BS3008: Computer Aided Drug Discovery

Kevin Fo

Table of contents

Preface

This is a Quarto book website (authored in the form of a website) that I (i.e., Kevin) has authored for the SBS module *BS3008: Computer Aided Drug Discovery* (this module was formerly known as "Computational Biology"). As of the time of writing, BS3008 is a three **a**cademic **u**nit (i.e., **AU**) module available to SBS students as a core module.

BS3008 taught by professor Mu Yuguang.

Part I

PART 1 : LECTURES

1 Theoretical Foundations of BS3008

This week's (i.e., week 1) lecture aims to provide an introduction to BS3008's course contents by explaining its various theoretical aspects.

1.1 Discplines in BS3008

BS3008 covers numerous disciplines, including the following. A brief explanation on what each discipline is is also provided for each of the disciplines:

1. **Chemoinformatics**

   This discipline deals with similarities and differences between chemical compounds.

   Chemoinformatics deals with compounds from 10^{-60} to 10^{-30} in magnitudes. Individuals who work in this field try to find "an island in an ocean" - they try to find a molecule that can do *some* purpose.

2. **Bioinformatics**

   This disciplines applies informatics tools (e.g., Python coding) to Biological molecules and data.

   Bioinformatics mainly focuses on Biological modelling.

3. **Theoretical Chemistry (i.e., Quantum Chemistry)**

   This discipline provides the theoretical foundations needed to understand the course's contents.

4. **Computational Chemistry and Biology**

   This discipline not only encompasses theoretical chemistry, but also molecular mechanics, minimization, simulations, and conformational analysis.

5. **Molecular Modelling**

   This discipline uses all of the above disciplines to represent and manipulate the structures of molecules.

   This also means that this discipline uses physics to model a system - that way, a model can be compared against experimental results.

Hence, BS3008 primarily focuses on molecular modelling (with emphasis on theoretical chemistry for the theoretical component of the course).

1.1.1 What is Molecular Modelling?

According to Tamar Schlick, molecular modelling is:

> "...the science and art of studying molecular structure and function through model building and computation."
>
> – Tamar Schlick

"Computation" in this sense refer to practices such as:

1. ab initio and semi-empirical quantum mechanics
2. Molecular mechanics
3. Monte Carlo simulations
4. Molecular dynamics
5. Free energy and solvation methods
6. Structure / activity relationships (i.e., SAR analyses)
7. Chemical / biochemical information and databases

It is important to understand that while "model building" can be as simple as using plastic or metal rods to depict molecules' structures, it can also be as sophisticated as an interactive, animated color graphics and lasers.

Nonetheless, the computational tools used in molecular modelling is just as, if not more complex than Biological systems. However, the concepts in molecular modelling must be carefully applied and one must also be wary of molecular modeling's strengths and weaknesses.

1.1.2 Important Databases and Tools

Professor Mu lists some important molecular modelling tools in this chapter - click on their hyperlinks to access them:

1. **PDB**

   This is a database with numerous entries on proteins' information.

2. **PDBBinding**

   This is another database that provides entries on the binding affinity for all biomolecular complexes.

3. **ZINK DOCK**

4. **Autodock Zina**

   This is an open-source program for performing molecular docking.

1.1.3 Molecular Mechanics in Molecular Modelling

Molecular modelling started with the idea that molecular geometry, energy, and other molecular properties could be calculated from models (that are influenced by basic forces).

A **molecule** - hence - is a system of particles (i.e., atoms) connected by springs (i.e., bonds). This molecule is free to rotate, vibrate, and adopt a favorable conformation in space as a result of the inter- *and* intramolecular forces acting upon it.

1.2 Structure, Topology, Motion, Functions, and Potential Energy

This section aims to illustrate how different energy functions can influence the behavior of particles in a system.

1.2.1 Case #1

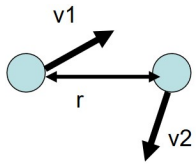

Figure 1.1: A Simple System to Consider

The potential and kinetic energy E_p and E_k respectively in this system is given by:

$$E_p = E_p(\vec{x}) = \sum_i f_i(x, y, z) \tag{1.1}$$

$$E_k = \frac{1}{2} \cdot (m_1 v_1^2 + m_2 v_2^2) \tag{1.2}$$

Where $\vec{x}$ is the system and $f_i(x, y, z)$ a function that calculates the potential energy for each particle in the system (i.e., each atom). Hence, we can say that:

$$E_{tot} = E_p + E_k \tag{1.3}$$

Where E_{tot} is the total energy of the system.

