

wykład 11

Aminokwasy, białka, struktury drugorzędowe

dr Jacek Śmietański

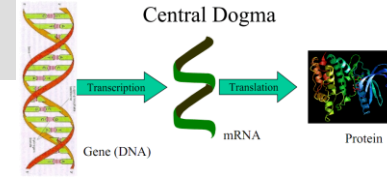
jacek.smietanski@ii.uj.edu.pl

<http://jaceksmietanski.net>

1. Znaczenie struktury makrocząsteczek
2. Poziomy organizacji strukturalnej
3. Klasyfikacja rodzin białkowych
4. Przewidywanie struktury drugorzędowej białek

1. Znaczenie struktury makrocząsteczek
2. Poziomy organizacji strukturalnej
3. Klasyfikacja rodzin białkowych
4. Przewidywanie struktury drugorzędowej białek

ZNACZENIE STRUKTURY MAKROCZĄSTECZEK

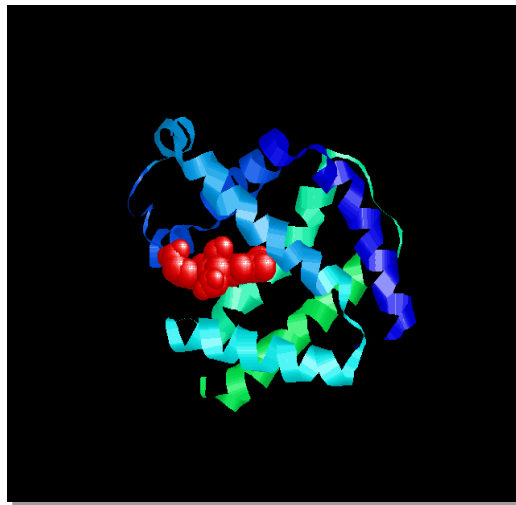


sekwencja -> struktura -> funkcja

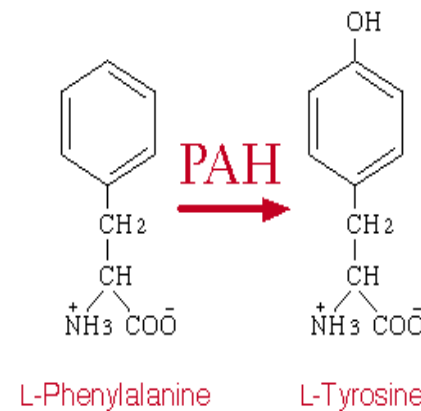
Genetic Information

MVHLTPEEKT
AVNALWGKVN
VDAVGGEALG
RLLVVYPWTQ
RFFESFGDLS
SPDAVMGNPK
VKAHGKKVLG
AFSDGLAHL
DLNLKGTF
SQLSELHCD
KLHVDPENF
RLLGNVL
LCVLARNF
GKEFTPQM
QAAAYQKV
VAGVANAL
AHKYH

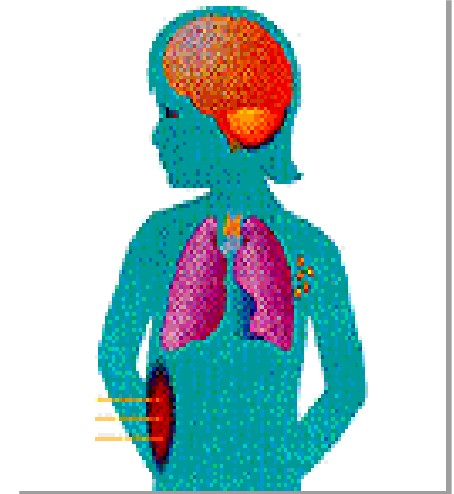
Molecular Structure



Biochemical Function

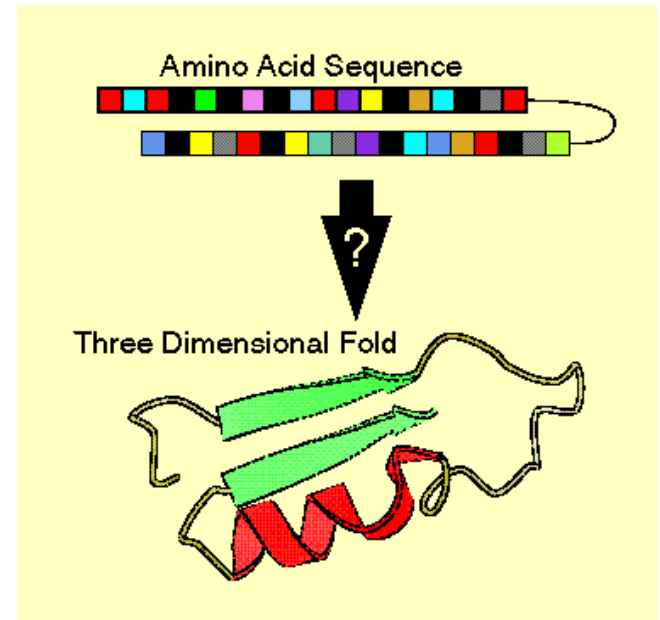


Phenotype (Symptoms)



Dotyczą zarówno białek, jak i kwasów nukleinowych (głównie RNA).

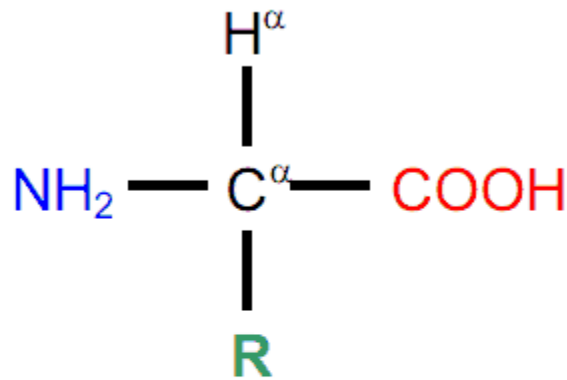
- przewidywanie struktury
- poszukiwanie miejsc wiążących
- predykcja oddziaływań
- modelowanie molekularne
- analiza relacji ewolucyjnych
- przewidywanie funkcji
- projektowanie leków (CADD)



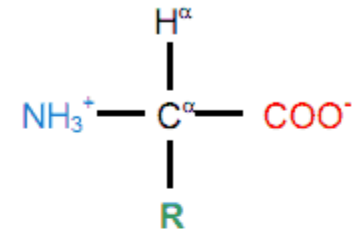
1. **Znaczenie struktury makrocząsteczek**
2. **Poziomy organizacji strukturalnej**
3. **Klasyfikacja rodzin białkowych**
4. **Przewidywanie struktury drugorzędowej białek**

POZIOMY ORGANIZACJI

Ogólna budowa aminokwasów:



- w neutralnym pH



grupa aminowa - NH_2

grupa karboksylowa - COOH

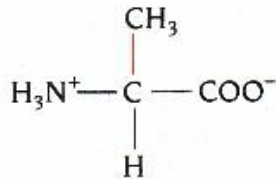
Aminokwasy białkowe

nazwa	symbol	skrót
alanina	A	Ala
arginina	R	Arg
asparagina	N	Asn
kw.asparaginowy	D	Asp
cysteina	C	Cys
glutamina	Q	Gln
kw.glutaminowy	E	Glu
glicyna	G	Gly
histydyna	H	His
izoleucyna	I	Ile

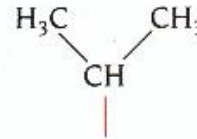
nazwa	symbol	skrót
leucyna	L	Leu
lizyna	K	Lys
metionina	M	Met
fenyloalanina	F	Phe
prolina	P	Pro
seryna	S	Ser
treonina	T	Thr
tryptofan	W	Trp
tyrozyna	Y	Tyr
walina	V	Val

Aminokwasy hydrofobowe

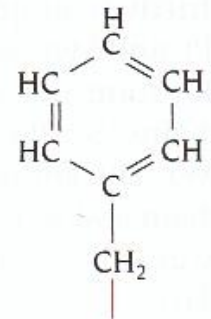
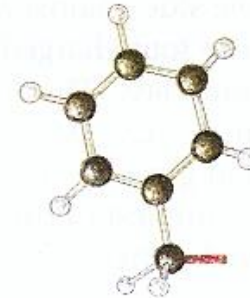
(a) Hydrophobic amino acids



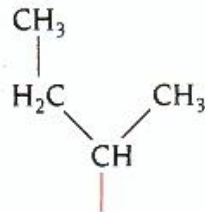
A Ala, Alanine



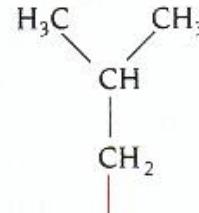
V Val, Valine



F Phe, Phenylalanine

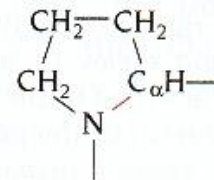
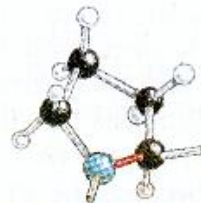


I Ile, Isoleucine

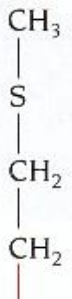


L Leu, Leucine

A, V, F, I, L, P, M
C, G, Y, W, H, K, T



P Pro, Proline



M Met, Methionine

Aminokwasy polarne

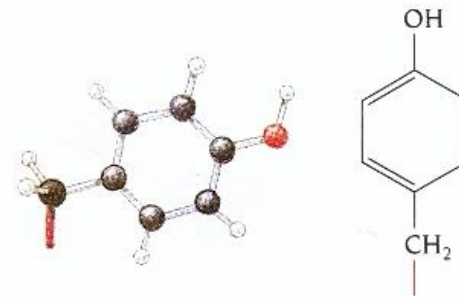
(c) Polar amino acids



S Ser, Serine



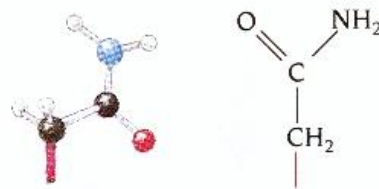
T Thr, Threonine



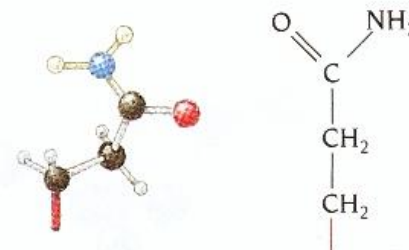
Y Tyr, Tyrosine



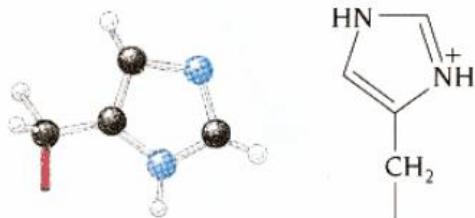
C Cys, Cysteine



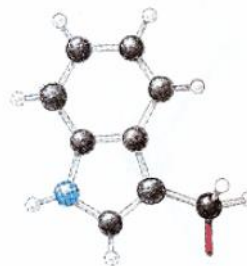
N Asn, Asparagine



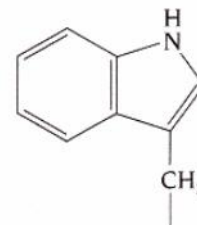
Q Gln, Glutamine



H His, Histidine



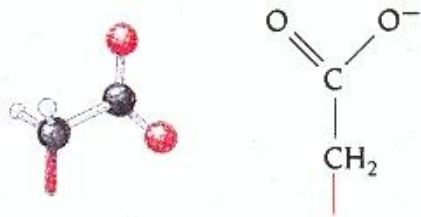
W Trp, Tryptophan



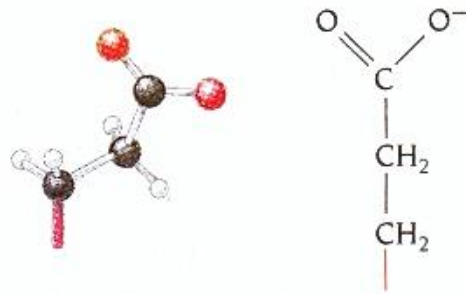
S, T, Y, C, N, Q, H, W
D, E, K, R

Aminokwasy naładowane

(b) Charged amino acids



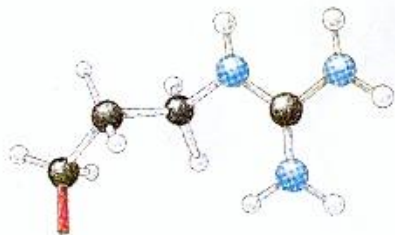
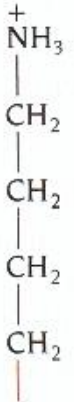
D Asp, Aspartic acid



