

Predikce struktury proteinů

C2131 Úvod do bioinformatiky

Jaro 2024

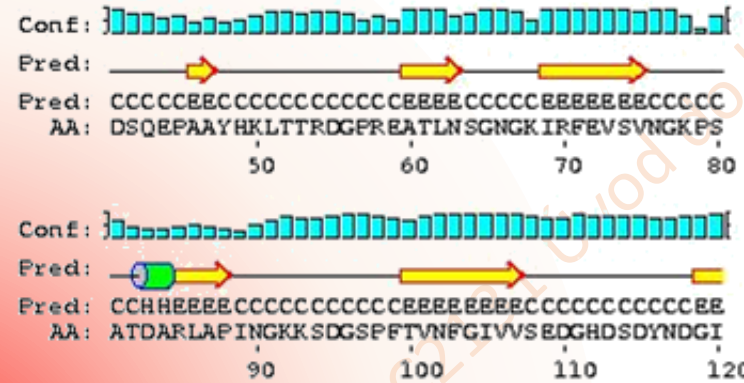
Mgr. Josef Houser, Ph.D.

Struktura proteinů

1D

ADSQTSSNRAGEFSIPPNTDFRAIFFANAAE
QQHIKLFIGDSQEPAAYHKLTTTRDGPREATL
NSGNGKIRFEVSVNGKPSATDARLAPINGK
KSDGSPFTVNFVIGVVSSEDGHDSYNDGIVV
LQWPIG

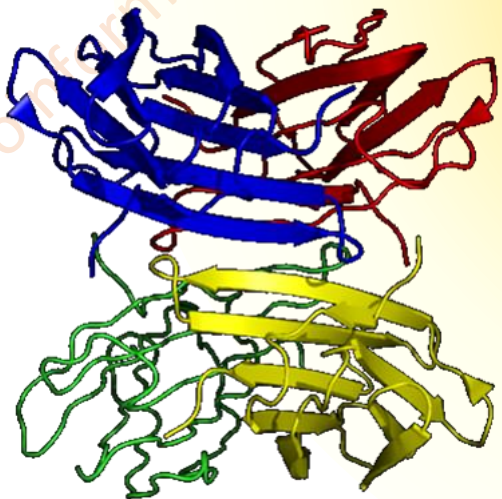
**primární
(sekvence)**



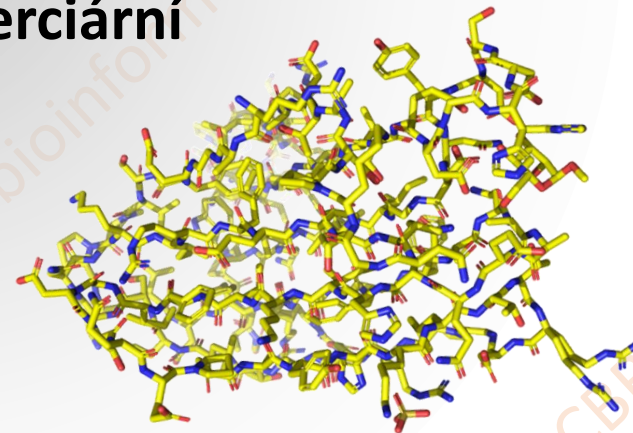
2D

sekundární

kvartérní



terciární

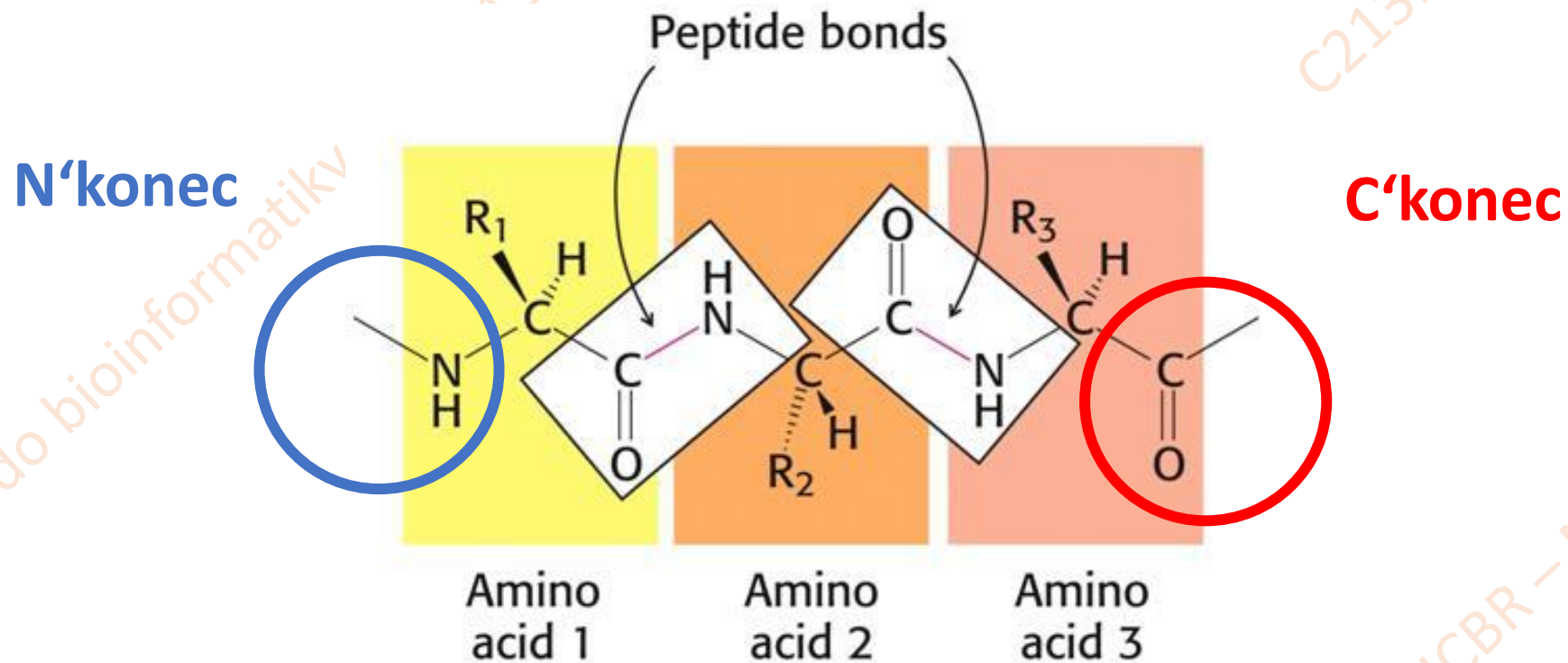


4D

3D

Primární struktura

- Sekvence aminokyselin zapsaná od N' konce k C' konci



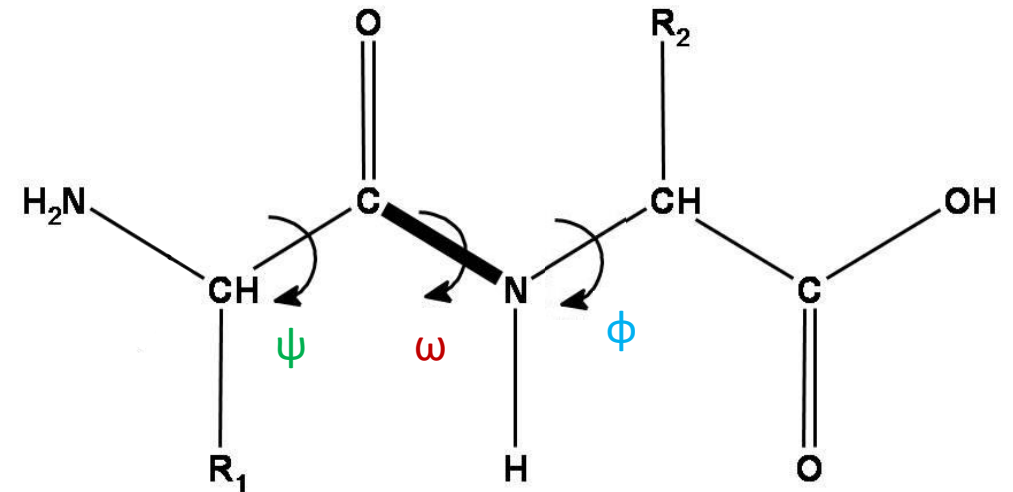
Sekundární struktura

Definována pomocí **torzních úhlů** peptidové páteře

Pro každou aminokyselinu lze definovat tři úhly:

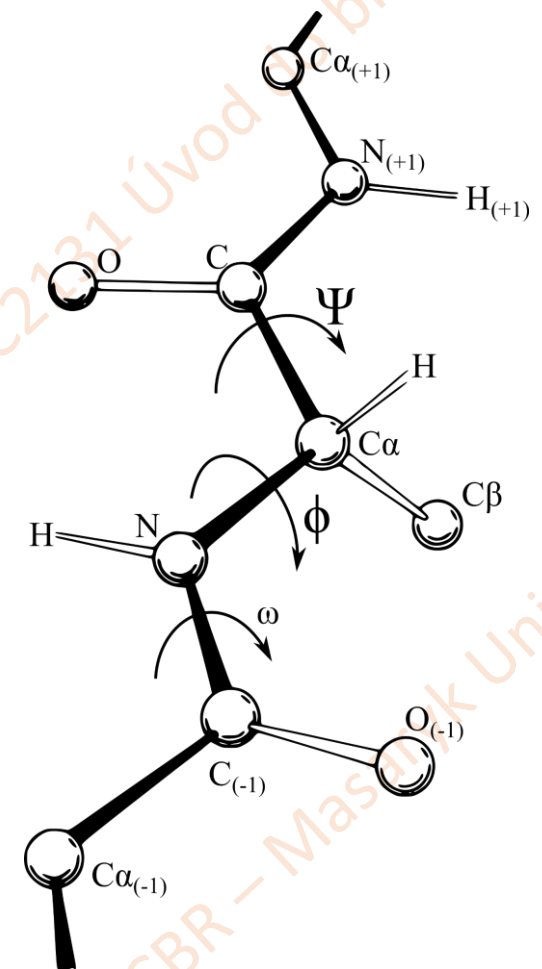
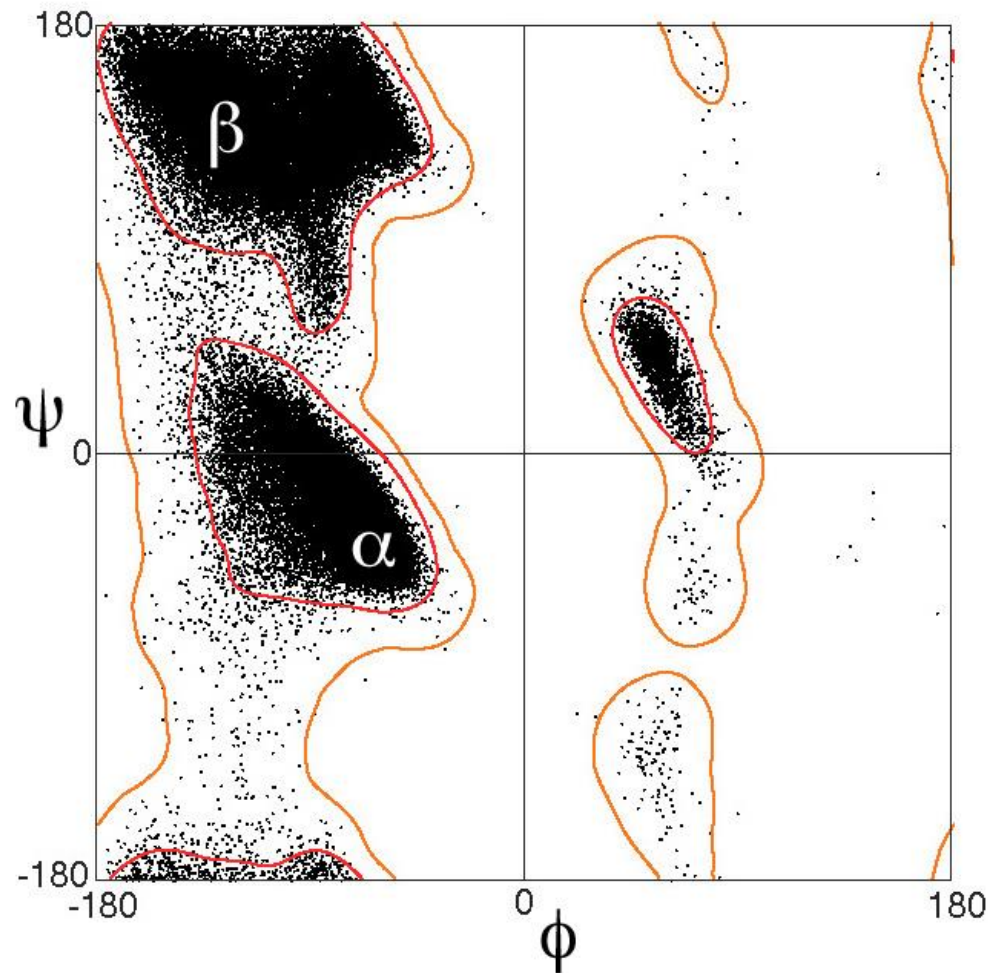
- ϕ – úhel kolem vazby N-C α
- ψ – úhel kolem vazby C α -C_(karbonyl)
- ω – úhel kolem peptidové vazby (180°, výjimečně 0°)

Stabilizována pomocí vodíkových můstků
mezi atomy peptidové kostry



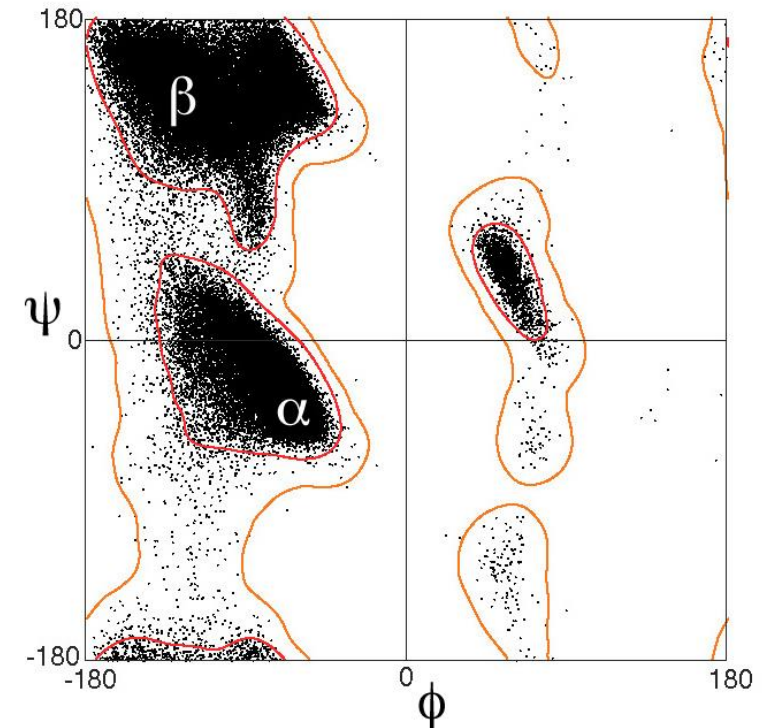
Ramachandranův diagram

Každé aminokyselině odpovídá jeden bod v diagramu



Sekundární struktura

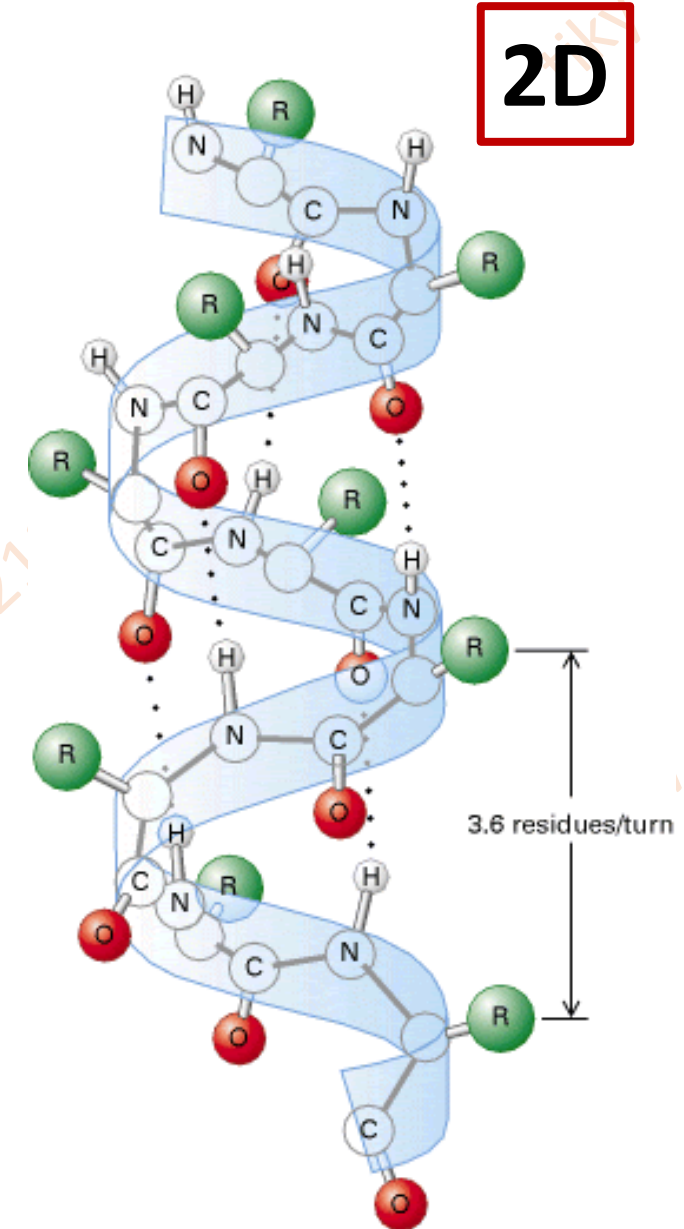
- Stabilní konformace polypeptidového řetězce
- Důležité pro udržení 3D struktury
- Základní typy:
 - α -šroubovice (helix)
 - β -skládaný list (sheet)
 - otáčky, smyčky
- Cca 50 % aminokyselin je součástí α a β struktur



Šroubovice (helix)

