Theoretical Prediction of Crystal Structure by Global Optimization of Potential Energy

Yelena Arnautova and Jaroslaw Pillardy

Script for hands-on session

Ethylene molecule has been chosen for this hands-on session, because it is a small, highly symmetric organic molecule, and all calculations with it are very fast. The global optimization for this molecule is not complicated, therefore the full prediction is possible in a very short time. For more complicated systems (less symmetric molecules containing heteroatoms) global optimization runs are longer by orders of magnitude (e.g., for imidazole with 4 molecules in the unit cell, full global optimization required 10 hours, using 30 processors of the Pentium III 650 MHz computer cluster).

The single global optimization run is designed to find a global minimum for a given number of molecules in the unit cell (Z). In order to predict the value of Z one should carry out several global optimization runs for different Z's and choose the Z, for which the energy per molecule for the global minimum is the lowest. We suggest carrying out global optimization runs for four values of Z=2, 3, and 4.

All files for hands-on session are located in subdirectory *tutorial*. There are 6 subdirectories there: local, 01, …, 05 corresponding to Part 0, …, Part 6 of this tutorial, respectively.

To run a job on Velocity cluster *i.bat*, *crystalg.bat* and *ccp.bat* files has to be changed. XXX should be replaced by the name of your account. YYY is the name of your current directory.

i.bat

```
REM CCS account = XXX
REM CCS type = batch
REM CCS nodes = 2
REM CCS requirements = 2@vplus
REM CCS minutes = 60
h:
cd \users\XXX\tutorial\YYY
call MachineMaker.bat
set MPI_COMM=TCT
mpirun –np 2 ccp.bat
time /T >log
date /T >>log
mpirun –np 4 crystalg.bat ethlen
time /T >>log
date /T >>log
Ccrelease
```

crystalg.bat

```
set BIN=t:\XXX
set PREFIX=%1
h:
cd \users\XXX\tutorial\YYY
%BIN%\crystalg.exe
```

ccp.bat

```
mkdir t:\XXX
copy h:\users\lenaa\crystalg\src\crystalg.exe t:\XXX
```

Part 0. Preparing input files for local and global optimizations, obtaining reference structure.

The input file for the global optimization program (*ethlen.inp*) consists of following parts (see Fig 1):

- title
- two sets of optimization parameters
- information about molecular geometry, atom types and connectivity (bonds)

Information about general and global optimization parameters and the structure of an input file is provided in Appendix A.

Molecular geometry may be obtained in different ways. One of them is to prepare the molecule using any molecular modeling package, and then put atom coordinates in proper format into the input file. It is also possible to use experimental molecular geometry taken from the Cambridge Structural Database (CSD).

For both global and local optimization runs the files with the force field parameters are necessary. The parameter files are *latf.data* and *tors.data. latf.data* file contains parameters of '6-12' or '6-exp' potential functions. *tors.data* file contains parameters for the Fourier series describing the torsional potential.

The files with potential parameters are already prepared, we have chosen the AMBER force field. For details about force field parameters' format see Appendix B.

Preparing reference structure: local minimization of the experimental structure.

In order to validate the crystal structure prediction we should prepare the experimental structure of the crystal minimized locally with the force field we have chosen. Program *crystalg* has options allowing producing such a structure using an additional input file **.coord* containing information about the crystal structure obtained by X-ray or neutron diffraction (from CSD). The file **.coord* contains parameters of the experimental unit cell, fractional or Cartesian coordinates of atoms and symmetry operators of the experimental space group (see Appendix B). The *ethlen.coord* file is already prepared and is being stored in *local* directory.

- go to *local* directory
- Set the parameter MODE in *ethlen.inp* file to 1 and run *ccsubmit i.bat*.
- Rename file *ethlen_W000.out* into *ethlen_exp.out*. To visualize results of your calculations FTP files *ethlen_exp.mol2* and *ethlen.mol2* to your local workstation and use *molmol* for visualization.

Part 1. Global optimization using initial parameters.

