

# **Základy bioinformatiky**

Evžen Wybitul      Kateřina Krausová

16. května 2019

# Obsah

<b>1</b>	<b>Úvod</b>	<b>2</b>
<b>2</b>	<b>Struktura nukleových kyselin</b>	<b>3</b>
<b>3</b>	<b>Struktura proteinů</b>	<b>9</b>
3.1	Primární struktura . . . . .	9
3.1.1	Seznam aminokyselin . . . . .	13
3.2	Další proteinové struktury . . . . .	18
<b>4</b>	<b>Sequence alignment</b>	<b>24</b>
4.1	Dotplot . . . . .	33
4.2	Pairwise sequence alignment . . . . .	38
4.2.1	Skórovací tabulky . . . . .	40
4.2.2	Algoritmy . . . . .	45
4.3	Multiple sequence alignment . . . . .	48
<b>5</b>	<b>Hledání v databázích</b>	<b>51</b>
5.1	Algoritmy . . . . .	53
5.1.1	FASTA . . . . .	53
5.1.2	BLAST . . . . .	54
5.1.3	Srovnání FASTA a BLAST . . . . .	57
5.1.4	Parametry významnosti alignmentu . . . . .	58
5.1.5	Profilové algoritmy . . . . .	61
<b>6</b>	<b>Analýza sekvencí</b>	<b>63</b>
6.1	Hledání motivů . . . . .	64
6.2	Další možnosti analýzy sekvencí . . . . .	67

<b>7 Databáze</b>	<b>69</b>
7.1 Strukturní databáze . . . . .	70
7.1.1 PDB . . . . .	72
<b>8 Strukturní alignment</b>	<b>74</b>
8.1 Klasifikace proteinů . . . . .	80
8.2 Predikce struktury . . . . .	81
8.2.1 Intrinsically disordered proteins . . . . .	83
8.2.2 Predikce sekundární struktury . . . . .	84
8.2.3 Predikce membránových proteinů . . . . .	90
8.2.4 Homologní modelování . . . . .	94
8.2.5 Fold recognition . . . . .	97
8.2.6 Ab initio predikce . . . . .	99
8.2.7 Predikce interakce . . . . .	100
<b>9 Souvislost struktury a funkce</b>	<b>100</b>
9.1 Hledání funkce . . . . .	102
9.1.1 Solvent-accessible surface area . . . . .	104
9.1.2 Enzymy . . . . .	105

# 1 Úvod

**META** Zbytečná kapitola? Kdepak! Ve zkouškovém testu jsou otázky na historický vývoj bioinformatiky běžné, stejně jako obecné otázky typu ”Čím se zabývá bioinformatika? (článek na 100 slov)“.

Bioinformatika je vědní disciplína, která se zabývá zpracováním biologických dat. Slovem ”zpracování“dat máme namysli jejich sběr, archivaci, organizaci a interpretaci.

## Jaká data zpracováváme?

- měření (délky křídel, váhy vůní [sic]...)
- sekvence (DNA, RNA, proteiny)
- 3D struktury
- genomická data
- příbuzenské vztahy
- interakce
- atd.

A jak velká? Například největší genom, patřící organismu Amoeba dubia, má 670GB. Stejně tak se v desítkách GB pohybují i 3D histologické skeny. EBI (European Bioinformatic Institute) měl v roce 2015 kapacitu 60 PB dat.

## Historie bioinformatiky

- (1707–1778) Carl Linne, první bioinformatik
- (1956) Fred Sanger, první sekvence (insulin)
- (1957) Perutz, Kendrew, první proteinová struktura
- (1965) Margarett Dayhoff , první sekvenční databáze
- (1970) Needlman, Wunsch, algoritmus sekvenčního srovnávání
- (1971) první strukturní databáze
- (1988) Hugo projekt
- (1990) Altschul, Lipman et al., BLAST
- (1992) NCBI, GenBankGenbank
- (1995) First genome, Haemophilus influenzae

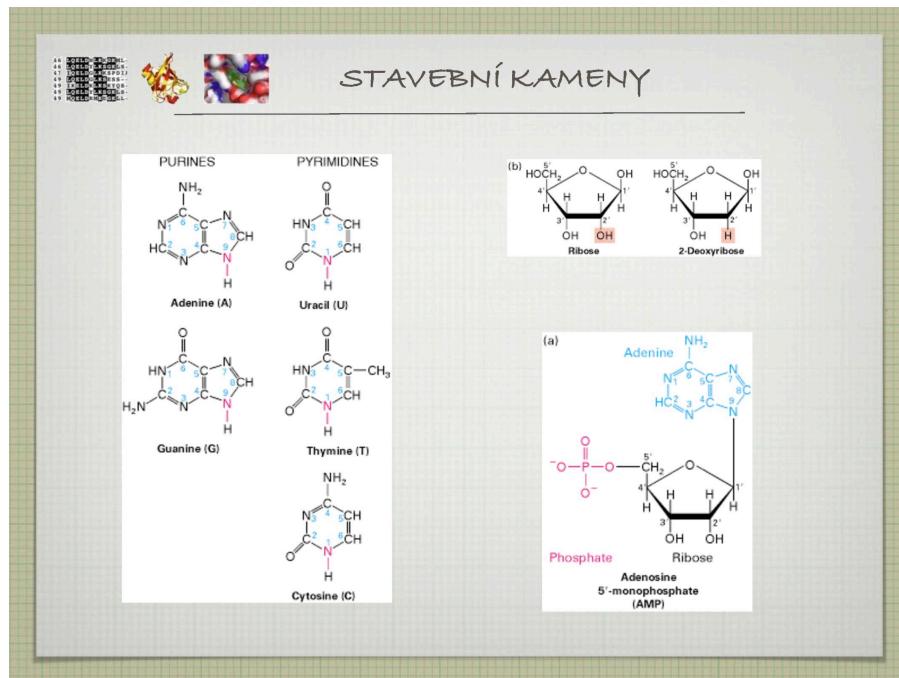
## 2 Struktura nukleových kyselin

Přednáška č.

Objev struktury DNA: Watson, Crick, Franklin (50. léta 20. století).

2

Obrázek 2.1: Prezentace č. 1, slide č. 10



### Centrální dogma molekulární biologie

1. transkripce DNA do RNA
2. translace RNA na proteiny
3. proteiny jsou finální manifestací genetické informace

### Stavební kameny

- puriny (adenosin, guanin), pyrimidiny (thyrosin, uracil, cytosin)
- ribosa, 2-deoxyribosa
- fosfát

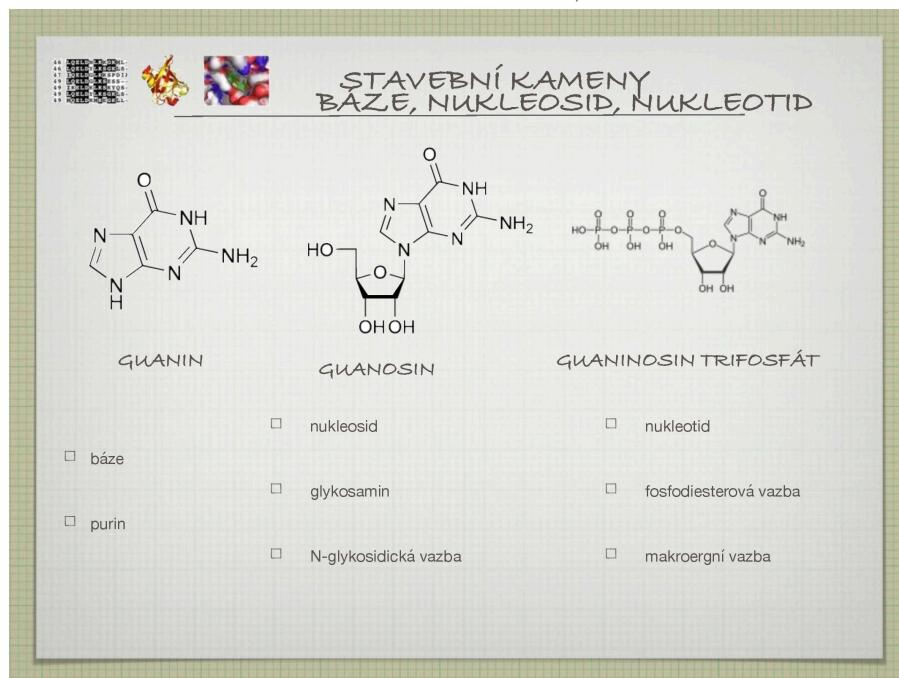
#### guanin

viz slide, běžná purinová báze

#### guanosin

nukleosid guaninu, tj. guanin + cukr vázaný N-glykosidickou vazbou

Obrázek 2.2: Prezentace č. 1, slide č. 11



### guanosin trifosfát

nukleotid guaninu, tj. guanosin + fosfát navázaný fosfodiesterovou vazbou

### Deoxynukleotid

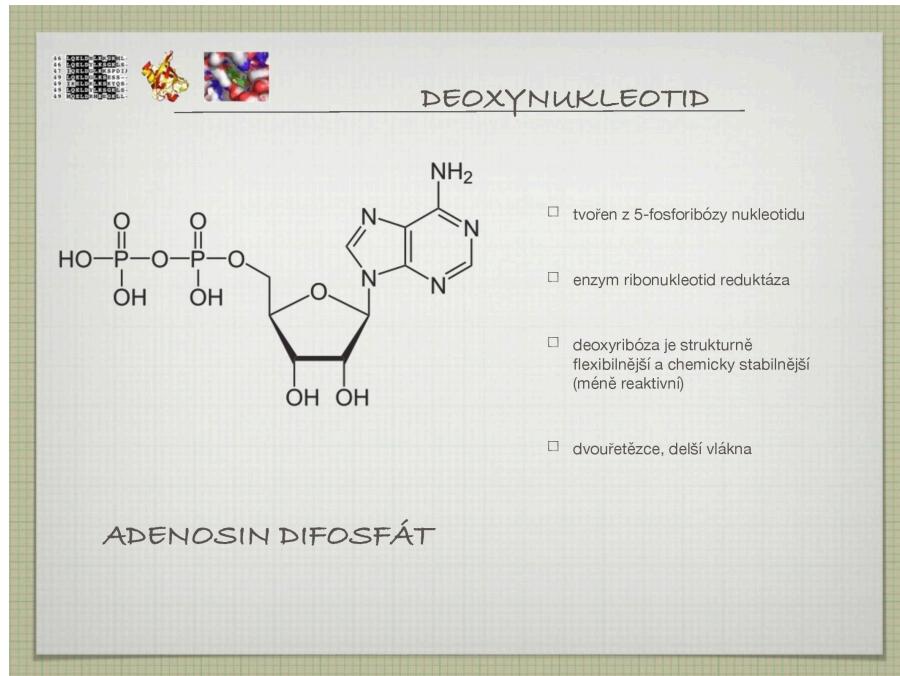
- na druhém uhlíku má cukr H místo OH
- vznik z nukleotidu redukovaného ribonukleotid reduktázou
- deoxyribóza je flexibilnější a chemicky stabilnější (výhoda pro DNA, které by se němelo měnit), protože OH skupina je reaktivní
- stabilizace vede k tvoření delších vláken

### Párování

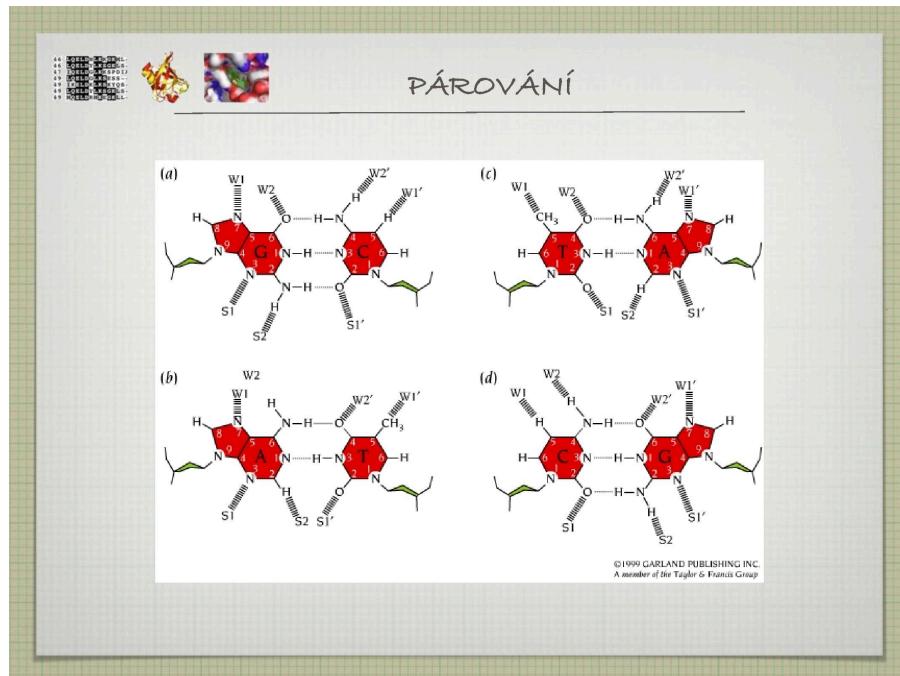
- díky němu vzniká sekundární struktura
- A + T páruje dvěma H můstky, C + G třemi H můstky
- AT bohaté úseky jsou tedy pružnější a GC úseky stabilnější

## 2. STRUKTURA NUKLEOVÝCH KYSELIN

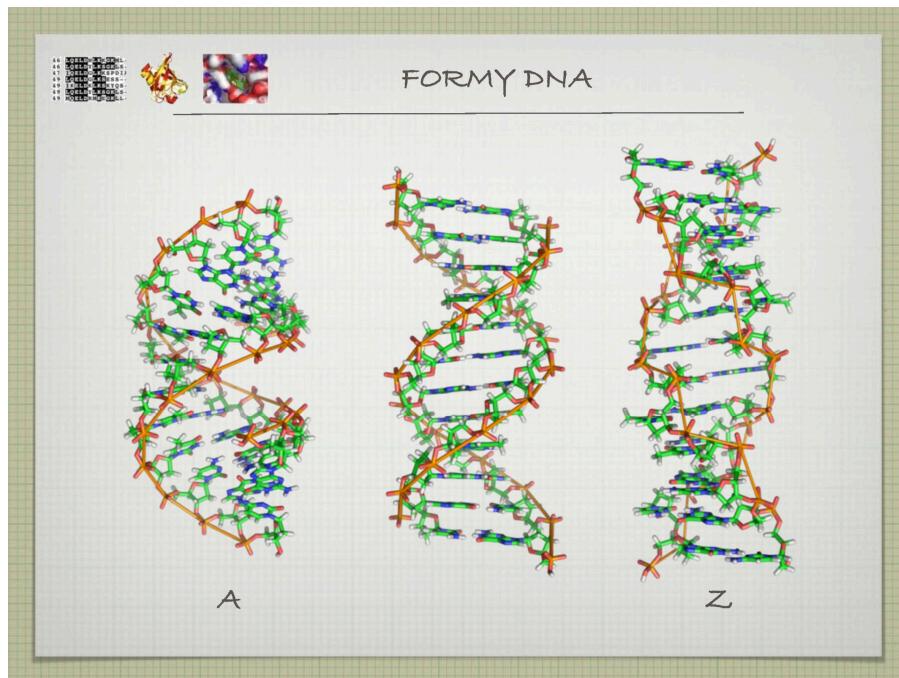
Obrázek 2.3: Prezentace č. 1, slide č. 12



Obrázek 2.4: Prezentace č. 1, slide č. 13



Obrázek 2.5: Prezentace č. 1, slide č. 17



### Vodíková vazba

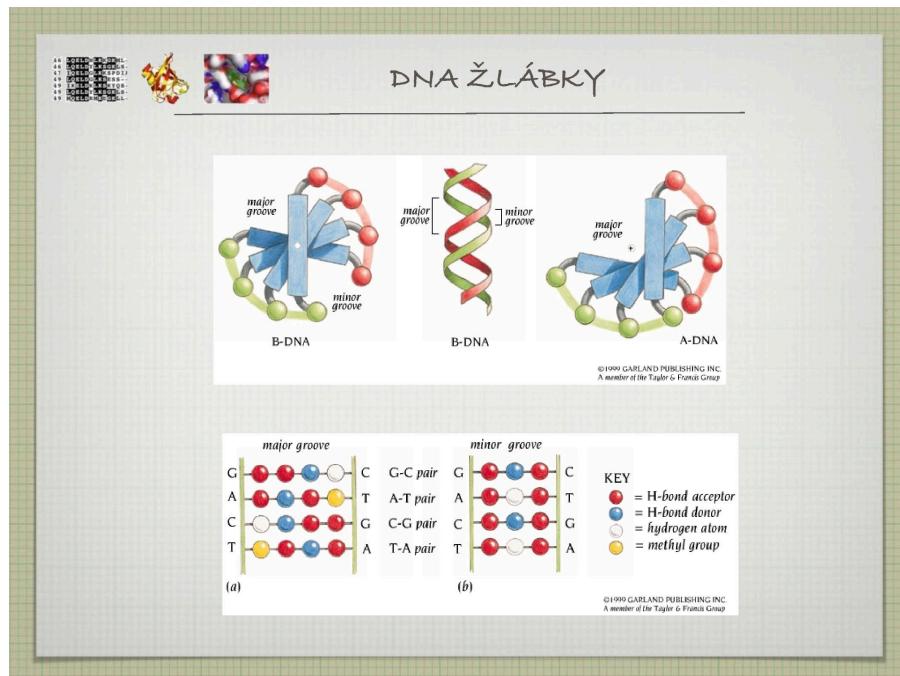
- nekovalentní přitažlivé interakce
- interakce dvou elektronegativních atomů, které jsou "spojeny" vodíkem
- vodík je připojen kovalentně k donoru a elektrostaticky k akceptoru (na vodíku vzniká parciální kladný náboj)
- délka 3 Å

### Struktura DNA

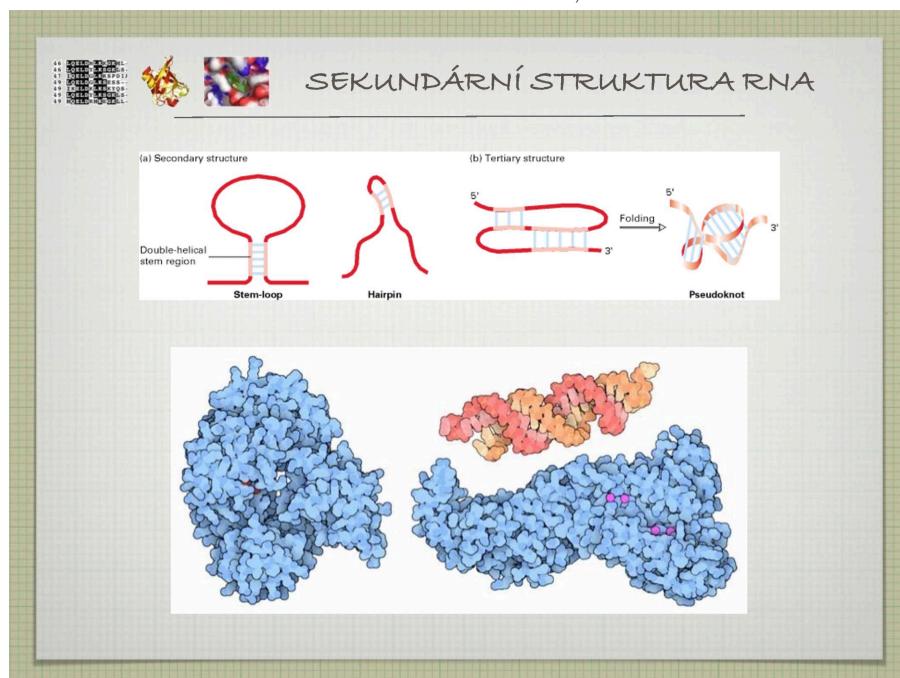
- dvoušroubovice (většinou pravotočivá)
- tři druhy: A, B, Z
  - A má skoro stejně velké žlábky, mezeru uprostřed
  - B je nejběžnější, má velký a malý žlábek
  - Z není příliš častá, je levotočivá (na rozdíl od zbytku)
- žlábky hrají důležitou roli při vázání enzymů (přes žlábek lze vidět, jaké nukleotidy v DNA jsou)

## 2. STRUKTURA NUKLEOVÝCH KYSELIN

Obrázek 2.6: Prezentace č. 1, slide č. 18



Obrázek 2.7: Prezentace č. 1, slide č. 19



## **Struktura RNA**

- loop, hairpin, pseudoknot
- většinou je tvořena pouze jedním vláknem

# **3 Struktura proteinů**

Přednáška č.

3

Primární až kvarterní; struktura určuje funkci proteinu (proto nás zajímá), například s čím reaguje, jakými membránami projde a za jakých podmínek atp. Další informace viz též zápisky z biopolymerů.

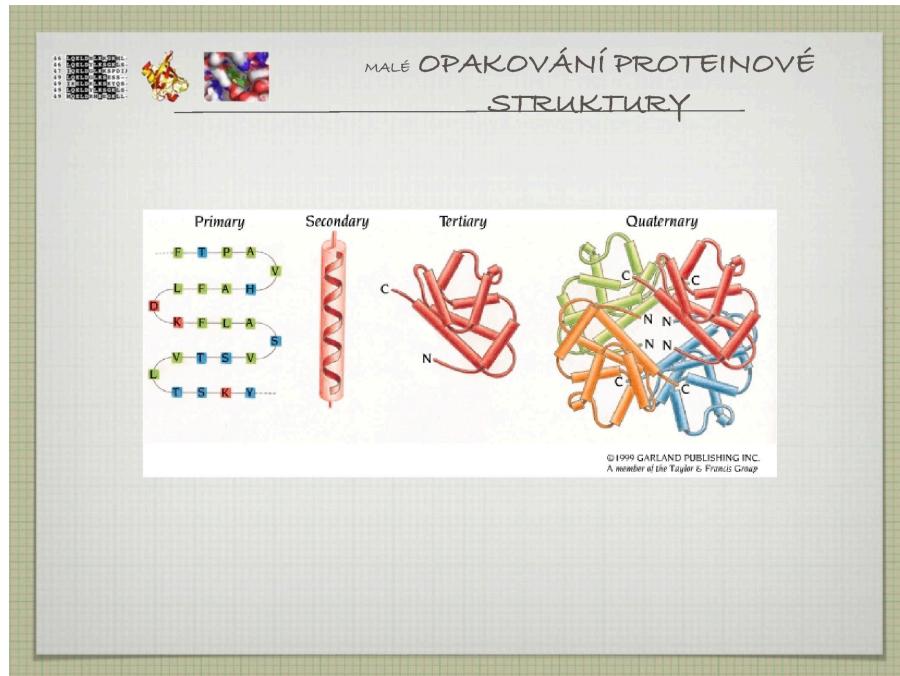
## **3.1 Primární struktura**

Primární struktura je určená pořadím aminokyselin (AK). AK je 20+2.

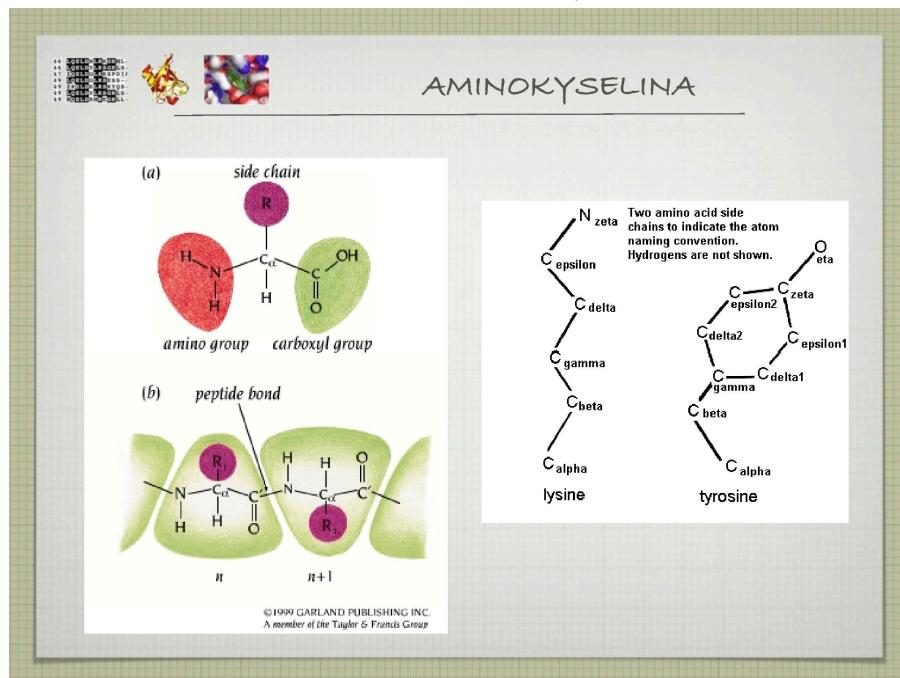
### **Struktura AK**

- $C_\alpha$  je chirální, jsou na něm navázány čtyři různé skupiny
- $NH_2$  se váže na  $COOH$  za vzniku peptidické vazby, uvolňuje se  $H_2O$ 
  - peptidická vazba je planární, vzniká pomyslný čtyřúhelník s rohy v  $C_\alpha$
  - ze 40% má charakter dvojně vazby
    - \* je kratší než jednoduchá
    - \* je planární
    - \* má cis a trans konfiguraci
  - rotace je tedy možná pouze v  $C_\alpha$ , existují dva torzní úhly ( $\phi, \psi$ )
    - \* teoreticky možné a prakticky spočítané hodnoty torzních úhlů se znamenávají do Ramachandranova diagramu

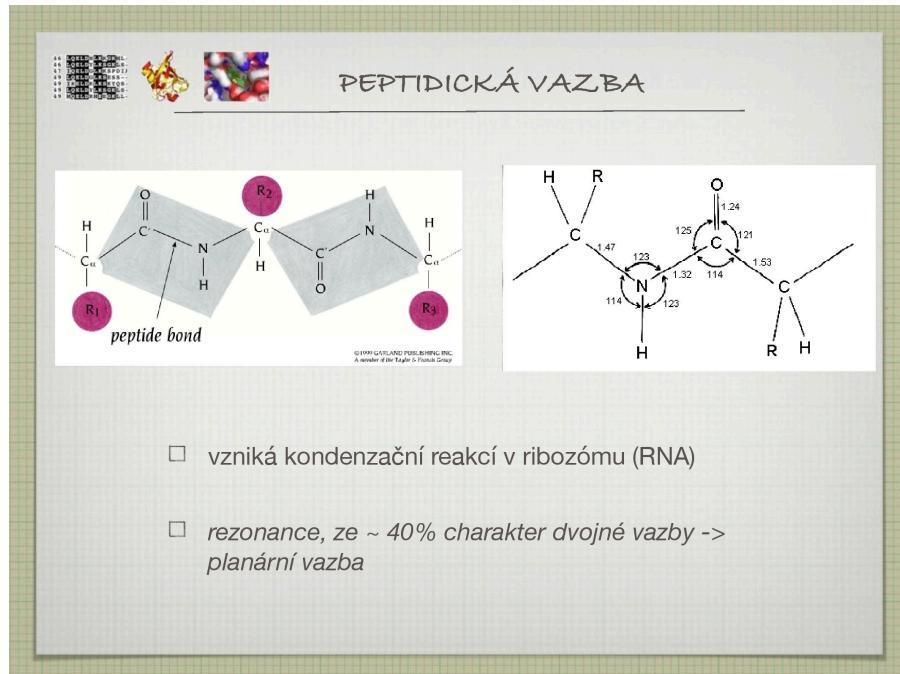
Obrázek 3.1: Prezentace č. 1, slide č. 21



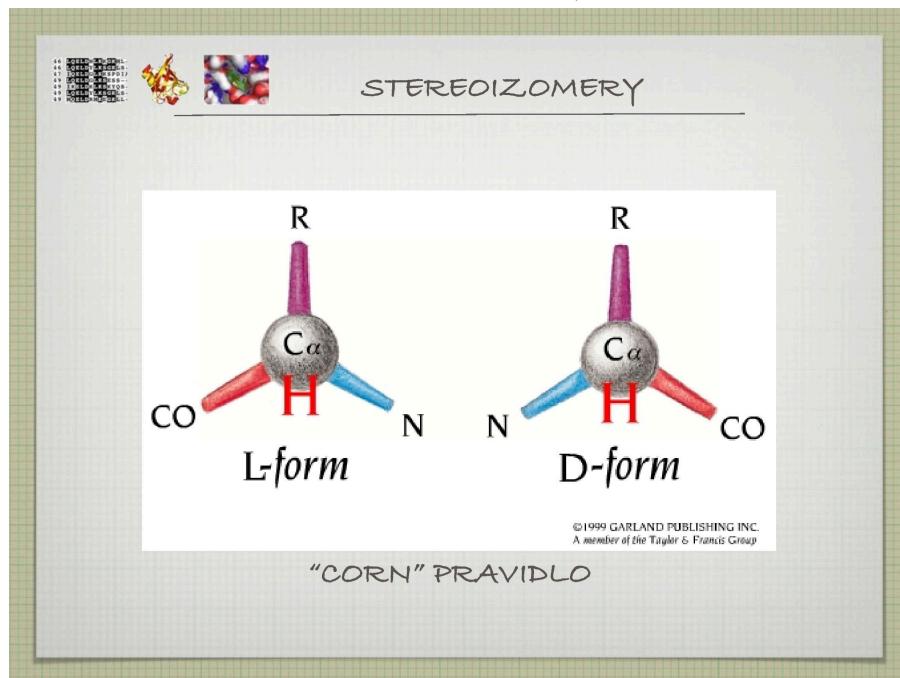
Obrázek 3.2: Prezentace č. 1, slide č. 22



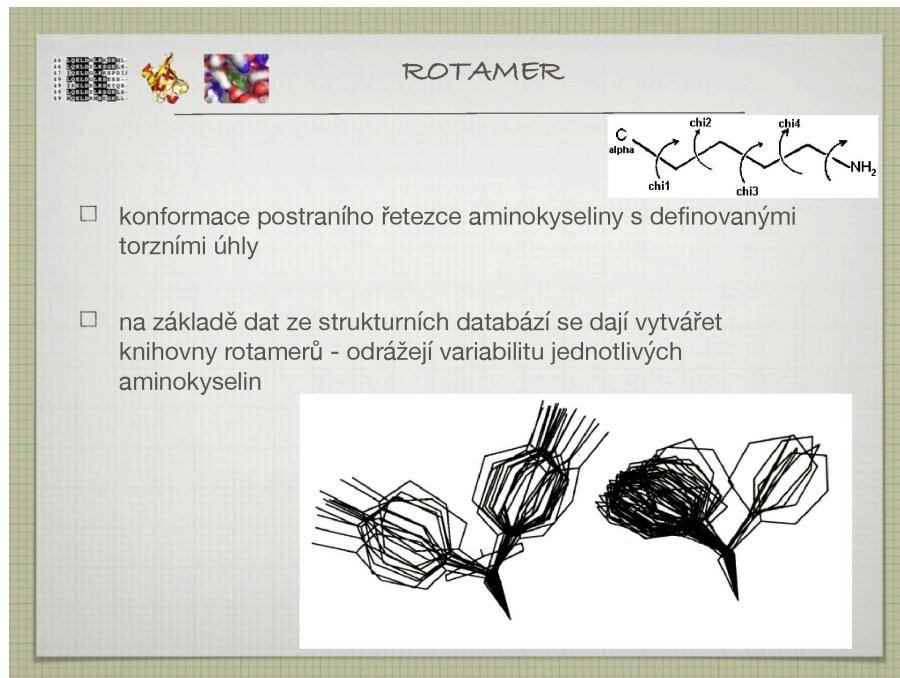
Obrázek 3.3: Prezentace č. 1, slide č. 51



Obrázek 3.4: Prezentace č. 1, slide č. 23



Obrázek 3.5: Prezentace č. 1, slide č. 26



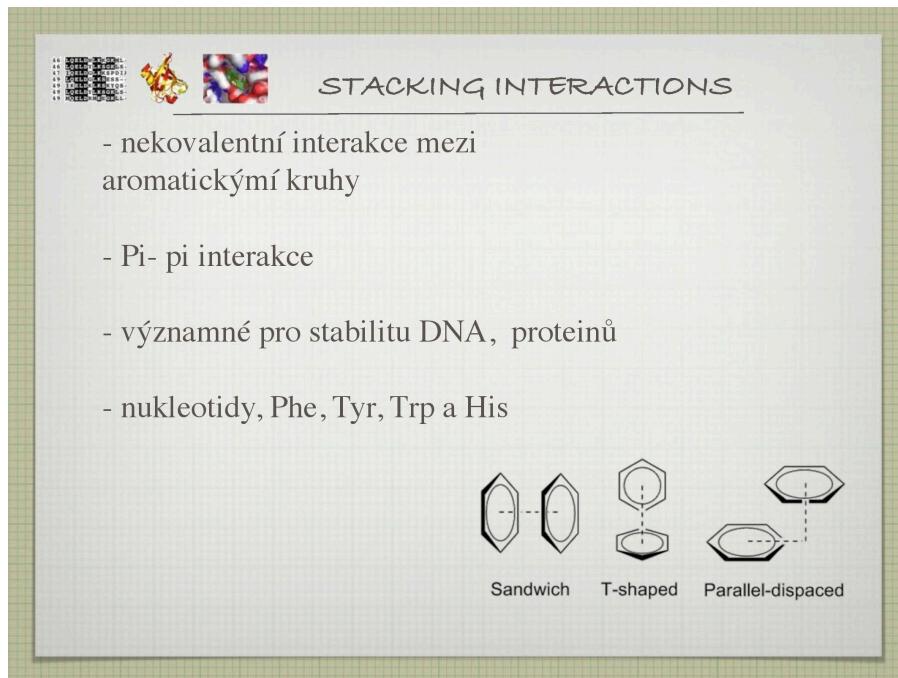
### Stereoizomery

- chirální uhlík stáčí rovinu polarizovaného světla
- rozlišujeme L a D enantiomery
  - v laboratoři vznikají přibližně v poměru 1:1
  - v živých organismech je většina AK druhu L
  - buněčná stěna baterií bývá často D, aby nebyla rozpoznána jinými (imunitními/nepřátelskými) buňkami

### Rotamery

- rotamery jsou AK se stejným složením, u nichž se liší konformace jejich postranního řetězce
- vytváří se knihovny rotamerů (dle naměřených dat), odráží variabilitu jednotlivých AK

Obrázek 3.6: Prezentace č. 1, slide č. 33



### Stacking interakce

- interakce mezi aromatickými kruhy ( $\pi$ - $\pi$  interakce)
- sandwich, T-shaped, parallel-displaced
- jsou významné pro stabilitu DNA i proteinů

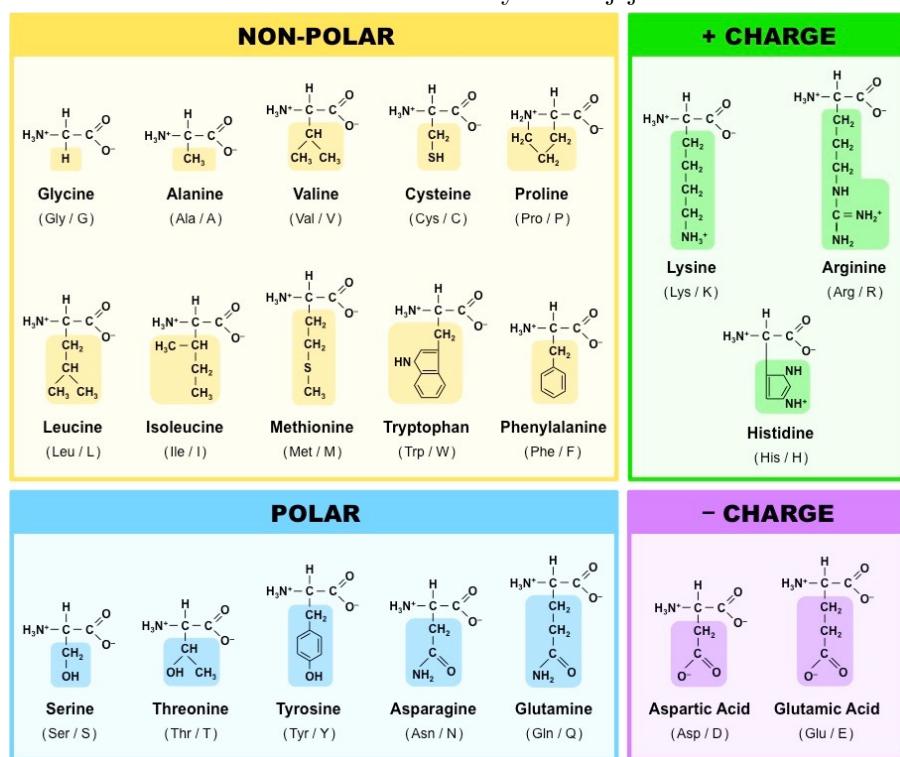
#### 3.1.1 Seznam aminokyselin

AK se dají rozdělit do několika skupin; nejdůležitější rozdělení je asi podle hydrofobicity, protože podle toho se poté jednotlivé AK vyskytují uvnitř nebo naopak na povrchu proteinů. Další významnou vlastností, která navíc s hydrofobicitou souvisí, je elektrický náboj.

