Zpracování informací a vizualizace v chemii

(C2150)

1. Úvod, databáze molekul

Organizační pokyny

- Přednášející: Martin Prokop
- Email: martinp@chemi.muni.cz
- Pracovna: INBIT/2.10 (v dubnu/květnu přesun do A4)
- Spoluautor přednášek: Zdeněk Kříž
- Web předmětu: **ncbr.chemi.muni.cz/~martinp/C2150**

- Účast na cvičeních je povinná
- Nepřítomnost může být omluvena pouze prostřednictvím studijního oddělení (omluva musí být zanesena v informačním systému MU)

- Podmínky pro udělení kolokvia:
  - účast na cvičeních
  - odevzdané úlohy ze cvičení
  - odevzdaný závěrečný projekt

Osnova kurzu

- Strukturní databáze molekul a jejich prohledávání
- Validace struktur získaných z databází a příprava dat pro molekulové modelování – program WHAT IF a jeho WWW server
- Vizualizace molekul pomocí programu VMD, vizualizace trajektorií molekulové dynamiky
- Zpracování dat a jejich vizualizace pomocí grafů (xmgrace)
- Program TRITON – možnosti vizualizace a aplikace v molekulovém modelování
- Vizualizace molekul v ostatních programech – Chimera, PyMol
- Převod 2D zobrazení molekul do 3D a naopak
- Příprava grafických prezentací (Open Office, PovRay, Inkscape)

Databáze molekul

- Metody pro získání 3D struktury molekul lze rozdělit na:
  - **Experimentální metody** (rengenová krystalografie, NMR, elektronová mikroskopie)
  - **Molekulové modelování**
- Struktury získané experimentálními metodami (někdy i molekulovým modelováním) se ukládájí do databází aby byly přístupné i ostatním uživatelům

Databáze malých molekul:

- Cambridge Structural Database - CSD (struktury z rentgenové krystalografie)
- Inorganic Crystal Structure Database - ICSD
- Glyco3D Database (zaměřená na cukry)
- 3Dchem database

Databáze biopolymerů (proteiny, DNA, RNA):

- Protein Data Bank – PDB
- Nucleic Acid Database – NDB

Jak hledat v databázi

- Znám název molekuly, nebo její sumární vzorec
- U malých molekul znám jejich SMILES kód (CCO – ethanol, c1CCCCc1 – cyklohexan a pod)
- CSD umí hledat podle funkčních skupin (umožňuje přes grafické rozhraní nakreslit vzorec)
- NDB databáze – hledání podle sekvence bazí
- PDB umožňuje několikastupňové prohledávání

Cambridge Structural Database - CSD

Dostupnost do CSD:

- Komerční licence
- Na MU dostupná na serveru xray.chemi.muni.cz (nutné zřídit účet – doc. Marek Nečas)

Co ukládá CSD:

- Data o organických molekulách
- Data o komplexech kovů s organickými molekulami

Jakými metodami jsou získávána data:

- RTG krystalografie
- Monokrystalů
- Prášková difrakce

Co v CSD nenajdeme:

- Data o polypeptidech, polysacharidech a oligonukleotidech
- Data o anorganických molekulách

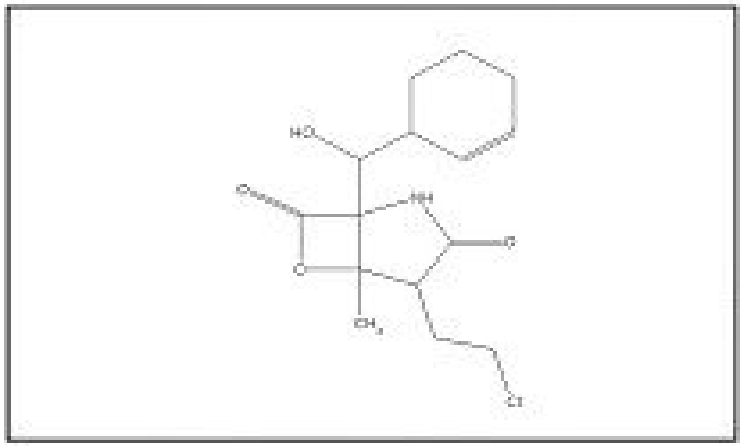
Cambridge Structural Database - CSD

1D Bibliographic Information

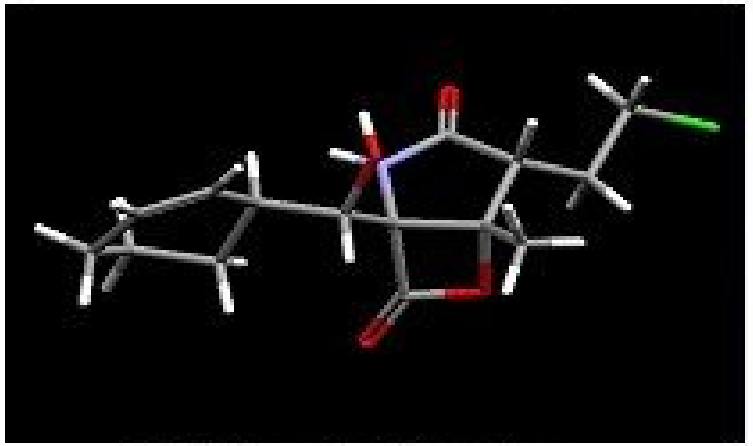
TUSJIA
4-(2-chloroethyl)-1-[cyclohex-2-en-1-yl(hydroxy)methyl]-5-methyl-6-oxa-2-azabicyclo[3.2.0]heptane-3,7-dione
Synonym: Salinosporamide A
Formula: C15 H20 Cl1 N1 O4
Source: Salinospora strain CNB-392 isolated from heat treated marine sediment
Extra info: absolute configuration; potent cancer cell cytotoxicity through inhibition of the 20S proteasome
R.H.Feling, G.O.Buchanan, T.J.Mincer, C.A.Kauffman, P.R.Jensen, W.Fenical

Angew.Chem., Int.Ed.Engl., **42**, 355, 2003.

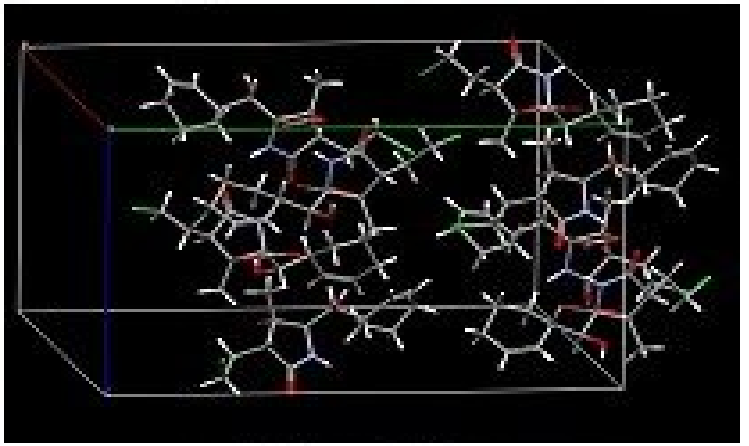
2D Chemical Connectivity



CSD



3D Molecular Structure



3D Crystal Structure

Cambridge Structural Database - CSD

CSD Entries: Summary Statistics (Historical Snapshots)

