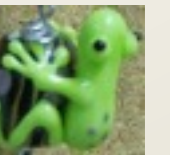


医学生物学分野におけるデータのオープン化と そのインセンティブ

仲里 猛留

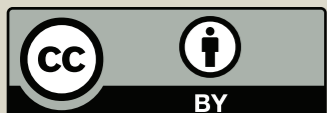
NAKAZATO, Takeru

@chalkless



情報・システム研究機構 データサイエンス共同利用基盤施設
ライフサイエンス統合データベースセンター

Database Center for Life Science (DBCLS),
Joint Support-Center for Data Science Research, Research Organization of Information and Systems (ROIS)



2016/10/26

国立情報学研究所

自己紹介

略歴

東工大院・生命理工

浸透圧調節・イオン輸送

イオントランスポーターの
クローニング

99.4

Wet

02.4

阪大院・情報科学

文献情報を利用した

マイクロアレイデータの
生物学的知見の付与

05.10

Dry

08.9

NEC バイオIT事業推進センター

文献検索（もどき）ツールの開発

遺伝子（群）への文献情報を用いた
アノテーションづけ

Dry

07.4

（部署解体 → 異動）

休眠時代 毎日、PowerPointで営業資料作成

07.9

ライフサイエンス

統合データベースセンター

遺伝子、疾患のアノテーション

キーワードづけ、用語整備
NGSデータの整理

Dry

16.10

昔は

ウナギの海水適応機構

血压調節

分子生物学っぽく
言ってみる

イオン濃度調節

mouse の系

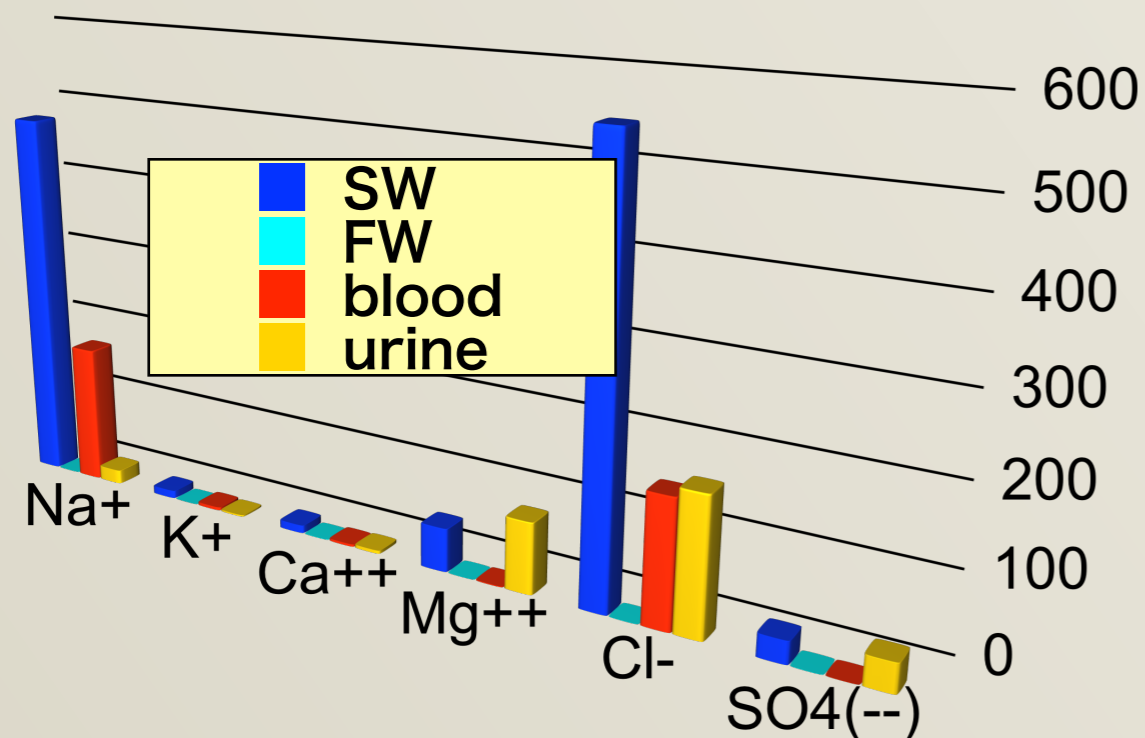
高Na食 or 高K食
変化が見にくい

ウナギ

淡水と海水を行き来
(サケ、マスと同じ)



Anguilla japonica



淡水／海水で遺伝子発現が
どうか変わるか。

(イオントランスポーター中心)

本業

DDBJの集めた公共NGSデータの検索サイト運用

The screenshot shows the DBCLS SRA website interface. At the top, there is a navigation bar with the site name 'DBCLS SRA' and a logo. Below the navigation bar, there are three main sections: 'Species', 'Study Type', and 'Platform', each with a bar chart showing the distribution of data. A 'Usage' section provides instructions on how to interact with the charts. At the bottom, there is a 'Free Keyword' search bar.

