Global Energy Minimization of Small Molecules Combining Constraint Logic Programming and Molecular Mechanics

Darko Zupanič[1], Milan Hodošček[2], Nada Lavrač[1]*, Igor Mozetič[3]

[1] Jožef Stefan Institute, Department of Intelligent Systems
Jamova 39, 1000 Ljubljana, Slovenia

[2] National Institute of Chemistry,
Laboratory of Molecular Modeling and NMR Spectroscopy
Hajdrihova 19, 1000 Ljubljana, Slovenia

[3] Center for Applied Mathematics and Theoretical Physics
University of Maribor, Krekova 2, 2000 Maribor, Slovenia

Abstract

The paper presents an approach to molecular energy minimization using Constraint Logic Programming (CLP) as a preprocessor to a molecular mechanics minimization program such as CHARMM. The use of CLP enables the definition of simple constraints that

*corresponding author, e-mail:nada.lavrac@ijs.si

sufficiently describe relations in a molecule, thus limiting the search space of the globally optimal solution. The approach was used on the problem of minimizing the energy of the alanine dipeptide and N-methylalanylacetamide. An approximate 3-D structure produced by a CLP system was used as an initial structure for the CHARMM program that considers all relations in such a structure and performs energy optimization. With this initial structure, computed in about one minute of CPU time, the optimal solution (global minimum) was found in only few seconds, while calculating the same solution with CHARMM alone took 15 hours. All calculations were performed on HP/735.

1 Introduction

Flexible molecular systems such as oligopeptides posses many local minima on the potential energy surface. Conventional algorithms, for example the steepest descent [15], conjugate gradients [16], Newton-Raphson, and others used for the minimization of the energy function on these surfaces usually find the nearest minimum from the initial molecular structure. Conformational studies of dipeptides [18] suggest that the lowest energy minima are the most populated ones. The information about the global energy minimum is also very useful in a variety of studies such as harmonic analysis of the vibrational frequencies[19]. The structure of the global minimum can be used for the evaluation of the quality of the parameterization of the empirical force field.

In the past there have been several attempts to deal with the global minimum problem (see [17, 21, 22] and references therein) but none is efficient enough to be routinely used in computational chemistry.

The motivation for this paper is to considerably speed up exhaustive conformational search for small molecules. The idea is to use Constraint Logic Programming (CLP) first, to find an

approximate, crude solution to the molecular energy minimization problem. This solution then provides an initial structure for a classical energy minimization system like CHARMM (Figure 1). The main advantage of using CLP is its flexibility in combining symbolic and numerical constraints, and the possibility to use different domain-specific constraint resolution strategies within the same framework.

The current implementation permits the use of the algorithm on molecules of up to 30 atoms only. The proposed method was tested on two small molecules, namely the alanine dipeptide and N-methylalanylacetamide shown in Figures 5b and 6b.

2 Problem formulation in CLP

2.1 Constraint Logic Programming

Constraint Logic Programming (CLP, [6, 4]) is a generalization of logic programming [7]. Unification, the basic operation in logic programs, is replaced by a more general mechanism of constraint satisfaction over a specific computation domain. An instance of the general CLP scheme is obtained by selecting a computation domain, a set of allowed constraints and designing a solver for the constraints. CLP combines the advantages of logic programming (declarative semantics, nondeterminism, partial answers) with the efficiency of specialized constraint satisfaction algorithms.

There exist several instances of the general CLP scheme, implemented in e.g., ECL^{i}PS^e [2], or SICStus Prolog [11]. The most common are: CLP($\mathcal{B}$) — a solver for constraints over the $\mathcal{B}$oolean domain, CLP($\mathcal{F}$) — a solver for constraints over $\mathcal{F}$inite domains, and a solver for systems of linear equations and inequalities over the domain of $\Re$eals — CLP($\Re$), or over the domain of rationals — CLP($\mathcal{Q}$).

In this study we used ECL^{i}PS^e [2] which provides an implementation of CLP($\mathcal{F}$) with a large repertoire of available predicates and constraints over finite domains. CLP($\mathcal{F}$) allows for an explicit manipulation of constraints over finite domains through the concept of *domain variables* [10]. Domain variables range over finite sets of atomic values and as more constraints are imposed on the variables these sets become smaller. Constraints are propagated in such a way to ensure some form of local consistency, e.g., *node-*, *arc-*, or *path-(of some length) consistency* [5] and are combined with backtrack search in the reduced solution space. For example, *forward-checking* as introduced in [9] essentially ensures arc-consistency.

