SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

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

Ivan Jakus

Voditelj: *Mr.sc. Krešimir Šikić*

Zagreb, svibanj, 2007

Sadržaj

Uvod

Proteini, iako mali veličinom imaju važnu ulogu. Sastavni su dijelovi svake stanice koja čini osnovu života na Zemlji. Sudjeluju u svim funkcijama unutar stanice. Glavni su izvor tvari za izgradnju svih tjelesnih tkiva u ljudskom organizmu.

Krajem prošlog stoljeća paralelno s ubrzanim razvojem računala i računarske znanosti razvijala se još jedna disciplina, poveznica biologije i informatike, bioinformatika. Bioinformatiku je moguće podijeliti u dva veća područja, genomiku i proteomiku. Genomika se bavi proučavanjem gena i genoma, dok se proteomika bavi proteinima.

Kao sto je već navedeno proteini su sastavni dio svake stanice, te je istraživanje proteina, njihovog nastajanja, funkcija i interakcija, kako međusobnog tako i sa drugim molekulama, od velike važnosti, pogotovo u području razvoja lijekova i cjepiva, analizi metaboličkih reakcija i praćenju razvoja organizama kao i u mnogim drugim područjima.

Promatranje proteina i njihovih interakcija iznimno je teško pa je stoga broj eksperimentalno pronađenih proteina i proteinskih kompleksa relativno mali. Kako bi se istraživanje proteina ubrzalo u posljednje se vrijeme sve više koriste računala i računalni modeli pri utvrđivanju izgleda proteina te utvrđivanja koji proteini i na koji način vrše interakciju.Ujedno se primjenom računalnih modela povećala raspoloživa količina podataka o proteinskim sekvencama i mogućim proteinskim interakcijama.

Proteini

Aminokiseline su kemijski spojevi koji svojim povezivanjem u linearne polimere stvaraju proteine. Svaka aminokiselina sastoji se od četiri dijela: središnje postavljen (alfa) atom ugljika, peptidna skupina (NH2), karboksilna skupina (COOH) i kiselinski ostatak. U svijetu je poznato 20 različitih aminokiselina koje se međusobno razlikuju samo po kiselinskom ostatku.

Aminokiseline se u polipeptidne lance povezuju stvaranjem peptidne veze između peptidne skupine jedne aminokiseline i karboksilne skupine druge aminokiseline. Na taj se način u polipeptidne lance može povezati proizvoljno velik broj aminokiselina.
Protein se može sastojati od samo jednog polipeptidnog lanca i tada govorimo o monomeru ili može nastati povezivanjem vise polipeptidnih lanaca te tada govorimo o multimeru (polimeru).
Veličina nastalih proteina mjeri se brojem aminokiselina koje sadrži i danas su poznati proteini sa oko 460 aminokiselina, pa sve do proteina sa gotovo 27 000 aminokiselina.

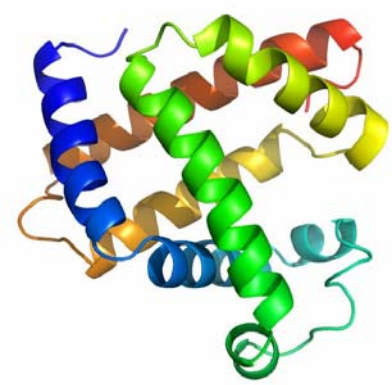
Kod proteina i njegove funkcije nije samo važna njegova građa odnosno aminokiseline od kojih se sastoji nego i način na koji će se proteinski lanac svinuti (svrtanje proteina).
Struktura proteina se najčešće promatra u četiri strukture:

1. Primarna stuktura – poredak aminokiselina u proteinu

2. Sekundarna struktura – uzima u obzir svijanje prteinskog lanca pod utjecajem vodikovih veza koje se stvaraju zbog blizine kiselinskih ostataka različitih susjednih aminokiselina unutar proteina

3. Tercijarna struktura – razmatra utjecaj međusobno udaljenih aminokiselina. Pojam svrtanje najčešće se povezuje sa ovom strukturom proteina

4. Kvartarna struktura – kvartarna struktura dodatno razmatra i utjecaj interakcije između dva ili više polipeptidnih lanaca unutar proteina



Slika 1. Prikaz strukture proteina mioglobina

Dokiranje proteina (pristupi)

Cilj proučavanja metoda dokiranja je pokušaj da se predvidi struktura proteinskih kompleksa ako je poznata struktura proteina koji su sastavni dio novog kompleksa. U najopćenitijem slučaju, metode dokiranja otkrivaju mjesto na kojem dolazi do spajanja i predviđaju oblike i položaje koje će ti proteini zauzeti u novonastalom proteinskom kompleksu.

