

## Oznamy

- DÚ 2 je na stránke, odovzdať do 3.12.
- Všetky stretnutia skupín oznámené, skupine 6 chýba krátka správa
  - niektoré skupiny nemali úplnú účasť
  - so zvyšnými členmi sa skúste spojiť inak
  - ak sa neozývajú, nepíšte ich na správu, dostanú 0 bodov
- Termíny na konci semestra
  - posledná prednáška 4.12., posledné cvičenia 11.12.
  - DÚ 3 streda 17.12., správy zo journal clubu piatok 19.12.
  - nepovinné prezentácie journal clubu 11.12. (prípadne 18.12.)

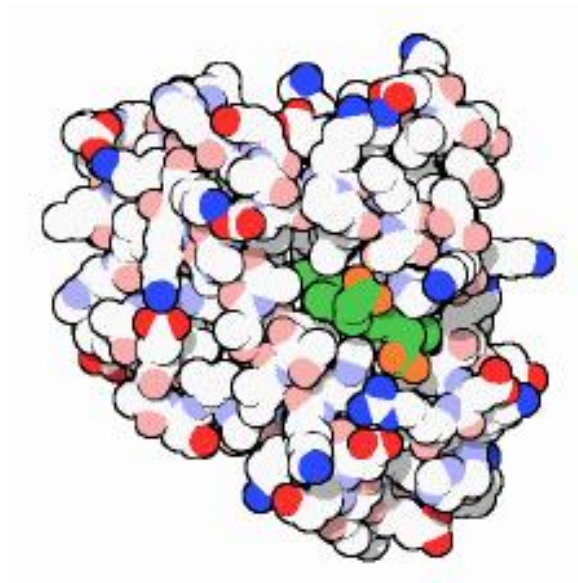
## Správa zo journal clubu

- Pochopiteľná pre študentov tohto predmetu (inf aj bio)
- Vysvetlite pojmy, ktoré sú nad rámec tohto predmetu
- Netreba pokryť všetko, môžete využiť aj iné zdroje
- Podrobne vysvetliť aspoň jednu bioinformatickú metódu a aspoň jeden biologický výsledok (alebo overovanie správnosti metódy na dátach)
- Ako článok súvisí s učivom preberaným na predmete
- Nájdite zopár citujúcich prác, ktoré výsledky využili alebo vylepšili
- Rozsah cca 1-2 strany na osobu, jeden ucelený text
- Píšte vlastnými slovami, citujte zdroje
- V správe vymenujte členov skupiny, ktorí sa podieľali na jej spísaní, dostanú rovnako bodov
- Pdf odovzdať cez Moodle (stačí 1 za skupinu)

# Štruktúra a funkcia proteínov

Broňa Brejová

20.11.2025



## Proteíny

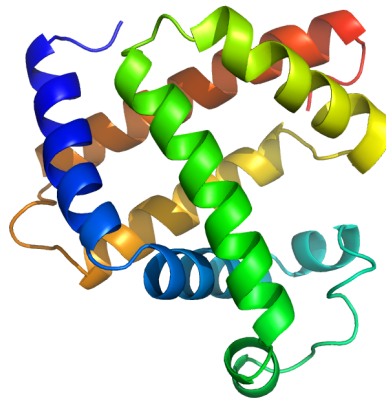
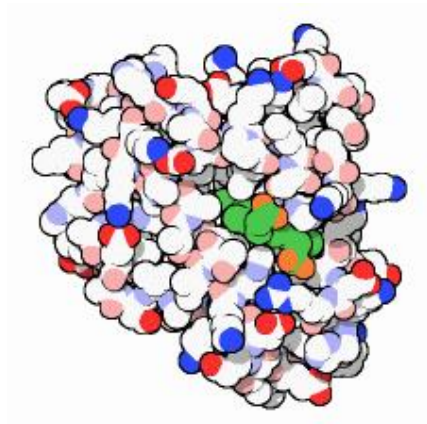
Reťazce 20 rôznych aminokyselín s rôznymi chemickými vlastnosťami:

| Aminokyselina           | Postranný reťazec  | Jeho vlastnosti |
|-------------------------|--|-----------------|
| Alanín (A)              | -CH <sub>3</sub>   | hydrofóbny      |
| Arginín (R)             | -(CH <sub>2</sub> ) <sub>3</sub> NH-C(NH)NH <sub>2</sub>       | bázický         |
| Asparagín (N)           | -CH <sub>2</sub> CONH <sub>2</sub>                             | hydrofilný      |
| Kyselina asparágová (D) | -CH <sub>2</sub> COOH  | kyslý           |
| Cysteín (C)             | -CH <sub>2</sub> SH  | hydrofóbny      |
| Kyselina glutámová (E)  | -CH <sub>2</sub> CH <sub>2</sub> COOH                          | kyslý           |
| Glutamín (Q)            | -CH <sub>2</sub> CH <sub>2</sub> CONH <sub>2</sub>             | hydrofilný      |
| Glycín (G)              | -H   | hydrofilný      |
| Histidín (H)            | -CH <sub>2</sub> -C <sub>3</sub> H <sub>3</sub> N <sub>2</sub> | bázický         |
| Izoleucín (I)           | -CH(CH <sub>3</sub> )CH <sub>2</sub> CH <sub>3</sub>           | hydrofóbny      |
| Leucín (L)              | -CH <sub>2</sub> CH(CH <sub>3</sub> ) <sub>2</sub>             | hydrofóbny      |
| Lyzín (K)               | -(CH <sub>2</sub> ) <sub>4</sub> NH <sub>2</sub>               | bázický         |
| Metionín (M)            | -CH <sub>2</sub> CH <sub>2</sub> SCH <sub>3</sub>              | hydrofóbny      |
| Fenylalanín (F)         | -CH <sub>2</sub> C <sub>6</sub> H <sub>5</sub>                 | hydrofóbny      |
| Prolín (P)              | -CH <sub>2</sub> CH <sub>2</sub> CH <sub>2</sub> -             | hydrofóbny      |
| Serín (S)               | -CH <sub>2</sub> OH  | hydrofilný      |
| Treonín (T)             | -CH(OH)CH <sub>3</sub>   | hydrofilný      |
| Tryptofán (W)           | -CH <sub>2</sub> C <sub>8</sub> H <sub>6</sub> N               | hydrofóbny      |
| Tyrozín (Y)             | -CH <sub>2</sub> -C <sub>6</sub> H <sub>4</sub> OH             | hydrofóbny      |
| Valín (V)               | -CH(CH <sub>3</sub> ) <sub>2</sub>                             | hydrofóbny      |

## Štruktúra proteínov

- **Primárna štruktúra:** sekvencia aminokyselín
- **Sekundárna štruktúra:** pravidelné útvary  
alfa-hélix, beta-skladaný list (beta sheet)
- **Terciálna štruktúra:** presné 3D rozloženie atómov
- **Kvartérna štruktúra:** interakcia viacerých proteínov v komplexe

Myoglobín, prvý proteín so známou štruktúrou [Kendrew et al 1958]



## Experimentálne určovanie štruktúry

- RTG kryštalografia (X-ray crystallography)  
vyžaduje proteín v kryštalickej forme
- NMR (nuclear magnetic resonance spectroscopy)  
hlavne používaná na kratšie proteíny
- Cryo-EM (cryogenic electron microscopy)  
vhodná na veľké proteínové komplexy, rastúca popularita
- Náročný a drahý proces
- Databáza štruktúr PDB  
228 000 proteínových štruktúr (83% X-ray)  
(UniProt má 250 miliónov sekvencií)

## Určovanie štruktúry proteínov ako bioinformatický problém

(protein structure prediction, protein folding)

**Vstup:** sekvencia proteínu  $X$

**Výstup:** 3D pozície atómov alebo aminokyselín

### (1) Ab initio metódy

- Nájsť štruktúru s najnižšou voľnou energiou
- Vzorce na približný výpočet energie založené na fyzike
  - sily medzi atómami v proteíne a okolitom roztoku
- Veľmi ťažký výpočtový problém
  - simulácia molekulárnej dynamiky
  - optimalizačné metódy, napr. gradientová metóda, simulované žíhanie
- Používané na malé proteíny a zlepšenie približných štruktúr

## Určovanie štruktúry proteínov ako bioinformatický problém

(protein structure prediction, protein folding)

**Vstup:** sekvencia proteínu  $X$

**Výstup:** 3D pozície atómov alebo aminokyselín

### (1) Ab initio metódy

### (2) Metódy založené na homológii

Hľadáme homológy proteínu  $X$ , t.j. podobné proteíny

Štruktúra sa väčšinou evolučne mení pomalšie ako sekvencia  
ak niektorý homológ má známu štruktúru, aj  $X$  má asi podobnú

Určovanie štruktúry proteínov bolo dlho považované za otvorený problém,  
ktorý nevieme bioinformaticky riešiť, ak sa nedá použiť metóda (2)

## Určovanie štruktúry proteínov ako bioinformatický problém

(protein structure prediction, protein folding)

**Vstup:** sekvencia proteínu  $X$

**Výstup:** 3D pozície atómov alebo aminokyselín

**(1) Ab initio metódy**

**(2) Metódy založené na homológii**

**(3) Metódy založené na hlbokých neurónových sieťach**

Od roku 2018 veľký pokrok

Hlavne program AlphaFold od firmy DeepMind/Google

Nobelova cena za chémiu 2024: dvaja z autorov, Demis Hassabis a John Jumper  
(a David Baker za návrh nových proteínov)

