Prirodno-matematički fakultet u Banjoj Luci,
Matematika i informatika – informatički smjer

BLAST algoritam za pretraživanje proteinskih sekvenci

Predmet: Informacione tehnologije i društvo
Profesor: Dragan Matić

Student: Uroš Bojić

Sadržaj:

1. *Pretraživanje sličnosti u bazi podataka sekvence*

Slične sekvence obično potiču od predačke sekvence. To znači da, one vjerovatno imaju zajedničkog pretka, sličnu strukturu i sličnu biološku funkciju te možemo reći da pripadaju istoj proteinskoj porodici.
Ako za jednog člana te porodice postoje podaci o strukturi i funkciji, oni se mogu primjeniti i na one članove za koje još ne postoji eksperimentalna potvrda datih informacija i to prema principu **homologije** (sve homologne sekvence imaju istog pretka, sličnu strukturu i sličnu funkciju). Ovo se odnosi i na sekvence koje potiču od različitih organizama.

Na primer, zamislite da vasa omiljena sekvenca izgleda baš kao neka druga sekvenca koju je neko proučio. Pošto su ove dvije sekvence slične, možete reći: "Ako je nešto tačno za tu sekvencu, to je vjerovatno tačno i za moju!" Zamislite koliko vremena možete uštedjeti: Studiranje gena u laboratoriji traje godinama, dok traženje sličnosti u bazi podataka traje nekoliko sekundi.

Ukoliko su dva proteina ili dvije sekvence gena veoma slične biolozi ih nazivaju homolognima (imaju iste pretke, slicnu funkciju i slicnu strukturu). Problem nastaje prilikom utvrđivanja sličnosti. Ako su vaše sekvence dugačke više od 100 aminokiselina, pravilo kaže da su protein homologni ako je 25 procenata aminokiselina identično.
Vrijednosti niže od ovih potpadaju u tzv. *„zonu sumraka“* – gdje:

- **Ništa nije sigurno u vezi s uočenim sličnostima**
- **Homologija odnosno nehomologija nikada nije zagarantovana**

Vrijednost od 25% je prag identičnosti kada je u pitanju homologija. U stvarnosti su stvari ipak mnogo komplikovanije. Ako dvije sekvence imaju manje od 25% identičnost, to ne znači da nisu homologe, već samo da nema dovoljno pouzdanosti tj. dokaza. U većini slučajeva, da bismo utvrdili da su dvije sekvence istinski homologne, trebaju nam još neke informacije koje nam BLAST pretraga daje. Ove informacije uključuju:

- **E-vrijednost (*Expectation value*)** koja nam govori koliko je vjerovatno da je sličnost između naše sekvence i sekvence iz baze podataka rezultat slučajnosti. Kao dobri rezultati prihvataju se E-vrijednosti manje od 0,001
- **dužina sličnih segmenata između dvije sekvence**,
- **struktura konzerviranosti aminokiselinske sekvence**.

Homologija je kvalitativan parametar, dok je sličnost kvantitativan. Drugim riječima, dvije sekvence su ili homologne ili nehomologne, a mogu biti slične u različitom stepenu (15%,50%,75.5%). Mi ne možemo reći da je nešto 40% homologno, baš kao što ne možemo reći da je žena 40% trudna.

2. BLAST: Alat za pretragu sličnosti sekvenci

Kada radimo sa proteinima želimo znati informaciju o njihovoj funkciji. Međutim, utvrđivanje funkcije nekog proteina zahtijeva i dosta vremena i dosta novca. Ukoliko smo sekvencirali neku proteinsku molekulu, ali ne znamo koja joj je funkcija, potražite njoj sličnu u bazi podataka. Na taj način možete dobiti dobru polaznu ideju o kojem proteinu se radi i u tom smislu osmisliti svoje dalje eksperimente.

Pre trideset godina biolozi su pretraživali bazu podataka štampanjem cjelokupnog sadržaja na baze na papiru, utisnuli otisak na zid kancelarije, zapisali svoju sekvencu na komad papira i nekoliko sati pokušavali poklopiti svoju sekvencu sa otiskom na zidu. A sada BLAST vrši poklapanje papira na zidu za vas, uz znatno manje vremena.

Varijante BLAST-a:
Postoje dvije varijante BLAST-a koje se bave analizom proteina i to su:

- **blastp** koji upoređuje proteinsku sekvencu s proteinskom bazom podataka,
- **tblastn** koji poredi proteinsku sekvencu s bazom podataka nukleotidnih sekvenci

Kada koristiti jednu, a kada drugu verziju? Ukoliko želimo da saznamo nešto o funkciji proteina koristimo **blastp** ,a ukoliko želimo da otkrijemo nove gene kodiranjem jednostavnih proteina onda koristimo **tblastn**.

Dvije najpopularnije blastp mrežne usluge su:

- Blastp server nacionalnog centra za biotehnološke informacije (NCBI) u SAD
- Blastp server sa švajcarskog EMBnet server (Swiss-Prot)

Ova dva servera imaju malo drugačije interfejse. Dobri razlozi za korištenje više BLAST server su:

- Baze podataka: BLAST serveri vam ne daju pristup istim bazama podataka. Ako na jednom serveru ne pronađete potrebnu bazu podataka, uvijek je možete potražiti na drugom serveru.
- Brzina: Najpopularniji serveri su često pretrpani. Kad se to dogodi, uvijek možemo pretražiti na drugom mjestu.

3. Izbor odgovarajućih parametara

Sada ćemo da vidimo glavne parametre u BLAST-u i kako ih prilagoditi našim potrebama. Kod većine BLAST servera ovi parametri se malo razlikuju, mi ćemo raditi na NCBI serveru. Neki od ključni razloga za promjenu BLAST parametara, kao i parametri koji trebamo promjeniti, su:

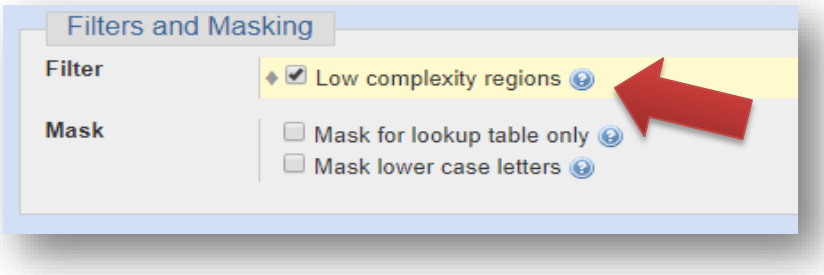
- Sekvenca koja nas zanima sadrži previše identičnih rezidua, u ovom slučaju je potrebno promjenuti sekvencijalni filter (automatic masking)
- BLAST nam ne prikazuje nikakve rezultate, tada su parametri podudaranja koje mijenjamo "matrix" i "gap costs"
- BLAST nam vraća previše rezultata, ukoliko nam se ovo desi tada možemo da promjenimo nekoliko stvari kao što su: baza podataka u kojoj tražimo poklapanja, unesene ključne riječi, broj maksimalnih izvještaja o poklapanju, expect (prag e-vrijednosti) itd.

3.1. Kontrola maskiranja sekvence

Kada BLAST pretražuje baze podataka,on čini važnu pretpostavku: BLAST pretpostavlja da su sve naše sekvence prosječne sekvence. Ukoliko pretražujemo proteinske sekvence, BLAST pretpostavlja da je prosječni sastav proteina jednak prosječnom sastavu cijele baze podataka. U praksi međutim nije sve tako savršeno.

Mnogi proteini sadrže oblasti niske složenosti (ili niske entropije). Na primjer, ove oblasti mogu biti segmenti koji sadrže mnogo prolina ili mnogo rezidua glutaminske kiseline. Ako BLAST poravna dva domena bogata prolinom, ovo poravnavanje dobija vrlo dobru E-vrijednost zbog velikog broja identičnih aminokiselina koje sadrži. Nažalost, postoji dobra šansa da preteini koji sadrže ova dva domena uopšte nisu povezana. U stvari, ovi domeni bogati aminokiselinama zavaravaju BLAST.

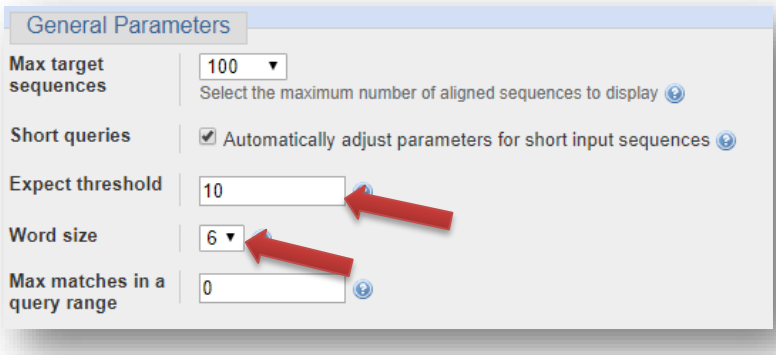
Da bi se izbjegao ovaj problem, BLAST filtrira oblasti male složenosti prilikom analize proteina. Da bi to učinio, zamjenjuje one regije u nizu sa X-om. Ukoliko ste posebno zainteresovani za ove oblasti i ne želite da ovi regioni budu filtrirani iz pretrage morate poništiti odabir odgovarajućeg čekboksa Low complexity regions u odjeljku Filters and Masking,koji se pojavljuje kada na stranici pretraživanja blastp-a otvorimo napredna podešavanja odnosno Algorithm parameters.



