UNIVERSITÁ DEGLI STUDI DI TORINO

SCUOLA DI SCIENZE DELLA NATURA

CORSO DI LAUREA IN CHIMICA E TECNOLOGIE CHIMICHE

Tesi di Laurea di I Livello

Semiempirical Quantum-mechanical Methods for the Study of Proteins

Candidato: Davide La Vardera 716297

Relatore: Prof. Piero Ugliengo

Anno Accademico: 2013-2014

Contents

Abstract

The direct application of the Hartree-Fock model to the study of systems of biological interest is impractical because of the high number of two-electron integrals that need to be calculated explicitly in order to construct the Fock matrix. A great computational reduction can be achieved with semiempirical SCF methods, which are derived from the Hartree-Fock model by introducing several approximations, the most important of which is the neglect of all three- and four-center two-electron integrals; some or all of the remaining integrals are parameterized in reference to experimental data, in order to improve the quality of the computation.

However, the performance of conventional semiempirical SCF methods is still insufficient to deal with biological macromolecules, as a consequence of the use of matrix algebra techniques to solve the Roothaan-Hall equations. By using the near-linear scaling method MOZYME, the computational effort can be significantly reduced, allowing systems containing up to about 15000 atoms to be treated with conventional semiempirical SCF methods.

The first chapter of this dissertation provides an overview of the theoretical basis underlying semiempirical SCF methods. Semiempirical SCF methods are examined in detail in the second chapter, with particular attention to modified NDDO methods. The third and most important chapter of this dissertation provides an essential guide covering the practical aspects and the common issues involved in modeling protein structures with MOPAC, using PM7 with MOZYME.

1 Theoretical Background

The purpose of this chapter is to provide the reader with the essential theoretical background underlying semiempirical SCF methods. Since the aim of this dissertation is to discuss the application of semiempirical SCF methods to the study of proteins, the large majority of which are closed-shell systems, only the restricted closed-shell formalism will be considered in the following paragraphs.

1.1 The Born-Oppenheimer Approximation

Each molecule has its own unique shape, from the simplest H_2 to the most complex biomolecules. Even though this may seem rather obvious at first glance, it is one of the deepest consequences of the Born-Oppenheimer approximation [1]. The essence of the Born-Oppenheimer approximation is the assumption that the nuclei in a molecule are heavy enough that they can be described quite accurately by classical mechanics. If they showed a significant quantum behavior, the concept of molecular structure would not have any meaning because of Heisenberg's uncertainty principle and tunneling effects.

In order to understand what are the consequences of the Born-Oppenheimer approximation from a more mathematical point of view, consider the molecular non-relativistic Schrödinger equation

$$\hat{\mathcal{H}}_{\mathrm{mol}}\Psi_{\mathrm{mol}}\left(\mathbf{R},\mathbf{r}\right)=\mathcal{E}_{\mathrm{mol}}\Psi_{\mathrm{mol}}\left(\mathbf{R},\mathbf{r}\right) \tag{1.1}$$

where $\hat{\mathcal{H}}_{\mathrm{mol}}$ represents the molecular Hamiltonian operator, as given in Eq. (1.2).

$$\hat{\mathcal{H}}_{\mathrm{mol}}=\hat{\mathcal{T}}_{\mathrm{N}}+\hat{\mathcal{T}}_{\mathrm{e}}+\hat{\mathcal{V}}_{\mathrm{ee}}+\hat{\mathcal{V}}_{\mathrm{NN}}+\hat{\mathcal{V}}_{\mathrm{eN}}=\hat{\mathcal{T}}_{\mathrm{N}}+\hat{\mathcal{H}}_{\mathrm{el}} \tag{1.2}$$

If there exist two complete sets of wave functions $\{\Psi_{\mathrm{N},j}\left(\mathbf{R}\right)\}$ and $\{\Psi_{\mathrm{el},i}\left(\mathbf{r}\right)\}$, the molecular wave function can be expressed, without any approximation, as

$$\Psi_{\mathrm{mol}}\left(\mathbf{R},\mathbf{r}\right)=\sum_{j}^{\infty}\sum_{i}^{\infty}\Psi_{\mathrm{N},j}\left(\mathbf{R}\right)\Psi_{\mathrm{el},i}\left(\mathbf{r}\right) \tag{1.3}$$

Because the nuclei can be assumed to behave, to a good approximation, as classical particles, it is meaningful to solve the molecular Schrödinger equation for different nuclear configurations. In other words, the dependence on the nuclear coordinates can be assumed to be parametric. Since the electronic Hamiltonian operator $\hat{\mathcal{H}}_{\mathrm{el}}$ depends on the nuclear coordinates via the term $\hat{\mathcal{V}}_{\mathrm{eN}}$, the electronic wave functions $\Psi_{\mathrm{el},i}$ depend on the nuclear configuration. For this reason, Eq. (1.3) can be rewritten as

$$\Psi_{\mathrm{mol}}\left(\mathrm{R},\mathbf{r}\right)=\sum_{j}^{\infty}\sum_{i}^{\infty}\Psi_{\mathrm{N},j}\left(\mathrm{R}\right)\Psi_{\mathrm{el},i}\left(\mathrm{R},\mathbf{r}\right) \tag{1.4}$$

Substituting Eq. (1.4) into Eq. (1.1) gives

$$\begin{gathered}\sum_{j}^{\infty}\sum_{i}^{\infty}\hat{\mathcal{T}}_{\mathrm{N}}\left[\Psi_{\mathrm{N},j}\left(\mathrm{R}\right)\Psi_{\mathrm{el},i}\left(\mathrm{R},\mathbf{r}\right)\right]+\sum_{j}^{\infty}\sum_{i}^{\infty}\Psi_{\mathrm{N},j}\left(\mathrm{R}\right)\hat{\mathcal{H}}_{\mathrm{el}}\Psi_{\mathrm{el},i}\left(\mathrm{R},\mathbf{r}\right)= \\ \mathcal{E}_{\mathrm{tot}}\sum_{j}^{\infty}\sum_{i}^{\infty}\Psi_{\mathrm{N},j}\left(\mathrm{R}\right)\Psi_{\mathrm{el},i}\left(\mathrm{R},\mathbf{r}\right)\end{gathered} \tag{1.5}$$

A full evaluation of Eq. (1.5) leads to the equations [2]

$$\begin{gathered}\left[\hat{\mathcal{T}}_{\mathrm{N}}+\mathcal{E}_{\mathrm{el},k}\left(\mathrm{R}\right)\right]\Psi_{\mathrm{N},k}\left(\mathrm{R}\right)+\sum_{\mathrm{A}}^{n_{\mathrm{nuclei}}}\sum_{j}^{\infty}\left[C_{1\mathrm{A},kj}\nabla_{\mathrm{N_A}}\Psi_{\mathrm{N},j}\left(\mathrm{R}\right)+C_{2\mathrm{A},kj}\Psi_{\mathrm{N},j}\left(\mathrm{R}\right)\right]= \\ \mathcal{E}_{\mathrm{mol},k}\Psi_{\mathrm{N},k}\left(\mathrm{R}\right)\end{gathered} \tag{1.6}$$

$$\hat{\mathcal{H}}_{\mathrm{el}}\Psi_{\mathrm{el},k}\left(\mathrm{R},\mathbf{r}\right)=\mathcal{E}_{\mathrm{el},k}\left(\mathrm{R}\right)\Psi_{\mathrm{el},k}\left(\mathrm{R},\mathbf{r}\right) \tag{1.7}$$

where the coupling elements $C_{1\mathrm{A},kj}$ and $C_{2\mathrm{A},kj}$ are defined as

$$\begin{aligned}C_{1\mathrm{A},kj}&=-\frac{\hbar^{2}}{M_{\mathrm{N_A}}}\left\langle\Psi_{\mathrm{el},k}\left(\mathrm{R},\mathbf{r}\right)\left|\nabla_{\mathrm{N_A}}\right|\Psi_{\mathrm{el},j}\left(\mathrm{R},\mathbf{r}\right)\right\rangle \\ C_{2\mathrm{A},kj}&=-\frac{\hbar^{2}}{2M_{\mathrm{N_A}}}\left\langle\Psi_{\mathrm{el},k}\left(\mathrm{R},\mathbf{r}\right)\left|\nabla_{\mathrm{N_A}}^{2}\right|\Psi_{\mathrm{el},j}\left(\mathrm{R},\mathbf{r}\right)\right\rangle\end{aligned} \tag{1.8}$$

It is important to notice that the electronic energy $\mathcal{E}_{\mathrm{el},k}\left(\mathrm{R}\right)$ in Eq. (1.7) define a potential energy surface (PES), i.e. a hypersurface that maps the electronic energy as a function of the nuclear coordinates. As long as the PESs are well

separated, the coupling elements in Eq. (1.6) can be neglected [3], leading to

$$\left[\hat{\mathcal{T}}_{\mathrm{N}} + \mathcal{E}_{\mathrm{el},k}\left(\mathrm{R}\right)\right] \Psi_{\mathrm{N},k}\left(\mathrm{R}\right) = \mathcal{E}_{\mathrm{mol}} \Psi_{\mathrm{N},k}\left(\mathrm{R}\right) \tag{1.9}$$

In conclusion, as shown by Eq. (1.9), within the Born-Oppenheimer approximation the nuclei move on a PES that is a solution to the electronic Schrödinger equation. This means that molecular properties can be determined by solving the electronic Schrödinger equation for different nuclear configurations. From the practical point of view, as this process is very demanding, even for relatively small systems, it is generally restricted to chemically meaningful regions of the PESs, such as regions near local minima, which correspond to stable molecules.

1.2 Symmetry of the Electronic Wave Function

Within the non-relativistic treatment of multi-electronic systems, the electron spin must be introduced as an ad hoc quantum observable. In particular, for each electron, there exist two spin wave functions, namely $\alpha\left(\sigma\right)$ and $\beta\left(\sigma\right)$, which satisfy the following equations [4]

$$\hat{s}^2\alpha\left(\sigma\right) = \frac{3}{4}\hbar^2\alpha\left(\sigma\right) \quad \hat{s}^2\beta\left(\sigma\right) = \frac{3}{4}\hbar^2\beta\left(\sigma\right) \tag{1.10}$$

$$\hat{s}_z\alpha\left(\sigma\right) = \frac{1}{2}\hbar\alpha\left(\sigma\right) \quad \hat{s}_z\beta\left(\sigma\right) = \frac{1}{2}\hbar\beta\left(\sigma\right) \tag{1.11}$$

$$\begin{aligned} \langle\alpha|\alpha\rangle = \langle\beta|\beta\rangle = 1 \\ \langle\alpha|\beta\rangle = \langle\beta|\alpha\rangle = 0 \end{aligned} \tag{1.12}$$

where $\hat{s}$ and $\hat{s}_z$ denote the spin angular momentum operator and its component along the z-axis, respectively, while σ represents the spin coordinate. As a consequence, each electron is not only described by the three spatial coordinates $\mathbf{r}$ but also by one spin coordinate σ.

With this assumption in mind, the electronic wave function describing an N-electron system must be written as

$$\Psi_{\mathrm{el}}\left(\boldsymbol{\xi}\right) = \Psi\left(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_N\right) \tag{1.13}$$

where $\boldsymbol{\xi} = \{\mathbf{r}, \sigma\}$. If there exist, for each electron, a complete set of wave functions $\{\phi_i\left(\boldsymbol{\xi}_k\right)\}$, the electronic wave function can be expanded as

$$\Psi_{\mathrm{el}}\left(\boldsymbol{\xi}\right) = \sum_{ij\ldots z}^{\infty} c_{ij\ldots z}\left[\phi_i\left(\boldsymbol{\xi}_1\right)\phi_j\left(\boldsymbol{\xi}_2\right)\ldots\phi_z\left(\boldsymbol{\xi}_N\right)\right] \tag{1.14}$$

The one-electron wave functions $\phi_i\left(\boldsymbol{\xi}_k\right)$ in the above equation are generally referred to as one-electron spin-orbitals, or, more commonly, as molecular orbitals (MOs).

Although the wave function given in Eq. (1.14) does not involve any approximation, it is not a suitable electronic wave function because it does not have the correct symmetry. In fact, the wave function describing an ensemble of identical half-integer spin particles, like the electrons in a molecule, must be antisymmetric with respect to an interchange of any two particles coordinates [4, 5]. The antisymmetry requirement can be achieved by applying the antisymmetrizing operator $\hat{\mathcal{A}}$ to the wave function given in Eq. (1.14).

$$\Psi_{\mathrm{el}}\left(\boldsymbol{\xi}\right) = \hat{\mathcal{A}}\left\{\sum_{ij\ldots z}^{\infty} c_{ij\ldots z}\left[\phi_i\left(\boldsymbol{\xi}_1\right)\phi_j\left(\boldsymbol{\xi}_2\right)\ldots\phi_z\left(\boldsymbol{\xi}_N\right)\right]\right\} \tag{1.15}$$

The antisymmetrizing operator $\hat{\mathcal{A}}$ is defined as

$$\hat{\mathcal{A}} = \frac{1}{\sqrt{N!}}\sum_{k}\left(-1\right)^k\hat{\mathcal{P}}_k \tag{1.16}$$

where the operator $\hat{\mathcal{P}}_k$ denotes the permutation over k-particle coordinates and the summation is carried out over all possible permutations.