U posljednjih 40 godina predloženi su mnogi pristupi rješavanju tog problema no većina ih se može svrstati u dvije glavne podskupine. Ove dvije skupine zapravo su jako slične jer su oba pristupa zasnovana na termodinamici intermolekularnih interakcija.
Prva skupina metoda koristi izravni termodinamički pristup u kojem se slobodna energija kompleksa, opisana pomoću raznih aproksimacija entalpije i entropije, pokušava smanjiti.

Druga skupina, koja je nama zanimljivija, koristi fenomenološke podatke kao što su geometrijska i kemijska podudarnost u proteinskim kompleksima. Algoritmi koji propadaju ovoj kupini molekule promatraju kao čvrsta tijela te samim time problem dokiranja svode na šest-dimenzijonalnu pretragu kroz rotacijsko-translacijski prostor. Iako se konformacije dokiranih proteina ne mijenjaju određena geometrijska nepodudaranja se toleriraju.

Iskorištavanje trodimenzionalnog prikaza molekula u području dokiranja proteina predstavili su Jiang i Kim te u približno isto vrijeme neovisno o njima i Katchalski-Katzir i suradnici. Njihovi algoritmi iako slični razlikuju se u određenim detaljima.
Jaing i Kim kombiniraju dva prikaza molekula, površinskih točaka, njima pripadajućih normala i volumena, te površinskih kocaka koje sadrže po dvije do tri točke.
Povezanost između dvije površine molekula na svakoj poziciji određuje se pomoću broja točaka koje se podudaraju, uz uvjet da se površinske kocke koje sadrže te točke preklapaju i da normale pokazuju u suprotnim smjerovima.
Katchalski-Katzir koristi jednostavniji pristup. Prvo on koristi samo jedan prikaz molekule; one se digitaliziraju u trodimenzionalne mreže, te se površina i unutrašnjost molekule razlikuju pomoću digitalnog procesa koji ne zahtjeva računanje površinskih točaka.
Za svaki položaj molekula funkcija koja povezuje međusobni položaj molekula računa se pomoću FFT (*Fast Fourier Transformations*) metode, te samim time implicitno pretražuje sve relativne translacije.
Jednostavan i izravan pristup ove metode prepoznat je od strane drugih znanstvenika te su mnoge istraživačke grupe prihvatile i modificirale ovu metodu u svojim istraživanjima.

3D strukture proteinskih kompleksa otkrivaju usku geometrijsku i kemijsku povezanost onih dijelova površine molekula koje su u kontaktu. Zbog toga oblik i ostalu fizičke karakteristike površine uvelike određuju prirodu pojedinih interakcija.

Opis algoritma

Prvi korak u algoritmu je stvaranje mrežnog prikaza proteinske molekule i liganda na temelju njihovih atomskih koordinata. Te dvije molekule označe se sa **a** i **b.** Molekule se projiciraju na trodimenzionalnu mrežu od N × N × N točaka gdje su one prikazane pomoću diskretnih funkcija

$$a_{l,m,n} = \begin{cases} 1 & \text{unutar molekule} \\ 0 & \text{izvan molekule} \end{cases} \qquad (1a)$$

$$b_{l,m,n} = \begin{cases} 1 & \text{unutar molekule} \\ 0 & \text{izvan molekule} \end{cases} \qquad (1b)$$

gdje l, m i n predstavljaju indekse 3D mreže dimenzija N×N×N; l,m,n = (1, ... , N).
Svaka točka mreže smatra se da je dio molekule ako je barem jedna atomska jezgra unutar udaljenosti **r** od molekule, gdje je **r** van der Waalsov radijus.

