Molecular Mechanics

I. Quantum mechanical treatment of molecular systems

The first principle approach for describing the properties of molecules, including proteins, involves quantum mechanics. For example, consider an atom, which consists of nuclei and electrons. On can write a time-independent (stationary) Schrödinger equation

$$\hat{H}\psi = E\psi ,$$

where

$$\hat{H} = -\frac{\hbar^2}{2m}\nabla^2 + V$$

is the Hamiltonian, which includes the contribution from kinetic energy (the first term) and potential energy (the second term). ψ and E are the wavefunction and energy of the system (including both, kinetic and potential, parts), which can be treated as the eigenfunction and eigenvalue of the operator $\hat{H}$. The wavefunction (more precisely $|\psi|^2$) describes the spatial probability distribution for a particle.

Generally, Schrödinger equation can be analytically solved for a very limited set of problems, such as hydrogen atom. For molecules, especially macromolecules, direct numerical solution of Schrödinger equation is computationally unfeasible. To simplify the quantum description of molecular systems, a Born-Oppenheimer (BO) approximation is applied, which assumes that motions of nuclei and electrons can be decoupled due to large disparity in their masses. This implies that nuclei is considered fixed and only the motions of electrons are considered. As a result E and ψ in the Schrödinger equation now represent electronic properties. BO approximation serves as foundation for *ab initio* and *semi-empiric* quantum approaches.

Ab initio approach neglects all the relativistic effects and further assumes that the molecular orbitals (i.e., the wavefunctions of the entire molecule) can be represented as a linear combination of atomic orbitals, which are essentially electron wavefunctions. The related coefficients in the linear combinations are determined by energy minimization. The energy of a molecule constructed in this way is referred to as Hartree-Fock energy. The Hartree-Fock energy can then be used to adjust the positions of nuclei to reflect the computed molecular orbitals. After this Hartee-Fock energy must be again recomputed. This process is repeated self-consistently until no further adjustments to the nuclei positions are necessary. The *ab initio* method is computationally expensive and is suitable for the systems containing < 1000 atoms. *Semi-empirical* method uses additional approximation by assuming that the matrix elements associated with the combinations of various wavefunctions ψ_n and ψ_m may be obtained from the predetermined set of atomic orbitals, which are consistent with the experimental data. Because *semi-empirical* method

does not require computations of complex integrals of wavefunctions, larger quantum systems can be examined (number of atoms < 10,000). An interesting simulation methodology is based on the combination of quantum and classical approaches. For example, protein atoms may be considered using quantum approach, whereas solvent and the interactions of solvent with a protein are treated classically. The advantage of quantum description is that high frequency motions or bond formation/breakage can be studied, yet classical simulations of solvent significantly reduce the computational burden. The example of such mixed, quantum/classical simulations is reported for crambin. Quantum computations were performed using semi-*empirical* method and the crambin with explicit water was simulated for 350 ps (*Protein Structure Function Genetics* **44**, 484 (2001)).

II. Applicability of classical approach

Generally, molecular properties have quantum characteristics at low temperatures or when high-frequency (fast) motions are involved. Let us determine an approximate timescale of characteristic motions in a protein, which serves as a border between quantum and classical descriptions. The classical physics is applicable, if the thermal energy

$$k_B T >> h\nu,$$

where $h\nu$ can be thought of as energy gap between quantum energy levels (h is the Planck's constant). Therefore, the frequency of molecular motions, at which classical description is not valid, is $\nu_q = k_B T/h = 1.38\cdot10^{-23}$ J $\cdot$K^{-1} $\cdot$300K/$6.62\cdot10^{-34}$ J$\cdot$s $\approx$6 ps^{-1}. Therefore, the characteristic timescale

$$\tau_q = \nu_q^{-1} \approx 0.2\text{ps} \sim 10^{-13}\text{s}$$

and any motions taking place on a timescale faster than τ_q are quantum in nature.

