

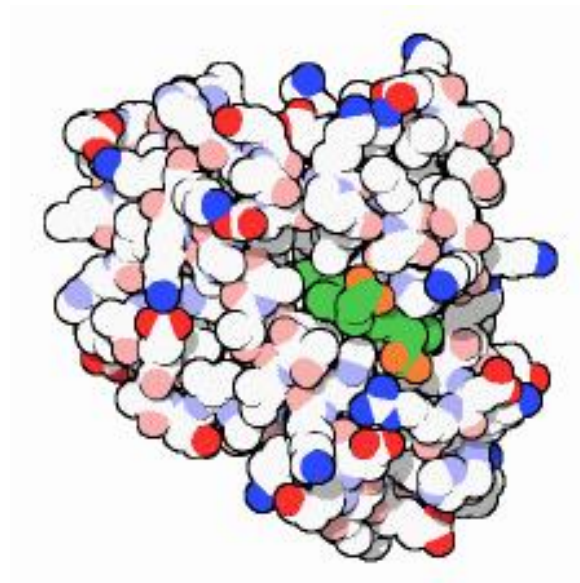
Oznamy

- DÚ 2 je na stránke, odovzdať do 29.11.
- Stretnutia journal clubu do stredy 22.11.
- Budúci štvrtok 23.11. nebudú prednáška ani cvičenia
Dáme vám video verziu
Dôležité: na cvičeniach úvod do bezkontextových gramatík,
pozrite si určite pred prednáškou 30.11., budete ich potrebovať
Obidva kvízy zverejnené dnes, termín o dva týždne
- Termíny na konci semestra
– DÚ3 streda 13.12., správy zo journal clubu piatok 15.12.

Štruktúra a funkcia proteínov

Broňa Brejová

15.11.2023



Proteíny

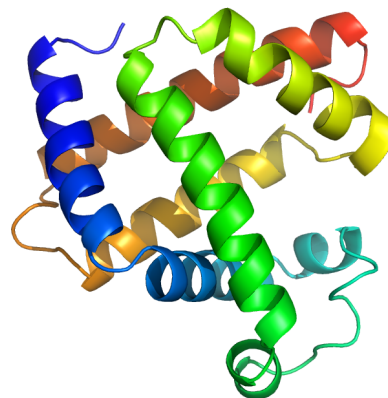
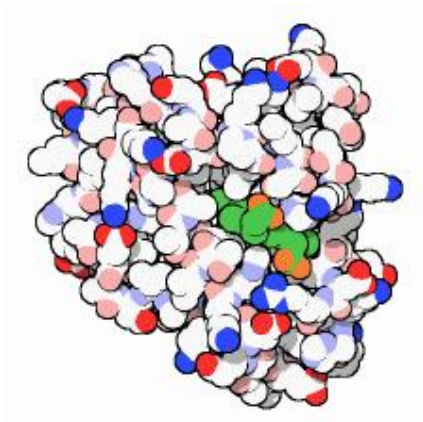
Reťazce 20 rôznych aminokyselín s rôznymi chemickými vlastnosťami:

Aminokyselina	Postranný reťazec	Jeho vlastnosti
Alanín (A)	-CH ₃	hydrofóbny
Arginín (R)	-(CH ₂) ₃ NH-C(NH)NH ₂	bázický
Asparagín (N)	-CH ₂ CONH ₂	hydrofilný
Kyselina asparágová (D)	-CH ₂ COOH	kyslý
Cysteín (C)	-CH ₂ SH	hydrofóbny
Kyselina glutámová (E)	-CH ₂ CH ₂ COOH	kyslý
Glutamín (Q)	-CH ₂ CH ₂ CONH ₂	hydrofilný
Glycín (G)	-H	hydrofilný
Histidín (H)	-CH ₂ -C ₃ H ₃ N ₂	bázický
Izoleucín (I)	-CH(CH ₃)CH ₂ CH ₃	hydrofóbny
Leucín (L)	-CH ₂ CH(CH ₃) ₂	hydrofóbny
Lyzín (K)	-(CH ₂) ₄ NH ₂	bázický
Metionín (M)	-CH ₂ CH ₂ SCH ₃	hydrofóbny
Fenylalanín (F)	-CH ₂ C ₆ H ₅	hydrofóbny
Prolín (P)	-CH ₂ CH ₂ CH ₂ -	hydrofóbny
Serín (S)	-CH ₂ OH	hydrofilný
Treonín (T)	-CH(OH)CH ₃	hydrofilný
Tryptofán (W)	-CH ₂ C ₈ H ₆ N	hydrofóbny
Tyrozín (Y)	-CH ₂ -C ₆ H ₄ OH	hydrofóbny
Valín (V)	-CH(CH ₃) 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 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
212 000 proteínových štruktúr (85% X-ray)
(UniProt má 250 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