**META** Na zkoušku bude požadována znalost všech AK včetně jejich vzorce, vlastností, a zkratky.

Polární AK jsou hydrofilní, nepolární jsou hydrofobní.

Obrázek 3.7: Seznam aminokyselin a jejich rozdělení



**AK s alifatickým postranním řetězcem****glycin**

často v kolagenu, často ve smyčkách, nejmenší a tedy dobře konzervovaný

**alanin**

také velice častý, existuje i D forma (buněčná stěna, antibiotika), také velice malý a tedy dobře konzervovaný

**valin**

často v helixech a listech

**isoleucin**

má dva chirální atomy a tedy čtyři formy, je častý v helixech i listech

**leucin**

součástí leucinového zipu při interakci proteinů s DNA

**AK s kyselou (karboxylovou/amidovou) skupinou****asparagová kyselina**

bývá v aktivních místech enzymů

**asparagin**

první izolovaná AK (z chřestu, viz jméno), tvoří vodíkové můstky, účastní se cappingu (neutralizuje parciální náboj na N' koncích alfa helixů)

**glutamová kyselina**

může fungovat jako neurotransmiter, je podobná ASP

**glutamin**

je zdrojem energie pro mozek

**AK se zásaditou (aminovou) skupinou****arginin**

může být methylován, bývá na povrchu, kvůli kladnému náboji tvoří vodíkové můstky se záporně nabitémi strukturami (DNA)

**lysin**

může být postrtranslačně modifikován

**AK s aromatickým jádrem nebo hydroxylovou skupinou****histidin**

tvoří imidazol (další nukleotid, někdy součástí wobblingu), má neutrální pKa — malá změna pH vede ke změně náboje, takže je často používán jako vypínač závislý na pH, účastní se koordinace kovů

**fenyalanin**

je prekurzorem neurotransmiterů

**serin**

katalyzuje reakce (je to alkohol), především O-glykosylace a fosforylace, nervové plyny jej blokují v acetylcholinesteráze

**threonin**

má dva chirální atomy, také účasten O-glykosylace a fosforylace (je to alkohol)

**tyrosin**

podobný PHE, prekurzor neurotransmiterů, účasten fosforylací (je to alkohol)

**tryptofan**

je největší a tedy dobře konzervovaný, účasten hydrofóbních interakcí (s cukry), prekurzor serotoninu a niacinu

**AK se sírou v postranním řetězci****methionin**

má jen jeden kodon, může být na povrchu oxidován

**cystein**

často v hydrofobním jádře proteinů (přestože je polární), tvoří disulfidické můstky, interaguje s ionty kovů (často v aktivních místech enzymů)

**AK obsahující sekundární amin****prolin**

nemá vodík na dusíku => netvoří vodíkové můstky, nebývá v alfa-helixech ani listech

může být i v konformaci cis (většinou uhlovodíková zbytek AK bývá trans) => může fungovat jako vypínač, protože mění konformaci

jeho cyklus je extrémě rigidní, tvoří zlomy v proteinech

Pro popis aminokyselin se někdy využívá i B (Asn/Asp) a Z (Gln/Glu). Kromě výše zmíněných dvaceti AK se vydělují ještě následující dvě.

**pyrolysin**

kódovaný UAG stop kodonem

**selenocystein**

kódovaný UGA stop kodonem, využíván pro určení struktury proteinů, je v řadě enzymů

## 3.2 Další proteinové struktury

Kromě primární struktury proteinu rozlišujeme ještě sekundární, teriární a kvarterní. Sekundární struktura proteinu je určena lokálními konformacemi jeho částí.

Obrázek 3.8: Prezentace č. 1, slide č. 49

  CO Z ABECEDY ZBÝVÁ?

B, J, Z, X

PO BROUZDÁNÍ DATABÁZEMI:

$$B = ASX = ASN = ASP$$

$$Z = GLX = GLN = GLU$$

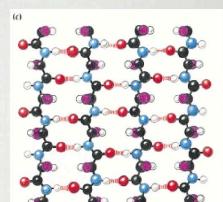
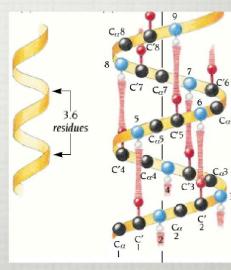
$$X = \text{NEZNÁMÁ AMINOKÝSELINA}$$

ZBÝVÁ TEDY J PRO POSLEDNÍ  
STOP (UAA, OCHRE) KODON

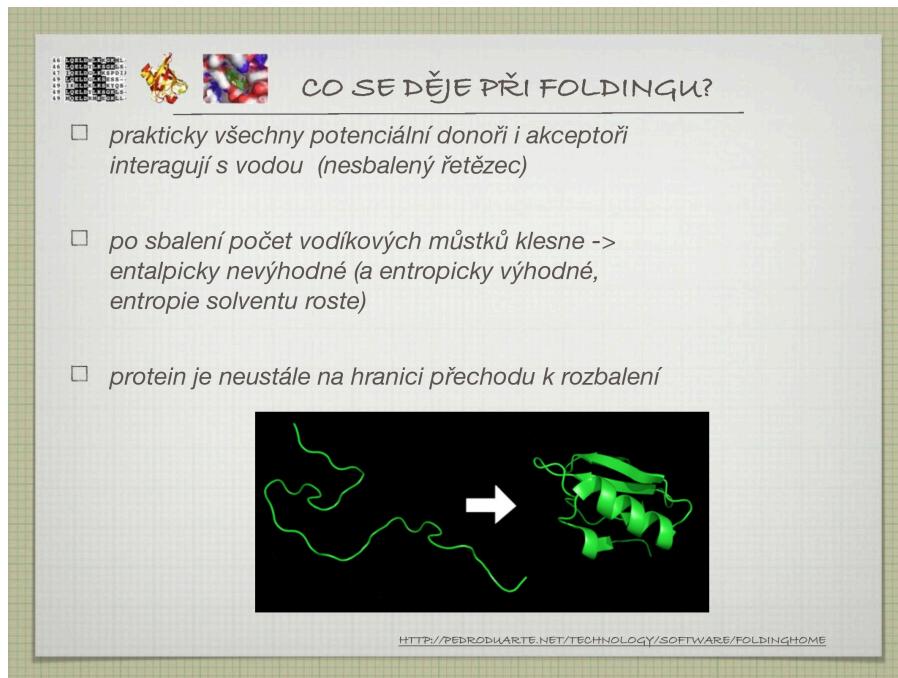
Obrázek 3.9: Prezentace č. 1, slide č. 55

  PROČ TVORÍ PROTEINY SEKUNDÁRNÍ STRUKTURU?

- tvorba stabilního hydrofóbního jádra
- brání amino a karboxylová skupina hlavního řetězce
- neutralizace polárních skupin hlavního řetězce pomocí vodíkových můstků
- omezení entropie vyvážené redukcí negativních entalpických příspěvků spojených s výskytem náboje v hydrofóbním jádře proteinu

Obrázek 3.10: Prezentace č. 1, slide č. 57



### Důvody vzniku

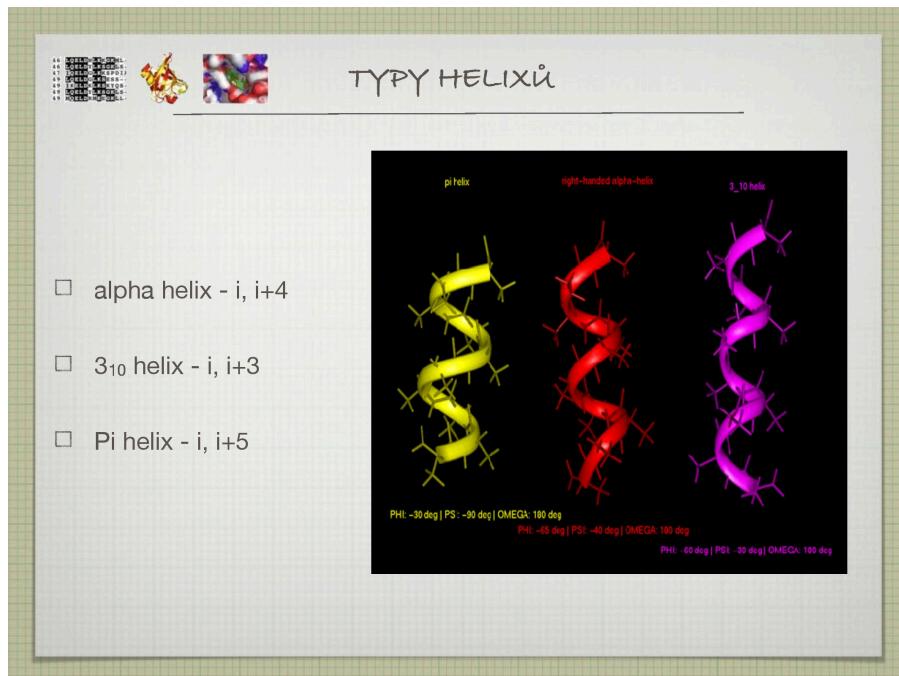
- snaha o tvorbu stabilního hydrofobního jádra
  - důvod: entropie klesne, ale tento pokles je vyvážen růstem entalpie, která je negativně ovlivněná výskytem náboje v jádře proteinu
  - způsob: neutralizace polárních amino a karboxylových skupin na hlavním řetězci vznikem vodíkových můstků

### Vodíková vazba a stabilizace

- síla vodíkové vazby závisí na typu atomu a geometrii vazby
  - cca 1–60 kJ/mol, v proteinech většinou okolo 10 kJ/mol
  - se zvětšujícím se úhlem vazby klesá její síla: odklon o 20° snižuje energii o 10%

### Folding

Obrázek 3.11: Prezentace č. 1, slide č. 59



1. protein je nesbalený, všichni donoři i akceptoři reagují s vodou
2. protein se sbalí, počet vodíkových můstků klesne
  - entalpicky nevýhodné, ale entropicky výhodné
3. protein je nyní neustále na hranici rozbalení, aby bylo možné jej případně rozložit (a nezůstal v buňce napořád)

## Helix

- teoreticky popsaný Linusem Paulingem
- model potvrzen strukturou myoglobinu
- může být levotočivý (Ala, Leu, Val) i pravotočivý (Gly, Pro)
- několikrát druhů
  - $\alpha$  helix, nad sebou jsou AK  $i$  a  $i + 4$
  - $3_{10}$  helix, nad sebou jsou AK  $i$  a  $i + 3$
  - $\pi$  helix, nad sebou jsou AK  $i$  a  $i + 5$
- zobrazování pomocí helix-wheel diagramu
  - barvení podle typu AK

Obrázek 3.12: Prezentace č. 1, slide č. 60

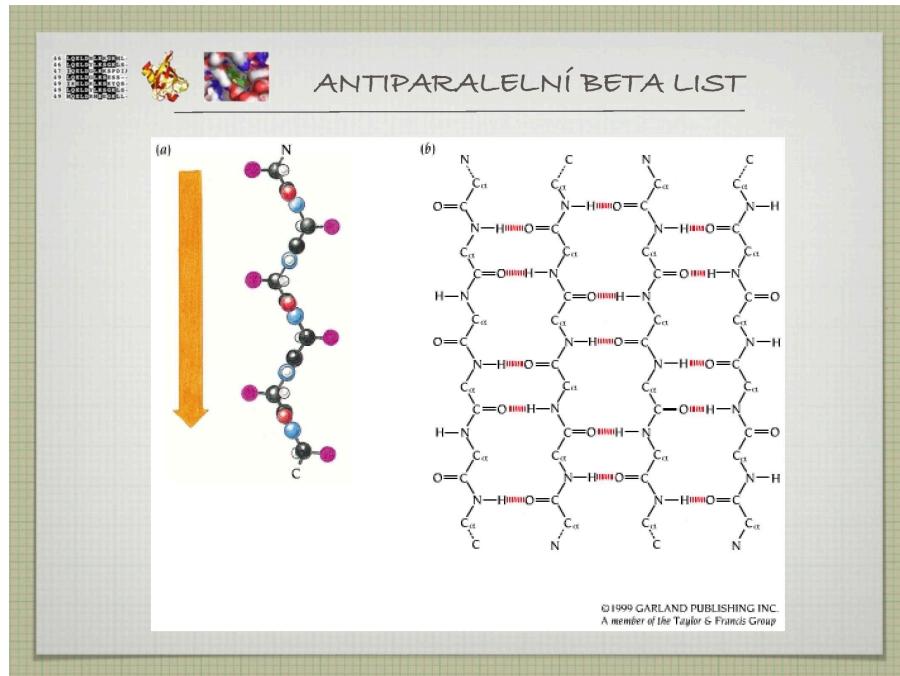


- všechny hydrofilní budou na jedné straně, hydrofobní na druhé (vznik snopců)

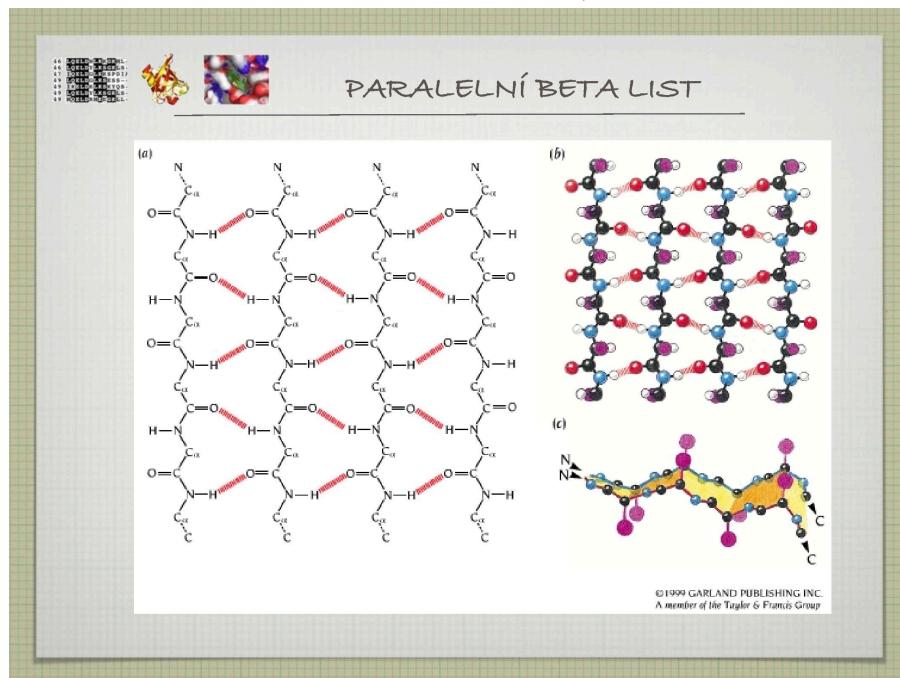
### Beta list

- teoreticky jej popsal William Astbury a Linus Pauling
- složen z  $\beta$  hřebenů (strand)
- uprostřed bývají Tyr, Thr, Trp, Val a Ile, na krajích spíše Pro
- vzdálenost mezi  $C_\alpha$  asi 3,5 Å
- dvě formy
  - paralelní, jsou méně stabilní (vázané atomy napříč hřebeny nejsou přesně naproti sobě)
  - antiparalelní, jsou stabilnější, planární
- vznik  $\beta$  barelu
  - poslední hřeben se váže na první ve stejném listu, vzniká kanál
  - často bývá stočený
  - často v membránách

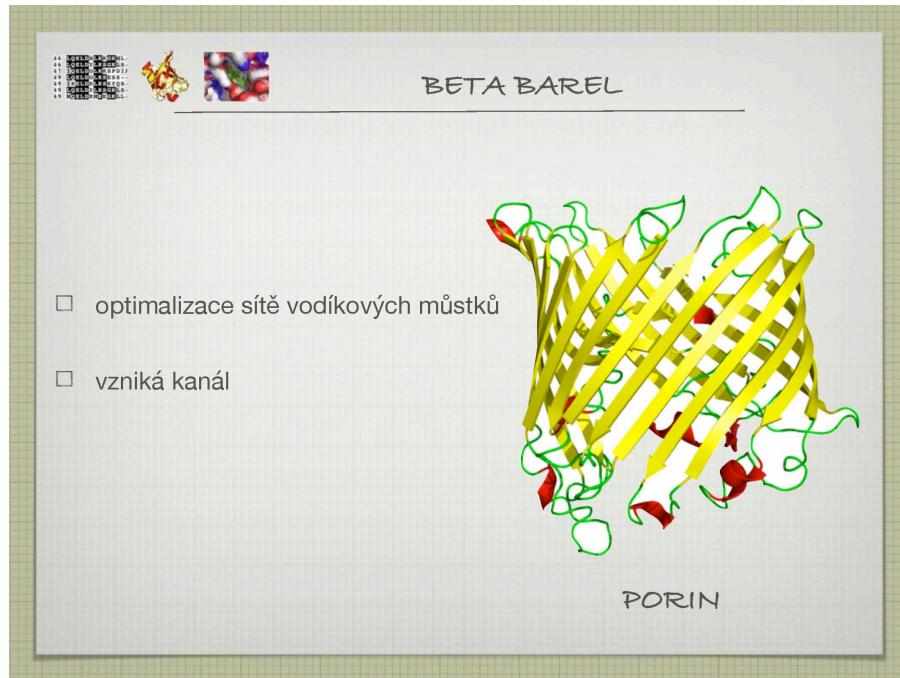
Obrázek 3.13: Prezentace č. 1, slide č. 63



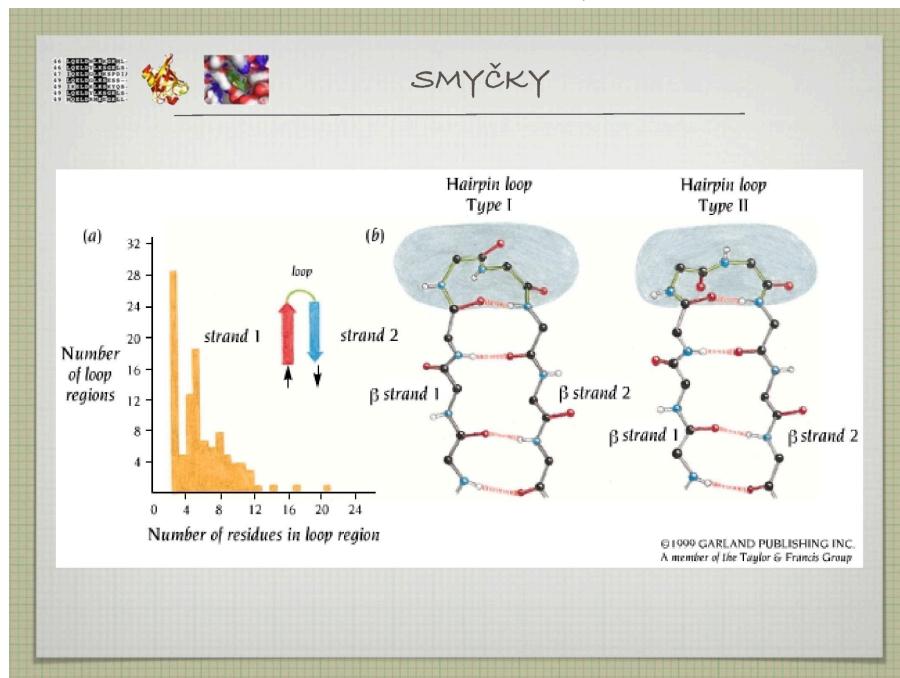
Obrázek 3.14: Prezentace č. 1, slide č. 64



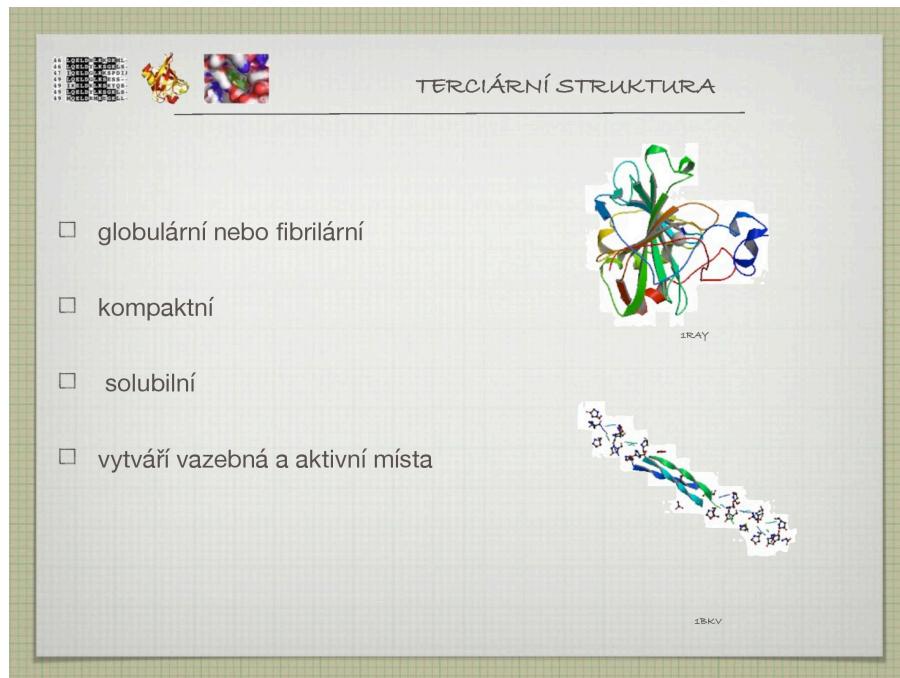
Obrázek 3.15: Prezentace č. 1, slide č. 65



Obrázek 3.16: Prezentace č. 1, slide č. 66



Obrázek 3.17: Prezentace č. 1, slide č. 67



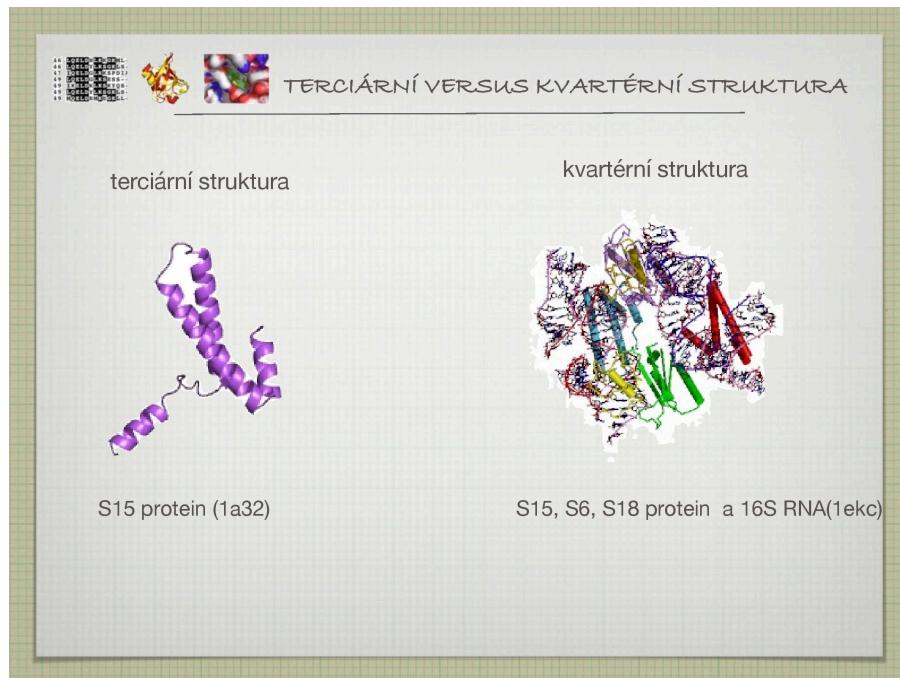
### Smyčky

- nepravidelné struktury
- často hodně Gly (protože je malý)
- spojují helixy a listy
- bývají krátké, většinou kolem 4,5 ÅK
- existuje více druhů (vlásenky, loopy, atd.)

### Terciární struktura proteinu

- někdy také konformace, topologie nebo folding
- celková 3D struktura proteinu
- dva typy
  - globulární, nepříliš uspořádaná
  - fibrilární, uspořádaná do vláken
- rozhoduje o solubilitě
- vytváří vazebná a antivazebná místa

Obrázek 3.18: Prezentace č. 1, slide č. 68



Kvarterní struktura proteinu popisuje uspořádání několika terciárních struktur (například v dimerech).

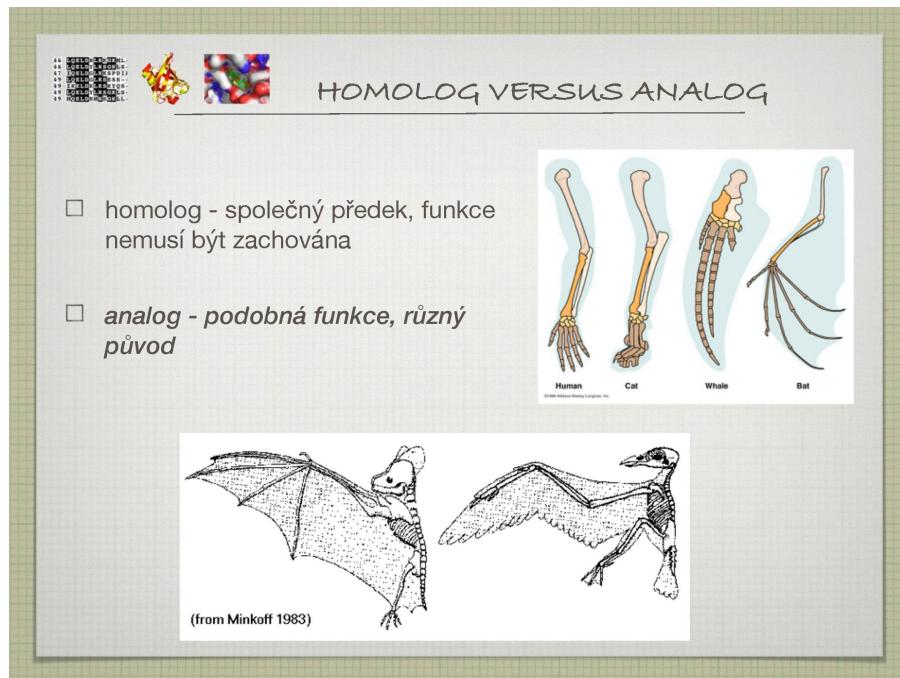
## 4 Sequence alignment

Přednáška č. 4

Základní bioinformatická metoda užíváná k porovnání dvou sekvencí (DNA, proteinů). Obecně se jedná o nějaké seřazení sekvencí pod sebe. Dobrý alignment dvou sekvencí má však důležitou vlastnost: pod sebou jsou jednotky (nukleotidy, AA), které se vyvinuly ze stejného předka. Někdy byly určité jednotky v průběhu evoluce přidány nebo odebrány, což se v rámci alignmentu značí pomlčkami (viz níže).

Než se dostaneme k samotnému procesu alignmentu (tj. zjištování, které jednotky a

Obrázek 4.1: Prezentace č. 2, slide č. 26



potažmo celé sekvence jsou evolučně spřízněné), ukážeme několik jeho praktických využití.

### A je homolog B

A a B mají společného předka, jejich původní funkce však nemusí být zachována. Homologii můžeme (opatrně) odvodit z vysokého procenta sekvenční identity A a B (viz dále). Musíme však dát pozor na paralogii.

Na základě homologie můžeme (opatrně) odvodit funkční a strukturní podobnost. Například můžeme hledat homology problémových lidských proteinů v modelových organismech, na které budeme cílit vyvýjená léčiva.

### A je ortolog B

Poddruh homologie; A a B vznikly speciací ze společného předka, jejich funkce by tedy měla být zachována.

### A je paralog B

Poddruh homologie; A a B vznikly genovou duplikací ze společného předka —

Obrázek 4.2: Prezentace č. 2, slide č. 27

**ORTOLOG VERSUS PARALOG**

- **ortolog** - vzniká speciací, funkce pravděpodobně zachována
- **paralog** - vzniká genovou duplikací, funkce nemusí být zachována
- **ohnolog** - vzniká celogenomovou duplikací
- **xenolog** - vzniká horizontálním transferem

jejich funkce tedy nemusí být zachována (protože jedna kopie genu ji zastane, zatímco A a B se mohli vyvinout v něco jiného).

### A je ohnolog B

Podobný vztah jako paralog, vzniká ale celogenomovou duplikací.

### A je xenolog B

A a B vznikly horizontálním transferem (například mezidruhovým).

### A je analog B

A a B mají podobnou funkci, avšak je to jen náhoda — společného předka nemají.