Consider now a covalent bond between an oxygen and hydrogen atoms and evaluate classically the frequency of vibration of this bond. In the CHARMM force field the hydrogen atom is linked to oxygen through the bond-length potential $V=k_a(r-a)^2$, where r is the H-O distance, k_a=400 kcal/(mol$\cdot$Å^2) is the spring constant, and a is the equilibrium distance between H and O. Applying Newton equation of motion

$$m\ddot{r} = -\frac{\partial V}{\partial r} = -2k_a(r-a)$$

and taking into account the fact that

$$\nu = \frac{1}{2\pi}\sqrt{\frac{2k_a}{m}},$$

where m =1.67·10^{-27} kg is the hydrogen mass, we find

$$\nu = \frac{1}{2\pi}\sqrt{\frac{2\cdot 400 J/(mol\cdot m^2)\cdot 10^{20}\cdot 4184}{1.67\cdot 10^{-27} kg\cdot 6.02\cdot 10^{23} mol^{-1}}} \approx 90\, ps^{-1}. \quad (1)$$

Therefore, the characteristic time scale is $\tau=\nu^{-1}$ = 0.01ps=10^{-14}s. We see now that the time scale of hydrogen motion is about an order of magnitude faster than the time scale of the crossover between quantum and classical physics. Therefore, the motions of hydrogen must be described by quantum physics. The same conclusion is applicable for the covalent bonds between heavy atoms (other than hydrogens) and for the bond angle potential.

There is, however, a solution designed to retain classical energy terms, which is based on applying a stiff spring potential (to prevent large bond length deviations) or even simply constraining the distance r to a. Indeed, rigorous quantum calculations show that the fluctuations in r, $<\Delta r^2> \sim 0.1$Å or about 10% of the equilibrium bond length that justifies keeping the stiff classical potential for bond lengths. Still one has to be cautions on drawing any conclusions from the fast motions (on the time scales less or about τ_q =0.2ps), because these may potentially be the artifacts of classical approximation.

If the vibrations of atom around their bond lengths fall in the realm of quantum mechanics, what are the types of motions, for which classical description suffices? The local motions in proteins span a wide time scale range from 10^{-15} to 10^{-1} s and include atomic fluctuations, side chain or loop motions. Thus, only extremely fast motions of this type are quantum in nature, while most of them can still be treated classically. Furthermore, classical description is clearly sufficient for rigid body motions, which involve helix and strand motions, rearrangements of several secondary structure elements in protein etc. The times scales of these motions is from 10^{-9} to 10^{1} s. Structural transitions in proteins (helix formation, folding and unfolding, aggregation) are also well above the boundary between classical and quantum physics.

The arguments presented above outline the applicability of classical description of molecules and demonstrate that many biological processes can be considered on the basis of classical physics. This lays the foundation of *molecular mechanics*.

III. Basic principles of molecular mechanics

The main idea behind the *molecular mechanics* (MM) is that a molecular system can be viewed as a microscopic mechanical system. According to this idea the atoms in the molecular system are linked by mechanical springs, which control their covalent bonds, angles between successive bonds, rotations around the bonds, etc. Atoms interact with each other (attract or repel) according to classical non-bonded potentials, which determine the non-bonded interatomic forces. In mathematical terms, this description requires the construction of a potential function incorporating exclusively classical terms.

This potential (or energy) function is then used to compute all the relevant forces for the Newton equation of motions, which ultimately describe the microscopic dynamics of the molecular system.

