Licensed under a Creative Commons表示4.0国際ライセンス ©2018 横地政志（大阪大学 蛋白質研究所）

PDBアーカイブの検証レポートのRDF化

横地政志[1]、金城玲[1]、藤原敏道[1]、中村春木[1,2]、栗栖源嗣[1]

[1]大阪大学 蛋白質研究所、[2]国立遺伝学研究所

発表要旨

PDBアーカイブの検証レポート（wwPDB validation reports）は、X線結晶回折、核磁気共鳴、電子顕微鏡の構造決定法毎の専門委員会（Validation Task Force）の意見に基づいて、広く受け入れられている評価基準を使って、PDBに登録された立体構造の品質を評価する文書です。特に論文の投稿・査読過程における利用が推奨されています。

PDBjは、検証レポートに含まれる評価指標の検索を容易にするため、機械処理しやすい検証レポートの代替フォーマットによるアーカイブ（PDBx/mmCIFと高い互換性のあるXML版とRDF版）を開発しました。現在、新しいアーカイブを利用して、検証レポートのリレーショナルデータベース（PostgreSQL）およびSPARQLエンドポイントを公開しています。これらは立体構造に基づくバイオインフォマティクス、創薬の分野において活用が期待されます。

wwPDB検証レポート（PDF）

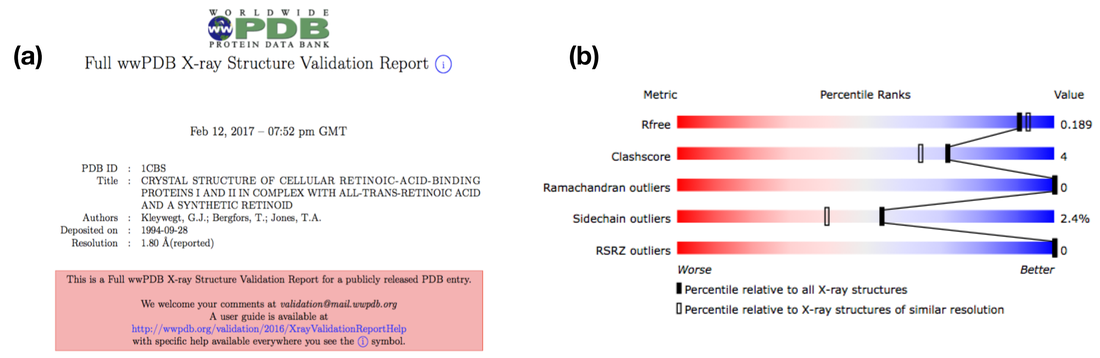


Fig.1 (a) wwPDB検証レポート（PDF版）の表紙。 **(b)** wwPDB検証レポートの最初のページには、主な検証項目に関して、全PDBエントリーを基準にした相対評価がパーセンタイルで表示されていて、注目している構造の品質の概要を知ることができます。パーセンタイル表示の青いほうが評価が良好であることを表します。塗りつぶした四角形は、全エントリーと比較した相対評価、穴あきの四角は同手法、同精度で決定された全エントリーと比較した相対評価です。