- Copy the file *ethlen.inp* prepared in Part 0 to your current directory (you should change directory to 01 for Part 1).
- Set parameter MODE to 0 in order to run global optimization. Check if all global optimization parameters are set to proper values:

- MAXFAM 30
- MAXSTR 150
- NMCMF 1
- INIMIN 10
- KT_MCMF1, KT_MCMF2 1.0; 1.0
- NTHREAD 1
- MAXSTEPMCMF 200
- MAXFAMUSE 200
- NRESET 200
- number of molecules in the unit cell 2

- Run *ccsubmit i.bat*. Watch progress of calculations: each newly found lowest-energy minimum geometry is stored in *ethlen@ZZZ.mol2* (ZZZ is the number), general information about progress in calculation current set of minima is shown in *ethlen.info and ethlen.stat*. The program is fully restartable, all necessary information for restarting is stored in *ethlen.rst.*
- Compare results of the global optimization with the reference structure. Has the experimental structure been found as the global minimum of the AMBER potential for Z=2?

For this extremely simple system the minimized experimental structure has been found as a global minimum even with the simplest set of parameters. However, it is not usually the case for more complicated systems. It is not possible to tell a priori if a given set of global optimization parameters is suitable for a given crystal, therefore few runs using different values of parameters may be necessary. Usually one starts with the set allowing very fast calculations, and then increases the numerical effort until the result (the lowest-energy minimum found) does not change. This is so-called *tuning-up*. In the case of ethylene, we will demonstrate the effectiveness of the search method by finding not only the global minimum of the ethylene crystal but also a complete set of the lowest-energy states. The completeness of this set may be evaluated by comparing the lowest-energy minima obtained during independent runs.

Part 2. Tuning-up the global optimization parameters: temperature, MAXFAMUSE and NRESET.

- Copy the file *ethlen.inp* from Part 1 to your current directory (you should change directory to 03 for Part 3).
- Change *KT_MCMF1* to 0.05, *KT_MCMF2* to 0.50, *MAXFAMUSE* to 25 and *NRESET* to 50. Run *ccsubmit i.bat*. Watch progress of calculations.
- Is the set of minima found different from the previous run?

Part 3. Tuning-up the global optimization parameters: focusing

- Copy the file *ethlen.inp* from Part 2 to your current directory (you should change directory to 03 for Part 3).
- Change *NMCMF* to 2, *MAXSTEPMCMF* to 100, *RCUT1S* to 5.0, *RCUT2S* to 3.0. Run *ccsubmit i.bat*. Watch progress of calculations.
- Is the set of minima found different from the previous run?

Part 4. Global optimization for four molecules in the unit cell (Z=4).

The global optimization for Z=4 takes about 1.5 hours, when using the input file from Part 3 and MAXSTEPMCMF=200. Therefore we have already carried out these calculations and the output files are placed in the directory 04.

The global optimization run with doubled number of molecules in the unit cell (Z=4 compared to Z=2) validates the prediction obtained previously. The minimum being the representation of the global minimum structure for smaller Z (Z=2) with the number of molecules in the unit cell doubled should be found as a global minimum in the run with larger Z (Z=4) in order to confirm the prediction for Z=2.

Unfortunately, in the case of ethylene with Z=4 the structure corresponding to the minimized experimental structure (Z=2) was found as a minimum number 2 (see output file *ethlen_W000.out*). The unit cell of this structure is equivalent to two unit cells for Z=2 concatenated together along the c axis. This shows that the AMBER potential is able to predict the crystal structure of ethylene as one of lowest energy minima.

(Please remember that program prints out ***energy per unit cell;*** *in order to compare energies you have to divide them by Z.)*

Part 5. Real-world example.

Results of the global optimization for PHYPHM are placed in directory 05. These calculations took about 18 hours using 30 processors of Pentium III 650 MHz computer cluster ("Matrix"). The minimized experimental structure has been found as global minimum, showing that the AMBER force field may be used for crystal structure calculations for PHYPHM. Progress of calculations is shown in *phyphm.info* and *phyphm.stat*. Twenty lowest-energy minima are stored in *phyphm_final_XXX.mol2 files*, transfer them to your local workstation and visualize using *molmol*.

