COMPUTATIONAL ORGANIC CHEMISTRY

Giuseppe Zampella, Luca De Gioia
Department of Biotechnology and Biosciences. University of Milano-Bicocca, Italy

Keywords: Molecular Mechanics, Molecular Dynamics, Quantum Chemistry, Density Functional Theory, Semiempirical Methods, Molecular Orbitals

Contents

Summary

The aim of this contribution is to provide the reader with the basic concepts of computational organic chemistry. In the first section the relevance of computational chemistry to modern organic chemistry is discussed. In the second section the theory and application of computational methods based on classical physics models are presented. In the third section, molecular orbitals theory is introduced and the Hartee-Fock implementation is discussed. Methods based on density functional theory are outlined in the fourth section and the contribution is closed with the description of semi-empirical methods and their application to organic chemistry.

1. Introduction

Computational organic chemistry is a branch of theoretical chemistry whose main objective is the development and use of reliable mathematical models and algorithms to calculate properties of organic molecules, with the final goal to apply these algorithms to concrete chemical problems, such as the investigation of reaction paths.

While the term theoretical chemistry may be defined as the mathematical description of chemistry, the term computational chemistry is usually used when a mathematical method is sufficiently well developed that it can be automated for implementation on a computer. Indeed, almost every aspect of chemistry can be described using a qualitative or approximate quantitative computational scheme, and present computational

chemistry methods can accurately compute the properties of molecules that contain more than 1000 atoms. Consequently, recent years have seen an increase in the number of people involved in computational organic chemistry investigations. Notably, many users of computer programs relevant to computational organic chemistry are not theoreticians and are mainly interested in applied chemistry.

The spread of computational organic chemistry as a useful tool in the organic chemistry laboratory has been also facilitated by the development of computer software which is increasingly easy to use. As a consequence, it is now relatively easy to carry out computational chemistry investigations even if one does not know all the theoretical chemistry that is at the base of computation. However, it is very important to be able to choose the most suited computational method to investigate a specific chemical problem. As an example, not all computational chemistry approaches are suited to investigate the reactivity of a molecule. On the other hand, it is also important to be able to choose the proper level of approximation to investigate a particular chemical problem, in light of the observation that the most accurate computational methods are also the most computationally demanding. Indeed, in principle it is possible to use only very accurate methods and apply it to all molecules. However, although such methods are well-known and available, the computational cost of their use grows sharply with the number of atoms. Therefore, a great number of approximate methods strive to achieve the best trade-off between accuracy and computational cost.

The methods of computational organic chemistry cover a broad range of applications, and can give valuable contribution in different fields such as: storing and searching for chemical data, identification and evaluation of correlations between chemical structures and properties, theoretical investigation of structures and electronic properties, design of synthetic pathways, design of molecules that interact in specific ways with other molecules. Moreover, the types of studies undertaken by people involved in computational organic chemistry can be roughly divided in two categories: computational studies carried out in order to drive laboratory experiments and computational studies carried out with the aim of exploring reaction mechanisms and help rationalizing experimental observations.

The computational organic chemistry methods that are more commonly used to investigate chemically relevant problems can be also classified according to the approach used to describe atoms and electrons in the molecules under investigation. The properties of large molecules can be often investigated using methods based on classical physics (molecular mechanics and molecular dynamics), in which the presence of electrons (that cannot be treated satisfactorily using Newton's mechanics but requires a quantum mechanical description) is treated implicitly. However, classical physics methods do not allow us to investigate chemical problems that are explicitly dependent on the electronic properties of a molecule, such as reactivity (which implies bond formation and cleavage). The description of the reactivity of molecules must be undertaken using quantum chemical methods (Hartree-Fock and post Hartree-Fock methods), even though it should be underlined that presently the investigation of the reactivity of molecules that contain many electrons can be computationally very demanding and tractable only using semiempirical methods or approaches based on the density functional theory.

