Fully Quantum Mechanical Energy Optimization for Protein–Ligand Structure

YUN XIANG, DA W. ZHANG, JOHN Z. H. ZHANG
Department of Chemistry, New York University, 100 Washington Square East, Room 1001, New York, New York 10003

Received 3 February 2004; Accepted 12 April 2004
DOI 10.1002/jcc.20069
Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: We present a quantum mechanical approach to study protein–ligand binding structure with application to a Adipocyte lipid-binding protein complexed with Propanoic Acid. The present approach employs a recently developed molecular fractionation with a conjugate caps (MFCC) method to compute protein–ligand interaction energy and performs energy optimization using the quasi-Newton method. The MFCC method enables us to compute fully quantum mechanical ab initio protein–ligand interaction energy and its gradients that are used in energy minimization. This quantum optimization approach is applied to study the Adipocyte lipid-binding protein complexed with Propanoic Acid system, a complex system consisting of a 2057-atom protein and a 10-atom ligand. The MFCC calculation is carried out at the Hartree–Fock level with a 3-21G basis set. The quantum optimized structure of this complex is in good agreement with the experimental crystal structure. The quantum energy calculation is implemented in a parallel program that dramatically speeds up the MFCC calculation for the protein–ligand system. Similarly good agreement between MFCC optimized structure and the experimental structure is also obtained for the streptavidin–biotin complex. Due to heavy computational cost, the quantum energy minimization is carried out in a six-dimensional space that corresponds to the rigid-body protein–ligand interaction.


Key words: quantum mechanical energy optimization; protein–ligand structure

Introduction

Determination of molecular structure through energy minimization is of fundamental importance and practical utility in chemistry. For a given molecular system, knowing its stable geometry is just the starting point for many chemical explorations. One class of system that is of significant interest is protein–ligand binding especially as related to the design of therapeutic inhibitors for specific protein targets. For example, how ligands or drug molecules bind to their target proteins, in what geometry, and due to what specific interactions, etc., are of fundamental importance to determine the best protein inhibitors with high selectivity. Thus, geometry optimization of a protein–ligand complex through energy minimization is crucial in the design of potent protein inhibitors.

Because of the large size of proteins (with thousands of atoms), current geometry optimization of the protein–ligand complex is almost exclusively based on energies derived from empirical force fields in which the potential energy is given by some predefined empirical formulas, such as in AMBER,1 CHARMM,2 and others.3–5 etc. Despite great success of the force field methods in computational study of biological systems, there are still many limitations. The simplistic atomic pairwise formalism of the force field limits the accuracy of the energy it gives. In particular, the current force field approach does not take into account the polarization effect. To give more reliable prediction of molecular structure, it is desirable to employ ab initio quantum mechanical calculation for geometry optimization of the protein–ligand complex. Although geometry optimization has become an integral part of many available quantum chemistry packages from the very beginning, these energy minimization calculations are limited to relatively small molecules, typically less than a hundred atoms in practical calculations. Recent development of linear scaling methods has made significant advance in extending quantum calculation to large molecular systems, but their current applications to biomolecular systems are largely limited to semiempirical methods due to huge computational costs.6–15

Recently, a new method has been proposed to perform full ab initio quantum mechanical calculation for protein interaction energy, in particular the protein–ligand interaction system.16–18 The basic idea behind the MFCC (molecular fractionation with conju-
gate caps) approach is to cut a protein chain into individual amino acid fragments and to insert pairs of proper molecular caps (conjugate caps or concaps) to cap the fragments. This method has been successfully tested on several small peptide–water interaction systems with excellent results,\textsuperscript{16,19} and has also been applied recently to a number of real protein–ligand systems with thousands of atoms including HIV-I gp41–water,\textsuperscript{20} streptavidin–biotin,\textsuperscript{17} and β-trypsin–benzamidine.\textsuperscript{18} A somewhat different approach was proposed by Kitaura and coworkers a few years ago to compute quantum mechanical energy for proteins.\textsuperscript{21} The approach of Kitaura and coworkers to quantum mechanical computation of protein energy also divides the protein into amino acid fragments but it uses an electric static field instead of conjugate caps.\textsuperscript{21} Another related method was recently proposed for quantum mechanical computation of large molecular interactions in which the fragments are capped with a single hydrogen atom, and its applicability is therefore limited.\textsuperscript{22}

