目標とする。



微生物学分野のオミックス研究の発展に寄与
データ駆動型研究による新しい仮説の提唱

研究開発メンバー

東京工業大学

黒川　顕：微生物DBにおける研究統括
小西史一：スパコンにおける解析システムの開発および実装
森　宙史：ゲノム、メタゲノムDB、メタデータの構築
吉野弘二，竹原潤一：メタデータDBの構築

国立遺伝学研究所

中村保一：微生物アノテーションリファレンスの整備と共用化
菅原秀明：微生物ゲノム基盤情報資源の共用化
神沼英里：KazusaAnnotationの拡張
藤澤貴智：モデル微生物情報の高度化

基礎生物学研究所

内山郁夫：比較ゲノム解析に立脚した微生物ゲノム情報の統合化
千葉啓和：MBGDの統合化

統合データベースセンター（技術アドバイザー）

岡本忍，片山俊明，川島秀一，川本祥子，山本泰智：技術協力

研究開発内容（全体概要）

- 日本をはじめ世界中に散在している細菌の各種オミックス情報を広く収集
- 上記データをホモロジー、オーソロジーに基づいて整理し、遺伝子、ゲノム（生物種）、環境の3つの軸に沿って整理統合
- 3つの軸に関わる、遺伝子機能、分類学的情報、菌株保存情報、表現型情報などの知識を整理し、ゲノム情報を核として統合
- 広く研究者コミュニティからのフィードバックを得るための仕組みを開発
- 研究者が活用しやすいインターフェース等を整備

目標とするデータベース

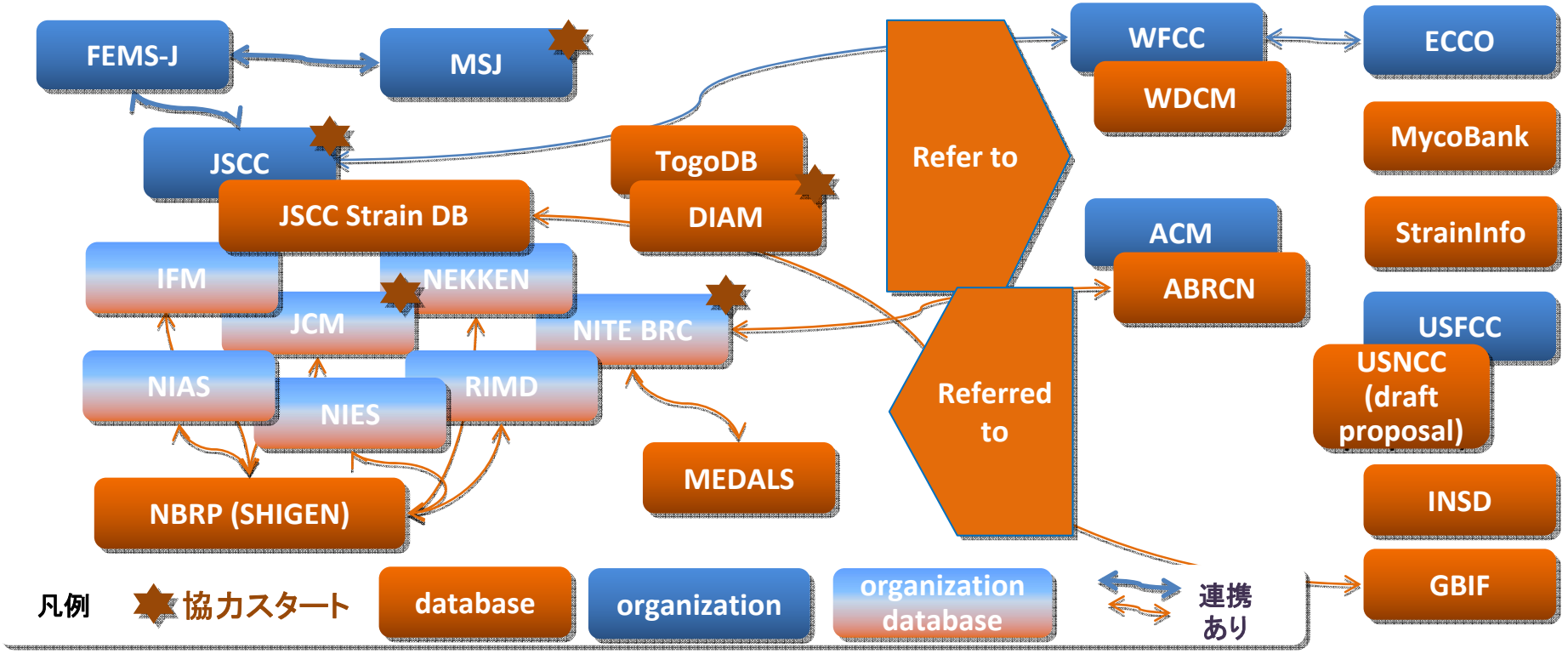
まず研究基盤となる6項目のDBを整備し、個々のDBを
ゲノム情報を核としてセマンティック技術でシームレスに連携する

1. 分類学的情報（16S rDNAを含む）
2. 菌株保存情報（培養条件含む）
3. モデル微生物（大腸菌、枯草菌、シアノバクテリア、放線菌）における高品質データ
4. 各種オミックスデータ
5. オーソログ遺伝子情報
6. メタゲノムデータ

H23年度開発計画

- 分類学的情報および保存菌株情報の整理
- モデル微生物ゲノムアノテーションの高度化
- 各種オミックスデータの整理
- オーソログ遺伝子情報の統合化検討
- メタゲノムデータの整理

Strain data are partially interlinked in many ways.
User need the **navigation that MicroTOGO will provide**



ABRCN: http://www.abrcn.net/
ACM: Asian Consortium for the Conservation and Sustainable Use of Microbial Resources
DIAM: DataBiosafety for the Industrrial Applciation of Microbes
ECCO: European Culture Collection Organization
FEMS-J: Federation of Microbiological Society of Japan
GBIF: Global Biodiversity Information Facility
IFM: 千葉大学真菌医学研究センター
INSD: DDBJ/EMBL/GenBank
JCM: Japan Collection of Microorganisms
JSCC: Japan Society for Culture Collections
MSJ: The Mycological Society for Japan

NBRP: National BioResource Project
NEKKEN:長崎大学熱帯医学研究所
NIAS:（独）農業生物資源研究所ジーンバンク微生物遺伝資源部門
NIES:（独）国立環境研究所微生物系統保存施設
NITE BRC: NITE Biological Resource Center
RIMD:阪大微研感染症国際研究センター病原微生物資源室
StrainInfo: http://www.straininfo.net/
USCCN: US Culture Collection Network
USFCC: US Federation for Culture Collections
WDCM: WFCC-MIRCEN World Data Center for Micoorganisms
WFCC: World Federation for Culture Collections

National Institute of Technology and Evaluation
独立行政法人
製品評価技術基盤機構
Biotechnology Center
NBRC
NITE Biological Resource Center

NBRC No.	NBRC 102086
Scientific Name of this Strain	***Burkholderia multivorans*** Vandamme et al. 1997
Synonymous Name	
Type Strain	
History	NITE <- Meijo Univ. (S. Ichihara, 25)
Other Culture Collection No.	
Other No.	NITE 02208=25
Rehydration Fluid	702
Medium	802
Cultivation Temp.	25 C
Source of Isolation	
Locality of Source	
Country of Origin	Japan
Biosafety Level	
Applications	Phenylacetic acid;degradation
Mating Type	
Genetic Marker	
Plant Quarantine No.	
Animal Quarantine No.	
Herbarium No.	
Restriction	
Comment	
References	
Sequences	16S rDNA

