Large-Scale Quantum Chemistry: A Practical Guide for Electronic Structure Calculations

Antonio M. Ferreira and Henry A. Kurtz

November 10, 2014

Contents

Preface

This is an attempt to start collecting the ideas for each chapter and what we may need. It serves as an organization of ideas mostly. It is also a good way for me to relearn to use TeX and the software again. I can use the practice.

I assume that best way to proceed will to make each chapter a different file and work on them separately. In fact on the publisher's web page of instructions they want each chapter as a separate file it looks like. We could divide them to do the first drafts. One start would be: Tony - 4,5; Henry - 2,3. We should talk a bit more about 1, 6, and 7. The sections are based on what was in the proposal and I expect they will change a lot as it goes.

Chapter 1

Why Bigger is Better

1.1 The role of modern computer architecture

1.2 The convergence of theory and experiment: the mesocopic world

Experiments are probing smaller but still a need for larger and larger theory to actually match.

1.3 Quantum effects in biology

Need a discussion of motivation. That computational biochemistry is largely non-electron and empirical. Can do better than that, particularly if interest in reactions not just structures.

1.4 Qualitative versus Quantitative Descriptions

1

Chapter 2

A Brief Review of Quantum Chemistry

The goal of theoretical chemistry has been to provide both correct and efficient methods for understanding and calculating the properties of chemical systems. An ideal goal is to provide the correct answer for any problem. This is obviously a work in progress but one with lots of progress. In this chapter the goal is to survey many methods for obtaining electronic properties and provide enough background to understand the approximations involved. There are many good textbooks and references that can provide the detailed derivations. [1–3]

A system composed of n electrons and N atoms is described by the time-independent electronic Hamiltonian,

$$H = -\frac{1}{2} \sum_i^n \nabla_i^2 - \sum_i^n \sum_\nu^N \frac{Z_\nu}{r_{i\nu}} + \sum_{i>j}^n \frac{1}{r_{ij}} \tag{2.1}$$

and is the Hamiltonian used within the scope of this book. We will use atomic units [4] throughout to simplify the expressions unless noted otherwise. The terms in Equation 2.1 contribute to the energy and are easily identified. The first term is the kinetic energy of the electrons, the second is the electrostatic attraction of the negative electrons with the positive nuclei, and the last term is the electrostatic repulsion of the electrons with each other.

We are focusing only on the electronic structure and not dynamics and have already simplified away the motion of the nuclei within the Born-Oppenheimer approximation [5]. Because the nulcei are at least 2000 times

heavier than an electron, this amounts to considering the electrons to move much faster and at any instance respond to a fixed set of nuclear coordinates. This removes the nuclear kinetic energy from the Hamiltoning for the electrons. To obtain the electornic energy of the system this term must be added to the electronic energy obtained by solving the electronic-Schrodinger equation:

$$H\Psi = E_{elec}\Psi \tag{2.2}$$

The remaining term non shown above is the nuclear-nuclear repulsions which for fixed nuclei is a constant term.

We have also neglected relativistic effects from this equation. Those may be important for some heavier elements and a discussion of how to include them will be given later.

The energy, $E_{Total} = E_{elec} + V_{NN}$, can be used to define a potential energy surface for the motion of the nuclei for optimization of structures, defining reaction paths, and even performing dynamics calculations. The wavefunction, Ψ, can be used to calculate various properties and to obtain the total electronic density via $|\Psi|^2$.

The earliest solutions encountered in most undergraduate physical chemistry or modern physics books are an exact solution to a non-relivistic hydrogen atom is shown. For chemical applications this example provides a way to introduce the nature and type of atomic orbitals, i.e, s, p, d ..., which provides the local symmetries of the bais functions for further refinements of multieletron systems. When the two-electron problem is next considered, the first approximations are introduced. One method of finding an approximate solution is to modifiy the Hamiltonian to remove the hard part such as dropping the electron-electron repulsion term. This give a simple independent particle approximation. While not computationally useful from numerical perspectives it gives insight into the requirements of multi-electron systems to meet the exclusion principle and provides a simple introduction into casting multielectron wavefunctions as determinants of single particle functions. For example the solution to the 2-electron problem with the electron repulsion removed is

$$H = -\frac{1}{2}\nabla_1^2 - \frac{2}{r_1} - \frac{1}{2}\nabla_2^2 - \frac{2}{r_2} \tag{2.3}$$

Because the Hamiltonian can be factored into separate terms for each electron with no interactions, the solution of the overall equation is a product of

the individual electron terms. Such a solution is the product of hydrogen-like orbitals with an increased nuclear charge. However, the Pauli Exclusion principle states that the wavefuntions must be anitsyemmetric under exchange of the labels for any two electrons —— add a lot more here about this ——

The way to ensure that the result satifyes the exclusiton principle is to take linear combinations of solutions with the same energy in the form

$$\Psi(1,2) = \begin{vmatrix} \psi_1(1) & \psi_2(1) \\ \psi_1(2) & \psi_2(2) \end{vmatrix} = (\psi_1(1)\psi_2(2) - \psi_1(2)\psi_2(1)) \qquad (2.4)$$

This representation of electronic wavefunctions in terms of determinantal expressions is the basis for many approaches beyond this simple independent particle model.

2.1 Hartree-Fock Theory

The variational principle states that any approximate wavefuction, Ψ_0 will provide an approximate energy, E_0 that is an upper bound to the true energy when used as an expectation value with the correct Hamiltonian. [6]

$$\frac{\int \Psi^* \widehat{H} \Psi d\tau}{\int \Psi^* \Psi d\tau} = E_0 \geq E \qquad (2.5)$$

If a wavefuction is given with parameters that can be varied then minimization of the energy with respect to those parameters give the closest energy possible to the exact energy of the system within the constraints of the wavefuction form. If this idea is applied to the choice of a the wavefuction being a single determinant

$$\Phi(1,2,...,n) = \begin{vmatrix} \psi_1(1) & \psi_2(1) & \ldots & \psi_n(1) \\ \psi_1(2) & \psi_2(2) & \ldots & \psi_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(n) & \psi_2(n) & \ldots & \psi_n(n) \end{vmatrix} \qquad (2.6)$$

and having the orbitals that form the determinant as the parameters that can be varied then the resulting description is known as the Hartree-Fock Approximation. The one-electron functions in the determinant that minimize the energy are solutions of the Hartree-Fock equation:

$$\widehat{F}_i \psi_i = \epsilon_i \psi_i \qquad (2.7)$$

with the Fock operator, $\widehat{F}$, given as:

$$\widehat{F}_i = -\frac{1}{2}\nabla_i^2 - \sum_{\nu}^{N} \frac{Z_{\nu}}{r_{i\nu}} + \sum_{j} \left(\widehat{J}_j - \widehat{K}_j \right) \tag{2.8}$$

A very good derivation of these equations can be found in the book by Szabo and Ostland. [1]

The solutions, ψ, are referred to as spin orbitals and there is one for each electron in the system. The operators shown as $\widehat{J}_j$ and $\widehat{K}_j$ are the coulomb and exchange operators and are given by:

$$\widehat{J}_j\psi_i(x_1) = \psi_i(x_1) \int \psi_j^*(x_2)\frac{1}{r_{12}}\psi_j(x_2)dx_2 \tag{2.9}$$

$$\widehat{K}_j\psi_i(x_1) = \psi_j(x_1) \int \psi_j^*(x_2)\frac{1}{r_{12}}\psi_i(x_2)dx_2 \tag{2.10}$$

Because the operators $\widehat{J}_J$ and $\widehat{K}_j$ contain the orbitals within the integrals, equation 2.7 must be solved in an iterative fashion by choosing a guess at the set of ψs and calculating new ones until a self-consistency is reached – the new orbitals are the same as the previously ones. Hence this procedure is also called the self-consistent field approach or SCF. The result is now another form of an independent particle approximation where one electron moves in the average field generated by the other N-1 electrons. This is the solution for an effective Hamiltonian given by

$$H_{eff} = \sum_{i} \widehat{F}_i \tag{2.11}$$

and the total electronic energy evaluated using the HF wavefuction with the correct Electronic Hamiltonian is

$$E = \int \Phi^* \widehat{H} \Phi d\tau$$

$$= \sum_i^n \int \psi_i(x)^* h \psi_i(x) dx$$

$$+ \frac{1}{2} \sum_i^n \sum_j^n \int \psi_i(x_1)^* \psi_j(x_2)^* \frac{1}{r_{ij}} \psi_i(x_1) \psi_j(x_2) dx_1 dx_2 \qquad (2.12)$$

$$- \frac{1}{2} \sum_i^n \sum_j^n \int \psi_i(x_1)^* \psi_j(x_2)^* \frac{1}{r_{ij}} \psi_j(x_1) \psi_i(x_2) dx_1 dx_2$$

where h is the one-electron part of the Fock operator with the kinetic energy and nuclear attractions operators.

