

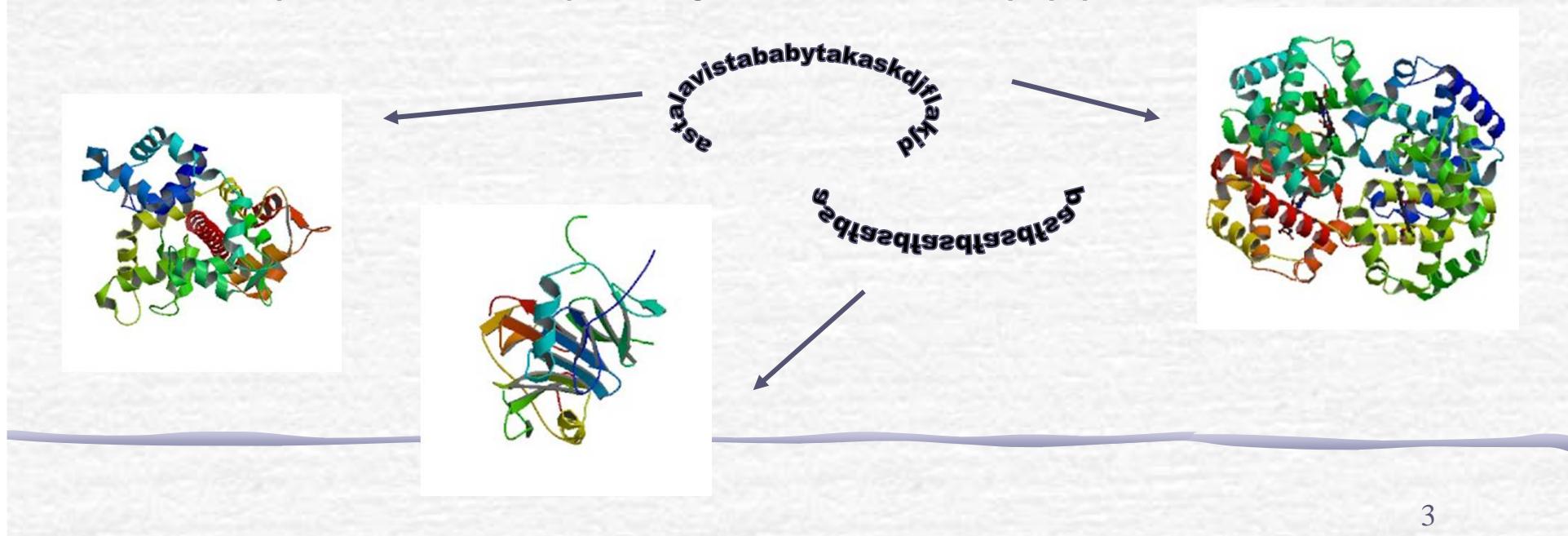
# Predikce struktury proteinů

# Obsah

- Motivace
- Základní pojmy
- Predikce elementů sekundární struktury
- Predikce celkové struktury proteinů

# Motivace

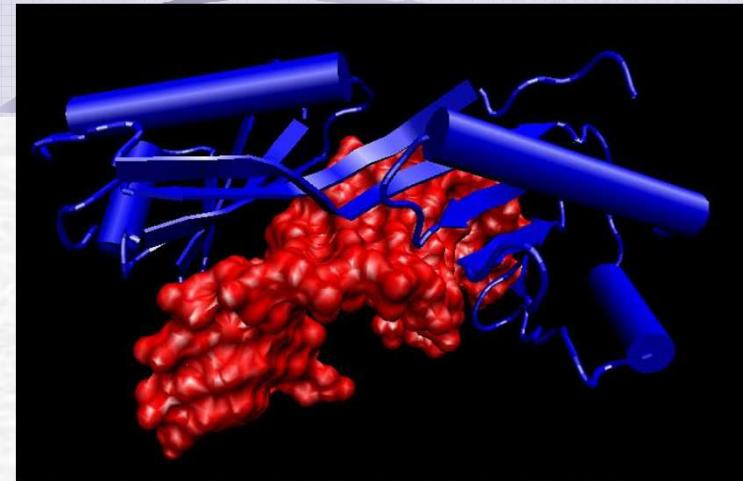
- ✓ Ze znalosti struktury proteinu je možné odhadnout jeho funkci
- ✓ Stanovit strukturu proteinu experimentálně je poměrně drahé, trvá to dlouho a výsledek nemusí být přesný
- ✓ Potřebujeme krystal proteinu!
- ✓ Rozdíl ve velikosti databází sekvencí a struktur
- ✓ Cíl: predikovat alespoň nějaké strukturní rysy proteinu



