BIOINFORMATYKA

edycja 2019 / 2020

wykład 11

Aminokwasy, białka, struktury drugorzędowe

dr Jacek Śmietański
jacek.smietanski@ii.uj.edu.pl
http://jaceksmietanski.net

Klasyfikacja rodzin białkowych

Klasyfikacja SCOP

SCOP = *Structural Classification Of Proteins*

Hierarchiczny schemat klasyfikacji obejmujący 4 poziomy:

Rodzina – grupa białek powiązana strukturalnie, ewolucyjnie i funkcjonalnie;

Superrodzina – zbiór rodzin o podobnej strukturze i funkcji;

Zwój – wspólna topologia na większym fragmencie łańcucha

Klasa – grupa zwojów charakteryzowanych strukturą 2-rzędową.

Klasy:

α (głównie α-helisy),

β (głównie β-kartki),

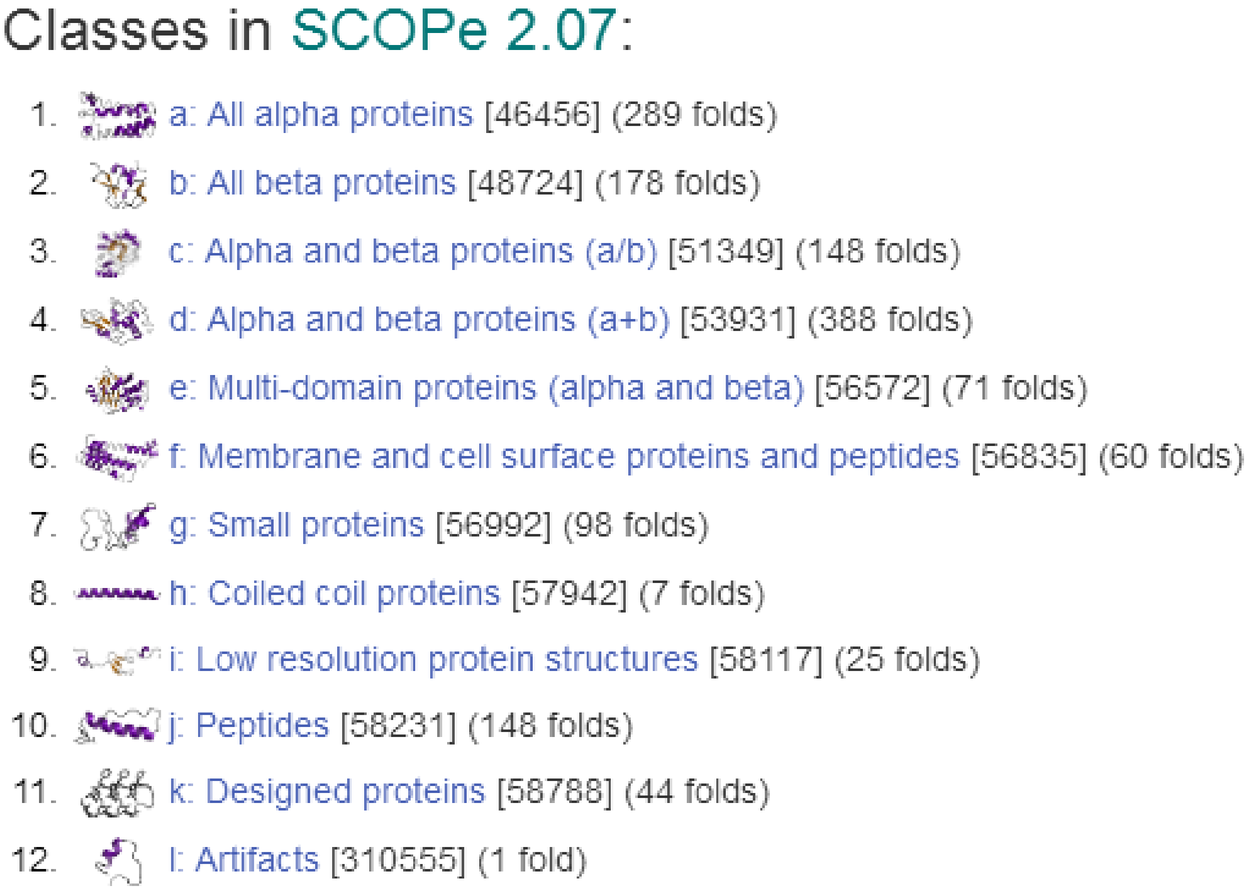
α/β (α-helisy i β-kartki w silnej interakcji),

α+β (α-helisy i β-kartki słabo lub w ogóle nie oddziałujące na siebie),

multidomain proteins (niehomologiczne białka, z różnymi zwojami)

SCOPe

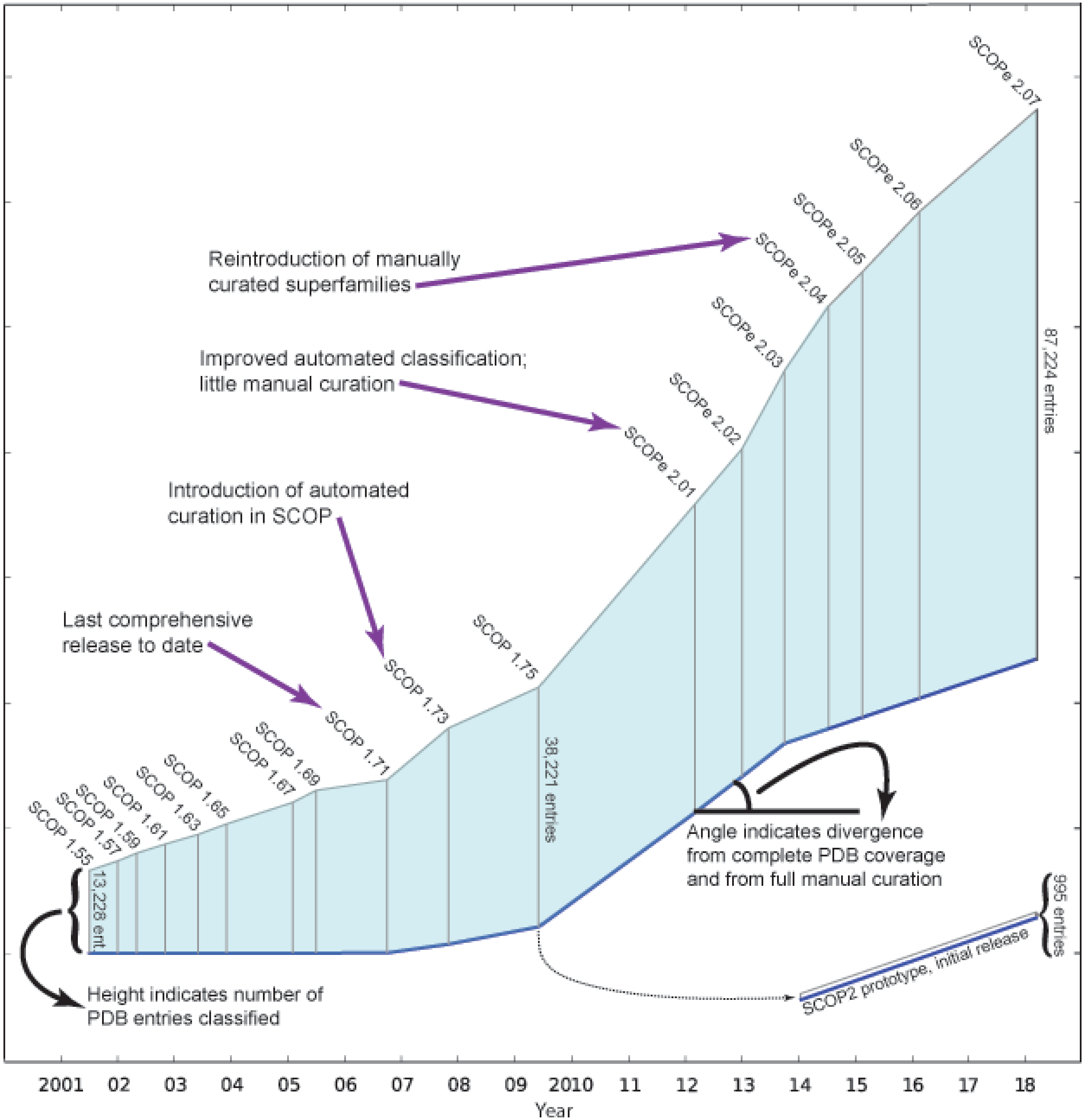
Kontynuacja i rozszerzenie pierwotnej klasyfikacji.
http://scop.berkeley.edu/



Classes in SCOPe 2.07:

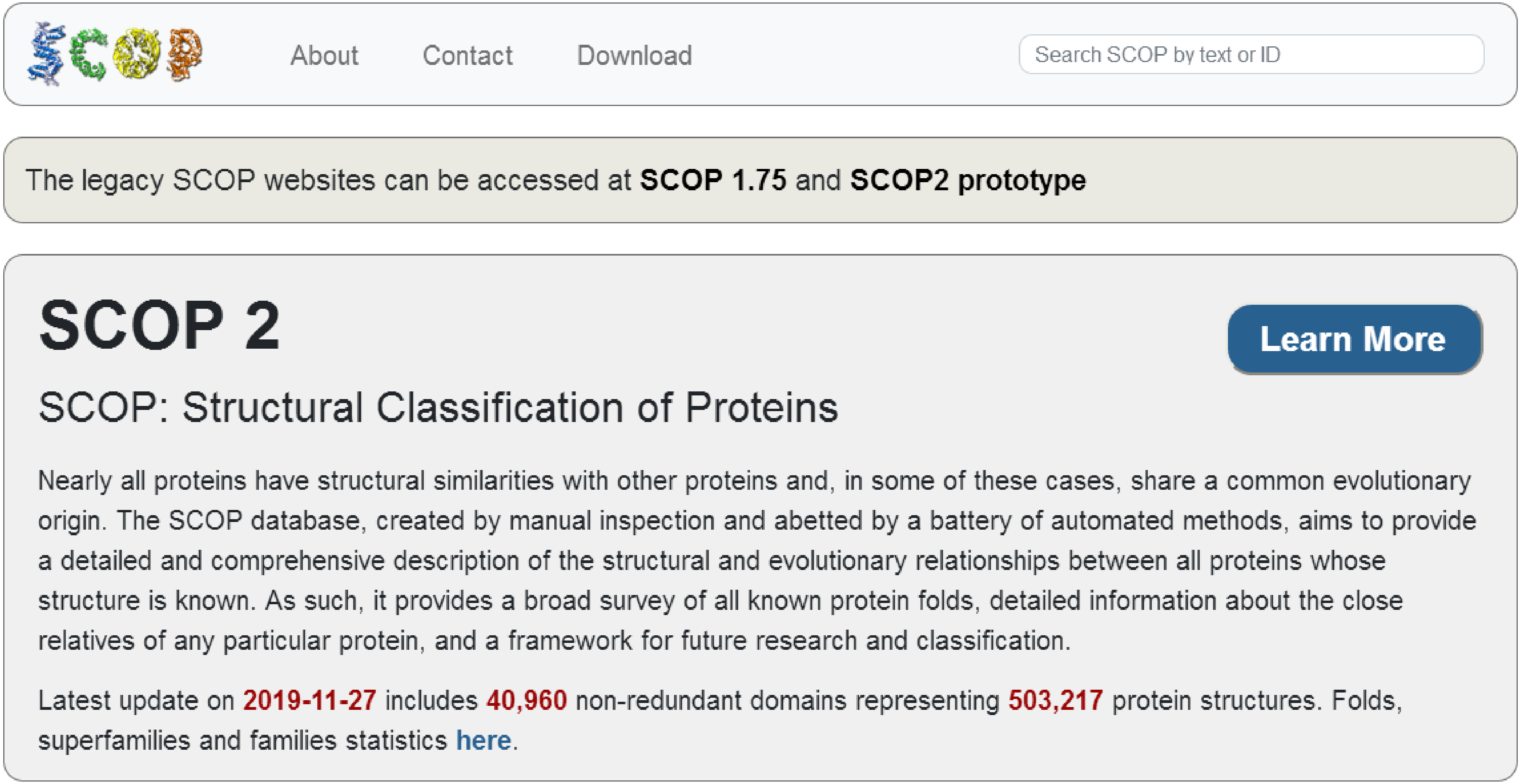
1. a: All alpha proteins [46456] (289 folds)
2. b: All beta proteins [48724] (178 folds)
3. c: Alpha and beta proteins (a/b) [51349] (148 folds)
4. d: Alpha and beta proteins (a+b) [53931] (388 folds)
5. e: Multi-domain proteins (alpha and beta) [56572] (71 folds)
6. f: Membrane and cell surface proteins and peptides [56835] (60 folds)
7. g: Small proteins [56992] (98 folds)
8. h: Coiled coil proteins [57942] (7 folds)
9. i: Low resolution protein structures [58117] (25 folds)
10. j: Peptides [58231] (148 folds)
11. k: Designed proteins [58788] (44 folds)
12. l: Artifacts [310555] (1 fold)

SCOP - rozwój


SCOPe 2.07
SCOPe 2.06
SCOPe 2.05
SCOPe 2.04
SCOPe 2.03
SCOPe 2.02
SCOPe 2.01
SCOP 1.75
SCOP 1.73
SCOP 1.71
SCOP 1.69
SCOP 1.67
SCOP 1.65
SCOP 1.63
SCOP 1.61
SCOP 1.59
SCOP 1.57
SCOP 1.55
Reintroduction of manually curated superfamilies
Improved automated classification; little manual curation
Introduction of automated curation in SCOP
Last comprehensive release to date
87,224 entries
38,221 entries
13,228 ent.
995 entries
Angle indicates divergence from complete PDB coverage and from full manual curation
Height indicates number of PDB entries classified
SCOP2 prototype, initial release
2001
02
03
04
05
06
07
08
09
2010
11
12
13
14
15
16
17
18
Year

SCOP2

http://scop2.mrc-lmb.cam.ac.uk/



The legacy SCOP websites can be accessed at **SCOP 1.75** and **SCOP2 prototype**

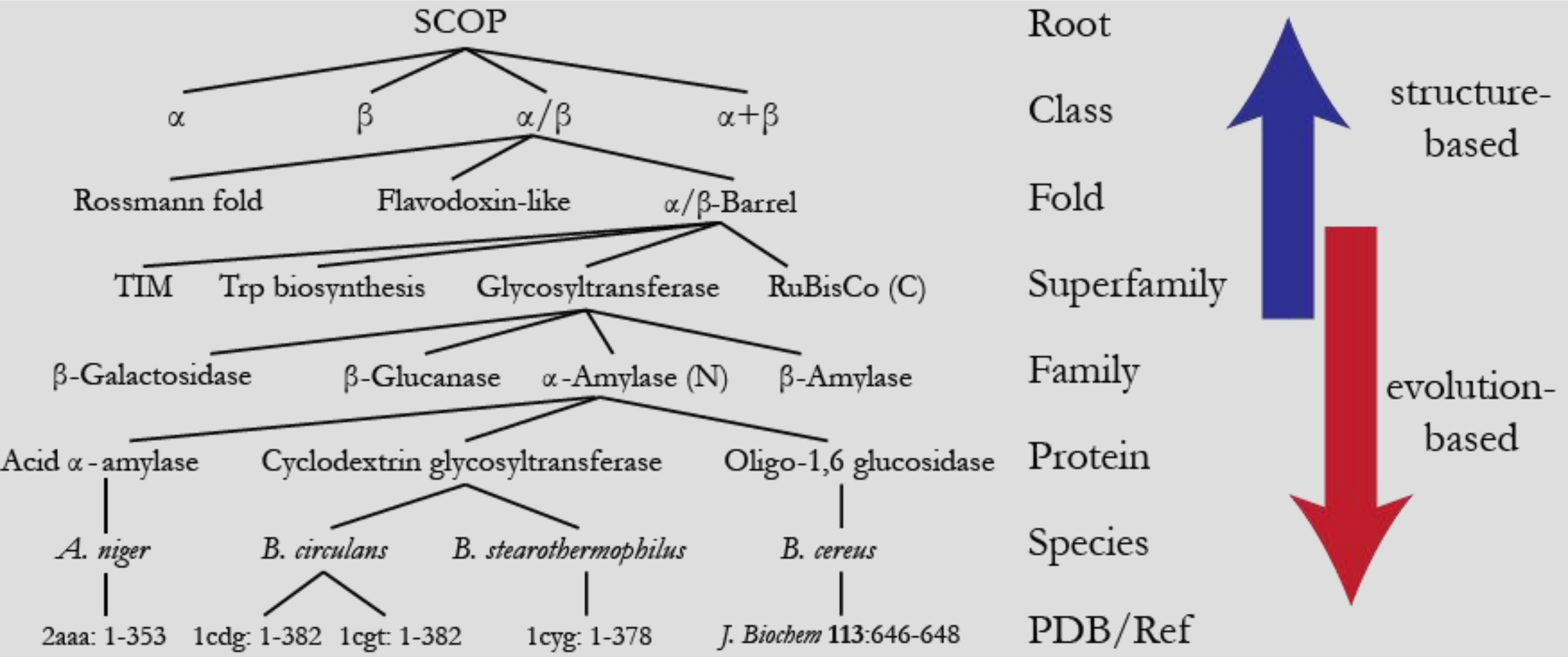
SCOP 2

SCOP: Structural Classification of Proteins

Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. The SCOP database, created by manual inspection and abetted by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification.

Latest update on **2019-11-27** includes **40,960** non-redundant domains representing **503,217** protein structures. Folds, superfamilies and families statistics **here**.

Przykładowa hierarchia SCOP



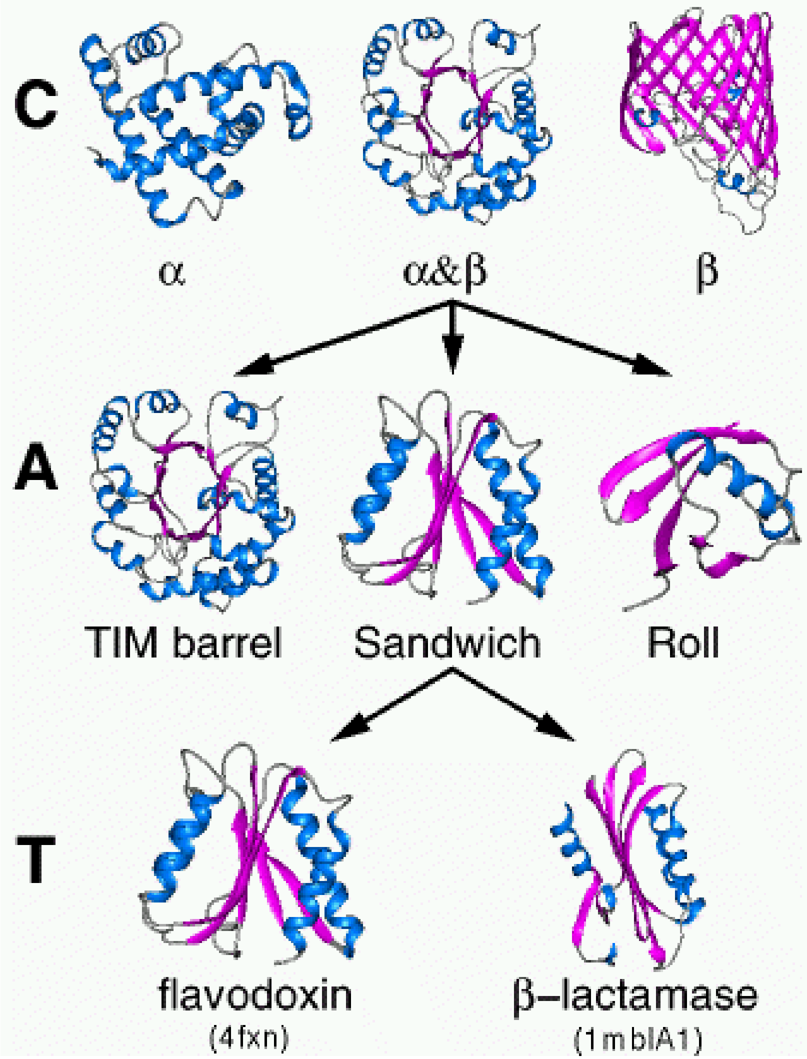
Klasyfikacja CATH

Class, Architecture, Topology, Homologous superfamily

Cztery poziomy hierarchii:

1. Klasa (poziom C): na podstawie typu struktury drugorzędowej: α, β, α&β (α/β i α+β), słaba lub nieokreślona struktura.
2. Architectura (poziom A): orientacja i topologia pomiędzy elementami struktury drugorzędowej.
3. Topologia (poziom T) – bazuje na typie pofałdowania.
4. Homologiczna superrodzina. (poziom H) – wysoka homologia wskazująca na wspólnego przodka:
   - > 30% identycznej sekwencji LUB
   - > 20% identycznej sekwencji i 60% strukturalnej homologii LUB
   - > 60% strukturalnej homologii i podobne domeny mają podobne funkcje.

