SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO - MATEMATIČKI FAKULTET

BIOLOŠKI ODSJEK

USPOREDBA 1D i 3D PORAVNANJA PROTEINA

COMPARISON OF 1D AND 3D PROTEIN ALIGNMENT

SEMINARSKI RAD

Maja Zagorščak

Preddiplomski studij molekularne biologije

(Undergraduate Study of Molecular Biology)

Mentor: Doc. Dr. Sc. Pavle Goldstein

Zagreb, 2009.

SADRŽAJ

1. UVOD

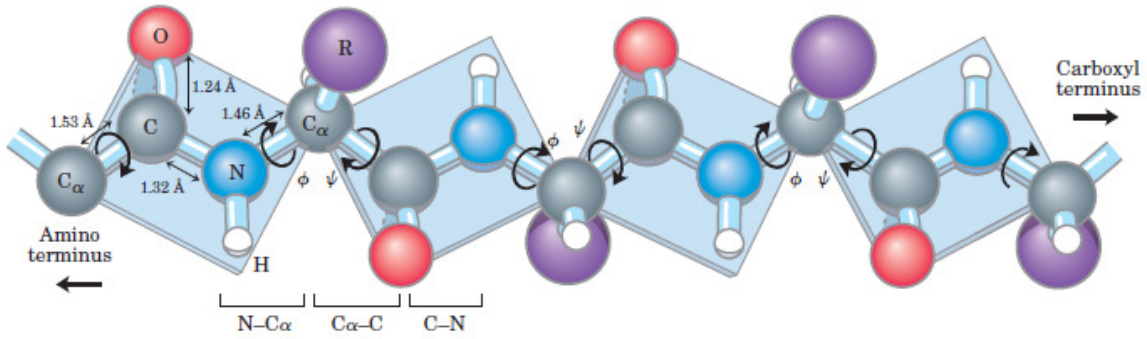
PROTEINSKE STRUKTURE

Svi proteini, izolirani iz najstarijih bakterijskih linija ili iz najsloženijih oblika živućeg svijeta, konstruirani su od istog sveprisutnog seta 20 aminokiselina, kovalentno povezanih u karakteristične linearne slijedove. Svaka od tih aminokiselina sadrži bočni lanac sa određenim biokemijskim svojstvima, te se grupa od 20 prekursornih molekula može gledati kao specifična vrsta pisma u kojem je zapisana struktura proteina. Stanice su sposobne stvarati proteine (polipeptide) različitih osobina i aktivnosti spajajući istih 20 aminokiselina u različite kombinacije i slijedove, te iz tih izgrađujućih blokova tvoriti raznolike produkte poput enzima, hormona, protutijela, transportne bjelančevine, mišićnih vlakana, očne leće, perja, paukovih mreža, roga nosoroga, mliječnih proteina, antibiotika, otrova gljiva i mnogih drugih supstanci koje imaju različite biološke funkcije. Sve od 20 aminokiselina su α-aminokiseline. Imaju karboksilnu grupu i amino grupu vezanu na isti atom ugljika. Međusobno se razlikuju u lancima bočnih ogranaka ili R grupama, koje variraju u strukturi, veličini i električnom naboju. Svaku aminokiselinu kraće možemo zapisati pomoću kombinacije tri slova ili pridodanog simbola (slika 1.2). Postoje dva načina obilježavanja ostalih ugljikovih atoma: dodatnim C-atomima u R-grupama dodajemo β, γ, δ,... slova grčkog alfabeta, počevši od unuarnjeg Cα-atoma, dok se u ostalim organskim molekulama C-atomi jednostavno označavaju brojevima od jednog kraja prema drugom, s time da karboksilni-atom ima najveći prioritet (C-1) a Cα-atom je označen kao C-2. U nekim slučajevima, poput aminokiselina sa heterocikličkim R grupama, sustav grčkog afabeta je prekompliciran te se upotrebljava standardno brojčano označavanje. U svih 20 aminokiselina (osim glicina) na Cα-atom su vezane četiri različite grupe: karboksilna grupa (-COOH) , amino grupa ($-NH_2$), R grupa i atom vodika (u glicinu R grupa je dodatni H-atom). Time Cα-atom (slika 1.1) postaje kiralni centar, a aminokisleine imaju dva moguća stereoizomera D i L, zrcalno simetrična, što nazivamo enantiomeri. Sve molekule s kiralnim centrom su optički aktivne.