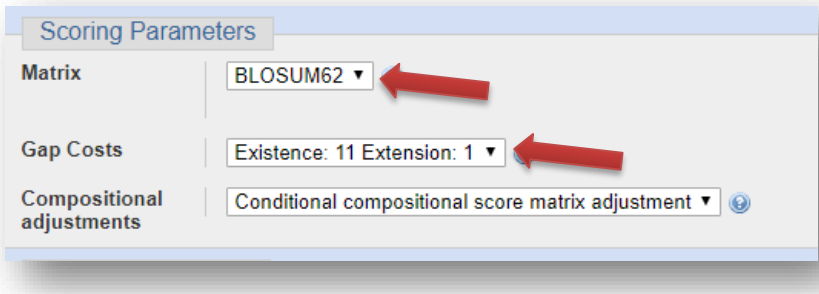
3.2. Promjena BLAST parametara za poravnavanje

Na NCBI BLAST serveru postoje parametri za poravnavanje kao što su očekivanji prag (Expect threshold), dužina tačnog poravnavanja sekvence (Word size).

Word Size je BLAST-ov „tajni recept", to predstavlja minimalnu veličinu elemenata sekvence koju BLAST pokušava do uklopi sa sekvencama iz baze podataka. Na primjer, ako postavimo Word size na 6, BLAST će razmatrati sve moguće podsekvence veličine 6 sadržane u našem upitu, a u bazi podataka će identifikovati sve sekvence koje sadrže riječi veličine 6 koje su slične onima u našoj sekvenci. Duge riječi čine BLAST bržim i manje osjetljivim, ali i manje pouzdanim. Kraće riječi rade suprotno.



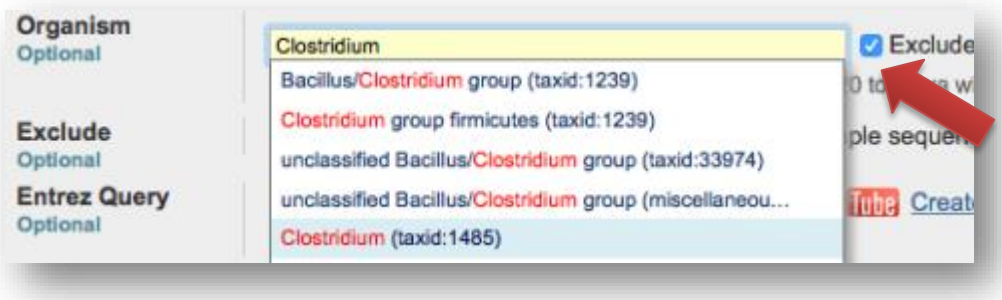
Pored ovih parametara tu su još i dva vrlo važna parametra za poravnavanje: gap costs (koriste se da prilagode poklapanje broju i dužini praznina u sekvenci) i matrix (poredi sve moguće parove rezidua i dodjeljuje im rezultat). Ako promjenimo bilo koji od ovih parametara, BLAST će nam vjerovatno vratiti različite rezultate. Kako je već naznačeno, BLAST poredi zadatu sekvencu s određenom bazom podataka, koju odabiremo u zavisnosti od konačnog cilja analize. Kada su u pitanju proteini, funkcija molekule direktno zavisi od njenog prostornog izgleda.



3.3. Kontrolisanje BLAST rezultata

Ako predmet vašeg upita pripada velikoj grupi proteina, BLAST rezultat može vam stvoriti probleme jer baze podataka sadrže previše sekvenci gotovo identične vašoj. Ponekad ovo bogatstvo homolognih nizova može da vas spreči da vidite homolognu sekvencu koja je manje srodna, ali još uvek povezana sa eksperimentalnim informacijama. Rješenja su:

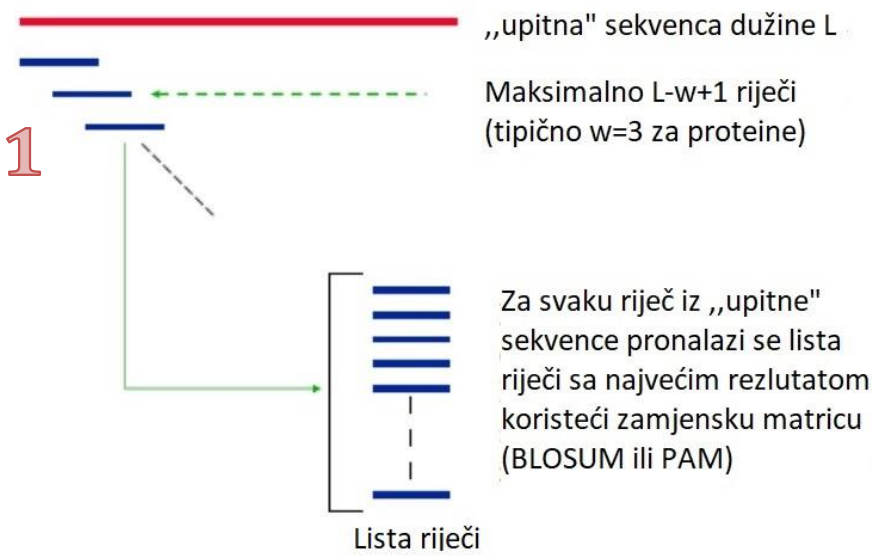
- Izbor odgovarajuće baze – ako BLAST prijavi previše poklapanja, možda ćete naići na ovaj problem tako što ćete pretražiti Sviss-Prot a ne NR (Sviss-Prot je 100 puta manji od NR)
- Unos u polje Organism – ako dobijete previse poklapanja, a zanima vas samo jedna vrsta organizma možete koristiti ovo polje za pretragu smao određenje protease ili za ignorisanje te proteaze. Da li ćemo pretraživati određenu proteasu ili je ignorisati zavisi od čekboksa exclude (isključi).

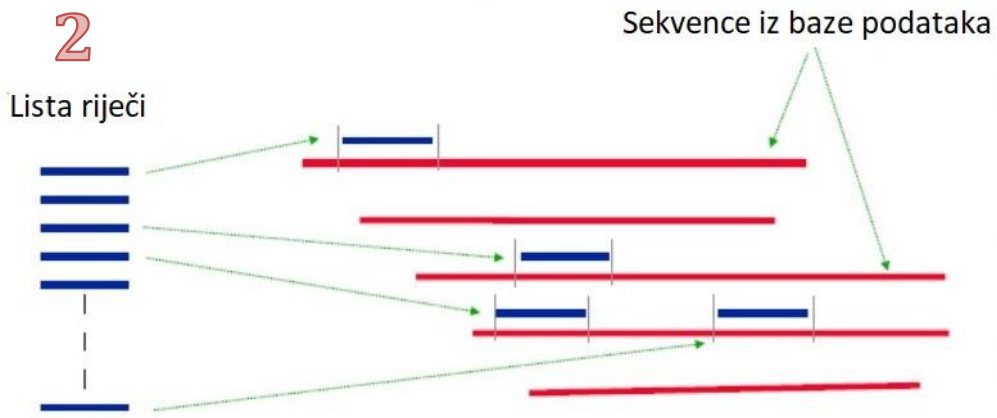


4. BLAST-ov algoritam

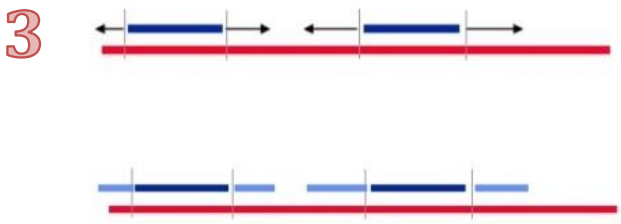
BLAST započinje pretragu indeksiranjem svih karaktera niske određene dužine unutar „upita“ prema njihovoj početnoj poziciji. Dužinu niza za indeksiranje „wordsize“, može da podesi korisnik. Dozvoljeni raspon za veličinu riječi varira u skladu sa BLAST programom koji se koristi, tipične vrednosti su 3 za pretragu sekvence protein-protein. Zatim BLAST skenira bazu podataka tražeći podudaranja između “riječi” indeksiranih u “upitu” i nizova pronađenih u nizovima baza podataka.

