Chapter 4

Input and output files

4.1 Input files

The input to MHOP can be divided into two groups: input, controlling the operation of the MHA, and unput required by the DYNAMO-related functionalities.

4.1.1 input.<nnn>

Control parameters for the MHA are read upon start of the program from the input.<nnn> file(s), where <nnn> is a numeric extension (000, 001,...) corresponding to the process rank when the program is run on multiple processors (or processing elements, PE). The program, however, does allow that the number of input files is less than the number of requested PE. Thus, if a parallel job on 4 PE is submitted, then the process of rank 3 will look for input.003. If it is not found in the current working directory it will check for input.002 and so on. This means that *at least one* input file, input.000, should be provided to MHOP.

The input(s) input.<nnn> defines the values of the control keywords for the MHOP program. The follwoing keywords at present are available:

Keyword	Type	Description
QRESTART	logical	Specifies the mode to run MHOP .FALSE.: start a new run .TRUE.: restart run. In this case MHOP will read all counters from input.<nnn>
QVERBOSE	logical	Determines the amount of information dumped to the diferent output units
FPREFIX	character	File-name prefix of data input files for DYNAMO. Atomic coordinates should be provided in <FPREFIX>.crd and residues sequence specifications in <FPREFIX>.seq
SCRATCH	character	Path to scratch space. MHOP will write the protocol and monitor files to the scratch space.
OUTFORM	character	Format of the geometry files for reading and writing. Allowed values at present include: "PDB": Protein Data Bank format "PLT": BALSAC (cluster) format "VISU": generic format compatible with the free visualization tool VISU.
NLMINX	integer	Maximum number of minima to be found in the run.
NRANDOFF	integer	Initial offset for the random number generator
MDMIN	integer	The potential energy minimum in which the MD trajectory should be stopped.
EDIFF	real	Initial value for the E_{diff} parametter in the MHA [1]. This paramter controls the acceptance/rejection process for a new minimum: if the energy of the new minimum relative to the current energy doesn not exceed the threshold value E_{diff} the minimum is accepted.
EKINETIC	real	Kinetic energy for generating the Boltzmann velocity distribution in the molecular dynamics step of the MHA.
DT	real	Initial value of the time step for the velocity Verlet integration scheme.
MEFF	real	Effective mass of the C_α (CA) and C_β (CB) atoms for the MD escape step. The masses of all other atoms is internally fixed to 1.0

Every keyword should be specified on a sperate line, followed by its value, and the order of keywords should be as given above. No other records are allowed on the same input line. Input lines starting with '!' are considered as comments. *Blank lines* are ignored by the parser. A sample input.000 file for MHOP is given below:

input.000

```
!=== Input file for MHOP ===
! Run mode: f= new run; t= restart
QRESTART  F

! Printing mode: t= verbose; f= minimal ouput
QVERBOSE  F

! Filename prefix DYNAMO input files (coordinates, sequences)
FPREFIX   "struct"

! Scratch space for protocol and monitoring files
SCRATCH   "/tmp"
```

```
! Geometry format: {PDB, PLT, VISU}
OUTFORM   "PDB"

! Maximum number of minima to be found
NLMINX    200

! Initial offset for random number generator
NRANDOFF  3972

! Number of minima after which MD simulation stops
MDMIN     6

! Energy difference parameter in the min. hopp. algorithm
EDIFF     5.000E+01

! Kinetic energy for initial velocities in the MD step
EKINETIC  1.000E+02

! MD integration time step (arb. units)
DT        1.000E-04

! Effective mass of Ca and Cb carbons in the MD step
MEFF      0.100000000E+02
```

It is good to keep in mind that in setting up a parallel run every process should be fed with a unique NRANDOFF value. Furthermore the FPREFIX keyword allows for different processes to be intialized with different starting geometries.