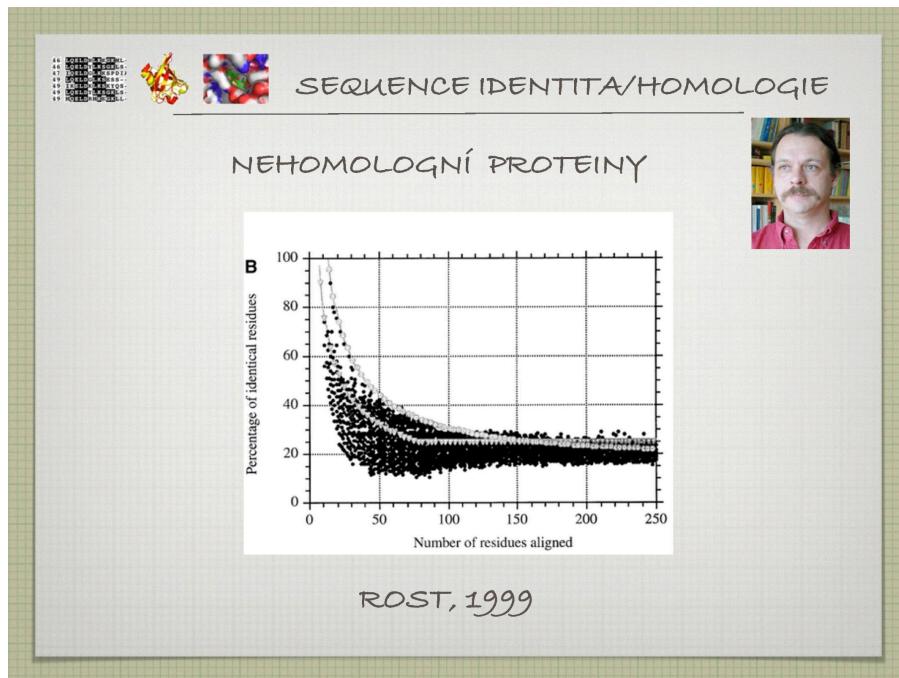
### globální alignment

Srovnávání celé sekvence.

### lokální alignment

Srovnávání pouze částí sekvence, vybírá kousky, které k sobě sedí nejlépe.

Obrázek 4.3: Prezentace č. 2, slide č. 31

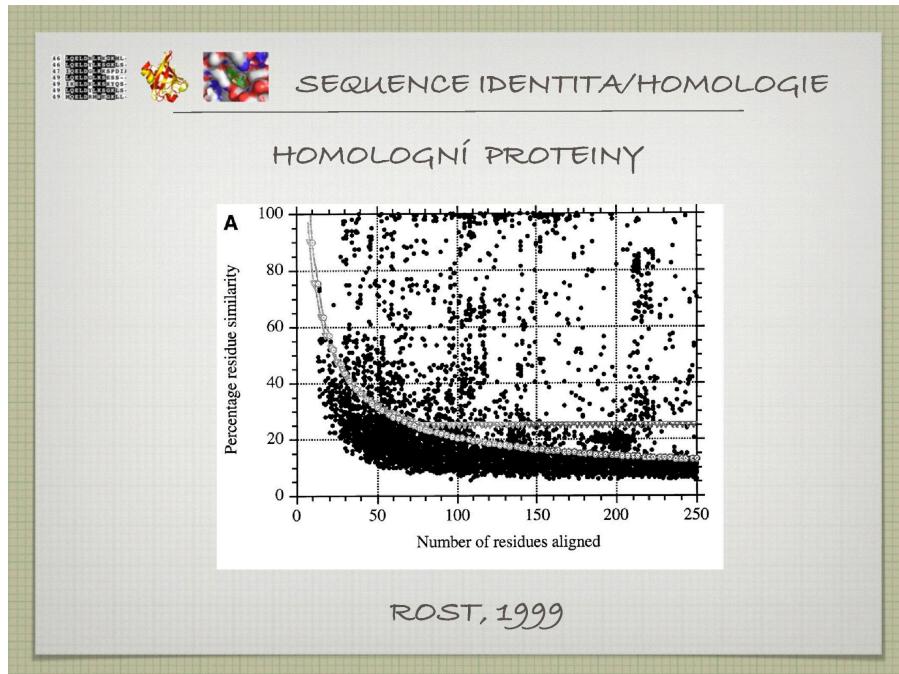


### Důvod srovnávání sekvencí

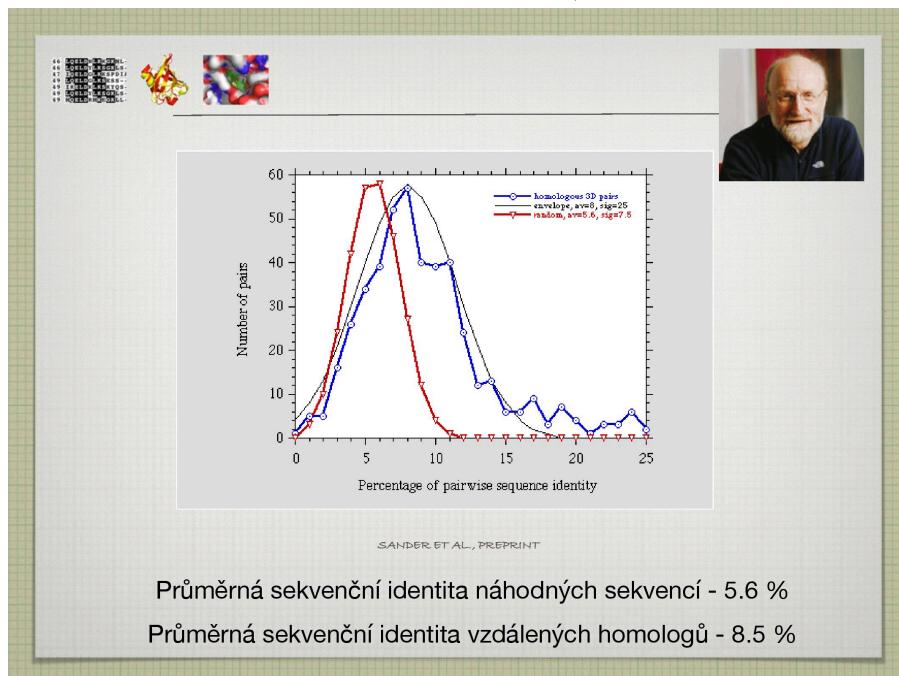
- nalezení evoluční podobnosti (analogie, homologie)
- získání informací o struktuře, funkci, a evolučním vývoji proteinu
  - pomocí srovnání s proteiny, které už mají známou strukturu, funkci a původ
- nalezení aktivních (konzervovaných) míst
- nalezení mutantů
- možno pomocí něj dát smysl velkému množství biologických dat

Snažíme se ze znalosti struktury a funkce určitého proteinu odvodit funkci jiného, podobného (homologního) proteinu. To, jestli je vůbec možné z kvantitativní veličiny sekvenční identity (SI) vyvodit kvalitativní rozhodnutí o homologii, zkoumali Chotia, Lest (1986) a Rost (1999). Zjistilo se, že změny ve struktuře jsou korelovány se změnami v sekvenci, neboť z %SI si můžeme troufnout odvodit homologii a podobné vztahy, a z nich poté hádat věci jako je funkce nebo evoluční původ.

Obrázek 4.4: Prezentace č. 2, slide č. 32



Obrázek 4.5: Prezentace č. 2, slide č. 39



### Sekvenční identita (SI)

- hranice relevantní pro potenciální homologii (určené Rostem)
  - $SI > 35\%$  naznačuje možnou homologii
  - $20\% < SI < 35\%$  je takzvaná twilight zone, homologie je možná
  - $SI < 20\%$ , o homologii nemůžeme s jistotou říct vůbec nic
- průměrná SI náhodných sekvencí je asi 5,9%
- průměrná SI vzdálených homologů je asi 8,5%
- existuje více metod, jak SI vypočítat (a SI se u každé metody liší)
  - počet náhodných pozic / délka alignmentu
  - počet shodných pozic / délka kratší sekvence
  - počet shodných pozic / průměrná délka sekvencí

Jakým způsobem se rozhodnout, když jsme v twilight zóně? Na to existuje několik triků.

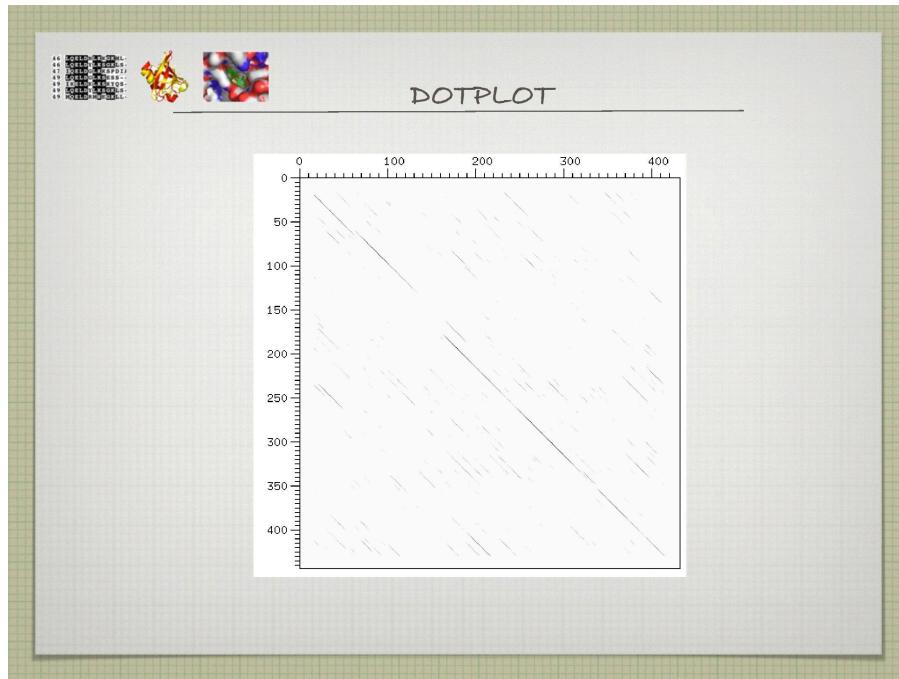
### Jak z twilight zóny?

- pokud jsou vyměněny kladně nabité AK za jiné kladně nabité AK, alanin za valin atp. (zkrátka tzv. konzervativní záměny), sekvence jsou nejspíše homologní
- pokud se snažíme zjistit něco více o homologii A a B, stačí najít C, které je homologické s A a zároveň je i homologické s B; z toho totiž plyne, že A i B jsou také homologické

Výše bylo zmíněno, že srovnávání sekvencí funguje jak pro proteiny, tak pro DNA. Přesto se ale častěji, minimálně k určování homologie, používají proteiny, a to ze dvou důvodů:

1. protože AK je dvacet, je menší šance, že budou na jednom místě dvě shodné AK náhodou (oproti čtyřem nukleotidům v DNA, kde je náhodná shoda pravděpodobnější)
2. různé kodony kódují stejné AK, čili určité změny v DNA kódu se vůbec nemusí projevit v jeho exprimaci; jinými slovy, i relativně hodně odlišné sekvence mohou kódovat stejné, nebo velice podobné proteiny

Obrázek 4.6: Prezentace č. 2, slide č. 45



Srovnávání sekvencí DNA ale má svá uplatnění. Používá se v místech, která se v proteomu vůbec neobjeví; při zkoumání regulačních oblastní genů a definování genů a při celogenomovém srovnávání.

Metody jsou v zásadě dvě, dotploty, které slouží spíše k hrubému odhadu situace, a pairwise sequence alignment, což je aby se řeklo the real deal.

## 4.1 Dotplot

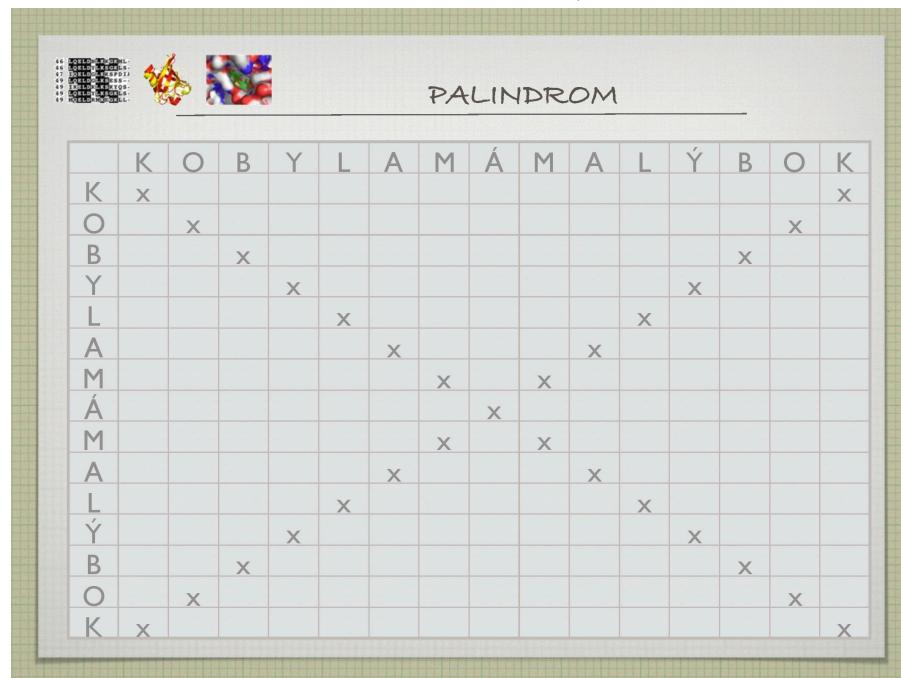
Nejpřímější a nejjednodušší metoda: do tabulky se zaznamenávají místa, na kterých jsou dvě sekvence shodné (viz slidy). Někdy se místo jednotlivých stavebních jednotek sekvencí používají celé domény na sekvencích.

Na dotplotu byly sledovány i první dvě známé struktury, hemoglobin a myoglobin.

Obrázek 4.7: Prezentace č. 2, slide č. 50



Obrázek 4.8: Prezentace č. 2, slide č. 55



Obrázek 4.9: Prezentace č. 2, slide č. 57



### Silné stránky

- jednoduchý a rychlý
- odhaluje repetice, inverze, přeházené domény, oblastní s nízkou komplexitou
- poskytuje návod, kde má vůbec smysl dělat podrobnější sequence alignment
- vhodný pro odhad podobnosti sekvencí

### Slabé stránky

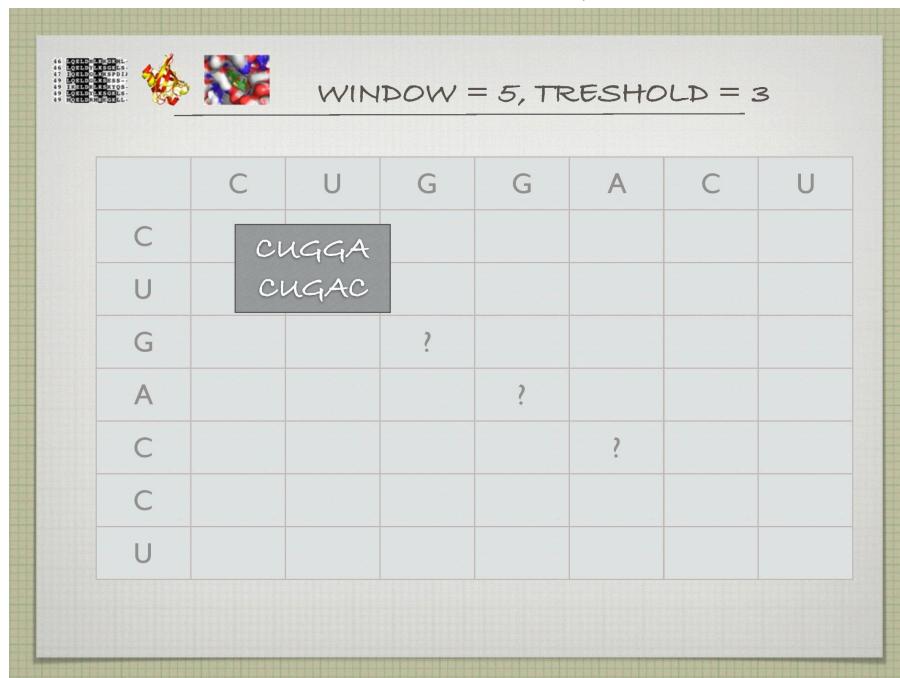
- neumí rekonstruovat evoluci (odhalovat homologii atp.)
- generuje příliš mnoho signálů, velký šum
  - toto se často ”řeší” tak, že se koukáme hned na několik nukleotidů za sebou, a křížek v daném políčku uděláme pouze tehdy, když v tomto našem klouzavém okně je více než  $k$  shod
- ukazuje i náhodné podobnosti

V praxi se často používá self-dotplot, tedy dotplot, kde je sekvence srovnávána sama

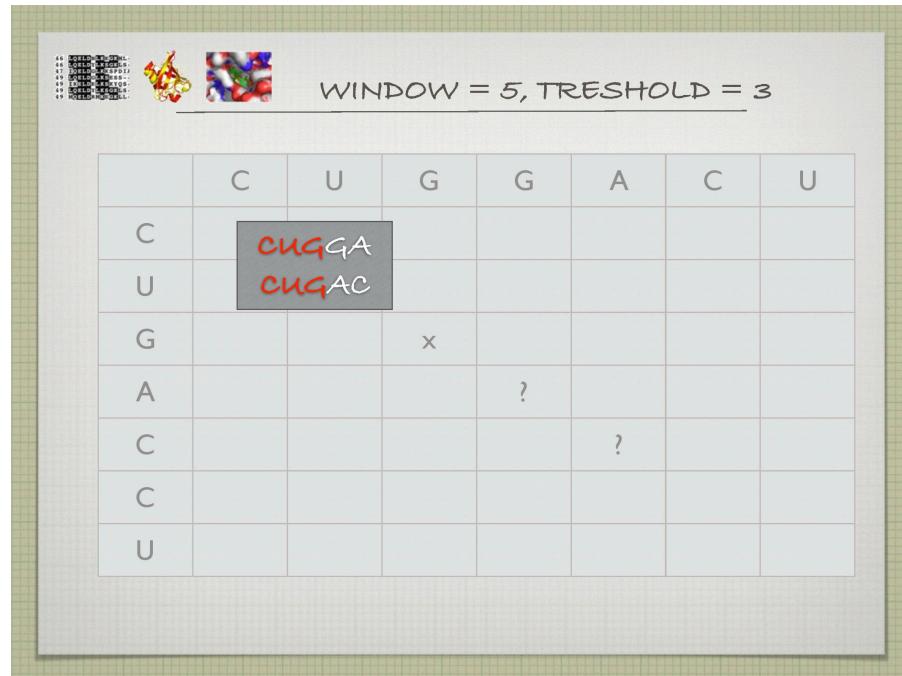
Obrázek 4.10: Prezentace č. 2, slide č. 60



Obrázek 4.11: Prezentace č. 2, slide č. 61



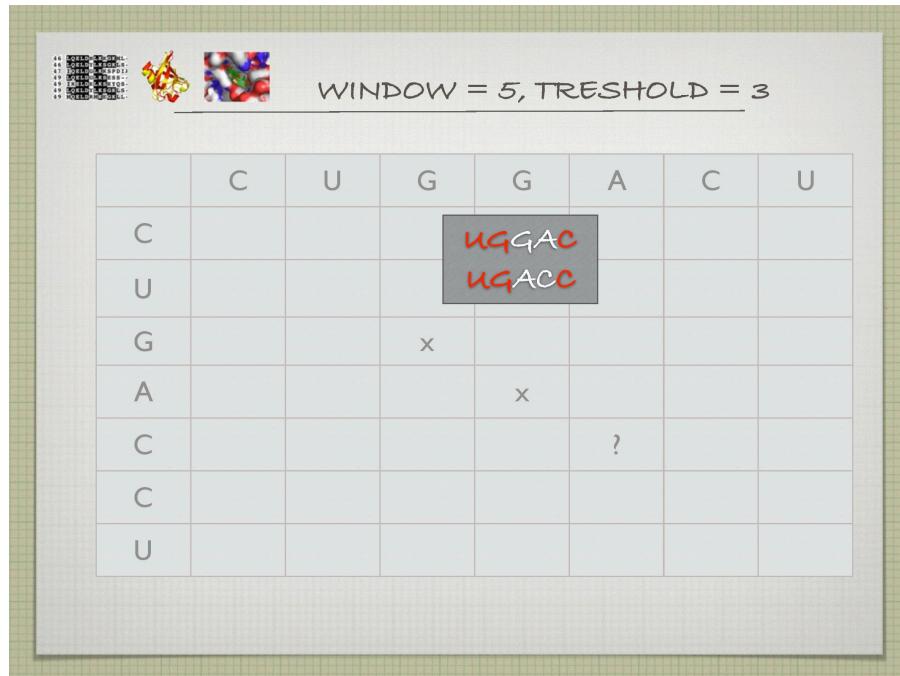
Obrázek 4.12: Prezentace č. 2, slide č. 62



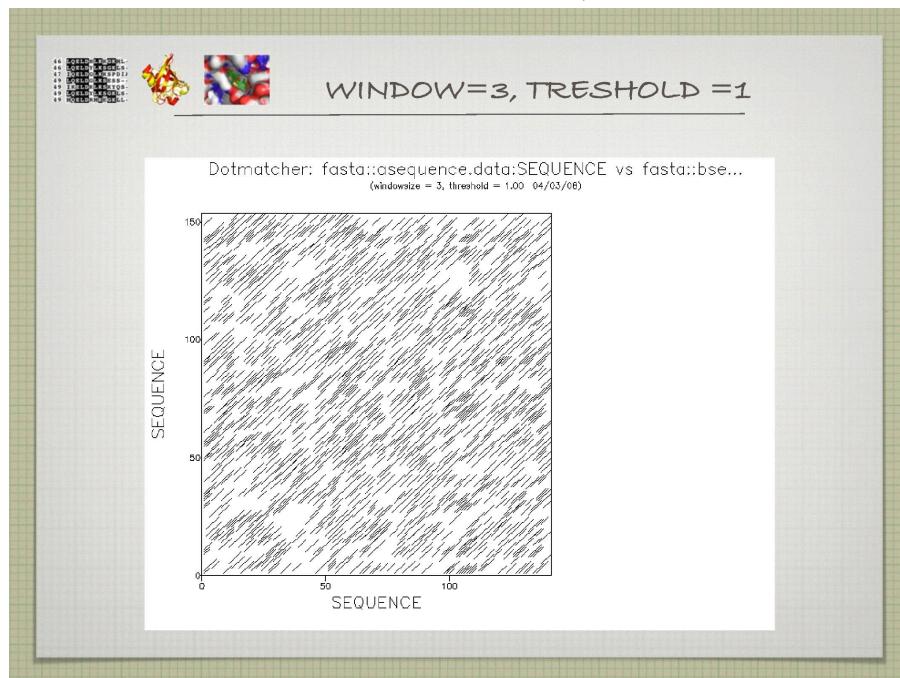
Obrázek 4.13: Prezentace č. 2, slide č. 63



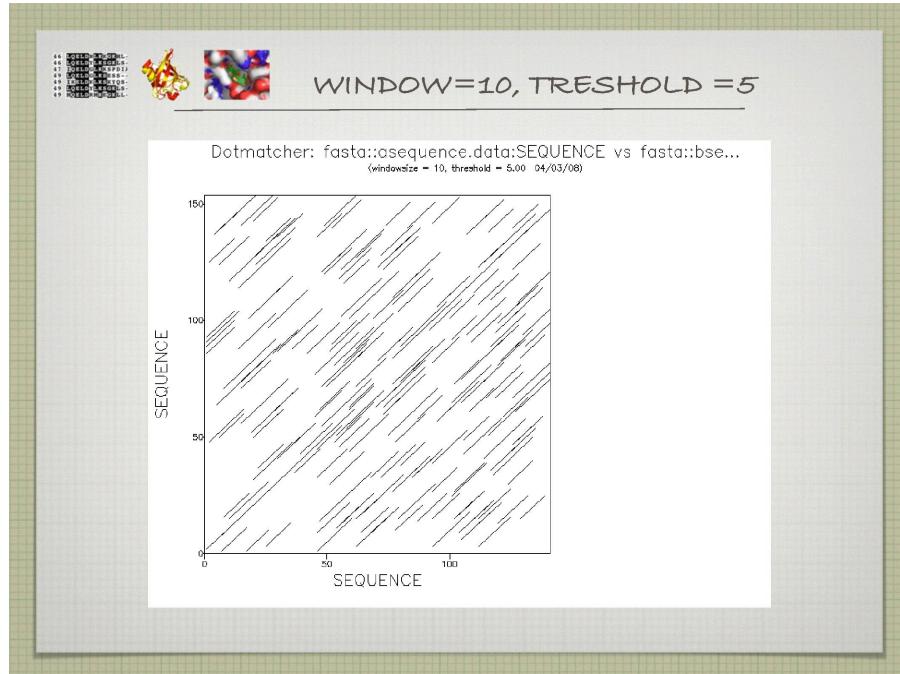
Obrázek 4.14: Prezentace č. 2, slide č. 64



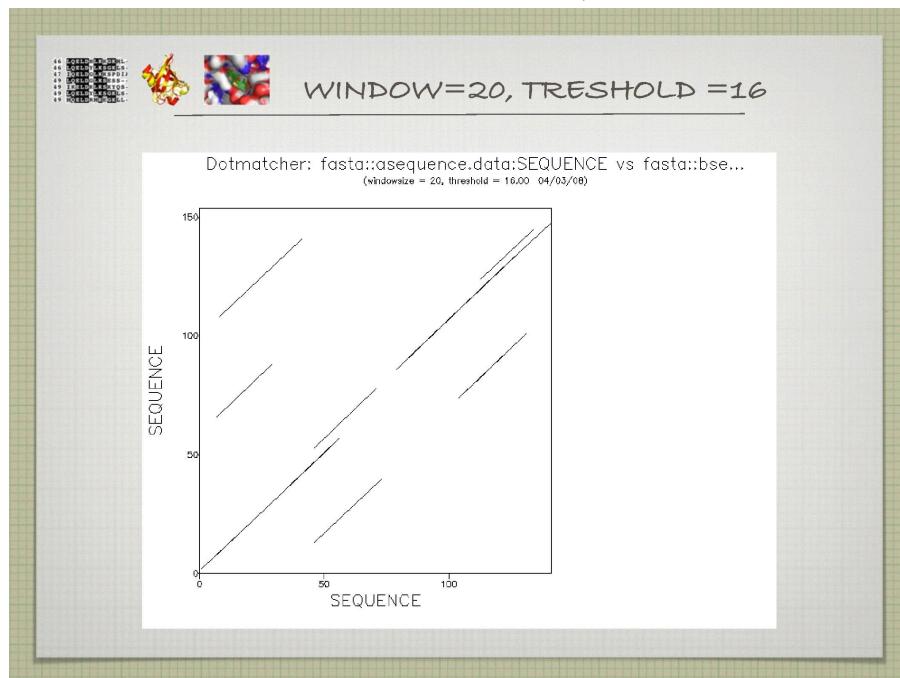
Obrázek 4.15: Prezentace č. 2, slide č. 67



Obrázek 4.16: Prezentace č. 2, slide č. 68



Obrázek 4.17: Prezentace č. 2, slide č. 69



se sebou. Ten opět vyhledává symetrické úseky, repetice, odhaluje místa s nízkou komplexitou a palidromy.

## 4.2 Pairwise sequence alignment

Může samozřejmě probíhat i na DNA, pro jednoduchost jej ale popíšeme pouze na proteinech. Pro DNA funguje analogicky.

Předpokládáme, že sekvence A a B mají společného předka. Poté, když je srovnáme (naalignujeme) "pod sebe", můžeme na každém jednotlivém místě pozorovat následující:

- shoda: AK v A i v B jsou na daném místě stejné
- neshoda: AK v A je na daném místě odlišná od AK v B
- mezera (gap): v jedné ze sekvencí došlo při vývoji od společného předka k inzerci nebo deleci

**META** Gap (mezera v sekvenci při procesu alignmentu) se Švédsky řekne lucka.

Proces alignmentu je vlastně proces umisťování mezer a pozorování toho, jak si poté dvě sekvence navzájem odpovídají. Příklad alignmentu:

```
VLSEGKTEAPV [ . . . ]
| | | .. | | |
VLSPA----PV [ . . . ]
```

Toto je další příklad alignmentu těchto sekvencí; tentokrát dosti nepovedeného:

```
---VLSEGKTEA--PV [ . . . ]
. .
V-LS--PA----PV-- [ . . . ]
```

Substituce jedné AK za jinou je pravděpodobnější než inzerce/delece. V rámci substitucí je pravděpodobnější substituce podobných jednotek (Val <-> Leu, G <-> A) než těch nepodobných (Trp <-> Gly, G <-> C).

Obrázek 4.18: Prezentace č. 2, slide č. 87



**SKÓRE PRO NÁŠ ALIGNMENT**

	A	C	G	T
A	3	-3	0	-3
C	-3	3	-3	0
G	0	-3	3	-3
T	-3	0	-3	3

SHODA = 3  
 NESHODA (Pu/Pu; Py/Py) = 0  
 NESHODA (Pu/Py) = -3  
 GAP PENALTY (KONSTANTNÍ) = -5

Obrázek 4.19: Prezentace č. 2, slide č. 97



Alignment	Shoda (+3)	Pu/Pu (0)	Pu/Py (-3)	Gap (-5)	Score
-----GTACGTACG GGCATGAGG-----	0	0	0	18	-90
GTACGTACG GGCATGAGG	3	0	5	0	-6
-GTACGTACG- GGC-ATG-AGG	2	1	4	4	-26
GT-ACGTACG GGCATG-AGG	5	1	2	2	-1
GTACG--T-ACG G---GCATGAGG	5	0	1	6	-18

### Měřítko kvality alignmentu

- alignmentů je nekonečně mnoho, musíme vybrat ty nejpravděpodobnější, po kud z nich chceme něco vyvzovat (společnou strukturu, funkci atp.)
- většinou alignment skórujeme po jednotkách
  - shoda/neshoda jsou za určitý počet bodů (klidně záporných)
  - mezery jsou za záporné body (tzv. gap penalty, GP), většinou za začátek mezery je více záporných bodů než za její rozšíření
- skóre pro všechny kombinace shod/neshod bývá uloženo v tabulkách
- tyto tabulky, společně s určením hodnot gap penalty, velice ovlivňují výsledný (vybraný) alignment
  - z toho plyne snaha o optimalizaci tabulek i GP tak, aby co nejvíce odpovídali biologickým empirickým datům (netvořily nesmyslné alignmenty)
  - optimální tabulky/GP se liší protein od proteinu
  - vznikají experimentální variabilní GP založeny na strukturních datech proteinů (v pravidelných sekundárních strukturách je nízší pravděpodobnost výskytu mezery)

K hledání optimálního (nebo suboptimálního) alignmentu používá algoritmus, který projde mnoho různých možných alignmentů a vybere z nich ten s nejvyšším skóre dle přidělené skórovací tabulky. Pozor, ani nejlepší alignment nemusí odpovídat reálu.

#### 4.2.1 Skórovací tabulky

Neboli scoring matrices.

#### Určení hodnot skóre

- skóre původně určeno z fyzikálněchemických, tedy teoretických, vlastností AK
- nyní máme mnoho empirických dat, skóre tvoříme na základě nich
  - skóre záměny jedné AK za jinou je tedy postaveno na základě pravděpodobnosti toho, že tato záměna v reálu proběhne, kterou zjistíme z pozorovaných sekvencí

Obrázek 4.20: Prezentace č. 3, slide č. 26



**PAM III**

□	současně používané PAM tabulky z roku 1978 (málo sekvencí)
□	nové tabulky generované stejným způsobem z mnohem většího množství dat nesou označení <b>PET (1992)</b>
	<b>PAM 250</b>
	A R N D C Q E G H I L K M F P S T W Y V B Z X * A -2 6 0 -1 -4 1 -1 -3 2 -2 -3 0 -4 0 0 -1 2 -4 -2 -1 0 -1 -8 R -2 6 0 -1 -4 1 -1 -3 2 -2 -3 0 -4 0 0 -1 2 -4 -2 -1 0 -1 -8 N 0 0 2 2 -4 1 1 0 2 -2 -3 1 -2 -3 0 1 0 -4 -2 -2 2 1 0 -8 D 0 -1 2 4 -5 2 3 1 1 -2 -4 0 -3 -6 -1 0 0 -7 -4 -2 3 3 -1 -8 C -2 -4 -4 -5 12 -5 -5 -3 -3 -2 -6 -5 -5 -4 -3 0 -2 -8 0 -2 -4 -5 -3 -8 Q 0 1 1 2 -5 4 2 -1 3 -2 -2 1 -1 -5 0 -1 -1 -5 -4 -2 1 3 -1 -8 E 0 -1 1 3 -5 2 4 0 1 -2 -3 0 -2 -5 -1 0 0 -7 -4 -2 3 3 -1 -8 G 1 -3 0 1 -3 -1 0 5 -2 -3 -4 -2 -3 -5 0 1 0 -7 -5 -1 0 0 -1 -8 H -1 2 2 1 -3 3 1 -2 6 -2 -2 0 -2 -2 0 -1 -1 -3 0 -2 1 2 -1 -8 I -1 -2 -2 -2 -2 -2 -3 -2 5 2 -2 2 1 -2 -1 0 -5 -1 4 -2 -2 -1 -8 L -2 -3 -3 -4 -6 -2 -3 -4 -2 2 6 -3 4 2 -3 -3 -2 -2 -1 2 -3 -3 -1 -8 K -1 3 1 0 -5 1 0 -2 0 -2 -3 5 0 -5 -1 0 0 -3 -4 -2 1 0 -1 -8 M -1 0 -2 -3 -5 -1 -2 -3 -2 2 4 0 6 0 -2 -2 -1 -4 -2 2 -2 -2 -1 -8 F -3 -4 -3 -6 -4 -5 -5 -2 1 2 -5 0 9 -5 -3 -3 0 7 -1 -4 -5 -2 -8 P 1 0 0 -1 -3 0 -1 0 0 -2 -3 -1 -2 -5 6 1 0 -6 -5 -1 -1 0 -1 -8 S 1 0 1 0 0 -1 0 1 -1 -3 0 -2 -3 1 2 1 -2 -3 -1 0 0 0 -8 T 1 -1 0 0 -2 -1 0 0 -1 -2 0 -1 -3 0 1 3 -5 -3 0 0 -1 0 -8 W -6 2 -4 -7 -8 -5 -7 -7 -3 -5 -2 -3 -4 0 -6 -2 -5 17 0 -6 -5 -6 -4 -8 Y -3 -4 -2 -4 0 -4 -4 -5 0 -1 -4 -2 7 -5 -3 -3 0 10 -2 -3 -4 -2 -8 V 0 -2 -2 -2 -2 -2 -2 -1 -2 4 2 -2 2 -1 -1 0 -6 -2 4 -2 -2 -1 -8 B 0 -1 2 3 -4 1 3 0 1 -2 -3 1 -2 -4 -1 0 0 -5 -3 -2 3 2 -1 -8 Z 0 0 1 3 -5 3 3 0 2 -2 -3 0 -2 -5 0 0 -1 -6 -4 -2 2 3 -1 -8 X 0 -1 0 -1 -3 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -8 * -8 -1

Tabulky jsou tedy symetrické — nejsme schopni z empirických dat zjistit, jakým směrem proběhla substituce (pokud máme na daném místě v jedné sekvenci Ille a v druhé Val, nevíme, které AK z těch dvou tam byla původně, a která tam bylo substituována dodatečně).

