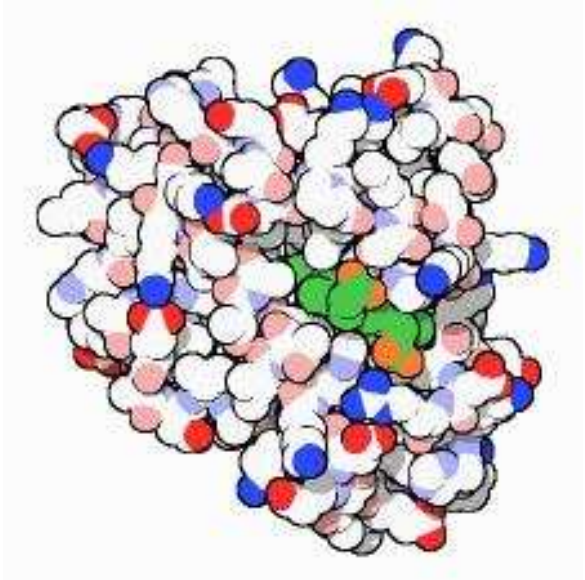
Štruktúra a funkcia proteínov

Tomáš Vinař

29.11.2018



Proteíny

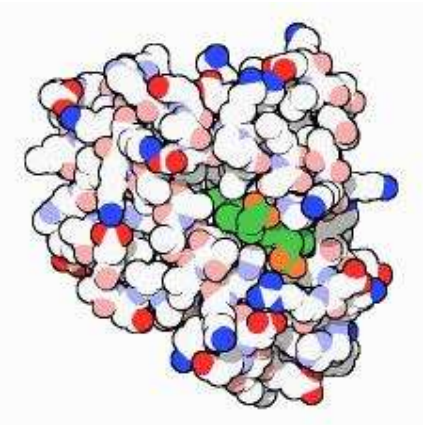
Reťazce 20 rôznych aminokyselín s rôznymi chemickými vlastnosťami:

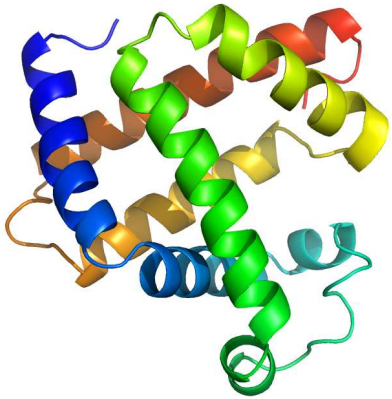
Aminokyselina	Postranný reťazec	Jeho vlastnosti
Alanín (A)	-CH3	hydrofóbny
Arginín (R)	-(CH2)3NH-C(NH)NH2	bázický
Asparagín (N)	-CH2CONH2	hydrofilný
Kyselina asparágová (D)	-CH2COOH	kyslý
Cysteín (C)	-CH2SH	hydrofóbny
Kyselina glutámová (E)	-CH2CH2COOH	kyslý
Glutamín (Q)	-CH2CH2CONH2	hydrofilný
Glycín (G)	-H	hydrofilný
Histidín (H)	-CH2-C3H3N2	bázický
Izoleucín (I)	-CH(CH3)CH2CH3	hydrofóbny
Leucín (L)	-CH2CH(CH3)2	hydrofóbny
Lyzín (K)	-(CH2)4NH2	bázický
Metionín (M)	-CH2CH2SCH3	hydrofóbny
Fenylalanín (F)	-CH2C6H5	hydrofóbny
Prolín (P)	-CH2CH2CH2-	hydrofóbny
Serín (S)	-CH2OH	hydrofilný
Treonín (T)	-CH(OH)CH3	hydrofilný
Tryptofán (W)	-CH2C8H6N	hydrofóbny
Tyrozín (Y)	-CH2-C6H4OH	hydrofóbny
Valín (V)	-CH(CH3)2	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 crystalography)
  vyžaduje proteín v kryštalickej forme
- NMR (nuclear magnetic resonance spectroscopy)
  hlavne používaná na kratšie proteíny
- Náročný a drahý proces
- Databáza štruktúr PDB
  146 000 proteínových štruktúr
  (UniProt má 134 miliónov sekvencií)

Určovanie štruktúry proteínov ako bioinformatický problém

(protein structure prediction, protein folding)

