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26年度キックオフミーティング, 東京JST, 2014年6月2日

# 蛋白質構造データバンクの高度化と統合的運用

代表研究者：中村 春木
分担研究者：藤原 敏道
（大阪大学蛋白質研究所・教授）

http://pdbj.org/
http://wwpdb.org/

©2014中村 春木(大阪大学) licensed under CC表示2.1日本

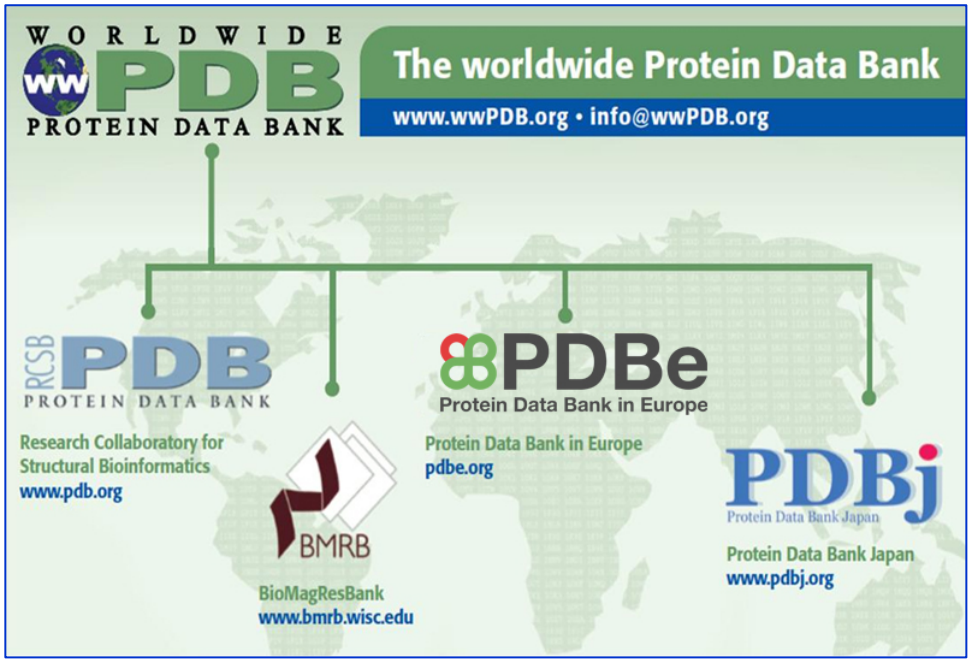
# 蛋白質の形のデータバンク:PDB

・生体高分子の構造(形)に関する全ての情報を集めた国際的データベース:データバンク方式(1971年から開始)

・情報は、研究者・教育者・学生・企業を問わず無償で利用される

・運営は各国(米国、欧州、日本)の政府機関による研究費用でまかなわれている

→　国際組織wwPDBとして活動(PDBjは2003年からの創立メンバー)




X線結晶回折実験,
核磁気共鳴(NMR)法,
超低温電子顕微鏡(EM),
による実験結果のアーカイブ

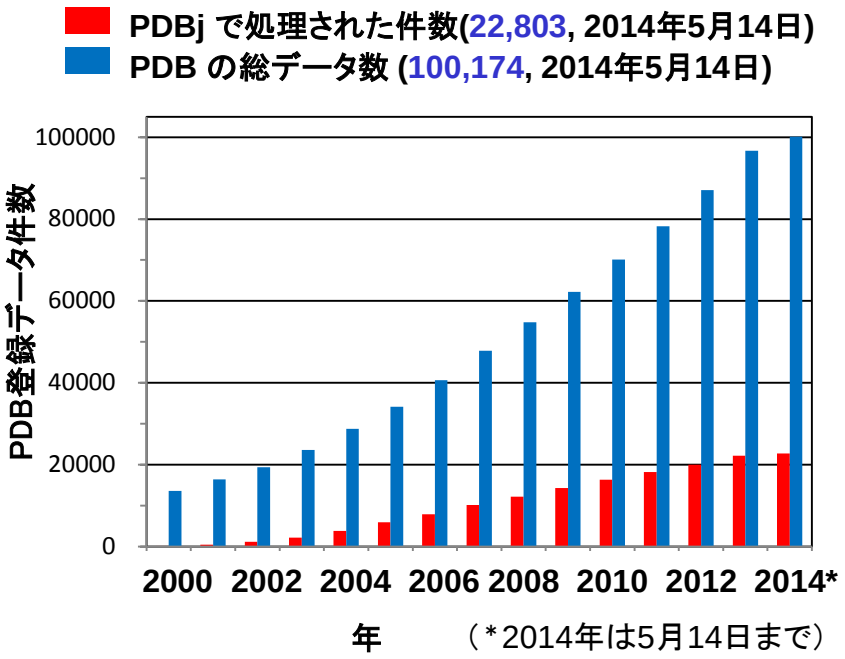
# PDBjの活動

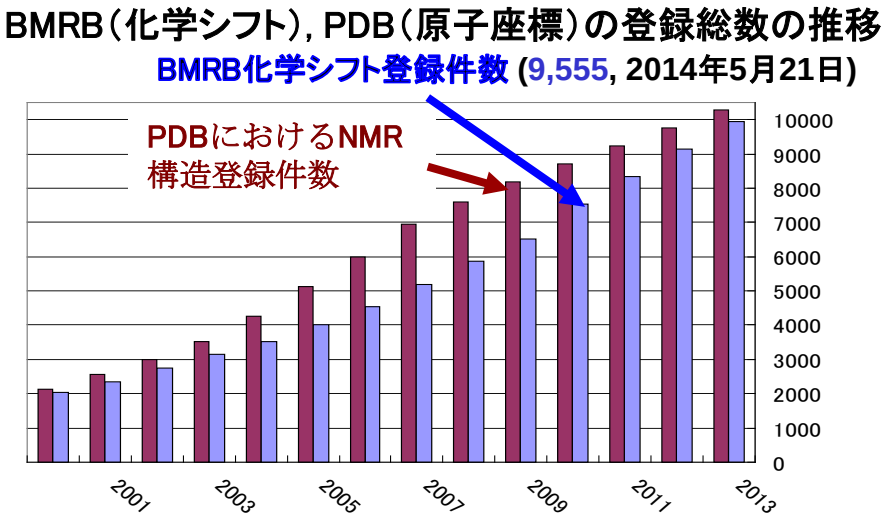
- Data-inの活動：
  wwPDBの一員として品質管理をしつつ登録作業を実施

  新たな標準フォーマット等の開発(PDB/RDF, BMRB/RDF)