## Najnovšie prístupy: hlboké neurónové siete

- Súťaž CASP raz za dva roky
- V roku 2018 a 2020 vyhral AlphaFold od firmy DeepMind/Google.  
V roku 2020 AlphaFold2 vyhral s veľkým náskokom.  
2/3 predpovedaných štruktúr mali vysokú presnosť.  
Využíva nové prvky, aj existujúce prístupy.  
V roku 2022 väčšina metód inšpirovaná AlphaFold2.
- Kľúčová myšlienka využitá aj pred AlphaFold-om: **detekcia ko-evolúcie**
  - k skladanému proteínu zarovnajú veľké množstvo homológov  
(aj bez známych štruktúr)
  - hľadajú dvojice pozícií, ktoré sa menia súčasne
  - takéto dvojice sú potenciálne v kontakte

## Najnovšie prístupy: hlboké neurónové siete

- **AlphaFold 1 (2018):**

- (1) Predikcia vzdialeností amino kyselín pomocou neurónovej siete
- (2) Hľadanie štruktúry, ktorá dobre sedí so vzdialenosťami a fyzikálnym modelom využitím štandardnej numerickej optimalizácie (gradientové metódy) [animácia]

- **AlphaFold 2 (2020):**

kombinuje oba kroky do jednej neurónovej siete, ktorá sa opakovane spúšťa na svojich výsledkoch

- **AlphaFold 3 (2024):**

Iná neurónová sieť, na finálnu úpravu štruktúry používa difúzne modely, ktoré sa používajú aj na generovanie obrázkov  
Umožňuje skladať aj komplexy kombinujúce viac proteínov, alebo proteín a inú molekulu (DNA, RNA, ióny, a pod.)

## Limitácie programu AlphaFold

Vyplývajú z dostupných dát pre tréning

- Nedá sa využiť na proteíny bez homológov (napr. umelo vytvorené alebo tie, ktoré rýchlo mutujú, napr. protilátky)
- Nie je úplne presný v predpovedaní vplyvu mutácie na štruktúru
- Predpovedá jednu štruktúru, ale veľa proteínov má viacero možných polôh
- Flexibilnejšie časti proteínov (disordered) sú často predpovedané s nízkou spoľahlivosťou (vyznačenou vo výsledkoch ako low confidence)
- AlphaFold3 nevie spracovať všetky typy molekúl viažúcich proteíny

## Praktické prístupy k určovaniu štruktúry proteínu

Pre daný proteín  $X$ :

- Pozrieme do PDB, či má  $X$  známu štruktúru
- V databázach môžeme nájsť aj štruktúru pre  $X$  od AlphaFold
- Môžeme spustiť AlphaFold na  $X$
- Môžeme hľadať homológy  $X$  so známou štruktúrou

## Hľadanie homológov proteínu

### Dôležité pre rôzne účely:

- určenie približnej štruktúry a funkcie proteínu
- štúdium evolúcie proteínu
- vstup pre AlphaFold

### Videli sme:

- dynamické programovanie
- heuristické zrýchlenia (BLAST a spol.)
- skórovacie matice (BLOSUM)

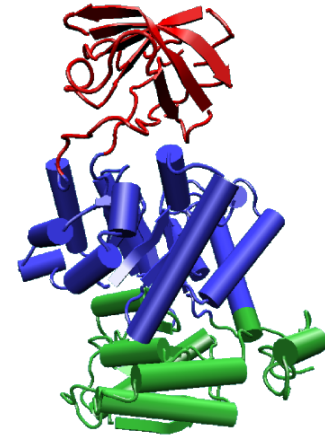
### Nevedia najst' vzdialenejšie homológy

Dnes si ukážeme prístupy založené na **pravdepodobnostných profiloch**

## Proteínové domény a rodiny

### Doména (domain)

- Časť proteínu s nezávislou štruktúrou
- Veľa proteínov sa skladá z viacerých domén
- Domény sa tiež v proteínoch preskupujú počas evolúcie



### Rodina (family)

- Skupina proteínov/domén s podobnou sekvenciou, štruktúrou, funkciou
- Ak poznáme štruktúru jedného člena rodiny, môžeme predpokladať, že ostatné majú podobnú

## Proteíny ako skladačka domén

### Databáza Pfam

Domény v proteínoch rozdelené do viac ako 20 tisíc rodín

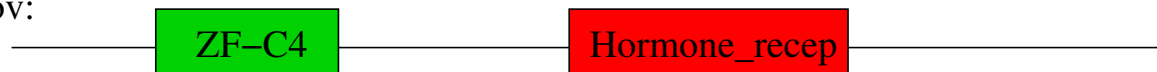
76% proteínov aspoň jedna známa doména

49% proteínových sekvencií pokrývajú známe domény

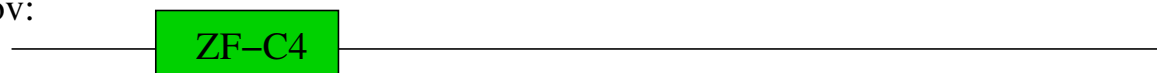
### Príklad:

4 z 654 architektúr obsahujúcich doménu Zinc finger, C4 type (Pfam)

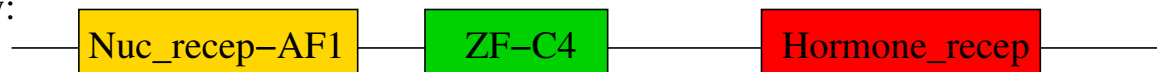
56171 proteínov:



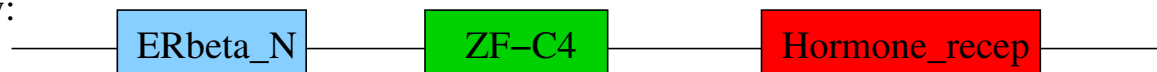
13525 proteínov:



3514 proteínov:



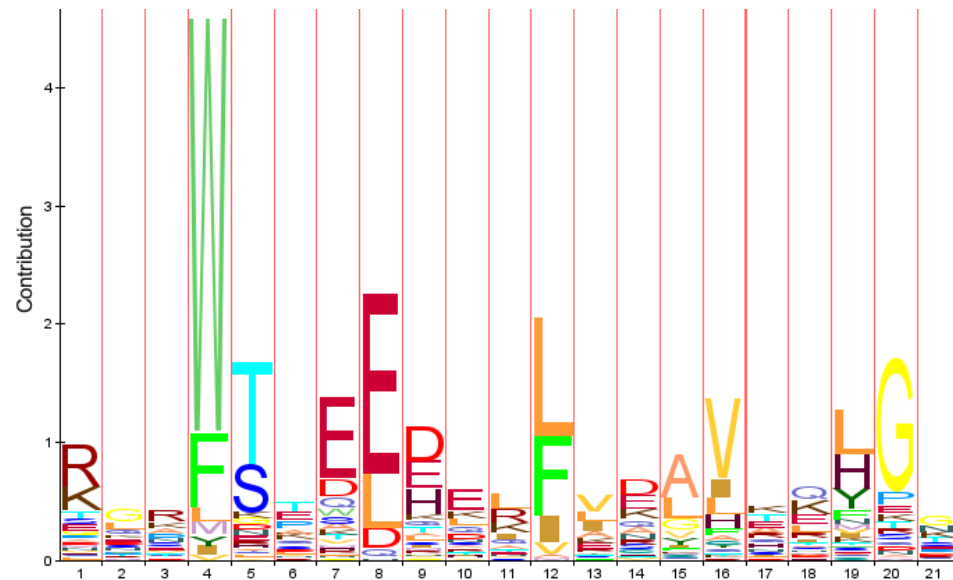
1574 proteínov:



## Charakterizácia rodín proteínov

- Zarovnania medzi známymi prvkami rodiny a novým proteínom nemusia nájsť vzdialených členov
- Viacnásobné zarovnanie rodiny ukáže dôležité evolučne zachované pozície

```
MEEW SASEANLFEEALEKY GKDF  
PDEWTVEDKVLFEQAFSFGKT.  
G TKWTA EENKKFENALAFYDKDT  
SKNWS EDDLQLLIKAVNLF PAGT  
EKP WSNQETLLLLLEAIETY GDD.  
AREWTDQETLLLLLE GLEMHKDD.  
KPE WSDKEILLLEAVMHY GDD.  
DDTWTAQELVLLSEGVEMYS...  
KKNWSDQEMLLLLLEGIEMYE...  
DENWSKEDLQKLLKGIQEF GAD.  
EDDWSQAEQKAFETALQKYPKGT  
EEAWTQSQQKLLLELALQQYPKGA  
EDVWSATEQKTLEDAIKKHKSSD  
AMSWTHEDEFELLKAAHKFKMG.
```



## Pravdepodobnostný profil rodiny

(profile, position specific score matrix PSSM)

- V zarovnaní spočítaj  $e_i(x)$ : frekvencia výskytu písmena  $x$  v stĺpci  $i$
- Dostaneme model, ktorý generuje sekvenciu  $x_1, x_2, \dots, x_n$  s pravdepodobnosťou

$$e_1(x_1) \cdot e_2(x_2) \cdots e_n(x_n)$$

- Nulová hypotéza: sekvencia bola vygenerovaná náhodne, kde písmeno  $x$  má frekvenciu  $q(x)$
- Skóre sekvencie  $x_1, \dots, x_n$ :  
logaritmus pomeru pravdepodobností v dvoch modeloch

$$\log \frac{e_1(x_1) \cdots e_n(x_n)}{q(x_1) \cdots q(x_n)}$$

(neskôr rozpíšeme na súčet dielčích skóre pre aminokyseliny)

