SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Pregled trenutnih dostignuća i problema u području dokiranja proteina (surface matching)

Martina Perše

Voditelj: *Krešimir Šikić*

Zagreb, svibanj, 2007.

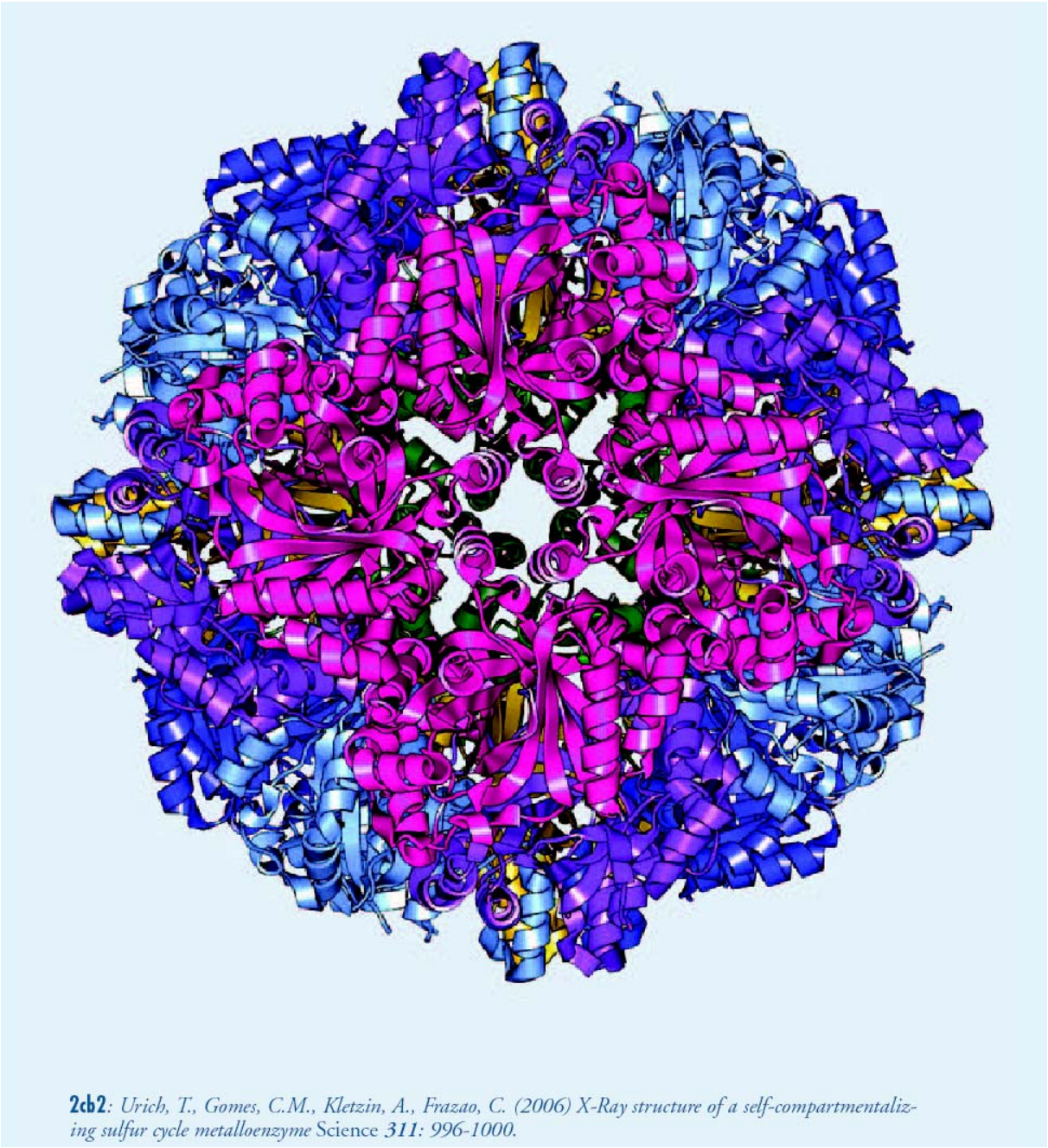
Sadržaj

Slike

Skraćenice

PDB – Protein Data Bank

NMR – Nulearna Magentska Rezonancija (eng. Nuclear magnetic resonance)



Slika 1. 3-D struktura proteina

1. Uvod

Bioinformatika je disciplina koja objedinjuje znanja, metode i tehnike iz primijenjene matematike, informatike, računarstva, umjetne inteligencije, kemije, biologije, biokemije i drugih srodnih znanosti. Dijeli se na protenomiku i genomiku.

Protenomika se bavi proučavanjem proteina, a genomika proučavanjem gena. Svakodnevna otkrića, mnoštvo raznolikosti i kompleksnost biomolekula zahtijevaju efikasne i sofisticirane tehnike i metode kako bi mogli organizirati, analizirati i interpretirati dobivene rezultate na biološki smislen način. Protein - protein interakcije, dokiranje proteina, predviđanje strukture proteina su neka od istraživanja koja se provode na području protenomike.

Proteini ili bjelančevine su sastavni dijelovi svake stanice koja čini osnovu života na Zemlji. Proteini su organske strukture građene od dvadesetak različitih aminokiselina, međusobno povezanih poput karika u lancu. Geni definiraju redoslijed i broj aminokiselina. Poznavanje strukture proteina je ključno za razumijevanje specifičnih karakteristika svake bjelančevine i njezine funkcije. Proteini mogu međusobno djelovati udružujući se u složenije strukture. Postoje primarne, sekundarne, tercijarne i kvarterne strukture. Biomolekularna međudjelovanja čine srž u svim metaboličkim procesima i regulacijama.

Promatrajući bilo koja dva proteina pred nama se otvaraju tri osnovna pitanja:

(1) Da li će dva proteina međusobno djelovati?

(2) Ako da, kako će djelovati?

(3) Kako će izgledati struktura njihovog kompleksnog spoja?

Relativno je lako eksperimentalno odgovoriti na ova pitanja. Zbog prevelikog broja različitih proteina i mogućih interakcija javlja se potreba za razvijanjem brze i učinkovite računalne metode, koja bi mogla predvidjeti protein - protein interakcije.