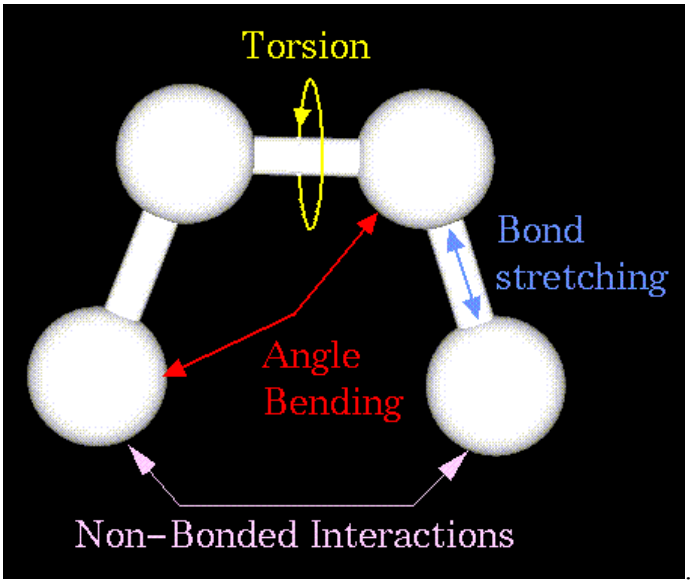

Fig. 1 The idea of MM is illustrated for four atom fragment of polypeptide chain. Four types of interactions indicated make up the basis of energy function in MM.

There are three fundamental principles of MM, namely, *thermodynamic hypothesis*, *additivity*, and *transferability*.

Thermodynamic hypothesis states that the native state of a protein corresponds to the minimum of potential energy. This is essentially a rephrased version of Anfinsen idea that folding of proteins is largely controlled by thermodynamic principles. A unique native state is encoded in the protein sequence, because the complex network interactions in the heterogeneous polypeptide chain guides a protein toward a native conformation, which is the structure of minimum energy. These ideas also lay in the center of "new view" on protein folding. Although it is impossible to prove this assertion rigorously, the experimental and theoretical data support it. Spontaneous folding of many single domain proteins does not rely on any "external" help and appears to be following "internal" sequence instructions. Native structures of protein are generally stable (albeit not always) against mutations, which tend to increase their native energies.

Additivity assumes that the total potential function E_p may be decomposed into individual terms associated with electrostatic, van-der-Waals, and mechanical-like terms describing bond stretching, bond angle fluctuations and rotations. In other words, E_p can be separated into local and non-local energy terms. The local part of E_p is

$$E_p^{local} = \sum_{bonds} V_{BL} + \sum_{angles} V_{BA} + \sum_{dihedrals} V_{DA}\,, \qquad (2)$$

where V_{BL}, V_{BA}, V_{DA} are bond-length, bond-angle, and dihedral potentials, respectively. Note that no cross terms are usually considered in Eq. (2), i.e., possible couplings between the bond stretching and bond angles are neglected. Because the total number of bonds, angles and dihedrals is proportional to the length of protein, E_p^{local} scales with N, the number of amino acids, as N^1. The non-local (non-bonded) part of E_p is

$$E_p^{non-local} = \sum_{ij} V_{ij} \,, \tag{3}$$

where V_{ij} are non-bonded interactions between atoms i and j, including electrostatic and/or van-der-Waals interactions. It is clear from Eq. (3) that $E_p^{non\text{-}local}$ scales as N^2 that turns their calculations for large molecular systems in computational bottleneck. Note also that Eq. (3) assumes that all terms in non-local potential energy are represented by pairwise interactions. This assumption is generally not correct, because, for example, electrostatic interactions between a pair of atoms affect the charge distribution on other atoms (through polarization effects). Therefore, the more consistent form of Eq. (3) should include multi-body terms as well. However, in the majority of current force fields these polarization effects are neglected and only two-body terms are incorporated. The *additivity* of potential functions may be exploited in the design of multistep algorithms, which take into account faster variations in local terms as compared with non-local interactions.

Transferability hypothesis suggests that the properties of atoms in large molecules can be deduced from the study of representative set of small molecules. The derived energy parameters are then transferred without modification to proteins or other larger and more complex molecular systems. Roughly speaking, the *transferability* implies that the bond lengths, angles etc are the same in small "test" molecules as well as in much larger proteins. This approach is employed in the development of all potential functions for proteins. For example, the parameters for the potential function ("force field") OPLS (Optimal Parameters for Liquid Simulations) were derived by Jorgensen and coworkers by performing Monte Carlo simulations of 36 organic molecules and their water solutions, which have similar structures to various protein side chains (*JACS* **110**, 1657 (1988)). The aim of Monte Carlo simulations was to choose such values of interaction parameters for the organic molecules, which closely reproduce their experimental data on density or heats of vaporization. The OPLS parameters for the potential function were then shown to reproduce the fragments of X-ray native structures of proteins.

