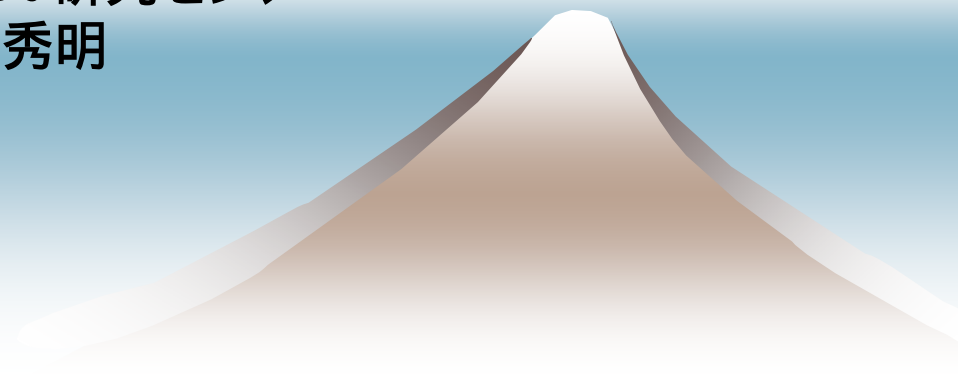


生物情報資源の相互運用性

〒411-8540 静岡県三島市谷田1111
大学共同利用機関法人 情報・システム研究機構 (ROIS)
国立遺伝学研究所
生命情報・DDBJ研究センター
菅原秀明



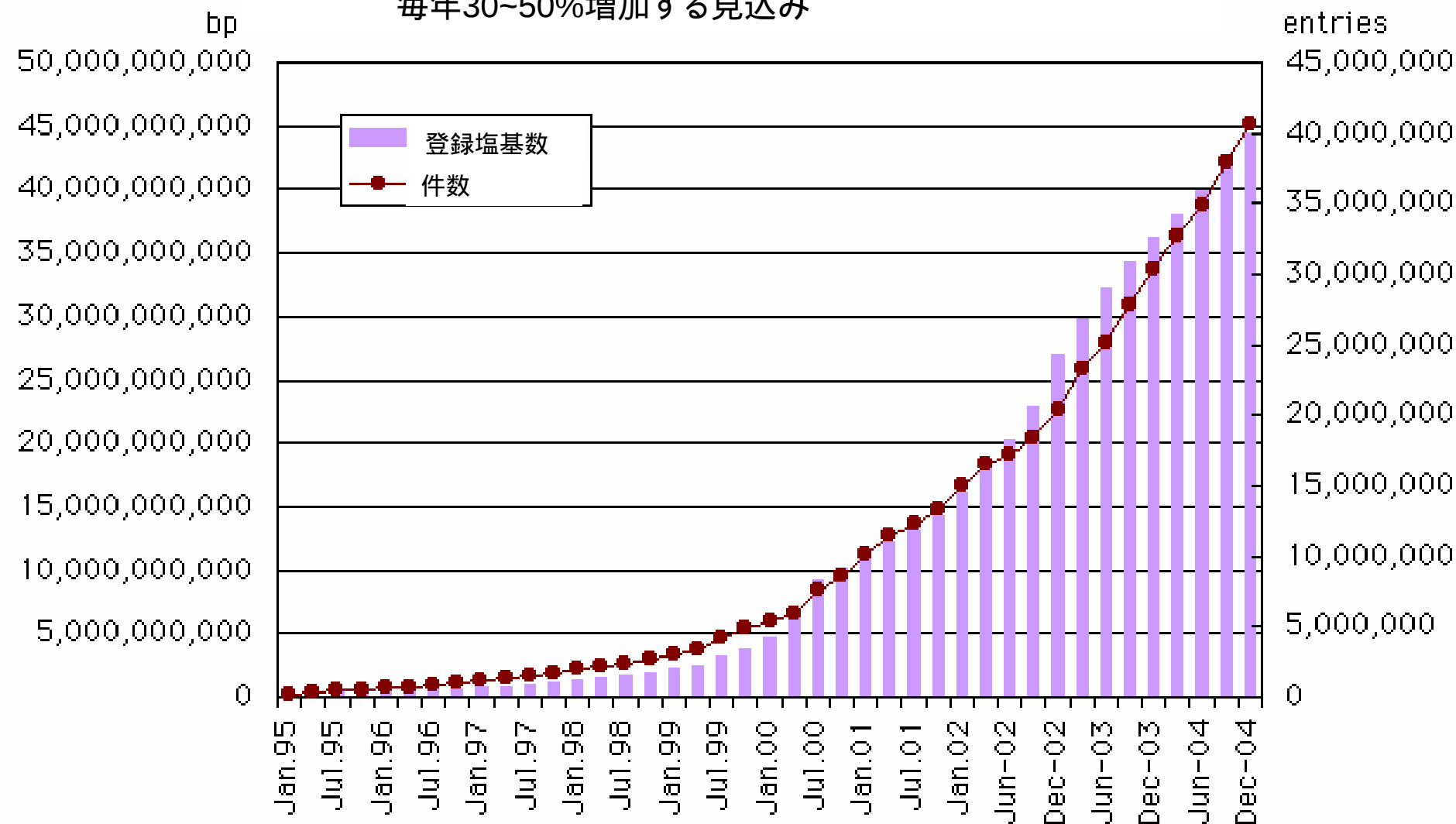
目次

- ◆ 生物情報資源の状況
 - 拡大
- ◆ 生物情報資源の信頼性
 - 大規模データの再検証：微生物ゲノムの例
- ◆ 生物情報資源の探索と統合利用
 - Webサービス