DBCLS SRA

DISCOVER
Interesting & Available SRA Data

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

Trends & Search SRA data

→ [for more detail](#)

Species

- Homo sapiens
- Mus musculus
- human gut metagenome
- Oryza sativa
- soil metagenome

Species	Count
Homo sapiens	301929
Mus musculus	88239
human gut metagenome	40061
Oryza sativa	29622
soil metagenome	22162

Q Search

Study Type

- Whole Genome Sequencing
- Other
- Transcriptome Analysis
- Metagenomics
- Population Genomics

Study Type	Count
Whole Genome Sequencing	27402
Other	15292
Transcriptome Analysis	7758
Metagenomics	5105
Population Genomics	723

Select your option

Platform

- Illumina HiSeq 2000
- Illumina MiSeq
- 454 GS FLX Titanium
- Illumina Genome Analyzer II
- Illumina HiSeq 2500

Platform	Count
Illumina HiSeq 2000	572366
Illumina MiSeq	97400
454 GS FLX Titanium	87105
Illumina Genome Analyzer II	68519
Illumina HiSeq 2500	59793

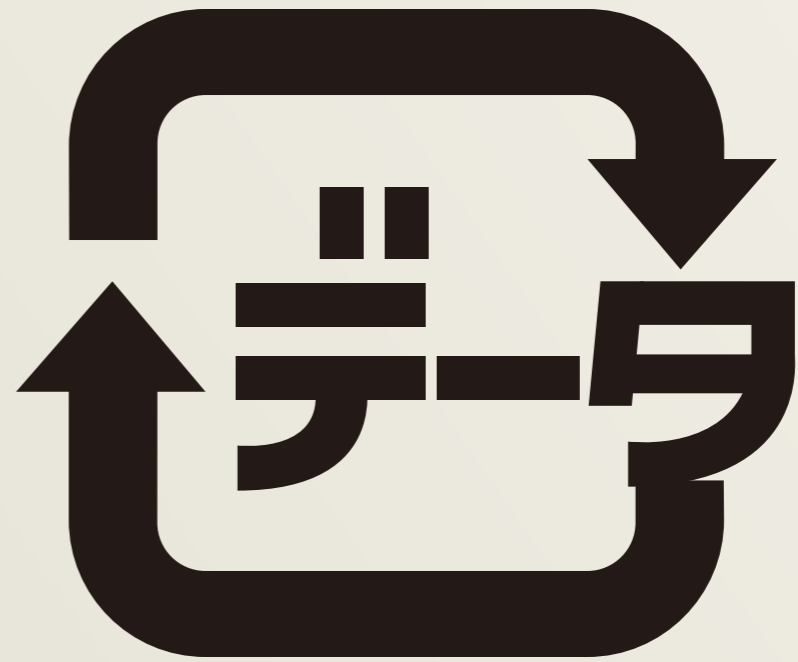
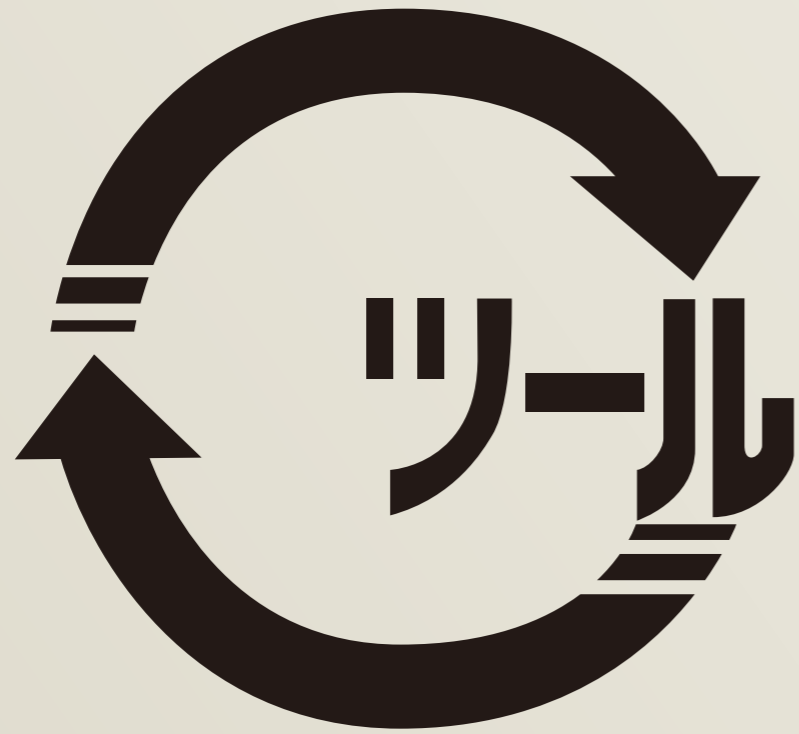
Q Search

Usage

- Click bars or bubbles and you can see more details of selected data.
- Input keywords and you can view ratio of feature selection.
- You can see result of combined search criteria too.

Search Conditions

Free Keyword



生命科学分野における データベース

現状：NCBIのデータベースと登録数

Literature

Books	536,435	books and reports
MeSH	265,382	ontology used for PubMed indexing
NLM Catalog	1,553,923	books, journals and more in the NLM Collections
PubMed	26,562,500	scientific & medical abstracts/citations
PubMed Central	4,114,647	full-text journal articles

Health

ClinVar	170,659	human variations of clinical significance
dbGaP	223,863	genotype/phenotype interaction studies
GTR	48,790	genetic testing registry
MedGen	293,286	medical genetics literature and links
OMIM	24,895	online mendelian inheritance in man
PubMed Health	63,329	clinical effectiveness, disease and drug reports

Genomes

Assembly	93,453	genome assembly information
BioProject	200,402	biological projects providing data to NCBI
BioSample	5,408,512	descriptions of biological source materials
Clone	38,083,623	genomic and cDNA clones
dbVar	6,164,814	genome structural variation studies
Genome	17,491	genome sequencing projects by organism
GSS	39,695,576	genome survey sequences
Nucleotide	218,585,723	DNA and RNA sequences
Probe	32,405,048	sequence-based probes and primers
SNP	819,309,821	short genetic variations
SRA	3,281,545	high-throughput DNA and RNA sequence read archive
Taxonomy	1,628,067	taxonomic classification and nomenclature catalog

Genes

EST	76,321,765	expressed sequence tag sequences
Gene	24,930,660	collected information about gene loci
GEO DataSets	2,054,326	functional genomics studies
GEO Profiles	128,414,055	gene expression and molecular abundance profiles
HomoloGene	141,268	homologous gene sets for selected organisms
PopSet	259,661	sequence sets from phylogenetic and population studies
UniGene	6,473,284	clusters of expressed transcripts