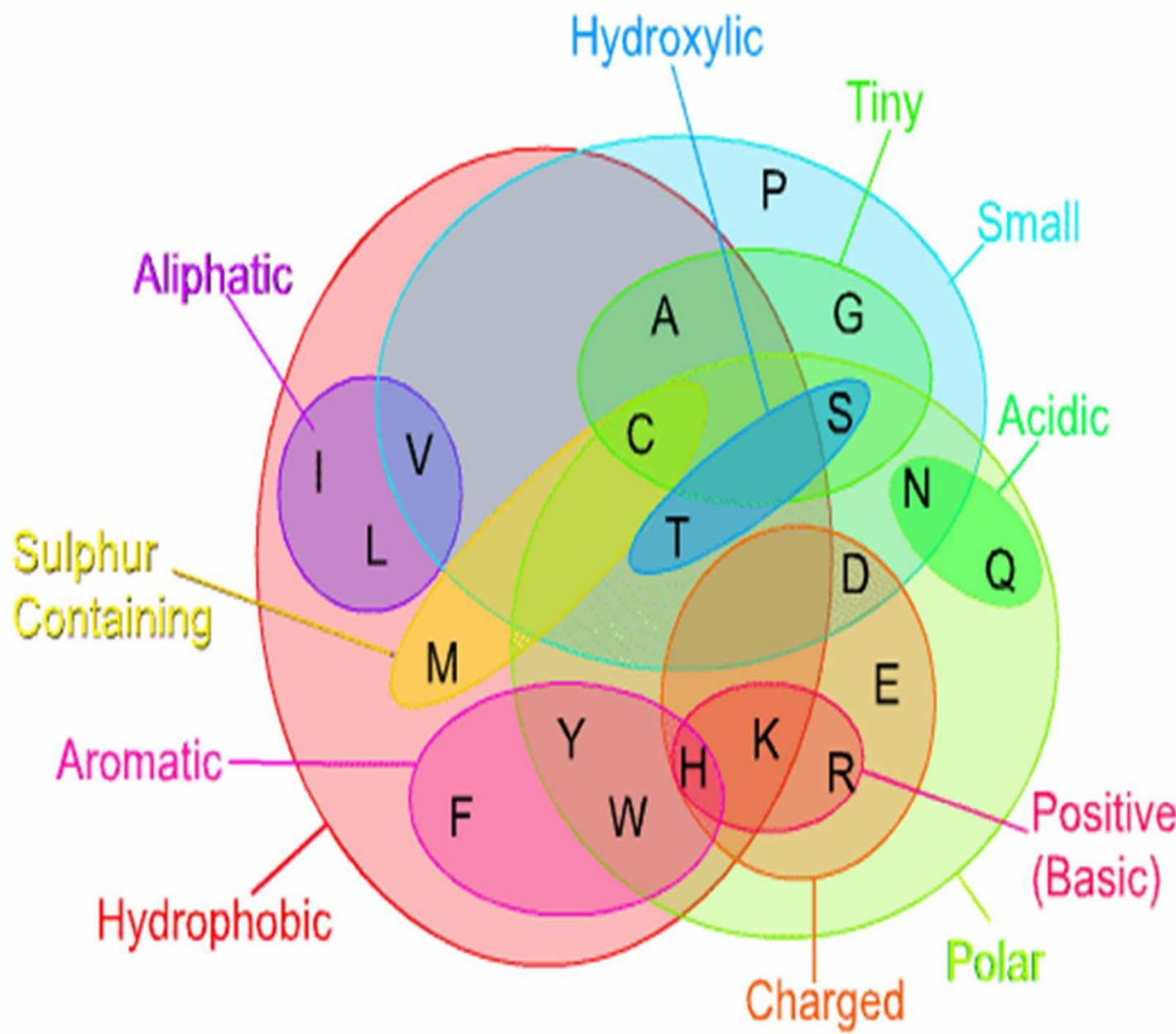
# Obsah

- Motivace
- Základní pojmy
- Predikce elementů sekundární struktury
- Predikce celkové struktury proteinů

# Proteiny



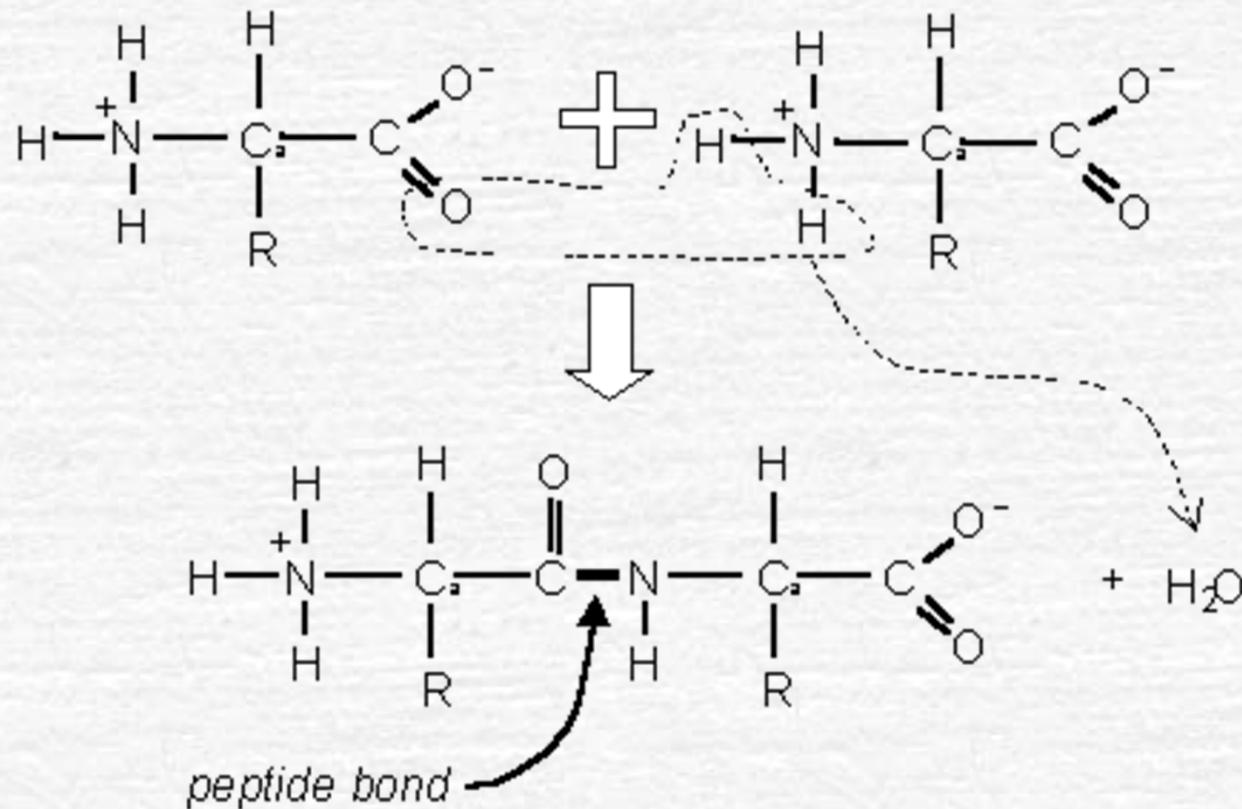
- Protein – sekvence aminokyselin, řetězec nad abecedou aminokyselin, 20 různých aminokyselin
- Sekvence aminokyselin = primární struktura proteinu
- Různé aminokyseliny mohou vytvářet různé vazby, tím ovlivňují strukturu a funkci proteinu (*hydrofobicia*)
- Primární struktura** určuje fyzikální a chemické vlastnosti proteinu, jeho prostorovou strukturu a biologickou funkci