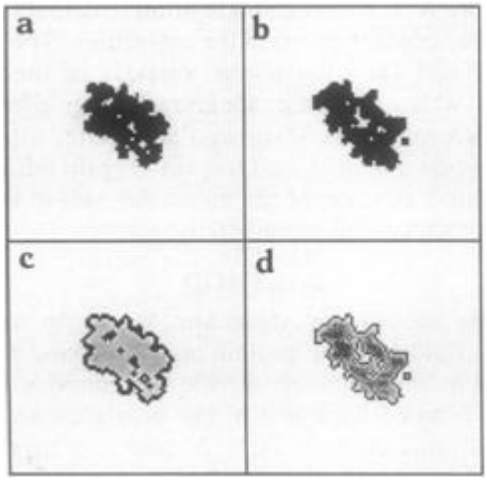
Potom je potrebno napraviti razliku između površine i unutrašnjosti svake molekule. To postižemo tako da uvodimo novu vrijednost koja predstavlja unutrašnjost molekule dok vrijednost jedan predstavlja tanak sloj uz površinu molekule.
Diskretne funkcije tako prelaze u

$$\bar{a}_{l,m,n} = \begin{cases} 1 & \text{na površini molekule} \\ \rho & \text{unutar molekule} \\ 0 & \text{izvan molekule} \end{cases} \qquad (2a)$$

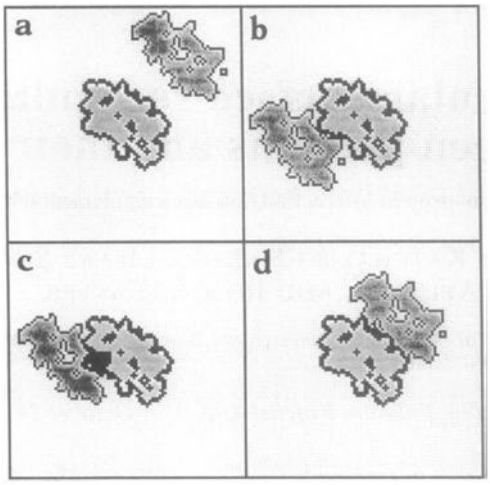
$$\bar{b}_{l,m,n} = \begin{cases} 1 & \text{na površini molekule} \\ \delta & \text{unutar molekule} \\ 0 & \text{izvan molekule} \end{cases} \qquad (2b)$$

gdje se površina definira kao granični sloj konačne širine između unutrašnjeg dijela i vanjskog dijela molekule.

Parametri opisuju vrijednost točaka unutar molekule dok je svim točkama izvan molekule dodijeljena vrijednost 0.



Slika 2. Različiti primjeri presjeka trodimentzionalnog prikaza proteina



Slika 3. Prikaz zrazličitih relativnih pozicija molekula **a** i **b**

Uspoređivanje površina postiže se računanjem korealcijske funkcije. Korelacijska funkcija između diskretnih funkcija $\bar{a}$ i $\bar{b}$, definirana je

$$\bar{c}_{\alpha,\beta,\gamma} = \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} \bar{a}_{l,m,n} \cdot \bar{b}_{l+\alpha,m+\beta,n+\gamma} \qquad (3)$$

