SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Diplomski seminar

Evolucija i uspoređivanje sljedova proteina

Ogrizek-Tomaš Mario
Mentor: Dr. Sc. Mile Šikić

Zagreb, svibanj 2011.

Sadržaj

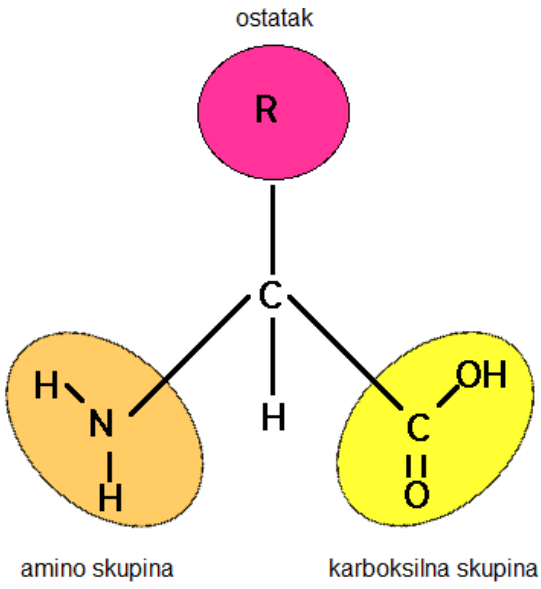
1. Uvod

Ubrzan razvoj tehnika molekularnog kloniranja, metoda sekvencioniranja DNA, algoritama za usporedbu sljedova, te samih računala uzrokovao je porast u važnosti usporedbe bioloških sljedova u molekularnoj biologiji. Kao posljedica, uloga podataka o sljedovima proteina se promijenila. Prije trideset godina, određivanje slijeda aminokiselina nekog proteina bila je posljednja stepenica pri njegovoj karakterizaciji. Danas, proces se okrenuo. Uobičajena je praksa kloniranje i sekvencioniranje gena kako bi otkrili povezanosti s nekom bolesti, ili za razvoj seruma, ljekova i slično. To je ujedno i osnovna premisa *"Human genome"* projekta: sekvencioniranje svih gena organizma, te određivanje njihovih funkcija pomoću analize sljedova. Usporedba sljedova proteina je snažna metoda karakterizacije proteina. Razlog tomu je golema količina informacije sačuvane kroz evoluciju. Za mnoge sljedove, evolucijska povijest se može pratiti jednu do dvije milijarde godina. Cilj ovog seminara je proučiti građu i strukturu proteina, shvatiti povezanost i utjecaj evolucije na njih, te vidjeti na koji način pomoću analize njihovih sljedova zaključiti nešto o njihovoj strukturi i evolucijskoj povijesti.

2. Proteini

2.1 Aminokiseline

Proteini su biokemijski spojevi sastavljeni od jednog ili više polipeptida savijenih u globularni ili vlaknasti oblik. Polipeptid je dugi lanac aminokiselina povezanih peptidnim vezama. Aminokiselina se sastoji od karboksilne grupe, amino grupe i R grupe koja im daje određene karakteristike. Te karakteristike određuju interakcije između atoma i molekula.



Slika 1 Struktura aminokiseline
Izvor: Wissman, 2007.

Aminokiseline se vežu u dugi lanac čvrstim peptidnim vezama što čini primarnu strukturu proteina. Uz primarnu strukturu, aminokiseline kao posljedica R ostatka, međusobno stvaraju dodatne slabe veze. To oblikuje trodimenzionalne strukture proteina, tj. sekundarnu, tercijarnu i kvartarnu strukturu.
Postoji 20 aminokiselina koje se često pojavljuju u prirodi. Svaka od njih posjeduje svojstva (definirana tipom R ostatka) koja imaju određenu ulogu u strukturi proteina.

2.2 Struktura proteina

Danski kemičar K.U.Linderstrøm-Lang (Linderstrøm-Lang, 1952.)dao je opis hijerarhijske prirode arhitekture proteina:

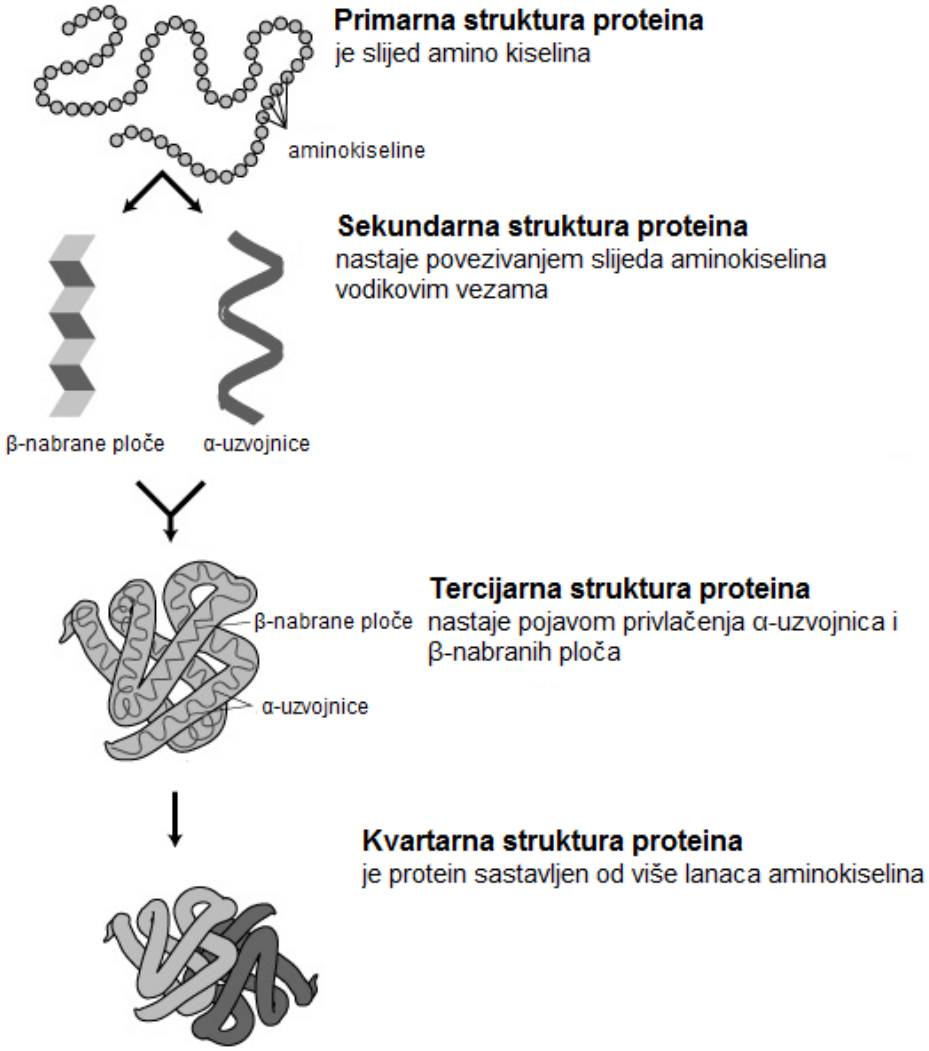
1. Primarna struktura
   poredak aminokiselina u lancu - skup primarnih kemijskih veza (peptidne veze)

2. Sekundarna struktura
   pridjeljivanje α-uzvojnice i β-nabrane ploče, tj. povezivanje dijelova lanca slabijim vodikovim vezama

3. Tercijarna struktura
   slaganje sekundarnih struktura u prostoru

4. Kvartarna struktura

slaganje proteinskih pod jedinica u jedinstvenu molekulu proteina



Slika 2 Strukture proteina
Izvor: Martz, 2009.

