SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Predikcija sekundarnog oblika proteina korištenjem *data mining* metode

Ivor Prebeg

Voditelj: *Krešimir Šikić*

Zagreb, travanj, 2007.

Sadržaj

1. Uvod

1.1. Zašto želimo znati sekundarni oblik proteina?

Proteinomika je grana bioinformatike, znanosti koja se služi računalnim metodama pri istraživanju proteina na razini molekularne strukture te prirode ponašanja proteina. Predikcija strukture proteina jedna je od najznačajnih metoda razvijenih unutar računalne biologije. Cilj joj je kompleksnim računalnim metodama što točnije odrediti trodimenzionalnu strukturu iz linearne sekvence osnovnih gradivnih jedinica proteina – alfa-aminokiselina. Poznavajući trodimenzionalnu strukturu moguće je klasificirati protein prema njegovoj funkciji i prirodi njegova ponašanja. Također možemo otkriti vrlo udaljene odnose između proteina, što bi globalno povećalo točnost predikcije. Što je veća baza podataka o ponašanju proteina, to je moguće točnije klasificirati nove proteine.

1.2. Primjena

Ova je metoda svojim rezultatima uvelike nadmašila i zamjenila dugotrajne i relativno skupe metode kao što su rendgenska kristalografija i NMR sprektroskopija. Poznavanje prirode ponašanja proteina od iznimne je važnosti u dizajnu lijekova ili otkrivanju uzroka bolesti. Na primjer, želimo li pronaći lijek za neku bolest, moramo pronaći takvu molekulu koja se ponaša točno onako kako bi spriječila bolest. Uštedjeti ćemo mnogo resursa ukoliko testiramo samo one molekule za koje možemo sa određenom sigurnošću reći da se ponašaju barem približno onome što je potrebno za daljnja testiranja.

1.3. Povijest

Teorija o tome da je primarni oblik proteina linearna sekvenca alfa aminokiselina izložena je gotovo istovremeno na istoj konferenciji od dvojice neovisnih znastvenika. Bili su to Franz Hofmeister i Emil Fischer, 1902. godine u Karlsbadu u Njemačkoj. Pedesetih godina prošlog stoljeća Linus Pauling otkrio je oblik *alfa-helix* (pužnica) i *beta-strand* (ploča). Nekoliko godina prije njegovog otkrića bilo je pokušaja predikcije strukture iz linerane sekvence putem rendgenskog zračenja. Prve metode predikcije koje su uslijedile tijekom šezdesetih i sedamdesetih godina prošlog stoljeća bazirale su se na sklonostima jedne aminokiseline. Druga generacija metoda koja dominira do ranih devedestih godina prošlog stoljeća bila je bazirana na sklonostima od tri do pedeset i jedne susjedne aminokiseline unutar proteina. Gotovo svaki mogući algoritam primjenjen je na predikciju sekundarnog iz primarnog oblika proteina. Činilo se da je točnost predikcije stala na oko 60 %. Proboj metoda treće generacije podigao je točnost predikcije preko 70 % uz uvijet korištenja većih baza podataka i naprednijih algoritama. Glavna novost u tim algoritmima je bilo korištenje podataka o evoluciji. Svi prirodno nastali proteini sa više od 35 % sličnosti u najmanje sto poravnatih sekvenci imaju sličnu strukturu. Metoda PSI-BLAST i HMM (eng. *Hidden Markov models*) metoda još uvijek podižu granice predikcije sekundarnog oblika proteina.

2. Seminarski rad

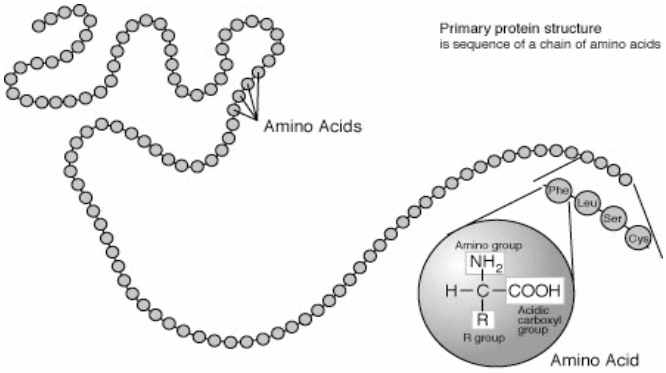
2.1. Protein

2.1.1. Uloga i važnost

Svaka ljudska stanica unutar svoje jezgre posjeduje čitavu molekulu DNA. Ona predstavlja jedinstveni genski zapis svakog organizma i jednoznačno ga određuje. Izgleda poput dvije trake savijene u obliku pužnice, međusobno spojene kemijskim vezama. To su vodikove veze koje vežu po dvije od četiri osnovne baze, adenin i timin te citozin i gvanin. Posredstvom molekule DNA nastaju proteini, manifestacija gena. Proteini su tvari koje imaju golemu ulogu u svima organizmima, pa je nužno da se generiraju ispravno i da rade točno ono što trebaju raditi. Jedna od funkcija proteina je da sudjeluju u stvarnju tkiva. Dakle, ako želimo da nam na prstima rastu nokti, stanice u području korijena nokta moraju generirati proteine koji će omogućiti stvaranje gradivnog materijala koji izgrađuju nokte. Kako DNA zna koje proteine treba generirati, tj. da nam umjesto noktiju na prstima ne raste, recimo, kosa, još uvijek je jedna od najvećih tajni moderne mikrobiologije. Enzimi, tvari koje ubrzavaju kemijske reakcije, također su proteini, kao i hemoglobini, molekule koje raznose kisik u ljudskom organizmu. Dakle, funkcije proteina od vitalne je važnosti za život bilo kojeg organizma.