In light of the above observations, the aim of this section is to give a basic description of key principles and methods relevant to computational organic chemistry. In particular, a concise description of the theoretical basis of classical, quantum chemical and semiempirical methods will be presented, in the context of organic chemistry applications.

2. Computational Approaches based on Classical Physics: Molecular Mechanics and Molecular Dynamics

2.1 Molecular Mechanics

Organic molecules can be often modeled successfully without explicitly considering electrons, and therefore avoiding computationally demanding quantum mechanical calculations. The most common approaches imply the use of the so-called Molecular Mechanics (MM) formalism, in which the energy of a molecule is computed according to a classical mechanics expression for the energy. In particular, in the MM approach electrons are not considered explicitly, but rather it is assumed that they will find their optimum distribution once the positions of the nuclei are known. This assumption is based on the Born-Oppenheimer approximation of the Schrödinger equation. The Born-Oppenheimer approximation states that nuclei are much heavier and move much more slowly than electrons. Thus, nuclear motions, vibrations and rotations can be studied separately from electrons; the electrons are assumed to move fast enough to adjust to any movement of the nuclei.

In a very crude way, it can be stated that MM treats a molecule as a collection of spherical rigid bodies connected with springs, where the rigid bodies represent the nuclei and the springs represent the bonds. According to this formalism, the atoms composing the system are treated as classical particles whose interactions are described by simple two-, three- and four-body potential energy functions. The simplifications adopted in MM result in less computationally demanding calculations and consequently allow us to handle complex organic molecules. Of course, the simplified treatment of a molecule in MM does not allow the investigation of all chemically relevant feature of a molecule. In particular, the non-explicit description of electrons hinders the study of chemical reactions (bond formation and cleavage), as well as the computation of spectroscopic properties that explicitly depend upon electrons.

In MM approaches, the potential energy of a molecule is calculated using a set of potential energy functions and parameters, which constitute the force field. Several force fields have been specifically developed to study different families of molecules, including biomolecules, organic molecules, polymers and materials. Typical MM force fields consist of terms that describe the energy associated with bond stretching, bond angles bending, torsions and non-bonded interactions, as usually expressed by the following equation:

$$V_{\mathrm{p}}\left(r^{N}\right)=\frac{1}{2}\sum_{bonds}K_{\mathrm{b}}\left(b_i-b_{i,0}\right)^2+\frac{1}{2}\sum_{angles}K_{\theta}\left(\theta_i-\theta_{i,0}\right)^2+\frac{1}{2}\sum_{torsions}K_N\left(1+\cos(n\omega-\gamma)\right)+$$
$$\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]+\frac{q_iq_j}{4\pi\varepsilon_0 r_{ij}}\right) \quad (1)$$

where $V_{\mathrm{p}}\left(r^{N}\right)$ indicates the potential energy as a function of the position (r) of the N atoms, N = number of atoms, K_{b} = bond force constant, b_i = bond distance, $b_{i,0}$ = reference bond distance, K_θ= angle force constant, θ_i = current bond angle, $\theta_{i,0}$ = reference bond angle, K_N = torsional force constant, ω = current torsional angle, γ = phase angle, ε = dielectric constant, σ = Lennard-Jones parameters, q = partial atomic charges. The first term of the Eq. (1) represents the energy required to stretch bonds from their reference bond lengths $b_{i,0}$, while the second terms describes the energy increase associated to bond angles distortion from their reference values $\theta_{i,0}$. K_{b} and K_θ are the bond-stretching and the angle-bending force constants, respectively, and each particular type of bond and angle is characterized by its peculiar force constant. Both the first and the second terms of Eq. (1) are generally described using a harmonic potential, but also other functional forms have been adopted. The third term is a torsional potential that describes how the energy changes as a function of a specific bond rotation, and is generally modeled by a periodic function. The last term of Eq. (1) is the non-bonded potential, which describes the interactions between atoms *i* and *j*. Generally, the non-bonded term is neglected when the atoms *i* and *j* belongs to the same molecule and are not separated by at least three bonds. In the Eq. (1) the van der Waals interactions are modeled by a Lennard-Jones potential, while the electrostatic interactions are described by a Coulombic potential.