株数：約16,000株
単離元：1,627
培地情報：432種類

ゲノムデータおよび
メタゲノムデータ等と
統合を目標にRDF化

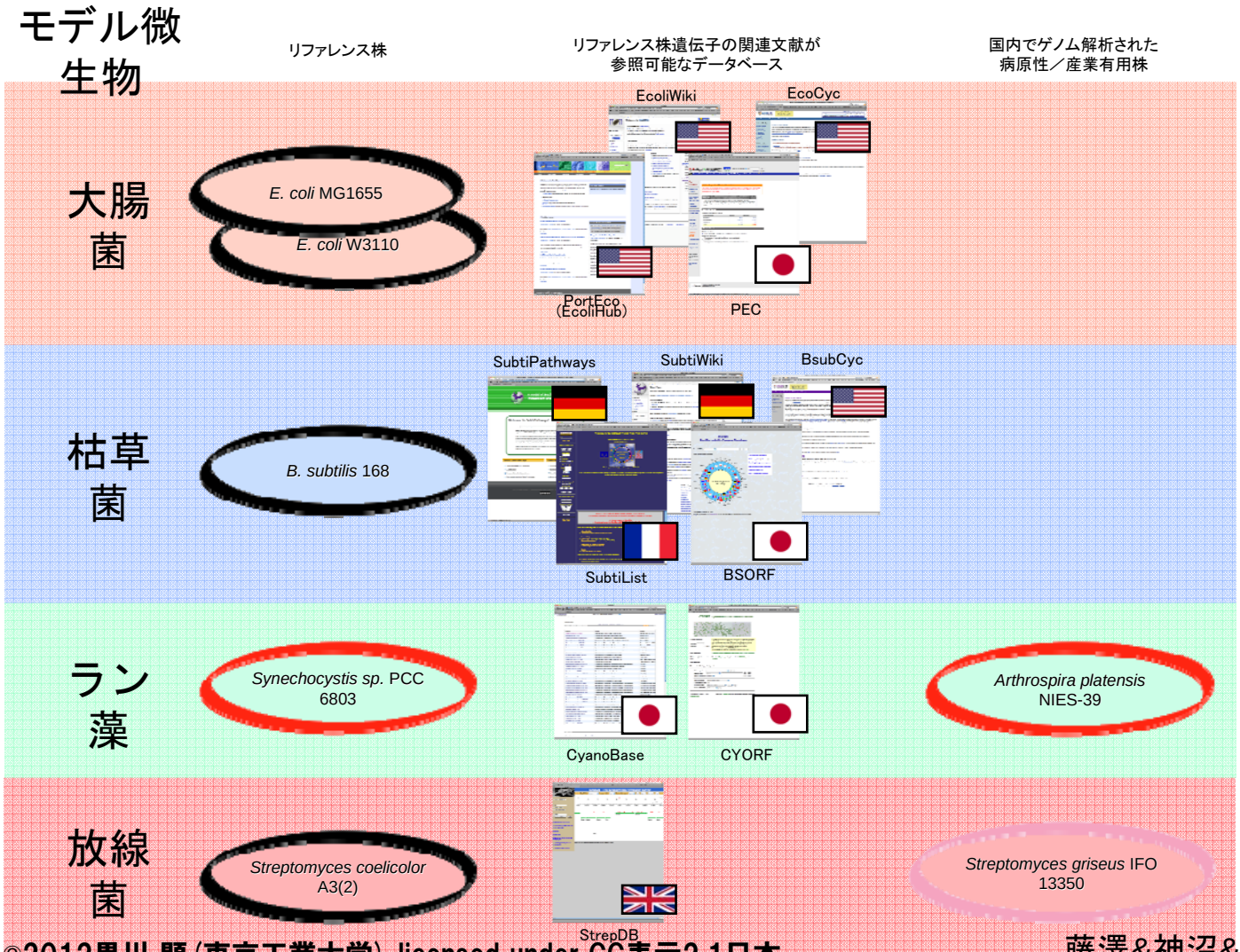
24年度はJCMも対象と
する（約14,000株）

菅原@遺伝研，市原@NBRC

H23年度開発計画

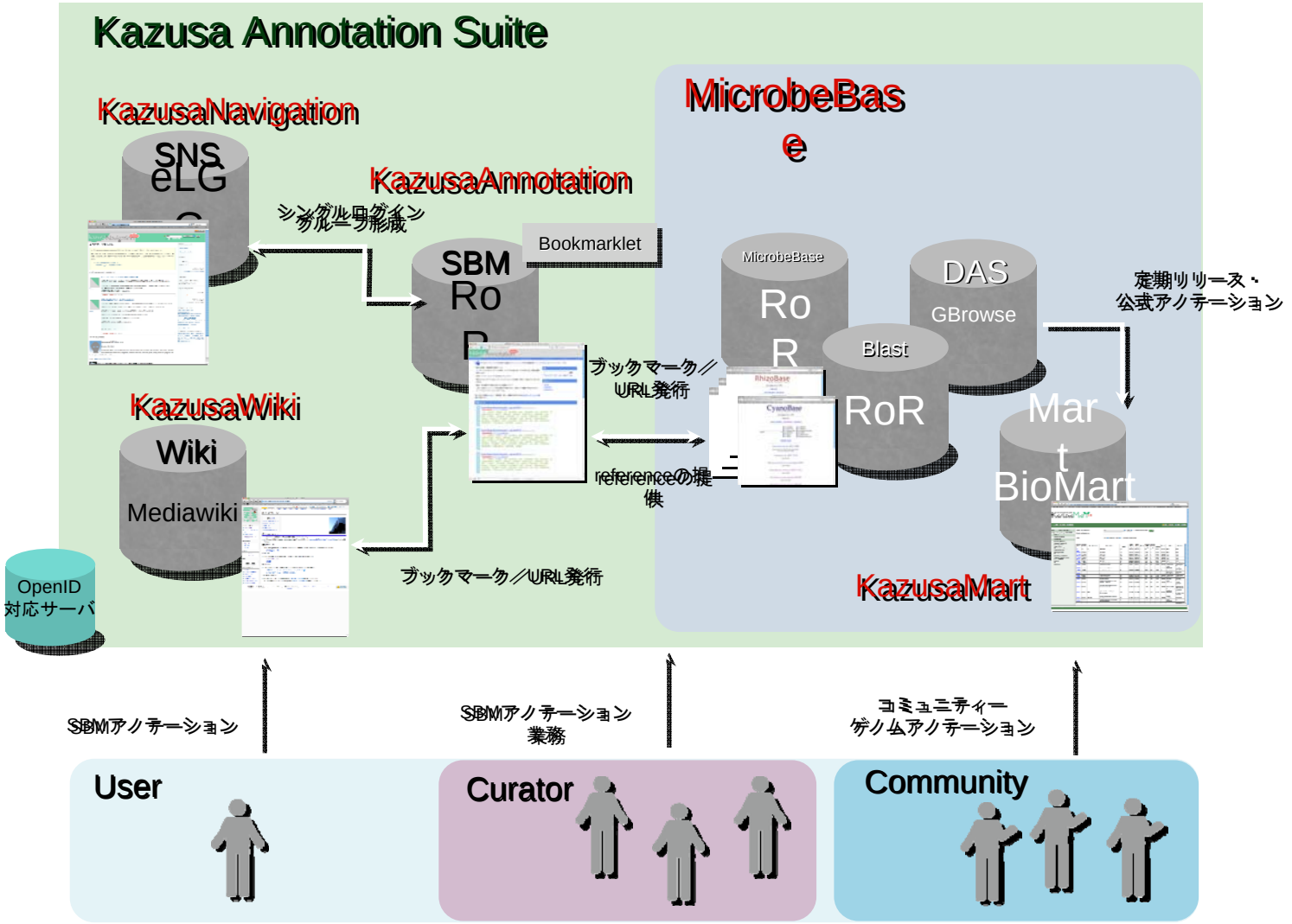
- 分類学的情報および保存菌株情報の整理
- モデル微生物ゲノムアノテーションの高度化
- 各種オミックスデータの整理
- オーソログ遺伝子情報の統合化検討
- メタゲノムデータの整理