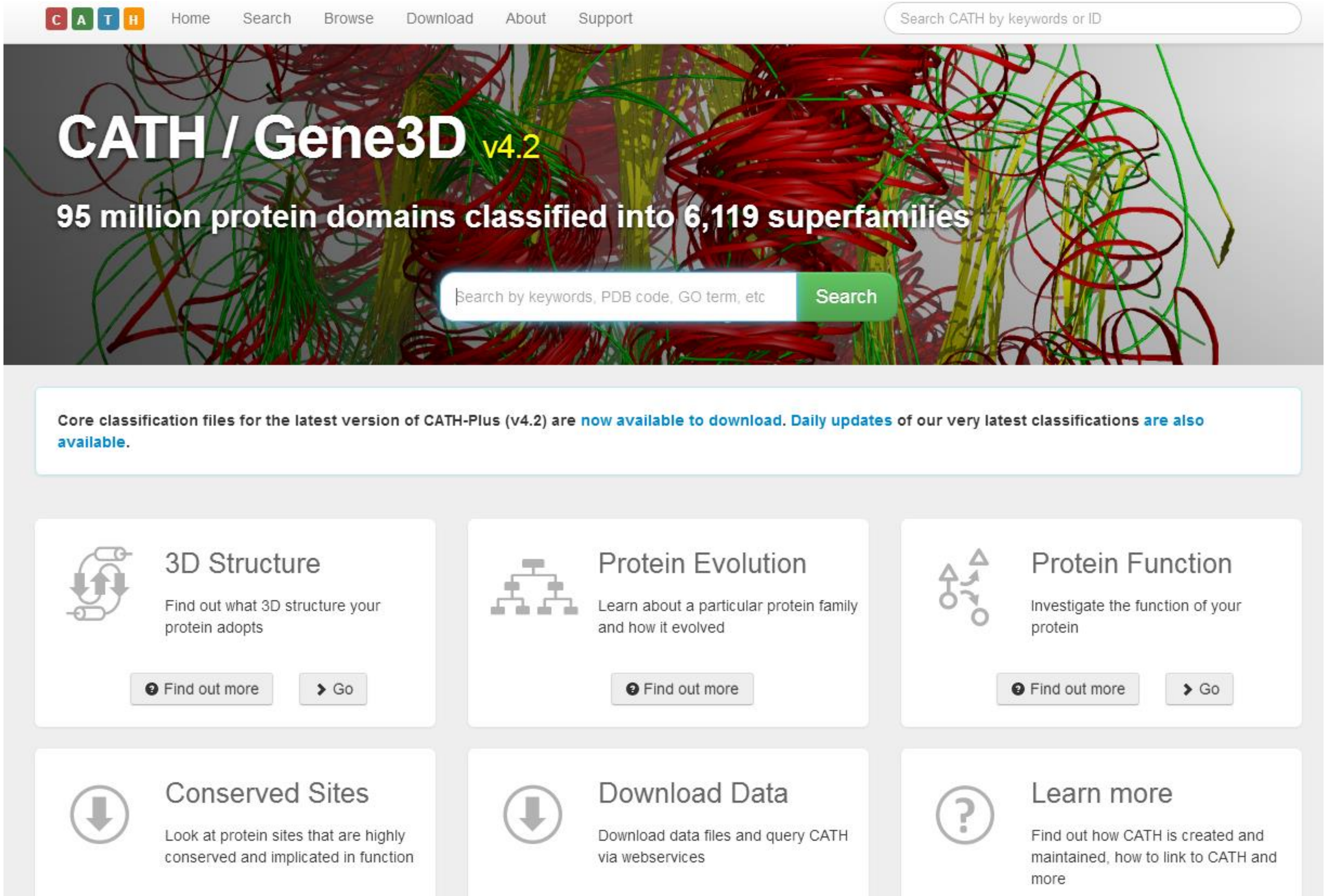
Klasyfikacja CATH



- Class(C)
  derived from secondary structure content is assigned automatically
- Architecture(A)
  describes the gross orientation of secondary structures, independent of connectivity.
- Topology(T)
  clusters structures according to their topological connections and numbers of secondary structures
- Homologous superfamily (H)

CATH on-line

http://www.cathdb.info



C A T H
Home
Search
Browse
Download
About
Support
Search CATH by keywords or ID

CATH / Gene3D v4.2

95 million protein domains classified into 6,119 superfamilies

Search by keywords, PDB code, GO term, etc
Search

Core classification files for the latest version of CATH-Plus (v4.2) are now available to download. Daily updates of our very latest classifications are also available.

3D Structure

Find out what 3D structure your protein adopts

Protein Evolution

Learn about a particular protein family and how it evolved

Protein Function

Investigate the function of your protein

Conserved Sites

Look at protein sites that are highly conserved and implicated in function

Download Data

Download data files and query CATH via webservices

Learn more

Find out how CATH is created and maintained, how to link to CATH and more

Przewidywanie struktury drugorzędowej białek

Do czego może się przydać informacja o strukturze 2-rzędowej?

- jest etapem tworzenia struktury przestrzennej i domenowej

- przydatna koncepcja dla zrozumienia struktury

- ma związek z funkcją białka

- przydatna w algorytmach przewidywania struktury przestrzennej (definiuje obszary na wzorcach)

Rozwój metod

1. Pierwsza generacja: statystyki pojedynczych aminokwasów

   np.: Chou-Fasman, LIM, GOR I, etc

   skuteczność: niska

2. Druga generacja: statystyki w oknach

   np.: ALB, GOR III, etc

   skuteczność: ~60%

3. Trzecia generacja: oddziaływania długodystansowe, metody homologiczne

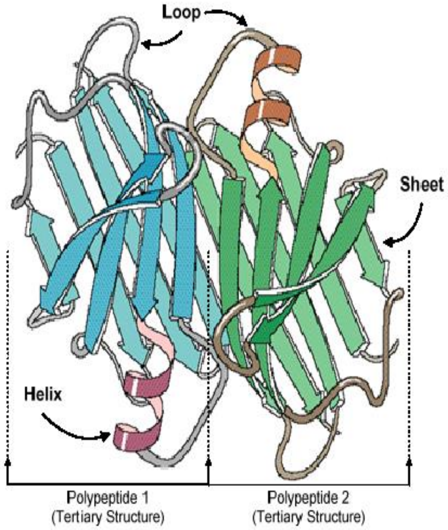
   np.: PHD

   skuteczność: ~70%

Najważniejsze techniki

- metody statystyczne;
- najbliższego sąsiada;
- sieci neuronowe;
- ukryte modele Markowa

Zwykle są to metody porównawcze, bazujące na homologii.



Strategia przewidywania struktury drugorzędowej:

- stosować jak najwięcej metod;
- wykorzystać zestawienie sekwencji homologicznych;
- złożyć przewidywania w jednokonsensusowe.

Zalety i wady

W kontekście poszukiwania struktury.

Zalety:

- Może być zastosowana do sekwencji całkowicie nieznanego białka
- Poprzedza rozpoznanie zwoju
- Uzupełnia inne metody modelowania

Wady:

- Najlepsze metody mają precyzję nie wyższą niż 80%
- „doskonale” przewidziana struktura II-rz nie zawsze prowadzi do rozpoznania zwoju

Metoda statystyczna (Chou, Fasman 1974)

Na podstawie analizy częstości występowania poszczególnych aminokwasów w poszczególnych typach struktur.

Przykład:

Total number of residues	2000
Number of alanines	100
Number of helical residues	500
Number of alanines in helices	50

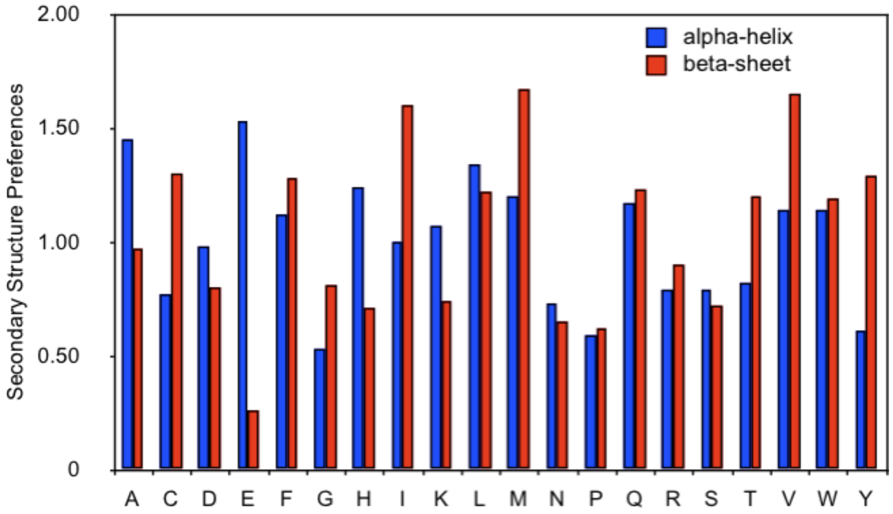
- P(Ala in Helix) = 50/500 = 0.1

  P(Ala) = 100/2000 = 0.05

  Helix propensity (PA) of Ala = P(Ala in Helix)/P(Ala) = 0.1/0.05 = 2

Tabela preferencji

	helix	strand	turn
•Alanine	1.42	0.83	0.66
•Arginine	0.98	0.93	0.95
•Aspartic Acid	1.01	0.54	1.46
•Asparagine	0.67	0.89	1.56
•Cysteine	0.70	1.19	1.19
•Glutamic Acid	1.39	1.17	0.74
•Glutamine	1.11	1.10	0.98
•Glycine	0.57	0.75	1.56
•Histidine	1.00	0.87	0.95
•Isoleucine	1.08	1.60	0.47
•Leucine	1.41	1.30	0.59
•Lysine	1.14	0.74	1.01
•Methionine	1.45	1.05	0.60
•Phenylalanine	1.13	1.38	0.60
•Proline	0.57	0.55	1.52
•Serine	0.77	0.75	1.43
•Threonine	0.83	1.19	0.96
•Tryptophan	1.08	1.37	0.96
•Tyrosine	0.69	1.47	1.14
•Valine	1.06	1.70	0.50



Metody okna, GOR

Rozwinięcie metody Chou-Fasmana. Wykorzystuje tabele preferencji dla pojedynczych aminokwasów, uwzględnia jednak również aminokwasy sąsiadujące:

- weźmy okno uwzględniające 16 sąsiadujących residuów (8 przed i 8 po badanym aminokwasie)

- dla każdego residuum w oknie analizujemy jego wpływ na konformację badanego (środkowego) aminokwasu.

- badany wpływ ewaluujemy na podstawie danych statystycznych.

Informacja ewolucyjna

- Pojedyncza sekwencja zastąpiona uliniowieniem spokrewnionych (homologicznych) sekwencji

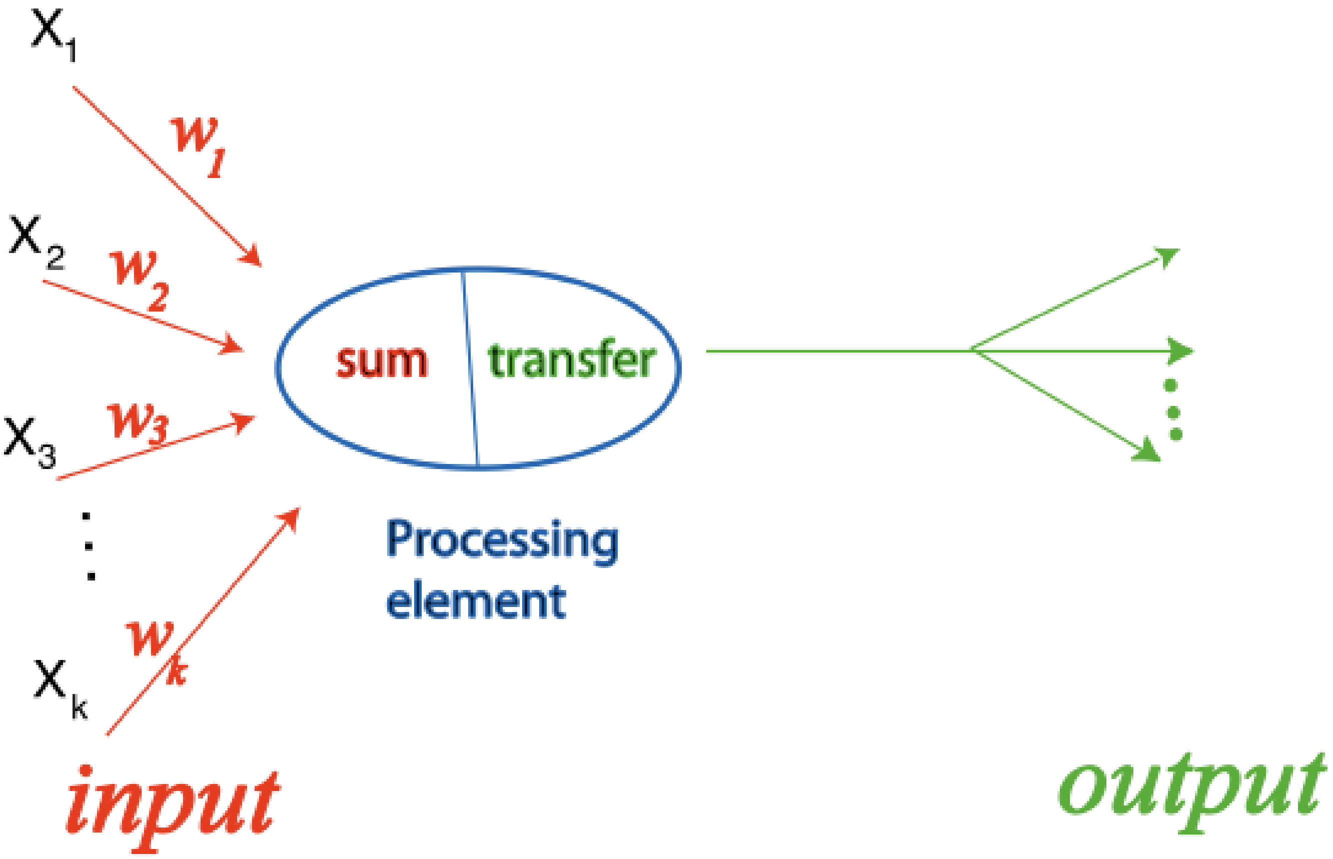
- Profil:

	A	C	D
ACAA	0.75	0.25	0
DDCA	0.25	0.25	0.5
ACDA	0.5	0.25	0.25
DAAA	0.75	0	0.25

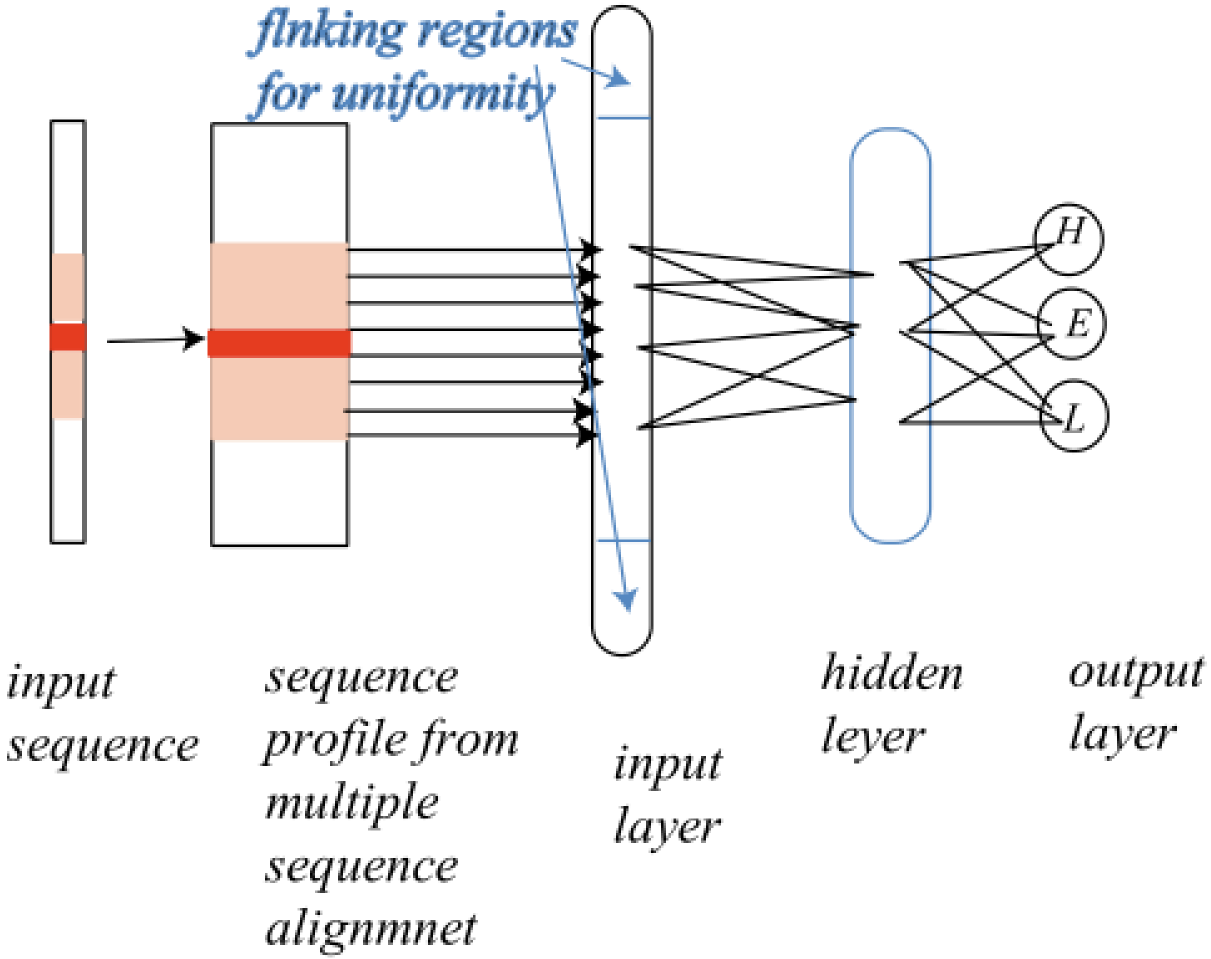
- Wartości binarne na wejściu sieci zastąpione wartościami rzeczywistymi z przedziału [0,1]

- Poprawa jakości predykcji (z 65% do >70%)

Sieci neuronowe



Sieci neuronowe (2)



Algorytm DSSP

- PDB nie zawiera jawnych danych na temat struktury 2-rzędowej
- Ustalenie struktury drugorzędowej na podstawie współrzędnych atomów w przestrzeni
- DSSP
  - Wolfgang Kabsch, Chris Sander;
  - Uzyskiwanie informacji o strukturze 2-rzędowej na podstawie danych z PDB;
  - 7 klas: H, G, I, E, B, T, S

7 klas	H	G	I	E	B	T	S
3 klasy	H	H	H	E	E	L	L

Przykład – sieć neuronowa

Wejście:

- Informacja na temat *w* sąsiednich aminokwasów (*w* nieparzysta) – tzw. okno wejściowe
- Kodowanie ortogonalne aminokwasów:
  - wektor o wymiarze 20
  - na jednej pozycji 1, a na pozostałych 0
- *(20*w)* elementów wejściowych

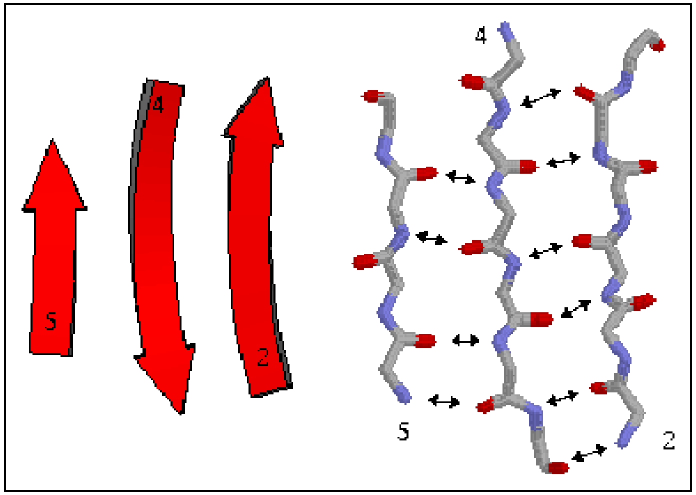
Wyjście:

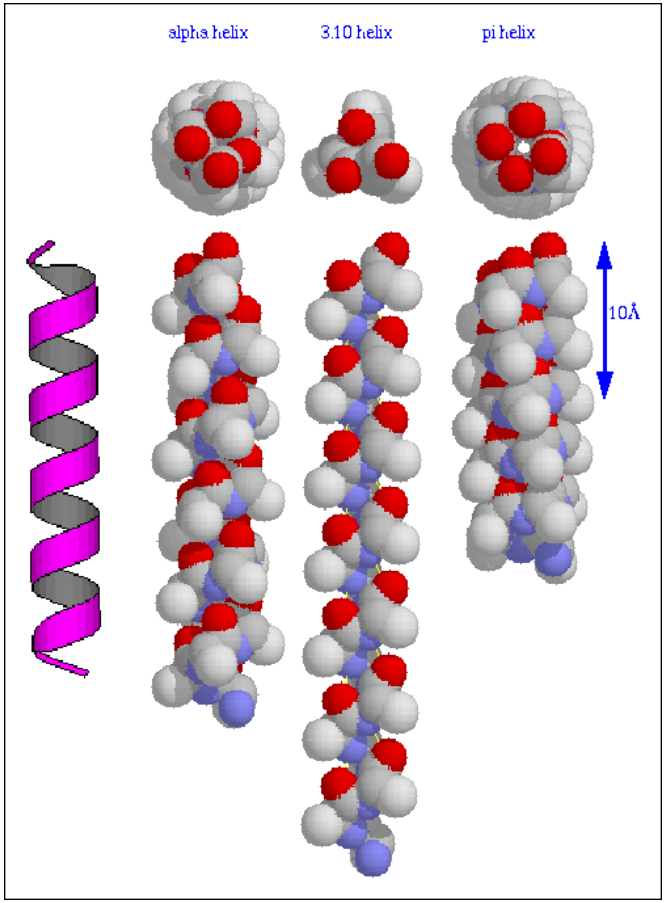
- 3 neurony wyjściowe odpowiadające poszczególnym klasom struktury 2-rzędowej (wartość rzeczywista z przedziału [0,1]):
  - H – α helisa
  - E – struktura β
  - L – pętla łącząca
- Wynik predykcji: klasa odpowiadająca neuronowi wyjściowemu o maksymalnej wartości
- Predykcja dla centralnego aminokwasu z okna wejściowego

Rozpoznawane klasy

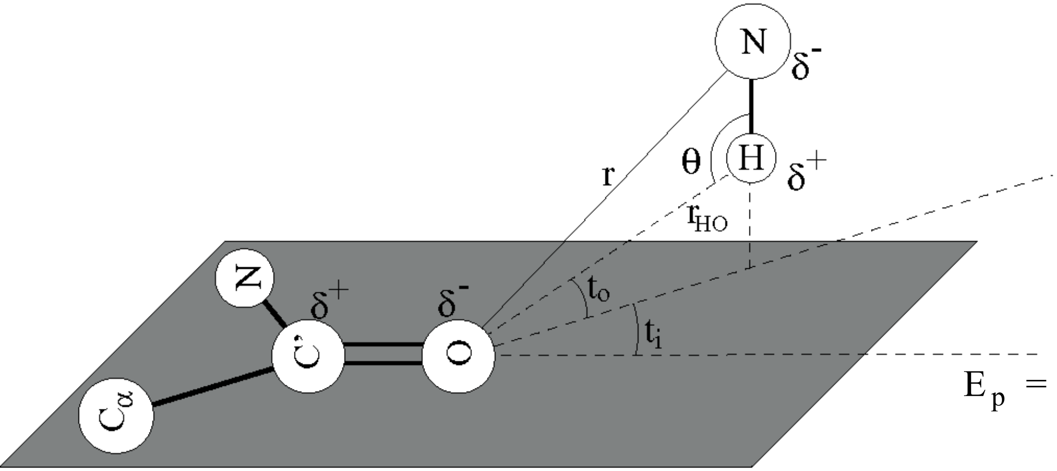
8 kategorii (DSSP):

- H: α - helisa
- G: 3_{10} – helisa
- I: π - helisa (bardzo rzadka)
- E: β - kartka
- B: β - most
- T: zwrot
- S: bend
- L: pozostałe





STRIDE – Empirical Hydrogen Bond Calculation



$$E_{hb} = E_r \cdot E_t \cdot E_p$$

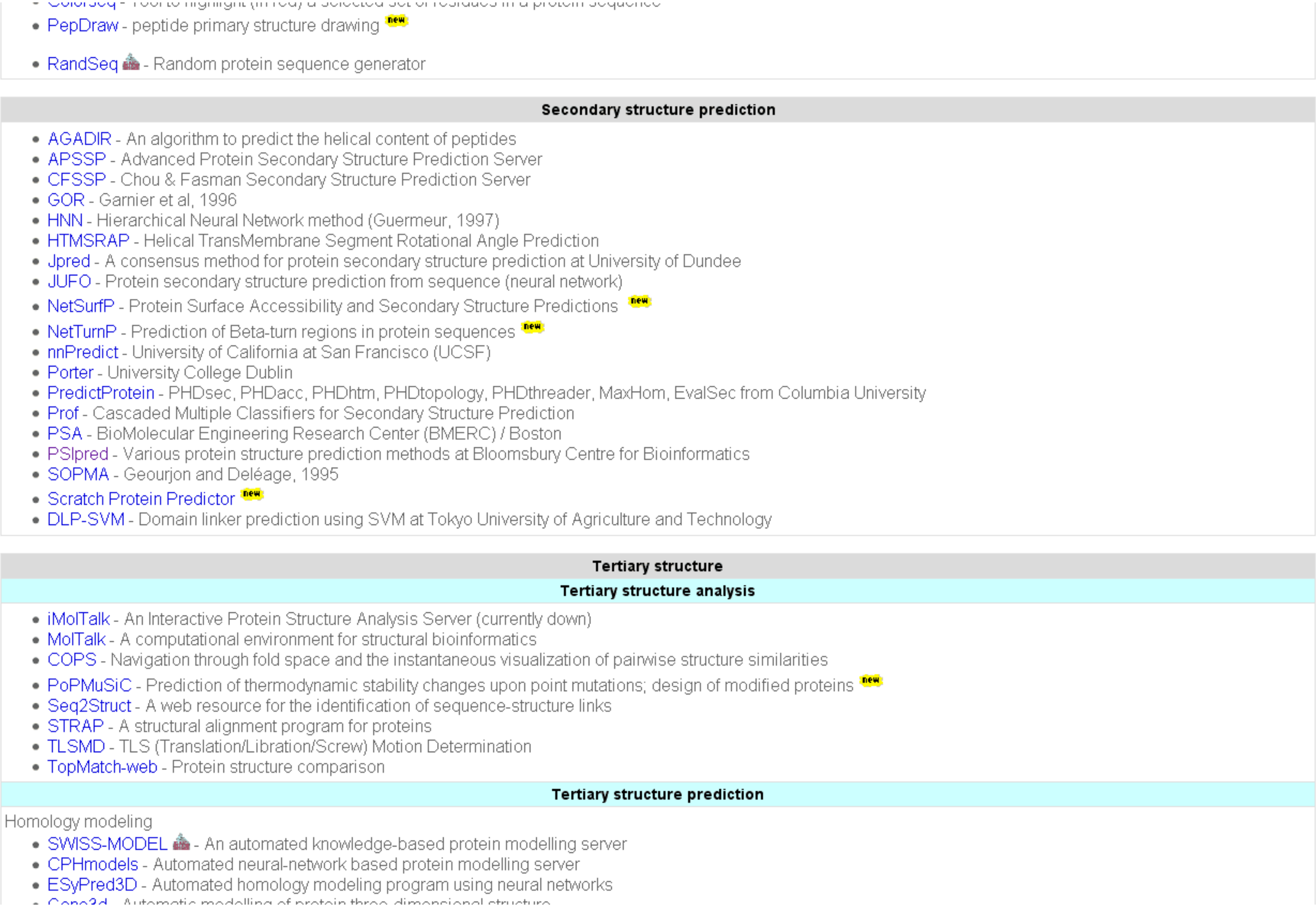
$$E_r = \left(\frac{4\, r_m^6}{r^6} - \frac{3\, r_m^8}{r^8} \right) E_m$$

$$E_p = \cos^2(\theta)$$

$$E_t = \begin{cases} [0.9 + 0.1 \sin(2t_i)] \cos(t_o) & 0° < t_i \leq 90° \\ K_1 [K_2 - \cos^2(t_i)] \cos(t_o) & 90° < t_i \leq 110° \\ 0 & 110° \leq t_i \end{cases}$$

- Derived from small molecule structures r_m (3.0A) and E_m (-2.8kcal/mole)
- Total energy E_{hb}

Expasy – odsyłacze do licznych narzędzi



- ColorSeq - Tool to highlight (in red) a selected set of residues in a protein sequence
- PepDraw - peptide primary structure drawing new
- RandSeq - Random protein sequence generator

Secondary structure prediction

- AGADIR - An algorithm to predict the helical content of peptides
- APSSP - Advanced Protein Secondary Structure Prediction Server
- CFSSP - Chou & Fasman Secondary Structure Prediction Server
- GOR - Garnier et al, 1996
- HNN - Hierarchical Neural Network method (Guermeur, 1997)
- HTMSRAP - Helical TransMembrane Segment Rotational Angle Prediction
- Jpred - A consensus method for protein secondary structure prediction at University of Dundee
- JUFO - Protein secondary structure prediction from sequence (neural network)
- NetSurfP - Protein Surface Accessibility and Secondary Structure Predictions new
- NetTurnP - Prediction of Beta-turn regions in protein sequences new
- nnPredict - University of California at San Francisco (UCSF)
- Porter - University College Dublin
- PredictProtein - PHDsec, PHDacc, PHDhtm, PHDtopology, PHDthreader, MaxHom, EvalSec from Columbia University
- Prof - Cascaded Multiple Classifiers for Secondary Structure Prediction
- PSA - BioMolecular Engineering Research Center (BMERC) / Boston
- PSIpred - Various protein structure prediction methods at Bloomsbury Centre for Bioinformatics
- SOPMA - Geourjon and Deléage, 1995
- Scratch Protein Predictor new
- DLP-SVM - Domain linker prediction using SVM at Tokyo University of Agriculture and Technology

Tertiary structure

Tertiary structure analysis

- iMolTalk - An Interactive Protein Structure Analysis Server (currently down)
- MolTalk - A computational environment for structural bioinformatics
- COPS - Navigation through fold space and the instantaneous visualization of pairwise structure similarities
- PoPMuSiC - Prediction of thermodynamic stability changes upon point mutations; design of modified proteins new
- Seq2Struct - A web resource for the identification of sequence-structure links
- STRAP - A structural alignment program for proteins
- TLSMD - TLS (Translation/Libration/Screw) Motion Determination
- TopMatch-web - Protein structure comparison

Tertiary structure prediction

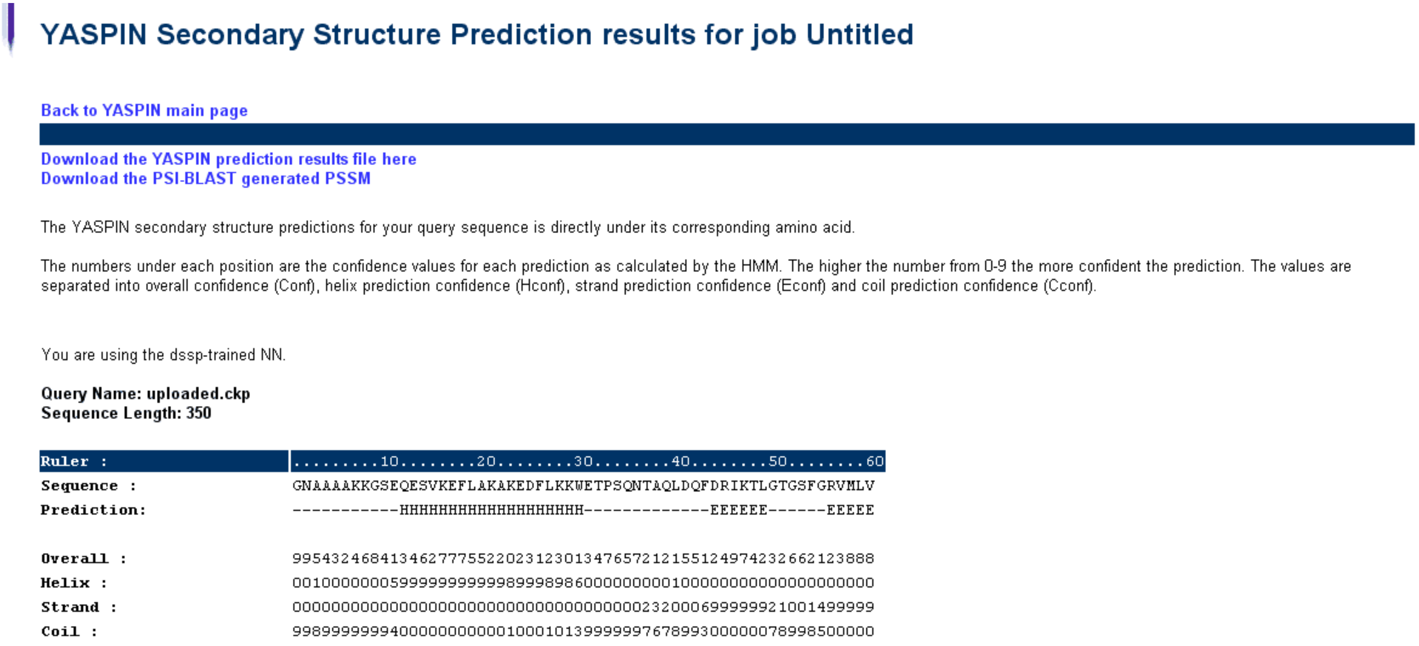
Homology modeling

- SWISS-MODEL - An automated knowledge-based protein modelling server
- CPHmodels - Automated neural-network based protein modelling server
- ESyPred3D - Automated homology modeling program using neural networks
- Geno3d - Automatic modelling of protein three-dimensional structure

http://expasy.org/tools/

Narzędzia - przykłady

PSIPRED - http://bioinf.cs.ucl.ac.uk/psipred/
YASPIN - http://www.ibi.vu.nl/programs/yaspinwww/
SSPRO - http://download.igb.uci.edu/sspro4.html
PROTEUS - http://wks16338.biology.ualberta.ca/proteus/



YASPIN Secondary Structure Prediction results for job Untitled

Back to YASPIN main page

Download the YASPIN prediction results file here
Download the PSI-BLAST generated PSSM

The YASPIN secondary structure predictions for your query sequence is directly under its corresponding amino acid.

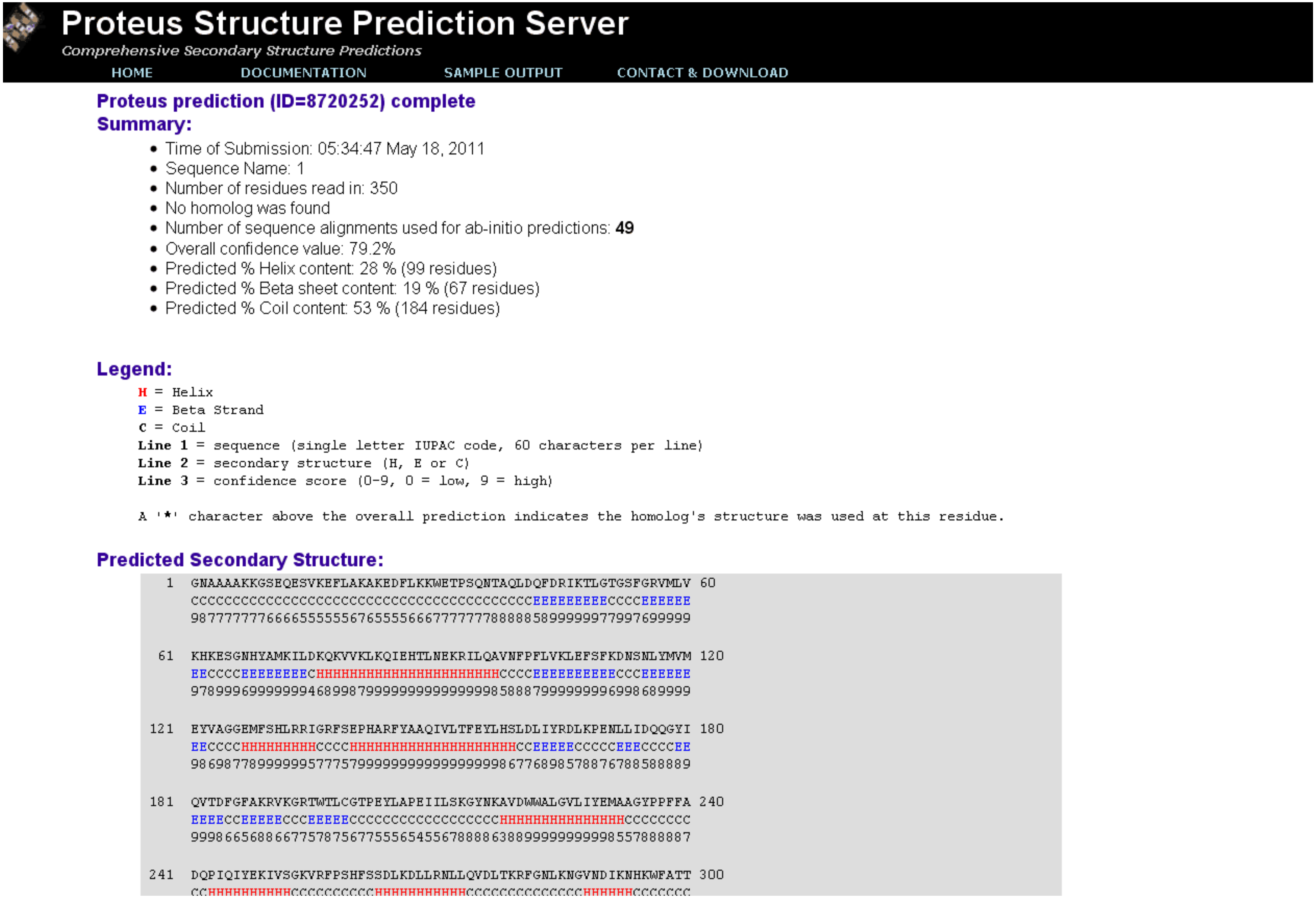
The numbers under each position are the confidence values for each prediction as calculated by the HMM. The higher the number from 0-9 the more confident the prediction. The values are separated into overall confidence (Conf), helix prediction confidence (Hconf), strand prediction confidence (Econf) and coil prediction confidence (Cconf).

You are using the dssp-trained NN.