2.2 Molecular energy minimization

There are several computer-based approaches to molecular energy minimization. The approach selected in this work can be viewed as a preprocessing step to energy minimization by CHARMM [1]. In this work, the idea is to constrain the space of possible solutions for CHARMM in such a way that the globally optimal solution will easily be reached from an approximate solution determined in preprocessing; CHARMM is then used just for the refinement of the approximate solution proposed by the preprocessor. Being a preprocessor to CHARMM, the proposed CLP approach thus assumes the same potential field (cost function) as defined by CHARMM.

2.2.1 Cost constraints

In the molecular energy minimization problem, there is a cost function which has to be minimized. The cost function is taken from the empirical potential used by CHARMM, which is composed of several terms [1]. In our approach, we consider constant bond lengths and bond angles, and the torsion energy is neglected for simplicity reasons. Consequently, the following

two terms are used for energy computation:

- Van der Waals interaction potential

$$E_{vdW} = \sum_{excl(i,j)=1} \frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}}$$

- electrostatic potential

$$E_{el} = C \sum_{excl(i,j)=1} \frac{q_i q_j}{r_{ij}}$$

where A and B are constants depending on atoms that appear in a pair (i, j) of atoms, C is a constant $C = \frac{1}{4\pi\epsilon_0}$, q_i and q_j are electric charges of atoms, r_{ij} is the distance between atoms, and $excl(i,j) = 1$ when the corresponding atoms are at least three bonds apart (i.e., $excl(i,j) = 0$ if atoms are connected by bonds or bond angles, or if $i \geq j$ in order to eliminate symmetrical computations).

The cost function $E = E_{vdW} + E_{el}$ is formulated in ECL^{i}PS^e[2]. It can be viewed as a function that needs to be minimized by a CLP constraint solver. However, due to its non-linearity, we had to implement a special search procedure for dealing with such complex constraints (see Section 2.3).

2.2.2 Distance constraints

The selected CLP model of a molecule consists of distances of pairs of atoms. The information about bonds and angles is taken from the CHARMM parameter file. The predefined bonds and angles constrain the possible distances of other atoms.

A constraint for two pairs of bonded atoms involves six distances (Figure 2). Simplified constraints between these six distances are all the possible triangular inequalities that can be stated between the two bonds. Since each distance appears in two consecutive bonds in a molecule,

each change of possible values of a distance propagates to all the involved constraints. A constraint that relates six distances is written in the CLP language ECL^{i}PS^e[2] as follows:

```
dists_cnstr( R12, R34, R14, R23, R24, R13 ) :-
  triangle_ineq_cnstr( R12, R23, R13 ),
  triangle_ineq_cnstr( R34, R14, R13 ),
  triangle_ineq_cnstr( R12, R14, R24 ),
  triangle_ineq_cnstr( R34, R23, R24 ).


triangle_ineq_cnstr( Dist1, Dist2, Dist3 ) :-
  Dist1 + Dist2 >= Dist3,
  Dist2 + Dist3 >= Dist1,
  Dist3 + Dist1 >= Dist2.
```

Note that the above variables `Rij` correspond to r_{ij} distances between atoms occurring in the Van der Waals and electrostatic interaction potential cost functions. The first constraint can be read as follows: a constraint between six distances of two pairs of bonded atoms holds if the triangular inequality constraints between triples of specified distances hold. The triangular inequality constraint, relating three distances forming a triangle, is formulated according to the obvious geometrical properties that hold for the edges of a triangle.

A disadvantage of these simple constraints is a wider interval of possible distance values as opposed to exact boundary values of intervals. The advantage of such simple constraints is their computational efficiency.[1] As a consequence, some distance values that are consistent with the

[1]The exact computation of distances would be possible by explicitly considering the given information about the angles; however, in order to reduce the computational complexity, we rather considered more relaxed constraints thus trading exactness for efficiency.

constrains are impossible in real situations. Therefore, the constraints should not be viewed as abstractions of real distance values but rather as their approximations.

2.3 Constraint propagation by CLP

Our implementation uses two types of constraints: arithmetic constraints and user-defined constraints.