Even though the terms in the Eq. (1) are common to almost every MM force fields, some force fields include additional terms or slightly different functional forms. As an example, MM forcefields used to model biomolecules can adopt a distant-dependent dielectric constant to model the electrostatic interactions when the solvent molecules are not considered explicitly. Other force fields contain an additional term to model the polarization effects due to neighboring atoms.

All constants appearing in the forcefield are obtained from experimental data or ab initio calculations. Indeed, it is important to note that the database of compounds used for parameterization of the forcefield is crucial to the success of MM calculations. In particular, a force field parameterized using a specific class of molecules, for instance proteins, is generally expected to perform satisfactorily only when used to describe other proteins.

The MM formalism is used in a wide range of contexts related to molecular modeling. The most common applications include energy minimization, which plays an important role in the refinement process of structure determination by X-ray crystallography or NMR, and docking calculations of drugs to biological macromolecules.

More generally, classical potential energy methods are used to study the conformational properties of molecules. This kind of investigations implies always the stepwise modification of the molecular structure, in order to minimize the potential energy, a process usually referred to as geometry optimization.

Keeping in mind that molecular energy is a function of the molecule degrees of freedom (i.e. bonds, angles, and dihedrals), conformational energy searching methods can be used to find the low-energy conformations of a molecule, which is mathematically equivalent to locating the low-lying minima of its energy function.

Systematic energy sampling is, in principle, the most suited method for searching conformational energy space. According to this approach, energy is sampled over the entire range of each degree of freedom at regularly spaced intervals. The major problem associated to systematic searching is related to the observation that the number of conformations required to adequately sample the conformational space of a molecule can be extremely large and computationally very demanding.

Energy minimization methods can locate minimum energy conformations by "homing in" on the energy function minima. The goal of energy minimization is to move from the initial conformation to the nearest minimum energy conformation using the smallest number of calculations. The minimization of the potential energy function (i.e., geometry optimization) involves a search for the minimum of a function, and, to be efficient, usually requires calculations of derivatives of the potential energy as a function of independent variables (atomic coordinates). In general, all geometry optimization methods are iterative. During minimization, the energy of molecules is lowered by changing atomic coordinates. These methods require on input the initial atomic coordinates of the molecule under investigation, and provide a better estimate for the energy-minimum structure as a result. This corrected estimate is used as an input into the next iteration and the process is continued until there is no significant improvement in the criteria adopted to evaluate convergence.

Various algorithms are used to determine how the geometry will change from one step of optimization to the next. The most common methods are: steepest descent, the Newton-Raphson, the simplex and the Fletcher-Powell. Often, the optimization procedure implies the combination of two methods in order to take advantage of the specific features of each approach, e.g. steepest descent is easy to implement in computer codes but it is very slow to converge on a shallow potential energy surface. However, steepest descent methods are excellent at correcting major abnormalities often present at the start of a calculation. During optimization, the program keeps altering the geometry until a specified cutoff value of a parameter is reached. The specified cutoff value is termed the convergence criterion. Common convergence criteria are based on change in energy or change in energy gradient between the last structures calculated.

Force fields used in MM calculations can be broadly divided into two classes: Forcefields used to investigate small molecules, in which all atoms are included in the calculation (referred to as “all atom” forcefields), or forcefields used to investigate large biological molecules, such as proteins and nucleic acids, in which the majority of hydrogen atoms are removed from the molecule in order to decrease computational cost

(referred to as “united atom” forcefields). In “united atom” forcefields, only hydrogen atoms bonded to atoms different from carbon are explicitly considered, because they can be involved in hydrogen bonding. To compensate for the lack of C-H bonds, carbon atoms have an expanded van der Waals radius which accommodates the missing hydrogen atoms. A widely used MM forcefield for small organic molecules is MM2, whereas frequently used forcefields suited for large molecules are AMBER, CHARMM, GROMOS and GROMACS.

