Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Departamento de Química Inorgánica, Analítica y Química Física

Computer Simulation of Chemical Reactivity in Biomolecules

Alejandro Crespo

Directores: Darío A. Estrin y Adrián E. Roitberg.

Lugar de Trabajo: Departamento de Química Inorgánica, Analítica y Química Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

Trabajo de Tesis para optar por el título de Doctor en Ciencias Químicas.

Marzo 2006

a mis viejos

a agus

Computer Simulation of Chemical Reactivity in Biomolecules

Abstract

The main goals of this thesis have been (i) to implement a quantum mechanical-molecular mechanical (QM-MM) methodology and (ii) to apply classical and QM-MM simulations schemes to the study of chemical reactivity in biomolecules.

Regarding the first goal, we have successfully implemented a density functional theory (DFT) hybrid QM-MM method developed for simulation of reactions in complex environments. It is particularly suited to study enzyme active sites or solutes in condensed phases. The method combines a QM description of the solute with a MM treatment of the environment. The QM fragment is treated using DFT as implemented in the computationally efficient numerical basis set program SIESTA, while the environment is treated using the classical Amber force field parameterization.

The second goal has been achieved by applying this methodology to the investigation of three biological problems, in order to shed light into the molecular basis of the protein effects and to obtain new insights that would not be extracted by experimental work only. In the first place, we have performed calculations on a well-known system in order to validate the implemented methodologies, the chorismate to prephenate conversion catalyzed by the *Bacillus subtilis* chorismate mutase. We have correctly predicted both the energetic and entropic experimental contributions to the enzyme catalytic activity. Secondly, we have performed calculations for the NO detoxification process catalyzed by *Mycobacterium tuberculosis* oxygenated truncated hemoglobin N (trHbN). Our results suggest that the essential dynamics of the protein leads to a mechanism suited for selective migration of both O_2 and NO ligands and that the reaction itself is catalyzed mainly by means of the heme moiety. Finally, we have investigated the hydroxylation reaction catalyzed by the binuclear copper enzyme peptidylglycine $\alpha-$hydroxylating monooxygenase (PHM), concluding that the active specie that abstracts the substrate-HA atom is $[CuO]^{+2}$ and that the process occurs concertedly with almost no activation energy.

Keywords: DFT, QM-MM, SIESTA, Amber, Chorismate Mutase, trHbN, PHM.

Simulación Computacional de Reactividad Química en Biomoléculas

Resumen

Los objetivos principales de esta tesis han sido (i) implementar un método cuántico-clásico (QM-MM) y (ii) la aplicación de esquemas de simulación computacional clásicos y QM-MM al estudio de la reactividad química en biomoléculas.

Respecto al primer objetivo, hemos implementado exitosamente un método híbrido QM-MM basado en la teoría de los funcionales de la densidad (DFT), desarrollado para simular reactividad química en medios complejos. En particular, éste método es adecuado para el estudio de sitios activos enzimáticos o solutos en fases condensadas. El método combina una descripción basada en la mecánica cuántica (QM) para el soluto, con un tratamiento clásico (MM) del medio. El subsistema QM es descripto a nivel DFT, mediante la implementación del eficiente programa de bases numéricas SIESTA, mientras que el entorno se describe utilizando la parametrización clásica basada en el campo de fuerzas Amber.

El segundo objetivo se ha logrado mediante la aplicación de dicha metodología a la investigación de tres problemas biológicos. Nuestra meta fue arrojar luz en las bases moleculares de los efectos del entorno y obtener nuevos resultados no obtenibles sólo de la experimentación. En primer lugar, hemos llevado a cabo cálculos sobre un sistema bien conocido para validar nuestro método, la conversión de corismato a prefenato catalizada por la enzima corismato mutasa del *Bacilus subtilis.* Hemos predicho correctamente tanto los factores energéticos como entrópicos experimentales para la actividad catalítica de esta enzima. Seguidamente, hemos llevado a cabo cálculos computacionales de la reacción de detoxificación del NO catalizada por la enzima hemoglobina truncada N (trHbN) oxigenada del *Mycobacterium tuberculosis*. Nuestros resultados sugieren que la dinámica de la enzima conlleva a un mecanismo de regulación de entrada selectiva de los ligandos O_2 y NO y que la reacción química es catalizada por el grupo hemo solamente. Finalmente, hemos investigado la reacción de hidroxilación llevada a cabo por la enzima binuclear de cobre monooxigenasa $\alpha-$hidroxilante de peptidil-glicinas (PHM), concluyendo que la especie activa capaz de abstraer el átomo HA del sustrato es $[CuO]^{+2}$ y que el proceso ocurre concertadamente y sin energía de activación.

Palabras Clave: DFT, QM-MM, SIESTA, Amber, Corismato Mutasa, trHbN, PHM.

Computer Simulation of Chemical Reactivity in Biomolecules

Index

Quien no conoce nada, no ama nada. Quien no puede hacer nada, no comprende nada. Quien nada comprende, nada vale. Pero quien comprende también ama, observa, ve... Cuanto mayor es el conocimiento inherente a una cosa, más grande es el amor... Quien cree que todas las frutas maduran al mismo tiempo que las frutillas nada sabe acerca de las uvas.

Paracelso

Yo me alejo con espanto y horror de la triste maldad de las funciones que no tienen derivadas.

Charles Hermite

Chapter 1: Introduction

Computational Biological Chemistry?

Biological systems are responsible for life. In the last fifty years, since the DNA double-helix structure discovery, elucidating and understanding the molecular basis of biomolecular function has played a central role in biochemical and biological sciences, starting from DNA transcription, DNA transduction and the biosynthesis of proteins to the understanding of what their principal functions are inside the cells. Proteins are then, the leading moieties in almost all biological processes, and are involved in catalysis, transport, growth control, cellular differentiation, besides their functions as mechanical support and molecular motors.[1]

Most biological transformations are performed by a special type of proteins called enzymes, which can catalyze biochemical reactions that would not occur at any appreciable rate if they were not present.[1] The most important properties of enzymes are their capability of catalyzing chemical reactions at moderate temperature and pressure conditions, their high rate constants and their substrate specificity. Detailed microscopic understanding of the enzymatic mechanisms, interaction with substrates and other agents, transmission of the information, correlation between their amino-acids sequence and conformation, or in general between their structure and function, are the prime objectives of contemporary biological chemistry research and nowadays represent one of the most attractive and ambitious challenges in science, leading to a wide variety of applications, such as drugs design or developing of new protein based catalysts.

The advance in the understanding of protein's behavior at a microscopic level has been impressive in the last decades. Experimental molecular biology, especially by means of sequencing and recombinant DNA techniques, is one of the key factors for such success. However, this evolution would not have been possible without the contributions arising from physics and chemistry, especially by techniques such as X-ray diffraction and NMR spectroscopy and the purification methods as chromatography or electrophoresis.[1]

On the other hand, in the recent years, a different approach has been used to explore the microscopic behavior of biomolecules: computer simulation techniques. Computational schemes have been adopted by the biological community and have been successfully implemented in a wide variety of examples, allowing us to obtain structural, thermodynamic, spectroscopic and dynamical information that in some cases would not be accessible by means of experimental techniques.[2,3] This impressive advance in computer simulation techniques applied to the study of complex chemical systems is due to the availability of efficient and economic computers and the

development of new methodologies. This has been evidenced by the 1998 Chemistry Nobel price awarded to professors John Pople and Walter Kohn for their fundamental contributions in this field. Computational chemistry is then a fundamental tool for elucidating enzymatic reactivity providing microscopic information hardly accessible by other techniques.

Basically, these kinds of simulations are based on the principles of quantum mechanics (QM) or classical mechanics (MM). In classical methods the atoms interact themselves by parameterized potentials, constituting the force field. The electrons are not explicitly included in the calculations, but they are implicitly present by means of the parameterization, and therefore these methods are not suited for studying chemical reactions or any other process involving charge transfer or electronic rearrangements. However, they can be correctly applied to calculations of structural or dynamical properties, especially when the number of atoms constituting the system is very high, such as biomolecules or polymers in general. Protein folding, conformational analysis and protein-ligand interactions, where there is no chemical reaction involved, are examples of phenomena which may be investigated by classical models.[4,5] Techniques based on QM, called electronic structure methods, are in principle capable to deal with all kind of phenomena, allowing us to obtain information that would not be obtained with classical mechanics, such as chemical reactivity or electronic excitations.

The dawn of quantum mechanics dates at the beginning of the 20th century with the contributions of great scientists such as Planck, Einstein, Schrödinger, Heisenberg, Dirac, and Feynman among others.[6] The quantum mechanics formalism originated a turning point in physics and in the way of thinking how to describe microscopic systems such as atoms or molecules, which classical mechanics fails to describe. In fact, solving Schrödinger equation provides us all system information or all electronic properties if applied to a molecule, including its energy and structure.[7] However, this equation has not analytical solution for systems containing more than one electron. Finding approximate methodologies in order to solve this equation gave rise to quantum chemistry. From the origin of such theory till today, there has been a great effort to develop efficient techniques in order to solve the electronic problem with a better degree of accuracy. Quantum chemistry has emerged as an essential tool in chemical research in the last decades, mainly due to the advances in computer hardware and methodology.

The Hartree-Fock method (HF),[8] developed around the 30's, was one of the most used methodologies for approximately solving Schrödinger equation. It is based on the variation principle and it is a self-consistent method, because the Hamiltonian explicitly depends upon its solution. A different approach is the density functional theory (DFT).[8] It was originally proposed by Hohenbeg, Kohn and Sham in the middle 60's and is based in using the electronic density for describing the system and solving the electronic problem. During a long time, the time required

and the hardware limitations, made the computer calculations almost impossible for systems interesting from the biological standpoint. However, since the beginning of the 80's, when the computational resources became both efficient and available, computer simulations based in HF of DFT schemes, have been widely adopted by the quantum chemistry community.

Application of quantum chemistry to biological systems has initially consisted in studying small model system of the relevant biochemical molecules. Despite the simplicity of these models, it has been possible to obtain information about reaction paths, conformational properties and specific interactions. However, in the last years, the situation has drastically changed, given the fact that there are new accessible methodologies to treat more complex systems at an adequate level of theory, such as techniques based in DFT. The exact description of an enzymatic system involving one enzyme, the substrate and the solvent, could be carried out by performing a molecular dynamics simulation employing the Schrödinger equation solution of the entire system. However, due to the nonlinear increase of computational cost of QM calculations, the treatment of these large systems is not presently feasible. One way to overcome this limitation has been oriented to the development of linear-scaling electronic structure methods.[9-14] The development and validation of theoretical models which balance accuracy and computational efficiency is nowadays a very active research area.

The model systems calculations previously discussed give us valuable information about the system properties, but they can difficultly describe neither issues such as microsolvation, enzyme-substrate specific interaction nor the dynamical behavior of the processes. For these reasons, new methodologies have been developed in order to describe more realistically environment effects on biochemical reactions. The most employed strategy is the so called hybrid quantum mechanical-molecular mechanical (QM-MM) scheme.[15-20] These schemes are adequate for the investigation of chemical events that take place in a certain region of a large system. The fundamental strategy of QM-MM methods is very attractive from a chemical point of view and is based in the fact that, in almost enzymes, the active site is limited to a small number of atoms. This subsystem, where most of the electronic density changes occur during the reaction, must be necessary described by means of QM, while the rest of the system, the environment with no bond breaking or forming, can be adequately represented by a classical force field parametrization (MM).

The use of these kind of techniques for the investigation of chemical reactivity in solution and in enzymes has increased enormously in the past years.[21-26] Due to the fact that a large degree of configurational sampling is necessary in most biological applications, the computational expense of QM-MM implementations remains a crucial issue, and the electronic structure scheme should be carefully chosen in order to provide accurate results at low cost for each particular problem.

The use of QM-MM schemes in which the QM calculation involves a linear scaling approach may be the best strategy for dealing with problems in which it is necessary to select a large QM subsystem.[9-14]

These kinds of schemes have been widely employed to the study of small molecules in condensed phase and in nanoaggregates.[27] Other applications include surface and zeolite studies.[28,29] However, in this thesis we will focus in applying QM-MM methodologies to biological systems. Illustrative examples of these kind of calculations are the binding energy and electron transfer between Putidarredoxine and Cytochrome P450cam,[30] modeling of the catalytic mechanism in Methane Monooxygenase,[31] and conformational equilibrium inside Chorismate Mutase active site.[32]

In most cases, the system description is achieved by employing classical force field parameterizations, such as Amber[33] or Charm,[34] for the classical subsystem and semiempirical or Hartree-Fock schemes for the active site. These approximations in describing the electronic structure of the QM subsystem can be improved employing DFT based schemes. Inclusion of these methods into the QM formalism gives us a more realistic description of the system, and is of prime interest for studying metalloenzymes,[35] where electronic correlation effects are essential for an accurate description of chemical reactivity.

Under these considerations, it is evident that computer simulation methods are now in a position where they can significantly contribute to the elucidation and interpretation of the enzymatic chemical reactivity, because they allow us to describe protein behavior at a molecular level. Even though nowadays they are usually employed as a complement to experimental research, it is not unreal to think that with better QM methods and faster computational implementations, computer simulations techniques will become a fundamental tool for accessing to chemical mechanisms and microscopic detailed information that would not be possible to obtain by any other method.

Objectives and Outline

The present work has two prime objectives:
1) The first is a methodological and more general one. It consists in developing a hybrid QM-MM methodology suited for describing chemical reactions in complex environments by coupling an efficient code based in DFT for describing the electronic structure of the active site together with a force field parameterization for the rest of the protein, the environment. DFT has shown to be an optimal cost-benefit alternative.[36] We will employ efficient DFT schemes, in order to make the electronic structure calculations feasible. In particular, we will use the SIESTA method (Spanish Initiative for Electronic Structure of Thousands of Atoms).[37] This electronic structure technique will

be coupled with classical force fields potentials as implemented in the Amber[33] or Charmm[34] codes. This part of the present thesis work is quite ambitious, because it involves developing new methodologies and their implementation in complex and efficient programs.

In Chapter 2 we will briefly introduce the theoretical foundations of this thesis. We will describe the QM description of matter, based in solving the Schrödinger equation. We will explain different methodologies, from HF to DFT, which is mostly used throughout this thesis. It will be also important to describe simulation techniques based on classical potentials. A complete description of hybrid QM-MM methodologies is given, taking particular attention to the implementation on the SIESTA code, which is the described objective. Finally, we will explore different techniques for exploring the potential energy surface.

2) The second objective of the present work is to apply our QM-MM technique to study relevant biological systems. We will start by performing calculations on well-known systems in order to validate the implemented methodologies. In the second place, we will apply our computational schemes to the investigation of open questions in chemical reactivity of biological systems. Our goal is to shed light into the molecular basis of the selected problems in order to both complement experimental information and obtain insight that is not evident from experimental results alone.

Chapter 3 is focused on the discussion of the chorismate to prephenate conversion catalyzed by the *Bacillus subtilis* Chorismate Mutase.[38] This was done in order to present and test our QM-MM scheme, because it is a problem for which there is an important body of theoretical and experimental data. This intramolecular transformation is of special interest because Chorismate Mutase is an essential enzyme in the Shikimate pathway, responsible for the biosynthesis of aromatic amino acids in bacteria, fungi, and plants, but not in mammals. So, this problem is of interest not only for an understanding of its catalytic rate enhancement, but also as a target for the discovery of antibiotics, antifungals, and herbicides.

In Chapter 4 we will present the results obtained for the NO detoxification process performed by *Mycobacterium tuberculosis* truncated hemoglobin N (trHbN).[39] *M. tuberculosis* is the causative agent of human tuberculosis and is forced into latency by nitric oxide produced by macrophages during infection. In response to nitrosative stress, *M. tuberculosis* has evolved a defense mechanism that relies on the oxygenated form of "truncated hemoglobin" N (trHbN), formally acting as NO-dioxygenase, yielding the harmless nitrate ion. We will shed light on the molecular basis for the NO scavenging reaction of oxygenated trHbN from *M. tuberculosis* by using both classical and QM-MM tools.

Chapter 5 is devoted to the investigation of the hydroxylation reaction catalyzed by peptidylglycine $\alpha-$hydroxylating monooxygenase (PHM).[40] Hydroxylation of the Cα of a C-terminal glycine is the first step in the amidation reaction of many peptide hormones and neuropeptides in order to be biologically active. This reaction is catalyzed by separate enzymatic domains of peptidylglycine α-amidating monooxygenase (PAM). The second step consists in disproportionation of the α-hydroxyglycine previously generated. PHM, a two-copper, ascorbate-dependent enzyme is the more extensively studied of the two PAM domains.[40] For that reason, we will shed light into the molecular basis of the hydrogen abstraction mechanism catalyzed by this enzyme.