- Data-out の活動：
  共通データのダウンロードサイト（毎週アプデート）の運営

  種々のサービスや二次データベースの開発・提供



# PDBj, PDBj-BMRBのData-out活動

## http://pdbj.org/

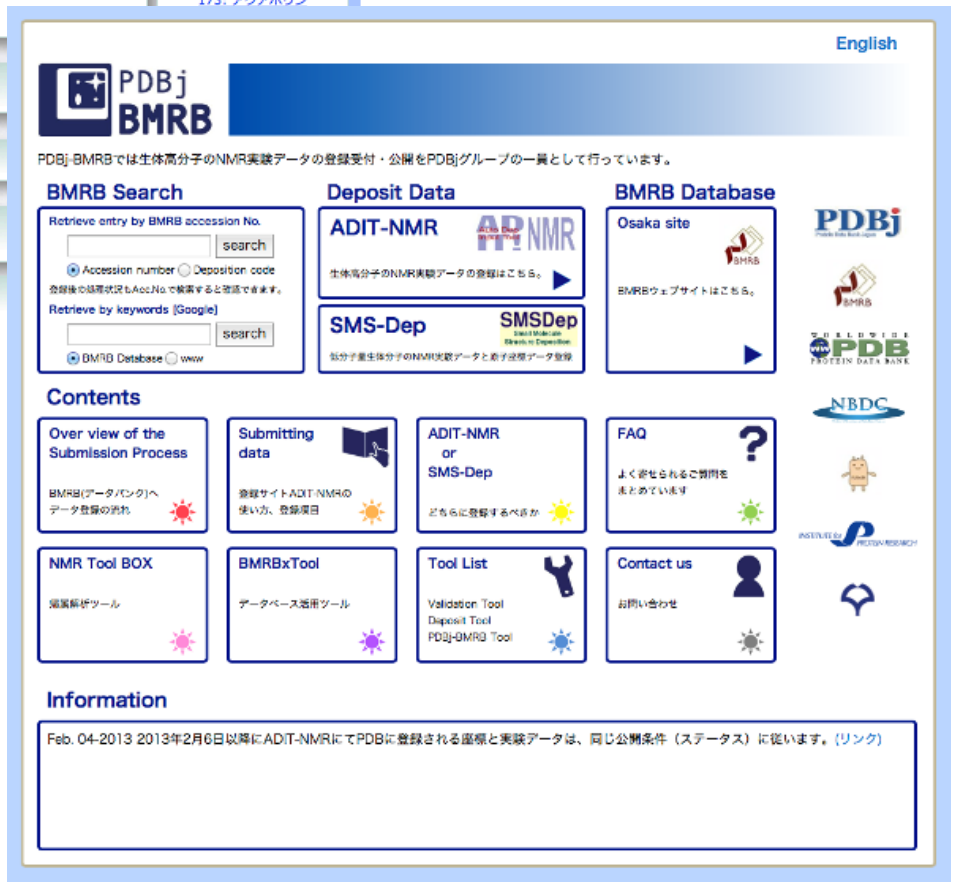
NBDC北海道
データセンター

NBDC
National Bioscience Database Center

大阪大学
蛋白質研究所

**ポータルを二重化**

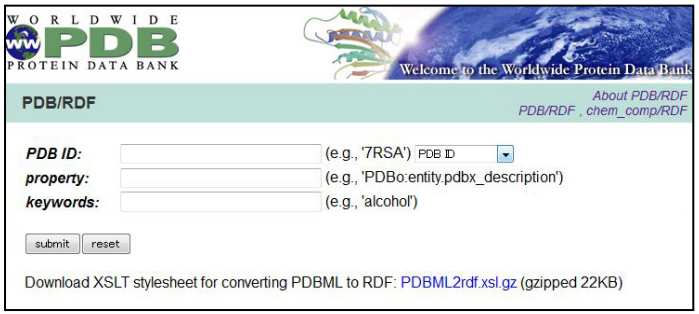
## PDBj-BMRB



# 世界標準のRDF format の開発

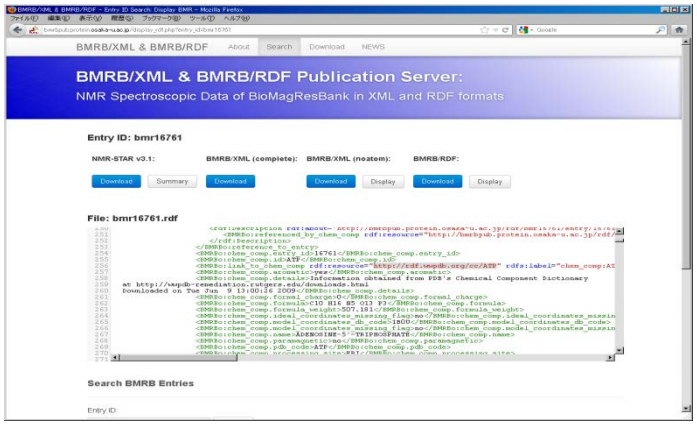
## PDB/RDF

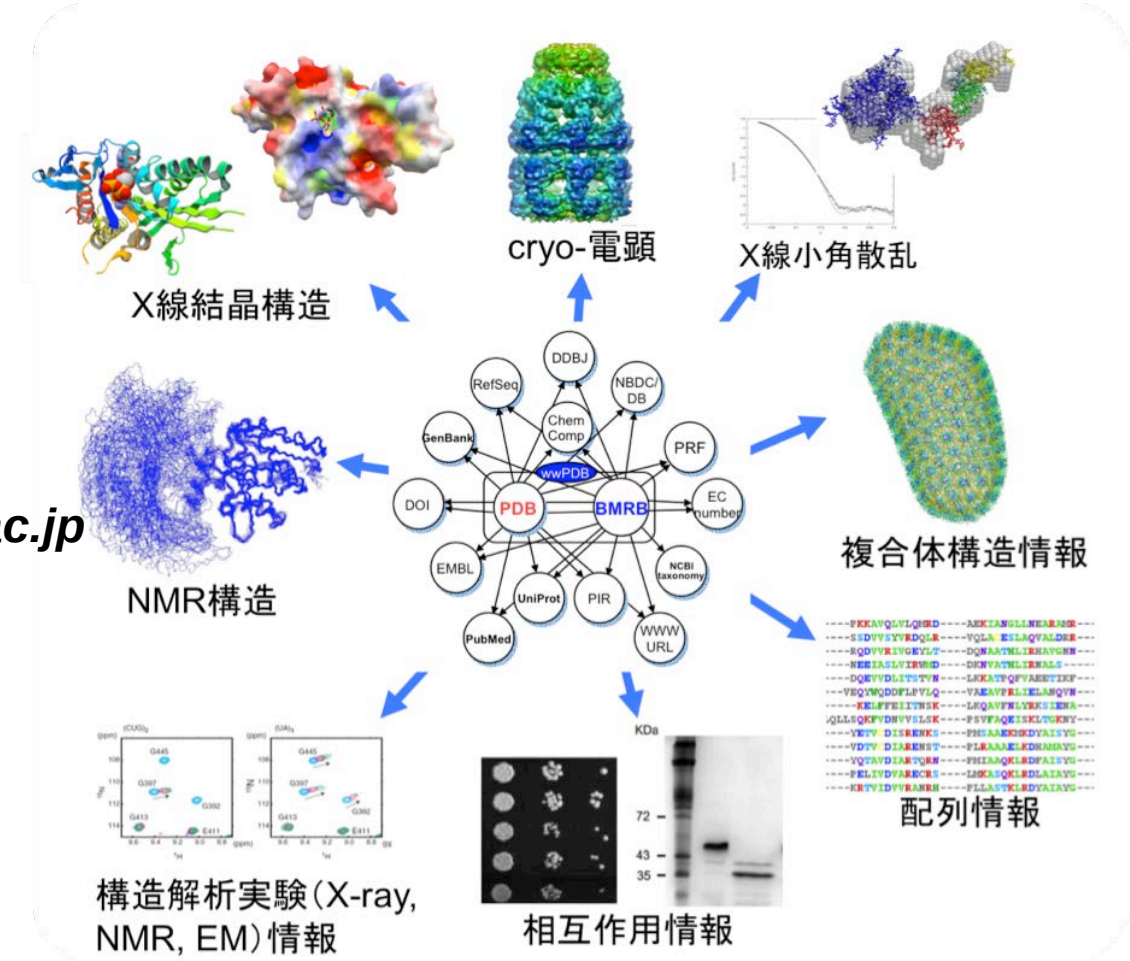
**http://rdf.wwpdb.org/**




## BMRB/RDF

**http://bmrbpub.protein.osaka-u.ac.jp**




```
<rdf:Description rdf:about="http://rdf.wwpdb.org/pdb/1BY4">
  <rdf:type rdf:resource="http://purl.uniprot.org/core/Structure_Resource"/>
  <database rdf:resource="http://purl.uniprot.org/database/PDB"/>
  <method rdf:resource="http://purl.uniprot.org/core/X-Ray_Crystallography"/>
  <resolution rdf:datatype="http://www.w3.org/2001/XMLSchema#float">2.10</resolution>
</rdf:Description>
```

