

6. vaja: Naloge in odgovori

V06-01

V bazi UniProtKB poiščite (<http://www.uniprot.org>) zapis za človeški E-kadherin. Oglejte si, kaj vse zapis v podatkovno bazo vsebuje (pomagajte si z navodili <http://www.uniprot.org/docs/manual>), podrobneje pa analizirajte določene dele zapisa in ugotovite (seznam skopirajte v polje za odgovor in dopišite manjkajoče):

- ali protein vsebuje signalni peptid in če ga, koliko ak ostankov vsebuje:
- ali se protein sintetizira v zreli obliki ali ne, ter koliko ak vsebuje zrelo obliko:
- koliko transmembranskih regij vsebuje:
- koliko glikozilacijskih mest je na polipeptidni verigi:
- koliko disulfidnih vezi je v zreli obliki proteina:
- kje v celici se protein nahaja:
- kaj je najpomembnejši biološki proces, v katerem protein sodeluje:
- s katerimi proteini interagira:
- zaporedje mRNA (napišite kodo zapisa v ustrezno bazo in ime baze):
- obstajajo kakšni strukturni podatki za ta protein:

Odgovor

V bazi UniProtKB zapis za človeški E-kadherin najdemo pod oznako **P12830**. Protein vsebuje signalni peptid, dolg 22 ak ostankov, kar lahko ugotovimo iz razdelka "Sequence annotation (Features)", v katerem so zbrani vsi podatki o signalnih peptidih, propeptidih, transmembranskih regijah, domenah v protein, položaju disulfidnih vezi, mestih za glikozilacijo... Iz istega razdelka tako tudi razberemo, da se protein ne sintetizira v zreli obliki, saj vsebuje propeptid, dolg 132 ak ostankov.

Sequence annotation (Features)

Feature key	Position(s)	Length	Description
Molecule processing			
<input type="checkbox"/> Signal peptide	1 – 22	22	Potential
<input type="checkbox"/> Propeptide	23 – 154	132	Potential
<input type="checkbox"/> Chain	155 – 882	728	Cadherin-1
<input type="checkbox"/> Chain	701 – 882	182	E-Cad/CTF1
<input type="checkbox"/> Chain	732 – 882	151	E-Cad/CTF2
<input type="checkbox"/> Chain	751 – 882	132	E-Cad/CTF3

Protein vsebuje eno transmembransko regijo,

Regions

<input type="checkbox"/> Topological domain	155 – 709	555	Extracellular	Potential
<input type="checkbox"/> Transmembrane	710 – 730	21	Helical;	Potential
<input type="checkbox"/> Topological domain	731 – 882	152	Cytoplasmic	Potential

ima 4 mesta za glikozilacijo, tvori pa tudi eno disulfidno vez z drugo molekulo E-kadherina (intermolekularna disulfidna vez).

Amino acid modifications

<input type="checkbox"/>	Modified residue	753	1	Phosphotyrosine; by SRC Ref.36
<input type="checkbox"/>	Modified residue	754	1	Phosphotyrosine; by SRC Ref.36
<input type="checkbox"/>	Modified residue	755	1	Phosphotyrosine; by SRC Ref.36
<input type="checkbox"/>	Modified residue	838	1	Phosphoserine By similarity
<input type="checkbox"/>	Modified residue	840	1	Phosphoserine By similarity
<input type="checkbox"/>	Modified residue	846	1	Phosphoserine By similarity
<input type="checkbox"/>	Glycosylation	558	1	N-linked (GlcNAc...) Ref.28
<input type="checkbox"/>	Glycosylation	570	1	N-linked (GlcNAc...) Ref.28
<input type="checkbox"/>	Glycosylation	622	1	N-linked (GlcNAc...) Ref.28
<input type="checkbox"/>	Glycosylation	637	1	N-linked (GlcNAc...) Ref.28
<input type="checkbox"/>	Disulfide bond	163		Interchain Ref.20

O lokalizaciji in funkciji posameznega proteina ponavadi zadosti informacij pridobimo že iz uvodnega poglavja "General annotation (Comments)". Tako ugotovimo, da se protein nahaja na celični membrani epiteljskih celic, kjer sodeluje pri tvorbi adherentnih stikov.

V istem poglavju lahko najdemo omenjene tudi nekatere proteine, s katerimi obravnavani protein interagira. Za natančnejše in bolj sistematično predstavljene podatke pa lahko pogledamo na dno zapisa o protein v razdelek "Cross-references", podpoglavje "Protein-protein interaction databases", kjer so nam na voljo povezave do različnih baz.

Protein-protein interaction databases	
BioGrid	107434 96 interactions.
DIP	DIP-477N .
IntAct	P12830 . 43 interactions.
MINT	MINT-119105 .
STRING	9606.ENSP00000261769 .

V istem razdelku najdemo tudi podatke o mRNA, na primer zapis z oznako Z13009 v bazi NCBI Nucleotide,

Sequence databases	
<input type="radio"/> EMBL	Z13009 mRNA. Translation: CAA78353.1 .
<input checked="" type="radio"/> GenBank	Z18923 mRNA. Translation: CAA79356.1 .
<input type="radio"/> DDBJ	L08599 mRNA. Translation: AAA61259.1 . Frameshift.
	AK290012 mRNA. Translation: BAF82701.1 .
	AK312551 mRNA. Translation: BAG35448.1 .
	DQ090940 Genomic DNA. Translation: AAY68225.1 .
	CH471092 Genomic DNA. Translation: EAW83243.1 .
	L34545 Genomic DNA. Translation: AAA21764.1 .
	D49685 Genomic DNA. Translation: BAA08537.1 .
	Z35402 Z35415 Genomic DNA. Translation: CAA84586.1 .
	X12790 mRNA. Translation: CAA31279.1 .
	X52279 mRNA. Translation: CAA36522.1 .
	S72492 , S72397 , S72491 Genomic DNA. Translation: AAD14108.1 .
PIR	JHUCE . S37654 .
RefSeq	NP_004351.1 . NM_004360.3 .
UniGene	Hs.461086 .

in pa podatke o strukturi.