Finally, in Chapter 6, we will present the general conclusions of the present work, as well as our future perspectives in computer simulation techniques for studying chemical reactivity in biological systems.

[1] Voet, D.; Voet, J. G. Biochemistry, John Wiley & Sons, New York, 1995.

[2] Loew, G. H.; Harris, D. L. Chem. Rev. 2000, 100, 407.

[3] Siegbahn, P. E. M.; Blomberg, M. R. A. Chem. Rev. 2000, 100, 421.

[4] Daggett, V. Methods. Mol. Biol. 2001, 168, 215.

[5] Wang, W; Donini, O.; Reyes, C. M.; Kollman P. A. Annu. Rev. Biophys. Biomol. Struct. 2001, 30, 211.

[6] Jammer, M. The Conceptual Development of Quantum Mechanics (The History of Modern Physics 1880-1950, Vol. 12), Tomash Publishers, 1989.

[7] Szabo, A.; Ostlund, N. S. Modern Quantum Chemistry, First Edition, Mc Graw-Hill Publishing Company, New York, 1989.

[8] Levine, I. N. Quantum Chemistry, Fifth Edition, Prentice-Hall, Inc., New Jersey, 2000.

[9] Scuseria, G. E. J. Phys. Chem. A 1999, 103, 4782.

[10] Gogonea, V.; Merz, K. M., Jr. J. Phys. Chem. A 1999, 103, 5171.

[11] Greatbanks, S. P.; Gready, J. E.; Limaye, A. C.; Rendell, A. P. J. Comput. Chem. 2000, 21, 788.

[12] York, D.; Lee, T.; Yang, W. Phys. Rev. Lett. 1998, 80, 5011.

[13] Lewis, J. P.; Carter, C. W., Jr.; Hermans, J.; Pan, W.; Lee, T.; Yang, W. J. Am. Chem. Soc. 1998, 120, 5407.

[14] Khandogin, J.; York, D. M. J. Phys. Chem. B 2002, 106, 7693.

[15] Kohanoff, J.; Koval, S.; Estrin, D.; Laria, D.; Abashkin, Y. J. Chem. Phys. 2000, 112, 9498.

[16] Elola, M. D.; Laria, D.; Marceca, E. J.; Estrin, D. Chem. Phys. Lett. 2000, 326, 509.

[17] Elola, M. D.; Laria, D.; Estrin, D. J. Phys. Chem. A. 1999, 110, 9039.

[18] Roitberg, A.; Worthington, S. E.; Holden, M. J. J. Am. Chem. Soc. 2000, 122, 7312.

[19] Field, M.; Bash, P.; Karplus, M. J. Comp. Chem. 1990, 11, 700.

[20] Jung, Y.; Choi, C.; Gordon, M. J. Phys. Chem B. 2001, 105, 4039.

[21] Merz, K. M., Jr. Encycl. Comput. Chem. 1998, 4, 2330.

[22] Monard, G.; Merz, K. M., Jr. Acc. Chem. Res. 1999, 32, 904.

[23] Walker, R. C.; de Souza, M. M.; Mercer, I. P.; Gould, I. R.; Klug, D. R. J. Phys. Chem. B 2002, 106, 11658.

[24] Bruice, T. C. Acc. Chem. Res. 2002, 35, 139.

[25] Orozco, M.; Luque, F. J. Chem. Rev. 2000, 100, 4187.

[26] Kohanoff, J.; Koval, S.; Estrin, D. A.; Laria, D.; Abaskin, Y. J. Chem. Phys. 2000, 112, 9498.

[27] Gao, J. Reviews in Computational Chemistry, VCH Publishers: New York, 1996.

[28] Carmer, C.; Weiner, B.; Frenklach, M. J. Chem. Phys. 1993, 99, 1356.

[29] Sherwood, P.; De Vries, A.; Collins, S. Faraday Discuss. 1998, 19, 38.

[30] Roitberg, A.; Holden, M.; Mayhew, M.; Kurnikov, I.; Beratan, D.; Vilker, V. J. Am. Chem. Soc. 1998, 120, 8927.

[31] Friesner, R.; Dunietz, D. Acc. Chem. Res. 2001, 34, 351.

[32] Marti, S.; Andrés, J.; Moliner, V.; Silla, E.; Tuñón, I.; Beltran, J. J. Phys. Chem B. 2000, 104, 11308.

[33] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J. Am. Chem. Soc. 1995, 117, 5179.

[34] Brooks, B.R.; Bruccoleri, R.E.; Olafson, B.D.; States, D.J.; Swaminathan, S.; Karplus, M. J. Comp. Chem. 1983, 4, 187.

[35] Scherlis, D. A.; Cymmeryng, C. B.; Estrin, D. Inorg. Chem. 2000, 39, 2352.

[36] Siegbahn, P. E. M.; Blomberg, M. R. A. Chem. Rev. 2000, 100, 421.

[37] Soler, J. M.; Artacho, E.; Gale, J. D.; García, A.; Junquera, J.; Ordejón, P.; Sánchez-Portal, D. J. Phys.: Condens. Matter 2002, 14, 745.

[38] Crespo, A.; Scherlis, D. A.; Martí, M. A.; Ordejón, P.; Roitberg, A. E.;Estrin, D. A. J. Phys. Chem. B 2003, 107, 13728.

[39] Crespo A.; Martí, M. A.; Kalko, S. G.; Morreale, A.; Orozco, M.; Gelpi, J. L.; Luque, F.J.; Estrin, D. A. J. Am. Chem. Soc. 2005, 127, 4433.

[40] Prigge, S. T.; Eipper, B. A.; Mains, R. E.; Amzel. L. M. Science. 2004, 304, 864.

Chapter 2: Methodology

The theoretical foundations of this thesis are briefly introduced in this chapter. We will start by presenting the quantum-mechanical description of matter, based in solving the Schrödinger equation. In this context, different techniques are explained, such as Hartree-Fock and density functional theory, which is the most used implementation of quantum mechanics employed throughout this thesis. It is also of particular interest introducing simulation schemes based on classical potentials, such as molecular mechanics. A complete description of hybrid quantum-classical methodologies is given, taking particular attention to the implementation of the SIESTA QM-MM code. Finally, the issue of sampling the potential energy surface is also explored, specifically energy minimization, molecular dynamics simulation and free energy calculations.

The Schrödinger equation and the variation principle

Any approximation to study atomic or molecular systems from first principles must be based in the theory which describes microscopic particles behavior, quantum mechanics. Quantum mechanics, developed at the beginnings of the last century, is based in solving the Schrödinger equation:

$$-\frac{\hbar}{i}\frac{\partial \Psi(\mathbf{r},t)}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\Psi(\mathbf{r},t) + V(\mathbf{r},t) \tag{2.1}$$

This is a differential equation whose resolution gives us the wave function $\Psi(\mathbf{r},t)$, which is the mathematical description of a quantum particles system, subject to the external potential $V(\mathbf{r},t)$. The wave function contains all the system information because it allows us to obtain any system property. The quantum chemistry main goal is then to solve this differential equation for a system containing nucleus and electrons, such as a molecule. For these kinds of systems, an exact solution is not possible and approximate solutions should be seeked. In the case of time independent potentials, $\Psi(\mathbf{r},t)$ can be rewritten as the product of a spatial and a time part. This procedure allows us to reduce the Schrödinger equation to a time independent form:

$$-\frac{\hbar^2}{2m}\nabla^2\Psi(\mathbf{r},t) + V(\mathbf{r},t) = E\Psi(\mathbf{r},t) \tag{2.2}$$

or, defining the Hamiltonian operator as $\hat{H} = -\frac{\hbar^2}{2m}\nabla^2 + \hat{V}$,

$$\hat{H}\,\Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \tag{2.3}$$

where E is the system energy and **r** represents the coordinates of the particles. The Schrödinger equation is then a partial differential eigenvalue equation in which an operator, **H**, acts on a function (the eigenfunction, $\Psi(\mathbf{r})$) and returns the function multiplied by a scalar (the eigenvalue, E). For the case of a system containing N electrons and M nucleus, the Hamiltonian operator in atomic units[1] is:

$$\hat{H} = -\sum_{i=1}^{N}\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{R_{AB}} \quad (2.4)$$

where i and j are the electron indexes and A and B are the nuclear ones. The first two terms are the electronic and nuclear kinetic energy, the third term is the nuclear-electron attraction, and the fourth and fifth terms are the electron-electron and nuclear-nuclear repulsion, respectively.

One common procedure to all quantum chemistry calculations is to uncouple the electron and nuclear motions, in order to solve the electronic problem alone. This procedure is better known as the Born-Oppenheimer approximation,[2] based in the fact that since the nucleus are much heavier than the electrons, they move more slowly. This means that the electrons are subject to a static potential of punctual charges or that they instantaneously respond to any nuclear motion. Under this approximation, the second term in equation (2.4) is zero and the last one becomes a constant. The rest of the terms are what is called the electronic Hamiltonian, which has an associated electronic wave function that is a function of the electrons coordinates only and depends parametrically on the nuclear coordinates.

If we take into account the spin of the electrons, we must satisfy the antisymmetric quantum mechanical principle for fermions, which demands that any well behaved wave function must be antisymmetric when interchanging the coordinates of any two electrons:

$$\Psi(x_1,...,x_i,...,x_j,...,x_N) = -\Psi(x_1,...,x_j,...,x_i,...,x_N) \quad (2.5)$$

where x_i denotes the position and the spin. If we define the spin-orbital, χ_i as the one electron wave function, the antisymmetric requirement for the all electron wave function is accomplished by using a Slater determinant of the form:

$$\Psi(x_1,x_2,...,x_N) = (N!)^{1/2}\begin{vmatrix} \chi_i(x_1) & \chi_j(x_1) & \cdots & \chi_k(x_1) \\ \chi_i(x_2) & \chi_j(x_2) & \cdots & \chi_k(x_2) \\ \vdots & \vdots & & \vdots \\ \chi_i(x_N) & \chi_j(x_N) & \cdots & \chi_k(x_N) \end{vmatrix} \quad (2.6)$$

This determinant involves N indistinguishable electrons distributed over N spin-orbitals, χ_i with x_i being the spatial and spin coordinates of electron i. This is the simplest way of generating an antisymmetric wave function.

So far, we have presented the Schrödinger equation and the mathematical form of its solution, the wave function. However, we have not said anything about how to find it. In this sense, the variation principle is the way to find an approximate solution to an eigenvalue problem, such as the Schrödinger equation, without solving it in a strict way. Let's first note that given the Hamiltonian operator, there is an infinite set of solutions to the time independent Schrödinger equation:

$$\hat{H}\left|\Psi_i\right\rangle = \varepsilon_i\left|\Psi_i\right\rangle \qquad i = 0,1,2,... \qquad \varepsilon_0 \leq \varepsilon_1 \leq ... \leq \varepsilon_i \leq ...$$

where ε_i are the exact eigenvalues, ε_0 the energy of the ground state, and $\left|\Psi_i\right\rangle$ are the orthonormal eigenvectors such that $\left\langle\Psi_i\middle|\Psi_j\right\rangle = \delta_{ij}$. Because $\left\{\left|\Psi_i\right\rangle\right\}$ is a complete set of functions, any other function, $\left|\Phi\right\rangle$ can be written as the linear combination:

$$\left|\Phi\right\rangle = \sum_i c_i\left|\Psi_i\right\rangle, \quad c_i = \left\langle\Psi_i\middle|\Phi\right\rangle \tag{2.7}$$

If the function $\left|\Phi\right\rangle$ is normalized, the variation principle states that the Hamiltonian expected value on $\left|\Phi\right\rangle$ will be greater or equal to the ground state energy. So, if $\left\langle\Phi\middle|\Phi\right\rangle = 1$, then:

$$\left\langle\Phi\middle|\hat{H}\middle|\Phi\right\rangle \geq \varepsilon_0 \tag{2.8}$$

The variation principle application to the electronic problem is as follows: we want to find a function that minimizes the energy of the system $\left\langle\Phi\middle|\hat{H}\middle|\Phi\right\rangle$. In practice, this trial wave function is expressed as a linear combination:

$$\left|\Phi\right\rangle = \sum_i^N c_i\left|\phi_i\right\rangle \tag{2.9}$$

being $\left\{\left|\phi_i\right\rangle\right\}$ a known set of N basis functions. The idea now is to find the set of coefficients c_i which minimize the expectation value of the energy by using (2.9).

$$\left\langle\Phi\middle|\hat{H}\middle|\Phi\right\rangle = \sum_{ij} c_i\left\langle\phi_i\middle|\hat{H}\middle|\phi_j\right\rangle c_j = \sum_{ij} c_i c_j H_{ij} \tag{2.10}$$

where $H_{ij} = \left\langle\phi_i\middle|\hat{H}\middle|\phi_j\right\rangle$ are the Hamiltonian matrix elements in the $\left\{\left|\phi_i\right\rangle\right\}$ basis. Taking into account that $\left|\Phi\right\rangle$ is normalized, the coefficients are given by:

$$\left\langle\Phi\middle|\Phi\right\rangle = \sum_{ij} c_i c_j\left\langle\phi_i\middle|\phi_j\right\rangle = \sum_i c_i^2 = 1 \tag{2.11}$$

In order to minimize the energy (2.10) subject to the constraint imposed in (2.11), we use the Lagrange's method of undetermined multipliers.[2] Let us construct a function λ subtracting the

term $E(\langle \Phi | \Phi \rangle - 1)$ to the original expression (2.10), which is equivalent to add zero because of

(2.11), and so the minimum of both $\langle \Phi | \hat{H} | \Phi \rangle$ and λ occur at the same values of the coefficients. This procedure introduces a new degree of freedom associated to the new multiplier E that allows us to obtain the best coefficients independently:

$$\frac{\partial \lambda}{\partial c_k} = 0 = \sum_j c_j H_{kj} + \sum_i c_i H_{ik} - 2E c_k \qquad k = 1, 2, ..., N \qquad (2.12)$$

but since $H_{ij} = H_{ji}$, if real basis functions are used, we have:

$$\sum_j H_{ij} c_j - E c_i = 0 \qquad i = 1, 2, ..., N \qquad (2.13)$$

or in matrix notation:

$$\mathbf{Hc} = E\mathbf{c} \qquad (2.14)$$

where $\mathbf{c}$ represents a column vector with elements c_i. The equation (2.14) is the standard eigenvalue problem for the matrix $\mathbf{H}$ and can be solved to yield N orthonormal eigenvectors $\mathbf{c}^\alpha$ and corresponding eigenvalues E_α. Introducing the diagonal matrix $\mathbf{E}$ containing the eigenvalues and the matrix of eigenvectors $\mathbf{C}$ defined as $C_{i\alpha} = c_i^\alpha$, the N relationships obtained from solving (2.14) can be written as:

$$\mathbf{HC} = \mathbf{EC} \qquad (2.15)$$

Solving (2.15) we obtain N orthonormal eigenfunctions,

$$|\Phi_\alpha\rangle = \sum_i^N c_i^\alpha |\phi_i\rangle \qquad \alpha = 0, 1, ..., N-1 \qquad (2.16)$$

each of them with an eigenvalue E_α associated. Taking into account that the eigenvectors $\mathbf{c}^\alpha$ are orthonormal, the eigenvalue E_α is the expectation value of the Hamiltonian with respect to $|\Phi_\alpha\rangle$:

$$\langle \Phi_\beta | \hat{H} | \Phi_\alpha \rangle = \sum_{ij} c_i^\beta \langle \phi_i | \hat{H} | \phi_j \rangle c_j^\alpha = \sum_{ij} c_i^\beta H_{ij} c_j^\alpha$$
$$= \mathbf{c}^\beta \mathbf{H} \mathbf{c}^\alpha = E_\alpha \mathbf{c}^\beta \mathbf{c}^\alpha = E_\alpha \delta_{\alpha\beta} \qquad (2.17)$$

In particular, the lowest eigenvalue is the best possible approximation to the ground state energy of H in the space spanned by the basis functions $\{|\Phi_\alpha\rangle\}$, and it can also be shown that the rest of the eigenvectors are upper bounds for the energy of the different excited states of the Hamiltonian.

In summary, the linear variational method is a procedure for finding, by diagonalizing the $\mathbf{H}$ matrix, the N solutions for the time independent Schrödinger equation. The wave function of the ground electronic state, $|\Phi_0\rangle$ is the eigenvector corresponding to the lower energy eigenvalue.

