Database-driven discovery

John Berrisford



Summary

- Searching the PDB
- Advanced search
- Searching the PDB programmatically







PDB over the years



In May 2014, we passed the 100,000 milestone As of today, ~150,000 structures



At the heart of EMBL-EBI resources

Genes, Genomes & Variation



Searching the PDB

The problem with the PDB is...

- I can't find what I want
- Too many false hits when I search
 - CaM, CaM-Kinase, CaM binding protein, Cam-like...
- Results are too complicated
 - Which lysozyme to use? 47 species!
 - Which is best?/what is best? >250 ligands
- Redundancy
 - How many unique human protein structures?







Searching the PDB made simple

At PDBe we've implemented:

- Auto-complete suggestions
- Facets to narrow down search results
- Quality presented on search results
- Four views of results
 - Moving away from entry-centricity





Protein Data Bank in Europe Enging Structure to Biology Description Structure to Biology Operation Operation Operation of Description Operation Operatio	EMBL-EBI		Service	Research Training About u
PDBe beind Pope de training Documentation About PDBe Abou	Protein Data Bank in Bringing Structure to Biology	Europe	Examples: hemoglobin, BRCA1_HUMAN	Search <u>EMsearch</u>
PDBe is the European resource for the collection, organisation and dissemination of data on biological macromolecular structures. Read more about PDBe. Featured structure Stop motion: The muscular system Ist May 2018 Image for May nour 2018 calendar captures a molecular snapshot of one of our primary groups of organs - the muscular system. Here we discuss the individual molecular snapshot of one of our primary groups of organs - the muscular system. Here we discuss the individual molecular snapshot of one of our primary groups of organs - the muscular system. Here we discuss the individual molecular snapshot of one of our primary groups of organs - the muscular system. Here we discuss the individual molecular snapshot of one of our molecular snapshot of one	PDBe home Deposition PDBe services PDBe training	Documentation About PDBe		< Share Share Seedbac
Featured structure Stop motion: The muscular system Ist May 2018 Image for May in our 2018 calendar captures a molecular system. Here we discuss the individual molecules responsible for the large motions of these muscles. Image for May in our 2018 calendar captures a molecular system. Here we discuss the individual molecules responsible for the large motions of these muscles. Image for May in our 2018 calendar captures a molecular system. Here we discuss the individual molecules responsible for the large motions of these muscles. Image for May in our 2018 calendar captures a molecular system. Here we discuss the individual molecules responsible for the large motions of these muscles. Image for May in our 2018 calendar captures a molecular system. Here we discuss the individual molecules responsible for the large motions of these muscles. Read more Previous featured structures Read more Previous featured structures Image for May in our 2018 calendar captures at the muscular system. Here we discuss the individual molecules responsible for the large motions of these muscles. Read more Previous featured structures Image for May in our 2018 calendar captures at the muscular system. Here we discuss the individual molecules responsible for the large motions of these muscles. News David Blow Poster Prize awarded at BCA 13 April, 2018 Indiversity of Cambridge, Cambridge, Dix 2018 PDBe Explores Art 13 April, 2018 PDBe Explores Art 29 March, 2018 20 March, 2018 20 March, 2018 20 March, 2018	PDBe is the European resource for the collection, organisation structures. Read more about PDBe.	and dissemination of data on biological macromolecular	Popular	
Previous featured structures Latest archive statistics News Events David Blow Poster Prize awarded at BCA University of Cambridge Protein Structure 13 April, 2018 University of Cambridge, Cambridge, UK PDBe Explores Art 24 May 2018 29 March, 2018 University of Cambridge, Cambridge, UK Analysis course Tweets by @PDBeurope March, 2018 Three Dimensional Electron Microscopy -	Featured structure Stop motion: The muscular system Image for May in our 20 primary groups of organs - indecules responsible for the melecules resp	1st May 2018 118 calendar captures a molecular snapshot of one of our the muscular system. Here we discuss the individual e large motions of these muscles.	 EMsearch PDBeFold PDBePISA PDBeChem Sequence search PDBe REST API EM resources NMR resources EMPIAR Coordinate Server PDB Component Library 	 News Events Training Contact us
News Events David Blow Poster Prize awarded at BCA University of Cambridge Protein Structure Analysis course University of Cambridge, VK Analysis course Itest to poster Prize awarded at BCA 13 April, 2018 University of Cambridge, Cambridge, UK 24 May 2018 Itest to poster Prize awarded at BCA Itest to poster Prize awarded at BCA 29 March, 2018 University of Cambridge, Cambridge, UK 24 May 2018 Itest to poster Prize awarded at Bank Itest to poster Prize awarded at BCA 29 March, 2018 Three Dimensional Electron Microscopy - Itest to poster Prize awarded treat pain. This structure, How snail toxins could treat	Previous featured structures		Latest archive statistics	S contains 140591 entries
University of Cambridge, Cambridge, UK Tweets by @PDBeurope Image: Cambridge, UK 29 March, 2018 24 May 2018 Protein Data Bank Image: Cambridge, UK 29 March, 2018 Three Dimensional Electron Microscopy - How snail toxins could treat pain. This structure,	News David Blow Poster Prize awarded at BCA	Events University of Cambridge Protein Structur Analysis course	(latest PDB entries, chemist contains 6215 entries (lates releases, latest updates).	try, <u>biology</u>) and EMDB st map releases, latest header
29 March, 2018 Protein Data Bank Protein Data Ba	PDBe Explores Art	University of Cambridge, Cambridge, UK 24 May 2018	Tweets by @PDBeurope	0
How snail toxins could treat pain. This structure,	29 March, 2018		Protein Data Bank @PDBeurope	¥ ^
	Access the data of Maltalation Demonts for	Three Dimensional Electron Microscopy	- How snail toxins could	d treat pain. This structure,



