An Efficient Algorithm for Conformational Search of Macrocyclic Molecules

by

Cheuk-san (Edward) Wang

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January 1995

Author ....
Department of Electrical Engineering and Computer Science
September 23, 1994

Certified by ....
Tomás Lozano-Pérez
Professor, Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ....
Frederic R. Morgenthaler
Chairman, Departmental Committee on Graduate Students

An Efficient Algorithm for Conformational Search of Macrocyclic Molecules

by

Cheuk-san (Edward) Wang

Submitted to the Department of Electrical Engineering and Computer Science
on September 23, 1994, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

This thesis describes a new algorithm for conformational search of macrocyclic molecules. It scans a large number of candidate conformations and minimizes only the promising ones. These candidates can be generated by two operators that construct new conformations from known minima. The candidates have similar bonded-interaction energy as the known minima and possibly lower non-bonded interaction energy. This algorithm is 9 to 11 times faster than the existing methods when tested on two large rings, cycloheptadecane and rifamycin SV.

Thesis Supervisor: Tomás Lozano-Pérez
Title: Professor, Electrical Engineering and Computer Science

Acknowledgments

I thank my advisor, Tomás Lozano-Pérez, for his continuous guidance and support. He has given me much freedom in pursuing my interests.

I also thank Tao Ke and Jeffery Song for discussion on stereochemistry, Carlo Maley and Kevin Lew for comments on the manuscript.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A *conformation* of a molecule describes its 3-dimensional shape. The conformation gives the positions of all atoms in a molecule. Properties of organic molecules are intimately related to the conformations they are able to attain. The shapes of organic molecules determine how they interact with proteins and other molecules. For example, if a molecule can fit precisely into the active site of a protein, it may inactivate this protein. The protein may participate in a disease such as the common cold, cancer, or AIDS. Thus such a molecule could potentially be a cure for a disease. Rational drug design [BCM93, SSS+93] is an approach that finds molecules to fit into the active sites of disease-causing proteins. In order to determine whether a molecule fits into an active site, the molecule's reachable conformations must first be found. Chemists and biologists are interested in the reachable conformations of molecules because the biological activities and reactions of molecules depend on their conformations.

These reachable conformations can be modeled by minimizing a *potential energy function*. For each arrangement of atoms in a molecule, the energy function gives its approximate energy value. The higher the energy, the less probable that this conformation is to occur. The probability that a molecule would take on a conformation with energy ε_i is given by the Boltzmann distribution

$$p_i = \frac{e^{-\varepsilon_i/kT}}{\sum_i e^{-\varepsilon_i/kT}}$$

in which k and T are the Boltzmann constant and absolute temperature respectively. Suppose two conformations differ in energy by E. Let P_h and P_l be the probability of a molecule having high and low energy conformation respectively. They are related by the following equation.

$$\frac{P_h}{P_l} = e^{-E/kT}$$

At 37°C, if two conformations' energies differ by 3 KCal/mol, $P_h/P_l = 0.0077$. Therefore, a conformation whose energy is more than 3 KCal/mol above the global minimum has very little chance of occurring in nature.

Given the energy function and a molecule, the problem of *conformational search* requires finding all possible conformations of a molecule that have energy close to the global minimum (typically within 3 KCal/mol). Unfortunately, the space of possible conformations has $3N-6$ dimensions where N is the number of atoms in a molecule. There is an enormous number of local minima in this space for even a small molecule with tens of atoms. Numerical minimization with gradient descent or conjugate gradient would be trapped in a local minimum close to the starting conformation. Other global optimization techniques like simulated annealing or genetic algorithms could find the global minimum, but they are very inefficient [MJ93, GW92]. The conformational search problem is generally believed to be NP-hard [UM93]. The best systematic method has 6^n complexity for tree-like molecules where n is the number of single bonds. This method cannot guarantee finding the global minimum because it scans only a few values for each torsional angle. The global minimum may not be accessible from the scanned values.

The goal of this thesis is to develop better methods for conformational search of *macrocyclic molecules*, which are ring-like molecules with 10 or more bonds in the ring. Figure 1-1 shows a conformation of a macrocyclic molecule. There are many macrocyclic molecules in nature, and they are usually biologically active. For example, a class of macrocyclic molecules called *macrolides* are important antibiotics [Omu84, BBNE93]. *Musks* are another class of macrocyclic molecules that have a musky aroma. Natural and synthetic musks are widely used in perfumery.

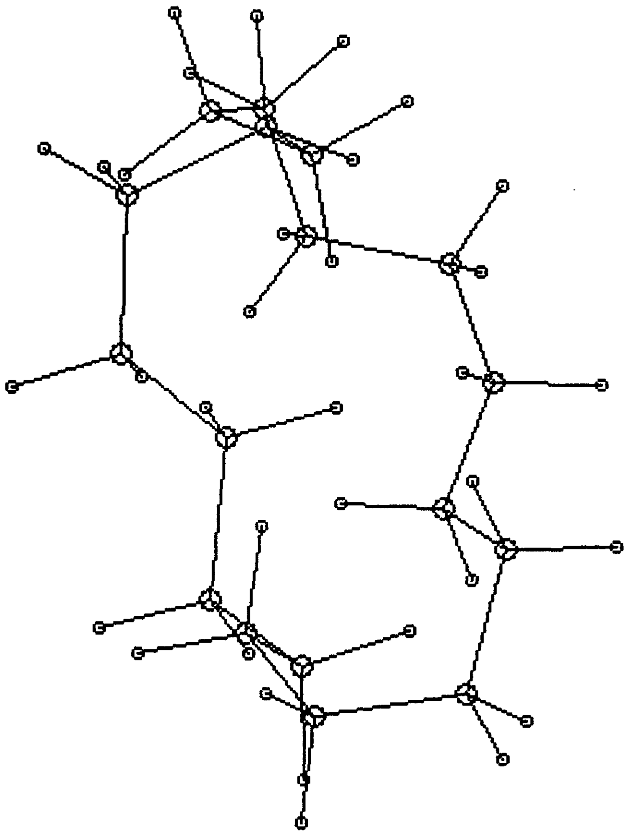


Figure 1-1: A conformation of cycloheptadecane, a 17 carbon cycloalkane.

Cyclic molecules are very different from acyclic ones because they have the ring closure constraint. Atoms in rings usually do not fall on any lattice. Energy in a cyclic molecule distributes quite evenly in stretching, bending, torsion and van der Waals energies. The van der Waals force is extremely repulsive at short distance. In rings, it is very important to avoid repulsive interactions because atoms are generally close to each other. All published algorithms take several days on typical workstations to find all stable conformations of a medium sized molecule. New ring-specific algorithms can perform much better than these general algorithms.

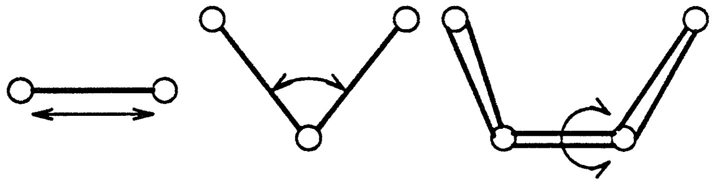


Figure 1-2: Different types of energy resulting from bonded interactions.

1.1 Energy Function

In a typical energy function[Tes79], the potential energy (E) is the sum of energies from bonded interactions (E_{val}) that depend on the specific bonds, and nonbonded interactions (E_{nb}) that depend only on the distances between atoms.

$$E = E_{val} + E_{nb}$$

Bonded interactions depend mainly on the length of bonds (bond stretch, E_B), angles between two adjacent bonds (bond-angle bend, E_A), and torsional angles among 3 adjacent bonds (dihedral angle torsion, E_T) (Figure 1-2).

$$E = E_B + E_A + E_T$$

Bond length is the most rigid local geometry. It deviates very little from the equilibrium value because a great deal of energy is required to compress or stretch a bond. For a single carbon-carbon (C-C) bond, $E_B \approx 350(R - R_e)^2$ KCal/mol/Å^2 where R and R_e are the actual and equilibrium bond lengths respectively. A bond deformation of 0.05 Å thus requires about 1 KCal/mol.

Bond angles are not as stiff as bond lengths. In some large molecules, bond angles could deviate up to 10 degrees from the optimal values. For a C-C-C angle, $E_A \approx 0.01(\theta - \theta_e)^2$ KCal/mol/deg^2 where θ and θ_e are the actual and equilibrium angles respectively. It can be seen that 1 KCal/mol can induce a 10° bending.