In this article, we extend the MFCC method for geometry optimization of the protein–ligand complex by combining energy and gradient calculation with a local optimization method using the quasi-Newton method. We describe the gradient calculation in the MFCC approach and apply the optimization method to a protein–ligand system. This system consists of a 10-atom ligand in complex with a 2057-atom protein (adipocyte lipid-binding protein or ALBP) whose structural data is taken from the Protein Data Bank (PDB id 1LIC). Adipocyte lipid-binding protein or ALBP is an intracellular fatty acid-binding protein, and is responsible for the intracellular trafficking of fatty acids.\textsuperscript{23} The protein is found in adipocytes and macrophages, and binds a large variety of intracellular lipids with high affinity, and has been linked to the development of number of diseases.\textsuperscript{24–26} In our study, the protein is first protonated, preoptimized with the force field, and then fixed in space. We then minimize the interaction energy between the ligand and the fixed protein by quasi-Newton method, with the energy and gradients calculated using the MFCC method at the Hartree–Fock level. The MFCC calculation for the protein–ligand interaction is implemented in a parallel scheme on multiple computers.

This article is organized as follows. The next section describes the methodology for full quantum mechanical computation of protein–ligand interaction energy. Algorithm for geometry optimization as well as parallel implementation of a numerical calculation of the MFCC method are described. Then we provide application to protein–ligand systems and discusses the results of calculation, we then conclude.

**Theoretical Approach**

**Energy and Gradient for Protein–Ligand Interaction**

The MFCC approach is developed to compute \textit{ab initio} energy for protein–ligand interaction in which the protein is kept rigid.\textsuperscript{16} Below, we give a brief description of the MFCC approach as applied to computing the protein–ligand interaction energy. The main idea of the MFCC approach is to divide a protein molecule into amino acid fragments that are properly capped. Using the fractionation scheme, the interaction energy between the protein and another molecule (ligand) can be computed by separate calculations of individual fragments interacting with the ligand. A crucial feature of the MFCC approach is that a pair of conjugate caps (or concaps) are inserted at the location of cut. These caps are introduced to serve two purposes. (1) They cap the cutoff fragments to satisfy the valency requirement. (2) They mimic the local chemical environment of the original protein to the capped fragments. In addition, the pair of caps is fused to form proper molecular species such that the extra interaction energy between caps of the fragments and the ligand can be essentially subtracted. Figure 1 illustrates the MFCC scheme in which a peptide bond is cut and the fragments are capped.

By cutting the peptide bond of the protein into amino acid fragments and inserting a pair of concaps CH\textsubscript{3}CO—NHCH\textsubscript{2} at the location of cut to cap the fragments, the MFCC method uses the following formula to compute the protein–ligand interaction energy $E_{P-L}$.\textsuperscript{16,18}

\begin{equation}
E_{P-L} = \sum_{i=1}^{n} E_{F_{i}} - \sum_{i=1}^{n-1} E_{CC_{i}} - \sum_{i=1}^{n} E_{F} + \sum_{i=1}^{n-1} E_{CC} - E_{L} \tag{1}
\end{equation}

Here, $E_{F_{i}}$ denotes the capped $i$th fragment-ligand energy, $E_{CC_{i}}$ the $i$th concap-ligand energy, $E_{F}$ and $E_{CC}$ are, respectively, the self-energy of the $i$th fragment and $i$th concaps, and $E_{L}$ is the ligand self-energy. For a protein with $n$ amino acids, there are $n – 1$ concaps needed. In situations where the protein has additional chemical bonds like disulfide bonds, additional cutting of these bonds are needed. For example, if the protein has any disulfide bond, eq. (1) is generalized to

\begin{equation}
E_{P-L} = \sum_{i=1}^{n} E_{F_{i}} - \sum_{i=1}^{n-1} E_{CC_{i}} - \sum_{i=1}^{n} E_{F} + \sum_{i=1}^{n-1} E_{CC} - \sum_{i=1}^{n} E_{DC_{i}} - E_{L} \tag{2}
\end{equation}

where $E_{DC_{i}}$ is the disulfide concaps-ligand energy as introduced in ref. 27. In the MFCC approach, the protein can also be cut at
other locations as has been tested before, but cutting the peptide bond seems to be preferable in overall considerations.

Following the above procedure to calculate protein–ligand interaction energy, it is relatively straightforward to obtain energy gradients or forces with respect to the ligand molecule. The energy gradients can be employed in applications such as geometry optimization or molecular dynamics studies. Currently, analytical calculation of energy gradient is quite routine in quantum chemistry calculations at the levels such as HF, DFT, MP2, et al. This functionality is widely available in some popular quantum chemistry packages like Gaussian98. Using the MFCC method of eq. (1) to calculate protein–ligand interaction energy, the energy gradient can be calculated as follows

\[ G_{P-L} = \frac{\partial E_{P-L}}{\partial q_L} = \sum_{i=1}^{n} G_{E-L} - \sum_{i=1}^{n-1} G_{G_{CC}-L} - G_L \]  (3)

where \( q_L \) denotes one of the Cartesian coordinates of the ligand and \( G_L \) is the gradient due to intraligand interaction.