The Hartree-Fock Approximation

Finding and describing approximate solutions to the electronic Schrödinger equation has been a major concern of quantum chemists since the birth of quantum mechanics. In general, quantum chemists are faced with many-electron problems. Central to attempt solving such problems, is the Hartree-Fock approximation, which has played an important role in the development of modern chemistry. In addition, it usually constitutes the first step towards more accurate approximations. The basic idea which underlies this approximation is to apply the variation principle to a system described by a Hamiltonian such as the one introduced in (2.4), by using the Born-Oppenheimer approximation and choosing the electronic wave function by means of a Slater determinant as defined in (2.6).

Let us define the electronic Hamiltonian operator $\hat{H}_{el}$, by taking into account only the first, third and forth terms of equation(2.4):

$$\hat{H}_{el} = -\sum_{i=1}^{N}\frac{1}{2}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} \tag{2.18}$$

If the wave function Ψ is defined by means of a Slater determinant in terms of the orbitals χ_i as in (2.6), the electronic energy is:

$$E = \left\langle \Psi \middle| \hat{H}_{el} \middle| \Psi \right\rangle = \sum_{i=1}^{N} H_{ii}^{core} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(J_{ij} - K_{ij}\right) \tag{2.19}$$

where we have introduced the Coulomb J_{ij} and exchange K_{ij} integrals, which depend on two electrons coordinates, and the term H_{ii}^{core} arising from the integral of only one electron coordinate. If we consider real orbitals, then:

$$H_{ii}^{core} = -\frac{1}{2}\left\langle \chi_i(1) \middle| \nabla_1^2 \middle| \chi_i(1)\right\rangle - \left\langle \chi_i(1) \middle| \sum_{\alpha}\frac{Z_\alpha}{r_{1\alpha}} \middle| \chi_i(1)\right\rangle \tag{2.20}$$

$$J_{ij} = \left\langle \chi_i(1)\chi_j(2) \middle| r_{12}^{-1} \middle| \chi_i(1)\chi_j(2)\right\rangle \tag{2.21}$$

$$K_{ij} = \left\langle \chi_i(1)\chi_j(2) \middle| r_{12}^{-1} \middle| \chi_j(1)\chi_i(2)\right\rangle \tag{2.22}$$

whereas H_{ii}^{core} is associated to the kinetic energy of electron *i* and its interaction with the nucleus, the Coulomb integral J_{ij} can be identified as the electron-electron repulsion. The exchange integral K_{ij}, however, comes from the wave function antisymmetric requirement and has no relationship with any quantity from classical mechanics. Nevertheless, its contribution is far from being negligible and should be taken into account.

According to the variation principle, the system wave function will be the one defined in terms of the spin-orbitals χ_i which minimize the energy of the system. If we minimize expression (2.19) subject to the restriction that the spin-orbitals χ_i are orthonormal, one can find a set of relationships for the best spin-orbitals χ_i. These particular relationships are known as the Hartree-Fock equations,[2] which at first glance are not in the standard eigenvalue form,

$\hat{F}(1)\left|\chi_i(1)\right\rangle = \sum_{j=1}^{N} \varepsilon_{ji} \left|\chi_j(1)\right\rangle$. However, one can introduce a new set of spin-orbitals by performing a unitary transformation to the original ones, without changing the absolute value of the Slater determinant, which allow us to rewrite the Hartree-Fock equations in the canonical eigenvalue form:[2]

$$\hat{F}(1)\left|\chi_i(1)\right\rangle = \varepsilon_i \left|\chi_i(1)\right\rangle \qquad i = 1, 2, ..., N \tag{2.23}$$

where the eigenfunctions are the N spin-orbitals χ_i, also called molecular or Hartree-Fock orbitals, whose energies are given by the eigenvalues ε_i. $\hat{F}$ is the Fock operator which can be written in terms of three different operators:

$$\hat{F}(1) = \hat{H}^{core}(1) + \sum_{j=1}^{N} [\hat{J}_j(1) - \hat{K}_j(1)] \tag{2.24}$$

each of them defined by its action over the orbital χ_i:

$$\hat{H}^{core}(1)\chi_i(1) = -\frac{1}{2}\nabla_1^2 \chi_i(1) - \sum_{\alpha} \frac{Z_\alpha}{r_{1\alpha}} \chi_i(1) \tag{2.25}$$

$$\hat{J}_j(1)\chi_i(1) = \left[\int dx_2 \chi_j^*(2) r_{12}^{-1} \chi_j(2)\right] \chi_i(1) \tag{2.26}$$

$$\hat{K}_j(1)\chi_i(1) = \left[\int dx_2 \chi_j^*(2) r_{12}^{-1} \chi_i(2)\right] \chi_j(1) \tag{2.27}$$

the first one, $\hat{H}^{core}$ operates over one electron coordinates, but $\hat{J}_j$ and $\hat{K}_j$, called the Coulomb and exchange operators, respectively, involve two electron integrals. The exchange operator, arising from the antisymmetric nature of the determinant, is different from the Coulomb operator because depends on the value of χ_i throughout all space, and since there does not exist a simple potential uniquely defined at each points of space, it is said to be a nonlocal operator. On the other hand, $\hat{J}_j$ which represents the repulsion between electron pairs assuming that for each electron i there is an effective or averaged local potential due to the presence of electron j, and so is a local operator.

So far, we have the equations that the spin-orbitals must fulfill in order to find the system wave function. However, one can see that in these eigenvalue equations, the Fock operator depends explicitly of its eigenfunctions, which are unknown. Obtaining an analytical solution for this problem seems quite complex, then approximate solutions are introduced. In order to find the orbitals from the resolution of the Hartree-Fock equations it is convenient to expand them into a finite basis of known functions:

$$\left|\chi_i\right\rangle = \sum_{\mu=1}^{K} C_{\mu i} \left|\phi_\mu\right\rangle \tag{2.28}$$

so now the problem is reduced to find the values of the coefficients $C_{\mu i}$. If we use a large number of K basis functions $\left|\phi_\mu\right\rangle$, this will lead us to a correct representation of the orbitals. Substitution of the expansion equation (2.28) into the Hartree-Fock eigenvalue equations leads us to:

$$\sum_{\mu=1}^{K} C_{\mu i} \hat{F} \left|\phi_\mu\right\rangle = \varepsilon_i \sum_{\mu=1}^{K} C_{\mu i} \left|\phi_\mu\right\rangle \tag{2.29}$$

and applying $\left\langle\phi_\nu\right|$ to both sides, we have a linear homogeneous system of K equations:

$$\sum_{\mu=1}^{K} C_{\mu i} \left(\left\langle\phi_\nu\left|\hat{F}\right|\phi_\mu\right\rangle - \varepsilon_i \left\langle\phi_\nu|\phi_\mu\right\rangle \right) = 0 \qquad \nu = 1, 2, ..., K \tag{2.30}$$

which are called the Roothaan equations and will lead us to find the orbitals coefficients. Defining the matrix elements of **F** as $F_{\nu\mu} = \left\langle\phi_\nu\left|\hat{F}\right|\phi_\mu\right\rangle$, and of **S** as $S_{\nu\mu} = \left\langle\phi_\nu|\phi_\mu\right\rangle$, the matrix form of the Roothaan equations will be:

$$\mathbf{FC} = \mathbf{SC\varepsilon} \tag{2.31}$$

where **C** is the coefficients matrix of $C_{\mu i}$, and $\mathbf{\varepsilon}$ is the diagonal matrix containing the K eigenvalues ε_i, which can be found by solving the secular equation:

$$\det\left|\mathbf{F} - \mathbf{\varepsilon S}\right| = 0 \tag{2.32}$$

Diagonalizing equation (2.31) leads us to the coefficients needed to construct the orbitals and calculate the energy, however, as the Fock matrix depends on such coefficients, as can be seen:

$$F_{\nu\mu} = -\frac{1}{2}\left\langle\phi_\nu(1)\left|\nabla_1^2\right|\phi_\mu(1)\right\rangle - \sum_\alpha\left\langle\phi_\nu(1)\left|\frac{Z_\alpha}{r_{1\alpha}}\right|\phi_\mu(1)\right\rangle$$
$$+\sum_a^N\sum_\lambda^K\sum_\sigma^K C_{\lambda a} C_{\sigma a}\left[\left\langle\phi_\nu(1)\phi_\mu(1)\left|r_{12}^{-1}\right|\phi_\sigma(2)\phi_\lambda(2)\right\rangle \right. \tag{2.33}$$
$$\left. -\left\langle\phi_\nu(1)\phi_\lambda(1)\left|r_{12}^{-1}\right|\phi_\sigma(2)\phi_\mu(2)\right\rangle\right]$$

the equations system resolution must be done in an iterative way. In practice, in order to make a Hartree-Fock calculation, you guess a starting tentative **C** matrix from which you can construct the

F matrix. Diagonalizing then (2.31) you find a new set of coefficients $C_{\mu i}$ that are now used to reconstruct the **F** matrix. The only term you have to reevaluate at each iteration is the last one of equation (2.33), because the first two terms of $F_{\mu\nu}$ and the $S_{\mu\nu}$ elements only depend on the basis functions, which are not modified during the iterative process. This procedure is repeated until the coefficients $C_{\mu i}$ do not differ significantly from the ones found in the step before, that is, when self-consistency has been reached. The computation of the two electron integrals in equation (2.33) is the most time consuming part of a Hartree-Fock calculation, because of the huge number of calculations that they demand.

The energy and the wave function quality obtained by this procedure are directly related with the size and capacity of the basis functions for describing correctly the electronic states. If the set of basis functions is more complete, the flexibility to expand the molecular orbitals increase, thus decreasing the energy value obtained. However, because the evaluation of the electronic repulsion is done by means of a mean field potential, the Hartree-Fock approximation lacks the correlation arising from their instantaneous positions. It can be thought then, that there is an exclusion region surrounding each electron, which is neglected in the mean field potential approximation, which makes an overestimation in the total energy. That is the reason why, even if using a complete basis set, the Hartree-Fock energy would be always greater than the exact energy in a quantity that is precisely called the correlation energy.

$$E_{corr} = E_0 - E_{HF} \tag{2.34}$$

E_0 represents the non-relativistic total energy corresponding to the exact Hamiltonian under the Born-Oppenheimer approximation, whereas E_{HF} is the Hartree-Fock energy limit. There are several ways to go beyond the Hartree-Fock approximation. One of them is based in expressing the wave function as a combination of several Slater determinants arising from different excited states configurations. This method is called configuration interaction (CI), whose computational cost is much greater than Hartree-Fock. Another method for including the correlation energy is the so called Møller Plesset perturbation theory.[2]

Density Functional Theory

A quantum mechanical description of matter based in the electronic density seems more appealing than the description using a wave function, because the density is related with a real physical observable. In this context there was one density based model contemporaneous to Hartree-Fock theory, the Thomas-Fermi model (1930), but it was unsuccessful to describe chemical bonding even qualitatively.[3] The really great contribution in this area was done in the sixties when Hohenberg and Kohn proved two theorems which allow us to rigorously describe the electronic problem by means of the charge density. Substituting the wave function Ψ , dependent

on N coordinates by the charge density ρ that depends only of the position, gives us a new particular way of treating atoms and molecules. The formalism adopted by this methodology is called Density Functional Theory (DFT).[4]

The first Hohenberg and Kohn[3] theorem (1964) establishes that the wave function, and therefore the energy and any electronic property of the system, is uniquely determined by the electronic density $\rho(x, y, z)$. It is interesting to note, that is the system Hamiltonian, or more precisely the external potential $v(\mathbf{r}_i)$[5] and the number of electrons, which determine, by solving the Schrödinger equation, the system wave function and energy levels. Hohenberg and Kohn demonstrated that the electronic density of the ground state $\rho_0(\mathbf{r})$ determines the external potential $v(\mathbf{r}_i)$ and the number of electrons. For that reason, there is a unique relationship between ρ_0 and Ψ_0, and then, the energy can be thought as a functional of the density, $E_0 = E_v[\rho_0(\mathbf{r})]$. The different contributions to the total energy can be written as:

$$E[\rho] = T[\rho] + V_{ee}[\rho] + V_{ne}[\rho] \tag{2.35}$$

T denotes the kinetic energy and V_{ee} the potential energy arising from the inter electronic interaction. V_{ne} represents the interaction between electrons and nucleus and can be written in terms of the density:

$$V_{ne} = \left\langle \Psi \Big| \sum_{i=1}^{N} v(\mathbf{r}_i) \Big| \Psi \right\rangle = \int \rho(\mathbf{r}) v(\mathbf{r}) d\mathbf{r} \tag{2.36}$$

The second great contribution to DFT is known as the variation Hohenberg and Kohn theorem. It is based on the fact that for any particular density ρ' where $\int \rho'(\mathbf{r}) d\mathbf{r} = N$ and $\rho'(\mathbf{r}) \geq 0$ in all space, then $E[\rho'] \geq E_o$. The equality holds when ρ' is the ground state density, which means that ρ_0 minimizes the energy functional $E[\rho]$ as well as Ψ_0 minimizes $\langle \Psi | H | \Psi \rangle$.

So far, we have a theory which relates the energy to the density, but it says nothing about how to calculate one or the other. En 1965 Kohn and Sham introduced an approximation for the functional form of the kinetic energy $T[\rho]$ which makes possible to find the density and the total energy in a straightforward way.[6] This procedure, known as the Kohn and Sham method, made DFT an extreme useful technique for electronic structure calculations.

The Kohn-Sham approximation considers a non interacting electrons, fictitious reference system (s sub index), whose density ρ_s, is equal to the real one, ρ_0. Doing this, the external

potential $v_s(\mathbf{r})$ must define, in the non interacting electrons system, a density which must be equal to the density of the N interacting electrons subject to the potential $v(\mathbf{r})$. For this reference system, the Hamiltonian $\hat{H}_s$ can be written as:

$$\hat{H}_s = \sum_{i=1}^{N}[-\frac{1}{2}\nabla_i^2 + v_s(\mathbf{r}_i)] = \sum_{i=1}^{N}\hat{h}_i^{KS} \tag{2.37}$$

where $\hat{h}_i^{KS} = -\frac{1}{2}\nabla_i^2 + v_s(\mathbf{r}_i)$ is the one electron Kohn-Sham operator. The wave function of the non interacting electrons system, Ψ_s is exactly a Slater determinant of the eigenfunctions of $\hat{h}_i^{KS}$ that are the spin-orbitals of the reference systems χ_i^{KS}, which satisfy:

$$\hat{h}_i^{KS}\chi_i^{KS} = \varepsilon_i\chi_i^{KS} \tag{2.38}$$

Then, the electronic density can be calculated as:

$$\rho_s = \sum_{i=1}^{N}\left|\chi_i^{KS}\right|^2 \tag{2.39}$$

The main reason that makes useful to employ a non interacting electrons system is because the kinetic energy is then simply calculated by:

$$T_s = -\frac{1}{2}\left\langle\Psi_s\left|\sum_i\nabla_i^2\right|\Psi_s\right\rangle = -\frac{1}{2}\sum_i\left\langle\chi_i^{KS}(1)\left|\nabla_1^2\right|\chi_i^{KS}(1)\right\rangle \tag{2.40}$$

if we define the kinetic energy difference ΔT between the real system and the reference system, ΔV_{ee} to the non classical contribution to the electron-electron interaction, and put them together in a new functional form $E_{xc}[\rho]$, such that:

$$\Delta T[\rho] = T[\rho] - T_s[\rho] \tag{2.41}$$

$$\Delta V_{ee}[\rho] = V_{ee} - \frac{1}{2}\int\int\frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{r_{12}}d\mathbf{r}_1 d\mathbf{r}_2 \tag{2.42}$$

defining the exchange-correlation functional $E_{xc}[\rho]$, as:

$$E_{xc}[\rho] = \Delta T[\rho] + \Delta V_{ee}[\rho] \tag{2.43}$$

then, the energy functional (2.35) can be re-written:

$$E[\rho] = T_s[\rho] + V_{ne}[\rho] + \frac{1}{2}\int\int\frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{r_{12}}d\mathbf{r}_1 d\mathbf{r}_2 + E_{xc}[\rho] \tag{2.44}$$

where T_s and V_{ne} are defined from (2.41) and (2.36). Then, E_0 could be calculated from ρ_0, except from the relatively small contribution of E_{xc}. The Kohn and Sham approximation involves a

functional exact by definition, where the exchange and correlation effects due to the quantum nature of the electrons are described by E_{xc}. However, the explicit dependence of E_{xc} with the density is unknown and then, the accuracy of DFT calculations will be related to the approximation used for this functional. The exact form of E_{xc} can be split into different energetic contributions, which are: i) the kinetic correlation defined in (2.41), ii) the Coulombic correlation, related with the electronic repulsion, iii) the exchange arising from the antisymmetric principle for electrons, iv) the auto interaction correlation: given the fact that the term $\int\int \rho(\mathbf{r}_1)\rho(\mathbf{r}_2)r_{12}^{-1}d\mathbf{r}_1 d\mathbf{r}_2$ includes the interaction between an electron with its own charge density, in order to describe the electron-electron interaction correctly, we need to introduce a correction if the approximate exchange does not compensate the Coulombic terms.