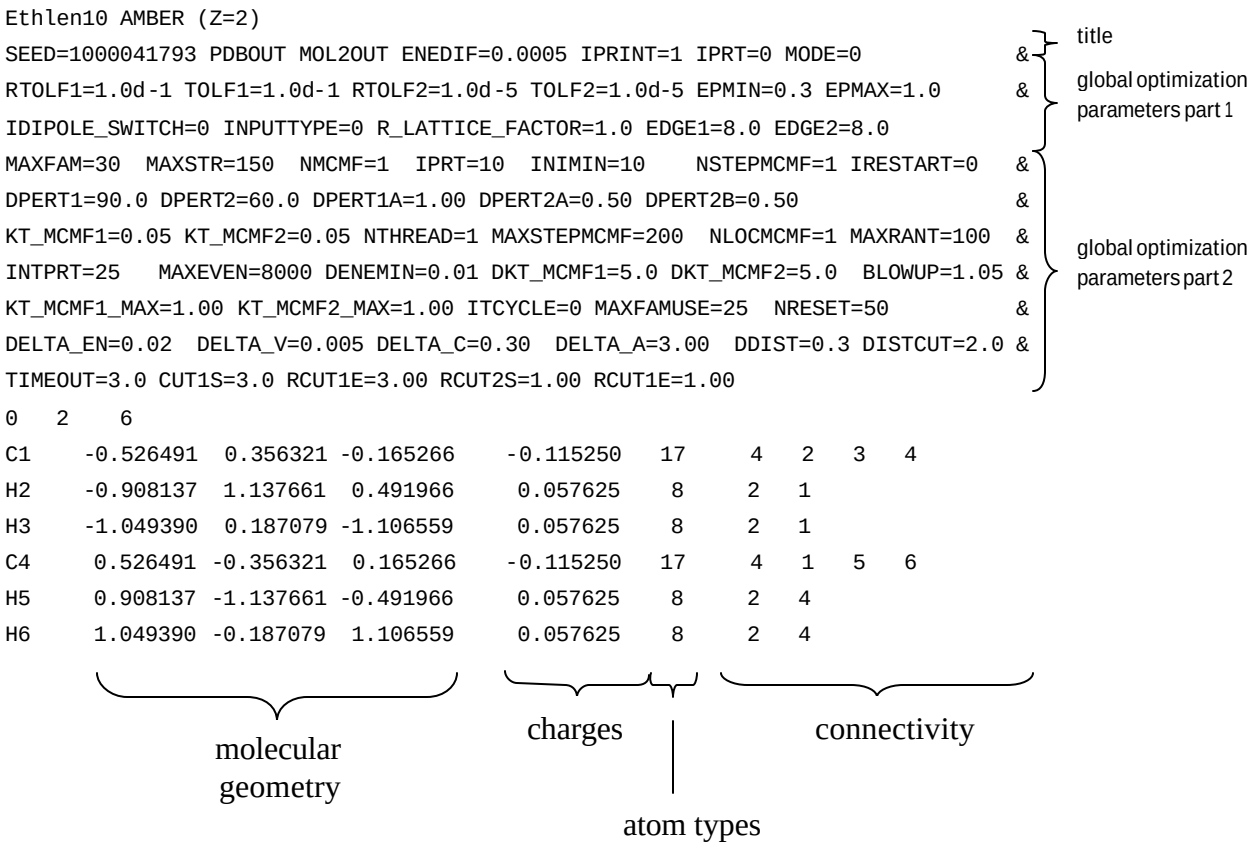

Fig 1. Input file

Appendix A. Ethlen.coord file

```
ETHLEN10                                                          } title
CELL  P     4.626   6.620   4.067  90.000  94.390  90.000        } parameters of the unit cell
f                                                                 } type of coordinates
C1          -0.11656   0.05382  -0.04075                         }
H2          -0.18690   0.16980   0.11850                         }
H3          -0.24300   0.02870  -0.26890                         } molecular geometry
C4           0.11656  -0.05382   0.04075                         }
H5           0.18690  -0.16980  -0.11850                         }
H6           0.24300  -0.02870   0.26890                         }
 #  1    0                                                       }
 #  2    1    5    R    R    T    T    T                         }
    3                                                             }
    x   0.5        y   0.5        z   0.5                       } symmetry operations
    2                                                             }
    x             z                                               }
    0                                                             }
```

Line 1: Title

Line 2: Lattice type (' ' or 'P' for primitive, 'F', 'C', 'I', 'R', etc. for centered), and starting lattice constants (a, b, c, alpha, beta, gamma).

Line 3: letter indicating whether the experimental crystal coordinates are Cartesian ('C' or 'c') or fractional ('F', 'f', or blank).

Lines 4ff: Fractional or Cartesian coordinates of all atoms in each molecule in the asymmetric unit, as given in or deduced from experimental data.

Next line: # of molecule (=1) in (' #',i3,4x,'0') format.

Next line: # of next molecule, # of molecule to which it is related by symmetry operations, # of symmetry operations, and type of each operation (T=translation, R=reversal of sign, P=permutation of coordinates) in (' #',i3,2i5,9(4x,a1)) format. Symbols for the symmetry operations must be in the order in which the operations are to be performed.

Next line: # of translations, in i5 format.

Next line: name of axis (x, y or z) and fractional translation, for each translation, in (5(4x,a1,f10.5)) format. If there are no translations, this line is omitted.