Prostorna organizacija atoma u proteinu naziva se konformacija. Moguće konformacije proteina uključuju sva strukturalna stanja koja mogu biti postignuta bez kidanja kovalentne veze. Do promjene u konformacije može doći i uslijed rotacije oko samo jedne veze. Konformacije postojane pod nizom uvjeta obično su one koje su termodinamički najstabilnije, imaju najnižu slobodnu Gibbsovu energiju. Proteini u funkcionalnoj, smotanoj konformaciji zovu se nativni proteini. Polipeptidi variaju u veličini od dvije, tri aminokiseline do nekoliko tisuća u svim vezanim aminokiselinskim ograncima. Kovalenta okosnica tipičnog proteina sadržava stotine pojedinačnih veza. Zbog slobodne rotacije oko svih tih veza protein je u mogućnosti zauzeti neograničen broj konformacija, no ipak svaki protein posjeduje specifične kemijske ili strukturalne osobine, iz čega proizlazi kako svaki ima i jedinstvenu trodimenzionalnu strukturu. S obzirom na činjenicu kako se usmjereno polje molekula u kristalu može formirati jedino ako su molekularne podjedinice identične, jednostavna činjenica što mnogi proteini mogu kristalizirati, pruža snažan dokaz da su čak i vrlo veliki proteini diskretni entiteti sa jedinstvenom

strukturom. Nekoliko je važnih činjenica vezanih uz trodimenzionalnu strukturu proteina: determinirana je aminokiselinskim slijedom, funkcija proteina ovisi o njegovoj strukturi, izolirani protein obično postoji u jednoj ili malom broju stabilnih konformacija, za održavanje stabilnosti određene strukture najvažnije su nekovalentne interakcije, između velikog broja jedinstvenih proteinskih struktura možemo prepoznati određene ponavljajuće uzorke koji nam pomažu razumjeti proteinsku arhitekturu.

Za pojedini protein, aminokiselinski ostatak (slijed) ključan za njegovu aktivnost, ostaje konzerviran u procesu evolucije. Aminokiselinski ostaci koji nisu ključni za njegovu funkcionalnost mogu varirati tijekom vremena – jedna aminokiselina može biti zamijenjena drugom (frameshift mutacije, tranzicije i transverzije...), dolazi do skraćivanja lanca (nonsense mutacije) ili produljenja (insercije). Aminokiselinske supstitucije nisu uvijek nasumične, uglavnom dolazi do zamjene aminokiselinom sa istim svojstvima (pr. leucin i izoleucin su slične hidrofobne aminokiseline koje obično mogu zamijeniti jedna drugu bez da se funkcija ili struktura proteina zbog toga bitnije promijeni). Sličnosti u strukturi i funkciji proteina ukazuju na evoluciju iz zajedničkog pretka. Proteini sa zajedničkim pretkom grupirani su u proteinske familije i nazvani homolozima. Ako su dva proteina unutar familije prisutna unutar iste vrste, nazivamo ih paralozima, dok homologe koji potječu iz različitih vrsta nazivamo ortolozima.



(Slika 1.1)

Amino acid	Abbreviation/ symbol		M_r	pK_1 (—COOH)	pK_2 ($—NH_3^+$)	pK_R (R group)	pI	Hydropathy index*	Occurrence in proteins (%)†
Nonpolar, aliphatic R groups									
Glycine	Gly	G	75	2.34	9.60		5.97	−0.4	7.2
Alanine	Ala	A	89	2.34	9.69		6.01	1.8	7.8
Proline	Pro	P	115	1.99	10.96		6.48	1.6	5.2
Valine	Val	V	117	2.32	9.62		5.97	4.2	6.6
Leucine	Leu	L	131	2.36	9.60		5.98	3.8	9.1
Isoleucine	Ile	I	131	2.36	9.68		6.02	4.5	5.3
Methionine	Met	M	149	2.28	9.21		5.74	1.9	2.3
Aromatic R groups									
Phenylalanine	Phe	F	165	1.83	9.13		5.48	2.8	3.9
Tyrosine	Tyr	Y	181	2.20	9.11	10.07	5.66	−1.3	3.2
Tryptophan	Trp	W	204	2.38	9.39		5.89	−0.9	1.4
Polar, uncharged R groups									
Serine	Ser	S	105	2.21	9.15		5.68	−0.8	6.8
Threonine	Thr	T	119	2.11	9.62		5.87	−0.7	5.9
Cysteine	Cys	C	121	1.96	10.28	8.18	5.07	2.5	1.9
Asparagine	Asn	N	132	2.02	8.80		5.41	−3.5	4.3
Glutamine	Gln	Q	146	2.17	9.13		5.65	−3.5	4.2
Positively charged R groups									
Lysine	Lys	K	146	2.18	8.95	10.53	9.74	−3.9	5.9
Histidine	His	H	155	1.82	9.17	6.00	7.59	−3.2	2.3
Arginine	Arg	R	174	2.17	9.04	12.48	10.76	−4.5	5.1
Negatively charged R groups									
Aspartate	Asp	D	133	1.88	9.60	3.65	2.77	−3.5	5.3
Glutamate	Glu	E	147	2.19	9.67	4.25	3.22	−3.5	6.3

(Slika 1.2)