Inequality constraints are built-in arithmetic constraints and are completely handled by the ECL^{i}PS^e system. Each inequality constraint consists of three finite domain variables representing distances. An upper and a lower bound of each variable are adjusted to upper and lower bounds of the other two variables. For instance[2], if domains of `Dist1, Dist2` and `Dist3` are `1..5, 1..2` and `4..5` respectively, and an inequality is `Dist1 + Dist2 >= Dist3` then the lower bound of the variable `Dist1` is increased to 2. The domain change causes the other constraints containing the `Dist1` variable to adjust their variables' domains according to the new domain of the variable `Dist1`.

On the other hand, the cost function constraints are user-defined. They are also incorporated into the ECL^{i}PS^e system propagation mechanism, but we had to implement procedure of variables' domains adjustment. Actually, only the terms $E_{vdW_{ij}} = \frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}}$ and $E_{el_{ij}} = C\frac{q_i q_j}{r_{ij}}$ are implemented as user-defined constraints. The summation used in the cost function ($E_{vdW} = \sum_{excl(i,j)=1} E_{vdW_{ij}}$ and $E_{el} = \sum_{excl(i,j)=1} E_{el_{ij}}$) is a built-in arithmetic constraint and is handled in a similar way as the inequality constraints. In the first term, the upper and lower bounds of the variables $E_{vdW_{ij}}$ and r_{ij} must be adjusted, while in the second term the upper and lower bounds of the variables $E_{el_{ij}}$ and r_{ij} must be adjusted. All other items in

[2]According to the lower bound of `Dist3` the sum of `Dist1 + Dist2` must be at least 4. The variable `Dist2` can contribute at most 2 to the sum and therefore, at least 2 must be contributed by the variable `Dist1`. The domain of the variable `Dist1` thus shrinks to `2..5`.

the terms are constants. The second term is linear and relations between variables' domains are quite straightforward. This is not the case in the first term which is non-linear. It must be considered that the function is not monotone. Therefore, not only upper and lower bounds of r_{ij} are computed but also a value of r_{ij} where the function $\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}}$ has the extreme.

A strategy for traversing the search space is branch-and-bound (see Figure 3). First, we select a distance variable with the largest interval of possible values. Afterwards, we halve the interval and continue with two subproblems, each having one half of values of the selected variable. This loop of partitioning the problem into smaller subproblems is repeated until only small intervals for distances remain. Exact distances (an interval with one value only) are almost impossible due to numerical errors.

At every step of partitioning the algorithm decides which subproblem will be solved first. The decision depends on electric charges of atoms associated with a selected distance. If the electric charges have the same sign then a subproblem with an upper half of the values of the selected distance is solved first, otherwise the search algorithm selects a subproblem with a lower half of the values of the selected distance. In principle, all branches of the search space have to be traversed in order to find the minimum energy. However, once an energy of the whole molecule is computed, all branches that lead to higher energies can be pruned. This cut-off energy is updated whenever a new minimum energy is found.

Searching for an optimal solution has an exponential complexity. There is no general way to overcome this complexity and still get an optimal solution. Nevertheless, the search space can be pruned with additional constraints. These constraints can be a result of some knowledge about the molecule (i.e., NMR distances), or energy bounds got from some heuristic algorithm which can highly reduce the number of 3-D structure candidates. The main advantage of our approach is that it is easy to incorporate various types of constraints into the CLP formulation

of the problem.

Our approach currently uses only the energy upper bound constraint. The size of molecules which can be solved with such search space pruning is thus limited to about 30 atoms.

3 Case studies

3.1 Alanine dipeptide

An alanine dipeptide consists of two alanine amino acids (Figure 4). There are altogether 23 atoms. Each of the 22 bonds has a known distance, and each of the 39 triples of connected atoms has a known angle.

A model of the alanine dipeptide, built from the distances of each pair of atoms, has 253 distances. Each pair of atoms determines the constants which are used in the cost function.

3.1.1 An approximate solution by CLP

The CLP system computes an approximate solution which minimizes the cost function. The output of the CLP program are 253 distances (note that some of these are known in advance, i.e., distances of bonded atoms). These distances are then input to the standard restraints facility in the CHARMM program. Since the number of distances is complete the program easily finds the unique solution to satisfy all the distances. The structure which satisfies CLP distances is shown in Figure 5a. CLP needed about one minute of CPU time on HP/735 to find the approximate solution.