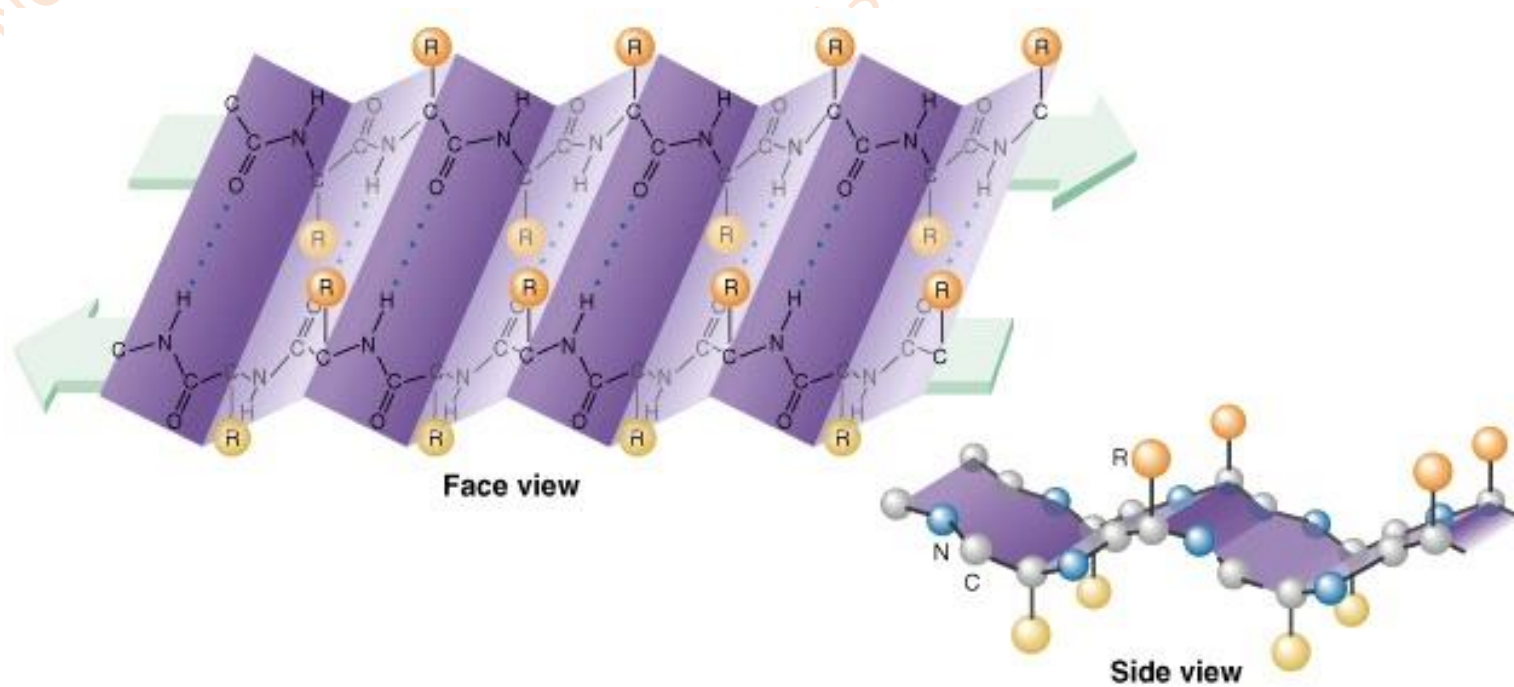
- α -helix – nejčastější
- 3_{10} -helix – obvykle na začátku nebo na konci α -helixu
- π -helix – málo stabilní, málo častý

	α -helix	3_{10} -helix	π -helix
Vodíkové můstky	$n \dots n+4$	$n \dots n+3$	$n \dots n+5$
Residua na otáčku	3,6	3	4,4
Vinutí (Å na 1 AK)	1,5	2	1,15

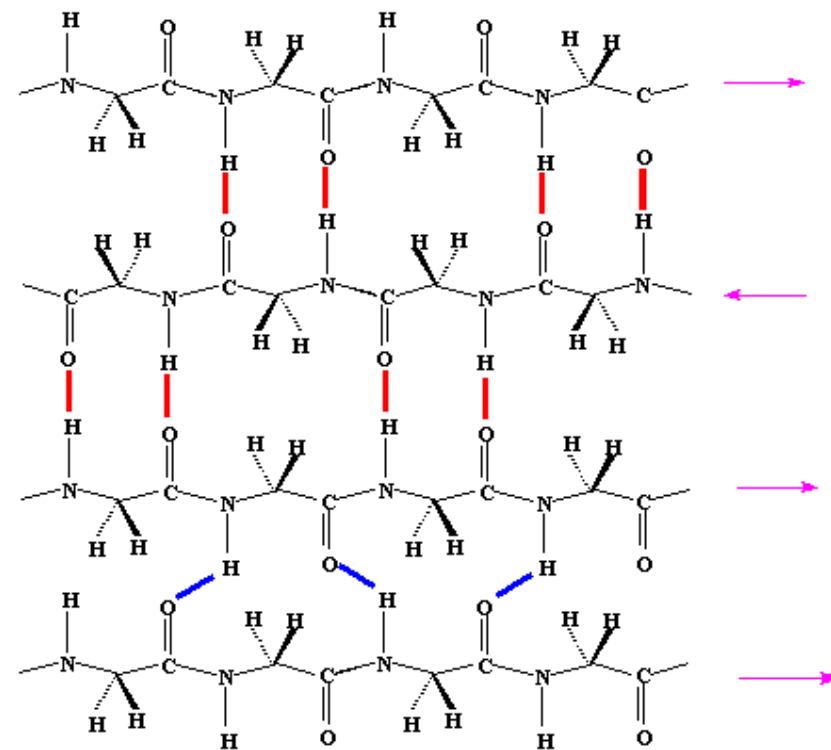


Skládaný list (extended β -sheet)

➤ Paralelní, antiparalelní, mix



Antiparalelní

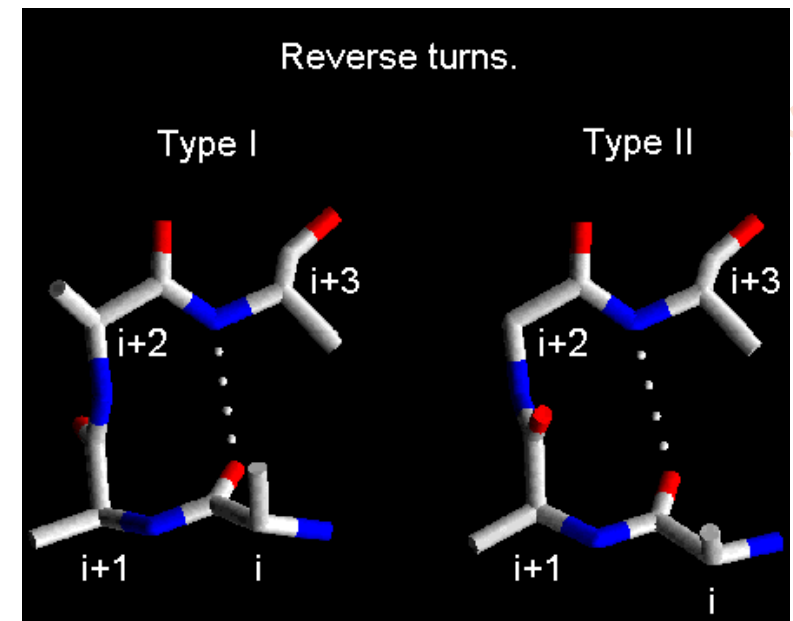


Paralelní

Ostatní

2D

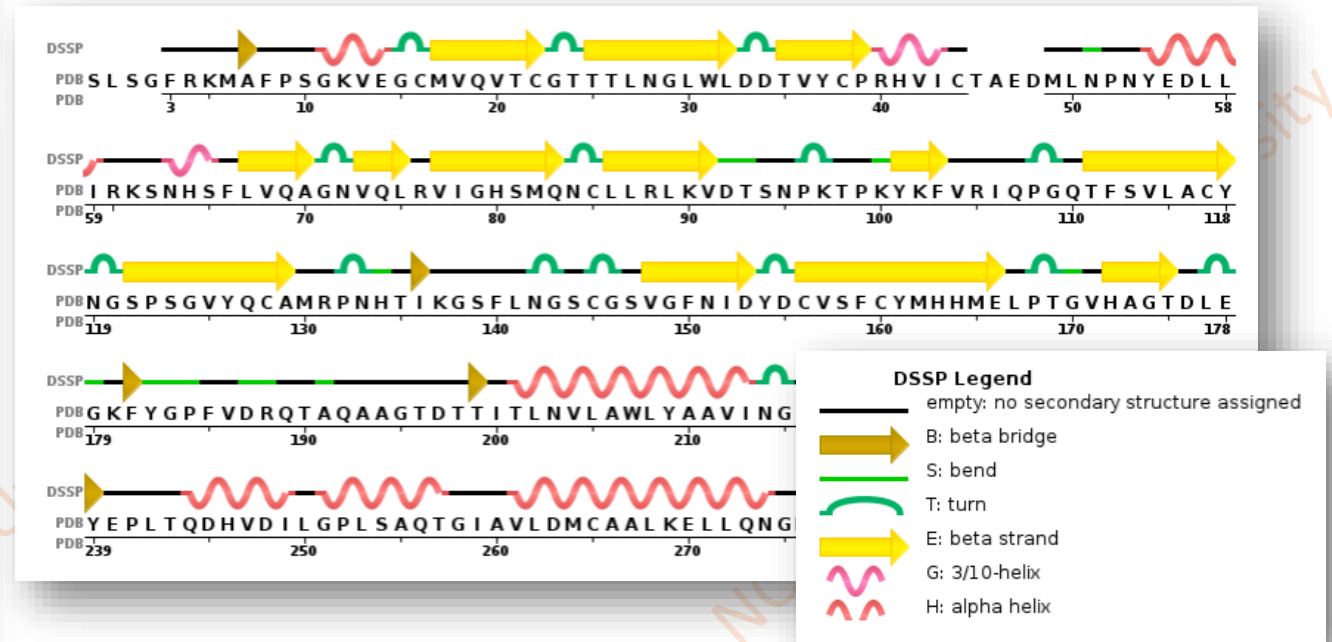
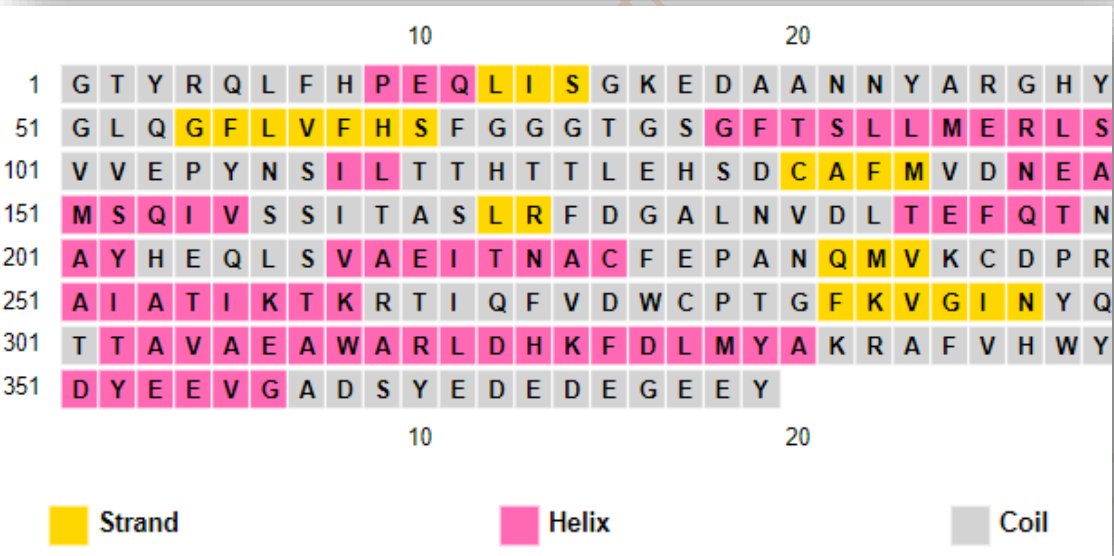
- Úseky které nespádají do kategorií helix nebo list
- **Kombinace** povolených torzních úhlů
- **Nestabilní** konformace
- **Nestandardní** konformace (glycin, prolin)
- **Otáčky** (turns), „náhodné klubko“ (random coil)



Znázornění 2D struktury

- **Písmeny** – H (helix), E (extended sheet), C (coil), ...
- **Barevně** – např. červená (helix), žlutá (skládáný list)
- **Grafickými elementy** – spirála/válec (helix), plochá šipka (skládáný list), linka (ostatní)

MQVWPIEGIKKFETLSYLPPLTVEDLLKQIEYLLRSKWVPCLEFSKVG
 -----EEEE-----HHHHHHHHHHHH-----EEEEEE-----

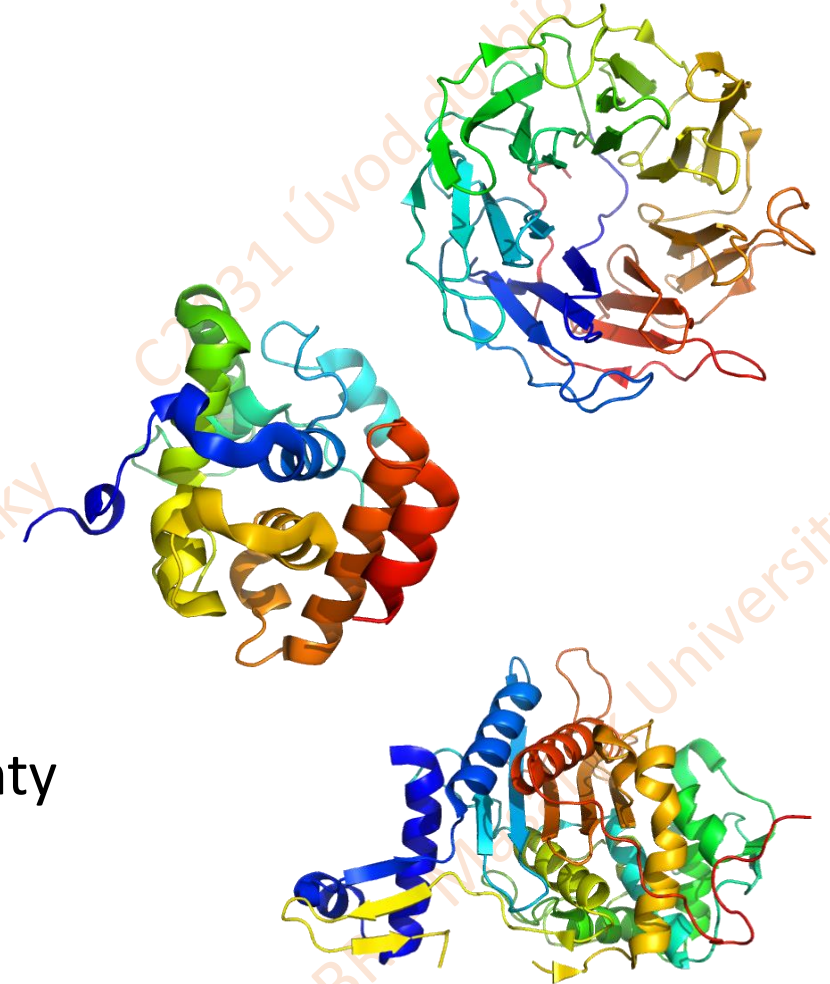


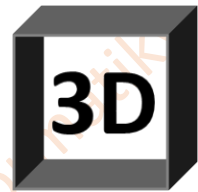
Dělení proteinů dle 2D struktury

Zejména pro účely klasifikace, hledání společných rysů

Každý protein obsahuje mj. smyčky a ohyby

- Jen α struktury
- Jen β struktury
- α/β – Motivy kombinující α i β struktury
- $\alpha + \beta$ – Oddělené domény tvořené jen α nebo jen β strukturami
- **Malé proteiny** – speciální případy, např. obsahující ionty kovů, stabilizované disulfidickými můstky





Terciární struktura

Konkrétní umístění jednotlivých atomů polypeptidového řetězce v prostoru

Stabilizována pomocí různých typů vazeb:

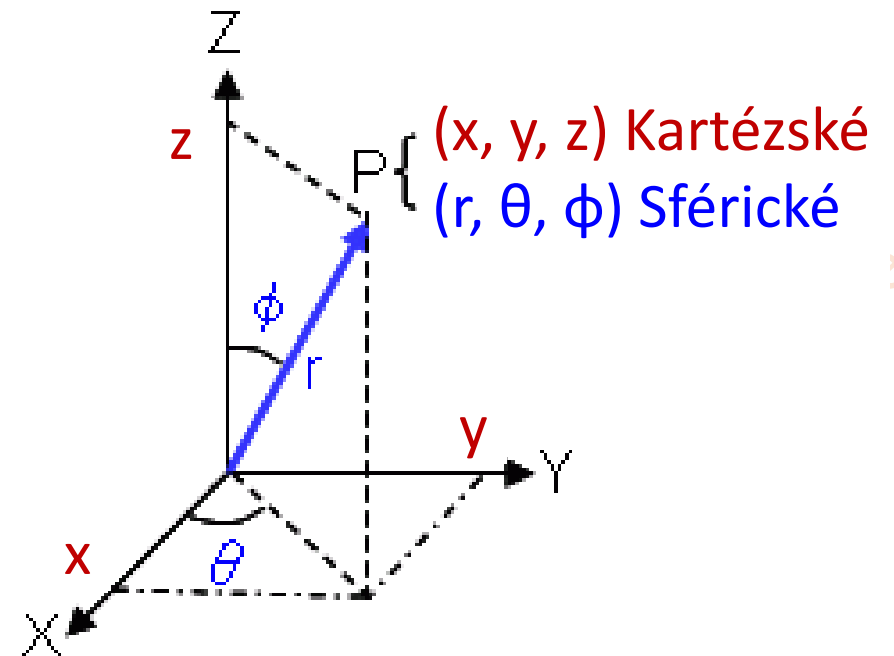
- **Vodíková vazba** (H-můstek)
mezi polárními AK, mezi N-H a C=O hlavního řetězce
- **Iontová** interakce – nabité AK
- **Hydrofobní** interakce – nepolární AK
- „**Stacking**“ (π - π , CH- π interakce) – aromatické AK
- Kovalentní vazba **síra-síra** – cystein / cystin
- Vazba **iontů kovů**

Absolutní souřadnice

Vztažené k definovanému počátku soustavy souřadnic $[0, 0, 0]$

- **Kartézské souřadnice** – x, y, z
- **Sférické souřadnice** – r, θ, ϕ nebo ρ, θ, ϕ

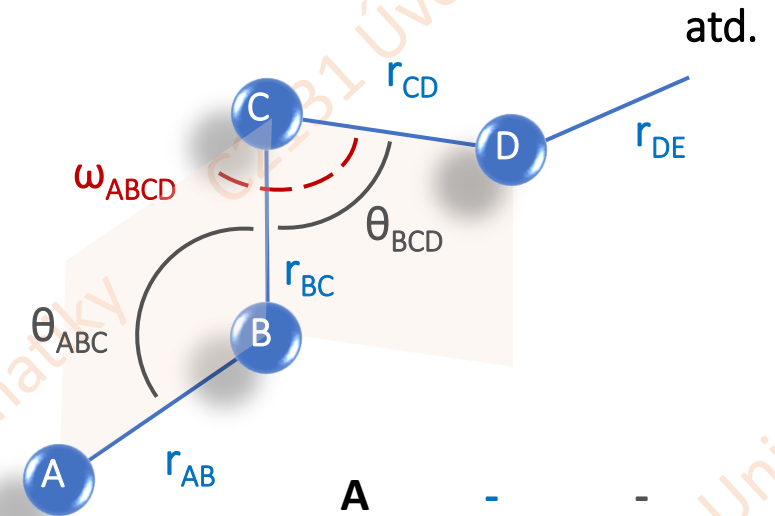
Pro N atomů $\rightarrow 3N$ souřadnic



Relativní souřadnice

Vztažené k předchozímu definovanému bodu (atomu)

- **Vzdálenost** od předchozího atomu
- **Úhel** mezi třemi atomy
- **Torzní úhel** mezi čtyřmi atomy



A	-	-	-
B	r_{AB}	-	-
C	r_{BC}	θ_{ABC}	-
D	r_{CD}	θ_{BCD}	ω_{ABCD}
E	r_{DE}	θ_{CDE}	ω_{BCDE}
...			