Podudaranja moraju biti tačna,za pretragu proteina-proteina, rezultat podudarnosti utvrđen pomoću zamjenske matrice, mora biti veći od određenog praga. Kad se nađe podudaranje riječi, BLAST pokušava produžiti i lijevo i desno od podudaranja kako bi se postigla poravnanja. BLAST će nastaviti ovo produženje sve dok se rezultat poravnanja nastavi povećavati ili dok ne padne za kritični iznos zbog negativnih rezultata koji su neusklađeni. Ova kritična količina poznata je i kao „dropoff“.



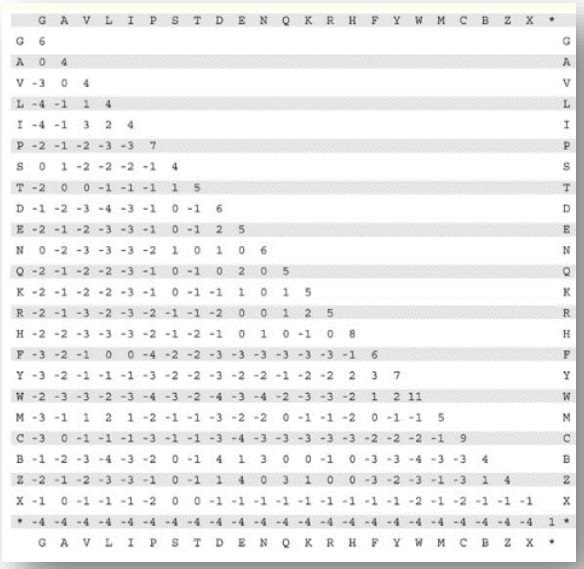


Tačna poklapanja riječi iz liste i sekvenci iz baze podataka



Za svako podudaranje riječi, poravnanje se produžavaju u oba pravca da bi se pronašla što bolja poravnanja

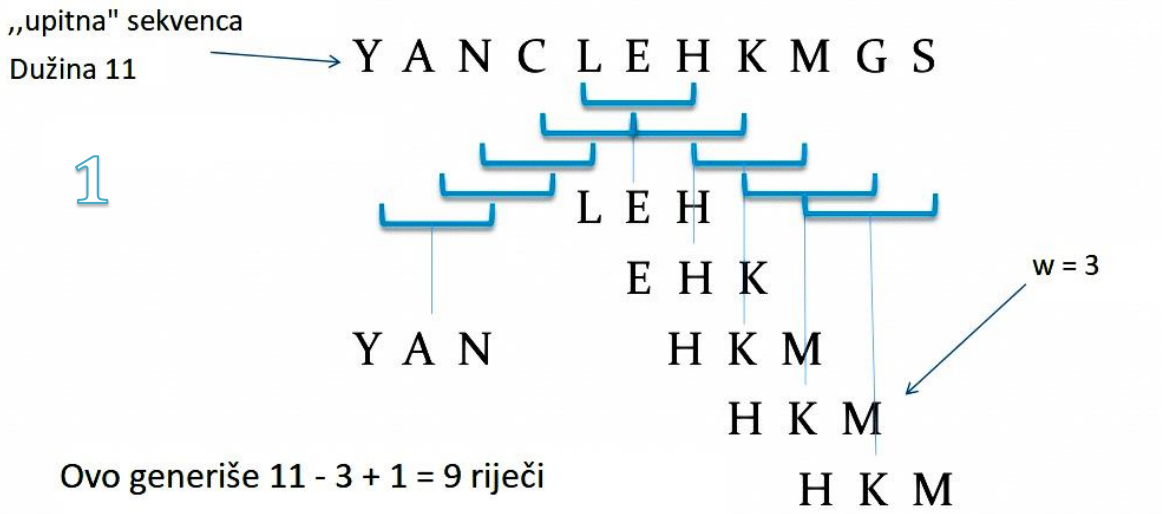
4.1. Bodovanje poravnanja i matrica zamjene



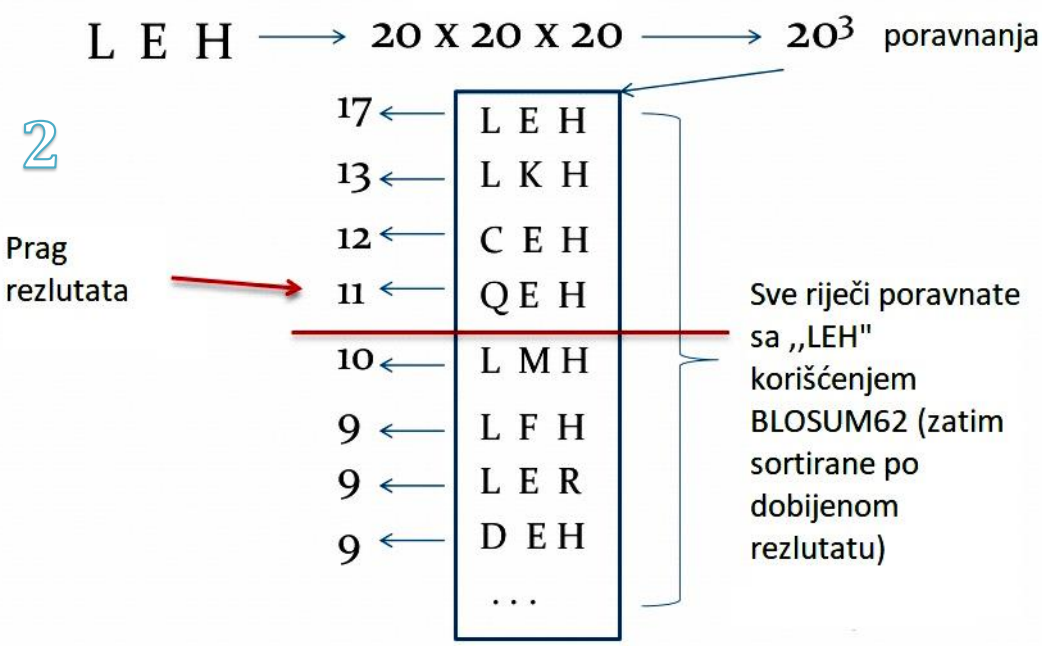
	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	
G	6																								G
A	0	4																							A
V	-3	0	4																						V
L	-4	-1	1	4																					L
I	-4	-1	3	2	4																				I
P	-2	-1	-2	-3	-3	7																			P
S	0	1	-2	-2	-2	-1	4																		S
T	-2	0	0	-1	-1	-1	1	5																	T
D	-1	-2	-3	-4	-3	-1	0	-1	6																D
E	-2	-1	-2	-3	-3	-1	0	-1	2	5															E
N	0	-2	-3	-3	-3	-2	1	0	1	0	6														N
Q	-2	-1	-2	-2	-3	-1	0	-1	0	2	0	5													Q
K	-2	-1	-2	-2	-3	-1	0	-1	-1	1	0	1	5												K
R	-2	-1	-3	-2	-3	-2	-1	-1	-2	0	0	1	2	5											R
H	-2	-2	-3	-3	-3	-2	-1	-2	-1	0	1	0	-1	0	8										H
F	-3	-2	-1	0	0	-4	-2	-2	-3	-3	-3	-3	-3	-3	-1	6									F
Y	-3	-2	-1	-1	-1	-3	-2	-2	-3	-2	-2	-1	-2	-2	2	3	7								Y
W	-2	-3	-3	-2	-3	-4	-3	-2	-4	-3	-4	-2	-3	-3	-2	1	2	11							W
M	-3	-1	1	2	1	-2	-1	-1	-3	-2	-2	0	-1	-1	-2	0	-1	-1	5						M
C	-3	0	-1	-1	-1	-3	-1	-1	-3	-4	-3	-3	-3	-3	-3	-2	-2	-2	-1	9					C
B	-1	-2	-3	-4	-3	-2	0	-1	4	1	3	0	0	-1	0	-3	-3	-4	-3	-3	4				B
Z	-2	-1	-2	-3	-3	-1	0	-1	1	4	0	3	1	0	0	-3	-2	-3	-1	-3	1	4			Z
X	-1	0	-1	-1	-1	-2	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	-2	-1	-1	-1		X
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1	*
	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*	