### Tabulka PAM

- název z percent accepted mutations
- autorkou je Margarette Dayhoff (70. léta)
- založená na pravděpodobnostních mírách mutace kalkulovaných z globálních alignmentů blízce podobných sekvencí

$$\text{hodnota}(X, Y) = \log \frac{\text{počet pozorovaných záměn } X \text{ za } Y}{\text{počet očekávaných záměn}}$$

- pozitivní je tedy skóre jen u zbytků, u kterých záměna proběhla častěji než by bylo očekáváno při náhodném zaměňování
- PAM tabulek je mnoho

Obrázek 4.21: Prezentace č. 3, slide č. 23

KALKULACE PRAVDĚPODOBNOSTI

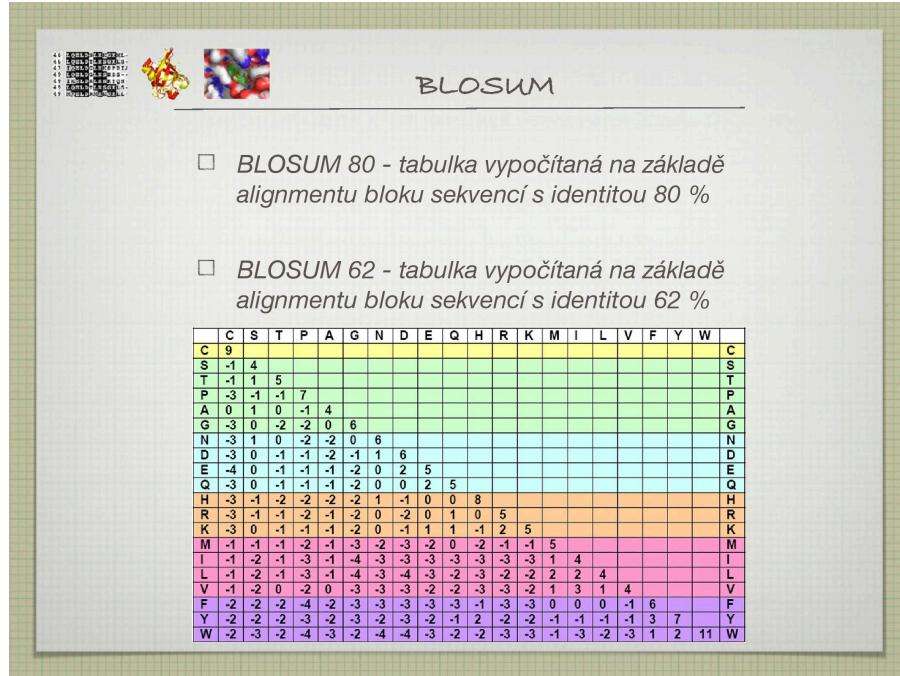
- skóre pro záměnu valinu za isoleucin
- pravděpodobnost záměny pokud jsou sekvence příbuzné (pozorovaná) = 0.03
- výskyt aminokyselin v populaci (databáze, proteom) = 0.1 a 0.05
- poměr pravděpodobností =  $0.03/(0.1 \cdot 0.05) = 6x$  větší pravděpodobnost záměny V -> I než očekáváno při náhodných záměnách
- skóre = desítkový logaritmus \* 10 a zaokrouhlen na nejbližší celé číslo =  $10 \cdot \log 6 = 10 \cdot 0.778 = 8$

- PAM  $x \rightarrow x$  AK ze 100 bylo nahrazeno
- nízké  $x$  se tedy hodí pro evolučně blízké sekvence, vysoké  $x$  pro ty vzdálené
- nejoblíbenější PAM 250
  - \* každá AK byla v průměru zaměněna 2,5 krát
  - \* některé AK jsou ale mutovány vícekrát, proto jsou i takové sekvence asi z 20% složené ještě z původních AK
- nyní už existují novější tabulky, které jsou generované stejným způsobem, ale z většího množství dat: PET

### Tabulka BLOSUM

- název z BLOck SUbstitution Matrix
- autoři Henihoff a Henihoff (90. léta)
- založena na experimentálních datech, není extrapolována jako některé PAM tabulky
- opět více druhů

Obrázek 4.22: Prezentace č. 3, slide č. 29



- BLOSUM  $x$ : založena na lokálních alignmentech bloků AK s SI =  $x$  (u homologních proteinů), bez mezer
- nejoblíbenější BLOSUM 62 (tedy popisující proteiny se sekvenční identitou 80%)

### Další tabulky

- STR, SDM
  - informace ze struktur
  - záměny ve smyčkách jsou pravděpodobnější než v helixech a listech
- PHAT, SLIM
  - vhodné pro specifický výběr proteinů (například hydrofobní)
  - SLIM je asymetrická

Výběr tabulky se snažíme přizpůsobit sekvencím, které srovnáváme, abychom získali co nejlepší výsledky; především rozlišujeme evolučně blízké sekvence od těch vzdálených. Krátké sekvence skórujeme podle tabulek pro krátký evoluční čas.

Obrázek 4.23: Prezentace č. 3, slide č. 33

## DALŠÍ SKÓROVACÍ TABULKY

- *STR, SDM* - využívají informace ze struktur (záměny ve smyčkách pravděpodobnější než v helixech nebo beta listech)
- tabulky pro specifický typ proteinů - např. hydrofóbní proteiny (*PHAT, SLIM*). *SLIM* - **nesymetrická**
- databáze dostupných maticí - AA index (9.1) - 94 matric **SLIM 161**

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A .5 -.8 -.5 -.10 .3 -.7 -.11 0 -.5 1 1 -.10 1 1 -.5 .2 1 -.2 -.5 2	R -.3 10 -.4 -.2 -.3 -.1 -.1 -.4 -.4 -.2 -.1 -.1 -.2 -.7 .4 -.4 -.3 -.2 -.3 .2	N -.1 5 -.8 -.2 0 -.2 -.2 -.7 -.2 -.2 -.1 -.1 -.6 0 1 -.5 .2 0 -.2 -.2 .2	D -.2 2 -.5 1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.2 -.1 -.1 -.1 -.1 -.1 -.1 .2	C 2 -.8 -.5 -.11 11 .1 -.7 -.12 -.3 -.8 0 1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 .1	Q -.1 -.2 0 -.3 0 7 -.4 -.2 0 0 0 0 -.4 2 2 -.4 0 1 -.1 4 0	E -.1 -.7 -.2 3 .2 -.1 7 -.3 -.1 -.2 -.2 -.2 -.8 -.1 0 -.4 -.1 -.3 -.1 -.1 -.2 .2	G 1 -.4 -.7 0 -.5 -.10 -.8 7 -.1 0 -.8 0 1 -.5 -.1 -.1 -.1 -.1 -.3 -.1 -.1 -.1 .2	H 1 -.5 3 -.5 -.2 1 -.5 -.4 10 .1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 .2	I 1 -.9 -.7 -.11 11 -.1 -.7 -.12 -.4 -.7 6 3 .1 4 2 2 -.3 -.1 -.1 -.1 -.1 -.4 .2	L 1 -.8 -.6 -.10 1 -.1 -.7 -.12 -.4 -.7 4 .5 11 4 3 -.7 -.3 -.1 -.1 -.1 -.2 .2	K 1 -.1 1 -.4 0 0 1 1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 .2	M 1 -.7 -.6 -.10 1 -.5 -.11 -.3 -.7 4 4 4 -.11 7 3 -.7 -.1 -.1 -.1 -.1 -.1 .2	F 1 -.9 -.5 -.10 2 -.6 -.11 -.3 4 1 2 2 2 8 -.7 -.2 -.2 -.2 -.3 4 0	P -.3 -.9 -.7 -.9 -.7 -.8 -.10 -.4 -.9 -.2 -.3 -.8 -.3 -.2 11 -.5 -.3 -.3 -.4 -.2 .2	S 2 2 -.1 4 -.4 -.5 -.9 0 0 0 0 0 0 0 0 0 0 0 0 0 0	T 2 -.7 -.3 -.8 2 .6 -.10 -.2 -.5 2 2 -.9 3 0 2 -.4 2 4 4 -.4 -.2 .2	W -.3 -.7 -.6 -.10 0 -.2 -.11 -.5 -.3 -.1 0 0 -.10 -.8 -.1 6 -.7 -.3 -.3 2 11 .2	Y 4 -.8 -.2 -.9 1 .5 -.10 -.5 0 2 -.1 -.1 -.8 -.1 -.1 -.1 -.1 -.1 -.1 -.1 -.1 .2	V 1 -.9 -.7 -.10 1 -.7 -.11 -.4 -.7 -.5 3 -.12 3 2 -.6 -.2 0 -.2 -.3 .5

Obrázek 4.24: Prezentace č. 3, slide č. 31

The diagram illustrates the relationship between BLOSUM and PAM matrices. At the top left, there are two small icons: a DNA sequence logo and a colorful protein structure. To the right, the text "BLOSUM VERSUS PAM" is written in a large, bold, black font, with a horizontal line underneath it.

Below this, four labels are arranged in a 2x2 grid:

- Top-left: "PAM 10"
- Top-right: "PAM 250"
- Bottom-left: "BLOSUM 90"
- Bottom-right: "BLOSUM 62"

Below the bottom row of labels is a large blue arrow pointing from left to right. The arrow consists of three segments: a long horizontal segment on the left, a shorter horizontal segment on the right, and a vertical segment connecting the end of the first horizontal segment to the start of the second.

At the bottom left, the text "VELMI PŘÍBUZNÍ" is written in a large, black, sans-serif font. At the bottom right, the text "VZDÁLENĚ PŘÍBUZNÍ" is written in a similar large, black, sans-serif font.

Obrázek 4.25: Prezentace č. 3, slide č. 44



The slide features a title 'ZAČÍNÁME VYPLŇOVAT' at the top right, with three small decorative icons (DNA helix, red flower, and colorful abstract shapes) to its left. Below the title is a grid representing a sequence alignment. The grid has a header row with letters G, A, A, T, T, C, A, G, T, T, T, A and a header column with scores 0, -3, -6, -9, -12, -15, -18, -21, -24, -27, -30, -33. The main body of the grid contains mostly empty cells, indicating the start of the filling process.

		G	A	A	T	T	C	A	G	T	T	A
	0	-3	-6	-9	-12	-15	-18	-21	-24	-27	-30	-33
G	-3											
G	-6											
A	-9											
T	-12											
C	-15											
G	-18											
A	-21											

## 4.2.2 Algoritmy

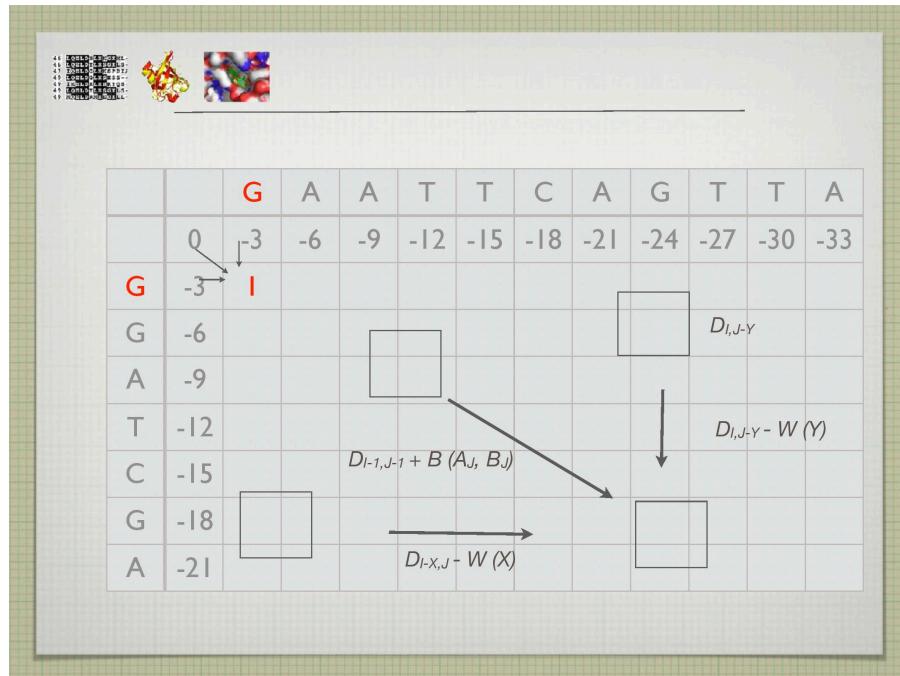
### Needleman-Wunsch

- hledá globální alignment
- pracuje na principu dynamického programování
- jeden z nejstarších (1970)
- zaručuje nalezení optimálního alignmentu (vzhledem k dané GP a skórovací tabulce)

### Průběh NW

1. první a druhou sekvenci napíšeme na první sloupec a řádek tabulky (respektive)
2. pro každou pozici v alignmentu s pomocí scoring matrix počítáme skóre, které bychom dostali:
  - při shodě

Obrázek 4.26: Prezentace č. 3, slide č. 45



Obrázek 4.27: Prezentace č. 3, slide č. 49

The completed dynamic programming matrix for sequence alignment is shown below. The rows represent sequence Y (G, A, T, T, C, A, G, T, T, A) and the columns represent sequence X (G, A, A, T, T, C, A, G, T, T, A). The matrix values are as follows:

	G	A	A	T	T	C	A	G	T	T	A	
0	0	-3	-6	-9	-12	-15	-18	-21	-24	-27	-30	-33
G	-3	I	-2	-5	-8	-11	-14	-17	-20	-23	-26	-29
A	-6	-2	I	-2	-5	-8	-11	-14	-16	-19	-22	-25
A	-9	-5	-1	2	-1	-4	-7	-10	-13	-16	-19	-21
T	-12	-8	-4	-1	3	0	-3	-6	-9	-12	-15	-18
C	-15	-11	-7	-4	0	3	I	-2	-5	-8	-11	-14
G	-18	-14	-10	-7	-3	0	3	I	-1	-4	-7	-10
A	-21	-17	-13	-9	-6	-3	0	4	I	-1	-4	-6

Obrázek 4.28: Prezentace č. 3, slide č. 51




**HLEDÁME ALIGNMENT**

	G	A	A	T	T	C	A	G	T	T	A	
	0	-3	-6	-9	-12	-15	-18	-21	-24	-27	-30	-33
G	-3	1	-2	-5	-8	-11	-14	-17	-20	-23	-26	-29
G	-6	-2	1	-2	-5	-8	-11	-14	-16	-19	-22	-25
A	-9	-5	-1	2	-1	-4	-7	-10	-13	-16	-19	-21
T	-12	-8	-4	-1	3	0	-3	-6	-9	-12	-15	-18
C	-15	-11	-7	-4	0	3	1	-2	-5	-8	-11	-14
G	-18	-14	-10	-7	-3	0	3	1	-1	-4	-7	-10
A	-21	-17	-13	-9	-6	-3	0	4	1	-1	-4	-6
SEKVENCE 1 A												
SEKVENCE 2 A												

Obrázek 4.29: Prezentace č. 3, slide č. 54




**RÉŠENÍ**

	G	A	A	T	T	C	A	G	T	T	A	
	0	-3	-6	-9	-12	-15	-18	-21	-24	-27	-30	-33
G	-3	1	-2	-5	-8	-11	-14	-17	-20	-23	-26	-29
G	-6	-2	1	-2	-5	-8	-11	-14	-16	-19	-22	-25
A	-9	-5	-1	2	-1	-4	-7	-10	-13	-16	-19	-21
T	-12	-8	-4	-1	3	0	-3	-6	-9	-12	-15	-18
C	-15	-11	-7	-4	0	3	1	-2	-5	-8	-11	-14
G	-18	-14	-10	-7	-3	0	3	1	-1	-4	-7	-10
A	-21	-17	-13	-9	-6	-3	0	4	1	-1	-4	-6
SEKVENCE 1 GAATTTCAGTTA												
SEKVENCE 2 GGA-TC-G--A												

- při neshodě
  - při inzerci nebo deleci
3. z těchto možností vždy vybereme tu nejvyšší na napíšeme šipku příslušného směru
  4. postupujeme od konce alignmentu (políčka vpravo dole), a uvažujeme, odkud jsme se na současné políčko dostali

### **Smith-Waterman**

- hledá lokální alignment
- k hledání podobnosti mezi proteiny, motivů a domén je vhodnější než NW
- funguje podobně jako NW, s několika rozdíly
  - všechna dílčí negativní skóre jsou nahrazena 0
  - při backtrackingu nezačínáme vpravo dole, ale na políčku s nejvyšším skórem
  - končíme, jakmile narazíme na 0
- vzniká alignment pouze dobře konzervovaných úseků

Nejznámějším programem na hledání alignmentu je Clustal  $\Omega$  (kdysi Clustal W).

## **4.3 Multiple sequence alignment**

Přednáška č.

5

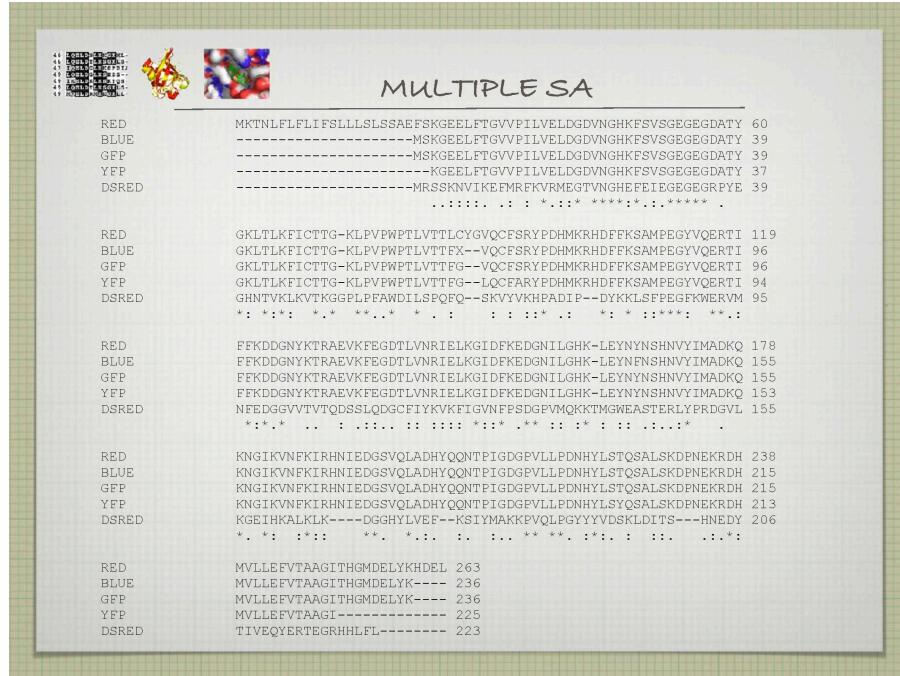
Když srovnáváme více sekvencí najednou, je to sice složitější, ale má to několik velkých výhod:

1. výsledný alignment je přesnější
2. data z alignmentu se dají použít pro fylogenetické studie
3. máme větší šanci nalézt strukturně nebo funkčně významné AK
  - takové AK budou v sekvencích konzervované
4. alignment slouží jako základ pro studie funkce proteinu

### **Jak dělat MSA**

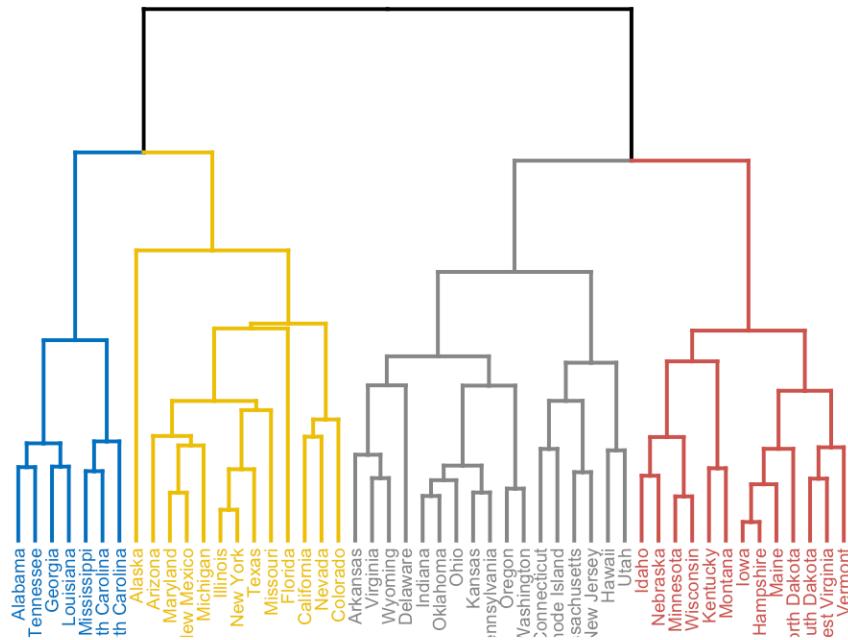
- dynamické programování už není tak dobrá volba

Obrázek 4.30: Prezentace č. 4, slide č. 16



- počet rozměrů matice roste lineárně s počtem sekvencí, čili počet nutných srovnání roste exponenciálně
- používají se hierarchické progresivní metody
  - všechny dvojice sekvencí jsou alignovány v rámci PSA
  - alignmenty jsou hierarchicky seřazeny dle míry podobnosti do fylogenetického stromu (viz níže)
  - finální MSA je budován v krocích — nejprve jsou naalignovány dvojice nejpřibuznějších sekvencí, poté jsou alignovány dvojice nejpřibuznějších sekvencí z těchto alignmentů atd.
  - Clustal Ω, T-Coffee
- nevýhody hiearchických metod
  - chyby vytvořené v úvodních alignmentech se dostanou až do finálního výsledku
    - \* zavedení iterativních metod (optimalizace ohodnocení pomocí objective function): Muscle, ProbCons
    - \* zavedení učících metod: hidden Markov models, genetické algoritmy,

Obrázek 4.31: Fylogenetický strom (dendrogram)



simulated annealing FSA

### Programy pro MSA

- ProbCons
  - využívá informace z MSA při PSA
  - mává nejlepší výsledky
- FSA (fast statistical alignment)
  - používá machine learning metodu simulated annealing na základě PSA
  - GP i skórovací tabulky jsou odhadovány pro každý set sekvencí individuálně
  - funguje i pro velice dlouhé sekvence
- MAFFT
  - metoda pro veliké soubory dat (například fylogenetické analýzy)
  - homologické oblasti identifikovány pomocí rychlých Fourierových transformací (objem a polarita AK)

- výsledný alignment je kombinací progresivních a iterativních metod

**Poznámka bokem — HMM** HMM je probabilistický model, který se využívá k tvoření obdoby skórovacích tabulek—takových, které jsou alignovaným sekvencím šité na míru. Pro každou pozici ukládá HMM, s jakou pravděpodobností se tam vyškytne jaká AK, s jakou pravděpodobností na daném místě dojde k inzerci a s jakou k deleci. Z těchto údajů dokáže HMM předpovědět sekvence, které do daného modelu zapadají, ale také určit, jak dobré do modelu ”sedí” nějaká zadaná sekvence.

Je samozřejmě velice důležité co možná nejlépe určit parametry HMM (ony pravděpodobnosti zmíněné výše). To se většinou dělá trénováním, kdy se HMM zadají nějaké sekvence a on z nich sám vypočítá potřebné pravdepodobnosti, které si poté uloží. HMM poté může rozhodnout, jak velká šance je, že je nějaká zadaná sekvence příbuzná s těmi, na kterých byl vytrénován.

### Kvalita alignmentu

- kvalitu lze hodnotit ze strukturních informací
- výsledný MSA je porovnáván s databází strukturních alignmentů BALiBase, HomFam
- hodnotící programy
  - APDB, který je součástí T-Coffee (což je program na MSA)
  - QuanTest (2017), za pomoci přesnosti predikce sekundárních struktur
- umožňuje vybrat nejlepší z alternativních alignmentů
- kvalita uvnitř alignmentu
  - není uniformní, MSA programy ale často neoznačují, kterým částem věří a kterým ne
  - pro účely fylogenetických analýz se často vyřazují oblasti se spoustou mezer
  - programy TrimAl, JalView, UGENE

## 5 Hledání v databázích

Jak roste množství biologických dat, roste i nutnost umět v nich dobře vyhledávat; zpravidla se snažíme najít sekvenci podobnou nějaké jiné, kterou zrovna máme. Je tedy samozřejmé, že alignment je součástí procesu vyhledávání, a to často i lokální alignment (vyhledávání na základě podobných domén, motivů).

### **true positive (TP)**

To, co jsme hledali a našli.

### **false positive (FP)**

To, co jsme nehledali a přesto našli.

### **true negative (TN)**

To, co jsme nehledali a nenašli.

### **false negative (FN)**

To, co jsme hledali, ale nenašli.

Při vyhledávání je nutno brát ohledy na selektivitu a senzitivitu: obě tyto veličiny ale nelze optimalizovat zároveň.

### **senzitivita**

Pravděpodobnost, s jakou budou nalezeny sekvence příbuzné k vyhledávané sekvenci. Čím nižší je, tím méně skutečných výsledků program najde.

$$\text{senzitivita} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

### **selektivita**

Pravděpodobnost, s jakou jsou nalezené sekvence příbuzné s vyhledávanou sekvencí. Čím nižší je, tím více nevýsledků se objevuje v rámci výsledku ( $\Rightarrow$  je těžší najít ve výsledcích zajímavé údaje).

$$\text{selektivita} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Obdobné veličiny existují i pro analýzu nenalezených sekvencí.

### **specifita**

Udává s jakou pravděpodobností nebudou nalezeny sekvence, které nejsou příbuzné s vyhledávanou sekvencí.

$$\text{specifita} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

### **negative predictive value (NPV)**

Udává s jakou pravděpodobností budou nenalezené sekvence nepříbuzné s vyhledávanou sekvencí.

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

## 5.1 Algoritmy

- tradiční algoritmy jsou příliš pomalé, využívají se heuristiky
  - vedou rychle k výsledku, který je blízko tomu optimálnímu
  - trocha přesnosti obětována pro rychlosť
  - FASTA, BLAST
  - obě použitelné pro proteiny i DNA
- někdy se používá i machine learning
  - HMM (hidden Markov models)
  - profiling methods

### 5.1.1 FASTA

V 80. letech byl vyvinut algoritmus FASTA, který využívá globální alignment. Funguje následovně:

1. známé sekvence v databázi jsou rozděleny na krátké úseky o délce  $k$  a uloženy do vyhledávací tabulky
  - u proteinů  $k \in \{2, 3\}$
  - u DNA  $4 \leq k \leq 6$
2. na stejně dlouhé úseky je nyní rozdělena i hledaná sekvence

3. úseky z hledané sekvence jsou porovnány s úseky uloženými ve vyhledávací tabulce, jsou zaznamenány shodné úseky i jejich offsety
  - například úsek AB je v hledané sekvenci na začátku, ale AB ve vyhledávací tabulce začíná až na pátém místě, offset je tedy 5
4. nejlepší matche dvojic úseků jsou rozšířeny a oskórovány příslušnou skórovací tabulkou (bez mezer)
5. nejlepší takové úseky jsou naalignovány s hledanou sekvencí (tentokrát už s mezerami)
6. výstupem jsou sekvence z databáze, jejichž úseky mají celkově nejvyšší skóre

Z výše zmíněného vyplývá, jaká je největší nevýhoda FASTA algoritmu. Může se stát, že FASTA některé příbuzné sekvence nenajde — konkrétně ty, které s tou hledanou nemají  $k$  identit v řadě. Jsou totiž srovnávány úseky o délce  $k$  a v bodě 3. postupují jen úseky 100% shodné s nějakým úsekem hledané sekvence.

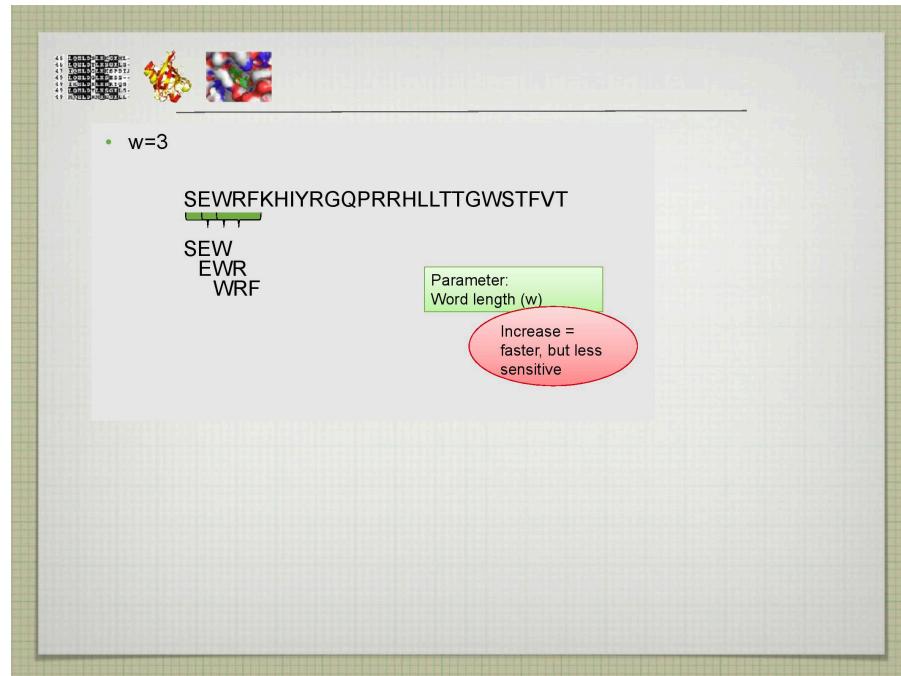
### 5.1.2 BLAST

V 90. letech následoval algoritmus BLAST (Basic Local Alignment Search Tool), který funguje na bázi lokálního alignmentu.