2.1.2. Sinteza

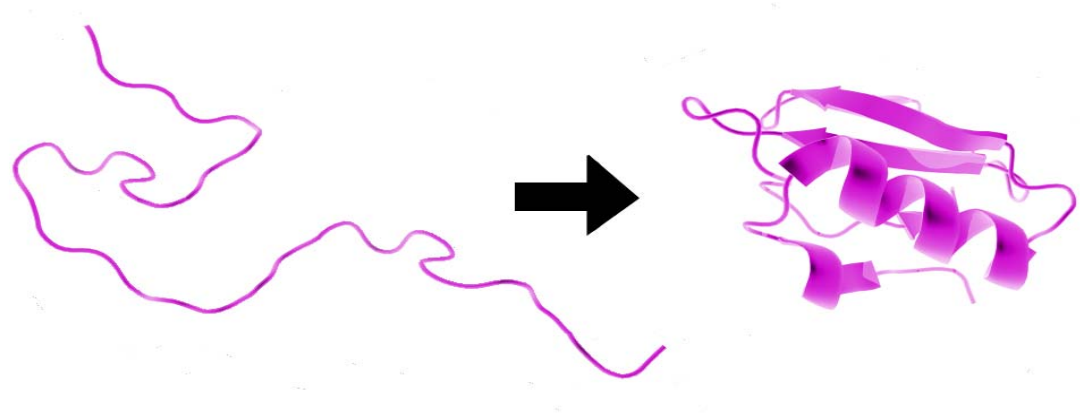
Sinteza proteina odvija se tako da se pod utjecajem enzima polimeraze molekula DNA se cijepa na dva dijela. Jedan dio se nadopuni slobodnim nukleotidima iz jezgre kako bi se DNA vratila u prijašnje stanje, dok drugi aktivno sudjeluje u sintezi i naziva se glasnička ili m-RNA (eng. *messenger*). M-RNA molekulu koristi ribosom kao uzorak za sintezu proteina. Triplet baza iz m-RNA naziva se kodon i čini osnovu za prevođenje u potpun protein. Prijenosna ili t-RNA (eng. *transfer*) dolazi iz jezgre i sastoji se od antikodona koji nosi jednu alfa-aminokiselinu. Na mjestu gdje kodon odgovara antikodonu, lijepi se t-RNA i gradi potpunu molekulu. Kako postoje četiri baze, moguće je šezdeset i četiri različita kodona. (4 x 4 x 4) Također, postoji dvadeset esencijalnih aminokiselina, pa tako je preslikavanje kodon -> esencijalna bjelančevina neinjektivno te čini genski kod uz zalihost odnosno zaštitu od pogrešaka. Genski je kod identičan za sav živi svijet, od bakterija do sisavaca. Kada se potpuno izgrađena molekula odvoji od ribosoma, naziva se protein u svom primarnom obliku.(Slika 2.1.)



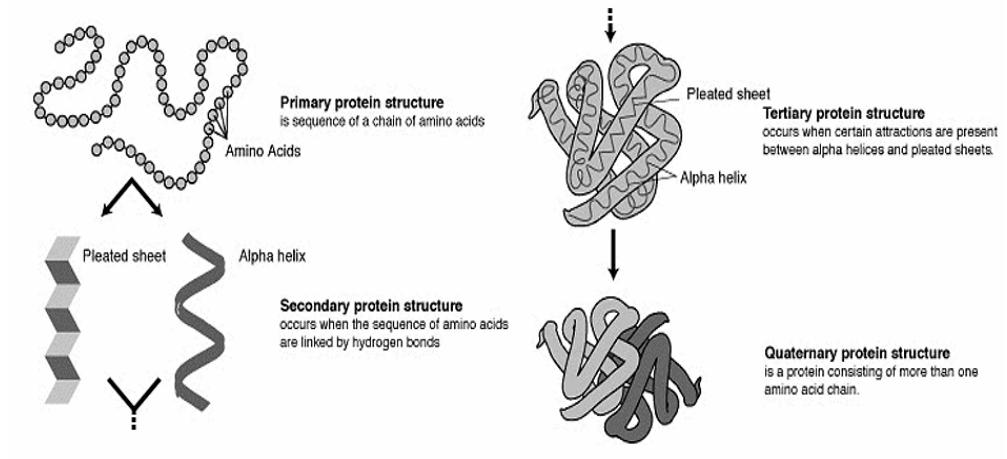
Slika 2.1. - Primarni oblik proteina – linearna sekvenca.

