

## Teoretična vprašanja in teme pri zagovorih vaj iz biokemijske informatike

### 1. Naštej Boolove operatorje

Osnovni Boolovi operatorji so AND, OR in NOT (NOT je tudi unitarni operator). Sprejmejo par argumentov in jih spremenijo v rezultat.

### 2. Kakšne načine naslavljanja (sklicevanja) poznaš?

Sklic določa celico ali obseg celic na delovnem listu in kaže Microsoft Excelu, kje iskati vrednosti ali podatke, ki jih želite uporabiti v formuli. S sklici lahko podatke, ki so v različnih delih delovnega lista, uporabite v eni formuli ali pa uporabite vrednost ene celice v več formulah. Prav tako se lahko sklicujete na celice na drugih listih istega delovnega zvezka in na druge delovne zvezke. Sklici na celice v drugih delovnih zvezkih se imenujejo povezave.

Relativni sklici Relativen sklic v formuli, kot je A1, temelji na relativnem položaju celice, ki vsebuje formulo in celico, na katero se sklic nanaša. Če se mesto celice, ki vsebuje formulo, spremeni, se spremeni tudi sklic. Če formulo prekopirate prek vrstic ali prek stolpcev, se sklic samodejno prilagodi.

Absolutni sklici Absoluten sklic na celico v formuli, kot je \$A\$1, se vedno sklicuje na celico na določenem mestu. Če se mesto celice, ki vsebuje formulo, spremeni, absoluten sklic ostane nespremenjen. Če formulo prekopirate prek vrstic ali prek stolpcev, se absolutni sklic ne prilagodi. Nove formule privzeto uporabljajo relativni sklic in jih morate preklopiti na absoluten sklic.

Mešani sklici Mešani sklic vsebuje ali absoluten stolpec in relativno vrstico ali absolutno vrstico in relativen stolpec. Sklic na absoluten stolpec jemlje oblike \$A1, \$B1 itd. Sklic na absolutno vrstico jemlje oblike A\$1, B\$1 itd. Če se položaj celice, ki vsebuje formulo, spremeni, se relativni sklic spremeni, medtem ko se absolutni sklic ne.

### 3. Optimalni začetni nukleotidi za PCR (določimo jih z GeneRunnerjem)

	<i>Pri DNA</i>	<i>Pri RNA</i>
Dolžina	18-30 nt	18-25 nt
T <sub>m</sub> (tališče)	52-65 °C	55-80 °C
stabilnost	Nizka vsebnost G	Zadnji nukleotid G ali C, vendar ne več zaporednih GC parov
% GC	40-60	45-55
Preprečevanje tvorbe nezaželenih sekundarnih struktur		
Nekomplementarnost samemu sebi ali drugemu nukleotidu predvsem na 3' koncu		
Specifičnost samo za izbrano DNA/RNA		

#### 4. Frekvenca uporabe genskega koda!

Povezava s koncentracijo specifičnih tRNA. Pomembno za optimalno ekspresijo rekombinantnih proteinov

Original assumption: if multiple codons for an amino acid, expect equal frequency of use.

Surprise! Codon use is often highly biased. Eg. *E. coli* can use AUA, AUU, or AUC to specify isoleucine; but only 1 in 20 times is an isoleucine coded by AUA; 19/20 times encoded by AUU or AUC. So AUA is only rarely used -- may allow evolution to develop alternate codes. Codone usage database: <http://www.kazusa.or.jp/codon/>  
Torej.. če uporabimo AUA za izolevcin v našem rekombinantnem proteinu, je zelo velika verjetnost, da bomo pri delu neuspešni!

#### 5. Restriksijske mape. Analiza mest za restriksijske encime. Metilazna občutljivost

Restriksijske mape povejo, katera mesta za restriktaze so prisotna in katerih ni. Posebno pomembna so mesta, ki nastopajo le enkrat ali dvakrat. Pri večjem številu encimov (in s tem fragmentov) je za analizo mest potreben zahteven računalniški program, ki upošteva razgradnjo neznanega zaporedja z več encimi hkrati, potem pa mora fragmente urediti po zaporedju.