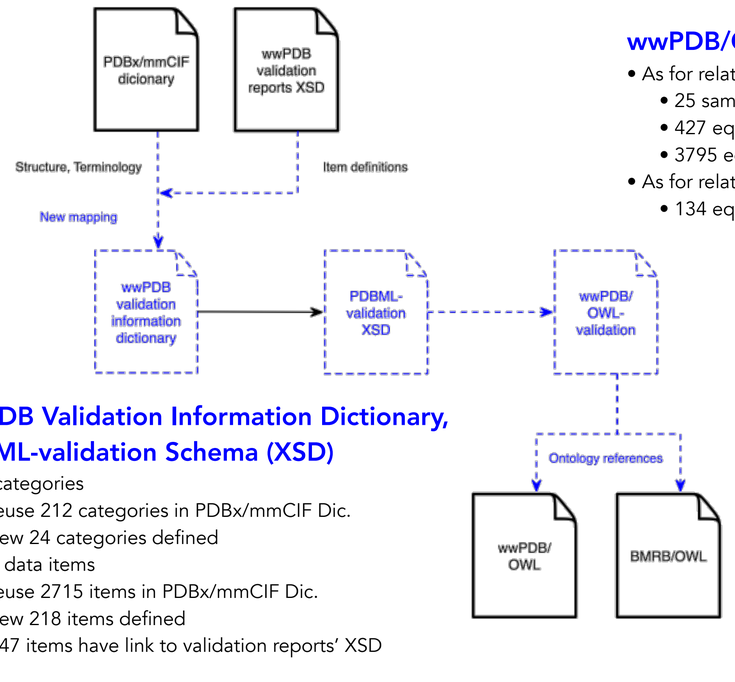
wwPDB検証レポートの主な評価指標

X-ray/EM/NMR
- Geometric & conformational
  - bond, angle, planarity
  - protein backbone conformation
  - protein side-chain conformation
- Atomic & molecular interaction
  - all-atom contacts
  - under packing
  - hydrogen bond quality
- Non-protein
  - nucleic acids (RNA pucker, suite)
  - carbohydrates (N-glycan core)
  - ligands (CSD)
  - ions & other solvent
- Incomplete model (e.g. CA_ONLY)

X-ray
- Structure factor & electron density
  - Wilson plot outliers, tNCS
  - wrong space group
  - twinning
  - agreement (R_{free}, RSR, RSCC)

NMR
- Chemical shifts
  - completeness
  - outliers
  - estimated reference error
  - random coil index
- Structure ensembles
  - representative model (medoid)
  - domain detection

代替wwPDB検証レポートのオントロジー作成スキーム



wwPDB Validation Information Dictionary, PDBML-validation Schema (XSD)
- 236 categories
  - reuse 212 categories in PDBx/mmCIF Dic.
  - new 24 categories defined
- 2922 data items
  - reuse 2715 items in PDBx/mmCIF Dic.
  - new 218 items defined
  - 447 items have link to validation reports' XSD

wwPDB/OWL-validation
- As for relation with wwPDB/OWL
  - 25 same classes
  - 427 equivalent classes
  - 3795 equivalent properties
- As for relation with BMRB/OWL
  - 134 equivalent properties

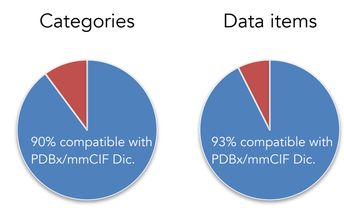


Fig.2 wwPDB検証レポートの定義を、PDBx/mmCIF辞書の文書構造、命名則で書き直します。新たにwwPDB validation information dictionaryを作り、PDBML互換の検証レポートのスキーマ、さらにRDF版のオントロジーを作成します。検証レポートの移植のため、新しく24のカテゴリーと、218のアイテムを辞書に追加しました。既存のPDBx/mmCIF辞書で記述可能なアイテムは、約200個存在しました。最終的に、補完した方が良い定義を加えると、PDBx/mmCIF辞書と90%以上の互換性がある、検証レポート専用の辞書を作成しました。