Dihedral angles are the most flexible bonded interactions. For single bonds, the

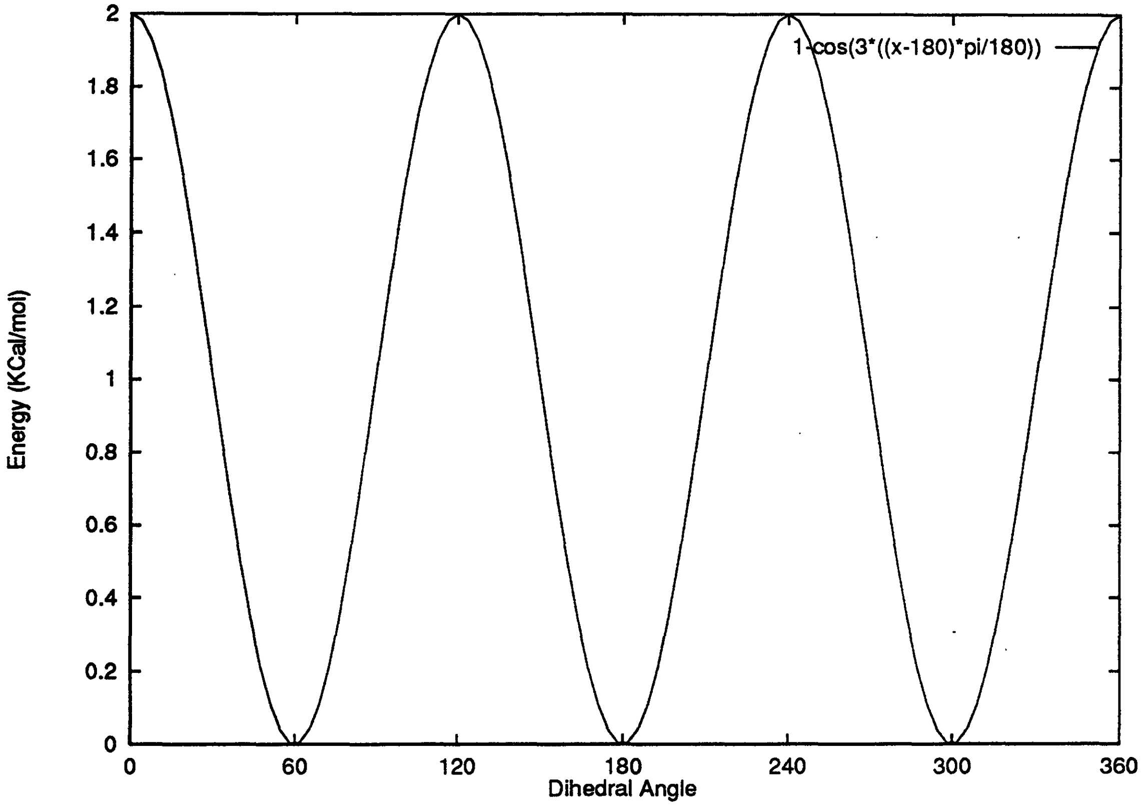


Figure 1-3: Torsional energy of a C-C-C-C dihedral angle.

dihedral angle can change with very little energy penalty. For instance, the torsional energy of C-C-C-C dihedral angle (Figure 1-3) has the form $E_T \approx 1 - \cos(3(\varphi - 180^o))$ KCal/mol where φ is the actual dihedral angle. There are three very different torsional angles that have zero energy. The more single bonds a molecule has, the more conformational energy minima it has.

Nonbonded interactions are affected by the distances between non-bonded atoms. They consist of van der Waals interactions (E_{vdw}), electrostatics (E_Q), and hydrogen bonds (E_{hb}).

$$E_{nb} = E_{vdw} + E_Q + E_{hb}$$

Van der Waals force is attractive at medium distance and extremely repulsive at short distance. The van der Waals energy between two carbon atoms is $E_{vdw} \approx 0.0951((\frac{R}{3.8983})^{-12} - 2(\frac{R}{3.8983})^{-6})$ KCal/mol where R is the actual distance between the

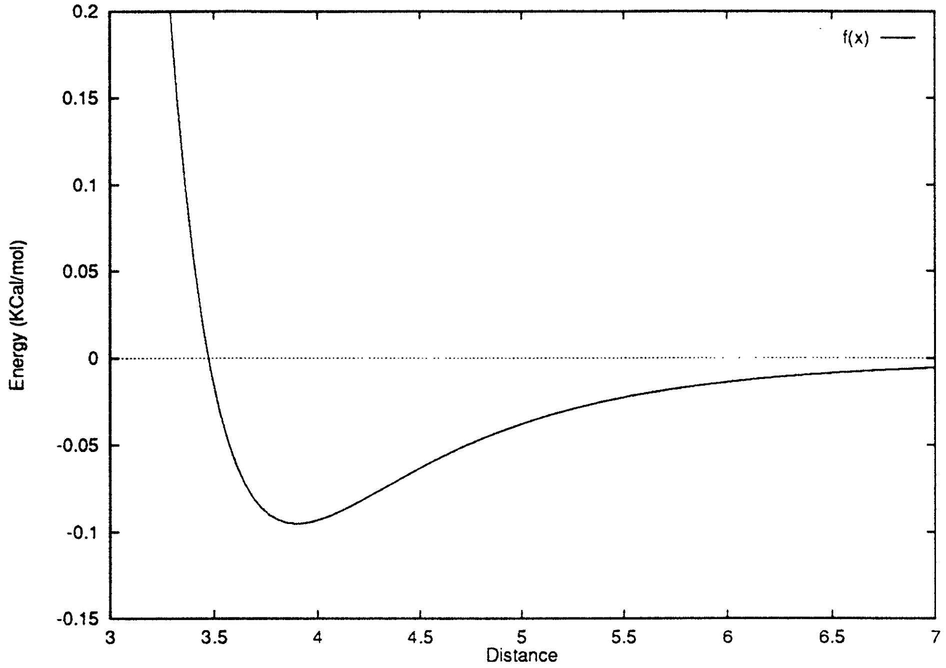


Figure 1-4: Van der Waals energy between two carbon atoms.

atoms (Figure 1-4).

Electrostatic interactions are caused by the attraction and repulsion of charged atoms. If an electron transfers from one atom to another, *formal charges* are added to these atoms. A polar covalent bond is modeled by putting *partial charges* on the atoms forming the bond because the shared electrons do not distribute evenly between the atoms. Electrostatic energy has the form

$$E_Q = 322.0637\frac{Q_iQ_j}{\epsilon R}$$

where Q_i and Q_j are charges of the atoms in electron units. ϵ is the dielectric constant and R is the distance in Å. Electrostatics has longer range than other nonbonded interactions.

Hydrogen bond is a very important stabilizing interaction in macromolecules. It is a complex interaction involving electrostatics, charge transfer, van der Waal's forces, etc. A hydrogen atom serves as a bridge between two electronegative atoms such as nitrogen, oxygen, or fluorine. The hydrogen atom holds one atom by a covalent bond

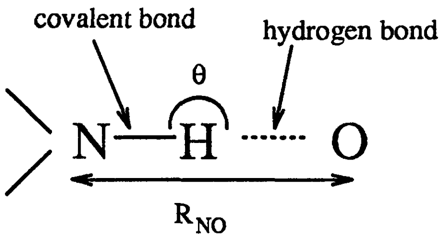


Figure 1-5: A hydrogen bond formed by N-H and O.

and the other by nonbonded forces (Figure 1-5). The factors affecting the strength of a hydrogen bond are interatomic distances, directionality, and linearity. For the hydrogen bond between N-H and O, $E_{hb} \approx 4[5(2.9/R_{NO})^{12}-6(2.9/R_{NO})^{10}]\cos^4\theta$ KCal/mol where R_{NO} is the actual distance between the nitrogen and oxygen atom, and θ is the bond angle between N, H, and O. A collinear configuration of the three atoms would have the lowest potential energy. However, due to the small energies involved, large deviations from collinearity sometimes occur.

Energy functions are also called *force fields* because the force acting on a molecule due to its conformation can be found by differentiating the energy function. ($F = -\nabla E$) There are several published energy functions with slightly different sets of parameters. We shall use the Dreiding force field [MOI90] and MM2 force field [All77]. MM2 is more widely used but also more complicated than Dreiding.

1.2 Chirality

All conformational search algorithms must preserve the *chirality* of molecules. A molecule is *chiral* if it is not superimposable on its mirror image. A chiral molecule (Figure 1-7) usually has at least one *chiral center*, which is a carbon atom bonded to 4 different groups. A conformational search method should find conformations with a fixed chirality because molecules normally change their conformations but not chiralities.

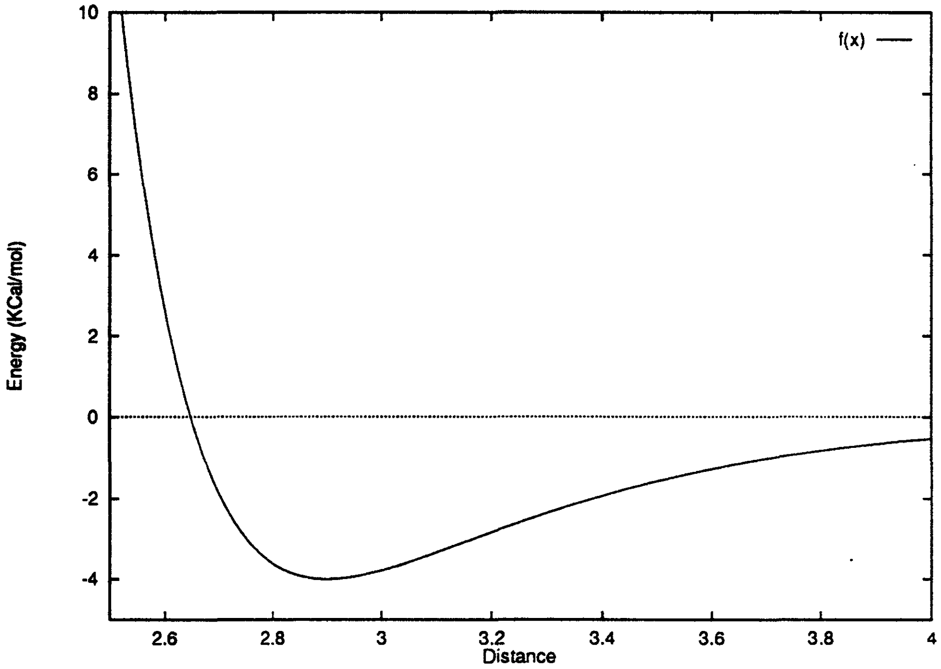


Figure 1-6: Hydrogen bond energy between N-H and O when $\theta = 180^o$.

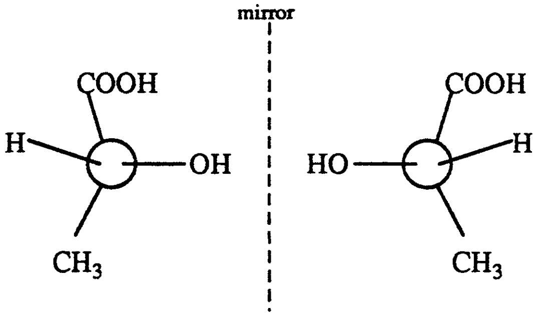


Figure 1-7: Lactic acid, a chiral molecule.