2. Surface matching metoda

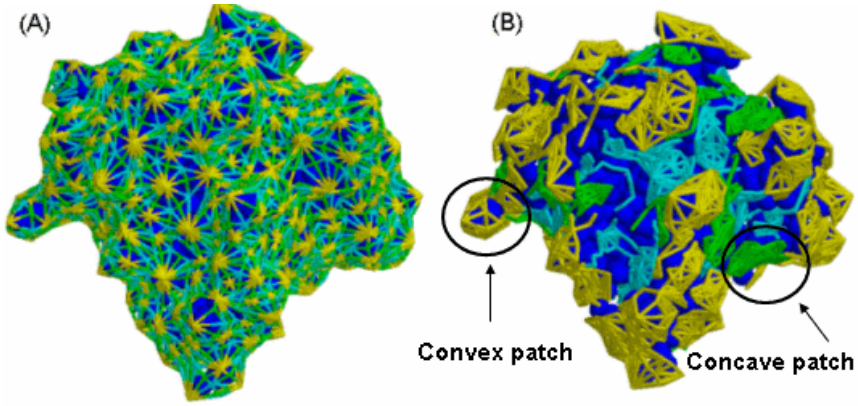
Poznavanjem strukture proteina od brojnih mogućnosti spajanja možemo različitim metodama predvidjeti koji bi proteini mogli međudjelovati. Cilj je pretraživanjem baze podataka poznatih proteina izdvojiti sve proteine, koji bi mogli reagirati s određenom strukturom. Proteini su na nekim mjestima fleksibilni i kod njihove interakcije oni mogu utjecati na strukturu proteina s kojim reagiraju. Time je pronalazak kompatibilnih mjesta na njihovoj površini otežan.

Metodom surface matching na temelju poznavanja površine dvaju proteina, koje promatramo možemo odrediti da li će se oni dokirati i na kojim mjestima. Često se kaže da se proteini spajaju po sistemu ključ - brava. Za istraživanja surface matching metodom važno je poznavanje svih relevantnih informacija o strukturama proteina.

Otkrivanje složene strukture proteina proučava strukturalna biologija. Neke od metoda koje se koriste su rendgenska kristalografija, NMR spektroskopija (Nuklearna magnetska rezonancija), elektronskim mikroskopom i dr. Od 42000 poznatih struktura proteina, njih 36000 dobiveno je rendgenskom kristalografijom, 6000 NMR spektroskopijom i preko 140 struktura proteina elektronskim mikroskopom.

Veliku količinu informacija o proteinima potrebno je na smislen način organizirati, kako bi bilo olakšano njihovo pregledavanje i pretraživanje. Postoji mnogo baza podataka koje sadrže podatke o 3-D strukturi proteina i nukleinskih kiselina. Jedna od najpoznatijih i najkorištenijih je PDB (Protein Data Bank). PDB nije baza podataka u pravom smislu riječi. Nad njome se ne mogu postavljati upiti. PDB sadrži skup datoteka s podacima o proteinima. Znanstvenici širom svijeta je svakodnevno zajednički upotpunjuju informacijama o novootkrivenim proteinima. Pomaci u znanosti i tehnologiji su uzrokovali nagli rast PDB-a. Kroz PDB dostupne su nam različite informacije povezane sa strukturom proteina, poput detalja o slijedu, koordinate atoma u prostoru, uvjeti kristalizacije, geometrijski podaci, strukturalni faktori, 3-D slike i mnogi drugi. Upravo podaci iz PDB-a su nam ključni za daljnja istraživanja protein-protein interakcija opisana u ovom seminaru.

Tehnološki razvoj, rast baze podataka o strukturama poznatih proteina, te povećanje brzine procesora kod kompjuterske analize omogućili su lakše otkrivanje molekularnih interakcija.



Slika 2. Konveksni i konkavni dijelovi proteina

2.1 Geomerijsko hashiranje

Geometrijsko hashiranje je tehnika koja je originalno osmišljena za rješenje problema računalnog vida. Pomoću nje je npr. robotu omogućeno da raspozna neki objekt na temelju slike, koju prima senzorima ili video kamerom i uspoređuje s podacima pohranjenim u njegovoj memoriji.

Postoji mogućnost da su promatrani objekti podlegli transformacijama u odnosu na inicijalnu poziciju u bazi podataka. Također je postoje slučajevi kada nam nisu dostupne sve informacije, jer su objekti djelomično prekriveni ili se neki njegovi dijelovi ne nalaze u bazi podataka. Tehnika geometrijskog hashiranja je u takvim situacijama primjenjiva i time vrlo učinkovita.

Susrećemo se sa dva osnovna problema: kako prikazati površinu promatranog objekta u računalu i kako raspoznati promatrani objekt algoritmom usporedbe.

2.1.1 Prikaz površine u računalu, preprocesna faza

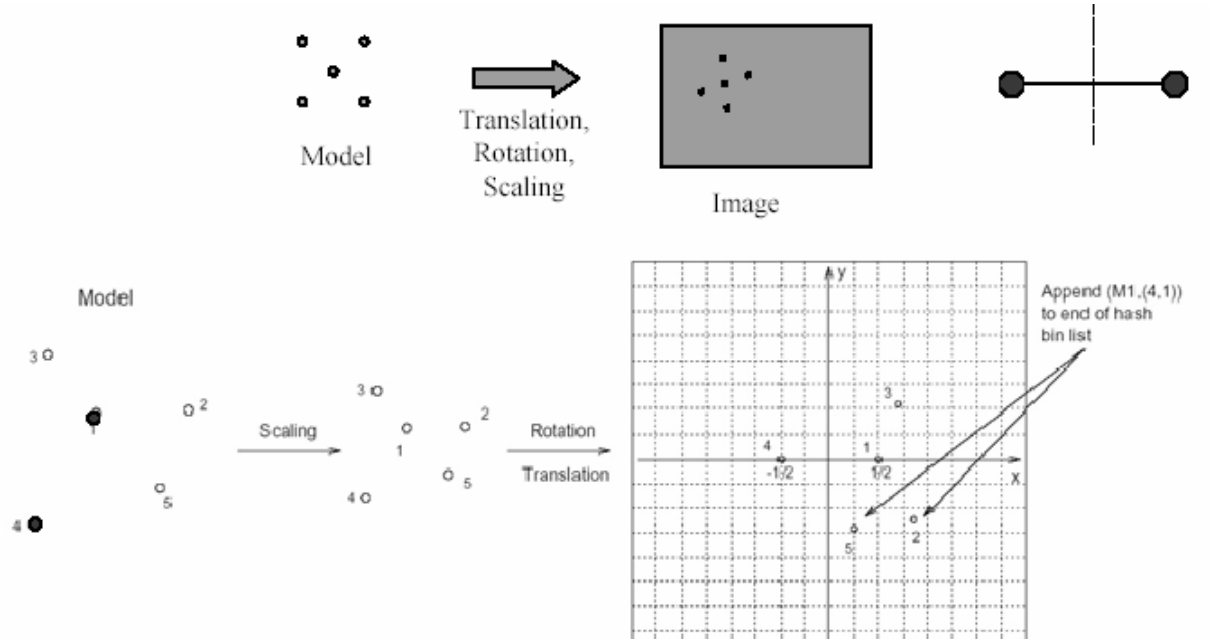
Svaki objekt, koji ćemo algoritmom usporedbe pokušati raspoznati je u računalu prikazan svojim modelom. Podatke o svim modelima čuvamo u bazi podataka. Moramo imat dovoljnu količinu podataka o modelu, tako da on jednoznačno određuje odgovarajući objekt. Podaci moraju istovremeno biti dovoljno kratki, kako bi algoritam pretraživanja bio efikasan.