Proteini su zaslužni za regulaciju različitih aktivnosti svih dosad poznatih organizama. Općenito, odgovorni su za upravljanje stanicom (npr. replikacija genetskog materijala, prenošenja kisika) te određivanje fenotipa. Sve funkcije postižu svojom trodimenzionalnom strukturom, tj. tercijarnim i kvartarnim interakcijama. Funkcionalna svojstva proteina, dakle, ovise o njihovoj trodimenzionalnoj strukturi.

Niz aminokiselina proteina određuje njegovu trodimenzionalnu strukturu. Kada se nađu u mediju odgovarajućih uvjeta, proteini se spontano smataju u svoja prirodna, aktivna stanja. Ukoliko bi nizovi aminokiselina sadržavali dovoljno informacija, mogli bi pomoću njih odrediti trodimenzionalnu

strukturu proteina. Trebalo bi biti moguće osmisliti algoritam kojim bi se moglo *a priori* predvidjeti struktura proteina iz njegovog niza aminokiselina.

Taj problem se pokazao previše složenim pa su znanstvenici podijelili problem na dijelove:

- *Predviđanje sekundarne strukture:*
  Koji segmenti niza sačinjavaju α-uzvojnice i β-nabrane ploče?

- *Raspoznavanje smatanja:*
  Uz biblioteku poznatih proteinskih struktura i njihovih nizova aminokiselina, pokušati pronaći strukturu proteina (kojem poznajemo samo niz aminokiselina) za koju je vjerojatnost sličnog obrasca smatanja najveća.

- *Modeliranje homologije:*
  Pretpostavimo da je ciljni protein (kojem poznajemo niz aminokiselina, no ne i strukturu) homologan jednom ili više proteina poznate strukture. Očekujemo da bi većina strukture ciljnog proteina nalikovala na onu poznatog homolognog proteina. To može činiti osnovu za predviđanje strukture ciljanoga proteina.

3. Evolucija

3.1 Biološka klasifikacija i nomenklatura

U osamnaestom stoljeću, biološka nomenklatura bazirala se na ideji kako su živući organizmi podijeljeni u grupe, znane kao vrste - grupe sličnih organizama sa zajedničkim skupom gena. Švedski naturalist, Carl Linnaeus, podijelio je živa bića hijerarhijski u kategorije (taksone) (Linnaeus, 1735.):

1. Carstvo (eng. *Kingdom*)

2. Koljeno (lat. *Phylum*)
3. Razred (eng. *Class*)
4. Red (eng. *Order*)
5. Porodica (eng. *Family*)
6. Rod (lat. *Genus*)
7. Vrsta (eng. *Species*)

Kako bi odredili o kojoj je vrsti riječ, dovoljno je navesti rod i vrstu, primjerice *Homo sapiens* za čovjeka ili *Drosophila melanogaster* za mušicu. U početku, Linnaeanov sustav je bio baziran na vidljivim sličnostima. Otkrićem evolucije, uvidjelo se da sustav u velikoj mjeri odražava biološko nasljedstvo.

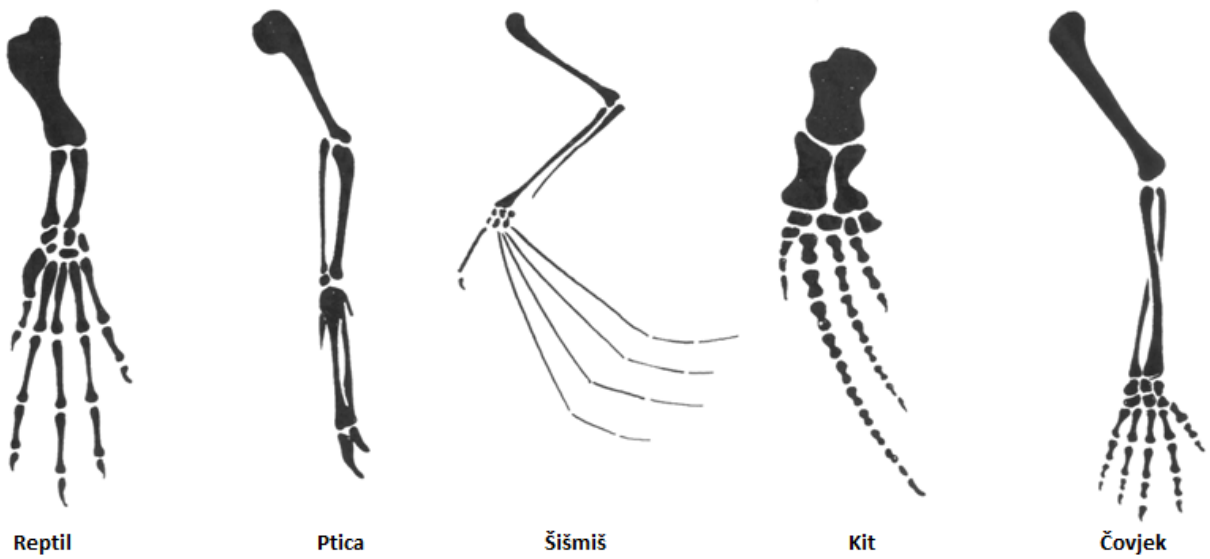
Tabela 1 Klasifikacija čovjeka i mušice

	Čovjek	Mušica
Carstvo	Animalia	Animalia
Koljeno	Chordata	Arthropoda
Razred	Mammalia	Insecta
Red	Primata	Diptera
Porodica	Hominidae	Drosophilidae
Rod	*Homo*	*Drosophila*
Vrsta	*sapiens*	*melanogaster*

3.2 Homologija

Slične karakteristike koje potječu od zajedničkog pretka nazivaju se homologne, primjerice orlovo krilo i ljudska ruka. Ostale, na prvi pogled slične, karakteristike mogle su se razviti neovisno konvergentnom evolucijom, primjerice, krilo orla i pčele (najbliži zajednički predak orla i pčele nije

posjedovao krila). Evolucijom, homologne karakteristike su divergirale u neprepoznatljivo različite oblike i funkcije.



