Bioinformatyka 2

ZBOiB WBBiB UJ
Adrian Kania

28 października 2025

NCBI

1. **NCBI** – National Center for Biotechnology Information
2. **NCBI Entrez** – zintegrowany system ponad 30 baz danych gromadzących informację z nauk biomedycznych, głównie w postaci sekwencji i literatury naukowej. Obejmuje m.in Pubmed (streszczenia publikacji), Nucleotide (sekwencje nukleotydowe), Protein (sekwencje aminokwasowe).
3. **Entrez Programming Utilities** – oprogramowanie umożliwiające dostęp do danych zintegrowanych w ramach systemu Entrez bez konieczności obsługi formularzy na stronach WWW
   1. **ESearch** – wyszukiwanie identyfikatorów/kodów dostępu rekordów wybranej bazy danych,
   2. **EFetch** – pobieranie rekordów o wskazanych identyfikatorach lub wyników wskazanego wyszukiwania

NCBI

NCBI Resources How To

NCBI
National Center for Biotechnology Information

All Databases | Search

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

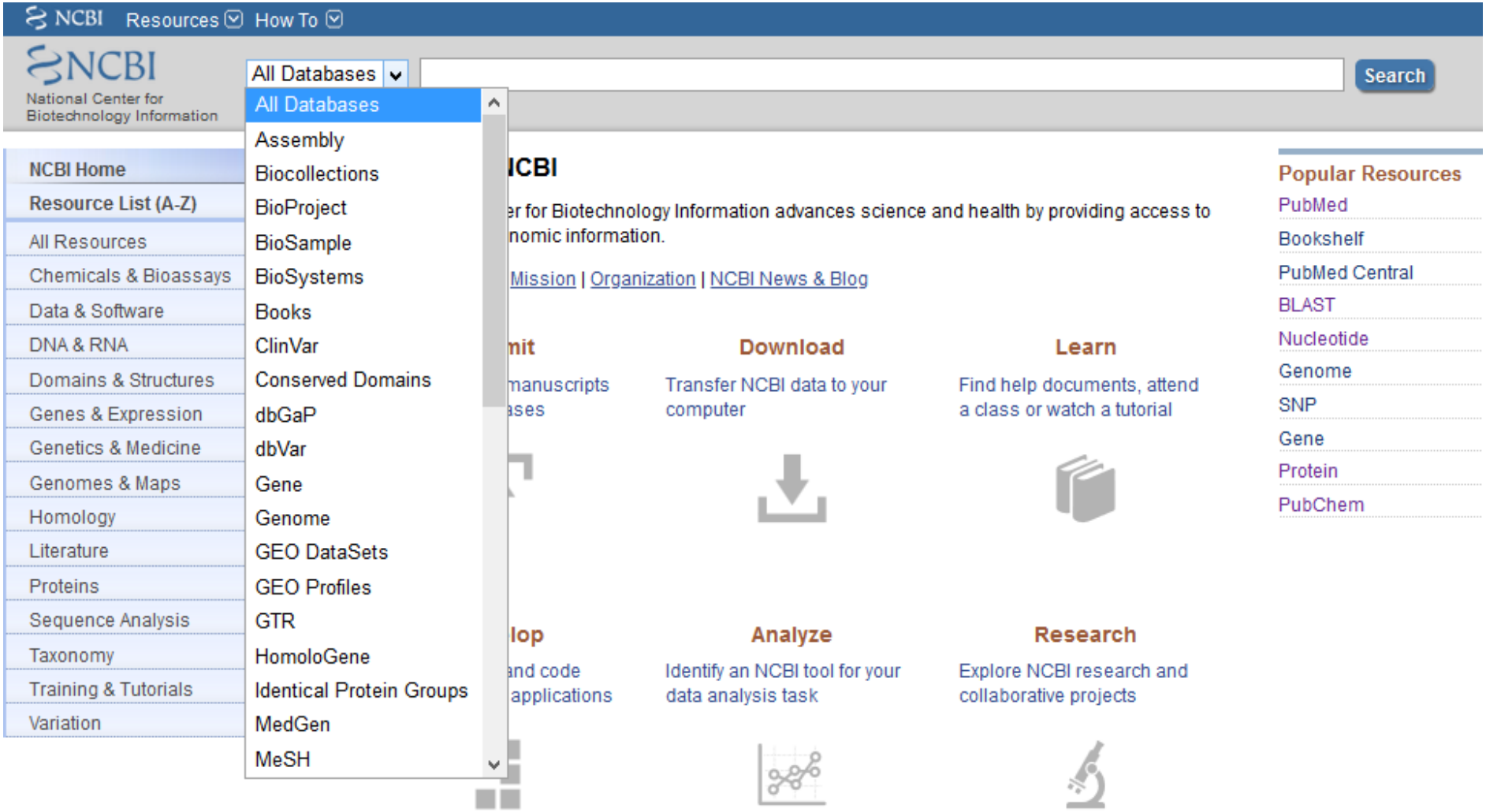
Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI


NCBI Resources How To
NCBI
National Center for Biotechnology Information
All Databases
Search
All Databases
Assembly
Biocollections
BioProject
BioSample
BioSystems
Books
ClinVar
Conserved Domains
dbGaP
dbVar
Gene
Genome
GEO DataSets
GEO Profiles
GTR
HomoloGene
Identical Protein Groups
MedGen
MeSH
NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
Mission | Organization | NCBI News & Blog
Download
Transfer NCBI data to your computer
Learn
Find help documents, attend a class or watch a tutorial
Analyze
Identify an NCBI tool for your data analysis task
Research
Explore NCBI research and collaborative projects
Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

EFetch

EFetch

EFetch (efetch.fcgi) returns full data records for a list of unique identifiers (UIDs) in a format specified in the parameters. The list of UIDs is either provided in the parameters, or is retrieved from the History server.

EFetch Parameters

EFetch Required Parameters

- `db` (required): Database containing the unique identifiers (UIDs) for which you wish to retrieve records. You can see NCBI's table of Entrez Unique Identifiers (UIDs) for a complete list of allowable database names, but some example values include:
  - `pubmed`: PubMed
  - `pmc`: PubMed Central
  - `nlmcatalog`: NLM Catalog
- `id` (required): Either a single unique identifier (UID) or a comma-delimited list of UIDs. All of the UIDs must be from the database specified by the `db` parameter.

EFetch Optional Parameters

- `retstart` (optional): Setting this parameter helps limit which records will be shown in the output, as it determines whether the record for the first input unique identifier (UID) is retrieved, or whether to skip to a later UID in the input list. For example, if `retstart` is set to `10`, the output will begin with the record for the tenth UID. The default of this parameter is `1`, corresponding to the first UID in the input list. This parameter can be used in conjunction with `retmax` to download an arbitrary subset of records.
- `retmax` (optional): Total number of records to be shown in the output, up to a maximum of 10,000. If the set of records you are trying to retrieve is larger than 10,000, you can submit multiple EFetch requests, and increase the `retstart` parameter each time.
- `retmode`/`rettype`: These two parameters determine how your results will be displayed. `retmode` determines the data format your records will be returned in (e.g. XML, plain text, etc.). `rettype` determines the specific view your records will be returned in (e.g. MEDLINE, Abstract, list of PMIDs, etc.). Different databases have different allowable data formats and record views, and not all `retmode` data formats are compatible with all `rettype` record views, and vice versa.

  The table below shows the allowable combinations of `retmode` and `rettype` for some of the databases. **Bold** `retmode` values are the default data format for the specified database. **Bold** `rettype` parameters are the default record view for the specified data format and database.