2. JEDNODIMENZIONALNO PORAVNANJE

Osnova sekvencijske analize je istraživanje jesu li dva aminokiselinska slijeda međusobno srodna. Svaku poziciju u kojoj su dva slijeda identična tretiramo pozitivnim bodovima (score), a konačan iznos govori nam o kvaliteti poravnanja. Dijelove sekvence koje ne možemo međusobno poravnati nadopunjujemo i poravnavamo prazninama (gapovima). Ukoliko bi bio uveden dovoljan broj praznina, gotove sve aminokiselinske sekvence, bez obzira na duljinu i aminokiselinski slijed, mogle bi biti dovedene u neku vrstu poravnanja. Kako bi izbjegli neinformativna poravnanja, uvodimo ‘’kaznu’’ (penalty) za svaku uvedenu prazninu, snižavajuću finalni score. Viši scorovi dodijeljeni su i neidentično poravnatim sekvencama koje imaju veću frekvenciju pojavljivanja, od onih koje se pojavljuju rjeđe. Steven Henikoff i Jorja Henikoff konstruirali su "blocks supstitution" matricu (BLOSUM, tablica 2.1) aminokiselina i pripadajućih vrijednosti baziranih na učestalosti supstitucije pojedine aminokiseline, i njihovim jedinstvenim kemijskim svojstvima (supstitucija različitijom aminokiselinom poput Cys->Trp supstitucije nosi više bodova od one konzervativne). Gap je označen " - " simbolom.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**5**	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	**7**	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	**7**	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	-2	**8**	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4
C	-1	-4	-2	-4	**13**	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	**7**	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	**6**	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	**8**	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
H	-2	0	1	-1	-3	1	0	-2	**10**	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	**5**	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	**5**	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	**6**	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	**7**	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	**8**	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	**10**	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	**5**	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	**5**	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	**15**	2	-3
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	**8**	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	**5**

(Tablica 2.1)

2.1 **Definicija poravnanja dva niza**:

Neka je A alfabet, te neka je $W(A)$ skup svih konačnih riječi nad alfabetom A. Neka je A' skup definiran kao $A \cup \{-\}$ te neka je $W(A')$ skup svih konačnih riječi nad alfabetom A'. Neka su $a = (a_1, a_2, a_3, ..., a_m), b = (b_1, b_2, b_3, ..., b_n) \in W(A)$ dvije konačne riječi nad alfabetom A'. Poravnanje od a i b označavamo s $A(a,b)$ te definiramo kao bilo koji par preslikavanja $a \rightarrow a'$ i $b \rightarrow b'$ gdje su $a' = (a'_1, a'_2, a'_3, ..., a'_m)$, $b' = (b'_1, b'_2, b'_3, ..., b'_n) \in W(A')$ te vrijedi:

$$a'|_A = a, \quad b'|_A = b$$
$$|a'| = |b'| = k$$
$$\forall\, i = 1, ..., k \quad a'_i \neq - \text{ ili } b'_i \neq -$$

Ako su a i b nizovi aminokiselina, $A(a, b)$ neko njihovo poravnanje, $|a'| = |b'| = k$. Za simbole a_i i b_j kažemo da su povezani ili spareni u poravnanju $A(a, b)$ ako postoji $I \in 1....., k$ takav da vrijedi:

$$a'_i = a_i \text{ i } b'_j = b_j$$

2.1 Needelman – Wunsch algoritam

Needleman-Wunsch algoritam globalno poravnava dva niza, a temelji se na dinamičkom programiranju.

Optimalno poravnanje se izgrađuje upotrebom rješenja za optimalno poravnanje manjih podnizova.

$x_1 x_2 \cdots x_m$ (duljina niza je m)

$y_1 y_2 \cdots y_n$ (duljina niza je n)

Konstruira se matricu F takva da $F(i, j)$ bude score najboljeg poravnanja između početnog dijela niza $x, x_{1...i,}$ i početnog dijela niza $y, y_{1...j}$.

Matricu F gradi se rekurzivno. Inicijalne vrijednosti su:

$F(0, 0) = 0$

$F(i, 0) = -i \cdot d$ (svi iz x do x_i su poravnati s −)

$F(0, j) = -j \cdot d$ (svi iz y do y_j su poravnati s −)

Zatim se gradi rekurzija za računanje elemenata matrice *F*. Matricu se puni počevši s gornjim lijevim kutem, a završavamo s donjim desnim kutem. Ako su poznati $F(i-1, j-1)$, $F(i-1, j)$ i $F(i, j-1)$, može se izračunati $F(i, j)$. Na tri načina dobiva se najbolji score do poravnanja x_i, y_j :

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j), & x_i \text{ poravnat s } y_j \\ F(i-1, j) - d, & x_i \text{ poravnat s gapom} \\ F(i, j-1) - d & y_j \text{ poravnat s gapom} \end{cases}$$

Vrijednost $s(x_i, y_j)$ pohranjena je u BLOSSUM matrici, a d je "cijena" stavljanja gap-a i optimalno se uzima 8. Vrijednost u $F(n,m)$ je po definiciji score najboljeg poravnanja.

Potrebno je "pamtiti" na koji način se dolazi do izračuna optimalnog score-a tj. koja od gornjih triju formula daje maksimalni score. To se pamti tako da se uvede još i matrica P za traceback. Pomoću nje rekonstruira se optimalno poravnanje iz matrice F. Matrica P je istih dimenzija kao i matrica dinamičkog programiranja. Mjesto (i, j) popunjava se sa :

0 ako je optimalni score u $F(i, j)$ dobiven tako da je x_i poravnat s −

1 ako je optimalni score u $F(i, j)$ dobiven tako da je x_i poravnat s y_j

2 ako je optimalni score u $F(i, j)$ dobiven tako da je y_j poravnat s −

U svakom koraku traceback-a najprije se pogleda $P(i, j)$, te ovisno o tamo pohranjenom broju, u nizove poravnanja od x i y stavlja se pripadno slovo iz x odnosno y ili − . Postupak traceback-a pronalazi samo jedno poravnanje s optimalnim scorom.