#### 6. EST?

An **expressed sequence tag** or **EST** is a short sub-sequence of a [protein-coding DNA sequence](#). It was originally intended as a way to identify gene [transcripts](#), but has since been instrumental in gene discovery and sequence determination. An EST is produced by one-shot [sequencing](#) of a cloned [mRNA](#), and the resulting sequence is a relatively low quality fragment whose length is limited by current technology to approximately 500 to 800 [nucleotides](#). ESTs are also a useful resource for designing probes for [DNA microarrays](#) used to determine [gene expression](#). (*vir: wikipedia*)

#### 7. Drseče okno

Uporaba drsečega okna povpreči neke vrednosti v aminokislinskem zaporedju. Velikost okna je odvisna od vprašanja, ki nas zanima. ( če iščemo transmembranske vijačnice ki so dolge ~20 AK, uporabimo velikost okna od 5 do 9 AK)

#### 8. Hidrofobnost

AK zaporedje določa prostorsko strukturo proteinov in hidrofobni efekt je ena najpomembnejših interakcij, ki so podlaga tvorbe 3D strukture. Fizikalna osnova je entropijski efekt ureditve topila okoli nepolarnih skupin , s katerimi topilo ne more tvoriti vodikovih vezi. Hidrofobni profil je pripomoček, s pomočjo katerega lahko napovemo lastnosti izbranih regij v AK zaporedju. (primer: najvišja hidrofobnost pogosto sovпада z beta strukturo).

## 9. Struktura proteinov

<i>primarna</i>	nukleotidno zaporedje, cisteinski mostički
<i>sekundarna</i>	$\alpha$ -vijačnica, $\beta$ -ploskev
<i>terciarna</i>	3D-zvitje ene polipeptidne verige
<i>kvararna</i>	Kako se več podenot organizira med sabo

## 10. Napovedovanje sekundarne strukture! Kateri programi, kako delujejo?

Programi, kot je Generunner (in ostala orodja na [www.expasy.ch](http://www.expasy.ch)) vsebujejo algoritme za iskanje določenih predelov v AK zaporedju, ki se zvijejo v značilne sekundarne strukture:

Algoritem Chou-Fasman: izmeri težnjo aminokislin, da se nahajajo v določenem tipu sekundarne strukture.. Podatki za to »težnjo« so bili dobljeni iz statistične analize vseh takrat znanih 3D struktur proteinov. Algoritem opazuje več sosednjih AK in če imajo vse podobno težnjo jim pripiše določeno sekundarno strukturo.

Garnier-Robson: Temelji na tem, kakšen vpliv ima določena AK, na ostale, bolj oddaljene AK

Sekundarno strukturo se določa še z: nevronskimi mrežami, na osnovi segmentov znanih struktur in z vključitvijo nelokalnih interakcij (H vezi med segmenti beta strukture in alfa vijačnico, na osnovi statistike 3D struktur)

## 11. Napovedovanje transmembranskih regij

Napovedovanje transmembranskih regij omogočajo programi (najdeš jih pa na [www.au.expasy.org/tools/#topology](http://www.au.expasy.org/tools/#topology)):

- [DAS](#) - Prediction of transmembrane regions in prokaryotes using the Dense Alignment Surface method (Stockholm University)
- [HMMTOP](#) - Prediction of transmembrane helices and topology of proteins (Hungarian Academy of Sciences)
- [PredictProtein](#) - Prediction of transmembrane helix location and topology (Columbia University)
- [SOSUI](#) - Prediction of transmembrane regions (TUAT; Tokyo Univ. of Agriculture & Technology)
- [TMAP](#) - Transmembrane detection based on multiple sequence alignment (Karolinska Institut; Sweden)
- [TMHMM \(TM-hidden Markovski Models\)](#) - Prediction of transmembrane helices in proteins (CBS; Denmark) → **tega smo mi uporabljali (definiramo stanja, ki odgovarjajo biološki realnosti, recimo TM-regije, notranje ali zunanje zanke; in vsaka AK ima določeno verjetnost, da se nahaja v enem od teh stanj; recimo za prolin je najbolj verjetno, da bo v zanki, ker je preokoren za heliks. Določi se verjetnost prehoda med stanjima uparabimo pa seveda udi bazo proteinov z znanimi TM-strukturami in te informacije povežemo z utežnimi matrikami)**
- [TMPred](#) - Prediction of transmembrane regions and protein orientation (EMBnet-CH)
- [TopPred 2](#) - Topology prediction of membrane proteins (Stockholm University)