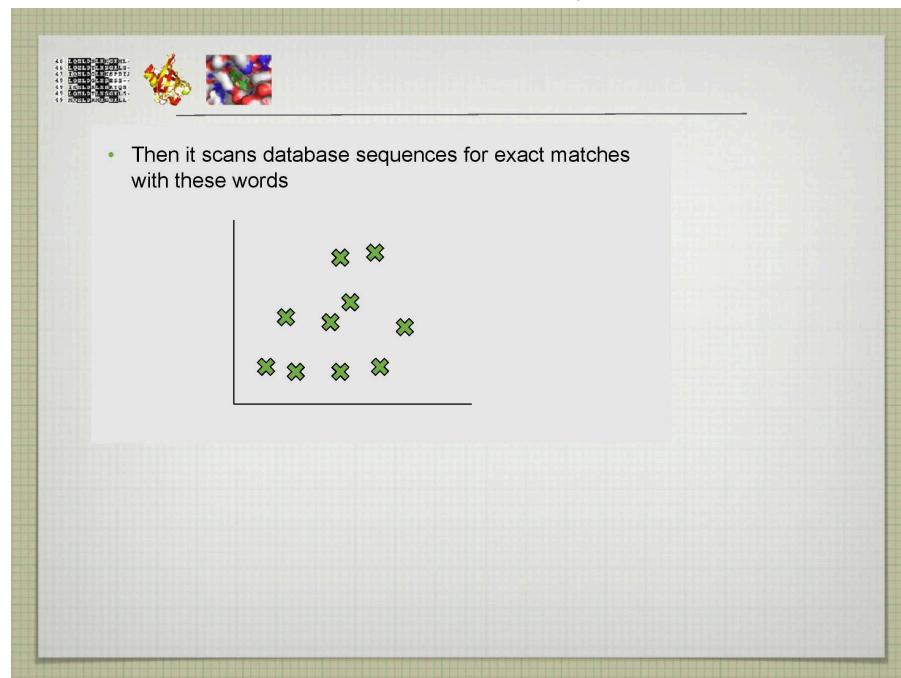
#### Funkce BLASTu

1. známé sekvence v databázi jsou rozděleny na úseky délky  $k$ , tzv. slova (words)
  - u proteinů je často  $k = 3$
2. na stejně dlouhé úseky je nyní rozdělena i hledaná sekvence
3. slova z hledané sekvence jsou porovnávána se slovy získanými ze sekvencí v databázi a podobnosti jsou oskórovány tabulkou (bez mezer); jsou vybrána taková slova z databáze, která dosáhnou předem naefinovaného minimálního skóre (threshold)
  - pro proteiny se většinou používá běžná Blosum 62 tabulka
4. vybraná slova (hity) jsou rozširována dokud skóre jejich alignmentu roste, dál postupují opět jen dvojice slov s určitým skóre, tzv. high scoring pairs (HSPs)
  - rozširování trvá dlouho, proto se k němu většinou přistupuje pouze tehdy, když jsou na najité sekvenci dva hita nedaleko od sebe

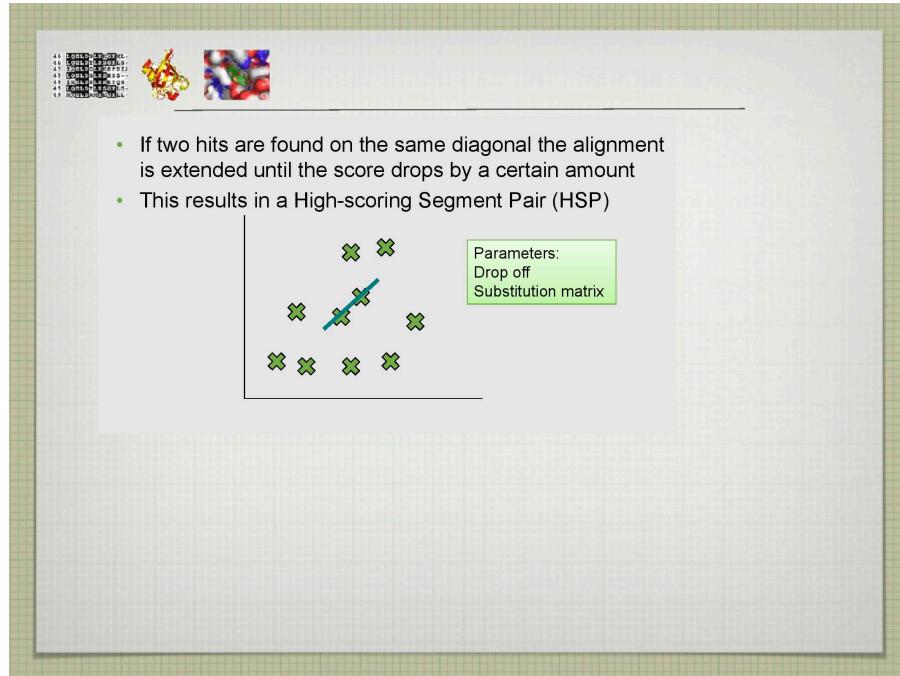
Obrázek 5.1: Prezentace č. 4, slide č. 44



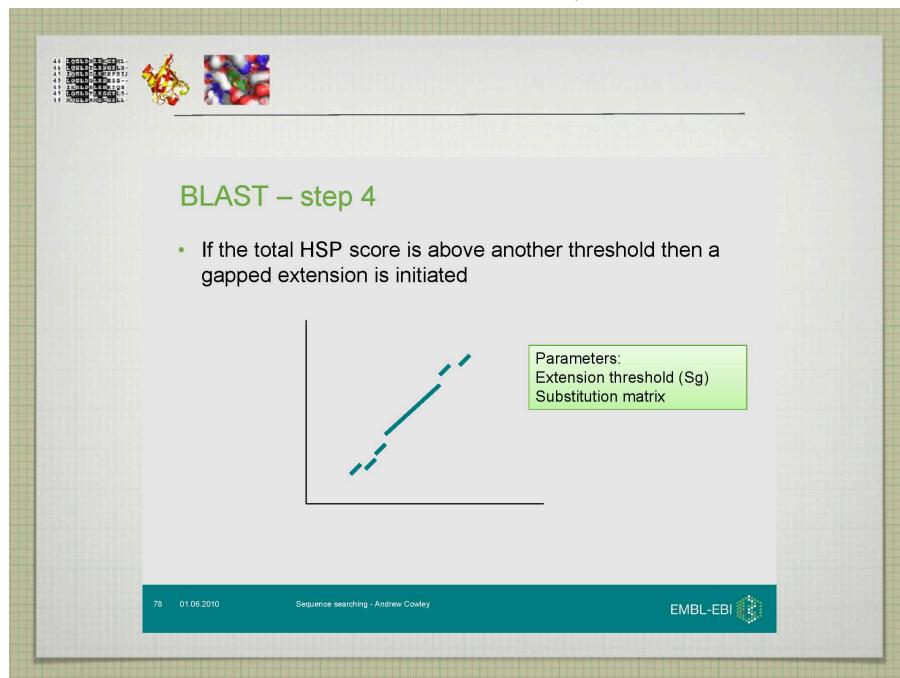
Obrázek 5.2: Prezentace č. 4, slide č. 46



Obrázek 5.3: Prezentace č. 4, slide č. 47



Obrázek 5.4: Prezentace č. 4, slide č. 48



- dvojicí slov je myšlen pár [slovo z hledané sekvence + odpovídající slovo z databáze slov známých sekvencí]
5. výstupem jsou HSPs seřazené podle svého skóre, je u nich dostupná i E-value

Základní rozdíl oproti FASTA tkví v bodě 3. Nejsou vybrána pouze 100% shodná slova, nýbrž všechna slova, která dosáhnou určitého bodového ohodnocení.

Algoritmus BLAST se vyskytuje v několika verzích, mnohé z nich jsou na internetu, například zde.

### Druhy BLASTu

- BLASTn: hledá DNA sekvenci v DNA databázi
- BLASTp: hledá proteinovou sekvenci v proteinové databázi
- BLASTx: hledá DNA sekvenci (6 čtecích rámců) v proteinové databázi
- tBLASTn: hledá proteinovou sekvenci v DNA databázi
- tBLASTx: překládá DNA v překládané DNA databázi
- megablast: zvládne více dotazů (queries) najednou

### 5.1.3 Srovnání FASTA a BLAST

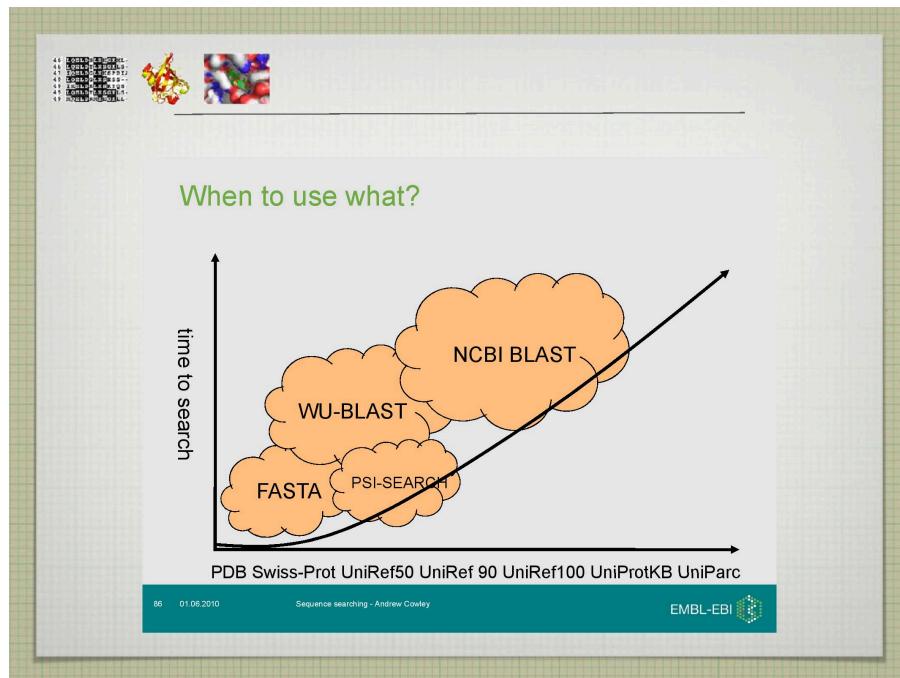
#### Výhody BLASTu

- je rychlejší
- lépe pracuje s proteiny
- má dobré lokální a krátké globální alignmenty
- vytváří HSP (high scoring pairs)
- umí najít blízké sourozence (co se evoluce týče)

#### Výhody FASTA

- lépe pracuje s DNA
- má dobré lokální a krátké globální alignmenty
- vytváří Smith-Waterman alignmenty
- umí najít vzdálenější sourozence ("sestřenice a bratrance")

Obrázek 5.5: Prezentace č. 4, slide č. 60



Existují také SSearch a GSearch, což jsou rigorózní globální/lokální alignmentovací algoritmy. Jejich běh trvá hodiny.

#### 5.1.4 Parametry významnosti alignmentu

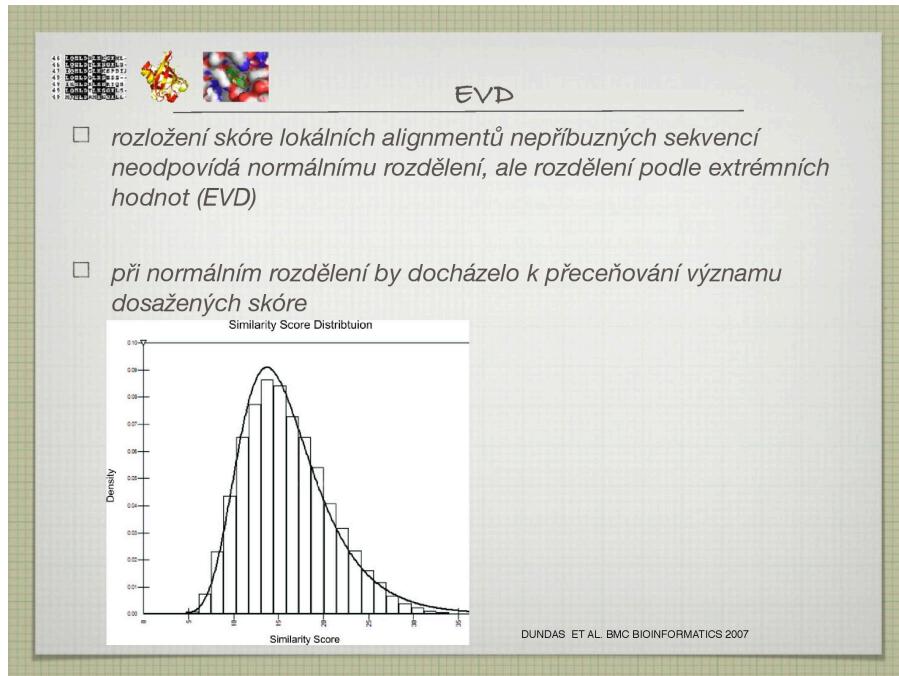
##### Z-score

Říká nám, jak moc je naše skóre odlišné od toho průměrného.  $ZS > 15$  je statisticky významné, pro  $5 \leq ZS \leq 15$  se pravděpodobně jedná o homology a při  $ZS < 5$  sekvence sice mohou, ale nemusí být homologní.

##### Postup výpočtu

1. uděláme alignment dvou sekvencí a zaznamenáme skóre
2. jednu sekvenci náhodně přeházíme
3. znova uděláme alignment a zaznamenáme skóre
4. spočítáme průměr a standardní odchylku skóre

Obrázek 5.6: Prezentace č. 4, slide č. 68



$$Z\text{-score} = \frac{\text{první skóre} - \text{průměr skóre}}{\text{standartní odchylka}}$$

### P-value

Existují dvě různé definice, přičemž druhá z nich lépe odpovídá realitě a poskytuje lepší výsledky.

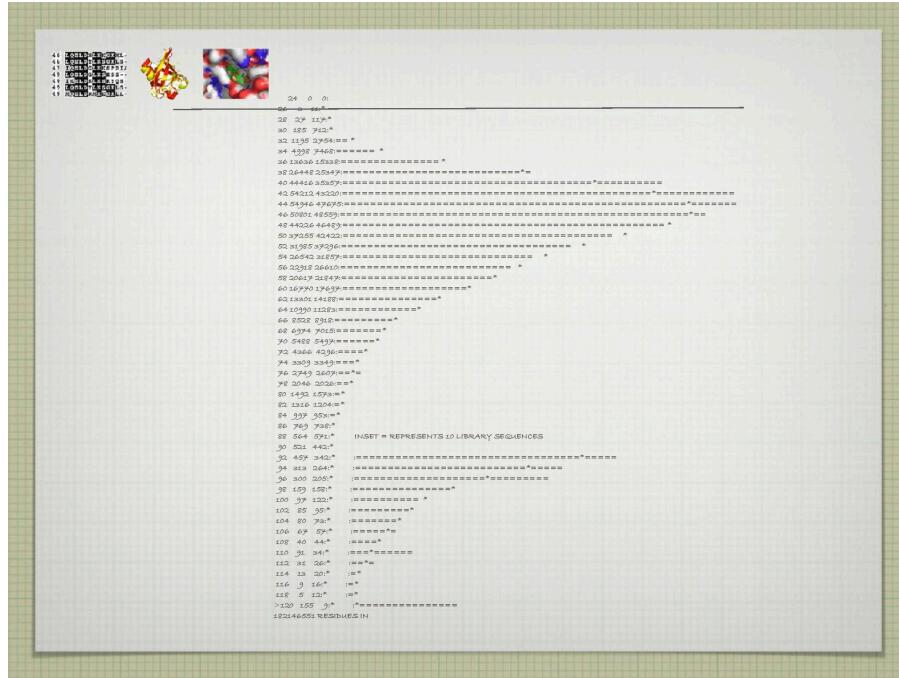
1. pravděpodobnost, že alignment nepříbuzných sekvencí (FP hit) bude mít stejné nebo vyšší skóre
2. pravděpodobnost, že bude stejněho nebo vyššího skóre dosaženo náhodou

Rozdělení skóre není normální (podle Gaussovy křivky), ale odpovídá EVD křivce (extreme value distribution). Při normálním rozdělení by docházelo k přečeňování významu dosažených skóre.

Pro skóre  $S > x$  platí

$$\text{P-value} = 1 - e^{-e^{-\lambda(x-u)}},$$

Obrázek 5.7: Prezentace č. 4, slide č. 69



kde

- $u$  je charakteristická hodnota,  $u = Kmn/\lambda$
- $m, n$  jsou délky sekvencí
- $K$  je konstanta
- $\lambda$  je "decay factor"

$K$  a  $\lambda$  se dají spočítat ze skórovací tabulky.

### E-value

Předpokládaný počet náhodných (FP) sekvencí se stejným nebo vyšším skóre v databázi o dané velikosti. Udává něco jako šum, chceme tuto hodnotu tedy co nejnižší.

$$\text{E-value} = \text{P-value} \cdot \text{velikost databáze}$$

Cutoff skóre v BLASTu udává, kolik lze v databázi o dané velikosti průměrně

Obrázek 5.8: Prezentace č. 5, slide č. 5

**PŘÍKLAD PROFILU**

- v alignmentu globinů se na pozici 3 vyskytuje 3x Ala, 6x Val, 1x Ile, používáme tabulku Blosum 62
- jaké bude profilové skóre pro výskyt Ile a His ?
  - $N(x, A) = 0.3, N(x, V) = 0.6, N(x, I) = 0.1$
  - $S(A, I) = -1, S(V, I) = 3, S(I, I) = 4$
  - $S(A, H) = -2, S(V, H) = -3, S(I, H) = -3$
- $Prof(x, I) = 0.3 \times -1 + 0.6 \times 3 + 0.1 \times 4 = 1.9 \times 10 \text{ (v profilu)} = 19 (-1, 3, 4)$
- $Prof(x, H) = 0.3 \times -2 + 0.6 \times -3 + 0.1 \times -3 = -2.7 \times 10 = -27 (-2, -3, -3)$

čekat FP. Je to vlastně způsob vyvažování selektivity a senzitivity (nižší cutoff zvyšuje selektivitu).

### 5.1.5 Profilové algoritmy

Přednáška č.

6

BLAST přistupuje ke všem sekvencím stejně, existují ale i citlivější metody — profilové.

#### Profile

- skórovací tabulka šitá na míru (pozičně specifická tabulka pro danou proteino-vou rodinu)
- pro každou pozici v alignmentu jsou generována specifická skóre (jak pro zá-měnu AK, tak pro inzerci a deleci)
- zvyšují citlivost dané metody

$$\text{profilové skóre} = 10 \cdot \sum_{p \in \text{pozice}} (\text{četnost AK na pozici } p) \cdot (\text{hodnota z tabulky})$$

Z roku 1997 pochází PSI-BLAST (position specific iterative BLAST). Oproti běžnému BLASTu používá position specific scoring matrix (PSSM), což je tabulka obsahující specifická skóre pro každou pozici v sekvenci.

### **PSI-BLAST**

1. průběh jako BLAST, z nejlepších výsledných alignmentů je vytvořena PSSM
2. je spuštěno další kolo BLASTu, pro počítání skóre je ale použita vypočítaná PSSM
  - po konci druhého kola je vytvořena nová PSSM
3. GOTO 2 (dokud nacházíme nové hity)

Z roku 2009 je CS/CSI BLAST, context-specific iterative BLAST.

### **CS/CSI BLAST**

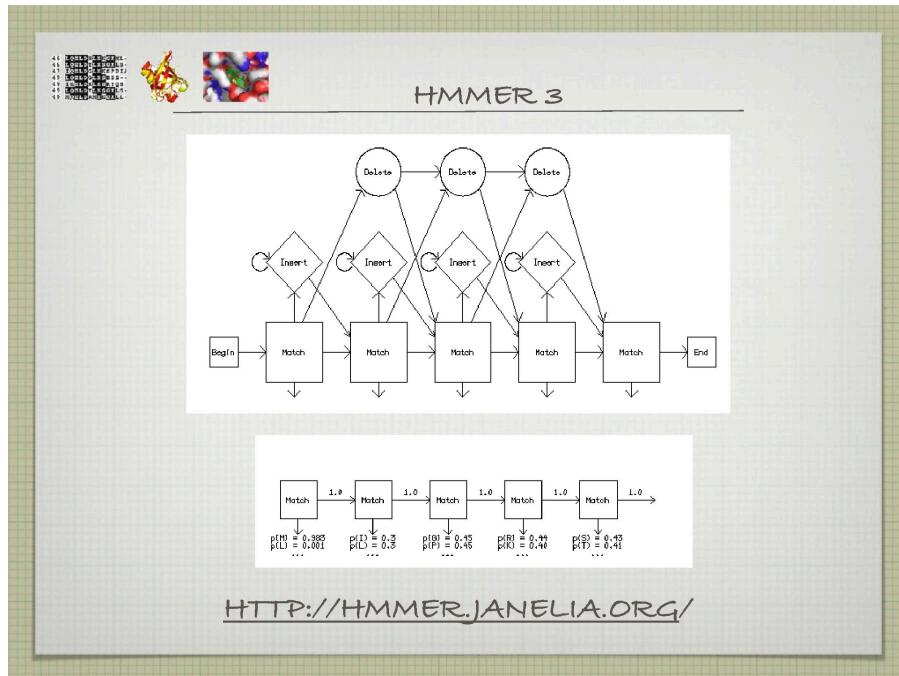
- ”kontext” je tvořen 12 AK v okolí sledované AK
- je schopen najít dvakrát více vzdálených homologů než běžný BLAST při zachování rychlosti a chybovosti
- po dvou iteracích CSI BLAST dostaneme stejné výsledky jako po pěti iteracích PSI-BLAST

Poslední profilovou metodou jsou HMM, hidden Markov models.

### **HMM**

- velice citlivá metoda, vytváří statistický model pro definovanou skupinu sekvencí
  - z modelu počítá pravděpodobnosti výskytu dané AK, ostatních AK, inserce, delece, ale i výskytu mezery a přechodu mezi jednotlivými stavami

Obrázek 5.9: Prezentace č. 5, slide č. 9



- používána při rozhodování, zda protein spadá do určité skupiny proteinů, typicky pro sekvence s nízkou %SI
- na základě “tréninku” na sekvencích patřících do jedné skupiny (globiny) generuje pravděpodobnost nejen pro jednotlivé záměny a inzerce a delece, ale i pro přechody mezi nimi
- doveď do modelu zahrnout i aminokyseliny, které se v tréninkové skupině nevyskytují

## 6 Analýza sekvencí

Co dělat, když vyhledávání v databázích nepřineslo nic zajímavého? Jak přesto nějak využít sekvenci, kterou máme?

**Co dělat se sekvencí?**

- „pattern search“ — hledání domén, motivů
- hledání posttranslačních modifikací — glykosylace, fosforylace, methylace
- hledání buněčné lokalizace
- určení, zda se nejedná o membránové proteiny
- HCA - Hydrophobic Cluster Analysis
- hledání procentuální zastoupení AK
  - např. v buňce existují dvě kategorie molekul, které reagují s vápníkem, jedny se zásobní a druhé se signalizační funkcí
  - obě kategorie můžeme rozlišit podle toho, jaké procento obsahují určitých AK
- promotorové oblasti — hledání DNA vazebních míst
- predikce struktury

Aneb, ano, máme co dělat, i když nám MSA nic moc neprozradil.

## 6.1 Hledání motivů

**PROSITE**

- databáze motivů spojená s databází Swissprot
- motivy jsou kontrolovány ručně
- umožňuje hledání motivů v sekvenci, hledání sekvencí se specifickým motivem, vytvoření vlastního motivu

**profile**

Kvantitativní popis toho, jak vypadá sekvence — udání vyskytujících se AK a frekvence výskytu. Většinou se jedná o domény.

**pattern**

Funguje podobně jako regex — udává, které AK se (ne)můžou vyskytovat na daném místě, případně kolikrát.

Obrázek 6.1: Prezentace č. 5, slide č. 17

The screenshot shows a presentation slide with the title 'PROFILE' at the top. Below the title are three small images: a grid of colored squares, a stylized red and yellow protein structure, and a cluster of colorful spheres representing amino acids.

*udává nejenom výskytující se aminokyseliny, ale i frekvence jejich výskytu -> citlivější než pattern*

*popisuje obvykle větší kusy proteinu (doménu)*

	A	K	L	S	H	C	L	L	V	
F	-18	-10	-1	-8	8	-3	3	-10	-2	-8
F	-22	-33	-18	-18	-22	-22	-24	-19	-7	
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-23	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-22	-23	-10
I	3	-27	21	28	-22	-23	-8	33	19	-23
K	-26	25	-25	-27	6	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	10	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	-10	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	3	1	-24	-19	-4
T	-10	-10	-6	-7	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	-1	-23	-12	-19	0	0	-18

[STAIV] -{ERDL} -[LIVMF] -[LIVM] -D-  
-[DSTA] -G-[LIVMFC] -X(2,3)-[DNH]

AK jsou označeny jednopísmenným kódem, mezi jednotlivými pozicemi jsou pomlčky. V hranatých závorkách jsou AK vyskytující se na určité pozici, naopak ve složených jsou AK, které se na ni nevyskytují. Číslo v závorce udává počet pozic.

Pattern jako jediný dovede jednoznačně přidělit či vyloučit motiv.

### Další databáze motivů

- PFAM: používá HMM, dobře anotovaná, informace o tom, jak dobře proteiny interagují, jestli mají známou strukturu atd.; jde o databázi proteinových rodin a domén
- BLOCKS: funguje na podobném principu jako BLAST: automaticky generovaná databáze alignmentů konzervovaných úseků

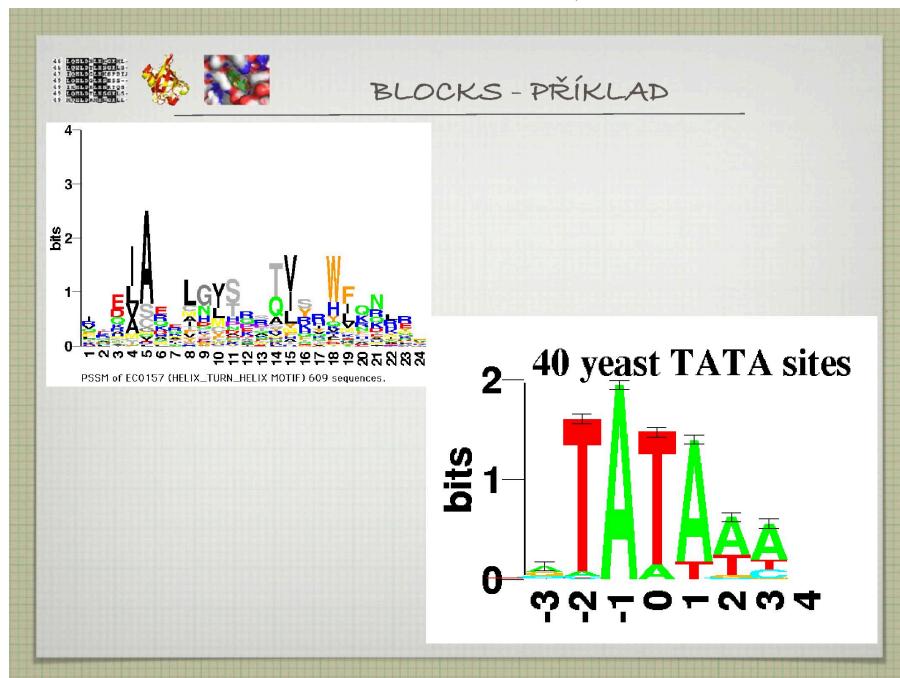
Obrázek 6.2: Prezentace č. 5, slide č. 23




### SPECIALISATION OF DATABASES

<b>PRINTS</b>	Describe sibling families
<b>PROSITE</b>	Identify binding and active sites
<b>PRODOM</b>	Describe conserved core of domains
<b>PFAM</b>	Wide coverage of domains & families
<b>SMART</b>	Signalling, extracellular & nuclear domains
<b>TIGRFAM</b>	Functional classification of families
<b>PIRSF</b>	Families conserved in domain composition
<b>PANTHER</b>	Functional classification of families
<b>HAMAP</b>	Functional classification of bacterial families
<b>Superfam</b>	Structural-based domain classification
<b>GENE3D</b>	Structural-based domain classification

Obrázek 6.3: Prezentace č. 5, slide č. 24



- PRINTS: kde je více motivů kombinovaných do fingerprints, které popisují protein
- PRODOM: oblíbeno strukturními biology
- Gene3D: založena na 3D strukturních alignmentech
- INTERPRO: shromažďuje informace z více databází, jde o metaserver

## 6.2 Další možnosti analýzy sekvencí

### Druhy posttranslačních modifikací

- fosforylace, methylace: součástí signálních kaskád, regulace exprese
- N-, O- glykosylace
- myristoylace, palmitoylace: uchycení proteinů do membrán
- biotynylace: na lysinech

### Hledání posttranslačních modifikací

- predikce probíhá na AutoMotif Serveru, nebo na některé z neuronových sítí: NetPhos, NetOGlyc, NetNGlyc
- často ale nalézáme false positives

### Zjišťování buněčné lokalizace

- nástroj určování funkce proteinu a může usnadnit vývoj nových léků
  - sekretované a membránové proteiny jsou dobré dostupné pro léky
  - bakteriální povrchové proteiny jsou dobrými kandidáty pro vakcíny
- proteiny obsahují signály (sekvenční i strukturní), které je návádějí z cytoplazmy do místa jejich určení
  - jaderné proteiny, membránové proteiny, sekretované proteiny, chloroplas-tové proteiny, proteiny v ER
- tradičním nástrojem je PSORT
  - dnes updatovaný na WoLF PSORT (80% přesnost, 14 000 sekvencí)
- BaCelLo, LOCtree, SherLoc2, SEcretomeP, PredictNLS

Obrázek 6.4: Prezentace č. 5, slide č. 30

The screenshot shows the PSORT software interface with the title "PSORT - VÝSTUP". It displays several analysis methods and their results:

- checkings**: 71 PROSITE ribosomal protein motifs: none
- checkings**: 33 PROSITE prokaryotic DNA binding motifs: none
- NNCN**: Reinhardt's method for Cytoplasmic/Nuclear discrimination  
Prediction: cytoplasmic  
Reliability: 89
- COIL**: Lupas's algorithm to detect coiled-coil regions  
total: 0 residues

### Results of the $k$ -NN Prediction

$k = 9/23$

55.6 %: extracellular, including cell wall  
11.1 %: Golgi  
11.1 %: mitochondrial  
11.1 %: vacuolar  
11.1 %: endoplasmic reticulum

>> prediction for QUERY is exc (k=9)

Obrázek 6.5: Prezentace č. 5, slide č. 31

The slide is titled "MEMBRÁNOVÉ PROTEINY". It contains the following text:

- důležité, velmi početné, málo struktur
- integrální proteiny musí alespoň jednou projít hydrofóbní membránou -> ~20 hydrofóbnych aminokyselin
- TMHMM, Phobius.....

On the right side of the slide is a portrait of a man.

Below the text is a diagram of a membrane protein embedded in a lipid bilayer. The protein structure consists of several alpha-helices (yellow) spanning the membrane. The regions outside the membrane are labeled "Outside", and the regions inside the membrane are labeled "Inside". The lipid tails (red) are shown extending from the membrane. The diagram is labeled with "Outside loop", "Tail", "Helix", "Tail", "Inside loop", "Tail", "Helix", "Tail", "Helix", "Tail", "Helix", "Tail", "Helix", "Tail", "Helix", "Tail", "Helix", "Tail", "Tail", "Outside", "Membrane", "Inside". Below the diagram, there is an "Amino acid seq." and a "State seq." followed by a "Topology" diagram.

### **Určování membránových proteinů**

- určování pomocí TMHMM (transmembrane helices HMM) nebo Phobius
  - membránové proteiny mají totiž často transmembránové helixy
- ne příliš početné, ale velice důležité
- integrální proteiny musí alespoň jednou projít hydrofóbní membránou
  - musí mít alespoň dvacet hydrofóbních AK

### **Určování promotorových oblastí**

- zajímá nás, zda se před určitým genem vyskytuje motiv, který by byl dobré rozpoznatelný nějakým transkripčním faktorem (TF)
- existuje databáze TRANSFAC obsahující DNA sekvence, které na sebe vážou TF
  - databázi můžeme využít při prohledávání promotorových oblastí
- existují i novější databáze, ale ty jsou placené a velice drahé

### **Další nástroje k analýze sekvencí**

- porovnání složení z hlediska AK
- vazba iontů
- řada specializovaných serverů pro specifické skupiny proteinů (například proti látky)

## **7 Databáze**

Databáze jsou strukturované soubory dat v počítači, které je možné prohledávat, měnit a ukládat. Zazkádají se z důvodů organizace a zálohování dat, a proto, aby měl k datům kdokoli relativně jednoduchý přístup.