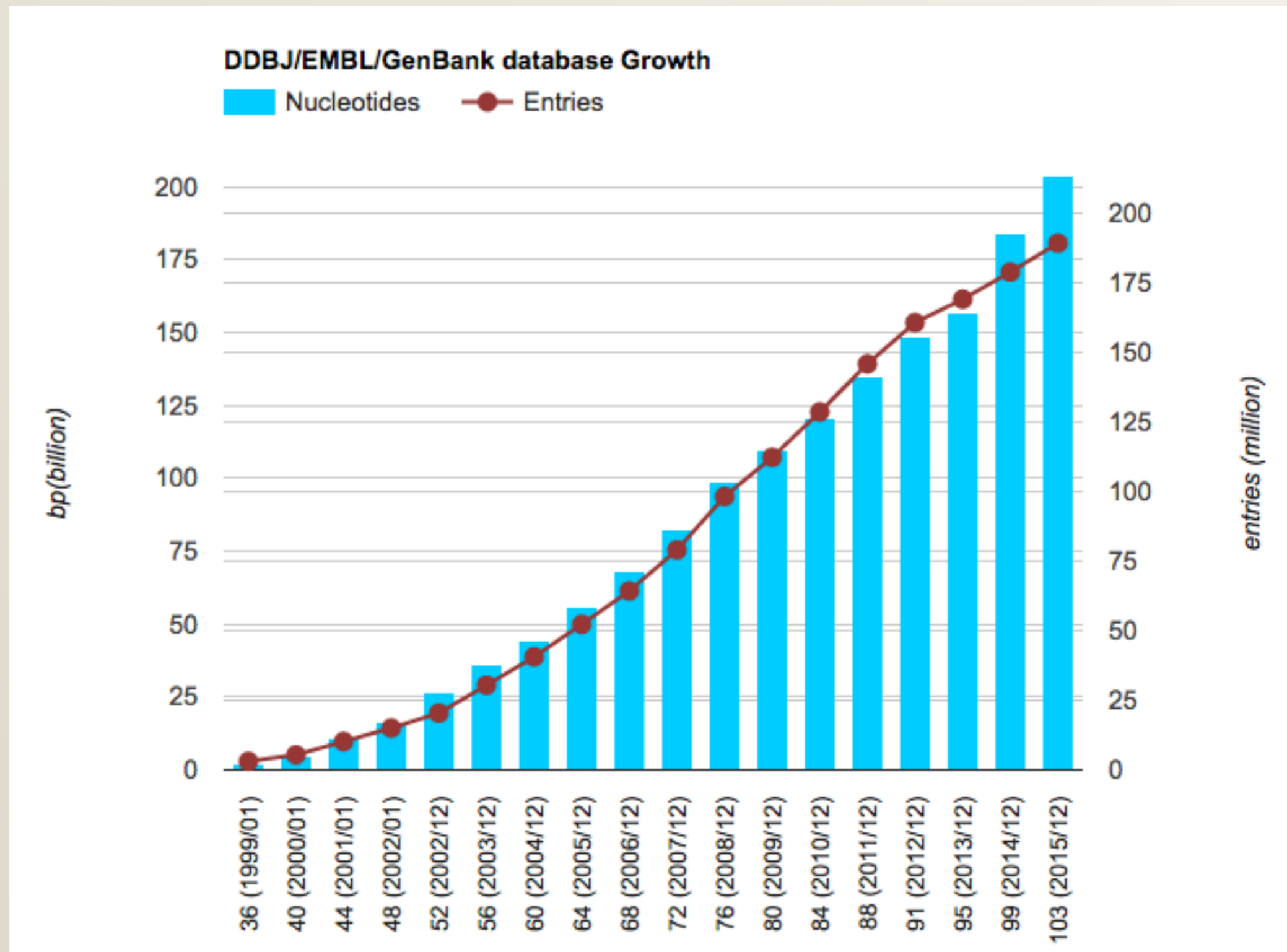
Proteins

Conserved Domains	52,411	conserved protein domains
Protein	317,695,190	protein sequences
Protein Clusters	820,546	sequence similarity-based protein clusters
Structure	122,523	experimentally-determined biomolecular structures