Since r represents the distance between both molecules, therefore $E_p(r) = 0$.

$$\vec{F} = \frac{\partial E_p(\vec{x})}{\partial \vec{x}} \tag{1.4}$$

The force[1] $\vec{F}$ on the system is denoted via the above equation.

Since $E_p(r) = 0$ in the first figure, it follows that $F(r) = 0$.

1.2.2 Case #2

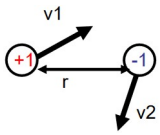

Figure 1.2: A System with Two Charged Molecules

Here, the charge V_{ele} and the potential energy $E_p(r)$ is given via the following equations:

$$V_{ele}(r) = \frac{1}{4\pi\epsilon_0} \cdot \frac{q_1}{r} \tag{1.5}$$

$$E_p(r) = \frac{1}{4\pi\epsilon_0} \cdot \frac{q_1 q_2}{r} \tag{1.6}$$

Some important considerations to think about include the variables and the parameters of the system.

1.2.3 Case #3

In this system, we let:

$$E_p(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{1.7}$$

In this system, we also note that $E_p(r) = 0$ when $r = \sigma$ and that $E_p(\sqrt[6]{2}\sigma) = -\epsilon$.

Do also consider the variables and parameters in this system (and whether or not both particles in this system can move freely).

[1] Note that $\vec{F}$ is caused by a change in potential energy, not by potential energy itself!

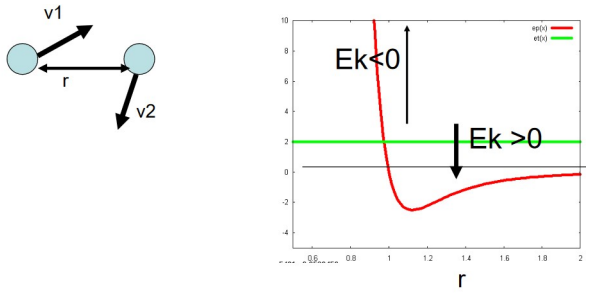

Figure 1.3: A System with Two Molecules and their Energy Graph

1.2.4 Case #4

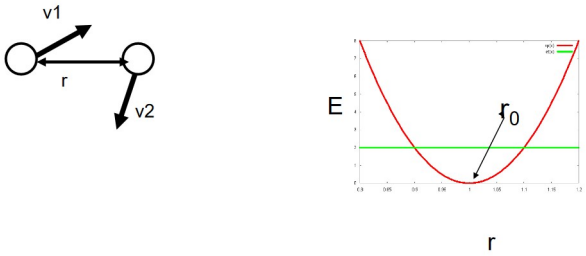

Figure 1.4: A System with Two Molecules and their Energy Graph

Here, we define the system's potential energy E_p as:

$$E_p(r) = \frac{1}{2}k(r - r_0)^2 \tag{1.8}$$

While the energy graph for this system appears to be that of bonding, it is still important to consider the variables and the parameters of the system.

We can also further decompose the above equation to its spatial components x, y, and z and say that:

$$E_p(r) = \frac{1}{2}k(x - x_0)^2 + \frac{1}{2}k(y - y_0)^2 + \frac{1}{2}k(z - z_0)^2 \tag{1.9}$$

In this sub-case, the system appears to be a lattice. However, are the particles still movable?

1.3 Professor Mu's Current Works

As of the time of writing, professor Mu's lab is currently focused on the following topics:

1. Amyloidogenic protein / peptide aggregation and misfolding
2. DNA-DNA, DNA-ions, and DNA-protein interactions
3. Drug-protein interaction and drug candidate screening
4. Peptide-membrane interactions.

For more information on professor Mu's current research topics, do visit his lab's homepage.

2 Fundamental Quantum Chemistry

Computational chemistry is a branch of chemistry that uses mathematical approximations and computer programs to obtain results (for chemical problems).

Computational *quantum* chemistry focuses specifically on equations that have been derived from principles in quantum mechanics (i.e., solving Schrodinger's equation for molecular systems).

Ab initio quantum chemistry uses methods that do not use any empirical data.