- 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

Praktické prístupy k určovaniu štruktúry proteínu

Pre daný proteín X :

- Pozrieme do PDB, či má X **známu štruktúru**
- Ak nič, hľadáme **proteín podobný na X** v PDB (BLAST)
 X má asi podobnú štruktúru (**homology modelling**)
- Ak nič, hľadáme vzdialenejšie homológy senzitivnejšími prístupmi, cez **profily** (táto prednáška)
- Ešte vzdialenejšie homológy sa dajú hľadať technikou **protein threading**
- V posledných rokoch veľké pokroky v predpovedaní štruktúry pomocou **hlbokých neurónových sietí**, ktoré využívajú veľké počty nájdených homológov
- **Predpovedané štruktúry** sa tiež dajú nájsť v databázach

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

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 zarovnaj veľké množstvo homológov
(aj bez známych štruktúr)
 - hľadaj 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

- Nedá sa využiť na proteíny bez homológov (napr. umelo vytvorené)

Praktické prístupy k určovaniu štruktúry proteínu

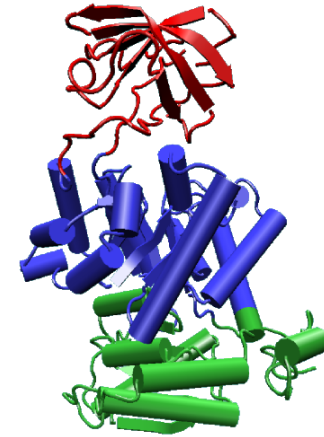
Pre daný proteín X :

- Pozrieme do PDB, či má X **známu štruktúru**
- Ak nič, hľadáme **proteín podobný na X** v PDB (BLAST)
 X má asi podobnú štruktúru (**homology modelling**)
- Ak nič, hľadáme vzdialenejšie homológy senzitivnejšími prístupmi, cez **profily** (táto prednáška)
- Ešte vzdialenejšie homológy sa dajú hľadať technikou **protein threading**
- V posledných rokoch veľké pokroky v predpovedaní štruktúry pomocou **hlbokých neurónových sietí**, ktoré využívajú veľké počty nájdených homológov
- **Predpovedané štruktúry** sa tiež dajú nájsť v databázach

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

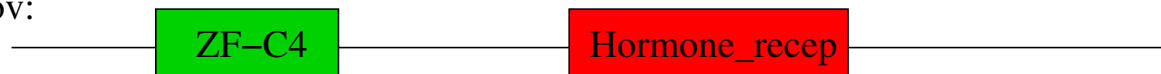
76% proteínov aspoň jedna známa doména

49% proteínových sekvencií pokrývajú známe domény

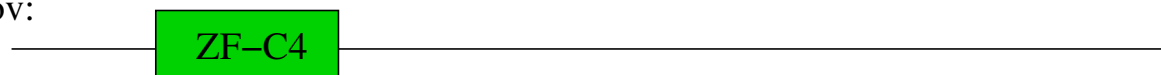
Príklad:

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

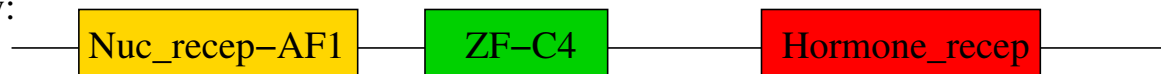
56171 proteínov:



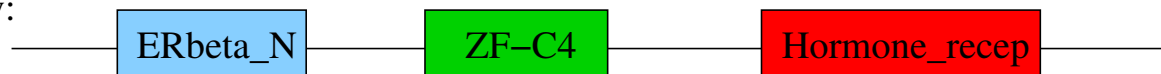
13525 proteínov:



3514 proteínov:



1574 proteínov:

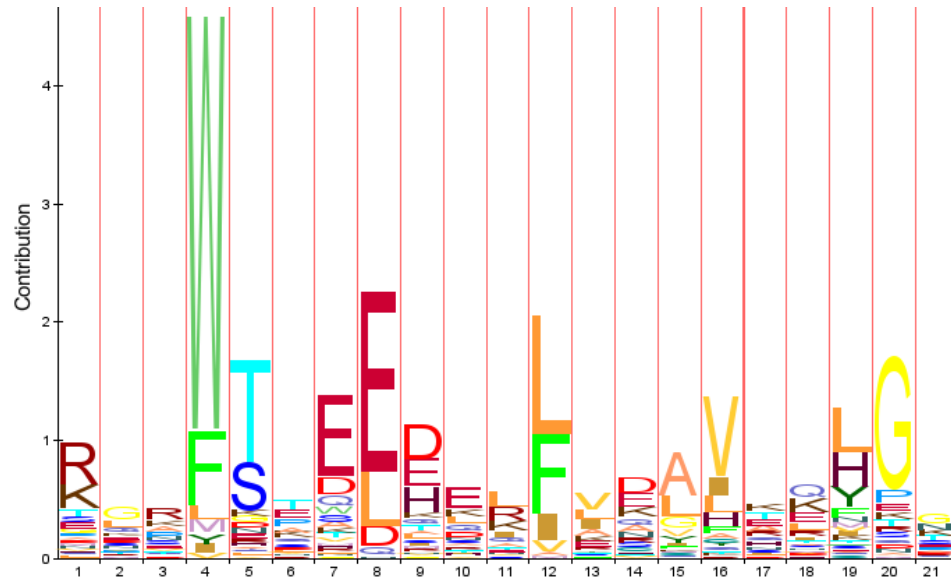


Charakterizácia rodín proteínov

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

```

MEEWSASEANLFEEALEKYGKDF
PDEWTVEDKVLFEQAFSFGKT.
GTKWTAENKKFENALAFYDKDT
SKNWSEDDLQLLIKAVNLFAGT
EKPWSNQETLLLLLEAIETYGDD.
AREWTDQETLLLLLEGLEMHKDD.
KPEWSDKEILLLEAVMHYGGDD.
DDTWTAQELVLLSEGVEMYS...
KKNWSDQEMLLLLLEGIEMYE...
DENWSKEDLQKLLKGIQEFGAD.
EDDWSQAEQKAFETALQKYPKGT
EEAWTQSQQKLELALQQYPKGA
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

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

- 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

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

(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
- $s_i(x) = \log \frac{e_i(x_i)}{q(x_i)}$ 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

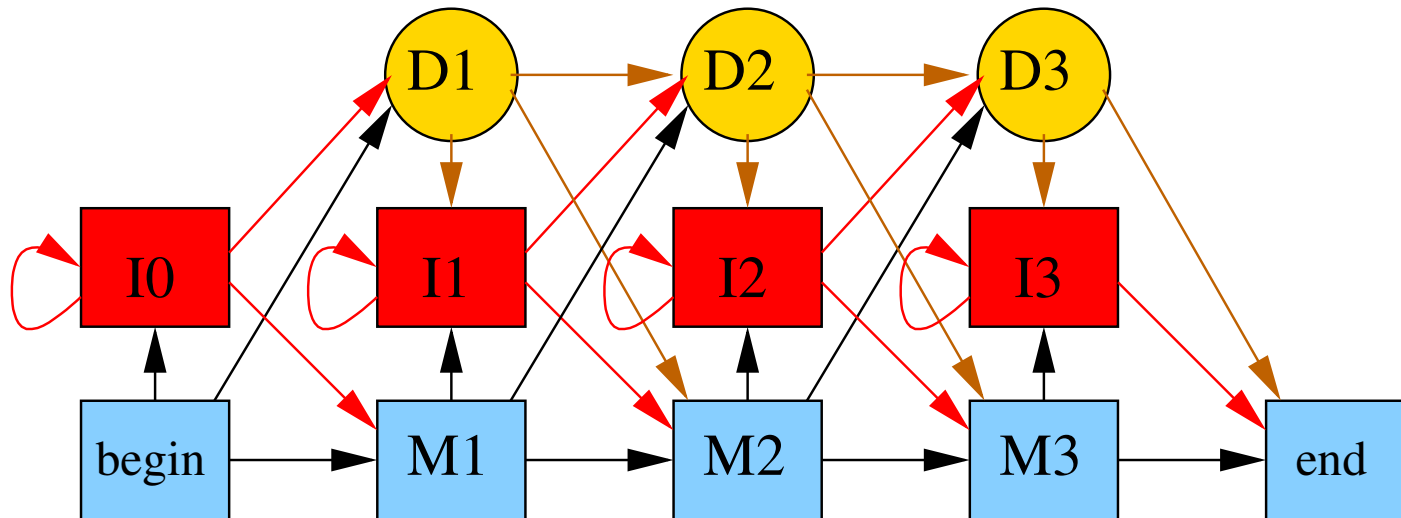
Profilové HMM

Rozšíříme profil o inzercie a delécie

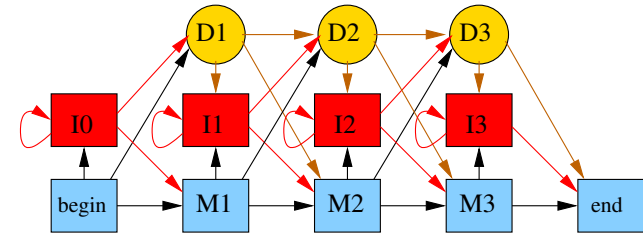
PSSM profil ako HMM:



Profilové HMM: match state, insert state, delete state



Konštrukcia profilového HMM



- Začneme z viacnásobného zarovnaní
- 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)