## Amino Acids

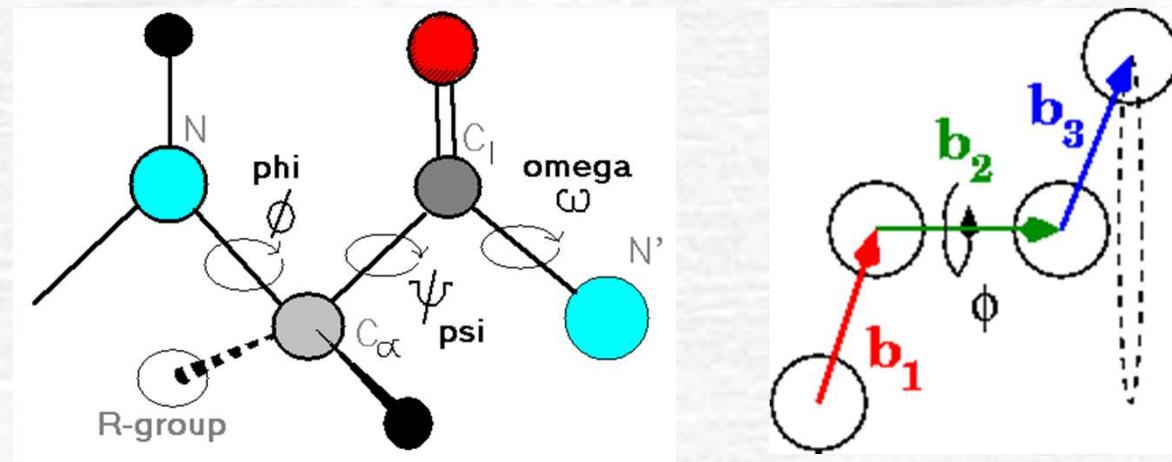
- A** alanine (ala)
- R** arginine (arg)
- N** asparagine (asn)
- D** aspartic acid (asp)
- C** cysteine (cys)
- Q** glutamine (gln)
- E** glutamic acid (glu)
- G** glycine (gly)
- H** histidine (his)
- I** isoleucine (ile)
- L** leucine (leu)
- K** lysine (lys)
- M** metioneine (met)
- F** phenylalanine (phe)
- P** proline (pro)
- S** serine (ser)
- T** threonine (thr)
- W** tryptophan (trp)
- Y** tyrosine (tyr)

# Proteiny

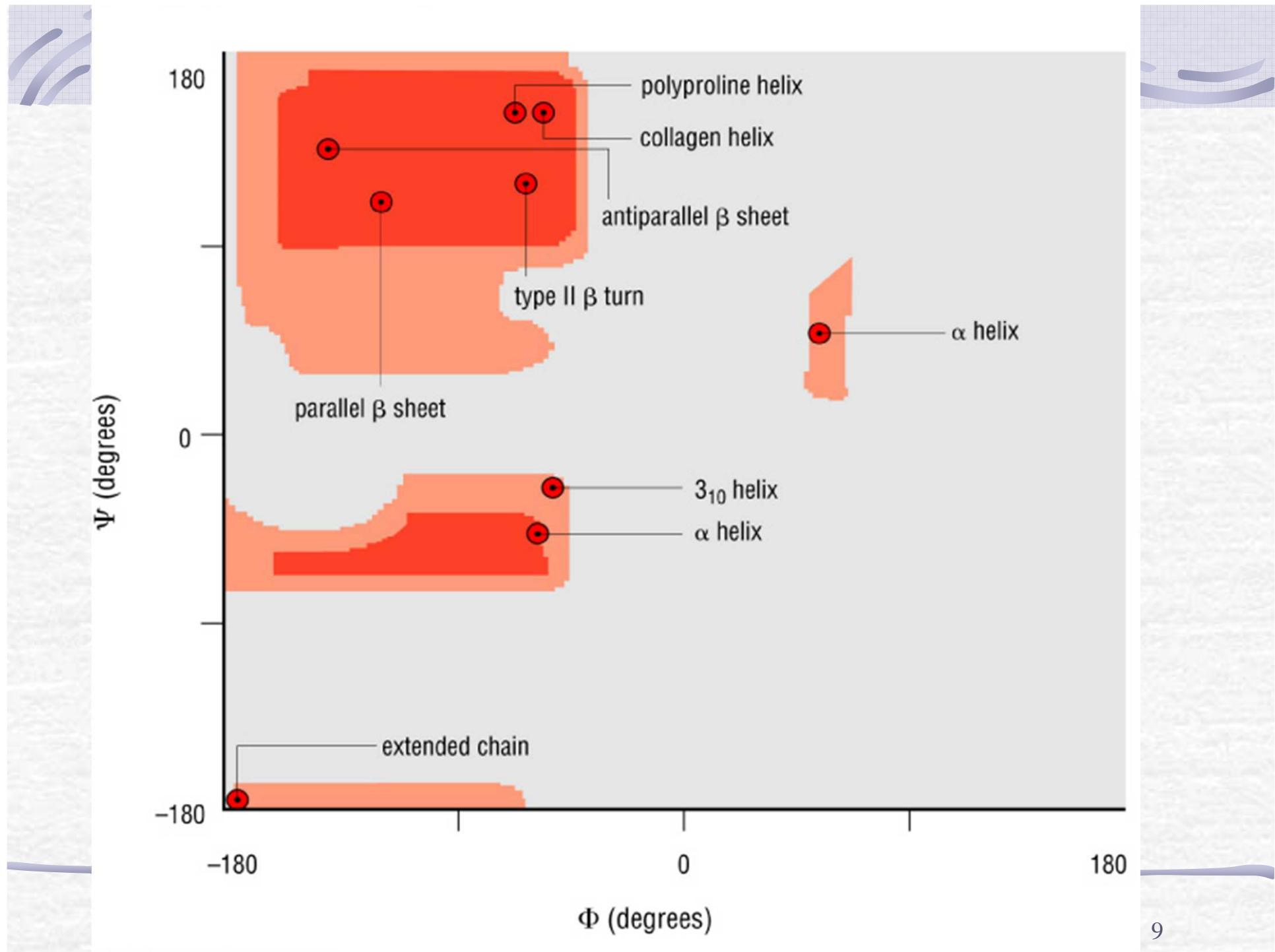


# Proteiny

- Atomy mimo postranní řetězce – kostra proteinu
- Délky vazeb a planární vazebné úhly vazeb atomů páteře proteinu jsou víceméně pevné
- Ohebnost páteře proteinu je odvozena od torzních úhlů  $\phi$  a  $\psi$