gdje α, β, γ predstavljaju broj koraka na mreži za koje se molekula **b** pomakne u odnosu na molekulu **a** u bilo kojem smjeru. Ako je vektor pomaka (α, β, γ) takav da nema kontakta između dviju molekula vrijednost korelacije je 0. Ako postoji kontakt između površina prinos korelaciji je pozitivan.
Kako ulazak jedne molekule unutra druge nije moguć potrebno je nekako postići jedinstvenu razliku između kontakta među površinama i ulaska jedne molekule u drugu. To ograničenje se postiže dodjeljivanjem velikih negativnih vrijednosti **ρ** u $\bar{a}$ i malih pozitivnih vrijednosti varijabli **δ** u $\bar{b}$. Zahvaljujući ovom ograničenju kada vektor pomaka (α, β, γ) postane takav da molekula **a** uđe u molekulu **b** množenje negativnih vrijednosti iz $\bar{a}$ i pozitivnih vrijednosti iz $\bar{b}$ za posljedicu ima to da je ukupni prinos u korelacijsku sumu negativan.
Pozitivne korelacijske vrijednosti postižu se kada prinos površinskog kontakta nadvlada negativan prinos ulaska jedne molekule u drugu.
Izravno računanje međudjelovanja dviju molekula jako je dug proces jer zahtjeva N^3 množenja i zbrajanja za svaki od mogućih N^3 relativnih pomaka (α, β, γ) što u konačnici iznosi N^6 koraka za računanje. Zbog toga se koristi Fourierova transformacija koja omogućava računanje korelacijske funkcije mnogo brže.
Diskretna Fourierova transformacija funkcije $x_{l,m,n}$ definira se

$$X_{o,p,q} = \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} \exp\left[-2\pi i\left(ol + pm + qn\right)/N\right] \cdot x_{l,m,n} \qquad (4)$$

Gdje su o, p, q = { 1 ... N } i gdje je $i = \sqrt{-1}$.
Primjenom ove jednadžbe na obje strane jednadžbe (3) dobijamo

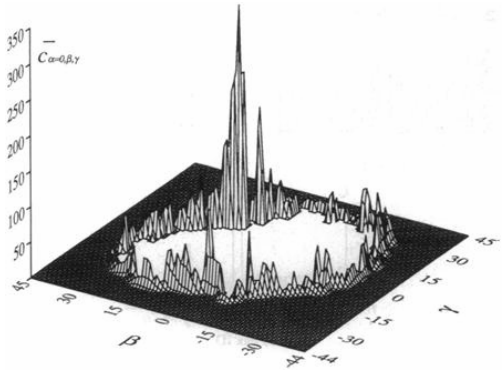
$$C_{o,p,q} = A^{*}_{o,p,q} \cdot B_{o,p,q} \qquad (5)$$

Gdje su C i B diskretne Fourierove transformacije funkcija $\bar{c}$ i $\bar{b}$, a A^{*} kompleksno konjugirana diskretna Forierova transformacija od $\bar{a}$.
Inverzna Fourierova transformacija(IFT) definirana je kao

$$\bar{c}_{\alpha,\beta,\gamma} = \frac{1}{N^3}\sum_{o=1}^{N}\sum_{p=1}^{N}\sum_{q=1}^{N} \exp\left[2\pi i\left(o\alpha + p\beta + q\gamma\right)/N\right] \cdot C_{o,p,q} \qquad (6)$$

i koristi se kako bi se postigla željena korelacija između dviju originalnih $\bar{a}$ i $\bar{b}$ funkcija.
Fourierove transformacije moguće je izvoditi pomoću brzog Fourierovog transformacijskog algoritma (*fast Fourier transform algorithm*) koji zahtjeva manje od $N^3\ln(N^3)$ koraka za računanje transformacije 3D funkcije od N × N × N vrijednosti.

Kako bi se napravila potpuna provjera poklapanja površina dviju molekula **a** i **b** potrebno je izračunati vrijednost korelacijske funkcije $\bar{c}$ za svaku moguću orijentaciju molekula. U praksi se najčešće fiksira molekula a dok se tri Eulerova kuta koja određuju orijentaciju molekule **b** mijenjaju unutar fiksnog intervala od Δ stupnjeva.
Riješenje analize prikazuje se u obliku grafa gdje svaki visoki i oštri vrh predstavlja potencijalno mjesto geometrijskog poklapanja i time predstavlja mogući spoj molekula. Relativna pozicija i orijentacija molekula može se odredit na osnovu koordinata vrha i iz tri Eulerova kuta na kojima je vrh pronađen.



Slika 4. Grafički prikaz analize međusobnih položaja proteina

Implementacija algoritma

Kako bi se algoritam uspješno implementirao potrebno je prvo odrediti neke posebne vrijednosti kao što su debljina površinskog sloja molekule, r, ρ, δ, η i još neke. Prilikom odabiranja ovih vrijednosti u obzir se uzima više parametara kako bi se vrijednosti mogle što bolje odrediti, te samim time dobiti algoritam sa sto većom učinkovitošću.