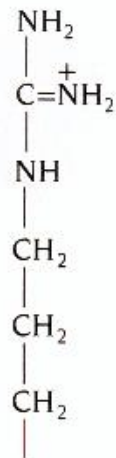
E Glu, Glutamic acid



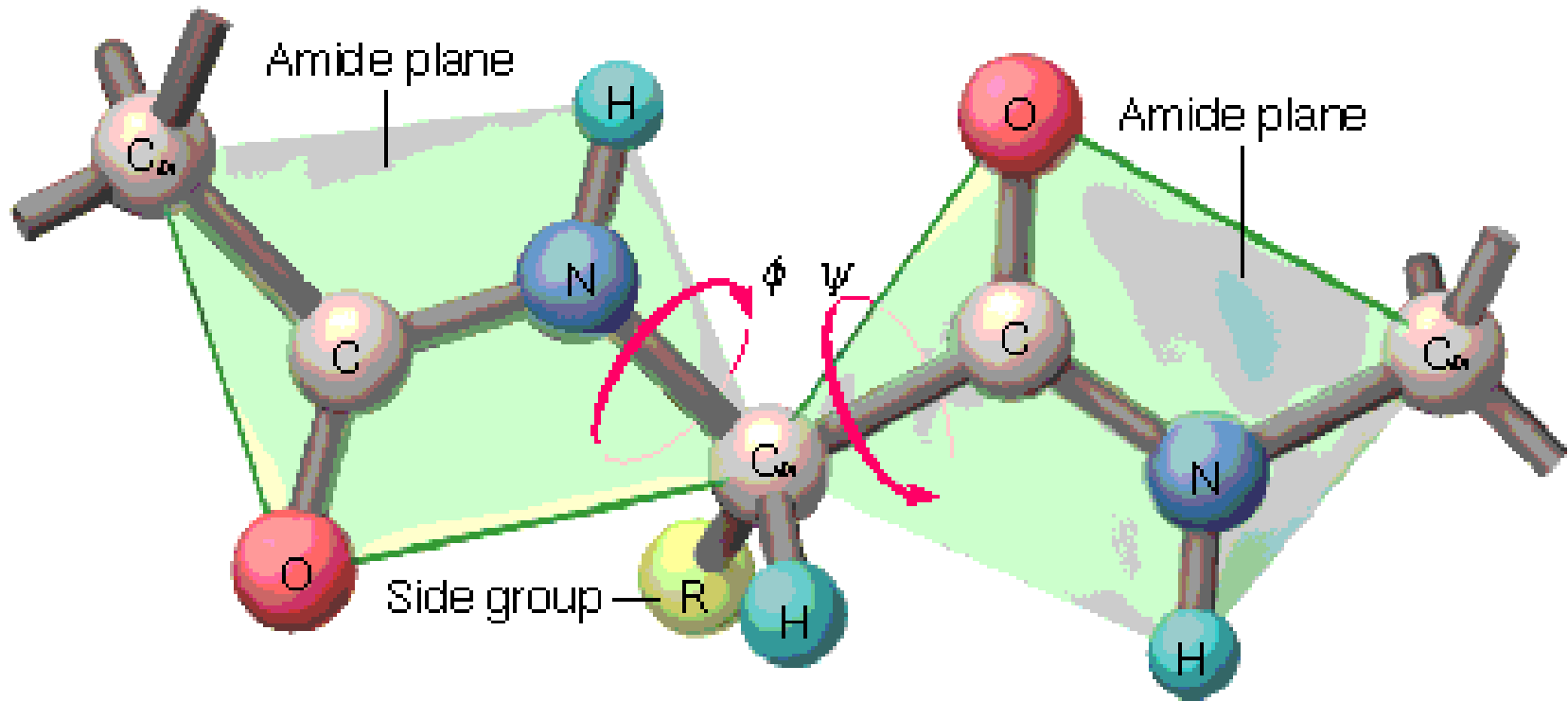
K Lys, Lysine



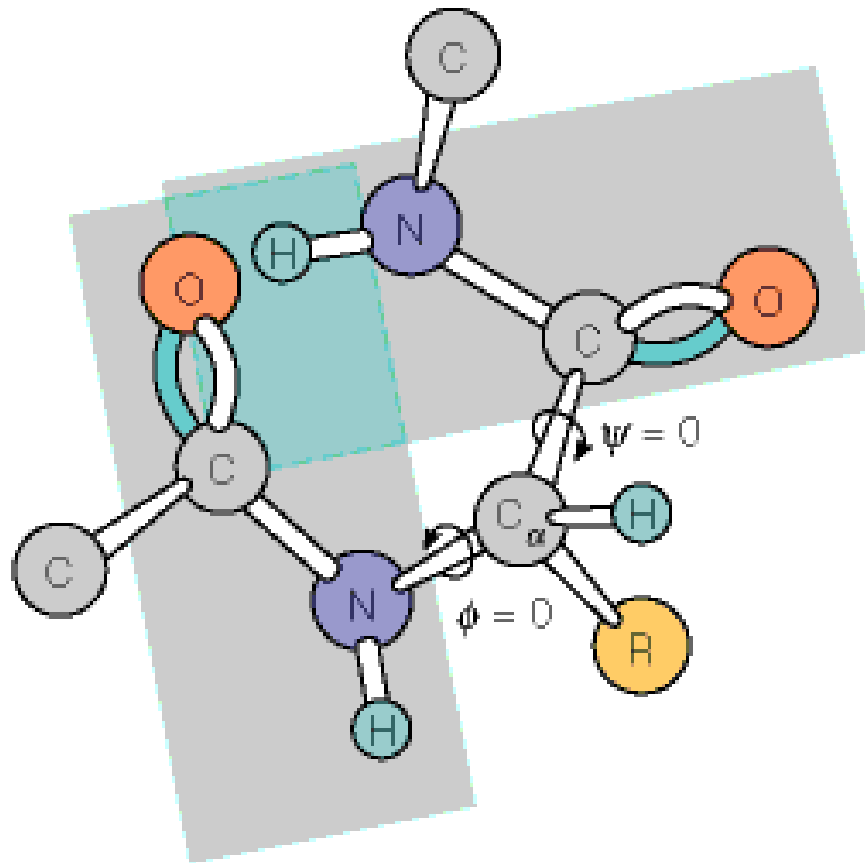
R Arg, Arginine



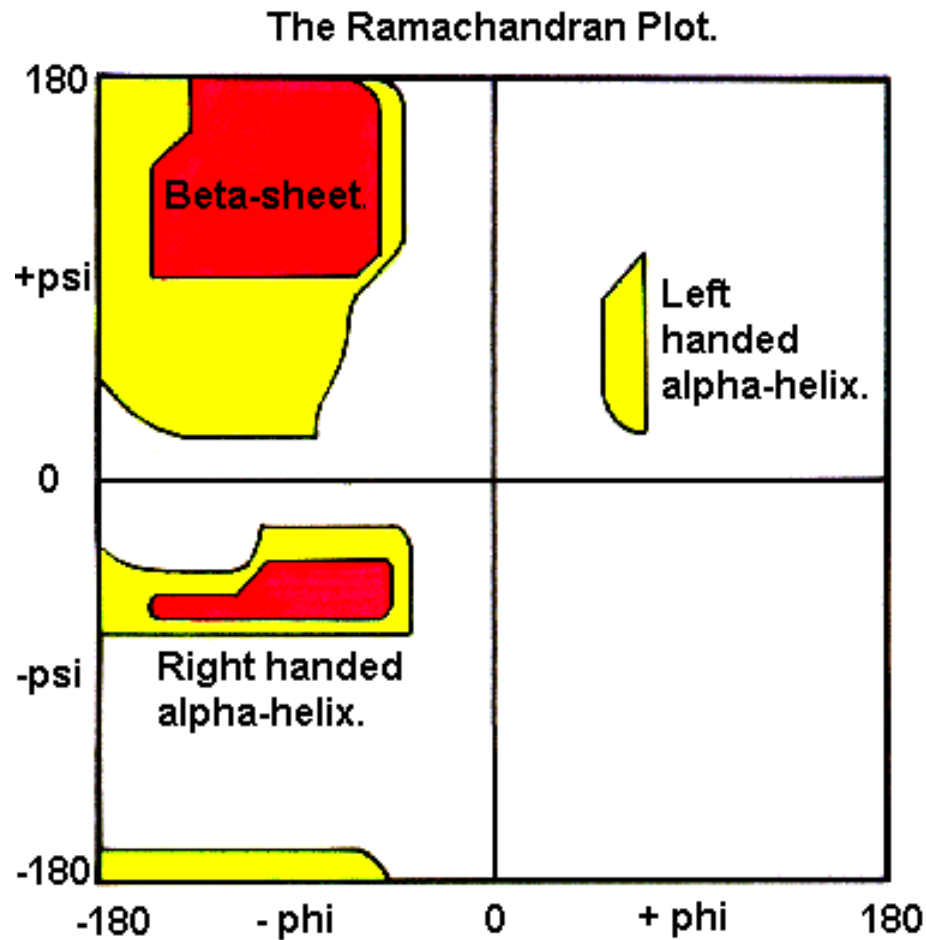
D, E, K, R, H



Ponieważ wiązanie peptydowe jest płaskie, konformacja głównego łańcucha może być wyznaczona przez kąty ϕ i ψ .



**Niektóre konformacje
są niedopuszczalne.**



Obrazuje dopuszczalne konformacje.

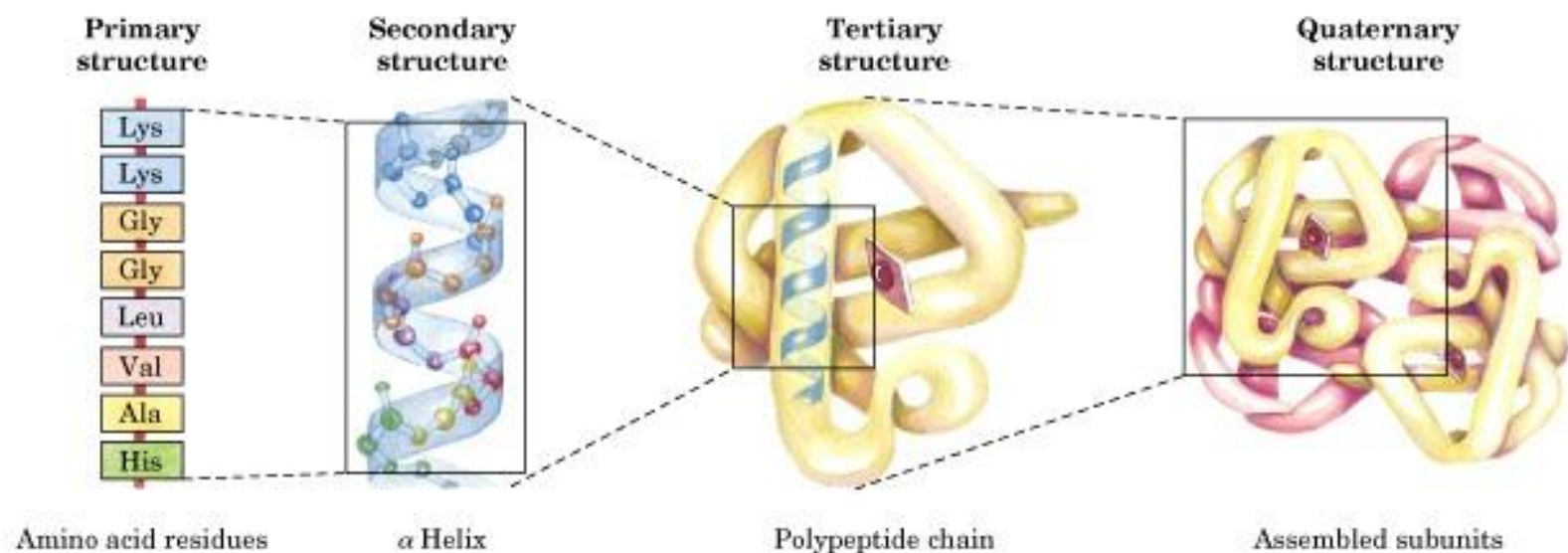
Poziomy przestrzennej organizacji białek:

I rządowa – liniowa sekwencja aminokwasów

II rządowa – opisuje lokalne pofałdowanie (α -helisy, β -kartki)

III rządowa – struktura 3D pojedynczego łańcucha

IV rządowa – struktura 3D całego białka (połączone wszystkie łańcuchy)