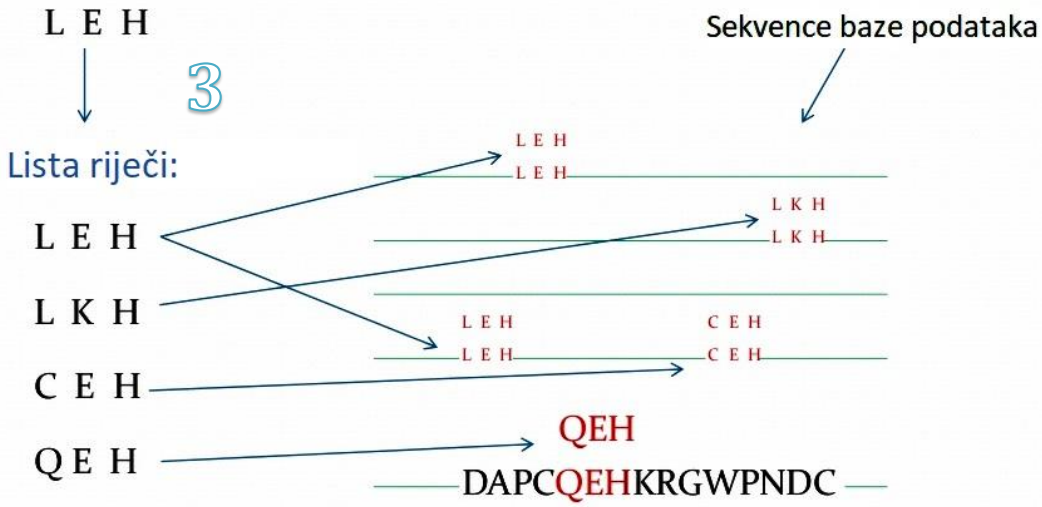
BLAST poravnavanje sastoji se od para sekvenci u kojima je svako slovo u jednom nizu upareno sa tačno jednim slovom ili prazninom u drugom. Rezultat poravnanja izračunava se dodjeljivanjem vrednosti svakom poravnatom paru slova, a zatim sabiranjem ovih vrednosti po dužini poravnanja. Za podešavanja poravnanja proteinskih sekvenci, ocjene za svaki mogući par slojeva aminokiselina su date u „zamjenskoj matrici“. Podrazumevano, BLAST koristi matricu „blosum62“, koja je najčešće korišćena serija zamjenskih matrica, međutim, dostupne su nam i druge matrice.

4.2. Primjer BLAST algoritma



Za svaku riječ za koju je w = 3 generišu se susjedne riječi korišćenjem BLOSUM62 matrica sa pragom 11