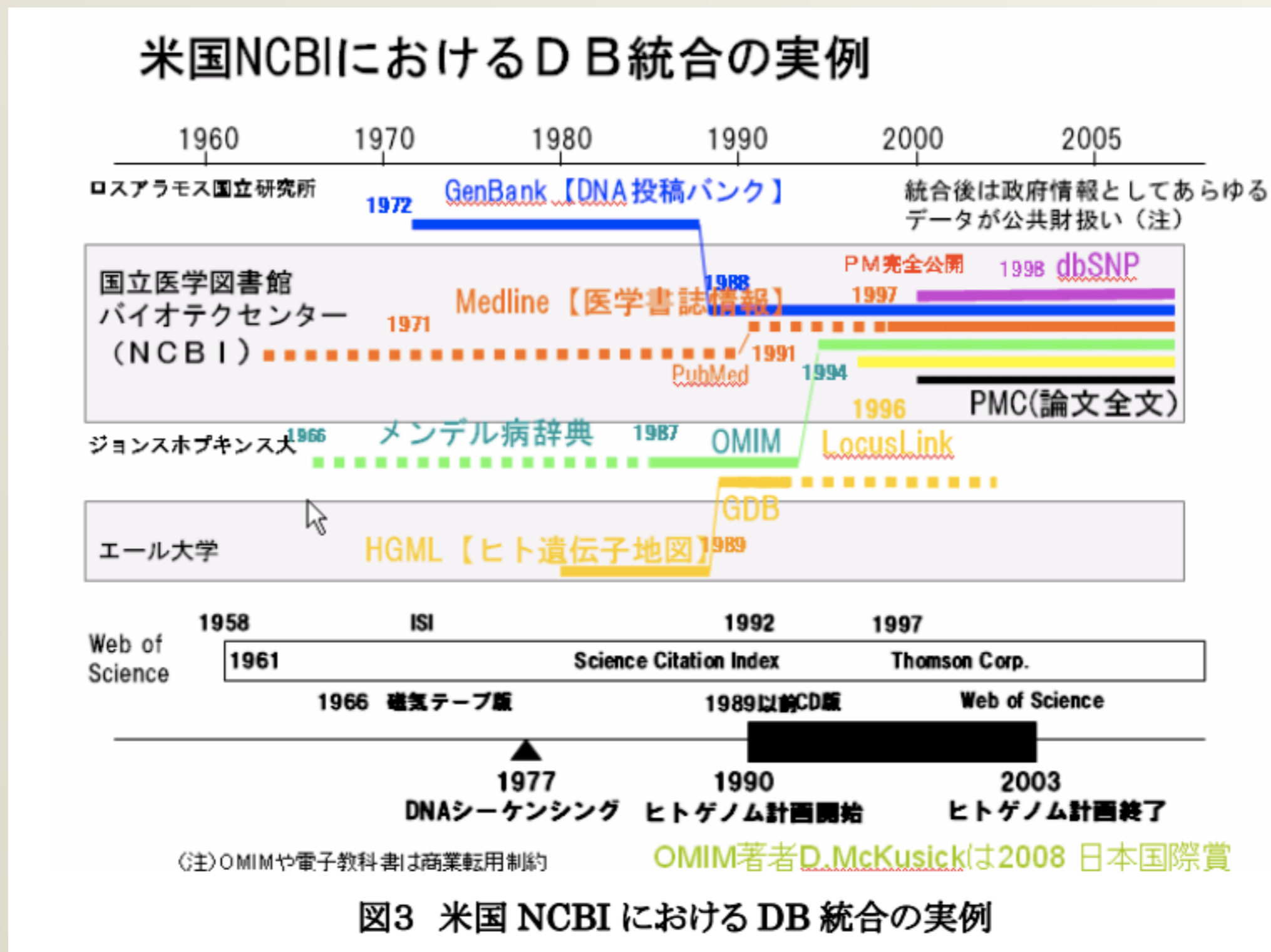
Chemicals

BioSystems	918,220	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	1,218,719	bioactivity screening studies
PubChem Compound	92,340,732	chemical information with structures, information and links
PubChem Substance	223,912,985	deposited substance and chemical information

現状：塩基（≒遺伝子）データの登録量



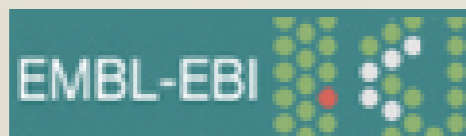
生命科学分野のデータベースの歴史



公共データベース ≡ INSDC

EMBL (European Mol. Biol. Lab.)

(欧)



INSDC
(Int'l Nucleotide Seq. DB Collab.)



ROIS

DDDBJ
DNA Data Bank of Japan

(米)

(日)

NIH/
NLM (Natl. Lib. of Med.)

情報・システム研究機構
国立遺伝学研究所

日々生まれるデータベース



Nucleic Acids Research

年に一度の Database Issue と Web Server Issue

The 2016 Nucleic Acids Research Database Issue is the 23rd annual collection of descriptions of various molecular biology databases. It includes 178 papers, of which 62 describe newly created databases (Table 1), 95 papers provide updates on databases that have been described in the previous NAR Database Issues and 17 contain updates on databases whose descriptions have previously been published in other journals (Table 2).

主要な生命科学データベース1:

PubMed: 生命科学文献検索サービス

PubMed(詳細画面)

論文PDF

The screenshot shows the PubMed interface for the article. The title is "Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive" by Nakazato T, Ohta T, Bono H. The abstract is visible, and there are links for "Full text links" and "PMC Full text". A red box highlights the "Full text links" section, with a red arrow pointing to the PDF viewer on the right. A yellow speech bubble with the word "リンク" (Link) is positioned over the "Full text links" section.

リンク

The screenshot shows the PLOS ONE website displaying the PDF of the article. The title is "Experimental Design-Based Functional Mining and Characterization of High-Throughput Sequencing Data in the Sequence Read Archive" by Takeru Nakazato, Tazuo Ohta, Hidemasa Bono. The abstract is visible, and there are links for "Full text links" and "PMC Full text". A red box highlights the abstract section, with a red arrow pointing from the PubMed screenshot on the left. A yellow speech bubble with the text "Abstract (要約) を収録" (Abstract included) is positioned over the abstract section.

Abstract (要約) を収録

主要な生命科学データベース1:

PubMed : 生命科学文献検索サービス

<http://pubmed.gov/>

(本当は <http://www.ncbi.nlm.nih.gov/pubmed/>)

- NIHの図書館部門 (National Library of Medicine) が生命科学系の雑誌記事を収集
- メインは1950年代～ (さかのぼって登録中)
- 現在、2600万件 (増加中)
- PubMed はAbstだけだが、15%は全文がPMCで閲覧可能

1879 : NLMがIndex Medicusを出版 (月刊の論文索引集)

1960 : コンピューター化 = MEDLARS

1965 : 検索サービススタート (郵送ベース)

1971 : オンライン化 : MEDLINE (MEDLAR Online)

1996 : インターネットで無料で検索 : PubMed (Public MEDLINE)