It is possible to design a procedure which would output the approximate solution directly from the CLP program by a transformation from the internal to Cartesian coordinates. This facility would eliminate the additional step of preparing the input for the standard restraints in

the CHARMM program.

3.1.2 The results of refinement using CHARMM

The coordinates, representing an approximate solution, are input to CHARMM [1] which computes, without any constraints, an actual 3-D structure of the alanine dipeptide, shown in Figure 5b.

Alanine dipeptide is a small molecule, so we can evaluate its 3-D structure by using a classical optimization approach by exhaustive conformational search. Using 3 torsional angles there are $36 \times 36 \times 36$ initial structures and all the resultant minima are presented in Table 1. To prove that the initial torsional angle step size of 10° is enough we repeated the calculation using the step size of 5°. The number and values of minima found remained the same as reported in Table 1. This result can be used for comparing our approach with other known approaches. With CLP we got the approximate structure within one minute computation time and only several additional seconds are necessary to get the proper 3-D structure with CHARMM. Running from scratch (without proposing an approximate initial structure), it took CHARMM 15 hours to find the optimal solution. The RMS difference between the approximate CLP structure from Figure 5a and the refined one in Figure 5b which represents the global minimum, is 0.88 Å. The approximate structure is significantly different from the global minimum, but they both rather lie in the region of the potential energy surface which has no high barrier between them. Therefore, the CHARMM ABNR [1] minimizer reached the final global minimum from the approximate structure without difficulties.

3.2 N-methylalanylacetamide

A N-methylalanylacetamide has altogether 22 atoms and 21 bonds. The global minimum of this structure was determined previously in reference [20] using the same parameters as in our study. The CLP structure and the corresponding refined CHARMM structure are shown in Figures 6a and 6b, respectively. The latter is identical to the one reported in [20]. The RMS difference between the approximate CLP structure and the refined CHARMM structure is 1.12 Å.

4 Conclusions

The paper proposes a method for speeding-up molecular energy minimization using a molecular mechanics minimization program CHARMM. For small molecules, substantial speed-up can be achieved by first computing an approximation to the global energy minimum, and then running CHARMM for the refinement of the approximate solution. This study proposes an approach to approximate minimal energy computation using Constraint Logic Programming (CLP). The CLP system succeeded to find approximate structures of the alanine dipeptide and the N-methylalanylacetamide which are close to their global minima so that a standard numerical minimization algorithm CHARMM is then able to locate the precise global minima on the potential energy surfaces. In both case studies, preprocessing using CLP resulted in a speed-up of three orders of magnitude (1 minute vs. 15 hours).

The method presented in this paper can easily be modified to other empirical potential functions used in program packages such as GROMOS [12], AMBER [13], DISCOVER [14] and others. Due to the large number of energy calculations the method is probably not suitable for ab initio methods using quantum potentials.

Due to the computational complexity of constraint propagation, the proposed CLP approach

is currently limited to small molecules of up to approximately 30 atoms. Extensions of the method to larger systems are currently being investigated. Heuristic search is applied instead of the limited exhaustive search as used in this study. Thus, in addition to the approximate molecular structure, the energy minimum is approximated as well. First experiments on a cyclic sextapeptide indicate that such an approximation also provides a good starting point for the CHARMM energy minimization. The estimated size of tractable problems can thus be increased to about 100 atoms.

Acknowledgement

We acknowledge the support of the Ministry of Science and Technology of Slovenia.

References

[1] Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M.: CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamic Calculations. *Journal of Computational Chemistry*, 4(2), pp. 187–217, John Wiley & Sons, Inc., 1983.

[2] Project Team. ECL^{i}PS^e 3.5.2, ECRC Common Logic Programming System. User Manual, January 1996, ©International Computers Limited and ECRC GmbH 1992.

[3] Carl Branden & John Tooze, "Introduction to Protein Structure", Garland Publishing, Inc., New York and London, 1991.

[4] Cohen, J. Constraint logic programming languages. *Communications of the ACM 33 (7)*, pp. 52-68, 1990.

[5] Guesgen, H. W., Hertzberg, J. Some fundamental properties of local constraint propagation. *Artificial Intelligence 36*, pp. 237-247, 1988.