A full evaluation of Eq. (1.15) shows that, as a result of the antisymmetry principle, the electronic wave function describing an N-electron system can be written exactly as a linear combination of N-electron determinants $\{\Phi_{\mathrm{el},i}\left(\boldsymbol{\xi}\right)\}$, also known as Slater determinants [4].

$$\Phi_{\mathrm{el}}\left(\boldsymbol{\xi}\right) = \sum_{i}^{\infty} a_i\Phi_{\mathrm{el},i}\left(\boldsymbol{\xi}\right) \tag{1.17}$$

In other words, the determinants $\{\Phi_{\mathrm{el},i}\left(\boldsymbol{\xi}\right)\}$ form a complete set. Each Slater determinant represents a specific configuration of MOs, i.e. a specific electronic configuration. Because the electronic wave function given in Eq. (1.17) is modeled as a sum of contributions from an infinite number of electronic configurations, it is generally referred to as full Configuration Interaction (full CI) wave function.

1.3 Energy of a Single Slater Determinant

The non-relativistic electronic energy can be evaluated exactly as

$$\mathcal{E}_{\mathrm{el}} = \left\langle \Phi_{\mathrm{el}}\left(\boldsymbol{\xi}\right) \middle| \hat{\mathcal{H}}_{\mathrm{el}} \middle| \Phi_{\mathrm{el}}\left(\boldsymbol{\xi}\right) \right\rangle \tag{1.18}$$

where $\Phi_{\mathrm{el}}\left(\boldsymbol{\xi}\right)$ represents the full CI wave function, as given in Eq. (1.17). However, from the computational point of view, infinite basis sets are impossible to handle. Furthermore, even for relatively small systems and minimal basis sets the number of determinants that need to be evaluated in Eq. (1.18) is huge.

The simplest approach in order to produce a computationally viable model is to describe the electronic wave function as a single Slater determinant $\Phi_{\mathrm{el},0}\left(\boldsymbol{\xi}\right)$, namely the Slater determinant for which the corresponding energy eigenvalue is the lowest [4, 5]. Within the single Slater determinant approximation, the electronic energy can be therefore evaluated as

$$\mathcal{E}_{\mathrm{el}} = \left\langle \Phi_{\mathrm{el},0}\left(\boldsymbol{\xi}\right) \middle| \hat{\mathcal{H}}_{\mathrm{el}} \middle| \Phi_{\mathrm{el},0}\left(\boldsymbol{\xi}\right) \right\rangle \tag{1.19}$$

For the purpose of evaluating Eq. (1.19), it is convenient to collect the operators in the electronic Hamiltonian operator according to the number of electron coordinates upon which they depend.

$$\begin{aligned}
\hat{h}_i &= -\frac{\hbar^2}{2m_{\mathrm{e}}}\nabla_i^2 - \frac{\mathrm{e}^2}{4\pi\varepsilon_0}\sum_{\mathrm{A}}^{n_{\mathrm{nuclei}}}\frac{Z_{\mathrm{A}}}{|\mathbf{R}_{\mathrm{A}} - \mathbf{r}_i|} \\
\hat{g}_{ij} &= \frac{\mathrm{e}^2}{4\pi\varepsilon_0}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \\
\hat{\mathcal{V}}_{\mathrm{NN}} &= \frac{\mathrm{e}^2}{4\pi\varepsilon_0}\sum_{\mathrm{A}}^{n_{\mathrm{nuclei}}}\sum_{\mathrm{B}>\mathrm{A}}^{n_{\mathrm{nuclei}}}\frac{Z_{\mathrm{A}}Z_{\mathrm{B}}}{|\mathbf{R}_{\mathrm{A}} - \mathbf{R}_{\mathrm{B}}|} \\
\hat{\mathcal{H}}_{\mathrm{el}} &= \sum_{i}^{n_{\mathrm{e}}}\left(\hat{h}_i + \sum_{j>i}^{n_{\mathrm{e}}}\hat{g}_{ij}\right) + \hat{\mathcal{V}}_{\mathrm{NN}}
\end{aligned} \tag{1.20}$$

The one-electron operator $\hat{h}_i$ describes the motion of the electron i subject to the electric field of the nuclei, while the two-electron operator $\hat{g}_{ij}$ describes the electron-electron repulsion between the electrons i and j. The operator $\hat{\mathcal{V}}_{\mathrm{NN}}$ represents the total nuclear repulsion energy and, since it does not depend upon the electron coordinates, it remains constant for a fixed nuclear configuration. A full evaluation of Eq. (1.19) shows that the energy of a single Slater determinant is given by [5]

$$\mathcal{E}_{\mathrm{el}} = \sum_{i}^{n_{\mathrm{e}}} h_i + \frac{1}{2}\sum_{i}^{n_{\mathrm{e}}}\sum_{j}^{n_{\mathrm{e}}} \left(J_{ij} - K_{ij}\right) + V_{\mathrm{NN}} \tag{1.21}$$

or in terms of operators by

$$\begin{gathered}\mathcal{E}_{\mathrm{el}} = \sum_{i}^{n_{\mathrm{e}}} \left\langle \psi_i\left(\mathbf{r}_1\right) \left| \hat{h}_i \right| \psi_i\left(\mathbf{r}_1\right) \right\rangle + \\ \frac{1}{2}\sum_{i}^{n_{\mathrm{e}}}\sum_{j}^{n_{\mathrm{e}}} \left(\left\langle \psi_j\left(\mathbf{r}_2\right) \left| \hat{\mathcal{J}}_i \right| \psi_j\left(\mathbf{r}_2\right) \right\rangle - \left\langle \psi_j\left(\mathbf{r}_2\right) \left| \hat{\mathcal{K}}_i \right| \psi_j\left(\mathbf{r}_2\right) \right\rangle \right) + V_{\mathrm{NN}}\end{gathered} \tag{1.22}$$

where the Coulomb operator $\hat{\mathcal{J}}_i$ and the Exchange operator $\hat{\mathcal{K}}_i$ are defined as

$$\begin{aligned}\hat{\mathcal{J}}_i \left|\psi_j\left(\mathbf{r}_2\right)\right\rangle &= \left\langle \psi_i\left(\mathbf{r}_1\right) \left| \hat{g}_{12} \right| \psi_i\left(\mathbf{r}_1\right) \right\rangle \left|\psi_j\left(\mathbf{r}_2\right)\right\rangle \\ \hat{\mathcal{K}}_i \left|\psi_j\left(\mathbf{r}_2\right)\right\rangle &= \left\langle \psi_i\left(\mathbf{r}_1\right) \left| \hat{g}_{12} \right| \psi_j\left(\mathbf{r}_1\right) \right\rangle \left|\psi_i\left(\mathbf{r}_2\right)\right\rangle\end{aligned} \tag{1.23}$$

It is important to notice that, as a consequence of the non-relativistic approximation, the spin wave functions in Eq. (1.22) and Eq.s (1.23) have been integrated out. In fact, since the electronic Hamiltonian does not depend on the spin coordinates, each MO can be written as the product between a spatial orbital $\psi\left(\mathbf{r}\right)$ and a spin wave function, $\alpha\left(\sigma\right)$ or $\beta\left(\sigma\right)$.

As shown by Eq. (1.21), the electronic energy is given as a sum of four contributions:

1. The one-electron integral h_i, which represents the energy of the electron i subject to the field of the nuclei

2. The two-electron Coulomb integral J_{ij}, which represents the repulsion energy between two charge distributions described by the probability density functions $\psi_i^{\star}\left(\mathbf{r}_1\right)\psi_i\left(\mathbf{r}_1\right)$ and $\psi_j^{\star}\left(\mathbf{r}_2\right)\psi_j\left(\mathbf{r}_2\right)$

3. The two-electron Exchange integral K_{ij}, which can be thought as a negative correction term to the total electronic energy. In fact, as a consequence of the antisymmetry principle, same spin electrons cannot occupy the same MO and thus they tend to avoid each other more frequently than expected from taking only the Coulomb repulsion into account

4. The term V_{NN}, which represents the total nuclear repulsion energy

From this careful analysis, it is evident that, within the single Slater determinant approximation, the electronic energy cannot be evaluated exactly. In fact, the

instantaneous electron-electron interaction is accounted only in an average way, via the operators $\hat{\mathcal{J}}_j$ and $\hat{\mathcal{K}}_j$. In other words, the electron correlation is partially neglected. Nevertheless, the single Slater determinant approximation takes some electron correlation into account. In fact, same spin electrons cannot occupy the same spin-orbital, otherwise the Slater determinant would vanish.

1.4 Restricted and Unrestricted Determinants

Within the non-relativistic approximation, the electronic Hamiltonian does not depend on the spin coordinates. For this reason, the operators $\hat{\mathcal{S}}^2$ and $\hat{\mathcal{S}}_z$ commute with the electronic Hamiltonian [4]

$$\begin{aligned} \left[\hat{\mathcal{H}}_{\text{el}}, \hat{\mathcal{S}}^2\right] &= 0 \\ \left[\hat{\mathcal{H}}_{\text{el}}, \hat{\mathcal{S}}_z\right] &= 0 \end{aligned} \tag{1.24}$$

where $\hat{\mathcal{S}}$ and $\hat{\mathcal{S}}_z$ represent the total spin angular momentum and its component along the z-axis, respectively. As a consequence, the exact eigenfunctions of the electronic Hamiltonian, namely the full CI wave functions, are also eigenfunctions of the operators $\hat{\mathcal{S}}^2$ and $\hat{\mathcal{S}}_z$ [4]

$$\begin{aligned} \hat{\mathcal{S}}^2 \Phi_{\text{el}}\left(\boldsymbol{\xi}\right) &= S\left(S+1\right)\Phi_{\text{el}}\left(\boldsymbol{\xi}\right) \\ \hat{\mathcal{S}}_z \Phi_{\text{el}}\left(\boldsymbol{\xi}\right) &= M_S \Phi_{\text{el}}\left(\boldsymbol{\xi}\right) \end{aligned} \tag{1.25}$$

where S and M_S are the total spin quantum numbers.

For the purpose of determining if Eq.s (1.25) hold within the single Slater determinant approximation, determinants can be classified into two categories:

1. Restricted determinants, which are formed from MOs whose spatial wave function is the same for different spin wave functions

$$\left\{\phi_i\left(\boldsymbol{\xi}_k\right)\right\} = \left\{\psi_i\left(\mathbf{r}_k\right)\alpha\left(\sigma_k\right), \psi_i\left(\mathbf{r}_k\right)\beta\left(\sigma_k\right)\right\} \tag{1.26}$$

2. Unrestricted determinants, which are formed from MOs whose spatial wave function is different for different spin wave functions

$$\left\{\phi_i\left(\boldsymbol{\xi}_k\right)\right\} = \left\{\psi_i^{\alpha}\left(\mathbf{r}_k\right)\alpha\left(\sigma_k\right), \psi_i^{\beta}\left(\mathbf{r}_k\right)\beta\left(\sigma_k\right)\right\} \tag{1.27}$$

Restricted determinants can be further classified into:

1. Closed-shell determinants, in which each spatial orbital is doubly occupied

2. Open-shell determinants, in which at least one spatial orbital is singly occupied

While every single Slater determinant is an eigenfuction for the operator $\hat{\mathcal{S}}_z$, this is not necessarily true for the operator $\hat{\mathcal{S}}^2$. In fact, it can be shown that only restricted closed-shell determinants and restricted open-shell determinants in which unpaired electrons have parallel spins are eigenfunctions for the operator $\hat{\mathcal{S}}^2$ [4]. As a consequence, only closed-shell molecules can be described correctly in terms of spin within the single Slater determinant approximation. Open-shell molecules, such as radicals and certain excited molecules, can be described correctly in terms of spin only by using restricted multi-determinant wave functions. As said in the introduction, only the restricted closed-shell formalism will be considered in the following paragraphs.

1.5 Restricted Closed-Shell Hartree-Fock Equations

Within the restricted closed-shell formalism, the summations in Eq. (1.21) and Eq. (1.22) can be carried out over occupied spatial orbitals rather than over electron coordinates, giving

$$\begin{gathered}\mathcal{E}_{\mathrm{el}} = 2\sum_{i}^{n} h_i + \sum_{i}^{n}\sum_{j}^{n} \left(2J_{ij} - K_{ij}\right) + V_{\mathrm{NN}} \\ \mathcal{E}_{\mathrm{el}} = 2\sum_{i}^{n} \left\langle \psi_i \left| \hat{h}_i \right| \psi_i \right\rangle + \sum_{i}^{n}\sum_{j}^{n} \left(2 \left\langle \psi_j \left| \hat{\mathcal{J}}_i \right| \psi_j \right\rangle - \left\langle \psi_j \left| \hat{\mathcal{K}}_i \right| \psi_j \right\rangle \right) + V_{\mathrm{NN}}\end{gathered} \tag{1.28}$$

where n represents the number of occupied spatial orbitals.

Having determined the electronic energy within the single Slater determinant approximation, the Langrangian multipliers method can be used to determine the set of occupied spatial orbitals that make the electronic energy a minimum, or at least a stationary point, under the constraint that the orbitals remain an orthonormal set. This is equivalent to require that the functional

$$\mathcal{L}\left[\{\psi_k\}\right] = \mathcal{E}_{\mathrm{el}}\left[\{\psi_k\}\right] - \sum_{i}^{n}\sum_{j}^{n} \lambda_{ij} \left(\langle \psi_i | \psi_j \rangle - \delta_{ij} \right) \tag{1.29}$$

is stationary with respect to any orbital variation.

$$\begin{cases} \dfrac{\delta \mathcal{L}\left[\{\psi_k\}\right]}{\delta \psi_1} = 0 \\ \vdots \\ \dfrac{\delta \mathcal{L}\left[\{\psi_k\}\right]}{\delta \psi_n} = 0 \end{cases} \tag{1.30}$$

A full evaluation of Eq.s (1.30) leads to the equations [4, 5]

$$\begin{cases} \hat{\mathcal{F}}\psi_1 = \sum_j^n \lambda_{1j}\psi_j \\ \vdots \\ \hat{\mathcal{F}}\psi_n = \sum_j^n \lambda_{nj}\psi_j \end{cases} \tag{1.31}$$

where the Fock operator $\hat{\mathcal{F}}$ is defined as

$$\hat{\mathcal{F}} = \hat{h} + \sum_j^n \left(2\hat{\mathcal{J}}_j - \hat{\mathcal{K}}_j\right) \tag{1.32}$$

The Fock operator is a one-electron operator, identical for every electron in the system, which describes the kinetic energy of an electron and its instantaneous attraction to the nuclei, via the operator $\hat{h}$, as well as its average interactions with the other electrons, via the operators $\hat{\mathcal{J}}_j$ and $\hat{\mathcal{K}}_j$.