参考 : <https://ja.wikipedia.org/wiki/MEDLINE>

主要な生命科学データベース2：

BLAST：類似遺伝子検索ツール

DNA/タンパク質配列を入力



データベース中から

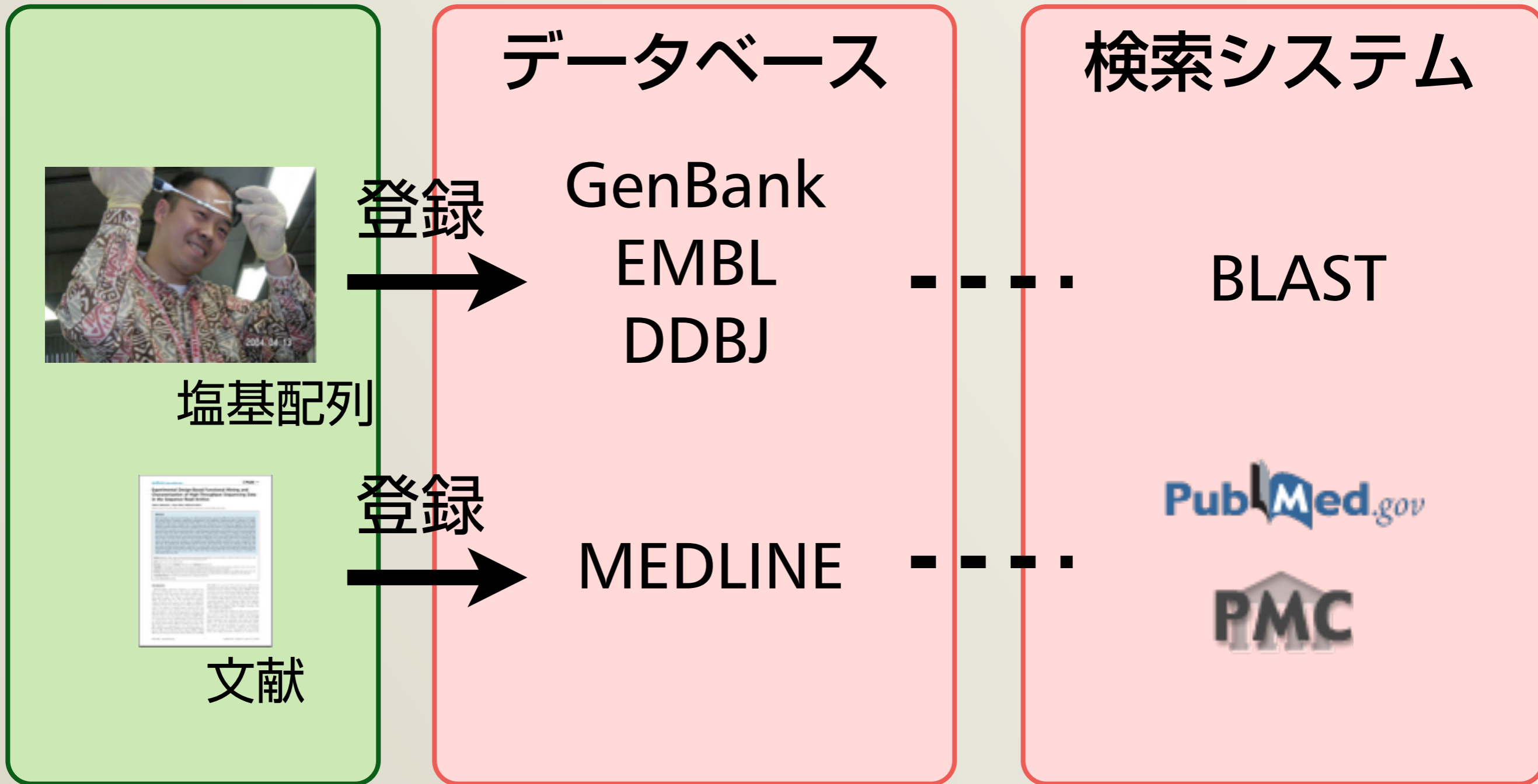
- ・ 配列の類似したエントリ
- ・ その類似度 などを表示

The screenshot shows the BLAST web interface. At the top, it says 'BLAST' and 'blastp suite'. Below that, there's a search bar and a 'BLAST Results' section. The 'Protein Sequence (78 letters)' is shown. There's a 'Graphic Summary' section with a 'Distribution of 47 Blast Hits on the Query Sequence' chart. The chart shows a color key for alignment scores: <math> <math>

Accession	Score	Expect	Method	Identical	Positives	Gaps
RefName: Full-Metallothionein-like protein 4B, AName: Full-Class I metallothionein-like protein 4B, AName: Full-OsMT1-4B	45.1	45.1	67%	8e-07	52%	Q226E3.1
RefName: Full-Metallothionein-like protein type 2	44.3	44.3	71%	1e-06	52%	Q435I2.1
RefName: Full-Metallothionein-like protein 3A, AName: Full-Class I metallothionein-like protein 3A, AName: Full-OsMT1-3a, Short-OsMT1a	43.9	43.9	62%	2e-06	44%	A32420.1
RefName: Full-Metallothionein-like protein 4C, AName: Full-Class I metallothionein-like protein 4C, AName: Full-MT-1-4c, AName: Full-OsMT1-4c, AName: Full-OsMT1c	43.5	43.5	67%	4e-06	50%	Q226C3.1
RefName: Full-Metallothionein-like protein 1, Short-MT1	43.1	43.1	67%	4e-06	45%	Q24528.1
RefName: Full-Metallothionein-like protein 1, Short-MT1	40.0	40.0	69%	7e-05	41%	F53455.1
RefName: Full-Metallothionein-like protein 2A, AName: Full-Class I metallothionein-like protein 2A, AName: Full-OsMT1-2a, Short-OsMT2a, AName: Full-OsMT2.1	38.5	38.5	91%	4e-04	57%	F5H029.1
RefName: Full-Metallothionein-like protein type 2 LSC210 (Stressis repeat)	37.0	37.0	30%	4e-04	71%	Q96353.1
RefName: Full-Metallothionein-like protein 1, Short-MT1	37.7	37.7	67%	5e-04	40%	F59571.1
RefName: Full-Metallothionein-like protein type 2 MET1	37.0	37.0	62%	0.001	53%	F50134.1
RefName: Full-Metallothionein-like protein 1, Short-MT1	33.1	33.1	67%	0.031	43%	F50571.1
RefName: Full-Metallothionein-like protein 3A, AName: Full-Class I metallothionein-like protein 3A, AName: Full-OsMT1-3a, Short-OsMT3, Short-OsMT3a	30.4	30.4	47%	0.34	45%	A26630.1
RefName: Full-Metallothionein-like protein type 2 (Andria delicosa)	30.0	30.0	43%	0.47	54%	F53369.1
RefName: Full-Metallothionein-like protein 3, Short-MT3	29.6	29.6	78%	0.75	36%	Q24433.1
RefName: Full-FoxO/FOXO domain protein	30.4	30.4	20%	1.1	56%	Q92150.1
RefName: Full-Metallothionein-like protein DMS3 (Pisum sativum)	28.5	28.5	78%	1.8	33%	Q4054.1