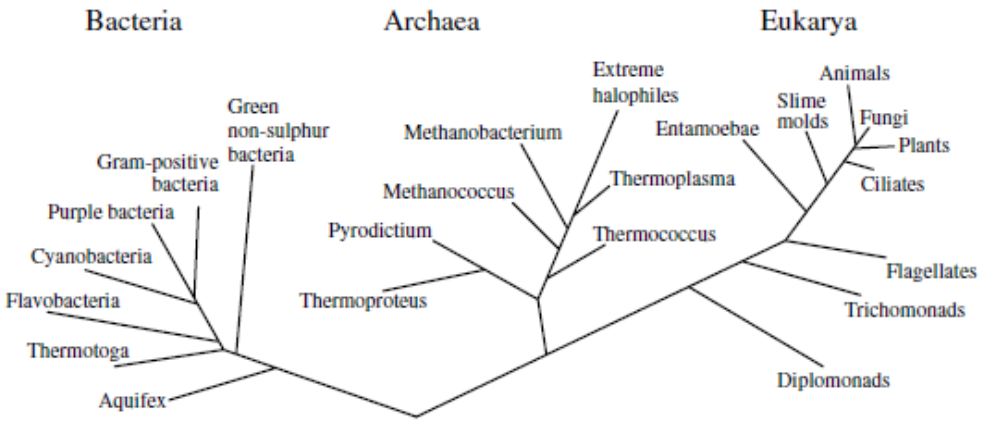
Slika 3 Primjer homologije kralježnjaka

Kada govorimo o homologiji, tada možemo govoriti i o sličnosti na razini DNA sljedova, a time i DNA sljedova za gene koji kodiraju proteine. Analiza sljedova daje jednoznačne dokaze odnosa među vrstama. Klasifikacija viših organizama je pouzdanija od klasifikacije mikroorganizama. Razlog tomu je što analiza sljedova uz tradicionalne discipline poput komparativne anatomije, paleontologije i embriologije daje konzistentnu sliku, dok se kod klasifikacije mikroorganizama javljaju problemi, primjerice odabir svojstava. Ribosomska RNA je ključna u postupku analize, iz razloga što je prisutna u svim organizmima te posjeduje dobar stupanj divergencije.

Analizom sljedova ribosomske RNA, Carl Woese (Woese, 1990.) podijelio je živa bića u tri domene (razina iznad Carstva):

1. Bakterije
2. Archaea
3. Eukarioti

gdje bakterije i Archaea pripadaju skupini prokariota, grupi jednostavnih organizama bez stanične jezgre i organela (posjeduju samo ribosome).



Slika 4 Stablo života
Izvor: Lesk, 2005.

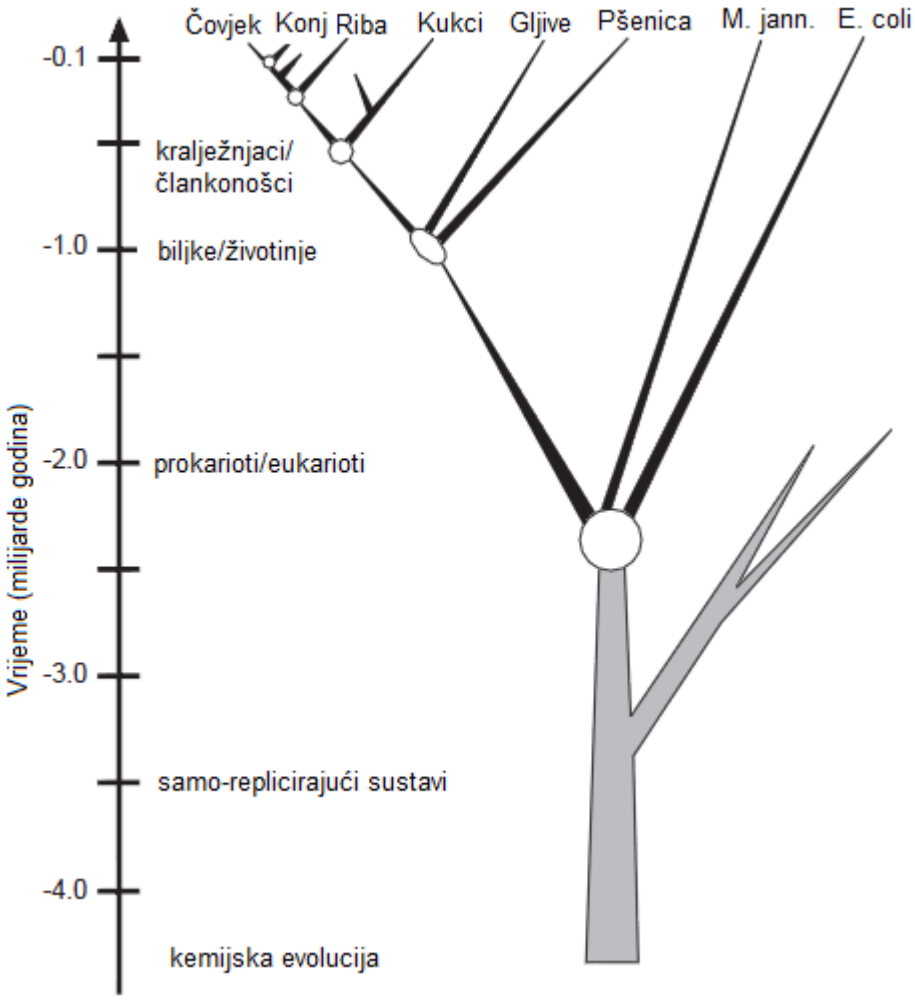
3.3 Evolucijske vremenske mjere

U potrazi za homolognim proteinima, pokušavamo otkriti proteine koji su u prošlosti dijelili zajedničkog pretka. Slika 5 prikazuje evolucijsko stablo koje seže do početaka zemljine povijesti. Cilj usporedbe sljedova je: za odabrani proteinski slijed pronaći (u bazi proteina) homologni slijed nekih drugih, divergentnih organizama. Primjerice, ako pretraga za neki ljudski protein daje značajne rezultate sličnosti s proteinom iz kvasca, tada je nasljeđen protein morao postojati u organizmu od prije naj manje milijardu godina, te su ga potomci tog organizma sačuvali u današnjem čovjeku i kvascu. Također, ako je protein kvasca homologan s nekim proteinom iz bakterije *E.coli*, tada je taj slijed (protein) morao postojati zadnjih 2 milijarde godina u praiskonskom organizmu koji je preteča bakterija i gljiva.

Pregledavanjem sljedova proteina i DNA, u većini slučajeva promatramo aktualne (današnje) sljedove. Dakle, nema smisla reći da je slijed kvasca ili bakterije primitivniji od slijeda sisavca, svi sljedovi su suvremeni. Također, može se vidjeti da postoje sljedovi koji se mogu pronaći samo u kralježnjacima ili samo u biljkama, ali ne u oboje. Takvi sljedovi su moderniji od onih koji se mogu naći i u sisavcima i u bakterijama.

Za organizme koji su divergirali u zadnjih 600 milijuna godina, zaključci o vremenu divergencije današnjih organizama dobivaju se iz geoloških podataka. Starija vremena divergencije zaključuju se ekstrapolacijom evolucijskih "satova" (Pearson, 2001).