2.1.3. Presavijanje

Presavijanje proteina je fizikalni proces tijekom kojeg se molekule proteina presavijaju iz primarnog u svoj karakterističan trodimenzionalni oblik, zvan i prirodno stanje. Presavijanje traje od nekoliko milisekundi do nekoliko sekundi. Šezdesetih godina prošlog stoljeća pokusima je zaključeno da primarni oblik uz parametre okoline određuje sekundarni oblik. Za neke bolesti (Alzheimerova bolest, kravlje ludilo) vjeruje se da su uzrokovane pogreškama pri presavijanju proteina. Sekundarni oblik je trodimenzionalna forma koja ne opisuje specifične pozicije atoma u trodimenzionalnom prostoru. (Slika 2.2.) Tercijarna struktura nastaje nakon što se dogode određene interakcije između *alfa-helix*-a i *beta-strand*-ova dok je kvartarna strukura protein koji je interagirao sa drugim proteinima s kojima je stvorio novi jedinstveni protein.(Slika 2.3)



Slika 2.2. Prijelaz proteina iz linearne sekvence u sekundarni oblik



Slika 2.3. Prijelaz proteina iz linearne sekvence u kvartarni oblik

2.2. *Data mining*

Još od najranijih dana čovjek je pokušavao naći uzorke i pravilnosti u prirodi, poput godišnje seobe životinja koje je lovio i perioda u kojem raste voće koje je sakupljao. Iako danas živimo na sasvim drugačiji način, još uvijek pokušavamo oponašati već postojeće načine ponašanja koje vidimo u svojoj okolini. Obaviti takav posao ručno veoma je teško i dugotrajno te neekonomično u današnjem vremenu. *Data mining* je automatska ili poluautomatska metoda te otkrivanje uzoraka radi računalo. U proteomici, *data mining* se koristi za predviđanje sekundarne strukture proteina te za njihovu klasifikaciju. *Data mining* je metoda uvelike primjenjena ne samo u bioinformatici, nego i u drugim znanostima poput ekonomije, medicine, psihologije...

2.2.1. Strukturalni uzorci

Kada tražimo vezu između unutar gomile podataka, njihov smisao ili korisnu informaciju, ono što zapravo tražimo su strukturalni uzorci. Strukturalni uzroci su set pravila koji opisuju način na koji različite kombinacije vrijednosti atributa koji opisuju niz entiteta utječu na to kako klasificirati novi entitet.

2.2.2. Strojno učenje

Strojno učenje je promatranje već poznatih načina ponašanja u različitim situacijama da bi se što točnije moglo pretpostaviti kako se ponašati u proizvoljnoj situaciji. Formalno možemo definirati učenje: stvari koje uvježbavamo kako bi promijenili svoje ponašanje na bolje u budućnosti. U tom slučaju bi riječ „učenje" bilo pravilnije zamijeniti riječju „obučavanje" ili „vježbanje".

2.2.3. *Data mining*

Data mining je proces otkrivanja uzoraka u velikim bazama podataka, a uzorci moraju biti korisna informacija. Nakon što smo „obučili" računalo raznim pravilima ponašanja u raznim situacijama, dobivene infromacije koristimo kako bismo mogli pobliže odrediti koje se ponašanje najčešće primjenjuje u situaciji nasličnijoj onoj u kojoj se mi nalazimo. Ono što nas zaista zanima jesu tehnike za nalaženje i opisivanje strukturalnih uzoraka kao alat koji će nam pomoći objasniti podatke i napraviti određena predviđanja na temelju tih podataka.

2.2.4 **Primjer *data mining*-a**

- Automobili – Recimo da je interesna sfera istraživanja automobilska industrija. Kako danas postoji čitav set novih modela automobila, bilo bi dobro imati nekakav aparat po kojem bi odredili kojoj klasi pripada neki automobil. Drugim riječima, želimo imati pravila na osnovu kojih ćemo uspoređujući karakteristike nekog automobila reći kojoj klasi pripada. Način na koji ćemo automobile opisati predstavlja set atributa. Neka to budu: veličina kotača, međuosovinski razmak, snaga motora, zapremnina prtljažnika, težina, maksimalna brzina, ubrzanje do 100 km/h i klasa. Što veći set atributa odaberemo, to će *data mining* biti djelotvorniji.

Veličina kotača [']	Međuosovinski razmak [cm]	Snaga motora [KS]	Zapremnina prtljažnika [L]	Težina [kg]	Maksimalna brzina [km/h]	Ubrzanje do 100 km/h [s]	Klasa
14	217	75	224	950	167	14	Gradski
16	429	105	750	2000	187	13	Kombi
19	219	280	130	1100	280	6	Sportski
15	313	170	540	1700	223	7	Limuzina
18	287	200	430	2500	190	14	Terenski
...	...	...	...	...	...	...	...

