

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

ANALÝZA NÁSTROJŮ PRO ZOBRAZOVÁNÍ STRUKTURY PROTEINŮ

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

JARMILA KLEMŠOVÁ

BRNO 2009



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

ANALÝZA NÁSTROJŮ PRO ZOBRAZOVÁNÍ STRUKTURY PROTEINŮ

ANALYSIS OF PROTEIN STRUCTURE VISUALISATION TOOLS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE

AUTHOR

JARMILA KLEMŠOVÁ

VEDOUcí PRÁCE

SUPERVISOR

Ing. IVANA RUDOLFOVÁ

BRNO 2009

Abstrakt

Tato práce se zabývá analýzou nástrojů pro prohlížení struktury proteinů. Pojednává o struktuře proteinů, databázích struktur proteinů a formátu dat v nich uložených. Dále popisuje vybrané prohlížečí nástroje a skripty pro ně vytvořené. Součástí je i ukázková vizualizace.

Abstract

In this thesis we analyze several tools for protein structure visualisation. We describe the structure of the proteins, the databases for storing protein structures and their data formats. Next part of the thesis, provides detailed information about selected visualisation tools and some scripts for them. The thesis also include an example of visualisation of the selected protein.

Klíčová slova

databáze, pdb, proteiny, prohlížečí nástroje, vizualizace

Keywords

database, pdb, proteins, visualisation tools, visualisation

Citace

Jarmila Klemšová: Analýza nástrojů pro zobrazování struktury proteinů, bakalářská práce, Brno, FIT VUT v Brně, 2009

Analýza nástrojů pro zobrazování struktury proteinů

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně pod vedením paní Ing. Ivany Rudolfové. Uvedla jsem všechny literární prameny a publikace, ze kterých jsem čerpala.

.....
Jarmila Klemšová
17. května 2009

Poděkování

Děkuji své vedoucí paní Ing. Ivaně Rudolfové za odbornou pomoc a podnětné návrhy.

© Jarmila Klemšová, 2009.

Tato práce vznikla jako školní dílo na Vysokém učení technickém v Brně, Fakultě informačních technologií. Práce je chráněna autorským zákonem a její užití bez udělení oprávnění autorem je nezákonné, s výjimkou zákonem definovaných případů.

Obsah

1	Úvod	3
2	Proteiny	4
2.1	Primární struktura	6
2.2	Sekundární struktura	6
2.3	Terciární struktura	7
2.4	Kvarterní struktura	7
3	Data – zdroje a formáty	8
3.1	Zdroje dat	8
3.1.1	RCSB PDB	8
3.1.2	NCBI Entrez	9
3.2	Formáty dat	9
3.2.1	Formát PDB	9
3.2.2	Formát mmCIF	11
4	3D vizualizace	13
4.1	RasMol	15
4.1.1	Výběr	15
4.1.2	Zobrazení	15
4.1.3	Barvy a barevná schémata	15
4.1.4	Skripty	16
4.1.5	Další funkce	16
4.1.6	Shrnutí	16
4.2	Cn3D	17
4.2.1	Výběr	17
4.2.2	Zobrazení	17
4.2.3	Barvy a barevná schémata	17
4.2.4	Skripty	17
4.2.5	Další funkce	18
4.2.6	Shrnutí	18
4.3	Swiss-PDB Viewer	19
4.3.1	Výběr	19
4.3.2	Zobrazení	19
4.3.3	Barvy a barevná schémata	19
4.3.4	Skripty	20
4.3.5	Další funkce	20
4.3.6	Shrnutí	20

4.4	Přehled vlastností nástrojů	21
5	Vizualizace hemoglobinu 2DN2	23
6	Implementace	28
6.1	RasMol	28
6.2	Swiss-PDB Viewer	28
6.2.1	Skript swissDemo1.txt	29
6.2.2	Skript swissDemo2.txt	29
6.2.3	Skript swissDemo3.txt	29
7	Závěr	31
	Literatura	33
	Seznam příloh	34
A	Obsah přiloženého CD	35
B	Obrázky vytvořené skriptem rasDemo.spt v programu RasMol	36

Kapitola 1

Úvod

Nástroje pro prohlížení prostorové struktury molekul jsou užitečnou pomůckou pro pochopení struktury a některých chemických zákonitostí uvnitř jinak velmi malých makromolekul. V dnešní době existuje mnoho nástrojů. Některé slouží pouze na prohlížení, jiné jsou obohaceny o další funkce jako například porovnávání dvou struktur. V rámci práce jsem se seznámila s několika nástroji a provedla jejich analýzu.

Kapitola 2 obsahuje teoretický úvod zaměřený na proteiny. Na začátku se zabývám jejich složením a některými nejvýznamnějšími chemickými vazbami, které mají vliv na celkové uspořádání proteinů. Také se zde blíže věnuji jednotlivým úrovním struktury proteinů.

Dále, v kapitole 3, popisuji vybrané databáze biologických dat. Součástí je i rozbor dvou základních formátů souborů pro popis trojrozměrné struktury.

V kapitole 4, 3D vizualizace, se věnuji vybraným nástrojům a popisu základních i některých rozšířených funkcí jednotlivých nástrojů. V závěru této kapitoly je pak uvedena tabulka s přehledem vybraných vlastností nástrojů.

Další dvě kapitoly jsou věnovány praktickým úlohám. Kapitola 5 ukazuje v několika obrázcích protein hemoglobin. Je zde vidět jeho povrch, vnitřní struktura, detail vazebného místa i struktura, která vznikne mutací jedné aminokyseliny. Kapitola 6 popisuje vytvořené skripty pro vybrané nástroje.

V závěru práce pak hodnotím dosažené výsledky.

Kapitola 2

Proteiny

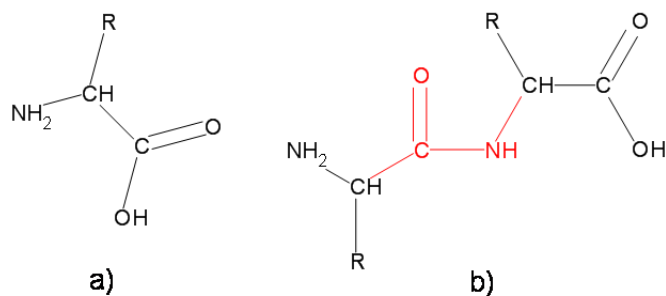
Proteiny jsou základem všech živých organismů [23]. Mohou mít funkci stavební (kolagen, keratin) či transportní (hemoglobin). Jsou součástí imunitního systému (imunoglobulin). Jiné (enzymy) zase působí jako katalyzátory. Po chemické stránce se jedná o vysokomolekulární látky složené z dvaceti různých aminokyselin [5]. Každá aminokyselina má svůj jednoznačný a tříznakový kód (viz tabulka 2.1).

aminokyselina	jednoznačný kód	tříznakový kód
Alanin	A	Ala
Arginin	R	Arg
Asparagin	N	Asn
Kyselina asparagová	D	Asp
Cystein	C	Cys
Glutamin	Q	Gln
Kyselina glutamová	E	Glu
Glycin	G	Gly
Histidin	H	His
Isoleucin	I	Ile
Leucin	L	Leu
Lysin	K	Lys
Methionin	M	Met
Fenylalanin	F	Phe
Prolin	P	Pro
Serin	S	Ser
Threonin	T	Thr
Tryptofan	W	Trp
Thyrosin	Y	Tyr
Valin	V	Val

Tabulka 2.1: Přehled aminokyselin

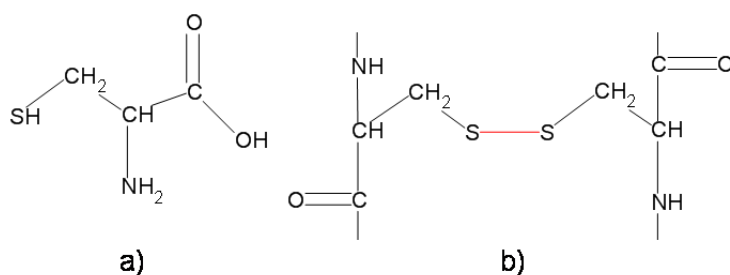
Kostru aminokyselin tvoří aminoskupina $-NH_2$, karboxylová skupina $-COOH$ a α uhlík C, na který je navázán postranní řetězec obvykle označovaný R [23]. Peptidovou vazbou se pak aminokyseliny mezi sebou váží do delších řetězců, kterým říkáme polypeptidy či proteiny. Peptidová vazba $-CO-NH-$ (obrázek 2.1) je kovalentní vazba (sdílení elektronů)

mezi karboxylem jedné aminokyseliny a aminoskupinou aminokyseliny druhé.

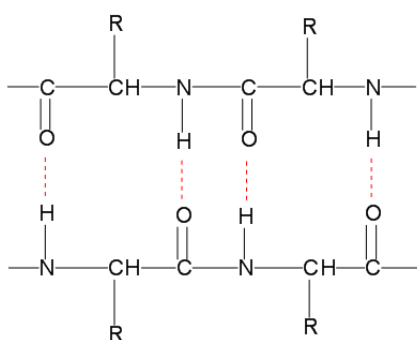


Obrázek 2.1: a) Základní struktura aminokyselin b) Peptidová vazba

V proteinech se také vyskytuje několik dalších vazeb, které se podílejí na stabilitě prostorového uspořádání [4]. Jsou to zejména disulfidová vazba 2.2 (kovalentní propojení cysteinových zbytků) a vazba vodíková 2.3 (nekovalentní vazba).



Obrázek 2.2: a) Cystein b) Disulfidová vazba



Obrázek 2.3: Vodíková vazba

Dále pak vazba iontová mezi karboxylem některých aminokyselin (Asp, Glu) a aminoskupinou jiné aminokyseliny (Lys). Případně van der Waalsovy disperzní síly u aminokyselin s delším postranním řetězcem jako Ile, Leu nebo Phe.

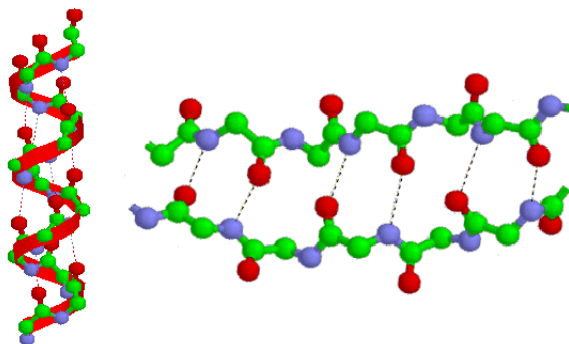
Struktura a funkce spolu velice úzce souvisí. Proteiny zaujímají v buňčném prostředí tvar kulovitý (enzymy, svalová tkáň) nebo vláknitý (kolagen, keratin). Rozrušením struktury (var, změna pH) dochází u většiny kulovitých proteinů ke ztrátě některých biologických funkcí. V souvislosti se strukturou a funkcí se můžeme setkat s pojmy doména a funkční doména. Proteinovou doménou rozumíme část proteinu s určitým složením a strukturou, která může existovat samostatně bez zbytku molekuly. Je-li taková doména nositelem funkce, říkáme jí funkční doména [10].

2.1 Primární struktura

Primární strukturou se rozumí pořadí aminokyselin v řetězci. Díky schopnostem aminokyselin vytvářet různé vazby (disulfidická, vodíková) má zásadní vliv na prostorovou strukturu. Mimo to určuje i fyzikální a chemické vlastnosti proteinu. Primární struktura je zapisována jako posloupnost jednoznakových nebo tříznakových kódů. Posloupnost se obvykle uvádí od N-konce (aminokyselina s volnou aminoskupinou) k C-konci (aminokyselina s volným karboxylem). Textový popis primární struktury jednoznakovými kódy se nazývá formát FASTA.