生物情報資源の状況

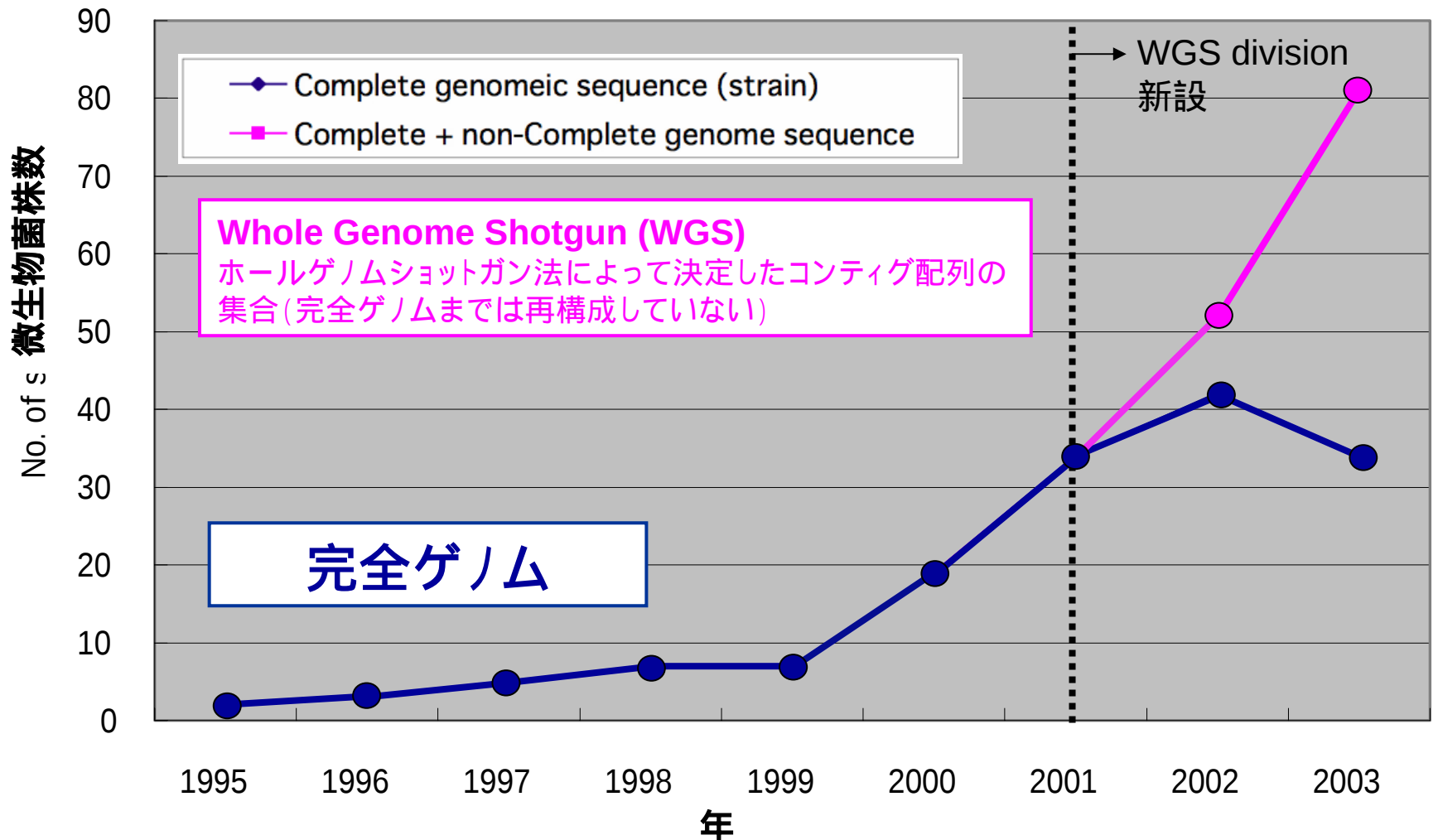
生物情報資源の膨張

- 国際塩基配列データベース (DDBJ/EMBL/GenBank)
2004年に4千万件400億塩基対を超え、これからも
毎年30~50%増加する見込み



生物情報資源の膨張

微生物ゲノム解析の成果の場合



生物情報資源の拡大

- **Nucleic Acids Research**

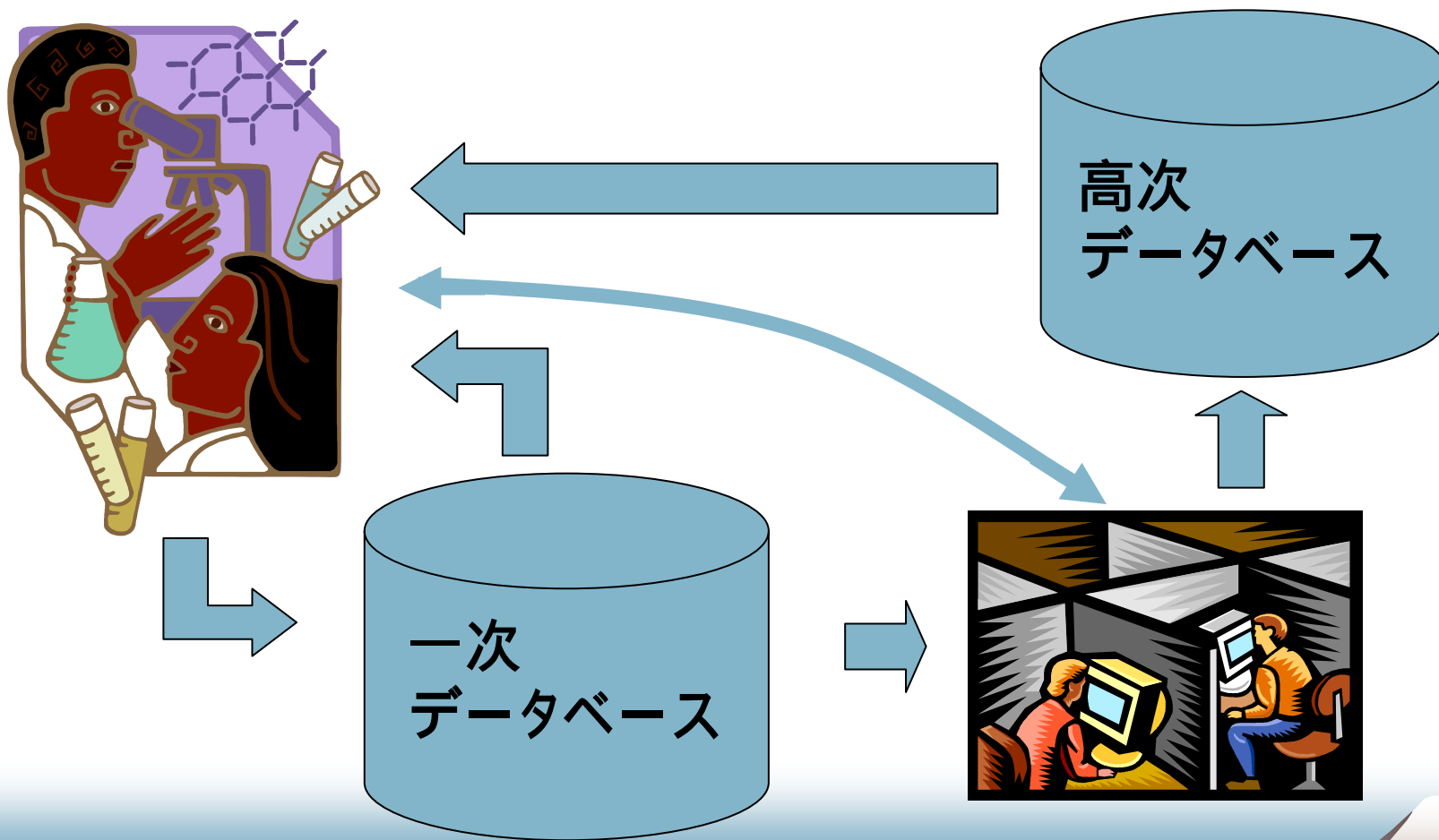
Database Categories List

収録件数年々増加
2005年: > 700件

- ▶ Major Sequence Repositories
- ▶ Comparative Genomics
- ▶ Gene Expression
- ▶ Gene Identification and Structure
- ▶ Genetic and Physical Maps
- ▶ Genomic Databases
- ▶ Intermolecular Interactions
- ▶ Metabolic Pathways and Cellular Regulation
- ▶ Mutation Databases
- ▶ Pathology
- ▶ Protein Databases
- ▶ Protein Sequence Motifs
- ▶ Proteome Resources
- ▶ RNA Sequences
- ▶ Retrieval Systems and Database Structure
- ▶ Structure
- ▶ Transgenics
- ▶ Varied Biomedical Content

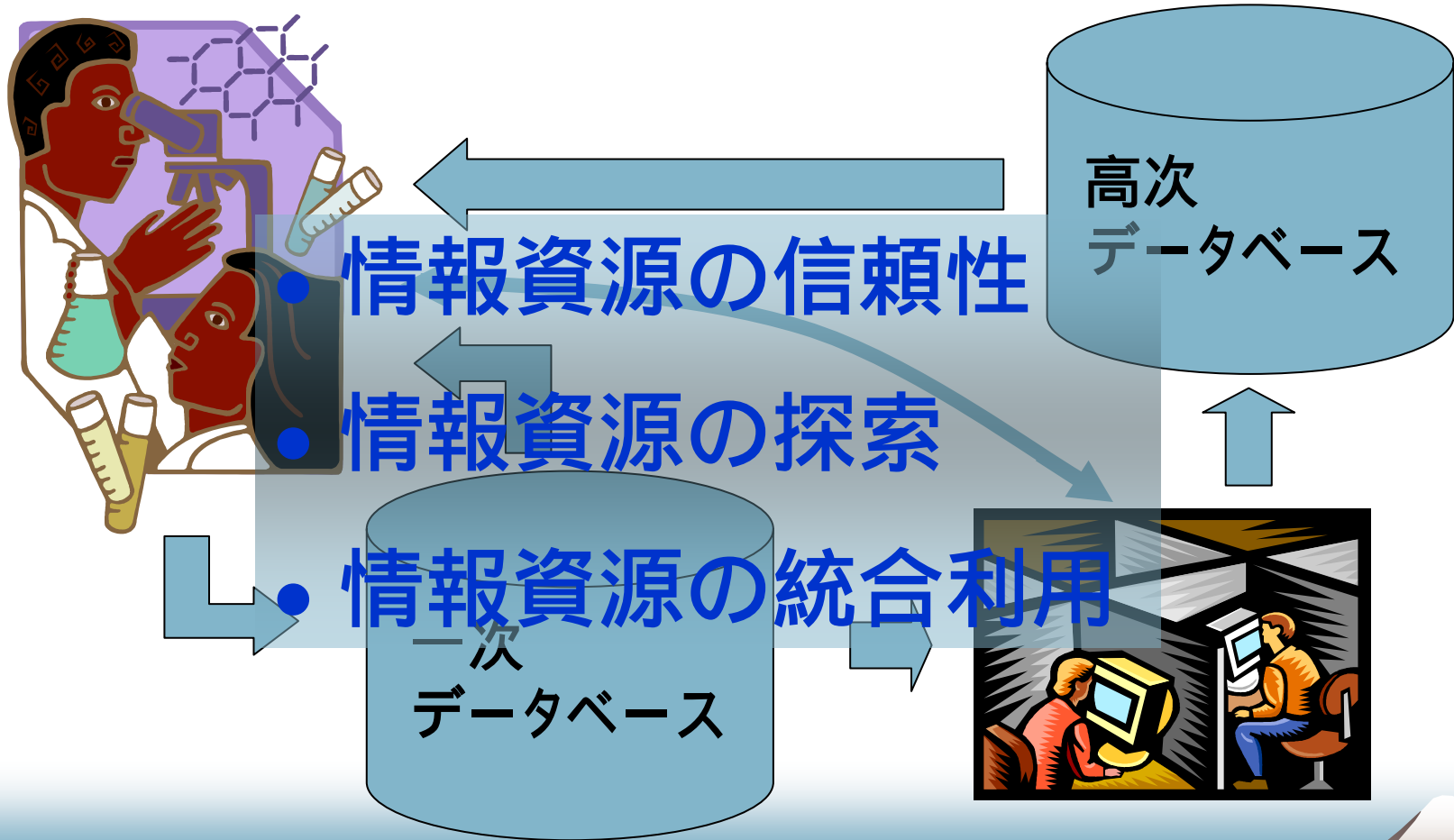
生物情報資源の増殖

データベース依存



生物情報資源の増殖

データベース依存



生物情報資源の信頼性

大規模データの再検証：微生物ゲノムの例

公開データの再評価

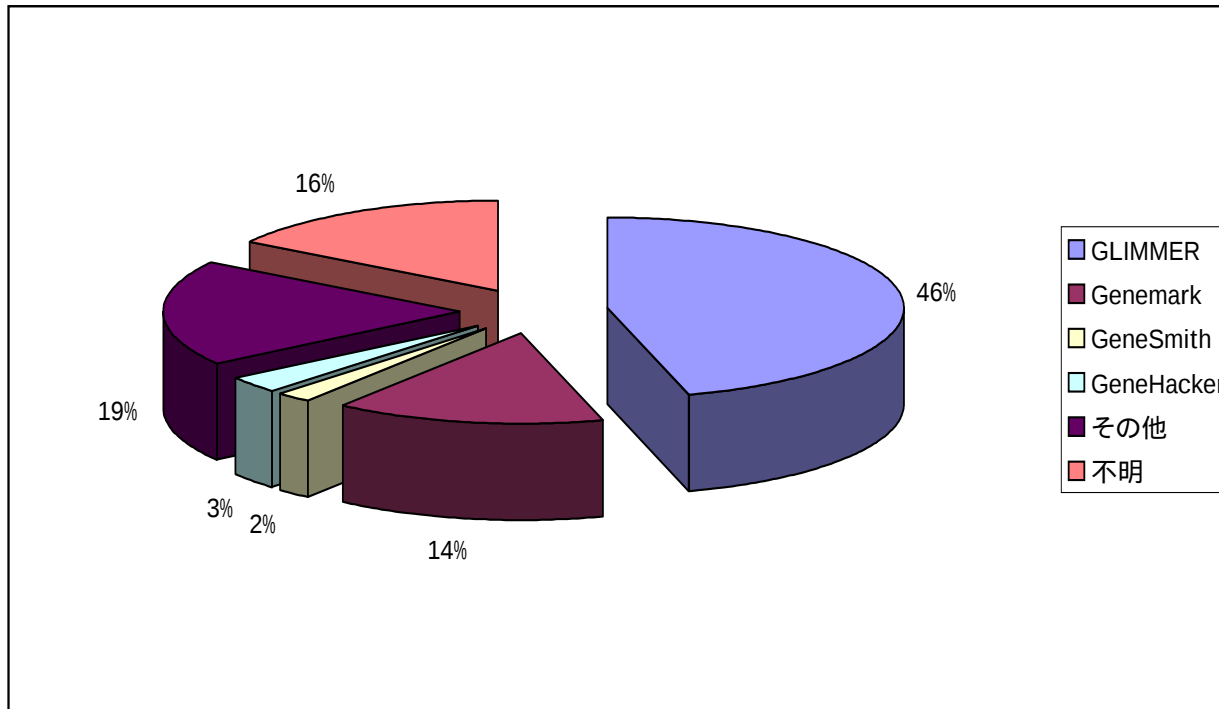
微生物ゲノムの遺伝子領域 (Open Reading Frames, ORFs)

データ解析プロトコールが多様

- ORF予測プログラムの多様性
 - Diversity of prediction programs
 - Diversity of parameters, e.g. the minimum length of ORFs
- 参照データベースの更新
 - Genes in DDBJ/EMBL/GenBank are continuously expanded and updated
 - Reference databases like InterPro are also continuously expanded and updated

ORF予測プログラムの状況

ORF Prediction programs used



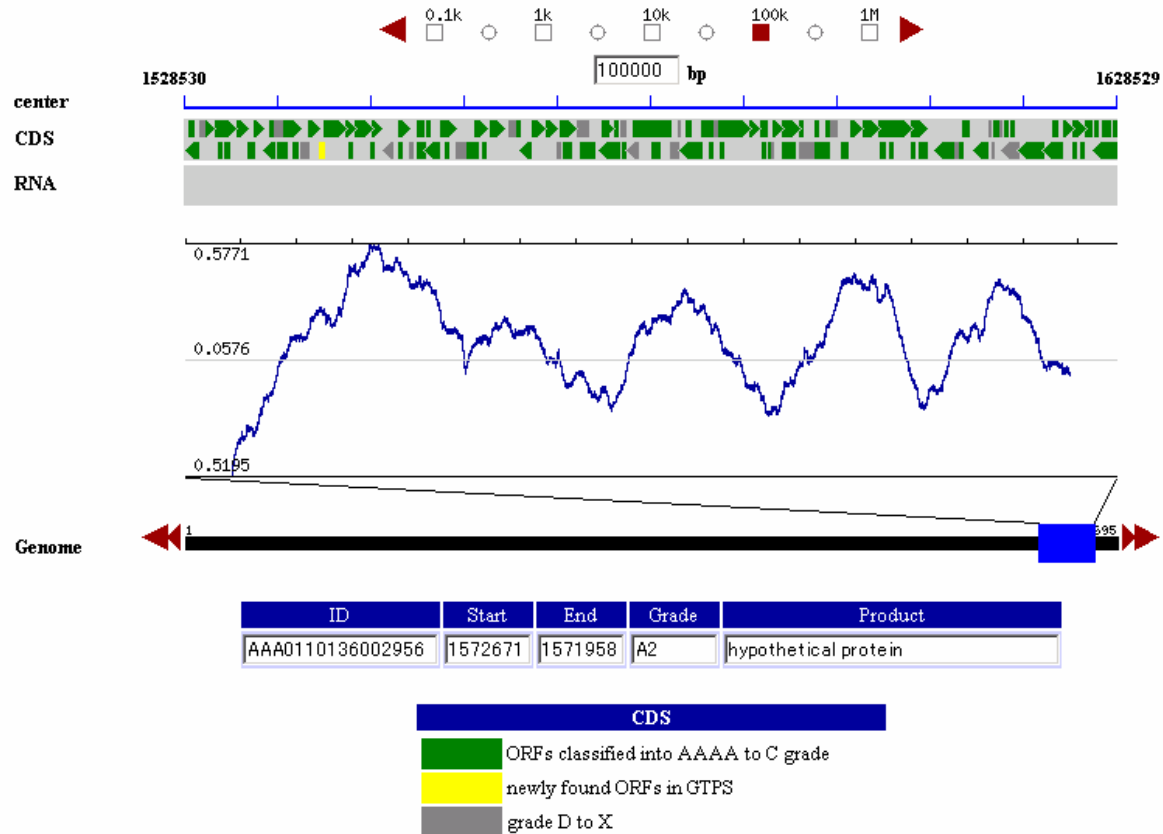
Threshold value to determine the ORF

>5aa	1
>20aa	1
>30aa	24
>33aa	1
>33.3aa(100bp)	3
>40aa	1
>50aa	8
>60aa	4
>66.6aa	1
>80aa	2
>100aa	6
>150aa	1
>200aa	1
>300aa	1
>400aa	1