Next line: # of reversals, in i5 format.

Next line: name of coordinate (x, y or z) for reflection, in (3(4x,a1,10x)) format. If there are no reversals, this line is omitted.

Next line: # of permutations, in i5 format.

Next line: names of coordinates (x, y or z) whose sum replaces the x coordinate, in (3(4x,a1,10x)) format.

Next line: names of coordinates (x, y or z) whose sum replaces the y coordinate, in (3(4x,a1,10x)) format.

Next line: names of coordinates (x, y or z) whose sum replaces the z coordinate, in (3(4x,a1,10x)) format. If there are no permutations, these lines are omitted. Repeat lines 2 and those that follow for each molecule of the same type in the unit cell. If the asymmetric unit contains more than one type of molecule, repeat all lines for each type.

Appendix B. General and global optimization parameters

The name of the input file defines names of other files used, they are created by adding appropriate suffixes to the core part of the input file name. Input file must be named according to the template _name_.inp, where _name_ is core name (e.g. ethene.inp).

Record 1: Title (1 line - character - free format)
Record 2: Program parameters/ global optimization parameters part 1
This record contains arbitrary number of 80-character lines terminated with '&', the first line shorter than 80 characters or not terminated with '&' ends this record. The record consists of expressions VAR_NAME1=VALUE or VAR_NAME separated by spaces. In the first case variable VAR_NAME1 is set to value VALUE, in the second case logical vaiable VAR_NAME2 is set to .true.
Record 3: Global optimization parameters part 2
Structure of the Record 2 is identical as for Record 1.

Detailed description of all variables and parameters from Record 1 & 2 is in the next section.

Record 4: Parameter defining if
number of molecules in the unit cell, number of atoms in the molecule, number of torsional angles.
Record 6: Geometry of a molecule including atom type and atom number; atomic Cartesian coordinates x,y,z; charge in electronic units; atom type number according to the chosen force field; connectivity information (number of atoms directly connected to the given atom plus 1, list of these atoms).