As a consequence of the Fock operator hermiticity, the matrix of Lagrange multipliers, whose elements are the matrix elements $\left\langle \psi_i \middle| \hat{\mathcal{F}} \middle| \psi_j \right\rangle$, can be diagonalized by a unitary transformation to give a set of pseudo-eigenvalue equations known as the restricted closed-shell Hartree-Fock (HF) equations [4, 5].

$$\begin{cases} \hat{\mathcal{F}}\psi'_1 = \varepsilon_1\psi'_1 \\ \vdots \\ \hat{\mathcal{F}}\psi'_n = \varepsilon_n\psi'_n \end{cases} \tag{1.33}$$

The occupied spatial orbitals $\{\psi'_k\}$ for which the matrix of Lagrange multipliers is diagonal are generally referred to as canonical molecular orbitals. It is important to notice that the Langrange multipliers have the physical interpretation of canonical MO energies.

$$\varepsilon_i = \left\langle \psi'_i \middle| \hat{\mathcal{F}} \middle| \psi'_i \right\rangle = \left\langle \psi'_i \middle| \hat{h}_i \middle| \psi'_i \right\rangle + \sum_j^n \left(2\left\langle \psi'_i \middle| \hat{\mathcal{J}}_j \middle| \psi'_i \right\rangle - \left\langle \psi'_i \middle| \hat{\mathcal{K}}_j \middle| \psi'_i \right\rangle\right) \tag{1.34}$$

The HF equations do not properly form a set of eigenvalue equations. In fact, since the Fock operator depends on all canonical MOs, via the operators $\hat{\mathcal{J}}_j$ and $\hat{\mathcal{K}}_j$, a specific canonical MO can only be determined if all the other canonical MOs are known. For this reason, an iterative procedure, generally referred to as self-consistent field (SCF) procedure, must be employed to solve the HF equations.

1.6 Restricted Closed-Shell Roothaan-Hall Equations

From the computational point of view, the HF equations do not provide a simple procedure to get the initial guesses for the canonical MOs, which are essential to start the SCF procedure. Moreover, they are non-linear integro-differential equations that can be solved only for small symmetric systems, such as atoms and diatomic molecules, by employing numerical methods [2, 5].

As pointed out independently by Roothaan and Hall in 1951 [6, 7], these problems can be solved by expanding the unknown canonical MOs in the HF equations in terms of a known basis set of spatial wave functions $\{\chi_\beta\}$, conventionally referred to as atomic orbitals (AO),

$$\psi_i = \sum_{\beta}^{m} c_{\beta i}\chi_\beta \tag{1.35}$$

where m is the number of basis functions. If the basis set $\{\chi_\beta\}$ is complete, the expansion does not involve any approximation. However, infinite basis sets are impossible to handle in actual calculations. It is important, therefore, to choose a set of AOs that are well behaved from a physical point of view in order to ensure an accurate representation of the canonical MOs with a finite number of basis functions.

By expanding each canonical MO in terms of AOs, the HF equations may be rewritten as

$$\begin{aligned} \sum_{\beta}^{m} c_{\beta 1}\hat{\mathcal{F}}\chi_\beta &= \varepsilon_1 \sum_{\beta}^{m} c_{\beta 1}\hat{\mathcal{F}}\chi_\beta \\ &\vdots \\ \sum_{\beta}^{m} c_{\beta n}\hat{\mathcal{F}}\chi_\beta &= \varepsilon_n \sum_{\beta}^{m} c_{\beta n}\hat{\mathcal{F}}\chi_\beta \end{aligned} \tag{1.36}$$

Multiplying from the left by each basis function and integrating over electron

coordinates leads to the restricted closed-shell Roothan-Hall (RH) equations [8]

$$\begin{array}{rl} \text{for the basis function } \chi_1: & \begin{cases} \sum_{\beta}^{m} F_{1\beta} c_{\beta 1} = \varepsilon_1 \sum_{\beta}^{m} S_{1\beta} c_{\beta 1} \\ \vdots \\ \sum_{\beta}^{m} F_{1\beta} c_{\beta n} = \varepsilon_n \sum_{\beta}^{m} S_{1\beta} c_{\beta n} \end{cases} \\ \vdots & \vdots \\ \text{for the basis function } \chi_m: & \begin{cases} \sum_{\beta}^{m} F_{m\beta} c_{\beta 1} = \varepsilon_1 \sum_{\beta}^{m} S_{m\beta} c_{\beta 1} \\ \vdots \\ \sum_{\beta}^{m} F_{m\beta} c_{\beta n} = \varepsilon_n \sum_{\beta}^{m} S_{m\beta} c_{\beta m} \end{cases} \end{array} \tag{1.37}$$

where the Fock matrix elements $F_{\alpha\beta}$ and the overlap matrix elements $S_{\alpha\beta}$ are the defined as

$$F_{\alpha\beta} = \left\langle \chi_\alpha \left| \hat{\mathcal{F}} \right| \chi_\beta \right\rangle = \left\langle \chi_\alpha \left| \hat{h} \right| \chi_\beta \right\rangle + \sum_{j}^{n} \left(2 \left\langle \chi_\alpha \left| \hat{\mathcal{J}}_j \right| \chi_\beta \right\rangle - \left\langle \chi_\alpha \left| \hat{\mathcal{K}}_j \right| \chi_\beta \right\rangle \right) \tag{1.38}$$

$$S_{\alpha\beta} = \langle \chi_\alpha | \chi_\beta \rangle \tag{1.39}$$

In order to evaluate the matrix elements $F_{\alpha\beta}$ explicitly, the two-electron integrals in Eq. (1.38) must be written in terms of AOs as [8]

$$F_{\alpha\beta} = h_{\alpha\beta} + \sum_{\gamma}^{m} \sum_{\delta}^{m} P_{\gamma\delta} \left[(\alpha\beta|\gamma\delta) - \frac{1}{2} (\alpha\delta|\gamma\beta) \right] \tag{1.40}$$

where $P_{\gamma\delta}$ is the density matrix element, as given in Eq. (1.41).

$$P_{\gamma\delta} = 2 \sum_{j}^{n} c_{\gamma j} c_{\delta j} \tag{1.41}$$

The two-electron integrals in Eq. (1.40) are written using the Mulliken notation, where the ordering of the AOs is given by the electron coordinates.

$$(\alpha\beta|\gamma\delta) = \iint \chi_\alpha^* (\mathbf{r}_1) \chi_\beta (\mathbf{r}_1) \hat{g}_{12} \chi_\gamma^* (\mathbf{r}_2) \chi_\delta (\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \tag{1.42}$$

The RH equations may be written in the more compact matrix form

$$\mathbf{F}(\mathbf{C})\,\mathbf{C} = \mathbf{S}\mathbf{C}\boldsymbol{\varepsilon} \tag{1.43}$$

where $\mathbf{F}(\mathbf{C})$ is the Fock matrix, $\mathbf{S}$ is the overlap matrix, $\mathbf{C}$ is coefficient matrix and $\boldsymbol{\varepsilon}$ is the diagonal matrix of canonical MO energies. The above equation shows that when the canonical MOs in the HF equations are expanded in terms of AOs, the HF equations are converted to a set of algebraic equations, the RH equations, which can be solved by standard matrix techniques to give the coefficients $c_{\beta i}$ and the canonical MO energies ε_i.

Since the Fock matrix elements depend on the coefficient matrix elements, as shown by Eq. (1.40), a SCF procedure must be employed to solve the RH equations. However, in order to diagonalize the Fock matrix to get the coefficients $c_{\beta i}$ and the canonical MO energies, the RH equations must be first transformed into the standard eigenvalue form by means of an orthogonalizing matrix $\mathbf{S}^{-1/2}$, which can be calculated from the overlap matrix $\mathbf{S}$ [8].

$$\begin{gathered}\left\{\mathbf{S}^{-1/2}\mathbf{F}(\mathbf{C})\,\mathbf{S}^{-1/2}\right\}\left\{\mathbf{S}^{1/2}\mathbf{C}\right\} = \left\{\mathbf{S}^{-1/2}\mathbf{S}\mathbf{S}^{-1/2}\right\}\left\{\mathbf{S}^{1/2}\mathbf{C}\right\}\boldsymbol{\varepsilon} \\ \mathbf{C}'^{-1}\mathbf{F}'(\mathbf{C})\,\mathbf{C}' = \boldsymbol{\varepsilon}\end{gathered} \tag{1.44}$$

The general RH SCF algorithm can be therefore summarized as follows:

1. Specify molecular geometry $\mathrm{R}^{(n)}$, charge and electronic state
2. Choose a suitable set of AOs
3. Program computes and stores all overlap, one-electron and two-electron integrals
4. Program computes and stores the orthogonalizing matrix $\mathbf{S}^{-1/2}$
5. Specify an initial guess for the density matrix $\mathbf{P}^{(n)}$
6. Program computes the Fock matrix $\mathbf{F}(\mathbf{C})$
7. Program computes the transformed Fock matrix as $\mathbf{F}'(\mathbf{C}) = \mathbf{S}^{-1/2}\mathbf{F}(\mathbf{C})\,\mathbf{S}^{-1/2}$
8. Program diagonalizes the transformed Fock matrix to obtain the transformed coefficient matrix $\mathbf{C}'$ and the canonical MO energy matrix $\boldsymbol{\varepsilon}$
9. Program computes the coefficient matrix $\mathbf{C}$ from the transformed coefficient matrix $\mathbf{C}'$ as $\mathbf{C} = \mathbf{S}^{-1/2}\mathbf{C}'$

10. Program computes the new density matrix $\mathbf{P}^{(n+1)}$ from the coefficient matrix $\mathbf{C}$

11. Program determines if the new density matrix $\mathbf{P}^{(n+1)}$ is sufficiently similar to the old density matrix $\mathbf{P}^{(n)}$. If the two density matrices are not similar enough, it restarts the SCF algorithm from Step 5 using the new density matrix $\mathbf{P}^{(n+1)}$ as initial guess. Otherwise, if the two density matrices are sufficiently close to each other the SCF algorithm has converged and the program stops

After each SCF cycle has converged, a geometry optimization procedure is performed, i.e. a specific algorithm evaluates if the current structure is a stationary point of the electronic PES. If the current structure is a stationary point of the PES, the calculation stops and the resulting matrices can be used to evaluate the quantities of interest. Otherwise, a new molecular geometry $\mathrm{R}^{(n+1)}$ is chosen according to the optimization algorithm and the SCF algorithm is restarted.

References

1. Born M, Oppenheimer JR. Zur Quantentheorie der Molekeln. Ann Physik. 1927;389(20):457-484. doi:10.1002/andp.19273892002.

2. Mayer I. Simple Theorems, Proofs, and Derivations in Quantum Chemistry. New York: Kluwer Academic/Plenum Publishers; 2003.

3. Butler LJ. Chemical reaction dynamics beyond the Born-Oppenheimer approximation. Annu Rev Phys Chem. 1998;49:125-171. doi:10.1146/annurev.physchem.49.1.125.

4. Szabo A, Ostlund NS. Modern Quantum Chemistry. Mineola: Dover Publications, INC.; 1996.

5. Jensen F. Introduction to Computational Chemistry. 2nd ed. Chisester: John Wiley & Sons Ltd; 2007.

6. Roothaan CCJ. New Developments in Molecular Orbital Theory. Rev Mod Phys. 1951;23(2):69. doi:10.1103/RevModPhys.23.69.

7. Hall GG. The Molecular Orbital Theory of Chemical Valency. VIII. A Method of Calculating Ionization Potentials. Proc R Soc A. 1951;205(1083):541. doi:10.1098/rspa.1951.0048.

8. Lewars E. Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics. Boston: Kluwer Academic Plubishers; 2003.

2 Semiempirical SCF Methods

The practical application of the RH SCF procedure to the calculation of molecular properties has resulted into two main methods: ab initio SCF methods and semiempirical SCF methods.

Ab initio SCF methods generate the solutions to the RH equations by computing all the required integrals without any reference to experimental data. Because of the high number of two-electron integrals that need to be evaluated explicitly for constructing the Fock matrix, ab initio SCF methods require a significant computational effort, even for relatively small systems.

A great computational reduction can be achieved with semiempirical SCF methods, which are derived from the RH SCF model by introducing several approximations, the most important of which is the neglect of all three- and four-center two-electron integrals. Some or all of the remaining integrals are parameterized in reference to experimental data in order to improve the quality of the computation.

Despite their inherent approximations, semiempirical SCF methods are fast and accurate enough for routine applications, even for quite large molecular systems, such as biological macromolecules and crystal systems, in their recent implementations.

The first part of this chapter will concentrate on the general aspects of semiempirical SCF methods. In the second part, the most popular semiempirical SCF methods to date will be considered more in detail, with particular attention to the most recently developed modified NDDO methods, namely PM6 and PM7.