Ideja evolucijskih satova bazira se na sporim promjenama proteinskih sljedova. Takve procjene vremena divergencije zahtjevaju da te promjene budu u prosjeku konstantne. Najstariji fosili pripadaju prokariotima, pronađeni su u kamenju starom oko 2.5 milijarde godina. To geološko vrijeme konzistentno je sa vremenom zaključenim evolucijskim satovima.



Slika 5 Vremenska lenta evolucije
Izvor: Dayhoff, 1978

Tabela 2 prikazuje prekretnice u evoluciji, te uz tabelu 3 daje bolju perspektivu o evolucijskim dosezima koje pružaju različite obitelji proteina. Teoretska vremena pogleda u prošlost prikazana u tabeli 3, baziraju se na pretpostavci da je moguće prepoznati proteine koji dijele oko 20% sljedova. Kasnije će se pokazati da su dva proteina homologna ako dijele 25% slijeda cijelom dužinom. Navedena vremena pogleda u prošlost mogu se pokazati u praksi, osjetljivim algoritmima usporedbe sljedova (Pearson, 2001).

Tabela 2 Povijesni događaji u biologiji

Podrijetlo svemira	-12 ± 2	(milijardi godina)
Nastanak sunčevog sustava	-4.6 ± 0.4	
Prvi samo-replicirajući organizam	-3.5 ± 0.5	
Divergencija prokariota i eukariota	-1.8 ± 0.3	
Divergencija biljaka i životinja	-1.0	
Divergencija kralježnjaka i beskralježnjaka	-0.5	
Početak širenja sisavaca	-0.1	

Preuzeto: Doolitle, 1986

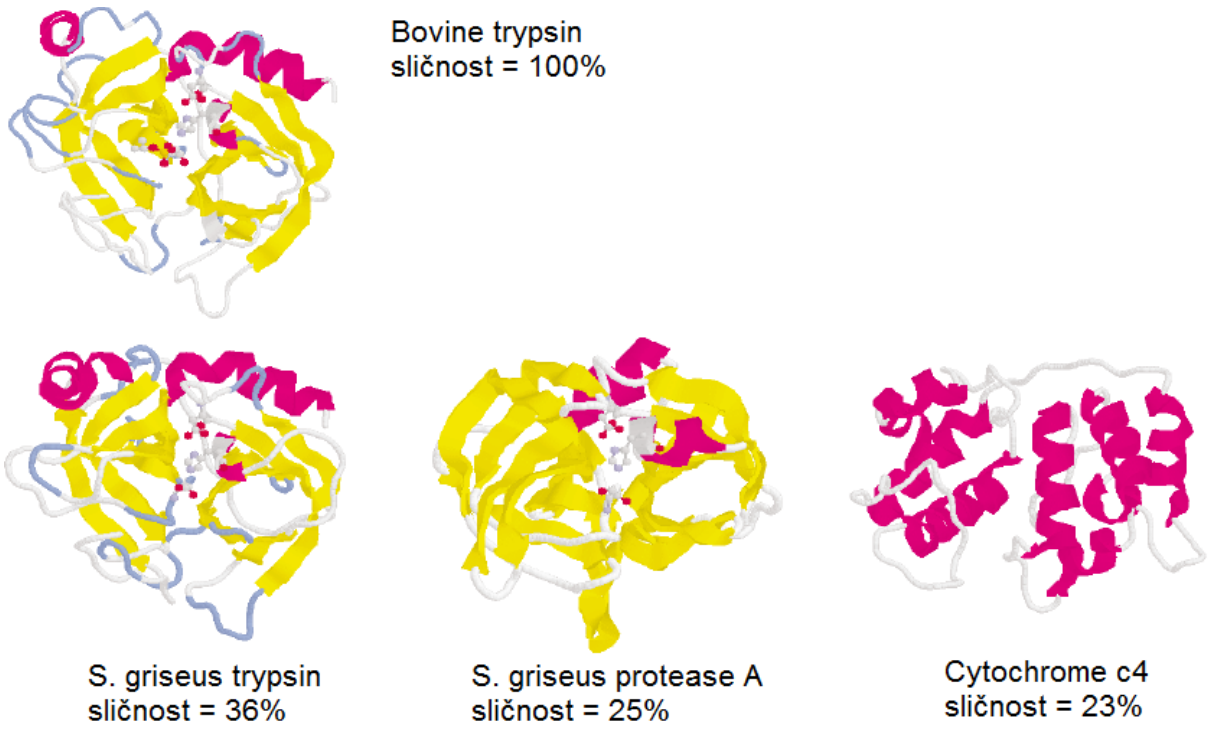
Tabela 3 Evolucijski opsezi

Protein	Teorijsko vrijeme pogleda u prošlost	Evolucijski opseg
Pseudogeni	45 (milijuna godina)	Primati, Glodavci
Fibrinopeptidi	200	Sisavci
Laktalbumini	670	Kralježnjaci
Ribonukleaze	850	Životinje
Hemoglobini	1.5 (milijarde godina)	Biljke/Životinje
Kisele proteaze	2.3	Prokarioti/Eukarioti
Triofosfat-izomeraze	6	Arhae
Glutamat-dehidrogenaze	18	

Preuzeto: Doolitle, 1986

3.4 Sličnost, nasljedstvo i struktura

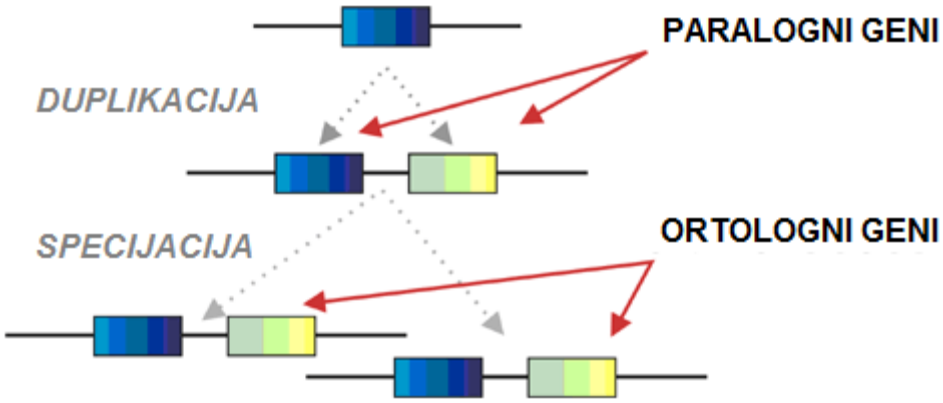
Kako homologni proteini dijele sličnu trodimenzionalnu strukturu, najbolji zaključak koji se može dobiti usporedbom sljedova proteina je upravo onaj koji potvrđuje homologiju. Primjeri tri člana superfamilije *serin proteaze* prikazani su na slici 6. Navedene su vrijednosti očekivanja i postotka sličnosti sa *goveđim tripsinom*. Dva proteina, *goveđi kimotripsin* i *S. griseus tripsin,* dijele veliku sličnost sljedova, dok treći slijed u rodu, *S. griseus proteaza A*, ne dijeli značajnu sličnost dok im je njihova struktura vrlo slična. To pokazuje da homologni proteini ne dijele nužno primjetnu sličnost sljedova. Protuprimjer je *citokroma c4*, koji postiže visoku sličnost sljedova. U ovom slučaju, protein ne dijeli strukturnu sličnost sa *tripsinom*. Ako dva proteina nisu homologni, ne mogu se vući zaključci o njihovoj strukturnoj sličnosti, makar oni imaju visoku sličnost sljedova.