Pro N atomů $\rightarrow 3N - 6$ souřadnic



Od 2D ke 3D

Motivy

- 2-3 prvky sekundární struktury

Foldy

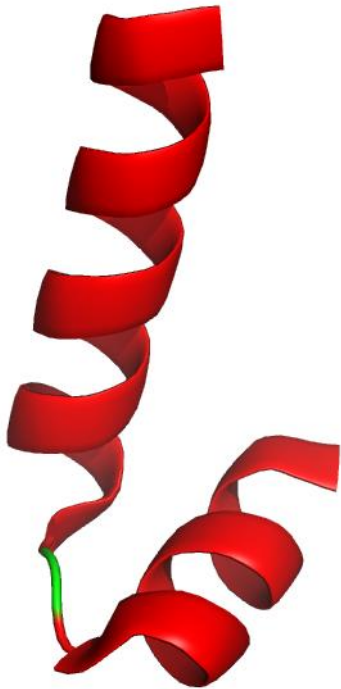
- Kombinace jednoduchých motivů

Domény

- Jsou tvořeny jedním nebo několika motivy/foldy
- Část proteinu s vlastní funkcí (nejmenší funkční jednotka)
- Nezávislá jednotka (alespoň částečně nezávislá)

Jednoduché motivy

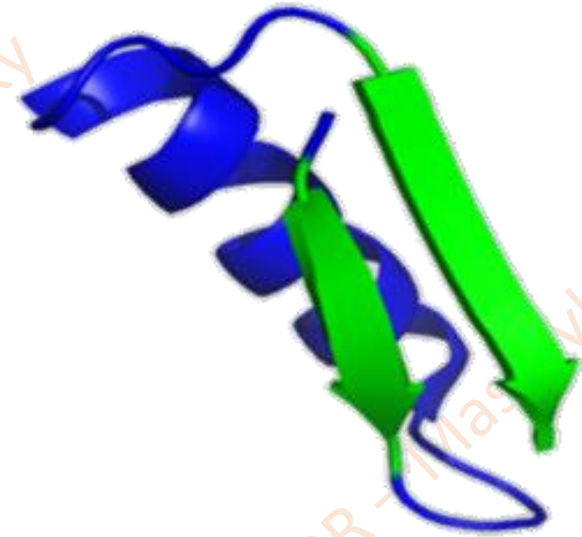
Helix-otáčka-helix



β -vlásenka

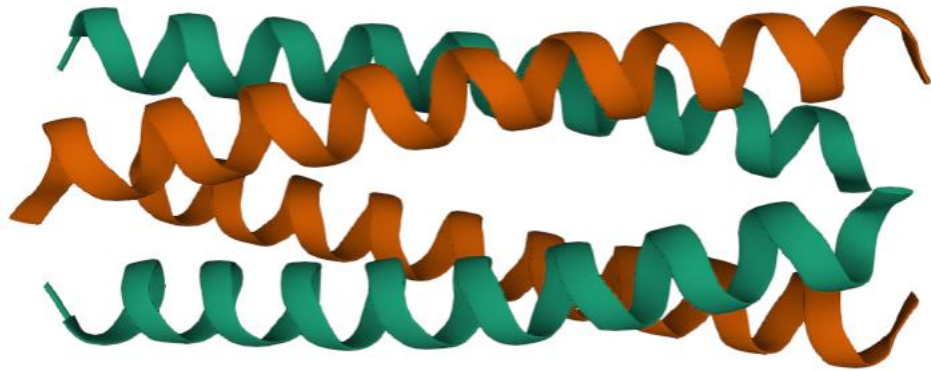


β - α - β

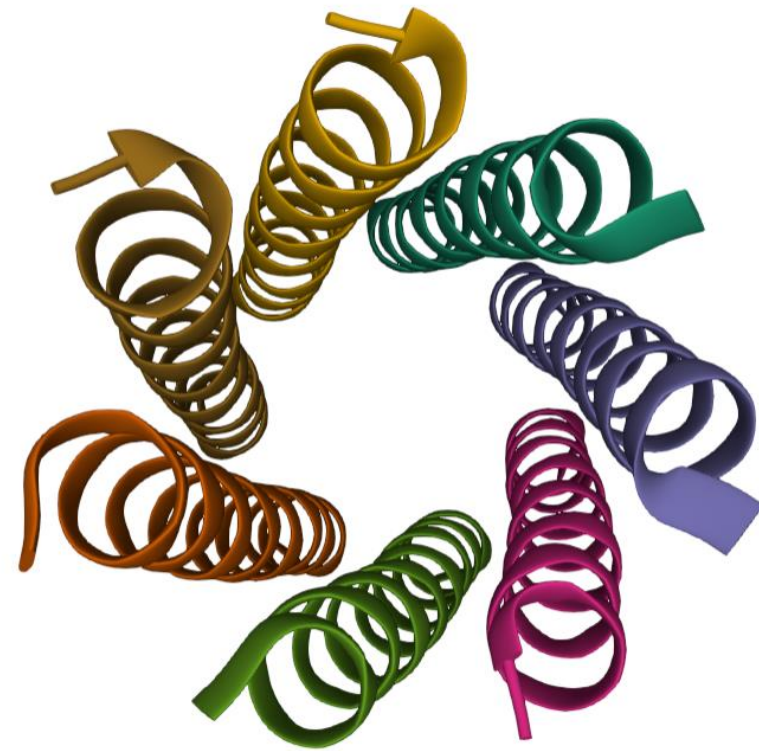


Složené α -motivy/foldy

4-helix bundle

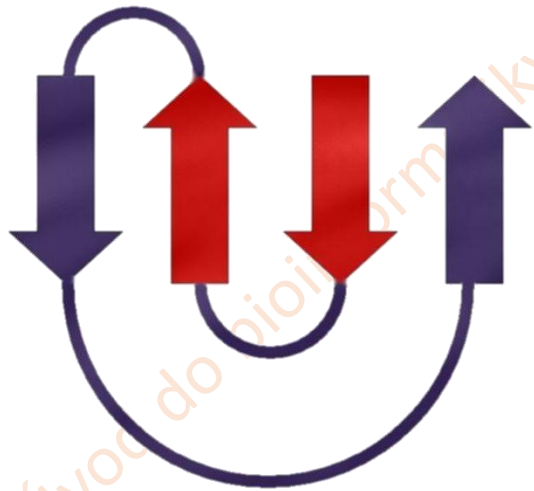


7-helix barrel



Složené β -motivy/foldy

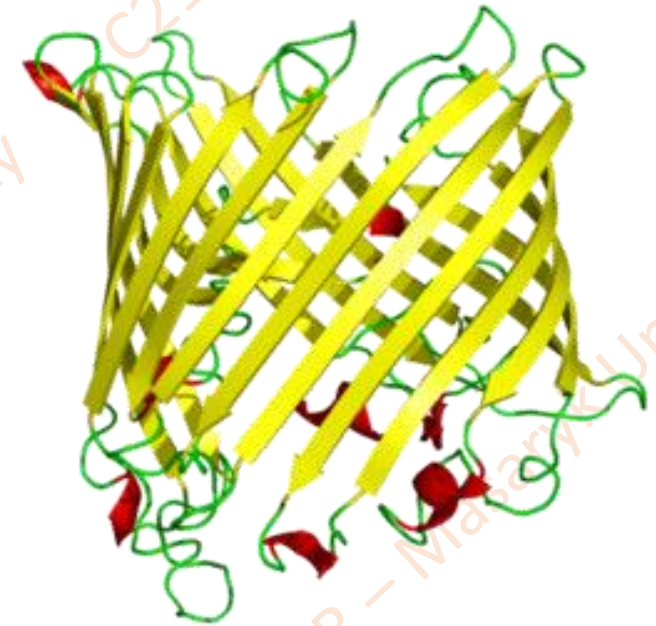
Řecký klíč



β -meandr



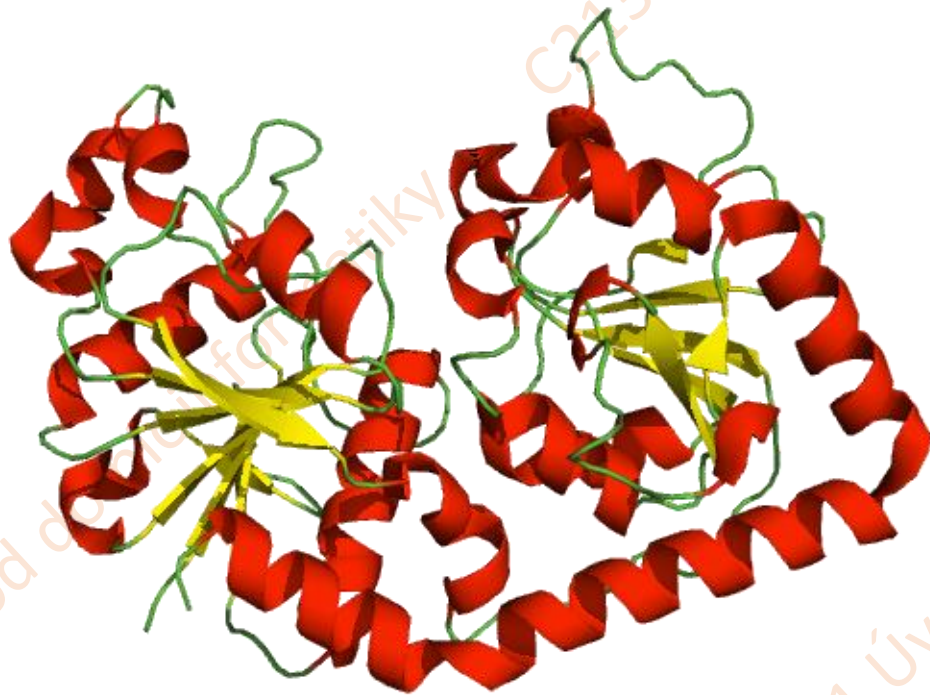
β -barel



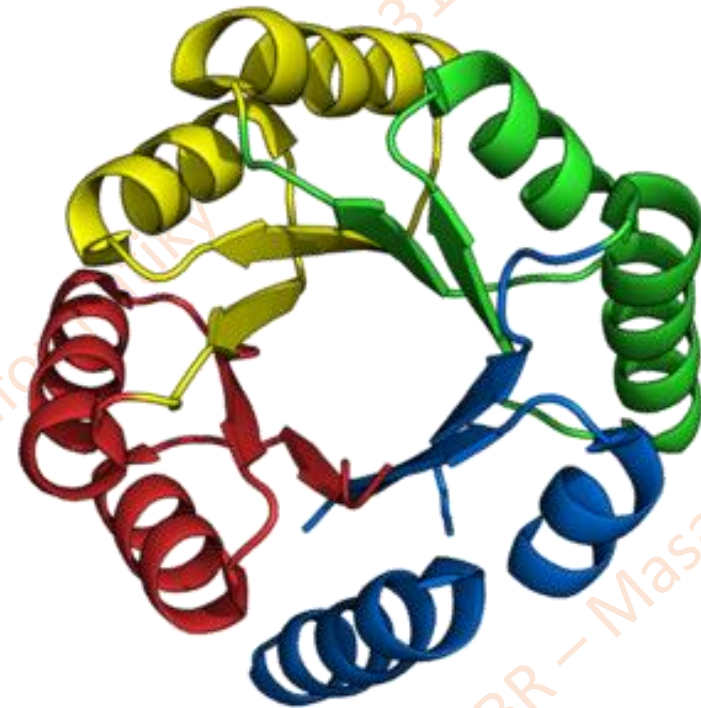


Složené α/β -motivy/foldy

Rossmannův fold

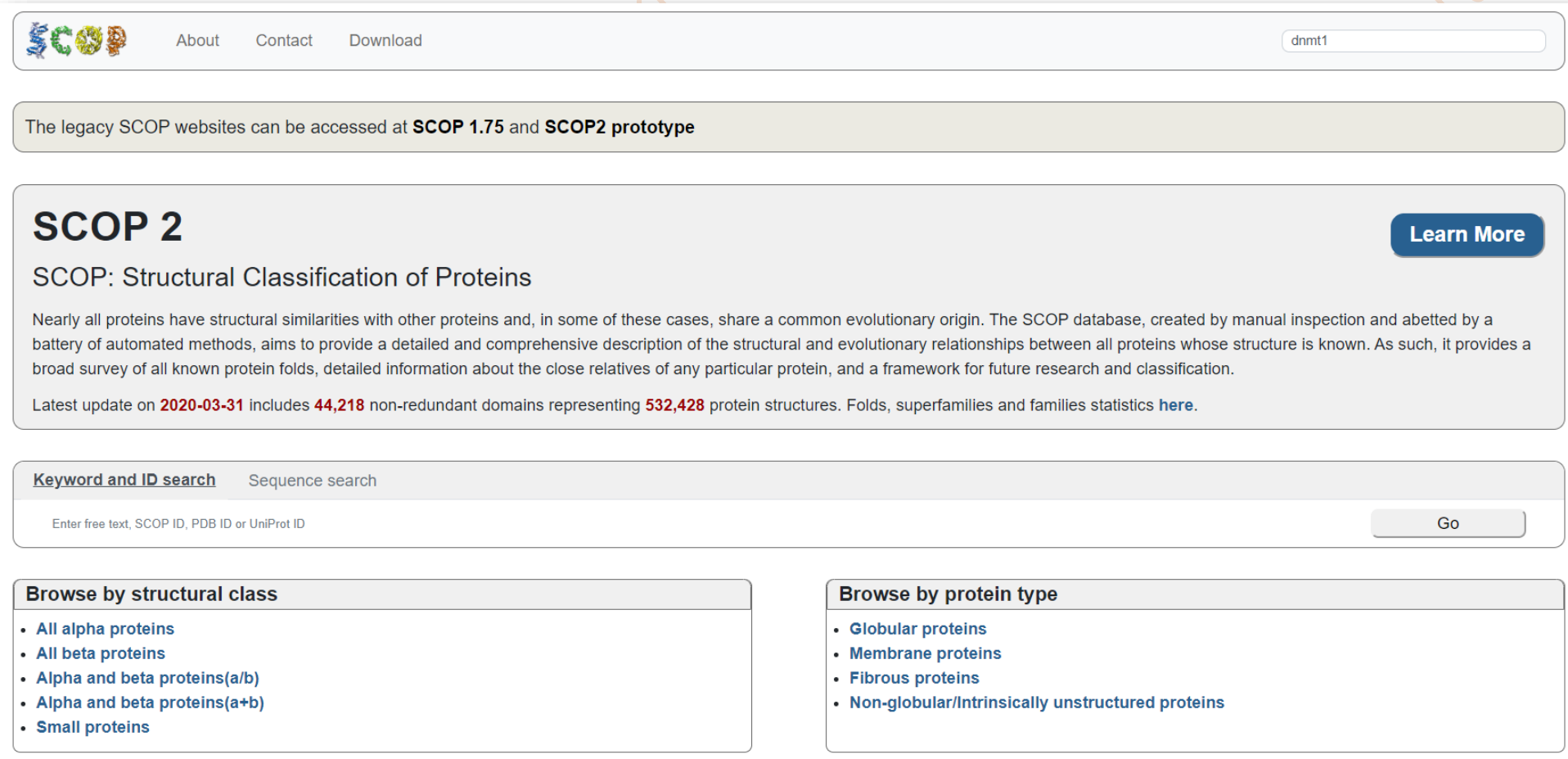


TIM-barel



Structural classification of proteins (SCOP)

<http://scop.mrc-lmb.cam.ac.uk/>



The screenshot shows the SCOP website interface. At the top, there is a navigation bar with the SCOP logo, links for 'About', 'Contact', and 'Download', and a search box containing 'dnmt1'. Below this is a banner for legacy websites. The main content area features the 'SCOP 2' title, a 'Learn More' button, and a descriptive paragraph about the database. A search bar is located below the text, with 'Keyword and ID search' selected. At the bottom, there are two columns of links for browsing by structural class and protein type.

SCOP

About Contact Download

dnmt1

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

SCOP 2

[Learn More](#)

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 **2020-03-31** includes **44,218** non-redundant domains representing **532,428** protein structures. Folds, superfamilies and families statistics [here](#).

