J O Y C E

August 28, 2007

Version 1.0

INSTALLATION AND USER GUIDE

Contents

1 General purposes

The energy and its first and second geometrical derivatives obtained by DFT calculations for a number of conformations of a single molecule are used to parameterize intramolecular force fields, suitable for computer simulations.

The JOYCE program first reads a Moscito3.9 system file [1] in which are specified the model potential functions defining the intramolecular potential to be used in the MD simulations.

The equilibrium values of the chosen internal coordinates (IC) are read by the Joyce program from a formatted checkpoint file (.fchk) produced by the Gaussian03 package [2].

The force constants are computed by an non standard procedure [3] from the first and second derivatives again read from the .fchk file.

The authors can be contacted at

http://www.dcci.unipi.it/≃ivo

email: ivo@dcci.unipi.it ; giacomo@dcci.unipi.it

2 Installation

The Joyce program is intended to run on any Linux platform. The only requirement is a Fortran 77 compiler, such as g77, ifort or any other.
Please note that the implemented compilers are:
g77 – free GNU compiler
ifort – Intel compiler, available for academic use
pgf90 – Portland compiler
lf95 – Lahey compiler

To install the program you should follow the next steps:

1. Unzip and untar the Joyce package: Joyce.v1-0.tgz with the command:
   > *tar -xzvf Joyce.v1-0.tgz*

2. Set the environmental variable JOYCE where the program was unpacked, *e.g.* if the program was unpacked in the user home directly write (in tcsh shell):
   > *setenv JOYCE /home/username/Joyce.v1.0*
   Alternatively, you can add the aforementioned veritable definition in your login script.

3. Copy the Joyce executable in your bin directory (/home/username/bin) or any other directory contained in your PATH:
   > *cp $JOYCE/exe/go.joyce /home/username/bin/.*

4. Install the program by writing
   > *go.joyce –g77*
   to use g77 as compiler. Just type
   > *go.joyce*
   to see other compiling options.

3 Theory

To make the formulae easier to be understood, the following notation will be adopted for the summation indeces and symbols

i, j are used for the Cartesian coordinates (CCs) x or mass weighted Cartesian coordinates $(1 \div 3N)$

μ, ν indicate the redundant internal coordinates [4,5] (RICs) q $(1 \div N_{RIC})$

K, L run over the normal coordinates (NCs) Q $(1 \div 3N-6)$ $(3N-5$ for linear molecules)

g run over the considered molecular geometries $(0 \div N_g)$

a, b indicate the functions f used to represent the empirical FF and/or the number of linear parameters of the FF $(1 \div N_{func})$

s, t run over the quantities to be represented by the FF (energies, energy gradients and Hessian) for the considered geometries $(1 \div N_{points})$

The QMD-FF, to be used in molecular dynamics or molecular mechanics, is expressed through a linear combination of functions f_a of a set of RICs

$$V(q) = \sum_{a=1}^{N_{func}} p_a \; f_a(q) \tag{1}$$

where the q symbol collects all RICs. The functions may conveniently be expressed in terms of displacements with respect to a given reference geometrical conformation identified by the vector q^0

$$\Delta q_\mu = q_\mu - q_\mu^0 \tag{2}$$

Usually the RICs consist in all bond stretches, angle bendings and dihedral torsions that can be obtained from a given connectivity criteria referred to the reference conformation. The inversion coordinate [6] can be included for atoms bonded to three other atoms. Nonbonded intramolecular interactions can also be added in order to make the FF more

accurate. In usual FFs the number of RICs exceeds $3N$-6 and therefore they form a redundant set of coordinates. Although equation (1) has been written in general form, each function f_a only depends on one or two RICs, as reported in detail later on (equations (36)–(41)).