The input.<nnn> file(s) are overwritten by MHOP in the course of the run, with all variable keyword values being updated. Aditionally an extra record is appended in the end of the file wich contains the current values of the following *real* counters, crucial for the operation of the MHA: HOPP, ESCAPE, ESCAPE_SAM, ESCAPE_OLD, and ESCAPE_NEW. Thus, the above sample input.000 for restart will read:

input.000

```
!=== Input file for MHOP ===
! Run mode: t= new run; f= restart
QRESTART  T
:
:
! Effective mass of Ca and Cb carbons in the MD step
MEFF      0.100000000E+02
 0.2738100E+05  0.3645100E+05  0.9070000E+04  0.9119000E+04  0.1826200E+05
```

4.1.2 Input to DYNAMO

Although all DYNAMO-related input is described in great details in Ref. [6] a short overview is given here for illustrative purposes. In a new run, MHOP reads the atomic coordinates in DYNAMO format from the file <FPREFIX>.crd, with the FPREFIX specified in input.<nnn>. Residue sequence information for the system is read from <FPREFIX>.seq. As a simple example we provide these files for the case of a peptide in a fully extended conformation, whose residue sequnce corresponds to that of the small opiate peptide Met-enkephalin: Tyr-Gly-Gly-Phe-Met, shown in Fig. 4.1.

metenk.seq

```
Sequence
  1
  Subsystem MET-ENKEPHALIN
    5
```

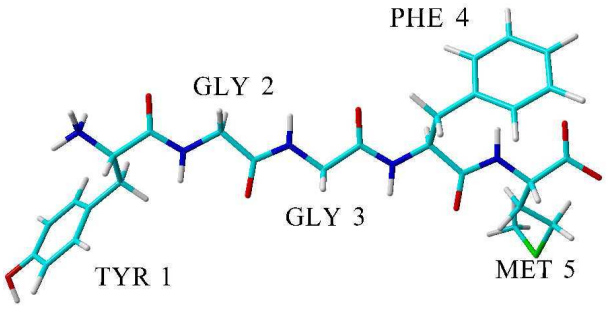

Figure 4.1: A fully extended structure of a small 5-residue peptide whose sequence corresponds to that of Met-enkephalin: Tyr-Gly-Gly-Phe-Met.

```
    TYR; GLY; GLY; PHE; MET
    Variant N_TERMINAL  TYR  1
    Variant C_TERMINAL  MET  5
  End
End
```

Note that the the sequence input data file is universal for all processes. Thus, once setup for a particular system one does not need to change it.

The coordinate file has a generic, self-explanatory format. All !-idented lines are treated as comments by the DYNAMO parser. Cartesian coordinates should be given in Å. Every atom must have a unique name within a given residue, and must match the OPLS-AA definitions as implemented in DYNAMO, cf. Ref. [6]. In the metenk.crd bellow, for brevity, coordinates of only 4 atoms per residue are indicated:

metenk.crd

```
!======================================================================
        75        5     1 ! # of atoms, residues and subsystems.
!======================================================================
Subsystem     1  MET-ENKEPHALIN
       5   ! # of residues.
!======================================================================
Residue       1  TYR
  23   ! # of atoms.
   1    N   7      0.078000     2.220000    -1.036000
   2  HT1   1      0.101000     2.814000    -1.853000
   :
  22   OH   8      3.293000    -3.064000     1.136000
  23   HH   1      2.969000    -3.950000     1.192000
!------------------------------------------------------
Residue       2  GLY
  7   ! # of atoms.
  24    N   7     -3.531000     1.550000    -1.036000
  25    H   1     -3.429000     0.556000    -1.036000
   :
  29    C   6     -5.973000     1.239000    -1.036000
  30    O   8     -5.760000     0.027000    -1.036000
!------------------------------------------------------
Residue       3  GLY
  7   ! # of atoms.
  31    N   7     -7.196000     1.759000    -1.036000
  32    H   1     -7.221000     2.758000    -1.036000
   :
  36    C   6     -9.658000     1.759000    -1.036000
  37    O   8     -9.600000     2.988000    -1.036000
!------------------------------------------------------
Residue       4  PHE
  20   ! # of atoms.
  38    N   7    -10.806000     1.089000    -1.036000
  39    H   1    -10.703000     0.094000    -1.036000
   :
  56   CZ   6    -15.873000     4.850000    -2.281000
```

```
  57   HZ   1    -16.833000     5.389000    -2.267000
!-------------------------------------------------------
Residue       5  MET
  18    ! # of atoms.
  58    N   7    -14.471000     1.297000    -1.036000
  59    H   1    -14.495000     2.297000    -1.036000
   :
  74  HE3   1    -16.125000    -0.391000     2.985000
  75  OT2   8    -18.080000     0.627000    -1.036000
!-------------------------------------------------------
```