EFetch Examples

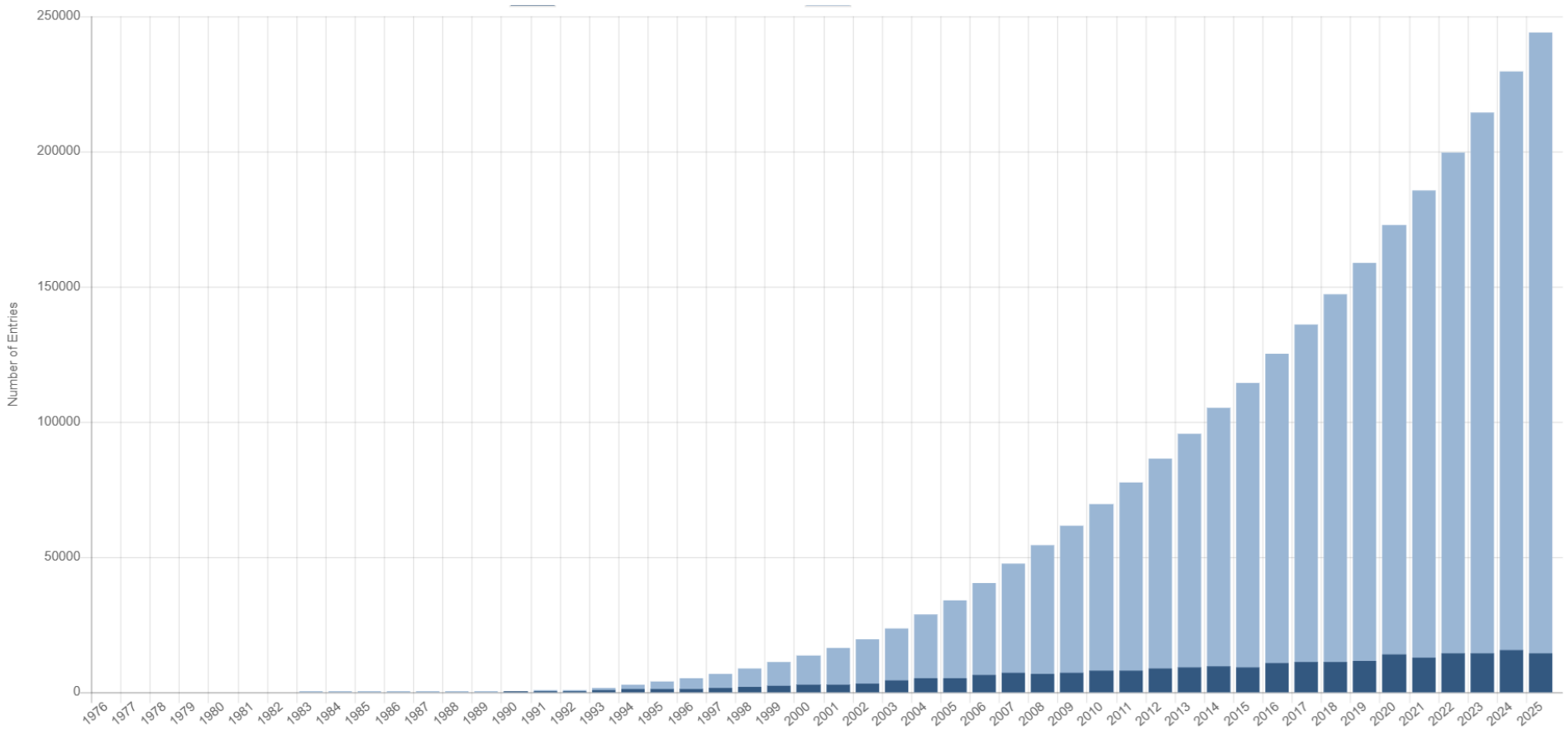
- Retrieve the abstract view (text format) of two PubMed records.
  - https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=17284678,9997&retmode=text&rettype=abstract
- Retrieve two PubMed records in XML format.
  - https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=11748933,11700088&retmode=xml

