

蛋白質構造データバンクのデータ検証 高度化と統合化

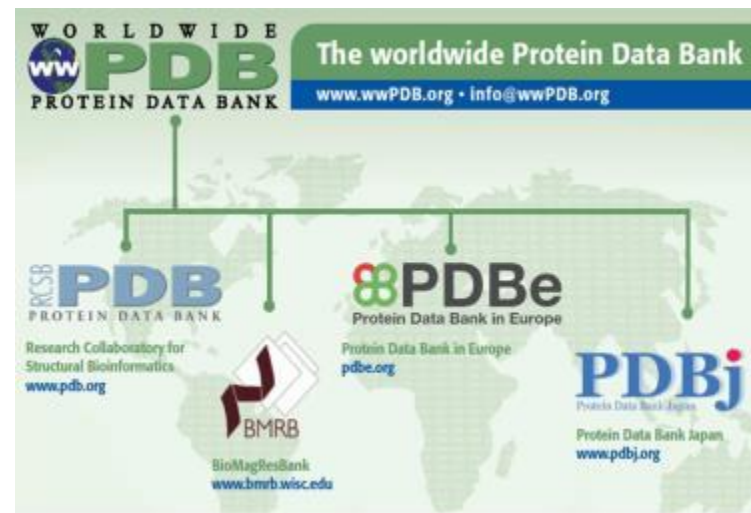
栗栖源嗣¹, 藤博幸²

1)大阪大学蛋白質研究所, 2)関西学院大学理工学部

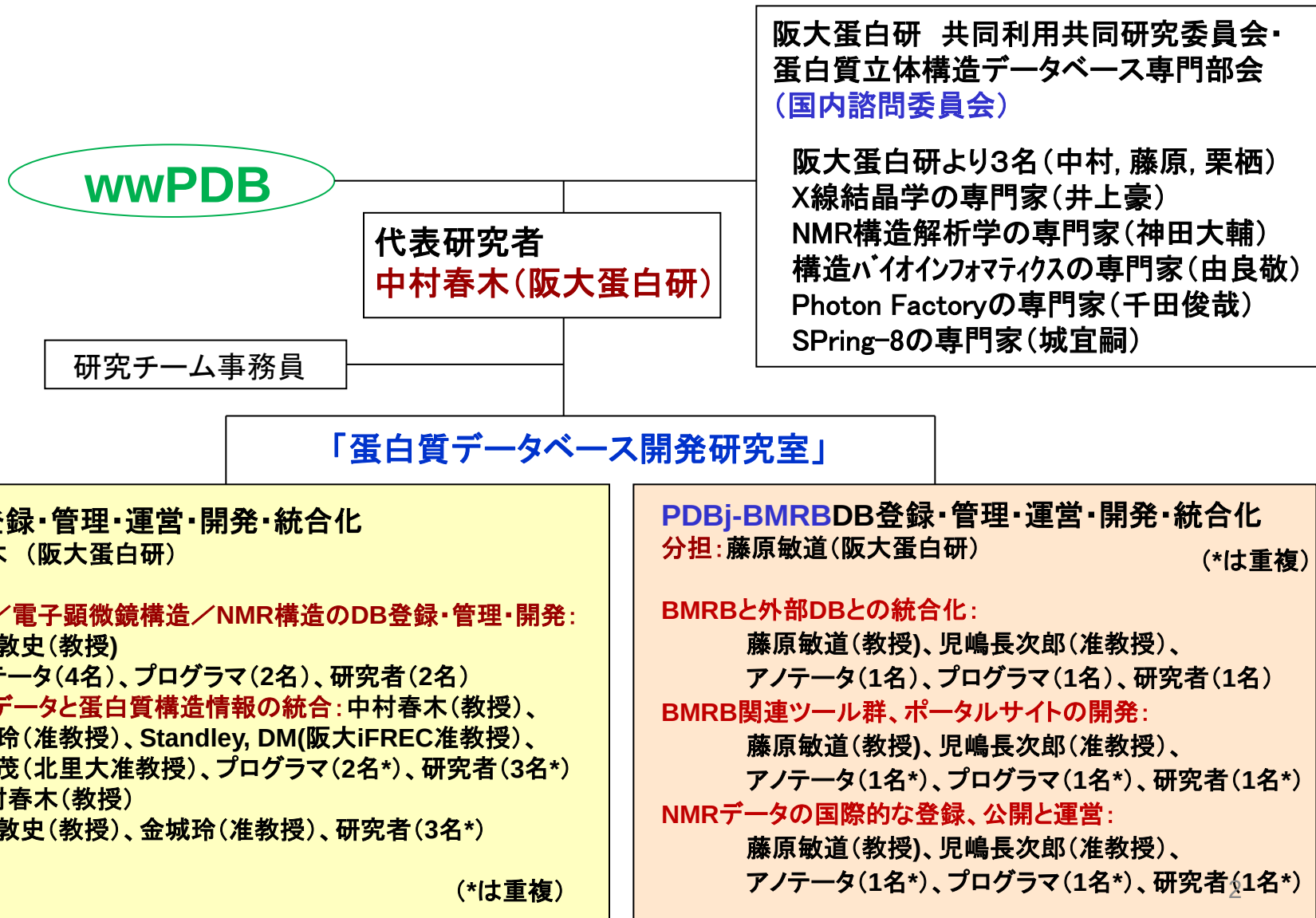


生体高分子の3次元構造(形) に関する情報を集めた国際的 データベース

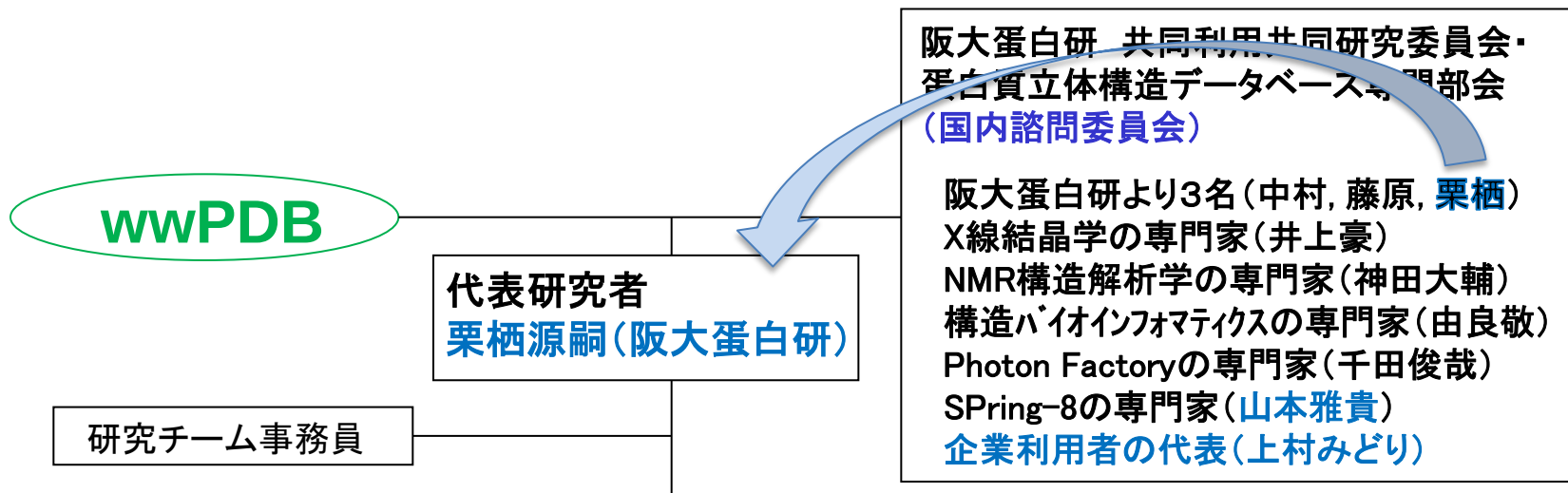
1971年からのデータが集積され、情報は無償で利用できる。運営は各国(米国, 欧州, 日本)の政府機関による研究費用でまかなわれている。
2003年からは国際組織wwPDBとして活動している(PDBjは創立メンバー)。



第二期までのPDBj運営体制(～2017.3)



今期のPDBj運営体制(2017.4～)



「蛋白質データベース開発研究室」

PDBj DB登録・管理・運営・開発・統合化

代表: 栗栖源嗣 (阪大蛋白研)

(*は重複)

PDBアーカイブの構築・検証: 中川敦史(教授),
アノテータ(4名), プログラマ(2名*), 研究者(2名*)

BMRBアーカイブの構築・統合化: 藤原敏道(教授),
宮ノ入洋平(准教授), アノテータ(1名), プログラマ(1名),
研究者(1名*)

セマンティック化と構造データのアノテーション: 金城玲(准教授)
プログラマ(2名*), 研究者(3名*)

国際組織(wwPDB)との連携人材養成: 中村春木(教授)
栗栖源嗣(教授), 中川敦史(教授), 金城玲(准教授)

配列・構造・機能解析高度化と可視化ツールの開発
分担: 藤博幸(関西学院大理工)

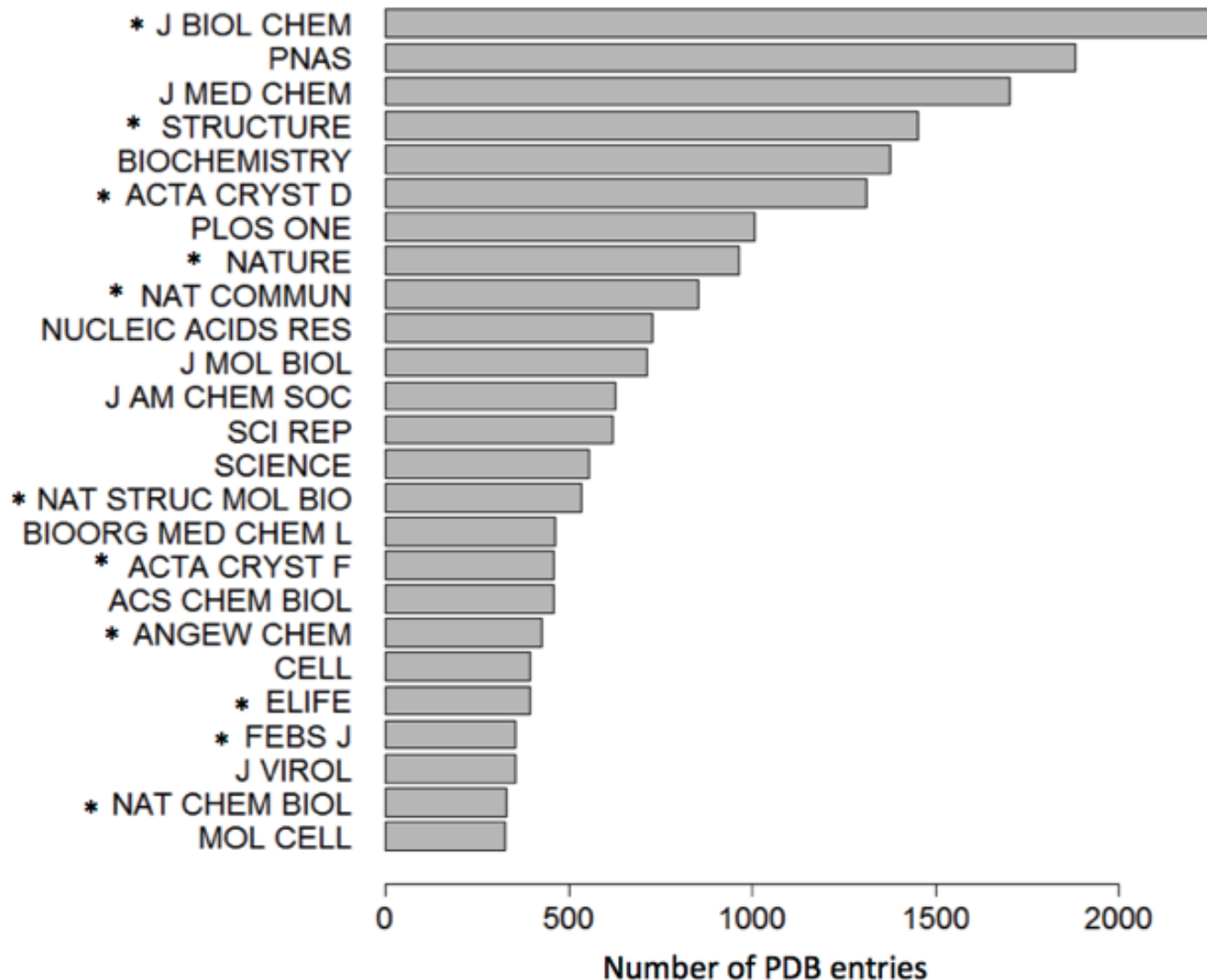
ASHビューアの開発: 藤博幸(教授),
研究者(2名)

蛋白質構造データバンクの データ検証高度化と統合化

- 1) wwPDBメンバーとしてのPDBアーカイブ構築・検証・公開
- 2) 他のデータベースとの統合化および高度化
 - 2-1) CSDとの統合化
 - 2-2) ASHビューアの開発→関西学院大学藤教授より
 - 2-3) アミノ酸の立体構造中での役割の自動アノテーション
 - 2-4) 統合化されたデータベースおよび関連ツール開発
- 3) データベースの利用促進・人材育成
 - 3-1) 利用促進・アノテータの育成
 - 3-2) 国際協力
- 4) 達成達成目標と選考時に頂いた意見への対応

構造生物学領域の基盤データとして貢献

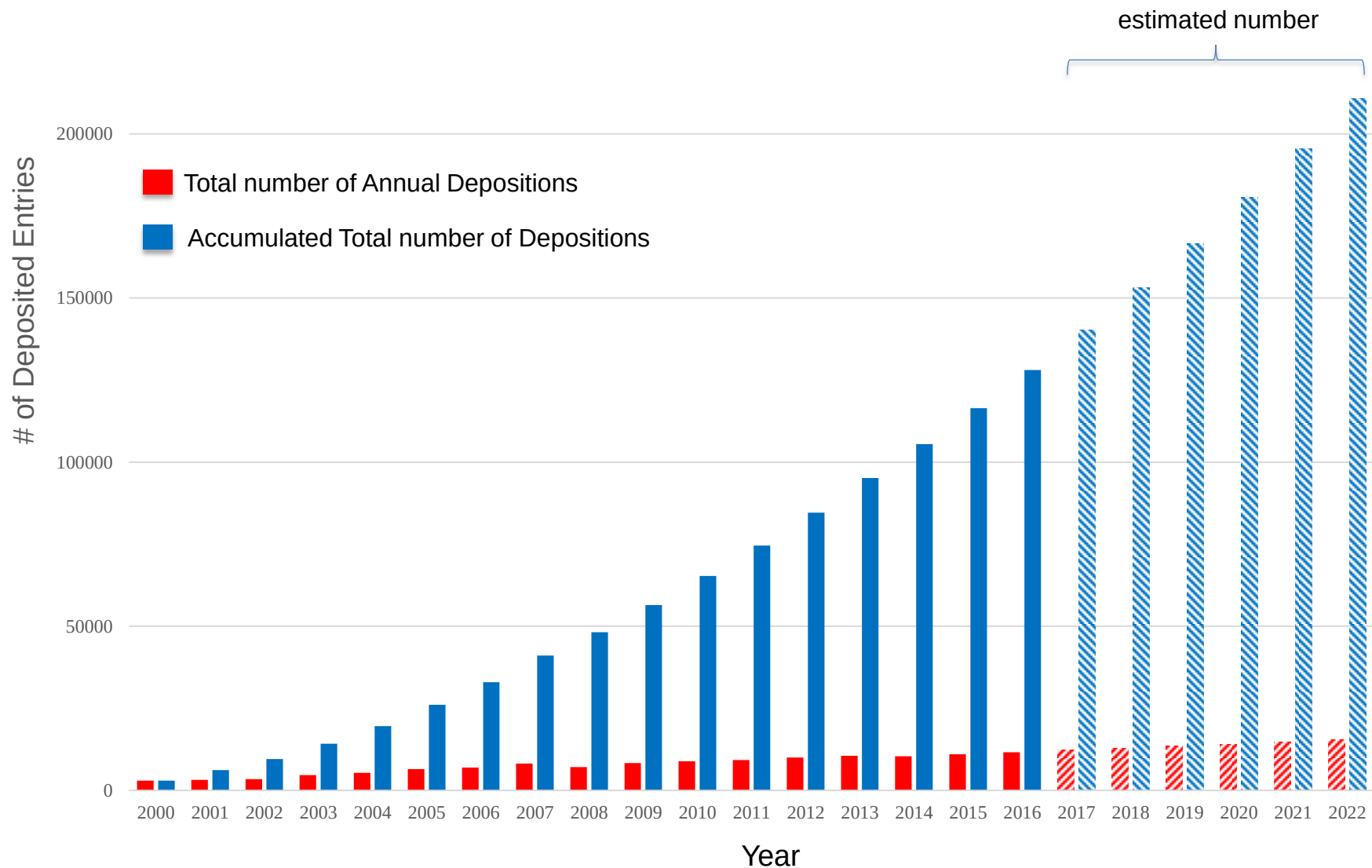
2012-2016



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(毎週アップデート)の運営
 - 関連DBとの統合化や二次データベース・ツールの開発