Keyword and ID search Sequence search

Enter free text, SCOP ID, PDB ID or UniProt ID

Go

Browse by structural class

- [All alpha proteins](#)
- [All beta proteins](#)
- [Alpha and beta proteins\(a/b\)](#)
- [Alpha and beta proteins\(a+b\)](#)
- [Small proteins](#)

Browse by protein type

- [Globular proteins](#)
- [Membrane proteins](#)
- [Fibrous proteins](#)
- [Non-globular/Intrinsically unstructured proteins](#)

CATH – Protein structure classification database

Domény jsou klasifikovány podle **CATH** hierarchie

<https://www.cathdb.info/>

- **Třída (Class)**
 - Podle sekundární struktury
 - Jen α , jen β , α i β , minimum sekundární struktury
- **Architektura**
 - 3D uspořádání sekundární struktury
- **Topologie/fold**
 - Jak jsou prvky sekundární struktury uspořádané za sebou
- **Homologní nadrodina**
 - V případě, že jsou domény evolučně příbuzné (homologní proteiny)

The screenshot displays the CATH database search results for a query. It is organized into three sections:

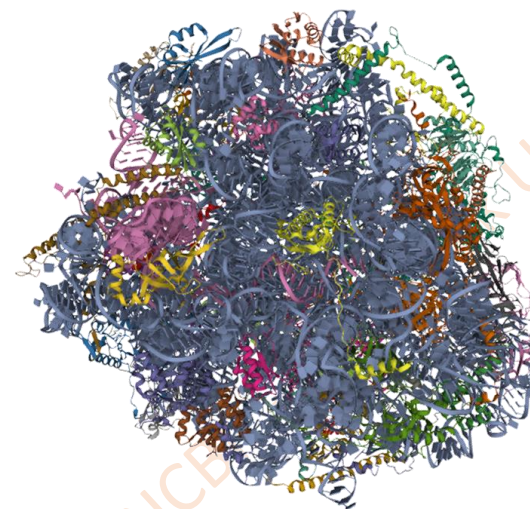
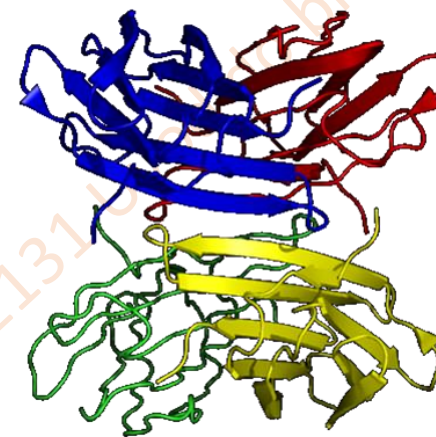
- Matching CATH Superfamilies:** Shows 1 result with CATH code 2.120.10.70, identified as 'Fucose-specific lectin'. It includes a 3D ribbon diagram of the protein structure.
- Matching CATH Domains:** Shows 4 results, with the first being '4agiA00' (PDB code 4agi, chain A, domain 00). It includes a 3D ribbon diagram of the domain structure.
- Matching PDB Structures:** Shows 1 result with PDB code 4agi, identified as '4agi'. It includes a 3D ribbon diagram of the full protein structure.

Each section includes a 'View all entries' button and an 'Info' icon.

Kvartérní struktura

4D

- Vzájemná **kombinace více řetězců** (monomerů)
- Podle typu podjednotek:
 - **Homooligomery** (identické jednotky)
 - **Heterooligomery** (alespoň dva různé typy jednotek)
- **Komplexy** proteinů s dalšími makromolekulami
 - Ribozom, proteazom, replikační komplex,...
- **Nadmolekulární komplexy**
 - Virové částice, buněčná membrána, organely,...



NCBR – Masaryk University

Strukturní data (3D)

C2131 Úvod do bioinformatiky

C2131 Úvod do bioinformatiky

C2131 Úvod do bioinformatiky

C2131 Úvod do bioinformatiky

NCBR – Masaryk University

Způsob uložení 3D (4D) strukturních dat



➤ Veřejně dostupné **databáze**

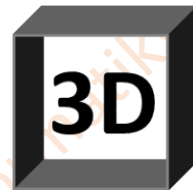
- **Protein Data Bank** (PDB), Biological Magnetic Resonance Data Bank, EMDataBank

➤ Několik typů dat:

- **Koordináty** atomů
- **Experimentální** data
- Doplňkové informace (**meta data**)

➤ Definovaný **formát**

- PDB
- **mmCIF**



Formát PDB

- Stále častý, dnes již zastaralý
- Fixní pozice sloupců, kapacitní omezení

```
HEADER      HYDROLASE                      14-MAR-03  1HL8
TITLE      CRYSTAL STRUCTURE OF THERMOTOGA MARITIMA ALPHA-FUCOSIDASE
COMPND     MOL_ID: 1;
COMPND     2 MOLECULE: PUTATIVE ALPHA-L-FUCOSIDASE;
COMPND     3 CHAIN: A, B;
COMPND     4 EC: 3.2.1.51;
COMPND     5 ENGINEERED: YES;
COMPND     6 OTHER_DETAILS: ORF TH0306
SOURCE     MOL_ID: 1;
SOURCE     2 ORGANISM_SCIENTIFIC: THERMOTOGA MARITIMA;
SOURCE     3 ORGANISM_TAXID: 243274;
SOURCE     4 STRAIN: MSB8;
SOURCE     5 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE     6 EXPRESSION_SYSTEM_TAXID: 511693;
SOURCE     7 EXPRESSION_SYSTEM_STRAIN: BL21;
SOURCE     8 EXPRESSION_SYSTEM_VECTOR: PDEST17
KEYWDS     HYDROLASE, GLYCOSIDE HYDROLASE, ALPHA-L-FUCOSIDASE, THERMOSTABLE
EXPDTA     X-RAY DIFFRACTION
AUTHOR     G.SULZENBACHER,C.BIGNON,V.BOURNE,B.HENRISSAT
REVDAT     5 13-JUL-11 1HL8 1 UERSN
REVDAT     4 24-FEB-09 1HL8 1 UERSN
REVDAT     3 25-MAR-04 1HL8 1 JRNL
REVDAT     2 13-FEB-04 1HL8 1 REMARK
REVDAT     1 15-JAN-04 1HL8 0
JRNL       AUTH G.SULZENBACHER,C.BIGNON,T.NISHIMURA,C.A.TARLING,S.G.WITHERS,
JRNL       AUTH 2 B.HENRISSAT,V.BOURNE
JRNL       TITL CRYSTAL STRUCTURE OF THERMOTOGA MARITIMA ALPHA-L-
JRNL       TITL 2 FUCOSIDASE. INSIGHTS INTO THE CATALYTIC MECHANISM AND THE
JRNL       TITL 3 MOLECULAR BASIS FOR FUCOSIDOSIS.
JRNL       REF J.BIOL.CHEM. U. 279 13119 2004
JRNL       REFN ISSN 0021-9258
JRNL       PHID 14715651
JRNL       DOI 10.1074/JBC.M313783200
REMARK     2
REMARK     2 RESOLUTION. 2.4 ANGSTROMS.
REMARK     3
REMARK     3 REFINEMENT.
REMARK     3 PROGRAM : REFMAC 5.1.24
REMARK     3 AUTHORS : MURSHUDDU,VAGIN,DODSON
REMARK     3
REMARK     3 REFINEMENT TARGET : MAXIMUM LIKELIHOOD
REMARK     3
REMARK     3 DATA USED IN REFINEMENT.
REMARK     3 RESOLUTION RANGE HIGH (ANGSTROMS) : 2.40
REMARK     3 RESOLUTION RANGE LOW (ANGSTROMS) : 37.27
```

Řetězec Číslo rezidua

Typ aminokyseliny Souřadnice

ATOM	1	N	ARG	A	7	-26.699	-11.392	48.842	1.00	56.84	N
ATOM	2	CA	ARG	A	7	-25.554	-10.912	49.663	1.00	55.29	C
ATOM	3	C	ARG	A	7	-24.623	-9.995	48.864	1.00	53.39	C
ATOM	4	O	ARG	A	7	-24.414	-10.191	47.661	1.00	54.25	O
ATOM	5	CB	ARG	A	7	-24.761	-12.105	50.193	1.00	56.00	C
ATOM	6	CG	ARG	A	7	-25.374	-12.749	51.426	1.00	58.45	C
ATOM	7	CD	ARG	A	7	-24.396	-12.945	52.578	1.00	59.72	C
ATOM	8	NE	ARG	A	7	-25.048	-12.736	53.869	1.00	61.30	N
ATOM	9	CZ	ARG	A	7	-24.413	-12.499	55.014	1.00	61.72	C
ATOM	10	NH1	ARG	A	7	-23.087	-12.440	55.065	1.00	61.05	N
ATOM	11	NH2	ARG	A	7	-25.115	-12.320	56.126	1.00	63.61	N
ATOM	12	N	TYR	A	8	-24.055	-9.007	49.545	1.00	50.83	N
ATOM	13	CA	TYR	A	8	-23.096	-8.100	48.940	1.00	48.87	C
ATOM	14	C	TYR	A	8	-21.680	-8.609	49.201	1.00	47.84	C
ATOM	15	O	TYR	A	8	-21.378	-9.123	50.279	1.00	47.98	O
ATOM	16	CB	TYR	A	8	-23.287	-6.680	49.481	1.00	47.56	C
ATOM	17	CG	TYR	A	8	-24.700	-6.147	49.294	1.00	48.37	C
ATOM	18	CD1	TYR	A	8	-25.123	-5.630	48.067	1.00	49.00	C
ATOM	19	CD2	TYR	A	8	-25.619	-6.180	50.332	1.00	48.91	C
ATOM	20	CE1	TYR	A	8	-26.419	-5.156	47.889	1.00	48.83	C
ATOM	21	CE2	TYR	A	8	-26.918	-5.707	50.160	1.00	50.24	C
ATOM	22	CZ	TYR	A	8	-27.306	-5.192	48.936	1.00	49.98	C
ATOM	23	OH	TYR	A	8	-28.589	-4.719	48.773	1.00	51.15	O
ATOM	24	N	LYS	A	9	-20.837	-8.493	48.178	1.00	46.89	N

Formát mmCIF



- Novější, preferovaný
- Bezkontextová gramatika, možnost rozšiřování o další typy údajů

```
data_4I1S
#
_entry.id 4I1S
#
_audit_conform.dict_name mmcif_pdbx.dic
_audit_conform.dict_version 4.027
_audit_conform.dict_location http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic
#
loop_
_database_2.database_id
_database_2.database_code
PDB 4I1S
RCSB RCSB076195
#
_database_PDB_rev.num 1
_database_PDB_rev.date 2013-01-30
_database_PDB_rev.date_original 2012-11-21
_database_PDB_rev.status ?
_database_PDB_rev.replaces 4I1S
_database_PDB_rev.mod_type 0
#
_pdbx_database_status.status_code REL
_pdbx_database_status.entry_id 4I1S
_pdbx_database_status.deposit_site RCSB
_pdbx_database_status.process_site RCSB
_pdbx_database_status.status_code_sf REL
_pdbx_database_status.status_code_mr ?
_pdbx_database_status.SG_entry ?
_pdbx_database_status.status_code_cs ?
_pdbx_database_status.methods_development_category ?
#
loop_
_audit_author.name
_audit_author.pdbx_ordinal
'Motz, C.' 1
'Witte, G.' 2
'Hopfner, K.P.' 3
#
_citation.id primary
_citation.title
'Paramyxovirus V Proteins Disrupt the Fold of the RNA Sensor MDA5 to Inhibit Antiviral Science
_citation.journal_abbrev
```

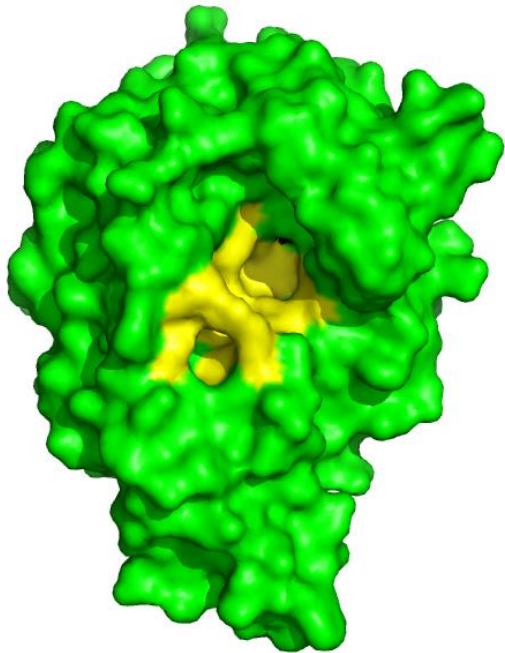
			Řetězec	Číslo rezidua	Souřadnice															
	Typ aminokyseliny																			
ATOM	1	N	N	. GLU A 1 1	? 7.254	11.020	4.888	1.00	61.38	?	?	?	?	?	?	546	GLU	A	N	1
ATOM	2	C	CA	. GLU A 1 1	? 6.404	12.200	5.071	1.00	67.04	?	?	?	?	?	?	546	GLU	A	CA	1
ATOM	3	C	C	. GLU A 1 1	? 7.111	13.526	4.729	1.00	59.60	?	?	?	?	?	?	546	GLU	A	C	1
ATOM	4	O	O	. GLU A 1 1	? 6.576	14.360	3.999	1.00	64.05	?	?	?	?	?	?	546	GLU	A	O	1
ATOM	5	C	CB	. GLU A 1 1	? 5.842	12.232	6.500	1.00	74.02	?	?	?	?	?	?	546	GLU	A	CB	1
ATOM	6	C	CG	. GLU A 1 1	? 5.625	13.627	7.094	1.00	74.52	?	?	?	?	?	?	546	GLU	A	CG	1
ATOM	7	C	CD	. GLU A 1 1	? 4.448	14.369	6.495	1.00	78.40	?	?	?	?	?	?	546	GLU	A	CD	1
ATOM	8	O	OE1	. GLU A 1 1	? 3.968	13.977	5.409	1.00	81.00	?	?	?	?	?	?	546	GLU	A	OE1	1
ATOM	9	O	OE2	. GLU A 1 1	? 3.997	15.354	7.118	1.00	79.97	?	?	?	?	?	?	546	GLU	A	OE2	1
ATOM	10	N	N	. ASP A 1 2	? 8.299	13.714	5.287	1.00	44.26	?	?	?	?	?	?	547	ASP	A	N	1
ATOM	11	C	CA	. ASP A 1 2	? 9.213	14.768	4.873	1.00	34.80	?	?	?	?	?	?	547	ASP	A	CA	1
ATOM	12	C	C	. ASP A 1 2	? 10.508	14.039	4.527	1.00	30.06	?	?	?	?	?	?	547	ASP	A	C	1
ATOM	13	O	O	. ASP A 1 2	? 11.245	13.650	5.424	1.00	29.92	?	?	?	?	?	?	547	ASP	A	O	1
ATOM	14	C	CB	. ASP A 1 2	? 9.460	15.735	6.039	1.00	34.15	?	?	?	?	?	?	547	ASP	A	CB	1
ATOM	15	C	CG	. ASP A 1 2	? 10.399	16.909	5.672	1.00	36.09	?	?	?	?	?	?	547	ASP	A	CG	1
ATOM	16	O	OD1	. ASP A 1 2	? 11.138	16.835	4.665	1.00	33.05	?	?	?	?	?	?	547	ASP	A	OD1	1
ATOM	17	O	OD2	. ASP A 1 2	? 10.397	17.917	6.418	1.00	36.96	?	?	?	?	?	?	547	ASP	A	OD2	1
ATOM	18	N	N	. LEU A 1 3	? 10.778	13.854	3.239	1.00	32.19	?	?	?	?	?	?	548	LEU	A	N	1
ATOM	19	C	CA	. LEU A 1 3	? 11.922	13.061	2.787	1.00	30.81	?	?	?	?	?	?	548	LEU	A	CA	1
ATOM	20	C	C	. LEU A 1 3	? 13.253	13.688	3.155	1.00	27.21	?	?	?	?	?	?	548	LEU	A	C	1

Zobrazení 3D struktury

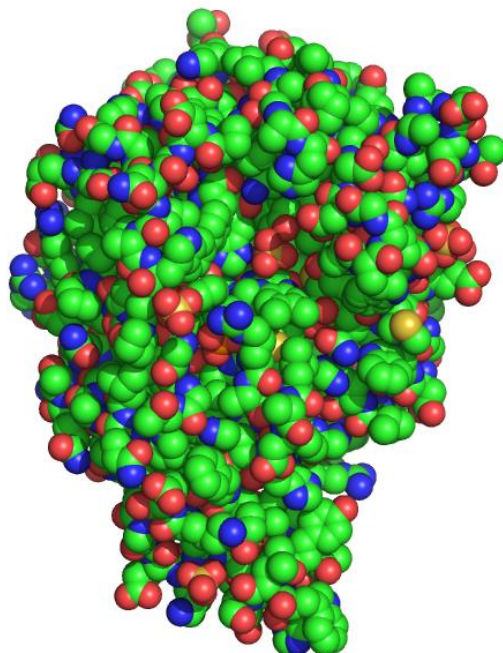
3D

- **Zobrazovací SW:** PyMol, LiteMol, Mol*, Jmol, Chimera, RasMol, VMD, ...
- Konkrétní styl záleží na účelu zobrazení

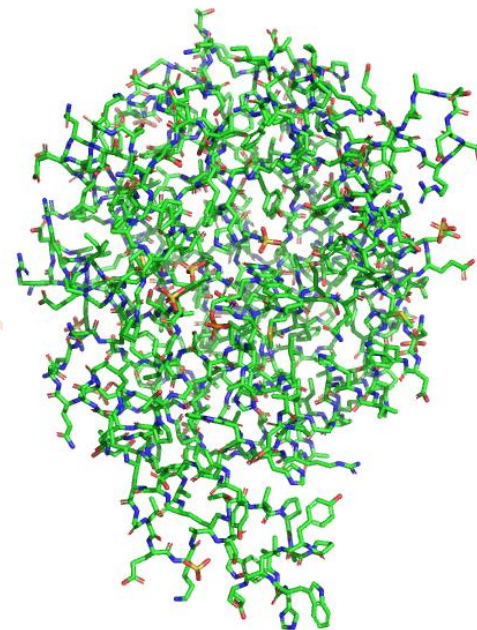
Povrch (surface)



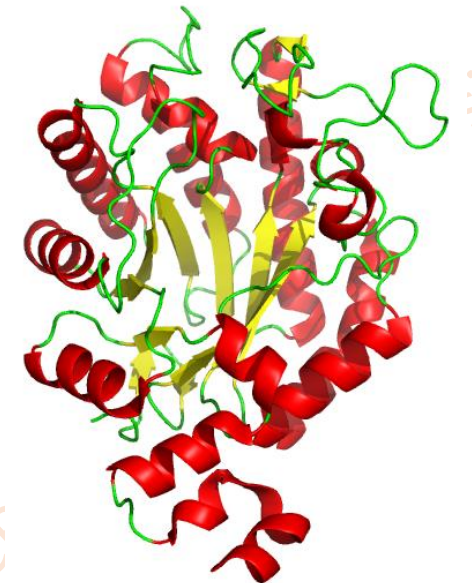
Kuličky (spheres)



Tyčky (sticks, balls and sticks)



