ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2006
Seria: AUTOMATYKA z. 145 Nr kol. 1728

Marcin PACHOLCZYK, Damian BERESKA
Politechnika Śląska

ALGORYTM HASZOWANIA GEOMETRYCZNEGO W DOKOWANIU MOLEKULARNYM

Streszczenie. W pracy przedstawiono implementację oraz wstępne wyniki działania zmodyfikowanego algorytmu wizji komputerowej - haszowania geometrycznego zastosowanego w problemie dokowania molekularnego. Dokowanie molekularne w ujęciu przedstawionym w pracy jest problemem znalezienia najlepszego dopasowania struktury przestrzennej białka oraz mniejszej molekuły – ligandu (molekuły lekopodobnej). Jako model interakcji białko ligand zastosowano powierzchnie interakcji – stanowiących geometryczną reprezentację zbioru reguł rządzących oddziaływaniami na poziomie molekularnym (wiązania wodorowe, oddziaływania hydrofobowe). Właściwości algorytmu zostały sprawdzone podczas próby rekonstrukcji naturalnego dopasowania struktury izomerazy oraz ligandu SO_4, pochodzących z kompleksu o oznaczeniu 5TIM (PDB).

THE GEOMETRIC HASHING ALGORITHM IN MOLECULAR DOCKING

Summary. The paper presents an implementation and preliminary results obtained with modified computer vision algorithm called geometric hashing applied to the problem of molecular docking. Molecular docking as presented here is the problem of finding the best possible matching between protein structure and a ligand (typically a smaller, drug like molecule) . As a model for protein – ligand interaction we use the interaction surfaces – geometric representation of rules governing intramolecular interactions (hydrogen bonds, hydrofobic interactions). We tested our method trying to reconstruct native binding pose of the SO_4 ligand and isomerase in 5TIM (PDB) complex.

1. Wprowadzenie

Technika zwana dokowaniem molekularnym może być rozpatrywana jako potencjalna metoda komputerowo wspomaganego projektowania i optymalizacji nowych leków. Problem dokowania molekularnego sprowadza się zwykle do odnalezienia właściwego ligandu (w naszym przypadku molekuły lekopodobnej), dopasowanego (pod względem zarówno geometrycznym, jak i energetycznym) do pewnego szczególnego miejsca w strukturze białka, zwanego miejscem wiązania lub

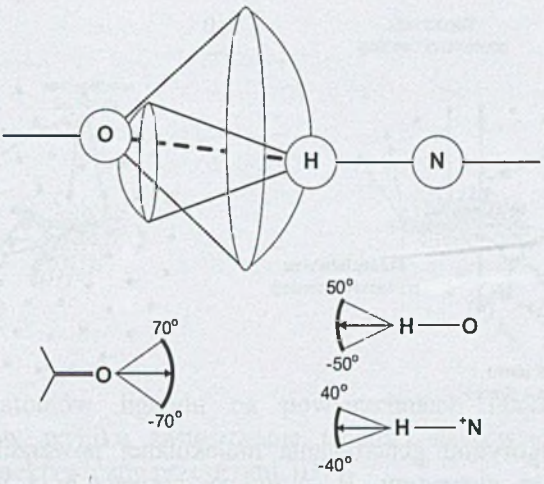
miejscem aktywnym. Takie dopasowanie powinno wywołać określoną reakcję biochemiczną, taką jak np. inhibicja funkcji białka wirusowego. Rozwiązania problemu dokowania molekularnego dokonuje się zazwyczaj stosując algorytmy poszukujące najlepszego dopasowania geometrycznego ligandu i białka, uzupełnione o algorytmy minimalizacji energii takiego dopasowania. W bieżącej pracy rozważa się rozwiązanie geometrycznego aspektu problemu dokowania molekularnego oparte na zmodyfikowanym, wywodzącym się z wizji komputerowej, algorytmie haszowania geometrycznego.

W celu uwzględnienia biochemicznej charakterystyki interakcji pomiędzy ligandem i białkiem, użyto powierzchni interakcji wprowadzonych przez H.-J. Bohma [1] i rozwiniętych przez M. Rareya i wsp. [2]. Powierzchnie interakcji to zbiór reguł biochemicznych rządzących podstawowymi typami oddziaływań na poziomie molekularnym (takich jak wiązania wodorowe czy oddziaływania hydrofobowe) zapisanych w postaci łatwych w implementacji ograniczeń geometrycznych. W zaprezentowanym podejściu przyjęto następujące założenia upraszczające: ligand jest małą molekułą nie posiadającą wewnętrznych stopni swobody, białko zaś jest nieruchomą bryłą sztywną. Trójwymiarowe struktury (współrzędne kartezjańskie lokalizacji poszczególnych atomów) ligandu i białka są uważane za znane i zostały pobrane z ogólnodostępnej bazy struktur białkowych The Protein DataBank (PDB)[3]. W celu identyfikacji miejsca wiązania oraz typów interakcji pomiędzy ligandem i białkiem, używanych jako punkt odniesienia, posłużono się programem LIGPLOT[4].

