

Deliverable D7.7

Deliverable D6.3

Project Title:	World-wide E-infrastructure for structural biology	
Project Acronym:	West-Life	
Grant agreement no.:	675858	
Deliverable title:	Assembly queries	
Lead Beneficiary:	EMBL	
Contractual delivery date:	Month 32	
Actual delivery date:	Month 32	
WP No.	6	
WP Title	Data management	
WP leader:	Thomas Kulhanek	CSIC
Contributing partners:	EMBL-EBI	

Deliverable written by Nurul Nadzirin

Work Package 6, Deliverable 6.3: Assembly queries

Deliverable D6.3	1
Work Package 6, Deliverable 6.3: Assembly queries	2
Macromolecular assembly queries	3
Technology	5
PDBe Advanced Search System.....	6
CAPRI Search System.....	8
References	11

Deliverable D7.7

Executive Summary

This deliverable describes the functionality of the query system provided by the PDBe (Mir, et al., 2018) to search for different criteria important for macromolecular assemblies available in the PDB and EMDB data archives, as well as value added data on complexes at PDBe.

It also reports the development of a search system for predicted model complexes based on the results of previous CAPRI experiments (Janin, et al., 2003). Taken together, the aim is to offer a comprehensive search system of macromolecular complexes comprising of both experimentally established and predicted structures. Some of these complexes fall into the realm of integrative/hybrid methods, as they integrate experimental data from multiple methods. In this report, these query systems will be referred to as the PDBe Advanced Search system, and the CAPRI Search system.

Macromolecular assembly queries

Macromolecular complexes are of central interest in structural biology (Bertoni, et al., 2017). Most, if not all, biological processes depend upon the ability of proteins to assemble in complexes (Tramontano, 2017). The function of proteins in the cell is almost always mediated by their interaction with different partners, including other proteins, nucleic acids or small organic molecules. Proteins acquire oligomeric organization for a variety of functional and biophysical advantages: modular elements are less prone to coding errors, oligomeric regulation add an additional level of control, large structures are more stable and can perform their function cooperatively (Goodsell and Olson, 2000).

The large majority of protein interactions are specific both in terms of their partners and of the regions of interaction. It is therefore important to be able to identify the interaction partners of a target protein and, given two interacting proteins, identifying the precise molecular details of their interaction, that is, building the three-dimensional model of the complex. The number of ways proteins interact in nature is probably limited, and it has been observed that similar binding modes can be identified for almost all known protein-protein interactions (Kundrotas, et al., 2012). The ability to provide an accounting of assembly types would allow the discovery of patterns in different criteria making up the knowledge of these.

A macromolecule can have one or multiple partners and form. The simplest form of a protein assembly is a *dimer*, consisting of two polypeptide chains. Protein-protein interactions can occur either between proteins made by identical polypeptide chains (homo-) or proteins made by different polypeptide chains (hetero-). Furthermore, a protein can interact with different types of macromolecule, either with another protein, RNA, DNA, carbohydrates and many types of ligands.

There are two broad aims for this deliverable. The first aim is to describe the functionality of the query system provided by the PDBe (Mir, et al., 2018) to search for different criteria important for macromolecular assemblies available in the PDB and EMDB data archives, as well as value added data on complexes at PDBe. The second aim is to build a search system for predicted model complexes based on the results of previous CAPRI experiments (Janin, et al., 2003). Taken together, the aim is to offer a comprehensive search system of macromolecular complexes comprising of both experimentally established and predicted structures. Some of these complexes fall into the realm of integrative/hybrid methods, as they integrate experimental data from multiple methods. In this report, these query systems will be referred to as the PDBe Advanced Search system, and the CAPRI Search system.

Technology

The backbone of the query system for both the PDBe Advanced Search system and the CAPRI Search system is based on Apache Solr. Apache Solr is a web application built around Lucene (<http://lucene.apache.org/solr/>), which facilitates the searching, filtering, as well as providing an easy platform to directly produce XML/HTTP and JSON APIs. Users can query directly based on a RESTful API, which makes integration into any web platform or application straightforward. The front-end application for both systems generate the vocabularies necessary to build specific query strings, which are then used to make HTTP GET Solr requests.

One of the most useful features of Solr is the faceted search, which has become a critical feature for enhancing find-ability and the user search experience for all types of search applications. Faceted search is the dynamic clustering of items or search results into categories that let users drill into search results by any value in any field. Each facet displayed also shows the number of hits within the search that match that category. Users can then “drill down” by applying specific constraints to the search results.