Slika 6 Primjeri strukturne sličnosti i sličnosti sljedova

3.5 Načini evolucije

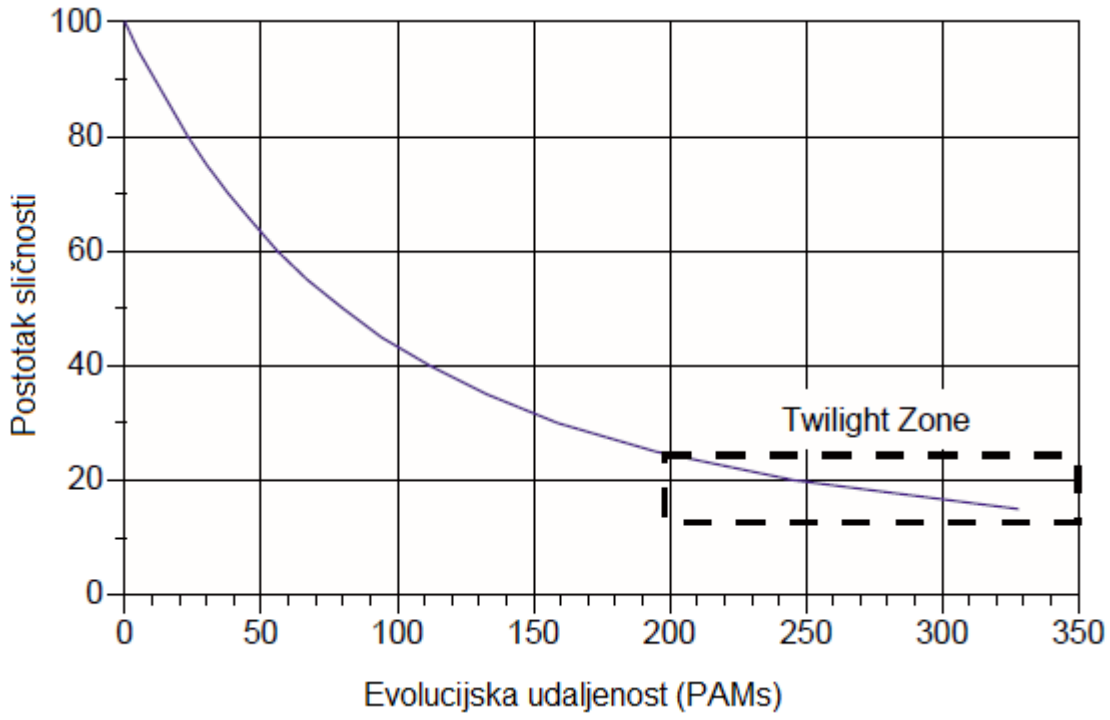
Homologni sljedovi mogu se podijeliti u dvije skupine: *ortologne i paralogne* sljedove. Ortologni sljedovi se razlikuju jer se pronalaze u različitim vrstama, dok su paralogni sljedovi nastali dupliciranjem gena unutar iste vrste. Primjerice, pripadnici obitelji globina su i ortologni, razlikuju se zbog specijacije, i paralogni, razlikuju se zbog duplikacije gena. Tako su ljudski α-globin, mišji α-globin i pileći α-globin svi ortologni, pošto se razlikuju zbog specijacije kojom su se razvili ljudi, glodavci i ptice. Sa druge strane, mišji β-globin i ljudski α-globin su paralogni radi duplikacije gena koja je stvorila α i β pod jedinice prije 600 milijuna godina. Evolucijsko stablo koje se bazira na ljudskom α, pilećem α i mišjem β-globinu dovelo bi do zaključka da je čovjek u jačem srodstvu sa piletom nego mišem. Takva se greška teško može dogoditi u slučaju dobro istražene obitelji poput globina, no može se češće pojaviti kod većih, raznolikijih i slabije opisanih obitelji proteina (Pearson, 2001).



Slika 7 Načini evolucije
Izvor: Šeda, 2006.

3.6 Divergencija obitelji proteina

Mnoge obitelji proteina imaju različite stope divergencije. Ipak, u većini slučajeva je stopa promjene u ovisnosti o evolucijskom vremenu konstantna (možemo promatrati kao brzinu promjene). Pomoću tih stopa možemo datirati događaje koji su se dogodili prije više od 600 milijuna godina, za koje ne postoje fosilni ostaci. Ipak, različite obitelji proteina divergiraju različitim stopama, te općenito, brojem razlika između dva slijeda ne možemo procijeniti vrijeme kada su se ti sljedovi razdvojili (divergirali). To ponajviše vrijedi za paralogne sljedove. Jednom kada se slijed duplicira, ubrzano se mijenja, sve dok ga ne uspori selekcijski pritisak na njegovu novu funkciju.



Slika 8 Ograničenja sličnosti sljedova
Izvor: Pearson, 2001.

3.7 Usporedba DNA i proteina

Sve metode usporedbe koje ćemo u nastavku opisati rade i na sljedovima proteina i na DNA sljedovima. Mogućnost otkrivanja daljeg srodstva drastično se smanjuje uporabom DNA sljedova. Razlog tomu je što DNA sljedovi daju puno manje informacije, zbog nedostatka inherentne biokemijske informacije koju sadrže PAM matrice vrednovanja, te zbog toga što mnoge promjene u DNA sljedovima ne mijenjaju protein kojeg kodiraju. Razlike u performansama algoritama su zanemarive u usporedbi sa gubitkom informacija pri usporedbi DNA sljedova.

4. Metode poravnanja

Postoji velik broj algoritama i parametara vrednovanja za procjenu sličnosti sljedova proteina ili DNA. Kao i u mnogim drugim disciplinama, odabir "najbolje" metode ovisi o problemu kojeg se rješava. Primjerice, algoritmi koji izračunavaju lokalnu sličnost obično su prikladni kod pretraživanja proteinskih i DNA baza podataka, dok su algoritmi za izračunavanje globalne sličnosti prikladniji kod izgradnje evolucijskih stabala, kada je homologija već uspostavljena. Metode zasnovane na uzorcima su prikladnije pri traženju funkcionalno očuvanih nehomolognih domena. Pretraživanjem proteinskih baza podataka, sa ciljem pronalaska udaljenih homolognih proteina, važnije je izbjeći visoke rezultate sličnosti međusobno nesrodnih sljedova od izračunavanja visokih rezultata za povezane sljedove. Neke baze podataka sadrže preko 50 tisuća zapisa, dok obitelji proteina uobičajeno sadrže manje od 100 članova. Dakle, algoritmi i matrice vrednovanja koje daju najbolja poravnanja i nisu najučinkovitiji u pretraživanju velikih proteinskih baza podataka (Pearson, 1995; Pearson, 1998).