3D structure databases						
<input checked="" type="radio"/> PDBe	Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
<input type="radio"/> RCSB PDB	1O6S	X-ray	1.80	B	156-255	[*]
<input type="radio"/> PDBj	2072	X-ray	2.00	A	155-367	[*]
	20MT	X-ray	2.00	B	156-255	[*]
	20MU	X-ray	1.80	B	156-255	[*]
	20MV	X-ray	1.90	B	156-255	[*]
	20MX	X-ray	1.70	B	156-258	[*]
	20MY	X-ray	1.70	B	156-255	[*]
	20MZ	X-ray	1.60	B	156-255	[*]
	3FF7	X-ray	1.80	A/B	155-253	[*]
	3FF8	X-ray	2.00	A/B	155-254	[*]
	3L6X	X-ray	2.40	B	756-773	[*]
	3L6Y	X-ray	3.00	B/D/F	756-773	[*]

V06-02

Oglejte si tudi zapise (jih je več!) za isti protein (človeški E-kadherin) v bazi NCBI Protein (<http://www.ncbi.nlm.nih.gov/protein>) ter primerjajte ta zapis z zapisom v bazi UniProtKB. Komentirajte. Pomagajte si z s FAQ za NCBI Proteins (http://www.ncbi.nlm.nih.gov/books/NBK49541/#NucProtFAQ.8_what_are_the_sources_of_th).

Odgovor

Zapisov v bazi NCBI Protein je več, ker nastanejo avtomatsko s prevajanjem kodirajočih regij v bazi GenBank Nucleotide. Najprej je navedenih nekaj splošnih podatkov o zapisu (dolžina, organizem v katerem se nahaja, reference...).

```

LOCUS       CAA78353                882 aa           linear   PRI 01-JUL-1999
DEFINITION  E-cadherin [Homo sapiens].
ACCESSION   CAA78353
VERSION     CAA78353.1  GI:31073
DBSOURCE    embl accession Z13009.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 882)
  AUTHORS   Bussemakers,M.J., van Bokhoven,A., Mees,S.G., Kemler,R. and
            Schalken,J.A.
  TITLE     Molecular cloning and characterization of the human E-cadherin cDNA
  JOURNAL   Mol. Biol. Rep. 17 (2), 123-128 (1993)
  PUBMED    8459805
    
```

Podobno kot v bazi UniProt sledijo informacije o lastnostih proteina, le da niso sistematično urejene po področjih, zaradi česar so podatki tudi manj pregledni.

```

FEATURES             Location/Qualifiers
  source              1..882
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
  Protein            1..882
                     /product="E-cadherin"
  Region            27..116
                     /region_name="Cadherin_pro"
                     /note="Cadherin prodomain like; pfam08758"
                     /db_xref="CDD:255020"
  mat\_peptide       155..882
                     /product="E-cadherin"
  Region            158..258
                     /region_name="CA_like"
                     /note="Cadherin repeat-like domain; cd00031"
                     /db_xref="CDD:206635"
  Site              order(165..166,221,223,254,256..257)
                     /site_type="other"
                     /note="Ca2+ binding site [ion binding]"
                     /db_xref="CDD:206635"
    
```

Na splošno lahko ugotovimo, da je za pridobivanje osnovnih informacij o posameznem proteinu baza UniProt uporabniku bolj prijazna.

V06-03

S pomočjo orodja ProtParam (<http://web.expasy.org/protparam/>) izračunajte molekulsko maso in ekstinkcijski koeficient za zrelo obliko človeškega proteina EpCAM.

Odgovor

Zapis za človeški protein EpCAM v bazi UniProt najdemo pod oznako **P16422**. Celoten zapis za protein vsebuje tudi zapis za signalni peptid, tako da moramo najprej pridobiti le zapis za zrelo obliko. Do zapisa v FASTA obliki najenostavneje pridemo tako, da najprej v razdelku "Sequence annotation (Features)" najdemo podpoglavje »Molecule processing« nato pa kliknemo na eno izmed mest na spodnji sliki označenih s puščico.



S pomočjo orodja ProtParam tako lahko ugotovimo, da je molekulska masa zrele oblike enaka 32733.6 Da. Za pravi izračun ekstinkcijskega koeficienta potrebujemo še informacijo o številu disulfidnih vezi. Iz rezultatov v ProtParam razberemo, da zapis vsebuje 12 cisteinov, iz zapisa v bazi UniProt pa, da protein tvori 6 disulfidnih vezi.