[6] Jaffar, J., Lassez, J.-L. Constraint logic programming. *Proc. 14th ACM Symp. on Principles of Programming Languages*, pp. 111-119, Munich, 1987.

[7] Lloyd, J.W. *Foundations of Logic Programming*. Springer, 1987.

[8] Schulze-Kremer, S.: *Molecular Bioinformatics: Algorithms and Applications*. Walter de Gruyter, 1995.

[9] Van Hentenryck, P. (1989). *Constraint Satisfaction in Logic Programming*. The MIT Press, Cambridge, MA.

[10] Van Hentenryck, P., Dincbas, M. (1986). Domains in logic programming. *Proc. 5th Natl. Conference on Artificial Intelligence, AAAI-86*, Philadelphia, PA, Morgan Kaufmann, pp. 759-765.

[11] SICStus Prolog user's manual, Release 3, Swedish Institute of Computer Science, Kista, Sweden, 1995.

[12] W. F. van Gunsteren, and H. J. C. Berendsen, Groningen Molecular Simulation (GROMOS) Library Manual, Nijenborg 16, Groningen, The Netherlands.

[13] D. A. Pearlman, D. A. Case, J. C. Caldwell, G. L. Seibel, U. C. Singh, P. Weiner, and P. A. Kollman, AMBER 4.0, University of California, San Francisco, 1991.

[14] DISCOVER/CFF91 forcefield, InsightII User Guide, October 1995. San Diego: Biosym/MSI, 1995.

[15] M. Levitt, S. Lifson, J. Mol. Biol., 46, 269 (1969).

[16] R. Fletcher and C. M. Reeves, Comput. J., 7, 149 (1964).

[17] H. Gotō, E. Ōsawa, J. Molec. Struct. (THEOCHEM), 285, 157 (1993).

[18] S. S. Zimmerman, H. A. Scheraga, Biopolymers, vol. 16, 811-843 (1997)

[19] B. R. Brooks, D. Janežič, M. Karplus, J. Comp. Chem, 16, 1522-1542 (1995)

[20] B. M. Pettitt, and M. Karplus, J. Am. Chem. Soc., 107, 1166, (1985).

[21] J. Kostrowicki, H. A. Scheraga, J. Phys. Chem., 96, 7442 (1992).

[22] J. Mestres, G. E. Scuseria, J. Comp. Chem., 16, 729 (1995).

Table caption

Table 1 All the minima found by exhaustive search at 10° step size for each of the three angles. Angles φ, ψ, ω have the usual meaning.

Figures captions

Fig. 1 Interaction between CLP and CHARMM.

Fig. 2 The distances between atoms that constitute two bonds: A1-A2 and A3-A4.

Fig. 3 Partitioning a problem into subproblems by halving variable domains and updating the cutoff minimum energy.

Fig. 4 The schematic structure of the alanine dipeptide.

Fig. 5 The structure of the alanine dipeptide (a) approximate found by CLP and (b) refined by CHARMM.

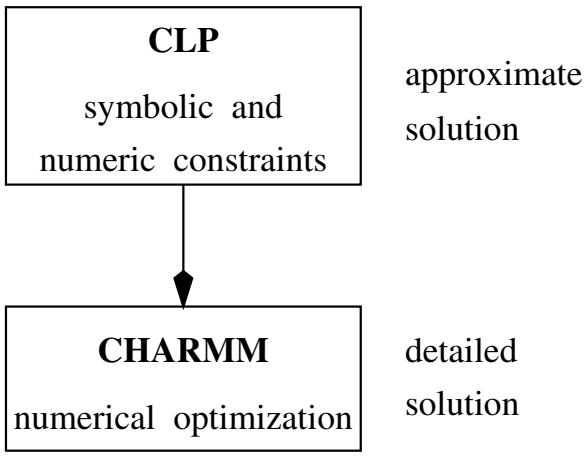
Fig. 6 The structure of the N-methylalanylacetamide (a) approximate found by CLP and (b) refined by CHARMM.