ESearch

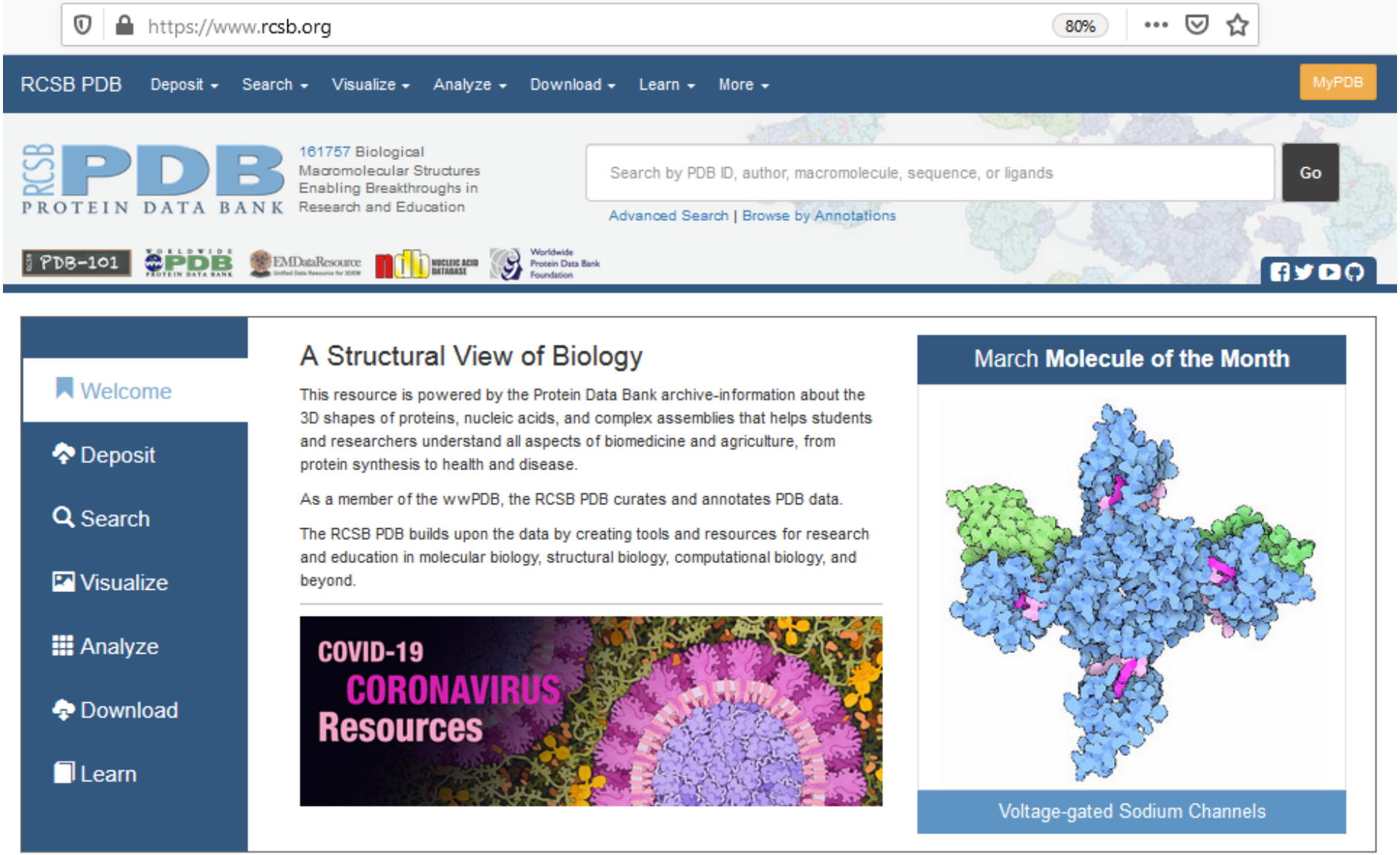
ESearch Formatting Parameters

- `retstart` (optional): Setting this parameter helps limit which of the unique identifiers (UIDs) in the results set will be shown in the output, as it determines whether the output begins at the first retrieved UID, or with a UID that is later in the results set. For example, if `retstart` is set to `10`, the first ten UIDs in the results set will be skipped, and the output will begin with the eleventh UID. The default of this parameter is `0`, corresponding to the first record in the entire set. This parameter can be used in conjunction with `retmax` to download an arbitrary subset of UIDs retrieved from a search.
- `retmax` (optional): Total number of unique identifiers (UIDs) from the retrieved set to be shown in the output (default=20). Increasing `retmax` allows more of the retrieved UIDs to be included in the output, up to a maximum of 100,000 UIDs. If you need to retrieve more than 100,000 UIDs, you can submit multiple ESearch requests, and increase the `retstart` parameter each time. For example:
  - https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=cancer&retmax=100: This URL will return results 1 through 100 of a search for "Cancer".
  - https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=cancer&retstart=100&retmax=100: This URL will return results 101 through 200 of the same search for "Cancer".
- `rettype` (optional): Retrieval type. There are two supported values:
  - `uilist` (default): Displays the standard XML output, including a list of unique identifiers (UIDs), the total number of results, and the query translation for the search.
  - `count`: Displays only the total number of results, without the list of UIDs or query translation.
- `retmode` (optional): Determines the format of the returned output. The default value is `xml`, but `json` is also supported.