Most large scale systems have a closed-shell singlet electronic structure. A further approximation is to require spin-paired spin orbitals to have the same spacial orbital part. For example $\psi_i = \phi_j \alpha$ and $\psi_i + 1 = \phi_j \beta$. Using this restriction and integrating over the spin part of the equations leads to the Restricted Hartree-Fock method (RHF) with the Fock operator written as:

$$E = \int \Phi_{RHF}^* \widehat{H} \Phi_{RHF} d\tau \qquad (2.13)$$

$$= \sum_i^{n/2} \int \phi_i(x)^* h \phi_i(x) dx$$

$$+ \frac{1}{2} \sum_i^{n/2} \sum_j^{n/2} 2 \int \phi_i(x_1)^* \phi_j(x_2)^* \frac{1}{r_{ij}} \phi_i(x_1) \phi_j(x_2) dx_1 dx_2 \qquad (2.14)$$

$$- \frac{1}{2} \sum_i^{n/2} \sum_j^{n/2} \int \phi_i(x_1)^* \phi_j(x_2)^* \frac{1}{r_{ij}} \phi_j(x_1) \phi_i(x_2) dx_1 dx_2$$

Introducing a common short hand representation for the integrals as,

$$\langle ij|kl \rangle = \int \phi_i(x_1)^* \phi_j(x_2)^* \frac{1}{r_{12}} \phi_k(x_1) \phi_l(x_2) dx_1 dx_2 \qquad (2.15)$$

this equation is written as

$$E = \sum_{i}^{n/2} \langle i|j \rangle + \sum_{i>j}^{n/2} 2 \langle ij|ij \rangle - \langle ij|ji \rangle \tag{2.16}$$

In the above equations the variables are the orbitals that are obtained by solving the Hartee-Fock equations

$$F\phi_i(x) = \epsilon_i \phi_i(x) \tag{2.17}$$

If the orbitals are expanded in terms of a fixed basis set $\{\chi_i\}$

$$\phi_i(x) = \sum_{j}^{M} \chi_j(x) c_{ji} \tag{2.18}$$

and this is inserted into equation 2.14, it gives the usual Hartree-Fock expressions

$$F_{ij} = \int \chi_i \widehat{F} \chi_j dx \tag{2.19}$$

$$F_{ij} = h_{ij} + 2 \sum_{\nu,l,m} c_{l\nu}^* c_{m\nu} \int \chi_i^* \chi_l^* \frac{1}{r_{12}} \chi_m \chi_j dx_1 dx_2$$

$$- \sum_{\nu,l,m} c_{m\nu}^* c_{m\nu} \int \chi_i^* \chi_l^* \frac{1}{r_{12}} \chi_j \chi_m dx_1 dx_2 \tag{2.20}$$

$$F_{ij} = h_{ij} + \sum_{\nu,l,m} c_{l\nu}^* c_{m\nu} \left(2 \langle il|mj \rangle - \langle il|jm \rangle\right) \tag{2.21}$$

$$F_{ij} = h_{ij} + \sum_{l,m} P_{lm} \left(\langle il|mj \rangle - \frac{1}{2} \langle il|jm \rangle \right) \tag{2.22}$$

where the density matrix, P, is defined as

$$P_{ij} = 2 \sum_{\nu} c_{i\nu}^* c_{j\nu} \tag{2.23}$$

These equations are solved iteratively by the following steps:

1. Choose initial guess for density matrix P

2. Calculate the Fock matrix, F

3. Solve euation xx to obtain orbials, C, and energies, ϵ

4. Form new density matrix, P, from orbitals

5. compare new density with previous step,

    (a) if changes greater than a chosen tolerance then use the new P to start again at step 2.

    (b) If density converged, then solution found

2.2 Basis Set Discussion

The first systematic use of basis sets was by Slater [**?**] to fit a set of functions to describe the atomic orbitals. These functions were of the form

$$\psi\left(r, \theta, \phi\right) = N_{nlm}\left(\zeta\right) r^{n-1} e^{-\zeta r} Y_{lm}\left(\theta, \phi\right) \tag{2.24}$$

where the $Y_{lm}\left(\theta, \phi\right)$ are the spherical harmonics giving the shape and correct angular momentum for each orbtial and $N_{nlm}\left(\zeta\right)$ is the appropriate normalization constant. The coeficient in the exponental, ζ, defines the radial extent of the funciton with larger values of ζ tigher, closer to the origin and smaller values of ζ more extended. Further refinements where added by splitting the functions to describe an orbital in to 2 or more parts with different exponents, threby giving greater ability to fit. These Slater-type-orbitals (STO) are rarely used in modern programs but the terminology carries on for "double zeta" basis sets and "'triple zeta"' basis sets, etc, to refer to the use of multiple basis functions for each atomic orbital.

Because of the difficulty in calculating the integrals needed for molecular systems, in particularly multicenter two-electron integrals, Boys [7] proposed the use of Gaussian-type orbitals (GTO). These are must commonly used as:

$$\gamma_i(x, y, z) = \left(x - x_A\right)^{k-i} \left(y - y_A\right)^{m_i} \left(z - z_A\right)^{l_i} e^{-\alpha_i |r - r_A|^2} \tag{2.25}$$

where the gaussian is centered at $\{x_A, y_A, z_A\}$, usually on an atom but not necessarily. This form replaces the "'spherical"' haromnic with a "'cartesian"' harmonic for easier integral evaluations.

There is still a single "s" function with $k + m + l = 0$ and three "p" functions with $k+m+l = 1$ but now there are 6 "d" functions from $k+l+m = 2$. These six functions include function combination $x + y + z$ which is an extra s-function. If someone used their own exponents or adds functions to standard supplied basis sets there is the possibility of linear dependency, which will be discussed below.

To provide better basis sets then just a single gaussian shape, several can be combined together in a fixed linear combination. These new functions are refered too as contracted gaussian basis functions,

$$\chi_i = \sum d_{ij} \gamma_j \tag{2.26}$$

Pople's influence (huzanaga?)
– polarization
Dunning terminology
Effective Core – below
Linear Dependency and how handled

2.3   Bottlenecks to HF

— Bottlenecks and needs to cover

- integral storage/direct methods

- integral accuracy and precision

- scaling with basis size – Ncube?

There are many parts to solving the Hartree-Fock equations that lead to difficulites as the size of the system grows. It is important to recognize this issues and also understand the way they are handled in modern quantum chemistry programs.

At first glance, the calculation of the 2-electron integrals seems to be the biggest hurdle. To calculate all possible integrals for a basis set of size N gives N^4 integrals. For a single water molecule with typcially xxx basis funcitons for a good calucaiton this is lkllj integrals but for a 10-basis pair DNA helix there are about yyyy basis functions needed in the simplest case and this would be $yyx10^{88}$ integrals. It is obvious way efficient ways to calculate these intergrals has been the area of a much research.

One problem is if someone would calculate all these integrals then how would you store them and retrieve them as needed? The answer to that is the implementation of "direct" method where the intergrals are never stored but calculated as needed.

As it turns out for Hartree-Fock caluclations now all the possible intergrals need to be calucalted either. Many of them are zero or too small to have an effect. This is usually checked before calculation of an integral but use of the Schwartz Inequatilyt

$$|\langle ij|kl \rangle| \leq \langle ii|kk \rangle^{1/2} \langle jj|ll \rangle^{1/2} \tag{2.27}$$

As the size of the system grows and primitive basis sets are farther apart then more and more of them are neglected and the N^4 scaling does not happen. The limited step becomes the matrix manipulations in sovling the Hartree-Fock equations and scales as N^3. This is still quite fast for very big systems and approzimations have been introduced to approach scaling as N, linear scaling. This approximations are discussed later in this chapter.

There are two important quantities that should be adjusted in large scale calculations: the adjaljfld and the ajdja;f

2.4 Electron Correlation and Post HF

2.4.1 Exact Answer

But there is a prescription for finding the exact solution to a mulitelectron problem. In 1955 Löwdin [8] shows such solutions can be expressed as sums of determinants built from a complete basis set. This has become know as the Full-CI method.

$$\Psi(1,2,3,..,n) = \sum_K C_K D_K(1,2,3,...,n) \tag{2.28}$$

where the D_K terms are all the possible determinants made up from all combinations of a complete basis set.

Of course there are an infinite number of these terms and this is not a practical way to get the exact answer but does show a method for approaching it. Eq. 2.28 is the basis for configuration interaction approaches where a finite number of terms are selected. If a fixed basis set is used and all determinants

generated by that basis are used then this is know as a FullCI calculation and the best result possible within the theory based on that basis set.

As discussed above the difference between the Hartree-Fock approximation and the exact solution of the electonic Schrodinger Equation is that in the former an electron only sees the average field of all the others, wereas in the later each electron's motion would be instantaneously correlated with the other individual electrons. The difference is energy between E_{HF} and E_{exact} is know as the correlation energy.

2.4.2 Configuration Interaction

There are many ways to organize and select terms to include in a CI expansion. The most common procedure used is to use the Hartree-Fock approximation as a starting point and with the set of orbitals obtained organize the other terms by differences from the Hartree-Fock determinant refreed to as excitations. For example

$$\Phi = \Phi_{HF} + \sum_{i,a} C_i^a \Phi_i^a + \sum_{i<j,a<b} C_{ij}^{ab} \Phi_{ij}^{ab} + \dots \tag{2.29}$$

this expansion would go up to n-electations. The coeffecnts $C_i^a, C_{ij}^{ab}, \dots$ are the linear variational prarmeters. Trucations after the third term is called CCSD, after the fourth CCSDT, etc. If only the first and third term are keep it is refreed to as the CCD method. The double excitaitons are the most important largest interatctions with the refrences Hartree-Fock state because it can be shown that the single exications do not interact directly becasue

$$\int \Phi_{HF}^* \widehat{H} \Phi_i^a d\tau = \left\langle \Phi_{HF} \left| \widehat{H} \right| \Phi_i^a \right\rangle = 0 \tag{2.30}$$

$$H = \begin{pmatrix} E_{HF} & 0 & H_{0D} & \dots \\ 0 & H_{SS} & H_{SD} & \dots \\ H_{D0} & H_{DS} & H_{DD} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{2.31}$$

MCSCF

Perturbation Theory

An alternative approach to variational methods like those discussed above is break the Hamiltonian into parts that can be solved exactly and an additional part that leads to the system of interest. This extra part is know as a perturbation and the approach is Perturbation Theory. The Hamiltonian is expressed as

$$\widehat{H} = \widehat{H}_0 + \widehat{V} \tag{2.32}$$

with the known solutions to the $\widehat{H}_0$ from

$$\widehat{H}_0 \Phi_i = E_i^{(0)} \Phi_i \tag{2.33}$$

following the outline in Shavitt and Bartlett [3], we can write the corrections as difference

$$\chi_n = \Psi_n - \Phi_n$$
$$\Delta E_n = E_n - E_n^{(0)} \tag{2.34}$$

lots of writing and figuring here

$$\Psi_n = \Phi_n + \lambda \Psi_n^{(1)} + \lambda^2 \Psi_n^{(2)} + \lambda^3 \Psi_n^{(3)} + \dots$$
$$E_n = E_n^{(0)} + \lambda E_n^{(1)} + \lambda^2 E_n^{(2)} + \lambda^3 E_n^{(3)} + \dots \tag{2.35}$$

Second-order engery is:

$$E_i^{(2)} = \sum_{j \neq i} \frac{\left| \int \Phi_i^* \widehat{V} \Phi_j d\tau \right|^2}{E_i^{(0)} - E_j^{(0)}} \tag{2.36}$$

IF the reference state is taken as the Hartree-Fock wavefuction then for a closed shell RHF wavefunction this equations becomes

$$E_i^{(2)} = 2 \sum_{abrs} \frac{\langle ab|rs \rangle \langle rs|ab \rangle}{\epsilon_a + \epsilon_b - \epsilon_r - \epsilon_s} - \sum_{abrs}^{N/2} \frac{\langle ab|rs \rangle \langle rs|ba \rangle}{\epsilon_a + \epsilon_b - \epsilon_r - \epsilon_s} \tag{2.37}$$

where the $\epsilon_i s$ are the orbital energies and the sums over a and b are occupied orbitals and r and s are unoccupied orbitals.

