第 13 卷第 3 期 应用泛函分析学报 Vol.13, No. 3
2011年 9 月 ACTA ANALYSIS FUNCTIONALIS APPLICATA Sept., 2011

DOI : 10.3724/SP.J.1160.2011.00303
文章编号: 1009-1327(2011)03-0303-07

Optimization Applications in Molecular Recognition

KUANG Zhifeng[1,2], Rajesh R Naik[1], Barry L. Farmer[1]

1. *Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright-Patterson AFB, OH* 45433, *USA*
2. *Universal Technology Corporation, Dayton, OH* 45432, *USA*

Abstract: Molecular recognition is the specific non-covalent binding of two or more molecules. Molecular recognition plays an important role in nature such as materials self-assembly, cellular signal transduction and the expression of genetic information, and drug design. Mathematical modeling and computational techniques from quantum mechanics to classical molecular mechanics are reviewed in understanding the molecular recognition. Mathematically the molecular recognition problem can be modeled as an optimization problem. Challenges in solving the optimization problem are reviewed. Novel clustering and binning selection schemes are reported to increase the flexible ligand-protein docking prediction success rate from 63% to 90%.

Keywords: optimization; molecular recognition; density functional theory; autodock; clustering and binning

CLC number: O224 **Document code:** A

1 Introduction

The integration of mathematics, computation and applications has yielded and will continue to yield results never before possible and ideas never before imagined. The cross-fertilization of mathematics, computation and chemistry has been recognized by the 1998 Nobel Prize in Chemistry shared by Professor John Pople and Walter Kohn. The applications of mathematics and computation in biology are the next big things.

A main goal in chemistry is to determine molecular structures of a molecule. In principle, this goal can be achieved by finding the ground state N-electron wave function $\Psi_0(\vec{r}_1, s_1, \vec{r}_2, s_2, \cdots, \vec{r}_N, s_N)$ which minimizes the functional minimization problem[1]

$$\min_{(\Psi,\Psi)=1} \left(\Psi, \hat{H}\Psi\right) \tag{1.1}$$

where $\vec{r}_i \in R^3$ and $s_i \in R$ are the positions and the spin coordinates of electrons; $\hat{H}$ is the Hamiltonian operator which reads in the Born-Oppenheimer non-relativistic approximation in atomic units,

$$\hat{H} = \sum_{i=1}^{N} \left(-\tfrac{1}{2}\nabla_i^2\right) + \sum_{i=1}^{N} v(\vec{r}_i) + \sum_{i<j}^{N} \frac{1}{|\vec{r}_i - \vec{r}_j|} \tag{1.2}$$

where $v(\vec{r}_i)$ is the external potential acting on electron i; and

$$\left(\Psi, \hat{\mathrm{H}}\Psi\right) = \int \cdots \int \Phi^*(\vec{r}_1, s_1, \vec{r}_2, s_2, \cdots, \vec{r}_N, s_N)\hat{H}\Phi(\vec{r}_1, s_1, \vec{r}_2, s_2, \cdots, \vec{r}_N, s_N) \mathrm{d}\vec{r}_1\mathrm{d}s_1\mathrm{d}\vec{r}_2\mathrm{d}s_2\cdots\mathrm{d}\vec{r}_N\mathrm{d}s_N \tag{1.3}$$

It has been known that in most cases, the minimization problem (1.1) is too complicated to allow analytical solution. The advances in computer technology and computational methods have made it possible to find approximate solutions for small molecules. The enormous progress in this area has

Received date: 2011-06-21

been implemented in the Gaussian computer program developed by John Pople[2]. However, there is still great need to further simplify the problem to enable the computational solutions of larger systems. The density functional theory (DFT) developed by Walter Kohn significantly simplified the problem[3]. The breakthrough theorem on which the density functional theory is founded is actually very simple[4]. Generally speaking, for a given function $\rho(\vec{r}_1)$, $\vec{r}_1 \in R^3$, there are many functions $\Psi(\vec{r}_1, s_1, \vec{r}_2, s_2, \cdots, \vec{r}_N, s_N)$ such that

$$\rho(\vec{r}_1) = N \int \cdots \int |\Psi(\vec{r}_1, s_1, \vec{r}_2, s_2, \cdots, \vec{r}_N, s_N)|^2 \, \mathrm{d}s_1 \mathrm{d}\vec{r}_2 \mathrm{d}s_2 \cdots \mathrm{d}\vec{r}_N \mathrm{d}s_N \tag{1.4}$$