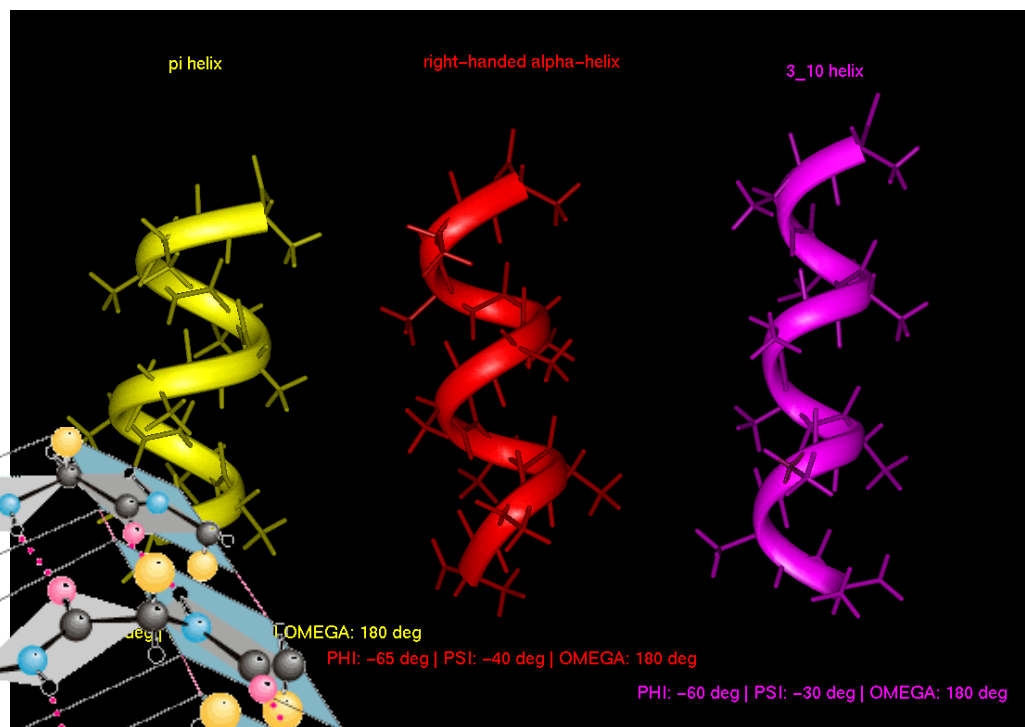
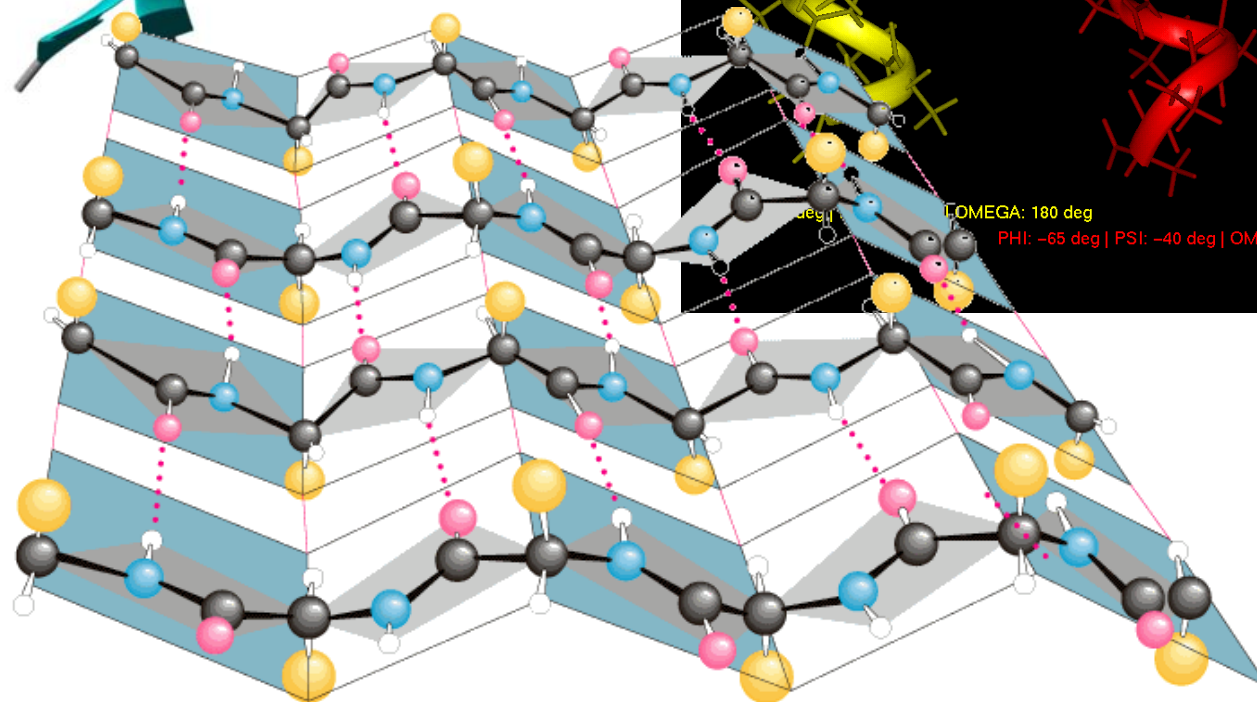


Zwijanie białka (film): <http://www.youtube.com/watch?v=fvBO3TqJ6FE>

Struktura 2-rzędowa



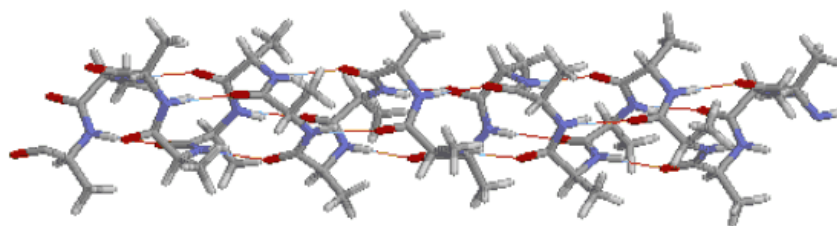
α -helisy
 β -kartki
zwroty



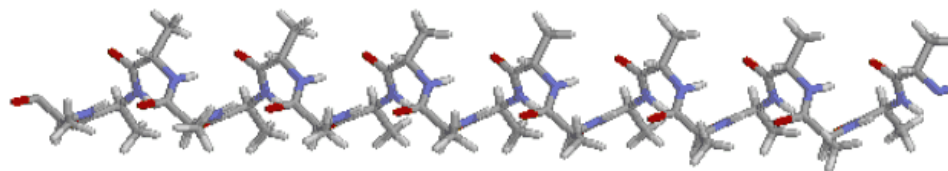
α -helisy

helisa	φ	ψ	ω	reszt aminkwasowych na skręt	przesunięcie na resztę (Å)	wiązania wodorowe
α helisa	-57	-47	180	3,6	1,5	i+4
3_{10} helisa	-49	-26	180	3,0	2,0	i+3
π helisa	-57	-70	180	4,4	1,2	i+5

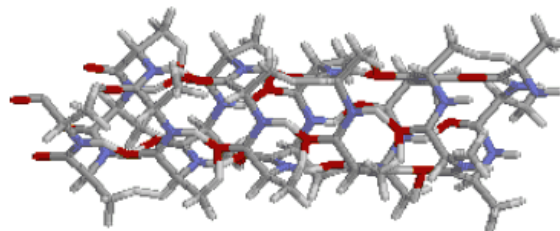
α - helisa



3_{10} - helisa



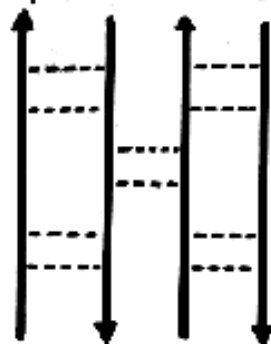
π - helisa
(22-reszty aminokwasowe)



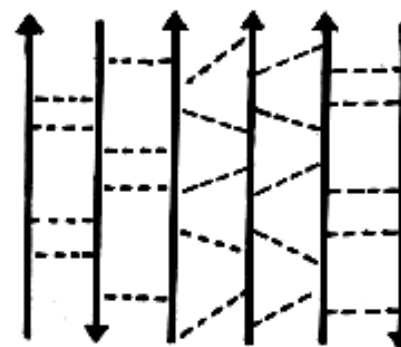
Są stabilizowane przez wiązania wodorowe i kontakt sąsiadujących łańcuchów

- równoległe
- antyrównoległe
- mieszane

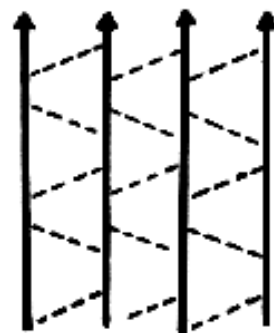
Antiparallel beta-sheet



The different types of beta-sheet. Dashed lines indicate main chain hydrogen bonds.

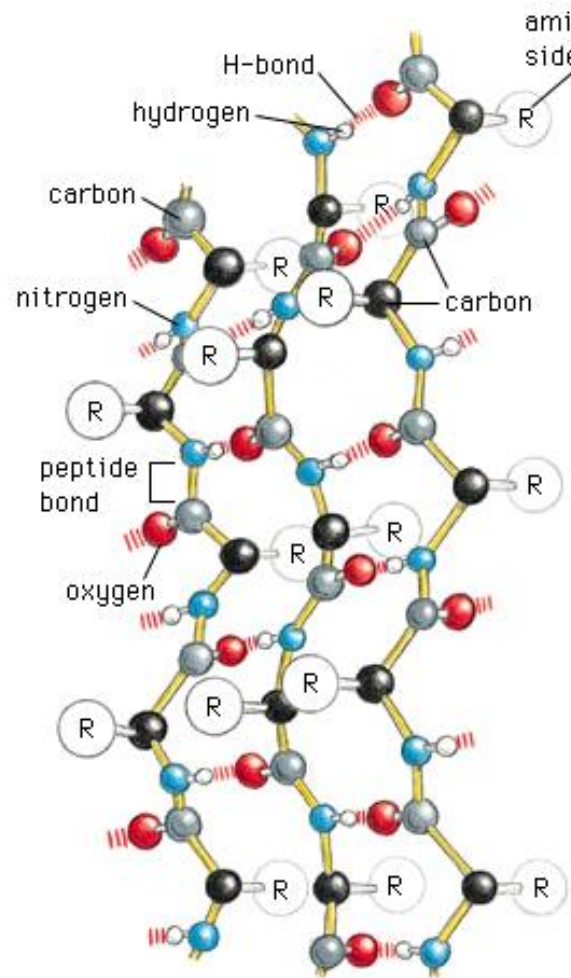


Mixed beta-sheet

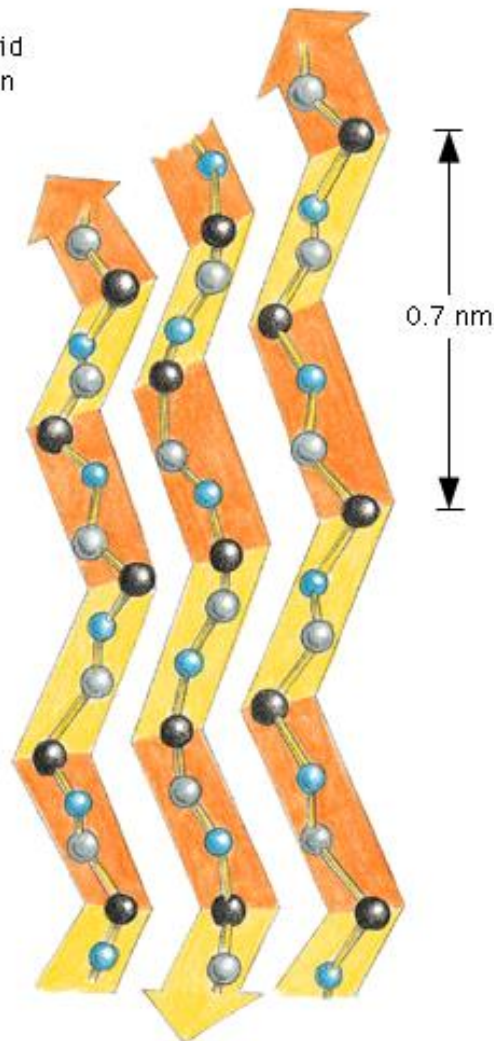


Parallel beta-sheet

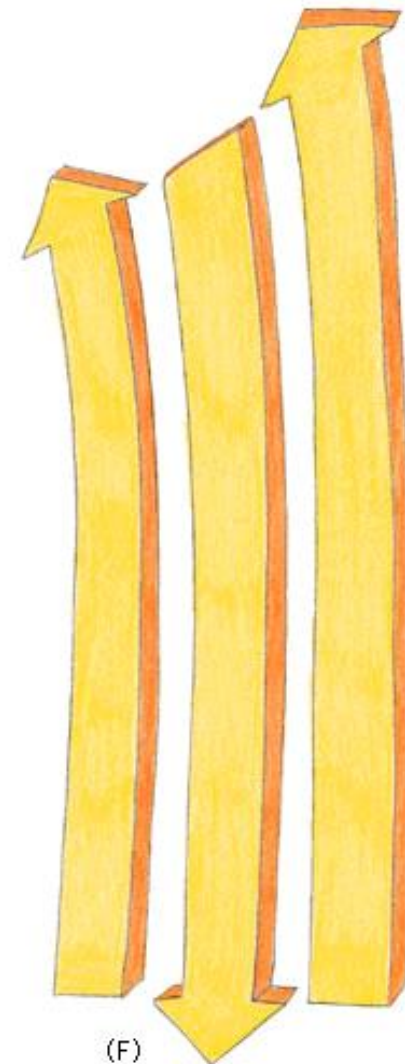
β -kartki – reprezentacja graficzna



(D)