## Hračársky príklad PSSM

- Uvažujme len leucín L a alanín A
- Majme zarovnanie 10 sekvencií s počtami / frekvenciami  $e_i(x)$  v tabuľke

|   | počty |   |   |   |  | frekvencie |     |     |     |
|---|-------|---|---|---|--|------------|-----|-----|-----|
|   | 1     | 2 | 3 | 4 |  | 1          | 2   | 3   | 4   |
| A | 2     | 6 | 9 | 1 |  | 0,2        | 0,6 | 0,9 | 0,1 |
| L | 8     | 4 | 1 | 9 |  | 0,8        | 0,4 | 0,1 | 0,9 |

- Nulová hypotéza  $q(A) = 0,3, q(L) = 0,7$
- Pravdepodobnosť sekvencie LAAL
  - v profile  $0,8 \cdot 0,6 \cdot 0,9 \cdot 0,9 = 0,3888$ ,
  - v nulovom modeli  $0,7 \cdot 0,3 \cdot 0,3 \cdot 0,7 = 0,0441$
- Skóre LAAL:  $\log_2(0,3888/0,0441) = 3,14$   
Skóre LALA:  $\log_2(0,0048/0,0441) = -3,20$

## Pravdepodobnostný profil rodiny

- $e_i(x)$ : frekvencia výskytu písmena  $x$  v stĺpci  $i$  zarovnania rodiny
- $q(x)$ : frekvencia výskytu písmena  $x$  v nulovom modeli
- $s_i(x) = \log \frac{e_i(x)}{q(x)}$  skóre písmena  $x$  v stĺpci  $i$  zarovnania rodiny
- Skóre sekvencie  $x_1, \dots, x_n$ :

logaritmus pomeru pravdepodobností v dvoch modeloch

$$\begin{aligned} & \log \frac{e_1(x_1) \cdot \dots \cdot e_n(x_n)}{q(x_1) \cdot \dots \cdot q(x_n)} \\ &= \log \left( \frac{e_1(x_1)}{q(x_1)} \cdot \dots \cdot \frac{e_n(x_n)}{q(x_n)} \right) \\ &= \log \frac{e_1(x_1)}{q(x_1)} + \dots + \log \frac{e_n(x_n)}{q(x_n)} \\ &= s_1(x_1) + \dots + s_n(x_n) \end{aligned}$$

## Hračársky príklad PSSM

- Majme zarovnanie 10 sekvencií s počtami / frekvenciami  $e_i(x)$  v tabuľke

|   | počty |   |   |   |  | frekvencie |     |     |     |
|---|-------|---|---|---|--|------------|-----|-----|-----|
|   | 1     | 2 | 3 | 4 |  | 1          | 2   | 3   | 4   |
| A | 2     | 6 | 9 | 1 |  | 0,2        | 0,6 | 0,9 | 0,1 |
| L | 8     | 4 | 1 | 9 |  | 0,8        | 0,4 | 0,1 | 0,9 |

- Nulová hypotéza  $q(A) = 0,3, q(L) = 0,7$
- Skóre alanínu v prvom stĺpci  $s_1(A) = \log_2(0,2/0,3) = -0,58$   
skóre leucínu v prvom stĺpci  $s_1(L) = \log_2(0,8/0,7) = 0,19$
- Dostávame tabuľku skór

|   | 1     | 2     | 3     | 4     |
|---|-------|-------|-------|-------|
| A | -0,58 | 1,00  | 1,58  | -1,58 |
| L | 0,19  | -0,81 | -2,81 | 0,36  |

- Skóre LAAL je  $0,19 + 1 + 1,58 + 0,36 = 3,13$   
Skóre LALA je  $0,19 + 1 - 2,81 - 1,58 = -3,2$

## Pseudocounts

Ak na niektorej pozícii určitá amino kyselina nebola pozorovaná, mala by v modeli pravdepodobnosť 0

|   | 1 | 2 | 3 | 4  |
|---|---|---|---|----|
| A | 2 | 6 | 9 | 0  |
| L | 8 | 4 | 1 | 10 |

Aby sme sa vyhli tomuto problému, pridáme ku každému políčku najskôr nejakú malú hodnotu, **pseudocount**, napr. 0,5:

|   | 1   | 2   | 3   | 4    |
|---|-----|-----|-----|------|
| A | 2,5 | 6,5 | 9,5 | 0,5  |
| L | 8,5 | 4,5 | 1,5 | 10,5 |