Amino acid modifications

<input type="checkbox"/>	Modified residue	24	1	Pyrrolidone carboxylic acid (Probable)
<input type="checkbox"/>	Glycosylation	74	1	N-linked (GlcNAc...); partial (Ref.12) (Ref.16)
<input type="checkbox"/>	Glycosylation	111	1	N-linked (GlcNAc...) (Ref.12) (Ref.16)
<input type="checkbox"/>	Glycosylation	198	1	N-linked (GlcNAc...) (Ref.16) (Ref.17)
<input type="checkbox"/>	Disulfide bond	27 ↔ 46		(Ref.12)
<input type="checkbox"/>	Disulfide bond	29 ↔ 59		(Ref.12)
<input type="checkbox"/>	Disulfide bond	38 ↔ 48		(Ref.12)
<input type="checkbox"/>	Disulfide bond	66 ↔ 99		(Ref.12)
<input type="checkbox"/>	Disulfide bond	110 ↔ 116		(Ref.12)
<input type="checkbox"/>	Disulfide bond	118 ↔ 135		(Ref.12)

To pomeni, da je pravi ekstinkcijski koeficient $28140 \text{ M}^{-1} \text{ cm}^{-1}$. Orodje ProtParam nam sicer samo izračuna le koeficienta za primera, da vsi cisteini tvorijo disulfidno vez ali pa nobeden. V kolikor naletimo na primer, ko le nekaj cistenov tvori disulfidne vezi lahko pravi ekstinkcijski koeficient izračunamo tako, da vrednosti, ki predpostavlja, da disulfidne vezi niso tvorjene, za vsako disulfidno vez prištejemo $125 \text{ M}^{-1} \text{ cm}^{-1}$.

Za pomoč pri delu z orodjem in odčitavanju rezultatov si pomagajte s PowerPoint predstavitevjo te vaje (5. diapozitiv).

V06-04

Bi lahko zunajcelični del človeškega proteina EpCAM (enak protein kot pri prejšnjem vprašanju) ločili od aktin vezavne domene človeškega alfa-aktinina 4 z uporabo ionsko-izmenjevalne kromatografije? Predlagajte nosilec in pufre za vezavo in elucijo. Za izračun izoelektrične točke uporabite ProtParam (<http://web.expasy.org/protparam/>) in še kakšen en program (<http://isoelectric.ovh.org>).

Odgovor

Najprej poiščemo zapis za zunajcelični del človeškega proteina EpCAM v FASTA obliki

Regions

<input type="checkbox"/>	Topological domain	24 – 265	242	Extracellular (Potential)
<input type="checkbox"/>	Transmembrane	266 – 288	23	Helical, (Potential)
<input type="checkbox"/>	Topological domain	289 – 314	26	Cytoplasmic (Potential)
<input type="checkbox"/>	Domain	63 – 135	73	Thyroglobulin type-1

in ga analiziramo z obema programoma. Orodje ProtParam nam poda izračunano vrednost izoelektrične točke 5,86, drugo orodje pa povprečno vrednost 5,82.

Sedaj v bazi UniProt poiščemo še zapis za aktin vezavno domeno človeškega alfa-aktinina 4. Zapis za celoten protein se v bazi nahaja pod oznako **043707**, zapis za aktin vezavno domeno v FASTA obliki pa poiščemo v razdelku "Sequence annotation (Features)", podpoglavje »Regions«.

Sequence annotation (Features)

Feature key	Position(s)	Length	Description
Molecule processing			
<input type="checkbox"/> Chain	1 – 911	911	Alpha-actinin-4
Regions			
<input type="checkbox"/> Domain	1 – 269	269	Actin-binding

Rezultat analize z orodjem ProtParam je 5,70, z drugim orodjem pa dobimo povprečno vrednost 5,79.

Izoelektrični točki imata, sploh po napovedi drugega programa, zelo podobno vrednost, kar bi oteževalo ločevanje z ionsko-izmenjevalno kromatografijo.

Če vseeno predpostavimo, da sta pravilni izoelektrični točki 5,86 za zunajcelični del EpCAM-a in 5,70 za aktin vezavno domeno alfa-aktinina 4 bi bile teoretične možnosti za izbiro nosilcev in pufrov za vezavo in elucijo sledeče:

zunajcelični del EpCAM-a – kratica zcEpCAM
 aktin vezavna domena alfa-aktinina 4 – kratica AvAA4

Anionski nosilec:

pH pufra za vezavo = 5,8 (pri teh pogojih se AvAA4 veže, zcEpCAM pa ne)
 pH pufra za elucijo = 5

Kationski nosilec:

pH pufra za vezavo = 5,8 (pri teh pogojih se zcEPCAM veže, AvAA4 pa ne)
 pH pufra za elucijo = 6,5

V06-05

V insektnih celicah (evkariontski ekspresijski sistem) želimo izraziti domeno Kazalnega tipa (Kazal-type) človeškega osteonektina. Je smiselno, da jo izrazimo v citosolu ali je bolje, da jo usmerimo v ER? V slednjem primeru načrtajte (ak zaporedje) signalni peptid, ki bi ga lahko za ta namen uporabili, ter preverite njegovo učinkovitost (tudi v smislu mesta odcepitve) z orodjem SignalP (<http://www.cbs.dtu.dk/services/SignalP/>) ter Phobius (<http://phobius.sbc.su.se>).