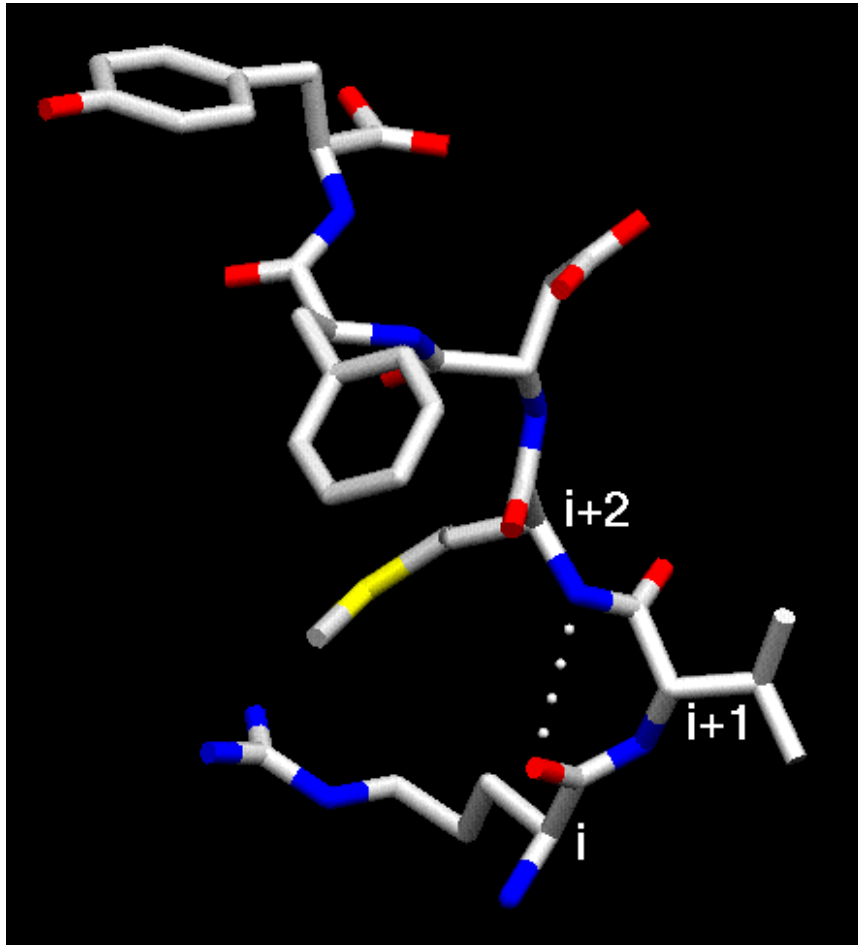
(E)



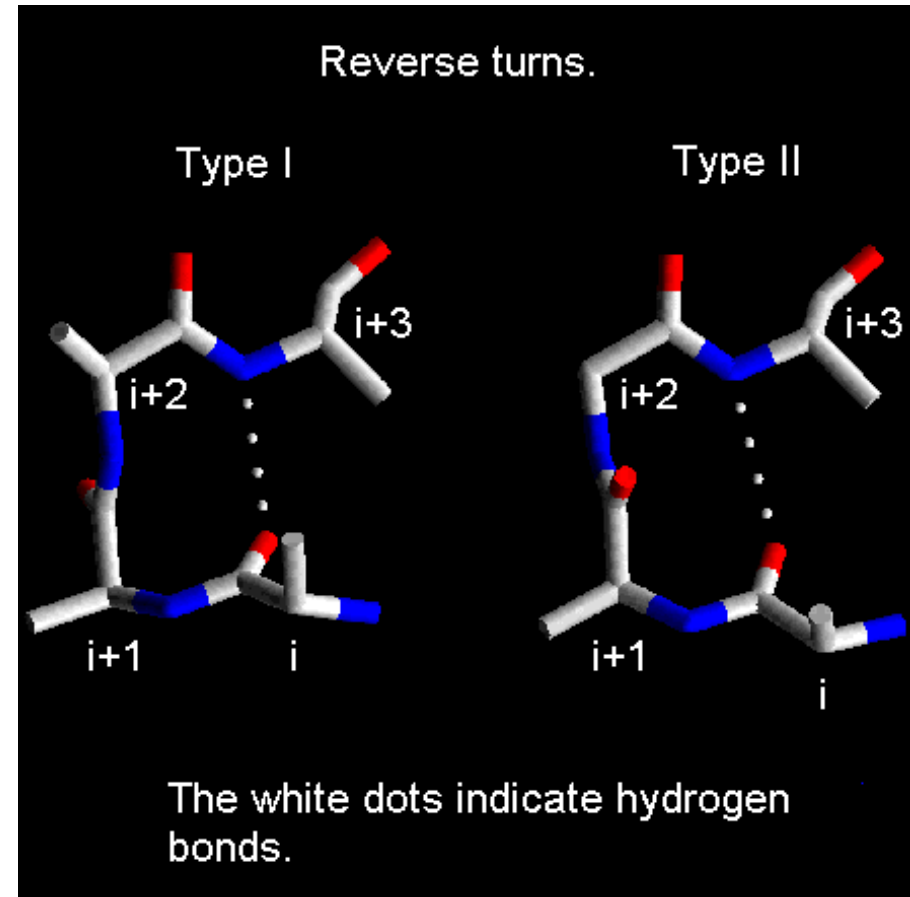
(F)

β sheet

©1998 GARLAND PUBLISHING



γ -zwrot ($\phi_{i+1} = -79^\circ$, $\psi_{i+1} = 69^\circ$)



β -zwroty

Typy β -zwrotów w białkach

Table 1. Frequency and mean dihedral angles for standard β -turn types

Turn type	Ramachandran nomenclature ^a	No. of turns		Mean dihedral angles ^d			
		b	c	$\phi(i+1)$	$\psi(i+1)$	$\phi(i+2)$	$\psi(i+2)$
I	$\alpha_R \alpha_R$	1,231	1,419	-64 (-60)	-27 (-30)	-90 (-90)	-7 (0)
II	$\beta \gamma_L$	405	489	-60 (-60)	131 (120)	84 (80)	1 (0)
VIII	$\alpha_R \beta$	325	451	-72 (-60)	-33 (-30)	-123 (-120)	121 (120)
I'	$\alpha_L \gamma_L$	127	142	55 (60)	38 (30)	78 (90)	6 (0)
II'	$\epsilon \alpha_R$	90	100	60 (60)	-126 (-120)	-91 (-80)	1 (0)
VIa1	$\beta \alpha_R$	15	17	-64 (-60)	142 (120)	-93 (-90)	5 (0)
VIa2	$\beta \alpha_R$	5	5	-132 (-120)	139 (120)	-80 (-60)	-10 (0)
VIb	$\beta \beta$	35	35	-135 (-135)	131 (135)	-76 (-75)	157 (160)
IV		1,666	1,241	-61	10	-53	17
TOTAL		3,899					

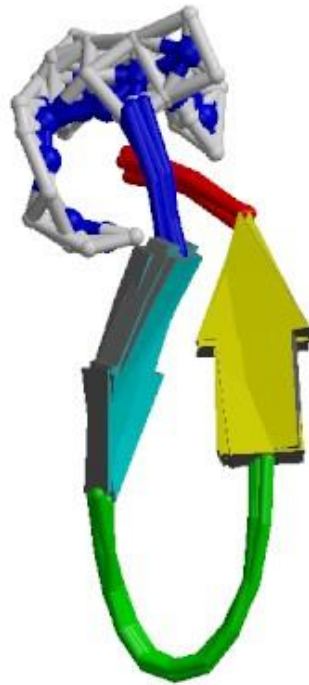
^a Ramachandran nomenclature for turn type as in Wilmot and Thornton (1990). The nomenclature describes the regions of the Ramachandran plot occupied by residues $i+1$ and $i+2$ of the turn.

^b Using normal cutoffs of 30° for deviation from standard angles, with one angle allowed to deviate by 45°.

^c Allowing up to 40° deviation from standard angles, with one angle allowed to deviate by 50°.

^d The idealized ϕ, ψ values as determined by Lewis et al. (1973) are given in parentheses after the averaged values determined from the data set. The values for the type VI turns are taken from Richardson (1981). Types VIa1 and VIa2 are the two subclasses of type VIa turns identified by Richardson (1981).

Odmiana β –zwrotu. Częsty wzór, zwykle złożony z 3 aminokwasów, łączący sąsiadujące, antyrównoległe β -kartki

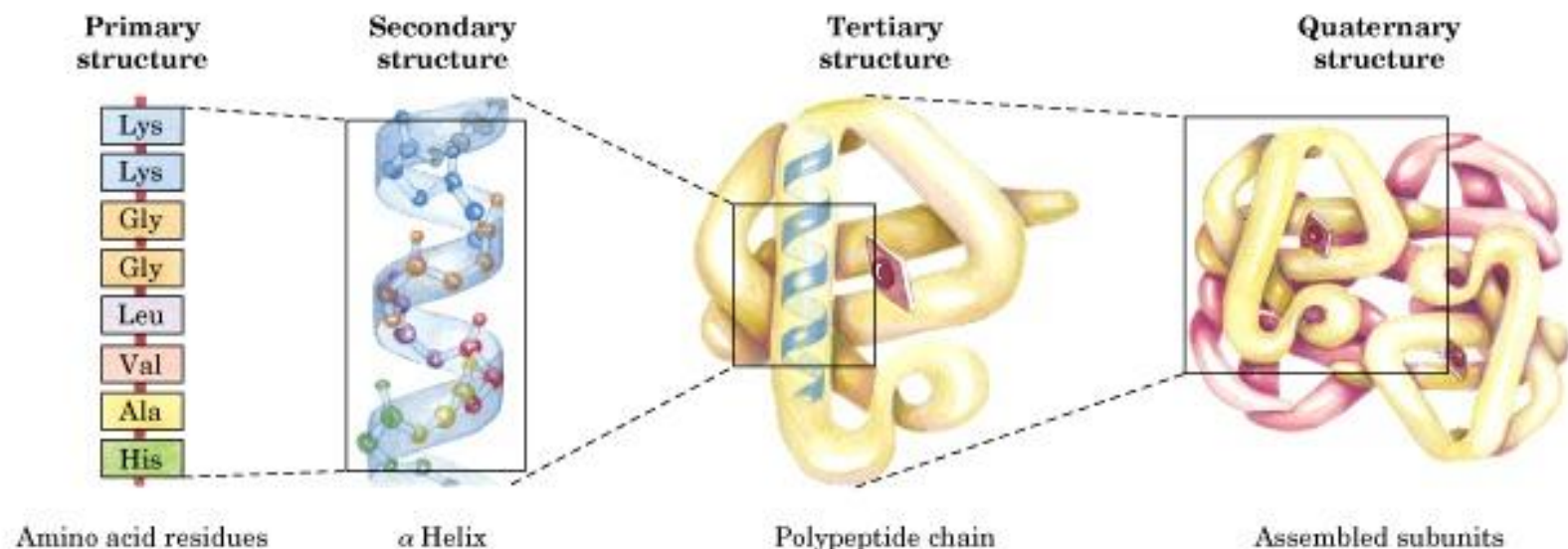


Jeśli każda reszta aminokwasowa może przybierać tylko trzy różne położenia przestrzenne. To dla białka o długości 100 aminokwasów, całkowita ilość struktur jakie może przyjąć wynosi 3^{100} , czyli ok. $5 \cdot 10^{47}$.

Jeżeli czas potrzebny na przekształcenie jednej struktury w drugą wynosi 10^{-13} s, to całkowity czas potrzebny na ustalenie optymalnej struktury wyniósłby $5 \cdot 10^{47} \cdot 10^{-13}$ s ($1,6 \cdot 10^{27}$ lat).

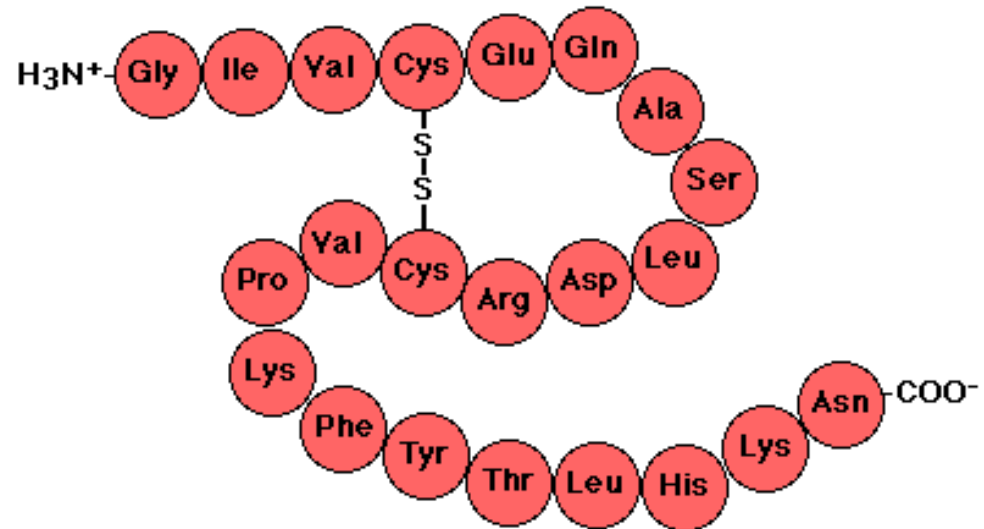
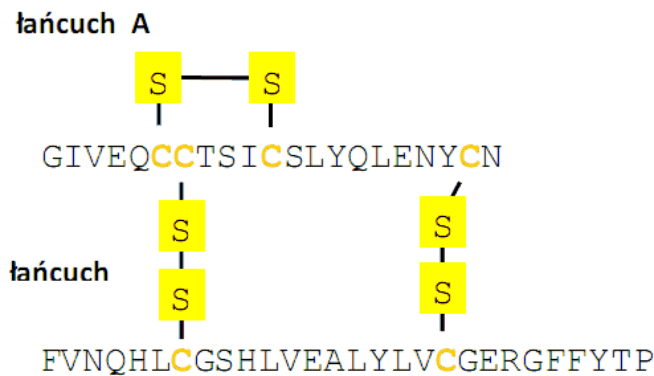
Tymczasem rzeczywisty czas potrzebny na zwinięcie się białka liczony jest w mikrosekundach.