Koristeći veliki broj ovakvih n-torki unutar tablice koje predstavljaju već postojeće podatke, *data mining* algoritmi kreiraju strukturalne uzorke. Moguće ih je opisati običnim izjavnim rečenicama tipa:

Za veličinu kotača od 14' do 15' i međuosovinski razmak od 195 do 225 cm automobil spada u klasu gradskih automobila.

Za težinu vozila od 2000 do 3000 kilograma i snagu motora veću od 150 konjskih snaga automobil spada u klasu terenskih vozila.

Za ubrzanje do 100 kilometara na sat manje od 7 sekundi i težinu manju do 1500 kilograma automobil spada u klasu sportskih vozila.

Na temelju seta ovakvih strukturalnih uzoraka moguće je ostvariti i automatizirati strojno učenje. Primjerice, kada se na tržištu pojavi novi automobil, možemo ga klasificirati koristeći se već naučenim pravilnostima, odnosno svrstati ga u klasu vozila kojoj pripada. Nakon toga, vrijednosti atributa novog automobila automatski se unose u bazu podataka te se ponavlja data mining kako bi strukturalni uzorci postali pametniji i točniji.

2.3. *Data mining* algoritmi

U bioinformatici se najčešće se koriste algoritmi koju spadaju u nagdledano učenje, granu indukcijskih data mining algoritama. Najbitniji za spomenti su r*andom forest, neural network i support vector machines*.

2.3.1. *Random forest*

Random forest je naziv za klasifikator koji se sastoji od više stabala odluke (eng. *decision tree*) i kao rezultat daje onu klasu koja se najviše puta pojavila kao rezultat pojedinog stabla odluke. Prednosti *random forest* algoritma su vrlo visoka točnost, mogućnost rukovanja velikim brojem ulaznih varijabli, procjenjivanje prioriteta varijabli te balansiranje pogrešaka u klasifikatorima za nebalansirane setove podataka.

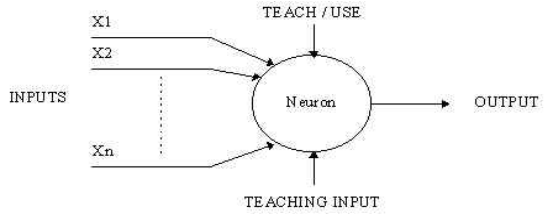
Svako posebno stablo se generira koristeći sljedeći algoritam

1. Neka je N broj ulaznih n-torki stabla, a broj varijabli u klasifikatoru M.

2. Neka je m broj ulaznih varijabli koje utječu u odlučivanju na čvoru stabla; m bi trebao biti puno manji od M
3. Odabire se set ulaznih podataka za trenutno stablo tako da se odabere N n-torki, ali sa zamjenama iz originalnog seta ulaznih podataka.
4. Za svaki čvor stabla, slučajnim odabirom odabrati m varijabli na osnovu kojih će se donijeti odluka na tom čvoru stabla. Izračunati najveću razliku dobivenu iz tih m varijabli ulaznog seta podataka.
5. Svako stablo se popunjava do kraja bez brisanja.

2.3.2. *Neural network*

Neural network ili mreža za paralelno distibuirano procesiranje je matematički model nastao oponašajući principe rada kore ljudskog mozga. Sastoji se od mreže međusobno gusto povezanih procesirajućih čvorova koji rade zajedno kako bi riješili problem. Razmjena informacija između čvorova odvija se paralelno, a proces učenja odvija se promatrajući primjere. Primjenjuje se u prepoznavanju uzoraka i klasificiranju podataka. Ovakav matematički model pristupa problemu na drukčiji način od klasičnih von Neumanovih računala, koja koriste standardni algoritamski pristup. Drugim riječima, von Neumanova računala ne znaju riješiti problem osim u slučaju kada im je zadan specifiačan set instrukcija dok se *neural network* ne može programirati u cilju rješavanja specifičnog problema, nego sama uči iz podataka koje joj se daju. Iz toga se da zaključiti da se set podataka iz kojeg *neural network* uči mora biti pažljivo odabran. U suprotnom, vrijeme utrošeno na strojno učenje je bačeno, ili još gore, mreža neće ispravno funkcionirati jer je naučila netočne podatke. Funkcija i način pristupanja problemu *neural networka* i von Neumanovih računala su komplementarni. Kako je operacije von Neumanovih računala moguće predvidjeti, obično se stvarne implementacije rade tako da von Neumanov model nadgleda *neural network* u svrhu ostvarivanja najveće iskoristivosti.