Kreiranje baze podataka se često naziva predprocesna faza. Sastoji se od generiranja hash tablice, koja se kasnije koristi u fazi usporedbe. Pohranjivanje podataka o modelima je dovoljno učiniti samo jednom.

Potrebno je izdvojiti osnovne karakteristike o objektu, kojeg želimo aproksimizirati modelom. One mogu biti točke, linije i druge prikladne značajke. Npr. Izdvojimo m točaka. Od m signifikantnih točaka odabiremo dvije točke $\overrightarrow{p_{\mu 1}}$ i $\overrightarrow{p_{\mu 2}}$, koje čine par baze.

Kreiramo kartezijev koordinatni sustav tako da kroz točke $\overrightarrow{p_{\mu 1}}$ i $\overrightarrow{p_{\mu 2}}$ položimo x os. Y os je okomica provučena kroz polovište dužine $\overline{p_{\mu 1} p_{\mu 2}}$. Duljinu dužine $\overline{p_{\mu 1} p_{\mu 2}}$ uzimamo kao jediničnu duljinu. Točkama $\overrightarrow{p_{\mu 1}}$ i $\overrightarrow{p_{\mu 2}}$ pridjeljene su koordinate (-0.5, 0) i (0.5, 0). S obzirom na opisani referentni koordinatni sustav svim preostalim m-2 točkama su pridjeljene odgovarajuće koordinate. Koordinatni sustav ostaje nepromijenjen kada je model podvrgnut transformacijama, jer ga par baze jednoznačno određuje. Opisani postupak kreiranja koordinatnog sustava je ilustriran na slici 3.

Algoritam treba omogućiti prepoznavanje i u slučaju kada je dio objekta zaklonjen. Tada ne možemo znati da li će obadvije točke para baze biti vidljive. Stoga je potrebno sačuvati podatke o modelu za sve moguće parove baze. Hash tablica sadrži zapise oblika (model, baza).



slika 3. Kreiranje koordinatnog sustava

Neka su $\overrightarrow{p_x}$ i $\overrightarrow{p_y}$ jedinični vektori, gdje $\overrightarrow{p_x}$ gleda u pozitivnom smjeru x osi, a $\overrightarrow{p_y}$ gleda u pozitivnom smjeru y osi. Neka su $\overrightarrow{p_{\mu 1}}$ i $\overrightarrow{p_{\mu 2}}$ točke baze. Svaka točka p u koordinatnom sustavu može se prikazati pomoću jednadžbe:

$$\overrightarrow{p_i} = u\overrightarrow{p_x} + v\overrightarrow{p_y} + \overrightarrow{p_0} \qquad (1)$$

gdje je $\overrightarrow{p_0} = \frac{\left(\overrightarrow{p_{\mu 1}} + \overrightarrow{p_{\mu 2}}\right)}{2}$. Skalarne vrijednosti u i v ostaju nepromijenjene nakon transformacija poput rotacije, translacije i skaliranja.

Njihovom kvantizacije računamo indekse $\left(u_q, v_q\right)$ koji nam koriste za pristupanje određenoj lokaciji u dvodimenzionalnoj hash tablici, koja sadrži podatke oblika (model, ($\overrightarrow{p_{\mu 1}}$, $\overrightarrow{p_{\mu 2}}$)).

Prethodno opisani postupak potrebno je ponoviti za svaki model. Zapisi se uređuju po principu hash tablice kako bi njihovo pretraživanje bilo što brže i efikasnije.

Za par baza moguće je uzeti više od dvije točke. Odaberemo li tri točke one će činiti trodimenzionalni koordinatni sustav s jediničnim vektorima $\overrightarrow{p_x}$, $\overrightarrow{p_y}$ i $\overrightarrow{p_z}$. Sve preostale točke u trodimenzionalnom koordinatnom sustavu se mogu prikazati pomoću $\overrightarrow{p_i} = u\overrightarrow{p_x} + v\overrightarrow{p_y} + w\overrightarrow{p_z} + \overrightarrow{p_0}$. (2)

Pri tome trojku (u, v, w) koristimo kao indeks u hash tablici. Tada je svaka točka prikazana pomoću više elemenata, čime je njezina preciznost veća, ali se povećava

vrijeme izvođenja algoritma. Stoga je potrebno pronaći kompromis, radi zadržavanja što boljih karakteristika.

2.1.2 Faza usporedbe

U fazi usporedbe potrebno je izdvojiti n značajki (točke, linije i dr.) sa slike. Kao u preprocesnoj fazi na isti način odrediti parove baze i koordinate preostalih točaka.

Na temelju indeksa dobivenih kvantizacijom pristupamo pojedinim zapisima u hash tablici. Svakom pronađenom zapisu (model, baza) pridjeljujemo glas. One parove (model, baza) s brojem glasova iznad određene granice, uzimamo kao kandidate za potencijalnu podudarnost između modela i promatranog objekta.

Ne očekujemo da se postupkom uspoređivanja izdvoji jedan kandidat. Cilj je signifikantno smanjiti broj mogućih kandidata za slijedeći korak.

Potrebno je uzeti u obzir moguće transformacije, poput rotacije, translcije i skaliranja. Za sve parove kandidata (model, baza) računamo moguće transformacije i tražimo one parove koji se najtočnije podudaraju sa slikom.