Apache Solr also allows an easy way for data indexing and loading, which can be done through a web administration interface. In addition, Solr allows range queries, which allows easy grouping of parameters to be queried, which would be particularly useful for queries involving numbers.

PDBe Advanced Search System

The advanced search interface for querying PDB and EMDB data archives can be found at the URL <https://www.ebi.ac.uk/pdbe/entry/search/>. This interface allows users to search for either an individual entry of interest, or a list of entries based on the criteria provided by the user. On the left hand panel are parameters chosen by the users to retrieve the results, which are shown on the right hand panel (see Figure 1).

The screenshot displays the PDBe Advanced Search System interface. At the top, there is a navigation bar with links to EMBL-EBI, Services, Research, Training, and About us. The main header features the PDBe logo and a search bar with the example text 'Ex. - hemoglobin, BRCA1_HUMAN'. Below the search bar, there are tabs for 'Entries', 'Macromolecules', 'Compounds', and 'Protein families'. The 'Macromolecules' tab is selected, showing a list of results. The first result is 'Protein: HR(MLZ)VLR' with a '3h6z' PDB ID. The description is 'Crystal Structure of the Four MBT Repeats of Drosophila melanogaster Smbt in Complex with Peptide RHR (me)K VLR'. The release date is 'Released: 16 Jun 2009'. The resolution is '2.8Å resolution'. The model geometry and fit model/data are shown with color-coded bars. The assembly composition is 'protein/protein complex'. The interface also includes a left sidebar with filters like 'Latest PDB release', 'New UniProt in PDB', and 'New ligands in PDB'. A 'Download' button is visible in the top right corner.

Figure 1: The PDBe Advanced Search system

The parameters associated with macromolecular assemblies queries are shown in Figure 2. Through this interface, macromolecular complexes can be queried based on assembly composition. Some of these compositions are “protein-protein complex”, “RNA/protein complex” and “DNA/protein complex”. In addition, the search system also allows grouping and faceting based on the assembly polymer count and whether the assembly is a homo/hetero assembly. This allows users to search for entries with assembly compositions such as hetero dimer, homo trimer, tetramer, and so on.

Deliverable D7.7

— Homo / hetero assembly (2)	
homo	(110602)
hetero	(30989)
— Assembly composition (24)	
protein structure	(108115)
protein/protein complex	(23307)
DNA/protein complex	(4312)
RNA/protein complex	(2233)
DNA structure	(1728)
RNA structure	(1064)
— Assembly polymer count (149)	
monomer	(62162)
dimer	(44471)
tetramer	(12779)
trimer	(9894)
hexamer	(4052)
octamer	(1352)

Figure 2: Faceted search on the PDBe Advanced Search interface

CAPRI Search System

Since its inception in 2001, the CAPRI (Critical Assessment of PRedicted Interactions) experiment has conducted 45 rounds comprising of 135 targets made up of protein-protein (which also includes protein-peptide experiments as this typically constitute a different kind of modelling challenge), protein-RNA and protein-polysaccharide complexes. The bulk of CAPRI data, most important being the results for each target, had never been made available for public searching before.

For this deliverable, a parser was written to extract data from several flat files stored in the EBI server. Some manual curation and checking were done before the parsed data were organized in a JSON schema and loaded into Solr (Figure 3).