## 12. Napovedovanje zavrtih vijačnic (coiled-coil)

Program COILS – na osnovi znanih obvitih vijačnic so pripravili frekvenčno pozicijsko matriko

### 13. Kaj je faktor vpliva

Faktor vpliva (IF=impact factor) je številka, ki nam pove kako pomembna je revija, glede na to, koliko njenih člankov je citiranih v drugih revijah. Večji kot je, boljša/bolj vplivna je revija. Izračunava jo ISI enkrat na leto po naslednjem postopku:

The impact factor is calculated by dividing the number of current citations to articles published in the two previous years by the total number of articles published in the two previous years.

### 14. Kaj je anotacija?

Dodajanje bioloških informacij nukleotidnemu oziroma AK zaporedju v bazah podatkov

### 15. Kaj je prileganje?

Prileganje (alignment) je medsebojna ureditev zaporedij tako, da se določi korespondenco med posameznimi elementi aminokislinskih ali nukleotidnih zaporedij. Optimalno prileganje ustreza največji podobnosti med zaporedjema.

### 16. Pomen poznavanja podobnosti zaporedij

Iz podobnosti zaporedij lahko sklepamo na podobnost terciarne strukture in podobno funkcijo (tudi v različnih organizmih).

### 17. Podobnost in homologija

PODOBNOŠT je *kvantitativna* oznaka, ki pove, kako podobni sta si zaporedji (lahko v %).

HOMOLOGIJA je *kvalitativna* oznaka, ki poudarja evolucijsko povezanost dveh zaporedij (ne moremo recimo reči, da sta si zaporedji 45% homologni, lahko edino rečemo ali sta ali nista).

### 18. Točkovne matrike in le-teh izboljšave

V matriki na mestu, kjer sta znaka enaka naredimo piko. Predeli, kjer sta dve zaporedji enaki se pokažejo kot črte. Pri nukleinskih kislinah je dosti šuma (25% naključna verjetnost da sta dve enaki). Matrike izboljšamo z uporabo oken in s tem odstranimo večino »šuma« (npr. točka le, če je enakih 5 zaporednih znakov)

### 19. Vrste algoritmov za prileganje:

<i><b>Rigorozni</b></i>	<i><b>Hevristični</b></i>
Vedno izračunajo optimalno prileganje	Reševanje problema po stopnjah (vsak korak je osnova za naslednjega, ki predlaga boljšo rešitev), vendar ni zagotovila, da najdejo optimalno prileganje
Časovno zahtevni	Hitrejši, ker primerjajo le del potencialnega prileganja
Needleman-Wunsch, Smith-Waterman	FASTA, BLAST

Delimo pa jih tudi glede na to, ali primerjajo globalno ali lokalno

## 20. Algoritmi za prileganje zaporedij

Globalno prileganje – primerja celotno dolžino dveh zaporedij, statistična teorija ocene prileganja ni poznana, najprimernejše za primerjanje homolognih zaporedij, težko zazna podobnost mozaičnih proteinov

Lokalno prileganje – primerja le najbolj podobne predele zaporedij, statistična ocena kvalitete prileganja, primeren za iskanje podobnih domen

## 21. Needleman-Wunsch

Needleman-Wunsch je rigorozen algoritem za prileganje zaporedij. Vedno izračuna optimalno prileganje, zato je časovno zahteven. Vendar pa uporablja dinamično programiranje. Koraki: inicializacija matrike (začne desno spodaj), sešteje vrednosti na vseh optimalnih delnih poteh (premika se levo navzgor), povratno sledi optimalno prileganje (maksimalno število v prvi vrstici in stolpcu). Glej še 8. profesorjevo predavanje z naslovom Prileganje uaporedij beljakobin in nukleinskih kislin!