文献情報に基づくモデル微生物ゲノムデータベースの現状



藤澤&神沼&中村@遺伝研

高度アノテーション情報集積データベース



ソーシャルブックマークシステムを利用した **KazusaAnnotation (http://a.kazusa.or.jp)** をはじめとする情報集積データベースを運用し、キュレーターによるゲノムデータベース上への遺伝子に言及した論文情報の蓄積を継続している。また、微生物ゲノムプロジェクトにおいて本システムを利用したコミュニティゲノムアノテーションへの応用も実証した。

「Kazusa Annotation Suite」系を拡張し微生物ゲノム基盤情報を拡充する

1. リファレンスとして重要な菌株あるいは現象について、信頼性の高いマニュアルキュレーションに基づいた既存の情報の高度化。**放線菌のアノテーション・キュレーションを開始**
2. 本システムで蓄積した信頼性の高い情報を元に、遺伝子の機能の記述などの固有表現を抽出するプログラムを開発、運用
3. コストと時間のかかる手動アノテーションの自動化を支援する系を開発、提供
4. 研究コミュニティに対しゲノムアノテーション支援を実施

来年度は、引続き放線菌ゲノムアノテーションの高度化を図るとともに、大腸菌、枯草菌、シアノバクテリアにも着手する

藤澤&神沼&中村@遺伝研

TogoAnnotationへの放線菌データの集積状況

Annotation Project	Entity	Database	Genome	Num of PubmedIDs	Num of URLs	Num of Bookmarks
Gene Attribute (GA)	gene	nih	*Streptomyces griseus* IFO 13350	22	91	2166
Gene Group (GG)	gene cluster	nih	*Streptomyces griseus* IFO 13350	8	1	65
	operon	nih	*Streptomyces griseus* IFO 13350	3	1	14
	regulon	nih	*Streptomyces griseus* IFO 13350	5	1	21
Strain Information (SI)	strain	nih	*Streptomyces griseus* IFO 13350	14	3	32

2012年02月21日現在

藤澤&神沼&中村@遺伝研

アノテーション・キュレーション実施体制

- 中村保一 (遺伝研)
  - 岡本忍 (DBCLS) -CyanoBase/RhizoBase
    - 遠隔雇用キュレータ(青字は男性)
      - 博士: 吉村(東京)山本(京都)矢野(川崎)
        鐘ヶ江(東京)
      - 修士: 谷中(つくば)　笠井(つくば)

  - 藤沢貴智 (遺伝研) - 統合DB微生物、システム運用
    - 遠隔雇用キュレータ
      - 博士: 照井(銚子)桧原(東京)
      - 修士: 加藤(名古屋)石井(奈良)

TogoAnnotation

H23年度開発計画

- 分類学的情報および保存菌株情報の整理
- モデル微生物ゲノムアノテーションの高度化
- 各種オミックスデータの整理
- オーソログ遺伝子情報の統合化検討
- メタゲノムデータの整理

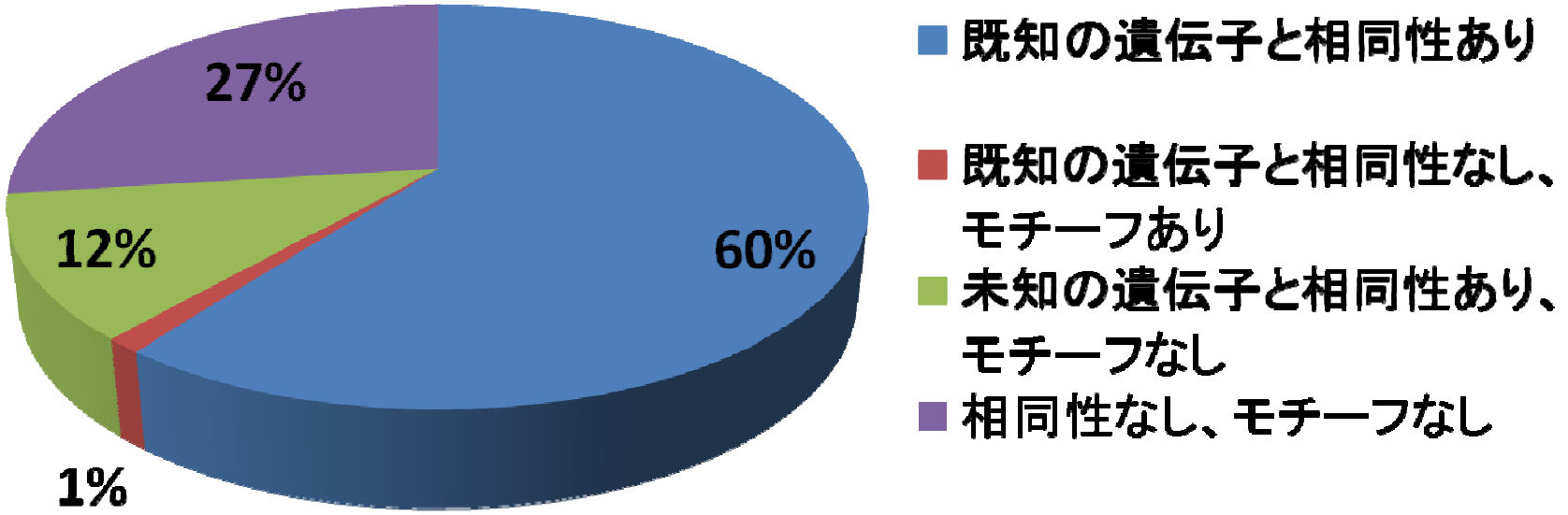
GTPS概要と2011年度の統計

- GTPS概要
  DDBJ / INSD の細菌・古細菌ゲノム配列を再アノテーションしたデータベース

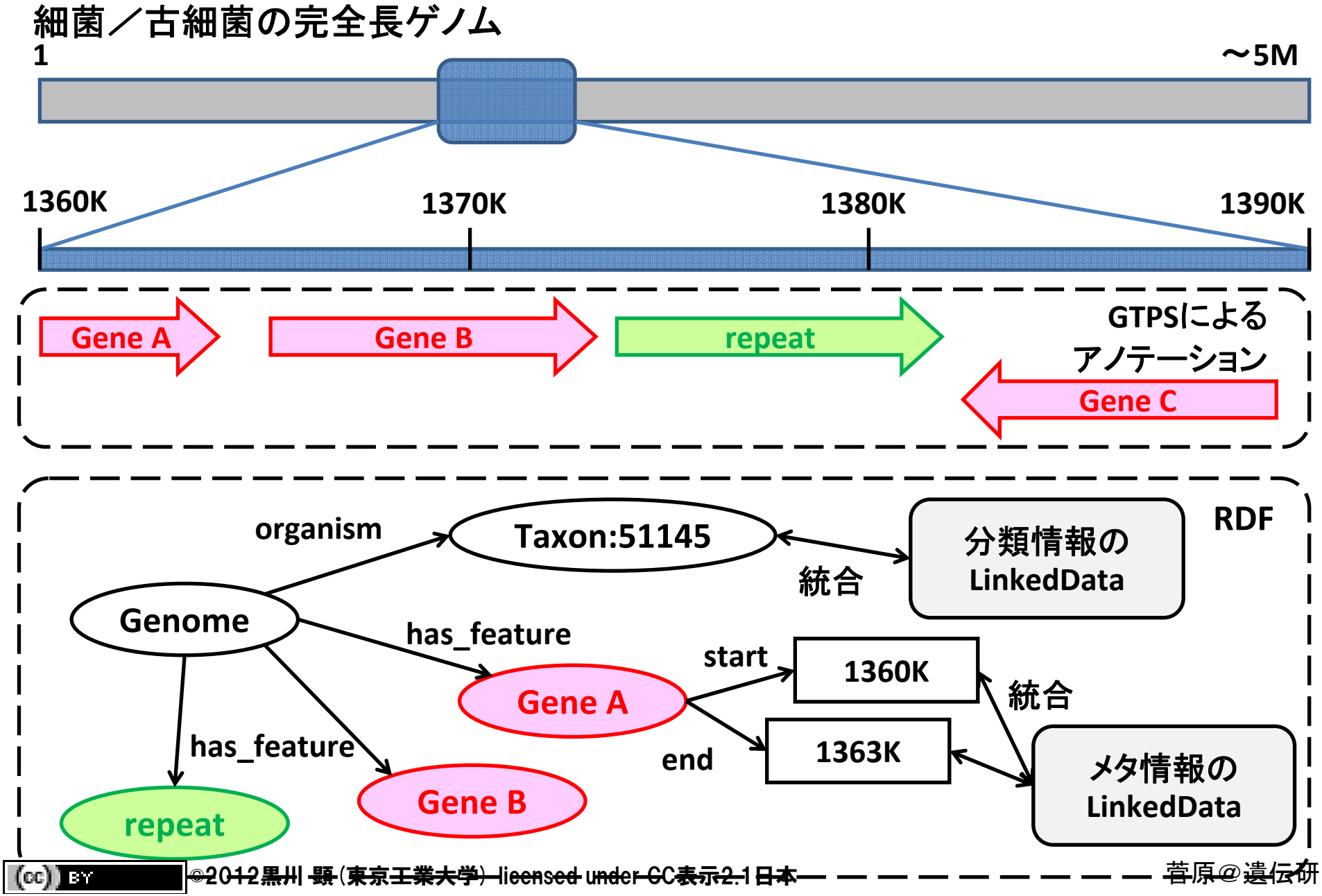
- 再アノテーション方法
  Glimmer, BLAST, InterProScan など