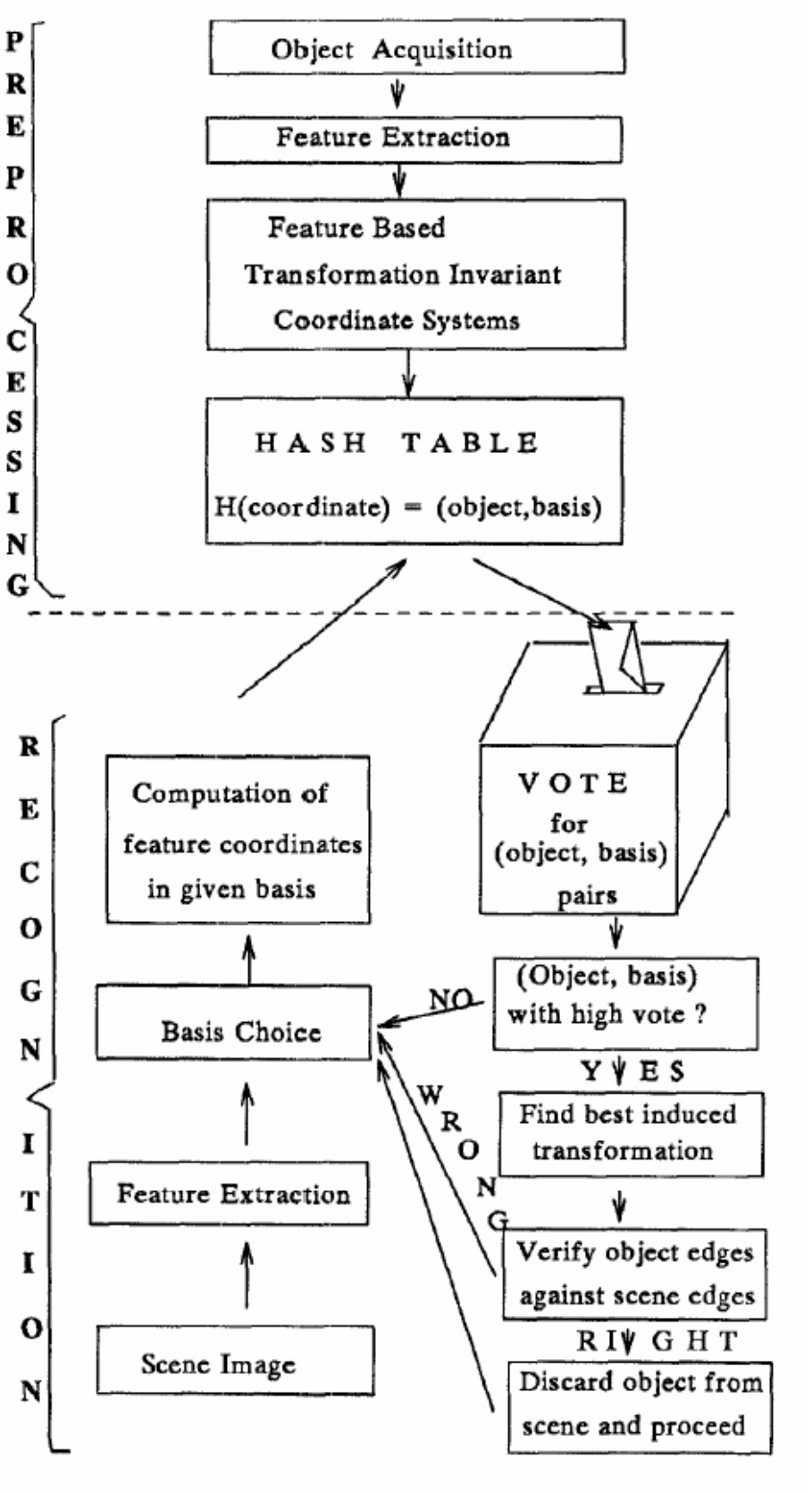
Ukoliko je objekt djelomično prekriven dio točaka modela nedostaje. Prepoznavanje je i dalje moguće sve dok odgovarajući parovi (model, baza) dobiju dovoljan broj glasova.

2.1.3 Analiza složenosti algoritma

Neka baza podataka sadrži M modela, pri čemu je svaki model prikazan pomoću m značajki, a c značajki čini bazu. Iz promatrane slike izdvajamo n značajki. Složenost predprocesne faze u najgorem slučaju je $O\left(Mm^{c+1}\right)$. Složenost faze usporedbe je $O\left(Hn^{c+1}\right)$. H predstavlja kompleksnost pristupa hash tablici. Ovisi o zauzeću tablice i distribuciji pretinaca. Uzmemo li da je tablica istog reda kao zapisi i da je uniformno distribuirana vrijedi jednakost $\mathrm{H} = \mathrm{O}(1)$. Ukoliko hash tablica sadrži mali broj pretinaca, vrijeme pristupa ovisi o broju elemenata.

U prethodno promatranom slučaju smo za bazu izdvojili dvije točke, stoga je c= 2. Za unos pojedinog modela u memoriju složenost iznosi $O\left(m^3\right)$. Uz pretpostavku da je $\mathrm{H} = \mathrm{O}(1)$ složenost faze prepoznavanja je $O\left(n^3\right)$.

Potrebno je primijetiti da se preprocesna faza izvodi neovisno o fazi usporedbe i najčešće se izvodi samo jedanput. Zato se vrijeme utrošeno na kreiranje baze podataka s modelima ne ubraja u vrijeme potrebno za raspoznavanje promatranih objekata.



slika 4. Geometrijsko hashiranje

2.1.4 Algoritam

1. Predprocesna faza:

1. Izdvojiti s uzorka m signifikantnih točaka.
2. Od m točaka izabrati dvije točke.
3. Dvije odabrane točke čine par baze i određuju referenti koordinatni sustav.
4. Za svih preostalih m-2 točaka odrediti koordinate (x, y) u referentnom koordinatnom sustavu.
5. Iz (x, y) koordinara i jednadžbe (1) izračunati (u_q, v_q)
6. Par (u_q, v_q) koristimo kao indeks za 2-D hash tablicu, te na određenu poziciju u hash tablici spremamo podatak o modelu i bazi (model, baza).
7. Ponavljati korake od 2 do 6 za svaki par baze.

