SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 68

ANALIZA SEKUNDARNE STRUKTURE PROTEINA METODOM OBRADE SIGNALA

Jura Ćurić

Zagreb, lipanj 2010

Hvala svim ljudima koji su mi pomogli tokom studija i prilikom izrade diplomskog rada, bilo svojim znanjem, bilo dobrom voljom i optimizmom. Posebno se zahvaljujem svom mentoru Mili Šikiću koji je uvijek bio spreman pomoći ne samo kao profesor, već kao i prijatelj.

Sadržaj

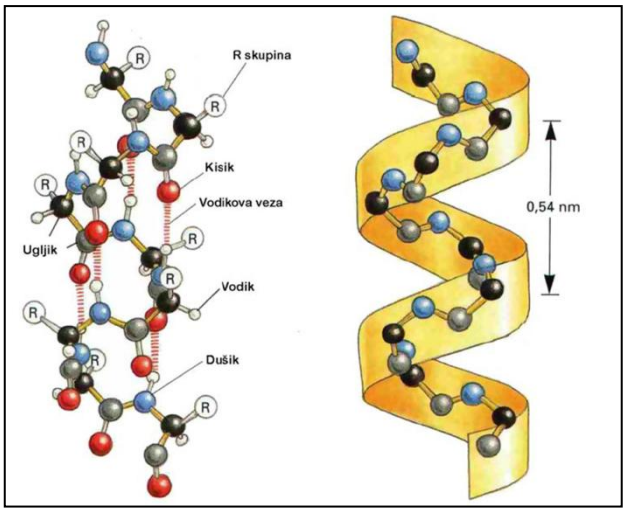
1. Uvod

Proteini imaju ključnu ulogu u gotovo svim biološkim procesima. Ta uloga definirana je funkcijom proteina koju s druge strane određuje njegova struktura. Proučavanje svojstava, interakcija i funkcije proteina zadatak je proteomike, znanstvene discipline čiji je cilj opisati ukupnost proteina koji tvore organizme (proteome). Proteini su kompleksne organske strukture koje se sastoje od aminokiselina povezanih peptidnim vezama, a čiji je slijed određen genima koji ih kodiraju. Istraživanje genoma urodilo je spoznavanjem ogromnog broja aminokiselinskih sljedova (sekvenci) koje geni kodiraju. Ipak, funkcija, struktura i interakcije proteina pripadnih sljedova uglavnom su nepoznate. Linearan niz aminokiselina koje tvore protein uvija se u specifičnu trodimenzionalnu strukturu koja određuje njegovu funkciju. Zbog toga se ulaže golem napor da se te strukture odrede, bilo eksperimentalno, bilo računski. Zadaća je biotehnologa da eksperimentalno odrede strukture proteina metodama rentgenske kristalografije i nuklearne magnetske rezonance. To su skupe, sofisticirane i vremenski zahtjevne metode pa je sve veći nesrazmjer između broja poznatih sljedova i pripadnih struktura. Stoga je nužno stvoriti, unaprijediti i iskoristiti računske metode koje će na temelju poznatih aminokiselinskih nizova vjerodostojno predvidjeti konačnu strukturu proteina. Unatoč desetljećima rada i istraživanja na tome području, problem predviđanja trodimenzionalne strukture na temelju slijeda ostaje neriješen.

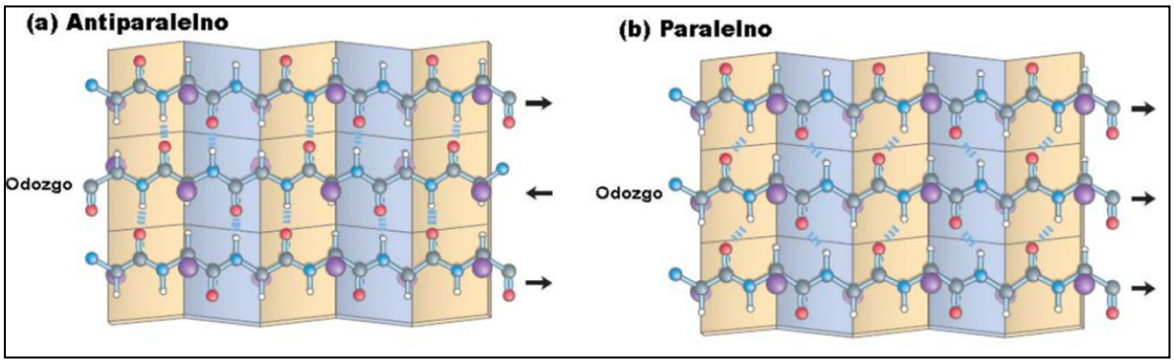
Ovaj rad je fokusiran na predviđanje sekundarne strukture kao i strukturnih razreda sekvenci aminokiselina proteinskih lanaca pomoću metoda obrade signala. Konkretno, na temelju primarne strukture proteina, tj. slijeda aminokiselina njegovih peptidnih lanaca stvaraju se na temelju nekog svojstva aminokiselina signali nad kojima će biti vršena obrada, ne bi li se uspjela izlučiti značajka koja bi pomogla u određivanju tipova sekundarne strukture pojedinih aminokiselina ili u nešto jednostavnijem slučaju, odredio strukturni razred pojedine sekvence.

2. Primarna i sekundarna struktura proteina

Proteini su građeni od jednog ili više polipeptidnih lanaca. Redoslijed aminokiselina u lancima proteina čini njegovu *primarnu* strukturu. Ukupno postoji 20 aminokiselina**.** Dijelovi polipeptidnog lanca pod utjecajem uglavnom vodikovih veza oblikuju (smotavaju se) specifične oblike. Takve lokalne konformacije dijelova polipeptidnog lanca čine elemente *sekundarne* strukture proteina. Uobičajene sekundarne strukture su α-uzvojince i β-ploče povezane zavojitim područjima različitih duljina i nepravilnih oblika. U prosjeku više od 50% svih aminokiselinskih rezidua u proteinima otpada upravo na te dvije strukture.



Slika 1 – α-uzvojnica



Slika 2 – β-ploča

2.1. Sekundarna struktura i njeno određivanje

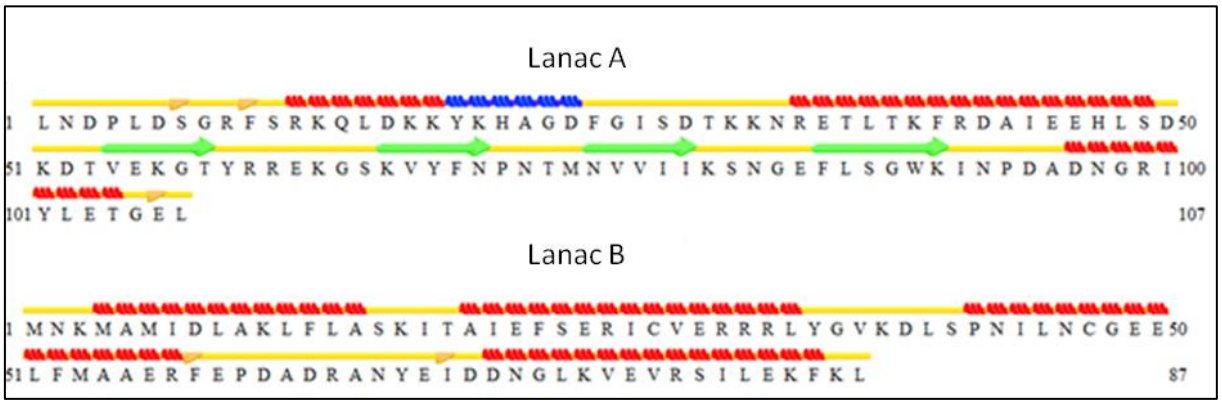
Trodimenzionalne strukture, odnosno položaji pojedinih atoma unutar lanaca ne pružaju direktnu informaciju o sekundarnoj strukturi nužnu za izgradnju skupova za treniranje i testiranje odabranih metoda predviđanja. Međutim, elementi sekundarne strukture definirani su svojim fizikalno-kemijskim parametrima i moguće ih je odrediti iz kristalne strukture proteina. Uz poznate 3D koordinate atoma proteinske strukture moguće je odrediti kojoj od sekundarnih struktura pripada pojedini aminokiselinski ostatak. Periodične sekundarne strukture proizvode pravilnosti koje se očituju u nekim prostorno-kemijskim uzorcima te se mogu iskoristiti za njihovo međusobno razlikovanje. Od mnogih postojećih računalnih metoda, za potrebe ovog rada odabrane su DSSP (*Dictionary of Secondary Structure in Proteins*) i STRIDE (*Structural identification*) metode za automatizirano pridruživanje sekundarne strukture. Obje metode na temelju prepoznavanja uzoraka vodikovih veza i geometrijskih svojstava razvrstavaju aminokiselinske ostatke u jedan od osam tipova sekundarne strukture. Sedam karakterističnih elemenata sekundarne strukture uključuju α-uzvojnicu (*H*), β-ploču (*E*), α_5 (*G*) i π-uzvojnicu (*I*), izolirani β-most (*B*), luk (*S*), zavoj (*T*), dok se ostale rezidue tretiraju kao nestandardne.

Tabela 1 – tipovi sekundarne strukture

DSSP kod	Opis
H	α-uzvojnica
G	α_5-uzvojnica
E	β-ploča
B	izolirani β-most
I	π-uzvojnica
T	zavoj
S	luk
-	Omča/namotaj

2.2 Od primarne do sekundarne strukture

Nakon što su objašnjeni pojmovi primarne i sekundarne strukture proteina, potrebno je sagledati sve moguće informacije koje nam nudi primarna struktura kako bi na temelju nje odredili sekundarnu. U konačnici je potrebno za svaku aminokiselinu odrediti kojem tipu sekundarne strukture pripada. Na sljedećoj slici prikazane su aminokiseline lanca A i B proteina 1V74 zajedno sa pripadnim sekundarnim strukturama:



Slika 3 – lanci proteina 1V74 sa pripadnom primarnom i sekundarnom strukturom (STRIDE)

gdje su sekundarne strukture predstavljene obojanim simbolima sa sljedećim značenjima:

Slika 4 – DSSP/STRIDE tipovi sekundarnih strukutra

Ono što će u daljnjem tekstu biti prezentirano jesu istraživanja i pokušaji određivanja gore navedenih tipova sekundarne strukture pomoću metoda obrade signala. Također, bit će objašnjeno kako primarne strukture, tj. niza aminokiselina dobiti signale pomoću kojih i nad kojima će se vršiti obrada.

2.3. Od primarne strukture do signala

Primarna struktura proteina, kako je navedeno predstavlja niz aminokiselina i najčešće je predstavljena simbolima aminokiselina. Npr:

LNDPLDSGRFSRKQLDKKYKHAGDFGISDTKKNRETLTKFRDAIEEHL

Naravno, ovakav slijed slova nije pogodan za numeričku obradu pa je stoga ovom nizu potrebno pridijeliti numeričke vrijednosti. Potrebno je uzeti nekakvu mjeru koja će biti povezana sa primarnom i sekundarnom strukturom te na pomoću te mjere niz aminokiselina transformirati u prikladan signal. Jedno moguće svojstvo jest polarnost bočnih lanaca aminokiselina koja određuje hoće li aminokiselina biti hidrofilna ili hidrofobna. Kao sastavne komponente proteina, aminokiseline se nalaze okružene vodom koja je polarno otapalo. Dakle, za očekivati je da će polarni dijelovi proteinskih lanaca težiti da budu okruženi vodom. Suprotno, nepolarni dijelovi će se savijati prema unutrašnjosti proteina kako bi smanjili površinu dodira s vodom. Budući da polarnost bočnog lanca utječe na oblik lanca, za pretpostaviti je da time utječe i na njegovu sekundarnu strukturu. Dakle, potrebne su numeričke informacije koje govore o polarnosti bočnog lanca aminokiseline. To svojstvo se naziva indeks hidrofobnosti. U sljedećoj tablici dane su vrijednosti indeksa hidrofobnosti za svaku aminokiselinu:

Tabela 2 – Popis aminokiselina zajedno s indeksom hidrofobnosti

Aminokiselina	Kratica	Simbol	Polarnost	Indeks hidrofobnosti
Alanin	ALA	A	Nepolarna, hidrofobna	1.8
Arginin	ARG	R	Polarna, hidrofilna	-4.5
Asparagin	ASN	N	Polarna, hidrofilna	-3.5
Asparginska kiselina	ASP	D	Polarna, hidrofilna	-3.5
Cistein	CYS	C	Polarna, hidrofilna	2.5
Glutamin	GLN	Q	Polarna, hidrofilna	-3.5
Glutaminska kiselina	GLU	E	Polarna, hidrofilna	-3.5
Glicin	GLY	G	Polarna, hidrofilna	-0.4
Histidin	HIS	H	Polarna, hidrofilna	-3.2
Izoleucin	ILE	I	Nepolarna, hidrofobna	4.5
Leucin	LEU	L	Nepolarna, hidrofobna	3.8
Lizin	LYS	K	Polarna, hidrofilna	-3.9
Metionin	MET	M	Nepolarna, hidrofobna	1.9
Fenilalanin	PHE	F	Nepolarna, hidrofobna	2.8
Prolin	PRO	P	Nepolarna, hidrofobna	1.6
Serin	SER	S	Polarna, hidrofilna	-0.8
Treonin	THR	T	Polarna, hidrofilna	-0.7
Triptofan	TRP	W	Nepolarna, hidrofobna	-0.9
Tirozin	TYR	Y	Polarna, hidrofilna	-1.3
Valin	VAL	V	Nepolarna, hidrofobna	4.2

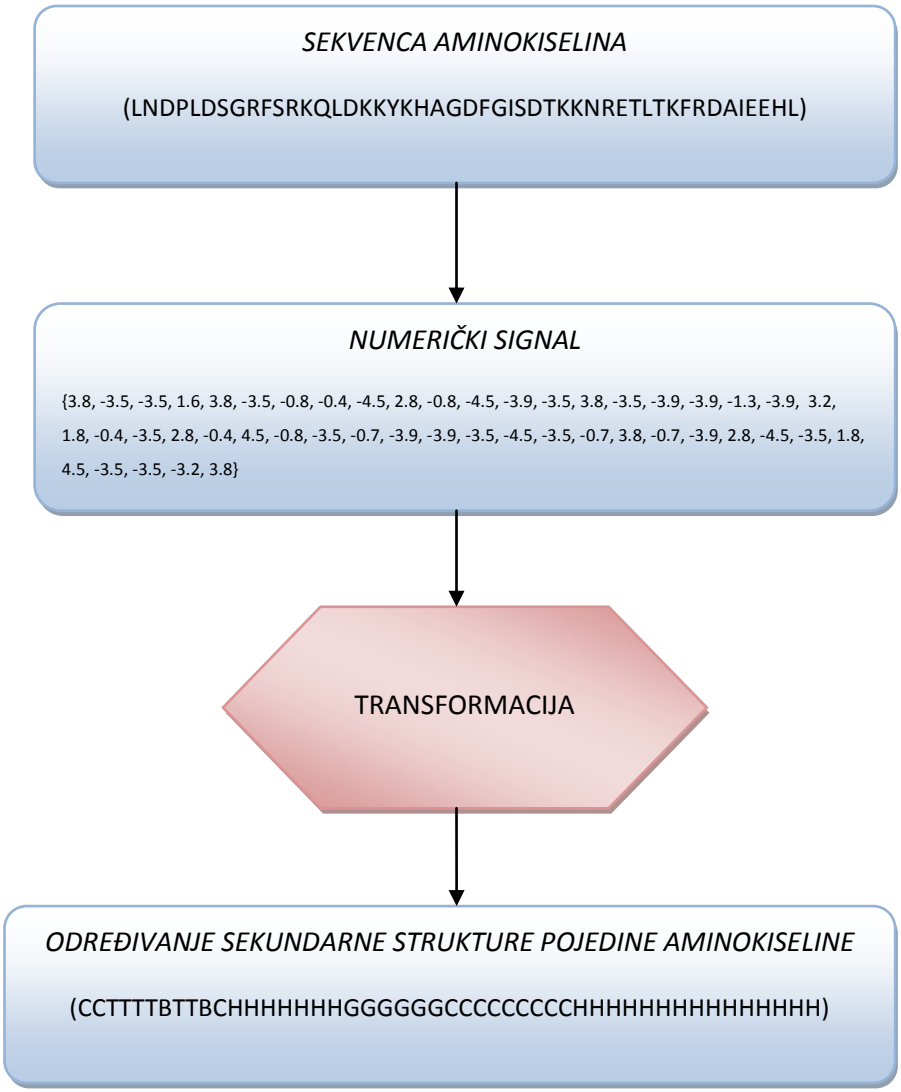
Na temelju danih vrijednosti možemo prije navedeni niz aminokiselina:

LNDPLDSGRFSRKQLDKKYKHAGDFGISDTKKNRETLTKFRDAIEEHL

transformirati u diskretni signal:

{3.8, -3.5, -3.5, 1.6, 3.8, -3.5, -0.8, -0.4, -4.5, 2.8, -0.8, -4.5, -3.9, -3.5, 3.8, -3.5, -3.9, -3.9, -1.3, -3.9, 3.2, 1.8, -0.4, -3.5, 2.8, -0.4, 4.5, -0.8, -3.5, -0.7, -3.9, -3.9, -3.5, -4.5, -3.5, -0.7, 3.8, -0.7, -3.9, 2.8, -4.5, -3.5, 1.8, 4.5, -3.5, -3.5, -3.2, 3.8}

Signal je dobiven tako da se svaka oznaka aminokiseline zamijenila s njenom vrijednosti hidrofobnosti. Očito, dobiveni signal je jednake duljine kao i znakovna sekvenca. Nad dobivenim signalom je sad moguće primijeniti razne transformacije koje mogu otkriti korisnu informaciju o sekundarnoj strukturi pojedinih aminokiselina. Dakle, konačni je cilj na temelju rezultata transformacije automatski odrediti sekundarnu strukturu svake aminokiseline:



3. Fourierova analiza

Osnovna transformacija koja se sama po sebi nameće kao prvi izbor jest brza Fourierova transformacija gdje se numerički signal x rastavlja na spektralne komponente oblika:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi k \frac{n}{N}} \qquad k = 0,...,N-1 \tag{1}$$

Izračunati spektar sadrži onoliko uzoraka koliko ih sadrži i ulazni signal. Pomoću amplitudnog spektra dobiva se uvid u zastupljenost pojedinih kompleksnih eksponencijala u originalnom signalu, tj. zastupljenost funkcija oblika:

$$e^{j2\pi \frac{k}{N}} \qquad k = 0,...,N \tag{2}$$

Dakle, ovisno u koliko točaka se računa FFT, tolika će biti numerička rezolucija spektra. Drugim riječima, frekvencija kompleksnih eksponencijala biti će finije podijeljena u intervalu $[0, 2\pi]$.

Pretpostavka korištenja gore navedene transformacije jest da neki signal dobiven na temelju njegove primarne strukture, a čiji uzorci (odn. aminokiseline) pripadaju nekom tipu sekundarne strukture ima izražene druge spektralne komponente od nekog drugog signala čiji uzorci pripadaju nekom drugom tipu sekundarne strukture. Kao što će kasnije biti pokazano, većina ovog rada biti će usmjerena upravo na pronalaženje specifičnih spektralnih komponenti koje karakteriziraju uzorke signala koji pripadaju pojedinom tipu sekundarne strukture.

Budući da će nam na raspolaganju biti razni skupovi podataka čiji stvoreni signali neće biti striktno sinusnog oblika (kao što je to slučaj s tonovima u glazbi) potrebno je definirati određene mjere nad skupom spektara kako bi lakše bilo odrediti neke zajedničke značajke, tj. spektralne komponente nekog proizvoljnog skupa signala. Tri mjere koje će se koristiti jesu *Multiple cross-power spectrum* i njena logaritamska inačica, *Energy spectrum* te statistika spektralnih komponenti (srednja vrijednost i standardna devijacija). Objašnjenja mjera su dana u nastavku.

3.1. *Multiple cross-power spectrum* (MCPS) – umnožak komponenti spektara

Pretpostavimo da na raspolaganju imamo n spektara od kojih je svaki duljine N uzoraka:

$$\begin{aligned} X^1 &= \left\{ X_1^{(1)}, X_2^{(1)}, ..., X_{N-1}^{(1)}, X_N^{(1)} \right\} \\ X^2 &= \left\{ X_1^{(2)}, X_2^{(2)}, ..., X_{N-1}^{(2)}, X_N^{(2)} \right\} \\ &... \\ X^n &= \left\{ X_1^{(n)}, X_2^{(n)}, ..., X_{N-1}^{(n)}, X_N^{(n)} \right\} \end{aligned} \tag{3}$$

Tada je njihov *MCPS* definiran kao signal koji na i-tom uzorku ima vrijednost umnoška i-tih komponenti svih spektara, tj:

$$MCPS_i = \prod_{k=1}^{n} X_i^{(k)} \tag{4}$$

Ova mjera je veoma korisna kod utvrđivanja zajedničkih spektralnih komponenata, tj. onih koje su prisutne u svim spektrima. Naime, dovoljno je da samo jedna i-ta komponenta nekog spektra ima vrijednost 0 da i-ta komponenta rezultirajućeg MCPS-a poprimi vrijednost 0. S druge strane, one komponente koje su prisutne kod svih spektara neće iščeznuti. Ova mjera je stoga pogodna za eliminaciju ne-zajedničkih spektralnih komponenata i još bitnije, određuje spektralne komponente koje se pojavljuju u svakom spektru s dovoljnim intenzitetom da se množenjem više istaknu od ostalih koje nisu nužno jednake nuli.

Budući da broj spektara može biti velik, zbog ograničene numeričke preciznosti komponente MCPS-a što će najčešće biti slučaj, mogu divergirati. Tada je poželjno prije računanja MCPS-a sve spektre normirati po najvećoj vrijednosti. Tada će se vrijednosti MCPS-a sigurno kretati u intervalu $[0,1]$. Budući da je nerijedak slučaj da će komponente tada biti veoma male (reda veličine 1e-100), korisno je logaritmirati vrijednosti MCPS-a te na taj način jasnije vidjeti zajedničke komponente.

3.2. Energy spectrum (ENE)

Kod MCPS-a su se spektralne komponente najprije normirale po amplitudi i potom pomnožile. Time se postigao efekt eliminacije ne-zajedničkih spektralnih komponenti i isticanja zajedničkih komponenti dovoljno velikog intenziteta. Kod energetskog spektra svaki se spektar najprije normira po energiji, tj. svaka komponenta spektra se podijeli korijenom ukupne energije spektra. Normiranje je korisno jer nam apsolutne veličine pojedinih komponenata nisu bitne, već je bitna distribucija energije po spektru, tj. relativni odnosi između pojedinih dijelova spektra. Valja uočiti da su spektri kao i signali diskretni i konačni. Energija diskretnog signala je definirana kao:

$$E = \sum_{i=1}^{L} x_i^2 \tag{5}$$

gdje je L duljina signala. Nakon normiranja po energiji zbrajaju se i-te komponente svih normiranih spektara. Dakle, energetski spektar na i-tom mjestu ima vrijednost:

$$ENE_i = \sum_{k=1}^{n} Xnorm_i^{(k)} \tag{6}$$

Ova mjera se dobiva sumacijom pa zbog toga nema eliminacijski karakter, već prikazuje raspodjelu energije zadanog skupa spektara po komponentama. Komponente koje su uočljive na MCPS-u će uglavnom biti uočene i na ENE mjeri uz jednu iznimku. Naime, moguće je da neki gotovo svi spektri skupa imaju veoma izraženu određenu komponentu spektra, a samo jedan spektar na toj komponenti ima vrijednost 0. Tada će MCPS mjera na tom mjestu imati vrijednost 0, no ENE mjera će zbog sumacijskog karaktera poprimiti veliku vrijednost. Veliki nesrazmjeri ENE i MCPS mjere upućuju na postojanje *outliera,* tj. spektara koji odudaraju od ostalih.

3.3. Statistika spektralnih komponenti

Statistika spektralnih komponenti uključuje izračun srednje vrijednosti i standardne devijacije pojedine spektralne komponente. Navedene vrijednosti se računaju na temelju svih spektara danog skupa:

$$mean_i = \frac{1}{N}\sum_{k=1}^{N} X_i^{(k)} \tag{7}$$

$$std_i = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(X_i^{(k)} - mean_i\right)^2} \tag{8}$$

I kod računanja ovih mjera je poželjno normirati spektre (npr. po amplitudi) kako bi izbjegli rezultate koji su posljedica različitih apsolutnih vrijednosti. Pomoću srednjih vrijednosti komponenta moguće je vidjeti trend koji slijede komponente spektara dok standardna devijacija opisuje odstupanje komponenti od srednje vrijednosti.

3.4. FFT spektar

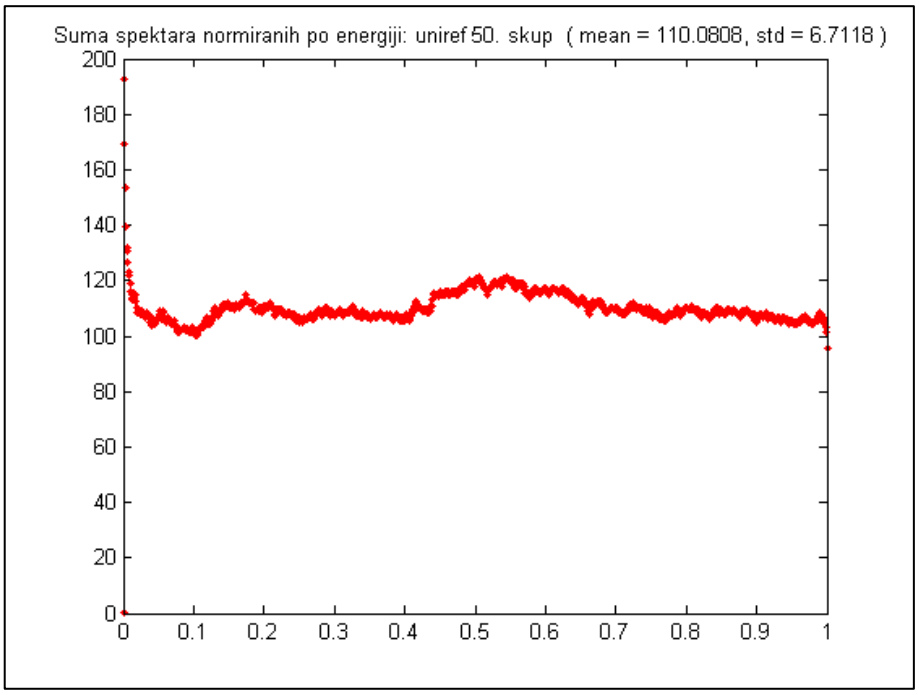
Sve navedene mjere biti će korištene nad **amplitudnim spektrom** diskretnih signala. Razlog tomu je taj što nas zanima isključivo udio kompleksnih eksponencijala u promatranom signalu. Fazni spektar govori o faznom pomaku kompleksnih eksponencijala koji nam za sad neće biti bitan. Dodatno, budući da su signali koji se dobivaju iz sekvenci realni, amplitudni spektar im je paran pa je dovoljno analizirati samo pola uzoraka dobivenog amplitudnog spektra. Stoga će sve mjere kao i sami spektri biti proučavani u frekvencijskom području $[0,\pi]$ što će na grafovima biti radi preglednosti prikazano intervalom $[0,1]$ gdje se podrazumijeva interval $[0,\pi]$. Također, bitno je uočiti da svi spektri moraju biti iste duljine. Budući da duljine proteinskih lanaca poprimaju različite vrijednosti, potrebno je sve signale dopuniti nulama (*zero padding*) tako da konačna duljina svakog signala bude sljedeća potencija broja dva veća od najdulje sekvence u promatranom skupu. Svođenjem signala na istu duljinu osiguravamo jednake duljine spektara, a ako je ta duljina potencija broja dva, izračun fft-a će biti brži. Također, dopunjavanjem signala nulama u spektar ne unosimo novu informaciju samo zato što se povećao broj uzoraka spektra. Na taj način se samo povećava numerička rezolucija spektra što je rezultat finijeg uzorkovanja frekvencija

kompleksnih eksponencijala (vidi formulu za FFT, frekvencija kompleksnih eksponencijala je u intervalu $[0, 2\pi]$ podijeljena na N dijelova gdje je N duljina signala).

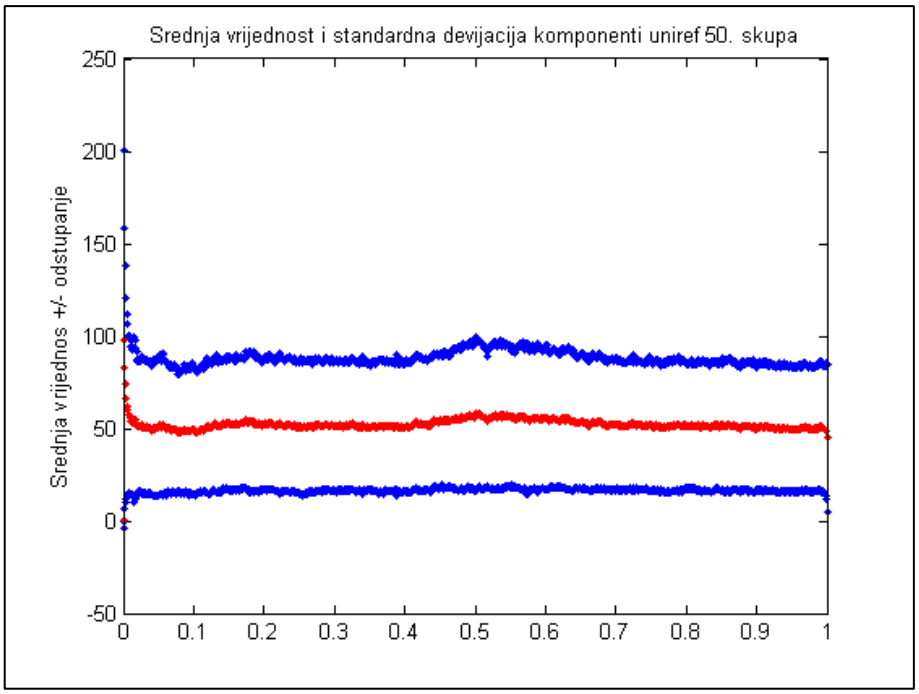
4. Baza sekvenci Uniref50

Uniref je baza proteinskih sekvenci gdje nam je dostupna informacija o primarnoj strukturi lanaca pojedinih proteina. Za inicijalno istraživanje korištena je baza Uniref50 koja sadrži sve poznate sekvence aminokiselina, a koje su maksimalno 50% slične čime je uvelike smanjena redundantnost.