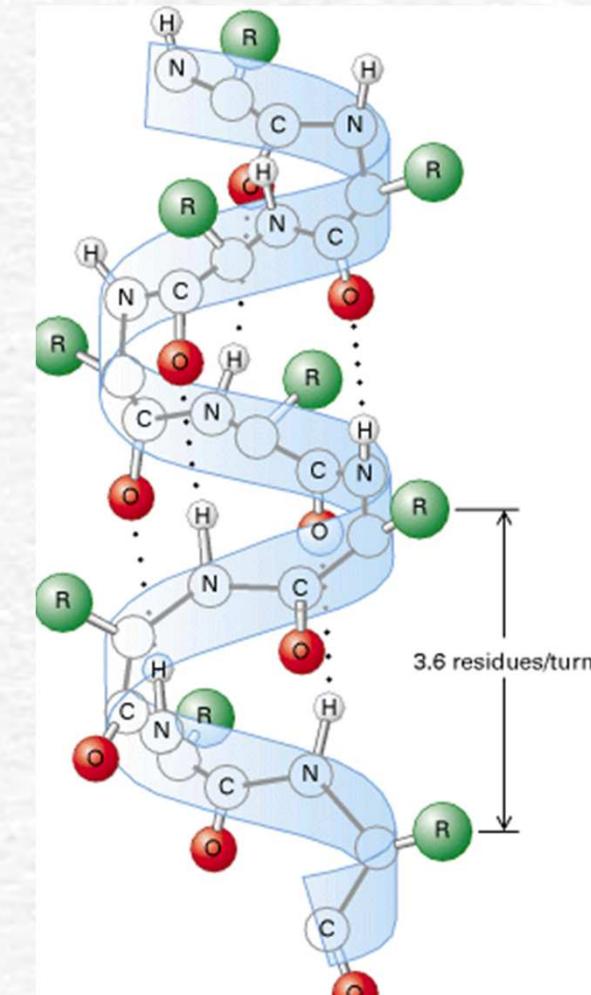
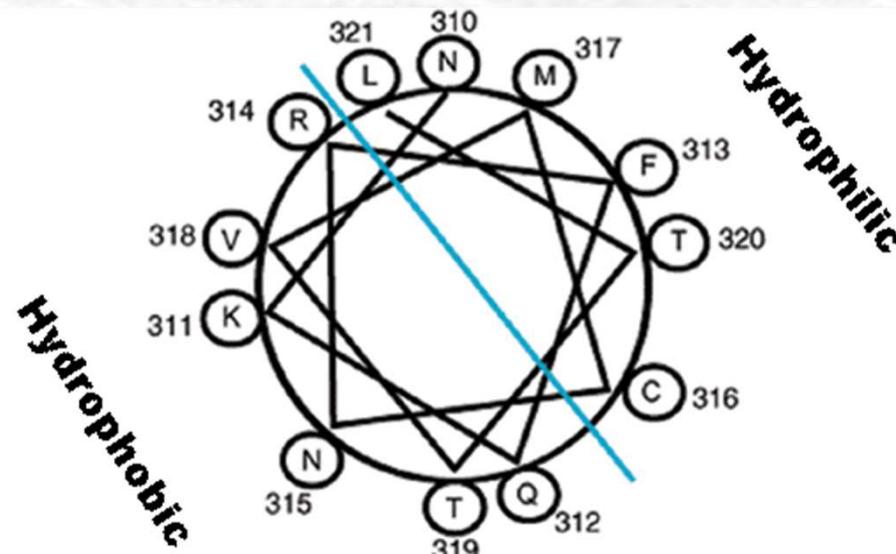
- Nejběžnější lokální struktury:  $\alpha$ -helix,  $\beta$ -sheet,  $\beta$ -turn
- Úhel  $\omega$  je planární – obvykle  $\sim 180^\circ$



# Proteiny

## A-helix:

- 3,6 AK na 1 otočku
- $\phi \approx -60^\circ$  a  $\psi \approx -50^\circ$
- Postranní řetězce trčí ze šroubovice ven



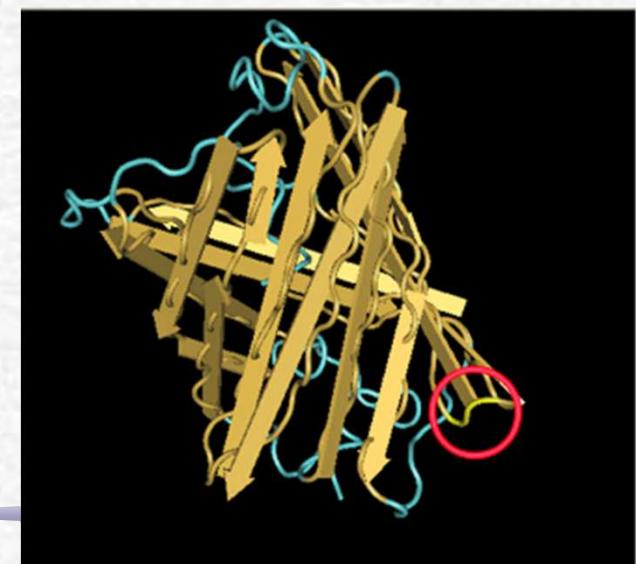
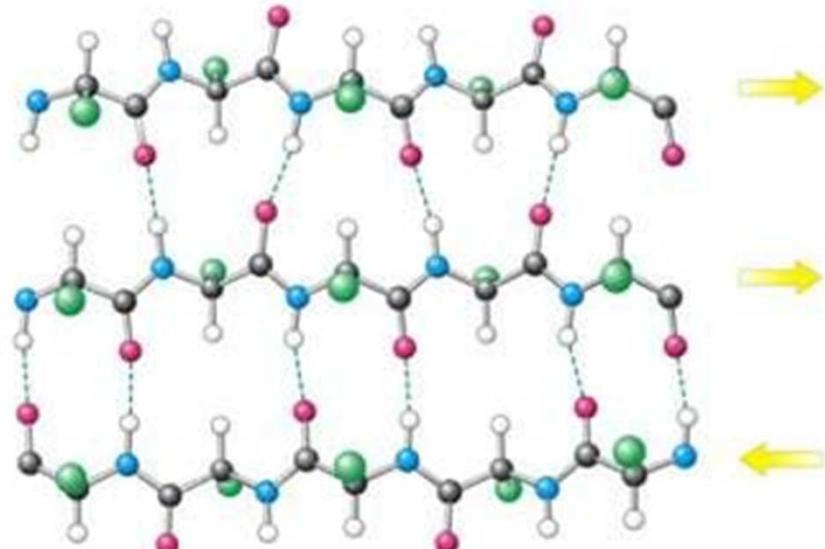
# Proteiny