Može se primijetit kako poklapanje $\bar{a}$ i $\bar{b}$ funkcija nije savršeno. Prilikom spajanja dviju molekula između njih ostaju male praznine koje imaju utjecaj na njihov matematički prikaz . Također, prilikom računanja funkcija $\bar{a}$ i $\bar{b}$ zanemaruju se vodikovi atomi koji također imaju utjecaja na molekule. Na posljetku, prilikom spajanja molekula dolazi do malih promjena na njihovim površinama. Sve te promjene u obliku molekula nisu uključene u prikaz funkcija $\bar{a}$ i $\bar{b}$. Kako bi se osiguralo da točan rezultat ne bude zanemaren zbog ignoriranja ovih malih promjena na molekulama potrebno je izvršiti neke promjene. Dodjeljuje se više od jednog sloja točaka na površinu u $\bar{a}$ tako da je debljina površine molekule **a** 1.5 – 2.5 Å. Zbog ove izmjene u funkciji **a** toleriraju se udubljena i praznine manje od ovih vrijednosti. Također je važno napomenuti da povećanjem debljine površine molekule **a** možemo očekivat i povećanje broja pogrešnih rješenja.

Debljina površinskog sloja ima utjecaj na kutnu toleranciju koja je definirana kao maksimalna devijacija od točne orijentacije molekula koja bi rezultirala jedinstvenim korelacijskim vrškom na konačnom grafu. Debljina površinskog sloja od 2 Å uzrokuje kutnu toleranciju od +/- 10°. Zbog toga se za pomak kuta uzima vrijednost $\Delta = 20°$.

Za parametar r se uzima vrijednost 1.8 Å sto je za 0.2 Å više od prosječnog van der Waalsovog radijusa za ugljik te se time nadoknađuje činjenica da vodikovi atomi nisu uzimani u obzir prilikom projiciranja molekule.

Parametri ρ i δ se postavljaju u vrijednosti -15 i 1 te time značajno smanjuju korelacijsku vrijednost ako dođe do proboja jedne molekule u drugu.

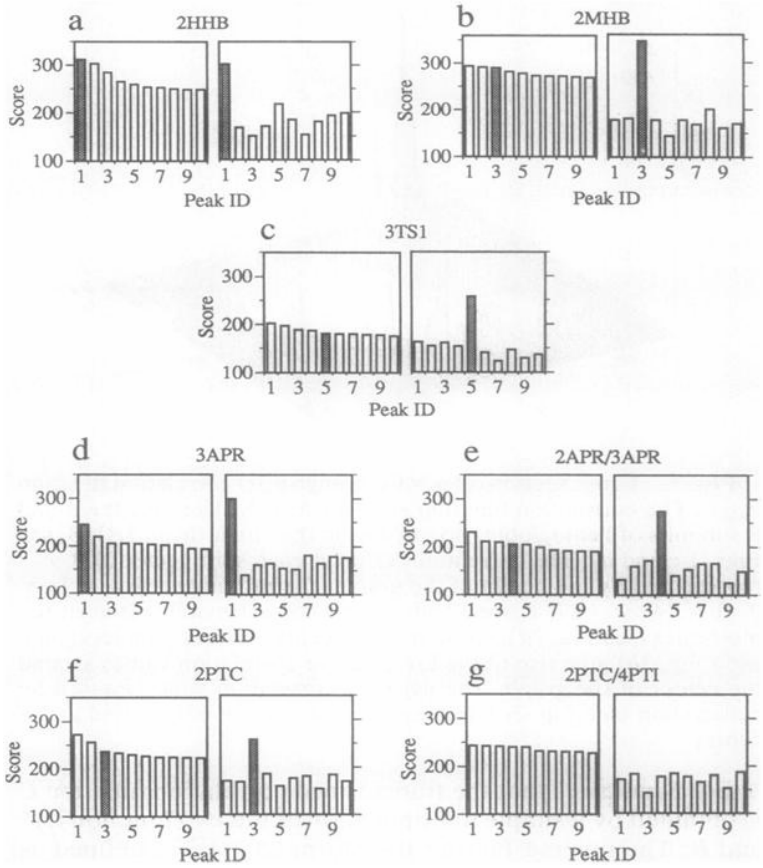
Još jedan važan parametar algoritma je η koji predstavlja veličinu koraka na mreži. Optimalni rezultati postizani su kada je iznos parametra η bio između 0.7 i 0.8 Å što odgovara polovici dužine veze između dvije molekule ugljika. Uz te parametre proces računanja mogućih rješenja trajao bi dugo, te je stoga sam proces pretraživanja podijeljen u dva dijela. Prvi dio u kojem se traže potencijalna rješenja i drugi dio u kojem se ponovno reevaluiraju rješenja koja su dobila najveću ocjenu u prvom prolazu. Prilikom prvog prolaza moguće je da rješenje koje je u konačnici točno ne dobije najveći rezultat te je zbog toga važno imati drugi prolaz koji će izdvojiti točno rješenje.
Za prvi prolazak se veličina η postavlja između 1.0 i 1.2 Å dok se za drugi prolazak koriste prije navedene vrijednosti.