The properties of atoms depend on the specific local environment, such as bond character or electron hybridization state. For example, the properties of a carbon atom depend on whether the atom is a part of aromatic or aliphatic side chain. Similarly, hydrogen may adopt polar or non-polar properties depending on the local environment. The solution to this problem is the use of *atom types*, i.e., to assign properties to the same atom depending on particular environment. One of the standard current force fields, CHARMM22, uses 57 different atom types. Specifically, there are 21 different carbons, 12 hydrogens, 11 nitrogens, seven oxygens, etc.

The fragment of the CHARMM22 topology file is shown in Fig. 2. Hydrogen atom H appears as polar hydrogen (atom type H), which is found in backbone amides (Figs. 2 and 3), or as non-polar hydrogen (atom type HA), which is used in Val or Ala side chains (Figs. 2 and 3). There are also aromatic hydrogens (atom type HP) used in aromatic side chains or backbone hydrogen (atom type HB) bonded to C_α-carbon (Figs. 2 and 3). Similarly, carbon C (atom type C) is used in peptide backbone (carbonyl group, Figs. 2 and 3), aromatic C (atom type CA) appears, e.g., in Phe side chain. Several aliphatic carbons (atom types CT1, CT2, CT3), which are used in backbone or Val side chains, are also included (Figs. 2 and 3).