Miyazawin koncept - Vjerojatnost povezivanja simbola:

Za nizove *a* i *b* definiramo particijsku funkciju Z formulom koja zbraja statističke težine za sva poravnanja nad zadanim nizovima

$$Z = \sum_1 \exp(F(Al))$$

Vjerojatnost $P(A_l)$ poravnanja Al definiramo formulom:

$$P(A_l) = 1/Z \exp(F(A_l))$$

Za ovako definiranu vjerojatnost, optimalno poravnanje je ujedno i najvjerojatnije.

Vjerojatnost povezivanja ili vjerojatnost sparivanja za par simbola a_i i bj iz nizova a i b duljina m i n definiramo formulom:

$$p(a_i, b_j) = 1/Z \; Z_{i-1\,j-1} \; \exp(s(a_i, b_j)) \; Z'_{i+1,j+1}$$

gdje je $Z_{i-1\,j-1}$ particijska funkcija za podnizove $a_{1\ldots i-1}$ i $b_{1\ldots j-1}$, a $Z'_{i+1,j+1}$ particijska funkcija za podnizove $a_{i+1\ldots m}$ i $b_{i+1\ldots n}$.

Simbol je u nekom poravnanju povezan s gapom ako nije povezan ni s jednim simbolom iz drugog niza. Vjerojatnost povezanosti simbola s gapom računa se na sljedeći način:

$$p(a_i, -) = 1 - \sum_{j=1}^{n} p(ai, bj)$$

$$p(-, b_j) = 1 - \sum_{i=1}^{m} p(ai, bj)$$

2.3 Miyazawin algoritam

Imamo dva niza aminokiselina a i b duljina m i n

1. Postavimo $i_1 = 1$, $i_2 = m$, $j_1 = 1$ i $j_2 = n$

2. Pronađemo par simbola takav da vrijedi:

i. $p(a_i, b_j) = \max_{i1 \le k \le i2 \le j1 \le l \le j2} p(a_k, b_l)$

ii. $p(a_i, b_j) \ge p(a_i, -)$

iii. $p(a_i, b_j) \ge p(-, b_j)$

3. Ako takav par simbola ne postoji, povezujemo sve a_i za $i_1 \le i \le i_2$ i sve b_j za $j_1 \le j \le j_2$ s gapom

4. Ako je (a_i, b_j) takav par, povezujemo ga u poravnanju, onda ponavljamo korake 2 do 4 za nastale podnizove

U ovako konstruiranom poravnanju mogu biti sparene aminokiseline a_i i b_j za koje vrijedi $p(a_i, b_j) < \max p(a_k, b_j)$ i $p(a_i, b_j) < \max p(a_i, b_l)$. Za takve, ali i druge aminokiseline sparene u tom poravnanju, vjerojatnost povezivanja može biti vrlo mala. Ako imamo više parova simbola čija je vjerojatnost povezivanja veća od 0.5, garantirano je moguće konstruiranje poravnanja u kojima su svi ti parovi povezani. Jednom kad su fiksirani parove simbola s velikom vjerojatnošću povezivanja, te vjerojatnosti mogu se ponovno izračunati i za nastale podnizove te tako doći do još nekih parova čije bi sparivanje moglo biti značajno.

3 STRUKTURALNO PORAVNANJE (3D PORAVNANJE)

Strukturalnim poravnanjem pokušava se uspostaviti ekvivalencija između dvije ili više polimernih struktura bazirana na njihovom obliku i trodimenzionalnoj konformaciji. Ovaj postupak obično se primjenjuje na proteinske tercijarne strukture, ali može se upotrijebiti i za velike RNA molekule. U odnosu na jednostavnu strukturalnu superpoziciju, u kojoj barem neki ekvivalenti ogranci dvije strukture moraju biti poznati, strukturalno poravnanje ne zathijeva *a priori* određene ekvivalentne pozicije. Strukturalno poravnanje je funkcionalni alat za usporedbu proteina male sličnosti u sekvencama i proteina između kojih se ne može uspostaviti evolucijska poveznica standardnim tehnikama poravnanja. Ipak, kod tumačenja rezultata važno je uzeti u obzir mogućnost konvergentne evolucije različitih aminokiselinskih slijedova prema sličnoj, stabilnijoj tercijarnoj srukturi. Kao ulazni podaci uzimaju se setovi prostornih koordinata atoma dvije ili više sekvence u svim mogućim konformacijama. Izlazna informacija uspješnog strukturalnog poravnanja je set superpozicioniranih trodimenzionalnih koordinata svake ulazne strukture, primijenom metode najmanjeg korijena aritmetičke sredine kvadrata standardnih devijacija svake optimalno rotirane ili translatirane prostorne koordinate – RMSD (root mean square distance, slika 3.1) ili sofisticiranije metode, testa globalne udaljenosti – GDT(global distance test). RMDS dvije poravnane strukture govori o njihovoj divergenciji. Zajedničko za sve tipove strukturalnih poravnanja uzimanje je u obzir koordinata atoma okosnice strukture tj. koordinata atoma uključenih u peptidnu vezu, za prvi stupanj poravnanja. Tek ako su strukture gotovo identične, uzimaju se u obzir koordinate atoma izvan proteinske okosnice, a RMDS se računa i za sva rotacijska stanja bočnih ogranaka.