# 第二期プロジェクトの紹介

## 1）蛋白質構造アーカイブの構築・公開

1−1）wwPDBメンバーの一員として、データ登録・編集・品質管理（remediationを含む）の実施

1−2）リガンド・蛋白質複合体結晶解析による医薬スクリーニングデータの精密化・登録支援

## 2）異なる階層のデータと蛋白質構造情報との統合化

2− 1）遺伝情報を活用した遠縁蛋白質検出と蛋白質機能推定サービスの拡張

2− 2）蛋白質複合体に対する多面的解析ツールおよびサービスの開発と公開

2− 3）統合化されたNMR実験データベースを有効利用するためのツール群開発と高度化

## 3)人材養成の実施

# 第二期プロジェクトの紹介

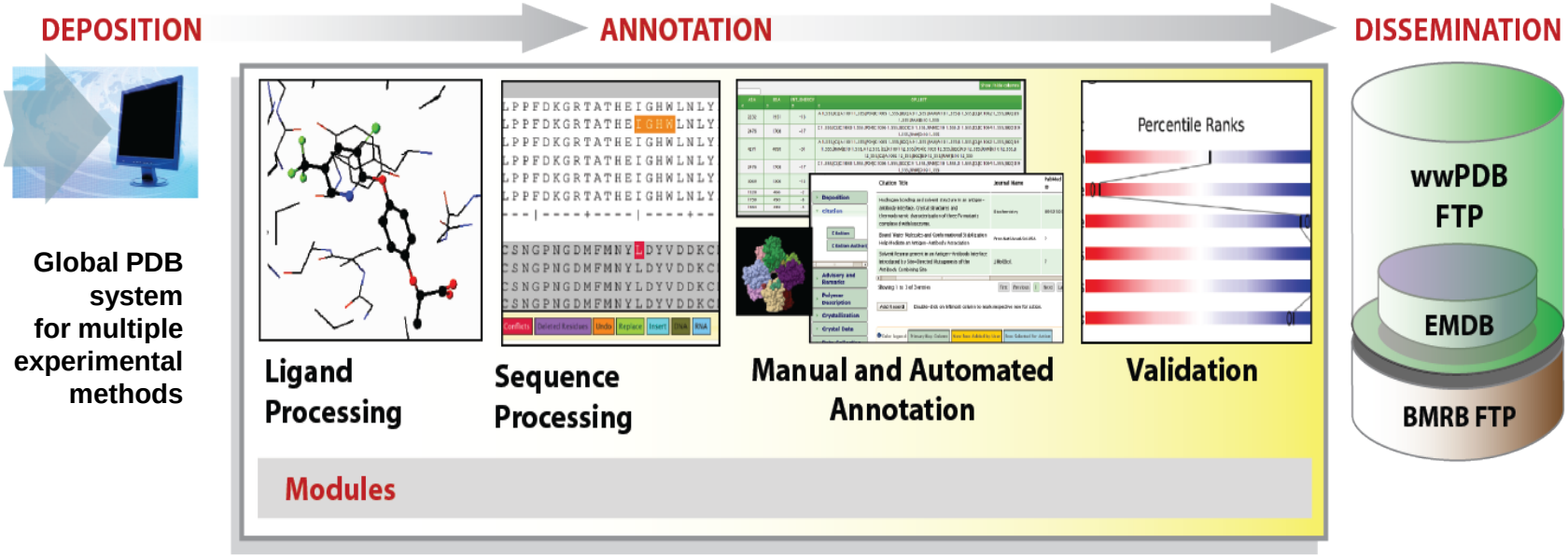
## 1) 蛋白質構造アーカイブの構築・公開

1−1）wwPDBメンバーの一員として、データ登録・編集・品質管理（remediationを含む）の実施

✓ 新たな自動登録・アノテーションシステムの導入と推進

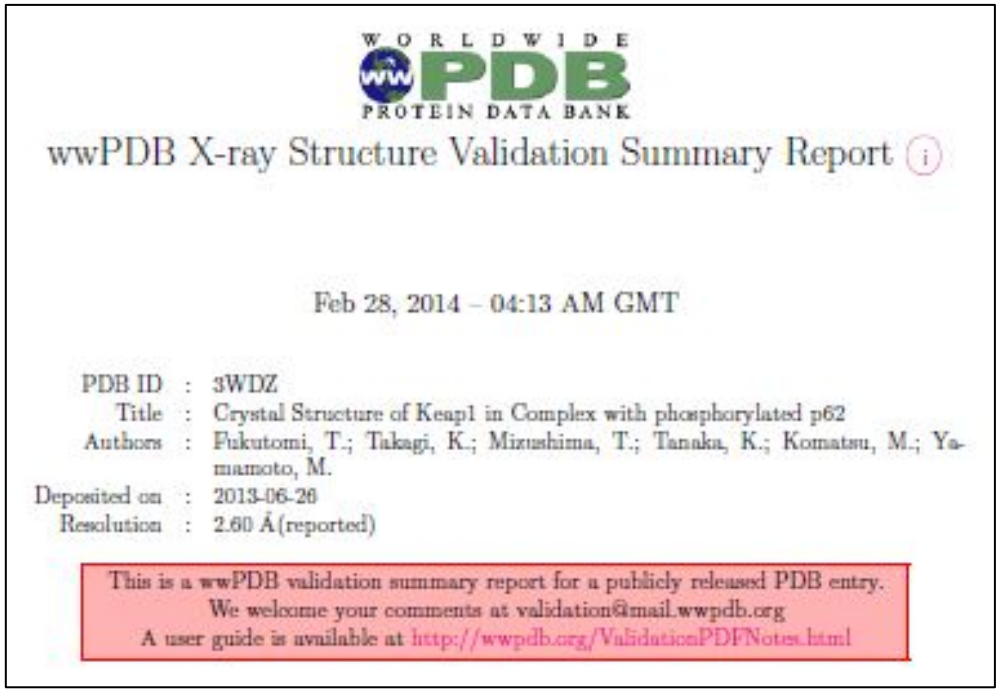
**新しいデータ処理システム**

**データ増大・高品質化に対処**



PDBj Protein Data Bank Japan　WORLDWIDE PDB PROTEIN DATA BANK

# ✓ Validation report（PDF版とXML版）の公開・配布



WORLDWIDE PDB PROTEIN DATA BANK

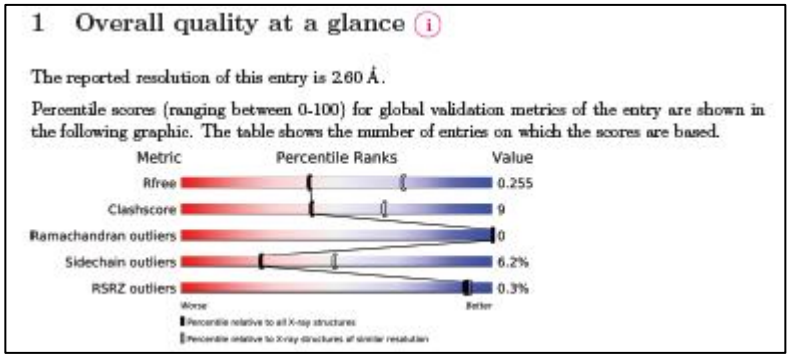
wwPDB X-ray Structure Validation Summary Report