2. Faza usporedbe:

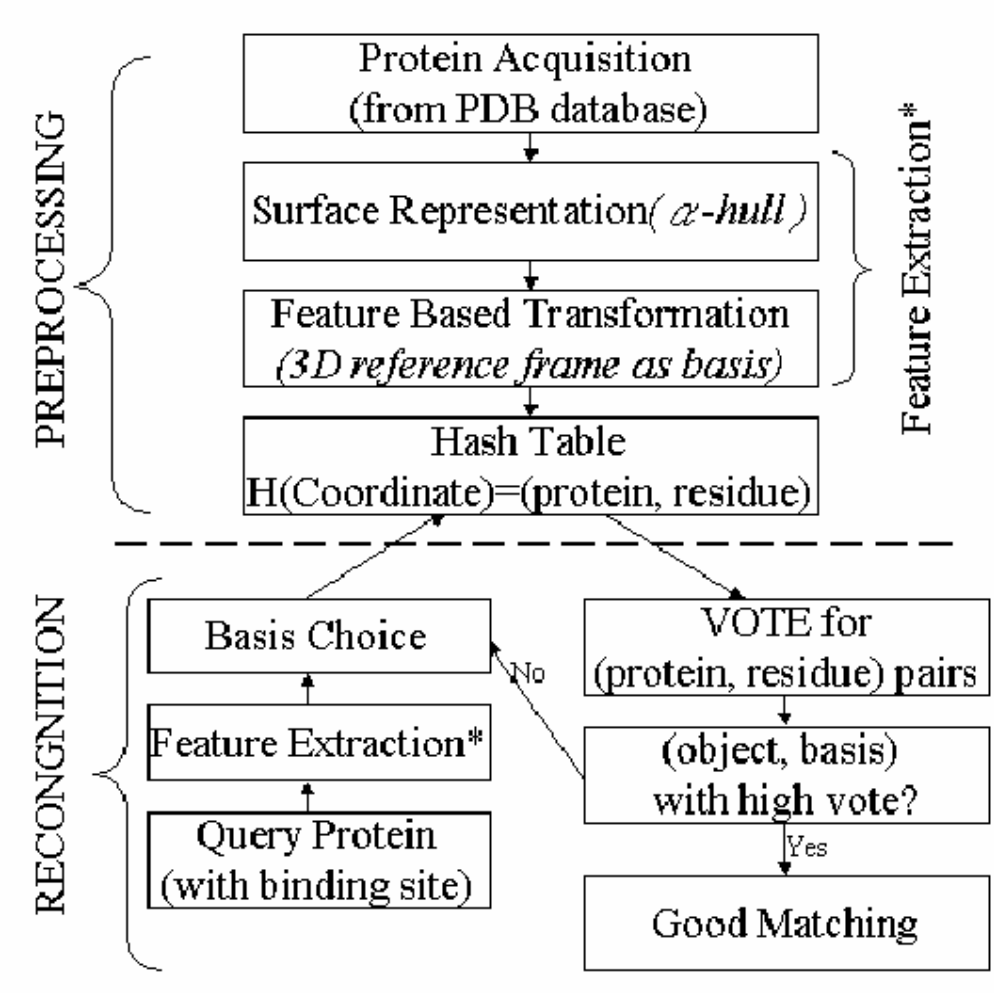
1. Izdvojiti n signifikantnih točaka iz promatrane slike.
2. Od n točaka odabrati proizvoljan par točaka za bazu i definirati koordinatni sustav.
3. Izračunati koordinate za preostalih n-2 točaka u referentom koordinatnom sustavu.
4. Kvantizirati izračunate koordinate.
5. Pristupiti odgovarajućim zapisima u hash tablici na temelju indeksa (u_q, v_q). Za svaki pronađeni zapisi oblika (model, baza) u hash tablici pridijeliti glas za određeni model i bazu.
6. Izdvojiti sve parove (model, baza) koji su dobili jedan ili više glasova. Uzimamo u obzir kao kandidate za potencijalnu podudarnost samo one parove, koji su dobili veći broj glasova od određene granice.
7. Za svakog kandidata dobivenog u koraku 6. izračunati transformacije
8. Pratiti sličnosti između kandidata i slike, izračunati sličnost za dodatne posebnosti, koje ta transformacija uzrokuje. Konačno pronaći transformaciju koja odgovara svim podudarnostima.
9. Ukoliko u koraku 8. nismo uspjeli pronaći sličnost između slike i modela za odabrani par baza, vratiti se na korak 2.

2.2 Primjena geometrijskog hashiranja u bioinformatici

Tehnika geometrijskog hashiranja je našla primjenu u mnogim područjima, pa tako i u bioinfomratici. Kod analize površine u svrhu pronalaženja kompatibilnih točaka na površini proteina koja bi mogla odgovarati po sistemu ključ-brava tehnika geometrijskog hashiranja se pokazala vrlo učinkovitom.

Protein podijelimo na manje dijelove od kojih su neki konveksni, konkavni, ravni. Zatim ih kao puzle slažemo tako da spajamo ravni s ravnim, konveksni s konkavnim dijelom i obratno.

Posljednjih godina zabilježen je nagli rast broja poznatih proteina i podataka o njihovim 3-D strukturama. Oni se čuvaju u Protein Data Bank (PDB). Iz PDB dokumenta je moguće dobiti sve podatke o 3-D strukturama proteina, o položaju pojedinih molekula u proteinu koje međudjeluju s drugim proteinima i molekulama. Na temelju informacija o strukturi, na prethodno opisan način izgrađuje se hash tablica i algoritmom usporedbe traže kompatibilna mjesta na površini proteina.



Slika 5. Geometrijsko hashiranje u bioinformatici

2.2.1 α -hull algoritam

Kako se većinom proteini spajaju na površini, postupak računanja možemo pojednostaviti tako da uzimamo u obzir samo površinske dijelove proteina. Time je broj zapisa u hash tablici smanjen, jer sadrži samo površinske dijelove. Za izdavanje površinskih dijelova se koristi algoritam α -hull.

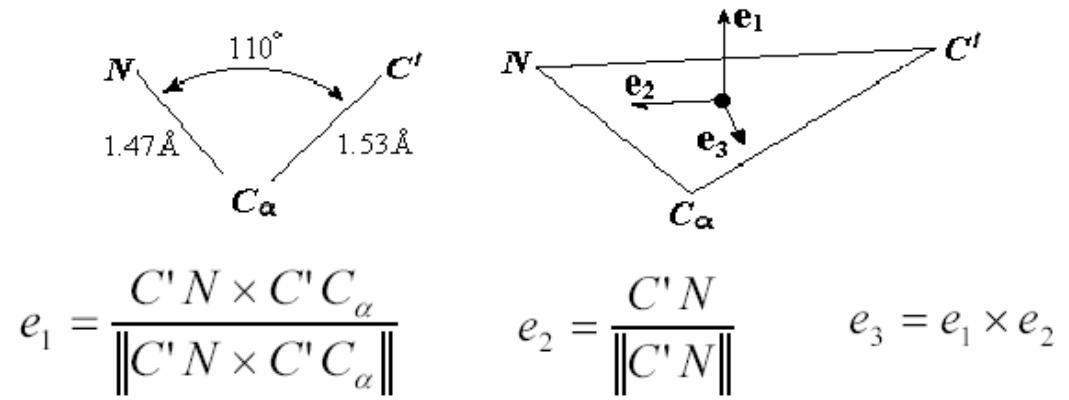
2.2.2 Dayhoffova matrica sličnosti

Kod predviđanja interakcija proteina nije dovoljno promatrati njihovu površinsku kompatibilnost. Moguće je algoritmom geometrijskog hashiranja dobiti proteine, koji se geometrijski podudaraju po sistemu ključ- brava, ali s potpuno neodgovarajućim kemijskim svojstvima. Stoga je u daljnjim analizama potrebno uzeti u obzir redoslijede sekvenci.