増え続けるPDB登録数



BMRBは構造以外の情報も含む

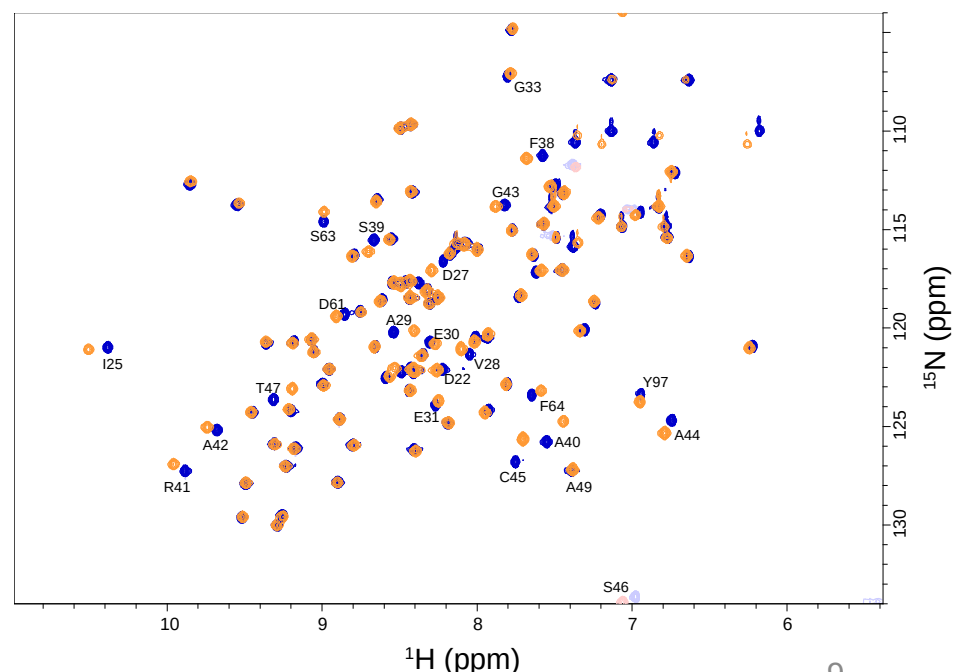
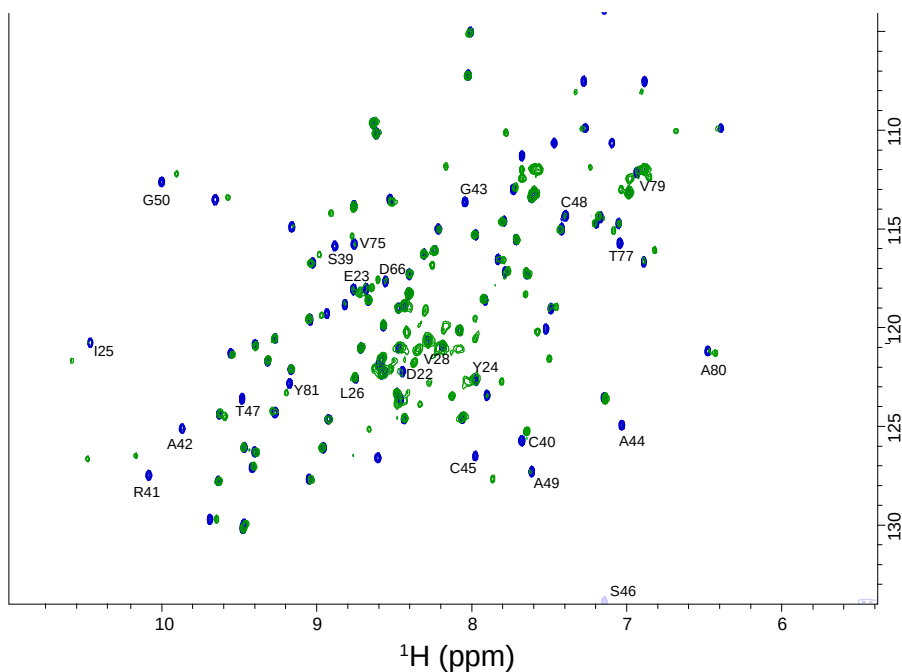
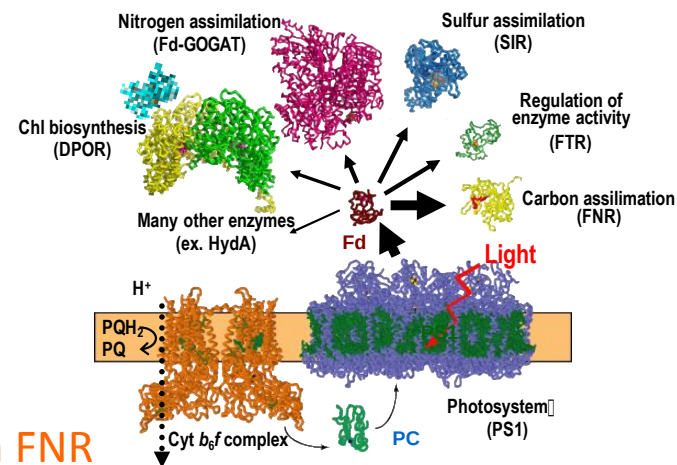
Year	NMRによって 決定された 溶液構造の数	構造情報を伴わな い化学シフトの登録 (揺らぎや相互作用 を示すデータ)	Total
2013	486	308	714
2014	506	240	746
2015	427	333	760
2016	507	234	741

化学シフトのみの登録例(1)

BMRB ID:11596

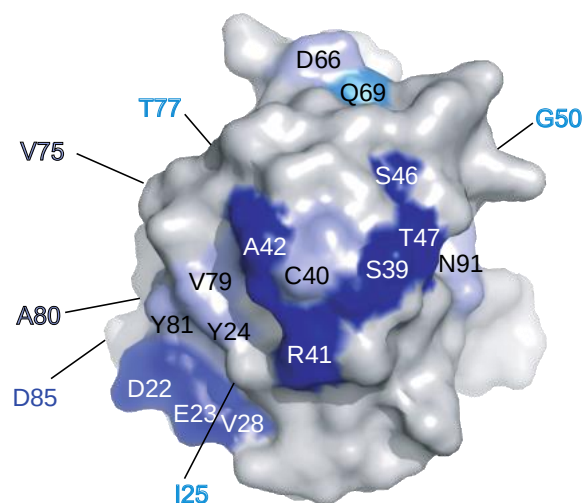
Green: with PSI
Blue: without PSI

Orange: with FNR
Blue: without FNR

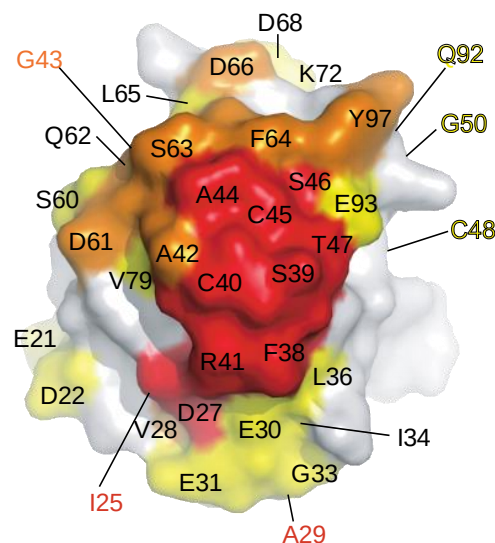
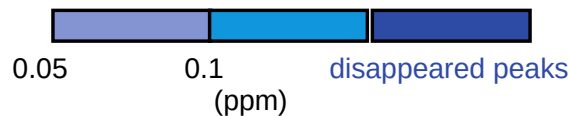


化学シフトのみの登録例(1)

PDB ID:5AUIにマッピング



T. elongatus Fd with PSI



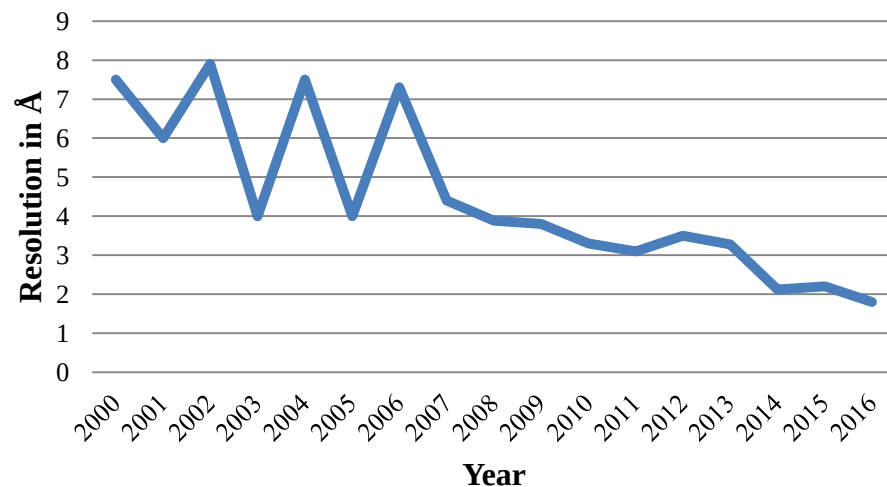
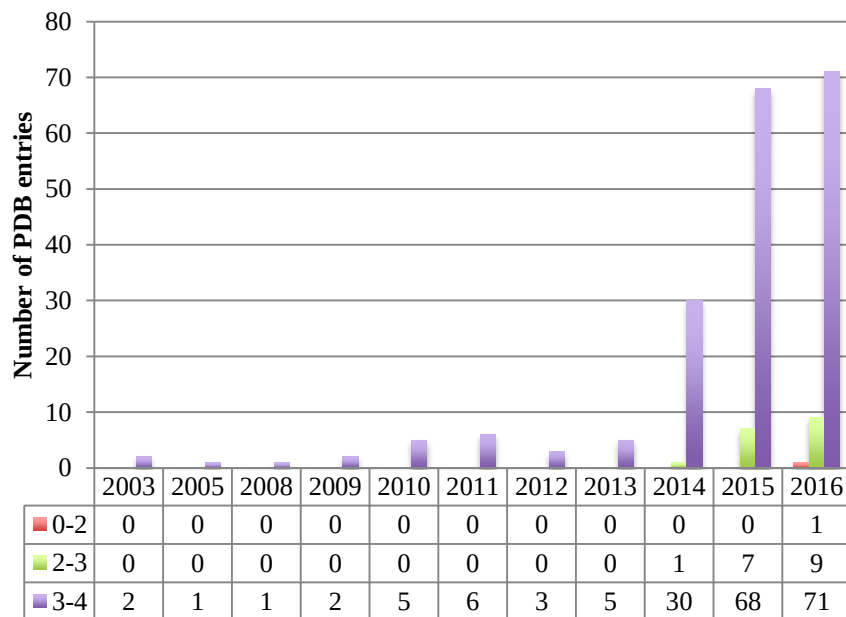
T. elongatus Fd with FNR_S



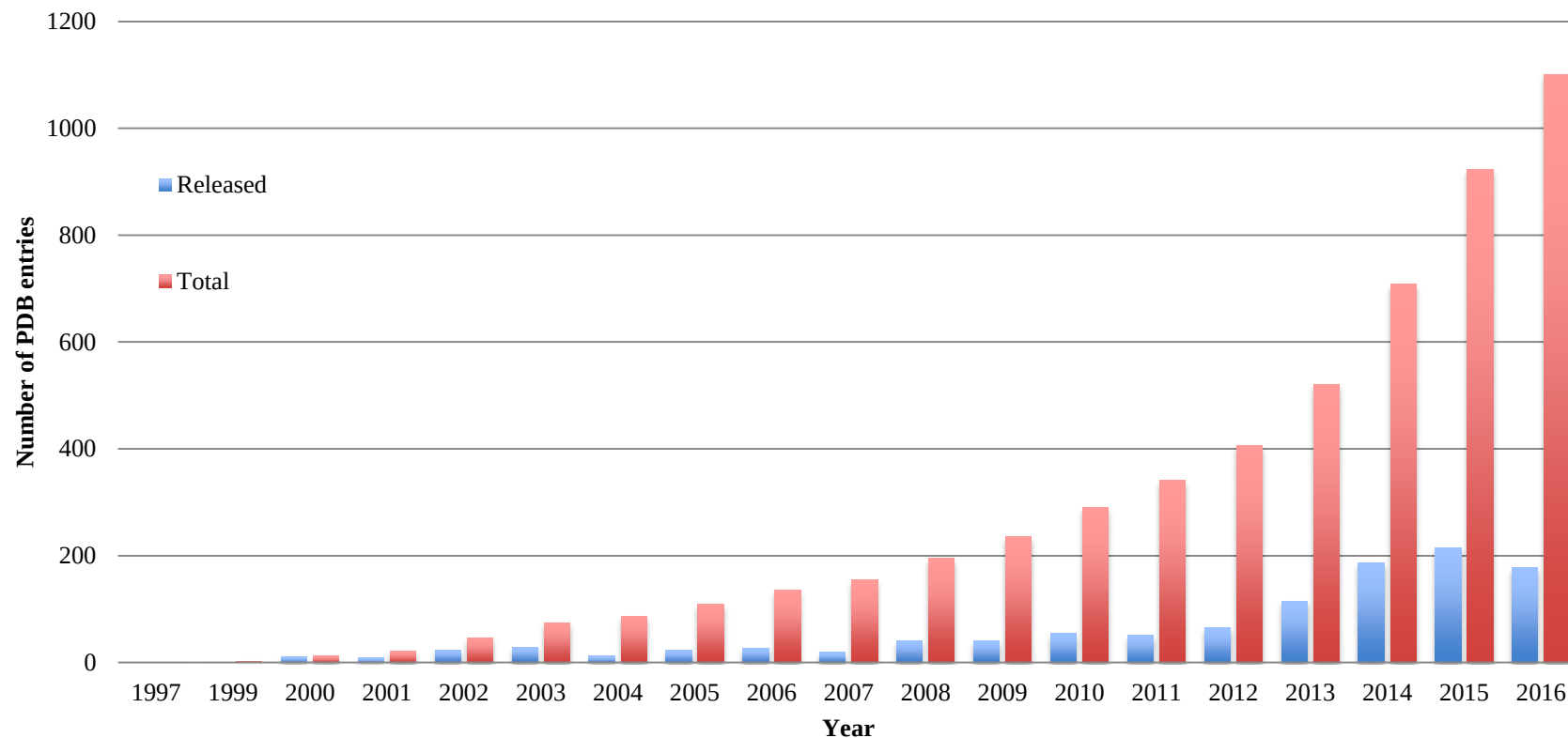
Mutoh, *et al*, *Biochem.* 2015

クライオ電子顕微鏡(3DEM)の進展

- 1.8Å structure in 2016
(PDB ID 5K12; EMD-8194)
- Increasing number of 3DEM structures at 2-4Å resolution
(75 in calendar 2015 and 80 in first 7 months of 2016)