1.3 Previous Work

Conformational search methods have been studied for many years. Most methods do not distinguish between cyclic and acyclic molecules. These techniques can be separated into two categories: stochastic methods and systematic methods. The simplest stochastic technique is *Cartesian stochastic search* [Sau87]. It represents a conformation by the Cartesian coordinates of each atoms. The method operates by taking a known conformation and applying limited, random translations ("kicks") to every atom in the molecule. The resulting conformation becomes the new starting geometry for energy minimization.

A similar method, *internal coordinate Monte Carlo search* [CGS89], represents a conformation by its internal parameters. That is, instead of Cartesian coordinates, it uses bond lengths, bond angles, and torsion angles. The torsion angles are less constrained than other parameters. In each step of the algorithm, several torsion angles are randomly varied. The resulting conformation becomes the new starting point of minimization. When applied to macrocyclic molecules, the random variation of torsional angles often produces conformations that violate the ring-closure constraint. Those structures cannot be used for minimization. The Monte Carlo step has to be repeated until a suitable structure is found.

Molecular dynamics is a stochastic method that simulates the physical interactions of molecules. During conformational search, it models the movement of atoms of a molecule in a thermal bath. The molecule changes its conformation due to thermal vibration. At certain time intervals, the algorithm collects the conformations of the molecule being simulated and minimizes them.

The systematic version of internal coordinate Monte Carlo search is *internal coordinate tree search* [LS88]. Each torsion angle of a single bond is searched through a series of possible values (e.g. $0^o, 60^o, 120^o, 180^o, \ldots$). Suppose there are N single bonds in a molecule and each torsion angle is searched through d values, N^d conformations will be generated and minimized. Its computational complexity is exponential in the number of single bonds. Just as this method's stochastic counterpart, it would also generate many ring structures that violates the closure constraint. Hence this method is quite wasteful of resources for macrocyclic molecules.

The *Distance geometry* [WJW$^+$83] method stores the ranges of distances between atoms in matrices. Each element of a matrix represents the distance between two atoms. A lower bound matrix and an upper bound matrix are used for the algorithm. Each element is computed using constraints like bond length and bond angle. Then one applies the triangle inequality and other higher order inequalities on the elements to tighten the bounds. These inequalities are repeatedly applied until the bounds cannot be tightened any further. Then, coordinates of conformations are generated systematically or stochastically satisfying the upper and lower bounds. These conformations are then minimized.

Expert-system-like approaches [DLP87, AW93] study the components of molecules and deduce their conformation with a rule set. It is not clear whether their approaches can scale to larger molecules.

Several people have developed ring-specific algorithms. They all tried to change local geometry of a ring and then minimize. They use operations like corner flapping [GO89], edge flipping [GO93], and torsion flexing [KG93]. The running time of these methods are similar to other techniques. Chapter 4 compares their performance to our new approach.

Biocad claims to have a conformational search algorithm superior to others, but the algorithm has not been published.

Chapter 2

The Complementarity Approach

The main problem with the existing approaches to conformational search is their inefficiency. For example, with the 51-atom molecule in figure 1-1, all methods use more than 30 CPU days on a MicroVAX II to find most low-energy conformations [SHW+90]. Most of the CPU cycles are spent on minimizing the energy function. Starting from a random conformation, each minimization takes about a minute on a SPARCstation 2 under BatchMin V3.5 [Dep90]. Cartesian stochastic search, internal coordinate tree search and Monte Carlo search, and distance geometry all require about 10,000 minimizations each for the 51-atom molecule. Therefore, these algorithms would take about 7 days on a SPARCstation 2. These methods do not attempt to generate starting geometries close to a local minima. They spend less than 5% of the time finding starting conformations and more than 95% of the time minimizing the energy function on all starting conformations. No selection of the starting conformations was attempted. Not surprisingly, only 2-3% of the minimized conformations are useful. Most minimizations result in high energy or duplicated conformations. The indiscriminate use of minimizations wastes a lot of time.

Naturally, one might try to increase the amount of time spent on generating and selecting starting conformations and reduce the time spent on minimization. We can quickly scan many conformations and minimize only the promising ones. Since each minimization requires thousands of energy function evaluations, if we efficiently scan 1000 conformations and minimize only one or two of them, the use of time is more balanced. We will minimize only those conformations that have low energy and close to a minimum. This will reduce the time for minimization and produce more useful conformations. Some studies have shown that initial energy is not always a good predictor of minimized energy. However, these studies are based on experiments where initial energies are much higher than the global minimum. When the limit is set to be a few KCal/mol above the global minimum, the initial energy becomes an accurate predictor of minimized energy.

This chapter describes a different approach to conformational search of macrocyclic molecules. We call it the *complementarity* approach. Given some known conformational minima of a macrocyclic molecule, one can apply some operators on these minima and generate many new candidate conformations. The quality of these candidates can be found by a single evaluation of the energy function. If we only minimize those candidates that have energy close to the current global minimum and are quite different from known minima, we have a much higher chance of obtaining a useful conformation. The time for minimization is also reduced because of the proximity of the candidate's structure to its local minimum. The operators may not have complete coverage of the conformational space, but we can use other techniques like Cartesian stochastic search to ensure completeness.

The remaining problem is finding a set of operators that can produce many new, low energy conformations given some known minima. For chain or tree-like molecules, one can simply change the torsional angle of some single bonds because these molecules do not have the closure constraint. Finding such operators for macrocyclic molecules is much harder. Other researchers have used operators like corner flapping or edge flipping [GO93], but the candidates they generate have much higher energy than the global minimum. This is because some local geometries (bond angles) are deformed. Before presenting the operators we use, we make two observations about the low-energy conformations.

1. Low-energy conformations have low-energy components.

2. Low-energy conformations share similar components.

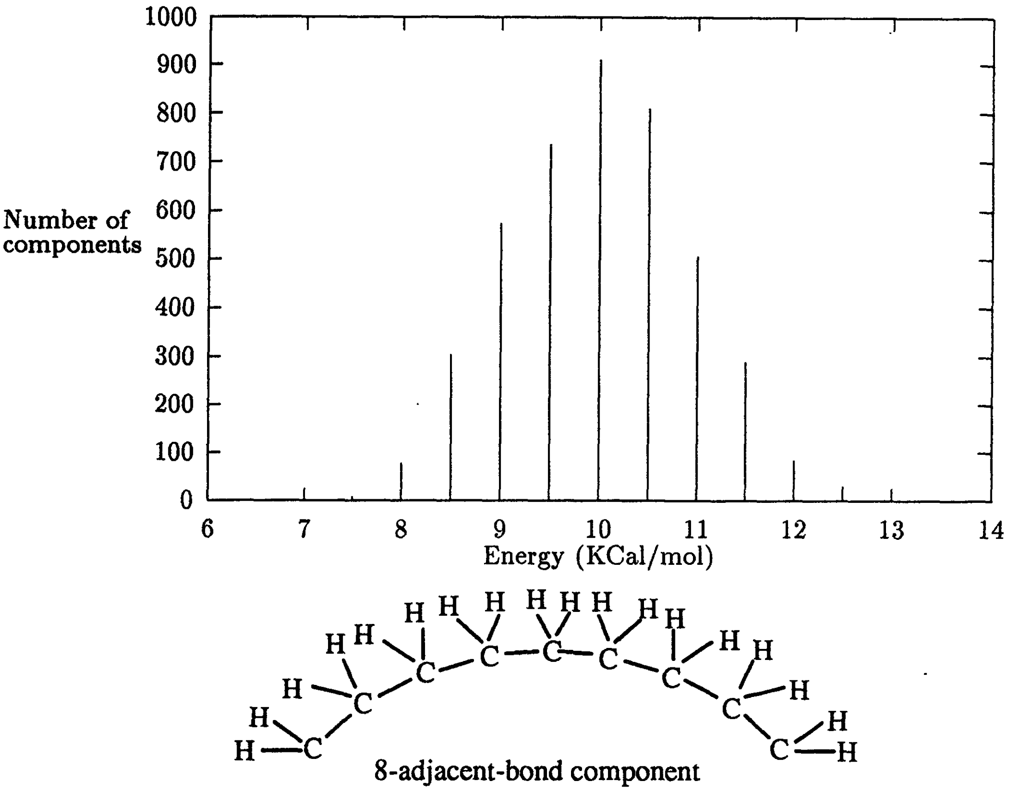


Figure 2-1: Energies of 4352 8-adjacent-bond components of low-energy conformers of cycloheptadecane.

We have found 256 conformations of cycloheptadecane within 3 KCal/mol of the global minimum[1] using the MM2 force field. Figure 2-1 shows a histogram of the MM2 energies of all 8-adjacent-bond components of these conformations. The mean and standard deviation of the energies are 10.257 KCal/mol and 0.921 KCal/mol respectively. Figure 2-1 shows that most of the components have similar energy. There are very few components with high energy. This data supports the first observation. Figure 2-2 shows a similar histogram for rifamycin SV where 42 low-energy conformations are found. The mean and standard deviation of the energies are 15.42 and 1.002 KCal/mol respectively. It is generally not the case that a low-energy component forms a low-energy conformer with a high energy component with the help of non-bonded interactions.

[1] The global minimum of cyclohpetadecane's conformational energy is 19.23 KCal/mol.