	31 Dec 1983		*31 Dec 1990*		*30 Oct 2001*		*31 Dec 2009*	
	Structures	*%*	*Structures*	*%*	*Structures*	*%*	*Structures*	*%*
Total structures	52 363	100.0	104 380	100.0	251 515	100.0	501 857	100.0
Organic structures	28 995	55.4	52 450	50.3	112 113	44.6	215 106	42.9
Transition metal present	20 439	38.9	45 588	43.7	120 638	48.0	266 333	53.1
Li – Fr or Be – Ra present	2 887	5.5	5 299	5.1	13 471	5.4	25 739	5.1
Main group metal present	2 206	4.2	5 024	4.8	16 171	6.4	31 470	6.3
3D coordinates present	37 318	71.0	83 884	80.4	223 920	89.0	462 146	92.0
Error-free coordinates	35 032	93.9†	80 372	95.8†	219 864	98.2†	453 422	98.1†
Neutron studies	567	1.08	786	0.8	1 062	0.4	1 437	0.3
Low/high temp. studies	3 275	6.2	9 943	9.5	55 752	22.2	192 352	38.3
Abs. config. determined	1 330	2.5	2 344	2.2	4 924	2.0	7 672	1.5
Disorder present	4 943	9.4	13 594	13.0	45 728	18.2	107 032	21.3
Polymorphic structures	3 231	6.1	4 618	4.4	7 892	3.1	16 253	3.2
R-factor < 0.100	37 190	70.8	85 389	81.8	227 181	90.3	467 547	93.2
R-factor < 0.075	29 937	57.0	73 424	70.3	202 848	80.7	422 654	84.2
R-factor < 0.050	15 974	30.4	42 996	41.2	125 112	49.7	268 930	53.6
R-factor < 0.030	2 231	4.2	7 150	6.8	22 346	8.9	52 693	10.5
n(atoms)/structure *	44	-	54	-	73	-	76	-
Mb data added in year+	14	-	29	-	86	-	171	-

† Percentage of structures for which coordinates are present in the CSD

* Average number of atoms per structure in the year cited in the column heading. The figure for 1970 was 27.

+ Number of Mb of data added to the CSD in the year cited in the table heading, the value in the 31 Dec 2009 column is for 2008 (the last complete year). The figure for 1970 was 2 Mb.

Cambridge Structural Database - CSD

Cambridge Structural Database
6 January 2014

CSD Entries: Summary Statistics

	Structures	%CSD
Total No. of structures	686 944	100.0
No. of different compounds	628 684	-
No. of literature sources	1 578	-
Organic structures	292 661	42.6
Transition metal present	369 682	53.8
Li – Fr or Be – Ra present	34 433	5.0
Main group metal present	41 711	6.1
3D coordinates present	643 032	93.3
Error-free coordinates	630 329	98.0†
Neutron studies	1 616	0.2
Powder diffraction studies	2 930	0.4
Low/high temp. studies	306809	44.7
Absolute configuration determined	14 752	2.1
Disorder present in structure	158 127	23.0
Polymorphic structures	20 753	3.0
R-factor < 0.100	645 809	94.0
R-factor < 0.075	585 333	85.2
R-factor < 0.050	378 391	55.1
R-factor < 0.030	78 594	11.4
No. of atoms with 3D coordinates	53 563 990	-

† Taken as a percentage of structures for which 3D coordinates are present in the CSD

Inorganic Crystal Structure Database - ICSD

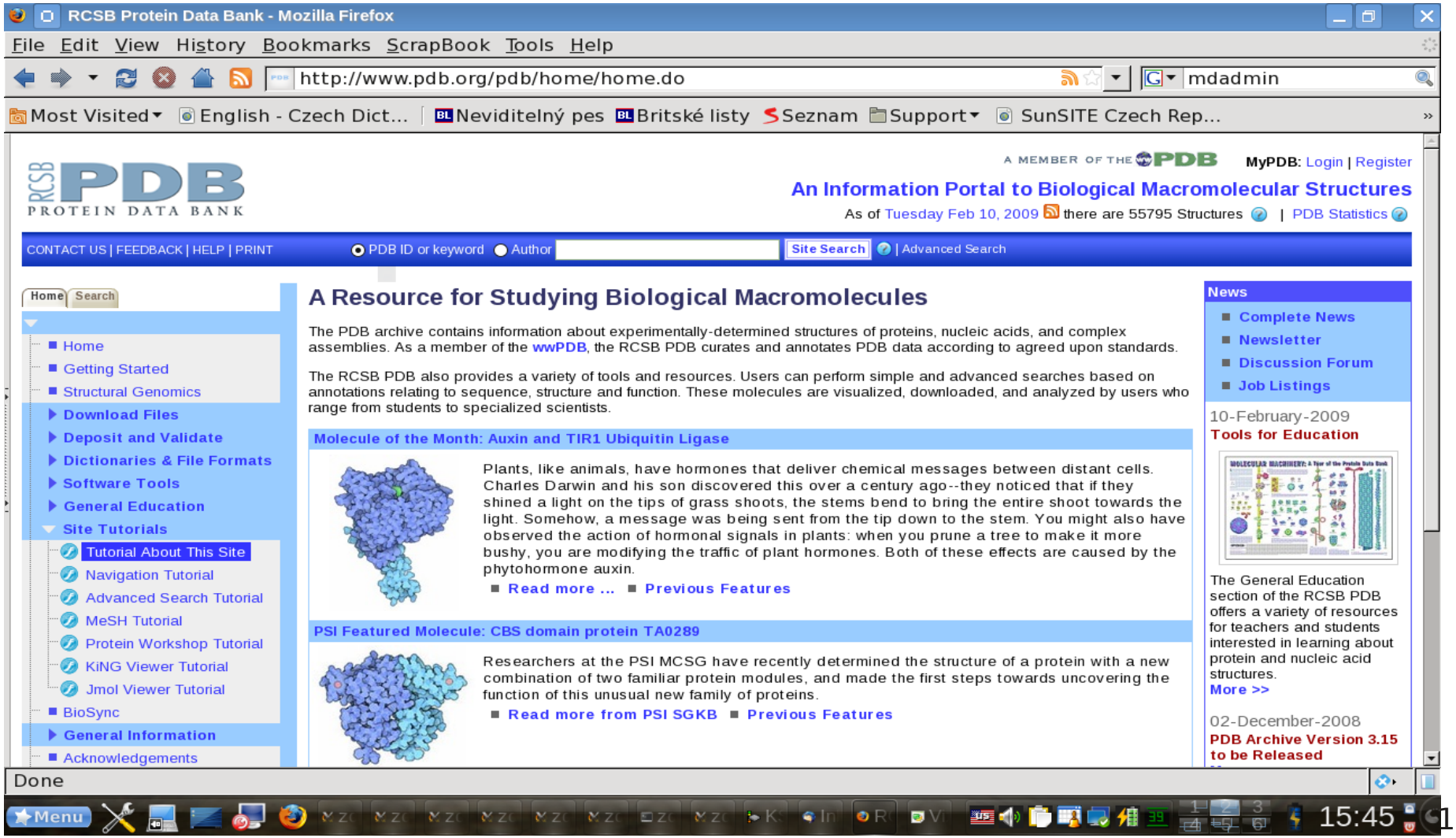
- Dostupná na adrese: www.fiz-karlsruhe.de/icsd.html
- ICSD obsahuje více než 166000 záznamů

Co najdeme v ICSD:

- Krystalové struktury prvků
- Krystalové struktury binárních sloučenin
- Krystalové struktury složitějších anorganických sloučenin