3DEM Entriesの増加

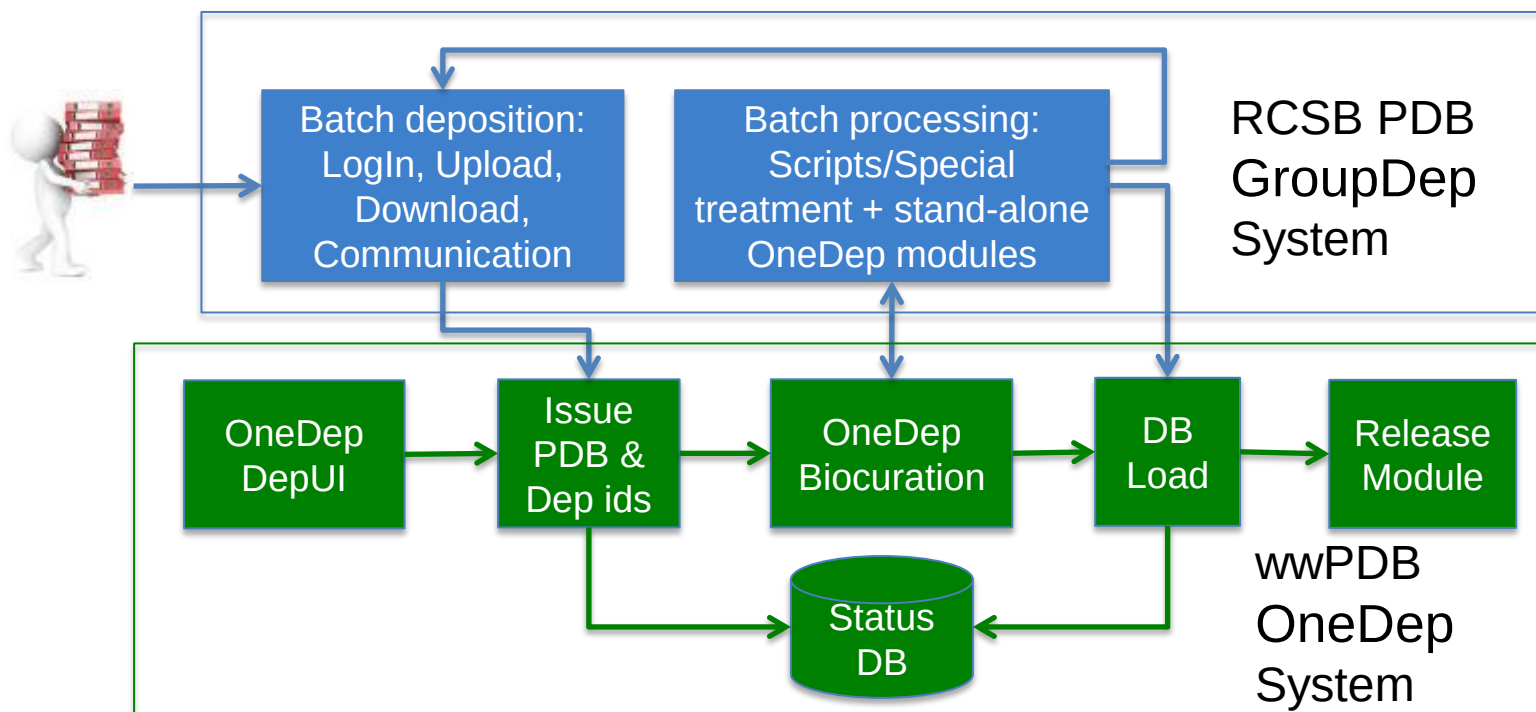


As of August 1, 2016, >1100 EM entries in the PDB archive

178 new entries released Jan 1 - Aug 1, 2016

企業ユーザーからのバッチ登録にも対応へ

- Group Deposition processing
 - Requirements set by wwPDB OneDep Team
 - Provided support for D3R Blind Challenges



PDB IDのバージョン化

- Enable revisions to entries updated by the Depositor of Record (e.g., Version 1-0 → 1-1; 1-0 → 2-0)
 - wwPDB will NOT assign a new PDB ID going forward (for Depositor of Record revision only)
- Introduce new PDB ID code format
 - Allow more informative and transparent delivery of revised data files
 - With PDB prefix and extension of 4 characters (e.g., from “1ABC” to “PDB_00001ABC”)
- Example: PDB_00001ABC_XYZ_V2-2.cif.gz

共通登録サイト導入による効率化

- wwPDB datacenterの共通登録サイト
 - X線, NMR, 電子顕微鏡の各手法で決定した座標を全て取り扱う
 - 実験データ(構造因子, 化学シフトと距離束縛情報, EMマップ)の登録も行う

wwPDB OneDep System

Existing deposition

Deposition ID

Password

Log in

Forgot Password

wwPDB regions

Start a new deposition

Welcome to the wwPDB OneDep system.

To continue with an existing deposition, please login on the left.
Please note that un-submitted sessions will expire 3 months after last login. Un-submitted sessions and uploaded files will be removed once they expire.

If you have any feedback, please write to us at deposit-help@mail.wwpdb.org
At this time this deposition system does not work with Internet Explorer versions 8 or less.

Warning: Please note the current system does not support multiple simultaneous depositions.

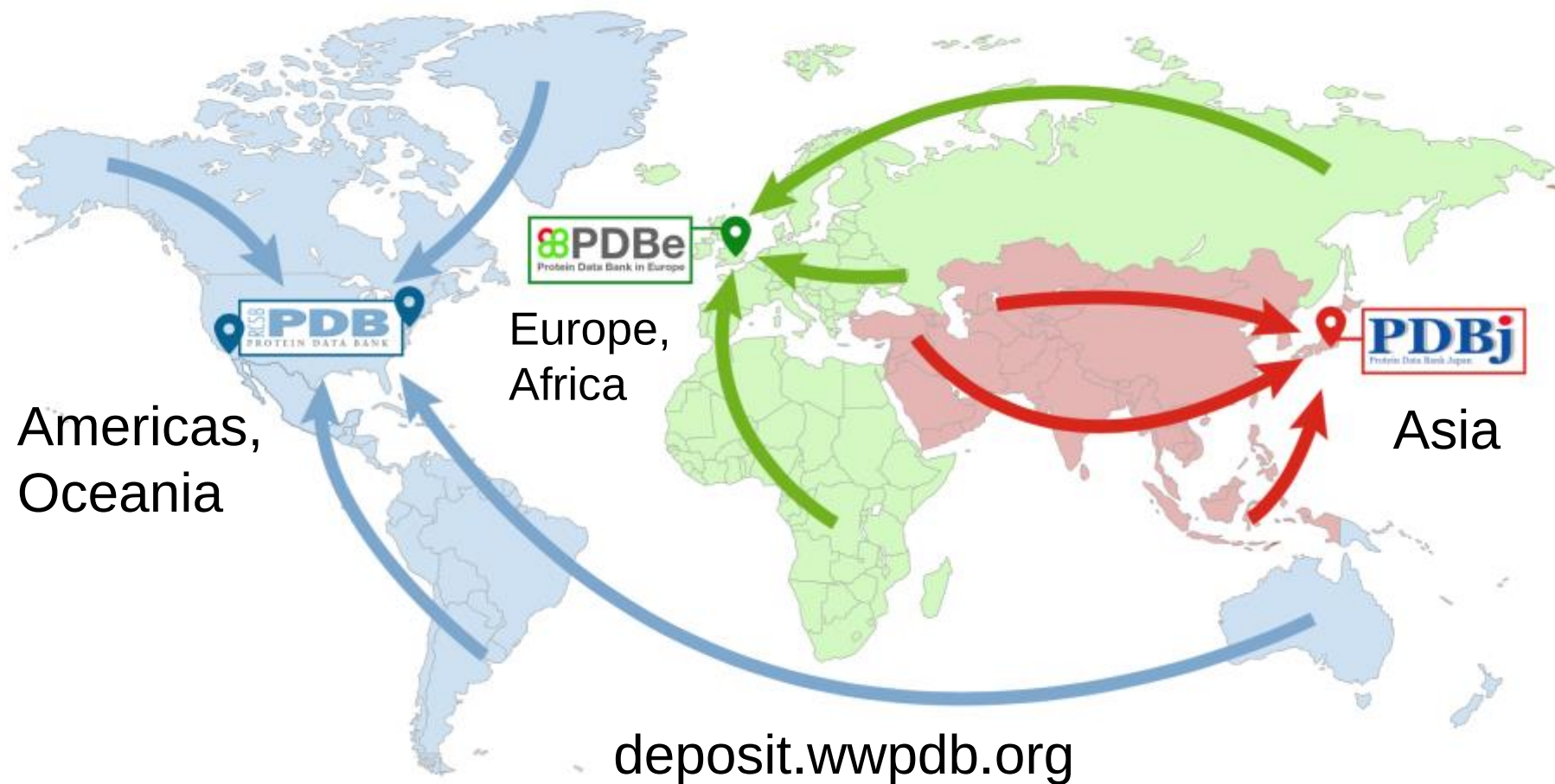
<http://deposit.wwpdb.org/deposition/>

Please select the location of the institute of your PI.
This will automatically direct to the closest wwPDB data center (RCSB PDB/US, PDBe/UK, or PDBj/Japan) for faster response times for communication and computation.

Country:

- ✓ Select...
- United Kingdom
- United States
- Japan**
- Afghanistan
- Aland Islands

PDBjは日本を中心にアジア地区 からのデータ登録に責任を持つ



引き続き、アジア地区のデータ登録とデータ検証に責任を持つ。

ORCID IDの収集

- Successfully Implemented Apr 11, 2016
- Metrics (Apr 11 – Aug 31, 2016):
 - ~8% of Depositions have ORCID ID (374/4713)
 - 170 unique ORCID IDs (92 identified as PIs)
- Plans to Increase ORCID Adoption
 - Expand to all entry authors to provide ORCID (2017)
 - Distribute collected ORCID IDs at ftp archive (2017)
 - Mandatory going forward (2018)

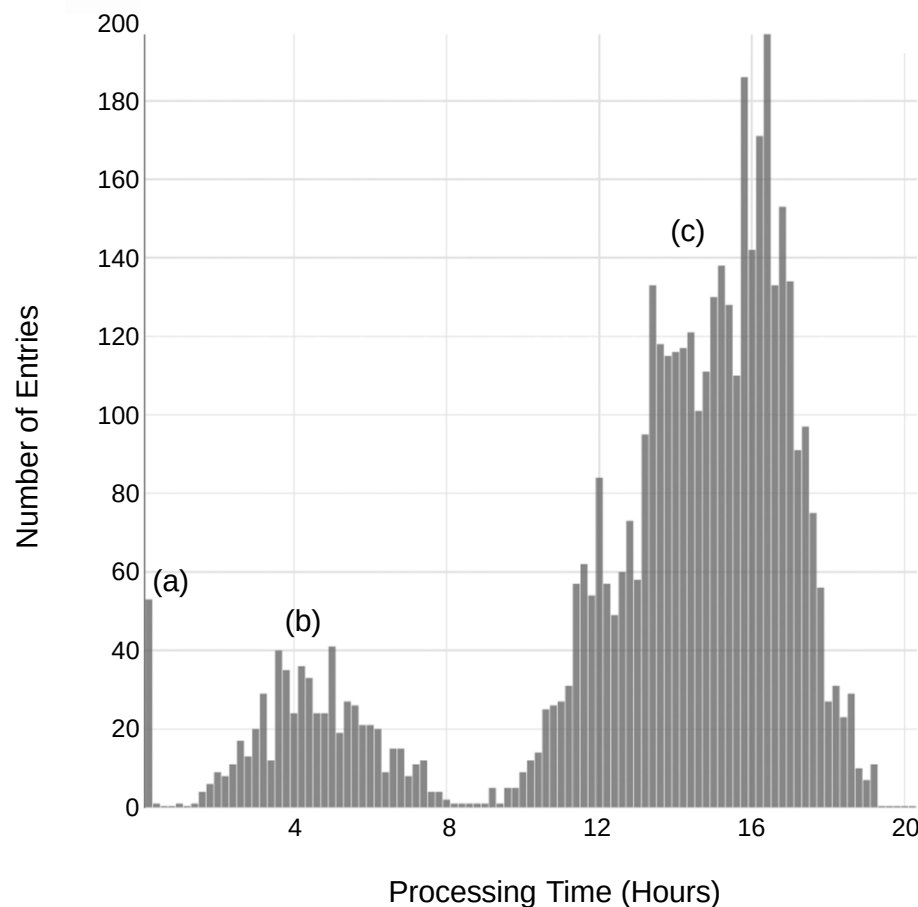
共通データ登録システム開発と高効率化

アノテーションに要する時間

(a) ~1hr: 修正依頼の必要ない単純な構造

(b) ~4 hrs: 修正依頼を必要としない複合体などの複雑な構造

(c) ~15 hrs: 登録者からの修正を必要とする構造



登録システムと品質管理システムの改良により、更に効率化を計る。

データ検証(Validation)レポートの 発行と実験データの公開



wwPDB X-ray Structure Validation Summary Report

Feb 28, 2014 - 04:13 AM GMT

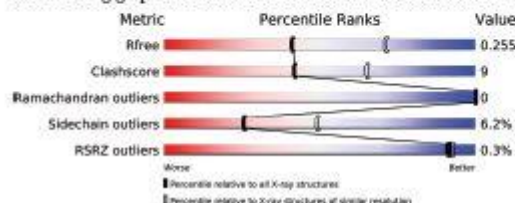
PDB ID : 3WDZ
Title : Crystal Structure of Keap1 in Complex with phosphorylated p62
Authors : Fukutomi, T.; Takagi, K.; Mizushima, T.; Tanaka, K.; Komatsu, M.; Yamamoto, M.
Deposited on : 2013-06-26
Resolution : 2.60 Å (reported)

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://www.pdb.org/ValidationPDFNotes.html>

1 Overall quality at a glance

The reported resolution of this entry is 2.60 Å.

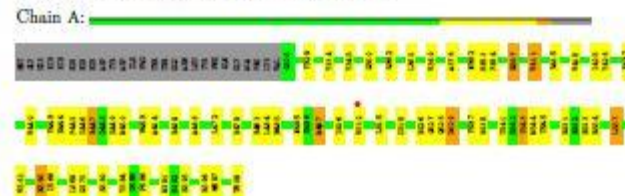
Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.



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



- Molecule 2: Peptide from Sequestosome-1



Validation report のweb公開

登録時に必須とされている実験情報

- X-ray: Structure Factor (構造因子)
- NMR: 化学シフトと距離拘束情報
- EM: 3DEM volume マップ

wwPDBによるValidation reportが 論文peer reviewのデフォルトに

EDITORIAL

nature
structural &
molecular biology

Nature Struct. Mol. Biology, 23 (10), 871, 2016

Where are the data?