アノテーションとは

Aeropyrum pernix K1

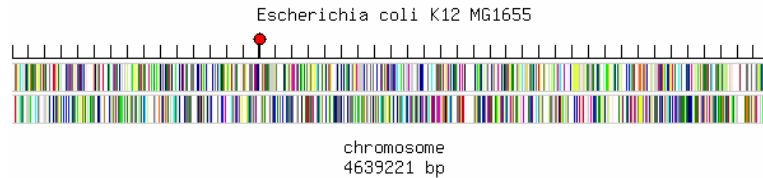
Feature View



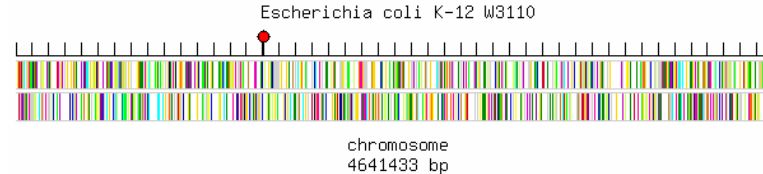
アノテーションの振れ

“virulence”によるキーワード検索

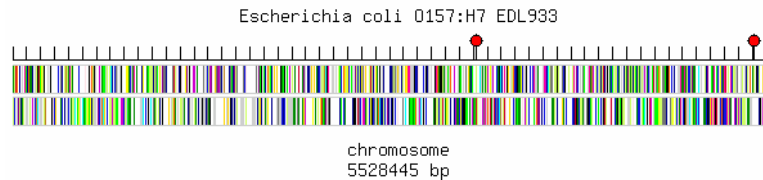
E. Coli MG1655



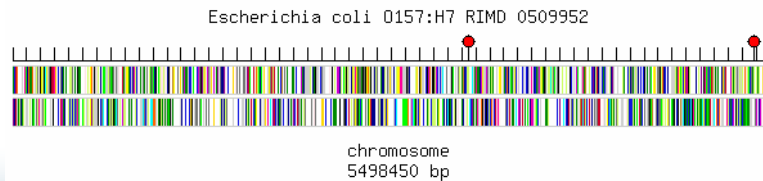
E. Coli W3110



O157 H7
EDL933



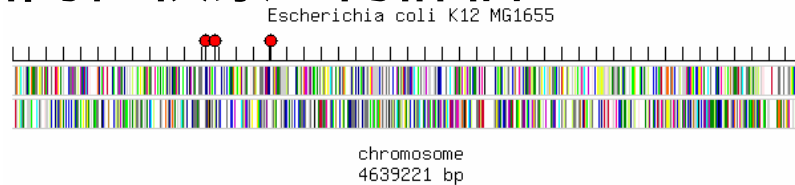
O157 H7 RIMD
509952



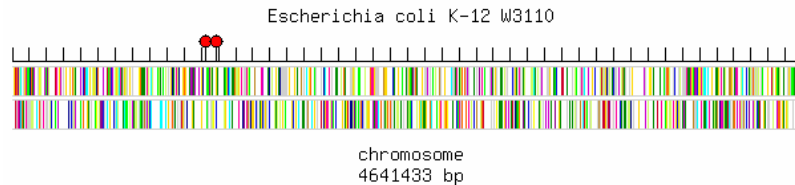
アノテーションの振れ

相同性検索で再評価

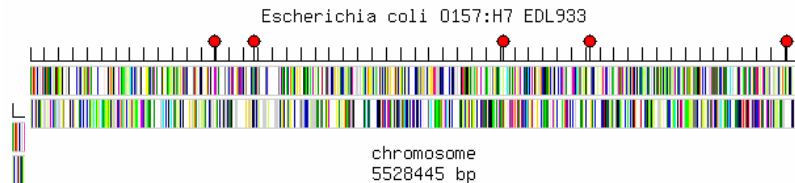
E. Coli MG1655



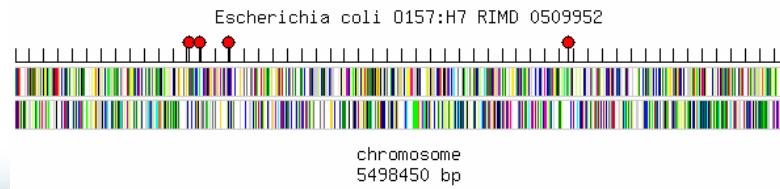
E. Coli W3110



O157 H7
EDL933



O157 H7 RIMD
509952



アノテーションの参照価値

putative protein

probable protein

predicted protein

uncharacterized protein

conserved protein

hypothetical protein

conservative hypothetical protein

conserved Hypothetical Protein

conserved hypothetical protein

ORF再評価のワークフローを確立

Prediction of CDS regions



Complete genome sequences registered
in the INSD (*Archaea* 16, *Bacteria* 107)

Validation and annotation of CDS



1. Homology search
2. Motif search
3. Check overlapped CDS
4. Check pseudogene

Ranking of CDS



Comparison CDS with ORFs of each genome in the INSD

ORFsのクラス分け

blastp results
coverage
protein name

InterProScan
motif names



AAAA, AAA, AA, A
BBBB, BBB, BB, B
C
D
E
X

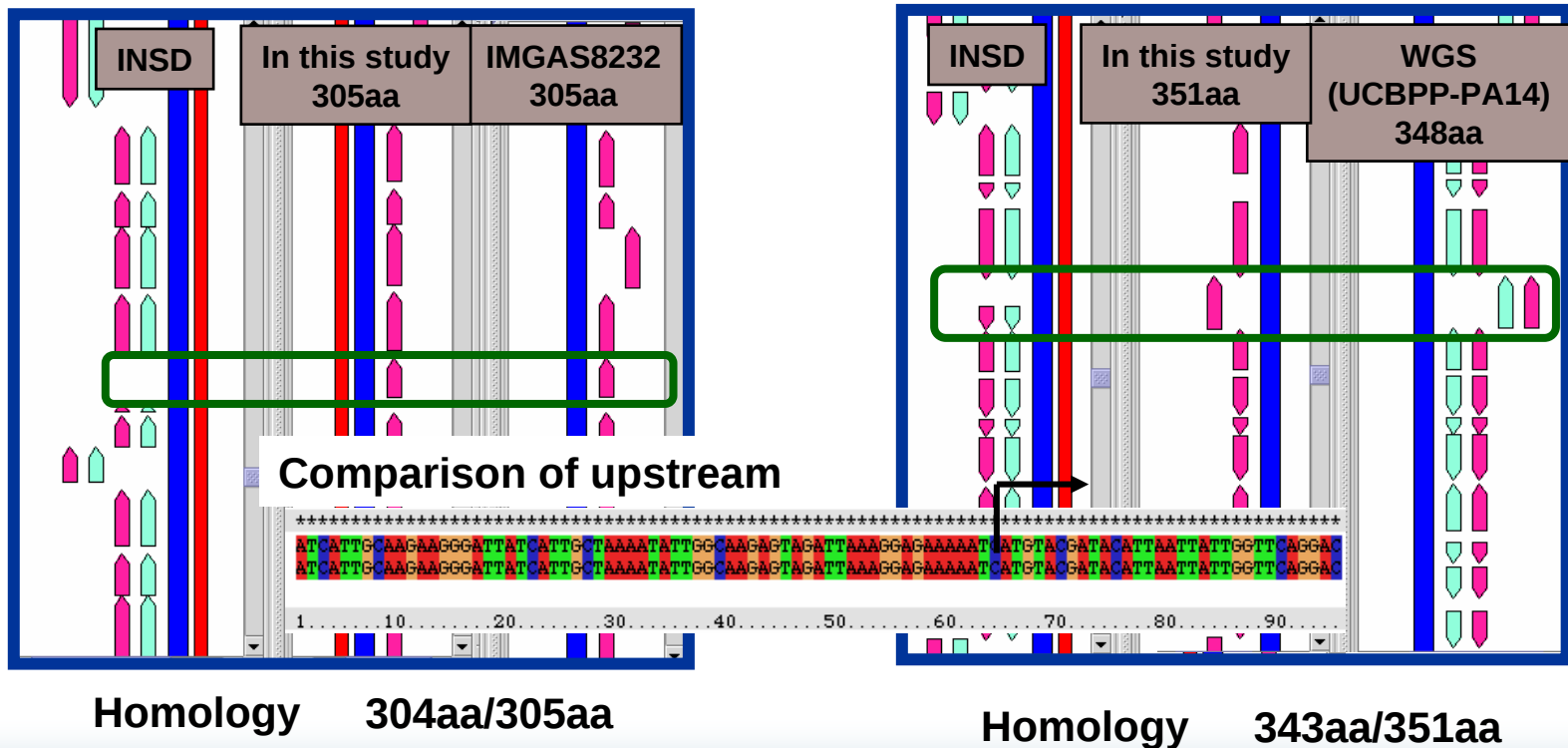
A-C: standard

D-X: non-standard

新規CDSs

Streptococcus pyogenes SF370
Complement (1375345..1376262)
Thioredoxin reductase
Grade: AAAA3

Pseudomonas aeruginosa PAO1
Complement (3607064..3608119)
Glycosyltransferase
Grade: AAAA3



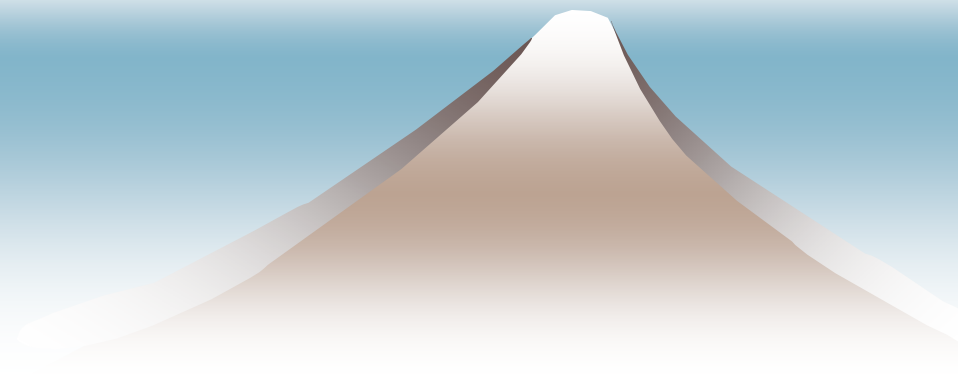
再評価ORFsに準拠した比較ゲノム

Clustering of ORFs by SODHO algorithm

to understand the structure and function of gene products

Horizontal Transfer of Genes (HTG)

to understand the evolution and diversity of microbes



SODHOアルゴリズム

1. All to all homology search by FASTA
2. ORFs of low homology to ORF X is discarded from the ORFs set for the ORF X
3. Pair-wise alignment of ORF X vs all the member ORFs in the set for ORF X
4. Evaluation of the alignment based on the score and matching length
5. ORFs of *high similarity* are preserved for the further analysis
6. Clusters are defined