3.1 Internal coordinates transformations

Since the Hessian and gradients are computed in CCs, whereas the FF is usually expressed through RICs, some coordinate transformation is required. For infinitesimal displacements with respect to a given geometrical conformation, the RICs are related to the nuclear CCs x through a non invertible transformation

$$\delta q = B\ \delta x \tag{3}$$

where δq and δx are colum vectors. The Wilson rectangular B matrix

$$B_{\mu i} = \left(\frac{\partial q_\mu}{\partial x_i}\right) \tag{4}$$

is related to the geometry the displacements are referred to, and can be accurately computed both in analytical [7] and numerical ways.

The normal coordinates are computed from the Hessian matrix in CCs

$$H_{ij} = \left(\frac{\partial^2 E}{\partial x_i \partial x_j}\right) = E''_{ij} \tag{5}$$

obtained by a QM calculation at a given geometry. H is transformed to the mass weighted CCs form and diagonalized by a unitary matrix C

$$M^{-1/2}\ H\ M^{-1/2} C = C\Lambda \tag{6}$$

The matrix M is diagonal and for each CC contains the mass m of the related atom. The columns of the C matrix are the linear combinations of the mass weighted CCs that correspond to the NCs displacements

$$\delta Q_K = \sum_{i=1}^{3N} \sqrt{m_i} C_{iK} \delta x_i \tag{7}$$

or in matrix form

$$\delta Q = \widetilde{C}\ M^{1/2} \delta x \tag{8}$$

where δQ and δx are column vectors. In the case the geometry corresponds to an absolute or local energy minimum, $3N-6$ eigenvalues Λ_K are positive and refer to vibrations, whereas the 3 translational and 3 rotational modes are identified by zero eigenvalues. In other cases negative eigenvalues can occur and these do not correspond to vibrational modes. If all the NCs are retained, the transformation of equation (7) is fully invertible

$$\delta x = M^{-1/2} C \delta Q \tag{9}$$

The relation between the RICs and the NCs can be easily obtained exploiting the completeness of the CCs basis set. Using equations (3) and (9)

$$\delta q = B M^{-1/2}\, C \delta Q = T \delta Q \tag{10}$$

where the T matrix is defined as

$$T_{\mu K} = \left(\frac{\partial q_\mu}{\partial Q_K} \right) \tag{11}$$

Thus the RICs may be expressed in terms of the NCs and the inclusion or not of the rotational and translational modes is uninfluential since they leave the RICs unchanged.

3.2 The optimal parameters of the Force Field

The best parameters for the QMD-FF in order to represent the internal molecular motion are obtained by minimizing the following merit function, written as a sum over the considered molecular geometries

$$I = \sum_{g=0}^{N_g} I_g \tag{12}$$

where

$$I_g = W_g \left[(E_g - E_0) - V_g \right]^2 + \sum_{K=1}^{3N-6} \frac{W'_{Kg}}{3N-6} \left[E'_{Kg} - V'_{Kg} \right]^2 + \sum_{K \le L}^{3N-6} \frac{2 W''_{KLg}}{(3N-6)(3N-5)} \left[E''_{KLg} - V''_{KLg} \right]^2 \tag{13}$$

The indeces K, L (capital letters) run over the normal coordinates and include all the modes except for the rotational and translational ones. E_g is the total energy obtained

by a QM calculation and E_0 is the same at the reference geometry ($g = 0$). E'_{Kg} (E''_{KLg}) is the energy gradient (Hessian) at a given geometry with respect to the NC evaluated at the same geometry. V, V' and V'' are the corresponding quantities calculated by the FF in equation (1). The constants W, W' and W'' weight the several terms at each geometry and can be chosen in order to drive the results depending on the circumstances. The energy, gradient and Hessian terms are normalized in order to account for the different number of terms and to make the weights independent from the number of atoms in the molecule.