III. Variables and parameters.

A. Record 2

SEED (double) random number initialization
PDBOUT (logical) produce output in PDB format
MOL2OUT (logical) produce output in MOL2 format
? ENEDIF (double) when energies of two structures differ less the ENEDIF they are treated as identical
IPRINT (integer) level of output from the program (except local minimization), 0-6
IPRT (integer) level of output from the local minimization (SUMSL)
MODE (integer) type of run:
0 - global optimization
1 - local minimization of an experimental structure (see tutorial for more details and examples)
2 - local minimization (starting structure must be provided in file *.rev2test)
3 – comparison of the structures from *.rev2test
RTOLF1 (double) local minimization accuracy parameters used during the global search
TOLF1 (double)

RTOLF2 (double) local minimization accuracy parameters used at the final minimizations of the global optimization procedure
TOLF2 (double)

EPMIN (double) parameter defining a minimum value for randomly generated unit cell parameters a, b,and c (randomly generated unit cell parameters will depend on size of the molecule and number of molecule in the unit cell).
EPMAX (double) parameter defining a maximum value for randomly generated parameters unit cell parameters a, b, and c.
IDIPOL_SWITCH (integer) 0 - no unit cell dipole moment correction is added
1 - spherical unit cell dipole moment correction is added

INPUTTYPE (integer) 1 – if special rules are used for calculating heteroatomic potential parameter; 0 – if no rules are used. In the current version of the program the rules are geometric average for the parameters A_{ij}, B_{ij} and e^0_{ij} and arithmetic average for the parameters C_{ij} and r^0_{ij}.
R_LATTICE_FACTOR (double) parameter defining a maximum possible value of unit cell parameters (EPMIN and EPMAX are multiplied by R_LATTICE_FACTOR). It

	is calculated as $1.5n_{mol}r_{mol}$, where n_{mol} is number of molecules in the unit cell and r_{mol} is radius of the molecule.
EDGE1	parameter used in cubic spline and defining the distance at which lattice E becomes equal to 0. Parameter EDGE1 is used during global search
EDGE2	same as EDGE2.Parameter EDGE2 is used for final minimizations and local minimization of experimental structure.

B. Record 3

MAXFAM	Maximum number of families
MAXSTR	Maximum number of structures
NMCMF	parameter defining number of step for geometric focusing
IPRT	Printout level (0-minimal)
INIMIN	Number of randomly generated structures (initialization)
NSTEPMCMF	Parameter regulating ratio of local/global moves
IRESTART	Restart if equal 1, fresh start if equal 0
DPERT1	Large-scale perturbation of rotations
DPERT2	Small-scale perturbation of rotations
DPERT1A	Large-scale perturbation of translations
DPERT2A	Small-scale perturbation of translations
DPERT2B	Perturbation of unit cell
DPERT1T	Perturbation of torsional angles
DPERT2T	Perturbation of torsional angles
KT_MCMF1	Reduced temperature for interfamily Metropolis criterion
KT_MCMF2	Reduced temperature for intrafamily Metropolis criterion
NTHREAD	Number of parallel threads
MAXSTEPMCMF	Maximum number of local minimizations (excluding initialization)
NLOCMCMF	Parameter regulating ratio of local/global moves
MAXRANT	Maximum number of iterations for random choice of structure/family
INTPRT	Interval for printouts
MAXEVEN	If the lowest-energy minimum does not change for this number of local minimization temperatures are increased by factor of DKT_MCMF*
DENEMIN	Minimum difference between energies allowing to print out structure of new lowest-energy structure
DKT_MCMF1	Factor for increasing KT_MCM1
DKT_MCMF2	Factor for increasing KT_MCM2
BLOWUP	Structure is expanded by factor of BLOWUP after perturbation but before local minimization
KT_MCMF1_MAX	Maximum value of KT_MCMF1
KT_MCMF2_MAX	Maximum value of KT_MCMF2
ITCYCLE	If KT_MCMF* reaches its maximum it is reset to initial value if ITCYCLE=1
MAXFAMUSE	Maximum number of times a given family is used for perturbations
NRESET	Threads are reset to lowest-energy minima after NRESET number of minimizations
DELTA_EN	maximum relative energy deviation (in %) between equivalent structures
DELTA_V	maximum deviation of unit cell parameters a, b, c for which structures are still considered equivalent
DELTA_C	maximum deviation of unit cell parameters a, b, c for which structures are still considered equivalent
DELTA_A	maximum deviation of unit cell angels a, b, c for which structures are still considered equivalent
DDIST	maximum value of an average deviation of interatomic distances which structures are considered equivalent
TIMEOUT	parameter preventing ‘infinitely’ long minimization in the case of the big initial unit cell. TIMEOUT multiplied by average time necessary for a single minimization defines longest allowed minimization time.
RCUT1S RCUT1E RCUT2S RCUT2E	parameters used in geometrical focusing to define starting and ending sizes of a family and interfamily distance