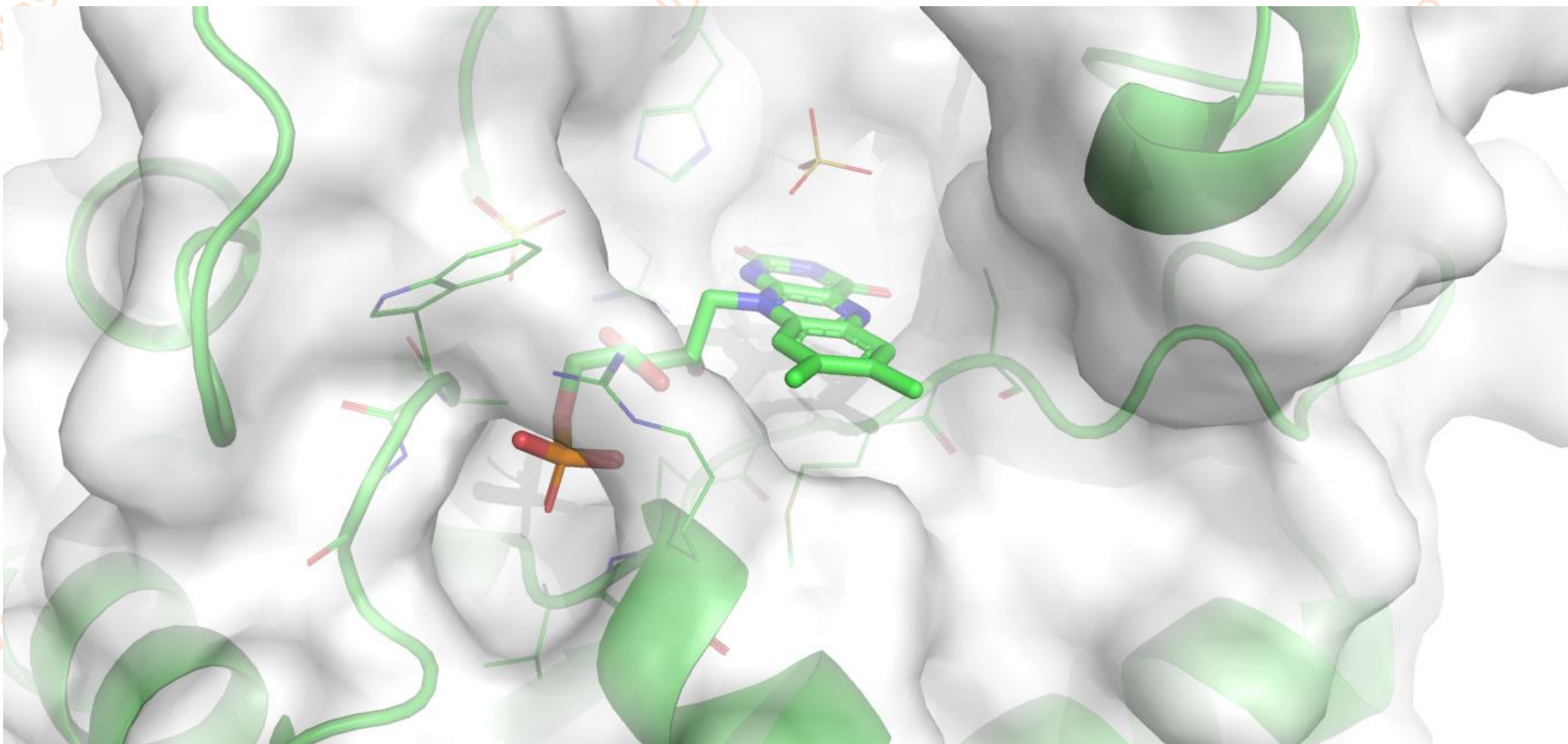
Stuha (cartoon/ribbon)



Zobrazení 3D struktury



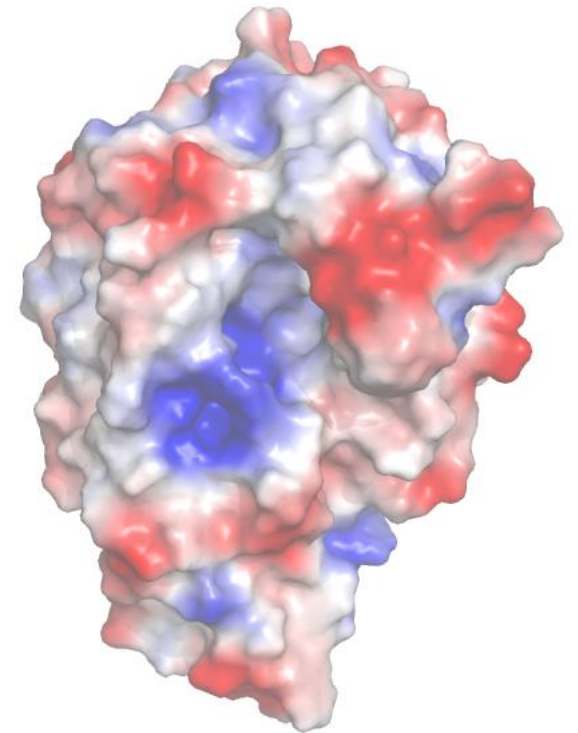
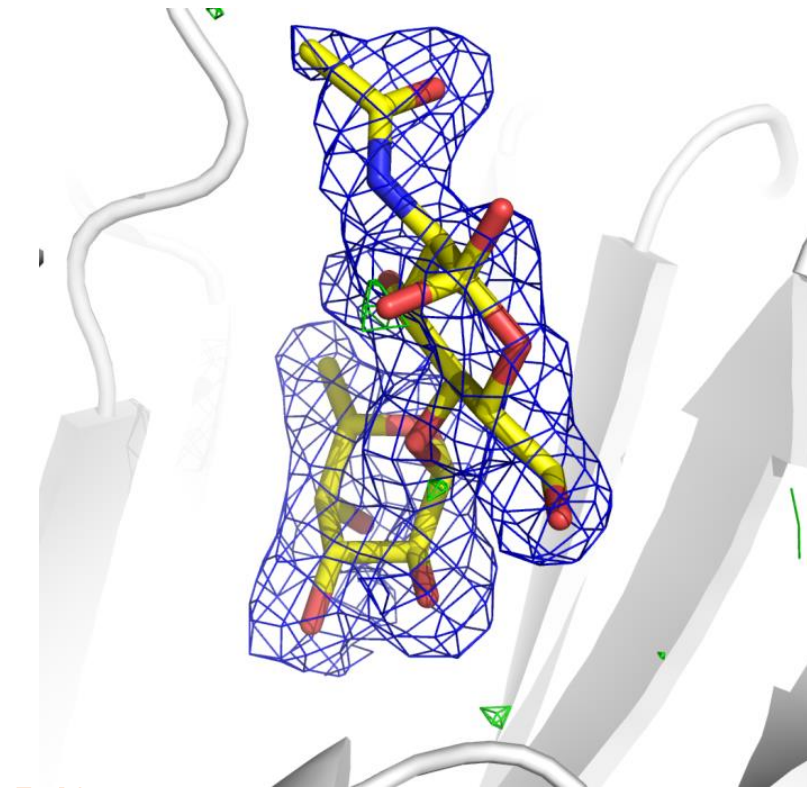
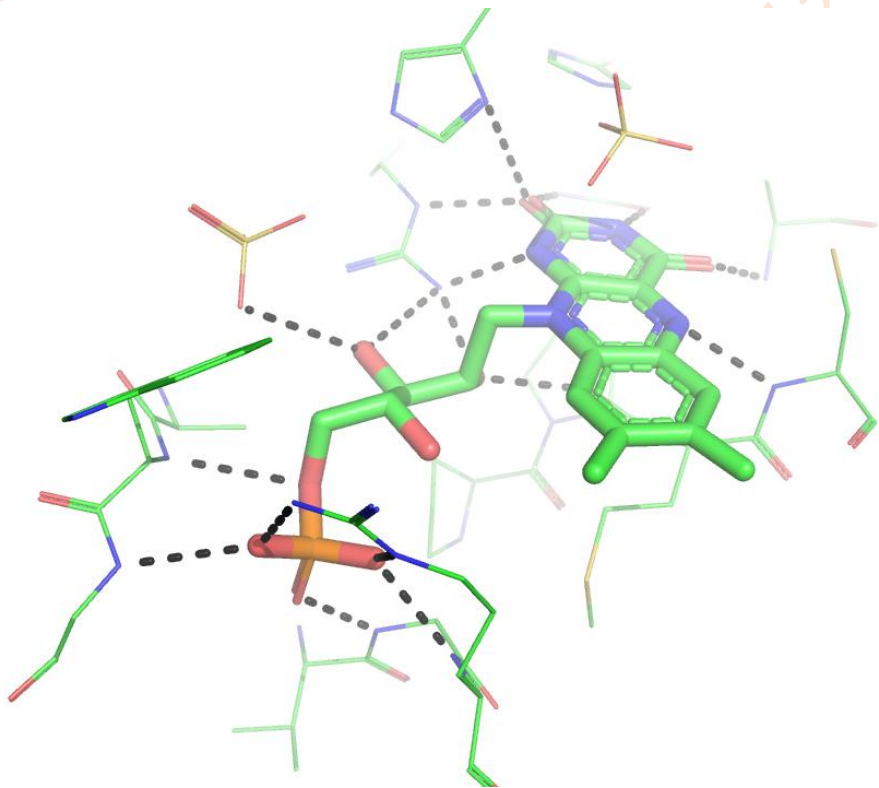
- Časté je kombinované zobrazení



Zobrazení 3D struktury

3D

- Možnost zobrazení dalších informací, např. vodíkové vazby, elektronová hustota, hydrofobicita povrchu



NCBR – Masaryk University

Predikce struktury

C2131 Úvod do bioinformatiky

C2131 Úvod do bioinformatiky

C2131 Úvod do bioinformatiky

C2131 Úvod do bioinformatiky

NCBR – Masaryk University

Predikce struktury

- Predikce struktury znamená přiřazení strukturních atributů jednotlivým aminokyselinám (2D struktura, koordináty – tvorba 3D modelu)
- Struktura 2D a 3D je konzervovaná více než samotná sekvence

➤ Vstupní informace

- **Sekvence**
- Fyzikálně-chemické parametry
- Informace v databázích

➤ Výstup

- **Model struktury** (2D, 3D, 4D)
- Doplnkové informace (např. spolehlivost predikce)

Proč predikovat strukturu?

- **Klasifikace** proteinů
- Vytvoření **modelu** struktury pro další studium
- **Předpověď funkce** proteinu
 - Homologní struktury
 - Vazebná místa
- **Analýza povrchu**
 - Přístupnost pro solvent, tunely, kavity

Predikce sekundární struktury

Predikce 3 základních typů: H (helix), E (β -list), C/– (smyčka/vše ostatní)

➤ 1. GENERACE

- *ab-initio*
- Vycházela z fyzikálně-chemických vlastností a ze statistik pro jednotlivé aminokyseliny

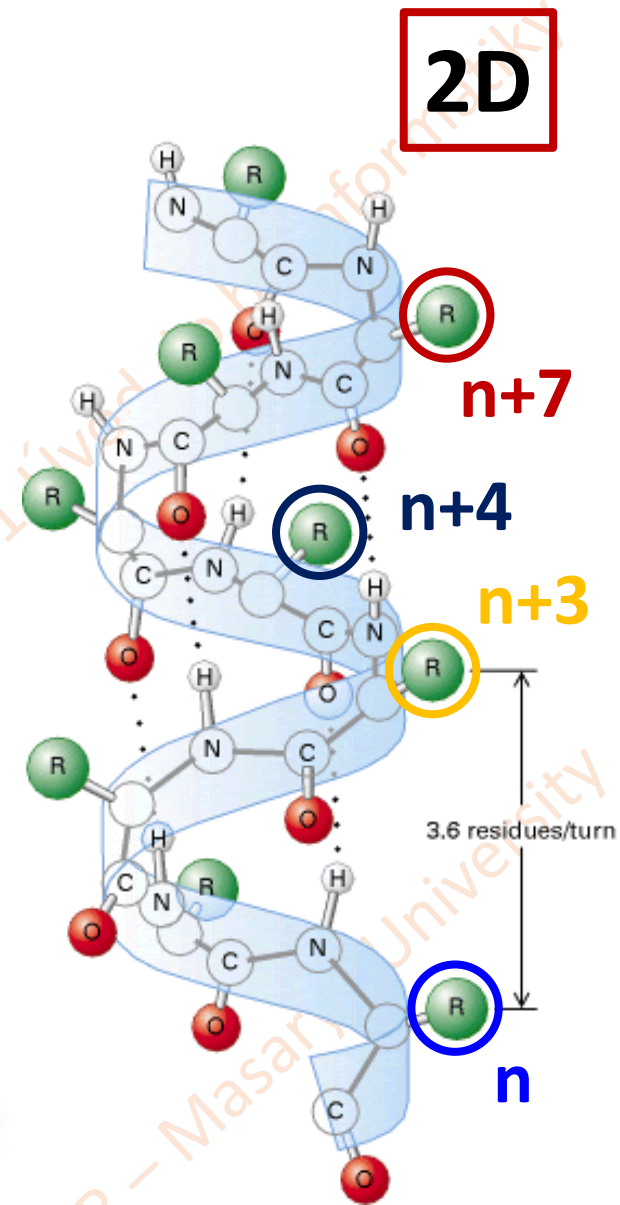
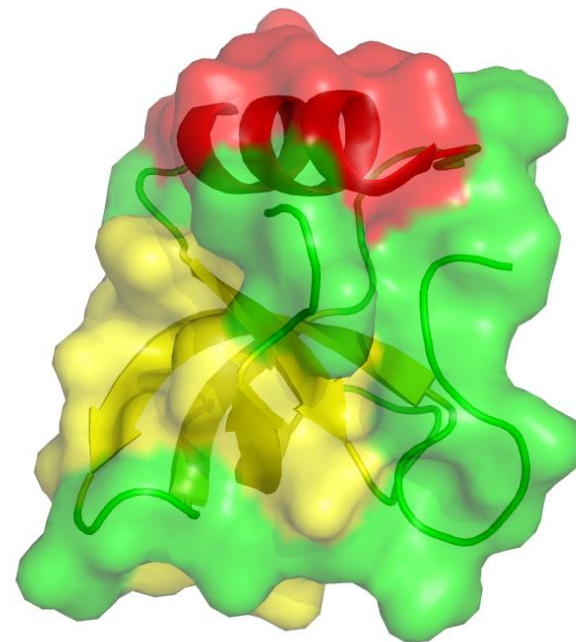
Typické znaky α -šroubovice

Často je částečně exponovaná

- Jedna strana je otočená dovnitř proteinu (hydrofobní) a druhá ven (hydrofilní)
- Residuum (aminokyselina) n , $n+3$, $n+4$, $n+7$ míří na stejnou stranu

Transmembránový helix

- Všechny aminokyseliny hydrofobní

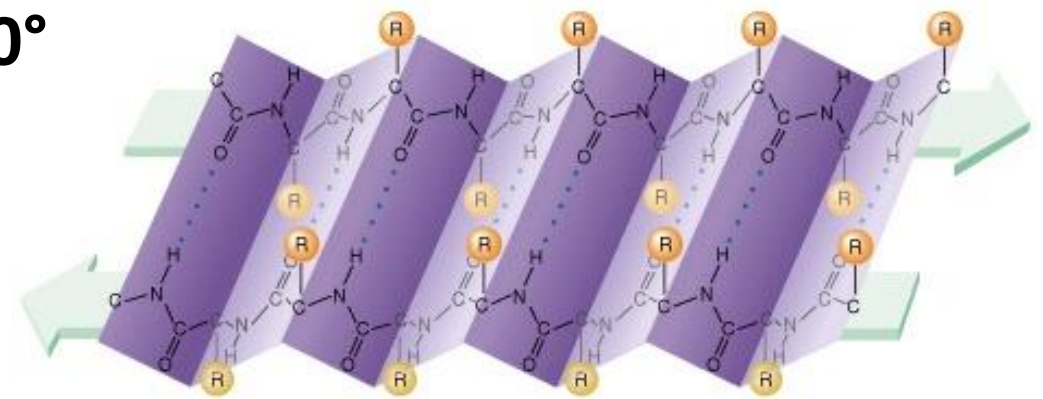


Typické znaky β -listu

Residua (aminokyseliny) se střídají po **180°**

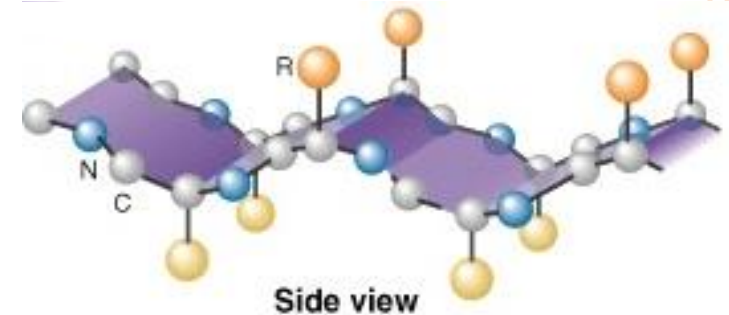
Částečně zanořený list

- Residua n , $n+2$, $n+4$ atd. jsou polární
- Residua $n+1$, $n+3$, $n+5$ atd. jsou nepolární



Úplně zanořený list

- Všechna residua jsou nepolární



α -šroubovice nebo β -list?

?

ELKAHIRVDLTQ α **ELKAHIRVDLTQ****ELKAHIRVDLTQ** β

Polární

Nepolární

α-šroubovice nebo β-list?

?