pdbe.org

EMBL-EBI

Services Research Training About us

1 W.Y 383

Protein Data Bank in Europe

hydro

And a	A A PAN	2
	Search]

Enzyme	
Hydrolases	(21099)
Glycosidases, i.e. enzymes hydrolyz	(3634)
Hydro-lyases	(1683)
Phosphoric monoester hydrolases	(1543)
Carboxylic ester hydrolases	(1038)
Carbonic acid hydro-lyase (carbon	(845)
Beta-lactam hydrolase	(537)
Phosphoric diester hydrolases	(499)
Hydrolyzing N-glycosyl compounds	(465)
3.4.19.12 : Ubiquitinyl hydrolase 1	(450)
	More

GO mapping	
GO:0016787 : hydrolase activity	(16113)
GO:0016798 : hydrolase activity, acti	(3749)
GO:0004553 : hydrolase activity, hyd	(2235)
GO:0090305 : nucleic acid phosph	(1925)
GO:0090502 : RNA phosphodiester	(775)
GO:0042542 : response to hydroge	(652)
GO:0090501 : RNA phosphodiester	(603)
GO:0042744 : hydrogen peroxide c	(590)
GO:0016788 : hydrolase activity, acti	(482)
GO:0052689 : carboxylic ester hydr	(479)
	More

Journal		Ligand	
int J Hydrogen Energy	(5)	NAG : N-[(2R,3R,4R,5S,6R)-6-(hydr	(6384)
		PEG : DI(HYDROXYETHYL)ETHER	(2501)
		PEG : 2-(2-hydroxyethyloxy)ethanol	(2501)
		MAN : (2S,3S,4S,5S,6R)-6-(hydroxy	(2087)
		BMA : (2R,3S,4S,5S,6R)-6-(hydroxy	(2012)
		FAD : [(2R,3S,4R,5R)-5-(6-aminopu	(1980)
		ADP : [(2R,3S,4R,5R)-5-(6-aminopu	(1892)
		NAD : [(2R,3S,4R,5R)-5-(3-aminoca	(1239)
		NAP : [(2R,3S,4R,5R)-5-(3-aminoca	(1181)
		TRS : 2-AMINO-2-HYDROXYMETHY	(1155)
			More