Feb 28, 2014 – 04:13 AM GMT

PDB ID : 3WDZ
Title : Crystal Structure of Keap1 in Complex with phosphorylated p62
Authors : Fukutomi, T.; Takagi, K.; Mizushima, T.; Tanaka, K.; Komatsu, M.; Yamamoto, M.
Deposited on : 2013-06-26
Resolution : 2.60 Å(reported)

This is a wwPDB validation summary report for a publicly released PDB entry.
We welcome your comments at validation@mail.wwpdb.org
A user guide is available at http://wwpdb.org/ValidationPDFNotes.html

Validation reportの表紙

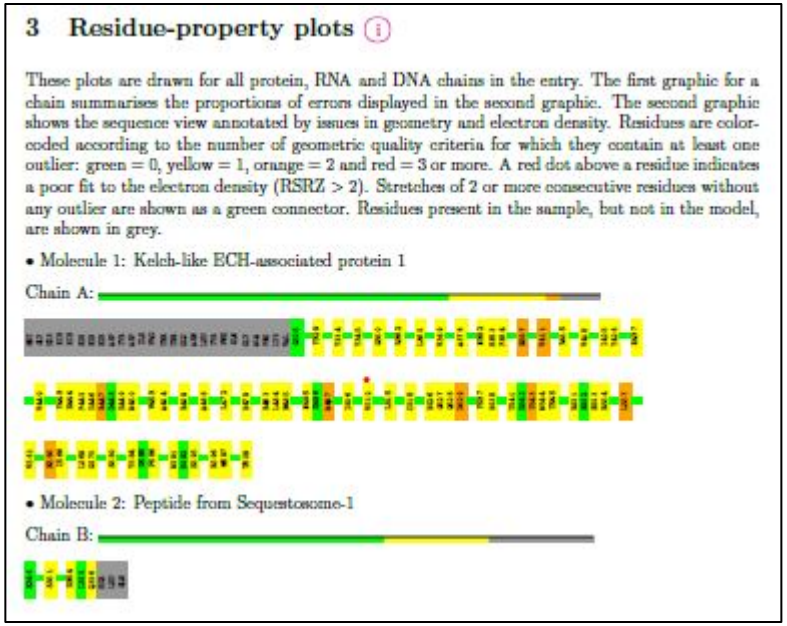


1 Overall quality at a glance

The reported resolution of this entry is 2.60 Å.

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.





3 Residue-property plots

These plots are drawn for all protein, RNA and DNA chains in the entry. The first graphic for a chain summarises the proportions of errors displayed in the second graphic. The second graphic shows the sequence view annotated by issues in geometry and electron density. Residues are color-coded according to the number of geometric quality criteria for which they contain at least one outlier: green = 0, yellow = 1, orange = 2 and red = 3 or more. A red dot above a residue indicates a poor fit to the electron density (RSRZ > 2). Stretches of 2 or more consecutive residues without any outlier are shown as a green connector. Residues present in the sample, but not in the model, are shown in grey.

• Molecule 1: Kelch-like ECH-associated protein 1

Chain A:

• Molecule 2: Peptide from Sequestosome-1

Chain B:

Validation report

PDBj Protein Data Bank Japan　WORLDWIDE PDB PROTEIN DATA BANK

## ✓ 新しいPDBフォーマット（PDBx/mmCIF）への移行：

ユーザに対するサービスサイト、コンバータプログラム、セミナーやワークショップの開催

**PDB**

```
ATOM      1  N   GLN A  39      24.690 -27.754  24.275  1.00 60.76          N
ATOM      2  CA  GLN A  39      23.581 -26.768  24.416  1.00 60.98          C
ATOM      3  C   GLN A  39      23.990 -25.379  23.905  1.00 59.98          C
ATOM      4  O   GLN A  39      25.070 -25.209  23.330  1.00 60.25          O
ATOM      5  CB  GLN A  39      23.136 -26.685  25.878  1.00 60.69          C
ATOM      6  N   VAL A  40      23.115 -24.395  24.122  1.00 59.58          N
ATOM      7  CA  VAL A  40      23.342 -23.010  23.690  1.00 57.26          C
ATOM      8  C   VAL A  40      24.000 -22.152  24.778  1.00 56.00          C
```

**PDBx**

```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.auth_atom_id
_atom_site.type_symbol
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_seq_id
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.pdbx_PDB_model_num
_atom_site.occupancy
_atom_site.pdbx_auth_alt_id
_atom_site.B_iso_or_equiv
ATOM      1  N    N  GLN  A   39   24.690  -27.754   24.275  1  1.00  .  60.76
ATOM      2  CA   C  GLN  A   39   23.581  -26.768   24.416  1  1.00  .  60.98
ATOM      3  C    C  GLN  A   39   23.990  -25.379   23.905  1  1.00  .  59.98
ATOM      4  O    O  GLN  A   39   25.070  -25.209   23.330  1  1.00  .  60.25
ATOM      5  CB   C  GLN  A   39   23.136  -26.685   25.878  1  1.00  .  60.69
ATOM      6  N    N  VAL  A   40   23.115  -24.395   24.122  1  1.00  .  59.58
ATOM      7  CA   C  VAL  A   40   23.342  -23.010   23.690  1  1.00  .  57.26
ATOM      8  C    C  VAL  A   40   24.000  -22.152   24.778  1  1.00  .  56.00
```

# 1−2）リガンド・蛋白質複合体結晶解析による医薬スクリーニングデータの精密化・登録支援

## 蛋白質・リガンド複合体や抗原・抗体複合体構造）における構造精密化と登録作業のパイプライン構築　→　将来ビッグデータに

**問題点：製薬企業において行われている医薬スクリーニングのための蛋白質・リガンド複合体構造データが公開されないまま企業内に死蔵されている。**

**1）知財権の問題**（特に多数のpositiveでない化合物との複合体）

**2）技術面の問題**

・製薬企業ではリガンド周辺のみを決定し蛋白質受容体全体の構造は精密化を行わない。
・positiveな化合物との複合体については構造全体の精密化が行われ、PDBへ登録したり特許申請が行われたりするが、negativeなもの放置されることが多い。

**具現化することの効果：**

1）リガンドと蛋白質とのドッキング過程に対して、**positiveとnegativeの双方の構造がpublicデータとして集積し、ドッキング技術が進む**。（将来はビッグデータ化の可能性もある）

2）パイプラインのシステムを公開することにより、**企業内でパイプラインを利用**することによって、リガンド結合部位を含む全体の構造が迅速に精密化され、スクリーニングが効率的に行われる。