Anfinsen (1961) – cała informacja potrzebna białku do przyjęcia ostatecznej konformacji zakodowana jest w jego strukturze pierwszorzędowej



Stabilizacja struktury przestrzennej

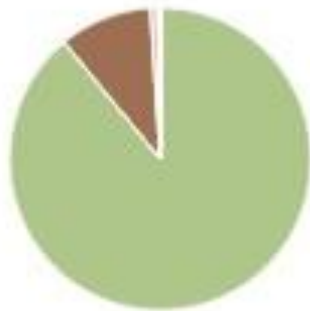
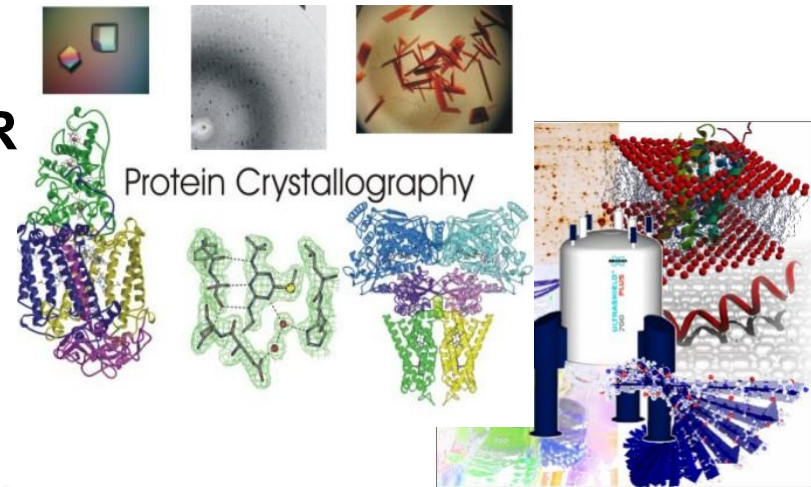
- wiązania wodorowe;
- mostki dwusiarczkowe
- oddziaływania elektrostatyczne;
- siły van der Waalsa



Ustalenie struktury 3D – techniki eksperymentalne

- Krystalografia rentgenowska (X-ray)
- Jądrowy rezonans magnetyczny (NMR)
- inne

Metody drogie i czasochłonne.



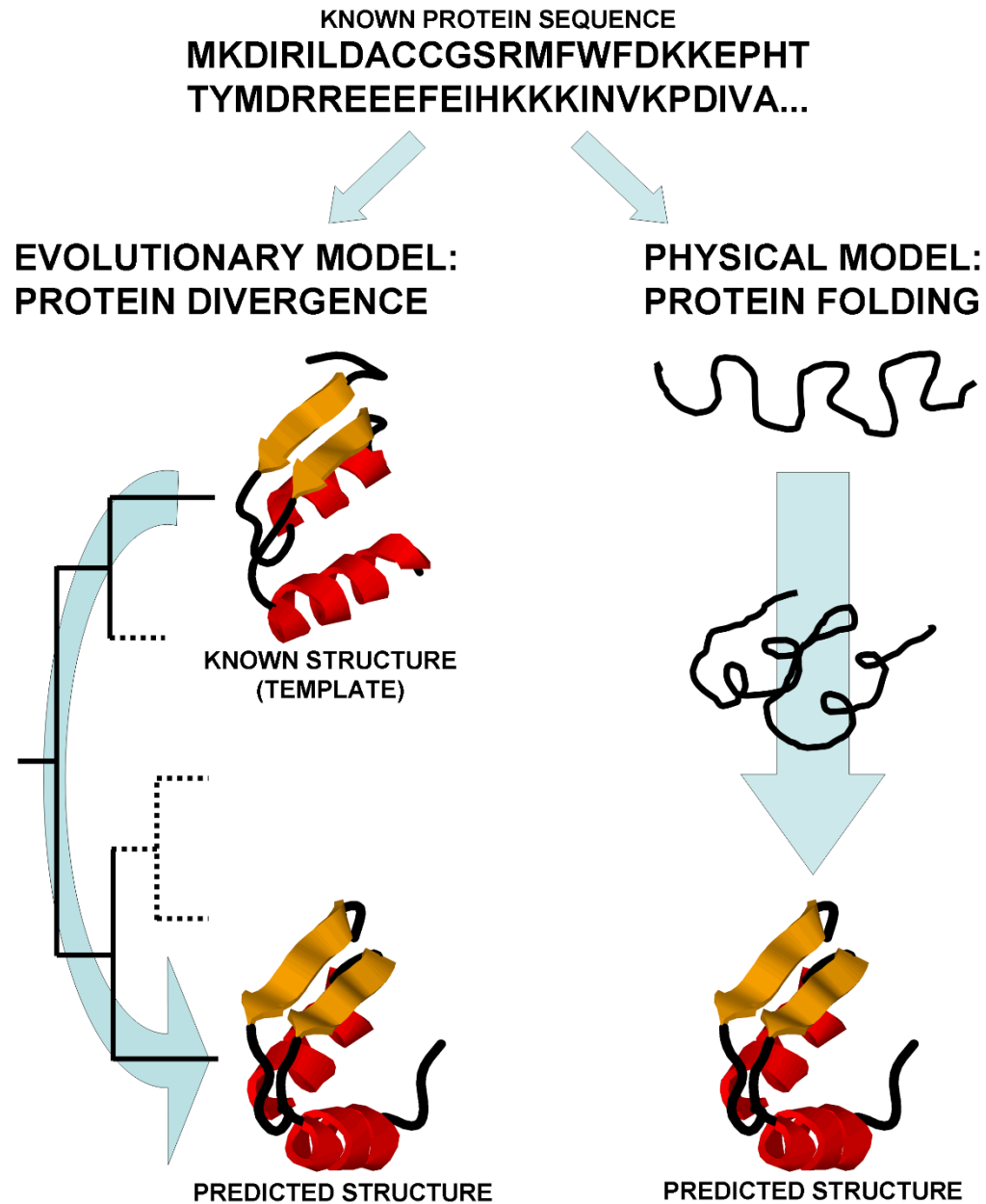
X-ray (96957)
Solution NMR (10879)
Electron Microscopy (781)
Solid-State NMR (76)
Hybrid (76)
Electron Crystallography (47)
Neutron Diffraction (47)
Fiber Diffraction (38)
Solution Scattering (32)
Other (24)

Podejście fizyczne – **szkoła boltzmannowska**

Modelowanie zwijania białka (procesu poszukiwania przez łańcuch konformacji o najniższej energii swobodnej, który w komórkach trwa zaledwie ułamki sekundy) korzystając z praw fizyki statycznej.

Podejście ewolucyjne – **szkoła darwinowska**

Rekonstrukcja procesu powstania sekwencji i struktury białka na drodze ewolucji (przyrodzie zabiera to miliony lat)



Klasyfikacja rodzin białkowych

SCOP = *Structural Classification Of Proteins*

Hierarchiczny schemat klasyfikacji obejmujący 4 poziomy:

Rodzina – grupa białek powiązana strukturalnie, ewolucyjnie i funkcjonalnie;

Superrodzina – zbiór rodzin o podobnej strukturze i funkcji;

Zwój – wspólna topologia na większym fragmencie łańcucha

Klasa – grupa zwojów charakteryzowanych strukturą 2-rzędową.

Klasy:

α (głównie α -helisy),

β (głównie β -kartki),

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

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

multidomain proteins (niehomologiczne białka, z różnymi zwojami)

Przykładowa hierarchia SCOP

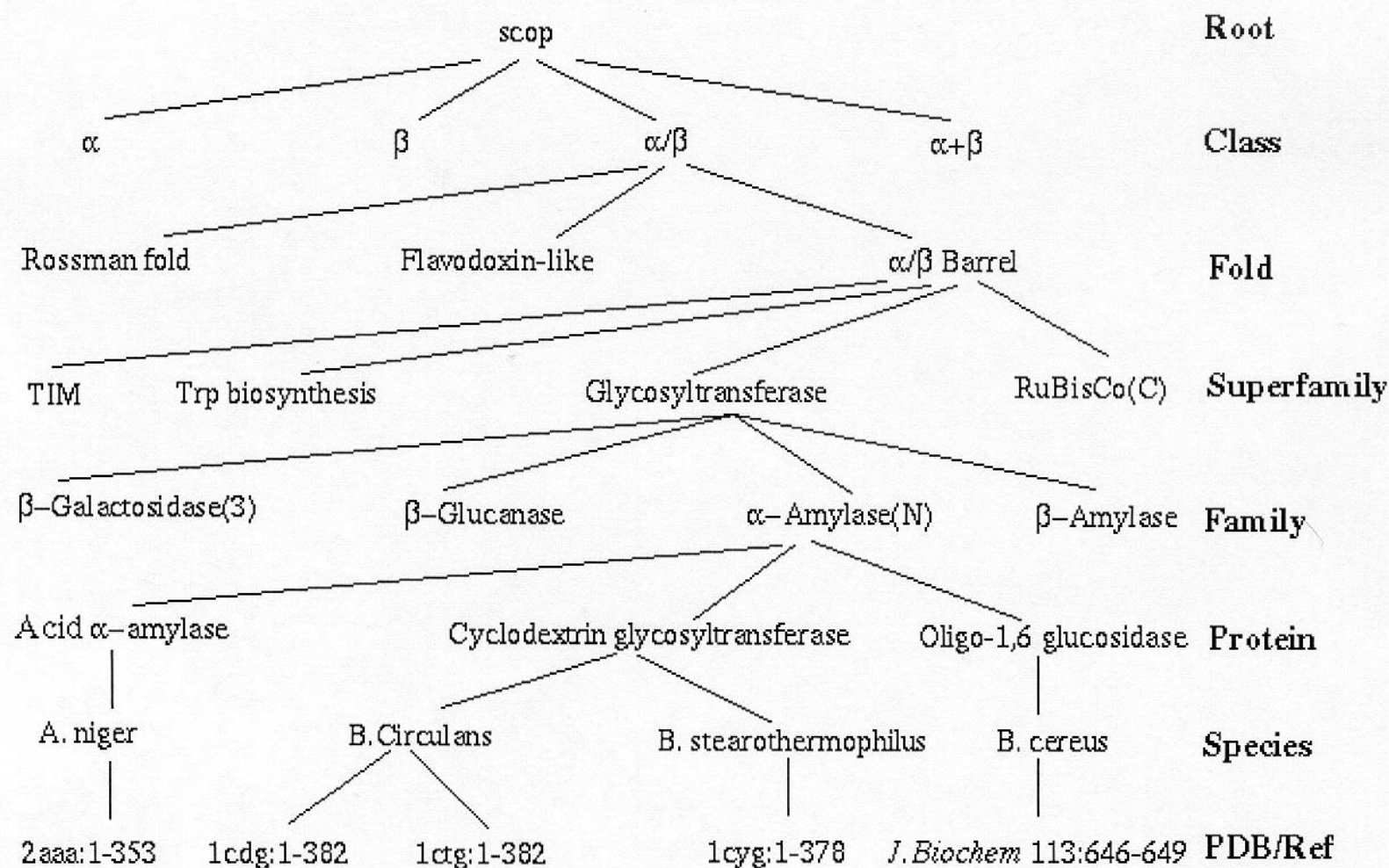
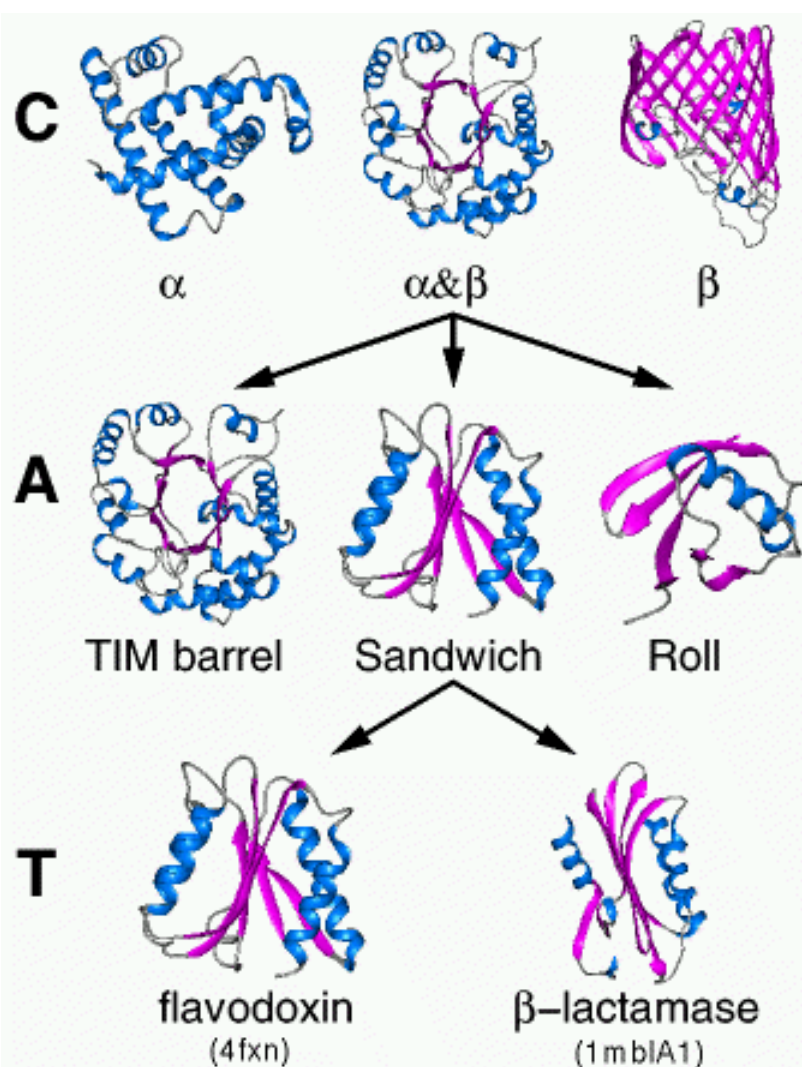


Figure 15.1: A Sample SCOP Hierarchy

Class, Architecture, Topology, Homologous superfamily

Cztery poziomy hierarchii:

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



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



Welcome to CATH

CATH is a manually curated classification of protein domain structures. Each protein has been chopped into structural domains and assigned into homologous superfamilies (groups of domains that are related by evolution). This classification procedure uses a combination of automated and manual techniques which include computational algorithms, empirical and statistical evidence, literature review and expert analysis.