Potom postupujeme ako predtým

**Problém:** ako vyriešiť inzercie a delécie?

## Ako inzercie a delécie?

Profil môžeme zarovnať k sekvencii dynamickým programovaním  
ako pre globálne alebo lokálne zarovnanie

Namiesto skóre zhody a nezhody použijeme čísla z profilu

Pridáme fixné penalty za medzery

Niektoré pozície však majú medzier viac, to nevieme zachytiť

|   | 0 | 1     | 2     | 3     | 4     |
|---|---|-------|-------|-------|-------|
| A |   | -0,58 | 1,00  | 1,58  | -1,58 |
| L |   | 0,19  | -0,81 | -2,81 | 0,36  |

0

1 A

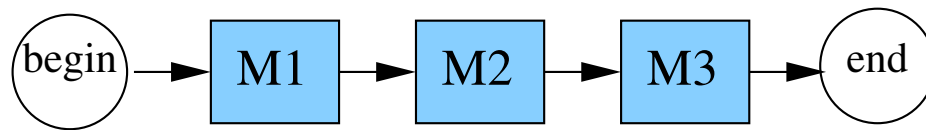
2 L

3 L

## Profilové HMM

Rozšíříme profil o inzercie a delécie

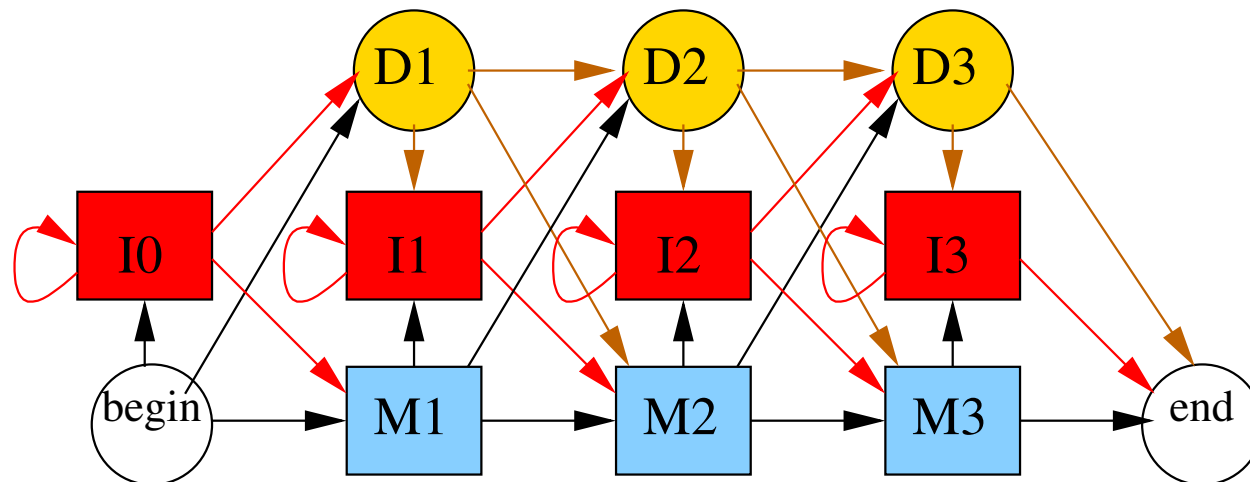
### PSSM profil ako HMM:



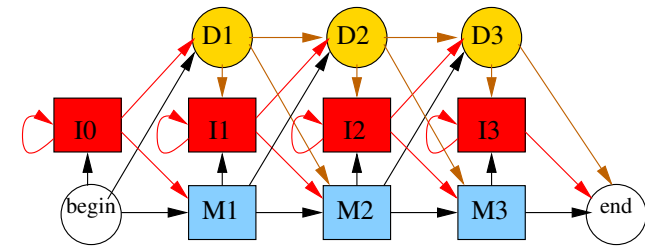
**Profilové HMM:** match state, insert state, delete state

krúžky sú tiché stavy, nič negenerujú

začínáme vždy v begin, končíme v end



## Konštrukcia profilového HMM



- Začneme z viacnásobného zarovnania
- Stĺpcom s málo medzerami priradíme match stavy, ostatné budú v insert stavoch
- V každom stĺpci zrátame  $E_i(a)$ : počet výskytov  $a$
- Pravdepodobnosť emisie  $e_i(a) = \frac{E_i(a)}{\sum_b E_i(b)}$
- Pridáme “pseudocounts”, aby sme nemali nulové položky
$$e_i(a) = \frac{E_i(a) + c}{\sum_b (E_i(b) + c)}$$
- Pravdepodobnosti prechodu nastavíme podľa medzier v zarovnaní
- Veľmi podobné sekvencie môžeme použiť s menšou váhou