$$\theta_1 = \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ \vdots \\ x_{1,n} \end{bmatrix} \quad \text{and} \quad \theta_2 = \begin{bmatrix} x_{2,1} \\ x_{2,2} \\ \vdots \\ x_{2,n} \end{bmatrix}.$$

$$\mathrm{RMSD}(\theta_1, \theta_2) = \sqrt{\mathrm{MSE}(\theta_1, \theta_2)} = \sqrt{\mathrm{E}((\theta_1 - \theta_2)^2)} = \sqrt{\frac{\sum_{i=1}^{n} (x_{1,i} - x_{2,i})^2}{n}}.$$

(Slika 3.1)

3.1 CATH

CATH baza je hijerarhijska domena klasifikacije proteinskih struktura u PBD-u (protein data bank). Postoje četiri glavne hijerarhijske razine: razred, arhitektura, topologija i homologne superfamilije.

3.2 DALI

DALI (distance alignment matrix method) je metoda strukturalnog poravnanja koja se bazira na cijepanju ulaznih struktura u heksapeptidne fragmente te izračunu matrica udaljenosti Cα-atoma svakog mogućeg poravnanja. 2D matrica udaljenosti dva proteina računa se kroz serije preklapajućih submatrica veličine 6x6, te na kraju predstavlja 3D strukturu. Udaljenijim atomima se proporcionalno smanjuje značaj kako bi se izbjegli negativni efekti nastali zbog mobilnosti omči, torzije heliksa i ostalih manjih strukturnih varijacija. Pošto se DALI bazira na svi-sa-svima matrici udaljenosti, postoji mogućnost strukturne preraspodjele dijelova struktura nakon poravnanja. Izlazni podaci uključuju globalno optimalno poravnanje, suboptimalno poravnanje te vrijednost koja predstavlja stupanj sličnosti dva proteina.

3.3 SPDBV

Swiss-PDBViewer je program za 3D molekularno modeliranje, sposoban dodati ili izrezati motiv, promijeniti kovalentne veze, kuteve veza, konformacije ili ne-kovalentne interakcije. Koristi se za strukturalno poravnanje, modeliranje homologije, mutacijske molekularne modele, minimizaciju energije i dr.

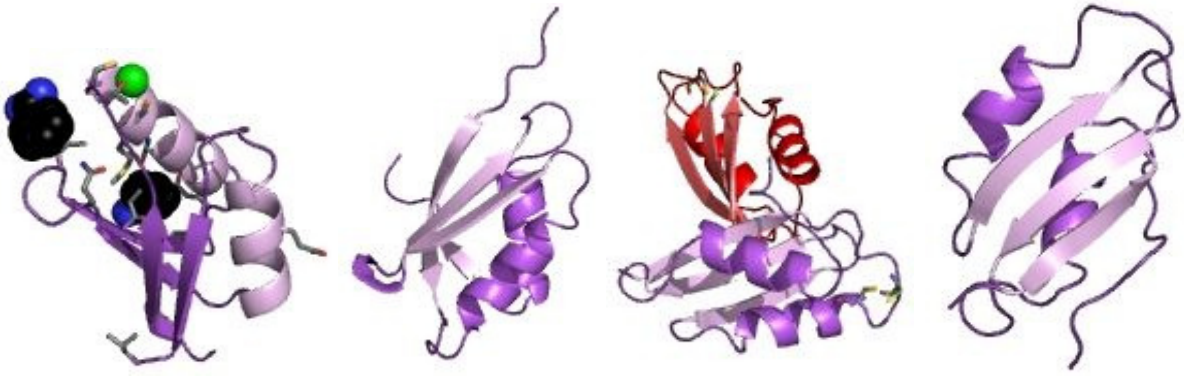
4. USPOREDBA 1D i 3D PORAVNANJA

Kada uspoređujemo sekvence dva srodna proteina, pretpostavljamo kako postoji jedinstveno optimalno poravnanje, iako ga je ponekad nemoguće naći. Oba proteina evoluirala su kroz serije mutacija-delecija-insercija iz zajedničkog pretka, te postoji precizan molekularni proces koji povezuje obje sekvence. Unatoč tome dva homologna proteina mogu evoluirati prema različitim konformacijama, njihovi fragmenti mogu promijeniti biološku funkciju, te strukturalno poravnanje dva udaljena homologa može biti različito od njihovog evolucijskog odnosa.