However, Professor Kohn and his coworker have found that there is a one-to-one correspondence between function ρ and Ψ if there is unique external potential $v(\cdot)$ such that the function Ψ is the strict minimizer to the minimization problem (1.1). Therefore, the formidable minimization problem (1.1) with respect to the $4N$-dimensional trial function Ψ can be transformed into a significantly simplified 3-dimensional problem of trial function $\rho(\vec{r})$, $\vec{r} \in R^3$

$$\min_{\rho} E[\rho] \tag{1.5}$$

where $E[\rho] = \int v(\vec{r})\rho(\vec{r})\, d\vec{r} + F[\rho(\vec{r})]$. F is a well-defined but not explicitly known universal functional.

Formally speaking, the problem of determining a molecular structure has been transformed into a seemingly trivial problem of finding the solution to the minimization problem (1.5). The difficulty towards applications is to construct the functional F. It has been an open question for mathematicians and physicists to find the explicit form of F since 1964. Another open question is asked what the sufficient and necessary conditions are of a general operator for the existence and uniqueness of the functional optimization problem (1.1).

With the improvements of available approximation methods, computational DFT has become a standard tool in estimating the ground state total energy of a system at absolute zero temperature. In this talk, we will show how the energy minimization theory can be used to quantify the binding affinity of molecules in the area of molecular recognition using ligand-protein docking as examples.

2 Molecular Recognition

By molecular recognition, here we refer to the specific binding of two molecules through non-covalent intermolecular interactions. The basic non-covalent intermolecular interactions include hydrogen bonding, metal coordination, hydrophobic forces, van der Waals forces, pi-pi interactions, and electrostatic effects. Molecular recognition plays an important role in biological structure and function. For example, a DNA molecule consists of two complementary chains of nucleotides. The antiparallel double helix structure of two polynucleotide chains is determined by the hydrogen bonds between adenine (A) and thymine (T) and between guanine (G) and cytosine (C). The binding of ligands to receptors regulates many biological functions such as signal transduction. Molecular recognition is a key to understand biological systems and materials self-assembly. Molecular recognition offers great potential for applications in various fields such as drug design, surface coating, catalysis, and molecular electronics[5–7].

According to thermodynamics, the Gibbs free energy change between the bounded states and the unbounded states determines whether a guest/ligand molecule (L) most likely forms a bound complex

(LR) with another host/receptor molecule (R). In the NVT configuration, $\Delta G = \Delta E - T\Delta S$ where T is the temperature; ΔE and ΔS are the conformational energy and entropy change between after and before binding, respectively. The calculation of the entropy is very difficult. In most cases, we have to resort to approximation methods[8].

At the zero temperature approximation, the binding affinity can be quantified by the energy difference:

$$\Delta E = E(LR) - E(L) - E(R) \tag{2.1}$$

The computational task is to calculate the ground state energy of complex LR and unbounded states L and R using the density functional theory equation (1.5). However, there are several challenges in implementing this procedure. 1) the exact universal functional F is unknown; 2) existing approximate forms of F ignore the London energy due to induced dipole interactions; 3) the basis set inconsistency between the LR complex system and the isolated subsystems L and R leads to the basis set superposition errors; 4) the convergence is very slow for large systems if they are convergent; 5) the computational load is very demanding. Therefore a practical way is to first construct the energy difference function and then minimize the energy function. The obtained energy from equation (1.5) is used to parameterize the energy function. Since mathematically this is completely different from first performing minimization and then calculating the difference as shown in (2.1), extensive parameterization and correction are usually included in the empirical energy difference function. One of those energy functions used in the popular AutoDock program[9] to study the ligand-protein recognition is as follows.