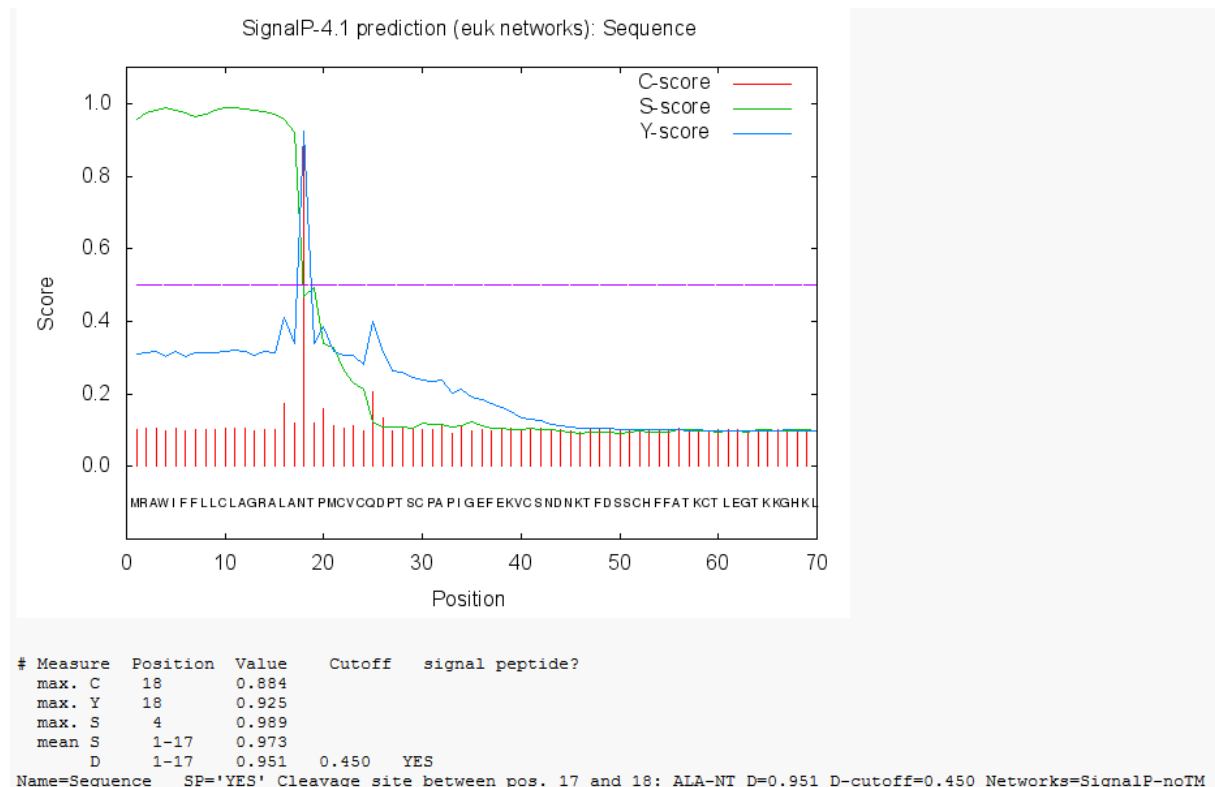
Odgovor

Zapis za čoveški osteonektin se v bazi UniProt nahaja pod oznako **P09486**. Zapis za domeno Kazalnega tipa pridobimo na enak način kot smo dele celotnega zapisa pridobili pri prejšnjih nalogah. Proteina ni smiselno izraziti v citosolu in ga je bolje usmeriti v ER. Do te ugotovitve pridemo zato, ker gre za zunajcelični protein (razvidno iz opisa) in ga je torej smiselno izražati v gojišče, kar pomeni da bo njegova pot vodila skozi ER. Še bolj pomembno pa je dejstvo, da je protein na 116 ak ostanku, ki je del domene Kazalnega tipa, N-glikozilirana, N-glikozilacija pa poteka izključno v ER.

Signalni peptid lahko načrtamo na podlagi skupnih značilnosti signalnih peptidov, kot so povzete v PowerPoint predstavitvi (7. diapozitiv). V praksi pogosto uporabimo že obstoječ signalni peptid istega ali pa kakega drugega proteina. V tem primeru lahko uporabimo kar signalni peptid osteonektina. Zapisa za signalni peptid in domeno Kazalnega tipa nato združimo (zapis za signalni peptid mora biti na N-koncu – torej na levi strani!). Združen zapis je sledeči (signalni peptid je podčrtan):

MRAWIFFLLCLAGRALANTPMVCVQDPTSCPAPIGEFEKVCSDNDNKTFDSSCHFFATKCTLEGTKKGHKLHLDYIG
PCKY

S pomočjo spletnih orodij nato preverimo, ali bo do cepitve res prišlo točno med signalnim peptidom in začetkom zapisa za del, ki ga želimo izraziti. Rezultat s programom SignalP:

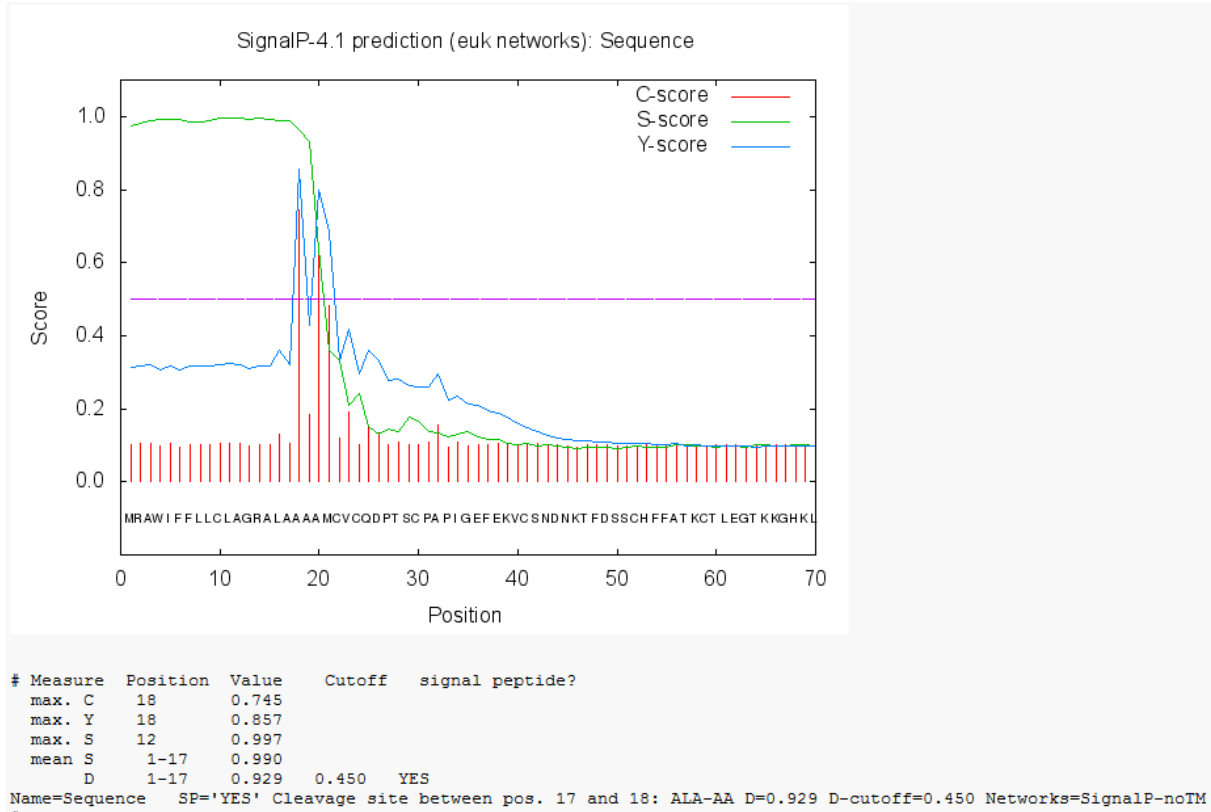


Program nam napove najverjetnejše mesto cepitve med 17. in 18. ak ostankom, torej za zapisom ALA in pred zapisom NT, kar ustreza mestu med signalnim peptidom in domeno. Na grafu pogledamo predvsem, če dobimo visok – mnogo višji od ostalih – vrh »C-score« na mestu prvega ak ostanka, ki ni več del signalnega peptida, ki sovpda s strmim padcem »S-score« na istem mestu. Tak rezultat pomeni visoko verjetnost cepitve točno pred tem mestom.

Podobno v programu Phobius preverimo, kje pride do strmega padca rdeče krivulje, ki označuje signalni peptid.

Za podrobnejše informacije o pomenu posameznih parametrov si lahko pogledate PowerPoint predstavitev.

Dodatno: Naloga se morda zdi trivialna, vendar načrtovanje signalnih peptidov ne poteka vedno tako gladko. Primer: če prve 3 ak ostanke domene Kazalnega tipa zamenjamo z alanini (NTP –AAA) je rezultat s programom SignalP veliko bolj dvoumen:



Pojavijo se dodatni vrhovi »C-score«, kar nakazuje, da bi do cepitve lahko prišlo tudi na drugih mestih. V takem primeru bi razmislili o izbiri drugega signalnega peptida ali pa dodajanju dodatnih ak preostankov na N-konec zapisa za domeno, ki jo želimo izraziti.

06-06

Določili smo celotno zaporedje cDNA, izolirano iz knjižnice cDNA mišjega epitelijskega tkiva. Vemo, da kodira za transmembranski protein, sestavljen iz najmanj 200 ak ostankov. Koliko transmembranskih vijačnic vsebuje? Na kateri strani membrane sta N- in C-konec proteina? Uporabite program TMHMM (<http://www.cbs.dtu.dk/services/TMHMM-2.0>).

Zaporedje:

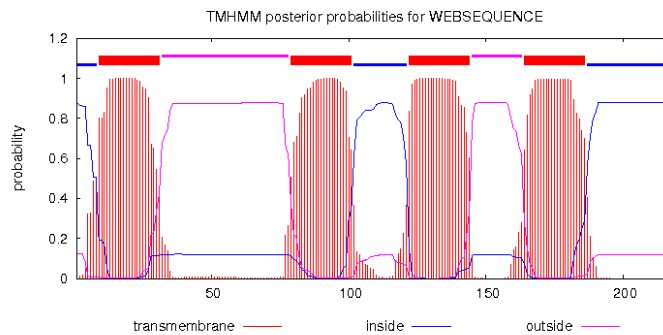
```
GTGAGCATT CAGTCTTTAGCCATGGGGTCTGCAGCGTTGGAAATTCTGGGTCTGGTGTGTGTCTGGTAGGATGG
GTGGGCTTGATCCTGGCGTGTGGGCTGCCATGTGGCAGGTGACTGCCTTCCTGGACCACAACATCGTGACGGCGC
AGACGACTTGAAGGGGCTGTGGATGTGCTGCGTGGTGCAGAGTACCGGGCACATGCAGTGCAAGGTGTATGAAT
CTGTGCTGGCGCTGAGTGCAGGAGGTGCAGGCAGCTCGGGCACTACCGTGGGCGCTGTGCTGCTGGCGCTGGTGGC
ACTCTTTGTTACCTTGACCGGCGCTCAGTGCACCACCTGCGTGGCCCCGGGCCAGTTAAGGCACGGGTAGCACTC
ACGGGAGGAGCGCTTTACGCGGTGTGCGGGCTGCTGGCACTCGTGCCGCTCTGCTGGTTCGCCAACATCGTTGTCC
GCGAGTTCTATGATCCGACGGTGCAGGTGTACAGAAGTACGAGCTGGGCGGGCGCTGTACATCGGCTGGGCGG
CCTCCGCACTGCTCATGTGCGGTGGCGGCCCTCGTGTGTTGCGGCGCTGGGTCTGCACCGGGCGCCCTGAGTTCAG
CTTCCCGGTCAAGTACTMTGCGCCGCGGCGGCCACGGCCAATGGCGATTACGACAAGAAGAACTATGTCTAAGG
GCGGGAGGCATGGCGGGGCTCTTCCCGCAGCTAAGCCCGCGATGGGAAAGACCGATGCGGGAAGCCGTGTGTGA
TGACGACCACCGCTGGGTTGCGCAGCGCAAGTCATGYTGGGTTGGGGCCAGACTTCCCCGCTCTCAGAGTCCGTT
WGACCATCACTARCCGGGCCYTGTCTAGAACAGACTACAGGCACCTTTAAGAAMTTGACCGACCTTTTTTTCTAT
GCGCAGTTGGCCACGACGTGGGTGGAACGCTCAGATTTTCATCGGTGAAGTTAGGCACCAAACCTGCCGGAACAGT
TCCTACTGAGACTCTGGGGGCACTAGATGCTGCCTTAATGTCCAGTGGCACCTGCTAACCTGAAAGGGCAGCTGG
AGAAACCCCGGGCTGCCAGAGGGANGTGTAAAAAGGGCATTTTTTTTTTKTRKTRGWGAAGAACCTACTGAAC
CAAAGGACTTAGCCTTGACAGCTGGTYTCACTCCAGCAWYCTCCCCAARGTTGGGGCCCTGTAGGTACCAGAGCCT
TAGAGGGGTTGCCTTCCTCCTGSAAGCTTGGGGCTGGGGGTTGGGCCGGGCAAGAATTTGCTCAGTAAATGGTT
TGAACACTTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

Odgovor

Najprej moramo cDNA zapis prevesti v proteinsko zaporedje (če smo podatek, da gre za DNA zapis spregledali v navodilih, lahko vseeno opazimo, da zapis vsebuje le znake GATC in se ne začne z M, tako da je zelo malo verjetno, da gre za protein). S pomočjo orodja ORF finder, ki smo ga že spoznali, poiščemo najdaljši bralni okvir. Najdemo ga v okvirju +1 z dolžno 657 ak ostankov.

Rezultat programa TMHMM:

```
# WEBSEQUENCE Length: 218
# WEBSEQUENCE Number of predicted TMHs: 4
# WEBSEQUENCE Exp number of AAs in TMHs: 90.67775
# WEBSEQUENCE Exp number, first 60 AAs: 22.88121
# WEBSEQUENCE Total prob of N-in: 0.87698
# WEBSEQUENCE POSSIBLE N-term signal sequence
WEBSEQUENCE TMHMM2.0 inside 1 8
WEBSEQUENCE TMHMM2.0 TMhelix 9 31
WEBSEQUENCE TMHMM2.0 outside 32 78
WEBSEQUENCE TMHMM2.0 TMhelix 79 101
WEBSEQUENCE TMHMM2.0 inside 102 121
WEBSEQUENCE TMHMM2.0 TMhelix 122 144
WEBSEQUENCE TMHMM2.0 outside 145 163
WEBSEQUENCE TMHMM2.0 TMhelix 164 186
WEBSEQUENCE TMHMM2.0 inside 187 218
```



ppmtogif: computing colormap... ppmtogif: 5 colors found

Ugotovimo, da vsebuje 4 transmembranske vijačnice (rdeče črtasti vrhovi na grafu ali pa zapis levo zgoraj) in da sta oba dela dela proteina na notranji strani membrane (na obeh koncih je modra črta nad roza ali pa zapis levo zgoraj).

Opomba: Če boste v programu ORF finder izbrali katerikoli drug odprt bralni okvir vam orodje TMHMM ne bo napovedalo nobene transmembranske vijačnice.

V06-07

Analizirajte aminokislinsko zaporedje mišjega proteina "Sushi, nidogen and EGF-like domain-containing protein 1" in sicer v smislu N-glikozilacije in domen. Rezultate analize primerjajte z anotacijami v podatkovni bazi UniProtKB ter komentirajte, ali se ujemajo in če se ne, zakaj menite, da ne.

Za analizo N-glikozilacije uporabite orodje NetNGlyc (<http://www.cbs.dtu.dk/services/NetNGlyc/>), za domene in funkcionalne regije pa Prosite (<http://prosite.expasy.org>).

Odgovor

Pravilen zapis za protein v bazi Uniprot ima oznako **Q70E20**. Iz vnosa najprej pridobimo zapis v FASTA obliki (pazite na signalni peptid!), ki ga nato vnesemo v obe spletni orodji.

NetNGlyc

Kot rezultat dobimo 11 kratkih zaporedij, ki v osnovi ustrezajo mestom za glikozilacijo.

(Threshold=0.5)					
SeqName	Position	Potential	Jury agreement	N-Glyc result	
Sequence	268 NPSY	0.6843	(9/9)	++	WARNING: PRO-X1.
Sequence	384 NYSC	0.7150	(9/9)	++	
Sequence	460 NTTL	0.6293	(8/9)	+	
Sequence	512 NISH	0.5963	(7/9)	+	
Sequence	688 NGTH	0.6812	(9/9)	++	
Sequence	862 NGSH	0.6624	(8/9)	+	
Sequence	953 NISV	0.7043	(9/9)	++	
Sequence	991 NISA	0.5039	(4/9)	+	
Sequence	1085 NLTA	0.5926	(7/9)	+	
Sequence	1115 NVTT	0.4109	(9/9)	--	
Sequence	1274 NSSK	0.4398	(8/9)	-	

Podrobneje si moramo ogledati še zadnji stolpec, kjer nam program poda svojo napoved ali posamezno zaporedje dejansko predstavlja mesto za glikozilacijo ali ne. Za zadnji dve (1115 in 1274) nam program napove, da glikozilacija tam najverjetneje ne poteče. Izmed ostalih zadetkov opazimo, da imajo nekateri ++ nekateri pa le +. Število + ustreza zanesljivosti napovedi – več jih je, bolj zanesljiva je napoved. Če si pogledamo le najbolj zanesljive napovedi nam tako ostanejo še štiri zaporedja, od katerih prvo (268) vsebuje prolin, zaradi česar glikozilacija tam najverjetneje ne poteče. Preostanejo še tri mesta (384, 688 in 953). Če jim prištejemo 24, kolikor je dolg signalni peptid, ki ga v program nismo vnesli, dobimo 408, 712 in 977. Ugotovimo, da ta tri mesta dejansko ustrezajo anotiranim v bazi UniProt

Amino acid modifications

<input type="checkbox"/>	Glycosylation	408	1	N-linked (GlcNAc...)	Potential
<input type="checkbox"/>	Glycosylation	712	1	N-linked (GlcNAc...)	Potential
<input type="checkbox"/>	Glycosylation	977	1	N-linked (GlcNAc...)	Potential

Prosite

Rezultati se popolnoma ujemajo z anotacijo. Opazimo le, da ima položaj vseh domen v primerjavi z anotacijo v bazi UniProt 24 ak ostankov manjšo številko, kar je posledica odsotnosti signalnega peptida (program Prosite »ne ve«, da je pred začetkom zapisa, ki smo ga vnesli, še signalni peptid).



V06-08