Provedena su 1D i 3D poravnanja nad različitim CATH homolognim superfamilijama proteina.

4.1 Membranski transportni proteini koji vežu atome metala

Strukture dvoslojnog beta|alpha "sendviča"



Saccharomyces cerevisiae (pivski/pekarski kvasac) | Bacillus subtilis (Gram + bakterija tla) | Cupriavidus metallidurans (nesporogen bacil) | Homo sapiens

(Slika 4.1.1)

<u>1D poravnanje Needleman-Wunsch algoritmom</u>

```
AEIKHYQFNVV-MTCSGCSGAVNKVLTKLEPDVSKIDISLEKQLVD-VYTT-L-PYDFIL    (S. Cerevisiae)
GEVV-LKMKVEGMTCHSCTSTIEGKIGKLQ-GVQRIKVSLDNQEATIVYQPHLISVEEMK    (H. Sapiens)

EKIKKTG-KE-VRSGKQL                                              (S. Cerevisiae)
KQIEAMGFPAFVKKIEGR                                              (H. Sapiens)
```

1D poravnanje Miyazawinim algoritmom

```
AEIKHYQFNVV-MTCSGCSGAVNKVLTKLEPDVSKIDISLEKQLVD-VYTT-L-PYDFIL     (S. Cerevisiae)
|||  |||||  ||||||||||||||||  ||||||||||||||  |   | ||||||
GEVV-LKMKVEGMTCHSCTSTIEGKIGKLQ-GVQRIKVSLDNQEATIVYQPHLISVEEMK     (H. Sapiens)

EKIKKTG-KE-VRSGKQL                                               (S. Cerevisiae)
|||||||  ||  |||||||
KQIEAMGFPAFVKKIEGR                                               (H. Sapiens)
```

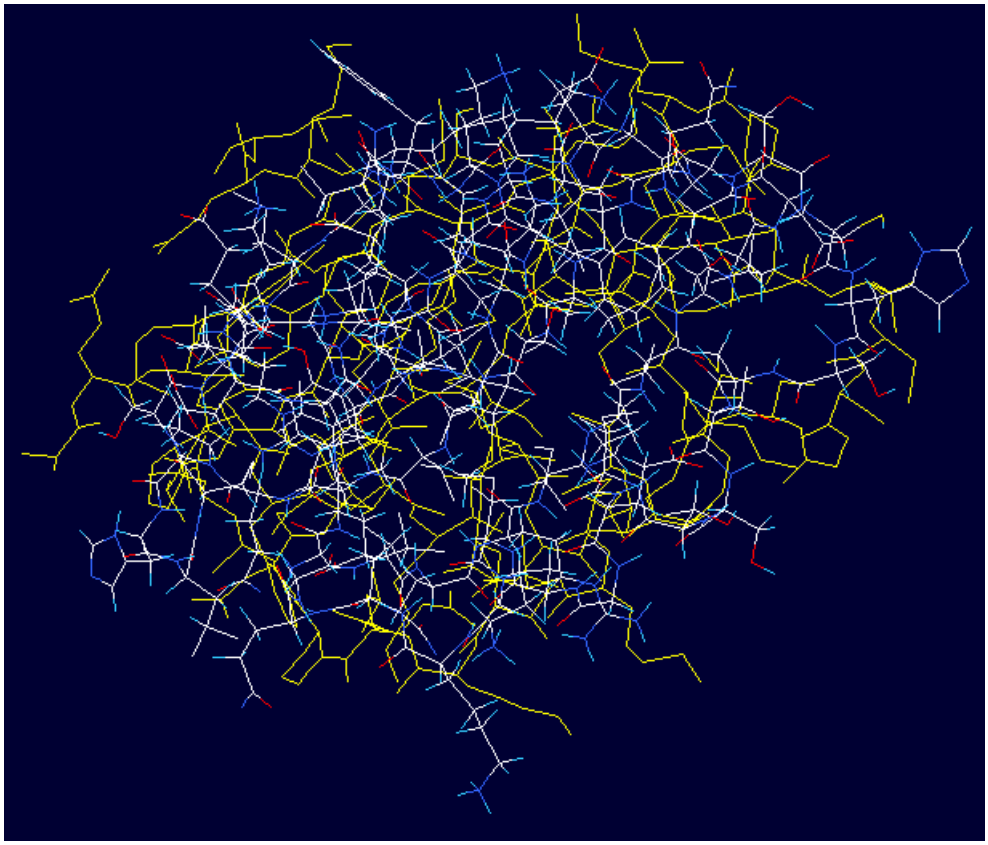
3D poravnanje DALI algoritmom

Z-score=9.3

```
DSSP  lLLEEEEEEEL.LLLHHHHHHHHHHHHLLLlLEEEEEEELLLLEEEEEEL...LLHHHHH
Query aEIKHYQFNVV.MTCSGCSGAVNKVLTKLEpDVSKIDISLEKQLVDVYTT...LPYDFIL   56
ident          |  |||  |         ||   |  |  ||  |
Sbjct .GEVVLKMKVEgMTCHSCTSTIEGKIGKLQ.GVQRIKVSLDNQEATIVYQphlISVEEMK   58
DSSP  .LLEEEEEELLlLLLLLLLHHHHHHHLLLL.LEEEEEEELLLLEEEEEELlllLLLLHHH
```

```
DSSP  HHHHLLLLLEEEEEEL..                                         (bočni lanci)
Query EKIKKTGKEVRSGKQL..   72                                    (S. Cerevisiae)
ident   |   |      |
Sbjct KQIEAMGFPAFVKKIEgr   76                                    (H. Sapiens)
DSSP  HHHHHHLLLLEELLLLll                                         (bočni lanci)
```