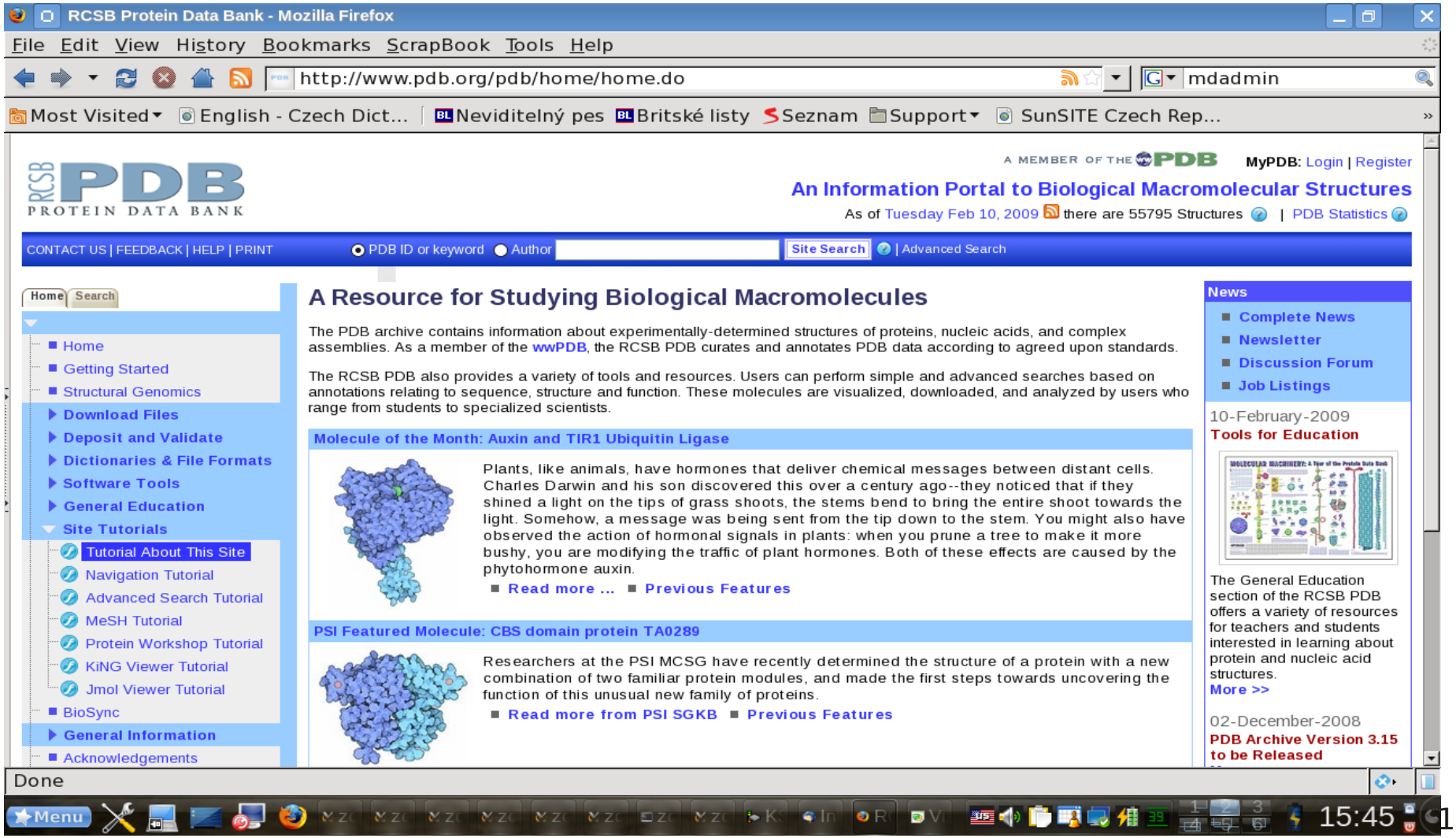
Protein Data Bank - PDB

- Dostupná na adrese: www.pdb.org



Protein Data Bank - PDB

- Dostupná na adrese: www.pdb.org



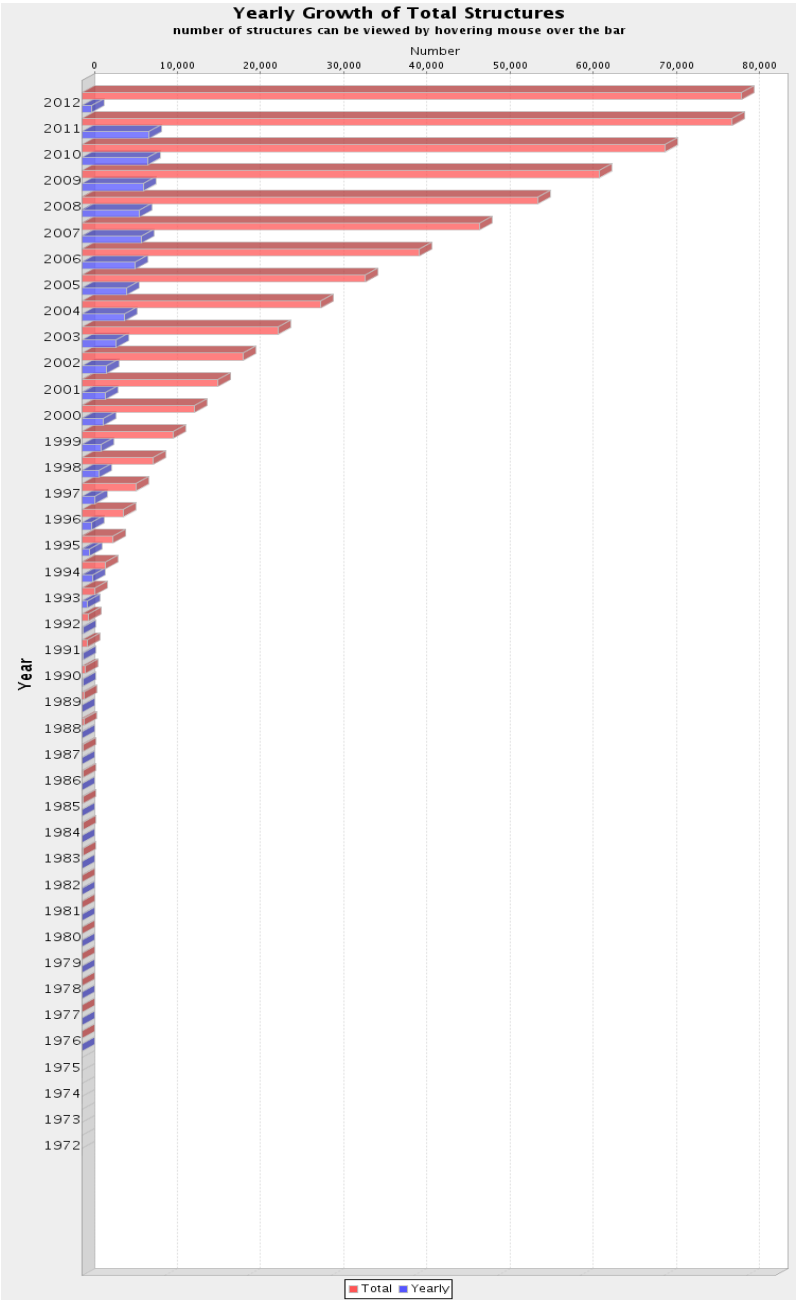
Protein Data Bank - PDB

Počet struktur: 88500

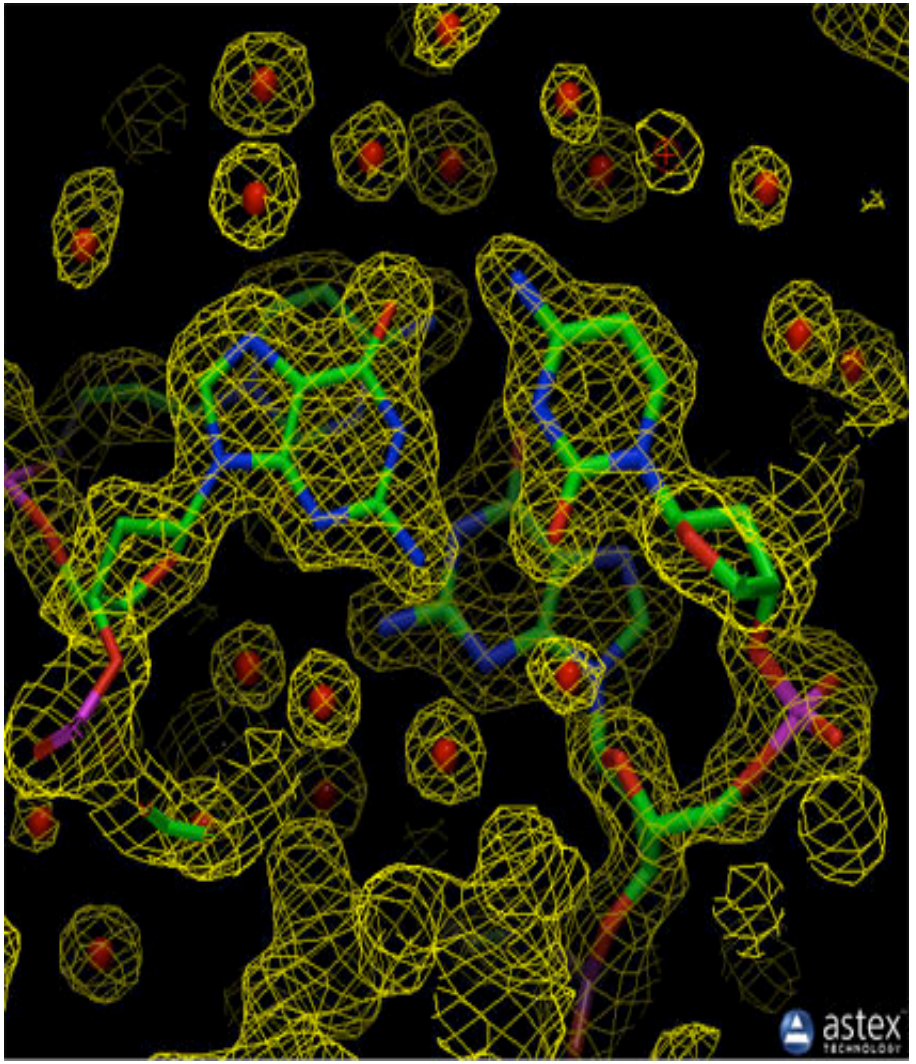
- 77975 – RTG
- 9862 – NMR
- 508 – EM

Proteiny/NA/komplexy