4.1 Algoritmi

Za izračunavanje sličnosti s ciljem određivanja homologije, koriste se dvije skupine algoritama za usporedbu. Strogi algoritmi koji garantirano izračunavaju optimalnu vrijednost sličnosti, npr. Needleman-Wunsch (Needleman i Wunsch, 1970) i Smith-Waterman (Smith i Waterman, 1981) algoritmi, te brzi heuristički algoritmi koji ne garantiraju izračun optimalne vrijednosti sličnosti za sve sljedove u bazi, npr. FASTA (Pearson i Lipman, 1988) i BLAST (Altschul, 1990).

U tabeli su pobrojani najčešće korišteni algoritmi.

Tabela 4 Algoritmi usporedbe sljedova

Algoritam	Rezultat izračunavanja	Matrica vrednovanja	Vremenska složenost (asimptotska)	Prostorna složenost	Autori
Needleman-Wunsch	Globalna sličnost	Proizvoljna	$O(n^2)$	$O(n^2)$	Needleman i Wunsch, 1970.
Smith-Waterman	Lokalna sličnost	$S_{ij} < 0.0$	$O(n^2)$	$O(n^2)$	Smith i Waterman, 1981..
FASTA	Približna lokalna sličnost	$S_{ij} < 0.0$	$O(n^2)/K$	$O(n^2)/K$	Lipman i Pearson, 1985.
BLAST	Najveća vrijednost segmenta	$S_{ij} < 0.0$	$O(n^2)/K$	$O(n^2)/K$	Altshul, 1990.

Dva optimalna algoritma za izračun sličnosti su Needleman-Wunsch algoritam, koji izračunava "globalni" rezultat sličnosti dva slijeda, te Smith-Waterman algoritam koji izračunava "lokalni" rezultat sličnosti. Izračun globalnog rezultata zahtjeva da poravnanje započne na početku svakog slijeda i da se proteže do kraja slijeda. Rezultati globalnog poravnanja mogu se računati sa i bez penala za praznine (eng. *gap penalty*) na krajevima sljedova. Algoritmi globalnog poravnanja ne mogu se koristiti za otkrivanje DNA vezivnih domena i sličnih karakterističnih struktura u sljedovima. Algoritmi lokalnog poravnanja otkrivaju najsličnije regije koje dijele dva slijeda. Algoritmi lokalnog poravnanja mogu se koristiti za otkrivanje DNA vezivnih domena i sličnih karakterističnih struktura u sljedovima.

Strogi algoritmi usporedbe sljedova, poput Smith-Waterman algoritma, zahtijevaju vrijeme proporcionalno složenošću *O(mN)*, gdje m predstavlja duljinu slijeda, a N broj aminokiselina u bazi proteinskih sljedova. Iako postoje vrlo brzi algoritmi za izračun optimalnog rezultata globalnog poravnanja dva slijeda (bilo kojeg tipa, ne specifično proteina), takvi algoritmi nisu primjereni za usporedbu bioloških sljedova. Razlog tomu je što zanemaruju važnu biokemijsku informaciju sadržanu u PAM250 matrici (vidi potpoglavlje "Matrice vrednovanja").

4.2 Algoritmi dinamičkog programiranja

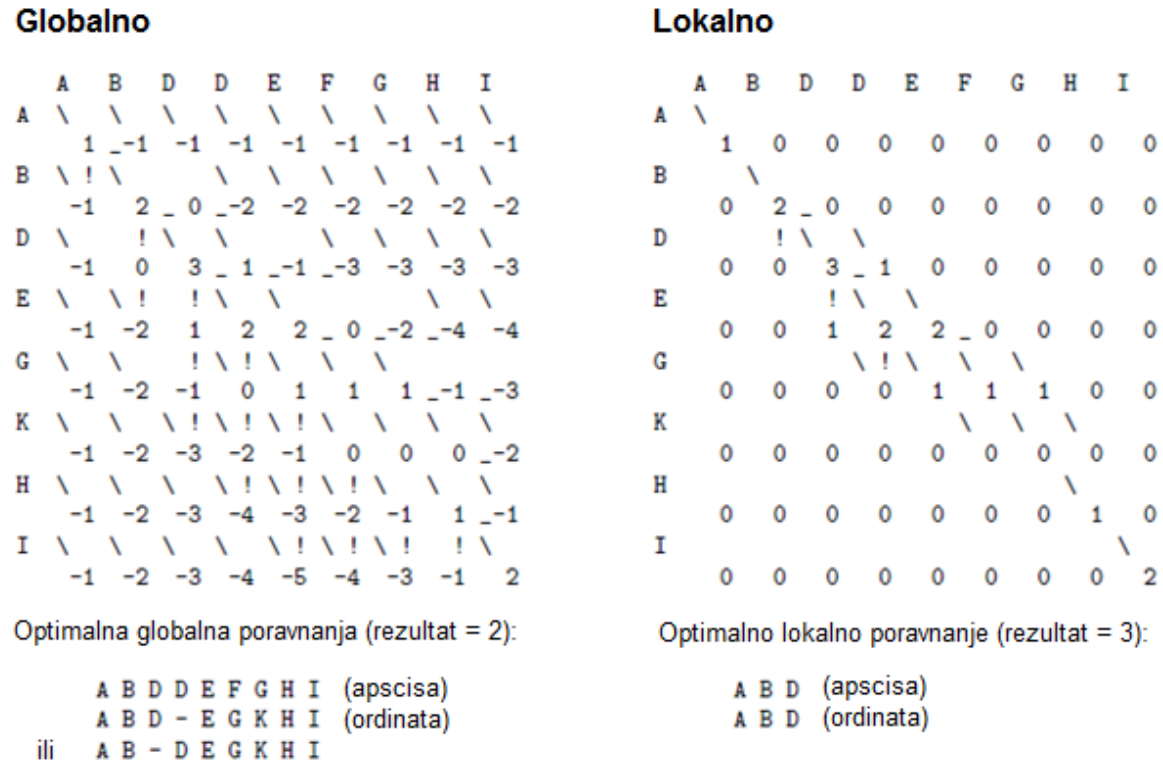
Postoji eksponencijalan broj mogućih poravnanja sa prazninama između dva slijeda proteina ili DNA. Uz algoritme dinamičkog programiranja moguće je izračunati optimalne rezultate u *O(MN)* koraka. Takvu efikasnost postiže se (1) određivanjem najboljeg rezultata za svaki prefiks svakog niza i potom (2) novim proširenjem svakog prefiksa odabirom jednog od tri puta kojima možemo proširiti poravnanje:

1) proširenjem poravnanja za 1 ostatak u svakom slijedu (dijagonalno)

2) proširenjem poravnanja za prvi ostatak u prvom slijedu i poravnavanjem sa prazninom u drugom (desno)