Vstup: sekvencia proteínu
Výstup: 3D pozície atómov alebo aminokyselín

Ab initio metódy

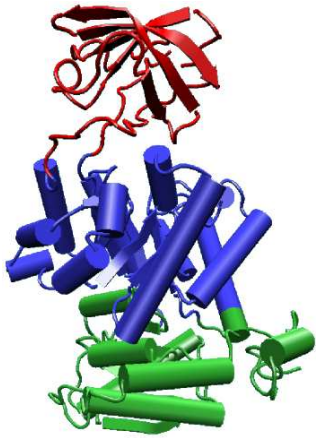
- Nájdi štruktúru s najnižšou voľnou energiou
- Vzorce na približný výpočet energie založené na fyzike
  — sily medzi atómami v proteíne a okolitom roztoku
- Štatistické vzorce merajúce typické vzialenosti medzi aminokyselinami na známych štruktúrach
- V oboch prípadoch veľmi ťažký výpočtový problém
  — simulácia molekulárnej dynamiky
  — optimalizačné metódy, napr. simulované žíhanie
- Používané na malé proteíny a zlepšenie približných štruktúr

Protein threading

- Aj proteíny s pomerne odlišnou sekvenciou môžu mať podobnú štruktúru
- Môžeme skúsiť "napasovať" proteín na každú známu štruktúru
- Určitý typ zarovnania, ale pri skórovaní uvažujeme aj interakcie medzi amino kyselinami blízko v štruktúre
- Výpočtovo ťažký problém

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

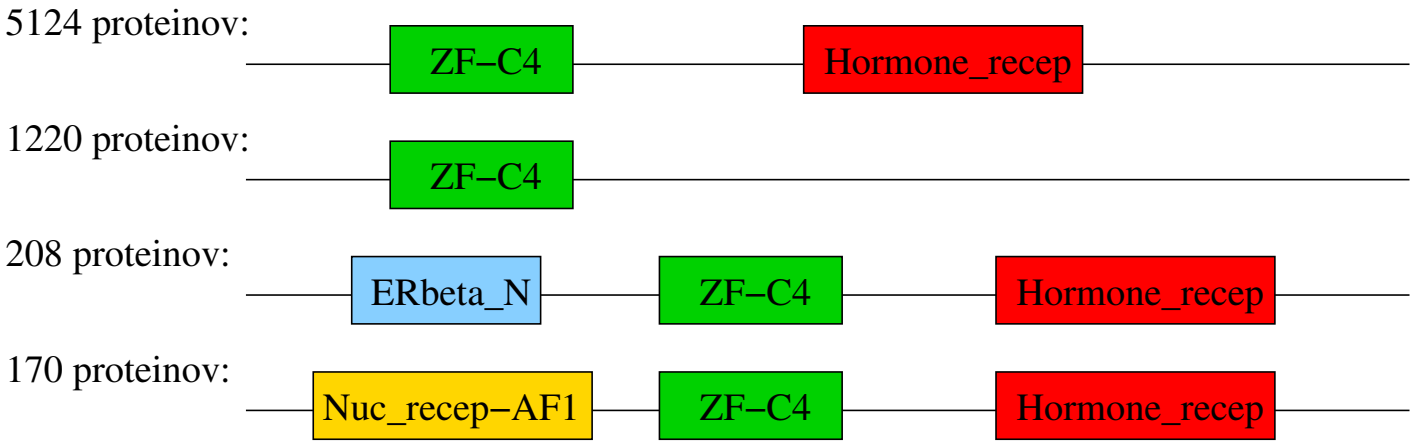
Domény v proteínoch rozdelené do rodín

80% proteínov aspoň jedna známa doména

58% proteínových sekvencií pokrývajú známe domény

Príklad:

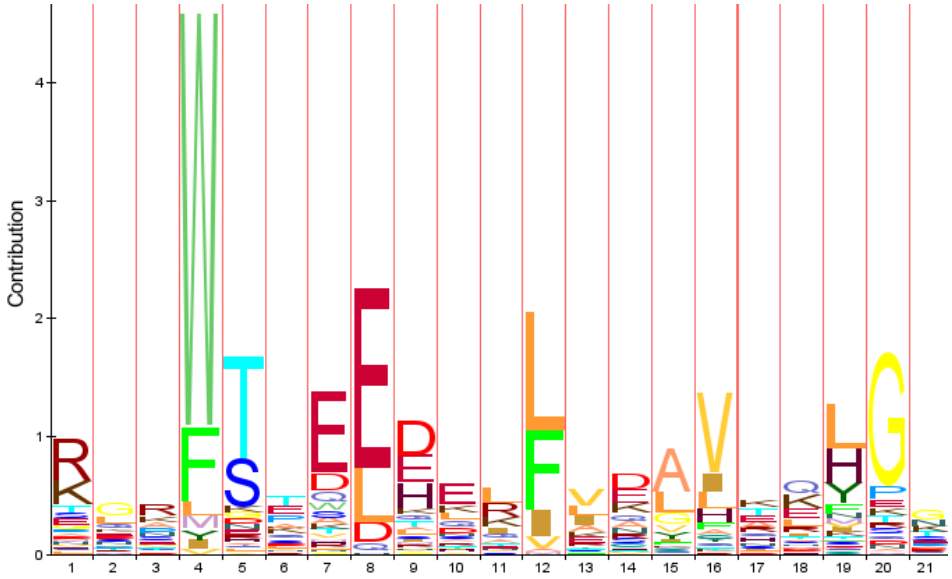
Rôzne architektúry obsahujúce doménu Zinc finger, C4 type (Pfam)