Typically, when setting up a new run one has the starting geometry for MHOP in PDB format. To facilitate the conversion of the atomic coordinates to the DYNAMO format the package provides a simple `awk`-based shell script, `pdb2crd`, that extracts the coordinates from a PDB file and writes them in DYNAMO coordinate format. Some additional editting of the resultant file may be needed for more specific customization and making the atomic labels compatible with the OPLS-AA definitions as used in DYNAMO.

Another important input file is needed by the DYNAMO routine `mm_system_construct` in order to construct the internal representation of the biomolecular system. It is a binary file that is created by another DYNAMO routine, `mm_file_process`, upon reading the OPLS force filed parameters from the protein.opls file distributed with the DYNAMO library. The sequence of calls that do the job for MHOP looks like

```
call mm_file_process     ( "protein.opls_bin", "<path_to>/protein.opls"   )
call mm_system_construct ( "protein.opls_bin", trim ( FPREFIX ) // ".seq" )
```

Note that the former call is needed only if any changes to the protein.opls have been made (e.g., new residues or variants are defined, atomic types added etc.). In practice, once the binary file protein.opls_bin (or whatever name is given) is generated only the later call is to be made. This why protein.opls_bin should be always available to MHOP.

4.1.3 Runtime control files

At present runtime control over MHOP allows only for program termination by means of requesting a CPU time limit (in seconds) in the file CPUlimit. Thus, if the file is available in the working directory and contains, e.g., the number `137.0`, MHOP will terminate whenever the next internal timecheck exeeds this limit. This check is made every iteration of the MHA. For "immediate" termination one simply requests `0`. Empty CPUlimit does not affect the program execution.

4.1.4 Input files for restart run

During runtime, the MHOP creates and regularly updates a number of files that can be used in a reastart/continuation run. MHOP can be instructed to perform a restart run by setting `QRESTART .TRUE.` in the input.<nnn> file(s), in which case the `HOPP`, `ESCAPE`, `ESCAPE_SAM`, `ESCAPE_OLD`, and `ESCAPE_NEW` counters, in this particular orderm, are initialized to the values given on the last record in the input.<nnn> file(s). Every process reads input atomic positions from corresponding posinp<nnn>.crd file, that has to be available from previous run of the program. Sequence structure data are read from the same file, <FPREFIX>.seq, that has been used in the previous run.

Furthermore, for initialization of the MHA, MHOP reads the earr.dat file, containing all energy minima found in a previous run, and also all available coordinate files poslow<nnnn>.<OUTFORM>.

4.2 Output files: protocol, monitoring etc.

4.2.1 earr.dat

This is the main output file created by MHOP. It contains the sorted (in ascending order) energy history stack. During the run it is regularly updated. The format is rather generic:

earr.dat

```
     193113     1000000    ! NLMIN, NLMINX
-0.11186506423950195E+04 -0.11186506423950195E+04   ! E, E - Eref
-0.11186506404876709E+04 -0.11186506404876709E+04
-0.11186506366729736E+04 -0.11186506366729736E+04
   :
   :
```

Upon termination, the first record in the file contains the number of energy minima found (the `NLMIN` variable) and the requested number of minima to be found (cf. the `NLMINX` keyword in input.<nnn>). It is followed by `NLMIN` data records where the left column gives the absolute energy of the minima (the default energy unit in DYNAMO is kJ/mol), and the right column is the same energy measured with respect to a reference value (so far it is hard-coded via the `eref` variable in the `MINHOPP` module). This file should exist when running MHOP in restart mode.