SODHOの計算結果例

InterProによる評価

Name of ORFs	IPR002528	IPR001064	IPR00687
PFDSM[1850] Pfur_DSM3638:.faa_C10	+	+	
AAVF5[101] Aaeo_VF5:.faa_C10	+		+
PAORS[366] Paby_ORsay:.faa_C10	+	+	
PHOT3[1861] Phor_OT3:.faa_C10	+	+	
TTMB4[1686] Tten_MB4T:.faa_C10	+		
TTMB4[234] Tten_MB4T:.faa_C10	+		

SODHO計算時間

Process Name	Principal Step	Computaiton by 1 CPU (2.0GHz)	Computation by OBIEnv/GRID
Homology search	FASTA (Algorism of Lipman&Pearson)	50 days	
Homology classification	*Gotoh Method (Advansed method of Needleman&Wuncsh) for pair- wise alignment *Homology classification	1 week	
Multiple Alignment (Not done this time)	HOMTRAL Method	~ 6 years	

* Time at 1.3GHz machine

HTG解析

A query genome (a complete genome sequence to be analyzed)

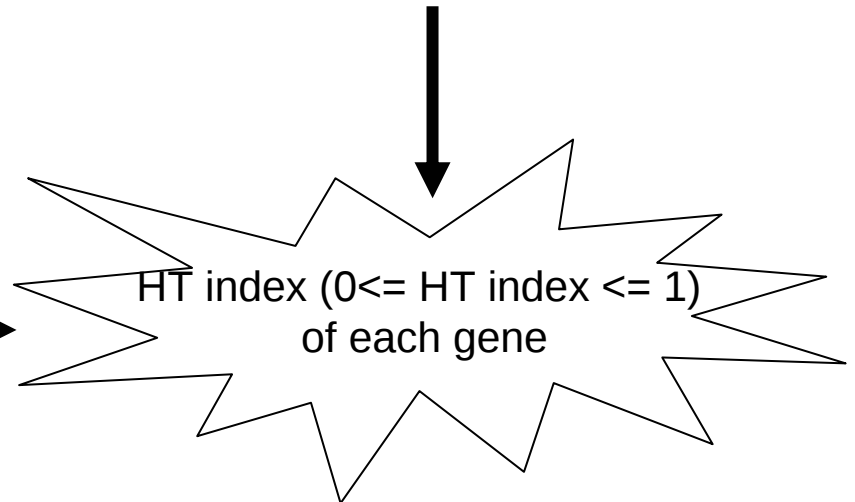
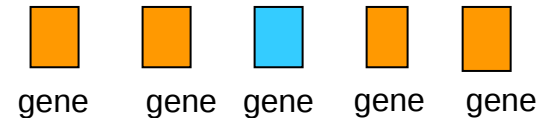


Construction of a model that represents nucleotide preferences in the query genome

The model

5bp-tuple frequency
ATATA --- 0.02
GGTCT --- 0.01
AGGTG --- 0.03
⋮

Gene sequences encoded in the query genome



orange high HT (close to 1): intrinsic
blue low HT (close to 0): extrinsic

HTGデータベース

Region Viewer

Detail Viewer for Searching Probability of HTG and Probability of Donor of the gene and near region.

ORF View

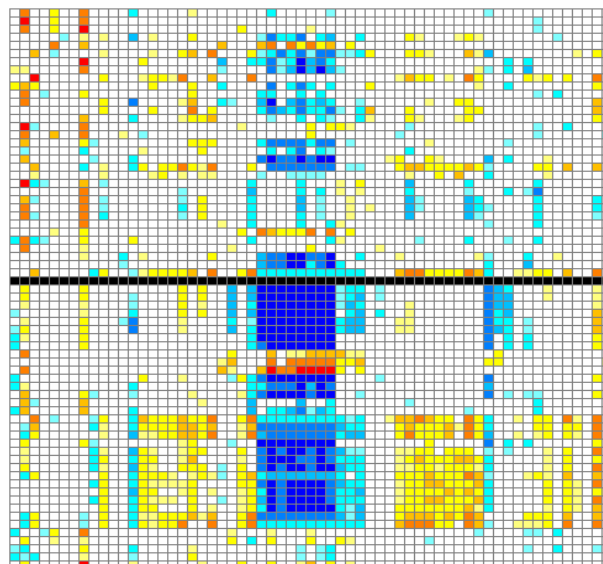
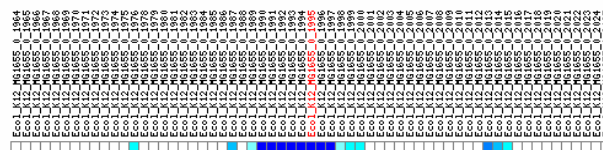
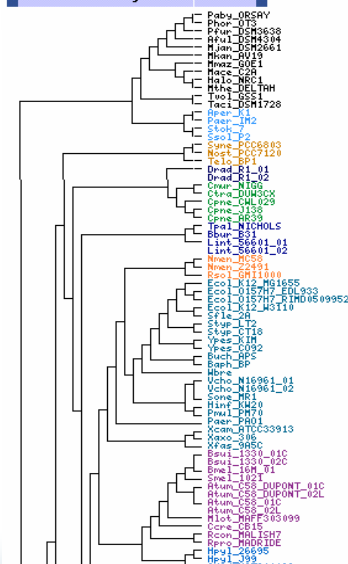
Probability of HTG

Base Information

Target Gene :
EcoI_K12_MG1655_0_1995

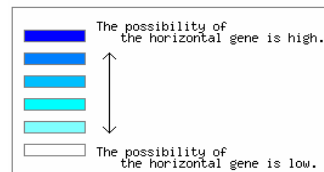
Species Name :
Ecol_K12_MG1655

Probability of Donor



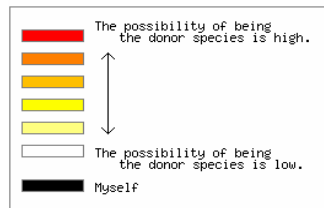
Region Length : 10[default] 30 100

Present Status: 30



Change View : default strict advanced

Preset Status: advanced



HTG計算時間

	by 1 CPU of 1GHz	by GRID
Detection of HTG	~ 20 days	
Detection of donors of HTG	4~5 years	

*

5 seconrds/ORF

 ×

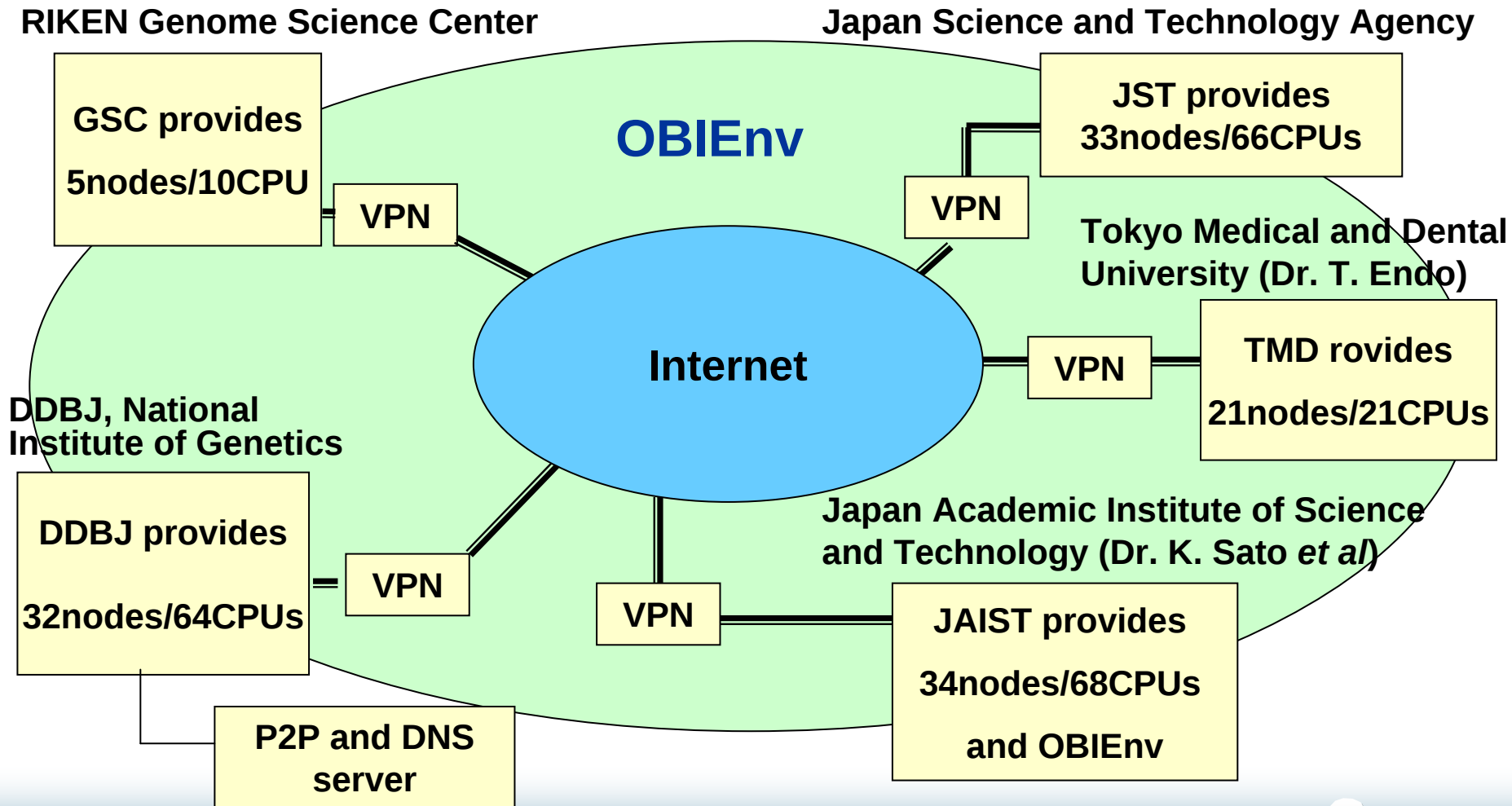
300K ORFs

 ×

120 micorbes

 ~4.5 years

GRID環境



稼働状況のモニター

OBIEnv machines for CGM

[illegible]

on : Fri Dec 05 09:12:29 JST 2003

summary

host list

■ <-- idle -- ■ -- busy -- ■ -- over load --> ■
over load : cpu < job

Green boxes stand for CPUs that are used

SODHO計算時間

Process Name	Principal Step	Computaiton by 1 CPU (2.0GHz)	Computation by OBIEnv/GRID
Homology search	FASTA (Algorism of Lipman&Pearson)	50 days	→ 1day
Homology classification	*Gotoh Method (Advansed method of Needleman&Wuncsh) for pair- wise alignment *Homology classification	1 week	→ 1hour
Multiple alignment (not done this time)	HOMTRAL Method	~ 6 years	30 days

HTG計算時間

	by 1 CPU of 1GHz	by GRID
Detection of HTG	~ 20 days →	~ 1 day
Detection of donors of HTG	4~5 years →	~ 20 days

*

5 seconrds/ORF

 X

300K ORFs

 X

120 micorbes

 ~4.5 years

プロジェクト期間

■ The 1st round: ORFs registered in GIB

Analysis	Microbial genomes	Duration
Clustering by SODHO [2]	119	6/18 ~ 6/19
HGT	112	6/20 ~ 7/20 *

■ The 2nd round: ORFs evaluated in GTMS

Analysis	Number of microbial genomes	Duration
Clustering by SODHO	124	12/4 ~ 12/5
HGT	124	11/5 ~ 12/11 *

* include interruption by maintenance of computers and network

GRIDによる解析のまとめ

◆ SODHO

- 999 jobs are done in ~ 1 day
- 354,605 ORFs are classified into 63,149 clusters
- All the clusters are referred to InterPro

◆ HTG

- 17,689 jobs were done in ~ 18days
- 45,136 HTGs are predicted from ~300K ORFs in 124 genomes (112 microbes)