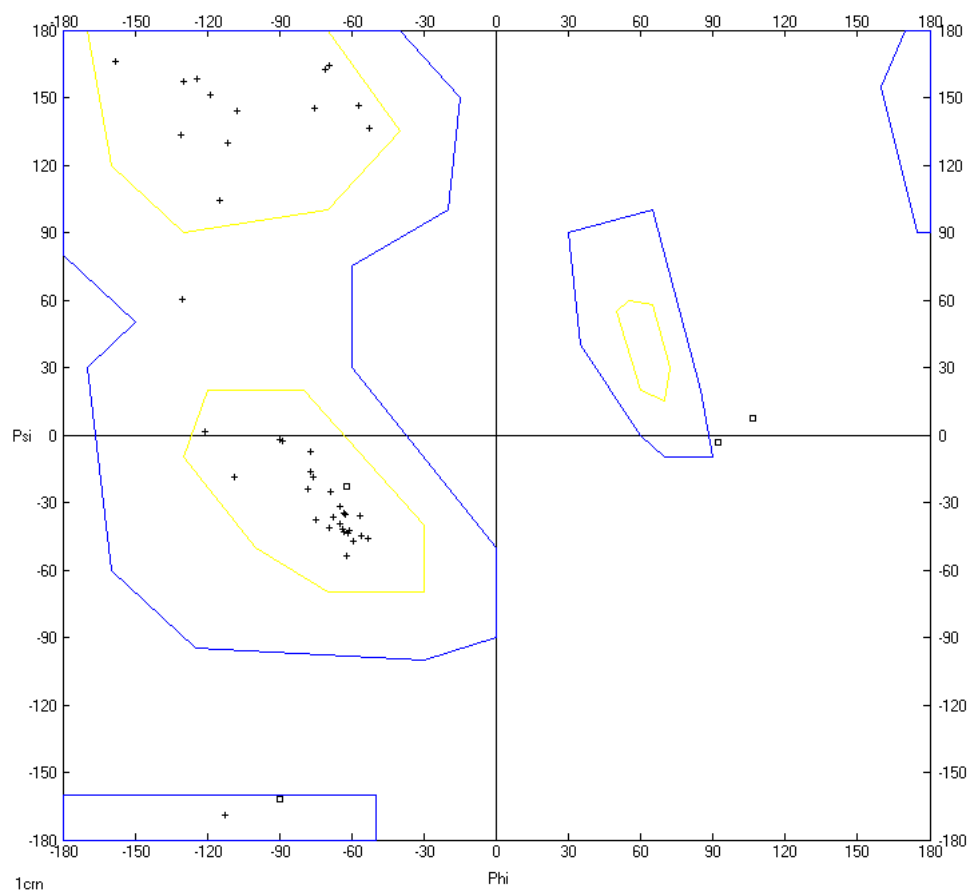
2.2 Sekundární struktura

Sekundární struktura popisuje prostorové uspořádání. Každý protein má tzv. páteř, kterou tvoří všechny atomy aminokyselin kromě postranních řetězců. Zatím co peptidová skupina má pevný tvar, u vazeb mezi α uhlíkem a sousedními atomy páteře může docházet k rotacím. Úhel rotace mezi α uhlíkem a dusíkem aminoskupiny nazýváme φ , úhel rotace mezi α uhlíkem a uhlíkem karboxylu ψ . Tyto rotace společně s vodíkovými vazbami tvoří několik lokálních prostorových struktur. Tato běžná uspořádání se vyskytují téměř ve všech proteinech. Mezi dvě nejběžnější struktury patří α -helix ($\varphi = \psi = -60^\circ$) a β -sheet ($\varphi = -135^\circ$, $\psi = 135^\circ$).



Obrázek 2.4: Schéma α -helix a β -sheet [13].

Sekundární strukturu můžeme zakreslit do Ramachandran mapy (obrázek 2.5). Jedná se o 2D zobrazení v kartézské soustavě, kde φ odpovídá ose x a ψ ose y.



Obrázek 2.5: Ramachandran mapa proteinu 1CRN [9] vytvořená programem Swiss-PDB Viewer.

2.3 Terciární struktura

Shlukováním sekundárních struktur v kombinaci s méně strukturovanými oblastmi páteře vzniká celkový prostorový tvar proteinu, terciární struktura.

2.4 Kvarterní struktura

Kvarterní strukturou se pak rozumí struktura vytvořená interakcemi mezi proteiny. Častým případem, kdy dva a více proteinů spolu tvoří jeden komplex, jsou aktivní enzymy.

Kapitola 3

Data – zdroje a formáty

V této kapitole se blíže věnuji databázím struktur proteinů, především databázi PDB, a webovým rozhraním, přes která k těmto databázím lze přistupovat. Dále pak popisují dva formáty dat pro popis struktury, a to standardní pdb a novější variantu mmCIF.

3.1 Zdroje dat

Velký rozvoj v oblasti molekulární biologie a genomického výzkumu vedl ke vzniku rozsáhlých databází biologických dat. Podle obsažených informací jsou tyto databáze rozděleny do několika skupin – sekvence proteinů, struktury proteinů, sekvence nukleotidů, genomové databáze a databáze s informacemi o expresi genů [23].

Asi největší databází pro proteinové struktury je Protein Data Bank (dále jen PDB). Tuto databázi zastřešuje organizace Worldwide PDB (wwPDB) [19], jejímž cílem je sdružovat data, která budou volně přístupná široké veřejnosti. wwPDB se skládá z několika členů. Jsou to RCSB (Research Colaboratory for Structural Bioinformatics) PDB a BMRB (Biological Magnetic Resonance Data Bank) s centrem v USA, evropská PDBe (dříve MSD) a japonská PDBj.

Základní přístup k datům velkých databází, nejen proteinovým strukturám, poskytuje Evropský ústav bioinformatiky (dále EBI) na adrese <http://www.ebi.ac.uk/>. Struktury jsou zde zastoupeny několika databázemi včetně PDBe. Další možností je rozhraní Entrez (<http://www.ncbi.nlm.nih.gov/Entrez>). Jádrem pro struktury proteinů je zde PDB. Přímý přístup k PDB pak poskytuje RCSB PDB na <http://www.rcsb.org/pdb/home/home.do>.

Více informací o dostupných zdrojích lze nalézt v knize Úvod do praktické bioinformatiky [6] a Bioinformatics: a practical guide to the analysis of genes and proteins [3].

3.1.1 RCSB PDB

RCSB PDB [18] je jeden z členů wwPDB poskytující informace o proteinových strukturách s vlastním webovým rozhraním. Každá struktura je zde jednoznačně identifikovatelná pomocí ID. PDB ID je čtyřmístné a má následující syntaxi:

[0–9] [a–z|0–9] [a–z|0–9] [a–z|0–9]

tedy např. 9ins, 4hhb. V systému lze vyhledávat nejen pomocí ID, ale i autora, klíčových slov, data uveřejnění, sekundární struktury, sekvencí a mnoho dalších. Vyhledaný záznam pak nabízí souhrnný přehled o struktuře (jako autor, klasifikace, experimentální metoda aj.),

detail o sekvenčních, biologické a chemické údaje, záznam o materiálech a metodách a nakonec podrobnou analýzu struktury. Nedílnou součástí je trojrozměrná vizualizace. Systém nabízí několik vizualizačních nástrojů, většinou jako plugin do prohlížeče. Mimo to, je zde možnost stáhnutí souborů popisující strukturu v několika formátech. Základním formátem je PDB, dále je zde k dispozici XML verze PDB – PDBML a novější typ mmCIF.

3.1.2 NCBI Entrez

NCBI (National Center of Biotechnology Information) Entrez [20] je webové rozhraní poskytující přístup k mnoha databázím různého zaměření. Vyhledávání probíhá přes všechny databáze a podle nalezených výsledků si pak lze zvolit oblast, která nás zajímá. Např. po zadání PDB ID 9ins bude v oblasti 3D struktur jeden záznam, v proteinových sekvencích dva záznamy, jeden nalezený záznam u PubMed (citace a abstrakty) a další.

Trojrozměrné struktury zde zastupuje MMDB (Molecular Modeling DataBase), jejíž jádro tvoří PDB. Vyhledávat lze pomocí klíčových slov nebo ID (PDB či MMDB). MMDB ID je číselný kód jako např. 3432, což odpovídá PDB ID 9ins. Samotný záznam pro 3D strukturu je oproti RCSB značně kratší z důvodu úvodního rozdělení do oblastí. Obsahuje jen nejdůležitější informace jako reference, popis, datum vložení, autor a zdroj. Součástí je přímý odkaz do databáze konzervovaných domén (uchovává domény, které se opakují v různých proteinech). K dispozici je zde pouze jeden prohlížeč nástroj – Cn3D. Nutná je však předchozí instalace. Dále je možno stáhnout soubor s popisem struktury a to ve dvou formátech, standardní PDB nebo typ ASN, který podporuje již zmíněný Cn3D.

3.2 Formáty dat

3.2.1 Formát PDB

PDB je standardní formát pro zápis trojrozměrné struktury. Jedná se o čistě textový zápis vyvinutý Brookhaven National Laboratory v roce 1971. MIME typ je chemical/x-pdb. Nevýhodou tohoto formátu je, že nenese informaci o chemických vazbách [21].

PDB formát obsahuje ASCII znaky a řádek je zarovnán na 80 znaků. Soubor je složen z množství záznamů, z nichž některé jsou povinné, jiné volitelné. Jeden typ záznamu se může v celém souboru vyskytnout pouze jednou nebo vícekrát. Některé zabírají prostor jednoho řádku, jiné jsou víceřádkové. Příkladem povinného záznamu, který se smí v souboru vyskytnout jedenkrát a zabírá prostor jednoho řádku je HEADER, úvodní záznam celého souboru. Formát záznamu je následující:

sloupec	datový typ	definice
1 – 6	název záznamu	„HEADER“
11 – 50	řetězec(40)	klasifikace
51 – 59	datum	datum vložení
63 – 66	ID kód	PDB ID

Podobným způsobem jsou definovány všechny typy záznamů. HEADER pro strukturu s PDB ID 9INS je vidět na obrázku 3.1. Protein s PDB ID 9INS [2] jsem vybrala pro ukázkou souborových formátů i pro některé vizualizace.

0	10	20	30	40	50	60	70	80
1	HEADER	HORMONE			23-OCT-91	9INS		

Obrázek 3.1: Záznam HEADER pro 9INS

Záznamy souboru pdb více méně odpovídají informacím, které lze vyhledat pomocí webového rozhraní. Ne všechny se přímo týkají struktury, jako tzv. souhrnně popisné (autor, název, zdroj aj.). Také zde existuje mnoho záznamů REMARK s pořadovým číslem 0 – 999, kde jsou doplňující informace, které nenáleží do žádného jiného záznamu. Záznamy, které se týkají popisu primární struktury jsou DBREF (křížové odkazy do sequence databází), SEQADV (rozdíl mezi záznamem SEQRES a záznamem v sequence databázi, kterou udává DBREF), SEQRES (sekvence aminokyselin, viz obrázek 3.2) a MODRES (identifikace modifikací standardních residuí).

0	10	20	30	40	50	60	70	80									
385	SEQRES	1	A	21	GLY	ILE	VAL	GLU	GLN	CYS	CYS	THR	SER	ILE	CYS	SER	LEU
386	SEQRES	2	A	21	TYR	GLN	LEU	GLU	ASN	TYR	CYS	ASN					
387	SEQRES	1	B	30	PHE	VAL	ASN	GLN	HIS	LEU	CYS	GLY	SER	HIS	LEU	VAL	GLU
388	SEQRES	2	B	30	ALA	LEU	TYR	LEU	VAL	CYS	GLY	GLU	ARG	GLY	PHE	PHE	TYR
389	SEQRES	3	B	30	THR	PRO	LYS	ALA									

Obrázek 3.2: Záznam SEQRES pro 9INS

Sekundární struktury se pak týkají záznamy HELIX (obrázek 3.3) a SHEET, které identifikují struktury.

	0	10	20	30	40	50	60	70	80
395	HELIX	1	A11	GLY	A	1	ILE	A	10
396	HELIX	2	A12	SER	A	12	GLU	A	6
397	HELIX	3	B11	SER	B	9	GLY	B	12

Obrázek 3.3: Záznam HELIX pro 9INS

Ze záznamu na obrázku 3.3 můžeme jednoduše vyčíst, že protein obsahuje celkem tři struktury typu helix. První začíná glycinem v řetězci A, sekvenční číslo glycinu je 1 (řetězec touto aminokyselinou začíná). Struktura končí isoleucinem v řetězci A, sekvenční číslo je 10. Třída helixu je 1, což odpovídá pravotočivému α -helixu a celková délka struktury je 10 aminokyselin. Obdobným způsobem lze přečíst zbylé dva záznamy.

Dále jsou tu záznamy o chemických propojeních jako SSBOND (identifikuje disulfidickou vazbu, jak je vidět na obrázku 3.4), LINK (meziresiduální spojení) a CISPEP (identifikuje cis uspořádání, což je méně časté prostorové uspořádání atomů).