Dayhoffova matrica sličnosti se često koristi za poravnavanje sekvenci. Njome se izražava sličnosti između dva podatka, pri čemu sličniji podaci dobivaju višu vrijednost, a oni manje slični manju vrijednost u matrici. Nakon što izdvojimo skup kandidata u matrici sličnosti u fazi usporedbe, provjeravamo njihovu kemijsku kompatibilnost. One molekule ili strukture čija je udaljenost veća od 2Å ne uzimamo u obzir za moguće interakcije. Biraju se molekule ili strukture koje u matrici sličnosti imaju najveće vrijednosti, koje podijeljene s udaljenošću daju vrijednost veću od 1 Å. Za vrijednosti manje od 1 Å, udaljenost ne utječe na glasovanje. Ako tim postupkom nismo izdvojili niti jednu strukturu ili molekulu, promatramo manje vrijednosti u matrici sličnosti. Postupak pokazuje uspješnost i za strukture proteina koje nisu precizno definirane, sve dok su uobličenja malena.

2.2.3 3-D referentni koordinatni sustav

Okosnica proteina je sastavljena od atoma ugljika (C_α). Geometrija atoma spojena s C_α je precizno određena. Tri atoma natrij (N), ugljik (C_α) i ugljik (C) čine trokut pomoću kojeg se može jednoznačno definirati koordinatni sustav opisan slikom. Na taj način se smanjuje broj mogućih odabira baza i olakšava pretraživanje.



Slika 6. 3-D koordinatni sustav

2.3 Poboljšanja algoritma

Postupak geometrijskog hashiranja je vrlo skup zbog velikog broja različitih proteina i njihovih složenih struktura. Zato se nastoji različitim metodama smanjiti vrijeme i složenost izvođenja algoritma. U ovom poglavlju razmatramo neke postupke kojim se algoritam geometrijskog hashiranja može dodatno pojednostaviti i time smanjiti njegovo vrijeme izvođenja.

2.3.1 Podmodeli i višebrojnost hash tablica

Koristimo li samo jednu hash tablicu za čuvanje svih modela koji posjeduju velik broj karakterističnih točaka, imati ćemo veliki broj kombinacija parova baza i koordinata za računanje. Neki pretinci će imati mnogo zapisa, čime se vrijeme pristupa pojedinom zapisu u predprocesnoj fazi i fazi prepoznavanja povećava. Smanjenjem broja karakterističnih točaka smanjili bismo broj kombinacija parova baza i koordinata, koje je potrebno računati, ali bismo istovremeno izgubili dio potrebnih informacija.

Postoje postupci kojima se može smanjiti broj kombinacija baza i vrijeme za računanje koordinata bez smanjenja broja karakterističnih točaka. Možemo razložiti svaki model u podmodele. Za svaki podmodel uzimamo u obzir samo one baze, koje se nalaze na istom podmodelu i za njih računamo lokalne koordinate. Time je uvelike smanjen broj mogućih baza i vrijeme računanja lokanih koordinata.

Moguće je različite podmodele razvrstati u više od jedne hash tablice, čime se smanjuje lista parova baza u pretincima. Kod faze prepoznavanja za svaki par baze nije potrebno računati lokalne koordinate onih točaka koje su izvan određene granice, jer neće pripadati niti jednom podmodelu. Time je smanjen broj koordinata za računanje i vrijeme pristupa memoriji u fazi prepoznavanja.

2.3.2 Distribucija indeksa i rehashiranje

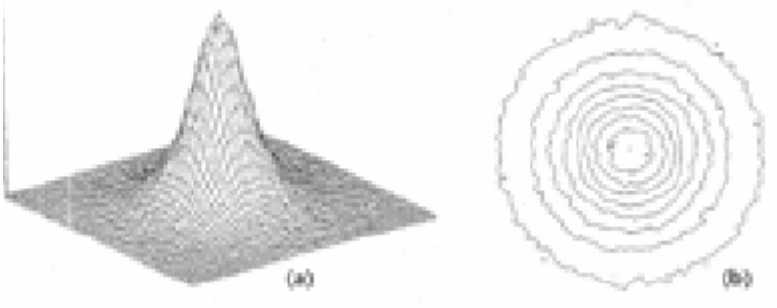
Postupkom kvantizacije smo dobili dvojku (u,v), koju koristimo kao indeks za pristup hash tablici. Pretpostavimo da su sve karakteristične točke identično i nezavisno distribuirane s poznatom funkcijom gustoće f slučajne varijable. Funkcija gustoće za poznate vrijednosti u i v se može izračunati integralom:

$$f(u,v) = \int_{R^4} f(x(u,v), y(u,v)) f(x_{\mu 1}, y_{\mu 1}) f(x_{\mu 2}, y_{\mu 2}) \left|\mathfrak{I}\right|^{-1} dx_{\mu 1} dx_{\mu 2} dy_{\mu 1} dy_{\mu 2} \quad (3)$$

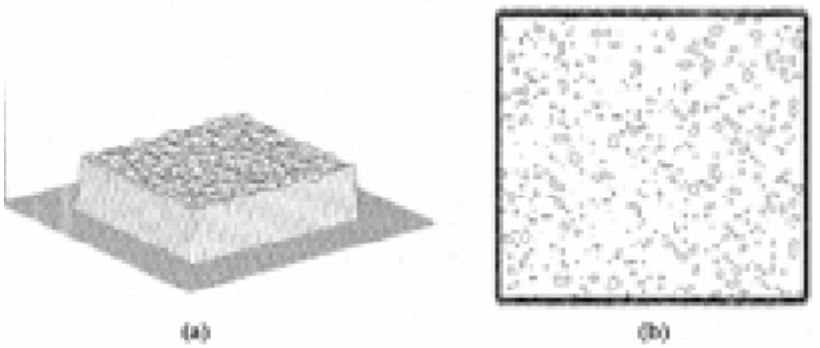
Na primjer Gaussovom normalnom razdiobom $N\left(0, \begin{pmatrix} \sigma & 0 \\ 0 & \sigma \end{pmatrix}\right)$ dobivamo da integral (3) iznosi $f(u,v) = \frac{12}{\pi} \frac{1}{\left(4\left(u^2 + v^2\right) + 3\right)}$.