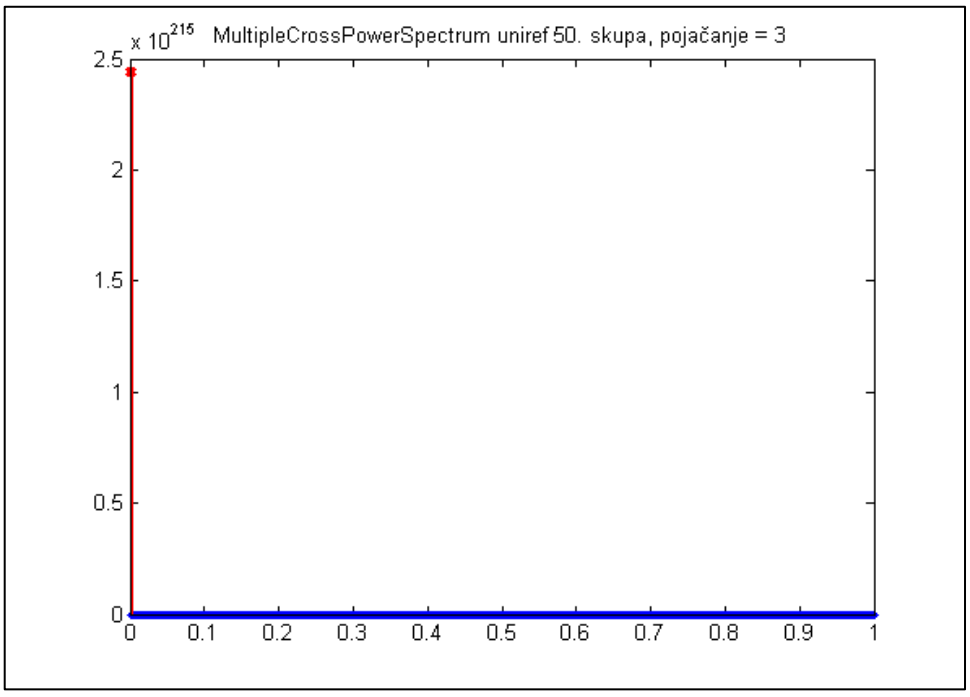
Inicijalna ideja jest sljedeća: uzeti skup sekvenci iz baze Uniref50 i pogledati MCPS i ENE mjere te statistiku spektralnih komponenti i vidjeti hoće li navedene mjere izlučiti neki skup frekvencija. Korišten je slučajni skup veličine 4000 signala od kojih je svaki duljine 2048 uzoraka (zajedno sa dodanim nulama). Rezultati su prikazani na sljedećim slikama gdje os apscise predstavlja diskretnu frekvenciju u intervalu od [0,1] pi.



Slika 5 – ENE mjera spektara signala sekvenci Uniref50 skupa



Slika 6 – statistika komponenti spektara signala sekvenci Uniref50 skupa



Slika 7 - MCPS mjera spektara signala sekvenci Uniref50 skupa

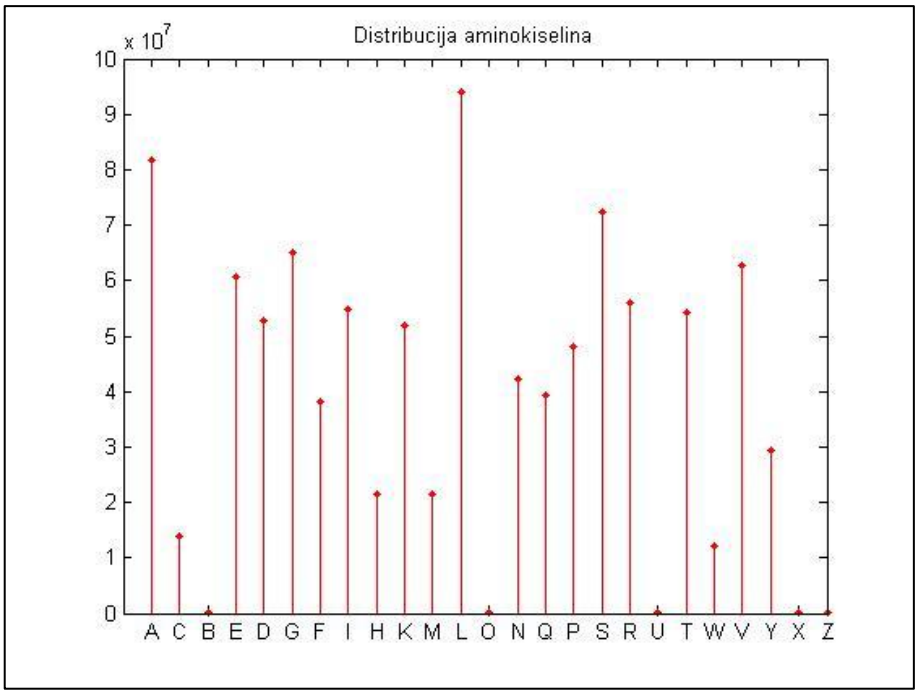
Na temelju vizualiziranih mjera može se zaključiti da su frekvencije oko 0.5*pi zajedničke. Na MCPS slici pojavljuje se i faktor pojačanja koji predstavlja faktor s kojim se spektar množi nakon normiranja kako uzastopnom multiplikacijom komponenta ne bi iščeznula zbog veoma malih vrijednosti bliskih nuli. MCPS mjera pokazuje isticanje samo niskih frekvencija pa njih u startu možemo smatrati irelevantnima u određivanju sekundarne strukture, budući da je MCPS računan na skupu sekvenci sa svim mogućim tipovima sekundarne struktura. Očito, niske frekvencije bliske nuli zajedničke su svim tipovima sekundarnih struktura.

4.1. Umjetno generirani signali na temelju Uniref50 razdiobe aminokiselina

Jedno od temeljnih pitanja koje se postavlja nakon promatranja slika prethodno izračunatih mjera jest nose li podatci na temelju kojih su mjere izračunate ikakvu informaciju? Naravno, u slučaju negativnog odgovora svako daljnje istraživanje temeljeno na FFT analizi bilo bi najvjerojatnije uzaludno. Zbog navedenog razloga proveden je izračun mjera, ali na umjetno stvorenim signalima.

4.1.1. Umjetno stvoreni signali

Ideja je sljedeća: potrebno je pogledati razdiobu aminokiselina sekvenci u cijeloj Uniref50 bazi i na temelju te razdiobe stvarati umjetne sekvence kod kojih će svaka aminokiselina imati vjerojatnost pojavljivanja proporcionalnu sa njenom zastupljenosti u Uniref50 bazi.

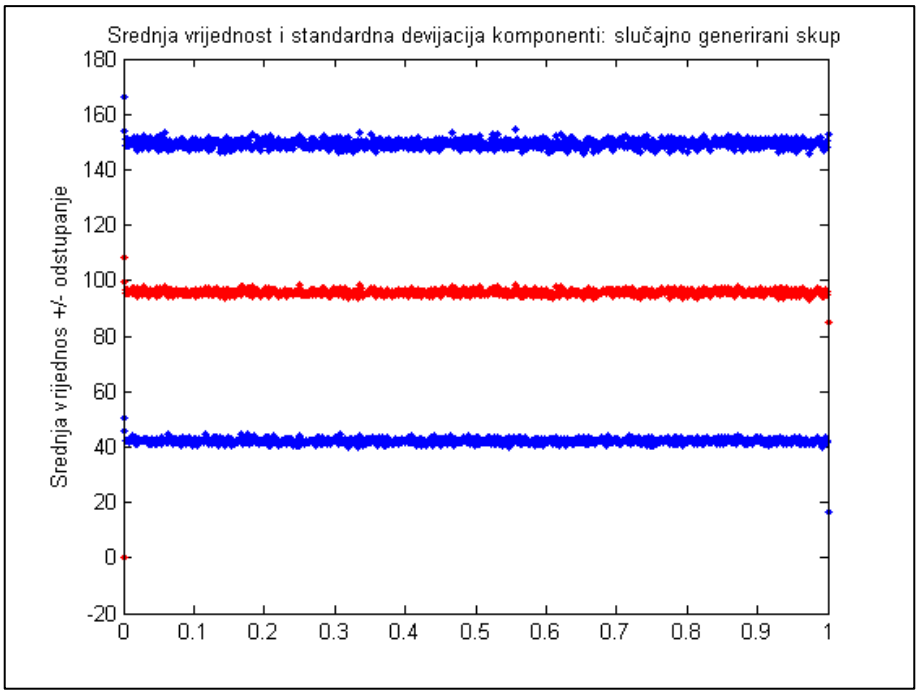


Slika 8 – zastupljenost aminokiselina u Uniref50 bazi

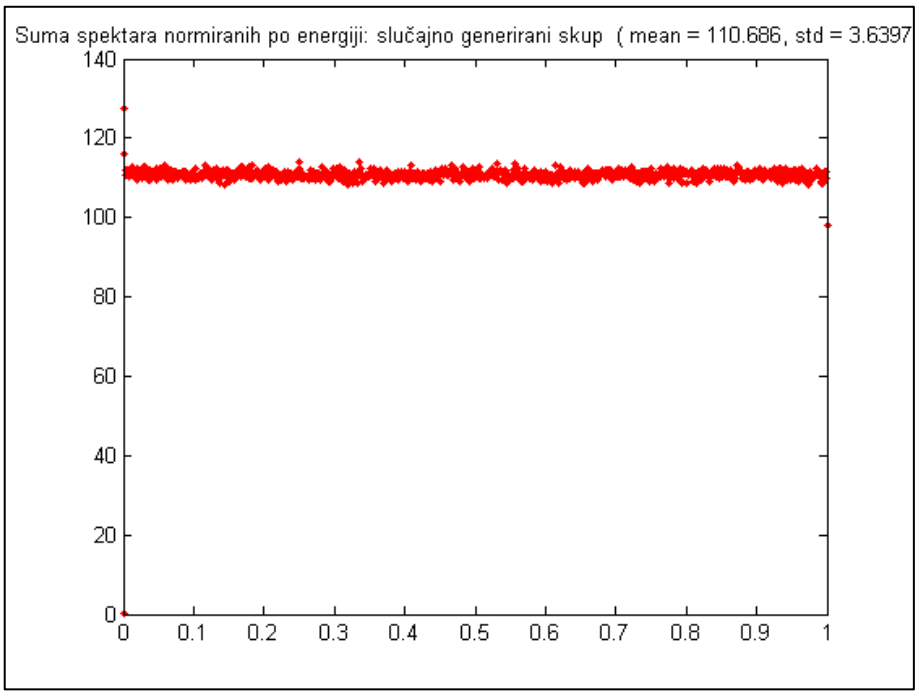
Dakle, što je broj pojavljivanja neke aminokiseline u Uniref50 bazi veći to će biti veća vjerojatnost da umjetni signal na i-tom mjestu ima upravo tu aminokiselinu. Ta vjerojatnost je jednaka:

$$p_{\text{aminokiseline}} = \frac{\text{broj pojavljivanja aminokiseline u bazi}}{\text{broj svih aminokiselina u bazi}} \tag{9}$$

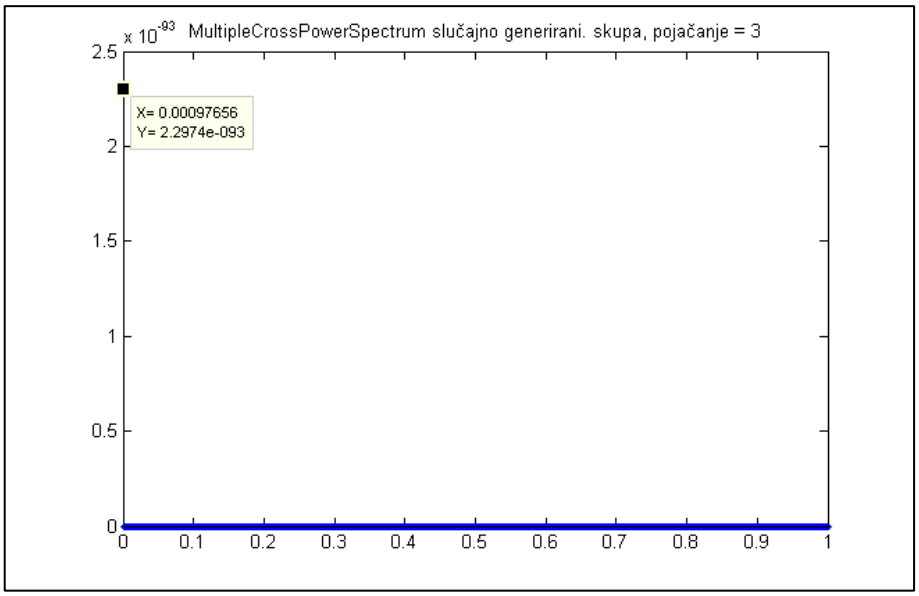
Duljine signala su pak generirane uniformno u intervalu između [512, 2048]. Na temelju slučajno generiranih signala dobivene su sljedeće mjere:



Slika 9 – Statistika spektara signala slučajno generiranih sekvenci



Slika 10 – ENE mjera spektara signala slučajno generiranih sekvenci



Slika 11 - MCPS mjera spektara signala slučajno generiranih sekvenci

Mjere dobivene na temelju umjetno generiranog skupa ne pokazuju nikakve razlike u različitim frekvencijskim pojasevima što vodi do zaključka da se slučajno generirani signali ponašaju kao šum dok se u originalnim podacima iz Uniref50 baze ipak nalazi nekakva informacija.

5. Strukturni razredi sekvenci

Jedna od zadaća kod određivanja sekundarne strukture proteina jest odrediti njihov *strukturni razred*. Postoje pet glavnih skupina, tj. strukturna razreda u koje možemo smjestiti sve sekvence: sve-α (skupina *a*), sve-β (skupina *b*), α/β (skupina *c*) i α+β (skupina *d*) i *ζ* (neraspoređeni). Razred *a* predstavlja sve sekvence koje su po tipu sekundarne strukture pretežno α-uzvojnice i obratno, razred *b* predstavlja sve sekvence koje većinski pripadaju β-pločama. Skupovi *c* i *d* sadrže manje više podjednaku količinu α–uzvojnica i β-ploča s tom razlikom da se u skupu *c* nalaze ispremiješane α–uzvojnice i β-ploče dok su u skupu *d* one striktno odvojene. Za ovo istraživanje korišten je skup sekvenci FC699 koji sadrži podatke o primarnoj i sekundarnoj strukturi te o pripadnosti svake sekvence pojedinom strukturalnom razredu. U ovoj bazi se koriste tri osnovna tipa sekundarne strukture (α–uzvojnice, β-ploče i coilovi). Raspodjela sekundarnih struktura u skupu FC699 dana je u sljedećoj tablici:

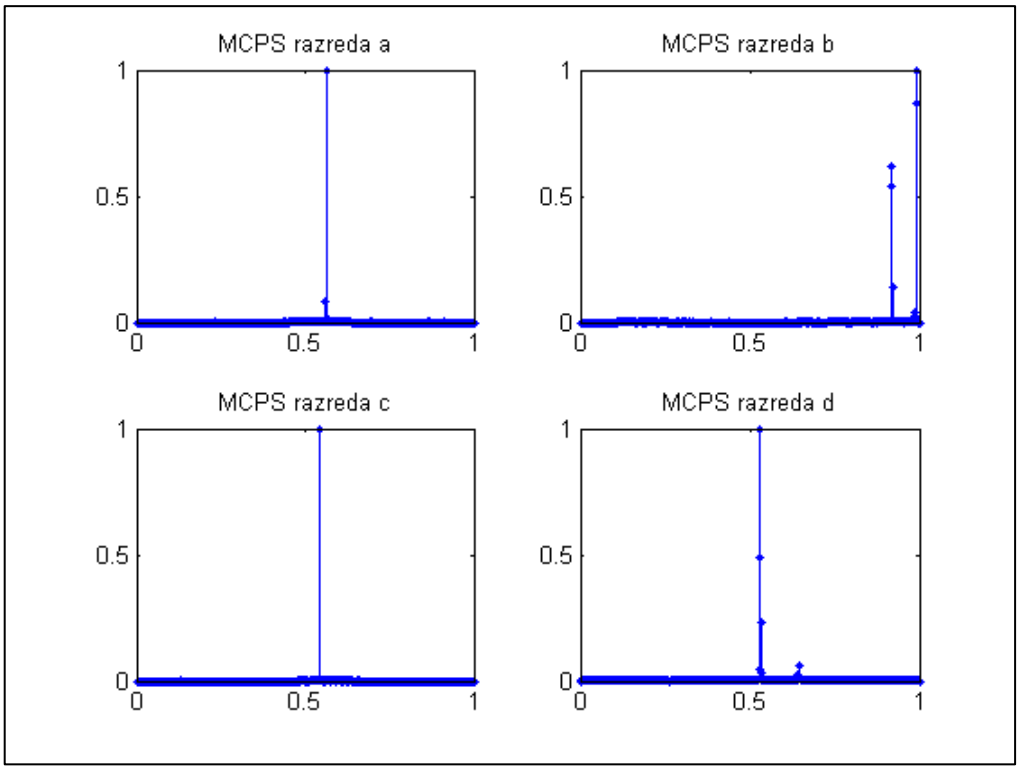
Tabela 3 – Zastupljenost sekvenci iz pojedinog strukturnog razreda u skupu FC699