## β-sheet

- 2 nebo více vláken
- Paralelní x antiparalelní
- Vodíkové vazby mezi jednotlivými vlákny
- $\phi \approx -130^\circ$  a  $\psi \approx 125^\circ$

## β-turn

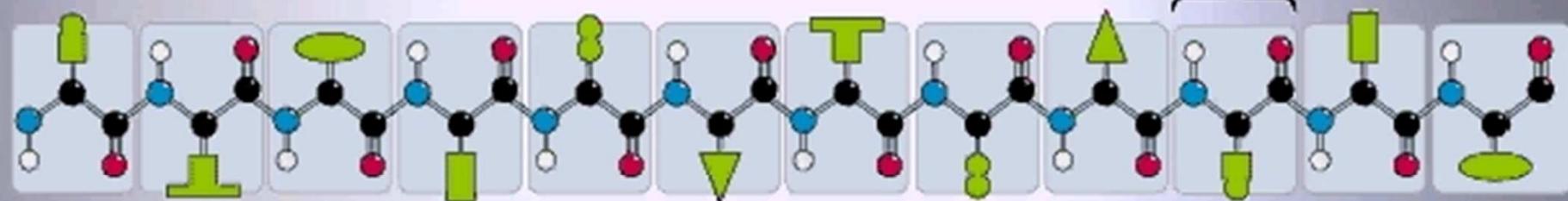
- 3-4 AK
- Obvykle na povrchu proteinu
- Otočení směru vlákna



# Struktura proteinů

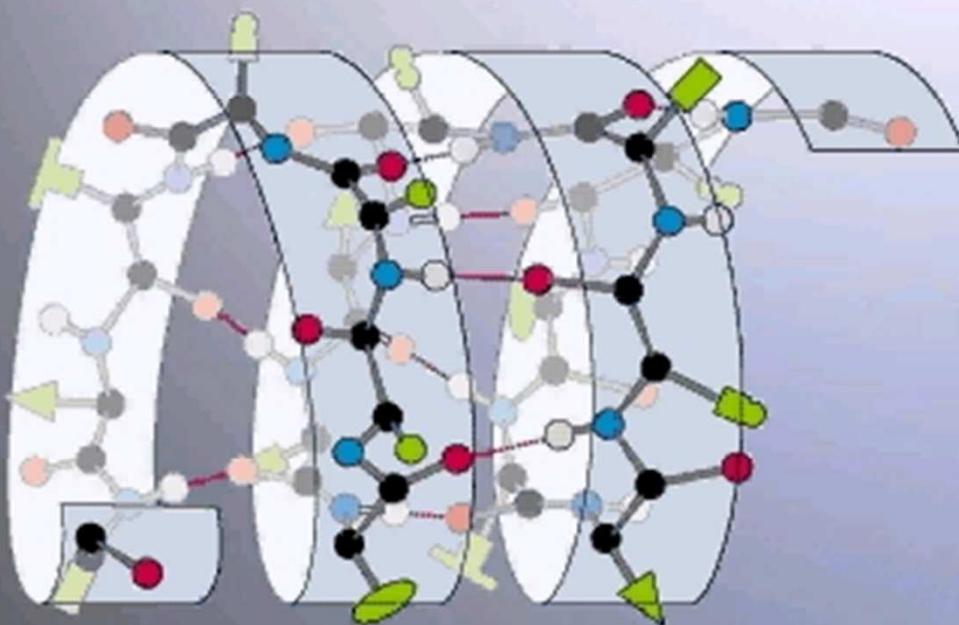
- ✓ V proteinech se opakují určité vzory uspořádání aminokyselin (charakteristické vodíkové vazby a hodnoty torzních úhlů)
- ✓ Nejběžnější **lokální struktury**:  $\alpha$ -helix,  $\beta$ -sheet (torzní úhly, vodíkové vazby, vzájemná pozice  $C_\alpha$ )
- ✓ Oblasti sekundární struktury a méně strukturované oblasti vytváří celkový prostorový tvar proteinu – **terciární struktura** (globulární tvar, hydrofóbní jádro)
- ✓ **Kvartérní struktura** – komplex tvořený více proteinovými řetězci

## Primary structure

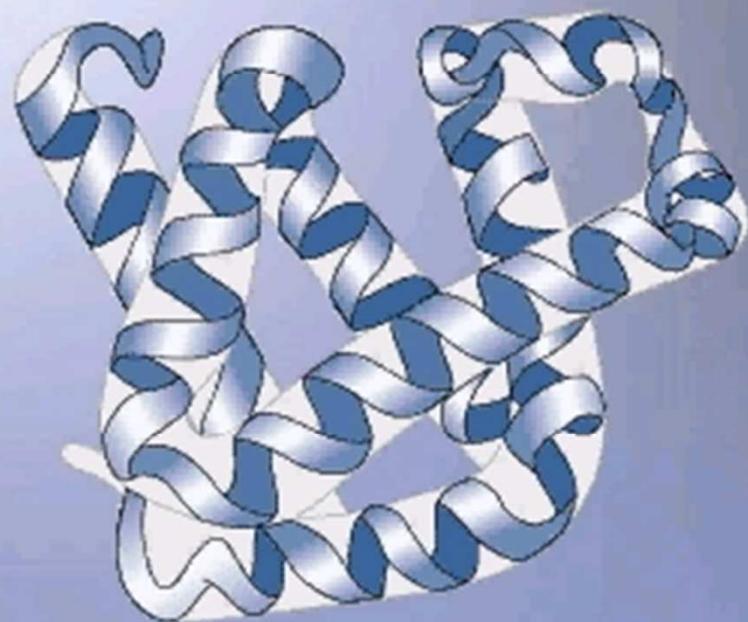


## Amino acid residue

## Secondary structure



## Tertiary structure



# Proteiny

- Na stabilizaci celkové struktury molekuly proteinu se podílí především slabé polární interakce:
  - Van der Waalsovy síly
  - Vodíkové vazby
  - Iontové vazby
  - Jsou to velmi slabé vazby, ale jeden protein mohou stabilizovat stovky až tisíce těchto vazeb
- Disulfidické vazby (vazby cystein-cystein)

# Obsah

- Motivace
- Základní pojmy
- **Predikce elementů sekundární struktury**
- Predikce celkové struktury proteinů

# Motivace

- ✓ Pro některé rodiny proteinů jsou typické určité vzory uspořádání elementů sekundární struktury.
- ✓ Pokud dokážeme předpovědět jaké elementy sekundární struktury se budou v proteinu vyskytovat (a jejich pořadí) můžeme předpovědět jeho přibližnou funkci. (např. katalýza chem. reakcí)
- ✓ Přesnost cca 70%
- ✓ Bohužel lze takto rozpoznat proteiny pouze z několika rodin proteinů a navíc funkci nelze predikovat přesně (Jaké reakce bude protein katalyzovat?)