The energy functional (2.44), can be explicitly written in function of the orbitals:

$$E[\rho] = -\sum_{\alpha} Z_{\alpha} \int \frac{\rho(\mathbf{r}_1)}{r_{1\alpha}} d\mathbf{r}_1 - \frac{1}{2}\sum_{i=1}^{N}\left\langle \chi_i^{KS}\left|\nabla_1^2\right|\chi_i^{KS}\right\rangle$$
$$+ \frac{1}{2}\int\int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_1 d\mathbf{r}_2 + E_{xc}[\rho]$$

(2.45)

Using this expression, if we knew the Kohn-Sham orbitals and the exchange-correlation functional, we could calculate the energy from the density. Because the density can be obtained from the Kohn-Sham orbitals χ_i^{KS} using (2.39), the problem is how to find these orbitals. If we use the variational Hohenberg and Kohn theorem, with the restriction of orthonormal orbitals, we can find the equations that must fulfill the orbitals which minimize (2.45) or:

$$\frac{\partial E[\rho(\mathbf{r})]}{\partial \rho(\mathbf{r})} = 0$$

(2.46)

The procedure is similar to obtaining the Hartree-Fock equations. In this case, the resulting eigenvalue system is known as the Kohn and Sham set of equations:

$$\left[-\frac{1}{2}\nabla_1^2 - \sum_{\alpha}\frac{Z_{\alpha}}{r_{1\alpha}} + \int \frac{\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_2 + v_{xc}(1)\right]\chi_i^{KS} = \varepsilon_i^{KS}\chi_i^{KS}$$

(2.47)

It can be seen that this expression is equivalent to equation (2.38), where $\hat{h}_i^{KS}$ is explicitly defined. The function v_{xc}, called the exchange-correlation potential, is the functional derivative of E_{xc} defined as:

$$v_{xc}(\mathbf{r}) = \frac{\delta E_{xc}[\rho(\mathbf{r})]}{\delta \rho(\mathbf{r})}$$

(2.48)

It is interesting to remark that the Kohn and Sham orbitals are defined for the non interacting electron reference system, so they do not have any other physical interpretation than allowing us to calculate the electron density. The wave function of the system is not a spin-orbital Slater determinant, in fact in DFT there is no wave function.

Which is the form of the exchange-correlation functional? Originally, Hohenberg and Kohn suggested this expression:

$$E_{xc}[\rho] = \int \rho(\mathbf{r})\varepsilon_{xc}(\rho)d\mathbf{r} = \int \rho(\mathbf{r})\varepsilon_x(\rho)d\mathbf{r} + \int \rho(\mathbf{r})\varepsilon_c(\rho)d\mathbf{r} \qquad (2.49)$$

where $\varepsilon_x(\rho)$ and $\varepsilon_c(\rho)$ are the exchange and correlation functionals, respectively. $\varepsilon_{xc}(\rho)$ is the exchange energy ε_x plus the correlation energy ε_c corresponding to an electron immersed into a homogeneous and neutral electron gas of density ρ, which in this case, it can be shown that:[7]

$$\varepsilon_x(\rho) = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}[\rho(\mathbf{r})]^{1/3} \qquad (2.50)$$

The correlation energy $\varepsilon_c(\rho)$ can be described by a function obtained by Vosko, Wilk and Nusair in the eighties, adjusting some parameters from Monte Carlo simulations.[7] The exchange-correlation functional obtained by combining these contributions is called Local Density Approximation (LDA) and Local Spin Density Approximation (LSDA), for the open shell case. This functional is useful when the density changes slowly with position, because it is generated from a uniform electron gas. Although this requirement is not valid for real systems in general, the LDA approximation allows us to obtain accurate molecular properties, such us geometries, vibrational frequencies or dipole moments, if we compare with Hartree-Fock or other methods that include correlation effects. However, it usually yields overestimated binding energies.

How does one find the Kohn and Sham orbitals? The key to do that is to express them as a linear combination of known basis functions $\left|\phi_\mu\right\rangle$, as denoted in (2.28). This approach allows us to find a system of equations similar to the Roothaan equations:

$$\sum_{\mu=1}^{K} C_{\mu i}\left(\left\langle\phi_\nu\left|\hat{h}^{KS}\right|\phi_\mu\right\rangle - \varepsilon_i\left\langle\phi_\nu\middle|\phi_\mu\right\rangle\right) \qquad \nu = 1, 2, .., K \qquad (2.51)$$

or in matrix form:

$$\mathbf{h}^{ks}\mathbf{C} = \mathbf{SC}\boldsymbol{\varepsilon} \qquad (2.52)$$

where $\mathbf{h}^{KS}$ matrix contains the elements $h_{\nu\mu}^{KS}$:

$$h_{\nu\mu}^{KS} = \left\langle\phi_\nu\left|\hat{h}^{KS}\right|\phi_\mu\right\rangle = T_{\nu\mu} + V_{\nu\mu}^{ne} + V_{\nu\mu}^{ee} + V_{\nu\mu}^{xc} \qquad (2.53)$$

with

$$T_{\nu\mu} = -\frac{1}{2}\int \phi_\nu(\mathbf{r}_1)\nabla_1^2\phi_\mu(\mathbf{r}_1)d\mathbf{r}_1 \tag{2.54}$$

$$V_{\nu\mu}^{ne} = -\sum_\alpha Z_\alpha \int \frac{\phi_\nu(\mathbf{r}_1)\phi_\mu(\mathbf{r}_1)}{r_{1\alpha}}d\mathbf{r}_1 \tag{2.55}$$

$$V_{\nu\mu}^{ee} = \int\int \frac{\rho(\mathbf{r}_2)\phi_\nu(\mathbf{r}_1)\phi_\mu(\mathbf{r}_1)}{r_{12}}d\mathbf{r}_1 d\mathbf{r}_2 \tag{2.56}$$

$$V_{\nu\mu}^{xc} = \int v_{xc}[\rho(\mathbf{r}_1)]\phi_\nu(\mathbf{r}_1)\phi_\mu(\mathbf{r}_1)d\mathbf{r}_1 \tag{2.57}$$

and $\mathbf{C}$ the matrix defined by the expansion coefficients. From a secular equation similar to (2.32) one can calculate $\boldsymbol{\varepsilon}$ and then diagonalize (2.52). The $\mathbf{h}^{KS}$ matrix elements are density dependent, and therefore $\mathbf{C}$ dependent, thus the resolution must be done in an iterative way. One starts with an initial set of coefficients that are used to calculate ρ with (2.39), and then the density is used to construct $\mathbf{h}^{KS}$, and finally the relationship (2.52) is solved to find a new set of coefficients $C_{\nu\mu}$. This procedure is repeated until the matrix $\mathbf{C}$ is self-consistently converged.

If we re write the density in (2.39) as a function of the basis functions $\{|\phi_i\rangle\}$ we have:

$$\rho(\mathbf{r}) = \sum_{i=1}^N |\chi_i(\mathbf{r})|^2 = \sum_{i=1}^N \sum_{\mu\nu} c_{i\mu}c_{i\nu}\phi_\mu(\mathbf{r})\phi_\nu(\mathbf{r}) = \sum_{\mu\nu} P_{\mu\nu}\phi_\mu(\mathbf{r})\phi_\nu(\mathbf{r}) \tag{2.58}$$

where we have introduced $P_{\mu\nu}$, the density matrix ($\mathbf{P}$) elements:

$$P_{\mu\nu} = \sum_{i=1}^N c_{i\mu}c_{i\nu} \qquad \mu,\nu = 1,2,...,K \tag{2.59}$$

Then, the total energy can be calculated as:

$$E = \sum_{\nu\mu} P_{\nu\mu}h_{\nu\mu}^{KS} - \frac{1}{2}\sum_{\nu\mu} P_{\nu\mu}V_{\nu\mu}^{ee} - \int \rho(\mathbf{r}_1)v_{xc}(\mathbf{r}_1)d\mathbf{r}_1 + \int \rho(\mathbf{r}_1)\varepsilon_{xc}(\mathbf{r}_1)d\mathbf{r}_1 \tag{2.60}$$

which is equivalent to equation (2.45).

The accuracy of DFT calculations is limited by the quality of the exchange-correlation functional. The principal limitation of the LDA and LSDA approximations is related to the uniform electron gas model, which is not accurate for describing large changes in ρ. In order to overcome this problem, functionals which incorporate an explicit dependency with the density gradient $\nabla\rho$ have been proposed. They are called gradient corrected or GGA (Generalized Gradient Approximation) functionals. Usually, the exchange and correlation contributions are modeled separately such that $E_{xc} = E_x + E_c$. Some models are only based in theoretical considerations, such as the Perdew and Wang (PW86) exchange functional,[8] or the PBE exchange-correlation one, proposed

by Perdew, Burke and Ernzerhof.[9] Other functionals also include terms in order to reproduce experimental results, such as the Becke exchange functional (B88 or B):[10]

$$E_x^{B88} = E_x^{LSDA} - b \sum_{\sigma=\alpha,\beta} \int \frac{\rho_\sigma^{-4/3} \left| \nabla \rho_\sigma \right|^2}{1 + 6b \left| \nabla \rho_\sigma \right| / \rho_\sigma^{4/3} \sinh^{-1} \left[\left| \nabla \rho_\sigma \right| / \rho_\sigma^{4/3} \right]} \tag{2.61}$$

where σ is the spin index and b is an empiric parameter chosen to reproduce the Hartree-Fock exchange energy of several atoms. Among the most popular correlation functionals we can mention the Lee, Yang and Parr (LYP),[11] the Perdew (P86),[12] or the Perdew and Wang (PW91)[13] functionals. Given the fact that the exchange and correlation contributions are constructed independently, in a DFT calculation is valid to combine any E_c^{GGA} functional with any other E_x^{GGA}.

Because the Hartree-Fock approximation includes the exact exchange as given in equation (2.22) this contribution has been added into DFT methods to generate the so called hybrid functionals. In these functionals, the Hartree-Fock exchange (E_x^{HF}) is combined with some of the GGA exchange and correlation functionals. For example, the B3LYP[14] hybrid functional, which has been widely used, has the form:

$$E_{xc}^{B3LYP} = (1 - a_o - a_x)E_{xc}^{LSDA} + a_o E_x^{exact} + a_x E_x^{B88} + (1 - a_c)E_c^{VWN} + a_c E_c^{LYP} \tag{2.62}$$

where E_c^{VWN} is the Vosko, Wilk and Nusair correlation functional.[7] The parameters a_0, a_x and a_c (0.20, 0.72 and 0.81, respectively) were chosen in order to optimize experimental atomization energies, ionization potentials and proton affinities of several atoms and molecules.

The SIESTA Method

In this section we will present the very efficient code for quantum mechanical calculations mostly used throughout this thesis, which is based on a very efficient implementation of DFT based on numerical basis sets called SIESTA (Spanish Initiative for Electronic Simulation of Thousands of Atoms).[15,16] The most remarkable feature of SIESTA is the use of flexible basis sets consisting of linear combinations of finite atomic orbitals defined in a real space grid.[17,18] The program has been optimized to yield order N scaling for large systems. Moreover, the program is extremely fast and efficient in comparison with conventional Gaussian or plane waves schemes for medium size species. The algorithms perform all the required calculations numerically, so that basis functions are not constrained to any analytical functional form, and the matching of the radial wave function to the core region, described by pseudopotentials, is inexpensive. In particular, it was found to be convenient to implement the so-called pseudoatomic orbitals (PAOs), which are the eigenfunctions of the atomic pseudopotential problem confined in an infinite potential sphere.[19] The confinement originates finite range PAOs with higher kinetic energies than the free eigenfunctions. By keeping this increment in the energy at the same amount for all the PAOs, the

basis is obtained in a balanced way. This is done by choosing appropriately the pseudoatomic orbital energy shift (ΔE_{PAO}), which specifies the energy increase of the orbitals due to confinement. Split valence bases may be generated by combining suitable Gaussian (numerical) orbitals on top of the minimal basis described above. For typical medium size molecules, values of ΔE_{PAO} from 100 to 200 meV give us a similar accuracy as that obtained with standard Gaussian basis set calculations.[18] For transition metals complexes containing Fe, Cu or Pt, among others, it is recommended to set the energy shift to lower values such as 20-30 meV.[15] In the program SIESTA the PBE functional[9] as well as LDA and LSDA functionals are available.

The nuclei and inner electrons are represented by norm conserving pseudopotentials to avoid the computation of core states, smoothing at the same time the valence charge density and therefore relaxing the quality requirements on the grid. Within the nonlocal pseudopotential approximation, the Kohn-Sham Hamiltonian may be written:

$$H = T + \sum_{\alpha}\left[V_{\alpha}^{local} + V_{\alpha}^{nl}\right] + V^{H}(\mathbf{r}) + V^{xc}(\mathbf{r}) \tag{2.63}$$

where α is an atom index, T is the kinetic energy operator, and V^{H} and V^{xc} are the Hartree (related to the electron repulsion, $V^{H}(\mathbf{r}) = \int \dfrac{\rho(\mathbf{r}_1)}{|\mathbf{r}-\mathbf{r}_1|}d\mathbf{r}_1$) and exchange-correlation potentials, respectively. The pseudopotential is split into two contributions: a local part V_{α}^{local} independent of the angular momentum *lm* of the core electrons, determining the long-range interaction; and a nonlocal part V_{α}^{nl} that operates selectively on the valence electrons, according to their angular momenta.

$$\hat{V_{\alpha}^{nl}} = \sum_{lm}\left|\phi_{\alpha lm}^{KB}\right\rangle\varepsilon_{\alpha l}^{KB}\left\langle\phi_{\alpha lm}^{KB}\right| \tag{2.64}$$

with

$$\phi_{\alpha lm}^{KB}(\mathbf{r}) = \left[V_{\alpha l}(\mathbf{r}) - V_{\alpha}^{local}(\mathbf{r})\right]\psi_{\alpha lm}(\mathbf{r}) \tag{2.65}$$

$$\varepsilon_{\alpha l}^{KB} = \left\langle\psi_{\alpha lm}\left|V_{\alpha l} - V_{\alpha}^{local}\right|\psi_{\alpha lm}\right\rangle^{-1} \tag{2.66}$$

where $\psi_{\alpha lm}$ are the atomic pseudoorbitals, for instance the eigenfunctions of the radial potentials $V_{\alpha l}(\mathbf{r})$. The terms $\phi_{\alpha l}^{KB}$ are known as the Kleinman-Bylander projectors.[20] Their use in V_{α}^{nl} allow us to discriminate the core effect over different angular momentum electrons: the non-local potential acting on the valence electrons comes mainly from the inner electrons with the same angular momentum. Moreover, electrons with different *lm* than the core ones will only see the

local potential. Note that since $V_{\alpha l}(\mathbf{r}) - V_{\alpha}^{local}(\mathbf{r})$ is nonzero only inside a small radius (r_{α}^{core}), V_{α}^{nl} will be zero beyond this radius, as expected for a short range operator.

If we call ρ_{α}^{atom} the charge density of the isolated atoms and,

$$\rho^{na}(\mathbf{r}) = \sum_{\alpha} \rho_{\alpha}^{atom}\left(|\mathbf{r} - \mathbf{r}_{\alpha}|\right) \tag{2.67}$$

then, it is possible to define:

$$\delta\rho(\mathbf{r}) = \rho(\mathbf{r}) - \rho^{na}(\mathbf{r}) \tag{2.68}$$

as the difference between the self-consistent and the non perturbed electronic density. $\delta\rho(\mathbf{r})$ represents a small value against $\rho(\mathbf{r})$, increasing the precision of numerical integrals and allowing us to rewrite the Hamiltonian in the following way:[16]

$$H = T + \sum_{\alpha} V_{\alpha}^{nl} + \sum_{\alpha} V_{\alpha}^{na}(\mathbf{r}) + \delta V^{H}(\mathbf{r}) + V^{xc}(\mathbf{r}) \tag{2.69}$$

with δV^{H} the potential associated to $\delta\rho$. The long range of V_{α}^{local} is eliminated by screening it with V_{α}^{atom}, created by the atomic electron density ρ_{α}^{atom}, to produce a short range "neutral atom" potential $V_{\alpha}^{na} \equiv V_{\alpha}^{local} + V_{\alpha}^{atom}$,

$$V_{\alpha}^{na}(\mathbf{r}) = V_{\alpha}^{local}(\mathbf{r}) + \int \rho^{na}(\mathbf{r}_{1})r_{\alpha 1}^{-1}d\mathbf{r}_{1} \tag{2.70}$$

Since the atomic basis orbitals are zero beyond the radius r_{α}^{core}, V_{α}^{na} is also zero beyond this radius.