Computational chemistry is a growing field: computers are getting faster and so are algorithms. This discipline can be used to calculate the following parameters:

1. NMR spectra
2. Vibrational frequencies
3. Thermochemical data

The goal of this chapter is to provide a basic, high-level overview of quantum chemistry: necessary to use GaussianView and Gaussian 16.

2.1 Born-Oppenheimer Approximation

This technique in computational chemistry leads to a later concept: **potential energy surfaces** (i.e., **PES**). A **PES** is a graph that describes the energy of a molecule (i.e., a system) in terms of certain parameters

For molecules that have many electrons, their wavefunction is a combination of electron and nuclear coordinates. Their wavefunctions can be represented as $\psi(R, r)$, where R are the nuclear coordinates and r the electron coordinates.

However, nuclei are *way* heavier than electrons, nuclei also move much slower than electrons.

$$\psi(R, r) = \psi_{et}(r; R)\psi_N(R) \tag{2.1}$$

The **Born-Oppenheimer approximation** allows us to separate electrons and nuclear motion via the above wavefunction approximation.

2.1.1 Schrodinger's Equation

The time-independent Schrodinger equation is:

$$\hat{H}\psi = E\psi \tag{2.2}$$

$$\hat{H} = \hat{T} + \hat{V} \tag{2.3}$$

Where $\hat{H}$ is the Hamiltonian, ψ the wavefunction, and E the total amount of energy in the system.

T represents kinetic energy and V the potential energy.

2.1.2 Solutions of the Schrodinger Equation

The equation can only be solved for simple cases (e.g., particle in a box, hydrogen atoms, rigid rotors, etc). For more complex solutions, some assumptions will need to be made.

Solving this equation also attempts to expand the wave function ψ into one of many Slater determinants - these are represented by **molecular orbitals**: linear combinations of atomic-like-orbital functions.

Yet, it is still possible to get *very* accurate results. Generally speaking, the cost of calculation increases with the accuracy of the calculation (and the size of the system).

2.1.3 Electronic Schrodinger Equation

This equation is as follows:

$$\hat{H}_{el}\psi_{el}(r;R) = E_{el}\psi_{el}(r;R) \tag{2.4}$$

$$\hat{H}_{el} = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 - \sum_\alpha \sum_i \frac{Z_\alpha e'^2}{r_{i\alpha}} + \sum_j \sum_{i>j} \frac{e'^2}{r_{ij}} \tag{2.5}$$

In a typical potential energy graph, the potential energy takes a dip within a certain distance before it goes up:

From the above graph, the following formulas can be derived:

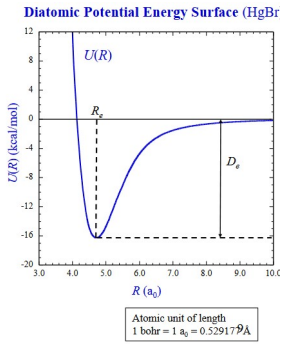

Figure 2.1: Potential Energy Graph for HgBr

$$U(R) = E_{el} + V_{NN} \tag{2.6}$$

$$V_{NN} = \sum_{\alpha} \sum_{\alpha>\beta} \frac{Z_\alpha Z_\beta e'^2}{r_{\alpha\beta}} \tag{2.7}$$

2.1.4 Nuclear Schrodinger Equation

Once a **p**otential **e**nergy **s**urface (i.e., **PES**) has been obtained for a molecule, one can solve the nuclear Schrodinger equation:

$$\hat{H}_N \psi_N(R) = E_N \psi_N(R) \tag{2.8}$$

$$\hat{H}_N = -\frac{\hbar^2}{2} \sum_{\alpha} \frac{1}{m_\alpha} \nabla_\alpha^2 + U(R) \tag{2.9}$$

The solutions of this equation allow one to determine a large amount of molecular properties - for instance, vibrational energy levels.

2.1.5 Polyatomic PESes

Where there are more than one parameters that are used to describe a molecule's PES, the degrees of freedom is given by $3n-6$, where N is the number of atoms in the system.