Molecule name	
Soluble epoxide hydrolase	(108)
Bifunctional epoxide hydrolase 2	(104)
Cytosolic epoxide hydrolase 2	(104)
Pyrophosphate phospho-hydrolase	(96)
4-hydroxy-tetrahydrodipicolinate syn	(73)

Organism	
Hydrozoa	(365)
Arthrobacter Hydrocarboglutamicus	(68)
Aeromonas hydrophila	(42)
Carboxydothermus hydrogenoform	(34)
Aeromonas Hydrophila	(28)

Sequence family	
IPR027417 : P-loop containing nucl	(4780)
IPR017853 : Glycoside hydrolase s	(2385)
CL0058 : Glyco_hydro_tim	(2139)
IPR029058 : Alpha/Beta hydrolase f	(1862)
CL0028 : AB_hydrolase	(1763)

Structure domain	
Methane Monooxygenase Hydroxyla	(1278)
P-loop containing nucleoside tripho	(1196)
alpha/beta-Hydrolases	(524)
Phosphorylase/hydrolase-like	(285)
Glycosyl hydrolase domain	(284)



Enzyme	
Hydrolases	(21099)
Glycosidases, i.e. enzymes hydrolyz	(3634)
Hydro-lyases	(1683)
Phosphoric monoester hydrolases	(1543)
Carboxylic ester hydrolases	(1038)
Carbonic acid hydro-lyase (carbon	(845)
Beta-lactam hydrolase	(537)
Phosphoric diester hydrolases	(499)
Hydrolyzing N-glycosyl compounds	(465)
3.4.19.12 : Ubiquitinyl hydrolase 1	(450)
	More



pdbe.org

Drill down with facets





Drill down with facets





- Search for queries by typing or scrolling
- Select a query to add it to your search form



 Begin typing to get autocomplete suggestions for that query

🚓 EMBL-EBI 🔦 Services 🛛 🍪 Research 📥 Training 🌖 About us	EMBL-EBI	Hinxton 👻
88PDBe	Ex hemoglobin, BRCA1_HUMAN	Search
Protein Data Bank in Europe	Examples: hemoglobin, BRCA1_HUMAN Advar	nced search
		Feedback
Advanced search form	×	
PDBE / SEARC Molecule name	^	
condition Contains Example: Carbonic anhydrase 2 hemoglobin bemoglobin	×	
seta bemoglobin subunit alpha (326)		
EMBL-E	~	-EBI
Submit Cancel		



Can also search using a protein sequence via Phmmer





☐ 5hph Structure of TRAP1	1 fragment	X-ray diffraction
		2.429Å resolution
Sung N, Chang C, Lee S, Tsai FTF		Released: 10 Aug 2016
Acta Crystallogr D Struct Biol (2016) [PN	AID: 27487821 ⊯ 1	Model geometry
		Fit model/data
Source organism: Homo sapiens e		
Assembly composition: protein only str	ructure	
Bound ligands: ANP SO4 MG PO	O4_GOL_	
	nload files	
Accuracy: 0.9	Identity percentage: 48.24%	
Bit score: 145.11	Score: 146.1	
E-value: 1.6e-41	Similarity count: 124	
Identity count: 82	Similarity percentage: 72.94%	
Chain ID: A		
Sequence Alignment:		
Query: FAFQAEIAQLMSLIINTFY	YSNKEIFLRELISNSSDALDKIRYESLTDPSKLDS	SGKELHINLIPNKQDRTLTIVDTGIGMTKADLINNL
Match: FQAE +L+ ++ ++YS KE+F+RELISN+SDAL+K+R++ ++D L E++I+L N + T+TI DTGIGMT+ +L++NLC		
Target: HEFQAETKKLLDIVARSLY	YSEKEVFIRELISNASDALEKLRHKLVSDGQALP-	EMEIHLQTNAEKGTITIQDTGIGMTQEELVSNL(
۲		>
Click here to check for other PDB ent	ries with identical sequence	



LiteMol and Downloads on search



Click here to check for other PDB entries with identical sequence



Programmatic access to the PDB

• What information can you get?