Matrix elements corresponding to the two former terms on the right side in the equation (2.69) involve two center integrals evaluated once in reciprocal space at the beginning of the iterative process. The other terms are obtained by numerical integration in real space. The self-consistent structure of the program is as follows:

1) Choice of the basis functions $\{|\phi_i\rangle\}$, and calculation of the non self-consistent matrix elements, T and V_{α}^{nl}.

2) Initial guess of the orbital coefficients C_{ij}, (for instance assuming that $\rho = \rho^{na}$).

3) Obtaining the charge distribution from the density matrix, $\rho = \sum_{\mu\nu} P_{\mu\nu}\phi_{\mu}\phi_{\nu}$, and the electronic density difference: $\delta\rho = \rho - \rho^{na}$.

4) Obtaining δV^{H}, the Hartree potential generated by $\delta\rho$, by inverting the Poisson equation ($\nabla^2 V = -\rho / \varepsilon_0$) in the reciprocal space.

5) Calculation of V^{xc} in the three dimensional grid.

6) Calculation of the integrals (real space numerical integration) associated to the matrix elements of $V_\alpha^{na}(\mathbf{r})$, $\delta V^H(\mathbf{r})$ and $V^{xc}(\mathbf{r})$.

7) Solution of the eigenvalue problem, to find a new set of coefficients C_{ij}.

8) Construction of the new density matrix $\mathbf{P}^{n+1}$, corresponding to the actual C_{ij}, comparison with $\mathbf{P}^n$, and determination if self-consistency has been achieved. If not, return to item 3.

9) Calculation of the energy and any other electronic property.

In this context, the expression for the Kohn-Sham energy may be written as:

$$E = \sum_{\mu\nu} P_{\mu\nu} H_{\mu\nu} - \frac{1}{2}\int V^H(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} + \int\left[\varepsilon^{xc}(\mathbf{r}) - V^{xc}(\mathbf{r})\right]\rho(\mathbf{r})d\mathbf{r} + \frac{1}{2}\sum_{\alpha\beta}\frac{Z_\alpha Z_\beta}{r_{\alpha\beta}} \quad (2.71)$$

where $P_{\mu\nu}$ and $H_{\mu\nu}$ are the density and the Hamiltonian matrix elements, Z_α the charge of the core, and $\varepsilon^{xc}(\mathbf{r})\rho(\mathbf{r})$ the exchange correlation energy density.

In order to avoid the long-range interactions of the last term, we construct from the local pseudopotential $V_\alpha^{local}(\mathbf{r})$, a diffuse ionic charge $\rho_\alpha^{local}(\mathbf{r})$, whose electrostatic potential is equal to $V_\alpha^{local}(\mathbf{r})$ and its integral is the core charge, $\int\rho_\alpha^{local}(\mathbf{r})d\mathbf{r} = Z_\alpha$. By doing this the core charges are not point charges, and $\rho_\alpha^{local}(\mathbf{r})$ can be thought as a core charge density (Gaussian like) defined over the real-space grid for each atom. This facilitates the calculation of the nuclear-nuclear and nuclear-electron interactions in the grid.

The Hartree potential $\delta V^H(\mathbf{r})$ is calculated from the self-consistent $\delta\rho(\mathbf{r})$ by inverting the Poisson equation in the reciprocal space using the standard fast Fourier transform method. The self-consistent potential is thus obtained as a numerical function in real space, avoiding the cost of evaluating the two electrons and four center integrals usually associated with the Coulombic terms. Since the potential and the electronic density are transformed back and forth between the real and the reciprocal space, they are periodic functions suitable for the treatment of solids and extended phases. Simulations in molecular systems may be performed as well, providing the size of the box is large enough to avoid self-interaction effects. Regarding the grid precision, it is defined by choosing the energy of the higher frequency plane wave which is possible to include inside the grid.[15,21] This energy is called cut-off energy (E_{cut}), and a high value of 150-200 Rydbergs ensures a quite precise grid.

Finally, the forces over the atoms are obtained by direct differentiation of equation (2.71) with respect to the atomic positions.

Classical Mechanics Methods

Many of the problems that we would like to tackle in molecular modeling are unfortunately too complex to be described by quantum mechanics due to the large expense of the associated calculations. However, in many situations simpler models may provide reasonable descriptions and it would suffice to obtain answers to many interesting chemical problems. Nowadays, classical mechanics based models allow us to perform computer simulations on systems containing a large number of atoms (up to several thousands) very efficiently. These methods are adequate for describing phenomena in which there are neither bond breaking/forming processes nor electronic excitations. For that reason, they have become the method of choice in the scientific community to study many biomolecular problems.

Classical Force Fields

Classical force fields (also known as molecular mechanics) ignore the electronic motions and calculate the energy of a system as a function of the nuclear positions only. However, molecular mechanics cannot of course provide description of properties that depend upon the electronic distribution in a molecule. The fact that molecular mechanics works at all is due to the validity of several assumptions. The first of these is the Born-Oppenheimer approximation, which allows us to define the potential energy surface (potential energy as a function of the nuclear coordinates). Molecular mechanics is based upon a rather simple model in which the energy is decomposed in contributions of bond stretching, angular bending, rotations about single bonds, electrostatic and non-bonding interactions. Despite their simplicity, these models have allowed us to understand and predict a large variety of phenomena.[22] Presently, there are several available classical force fields such a MM3,[23] CHARMM,[24] GROMOS[25] and AMBER.[26] These potentials are included inside the codes which allow us to simulate very complex systems. Transferability is a key attribute of a force field, for it enables a set of parameters developed and tested on a relatively small number of cases to be applied to a much wider range of problems. In this thesis, we have focused primarily in the AMBER force field parameterization. The AMBER force field could be described as 'minimalist' in its functional form, with the bonds and angles represented by a simple diagonal harmonic expression, the dihedral angles energies represented with a simple set of parameters, often only specified by the two central atoms, the van der Waals (VDW) interactions represented by a 6-12 Lennard-Jones potential, and the electrostatic interactions modeled by a Coulombic interaction of atom-centered point charges. The VDW and electrostatic interactions are only calculated between atoms in different molecules or for atoms in the same molecule separated by at least three bonds. Those non-bonded interactions separated by exactly three bonds (1-4 interactions) are reduced by the application of a scale factor. Then, the potential energy, which is a function of the positions of the N atoms, could be written as:

$$E(\mathbf{r}^N) = \sum_{bonds} K_b \left(b - b_0\right)^2 + \sum_{angles} K_\theta \left(\theta - \theta_0\right)^2 +$$

$$\sum_{torsions} \frac{V_n}{2} \left[1 + \cos(n\phi - \gamma)\right] + \sum_{i}^{N} \sum_{j>i}^{N} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right]$$

(2.72)

This force field allows us to model the conformational energies and the intra-molecular interactions of proteins and nucleic acids, and other molecules with similar functional groups.[26] The considered terms inside this parameterization are the following:

1) **Bond Stretching**: The potential energy curve for a typical bond is usually modeled using the Hooke's law formula. In equation (2.72) b_0 is the equilibrium bond distance of the particular bond considered, and K_b represents the force constant of the harmonic oscillator, related to the energy required to cause a bond to deviate from its equilibrium value. A true bond stretching potential is not harmonic but has a similar shape, which means that the average length of the bond in a vibrating molecule will deviate from the equilibrium value for the hypothetical motionless state. Given the fact that this kind of potential does not allow the bond forming or breaking, systems involving a chemical reaction could not be modeled.

2) **Angle bending**: The deviation of angles from their reference values is also frequently described using a Hooke's law or harmonic potential. The contribution of each angle is characterized by a reference value θ_0, and a force constant K_θ, which represents a measure of how soft, is the associated vibrational mode. Rather less energy is required to distort an angle away from equilibrium than to stretch or compress a bond, so the force constants are typically smaller.

3) **Torsions**: The third term of equation (2.72) refers to the torsion energies of dihedral angles, where V_n is often referred to as the barrier height. However, there are other terms that could contribute to the barrier as a bond is rotated such as the electrostatic and VDW. n is the multiplicity; its value gives the number of minimum points in the function that the dihedral angle has when it is rotated from 0 to 360°. γ is the angle phase and determines where the torsion angle passes through its minimum value.

4) **VDW and electrostatic interactions**: Independent molecules and atoms interact through non-bonded forces, which also play an important role in determining the structure of individual molecular species. The non-bonded interactions do not depend upon a specific bonding relationship between atoms. They are through-space interactions and are usually modeled as a function of some inverse power of the distance. The non-bonded terms, given by the last term in equation (2.72), are usually considered in two groups, the electrostatic and VDW interactions.

Electronegative elements attract electrons more than less electronegative elements, giving rise to an unequal distribution of charges in a molecule. This charge distribution can be represented as an arrangement of fractional point charges throughout the molecule. These charges are designed to reproduce the electrostatic properties of the molecule and are restricted to the nuclear centers.

Then, the electrostatic interaction between two molecules is calculated as a sum of interactions between these pairs of point or atomic charges using the Coulomb's law.

Electrostatic interactions cannot account for all the non-bonded interactions in a system, because, for example, in a rare gas, where all the multipole moments are zero, there clearly must be interactions between atoms. The way the interaction energy varies with the separation is as follows: the energy is zero at infinite distance. As the separation is reduced, the energy decreases (attractive forces), passing through a minimum (equilibrium distance), and then rapidly increases as the separation decreases further (repulsive forces). These kinds of interactions are known as VDW interactions, which include the dispersive (attractive) long-range and repulsive short-range interactions. The VDW potential is added to the electrostatic one in order to determine the inter-molecular interactions and have a particular role in intra-molecular interactions for atoms separated for more than three bonds. The best known of the VDW potential functions is the Lennard-Jones 12-6 function. The 12-6 Lennard-Jones potential contains two adjustable parameters for each pair of interacting atoms: the separation at which the energy passes through a minimum R_{ij}^*, and the VDW well depth ε_{ij} (minimum energy value). These parameters are tabulated by atom, so the values used in equation (2.72) for the interaction between atom i and atom j are $R_{ij}^* = (R_i^* + R_j^*)$ and $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$. Note that $A_{ij} = \varepsilon_{ij}(R_{ij}^*)^{12}$ and $B_{ij} = 2\varepsilon_{ij}(R_{ij}^*)^6$. Lennard-Jones potentials allow for the description of both long-range dispersive and short-range repulsive interactions.

Derivatives of the Energy Function

Many molecular modeling techniques that use force-field models require the derivatives of the energy (i.e. the force) to be calculated with respect to the coordinates. A molecular mechanical energy is usually expressed as a combination of internal coordinates of the system; function of the atomic positions expressed in terms of Cartesian coordinates. The calculation of derivatives with respect to the atomic coordinates for an energy functional that depends upon the separation between two atoms can be written:

$$\frac{\partial E}{\partial \xi_i} = \frac{\partial E}{\partial r_{ij}} \frac{\partial r_{ij}}{\partial \xi_i} = \frac{\partial E}{\partial r_{ij}} \frac{(\xi_i - \xi_j)}{r_{ij}}$$

(2.73)

where $r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$ is the distance between atom i and j and ξ denotes a generic coordinate (x, y, or z). Analytical expressions for the terms involved in the force field whose energy form is given in equation (2.72) are available. Then the force in the ξ direction acting on atom i could be calculated as $\mathbf{f}_{\xi_i} = -\partial E / \partial \xi_i$.

Truncating the Potential

It is interesting to discus how the potential could be truncated in order to decrease the computational cost. The most time-consuming part of a molecular mechanics simulation is the calculation of the non-bonded energies and/or forces. The number of bond stretching, angle bending and torsional terms in a force field model are all proportional to the number of atoms but the number of non-bonded terms that need to be evaluated increases as the square of the number of atoms (for a pairwise model) and is thus of order N^2. In principle the non-bonded interactions are calculated between every pair of atoms in the system. However, for many interaction models this is not justified and the most popular way to deal with the non-bonded interactions is to use a non-bonded cutoff. When a cutoff is employed, the interactions between all pairs of atoms that are further apart than the cutoff value are set to zero. In simulations with both electrostatic and VDW non-bonded interactions, a value of at least 10 Å is generally recommended. By itself, the use of a cutoff may not dramatically reduce the time taken to compute the number of non-bonded interactions, because we would still have to calculate the distance between every pair of atoms simply to decide whether they are close enough to calculate their interaction energy. Calculating all the $N(N-1)$ distances takes almost as much time as calculating the energy itself. However, if we knew which atoms to include in the non-bonded calculation, then it would be possible to identify directly each atom's neighbors without having to calculate the distances to all other atoms. The non-bonded neighbor list stores for each atom, all atoms within the cutoff distance, together with all atoms that are slightly further away than the cutoff distance. The neighbor list is then updated at regular intervals throughout the simulation. An update frequency of 20 fs is common. The distance used to calculate each atom's neighbors should by larger than the actual non-bonded cutoff distance.

A cutoff introduces a discontinuity in both the potential energy and the force near the cutoff value. This creates problems, especially in molecular dynamics simulations where energy conservation is required. There are several ways in which the effect of this discontinuity can be counteracted. One of the most common approaches is to use a switching function. A switching function is a polynomial function of the distance by which the potential energy function is multiplied $E'(\mathbf{r}) = E(\mathbf{r})S(\mathbf{r})$. It is convenient to gradually taper the potential between two cutoff values. The potential takes its usual value until the lower cutoff distance. Between the lower (r_l) and upper (r_u) cutoff distance the potential is multiplied by the switching function, which takes the value 1 at the lower cutoff and 0 at the upper one. Finally, the potential is exactly zero for distances larger than the upper cutoff. An acceptable switching function smoothly changes from 1 to 0 between r_l and r_u, ensuring that the first derivative is zero at the endpoints, so the forces approaches to zero smoothly, and a continuous second derivative, so the integration algorithm works properly. Then, the switching function can be written as:

$$S(\mathbf{r}) = 1 - 10\left[\frac{\mathbf{r} - r_l}{r_u - r_l}\right]^3 + 15\left[\frac{\mathbf{r} - r_l}{r_u - r_l}\right]^4 - 6\left[\frac{\mathbf{r} - r_l}{r_u - r_l}\right]^5 \tag{2.74}$$

Throughout the thesis, all calculations were performed using a switching function where r_l, r_u and the non-bonded cutoff distances were 8, 10 and 12 Å, respectively.

Hybrid QM-MM Methodologies

Hybrid quantum mechanical-molecular mechanical (QM-MM) schemes are adequate for the investigation of reactions in complex environments.[27,28] They are particularly suited to study enzyme active sites or solutes in condensed phases. The method combines an electronic structure description of the solute (QM subsystem) with a less expensive molecular mechanical treatment of the environment (MM subsystem), by coupling these subsystems through a hybrid Hamiltonian H_{QM-MM}.

In general we denote the total Hamiltonian H_{TOT}, which operates over the whole system wave function Ψ:

$$H_{TOT}\Psi(\mathbf{r}, \mathbf{R}, \boldsymbol{\tau}) = E_{TOT}(\mathbf{R}, \boldsymbol{\tau})\Psi(\mathbf{r}, \mathbf{R}, \boldsymbol{\tau}) \tag{2.75}$$

where $\boldsymbol{\tau}$ represents the classical atoms coordinates, $\mathbf{R}$ the quantum nucleus coordinates and $\mathbf{r}$ the electron ones. This Hamiltonian involves three terms: a quantum Hamiltonian, as defined in equation (2.18) or (2.37), a classical Hamiltonian and a coupling term:

$$H_{TOT} = H_{QM} + H_{MM} + H_{QM-MM} \tag{2.76}$$

and the same is valid for the total energy:

$$E_{TOT} = E_{QM} + E_{MM} + E_{QM-MM} \tag{2.77}$$

In the case of a hybrid method adapted to DFT, as is the case of this thesis, E_{QM} is given in (2.45).