2.2 Basic Methods

2.2.1 Solving the Electronic Schrodinger Equation

$$\hat{H}_{el} = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 - \sum_\alpha \sum_i \frac{Z_\alpha e'^2}{r_{i\alpha}} + \sum_j \sum_{i>j} \frac{e'^2}{r_{ij}} \tag{2.10}$$

The final term $\sum_j \sum_{i>j} \frac{e'^2}{e_{ij}}$ represents electron interactions. - this is the reason that it is impossible solve the electron schrodinger's equation.

2.2.1.1 Electron Spin and Antisymmetry

All electrons are described by a spin quantum number. The eigenfunctions describing these spins are denoted as α and β.

Electron spin also obeys the following principles:

1. **Indisguinshability**

   All electrons that are spin-up are identical to one another.

2. **Pauli's Exclusion Principle**

   No two electrons can be described by the same set of quantum numbers.

Whenever two electrons are interchanged, the signs of their wavefunctions are also changed - for instance:

If a wave function $\psi = \psi_a(1)\alpha(1)\psi_b(2)\alpha(2)$ is inverted, the result is $\psi_a(2)\alpha(2)\psi_b(1)\alpha(1) - \psi_a(1)\alpha(1)\psi_b(2)\alpha(2)$.

2.2.2 Slater Determinants

After performing the Born-Oppenheimer approximation, one then determines the expansion of the electron wavefunction ψ_{el} via Slater determinants:

$$\psi_{el} = \sum_i d_i\phi_i = d_0\phi_0 + d_1\phi_1 + d_2\phi_2 + ... \tag{2.11}$$

The Slater determinant ϕ_0 is given via the below matrix. One can also choose to think about Slater determinants as a kind of "configuration" (i.e., ground-state neon might go something like $\psi_0 = 1s^2 2s^2 2p^6$ and so on).

$$\phi_0 = \frac{1}{\sqrt{N!}} \begin{bmatrix} \phi_1\alpha(1) & \phi_1\beta(1) & \phi_2\alpha(1) & ... & \phi_M\beta(1) \\ \phi_1\alpha(2) & \phi_1\beta(2) & \phi_2\alpha(2) & ... & \phi_M\beta(1) \\ ... & .... & ... & ... & ... \\ \phi_1\alpha(N) & \phi_1\beta(N) & \phi_2\alpha(N) & ... & \phi_M\beta(N) \end{bmatrix} \tag{2.12}$$

Where:

1. α and β are the spin-up and spin-down functions respectively.
2. ψ_i the spatial functions
3. $\psi_i\alpha$ and $\psi_i\beta$ the spin-orbitals

Slater determinants give proper anti-symmetry (i.e., the Pauli Exclusion Principle).

2.2.3 Hartree-Fock Approximations

It is impractical (and impossible) to consider all configurations of a system ϕ_i. Hence, the **Hartree-Fock** approximation is often used to approximate the wavefunction for the complete set of ϕ_is using a single determinant ψ_0.

In this method, **S**elf-**c**onsistent **f**ield energies (i.e., SCF energies) are used instead to find an optimal set of molecular orbitals for ψ_0.

Each electron in this approximation only sees an *average* repulsion of the remaining electrons.

2.2.3.1 How Accurate is the Approximation?

Spectroscopic Constants of CO (Total E_e≈-300,000 kJ/mol)

	R_e (Å)	ω_e (cm^{-1})	D_e (KJ/mol)
HF/cc-pV6Z	1.10	2427	185
Experiment	1.13	2170	260
% Error	2.7%	11.8%	28.8%

Figure 2.2: Total Accuracy of a Hartree-Fock Approximation on a CO Molecule

Hartree-Fock wavefunctions usually approximate about 99% of the total energy. Hartree-Fock approximations can also be used to predict bond angles, thermochemistry measurements (e.g., enthalpy), and even vibrational force constants.

Quantum chemists are typically interested in energy differences and not total energies.

2.2.3.2 Electron Correlations

The **electron correlation** is the energy difference between an experimentally measured value (i.e., the "exact" value) and value obtained from a Hartree-Fock approximation. In more empirical terms:

$$E_{corr} = E_{exact} - E_{HF} \tag{2.13}$$

The E_{corr} accounts for missing electron-electron interactions in the Hartree-Fock method.

Because of this, the Hartree-Fock approximation is often used as a "starting point" for finding the wavefunction.

Do also note that different correlation methods also exist - depending on which values of ϕ_i and d_i to use, the value of E_{corr} might change.