Coupled Cluster

In the coupled cluster approach the exact wavefuction is related to a refernece wavefucnion by an expoential operators of the form

$$\Psi = e^C \Phi_0 \tag{2.38}$$

where the Φ_0 is most often the Hartree-Fock single determinant wavefuction. The expoential operator is defined by the series expansion

$$e^C - 1 + C + \frac{1}{2!}C^2 + \frac{1}{3!}C^3 + \frac{1}{4!}C^4 + \dots \tag{2.39}$$

The terms in the operator are identified with the excitation level out of the reference state amd the operator C can be broken into various orders

$$C = C_1 + C_2 + C_3 + \dots \tag{2.40}$$

with each term corresponding to the number of electrons excited to the new configurations. For eample C_1 is the set of single excitations and C_2 is the set of double excitations and represented as

$$C_1 \Phi_0 = \sum_{ap} t_a^p \Phi_a^p$$

$$C_2 \Phi_0 = \sum_{a,b,p,q} t_{ab}^{pq} \Phi_{ab}^{pq} \tag{2.41}$$

The coefficients $t_a^p, t_{ab}^{pq}, \dots$ are refered to as amplitudes and are the paramters in this method.

Equations for t and be developed

If only the double excitation operator C_2 is keep the resulting approximation is referred to as CCD. If single and double excitaitons are used it is know as CCSD and if triples are used it is CCSDT, and so on.

2.5 Density Functional Theory

Kohn-Sham [9]

$$\left\{ -\frac{1}{2}\nabla_1^2 - \sum_\nu^N \frac{Z_\nu}{r_{1\nu}} + \int \frac{\rho(r_2)}{r_{12}} dr_2 + V_{XC}(r_1) \right\} \psi_i(r_1) = -\epsilon_i \psi_i(r_1) \tag{2.42}$$

2.6   Semiempirical Quantum Chemistry

Short discussion of popular schemes
     Mozeyme

2.7   Approaches to Large Molecules

2.7.1   Effective Core Potentials

Replace Core with effective potential
     Averaged relativistic effectoe potential:

$$U_l^{AREP} = r^{-2} \sum_i C_{li} r^{n_{li}} e^{-\zeta_{li} r^2} \tag{2.43}$$

Hay Wayt type [10–12], and SKJC type [13–15], Dolg overview [16]
Tom Cundari's chapter

2.7.2   Fragment-based Methods

There are several good refences on the Frangment Molecular Orbital Method (FMO) and this section is based on them. [17–19]
     One simple way to make a large problem more tractable is to break it up into smaller pieces, or fragments. The whole can then be obtained by adding all the pieces back together. For example, the energy would be obtained via

$$E^{FMO} = \sum_i^N E_i \tag{2.44}$$

For non-interacting molecules or fragments this would be the right answer but not for interacting systems or, particularly for molecular systems broken into fragments across bonds. The simplest FMO method uses the energy decomposition idea above and recognizes that exchange interactions are fairly local and that the dominate interaction is due to the Coulomb fields from the other fragments. This basic method requires each E_i is a calculation of fragment i in the presence of the electrostatic fields from all the other fragments. Practically that requires forming an initial guess for the density of each fragment and calculating the electrostatic potentials then solving for new densities in a self-consistent procedure.

In general, the energies of a fragment X is obtained from the Hamiltonian given as

$$\widehat{H}_{ij}^X = H_{ij}^X + V_{ij}^X + P_{ij}^X \tag{2.45}$$

where, V_{ij}^X is the electrostatic potential term and P is the projection operator.

$$V_{ij}^X = \sum_{K \neq X}^{N} \left\{ \sum_{A \in K} \left\langle i \left| -\frac{Z_A}{|r - R_A|} \right| j \right\rangle + \sum_{kl \in K} D_{kl}^K (ij|kl) \right\} \tag{2.46}$$

$$P_{ij}^X = B \sum_{k \notin X} \left\langle i | \phi_i^k \right\rangle \left\langle \phi_i^k | j \right\rangle \tag{2.47}$$

includes electrostatics from other fragments solved self consistently
Improvements: 1) fragments size 2) Two and three-body interactions
Need for accuracy improvements
Fragment formations – bond breaking methods
Two-body FMO expansion

$$E^{FMO2} = \sum_i^N E_i + \sum_{i>j}^N (E_{ij} - E_i - E_j) \tag{2.48}$$

$$= \sum_{i>j}^N E_{ij} - (N_f - 2) \sum_i E_i \tag{2.49}$$

or a three-body expansion

$$E^{FMO3} = \sum_i E_i + \sum_{i>j} (E_{ij} - E_i - E_j)$$

$$- \sum_{i>j>k} ([E_{ijk} - E_i - E_j] - [E_{ij} - E_i - E_j] - [E_{ik} - E_i - E_k] - [E_{jk} - E_j - E_k])$$

$$= \sum_{i>j>k} E_{ijk} - (N_f - 3) \sum_{i>j} E_{ij} + \frac{(N_f - 2)(N_f - 3)}{2} \sum_i E_i \tag{2.50}$$

2.7.3 Divide and Conquer

2.7.4 Linear-Scaling Methods

2.7.5 The ONIOM method

$$E^{high}(R) \approx E^{low}(SM) + \left\{ E^{low}(R) - E^{low}(SM) \right\}$$
$$+ \left\{ E^{high}(SM) - E^{low}(SM) \right\} \qquad (2.51)$$

$$E^{ONIOM} = E^{low}(R) - E^{high}(SM) - E^{low}(SM) \qquad (2.52)$$

2.7.6 Local Orbitals versus Canonical Orbitals

2.7.7 NBO Analysis

2.8 Something about Molecular Mechanics?

2.8.1 Typical Force Fields and Programs

2.8.2 Polarizable Force Fields

Chapter 3

A Primer for Condensed Matter Physics

Chapter 4

Electronic Structure for Nucleic Acid Systems

Having established the foundations of quantum chemistry for condensed matter systems, we now turn our attention to nucleic acids. Modern molecular biology has focused a great deal of attention toward understanding the structure and function of this class of molecule. The very first quantum chemical calculations on DNA [20] by Ladik and Hoffman date back to 1964 and were focused primarily on semiempirical treatments of the band structure.

4.1 Start Small: Early Studies of DNA Bases

An impressive body of work concerning small numbers of DNA bases has come from the collaboration of Jiří Šponer, Pavel Hobza, and Jerzy Leszczynski. [21–28]

Our examination of larger systems relies on the fundamental description of the electronic structure of smaller

Choice of basis set has a tremendous influence on the quality of the results as illustrated by Figure 4.1. The two density of states plots were generated from the same d(CGTAAGGGTTACA) canonical double-helix with counterions using the B3LYP density functional, with only the basis set differing between the two calculations. The most obvious defference between these two calculations is the distribution of states around the HOMO–LUMO gap. In the minimal basis calculation, this gap is nearly 3 eV, whereas using the ECP basis set produces a gap of less than 1 eV. Furthermore, the energy
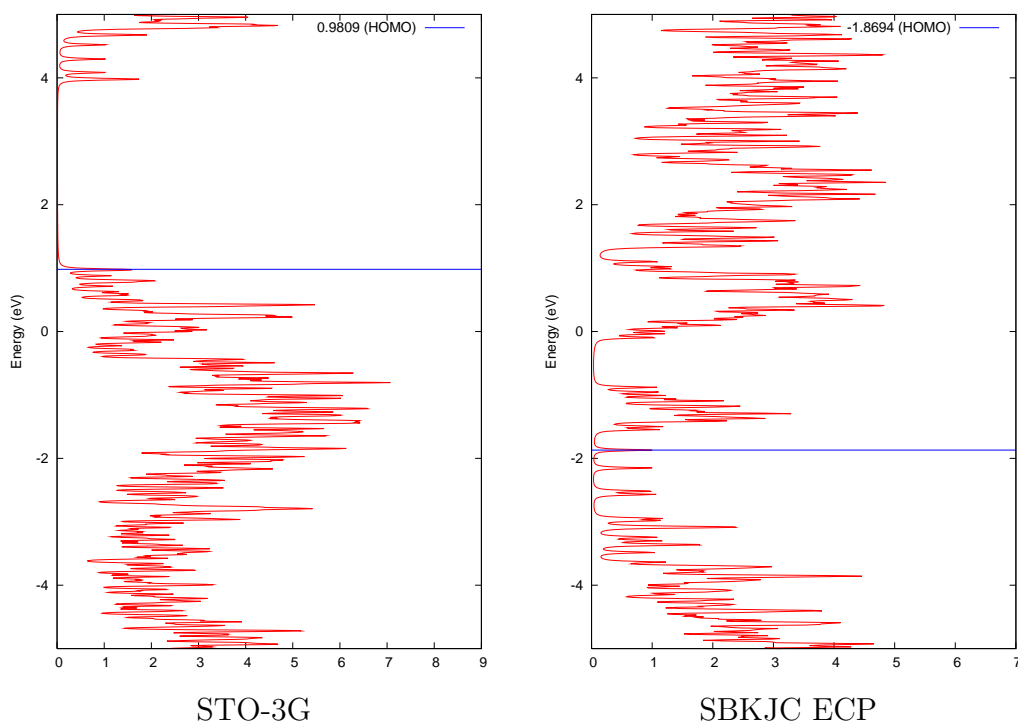
Figure 4.1: Comparison of the density of states plots for the sequence CG-TAAGGGTTACA generated using the B3LYP density functional.

of the HOMO in the STO-3G calculation is positive, suggesting that the HOMO is an unbound electron state.