The screenshot shows the 'Alignments' section of the BLAST results. It displays two alignment entries. The first entry is for 'Full-Metallothionein-like protein type 2 [Andria delicosa]' with sequence ID 'M33300.1' and length 78. The alignment shows a perfect match (100%) between the query and subject sequences. The second entry is for 'Full-Metallothionein-like protein type 2 [Malus domestica]' with sequence ID 'Q24058.1' and length 79. The alignment shows a high similarity (89%) between the query and subject sequences. Each entry includes a 'Download' button and a 'Related Information' link.

BLAST: 類似遺伝子検索ツール



**なぜ公共データベースに
データが集まるのか？**

投稿規定での要求

Mandates for specific datasets

<http://www.nature.com/authors/policies/availability.html>

For the following types of data set, submission to a community-endorsed, public repository is mandatory. Accession numbers must be provided in the paper. Examples of appropriate public repositories are listed below.

Mandatory deposition	Suitable repositories
Protein sequences	Uniprot
DNA and RNA sequences	Genbank
	DNA DataBank of Japan (DDBJ)
	EMBL Nucleotide Sequence Database (ENA)
DNA and RNA sequencing data	NCBI Trace Archive
	NCBI Sequence Read Archive (SRA)
Genetic polymorphisms	dbSNP
	dbVar
	European Variation Archive (EVA)
Linked genotype and phenotype data	dbGAP
	The European Genome-phenome Archive (EGA)
Macromolecular structure	Worldwide Protein Data Bank (wwPDB)
	Biological Magnetic Resonance Data Bank (BMRB)
	Electron Microscopy Data Bank (EMDB)
Microarray data (must be MIAME compliant)	Gene Expression Omnibus (GEO)
	ArrayExpress
Crystallographic data for small molecules	Cambridge Structural Database

Funding Agencyからの要求

米 NIH

To facilitate data sharing, investigators submitting a research application requesting \$500,000 or more of direct costs in any single year to NIH on or after October 1, 2003 are expected to include a plan for sharing final research data for research purposes, or state why data sharing is not possible.

参考：NIH Data Sharing Policy: https://grants.nih.gov/grants/policy/data_sharing/

日本でも

4 バイオサイエンスデータベースセンターへの協力

バイオサイエンスデータベースセンター (URL:<http://biosciencedbc.jp/>) は、様々な研究機関等によって作成されたライフサイエンス分野データベースの統合的な利用を推進するために、国立研究開発法人科学技術振興機構に設置されています。

同センターでは、関連機関に積極的な参加を働きかけるとともに、戦略の立案、ポータルサイトの構築・運用、データベース統合化基盤技術の研究開発、バイオ関連データベース統合化の推進を4つの柱として、ライフサイエンス分野データベースの統合化に向けて事業を推進しています。これによって、我が国におけるライフサイエンス分野の研究成果が、広く研究者コミュニティに共有かつ活用されることにより、基礎研究や産業応用研究につながる研究開発を含むライフサイエンス分野の研究全体が活性化されることを目指しています。

ついては、ライフサイエンス分野に関する論文発表等で公表された成果に関わる生データの複製物、又は構築した公開用データベースの複製物について、同センターへの提供に御協力をお願いします。

なお、提供された複製物については、非独占的に複製・改変その他必要な形で利用できるものとします。また、複製物の提供を受けた機関の求めに応じ、複製物を利用するに当たって必要となる情報の提供にも御協力

データのオープン化へのインセンティブ

- **自分の論文が掲載される = 研究者の究極の目的**
- **自分のデータが使ってもらえる、論文が引用される**
 - 昔：データの囲い込み（ジャイアニズム）
 - 今：オープンにした方がプレゼンスが上がる
(世の中を動かせる)
- **研究費がもらえる → 次の成果へ**
- **付加価値の付与**
 - データを登録することでウェブツールで解析可能に
他のデータベース、ツールとのリンク・連携

研究のプレゼンスの例（研究の再現性）

The screenshot shows the DRASearch interface for study SRP038104. The browser address bar shows the URL: trace.ddbj.nig.ac.jp/DRASearch/study?acc=SRP038104. The page title is "DRASearch" and the study ID is "SRP038104".

Study Detail	
Title	Mus musculus 1)Transcriptome or Gene expression; 2)Histone Modification H3K4me3, H3K27me3
Study Type	Transcriptome Analysis
Abstract	Global Expression Profile and Epigenetic profile of STAP and other types of ES cells.