	0	10	20	30	40	50	60	70	80
400	SSBOND	1 CYS A	6	CYS A	11				
401	SSBOND	2 CYS A	7	CYS B	7				
402	SSBOND	3 CYS A	20	CYS B	19				

Obrázek 3.4: Záznam SSBOND pro 9INS

Zde na obrázku 3.4 je vidět název aminokyseliny, označení řetězce, ve kterém se nalézá a její pořadové číslo.

Samotné umístění jednotlivých atomů nese záznam ATOM, jehož syntaxe je popsána v následující tabulce:

sloupec	datový typ	definice
1 – 6	název záznamu	„ATOM“
7 – 11	integer	pořadové číslo atomu
13 – 16	atom	název atomu
17	char	indikátor alternativní lokace
18 – 20	název residua	tříznakový kód aminokyseliny
22	char	identifikátor řetězce
23 – 26	integer	pořadí aminokyseliny
27	Achar (znak abecedy)	kód pro vložení residua
31 – 38	real(8.3)	x souřadnice v Å ¹
39 – 46	real(8.3)	y souřadnice v Å
47 – 54	real(8.3)	z souřadnice v Å
55 – 60	real(8.2)	obsazenost
61 – 66	real(8.2)	teplotní faktor
77 – 78	LString(2)(literal string)	symbol elementu
79 – 80	LString(2)	zátěž pro atom

Obázek 3.5 pak ukazuje několik prvních záznamů ATOM u struktury 9INS. Další záznam popisující umístění je HETATM (umístění heteroatomů, tedy atomů, které nejsou přímo součástí aminokyselin).

	0	10	20	30	40	50	60	70	80			
410	ATOM	1	N	GLY	A	1	13.749	47.133	32.172	1.00	29.44	N
411	ATOM	2	CA	GLY	A	1	13.950	46.145	31.092	1.00	27.61	C
412	ATOM	3	C	GLY	A	1	15.306	45.457	31.262	1.00	38.42	C
413	ATOM	4	O	GLY	A	1	15.914	45.566	32.339	1.00	24.76	O
414	ATOM	5	N	ILE	A	2	15.821	44.792	30.234	1.00	25.40	N

Obrázek 3.5: Záznam ATOM pro 9INS

Podrobný popis pdb souboru a všech jeho záznamů je k dispozici na webu wwPDB. V mé práci jsem čerpala z verze 3.2 [1].

3.2.2 Formát mmCIF

Macromolecular Crystallographic Information File (mmCIF), je novější alternativa k pdb souboru. Byl vyvinut v 90. letech 20. století organizací International Union of Crystallography. Jedná se o čistě textový zápis s MIME typem chemical/x-cif nebo chemical/x-mmCIF.

Data jsou zde oproti pdb poskládána poněkud odlišně. Pro srovnání uvedu záznam HEADER, který obsahuje informaci o klasifikaci, datumu vložení a PDB ID. Kdybychom chtěli tyto údaje najít v souboru mmCIF, museli bychom hledat položky _struct_keywords.pdbx_keywords, _database_PDB_rev.date_original a _entry.id. Jak je vidět na obrázku 3.6, jednotlivé informace se ani nevyskytují na stejném místě (čísla řádů jsou uvedeny vlevo).

¹Angstrom – jednotka délky využívaná v chemii pro velikost atomů či délku vazby. 1Å = 0,1nm.

Dále je zde patrné, že některé údaje jsou seskupeny a uvozeny slovem `loop_` a výčtem položek, za nimiž se nachází jednotlivá data, tak jak určuje výčet. Tento způsob zápisu mi přijde užitečný, protože nemusím znát přesné zarovnání do sloupců jako v pdb. Přesto mi pdb formát přijde čistší a přehlednější.

Stručná charakteristika formátu je dostupná na webu Wolfram Mathematica 7 [22], korespondenci pdb záznamů a mmcif položek pak lze najít na stránkách RCSB PDB [15].

0	10	20	30	40	50	60	70	80
1	HEADER	HORMONE			23-OCT-91	9INS		

0	10	20	30	40	50	60	70	80
382	_struct_keywords.pdbx_keywords	HORMONE						

0	10	20	30	40	50	60	70	80
12	#							
13	loop_							
14	_database_PDB_rev.num							
15	database_PDB_rev.date							
16	_database_PDB_rev.date_original							
17	_database_PDB_rev.status							
18	_database_PDB_rev.replaces							
19	_database_PDB_rev.mod_type							
20	1	1991-11-07	1991-10-23	?	9INS	0		
21	2	2003-04-01	?	?	9INS	1		
22	#							

0	10	20	30	40	50	60	70	80
2	#							
3	_entry.id	9INS						
4	#							

Obrázek 3.6: Srovnání pdb záznamu HEADER (nahore) a odpovídajících mmCIF položek pro 9INS

Kapitola 4

3D vizualizace

V úvodu této kapitoly popisují některá standardní zobrazení. Dále je zde jemný průřez vývojem nástrojů počínaje RasMolem. Většina kapitoly je pak věnována podrobnému popisu zkoumaných nástrojů. V poslední části je tabulka se stručným přehledem vybraných vlastností nástrojů.

Z historického hlediska byla vizualizace proteinů záležitostí velmi drahých a specializovaných počítačů. Dnes, díky velkému rozvoji v oblasti hardware, je možné prohlížet a pracovat s tojrozměrným obrazem na běžném počítači. Nástroje nám umožňují rotovat se strukturou, používat barvy, pracovat se světlem a mnohé další.

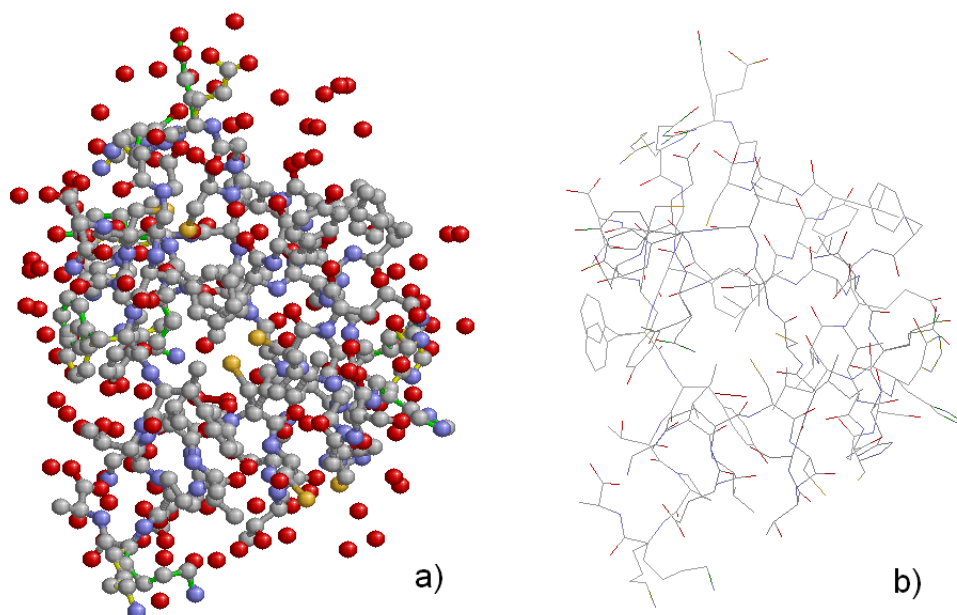
Pro zobrazení struktury proteinů jsou typické čtyři varianty. První dvě (obrázek 4.1) zobrazují celou páteř proteinu včetně postranních řetězců. Jsou to wireframe, nebo-li drátový model, a ball-and-stick, který znázorňuje atomy jako kuličky (ball) a propojení mezi nimi jako tyčky (stick). Třetím zobrazením je ribbons (obrázek 4.2), které kopíruje páteř proteinu, přičemž dobře znázorňuje sekundární struktury. A nakonec spacefill (obrázek 4.2), někdy též nazývaný CPK (Corey, Pauling, Koltun), který vyobrazuje pouze atomy jako velké koule bez propojení. Velikost koule je dána van der Waalovým poloměrem, což je nejmenší vzdálenost, na kterou se k atomu může přiblížit jiný atom.

Označení CPK se používá kromě zobrazení i pro způsob obarvení molekuly. Uhlík bývá šedý, dusík modrý, kyslík červený, vodík bílý, síra žlutá apod. Barevné schéma CPK jsem použila na všech obrázcích ukazující typy zobrazení.

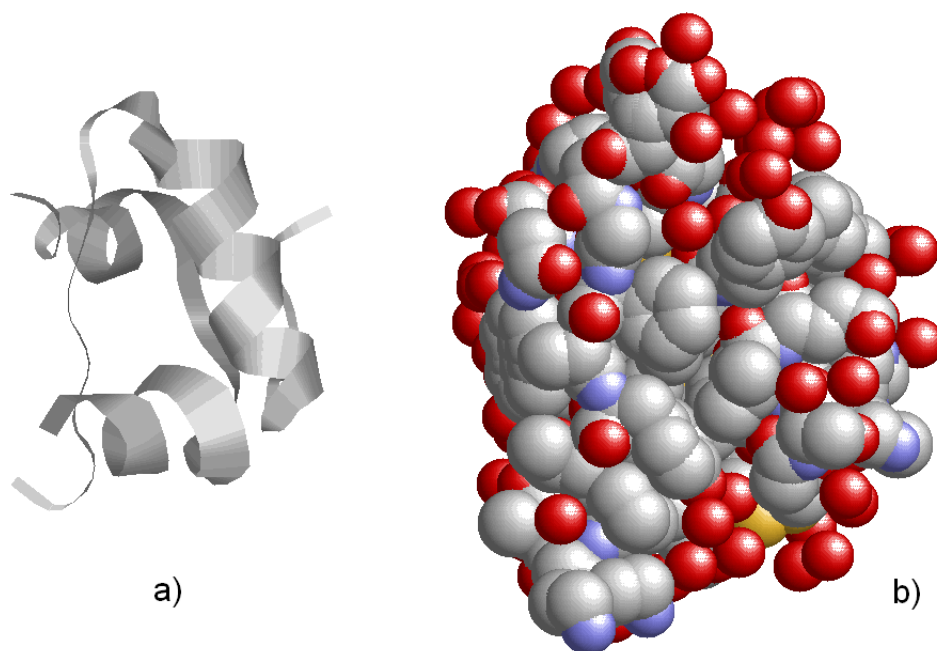
Prvním novodobým softwarem, který umožňoval práci s trojrozměrnou strukturou široké veřejnosti, byl v roce 1993 RasMol (Raster Molecule). Byl napsán v jazyce C s veřejným kódem, což umožnilo vývoj dalších nástrojů, které z RasMolu vycházely. Přepsáním do jazyka C++ a přidáním dalších funkcí vznikl plug-in Chime. Dalším nástrojem, který vychází z RasMolu je např. java applet WebMol. Poměrně dobrým nástrojem hlavně po grafické stránce je Cn3D („see in 3D“). Po předchozí instalaci je možno jej spouštět přímo přes rozhraní Entrez. Asi nejmocnějším nástrojem, z volně přístupných, je Swiss-PDB Viewer (nebo-li Deep View). Některé z nástrojů nabízí přímo RCSB PDB. Z výše zmiňovaných je to WebMol, dále pak např. Jmol nebo QuickPDB.

Pro analýzu jsem si vybrala celkem tři nástroje. Je to RasMol, který je nejstarší a položil základ dalším nástrojům. Dále Cn3D a to hlavně pro jeho úzké propojení s rozhraním Entrez, kterým jsem se v rámci práce také zabývala. A nakonec Swiss-PDB Viewer, zejména kvůli širším možnostem v oblasti tvorby skriptů.

Další informace o 3D vizualizaci, stejně jako přehled a základní charakteristiky vybraných nástrojů lze najít v knize Bioinformatics: a practical guide to the analysis of genes and proteins [3]. Pro ukázkou typů zobrazení jsem opět vybrala protein 9INS [2].



Obrázek 4.1: Standardní zobrazení struktury proteinu: a) ball-and-stick b) wire-frame. Vykresleno programem RasMol.



Obrázek 4.2: Standardní zobrazení struktury proteinu: a) ribbons b) spacefill. Vykresleno programem RasMol.