ELKAHIRVDLTQ

ELKAHIRVDLTQ

✓✓ x x ✓✓ x x x ✓✓ x x

α



ELKAHIRVDLTQ

✓✓✓✓✓✓✓✓✓✓✓✓✓

β



Polární

Nepolární

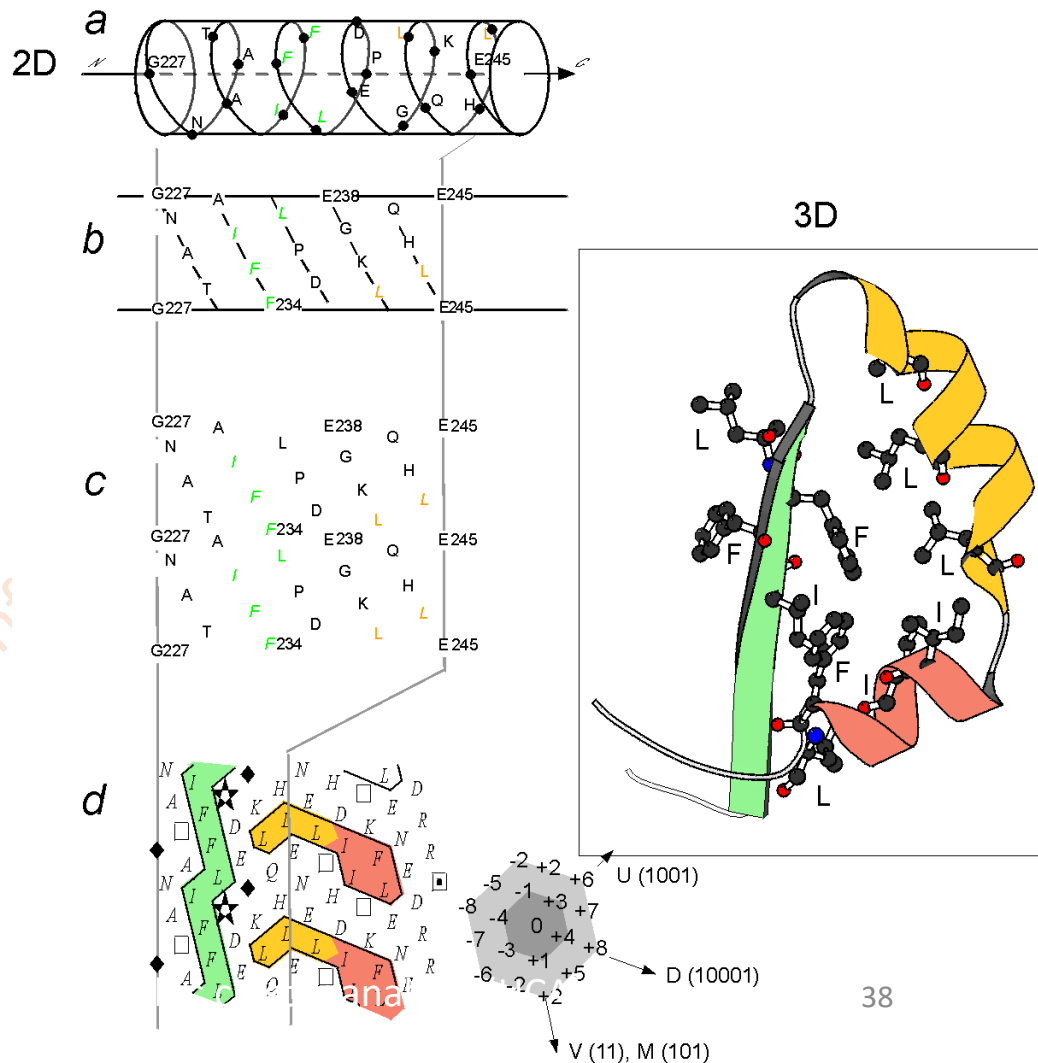
Analýza hydrofobních klastrů (HCA)

Hydrophobic Cluster Analysis 2D support

2D

human $\alpha 1$ antitrypsin
 1D 227...GNATA**IFFL**PDEGK**LQHL**ENE**L**THD**II**TK**FL**ENEDRR...263
 ...◆NA□A**IFFL**★DE◆K**LQHL**ENE**L**□HD**II**□K**FL**ENEDRR...
 ...00000**1111**00000**100**1000**1000**1100**11**000000...

- Sekvence „se namotá“ na válec (α -helix)
- HCA graf je zobrazení válce v rovině
- Hydrofobní aminokyseliny jsou ohraničeny a tvoří specifické tvary pro α -helixy a β -listy



RPBS Web Portal – HCA

<https://mobyli.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py?form=HCA#forms::HCA>

The screenshot displays the RPBS Web Portal interface. The top navigation bar includes the RPBS logo, the text "RPBS Web Portal", and user options: "(guest)", "set email", "sign-in", "sign-out", and "refresh workspace". Below the navigation bar, there is a search bar and a menu with tabs: "Welcome", "Forms", "Data Bookmarks", "Jobs", and "Tutorials". The "Forms" tab is active, showing a sub-tab for "HCA" with a red 'x' icon. The main content area displays "HCA 1.0.2" and "Hydrophobic Cluster Analysis." with buttons for "Run", "Reset", and "Help pages". Below this, there is an "Input Data" section. A large window in the foreground shows a sequence visualization titled "query.data.seq" and "Drawn by Luc Cazard". The visualization displays a protein sequence with residues numbered from 10 to 310. Residues are represented by colored letters (A, C, G, T, N) and symbols (circles, squares, stars) indicating hydrophobic clusters. The sequence is shown in a 3D-like perspective, with residues connected by lines representing the protein backbone.

Predikce sekundární struktury

Predikce 3 základních typů: H (helix), E (β -list), C/- (smyčka/vše ostatní)

➤ 1. GENERACE

- *ab-initio*
- Vycházela z fyzikálně-chemických vlastností a ze statistik pro jednotlivé aminokyseliny

➤ 2. GENERACE

- Zahrnovala i vliv okolních aminokyselin

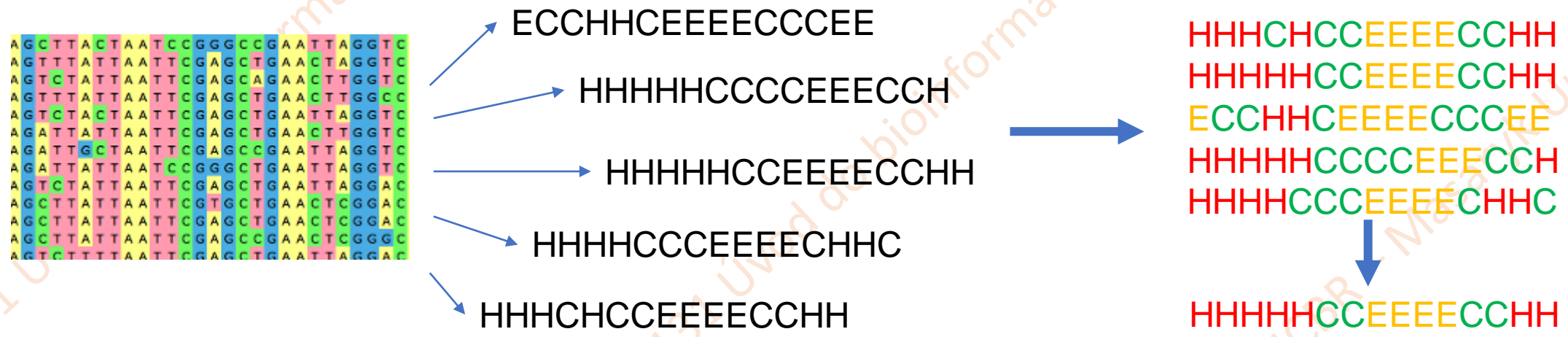
➤ 3. GENERACE

- *Homology-based models*
- Metody strojového učení
- Využívá multiple sequence alignmentu a toho, že 2D struktura je více konzervovaná než sekvence

Metody založené na homologii (*Homology-based*)

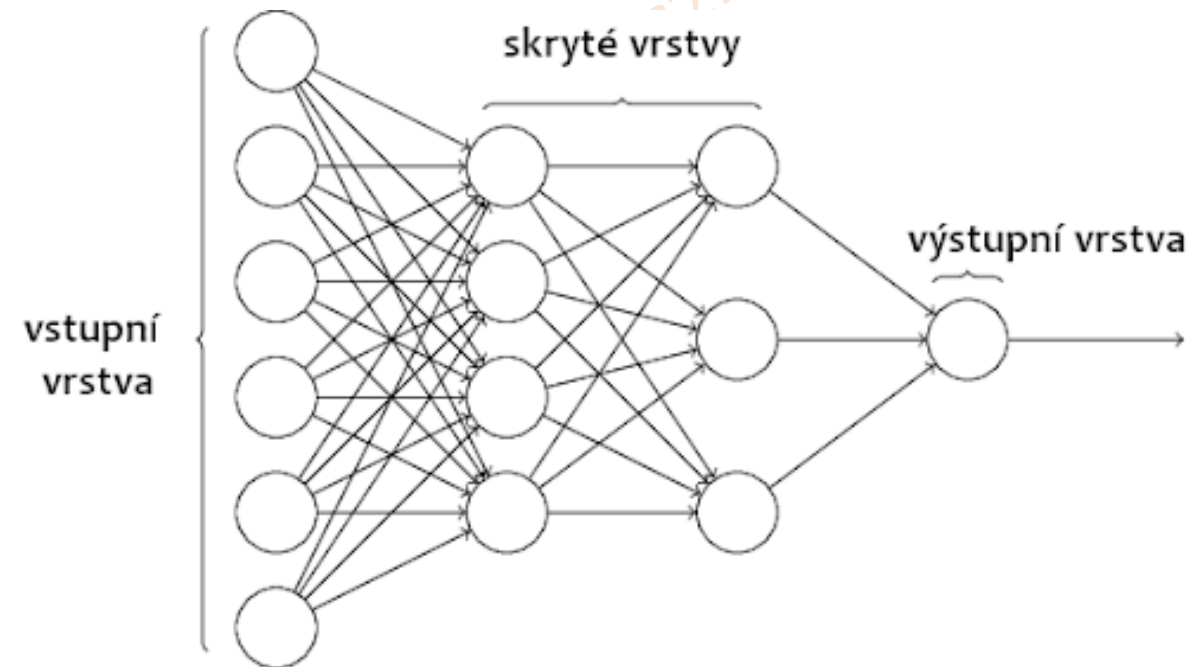
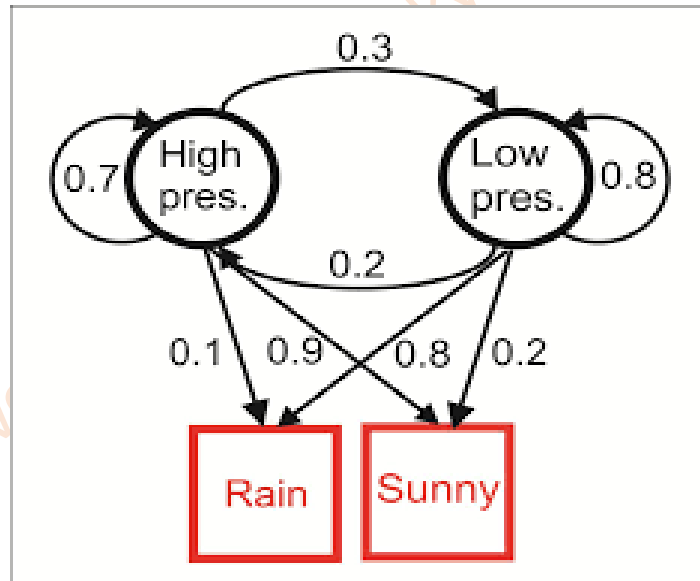
Vychází z předpokladu, že 2D struktura je více konzervovaná než sekvence

1. Multiple sequence alignment
2. Predikce sekundárních struktur pro každou sekvenci zvlášť
3. Porovnání predikovaných sekundárních struktur s alignmentem
4. Konsenzus sekundární struktury



Metody strojového učení (*Machine learning*)

- Model, který je natrénovaný na známé sadě dat
- Neuronové sítě
- Skryté Markovovy modely



PSIPRED

- Predikce sekundární struktury pomocí 2 neuronových sítí
- Časově náročnější
- Ve srovnání s většinou programů na predikci sekundární struktury má lepší výsledky

<http://bioinf.cs.ucl.ac.uk/psipred/>

Choose prediction methods (hover for short description)

Popular Analyses

<input checked="" type="checkbox"/> PSIPRED 4.0 (Predict Secondary Structure)	<input type="checkbox"/> DISOPRED3 (Disopred Prediction)
<input type="checkbox"/> MEMSAT-SVM (Membrane Helix Prediction)	<input type="checkbox"/> pGenTHREADER (Profile Based Fold Recognition)

Contact Analysis

<input type="checkbox"/> DeepMetaPSICOV 1.0 (Structural Contact Prediction)	<input type="checkbox"/> MEMPACK (TM Topology and Helix Packing)
---	--

Fold Recognition

<input type="checkbox"/> GenTHREADER (Rapid Fold Recognition)	<input type="checkbox"/> pDomTHREADER (Protein Domain Fold Recognition)
---	---

Structure Modelling

<input type="checkbox"/> Bioserf 2.0 (Automated Homology Modelling)	<input type="checkbox"/> Domserf 2.1 (Automated Domain Homology Modelling)
<input type="checkbox"/> DMPfold 1.0 Fast Mode (Protein Structure Prediction)	

Domain Prediction

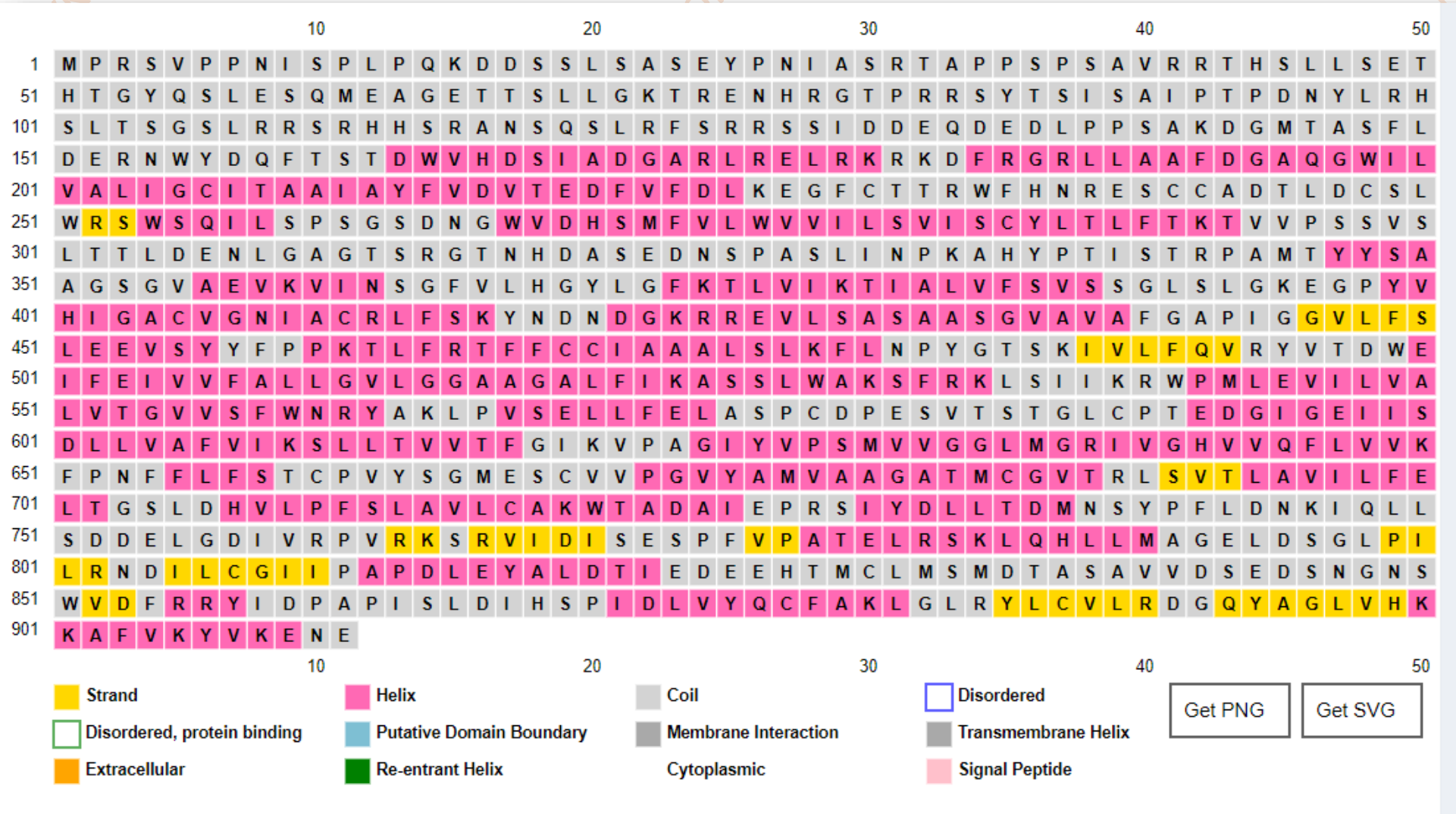
<input type="checkbox"/> DomPred (Protein Domain Prediction)	
--	--

Function Prediction

<input type="checkbox"/> FFPred 3 (Eukaryotic Function Prediction)	
--	--

[Help...](#)

PSIPRED



Rozšíření predikce 2D struktury

- Predikce **více typů** 2D struktury (dle DSSP – Database of Secondary Structure Assignments)
 - α -helix (H)
 - 3_{10} -helix (G)
 - π -helix (I)
 - β -řetězec, extended strand (E)
 - β -bridge (B)
 - turn (T)
 - bend (S)
 - ostatní, coil (C)
- Predikce **přístupnosti solventu**
- Predikce **transmembránových helixů**

Predikce terciární struktury

- ***Ab initio***
- **Homologní modelování**
- **Threading („navlékání“)**





Ab initio

- Nejuniverzálnější – vychází pouze ze sekvence
- Výpočetně **nejnáročnější**
- Zahrnuje řadu kroků:
 - Predikce 2D struktury
 - Modelování jednotlivých fragmentů
 - Kombinace fragmentů navzájem
 - Doplnění smyček a flexibilních úseků
- **Nízká spolehlivost** zejm. pro větší proteiny



Ab initio

- Quark
- RaptorX

User Input

```
>1ci4A (87 residues)
TTSQKHRDFVAEPGEKPVGSLAGIGEVLGKKLEERGFDKAYVVLGQFLVKKDEDLFREW
LKDTCGANAKQSRDCFGLREWCDACL
```

Predicted Secondary Structure

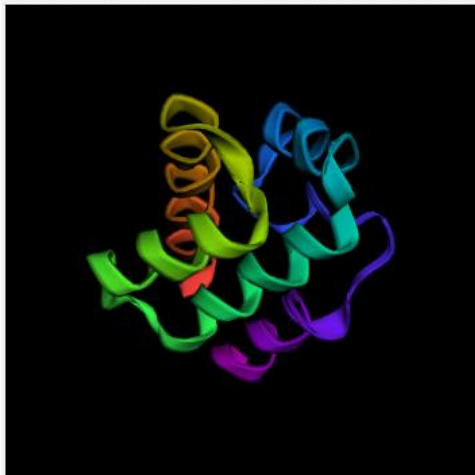
```
                20          40          60          80
Sequence  TTSQKHRDFVAEPGEKPVGSLAGIGEVLGKKLEERGFDKAYVVLGQFLVKKDEDLFREW
Prediction CCCHHHHHHCCCCCCCCCCCCCHHHHHHHHCCHHHHHHHHHHHHCCCHHHHHHHHHHHHH
Conf.Score 988899999879999987447898899999999979659999999999588899999999996889999999999999999859
          H:Helix; S:Strand; C:Coil
```

Predicted Solvent Accessibility

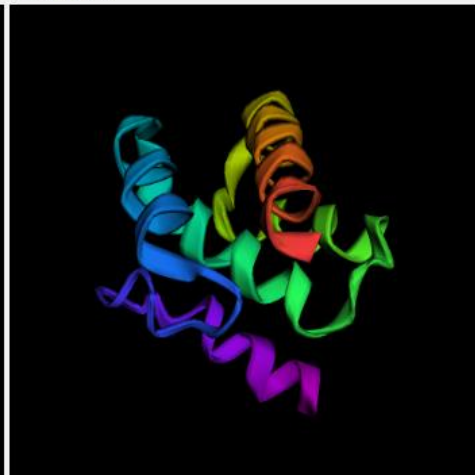
```
                20          40          60          80
Sequence  TTSQKHRDFVAEPGEKPVGSLAGIGEVLGKKLEERGFDKAYVVLGQFLVKKDEDLFREW
Prediction 553330221123223321120110032002102421132002000200113232310222102031310310010022003324
          Values range from 0 (buried residue) to 9 (highly exposed residue)
```

Top 5 Final Structure Model

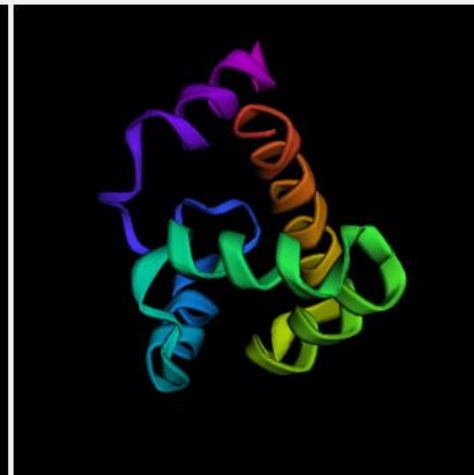
By dragging your mouse on the images, you rotate and zoom the structure.