Appendix B. Force field parameters

File latf.data - Van der Waals parameters (formatted)

Record 1: title
Record 2: type of potential function (1 – '6-12'; 2 – '6-exp'), number of atom types (format 3x, 2i7)
Record 3: atomic weights
Record 4: In the case of '6-12' potential, record contains number-of-atom-types lines, each containing ε_0 and r_0^2, where e_0 is potential well depth (kcal/mol) and r_0 is interatomic distance at the minimum (Angstrom) (format 2f12.5).

```
6-12 potential (AMBER)
         1     24
  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  8.0  8.0  8.0  8.0  6.0  6.0  6.0
           7.0 16.0 16.0 53.0  9.0 35.0
     0.01570     1.44000    0          ( 1  1)   H --- H
     0.00000     0.36000    0          ( 1  2)   H --- HO
     0.01570     1.44000    0          ( 1  3    H --- HS
     0.01570     4.35557    0          ( 1  4    H --- HC
     0.01570     3.94817    0          ( 1  5    H --- H1
     0.01570     3.56077    0          ( 1  6    H --- H2
     0.01570     3.19337    0          ( 1  7    H --- H3
     0.01535     4.23948    0          ( 1  8    H --- HA
     0.01535     4.03608    0          ( 1  9    H --- H4
     0.01535     3.83768    0          ( 1 10    H --- H5
```

In the case of '6-exp' potential, each line containes parameter A, B, and C (format f12.5, f15.3, f9.3).

```
W99 (Williams, 2001)
         2     14
  1.0  1.0  1.0  1.0  6.0  6.0  6.0  7.0  7.0  7.0  7.0  8.0  8.0  0.0
    66.53200       3030.593    3.560         1    1
     0.00000        511.567    3.560         1    2
     0.00000        289.491    3.560         1    3
     0.00000        744.337    3.560         1    4
   124.74975       9762.207    3.580         1    5
   164.49953      13993.991    3.580         1    6
   151.06338       8647.318    3.580         1    7
   149.60781       8353.944    3.520         1    8
   149.10645       8610.972    3.520         1    9
   194.39849      11790.848    3.520         1   10
   299.20349      17134.768    3.520         1   11
```

File tors.data (formatted). Parameters for the 3 terms of the Fourier series describing torsional potential (format 5x, 3F10.5)

```
   1   -0.394     0.5185    1.3295         C-C
   2    0.0       0.0       1.54           C-CM
   3    0.0       1.051     0.0            C-CR
   4    0.9245    0.475     0.42           C-CB
   5    0.0       0.0       0.333          CM-CB
   6    0.027    -0.0025    0.8745         C-CA (psi)
   7    0.007    -0.0025   -0.042          acetamide
   8   -0.8325    1.5095   -1.279          C-NE (Arg,chi-4)
   9    0.1585   -0.0095    0.869          C-N1
  10    0.0       0.0       0.956          CM-N1
  11   -0.33      0.6995    0.632          C-N2
  12    0.1375   -0.0945    1.399          CM-N2
  13    0.0       9.0095    0.0            CA-NA
  14    0.638    11.6205    0.0            CA-NH (omega)
  15    0.0       8.972     0.0            CA-NC (pre-Pro omega)
  16    0.0       0.0       0.5535         NH-C (phi)
  17    0.0       0.0       0.749          NH-CM
```