## 22. Matrike zamenjav

Za izračun podobnosti med zaporedji potrebujemo neko merilo za izračun razdalje (matrika)

Lahko upoštevamo identiteto, upoštevamo da določeni tipi mutacij nastopajo pogosteje, upoštevamo minimalno število nukleotidnih zamenjav za AK spremembo, kemijsko sorodnost, evolucijske zamenjave, prileganje sorodnih proteinov (PAM), prileganje segmentov sorodnih proteinov (BLOSUM)

## 23. ClustalX. Kaj so drevesa? Oblike dreves

Program **ClustalX** omogoča prileganje več zaporedij: najprej primerja pare zaporedij, tvori drevo za vodenje prileganja, bolj podobna zaporedja si ležijo bližje.

Taka prileganja uporabimo za tvorbo **dreves**, s katerimi ugotavljamo filogenetske odnose med zaporedji (oz. posredno organizmi). Drevesa so grafi sestavljeni iz vozlov in vej. Delimo jih na zakoreninjena in nezakoreninjena. Predstavimo jih lahko kot kladogram, dendrogram, radioalno nezakoreninjeno drevo ali filogram. Drevo lahko vsebuje še več informacij: če je to čas divergence imamo ultrametrično drevo, če odraža evolucijske spremembe je to aditivno drevo

## 24. Pristopa pri konstrukciji filogenetskih dreves

Fenetični pristop – drevo zgradimo na osnovi (fenotipskih) podobnosti podatkov brez razumevanja evolucijskih mehanizmov

Kladistični pristop – drevo zgradimo ob upoštevanju različnih možnih evolucijskih poti, med katerimi izberemo optimalnega – kladogram

## 25. Homologi, ortologi, paralogi in ksenologi

HOMOLOGA=dva proteina sta homologna če imata skupnega prednika, ni pa nujno, da imata isto funkcijo. Homologe delimo na paraloge in ortologe (in ksenologe):

ORTOLOGA=homologi, ki nastanejo z razvojem novih vrst (isti prednik, drugačna funkcija)

PARALOGA=homologi, ki nastanejo z gensko duplikacijo (isti prednik, ista funkcija)

KSENOLOGA=homologi nastali s horizontalnim prenosom genov (to je takrat kadar se % GC-parov v zaporedju zelo razlikuje od ostalih genov za proteine v določenem organizmu)

## 26. Združevanje sosedov (NJ), metoda vezanja (bootstrapping)

NJ- vsota dolžin vej je minimalna, sosedje so elementi v drevesu, ki so povezani preko enega notranjega vozla. Izračunamo dolžino vej za vse pare i-j, in par z najmanjšo vsoto dolžin vej izberemo za soseda. Nov par obravnavamo kot en element, nove razdalje izračunamo iz povprečja starih.

Metoda vezanja – analiziramo, če je topologija drevesa ustrezna tudi za omejen izbor znakov iz primerjalnih vrst. Iz izvirnih podatkov vzorčimo nove podatke tako, da stolpci v prileganju ostajajo enaki. Vzorčenje in tvorbo dreves ponovimo velikokrat in beležimo, kako pogosto se ponovi določena topologija. Pri bootstrappingu nove podatke kreiramo na osnovi izbora stolpcev s tem, da je obseg novih podatkov enak kot izhodni in lahko določen stolpec izberemo večkrat kakšnega pa nikoli.