- 対象件数：　菌株数：1,743、 DDBJエントリ数：3,265

- ORF数（総数） 約780万ORF

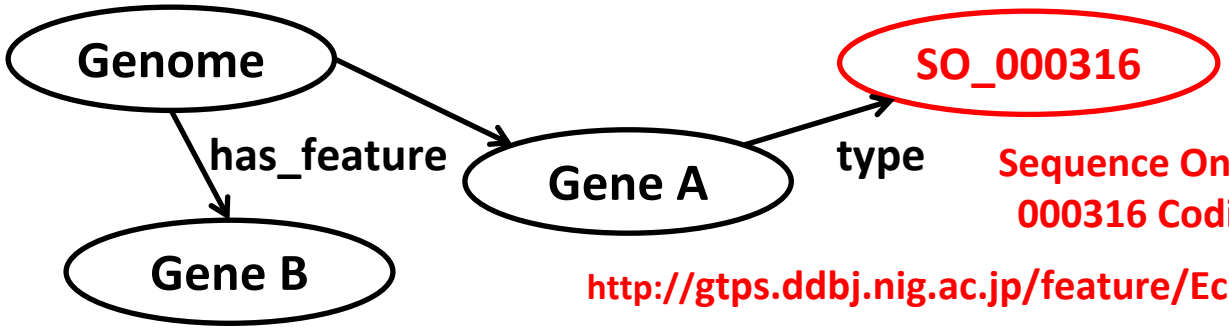


GTPSのRDF化による統合イメージ



GTPSのRDF（トリプル）例

■ フィーチャにURI設定、シーケンスオントロジー付与



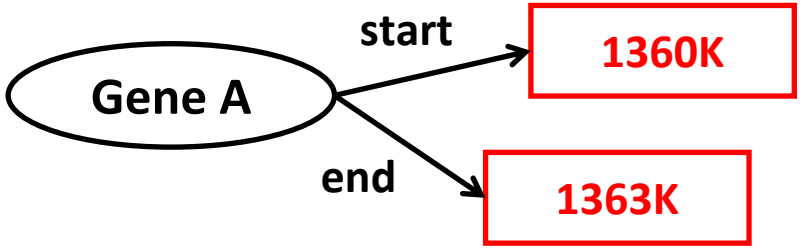
フィーチャの
外部参照が容易に

Sequence Ontology
000316 Coding sequence

http://gtps.ddbj.nig.ac.jp/feature/Ecol_K12_MG1655:ST1528

http://gtps.ddbj.nig.ac.jp/feature/Ecol_K12_MG1655:ST2749

■ 各フィーチャのゲノム位置（開始、終了）をトリプルに



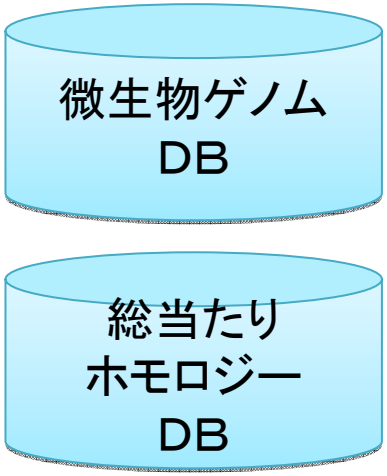
位置情報での統合が容易に

GTPSのRDF化により他のオミックス情報との統合が可能
来年度はメタゲノム情報との統合を開始する

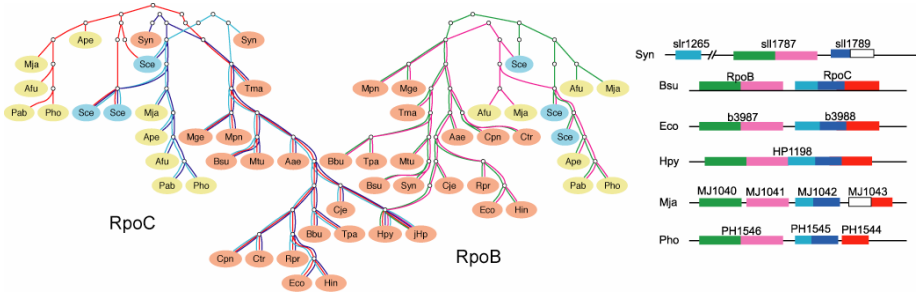
H23年度開発計画

- 分類学的情報および保存菌株情報の整理
- モデル微生物ゲノムアノテーションの高度化
- 各種オミックスデータの整理
- オーソログ遺伝子情報の統合化検討
- メタゲノムデータの整理