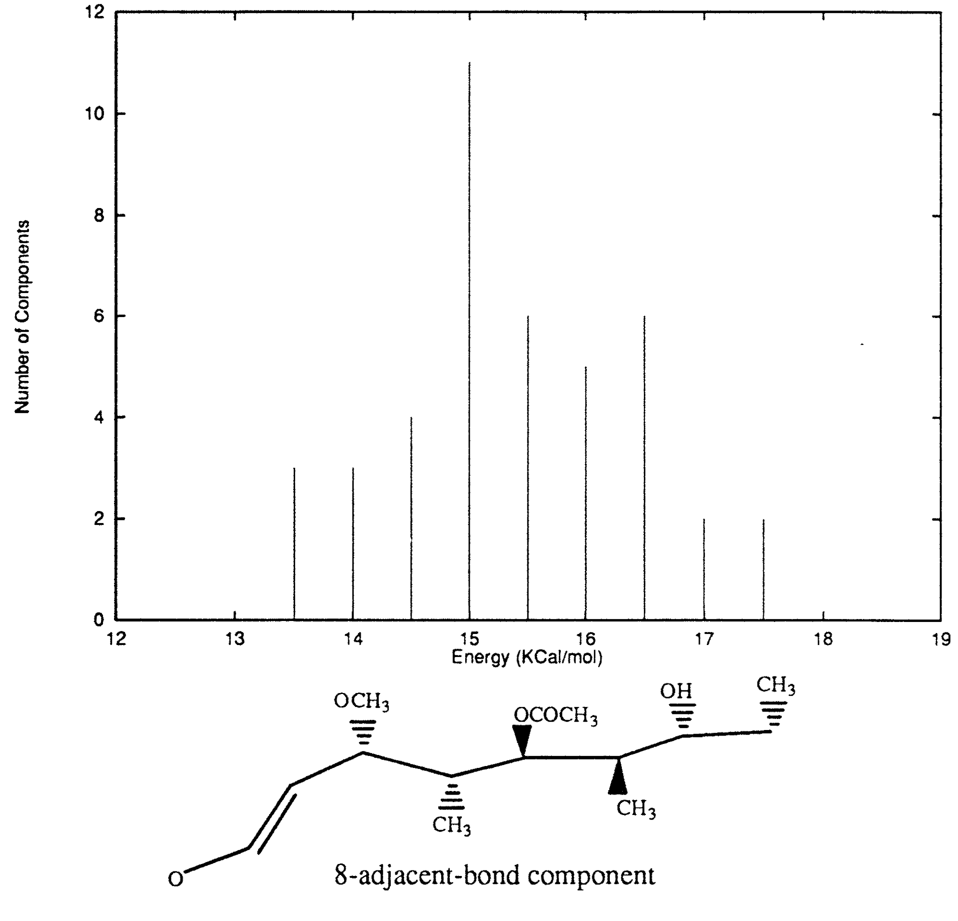


Figure 2-2: Energies of 42 8-adjacent-bond components of low-energy conformers of rifamycin SV.

The second observation is less obvious than the previous one. We want to show that components are "reused" in different conformers. We randomly take 300 8-adjacent-bond components from the low-energy conformers of cycloheptadecane. Then we compare the 6 dihedral angles against the dihedral angles of the other components. On average, there are 20.05 other pieces that have angles all within 20 degree of a selected piece.[2] This is much higher than what would be the result of a uniform distribution of angles. Suppose each dihedral angle can randomly take on one of 3 values (e.g. 60^o, 180^o, 300^o). The probability that two pieces would have the same angles (within 20^o) is $1/3^6 = 0.00137$, whereas the probability for two components of different low-energy conformers to have similar dihedral angles is about $\frac{20.05}{17 \times 256} = 0.0046$. This shows that the distribution of dihedral angles is far from random. One can generalize this observation to conformations of different molecules. We postulate that if two different molecules have a large connected component in common, the low-energy conformations of this shared component in one molecule will very likely appear in the conformations of the other.

Given the above observations, we can use two operators, *combine* and *mirror*, to generate starting conformations with low energy.

2.1 The Combine Operator

The combine operator recombines components from conformational minima. This is similar to the crossover operator in genetic algorithms. We have observed that low-energy conformations have low-energy components. New conformations can be found by recombining these components. Given a set of conformational minima, we can compute all relative positions of every pair of bonds that are several bonds apart. If two pairs of end bonds in two different components have the same relative positions and orientations, then the components can substitute each other (Figure 2-3). The bond lengths and bond angles are preserved but van der Waals interactions can raise

[2] The median and standard deviation are 15 and 17.8 respectively. This distribution has a long tail.

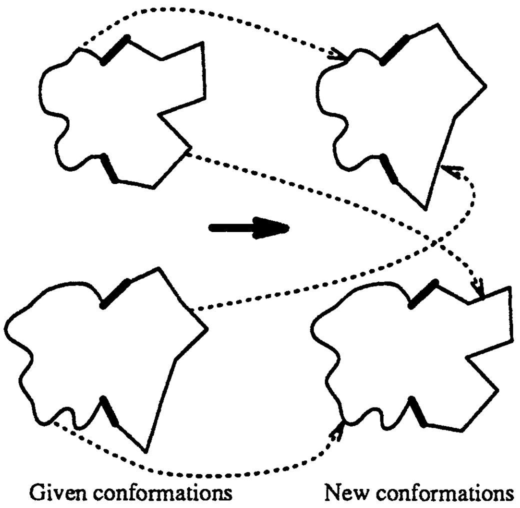
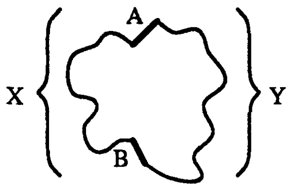


Figure 2-3: The **combine** operator that mixes components from two conformations.

or lower the total energy. If the given conformations have a fixed chirality, the new candidates are guaranteed to have the same chirality. Given m conformations of a cyclic molecule of size n, $O(m^2n)$ candidates will be examined. If the molecule is symmetric (all components have the same chemical formula), there will be $O(m^2n^2)$ candidates to examine.

2.2 The Mirror Operator

The mirror operator substitutes a component by its mirror image. Given a conformation, its mirror image (enantiomeric conformation) has exactly the same energy because all distances among atoms are unchanged. We have observed that a low-energy conformation must be composed of low-energy components. Hence the mirror images of the components must have low energy. Given a ring conformation, we can partition the molecule into two components. If we retain the conformation of one component and "glue" on the mirror image of the other component, the result would be a very different conformation (Figure 2-4). The local geometry of each individual component is unchanged. The components would have the same energy as before. Additional energy can only arise from bonded interactions at the junctions between the components and new van der Waals interactions between the components. We can minimize the change in bonded interactions at the junctions if we retain their local geometries. In other words, we want the bonds at the junctions to have the same bond lengths and angles as the given conformation. This is true if the end bonds of the retained component are coplanar and we use the plane of these end bonds as the plane of reflection of the other component. To prove this, consider a conformation consisting of components X and Y where A and B are the end bonds of X.



Suppose A and B are coplanar. Let R be the operator that reflects a component about the AB plane. We use "+" to denote the joining of two components and "=" to denote the equivalence between two structures. Clearly R is distributive over +. Because A and B must be on the AB plane, $R(A) = A$ and $R(B) = B$. Therefore $R(Y + A + B) = R(Y) + R(A) + R(B) = R(Y) + A + B$. This says that reflecting Y produces the structure as reflecting Y and A and B. Therefore the bond lengths and angles at the junction are unchanged after replacing Y by the mirror image of Y. The torsional angles at A and B may be different but they introduce very little energy. There would be new van der Waals interactions which can increase or decrease the total energy.

Notice that a component would have the opposite chirality of its mirror image. To preserve the chirality of new conformations, the reflected component cannot have any chiral center.

The mirror operator is a specialized form of the combine operator. It is equivalent to applying combine to a conformation and its mirror image. Given a conformation, the mirror operator looks at every pair of bonds. If they are approximately coplanar, one side of the ring is reflected about the plane. If the resulting conformation has low energy and is different from known minima, it will be minimized. The scanning process has $O(mn^2)$ complexity where m is the number of conformations given and

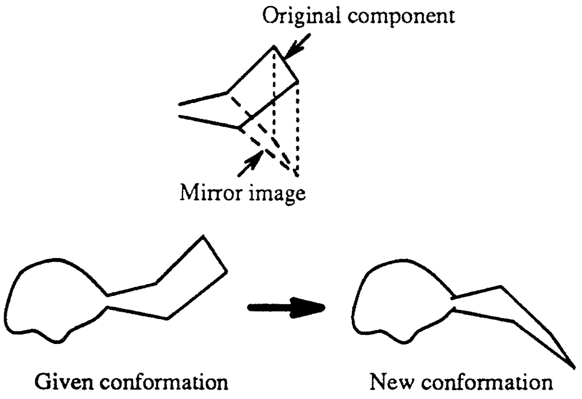


Figure 2-4: The **mirror** operator that replaces a component of a conformation by the component's mirror image.

n is the number of bonds forming the ring. Goto's operators are special cases of this operator. Corner flapping [GO89] reflects two bonds while edge flipping [GO93] reflects three bonds. By looking at all pairs of bonds and computing their planarity, we explore more conformational space more efficiently. There is also a bigger chance of lowering total energy because more bonds are changed. Section 4.1 compares the complementarity approach with Goto's algorithm.

Can the operators be applied to components of any size? With a few approximations, we can find the limitation to the sizes of components. We assume that all bond lengths and bond angles are fixed, but the dihedral angles are free to change. Go and Scheraga [GS70] have shown that two fixed end bonds would generate six nonlinear equations on the dihedral angles. Therefore at least six dihedral angles are needed to satisfy the equations. To apply the operators, the smallest component would have six bonds (Figure 2-5). As a consequence, the smallest ring on which we can apply the operators has two minimal size components and $6 + 6 - 2 = 10$ bonds.

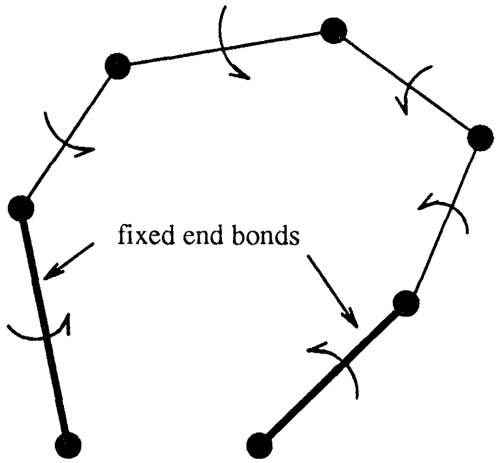


Figure 2-5: A minimal size component

Chapter 3

The Algorithm in Detail

Before presenting the algorithm, we first look at the underlying data structures. There are two data structures used: a priority queue and a database of components. The priority queue (heap) stores the starting conformations prioritized by their energy values. The lowest-energy starting conformation can be removed and new ones can be added very efficiently. This data structure allows us to find and minimize the best starting conformation first.