To compute the energy derivatives entering the merit function (13) we have to performs some transformations since no derivative is originally expressed with respect to the NCs. Indeed standard quantum chemistry programs provide derivatives E' and E'' with respect to CCs. Using the above relations and exploiting the completeness of the CCs, the transformation is simple

$$E'_K = \left(\frac{\partial E}{\partial Q_K}\right) = \sum_{i=1}^{3N} \left(\frac{\partial E}{\partial x_i}\right) \left(\frac{\partial x_i}{\partial Q_K}\right) = \sum_{i=1}^{3N} E'_i \; m_i^{-1/2} \; C_{iK} \tag{14}$$

or, in matrix form

$$[E']_{NC} = \widetilde{C} M^{-1/2} [E']_{CC} \tag{15}$$

where the square parentheses indicates column vectors of energy gradients computed with respect to the NCs and the CCs. The QMD-FF energy gradients at a given geometry

$$V'_K = \sum_{a=1}^{N_{func}} p_a \left(\frac{\partial f_a}{\partial Q_K}\right) = \sum_{a=1}^{N_{func}} p_a \; f'_{aK} \tag{16}$$

can be conveniently computed using the derivatives of the basis function with respect to the RICs, that is

$$\left(\frac{\partial f_a}{\partial Q_K}\right) = \sum_{\mu=1}^{N_{RIC}} \left(\frac{\partial f_a}{\partial q_\mu}\right) \left(\frac{\partial q_\mu}{\partial Q_K}\right) = \sum_{\mu=1}^{N_{RIC}} \sum_{i=1}^{3N} \left(\frac{\partial f_a}{\partial q_\mu}\right) T_{\mu K} \tag{17}$$

or in matrix form

$$[f'_a]_{NC} = \widetilde{T} [f'_a]_{RIC} \tag{18}$$

The Hessian matrix of the QM calculation in NCs

$$E''_{KL} = \left(\frac{\partial^2 E}{\partial Q_K \partial Q_L} \right) \tag{19}$$

is obtained from the Hessian matrix in the CC basis according to

$$[E'']_{NC} = \widetilde{C} M^{-1/2} [E'']_{CC} M^{-1/2} C \tag{20}$$

The second derivatives of the FF are a bit more complicated since they involve derivatives of the B matrix and are conveniently expressed in explicit form

$$\left(\frac{\partial^2 f_a}{\partial Q_K \partial Q_L} \right) = \sum_{\mu\nu=1}^{N_{RIC}} T_{\mu K} \left(\frac{\partial^2 f_a}{\partial q_\mu \partial q_\nu} \right) T_{\nu L} + \sum_{\mu\nu=1}^{N_{RIC}} T_{\mu K} \left(\frac{\partial f_a}{\partial q_\nu} \right) \left(\frac{\partial T_{\nu L}}{\partial q_\mu} \right) \tag{21}$$

As shown in equation (1), the QMD-FF is linear in the p parameters, thus the least squares minimization of functional (13) can be written as

$$\sum_{a}^{N_{func}} \sum_{s}^{N_{point}} \alpha_{bs} \, W_s \, \alpha_{as} \, p_a = \sum_{s}^{N_{point}} \alpha_{bs} \, W_s \, \beta_s \tag{22}$$

where the index s runs over the collections $[g]$, $[Kg]$ and $[KLg]$ defined in equation (13) for energy, gradient and Hessian, respectively. Following this notation the matrix α and the vector β are defined as

$$\alpha_{as} = f_{as} \text{ or } f'_{as} \text{ or } f''_{as} \quad ; \quad \beta_s = E_s \text{ or } E'_s \text{ or } E''_s$$

and

$$W_s = W_s \text{ or } \frac{W'_s}{3N-6} \text{ or } \frac{W''_s}{(3N-6)(3N-5)}$$

where f's are the functions of equation (1), E, E', E'' the QM data and W, W', W'' the weights of the merit function (13). Thus, defining

$$A = \alpha W \widetilde{\alpha}$$

$$b = \alpha W \beta$$

one has to solve a standard linear equation in the form

$$Ap = b \tag{23}$$

where A is a symmetric matrix.

In usual FF it is convenient for practical purposes, to employ functions of the RIC that will be in general redundant over the considered points. The scalar product between the FF functions is defined as

$$f_a \cdot f_b = \sum_{s=1}^{N_{point}} W_s f_{as}\ f_{bs} \tag{24}$$

and the redundancy strongly depends on the number and type of points included in the fitting. However in general the f set might not be linearly independent. This leads to a singular A matrix and the direct inversion method can not be used to solve the linear system (23). On the contrary, the Singular Value Decomposition method [8, 9] adapted to symmetric matrices is adequate and provides a stable solution of the linear system.