	α-uzvojnice	β-ploče	ostale
a	63%	4%	33%
b	9%	43%	48%
c	43%	17%	40%
d	28%	28%	44%

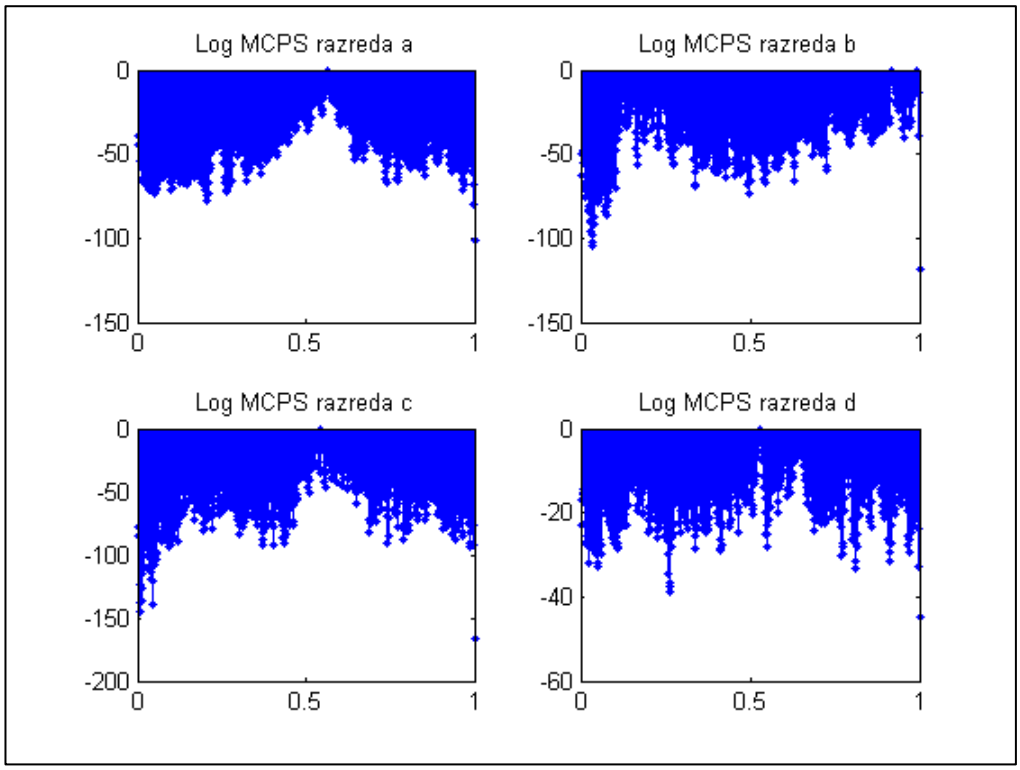
Tabela 4 – Broj sekvenci pojedinog strukturnog razreda u skupu FC699

Broj raspoloživih sekvenci iz pojedinih razreda u skupu FC699			
a	**b**	**c**	**d**
123	233	301	76

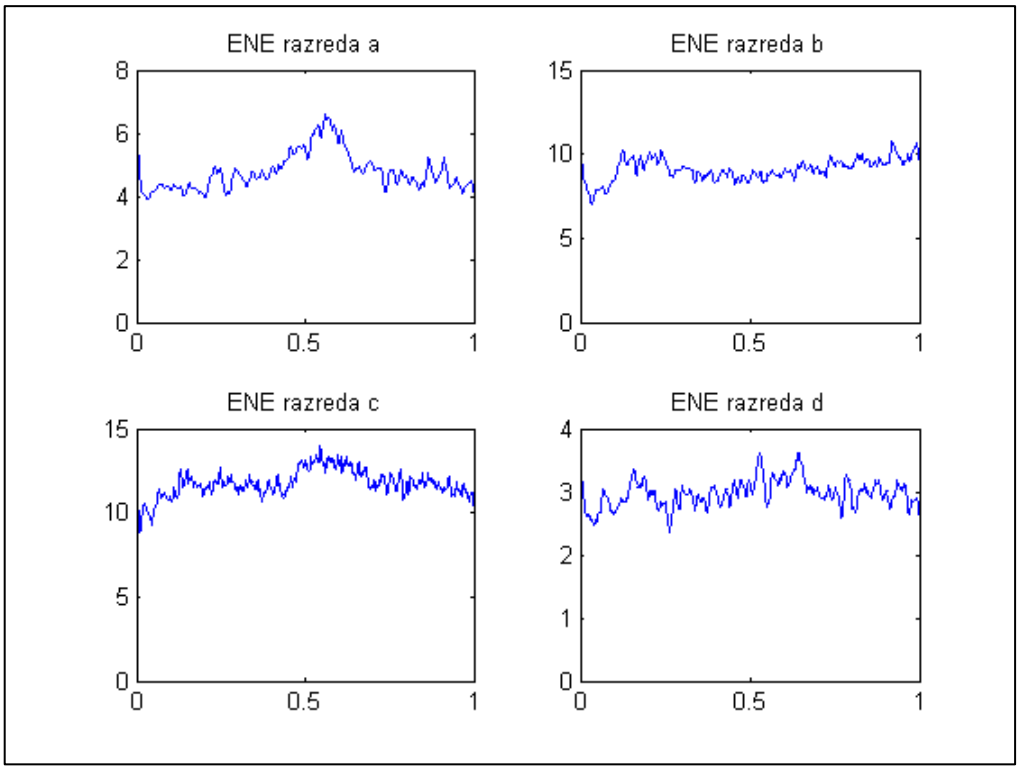
U nastavku su prikazane vizualizacije mjera skupa FC699, posebno za svaki strukturni razred kako bi se dobio osnovni uvid u frekvencijsku karakteristiku signala sekvenci iz pojedinih strukturnih razreda. Iz sljedećih slika vidi se da spektar skupa a ima izraženije frekvencije u središnjem pojasu frekvencija (oko $\pi/2$) , dok skup b ime izraženije frekvencije u visoko i nisko-frekvencijskom području. Lako je uočljivo i to da su mjere skupa c dosta slične mjerama skupa *a* što je i razumno budući da su slični po udjelu tipova sekundarnih struktura.



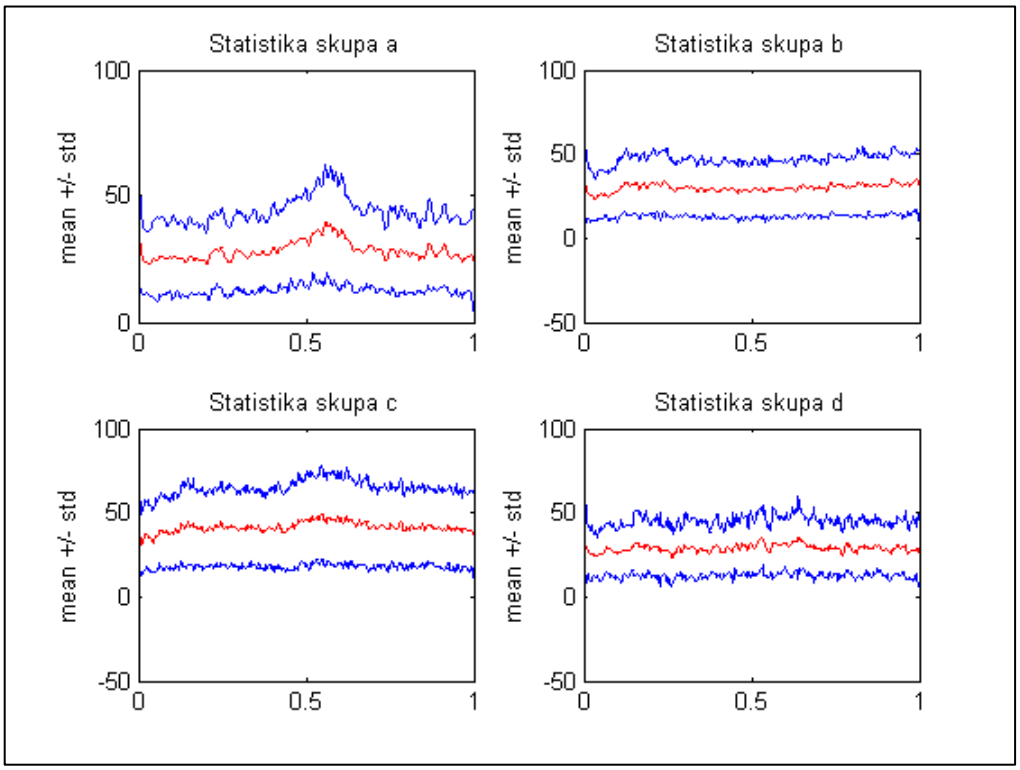
Slika 12 – MCPS mjera spektara signala sekvenci pojedinih strukturnih razreda



Slika 13 - log MCPS mjera spektara signala sekvenci pojedinih strukturnih razreda



Slika 13 - ENE mjera spektara signala sekvenci pojedinih strukturnih razreda



Slika 14 - statistika spektara signala sekvenci pojedinih strukturnih razreda

5.1. Klasifikacija sekvenci

Jedan od ciljeva ovog rada jest odrediti strukturni razred neke sekvence na temelju njezinih aminokiselina. Određivanje strukturnog razreda (klasifikacija) temelji se na amplitudnim spektrima signala sekvenci i njihovim mjerama. Ovaj problem je nešto jednostavnije prirode od određivanja egzaktne sekundarne strukture svake aminokiseline, no ipak, njegovo uspješno rješenje bi bilo od velike koristi jer bi se znalo koji tipovi sekundarnih struktura prevladavaju u pojedinoj sekvenci.

5.1.1. Implementacija klasifikatora temeljena na MCPS mjeri

Skup sekvenci iz FC699 zajedno sa pripadnim skupovima pripadnosti (*a, b, c, d*) čini ukupan skup uzoraka kojeg treba podijeliti na skup za učenje i skup za testiranje. Prikladan odabir je za učenje uzeti 80% svih uzoraka, odnosno za testiranje preostalih 20% uzoraka.

Za svaki od razreda (*a, b, c, d*) se na temelju 80% uzoraka koji pripadaju pojedinom razredu računa MCPS mjera. Rezultat su četiri vektora, tj. MCPS za svaki od razreda (*a, b, c, d*) koji predstavljaju vektore težina klasifikatora. Potom se svi od uzoraka za testiranje klasificiraju u jedan od razreda na način da se spektar svakog od signala za testiranje po komponentama skalarno pomnoži sa svakim od četiri MCPS-a te se tako dobiju četiri vrijednosti decizijskih funkcija. Pritom je potrebno amplitudni spektar signala kao i same MCPS-ove normirati po amplitudi kako apsolutne veličine ne bi utjecale na klasifikaciju.

Pripadni signal se razvrstava u onaj razred za čiji MCPS (dobiven na temelju 80% uzoraka iz tog razreda) skalarno pomnožen sa amplitudnim spektrom daje najveći iznos. Drugim riječima, sekvenca se svrstava u onaj razred čiji je iznos decizijske funkcije najveći.

5.1.1.1. Rezultati

Kako bi se dobili što vjerniji rezultati, klasifikator je potrebno pokrenuti više puta mijenjajući pritom uzorke za učenje i testiranje. Odnos je i dalje 80% ukupnog broja uzoraka za učenje i 20% za testiranje, no u svakom novom pokretanju klasifikatora pojedini uzorci koji su u prethodnoj iteraciji služili za testiranje u sljedećoj služe za učenje i obratno. Na taj način se izbjegava pristranost uzrokovana učenjem i testiranjem klasifikatora istim uzorcima. Sljedeća tablica

prikazuje postotak pogreške klasifikatora nakon 100 iteracija, tj. prikazuje statistiku točnosti u 100 pokretanja klasifikatora (čime se dobije vektor od 100 postotaka točnosti klasifikatora):

Tabela 5 – statistika rezultata klasifikacije sekvenci u strukturne razrede na temelju 100 iteracija

	a	b	c	d	ukupno
Max	58%	93%	63%	100%	75%
Mean	28%	34%	20%	50%	29%
Median	29%	37%	20%	37%	29%
Min	8%	9%	13%	13%	11%

5.1.2. Random class

Budući da rezultati nisu na zadovoljavajućoj razini, provedeno je tzv. *random class* testiranje kako bi se utvrdilo postoji li informacija u odabranoj težinskoj funkciji koju koristi klasifikator. Svim spektrima za testiranje se dodjele ne stvarni, već nasumce odabrani razredi. Dakle, svim uzorcima smo nasumice pridijelili razred pripadnosti. U slučaju da su nad takvim podacima rezultati klasifikacije lošiji, informacija u odabranoj težinskoj funkciji postoji pa i model ima smisla. U protivnom nastupa slučaj gdje je klasifikacija jednaka pogađanju rezultata, tj. težinska funkcija ne sadrži korisnu informaciju o razredu.

Tabela 6 - statistika rezultata klasifikacije sekvenci u strukturne razrede na temelju 100 iteracija (slučajno pridijeljeni razredi)

	a	b	c	d	ukupno
Max	42%	67%	45%	80%	55%
Mean	18%	22%	15%	42%	20%
Median	17%	20%	15%	40%	20%
Min	0%	4%	0%	0%	1%

Dobiveni rezultati su lošiji što znači da spektri dobiveni na temelju indeksa hidrofobnosti aminokiselina ipak nose informaciju o sekundarnoj strukturi. Ipak, zbog veoma niske točnosti klasifikacije je nužno isprobati druge metode klasifikacije.

5.1.3. Leave-one-out

Leave-one-out predstavlja način validacije kod kojeg se testiranje klasifikatora nad nekim skupom podataka vrši točno onoliko kolika je veličina tog skupa. U svakom koraku se N-1 uzorak (N je veličina skupa) koristi za treniranje klasifikatora, a preostali uzorak se koristi kao testni uzorak. U svakoj iteraciji drugi uzorak postaje testni pa se klasifikacija vrši ukupno N puta. Ovaj princip je testiran na skupovima *a* i *b*, budući da su po svom sastavu dovoljno različiti pa ima nade da će ovo testiranje biti relativno uspješno. Rezultati su dobiveni na temelju 100 epoha testiranja (100 puta je pokrenut gore opisani klasifikator) te su prikazani u sljedećoj tablici:

Tabela 7 – rezultati leave-one-out klasifikacije sekvenci u strukturne razrede a i b

Postotak točno klasificiranih uzoraka (%)			
	razred a	razred b	ukupno
Max	96	98	97
Mean	64	62	63
Median	67	61	63
Min	21	24	23

Također, kao i u prethodnom primjeru, testirana je i klasifikacija uzoraka nakon što se uzorcima pridijele slučajno odabrani razredi.