In the MFCC approach, we are only concerned with the interaction energy between a rigid protein and a ligand, not the internal energy of the protein. Thus, by neglecting the protein internal energy, the computation of the protein–ligand interaction energy scales linearly with the protein size, and therefore, can be efficiently carried out on modern computers. Thus, the protein structure is prefixed, either from crystal structures or from computer simulations, before the protein–ligand interaction energy is calculated.

**Geometry Optimization using the Quasi-Newton Method**

Once the energy and gradient for the protein–ligand interaction are obtained by quantum mechanical calculations at desired levels of accuracy, we can carry out geometry optimization by searching for energy minimums. Geometry optimization is of significant importance, as it is directly related to the structure of the complex. An important area of such application is search for protein–ligand binding structures that are of direct interest to drug design. Due to the large size of the protein, current popular geometry optimization methods in drug design for protein–ligand binding almost depend exclusively on empirical force fields such as programs like DOCK, GOLD, and AUTODOCK. At present, it is computationally intractable to employ standard quantum chemistry methods to compute protein interaction energy. A promising approach is to use the mixed quantum/classical approach like QM/MM methods in which a small part of the protein is treated by quantum mechanics while the remaining part is treated by classical mechanics.

In our approach, we employ the MFCC method to compute quantum mechanical energy and its derivative for use in geometry optimization of the protein–ligand complex. Because energy optimization for the protein–ligand complex is computationally expensive even using empirical force fields, we limit our geometry search to that of a rigid ligand relative to a rigid protein at a given crystal structure in the present calculation. However, the algorithm is applicable to energy optimization calculation for a flexible ligand. By limiting our search to rigid bodies, the problem becomes a simplified six-dimensional optimization problem. The six degrees of freedom in this approach include three center-of-mass coordinates (\( \mathbf{X} \), \( \mathbf{Y} \), \( \mathbf{Z} \)) and three Euler angles (\( \Phi \), \( \Theta \), \( \Psi \)) as shown in Figure 2, where the protein is fixed in space. After the energy gradients are computed as typically in Cartesian coordinates of all atoms in the ligand, coordinate transformation between is needed to transform the Cartesian gradients to that of the center of mass and Euler angles. This is carried out by utilizing the rotation matrix and the details can be found from the standard text book.55

For local geometry optimization, a variety of methods are currently available. The purpose of this study is to find the local energy minimum using the quantum mechanical energy and gradient for the protein–ligand interaction. A highly accurate and efficient approach to find local energy minimum is the Newton method. However, the Newton method requires second analytical derivatives that can be expensive to compute in ab initio calculations. We choose to use the quasi-Newton method, which is an approximately Newton method, for geometry optimization. In quasi-Newton method, only first-order energy derivative is needed to construct an approximate Hessian matrix, and is computationally efficient, especially for large molecular systems for which the exact Hessian matrix is very expensive to calculate.

In the quasi-Newton approach, a quadratic approximation of the potential energy function near a stationary point is assumed

\[ E(q) = E(q_i) + (q - q_i) g(q_i) + \frac{1}{2} (q - q_i) A(q - q_i) \]  (4)

where \( q \) is the coordinate, \( g \) is the energy gradient, and \( A \) is the Hessian matrix. At the stationary point, the gradient \( g \) vanishes. A step is taken toward the stationary point by

\[ q - q_i = A^{-1} g(q_i) \]  (5)

The basic idea of the quasi-Newton method is to build up, iteratively, a good approximation to the inverse Hessian matrix \( A^{-1} \), that is, to construct a sequence of matrices \( H_i \) with the property,
The close agreement between the theoretical calculation and experimental measurement. This is quite reasonable considering the fact that the current energy optimization is carried out in the gas phase. Figure 3 shows the structures of ALBP–propanoic acid binding complex before and after the MFCC optimization at the HF/3-21G level. There are small deviations from the experimental crystal structure. We measure the RMSD (root mean square deviation) according to the standard definition. The MFCC optimized ligand structure has a RMSD of 2.98 Å from the crystal structure. The resolution of the crystal structure is 1.6 Å. Thus, although the deviation of the optimized binding structure from the crystal structure is nonnegligible, it is still generally close to the experimental measurement. This is quite reasonable considering the fact that the current energy optimization is carried out in the gas phase without including any solvent molecules, and the protein is rigid. The close agreement between the theoretical calculation and experimental measurement in the binding structure is very encouraging. It shows that as far as the ligand binding structure is
concerned, the gas-phase calculation can provide reasonably correct binding structure.