3.3 United Atom Theory

In many molecular simulations a group of atoms whose individual behavior is considered not to be crucial for the properties to be investigated, can be grouped in a single interaction site. This approach, henceforth named United Atom (UA), allows saving computational time and simultaneously removes some high frequency vibrational modes which can limit the integration time step in MD simulations. The most common example concerns aliphatic chains where each CH_2 group is treated as a single interaction site (C_2) with FF parameters accounting for the effect of the hydrogen atoms both in the non-bonded interactions and electrostatic charge. Despite recent work has been done for some torsional potentials [10], usually the intramolecular FF parameters of "hard" IC are not changed in the UA approach, thus the parameters driving the C_2-C_2-C_2 stretching and bending motion in the aliphatic chains are the same as those commonly employed in the FA description.

In the UA approximation the involved atoms are considered to move as a single point with the consequence that the translational movements with respect to the rest of the molecule can be somehow taken into account, but the relative rotational movements are irreparably lost. In other words a three dimensional object described by 6 coordinates is transformed into a single point described by 3 coordinates. Even in the (non realistic)

hypothesis that there exists some local vibrational modes much faster than those involving the atoms close to the UA, this approximation affects the motion of the neighboring atoms. Thus the remaining vibrational frequencies are altered by the UA approach and it is convenient focusing on the representation of the intra-molecular potential energy rather than on the vibrational analysis.

In the JOYCE program the UA atom approach, consistently with the previous FA approach, is treated on the basis of *ab initio* calculation of energies, gradients and Hessian. The main problem concerns with the transformation of the gradient vector and Hessian matrix in equation (13) in the case the number of effective atoms is less than than the number of true atoms in the molecule. Let consider for simplicity the case of a single UA in which N_{UA} atoms are grouped together. We use the indeces μ, ν for the Cartesian coordinates referred to the atoms involved in the UA and the indeces a, b for those of the remaining atoms not involved in the UA (in this section we are forced to change the previous notation). For simplicity we suppose that only one atom in the UA group is linked to the unaltered atoms. The first order energy expansion around a given geometry is

$$E^{(1)} = \sum_{a} \sum_{s}^{x,y,z} E'_{as}\ \delta t_{as} + \sum_{\mu} \sum_{s}^{x,y,z} E'_{\mu s}\ \delta t_{\mu s} \tag{25}$$

where t_{as} represents the s-th component of the CC of the a-th atom. The new gradient vector of the united atom U for a given geometry is transformed according to the simple expression

$$E'_{Us} = \sum_{\mu} E'_{\mu s} \qquad\qquad (s = x, y, z) \tag{26}$$

where E'_U represents the energy gradient with respect to the UA displacements. This expression is consistent with the hypothesis that the UA represents a set of internally frozen atoms: $\delta t_{Us} = \delta t_{\mu s}$ $(\mu = 1...N_{UA})$ and holds for simultaneous translations but not for rotations of the grouped atoms.

The second order energy is

$$E^{(2)} = \frac{1}{2} \sum_{ab} \sum_{sr}^{x,y,z} E''_{as,br} \delta t_{as} \delta t_{br} + \frac{1}{2} \sum_{\mu\nu} \sum_{sr}^{x,y,z} E''_{\mu s,\nu r} \delta t_{\mu s} \delta t_{\nu r} + \sum_{a\mu} \sum_{sr}^{x,y,z} E''_{as,\mu r} \delta t_{as} \delta t_{\nu r} \tag{27}$$