PDB


Number of Entries
250000
200000
150000
100000
50000
0
1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025

PDB


https://www.rcsb.org
80%
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB
RCSB PDB
PROTEIN DATA BANK
161757 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education
Search by PDB ID, author, macromolecule, sequence, or ligands
Go
Advanced Search | Browse by Annotations
PDB-101
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
A Structural View of Biology
This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.
As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.
The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.
COVID-19
CORONAVIRUS
Resources
March Molecule of the Month
Voltage-gated Sodium Channels

The PDB file – text format

Atom number
Atom identity
Residue identity
chain
Residue number
The coordinates for each residue in the structure
x y z

ATOM:

Usually protein or DNA

ATOM	2	CA	GLY	A	672	54.168	8.340	69.707	1.00104.94	C
ATOM	3	C	GLY	A	672	52.692	8.194	69.380	1.00105.46	C
ATOM	4	O	GLY	A	672	51.877	9.045	69.750	1.00108.67	O
ATOM	5	N	GLU	A	673	52.359	7.101	68.691	1.00102.41	N
ATOM	6	CA	GLU	A	673	50.994	6.785	68.274	1.00 89.17	C
ATOM	7	C	GLU	A	673	50.624	5.325	68.585	1.00 81.77	C
ATOM	8	O	GLU	A	673	51.438	4.411	68.405	1.00 81.88	O
ATOM	9	CB	GLU	A	673	50.850	7.050	66.777	1.00 96.53	C
ATOM	10	CG	GLU	A	673	50.252	8.399	66.438	1.00 99.19	C
ATOM	11	CD	GLU	A	673	48.788	8.486	66.827	1.00115.45	C
ATOM	12	OE1	GLU	A	673	48.062	7.477	66.681	1.00116.71	O
ATOM	13	OE2	GLU	A	673	48.356	9.561	67.286	1.00113.58	O
ATOM	14	N	ALA	A	674	49.387	5.109	69.023	1.00 67.27	N
ATOM	15	CA	ALA	A	674	48.912	3.768	69.370	1.00 63.11	C
ATOM	16	C	ALA	A	674	48.702	2.826	68.174	1.00 58.54	C
ATOM	17	O	ALA	A	674	48.064	3.183	67.186	1.00 62.02	O
ATOM	18	CB	ALA	A	674	47.616	3.866	70.189	1.00 47.04	C
ATOM	19	N	PRO	A	675	49.260	1.612	68.240	1.00 55.66	N
ATOM	20	CA	PRO	A	675	49.087	0.665	67.134	1.00 52.95	C
ATOM	21	C	PRO	A	675	47.629	0.261	66.997	1.00 48.19	C

HETATM:

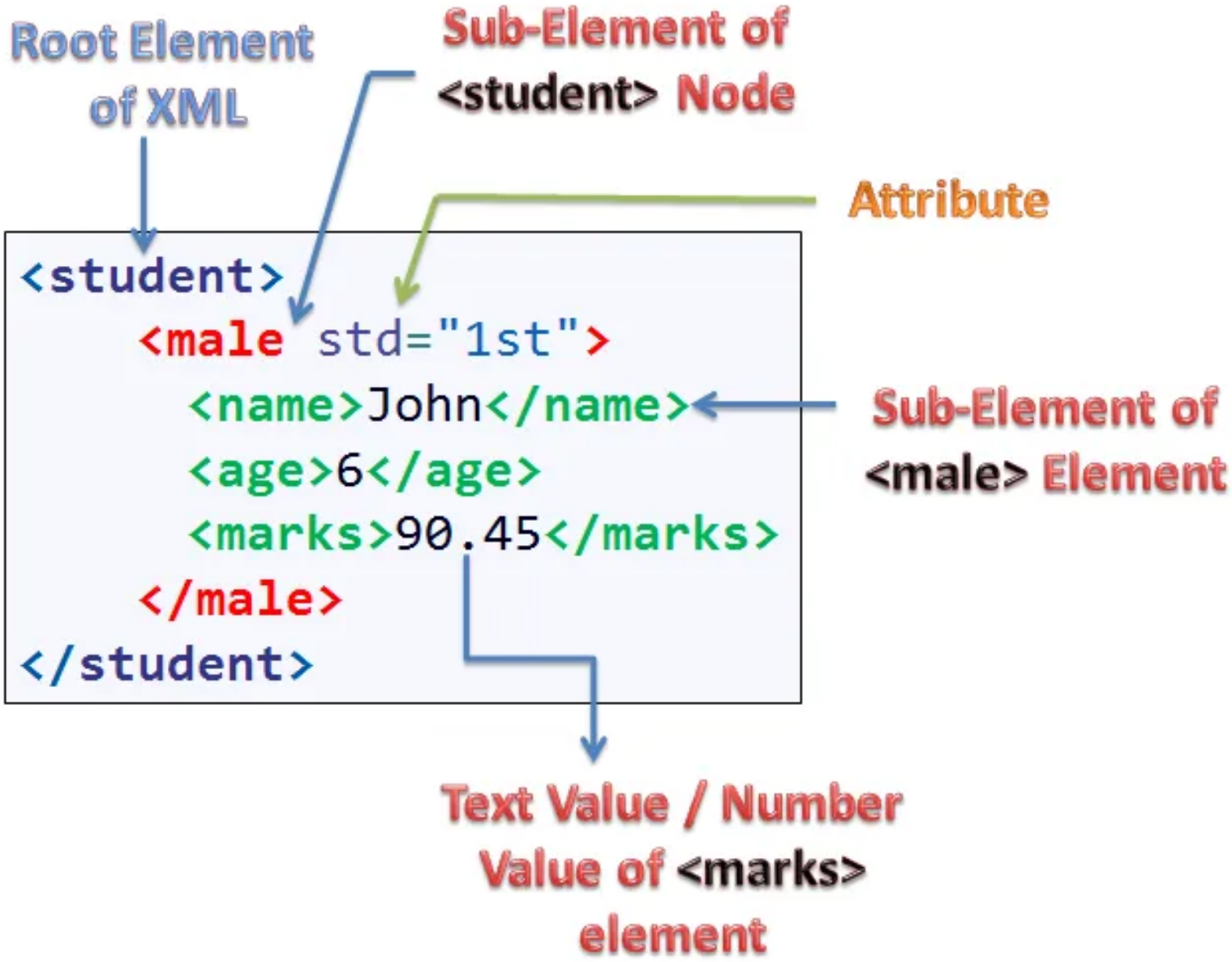
Usually Ligand, ion, water

HETATM	2517	C5	AQ4		774	25.725	0.972	53.258	1.00 75.72	C
HETATM	2518	N1	AQ4		774	24.289	0.712	53.215	1.00 63.33	N
HETATM	2519	C6	AQ4		774	23.410	-0.217	53.908	1.00 56.92	C
HETATM	2520	C7	AQ4		774	22.037	-0.309	53.572	1.00 52.23	C
HETATM	2521	C8	AQ4		774	21.501	0.476	52.546	1.00 48.18	C
HETATM	2522	C9	AQ4		774	20.143	0.376	52.218	1.00 52.11	C
HETATM	2523	O1	AQ4		774	19.589	1.220	51.120	1.00 82.48	O
HETATM	2524	C10	AQ4		774	20.550	1.362	50.041	1.00 83.98	C
HETATM	2525	C11	AQ4		774	20.235	2.645	49.262	1.00 91.80	C

XML

1. **XML** – Extensible Markup Language – rozszerzalny język znakowania przeznaczony do reprezentowania różnych danych w ustrukturalizowany sposób
   1. **znacznik** – ciąg liter, cyfr i innych znaków (bez spacji i zaczynający się od litery) ograniczony znakami $<$ i $>$. Tagowi (znacznikowi) otwierającemu zawsze towarzyszy tag zamykający zaczynający się od znaków $</$ i kończący się znakiem $>$.
   2. **atrybut** – definiowany w obrębie znacznika, składa się zwykle pary *nazwa* $=$ *wartość*, przy czym wartość zwykle podawana jest w cudzysłowach,
   3. **encja** – jest elementem treści dokumentu i reprezentuje pojedynczy symbol, encje definiuje się podając po znaku & numer symbolu i kończąc znakiem średnika, np & lt;