- 81922 – protein
- 2500 – DNA/RNA
- 4057 – komplexů

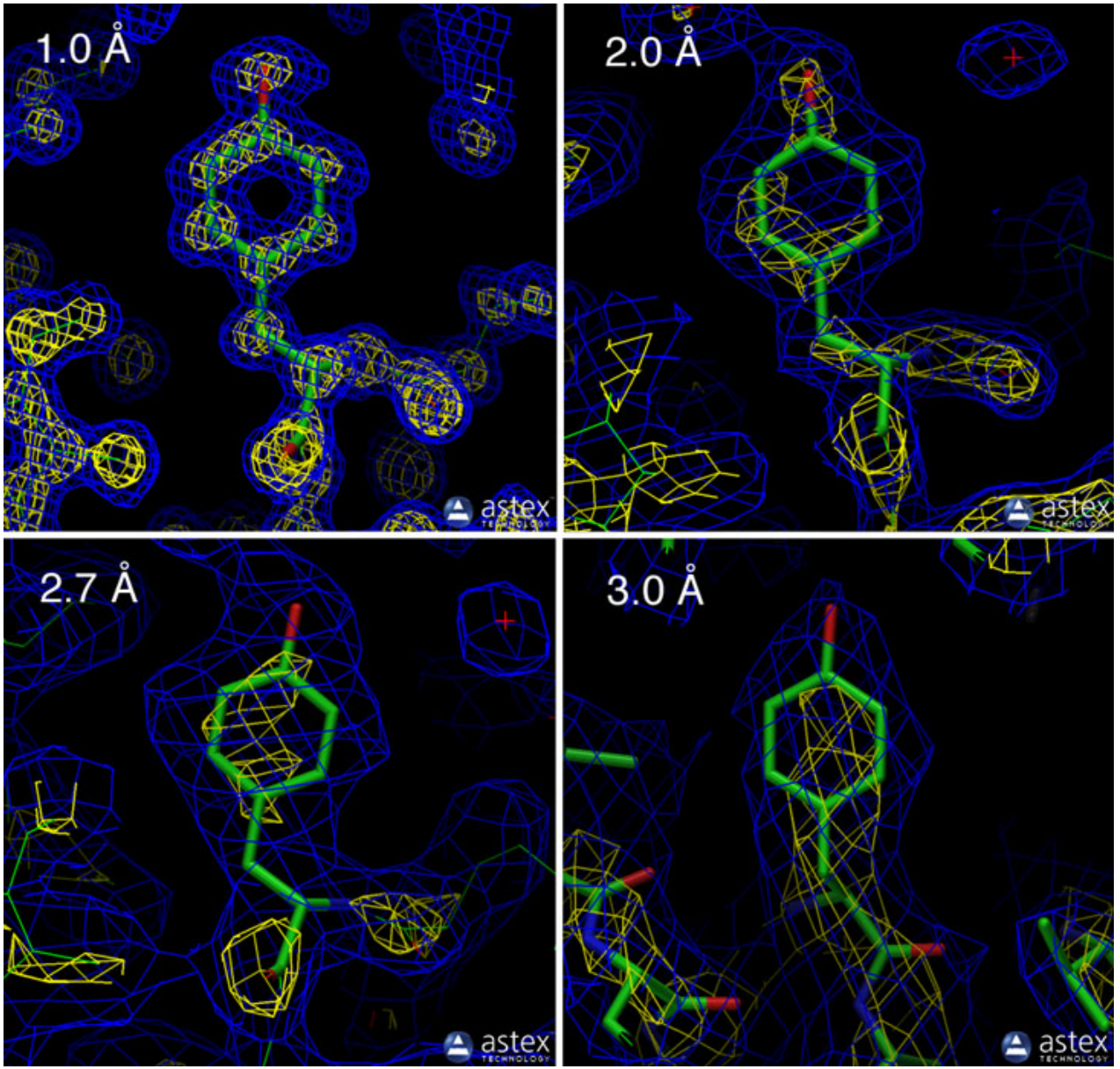


Rentgenová krystalografie

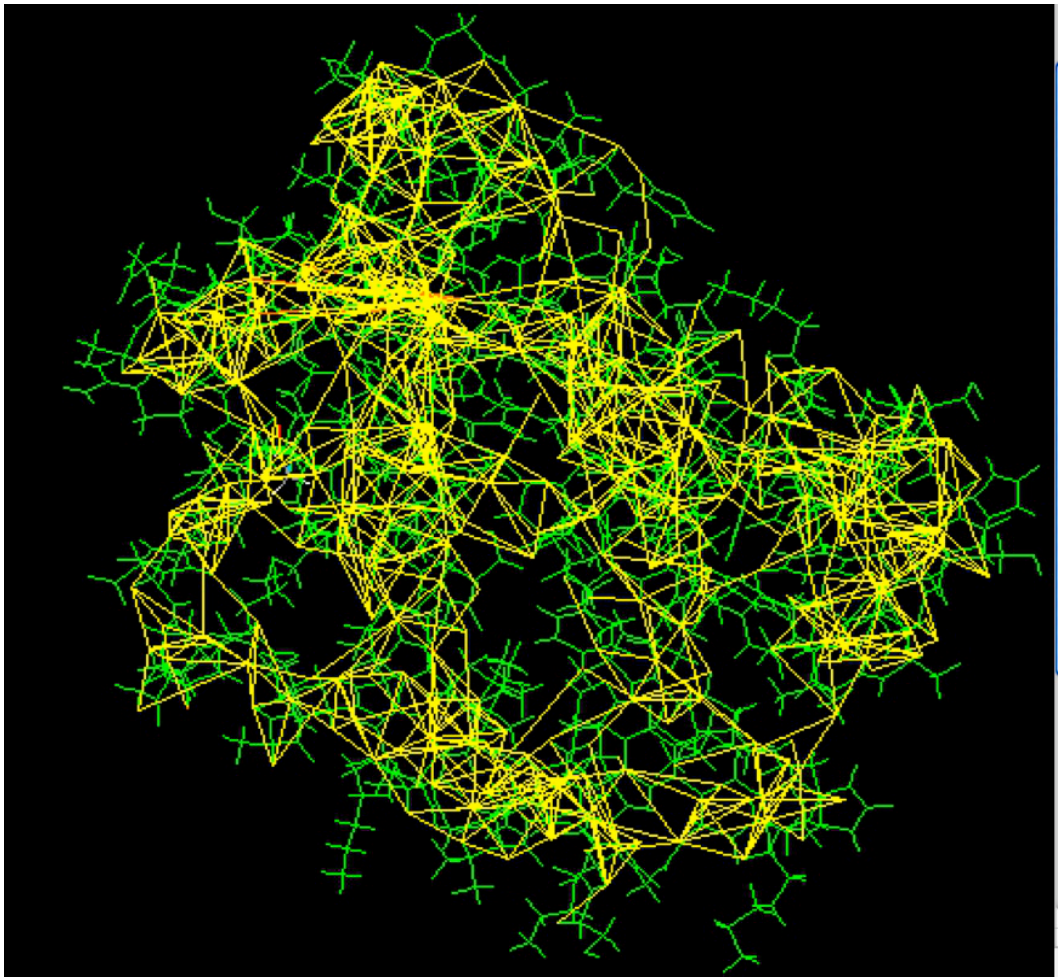


- Velmi detailní informace o struktuře molekul
- Pro rigidní systémy nejkvalitnější informace
- U flexibilních systémů někdy špatně rozeznatelné části
- O kvalitě struktury hovoří rozlišení a R-faktor.