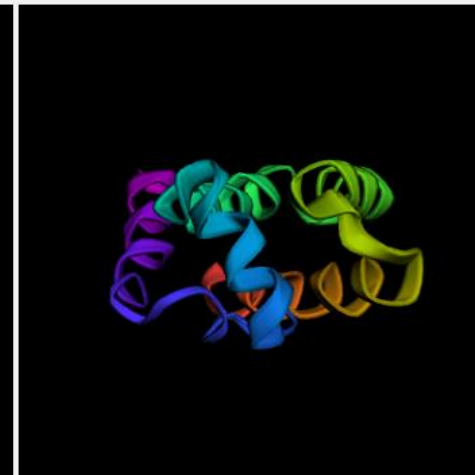
[Download Model 1](#)



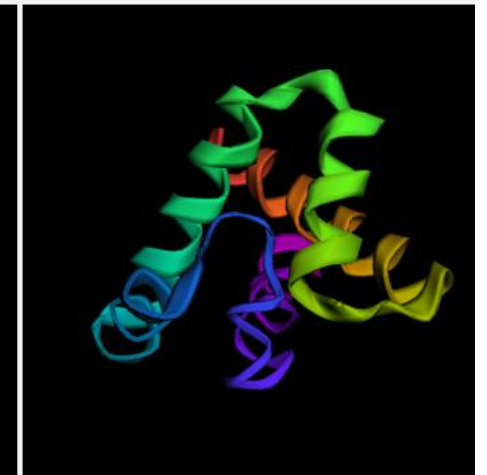
[Download Model 2](#)



[Download Model 3](#)



[Download Model 4](#)

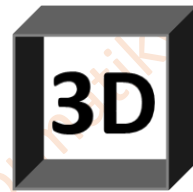


[Download Model 5](#)

Homologní modelování

A 3D isometric icon consisting of a dark grey cube with a white square in the center containing the text '3D' in black.

- Leží na opačném konci spektra než *ab initio*
- Je založeno na existenci blízkého **strukturního homologu** (typicky 50 % sekvenční podobnosti a více)
- Využívá skutečnosti, že dva proteiny ze stejné rodiny a s podobnou sekvencí mají i podobnou 3D strukturu
- Kromě sekvence modelovaného proteinu potřebujeme znát strukturu homologního proteinu = **templát**
- Pro vysoce homologní sekvence je spolehlivost velmi vysoká



Homologní modelování

1. **Alignment** zadané sekvence a sekvence templátu
2. Extrakce proteinové **páteře** ze struktury templátu a umístění **postranních řetězců**
3. **Modelování** otoček a smyček
4. **Minimalizace** energie
5. **Validace** vytvořené struktury

Swiss-Model

3D

- Výběr modelu (manuální, automatický)
- Podle vybraného modelu pak predikuje strukturu zadané sekvence
- Součástí výstupu je sada parametrů hodnotících **kvalitu** modelu. Při využití více templátů je tak možno porovnat jednotlivé modely

The screenshot displays the SWISS-MODEL web interface. The top navigation bar includes 'Modelling', 'Repository', 'Tools', 'Documentation', 'Log in', and 'Create Account'. The main content area is divided into three sections:

- Start a New Modelling Project:** A form for entering a target sequence (FASTA, Clustal, or UniProtKB AC) and a project title. It includes an 'Upload Target Sequence File...' button and a 'Validate' button.
- Template Results:** A table listing potential templates for structure prediction. The table has columns for Name, Title, Coverage, GMQE, QSQE, Identity, Method, and Oligo State. The first row is highlighted:

Name	Title	Coverage	GMQE	QSQE	Identity	Method	Oligo State
2ezy.1.A	BARRIER-TO-AUTOINTEGRATION FACTOR	100.00	0.99	0.56	100.00	NMR	homo-dimer
2ezx.1.A	BARRIER-TO-AUTOINTEGRATION FACTOR	100.00	0.99	0.56	100.00	NMR	homo-dimer
2e2x.1.A	BARRIER-TO-AUTOINTEGRATION FACTOR	100.00	0.99	0.70	100.00	NMR	homo-dimer
2odg.1.A	Barrier-to-autointegration factor	100.00	0.99	0.53	100.00	NMR	hetero-trimer
2baf.1.A	BARRIER-TO-AUTOINTEGRATION FACTOR	100.00	0.99	0.66	100.00	X-ray, 2.9Å	homo-dimer

- Model Results:** A detailed view of the selected model (2ezy.1.A). It shows a 3D ribbon diagram of the protein structure. Key quality metrics are displayed: GMQE (0.99) and QMEAN (-1.17). A 'Global Quality Estimate' section includes a heatmap for C β and All Atom, and a 'Local Quality Estimate' plot showing quality across the residue number. A 'Comparison' plot shows the protein size in residues. The description is 'BARRIER-TO-AUTOINTEGRATION FACTOR'. The model-template alignment is shown at the bottom.

The URL <http://swissmodel.expasy.org/> is provided for reference.

Threading

A 3D-style icon consisting of a black square with a white center containing the text '3D' in black. The square has a slight 3D effect with a grey shadow on the right side.

- Z hlediska náročnosti i spolehlivosti leží mezi *ab initio* a homologním modelováním
- Používá se pro případy, kdy zkoumaný protein má **nízkou homologii** s proteiny se známou strukturou (typicky cca 15-40 %)
- Porovnává možnost přiložení sekvence na proteiny známých **foldů**



Threading

1. S využitím strukturních databází (PDB, SCOP, CATH) je vytvořena databáze existujících foldů
 2. Sekvence je porovnána s potenciálními templáty
 - Alignment
 - Každou aminokyselinu se pokusí umístit do pozice aminokyseliny v templátu
 - Hodnocení umístění
 3. Výběr templátu pro výsledný model
- Proteiny s více doménami je nutné rozdělit a modelovat zvlášť

Phyre2

3D

- Server pro 3D predikci struktur pomocí **threadingu**
- Vysoce výkonný – poměrně spolehlivá detekce foldu i při **nízké homologii** (i pod 15%)

<http://www.sbg.bio.ic.ac.uk/phyre2/>

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c1r1zB 	Alignment		100.0	31	PDB header: sugar binding protein Chain: B; PDB Molecule: ergic-53 protein; PDBTitle: the crystal structure of the carbohydrate recognition2 domain of the glycoprotein sorting receptor p58/ergic-533 reveals a novel metal binding site and conformational4 changes associated with calcium ion binding
2	c2a6yA 	Alignment		100.0	21	PDB header: sugar binding protein Chain: A; PDB Molecule: emp47p (form1); PDBTitle: crystal structure of emp47p carbohydrate recognition domain2 (crd), tetragonal crystal form
3	d2a6za1 	Alignment		100.0	20	Fold: Concanavalin A-like lectins/glucanases Superfamily: Concanavalin A-like lectins/glucanases Family: Lectin leg-like
4	c2dupB 	Alignment		100.0	42	PDB header: protein transport Chain: B; PDB Molecule: vesicular integral-membrane protein vip36; PDBTitle: crystal structure of vip36 exoplasmic/luminal domain, metal-free form

Phyre2

I-TASSER



- Několikrát vyhodnocen jako nejlepší predikční server

<https://zhanglab.ccmb.med.umich.edu/I-TASSER/>

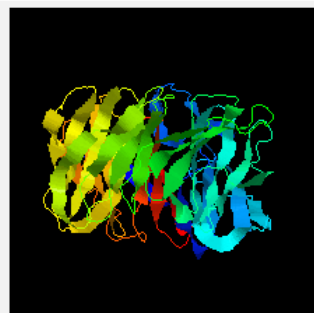
Predicted Secondary Structure

	20	40	60	80	100	120	140
Sequence	Y P F F D N P N Y T N T Y A T N E D F V C P Y F L D Y Y N S Q D D Y K N F R G E N Y D F E D T E E N I E N R N I E T E Y E G L F R A W N P W N L G G N I T S G L G A S S W A A N R I D L F A R G R G G E L I H N W F D N G K W N Y W E N L G G I L T S S P K A V S W G F N R I D V V C R G I						
Prediction	C C C C C C C C C C C C C C C C S S C C C H H H H H H C C C C C C C C C C C C S S C C S S S S S S S S C C C S S S S S C C C C C C C C C C C C S S C C C C C C C C C C S S S C C C C C C C C C C C C S C C C C C C C C C C S S S S C C C C S S S S S C C						
Conf. Score	9956778667888866536775235655126000458983036011147999825725887516989888771189977898747980899689999959995899982599876664068976588678993899489999989						

Predicted Solvent Accessibility

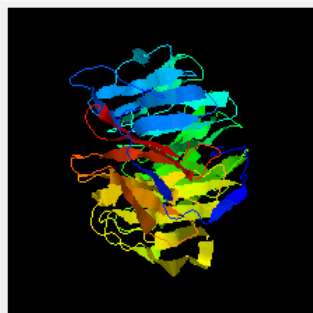
	20	40	60	80	100	120	140
Sequence	Y P F F D N P N Y T N T Y A T N E D F V C P Y F L D Y Y N S Q D D Y K N F R G E N Y D F E D T E E N I E N R N I E T E Y E G L F R A W N P W N L G G N I T S G L G A S S W A A N R I D L F A R G R G G E L I H N W F D N G K W N Y W E N L G G I L T S S P K A V S W G F N R I D V V C R G I						
Prediction	40103001223323343301011105412234741731433313143323103233122321323323123123031322120000023721000000034111010122332303213312232222000001372200000103						
	Values range from 0 (buried residue) to 9 (highly exposed residue)						

Top 5 Models predicted by I-TASSER



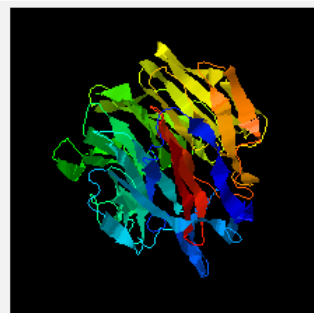
[Download Model 1](#)

C-score=-2.09



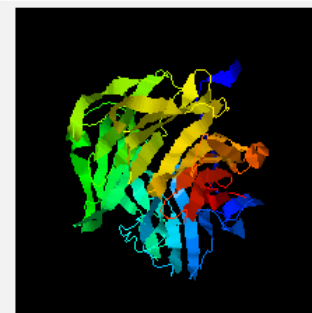
[Download Model 2](#)

C-score=-2.42



[Download Model 3](#)

C-score=-3.44



[Download Model 4](#)

C-score=-3.46



[Download Model 5](#)

C-score=-3.53

Estimated accuracy of Model1: 0.47±0.15 (TM-score) 11.3±4.5Å (RMSD) ([Read more about C-score of generated models](#))



Jakou metodu zvolit?

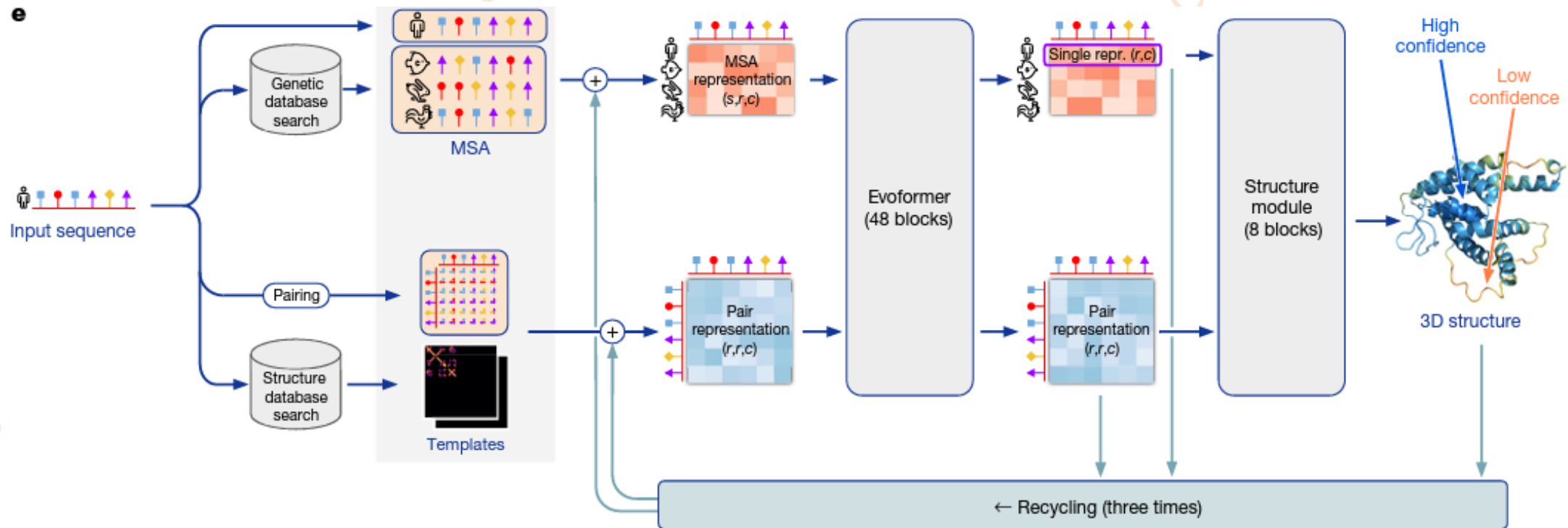
1. Mám homologní protein se známou strukturou → homologní modelování
2. Využiji experimentální data
 - Threading
 - Kombinace více templátů pro jednotlivé části struktury
 - Různé predikční nástroje
3. *Ab initio* modelování smyček a částí sekvence bez vhodného templátu
4. Mám unikátní sekvenci – *ab initio*

AlphaFold

Kombinace strukturních dat, alignmentu a neuronových sítí

První verze AlphaFold – spolehlivost predikce < 60 %

Od roku 2020 **AlphaFold2** – spolehlivost predikce > 90 %



AlphaFold Protein Structure Database



2021: Predikovaná struktura proteomu člověka a 47 dalších klíčových organismů (celkem 992 316 predikovaných struktur)

2023: > 200 000 000 predikovaných struktur

<https://alphafold.ebi.ac.uk/>

A screenshot of the AlphaFold Protein Structure Database website. The page has a dark blue header with the text 'AlphaFold Protein Structure Database' on the left and navigation links 'Home', 'About', 'FAQs', and 'Downloads' on the right. The main content area is also dark blue and features the title 'AlphaFold Protein Structure Database' in large white font, with 'Developed by DeepMind and EMBL-EBI' below it. A search bar is present with the placeholder text 'Search for protein, gene, UniProt accession or organism' and a 'BETA' label. Below the search bar, there are examples: 'Free fatty acid receptor 2', 'At1g58602', 'Q5VSL9', and 'E. coli', along with a 'Help: AlphaFold DB search help' link. At the bottom, there is a 'Feedback on structure: Contact DeepMind' link.



AlphaFold

Výhody

- Vysoká přesnost určení foldu
- Známé sekvence jsou již predikovány
- Dostupný pro širokou veřejnost

Rizika

- Nízká přesnost určení pozice bočních řetězců
- Přesnost klesá u unikátních sekvencí
- Bez posttranslačních modifikací
- Nevhodné pro komplexy (aktuálně v řešení)



Predikce kvartérní struktury

Zahrnuje různé úrovně, např.:

- Predikce vazebných míst
 - Predikce aminokyselin podílejících se na interakci
 - Odhad oligomerního stavu
 - Protein-protein docking (protein-nukleová kyselina docking)
-
- SW dosud často nedokonalý, **nízká spolehlivost** predikce
 - Složitější postupy většinou nejsou automatizované

Predikce kvartérní struktury

4D

Programy většinou vycházejí z podobnosti sekvence a/nebo 3D struktury se známými proteiny

Příklady SW:

- QuatIdent
- QuaBingo
- M-TASSER
- Quad-PRE
- **AlphaFold-Multimer**

Journal List > Biophys J > v.94(3); Feb 1, 2008 > PMC2186260

Biophysical Journal

Biophys J. 2008 February 1; 94(3): 918–928.
doi: [10.1529/biophysj.107.114280](https://doi.org/10.1529/biophysj.107.114280)

PMCID: PMC2186260

M-TASSER: An Algorithm for Protein Quaternary Structure Prediction

Huiling Chen and Jeffrey Skolnick*

[Author information](#) [Article notes](#) [Copyright and License information](#)

This article has been [cited by](#) other articles in PMC.

Abstract

In a cell, it has been estimated that each protein on average interacts with rough in tens of thousands of proteins known or suspected to have interaction partners; fraction have solved protein structures. To partially address this problem, we have developed TASSER, a hierarchical method to predict protein quaternary structure from sequence template identification by multimeric threading, followed by multimer model assembly. The final models are selected by structure clustering. M-TASSER has been tested comprising 241 dimers having templates with weak sequence similarity and 246

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 715494, 9 pages
<http://dx.doi.org/10.1155/2014/715494>

Research Article

Quad-PRE: A Hybrid Method to Predict Protein Quaternary Structure Attributes

Yajun Sheng,¹ Xingye Qiu,¹ Chen Zhang,¹ Jun Xu,¹ Yanping Zhang,¹ Wei Zheng,¹ and Ke Chen²

¹ School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China
² School of Computer Science and Software Engineering, Tianjin Polytechnic University, No. 399 Binshui Road, Tianjin 300387, China

Correspondence should be addressed to Ke Chen; kchen1.tj@gmail.com

Received 27 February 2014; Revised 24 April 2014; Accepted 27 April 2014; Published 18 May 2014

Academic Editor: Tao Huang

Copyright © 2014 Yajun Sheng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The protein quaternary structure is very important to the biological process. Predicting their attributes is an essential task in computational biology for the advancement of the proteomics. However, the existing methods did not consider sufficient properties of amino acid. To end this, we proposed a hybrid method Quad-PRE to predict protein quaternary structure attributes using the properties of amino acid, predicted secondary structure, predicted relative solvent accessibility, and position-specific scoring matrix profiles and motifs. Empirical evaluation on independent dataset shows that Quad-PRE achieved higher overall accuracy 81.7%.