2.1 General Aspects of Semiempirical SCF Methods

2.1.1 Valence Electron Approximation

Only valence electrons are involved in ordinary chemical reactions. For this reason, in semiempirical SCF methods the first step in computational reduction

is to consider only the valence electrons explicitly.

Within the valence electron approximation, a molecular system is modeled as a cloud of valence electrons moving in the field of atomic cores, i.e. the nuclei plus their core electrons. The electronic Hamiltonian operator $\hat{\mathcal{H}}_{\mathrm{el}}$ can be therefore written as the sum between the valence electron Hamiltonian operator $\hat{\mathcal{H}}_{\mathrm{val}}$ and the total core-core repulsion operator $\hat{\mathcal{V}}_{\mathrm{CC}}$.

$$\hat{\mathcal{H}}_{\mathrm{el}} = \hat{\mathcal{H}}_{\mathrm{val}} + \hat{\mathcal{V}}_{\mathrm{CC}} = \sum_{i}^{n_{\mathrm{val\,e}}} \left(\hat{h}_{\mathrm{val},i} + \sum_{j>i}^{n_{\mathrm{val\,e}}} \hat{g}_{ij} \right) + \sum_{\mathrm{A}}^{n_{\mathrm{nuclei}}} \sum_{\mathrm{B>A}}^{n_{\mathrm{nuclei}}} \hat{\mathcal{V}}_{\mathrm{AB}} \tag{2.1}$$

The one-electron operator $\hat{h}_{\mathrm{val},i}$, which describes the motion of the valence electron i subject to the field of all atomic cores, is defined as

$$\hat{h}_{\mathrm{val},i} = -\frac{\hbar^2}{2m_{\mathrm{e}}} \nabla_i^2 - \sum_{\mathrm{A}}^{n_{\mathrm{nuclei}}} \hat{\mathcal{V}}_{i\mathrm{A}} \tag{2.2}$$

where the operator $\hat{\mathcal{V}}_{i\mathrm{A}}$ describes the interaction between the valence electron i and the core A.

The Fock matrix elements and the electronic energy can be evaluated within the valence electron approximation as

$$F_{\mathrm{val},\alpha\beta} = h_{\mathrm{val},\alpha\beta} + \sum_{\gamma}^{m} \sum_{\delta}^{m} P_{\gamma\delta} \left[(\alpha\beta|\gamma\delta) - \frac{1}{2} (\alpha\delta|\gamma\beta) \right] \tag{2.3}$$

$$\mathcal{E}_{\mathrm{el}} = \frac{1}{2} \sum_{\alpha}^{m} \sum_{\beta}^{m} P_{\alpha\beta} \left(F_{\mathrm{val},\alpha\beta} + h_{\mathrm{val},\alpha\beta} \right) + \sum_{\mathrm{A}}^{n_{\mathrm{nuclei}}} \sum_{\mathrm{B>A}}^{n_{\mathrm{nuclei}}} V_{\mathrm{AB}} \tag{2.4}$$

As described later in this chapter, the core-core interaction terms V_{AB}, which represent the interaction energies between atomic cores, are modeled as parameterized functions, whose specific functional form depends on the particular semiempirial SCF method considered.

2.1.2 Minimal Basis Set

In ab initio SCF methods the computational effort scales as the fourth power of the number basis functions because of the number of two-electron integrals that need to be evaluated in order to construct the Fock matrix. A great computational reduction can be achieved by employing an orthonormal minimal basis set, i.e. a set of AOs that is composed of the minimum number of basis functions required to describe the valence electrons of each atom in a molecule.

The large majority of semiempirical SCF methods to date use Slater-type orbitals (STOs) as basis functions. The functional form of a STO centered on an atom is given by

$$\chi_{\zeta,n,l,m}(r,\,\theta,\,\varphi) = N\,Y_{l,\,m}(\theta,\,\varphi)\,r^{n-1}\exp(-\zeta r) \tag{2.5}$$

where N is a normalization constant, $Y_{l,\,m}(\theta,\,\varphi)$ are the spherical harmonics, n is the principal quantum number and ζ is a parameter that can be regarded as the effective nuclear charge. STOs are preferred to Gaussian-type orbitals (GTOs) in semiempirical SCF methods because they show the correct radial behavior, ensuring an accurate representation of the MOs with a finite number of basis functions. The major drawback of the use of the STOs is that three- and four-center two-electron integrals cannot be evaluated analytically. However, this does not constitute a problem in semiempirical SCF methods because all three- and four-center two-electron integrals are neglected as a consequence of the Zero Differential Overlap approximation.

2.1.3 Zero Differential Overlap Approximation

The most important approximation of semiempirical SCF methods in terms of computational cost reduction is the Zero Differential Overlap approximation (ZDO), which neglects all products of AOs that depend on the same electron coordinate when the AOs are located on different atoms.

$$dS_{\alpha_{\mathrm{A}}\beta_{\mathrm{B}}} = \chi_{\alpha}^{\mathrm{A}}(\mathbf{r}_1)\,\chi_{\beta}^{\mathrm{B}}(\mathbf{r}_1)\,d\mathbf{r}_1 = \delta_{\mathrm{AB}}\chi_{\alpha}^{\mathrm{A}}(\mathbf{r}_1)\,\chi_{\beta}^{\mathrm{A}}(\mathbf{r}_1)\,d\mathbf{r}_1 \tag{2.6}$$

As a consequence of the ZDO approximation:

1. All three- and four-center two-electron integrals, which are the most numerous two-electron integrals, are neglected

2. The overlap matrix $\mathbf{S}$ is reduced to the identity matrix, as the AOs employed in semiempirical SCF methods are assumed to be orthonormal.

$$S_{\alpha_{\mathrm{A}}\beta_{\mathrm{B}}} = \delta_{\mathrm{AB}}\delta_{\alpha\beta} \tag{2.7}$$

   The RH equations are therefore transformed into the standard eigenvalue form without requiring an orthogonalization procedure

3. One-electron integrals involving three centers are neglected

It is important to notice that, unlike the valence electron approximation, which is a fairly reasonable chemical approximation, the ZDO approximation is only justified by explicit calculations, which show that three- and four-center two-electron integrals are indeed very small compared to the other integrals.

As a result of the ZDO approximation, semiempirical SCF methods scale as second power of the number of basis functions. However, as a consequence of the use of matrix algebra techniques to solve the RH equations, the observed scaling, in the limit of large molecular systems, is M^3_{basis}. Standard semiempirical SCF methods are therefore limited to molecular systems containing up to about a thousand atoms; larger systems can only be treated by employing linear-scaling methods, such as the MOZYME method, which will be considered in the next chapter.

2.1.4 NDDO, CNDO and INDO Semiempirical Models

According to how the remaining one- and two-center two-electron integrals are treated, three semiempirical models can be defined:

1. Neglect of Diatomic Differential Overlap (NDDO) model
2. Complete Neglect of Differential Overlap (CNDO) model
3. Intermediate Neglect of Differential Overlap (INDO) model

2.1.4.1 NDDO Model

In the NDDO model there are no further approximations than those mentioned above, meaning that all one- and two-center two-electron integrals are non-zero.

$$(\alpha_{\text{A}}\beta_{\text{B}}|\gamma_{\text{C}}\zeta_{\text{D}}) = \delta_{\text{AB}}\delta_{\text{CD}}\,(\alpha_{\text{A}}\beta_{\text{A}}|\gamma_{\text{C}}\zeta_{\text{C}}) \tag{2.8}$$

As for the one-electron integrals $h_{\text{val},\alpha\beta}$, it is convenient to distinguish between cases where the AOs are located on the same or on different atoms.

If the AOs belong to the same atom, the integrals $h_{\text{val},\alpha\beta}$ are evaluated as

$$h_{\text{val},\alpha_{\text{A}}\beta_{\text{A}}} = \delta_{\alpha\beta}\left\langle \alpha_{\text{A}} \left| -\frac{1}{2}\nabla^2 - \hat{\mathcal{V}}_{\text{A}} \right| \alpha_{\text{A}} \right\rangle - \sum_{\text{B}\neq\text{A}}^{n_{\text{nuclei}}} \left\langle \alpha_{\text{A}} \left| \hat{\mathcal{V}}_{\text{B}} \right| \beta_{\text{A}} \right\rangle \tag{2.9}$$

If the AOs belong to different atoms, the integrals $h_{\text{val},\alpha\beta}$ are evaluated as

$$h_{\text{val},\alpha_{\text{A}}\beta_{\text{B}}} = \left\langle \alpha_{\text{A}} \left| -\frac{1}{2}\nabla^2 - \hat{\mathcal{V}}_{\text{A}} - \hat{\mathcal{V}}_{\text{B}} \right| \beta_{\text{B}} \right\rangle \tag{2.10}$$

2.1.4.2 CNDO and INDO Models

The CNDO and the INDO models involve more drastic approximations. In addition to the NDDO model approximations, the CNDO model neglects all products of different AOs that depend on the same electron coordinate, meaning that only one- and two-center Coulomb integrals are non-zero.

$$dS_{\alpha_\mathrm{A}\beta_\mathrm{B}} = \chi_\alpha^\mathrm{A}\left(\mathbf{r}_1\right)\chi_\beta^\mathrm{B}\left(\mathbf{r}_1\right)d\mathbf{r}_1 = \delta_\mathrm{AB}\delta_{\alpha\beta}\chi_\alpha^\mathrm{A}\left(\mathbf{r}_1\right)\chi_\alpha^\mathrm{A}\left(\mathbf{r}_1\right)d\mathbf{r}_1 \tag{2.11}$$

In addition, in order to maintain rotational and hybridizational invariance, the two-electron Coulomb integrals and the integrals $\left\langle \alpha_\mathrm{A}\left|\hat{\mathcal{V}}_\mathrm{B}\right|\alpha_\mathrm{A}\right\rangle$ must be assumed to depend only on the atoms to which the AOs belong and not on the actual type of AO [1].

$$\begin{aligned}
(\alpha_\mathrm{A}\alpha_\mathrm{A}|\beta_\mathrm{A}\beta_\mathrm{A}) &= \gamma_\mathrm{AA} \\
(\alpha_\mathrm{A}\alpha_\mathrm{A}|\beta_\mathrm{B}\beta_\mathrm{B}) &= \gamma_\mathrm{AB} \\
\left\langle \alpha_\mathrm{A}\left|\hat{\mathcal{V}}_\mathrm{B}\right|\alpha_\mathrm{A}\right\rangle &= V_\mathrm{AB}
\end{aligned} \tag{2.12}$$

As a consequence, the one-electron integrals $h_{\mathrm{val},\alpha_\mathrm{A}\beta_\mathrm{A}}$ are evaluated as

$$h_{\mathrm{val},\alpha_\mathrm{A}\beta_\mathrm{A}} = \delta_{\alpha\beta}\left(\left\langle \alpha_\mathrm{A}\left|-\frac{1}{2}\nabla^2 - \hat{\mathcal{V}}_\mathrm{A}\right|\alpha_\mathrm{A}\right\rangle - \sum_{\mathrm{B}\neq\mathrm{A}}^{n_\mathrm{nuclei}} V_\mathrm{AB}\right) \tag{2.13}$$

In the INDO model, the CNDO approximation for the two-electron integrals is not applied to the one-center two-electron integrals $(\alpha_\mathrm{A}\beta_\mathrm{A}|\gamma_\mathrm{A}\zeta_\mathrm{A})$. The reason for this choice is the fact that these integrals are the largest two-electron integrals after the Coulomb integrals.

2.1.5 Parameterization

Because of their inherent approximations, the models described above are not able to give a reliable prediction of molecular properties. In order to transform them into computational methods that are capable to give accurate results, some or all of the integrals must be parameterized in reference to experimental data.

The parameters can be either assigned according to atomic experimental data, such as ionization energies, or optimized to fit to experimental data of a set of molecules, known as reference set. In general, the molecular properties used for the optimization procedure are heats of formation, geometries, dipole moments and ionization potentials. The integrals that are not made into parameters can be directly evaluated from the AOs.

2.2 Modified NDDO Methods

Modified NDDO methods are parameterizations of the NDDO model where some or all of the integrals have been parameterized with the fitting procedure. Since the fitting procedure is inherently designed to give the best agreement with experimental data, modified NDDO methods generally perform better than non-NDDO methods, provided that the molecular system under study is sufficiently similar to the reference set molecules. In addition, they involve fewer approximations compared to non-NDDO methods, as they are derived from the NDDO model.

Some of the most commonly used semiempirical SCF methods to date, such as PM3, PM6 and PM7, belong to this category. The only differences between these methods lie in the functional form of the core-core interaction terms V_{AB} and in the way the optimization procedure has been performed.

All modified NDDO methods consider only valence s- and p-type STOs as basis functions, whose exponents are, respectively, ζ_s and ζ_p.

The integrals $\left\langle \alpha_{\mathrm{A}} \left| \hat{\mathcal{V}}_{\mathrm{B}} \right| \beta_{\mathrm{A}} \right\rangle$ in Eq. (2.9) are evaluated as Coulomb integrals in which the charge distribution of the atomic core is described by a valence s-type AO

$$\left\langle \alpha_{\mathrm{A}} \left| \hat{\mathcal{V}}_{\mathrm{B}} \right| \beta_{\mathrm{A}} \right\rangle = C_{\mathrm{B}} \left(\alpha_{\mathrm{A}} \beta_{\mathrm{A}} | s_{\mathrm{B}} s_{\mathrm{B}} \right) \tag{2.14}$$

where C_{B} represents the charge of the atomic core, i.e. the nuclear charge plus the charge of core electrons. The one-electron integrals in which the AOs are located on the same atom are therefore evaluated as

$$h_{\mathrm{val},\alpha_{\mathrm{A}}\beta_{\mathrm{A}}} = \delta_{\alpha\beta} U_{\alpha_{\mathrm{A}}\alpha_{\mathrm{A}}} - \sum_{\mathrm{B}\neq\mathrm{A}}^{n_{\mathrm{nuclei}}} C_{\mathrm{B}} \left(\alpha_{\mathrm{A}} \beta_{\mathrm{A}} | s_{\mathrm{B}} s_{\mathrm{B}} \right) \tag{2.15}$$

where $U_{\alpha_{\mathrm{A}}\alpha_{\mathrm{A}}}$ represents the integral $\left\langle \alpha_{\mathrm{A}} \left| -\frac{1}{2}\nabla^2 - \hat{\mathcal{V}}_{\mathrm{A}} \right| \alpha_{\mathrm{A}} \right\rangle$.