Query Name: uploaded.ckp
Sequence Length: 350

```
Ruler :        .........10........20........30........40........50........60
Sequence :     GNAAAAKKGSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVMLV
Prediction:    -----------HHHHHHHHHHHHHHHHHHH-------------EEEEEE------EEEEE

Overall :      995432468413462777552202312301347657212155124974232662123888
Helix :        001000000059999999999989998986000000000100000000000000000000
Strand :       000000000000000000000000000000000000232000699999921001499999
Coil :         998999999940000000000010001013999999767899300000078998500000
```

Proteus - przykład



Proteus Structure Prediction Server

Comprehensive Secondary Structure Predictions

Proteus prediction (ID=8720252) complete

Summary:

- Time of Submission: 05:34:47 May 18, 2011
- Sequence Name: 1
- Number of residues read in: 350
- No homolog was found
- Number of sequence alignments used for ab-initio predictions: **49**
- Overall confidence value: 79.2%
- Predicted % Helix content: 28 % (99 residues)
- Predicted % Beta sheet content: 19 % (67 residues)
- Predicted % Coil content: 53 % (184 residues)

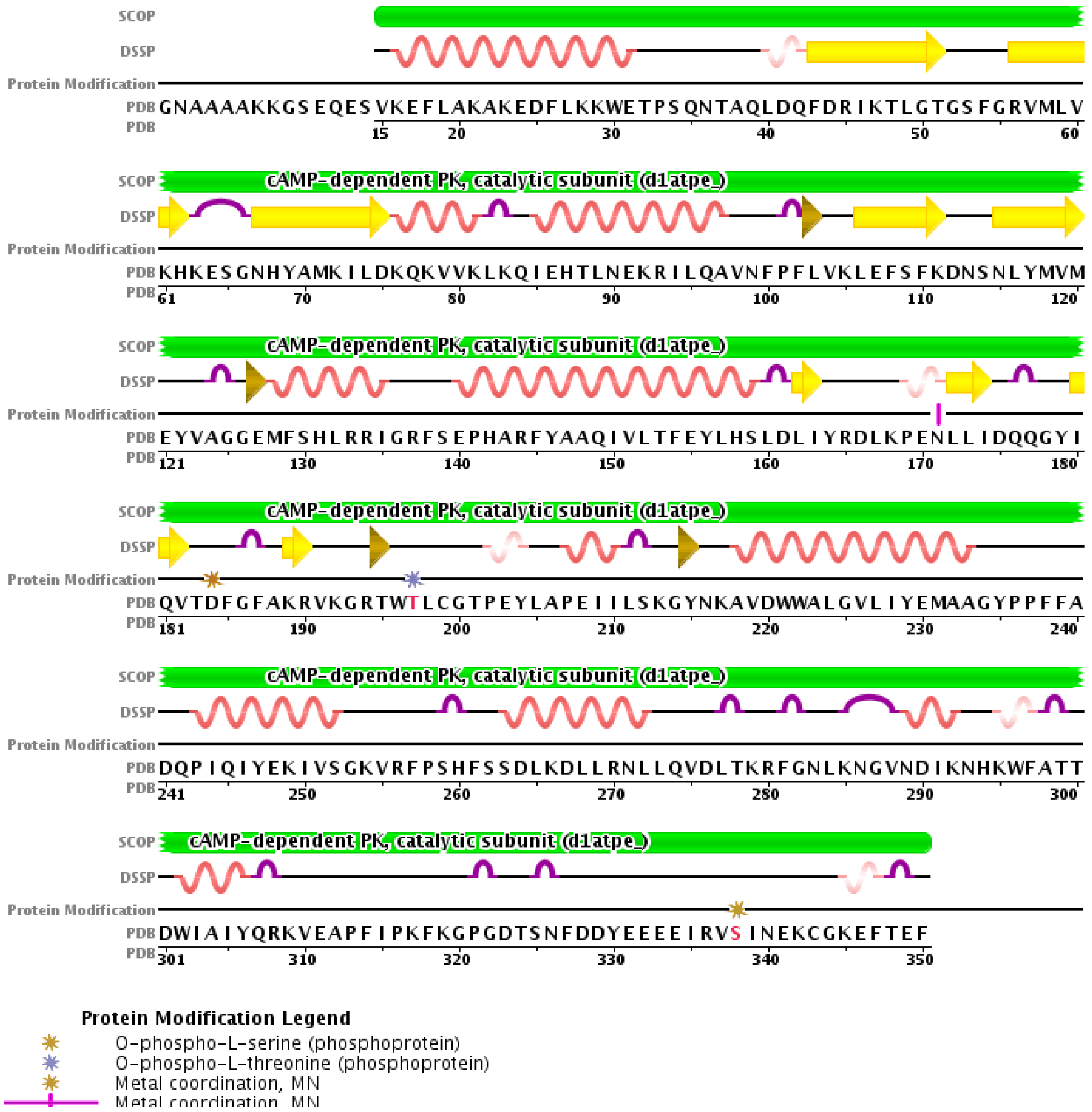
Legend:

```
H = Helix
E = Beta Strand
C = Coil
Line 1 = sequence (single letter IUPAC code, 60 characters per line)
Line 2 = secondary structure (H, E or C)
Line 3 = confidence score (0-9, 0 = low, 9 = high)

A '*' character above the overall prediction indicates the homolog's structure was used at this residue.
```

Predicted Secondary Structure:

```
  1  GNAAAAKKGSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVMLV  60
     CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEEEEEECCCCEEEEEE
     987777777666655555567655556667777777888885899999977997699999

 61  KHKESGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLYMVM 120
     EECCCCEEEEEEEECHHHHHHHHHHHHHHHHHHHHHHCCCCEEEEEEEEEECCCEEEEEE
     978999699999994689987999999999999999858887999999996998689999

121  EYVAGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYI 180
     EECCCCHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHHHCCEEEEECCCCCEEECCCCEE
     986987789999995777579999999999999999986776898578876788588889

181  QVTDFGFAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA 240
     EEEECCEEEEECCCEEEEECCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHCCCCCCCC
     999866568866775787567755565455678888638899999999998557888887

241  DQPIQIYEKIVSGKVRFPSHFSSDLKDLLRNLLQVDLTKRFGNLKNGVNDIKNHKWFATT 300
     CCHHHHHHHHHHCCCCCCCCCCHHHHHHHHHHHCCCCCCCCCCCCCCHHHHHHCCCCCCC
```

PDB - DSSP


SCOP
DSSP
Protein Modification
PDB GNAAAAKKGS EQES VKEF LAKAKEDF LKKWETPS QNTAQLDQFDR IKTLGTGS FGRVMLV
PDB 15 20 30 40 50 60
SCOP cAMP-dependent PK, catalytic subunit (d1atpe_)
DSSP
Protein Modification
PDB KHKESGNHYAMK I LDKQKVVKLKQ I EHTLNEKR I LQAVNF PFLVKLEFS FKDNSNLYMVM
PDB 61 70 80 90 100 110 120
SCOP cAMP-dependent PK, catalytic subunit (d1atpe_)
DSSP
Protein Modification
PDB EYVAGGEMFSHLRR I GRFS EPHARFYAAQ I VLTFEYLHS LDL I YRDLKPENLL I DQQGY I
PDB 121 130 140 150 160 170 180
SCOP cAMP-dependent PK, catalytic subunit (d1atpe_)
DSSP
Protein Modification
PDB QVTDFGFAKRVKGRTWTLCGTPEYLAPE I I LSKGYNKAVDWWALGVL I YEMAAGYPPFFA
PDB 181 190 200 210 220 230 240
SCOP cAMP-dependent PK, catalytic subunit (d1atpe_)
DSSP
Protein Modification
PDB DQP I Q I YEK I VSGKVRF PSHFSSDLKDLLRNLLQVDLTKRFGNLKNGVND I KNHKWFATT
PDB 241 250 260 270 280 290 300
SCOP cAMP-dependent PK, catalytic subunit (d1atpe_)
DSSP
Protein Modification
PDB DWI A I YQRKVEAPF I PKFKGPGDTSNFDDYEEEE I RVS INEKCGKEFTEF
PDB 301 310 320 330 340 350
Protein Modification Legend
O-phospho-L-serine (phosphoprotein)
O-phospho-L-threonine (phosphoprotein)
Metal coordination, MN
Metal coordination, MN

białko 1ATP

Najlepsze metody

- PredictProtein-PHD (72%)
  - http://www.predictprotein.org/
- Jpred (73-75%)
  - http://jura.ebi.ac.uk:8888/
- PREDATOR (75%)
  - http://www.embl-heidelberg.de/cgi/predator_serv.pl
- PSIpred (77%)
  - http://insulin.brunel.ac.uk/psipred

PDB

Protein Data Bank

wwPDB *(Worldwide Protein Data Bank)***:**

Organizacja utrzymująca bazę danych struktur makromolekularnych (DNA, RNA, białka, związki hybrydowe, kompleksy białkowe).

Skład wwPDB:

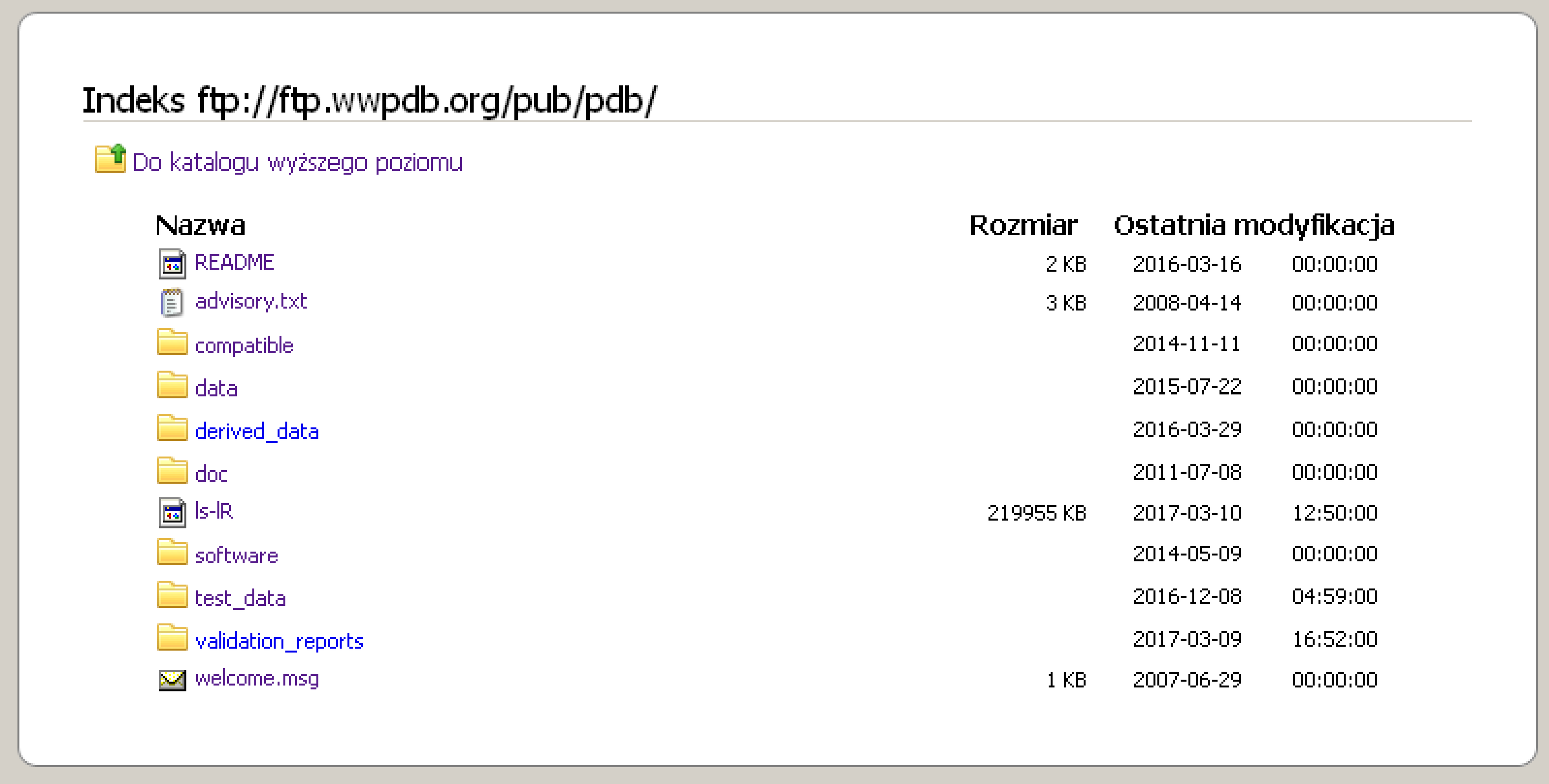
- Research Collaboratory for Structural Bioinformatics Protein Database (RCSB PDB)
- Protein Data Bank in Europe (PDBe)
- Protein Data Bank Japan (PDBj)
- Biological Magnetic Resonance Data Bank (BMRB)

Centralne archiwum znajduje się w RCSB. Dostęp do wszystkich danych jest możliwy również poprzez serwisy poszczególnych członków.

Dostęp do danych strukturalnych jest powszechny i bezpłatny.

FTP

wwPDB: ftp.wwpdb.org
RCSB PDB (USA): ftp.rcsb.org
PDBe (UK): ftp.ebi.ac.uk/pub/databases/pdb/
PDBj (Japan): ftp.pdbj.org

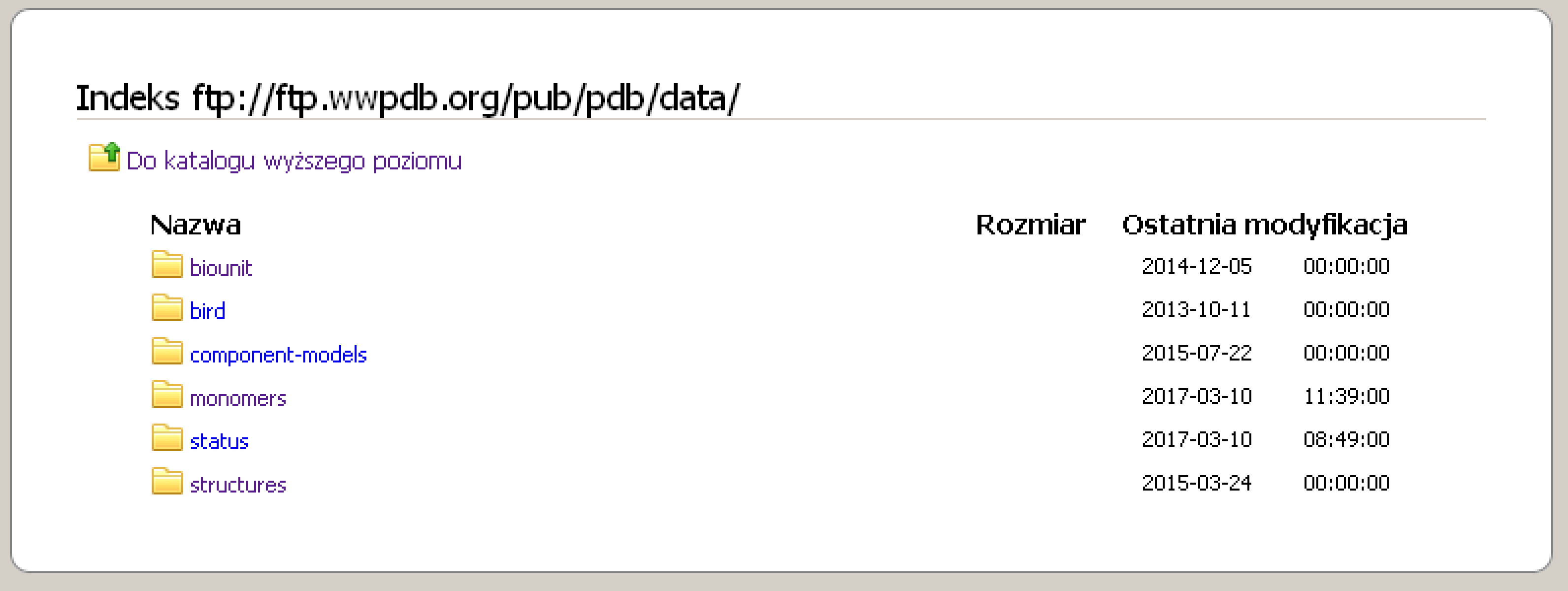


Indeks ftp://ftp.wwpdb.org/pub/pdb/

Do katalogu wyższego poziomu

Nazwa	Rozmiar	Ostatnia modyfikacja	
README	2 KB	2016-03-16	00:00:00
advisory.txt	3 KB	2008-04-14	00:00:00
compatible		2014-11-11	00:00:00
data		2015-07-22	00:00:00
derived_data		2016-03-29	00:00:00
doc		2011-07-08	00:00:00
ls-lR	219955 KB	2017-03-10	12:50:00
software		2014-05-09	00:00:00
test_data		2016-12-08	04:59:00
validation_reports		2017-03-09	16:52:00
welcome.msg	1 KB	2007-06-29	00:00:00

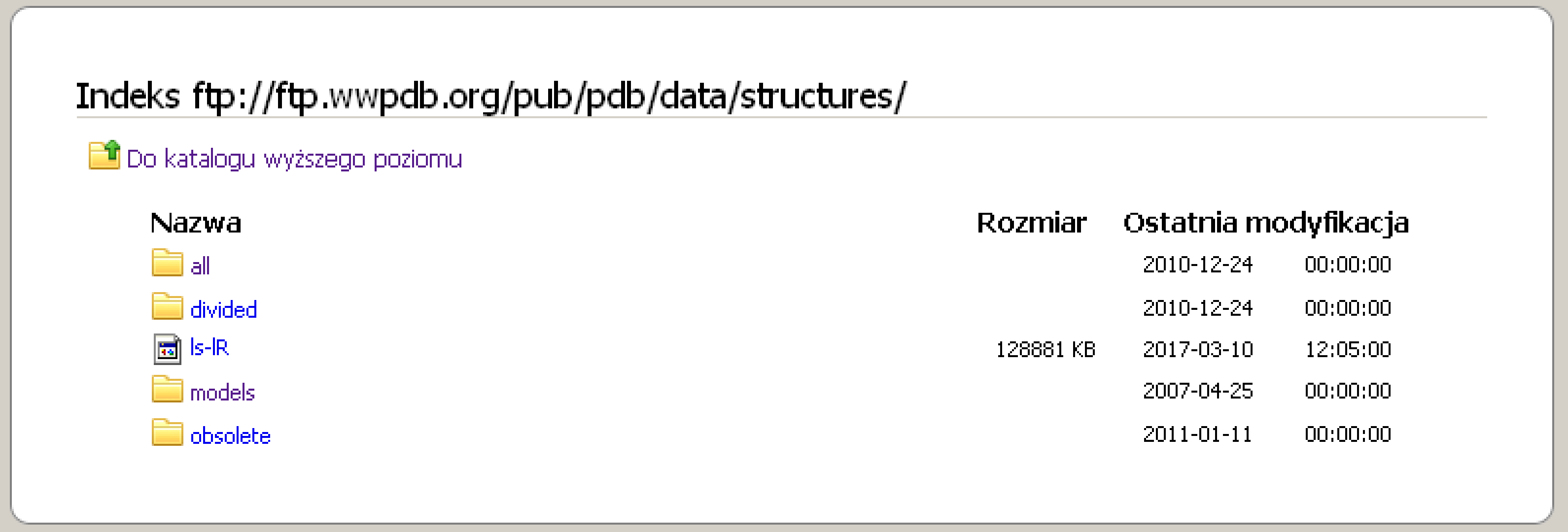
Struktura bazy danych



Indeks ftp://ftp.wwpdb.org/pub/pdb/data/

Do katalogu wyższego poziomu

Nazwa	Rozmiar	Ostatnia modyfikacja
biounit		2014-12-05 00:00:00
bird		2013-10-11 00:00:00
component-models		2015-07-22 00:00:00
monomers		2017-03-10 11:39:00
status		2017-03-10 08:49:00
structures		2015-03-24 00:00:00



Indeks ftp://ftp.wwpdb.org/pub/pdb/data/structures/

Do katalogu wyższego poziomu

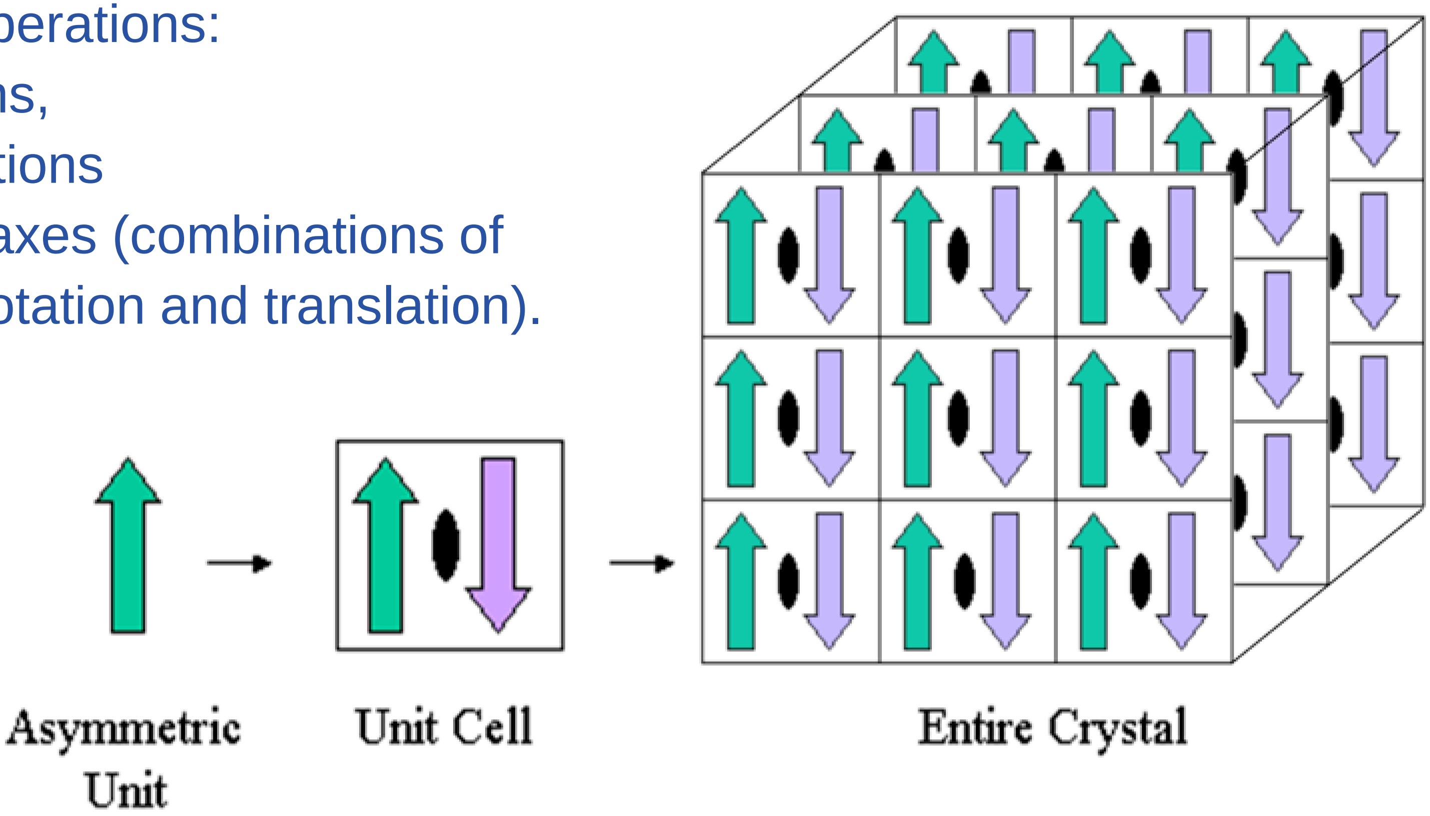
Nazwa	Rozmiar	Ostatnia modyfikacja
all		2010-12-24 00:00:00
divided		2010-12-24 00:00:00
ls-lR	128881 KB	2017-03-10 12:05:00
models		2007-04-25 00:00:00
obsolete		2011-01-11 00:00:00

Jednostka asymetryczna

The asymmetric unit is the smallest portion of a crystal structure to which symmetry operations can be applied in order to generate the complete unit cell.