The other data structure, a database, stores components of conformational minima. It is indexed by the type of components (their atoms and bonds), and the relative orientation of end bonds. In the current implementation, the relative orientation is encoded with the distances among atoms forming the end bonds[1] (Figure 3-1). The database is implemented as a hash table. The hash function applies to the types of components and a discretization of the distances. Given a component of a conformer, we can find all its *complementary components* from the database efficiently. A component's complement is another component with the complementary bonds in the ring, and whose end bonds have the same relative orientation (Figure 3-2). To find the complements of a component, we simply look up the complementary bonds and apply the hash function.

With these data structures, we define a procedure **Generate_starting_confor-**

[1] This encoding does not distinguish between a component and its enantiomer. We make use of this fact so as not to store the enantiomers.

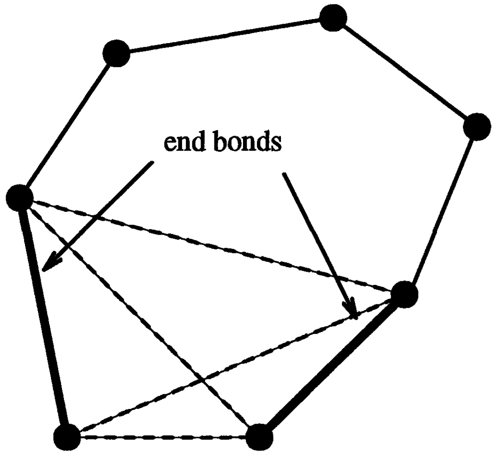


Figure 3-1: Four distances (indicated by dotted lines) are used to encode the relative configuration of the end bonds.

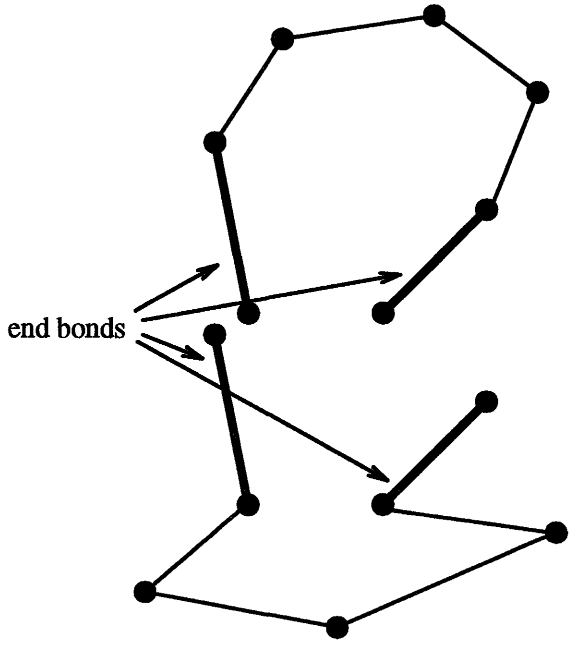


Figure 3-2: Two components that are complements of each other.

mations that finds starting conformations from a newly discovered minimum. Basically, it systematically applies the operators described in Chapter 2 to the new conformation.

Generate_starting_conformations(conformation)

For all connected components C with 6 or more single bonds in the conformation do the following:

1. If C is already in the component database, stop.
2. Otherwise, add C to the database.
3. Find all complementary components of C from the database.
4. Find transformations of the complementary components such that C and the components would form new starting conformations and the ring closure constraint is satisfied.
5. Compute the energies of the starting conformations.
6. Add the starting conformations that have energy within ΔE of the global minimum to the priority queue. ΔE must be bigger than the ultimate energy window and depends on the complexity of the molecule. If the molecule is small and symmetric, 5 KCal/mol is sufficient. If the molecule is large and asymmetric, ΔE can be as high as 15 KCal/mol. One can overestimate ΔE because this only enlarges the priority queue.

Every step of the procedure is straightforward except Step 4. In the current implementation, the conformations are represented by the Cartesian coordinates of the atoms. The end bonds of C and its complementary components have the same relative configurations, but their Cartesian coordinates do not necessarily match. To form a new starting conformation, a rigid transformation is needed to transform the end bonds and other atoms of the complementary component. This is achieved by the algorithm in [FH77]. A 4×4 transformation matrix can be computed to best match (in the least squared distance sense) the coordinates of the atoms on the end bonds of C and the complementary component.

It may seem like only the combine operator that recombines components from conformational minima is used in this procedure, but there is another twist to Step 4. A component can be transformed to its mirror image by a transformation matrix whose determinant is -1. This matrix is also computed and returned if it can produce a close match of the end bonds. Thus the mirror image of a complementary component is used to generate a starting conformation. Because the two halves of a conformer are always complement of each other, the mirror operator is also applied at this step.

With the procedure that generates starting conformations, we can describe the top-level algorithm.

1. A few conformational minima (typically below 100) are found by a randomized method such as Cartesian stochastic search or internal coordinate Monte Carlo search. These minima should have the correct chirality and be dispersed in the conformational space. If internal coordinate Monte Carlo search is used, 6 or more torsional angles should be changed at once.

2. **Generate_starting_conformations** is called with these minima as arguments. The main purpose of this step is to fill the component database with some entries.

3. The starting conformation with the lowest energy before minimization is removed from the priority queue. If the queue is empty, a Monte Carlo operation is performed instead.

4. If the starting conformation is sufficiently different from the known minima, it is minimized. This check is necessary because many similar starting conformations are found by the algorithm.

5. If the new minimum has not been found before, **Generate_starting_conformations** is called with it as the argument.

6. Go to 3.

This algorithm has been implemented in Common Lisp. Dreiding or the MM2 force field can be used for minimization and energy evaluation. The next chapter will evaluate the performance of the algorithm.

Chapter 4

Performance Evaluation

To compare the new algorithm with the existing methods, the best measure is the ratio of CPU time to number of "useful" conformational minima found. Chemists want to find the naturally occurring conformational minima of molecules in the shortest possible time. The "useful" minima are conformations that have energy within a few KCal/mol of the global minimum. Any conformation with higher energy would have little probability of occurring under normal temperature. However, CPU time varies with computer hardware and programming languages. Therefore, we would replace CPU time with the number of energy function evaluations. For existing algorithms, most of the resources are used for minimization. The time for minimization is directly proportional to the number of function evaluations. For the complementarity approach, function evaluations are also used for selecting starting conformations. This measure takes into account the overhead for the complementarity algorithm. Sometimes the number of function evaluations is not available because the minimizing program does not return this information. The number of minimizations would be used in those cases to measure performance.

4.1 Cycloalkane

The cycloalkanes have a regular ring structure of $(CH2)n$ (Figure 4-1). Chemists have extensively studied their conformations. They are good benchmarks for evaluating

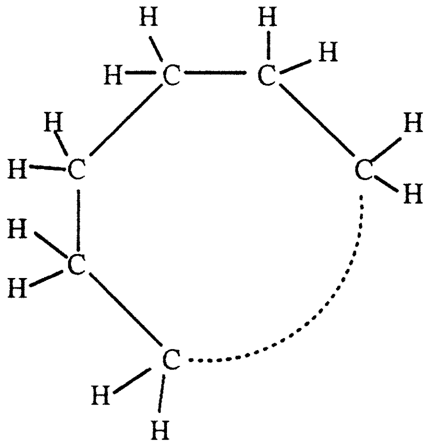


Figure 4-1: Chemical formula of cycloalkane.

conformational search algorithms. Cycloalkanes have neither electrostatic interactions nor hydrogen bonds. Therefore, non-bonded interactions are limited to van der Waal's forces. Additionally, their structures are completely symmetric. All components with the same number of bonds are of the same type in the database. More starting conformations could be generated because our algorithm exploits this symmetry. No other method can make explicit use of this symmetry. Because of these features of cycloalkanes, the complementarity algorithm vastly outperforms existing methods.

For the following results on cycloalkane, only 30 minima are found using the randomized method in Step 1 of the top level procedure.

Our program is first run on cyclopropadecane (Figure 4-2), the 13-carbon cycloalkane, to validate its completeness. We use BatchMin V3.5 [Dep90] and its MM2 force field for minimization. ΔE is set to 6 KCal/mol. Figure 4-3 shows the results of the new algorithm and Cartesian stochastic search running on the molecule. After 2000 Monte Carlo steps, 15 conformers are found to be within 3 KCal/mol of the global minimum. We believe that these are all the conformers because every minimum has been found several times. Cartesian stochastic search finds all useful conformations with 341 minimizations against 98 minimizations for the complementarity algorithm. The complementarity method is 3.48 times faster. More importantly, the result suggests that the algorithm completely explores the low-energy conformational

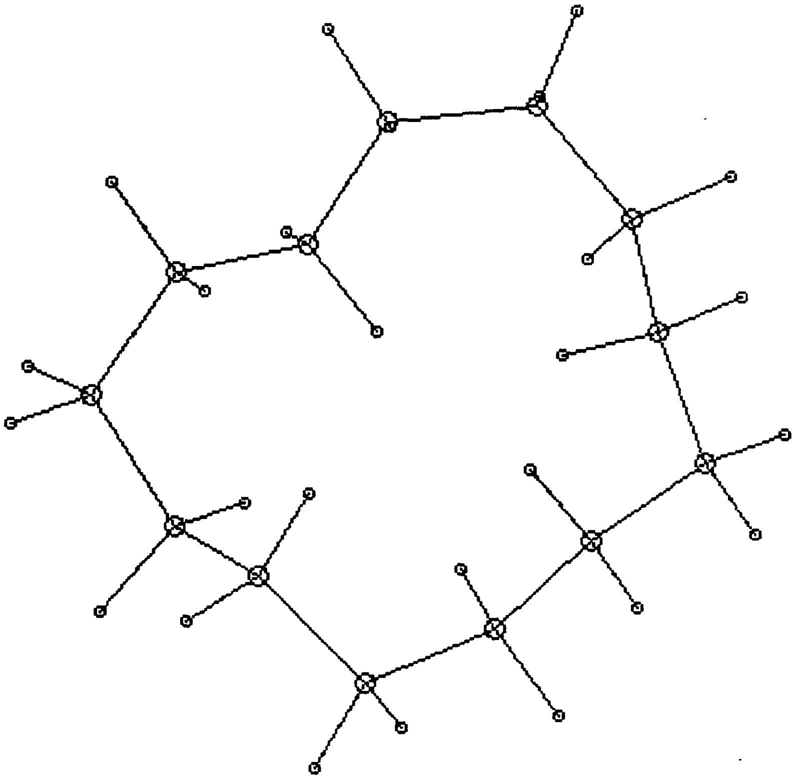


Figure 4-2: The lowest-energy conformer of cyclopropadecane with 20.415 KCal/mol.

space.