Charakterizácia rodín proteínov

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

```
MEEWSASEANLFEEALEKYGKDF
PDEWTVEDKVLFEQAFSFHGKT.
GTKWTAEENKKFENALAFYDKDT
SKNWSEDDLQLLIKAVNLFPAGT
EKPWSNQETLLLLEAIETYGDD.
AREWTDQETLLLLEGLEMHKDD.
KPEWSDKEILLLLEAVMHYGDD.
DDTWTAQELVLLSEGVEMYS...
KKNWSDQEMLLLLEGIEMYE...
DENWSKEDLQKLLKGIQEFGAD.
EDDWSQAEQKAFETALQKYPKGT
EEAWTQSQQKLLELALQQYPKGA
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 $x_1, x_2, \ldots, x_n$ 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: logaritmus pomeru pravdepodobností v dvoch modeloch

$$\log \frac{\prod_{i=1}^{n} e_i(x_i)}{\prod_{i=1}^{n} q(x_i)} = \sum_{i=1}^{n} \log \frac{e_i(x_i)}{q(x_i)} = \sum_{i=1}^{n} s_i(x_i)$$

Hračkársky príklad PSSM

- Uvažujme len leucín L a alanín A
- Majme zarovananie 10 sekvencií s nasledujúcimi počtami

	1	2	3	4
A	2	6	9	1
L	8	4	1	9

- Nulová hypotéza $q(A) = 30\%$, $q(L) = 70\%$
- Sekvencia LAAL má v profile pravdepodobnosť
  $0.8 \cdot 0.6 \cdot 0.9 \cdot 0.9 = 0.3888$,
  v nulovom modeli $0.7 \cdot 0.3 \cdot 0.3 \cdot 0.7 = 0.0441$
- Skóre $\log_2(0.3888/0.0441) = 3.14$

Hračkársky príklad PSSM

- Majme zarovananie 10 sekvencií s nasledujúcimi počtami

	1	2	3	4
A	2	6	9	1
L	8	4	1	9

- Nulová hypotéza $q(A) = 30\%$, $q(L) = 70\%$
- 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 ALAL je $-0.58 - 0.81 + 1.58 + 0.36 = 0.55$