セマンティック拡張された検証レポート作成と応用

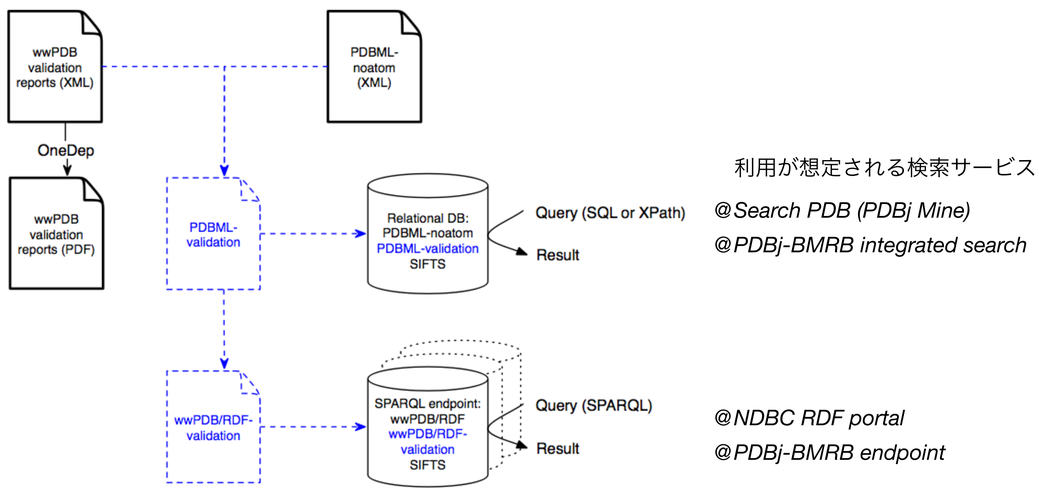


Fig.3 既存のXML版のwwPDB検証レポートとPDBMLを編集して、PDBML互換のwwPDB検証レポート、PDBML-valiationを作成しました。さらに、RDF版のwwPDB検証レポート、wwPDB/RDF-validationを作成しました。得られた２つのアーカイブの一方は、Relationalデータベース化して、SQLで検索可能です。RDF版は、SPARQLエンドポイントに入れて他のデータベースへ連携させることが可能になります。

セマンティック拡張された検証レポートの特徴

Archive name	wwPDB validation reports (PDF)	wwPDB validation reports (XML)	PDBML-validation	wwPDB/RDF-validation
Human-readability	yes	-	-	-
File extension	###_validation.pdf	###_validation.xml	###_validation_full.xml	###_validation.rdf
Metadata (author, entity, …)	partial	-	yes (from PDBML)	yes (from PDBML)
Validation information	summary	full	full	full
Searchable	no	yes (XQuery)	yes (SQL, XQuery)	yes (SPARQL)
PDBx/mmCIF	-	-	~90% compatible	~90% compatible
URI	-	-	-	supported
Purpose	**peer review**	**data exchange**	**data exchange & quick search**	**knowledge sharing**

wwPDB/RDF-validationのSPARQL検索　例１

検索：リガンドのRSR (Real space R-factor、残基毎に実験的に求めた電子密度と構造モデルの電子密度の残差の割合) が10%より小さい全ての酵素-リガンド複合体を選択

```
PREFIX PDBov: <https://rdf.wwpdb.org/schema/pdbx-validation-v1.owl#>

SELECT ?PDB_ID ?enzyme ?ligand ?comp_id MIN(?RSR AS ?minRSR)
FROM <https://rdf.wwpdb.org/pdb-validation>
WHERE {

 ?entity PDBov:link_to_enzyme ?link_to_enzyme ;
         PDBov:entity.pdbx_description ?enzyme ;
         PDBov:of_datablock ?datablock .                          ✍ selection of enzyme

 BIND (SUBSTR(STR(?datablock),38,4) AS ?PDB_ID)

 BIND (IRI(CONCAT(?datablock, "/pdbx_entity_nonpolyCategory")) AS ?entity_nonpoly_category)

 ?entity_nonpoly_category PDBov:has_pdbx_entity_nonpoly ?entity_nonpoly .

 ?entity_nonpoly PDBov:pdbx_entity_nonpoly.name ?ligand ;
                 PDBov:pdbx_entity_nonpoly.entity_id ?entity_id ;
                 PDBov:pdbx_entity_nonpoly.comp_id ?comp_id .     ✍ ligand selection: non-polymer, not water, not ion

 FILTER (?ligand!="water" && !STRENDS(?ligand, " ION"))

 BIND (IRI(CONCAT(?datablock, "/pdbx_nonpoly_schemeCategory")) AS ?nonpoly_scheme_category)

 ?nonpoly_scheme_category PDBov:has_pdbx_nonpoly_scheme ?nonpoly_scheme .

 ?nonpoly_scheme PDBov:pdbx_nonpoly_scheme.pdb_strand_id ?asym_id ;
                 PDBov:pdbx_nonpoly_scheme.pdb_seq_num ?seq_id ;
                 PDBov:pdbx_nonpoly_scheme.entity_id ?entity_id ;
                 PDBov:pdbx_nonpoly_scheme.mon_id ?comp_id .

 BIND (IRI(CONCAT(?datablock, "/pdbx_dcc_mapCategory")) AS ?dcc_map_category)

 ?dcc_map_category PDBov:has_pdbx_dcc_map ?dcc_map .

 ?dcc_map PDBov:pdbx_dcc_map.auth_asym_id ?asym_id ;
          PDBov:pdbx_dcc_map.auth_comp_id ?comp_id ;
          PDBov:pdbx_dcc_map.RSR ?RSR .

 FILTER (xsd:float(?RSR) < 0.1)                                   ✍ RSR < 0.1

} GROUP BY ?PDB_ID ?enzyme ?ligand ?comp_id
```