Here, we announce two policy changes across Nature journals: data-availability statements in all published papers and official Worldwide Protein Data Bank (wwPDB) validation reports for peer review.

“We are now taking a further step and are requesting official wwPDB validation reports for peer review.

These reports are made available by the wwPDB after data deposition

(<http://www.wwpdb.org/validation/validation-reports>).

Other Nature journals will soon follow suit.”

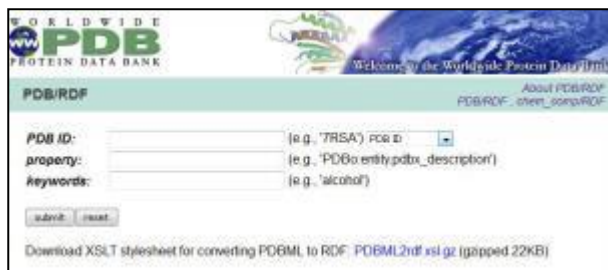
PDBjの活動

- 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(毎週アップデート)の運営
 - 関連DBとの統合化や二次データベース・ツールの開発

wwPDBにおけるRDF化の現状

wwPDB/RDF

<http://rdf.wwpdb.org/>



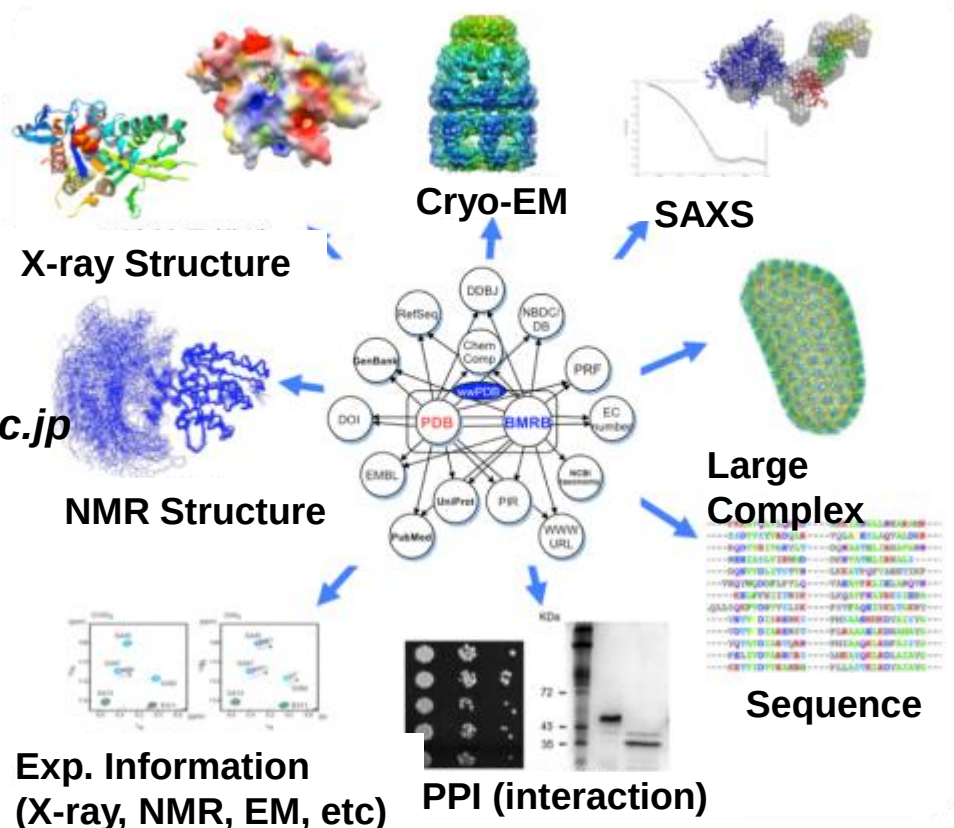
BMRB/RDF

<http://bmrbpub.protein.osaka-u.ac.jp>



Kinjo et al. (2012) Nucl. Acids Res. 40, D453-D460.

Yokochi et al. (2016) J. Biomed. Semantics, 7:16.



```
<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>
```

wwPDBにおけるRDF化の現状

wwPDB/RDF

<http://rdf.wwpdb.org/>

BMRB/RDF

<http://bmrbpub.protein.osaka-u.ac.jp>

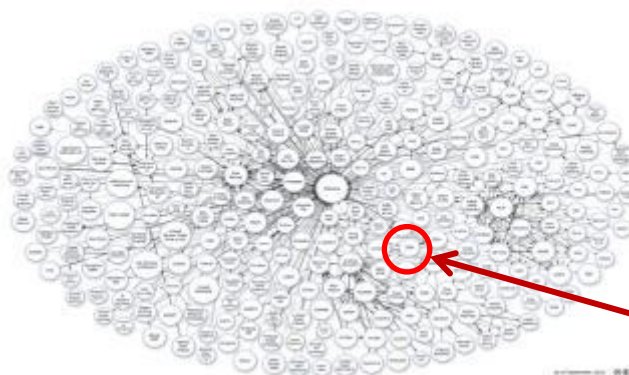
Kinjo et al. (2012) *Nucl. Acids Res.* 40, D453-D460.

Yokochi et al. (2016) *J. Biomed. Semantics*, 7:16.

Linked Data:

Connect Distributed Data across the Web

<http://linkeddata.org>



PDB

```
<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>
```


PDB/RDF example

By accessing <http://rdf.wwpdb.org/pdb/1GOF>, a list of category holders for the PDB entry 1GOF can be retrieved in the RDF/XML format. Then, a list of category elements can be retrieved (again in the RDF/XML format). Finally, for a particular category element, the list of properties of that element is retrieved.

Subject: <http://rdf.wwpdb.org/pdb/1GOF>

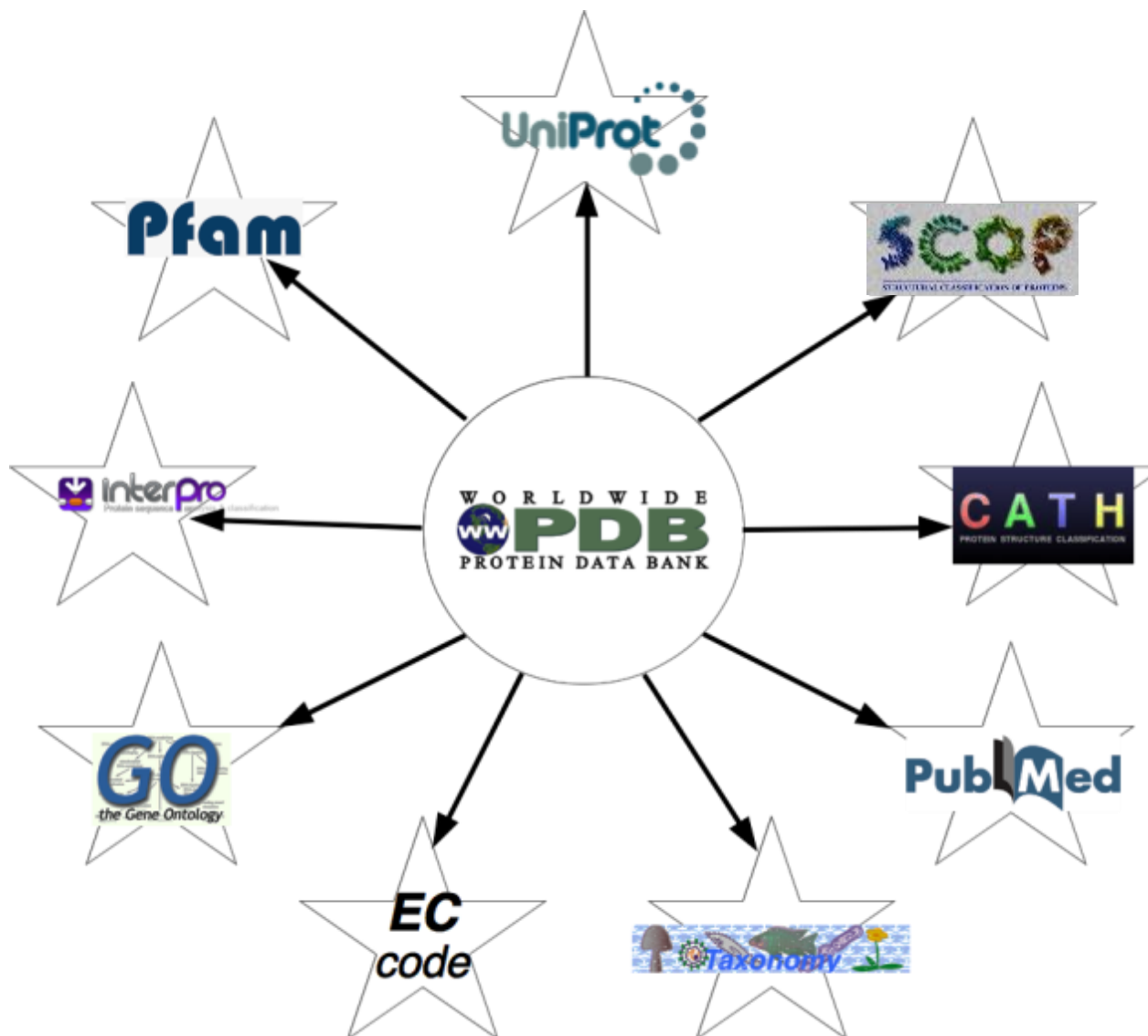
Predicate	Object
PDBo:datablockName	1GOF-noatom
PDBo:has_atom_sitesCategory	PDBr:1GOF/atom_sitesCategory
PDBo:has_atom_sites_footnoteCategory	PDBr:1GOF/atom_sites_footnoteCategory
PDBo:has_atom_typeCategory	PDBr:1GOF/atom_typeCategory
PDBo:has_audit_authorCategory	PDBr:1GOF/audit_authorCategory
PDBo:has_audit_conformCategory	PDBr:1GOF/audit_conformCategory
PDBo:has_cellCategory	PDBr:1GOF/cellCategory
PDBo:has_chem_compCategory	PDBr:1GOF/chem_compCategory
PDBo:has_citationCategory	PDBr:1GOF/citationCategory
PDBo:has_citation_authorCategory	PDBr:1GOF/citation_authorCategory
PDBo:has_computingCategory	PDBr:1GOF/computingCategory
PDBo:has_database_2Category	PDBr:1GOF/database_2Category
PDBo:has_database_PDB_matrixCategory	PDBr:1GOF/database_PDB_matrixCategory
PDBo:has_database_PDB_revCategory	PDBr:1GOF/database_PDB_revCategory
PDBo:has_database_PDB_rev_recordCategory	PDBr:1GOF/database_PDB_rev_recordCategory
PDBo:has_diffnCategory	PDBr:1GOF/diffrnCategory
PDBo:has_diffn_radiationCategory	PDBr:1GOF/diffrn_radiationCategory
PDBo:has_diffn_radiation_wavelengthCategory	PDBr:1GOF/diffrn_radiation_wavelengthCategory
PDBo:has_entityCategory	PDBr:1GOF/entityCategory
PDBo:has_entity_keywordsCategory	PDBr:1GOF/entity_keywordsCategory
PDBo:has_entity_polyCategory	
PDBo:has_entity_poly_seqCate	
PDBo:has_entity_src_genCate	
PDBo:has_entryCategory	
PDBo:has_exptlCategory	

Subject: <http://rdf.wwpdb.org/pdb/1GOF/entityCategory>

Predicate	Object
PDBo:has_entity	PDBr:1GOF/entity/1
PDBo:has_entity	PDBr:1GOF/entity/2
PDBo:has_entity	PDBr:1GOF/entity/3
PDBo:has_entity	PDBr:1GOF/entity/4
PDBo:has_entity	PDBr:1GOF/entity/5
rdf:type	http://rdf.wwpdb.org/pdb/1GOF/entityCategory

Subject: <http://rdf.wwpdb.org/pdb/1GOF/entity/1>

Predicate	Object
PDBo:of_datablock	PDBr:1GOF
PDBo:entity.formula_weight	68579.250
PDBo:entity.id	1
PDBo:entity.pdbx_description	GALACTOSE OXIDASE
PDBo:entity.pdbx_ec	1.1.3.9
PDBo:entity.pdbx_number_of_molecules	1
PDBo:entity.src_method	man
PDBo:entity.type	polymer
PDBo:link_to_enzyme	http://purl.uniprot.org/enzyme/1.1.3.9
PDBo:referenced_by_entity_keywords	PDBr:1GOF/entity_keywords/1
PDBo:referenced_by_entity_poly	PDBr:1GOF/entity_poly/1
PDBo:referenced_by_entity_src_gen	PDBr:1GOF/entity_src_gen/1
PDBo:referenced_by_struct_asym	PDBr:1GOF/struct_asym/A
PDBo:referenced_by_struct_ref	PDBr:1GOF/struct_ref/1
rdf:type	http://rdf.wwpdb.org/schema/pdbx-v40.owl#entity





ダウンロード数の推移

Year	Total	Total FTP Archive	Total Website	RCSB PDB FTP Archive	RCSB PDB Website	PDBe FTP Archive	PDBe Website	PDBj FTP Archive	PDBj Website
2011	383,131,048	276,952,286	106,178,762	204,939,406	81,560,098	40,960,368	18,515,245	31,052,512	6,103,419
2012	376,944,070	255,837,735	121,106,335	213,510,347	90,438,501	21,601,103	23,982,801	20,726,285	6,685,033
2013	441,262,210	296,176,290	145,085,920	215,331,908	97,549,580	43,684,850	37,762,496	37,159,532	9,773,844
2014	512,227,251	339,193,721	173,033,530	237,168,615	110,115,316	52,362,370	48,031,414	49,662,736	14,886,800
2015	534,339,871	368,244,766	166,095,105	255,346,630	111,802,897	48,544,330	41,127,219	64,353,806	13,164,989
2016	591,876,087	366,677,897	225,198,190	293,648,366	161,208,456	30,274,284	44,432,830	42,755,247	19,556,904
2017	398,016,597	265,239,786	132,776,811	213,942,495	84,649,397	20,630,847	37,100,965	30,666,444	11,026,449