4.2.2 global_mon.<nnn>

The operation of the MHA can be monitored during runtime by means of the global_mon.<nnn> file(s). It is created in the initialization phase and is then updated every time a local energy minimum is accepted by the MHA. The current format of the monitor file is illustrated below:

global_mon.<nnn>

```
#-------------------------------------------------------------------------------
# hopp  |   esc   |  E_wpos-Eref  |  Ediff   |  Ekin  | % same | % old | % new
#-------------------------------------------------------------------------------
   0.         0.     -945.195236    0.500E+01    0.100E+03  0.00  0.00  0.00
   1.         1.    -1021.593066    0.500E+01    0.952E+02  0.00  0.00  1.00
   5.         5.    -1055.120132    0.520E+01    0.784E+02  0.00  0.00  1.00
  23.        23.    -1070.272233    0.714E+01    0.326E+02  0.00  0.00  1.00
  28.        29.    -1065.184298    0.758E+01    0.268E+02  0.03  0.00  0.97
  32.        33.    -1062.113892    0.788E+01    0.220E+02  0.03  0.00  0.97
  33.        34.    -1070.313716    0.773E+01    0.210E+02  0.03  0.00  0.97
  35.        36.    -1075.532092    0.773E+01    0.190E+02  0.03  0.00  0.97
  37.        38.    -1070.313716    0.773E+01    0.190E+02  0.03  0.03  0.95
  38.        39.    -1076.719146    0.758E+01    0.181E+02  0.03  0.03  0.95
  39.        40.    -1076.689428    0.743E+01    0.173E+02  0.03  0.03  0.95
  48.        51.    -1080.182605    0.853E+01    0.123E+02  0.06  0.02  0.92
  53.        56.    -1085.690755    0.906E+01    0.961E+01  0.05  0.02  0.93
  54.        59.    -1087.814030    0.888E+01    0.101E+02  0.08  0.02  0.90
  :
  :
```

Every data record consits of 8 fields giving the following variables (values):
`hopp, escape, e_wpos-eref, ediff, ekinetic, escape_sam/escape, escape_old/escape, escape_new/escape`.

4.2.3 protocol.<nnn>

This is a generic protocol file where MHOP dumps information oabout varios aspects of the MHA. At present there is no well-defined format for the protocol. The amount of information dumped to protocol.<nnn> is controled by the keyword `QVERBOSE`. For `QVERBOSE = .FALSE.` only minimal output is generated. In long runs verbose information dump results in huge file size that are awkward to handle, hence `QVERBOSE = .TRUE.` might be useful choice for short test runs only.

Typical contents of the protocol file is shown below:

protocol.002

```
############################################################
                    _
 _ __ ___  | |__   ___  _ __
| '_ ` _ \| '_ \ / _ \| '_ \
| | | | | | | | | (_) | |_) |
|_| |_| |_|_| |_|\___/| .__/
                       |_|
############################################################

First call to GRADIENT():  0.000000E+00 sec.
Einp= -0.2647507E+03 [ kJ/mol ]
New run with nlminx=      1000000
Iteration results written to: /scratch/tmp/penev/global_mon.002
+---------------------------------------------------------------------------+
| beta1,  beta2, beta3:   0.105E+01  0.105E+01  0.952E+00
| alpha1, alpha2      :   0.980E+00  0.102E+01
| mdmin, meff         :      6      20.000
+---------------------------------------------------------------------------+
Perform initial geometry optimization
Eref=        0.000000
Initial re_sm=   -945.195236206
+---------------------------------------------------------------------------+
MD_ESCAPE>      Number of heavy atoms=      8.0
MD_ESCAPE>      istep=   9638, t=     0.964
+---------------------------------------------------------------------------+
|> New lowest      1     1 -0.10215930671691895E+04
+---------------------------------------------------------------------------+
MD_ESCAPE>      istep=   8332, t=     0.875
MD_ESCAPE>      istep=   8423, t=     0.929
MD_ESCAPE>      istep=   4985, t=     0.636
:
:
:
MD_ESCAPE>      istep=    561, t=     0.862
MD_ESCAPE>      istep=    754, t=     1.217
+---------------------------------------------------------------------------+
|> New lowest     47   306 -0.11186506423950195E+04
+---------------------------------------------------------------------------+
MD_ESCAPE>      istep=    666, t=     1.128
MD_ESCAPE>      istep=    358, t=     0.637
:
:
+---------------------------------------------------------------------------+
CPU timelimit exceeded => Finish
+---------------------- Final collect & info dump ----------------------
            Minama found [ local / global ]:      23655/    193113
                     Ratios: same, old, new:   0.102E+00  0.398E+00  0.500E+00
                  Ratio: accepted, rejected:   0.500E+00  0.500E+00
               COUNT_MD, COUNT_SD, COUNT_CG:   103071105   16806807  744819332
 energy/force eval, cpu/eval[sec], cpu[hrs]:    864697244        0.000       68.592
                    Average EDIFF, EKINETIC:   0.971E+01  0.175E+02
                  Final EDIFF, EKINETIC, DT:   0.221E+02  0.231E+02  0.154E-02
                     Final lowest E, E/atom:  -0.1118651E+04 -0.1491534E+02
+---------------------------------------------------------------------------+
| Minima Hoping Simulator finished
+---------------------------------------------------------------------------+
```

In the present version of the protocol after the initial header informaion, only the number of MD steps in each iteration of the MHA are printed by the `md_escape` subroutine. In case the process finds a new lowest energy minimum, this is acknowledged by:

```
|> New lowest      47     306 -0.10215930671691895E+04
```

where the first number is the local for the processor minimum index (`nlmin_l`), the second is the global minimum index (`nlmin`) followed by the corresponding energy.

Final information dump is made upon "manual termination" (via the CPUlimit file) or once NLMINX minima have been found. In the above example, MHOP have been terminated by requesting a 0 time-limit since the the run has apprached an exaustive minima search (fraction of old minima revisited amounted to 40 %).

4.2.4 posbest<nnnn>_<nnn>.<OUTFORM>

These files are created for every new lowest-energy minimum. the 4-digit part of the filename, <nnnn>, indicates the sequential index of the minimum in process-local ranking and <nnn> is the rank of the process that found the minimum. Thus if the PDB format is requested (OUTFORM "PDB"), the 6th lowest-energy minimum for process 4 will be saved as posbest0006_004.pdb.

In the case of PDB format, the program adds the corresponding energy and the process rank as a non-standard 'REMARK 6' line in the PDB file. The first line of posbest0006_004.pdb may okk like:

```
REMARK   6 -1116.16331100   4
```

We note that in the PDB format Cartesian coordinates are given with 3 decimal digits precission. Ofte, however, the initial input coordinates may be provided with higher precission. This remark line can then be used to estimate the effect of the coordinates precission on the total energy of the system.

4.2.5 poslow<nnnn>.<OUTFORM>

This files contain the geometries of the lowest NPMINX energy minima, with the corresponding energies and process ranks included as a non-standard remark in the case of PDF format. Thus, if NPMINX = 200, MHOP will create all posbest0001.pdb, . . . ,posbest0200.pdb files in the current working directory.

4.2.6 Geometries for restart run: posinp<nnn>.crd

Atomic positions that can be used as structural input in restart/continuation run are saved in DYNAMO coordinate format in posinp<nnn>.crd files, where <nnn> is the process rank. This file is regularly updated in the course of the run.