Table

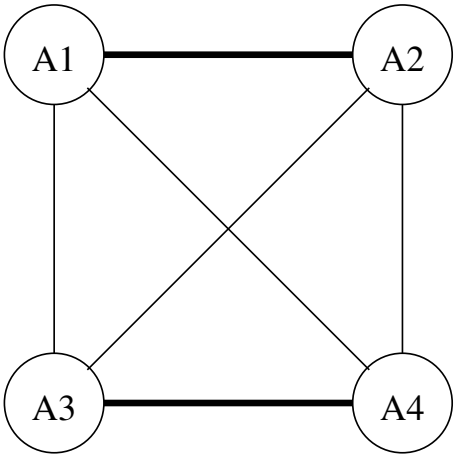
Table 1. All the minima found by exhaustive search at 10° step size for each of the three angles. Angles φ, ψ, ω have the usual meaning.

Angles [°]			Energy
ψ	ω	φ	kcal/mole
81.8	-1.7	-43.3	-54.0511*
-61.8	9.5	30.3	-49.0296
-74.7	111.3	-72.8	-42.5722
83.3	-104.9	64.7	-33.8062
96.9	-96.8	62.8	-33.5976
-73.3	-8.9	-165.8	0.7919

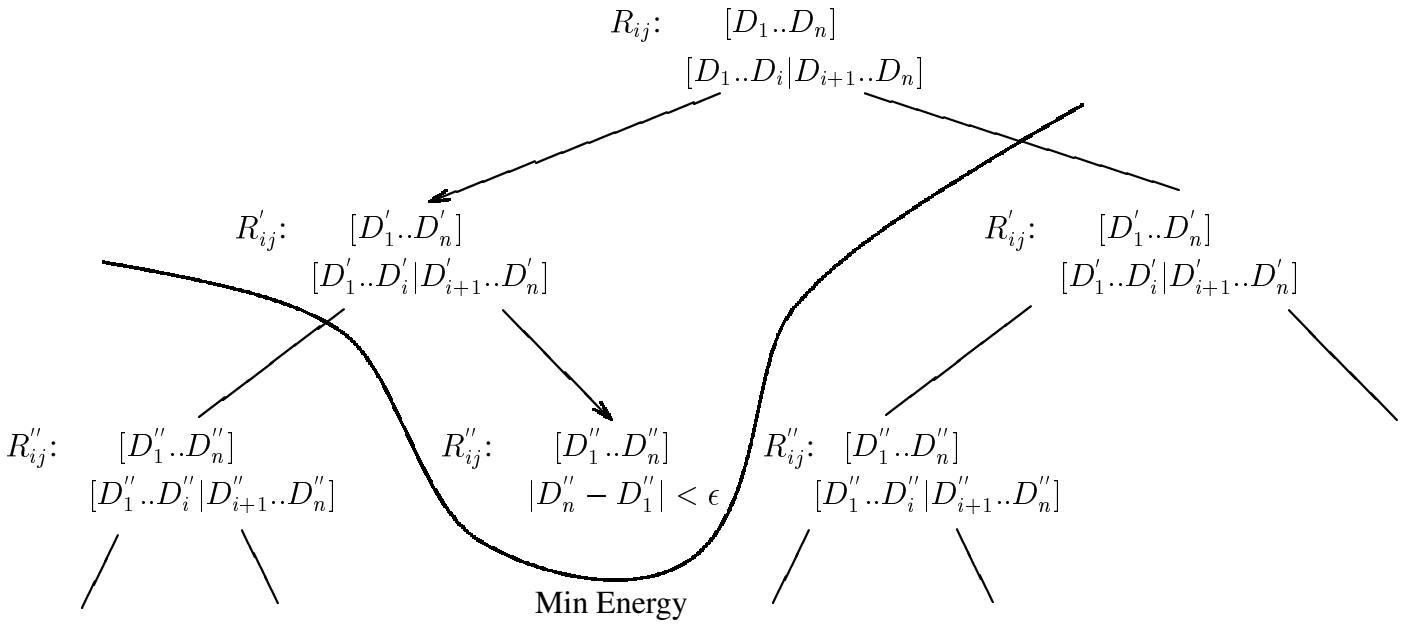
* This structure is shown in Figure 5b.