Symmetry operations:

- rotations,
- translations
- screw axes (combinations of rotation and translation).

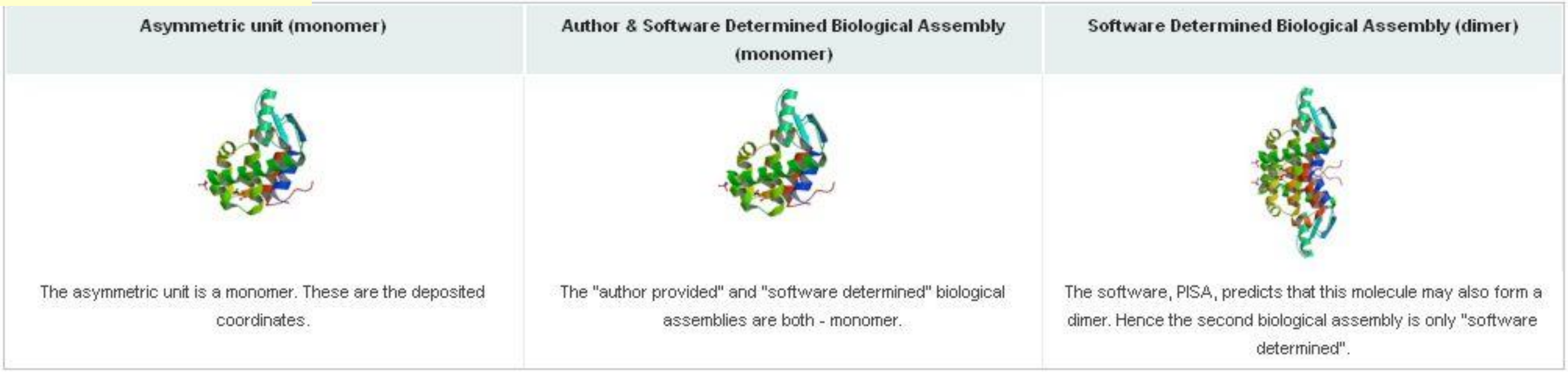


Jednostka biologiczna *(Biological Assembly)*

(A) author determined

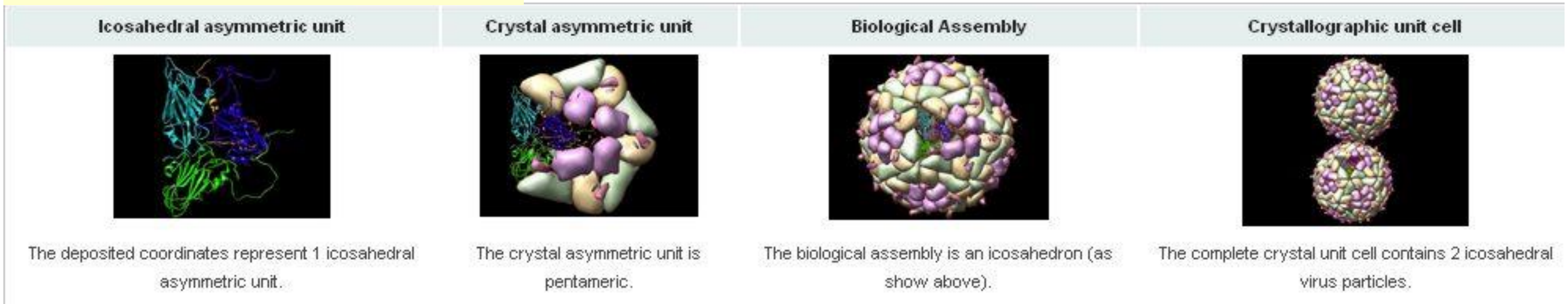
(S) software determined

Example: 3FAD



Asymmetric unit (monomer)	Author & Software Determined Biological Assembly (monomer)	Software Determined Biological Assembly (dimer)
The asymmetric unit is a monomer. These are the deposited coordinates.	The "author provided" and "software determined" biological assemblies are both - monomer.	The software, PISA, predicts that this molecule may also form a dimer. Hence the second biological assembly is only "software determined".

Example: 1QQP - viral capsid



Icosahedral asymmetric unit	Crystal asymmetric unit	Biological Assembly	Crystallographic unit cell
The deposited coordinates represent 1 icosahedral asymmetric unit.	The crystal asymmetric unit is pentameric.	The biological assembly is an icosahedron (as show above).	The complete crystal unit cell contains 2 icosahedral virus particles.

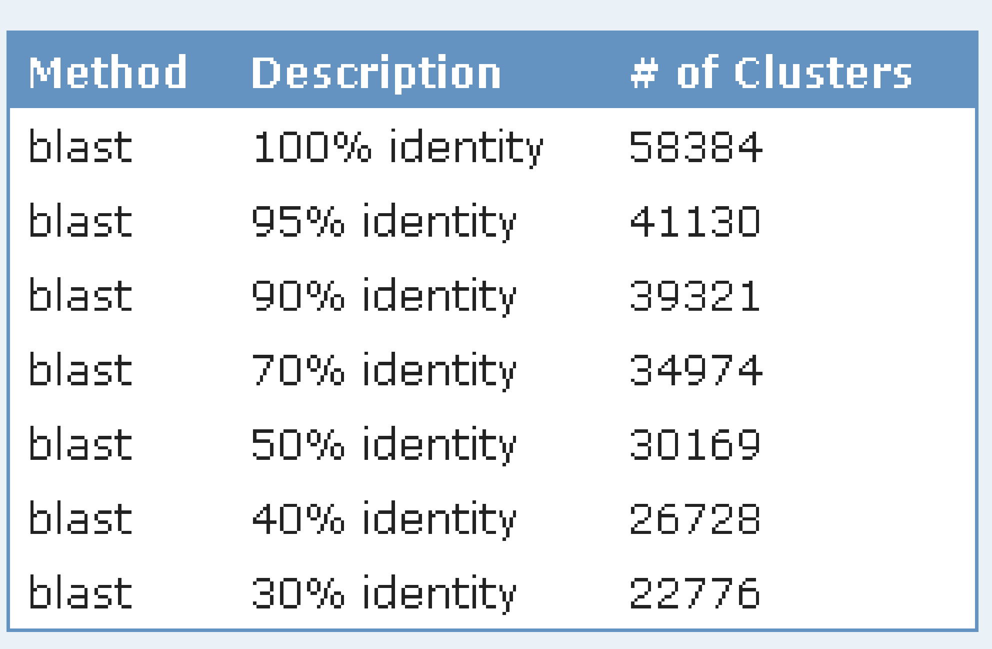
http://pdb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/bioassembly_tutorial.html

Redundancja

Important notes:

- Sequence similarity is defined on a chain basis, but results are returned on a structure basis.
- Many structures in the PDB contain multiple protein chains, or even hybrids of DNA or RNA and protein chains.
- Sequence similarity is only assessed for protein chains.
- The relationship between sequence similarity and structure similarity is complex.

Clasterization algorithm is described here:
http://www.pdb.org/pdb/statistics/clusterStatistics.do

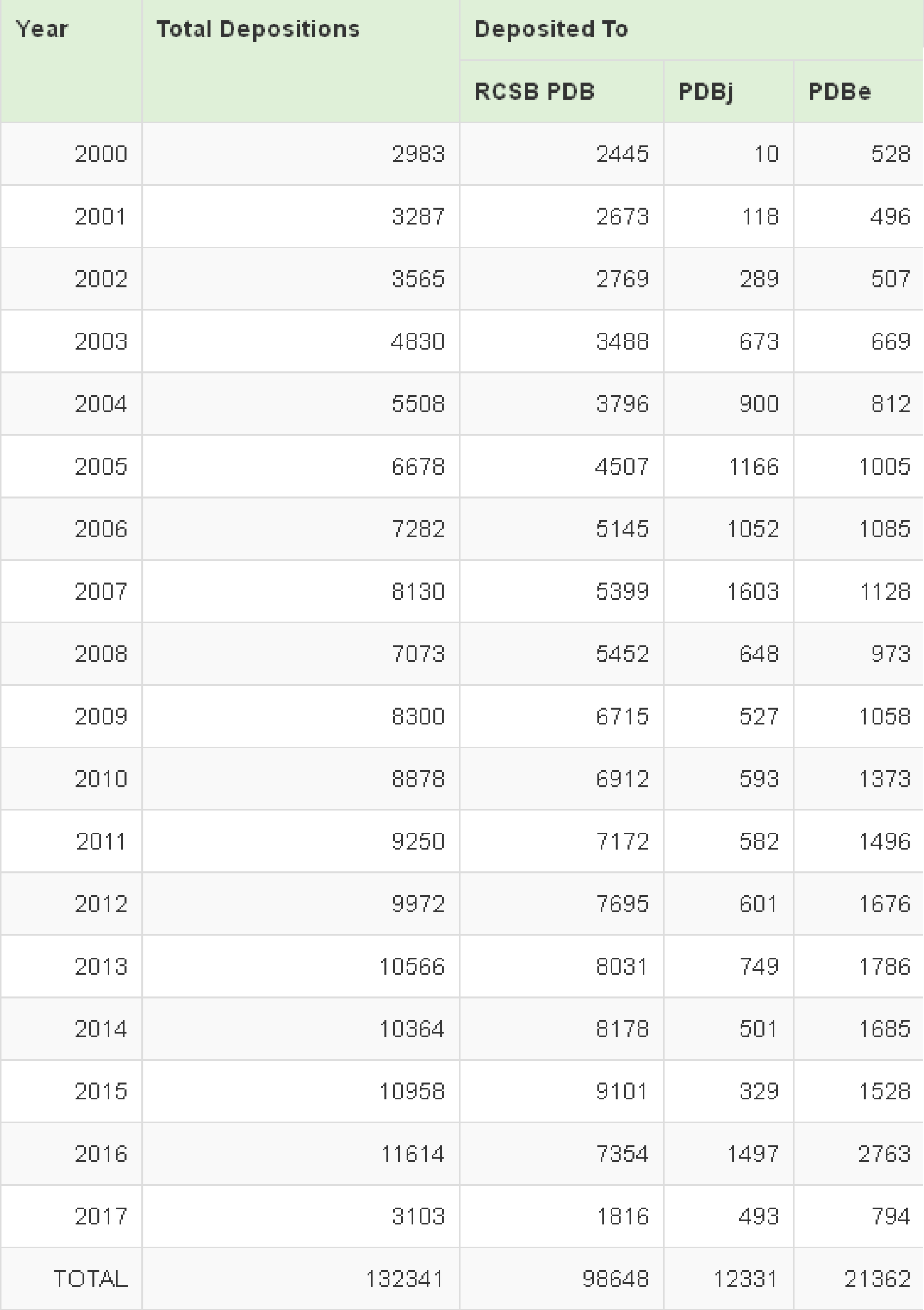


Method	Description	# of Clusters
blast	100% identity	58384
blast	95% identity	41130
blast	90% identity	39321
blast	70% identity	34974
blast	50% identity	30169
blast	40% identity	26728
blast	30% identity	22776

ftp://resources.rcsb.org/sequence/clusters/

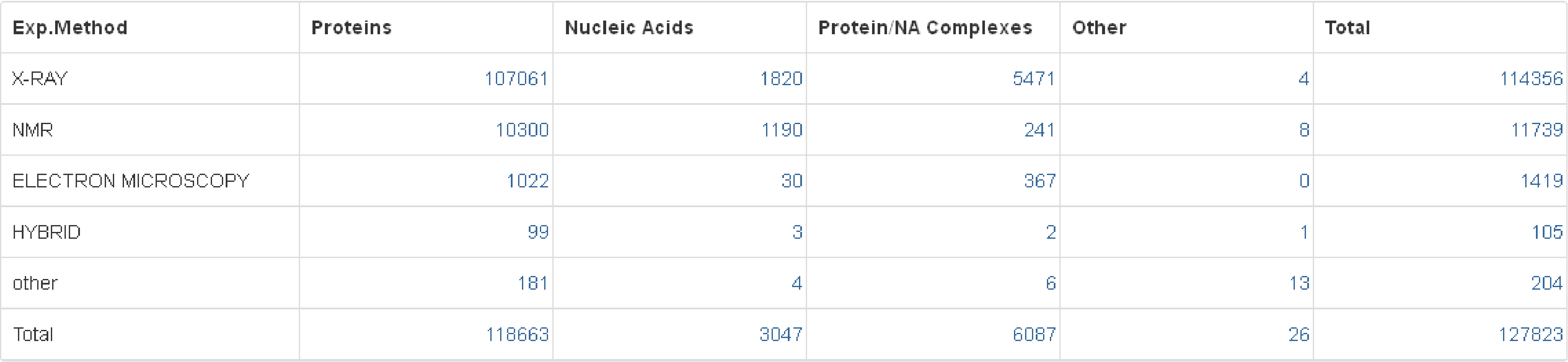
Aktualizacje

- Baza aktualizowana jest regularnie co tydzień
- Poszczególne „oddziały” wymieniają się informacjami
- Możliwość ustawienia powiadomień o nowych strukturach
- Możliwość automatycznej aktualizacji lokalnej kopii danych (np. za pomocą Bio.PDBList)



Year	Total Depositions	Deposited To		
		RCSB PDB	PDBj	PDBe
2000	2983	2445	10	528
2001	3287	2673	118	496
2002	3565	2769	289	507
2003	4830	3488	673	669
2004	5508	3796	900	812
2005	6678	4507	1166	1005
2006	7282	5145	1052	1085
2007	8130	5399	1603	1128
2008	7073	5452	648	973
2009	8300	6715	527	1058
2010	8878	6912	593	1373
2011	9250	7172	582	1496
2012	9972	7695	601	1676
2013	10566	8031	749	1786
2014	10364	8178	501	1685
2015	10958	9101	329	1528
2016	11614	7354	1497	2763
2017	3103	1816	493	794
TOTAL	132341	98648	12331	21362

Statystyki



Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	107061	1820	5471	4	114356
NMR	10300	1190	241	8	11739
ELECTRON MICROSCOPY	1022	30	367	0	1419
HYBRID	99	3	2	1	105
other	181	4	6	13	204
Total	118663	3047	6087	26	127823

wg stanu na 5.03.2018: **138270** struktur
(http://www.pdb.org/pdb/statistics/holdings.do)

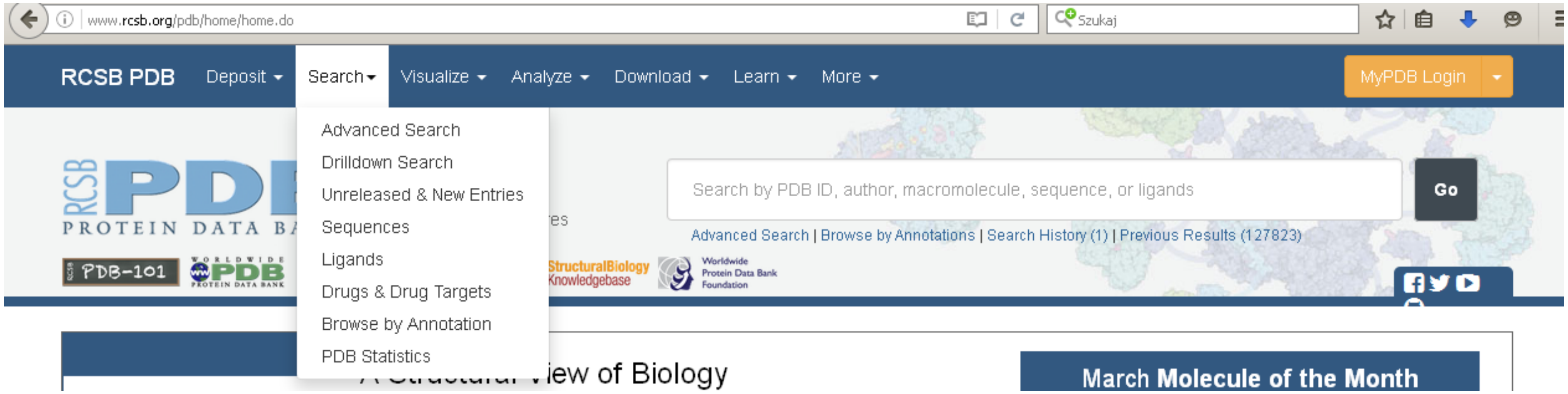
ale tylko **71153** różnych struktur
(http://www.pdb.org/pdb/statistics/clusterStatistics.do)

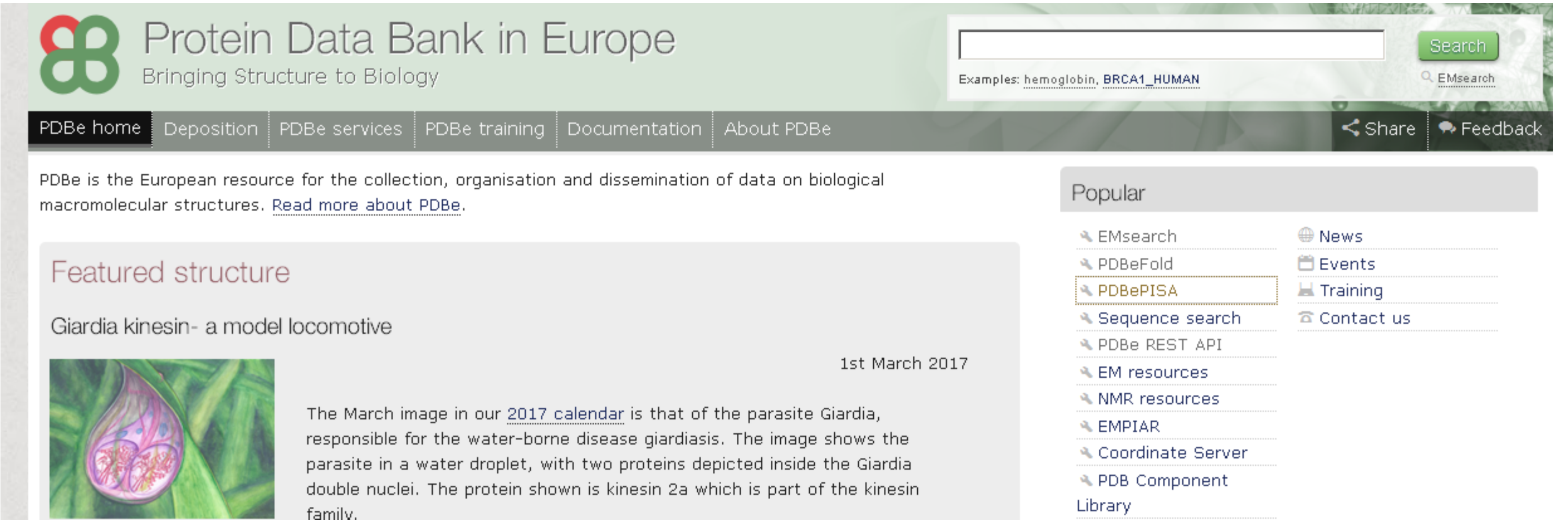
tylko **1195** różnych zwojów *(1205 wg SCOPe)*
(http://scop.mrc-lmb.cam.ac.uk/scop/count.html)

podczas, gdy znamy 80204459 sekwencji
(http://www.ebi.ac.uk/uniprot/TrEMBLstats/)

w tym 553941 sekwencji zweryfikowanych
(http://web.expasy.org/docs/relnotes/relstat.html)

Serwisy www


www.rcsb.org/pdb/home/home.do
Szukaj
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB Login
Advanced Search
Drilldown Search
Unreleased & New Entries
Sequences
Ligands
Drugs & Drug Targets
Browse by Annotation
PDB Statistics
PROTEIN DATA BA
PDB-101
WORLDWIDE PDB PROTEIN DATA BANK
StructuralBiology Knowledgebase
Worldwide Protein Data Bank Foundation
Search by PDB ID, author, macromolecule, sequence, or ligands
Go
Advanced Search | Browse by Annotations | Search History (1) | Previous Results (127823)
A Structural View of Biology
March Molecule of the Month



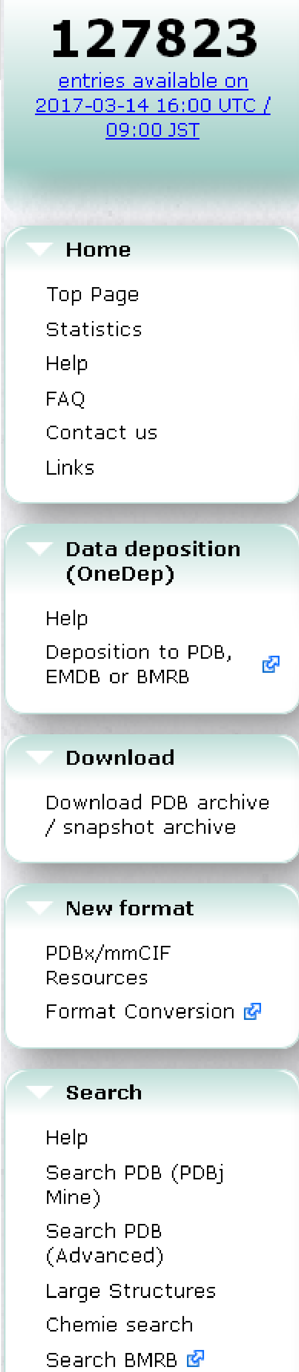
PDBe is the European resource for the collection, organisation and dissemination of data on biological macromolecular structures. Read more about PDBe.