E_{MM} is represented by an appropriate classical potential, as given in (2.72). For the treatment of biological macromolecular environments, we implemented the Wang et al. force field parameterization.[29] Finally, the coupling term E_{QM-MM} may be decomposed into three contributions:

$$E_{QM-MM} = \sum_{i=1}^{C} q_i \int \frac{\rho(\mathbf{r})}{|\mathbf{r} - \boldsymbol{\tau}_i|} dr + \sum_{i=1}^{C}\sum_{\alpha=1}^{Q} \frac{q_i Z_\alpha}{|\mathbf{R}_\alpha - \boldsymbol{\tau}_i|} + E_{QM-MM}^{LJ} \tag{2.78}$$

where C is the number of atoms in the classical region, whose charges q_i are determined by the force field chosen to model the environment and α is an index on the Q nuclei inside the quantum subsystem with core charges Z_α. The first term on the right represents the electrostatic

interaction between the electrons and the classical charges and is obtained by numerical integration over the grid. The next term stands for the electrostatic interactions between the nuclei in the quantum subsystem and the classical point charges. Finally, E_{QM-MM}^{LJ} models the van der Waals interactions between the atoms in the quantum and classical regions through a 6-12 Lennard-Jones potential. The Lennard-Jones parameters for the QM atoms have been also taken from the Wang et al. force field parametrization.[29] A complete description of these methodologies, covering theory and applications can be found in reference 28.

Implementation in SIESTA

Let us now consider how the computational implementation of the QM-MM formalism is done inside the SIESTA method, which is one of the main goals of this thesis. In the numeric SIESTA method, the potentials and density are defined for the L subdivisions of the grid (n subindex). Thus, the environment affects the charge density in a self-consistent fashion by the addition of the external point charge potential to the Hartree potential δV^H :

$$\delta V^H(\mathbf{r}_n) = \delta V^H(\mathbf{r}_n) + V^{EXT}(\mathbf{r}_n) \qquad n = 1, 2, ..., L \tag{2.79}$$

where:

$$V^{EXT}(\mathbf{r}_n) = \sum_{i=1}^{C} \frac{q_i}{\left|\boldsymbol{\tau}_i - \mathbf{r}_n\right|} \qquad n = 1, 2, ..., L \tag{2.80}$$

Since the potential δV^H is evaluated numerically, the computational expense associated with this sum is proportional to C times the number of elements in the grid (L). Because of this, the incorporation of a classical potential is more expensive than in the case of a Gaussian scheme, where the integrals are performed analytically.[30] Still, the cost of the numerical QM-MM algorithm remains relatively small in comparison with the Kohn-Sham diagonalization procedure. The sum given in equation (2.80) is performed in a step preceding the calculation of the Hamiltonian and other potential-dependent matrix elements, hence introducing the effect of the external classical potential in a self-consistent fashion in the electronic wave function and in the forces in a direct way.

The potential due to the classical charges, while proportional to the inverse of the distance, may become too steep at certain points inside the grid, affecting the numerical integrations. To preserve the smoothness of $\delta V^H(\mathbf{r})$, the classical potential is truncated at short distances so that:

$$V^{EXT}(\mathbf{r}_n) = \begin{cases} \sum_{i=1}^{C} \dfrac{q_i}{|\boldsymbol{\tau}_i - \mathbf{r}_n|} & |\boldsymbol{\tau}_i - \mathbf{r}_n| > R_{cut} \\ \sum_{i=1}^{C} \dfrac{q_i}{R_{cut}} & |\boldsymbol{\tau}_i - \mathbf{r}_n| \leq R_{cut} \end{cases} \qquad n = 1, 2, ..., L \qquad (2.81)$$

It has been found that a value of R_{cut} close to 0.2 or 0.3 Å is enough to avoid numerical problems without significantly affecting geometries or energies.[31] Notice that the point charges may be placed as far from the QM region as desired, regardless of the boundaries of the grid. In fact, whether the MM atoms are inside or outside the grid does not make any difference in terms of computational effort, neither if they are scattered along a large region in space. This is not true for the QM part, which size actually determines the size of the grid and therefore the cost of the numerical integrals.

The computational expense associated with the addition of the external potential to δV^H is not inexpensive. For that reason, it has been convenient to truncate to zero the external potential for large distances, diminishing the computational cost of the summation performed in equation (2.80).

The core charges are represented as a diffuse ionic charge $\rho_\alpha^{local}(\mathbf{r})$ (which integrate to Z_α), defined over the real-space grid for each atom. For this reason, the electrostatic interaction between the QM atoms and the partial charges, corresponding to the first two terms in equation (2.78), is calculated as an integral of the total (electronic plus nuclear) density over the grid. If v_n represents the volume of each grid element, this integral can be calculated as the following summation:

$$E_{QM-MM}^{electrostatic} = \sum_{n=1}^{L} \left[\sum_{\alpha=1}^{Q} \rho_\alpha^{atom}\left(|\mathbf{r}_n - \mathbf{R}_\alpha|\right) + \delta\rho(\mathbf{r}_n) + \sum_{\alpha=1}^{Q} \rho_\alpha^{local}\left(|\mathbf{r}_n - \mathbf{R}_\alpha|\right) \right] V^{EXT}(\mathbf{r}_n) v_n \quad (2.82)$$

where the first two terms represent the total electronic density and the last term the nuclear density. Calculating the electrostatic interaction as the product of the total density and the external classical potential is very useful because the former represents a smaller value with respect to the total electronic density, increasing the precision of the numerical integrals. The reason for this, is that the total density includes the contribution of the small term $\delta\rho$, and the sum of ρ_α^{atom} and ρ_α^{local} is almost zero for an isolated atom. This is equivalent to say that a partial charge does not "see" the neutral atom if it is far away enough (for distances longer than r_α^{core}).

Finally, the last term of E_{QM-MM} is given by:

$$E_{QM-MM}^{LJ} = \sum_{i=1}^{C} \sum_{\alpha=1}^{Q} \left(\frac{A_{i\alpha}}{|\boldsymbol{\tau}_i - \mathbf{R}_\alpha|^{12}} - \frac{B_{i\alpha}}{|\boldsymbol{\tau}_i - \mathbf{R}_\alpha|^{6}} \right)$$

(2.83)

where $A_{i\alpha}$ and $B_{i\alpha}$ are taken from the force field parametrization.

Forces on the QM nuclei are obtained by differentiation of equation (2.71) with respect to atomic positions, plus the derivative of the coupling energy E_{QM-MM} :

$$\mathbf{F}_\alpha = -\frac{\partial E_{KS}[\mathbf{R}]}{\partial \mathbf{R}_\alpha} - \frac{\partial E_{QM-MM}[\mathbf{R}, \boldsymbol{\tau}]}{\partial \mathbf{R}_\alpha}$$

(2.84)

E_{KS} depends only on the molecular mechanics subsystem through $V^H(\mathbf{r})$, contained in the second term on the right in (2.71). It can be shown[18] that the differentiation of $\int V^H(\mathbf{r})\rho(\mathbf{r})d\mathbf{r}$ ultimately leads to an expression whose overall dependence on $V^H(\mathbf{r})$ is through a term of the form $\mathrm{Re}\left(\sum_\mu \sum_{v \in \alpha} P_{\mu v} \langle \phi_\mu | V^H(\mathbf{r}) | \phi_v \rangle \right)$, meaning that the action of the MM potential on the forces is implicitly included if the contribution of $V^{EXT}(\mathbf{r})$ is previously added. Regarding the contribution of E_{QM-MM} to the QM forces, the first term in equation (2.78) vanishes upon differentiation with respect to the QM atomic coordinates. The second contribution, the electrostatic force induced in the core by the partial charges, is calculated from the last term in equation (2.82) by numerical interpolation on the grid:

$$\sum_{n=1}^{L} \left[\sum_{\alpha=1}^{Q} \frac{\partial \rho_\alpha^{local}(\mathbf{r}_n)}{\partial \mathbf{R}_\alpha} \right] V^{EXT}(\mathbf{r}_n) v_n$$

(2.85)

The term arising from E_{QM-MM}^{LJ} has the analytical derivative of:

$$\sum_{i=1}^{C} \sum_{\alpha=1}^{Q} \left(\frac{12 A_{i\alpha}}{|\boldsymbol{\tau}_i - \mathbf{R}_\alpha|^{14}} - \frac{6 B_{i\alpha}}{|\boldsymbol{\tau}_i - \mathbf{R}_\alpha|^{8}} \right) (\xi_i - \xi_\alpha)$$

(2.86)

where ξ denotes a generic coordinate (x, y, or z).

On the other hand, forces on the MM atoms are computed as the derivative of E_{MM} and E_{QM-MM} with respect to the classical atomic positions:

$$\mathbf{F}_i = -\frac{\partial E_{MM}[\tau_i]}{\partial \tau_i} - \frac{\partial E_{QM-MM}[\mathbf{R}, \tau_i]}{\partial \tau_i}$$

(2.87)

The first term on the right in the equation above is calculated as the analytical derivative of equation (2.72), as well as the term arising from E_{QM-MM}^{LJ}, equation (2.86). The electrostatic force induced in the MM atoms due to interaction with the QM atoms needs to be evaluated numerically and it is done after self-consistency is achieved. The forces arising from this contribution are obtained by numerical derivation of equation (2.82) with respect to the classical atoms positions. The use of the total density (electronic plus nuclear), which is a smaller value, increases the precision of the numerical integrals. Then, this term has the form:

$$\sum_{i=1}^{C}\sum_{n=1}^{L}\left[\sum_{\alpha=1}^{Q}\rho_{\alpha}^{atom}\left(\left|\mathbf{r}_n-\mathbf{R}_\alpha\right|\right)+\delta\rho(\mathbf{r}_n)+\sum_{\alpha=1}^{Q}\rho_{\alpha}^{local}\left(\left|\mathbf{r}_n-\mathbf{R}_\alpha\right|\right)\right]\frac{q_i}{\left|\boldsymbol{\tau}_i-\mathbf{r}_n\right|^3}v_n\left(\xi_i-\xi_n\right) \quad (2.88)$$

where ξ denotes a generic coordinate (x, y, or z). By means of the forces calculated for each atom, it is possible to perform coordinate relaxations or molecular dynamics simulations as will be described shortly.

Frontier Between Subsystems

It is interesting to discuss what happens if the frontier between both QM and MM subsystem involves the breaking of a covalent bond. This bond between the QM and MM portions of the system can be treated by several approaches.[32,33] In our case, the scaled position link atom method (SPLAM)[34] was adapted to our hybrid code. In this method, the bonds between a carbon atom of the QM subsystem (C_{QM}) and a carbon of the MM subsystem (C_{MM}) are replaced with a carbon-hydrogen bond. This hydrogen link atom (H_{link}) fills the valence of the QM subsystem, and its position is superimposed on the C_{QM}-C_{MM} bond. Forces exerted on the H_{link} are divided in parallel and perpendicular components, the former is added to the C_{MM} and the latter is scaled (by a factor 10) and summed over C_{QM} and C_{MM}, to avoid including a torque. All classical bond terms (bonds, angles, and dihedrals) involving at least one C_{MM} central atom are computed normally and are added to the MM potential. Lennard-Jones interactions between the C_{QM}, H_{link}, and C_{MM} atoms are omitted for atom pairs separated by less than three bonds. The charge of the C_{MM} atom is divided and summed over its C_{MM} neighbors to maintain the total charge unaltered. This method is efficient and has already been used in previous calculations with different systems including biomolecules.[31,34 -35]

Exploring the Potential Energy Surface

For all except the very simplest systems the potential energy is a complicated, multidimensional function of the coordinates. The way in which the energy varies with the coordinates is usually referred to as the potential energy surface. For a system with *N* atoms the energy is thus a

function of *3N - 6* internal coordinates, so it is therefore impossible to visualize the entire energy surface.

Energy Minimization

In molecular modeling we are especially interested in minimum points on the potential energy surface. Minimum energy arrangements of the atoms correspond to stable states of the system; any movement away from a minimum gives a configuration with higher energy. There may be a very large number of minima on the energy surface, and the minimum with the very lowest energy is known as the global energy minimum. To identify those geometries of the system that correspond to minimum points on the potential energy surface we use a minimization algorithm. We may be also interested to know how the system changes from one minimum energy structure to another. The highest point on the pathway between two minima is of special interest and is known as the saddle point, with the rearrangement of the atoms being the transition structure. Both minima and saddle points are stationary points on the potential energy surface, where the first derivative is zero with respect to all coordinates.

The minimization problem can be formally stated as follows: given a function *f* which depends on the independent variables $x_1, x_2, ..., x_i$, find the values of those variables where *f* has a minimum value. At a minimum point the first derivative of the function with respect to each of the variables is zero and the second derivatives are all positive. The function of most interest to us will be the quantum mechanics or molecular mechanics energy with the variables being the coordinates of the atoms. The first-order minimization algorithms (those that only use the first derivatives) that are frequently used are the steepest descend and conjugate gradient methods.[22] These gradually change the coordinates of the atoms as they move the system closer to the minimum point. The starting point for each iteration (*k*) is the molecular configuration obtained from the previous step, which is represented by the multidimensional vector $\mathbf{x}_{k-1}$. For the first iteration the starting point is the initial configuration of the system provided by the user, the vector $\mathbf{x}_1$.

The Steepest Descend Method

This method moves in the direction parallel to the net force, which corresponds to walking straight downhill. For 3*N* coordinates this direction is most conveniently represented by a 3*N*-dimensional unit vector, $\mathbf{s}_k = -\mathbf{g}_k / |\mathbf{g}_k|$, where $\mathbf{g}_k$ is the gradient at the point *k*, a 3*N* vector, each element of which is the partial derivative of the potential energy with respect to the appropriate coordinate, $\partial E / \partial x_i$. Having defined the direction along which to move it is necessary to decide how far to move along the gradient. We can obtain the new coordinates be taking a step of

arbitrary length along the gradient unit vector $\mathbf{s}_k$. The new set of coordinates after step k would then be given by,

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{s}_k \qquad (2.89)$$

where λ_k is the step size. In most applications of the steepest descend method the step size initially has a predetermined value. If the first iteration leads to a reduction in energy, the step size is increased by a multiplicative factor (e.g. 1.2) for the second iteration. When a step produces an increase in energy, it is assumed that the algorithm has leapt across the valley up to the opposite face of the minimum, then the step size is reduced by a multiplicative factor (e.g. 0.5). The step size depends upon the nature of the energy surface.

Conjugate Gradients Minimization

This method produces a set of directions which does not show the oscillatory behavior of the steepest descend method in narrow valleys. In the steepest descend method both the gradients and the direction of successive steps are orthogonal. In conjugate gradients, the gradients at each point are orthogonal but the directions are conjugate. A set of conjugate directions has the property that for a quadratic function of M variables, the minimum will be reached in M steps. The conjugate gradients method moves in a direction $\mathbf{v}_k$ from point $\mathbf{x}_k$ where $\mathbf{v}_k$ is computed from the gradient at the point ant the previous direction vector $\mathbf{v}_{k-1}$:

$$\mathbf{v}_k = -\mathbf{g}_k + \gamma_k \mathbf{v}_{k-1} \qquad (2.90)$$

where γ_k is a scalar constant given by:

$$\gamma_k = \frac{\mathbf{g}_k \cdot \mathbf{g}_k}{\mathbf{g}_{k-1} \cdot \mathbf{g}_{k-1}} \qquad (2.91)$$

In the conjugate gradients method all of the directions and gradients satisfy:

$$\mathbf{g}_i \cdot \mathbf{g}_j = 0$$
$$\mathbf{v}_i \cdot \frac{\partial^2 E}{\partial x_i \partial x_j} \cdot \mathbf{v}_j = 0 \qquad (2.92)$$
$$\mathbf{g}_i \cdot \mathbf{v}_j = 0$$

Clearly, equation (2.90) can only be used from the second step onwards and so the first step is the same as the steepest descend (i.e. in the direction of the gradient). This method is the one implemented within the SIESTA code.

Reaction Pathway Search

Chemists are interested not only in the thermodynamics of a process, but also in its kinetics. Knowledge of the minimum points on an energy surface enables thermodynamics data to be interpreted, but for the kinetics it is necessary to investigate the nature of the energy surface away from the minimum points. In particular, we would like to know how the system changes from one minimum to another. The minimum points on the energy surface may be the reactants and products of a chemical reaction, two different conformations, or two molecules that non-covalently associate, and the term reaction pathway is used to describe the path between two of these minima. As a system moves from one minimum to another, the energy increases to a maximum at the transition structure and then it falls. The energy passes through a maximum for the motion along the pathway, but is a minimum for displacements in all other directions perpendicular to the path.