[Find out more about CATH >>](#)

New in CATH v3.4

CATH v3.4 is built from 104,238 PDB chains. We have added the following data since v3.3:

- 49 folds (total 1,282)
- 163 superfamilies (total 2,549)
- 1,311 sequence families (total 11,330)
- 24,232 domains (total 152,920)

[Download CATH data >>](#)

Using CATH

[Search](#)

[Browse](#)

[Download](#)

[Tutorials](#)

[Introduction to CATH](#)

[FAQ](#)

CATH Tools

[Find My Sequence](#)

[Find My Structure](#)

[Linking to CATH](#)

About CATH

[Release Statistics](#)

[Glossary](#)

[CATH Team](#)

[References](#)

Przewidywanie struktury drugorzędowej białek

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

- jest etapem tworzenia struktury przestrzennej i domenowej**
- przydatna koncepcja dla zrozumienia struktury**
- ma związek z funkcją białka**
- przydatna w algorytmach przewidywania struktury przestrzennej (definiuje obszary na wzorcach)**

1. Pierwsza generacja: statystyki pojedynczych aminokwasów

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

skuteczność: niska

2. Druga generacja: statystyki w oknach

np.: ALB, GOR III, etc

skuteczność: ~60%

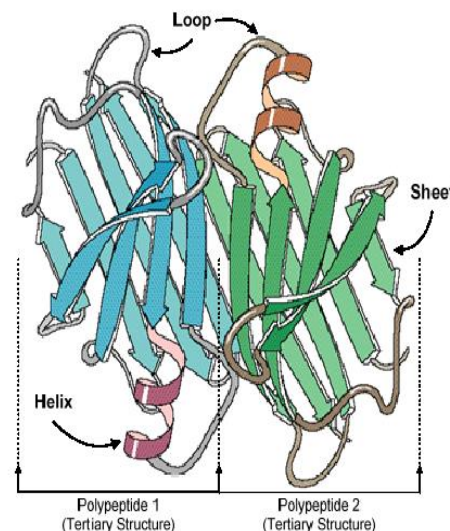
3. Trzecia generacja: oddziaływania długodystansowe, metody homologiczne

np.: PHD

skuteczność: ~70%

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

Zwykle są to metody porównawcze bazujące na homologii.



Strategia przewidywania struktury drugorzędowej:

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

W kontekście poszukiwania struktury.

Zalety:

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

Wady:

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

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

Przykład:

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

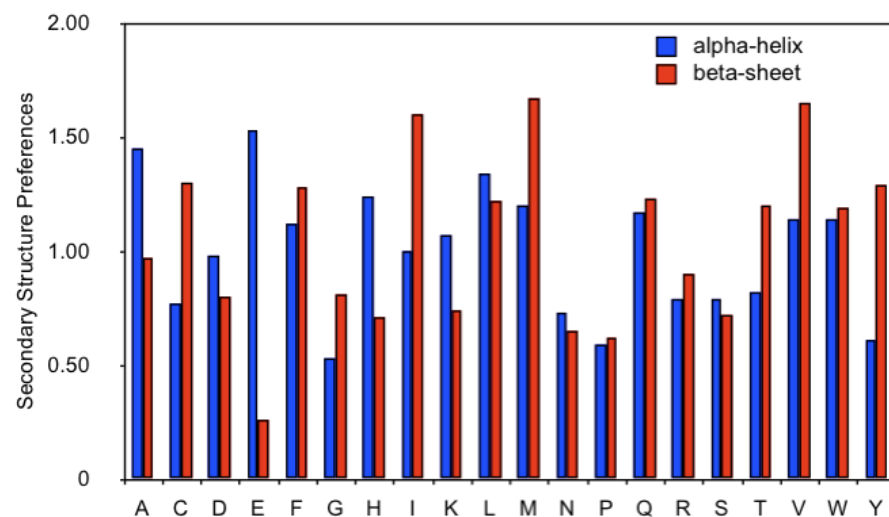
- $P(\text{Ala in Helix}) = 50/500 = 0.1$

$$P(\text{Ala}) = 100/2000 = 0.05$$

$$\text{Helix propensity (PA) of Ala} = P(\text{Ala in Helix})/P(\text{Ala}) = 0.1/0.05 = 2$$

Tabela preferencji

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

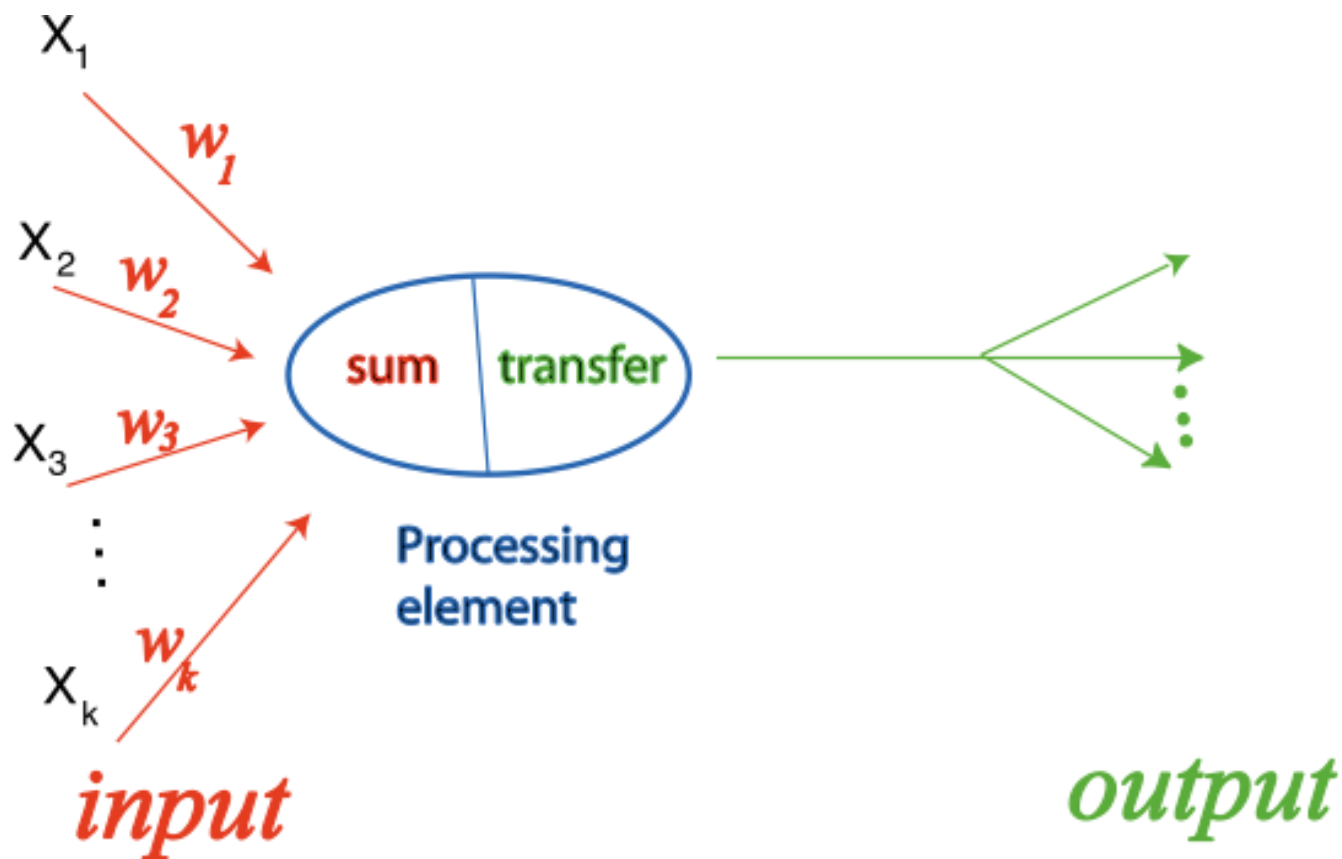


1. Individual propensity of amino acid R to be in secondary structure S. – same idea as in Chou – Fasman
2. Contribution of 16 neighbors.
 - take the window of radius 8 around the residue in question (8 before and 8 after the residue)
 - for each residue in the window consider it's contribution to the conformation of the middle residue and this it's value to PH, PS, PT.
 - like in Chou-Fasman the values of all contributions are based on statistics.

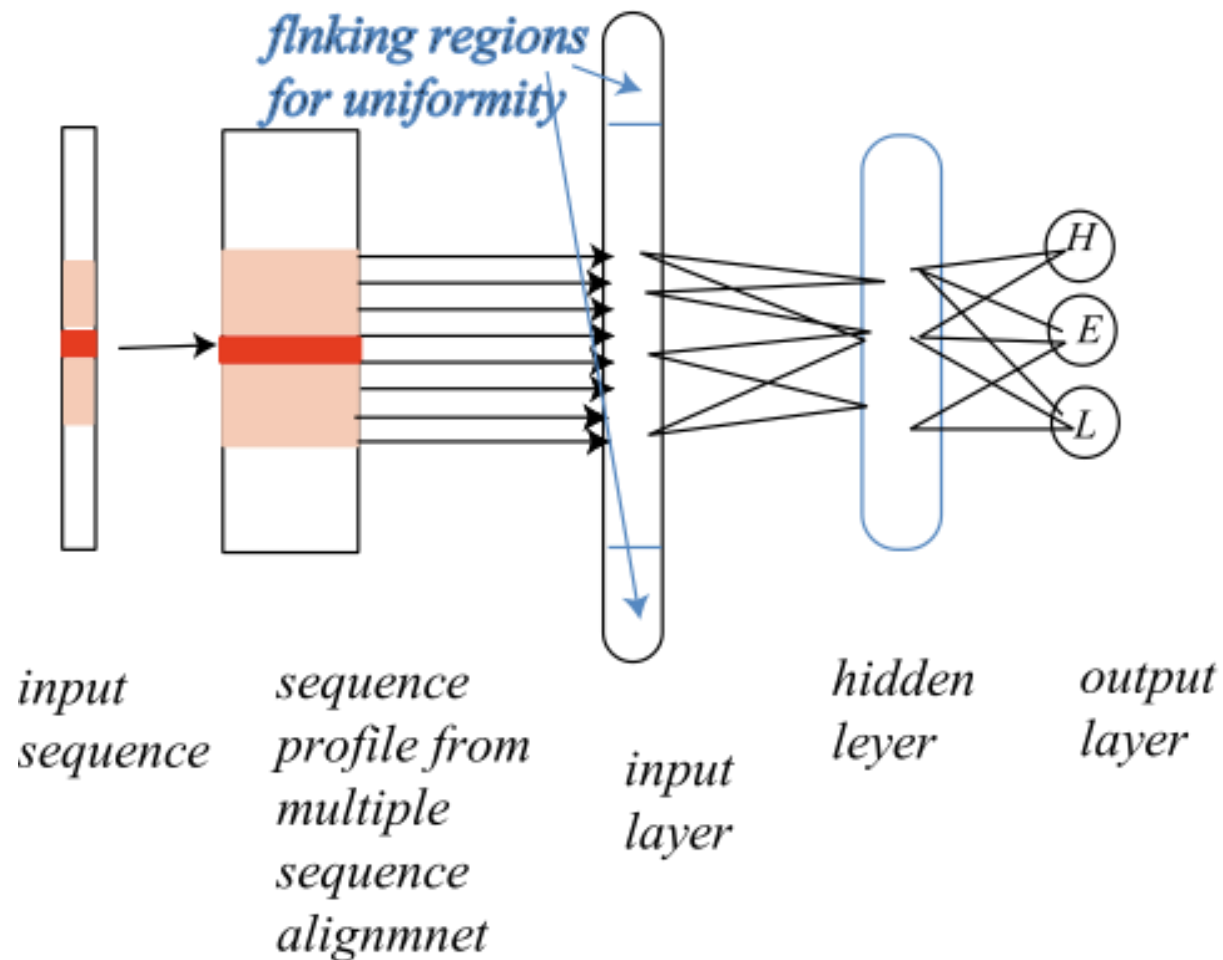
- **Pojedyncza sekwencja zastąpiona uliniowaniem spokrewnionych (homologicznych) sekwencji**
- **Profil**

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

- **Wartości binarne na wejściu sieci zastąpione wartościami rzeczywistymi z przedziału [0,1]**
- **Poprawa jakości predykcji (z 65% do >70%)**



Sieci neuronowe (2)



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

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

Wejście:

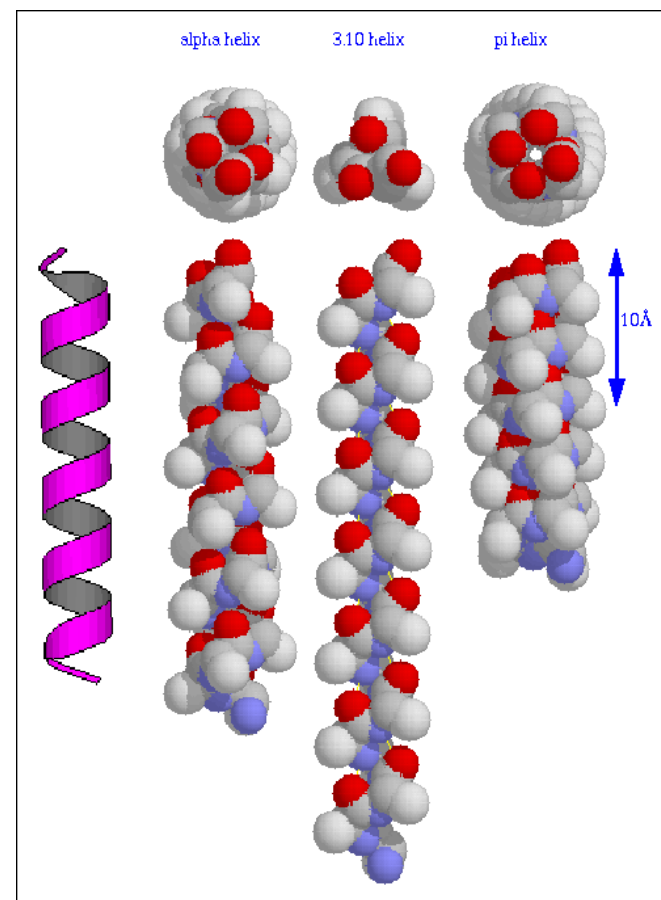
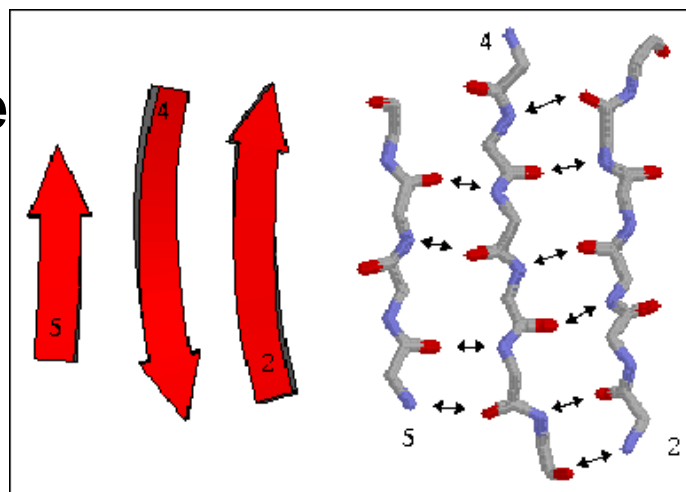
- Informacja na temat w sąsiednich aminokwasów (w nieparzystych) – tzw. okno wejściowe
- Kodowanie ortogonalne aminokwasów:
 - wektor o wymiarze 20
 - na jednej pozycji 1, a na pozostałych 0
- $(20 \times w)$ elementów wejściowych

Wyjście:

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

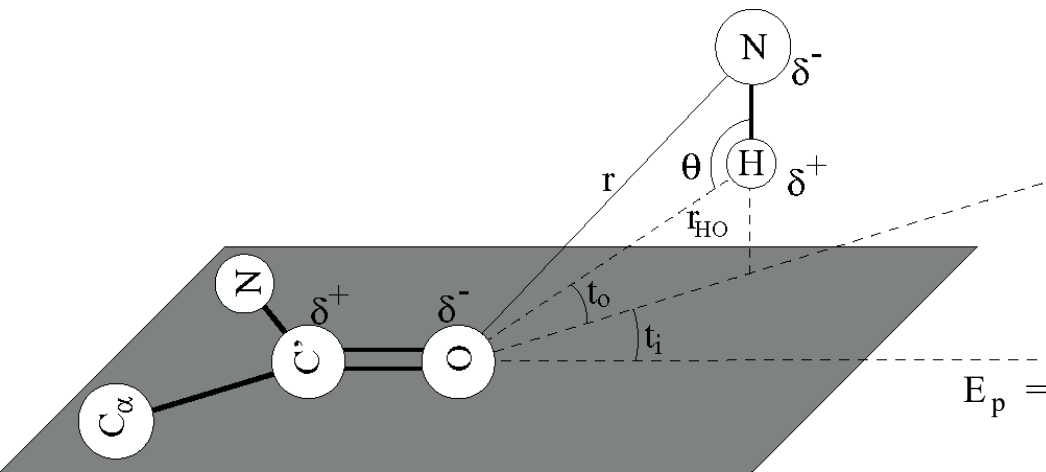
8 kategorii (DSSP):

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



STRIDE – Empirical Hydrogen Bond Calculation

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




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

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

$$E_t = \begin{cases} [0.9 + 0.1 \sin(2t_i)] \cos(t_0) & 0^\circ < t_i \leq 90^\circ \\ K_1 [K_2 - \cos^2(t_i)] \cos(t_0) & 90^\circ < t_i \leq 110^\circ \\ 0 & 110^\circ \leq t_i \end{cases}$$

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

Expasy – odsyłacze do licznych narzędzi

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

Secondary structure prediction

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


Tertiary structure

Tertiary structure analysis

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

Tertiary structure prediction

Homology modeling

- [SWISS-MODEL](#)  - An automated knowledge-based protein modelling server
- [CPHmodels](#) - Automated neural-network based protein modelling server
- [ESyPred3D](#) - Automated homology modeling program using neural networks
- [Cora3d](#) - Automatic modelling of protein three dimensional structure

<http://expasy.org/tools/>

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

YASPIN - <http://www.ibi.vu.nl/programs/yaspinwww/>

SSPRO - <http://download.igb.uci.edu/sspro4.html>

PROTEUS - <http://wks16338.biology.ualberta.ca/proteus/>

YASPIN Secondary Structure Prediction results for job Untitled

[Back to YASPIN main page](#)

[Download the YASPIN prediction results file here](#)

[Download the PSI-BLAST generated PSSM](#)

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

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

You are using the dssp-trained NN.

Query Name: uploaded.ckp

Sequence Length: 350

Ruler :10.....20.....30.....40.....50.....60
Sequence :	GNAAAAKKGSQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVMLV
Prediction:	-----NNNNNNNNNNNNNNNNNNNN-----EEEEEE-----EEEE
Overall :	995432468413462777552202312301347657212155124974232662123888
Helix :	0010000000059999999999999998600000000010000000000000000000
Strand :	00
Coil :	99899999994000

Comprehensive Secondary Structure Predictions

DOCUMENTATION

CONTACT & DOWNLOAD

Summary:

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

H = Helix

E = Beta Strand

C = Coil

Line 1 = sequence (single letter IUPAC code, 60 characters per line)

Line 2 = secondary structure (H, E or C)

Line 3 = confidence score (0-9, 0 = low, 9 = high)

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

Predicted Secondary Structure:

1 GNAAAAKKGSGEQSVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLTGTSFGVRMLV 60
CCEEEEEEBCCCBBBBE
987777777666655555676555566777777888885899999977997699999

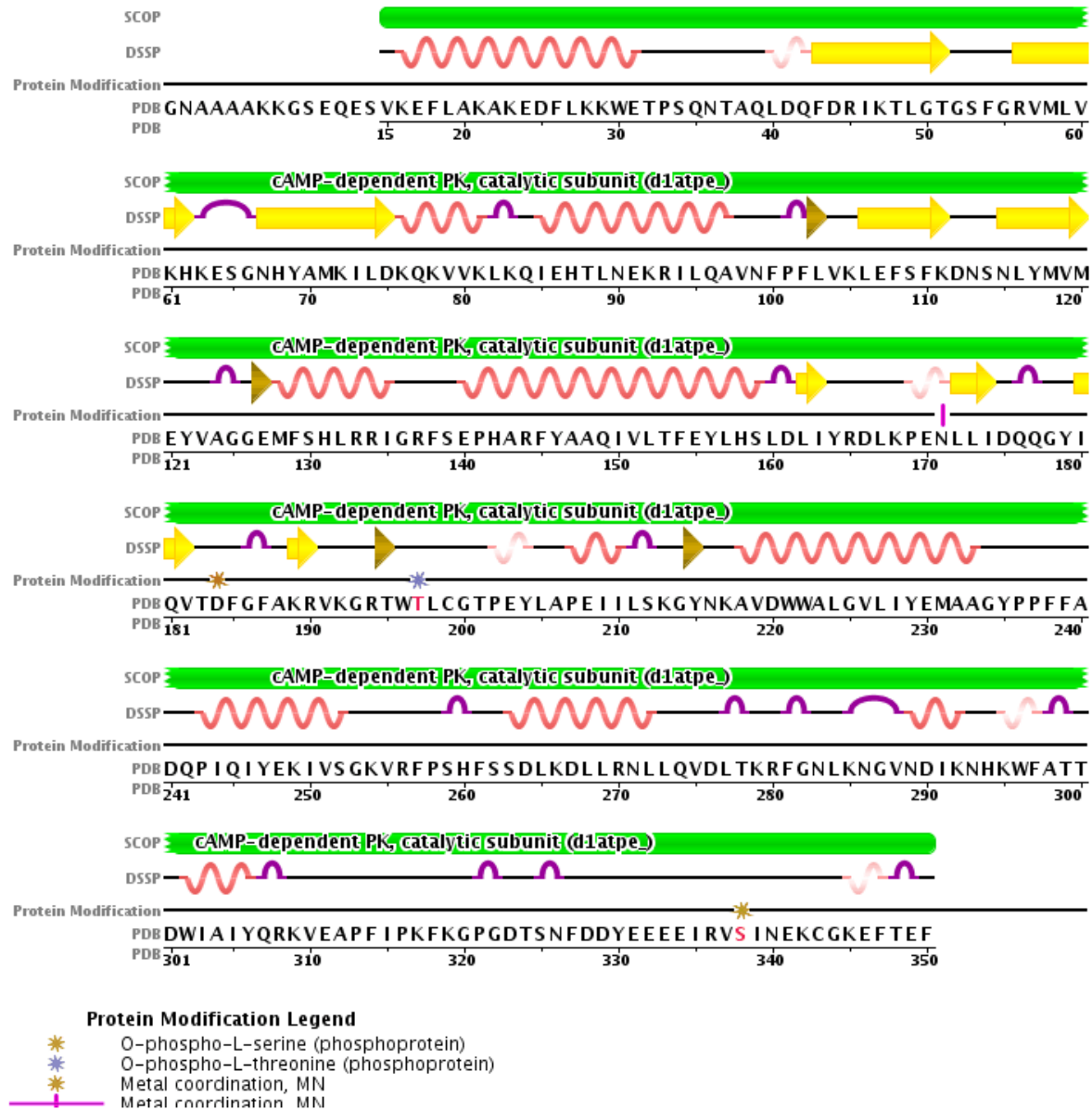
61 KHKESGNHYAMKILDQKVVLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLVMVM 120
EECCCCEEEEEEBCHHNNNNNNNNNNNNNNNNNNNNNNCCCCEEEEEEEBCCCCBBBEE
978999699999994689987999999999999999858887999999996998689999

121 ETVAGGEMFSLRRIGRFSPEHARFYAAQIVLTPEYLHSLDLIYRDLPKNLLIDQQGYI 180
EECCCCHHNNNNNNHHCCCCHHNNNNNNNNNNNNNNNNNNNNCCBBBBCCCCCBBBCCCCBE
986987789999995777579999999999999999986776898578876788588889

181 QVTDFGFAKRVKGRWTWLCTPEYLAPBIILSKGYNKAVDWALGLVIYEEMAAGYPFFFA 240
BBBBCCCCCECCCCCCCCCCCCCCCCCCCCCCCCCHNNNNNNNNNNNNNNCCCCCCCC
999866568866775787567755564556788886388999999999998557888887

241 DQPIQIYEKIVSGKVRFPSPHFSSDLKDLLRNLLQVDLTRKFNLKNGVNDIKNHKWPFATT 300
CSHHNNNNNNHHCCCCCCCCCHNNNNNNNNHHCCCCCCCCCCCCCHNNNNHHCCCCC



**białko 1ATP**

- PredictProtein-PHD (72%)

- <http://www.predictprotein.org/>

- Jpred (73-75%)

- <http://jura.ebi.ac.uk:8888/>

- PREDATOR (75%)

- http://www.embl-heidelberg.de/cgi/predator_serv.pl

- PSIPred (77%)


- <http://insulin.brunel.ac.uk/psipred>

1. **Znaczenie struktury makrocząsteczek**
2. **Poziomy organizacji strukturalnej**
3. **PDB (Protein Data Bank)**
4. **Klasyfikacja rodzin białkowych**
5. **Przewidywanie struktury drugorzędowej białek**

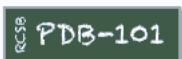


PDB




PDB – baza struktur molekularnych





PROTEIN DATA BANK



PDB-101

A MEMBER OF THE **PDB**
An Information Portal to Biological Macromolecular Structures
As of **Tuesday May 17, 2011 at 5 PM PDT** there are **73153** Structures   | [PDB Statistics](#)

Contact Us | Print

MyPDB Hide

Login to your Account
Register a New Account

Home Hide

News & Publications
Usage/Reference Policies
Deposition Policies
Website FAQ
Deposition FAQ
Contact Us
About Us
Careers
External Links
Sitemap
New Website Features

Deposition Hide

All Deposit Services
Electron Microscopy
X-ray | NMR
Validation Server
BioSync Beamlines/Facilities
Related Tools

Search Hide

Advanced Search
Latest Release
New Structure Papers
Sequence Search
Chemical Components
Unreleased Entries
Browse Database
Histograms

Tools Hide

Download: Entries | Ligands
Compare Structures
FTP Services
File Formats
Services: RESTful | SOAP
Widgets

A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the [wwPDB](#), the RCSB PDB curates and annotates PDB data according to agreed upon standards.