# 1−2)リガンド・蛋白質複合体結晶解析による医薬スクリーニングデータの精密化・登録支援

**蛋白質・リガンド複合体や抗原・抗体複合体構造)における構造精密化と登録作業のパイプライン構築 → 将来ビッグデータに**

**(H26年度:パイプライン構築と試行, H27年度:公開・実施)**

製薬企業等における化合物・蛋白質受容体の複合体結晶構造
:構造因子、アポ体(疑似アポ体)の原子構造

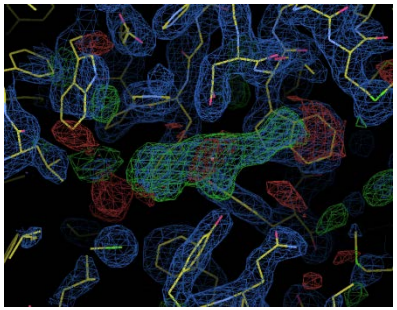
→ 構造精密化 → PDBが許容する精度に

→ PDBへの登録用ファイル → PDBj 登録サイト

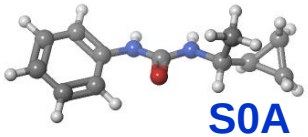
**入力情報**

化合物情報

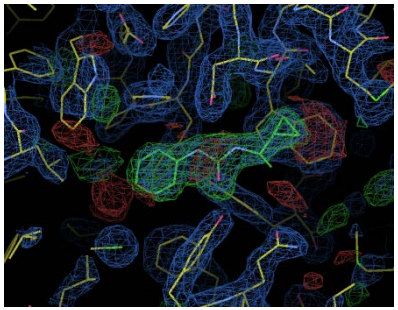
*ChemComp, LigandBOX, PubChem*

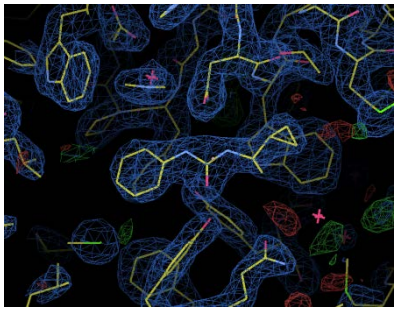


Epoxide hydrolase
(含阻害剤電子密度＋アポ体構造モデル)



S0A

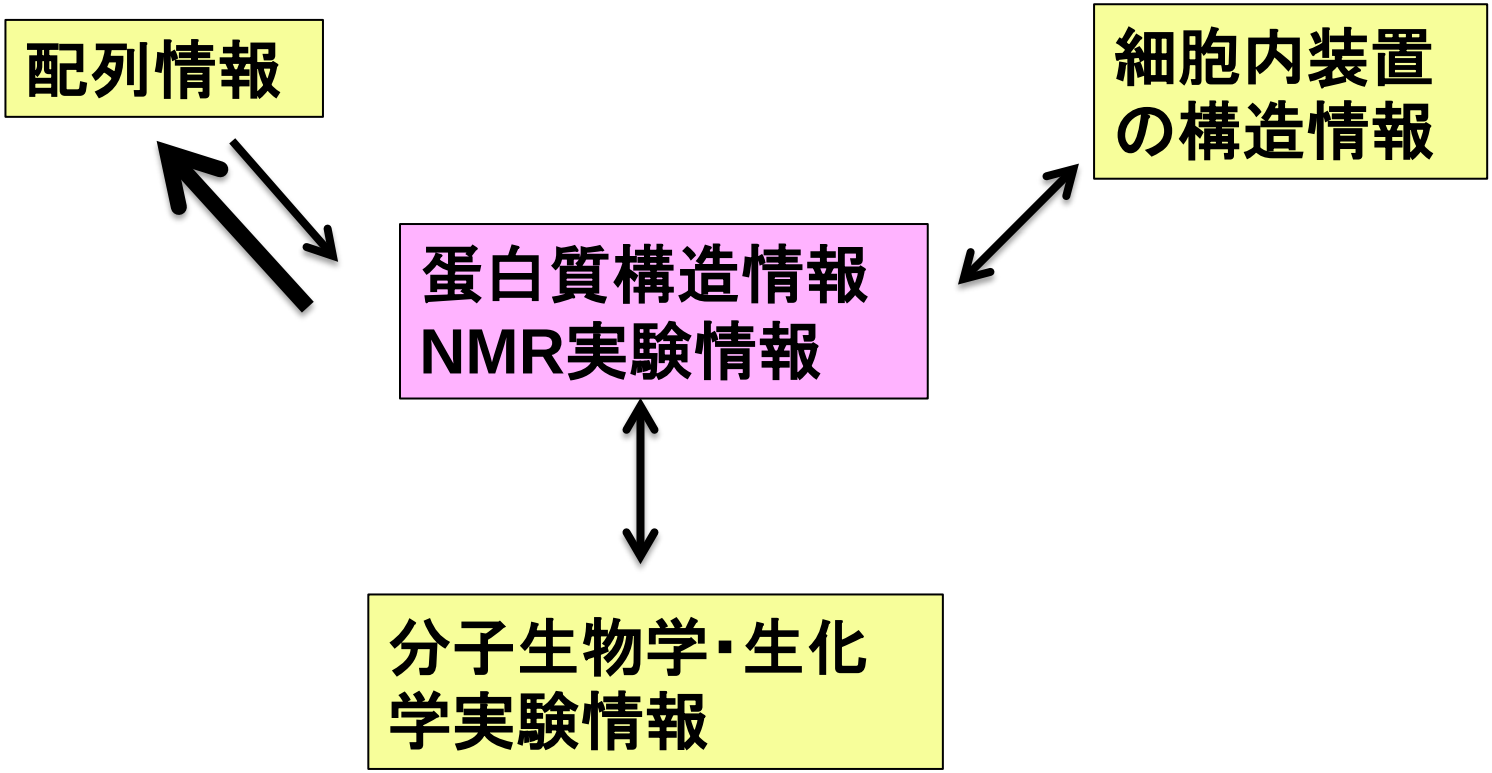




Epoxide hydrolaseの阻害剤(S0A: 1-[(1R)-1-cyclopropylethyl]-3-phenylurea)構造を電子密度にフィットして構造精密化