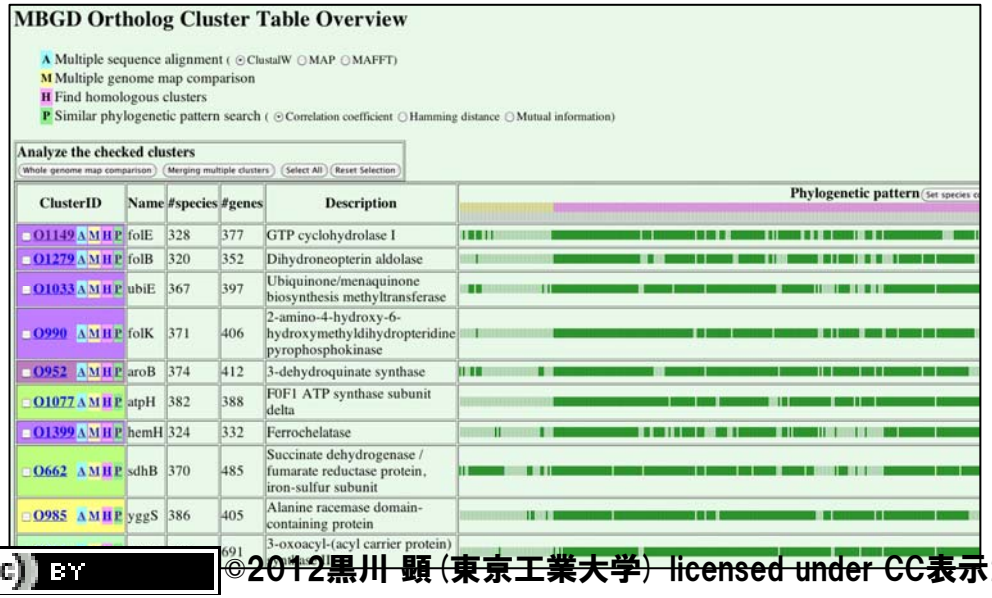
微生物比較ゲノムデータベースMBGD



オーソログ分類アルゴリズムDomClust



オーソログテーブル



MBGD Ortholog Cluster Table Overview

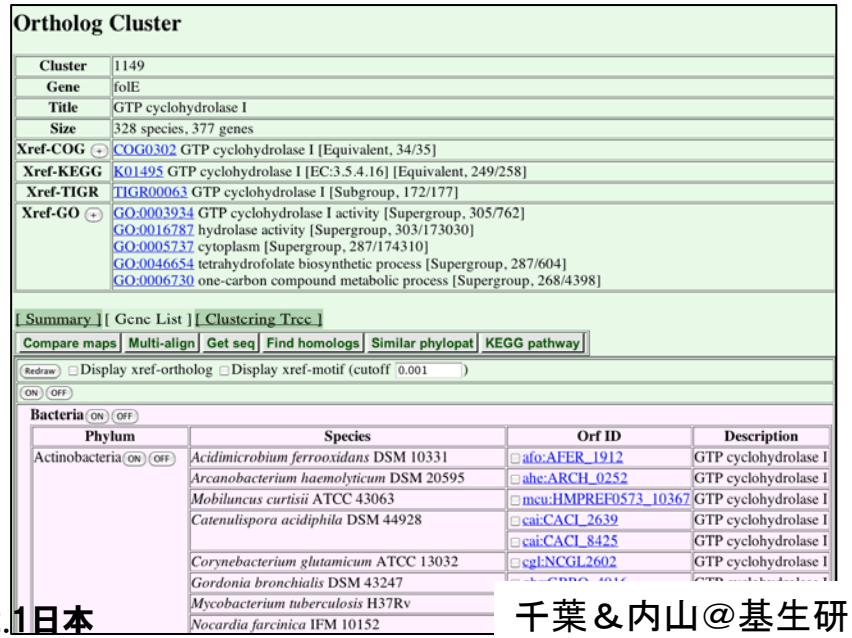
A Multiple sequence alignment (⊙ClustalW ○MAP ○MAFFT)
M Multiple genome map comparison
H Find homologous clusters
P Similar phylogenetic pattern search (⊙Correlation coefficient ○Hamming distance ○Mutual information)

Analyze the checked clusters

Whole genome map comparison | Merging multiple clusters | Select All | Reset Selection

ClusterID	Name	#species	#genes	Description	Phylogenetic pattern
O1149 A M H P	folE	328	377	GTP cyclohydrolase I	
O1279 A M H P	folB	320	352	Dihydroneopterin aldolase	
O1033 A M H P	ubiE	367	397	Ubiquinone/menaquinone biosynthesis methyltransferase	
O990 A M H P	folK	371	406	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	
O952 A M H P	aroB	374	412	3-dehydroquinate synthase	
O1077 A M H P	atpH	382	388	F0F1 ATP synthase subunit delta	
O1399 A M H P	hemH	324	332	Ferrochelatase	
O662 A M H P	sdhB	370	485	Succinate dehydrogenase / fumarate reductase protein, iron-sulfur subunit	
O985 A M H P	yggS	386	405	Alanine racemase domain-containing protein	
			691	3-oxoacyl-(acyl carrier protein)	

オーソロググループ



Ortholog Cluster

Cluster	1149
Gene	folE
Title	GTP cyclohydrolase I
Size	328 species, 377 genes
Xref-COG	COG0302 GTP cyclohydrolase I [Equivalent, 34/35]
Xref-KEGG	K01495 GTP cyclohydrolase I [EC:3.5.4.16] [Equivalent, 249/258]
Xref-TIGR	TIGR00063 GTP cyclohydrolase I [Subgroup, 172/177]
Xref-GO	GO:0003934 GTP cyclohydrolase I activity [Supergroup, 305/762] GO:0016787 hydrolase activity [Supergroup, 303/173030] GO:0005737 cytoplasm [Supergroup, 287/174310] GO:0046654 tetrahydrofolate biosynthetic process [Supergroup, 287/604] GO:0006730 one-carbon compound metabolic process [Supergroup, 268/4398]

[Summary] [Gene List] [Clustering Tree]

Compare maps | Multi-align | Get seq | Find homologs | Similar phylopat | KEGG pathway

Redraw ☐Display xref-ortholog ☐Display xref-motif (cutoff 0.001)