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

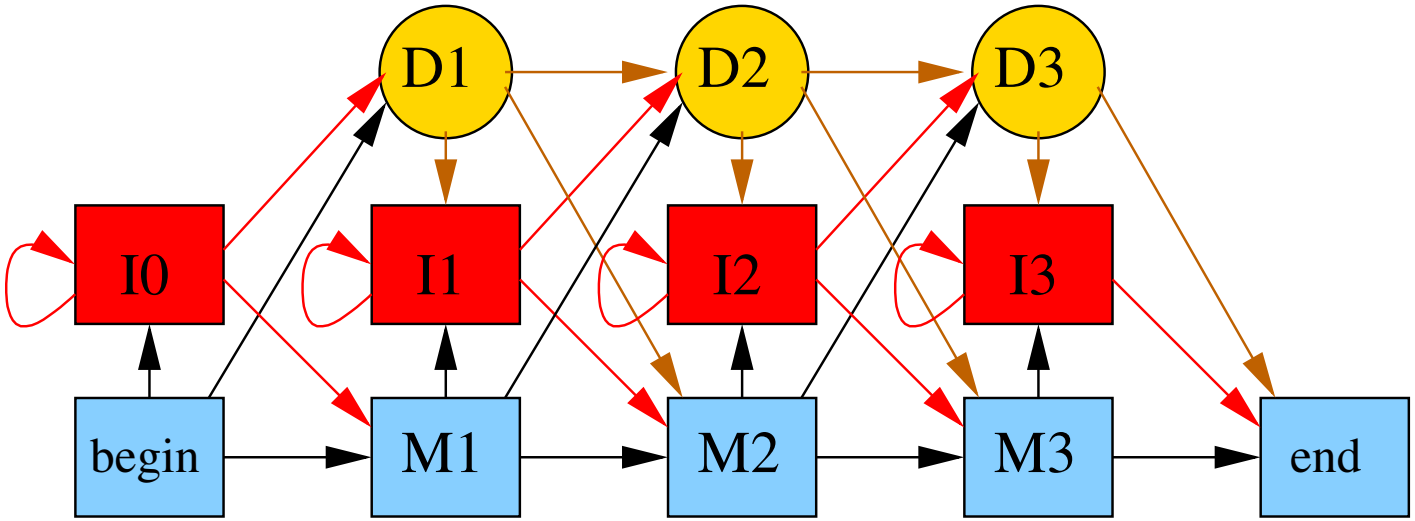
Potom postupujeme ako predtým

Profilové HMM

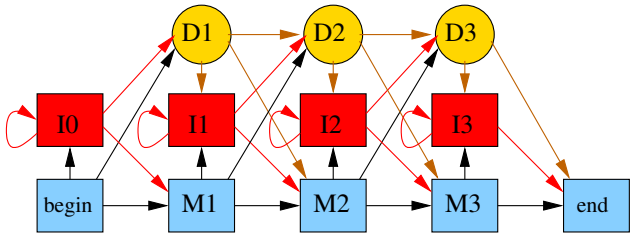
Rozšír profil o inzercie a delécie

PSSM profil ako HMM:

Profilové HMM: match state, insert state, delete state



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?

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

Nájdi výskyty profilu v proteínovej sekvencii

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

Patrí proteín do rodiny?

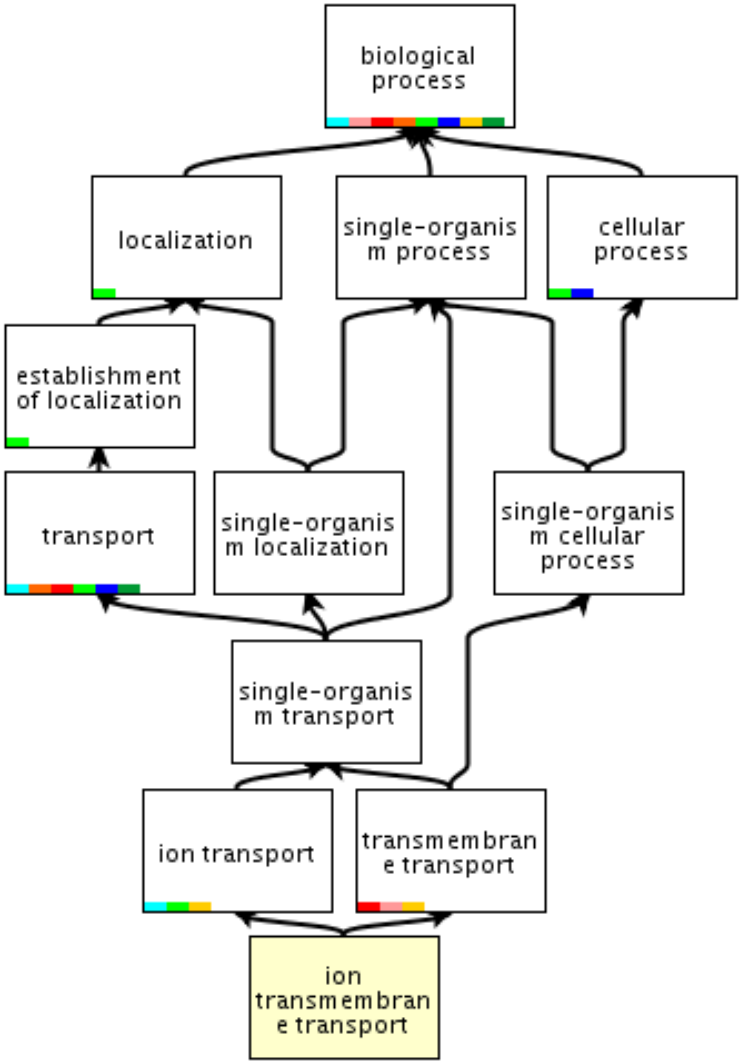
- PSSM profily: vysoko kladné skóre
- Profilové HMM: pravdepodobnosť vygenerovania sekvencie, normalizovaná na dĺžku

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.

Gene ontology (GO)

Hierarchická štruktúra pojmov:



Zhrnutie: akú štruktúru má proteín?

- Pozriem do PDB, či má známu štruktúru
- Ak nie, skúsim BLAST voči proteínom so známou štruktúrou
- Ak nič, skúsim hľadať domény so známou štruktúrou
- Ak nič, skúsim protein threading
- Pre krátke proteíny môžem skúsiť minimalizovať energiu,
  inak získané štruktúry doplniť/vylepšiť minimalizáciou energie

Minimalizácia energie je výpočtovo veľmi náročná
Súťaž CASP raz za dva roky

Zaujímavosti: Folding@home, Foldit