The second cycloalkane used as a benchmark is cycloheptadecane, the 17-carbon cycloalkane. Saunders [SHW$^+$90] applied seven methods to search for its conformations. He found that the methods have similar performance. First, we try several techniques using the Dreiding force field. Figure 4-4 compares their performance on this molecule. Our algorithm finds the most useful conformations in the shortest time. As a basis of comparison, we examine the resource each algorithm needs to find the 70th useful conformation in Table 4.1. Clearly the complementarity algorithm performs better than existing ones.

To evaluate the performance more precisely, we try the algorithm using BatchMin

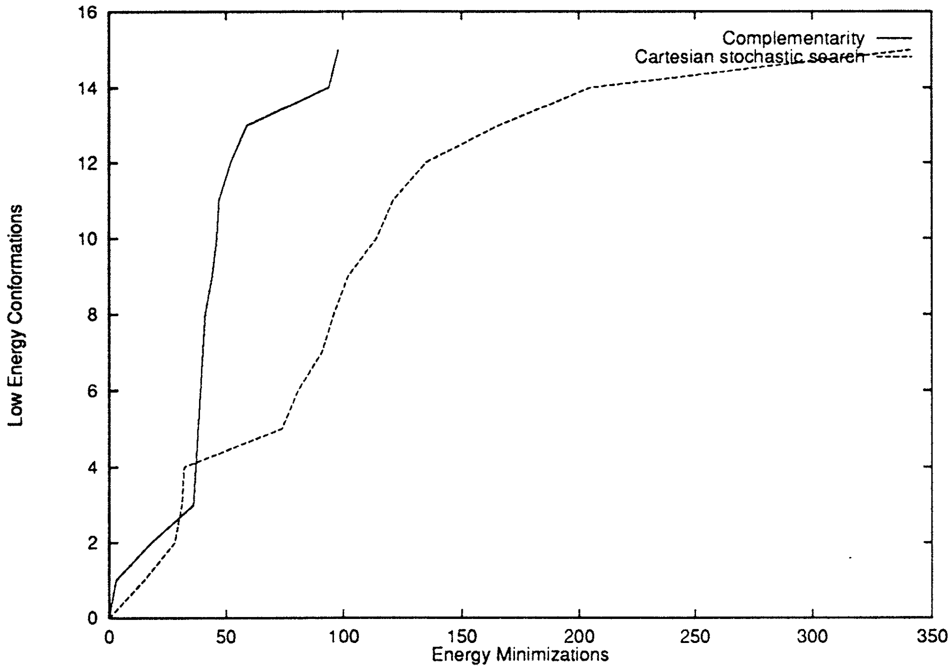


Figure 4-3: Performance of Cartesian stochastic search and the complementarity algorithm on cyclopropadecane.

Algorithm	Number of function evaluations	Approximate CPU time on Sparcstation ELC	Slow down factor against the new algorithm
Cartesian Stochastic search	8672057	6.1 days	35.6
Goto and Osawa's algorithm	1514372	25.5 hours	6.2
Complementarity	243539	4.1 hours	

Table 4.1: Resources needed to find the 70th low-energy conformation of cycloheptadecane (Dreiding force field).

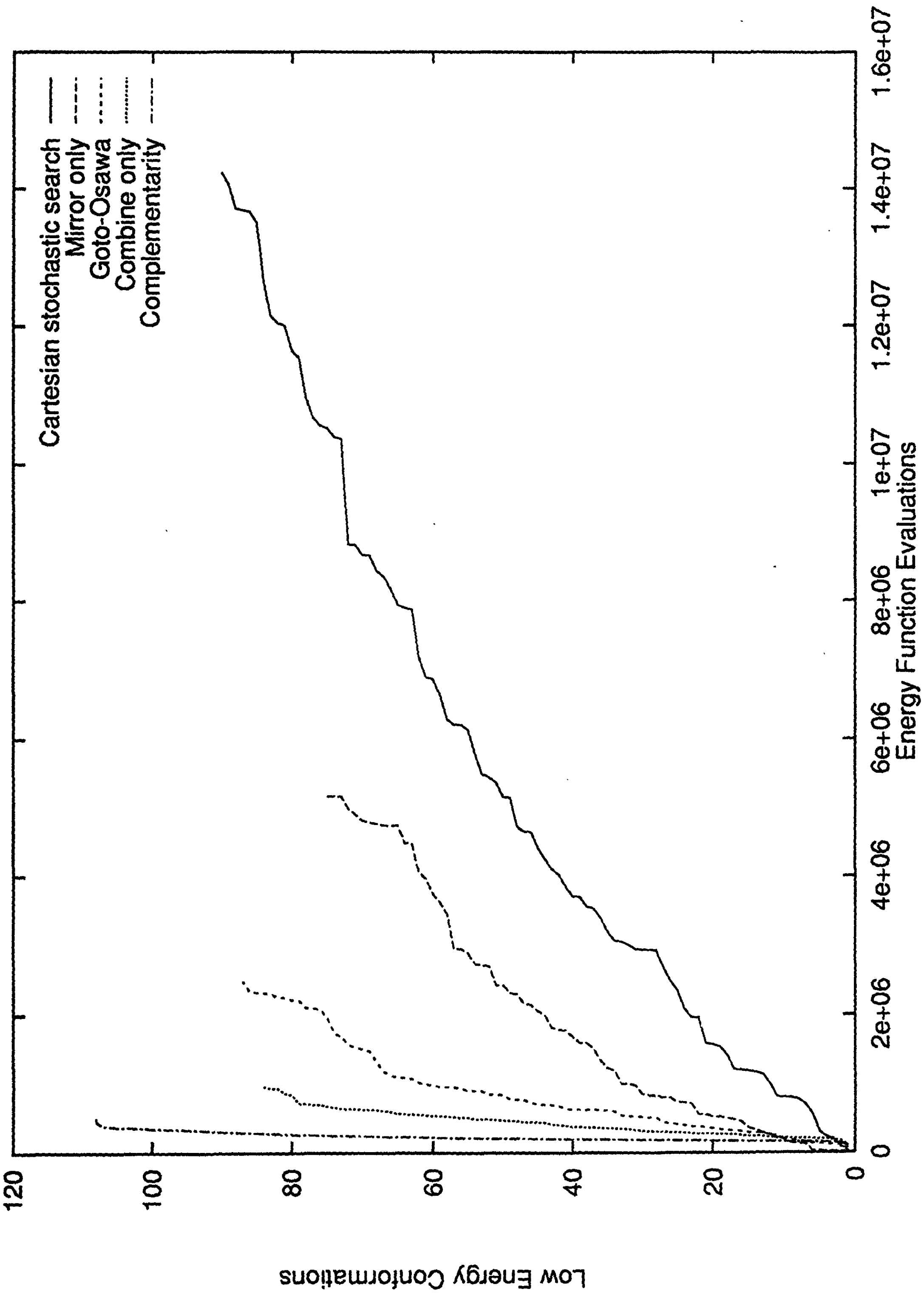


Figure 4-4: Performance of different techniques on cycloheptadecane using the Dreiding force field.

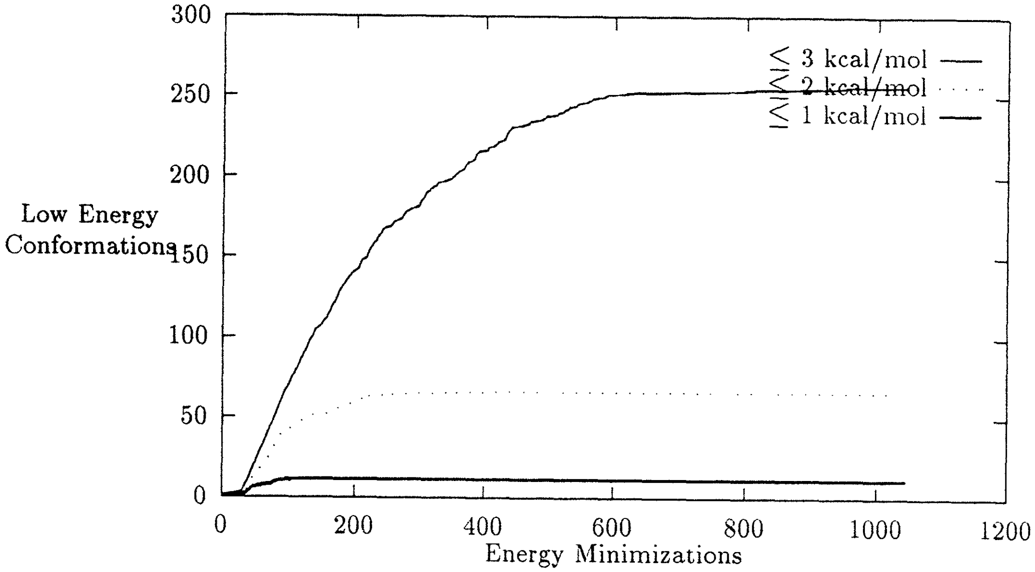


Figure 4-5: Performance of the new algorithm on cycloheptadecane with the MM2 force field.