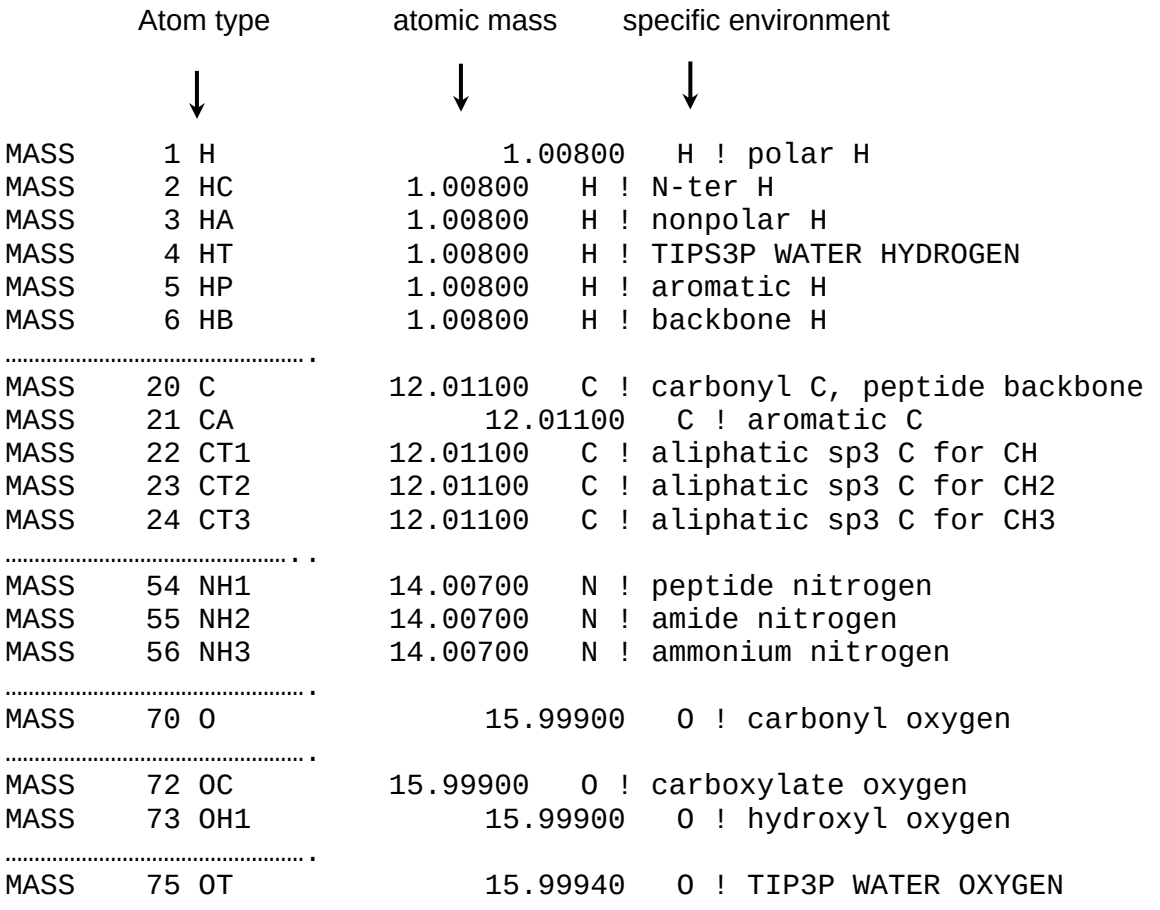

Fig.2 Fragment of CHARMM22 topology file, showing different atom types.

It is necessary to differentiate the atom types from atom names as they are used in the coordinate files (Figs. 3 and 4). The atoms names must be unique for a specific residue, whereas an atom of a particular type may appear in many amino acids in a given protein. Note also that atom type is defined by a particular set of van-der-Waals parameters, whereas the specific partial charge depends in a particular environment. Consider, for example, C_α- and C_β- carbons in Val amino acid. Both are represented by an aliphatic atom type CT1, which carry nevertheless different partial charges (Fig. 3).

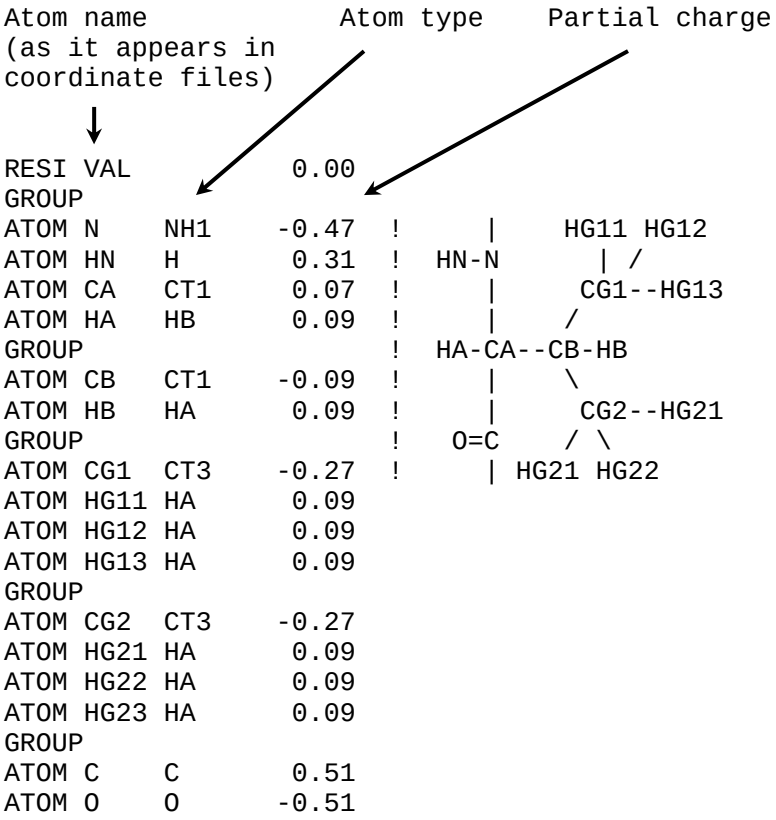

Fig. 3. Structure and composition of Val residue from CHARMM22 force field.