The MFCC method provides a convenient means to analyze detailed protein–ligand interaction energies. The specific interaction energies between the ligand and various amino acid groups given by the MFCC calculation gives insight on detailed mechanism of protein–ligand binding. For example, using the MFCC results, we can plot the quantum “interaction spectrum” or “map” for ALBP–propanoic acid binding. This interaction spectrum is obtained from just a single point calculation at the binding structure, but it provides detailed information for protein–ligand binding as was shown in a previous study for β-trypsin/benzamidine complex. It is useful to point out that in traditional quantum chemistry calculation for the binding complex, if it can ever be done for the protein system, one would simply obtain a single-point energy without much detailed interaction information between the protein and the ligand. The MFCC calculation, on the other hand, provides much detailed specific interaction information regarding individual protein residues and the ligand. This enables us to naturally obtain an interaction spectrum as shown in Figure 4 for ALBP–propanoic acid binding. The spectrum is generated from just a single point energy calculation at the binding geometry. Here, we obtained the interaction spectrum from two geometries: the crystal structure from PDB, and the optimized structure as shown in Figure 3.

The spectrum in Figure 4 shows a number of useful features. First, it clearly identifies the most important protein residues that have strong interactions with the ligand (attractive or repulsive). Secondly, the spectrum shows the relative strength of interaction between the residue and the ligand. These quantitative information on interactions between specific protein components and the ligand is very useful such as in the structure-based drug design. Because the ligand (propanoic acid) is charged in this system, it has strong interactions with many charged groups of the protein as shown in Figure 4. Of course, it is useful to understand that in a solvent environment, there are many water molecules that could screen out interactions between these charged groups, especially for those charged residue that are out in the solvent. For groups inside the active site, few water molecules are present and the dielectric effect of water is much smaller.

We also carried out MFCC optimization for a previously studied system—the streptavidin–biotin complex. The streptavidin–biotin system is one of the most tightly binding complexes for noncovalent binding of a protein and small ligand. The high-affinity binding of biotin to streptavidin or closely related avidin has been exploited for decades. The highly specific and strong binding between these two molecules has remarkable technological utility. For example, the streptavidin–biotin system provides a means to examine a number of biochemical systems in biosensor applications, and it is one of the best characterized systems based on self-assembled monolayers.

We begin with the crystal structure of streptavidin–biotin complex (PDB code: 1 stp). This complex consists of a 121-amino acid protein (streptavidin) and a 31-atom ligand (biotin). We follow the same procedure for protein protonation, hydrogen preoptimization, and MFCC procedure, as is done for the ALBP–propanoic acid system. Because in this system the ligand biotin is about three times that of the first system, the computation took relatively longer time for the MFCC optimization to reach convergence. The quantum optimized structure is found to be very close to the experimental structure (PDB code: 1 stp) and the RMSD is only 0.19 Å. This result is not unexpected. From our previous calculation and analysis in ref. 17, there are several strong hydrogen bonds formed between the protein and the ligand in addition to the strong ionic interaction between them. The resulting structure is shown in Figure 5.

Conclusions

In this article, we presented a fully quantum mechanical method to obtain an optimized protein–ligand binding structure. The optimization method is based on computed quantum mechanical energy and gradient for protein–ligand interaction using a recently developed MFCC method. The present quantum optimization method is particularly suited for finding ligand-binding structure to protein with a fixed structure. In the MFCC approach, the gradient exerted by the protein on the ligand is obtained by the vector sum of gradients exerted by individual protein residues. The minimum energy search is performed by the quasi-Newton method, and numerical calculation is implemented for parallel computation using MPI or PVM. Although the protein structure is fixed (usually at crystal structure), the ligand can be flexible using the present approach. Although the MFCC method makes ab initio calculation for protein–ligand interaction energy realistic, full quantum geometry optimization is still computationally expensive on small computer systems. It is therefore desirable to use quantum MFCC.
energy optimization for a protein–ligand system only when a highly accurate structure is desired.

Although the ab initio level of theory at HF/3-21G is not considered high level, comparisons of the calculated energies for protein–ligand binding with those from higher B3LYP/6-31G and MP2/6-31G levels show that the energy profile (or relative energy) is very similar.\(^{16,27}\) Thus, it is quite reasonable to employ an HF/3-21G calculation to search energy minimums for protein interaction.

It is useful to point out that the current quantum optimization method only gives local minimum energy geometry, not a global minimum, for protein–ligand system. For global minimum energy search, it will be desirable to employ other global methods like Monte Carlo or molecular dynamics methods to carry out a global energy search in combination with the current local quantum optimization method.

### References