Defining the UA Hessian matrix as

$$E''_{Us,Ur} = \sum_{\mu\nu} E''_{\mu s,\nu r} \tag{28}$$

$$E''_{as,Ur} = \sum_{\mu} E''_{as,\mu r} \tag{29}$$

the energy expression becomes

$$\begin{aligned} E^{(2)} &= \frac{1}{2}\sum_{ab}\sum_{sr} E''_{as,br}\delta t_{as}\delta t_{br} + \frac{1}{2}\sum_{sr} E''_{Us,Ur}\delta t_{Us}\delta t_{Ur} + \sum_{a}\sum_{sr} E''_{as,Ur}\delta t_{as}\delta t_{Ur} \\ &= \frac{1}{2}\widetilde{\delta t}E''\delta t \end{aligned} \tag{30}$$

It is easy to verify that such a transformation of the Hessian matrix will preserve the three null eigenvalues due to translations, whereas the rotational modes of a molecule with UA included may lead to small (unphysical) energy contributions with the further undesirable consequence of small mixing between rotational and vibrational modes.

The two other quantities of the UA to be defined are the mass and the position. For the UAs considered in this paper (methylene and methyl groups) the mass was taken as the sum of the involved atoms. In the case only one atom of the grouped atoms forms bonds with the rest of molecule, the natural choice for the position seems to make the UA coincident with that atom. However other choices are possible, for example the UA may be placed in the center of mass of the grouped atoms at the equilibrium geometry and/or its mass may be chosen in order to preserve the original inertia moments. Taking as criteria the magnitude of the rotational eigenvalues and the perturbation of the vibrational modes, these attempts do not lead to any improvement and were rejected. With the original choice the rotational eigenvalues at the equilibrium geometry are found to be much lower than the low frequency vibrational modes and the contamination is very small.

In summary the UA approach preserves some of the original atom-atom interactions contained in the Hessian matrix and leads to a useful simplification of the intra-molecular energy hyper-surface but does not allow conserving the rigorous implementation of the all-atom force field presented in this paper.

3.4 MD model Force-field

The FF employed in MD simulations has the following expression:

$$E_{tot} = E_{inter} + E_{intra} \tag{31}$$

The intermolecular part, E_{inter}, is computed as

$$E_{inter} = E_{LJ} + E_{Coul} \tag{32}$$

where the long-range electrostatic term is

$$E_{Coul} = \sum_{i=1}^{N_{sites}} \sum_{j=1}^{N_{sites}} \frac{q_i q_j}{r_{ij}} \tag{33}$$

and a Lennard-Jones term has been employed for the short range part, *i.e.*

$$E_{LJ} = \sum_{i=1}^{N_{sites}} \sum_{j=1}^{N_{sites}} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{34}$$

where i and j belong to different molecules and N_{sites} is the total number of interacting sites. The intermolecular parameters q_{ij}, σ_{ij} and ϵ_{ij} were taken for all molecules from the OPLS [11, 12] literature force-field.

The intramolecular part of the QMD-FF is expressed as a sum of different terms, namely

$$E_{intra} = V(q) = E_{stretch} + E_{bend} + E_{Rtors} + E_{Ftors} + E_{Coupl} \tag{35}$$

The first three terms count for the "hard" IC, *i.e.* bond stretchings, angle bendings and stiff angle dihedrals (Rdihedrals), as those that drive the planarity of aromatic rings and are expressed with harmonic potentials:

$$E_{stretch} = \frac{1}{2} \sum_{\mu}^{N_{bonds}} k_{\mu}^{s} (r_{\mu} - r_{\mu}^{0})^2 \tag{36}$$

$$E_{bend} = \frac{1}{2} \sum_{\mu}^{N_{angles}} k_{\mu}^{b} (\theta_{\mu} - \theta_{\mu}^{0})^2 \tag{37}$$

$$E_{Rtors} = \frac{1}{2} \sum_{\mu}^{N_{Rdihedrals}} k_{\mu}^{t} (\phi_{\mu} - \phi_{\mu}^{0})^2 \tag{38}$$