$$\begin{aligned}\Delta G(\vec{r}_i, \vec{r}_j) = {} & \Delta G_{vdW} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \Delta G_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ & + \Delta G_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij})\, r_{ij}} + \Delta G_{tor} N_{tor} + \Delta G_{sol} \sum_{i,j} (S_i V_j + S_j V_i)\, \mathrm{e}^{\left(-r_{ij}^2/2\sigma^2\right)}\end{aligned} \tag{2.2}$$

where $\vec{r}_i$ and $\vec{r}_j$ are the atomic positions of guest and host molecules; r_{ij} is the distance between $\vec{r}_i$ and $\vec{r}_j$; the five ΔG terms, A_{ij}, B_{ij}, C_{ij} and D_{ij} are fitting parameters; $E(t)$ is a weight function depending the hydrogen bond angle t; q_i and q_j are the atomic partial charges of guest and host molecules; $\varepsilon(r_{ij})$ is the screening coefficients; N_{tor} is the number of sp^3 bonds in the ligand to account for the conformational entropy change; S_i and S_j are atomic solvation parameters; V_i and V_j are atomic volume; σ is a distance-weighting factor.

Therefore the computational task for the molecular recognition of ligand-protein binding is to find partners and their conformations to minimize the energy function (2.2). In the next sections, we will point out the difficulty in this minimization problem and some progress that we have made.

3 Challenges in Optimization Methods

In the previous section, we have shown how a molecular recognition problem can be transformed into a minimization problem. However, this is a very difficult minimization problem because 1) the sampling space suffers from the astronomical number of combinations due to positional, orientational, and conformational possibilities; 2) the energy function (2.2) usually has many local minima; 3) the energy function (2.2) is discontinuous at $r_{ij} = 0$[10].

The discontinuity implies that the deterministic optimization methods which usually require

gradient information are not applicable to our cases. Therefore random optimization methods, especially biologically inspired optimization methods are often used in the molecular recognition field. The advantage of random search algorithms is that they are able to sample large space without the requirement of gradient information of the energy function. The disadvantage is that there is no guarantee for finding the global minimum. In the AutoDock program, hybrid search methods including global and local search aspects are implemented for finding the global minimum. Genetic algorithms are used for global searching. The Solis and Wets method[11] is used for local searching. It has been shown this method is efficient at finding the global minimum.

Unfortunately, due to computational uncertainty and inaccurate description of energy functions, sometimes the conformation corresponding to the global minimum is not necessarily the true conformation corresponding to the natural binding state. This phenomenon is called false positive prediction. By a true binding conformation, we mean those conformations whose root mean square deviation (RMSD) of heavy atoms is less than 2 Å from the experimentally determined crystal structure. When the solution corresponding to the true conformation is excluded in the selection process, it is called false negative prediction. In a molecular recognition study, a main goal is to avoid the false positive predictions and save the conformations that may have been falsely excluded. In the next section, we will demonstrate how effective selection methods can help achieve the goal in the ligand-protein docking.

4 Novel Selection Methods and Tests

Many molecular docking programs have been developed to predict the binding conformations and affinities of ligands associated with proteins. Among them, AutoDock is the most popular free software[12]. As we mentioned in section 2, AutoDock aims at finding the conformation to minimize the interaction energy function (2.2) between a ligand and a protein through exploring all possible positions, orientations and conformations of the ligand paired with the protein. It employs the hybrid searching methods mentioned in section 3 to efficiently generate many binding conformations through translational, rotational and conformational change of a ligand. It is assumed that the true binding conformation is the one corresponding to the lowest energy among the randomly generated conformations.

We have tried to use the AutoDock3.0.5 program to predict the ligand-protein binding conformations and compare the predicted conformations with known X-ray structures of 114 non-covalent ligand-protein complexes listed on Table 1[13]. All files prepared in the Tripos mol2 format were downloaded from the official UCSF Dock website (http://dock.compbio.ucsf.edu/Test_Sets/index.htm). To meet AutoDock convention of atom types, ions such as calcium and zinc were replaced by M. Phosphorus, fluorine and iodine atoms were replaced by X. Electrostatic partial charges and atomic coordinates of all atoms including hydrogen were retained. The AutoDock3.0.5 tool "addsol" was used to add solvent parameters to the atoms of receptors. Rotatable bonds of ligands were automatically determined by the tool "autotors" using flag -A +15.0 -a –h.