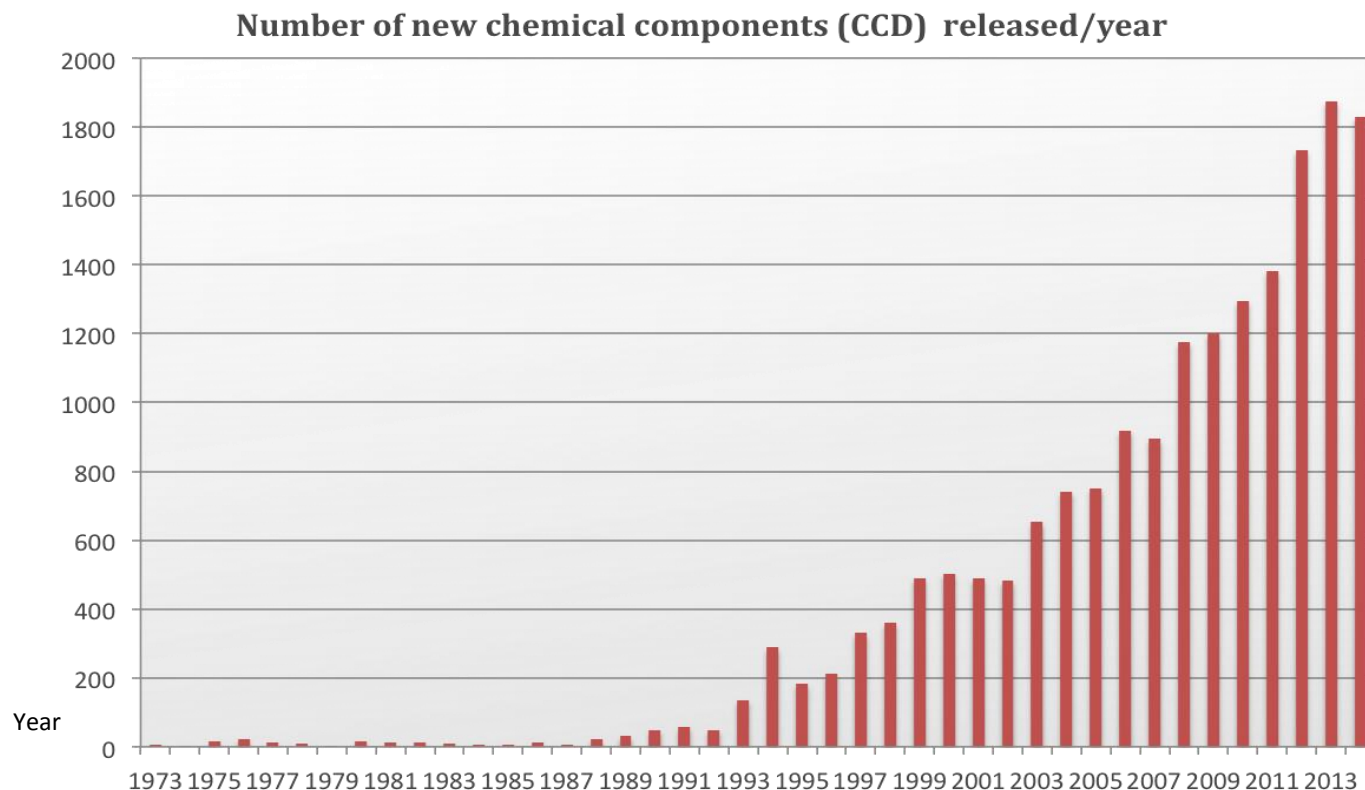
More than 1.6 million / day



蛋白質構造データバンクの データ検証高度化と統合化

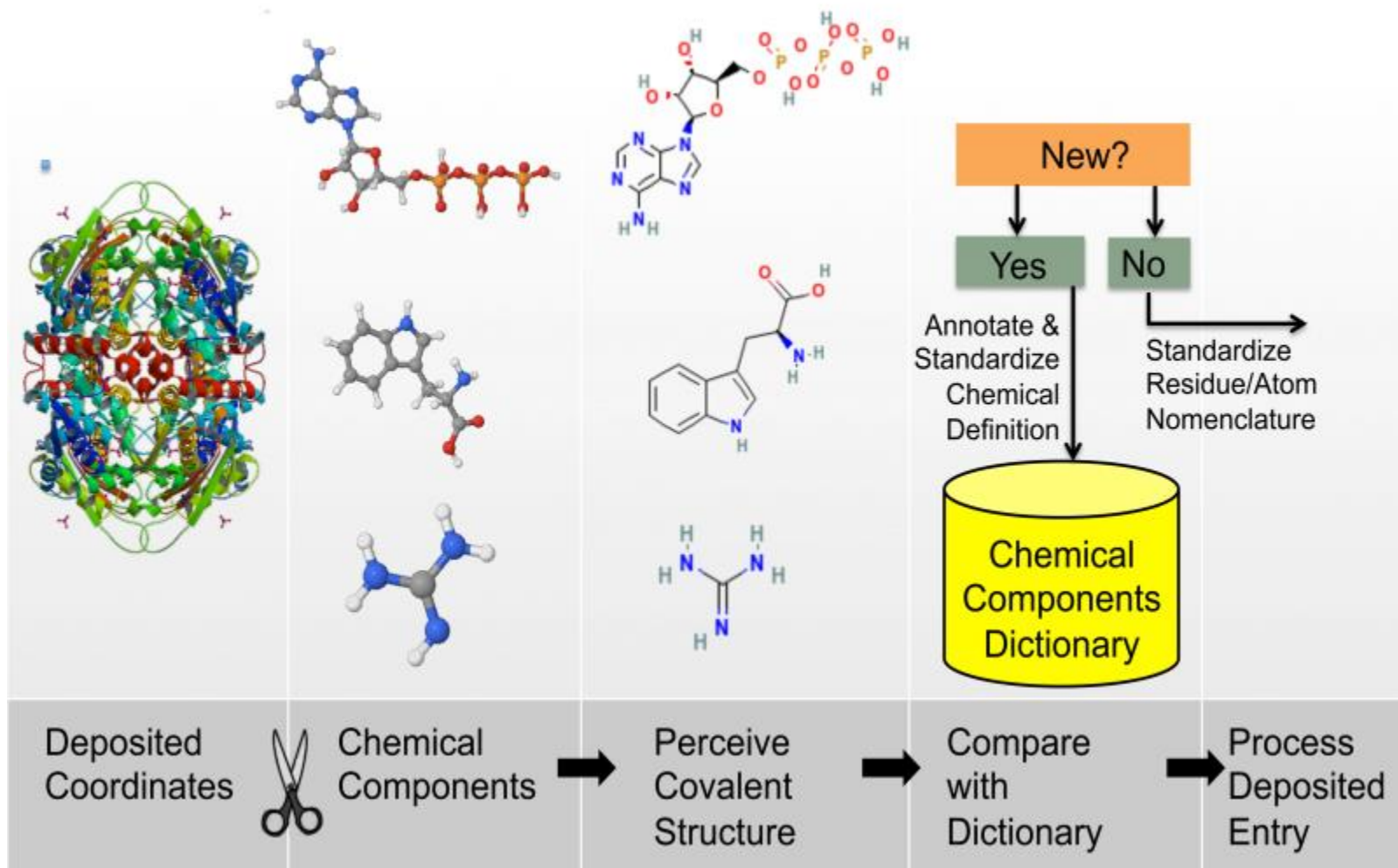
- 1) wwPDBメンバーとしてのPDBアーカイブ構築・検証・公開
- 2) 他のデータベースとの統合化および高度化
 - 2-1) CSDとの統合化
 - 2-2) ASHビューアの開発→関西学院大学藤教授より
 - 2-3) アミノ酸の立体構造中での役割の自動アノテーション
 - 2-4) 統合化されたデータベースおよび関連ツール開発
- 3) データベースの利用促進・人材育成
 - 3-1) 利用促進・アノテータの育成
 - 3-2) 国際協力
- 4) 達成達成目標と選考時に頂いた意見への対応

2-1) Cambridge Structural Data (CSD) との統合化

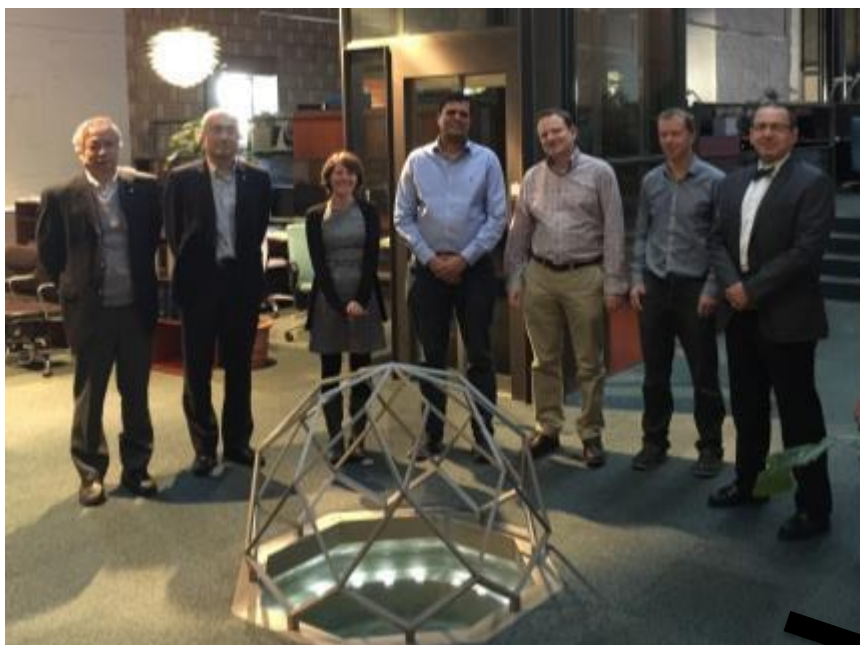


PDBに含まれる化合物データ(CCD)を製薬企業などの創薬ユーザーが積極的に活用

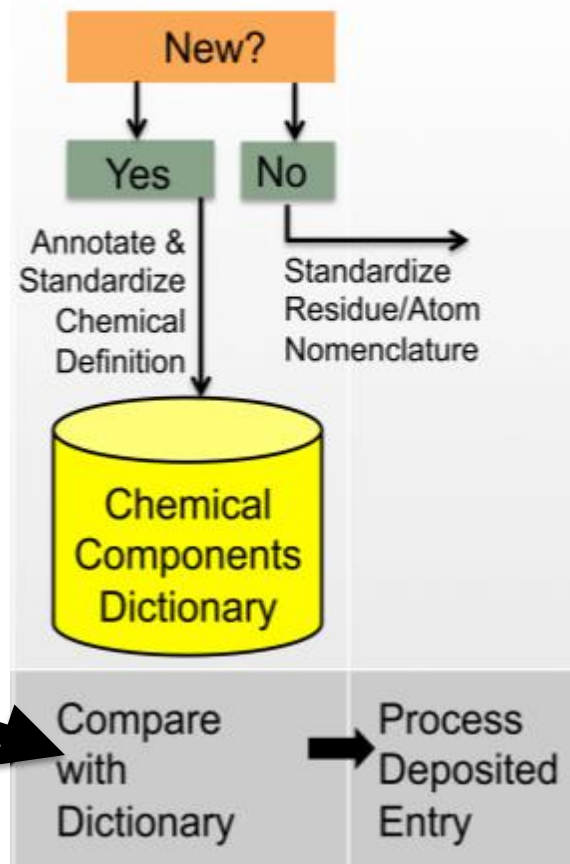
の流れ



PDBとCSDの連携を データ統合(RDF化)に生かす



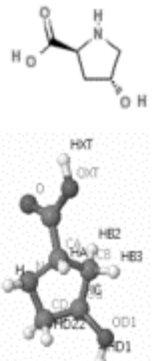
データ検証の段階で、
緊密な連携体制を構築



PDB中の化合物情報(CCD)と Cambridge Structural Data(CSD)との統合化

a)

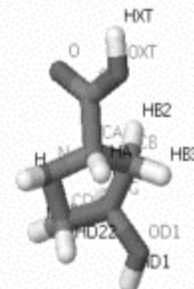
```
data_HYP
#
_chen_comp.id          HYP
_chen_comp.name        4-HYDROXYPROLINE
_chen_comp.type        "L-PEPTIDE LINKING"
_chen_comp.pdbx_type   ATOMP
_chen_comp.formula      "C5 H9 N O3"
_chen_comp.mon_std_parent_comp_id PRO
_chen_comp.pdbx_synonyms HYDROXYPROLINE
_chen_comp.pdbx_formal_charge 0
_chen_comp.pdbx_initial_date 1999-07-08
_chen_comp.pdbx_modified_date 2008-04-29
_chen_comp.pdbx_ambiguous_flag N
_chen_comp.pdbx_release_status REL
_chen_comp.pdbx_replaced_by ?
_chen_comp.pdbx_replaces ?
_chen_comp.formula_weight 131.130
_chen_comp.one_letter_code ?
_chen_comp.three_letter_code HYP
_chen_comp.pdbx_model_coordinates_details ?
_chen_comp.pdbx_model_coordinates_missing_flag N
_chen_comp.pdbx_ideal_coordinates_details Corina
_chen_comp.pdbx_ideal_coordinates_missing_flag N
_chen_comp.pdbx_model_coordinates_db_code 10B6
_chen_comp.pdbx_processing_site RCSB
```



b)

```
loop_
_chen_comp_atom.comp_id
_chen_comp_atom.atom_id
_chen_comp_atom.alt_atom_id
_chen_comp_atom.type_symbol
_chen_comp_atom.charge
_chen_comp_atom.pdbx_align
_chen_comp_atom.pdbx_aromatic_flag
_chen_comp_atom.pdbx_leaving_atom_flag
_chen_comp_atom.pdbx_stereo_config
_chen_comp_atom.model_Cartn_x
_chen_comp_atom.model_Cartn_y
_chen_comp_atom.model_Cartn_z
_chen_comp_atom.pdbx_model_Cartn_x_ideal
_chen_comp_atom.pdbx_model_Cartn_y_ideal
_chen_comp_atom.pdbx_model_Cartn_z_ideal
_chen_comp_atom.pdbx_ordinal
```

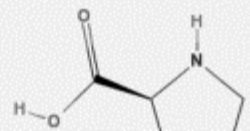
HYP	N	H	N	O	1	N	N	-3.366	16.585	44.188	0.168	1.360	-0.282	1
HYP	CA	CA	C	O	1	N	N	-2.955	15.768	43.044	-0.384	-0.003	-0.493	2
HYP	C	C	C	O	1	N	N	-1.447	15.609	43.030	-1.811	-0.072	-0.013	3
HYP	O	O	O	1	N	N	N	-0.722	16.484	43.503	-2.233	0.764	0.750	4
HYP	CB	CB	C	O	1	N	N	-3.408	16.578	41.829	0.515	-0.924	0.359	5
HYP	CG	CG	C	O	1	N	N	-4.437	17.482	42.330	1.847	-0.159	0.505	6
HYP	CD	CD	C	O	1	N	N	-4.068	17.803	43.753	1.640	1.159	-0.271	7
HYP	OD1	OD	O	1	N	N	N	-5.693	16.815	42.294	2.917	-0.913	-0.071	8
HYP	OXT	OXT	O	1	N	N	N	-0.976	14.502	42.469	-2.614	-1.063	-0.433	9
HYP	N	N	N	O	1	N	N	-3.980	16.047	44.765	-0.107	1.981	-1.020	10
HYP	HA	HA	H	O	1	N	N	-3.385	14.756	43.068	-0.329	-0.278	-1.546	11
HYP	HB2	HB	H	O	1	N	N	-2.567	17.141	41.398	0.066	-1.692	1.337	12
HYP	HB3	HB	H	O	1	N	N	-3.790	15.930	41.026	0.678	-1.873	-0.153	13
HYP	HC	HC	H	O	1	N	N	-4.508	18.399	41.726	2.052	0.048	1.555	14
HYP	HD22	HD	H	O	0	N	N	-4.956	18.005	44.370	2.018	1.065	-1.289	15
HYP	HD23	HD	H	O	0	N	N	-3.457	18.713	43.848	2.132	1.965	0.243	16
HYP	HD1	HD	H	O	1	N	N	-5.999	16.666	43.181	3.780	-0.479	-0.009	17
HYP	HXT	HXT	H	O	1	N	N	-0.027	14.511	42.499	-3.520	-1.066	-0.098	18



c)

```
loop_
_chen_comp_bond.comp_id
_chen_comp_bond.atom_id_1
_chen_comp_bond.atom_id_2
_chen_comp_bond.value_order
_chen_comp_bond.pdbx_aromatic_flag
_chen_comp_bond.pdbx_stereo_config
_chen_comp_bond.pdbx_ordinal
```