◆ Update

- InterPro for 160 genomes (~1.2M ORFs):
→ 4,000 hours by a 2.2GHz PC

Behind the scenes

pre-processing of data

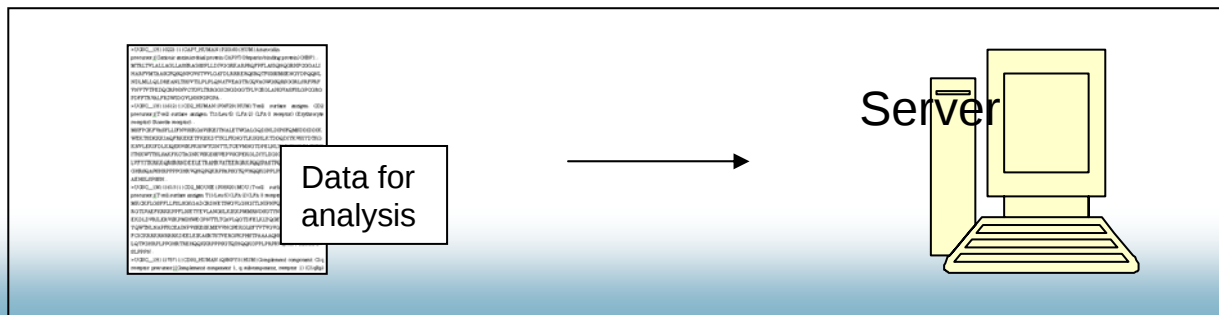


GRID解析のための前処理

- Data set for the analysis
 - The lead time to start computation by a stand-alone machine is relatively short
 - To do list to start GRID computing
 - To divide data and create a list for GRID. (3h)
 - To verify of the division and the list (1h)
 - To compress the data for FTP.(5h)
 - To distribute the data to OBIGRID sites (1h ~ 2 days)
 - To manage data in OBIGRID (e.g. preparation of RDB) (half a day)
 - To arrange data in OBIGRID (2-3hour)
 - To arrange data at each site (2 days)
 - To check the arrangement (1h)
 - To start JOBS

Total 5~6 days

Computing by a conventional server



生物情報資源の探索と統合利用

Webサービス

Behind the scenes

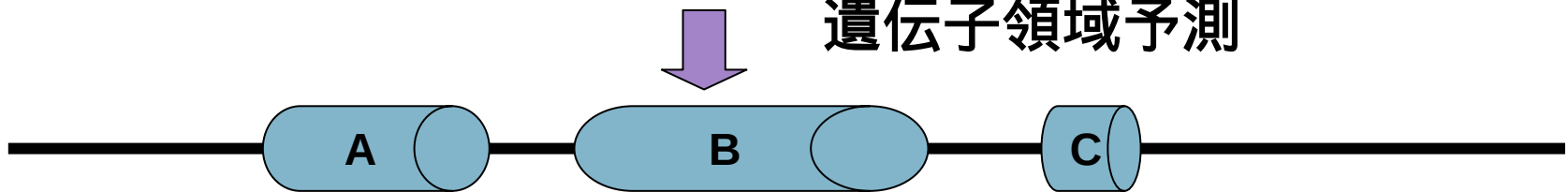
Workflow composed of Web services



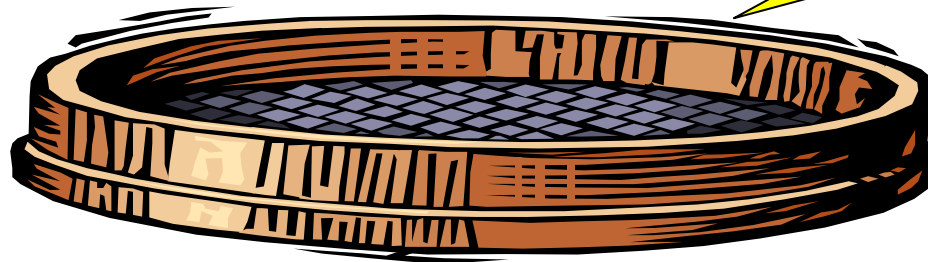
WebサービスによってORFs再評価 ワークフローを効率的開発

ゲノムDNA配列

遺伝子領域予測



相同性検索、
モチーフ探索による
ふるい分け



DNA

ORF



妥当性の高い遺伝子領域
比較ゲノム解析へ

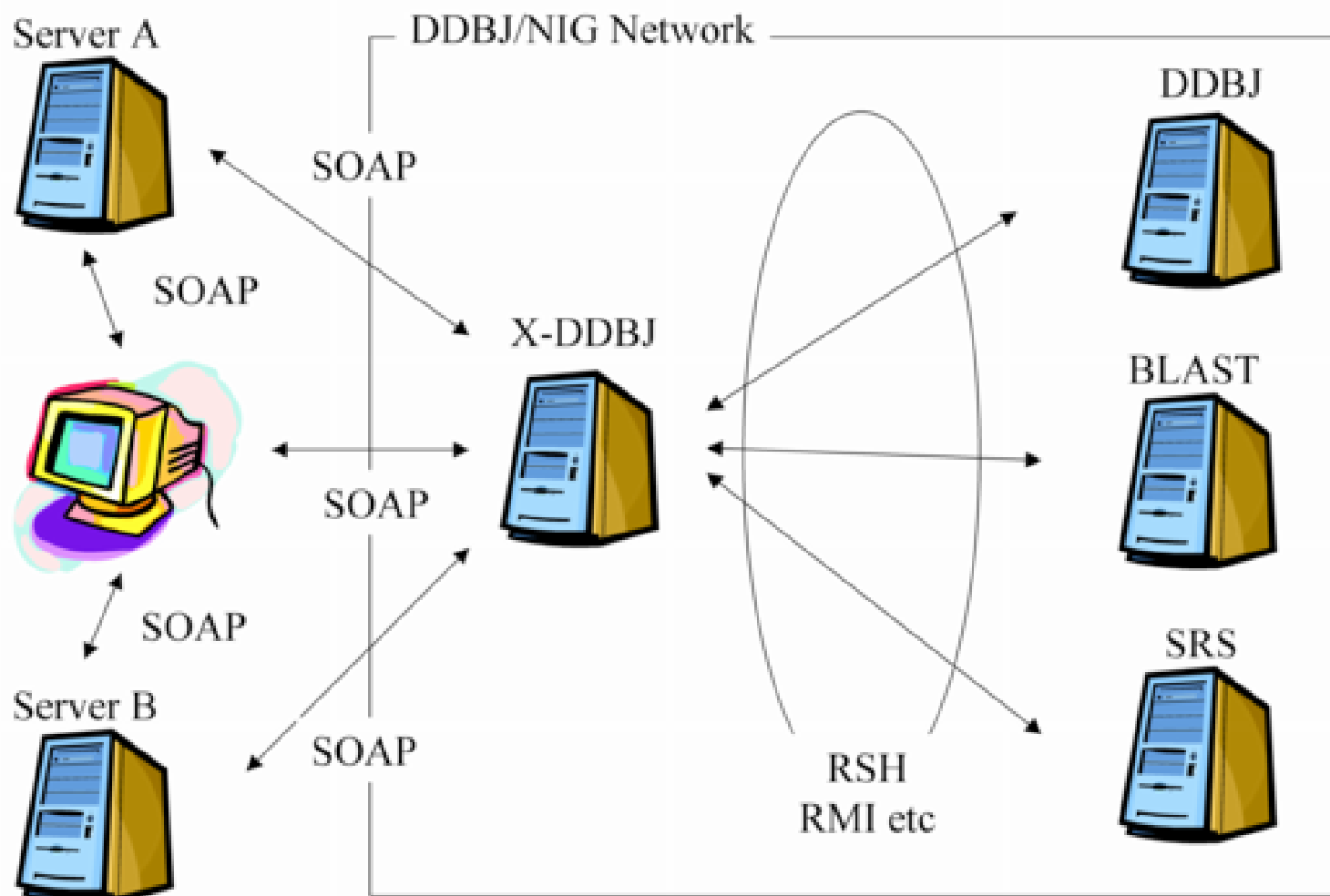
DDBJのWebサービス

Name	URL	Document	Registrant	Methods
BLAST Demo	http://xml.nig.ac.jp/wsdl/BlastDemo.wsdl	document javadoc	XML Central of DDBJ	List to execute
Blast	http://xml.nig.ac.jp/wsdl/Blast.wsdl	document javadoc	XML Central of DDBJ	List to execute
ClustalW	http://xml.nig.ac.jp/wsdl/ClustalW.wsdl	document javadoc	XML Central of DDBJ	List to execute
DDBJ	http://xml.nig.ac.jp/wsdl/DDBJ.wsdl	document javadoc	XML Central of DDBJ	List to execute
ExClustalW	http://xml.nig.ac.jp/wsdl/ExClustalW.wsdl	document javadoc	XML Central of DDBJ	List to execute
Fasta	http://xml.nig.ac.jp/wsdl/Fasta.wsdl	document javadoc	XML Central of DDBJ	List to execute
GetEntry	http://xml.nig.ac.jp/wsdl/GetEntry.wsdl	document javadoc	XML Central of DDBJ	List to execute
Gib	http://xml.nig.ac.jp/wsdl/Gib.wsdl	document javadoc	XML Central of DDBJ	List to execute
Gtop	http://xml.nig.ac.jp/wsdl/Gtop.wsdl	document javadoc	XML Central of DDBJ	List to execute
PML	http://xml.nig.ac.jp/wsdl/PML.wsdl	document javadoc	XML Central of DDBJ	List to execute
RefSeq	http://xml.nig.ac.jp/wsdl/RefSeq.wsdl	document javadoc	XML Central of DDBJ	List to execute
SRS	http://xml.nig.ac.jp/wsdl/SRS.wsdl	document javadoc	XML Central of DDBJ	List to execute
TxSearch	http://xml.nig.ac.jp/wsdl/TxSearch.wsdl	document javadoc	XML Central of DDBJ	List to execute

13種類のWebサービスに
それぞれメソッドが
用意されている

Method
<u>searchParam(program, database, query, param)</u>
<u>searchSimple(program, database, query)</u>
<u>extractPosition(result)</u>