Showing sequence annotation at PDBe.org

- Component for displaying sequence annotation
- Uses API to get this information





LiteMol – powered by PDBe REST API

- New, lightweight structure viewer at PDBe
- Uses API to get structure and annotation data
 - Including residue-level validation information



"author_insertion_code": "", "author_residue_number": 120, "alt_code": "", "outlier_types": ["* "clashes"

"author_insertion_code": "", "author_residue_number": 156, "alt_code": "", "outlier_types": [,* "bond_lengths", "planes"

"author_insertion_code": "", "author_residue_number": 191, "alt_code": "", "outlier_types": ["* "clashes", "bond_lengths", "bond_angles", "sidechain_outliers"



Searching the PDB programmatically

 It is possible to search the PDB programmatically and get access to the data in the PDB archive



RESTful API for PDBe data

- Allows access to vast amounts of PDBe data
- Powers the PDBe search and web pages
 - Will always be well maintained!

EMBL-EBI	Service	s Research Training About us
Protein Data Bank in Europe Bringing Structure to Biology	Examples: hemoglobin, BRCA1_HUMAN	Search Q EMsearch
PDBe home Deposition PDBe services PDBe training Documentation About PDBe	ATT De	< Share 🔍 Feedback
PDBe is the European resource for the collection, organisation and dissemination of data on biological macromolecular structures. Read more about PDBe.	Popular	
	EMsearch	News
Featured structure	A PDBeFold	Events
	N PDBePISA	🗏 Training
Focusing on Crystallin	Sequence search	The Contact us
Ist June 2017 As you read the words on this page, the light enters your eyes and the eye lens focuses it on the retina. From here the image is sent to your brain for processing. The major protein from the eye lens is shown here in June's featured structure from	PDBe REST API	
	NMP resources	
	EMDIAR	
	Coordinate Server	
	PDB Component Library	
our <u>2017 calendar</u> .		



RESTful API for PDBe data

- Allows access to vast amounts of PDBe data
- Powers the PDBe search and web pages
 - Will always be well maintained!

EMBL-EBI	Services Research	Training About us
Protein Data Bank in Europe		Search
Bringing Structure to Biology	Examples: hemoglobin, BRCA1_HUMAN	R EMsearch
PDBe home Deposition PDBe services PDBe training Documentation About PDBe	< sh	are 👒 Feedback
PDBe is the European resource for the collection, organisation and dissemination of data on biological macromolecular structures. Read more about PDBe.	Popular	
	EMsearch News	
Featured structure	sequence search	
Focusing on Crystallin		
	PDBE REST API	
As you read the words on this page, the light enters your eyes and the	EM recourses	
The major protein from the eye lens is shown here in June's featured structure from our 2017 calendar.	Coordinate Server	

PDBe.org/API

EMBL-EBI

What can the REST API give you?

Split into 8 groups



- These REST calls are based on different subsets of data
 - These are grouped for ease of use
- These are available on the documentation page
 - Thorough set of example calls
 - Give URL and output JSON on the page



PDBe web search is based on the API

- Initial queries typed into autocomplete search bar
- Can then refine search by additional queries
 - These are the facets on the left-hand side
- URL contains the queries involved in the search
 - Directly corresponding to API calls