The one-electron integrals in which the AOs are located on different atoms are assumed to be proportional to the overlap integral between the AOs involved

$$h_{\mathrm{val},\alpha_{\mathrm{A}}\beta_{\mathrm{B}}} = \frac{1}{2} \left(\beta_{\alpha_{\mathrm{A}}} + \beta_{\beta_{\mathrm{B}}} \right) S_{\alpha_{\mathrm{A}}\beta_{\mathrm{B}}} \tag{2.16}$$

where the proportionality constant is taken as the arithmetic mean between the atomic parameters $\beta_{\alpha_{\mathrm{A}}}$ and $\beta_{\beta_{\mathrm{B}}}$. This assumption appears to be reasonable from a physical point of view as the integral $h_{\mathrm{val},\alpha_{\mathrm{A}}\beta_{\mathrm{B}}}$ represents the interaction energy of the electronic distribution $\alpha^*_{\mathrm{A}}\beta_{\mathrm{B}}$ with the atomic cores A and B. It is important

to notice that the overlap elements $S_{\alpha_A \beta_B}$ are evaluated explicitly, even thought the overlap matrix is reduced to the identity matrix as a consequence of the ZDO approximation.

The one-center two-electron integrals, namely the Coulomb integrals $g_{\alpha\beta} = (\alpha_A \alpha_A | \beta_A \beta_A)$ and the Exchange integrals $h_{\alpha\beta} = (\alpha_A \beta_A | \beta_A \alpha_A)$, may be either directly evaluated from atomic experimental data or made into parameters to be optimized.

The two-center two-electron integrals are modeled as multipole-multipole interactions that are damped according to the Dewar-Sabelli-Klopman (DSK) approximation $[2-4]$ in order to account for the proper limiting behavior, i.e. classical electrostatic interactions for long enough interatomic distances and one-center two-electron integrals for $r_{AB} \to 0$

$$(\alpha_A \beta_A | \gamma_B \zeta_B) = \sum_{i \in A} \sum_{j \in B} \frac{Q_i Q_j}{\sqrt{r_{ij}^2 + \frac{1}{4}\left(\frac{1}{\rho_i} + \frac{1}{\rho_j}\right)^2}} \tag{2.17}$$

where the summations are over the charges used to mimic the AOs, while the atomic parameters ρ_i and ρ_j represent the one-center two-electron integrals. The main reason behind this approximation can be traced to the limited computational resources available when modified NDDO methods were initially developed.

The core-core interaction terms V_{AB} in Eq. (2.4) cannot be simply evaluated as classical electrostatic interactions between cores. In fact, because of the inherent approximations of modified NDDO methods, these terms would not be canceled out by the electron–electron interaction terms $P_{\alpha\beta}\left(F_{\text{val},\alpha\beta} + h_{\text{val},\alpha\beta}\right)$ at long distances, resulting in a net repulsion between uncharged atoms. In order to solve this problem, the core-core interaction term V_{AB} is evaluated as the sum between the Coulomb integral $[C_A C_B\,(s_A s_A | s_B s_B)]$, in which the charge distributions of the atomic cores are described by valence s-type AOs centered on their respective atoms, and a parameterized function $f\left(r_{AB}\right)$, which accounts for the proper limiting behavior.

$$V_{AB} = [C_A C_B\,(s_A s_A | s_B s_B)] + f\left(r_{AB}\right) \tag{2.18}$$

where C_A and C_B represent the charges of the atomic cores. The specific functional form of the correction term $f\left(r_{AB}\right)$ depends on the particular method considered.

2.2.1 MNDO

MNDO [5], published in 1977 by Dewar and Thiel, was the first reported modified NDDO method. In its first implementation, the parameters were optimized according to a reference set of small, common molecules containing only H, C, N and O; more recently, the parameterization for all the main group elements, except for the noble gases, has been reported [6].

The core-core interaction terms in MNDO are modeled as

$$V_{\mathrm{AB}}^{\mathrm{MNDO}} = \left[C_{\mathrm{A}} C_{\mathrm{B}} \left(s_{\mathrm{A}} s_{\mathrm{A}} | s_{\mathrm{B}} s_{\mathrm{B}}\right)\right] \left(1 + \mathrm{e}^{-\alpha_{\mathrm{A}} r_{\mathrm{AB}}} + \mathrm{e}^{-\alpha_{\mathrm{B}} r_{\mathrm{AB}}}\right) \tag{2.19}$$

where the atomic parameters α are fitted to molecular data. For the interactions involving $\mathrm{O}-\mathrm{H}$ and $\mathrm{N}-\mathrm{H}$ bonds, a slightly modified version is used

$$V_{\mathrm{AH}}^{\mathrm{MNDO}} = C_{\mathrm{A}} C_{\mathrm{B}} \left(s_{\mathrm{A}} s_{\mathrm{A}} | s_{\mathrm{H}} s_{\mathrm{H}}\right) \left(1 + r_{\mathrm{AH}} \mathrm{e}^{-\alpha_{\mathrm{A}} r_{\mathrm{AH}}} + \mathrm{e}^{-\alpha_{\mathrm{H}} r_{\mathrm{AH}}}\right) \qquad \mathrm{A} = \mathrm{N},\ \mathrm{O} \tag{2.20}$$

The one-center two-electron integral parameters are assigned according to atomic spectra, while the other parameters are fitted to molecular data. In addition, MNDO uses the following approximations

$$\zeta_{s_{\mathrm{A}}} = \zeta_{p_{\mathrm{A}}} \qquad \mathrm{A} = \mathrm{C},\ \mathrm{N},\ \mathrm{O} \tag{2.21}$$

$$\beta_{s_{\mathrm{A}}} = \beta_{p_{\mathrm{A}}} \qquad \mathrm{A} = \mathrm{N},\ \mathrm{O} \tag{2.22}$$

which save time on the parameterization procedure, without impairing the quality of the results.

The optimization procedure contributed to a large increase in accuracy over earlier, non-NDDO methods. Unfortunately, soon after MNDO was completed, several systematic errors were found, the most important of which is the almost complete lack of hydrogen bonding [7, 8]. This essentially precludes MNDO being used to study systems of biological interest, such as proteins, nucleic acids and polysaccharides.

2.2.2 AM1

Austin Model 1 (AM1) [9], published by Dewar, Zoebisch, Healy, and Stewart in 1985, is an improved version of MNDO, specifically designed in order to correct the MNDO tendency to overestimate the repulsion between atoms separated by

about $2-3$ Å. The main consequences of this fault are failure to reproduce intermolecular interactions, such as van der Waals attractions and hydrogen bonds, and predicted activation energies that are generally too large [7, 8].

In attempt to solve these problems, the MNDO core-core repulsion terms were modified by adding Gaussian functions centered at internuclear points

$$V_{\mathrm{AB}}^{\mathrm{AM1}} = V_{\mathrm{AB}}^{\mathrm{MNDO}} + \frac{C_{\mathrm{A}}C_{\mathrm{B}}}{r_{\mathrm{AB}}}\sum_{k}\left[a_{k_{\mathrm{A}}}\mathrm{e}^{-b_{k_{\mathrm{A}}}\left(r_{\mathrm{AB}}-c_{k_{\mathrm{A}}}\right)^2} + a_{k_{\mathrm{B}}}\mathrm{e}^{-b_{k_{\mathrm{B}}}\left(r_{\mathrm{AB}}-c_{k_{\mathrm{B}}}\right)^2}\right] \quad (2.23)$$

where the parameters a_k, b_k and c_k are fitted to molecular data. The number of Gaussian functions per atom varies between two and four, depending on the atom. In addition, the STO exponents $\zeta_{s_{\mathrm{A}}}$ and $\zeta_{p_{\mathrm{A}}}$ are not taken to be the same.

In its first implementation AM1 was parameterized only for H, C, N and O; more recently, the parameterization for all the main group elements, except for the noble gases, has been reported [6].

Compared to MNDO, AM1 led to a marked improvement in modeling hydrogen bonding and predicting activation energies, with no increase in computational time. However, there still remain some well known limitations, such as the failure to predict correctly the hydrogen bond energies and geometries [7, 8].

2.2.3 PM3

Modified Neglect of Diatomic Overlap, Parametric Method Number 3 (PM3) [10 – 13], published in 1989 by Stewart, is a variation of AM1. The main difference between AM1 and PM3 lies in the way the parameterization was performed. In fact, while in AM1 the parameterization was done essentially by hand according to a small reference set in which the molecules were chosen with great guidance from chemical intuition, PM3 employed a more statistical approach where all parameters, including one-center two-electron integrals, were optimized simultaneously according to a significantly larger reference set. In addition, only two Gaussian functions per atom were added to the MNDO core-core interaction terms.

In its first implementation, PM3 was parameterized for H, C, N, O, F, Al, Si, P, S, Cl, Br, and I; more recently, the parameterization for all the main group elements, except for the noble gases, has been reported [6].

Even though PM3 provides, in general, the best set of parameters for a particular reference set, human intervention is still required for the selection of the experimental data. For this reason, PM3 will not necessarily perform better than AM1, although it is likely to give the best results, provided that the

molecular system under study is sufficiently similar to the reference set molecules [8]. Possibly because of this, PM3 has not been used as widely as AM1 has been.

2.3 Recent Developments: PM6 and PM7

2.3.1 PM6

After two unpublished methods, namely PM4 and PM5, PM6 [14] was published by Stewart in 2007. Compared with earlier NDDO methods, only minor changes were introduced in PM6.

In earlier modified NDDO methods, the two-center two-electron integrals $(s_\mathrm{A}s_\mathrm{A}|s_\mathrm{B}s_\mathrm{B})$ and the electron-core interactions do not converge to the exact electrostatic interaction as the interatomic distance increases. As a consequence, there is always a small net repulsion between uncharged atoms separated by distances greater than about 5 Å. In order to solve this problem, in PM6 the core-core interaction terms were modified according to Eq. (2.24).

$$V_\mathrm{AB}^\mathrm{PM6} = \left[C_\mathrm{A}C_\mathrm{B}\left(s_\mathrm{A}s_\mathrm{A}|s_\mathrm{B}s_\mathrm{B}\right)\right]\left(1 + x_\mathrm{AB}\mathrm{e}^{-\alpha_\mathrm{AB}r_\mathrm{AB}-\alpha_\mathrm{AB}0.0003r_\mathrm{AB}^6}\right) \tag{2.24}$$

As the interatomic distance increases, Eq. (2.24) converges to the exact electrostatic interaction. The rate of convergence, and therefore the overall accuracy of the method, were found to be increased by the addition of the perturbation term $-\alpha_\mathrm{AB}0.0003r_\mathrm{AB}^6$. In addition, the introduction of the diatomic parameters x_AB, which are essential for the correct treatment of Group IA elements and molybdenum, was also found to be generally beneficial.

Specific corrections to Eq. (2.24) were made for a small number of interactions, namely that involving $\mathrm{O}-\mathrm{H}$ and $\mathrm{N}-\mathrm{H}$, $\mathrm{C}-\mathrm{C}$, and $\mathrm{Si}-\mathrm{O}$ bonds. In particular, the core-core interactions involving $\mathrm{O}-\mathrm{H}$ and $\mathrm{N}-\mathrm{H}$ bonds are evaluated as

$$V_\mathrm{AH}^\mathrm{PM6} = \left[C_\mathrm{A}C_\mathrm{H}\left(s_\mathrm{A}s_\mathrm{A}|s_\mathrm{H}s_\mathrm{H}\right)\right]\left(1 + x_\mathrm{AH}\mathrm{e}^{-\alpha_\mathrm{AH}r_\mathrm{AH}^2}\right) \tag{2.25}$$

for a better agreement with reference hydrogen bonding energies.

d-type STOs were added to expand the range of applicability of earlier modified NDDO methods to transition metals. Moreover, the addition of *d*-type STOs to several main-group elements that are potentially hypervalent resulted in a marked increase in accuracy with little extra computational cost.

The parameterization was performed for 70 elements, namely all main group elements plus several transition metals and a lanthanide, according to a reference

set consisting of several thousand molecules. During the parameterization procedure, particular attention was given to molecules of biological interest.

As shown by the extensive comparison reported by Stewart, PM6 represents a large improvement over earlier modified NDDO methods. However, there still remain some well known systematic errors, such as the failure to accurately reproduce hydrogen bonding and infinite errors arising in the application to crystalline solids.

2.3.2 PM7

PM7 [15], the most recent semiempirical modified NDDO method, published in 2012 by Stewart, was specifically designed to expand the range of applicability of earlier modified NDDO methods to crystalline systems. In fact, these kinds of systems cannot be modeled with modified NDDO methods prior to PM7 because they give rise to infinite errors.