4.1 RasMol

RasMol [17] je nejstarší prohlížeč nástroj prostorové struktury molekul. Nejaktuálnější je verze 2.7.4.2 z března 2008. Je běžně dostupný pro široké spektrum operačních systémů (MS Windows, Unix, Apple Macintosh). Podporuje velké množství formátů pro popis struktury, včetně pdb a mmCIF. Po spuštění se skládá ze dvou oken. Hlavní vykreslovací okno s menu, kde můžeme manipulovat se strukturou myši (rotace, posuvy, zoom) a vykonávat základní operace, jako je změna zobrazení či barevného schématu. Druhé terminálové okno slouží pro práci se strukturou pomocí příkazů.

4.1.1 Výběr

Všechny změny, které můžeme v nástroji provádět, se aplikují pouze na vybranou část molekuly. Výběr se provádí příkazem `select <výraz>`. Výrazem může být aminokyselina, řetězec, atom, pořadové číslo residua nebo jejich rozmezí. Navíc jsou zde i předdefinovány tzv. sety jak na základě struktury (helix, backbone), tak chemických vlastností (polar, acidic). Uživatel si může definovat i vlastní set. Ve výrazech lze uplatňovat logické operátory and, or a not. Například příkaz `select backbone and not helix` vybere všechny atomy páteře mimo strukturu helix. Pro kontrolu výběru slouží příkaz `set display selected`, který zobrazí celou strukturu modře, jen vybrané části žlutě. Pro zobrazení určité části pak slouží příkaz `restrict <výraz>`.

Jiný způsob výběru je přímo ve vykreslovacím okně pomocí myši. Samotnému výběru však musí předcházet nastavení myši příkazem `set picking atom|group|chain`. Další možnosti nastavení myši jsou probrány v části 4.1.5.

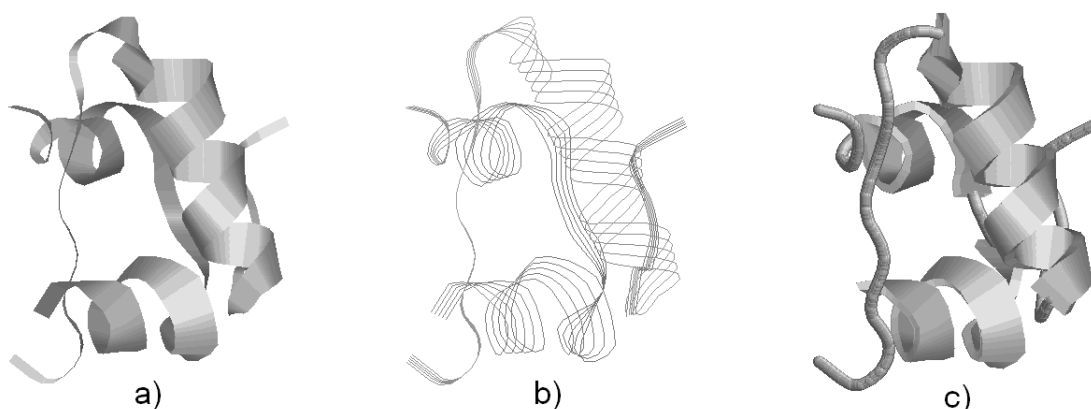
4.1.2 Zobrazení

RasMol nabízí všechna standardní zobrazení, která byla jmenována již v úvodu kapitoly. Jsou přístupná v menu nebo pomocí příkazů, případně kombinacemi příkazů. Zobrazením navíc můžeme určit velikost či šířku. Tedy např. `wireframe on` aktivuje wireframe zobrazení, `wireframe 50` přidá tloušťku jinak velice úzké čáry a vznikne zobrazení stick (pod tímto jménem je dostupné v menu). Kombinací `wireframe 50` a `spacefill 130` docílíme vzhledu ball-and-stick, pro které neexistuje samostatný příkaz.

Kromě toho nástroj nabízí některá další zobrazení dostupná i v menu. Je to backbone, které vykresluje pouze páteř proteinu. Jedná se o spojnice α uhlíků. Příkazem `trace` je možno páteř zobrazit jako splajn. Dalším je strands, které odpovídá zobrazení ribbon, místo pásku je však struktura vykreslena několika souběžnými křivkami. Ribbon odpovídá i zobrazení cartoon, které přidává hloubku vykreslení. V některé literatuře je cartoon a ribbon rozuměno jako identické zobrazení. Porovnání zobrazení ribbon, strand a cartoon je vidět na obrázku 4.3. Posledním nabízeným zobrazením je molecular surface.

4.1.3 Barvy a barevná schémata

Barvu lze měnit příkazem `colour [objekt] <colour>`. Objektem může být atom, popisek, vodíkový můstek aj. Barvu můžeme určit slovy (některé jsou nadefinované jako blue, orange, magenta aj.) nebo trojicí RGB. Existují i určitá barevná schémata. Je to již na úvodu kapitoly zmiňované CPK, dále amino, které barví dle druhu aminokyselin (cys žlutě, leu zeleně), structure (dle sekundární struktury), temperature (dle teplotního faktoru) a mnohé další.



Obrázek 4.3: Zobrazení proteinu 9INS: a) ribbons b) strand c) cartoon

4.1.4 Skripty

Součástí je podpora skriptů. Soubory mohou mít příponu txt, pak je lze spustit přes terminálové okno, nebo spt, což je spustitelný soubor. Nástroj navíc nabízí možnost nechat si skript vygenerovat podle již upravené struktury příkazem `write script <soubor>`. RasMol skripty jsou pouze posloupností příkazů a nejsou zde podporovány žádné konstrukce vyšších programovacích jazyků jako podmínky či cykly. Přesto lze skript využít pro inicializaci vlastního nastavení (picking, zobrazení, barvy apod.) či k uložení rozpracované struktury.

4.1.5 Další funkce

Mocným příkazem při zkoumání struktury je `set picking`, který nastavuje vlastnost myši při kliknutí na atom. Výchozí nastavení je `set picking ident`, který po kliknutí na atom zobrazí v terminálovém okně identifikační údaje. Zajímavým příkazem je pak `set picking distance`, který umožňuje změření vzdálenosti mezi atomy. Podobně fungují příkazy `set picking angle` a `set picking torsion`.

RasMol dále nabízí zobrazování vodíkových a disulfidických můstků, van der Waalsova povrchu či popisků atomů. Umožňuje některé manipulace s molekulou jako rotovat s částí molekuly v místě vybrané vazby. Má i několik příkazů pro změny vykreslování (světlo, stíny). Nabízí export vytvořeného obrazu do mnoha formátů jako gif, ps, eps, bmp. Možný je i export ramachandran mapy, jedná se však o textový soubor, který můžeme dále zpracovat programem GNUpot.

4.1.6 Shrnutí

Rasmol se mi zdá jako dobrý nástroj s přehledným rozhraním. Menu sice nenabízí všechny operace, ale používání příkazů je velice jednoduché. Program je navíc doplněn přehledným manuálem. Skriptování mi také nečinilo problém, i když nemožnost použití konstrukcí vyšších programovacích jazyků jej dosti omezuje.

K záporům bych zařadila menší funkčnost v oblasti porovnávání struktur. Je zde sice možnost otevření více struktur v jednom okně, ale kromě vlastního pozorování není k dispozici žádná porovnávací funkce. Po grafické stránce program také příliš nevyniká, některá

zobrazení jsou dosti kostrbatá. Navíc zobrazení typu molecular surface je náročné, načtení trvá nějakou dobu a pohyb s molekulou není plynulý.

4.2 Cn3D

Cn3D (nyní verze 4.1) [16] je další nástroj pro prohlížení a práci s 3D strukturou. Je dostupný pro řadu operačních systémů jako Windows, Macintosh a Unix. Typicky je spouštěn přes rozhraní Entrez, se kterým je úzce spjat. Strukturu samozřejmě lze načíst i lokálně z disku. Jediný podporovaný formát je ASN. Po spuštění se skládá z hlavního vykreslovacího okna s menu. Dále je možno si spustit Message Log, kde se vypisují informace o prováděných změnách, a Sequence/Alignment Viewer, kde vidíme sekvence aminokyselin jednotlivých řetězců. Cn3D neobsahuje žádné terminálové okno, všechny operace jsou tedy prováděny pomocí menu.

4.2.1 Výběr

Díky Sequence/Alignment Viewer oknu je výběr velice jednoduchý, a to označením aminokyseliny či sekvence aminokyselin myší. Případně je také možno vybrat určitou část přímo dvojklikem ve vykreslovacím okně. Položka menu *Show/Hidden* pak nabízí zobrazení pouze vybrané části, zobrazení vybrané části s okolím (zadáva se v Å), či zobrazení zarovnaných úseků (při zarovnání a překrytí dvou molekul). Je zde i volba *Show/Hide: Show Everything*, která zobrazí opět celou molekulu.

4.2.2 Zobrazení

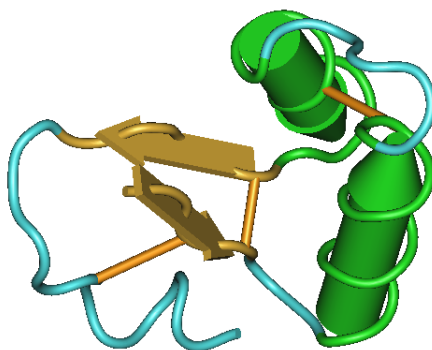
Cn3D nabízí celkem pět různých typů vykreslení molekuly dostupných v menu přes *Style: Rendering Shortcuts*. Jsou to již známé spacefill, ball-and-stick a wireframe (zde uvedeno jako wire). Standardní zobrazení ribbon je nahrazeno Worms (obrázek 4.4), kde je viditelná páteř a sekundární struktury jsou zvýrazněny válcem (helix) či plochým hranolem (sheet) s šipkou. Posledním zobrazením je Tube, které znázorňuje páteř. Doplnující volbou je pak *Style: Rendering Shortcuts: Toggle Sidechains*, která umožňuje skrýt či odkrýt postranní řetězce.

4.2.3 Barvy a barevná schémata

Obarvování molekuly nabízí menu přes *Style: Coloring Shortcuts*. Klasické CPK zbarvení je zde pod volbou Element. Zastoupeno je i barvení podle sekundární struktury (Secondary Structure), řetězců (Molecule), teplotního faktoru (Temperature) či hydrofóbnosti aminokyselin (Hydrophobicity). Při porovnávání dvou molekul je užitečnou volbou barevné schéma *Style: Coloring Shortcuts: Object*, které barví každou strukturu (míněno jako záznam MMDB) jinou barvou. Dále pak *Style: Coloring Shortcuts: Aligned*, které barví červeně zarovnané aminokyseliny. Také je tu další submenu, *Style: Coloring Shortcuts: Sequence Conservation*, jenž nabízí další volby obarvování při zarovnání a překrytí dvou molekul. Při barvení se obarvuje jak obraz ve vykreslovacím okně, tak znaky aminokyselin v okně Sequence/Alignment Viewer, což výrazně přispívá k lepší orientaci ve struktuře.

4.2.4 Skripty

Cn3D nepodporuje skriptování.



Obrázek 4.4: Protein 1CRN [9] ve zobrazení Worm

4.2.5 Další funkce

Jak možná již napovídaly předchozí části, hlavní funkcí nástroje Cn3D je porovnávání molekul. Jednou možností jak zobrazit dvě zarovnané a překryté molekuly je přes rozhraní Entrez přes odkaz do databáze konzervovaných domén. Zde si k původně vybrané struktuře vybereme ze seznamu struktur tu, se kterou chceme porovnávat. Volbou View 3D alignment se nám pak obě struktury již zarovnané zobrazí v Cn3D. Druhou možností je zarovnání, kdy nevyužijeme databázi konzervovaných domén, ale zarovnáme struktury sami. Nejprve načteme do Cn3D strukturu, kterou chceme porovnávat. Pak přes menu v Sequence/Alignment Viewer oknu zvolíme *Imports: Show Import*. Otevře se nám nové okno Import Viewer. Zde v menu vybereme položku *Edit: Import Structure*. Importovat můžeme buď lokálně z disku nebo přímo z internetu. Po importu se nám v okně objeví sekvence které budeme zarovnávat. Nástroj nabízí několik algoritmů pro zarovnání, případně můžeme zarovnání provést ručně. Pro zobrazení zarovnaných struktur zvolíme v menu Import Viewer okna *Alignments: Merge*.

Mimo to je zde možnost uživatelského nastavení přes *Style: Edit Global Style*. Nastavení se týká hlavně barev a typu zobrazení. Také se zde nastavují popisky. Dále nástroj nabízí animaci, kdy se molekula otáčí kolem svislé osy, či export obrázku do png.