**Vlastnosti databáze**

- četnost aktualizace dat
- četnost aktualizace software
- redundance dat
- anotace dat (přidělení biologického významu sekvencím)
- anotace databáze (kdo databázi vytvořil a co bylo jeho cílem, jak se s daty nakládá, jestli existuje kontrola dat)

V databázích nejsou uložena jen data o proteinech:

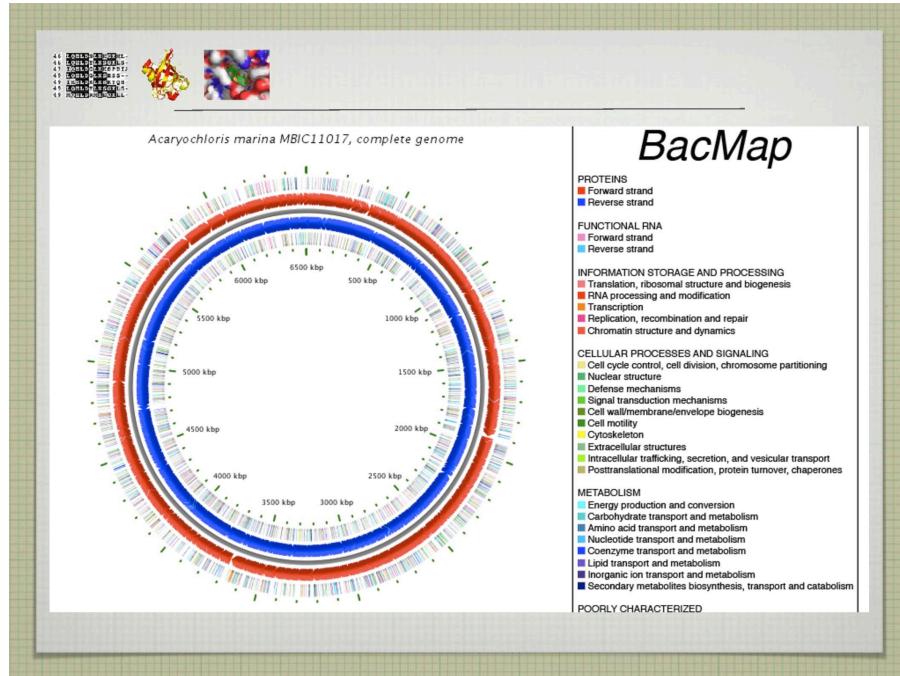
- databáze DNA
  - GenBank, EMBL, DDJB
  - data si denně vyměňují, takže mají stejný obsah
- databáze proteinů
  - UniProt (tj. Swissprot + TrEMBL + PIR): lepší než americké databáze
  - SwissProt, GenPept, PRF
- genomové databáze (obsahují nukleotidové sekvence a mapování, z genové mapy umíme předpovědět funkci)
  - Ensembl, Sanger
- strukturní databáze (obsahují 3D struktury molekul)
  - primární: RCSB (USA), PDBe (EU), PDBJ (Japonsko)
    - \* pravidelně si vyměňují data
    - \* všechny tři výše zmíněné jsou součástí obecné PDB (protein data bank), která je spravována mezinárodní organizací Worldwide Protein Data Bank
  - ”added-value” databáze: OCA, PDBSum
  - odvozené databáze, které hodnotí kvalitu dat: EDS, WhatCheck

## 7.1 Strukturní databáze

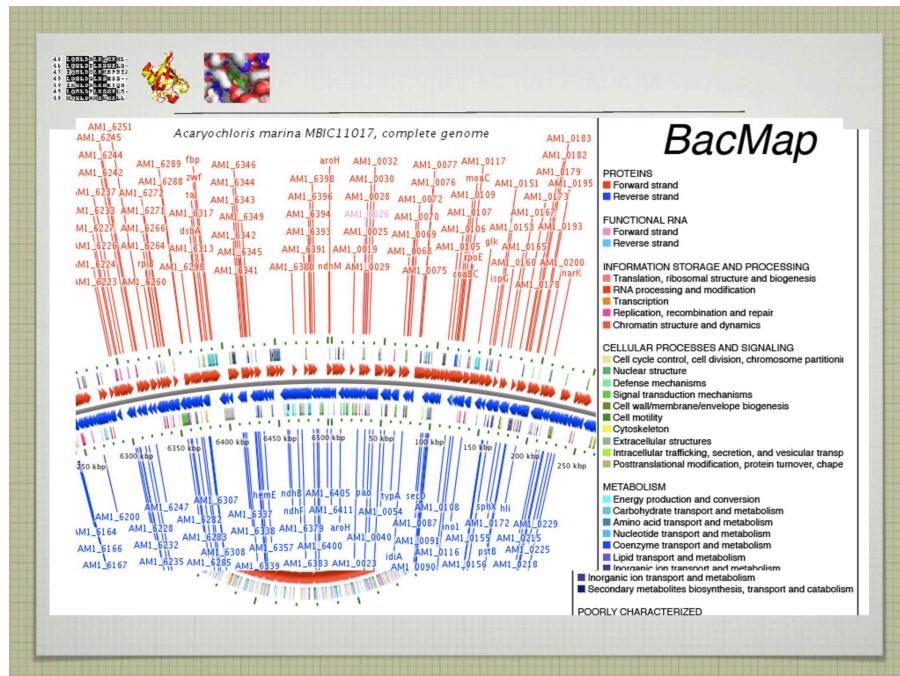
### Způsoby získávání dat

- rentgenová krystalografie
  - libovolná velikost proteinu nebo komplexu
  - potřebujeme ale krystal, který je velmi složité vyrobit

Obrázek 7.1: Prezentace č. 5, slide č. 52



Obrázek 7.2: Prezentace č. 5, slide č. 53



- vhodná pro statické struktury
- má velké rozlišení
- NMR (nuclear magnetic resonance)
  - limitovaná velikostí proteinu (kolem 50 kDa)
  - potřebujeme čistý vzorek v roztoku
  - nezískáme tolik detailů jako u krystalografie
  - vidíme i vzdálenosti vodíkových atomů a torzní úhly, výsledkem je několik struktur => vidíme i dynamiku
- elektronová mikroskopie
  - má limitované rozlišení
  - vhodná pro velké komplexy
  - většinou používána v kombinaci s krystalografií pro dosažení velkého rozlišení

**TODO** Doplnit odkazy na zápisky ze strukturní biologie (až zde budou).

### 7.1.1 PDB

PDB (protein data bank) je strukturní databáze.

#### Historie PDB

- založena 1971 Walterem Hamiltonem v Brookhaven National Laboratory
- na začátku 7 struktur, nyní přes 150 000 struktur
- v dnešní době ji řídí konsorcium tří institucí

Soubory jsou v PDB uloženy ve formátu .pdb. Ten má dvě části; první popisuje, o jakou strukturu se jedná, druhá popisuje už samotnou strukturu (rozložení atomů a vazeb v prostoru). Ve sloupcích jsou pak zapsaná různá další data, viz slide. Temperature factor určuje plochu, kde se popisovaný atom vyskytuje; buďto důsledkem nepřesnosti našich měření, či jeho dynamikou.

V databázích nejsou uloženy jen struktury samotné, ale i daší doplňující informace, například:

Obrázek 7.3: Prezentace č. 5, slide č. 63



Chain								Occupancy			Atom type	
Atom identifier												
1	2	3	4	5	6	7	8					
ATOM	340	N	PHE	A	43	3.853	28.346	32.161	1.00	10.57		N
ATOM	341	CA	PHE	A	43	3.839	29.688	32.724	1.00	12.33		C
ATOM	342	C	PHE	A	43	3.096	29.747	34.047	1.00	13.20		C
ATOM	343	O	PHE	A	43	2.361	28.823	34.393	1.00	12.52		O
ATOM	344	CB	PHE	A	43	3.228	30.659	31.700	1.00	10.99		C
ATOM	345	CG	PHE	A	43	3.993	30.709	30.401	1.00	9.80		C
ATOM	346	CD1	PHE	A	43	3.743	29.794	29.386	1.00	9.85		C
ATOM	347	CD2	PHE	A	43	5.032	31.615	30.233	1.00	11.37		C
ATOM	348	CE1	PHE	A	43	4.528	29.781	28.220	1.00	10.71		C
ATOM	349	CE2	PHE	A	43	5.816	31.612	29.075	1.00	10.61		C
ATOM	350	CZ	PHE	A	43	5.569	30.697	28.067	1.00	10.48		C

Atom identifier

Atom number

Residue number

X.Y.Z coordinates

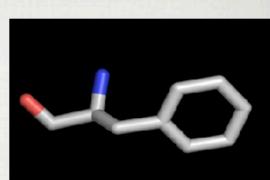
Residue type

Temperature factor

Occupancy

Atom type

Obrázek 7.4: Prezentace č. 5, slide č. 65



1	2	3	4	5	6	7	8					
ATOM	340	N	PHE	A	43	3.853	28.346	32.161	1.00	10.57		N
ATOM	341	CA	PHE	A	43	3.839	29.688	32.724	1.00	12.33		C
ATOM	342	C	PHE	A	43	3.096	29.747	34.047	1.00	13.20		C
ATOM	343	O	PHE	A	43	2.361	28.823	34.393	1.00	12.52		O
ATOM	344	CB	PHE	A	43	3.228	30.659	31.700	1.00	10.99		C
ATOM	345	CG	PHE	A	43	3.993	30.709	30.401	1.00	9.80		C
ATOM	346	CD1	PHE	A	43	3.743	29.794	29.386	1.00	9.85		C
ATOM	347	CD2	PHE	A	43	5.032	31.615	30.233	1.00	11.37		C
ATOM	348	CE1	PHE	A	43	4.528	29.781	28.220	1.00	10.71		C
ATOM	349	CE2	PHE	A	43	5.816	31.612	29.075	1.00	10.61		C
ATOM	350	CZ	PHE	A	43	5.569	30.697	28.067	1.00	10.48		C

Atom identifier

Atom number

Residue number

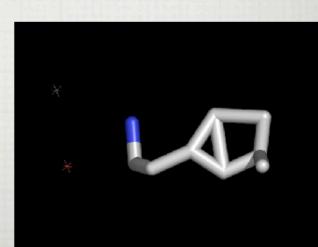
X.Y.Z coordinates

Residue type

Temperature factor

Occupancy

Atom type



- popis kvality struktury
- seznam strukturních motivů
- seznam 3D modelů celých proteinů i s ligandy

### Problémy databáze PDB

- nemůže odmítnout žádná data
  - tím pádem může obsahovat — a také vskutku obsahuje — mnoho chyb
  - kontrola přes WhatCheck, ProCheck, kontrolou Ramachandranova diagramu, nebo použití EDS (electron density server)
- struktury jsou pouze modely, které ne nutně vyhovují experimentálním datům (existují více interpretací těchto dat)
- změnit data může jen jejich autor (po smrti autora už nikdo)

### Problémy .pdb formátu

- počet atomů ve struktuře je omezený na 99 999 ÅK
- málo strukturovaný typ souborů, což je nevýhoda při automatické extrakci dat
- nekonzistentní pojmenování polí v rádcích

Existují ale nové strukturní formáty, jako mmCIF nebo XML, které jsou pro počítače dobře čitelné.

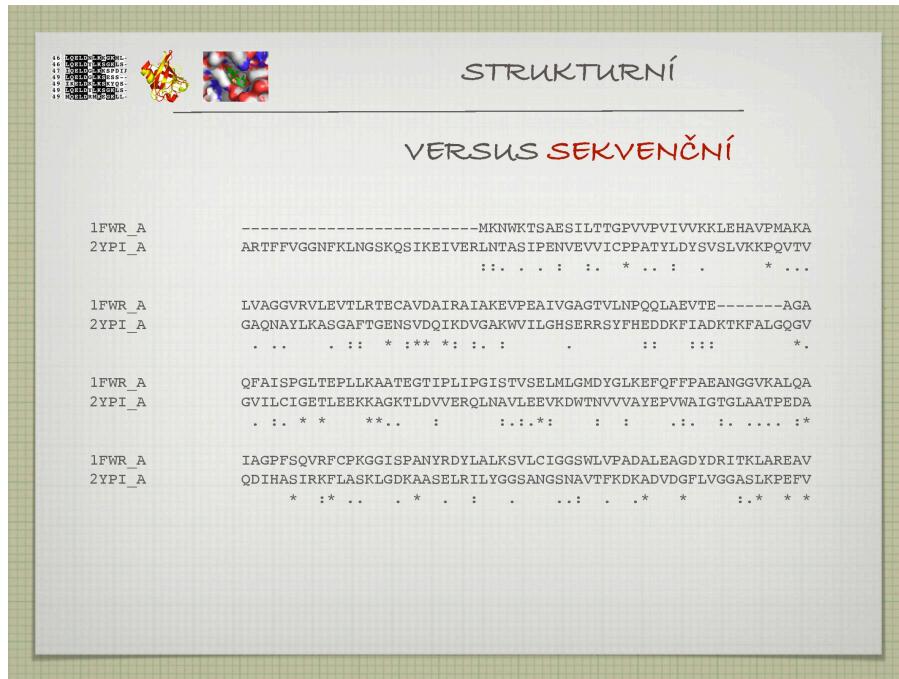
## 8 Strukturní alignment

Přednáška č.

7

Struktura proteinu je lépe konzervovaná, než sekvence — struktura totiž určuje jeho funkci, jejíž změna je jen zřídkakdy výhodná, naproti tomu i různé sekvence mohou mít podobnou strukturu, a tedy funkci.

Obrázek 8.1: Prezentace č. 6, slide č. 41



### Proč se zajímat o strukturu?

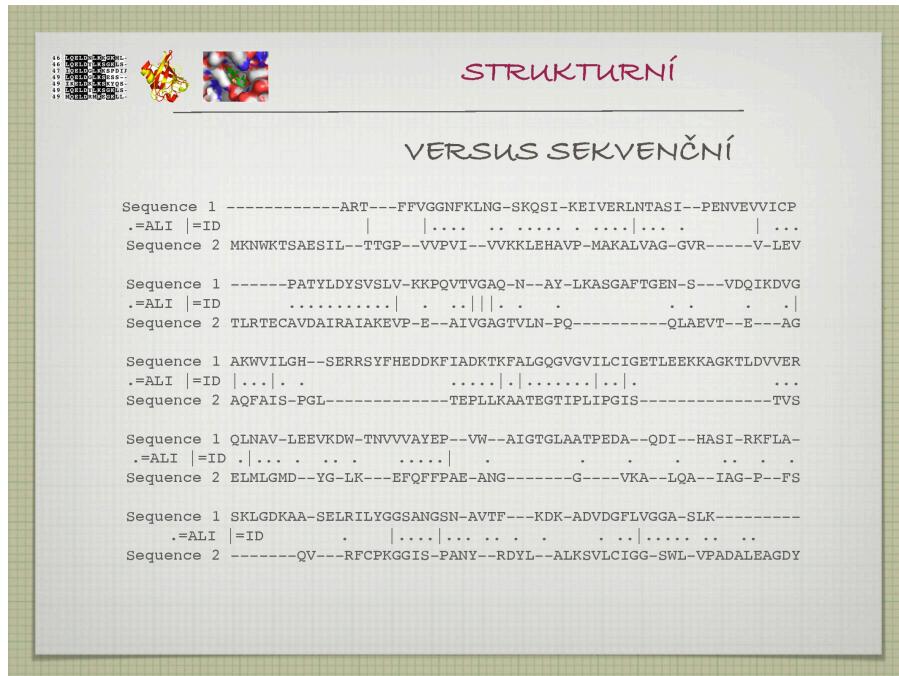
- můžeme pozorovat změny konformace při vazbě s ligandem
- můžeme odhalit evoluční vztahy mezi proteiny
  - můžeme dokonce odhalovat homologie v twilight (a midnight) zone
- proteiny lze na základě struktury dále třídit, můžeme v nich vyhledávat motivy atd.
- můžeme pomocí ní vylepšit MSA
  - pozice mezer závisí na sekundární struktuře (viz oddíl o PSA)

Najít strukturní alignment je složité (NP-složité), navíc ani optimální alignment (podle nějaké naší metriky) nemusí odpovídat reálným biologickým poznatkům.

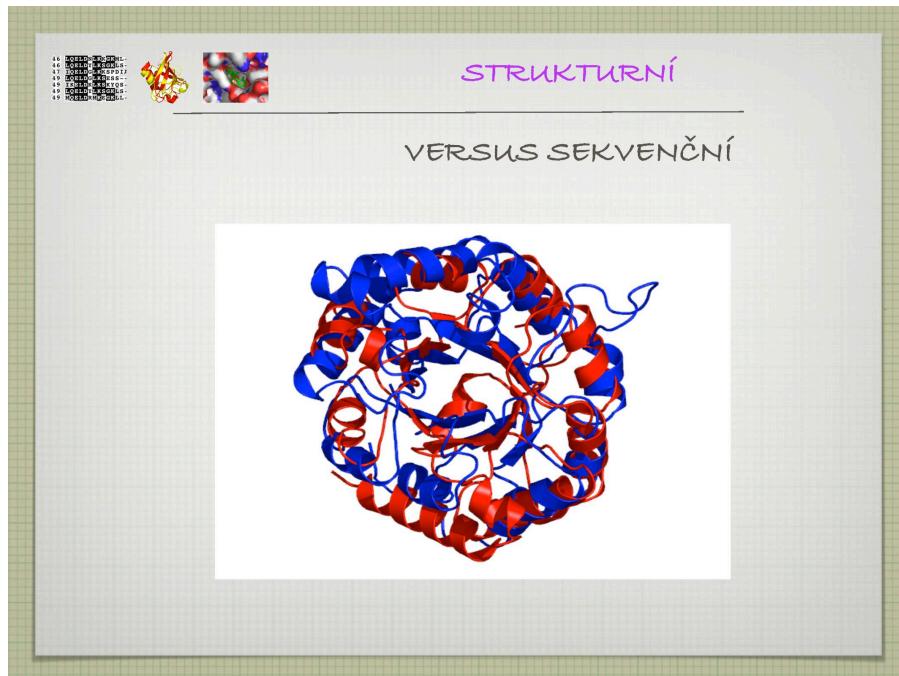
### Postup strukturního alignmentu

1. najdeme nějaký alignment pomocí heuristických metod
2. optimalizujeme jej dle předem stanovených kritérií

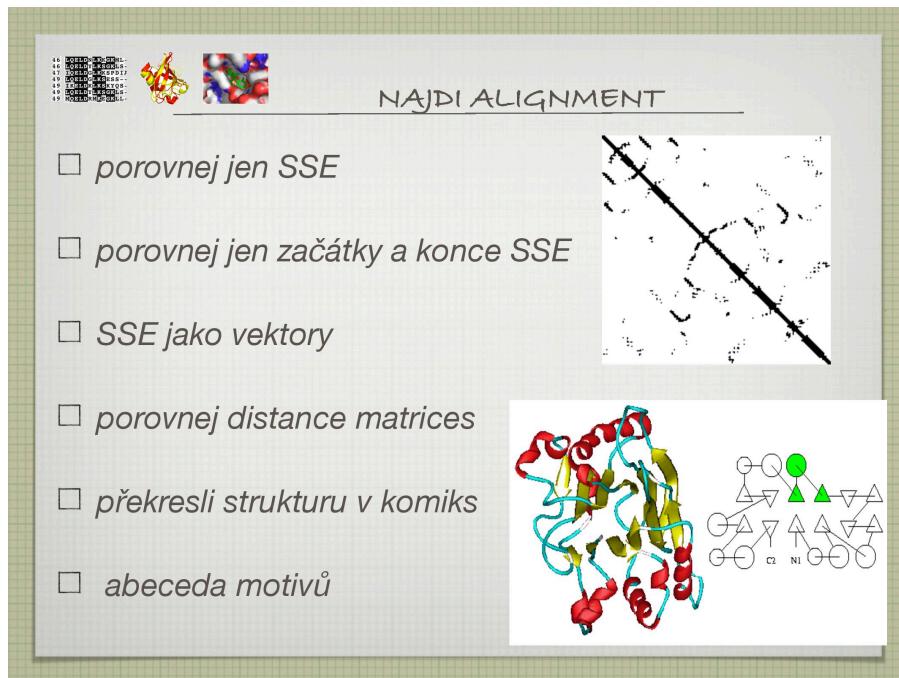
Obrázek 8.2: Prezentace č. 6, slide č. 43



Obrázek 8.3: Prezentace č. 6, slide č. 45



Obrázek 8.4: Prezentace č. 6, slide č. 50



3. zhodnotíme jeho statistickou významnost

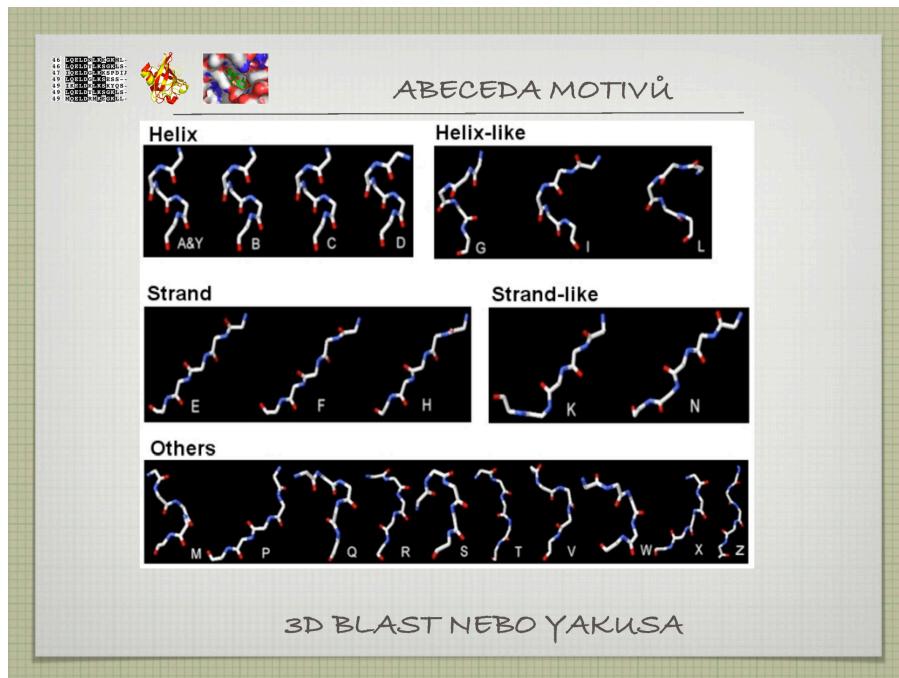
ad 1) Toto lze dělat několika způsoby:

- srovnáním pravidelných úseků sekundární struktury (SS), případně jen jejich začátků a konců
- srovnáním tabulek vzájemných vzdáleností (distance matrices) návzájem si odpovídajících atomů
- rozebráním struktury na jednotlivé motivy, z nichž každému je přiřazeno jedno písmeno, a přepsáním proteinů do této nové abecedy; textové sekvence motivů jsou poté srovnány běžným PSA (BLAST, Yakusa)

ad 2) Optimalizovány většinou bývají superpozice atomů. Superpozice je vzdálenost dvou  $C\alpha$ , která je pak přes všechny  $C\alpha$  měřena jako RMSD (root mean square distance). Hledají se pak takové konformace/rotace, bylo RMSD minimalizováno.

$$\text{RMSD} = \sqrt{\frac{\sum_i d_i^2}{N}},$$

Obrázek 8.5: Prezentace č. 6, slide č. 51



kde  $d$  je (Euklidovská) vzdálenost dvou atomů  $C\alpha$  a  $N$  je počet atomů  $C\alpha$ .

ad 3) RMSD je k hodnocení statistické významosti nevhodné, protože je to globální parametr citlivý na lokální změny a protože koreluje s délkou alignmentu. Existuje ale několik alternativ:

- SAS: řeší problém korelace hodnoty RMSD s délkou sekvence

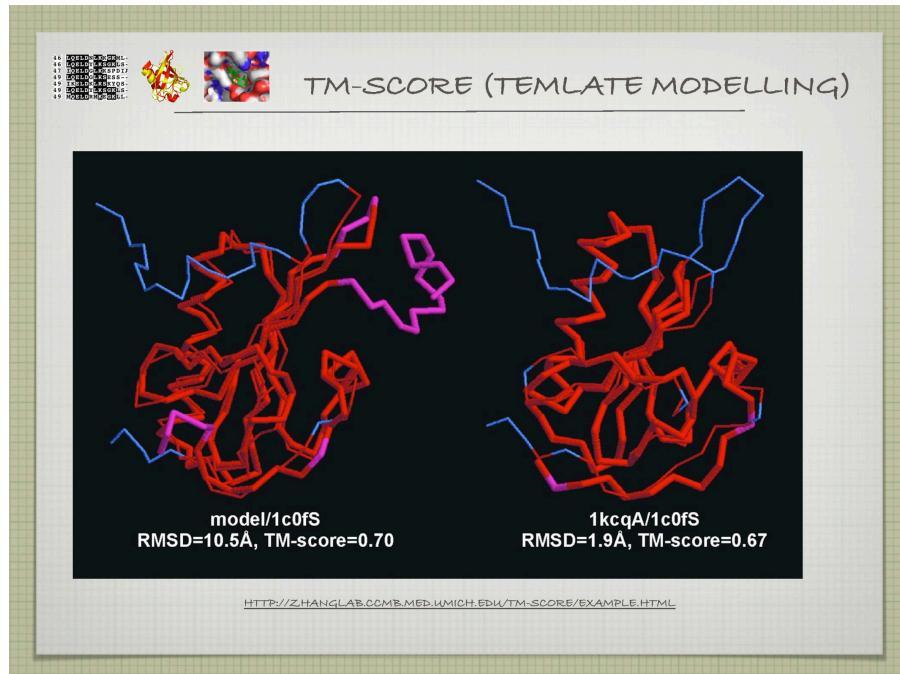
$$\text{SAS} = \frac{\text{RMSD}}{N}$$

- Z-score a E-value, viz parametry významnosti alignmentu
- TM-score: není závislé na délce, je citlivější (od 0=úplně jiné do 1=shodné)

### Metody (stránky, algoritmy)

- vyplatí se použít více metod
- CE, DALI, MATRAS, atd. (viz slide)

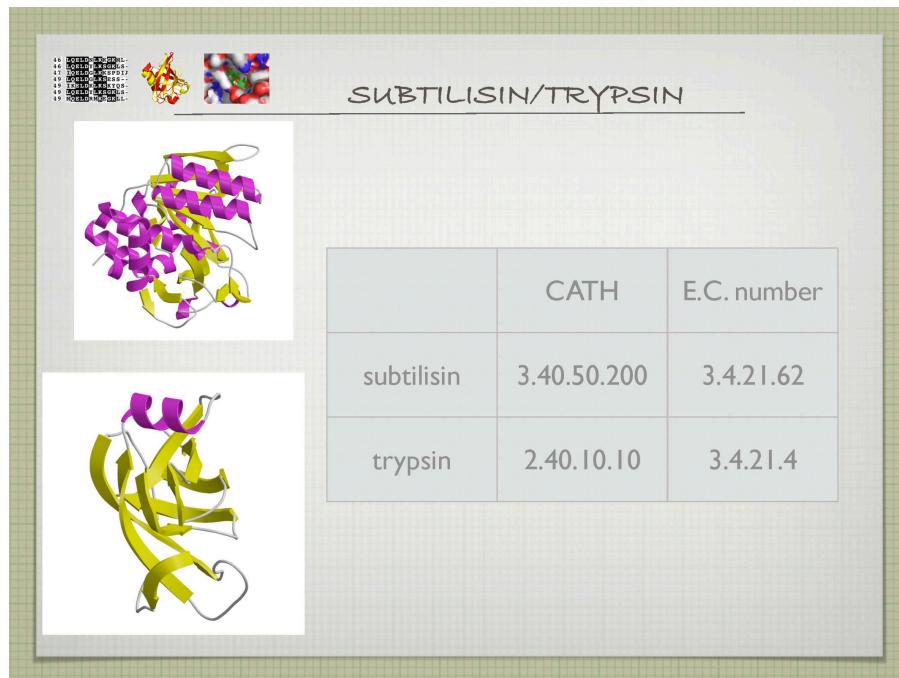
Obrázek 8.6: Prezentace č. 6, slide č. 57



Obrázek 8.7: Prezentace č. 6, slide č. 58

Server	Location	Method
CE	http://cl.sdsc.edu	Extension of optimal path <sup>1</sup>
DALI	http://www2.ebi.ac.uk/dali	Distance-matrix alignment <sup>2</sup>
DEJAVU	http://portray.bmc.uu.se/cgi-bin/dennis/dejavu.pl	SSE alignment with C $\alpha$ -atom optimisation <sup>3</sup>
LOCK	http://gene.stanford.edu/LOCK/	Absolute orientation of corresponding points <sup>4</sup>
MATRAS	http://bongo.lab.nig.ac.jp/~takawaba/Matras.html	Markov transition model of evolutions <sup>5</sup>
PRIDE	http://hydra.icgeb.trieste.it/pride/	C $\alpha$ - C $\alpha$ atom distances <sup>6</sup>
SSM	http://www.ebi.ac.uk/msd-srv/ssm/ssmstart.html	Graph matching algorithm
TOP	http://bioinfo1.mbfys.lu.se/TOP	SSE alignment <sup>7</sup>
TOPS	http://tops.ebi.ac.uk/tops/compare1.html	TOPS-diagram alignment <sup>8</sup>
TOPSCAN	http://www.rubic.rdg.ac.uk/~andrew/bioinf.org/topscan	Secondary topology-string alignment <sup>9</sup>
VAST	http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html	Vector alignment <sup>10</sup>

Obrázek 8.8: Prezentace č. 6, slide č. 66



## 8.1 Klasifikace proteinů

Strukturní alignment lze využít k tvorbě systému struktur (většinou podle domén). Takovýchto „opakoványch“ struktur je konečné množství, proto se vyplatí je klasifikovat.

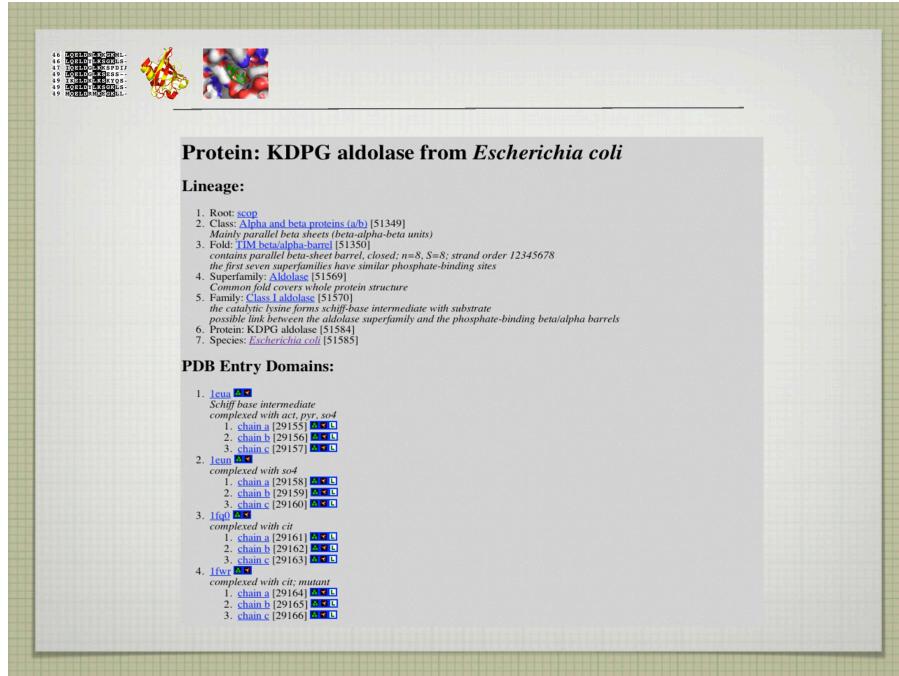
### doména

Někdy je uvažována jako jednotka evoluce. Je globulární, při foldingu nezávislá na zbytku proteinu, má více kontaktů uvnitř sebe než se zbytkem proteinu. Může se vyskytovat i samostatně.

### Klasifikační systémy

- SCOP (Single Curious Overworked Person?)
  - spíše historická kuriozita
  - srovnávání struktur bylo manuální, o klasifikaci rozhodoval člověk na základě svých znalostí a zkušeností

Obrázek 8.9: Prezentace č. 6, slide č. 62



- CATH (Class Architecture Topology Homology)
  - class: jsou struktury proteinu spíše alfa nebo beta
  - architecture: kolik jakých foldů protein obsahuje (sandwich, roll, TIM barrel atd.)
  - topology: jak vypadají smyčky propojující jednotlivé SS
  - homology: jak jsou si struktury sekvenčně podobné

Jak je ze slidů vidět, skoro třetina známých super-rodin spadá do deseti foldovacích skupin. Konkrétně TIM barrel se často vyskytuje u struktur, které mohou mít mnoho různých enzymatických funkcí. Není to ale možno říct s jistotou, stejně jako u jiných složitých struktur.

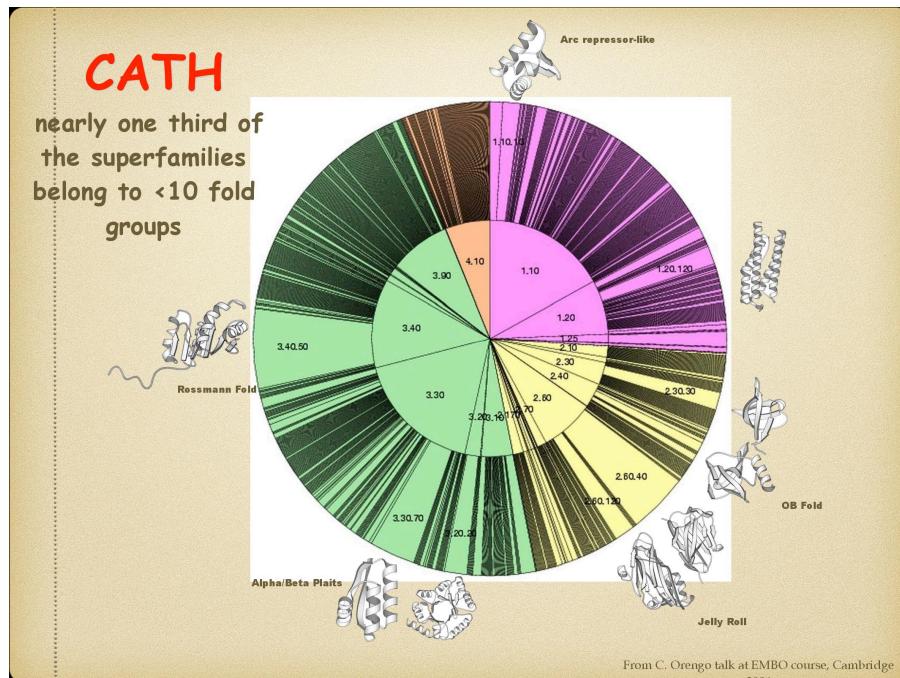
Přednáška č.