The cause of this fault was traced to the fact that the rate of convergence of the nucleus–nucleus interactions to the exact electrostatic interactions depends on the specific atoms involved. Although these differences are very small for isolated species, when infinite sums are involved, as in the case of solids, such small differences generate infinite errors. In PM7, in order to avoid this kind of error, the two-center two-electron integrals are evaluated with a modified version of the DSK approximation; the correction term for the two-center two-electron integral $(s_\mathrm{A}s_\mathrm{A}|s_\mathrm{B}s_\mathrm{B})$ is given in Eq. (2.26).

$$(s_\mathrm{A}s_\mathrm{A}|s_\mathrm{B}s_\mathrm{B}) = \frac{1}{r_\mathrm{AB}} e^{-0.22(r_\mathrm{AB}-7)^2} + \left(\frac{1}{\sqrt{r_\mathrm{AB}^2 + \frac{1}{4}\left(\frac{1}{\rho_\mathrm{A}} + \frac{1}{\rho_\mathrm{B}}\right)^2}} \right) \left(1 - e^{-0.22(r_\mathrm{AB}-7)^2}\right) \tag{2.26}$$

The failure of PM6 to accurately reproduce intermolecular interactions, namely van der Waals attractions and hydrogen bonding, greatly limits its applicability to model biochemical systems. Since only minor improvements were attained in earlier modified NDDO methods by modifying the core-core interaction terms, post-SCF corrections were implemented into PM7. These corrections were incorporated into the method before the parameterization procedure was performed.

Korth's model [16] was used to construct the hydrogen bonding interaction

term, as given in Eq. (2.27).

$$V_{\mathrm{AB,\,H-bond}} = -2.5\,(\cos\theta)^4\, e^{-80(r_{\mathrm{AB}}-2.67)^2} \tag{2.27}$$

It is important to notice that the correction term in the above equation depends only on the geometry and not on the specific donor-acceptor couple involved in the hydrogen bond.

The dispersion correction term is evaluated as a modified version of the Jurečka's dispersion term [17], in order to account for the reduction in instantaneous correlation energy for distant atoms in solids arising from short-range correlation.

$$V_{\mathrm{AB,\,disp}} = V_{\mathrm{Jurečka}} \left(1 - \mathrm{e}^{-(r_{\mathrm{AB}}-6.5)^2}\right) \tag{2.28}$$

Since the dispersion interactions and the electron correlation effects are accounted for by the dispersion correction term, the Gaussian functions in the core-core interaction terms were deleted, except for H, C, N, and O, for which sufficient high-quality reference data were available.

With the post-SFC corrections for the intermolecular interactions, PM7 reaches an accuracy comparable to that of much more expensive ab initio SCF methods, while being applicable to much larger molecular systems, containing up to about a thousand atoms.

References

1. Pople JA, Santry DP, Segal GA. Approximate Self-Consistent Molecular Orbital Theory. I. Invariant Procedures. J Chem Phys. 1965;43(10):S129-35. doi:10.1063/1.1701475.

2. Dewar MJS, Sabelli NL. The s.p-o. (Split-p-Orbital) Method. II. Further Definition and Application to Acetylene. Proc R Soc London, Ser A. 1961;264(1319):431-44. doi:10.1098/rspa.1961.0209.

3. Dewar MJS, Sabelli NL. THE SPLIT p-ORBITAL (S.P.O.) METHOD. III. RELATIONSHIP TO OTHER M.O. TREATMENTS AND APPLICATION TO BENZENE, BUTADIENE, AND NAPHTHALENE. J Phys Chem. 1962:66(12);2310-6. doi:10.1021/j100818a007

4. Klopman G. A Semiempirical Treatment of molecular Structures. II. Molecular Terms and Application to diatomic Molecules. J Am Chem Soc. 1964;86(21):4550-7. doi:10.1021/ja01075a008.

5. Dewar MJS, Thiel W. Ground states of molecules. 38. The MNDO method. Approximations and parameters. J Am Chem Soc. 1977;99(15):4899-907. doi:10.1021/ja00457a004.

6. Stewart JJP. Optimization of parameters for semiempirical methods IV: extension of MNDO, AM1, and PM3 to more main group elements. J Mol Model. 2004;10(2):155-64. doi:10.1007/s00894-004-0183-z.

7. Levine IN. Quantum Chemistry. 5 ed. New Jersey: Prentice Hall; 2000.

8. Jensen F. Introduction to Computational Chemistry. 2nd ed. Chisester: John Wiley & Sons Ltd; 2007.

9. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. J Am Chem Soc. 1985;107(13):3902-9. doi:10.1021/ja00299a024.

10. Stewart JJP. Optimization of parameters for semiempirical methods I. Method. J Comp Chem. 1989;10(2):209-20. doi:10.1002/jcc.540100208.

11. Stewart JJP. Optimization of parameters for semiempirical methods II. Applications. J Comp Chem. 1989;10(2):221-64. doi:10.1002/jcc.540100209.

12. Stewart JJP. Optimization of parameters for semiempirical methods. III Extension of PM3 to Be, Mg, Zn, Ga, Ge, As, Se, Cd, In, Sn, Sb, Te, Hg, Tl, Pb, and Bi. J Comp Chem. 1991;12(3):320-41. doi:10.1002/jcc.540120306.

13. Anders E, Koch R, Freunscht P. Optimization and application of lithium parameters for PM3. J Comp Chem. 1993;14(11):1301-12. doi:10.1002/jcc.540141106.

14. Stewart JJP. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. J Mol Model. 2007;13(12):1173-213. doi:10.1007/s00894-007-0233-4.

15. Stewart JJP. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. J Mol Model. 2013;19(1):1–32.

16. Korth M. Third-Generation Hydrogen-Bonding Corrections for Semiempirical QM Methods and Force Fields. J Chem Theory Comput. 2010;6(12):3808-16. doi:10.1021/ct100408b.

17. Jurečka P, Černý J, Hobza P, Salahub DR. Density functional theory augmented with an empirical dispersion term. Interaction energies and geometries of 80 noncovalent complexes compared with ab initio quantum mechanics calculations. J Comp Chem. 2007;28(2):555-69. doi:10.1002/jcc.20570.

3 Studying Proteins with Semiempirical SCF Methods

As shown in the previous chapter, the general SCF RH problem can be formulated in an orthonormal AO basis as the non-linear diagonalization problem

$$\mathbf{C}^{\dagger}\mathbf{F}\left(\mathbf{C}\right)\mathbf{C} = \boldsymbol{\varepsilon} \tag{3.1}$$

where the coefficient matrix $\mathbf{C}$ is the change of basis matrix that diagonalizes the Fock matrix $\mathbf{F}\left(\mathbf{C}\right)$.

As a consequence of the use of matrix algebra techniques to solve the SCF RH equations, standard semiempirical SCF methods scale as the third power of the number of basis functions. These methods are therefore limited to molecular systems containing up to about a thousand atoms; larger systems, such as common biomolecules, are out of reach for current computers.

Over the past two decades different methods have been developed to solve this problem. This chapter will concentrate on the localized molecular orbital (LMO) method MOZYME and on its application to the study of γ-chymotrypsin.

3.1 MOZYME

MOZYME [1] is a near-linear scaling method developed by Stewart in 1995 for the study of large molecular systems within the framework of semiempirical NDDO methods. The near-linear scaling is achieved in MOZYME by employing a pseudodiagonalization procedure and by exploiting the local character of LMOs interactions and the sparsity of the associated matrices.

3.1.1 Pseudodiagonalization

In order to understand the pseudodiagonalization procedure, it is necessary to examine the matrix $\mathbf{C}^{\dagger}\mathbf{F}\left(\mathbf{C}\right)\mathbf{C}$ in Eq. 3.1 more in detail. The matrix $\mathbf{C}^{\dagger}\mathbf{F}\left(\mathbf{C}\right)\mathbf{C}$ is symmetric and shows the following block structure

$$\mathbf{C}^{\dagger}\mathbf{F}\left(\mathbf{C}\right)\mathbf{C} = \begin{pmatrix} \mathbf{C}_o^{\dagger}\mathbf{F}\left(\mathbf{C}\right)\mathbf{C}_o & \mathbf{C}_o^{\dagger}\mathbf{F}\left(\mathbf{C}\right)\mathbf{C}_v \\ \mathbf{C}_v^{\dagger}\mathbf{F}\left(\mathbf{C}\right)\mathbf{C}_o & \mathbf{C}_v^{\dagger}\mathbf{F}\left(\mathbf{C}\right)\mathbf{C}_v \end{pmatrix} \tag{3.2}$$

where the matrices $\mathbf{C}_o$ and $\mathbf{C}_v$ denote the occupied and virtual blocks, respectively, of the coefficient matrix $\mathbf{C}$ [2]. It is important to notice that the coefficient matrix $\mathbf{C}$ is unitary in order to preserve the orthonormality of the MOs.

Since the total electronic energy is invariant under a unitary transformation among occupied or virtual orbitals, the conventional full diagonalization procedure can be replaced by a pseudodiagonalization procedure, which consists in the annihilation of the blocks $\mathbf{C}_o^{\dagger}\mathbf{F}\left(\mathbf{C}\right)\mathbf{C}_v$ and $\mathbf{C}_v^{\dagger}\mathbf{F}\left(\mathbf{C}\right)\mathbf{C}_o$ [2]: as the matrix $\mathbf{C}^{\dagger}\mathbf{F}\left(\mathbf{C}\right)\mathbf{C}$ is symmetric, it is necessary to annihilate only one of these two blocks. In other words, the necessary and sufficient condition for solving an SCF problem is that all virtual-occupied Fock matrix elements in the MO basis must be zero

$$\sum_{i=1}^{n_{occ}}\sum_{j=1}^{n_{vir}}\left|\left\langle \psi_i \left| \hat{\mathcal{F}} \right| \psi_j \right\rangle\right| = 0 \tag{3.3}$$

where the summations are over the occupied and the virtual LMOs, respectively.

3.1.2 Localized Molecular Orbitals

In standard SCF methods the total electronic wavefunction is described in terms of canonical MOs, which are delocalized over the entire molecular system. However, as the total electronic wavefunction is invariant under a unitary transformation of the occupied MOs among themselves, a molecular system may be equivalently described by other sets of MOs that are linear combinations of the canonical MOs.

A possible choice is the use of LMOs, which, unlike canonical MOs, are highly localized in space. Compared to canonical MOs, LMOs present several advantages in the study of large molecular systems. In particular:

- Virtual-occupied interactions between LMOs located on atoms separated by distances greater than a user-defined cutoff distance are neglected, i.e. the blocks $\mathbf{C}_o^{\dagger}\mathbf{F}\left(\mathbf{C}\right)\mathbf{C}_v$ and $\mathbf{C}_v^{\dagger}\mathbf{F}\left(\mathbf{C}\right)\mathbf{C}_o$ are sparse. As a consequence, the annihilation procedure is made much faster

- The evaluation of density matrix elements is limited to those atoms upon which the LMOs are localized, i.e. the density matrix $\mathbf{P}$ is sparse. Since the total electronic energy and the Fock matrix elements depend on the density

matrix elements as

$$\begin{aligned} \mathcal{E}_{\mathrm{el}} &= \frac{1}{2}\sum_{\alpha}^{m}\sum_{\beta}^{m} P_{\alpha\beta}\left(F_{\mathrm{val},\alpha\beta} + h_{\mathrm{val},\alpha\beta}\right) + \sum_{\mathrm{A}}^{n_{\mathrm{nuclei}}}\sum_{\mathrm{B>A}}^{n_{\mathrm{nuclei}}} V_{\mathrm{AB}} \\ F_{\mathrm{val},\alpha\beta} &= h_{\mathrm{val},\alpha\beta} + \sum_{\gamma}^{m}\sum_{\delta}^{m} P_{\gamma\delta}\left[(\alpha\beta|\gamma\delta) - \frac{1}{2}(\alpha\delta|\gamma\beta)\right] \end{aligned} \tag{3.4}$$

many one- and two-electron integrals do not need to be calculated

As a consequence, for sufficiently large systems, the computational effort required for the annihilation procedure and for building the density matrix scales linearly with the number of basis functions.

An overall linear scaling can only be achieved if each step involved in the SCF procedure scales linearly with the size of the system. In order to avoid the initial full diagonalization procedure employed in conventional SCF methods, which scales as the third power of the number of basis functions, a special method has been developed to generate the starting LMOs.

For each atom, a set of hybrid AOs is constructed from a minimal basis set of AOs. The hybrid AOs are oriented in the direction of bonded atoms, according to the Lewis structure of the system. If the number of bonded atoms is less than the number of hybrid AOs, the remaining hybrid AOs are oriented in such a way that the hybrid AOs form an orthonormal set. As a consequence of the orthonormality, the hybrid AOs generated with this procedure are ideally suited for constructing the starting LMOs.

For each bonding pair of electrons in the Lewis structure, an occupied and virtual LMOs are inferred to exist. In order for these LMOs to form an orthonormal set, pairs of hybrid AOs are combined to form diatomic occupied and virtual LMOs that diagonalize the corresponding Fock matrices.

The hybrid AOs that are not involved in bonds are assigned either to unoccupied LMOs if they correspond to a lone pair or to an unused orbital on an anion, or to virtual LMOs if they correspond to an unused orbital on a cation.

The major drawback of this procedure is that it is limited to closed-shell systems that can be represented by a Lewis structure. However, this does not constitute a problem for most molecules of biological interest.

3.1.3 Annihilation in the LMO Basis

As noted before, the necessary and sufficient condition for solving an SCF problem is that all virtual-occupied Fock matrix elements in the MO basis must be zero.