## 27. Načini predstavitve molekule

Molekulo lahko predstavimo s programi Swiss-PdbViewer, Rasmol, Chime, Cn3D, VRML, MOLMOL, VMD. Vrste prikazov, če nas zanimajo:

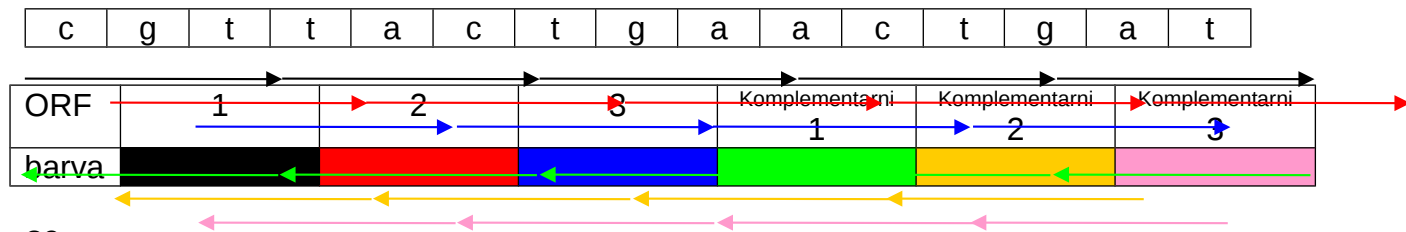
- **Atomi:** CPK-model, žična mreža (slaba za proteine), palčke in kroglice (stick&ball), navidezne vezi
- **Površine:** van der Waalsova, Connolyeva, površina dostopna topilu ali encimu, izopotencialne (ki imajo enak elektrostatski, lipofilni potencial ali interakcijsko energijo)
- **Vezi:** predvsem skeletni modeli
- **Shematsko:** trakovi, cilindri

## 28. RMSD?

The result of a structural alignment of two proteins is a superposition of their atomic [coordinate](#) sets with a minimal [root mean square](#) deviation (RMSD) (ja..wtf? ☺ ) between the two structures. (*Wikipedia*)

## 29. ORF (Open Reading Frame)

Je način na katerega lahko beremo zaporedje



## 30.

### Izoelektrična točka – kako jo določamo eksperimentalno, kje jo je potrebno poznati?

V izoelektrični točki je vsota vseh nabojev posameznih aminokislin v proteinu enaka nič. Določamo jo z uporabo titracijskih krivulj. Koristno jo je poznati pri ločevanju in čiščenju proteinskih vzorcev.

## 31. Oblike programa blast

BLAST je program s katerim iščemo prileganja med našim danim zaporedjem in vsemi ostalimi zaporedji v zbirki. Odvisno od vrste zaporedja, ki ga imamo, in od tega, kako želimo imeti podan rezultat, obstajajo sledeče oblike programa BLAST:

<b>Ime oblike</b>	<b>Katero zaporedje imaš?</b>	<b>Po kateri bazi zaporedij išče?</b>
<b>blastp</b>	aminokislinsko	aminokislinsko
<b>tblastn</b>	aminokislinsko	nukleotidno ( <i>prevedeno</i> )
<b>blastX</b>	nukleotidno ( <i>prevedeno</i> )	aminokislinsko
<b>blastn</b>	nukleotidno	nukleotidno
<b>tblastX</b>	nukleotidno ( <i>prevedeno</i> )	nukleotidno ( <i>prevedeno</i> )

## 32. E-vrednost. Od česa je odvisen rezultat pri programu BLAST

**E-** porazdelitev ekstremne vrednosti. Verjetnost da najdemo zadetek z rezultatom  $S$  je  $e^E$ . Manjši kot je  $E$ , boljše je ujemanje dveh zaporedij. Rezultat v programu BLAST je odvisen od iskanega zaporedja, zbirke zaporedij ki jih preiskujemo, uporabljene matrike zamenjav, definirane  $E$  vrednosti, filtriranja predelov z nizko kompleksnostjo, omejitev na organizem oziroma taksonomsko skupino, parametrov za vrzeli, dolžine besed..

## 33. Uporaba filtrov pri BLAST

Iščemo samo po določenih organizmih, ne upoštevamo regij z nizko kompleksnostjo..