# Cíl

- Pro každou aminokyselinu z cílové sekvence předpovědět, v jakém elementu sekundární struktury se bude vyskytovat
  - H – helix
  - B nebo E – sheet
  - T – turn
  - C – random coil
  - malá písmena – predikce s menší jistotou



# Používané metody

- 🕒 Statistické metody (pravděpodobnostní)
  - Pravděpodobnost výskytu AK v daném strukturním elementu
  - CF, GOR
- 🕒 Knowledge-base
  - PREDATOR, GOR V
  - + tvar, velikost, fyzikálně-chemické vlastnosti
  - + nejbližší sousedé
- 🕒 Metody založené na strojovém učení
  - Využívají neuronové sítě nebo HMM
  - PSIPRED
- 🕒 Konsenzuální metody

# Statistické metody

- ✓ (single residue statistics)
- ✓ Využívají pravděpodobnosti výskytu (preference výskytu) jednotlivých aminokyselin v různých elementech sekundární struktury
- ✓ Výsledky se dost liší
- ✓ Často se využívá více nástrojů a hledá se konsenzus
- ✓ CF algoritmus, GOR I algoritmus

# Chou-Fasman metoda (CF)

- Jedna z prvních metod (1970)
- Využívá relativní frekvence výskytu jednotlivých AK v různých elementech (helix, sheet, turn)
- Potřebuje trénovací množinu
- Relativní frekvence:

$$f_i^h = \frac{n_i^h}{n^h}$$

- Počet výskytů dané AK ve šroubovici / počet všech residuí ve šroubovicích
- Obdobně  $f_i^b$  a  $f_i^t$

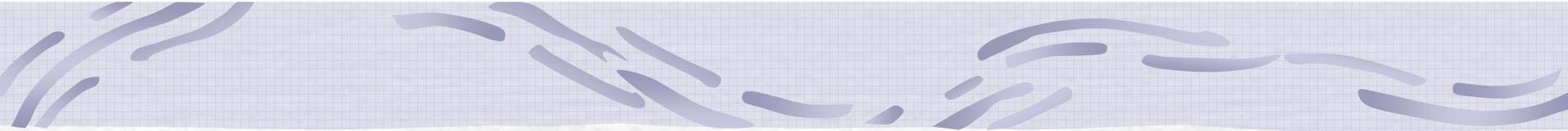
# CF metoda

- Parametr konformační preference:

$$P_i^h = \frac{f_i^h}{\langle f_i^h \rangle}$$

- relativní frekvence / průměrná relativní frekvence (pro všechny AK)
- Opět obdobně  $P_i^b$  a  $P_i^t$
- Každá AK je potom klasifikována *tvůrce* nebo *lamáč šroubovice* nebo listu

Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.070	0.106	0.099	0.085
Aspartic Acid	101	54	146	0.147	0.110	0.179	0.081
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Glutamic Acid	151	37	74	0.056	0.060	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.140	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.070
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.120	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053 <sup>22</sup>



# CF metoda

1. Každé AK přiřadí hodnoty z tabulky.
  2. Hledá regiony, ve kterých 4 ze 6 po sobě jdoucích AK mají  $P(h) > 100$  (tvůrci šroubovice).
  3. Rozšiřuje šroubovici v obou směrech, dokud 4 po sobě následující AK mají průměrné  $P(h) \geq 100$ . V opačném případě nalezen lamač šroubovice.
  4. Jestliže takto nalezený segment je delší než 5 AK a průměrné  $P(h) > P(b)$   $\rightarrow$  oblast šroubovice
  5. Hledá regiony, ve kterých 3 z 5 po sobě jdoucích AK mají  $P(b) > 100$  (tvůrci listu).
- 

# CF metoda

1. Rozšiřuje list v obou směrech, dokud 4 po sobě následující AK mají průměrné  $P(b) \geq 100$ . V opačném případě nalezen lamač listu.
2. Jestliže takto nalezený segment má průměrné  $P(b) > 105$  a průměrné  $P(b) > P(h) \rightarrow$  oblast listu
3. Každý region označený zároveň jako šroubovice i jako list, je označen jako šroubovice, jestliže průměrné  $P(h) > P(b)$ , v opačném případě je označen jako list.
4. Identifikuje  $\beta$ -turn.  
 $p(t) = f(j)f(j+1)f(j+2)f(j+3)$   
Jestliže  $p(t) > 0,000075$  a průměrná  $P(\text{turn}) > 100$  a  $P(h) < P(\text{turn}) > P(b) \rightarrow \beta\text{-turn}$ .

# GOR I

- ✓ *Garnier-Osguthorpe-Robson*
- ✓ Využívá teorii informace
- ✓ S – 4 různé strukturální elementy (alpha, beta, turn, coil)
- ✓ R – 20 různých AK

$$I(S; R) = \log_e \frac{P(S / R)}{P(S)}$$

- ✓  $P(S/R)$  – pravděpodobnost výskytu stavu S, jestliže je na dané pozici aminokyselina R
- ✓ Z trénovací množiny je třeba zjistit:  
20 AK x 4 strukt. Elementy = 80 hodnot

# GOR I

- ✓ Později rozšířena o tzv. řídící informaci
- ✓ Vliv sousedních AK na strukturální uspořádání v daném místě sekvence
- ✓ Bere v úvahu sousedních 8 AK z každé strany

$$I(S_j, R_1 \dots R_N) \approx I(S_j, R_j) + \sum_{m=-8 \dots +8} I(S_j, R_{j+m})$$

- ✓ Řídící informace je nezávislá na prostřední AK
- ✓ 16 pozic x 20 AK x 4 stavy = 1280 parametrů
- ✓ 1280 + 80 = 1360 parametrů celkem