Tabela 8 - rezultati leave-one-out klasifikacije sekvenci u strukturne razrede a i b (slučajno pridijeljeni razredi)

Postotak točno klasificiranih uzoraka (%)			
	razred a	razred b	ukupno
Max	87	89	88
Mean	47	53	51
Median	46	52	50
Min	4	20	14

Rezultati nisu na željenoj razini, no ipak, opet se vidi da su bolji u slučaju kad uzorke klasificiramo na temelju njihovih pravih razreda što znači da odabrana značajka sekvenci (spektar signala sekvenci) ipak nosi određenu informaciju o sekundarnoj strukturi.

5.1.4. Electron-Ion Interaction Potential (EIIP)

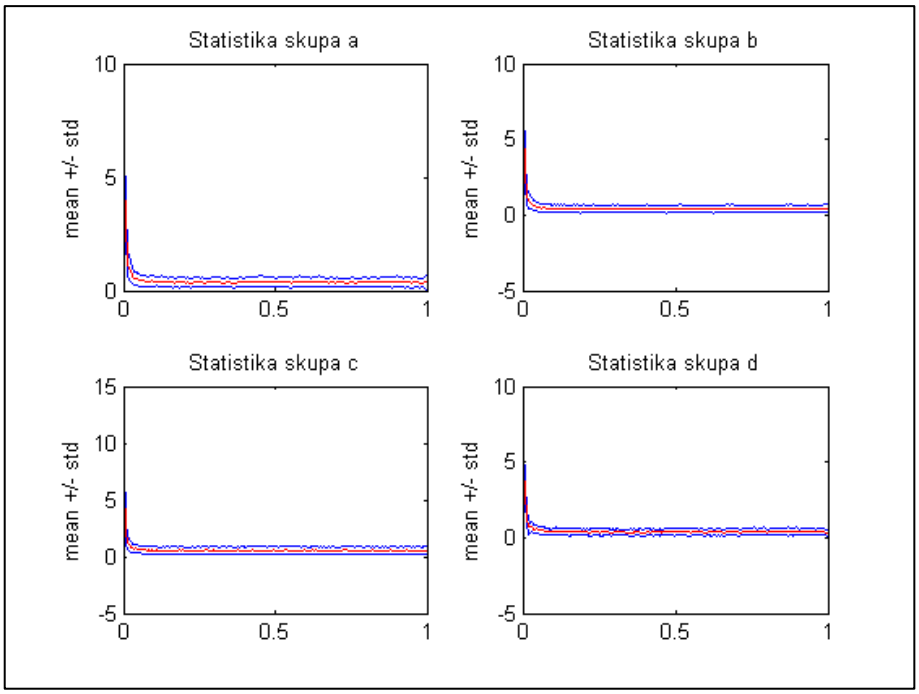
Elektro-ionski interakcijski potencijal je također svojstvo aminokiselina koje po svom nazivu upućuje da se radi o svojstvu, odnosno silama elektrostatske prirode kojima djeluje pojedina aminokiselina sekvenci na svoju okolinu. To pak upućuje da će to svojstvo u nekoj mjeri utjecati na prostorni oblik proteinskih lanaca, tj. na njihovu sekundarnu strukturu. Zbog ove pretpostavke su izračunate mjere nad spektrima signala sekvenci te su signali umjesto pomoću indeksa hidrofobnosti stvoreni pomoću vrijednosti elektro-ionskog interakcijskog potencijala koji je određen za svaku aminokiselinu kako je prikazano u tablici.

Tabela 9 – Vrijednosti EIIP za pojedinu aminokiselinu

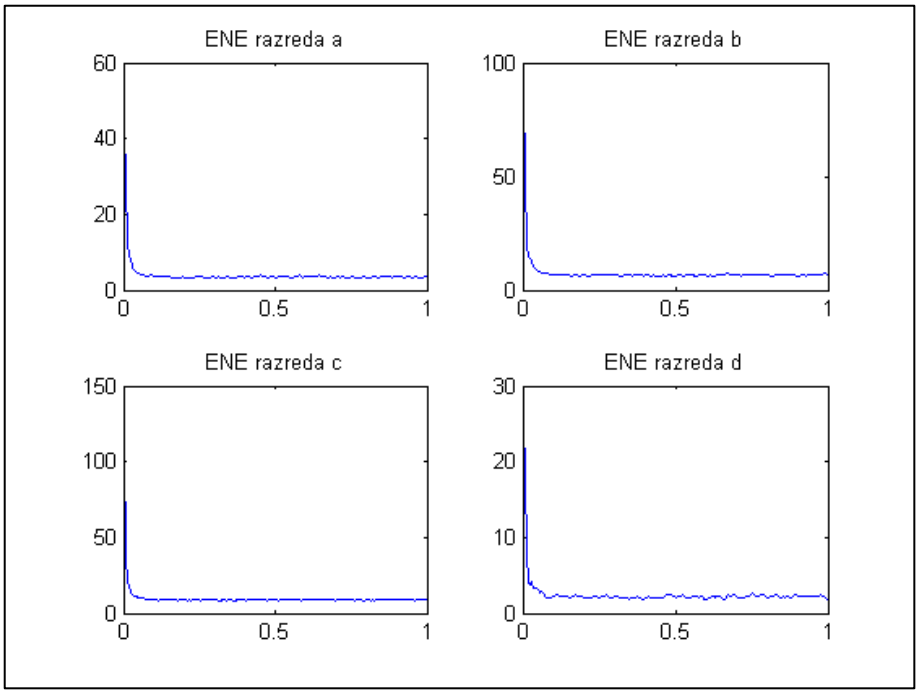
Leu(L)	Ile(I)	Asn(N)	Gly(G)	Val(V)	Glu(E)	Glu(E)	His(H)	Lys(K)	Ala(A)
0	0	0,0036	0,005	0,0057	0,0058	0,0198	0,0242	0,0371	0,0373

Tyr(Y)	Trp(W)	Gln(Q)	Met(M)	Ser(S)	Cys(C)	Thr(T)	Phe(F)	Arg(R)	Asp(D)
0,0561	0,0548	0,0761	0,0823	0,0829	0,0829	0,0941	0,0946	0,0959	0,1263

Na temelju signala nastalih pomoću EIIP vrijednosti prikazane su mjere na sljedećoj stranici. Obje mjere poprimaju gotovo jednake vrijednosti za sva četiri strukturalna razreda. To pak bi značilo da se FFT analizom ne može izlučiti korisna informacija koja bi razlikovala sekvence iz različitih strukturnih razreda koristeći informaciju o EIIP. Ipak, to ne znači nužno da ni jedna druga transformacija ne može polučiti bolje rezultate.



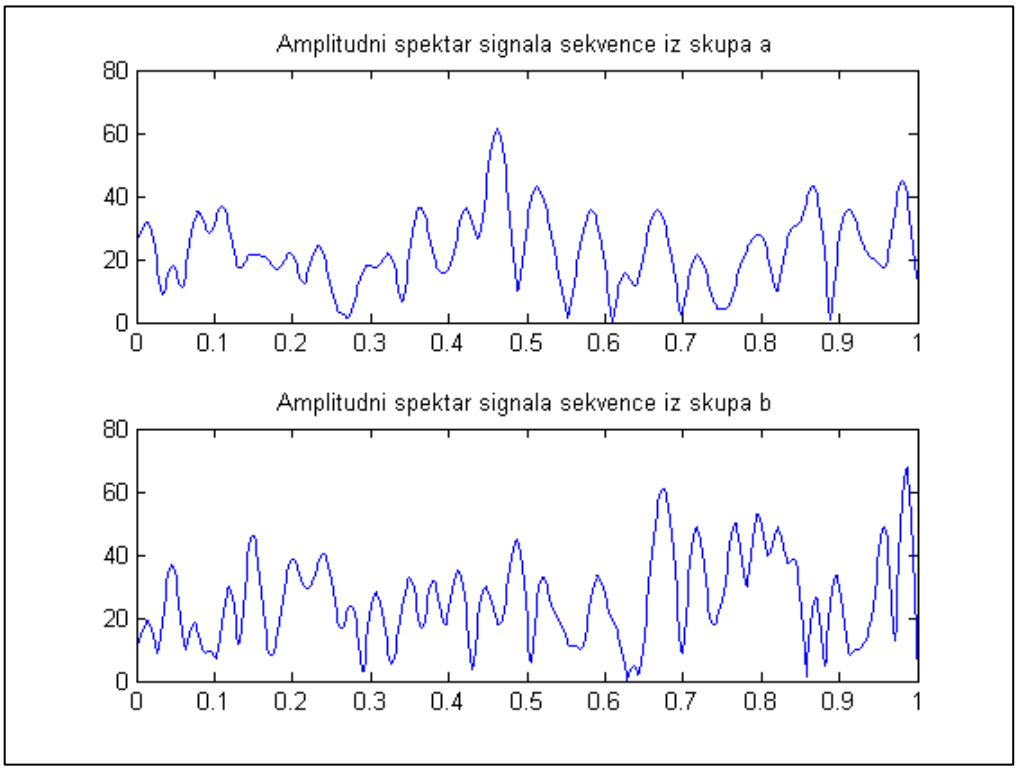
Slika 15 - statistika spektara EIIP signala sekvenci pojedinih strukturnih razreda



Slika 16 – ENE mjera spektara EIIP signala sekvenci pojedinih strukturnih razreda

5.1.5. Klasifikacija na temelju isključivo amplitudnog spektra

Do sada se klasifikacija vršila na temelju MCPS mjere koja je više-manje poprimala specifičan oblik za neki od četiri strukturna razreda. Ipak, kao što je vidljivo iz rezultata, MCPS kao i ostale mjere nisu uspjele uspješno opisati strukturne razrede. Jedan od ključnih razloga je taj što sve mjere uzimaju u obzir sve komponente spektra pojedinog signala sekvence. A ako uzmemo u obzir da u pojedinim skupovima postoje *outlieri* stvar postaje još kompliciranija. No, pogledajmo izgled amplitudnog spektra predstavnika strukturnih razreda *a* i *b*.

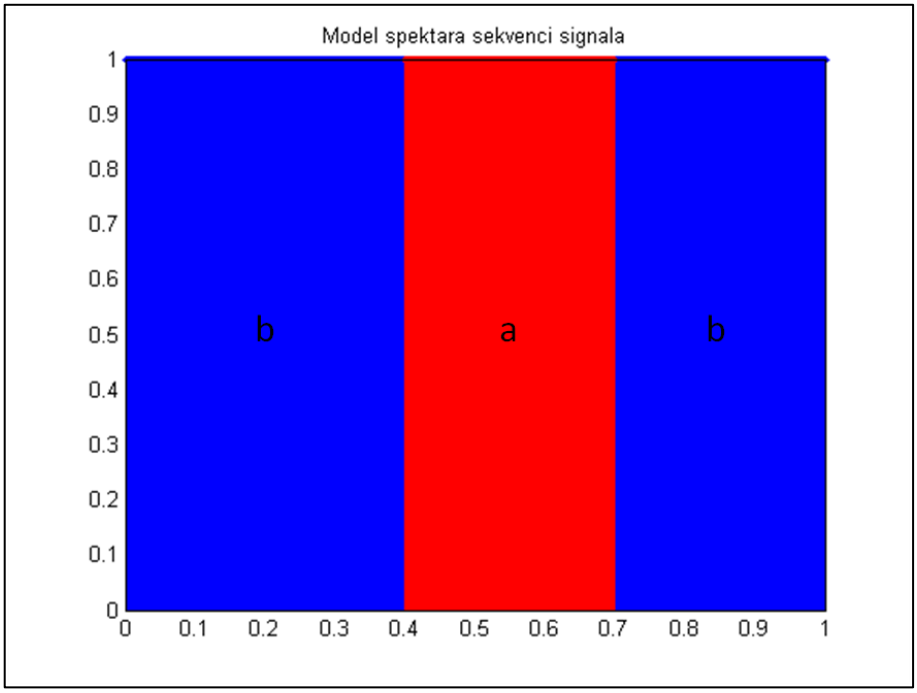


Slika 17 – primjeri amplitudnih spektara iz strukturnih razreda a i b

Uočljivo je da amplitudni spektar signala sekvenci iz spektralnog razreda *a* ima nešto izraženije frekvencijske komponente u središtu spektra, dok amplitudni spektar signala sekvenci iz spektralnog razreda b ima izraženije nisko i visoko-frekvencijske komponente. Ideja koja se sama po sebi nameće jest da se pri klasifikaciji sekvenca koristi informacija o dominantnom frekvencijskom području amplitudnog spektra.

5.1.6. Pojasni klasifikator

Na temelju opisanih razmatranja vezanih uz frekvencijsko područje zastupljenosti spektralnih komponenti nije na odmet provjeriti koliko se navedena pretpostavka poklapa sa stvarnim uzorcima, tj. spektrima signala sekvenci. Pretpostavka se najlakše može prikazati sljedećom slikom:



Slika 18 – Model raspodjele energije spektara signala sekvenci strukturnih razreda a i b

Ideja je sekvenci s nepoznatim strukturnim razredom sumirati komponente amplitudnog spektra u područjima koje karakteriziraju pojedine strukturne razrede i vidjeti za koja područja je navedena suma veća. Npr, za slučaj opisan slikom 20 računaju se dvije sume. Jedna od suma (suma *b*) nastaje zbrojem svih komponenti amplitudnog spektra signala na frekvencijskom području od $[0, 0.4\pi]$ i $(0.7\pi, \pi]$. Druga suma (suma *a*) nastaje zbrojem svih komponenti na frekvencijskom području $(0.4\pi, 0.7\pi]$. Ovisno o tome koja je suma veća, testnoj sekvenci se pridaje odgovarajući strukturni razred. Naravno, navedene granice u frekvencijskom području su dane samo ilustrativno i klasifikator će na temelju minimalne pogreške modela odrediti koje granice najbolje opisuju dane podatke.

Pretpostavka koja je fiksna jest ta da spektri signala sekvenca strukturnog razreda b imaju dominantne nisko i visoko-frekvencijske komponente amplitudnog spektra, dok one iz razreda a imaju dominantne središnje komponente.

5.1.6.1. Primjena praga na amplitudni spektar

Iz vizualizacije amplitudnih spektara kao i iz vizualizacije mjera, vidljivo je da nema naročito istaknutih frekvencija. Zbog postojanja *outliera* se ne može s velikom sigurnošću tvrditi da postoje jednoznačna dominantna područja u spektri kao što je to bilo rečeno u prethodnom poglavlju. Spektralne komponente se nalaze po cijelom frekvencijskom području te su iako manjih intenziteta od najistaknutijih, često presudne kod klasifikacije sekvenci na temelju sume spektralnih komponenata. Takve komponente nisu poželjne pa je neke od njih, kako će se pokazati, najbolje odstraniti. Ovaj pristup se naziva *thresholding* i podrazumijeva postavljanje na nulu svih komponenata koje su po iznosu manje od neke zadane vrijednosti.

5.1.6.2 Rezultati

Klasifikator je pokrenut za sve moguće kombinacije parametara granica frekvencijskih područja, tj. gornje granice NF područja (ujedno i donja granica PP područja), donje granice VF područja (ujedno i gornja granica PP područja) te praga ispod čije vrijednosti se spektralne komponente postavljaju na nulu. Donja granica NF područja je fiksirana na vrijednost 0, a gornja granica VF na π.