Featured structure

Giardia kinesin- a model locomotive

1st March 2017

The March image in our 2017 calendar is that of the parasite Giardia, responsible for the water-borne disease giardiasis. The image shows the parasite in a water droplet, with two proteins depicted inside the Giardia double nuclei. The protein shown is kinesin 2a which is part of the kinesin family.

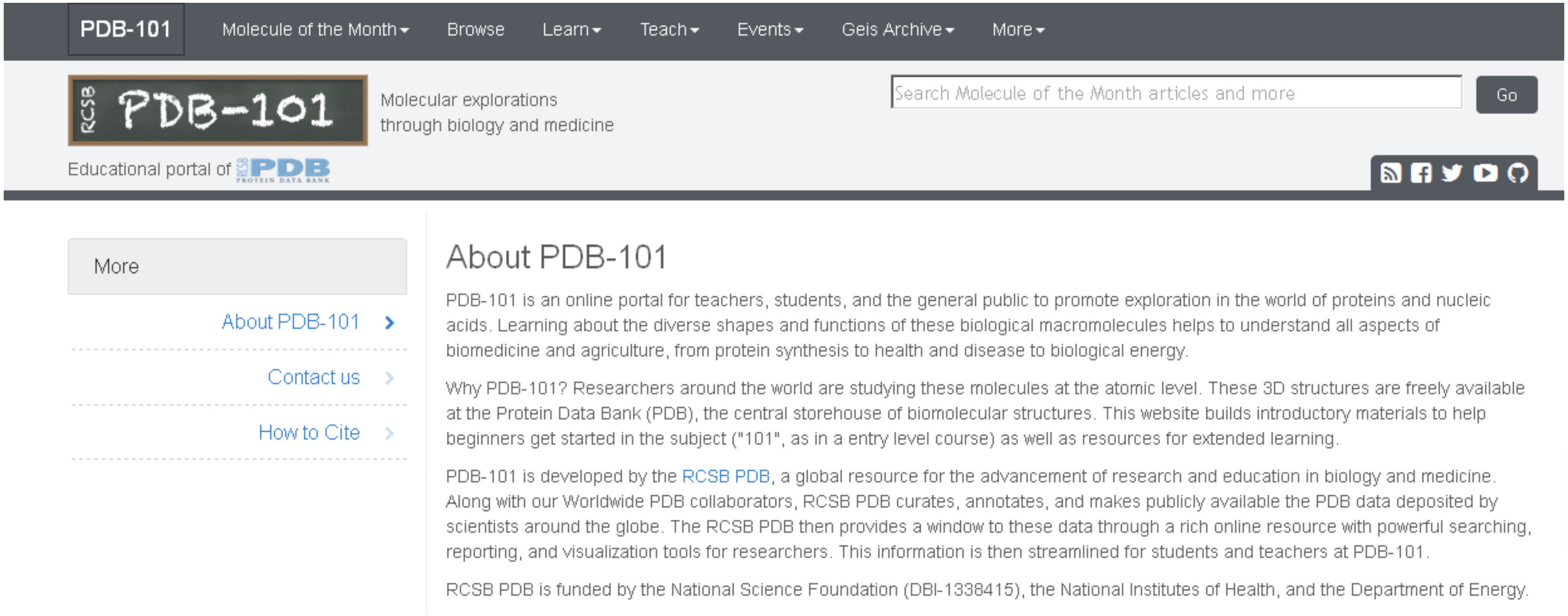


http://www.rcsb.org

https://www.ebi.ac.uk/pdbe/

https://pdbj.org/

PDB101



About PDB-101

PDB-101 is an online portal for teachers, students, and the general public to promote exploration in the world of proteins and nucleic acids. Learning about the diverse shapes and functions of these biological macromolecules helps to understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease to biological energy.

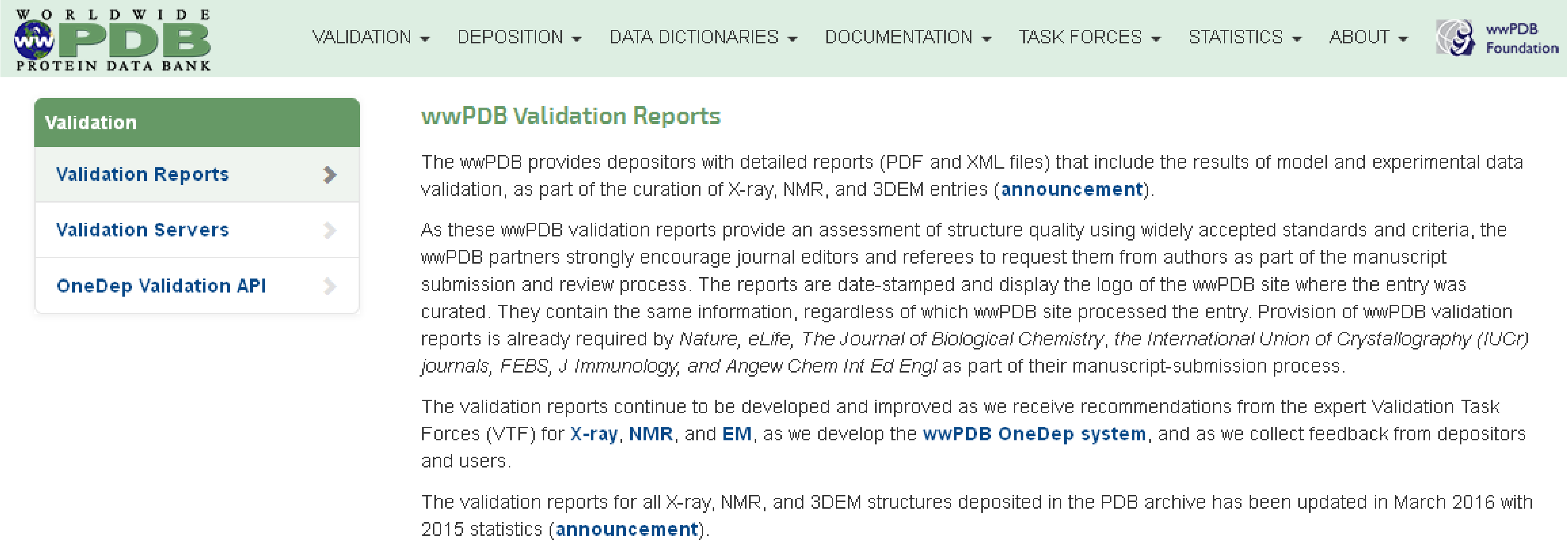
Why PDB-101? Researchers around the world are studying these molecules at the atomic level. These 3D structures are freely available at the Protein Data Bank (PDB), the central storehouse of biomolecular structures. This website builds introductory materials to help beginners get started in the subject ("101", as in a entry level course) as well as resources for extended learning.

PDB-101 is developed by the RCSB PDB, a global resource for the advancement of research and education in biology and medicine. Along with our Worldwide PDB collaborators, RCSB PDB curates, annotates, and makes publicly available the PDB data deposited by scientists around the globe. The RCSB PDB then provides a window to these data through a rich online resource with powerful searching, reporting, and visualization tools for researchers. This information is then streamlined for students and teachers at PDB-101.

RCSB PDB is funded by the National Science Foundation (DBI-1338415), the National Institutes of Health, and the Department of Energy.

http://pdb101.rcsb.org/more/about-pdb-101

Ocena jakości struktur



WORLDWIDE PDB PROTEIN DATA BANK
VALIDATION ▾ DEPOSITION ▾ DATA DICTIONARIES ▾ DOCUMENTATION ▾ TASK FORCES ▾ STATISTICS ▾ ABOUT ▾
wwPDB Foundation
Validation
Validation Reports
Validation Servers
OneDep Validation API

wwPDB Validation Reports

The wwPDB provides depositors with detailed reports (PDF and XML files) that include the results of model and experimental data validation, as part of the curation of X-ray, NMR, and 3DEM entries (**announcement**).

As these wwPDB validation reports provide an assessment of structure quality using widely accepted standards and criteria, the wwPDB partners strongly encourage journal editors and referees to request them from authors as part of the manuscript submission and review process. The reports are date-stamped and display the logo of the wwPDB site where the entry was curated. They contain the same information, regardless of which wwPDB site processed the entry. Provision of wwPDB validation reports is already required by *Nature, eLife, The Journal of Biological Chemistry*, *the International Union of Crystallography (IUCr) journals, FEBS, J Immunology, and Angew Chem Int Ed Engl* as part of their manuscript-submission process.

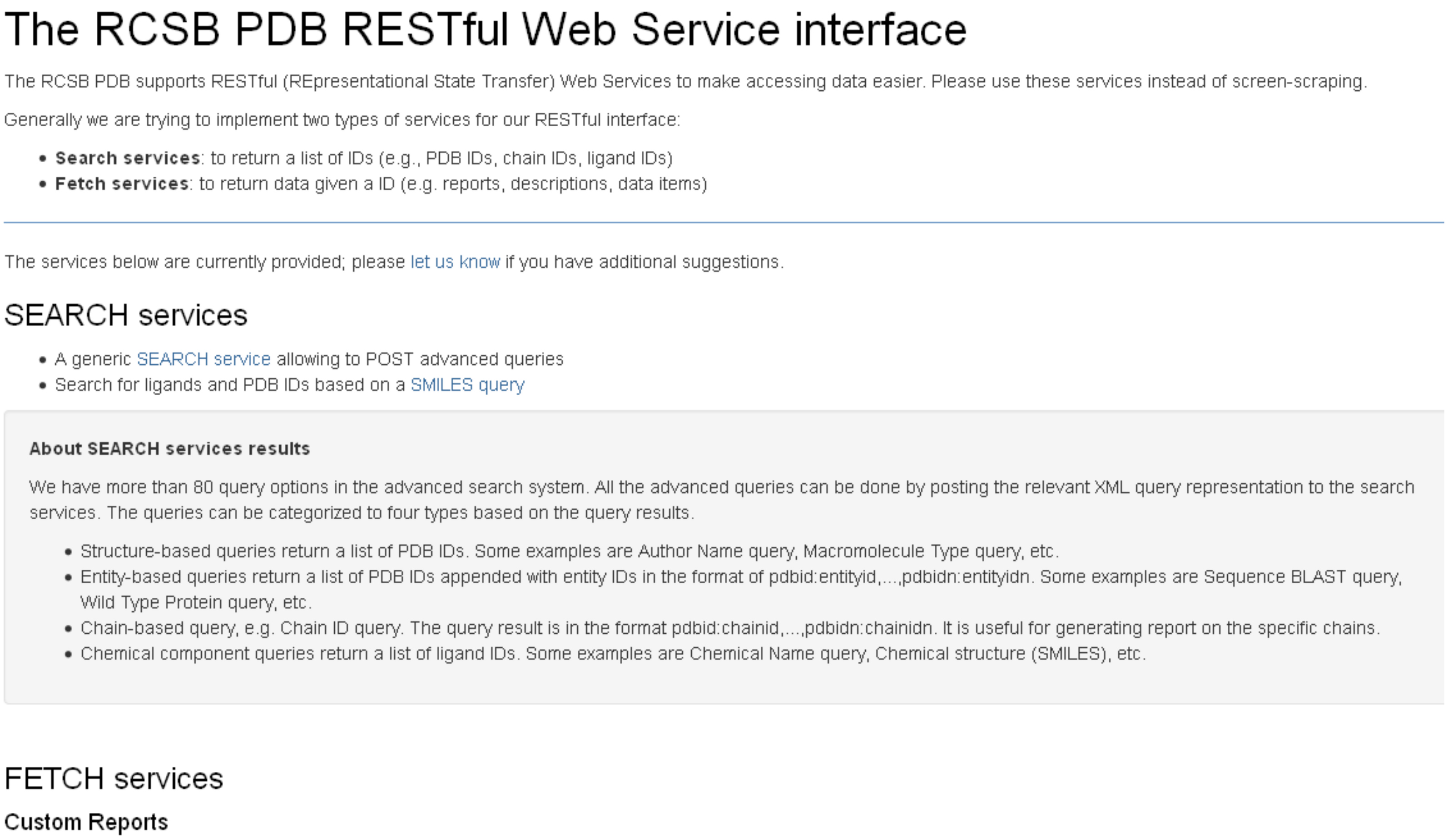
The validation reports continue to be developed and improved as we receive recommendations from the expert Validation Task Forces (VTF) for **X-ray**, **NMR**, and **EM**, as we develop the **wwPDB OneDep system**, and as we collect feedback from depositors and users.

The validation reports for all X-ray, NMR, and 3DEM structures deposited in the PDB archive has been updated in March 2016 with 2015 statistics (**announcement**).

https://www.wwpdb.org/validation/validation-reports

RCSB API

http://www.rcsb.org/pdb/software/rest.do



The RCSB PDB RESTful Web Service interface

The RCSB PDB supports RESTful (REpresentational State Transfer) Web Services to make accessing data easier. Please use these services instead of screen-scraping.

Generally we are trying to implement two types of services for our RESTful interface:

- **Search services**: to return a list of IDs (e.g., PDB IDs, chain IDs, ligand IDs)
- **Fetch services**: to return data given a ID (e.g. reports, descriptions, data items)

The services below are currently provided; please let us know if you have additional suggestions.

SEARCH services

- A generic SEARCH service allowing to POST advanced queries
- Search for ligands and PDB IDs based on a SMILES query

About SEARCH services results

We have more than 80 query options in the advanced search system. All the advanced queries can be done by posting the relevant XML query representation to the search services. The queries can be categorized to four types based on the query results.

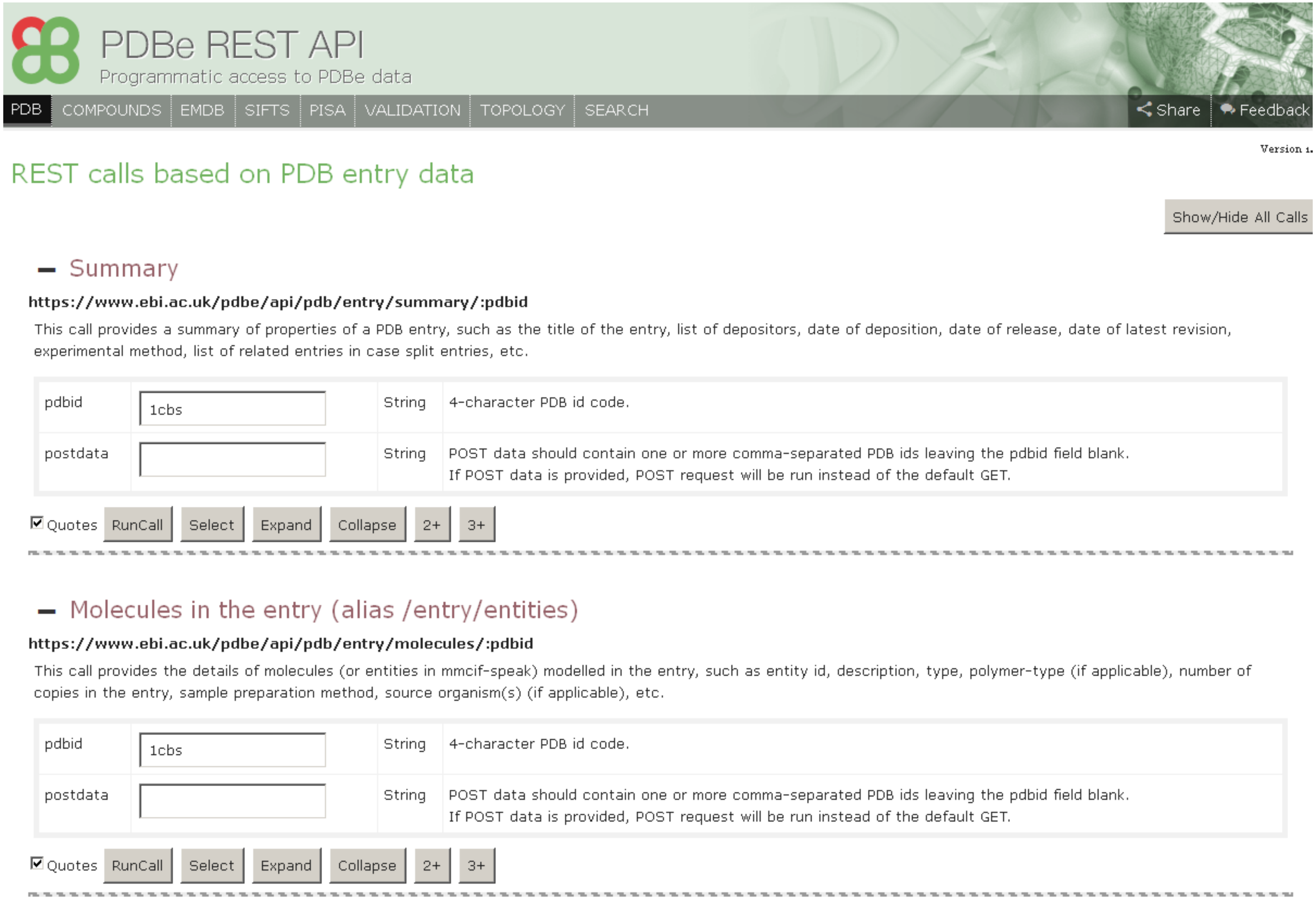
- Structure-based queries return a list of PDB IDs. Some examples are Author Name query, Macromolecule Type query, etc.
- Entity-based queries return a list of PDB IDs appended with entity IDs in the format of pdbid:entityid,...,pdbidn:entityidn. Some examples are Sequence BLAST query, Wild Type Protein query, etc.
- Chain-based query, e.g. Chain ID query. The query result is in the format pdbid:chainid,...,pdbidn:chainidn. It is useful for generating report on the specific chains.
- Chemical component queries return a list of ligand IDs. Some examples are Chemical Name query, Chemical structure (SMILES), etc.

FETCH services

Custom Reports

PDBe API

https://www.ebi.ac.uk/pdbe/api/doc/



REST calls based on PDB entry data

— Summary

https://www.ebi.ac.uk/pdbe/api/pdb/entry/summary/:pdbid

This call provides a summary of properties of a PDB entry, such as the title of the entry, list of depositors, date of deposition, date of release, date of latest revision, experimental method, list of related entries in case split entries, etc.

pdbid	1cbs	String	4-character PDB id code.
postdata		String	POST data should contain one or more comma-separated PDB ids leaving the pdbid field blank. If POST data is provided, POST request will be run instead of the default GET.