Jako rozwiązanie problemu poszukiwania najlepszego, pod względem geometrycznego dopasowania, położenia ligandu w miejscu wiązania białka, proponuje się wykorzystanie zmodyfikowanego algorytmu haszowania geometrycznego, którego oryginał został opisany w [5]. Zastosowanie haszowania geometrycznego w dokowaniu molekularnym zostało w literaturze zaproponowane już wcześniej [6], jednakże bez zastosowania jakiegokolwiek modelu oddziaływań białko-ligand. Zarówno białko, jak i ligand zostały potraktowane jako dowolne ciała sztywne.

2. Model oddziaływania białko-ligand

Przyjęty model zakłada kilka możliwych typów interakcji, takich jak wiązania wodorowe czy oddziaływania hydrofobowe. Sama interakcja modelowana jest za pomocą centrum interakcji i powierzchni interakcji umiejscowionej na sferze o środku w centrum interakcji. Przyjmuje się, że interakcja zachodzi w przypadku, gdy centrum interakcji ligandu leży na powierzchni interakcji receptora i odwrotnie. Powierzchnia interakcji modelowana jest za pomocą dyskretnego zbioru punktów. Tak scharakteryzowany model ma istotną zaletę, gdyż akumuluje w łatwej do zapisania i późniejszej obróbki, postaci geometrycznej, szereg reguł z dziedziny chemii organicznej, rządzących powstawaniem wiązań atomowych i opisujących oddziaływania międzycząsteczkowe. Rysunek 1 przedstawia kilka przykładów molekularnych powierzchni interakcji. Dane dotyczące powierzchni interakcji dla innych konfiguracji atomów czytelnik odnajdzie w pracach [1,2].



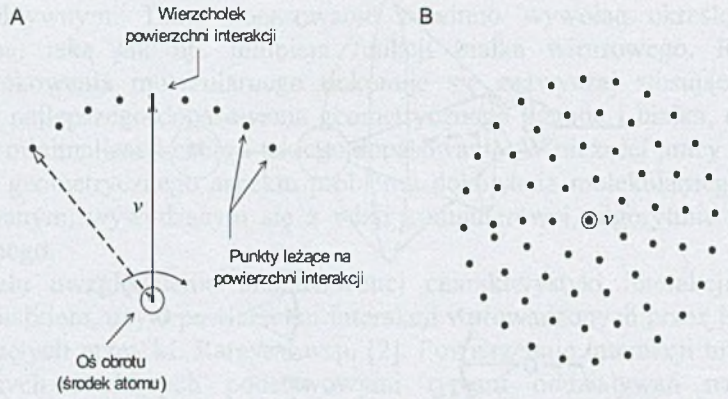
Rys. 1. Idea modelowania interakcji molekularnych za pomocą powierzchni interakcji. Przykłady powierzchni interakcji

3. Algorytm generacji molekularnej powierzchni interakcji

Pierwszy krok algorytmu stanowi odpowiednie rozmieszczenie atomów wodoru. Następnie, określana jest geometria powierzchni interakcji dla poszczególnych atomów (w szczególności dla atomów wchodzących w skład miejsca aktywnego). Rodzaj geometrii powierzchni interakcji [1, 2] zdeterminowany jest przez rodzaj atomu oraz typ konfiguracji, w jakiej dany atom występuje. Dyskretne punkty składające się na powierzchnię interakcji generowane są w dwóch krokach z zastosowaniem kwaternionowego opisu rotacji w przestrzeni kartezjańskiej xyz. W pierwszym kroku wyliczane są współrzędne punktów leżących na płaszczyźnie wektora $\vec{v}$ łączącego środek atomu z wierzchołkiem powierzchni interakcji (rys. 2A). Obrotu dokonuje się ze stałym krokiem, który jest parametrem algorytmu, wokół wektora prostopadłego do wspomnianego wcześniej wektora $\vec{v}$, o wartości kątów z zakresu charakterystycznego dla konfiguracji, w której występuje dany atom (rys. 1). W drugim kroku generowana jest właściwa powierzchnia interakcji. Wyliczone uprzednio współrzędne punktów poddawane są rotacji wokół wektora $\vec{v}$ ze zmiennym, malejącym krokiem. Etapy opisanego algorytmu wykonywane są dla każdego z atomów.