SOAPによるWebサービス

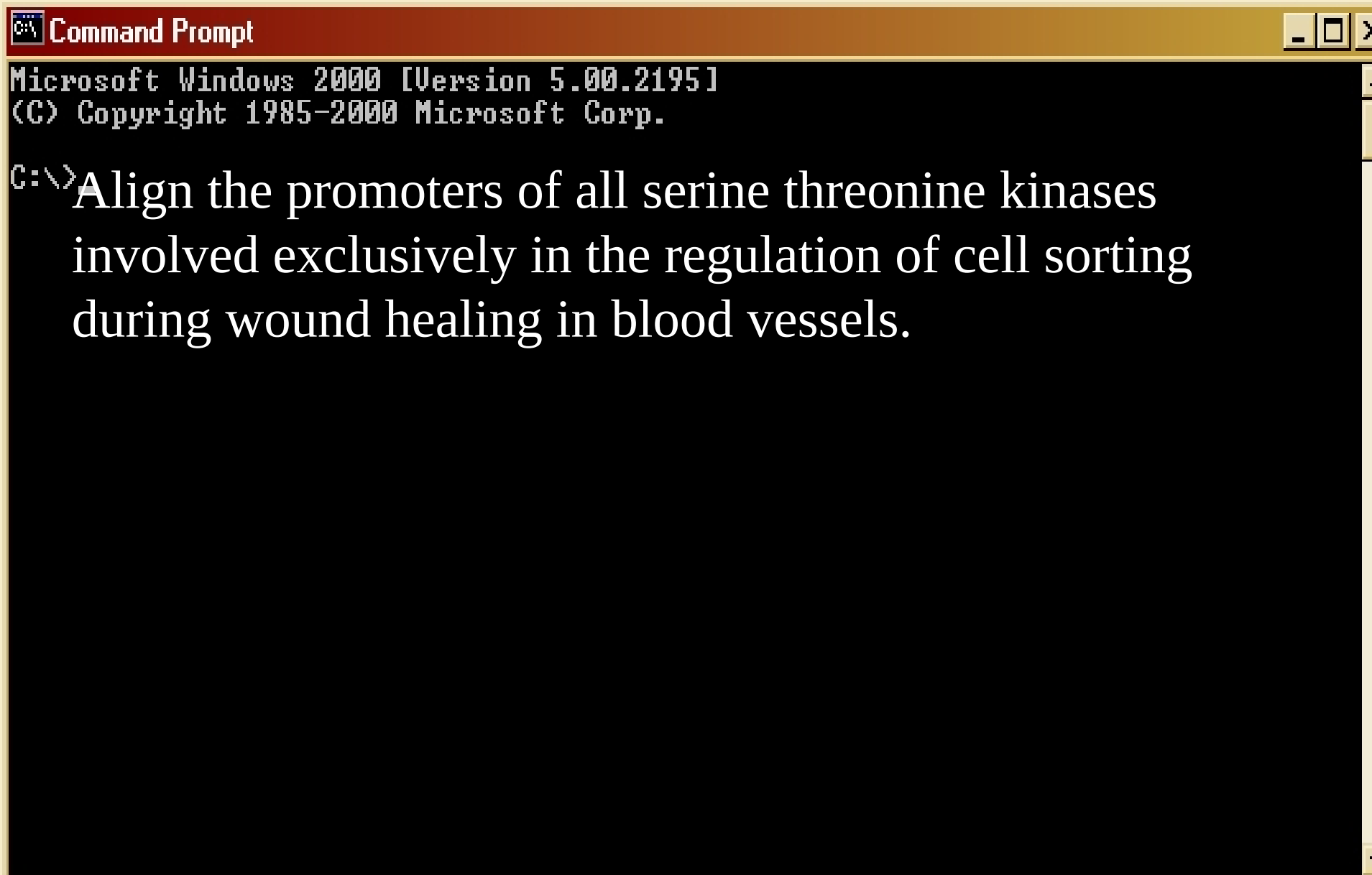


生物情報資源利用における 問題点の再確認

- 情報は分散している
さまざまなデータベースが世界中で構築されている
すべての関連データを保持することは不可能
今後は実験技術の向上によりさらに増える？
- 統合的に扱うインタフェースが無い
WEBサイトでデータベースを公開
人が見るのには良い、機械的に扱うには不便

→ 相互運用性の向上が不可欠
WEBサービス技術により解決

問題設定 (BioMOBYより引用)



```
Command Prompt
Microsoft Windows 2000 [Version 5.00.2195]
(C) Copyright 1985-2000 Microsoft Corp.

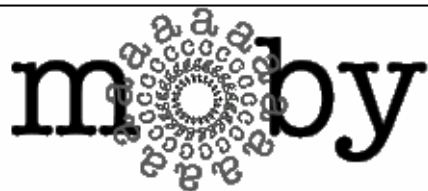
C:\>Align the promoters of all serine threonine kinases
involved exclusively in the regulation of cell sorting
during wound healing in blood vessels.
```

問題設定 (BioMOBYより引用)

```
Command Prompt
Microsoft Windows 2000 [Version 5.00.2195]
(C) Copyright 1985-2000 Microsoft Corp.

C:\> Align the promoters of all serine threonine kinases
C:\> involved exclusively in the regulation of cell sorting
C:\> during wound healing in blood vessels.
C:\>
C:\>Bad Command or File Name

Retrieve and align 2000nt 5' from every serine/threonine kinase
in Mus musculus expressed exclusively in the tunica [I | M |A]
whose expression increases 5X or more within 5 hours of
wounding but is not activated during the normal development of
blood vessels, and is <40% homologous in the active site to
kinases known to be involved in cell-cycle regulation in any
other species.
```



BioMOBY.org

General Documentation

(contact [Mark Wilkinson](#))

[FAQ](#)

[Published Manuscripts & Press](#)

[BioMOBY Project Analysis \(Use Cases, Scenarios, etc.\)](#)

[Getting the Code](#)

[Mark's Latest Presentation](#)
(Max Planck Institute, Cologne, Germany, January, 2005)

[Mailing Lists](#)

[Acknowledgements](#)

[Acknowledging the BioMOBY Project!](#)

[BUGS!](#)

MOBY Services Branch

(contact [Mark Wilkinson](#))

[NEWS!](#)

[Tutorials and How-To's](#)

[Browse Data using MOBY Services](#)

[Project Overview](#)

[MOBY-S 0.8 API **NEW!**](#)

[MOBY-S Java Library Docs](#)

[MOBY-S Perl Library Docs](#)

[Installing a Local Registry](#)

[Cool Tools! \(new Jan. 11, '05\)](#)

[MOBY Central Connection Details](#)

Semantic MOBY Branch

(contact [Lincoln Stein](#), [Damian Gessler](#), or [Gary Schultz](#))

Featured Link:

"W3C Semantic Web Activity"

Introduction to the W3C's work in the area of the Semantic Web.

[[Previous Featured Links](#)]

[S-MOBY Design Overview](#)

[S-MOBY Design Document](#) 

[S-MOBY Tech Assessment](#)

[S-MOBY Requirements Analysis](#) 

[BioMOBY Use Cases](#)

BioOBY

◆ 複数のサービスを統合して利用する環境を提供

- ・ 各データベースのフォーマットやスキーマに依存しない
- ・ WEBサービスを適用し、サービスの動的な発見、ワークフロー作成を可能に

◆ データベースのデータベース (MobyCentral)

- ・ だれがどこで提供しているのか
- ・ どのようなデータなのか (配列、文献、形質、変異)
- ・ どのように利用できるか (サービス名、入力、出力)

◆ MobyCentralのAPIを提供

- ・ サービスの登録、削除
- ・ 入出力データの登録、削除
- ・ サービス名や入出力データを条件にした検索
- ・ WSDLの取得
- ・ サービスの呼び出し

Overview

User's Guide

Architecture

Download

ToDo

API

Perl Client

Soaplab SOAP-based Analysis Web Service

Soaplab is a set of Web Services providing a programatic access to some applications on remote applications, especially in the scientific environment where Soaplab was born, usually analyze data. **Analysis (Web) Service.**

Soaplab was developed in the European Bioinformatics Institute (EBI), within the eScience initiative and the myGrid project.

Soaplab is both a specification for an Analysis Service (based on other approved specifications, e.g. SOAP) and an implementation. It is freely available for downloading - but bear in mind that the installation of this service requires some analyses - they are not part of the Soaplab. The EBI has Soaplab service running on top of several servers coming from EMBOSS, an independent package of high-quality FREE Open Source software for sequence analysis. It is an experimental service which may not have 24/7 availability.

Recently, the same API was used to access other resources than command-line tools. A sub-project was created to access ordinary web pages and see them as web services.

The implementation is written in Java using Apache Axis framework.

Soaplab What's New

18th March 2004

Code was revamped to accomodate a new sub-project Gowlab. But the API remained the same. Also that Soaplab uses now Ant for building and deploying services.

Biological SOAP servers and web services provided by the public sequence data bank

H. Sugawara^{1,2,*} and S. Miyazaki¹

¹Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan and ²SOKENDAI, Department of Genetics, Hayama, Kanagawa 240-0193, Japan

Received February 10, 2003; Revised March 26, 2003; Accepted April 3, 2003

ABSTRACT

A number of biological data resources (i.e. databases and data analytical tools) are searchable and usable on-line thanks to the internet and the World Wide Web (WWW) servers. The output from the web server is easy for us to browse. However, it is laborious and sometimes impossible for us to write a computer program that finds a useful data resource, sends a proper query and processes the output. It is a serious obstacle to the integration of distributed heterogeneous data resources. To solve the issue, we have implemented a SOAP (Simple Object Access Protocol) server and web services that provide a *program-friendly* interface. The web services are accessible at <http://www.xml.nig.ac.jp/>.

same data structure as used in the exchange. The most popular INSD data format visible to anonymous users is the so-called flat file format (FF format). Therefore, it is possible to write a script to parse the FF format. It is ironical that the quite similar script for a slightly different aim has been repeatedly written by a number of groups in the world. 'There is massive duplication of efforts' (1) in parsing many other web pages too.

It is also to be noted that the INSD has modified and expanded the features and qualifiers in accordance with the progress of biotechnology and life sciences (i.e. the FF format has changed). Thus a script that parses FF format has to be updated by referring to the document provided by the INSD (<http://www.ddbj.nig.ac.jp/sub/ref1-e.html>) sometimes. Nevertheless, the INSD may claim that it provides comparatively stable data structure and format. There are many other biological data sources that are valuable but do not provide a sufficient document for users and change the access method



NCBI News

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Department of Health and Human Services

Summer/Fall
2004

Past Issues

In this issue

Entrez Programming Utilities (E-Utils)

[PubChem](#)

[GenePlot](#)

[New NLM Catalog in Entrez](#)

Entrez Programming Utilities

In dealing with specialized datasets, researchers are often restricted to one of two unattractive choices: either to download an ftp archive containing far more than the data of interest, followed by a round of local parsing, or to access the data interactively, even though the volume of data may render this method cumbersome. To help with the latter method, NCBI provides a suite of programs called the Entrez Programming Utilities (E-Utilities) that allow automated access to the Entrez databases.

What Are the E-Utilities?

The E-Utilities are a set of seven server-side programs that provide a stable interface to the search, retrieval, and linking functions of the Entrez system, using a fixed URL syntax. The output provided by the E-Utilities is in XML format, with the notable exception of the EFetch utility, which returns data in a variety of formats. The E-Utilities are designed to be called from within a computer program that can process their output. Calling an E-Utility from any of the common programming languages—including Perl, Python, and Java—is a simple matter of posting a URL.

DDBJ Web servicesサイト



XML Central of DDBJ

"XML Central of DDBJ" has been partly supported by [BIRD](#) of Corporation (JST)

[Japanese](#)

What's New

DDBJ-XML **XML**

DDBJ-XML is a new output format of DDBJ entries. It is readable

Web services **SOAP**

This is the first public SOAP service for biology in Japan. The project aims at the standardization of bioinformatics services interoperability. This page also provides you a Web interface of the

Registration/Publication of your Web services

You are courteously invited to register your Web service(s) in the to the public.

Sample program to bind a web service to your program

You can use a Web service by specifying the URL address, method and parameter(s) in your program such as Java or perl.

Demo system using SOAP technology

You can get the seamless access to the retrieval of DDBJ entry and BLAST execution. Enjoy SOAP technology.