4.2.6 Shrnutí

Cn3D má velice přehledné rozhraní a díky Sequence/Alignment Viewer oknu poskytuje i dobrou orientaci v molekule. Po grafické stránce je o mnoho zdařilejší než RasMol. Textura je uhlazenější a pohyb s molekulou je plynulý. Funkčnost v ohledu porovnávání struktur je velice dobrá. Užitečné je i úzké propojení s webovým rozhraním Entrez, zvláště pak s databázemi konzervovaných domén.

Velkým nedostatkem je nemožnost skriptování. Dále zde není možno zobrazit vodíkové vazby či povrchu molekuly. Podpora pouze jednoho formátu popisu struktury, a to ASN, je také spíše záporem.

4.3 Swiss-PDB Viewer

Swiss-PDB Viwer nebo-li Deep View [7] je posledním nástrojem pro prohlížení prostorové struktury molekul, se kterým jsem se v rámci práce seznámila. Nejnovější verze je 4.0.1 z roku 2008. Je dostupný pro operační systémy Windows, Linux i Macintosh. Podporuje soubory typu pdb a mol. V menu je možnost výběru i mmCIF, avšak podpora tohoto formátu je ještě ve vývoji. Po spuštění se skládá z hlavního okna, kde je menu a tool bar s nejčastěji používanými funkcemi. Po načtení struktury je okno doplněno vykreslovací plochou. Dále si uživatel může nechat zobrazit Control panel, kde je vidět sekvence aminokyselin a dají se zde provádět i některé základní operace. K dispozici je i Layer info, kde jsou informace o načtených vrstvách, případně okno Alignment, které ukazuje zarovnání struktur. Po zobrazení povrchu se otevře okno Surface and Cavities s informacemi o povrchu a výdutích.

4.3.1 Výběr

Nejjednodušší způsob výběru je přes Control panel, kde pouhým kliknutím můžeme vybrat podle sekundární struktury (informace o ní je u každé aminokyseliny), řetězce nebo jednotlivé aminokyseliny. Užitečné předvolby pak obsahuje položka menu *Select*, kde je nepřehledné množství způsobů výběru. Lze zde vybírat podle chemických vlastností, sekundární struktury či v určitém okruhu zvolené části. Volbou *Select: Pick on screen* můžeme vybírat aminokyseliny přímo kliknutím ve vykreslovacím okně. Ve skriptu pak vybíráme příkazem `$var = select [in <layer>] <selection>`. V části selection může být přímo název residua, řetězec, číslo pozice, sekvence aj.

4.3.2 Zobrazení

Nástroj nabízí tři standardních zobrazení. Je to wireframe, který je nastaven jako výchozí, ribbons a ball-and-stick. Hlavní ovládání zobrazení je v Control panelu. Pro každou aminokyselinu a heteroatom je zde položka *show*, pro zobrazení atomů páteře, a *side*, pro postranní řetězce. Tyto položky se týkají zobrazení wireframe. Pro ribbons slouží položka *rib*. Další dvě položky v panelu jsou *lab*, pro popisky, a *::v*, která zobrazí van der Waalsův povrch. Zobrazení ball-and-stick je možno dosáhnout přes menu *Prefs: 3D Rendering*, kde se nastaví radius atomů, v kombinaci s oběma standardními vykreslovacími módy. O vykreslovacích módech se blíže zmiňuji v části 4.3.5.

V menu pod *Display: Show Backbone As Carbon Alpha Trace* je možnost zobrazení páteře jako spojnice α uhlíků. *Display* také nabízí zobrazení vodíkových vazeb. Před jejich zobrazením je však nutno vazby nejprve vypočítat. Výpočet vodíkových vazeb, ale třeba i povrchu molekuly najdeme v menu pod *Tools: Compute H-bonds* nebo *Tools: Compute Surface*. Disulfidické vazby jsou zobrazeny stále.

Všechny možnosti zobrazení jsou samozřejmě dostupné i skriptem. Základním příkazem je `show <part> of <selection>`, kde *part* může být právě *res*, *side*, *ribbon* aj. Podobně funguje příkaz `hide`, který naopak skryje určenou část.

4.3.3 Barvy a barevná schémata

Výchozím barevným schématem je CPK. Širokou nabídku způsobů barvení nabízí položka menu *Color*. Můžeme barvit podle typu aminokyselin, sekundární struktury, teplotního faktoru, řetězců, vrstvy (myšleno jako PDB záznam) či hydrofóbnosti. Důležitou položkou při barvení je *Color: act on <string>*, kde *<string>* určuje právě část na kterou se bude

změna barvy aplikovat. Rychlá volba změny barvy je přímo přes Control panel, kde každá aminokyselina má volbu *col*. Zde si může uživatel zvolit vlastní barvu.

Pomocí skriptu se barví příkazem `color <part> of <selection> by <vector>`. Selection značí proměnnou, do které jsme předtím vybrali příkazem `select`. Vector je hodnota RGB, ale přepočítaná do rozmezí hodnot 0.0 – 1.0. Příkaz má i svou obdobu, kde místo vektoru můžeme použít přímo název jedné z předdefinovaných barev. V tomto případě však místo proměnné, kterou jsme získali výběrem vypíšeme přímo výběr, tak jak ho používáme v příkazu `select`.

4.3.4 Skripty

Deep View podporuje tvorbu skriptů. Skript je typu `txt` a spouští se přes menu *File: Run Script*. Vždy musí začínat příkazem `please do` a končit `thank you`. Narozdíl od RasMol skriptů podporuje i konstrukce vyšších programovacích jazyků jako podmínky a cykly. Proměnné mohou být různých datových typů jako `vector`, `float`, `int`, `string` či `file`. Zvláštním datovým typem jsou `layer`, což je vlastně PDB ID, a `selection`, který je výsledkem příkazu `select`. Posledním datovým typem je `internal variable`, jako např. `gCurrentOS`, která nese název operačního systému. Každý příkaz je standardně ukončen středníkem.

Strukturu lze načíst buď z lokálního disku, nebo je přístupná funkce stáhnutí a otevření struktury z internetu (`open [pdb] from „,net“ <string>`). Skripty umožňují díky příkazu `pause <float>`, který pozastaví chod programu na daný počet sekund, tvořit animace. Umí zapisovat, číst či zobrazovat soubory typu `txt`. Je zde také možnost otevřít dialog box, pro komunikaci s uživatelem. Příkazy `phi <selection>` a `psi <selection>` vypočítají torzní úhly. Skriptem lze i zarovnat a překrýt dvě struktury. Možnosti skriptování jsou opravdu široké.

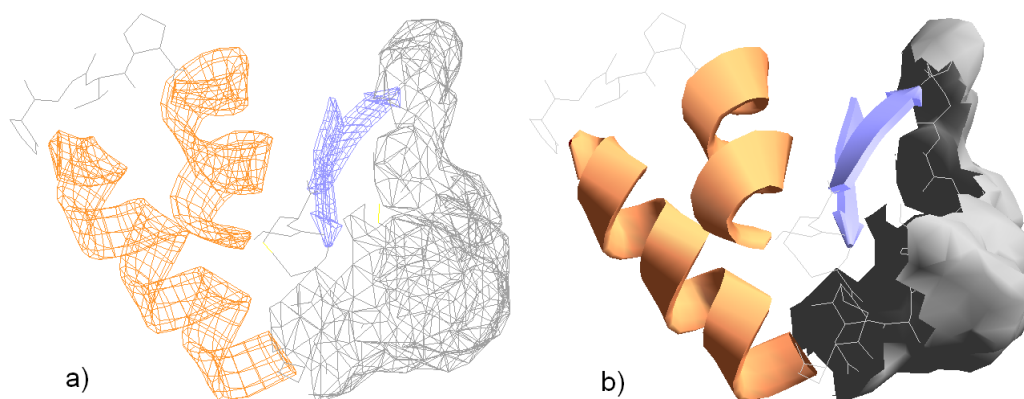
4.3.5 Další funkce

Nástroj má nepřeborné množství funkcí. Kromě již zmiňovaných základních je významná schopnost zarovnat a překrýt struktury dle různých kritérií. V menu tyto funkce můžeme nalézt pod položkou *Fit*. Nástroj umožňuje také homologní modelování, což je předpověď struktury na základě známé sekvence. Další zajímavou funkcí je zobrazení ramachandran mapy, ve které se dají přímo měnit úhly vybrané aminokyseliny. Tato změna má okamžitý vliv na strukturu v hlavním vykreslovacím okně. Ramachandran mapu vygenerovanou Deep View jsem použila pro ilustraci v úvodní sekci o sekundární struktuře (obrázek 2.5).

Uživatelé také jistě ocení velké množství nastavení přes *Prefs*, které se dá exportovat a kdykoli znova nahrát. Můžeme zde měnit použité barvy v schématech, chování molekuly při pohybu, vzhled sekundárních struktur při zobrazení ribbon a další. Dále třeba okno pro přímé zadávání skriptovacích příkazů (*Edit: Script Commands: Execute Script Command*). Při vykreslování nabízí nástroj dva módy dostupné v *Display*, a to `solid 3D` a `3D`, které využívá OpenGL. Rozdíl mezi vykreslením je patrný z obrázku 4.5. Poslední nadstandardní mód je vykreslení pomocí POV-Ray.

4.3.6 Shrnutí

Swiss-PDB Viewer je jednoznačně velice mocným nástrojem. Stejně jako Cn3D poskytuje lepší orientaci v molekule díky Control panelu, který zároveň poskytuje základní operace jako zobrazení, popisky či změna barvy. Skriptování je mnohem dokonalejší díky podpoře



Obrázek 4.5: Protein 1CRN [9]. Sekundární struktury ve zobrazení ribbon a část povrchu molekuly; vše v módu: a) solid 3D b) 3D

konstrukcí vyšších programovacích jazyků. Po grafické stránce je velice zdařilý. K dispozici je velké množství nastavení a vytvořené obrázky mohou dosahovat opravdu vysoké kvality.

Záporem je docela rozsáhlé menu, které z počátku dělá problémy. K nástroji je však k dispozici dobře zpracovaný manuál i tutoriál, který uživatele seznámí s hlavními funkcemi. Po překonání počátečních obtíží s ovládáním jsem si nástroj oblíbila a hodnotím ho jako nejlepší.

4.4 Přehled vlastností nástrojů

Během práce jsem se seznámila se třemi nástroji pro prohlížení prostorové struktury. Byly to RasMol, Cn3D a Swiss-PDB Viewer. V následující tabulce 4.1 jsem shrnula všechny základní funkce a vlastnosti, které nástroje nabízejí a které by měl dobrý nástroj obsahovat. Hodnocení jsem rozdělila následovně: buď nástroj tuto funkci vůbec nenabízí (znak 0), nebo ji nabízí, ale není příliš zdařilá (znak -), nebo ji nabízí (znak +), případně ji nabízí a je vyjímečná (znak ++). Na nezdar či naopak vyjímečnost dané vlastnosti měla především vliv možnost dalšího použití prvku. V případě porovnávání molekul jsem hodnotila vyjímečně nástroj Cn3D a to hlavně díky propojení s rozhraním Entrez a databází konzervovaných domén. Naopak ne příliš zdařile hodnotím zobrazení ball-and-stick v nástroji Swiss-PDB Viewer, kde je dostupné až přes kombinaci různých nastavení v menu, zatím co ostatní zobrazení jsou jednoduše přístupná přes Control Panel.

Celkově hodnotím jako nejlepší nástroj Swiss-PDB Viewer a to i přes jeho složitější ovládání. Má široké možnosti použití, a to jak ve zkoumání a prohlížení molekul jako celku, tak i menších úseků.

	RasMol	Cn3D	Swiss-PDB Viewer
Podpora souborů pro popis struktury			
pdb	+	0	+
mmCIF	+	0	0
Typy zobrazení			
ribbon	+	0	+
wireframe	+	+	+
ball-and-stick	+	+	-
spacefill	+	+	0
Způsoby barvení			
schémata	+	+	+
vlastní	+	+	+
Způsob výběru			
sety	+	+	+
vlastní	+	+	+
Možnosti skriptování			
podpora skriptů	-	0	+
okno pro příkazy	+	0	+
Ostatní			
okno se sekvencí	0	+	+
ramachandran mapa	-	0	++
porovnávání molekul	0	++	+
nastavení vykreslování	-	+	++
export obrázku	+	+	++

Tabulka 4.1: Přehled vybraných vlastností a jejich zastoupení v nástrojích.