☑ Quotes

— Molecules in the entry (alias /entry/entities)

https://www.ebi.ac.uk/pdbe/api/pdb/entry/molecules/:pdbid

This call provides the details of molecules (or entities in mmcif-speak) modelled in the entry, such as entity id, description, type, polymer-type (if applicable), number of copies in the entry, sample preparation method, source organism(s) (if applicable), etc.

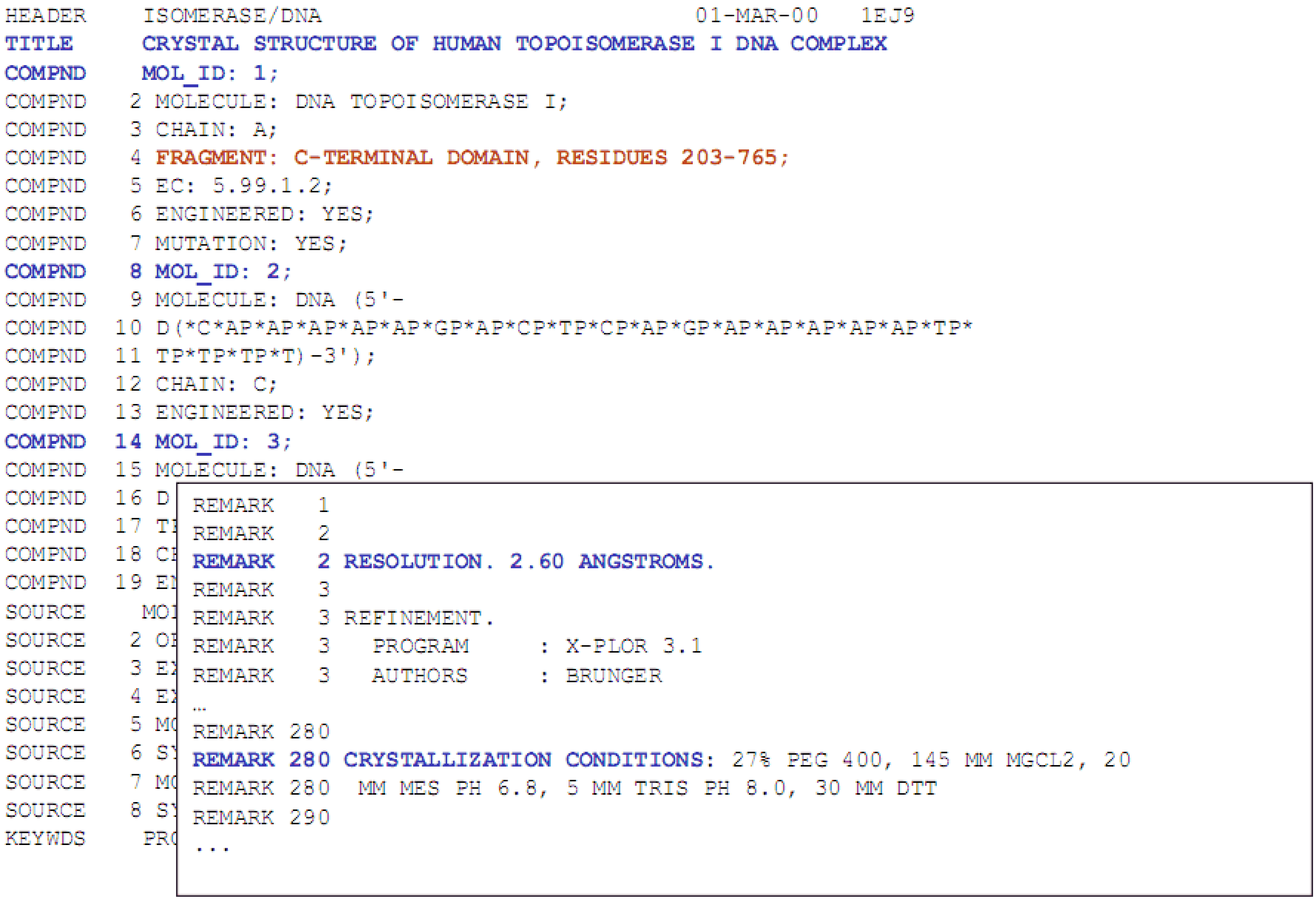
pdbid	1cbs	String	4-character PDB id code.
postdata		String	POST data should contain one or more comma-separated PDB ids leaving the pdbid field blank. If POST data is provided, POST request will be run instead of the default GET.

☑ Quotes

Formaty danych strukturalnych

- PDB
- PDBML (XML)
- PDBx / mmCif

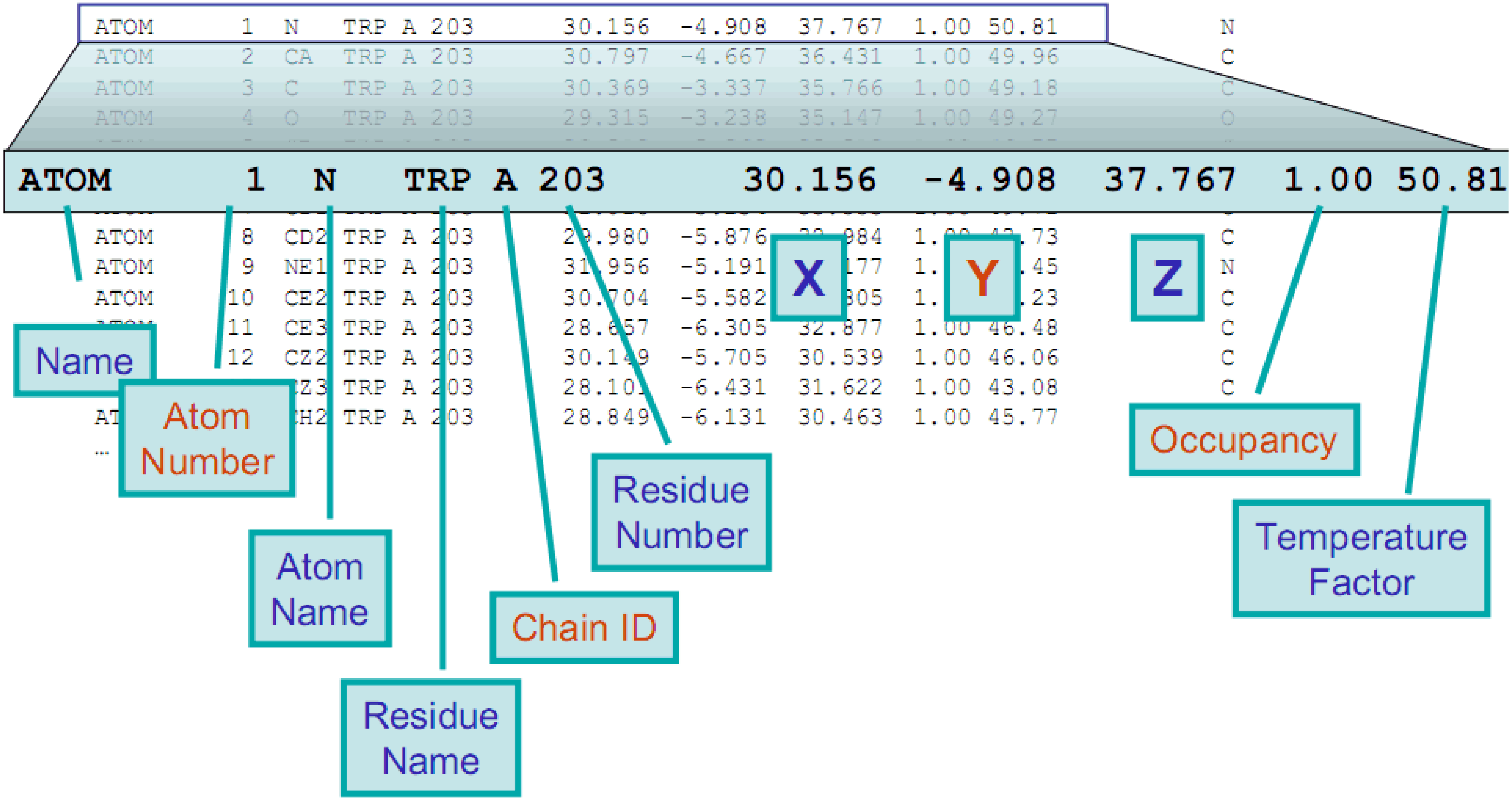
Format PDB - nagłówek



```
HEADER    ISOMERASE/DNA                          01-MAR-00   1EJ9
TITLE     CRYSTAL STRUCTURE OF HUMAN TOPOISOMERASE I DNA COMPLEX
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: DNA TOPOISOMERASE I;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: C-TERMINAL DOMAIN, RESIDUES 203-765;
COMPND   5 EC: 5.99.1.2;
COMPND   6 ENGINEERED: YES;
COMPND   7 MUTATION: YES;
COMPND   8 MOL_ID: 2;
COMPND   9 MOLECULE: DNA (5'-
COMPND  10 D(*C*AP*AP*AP*AP*AP*GP*AP*CP*TP*CP*AP*GP*AP*AP*AP*AP*AP*TP*
COMPND  11 TP*TP*TP*T)-3');
COMPND  12 CHAIN: C;
COMPND  13 ENGINEERED: YES;
COMPND  14 MOL_ID: 3;
COMPND  15 MOLECULE: DNA (5'-
COMPND  16 D
COMPND  17 TI
COMPND  18 CI
COMPND  19 EI
SOURCE    MOI
SOURCE   2 OI
SOURCE   3 EX
SOURCE   4 EX
SOURCE   5 MC
SOURCE   6 SY
SOURCE   7 MC
SOURCE   8 SY
KEYWDS    PRC
```

```
REMARK   1
REMARK   2
REMARK   2 RESOLUTION. 2.60 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : X-PLOR 3.1
REMARK   3   AUTHORS     : BRUNGER
…
REMARK 280
REMARK 280 CRYSTALLIZATION CONDITIONS: 27% PEG 400, 145 MM MGCL2, 20
REMARK 280  MM MES PH 6.8, 5 MM TRIS PH 8.0, 30 MM DTT
REMARK 290
...
```

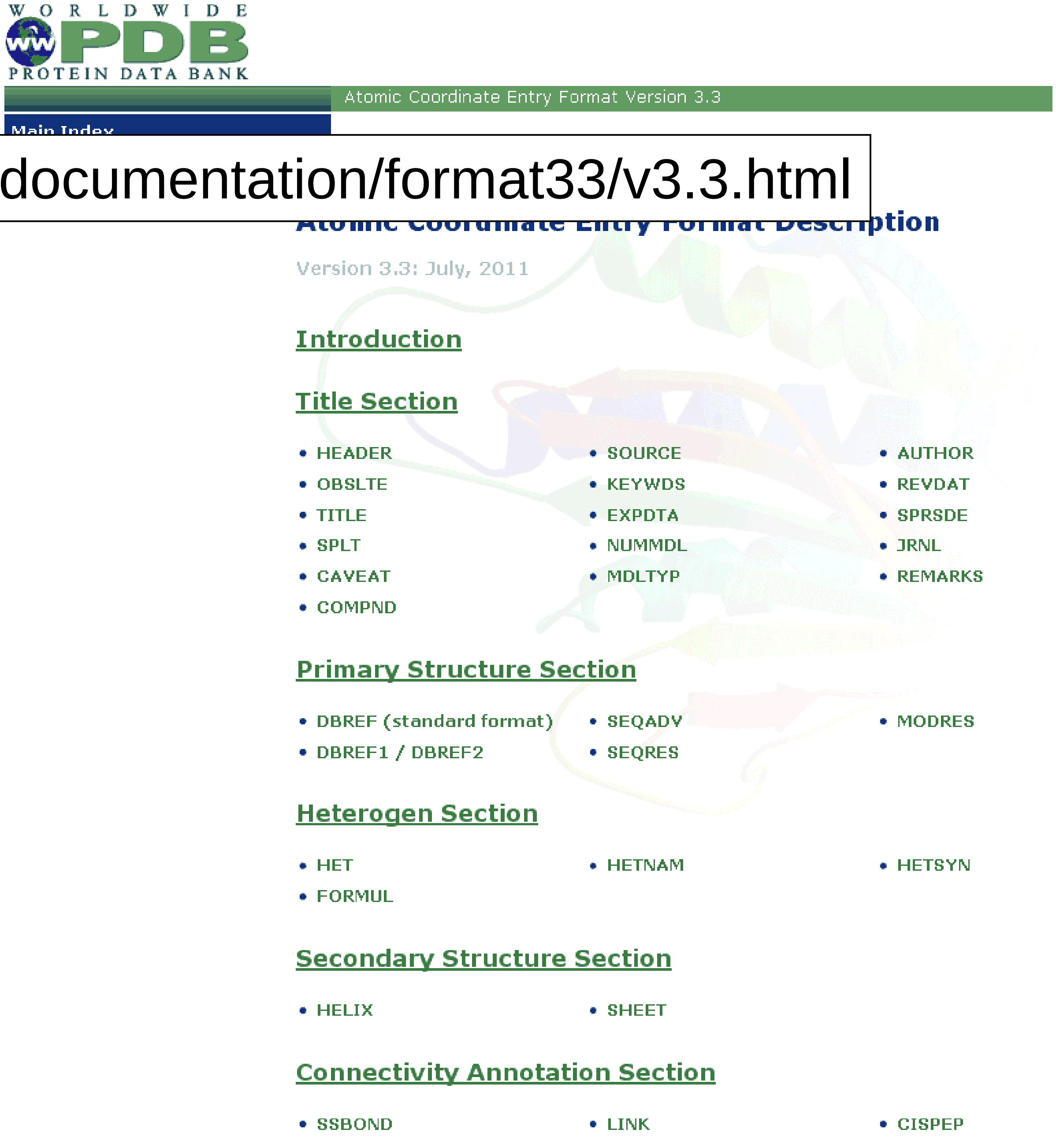
Format PDB - współrzędne



```
ATOM      1  N   TRP A 203      30.156  -4.908  37.767  1.00 50.81           N
ATOM      2  CA  TRP A 203      30.797  -4.667  36.431  1.00 49.96           C
ATOM      3  C   TRP A 203      30.369  -3.337  35.766  1.00 49.18           C
ATOM      4  O   TRP A 203      29.315  -3.238  35.147  1.00 49.27           O

ATOM      1  N   TRP A 203      30.156  -4.908  37.767  1.00 50.81

ATOM      8  CD2 TRP A 203      2[illegible].980  -5.876  [illegible]984  1.[illegible]  [illegible].73           C
ATOM      9  NE1 TRP A 203      31[illegible].956  -5.191  [illegible]177  1.[illegible]  [illegible].45           N
ATOM     10  CE2 TRP A 203      30.[illegible]704  -5.582  [illegible]305  1.[illegible]  [illegible].23           C
ATOM     11  CE3 TRP A 203      28.6[illegible]57  -6.305  32.877  1.00 46.48           C
ATOM     12  CZ2 TRP A 203      30.1[illegible]49  -5.705  30.539  1.00 46.06           C
[illegible]  CZ3 TRP A 203      28.10[illegible]  -6.431  31.622  1.00 43.08           C
AT[illegible]  CH2 TRP A 203      28.849  -6.131  30.463  1.00 45.77
…
```

Format PDB - dokumentacja

http://wwpdb.org/documentation/format33/v3.3.html



Atomic Coordinate Entry Format Version 3.3

Main Index

Atomic Coordinate Entry Format Description

Version 3.3: July, 2011

Introduction

Title Section

- HEADER
- OBSLTE
- TITLE
- SPLT
- CAVEAT
- COMPND
- SOURCE
- KEYWDS
- EXPDTA
- NUMMDL
- MDLTYP
- AUTHOR
- REVDAT
- SPRSDE
- JRNL
- REMARKS

Primary Structure Section

- DBREF (standard format)
- DBREF1 / DBREF2
- SEQADV
- SEQRES
- MODRES

Heterogen Section

- HET
- FORMUL
- HETNAM
- HETSYN

Secondary Structure Section

- HELIX
- SHEET

Connectivity Annotation Section

- SSBOND
- LINK
- CISPEP

Format PDB - problemy

Związane z błędami w strukturze pliku PDB:

- powtórzone residuum;

- powtórzony atom;

- brak alternatywnej informacji dla niejednoznacznego atomu;

- urwany łańcuch.

Związane z niedoskonałością formatu:

- brak informacji o wiązaniach;

- maksymalna liczba atomów w modelu: 99999 (pięcioznakowe pole na numer atomu);

- maksymalna liczba łańcuchów: 26 (identyfikator jednoliterowy).

Odmiany formatu PDB

Istnieją narzędzia stosujące własne odmiany plików PDB (niezgodne ze standardem). Stąd te same pliki mogą nie być kompatybilne z różnymi programami.

Format mmCIF – informacje dla użytkowników

1. **The PDB file format will be phased out in 2016.**
2. PDBx/mmCIF will become the standard PDB archive format in 2014.
3. All PDB data processing and annotation will be performed using PDBx/mmCIF at all wwPDB sites.
4. PDBx/mmCIF consists of categories of information represented as tables and keyword value pairs.
5. The categories in PDBx/mmCIF have explicit relationships with one another.
6. **PDBx/mmCIF imposes no limitations for the number of atoms, residues or chains that can be represented in a single PDB entry (no split entries!).**
7. Each data item in a PDBx/mmCIF file is precisely defined in a PDBx Exchange Data Dictionary The content of data dictionary is fully software accessible.
8. All of the data items in the current PDB format have corresponding items data items in the PDBx/mmCIF format.
9. Chemical descriptions of all of the monomers and ligands in PDB entries are provided in the PDB Chemical Component Dictionary which is in PDBx/mmCIF format.
10. PDBx/mmCIF is supported by visualization applications such as Jmol, Chimera, and OpenRasMol as well as structure determination systems such as CCP4 and Phenix.

Format mmCIF – ważne dla programistów

1. The format is based on a context-free grammar. PDBx/mmCIF has a simple grammar. Data are presented in either key-value or tabular form. It is much easier to parse than the record-oriented PDB format. Say good-bye to "exception" handling when reading old-style PDB flat files!
2. There are no column width limitations.
3. All relationships between common data items (e.g. atom and residue identifiers) are explicitly documented within the PDBx Exchange Dictionary. This permits software applications to evaluate and validate referential integrity with any PDB entry.
4. The mmCIF/PDBx Exchange Dictionary provides metadata (e.g. data types, allowed ranges, controlled vocabularies) which can be used to generate a validating mmCIF parser or a database loader.
5. Parsing tools are available in most popular languages (e.g. C/C++, Java, Python, Perl, FORTRAN) and toolkits (e.g. BioJava and biopython).
6. Mapping information between the residue sequences of the experimental sample and the model coordinates is included within each entry.
7. PDB Chemical reference data are maintained and distributed in PDBx/mmCIF format.