Navigation links: [Send Feedback](#), [Search Home](#), [DRA Home](#)

The screenshot shows the NCBI BioProject page for project PRJNA238286. The browser address bar shows the URL: <https://www.ncbi.nlm.nih.gov/bioproject/238286>. The page title is "BioProject" and the project name is "Mus musculus strain:C57BL/6, C57BL/6x129sv (house mouse)".

Accession: PRJNA238286 ID: 238286

Mus musculus 1)Transcriptome or Gene expression; 2)Histone Modification H3K4me3, H3K27me3

Global Expression Profile and Epigenetic profile of STAP and other types of ES cells.

Accession	PRJNA238286
Data Type	Transcriptome or Gene expression
Scope	Multisolate
Organism	Mus musculus [Taxonomy ID: 10090] Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus; Mus musculus
Publications	1. Published online: title: Nature Letter Print Email Share/bookmark 日本語要約 Bidirectional developmental potential in reprogrammed cells with acquired pluripotency; journal: Nature; year: 2014; volume: 505; issue: 7485; pages_from: 676; pages_to: 680; multiple_authors: True; author: Haruko Obokata
Submission	Registration date: 13-Feb-2014

Display Settings: [Send to](#)

[See Genome Information for Mus musculus](#)

[NAVIGATE ACROSS](#)
18745 additional projects are related by organism.

Navigation		
Submission	SRA110029	FTP
Experiment	SRX472627	FASTQ SRA
	SRX472628	FASTQ SRA
	SRX472629	FASTQ SRA
	SRX472630	FASTQ SRA
	SRX472631	FASTQ SRA
	SRX472632	FASTQ SRA
	SRX472633	FASTQ SRA
	SRX472634	FASTQ SRA
	SRX472635	FASTQ SRA
	SRX472636	FASTQ SRA
	SRX472637	FASTQ SRA
	SRX472638	FASTQ SRA
	SRX472639	FASTQ SRA
	SRX472640	FASTQ SRA
	SRX472641	FASTQ SRA
	SRX472642	FASTQ SRA
	SRX472643	FASTQ SRA
SRX472644	FASTQ SRA	
SRX472645	FASTQ SRA	
SRX472646	FASTQ SRA	
SRX472647	FASTQ SRA	
SRX472648	FASTQ SRA	
SRX472649	FASTQ SRA	
SRX472650	FASTQ SRA	

データのオープン化の課題

データをオープンにする手段

- 公共データベースに登録
- データジャーナルにsubmit
(Scientific Data, GigaScience, ...)
- 機関レポジトリを利用
- 自分でデータベースを作成して公開