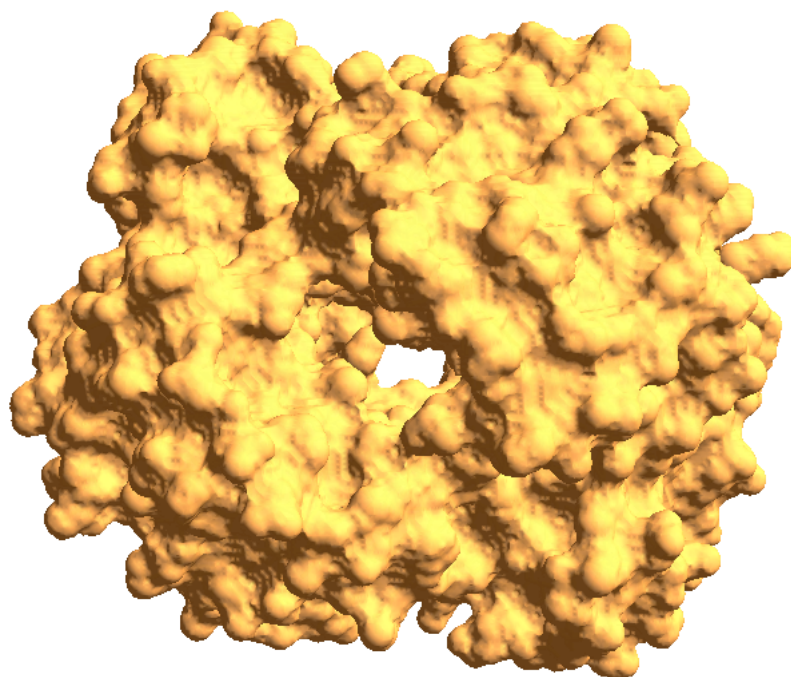
Kapitola 5

Vizualizace hemoglobinu 2DN2

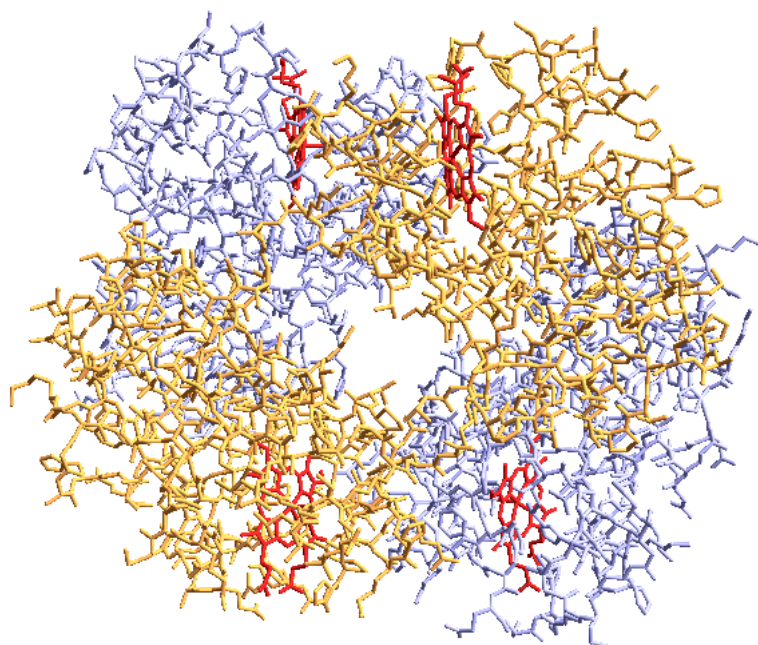
Pro ukázkou vizualizace jsem si vybrala hemoglobin 2DN2 [12]. Jedná se o protein obsažený v červených krvinkách, jehož hlavní úlohou je transport kyslíku. Je typicky kulovitý a skládá se celkem ze čtyř řetězců (dva α a dva β). Každý bílkovinný řetězec obsahuje jednu nebílkovinnou skupinu hem. Hem obsahuje iont železa Fe^{2+} , který poskytuje vazebné místo pro přenos kyslíku. Co se sekundární struktury týká, je zde velké zastoupení struktur helix.

Sekvence aminokyselin a jejich vlastnosti jsou v každé molekule klíčové. U hemoglobinu pak hraje velkou roli hydrofóbnost či hydrofilnost jednotlivých aminokyselin. Záměnou jediné hydrofilní aminokyseliny za hydrofóbní se celkový charakter mění a vzniká nové místo, které má tendenci tvořit komplex s jiným hydrofóbním místem [14]. Přesněji se jedná o záměnu kyseliny glutamové se sekvenčním číslem 6 v β řetězci za valin. Tato mutace má za následek tvorbu vláknitých komplexů molekul hemoglobinu (obrázek 5.7), které sice dále naztrácejí svoji transportní funkci, ale zásadně ovlivňují konečný tvar červenýchrvinek. Místo kulatých ve středu sploštělých buněk vznikají buňky větší a srpkovitého tvaru. Stav, kdy vznikají tyto neplnohodnotné krvinky se nazývá srpkovitá anémie [11] a v některých případech se jedná o smrtelné onemocnění. Důsledek mutace jediné aminokyseliny v řetězci jen dokazuje, jak důležité je znát strukturu proteinů.

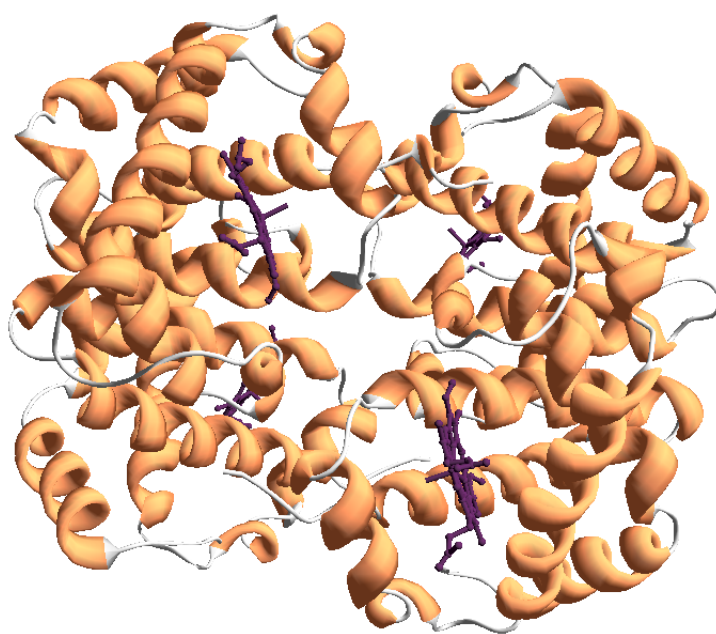
Následující obrázky ukazují molekulu 2DN2, její povrch, sekundární strukturu a detail vazebného místa. Dále mutovaný hemoglobin 2HBS [8], jeho povrch a detail propojení.



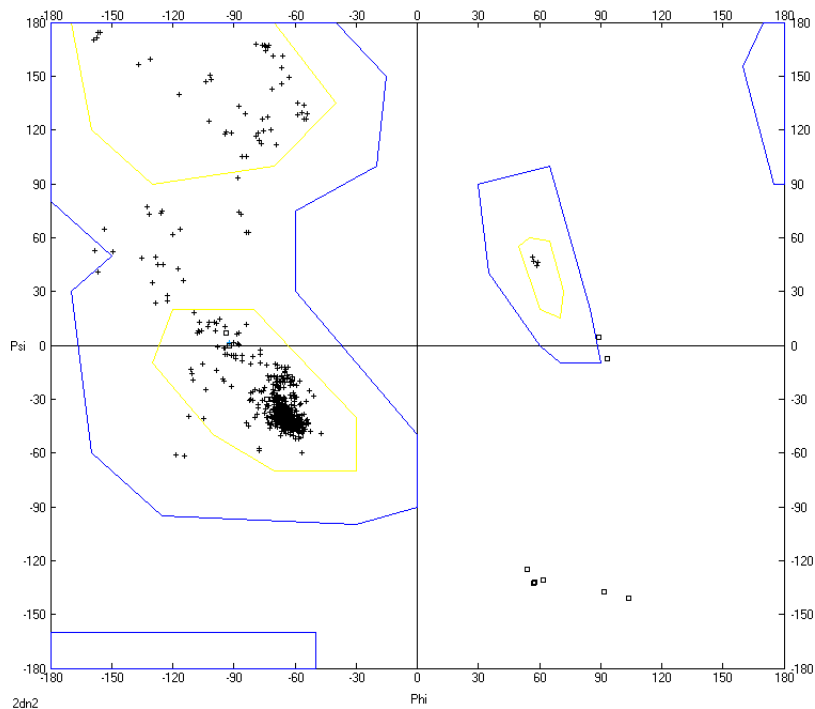
Obrázek 5.1: Protein 2DN2: Povrch molekuly v nástroji Swiss-PDB Viewer.



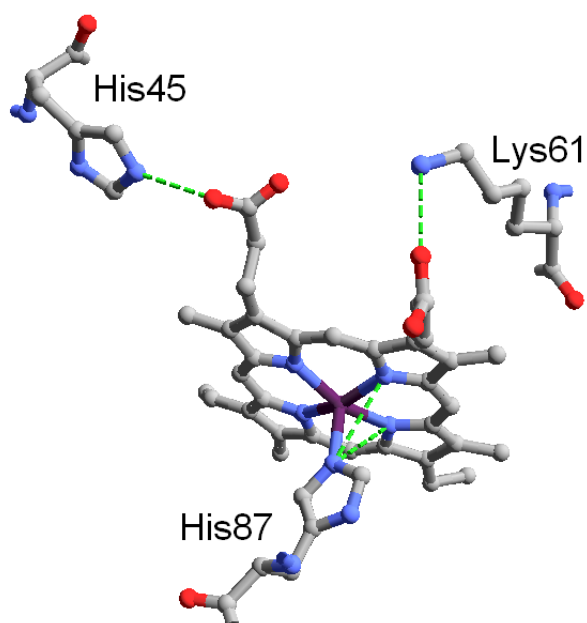
Obrázek 5.2: Protein 2DN2: Vnitřní struktura ve zobrazení wireframe. Barevně odlišeny řetězce (α oranžově, β modře). Nebílkovinný hem zvýrazněn červeně. Vytvořeno v nástroji Swiss-PDB Viewer.



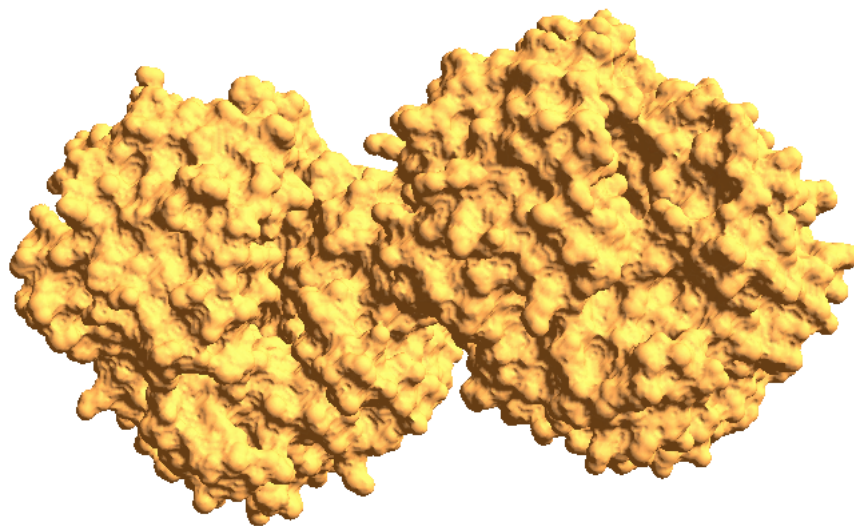
Obrázek 5.3: Protein 2DN2: Sekundární struktura molekuly v nástoji Swiss-PDB Viewer. Zobrazení ribbons s barevným zvýrazněním sekundární struktury. Hem zvýrazněn fialově.