Napovejte sekundarno strukturo človeškega proteina Trop2 in sicer z Jpred (<http://www.compbio.dundee.ac.uk/www-jpred/>) in PSIPRED (<http://bioinf.cs.ucl.ac.uk/psipred/>). S slednjim napovejte tudi neurejene regije (DISOPRED2) in domene (DomPred). Primerjajte rezultate analize (nekateri lahko primerjate z anotacijami v bazi podatkov) ter komentirajte, kje (v katerih regijah) se napovedi razlikujejo in zakaj menite, da se (pomagajte si s pomočjo za posamezne programe).

PSIPRED - pomembno:

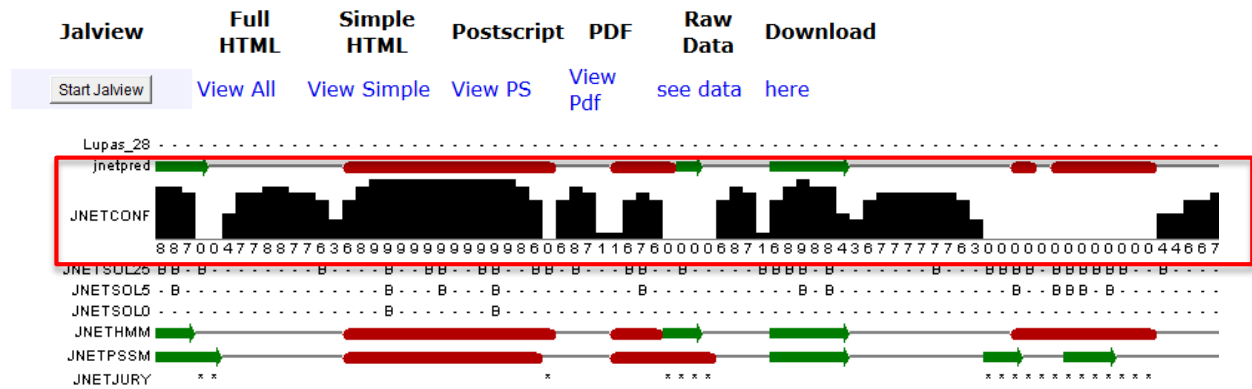
- vpišite tudi vaš e-poštni naslov, saj lahko napoved traja dlje časa (> 1 h),
- zaporedju morate dati ime (*Short identifier for submission*).

Odgovor

Zapis za človeški protein Trop2 se v bazi UniProt nahaja pod oznako **P09758**. Iz vnosa pridobim zapis za protein brez signalnega peptida v FASTA obliki in ga uporabimo za analizo z orodji Jpred in PSIPRED.

Jpred

Ko je napoved končana si rezultate lahko ogledamo v več formatih. Če imamo na računalniku nameščeno podporno Javo izberemo Jalview.



Na sliki vidimo rdeče in zelene regije. Rdeče regije ustrezajo napovedanim alfa-vijačnicam, zelene pa beta-listom. Rezultate odčitamo iz vrstice »jnetpred«, v ostalih vrsticah pa imamo še napovedi po različnih algoritmih. Pri napovedi je pomembno upoštevati tudi zanesljivost, saj orodje ne napove vseh delov z enako zanesljivostjo. Zanesljivost lahko odčitamo bodisi iz višine črnih stolpcev (višji stolpec pomeni bolj zanesljivo napoved) bodisi iz številčk pod njimi (9 - najvišja zanesljivost, 0 - najnižja zanesljivost).

Na zgornji sliki je tudi prikazan primer napovedi dveh alfa vijačnic, pri čemer je ena (leva) izmed napovedi zelo zanesljiva, druga (desna) pa ne. Pri interpretaciji upoštevamo le dele z visoko zanesljivostjo!

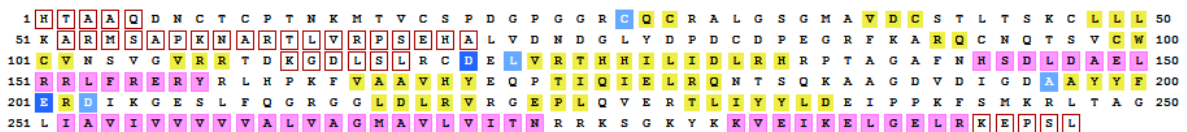
V našem primeru je zanesljivo napovedanih več beta-listov in 2 daljši alfa-vijačnici (ostale napovedi alfa-vijačnic so nezanesljive). Opazimo da na območju, kjer je v bazi UniProt anotiran transmembranski alfa-heliks (AGLIAVIVVVVALVAGMAVLVI) program sicer napove pravilno sekundarno strukturo, vendar napoved označi kot popolnoma nezanesljivo.

PSIPRED

Ločeno si pogledamo napoved sekundarne strukture, domen in neurejenih regij.

Secondary Structure Map

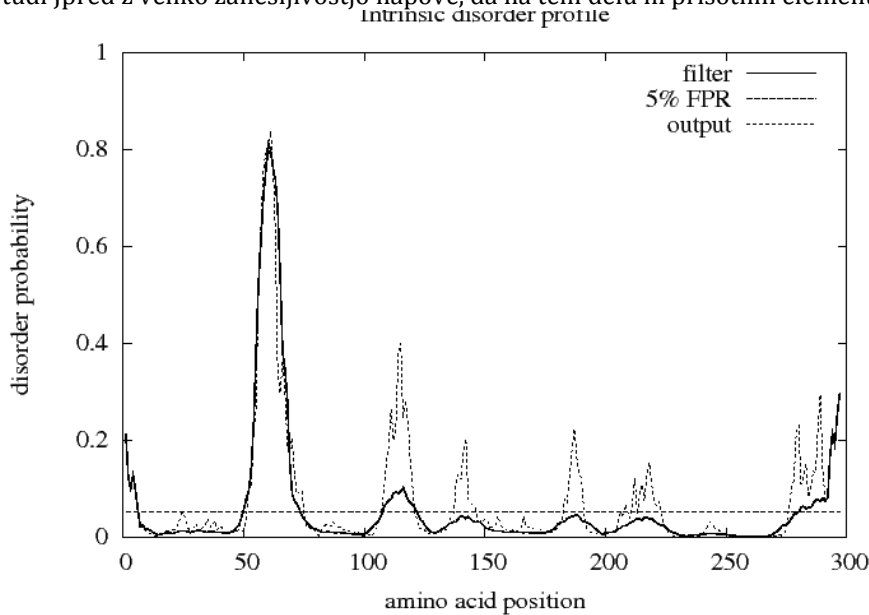
Feature predictions are colour coded onto the sequence according to the sequence feature key shown below.



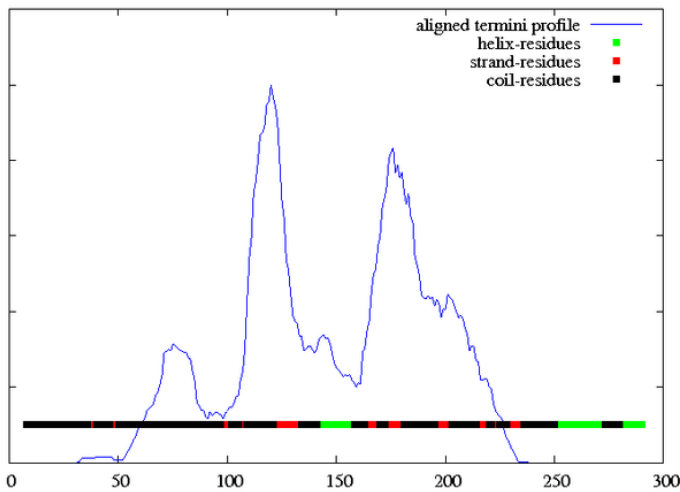
Key	Helix	Sheet	Disordered	Dompred Boundary	DomSSEA Boundary
Annotations	⊗	⊞	⊠	⊡	⊢

Orodje je napovedalo 3 alfa-vijačnice, vključno s tisto, ki je anotirana kot transmembranski alfa-heliks. Ostali dve približno sovpadata z napovedjo v Jpred. Ugotovimo pa, da se točni položaji med obema programoma in anotacijo pogosto razlikujejo za par ak ostankov, kar pa je posledica različnih algoritmov napovedi. O točnem položaju elementov sekundarne strukture tako ne moremo sklepati, lahko pa dobimo informacijo o približnem položaju.

Napoved neurejenih regij (DISOPRED2) nam z največjo zanesljivostjo napove neurejeno regijo na položaju nekje od 50-75 ak ostanka v vneseni sekvenci. Če si isto regijo ogledamo tudi v programu Jpred ugotovimo, da tudi Jpred z veliko zanesljivostjo napove, da na tem delu ni prisotnih elementov sekundarne strukture.



DomPred nam napove 3 domene, kar vidimo kot tri vrhove v diagramu.



Napovedi ne moremo primerjati z rezultati programa Jpred, lahko pa preverimo, če je napovedal tudi tiroglobulinsko domeno, ki je bazi UniProt anotirana med 70 in 145 ak ostankom. Da lahko primerjamo anotacijo z napovedjo moramo položaju ak ostankov v napovedi prišteti 26 ak ostankov za signalni peptid. Vrhovi tako ustrezajo približno 70-110, 130 do 160 in 185 do 240. Med ponujenimi rezultati tiroglobulinske domene ne najdemo.