A closer examination of the DOS plot for the B3LYP/SBKJC calculation reveals some interesting and important features. A collection of states creates a peak between -2 and -1 eV just above the HOMO. This collection of empty states represents the unoccupied s orbitals on the Na counterions. They are bound states and there is a small gap between the HOMO of the oligonucleotide and the lower edge of this unoccupied band. Another important feature is the small collection of unoccupied bound states just below 0 eV. They are part of a much larger set of states with a peak near 0.5 eV, but their physical importance is tied to their negative energies.

As was discussed in the theoretical background, a semiconductor system contains a set of unoccupied states with negative energy with the Fermi

gap separating the band of occupied and unoccupied states. What we see emerging in the calculations is the electronic structure of a semiconductor system and the failure of the minimal basis set treatment to describe this represents a serious shortcoming. While the ECP basis set neglects a direct treatment of the core orbitals, it does describe the important features of the frontier molecular orbitals with only a double-ζ quality basis set.

4.2 Large-scale Calculations of Electronic Structure

Ultimately, the goal of any computational technique is to provide insight into experiments that take place in the laboratory. Biological system present an array of modelling challenges with scale and complexity being foremost among them. Molecules important to biologists function in living systems and are therefore subject to an extraordinarily complex network of interactions. Properly describing the environmental conditions in which these molecular species function is often difficult, if not impossible.

4.2.1 Convergence Issues

Because nucleic acid systems carry significant negative charge in the phosphate backbone, the treatment of increasing numbers of nucleobases requires the solution of Hamiltonians associated with increasing diffuse orbitals.

In practice, this means that systems containing more than eight base pairs are impossible to converge to a stable wavefunction. Convergence of the SCF procedure can be achieved for sequences of up to eight base pairs by treating only the DNA atoms with an appropriate total charge derived from the number of phosphate groups in the backbone. This is possible because the excess negative charge resides in the phosphate oxygen atoms rather than throughout the molecule. Unfortunately this phenomenon, which allows the treatment of medium-sized systems, causes problems when going beyond eight base pairs. Mixing of the valence shell AOs on the exocyclic oxygen atoms occurs with larger systems and hinders SCF convergence to the point of failure, even when given very large ($\dot{c}$ 200) numbers of cycles in which to perform the SCF procedure. An examination of such calculations reveals that the mixing of atomic orbitals associated with the exocyclic oxygen atoms leads to eigenvectors with nearly degenerate energies.

An obvious solution to this problem is to introduce point charges neart these phosphate O atoms to localize the electron density and ameliorate the degeneracy problem. The challenge is to avoid perturbing the electronic structure of the dsDNA system in the process. A series of calculations can be performed using dsDNA, in which Na cations are placed at varying distances along the vector bisecting the O-P-O bond of the exocyclic oxygen atoms (See Fig. 4.6). Fitting the data points to a Morse potential identifies the point at which there is the strongest interaction between the dsDNA and the Na cations. For the B3LYP calculations, this distance works out to be 2.37 Å. While all these calculations converge in the SCF procedure, the challenge is to balance convergence against interaction and total time for the computation. Larger distances converge more slowly as the eigensolver deals with the nearly degenerate oxygen AOs. We have found the best balance to be at a distance of about 3.4 Å. This produces reliable convergence without interacting too strongly with the dsDNA of interest. This is borne out in the density of states plots shown below.

4.2.2 Linear Scaling Approximations

Linear scaling approximations have become a popular method for treating systems with many hundreds to thousands of atoms quantum mechanically. The two most popular implementations are MOZYME method in Jimmy Stewart's MOPAC2009 package (and previous MOPAC versions from Fujitsu) and the divide and conquer approach as implemented in DivCon, which was developed by Kenneth Merz and Lance Westerhoff. Additionally, the LoaclSCF approach from Victor Anisimov has gained popularity in recent years. As the theory behind these approximations has already been described, we turn our attention to an assessment of the results obtained from these approaches.

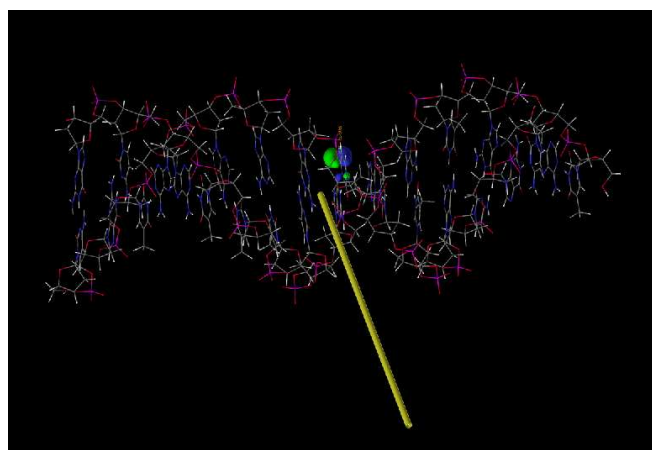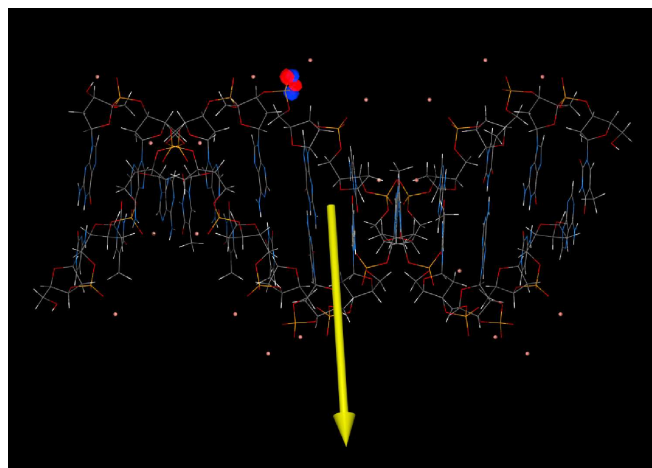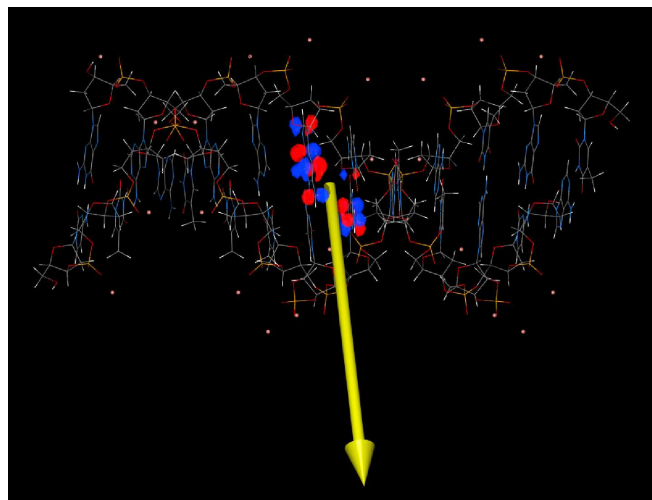## 4.2.3 Fragment Based Treatments

[WAITING FOR CALCULATIONS]

Table 4.1:

4.3 Case Study: Ecteinascidin 743

Having dicussed the considerations necessary for a proper treatment of large nucleic acid systems, we are now in a position to examine a particular problem for which electronic structure is of critical importance. Namely, the chemical reaction between dsDNA and the natural product Ecteinascidin 743 (Et-743). Et-743 is an isolate from sea squirts and has been prepared by *de novo* synthesis by Corey and coworkers. This compound is a DNA methylator, which forms a bond with the N2 carbon of Guanine bases via minor groove intercalation.

Pommier [29] studied the relative kinetics of this reaction among the sixteen possible sequences in which the central three base pairs were systematically permuted. It is somewhat remarkable that minor permutations in the sequence can lead to drastically different reaction kinetics. The explanation for this has been a theory called *direct readout.* [30] Under the hypothesis of direct readout an approching chemical enitity is influenced by changes in electrostatic potential at the exterior of the double helix caused by variations in the nucleobases at the interior. The varying electrostatic interactions allow for favorable or unfavorable interactions with the minor groove leading to differential intercalation (and thereby methylation) energetics.

Wether or not subtle sequence variations are sufficient to cause the requisite charge variation along the phosphate backbone is an important question that can be addressed using electronic structure calculations. As discussed previously, there are several models for computing atomic charges withing the Hartree-Fock framework. While the strengths and shortcomings of each are well known, a qualitative description is well withing the capabilities of nearly all of the atomic charge models. Here we will examine Mulliken charges calculated using the B3LYP/SBKJC methodology discussed earlier.

For each of the sixteen sequences used in the Pommier experiments, a B3LYP/SBKJC single-point energy calculation was performed. Thus, we have a set of molecular orbitals and a set of atomic charges from which to work.