Search in the documentation page

PDBe REST API Programmatic access to PDBe data PDB COMPOUNDS EMDB SIFTS PISA VALIDATION TOPOLOGY SEARCH	Share Seedback
REST calls related to PDBe search service	Version 1.5 Show/Hide All Calls
 Search on Solr instance based on polymeric entities http://www.ebi.ac.uk/pdbe/search/pdb/select?:query A document in this boin instance represents a polyment entry of type protein, DNA, I Output from the call depends on the query sent to Solr. Query parameters are well do Each document has a wide range of properties, grouped as follows: Basic information about the entry pdb_id : PDB entry id code. number_of_polymer_entities : Number of unique polymers in the entry. number_of_bound_entities : Number of unique bound molecules (excluding wat onumber_of_polymers : Number of polymer chains in the entry. number_of_polymer_residues : Number of bound molecules in the entry. number_of_polymer_residues : Number of polymer residues in all polymers. entry_authors : List of depositors who deposited this entry. all_authors : Combined list of depositors and authors of primary citation. title : Title of the entry as provided by the depositor. revision_date : Date of latest revision of the entry. deposition_date : Date when entry was deposited - this is a timestamp in for deposition_year : Year of deposition of the entry. status : Release state of the entry, e.g. REL, OBS etc. release_date : Date of release of the entry - this is a timestamp in format YY release_year : Year of release of the entry. 	s in the PDB. VA or sugar. umented in Solr documentation. er) in the entry. nat YYYY-MM-DDThh:mm:ssZ. hat YYYY-MM-DDThh:mm:ssZ. (Y-MM-DDThh:mm:ssZ.
PD	Be.org/API EMBL <u>-EBI</u>

Search queries in the documentation

- Search on Solr instance based on polymeric entities in the PDB.

http://www.ebi.ac.uk/pdbe/search/pdb/select?:query

0

A document in this Solr instance represents a polymeric entity of type protein, DNA, RNA or sugar. Output from the call depends on the query sent to Solr. Query parameters are well documented in Solr documentation. Each document has a wide range of properties, grouped as follows:

o pdb_id : PDB entry id code.

entry.

o number_of_bound_entities : Number of unique bound molecules (excluding water) in the entry.

o number_of_polymers : Number of polymer chains in the entry.

o number_of_bound_molecules : Number of bound molecules in the entry.

o number_of_polymer_residues : Number of polymer residues in all polymers.

o entry_authors : List of depositors who deposited this entry.

• all_authors : Combined list of depositors and authors of primary citation.

o title : Title of the entry as provided by the depositor.

• revision_date : Date of latest revision of the entry - this is a timestamp in format YYYY-MM-DDThh:mm:ssZ.

o revision_year : Year of latest revision of the entry.

o deposition_year : Year of deposition of the entry.

release_date : Date of release of the entry - this is a timestamp in format YYYY-MM-DDThh:mm:ssZ.

release_year : Year of release of the entry.

o number_of_protein_chains : Number of protein chains in the entry.

o number_of_DNA_chains : Number of DNA chains in the entry.

o number_of_RNA_chains : Number of RNA chains in the entry.



Search in the documentation page

String Options string allowed by Solr syntax. Details about constructing Solr queries can be found from webpages such as this.
postdata String Send one or more query options in post data instead of appending to URL.
Quotes RunCall Select Expand Collapse 2+ 3+
GET : http://www.ebi.ac.uk/pdbe/search/pdb/select?q=molecule_name:"Dihydrofolate reductase" AND organism_scientific_name:"Homo sapiens"&wt=json
HTTP status : 200 : OK (Hover to find undocumented bits in the output.)
{s*
"responseHeader": {
"status": 0,
"QTime": 0,
"params": {
"q": "molecule_name:\"Dihydrofolate reductase\" AND organism_scientific_name:\"Homo sapiens\"",
"wt": "ison"
}
h
"response": { *
"numFound": 75,
"start": 0,
"docs": [_*
{s*
"all_authors": [,*
"Borhani DW"
"Campbell SR",
"Dixon RW"
"Fairchild RG"
"Kisliuk RL",
"Leung AK".



Practical example of how to get data from the PDB

- For this tutorial we will use Jupyter notebooks to use the PDBe search API
- <u>https://github.com/PDBeurope/pdbe-api-training</u>