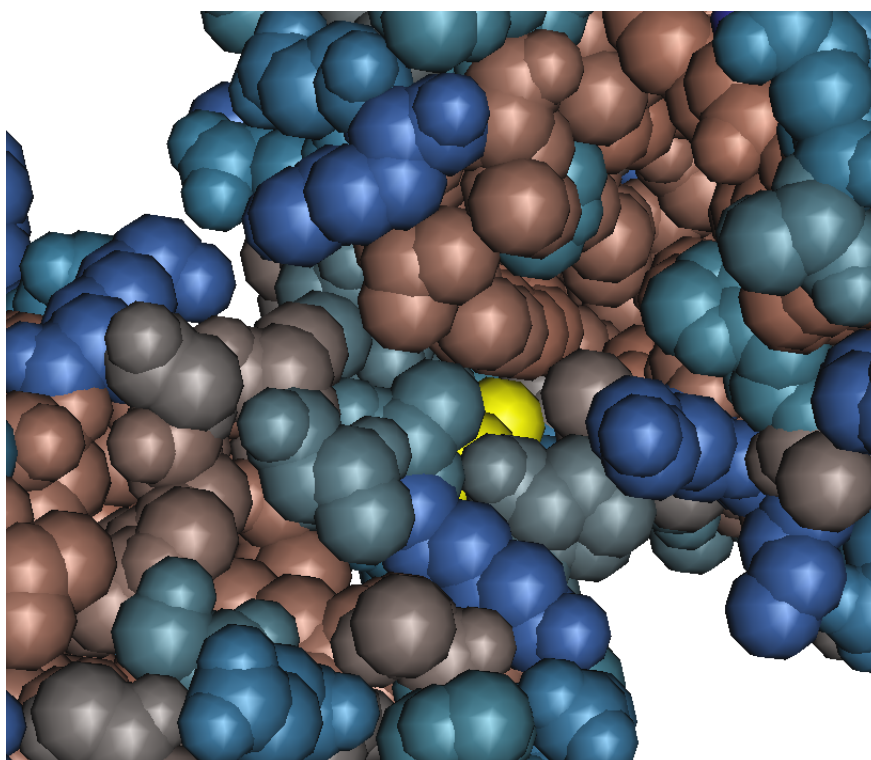
Obrázek 5.4: Protein 2DN2: Ramachandran mapa vytvořena nástojem Swiss-PDB Viewer.



Obrázek 5.5: Protein 2DN2: Detail hemu v řetězci A a jeho vazby na aminokyseliny bílkovinné části. Patrné vodíkové vazby (zeleně) a pevná vazba s histidinem 87. Vytvořeno v programu Swiss-PDB Viewer.



Obrázek 5.6: Protein 2HBS: Povrch molekuly v nástoji Swiss-PDB Viewer.



Obrázek 5.7: Protein 2HBS: Detail propojení mezi molekulami hemoglobinů. Zobrazení spacefill v barvách hydrophobicity (hnědá nejvíce hydrofóbní, modrá nejméně hydrofóbní). Valin 6 v β řetězci zvýrazněn žlutě. Patrné přiblížení k hydrofóbní oblasti (hnědě).

Kapitola 6

Implementace

V rámci bakalářské práce jsem vytvořila několik skriptů pro nástroje RasMol a Swiss-PDB Viewer. V následujících částech popisují jednotlivé skripty.

6.1 RasMol

Protože možnosti skriptování v tomto nástroji jsou dosti omezené, vytvořila jsem jeden soubor `rasDemo.spt`, který ukazuje vybrané možnosti nástroje. Skript se skládá celkem ze šesti obrazů. Po každém obrazu je uživatel vyzván ke stisku klávesy, aby skript pokračoval. Během pauzy (příkaz `pause`) je možno se strukturou pomocí myši rotovat či si ji přiblížovat. Každý nový obraz pak začíná příkazem `reset`, který vrátí molekulu do počáteční polohy. Vše je doplněno vhodným komentářem v terminálovém okně.

Nultý obraz je prázdný a v terminálu se vypíše úvodní informace. První obraz již znázorňuje vybranou molekulu s PDB ID 9INS ve zobrazení ribbons a v barevném schématu chain, ze kterého jsou jasně patrné dva řetězce proteinu. Druhý obraz ukazuje strukturu ve zobrazení wireframe včetně vodíkových a disulfidických vazeb. Vše je barveno ve schématu CPK, který rozlišuje jednotlivé atomy. Třetí obraz se soustředí na detail páteře řetězce B. Pro páteř jsem vybrala příkaz `backbone`. Barevné schéma jsem použila `structure`, který zvýrazňuje sekundární strukturu. Kromě růžového helixu je zde obarven ještě jeden kus modře. Je to tzv. turn, jedna z dalších sekundárních struktur. Čtvrtý obraz se soustřeďuje na helixovou strukturu téhož řetězce. Pro názornou ukázkou jsem ponechala vodíkové vazby. Vše je ve zobrazení wireframe, tvar helixu pak zdůrazňuje zobrazení ribbons. Barevné schéma jsem použila opět CPK. Pátý obraz se vrací k celé struktuře. Ukazuje páteř, tentokrát jsem však použila příkaz `trace`, a cysteiny mezi nimiž jsou disulfidické vazby. Cysteiny jsem zvýraznila zobrazením ball-and-stick v barvách CPK. Atomy síry mezi nimiž jsou patrné vazby jsem doplnila popiskem. Poslední šestý obraz ukazuje možnost vykreslení aminokyseliny a atomů v jejím okolí. Vybrala jsem cystein 6 v řetězci A s atomy v okolí 3Å. Je zde změřena vzdálenost mezi vybranými atomy síry a celý obraz je doplněn tečkovaným van der Waalovým povrchem. Vše ve zobrazení ball-and-stick a barvách CPK.

6.2 Swiss-PDB Viewer

Nástroj Swiss-PDB Viewer poskytuje mnohem širší možnosti ve skriptování. V rámci práce jsem vytvořila celkem tři skripty pro systém MS Windows, které demonstrují vybrané vlastnosti nástroje.

6.2.1 Skript swissDemo1.txt

První skript ukazuje možnost tvoření animací a některé druhy zobrazení na struktuře s PDB ID 1BHP.

Celý skript je tvořen jako animace. Té jsem docílila celkem čtyřmi cykly do `while`. Na úvodu je molekula zobrazena včetně svého povrchu, který je obarven žlutě. První cyklus rotuje s molekulou v ose `y` a v polovině svého trvání skryje povrch molekuly. Duhý cyklus pokračuje v rotaci a postupně mění zobrazení z wireframe na ribbons. Počítadlo, které zde cyklus ukončuje navíc funguje jako sekvenční číslo aminokyseliny, která zrovna mění své zobrazení. Po ukončení přeměny se zobrazí heteroatomy a aminokyseliny se kterými jsou tyto atomy vázány vodíkovými vazbami. Současně se zobrazí i popisky heteroatomů a vybraných aminokyselin. Na to začíná třetí cyklus, který molekulu přibližuje. Polední cyklus pak rotuje s molekulou ve všech třech osách až do ukončení.

Základními příkazy pro animaci jsou rotace (`rotate <x, y, z>`) či zoom (`zoom <float>`) a následná pauza (`pause <float>`). Pro docílení plynulých přechodů mezi jednotlivými obrazy jsem zvolila délku pauzy 0,1.

6.2.2 Skript swissDemo2.txt

Druhý skript ukazuje základní práci s textovým souborem a možnost počítat torzní úhly. Pro ukázkou jsem opět vybrala protein 1BHP.

Na úvodu je molekula ve zobrazení ribbons a obarvena dle sekundární struktury. Skript pak vybere jednu strukturu helix, která zabliká žlutě. Pro každý α uhlík ve vybraném helixu nechám spočítat torzní úhly, které následně2 zapíšu do souboru. Jakmile je soubor vytvořen, nástroj ho otevře pro nahlédnutí v novém okně. Skript pak pokračuje pootočením molekuly a výběrem struktury sheet. Vybraná část opět zabliká, nechám spočítat torzní úhly a zapíšu do souboru, který je v zápětí otevřen pro nahlédnutí. Oba vytvořené soubory jsou pak kdykoli k dispozici ve složce `usrstuff` programu Swiss-PDB Viewer.

Základněm pro práci se soubory je příkaz `$var = open file <string> in usrstuff` `for [reading | writing]` případně `open text <string> in usrstuff` pro otevření v novém okně. Zápis do souboru je možný příkazem `print on $var <string>`.

6.2.3 Skript swissDemo3.txt

Poslední skript ukazuje možnost větší komunikace s uživatelem prostřednictvím dialog boxu. Zároveň demonstruje funkci zarovnání a překrývání struktur. Pro tento případ jsem vycházela ze struktury 1BHP, ke které jsem si prostřednictvím Entrez a databáze konzervovaných domén vyhledala několik dalších struktur se kterými lze protein 1BHP překrýt. Překrytí 1BHP s jakoukoli jinou strukturou tedy funguje vždy. Pokud uživatel však vybere náhodně ze seznamu poskytovaných struktur, nemusí se zarovnání povést. Skript na tuto skutečnost upozorní a skončí předčasně.

Po spuštění skript otevře vytvořený seznam poskytovaných struktur, ze kterých může uživatel vybírat. Dialog box vyzve pro vypsání názvu první a poté druhé struktury. Pokud je zadán špatný řetězec, skript na to upozorní a čeká na nový vstup. Po úspěšném zadání a otevření struktur se otevře nový dialog box, který čeká na řetězec dle nabídky, který reprezentuje jednu z možností zarovnání. V případě, že si uživatel bude přát ukončit skript dříve, vpiše do dialog boxu QUIT. Po úspěšném výběru typu zarovnání jsou molekuly překryty a všechny zarovnané části jsou obarveny. Struktura první červeně a struktura druhá

zeleně. Nakonec následuje krátká rotace zarovnaných molekul kolem osy y a vypsání hodnoty RMSD (root mean square deviation). Jak jsem již zmiňovala při neúspěšném zarovnání skript končí předčasně.

Základem pro vytvoření skriptu byla opět práce se souborem, jak jsem již zmiňovala u skriptu swissDemo2.txt. Kontrolu, zda byla vybrána struktura ze seznamu, jsem provedla jednoduše pomocí příkazu `$var = readln from file $var`. Dialog box tvoří obdobný příkaz, a to `$var = readln from user <string>`. Strukturu jsem zarovnávala příkazem `$var = superpose <layer> onto <layer> using <string>`. RMSD jsem vypočítala pomocí `$var = rms of <layer> and <layer> using <string>`.

Kapitola 7

Závěr

V rámci bakalářské práce jsem se seznámila s proteiny, jejich složením a strukturou. Dále jsem prostudovala databáze, které poskytují přístup k biologickým datům. Zde jsem se soustředila na databázi Protein Data Bank (PDB) a na Molecular Modeling DataBase (MMDB), která je přístupná přes webové rozhraní Entrez. Také jsem se seznámila s formátem dat pro popis trojrozměrné struktury, a to standardním pdb a novějším mmCIF. Značnou část práce pak tvořilo zkoumání prohlížečích nástrojů.

Během práce jsem se seznámila celkem se třemi vizualizačními nástroji. Byly to RasMol, Cn3D a Swiss-PDB Viewer. Souhrn jejich vlastností jsem pak zaznamenala do tabulky 4.1. Jako nejlepší jsem ohodnotila nástroj Swiss-PDB Viewer, který má sice složitější ovládání, ale jeho funkčnost je opravdu nejširší.

Součástí práce bylo i vytvoření několika skriptů. Pro nástroj RasMol to byl skript jeden, protože možnosti nástroje v tomto ohledu jsou poněkud omezené. V případě Swiss-PDB Vieweru jsem pak vytvořila skripty tři. První dva ukazují možnosti tvorby animací či práce se soubory. Skript třetí je obsáhlejší a ukazuje možnost vytvoření skriptu, který komunikuje s uživatelem. Cíl takového skriptování vidím ve zautomatizování některých operací.

Pro ukázkou jsem začlenila i vizualizaci vybraného proteinu, kterým byl hemoglobin. V obrázcích jsem ukázala strukturu jako celek, ale i detail vazebného místa. Také jsem se zde snažila zdůraznit celý význam znalosti struktury a její souvislost s funkcí.

Hlavní přínos práce vidím ve zhodnocení kvality a funkčnosti jednotlivých nástrojů, který by mohl čtenáři pomoci při výběru nástroje. Samozřejmě by bylo vhodné začlenit více nástrojů. Také tvorbu skriptů by bylo možné do budoucna rozšířit, kdy ve spolupráci s uživatelem by bylo možné tvořit i rozsáhlejší účelové skripty.