Rentgenová krystalografie - rozlišení

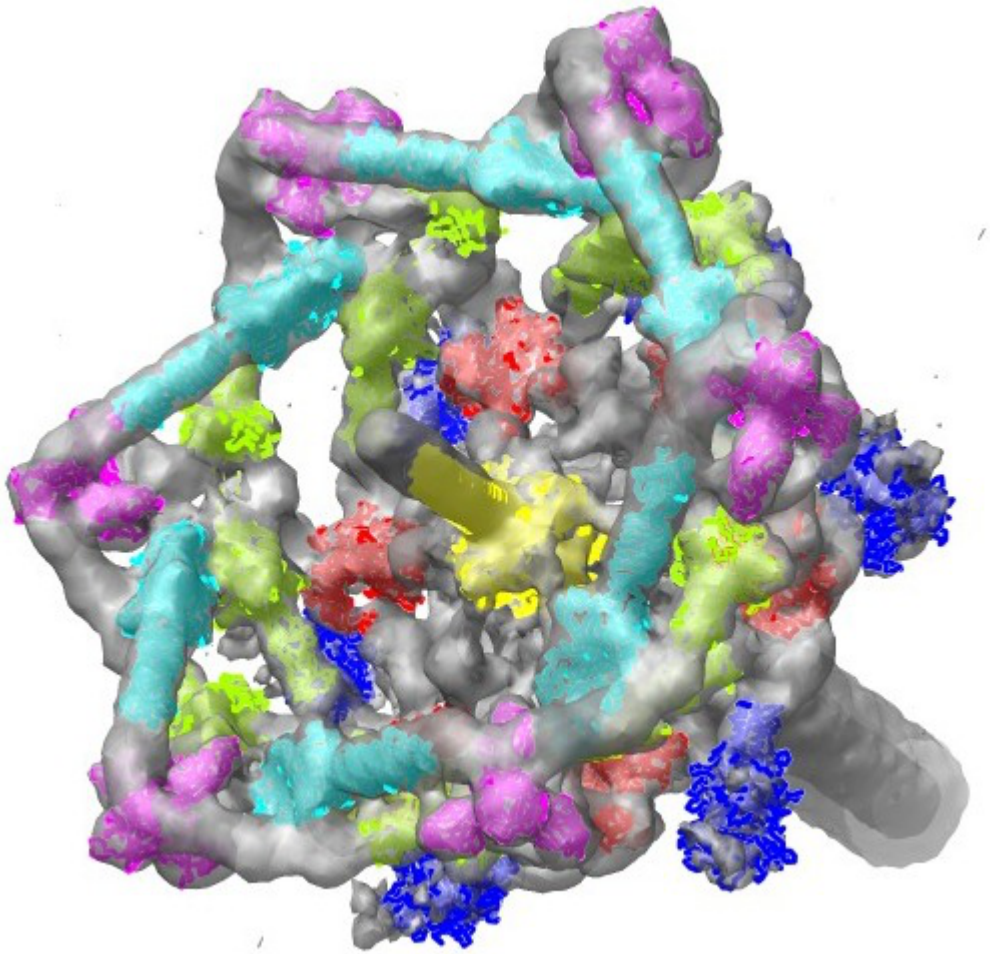


Nukleární magnetická rezonance - NMR



- Struktura v roztoku, ne v pevné fázi.
- Možnost studia flexibilních systémů.
- Výsledkem je set konformací molekuly, který splňuje experimentální kriteria.
- V databázi naleznete jak set 10, 30 , 50 struktur, tak tzv. průměrnou strukturu.