Tačna poklapanja riječi iz liste i sekvenci iz baze podataka

Upit = Y A N C L E H K M G S

Produžavanje u desnu stranu

Maksimalno opadanje(dropoff) rezlutata X = 2

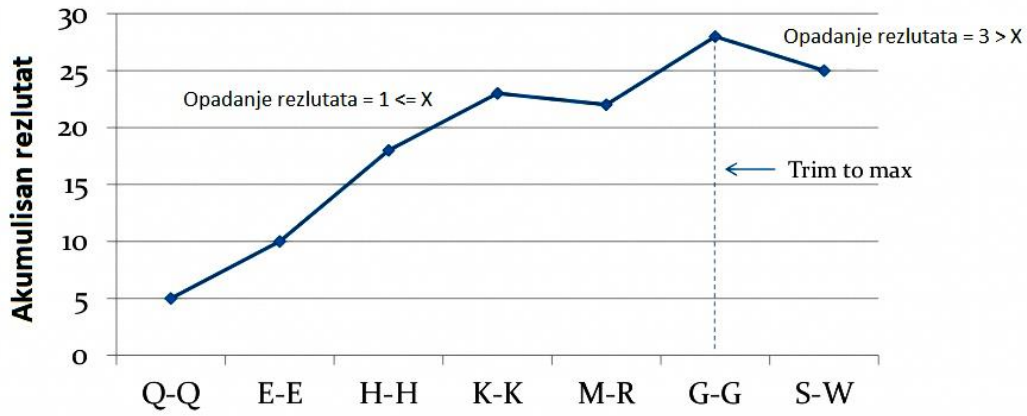
4

Q E H K M G S

D A P C Q E H K R G W P N D C

5 5 8 5 -1 6 -3

5 10 18 2322 28 25



Upit = Y A N C L E H K M G S

Produživanje u lijevu stranu

Maksimalno opadanje(dropoff) rezlutata X = 2

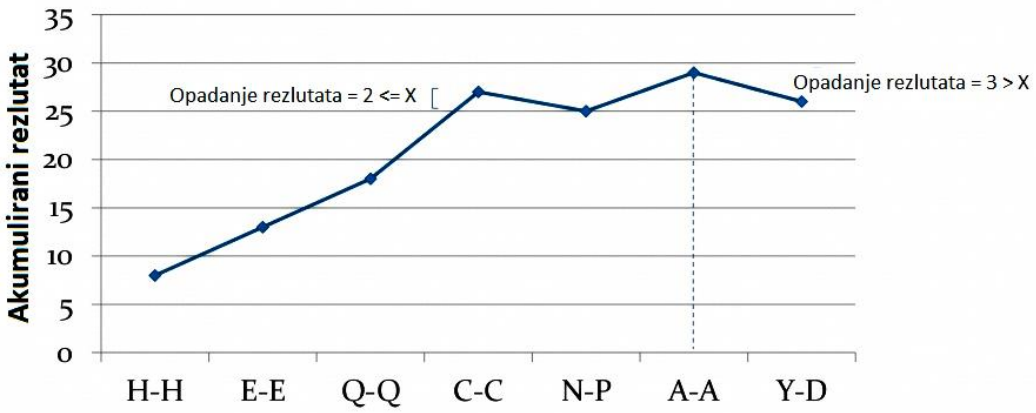
5

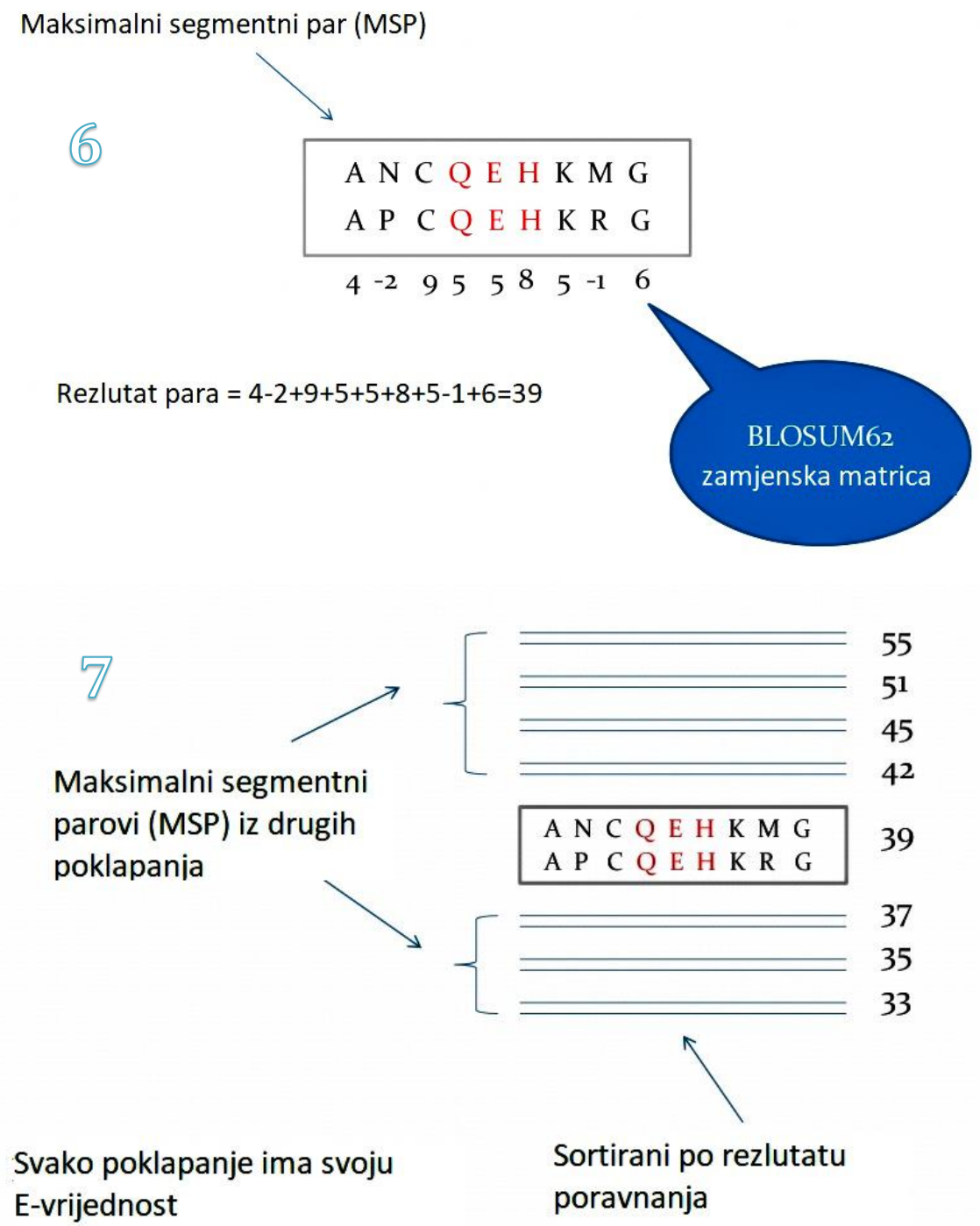
Y A N C Q E H K M G

D A P C Q E H K R G W P N D C

-3 4 -2 9 5 5 8

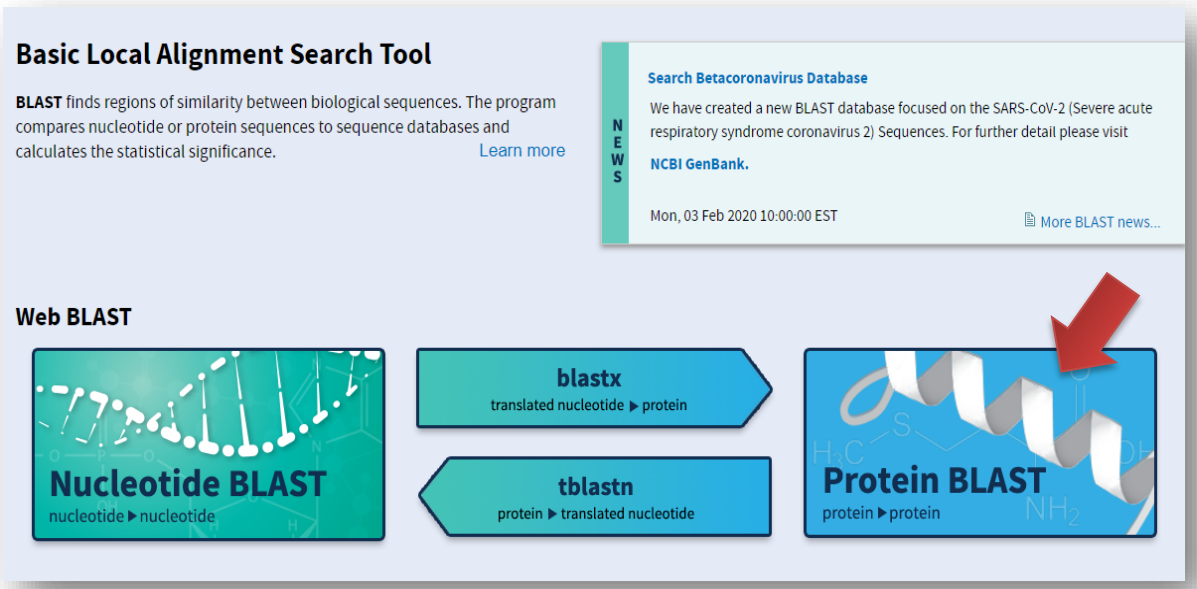
26 2925 27 18 13 8



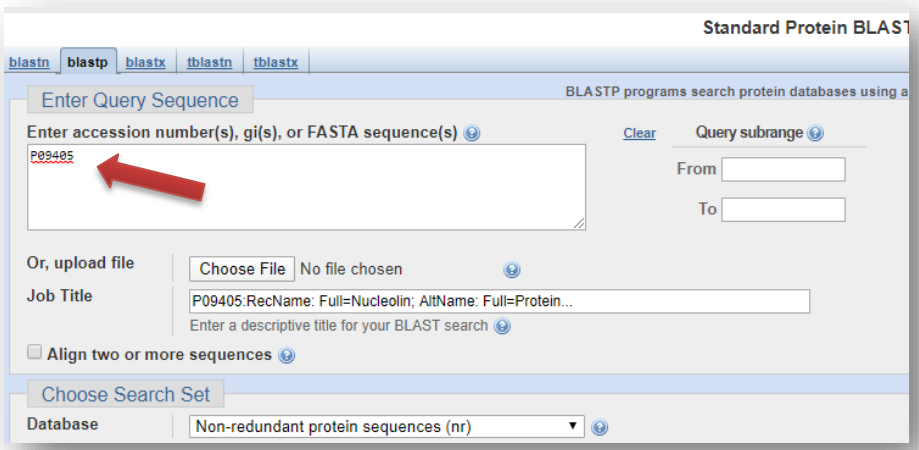

Maksimalni segmentni par (MSP)
6
A N C Q E H K M G
A P C Q E H K R G
4 -2 9 5 5 8 5 -1 6
Rezlutat para = 4-2+9+5+5+8+5-1+6=39
BLOSUM62
zamjenska matrica
7
Maksimalni segmentni parovi (MSP) iz drugih poklapanja
55
51
45
42
A N C Q E H K M G
A P C Q E H K R G
39
37
35
33
Svako poklapanje ima svoju E-vrijednost
Sortirani po rezlutatu poravnanja

Primjer1. Uz pomoć BLAST-a gledamo postoje li proteini koji su slični nukleolinu hrčka:

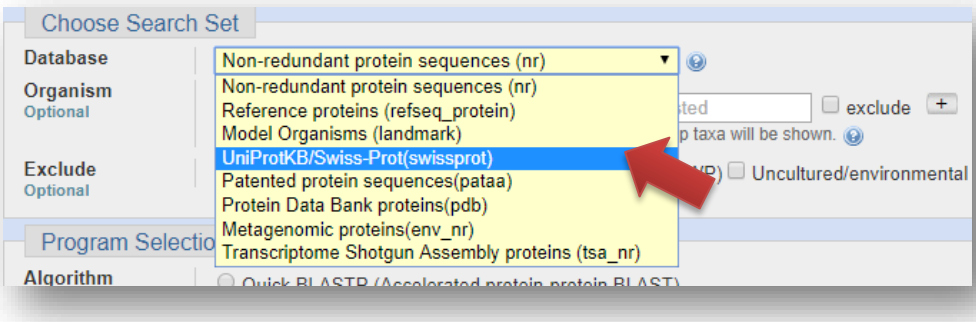
1. Otvaramo stranicu https://blast.ncbi.nlm.nih.gov/Blast.cgi
2. Biramo opciju proteinskog BLAST-a (blastp)



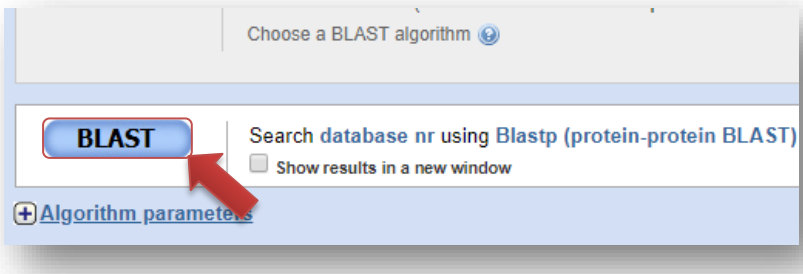
3. U search prozor unosimo pristupni broj, gi broj ili sekvencu u FASTA formatu, u našem slučaju unosimo pristupni broj P09405.



4. Biramo UniProtKB/Swiss-Prot iz Database menija.



5. U dnu stranice kliknemo na BLAST i sačekamo.



5. *Razumijevanje rezultata*

Sve BLAST verzije vraćaju sličan rezultat. Rezultat se sastoji iz 4 dijela kod većine BLAST servera. Ovi dijelovi se uvijek pojavljuju istim redosledom i to:

1. Grafički prikaz: prikazuje koje sekvence liče našoj. U zavisnosti od server koji koristimo, ovaj prikaz se veoma razlikuje.

2. Lista rezultata: imena sekvenci koje su slične našoj (poredane po sličnosti)

3. Poravnavanje sekvenci: svako poravnanje između naše sekvence i sekvence iz baze podataka

4. Parametri: Lista različitih parametara koji se koriste za pretragu.

Svaki od ovih elemenata sadrži puno informacija. Znajući šta je važno na svakom od ovih prikazivanja, može biti korisno za vaše istraživanje.

5.1. Grafički prikaz

Ova sekcija omogućava vizualizaciju.Na vrhu se nalazi vaša sekvenca, a svaka traka predstavlja dio druge sekvence koja je slična vašoj i region u kojem se sličnost javlja. Crvene trake predstavljaju najsličnije sekvence, ružičaste trake ukazuju na malo lošija poklapanja, zelene trake predstavljaju rezultate koji nisu impresivni. Ove boje uglavnom predstavljaju dobre rezultate. Plave i crne trake predstavljaju najlošije rezultate. Crne trake predstavljaju proteine koji imaju tako malo zajedničkog s vašom sekvencom da njihovo poravnanje vjerovatno nema biološkog značaja (spadaju u „zonu sumraka“).

Prednost grafičkog prikaza rezultata jeste što nam dozvoljava da vidimo kolikom se dužinom date sekvence poklapaju, što je velika pomoć prilikom otkrivanje proteinskih domena.