Tabela 10 – parametri modela koji se najbolje poklapa s podacima (sekvencama)

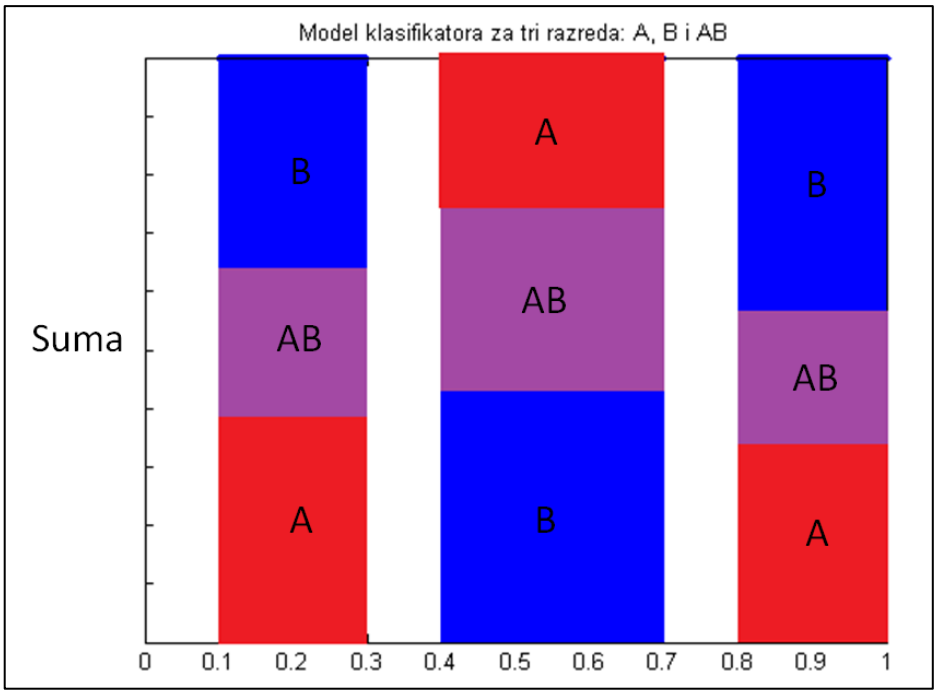
Frekv. područje razreda a	Frekv. područje razreda b	Vrijednost praga (threshold)
$(0.3\pi, 0.7\pi]$	$[0, 0.3\pi] \cup (0.7\pi, \pi]$	0.4
Točnost na skupu a	Točnost na skupu b	Ukupna točnost
66%	92%	83%

Rezultati su unatoč daleko od zadovoljavajućeg, mnogo bolji od onih dobivenih na temelju MCPS mjere. Postavlja se pitanje je li moguće pomoću ovakvog klasifikatora također klasificirati sekvence iz razreda *c* i *d*. Razredi *c* i *d* su podjednako građeni od α-uzvojnica i β-ploča, dakle po svojoj su građi mješavina razreda *a* i *b*. Zbog jednostavnosti, sekvence iz razreda *c* i *d* će se spojiti u jedan miješani razred *ab* te će se klasifikator prilagoditi tako da nepoznate sekvence klasificira u tri razreda: *a*, *b* i miješani razred *ab*.

5.1.6.3. Klasifikator razreda *a, b, ab*

Kod klasifikacije spektralnih razreda *a* i *b* eksperimentalno su na temelju minimalne pogreške klasifikacije određena disjunktna frekvencijska područja koja ih opisuju. Iz rezultata je vidljivo da je u 92% slučajeva suma spektralnih komponenti signala sekvenci iz razreda *b* veća u području $[0,0.3\pi]\cup(0.7\pi,\pi]$ nego u području $(0.3\pi,0.7\pi]$. Također, u 66% slučajeva je suma spektralnih komponenti signala sekvenci razreda *a* veća u području $(0.3\pi,0.7\pi]$ nego u području $[0,0.3\pi]\cup(0.7\pi,\pi]$. Uzme li se u obzir činjenica da je skup *ab* po sastavu tipova sekundarnih struktura zapravo mješavina skupa *a* i *b*, teško da bi novi klasifikator mogli oblikovati na način da se u model doda novo frekvencijsko područje u kojem bi pretpostavili da je suma komponenti iz tog frekvencijskog područja veća nego na ostalima, kao što je to bio slučaj kod klasifikacije dva razreda, *a* i *b*. Umjesto gledanja na kojim područjima je suma spektralnih komponenata veća, ideja je na određenim frekvencijskim područjima odrediti pragove sumacije i na temelju njih odrediti kojem strukturnom razredu pripada neka sekvenca.

Ideja je sljedeća: kao i na prethodnom primjeru klasifikacije, postoje tri frekvencijska područja u kojima se vrši sumacija spektralnih komponenti (NF, PP, VF). Pretpostavka modela jest ta da će omjer sume spektralnih komponenti signala sekvenci razreda *a* na pojasno-propusnom području te one na nisko i visoko-frekvencijskom području biti velik. Obrnuto, očekuje se da će isti omjer za sekvence razreda *b* biti malen te konačno, očekuje se da će biti blizak jedinici za sekvence razreda *ab*. Sljedeća slika ilustrira navedeni model:



Slika 19 – model raspodjele energije spektara signala sekvenci strukturnih razreda a, b i ab

Slika ilustrira sume spektralnih komponenata na svakom od tri frekvencijska područja. Na svakom od područja su prikazani očekivani odnosi suma za sekvence iz sva tri razreda. Za nisko i visoko-frekvencijsko područje se očekuje da će najveće sume postizati sekvence iz razreda B pa zatim sekvence iz razreda *ab* te naposljetku sekvence iz razreda *a*. Obrnuti slučaj nastupa na pojasno-propusnom području. Također, ovaj model ima „slobodnije“ granice nego prethodni za dva razreda gdje su sva područja bila spojena. Donja i gornja granica svakog frekvencijskog područja su sada slobodni parametri te se određuju na temelju minimalne pogreške klasifikatora. Naravno, frekvencijska područja moraju i dalje ostati disjunktna. Dodatni parametri su omjeri suma o_d i o_g na temelju kojih se sekvence klasificiraju prema sljedećem pravilu (sume pojaseva su označene kao imena pojaseva NF,PP,VF) :

Tabela 11 – kriterij klasifikacije sekvence u pojedini razred (a, b, ab)

$PP / (NF+VF) > o_g$	Sekvenca pripada razredu a
$o_d < PP / (NF+VF) < o_g$	Sekvenca pripada razredu ab
$PP / (NF+VF) < o_d$	Sekvenca pripada razredu a

Klasifikator se pokreće za sve moguće granice pojaseva (uz uvjet disjunktnosti pojaseva), donje i gornje granice omjera od i og te spektralnog praga. Pamte se parametri koji su rezultirali najboljom klasifikacijom sekvenci. Ovisno o postotku točnosti može se zaključiti koliko predloženi model odgovara podacima. Rezultati i iznos točnosti najbolje klasifikaciji su dani u sljedećim tablicama:

Tabela 12 - parametri modela koji se najbolje poklapa s podacima (sekvencama) za slučaj klasifikacije u tri strukturna razreda

Frekvencijsko područje	Donja granica	Gornja granica
NF	0	0.3π
PP	0.4π	0.7π
VF	0.9π	π
Donji omjer (o_d)	Gornji omjer (o_g)	Vrijednost praga (threshold)
0.7	2.4	0.4

Tabela 13 – kvaliteta poklapanja modela (rezultat klasifikacije) s podacima (sekvencama)

Točnost na a	Točnost na b	Točnost na ab	Ukupna točnost
32%	61%	78%	65%

Iz točnosti klasifikacije se može vidjeti koliko predloženi model odgovara podacima. U ovom modelu sekvence iz razreda *a* su klasificirane sa poprilično niskom točnošću. Ovaj model ne pruža zadovoljavajući opis sekundarne strukture sekvenci pa se ne može koristiti za određivanje strukturnog razreda sekvenci.

5.1.7. Uporaba SVM klasifikatora za određivanje strukturnog razreda sekvenci

Izvorno, algoritam s potpornim vektorima (eng. Support Vector Machine) je linearni klasifikator, dakle u najjednostavnijem slučaju, tj. klasifikaciji dva razreda pronalazi se decizijska funkcija, tj. pravac koji odjeljuje dva razreda, dok se u slučaju više razreda traži hiperravnina. Za razliku od najobičnijeg perceptrona, SVM pronalazi optimalnu hiperravninu. Optimalna hiperravnina je ona ravnina za koju su udaljenosti najbližih uzoraka iz svakog razreda do hiperravnine maksimalne.

Za potrebe klasifikacije spektara signala sekvenci linearan klasifikator nije bio odgovarajuć jer se pokazalo da uzorci nisu linearno odvojivi. Stoga je korišten SVM klasifikator sa Gaussovom jezgrenom funkcijom oblika:

$$K(x_i, x_j) = \exp(-\frac{\left\| x_i - x_j \right\|^2}{2\sigma}, \quad \sigma > 0 \tag{10}$$

Preporučljivo je da se širina postavi na vrijednost sumjerljivu sa srednjom vrijednošću udaljenosti uzoraka u skupu za učenje. Također, kako bi se izbjegli neželjeni efekti utjecaja različitih skala komponenti spektra, i-ta komponenta svakog spektra u skupu spektara se podijeli sa najvećom i-tom komponentom koja postoji u skupu spektara. Ovo dakle nije normiranje spektra po najvećoj komponenti pojedinog spektra (horizontalno normiranje), već se radi o normiranju i-tih komponenti svih spektara (vertikalno normiranje).

S obzirom na broj raspoloživih sekvenci s poznatim strukturnim razredom odabran je iz svakog strukturnog razreda fiksan broj od sto sekvenci za treniranje klasifikatora, dok je ostatak korišten za testiranje. Klasifikator je testiran na temelju dva razreda, a rezultati su prikazani u tablicama:

Tabela 14 – rezultat klasifikacije sekvenci u strukturni razred a ili b

Klasifikacija razreda a i b		
Širina Gaussove funkcije	Točnost na a	Točnost na b
100	70%	82%

Tabela 15 - rezultat klasifikacije sekvenci u strukturni razred a ili ab

Klasifikacija razreda a i ab		
Širina Gaussove funkcije	Točnost na a	Točnost na ab
100	70%	38%

Tabela 16 - rezultat klasifikacije sekvenci u strukturni razred b ili ab

Klasifikacija razreda b i ab		
Širina Gaussove funkcije	Točnost na b	Točnost na ab
100	70%	34%

SVM klasifikator s Gaussovom jezgrenom funkcijom je do sad dao najbolje rezultate u slučaju klasifikacije dva razreda, *a* i *b*. Ipak, uvede li se u cijelu priču miješani skup *ab* rezultati postaju mnogo lošiji.

5.1.8. PCA (Principal component analysis)

Analiza glavnih komponenti (*Principal component analysis, PCA*), poznata kao i Karhunen-Loeveova transformacija (*KLT*), je postupak kojim se izvorni prostor značajki preslikava u prostor čije su baze ortogonalni vektori čiji smjerovi odgovaraju smjeru najvećeg raspršenja uzoraka, koje se računa empirijski, na osnovu danih uzoraka. Tim se postupkom dobiva prikaz prostora stanja u kojem su naglašene različitosti uzoraka, čime se pojednostavljuje postupak raspoznavanja uzoraka. Analiza omogućuje i učinkovito smanjenje dimenzionalnosti prostora uzoraka.

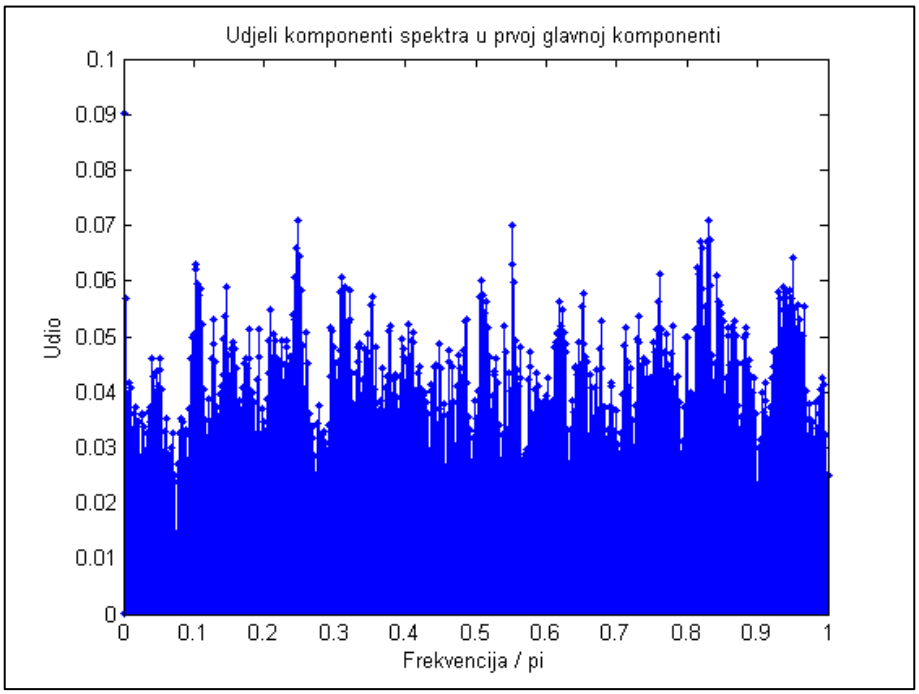
Uzmemo li naše skupove *a* i *b* te ih pomoću PCA transformiramo u nove uzorke od kojih zadržimo samo pet prvih (najbitnijih) komponenti dobivaju se sljedeći rezultati:

Tabela 17 - rezultat klasifikacije sekvenci u strukturni razred a ili b na temelju prvih pet glavnih komponenti

Klasifikacija razreda a i b (korišteno prvih pet glavnih komponenti)		
Širina Gaussove funkcije	Točnost na a	Točnost na b
10	74%	76%

Impresivna je činjenica da se uzimajući samo pet značajki dobivaju skoro pa jednaki rezultati klasifikacije. Prostor dimenzionalnosti uzoraka se pritom smanjio čak 102.4 puta!

Zanimljivo je također vidjeti od kojih se sve komponenata originalnog spektra (veličine 512 komponenti) sastoji prva i najznačajnija komponenta novih uzoraka.



Slika 20 – udjeli komponenta originalnog spektra u prvoj glavnoj komponenti

Iz slike se teško može uočiti neka ključna frekvencija koja bi bila sastavni dio prve glavne komponente, već je ona (prva glavna komponenta) sa relativno malim udjelom sastavljena od svih komponenti amplitudnog spektra.

Unatoč relativno dobroj klasifikaciji razreda *a* i *b*, rezultati klasifikacije razreda *a* i *ab* nisu tako uspješni. Razlog je opet vjerojatno, prevelika sličnost razreda *ab* sa razredom *a*, odnosno sa razredom *b*.

Tabela 18 - rezultat klasifikacije sekvenci u strukturni razred a i ab na temelju prvih pet glavnih komponenti

Klasifikacija razreda a i ab (korišteno prvih pet glavnih komponenti)		
Širina Gaussove funkcije	Točnost na a	Točnost na ab
1	83%	43%

Pokušaju li se klasificirati sva tri razreda dobivaju se sljedeći rezultati:

Tabela 19 - rezultat klasifikacije sekvenci u strukturni razred a, b i ab na temelju prvih pet glavnih komponenti

Klasifikacija razreda a, b i ab (korišteno prvih pet glavnih komponenti)			
Širina Gaussove funkcije	Točnost na a	Točnost na b	Točnost na ab
10	40%	68%	76%

Tabela 20 – matrica zabune klasifikacije sekvenci u strukturne razrede a, b i ab

Stvarno stanje\ Klasificirano stanje	a	b	ab
a	49	28	46
b	14	159	60
ab	28	64	285

Očekivano, postoji dosta sekvenci iz razreda *a* i *b* kojima je pridijeljen razred *ab*. Unatoč tomu, rezultati klasifikacije su relativno dobri s obzirom na to da je korišteno samo prvih pet glavnih spektralnih komponenti.