Table 1 The protein data bank ID code of tested 114 complexes

1A28	1COM	*1FLR*	1OKL	1TYL	2MCP
1A6W	1COY	*1HAK*	1PBD	1UKZ	2PCP
1A9U	1CPS	*1HDC*	1PDZ	1ULB	2PHH
1ABE	1D3H	*1HSL*	1PHD	1WAP	2PK4
1ABF	1D4P	*1HYT*	1PHG	1XID	2TMN
1ACJ	1DBB	*1IMB*	1PTV	1XIE	2YPI
1ACM	1DBJ	*1IVB*	1QCF	1YDR	3CPA
1ACO	1DG5	*1LAH*	1QPE	2AAD	3ERD
1AI5	1DID	*1LCP*	1QPQ	2ACK	3GPB
1AOE	1DOG	*1LDM*	1RNT	2ADA	3HVT
1AQW	1DR1	*1LST*	1ROB	2AK3	4AAH
1AZM	1DWB	*1LYL*	1RT2	2CHT	4COX
1BYG	1EBG	*1MDR*	1SNC	2CMD	4CTS
1C5C	1ETT	*1MLD*	1SRJ	2CPP	4FBP
1C5X	1F0R	*1MRG*	1TDB	2CTC	4LBD
1C83	1F0S	*1MRK*	1TNG	2DBL	5ABP
1CBX	1F3D	*1MUP*	1TNH	2GBP	5CPP
1CIL	1FGI	*1NGP*	1TNI	2H4N	6RNT
1CKP	1FKI	*1NIS*	1TNL	2LGS	7TIM

For each ligand-protein complex, the AutoDock program was performed six times to find the lowest energy conformations using default parameters. The first search started at the mass-center of the experimentally determined ligand. For the remaining five times, at each time a random starting point was chosen on a spherical surface of radius 5 Å, centered at the mass-center of the ligand. At each time ten different conformations with the lower energies were retained. At the end, a total of 60 conformations were obtained. We find that in 107 out of 114 test cases, the true conformation is among the total 60 retained predicted conformations. However, in only 72 out of 114 test cases (only a 63% success rate, see column 2 in table 2), the true conformations were found to be the conformations of the lowest energy among the 60 predicted conformations. That means the global minimum energy criterion produces 42 (114-72) false positive predictions and excludes 35 (107-72) true conformations. It is crucial to reduce the number of false positive predictions and retain the true conformations for real applications. This problem has been addressed in the clustering technique, the statistical rescoring method and the consensus clustering method to analyze pre-generated conformations[14−18].

In the clustering technique, starting from the lowest energy sample, the RMSDs of the other samples from it are calculated. If the RMSD is less than a given threshold 2 Å, the sample is grouped into the same cluster with the lowest energy sample. The procedure is repeated for the remaining ungrouped samples whose RMSDs are beyond the given threshold until all samples are grouped into a list of distinct clusters. The clusters are then ranked by their population. The most populated one is called the top cluster. If a representative is chosen from each cluster, it was reported that 87% true binding conformations for their test complexes are among the first five representatives[14]. Here we report that 88% true binding conformations for our 114 test complexes are among the lowest and

the highest energy samples in the top cluster (column 3 in table 2). We have also found a binning technique can help increase the prediction success rate.

In the binning technique, the RMSDs of all samples from the lowest energy sample are binned into an interval of 0.5 Å bin width. All the samples corresponding to the highest frequency belong to the top bin. If the lowest and the highest energy samples in the top bin are selected as the candidates of the true conformation, 86% successful prediction rate can be achieved (column 4 in table 2).

If all the four candidates selected from the top cluster and top bin are considered, 103 out of 114 test cases (90%) can find their true conformations among the four candidates (column 5 in table 2). That means the top cluster and top bin methods can't replace each other. Together, the successful prediction rate increases from 63% to 90%. This is a significant improvement in the prediction success rate.

Table 2 Success rate depends on selection methods

number of candidates	1	2		4	60
		top cluster	top bin		
number of success cases	72	100	98	103	107
success rate	63%	88%	86%	90%	94%

1-lowest energy conformation; 2-lowest and highest energy conformation in the top group; 4-all the conformations from 2; 60-total obtained conformations

5 Conclusions

It is shown that molecular recognition can be mathematically modeled as a minimization problem. There are many challenges in solving the minimization problem at the quantum mechanics level and the classical pairwise potential level. Any breakthrough in the two fundamental open questions mentioned in this paper will lead to a worldwide recognition. Applications of mathematics and computer science in biology will lead the future research. For AutoDock, a popular molecular docking program in the area of ligand-protein recognition, it is shown that the predicted conformation with the lowest energy does not necessarily correspond to the natural binding state. A post selection scheme combining the clustering technique with the binning procedure has been proposed to be a good remedy for this shortcoming. Using this scheme, 90% natural binding states can be found among four candidates. This selection method will add significant value to the computational drug design community.