Grafički prikaz rezultata BLAST-a je interaktivan: prelaskom mišem preko traka dobiva se ime odgovarajuće sekvence. Tanka crna linija koja povezuje dvije trake znači da se isti protein poklapa sa sekvencom, ali na dvije odvojene lokacije.

5.2. Lista rezultata

Lista BLAST rezultata daje potrebne informacije na osnovu kojih mogu odlučiti da li naša sekvenca liči na neku sekvencu iz baze. Svaka linija sadrži sljedeće informacije:

- **Opis sekvence**
- **Max score** – statistička mjera poklapanja, koja se izračunava na osnovu homologije i dužine. Što je ova vrijednost viša to su dvije sekvence sličnije.
- **Total score** –ukupno poravnata sekvenca
- **Query coverage** – dužina poravnate sekvence izražena u procentima
- **E-vrijednost (*E-value*)** – najvažnija mjera statističke značajnosti. Što je E vrijednost manja, to su sekvence sličnije i vjerovatnije je da se radi o homolognim sekvencama. Sekvence iz baze podataka koje su identične traženoj sekvenci imaju E vrijednost 0. Rezultati preko 0,001 ne ukazuju na homologiju.
- **Procentni identitet** – pokazuje u kojoj mjeri dvije imaju iste rezidue na istim pozicijama u poravnanju, i često se izražava kao postotak.
- **Pristupni broj sekvence**

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
RecName: Full=Nucleolin; AltName: Full=Protein C23 [Mus musculus]	1379	1379	100%	0.0	100.00%	P09405.2
RecName: Full=Nucleolin; AltName: Full=Protein C23 [Rattus norvegicus]	941	941	91%	0.0	90.84%	P13383.3
RecName: Full=Nucleolin; AltName: Full=Protein C23 [Mesocricetus auratus]	919	919	91%	0.0	85.89%	P08199.2
RecName: Full=Nucleolin [Macaca fascicularis]	768	768	91%	0.0	80.09%	Q4R4J7.3
RecName: Full=Nucleolin; AltName: Full=Protein C23 [Homo sapiens]	657	657	60%	0.0	83.22%	P19338.3
RecName: Full=Nucleolin; AltName: Full=Protein C23 [Gallus gallus]	448	448	53%	2e-147	63.59%	P15771.1
RecName: Full=Nucleolin; AltName: Full=Protein C23 [Xenopus laevis]	391	391	52%	8e-126	57.85%	P20397.3
RecName: Full=Nucleolin [Pongo abelii]	136	136	20%	3e-32	82.39%	Q5RF26.3
RecName: Full=Polyadenylate-binding protein 4; Short=PABP-4; Short=Poly(A)-binding protein 4; AltName: Full=Activated-platelet protein 1; Sl	104	104	41%	5e-22	28.79%	Q13310.1
RecName: Full=Nucleolin 2; AltName: Full=Protein NUCLEOLIN LIKE 2; Short=AtNUC-L2; AltName: Full=Protein PARALLEL LIKE 1; Short=At	103	166	53%	1e-21	34.55%	Q1PEP5.1
RecName: Full=Polyadenylate-binding protein 1A; Short=PABP-1A; Short=Poly(A)-binding protein 1A [Danio rerio]	100	230	48%	7e-21	27.05%	F1QB54.1
RecName: Full=Polyadenylate-binding protein 1; Short=PABP-1; Short=Poly(A)-binding protein 1 [Mus musculus]	100	227	48%	1e-20	27.32%	P29341.2
RecName: Full=Polyadenylate-binding protein 1; Short=PABP-1; Short=Poly(A)-binding protein 1 [Rattus norvegicus]	100	226	48%	1e-20	27.32%	Q9EPH8.1
RecName: Full=Polyadenylate-binding protein 1-B; Short=PABP-1-B; Short=Poly(A)-binding protein 1-B; Short=xPABP1-B; AltName: Full=Cyto	99.4	218	48%	2e-20	26.78%	Q6IP09.1
RecName: Full=Polyadenylate-binding protein 1; Short=PABP-1; Short=Poly(A)-binding protein 1 [Pongo abelii]	99.0	221	48%	2e-20	27.05%	Q5R8F7.1
RecName: Full=Polyadenylate-binding protein 1; Short=PABP-1; Short=Poly(A)-binding protein 1 [Homo sapiens]	98.6	222	48%	3e-20	27.05%	P11940.2

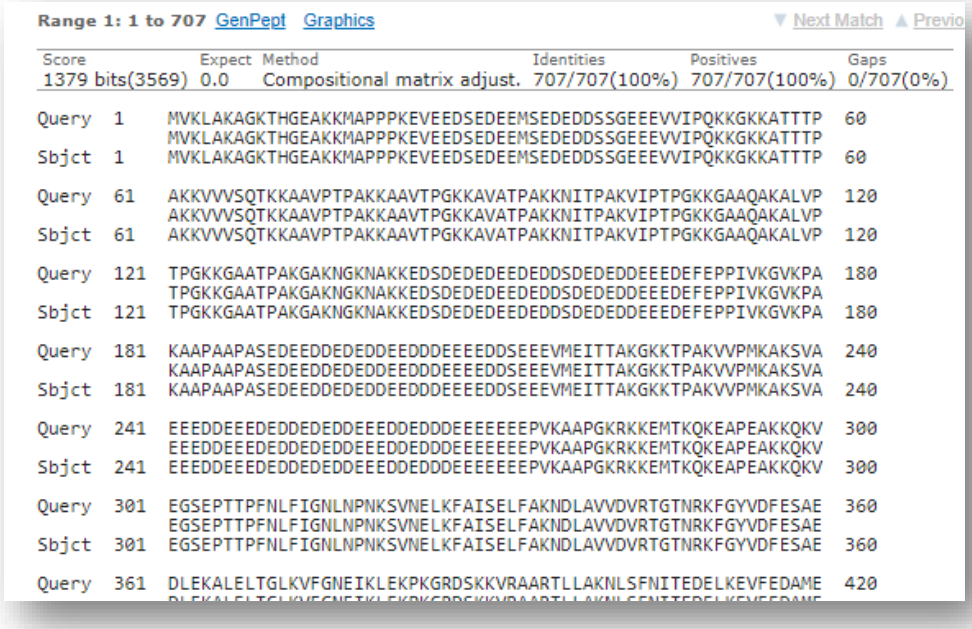
5.3. Poravnanje (alignment)

Kada treba donijeti konačan zaključak, biolozi vjeruju poravnanju i smatraju kako ono ne daje lažne rezultate. Što je u većini slučajeva tačno ako znate da ga čitate i tumačite.

Svaki prikaz BLAST poravnanja sadrži sljedeće informacije:

- **Dužina (Range)** – predstavlja dužinu poravnanja koja pokazuje koliko su dugi segmenti dvije sekvence koje je BLAST poravnao. To je najbolji način da se vidi poklapanje.

- **Rezultat (score)** –broj koji se koristi za procjenu biološke važnosti pretrage
- **Identiteti** – broj identičnih rezidua podijeljen s brojem poravnatih rezidua
- **Praznine (Gaps)** – ukazuje na rezidue koji nisu mogli biti poravnati
- **Positives** – mjera identičnih ili sličnih rezidua
- **Query (gornja sekvenca)** – naša sekvenca
- **Sbjct(donja sekvenca)** – sekvenca iz baze podataka
- **Linija između sekvenci** – sadrži slovo za identične rezidue, znak + za slične aminokiseline ili razmak za pogrešno sparivanje.
- **XXXXX regioni** označavaju regione koji sadrže mnogo identičnih rezidua. S obzirom da mogu prouzrokovati probleme prilikom pretrage, BLAST ih automatski maskira i to samo u vašoj sekvenci.