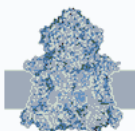
The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

[Hide Welcome Message](#)

Featured Molecules Hide

Structural View of Biology

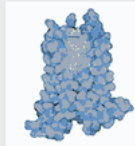
List View of Archive By: [Title](#) | [Date](#) | [Category](#)



Molecule of the Month:
Cytochrome bc1

Cells are masters at squeezing every drop of energy out of their food. They disassemble the molecules in food atom by atom, driving a variety of unusual energy transformations in the process. At the end, all of the hydrogen atoms have been separated from the food molecules and are used to turn the rotary motor of **ATP synthase**. To do this, the electrons are stripped from these hydrogen atoms and used to power huge protein pumps that transport protons across a membrane. These protons then power the rotation of ATP synthase as they return to their original positions.

[Full Article...](#)



Protein Structure Initiative Featured Molecule:
A2A Adenosine Receptor

The structure of A2A adenosine receptor with a powerful agonist reveals a GPCR in action.

[Full Article](#) | [PSI Featured Molecule Archive](#) | [PSI Structural Biology Knowledgebase](#)


Latest Structures Hide

Customize This Page

New Features Hide

Sequence Page: Improved Layout

Latest features released:

Website Release Archive: 

RCSB PDB News Hide

[Weekly](#) | [Quarterly](#) | [Yearly](#)

2011-05-17

Validate Entries Before Deposition

Use the [Validation Server](#) to create reports (plain text and PDF) that include results of format, geometric, and experimental data checking.
[more...](#)

- Educational Resources to Explore a Structural View of Biology
- Newsletter published
- Rutgers Day (April 30)
- New Publication

wwPDB News Hide

Statement on Retraction of PDB Entries

2011-04-13

Announcement: PDB Archive Version 4.0 to be Released June 2011

- Special Symposium Celebrating the 40th Anniversary of the PDB
- Full wwPDB News

www.pdb.org

Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	90814	1622	4517	4	96957
NMR	9603	1119	225	8	10955
ELECTRON MICROSCOPY	567	29	185	0	781
HYBRID	70	3	2	1	76
other	165	4	6	13	188
Total	101219	2777	4935	26	108957

wg stanu na 26.V.2015 – 108957 struktur (dwa lata wcześniej 89740)

źródło: <http://www.pdb.org/pdb/statistics/holdings.do>,

ale tylko 61201 różnych struktur (50580)

(<http://www.pdb.org/pdb/statistics/clusterStatistics.do>)

tylko 1375 różnych zwojów wg klasyfikacji CATH (1375)

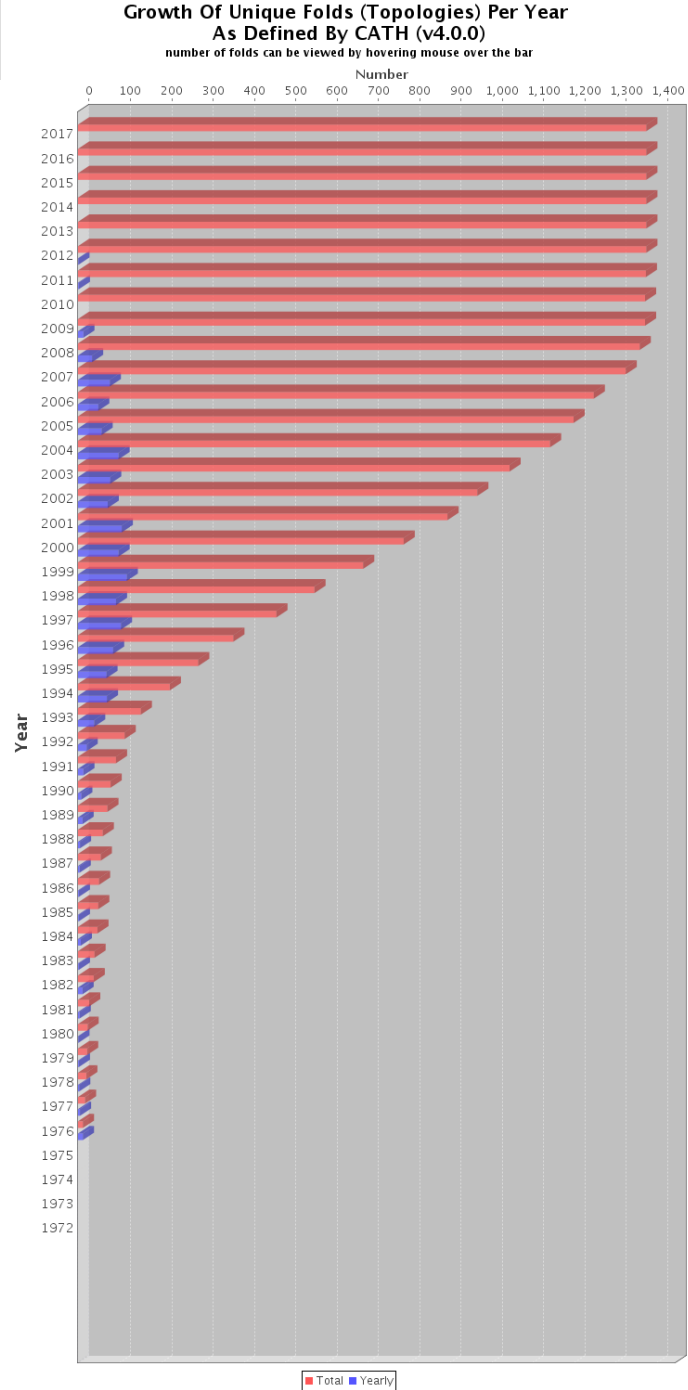
(<http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=fold-cath>)

podczas, gdy znamy 47452313 sekwencji (33106277)

(<http://www.ebi.ac.uk/uniprot/TrEMBLstats/>)

w tym 548454 sekwencji z weryfikowanych (539829)

(<http://web.expasy.org/docs/relnotes/relstat.html>)



Plik PDB - nagłówek

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

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

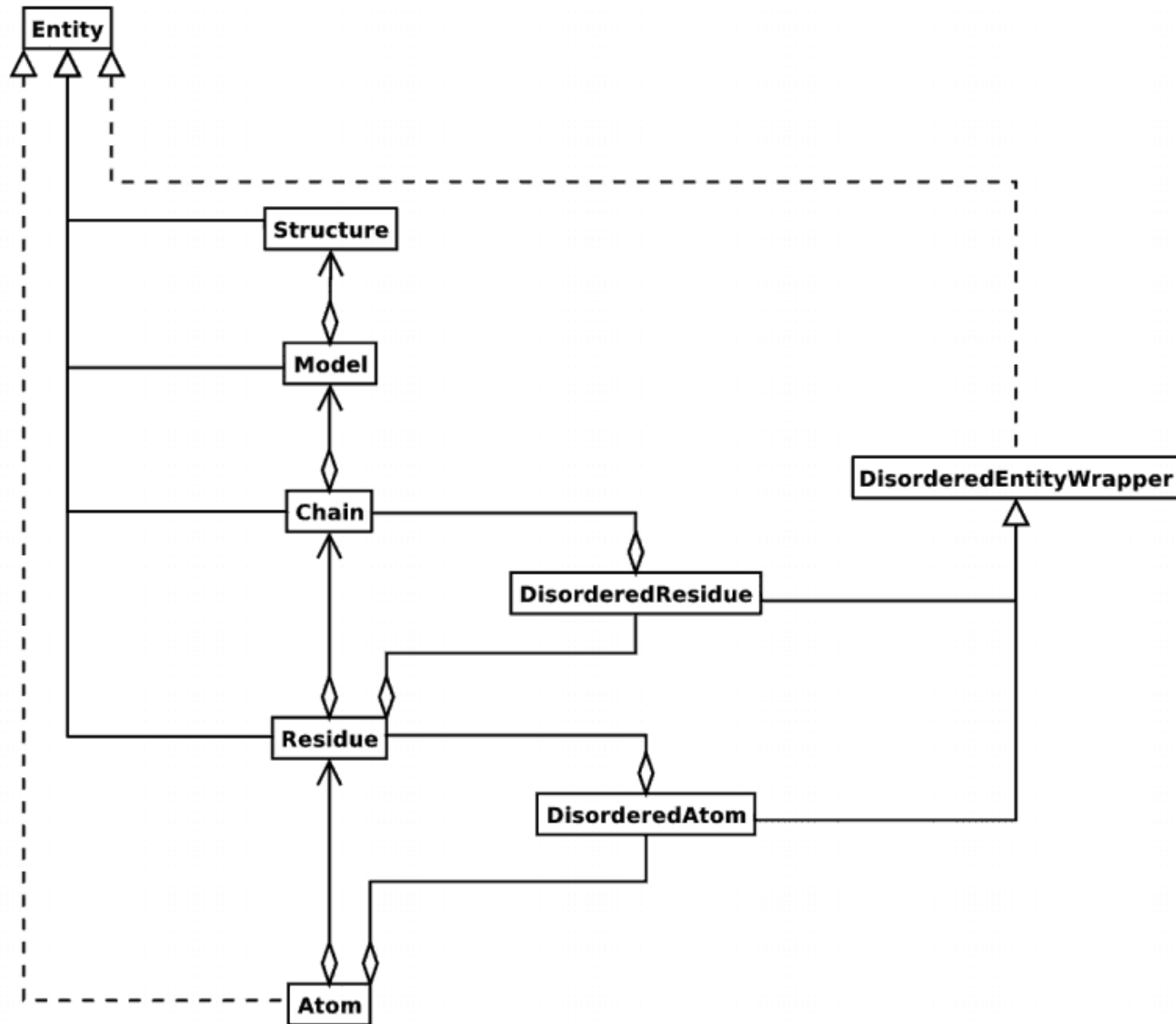
Plik PDB – dane

ATOM	1	N	TRP	A	203	30.156	-4.908	37.767	1.00	50.81	N
ATOM	2	CA	TRP	A	203	30.797	-4.667	36.431	1.00	49.96	C
ATOM	3	C	TRP	A	203	30.369	-3.337	35.766	1.00	49.18	C
ATOM	4	O	TRP	A	203	29.315	-3.238	35.147	1.00	49.27	O
...
ATOM	1	N	TRP	A	203	30.156	-4.908	37.767	1.00	50.81	N
ATOM	8	CD2	TRP	A	203	29.980	-5.876	32.984	1.00	42.73	C
ATOM	9	NE1	TRP	A	203	31.956	-5.191	31.177	1.00	41.45	N
ATOM	10	CE2	TRP	A	203	30.704	-5.582	30.805	1.00	41.23	C
ATOM	11	CE3	TRP	A	203	28.657	-6.305	32.877	1.00	46.48	C
ATOM	12	CZ2	TRP	A	203	30.149	-5.705	30.539	1.00	46.06	C
ATOM	13	CZ3	TRP	A	203	28.101	-6.431	31.622	1.00	43.08	C
ATOM	14	CH2	TRP	A	203	28.849	-6.131	30.463	1.00	45.77	C
...

Diagram illustrating the structure of a PDB file entry (ATOM) and its corresponding fields:

- Name
- Atom Number
- Atom Name
- Residue Name
- Chain ID
- Residue Number
- X
- Y
- Z
- Occupancy
- Temperature Factor

BioPython – PDBParser: diagram UML



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

Wypisuje wszystkie aminokwasy w strukturze białka 1FAT, zawierające nieuporządkowany atom C α

Zwykle wynikają z błędu w strukturze pliku PDB:

- powtórzone residuum;**
- powtórzony atom;**
- brak alternatywnej informacji dla niejednoznacznego atomu;**
- urwany łańcuch.**

- brak informacji o wiązaniach
- maksymalna liczba atomów w modelu: 99999 (pięciocyfrowe pole na numer atomu)
- maksymalna liczba łańcuchów: 26 (identyfikator jednoliterowy)

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

Atomic Coordinate Entry Format Description

Version 3.3: July, 2011

Introduction

Title Section

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

Primary Structure Section

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

Heterogen Section

- HET
- FORMUL
- HETNAM
- HETSYN

Secondary Structure Section

- HELIX
- SHEET

Connectivity Annotation Section

- SSBOND
- LINK
- CISPEP

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



Dictionary Resources

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

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

Dictionary Content and Representation

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

Data Dictionaries

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

format mmCIF -przykład

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

HEADER ISOMERASE/DNA

01-MAR-00 1EJ9



Protein Structures