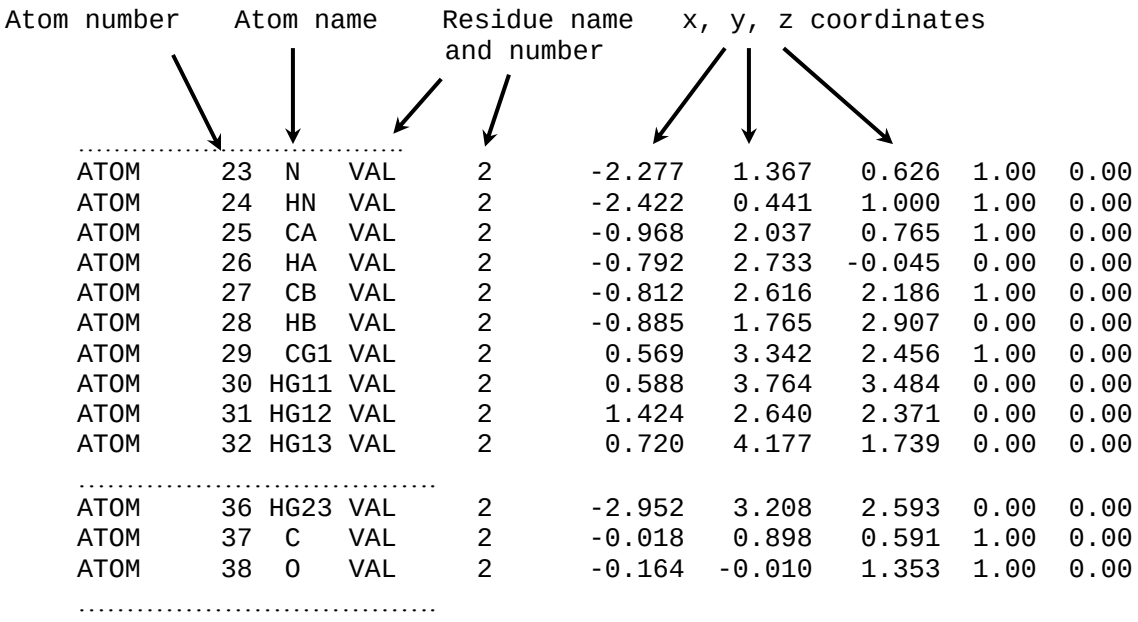

Fig. 4 Fragment of coordinate file containing Val coordinates as used in CHARM22 MD simulations.

The CHARMM22 or OPLS force fields are reasonably successful in capturing many properties of proteins. One of the directions of further improvement is the incorporation of more realistic distribution in partial charges. In CHARMM22, OPLS, and many other force fields partial charges are fixed at the centers of atoms. For example, in TIP3P water (Fig. 5) $+q$=0.417e charges are assigned to hydrogens and $-2q$=-0.834e is placed on oxygen. This assumption distorts higher order electric moments, which could be correctly included if partial charges are placed off the nuclei centers. Furthermore, in the force fields with fixed partial charges the radial distribution of electrons is assumed isotropic (spherical). For example, the water oxygen has four sp3 pairs of electrons, two of which are engaged in covalent bonding with hydrogens. Two other electrons make up the two lone pairs used in hydrogen bond formation, but their spatial distribution is highly anisotropic. Yet in the TIP3P water model in CHRAMM22 this anisotropy and even the total number of free lone pairs are not taken into account. As a result the number and geometry of HBs may become distorted.

The fact that water forms the right number of HBs in CHARMM22 force field (i.e., four) is a consequence of purely steric effects. Finally, many atoms or molecules may experience polarization effects when subject to external electric field. The polarization would change the distribution and the value of partial charges. New versions of force fields, which take into account the effects outlined above, are being developed, but they have not become the mainstream of MD simulations as yet.

```
RESI TIP3         0.000 ! tip3p water model
GROUP
ATOM OH2  OT     -0.834
ATOM H1   HT      0.417
ATOM H2   HT      0.417
```

Fig. 5 Representation of TIP3P water in CHARMM22 force field.

It is important to keep in mind that MM is used to develop a force field (i.e., a potential function), which can later be used for Monte Carlo, molecular dynamics, minimization calculations, or Langevin dynamics. In other words, MM itself does not make any assumptions about the type of dynamics used in the molecular simulations. In the case of molecular dynamics, the potential function constructed using MM principles is used to compute the forces in the Newton equation of motions which act upon the atom i

$$m_i \ddot{\vec{r}}_i = \vec{F}_i = -\frac{\partial E_p}{\partial \vec{r}_i} \quad ,$$

where $\vec{r}_i$ is the radius vector of an atom i.