結果：約15000件の酵素-リガンド複合体の組み合わせが存在

代替検証レポートのダウンロード、開発状況、ご意見はこちらへ、

https://github.com/yokochi47/pdbx-validation

wwPDB/RDF-validationのSPARQL検索　例２

検索：蛋白質配列中のRSRの異常値　(RSRZ>2、残基毎のRSRのZスコアが2σより大きい) の割合が1%より小さい全ての酵素-リガンド複合体を選択

```
PREFIX PDBov: <https://rdf.wwpdb.org/schema/pdbx-validation-v1.owl#>

SELECT ?PDB_ID ?enzyme (GROUP_CONCAT(?ligand; SEPARATOR=",") AS ?ligands) ?RSRZ_outliers_percent
FROM <https://rdf.wwpdb.org/pdb-validation>
WHERE {

 ?map_overall PDBov:pdbx_dcc_map_overall.entry_id ?PDB_ID ;
              PDBov:pdbx_dcc_map_overall.RSRZ_outliers_percent ?RSRZ_outliers_percent .

 FILTER (xsd:float(?RSRZ_outliers_percent) < 0.1)                 ✍ % of outliers in RSRZ < 1%

 BIND (IRI(CONCAT("https://rdf.wwpdb.org/pdb-validation/", ?PDB_ID, "/entityCategory")) AS ?entity_category)

 ?entity_category PDBov:has_entity ?entity .

 ?entity PDBov:link_to_enzyme ?link_to_enzyme ;
         PDBov:entity.pdbx_description ?enzyme .                  ✍ selection of enzyme

 BIND (IRI(CONCAT("https://rdf.wwpdb.org/pdb-validation/", ?PDB_ID, "/pdbx_entity_nonpolyCategory")) AS ?entity_nonpoly_category)

 ?entity_nonpoly_category PDBov:has_pdbx_entity_nonpoly ?entity_nonpoly .

 ?entity_nonpoly PDBov:pdbx_entity_nonpoly.name ?ligand .         ✍ ligand selection: non-polymer, not water, not ion

 FILTER (?ligand!="water" && !STRENDS(?ligand, " ION"))

}
```

結果：約5000件の酵素-リガンド複合体の組み合わせ、0%に制限した場合は約1000件の組み合わせが存在

まとめ

バルクデータとして検索が容易で、PDBx/mmCIFと高い互換性のある検証レポートの拡張を提案しました。現在、PDBML-validation，wwPDB/RDF-validation及びPostgreSQLデータベースのスナップショットが入手できます。今後は、wwPDBによる承認と公開、PDBj Mine、PDBj-BMRBの検索サービスと統合を進めていく予定です。