4. Algorytm haszowania geometrycznego

Algorytm haszowania geometrycznego został zaproponowany jako efektywny sposób wyszukania wzorców (modeli) w scenie niezależnie od zmiany ich położenia, orientacji i skali [4]. W zastosowaniu do problemu dokowania molekularnego pierwotny algorytm został jednak zmodyfikowany tak, aby można go było stosować do zbiorów punktów z przestrzeni trójwymiarowej przy możliwej rotacji i translacji poszukiwanego modelu (zmiana skali w problemie dokowania molekularnego nie zachodzi, gdyż rozmiary molekuł są ściśle określone).



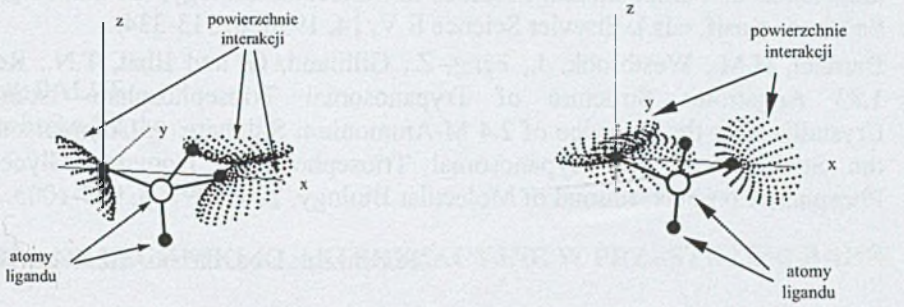
Rys. 2. Ilustracja algorytmu generowania molekulanej powerzchni interakcji. A – pierwsza faza algorytmu. B – gotowa powierzchnia interakcji (w rzucie płaskim)

Algorytm haszowania geometrycznego składa się z dwóch etapów: etapu przetwarzania wstępnego i etapu rozpoznawania [4,5].

Pierwszy etap rozpoczyna wyznaczenie punktów charakterystycznych a^m modelu (współrzędne przestrzenne atomów ligandu mogących wchodzić w interakcje z atomami białka). Następnie dla każdej uporządkowanej trójki takich punktów $(a^m{}_i, a^m{}_j, a^m{}_k)$ definiuje się bazę nowego układu współrzędnych w taki sposób, że pierwszy z punktów stanowi początek tego układu, drugi wyznacza kierunek osi OX, trzeci zaś wyznacza płaszczyznę XY nowego układu współrzędnych. Możliwe staje się zatem wyznaczenie kierunków osi OY i OZ nowego układu oraz przyporządkowanie pozostałym punktom modelu współrzędnych w nowym układzie. Każdy punkt modelu po przekształceniu do nowego układu współrzędnych oprócz trójki współrzędnych kartezjańskich (p^m, q^m, r^m) otrzymuje etykietę m_{ijk} określającą, które punkty modelu stanowiły bazę układu. Działania te poprzedzone są procedurą dyskretyzacji przestrzeni z zadanym krokiem d, mającą na celu ograniczenie ilości przetwarzanych danych.

Ostatnim krokiem w tym etapie jest wypełnienie tablicy haszującej H. Jeżeli w punkcie o współrzędnych (p^m, q^m, r^m) w układzie zbudowanym na bazie punktów $(a^m{}_i, a^m{}_j, a^m{}_k)$ znajduje się punkt modelu, to w komórkę tablicy H o adresie (p^m, q^m, r^m) wpisywana jest etykieta m_{ijk}, w przeciwnym przypadku komórka pozostaje pusta (w jednej komórce może znajdować się wiele etykiet).

Etap rozpoznawania podobnie jak etap przetwarzania wstępnego rozpoczyna wyznaczenie punktów charakterystycznych a^b na powierzchniach interakcji. Następnie analogicznie jak w pierwszym etapie dokonuje się transformacji tych punktów do nowych układów współrzędnych, których bazami są uporządkowane trójki $(a^b{}_i, a^b{}_j, a^b{}_k)$ oraz dyskretyzacji otrzymanych wyników z krokiem d. W efekcie otrzymujemy nowe współrzędne (p^b, q^b, r^b) punktów a^b. Kolejnym krokiem algorytmu jest procedura głosowania, która sprowadza się do przeglądnięcia komórek tablicy H o adresach równych współrzędnym (p^b, q^b, r^b) i oddanie głosu na model, którego etykieta m_{ijk} znajduje się w tej komórce. Modele o największej liczbie głosów rozpatrywane są jako potencjalne rozwiązanie zadania dokowania molekularnego.