Workflow example using web services

A WORK FLOW composed of the Web services

SOAP tutorial

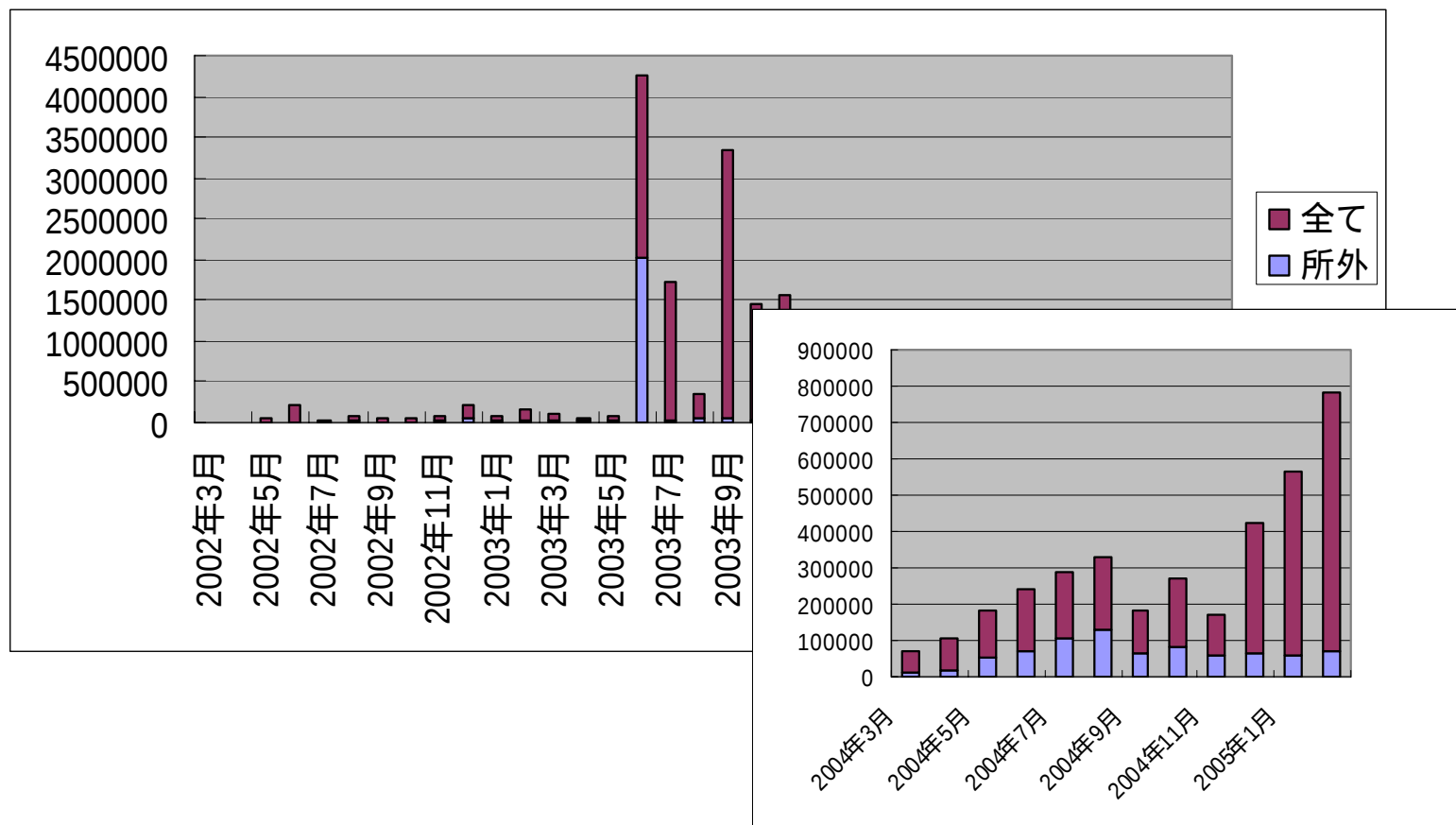
It's easy to access the SOAP services.
This is the first step to try web services.

Links

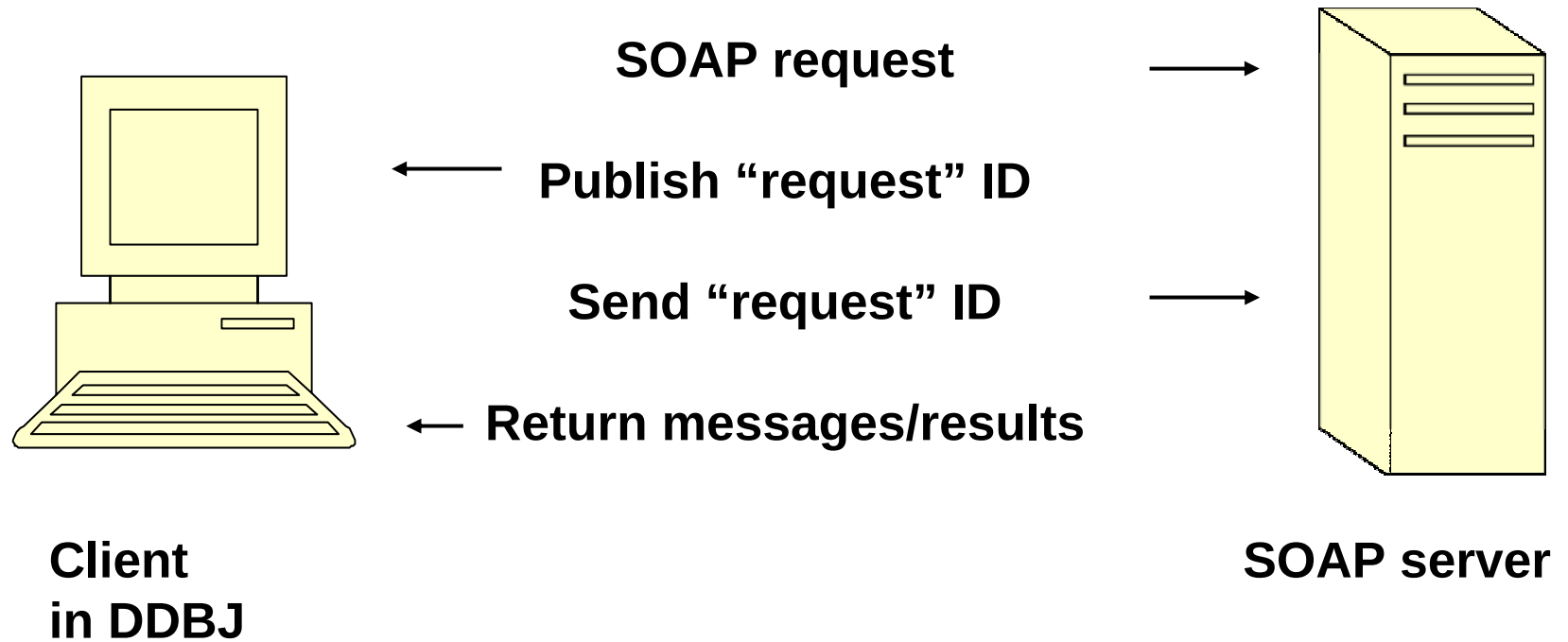
Gateway to biological XMLs and retrieval sites for XML documents

<http://www.xml.nig.ac.jp/>

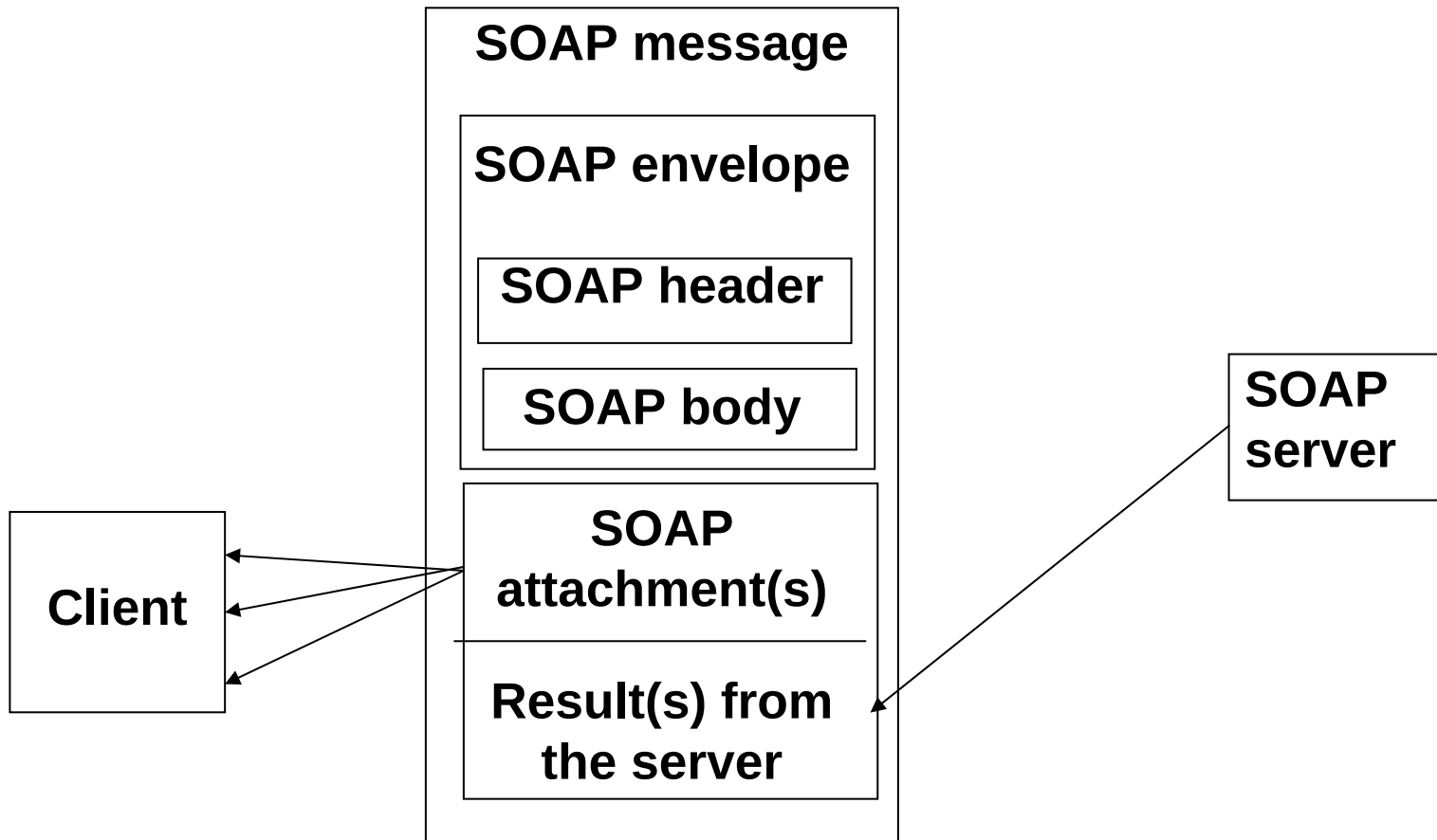
Web servicesの利用状況



長時間ジョブ対応：非同期通信



大量データ転送対応



Webサービスの日本語 検索インタフェース

Jabion-日本語バイオポータルサイト - Microsoft Internet Explorer

ファイル(E) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

Jabion このサイトは試作中のため一部ご利用にならないメニューがあります。
サイトに関するご意見、お問い合わせは biportal@nii.ac.jp まで。
<プロジェクトの詳細はこちら(pdfファイル)>

— 日本語で見る、入門書であり、専門書である。 —

Jabio 日本語バイオポータル

用語辞書 コラム一覧 遺伝子百科 文献検索 研究プロジェクト **道具箱** お役立ちリンク トップ

DDBJが提供するWebサービス

DDBJが提供するWebサービスから必要な機能をキーワードで検索することができます。
たとえば、「相同性」で検索すると相同性検索プログラムBlastおよびFastaのさまざまな
使い方をメソッドとして自作のプログラムから呼び出す方法へと辿ることができます。

検索
遺伝子

検索例: [遺伝子](#) [フィーチャー](#) [系統樹](#) [相同性検索](#) [塩基配列](#) [Feature](#)

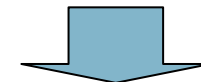
説明表示: ☒ 日本語 ☐ English

全メソッド表示 [日本語](#) [English](#)

[>>各種Webサービスとメソッドのリストへ](#)
[>>To the complete list of Web services and methods](#)

インターネット

国立遺伝学研究所が提供するWebサービスを、「遺伝子」、「系統樹」、「塩基配列」、「Feature」などの日本語及び英語のキーワードで検索することが可能。



サービス・オントロジーに準拠した記述へと展開

Jabion-日本語バイオポータルサイト - Microsoft Internet Explorer

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)



このサイトは試作中のため一部ご利用にならないメニューがあります。
 サイトに関する、ご意見、お問い合わせは bioportal@nii.ac.jp まで。
 <プロジェクトの[詳細はこちら](#)(pdfファイル)>