XML


Root Element of XML
Sub-Element of <student> Node
Attribute
<student>
<male std="1st">
<name>John</name>
<age>6</age>
<marks>90.45</marks>
</male>
</student>
Sub-Element of <male> Element
Text Value / Number Value of <marks> element

XML

```
<eInfoResult>
  <DbInfo>
    <DbName>pubmed</DbName>
    <MenuName>PubMed</MenuName>
    <Description>PubMed bibliographic record</Description>
    <Count>22595116</Count>
    <LastUpdate>2013/03/19 03:50</LastUpdate>
    <FieldList>
      <Field>
        <Name>ALL</Name>
        <FullName>All Fields</FullName>
        <Description>All terms from all searchable fields</Description>
        <TermCount>139835112</TermCount>
        <IsDate>N</IsDate>
        <IsNumerical>N</IsNumerical>
        <SingleToken>N</SingleToken>
        <Hierarchy>N</Hierarchy>
        <IsHidden>N</IsHidden>
      </Field>
      <Field>
      ...
</eInfoResult>
```

Wektory 2D - podstawowe operacje

Niech dane są $v = (1, 2)$ oraz $w = (4, 0)$. Wtedy:

1. **suma wekorów** v i w wynosi:
$$v + w = (1 + 4, 2 + 0) = (5, 2),$$
2. **iloczyn skalarny wektorów** v i w wynosi:
$$< v, w >= 1 \cdot 4 + 2 \cdot 0 = 4 + 0 = 4$$
3. **długość wektora** v wynosi:
$$||v|| = \sqrt{< v, v >} = \sqrt{1 \cdot 1 + 2 \cdot 2} = \sqrt{5}.$$

Wektory 3D - podstawowe operacje

Niech dane są $v = (1, 2, 3)$ oraz $w = (4, 0, -2)$. Wtedy:

1. **suma wekorów** v i w wynosi:
$$v + w = (1 + 4, 2 + 0, 3 + (-2)) = (5, 2, 1),$$
2. **iloczyn skalarny wektorów** v i w wynosi:
$$< v, w >= 1 \cdot 4 + 2 \cdot 0 + 3 \cdot (-2) = 4 + 0 - 6 = -2,$$
3. **długość wektora** v wynosi:
$$||v|| = \sqrt{< v, v >} = \sqrt{1 \cdot 1 + 2 \cdot 2 + 3 \cdot 3} = \sqrt{14}.$$

Wektory ND - podstawowe operacje

Niech dane są $v = (v_1, v_2, ..., v_n)$ oraz $w = (w_1, w_2, ..., w_n)$. Wtedy:

1. **suma wekorów** v i w wynosi:
$$v + w = (v_1 + w_1, v_2 + w_2, ..., v_n + w_n),$$
2. **iloczyn skalarny wektorów** v i w wynosi:
$$< v, w >= v_1 w_1 + v_2 w_2 + ... + v_n w_n,$$
3. **długość wektora** v wynosi:
$$||v|| = \sqrt{< v, v >} = \sqrt{v_1 v_1 + v_2 v_2 + ... + v_n v_n} = \sqrt{v_1^2 + v_2^2 + ... + v_n^2}.$$

Przestrzenie wektorowe a analiza tekstu

Zdanie1: new york times
Zdanie2: new york post
Zdanie3: los angeles times

Przestrzenie wektorowe a analiza tekstu

Dokument1: new york times
Dokument2: new york post
Dokument3: los angeles times

Występujące słowa : angeles, los, new, post, times, york

Przestrzenie wektorowe a analiza tekstu

Dokument1: new york times
Dokument2: new york post
Dokument3: los angeles times

Występujące słowa (w ilu tekstach): angeles (1), los (1), new (2), post (1), times (2), york (2)

IDF (Inverse Document Frequency)

Liczba zdań $N = 3$.