8

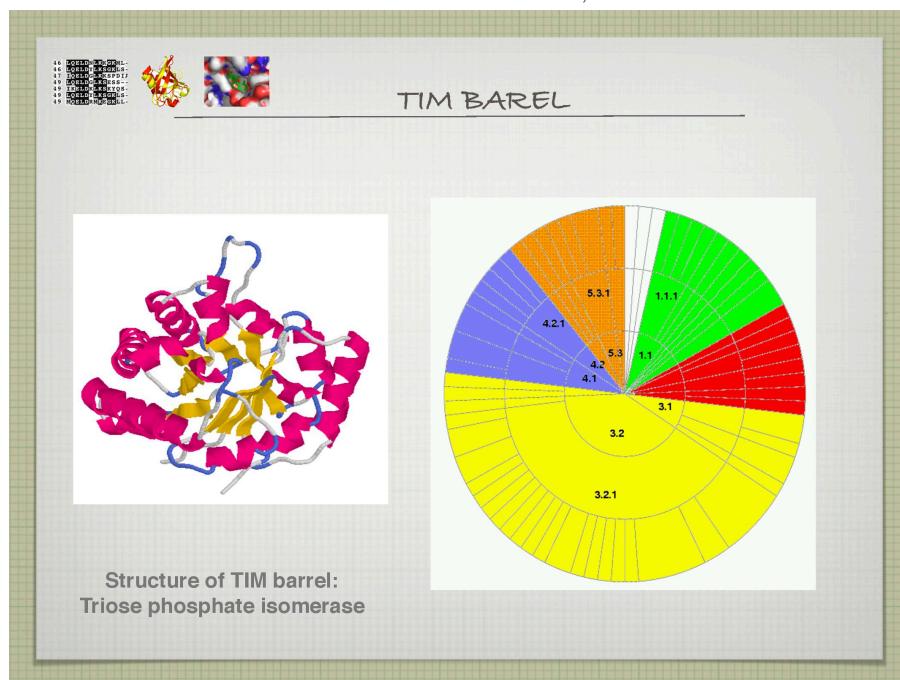
## 8.2 Predikce struktury

Primární struktura (sekvence) proteinu bývá často určena experimentálně, můžeme se tedy pokusit predikovat vyšší struktury. Tato predikce nebývá příliš přesná, míívá

Obrázek 8.10: Prezentace č. 6, slide č. 64



Obrázek 8.11: Prezentace č. 6, slide č. 65



tzv. confidence level, který udává, jak moc je odhad pravděpodobný.

Anfinsen ukázal (1973), že se ribonukleáza po denaturaci sama renaturuje tak, že je schopna vykonávat svou původní funkci a z toho usoudil, že veškerá informace potřebná pro zaujetí struktury je obsažena v sekvenci.

Určení struktury ze sekvence je ale výpočetně velice náročné a někdy to ani není možné.

### 8.2.1 Intrinsically disordered proteins

Proteiny (nebo jejich části), které nemají v nepřítomnosti vazebného partnera nebo ligandu pevnou sekundární a terciární strukturu.

#### Proč jsou zajímavé?

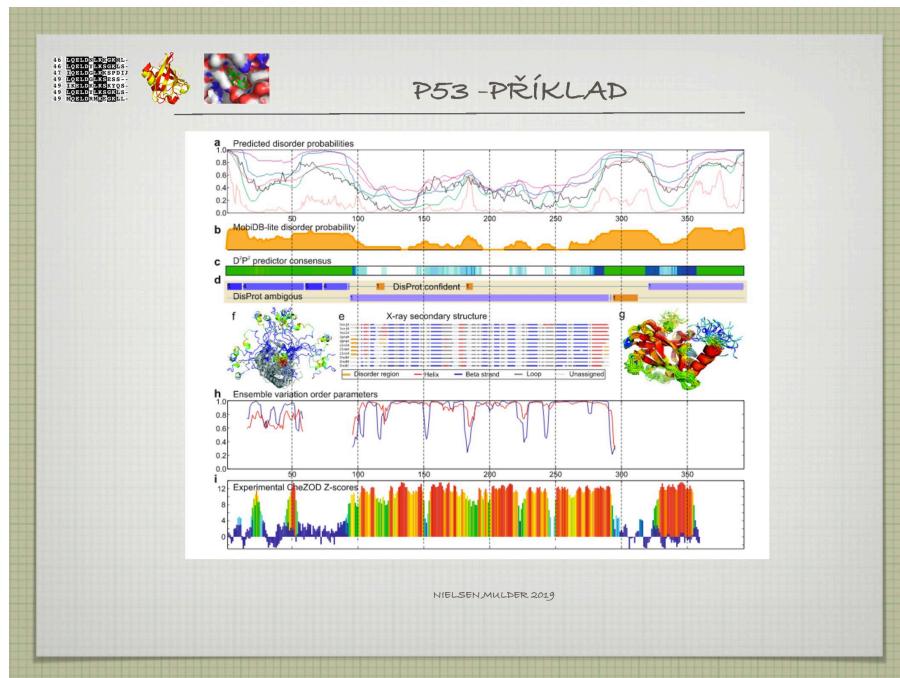
- bývají pro protein (nebo minimálně pro vědce) důležité
- přechod z nestrukturované do strukturované formy je často nezbytný pro funkci proteinu
- komplikují alignmenty, znemožňují krystalizaci
  - je tedy dobré je před krystalizací oddělit

V rámci proteinu jdou části bez pevné struktury často alespoň přibližně poznat, protože mají několik specifických vlastností.

#### Vlastnosti oblastí bez struktury

- mají typické složení
  - malé AK
  - jen málo hydrofobních AK, jinak by se daná část sbalila do SS
  - často se opakují stejné AK, mají nízkou sekvenční komplexitu
- nejsou moc dobře konzervované

Obrázek 8.12: Prezentace č. 7, slide č. 15



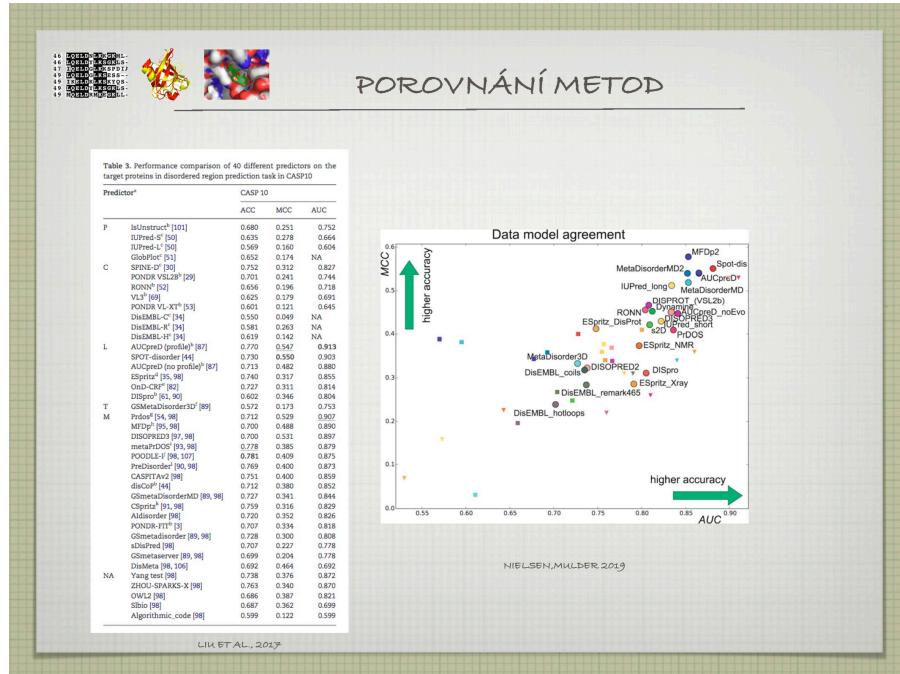
### Predikce oblastí bez struktury

- machine learning, meta servery (spojující několik metod dohromady)
- predikuje se sekundární struktura, AK složení, dostupnost AK pro rozpuštědlo, hot loops atd.
- typická přesnost předpovědí je mezi 60% a 70%
- DisEMBL, FoldIndex, DisoPred, SEG, SPOT-dis, AUCpreD
  - vyplatí se používat kombinaci těchto programů

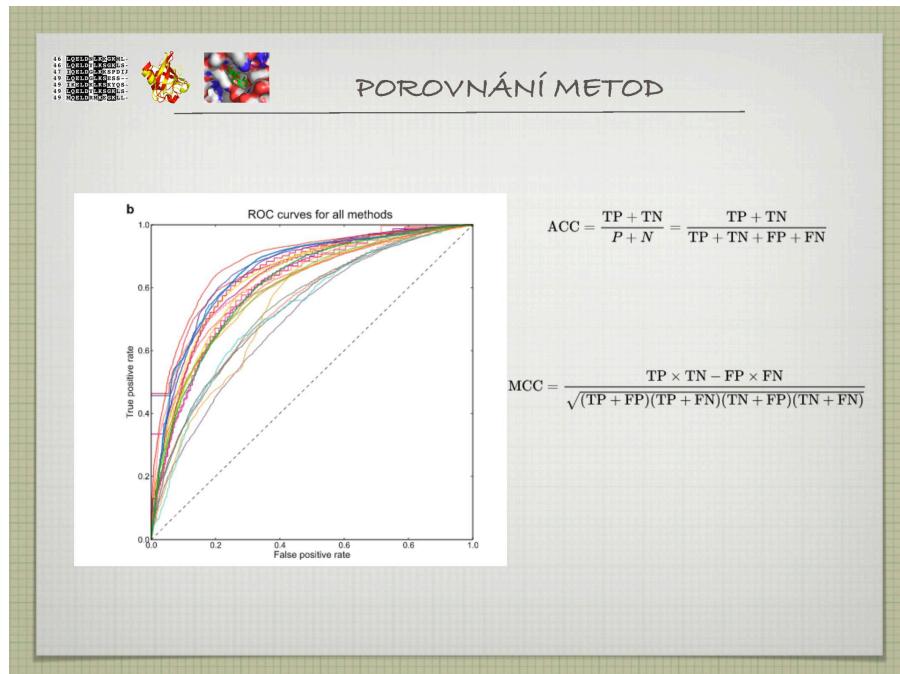
### 8.2.2 Predikce sekundární struktury

Často chceme určit, který druh SS se v proteinu vyskytuje nejčastěji, případně na kterém místě je jaká SS, abychom podle toho mohli vylepšit alignment, či abychom dané informace využili při stavění kompletního 3D modelu proteinu. Druhy SS většinou rozlišujeme pouze tři: helix, list a "zbytek".

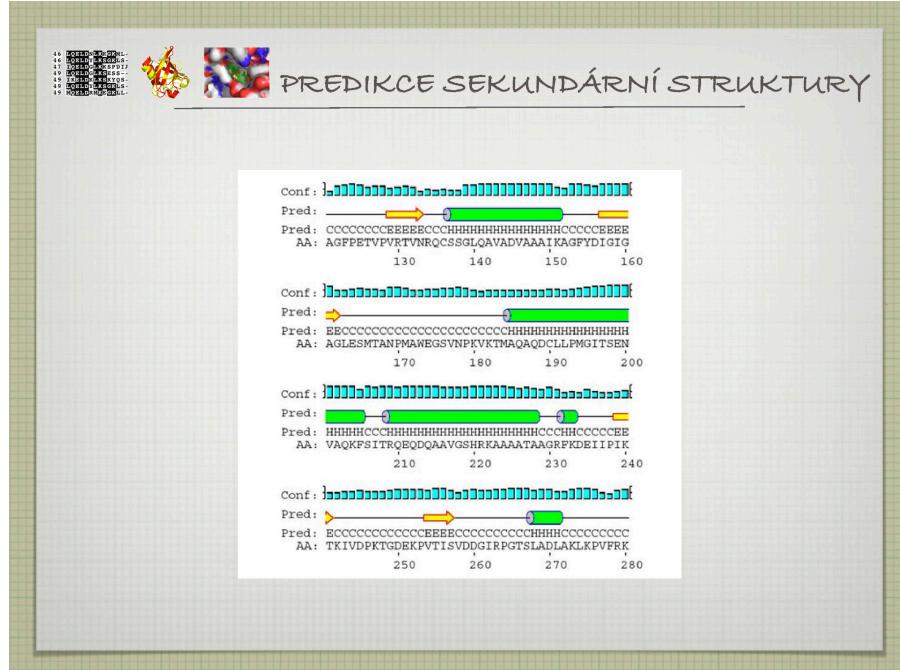
Obrázek 8.13: Prezentace č. 7, slide č. 16



Obrázek 8.14: Prezentace č. 7, slide č. 17



Obrázek 8.15: Prezentace č. 7, slide č. 18



Obrázek 8.16: Prezentace č. 7, slide č. 20



### Metody dříve

- predikce založená pouze na naší sekvenci
- odvozeno z preferencí jednotlivých aminokyselin být v určité SS, které byly experimentálně zjištěny a statisticky zpracovány
  - struktur, ze kterých jsme tato data získávali, bylo sedm
- pouze semiautomatické
- přesnost predikce kolem 60 %
- Chou-Fasman, GOR

### Metody dnes

- známe více sekvencí a jejich struktur, máme tedy více dat
- nové ”učící se” algoritmy, jako HMM a neuronové sítě
  - často využití MSA, které napomáhá správné predikci SS
- přesnost predikce 75%–80% (Q3 skóre, neboli predikce tří různých stavů), navíc dostaneme i odhad významnosti predikce pro každou aminokyselinu
- JPred, PsiPred, APSSP2
- metody jsou benchmarkovány, například benchmarkem EVA

### SS propensity

Udává, v jaké SS se daná AK nejčastěji vyskytuje; to zjistíme z experimentálně naměřených dat.

$$\text{propensita } X \text{ k helixu} = \frac{\text{frekvence } X \text{ v helixu}}{\text{frekvence } X}$$

### Průkopníci

- Chou-Fasman (1974,1978) — původně na 15 strukturách
- klasifikuje AK dle statistik jako [silné, slabé] [makers, breakers] [helixu, listu]
  - skóre 1/0/-1 (breakers, ani-ani, makers)
- postup (dva kroky)
  1. počátek (tzv. nukleace)
    - helix, když má okno o velikosti šest skóre alespoň 4
    - list, když má okno o velikosti pět skóre aspoň 3

Obrázek 8.17: Prezentace č. 7, slide č. 25



PARAMETRY

NAME	P(A)	P(B)	P(TURN)	F(I)	F(I+1)	F(I+2)	F(I+3)
ALANINE	1.42	0.83	0.66	0.06	0.076	0.035	0.058
ARGININE	0.98	0.93	0.95	0.070	0.106	0.099	0.085
ASPARTIC ACID	1.01	0.54	1.46	0.147	0.110	0.179	0.081
ASPARAGINE	0.67	0.89	1.56	0.161	0.083	0.191	0.091
CYSTEINE	0.70	1.19	1.19	0.149	0.050	0.117	0.128
GLUTAMIC ACID	1.39	1.17	0.74	0.056	0.060	0.077	0.064
GLUTAMINE	1.11	1.10	0.98	0.074	0.098	0.037	0.098
GLYCINE	0.57	0.75	1.56	0.102	0.085	0.190	0.152
HISTIDINE	1.00	0.87	0.95	0.140	0.047	0.093	0.054
ISOLEUCINE	1.08	1.60	0.47	0.043	0.034	0.013	0.056
LEUCINE	1.41	1.30	0.59	0.061	0.025	0.036	0.070
LYSINE	1.14	0.74	1.01	0.055	0.115	0.072	0.095
METHIONINE	1.45	1.05	0.60	0.068	0.082	0.014	0.055
PHENYLALANINE	1.13	1.38	0.60	0.059	0.041	0.065	0.065
PROLINE	0.57	0.55	1.52	0.102	0.301	0.034	0.068
SERINE	0.77	0.75	1.43	0.120	0.139	0.125	0.106
THREONINE	0.83	1.19	0.96	0.086	0.108	0.065	0.079
TRYPTOPHAN	1.08	1.37	0.96	0.077	0.013	0.064	0.167
TYROSINE	0.69	1.47	1.14	0.082	0.065	0.114	0.125
VALINE	1.06	1.70	0.50	0.062	0.048	0.028	0.053

Obrázek 8.18: Prezentace č. 7, slide č. 26



PŘÍKLAD (HELIX)

G	P	S	R	Y	I	V	T	L	A	N	G	K
-I	-I	0	0	-I	I	I	0	I	I	-I	-I	I

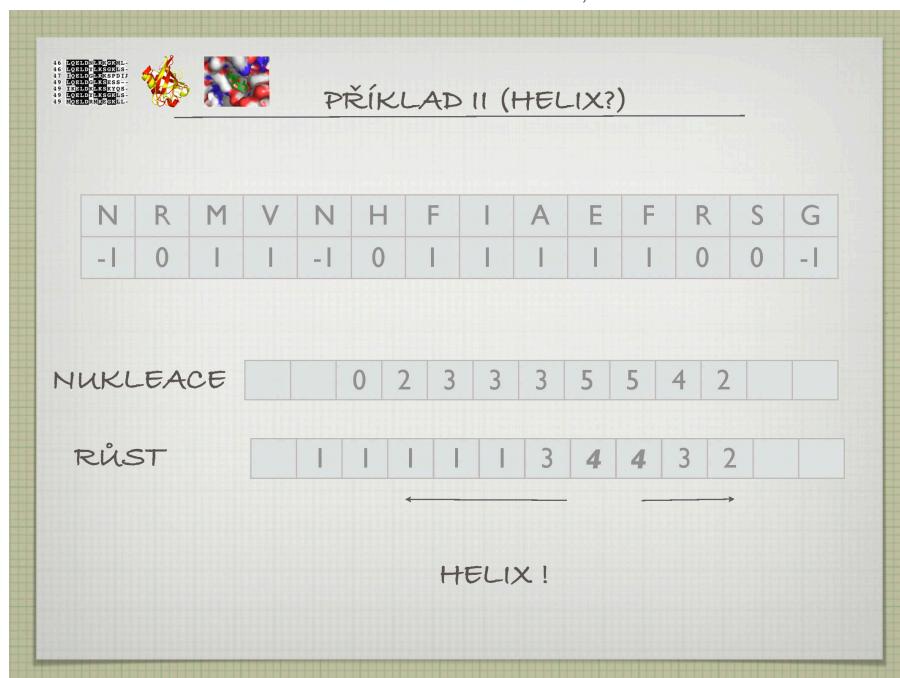
NUKLEACE

			-2									
--	--	--	----	--	--	--	--	--	--	--	--	--

Obrázek 8.19: Prezentace č. 7, slide č. 28



Obrázek 8.20: Prezentace č. 7, slide č. 29



## 2. růst

- postupuj oběma směry od počátku tak dlouho dokud je v okně o velikosti čtyři skóre alespoň +1
- má omezenou přesnost, kolem 60%
  - částečně způsobená malým datasetem, ze kterého byly vypočítány parametry
  - SS je určena i jinými věcmi než jen propensitami AK
  - existují "chameleón" sekvence, ve kterých je na stejné místo predikován list i helix

Trochu lepší výsledky než Chou-Fasman má metoda GOR, která sice také počítá propensities pro všechn 20 AK na určité pozici, ale výpočty u ní závisí i na 16 okolních AK. Výsledná tabulka s čísly je tedy  $20 \times 17$ , místo  $20 \times 1$ .

## Moderní metody

- například PHD, která má úspěšnost přes 70%
  - používá neurální sítě na dvou stupních, MSA
  - vychází z databáze nepříbuzných proteinů

**Vsuvka** Arteficiální neurální sítě (ANN) jsou adaptivní systémy založené na biologickém modelu nervové soustavy. Dají se trénovat: máme určitý test dataset, na kterém daná ANN optimalizuje své parametry.

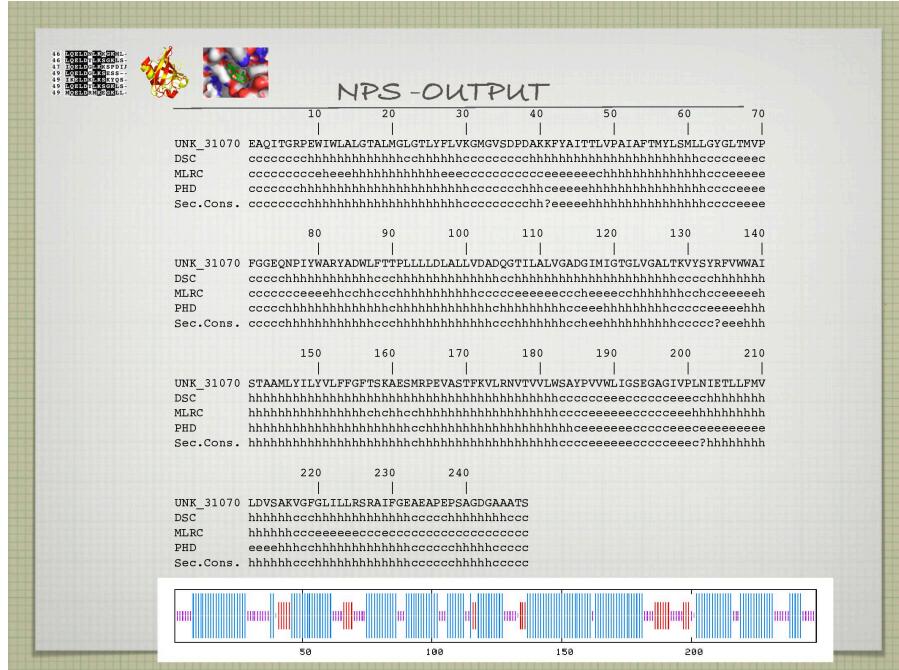
Kromě ANN se používají i metody konsenzu: kombinace několika různých metod pro dosažení optimálního výsledku. Například JPRED a NPS.

### 8.2.3 Predikce membránových proteinů

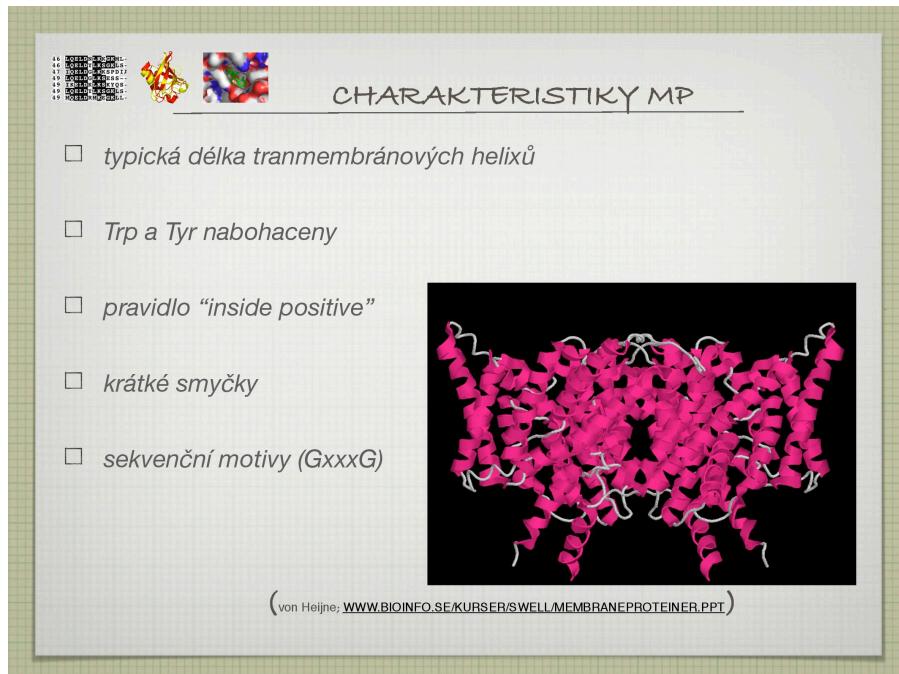
#### Charakteristiké vlastnosti MP

- transmembránové helixy mají typickou délku 20–30 AK
- AK jsou hydrofobní, aby mohly protnout membránu
  - na konci bývají Trp a Tyr (procházejí kolem polárních hlav lipidů)
  - na vnitřní straně bývají Lys a Arg

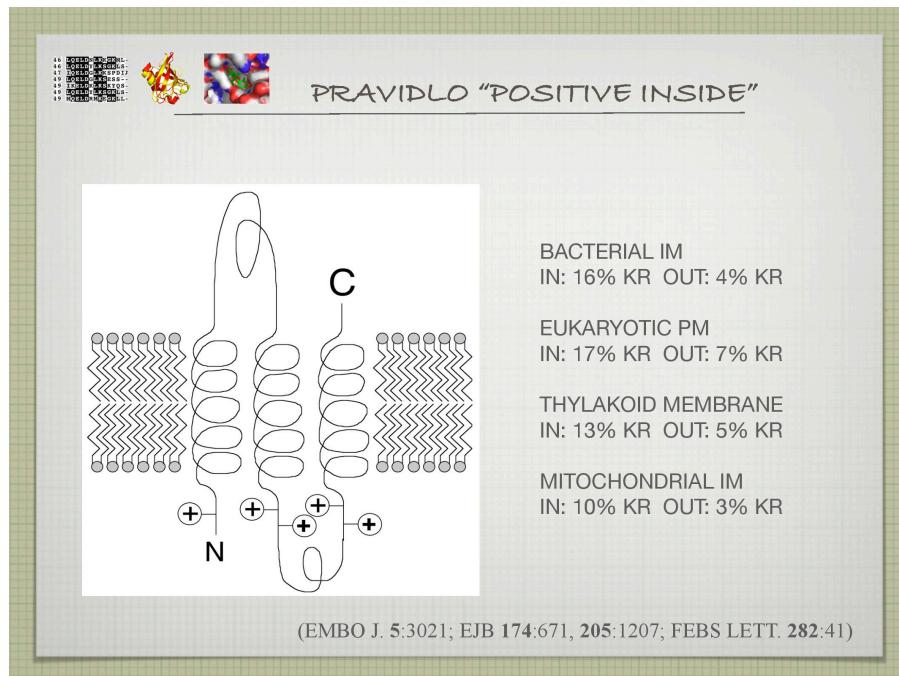
Obrázek 8.21: Prezentace č. 7, slide č. 37



Obrázek 8.22: Prezentace č. 7, slide č. 41



Obrázek 8.23: Prezentace č. 7, slide č. 44



- mívají krátké smyčky
- pravidlo pozitve inside, kladně nabité AK jsou uvnitř buňky

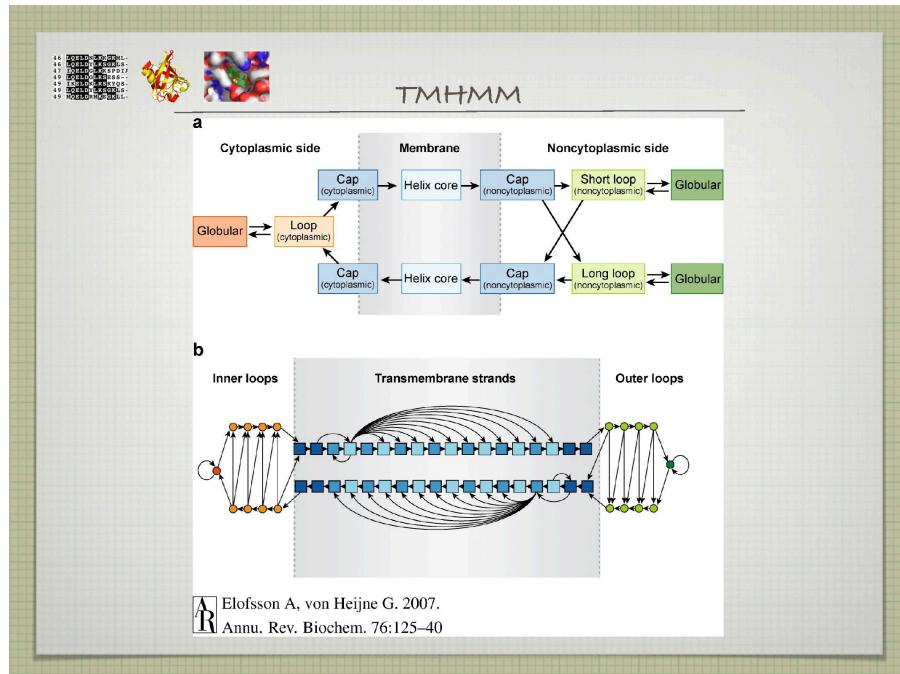
### Predikce MP

- viz také určování membránových proteinů
- TMHMM, PHOBIUS, HMMTOP, TMAP, PHD, SPLIT
- metody s vysokou (> 98%) pravděpodobností rozlišují membránové proteiny od globulárních
  - topologii určí správně v 70% případů (horší pro začátky a konce helixů)
  - úspěšnost se dá zvýšit použitím více metod najednou

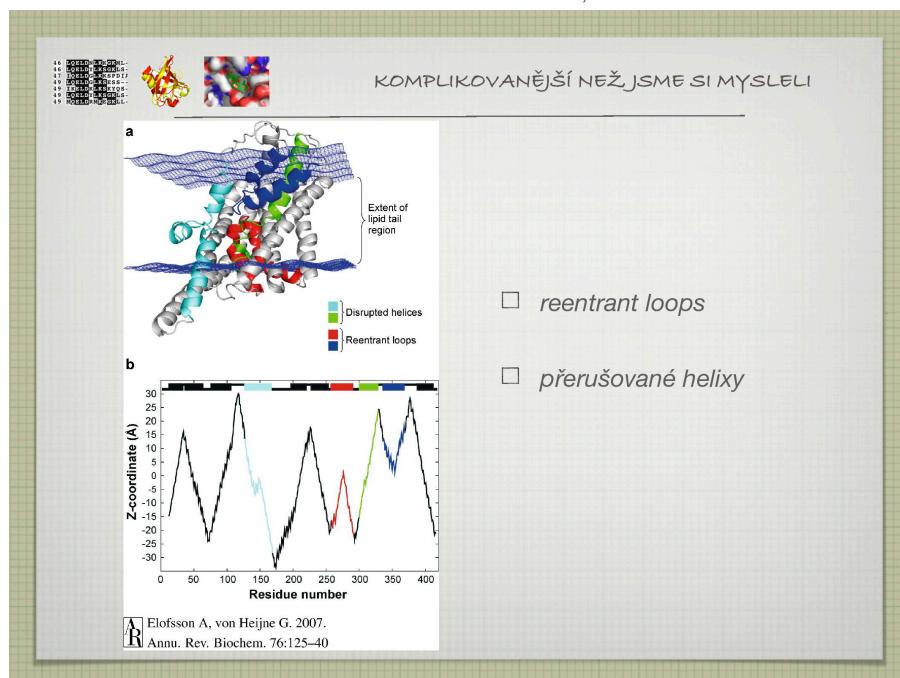
Predikce je komplikována tím, že ne všechny helixy procházejí celou membránou: existují přerušované helixy, které jsou přerušeny uvnitř membrány, a reentrant loops, což jsou helixy, které se vrací zpět na stranu, ze které vyšly.

Beta barrel je věnována menší pozornost, jelikož je jich málo a jsou často bakteriální či mitochondriální.

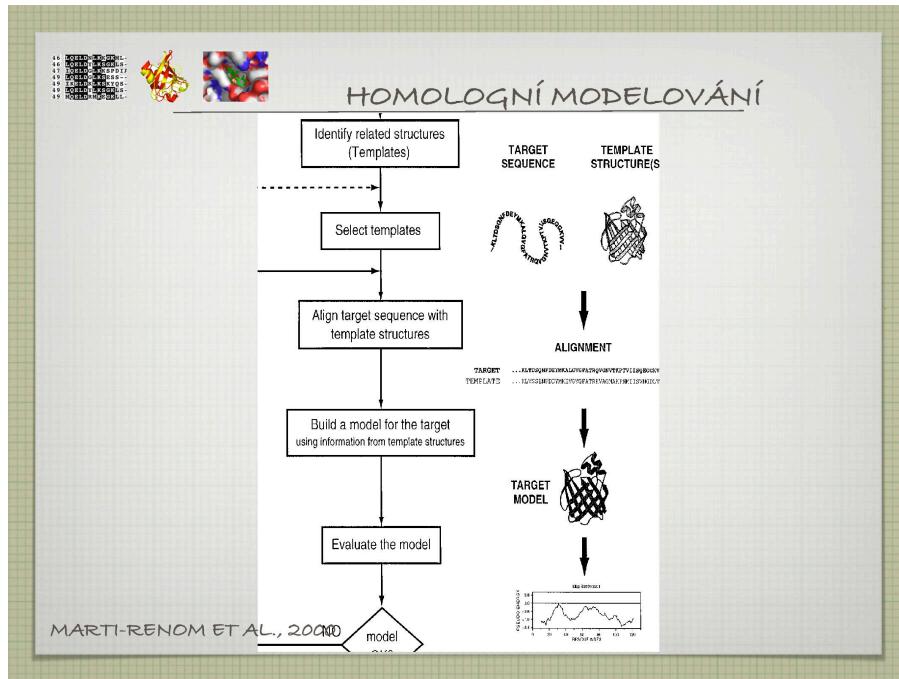
Obrázek 8.24: Prezentace č. 7, slide č. 47



Obrázek 8.25: Prezentace č. 7, slide č. 50



Obrázek 8.26: Prezentace č. 7, slide č. 58



### Vlastnosti beta barelů

- jsou méně hydrofobní než transmembránové helixy
- antiparalelní beta listy jsou spojené s nejbližším sousedem
- N- a C- konec jsou v periplazmatickém prostoru
- mají lichý počet hřebenů (strands)

#### 8.2.4 Homologní modelování

Homologní modelování umožňuje predikci 3D struktury proteinu na základě evoluční příbuznosti (homologie) s proteinem, jehož strukturu už známe (templátem). Ta je většinou stanovena pomocí sequence alignmentu.