—最新バイオテクノロジーの世界を日本語でご自由にお楽しみください—

日本語バイオポータルサイト

用語辞書 コラム一覧 遺伝子百科 文献検索 研究プロジェクト **道具箱** お役立ちリンク トップページ

DDBJが提供するWebサービス

Search word : 塩基配列

name	method	説明
DDBJ	getFFEntry	Accession numberにより、フラットファイル形式でDDBJの検索結果を表示します。 Get Flat File DDBJ Entry by accession number
DDBJ	getXMLEntry	Accession numberによりXML DDBJ Entryを取得します。 Get XML DDBJ Entry by accession number
DDBJ	getFeatureInfo	DDBJ Entryより詳細なFeature Informationを取得します。 Get the detail Feature Information of DDBJ Entry
DDBJ	getAllFeatures	特定のDDBJ Entryを含む特徴的タイプを返す。 Get the feature type contains specified DDBJ Entry
DDBJ	getRelatedFeatures	特定のDDBJ locationを含む関係のある特徴を取得する。 Get related features contains specified DDBJ locaiton
DDBJ	getRelatedFeaturesSeq	特定のDDBJ locationを含む関係のある特徴とシーケンスを取得する。

ページが表示されました

インターネット

検索結果画面 (塩基配列)

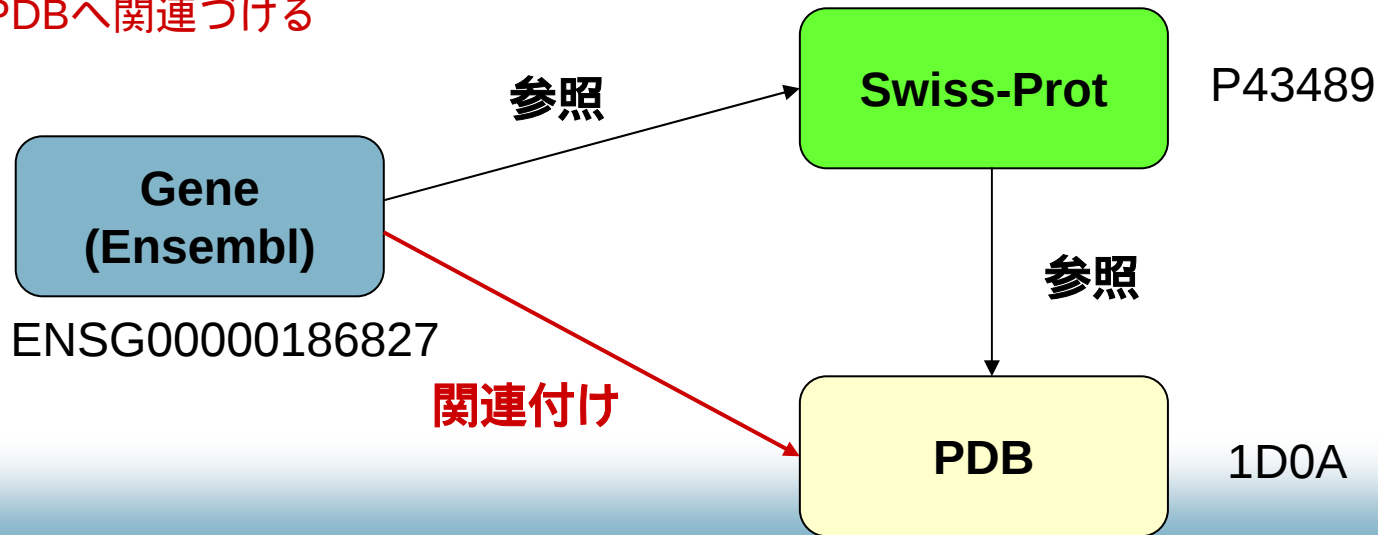
ワークフロー構築例

Ensembl遺伝子の立体構造

- クロスリファレンス情報を元にゲノムの全遺伝子とPDBと関連付け、立体構造まで分かっている遺伝子、タンパク質を分かりやすく整理する。
- Ensemblのヒトの全遺伝子とDDBJ、SwissProt、PDBとの関連付けを行う。
- Webサービス技術を用いて、各データベースのクロスリファレンス情報を取得し、染色体ごとにまとめる。

関連付け詳細

- Ensemblからクロスリファレンスのリストを作成
 - Ensemblの各遺伝子とDDBJエントリ
 - Ensemblの各タンパク質とSwiss-Protエントリ
 - Ensemblの各タンパク質とPDBエントリ
- SRS(Sequence Retrieval System)を用いてクロスリファレンスのリストを作成
 - Swiss-protから、ヒトのエントリでPDBのクロスリファレンスを持つものをピックアップ
 - PDBから、ヒトのデータでSwiss-Protのクロスリファレンスを持つものをピックアップ
- それぞれの結果からSwiss-protとPDBのリスト作成し、間接的に関係するエントリも含めてSwiss-Prot、PDBへ関連づける



実行結果

1番から22番、X、Y染色体にマッピングされている 22,037 個の遺伝子についてDDBJ、Swiss-prot、PDBのエントリへ関連付け

関連付けられた件数

	DDBJ	Swiss-prot	PDB
Ensemblの 情報のみ	18,695	18,713	1,273
ワークフロー の結果	18,695	18,713	1,320

ワークフローを組み立てたい方に

XML Central of DDBJ <http://xml.nig.ac.jp/>

1. ヒトのエントリでPDBデータベースからSwiss-protのクロスリファレンスを持つエントリの一覧取得

SRS `searchSimple("[pdb-Source:HUMAN] & [pdb-Dbref:SWS*]")`

2. ヒトのエントリでSwiss-protデータベースからPDBのクロスリファレンスを持つエントリの一覧取得

SRS

`searchParam("[swissprot-Organism:human*" & [swissprot-DbName:pdb]", "-f acc")`

3. PDB, Swiss-protエントリの取得

GetEntry `getPDBEntry(<accession>), getSWISSEntry(<accession>)`

ワークフローを組み立てたい方に

4 . Ensemblの遺伝子一覧取得

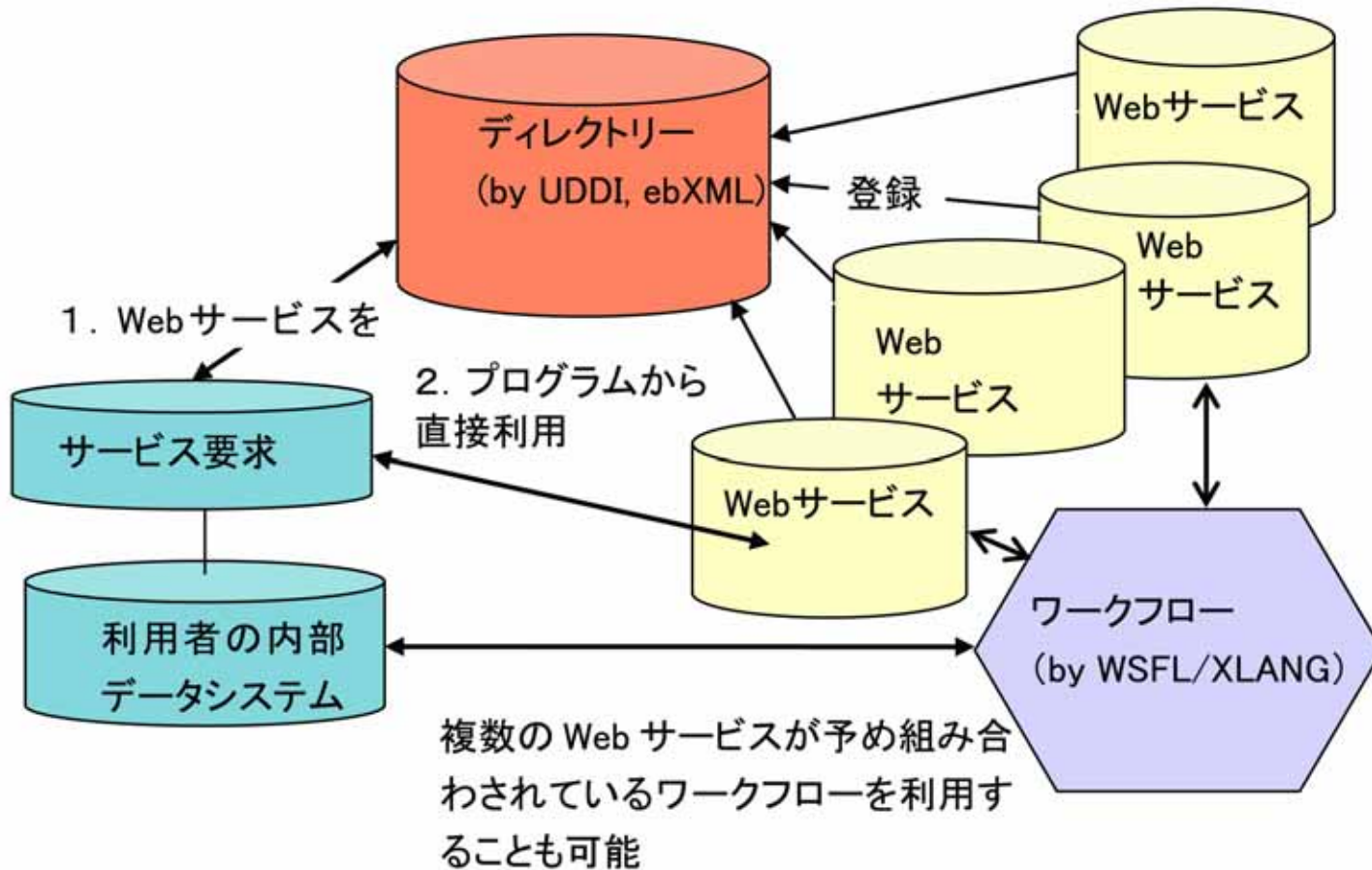
```
Ensembl getGeneList("human")
```

5 . Ensemblの遺伝子情報(遺伝子シンボル名、ゲノム上の位置、タンパク質)の取得

```
Ensembl getGeneInfo(< 遺伝子ID >)
```

```
Ensembl getGeneInfo(< 遺伝子ID >, < 染色体番号 >)
```

Webサービスの理想郷



今後の展開

- ◆ 国内外のWeb services/GRID環境との連携
- ◆ UDDIの整備
- ◆ サービス・オントロジーの整備と普及
- ◆ Semantic WEBへの展開

謝 辞

- ◆ 大学共同利用機関法人情報・システム研究機構
(ROIS)
国立遺伝学研究所生命情報・DDBJ研究センター
- ◆ 科学技術振興機構
バイオインフォマティクス推進事業
(JST BIRD事業)
- ◆ 科学技術振興調整費先導的研究
「新世代バイオポータルの研究開発」

GRIDによる解析グループ

Bioinformatics Group, RIKEN GSC

Akihiko KONAGAYA
Fumikazu KONISHI
Shingo OKI
Hiroshi UMEDA (IBM)

Graduate School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST)

Kenji SATOU
Shinichi TSUJI (NEC SOFTWARE HOKURIKU)
Yasuhiko NAKASHIMA (NEC SOFTWARE
HOKURIKU)

Tokyo Medical and Dental University

Toshinori ENDO (Hokkaido University)
Tekehiro FURUDATE (Hitachi SK)

Japan Science and Technology Agency

Masako KURODA
Kyoko SUZUKI
Toshiyuki KOIKE
Shinya NODAGUCHI
Munakata YOSHIHISA
Shunji KOHNO (Hitachi)
Shinsuke Dohkan (Hitachi)

Center for Information Biology and DDBJ, National Institute of Genetics (NIG)

Hideaki SUGAWARA
Satoru MIYAZAKI (Tokyo Science University)
Masahito YAMAGUCHI (Fujitsu)
Yasumasa SHIGEMOTO (Fujitsu)
Masashi MATSUO (Fujitsu)
Keiichi IDA (Fujitsu IST)
Kazutaka SUGIMOTO (Fujitsu IST)

微生物ゲノムORF再評価グループ

Takashi ABE (2)
Masaki HIRAHATA (1,2)
Takehide KOSUKUGE (2)
Satoru MIYAZAKI (4)
Toshihisa OKIDO (2)
Yasumasa SHIGEMOTO (3)
Naoto TANAKA (1,2)

- (1) JST BIRD
- (2) NIG CIB・DDBJ
- (3) FUJITSU LTD
- (4) Fac. Pharm. Sci., Tokyo Univ. of Sci.