Format mmCIF – dodatkowe informacje

Dictionary index:
http://mmcif.wwpdb.org/dictionaries/mmcif_pdbx_v40.dic/Index/

PDB to PDBx/mmCIF Data Item Correspondences:
http://mmcif.wwpdb.org/docs/pdb_to_pdbx_correspondences.html

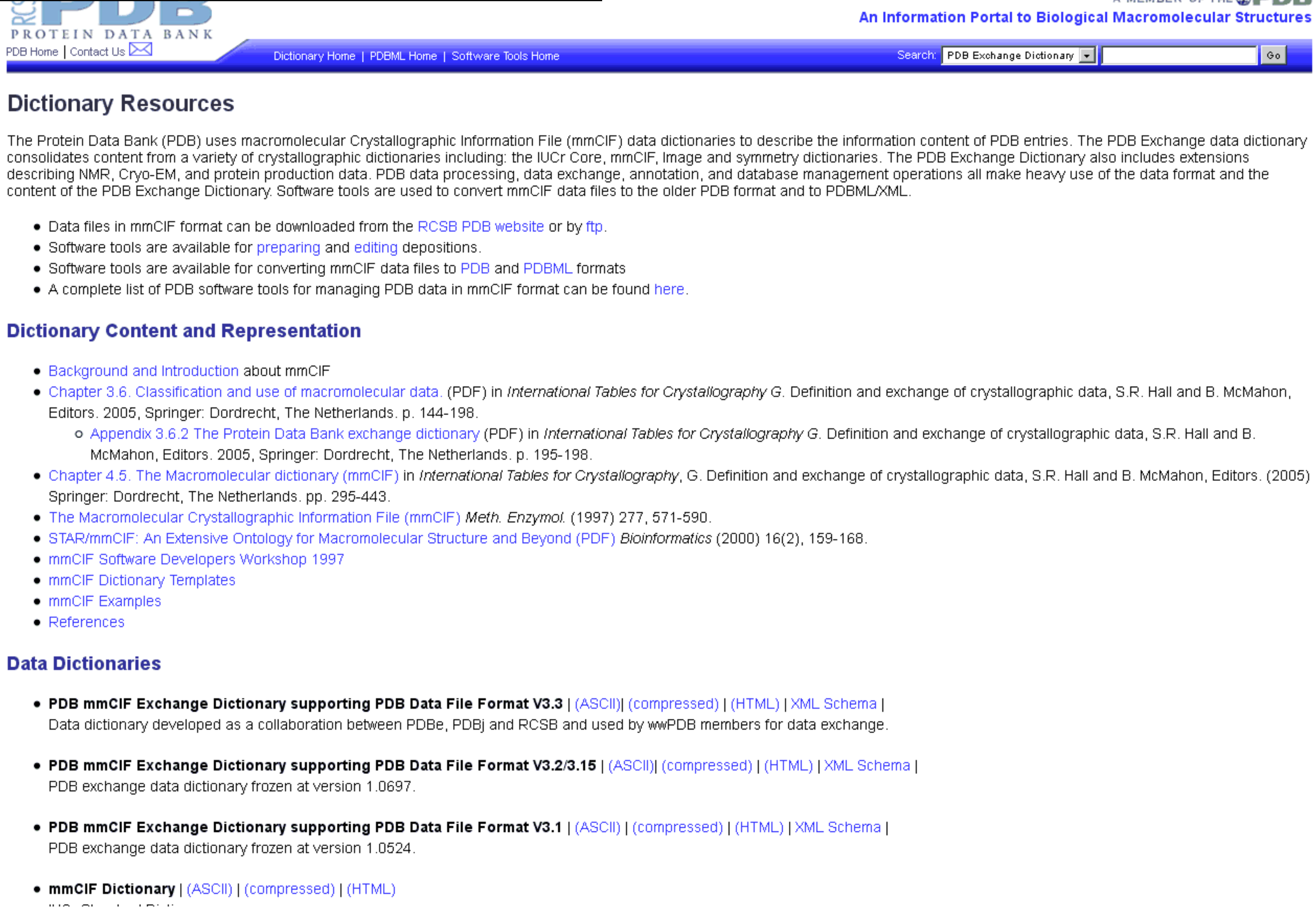
Large Structure mmCIF/PDBx Examples:
http://mmcif.wwpdb.org/docs/large-pdbx-examples/index.html

PDBx/mmCIF Software Resources:
http://mmcif.wwpdb.org/docs/software-resources.html

Crystallographic Information File (CIF) Specification:
http://www.iucr.org/resources/cif/spec/version1.1

Format mmCIF – online

http://mmcif.pdb.org/index.html


PROTEIN DATA BANK
PDB Home | Contact Us
A MEMBER OF THE PDB
An Information Portal to Biological Macromolecular Structures
Dictionary Home | PDBML Home | Software Tools Home
Search: PDB Exchange Dictionary
Go

Dictionary Resources

The Protein Data Bank (PDB) uses macromolecular Crystallographic Information File (mmCIF) data dictionaries to describe the information content of PDB entries. The PDB Exchange data dictionary consolidates content from a variety of crystallographic dictionaries including: the IUCr Core, mmCIF, Image and symmetry dictionaries. The PDB Exchange Dictionary also includes extensions describing NMR, Cryo-EM, and protein production data. PDB data processing, data exchange, annotation, and database management operations all make heavy use of the data format and the content of the PDB Exchange Dictionary. Software tools are used to convert mmCIF data files to the older PDB format and to PDBML/XML.

- Data files in mmCIF format can be downloaded from the RCSB PDB website or by ftp.
- Software tools are available for preparing and editing depositions.
- Software tools are available for converting mmCIF data files to PDB and PDBML formats
- A complete list of PDB software tools for managing PDB data in mmCIF format can be found here.

Dictionary Content and Representation

- Background and Introduction about mmCIF
- Chapter 3.6. Classification and use of macromolecular data. (PDF) in *International Tables for Crystallography G*. Definition and exchange of crystallographic data, S.R. Hall and B. McMahon, Editors. 2005, Springer: Dordrecht, The Netherlands. p. 144-198.
  - Appendix 3.6.2 The Protein Data Bank exchange dictionary (PDF) in *International Tables for Crystallography G*. Definition and exchange of crystallographic data, S.R. Hall and B. McMahon, Editors. 2005, Springer: Dordrecht, The Netherlands. p. 195-198.
- Chapter 4.5. The Macromolecular dictionary (mmCIF) in *International Tables for Crystallography*, G. Definition and exchange of crystallographic data, S.R. Hall and B. McMahon, Editors. (2005) Springer: Dordrecht, The Netherlands. pp. 295-443.
- The Macromolecular Crystallographic Information File (mmCIF) *Meth. Enzymol.* (1997) 277, 571-590.
- STAR/mmCIF: An Extensive Ontology for Macromolecular Structure and Beyond (PDF) *Bioinformatics* (2000) 16(2), 159-168.
- mmCIF Software Developers Workshop 1997
- mmCIF Dictionary Templates
- mmCIF Examples
- References

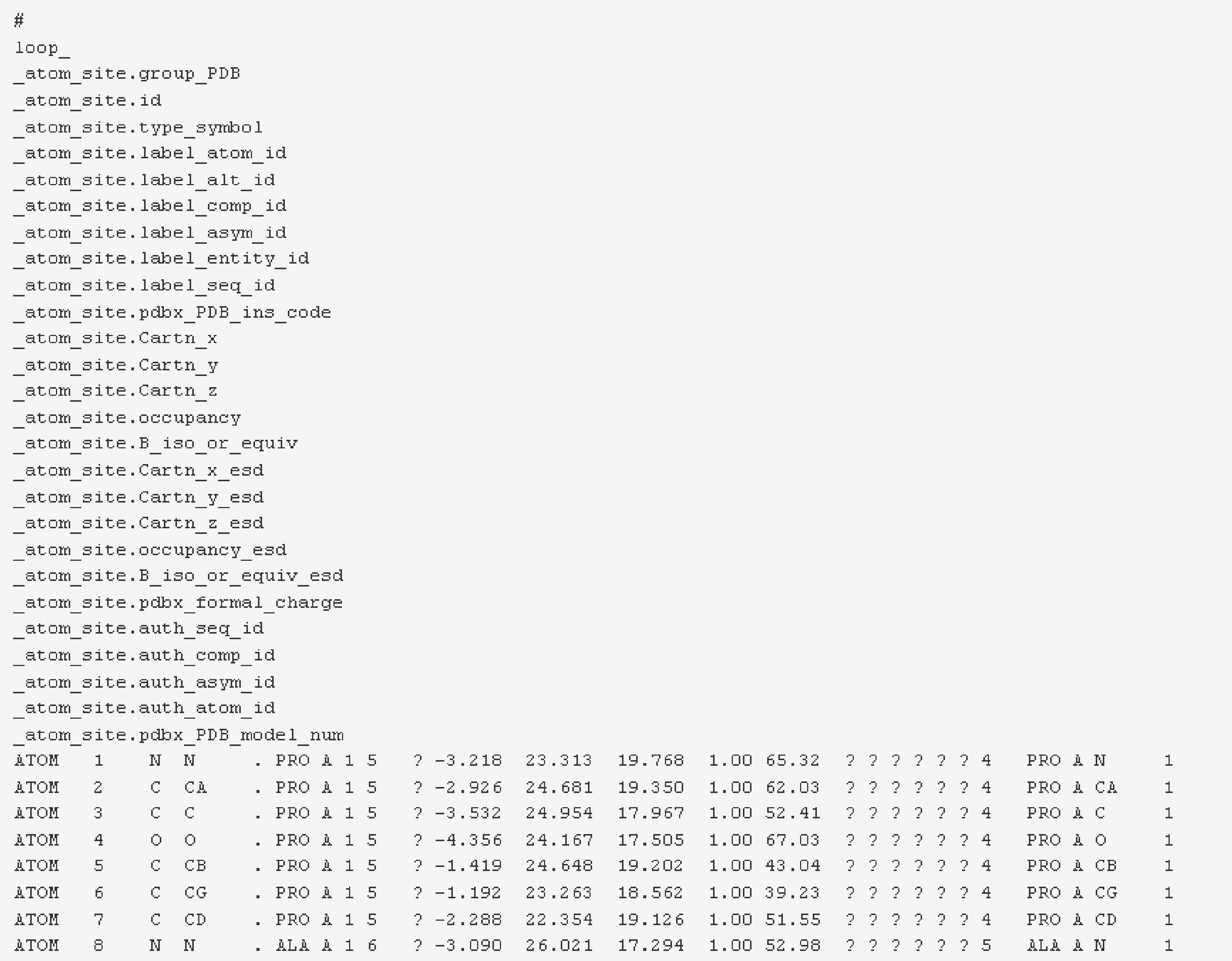
Data Dictionaries

- **PDB mmCIF Exchange Dictionary supporting PDB Data File Format V3.3** | (ASCII)| (compressed) | (HTML) | XML Schema |
  Data dictionary developed as a collaboration between PDBe, PDBj and RCSB and used by wwPDB members for data exchange.
- **PDB mmCIF Exchange Dictionary supporting PDB Data File Format V3.2/3.15** | (ASCII)| (compressed) | (HTML) | XML Schema |
  PDB exchange data dictionary frozen at version 1.0697.
- **PDB mmCIF Exchange Dictionary supporting PDB Data File Format V3.1** | (ASCII) | (compressed) | (HTML) | XML Schema |
  PDB exchange data dictionary frozen at version 1.0524.
- **mmCIF Dictionary** | (ASCII) | (compressed) | (HTML)

Format mmCIF – nagłówek

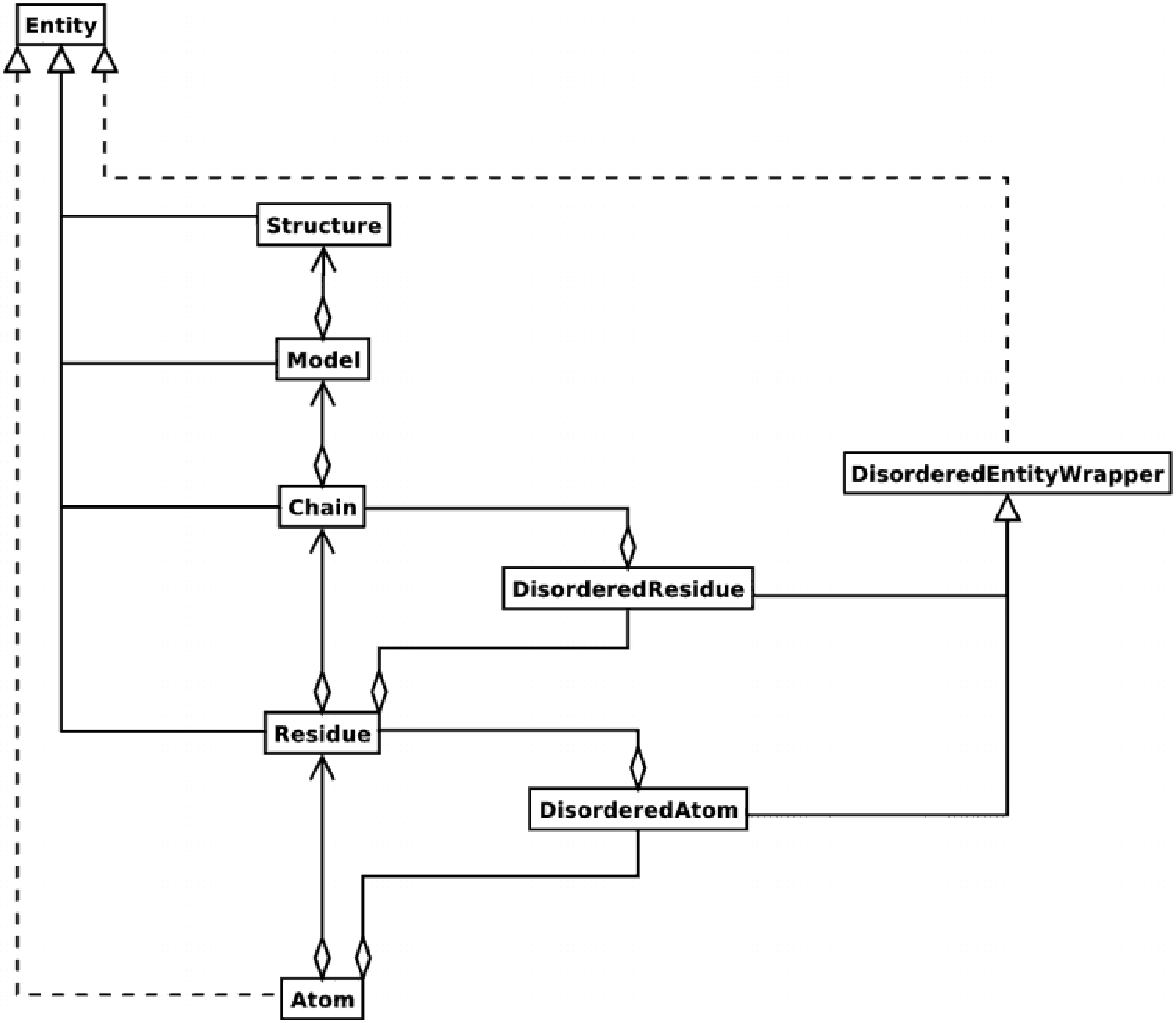
```
data_1EJ9
#
_entry.id   1EJ9
#
_audit_conform.dict_name       mmcif_pdbx.dic
_audit_conform.dict_version    4.007
_audit_conform.dict_location
http://mmcif.pdb.org/dictionaries/ascii/mmcif_p
dbx.dic
#
loop_
_database_2.database_id
_database_2.database_code
PDB   1EJ9
NDB   PD0125
RCSB RCSB010631
#
loop_
_database_PDB_rev.num
_database_PDB_rev.date
_database_PDB_rev.date_original
_database_PDB_rev.status
_database_PDB_rev.replaces
_database_PDB_rev.mod_type
1 2000-08-03 2000-03-01 ? 1EJ9 0
2 2009-02-24 ?          ? 1EJ9 1
#
_database_PDB_rev_record.rev_num   2
_database_PDB_rev_record.type      VERSN
_database_PDB_rev_record.details   ?
```

Format mmCIF – sekcja współrzędnych

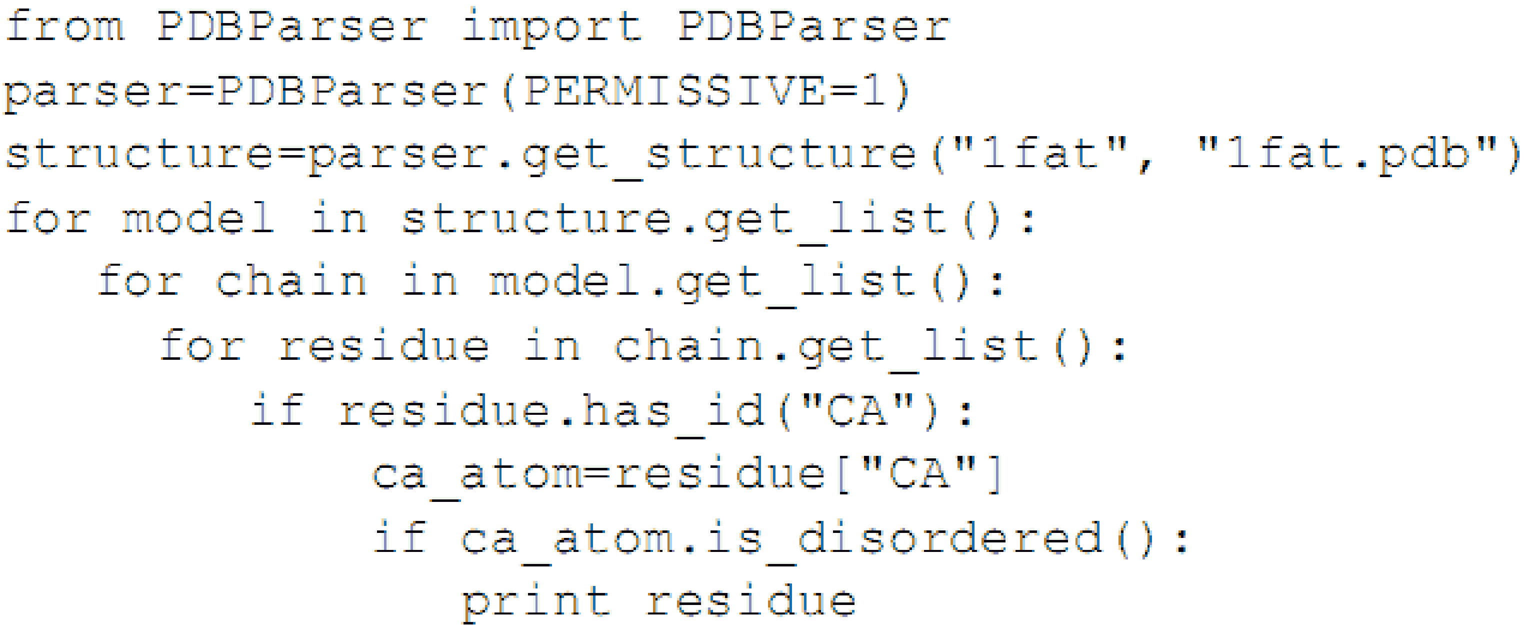


```
#
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.Cartn_x_esd
_atom_site.Cartn_y_esd
_atom_site.Cartn_z_esd
_atom_site.occupancy_esd
_atom_site.B_iso_or_equiv_esd
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM   1    N  N     . PRO A 1 5   ? -3.218  23.313  19.768  1.00 65.32  ? ? ? ? ? ? 4   PRO A N      1
ATOM   2    C  CA    . PRO A 1 5   ? -2.926  24.681  19.350  1.00 62.03  ? ? ? ? ? ? 4   PRO A CA     1
ATOM   3    C  C     . PRO A 1 5   ? -3.532  24.954  17.967  1.00 52.41  ? ? ? ? ? ? 4   PRO A C      1
ATOM   4    O  O     . PRO A 1 5   ? -4.356  24.167  17.505  1.00 67.03  ? ? ? ? ? ? 4   PRO A O      1
ATOM   5    C  CB    . PRO A 1 5   ? -1.419  24.648  19.202  1.00 43.04  ? ? ? ? ? ? 4   PRO A CB     1
ATOM   6    C  CG    . PRO A 1 5   ? -1.192  23.263  18.562  1.00 39.23  ? ? ? ? ? ? 4   PRO A CG     1
ATOM   7    C  CD    . PRO A 1 5   ? -2.288  22.354  19.126  1.00 51.55  ? ? ? ? ? ? 4   PRO A CD     1
ATOM   8    N  N     . ALA A 1 6   ? -3.090  26.021  17.294  1.00 52.98  ? ? ? ? ? ? 5   ALA A N      1
```

Parsowanie pliku PDB - biopython



Parsowanie pliku PDB – biopython, przykład



```
from PDBParser import PDBParser
parser=PDBParser(PERMISSIVE=1)
structure=parser.get_structure("1fat", "1fat.pdb")
for model in structure.get_list():
   for chain in model.get_list():
     for residue in chain.get_list():
        if residue.has_id("CA"):
           ca_atom=residue["CA"]
           if ca_atom.is_disordered():
              print residue
```

Prints all amino acids in 1FAT protein structure, which include disordered Cα atom.

Parsowanie pliku mmCIF – biopython

```
# Create an MMCIFParser object:
>>> from Bio.PDB.MMCIFParser import MMCIFParser
>>> parser = MMCIFParser()

# Create a structure object from the mmCIF file
>>> structure = parser.get_structure('1fat', '1fat.cif')
```

```
# To have some more low level access to an mmCIF file, you can use
# the MMCIF2Dict class to create a Python dictionary that maps all
# mmCIF tags in an mmCIF file to their values.
# If there are multiple values (like in the case of tag _atom_site.Cartn_y,
# which holds the y coordinates of all atoms), the tag is mapped
# to a list of values.
>>> from Bio.PDB.MMCIF2Dict import MMCIF2Dict
>>> mmcif_dict = MMCIF2Dict('1FAT.cif')
```

Parsowanie pliku mmCIF – biopython, przykłady

```
# Example1: get the solvent content
>>> sc = mmcif_dict['_exptl_crystal.density_percent_sol']

#Example2: get the list of the y coordinates of all atoms
>>> y_list = mmcif_dict['_atom_site.Cartn_y']
```