Rezultati

Kako bi se testirala uspješnost algoritma prilikom predviđanja strukture proteinskih kompleksa korišteno je nekoliko kompleksa čija je struktura poznata. Kako bi se algoritam što bolje testirao korišteni su kompleksi čije veličine variraju od 30 do 2500 atoma.

Korištene su dvije verzije hemoglobina, ljudski deoksihemoglobin (engl. *deoxyhemoglobin*)(2HHB) i konjski methemoglobin (engl. *methemoglobin*)(2MHB), koji predstavljaju heterodimere. Korišteni su i tirozin sintetaza (engl. *tRNA synthetase-tyrosinyl adenylate*)(3TS1), aspartat proteinaza inhibitor (engl. *Aspartic proteinase-peptide inhibitor*)(3APR), tripsin-tripsin inhibitor (engl. *trypsin-trypsin inhibitor*)(2PTC). Dodatni testovi su provedeni sa aspartat proteinazi (engl. *native aspartic proteinase*) i njenim peptidnim inhibitorom (2APR), te tripsinom (engl. *trypsin*) i njegovim inhibitorom (4PTI).

Uspješnost algoritma određivala se na osnovu dobivene relativne pozicije molekula, koja je postigla najvišu ocjenu, korištenjem algoritma i naknadnom usporedbom s poznatim izgledom proteinskog kompleksa. Rezultati provedenih testova prikazani su na slici 5. Histogrami pokazuju 10 korelacijskuh vrhova za svaki par molekula. Lijevi histogram pokazuje rezultate dobivene postupkom pretraživanja dok desni prikazuje rezulate dobivene reevaluacijom.



Slika 5. Histogrami koji prikazuju rezulate algoritma za različite parove proteina

Kao što se iz slika može vidjeti točan korelacijski vrh (tamnije obojan) za kompleks ne mora biti najveći u prvom prolasku algoritma, ali zato u drugom, reevaluacijskom, dijelu postiže znatno bolju ocjenu od ostalih potencijalnih pozicija molekula u kompleksu.

Analiza kompleksa 2PTC izabrana je zbog toga sto je prirodna sturktura molekule inhibitora različita od onog oblika koji ona ima u kompleksu. Kada se u proračunu koristila struktura inhibitora koju on ima u kompleksu najviši vrh nakon koraka reevaluacije je odgovarao stvarnoj poziciji inhibitora u kompleksu, no kada se za proračun koristio prirodni oblik inhibitora algoritam nije izdvojio niti jedan rezultat kao ispravan, niti u koraku pretraživanja niti u koraku reevaluacije. Razlog zbog kojeg nije dobiveno rješenje u ovom slučaju je taj sto su promjene koje su nastale na strukturi inhibitora prilikom stvaranja kompleksa veće od onih koje koje algoritam tolerira.