Neuniformna razdioba zauzeća pretinaca u hash tablici može uzrokovati da neki pretinci sadrže velik broj zapisa. S obzirom da pretinci s najvećim brojem zapisa izravno utječu na vrijeme izvođenja kod postupka glasovanja, poželjno je uniformno distribuirati zapise u pretince. Potrebno je pronaći funkciju $h : R^2 \to R^2$ koja će

jednoliko rasporediti pretince u pravokutnoj hash tablici. Iz prethodnog primjera dobivamo: $h(u,v)=\left(1-\frac{3}{4(u^2+v^2)+3}, a\tan 2(v,u)\right)$. Pri čemu je kodomena funkcije atan2(v, u) interval $[-\pi,\pi]$. Rehashiranje je vrlo efikasno i njime se može smanjiti vrijeme izvođenja.



Slika 7. Distribucija indeksa Gaussovom normalnom razdiobom



Slika 8. Distribucija indeksa Gausovim procesom $N\left(0,\begin{pmatrix}\sigma & 0\\ 0 & \sigma\end{pmatrix}\right)$

2.3.3 Iskorištavanje simetričnosti

Za svaki zapis oblika $(\mathrm{mod}el,(\mu 1,\mu 2))$ na lokaciji s indeksom (u, v) u hash tablici postoji zapis oblika $(\mathrm{mod}el,(\mu 2,\mu 1))$ na lokaciji (-u, -v). Iskorištavanjem simetričnosti i uklanjanjem dodatnih zapisa dobivamo upola manju listu zapisa, što vodi dvostruko većoj brzini izvođenja programa.

2.3.4 Modeliranje smetnji

Do sada smo pretpostavljali da su karakteristične točke izdvojene u preprocesnoj fazi i fazi usporedbe bez mogućnosti greške. Zbog različitih smetnji, u praksi pogreške nisu rijetkost. Greške kod izdvajanja značajki sa slike rezultiraju krivim rezultatima. Potrebno je u hash tablicu ugraditi određeni prag tolerancije koji će za male pogreške kod unosa rezultirati točnim rezultatima. Greške kod unosa će uzrokovati pogrešni izračun indeksa u hash tablici. Uobičajeno je da su pogrešni pretinci susjedni s točnim pretincima kojima bi se pristupilo da nije nastupila greška. Moguće rješenje problema je da se pristupa pravokutnoj regiji tablice, umjesto samo jednom pretincu. Određivanje oblika i veličine regije je kompliciran postupak.

2.4 Usporedba s drugim metodama

Velika prednost geometrijsko hashiranja u odnosu na strukturalne usporedbe temeljene na analizi redoslijeda sekvenci je neosjetljivost na pogreške, umetanja ili brisanja slijeda. Algoritam se može koristiti za proteine, DNA, RNA molekule, sekundarne, tercijarne i druge strukture u analizi strukturalne kompatibilnosti.

Metoda poravnanja i generalizirana Houghova transformacija rade na sličnom principu kao metoda geometrijskog hashiranja, zato ih u ovom poglavlju uspoređujemo.

2.4.1 Metoda poravnanja

Metoda poravnanja koristi jednake geometrijske tehnike za određivanje mogućih kandidata za sličnost između slike i modela. Složenost metode poravnanja iznosi $O\left(Nn^k m^k t\right)$, gdje je N broj modela, O(t) složenost verifikacije pojedinog modela, m broj značajki modela, n broj izdvojenih značajki slike i k red veličine baze.

Za razliku od geometrijskog hashiranja ,koje koristi prethodno pripremljenu hash tablicu, metoda poravnanja vrši dugotrajno izračuna za sve moguće parove baza slike i modela. Velika prednost geometrijskog hashiranja je što može usporedno procesuirati sve modele, a metoda poravnanja obrađuje modele slijedno. Metoda poravnanja ne zahtjeva dodatne količine memorije, dok geometrijsko hashiranje koristi velike količine memorije za pohranjivanje hash tablica. Geometrijsko hashiranje će pokazati bolje karakteristike od metode poravnanja u slučajevima kada imamo veći broj modela i dovoljan broj jedinstvenih značajki slike za efikasno uspoređivanje glasovanjem.

2.4.2 Generalizirana Houghova transofrmaicja

Generaliziranom Hougovom transformacijom računaju se sve moguće kontinuirane transformacije između modela i slike i raspoređuje se u pretince. Također se koristi postupak glasovanja za one pretince koji su u skladu s izračunatim podacima. Za razliku od generalizirane Houghove transformacije geometrijsko hashiranje kvantizira samo diskretne transformacije definirane pomoću baze. Time je postupak računanja skraćen.