HYP	N	CA	SING	N	1
HYP <th>N</th> <th>CD</th> <th>SING</th> <th>N</th> <th>2</th>	N	CD	SING	N	2
HYP <th>N</th> <th>H</th> <th>SING</th> <th>N</th> <th>3</th>	N	H	SING	N	3
HYP <th>CA</th> <th>C</th> <th>SING</th> <th>N</th> <th>4</th>	CA	C	SING	N	4
HYP <th>CA</th> <th>CB</th> <th>SING</th> <th>N</th> <th>5</th>	CA	CB	SING	N	5
HYP <th>CA</th> <th>HA</th> <th>SING</th> <th>N</th> <th>6</th>	CA	HA	SING	N	6
HYP <th>C</th> <th>O</th> <th>DOUB</th> <th>N</th> <th>7</th>	C	O	DOUB	N	7
HYP <th>C</th> <th>OXT</th> <th>SING</th> <th>N</th> <th>8</th>	C	OXT	SING	N	8
HYP <th>CB</th> <th>CG</th> <th>SING</th> <th>N</th> <th>9</th>	CB	CG	SING	N	9
HYP <th>CB</th> <th>HB2</th> <th>SING</th> <th>N</th> <th>10</th>	CB	HB2	SING	N	10
HYP <th>CB</th> <th>HB3</th> <th>SING</th> <th>N</th> <th>11</th>	CB	HB3	SING	N	11
HYP <th>CG</th> <th>CD</th> <th>SING</th> <th>N</th> <th>12</th>	CG	CD	SING	N	12



d)

```
#
loop_
_pdbx_chen_comp_descriptor.comp_id
_pdbx_chen_comp_descriptor.type
_pdbx_chen_comp_descriptor.program
_pdbx_chen_comp_descriptor.program_version
_pdbx_chen_comp_descriptor.descriptor
```

HYP	SMILES	ACDLabs	12.01	O=C(O)C1NCC(O)C1
HYP <th>SMILES_CANONICAL</th> <th>CACTVS</th> <td>3.385</td> <td>O[C@H]1CN[C@@H](C1)C(O)=O</td>	SMILES_CANONICAL	CACTVS	3.385	O[C@H]1CN[C@@H](C1)C(O)=O
HYP <th>SMILES</th> <th>CACTVS</th> <td>3.385</td> <td>O[C@H]1CN[C@H](C1)C(O)=O</td>	SMILES	CACTVS	3.385	O[C@H]1CN[C@H](C1)C(O)=O
HYP <th>SMILES_CANONICAL</th> <th>"OpenEye OEToolkits"</th> <td>1.7.5</td> <td>C1[C@H](CN[C@@H]1C(O)=O)O</td>	SMILES_CANONICAL	"OpenEye OEToolkits"	1.7.5	C1[C@H](CN[C@@H]1C(O)=O)O
HYP <th>SMILES</th> <th>"OpenEye OEToolkits"</th> <td>1.7.5</td> <td>C1C(CNC1C(O)=O)O</td>	SMILES	"OpenEye OEToolkits"	1.7.5	C1C(CNC1C(O)=O)O
HYP <th>InChI</th> <th>InChI</th> <td>1.03</td> <td>"InChI=1S/C5H9NO3/c7-3-1-4(5(8)9)6-2-3/h3-4,6-7H,1 -"</td>	InChI	InChI	1.03	"InChI=1S/C5H9NO3/c7-3-1-4(5(8)9)6-2-3/h3-4,6-7H,1 -"
HYP <th>InChIKey</th> <th>InChI</th> <td>1.03</td> <td>PMYEEVYMNWASQ-N-DW7CNVQSA-N</td>	InChIKey	InChI	1.03	PMYEEVYMNWASQ-N-DW7CNVQSA-N

SMILES

InChI

2-2) ASHビューアの開発

背景

アラインメントによる相同な立体構造やアミノ酸配列の比較

配列と構造を結びつけてタンパク質の機能情報を抽出するのに有効
(ASHなどの比較ツール開発)

→ 関西学院大学 藤博幸教授より

ASH VIEWER

大容量データに対応した可視化
ユーザインタラクティブな機能解析のサポート

ASH Viewer

情報統合のプラットフォーム

大容量データに対応した可視化
ユーザインタラクティブな機能解析のサポート

ALIGNMENT

A base for computational analysis of proteins

...LIFDQALNKWGHAE...
...LMFDQALARWGSAAE...
...LIYDQAIGKWGHGAE...
...MIYDQALCKWGAHAE...
...MIFNQAIKWGHPAE...

PDBj

- Structural Alignment: **ASH**, **GASH**, ...
- Conservation/Divergence Analysis: **SEALA**

PROBLEM

Increase of sequence and structure data

Several alignment tools have been developed to handle enormous amount of data

TIPA_ARATH	RADEAT-----HPDSIRATLAEFLSTFVFVFAAEGSIL
TIPG_ARATH	RPDEAT-----RPDALKAAALAEFISTLIFVVAGSGSGM
S47037	SHREVY-----EVGALKAAALAEFISTLIFVFAGQGSGM
T12439	RAEEAA-----HPETLKQALAEFISTLIFVFAGEGSGM
T48886aqup[imported]	RFDDSF-----SLGSFKAYLAEFISTLLYVVFAGVGSAT
T07819	SFDDSF-----SLASLRAYLAEFISTLLFVFAGVGSAT
TIP1_TOBAC	SIGDSF-----SVGSLKAYVAEFIATLLFVFAGVGSAT
T01648	SFRDSL-----SAASLKAYVAEFIATLLFVFAGVGSAT
I52366utwchuman	MASEFK-----KKLFWRAVVAEFLATTLFVFISIGSAL
A41616eimphuman	MASEFK-----KKLFWRAVVAEFLATTLFVFISIGSAL
AQP1_MOUSE	MASEIK-----KKLFWRAVVAEFLAMTLFVFISIGSAL
AQP2_HUMAN	-MWELR-----SIAFSRAVFAEFLATLLFVFFGLGSAL
AQP5_RAT	MKKEVC-----SLAFFKAVFAEFLATLIFVFFGLGSAL
MIP_HUMAN	-MWELR-----SASFWRAIFAEEFATLIFYVFFGLGSSL
AQP4_HUMAN	AFKGWV-----TQAFWKAVTAEEFLAMLIFVLLSLGSTI
AQP4_RAT	AFKGWV-----TQAFWKAVTAEEFLAMLIFVLLSVGSTI
TIPW_PEA	EPSELT-----SWSFYRAGIAEFIATFLFLYITVLTVM
T02095tmrice	EPGELK-----SWSFYRAGIADFVATFLFLYITILTVM
T04139tmrice	EPGELK-----SWSFYRAGIAEFMATFLFLYITVLTVM
WC1A_ARATH	EPGELS-----SWSFWRAGIAEFIATFLFLYITVLTVM
T01528	EPGELS-----SWSFYRAGIAEFIATFLFLYITVLTVM
T03794aqupNT2	WR-QKS-----CRHGLFTELEFSMATFLFLYITILTVM
T09794mip	EPGELK-----SWSFYRAGIAEFVATFLFLYITILTVM
T14599mip1	DMEELK-----KWSFYRAIIAEFVATLLFLYVTVLTVI
WC2A_ARATH	DGAELK-----KWSFYRAVIAEFVATLLFLYITVLTVI
T14600	DMGELK-----LWSFWRAVIAEFIATLLFLYITVATVI
JC5791aqup9human	KIQEIL-----QRKMVREFLAEFMSTYVMVMVFLGGSVA
T04053nodu26homolog	KCLPVMGSTWGQHDTCFTDFPSPDVSRLTRKLGAEFVGTFILIFTATAGPI
AQY1_YEAST	KHHRFKISR-----DTLRDHFIAAVGEFCGTFMFLWCAYVICN
aqp1_humanN	MASEFK-----KKLFWRAVVAEFLATTLFVFISIGSAL
2T22051hypoC	ATLLSP-----GSQWWQGLIAETVVTFFLVHTILITAA
2AQP8_MOUSE	FAIVQE-----QEQVAEALGIEIILTMLLVLAFCMGAV
2S01444nodu26	GTVPN-----G-TNLQAFVFEFIMTFFLMFVICGVAT
2JQ2286nodu26	GTVPN-----G-TNLQAFVFEFIMTFFLMFVICGVAT
2JQ2285nodu26	GTLPN-----G-TNLQAFVFEFIITFLLMFVISGVAT
2T05028nodu26-like	GTLPS-----G-SNLQSFVIEFIITFYLMFVISGVAT
2D69004hypoMTH	ATAPFP-----GIGYWQAMLAEVGTGFLLLMITIMGIAV
2A69428glpfhomolog	ATAPFP-----GIGYGQAAILTEAIGTFLLMLVIMGVAV
2AQP2_ECOLI	EHSPG-----GYSMLSALVVELVLSAGFLLVIHG-AT
2AQP3_HUMAN	ATYPSG-----HLDMINGFFDQFIGTASLIVCVLAIVD
2AQP9_HUMAN	ATYPAP-----YLSLANAFADQVVATMILLIIVFAIFD
2GLPF_ECOLI	STYPNP-----HINEVQAFAVEMVITAILMGLILALTD
2GLPF_HAEIN	STYPHP-----SLSIGGAFAVEFVITAILMALIMALTD
2T18834hypoC	ASYPAP-----HLGLVNGFVDQFVATAVVFVFLIAHIVD
2T16635hypoC	CSYPAL-----HVSNTTAFDQFAGTALLVLVFCVVID
2GLPF_MYCPN	CTNPA-----IFNIPRNFATEFVATSVLIASLLVAGS
2GLPF_THEMA	CTGPA-----VRKYGANLLTEIIGTMVLLMGVLGIGA
2GLPF_BACSU	STGPS-----IPHTFANVLSEVIGTFVLVLGILAIGA
2T10253mp28cucurbit	GFFVSS-----GISELHGFLEILLTFALVYVYATAV
2TIPA_ARATH	GFRLAS-----GVGAVNGLVLEIILTFGLVYVYSTLI
2TIPG_ARATH	AFGLSA-----GVGVNLNAFVFEIVMTFGLVYTVYATAI
2S47037	TFGLTG-----IGAWEAUVLEIVMTFGLVYTVYATAV
2T12439	AFALSA-----DVTVMNALVFEIVMTFGLVYTVYATAI
2T48886aqup[imported]	THSVAG-----RVGAIEGVVMEIITFALVYTVYATAA
2T07819	THSVAA-----GVGAIEGVVMEIITFSLVYTVYPTAA
2TIP1_TOBAC	THGVAA-----GLNGLQGVVMEIITFALVYTVYATAA

PROBLEM

It is difficult to analyze the alignment, due to the huge volume.

There are several alignment viewers. But, their functions are not enough for the investigation of the alignment.

どのように大量配列を可視化するか

機能的に関連する類縁タンパク質をグループ化
して比較解析が行われる



系統樹あるいはユーザ指定のクラスタ
に従って、グループ化して表示

3種の入力データから可視化/解析

- (1) アライメント
- (2) 立体構造
- (3) 系統樹あるいはクラスタ情報

系統樹のサブツリーの再構成と、 それに連動するアライメントの表示の変更



viewer

File Search

アライメント

KLMTGDFREDDRHGFPQWTYSAED

クリック可能なノード

ノードに対応する
コンセンサスアライメント

系統樹のサブツリーの再構成と、 それに連動するアライメントの表示の変更

The screenshot shows a software window titled "viewer" with a menu bar containing "File" and "Search". The window is divided into two main panels. The left panel displays a phylogenetic tree with several nodes; some nodes are blue circles, and others are white circles. A yellow arrow points to a specific blue node, which is also enclosed in a red square. Below this node, a red-bordered box contains the text "サブツリーの展開" (Expand Subtree). The right panel is titled "アライメント" (Alignment) and displays two lines of sequence alignment: "KLMDGDFREDQQRHGFPQWTYSAQD" and "KLMTPDFRENDRHGFSQWGFGAED". These two lines are enclosed in a red-bordered box. Below this box, a blue bracket connects it to the text "アラインメントの展開" (Expand Alignment).

viewer

File Search

アライメント

KLMDGDFREDQQRHGFPQWTYSAQD
KLMTPDFRENDRHGFSQWGFGAED

サブツリーの
展開

アラインメントの
展開

ユーザインタラクティブな比較解析

保存度/変異度（配列/構造）

機能差を決定するサイトの推測（配列/構造）

オーソログの推定

立体構造へのマッピング（Molmil使用）

分子進化工学データの解析

クラスタ情報の要約

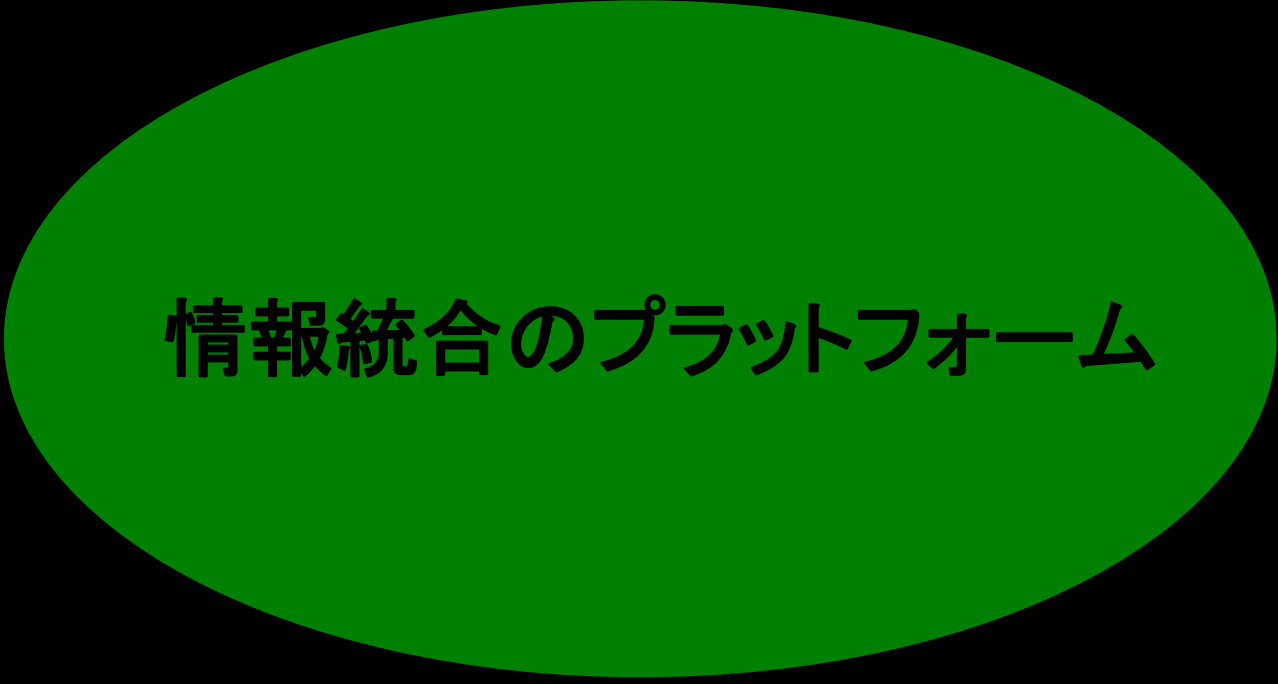
サブツリー、コンセンサス配列に含まれる

配列/構造の情報（本数、生物種、タンパク質名）

大容量データに対応した可視化
ユーザインタラクティブな機能解析のサポート

ASH Viewer

情報統合のプラットフォーム



情報統合のプラットフォーム

情報統合のプラットフォーム

各種データベースにある情報をネットを介してSPARQL検索により収集し、それに基づいて解析

- 定型検索（組み込み）

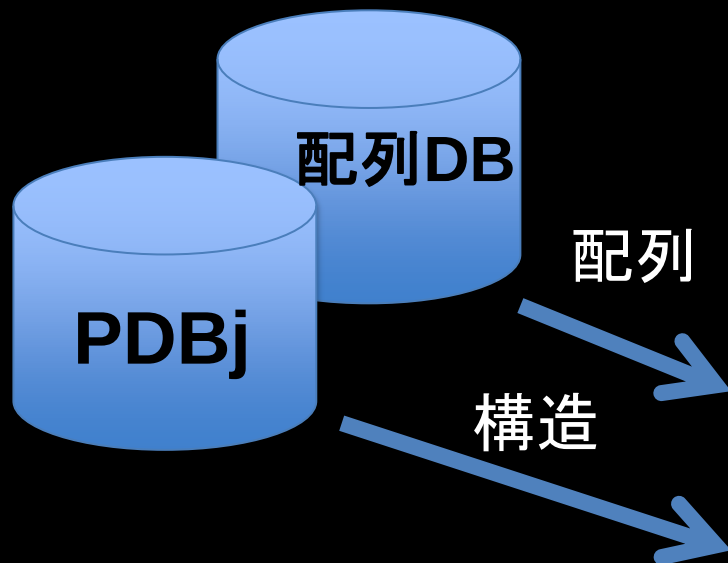
機能情報(PDBJ)

- 非定型検索 (endpointとSPARQLを入力)

目的に応じてSPARQL検索

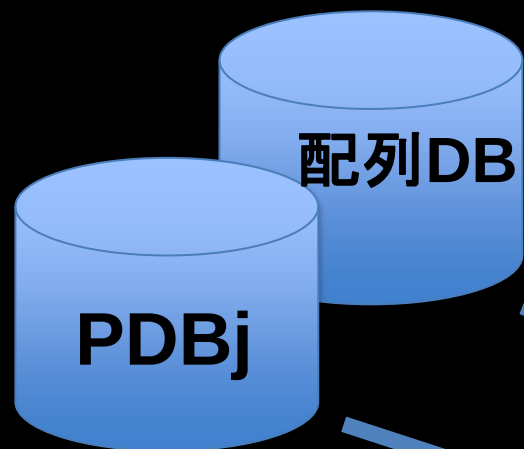
ex) 生物種の情報から至適生育温度を得て、
アミノ酸組成の違いを解析
(Microme.db)

想定される利用方法



アラインメント

...LIFDQALNKWGHAE...
...LMFDQALARWGSAAE...
...LIYDQAIGKWGHGAE...
.
.
.
...MIYDQALCKWGAHAE...
...MIFNQAIKWGHPAE...