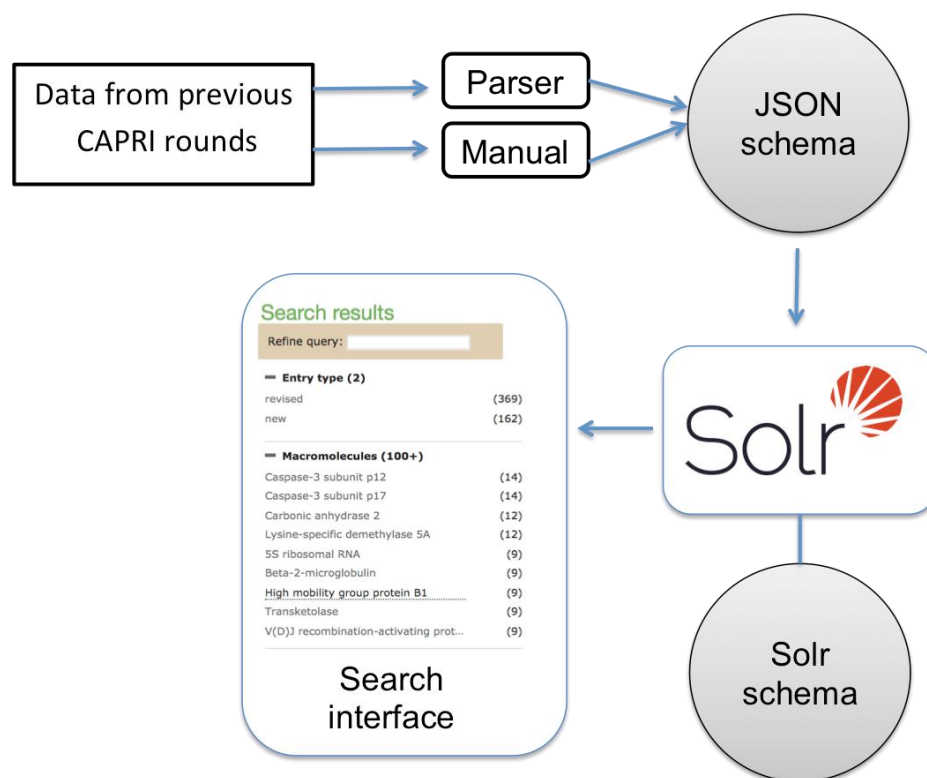


Figure 3: Steps taken in the development of the CAPRI Search system

The parameters for CAPRI search were determined based on a survey, which was documented in West-Life Milestone 6.4. This survey was done to gain the input and feedback from several members of the CAPRI Management Committee as well as several previous participants of CAPRI challenges. The parameters make up the JSON key in the documents loaded into Solr (Figure 4), which includes `assembly_composition`, `oligomeric_state` (assembly polymer count), and `homo_hetero_assembly`. These parameters, shown as facets in the user interface, can be seen in Figure 5.

Deliverable D7.7

Request-Handler (qt)

— common —
q

fq

sort

start, rows

fl

df

Raw Query Parameters

wt

[http://wp-np2-8c.ebi.ac.uk:8080/solr/capri-loading/select?q=*. *](http://wp-np2-8c.ebi.ac.uk:8080/solr/capri-loading/select?q=*.)

```

{
  "responseHeader":{
    "status":0,
    "QTime":1,
    "params":{
      "q": "*. ",
      "_: "1530283271702"}},
  "response":{ "numFound":17557, "start":0, "docs":[
    {
      "I_rmsdbb": [9.907],
      "I_rmsdsc": [11.636],
      "L_rmsd": [22.972],
      "M_rmsd": [1.968],
      "assembly_composition": [ "protein/protein complex" ],
      "average": [13.58],
      "clash_threshold": [61.2],
      "clashes": [21],
      "classification": [ "incorrect" ],
      "distance": [18.573],
      "engineered": [ "yes" ],
      "fIR": [0.225],
      "fIR_2": [0.581],
      "fOP": [0.82],
      "fOP_2": [0.49],
      "file_brk": [ "https://www.ebi.ac.uk/msd-srv/capri/orig/capri_33_103.brk" ],

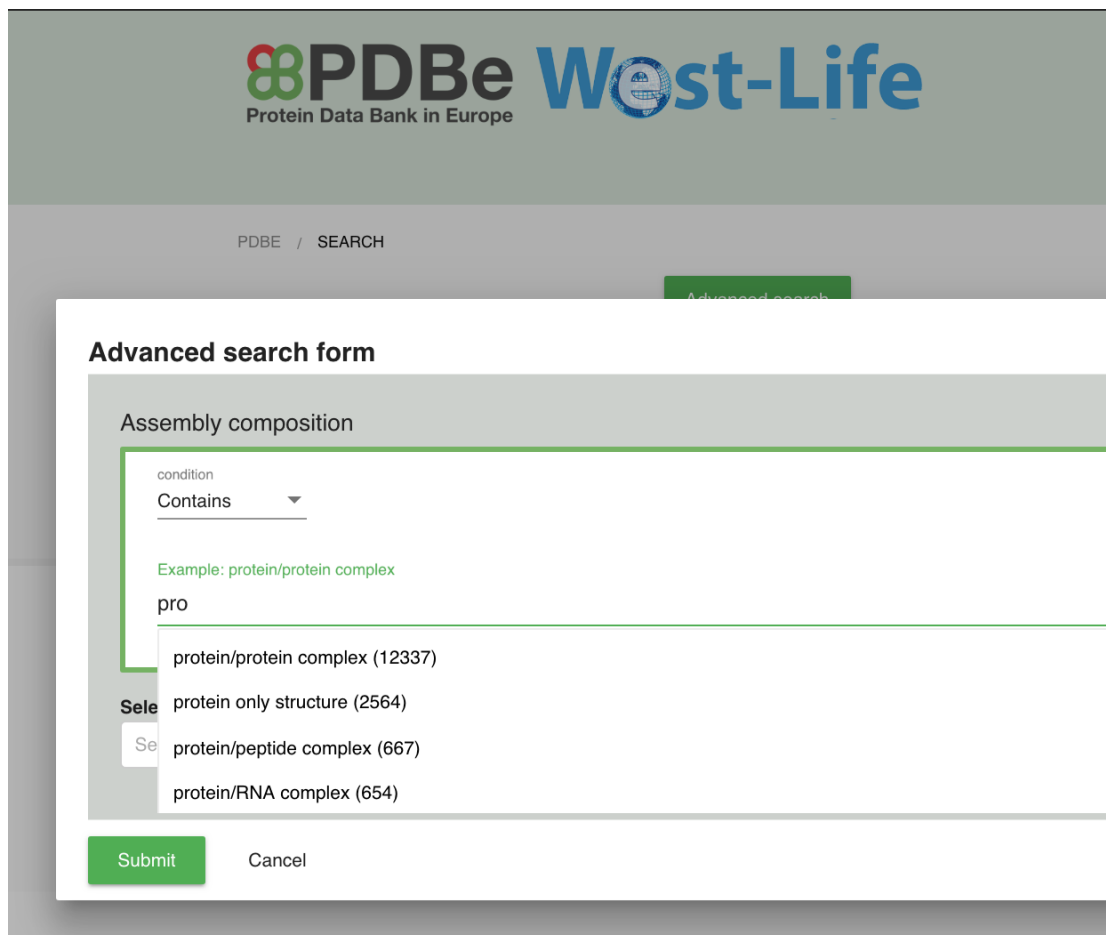
```