## Použitie profilov a profilových HMM

### Odkiaľ vziať profily/profilové HMM?

- Databáza Pfam: rodiny domén reprezentované ako profilové HMM
- PSI-Blast: PSSM iteratívne zo skupiny podobných proteínov
- PSSM sa používajú aj na reprezentáciu motívov v DNA  
(napr. väzobné miesta transkripčných faktorov)

### Nájsť výskyty profilu v proteínovej sekvencii

- Podobné problému lokálneho zarovnania
- PSSM profily: dynamické programovanie, penalta za medzery
- Profilové HMM: Viterbiho alebo dopredný algoritmus (mierne modifikovaný)

Výsledné skóre alebo pravdepodobnosť sa použije na rozhodnutie, či proteín patrí do rodiny

## Štruktúra proteínov, zhrnutie

(protein structure prediction, protein folding)

**Vstup:** sekvencia proteínu  $X$

**Výstup:** 3D pozície atómov alebo aminokyselín

- (1) Ab initio metódy
- (2) Metódy založené na homológii
- (3) Metódy založené na hlbokých neurónových sieťach

### Praktické prístupy k určovaniu štruktúry proteínu $X$

- Pozrieme do PDB, či má  $X$  známu štruktúru
- V databázach môžeme nájsť aj štruktúru pre  $X$  od AlphaFold
- Môžeme spustiť AlphaFold na  $X$
- Môžeme hľadať homológy  $X$  so známou štruktúrou resp. domény v  $X$  pomocou profilov

## Využitie proteínových štruktúr

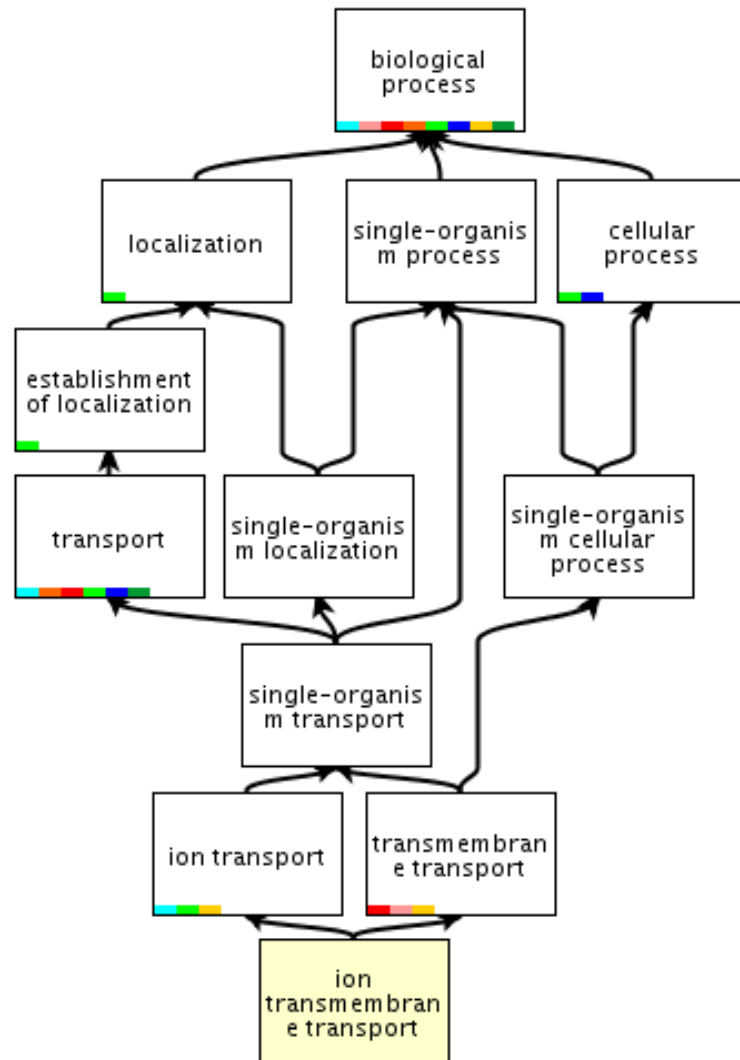
- Presnejšie definovanie domén v databázach ako Pfam
- Skúmanie efektu mutácií na štruktúru / funkciu
- Modelovanie interakcií medzi proteínmi, proteínových komplexov
- Objavovanie nových liečiv, ktoré sa budú viazať na určitý proteín
- Dizajn umelých proteínov s vhodnými vlastnosťami