3) proširenjem poravnanja za drugi ostatak u prvom slijedu i poravnavanjem sa prazninom u prvom (dolje)

ova odluka mora se učiniti za svaki od *MN* prefiksa sljedova duljina *M* i *N.*

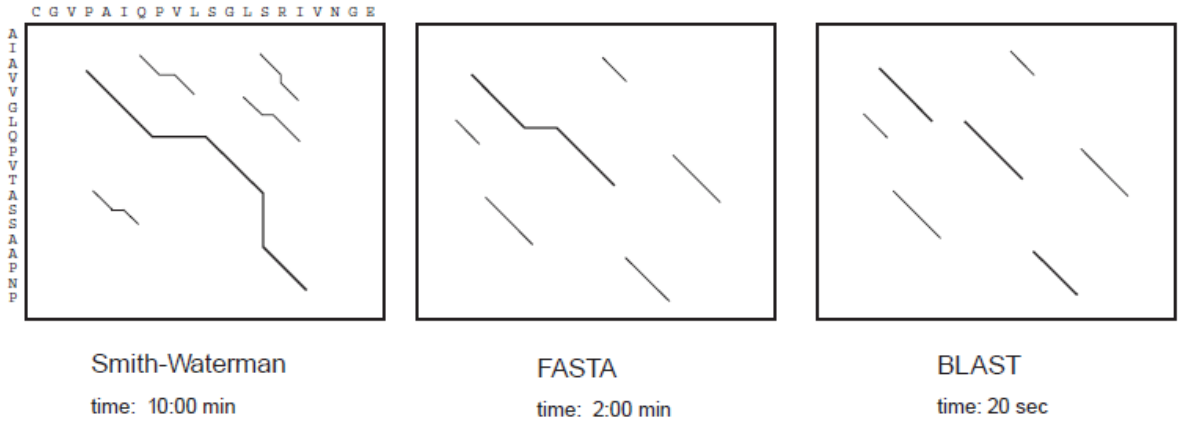


Slika 9 Primjeri dinamičkog programiranja: gloabalni i lokalni putevi poravnanja
Izvor: Pearson, 2001.

4.3 Heuristički algoritmi

Dva često korištena heuristička algoritma za pretragu baza proteina i DNA sljedova su FASTA (Pearson i Lipman, 1988) i BLAST(Altschul, 1990). Ove su metode 5 do 50 puta brže od strogih algoritama, npr. Smith-Waterman, dok su rezultati koje daju u većini slučajeva slične kvalitete.

Na slici su prikazane razlike između FASTA, BLAST i Smith-Waterman algoritama. BLAST i FASTA su brži od Smith-Watermana jer provjeravaju samo dio poklapanja dva niza proteina. FASTA se oslanja na regije koje su pojedinačno ili u parovima identične, dok BLASTP pretražuje samo triplete aminokiselina.



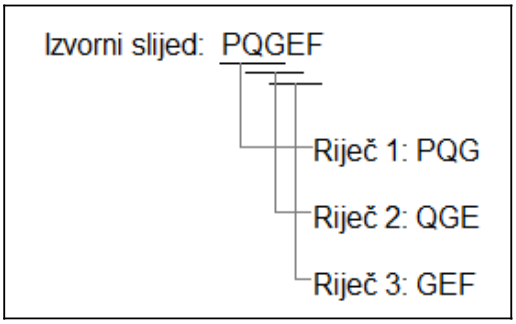
Slika 10 Usporedba heurističkih algoritama
Izvor: Pearson, 2001.

BLAST

Napredci u statističkoj teoriji poravnavanja sljedova bez pukotina(Karlin i Altschul, 1990) pružila je teorijsku podlogu za razvoj BLAST algoritma. Danas je jedan od najšire korištenih programa za brzu usporedbu sljedova, ponajviše zbog točnosti procjene statističke značajnosti rezultata sličnosti. BLAST (kao i FASTA), koristi pretraživanje temeljeno na usporedbi riječi, kako bi otkrio regije lokalne sličnosti (bez pukotina). Njegova učinkovitost posljedica je kombinacije visoke osjetljivosti i odlične selektivnosti.
Pregled BLASTP algoritma (BLAST algoritam koji uspoređuje proteine):

1. Uklanjanje regija niske složenosti ili ponavljajućih sljedova
   -odnosi se na regije slijeda sastavljenih od malo različitih elemenata, uzrokuju visoke rezultate sličnosti koji navode na pogrešne zaključke

2. Stvaranje liste riječi od k slova
   -za k se uzima najčešće 3 (slika 10)

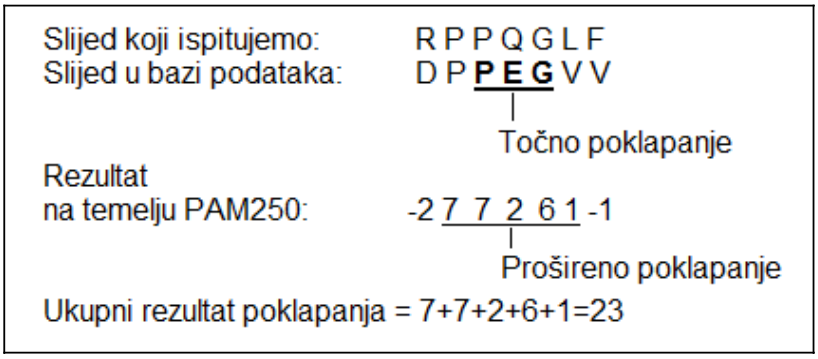


Slika 11 Primjer stvaranja liste riječi

3. Izračunavanje svih mogućih poklapanja riječi
   -za sve riječi iz liste (korak 2.) se prođe po svim riječima od 3 slova, te na temelju matrice vrednovanja (poglavlje 4.4) se izračuna rezultat poklapanja
   -na daljnje razmatranje se uzimaju sve riječi (obje)čiji rezultat prelazi određeni prag
   -npr. riječ PQG se usporedi sa PEG i PQA, te se na temelju PAM250 matrice vrednovanja dobivaju rezultati 13 i 11

4. Pretraživanje baze sljedova proteina za točnim poklapanjem
   - za sve riječi iz liste riječi sa visokim rezultatom, pretražuju se svi proteini iz baze podataka
   - sva točna preklapanja predstavljaju moguća poravnanja bez pukotina

5. Proširivanje točnih poklapanja parovima segmenata visokih rezultata (eng. *High-scoring segment pair,* ***HSP***)
   - poravnanje riječi između slijeda koji ispitujemo i slijeda iz baze se proširuje na lijevu i desnu stranu sve dok novi parovi imaju pozitivnu vrijednost u matrici vrednovanja (slika 11)



Slika 12 Primjer proširenja parovima segmenata visokih rezultata

6. Određivanje liste HSP-ova koji imaju dovoljno visok rezultat
   -odabiru se oni čiji je rezultat viši od nekog empirijski određenog praga

7. Procjena statističke značajnosti HSP rezultata
   - ravnanjem prema Grumbelovoj distribuciji (eng. *Grumbel extreme value distribution*, prema rezultatima Smith-Waterman algoritma) vrši se procjena

8. Povezivanje dvije ili više HSP regije
   -ponekad se mogu dvije ili više HSP regije povezati u dulje poravnanje

9. Prikaz lokalnog poravnanja s pukotinama za pronađene proteine pomoću Smith-Waterman algoritma

FASTA

FASTA koristi algoritam sličan BLAST-u, odnosno, pretraživanje temeljeno na usporedbi riječi, te dodatne optimizacije rezultata. Također, koristi i Smith-Waterman algoritam kako bi proizveo završna poravnavanja. Rezultati koje daje FASTA su bolji od rezultata BLAST algoritma.