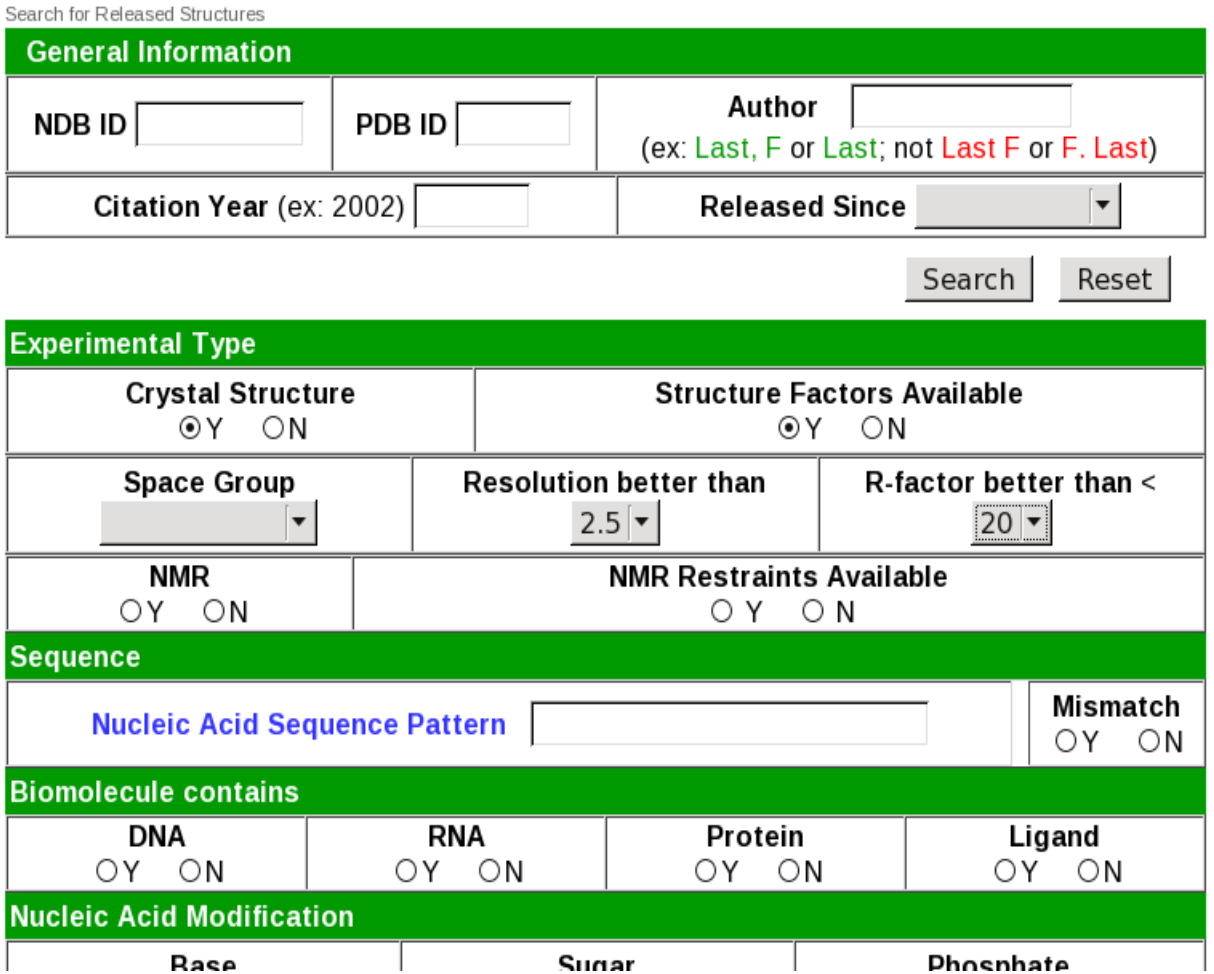
Elektronová mikroskopie - EM



- Struktura velkých komplexů
- Samostatně neumožňuje atomární rozlišení, proto se kombinuje s dalšími metodami

NDB - prohledávání

- Dostupná na adrese: ndbserver.rutgers.edu/



NDB - prohledávání

NUCLEIC ACID DATABASE

Results **1** - **20** of **374** records in **NDB** Search took **0.595** seconds.

There are **19** pages in total. Help Close

1. Display Dynamically Generated Atlas of Checked Structures Display Atlas

2. View NDB Status report of View Report

#	ID	DESCRIPTION
☐ All		
☐ 1	AD0007	A-DNA DECAMER GCGTA(T23)TACGC WITH INCORPORATED 2'-METHOXY-3'-METHYLENEPHOSPHATE-THYMIDINE
☐ 2	AD0010	ALTERNATION OF DNA AND SOLVENT LAYERS IN THE A FORM OF D(GGCGCC) OBTAINED BY ETHANOL CRYSTALLIZATION
☐ 3	AD0011	1.6 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-AMINOOXYETHYL THYMINE IN PLACE OF T6, BA-FORM
☐ 4	AD0012	1.45 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-FLUOROETHYL THYMINE IN PLACE OF T6, MEDIUM NA-SALT
☐ 5	AD0013	1.06 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-METHYL-3'-METHYLENEPHOSPHONATE (T23) THYMINE IN PLACE OF T6, HIGH CS-SALT
☐ 6	AD0014	1.05 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-METHYL-[TRI(OXYETHYL)] THYMINE IN PLACE OF T6, MEDIUM CS-SALT
☐ 7	AD0015	1.05 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-FLUOROETHYL THYMINE IN PLACE OF T6, HIGH RB-SALT
☐ 8	AD0016	1.3 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-METHYL-[TRI(OXYETHYL)] THYMINE IN PLACE OF T6, MEDIUM RB-SALT
☐ 9	AD0018	1.3 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-METHYL-[TRI(OXYETHYL)], MEDIUM K-SALT

NDB - prohledávání

NUCLEIC ACID DATABASE

Results **1** - **20** of **374** records in **NDB** Search took **0.595** seconds.

There are **19** pages in total. Help Close

1. Display Dynamically Generated Atlas of Checked Structures Display Atlas

2. View NDB Status report of View Report

#	ID	DESCRIPTION
☐ All		
☐ 1	AD0007	A-DNA DECAMER GCGTA(T23)TACGC WITH INCORPORATED 2'-METHOXY-3'-METHYLENEPHOSPHATE-THYMIDINE
☐ 2	AD0010	ALTERNATION OF DNA AND SOLVENT LAYERS IN THE A FORM OF D(GGCGCC) OBTAINED BY ETHANOL CRYSTALLIZATION
☐ 3	AD0011	1.6 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-AMINOOXYETHYL THYMINE IN PLACE OF T6, BA-FORM
☐ 4	AD0012	1.45 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-FLUOROETHYL THYMINE IN PLACE OF T6, MEDIUM NA-SALT
☐ 5	AD0013	1.06 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-METHYL-3'-METHYLENEPHOSPHONATE (T23) THYMINE IN PLACE OF T6, HIGH CS-SALT
☐ 6	AD0014	1.05 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-METHYL-[TRI(OXYETHYL)] THYMINE IN PLACE OF T6, MEDIUM CS-SALT
☐ 7	AD0015	1.05 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-FLUOROETHYL THYMINE IN PLACE OF T6, HIGH RB-SALT
☐ 8	AD0016	1.3 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-METHYL-[TRI(OXYETHYL)] THYMINE IN PLACE OF T6, MEDIUM RB-SALT
☐ 9	AD0018	1.3 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-METHYL-[TRI(OXYETHYL)], MEDIUM K-SALT

NDB - prohledávání

NDB ID: AD0014

Home
NMR Atlas X-Ray Atlas

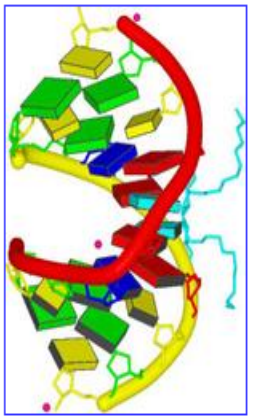
Title: 1.05 A STRUCTURE OF THE A-DECAMER GCGTATACGC WITH A SINGLE 2'-O-METHYL-[TRI(OXYETHYL)] THYMINE IN PLACE OF T6, MEDIUM CS-SALT

Molecular Description: 5'-D(GpCpGpTpAp(126)pApCpGpC)-3'

Structural Features: A DOUBLE HELIX

Nucleic Acid Sequence: ChainsA, B: (DG) (DC) (DG) (DT) (DA) (126) (DA) (DC) (DG) (DC)

Primary Citation: Tereshko, V., Wilds, C.J., Minasov, G., Prakash, T.P., Maier, M.A., Howard, A., Wawrzak, Z., Manoharan, M., Egli, M. Detection of alkali metal ions in DNA crystals using state-of-the-art X-ray diffraction experiments. *Nucleic Acids Res.* , **29**, pp. 1208 - 1215, 2001.