Předpokládáme, že určitý sekvenční motiv má dobře známou strukturu, a není tak důvod, aby podobný protein se stejnou sekvencí měl úplně jinou strukturu.

### **Postup**

1. najdeme příbuzné proteiny
2. vybereme z nich vhodný templát
3. uděláme alignment templátu a modelu
4. postupně tvoříme model (4.1 model building) a kontrolujeme jeho kvalitu (4.2 model evaluation)

ad 1. Nalezení příbuzných proteinů) Hledáme ve velkých databázích (pomocí BLAST, FASTA, PSI-BLAST), nejlépe včetně strukturních (PDB, pomocí HMM). Nevhodný templát nám zkazí celou budoucí práci, proto je důležité volit dobře.

ad 2. Výběr templátu) Obecně čím vyšší %SI, tím lepší daný templát je.

### **Strategie výběru**

- pokud je na výběr z několika stuktur stejného proteinu
  - EDS (electron density server): jak moc odpovídá model proteinu experimentálně zjištěným datům
  - B-faktor: jak je struktura stabilní; čím vyšší, tím je nestabilnější
  - rozlišení
- roli hrají i biologické faktory: jakou má kvarterní strukturu, jestli váže ligandy, jestli tvoří komplexy
  - může pro nás být zajímavější templát vázající GTP, takže nevybereme jiný templát, i když má vyšší %SI

Lze vybrat i více templátů, nebo použít různé templáty pro různé části proteinu.

### **Validace výběru**

- zjištění normálnosti
- srovnání délky vazeb, Ramachandranova diagramu
- WhatCheck, ProCheck, EDS

ad 3. Alignment) Toto je nejdůležitější část celého procesu. MSA algoritmy předpokládají, že jsou jejich proteiny homologické, proto je důležité dobře volit templát, aby vzniklý model nebyl nesmyslný.

Je dobré alignment ručně upravit (pokud například známe konzervovaná aktivní místa), případně k jeho tvorbě využít znalost SS. Zbytek informací viz MSA.

ad 4.1 Model building) Na základě alignmentu můžeme vytvořit 3D model sekvence; záleží hlavně na kvalitě templátu, existující programy a modelovací postupy se přesností příliš neliší. Počátečním modelem bývají SS templátu, ve kterých se poté doplňují nebo upravují AK.

### Doplnění modelu

- použití energetické minimalizace
  - vazeb, torzních úhlů, smyček
  - není garantováno, že přinese lepší model
  - není nikdy úplně přesná (ignoruje se roztok atd)
- modelování smyček
  - smyčky se často podílejí na vazbě ligandů, udělují specifitu nebo jsou součástí aktivních míst
  - často nemají protějška v templátu
  - je složité je nafoldovat
    - \* ab initio: fold začínáme od nuly, hledáme ten s nejnižší energií
    - \* databázové modely: v PDB hledáme podobné sekvence smyček a jejich struktury
- hledání rotamerů
  - z možných orientací vybíráme rotamer podle podobnosti s templátem a podle energetických preferencí
  - platí i pro disulfidické můstky
  - použijeme obvykle stejný rotamer jako u templátu
    - \* pokud AK není konzervována, tak se použije nejčastější rotamer
    - \* pokud je nejčastější rotamer v kolizi s jinou AK, použijeme druhý nejčastější rotamer
    - \* atd. => dead-end elimination theorem

Model je nutné nějak zkонтrolovat; nikdy ale nebude zcela odpovídat pravdě. Časté chyby (množství a závažnost roste s klesající %SI): chybný rotamer či pozice AK, chyby v oblastech s nedostatečnou homologií (smyčky), chyby v alignmentu.

ad 4.2 Model evaluation) Modely se hodnotí jednodušeji než alignmenty, proto se často na chybu přijde až v této fázi. Nejlepší je použít WhatCheck, který zkонтroluje celou škálu veličin, které nás zajímají.

### Programy pro homologní modelování

- Swiss-Model: plně automatický
- WhatIf: umožňuje vytvářet vlastní alignment
- Modeller: standartní nástroj
- Phyre, Tasser

WhatIf a Modeller vyžadují větší zkušenosti, jsou ale věrohodnější.

#### 8.2.5 Fold recognition

Fold recognition metody používáme, když neumíme najít templát se známou strukturou, který by byl homologní k naší sekvenci. Snažíme se najít nehomologní proteiny, které přesto mají alespoň část své struktury shodnou s částí struktury naší sekvence. V tom nám pomáhá to, že dovolených foldů je omezené množství a stejně foldy se často opakují (na 130000 známých strukturách je jen 1375 různých foldů) — pokud uvažujeme nějaký protein bez detekovatelného %SI, ze 70–80% bude mít fold, který už je známý.

Existují dva základní postupy, které se liší svou metodikou i úspěšností: profile a threading metody.

#### Profile metody

1. uděláme profil naší sekvence
  - každá AK zařazena do jedné z 18 skupin na základě predikce její oblíbené SS (helix, beta list, zbytek) a toho, kde se nachází (uvnitř, na povrchu, atd. — 6 skupin)
2. stejný profil uděláme pro všechny známé sekvence
  - z 3D informací (struktura) tedy tvoříme 1D informace (profil)

Obrázek 8.27: Prezentace č. 8, slide č. 46



3. pro takto vzniklé profily počítáme alignment a z něj pak predikujeme vlastnosti struktury naší sekvence
  - například programy 3D PSSM, Phyre

Na rozdíl od profile metod se threading metody nesnaží ze známých struktur vytvořit profily (3D -> 1D), ale naopak chtějí z naší sekvence získat nějaké informace o struktuře (1D -> 3D).

## Postup

1. na naší sekvenci "navlékáme" nějaký fold z databáze foldů (tzv. threading)
2. tento fold zkoušíme na naší sekvenci různě naalignovat a pro každý alignment spočítáme jeho skóre
  - skóre se většinou počítá energetickou funkcí, která optimalizuje energii párových interakcí a solvatace
  - oskórováním zjistíme, jak moc je naše sekvence kompatibilní se strukturou, kterou jsme jí přisoudili

3. výsledný fold a alignment použijeme pro tvorbu modelu, která probíhá podobně jako při homologním modelování

Threading metody jsou sofistikovanější než profile metody a přináší lepší výsledky (například program Threader2). Alfa proteiny (proteiny s více helixy než beta listy) se predikují lépe než beta proteiny, protože alfa helixy tvoří lokální vodíkové můstky, zatímco beta listy tvoří vodíkové můstky spíše mezi AK, které jsou od sebe v sekvenci více vzdálené.

### 8.2.6 Ab initio predikce

Homologní modelování ani fold recognition nemohou uspět, pokud má protein zcela nový fold. Když hledáme takovýto nový fold, řešíme vlastně nejobecnější problém foldingu (hledání struktury pouze se znalostí dané sekvence).

Předpokládáme, že správně nafoldovaný protein bude mít nejnižší energii; náš problém tedy převedeme na problém hledání struktury s nejnižší energií.

#### Problémy s hledáním struktury

- je to výpočetně velice náročné
- je to stále dosti nepřesné, protože nemáme dost detailní rovnice na výpočet energie proteinu
- nativní konformace proteinu často není energeticky příliš odlišná od jeho nestabilní (nesbalené) konformace

#### Způsoby hledání struktury

- používáme pravidla, která jsme odvodili z pozorovaných struktur
- užíváme zjednodušené reprezentace
  - rozpouštědlo je pouze jedna entita
  - nemodelujeme celé AK, ale pouze C $\alpha$
- definujeme energetické funkce, které popisují fyzikálně-chemické vlastnosti proteinu
- hledáme konformaci, pro kterou bude hodnota takové funkce minimální

- používáme Monte Carlo metody, abychom unikli z lokálních minim (a měli šanci najít globální minimum)
- nejsme schopní identifikovat vhodný templát pro celou strukturu, ale pro 5 AK jej umíme poskládat

Jeden z nejlepších nástrojů pro predikci struktur je Rosetta.  $\text{C}\alpha$  RMSD je méně než 1,5 Å mezi modelem a experimentálně určenou strukturou. Rosetta kombinuje fragmenty, používá NMR a energetické funkce.

## CASP

- Critical Assessment of Techniques for Protein Structure Prediction
- soutěž prediktivních metod
- sekvence, jejichž struktury jsou těsně před objevením, se zašlou několika výzkumným týmům, které poté predikují jejich strukturu
- vypočítaný model je poté porovnán s experimentálně objevenou strukturou

### 8.2.7 Predikce interakce

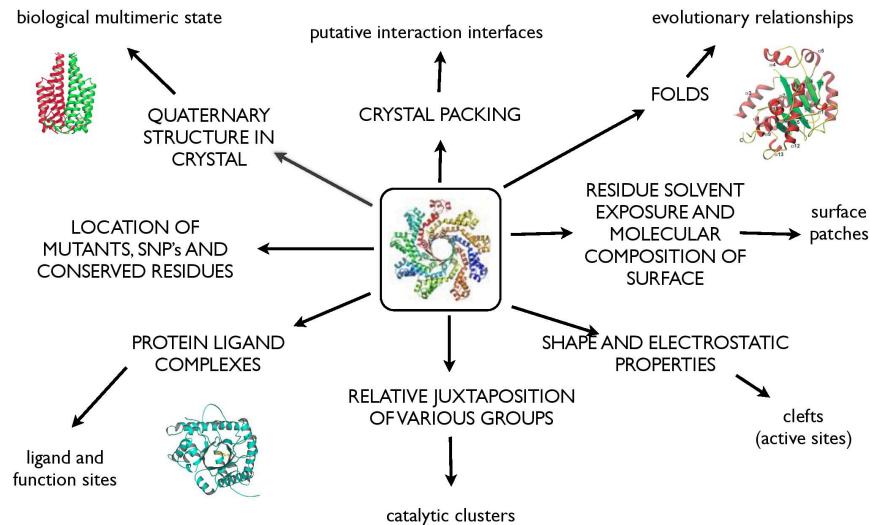
Proteiny, které spolu interagují, se obvykle vyvíjejí společně a synchronně; mutace v jednom z proteinů jsou kompenzovány mutacemi v druhém. Používá se proto *in silico* dvouhybridní systém: udělá se MSA obou proteinů a pokud vykazují podobnou frekvenci mutací, může se jednat o interakční páry.

#### Nástroje na predikci interakcí

- Bayesiánské metody (někdy kombinují i více přístupů)
- InterProSurf, PIP

**TODO** Udělat pořádek v nadpisech a hierarchii: například na tomto místě bylo srhnutí predikce struktury.

Obrázek 9.1: Prezentace č. 8, slide č. 69



Adapted from Bartlett et al., in Structural Bioinformatics

## 9 Souvislost struktury a funkce

Ze sekvence proteinu lze (s omezenou přesností, viz předchozí část) odvodit jeho strukturu. Podobně lze, opět s omezenou přesností, z jeho struktury odvodit jeho funkci.

Hlavní paradigma tedy zní: podobná sekvence  $\Leftrightarrow$  podobná struktura  $\Leftrightarrow$  podobná funkce.

### Co je funkce? (příklad: alkohol dehydrogenása)

- biochemická funkce: enzymatická, na zinku závislá, alcohol dehydrogenásová aktivita
- buněčná funkce: metabolizmus alkoholu
- buněčná lokalizace: cytoplazma
- fenotypická funkce: alkoholismus

Existuje databáze Gene Ontology, která ukládá definované atributy genů a proteinů; popisuje proteiny na třech úrovních: molecular function, biological process, cellular component.

Bohužel, hlavní paradigma ne vždy funguje; jeden protein (jedna struktura) může mít více různých funkcí a jedna funkce může být splněna několika různými strukturami.

**TODO** Jak je to s lysozymem a alpha-lactalbuminem? Jsou nebo nejsou to enzymy?

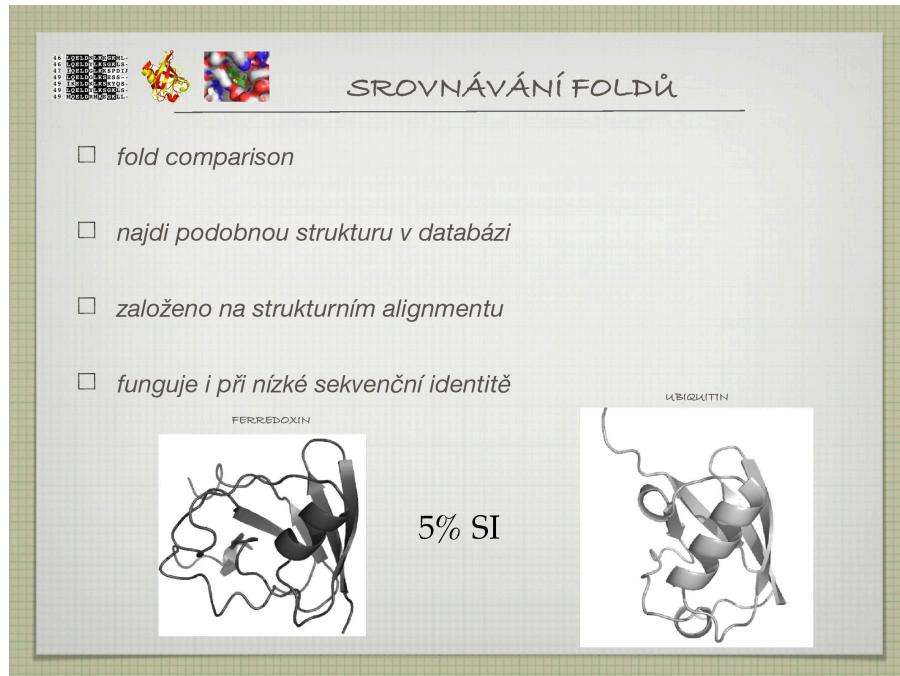
### Vady v paradigmatu

- jeden protein více funkcí, neboli moonlighting
  - dva crystalliny, 94% SI, jeden je pouze v čočce, druhý zvládne být i argininosukcinát
- jeden fold více funkcí
  - TIM barell (25 EC čísel), alpha/beta hydroláza (17 EC čísel)
    - \* jedno EC (enzyme commission) číslo popisuje jednu konkrétní enzymatickou reakci
  - lysozyme a alpha-lactalbumin, pouze 40% SI, ale oba jsou enzymy
- jedna funkce více struktur
  - beta-lactamáza A a beta-lactamáza B

## 9.1 Hledání funkce

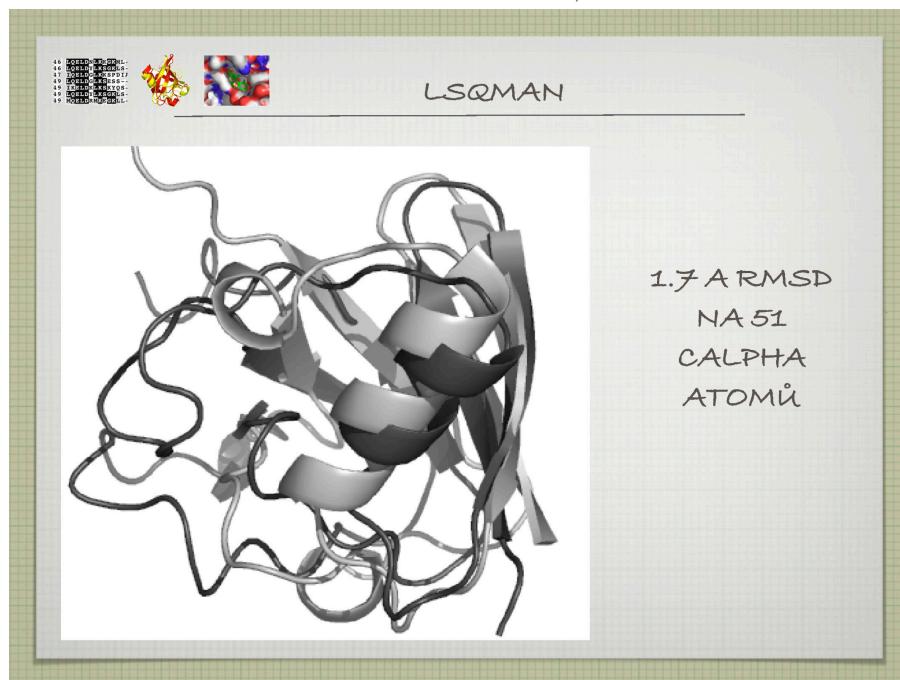
- analýza kvarterní struktury proteinu
  - často ji neznáme, protože struktury proteinů nejsou vždy určeny ve své nativní konformaci, musíme ji tedy odhadnout
  - zjistíme, zda protein funguje jako monomer, dimer, atd.
  - provedeme analýzu intermolekulárních kontaktů
  - nástroje PQS, PISA
- fold comparison se známými strukturami
  - hledání podobné struktury v databázi

Obrázek 9.2: Prezentace č. 8, slide č. 81



- funguje na základě strukturního alignmentu, tedy i pro sekvence s nízkým %SI
  - program LSQMAN
  - kombinace strukturních a evolučních metod
  - hledání 3D motivů
    - motivy lze automaticky extrahovat z dobře anotovaných struktur
    - hledání odpovědi na otázku: objevuje se alespoň jeden z takových motivů v nové struktuře?
      - \* programy JESS, PINTS
    - definování libovolného strukturního motivu
      - \* program SPASM
    - "reverse templates"- rozsekání struktury na motivy a hledání podobných fragmentů v databázi
      - \* program SiteSeer
  - kombinace výše uvedených, například server ProFunc

Obrázek 9.3: Prezentace č. 9, slide č. 22



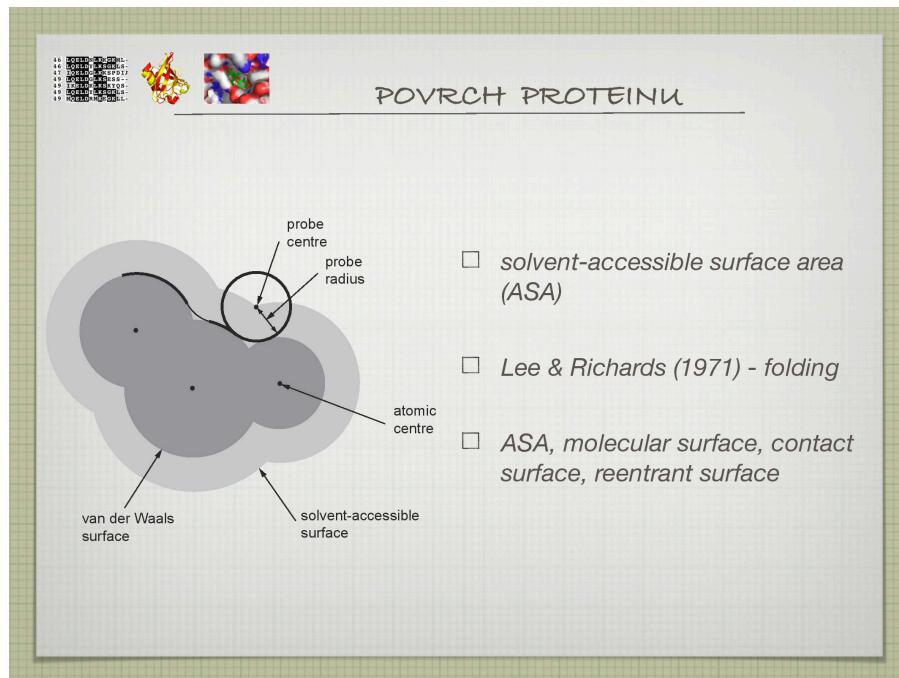
### Zjišťování protein-protein interakcí

- najít interakce s malými molekulami (ionty) je relativně snadné, neboť jsou na povrchu proteinu často "výdutě" speciálně přizpůsobené danému ligandu
- najít interakce s jiným proteinem je naopak složité, protože často interagují velkou částí svých povrchů
- tyto interakce se pozorují pomocí two-hybrid interactions
  - velké množství proteinů "z knihovny" je exprimováno v mnoha buňkách, společně s jedním bait proteinem
  - trikem zařídíme, že proliferují jen ty buňky, kde protein "z knihovny" interahuje s bait proteinem

#### 9.1.1 Solvent-accessible surface area

Solvent-accessible surface area (ASA) je metoda, kterou se zjišťuje "povrch" proteinu (viz slide).

Obrázek 9.4: Prezentace č. 8, slide č. 84



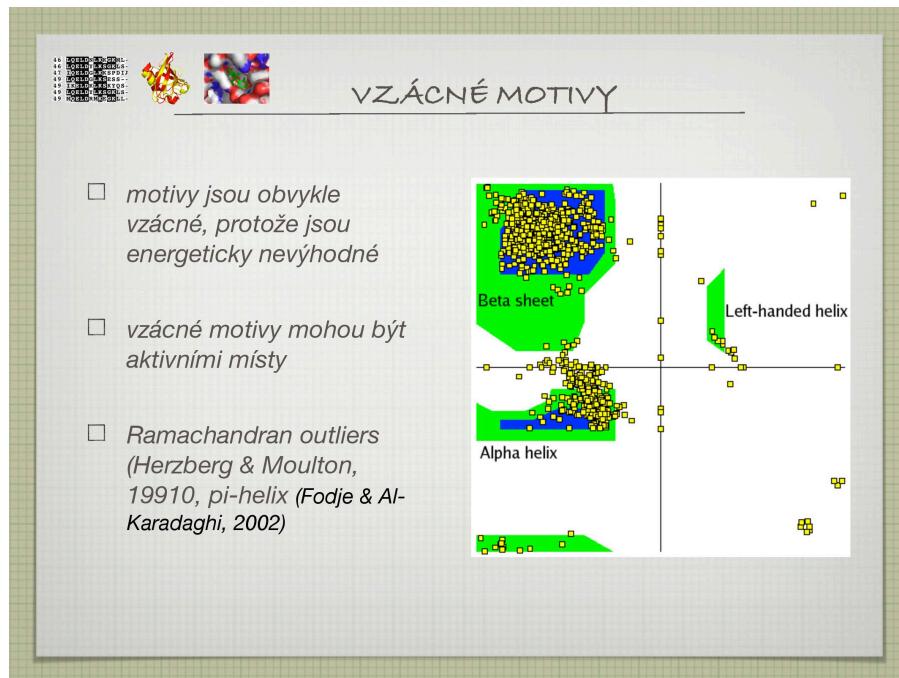
### Využití ASA

- určování kvartérní struktury proteinů
- skórování "docking solutions"
  - docking solution je predikce místa a způsobu vázání ligandu
- srovnávání příbuzných struktur
- charakterizace interakčních povrchů

### 9.1.2 Enzymy

Zjišťování funkce enzymů je pro vědce běžně nejdůležitější. Práci nám někdy ulehčuje to, že určité motivy či části sekvencí se vyskytují pouze ve spojitosti s určým ligandem nebo konkrétní funkcí.

Obrázek 9.5: Prezentace č. 9, slide č. 33



### Speciální případy motivů

Motivy obecně bývají spíše vzácné, protože jsou energeticky nevýhodné; když už tedy v proteinu jsou, je to často v jeho aktivním místě (kde jsou nezbytné).

### HTH motiv

- nejčastější DNA vazebná doména
- pouze 0,5% FP, když skenujeme strukturu proti databázi HTH templátů (má-li naše struktura co dočinění s DNA)

### Katalytická triáda

- specifická trojice AK, která se často vyskytuje v aktivních místech enzymů, hlavně hydroláz a transferáz
- trojici tvoří kyselá, zásaditá a nukleofilní AK
  - například Ser-His-Asp, Cys-His-Asp

Obrázek 9.6: Prezentace č. 9, slide č. 29

**SPECIÁLNÍ PŘÍPADY**

- HTH motiv je nejčastější DNA-vazebnou doménou*
- skenuj strukturu proti databázi HTH templátů*
- RMSD, ASA, elektrostatický potenciál*
- 0.5 % false positive*

Obrázek 9.7: Prezentace č. 9, slide č. 31

**KATALYTICKÁ TRIÁDA**

EBI Home About EBI Groups Services Toolbox Databases Downloads Submissions  
Catalytic Site Atlas Version 2.1.4

Find Annotated Site: PDB code:  Search Swiss-Prot code:  Search EC number:  Search Browse literature entries Help

**CSA entry for 1coy**

Title:	Oxidoreductase(oxygen receptor)
Compound:	Cholesterol oxidase (e.c.1.1.3.6) complex with 3-beta-hydroxy-5-androsten-17-one (dehyd)
Mutant:	No
UniProt/Swiss-Prot:	P22637-CHOD_BREST
Other CSA Entries:	Homologues of 1coy Entries for UniProt/Swiss-Prot: P22637 Entries for EC: 1.1.3.6
EC Class:	
Other Databases:	

**Introduction:** EC 1.1.3.6, cholesterol oxidase is the first step in cholesterol catabolism used by a wide range of soil bacteria which can use cholesterol as sole carbon source. The enzyme may be used as a potent insecticidal agent, active against boll weevil larvae and other insects. The enzyme acts by lysing the cells of the mid-gut epithelium resulting in larval death.

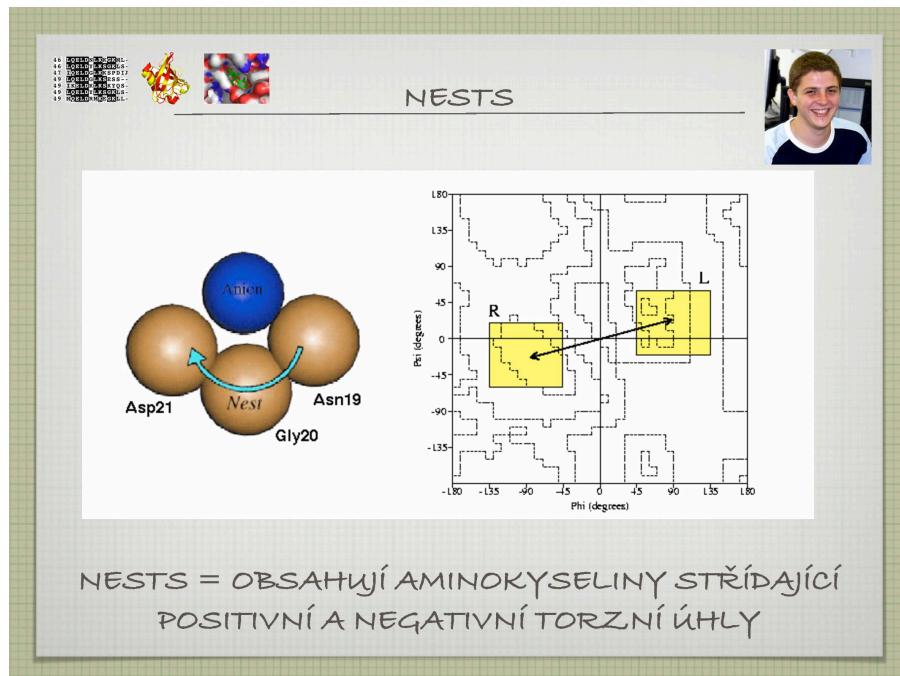
This enzyme is part of a wider family of Cholesterol-Methane-Glucose oxidoreductases or GMC oxidoreductases. The enzyme also performs an isomerisation of the cholesterol to a cholesterol-4-enol.

**Mechanism:** Structural and kinetic data suggests His447 and Glu361 act as general base catalysts in association with H(2)541 and Asn485. Site-directed mutagenesis confirms that these residues are important. Glu361 may be important in the isomerisation step

**Sites:**

<input checked="" type="checkbox"/> <b>Catalytic Site</b> (Get help with this section)
Found by: Literature reference (Structural analysis and templates exist for the 1coy family)
Residue Chain Number UniProt number Functional part
GLU 361 406 Sidechain
HIS 447 492 Sidechain

Obrázek 9.8: Prezentace č. 9, slide č. 32



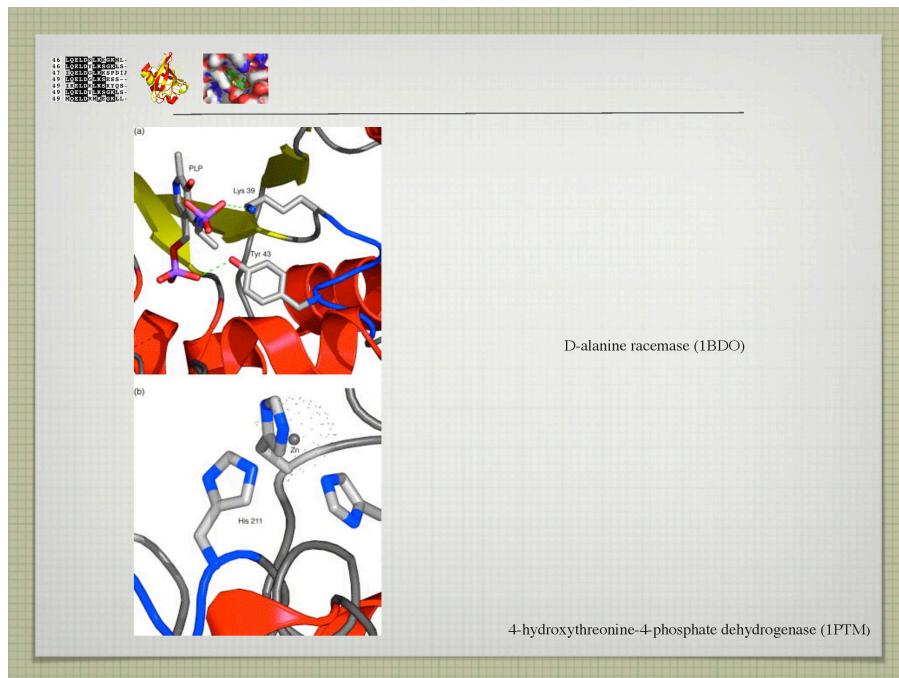
## NESTS

- oddíl 3 AK
  - první a třetí AK míří stejným směrem, tvoří "dutinu", na jejich NH skupiny se váže anion
  - druhá AK většinou míří opačným směrem
- v nestu (hnízdě) se tedy střídají AK s pozitivními a negativními torzními úhly
- v jednom "hnízdě" může být i více aniontů

## Levotočivý helix

- velice vzácný; když už se někde vyskytuje, nejpíše to nebude náhoda
- a vskutku, velice často je levotočivý helix přímo součástí aktivního místa nebo je v jeho bezprostřední blízkosti

Obrázek 9.9: Prezentace č. 9, slide č. 37



### Detekce ligand-vazebných míst

#### Základní způsoby detekce

- u enzymů bývá obvykle největší žlábek na povrchu => detekujeme žlábky
  - zpřesnění při použití evolučních informací (SurfNet + Consurf)
    - \* po odstranění slabě konzervovaných aminokyselin zmenšuje FP o 30%
    - \* přitom FN jsou pouze na 13%

Pokud najdeme podobnost v ligand-vazebných místech dvou různých proteinů, dá se předpokládat, že váží podobné ligandy a mají podobnou funkci. Vazebná místa proto popisujeme matematickými metodami — získáme real sphere harmonic coefficient — a srovnáváme tento koeficient mezi proteiny.

#### **promiskuita**

Jev, kdy se některé ligandy (např. ATP, NAD) na proteiny vážou v mnoha různých konformacích, viz slide.

Obrázek 9.10: Prezentace č. 8, slide č. 42



- používá se, když není nalezen žádný věrohodný templát pro homologní modelování
- hledá odpověď na otázku: Která struktura je kompatibilní s danou sekvencí?
- pokud je nějaká nalezena, lze použít homologní modelování pro vytvoření 3D modelu

### virtual screening

Bioinformatická metoda, jejímž cílem je odhadnout, jak dobře se daná nízko-molekulární sloučenina váže na protein; lze ji tedy v principu využít k predikci ligandů pro danou strukturu. Pro svou funkci používá docking (viz níže).

Je využívána farmaceutickými firmami, které navrhnu mnoho takovýchto látek, pro všechny udělají virtual screening a z nich vyberou několik nejlepších kandidátů, kteří půjdou do dalších testů.

### docking

Molecular docking je proces, který se pokouší nalézt nízkoenergetické vazebné módy dvou molekul (obvykle proteinu a jeho ligandu, případně dvou proteinů). Je to spíše chemická ne bioinformatická metoda.

### Postup při dockingu

- konformační hledání (binding poses)

- cílem je efektivně obsáhnout možné rotace a translace ligandu i proteinu, aby byla mezi vzniklými řešeními i nativní konformace ligandu a proteinu
- náročné na výpočetní techniku
  - \* kdysi byly jen metody, kdy je ligand i protein rigidní, dnes už i semi-flexibilní metody (ligand flexibilní), a objevují se i metody flexibilní (ve kterých alespoň část proteinu může během hledání měnit konformaci)
- nejlepší metody se dostávají na  $1,5\text{\AA}$  až  $2\text{\AA}$  RMSD mezi predikovanou a skutečnou pozicí ligandu
- skórování vzniklých řešení
  - obvykle se jedná o energetickou funkci, která počítá energii vazby a vybírá orientace s nejnižší energií
    - \* nejlepší metody se dostávají na 7–10 kJ/mol od experimentálně měřených volných energií vazby
  - nejsložitější část dockingu

Někdy lze k predikci vazebných partnerů využít i strukturní informace.

**META** Tak, a teď si dejte čokoládu.