Slika 2.4. Model čvora u *neural networku*

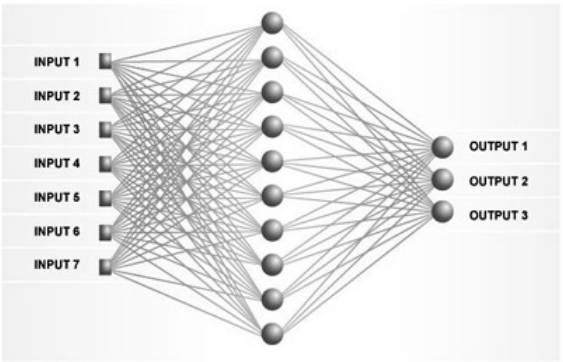
Npr. kada imamo čvor sa 3 ulaza i čvor je naučen da daje izlaz 1 (Y) za ulaze 111 ili 101 (X1,X2,X3) i da daje izlaz 0 kada je ulaz 000 ili 001. Tada ovako izgleda tablica znanja čvora

X1	0	0	0	0	1	1	1	1
X2	0	0	1	1	0	0	1	1
X3	0	1	0	1	0	1	0	1
Y	0	0	0/1	0/1	0/1	1	0/1	1

Za situacije za koje ne postoji naučeno pravilo, traži se najsličnija situacija za koju postoji definirano pravilo. Prvi nedefinirani ulaz u tablici je 010. Sada tražimo najsličniji ulaz i koristimo njegovo pravilo. Definirani ulazi su 000,001,101 i 111. Niz 010 i 111 razlikuju se u 2 elementa. Niz 010 i 101 razlikuju se u 3 elementa. Niz 010 i 001 ralikuju se u 2 elementa. Niz 010 i 000 razlikuju se u samo jednom elementu te vidimo da je to najbolja aproksimacija. Dakle odlučujemo da se za niz 010 imitira izlaz ulaza 000, što je 0. Isti se postupak primjenjuje za ostale ulaze. Za one ulaze za koje ne postoje definirano pravilo, a razlikuje se za isti minimalan broj znakova sa više definiranih različitih nizova koji generiraju različit izlaz, izlaz ostaje nedefiniran. Tada tablica izgleda ovako.

X1	0	0	0	0	1	1	1	1
X2	0	0	1	1	0	0	1	1
X3	0	1	0	1	0	1	0	1
Y	0	0	0	0/1	0/1	1	1	1

Jedna od najvećih prednosti ovakve mreže je da može dodjeljivati i oduzimati zadatke pojedinom čvoru, tako da ako neki čvor umre, drugi može preuzeti njegovu funkciju te funkcionalnost mreže nije ugrožena. Prednosti ovog algoritma su mogućnost paralelnog procesiranja, otpornost na pogreške i brzina procesiranja te prilagođenost *fuzzy logic* računalnim problemima. Nedostatak je to što mreža sama uči kako riješiti problem te stoga njezine operacije nije moguće predvidjeti.



Slika 2.5. Model *neural networka*

2.3.3. *Support vector machine*

Support vector machines je set povezanih nagledanih metoda učenja koje se koriste za klasifikaciju i regresiju. *Support vector machines* mapiraju ulazne vektore u višedimenzionalni prostor te tvoreći višedimenzionalnu plohu koja dijeli klase ulaznih vektora. Dvije nove paralelne višedimenzionalne plohe kreiraju se na svakoj strani višedimenzionalne plohe koja razdvaja podatke. Što je veća udaljenost između tih dviju paralelnih višedimenzionalnih ploha to je bolja klasifikacija. Prednosti *support vector machines* klasifikatora je simultano minimiziranje emipirijskih pogrešaka klasifikacije.

Podaci se prikazuju skupovima oblika

$$\{(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \ldots, (\mathbf{x}_n, c_n)\}$$

gdje je broj c 1 ili -1 i govori kojoj klasi pripada x, gdje je x p-dimenzionalni realni vektor, obično normaliziran na vrijednosti iz skupa [0,1] ili [-1,1]. Te podatke možemo promatrati kao skup podataka iz kojih se izvlače strukturalni uzorci koji dijele višedimenzionalnu plohu koja poprima oblik

$$\mathbf{w} \cdot \mathbf{x} - b = 0.$$

Vektor w ima smjer okomice na podijeljenu višedimenzionalnu plohu. Pomak b omogućuje nam povećavanje granica odnosno udaljenosti između dviju novih višedimenzionalnih ploha koje poprimaju oblik

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x} - b &= 1, \\ \mathbf{w} \cdot \mathbf{x} - b &= -1. \end{aligned}$$

U slučaju da su ulazni podaci linearno djeljivi, možemo odabrati granice, odnosno dvije paralelene višedimenzionalne plohe tako da između njih nema niti jednog ulaznog vektora. Tada te plohe možemo opisati jednadžabama

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}_i - b &\geq 1 \quad \text{or} \\ \mathbf{w} \cdot \mathbf{x_i} - b &\leq -1 \end{aligned}$$

što se da zapisati u obliku

$$c_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1, \quad 1 \leq i \leq n. \qquad (1)$$

što predstavlja osnovnu formu zapisa klasificikacijskih pravila. U osnovnoj se formi javlja problem minimizacije |w| člana u nejednadžbi (1), a minizira sve ovako

$$(1/2)||\mathbf{w}||^2$$

gdje je ½ zbog konzistentnosti s matematičkom teorijom.