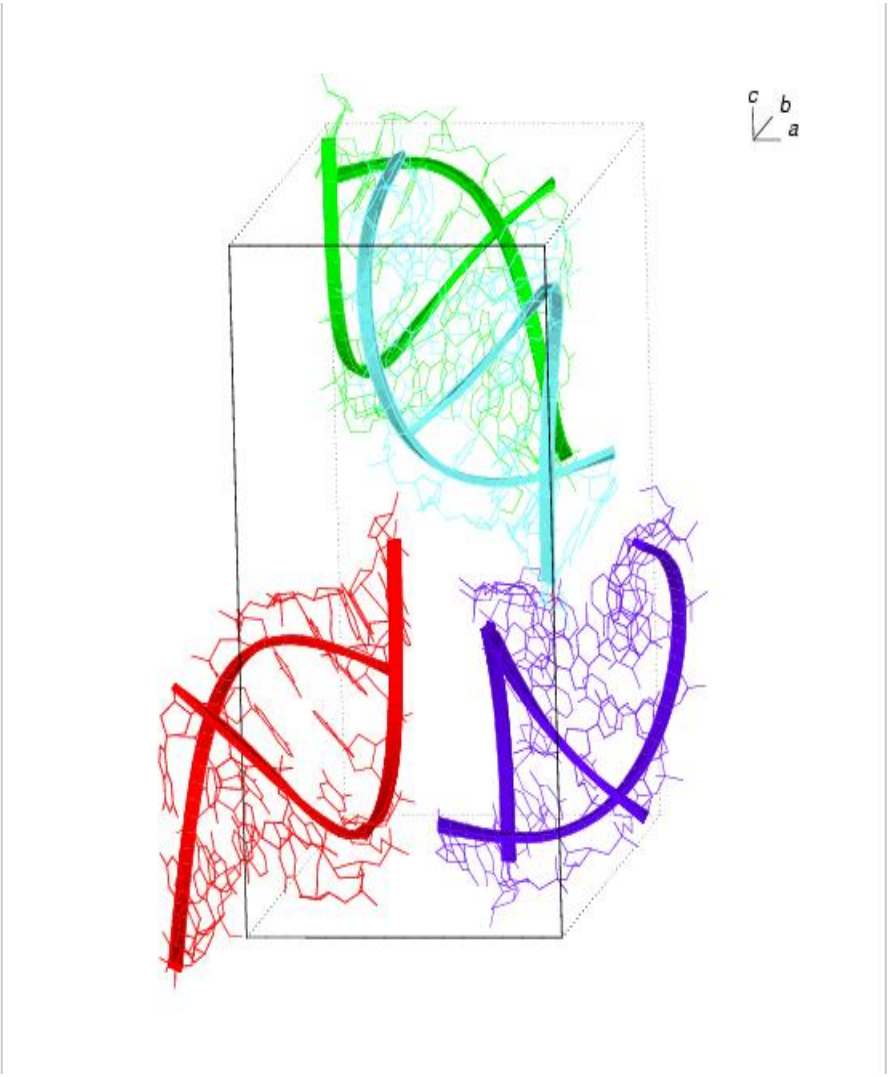


Biological Unit 1

Other Views

Asymmetric Unit
Crystal Packing
Enlarge Biological Unit 1

NDB - prohledávání



Glyco3D databáze

- Dostupná na adrese: glyco3d.cermav.cnrs.fr

Cermav

Home
- Home page

Databases
- Monosaccharides
- Disaccharides
- Oligosaccharides
- Polysaccharides
- Lectines
- Glycosyltransferases
- Gag binding proteins

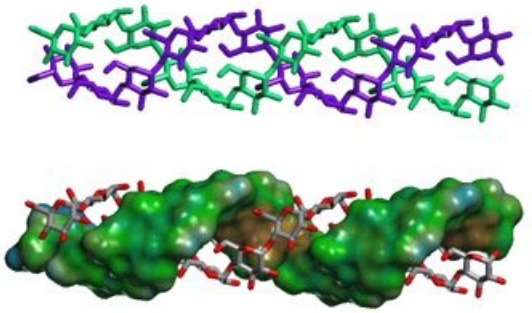
Book chapters
- Modeling
- Cellulose
- Starch

Gallery
- Models

Tools
- Useful links

GLYCO3D

A site for glycosciences



BP 53 - 38041 Grenoble Cedex 9 - France

Domaine Universitaire - 601 Rue de la Chimie
38400 St Martin d'Hères - France

tél : 33 (0)4 76 03 76 03 - fax : 33 (0)4 76 54 72 03

Glyco3D databáze

Mozilla Firefox <2>

http://www.cermav.cnrs.fr/cgi-bin/monos/monos.cgi

Google

MONOSACCHARIDES

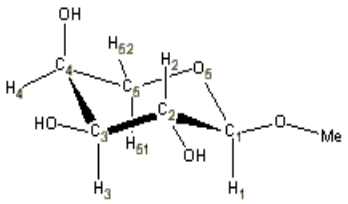
- ABEQUOSE
- ALTROSE
- APIOSE
- ARABINOSE
  - Alpha
    - a-L-Arabinopyranose
  - Beta
- FRUCTOSE
- FUCOSE
- GALACTOSE
- GLUCOSE
- GULOSE
- IDOSE
- LYXOSE
- MANNOSE
- MISC
- RHAMNOSE
- RIBOSE
- SIALIC
- TALOSE
- XYLOSE

a-L-Arabinopyranose

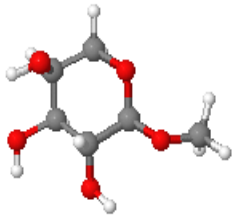
Family : Arabinose
Anomery : Alpha

Carbons :	5
Cycle :	Pyranose
Configuration :	Levogyre
Comment :	