# GOR I

- Pro každou AK v cílové sekvenci se vypočítá informace pro všechny čtyři stavy
- Stav s největší informací je dané pozici přiřazen
- Intrepretace – uvnitř jednoho elementu některá residua mají jinou předpověď – neurčitá předpověď
- GOR II – větší databáze pro výpočet parametrů

# GOR III

- Metoda druhé generace
- Predikuje pouze 3 stavy: helix, sheet, coil
- Korelace mezi typem AK na jednotlivých sousedních pozicích a centrální AK (okno přes 17 AK)

$$I(S_j, R_1 \dots R_N) \approx I(S_j, R_j) + \sum_{m=-8 \dots +8} I(S_j, R_{j+m} | R_j)$$

- Problém: dostatek dat v trénovací množině
- GOR IV – lepší trénovací množina pro výpočet parametrů

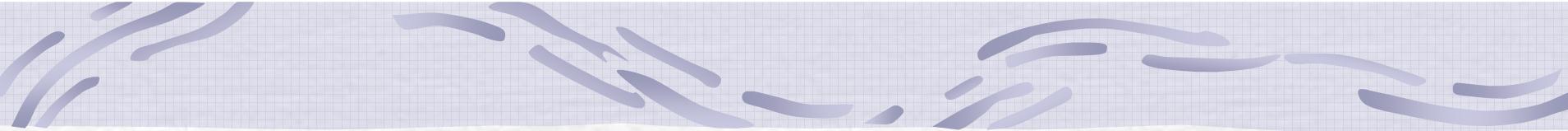
# Knowledge-base metody

- ✓ Využívají další informace o AK – tvar, velikost, fyzikálně-chemické vlastnosti
- ✓ Využívají evoluční informaci pro predikci
- ✓ Metoda nejbližších sousedů
  - Sekundární struktura se určí podle struktury nejpodobnějších sekvencí
  - Pro každou podsekvenci délky 17 se najdou nejpodobnější podsekvence z trénovací množiny
  - Centrální AK se přiřadí nejpočetnější struktura
- ✓ GOR V
  - PSI-BLAST pro zarovnání sekvencí
- ✓ PREDATOR (nearest-neighbor + interakce mezi vzdálenými AK)
- ✓ ZPRED (založen na GOR)

# Metody založené na strojovém učení

## ■ Využití neuronových sítí

- multi-layered feed-forward networks
- vytváří zarovnání s nejpodobnějšími sekvencemi (evoluční informace)
- pro predikci bere v úvahu 6 sousedních AK z každé strany ( $13*20=260$  neuronů ve vstupní vrstvě)
- Výstupní vrstva – obvykle 3 neurony ( $\alpha$ ,  $\beta$ , coil)
- PSIPRED, PHD, NNSSP
- PROF, Jnet



# Konsenzuální metody

- ✓ JPRED, NPS
  - ✓ Predikce pomocí více různých metod
  - ✓ Výběr nejčastějších přiřazení - konsenzu
  - ✓ Při nerozhodném výsledku – předpověď metody PHD
- 



# Trénovací a testovací data

## Trénovací data

- Rovnoměrné zastoupení všech různých struktur
- Hodně podobných struktur – ovlivnění predikce
- $\alpha$ - šroubovice častější než  $\beta$ -list
- Různé sekvence => stejná struktura

## Testovací data

- Stanovení přesnosti modelu
- Nezávislé na trénovacích datech
- Neobsahovat homology k trénovacím datům (podobnost sekvencí pod 25 – 30%)

# Přesnost predikce

- Vždy bude menší než 100% - dáno povahou dat
- $Q_3$ 
  - poměr mezi počtem správně predikovaných AK a celkovým počtem AK
  - nejlepší metody:  $Q_3=80\%$
- Sov (segment overlap)
  - procento správně predikovaných elementů sekundární struktury
  - důraz kladen na správný počet, typ a pořadí elementů sekundární struktury

# Obsah

- Motivace
- Základní pojmy
- Predikce elementů sekundární struktury
- Predikce celkové struktury proteinů

# Predikce struktury proteinů

- Modelování struktury na základě homologie  
(porovnávání primárních struktur, databáze sekvencí, nástroje BLAST) (*comparative modeling, knowledge-based modeling*)
- Threading  
(porovnání energetické výhodnosti uspořádání sekvence do jednotlivých známých struktur, rodiny proteinů)
- Ab initio modelování  
(modelování na základě energetické výhodnosti)
- Skládání ze sekvenčně-strukturních fragmentů  
(I-sites library, Ch. Bystroff a D. Baker )

# Modelování na základě homologie

- ✓ Využívá robustnost genetického kódu:
  - malé změny v sekvenci proteinu obvykle vedou pouze k malým změnám ve struktuře
  - radikální změnu struktury proteinu obvykle způsobí až kumulace mnoha změn v sekvenci
  - Struktura je lépe konzervovaná než sekvence
- ✓ Predikce struktury na základě struktury podobných proteinů
- ✓ Použitelná pokud k cílové sekvenci existuje dostatečně podobná sekvence (templát) jejíž struktura je známá
- ✓ Čím více struktur experimentálně zjištěno, tím lépe je metoda použitelná
- ✓ Homologní sekvence – sekvence odvozené od společného předchůdce

# Modelování na základě homologie

## ■ Předpoklady:

- Jádro proteinu
  - Obvykle dobře konzervované
  - Stejné prostorové uspořádání atomů (nebývá identické, ale pouze podobné)
  - Strukturní elementy  $\alpha$ -helix,  $\beta$ -sheet – pokud možno bez mezer v zarovnání sekvencí
- Smyčky
  - Obvykle málo konzervované, velmi variabilní
  - Spadají do nich mezery v zarovnání sekvencí
  - Mohou obsahovat doplňkové strukturní elementy
  - Často obsahují vazebná místa molekuly

# Modelování na základě homologie

## 1. Výběr templátu

- Nejdůležitější krok, špatný výběr nelze v dalších krocích napravit
- Problém podobnosti sekvencí (BLAST, FASTA)
- Vyhledávání v PDB
- Existuje-li více homologů:
  - Výběr nejpodobnější sekvence
  - Použítí průměrné struktury všech templátů
  - Použítí fragmentů struktury z různých templátů (nutno spojit fragmenty – vznik chyb)

## 2. Zarovnání cílové sekvence s templátem

- Globální zarovnání (Needleman-Wunsch metoda)
- Odpovídající AK – stejná struktura
- Zásadní vliv na správnost modelu

# Modelování na základě homologie

## 3. Vytvoření modelu

- Struktura vybraného templátu = struktura cílové sekvence
- Struktura templátu slouží pro definici omezení na cílovou strukturu
- Nejprve modelování jádra proteinu (úseky bez mezer)
- Ohodnocení správnosti modelu
- Následné modelování smyček a postranních řetězců

## 4. Evaluace modelu

- Zda model splňuje fyzikální zákony, principy stereochemie, biologie a biochemie (povolené hodnoty torzních úhlů, dostatek prostoru pro jednotlivé atomy, ...)