2.2 Molecular Dynamics

It should be noted that MM calculations are restricted to the evaluation of static properties of molecules and information related to properties that depend on the dynamics of the molecular systems cannot be directly inferred by MM calculations. As a matter of fact, the classical physics approach can be used not only to infer structural (static) properties of organic molecules, but also to investigate their dynamic behavior. In particular, in molecular dynamics (MD) simulations, the Newton equation of motion of an atomistic model is solved in order to obtain information about the dynamic and conformational properties of a molecule. MD simulations are widely used in several fields such as material science, biochemistry and biophysics, and are considered another important tool in structure determination and refinement by crystallography and NMR.

In MD simulations, the Newtonian equations of motion, a potential energy function and the associated force field are used to follow the movement of atoms in a molecule over a certain period of time, at a certain temperature and a certain pressure. The interaction between the atoms is usually described using the forcefields employed in MM. Atomic positions and velocities are computed at discrete and small time intervals and are used to calculate positions and velocities for the next step. Starting velocities can be computed randomly from a Maxwellian distribution at a fixed temperature, or by scaling the initial forces on the atoms.

The generation of several conformations of the system, obtained integrating the Newton law of motion, results in the collection of a trajectory, which contains information about the instantaneous position and velocities of the atoms in the system, as well as their evolution as a function of time.
The motion of the atoms is computed by calculating the forces on each atom, calculated using the adopted MM force field, according to the Newton’s second law:

$$F_i = m_i a_i \qquad (2)$$

where F_i is the force exerted on the atom, m_i the atom mass and a_i the acceleration of the atom i.
By numerical integration, F is evaluated at every integration step. The force can also be expressed as the gradient of the potential energy:

$$F_i = dV / dr_i \qquad (3)$$

where V is the potential energy. Therefore, the Newton's equation of motion can relate the derivative of the potential energy to the changes in position, as a function of time:

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \tag{4}$$

When the Newton equations of motion are integrated, the limiting factor for the simulation time step is the highest vibrational frequency that occurs in the system. In organic molecules, the vibrations of bonds involving hydrogen atoms are characterized by the highest frequency, thus limiting the time-step in MD simulations to about 0.5-1 fs. The introduction of restraints on bond lengths involving hydrogen atoms allows us to increase the time step to values larger than 2 fs. Since bond vibrations are generally uncoupled from other vibrations in the system, the restraints introduced on bond vibrations do not alter the relevant dynamic behavior of the system. However, the above observation does not hold true for bond-angle bending. In fact, bond-angles restraining can have a severe effect on relevant (global) dynamic properties of molecules.

Another strategy to reduce the computational cost due to the high frequency vibrations of bonds including hydrogen atoms implies the exclusion from the actual integration of hydrogen atoms, whose position is regenerated at the end of every time step from the positions of the heavy atoms to which they are bonded. Although the reliability of this approach has yet to be fully explored, results suggest that the integration time can be extended up to 6-8 fs.

The most computationally demanding step in MD simulations is the evaluation of electrostatic contributions, since computation of all pairwise non bonded interactions is time consuming and Coulomb terms are inversely proportional to the inter-atomic distances of charged particles. The latter factor makes electrostatic contribution to the total force non-negligible even at fairly large distance (above 10 Å). Several methods have been proposed to reduce the computational cost to calculate long-range electrostatic interactions. The most straightforward of these approaches are the so-called cut-off methods, in which all interactions beyond a certain radius are simply neglected. This approximation reduces the order of complexity from N^2 to N (where N is the number of atoms). However, artifacts can be observed at the edge of the cut-off radius. The Ewald method allows us to calculate electrostatic interactions in a more elegant way by calculating infinite lattice sums, but the $N^{3/2}$ order of complexity makes the method computationally expensive for simulation of large biomolecular systems. However, particle-particle and particle-mesh approximations scale with $N\log(N)$ and have recently given encouraging results.