アラインメント

39 . . . L I F D Q A L N K W G H H A E . . .
75 . . . L M F D Q A L A R W G S H A E . . .
81 . . . L I Y D Q A I G K W G H G A E . . .

. . .
. . .
. . .

33 . . . M I Y D Q A L C K W G A H A E . . .
24 . . . M I F N Q A I A K W G H P A E . . .

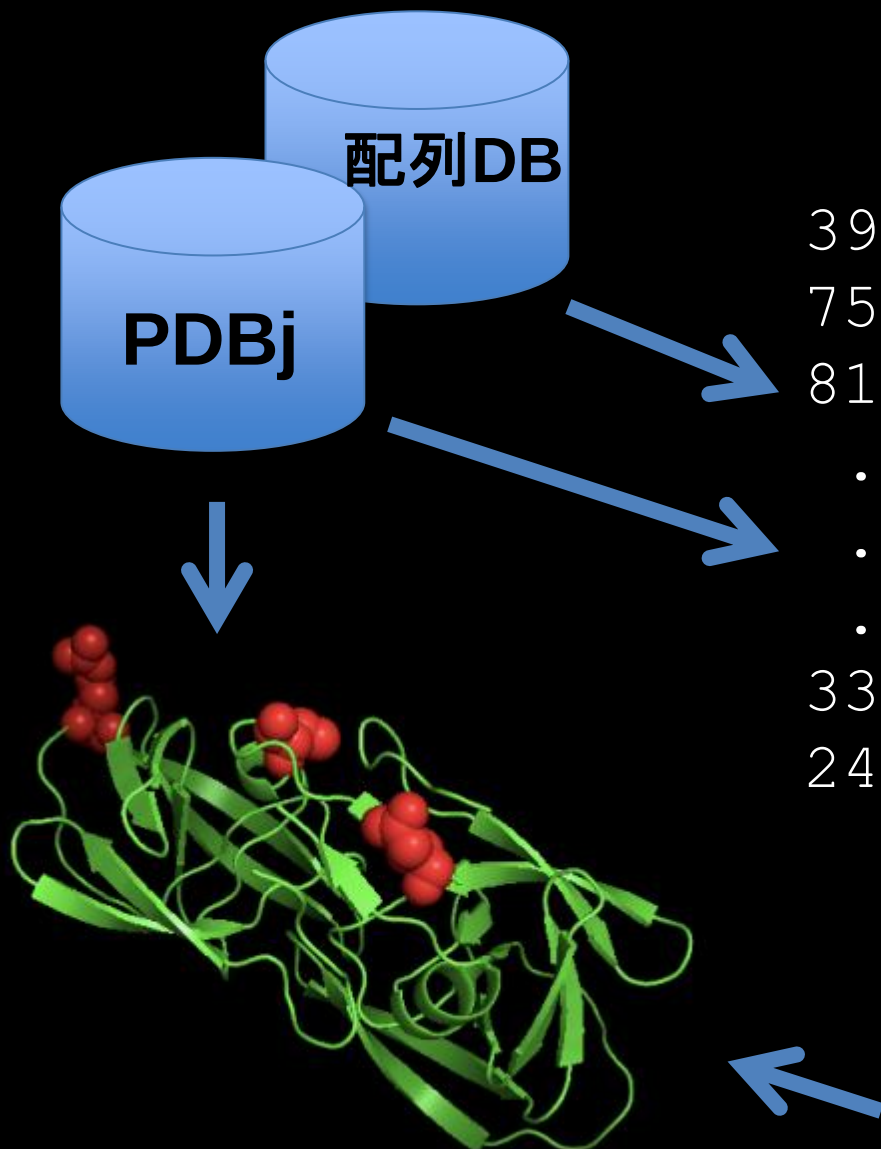
ソースの至適生育温度



ベイズ最適化

. . . M L F N Q A I A K W S H S A E . . .

好熱タンパク質の設計



立体構造上での設計の検討

rational designへの手がかり

アラインメント

39 . . . L I F D Q A L N K W G H H A E . . .
 75 . . . L M F D Q A L A R W G S H A E . . .
 81 . . . L I Y D Q A I G K W G H G A E . . .

.

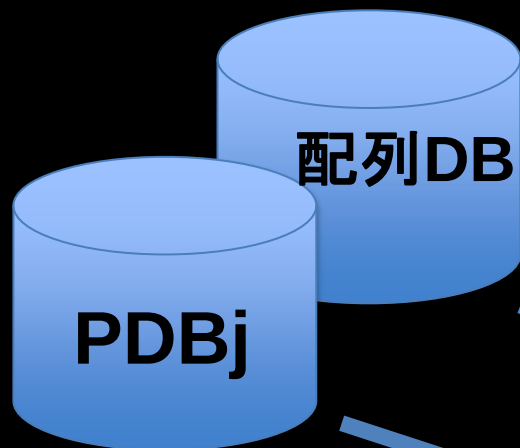
33 . . . M I Y D Q A L C K W G A H A E . . .

24 . . . M I F N Q A I A K W G H P A E . . .

ベイズ最適化

. . . M L F N Q A I A K W S H S A E . . .

好熱タンパク質の設計



アラインメント

39 . . . L I F D Q A L N K W G H H A E . . .
75 . . . L M F D Q A L A R W G S H A E . . .
81 . . . L I Y D Q A I G K W G H G A E . . .

. . .
. . .
. . .

33 . . . M I Y D Q A L C K W G A H A E . . .
24 . . . M I F N Q A I A K W G H P A E . . .

ソースに関連する情報



ベイズ最適化

. . . M L F N Q A I A K W S H S A E . . .

設計

Javaで作成

- JavaFX でGUIを作成
- 要約機能にはJDBCを使用
- SPAQL検索にはJenaを使用

開発体制

藤 全体の取りまとめ、
機能解析部位, GUIの作成

山下 Jenaを用いたSPARQL検索部分

工藤 Molmilによる立体構造表示部分

研究開発テーマ	タスク	小タスク	2017年度	2018年度	2019年度	2020年度	2021年度
---------	-----	------	--------	--------	--------	--------	--------

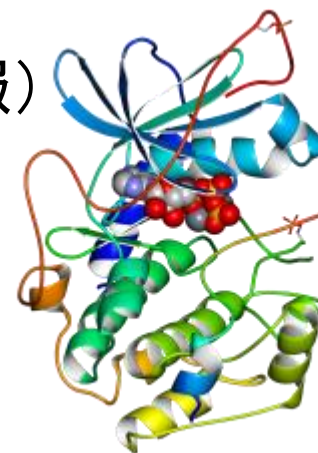
ASHの高度化	入力機能実装																			
	クラスタ化による可視化機能実装	クラスタ同定機能作成		—	→															
		表示部分作成		—	→															
	クラスタの圧縮、展開、要約機能実装	クラスタ圧縮、展開機能作成				—	→													
		クラスタ圧縮、展開機能作成					—	→												
	アラインメントからの機能部位推測機能と、その結果の立体構造への反映機能の実装	アラインメントからの機能部位推測機能作成							—	→										
		機能部位の立体構造への反映機能の作成									—	→								
	ネットを介してRDF化されたデータベースからの関連情報の取得機能の実装	SPARQL検索機能作成		—	—	→						—	→							
		検索に使用されるDBの調査												—	→					
		各種検索情報の機能解析への連結														—	—	—	→	

2-3) アミノ酸の立体構造中での 役割の自動アノテーション



SIFTS by PDBe/EBI

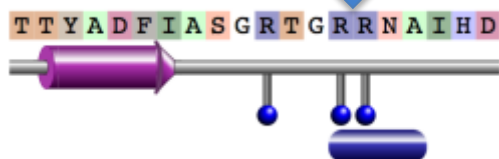
PDB (座標情報)



分子内相互作用

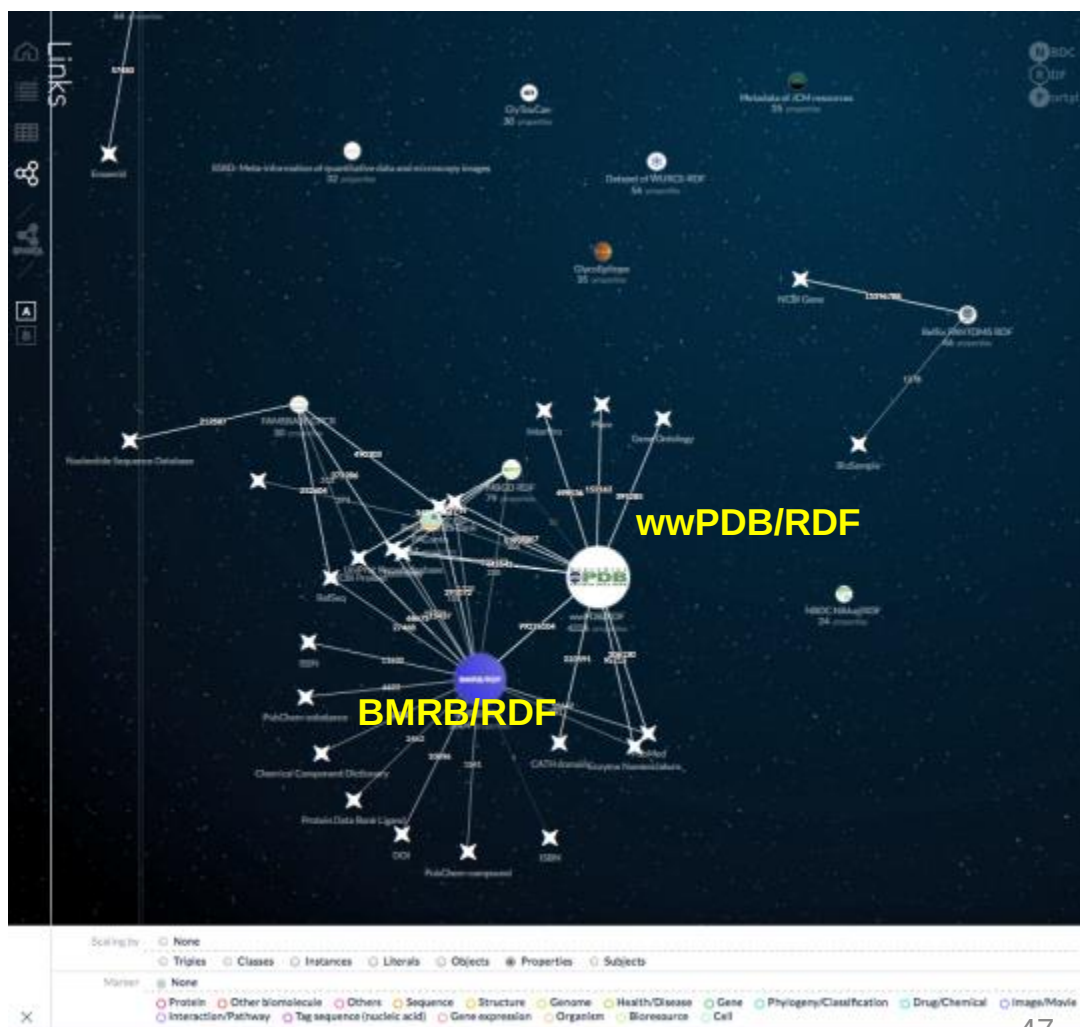


分子間相互作用



外部リソースと座標情報の統合により各残基の立体構造中での役割(分子内・分子間コンタクト)と機能を自動的にアノテーション
→残基レベルのデータベース解析が可能に！

発



NBDC RDF-Portal
<https://integbio.jp/rdf/>

セマンティック技術を利用した

データ検証の高度化



データの信頼性をマシン(AI)が判断できるようになり、PDBのデータ統合化には極めて重要.

※データ検証ファイルのRDF化

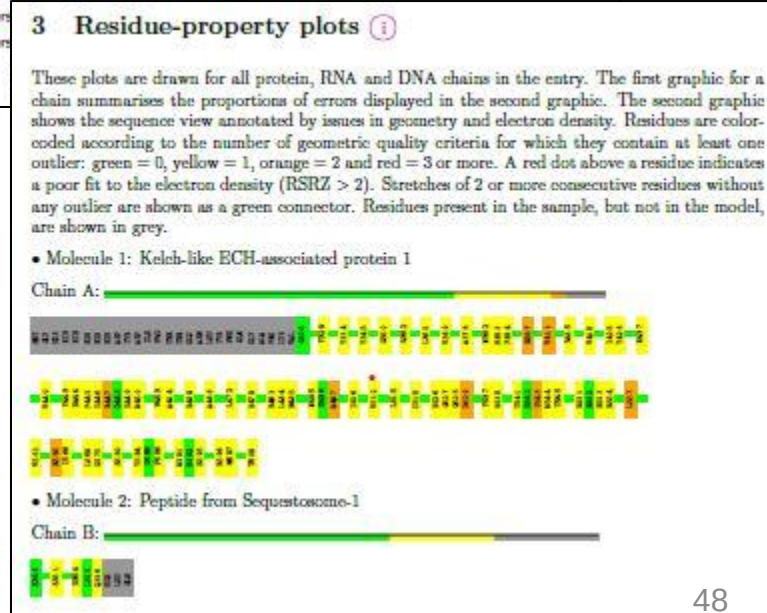
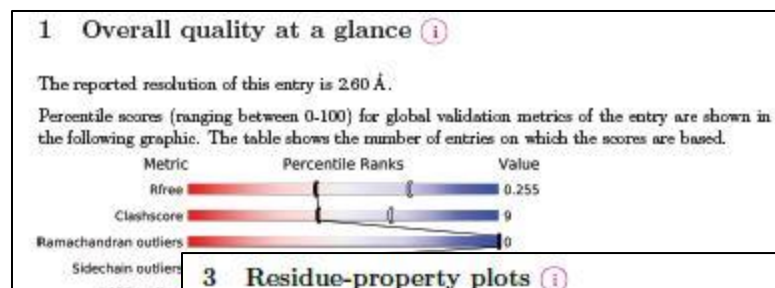
WORLDWIDE PDB
PROTEIN DATA BANK

wwPDB X-ray Structure Validation Summary Report ⓘ

Feb 28, 2014 - 04:13 AM GMT

PDB ID : 3WIZ
Title : Crystal Structure of KcsA 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@gmail.com.
A user guide is available at <http://www.pdb.org/ValidationPDFs.html>



揺らぎや相互作用などの情報も新フォーマットへ

Yokochi et al., *J. Biomed. Sem.* (2016)

➡ 更に、機械が理解出来るフォーマットへ

NMR-STAR v3

BMRBxTool

BMRB/XML

BMRBoTool

BMRB/RDF

多機能検索エンジン

Search

Everything Sequence

e.g., Entry ID, Macromolecule, Gene Ontology or Author

Search help

Found 467 Documents (0.735 - 1.365 seconds)

Showing 1 - 10 of 98 Documents, order by Relevance

BMRB: 17103

The structure of the calcium-sensitizer, dfbp-o, in complex with the N-domain of troponin C and the switch region of troponin I

Authors: Robertson, L.M., Sun, Y., Li, M.X., Sykes, B.D.