From the practical point of view, the annihilation procedure is performed by means of a series of 2×2 Jacobi rotations [3]. The result of each rotation is to mix an occupied and a virtual LMO in order to obtain two new LMOs for which the interaction is zero. In particular, to annihilate the matrix element $\left\langle \psi_i \left| \hat{\mathcal{F}} \right| \psi_j \right\rangle = F_{ij}$ the two new LMOs are given by

$$\begin{aligned} \psi'_i &= \alpha\psi_i + \beta\psi_j \\ \psi'_j &= -\beta\psi_i + \alpha\psi_j \end{aligned} \tag{3.5}$$

where the coefficients α and β are defined as

$$\begin{aligned} \alpha &= \sqrt{\frac{1}{2}}\left(1 + \frac{D}{\sqrt{4F_{ij}^2 + D^2}}\right) \\ \beta &= \sqrt{1-\alpha^2} \quad \text{if } F_{ij} < 0 \\ \beta &= -\sqrt{1-\alpha^2} \quad \text{if } F_{ij} \geq 0 \end{aligned} \tag{3.6}$$

and

$$D \equiv F_{ii} - F_{jj} \tag{3.7}$$

As the annihilation procedure proceeds, the number of atoms spanned by each LMO increases until every LMO includes contributions from every atom in the system. However, the atoms for which the contribution to an LMO is sufficiently small, i.e. smaller than a user-defined threshold, can be deleted from the LMO without significantly affecting the results [3]. As a consequence, for sufficiently large systems, the number of atoms involved in each LMO becomes independent of the size of the system; in practice, each LMO spans on average up to about 100-130 atoms.

It is therefore evident that for small systems the MOZYME method offers no advantages in terms of computational efficiency over conventional NDDO methods.

3.1.4 Long-Range Interactions

As shown before, the evaluation of density matrix elements is limited to those atoms upon which the LMOs are localized. As a consequence, many one- and two-electron integrals do not need to be computed. Moreover, most of the remaining integrals can be safely neglected in order to save computational resources. In fact, these additional approximations have a negligible effect on the results compared to the NDDO approximations.

In semiempirical methods, the one-electron integrals in which the AOs are located on different atoms are assumed to be proportional to the overlap integral between the AOs

$$h_{\mathrm{val},\alpha_\mathrm{A}\beta_\mathrm{B}} = \frac{1}{2}\left(\beta_{\alpha_\mathrm{A}} + \beta_{\beta_\mathrm{B}}\right) S_{\alpha_\mathrm{A}\beta_\mathrm{B}} \tag{3.8}$$

where the proportionality constant is taken as the arithmetic mean between the atomic parameters $\beta_{\alpha_\mathrm{A}}$ and β_{β_B}. For this reason, one-electron integrals involving atoms separated by distances greater than about 6-7 Å have negligible values and do not need to be calculated.

The two-center two-electron integrals are modeled in conventional NDDO methods as multipole-multipole interactions. A detailed analysis of these interactions suggests that at separation distances between 6 and 30 Å, most of the two-center two-electron integrals can be neglected, while for distances greater than 30 Å the remaining integrals are reduced to simple electrostatic interactions between the net charges located on the atoms involved. These interactions are evaluated as Coulomb integrals in which the charge distributions are considered to be described by s-type AOs centered on their respective atoms.

Another long-range interaction that needs to be considered is the interaction between a point-charge and a lone pair. In fact, a lone-pair pointing in the direction of a positive charge will have a stabilizing effect, while a lone-pair pointing in the opposite direction will have a destabilizing effect; the situation is reversed for a negative charge. In addition, a lone-pair oriented at 90° with respect to a charge will be subjected to a torque. Since the charge density of a lone pair on an atom A depends on the density matrix elements $P_{s_\mathrm{A}p_{x,\mathrm{A}}}$, $P_{s_\mathrm{A}p_{y,\mathrm{A}}}$ and $P_{s_\mathrm{A}p_{z,\mathrm{A}}}$, these effects can be expressed in terms of Fock matrix elements as

$$\begin{aligned}
F'_{s_\mathrm{A}p_{x,\mathrm{A}}} &= F_{s_\mathrm{A}p_{x,\mathrm{A}}} - \sum_\mathrm{B} Q_\mathrm{B}\left(s_\mathrm{A}p_{x,\mathrm{A}}|s_\mathrm{B}s_\mathrm{B}\right) P_{s_\mathrm{A}p_{x,\mathrm{A}}} \\
F'_{s_\mathrm{A}p_{y,\mathrm{A}}} &= F_{s_\mathrm{A}p_{y,\mathrm{A}}} - \sum_\mathrm{B} Q_\mathrm{B}\left(s_\mathrm{A}p_{y,\mathrm{A}}|s_\mathrm{B}s_\mathrm{B}\right) P_{s_\mathrm{A}p_{y,\mathrm{A}}} \\
F'_{s_\mathrm{A}p_{z,\mathrm{A}}} &= F_{s_\mathrm{A}p_{z,\mathrm{A}}} - \sum_\mathrm{B} Q_\mathrm{B}\left(s_\mathrm{A}p_{z,\mathrm{A}}|s_\mathrm{B}s_\mathrm{B}\right) P_{s_\mathrm{A}p_{z,\mathrm{A}}}
\end{aligned} \tag{3.9}$$

where the summation is over the point-charges.

3.1.5 SCF Procedure in the LMO Basis

The SCF procedure in the LMO basis is similar to the conventional SCF procedure illustrated in Section 1.6. The only step for which there is no equivalent in the conventional SCF procedure is the renormalization step, which is necessary after

tidying up the LMOs.

An unexpected consequence of the use of the LMOs is that the MOZYME method is more stable compared to conventional NDDO methods. In fact, sometimes using these methods the SCF procedure fails to converge. In addition, the SCF procedure in the LMO basis converges faster than the conventional SCF procedure, requiring typically only three or four iterations.

Calculations on selected proteins show that the energies obtained with the MOZYME method are almost identical to those obtained with conventional NDDO methods, provided that the cut-off distance is set to large enough values. These calculations also show that the computational effort scales between M_{basis} and $M_{\text{basis}}^{1.5}$. In fact, the linear scaling is observed only for sufficiently large systems, for which the LMO properties that depend on the size of the system become constant.

3.2 Modeling Protein Structures with MOPAC

This section provides an essential guide covering the practical aspects and the common issues involved in modeling protein structures with MOPAC. The procedure for building a starting model is provided, along with a complete worked example for the γ-chymotrypsin.

3.2.1 MOPAC

MOPAC [4] (Molecular Orbital PACkage) is a computer program that implements several semiempirical modified NDDO methods. The first version of MOPAC, published in 1983, was developed by Stewart while working in Dewar's research group at the University of Texas at Austin.

All the calculations reported in the following section were performed with MOPAC2012, the latest version of MOPAC. The methods available in MOPAC2012 are MNDO, AM1, PM3, RM1, PM6 and PM7. For all these methods, except for MNDO and RM1, the Sparkle Model for calculations on lanthanide complexes is also available.

The near-linear scaling method MOZYME is fully implemented in MOPAC2012, allowing calculations on large closed-shell molecular systems, such as common biomolecules and covalently bonded inorganic solids, containing up to about 15000 atoms. The implementation of MOZYME, in conjunction with that of the most recent semiempirical modified NDDO methods, namely PM6 and PM7, provides a tool that may become routine in the future for the study of

large systems of biological interest, even using relatively modest machines, such as personal computers.

Several types of calculations can be performed with MOPAC2012, the most important of which are geometry optimization, transition states localization, vibrational frequencies calculation, thermodynamic properties calculation and evaluation of solvent effects with COSMO (COnductor-like Screening MOdel).

MOPAC2012 is available for all major operating systems (Linux, Mac OS X and Windows) and it is free for academic, non-profit use. More information on MOPAC can be found at http://openmopac.net.

3.2.2 Building a Starting Model

The building of a starting model, i.e. a good energy minimum that can be used as starting point for a wide range of calculations, is of paramount importance in the study of proteins with MOPAC. In fact, if the starting model is faulty, e.g. the structure is not fully optimized, subsequent calculations are more likely to give inaccurate results.

The procedure for building a starting model for the γ-chymotrypsin is provided in the following section, but the process is very similar for other proteins. The choice to provide an example for the γ-chymotrypsin was dictated by the fact that this protein is a relatively small closed-shell protein, and therefore it can be modeled with MOZYME on inexpensive personal computers. In addition, the procedure for the γ-chymotrypsin is well documented on the MOPAC site.

3.2.2.1 PDB Structure

A de novo structure prediction for proteins is not feasible with MOPAC because the construction of a complete PES would require a virtually infinite amount of computational time. The building of a starting model therefore requires an initial geometry that is likely to lie near an energy minimum on the PES. Protein Data Bank structures are ideally suited for this purpose.

Protein Data Bank (PDB) [5] is the biggest free repository for three-dimensional structural data of large biological molecules. Each structure in the PDB is identified by a unique alphanumeric code, e.g. 8GCH. For each structure, a .pdb file, which is essentially a text file, contains all relevant information: the researchers who determined the structure, the method of structure determination, the structure itself and several additional information, such as missing residues and hetero molecules.

PDB structures are typically obtained either by X-ray diffraction or NMR spectroscopy. When selecting the initial geometry for building a starting model, NMR structures should be avoided because they tend to give less accurate results. In addition, if two or more PDB structures are available for a protein, the most complete should be used.

Because of the way in which they are determined, the large majority of X-ray PDB structures suffers from several errors, which must be corrected in order to perform any meaningful calculation. Four types of errors are found in X-Ray PDB structures:

- Positional disorder, which occurs when an atom might be found in more than one position
- Structural disorder, which occurs when a residue is mistaken for another
- Missing residue
- Absence of hydrogen atoms

Positional and structural disorders are resolved by choosing one of the possible structures, while in the case of a missing residue hydrogen atoms are added in order to satisfy the valence requirements. Fortunately, these kinds of errors generally occur in positions where their presence is unimportant, e.g. far from the protein active site.

3.2.2.2 Adding the Hydrogen Atoms

The main structural flaw, which occurs in the large majority of X-Ray PDB structures, is the absence of hydrogen atoms. The most straightforward way to solve this problem is to add the hydrogen atoms directly with MOPAC. Although several third-party programs offer the possibility of performing this procedure, their use is discouraged as they are known to cause several issues during the process.

With MOPAC the hydrogen atoms are added to both the protein and the hetero molecules only by satisfying the valence requirements. In fact, during this process, no charged sites are generated, in order to avoid the development of unrealistic charge distributions. As for the determination of charged sites, this fundamental problem is addressed during the unconstrained optimization.

Let us now examine the procedure for adding the hydrogen atoms with MOPAC. The first step is to download the .pdb file from the Protein Data Bank site. For the γ-chymotrypsin, the ideally suited structure for building

the starting model was determined to be 8GCH [6]. Since MOPAC generates a .pdb file after the addition of the hydrogen atoms, it is important to rename the original PDB file, for example as 8GCH_from_PDB.pdb, in order to avoid overwriting it. A copy of 8GCH_from_PDB.pdb is then created and renamed as 8GCH_from_PDB.mop. In order to add the hydrogen atoms, 8GCH_from_PDB.mop is run using MOPAC.

When the run is completed, MOPAC generates four new text files:

- 8GCH_from_PDB.log - A file that contains the empirical formula of 8GCH with the added hydrogen atoms and lists the problems that are eventually found the after hydrogenation, such as unusually short hydrogen bond distances and bridging hydrogen atoms. These errors are generally unimportant as they can be corrected by the optimization of the positions of the hydrogen atoms

- 8GCH_from_PDB.out - A file that contains detailed information about the procedure performed

- 8GCH_from_PDB.pdb - A file that contains the PDB structure with the added hydrogen atoms

- 8GCH_from_PDB.arc - A file that is similar to 8GCH_from_PDB.pdb, but in the MOPAC format

As shown by 8GCH_from_PDB.log, a total of 2444 hydrogen atoms are added to the original PDB structure, to give $C_{1111}H_{2444}S_{16}N_{300}O_{704}$.

After the addition of the hydrogen atoms, their positions must be optimized. This optimization procedure, like all the others described in the following paragraph, is performed using PM7 in conjunction with MOZYME. The starting point for the optimization procedure is 8GCH_from_PDB.pdb. Three lines are added to this file: the first line contains the keywords PDBOUT MOZYME GNORM=10 PL NOOPT OPT-H, while the second and the third lines can be either left blank or contain comments. The file is then saved as 8GCH_opt-H.mop and run with MOPAC.

The optimization will take in general several hours and up to about 200 iterations. During the initial stages of the run, the iterating heat of formation, which is printed in the .out file as the optimization proceeds, must be monitored carefully. If the iterating heat of formation does not drop to large negative values after about 10 iterations, the run is stopped and 8GCH_from_PDB.pdb is examined for faults.

When the run is completed, MOPAC generates two new text files:

- 8GCH_opt-H.out
- 8GCH_opt-H.pdb - A file that contains the PDB structure in which the positions of the hydrogen atoms are optimized

Even thought during the optimization procedure some salt bridges might spontaneously form, in general they are not likely to because of the energy barrier for proton migration. As for 8GCH, no salt bridges are observed to form.

3.2.2.3 Unconstrained Optimization

Once the optimization of the positions of the hydrogen atoms is completed, an unconstrained optimization, i.e. the optimization of the entire structure, hetero molecules included, is performed. Since the result of the unconstrained optimization will be the starting model for the protein under study, particular care must be used during this process.

In order to generate a realistic starting model, solvation must be taken into account during the unconstrained optimization. The solvation model implemented in MOPAC is the Conductor-like Screening MOdel (COSMO) [7], a model in which the solute is considered to be embedded into a void cavity surrounded by a dielectric continuum of permittivity ε representing the solvent. During the unconstrained optimization of proteins, COSMO is used by adding the keyword EPS=78.4, where 78.4 is the relative permittivity of water at 298 K.