MD calculations on biological molecules are frequently carried out considering explicitly the presence of solvent. However, this brings further complications due to two main problems. The first being increased CPU time due to the larger number of atoms. The second is related to the fact that the water molecules surrounding the molecule tend to drift away from the molecule of interest, causing nasty "edge effects". The most commonly adopted strategy to get rid of this problem implies to place the solute,

surrounded by a suitable number of water molecules, in a box of specific size and then to surround that box with an image of itself in all directions (periodic boundary conditions). The solute in the box of interest only interacts with its nearest neighbor images. Since each box is an image of the other, when a molecule leaves a box its image enters from the opposite side of the box in order to assure conservation of the total number of molecules and atoms in the box.

-
-
-

Bibliography

Karplus M. and McCammon J. A. (2002) Molecular dynamics simulations of biomolecules. *Nature Structural Biology* **9**, 646-652 [This article provides an overview of the applications of molecular dynamics in the investigation of biomolecules]

van Gunsteren W. F. and Berendsen H. J. C. (1990) Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angew. Chem. Int. Ed. Engl*. **29**, 992-1023 [This article provides an overview of molecular dynamics simulation methods and their relevance in chemistry]

Rauk A. (1994) *Orbital Interaction Theory of Organic Chemistry*, John Wiley & Sons, Inc., New York. [This book provides the fundamental notions of orbital interaction theory and its application to organic chemistry].

Cramer C. J. (2002) *Essentials of Computational Chemistry – Theories and Models*, John Wiley & Sons Ltd, Chichester, West Sussex, England. [This book provides the fundamental notions of several aspect of computational chemistry].

Koch W. and Holthausen M. C. (2002) *A Chemists's Guide to Density Functional Theory*, Second Edition, Wiley-VCH Verlag GmbH, Weinheim, Germany. [This book provides the fundamental notions of density functional theory relevant to investigate chemical problems].

Zerner M. (1991) *Reviews in Computational Chemistry*, Vol. 2, D.B. Boyd & K.B. Lipkowitz, Eds., VCH Publishers, New York. [This contribution provides the fundamental notions of Semiempirical Molecular Orbital Methods]

Stewart J. P. (1990) *Reviews in Computational Chemistry*, Vol. 1, D.B. Boyd & K.B. Lipkowitz, Eds., VCH Publishers, New York. [Another contribution providing the fundamental notions of Semiempirical Molecular Orbital Methods]

Biographical Sketches

Giuseppe Zampella was born in 1973 in Milan, Italy. He received his undergraduate education in chemistry at the university of Milan. In 1998, he joined the molecular modeling group of Prof. Piercarlo Fantucci at the university of Milano-Bicocca, where he got his Master degree in Bioinformatics in 2002. Since 2003 he has been research scientist in the same university and has been collaborating with Prof. Luca De Gioia on research topics covering several areas of computational chemistry. His interests are

focused on ligand-receptor interactions in biosystems such as the G-Protein Coupled Receptors (GPCRs) and structure-function relationships in metallo-enzymes cofactors such as the Vanadium dependent Haloperoxidases, [NiFe]-hydrogenases and [Fe]-hydrogenases, investigated by means of quantum chemistry.

Luca De Gioia is presently professor of General and Inorganic Chemistry at the University of Milan-Bicocca, where he is also involved in teaching computational chemistry and bioinformatics. Recently, he has been involved in the investigation of structure-activity relationships in organic biomolecules and metal-containing proteins and models systems. He is co-author of more than 70 papers published on international scientific journals in the fields of protein chemistry and computational chemistry.