The conversion of one minimum-energy structure into another may sometimes be represented by a combination of atomic coordinates, called reaction coordinate. In such cases, an approximation to the reaction pathway, called restrained minimization, can be obtained by gradually changing this coordinate, allowing the system to relax at each stage using minimization while keeping the chosen reaction coordinate fixed. The point of higher energy on the path is an approximation of the transition state and the structures generated during the course of the calculation can be considered to represent a sequence of points on the interconversion pathway.

For performing the restrained minimizations, an additional term is added to the potential energy. This term is chosen as:

$$V_R = \frac{1}{2} k \left(\xi - \xi_0 \right)^2 \tag{2.93}$$

where k is an adjustable force constant, ξ is the value of the reaction coordinate, and ξ_0 is the value of the reaction coordinate for a particular configuration. The path is constructed as follows: first, an unrestricted minimization is performed for the reactant or product to generate an initial configuration for the reaction path. The reaction path is mapped out by adding the V_R to the potential energy, varying ξ_0, and performing energy minimizations at each step. The actual energy of each configuration along the reaction path is obtained by subtracting the restraint term from the total energy.

Molecular Dynamics Simulation

The molecular modeler often wants to understand and predict the properties of very complex systems, such as liquids, solids or macromolecules. In such systems, the experimental measurements are made on macroscopic samples that contain large number of atoms with an

enormous number of minima on their energy surfaces. A full quantification of the energy surface of such systems is not possible. Computer simulation methods enable to study such systems by considering small replications of the macroscopic system, generating representative configurations of them in a way that accurate thermodynamics or structural values can be obtained. Simulation techniques also enable the time-dependent behavior of systems to be determined, providing a detailed picture of the way in which it changes from one conformation to another. One of the most used simulation technique is the molecular dynamics method.

Molecular dynamics calculates the 'real' dynamics of the system, from which time averages of properties can be calculated. Sets of atomic positions are derived in sequence by applying the Newton's equation of motion. It is a deterministic method, because the state of the system at any future time can be predicted from its current state. At each step, the energy and the forces on the atoms are computed and combined with the current positions and velocities to generate new ones a short time ahead. The forces are assumed to be constant during the time interval. The atoms are then moved to the new positions, an updated value of energy and forces are computed, and so on. In this way, a molecular dynamics simulation generates a 'trajectory' that describes how the dynamic variables change with time. The trajectory is obtained by solving the differential equations embodied in Newton's second law, $F = ma$:

$$\frac{d^2 x_i}{dt^2} = \frac{F_{x_i}}{m_i}$$

(2.94)

This equation describes the motion of a particle of mass m_i along the coordinate x_i with F_{x_i} being the force on the particle in that direction.

In the quantum mechanics or molecular mechanics interaction potentials, which are of primarily interest when describing the motion atoms or molecules, the force on each particle will change whenever the particle changes its position, or any other interacting particles change position. Under the influence of a continuous potential the motions of all the particles are coupled together, giving rise to a many-body problem that cannot be solved analytically. Under such circumstances, the equations of motion are integrated using a finite difference method. The essential idea of this method is that the integration is broken down into many small stages, each separated in time by a fixed time δt. The total force on each particle in the configuration at time t is calculated as the vector sum of its interactions with other particles. From the force we can determine the accelerations, which are combined with the positions and velocities to calculate them at a time $t + \delta t$. The forces in their new positions are then determined, leading to new positions and velocities at time $t + 2\delta t$, and so on. There are many algorithms for integrating the equations of motion using finite difference methods, but all assume that the positions and dynamic properties (velocities, accelerations, etc.) can be approximated as Taylor series expansions:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2}\delta t^2 \mathbf{a}(t) + ... \tag{2.95}$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \delta t \mathbf{a}(t) + \frac{1}{2}\delta t^2 \mathbf{b}(t) + ... \tag{2.96}$$

$$\mathbf{a}(t+\delta t) = \mathbf{a}(t) + \delta t \mathbf{b}(t) + \frac{1}{2}\delta t^2 \mathbf{c}(t) + ... \tag{2.97}$$

where $\mathbf{v}$ is the velocity (the first derivative of the positions with respect to time), $\mathbf{a}$ is the acceleration (the second derivative), $\mathbf{b}$ is the third derivative, and so on. The Verlet algorithm[36] is probably the most widely used method for integrating the equations of motion in a molecular dynamics simulation. It uses the positions and accelerations at time t, and the positions from the previous step, $\mathbf{r}(t-\delta t)$, to calculate the new positions at $t+\delta t$, $\mathbf{r}(t+\delta t)$. We can write down the following relationships between these quantities and the velocities at time t:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2}\delta t^2 \mathbf{a}(t) + ... \tag{2.98}$$

$$\mathbf{r}(t-\delta t) = \mathbf{r}(t) - \delta t \mathbf{v}(t) + \frac{1}{2}\delta t^2 \mathbf{a}(t) - ... \tag{2.99}$$

Adding these two equations gives

$$\mathbf{r}(t+\delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + \delta t^2 \mathbf{a}(t) \tag{2.100}$$

The velocities do not explicitly appear in the Verlet integration algorithm, however they can be calculated by dividing the difference in positions at times $t-\delta t$ and $t+\delta t$ by $2\delta t$:

$$\mathbf{v}(t) = \frac{\mathbf{r}(t+\delta t) - \mathbf{r}(t-\delta t)}{2\delta t} \tag{2.101}$$

Implementation of the Verlet algorithm is straightforward and the storage requirements are modest. This method is included in both the SIESTA and AMBER packages.

Which is the most appropriate time step to use in a molecular dynamics simulation? If it is too small the trajectory will cover only a limited proportion of the phase space, and if it is too large instabilities may arise in the integration algorithm. The total error is correlated to the time step, with the largest errors arising for the largest time steps, however, with a small time step much more computer time will be required for a given length of calculation. The aim is then to find the correct balance between simulating the correct trajectory and covering the phase space. When simulating molecules, a useful guide is that the time step should be approximately one-tenth the time of the shortest period of motion. In molecules, the highest-frequency vibrations are due to bond stretches, especially those including hydrogen atoms, which vibrate with a period of approximately 10 fs. The requirement that the time step is approximately one order of magnitude smaller than the shortest motion is clearly a severe restriction. One solution to this problem is to

'freeze out' the high-frequency vibrations by constraining the appropriate bonds to their equilibrium values while still permitting the rest of the degrees of freedom to vary under the intramolecular and intermolecular forces present. This enables a longer time step to be used. One way of performing such constraints is applying the SHAKE algorithm.[37] Another solution is to use deuterium atoms (mass 2 a.m.u.) instead of hydrogen throughout the simulation.

Many thermodynamic properties can be calculated from a molecular dynamics simulation. The instantaneous value of the temperature is related to the kinetic energy via the particles' momenta as follows:

$$K = \sum_{i=1}^{N} \frac{|\mathbf{p}_i|^2}{2m_i} = \frac{k_B T}{2}(3N - N_c) \tag{2.102}$$

where $\mathbf{p}_i = m_i \mathbf{v}_i$, N_c is the number of constraints and $3N - N_c$ is the total number of degrees of freedom. There are several reasons why to maintain or control the temperature during a molecular dynamics simulation. A constant temperature simulation may be required if we wish to determine how the behavior of the system changes with temperature and because many chemical processes occur at a constant temperature. Because the temperature is related to the kinetic energy, an obvious way to alter the temperature is thus to scale the velocities. One of the most used methods to maintain the temperature is the Berendsen algorithm,[38] which couples the system to an external bath that is fixed at the desired temperature. The bath acts as a source of thermal energy, supplying or removing heat from the system as appropriate. The velocities are scaled at each step, such that the rate of change of temperature is proportional to the difference in the temperature between the bath and the system:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}\left[T_{bath} - T(t)\right] \tag{2.103}$$

τ is a coupling parameter whose magnitude determines how tightly the bath and the system are coupled together. If it is large, the coupling will be weak, if it is small, the coupling will be strong. The change in temperature between successive time steps is:

$$\Delta T = \frac{\delta t}{\tau}\left[T_{bath} - T(t)\right] \tag{2.104}$$

and the scaling factor for the velocities is thus:

$$\lambda^2 = 1 + \frac{\delta t}{\tau}\left(\frac{T_{bath}}{T(t)} - 1\right) \tag{2.105}$$

Coupling constants generally used vary in order that $\delta t / \tau$ range from 10^{-2} to 10^{-3}. This thermostat is also implemented in both the SIESTA and AMBER packages.

The Monte Carlo Method

In a molecular dynamics simulation the successive configurations of the system are connected in time. This is not the case in a Monte Carlo simulation, where each configuration depends only upon its predecessor and not upon any other of the configurations previously visited. The Monte Carlo method generates configurations randomly and uses a special set of criteria to decide whether or not to accept each new configuration. These criteria ensure that the probability of obtaining a given configuration is equal to its Boltzmann factor, $\exp\left[-E(\mathbf{r}^N)/k_B T\right]$, where $E(\mathbf{r}^N)$ is calculated using the potential energy function. States with a low energy are thus generated with a higher probability than configurations with a higher energy. For each configuration that is accepted the values of the desired properties are calculated and at the end of the calculation the average of these properties is obtained by simply averaging over the number of values calculated, *M*:

$$\langle B \rangle = \frac{1}{M} \sum_{i=1}^{M} B(\mathbf{r}^N) \qquad (2.106)$$

In a Monte Carlo simulation each new configuration of the system may be generated by randomly moving a single or several atoms or molecules. The energy of the new configuration is then calculated using the potential energy function. If the energy of the new configuration is lower than the energy of its predecessor then the new configuration is accepted. If the energy of the new configuration is higher than the energy of its predecessor then the Boltzmann factor of the energy difference is calculated: $\exp\left[-\left(E_{new}(\mathbf{r}^N) - E_{old}(\mathbf{r}^N)\right)/k_B T\right]$. A random number is then generated between 0 and 1 and compared with this Boltzmann factor. If the random number is higher than the Boltzmann factor then the move is rejected and the original configuration is retained for the next iteration; if the random number is lower then the move is accepted and the new configuration becomes the next state. This procedure has the effect of permitting moves to states of higher energy. The smaller the uphill move (i.e. the smaller the value of $E_{new}(\mathbf{r}^N) - E_{old}(\mathbf{r}^N)$) the greater is the probability that the move will be accepted.

Free Energy Calculations

The free energy is often considered to be the most important thermodynamical property in chemistry. The free energy is usually expressed as the Helmholtz function, *A*, or the Gibbs function, *G*. The Helmholtz free energy is appropriate to a system with constant number of particles, temperature and volume (constant NVT), whereas the Gibbs free energy is appropriate to constant number of particles, temperature and pressure (constant NTP). Most experiments are conducted under conditions of constant temperature and pressure, where the Gibbs function is

the appropriate free energy quantity. Unfortunately, the free energy is a difficult quantity to obtain for systems such as liquids or flexible macromolecules that have many minimum energy configurations separated by low-energy barriers. Associated quantities such as the entropy and the chemical potential are also difficult to calculate. The free energy cannot be accurately determined from a 'standard' molecular dynamics simulation, because such simulation does not adequately sample from those regions of phase space that make important contributions to the free energy.

For any property (B) of the system that depends upon the positions and momenta of the N particles, the ensemble average, or expectation value (the average value of the property B over all replications of the ensemble generated by the simulation) can be written as:

$$\langle B \rangle = \iint d\mathbf{p}^N d\mathbf{r}^N B(\mathbf{p}^N, \mathbf{r}^N) \rho(\mathbf{p}^N, \mathbf{r}^N) \tag{2.107}$$

where $\rho(\mathbf{p}^N, \mathbf{r}^N)$ is the probability density of the ensemble, that is, the probability of finding a configuration with momenta $\mathbf{p}^N$ and positions $\mathbf{r}^N$. The ensemble average of the property is then determined by integrating over all possible configurations of the system. In accordance with the 'ergodic hypothesis' the ensemble average is equal to the time average, and the probability density is the familiar Boltzmann distribution:

$$\rho(\mathbf{p}^N, \mathbf{r}^N) = \exp\left[-H(\mathbf{p}^N, \mathbf{r}^N)/k_B T\right]\Big/Q \tag{2.108}$$

where Q, the partition function, is given by:

$$Q = \frac{1}{N! h^{3N}} \iint d\mathbf{p}^N d\mathbf{r}^N \exp\left[-H(\mathbf{p}^N, \mathbf{r}^N)/k_B T\right] \tag{2.109}$$

The Helmholtz free energy is related to the partition function Q, by:

$$A = -k_B T \ln Q \tag{2.110}$$

or, rearranging:

$$A = k_B T \ln\left\langle \exp\left[H(\mathbf{p}^N, \mathbf{r}^N)/k_B T\right]\right\rangle =$$
$$k_B T \ln\left(\iint d\mathbf{p}^N d\mathbf{r}^N \exp\left[H(\mathbf{p}^N, \mathbf{r}^N)/k_B T\right] \rho(\mathbf{p}^N, \mathbf{r}^N)\right) \tag{2.111}$$

The configurations with high energy make a significant contribution to the integral due to the presence of the exponential term. A molecular dynamics simulation preferentially samples the lower-energy regions of the phase space. An ergodic trajectory would visit all of these high-energy regions, but in practice these will never be adequately sampled by a real simulation. Consequently, the results for the free energy will be poorly converged and inaccurate.

Let us now consider the calculation of the free energy difference between states. This problem can be tackled using a molecular dynamics simulation method. Next, we will describe several methods proposed for calculating the free energy differences.

Thermodynamic Perturbation

Consider two well-defined states X and Y, containing *N* particles interacting according to H_X and H_Y, respectively. The free energy difference between the two states is as follows:

$$\Delta A = A_Y - A_X = -k_B T \ln \frac{Q_Y}{Q_X} \tag{2.112}$$

or,

$$\Delta A = -k_B T \ln \left\langle \exp\left(-\left[H_Y(\mathbf{p}^N,\mathbf{r}^N) - H_X(\mathbf{p}^N,\mathbf{r}^N)\right]/k_B T\right) \right\rangle_0$$
$$= k_B T \ln \left\langle \exp\left(-\left[H_X(\mathbf{p}^N,\mathbf{r}^N) - H_Y(\mathbf{p}^N,\mathbf{r}^N)\right]/k_B T\right) \right\rangle_1 \tag{2.113}$$

The subscripts 0 and 1 indicate averaging over the ensemble of configurations representative of the initial state X and final state Y, respectively. To perform a thermodynamic perturbation[39] calculation we must first define H_X and H_Y and then run a simulation, let us say at the state X. We form then the ensemble average given in equation (2.113) as we proceed in the simulation. The free energy difference could be obtained from a simulation of X, where for each configuration in the simulation, we calculate the value of the energy for every instantaneous conformation of X, in which X is temporarily assigned the potential energy of Y, or equivalently 'mutated' to Y. The same procedure can be done starting from a simulation at the state Y.

Thermodynamic Integration

An alternative way to calculate the free energy difference uses thermodynamic integration. The formula for the free energy difference is:

$$\Delta A = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial H(\mathbf{p}^N,\mathbf{r}^N)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{2.114}$$

To calculate the free energy difference using thermodynamic integration we thus need to determine the integral in equation (2.114). In practice, this is achieved by performing a series of simulations corresponding to discrete values of λ between 0 and 1, and for each value, the average of $\left\langle \frac{\partial H(\mathbf{p}^N,\mathbf{r}^N)}{\partial \lambda} \right\rangle_\lambda$ is determined. The total free energy difference is then equal to the area under the graph of $\left\langle \frac{\partial H(\mathbf{p}^N,\mathbf{r}^N)}{\partial \lambda} \right\rangle_\lambda$ versus λ.

The 'Slow Growth' Method

A third approach for the calculation of free energy differences from computer simulation is the slow growth method. Here, the Hamiltonian changes by a very small, constant amount at each step of the calculation. This means that at each stage Hamiltonian $H(\lambda_{i+1})$ is very nearly equal to $H(\lambda_i)$. The free energy difference is given by:

$$\Delta A = \sum_{i=1;\lambda=0}^{i=N_{step};\lambda=1} \left(H_{i+1} - H_i \right) \tag{2.115}$$

Potentials of Mean Force

The free energy changes that we have considered so far correspond to chemical 'mutations'. We may also be interested to know how the free energy changes as a function of some inter or intra molecular coordinate, such as the distance between two atoms, or the torsion angle of a bond within a molecule. The free energy surface along the chosen coordinate is known as a 'potential of mean force' (PMF). Unlike the mutations so common in free energy perturbation calculations, which are often along non-physical pathways, the PMF is calculated for a physically achievable process. Consequently, the point of higher energy on the free energy profile that is obtained from a PMF calculation corresponds to the transition state for the process. Various methods have been proposed for calculating the PMF. The simplest of them is the free energy change as the separation *(r)* between two particles is changed. We can calculate the PMF from the radial distribution function *g(r)*, which gives the probability of finding a particle a distance *r* from another particle compared to the ideal gas distribution, using the following expression for the Helmholtz free energy:

$$A(\mathbf{r}) = -k_B T \ln g(\mathbf{r}) + \text{constant} \tag{2.116}$$

The constant is often chosen so that the most probable distribution corresponds to a free energy of zero.