# 2）異なる階層のデータと蛋白質構造情報との統合化

第１期：世界標準のRDFフォーマット（PDB/RDF, BMRB/RDF）
　　など技術的基盤を開発・確立
第２期：コンテンツ・レベルでの階層間の統合化



# 2）異なる階層のデータと蛋白質構造情報との統合化

第１期：世界標準のRDFフォーマット（PDB/RDF, BMRB/RDF）
　　など技術的基盤を開発・確立
第２期：コンテンツ・レベルでの階層間の統合化

配列情報

配列→機能の高度化
MAFFTash, DASH, SFAS

細胞内装置の構造情報

蛋白質構造情報
NMR実験情報

分子生物学・生化学実験情報

統合化の障害
①PDB構造の
　Redundancy
②PDB中のアミノ酸番号
③構造解析のための
　多数の変異
④非天然のﾘｶﾞﾝﾄﾞ

# 2）異なる階層のデータと蛋白質構造情報との統合化

配列情報

細胞内装置の構造情報

蛋白質構造情報
NMR実験情報

分子生物学・生化学実験情報

構造生命科学の課題の一つ

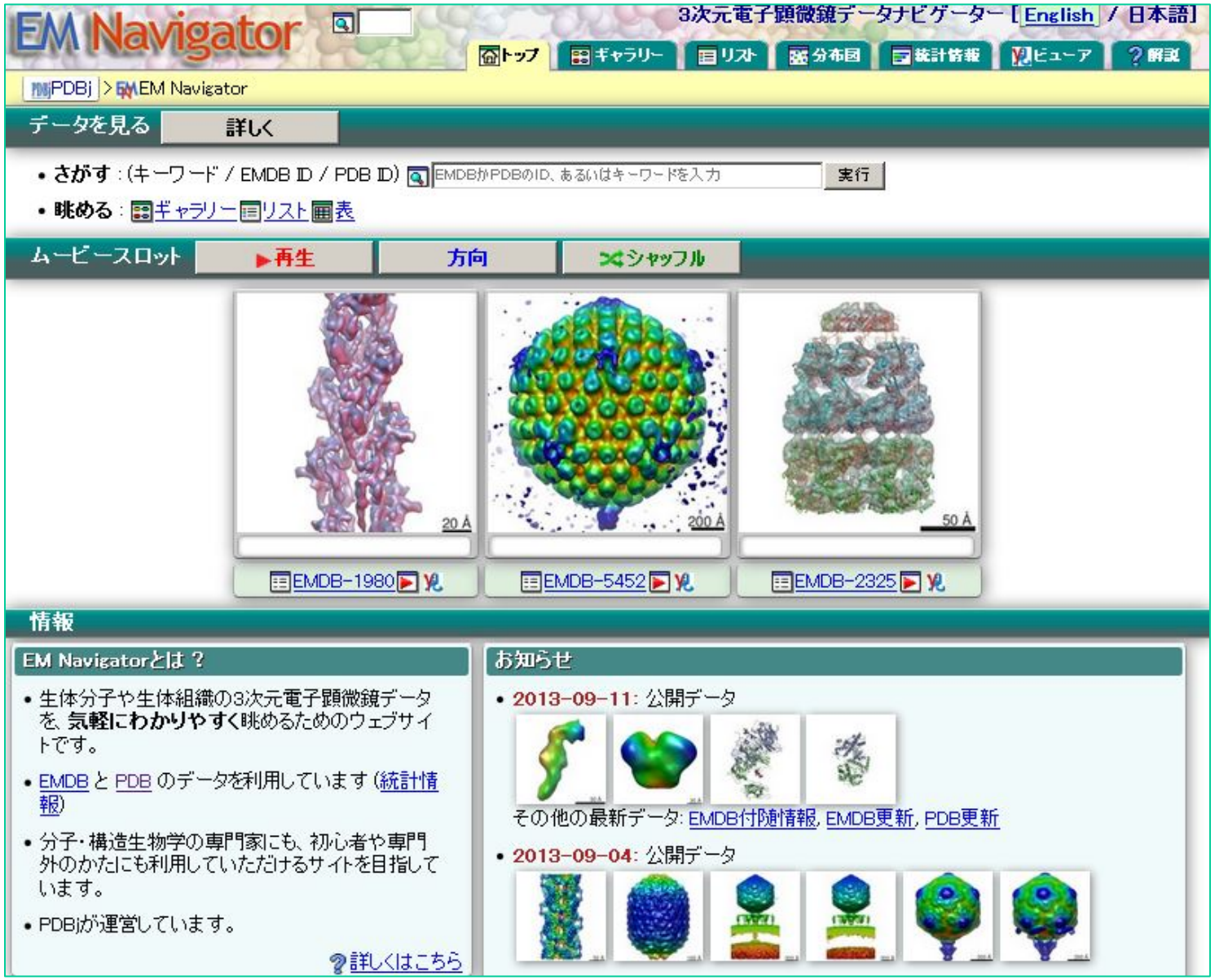
①Large structures in PDB
②EMDB, SSBD（統合生命動態ﾃﾞｰﾀﾍﾞｰｽ）との統合化
③EM画像同士、EM画像と原子構造の形の類似性検索
④eF-site (静電的分子表面) for large structures
⑤ProMode (基準振動ダイナミクス) for large structures

# ①Large structures in PDB

HIV-1 Capsid 3J3Q

- 1356 chains
- >2M atoms
- 25 – PDB format entries

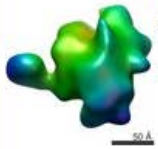
# ②EMDB，SSBD（統合生命動態ﾃﾞｰﾀﾍﾞｰｽ）との統合化



# ③EM画像同士、EM画像と原子構造の形の類似性検索

**Query: human RNA polymerase II with RNA (EMDB: 2189)**
**Similar str. from 185,558 images**

**Subject structure**

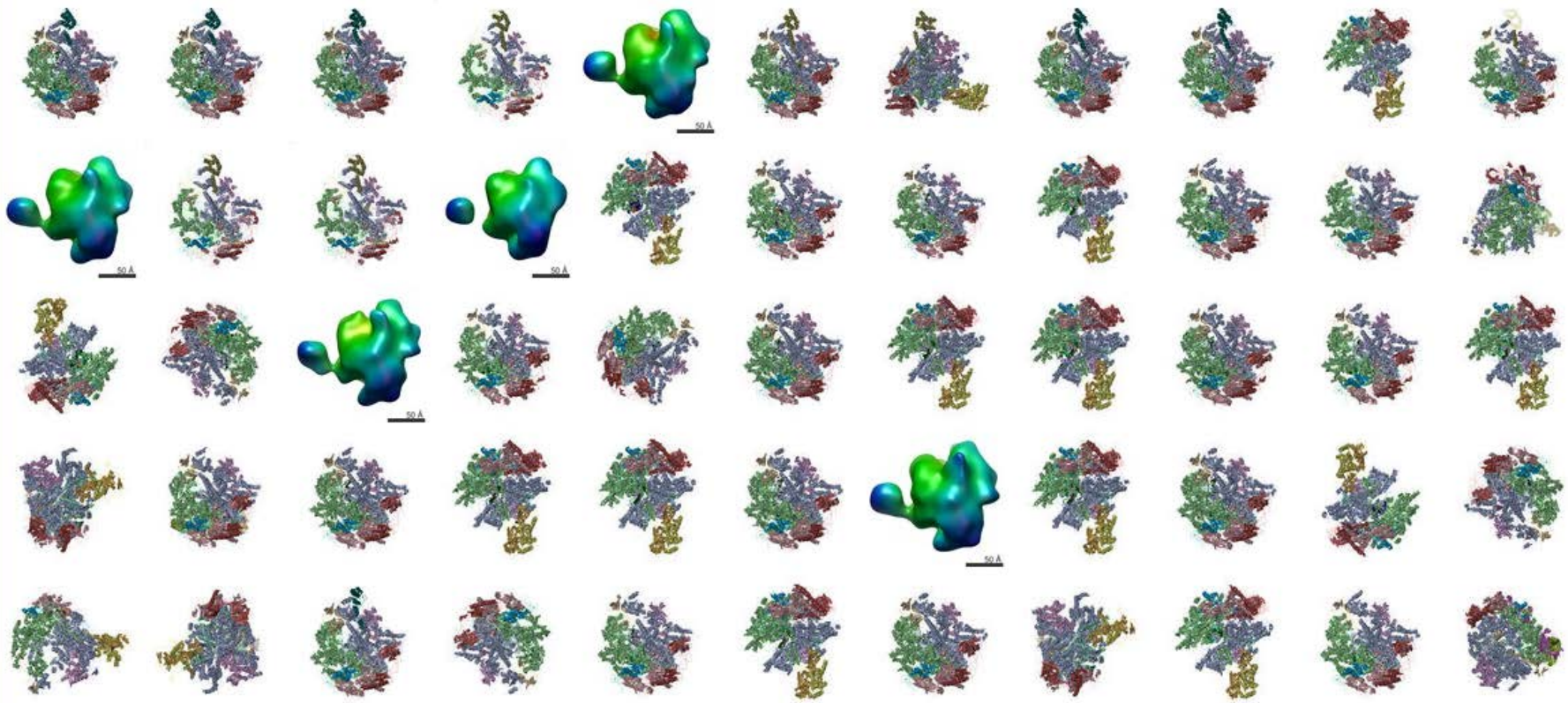


**EMDB-ID**: 2189
human RNA polymerase II in complex with Alu RNA
Yorodumi

**Search Result** [Display as list]

3000 similar structures found from all 185558 (filter level: 0.5)
showing 1 - 100
Pages: [1] 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