2D view



3D view



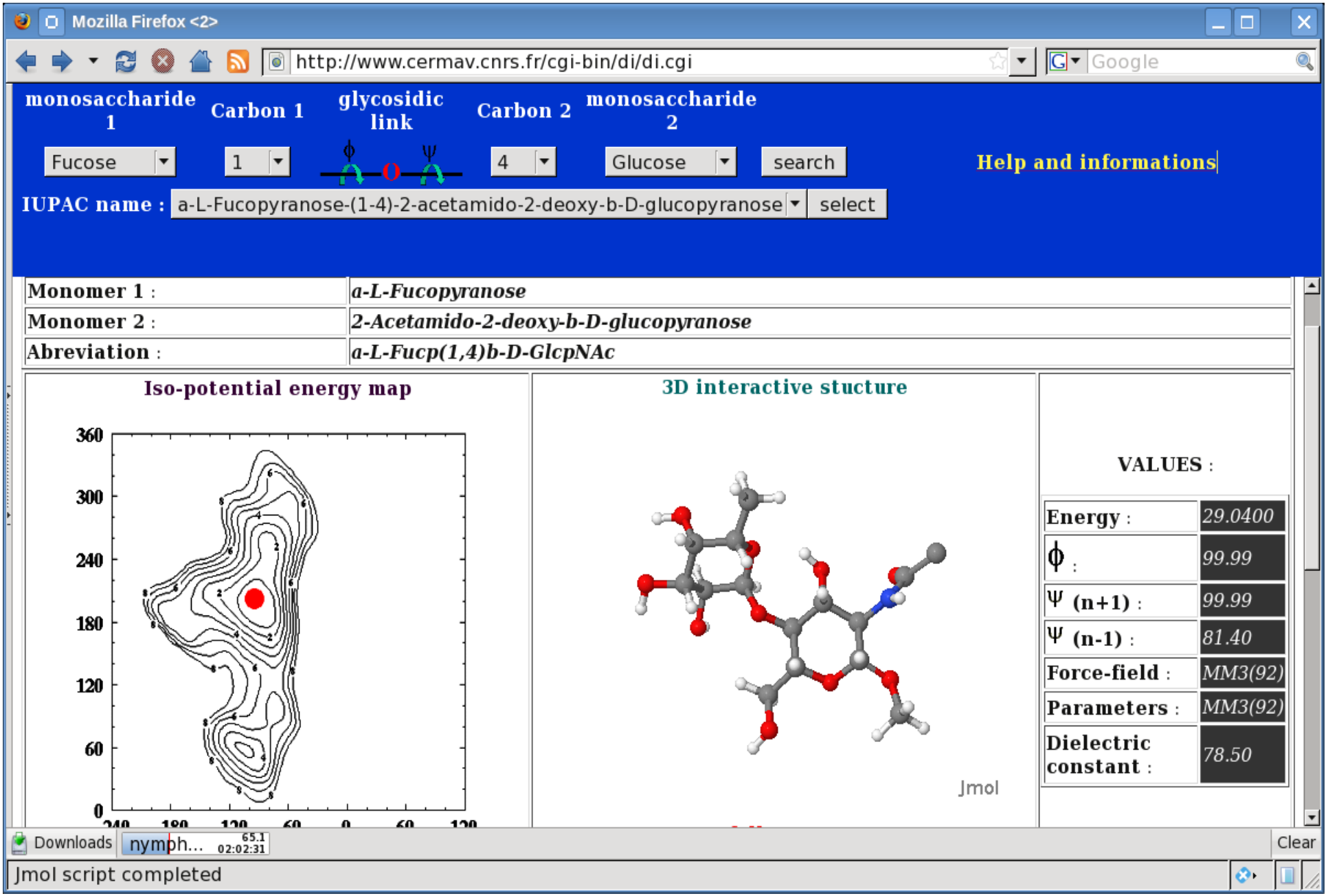
Jmol

PDB file

Downloads | nymph... 65.9 02:05:08 | Clear

http://www.cermav.cnrs.fr/cgi-bin/monos/monos.cgi/response?msiupac=a-L-Arabinopyranose

Glyco3D databáze



Glyco3D databáze

Mozilla Firefox <2>

http://www.cermav.cnrs.fr/lectines/

Google

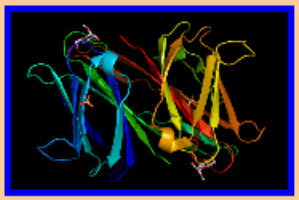
LECTINES

- Algal lectins
- Animal lectins
- Bacterial lectins
  - 1-Ca b-sandwich
  - 2-Ca b-sandwich
    - Burkholderia lectins
    - Chromobacterium CV-IIL
    - Pseudomonas PA-IIL
    - Ralstonia RS-IIL
  - ADP-ribosylating toxin (AB5 toxin)
  - b-propeller
  - b-trefoil hemagglutinin
  - Bacterial neurotoxin
  - cyanobacterial lectins
  - Pili adhesin
  - Staphylococcal toxin
  - Toxin repetitive domain

Bacterial lectins - 2-Ca b-sandwich

Burkholderia lectins

PDB code : 2VNV
Species : Burkholderia cenocepacia
Commentaire : Complexed with **aManOMe**
Resolution : 1.70
Reference : Lameignere E, Malinovská L, Sláviková M, Duchaud E, Mitchell EP, Varrot A, Sedo O, Imberty A, Wimmerová M., Structural basis for mannose recognition by a lectin from opportunistic bacteria Burkholderia cenocepacia:, *Biochem. J.*, **411**, 307-318, 2008



Medline

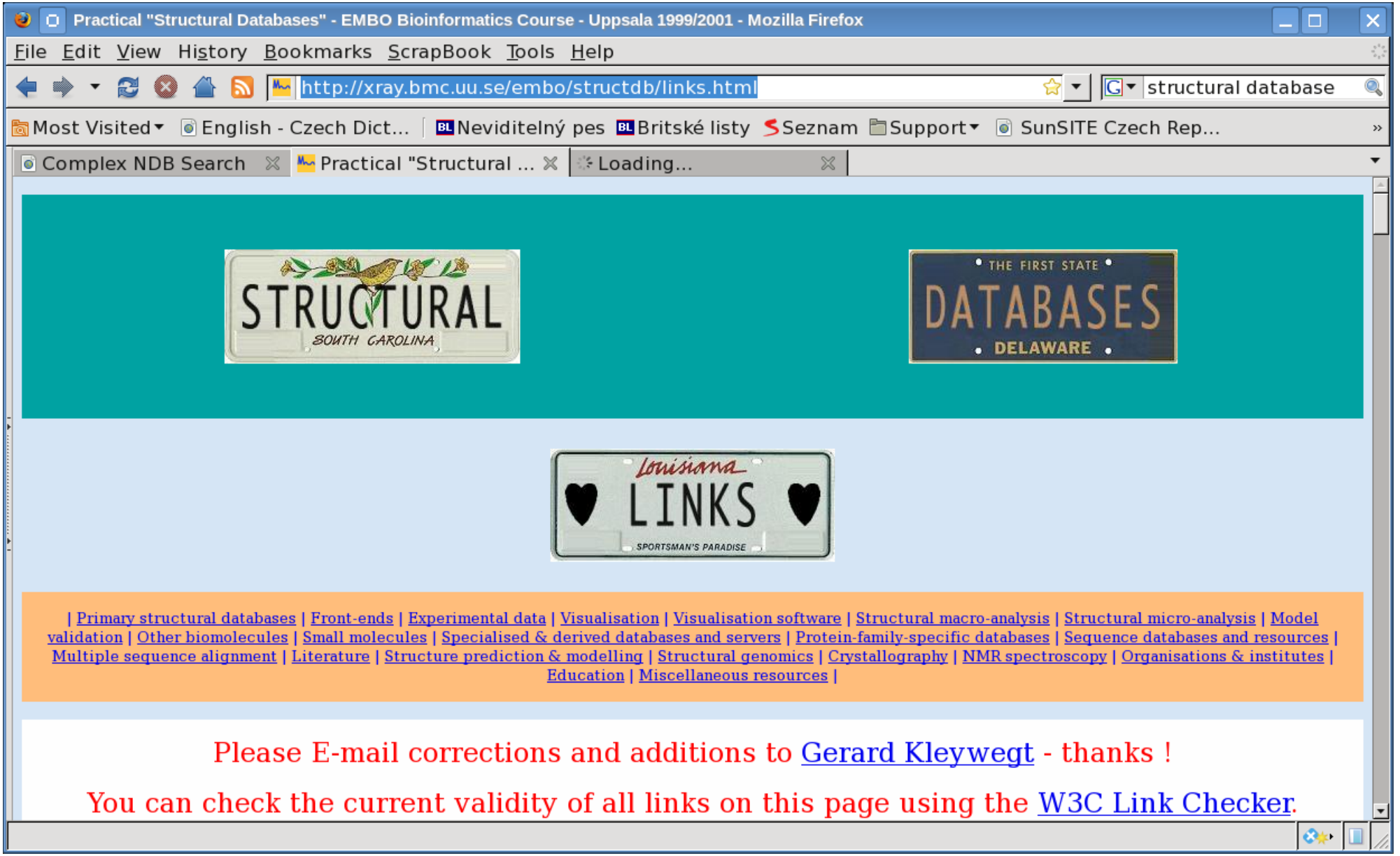
Link to Databases :
[PDB ENTRY (RCSB Site) / HSSP ENTRY / SCOP / RELIBASE]

PDB Files :
(RCSB (USA CA) / Rutgers University (USA NJ) / FRANCE / SINGAPORE / UK / JAPAN / BRAZIL)

Done

Practical Structural Databases

- Webová adresa: xray.bmc.uu.se/embo/structdb/links.html



Úkoly

- Nalezněte v PDB databázi struktury amyloid beta peptidu, vygenerujte report, kerý bude obsahovat: pdb kód, způsob určení 3D dat, rozlišení, biologický zdroj amyloidu.
- Zjistěte kolik struktur je v současné době uloženo v PDB databázi, kolik je určeno pomocí X-ray, NMR a EM. Kolik struktur bylo v databázi uloženo v letech 2010, 2005 a 2001. Pro vyhledání těchto iformaci použijte Advanced Search.
- Úkoly odevzdejte do odevzdávárny