Range 1: 1 to 707 GenPept Graphics ▼ Next Match ▲ Previo

Score	Expect	Method	Identities	Positives	Gaps
1379 bits(3569)	0.0	Compositional matrix adjust.	707/707(100%)	707/707(100%)	0/707(0%)

```
Query  1    MVKLAKAGKTHGEAKKMAPPPKEVEEDSEDEEMSEDEDDSSGEEEVVIPQKKGKKATTTP  60
            MVKLAKAGKTHGEAKKMAPPPKEVEEDSEDEEMSEDEDDSSGEEEVVIPQKKGKKATTTP
Sbjct  1    MVKLAKAGKTHGEAKKMAPPPKEVEEDSEDEEMSEDEDDSSGEEEVVIPQKKGKKATTTP  60

Query  61   AKKVVVSQTKKAAVPTPAKKAAVTPGKKAVATPAKKNITPAKVIPTPGKKGAAQAKALVP  120
            AKKVVVSQTKKAAVPTPAKKAAVTPGKKAVATPAKKNITPAKVIPTPGKKGAAQAKALVP
Sbjct  61   AKKVVVSQTKKAAVPTPAKKAAVTPGKKAVATPAKKNITPAKVIPTPGKKGAAQAKALVP  120

Query  121  TPGKKGAATPAKGAKNGKNAKKEDSDEDEDEEDEDDSDEDEDDEEEDEFEPPIVKGVKPA  180
            TPGKKGAATPAKGAKNGKNAKKEDSDEDEDEEDEDDSDEDEDDEEEDEFEPPIVKGVKPA
Sbjct  121  TPGKKGAATPAKGAKNGKNAKKEDSDEDEDEEDEDDSDEDEDDEEEDEFEPPIVKGVKPA  180

Query  181  KAAPAAPASEDEEDDEDEDDEEDDDEEEEDDSEEEVMEITTAKGKKTPAKVVPMKAKSVA  240
            KAAPAAPASEDEEDDEDEDDEEDDDEEEEDDSEEEVMEITTAKGKKTPAKVVPMKAKSVA
Sbjct  181  KAAPAAPASEDEEDDEDEDDEEDDDEEEEDDSEEEVMEITTAKGKKTPAKVVPMKAKSVA  240

Query  241  EEEDDEEEDEDDEDEDDEEEDDEDDDEEEEEEEPVKAAPGKRKKEMTKQKEAPEAKKQKV  300
            EEEDDEEEDEDDEDEDDEEEDDEDDDEEEEEEEPVKAAPGKRKKEMTKQKEAPEAKKQKV
Sbjct  241  EEEDDEEEDEDDEDEDDEEEDDEDDDEEEEEEEPVKAAPGKRKKEMTKQKEAPEAKKQKV  300

Query  301  EGSEPTTPFNLFIGNLNPNKSVNELKFAISELFAKNDLAVVDVRTGTNRKFGYVDFESAE  360
            EGSEPTTPFNLFIGNLNPNKSVNELKFAISELFAKNDLAVVDVRTGTNRKFGYVDFESAE
Sbjct  301  EGSEPTTPFNLFIGNLNPNKSVNELKFAISELFAKNDLAVVDVRTGTNRKFGYVDFESAE  360

Query  361  DLEKALELTGLKVFGNEIKLEKPKGRDSKKVRAARTLLAKNLSFNITEDELKEVFEDAME  420
```

5.4. Parametri

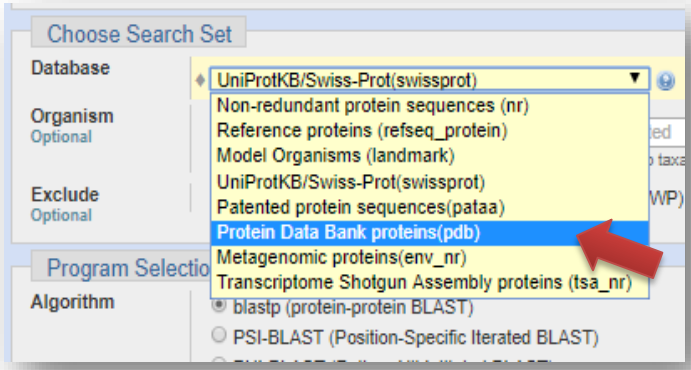
Ne morate se brinuti puno o značenju informacija koje se nalaze pri dnu stranice rezultata. Ako ste promjenili podrazumijevane parametre BLAST-a, ovaj dio prati to za vas.

Niko ne može garantovati da ćete za dva skreniranja dobiti isti rezultat, čak i ako koristite isti BLAST server. Nadogradnja bilo koje komponente na serveru može da modifikuje rezultate pretrage. Komponente koje se mogu nadograditi uključuju:

- Baze podataka
- BLAST program
- Zadani parametri na serveru

Primjer2. Korišćenje principa homologije da vidimo kako bi naš protein, kojeg smo već ranije sekvencirali, mogao da izgleda u trodimenzionalnom prostoru. Recimo da smo već sekvencirali protein miostatin vodenog goveda (Bubalus bubalis), u ovom primjeru koristićemo sekvencu sa pristupnim brojem AAW50584 u GenBank bazi.

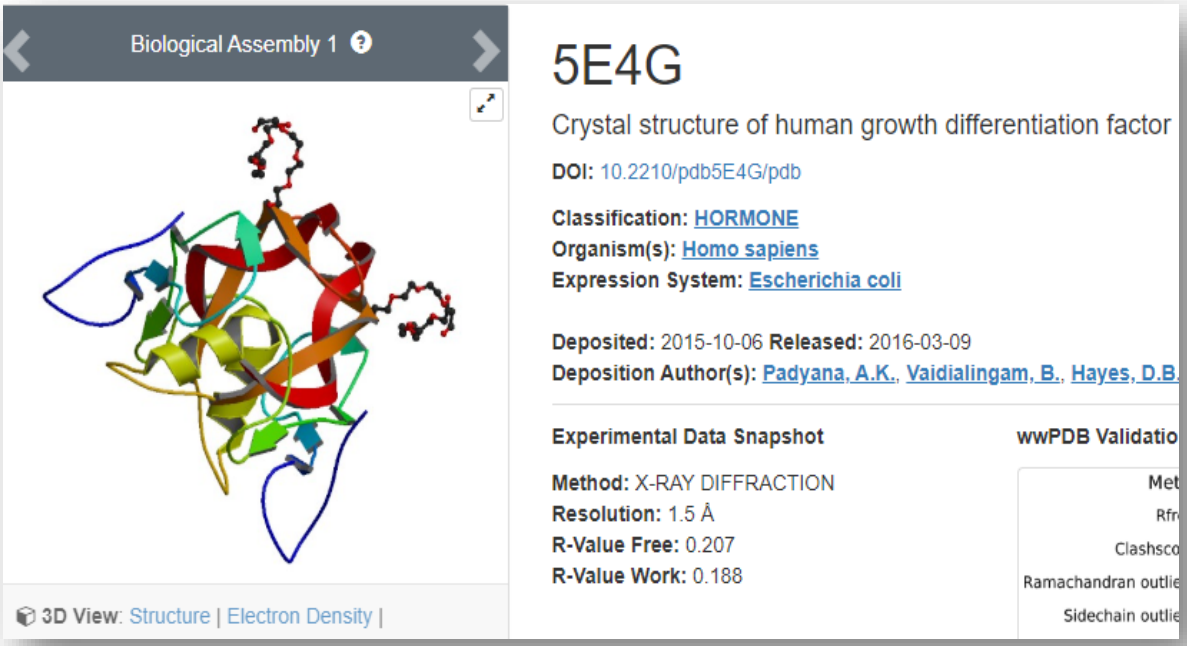
1. Otvaramo stranicu https://blast.ncbi.nlm.nih.gov/Blast.cgi
2. Biramo opciju preteinskog BLAST-a (blastp)
3. U search prozor unosimo pristupni broj AAW50584.
4. Biramo pdb iz Database menija.



5. U dnu stranice kliknemo na dugme BLAST i sačekamo.

Najbolji rezultat koji smo dobili ima dužinsko poravnanje od 93% i ukoliko slijednimo

pristupni broj ove proteinske strukture u PDB bezi podataka možemo dobiti ideju kakva bi mogla biti 3D struktura našeg proteina.



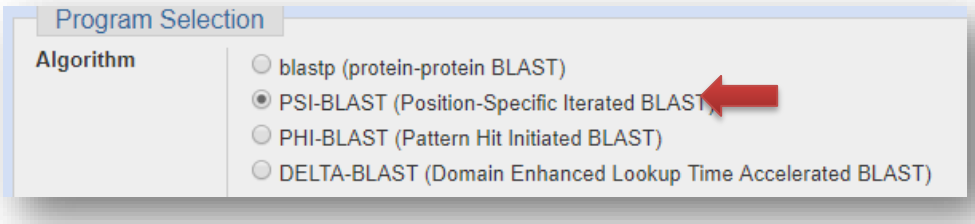
6. *Omogućavanje više iteracija BLAST-a*

Ponekad BLAST nije dovoljan. Zamislite da želimo da pronađemo sve članove proteinske porodice, počevši od naše proteinske sekvence. Upotrebom BLAST-a mi pronalazimo samo blisko povezane sekvence. Ali kako da pronađemo udaljene sekvence koje su takođe članovi te porodice, ali ih naša pretraga nije pronašla? To je upravo ono šta PSI-BLAST radi. PSI-BLAST (Position-Specified Iterated BLAST) prvo pranalazi sekvencu koja je blisko povezana sa našom, a onda poredi taj opšti profil sa bazom podataka i pronazi veću grupu proteina. Iz ovih proteina izvodi sljedeći profil i tako dalje. Pored PSI-BLAST-a koristi se i PHI-BLAST (Pattern Hit Initialized BLAST) koji je pogodan za pronalaženje proteina sa specifičnim motivom, međutim on nam ne omogućava pronalaženje udaljenih sekvenci ne koje sadrže takav motiv.