Evaluace kvality struktur a modelů



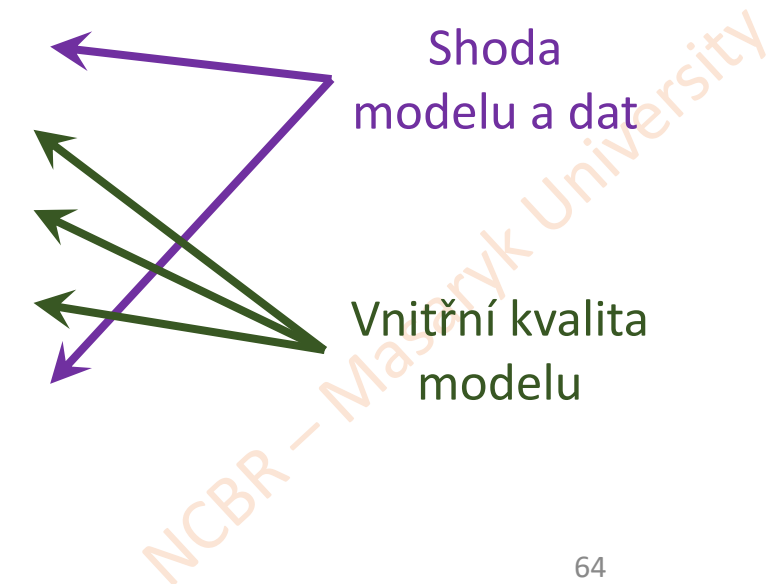
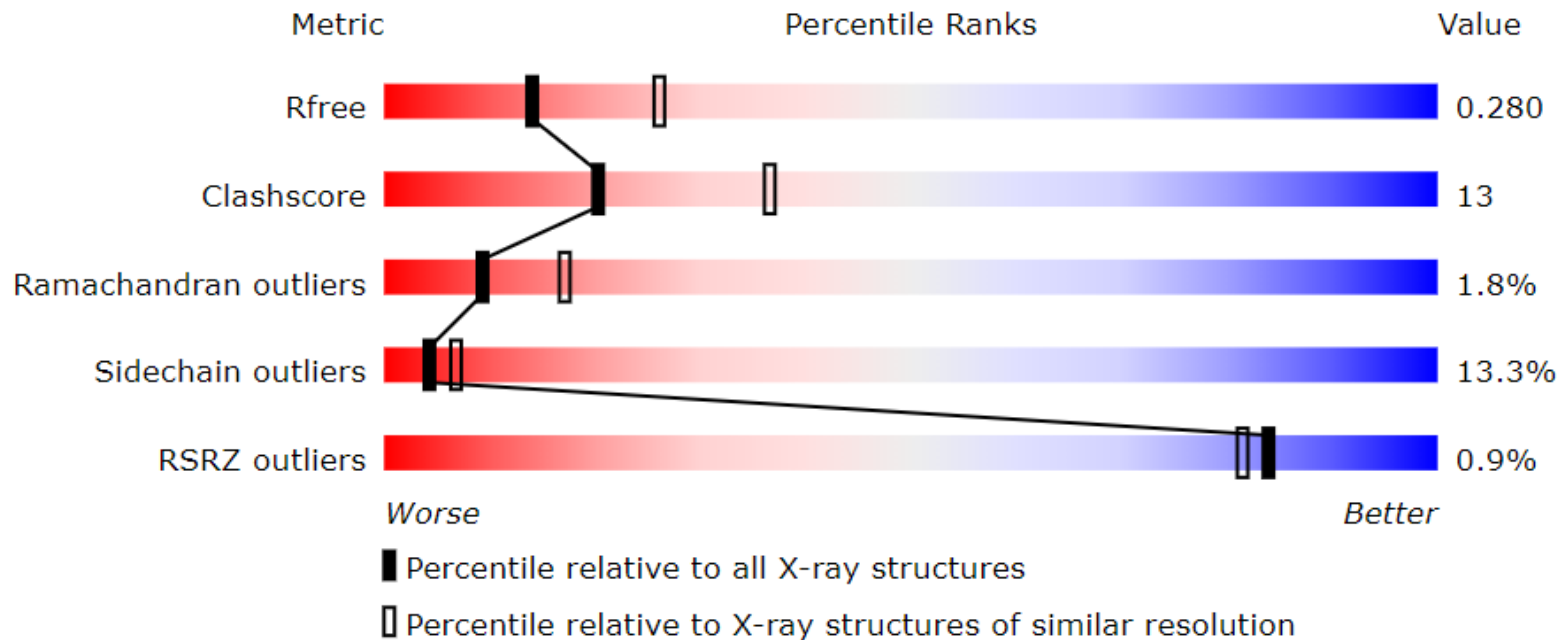
Evaluace kvality struktur a modelů

- **Shoda** strukturního **modelu** a vstupních **dat**
- Základní **fyzikální principy**

- Správně získaná **Experimentální** struktura vs. **Predikce**

PDB – validace dat

- Struktury vytváří lidé → mohou obsahovat chyby
- Kontrola při nahrávání struktur do databáze
- Informace o kvalitě u každé struktury

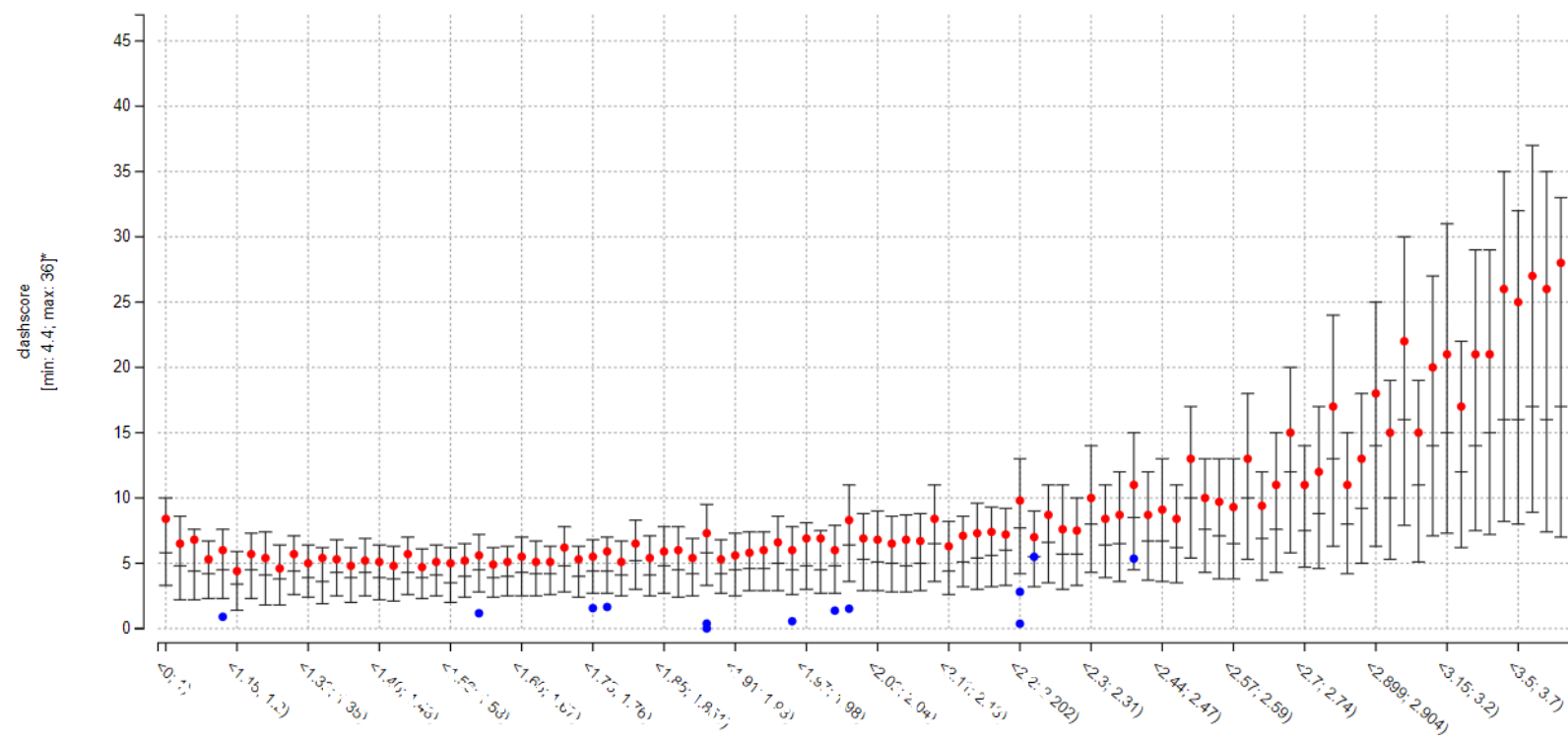


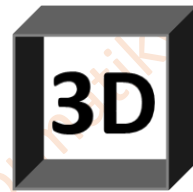
ValTrends DB

<https://webchem.ncbr.muni.cz/ValTrendsDB/>

Přehledová analýza struktur v PDB databázi

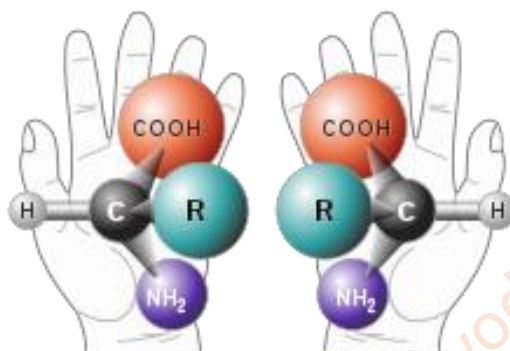
The plot below shows a relationship between **clashscore structure quality factor** and **structure resolution factor**. Values of **13 PDB** entries that have been deposited to the PDB database by Dr. Josef Houser are visualized in the plot by blue points. The plot demonstrates that **structures deposited by Dr. Houser have markedly better quality than average**.





MotiveValidator

- Kontrola struktury malých molekul – ligandů
- Úplnost struktury
- Správná chiralita
- Anotace



<http://webchem.ncbr.muni.cz/Platform/MotiveValidator>

MotiveValidator

Validate ligand and residue structure in biomolecular complexes.

MotiveValidator is a platform for a set of applications designed to help you determine whether a residue or a ligand in a biomolecule or biomolecular complex is structurally complete and correctly annotated according to its models stored in the **wwPDB Chemical Component Dictionary (wwPDB CCD)**.

The applications provided within the **MotiveValidator** platform cover all residues and ligands defined in the **wwPDB CCD**, and available via **PDBChem**. In addition, you may specify your own model residue if it is not available in **wwPDB CCD**.

Are you interested in validating ligands and non-standard residues in existing PDBe.org entries? Check out [Validator^{DB}](#).

[Quick Help](#)

[Residue Validation](#)

[Sugar Validation](#)

[Motif/Fragment Validation](#)

[Command Line Version](#)

Automatic custom residue validation in one or more biomolecules

- Reads the structure of an input biomolecule or biomolecular complex, and an input model residue to serve as reference template for validation.
- Scans the **entire biomolecule(s)**, automatically detects all residues in the input biomolecule(s) with the same annotation (i.e., the same 3-letter code) as the model residue, and subsequently validates them by comparison to the model.

Model Residue(s):

Select a single file or a ZIP file containing model residues(s) (a model must contain exactly one residue) in PDB or PDBx/mmCIF format. When using the PDB format, it is recommended that the input is a ZIP archive with both PDB and SD/SDF/MOL (for bonds) versions present.

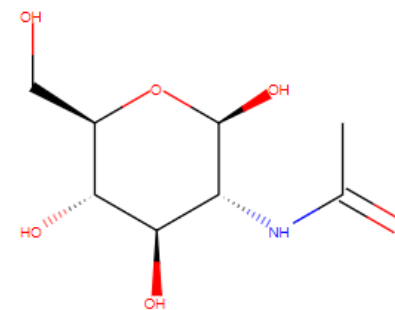
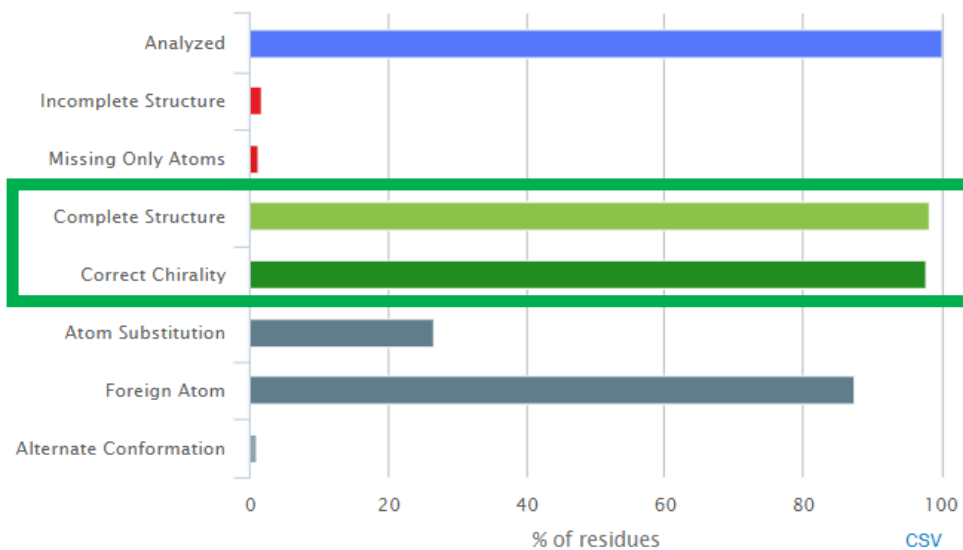
Biomolecule(s):

Select a single file or a ZIP file containing entire biomolecule(s) in PDB or PDBx/mmCIF format (300MB limit).

PDB identifiers are used only if no file is selected. Loaded from PDBx/mmCIF format.

MotiveValidator

NAG [[PDBChem](#) | [PDB](#)] [229 molecules in 46 PDB entries | 3 warnings]
 [C₈H₁₅NO₆ | n-acetyl-d-glucosamine] [Chiral Atoms (5): C1, C2, C3, C4, C5] [Experimental Coordinates]



Incomplete Structure			Complete Structure				
1.7% 4			98.3% 225				
Missing Only Atoms	Missing Rings	Degenerate	Correct Chirality (Tolerant)	Wrong Chirality	Atom Substitution	Foreign Atom	Different Naming
1.3% 3	0.4% 1	-	97.8% 224	0.4% 1	26.6% 81	87.3% 200	-
Missing Atoms in 4 molecule(s)							
C1 25.0% 1	C2 25.0% 1	C3 25.0% 1	C4 25.0% 1	C5 25.0% 1	C6 25.0% 1	O1 100.0% 4	O3 25.0% 1
O4 25.0% 1	O5 25.0% 1	O6 25.0% 1					
Chirality Errors in 1 molecule(s)							
C1 100.0% 1							

Kontrola predikovaného modelu

➤ Programy vytvoří nějaký model vždy → NUTNÁ KONTROLA

➤ **Vizuální kontrola struktury**

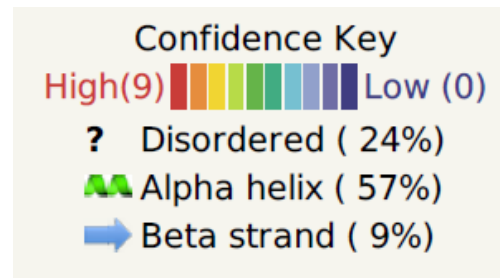
- Je roztržená?
- Dochází k překrytí aminokyselin?

➤ **Nastavení programu**

- Obsahuje model celou zadanou sekvenci?
- Byl zvolen smysluplný templát?

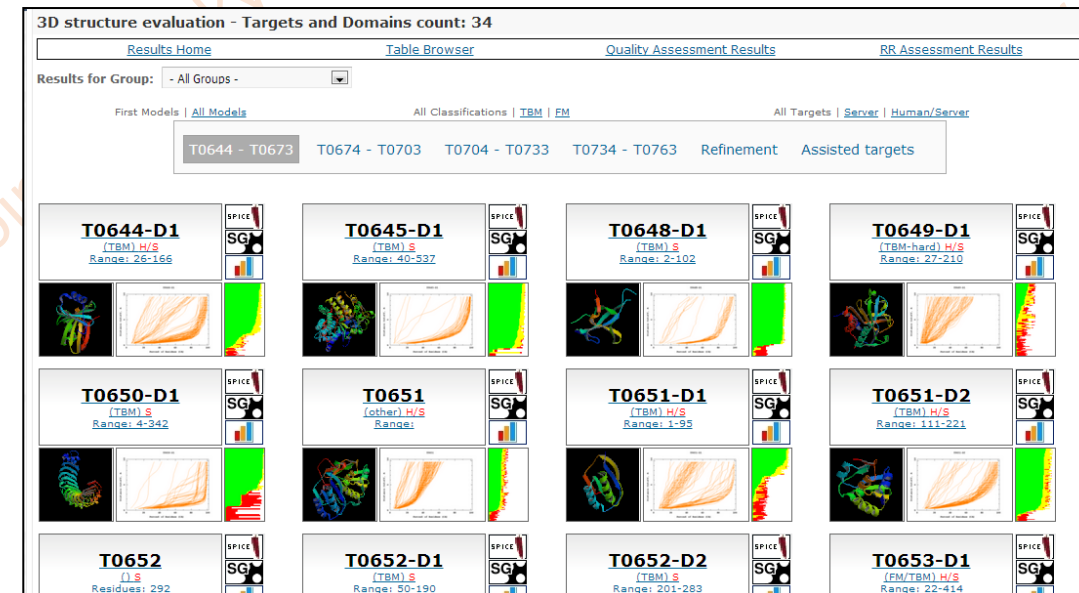
➤ **Skóre**

- QMEAN (<https://swissmodel.expasy.org/qmean/>)



Hodnocení kvality predikčních nástrojů - CASP

- *Critical Assessment of Techniques for Protein Structure Prediction*
- 2022 – CASP15
- Predikce vyřešených, ale zatím nepublikovaných struktur
- **Rozsáhlá analýza predikčních programů**
 - Predikce terciárních struktur
 - Identifikace neuspořádaných oblastí
 - Funkční predikce (predikce vazebných míst)
 - Interakce mezi doménami, podjednotkami a proteiny
 - Hodnocení spolehlivosti



Závěrem

- Struktura je klíčová pro správnou funkci proteinu
- Predikovat na základě sekvence (1D) lze 2D, 3D i 4D strukturu
- Vždy je nutné **kriticky kontrolovat** výstupy programů
- Ideální je využít více predikčních programů s různou metodologií a porovnat výsledky