5.1.9. Klasifikacija sekvenci skupa 25PDB

Kako bi ispitali utjecaj glavnih spektralnih komponenti na klasifikaciju sekvenci u spektralne razrede korišten je skup značajki koje je predložio Kurgan u radu [1] te su im dodane glavne spektralne komponente.

Rezultati bez dodanih spektralnih komponenti:

Tabela 21 - rezultat klasifikacije sekvenci u strukturni razred a, b , c i d bez dodanih prvih pet glavnih spektralnih komponenti

Klasifikacija razreda a, b, c i d (bez prvih pet glavnih spektralnih komponenti)					
Širina Gaussove fje	Točnost na a	Točnost na b	Točnost na c	Točnost na d	Ukupno
0.6	92%	80%	71%	71%	79%

Tabela 22 - matrica zabune (klasifikacija bez dodanih prvih pet glavnih spektralnih komponenti)

Stvarno stanje\ Klasificirano stanje	a	b	c	d
a	407	2	11	23
b	8	355	13	67
c	16	3	246	81
d	30	45	54	312

Rezultati s dodanih prvih pet spektralnih komponenti:

Tabela 23 - rezultat klasifikacije sekvenci u strukturni razred a, b , c i d s dodanih prvih pet glavnih spektralnih komponenti

Klasifikacija razreda a, b, c i d (dodano prvih pet glavnih spektralnih komponenti)					
Širina Gaussove fje	Točnost na a	Točnost na b	Točnost na c	Točnost na d	Ukupno
0.6	93%	80%	77%	73%	81%

Tabela 24 - matrica zabune (klasifikacija s dodanih prvih pet glavnih spektralnih komponenti)

Stvarno stanje\ Klasificirano stanje	a	b	c	d
a	410	2	5	26
b	7	352	6	78
c	14	4	266	62
d	32	43	42	324

Dodavajući prvih pet glavnih spektralnih komponenti rezultati postaju nešto bolji. Ukupna točnost klasifikacije povećala se za 2%. Najizraženija promjena točnosti vidi se na sekvencama iz strukturnog razreda *c*.

6. Razlaganje sekvenci na očišćene signale

Primarni cilj ovog rada je određivanje egzaktne sekundarne strukture, tj. nastoji se za svaku aminokiselinu proteinskog lanca odrediti kojem tipu sekundarne strukture pripada. Očito, radi se o mnogo složenijem problemu od određivanja strukturnog razreda gdje potrebno samo odrediti koji tipovi sekundarnih struktura prevladavaju, odnosno kako su posloženi.

6.1. Očišćeni signali

Sekvence koje su nam na raspolaganju sastoje se od različitih tipova sekundarnih struktura. Mjere i spektri signala dobivenih na temelju takvih sekvenci nam ne daju informaciju o pojedinim tipovima sekundarnih struktura. Kako bi dobili informaciju o pojedinom tipu sekundarne strukture potrebno je signal sekvence pročistiti. Npr. želi li se iz spektara izvući informacija koja (eventualno) karakterizira H tip sekundarne strukture potrebno je u signalu ostaviti samo one uzorke koji po sekundarnoj strukturi odgovaraju tipu H, a ostale postaviti na nulu. U sljedećem primjeru sekvenca se pročišćava za tip H sekundarne strukture. Crvenom bojom su označene aminokiseline sekvence čiji će se indeks hidrofobnosti uzeti u obzir prilikom stvaranja signala, dok su crnom bojom označene aminokiseline čiji će pripadni uzorci signala poprimiti vrijednost nula.

Sekvenca aminokiselina : LNDPLDSGRFSRKQLDKKYKHAGDFGISDTKKNRETLTKFRDAIEEHL

Sekundarna struktura : CCTTTTBTTBCHHHHHHHGGGGGGCCCCCCCCCHHHHHHHHHHHHHHH

Rezultirajući signal:

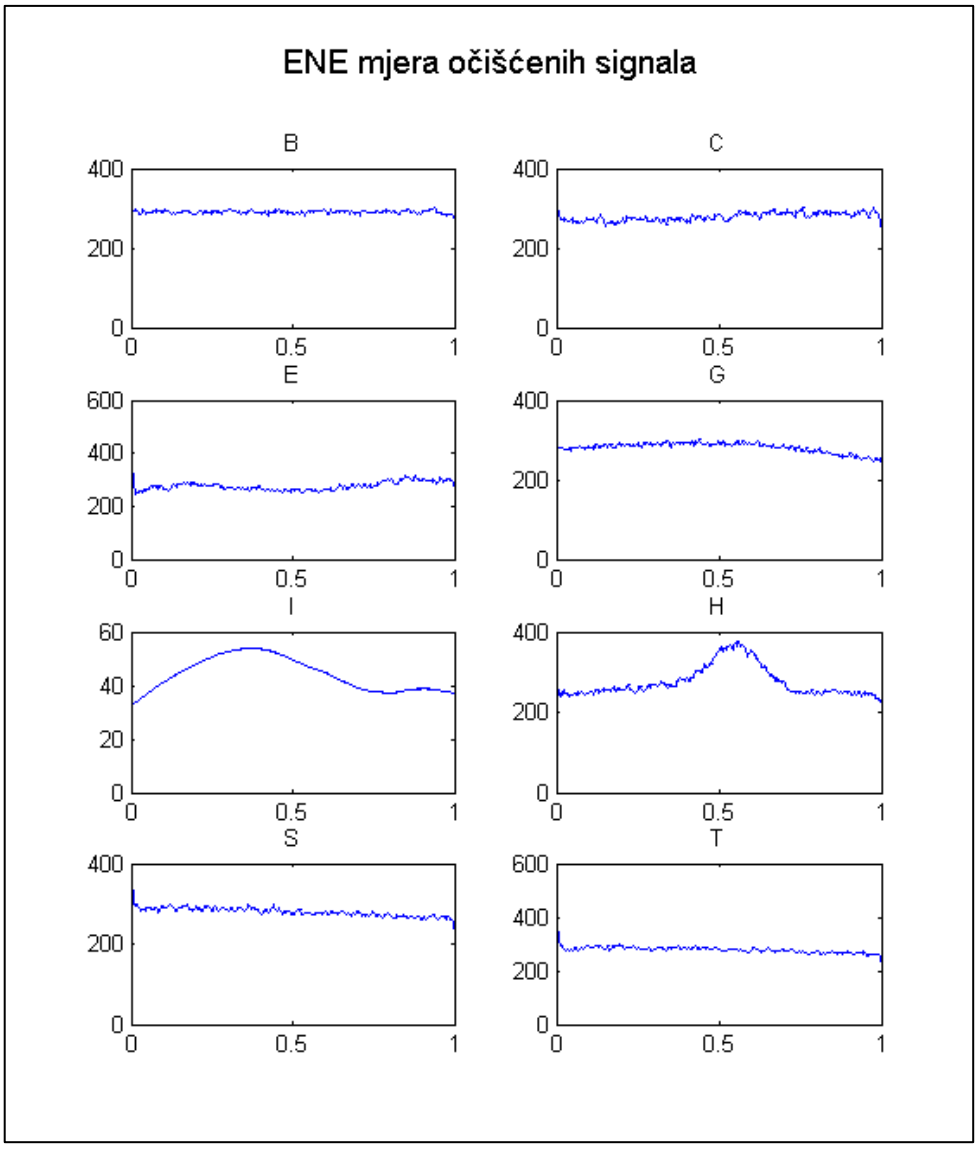
{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -4.5, -3.9, -3.5, 3.8, -3.5, -3.9, -3.9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -4.5, -3.5, -0.7, 3.8, -0.7, -3.9, 2.8, -4.5, -3.5, 1.8, 4.5, -3.5, -3.5, -3.2, 3.8}

Korištene su sekvence RCSB (*A Resource for Studying Biological Macromolecules*) baze. Na raspolaganju je 157.372 sekvenci zajedno sa primarnom i sekundarnom strukturom. Sljedeća tablica daje prikazuje zastupljenost pojedinih tipova sekundarnih struktura u cijeloj bazi:

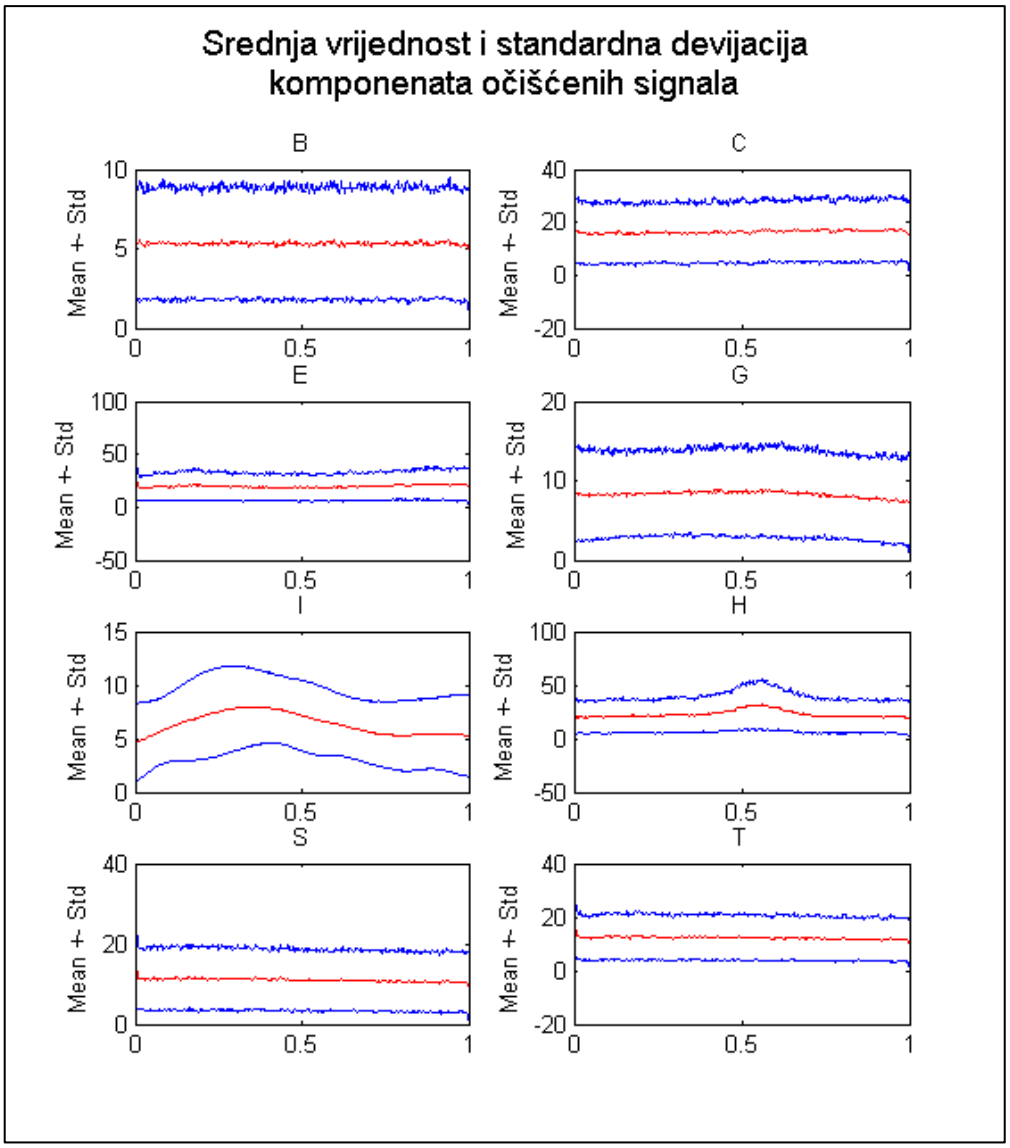
Tabela 25 – zastupljenost pojedinih tipova sekundarne strukture u RCSB bazi

Tip	C	B	E	G	I	H	S	T
Zastupljenost	21.93%	1.26%	20.84%	3.66%	0.02%	31.86%	9.11%	11.33%

Kako bi se uvidjele spektralne karakteristike pročišćenih signala prikazat će se ENE mjera te statistika komponenti. MCPS mjera nije pogodna zbog velikog skupa sekvenci pa je malo vjerojatno da će neka od frekvencija „opstati“ uzastopnim množenjem.



Slika 21 – ENE mjere skupova očišćenih signala za svaki tip sekundarne strukture



Slika 22 - statistika skupova očišćenih signala za svaki tip sekundarne strukture

Iz priloženih vizualizacija mjera vidi se da jedino spektri očišćenih signala za H i I tip ističu u nekim frekvencijskim područjima. Ipak, budući da je tip I sekundarne strukture zastupljen u bazi s veoma malim postotkom, njegove se mjere ne mogu uzeti kao reprezentativne. Budući da se za sve očišćene signale tipova sekundarnih struktura dobivaju mjere sa podjednakom zastupljenošću spektralnih komponenti na svim frekvencijskim područjima, ne može se očekivati određivanje sekundarne strukture sa velikom preciznošću.

6.2. Short-time Fourier transform

Do sad su se u radu proučavale frekvencijske karakteristike generiranih signala sekvenci ne bi li se uspjela uočiti neka frekvencija koja bi bila karakteristična za pojedini tip sekundarne strukture ili strukturnog razreda. Ipak, na taj način nije moguće znati gdje se navedena frekvencija nalazi u originalnom signalu, odnosno gdje u se u sekvenci nalazi pojedini tip sekundarne strukture. Npr. iz prijašnjih vizualizacija mjera očekuje se da će amplitudni spektar α-uzvojnica istaknut oko frekvencije $\pi/2$. Ako područje navedene frekvencije uzmemo kao karakterističnо za α-uzvojnice, postavlja se pitanje kako navedene frekvencije lokalizirati u polaznoj sekvenci. Osnovni pristup je STFT (Short-time Fourier transform).