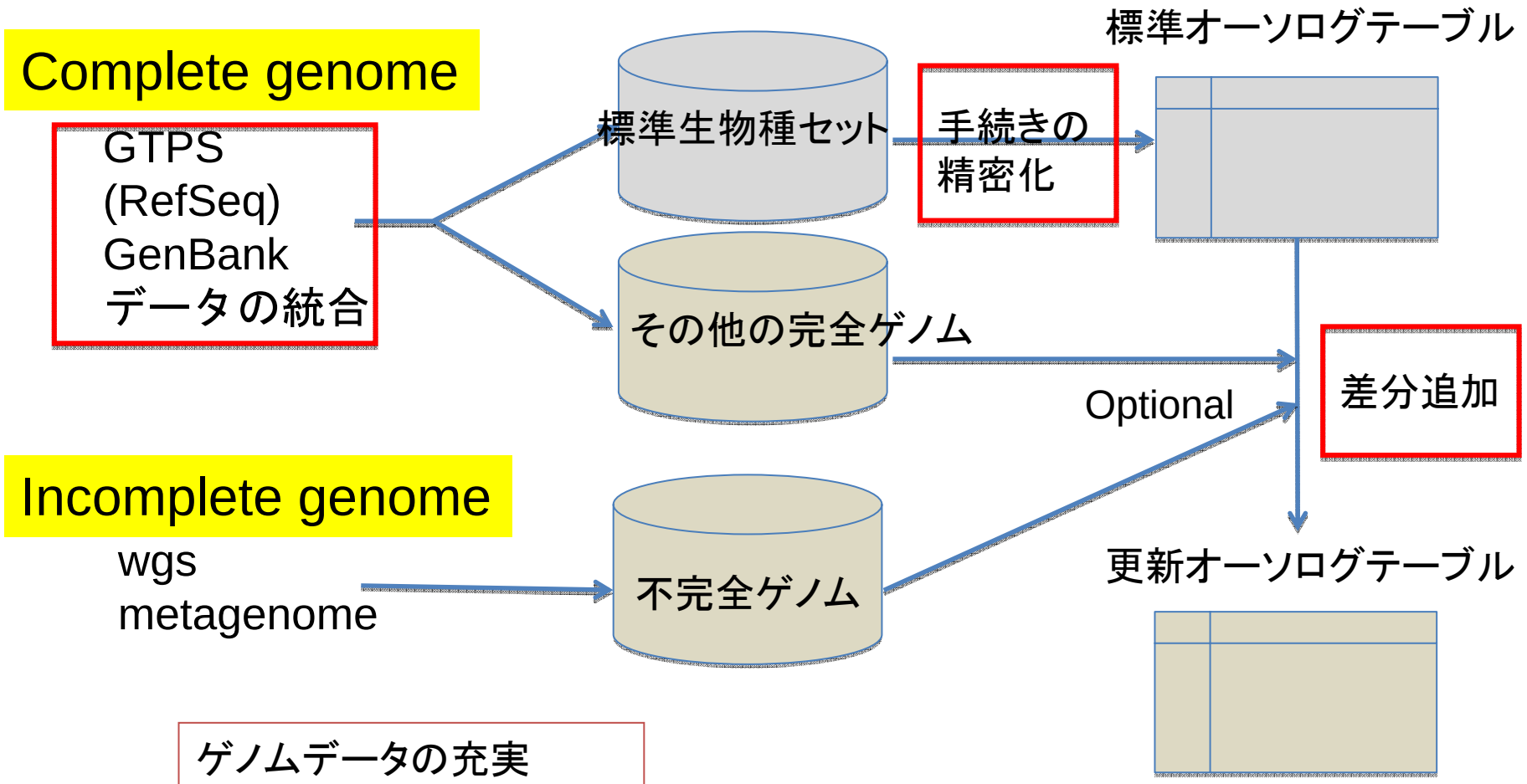
ON OFF

Bacteria ON OFF

Phylum	Species	Orf ID	Description
Actinobacteria ON OFF	*Acidimicrobium ferrooxidans* DSM 10331	afo:AFER_1912	GTP cyclohydrolase I
	Arcanobacterium haemolyticum DSM 20595	ahe:ARCH_0252	GTP cyclohydrolase I
	Mobiluncus curtisii ATCC 43063	mcu:HMPREF0573_10367	GTP cyclohydrolase I
	Catenulispora acidiphila DSM 44928	cai:CACI_2639	GTP cyclohydrolase I
		cai:CACI_8425	GTP cyclohydrolase I
	Corynebacterium glutamicum ATCC 13032	cgl:NCGL2602	GTP cyclohydrolase I
	Gordonia bronchialis DSM 43247		
	Mycobacterium tuberculosis H37Rv		
	Nocardia farcinica IFM 10152		

千葉＆内山@基生研

対象ゲノムデータの拡大と効率的なオーソログ解析



ゲノムデータの充実
データ更新の迅速化
オーソログ分類の精密化

千葉＆内山@基生研

GTPS/GenBank/RefSeqの対応付け

- 染色体レベル、遺伝子レベルで3データベース間の対応付け手続きは作成済み。
- 実際のデータ更新については来年度の次期バージョンに合わせて行う。
- 今のところ、GTPSのGrade X以外を標準遺伝子セットとして採用し、GTPSにないゲノムについてはRefSeq, GenBankの順に取り込む方針。

オーソログ分類の精密化

- DomClustによる分類結果を、マルチプルアライメントに基づく方法によって改善する
- ドメイン分割の改善（今年度実施）
  - ドメイン境界の改善
  - 不必要なドメイン分割の取り消し
- グルーピングの改善（来年度実施）
  - マルチプルアライメントに基づく系統樹を使って改善
  - 機能アノテーションの情報を使った改善

H23年度開発計画

- 分類学的情報および保存菌株情報の整理
- モデル微生物ゲノムアノテーションの高度化
- 各種オミックスデータの整理
- オーソログ遺伝子情報の統合化検討
- メタゲノムデータの整理

メタゲノムメタデータの集計結果

	サンプル数	メタデータのカテゴリー数	メタデータカテゴリーの例
ヒト共生細菌群集	69,968	85	Age , Sex ,Disease stage , Country , Body Habitat , Diet 等
環境共生細菌群集	4,054	627	pH , Temperature , Wind Speed , Dissolved Oxygen 等

（2011年7月7日時点）

	Age	Body Habitat	Body Site	Collection Date	Country	Disease Stage
Sample 1	22	Feces		2008		Obese
Sample 2					Japan	
Sample 3			Scalp			
Sample 4			Skin		USA	
Sample 5	1years		Gut	2011/8/8		Healthy

登録されるカテゴリーや値の語彙は登録者によってバラバラ

データを検索する際の問題点と解決策

ヒト腸内環境に関連した語彙

- human gut (ヒト消化官)
- human digestive tract (ヒト消化器官)
- human gastrointestinal tract (ヒト消化官)
- human intestine (ヒト腸) : gutの一部
- human intestinal lumen (ヒト腸管腔) : gutの一部
- human colon (ヒト大腸) : gutの一部
- human stomach (ヒト胃) : gutの一部
- human feces (ヒト糞便) : gutと関連
- human stool (ヒト糞便) : gutと関連

フリーワード検索の問題点

- 同義語や関連語が多数存在するため**対象の選択的取得が不可能**
- 文字の一致だけを調べるため**不要な情報も取得してしまう**

メタゲノムデータを網羅的に取得するためには

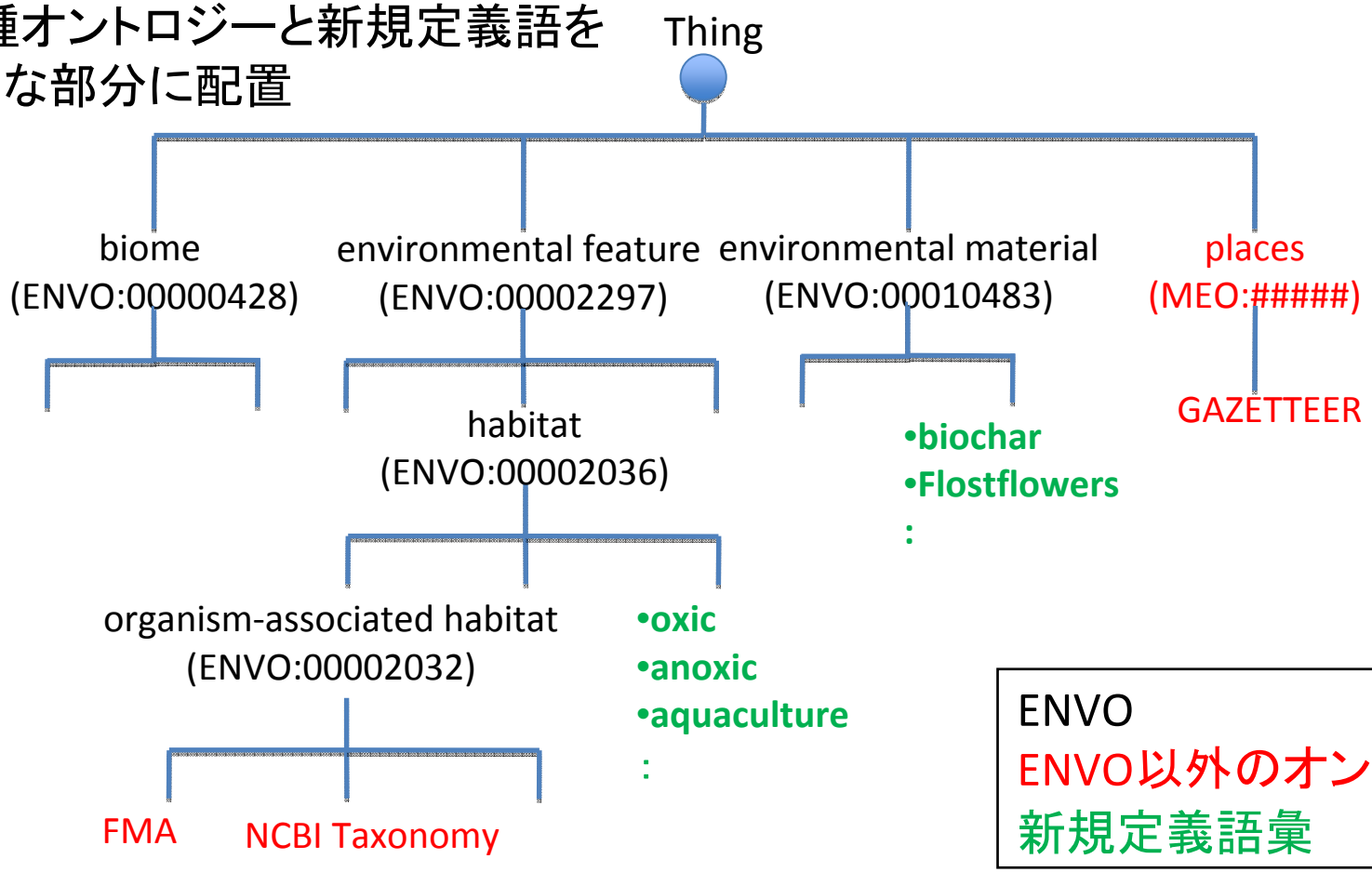
単語の意味、単語間の意味的関係性、階層性を定義する必要がある

＝オントロジー

Metagenome/Microbes Environmental Ontology (MEO)