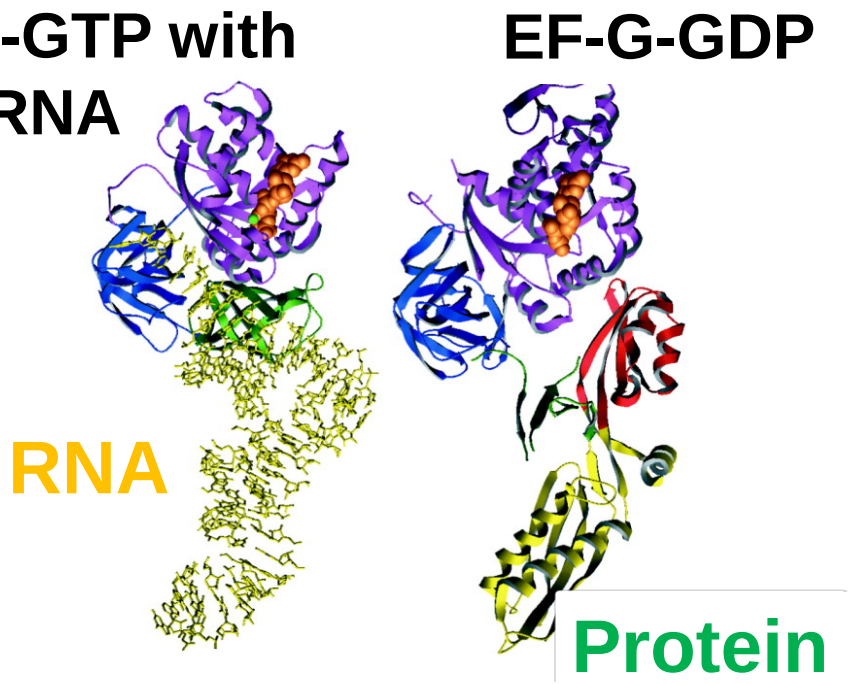
>100 RNA polymerase II found



PDBj Protein Data Bank Japan  WORLDWIDE PDB PROTEIN DATA BANK

# Molecular mimicryの解析

分子擬態：R. Green, *Curr Biol*., 2000



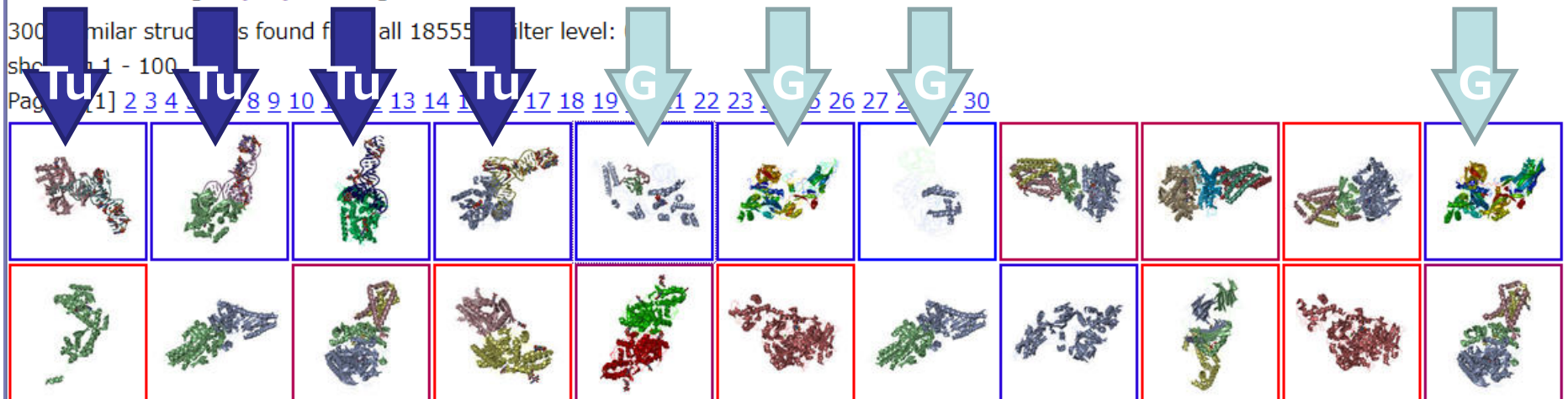

## Query: EF-Tu-GTP with Phe-tRNA

**Subject structure**

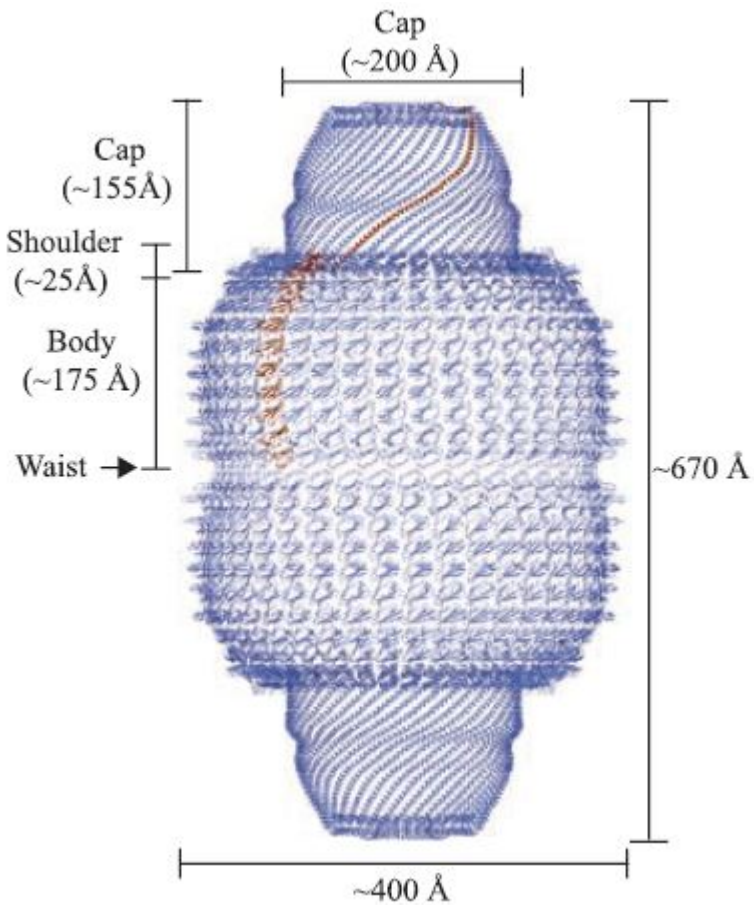
**PDB-ID**: 1ob2, **Assembly**: deposited form
E. coli elongation factor EF-Tu complexed with the antibiotic kirromycin, a GTP analog, and Phe-tRNA
Yorodumi

**Search Result** [Display as list]

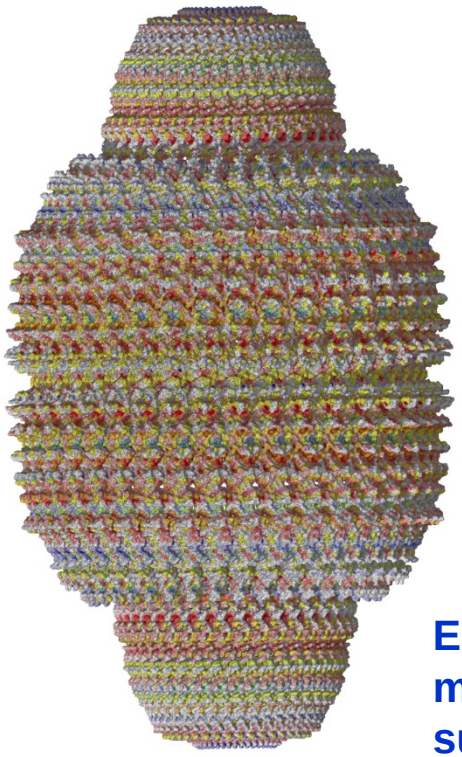
300 similar structures found from all 18555 filter level:
show 1 - 100
Page [1] 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30