4.3.1 How Big is Big Enough

Transcription of information coded in DNA to proteins is a fundamental process in biological systems. Codons represent the smallest distinct unit of information relevant to this process. Therefore, it is reasonable to view a

codon (three base pairs) as the fundamental unit that must be accurately described in a treatment of the electronic structure of DNA. We wish, then, to answer the following question: how do we accurately treat the electronic structure of a codon-size piece of DNA? Similarly, what is the smallest ds-DNA system required to produce a self-consistent description of the orbitals of a codon?

Developing a practical answer to this question requires an understanding of what the orbital structure of a three base pair segment looks like in the context of increasingly extended treatments with respect to the length of the surrounding sequence.

Accurate calculations of orbital energies are challenging even on small test systems of dsDNA when counterions are not present. In fact, we can see from the results in Table 4.2 that density functional theory fails to predict appropriate orbital energies even when given a large basis set like the polarized Ahlrichs basis set.**??** Fortunately, removing the core orbitals from the DFT treatment results in a dramatic improvement of the orbital energies but fails to improve the difference in the frontier molecular orbitals. The effective core potential approach combined with the B3LYP functional seems to give a fortuitous cancellation of errors and suggests a practical method we may use to treat larger systems. .

4.3.2 It's all about the orbitals

In order to examine the question about charge variation along the DNA backbone, we calculated the Mulliken charges for ds(CTAA**XXX**TTACA), where **XXX** = TGT, TGG, and AGC, using B3LYP/SBKJC. Results for the P atoms and the Na counterions are reported in Table 4.3. We have chosen to examine the charge density on the P atoms for two reasons. Firstly, they are directly bonded to the atoms that should show the greatest charge variation along the backbone. Namely the exocyclic oxygen atoms, which are highly polarizable because of the *d*-bonding present in phosphates. If there is charge variation along the backbone these oxygens should be perturbed by virtue of their polarizable electron density. Secondly, if the bases interior to the double helix are responsible for creating charge variation along the backbone, they will have to do so (at least in part) through an inductive effect. Any perturbation of the backbone charges must effect the P atoms either by perturbing the charge in the ribose rings or through direct inductive effects exhibited through bond polarization. Examining the charge on the Na cations is done

NOTE: We need to put the orbital energies obtained **with** the counterions into this table! Q: Should we report the average charge on the exocyclic O atoms as well?

Orbital	B3LYP/ 3-21G	B3LYP/ 6-31G(d)	B3LYP/ cc-pVDZ	HF/ cc-pVDZ	MP2/ cc-pVDZ	B3LYP/ cc-pVTZ	B3LYP/ p-Ahlrichs	B3LYP/ SBKJC
LUMO + 5	0.2144	0.2087	0.2083	0.3550	0.3550	0.1981	0.2032	0.1937
LUMO + 4	0.2123	0.2042	0.2031	0.3478	0.3478	0.1894	0.1988	0.1884
LUMO + 3	0.2021	0.1917	0.1903	0.3319	0.3319	0.1780	0.1855	0.1774
LUMO + 2	0.1898	0.1849	0.1839	0.3273	0.3273	0.1730	0.1793	0.1712
LUMO + 1	0.1874	0.1809	0.1807	0.3176	0.3176	0.1691	0.1752	0.1652
LUMO	0.1857	0.1776	0.1766	0.3163	0.3162	0.1655	0.1716	0.1641
HOMO	0.0471	0.0407	0.0415	-0.0465	-0.0465	0.0289	0.0357	0.0222
HOMO − 1	0.0348	0.0126	0.0154	-0.0763	-0.0763	0.0019	0.0115	-0.0038
HOMO − 2	0.0318	0.0123	0.0128	-0.0925	-0.0925	-0.0052	0.0075	-0.0139
HOMO − 3	0.0247	0.0057	0.0077	-0.1074	-0.1074	-0.0100	0.0033	-0.0166
HOMO − 4	0.0243	0.0040	0.0062	-0.1182	-0.1182	-0.0111	0.0020	-0.0177
HOMO − 5	0.0223	0.0027	0.0057	-0.1240	-0.1240	-0.0124	0.0015	-0.0191
$\varepsilon_{\text{LUMO}} - \varepsilon_{\text{HOMO}}$	0.1386	0.1369	0.1351	0.3628	0.3627	0.1366	0.1359	0.1419

Table 4.2: Comparison of the calculated orbital energies for the d(AGC) system using various basis sets. It is interesting to note that the density functional theory calculations consistently underestimate the HOMO–LUMO gap and that only the Hartree–Fock methods predict that all occupied states are bound.

	CTAA<u>TGT</u>TTACA		CTAA<u>TGG</u>TTACA		CTAA<u>AGC</u>TTACA	
	P	Na	P	Na	P	Na
1	0.4510	0.8589	0.4510	0.8591	0.4544	0.8589
2	0.4507	0.8567	0.4507	0.8570	0.4517	0.8550
3	0.4507	0.8556	0.4508	0.8561	0.4357	0.8562
4	0.4540	0.8559	0.4538	0.8566	0.4510	0.8542
5	0.4522	0.8561	0.4489	0.8544	0.4485	0.8520
6	0.4479	0.8528	0.4467	0.8518	0.4468	0.8557
7	0.4503	0.8529	0.4523	0.8518	0.4512	0.8564
8	0.4504	0.8530	0.4507	0.8523	0.4518	0.8566
9	0.4498	0.8550	0.4498	0.8544	0.4549	0.8555
10	0.4552	0.8574	0.4551	0.8569	0.4538	0.8547
11	0.4572	0.8611	0.4572	0.8609	0.4483	0.8568
Average	0.4518	0.8559	0.4515	0.8556	0.4515	0.8556
Std. Dev.	0.0027	0.0026	0.0029	0.0030	0.0027	0.0017
Range	0.0093	0.0083	0.0105	0.0091	0.0080	0.0069

Table 4.3: Comparison of the charge variation along the backbone of three distinct dsDNA sequences. The central three bases correspond to sequences with poor, moderate, and good alkylation as observed in the experiments of Pommier. Interestingly, the sequence with the most efficient alkylation shows the lowest charge variation among the backbone phosphorous atoms and the sodium counterions.

primarily as a reference. These cations are closed-shell species and do not interact with the orbitals of the dsDNA directly, an assertion that is easily verified by examining which orbitals contain non-zero AO coefficients from the Na orbitals. There should be no appreciable charge variation for these cations and their interaction with any electron in the nucleic acid system will be purely electrostatic.