V3.5 and the MM2 force field. These are the exact program and energy function used in Saunders' paper. We can directly compare our data with those in his paper. Figure 4-5 shows the performance of the algorithm using MM2. All minima are tested by BatchMin's normal mode analysis. Table 4.2 is a comparison of the best method in his paper, usage-directed torsional Monte Carlo search, against our algorithm[1]. We can compare the number of minimizations each algorithm needs to find the 203rd conformation in the 3 KCal/mol bracket, where about 80% of the useful conformers are found. Our algorithm is 9.4 times faster than usage-directed torsional Monte Carlo Search. The complementarity algorithm is 12.2 and 14.7 times faster in finding the 232nd and 249th useful conformation.

Figure 4-5 also shows that conformers within 1 and 2 KCal/mol of the global minimum are found very rapidly in the beginning of the search. In addition, at the early stage of the run, nearly every minimization produces a useful conformation. Table 4.3 is a comparison of the rate of conformational search during the early stage.

The initial rate of conformation discovery is much higher using the complementar-

[1] Data on usage-directed torsional Monte Carlo search is from [SHW+90].

	Energy Minimizations					
kcal/mol	359	847	1045	3388	5647	8471
	Still/Chang/Guida Usage-Directed Torsional Monte Carlo Search					
1		10		10	11	11
2		44		61	66	69
3		110		203	232	249
	Complementarity					
1	11	11	11			
2	65	66	66			
3	203	254	256			

Table 4.2: Unique Conformers Found versus Energy Minimizations during Conformational Searches of Cycloheptadecane.

	percentage of total minima		
	found/100 starting geometries		
Method	1 kcal/mol	2 kcal/mol	3 kcal/mol
Usage-directed torsional Monte Carlo Search	9.1	6.4	4.4
Complementarity	90.9	59.4	26.3

Table 4.3: Rate of conformational search of cycloheptadecane.

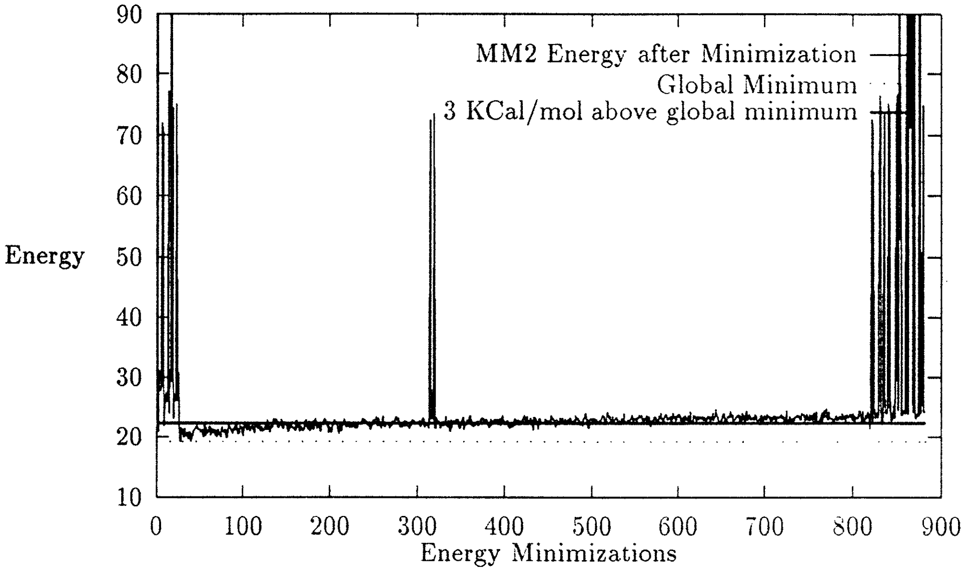


Figure 4-6: MM2 energies of cycloheptadecane conformers after minimizations.

ity algorithm. This is because the operators are able to generate low energy starting conformations and the use of the priority queue. The operators are very effective in finding candidates that minimize to very low energy. The priority queue allows us to minimize the starting conformations in order of their energy. Thus the energies of the minima are roughly in increasing order. Figure 4-6 illustrates this point. The sharp peaks in the graph are caused by the Monte Carlo steps of the algorithm. The rest of the minima have a slow upward trend in energy. This trend shows that the best minima are likely to be found early in the search.

4.2 Rifamycin SV

Our algorithm performs very well on cycloalkanes. To evaluate our method more completely, we try it on a radically different molecule, rifamycin SV. It is a well known representative of the ansamycin family [Aro83] (Figure 4-7). Kolossvary [KG93] has also worked on the conformations of this molecule. He said that it is an extremely difficult conformational search problem for the current technology. The molecule is completely asymmetric. There are strong electrostatic interactions and several

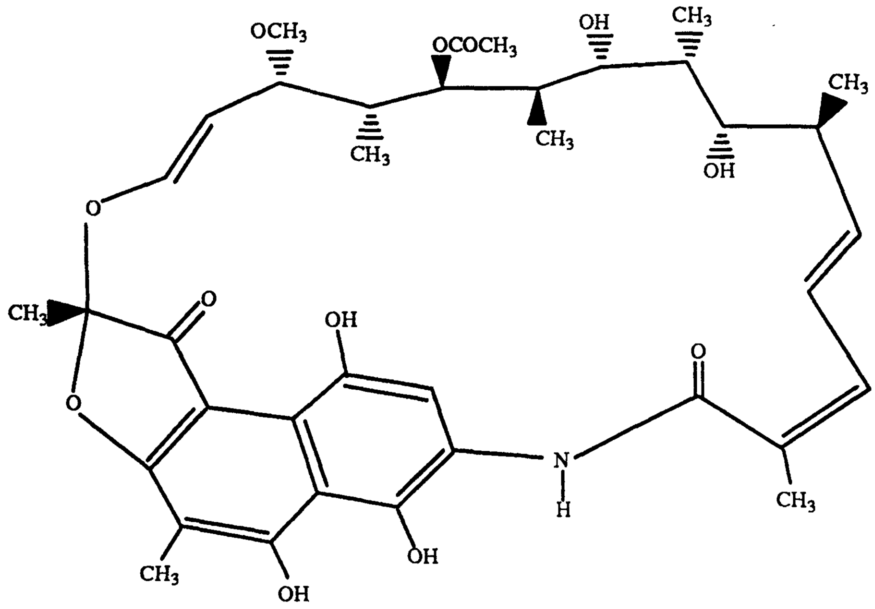


Figure 4-7: Chemical formula of rifamycin SV.

intramolecular hydrogen bonds. Our algorithm cannot exploit any symmetry as in cycloalkane, but the method still outperforms others.

We use the MM2 force field and Batchmin V3.5 for the search. The dielectric constant was attenuated by a factor of 10. These are the same parameters used by Kolossvary. We do not use any united atoms for the computation. 100 minima are found using the randomized method in step 1 of the top level procedure. ΔE is set to 15 KCal/mol. We can only use the combine operator because all components have at least one chiral center. The lowest energy conformer found by our algorithm has an energy of 50.006 KCal/mol, which is lower than the 55.69 KCal/mol found by Kolossvary. Figure 4-9 shows the result of running our algorithm for 892 energy minimizations. Table 4.4 is a comparison of methods for conformational search of rifamycin SV[2]. Complementarity has vastly outperformed the other methods. It found the 42nd conformer within 3 KCal/mol of the global minimum after 849 minimizations versus 10000 for FLEX, the best of the other algorithms. It is 11.8 times faster than FLEX.

[2] Data on SUMM and FLEX are from [KG93]

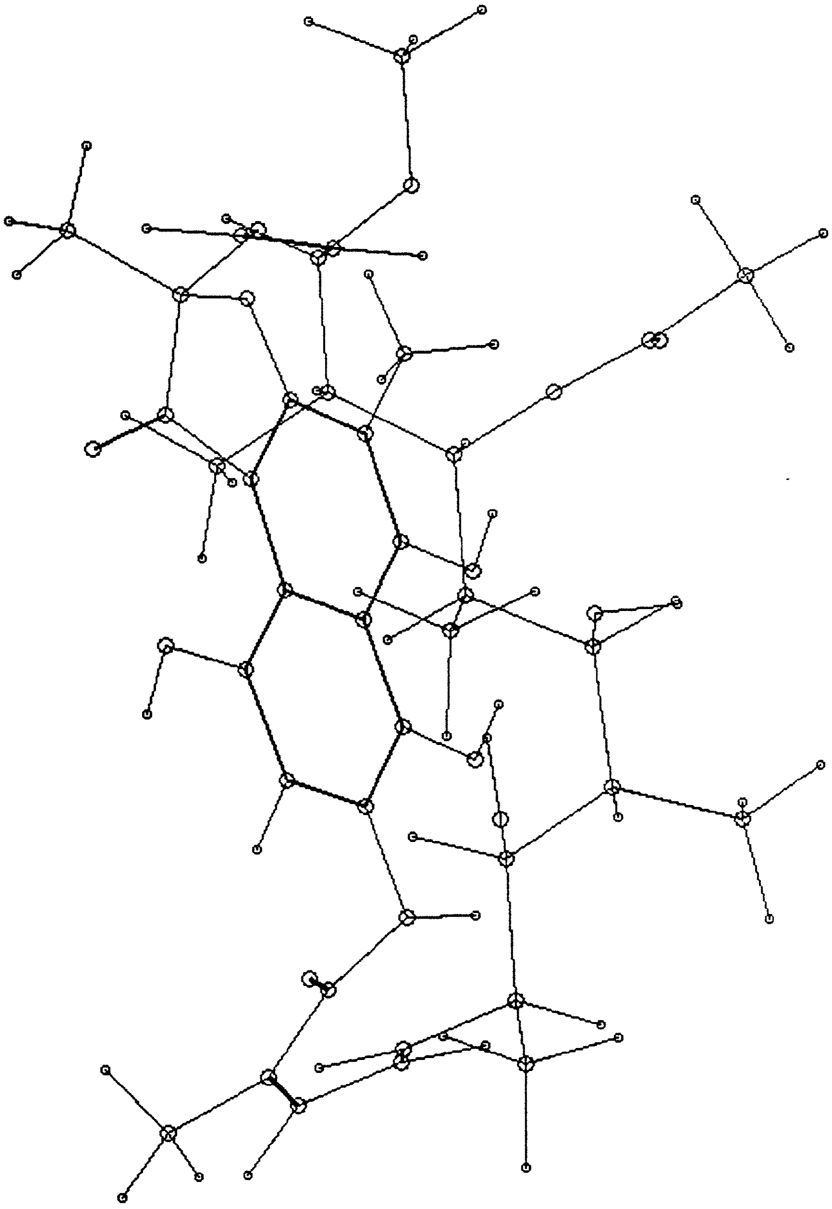


Figure 4-8: The lowest-energy conformation of rifamycin SV with 50.006 KCal/mol.