References:

[1] Parr R G, Yang W. Density Functional Theory of Atoms and Molecules [M]. New York: Oxford University Press, 1989.

[2] Frisch M, Trucks G W, Schlegel H B, et al. Gaussian, Inc. Wallingford CT, 2004.

[3] Kohn W. Nobel Lecture: electronic structure of matter-wave functions and density functionals [J]. Reviews of Modern Physics, 1999, 71(5): 1253.

[4] Hohenberg P, Kohn W. Inhomogeneous electron gas [J]. Physical Review, 1964, 136(3B): B864.

[5] Fritz J, Baller M K, Lang H P, et al. Translating biomolecular recognition into nanomechanics [J]. Science, 2000, 288(5464): 316–318.

[6] Silva A M, Cachau R E, Sham H L, et al. Inhibition and catalytic mechanism of HIV-1 aspartic protease [J]. Journal of Molecular Biology, 1996, 255(2): 321–40.

[7] Kuang Z, Kim S N, Crookes-Goodson W J, et al. Biomimetic chemosensor: Designing peptide recognition

elements for Surface functionalization of carbon nanotube field effect transistors [J]. ACS Nano, 2009, 4(1): 452–458.

[8] Meirovitch H, Cheluvaraja S, White R P. Methods for calculating the entropy and free energy and their application to problems involving protein flexibility and ligand binding [J]. Curr Protein Pept Sci, 2009, 10(3): 229–243.

[9] Morris G M, Goodsell D S, Halliday R S, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function [J]. Journal of Computational Chemistry, 1998, 19(14): 1639–1662.

[10] Brooks C L, Onuchic J N, Wales D J. Taking a walk on a landscape [J]. Science, 2001, 293(5530): 612–613.

[11] Solis F J, Wets R J-B. Minimization by random search techniques [J]. Mathematics of Operations Research, 1981, 6(1): 19–30.

[12] Sousa S F, Fernandes P A, Ramos M J. Protein-ligand docking: Current status and future challenges [J]. Proteins: Structure, Function, and Bioinformatics, 2006, 65(1): 15–26.

[13] Moustakas D, Lang P, Pegg S, et al. Development and validation of a modular, extensible docking program: DOCK 5 [J]. Journal of Computer-Aided Molecular Design, 2006, 20(10): 601–619.

[14] Kallblad P, Mancera R L, Todorov N P. Assessment of multiple binding modes in ligand-protein docking [J]. J Med Chem, 2004, 47(13): 3334–3337.

[15] Stahl M, Rarey M. Detailed analysis of scoring functions for virtual screening [J]. J Med Chem, 2001, 44(7): 1035–1042.

[16] Wang R, Lu Y, Wang S. Comparative evaluation of 11 scoring functions for molecular docking [J]. J Med Chem, 2003, 46(12): 2287–2303.

[17] Paul N, Rognan D. ConsDock: A new program for the consensus analysis of protein–ligand interactions [J]. Proteins: Structure, Function, and Bioinformatics, 2002, 47(4): 521-533.

[18] Lee J, Seok C. A statistical rescoring scheme for protein–ligand docking: Consideration of entropic effect [J]. Proteins: Structure, Function, and Bioinformatics, 2008, 70(3): 1074-1083.

最优化方法在分子识别中的应用

Zhifeng Kuang[1,2], Rajesh R. Naik [1], Barry L. Farmer [1]

1. 美国空军研究实验室材料与制造部
2. 美国宇航技术公司

摘要: 分子识别是指两个或多个分子靠非共价键专一地结合在一起. 分子识别在物质的形成, 细胞信号的传递, 基因信息的表达和药物的设计等方面起重要的作用. 我们首先对数学和计算方法在分子识别上的应用作了回顾, 并从量子力学, 经典分子力学和热力学上解释分子识别可转换成一类最优化问题. 其次, 我们指出了解决这类最优化问题的困难. 最后, 我们报告了在预报配体与蛋白质识别上所获得的一类新的选择方法. 这类方法可将预报的成功率从63% 提高到90%.

关键词: 最优化方法; 分子识别; 密度泛函理论; 自动对接; 分簇聚类法; 分段聚类法