These calculations show remarkably little charge variation among the P atoms, with varations measured in the hundreths to thousandsth of a charge unit. Comparing the charge variation to that of the reference cations suggests that there is actually no appreciable charge variation along the phosphate backbone. In order for the hypothesis of direct readout to work, charge variations of at most $0.105\,e$ would have to influence the electrostatic interactions between the intercalating ligand and the dsDNA substrate. This seems highly unlikely given the solvent dielectric constant screening the backbone charge from any approaching ligand. In the absence of an appreciable charge variation along the backbone to explain the sequence specificity of Et-743 as an intercalating species, we turn to electronic structure theory for a possible explanation.
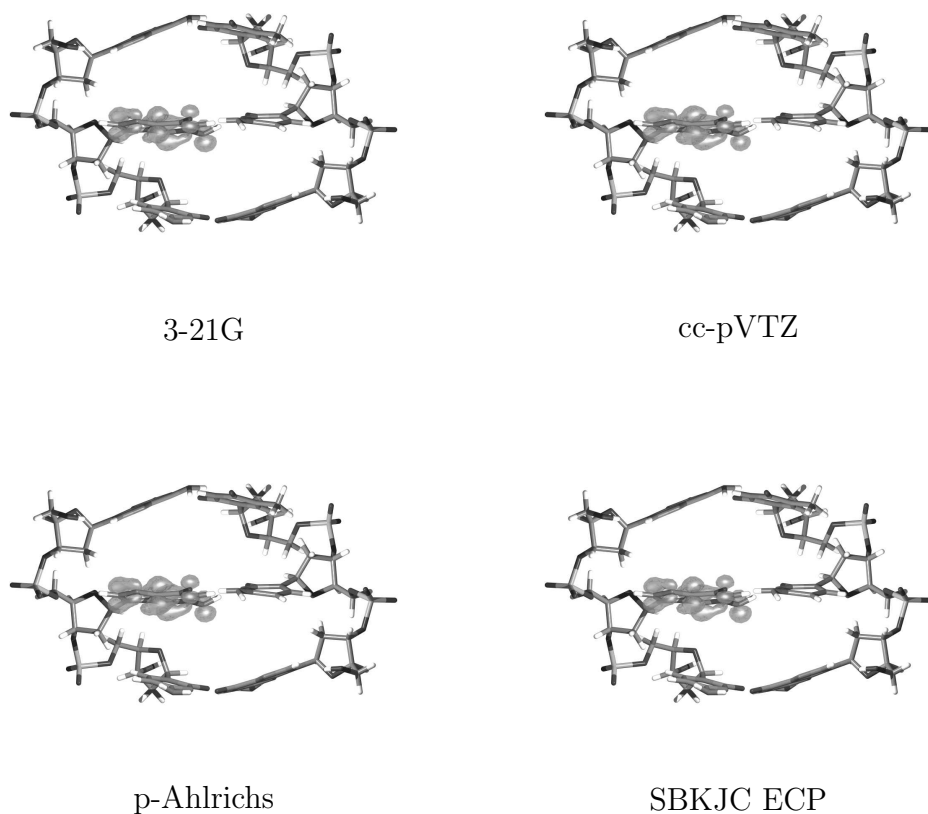
Figure 4.2: Comparison of the highest occupied molecular orbital (HOMO) for the three base pair dsDNA (AGC) using four different AO basis sets. It is imporatnt to note that the qualitative description of the HOMO differs significantly for the 3-21G basis set versus the other three basis sets. However, the double-ζ quality effective core potential gives a qualitative description nearly identical to that of the much higher-quality polarized basis sets.
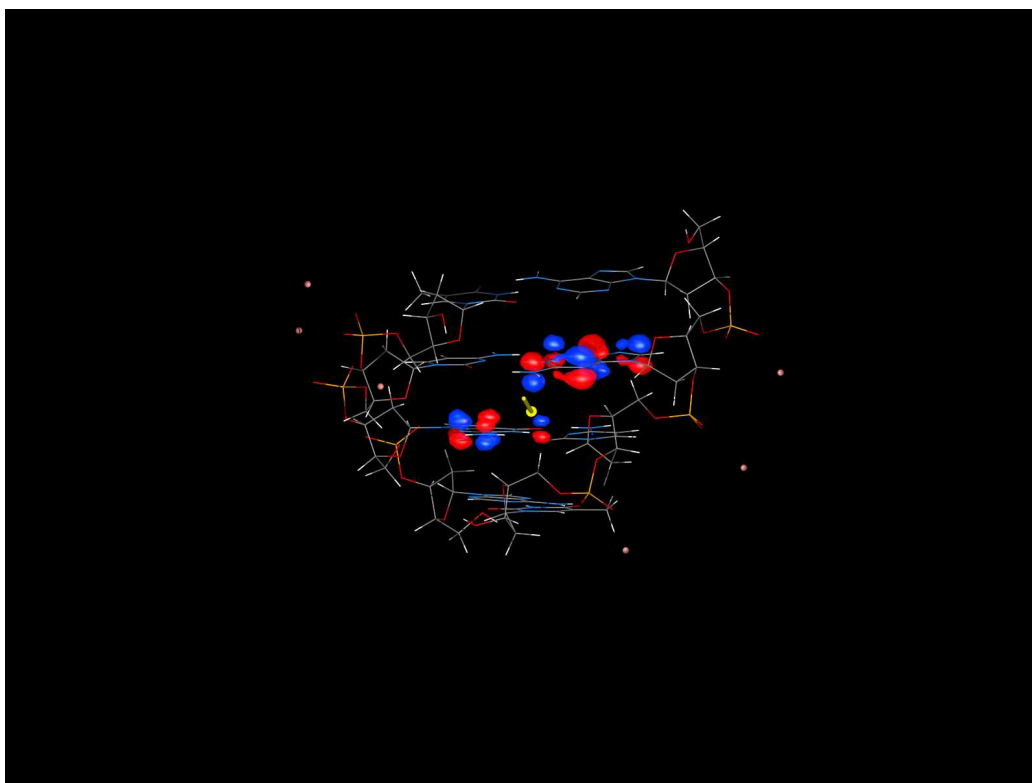
Figure 4.3: Orbital diagram for the HOMO of d(AGCT) obtained from B3LYP/SBKJC.
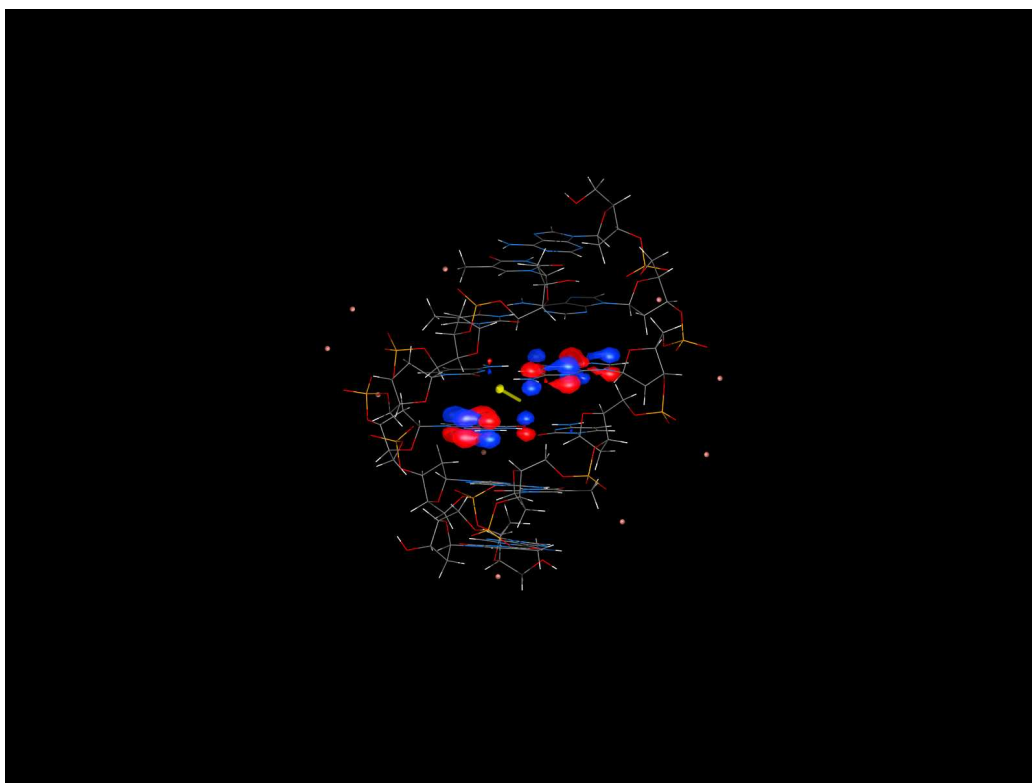
Figure 4.4: Orbital diagram for the HOMO of d(AAGCTT) obtained from B3LYP/SBKJC.