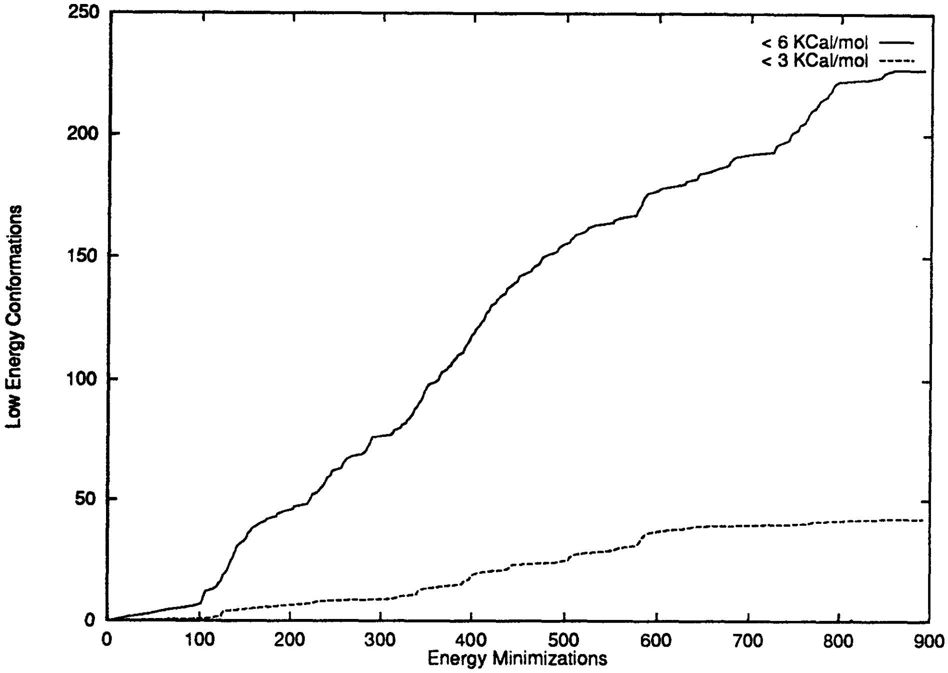


Figure 4-9: Performance of the complementarity algorithm on rifamycin SV.

	Energy Minimizations			
kcal/mol	892	1000	2000	10000
	SUMM			
3				39
6		63		186
	FLEX			
3				42
6		70		305
	Usage-Directed Torsional Monte Carlo Search			
3			0	
6			17	
	Complementarity			
3	42			
6	227			

Table 4.4: Unique Conformers Found versus Energy Minimizations during Conformational Searches of Rifamycin SV.

Chapter 5

Conclusion

This thesis has described a new algorithm, complementarity, for conformational search of macrocyclic molecules. It scans a large number of candidate conformations and minimizes only the promising ones. These candidates can be generated by two operators that construct new conformations from known minima. The candidates have similar bonded-interaction energy as the known minima and possibly lower non-bonded interaction energy. The components of conformers are accessed efficiently from a database. The starting conformations are ordered by energy in a priority queue. On the examples tested, this algorithm is 9 to 11 times faster than the existing methods for large symmetric and asymmetric rings.

There are several reasons for the efficiency of the algorithm.

1. The most important reason is that the operators are able to generate good conformations close to the local minimum. Thus we can use initial energy of the conformations for selection. Low-energy starting conformations increases the chance of finding useful minima and reduces the time of minimization.

2. The database allows the systematic retrieval of the complements of any component. The operators can be applied systematically to generate good starting conformations. There is no redundancy in generation.

3. The priority queue allows us to minimize the starting conformations in order of their energy. Thus lowest energy minima are likely to be found early in the

search.

To further improve the efficiency of conformational search of macrocyclic molecules, we envision the construction of a massive component database. If all common types of components are present in the database, the best starting conformations can be generated and minimized very quickly. This may be the next step of research.

In addition, we have shown that initial energy is a good predictor of minimized energy if the starting conformations of macrocyclic molecules satisfy the bond length and angle constraints. This may also apply to acyclic molecules. We believe that if the starting conformations of acyclic molecules are generated carefully and selectively minimized based on their initial energy, their low-energy conformations can also be found efficiently.

Bibliography

[All77] Norman L. Allinger. MM2. a hydrocarbon force field utilizing v_1 and v_2 torsional terms. *Journal of the American Chemical Society*, 99(25):8127, December 1977.

[Aro83] S. K. Arora. Correlation of structure and activity in ansamycins, molecular structure of sodium rifamycin SV. *Molecular Pharmacology*, 23:133–140, 1983.

[AW93] Zhuming Ai and Yu Wei. Knowledge based method for building molecular models. *Journal of Chemical Information and Computer Sciences*, 33(4):635–638, 1993.

[BBNE93] A. J. Bryskier, J.-P. Butzler, H. C. Neu, and P. M. Tulkens Ed. *Macrolides, Chemistry, pharmacology and clinical uses.* Arnette Blackwell, Paris, France, 1993.

[BCM93] Charles E. Bugg, William M. Carson, and John A. Montgomery. Drugs by design. *Scientific American*, 269(6):92, December 1993.

[CGS89] George Chang, Wayne E. Guida, and W. Clark Still. An internal coordinate Monte Carlo method for searching conformational space. *Journal of American Chemical Society*, 111:4379–4386, 1989.

[Dep90] Department of Chemistry, Columbia University, New York, NY 10027. *BatchMin Documentation*, April 1990.

[DLP87] Daniel P. Dolata, Andrew R. Leach, and Keith Prout. Wizard: AI in conformational analysis. *Journal of Computer-Aided Molecular Design*, pages 73–85, 1987.

[FH77] Dino R. Ferro and Jan Hermans. A different best rigid-body molecular fit routine. *Acta Crystallographica*, A33:345–347, 1977.

[GO89] Hitoshi Goto and Eiji Osawa. Corner flapping: A simple and fast algorithm for exhaustive generation of ring conformations. *Journal of American Chemical Society*, 111:8950–8951, 1989.

[GO93] Hitoshi Goto and Eiji Osawa. An efficient algorithm for searching low-energy conformers of cyclic and acyclic molecules. *Journal of the Chemical Society-Perkin Transactions II*, pages 187–198, 1993.

[GS70] Nobuhiro Go and Harold A. Scheraga. Ring closure and local conformational deformation of chain molecules. *Macromolecules*, 3(2):178–187, March 1970.

[GW92] F. Guarnieri and S. R. Wilson. Simulated annealing of rings using an exact ring-closure algorithm. *Tetrahedron*, 48(21):4271–82, 1992.

[KG93] Istvan Kolossvary and Wayne C. Guida. Torsional flexing: Conformational searching of cyclic molecules in biased internal coordinate space. *Journal of Computational Chemistry*, 14(6):691–698, 1993.

[LS88] Mark Lipton and W. Clark Still. The multiple minimum problem in molecular modeling. tree searching internal coordinate conformational space. *Journal of Computational Chemistry*, 9(4):343–355, 1988.

[MJ93] D. B. McGarrah and R. S. Judson. Analysis of the genetic algorithm method of molecular-conformation determination. *Journal of Computational Chemistry*, 14(11):1385–1395, 1993.

[MOI90] Stephen L. Mayo, Barry D. Olafson, and William A. Goddard III. Dreiding: A generic force field for molecular simulations. *Journal of Physical Chemistry*, 94:8897–8909, 1990.

[Omu84] Satoshi Omura. *Macrolide antibiotics, chemistry, biology, and practice.* Academic Press, Orlando, Florida 32887, 1984.

[Sau87] Martin Saunders. Stochastic exploration of molecular mechanics energy surfaces. hunting for the global minimum. *Journal of American Chemical Society*, 109:3150–3152, 1987.

[SHW$^+$90] Martin Saunders, K. N. Houk, Yun-Dong Wu. W. Clark Still, Mark Lipton, George Chang, and Wayne C. Guida. Conformations of cycloheptadecane. a comparison of methods for conformational searching. *Journal of American Chemical Society*, 112:1419–1427. 1990.

[SSS$^+$93] B. K. Shoichet, R. M. Stround, D. V. Santi. Irwin D. Kuntz, and K. M. Perry. Structure-based discovery of inhibitors of thymidylate synthase. *Science*, 259(5100):1445–1450, 1993.

[Tes79] Bernard Testa. *Principles of organic stereochemistry.* Marcel Dekker, inc., 1979.

[UM93] R. Unger and J. Moult. Finding the lowest free-energy conformation of a protein is an np-hard problem–proof and implications. *Bulletin of mathematical biology*, 55(6):1183–1198, 1993.

[WJW$^+$83] Paul K. Weiner, Salvatore Profeta Jr., Georges Wipff, Tim Havel, Irwin D. Kuntz, Robert Langridge, and Peter A. Kollman. A distance geometry study of ring systems. *Tetrahedron*, 39(7):1113–1121, 1983.