STFT – ideja

STFT je zapravo računanje FFT-a na vremenskom otvoru. Umjesto da se računa FFT nad cijelim signalom, računa se na manjim segmentima definiranim vremenskim otvorom. Transformacija je oblika:

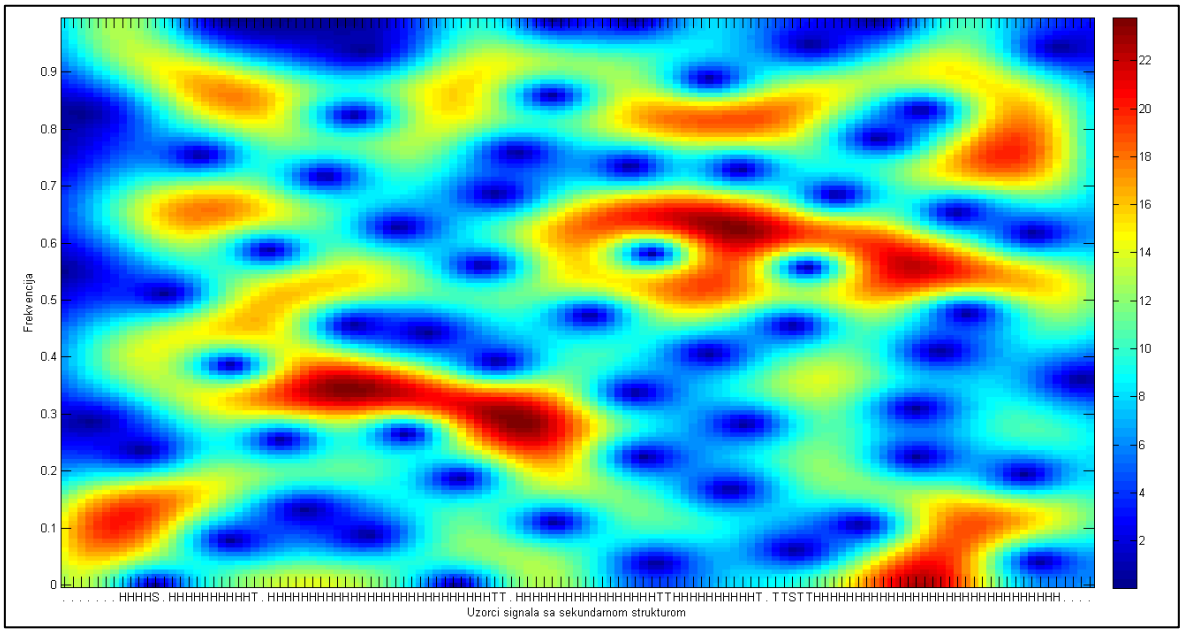
$$X(\tau,\omega)=\int_{-\infty}^{\infty} x(t)\,g^{*}(t-\tau)\,e^{-j\omega t}dt \qquad (11)$$

gdje je $g(t)$ lokalni analizirajući otvor željenih svojstava u obje domene. Poznato je da bolja određenost u jednoj domeni povlači lošiju određenost u drugoj domeni. Stoga je bitno poznavati prirodu signala kako bi se odabrali odgovarajući otvori. Kao što se dalo vidjeti, ne postoje točno određene frekvencije koje bi karakterizirale neki tip sekundarne strukture, već se radi o područjima karakterističnih frekvencija. S druge strane, za svaku aminokiselinu je potrebno odrediti kojem tipu sekundarne strukture pripada. Dakle, potrebna je bolja razlučivost u vremenskoj domeni. To znači da će se nastojati odabrati uži vremenski otvor kako bi se jasnije odredila željena područja u vremenskoj domeni. S druge strane, uži otvor će obuhvatiti šire frekvencijsko područje. Gaussov otvor je veoma čest izbor jer daje najmanji produkt efektivnih širina u vremenskom i frekvencijskom području, dakle, ukupna kvaliteta lokalizacije je optimalna za Gaussov otvor.

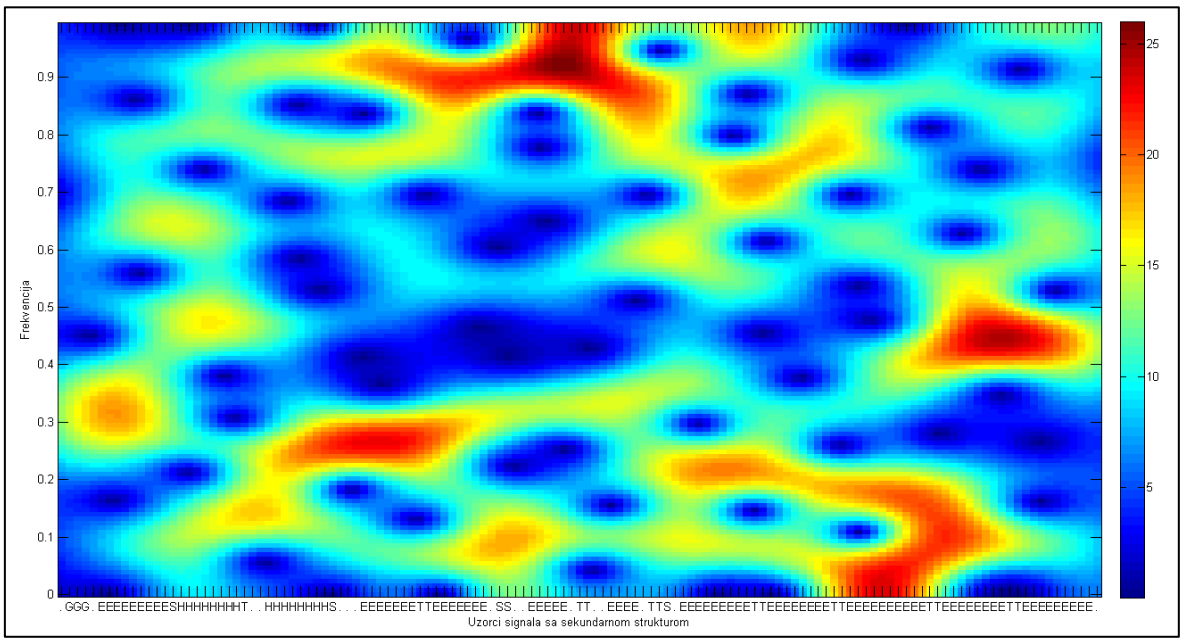
U programskoj implementaciji korištena je diskretna inačica STFT-a:

$$X\left(kT,\omega\right)\approx T\sum_{n=0}^{N-1}x\left(nT\right)g^{*}\left(nT-kT\right)e^{-j\omega nT} \tag{12}$$

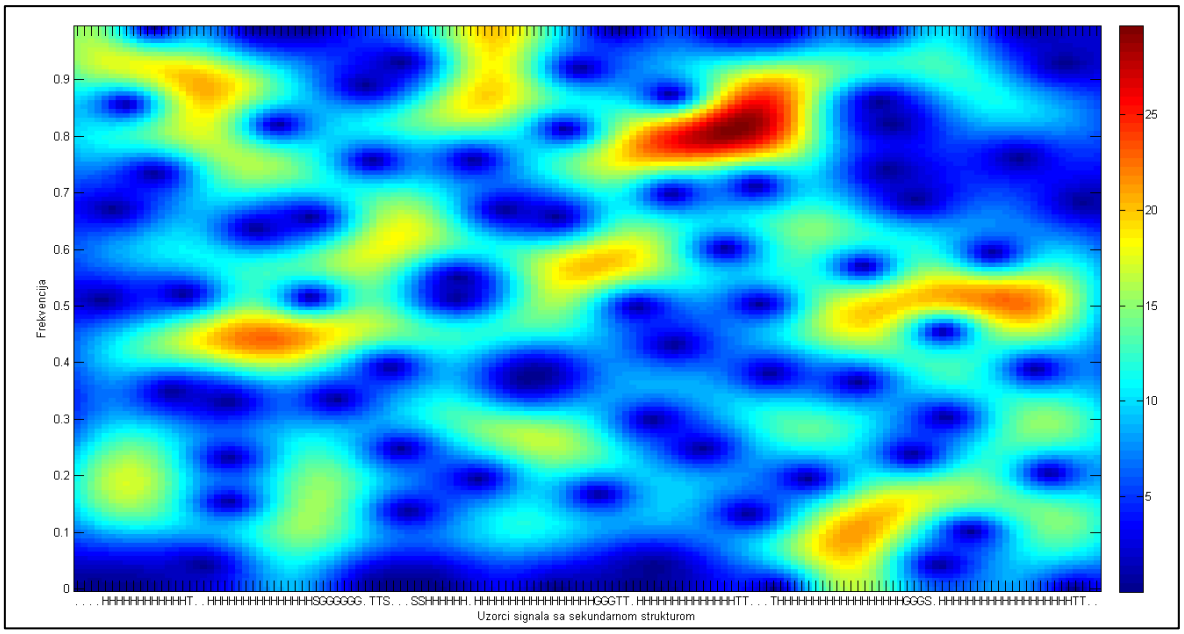
kao i Gaussov otvor. Transformacija je isprobana na nekoliko sekvenci i dobiveni su sljedeći rezultati:



Slika 23 – STFT signala sa dominantnim α-uzvojnicama



Slika 24 - STFT signala sa dominantnim β-pločama



Slika 25 - STFT signala sa dominantnim α-uzvojnicama

Iz navedenih rezultata se vidi da je određivanje egzaktne sekundarne strukture pomoću STFT-a gotovo nemoguće. Iz slika slijedi da isti tipovi sekundarnih struktura imaju više različitih frekvencijskih područja. S druge strane, vidi se da se frekvencijska područja α-uzvojnica i β-ploča na nekim mjestima i preklapaju. Neki od razloga tomu su sigurno: sekvence na kojima se računa FFT su relativno malih duljina pa je i sama informacija dobivena na temelju spektra dosta neprecizna. Drugo, vidljivo je da tipovi sekundarnih struktura u sekvencama nisu uvijek striktno složeni u veće grupe, već se pojavljuju manje grupe pojedinih tipova sekundarnih struktura koje je stoga veoma teško detektirati. STFT kao i FFT pristup nije pogodan za određivanje egzaktne sekundarne strukture već je pogodniji za okvirnu informaciju od strukturnom razredu.

7. Rezultati i diskusija

Analiza amplitudnih spektara signala sekvenci pokazala je da postoji generalni trend sekvenci koje odgovaraju strukturnim razredu *a*, odnosno *b*. Također postoji veoma sličan trend očišćenih sekvenci za tip *H* i tip *E*. Dominantno frekvencijsko područje spektara signala sekvenci strukturnog razreda *a* i spektara očišćenih signala tipa *H* jest pojasno propusno, dok je za strukturni razred b i spektre očišćenih signala tipa E karakteristično nisko i visoko-frekvencijsko područje.

Za sve sekvence iz pojedinih strukturnih razreda računate su spektralne mjere od kojih se MCPS koristila pri korelacijskoj klasifikaciji sekvenca u strukturne razrede. Korelacijski klasifikator pokazao se najslabijim, pojasni klasifikator boljim, a SVM najboljim. Ipak, zadovoljavajuće točne klasifikacije dobivaju se samo kod klasifikacije strukturnih razreda *a* i *b*. Miješani skupovi *c* i *d* su po svom sastavu tipova sekundarne strukture slični razredima *a* i *b* te samim time nisu dovoljno spektralno različiti da bi se mogli ispravno klasificirati.

Nad spektrima je provedena analiza glavnih komponenti te se odabirom prvih pet glavnih komponenti dobio čak i bolji rezultat nego u slučaju kad se koriste sve komponente amplitudnog spektra. Također, pomoću algoritma i značajki uzoraka korištenih u [1] točnost klasifikacije sekvenci u strukturne razrede porasla je za 2% dodavanjem prvih pet glavnih komponenti.

Određivanje egzaktne sekundarne strukture na temelju spektralne analize nije polučilo dobre rezultate. Razlog tomu leži u činjenici da spektri signala sekvenci pojedinih tipova sekundarne strukture nisu dovoljno različiti, što više, velik dio dominantnog frekvencijskog područja im se preklapa.

Kod stvaranja signala iz sekvenci korištena je Kyte-Doolittle skala hidrofobnosti. Također postoji i Hopp-Woods skala koja daje gotovo jednake rezultate kao i Kyte-Doolittle pa nije detaljnije obrađivana i korištena u radu.

8. Zaključak

Budući da polarnost bočnog lanca aminokiselina utječe na oblik cijelog proteinskog lanca, samim time utječe i na sekundarnu strukturu proteina. Stoga je indeks hidrofobnosti aminokiselina smislena značajka koja je korištena u ovom radu. Na temelju vrijednosti indeksa hidrofobnosti aminokiselina sekvence aminokiselina transformirane su u realne diskretne signale nad kojima je potom provedena frekvencijska analiza.

Korištene spektralne mjere: MCPS, ENE kao i statistika spektralnih komponenti pružaju dobar uvid u komponente spektara koje su istaknute na cijelom skupu sekvenci. Ipak, pokazalo se da je mali skup komponenti amplitudnog spektra specifičan za pojedine strukturne razrede, odnosno očišćene signale za pojedine tipove sekundarne strukture.

Spektri signala sekvenci strukturnih razreda a i b, odnosno očišćenih signala tipa H i E pokazali su generalni trend dominantnih spektralnih komponenata u disjunktnim frekvencijskim područjima pa je njih najlakše bilo klasificirati. S druge strane, u svim skupovima postoji poprilično broj *outliera* što dodatno otežava klasifikaciju.

Spektar signala sekvenci nosi određenu informaciju o sekundarnoj strukturi, no sam po sebi nije dovoljna značajka za određivanje sekundarne strukture na temelju primarne. Ipak, zbog malog, ali uočljivog (2%) poboljšanja klasifikacije sekvenci nakon što se značajkama korištenim u [1] doda prvih pet glavnih komponenti dobivenih na temelju amplitudnog spektra, zaključuje se kako spektar sadrži barem dio informacije o sekundarnoj strukturi.

9. Literatura

[1] LUKASZ KURGAN, KRZYSTOF CIOS, KE CHEN: „Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences“, University of Alberta, 2008.

[2] HANG CHEN, FEI GU, ZHENGGE HUANG: „Improved Chou-Fasman method for protein secondary structure prediction“, Zhejiang University, Hangzhou, China, 2006.

[3] CHAFIA HEJASE DE TRAD, QIANG FANG, IRENA COSIC: „Protein sequence comparision based on the wavelet transform approach“, BioElectronics Group, Department of Electrical and Computer Systems Engineering, Monash University, Australia, 2002.

[4] SAŠA JANJIĆ: „ Predviđanje sekundarne strukture proteina“, Seminarski rad, Fakultet elektrotehnike i računarstva, 2009.

[5] SAŠA JANJIĆ: „ Predviđanje mjesta sekundarne strukture proteina iz slijeda aminokiselinskih ostataka“, Diplomski rad, Fakultet elektrotehnike i računarstva, 2009.

[6] P.P VAIDYANATHAN, BYUNG-JUN YOON: „Gene and exon prediction using allpass-based filters“, Dept. of Electrical Engineering, California Institute of Technology, Pasadena, USA

10. Naslov, sažetak i ključne riječi

Analiza sekundarne strukture proteina metodom obrade signala

Kako bi mogli odrediti ulogu i funkciju proteina potrebno je poznavati njegovu sekundarnu strukturu. Budući da su eksperimentalne metode određivanja sekundarne strukture veoma skupe i zahtjevne, sve više se teži unapređenju računalnih metoda za određivanje sekundarne strukture. Jedan način jest poznavajući primarnu strukturu iz nje izlučiti određene značajke i pomoću njih odrediti sekundarnu strukturu. Jedno od bitnih svojstava koje utječe na trodimenzionalnu, a time i na sekundarnu strukturu proteina jest polarnost bočnog lanca aminokiselina koja definira indeks hidrofobnosti aminokiselina. Pomoću tih vrijednosti sekvence aminokiselina se transformiraju u diskretne realne signale nad kojima je moguće vršiti razne transformacije ne bi li se došlo do korisne informacije o sekundarnoj strukturi.

Ključne riječi: protein, proteinski lanac, sekvenca aminokiselina, polarnost bočnog lanca aminokiselina, indeks hidrofobnosti aminokiselina, primarna i sekundarna struktura, brza Fourierova transformacija, amplitudni spektar, spektralne mjere, tipovi sekundarne strukture, analiza glavnih komponenti.

The title, summary and key words

Secondary protein structure analysis by the method of signal processing

In order to determine the role and function of a protein it's necessary to know its secondary structure. Because experimental methods of secondary structure determination are extremely expensive and demanding, the tendency is to further develop computing methods for secondary structure determination. Knowing the primary structure, certain characteristics can be extracted and used to determine the secondary structure. One of the important properties influencing the three-dimensional (and therefore also the secondary) protein structure is the sideward amino-acid chain polarity which defines the amino-acid hydrophobicity index. Using those values, amino-acid sequences can be transformed into discrete real signals. We can perform different transforms to these signals in order to get some useful information about the secondary structure.

Key words: protein, protein chain, amino-acid sequence, sideward amino-acid chain polarity, amino-acid hydrophobicity index, the primary and the secondary structure, Fast Fourier Transform, amplitude spectrum, spectral measurements, secondary structure types, principle component analysis.