Ako se klasifikacijska pravila napišu u dualnoj formi, primjećuje se da klasifikacija ovisi samo o *support vectors,* odnosno o podacima koji žive točno na granicama koje određuju gore navedene jednadžbe. Dualna forma je oblika

$$\max \sum_{i=1}^{n} \alpha_i - \sum_{i,j} \alpha_i \alpha_j c_i c_j \mathbf{x}_i^T \mathbf{x}_j$$

gdje alfa izrazi ukazuju na dualnu reprezentaciju težinski vektor u terminima ulaznih podataka

$$\mathbf{w} = \sum_i \alpha_i c_i \mathbf{x}_i$$

2.4. Predikcija sekundarnog oblika proteina

2.4.1. Kako?

Postoji nekolicina razloga zašto je predikcija sekundarnog oblika proteina netrivijalna i vrlo teška.

- Broj mogućih struktura koje proteini mogu imati je ekstremno velik.
- Fizikalna osnova stabilnosti strukture proteina nije sasvim istražena i objašnjena.
- Tercijarna struktura nekih proteina ne može biti ispravno definirana bez poznavanja drugih proteina koji utječu na presavijanje dotičnih proteina.
- Određena sekvenca može poprimiti više uobličenja, ovisno o okolini u kojoj se nalazi, te biološki aktivno uobličenje ne mora nužno biti termodinamički najpogodnije.
- Direktna simulacija presavijanja proteina metodama kao što je molekularna dinamika (interakcija molekula po zakonima fizike) nije jednostavna za izvođenje na kućnim računalima, osim kada se radi o vrlo malim proteinima

Poznavajući primarni oblik, moguće je sa određenom točnošću pretpostaviti sekundarni oblik. Sekundarni je oblik trodimenzionalna struktura predstavljena uzorcima vodikovih veza između osnovica sastavljenih od skupina amida. Točnije, protein se nalazi u sekundarnom obliku nakon što se dogodilo presavijanje. To se događa uslijed težnje za minimalnom energijom. Predikcija se odvija promatranjem parametara proteina koji bitno utječu na to kako će se protein presaviti. Bitno je odrediti na kojim je mjestima protein fleksibilan, a na kojima je rigidan. Fleksibilna mjesta su oni dijelovi sekvence aminokiselina na kojima se prostorno savijanje odvija uz malu potrošnje energije i minimalne promjene molekularne strukture, dok su rigidni sektori oni gdje je sekvenca aminokiselina takva da ne dopušta značajnije pomake u prostoru.

2.4.2. Weka

Istraživanja pokazuju da ne postoji jedinstvena shema strojnog učenja koja bi bila primjenjiva na sve probleme koje želimo riješiti. Weka (eng. *Waikato Environment for Knowledge Analysis*) je programski paket koji nudi već gotov set algoritama strojnog učenja i alata za procesiranje podataka te podršku za datoteke ARFF (eng. *Attribute-Relation File Format*) tipa, kao i rezultate dobivene vršenjem upita na bazu podataka. Kroz nekoliko vrlo intuitivnih grafičkih sučelja moguće je odabrati i primjeniti neki od algoritama i filtera na vlastiti set podataka bez pisanja i jedne linije programskog koda. U primjeru u slijedećem poglavlju služiti ćemo se Explorer grafičkim sučeljem unutar kojeg ćemo odabrati shemu strojnog učenja te eventualno filtre kojima želimo klasificirati odabrani set podataka. Weka također nudi opcije vizualizacije dobivenih rezultata.

2.4.3. Primjer data mining-a u bioinformatici

Korisiti ćemo Weka alatom za data mining koji ćemo provesti nad unaprijed pripremljenim podacima zapisanim u ARFF datoteci.

Izgled ARFF datoteke:

```
%navođenje relacije
@RELATION asa
%navođenje atributa
@ATTRIBUTE pdb_id STRING
@ATTRIBUTE chain STRING
@ATTRIBUTE residue1 {ALA,CYS,ASP,GLU,PHE,GLY,HIS,ILE,LYS,LEU,MET,ASN,PRO,GLN,ARG,SER,THR,VAL,TRP,TYR}
@ATTRIBUTE residue2 {ALA,CYS,ASP,GLU,PHE,GLY,HIS,ILE,LYS,LEU,MET,ASN,PRO,GLN,ARG,SER,THR,VAL,TRP,TYR}
...
@ATTRIBUTE residue_number STRING
@ATTRIBUTE asa NUMERIC
@ATTRIBUTE backbone_asa NUMERIC
@ATTRIBUTE hydrophobicity NUMERIC
...%slijedi još atrubuta
%unos podataka
@DATA
1A0O,A,ALA,ASP,LYS,GLU,LEU,LYS,PHE,LEU,VAL,6,435.906,275.22182,77.4114,196.15806,358.4942,278.67347,...
1A0O,A,ASP,LYS,GLU,LEU,LYS,PHE,LEU,VAL,VAL,7,376.5934,219.37799,36.7747,104.15183,339.8183,249.45587,...
1A0O,A,LYS,GLU,LEU,LYS,PHE,LEU,VAL,VAL,ASP,8,287.0741,157.23766,24.4546,70.07133,262.6191,177.91546,...
1A0O,A,GLU,LEU,LYS,PHE,LEU,VAL,VAL,ASP,ASP,9,262.4328,157.09606,46.0523,129.63113,216.3801,161.84016,...
1A0O,A,LEU,LYS,PHE,LEU,VAL,VAL,ASP,ASP,PHE,10,268.5538,150.87696,34.7397,98.3915,233.814,159.09416,...
1A0O,A,LYS,PHE,LEU,VAL,VAL,ASP,ASP,PHE,SER,11,311.5575,194.90286,35.2666,97.88325,276.2908,228.67186,...
A0O,A,PHE,LEU,VAL,VAL,ASP,ASP,PHE,SER,THR,12,320.6841,213.62586,36.0506,100.39727,284.6334,255.29656,...
%nisu navedene vrijednosti za sve atribute zbog nečitkog izgleda
%broj unesenih n-torki je mnogo veći od onoga što je ovdje vidljivo (tekstualna datoteka od cca 30 MB)
```