3D poravnanje SPDBV algoritmom



(Slika 4.1.2)

Žuto obojena struktura pripada Saccharomyces cerevisiae, a ostalo Homo sapiensu

```
 AEIKHYQFN VVMTCSGCSG AVNKVLTKLE PDVSKIDISL EKQLVDVYTT
GEVVLKMKVE GMTCHS-CTS TIEGKIGKLQ GVQRIKVSLD NQEATIVYQP
   -     -  -    * *-- ---  - **-            - -   **


LPYDFILEKI KKTGKEVRSG KQL
HLISVEEMKK QIEAMGFPAF VKKIEGR-
        *     --   -
```

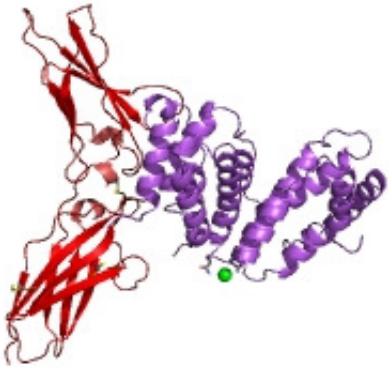
Rezultati dobiveni Needleman-Wunsch algoritmom i Miyazawinim algoritmom nad ovim homolognim strukturama su identični. Z-score vraćen DALIjevim algoritmom koji ima višu vrijednost od 2.0 je relevantan i ukazuje na homologiju nekih sekvenci unutar proteina. Z-score visoko preko 20.0 imaju strukture sa skoro identičnom okosnicom i manjom razlikom izmedu bočnih ogranaka. Iz Z-scora 9.3 može se zaključiti kako su metalovezajući proteini kvasca i čovjeka uistinu homologni. 1D poravnanja u odnosu na DALI poravnanje malo odstupaju, ali globalno gledajući poprilično su slična.

4.2 Citokini

Strukture alfa-zavojnice (up-and-down helix bundle)

Citokin - Granulocyte colony-stimulating factor

Bos taurus (domaća krava)

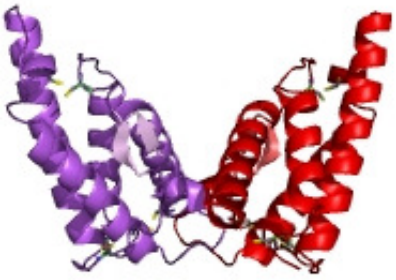


Interferon-gamma

Homo sapiens

HGH - somatotropin

Homo sapiens



Citokin - Interleukin-4 mutant

Homo sapiens

Citokin - rekombinantan Interleukin-22

Homo sapiens

(Slika 4.2)

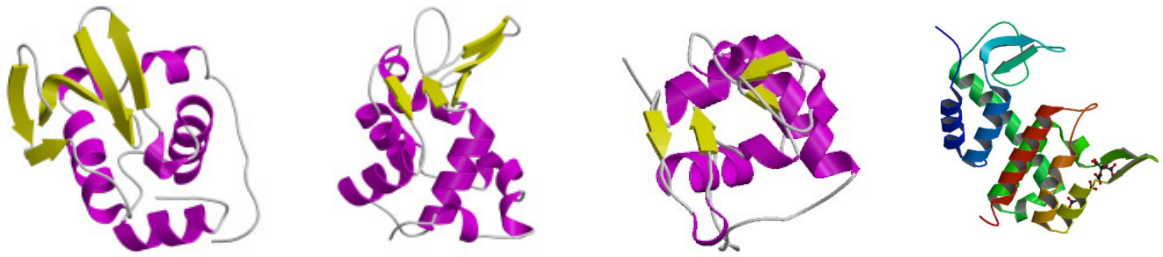
Miyazawin algoritam ne nalazi nikakvo optimalno strukturalno poravnanje interferona-γ sa somatotropinom, dok DALI daje loš izlazni Z-score. Needleman-Wunsch pri poravnanju u prvi niz ubacuje mnogo praznina. Ista stvar događa se kod 1D poravnanja somatotropina s interleukinom te somatotropina s rekombinantnim interleukinom. U ovim slučajevima izlazni score za 3D poravnanje iznosi oko 6.0. To bi moglo objasniti par fizioloških činjenica o tim proteinima. Interleukini (danas se preferira naziv citokini) čine grupu proteinskih hormona koji reguliraju efektorsku fazu imunosti (imunološki odgovor neposredno nakon "prepoznavanja" antigena od strane imunološkog sustava). Sintetiziraju ih mnoge stanice, a i ciljne su im stanice raspoređene po cijelom organizmu. Somatotropin ili hormon rasta hormon je prednjeg režnja hipofize. Njegova funkcija u organizmu je prije svega rast kostuju po dužini, povećani rast i razvoj tkiva, povećava iskorištavanje glukoze i aminokiselina u organizmu, dakle rast. Luči se tijekom djetinjstva, a prestaje se lučiti nakon zarastanja epifiza kostiju