# ④eF-site (静電的分子表面) for large structures
# ⑤ProMode (基準振動ダイナミクス) for large structures




**Connolly Surface / Poisson-Boltzmann eq.**

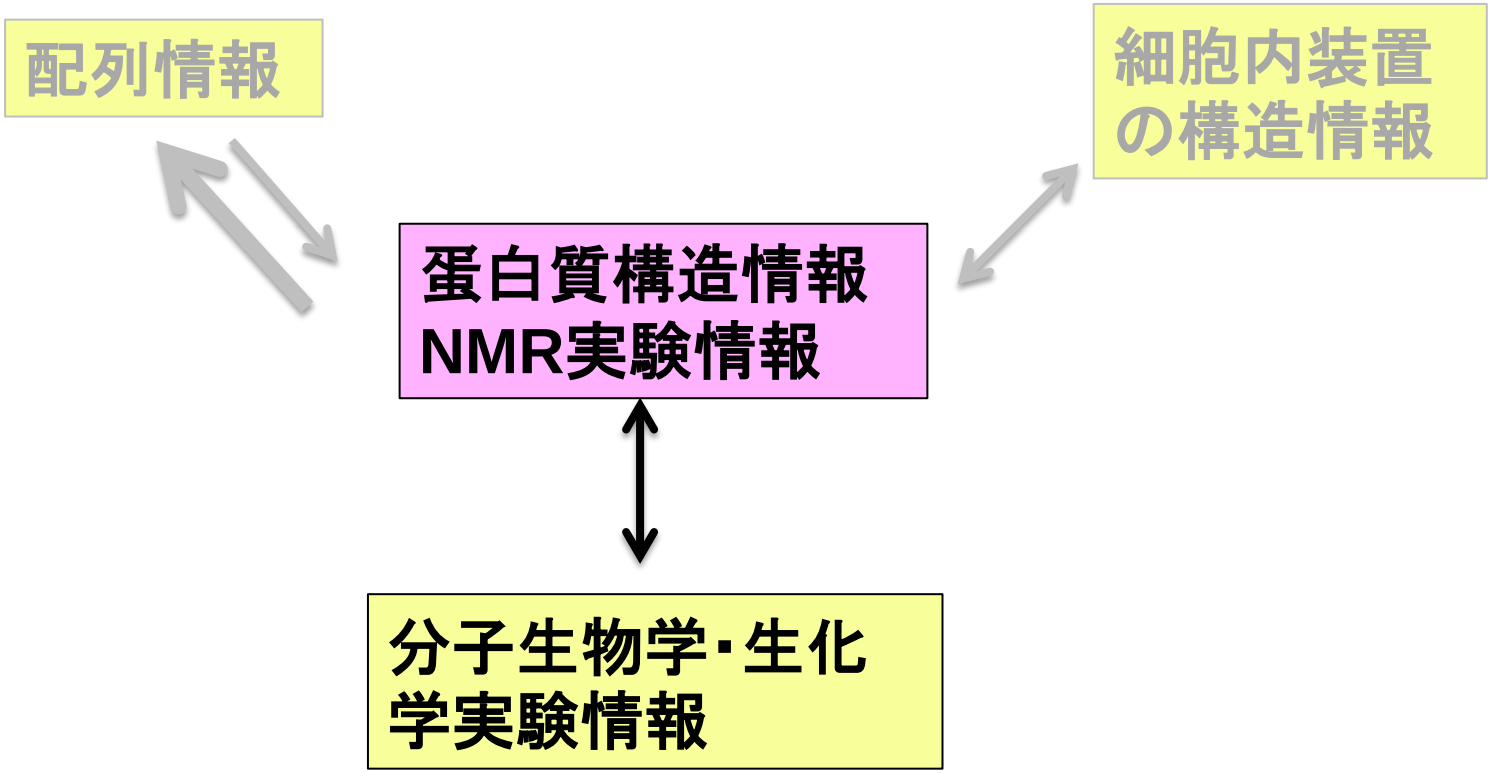


**Electrostatic molecular surface**

**Vault (2zuo, 2zv4, 2zv5) : more than 10,000,000 Da (78 chains)**

*Tanaka et al. (2009) Science, 323. 384-388*

# 2）異なる階層のデータと蛋白質構造情報との統合化




解析対象分子の配列相同性情報や、一塩基多型 (SNP: OMIM)、Yeast-2-hybridやpull-down assayなどの相互作用情報（UniProt, IntAct）、疾患関連情報(CTD)、文献情報(PubMed)などを統合的に検索・表示できるWebツールの開発

# 3）人材養成

## 3−1）構造解析研究者・DB利用者・アノテータへの教育・情報提供

- ✓ 構造解析研究者への情報提供・交流（日本結晶学会、アジア結晶学会、NMR討論会等）
- ✓ 蛋白質構造を利用する研究者への情報提供・交流（日本バイオインフォマティクス学会、CBI学会、日本分子生物学会、日本生物物理学会、蛋白質科学会、アジアオセアニア蛋白質連合等）
- ✓ wwPDBメンバーとのタイトな連携・交流によるアノテータおよび技術者・研究者のon-the-job-training

2013年11月にEBIで行われたNMR restraints に関するﾜｰｸｼｮｯﾌﾟ

2013年11月にEBIで行われたPDBx/mmCIFのﾜｰｸｼｮｯﾌﾟ

PDBj Protein Data Bank Japan　WORLDWIDE PDB PROTEIN DATA BANK

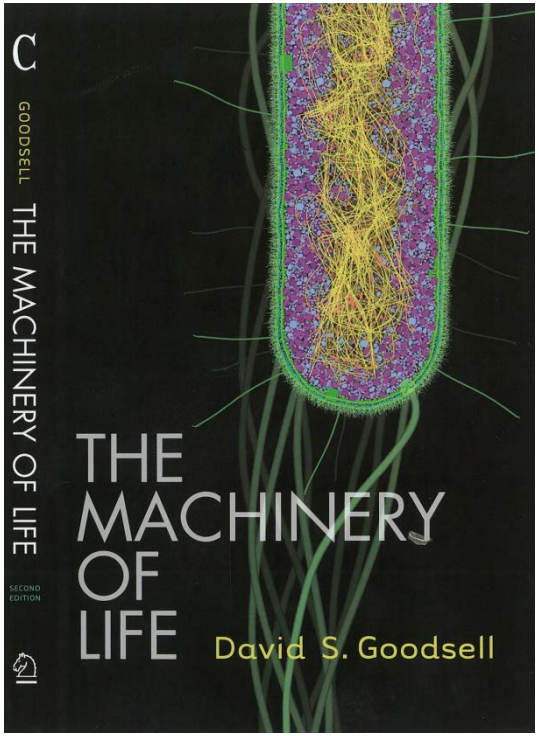
## 3−2）一般社会へのアウトリーチ活動

✓ 一般社会人・小学生以上の学生に対するサイエンスアゴラ（東京お台場）(A)などの展示や一般向け講習会 教材の出版

(B) 見てわかる構造生命科学（化学同人：2014年4月刊）

(C) The Machinery of Life (by David S. Goodsell)
（第二版の日本語訳：2014年出版予定）

(A)2013年11月サイエンスアゴラ
（東京・日本未来科学館）

(B)

(C)