Atributi koji su navedeni u ARFF datoteci opisuju pojedini protein. Svaki unos, odnosno n-torka vrijednosti atributa predstavlja pojedini protein. Opisati ćemo značenja nekih važnijih atributa.

- pdb_id: znakovni niz koji predstavlja jedinstveni naziv proteina predstavljenog pdb (eng. *Protein Data Bank*) datotekom. U njoj se nalazi detaljniji opis proteina.
- chain: znakovni niz koji predstavlja na koji lanac aminokiselina se odnose ostali atributi
- residue(i): opisuje od kojih se aminokiselina sastoji referentni lanac
- asa: asa(eng. *Accessible Surface Area*) je numerička vrjednost koja predstavlja površinu molekule dostupne otapalu (vodi). Obično je zadana u angstremima, što je standarnda jedinica u bioinformatici (0.1 nm)
- backbone_asa: asa na okosnici proteina

- hydrophobicity: numerička vrijednost koja predstavlja hidrofobnost, odnosno koeficijent koji opisuje koliko molekula odbija ili prihvaća interakciju s molekulama vode

Potrebno je učitati datoteku ARFF tipa. Nakon preprocesiranja podataka odabiranjem atributa koje Weka može analizirati (odstranjivanje atributa koji su znakovni nizovi), potrebno je odabrati klasifikator. Za klasifikaciju podataka koristiti ćemo standardni i zbog jednostavnosti pogodniji tree.J48 klasifikator. Nakon dugotrajnog procesiranja (>1h, >150000 ulaznih n-torki, procesor od 1.5 Ghz), Weka je dala ovakav izlaz:

```
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      136802               80.381  %
Incorrectly Classified Instances     33390               19.619  %
Kappa statistic                          0.2451
Mean absolute error                      0.2937
Root mean squared error                  0.3918
Relative absolute error                 86.374  %
Root relative squared error             95.0125 %
Total Number of Instances           170192
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.965    0.778      0.817     0.965   0.885      0.677     0
                 0.222    0.035      0.639     0.222   0.33       0.677     1
=== Confusion Matrix ===
     a      b   <-- classified as
 128592   4630 |     a = 0
  28760   8210 |     b = 1
```

Nula u matrici i tablici označava neispravno, dok jedinica označava ispravno klasificirane uzorke. Najbitniji stupci tablice su preciznost i odziv, a računaju se iz dane matrice gdje se pojedini elementi interpretiraju ovako

TN(True Negative) FP(False Positive)

FN(False Negative) TP (True Positive)

Točnost dobivamo iz formule: Precision = TP / (FP + TP),

a odziv Recall : TP / (FN + TP)

Što su ti koeficijenti veći, klasifikacija je bolja.

3. Zaključak

Velika istraživanja u bioinformatici trenutno se obavljaju na polju presavijanja i prepoznavanja proteina što doprinosi točnosti predikcije trodimenzionalnig oblika. Bioinformatika bavi se dizajniranjem proteina koji zadovoljavaju željene funkcije. Postoje dvije strategije. Prva se naziva racionalno dizajniranje te se kod nje koristi detaljan opis proteina da bi se napravile minimalne promjene i ostvarila tražena funkcija. Ova je strategija jeftina te su mutacije proteina dobro istražene. Glavni je problem to što detaljna trodimenzionalna struktura proteina često nije dostupna. Druga metoda naziva se direktna evolucija gdje se slučajne mutageneze primjenjuju na protein, a provjeravajući određene granice kvalitete iterativno se odbacuju one varijante koje nemaju tražene kvalitete. Ova metoda oponaša prirodnu evoluciju i u većini slučajeva daje bolje rezultate nego racionalno dizajniranje. Glavna prednost ove metode je što ne zahtjeva prethodno poznavanje trodimenzionalne strukture niti mora predviđati kakav će biti rezultat promjene, dok je nedostatak skup i dugotrajan proces.