# Modelování na základě homologie

## Problémy:

- Co s residui, které se liší od templátu? Jaké bude uspořádání jejich postranních řetězců v prostoru?
- Inzerce a delece – dlouhé oblasti mezer
- Nesprávné zarovnání sekvencí při malé podobnosti

• Nicméně tato metoda je velmi užitečná



# Threading

- *Fold recognition, inverse folding*
  - Struktury proteinů jsou často mnohem více podobné než jejich sekvence
  - V přírodě se vyskytuje pouze relativně malé množství různých celkových uspořádání proteinů
  - Nový protein nemusí představovat novou topologii (fold)
  - Jaká konkrétní topologie se nejlépe hodí pro daný protein?
- 

# Threading

- Bereme různé známé struktury a hledáme, jak nejlépe cílovou sekvenci namapovat na danou strukturu tak, abychom maximalizovali vzájemné interakce mezi AK
- Potřebujeme:
  - Kompletní knihovnu různých topologií (strukturální templaty)
  - Skórovací funkci, která rozhodne, které zarovnání je nejlepší
  - Algoritmus pro mapování cílové sekvence na jednotlivé topologie

# Ab-initio modelování

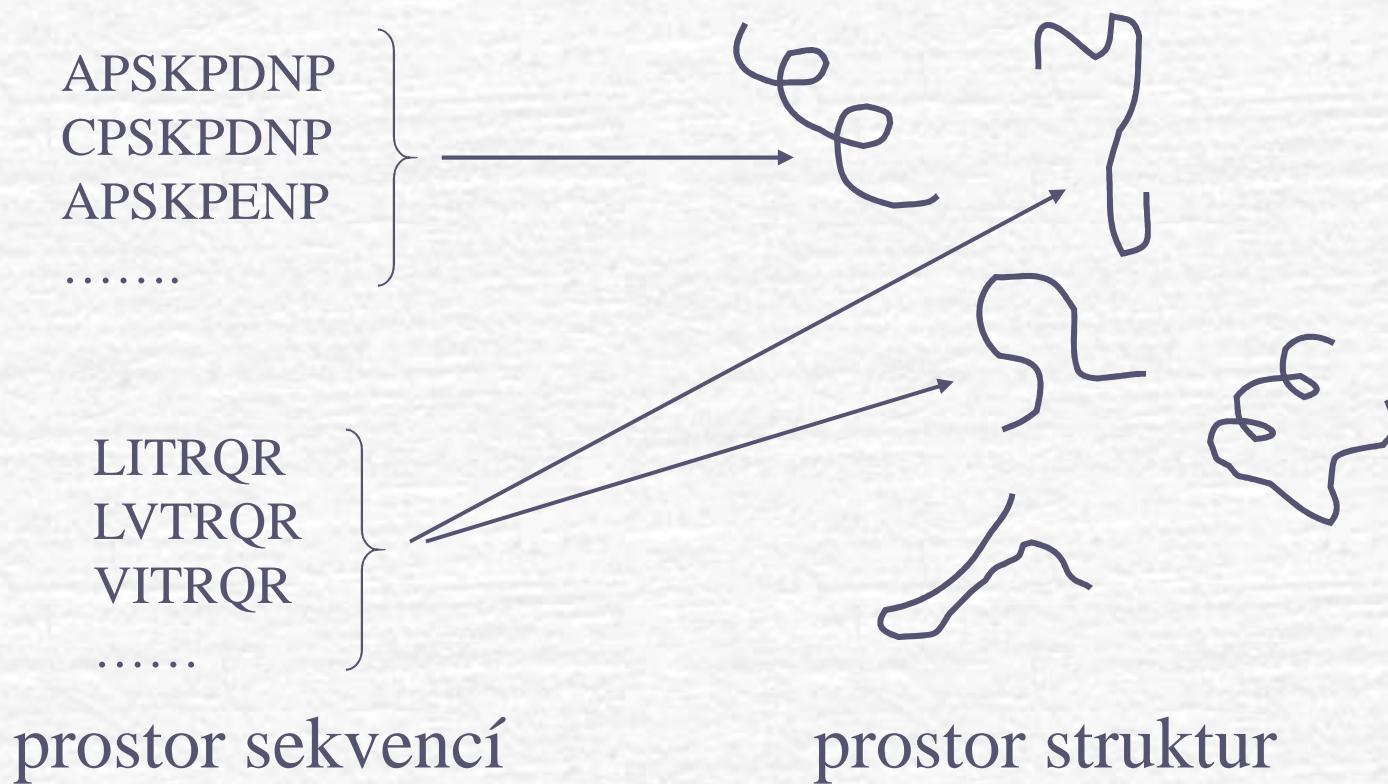
- ✓ Založeno pouze na znalosti sekvence a přírodních zákonů a principů (fyzikálních, chemických)
- ✓ Finální struktura proteinu je struktura s nejmenší volnou energií
  - Musí respektovat stereochemická omezení – reálné hodnoty torzních úhlů
- ✓ Problémy:
  - Jak definovat energii proteinu?
  - Jak určit konformaci s nejmenší energií
  - Obrovské množství možných konformací



# Skládání ze sekvenčně-strukturálních fragmentů

- ✓ Potřebujeme databázi fragmentů
  - ✓ Umožňuje nalézt nové topologie
  - ✓ Není tak výpočetně složité jako ab-initio modelování
  - ✓ V cílové sekvenci se naleznou úseky, které odpovídají některému z fragmentů
  - ✓ Hledá se vzájemná poloha fragmentů a struktura oblastí, které nelze pokrýt fragmenty
- 

# Sekvenčně-strukturní fragmenty



# CASP

- Comparative Assessment of Structure Prediction
- Každoroční soutěž o nejlepší metodu pro predikci struktury proteinů
- Ohodnocení přesnosti predikce různých metod – srovnání přesnosti metod
- Predikce sekundární struktury nebo i terciální struktury