(to se događa nakon puberteta) nakon čega više ne postoje mogućnosti za longitudinalan rast kostiju. Dakle, somatotropin možda ima neke strukturalne sličnosti s interleukinima, ali iz interpretacije rezultata 1D poravnanja proizlazi da njihove primarne strukture, tj. sekvence, ipak ne nalikuju jedna drugoj, ni prema aminokiselinskom slijedu ni prema funkciji.

Iz prethodnih rezultata možemo zaključiti kako je Miyazawin algoritam pouzdan za poravnanje visoko homolognih proteinskih nizova na 3D razini, tj. ako izlazni rezultat algoritma prikaže kvalitetno poravnate nizove mogli bismo zaključiti kako su te strukture poravnate i na 3D razini.

4.3 Enolaze i Lizozimi

Funkcionalnost poravnanja Miyazawinim algoritmom dodatno je provjerena na CATH superfamiliji enolaza koje imaju konformaciju TIM bureta, i lizozima. DALI poravnanje nekih enolaza rezultira vrlo visokim Z-scoreovima (i do 22.9), i u tom slučaju 1D poravnanje daje izlazni rezultat u obliku točno poravnatih nizova sa minimalnim brojem gapova. Lizozimi, također poznati pod nazivom N-acetilmuramid glikanhidrolaza, su familija enzima koja služi obrani organizma od bakterijskih stanica metodom oštećivanja njihove stanične stijenke. Nalaze se u mnogim tjelesnim izlučevinama poput suza i mlijeka, oplođenim jajnim stanicama domaće kokoši, probavnom traktu kućnih muha, a također su otkriveni i u nekim bakteriofagima.



H. sapiens sapiens Mus Musca domestic Enterobacteria phage p1

(Slika 4.3)

Već iz dobrog 1D poravnanja ljudskog i mišjeg lizozima mogli smo pretpostaviti dobro poravnanje na 3D razini, što se kasnije i potvrdilo. Z-score DALI poravnanja ljudskog i lizozima P1 faga iznosi 0.8, dok Miyazawin algoritam ne daje nikakve rezultate.

5. LITERATURA

I. David L. Nelson, Michael M. Cox : ''Lehninger Principles of Biochemistry'', Fourth Edition

II. R. Durbin, S. Eddy, A. Krogh, G Mitchison (2002) : ''Biological Sequence Analysis ''

III. Liisa Holm and Jong Park (Bioinformatics application notes, 2002) : ''DaliLite workbench for protein structure comparison''

IV. S Subbiah (Department of Applied Physics Stanford University & Bioinformatics Centre National University of Singapore) : ''An overview of the computational analysis of biological sequences''

V. Chan-Yong Park, Sung-Hee Park, Dae-Hee Kim, Seon-Hee Park, Chi-Jung Hwang : ''A Protein structure Retrieval System Using 3D LRA ''

VI. Liisa Holm, Chris Sander (Oxford university press,1999): ''Protein folds and families: sequence structure alignments''

VII. Adam Goodzik (Cambridge University Press 1996): ''The structural alignment between two proteins: Is there a unique answer?''

VIII. Shann-Ching Chen and Tsuhan Chen (Carnegie Mellon University) : ''Retrieval of 3D protein structures''

IX. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/

X. Wikipedia, the free encyclopedia, http://en.wikipedia.org

XI. Algoritmi iz bioinformatike, http://web.math.hr/~payo/vjezbe.pdf

XII. CATH, http://www.cathdb.info/

XIII. PDB sum, DaliLite, http://www.ebi.ac.uk

XIV. Swiss-PdbViewer DeepView, http://spdbv.vital-it.ch/

6. SAŽETAK

U ovom radu bavimo se usporedbom strukturalnog poravnanja (3D) i poravnanja nizova (1D). Poznato je kako proteinske strukture evoluiraju mnogo sporije nego nizovi, stoga pronalazak “pravog” (strukturalnog) poravnanja iz primarne strukture postaje teže kada se bavimo udaljenim homolozima. Pokazano je kako Miyazawin algoritam1D poravnanje provodi točnije nego standardne metode poput Needleman-Wunsch, Smith-Waterman ili BLAST algoritma.

7. SUMMARY

This work is concerned with a comparison between structural and sequence alignment. It is well known that protein structures evolve more slowly than sequences, so detecting the “right” (i.e. structural) alignment from the sequence information only becomes increasingly difficult when dealing with distant homologues. It is shown that Miyazaw’s algorithm for sequence alignment performs this task more accurately than standard methods, such as Needleman-Wunsch, Smith-Waterman or BLAST.