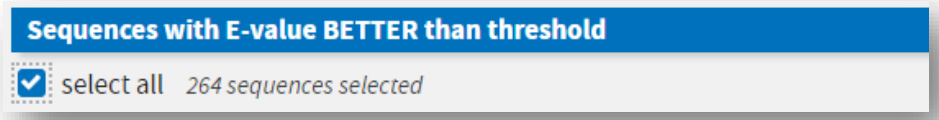
Primjer3. Korišćenje PSI-BLAST-a za pronalaženje članova porodice

1. Otvaramo stranicu https://blast.ncbi.nlm.nih.gov/Blast.cgi
2. Biramo opciju proteinskog BLAST-a (blastp).
3. U search prozor unosimo pristupni broj P09405.

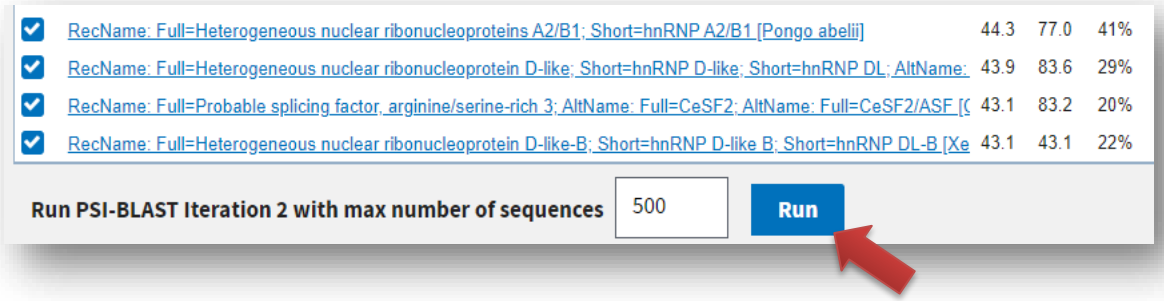
4. Biramo UniProtKB/Swiss-Prot iz Database menija.
5. Za algoritam sada biramo PSI-BLAST (Position-Specific Iterated BLAST), umjesto blastp (protein-protein BLAST) koji smo do sada koristili.



6. U dnu stranice kliknemo na dugme BLAST i sačekamo.
7. Trenutni rezultati predstavljaju blisko povezane sekvence, sada biramo koje od tih sekvenci PSI-BLAST uzima za drugu interaciju.



8. Zatim biramo broj maksimalni sekvenci koje PSI-BLAST uzima i kliknemo na dugme RUN da pokrenemo drugu interaciju.



9. Nakon određenog vremena prikazuju nam se rezultati druge iteracije PSI-BLAST-a koji uključuju i nove sekvence, koje se nisu pojavile prilikom prve iteracije.Da bi se razlikovale, nove sekvence su markirane žutom bojom.

☑	RecName: Full=ELAV-like protein 1-A; AltName: Full=36 kDa embryonic-type cytoplasmic polyadenylation element-	183	679	55%	7e-51	17.67%	Q1JQ73.1	☑	✓	
☑	RecName: Full=31 kDa ribonucleoprotein, chloroplastic; Flags: Precursor [Nicotiana sylvestris]	179	653	65%	9e-50	23.86%	P19683.1	☑	✓	
☑	RecName: Full=Polyadenylate-binding protein 1-like 2; AltName: Full=RNA-binding motif protein 32; AltName: Full=	171	369	57%	3e-48	17.68%	Q5JQF8.1	☑		✓
☑	RecName: Full=APOBEC1 complementation factor; AltName: Full=APOBEC1-stimulating protein [Pongo abelii]	182	327	53%	3e-48	20.51%	Q5R9H4.1	☑		✓
☑	RecName: Full=28 kDa ribonucleoprotein, chloroplastic; Short=28RNP; Flags: Precursor [Nicotiana sylvestris]	173	651	65%	4e-48	22.58%	P19682.1	☑	✓	
☑	RecName: Full=APOBEC1 complementation factor; AltName: Full=APOBEC1-stimulating protein [Homo sapiens]	180	322	53%	2e-47	20.79%	Q9NQ94.1	☑		✓
☑	RecName: Full=Splicing factor 3B subunit 4 [Rattus norvegicus]	175	638	50%	4e-47	25.84%	Q6AYL5.1	☑	✓	
☑	RecName: Full=Polyadenylate-binding protein RBP45C; Short=Poly(A)-binding protein RBP45C; AltName: Full=RN	174	711	58%	6e-47	22.65%	Q93W34.1	☑	✓	
☑	RecName: Full=28 kDa ribonucleoprotein, chloroplastic; Short=28RNP [Spinacia oleracea]	169	637	69%	7e-47	25.93%	P28644.1	☑	✓	
☑	RecName: Full=Splicing factor 3B subunit 4; AltName: Full=Pre-mRNA-splicing factor SF3b 49 kDa subunit; AltNam	174	635	50%	7e-47	25.84%	Q15427.1	☑	✓	
☑	RecName: Full=30 kDa ribonucleoprotein, chloroplastic; AltName: Full=CP-RBP30; Flags: Precursor [Nicotiana plur	169	641	56%	2e-46	22.99%	P49313.1	☑	✓	
☑	RecName: Full=Nuclear and cytoplasmic polyadenylated RNA-binding protein PUB1; AltName: Full=ARS consensu	174	792	72%	3e-46	26.26%	P32588.4	☑	✓	
☑	RecName: Full=29 kDa ribonucleoprotein A, chloroplastic; AltName: Full=CP29A; Flags: Precursor [Nicotiana sylve	168	635	64%	4e-46	21.89%	Q08935.1	☑	✓	
☑	RecName: Full=RNA-binding protein CP31B, chloroplastic; Flags: Precursor [Arabidopsis thaliana]	168	542	63%	5e-46	25.48%	Q9FGS0.1	☑	✓	
☑	RecName: Full=Probable RNA-binding protein 46; AltName: Full=Cancer/testis antigen 68; Short=CT68; AltName: F	174	435	47%	7e-46	22.26%	Q8TBY0.1	☑		✓
☑	RecName: Full=Probable RNA-binding protein 46; AltName: Full=RNA-binding motif protein 46 [Macaca fascicularis	172	430	47%	2e-45	22.33%	Q4R2Z0.2	☑		✓
☑	RecName: Full=RNA-binding protein 47; AltName: Full=RNA-binding motif protein 47 [Mus musculus]	174	309	54%	3e-45	22.08%	Q91WT8.1	☑		✓
☑	RecName: Full=RNA-binding protein CP29B, chloroplastic; AltName: Full=Ribonucleoprotein At2g37220; Flags: Pre	166	546	49%	5e-45	18.64%	Q9ZUU4.1	☑	✓	
☑	RecName: Full=Probable RNA-binding protein 46; AltName: Full=RNA-binding motif protein 46 [Mus musculus]	172	430	47%	5e-45	22.33%	P86049.1	☑		✓
☑	RecName: Full=RNA-binding protein 47; AltName: Full=RNA-binding motif protein 47 [Rattus norvegicus]	172	307	54%	7e-45	22.08%	Q66H68.1	☑		✓
☑	RecName: Full=Sex-lethal homolog [Megaselia scalaris]	165	447	60%	1e-44	23.66%	O01671.3	☑	✓	

10. Ponavljamo korake 8. i 9. potreban broj puta.

7. Otkrivanje proteinskih domena sa BLAS-om i PSI-BLAST-om

Možemo pronalaziti poznate proteinske domene koristeći CD server (također poznat kao obrnuti-PSI-BLAST ili rps-blast). Možemo čak i pomoću BLAST-a da otkriti domene i upotrijebiti ih za skeniranje baza podataka. U BLAST-u, domena se naziva PSSM (Matrica za zamjenu specifične pozicije).

Kad god pokrenete PSI-BLAST, svoje rezultate možete prikazati u obliku PSSM-a. Jednom kada je vaš PSSM spreman, sve što trebate je isjeći i zalepiti ga u odgovarajući odeljak naprednog BLAST parametra.

Kada pružite PSSM, BLAST će ga koristiti umjesto jedne sekvence upita - i uporediće ga sa svakim nizom u bazi koju odaberete.

8. Literatura

- Claverie, J-M., Notredame, C. (2007): Bioinformatics for dummies, 2. izdanje. Wiley Publishing, Hoboken, USA.

- Grzegorz M. Boratyn, Christiam Camacho, Peter S. Cooper, George Coulouris, Amelia Fong, Ning Ma, Thomas L. Madden, Wayne T. Matten, Scott D. McGinnis, Yuri Merezhuk, Yan Raytselis, Eric W. Sayers, Tao Tao, Jian Ye, and Irena Zaretskaya (2013): BLAST: a more efficient report with usability improvements. Nucleic Acids Research, 41: W29-W33.

- Belma Kalamujić Stroil, Semir Dorić, Lada Lukić Bilela, Naris Pojskić (2018): Aplikativna bioinformatika – Praktikum