微生物の生息環境メタデータのオントロジー

- ENVOを基礎とした構造
- 各種オントロジーと新規定義語を適切な部分に配置



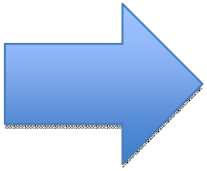
ENVO
ENVO以外のオントロジー
新規定義語彙

BioPortal http://bioportal.bioontology.org/ 及び
プロジェクトページ http://mdb.bio.titech.ac.jp/meo/ でMEOを公開

Metagenome/Microbes Environmental Ontology (MEO)

微生物の生息環境メタデータのオントロジー

- MEOの統計情報
  - ターム数：1,318,245 ターム
  - 新規定義ターム数：10 ターム
  - Base URI: http://mdb.bio.titech.ac.jp/meo/meo.owl
  - ファイル形式：OWL
  - ファイルサイズ：1.5GB



BioPortalへ登録（2012.2.22）

MEOをBioPortalへ登録

実際の検索クエリと検索結果

```
SPARQL

prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix srs: <http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?sample=>
prefix envo: <http://purl.obolibrary.org/obo/>
prefix gaz: <http://purl.org/obo/owl/GAZ#>
prefix fma: <http://sig.uw.edu/fma#>
prefix dc: <http://purl.org/dc/elements/1.1/>
prefix ncbitaxon: <http://purl.org/obo/owl/NCBITaxon#>
prefix meo: <http://mdb.bio.titech.ac.jp/meo/>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?SRS ?label
FROM <http://localhost:8890/DAV/test/meo2>
WHERE {
?SRSID meo:environmentalMaterial ?blanknode.
?SRSID rdfs:label ?SRS.
?blanknode rdf:type envo:ENVO_00001998.
?blanknode rdfs:label ?label.
};
```

Query

推論によって関連する語も同時に検索される

SRS074564	soil
SRS074565	soil
SRS074566	soil
SRS074567	soil
SRS074568	soil
SRS074569	soil
SRS074570	soil
SRS074571	soil
SRS074572	soil
SRS074573	soil
SRS074574	soil

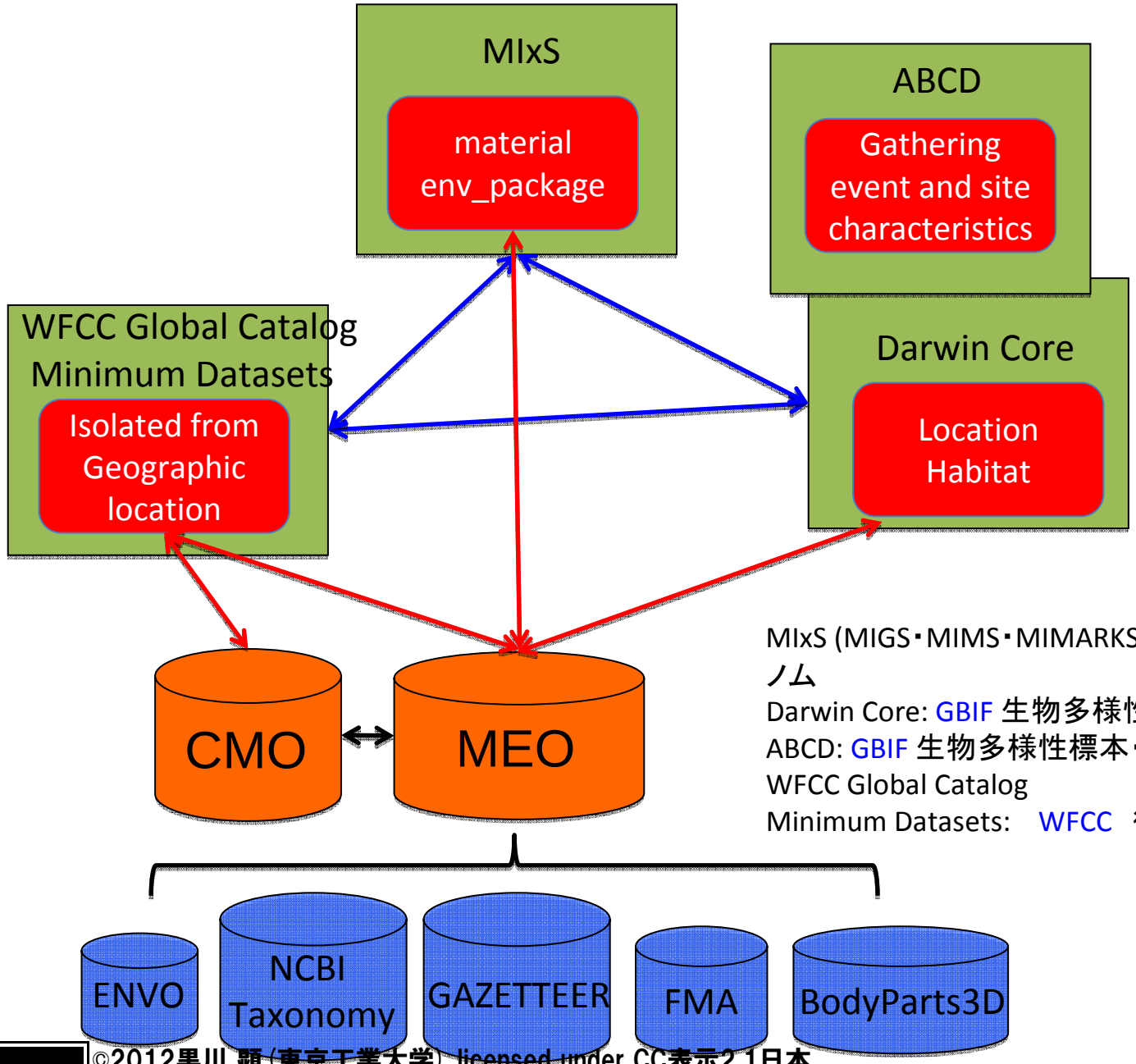
109 Rows. -- 6 msec.