Assembly: calcium-sensitizer, dfbp-o, in complex with troponin C and the switch region of troponin I

Entity:

- cNTnC (polymer), 89 monomers, 10062.23 Da [Detail](#)
- cTnI(144-163) (polymer), 20 monomers, 2214.660 Da [Detail](#)
- CA (non-polymer), 40.078 Da
- dfbp-o (non-polymer), 197.239 Da

Total weight: 12514.206 Da

Max. entity weight: 10062.23 Da

Source organism: Homo sapiens (H)

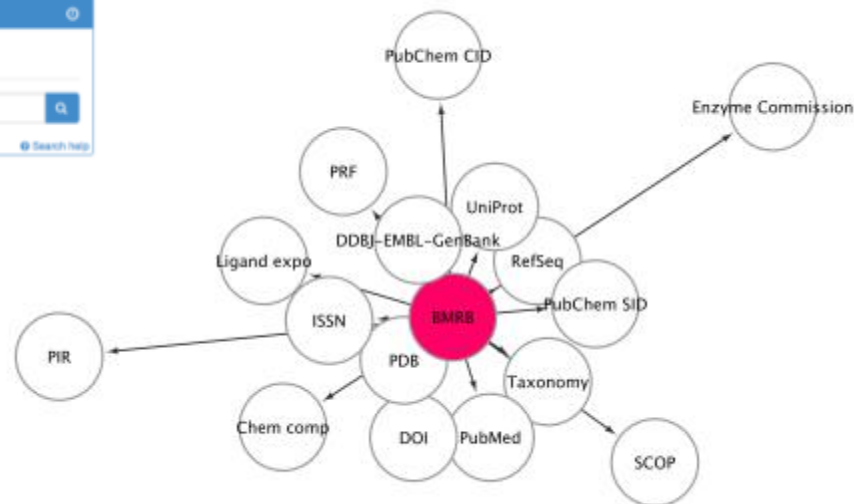
Exptl. method: SOLUTION NMR

Refine. method: simulated annealing

Data set: assigned_chemical_shifts

Chem. Shift Compl.: Sequence coverage: 97.2 %, Completeness: 72.7 %, Completeness [bb]: 73.4 % [Detail](#)

Release date: 2009-12-08



多岐に及ぶデータベースリンク

蛋白質構造データバンクの データ検証高度化と統合化

- 1) wwPDBメンバーとしてのPDBアーカイブ構築・検証・公開
- 2) 他のデータベースとの統合化および高度化
 - 2-1) CSDとの統合化
 - 2-2) ASHビューアの開発→関西学院大学藤教授より
 - 2-3) アミノ酸の立体構造中での役割の自動アノテーション
 - 2-4) 統合化されたデータベースおよび関連ツール開発
- 3) データベースの利用促進・人材育成
 - 3-1) 利用促進・アノテータの育成
 - 3-2) 国際協力
- 4) 達成達成目標と選考時に頂いた意見への対応

3) データベースの利用促進・人材育成

3-1) 利用促進・人材育成

- **アノデータ育成**: 継続的な基盤データベースとして, wwPDBメンバーとしての共通のデータ品質管理を行えるアノデータを育成
- **データ寄託者・登録者に対する講習会等**: 日本結晶学会、日本蛋白質科学会でのランチョンセミナーやブース発表等で完成度の高いデータの登録を依頼(毎年開催)
- **データ利用者に対する講習会等**: 日本生物物理学会、生命医薬情報学連合大会でランチョンセミナー&講習会を開催(毎年開催)
- **国際会議・シンポジウムの実施**: 国際運営諮問会議開催(2019年度予定), Hybrid Methodの国際集会を主催予定(2019年度)

3-2) 国際協力



PRESIDENT: Professor Jennifer Martin, Director, GRIH Institute for Drug Discovery, Griffith University, Nathan, QLD 4111, Australia. Tel: +61 3725 9058 email: j.martin@griffith.edu.au

VICE PRESIDENT: Professor Xiao-Dong Gu, Peking University, No.5 Yiheyuan Road Haidian District, Beijing, P.R. China 100871, P.R. Tel: +86 10 6275 8743 email: xdgu@pku.edu.cn

SECRETARY-TREASURER: Professor Edward R.T. Tiekink, Head Research Centre for Crystalline Materials, Faculty of Science and Technology, Sunway University, 47500 Bandar Sunway, Selangor Darul Ehsan, Malaysia. Tel: +60 3 7491 7172 email: edward@sunway.edu.my



wwpdb.org • info@wwpdb.org

Professor Genji Kurisu
Institute for Protein Research
Osaka University
3-2 Yamadaoka, Suita, Osaka 565-0871, JAPAN

January 21, 2017

Dear Professor Kurisu:

We write in strongest possible support for your application to the Japan Science and Technology Agency (JST) for competitive renewal of Protein Data Bank Japan (PDBj) funding. As a founding member of the Worldwide Protein Data Bank organization (wwpdb.org), PDBj plays an essential role in ensuring the integrity and utility of the PDB archive. The PDB archive, which serves as the sole global repository for three-dimensional structures of biological macromolecules, represents a critical resource that is freely available to researchers and educators worldwide. Since it was established in 1971 with just 7 protein structures, the PDB has grown to more than 126,000 entries. Conservative estimates put the replacement value of the information in the archive today at more than US\$12,000,000,000. During 2016, ~570 million PDB data files were downloaded or viewed online at wwpdb.org partner websites. PDB Users include structural biologists, computational biologists, biochemists, biophysicists, molecular and cellular biologists, and clinicians in universities, hospitals, government laboratories, and industry (including pharmaceutical, chemical, and biotechnology companies). PDB data are also used extensively by educators and students to further their understanding of biology and human health and disease.

In 2003, organizations from the US (RCSB PDB, Rutgers University; rcsb.org), Europe (PDBe, European Bioinformatics Institute; pdbe.org), and Japan (PDBj, Osaka University; pdbj.org) came together to establish the wwpdb.org as the global custodian of the PDB archive. These three founding members plus BMRB at the University of Wisconsin-Madison (bmrwisc.edu), which also has a portal at Osaka University (PDBj-BMRB), now work closely with an international Advisory Committee that represents stakeholder communities to ensure unrestricted access to a comprehensive biomacromolecular structure database of the highest quality. In addition, the work of the wwpdb.org is informed by a variety of special-purpose task forces, made up of experts in a variety of structural biology areas. Each partner organization is responsible for coordinated data deposition, annotation, exchange, and dissemination. In 2016 alone, the archive grew by more than 11,500 entries. The current size and growth rate of the PDB make it impossible for any one country to fund and manage the archive alone. International collaboration is of crucial importance. PDBj currently processes nearly 20% of all worldwide depositions, including all of those originating from researchers in Asia. PDBj plays a similarly important role in making the PDB archive available to the global user community.

The success of your funding application to JST is of paramount importance to the global scientific community. The wwpdb.org is an international organization supporting a globally accessible data archive, which depends critically on the success and stability of PDBj and the other member organizations. Any interruption in or reduction of PDBj funding would represent a very serious threat to the integrity and operations of the PDB archive. Any compromise of the PDB archive would in turn have extremely damaging effects on researchers and educators worldwide.

In closing, we wish to express our gratitude to you and your team at Osaka University for all of your efforts on behalf of the wwpdb.org and the global scientific community, and reiterate strongest support for your application.

Yours faithfully,

Stephen K. Burley, M.D., D.Phil.
Distinguished Professor of
Chemistry and Chemical Biology
Rutgers University
Director, RCSB PDB

Sameer Velankar, Ph.D.
Team Leader, EMBL-EBI
Head, PDBe

John L. Markley, Ph.D.
Steenbock Professor of Biomolecular
Structure
University of Wisconsin-Madison
Head, BMRB

Professor Genji Kurisu, PhD
Laboratory of Protein Crystallography,
Institute for Protein Research,
Osaka University, Japan

Dear Professor Kurisu

I am writing to you as the incoming President of the Asian Crystallographic Association, which is the regional affiliate representative for crystallographic societies and organisations in Asia and Oceania.

This letter formally expresses our sincere appreciation of the work being done at the Protein Data Bank Japan (PDBj), which has enriched the work of scientists around the world. PDBj is part of the worldwide wwpdb.org archive, together with PDBe (Europe) and RCSB-PDB (Americas). The wwpdb.org is the sole global repository of atomic coordinates and experimental diffraction data of biological macromolecules.

PDBj handles protein and nucleic acid depositions originating in Asia and Oceania regions, and currently processes 20% of the depositions worldwide. With the scientific activity accelerating dramatically in many Asian countries, the increasing workload at PDBj has been challenging. I note that handling of depositions from Australia and New Zealand have shifted to RCSB-PDB to re-balance the load across the three wwpdb.org sites. PDBj also develops and maintains open access software for the global scientific community, including chemists and biologists in academic institutes, the pharmaceutical and biotechnology industries, as well as clinicians in universities and hospitals. It is a fundamentally important data resource for basic and applied research. Moreover, it is widely used for teaching and training purposes, and for understanding health and disease. PDBj embodies the legacy of the UNESCO International Year of Crystallography (IYCr 2014) including building capacity in developing regions of the world, and further expanding the public understanding of science. Consequently, we must continue to provide our strongest possible commitment to the PDBj.

PDBj is a key Asian endeavor, bringing together the scientific output of diverse countries and peoples to serve the international community. Members of the Asian Crystallographic Association are extremely grateful to you and your team for the professional and efficient handling, checking and validating of data that is made available for everyone to use. We strongly encourage JST to provide continued financial support to maintain this essential international resource.

Kind regards

Professor Jennifer Martin
President of the Asian Crystallographic Association (2016-2019)
11 January 2017

PAST PRESIDENT: Professor Pinal CHAKRABARTI, Department of Biochemistry and Bioinformatics Centre Bose Institute, P/112 CIT Scheme IV M, Calcutta 700 075, India. Tel: +91-33-2960-3263 email: pinal@boseindia.ac.in

蛋白質構造データバンクの データ検証高度化と統合化

- 1) wwPDBメンバーとしてのPDBアーカイブ構築・検証・公開
- 2) 他のデータベースとの統合化および高度化
 - 2-1) CSDとの統合化
 - 2-2) ASHビューアの開発→関西学院大学藤教授より
 - 2-3) アミノ酸の立体構造中での役割の自動アノテーション
 - 2-4) 統合化されたデータベースおよび関連ツール開発
- 3) データベースの利用促進・人材育成
 - 3-1) 利用促進・アノテータの育成
 - 3-2) 国際協力
- 4) 達成達成目標と選考時に頂いた意見への対応

第3年時末, 第5年時末の達成目標

研究開発テーマ	タスク	小タスク	2017年度	2018年度	2019年度	2020年度	2021年度
PDBアーカイブ構築	メインアーカイブの継続的構築						
	検証ファイルのRDF化						
	CSDとの統合化						
	3D自動アノテーション						
BMRBデータ構築	管理, 登録業務の効率化	プロトタイプ					
		実装, 運用開始					
		永続化					
	XML, RDF化と応用	プロトタイプ					
		実装, Webツール公開					
		長期運用化					
	解析支援ツール	開発, デモ試用					
		民間など支援					
		長期運用化					

選考時に頂いた意見への対応

【1】Cryo-EMによって得られたデータ，1分子単位の揺らぎを含む動的データの取り扱い

Cryo-EMによって得られたデータは，既にwwPDBの枠組みにおいてEMDBとの協力体制が確立されており，現在でもPDBjのアノータがCryo-EMの3D-volume mapのデータベース化を実施していますが，今後ともさらに拡充します。また、2-3)で示した『アミノ酸の立体構造中での役割の自動アノテーション』の開発項目が可能となれば，PDBの座標データと揺らぎを含む他の動的データの統合化は可能であると考えています。PDBjで開発済みのOmokage Search (EMデータとPDBデータのクロス検索ツール)，ProMode (揺らぎ情報のデータベース) やeF-site (表面情報のデータベース) などの関連データベースを積極的に活用することで，総合的な立体構造データベースとして運営します。

具体的には，OmokageSearchには，複合体の場合に部分の「かたち」の類似性を検索できるようにして，ProModeによる基準振動によって変形した「かたち」による複合体構造の検索も試みます。また，ProModeの結果をアノートして「アミノ酸残基毎の揺らぎの大きさ」や「複数のアミノ酸間の動的性質の正と負の相関」を抽出したり，現在は青や赤の色として表示しているeF-siteによる「静電ポテンシャルの値」を抽出してアミノ酸残基毎の情報として付与したりして，これらについてもRDFによる統合化を進めたいと考えています，

選考時に頂いた意見への対応

【2】CSD以外の生物学的データ[ゲノム, 発現, 相互作用など]との統合についてどのように対応するか

これまでのプロジェクト(第二期)で, データ検証ファイル以外のセマンティック・ウェブ化が完了しており, コミュニティーの大きな財産となっています. 引き続き, DBCLSと一緒に他の生物学データとの統合化を計り, より一層の統合を目指して運営したいと考えております. すなわち, 利用者が自ら探しに行かなくても, 自動的に関連データベース(ゲノム, 発現, 相互作用など)から情報を抽出できる統合化を目指します.

具体的には, 今回提案するASHの高度化とその統合的運用によって, 最も利用者からのリクエストが多い蛋白質の立体構造およびアミノ酸配列のマルチプル・アラインメントを実施するASHのビューアを, 情報統合プラットフォームとしての機能を持たせるようにいたします. これにより, 情報統合で進められている様々なRDF化されたデータ(ゲノム, 発現, 相互作用など)をネットワークから抽出し, それらをアラインメントや立体構造に反映させようとする計画です.

選考時に頂いた意見への対応

【3】3DEMの構造データの現状のデータ検証について

3DEMによる構造解析のデータは、現状では立体化学的妥当性のみを検証し、実験データに基づいた検証は行われていません。しかし、wwPDBでは出来るだけ早く実験データ(マップ)に基づいた検証を導入する必要があると考えています。専門家によるTFを組織し、マップを表すパラメーター(voxel size, density statistics, projections, central slices, surface view at recommended contour level), や一致度を示す指標(map-density distribution, volume estimate vs. contour level, rotationally averaged power spectrum, depositor-provided Fourier Shell Correlation curve), などに基づいた検証の実装に向けて作業を進めています。