## Funkcia proteínu

- Pre niektoré proteíny určená laboratórne
- Na ďalšie proteíny prenášame bioinformaticky pomocou podobnosti sekvencie, prítomnosti domén, polohy v genóme a ďalších dát
- Swissprot/Uniprot zhromažďuje údaje o funkcii proteínov
- Klasifikácia proteínov pomocou Gene ontology (GO)  
Príklad pojmu v GO:  
Accession: GO:0034220  
Name: ion transmembrane transport  
Ontology: biological\_process  
Definition: A process in which an ion is transported from one side of a membrane to the other by means of some agent such as a transporter or pore.  
Comment: Note that this term is not intended for use in annotating lateral movement within membranes.
- Príklad použitia: nadreprezentácia určitej funkcie medzi génmi s pozitívnym výberom alebo zmenenou expresiou

## Gene ontology (GO)

Hierarchická štruktúra pojmov:



## Učenie s učiteľom

### Príklad úlohy:

pre každú amino kyselinu vo vstupnej sekvencii proteínu určte jednu z 8 známych sekundárnych štruktúr

$\alpha$ -helix, 3-helix,  $\pi$ -helix,  $\beta$ -strand,  $\beta$ -bridge,  $\beta$ -turn, bend, random coil

**Čo by sme potrebovali** od zadávateľa a čo by sme museli urobiť, aby sme takúto úlohu vyriešili strojovým učením?

## Učenie s učiteľom

### Príklad úlohy:

pre každú amino kyselinu vo vstupnej sekvencii proteínu určte jednu z 8 známych sekundárnych štruktúr

$\alpha$ -helix, 3-helix,  $\pi$ -helix,  $\beta$ -strand,  $\beta$ -bridge,  $\beta$ -turn, bend, random coil

### Potrebuje dostat':

dáta, t.j. príklady proteínov so známou správnou odpoveďou

### Potrebuje spraviť: (trochu zjednodušené)

- rozdeliť dáta na trénovacie a testovacie
- vybrať konkrétny typ modelu strojového učenia, napr. typ neurónovej siete
- zapísať vstup ako postupnosť čísel pevnej dĺžky
- spustiť trénovací algoritmus na trénovacích dátach
  - aby sa naučil parametre, ktoré dobre fungujú
- vyhodnotiť úspešnosť modelu na testovacích dátach

## Problém 1

### Zapísať vstup ako postupnosť čísel pevnej dĺžky

Príklad riešenia: zoberieme okno sekvencie určitej dĺžky, napr. 100 aminokyselín

Každú aminokyselinu zapíšeme ako 20 núl/jedničiek

One-hot encoding: 19 núl a 1 jednotka, ktorej poloha určuje o ktorú kyselinu ide

Pomerne dlhý a málo informatívny zápis

## Problém 2

**Vybrať konkrétny typ modelu strojového učenia, napr. typ neurónovej siete**

Zložitejšie modely často majú lepšie výsledky, ale potrebujú viac trénovacích dát

Trénovacie dáta so známou odpoveďou je často ťažké získať

## Vyriešenie obidvoch problémov: proteínové jazykové modely

Podobné na veľké jazykové modely prirodzeného jazyka, ako ChatGPT

Máme síce málo trénovacích dát so známou odpoveďou na konkrétnu otázku, ale celkovo máme veľa známych proteínov

Jazykový model je veľká neurónová sieť, ktorá sa na všetkých známych proteínoch učí riešiť umelú úlohu:

Na vstupe zmažeme niektoré aminokyseliny v proteíne a má uhádnuť, aké boli

Použijeme veľa dát a veľkú výpočtovú kapacitu na tréning

Ako medzivýsledok má táto sieť zakódovanie každej aminokyseliny vo vstupnej sekvencii pomocou postupnosti čísel (**vnorenie, embedding**)

Toto vnorenie závisí aj od ostatných aminokyselín na vstupe

Vyjadruje vzory, ktoré sa sieť naučila zo všetkých proteínov

## Využitie proteínového jazykového modelu na iné úlohy

Spustíme natrénovaný jazykový model

Dá nám vnorenie každej aminokyseliny

Tieto vnorenia použijeme ako vstup pre jednoduchší model (malú sieť, logistickú regresiu a pod.) ktorý trénujeme na našich menších dátach

## **Príklad proteínového jazykového modelu: ESM-2, Rives et al. 2020**

Jazykový model natrénovaný na 250 miliónoch proteínov

### **Použitie vnorenia na určenie sekundárnej štruktúry:**

Jednoduchý model (logistická regresia):

Profil podobných sekvencií 47.9% presnosť

Vnorenie jazykového modelu 71.1% presnosť

Zložitý model (hlboká neurónová sieť), iné dáta:

Profil podobných sekvencií 71.2% presnosť

Vnorenie jazykového modelu 71.6% presnosť

Vnorenie veľmi pomôže jednoduchému modelu

zložitému modelu nahradí dáta z príbuzných proteínov (profil)

lebo jazykový model dostane pri použití len jeden proteín