Rys. 3. Położenie atomów ligandu na powierzchniach interakcji –rozwiązania otrzymane w wyniku zastosowania techniki haszowania geometrycznego (widok perspektywiczny przestrzeni 3D)

5. Wyniki i wnioski z przeprowadzonych symulacji

Przedstawiony powyżej algorytm zastosowano w próbie rekonstrukcji naturalnego dopasowania struktury izomerazy oraz ligandu SO_4 (grupa siarczanowa), pochodzących z kompleksu o oznaczeniu 5TIM [7]. Jako punkty charakterystyczne modelu wybrano współrzędne przestrzenne środków atomów tlenu ligandu. Dla uproszczenia obliczeń jako punkty charakterystyczne wybrano wierzchołki powierzchni interakcji. W wyniku działania algorytmu otrzymano cztery potencjalne rozwiązania, które zostały zweryfikowane w celu odrzucenia przypadków, w których ligand koliduje ze strukturą białka. W ten sposób odrzucono dwa przypadki otrzymując ostatecznie wyniki prezentowane na rysunku 3. Przeprowadzone badania wskazują na duże możliwości implementowanego algorytmu. Mimo rozbudowanego etapu przetwarzania wstępnego znacząco maleje czasochłonność wyszukiwania wzorców i ilość potencjalnych rozwiązań. Ewidentna przewaga algorytmu objawia się zwłaszcza w przypadku poszukiwania dopasowania kilku modeli (ligandów), gdyż cała procedura może być przeprowadzona równolegle dla wszystkich wzorców.

LITERATURA

1. Bohm, H-J.: LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. Journal of Computer Aided Molecular Design, 6,1992, p.593-606.
2. Rarey, M., Wefing, S., Langauer, T.: Placement of medium-sized molecular fragments into active sites of proteins. Journal of Computer Aided Molecular Design, 10, 1996, p. 41-54.
3. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Research, 28, 2000, p. 235-242.
4. Wallace, A. C., Laskowski, R. A. & Thornton, J. M.: LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions. Protein Engineering, 8, 1995, p. 127-134.
5. Wolfson H.J. Geometric hashing: an overview. IEEE Computational Science and Engineering, 13, 1997, p. 10-21.

6. Wolfson H.J., Nussinov R. From computer vision to protein structure and association in Computational Methods in Molecular Biology, (S. Salzberg, D.B. Searls, S. Kasif, eds.), Elsevier Science B.V, 14, 1998, p. 313-334.
7. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G. and Bhat, T.N.: Refined 1.83 Angstroms Structure of Trypanosomal Triosephosphate Isomerase, Crystallized in the Presence of 2.4 M-Ammonium Sulphate. A Comparison with the Structure of the Trypanosomal Triosephosphate Isomerase-Glycerol-3-Phosphate Complex, Journal of Molecular Biology, 220, 1991, p. 995-1005.

Recenzent: Doc. dr hab. inż. Piotr Widlak

Abstract

The technique called molecular docking can be regarded as a potential method for computer aided drug design and optimization. The problem of molecular docking is usually referred to as a process of finding a proper ligand (in our case a drug molecule) that fits (geometrically and energetically) a specific region of protein designated as protein binding site. Such matching should cause particular biochemical response i.e. viral protein function inhibition.

The problem can be solved using either geometric or energetic approach.

In current paper we consider geometric aspect of molecular docking using methods derived from the domain of robotic vision.

In order to account for interactions between a protein and a ligand we use interaction surfaces introduced by H.-J. Bohm [1]. The interaction surfaces represent a number of biochemical rules governing fundamental types of possible molecular interactions (i.e. hydrogen bonds or hydrophobic interactions) stored in the form of easy to use set of geometric constraints. In our approach we assume ligand is a small molecule without internal degrees of freedom and protein is a rigid immobilized body. The 3D structure (cartesian coordinates of atoms) of both the ligand and the protein is considered known and taken form public database of protein structures The Protein DataBank (PDB)[3]. The location of protein binding site and types of protein-ligand interactions, used as reference, are obtained thanks to the LIGPLOT software[4].

As a solution to the problem of finding a correct pose of the ligand in the binding site we study a robotic vision method called geometric hashing introduced in [5]. Geometric hashing was previously used in molecular docking domain, however without using any biochemical model of protein-ligand interaction [6]. The protein and the ligand were simply considered as arbitrary rigid bodies. Basically the idea of geometric hashing is to find an object in scene using a database of models of objects. In our case the scene is understood as protein binding site and the object as a ligand, both represented as discrete sets of points in 3D space. The geometric hashing algorithm consists of consecutive preprocessing and recognition phases. During the preprocessing phase ligand geometric features of interest are stored in an array called hash-table, while during the recognition phase the binding site features are matched with the ligand features in a voting process.

We tested our method trying to reconstruct native binding pose of the SO_4 ligand in 5TIM [7] complex of trypanosomal triosephosphate isomerase structure downloaded form PDB database.