Cation Interaction Energy
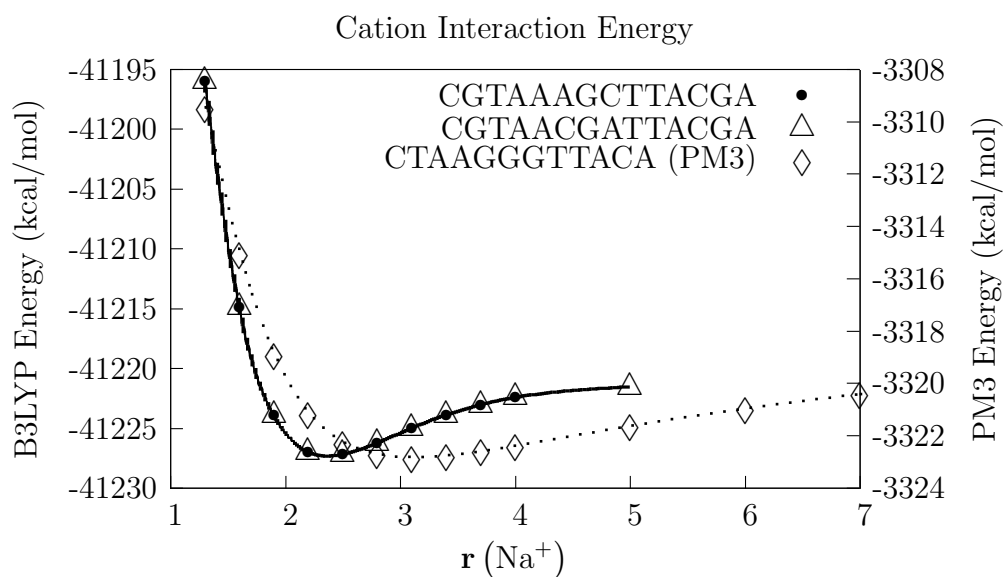
Figure 4.5: The cation interaction energies for the two sequences calculated at the B3LYP/STO-3G level of theory are essentially identical, while the PM3 result shows the lack of dispersion treatment inherent in the semiempirical treatment.
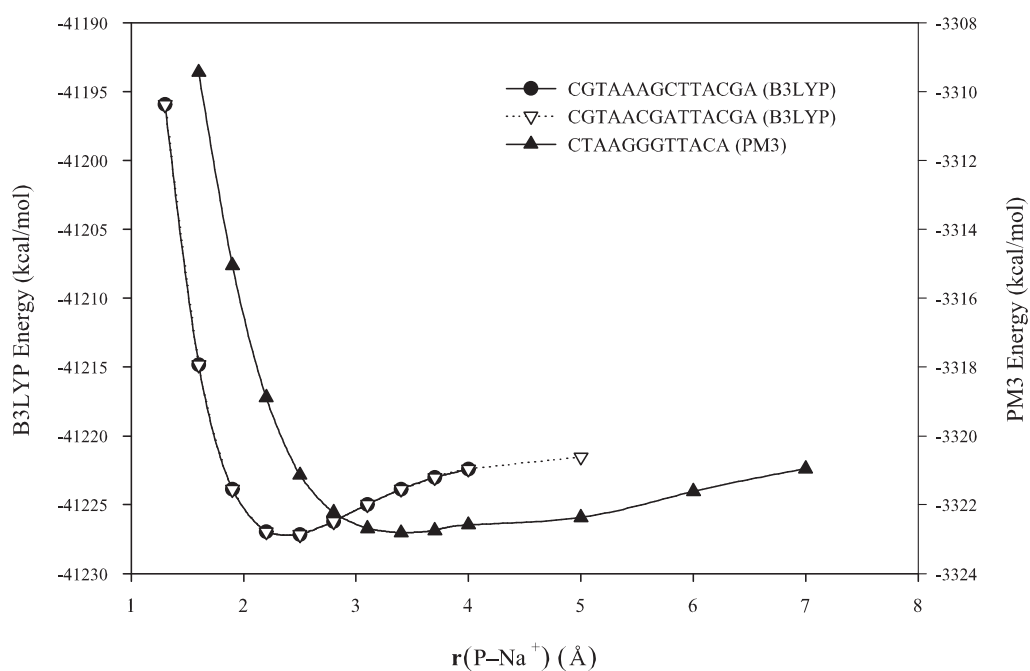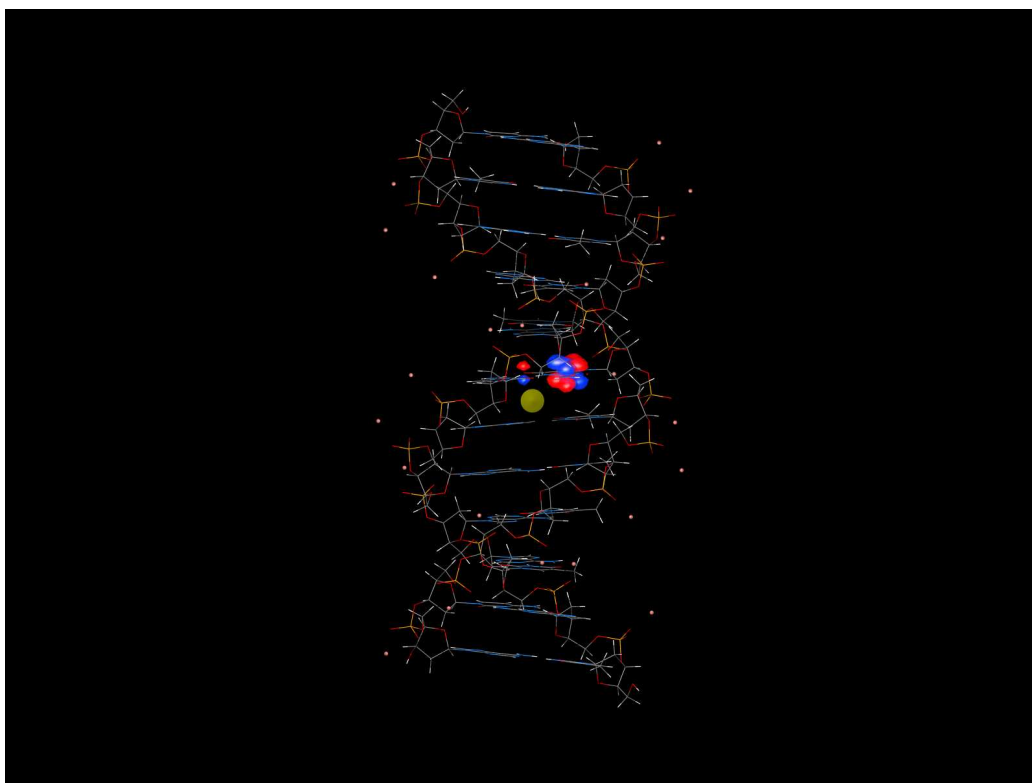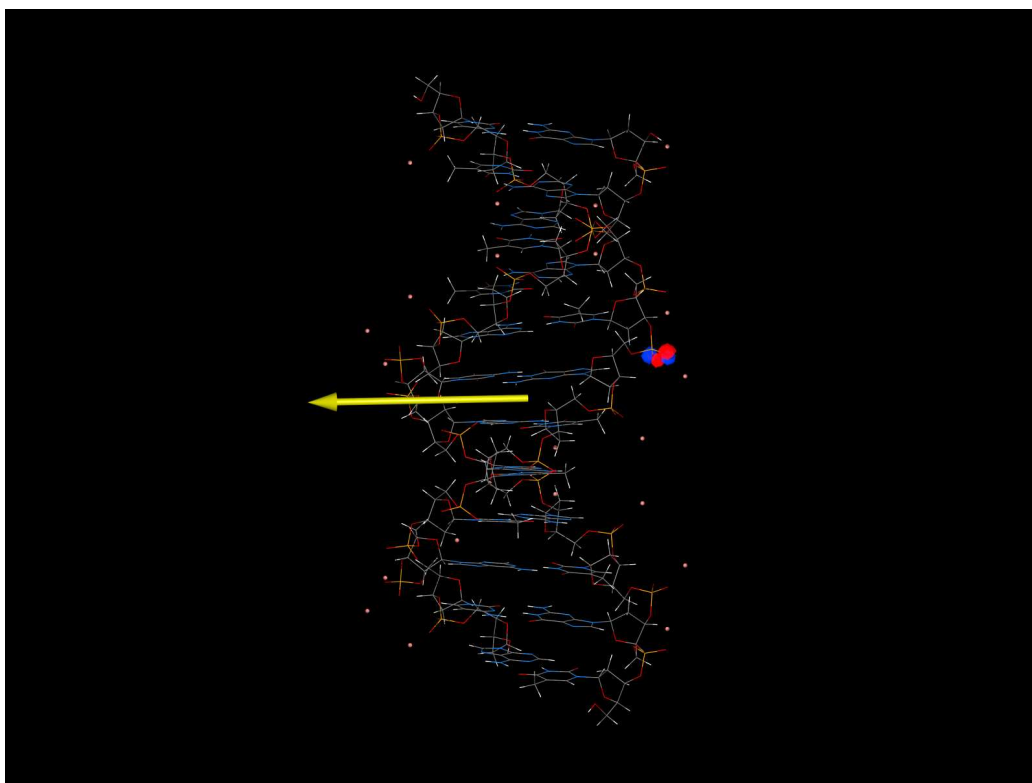
Figure 4.6: The cation interaction energies for the two sequences calculated at the B3LYP/STO-3G level of theory are essentially identical, while the PM3 result shows the lack of dispersion treatment inherent in the semiempirical treatment.
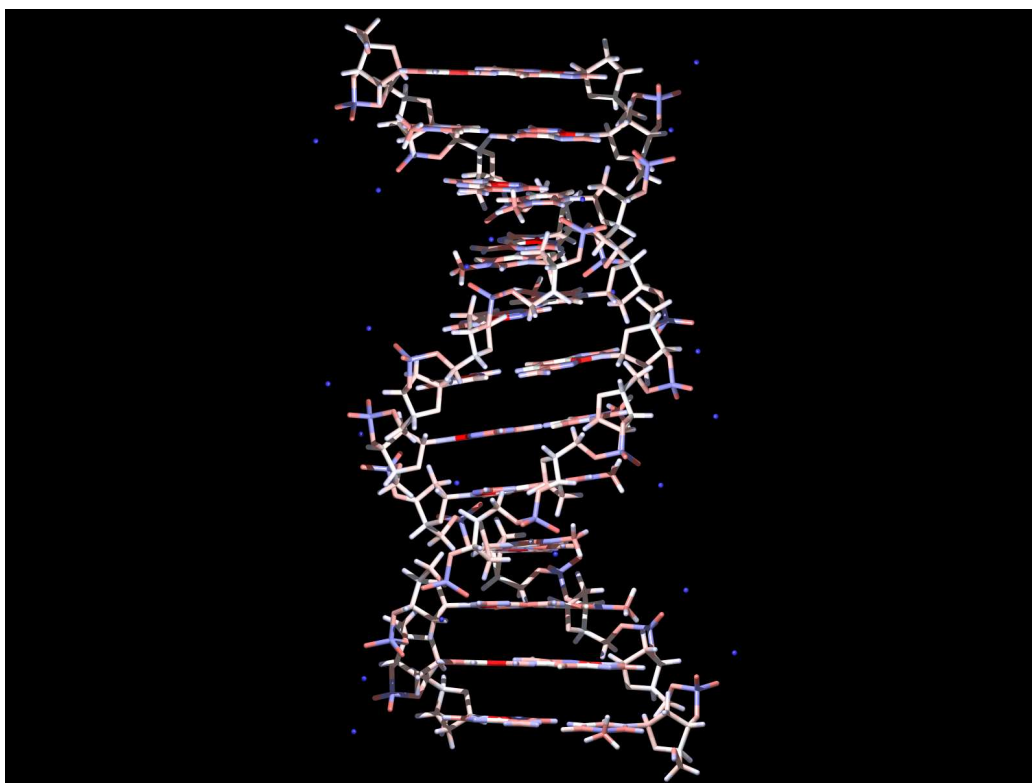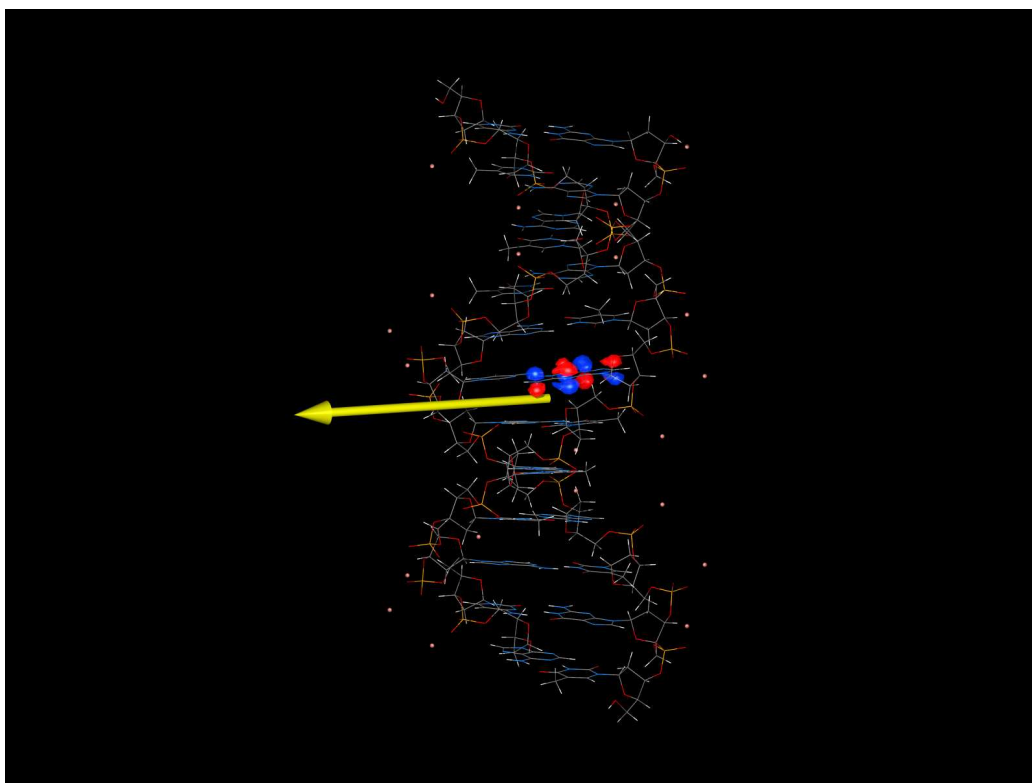
Figure 4.7:
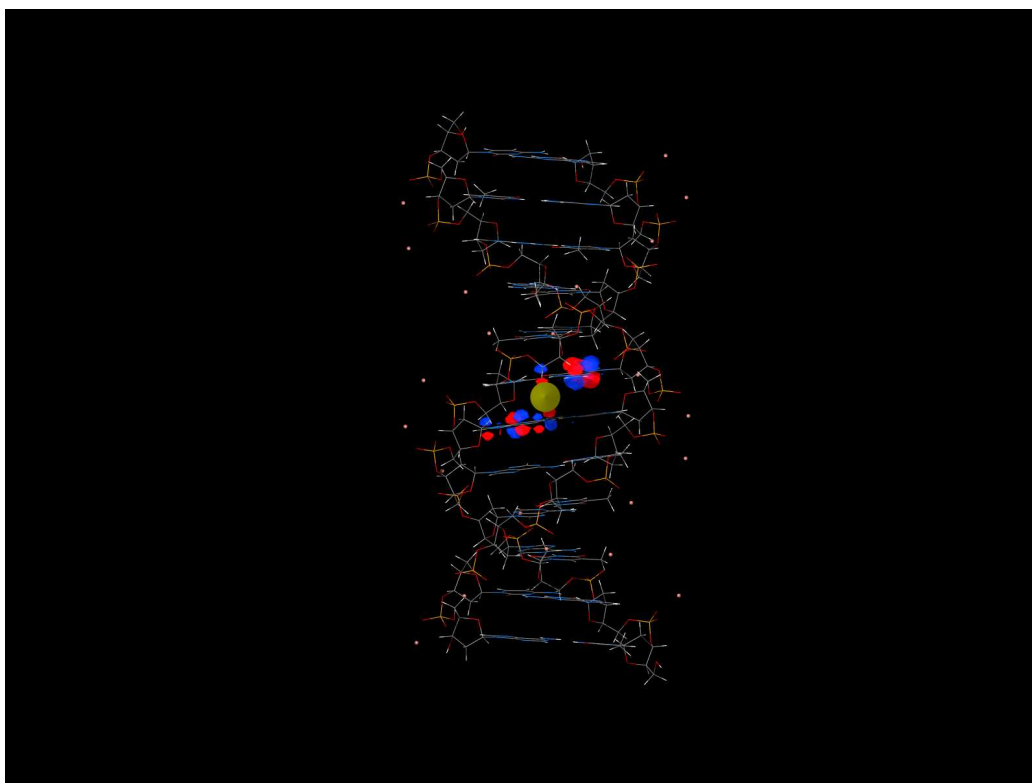
Figure 4.8:

Figure 4.9:

Figure 4.10:

Figure 4.11:

Chapter 5

Electronic Structure for Amino Acid Systems

5.1 Why Proteins Differ from DNA: An Electronic Structure Perspective

5.1.1 Nitrogen as an insulator?

5.2 Taking a Big Bite

5.3 Convergence Characteristics

5.4 Comparison of Linear Scaling and FMO Methods

5.5 Case Study: Dihydopteroate Synthase

5.5.1 When do you really need all the protein

5.5.2 Orbital Steering and Quantum Chemistry

Need List of calculations done that will be included and discussed.

Chapter 6

Electronic Structure for Semiconductor Systems

6.1 Periodic Treatments

6.2 Can we get Band Stricture Directly?

6.3 What do we lose in the approximations

6.4 Case Study

Options:
    another system

Chapter 7

General Considerations

Bibliography

[1] Attila Szabo and Neil S. Ostland. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. McGraw-Hill, 1989.

[2] Peter Atkins and Ronald Friedman. *Molecular Quantum Mechanics, 4th Edition*. Oxford University Press, 2005.

[3] I Shavitt and Rodney J. Bartlett. *Many-Body Methods in Chemistry adn Physics MBPT and Coupled-Cluster Theory*. Cambridge University Press, 2009.

[4] who knows. Atomic unit reference. *some jounral*, 0:0, 1900.

[5] whoknows. Bo reference. *some journal*, 0:0, 1900.

[6] unknown. variational principle. *none yet*, 0:0, 1900.

[7] S. F Boys. *Proc. R. Sco. (London)*, A200:542, 1950.

[8] P.-O. Lőwdin. *Physical Review*, 97:1474–1489, 1955.

[9] W. Kohn and I. J. Sham. *Physical Review*, A140:1133, 1963.

[10] P. J. Hay and W. R. Wadt. Ab initio effective core potentials for molecular calculations. potentials for the trnasition metal atoms sc to hg. *J. Chem. PHys.*, 82:270, 1985.

[11] P. J. Hay and W. R. Wadt. Ab initio effective core potentials for molecular calculations. potentials for main group elements na to bi. *J. Chem. Phys.*, 82:284, 1985.

[12] P. J. Hay and W. R. Wadt. Ab initio effective core potentials for molecular calculations. potentials for k to au including the outermost core orbitals. *J. Chem. Phys.*, 82:299, 1985.

[13] W. J. Stevens, Basch H., and Krauss M. Compact effective potentials and efficient shared-exponent basis sets for the first- and second-row atoms. *J. Chem. Phys.*, 109:3875, 1984.

[14] W. J. Stevens, M. Krauss, H. Basch, and P. G. Jasien. *Can. J. Chem.*, 70:612, 1992.

[15] T. R. Cundari and W. J. Stevens. Effective core potential methods for the lanthanides. *J. Chem. Phys.*, 98:5555, 1993.

[16] Michael Dolg. *Modern Methods and Algoritms of Quantum Chemistry, Proceedings*, chapter Effective Core Potentials, pages 507–540. John von Neumann Institute for Computing, 2000.

[17] D. G. Fedorov and K. Kitaura. The importance of three-body erms in the fragment molecular orbital method. *J. Chem. Phys.*, 120:6832–6840, 2004.

[18] Dmitri G. Fedorov and Kazuo Kitaura. Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. *J. Phys. Chem. A*, 111:6904–6914, 2007.

[19] Dmiitr G. Fedorov and Kazuo Kitaura, editors. *The Fragment Molecular Orbital Method: Practical Applications to Large Molecular Systems.* CRC Press, 2009.

[20] J LADIK and T A HOFFMANN. QUANTUM MECHANICAL CALCULATION OF THE ELECTRONIC STRUCTURE OF DNA. *Biopolymers symposia*, 13:117–127, 1964.

[21] JiÅ™í Šponer and Pavel Hobza. Nonplanar Geometries of DNA Bases. Ab Initio Second-Order Møller-Plesset Study. *J. Phys. Chem.*, 98:3161–3164, 1994.

[22] JiÅ™í Šponer, Jerzy Leszczynski, and Pavel Hobza. Hydrogen bonding, stacking and cation binding of DNA bases. *J. Mol. Struct. (Theochem)*, 573:43–53, 2001.

[23] Jiří Šponer, Jerzy Leszczynski, and Pavel Hobza. Electronic Properties, Hydrogen Bonding, Stacking, and Cation Binding of DNA and RNA Bases. *Biopolymers (Nucleic Acid Sciences)*, 61(1):3–31, 2002.

[24] David Řeha, Martin Kabeláč, Filip Ryjáček, Jiří Šponer, Judit E. Šponer, Marcus Elstner, Sándor Suhai, and Pavel Hobza. Intercalators. 1. Nature of Stacking Interactions between Intercalators (Ethidium, Daunomycin, Ellipticine, and 4'6-Diaminide-2-phenylindole) and DNA Base Pairs. *J. Am. Chem. Soc.*, 124(13):3366–3376, 2002.

[25] Jiří Šponer, Petr Jurečka, and Pavel Hobza. Accurate Interaction Energies of Hydrogen-Bonded Nucleic Acid Base Pairs. *J. Am. Chem. Soc.*, 126(32):10142–10151, 2004.

[26] Judit E Šponer, Nad'a Špačková, Jerzy Leszczynski, and Ji\vr\'i Šponer. Principles of RNA Base Pairing: Structures and Energies fo the Trans Watson-Crick/Sugar Edge Base Pairs. *J. Phys. Chem. B*, 109:11399–11410, 2005.

[27] Alberto Pérez, Jiri Sponer, Petr Jurecka, Pavel Hobza, F Javier Luque, and Modesto Orozco. Are the hydrogen bonds of RNA (AU) stronger than those of DNA (AT)? A quantum mechanics study. *Chemistry (Weinheim an der Bergstrasse, Germany)*, 11(17):5062–6, August 2005.

[28] Martin Kabelác, Edward C Sherer, Christopher J Cramer, and Pavel Hobza. DNA base trimers: empirical and quantum chemical ab initio calculations versus experiment in vacuo. *Chemistry (Weinheim an der Bergstrasse, Germany)*, 13(7):2067–77, 2007.

[29] Yves Pommier, Glenda Kohlhagen, Christian Bailly, Michael Waring, Abhijit Mazumder, and Kurt W. Kohn. DNA Sequence- and Structure-Selective Alkylation of Guanine N2 in the DNA Minor Groove by Ecteinascidin 743, a Potent Antitumor Compound from the Caribbean Tunicate *Ecteinascidia turbinata*. *Biochemistry*, 35(41):13303–13309, 1996.

[30] Frederick C. Seaman and Laurence H. Hurley. Molecular Basis for the DNA Sequence Selectivity of Ecteinascidin 736 and 743: Evidence for the Dominant Role of Direct Readout via Hydrogen Bonding. *J. Am. Chem. Soc.*, 120(50):13028–13041, 1998.

Index