Svaka pretraga baze za članovima raznolike obitelji proteina uključuje *"tradeoff"* između osjetljivosti (mogućnosti otkrivanja udaljenih srodnih sljedova) i selektivnosti (mogućnost izbjegavanja dobrih rezultata za nesrodne sljedove).

4.4 Matrice vrednovanja

Matrice vrednovanja (eng. *scoring matrix*) koje se koriste za usporedbu proteina su sofisticiranije od "+1 ako se poklapa i -1 ako ne". Najučinkovitije matrice se baziraju na stvarnoj učestalosti zamjena koje se javljaju kod srodnih proteina. Matrice vrednovanja se razlikuju na tri načina:

- načinu konstrukcije
- sadržanim informacijama, što je povezano sa brojem ostataka aminokiselina koji se moraju podudarati da bi se došlo do statistički značajnog rezultata
- veličini, tj. količini informacije po jedinici vrednovanja

Osnovna PAM250 matrica je nastala promatranjem nekoliko stotina poravnanja vrlo srodnih proteina, te izračuna učestalosti promjene svakog ostatka aminokiseline u drugi, u vrlo kratkom evolucijskom vremenu (gdje se najviše promijenilo 1% ostataka) (Dayhoff, 1978). Ta učestalost promjena se potom iskoristi za izračun PAM1 matrice (PAM je eng. *Point Accepted Mutation*). Ako PAM1 matricu pomnožimo 250 puta samu sa sobom, dobivamo PAM250 maticu koja reflektira učestalost promjena proteina koji su divergirali 250%. Očekivanje sličnosti dva proteina koji su divergirali 250% iznosi 20% istovjetnog niza, što je na rubu otkrivanja sličnosti proteina (vidi poglavlje 3. Evolucija). Uz PAM matrice, često su korištene BLOSUM matrice (koristi ih BLASTP program).

Na slici 10 je prikazana PAM250 matrica sličnosti. Matrica je simetrična, elementi na dijagonalama predstavljaju vrijednosti pridijeljene identičnim amino-kiselinama, dok su ostali elementi vrijednosti pridijeljene supstituiranim aminokiselinama.

Cys	12																			
Ser	0	2																		
Thr	-2	1	3																	
Pro	-1	1	0	6																
Ala	-2	1	1	1	2															
Gly	-3	1	0	-1	1	5														
Asn	-4	1	0	-1	0	0	2													
Asp	-5	0	0	-1	0	1	2	4												
Glu	-5	0	0	-1	0	0	1	3	4											
Gln	-5	-1	-1	0	0	-1	1	2	2	4										
His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

Slika 13 PAM250 Matrica vrednovanja

5. Zaključak

Usporedba sljedova proteina je jedan od najjačih alata za predviđanje strukture i funkcije proteina iz njegovog slijeda. Razlog tomu leži u ograničenjima evolucije proteina, tj. potrebe za savijanjem u funkcionalnu strukturu. Pomoću sličnosti sljedova proteina, možemo otkriti rodbinske veze proteina (homologe) čiji je zadnji zajednički predak postojao prije 1-2.5 milijarde godina. Većina sljedova koji su statistički vrlo slični su homologni, no mnogi dalji homolozi ne dijele veliku količinu slijeda te ih se teže otkriva. Homologni proteini dijele zajedničkog pretka, time i sličnu trodimenzionalnu strukturu.

Mogućnosti otkrivanja daljih homologa je unaprijeđen razvojem točnih statističkih predviđanja, i tehnika normalizacije koje su iskoristili mnogi alati za poravnavanje sljedova proteina. Spomenute su najčešće korištene tehnike za usporedbu i poravnanje sljedova, algoritmi, dinamičko programiranje, heuristike te objasnili na koji način koristimo statističke informacije o proteinima - matrice vrednovanja. Tehnike usporedbe sljedova posjeduju dva važna svojstva: osjetljivost - mogućnost otkrivanja udaljenih homologa, te selektivnost - prepoznavanje nehomolognih proteina s velikim postotkom sličnosti.

6. Literatura

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. Basic local alignment search tool, *J. Mol. Biol.* 1990.

2. Bujnicki , J.M. Prediction of protein structures and functions, *Wiley*, 2009.

3. Dayhoff, M., Schwartz, R. M., Orcutt, B. C. A model of evolutionary change in proteins, *Atlas of Protein Sequence and Structure*, 1978.

4. Doolittle , R. F., Feng , D. F., Johnson , M. S., McClure, M. A. Relationships of human protein sequences to those of other organisms, *Cold Spring Harb. Symp. Quant. Biol.,* 1986.

5. Durbin, R., Eddy, S. E., Krogh, A., Mitchison, G. Biological sequence analisys, *Cambridge*, 1998.

6. Karlin, S. & Altschul, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proc. Natl. Acad. Sci. USA,* 1990.

7. Linderstrøm-Lang, K.U. Proteins and Enzymes, *Lane Medical Lectures*, Stanford University Publications, University Series, Medical Sciences, vol. 6, Stanford University Press, 1952.

8. Linnaeus, C. Systema Naturae, Švedska, 1735.

9. Lesk, A.M. Introduction to bioinformatics, *Oxford press*, 2005.

10. Martz, E. Practical Protein 3D Structure & Structural Bioinformatics Workshops, 1.4.2009., *http://www.umass.edu/molvis/workshop/prot1234.htm*, 28.3.2011.

11. Needleman, S., Wunsch, C. A general method applicable to the search for similarities in the amino acid sequences of two proteins, *J. Mol. Biol.*, 1970.

12. Pearson, W. R., Lipman, D. J. Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci. USA*, 1988.

13. Pearson, W. R. Comparison of methods for searching protein sequence databases, *Prot. Sci.,* 1995.

14. Pearson,W. R. Empirical statistical estimates for sequence similarity searches, *J. Mol. Biol.,* 1998.

15. Pearson, W.R. Protein Sequence comparison and Protein evolution, *Univ. of Virginia*, 2001.

16. Smith, T. F. , Waterman, M. S. Identification of common molecular subsequences*, J. Mol. Biol.*, 1981.

17. Šeda, O. , Liška, F., Šedová, L., 20.11.2006., *http://biol.lf1.cuni.cz/ucebnice/en/glossary.htm,* 28.3.2011.

18. Wissman, P. Amino Acids and Proteins, 1.2.2007., *http://homepage.smc.edu/wissmann_paul/anatomy2textbook/AAcidsProteins.html* , 28.3.2011.

19. Woese C., Kandler O., Wheelis M. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, *Proc. Natl. Acad. Sci.*, 1990.