Data mining se koristi u racionalnom dizajniranju za predikciju trodimenzionalne strukture i radi to vrlo dobro. Velikoj točnosti predikcije treba zahvaliti naprednim računalnim algoritmima te velikim bazama podataka o proteinima koje rastu iz dana u dan. *Data mining* našao je primjenju i u mnogim drugim znanstvnim disciplinama. Npr. primjenujući *data mining* na podatke o osobama koje koriste određeni lijek, moguće je otkriti koji lijekovi vrlo štetno utječu na osobe sa određenim karakterstikama. Također, primjena *data mininga* na setove podataka koji se tiču individualnih osoba pokreće dvojbe oko legalnosti, etičnosti takve metode te zadiranja u privatni život pojedinca.

4. Literatura

[1]Aydin,Z. and Altunbasak,Y., A signal processing aplication in genomic research: Protein Secondary Structure Prediction, IEEE Signal Processing Magazine, July 2006, str. 128. - 131.

[2]Vaidyanathan, P.P., Genomics and Proteomics:A Signal Processor's Tour, IEEE Circuits and Systems Magazine, Fourth Quarter 2004, str. 6. – 29.

[3]Witten,I.H. and Frank,E., Data Mining: Practical Machine Learning Tools, Second Edition, San Francisco: Elsevier Inc, 2005

[4]Li,J., Wong.L and Yang Q., Data Mining in Bioinformatics, IEEE Inteligent Systems, November/December 2005, str. 16. - 18.

[5]Rost.B, Review: Protein Secondary Structure Prediction Continues to Rise, Journal of Structural Biology, November 2000,str. 1. - 15.

[6]Jones N.C. and Pevzner P., An Introdcution to Bioinformatics Algorithms, Cambridge, Massachusetts, London, England: The MIT Press, 2004

[7]Primary Structure,21 April 2007, http://en.wikipedia.org/wiki/Primary_structure, 3 May 2007

[8]Secondary Structure,11 April 2007,http://en.wikipedia.org/wiki/Secondary_structure, 1 May 2007

[9]Protein Folding ,26 April 2007, http://en.wikipedia.org/wiki/Protein_folding, 28 April 2007

[10]Data Mining ,3 May 2007, http://en.wikipedia.org/wiki/Data Mining, 4 May 2007

[11]Machine Learning ,1 May April 2007, http://en.wikipedia.org/wiki/Machine_learning, 4 May 2007

[12]Random Forest ,1 April 2007, http://en.wikipedia.org/wiki/Random_forest, 4 May 2007

[13]Support Vector Machine ,1 May 2007, http://en.wikipedia.org/wiki/Support_Vector_Machine, 5 May 2007

[14]Neural Network ,2 May 2007, http://en.wikipedia.org/wiki/Neural_network, 5 May 2007

5. Sažetak

Proteini su molekule koje nastaju prepisivanjem dijelova molekule DNA te prevođenjem genskog koda. Njihove su funkcije od vitalne važnosti za ispravan rad organizma te je stoga bitno da se proces sinteze odvija ispravno. Sinteza se odvija na staničnom organelu zvanom ribosom, gdje se posredstvom molekula mRNA i tRNA stvara niz kodona na koje su povezane esencijanle ili alfa-aminokiseline tvoreći sekvencu odnosno protein u njegovom primarnom obliku. Uz parametre okoline u kojoj se takav protein nalazi, raspored gradivnih jedinica sekvence određuje u kakav trodimenzionalni oblik će se presaviti. Presavijanje se odvija uslijed težnje za minimalnom energijom mirovanja te se proces presavijanja u stvarnosti odvija u rasponu od nekoliko milisekundi do nekoliko sekundi. Kada se protein presavije, kažemo da se nalazi u sekundarnom obliku. Sekundarni oblik proteina određuje kako će se protein ponašati, koju će funkciju obavljati, s kojim će proteinima interagirati i tvoriti nove... Koristeći se bazama podataka o proteinima čiji su primarni oblici, trodimenzionalne strukture i načini ponašanja već proučeni te razotkriveno koje sekvence kako utječu na presavijanje, moguće je uvidjeti pravilnosti ili uzorke. Poznavajući takve uzorke, moguće je sa određenom vjerojatnošću pretpostaviti kako bi se presavio i ponašao neki novi upravo otkriveni ili onaj kojeg ćemo možda tek dizajnirati da ostvarimo funkcionalnost koja nam je potrebna. Predikcija se vrši korištenjem naprednih računalnih metoda. Jedna od takvih metoda je *data mining*. Izvodi se tako da se primjenom matematičkih algoritama ustanove strukturalni uzorci odnosno pravilnosti u ponašanju entiteta nad kojima se provodi *data mining*. Takvi strukturalni uzorci omogućuju određivanje kojoj od klasa pripada neki novi entitet, čime je moguće ostvariti i automatizirati strojno učenje.