Figure 4: Solr admin interface showing some of the parameters in CAPRI search system

CAPRI related	Assessment information per target
+ Round (25)	+ Clash threshold
+ Target (46)	+ Standard deviation
+ Modelling type (8)	+ Average of number of clashes
Target information	Assessment information per model
+ Organism (23)	+ f(nat)
+ Assembly composition (7)	+ L-rmsd
+ Homo/hetero assembly (1)	+ I-rmsd (backbone-only fit)
+ Assembly polymer count (3)	+ I-rmsd (side chain-only fit)
+ Molecule name (80)	+ f(non-nat)
+ Published structure (37)	+ fIR (ligand)
+ Target contributor (26)	+ fIR (receptor)
+ Engineered (1)	+ fOP (ligand)
	+ fOP (receptor)
Participant information	+ IA
+ Group (193)	+ d(L)
Model information	+ nclash
+ Model classification (12)	+ theta(L)

Figure 5: Faceting for the CAPRI search system

Subsequently, data configurations were made to allow the same components developed in the PDBe Advanced Search interface to be used for this system. The Advanced Search form allows users to filter results based on the facets they're interested in. For example, in Figure 6, a user can choose from a drop-down menu to search for entries with specific assembly compositions. The drop-down menu is generated as an autocomplete feature in the Advanced Search form.



The screenshot shows the PDBe West-Life Advanced search form. At the top, the PDBe logo (Protein Data Bank in Europe) and West-Life logo are displayed. Below the logos, the breadcrumb "PDBe / SEARCH" is visible. The main section is titled "Advanced search form". Under the "Assembly composition" heading, there is a dropdown menu labeled "condition" with "Contains" selected. Below the dropdown, an example "protein/protein complex" is shown. The text "pro" is entered in the search field, and a list of suggestions is displayed: "protein/protein complex (12337)", "protein only structure (2564)", "protein/peptide complex (667)", and "protein/RNA complex (654)". To the left of the suggestions, there are labels "Sele" and "Se". At the bottom of the form, there are "Submit" and "Cancel" buttons.

Figure 6: The Advanced search form showing an example for assembly composition query

The search system is, at the time of writing, available for testing on a development server at the URL <https://wwwdev.ebi.ac.uk/pdbe/widgets/capri-search/>.

References

- Bertoni, M., *et al.* Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci Rep* 2017;7(1):10480.
- Goodsell, D.S. and Olson, A.J. Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct* 2000;29:105-153.
- Janin, J., *et al.* CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* 2003;52(1):2-9.
- Kundrotas, P.J., *et al.* Templates are available to model nearly all complexes of structurally characterized proteins. *Proceedings of the National Academy of Sciences of the United States of America* 2012;109(24):9438-9441.
- Mir, S., *et al.* PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. *Nucleic acids research* 2018;46(D1):D486-D492.
- Tramontano, A. The computational prediction of protein assemblies. *Curr Opin Struct Biol* 2017;46:170-175.