Conversely, the model functions employed for more flexible dihedrals (Fdihedrals) are sums of periodic functions, namely

$$E_{Ftors} = \sum_{\mu}^{N_{Fdihedrals}} \sum_{j=1}^{N_{cos_\mu}} k^d_{j\mu}[1 + cos(n^\mu_j \phi_\mu - \gamma^\mu_j)] \tag{39}$$

were N_{cos_μ} is the number of cosine function employed to describe the potential of the ϕ_μ dihedral. It is worth noticing that equations (36) - (39) can be easily expressed in the formalism of equation 1 by setting $q_\mu = r_\mu, \theta_\mu, \phi_\mu$ respectively. Following the notation introduced in equation (1), the last term of equation (35) can be written as

$$E_{Coupl} = \sum_{i}^{N_{Coupl}} V_i(q_\mu, q_\nu) \tag{40}$$

and may contain specific cross terms between the ICs q_μ, q_ν. The presence or absence of these off-diagonal coupling terms, which for instance may be of the form proposed in Ref. [13], discriminates between QMD-FFs of class II or class I. In this paper only couplings between soft dihedrals have been tested, which take the following expression

$$V_i(\phi_\mu, \phi_\nu) = \sum_{j=1}^{N_{sin^i_\mu}} \sum_{k=1}^{N_{sin^i_\nu}} k^{tc}_{ijk} sin(n^i_j \phi_\mu - \gamma^i_j) sin(m^i_k \phi_\nu - \gamma^i_k) \tag{41}$$

4 User Guide

A complete version of the JOYCE manual is under construction. A brief summary of the main commands to use JOYCE follows.

All commands are contained in the Joyce input section, which is named joyce.*jobname*.inp; to launch the program write

> *go.joyce jobname -e*

The Joyce input file is automatically edited (-e option), and the commands to perform the intramolecular parameterization are all contained in this file. For a list of these commands write

> *go.joyce –h*

Some examples are given in the examples directory, in the Joyce package.

The program creates two main outputs: a joyce.*jobname*.out file which contains all the numerical details of the fitting procedure, and a new.system file which contains all the fitted parameters in the Moscito format.

More details about the JOYCE procedure and the examples can be found in Ref. [3]. The authors can be contacted at the following email-addresses:

ivo@dcci.unipi.it ; giacomo@dcci.unipi.it

References

[1] Paschen, D.; Geiger, A. *MOSCITO 3.9;* Department of Physical Chemistry: University of Dortmund, 2000.

[2] *Gaussian 03 (Revision A.1)* M. J. Frisch *et al.*. *Gaussian, Inc., Pittsburgh PA* **2003**.

[3] Cacelli, I.; Prampolini, G. *J. Chem. Theory Comput.* **2007**, page published on the web.

[4] Pulay, P.; Fogarasi, G. *J. Chem. Phys.* **1992**, *96*, 2856.

[5] Peng, C.; Ayala, P.; Shlegel, H.; Frisch, M. *J. Comp. Chem.* **1996**, *17*, 49.

[6] Dasgupta, S.; Goddard III, W. *J. Chem. Phys.* **1989**, *90*, 7207.

[7] Bakken, V.; Helgaker, T. *J. Chem. Phys.* **2002**, *117*, 9160.

[8] Dasgupta, S.; Yamasaki, T.; Goddard III, W. *J. Chem. Phys.* **1996**, *104*, 2898.

[9] Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical Recipies in Fortran 77;* Cambridge University Press: Cambridge, 1992.

[10] Yang, L.; Tan, C.; Hsieh, M.; Wang, J.; Duan, Y.; Cieplak, P.; Caldwell, J.; Kollmann, P.; Luo, R. *J. Phys. Chem. B* **2006**, *110*, 13166.

[11] Jorgensen, W.; Maxwell, D.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

[12] Damm, W.; Frontera, A.; Tirado-Rives, J.; Jorgensen, W. *J. Comp. Chem.* **1997**, *18*, 1955.

[13] Maple, J.; Hwang, M.-J.; Stockfish, T.; Dinur, U.; Waldman, M.; Ewig, C.; Hagler, A. *J. Comp. Chem.* **1994**, *15*, 162.