Występujące słowa (w ilu tekstach): angeles (1), los (1), new (2), post (1), times (2), york (2)

$$idf_i = \log_2(\frac{N}{df_i})$$

gdzie df_i - liczba dokumentów zawierających i-te słowo. Wtedy:

- angles, $\log_2(\frac{3}{1}) = 1.584$,
- los, $\log_2(\frac{3}{1}) = 1.584$,
- new, $\log_2(\frac{3}{2}) = 0.584$,
- post, $\log_2(\frac{3}{1}) = 1.584$,
- times, $\log_2(\frac{3}{2}) = 0.584$,
- york, $\log_2(\frac{3}{2}) = 0.584$.

TF (Term Frequency)

Niech:

$$f_{ij} \text{ - częstość } i \text{ - tego słowa w dokumencie } j.$$

Wtedy:

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{ij}\}}.$$

W naszym przypadku:

	angeles	los	new	post	times	york
d1	0	0	1	0	1	1
d2	0	0	1	1	0	1
d3	1	1	0	0	1	0

Ważenie TF-IDF

Definiujemy:

$$w_{ij} = tf_{ij} idf_i = tf_{ij} \log_2(\frac{N}{df_i}).$$

Uwaga:

- słowa występujące często w jakimś dokumencie, ale rzadko w pozostałych dokumentach będą miały wysoką wagę
- eksperymentalnie takie ważenie się sprawdza

Ważenie TF-IDF

Definiujemy:

$$w_{ij} = tf_{ij} idf_i = tf_{ij} \log_2(\frac{N}{df_i}).$$

W naszym przypadku:

	angeles	los	new	post	times	york
d1	0	0	0.584	0	0.584	0.584
d2	0	0	0.584	1.584	0	0.584
d3	1.584	1.584	0	0	0.584	0

Zapytanie

Powiedzmy, że mamy nowy dokument:

new new times

Pytanie: Który dokument najbardziej przypomina?

Zapytanie

Powiedzmy, że mamy nowy dokument:

new new times

Pytanie: Który dokument najbardziej przypomina? Należy obliczyć ważenie $TD - IDF$ dla zapytania.

Zapytanie

Powiedzmy, że mamy nowy dokument:

new new times

Pytanie: Który dokument najbardziej przypomina? Należy obliczyć ważenie $TD - IDF$ dla zapytania.

q	0	0	(2/2)*0.584=0.584	0	(1/2)*0.584=0.292	0

Wyznaczanie długości wektorów

Mamy:

- $||d1|| = \sqrt{0.584^2 + 0.584^2 + 0.584^2} = 1.011,$
- $||d2|| = \sqrt{0.584^2 + 1.584^2 + 0.584^2} = 1.786,$
- $||d3|| = \sqrt{1.584^2 + 1.584^2 + 0.584^2} = 2.316,$
- $||q|| = \sqrt{0.584^2 + 0.292^2} = 0.652.$

Podobieństwo pomiędzy wektorami

Mamy:

$$\cos(d, q) = \frac{<d,q>}{||d|| \cdot ||q||}$$

Zadanie:
Wyznaczy miarę kosinus pomiędzy sekwencjami $d1$, $d2$, $d3$ a q.

Podobieństwo pomiędzy wektorami

Mamy:

$$\cos(d, q) = \frac{<d,q>}{||d|| \cdot ||q||}$$

- $\cos(d1, q) = \frac{0 \cdot 0 + 0 \cdot 0 + 0.584 \cdot 0.584 + 0 \cdot 0 + 0.584 \cdot 0.292 + 0.584 \cdot 0}{1.011 \cdot 0.652} = 0.776$
- $\cos(d2, q) = \frac{0 \cdot 0 + 0 \cdot 0 + 0.584 \cdot 0.584 + 1.584 \cdot 0 + 0 \cdot 0.292 + 0.584 \cdot 0}{1.786 \cdot 0.652} = 0.292$
- $\cos(d3, q) = \frac{1.584 \cdot 0 + 1.584 \cdot 0 + 0 \cdot 0.584 + 0 \cdot 0 + 0.584 \cdot 0.292 + 0 \cdot 0}{2.316 \cdot 0.652} = 0.112$

W jakiej kolejności należałoby zwrócić dokumenty na podane zapytanie?