データのオープン化に求められるもの

- データを参照するしくみ

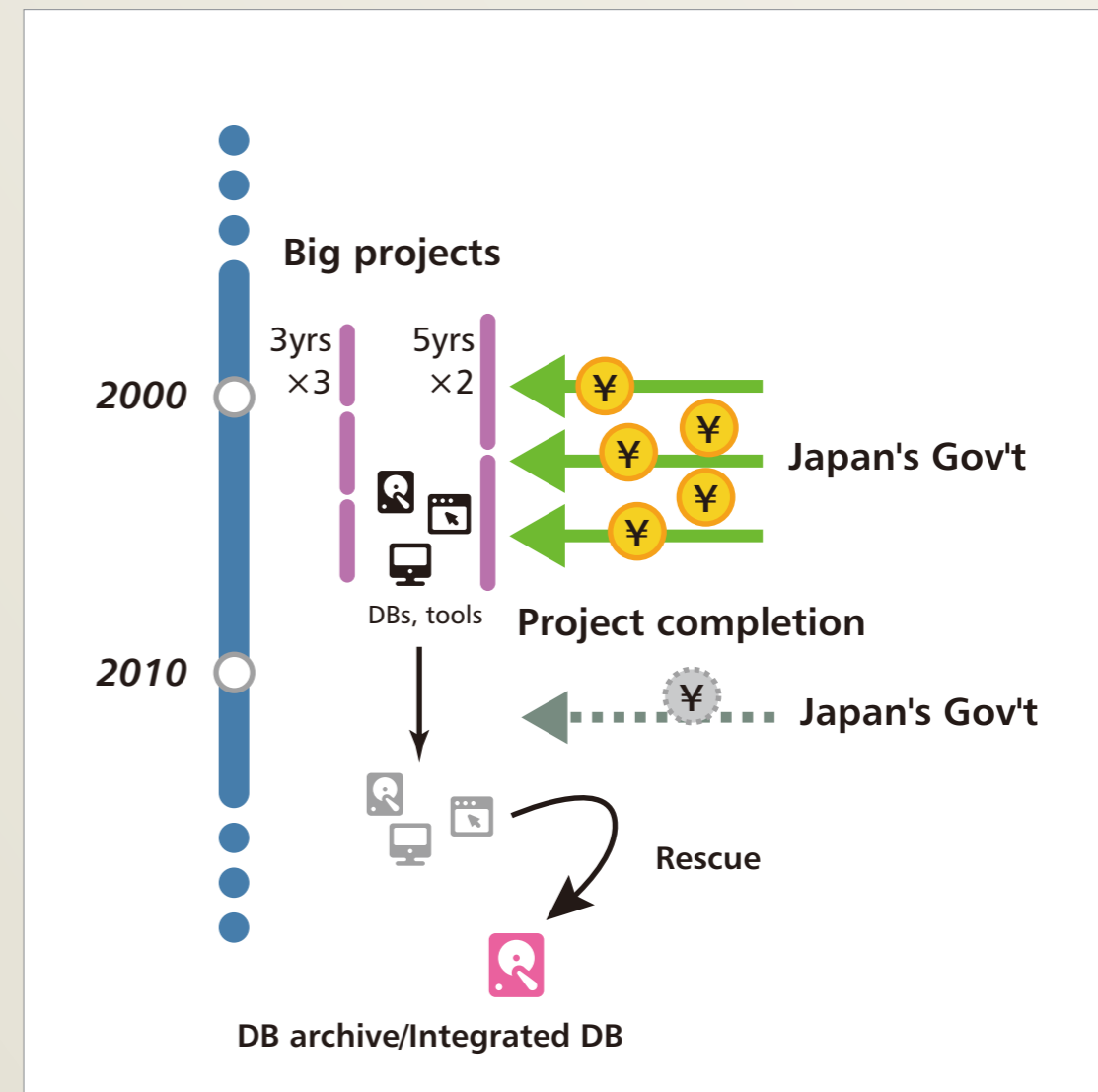
アクセッション番号（登録ID、文献ID、…）

DOI

URL などなど

- 永続性

- 維持費用



インセンティブの面から見たオープン化

- 自分の論文が掲載される = 研究者の究極の目的
- 自分のデータが使ってもらえる、論文が引用される
 - 昔：データの囲い込み（ジャイアニズム）
 - 今：オープンにした方がプレゼンスが上がる
(世の中を動かせる)
- 研究費がもらえる → 次の成果へ

研究者

出版社

機関
レポジトリ

リスペクトするしくみを！

- データを参照する = リスペクトする
- 参考例：計算機資源の提供（遺伝研のスパコン）

Acknowledgments

We thank K. Osaki and M. Kitazume at Tomy Digital Biology Co. Ltd for their technical support with sequencing and *de novo* assembly, and M. Ezure for technical support and helpful discussions. The computational analysis was performed using the supercomputer system at the National Institute of Genetics, the Research Organization of Information and Systems. This study was supported partly by a Grant-in-Aid for Young Scientists (A) (25712032) from the Japan Society for the Promotion of Sciences and an NIG Collaborative Research Program (2012–2088, 2013–2070) from the National Institute of Genetics.

- 使われている感があれば予算につながる???

データのオープン化の弊害

ヒト疾患研究

- ・データの解像度がよくなりすぎて、個人が識別できるレベルに
 - ・稀な疾患だと、その患者というだけで個人が特定されかねない
- ヒトデータ用のデータベースを用意。

Controlled Accessで

世界的にはGA4GH (Global Alliance for Genomics and Health) で議論

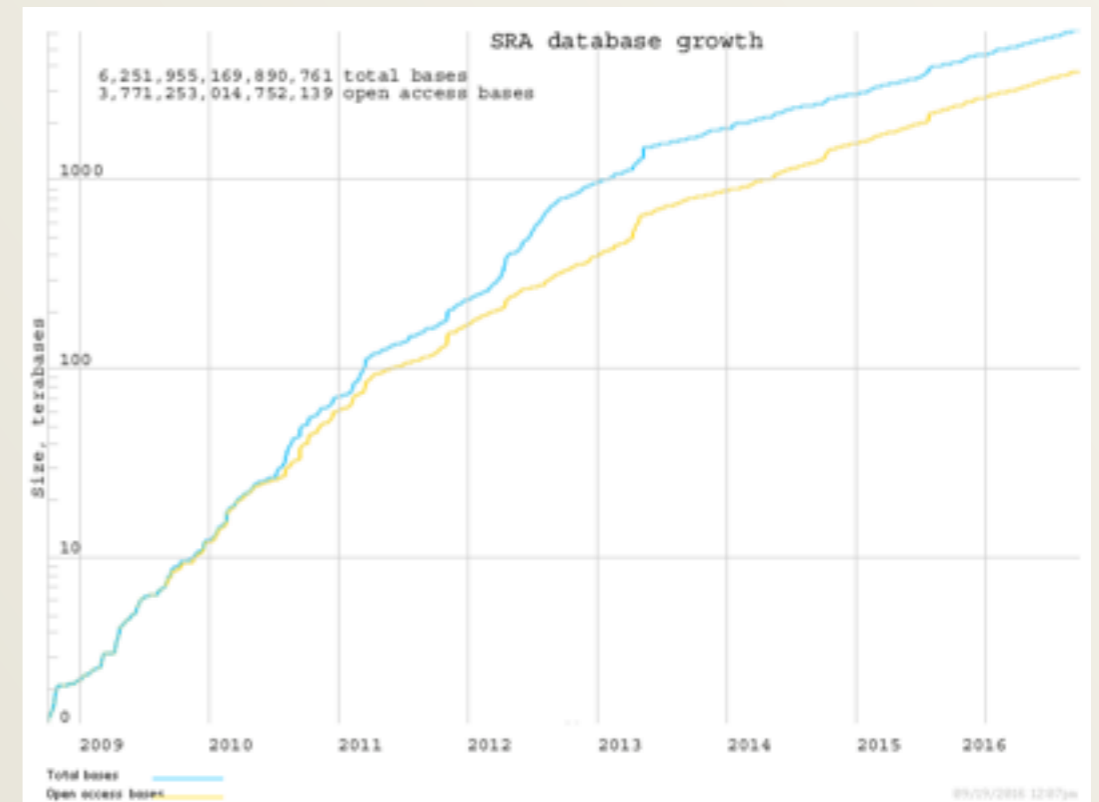
(Data WG, Security WG, Clinical WG, Regulatory and Ethics WG)

生態学・博物館

希少な動植物の採集地を見ての乱獲

→ 市町村・地名、緯度経度高度は書かない。

DB中では隠すでなく消しておく



<https://trace.ncbi.nlm.nih.gov/Traces/sra/>

NIPPON: Chiba-ken,
Sakura-shi, Nishimikado,
24 V 2013, MARUYAMA M.
35°38'7" N 140°15'12" E
20 m 佐倉市 西御門

図 15. 筆者の作成する典型的なラベル。

研究現場のデータの現状



← せっかくの宝の山も
持ち腐れに↓



カタツケていきましよう