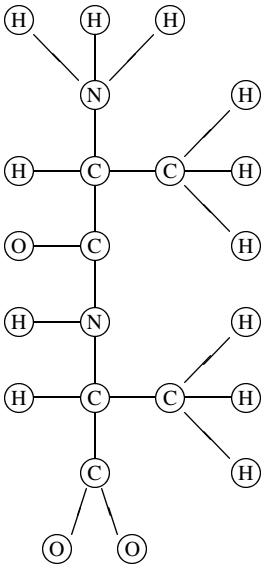
D. Zupanič, et al **Fig. 1**



D. Zupanič, et al **Fig. 2**

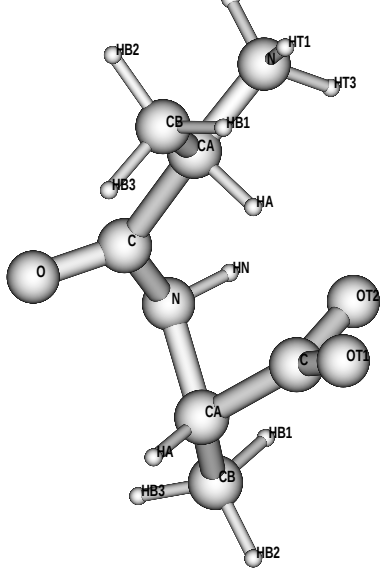


D. Zupanič, et al **Fig. 3**

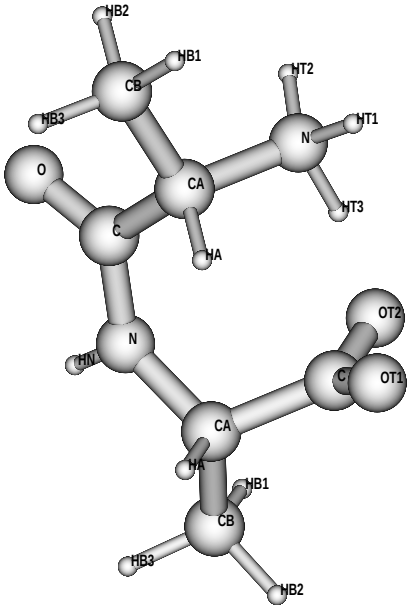


D. Zupanič, et al **Fig. 4**

(a)

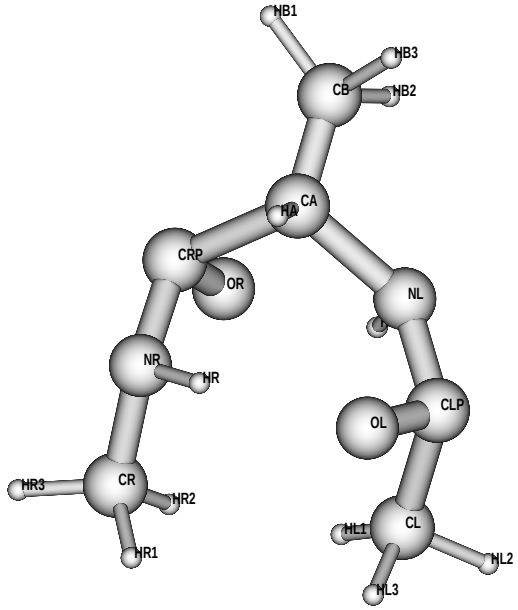


(b)

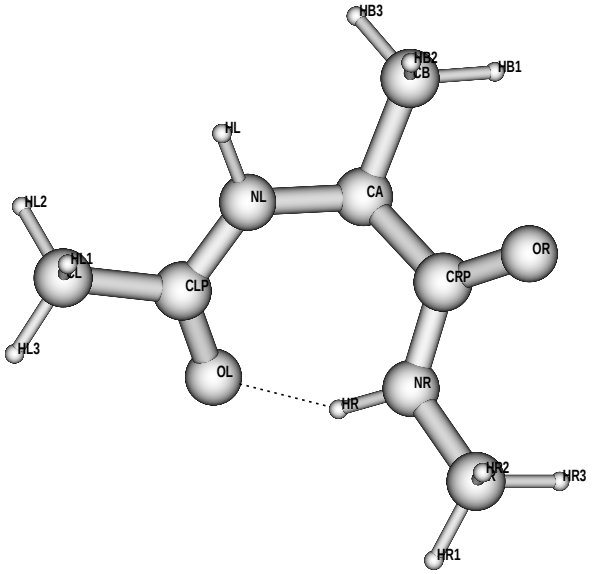


D. Zupanič, et al **Fig. 5**

(a)



(b)



D. Zupanič, et al **Fig. 6**