Literatura

- [1] *PDB File Format - Contents Guide Version 3.20 [online]*. 2008-09-15 [cit. 2009-02-28], <http://www.wwpdb.org/docs.html>.
- [2] Badger, J.; Dodson, G.: Monovalent cation binding in cubic insulin crystals. PDB ID: 9INS.
- [3] Baxevanis, A. D.: *Bioinformatics: a practical guide to the analysis of genes and proteins*. John Wiley and Sons, 2005, ISBN 0-471-47878-4.
- [4] Clarc, J.: The Structure of Protein [online].
<http://www.chemguide.co.uk/organicprops/aminoacids/proteinstruct.html>, 2004 [cit. 2009-02-17].
- [5] Clavier, J.-M.; Notredame, C.: *Bioinformatics for Dummies*. Wiley Publishing, 2003, ISBN 0-7645-1696-5.
- [6] Cvrčková, F.: *Úvod do praktické bioinformatiky*. Academia, 2006, ISBN 80-200-1360-1.
- [7] Guex, N.; Schwede, T.: Swiss-Pdb Viewer Deep View [online].
<http://spdbv.vital-it.ch/>, 2008-06-27 [cit. 2009-04-15].
- [8] Harrington, D.; Adachi, K.; Jr., W. R.: The high resolution crystal structure of deoxyhemoglobin S. PDB ID: 2HBS.
- [9] Hendrickson, W.; Teeter, M.: Water molecule of a hydrophobic protein at atomic resolution. Pentagon rings of water molecules in crystal of crambin. PDB ID: 1CRN.
- [10] Kodíček, M.: Biochemické pojmy: výkladový slovník [online].
http://vydavatelstvi.vscht.cz/knihy/uid_es-002/ebook.html?p=bilkoviny-_domena, 2007 [cit. 2009-04-14].
- [11] Kodíček, M.: Biochemické pojmy: výkladový slovník [online].
http://vydavatelstvi.vscht.cz/knihy/uid_es-002/ebook.html?p=anemie_srpkovita, 2007 [cit. 2009-05-01].
- [12] Park, S.; Yokoyama, T.; Shibayama, N.; aj.: 1.25Å resolution crystal structure of human hemoglobin in the deoxy form. PDB ID: 2DN2.
- [13] Publishing, W. B.: Molecular Biology Web Book [online].
<http://www.web-books.com/MoBio/Free/Chap2.htm>, [cit. 2009-02-17].
- [14] WWW stránky: Sick Cell Hemoglobin.
<http://www.wellesley.edu/Chemistry/chem227/bindprotns/hbsickle.htm>, 1999-02-12 [cit. 2009-05-01].

- [15] WWW stránky: Recommended PDB – > mmCIF Data Item Correspondences [online]. <http://mmcif.pdb.org/dictionaries/pdb-correspondence/pdb2mmcif.html>, 2005-05-27 [cit. 2009-03-03].
- [16] WWW stránky: Cn3D Home Page [online]. <http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml>, 2008-04-24 [cit. 2009-04-15].
- [17] WWW stránky: RasMol and OpenRasMol [online]. <http://www.openrasmol.org/>, 2008 [cit. 2009-03-22].
- [18] WWW stránky: RCSB Protein Data Bank [online]. <http://www.rcsb.org/pdb/home/home.do>, 2009 [cit. 2009-02-21].
- [19] WWW stránky: Worldwide Protein Data Bank [online]. <http://www.wwpdb.org/index.html>, 2009 [cit. 2009-02-21].
- [20] WWW stránky: Entrez, The Life Sciences Search Engine [online]. <http://www.ncbi.nlm.nih.gov/sites/gquery>, 2009 [cit. 2009-02-28].
- [21] WWW stránky: Wolfram Mathematica 7 Documentation [online]. <http://reference.wolfram.com/mathematica/ref/format/PDB.html>, 2009 [cit. 2009-02-28].
- [22] WWW stránky: Wolfram Mathematica 7 Documentation [online]. <http://reference.wolfram.com/mathematica/ref/format/MMCIF.html>, 2009 [cit. 2009-03-03].
- [23] Zendulka, J.; Bártík, V.; Lukáš, R.; aj.: Získávání znalostí z databází ZZN, 2006, studijní opora.

Seznam příloh

Dodatek A: Obsah přiloženého CD

Dodatek B: Obrázky vytvořené skriptem rasDemo.spt v programu RasMol

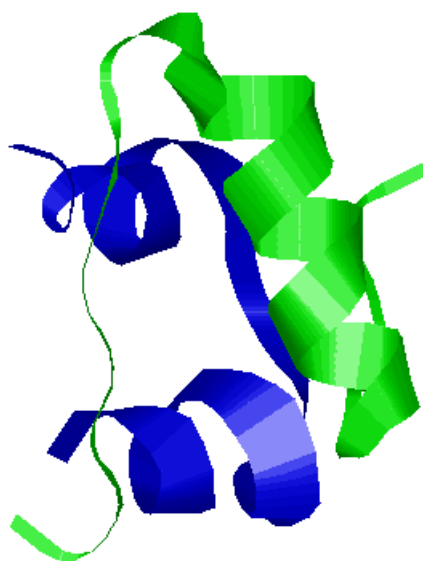
Dodatek A

Obsah příloženého CD

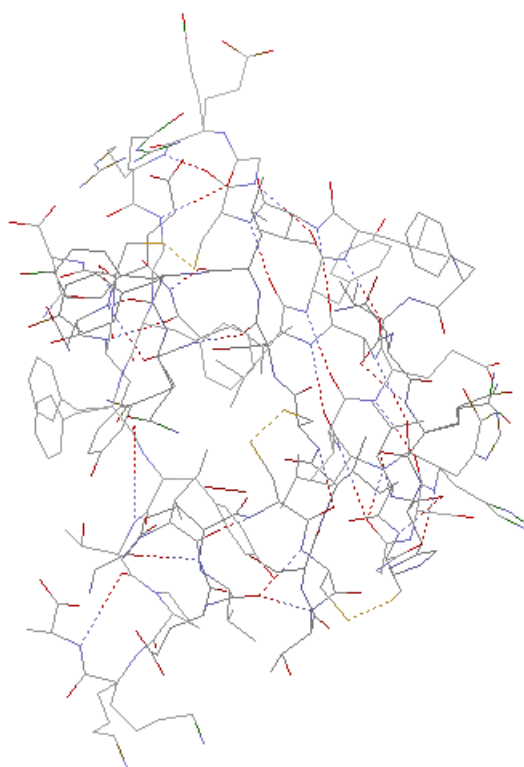
- Bakalářská práce v elektronické podobě
- Skripty
 - rasDemo.spt
 - swissDemo1.txt
 - swissDemo2.txt
 - swissDemo3.txt
 - soubory *.pdb
 - soubor seznam.txt – seznam struktur pro swissDemo3.txt
 - readme s pokyny pro uložení jednotlivých souborů

Dodatek B

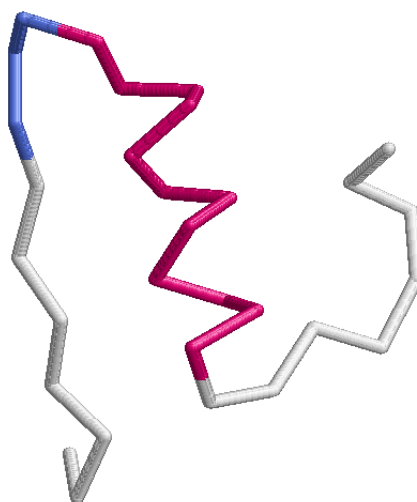
Obrázky vytvořené skriptem rasDemo.spt v programu RasMol



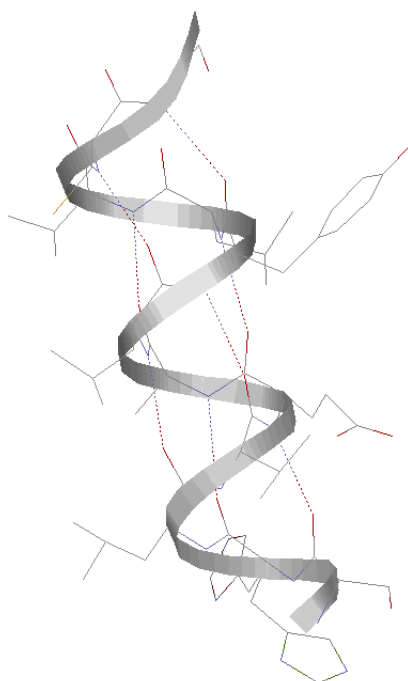
Obrázek B.1: Protein 9INS: Zobrazení ribbon, barevné schéma chain.



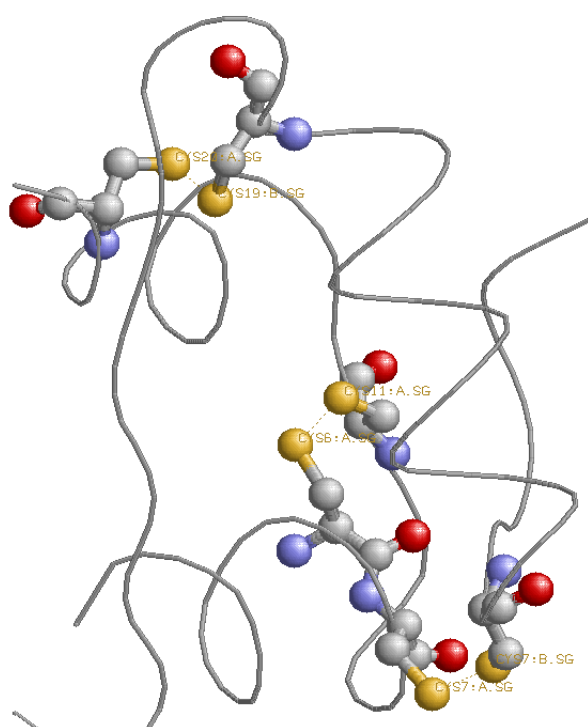
Obrázek B.2: Protein 9INS: Zobrazení wireframe, barevné schéma cpk. Viditelné disulfidické a vodíkové vazby.



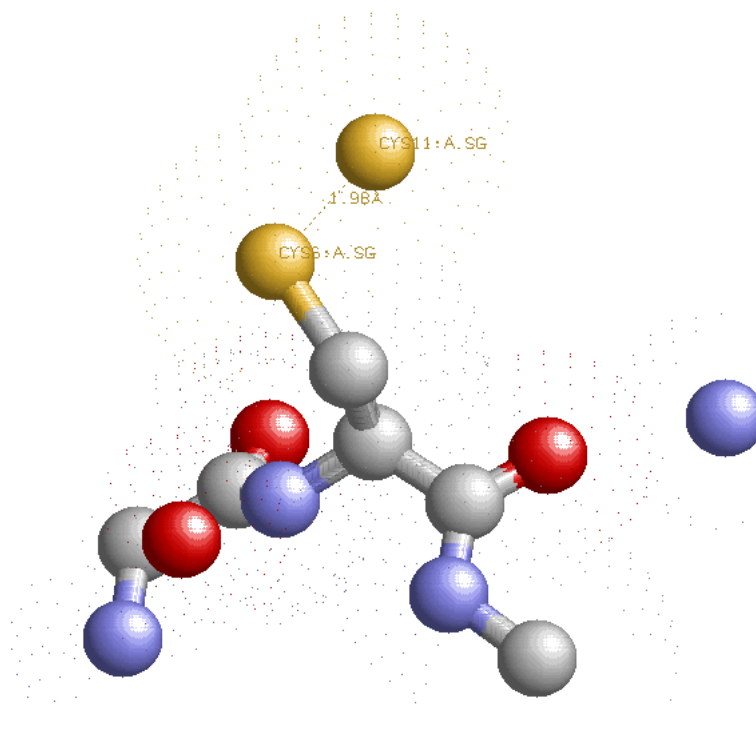
Obrázek B.3: Protein 9INS: Detail řetězce B. Zobrazení backbone, barevné schéma structure.



Obrázek B.4: Protein 9INS: Helix řetězce B s vodíkovými můstky. Zobrazení wireframe a ribbon, barevné schéma cpk.



Obrázek B.5: Protein 9INS: Páteř s disulfidickými vazbami. Zobrazení trace a ball-and-stick (cysteiny), barevné schéma cpk.



Obrázek B.6: Protein 9INS: Cystein 6 v řetězci A a okolní atomy ve vzdálenosti 3Å. Zobrazení ball-and-stick, barevné schéma cpk. Viditelný van der Waalsův povrch a změřená vzdálenost mezi atomy síry.