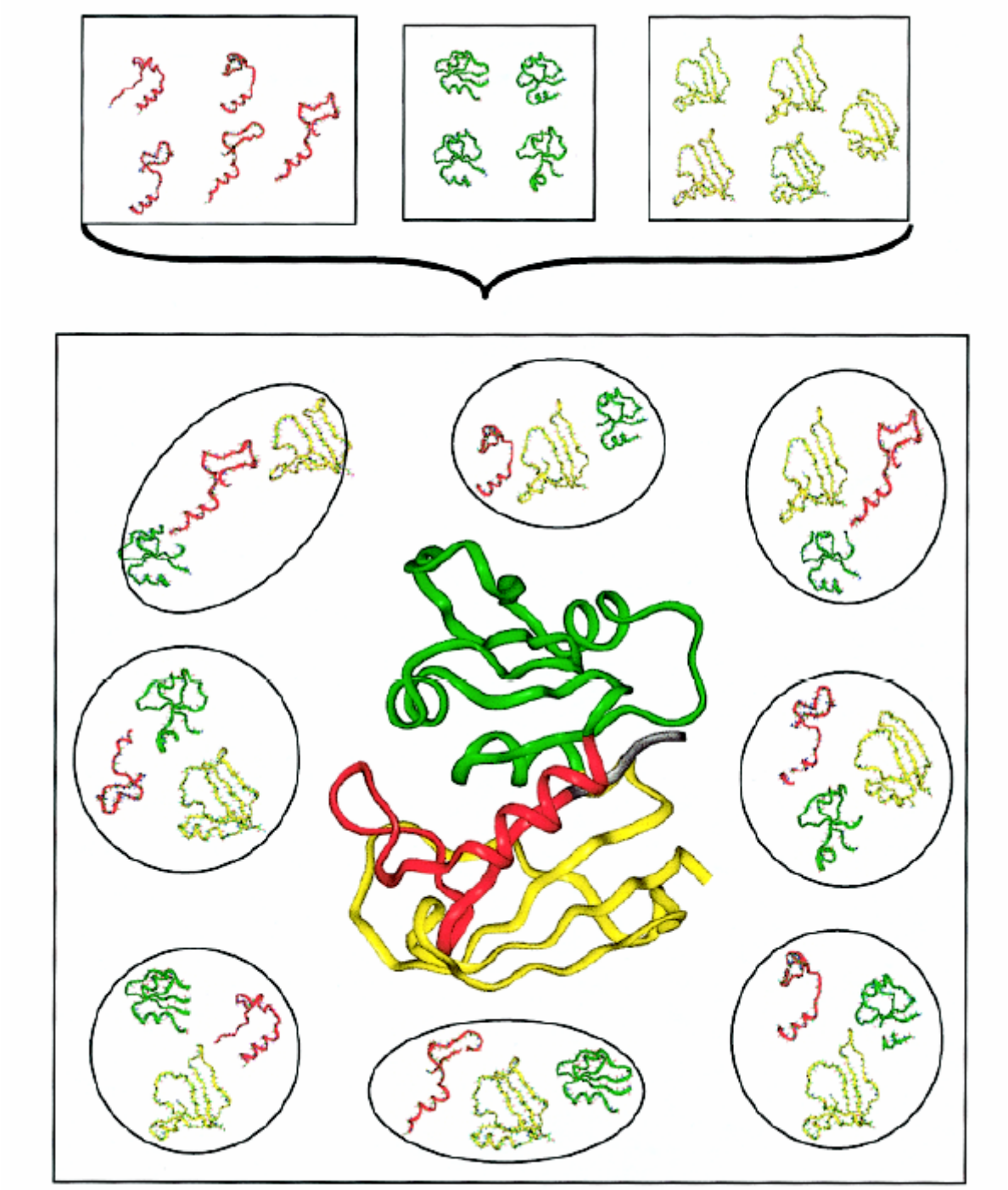
3. Zaključak

Proteini su odgovorni za gotovo sve metaboličke procese i regulacije u živim stanicama. Spajanjem s drugim strukturama i molekulama čine kompleksne strukture, koje mogu imati nove funkcije u organizmu. Predviđanjem protein - protein interakcija i kompleksnih proteinskih struktura možemo otkriti funkcije nepoznatih i nove funkcije poznatih proteina. Identifikacija proteinskih funkcija je okosnica za dizajniranje lijekova i predviđanje bolesti.

Bioinformatičari i biolozi predviđaju otkrića mnogih struktura proteina bez poznavanja njihove funkcije. U tom slučaju algoritam geometrijskog hashiranja bi mogao biti od pomoći. Ako se utvrdi sličnost između pojedinih proteina moguće je da smo pronašli neke njihove nove funkcije. Istovremeno je moguće da proteini potpuno različitog oblika izvršavaju iste funkcije.

Prethodno opisana metoda geometrijskog hashiranje je samo jedna od brojnih tehnika u području dokiranja proteina, konkretno površinskom analizom. Izdvojili smo samo neke postupke, kojima se nastoji poboljšati algoritam, no i dalje se radi na njegovom usavršavanju. Pri spajanju proteina oni mogu potpuno promijeniti svoj oblik. Algoritam geometrijskog hashiranja može podnijeti samo manja uobličenja, no znanstvenici rade na razvijanju metoda, koje bi davale ispravne rezultate za veća uobličenja. Geometrijskim hashiranjem se mogu dobiti proteini s neodgovarajućim kemijskim svojstvima, stoga se u algoritam nastoje implementirati postupci, koji bi uzeli u obzir kemijska svojstva proteina.

Istraživanja u području dokiranja proteina su posebice važna za dizajniranje lijekova. Cilj je poznavanjem strukture štetnog proteina (uzročnika bolesti) i metodom surface matching pronaći proteine, koji će se s njime reagirati. Ukoliko se otkriveni protein spoji s štetnim proteinom na određenim mjestima, tako da ga onemogući i poništi njegovo štetno djelovanje, možemo reći da smo pronašli lijek za tu bolest.



Slika 9. Dokiranje proteina

4. Literatura

1. Yehezkel Lamdan, Haim J. Wolfson: Geometric Hashing: A General And Efficient Model-Based Recognition Scheme, Dec 1988

2. James Bradford, Nicola D. Gold, Steven J. Pickering And David R. Westhead: Protein Surface Descriptions And Function, Dec 2005

3. *Shann-Ching Chen And Tsuhan Chen:* Protein Retrieval By Matching 3d Surfaces, Aug 2002

4. Ruth Nussinov, Haim J. Wolf Son: Efficient Detection Of Three-Dimensional Structural Motifs In Biological -Macromolecules By Computer Vision Techniques, July 1991

5. Ling Ma K. Masuda, J. Sugano, A. Yamaguchi: A New Object Detection Technique Based On Geometric Hashing, 2005

6. Haim J. Wolfson: Geometric Hashing: An Overview, Oct.-Dec. 1997

7. Yusu Wang: Geometric And Topological Methods In Protein Structure Analysis, 2004

8. Pdb Annual Report July 2005- June 2006

5. Sažetak

Danas aktualno područje istraživanja bioinformatike je dokiranje proteina. Surface matching metoda posebim postupcima nastoji otkriti da li će neka dva proteina međusobno djelovati i na kojim mjestima će se se spojiti. Vrlo efikasna tehnika za otkrivanje spojnih mjesta na površini proteina je geometrijsko hashiranje.

Geometrijsko hashiranje je metoda prvotno osmišljena za kompjuterski vid, no našla je primjenu u mnogim područjima. Ona nudi mogućnost raspoznavanja u slučajevima kada je promatrani objekt djelomično zaklonjen ili je slika podlegla transformacijama. U nekim slučajevima raspoznavanje je moguće i kod nedostatka dijela informacija. Aloritam se sastoji od dvije osnovne faze: preprocesna faza i faza usporedbe. Predprocesnoj faza se obuhvaća kreaciju hash tablice. U fazi usporedbe se na temelju hash tablice i slike nastoji raspoznati sličnosti ili razlke.

Primjena geometrijskog hasiranja u protenomici zahtjeva implemetaciju nekih dodatnih mogućnosti, zbog velikog broja proteina, kompleknosti strukture, te bioloških i kemijskih svojstava.

Moguća su dodatna poboljšanja algoritma različitim postupcima, poput povećanja broja hash tablica, podijele modela na podmodele, rehashiranje, modeliranje smetnji itd. Metoda geometrijskog hashiranja se u odnosu na druge metode pokazala vrlo efikasnom i široko primjenjivom.

Istraživanja u području dokiranja proteina su temelj dizajniranja lijekova i predviđanja ponašanja bolesti.