2.2.3.3 Configuration Interactions

First suppose the following about electron wavefunction:

$$\psi_{el} = d_0\phi_{HF} + \sum_{i=1} d_i\phi_i \tag{2.14}$$

The above assumption uses the **linear variation principle**: the amount energy in a wavefunction is always equal to or great than the true energy.

Some other possible configurations (i.e., methods) include:

1. **CISD**

   This is short for **C**onfiguration **I**nteraction with **S**ingle and **D**ouble excitations. All determinants of "s" and "d" type orbitals are used.

2. **MRCI**

   This is short for **M**ulti**r**eference **C**onfiguration **I**nteraction.

Both CISD and MRCI can be very accurate, but they also take a long time to process.

2.2.4 Mollet-Plesset (i.e., MP) Perturbation Theory

Perturbation methods such as these assume that the problem at hand (i.e., ψ and E) only differ *slightly* from a solved problem (i.e., HF_{ψ} and E).

The energy is calculated via various orders of approximations. At MP2, the calculation involves single and double excitations; higher MPs result in costlier calculations.

2.2.5 Coupled Cluster (i.e., CC) Theory

This theory leads to accurate wavefunction expansions (and yields accurate electronic charges).

Some common variants of this include:

1. **CCSD**

   Which includes single and double CCs, hence its name.

2. **CCSD(T)**

   This is the same as CCSD, albeit with treatments of triple excitations. This method uses accurate results when used on large basis sets.

   It's possible to get thermochemical results (i.e., enthalpy) with good accuracy.

2.2.6 Frozen Core Approximations

Here, only valence orbitals are involved in chemical bonding processes. Core orbitals don't change much when atoms are involved in molecules (as opposed to when atoms are free).

Because of this, most electronic structure calculations only correlate with the frozen electrons (i.e., the core orbitals are kept frozen).

2.2.7 Density Frozen Theory

The wavefunction is generally uninterpretable[1].

Nonetheless, a useful physical observation would be an atom's electron density ρ as it gives the total number of electrons over space.

Density Functional Theory solves for electron density. The cost of such method is similar to Hartree-Fock approximations (and usually involve some of empirical parameterization).

This approach is often the best choice when dealing with large molecules.

[1] i.e., even scientists don't fully know what it means - this is still a matter of debate over quantum chemists.

2.3 Basis sets

2.3.1 Linear Combination of Atomic Orbitals-Molecular Orbitals

Given the Slater determinant ϕ_0:

$$\phi_0 = \frac{1}{\sqrt{N!}} \begin{bmatrix} \phi_1\alpha(1) & \phi_1\beta(1) & \phi_2\alpha(1) & ... & \phi_M\beta(1) \\ \phi_1\alpha(2) & \phi_1\beta(2) & \phi_2\alpha(2) & ... & \phi_M\beta(1) \\ ... & .... & ... & ... & ... \\ \phi_1\alpha(N) & \phi_1\beta(N) & \phi_2\alpha(N) & ... & \phi_M\beta(N) \end{bmatrix} \tag{2.15}$$

The spatial functions ϕ_i can be calculated via the following equation:

$$\phi_i = \sum_k^M c_{ki}\chi_k \tag{2.16}$$

Where c_{ki} molecular orbital coefficients that are calculated in the SCF procedure previous outlined and χ_k atom-centric functions that mimic solutions of hydrogen atoms (i.e., s and p orbitals).

Just like other calculations, the more accurate a calculation is, the more expensive it becomes.

2.3.2 Gaussian Type Orbitals

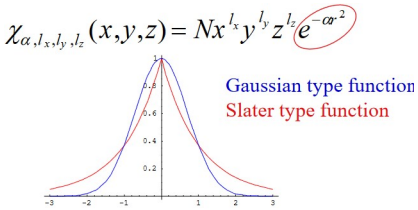

Figure 2.3: Graph of a Slater-Type Function Against a Gaussian-Type Function

l_x, l_y, and l_z in the above graphic determines the kinds of orbitals in question (e.g., $l = 1$ represents a "p" orbital).

The solutions of a hydrogen atom come in the form of a Slater-type function (most of which are electronic structure theory calculations):

$$\chi_{\alpha,n,l,m}(r,\theta,\phi) = NY_{l,m}(\theta,\phi)r^{n-1}e^{-\alpha r} \tag{2.17}$$