SRS074573	soil
SRS074574	soil
ERS025421	rhizosphere
ERS009173	rhizosphere
ERS009174	anthrosol
ERS009177	agricultural soil
ERS009178	forest soil
ERS009185	oil contaminated soil
ERS040660	greenhouse soil
ERS017979	clay soil
ERS017981	clay soil

164 Rows. -- 161 msec.

Result

国際データ標準化グループとの連携



MIxS (MIGS・MIMS・MIMARKS): GSC ゲノム・メタゲノム
Darwin Core: GBIF 生物多様性標本・観測
ABCD: GBIF 生物多様性標本・観測
WFCC Global Catalog
Minimum Datasets: WFCC 微生物保存株

東工大&遺伝研

http://microbedb.jp/MDB/

Microbe DB .JP

http://microbedb.jp/

Microbe DB.jp

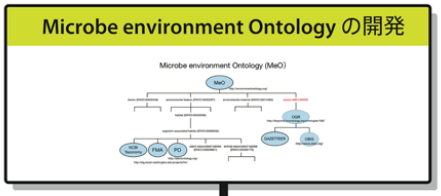
MicrobeDB プロジェクトでは様々な微生物学上の知識を、ゲノム情報を核として遺伝子、系統、環境の３つの軸に沿ってセマンティックウェブの技術を駆使して整理統合し、幅広い分野での微生物学の発展に資することの出来るデータベースの構築を目標としています。

遺伝子名 or 系統名 or 環境メタデータ or 配列

Ontology

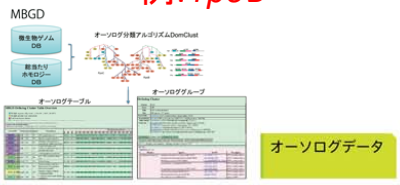
オントロジー: 検索タームの柔軟化＆明確化

Microbe environment Ontology の開発

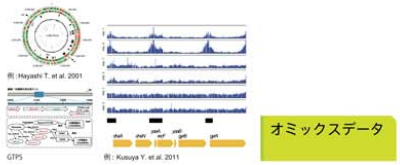


遺伝子

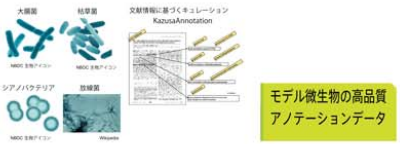
例: *rpo*B



オーソログデータ



オミックスデータ



モデル微生物の高品質アノテーションデータ

系統

例: *Escherichia coli*

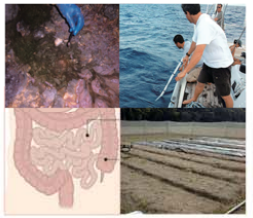
系統分類データ

菌株データ

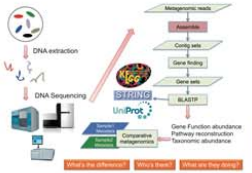
菌株保存情報（培養条件含む）

環境

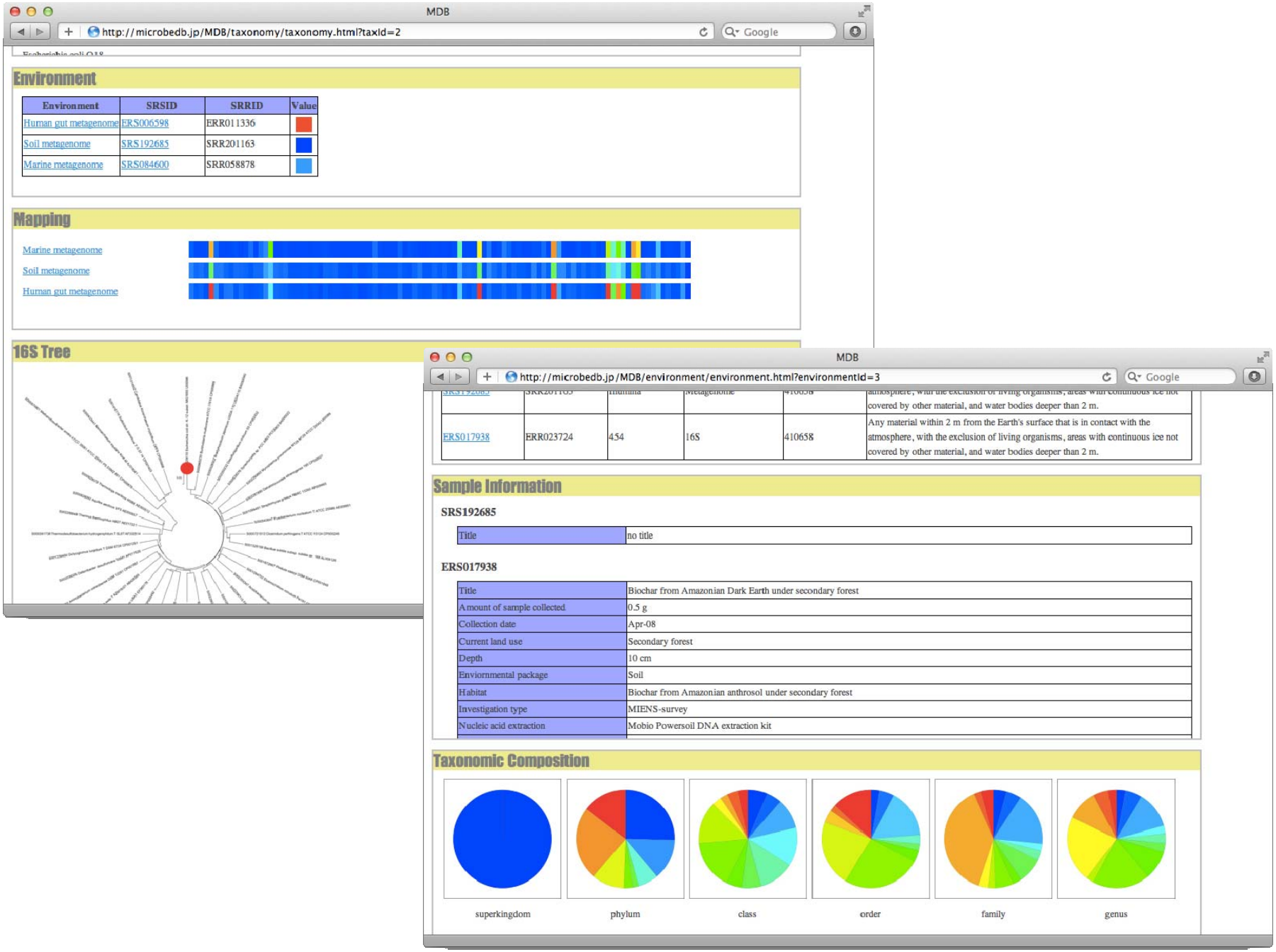
例: soil



環境のメタデータ



メタゲノムデータ



H24年度の計画

- 保存菌株情報（NBRC, JCM）のRDF化
- 培地情報オントロジーCMOの整備（w/ DBCLS）
- GTPSのRDF化および各オミックスデータの統合
- MBGDオーソログ情報のRDF化
- 放線菌以外のモデル微生物ゲノムアノテーション高度化
- メタゲノムデータのRDF化およびGTPSとの統合
- 各種アプリケーション、結果表示要素「Stanza」の標準化および開発（w/ DBCLS）

直面している課題

- 文献からの情報抽出
  - 細菌種名、菌株番号、遺伝子名などの文献からの自動抽出
- データストアおよび検索技術（Triplestore）
- URIの一元化、安定維持管理
- 大規模データに耐えうるDBシステム
  - ヒトメタゲノムだけで現在約600億配列
- 計算機資源の不足

現在保有しているDB

- CyanoBase, RhizoBase, KazusaAnnotation, KazusaMart
- HumanMetaBodyMap
- GIB-M, GTPS
- MBGD

構築中または構築予定のDB

- TogoAnnotation (モデル微生物群)+RDF
- GTPS, Metagenome, MBGD 各RDF DB
- NBRC, JCM 等の保存菌株 RDF DB
- MEO, CMO 各オントロジー

- 微生物統合DB：MicrobeDB

DB公開の状況

- MicrobeDB.jpを立上げ、限定的なデータのみでMicrobeDBのプロトタイプを公開中（http://microbedb.jp/MDB/）
- TogoAnnotationの公開（http://genome.microbedb.jp/）

- GTPSのRDF化が完了した段階で、DBの拡充を図る

H23年度 主な活動状況

- 微生物統合DB全体会議（3回開催）
- RDF会議（7回開催）
- BioHackathon11.11 (11/21-25@修善寺）

- ゲノム微生物学会シンポジウム（8/20-21@仙台）
- 分子生物学会（12/13-16@横浜）
- 微生物生態学会・微生物多様性部会（1/24@東京）
- DwC-MIxS alignment workshop（2/27-29@Oxford）
- The 13th workshop of the GSC (3/4-7@Shenzhen)