Nájdí výskyty profilu v proteínovej sekvencii

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

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

Praktické prístupy k určovaniu štruktúry proteínu

Pre daný proteín X :

- Pozrieme do PDB, či má X **známu štruktúru**
- Ak nič, hľadáme **proteín podobný na X** v PDB (BLAST)
 X má asi podobnú štruktúru (**homology modelling**)
- Ak nič, hľadáme vzdialenejšie homológy senzitivnejšími prístupmi, cez **profily** (táto prednáška)
- Ešte vzdialenejšie homológy sa dajú hľadať technikou **protein threading**
- V posledných rokoch veľké pokroky v predpovedaní štruktúry pomocou **hlbokých neurónových sietí**, ktoré využívajú veľké počty nájdených homológov
- **Predpovedané štruktúry** sa tiež dajú nájsť v databázach

Využitie proteínových štruktúr

- Presnejšie definovanie domén v databázach ako Pfam
- Skúmanie efektu mutácií na štruktúru
- 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.

Ďalšie použitia HMM a profilov na proteíny

- Určovanie sekundárnej štruktúry
- Určovanie transmembránových proteínov a signálnych peptidov
- Určovanie funkčných motívov a posttranslačných modifikácií (databáza PROSITE)

Cyclic nucleotide-binding domain signature 1:

[LIVM] - [VIC] -x- {H} -G- [DENQTA] -x- [GAC] -{L} -x- [LIVMFY] (4) -x (2) -G