In addition to the use of COSMO, it is of paramount importance to ensure that the starting model is a fully optimized structure. For this reason, the unconstrained optimization is run until the rate of decrease of the iterating heat of formation is very low, usually less than about $1\,\mathrm{kcal\,mol^{-1}}$ per 50 iterations.

Sites that are known to be charged are ionized before performing the unconstrained optimization. The γ-chymotrypsin has three known ionized sites: Ile16 and Asp194, which form a salt bridge, and Asp102. As shown by the PDB file, these residues occur in chains F, G and F, respectively. In order to ionize these residues, three lines are added to 8GCH_opt-H.pdb: the first line contains the keywords PDBOUT SITE=(F16(+),G194(-),F102(-)), while the second and the third lines can be either left blank or contain comments. The file is then saved as 8GCH_make_F16(+)_G194(-)_F102(-).mop and run with MOPAC. When the run is completed, MOPAC generates three new text files.

- 8GCH_make_F16(+)_G194(-)_F102(-).log

- 8GCH_make_F16(+)_G194(-)_F102(-).out
- 8GCH_make_F16(+)_G194(-)_F102(-).pdb

8GCH_make_F16(+)_G194(-)_F102(-).pdb is used as starting point for the unconstrained optimization. During the unconstrained optimization, the keyword SETUP, which allows MOPAC to read the keywords from a .txt file, is used in order to reduce the risk of faulty or missing keywords. In the same folder as 8GCH_make_F16(+)_G194(-)_F102(-).pdb, a new text file named chymotrypsin.txt is created. The line EPS=78.4 MOZYME GNORM=10 CHARGE=-1 PDBOUT is added to this file and the file is saved. Three lines are then added to 8GCH_make_F16(+)_G194(-)_F102(-).pdb: the first line contains the keyword SETUP=chymotrypsin.txt, while the second and the third lines can be either left blank or contain comments. This file is saved as 8GCH_F16(+)_G194(-)_F102(-).mop and run with MOPAC.

The unconstrained optimization will generally take several days, and probably will need to be restarted a couple of times, as the number of iterations allowed for a single run is limited. When the maximum number of iterations for the optimization of 8GCH is exceeded, MOPAC stores the intermediate results in 8GCH_F16(+)_G194(-)_F102(-).mop. In order to restart the optimization, the keyword RESTART is added to chymotrypsin.txt and 8GCH_F16(+)_G194(-)_F102(-).mop is rerun with MOPAC.

After the optimization is completed, 8GCH_F16(+)_G194(-)_F102(-).pdb is used as starting point for a second unconstrained optimization with a finer optimization criterion, usually either GNORM=5 or GNORM=8. This second optimization, which is most likely to not finish successfully, is allowed to run until the rate of decrease of the iterating heat of formation is very low, usually less than about $1\,\mathrm{kcal\,mol^{-1}}$ per 50 iterations. At this point the run is stopped with the command script SHUT and a final unconstrained optimization with the additional keywords RESTART and 1SCF is performed. The resulting structure, which should be a fully optimized structure, can be used as starting model for a wide range of simulations.

When the final optimization is completed, MOPAC generates two new important text files

- 8GCH_F16(+)_G194(-)_F102(-).pdb
- 8GCH_F16(+)_G194(-)_F102(-).arc

which contain all the relevant information on the starting model, such as the structure and the heat of formation. The optimized structure for 8GCH_F16(+)_G194(-)_F102(-) is shown in Fig. 3.1.

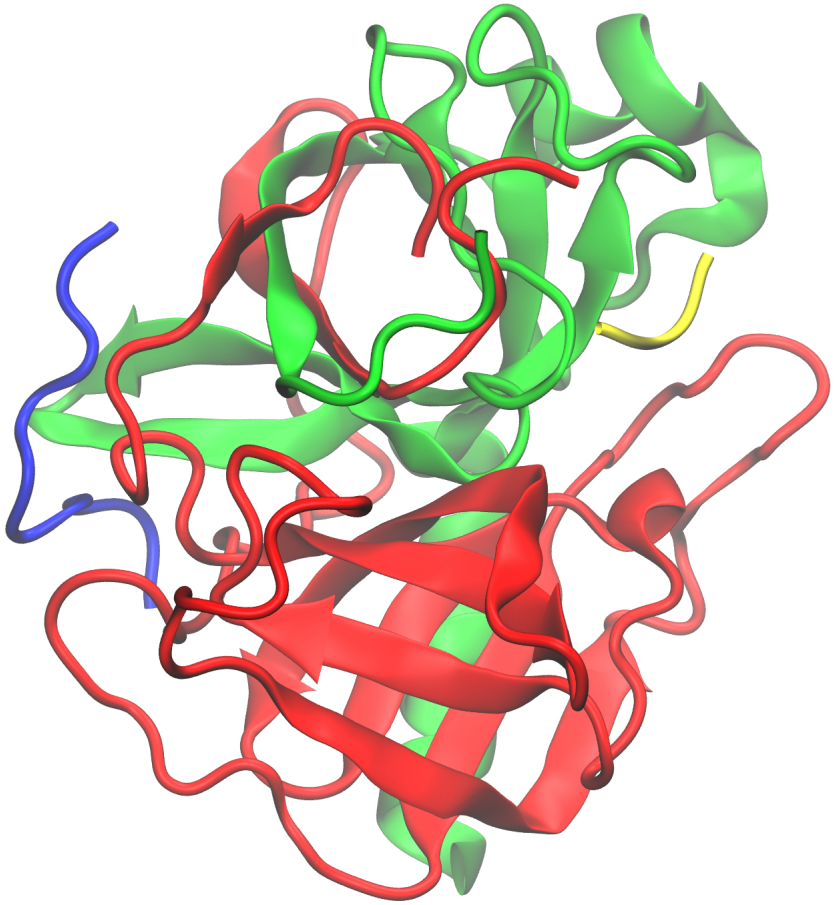

Figure 3.1 Optimized structure for 8GCH_F16(+)_G194(-)_F102(-). The E, F, G and C chains are colored in blue, red, green and yellow, respectively. This image was made with VMD [8] and Tachyon [9].

At this point it is interesting to compare the starting model, i.e. the fully optimized structure, with the hydrogenated X-Ray PDB structure, in which only the positions of the hydrogen atoms were optimized. It is important to notice that, in order for the comparison between two structures to be meaningful, they must have exactly the same empirical formula and the same charged sites, and their optimization procedures have to be performed using the same conditions. For the two structures compared here, COSMO was used to model solvation effects and the optimization criterion was set to GNORM=5.

An important means of comparison between the two structures is the difference between their heats of formation. The change in heat of formation

$(-5434.37\,\mathrm{kcal\,mol^{-1}})$ in going from the hydrogenated PDB X-Ray structure $(-42423.37\,\mathrm{kcal\,mol^{-1}})$ to the fully optimized structure $(-47857.74\,\mathrm{kcal\,mol^{-1}})$ shows that, as expected, the structure obtained with the unconstrained optimization is a better starting model for the γ-chymotrypsin. In fact, if the hydrogenated PDB X-Ray structure were used as starting model, the change in heat of formation due to the structure moving towards a lower energy electronic state would more than likely to invalidate the results of subsequent calculations.

Another means of comparison between the two structures is the root mean square difference (RMSD), which is the measure of the average distance between the atoms of the superimposed protein structures. The RMSDs were evaluated with VMD. The RMSD for the two complete structures, i.e. the protein with the hetero molecules, is 1.04 Å while if only the backbone is considered, the RMSD drops to 0.59 Å. This difference is mainly due to the motion of the water molecules. An image of the superimposed backbone structures is shown in Fig. 3.2.

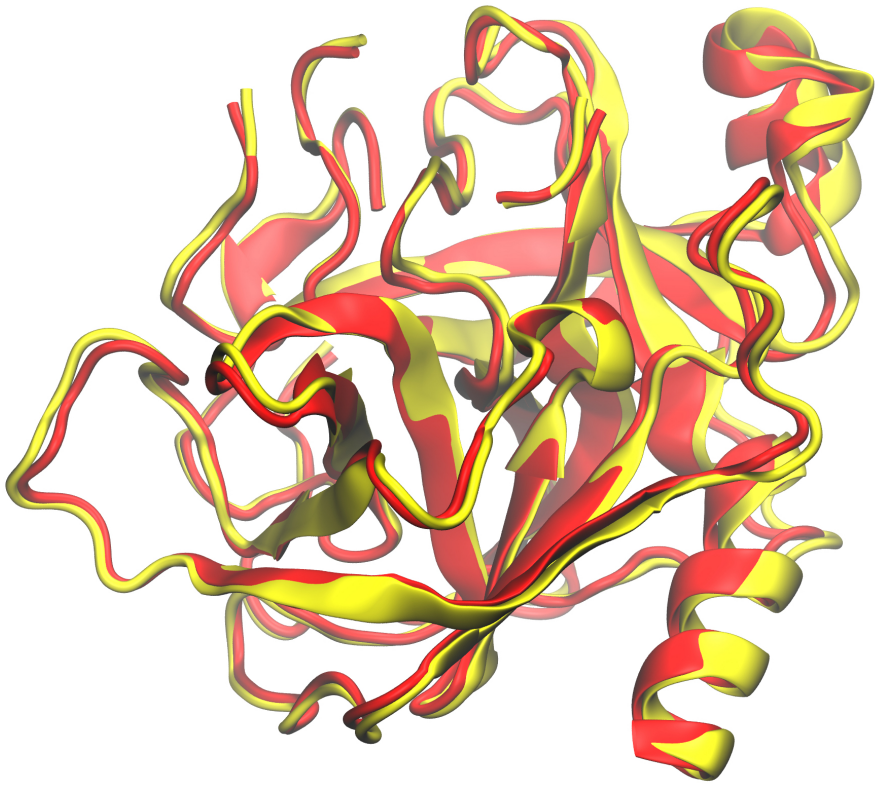

Figure 3.2 Structural superposition of the backbone structures for 8GCH_F16(+)_G194(-)_F102(-). The optimized structure and the hydrogenated PDB X-Ray structure are colored red and yellow, respectively. This image was made with VMD and Tachyon.

3.2.2.4 Testing for Salt Bridges

After the final optimization is completed, in order to generate a more realistic starting model, it is possible to investigate whether specific salt bridges exist or not. Even thought this problem might seem very hard to solve, as the ionization of the residues of a protein depends on several factors, such as the pH, the solvent and the chemistry of the protein, in practice the criterion for deciding if a salt bridge should exist is pretty straightforward: if the presence of a salt bridge lowers the heat of formation, the salt bridge is more than likely to exist.

For the purpose of illustrating the procedure, 8GCH_F16(+)_G194(-)_F102(-) can be tested for the existence of the salt bridge involving the residues Ser16(+) and Glu194(-). 8GCH_F16(+)_G194(-)_F102(-).pdb is used as starting point for this procedure. In order to ionize the residues involved in the salt bridge, three lines are added to 8GCH_F16(+)_G194(-)_F102(-).pdb: the first line contains the keywords PDBOUT SITE=(E11(+),F20(-)), while the second and the third lines can be either left blank or contain comments. The file is then saved as 8GCH_E11(+)_F20(-)_F102(-).mop and run with MOPAC.

An unconstrained optimization is then performed on 8GCH_E11(+)_F20(-)_F102(-). When the final optimization is completed, the heats of formation of 8GCH_F16(+)_G194(-)_F102(-) and 8GCH_E11(+)_F20(-)_F102(-) are compared. As the heat of formation of 8GCH_E11(+)_F20(-)_F102(-) ($-47825.79\,\mathrm{kcal\,mol^{-1}}$) is more positive than the heat of formation of 8GCH_F16(+)_G194(-)_F102(-), the salt bridge should not exist.

References

1. Stewart JJP. Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. Int J Quantum Chem. 1996;58(2):133-46. doi: 10.1002/(SICI)1097-461X(1996)58:2<133::AID-QUA2>3.0.CO;2-Z.

2. Scemama A, Renon N, Rapacioli M. A Sparse Self-Consistent Field Algorithm and Its Parallel Implementation: Application to Density-Functional-Based Tight Binding. J Chem Theory Comput. 2014;10(6):2344-54. doi:10.1021/ct500115v.

3. Stewart JJP, Császár P, Pulay P. Fast semiempirical calculations. J Comput Chem. 1982;3(2):227-8. doi:10.1002/jcc.540030214.

4. Stewart JJP. Stewart Computational Chemistry - MOPAC Home Page. [Online] Available from: http://openmopac.net/ [Accessed June 2014].

5. Research Collaboratory for Structural Bioinformatics, The San Diego Supercomputer Center. RCSB Protein Data Bank - RCSB PDB. [Online] Available from: http://www.rcsb.org/ [Accessed June 2014].

6. PDB ID: 8GCH. Harel M, Su CT, Frolow F, Silman I, Sussman JL. Gamma-chymotrypsin is a complex of alpha-chymotrypsin with its own autolysis products. Biochemistry. 1991;30(21):5217-25. doi:10.1021/bi00235a015.

7. Klamt A, Schüürmann G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. J Chem Soc Perkin Trans 2. 1993;1(5):799-805. doi:10.1039/P29930000799.

8. Theoretical and Computational Biophysics group, NIH Center for Macromolecular Modeling and Bioinformatics, at the Beckman Institute, University of Illinois at Urbana-Champaign. VMD - Visual Molecular Dynamics. [Online] Available from: http://www.ks.uiuc.edu/Research/vmd/ [Accessed June 2014].

9. Stone JE. Tachyon Parallel / Multiprocessor Ray Tracing System. [Online] Available from: jedi.ks.uiuc.edu/~johns/raytracer/ [Accessed June 2014].