Zaključak

Algoritam koji su predstavili Katchalski-Katzir i suradnici bio je među prvima koji su se bavili problemom dokiranja proteina. Oni su prvi iznijeli ideje koje će kasnije postati temelj mnogih istraživanja na području dokiranja proteina. Mnogi algoritmi koji su kasnije predstavljeni bili su zapravo nadogradnje ovog algoritma.

Najveći nedostatci ovog algoritma su njegova brzina i nepreciznost kada se radi sa proteinima koji prilikom stvaranja kompleksa doživljavaju veće promjene u obliku. Mnoge su ideje kako poboljšati ovaj algoritam. Metode koje su se koristile kako bi se poboljšala uspješnost algoritma su uvođenje dodatnih testova koji se oslanjaju na elektrostatsku komplementarnost proteina ili uvođenje računanja komplementarnosti hidrofobnih dijelova proteina. Neki od dodataka algoritmu pokazali su uspješnima dok je bilo i pokušaja koji su rezultirali lošijim rezultatima od originalnog algoritma.

Razvijanje algoritama za dokiranje proteina zadnjih je godina jako brzo napredovalo, no pokazalo se da je određivanje geometrijske sličnosti među proteinima temeljni korak u analizi proteinskih kompleksa. Postoji još dosta mjesta za razvijanje novih metoda za analizu dokiranja, ali također i za unaprjeđenje već postojećih. To se pogotovo odnosi na poboljšanje u proučavanju promjena oblika proteina prilikom stvaranja kompleksa.

Bioinformatika je još relativno mlada grana znanosti tako da se nove ideje i metode objavljuju svakim danom i stoga se može očekivati veliki napredak u svim područjima njenog istraživanja pa samim time i u dokiranju proteina.

Popis literature

1. Protein, 22. 04. 2007., *Protein*, http://en.wikipedia.org/wiki/Proteins, 02.05.2007
2. Protein structure, 22. 04. 2007., *Protein structure*, http://en.wikipedia.org/wiki/Protein_structure , 03.05.2007
3. Protein folding, 20. 04. 2007., *Protein folding*, http://en.wikipedia.org/wiki/Protein_folding , 02.05.2007
4. Katchalski-Katzir E., Shariv I., Eisenstein M., Friesem A.A., Aflalo C., Vakser I.A., Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques, Proc. Natl. Acad. Sci. USA 89 (1992), 2195-2199
5. Eisenstein M., Katchalski-Katzir E., On proteins, grids, correlations, and docking, C. R. Biologies 327 (2004), 409–420
6. Vajda S., Camacho C.J., Protein–protein docking: is the glass half-full or half-empty?, TRENDS in Biotechnology Vol.22 No.3 March 2004, 110-115
7. Halperin I., Ma B., Wolfson H., Nussinov R., Principles of Docking: An Overviewof Search Algorithms and a Guide to Scoring Functions**,** PROTEINS: Structure, Function, and Genetics 47 (2002), 409–443
8. Gabb H.A., Jackson R.M., Sternberg M.J.E., Modelling Protein Docking using Shape Complementarity, Electrostatics and Biochemical Information, J. Mol. Biol. (1997) 272, 106-120
9. Vaidyanathan P.P., Genomics and Proteomics: A Signal Processor's Tour, IEEE circuits and systems magazine, 1531-6364/04/2004, 6-29

Sažetak

Proteini su sastavni dio svake stanice. Zbog toga se u posljednje vrijeme sve veća pažnja posvećuje njihovom proučavanju; njihovom obliku i načinu na koji međusobno reagiraju te obliku koji pritom nastaje.

Ovdje opisani algoritam za dokiranje proteina se zasniva na digitaliziraju proteina u trodimenzionalne mreže te analiziranja međusobnog položaja dviju molekula pomoću korelacijske funkcije. Dodatna ubrzanja algoritma postižu se korištenjem Fourierovih transformacija prilikom računanja korelacijske funkcije.

Algoritam se pokazao učinkovitim kod kompleksa u kojima proteini koji ih stvaraju ne doživljavaju veće promjene u obliku izvan kompleksa i u kompleksu.

Određene promjene algoritma su moguće kako bi se povećala njegova učinkovitost.

.