Unfortunately, the PMF may vary by several multiples of $k_B T$ over the relevant range of the parameter *r*. The logarithmic relationship between the PMF and the radial distribution function means that a relatively small change in the free energy may correspond to *g(r)* changing by an order of magnitude from its most likely value. Unfortunately, standard molecular dynamics simulation method does not adequately sample regions where the radial distribution function differs drastically from the most likely value, leading to inaccurate values for the PMF. The traditional way to avoid this problem uses the 'umbrella sampling' technique.

Umbrella Sampling

This method attempts to overcome the sampling problem by modifying the potential function so that the unfavorable states are sampled sufficiently. The modification of the potential function can be written as a perturbation:

$$E'(\mathbf{r}^N) = E(\mathbf{r}^N) + W(\mathbf{r}^N) \tag{2.117}$$

where $W(\mathbf{r}^N)$ is a weighting function, which often takes a quadratic form:

$$W(\mathbf{r}^N) = k_W \left(\mathbf{r}^N - \mathbf{r}_0^N \right)^2 \tag{2.118}$$

For configurations that are far from the equilibrium state $\mathbf{r}_0^N$ the weighting function will be large and so a simulation using the modified energy function $E'(\mathbf{r}^N)$ will be biased along some relevant 'reaction coordinate' away from the configuration $\mathbf{r}_0^N$. The resulting distribution will be non-Boltzmann. To obtain the PMF via the radial distribution function, equation (2.116) the distribution function with the forcing potential would be determined and then corrected to the 'true' radial distribution function, from which the free energy can be calculated as a function of the separation. It is usual to perform an umbrella sampling calculation in a series of stages, each of which is characterized by a particular value of the coordinate and an appropriate value of the forcing potential $W(\mathbf{r}^N)$. The PMF is then obtained by superposing the results obtained for all the series of stages. Although the method itself seems to be easily implemented in parallel computers, its principal drawback are that two consecutive series must overlap in order to correctly superpose them, the force constant for each stage has to be chosen carefully in order to make an efficient sampling of the surface, and the initial thermalization simulation must be performed for all the series of stages. These last reasons make umbrella sampling a technique for which the data managing becomes quite difficult and tedious.

Multiple Steering Molecular Dynamics

The multiple steering molecular dynamics (MSMD) approach is a novel technique for computing free energy profiles, and was originally proposed by C. Jarzynski.[40] The MSMD approach establishes a relation between the non-equilibrium dynamics and equilibrium properties.[41,42] We present the relevant equations which allow for efficient calculation of free energy profiles. Let $H(\mathbf{r}, \lambda)$ be the Hamiltonian of a system that is subject to an external time-dependent perturbation $\left(\lambda = \lambda(t) \right)$ and respectively let $\Delta A(\lambda)$ and $W(\lambda)$ be the change in free energy and the external work performed on the system as it evolves from λ_0 to λ. Here $\mathbf{r}$ indicates a configuration of the whole system, while λ is a reaction coordinate. Then, $\Delta A(\lambda)$ and $W(\lambda)$ are related to each other by the following identity:

$$\exp\left[-\Delta A(\lambda)/k_B T\right] = \left\langle \exp\left[-W(\lambda)/k_B T\right]\right\rangle \tag{2.119}$$

where the brackets represent an average taken over an ensemble of molecular dynamics trajectories, provided the initial ensemble is equilibrated. Equation (2.119) is valid under the assumption of a converged average, formally needing an infinite number of realizations of the process. The Hamiltonian $H(\mathbf{r}, \lambda)$ can be written as the sum of the time independent Hamiltonian of the unperturbed system, $H_0(\mathbf{r})$, plus a time-dependent external potential. As usual, the perturbation has been chosen to be a harmonic potential, whose minimum position moves at constant velocity v according to:

$$H(\mathbf{r}, \lambda) = H_0(\mathbf{r}) + \frac{k}{2}\left[\lambda(\mathbf{r}) - \lambda_0 - vt\right]^2 \tag{2.120}$$

where $\lambda(\mathbf{r})$ represents a chosen reaction path. Thus, the free energy of a process along a reaction coordinate can be computed performing a number of finite time transformations, collecting the work done at each time step, and then properly averaging it as in equation (2.119). This technique is much more straightforward than others, because is easily parallel, and the initial thermalization procedure has to be only performed at the initial or final stage, making this technique a powerful tool for calculating free energy profiles.

Finally, it is interesting to note that all free energy calculations performed throughout this thesis were done using the umbrella sampling and/or the MSMD methods, which were included inside the SIESTA package. Moreover, the MSMD method was introduced in the AMBER code, and will be released in the new version of AMBER9.

---

[1] Atomic units are used throughout the thesis in order to simplify the notation. In this convention the charge unit is the electron charge *(e)*, the mass unit is the electron mass *(m)* and the angular moment is defined in $\hbar$ multiples. Then, the energy and distance units are, respectively:

$$1\,hartree = e^2/4\pi\varepsilon_0 a_0 \ , 1\,bohr = \hbar^2/m_e e^2$$

[2] Szabo, A.; Ostlund, N. S. Modern Quantum Chemistry, First Edition, Mc Graw-Hill Publishing Company, New York, 1989.

[3] Parr, R. G.; Yang, W. Density-Functional Theory of Atoms and Molecules, Oxford University Press, New York, 1989.

[4] Levine, I. N. Quantum Chemistry, Fifth Edition, Prentice-Hall, Inc., New Jersey, 2000.

[5] The external potential is the energy of the interaction between the electron i and the nucleus:

$$v(\mathbf{r}_i) = -\sum_\alpha \frac{Z_\alpha}{r_{i\alpha}}$$

[6] Kohn, W.; Sham, L.J. Phys. Rev. 1965, 140, A1133.

[7] Vosko, S. H.; Wilk, L.; Nusair, M. Can. J. Phys. 1980, 58, 1200.

[8] Perdew, J. P.; Wang, Y. Phys. Rev. B 1986, 33, 8800.

[9] Perdew, J. P.; Burke, K.; Ernzerhof, M. Phys. Rev. Lett. 1996, 77, 3865.

[10] Becke, A. D. Phys. Rev. 1988, A38, 3098.

[11] Lee, C.; Yang, W.; Parr, R. Phys. Rev. B 1988, 37, 785.

[12] Perdew, J. P. Phys. Rev. B 1986, 33, 8822.

[13] Perdew, J. P.; Wang, Y. Phys. Rev. B, 1992, 45, 13244.

[14] Becke, A. D. J. Chem. Phys. 1993, 98, 5648.

[15] Soler, J. M.; Artacho, E.; Gale, J. D.; García, A.; Junquera, J.; Ordejón, P.; Sánchez-Portal, D. J. Phys.: Condens. Matter 2002, 14, 745.

[16] Sánchez-Portal, D.; Ordejón, P.; Artacho, E.; Soler, J. M. Int. J. Quantum Chem. 1997, 65, 453.

[17] Reich, S.; Thomsen, C.; Ordejón, P. Phys. Rev. B 2002, 65, 155411.

[18] Artacho, E.; Sánchez-Portal, D.; Ordejón, P.; García, A.; Soler, J.M. Phys. Status Solidi B 1999, 215, 809.

[19] Sankey, O. F.; Niklewski, D. J. Phys. Rev. B 1989, 40, 3979.

[20] Kleinman, L.; Bylander, D. M. Phys. Rev. Lett.1982, 48, 1425.

[21] Press, W. H.; Teulosky, S. A.; Vetterling, W. T.; Flannery, B. P. Numerical Recipes. The Art of Scientific Computing. Cambridge University Press, Cambridge, 1992.

[22] Leach. A. R. Molecular Modeling. Eds; Addison Wesley Longman, 1997.

[23] Allinger, N. L.; Yuh, Y. H.; Lii, J-H. J. Am. Chem. Soc. 1989, 111, 8551, 8566, 8576.

[24] Brooks, B.R.; Bruccoleri, R.E.; Olafson, B.D.; States, D.J.; Swaminathan, S.; Karplus, M. J. Comp. Chem. 1983, 4, 187.

[25] Ott, K-H.; B. Meyer J. Comp. Chem. 1996, 17, 1068.

[26] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J. Am. Chem. Soc. 1995, 117, 5179.

[27] Monard, G.; Merz, K. M. Jr. Acc. Chem. Res. 1999, 32, 904.

[28] Gao, J. Methods and applications of combined quantum mechanical and molecular mechanical Potentials. En Reviews in Computational Chemistry; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1996; Vol. 7.

[29] Wang, J.; Cieplak, P.; Kollman, P. A. J. Comput. Chem. 2000, 21, 1049.

[30] Elola, M. D.; Laria, D.; Estrin, D. A. J. Phys. Chem. A 1999, 103, 5105.

[31] Crespo, A.; Scherlis, D. A.; Martí, M. A.; Ordejón, P.; Roitberg, A. E.;Estrin, D. A. J. Phys. Chem. B 2003, 107, 13728.

[32] Lee, Y. S.; Worthington, S. E.; Krauss, M.; Brooks, B. R. J. Phys. Chem. B 2002, 106, 12059.

[33] Reuter, N.; Dejaegere, A.; Maigret, B.; Karplus, M. J. Phys. Chem. A 2002, 104, 1720.

[34] Eichinger, M.; Tavan, P.; Hutter, J.; Parrinello, M. J. Chem. Phys. 1999, 110, 10452.

[35] Rovira, C.; Schultze, B.; Eichinger, M.; Evanseck, J. D.; Parrinello, M. J. Biophys. J. 2001, 81, 435.

[36] Verlet L. Phys. Rev. 1967, 159, 98.

[37] Ryckaert, J. P.; Ciccotti G.; Berendsen, H. J. C. J. Comp. Phys. 1977, 23, 327.

[38] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W.; DiNola, A.; Haak, J. R. J. Chem. Phys. 1984, 81, 3684.

[39] Zwanzig, R. W. J. Chem. Phys. 1954, 22, 1420.

[40] Jarzynski, C. Phys. Rev. Lett. 1997, 78, 2690.

[41] Vandevondele, J.; Rothlisberger, U. J. Am. Chem. Soc. 2002, 124, 8163, and references therein.

[42] Hummer, G.; Szabo, A. Proc. Natl. Acad. Sci. U.S.A. 2001, 98, 3658.

Chapter 3: Chorismate to Prephenate Conversion Catalyzed by Chorismate mutase

Introduction

The conversion of chorismate to prephenate (Figure 3.1) is a problem for which there is an important body of theoretical and experimental data.
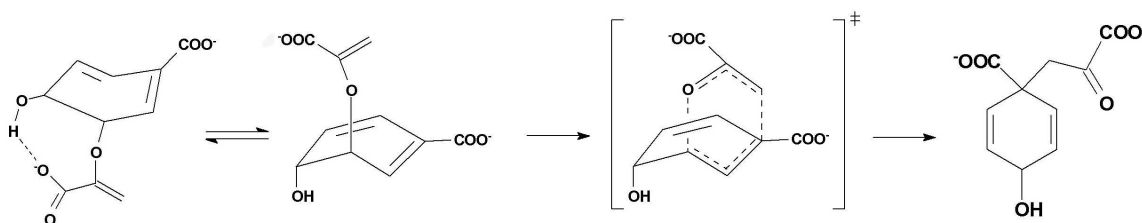

Figure 3.1. Chorismate to prephenate conversion. Species from left to right: diequatorial chorismate conformer, diaxial chorismate conformer, transition state and prephenate.

This intramolecular transformation is of special interest because it proceeds both uncatalyzed in water solution and catalyzed by the enzyme chorismate mutase, providing a unique opportunity to compare the catalyzed and uncatalyzed reactions.[1] Chorismate mutase is an essential enzyme in the Shikimate pathway. This pathway is responsible for the biosynthesis of aromatic amino acids in bacteria, fungi, and plants, but not in mammals.[2] The enzyme is of interest not only for an understanding of its catalytic rate enhancement, but also as a target for the discovery of antibiotics, antifungals, and herbicides. For these reasons, chorismate mutase has been the focus of numerous studies in the past decades.[3-20]

Our calculations provide a clear improvement with respect to previous AM1[10,11,17,18] and HF-based calculations,[14,20] which may present flaws in the description of transition states and overestimate the activation energy. The experimental activation parameters for the uncatalyzed reaction of chorismate to form prephenate in aqueous solution at pH 7.5 are $\Delta H^{\neq}$ = 20.7 kcal/mol and $\Delta S^{\neq}$ = -12.9 eu.[3] The enzyme chorismate mutase accelerates the reaction by a factor of about 10^6 over the uncatalyzed reaction. The activation parameters for the *B. subtilis* chorismate mutase reported by Kast et al.[7] are $\Delta H^{\neq}$ = 12.7 kcal/mol and $\Delta S^{\neq}$ = -9.1 eu. They concluded that lowering the entropy barrier of the reaction is not a major factor in catalysis in *B. subtilis* chorismate mutase, because $\Delta S^{\neq}$ in this case is unfavorable and comparable to that for the

uncatalyzed reaction. Chorismate has been found to adopt several conformations in vacuum and in condensed phase. The more populated conformations in vacuum are the pseudodiequatorial and pseudodiaxial, with the pseudodiequatorial being the most stable.[4,5,7,11] The rearrangement of chorismate to prephenate requires a conformational change in which the diequatorial conformer is converted to the diaxial one, which is the reactive conformer capable of progressing via the pericyclic rearrangement to prephenate (Figure 3.1) with a chair-like transition state. Several researchers have applied theoretical methods to study the conformational preequilibrium of chorismate. One of the first theoretical works was done by Wiest and Houk, who located and characterized as minima two chorismate conformers at the Hartree-Fock and density functional theory level (RHF/6-31G* and BLYP/6-31G*, respectively).[5,7] The mechanism of chorismate mutase has been studied by Khanjin et al. by performing B3LYP/6-31G* calculations on several chorismate conformationally restricted analogues.[9] Marti et al.[11] found five chorismate conformers as minima in vacuum. Hillier and co-workers[8] examined the effect of water on the conformational energetics of chorismate conformers by both a polarized continuum model (PCM) and explicit solvation within a Monte Carlo free energy perturbation (MC/FEP) treatment, and concluded that solvation reduces the energy difference between the diequatorial and diaxial structures, but still the diequatorial form was the most stable. Guo et al.[19] found two nonreactive conformers that are more stable than the reactive diaxial conformer in solution and which resemble the diequatorial conformers, using B3LYP/6-31G* and PCM. However, they showed that these conformers rapidly convert (approximately in 5 ps) to the active diaxial conformer in the active site of chorismate mutase using self-consistent charge density functional tight-binding (SCC-DFTB) calculations.

More recently, Jorgensen and co-workers, studied the environment effects of the reaction in water and in the active site of chorismate mutase. They used a combined QM-MM MC/FEP method and applied the concept of near attack conformations (NACs) to chorismate.[17,18] They showed that 1%, 82%, and 100% of chorismate conformers were found to be NACs structures, or active conformers, in vacuum, water, and in the enzyme, respectively. Moreover, they calculated the activation free energy of the conformational preequilibrium of chorismate, showing that only in the gas phase does this step provide a positive contribution to the total activation free energy of the reaction (approximately 3, -1, and -9 kcal/mol in a vacuum, water and enzyme, respectively). They proposed that the rate enhancement by chorismate mutase over the uncatalyzed aqueous reaction results primarily from conformational compression of the reactant by the enzyme and suggested that the preferential stabilization of the transition state in the enzyme environment relative to water played a secondary role.[18] On the other hand, Marti et al.[10] calculated the free energy profile for the reaction in aqueous solution and in the enzyme, and concluded that the enzyme preferentially binds the reactive diaxial conformer of chorismate, and it reduces the activation free energy of the reaction relative to that in solution by providing an environment which preferentially stabilizes the transition state. Mulholland and co-workers studied the energetic profile of the reaction at the RHF/6-31G(d) QM-MM using CHARMM and corrected it by ab initio

calculations in the gas phase (DFT and MP2).[20] Lee et al.[14] assessed the role of different residues in the stabilization of the transition state by performing QMMM calculations at the HF level. Recently, a QM-MM pathway optimization study of the conversion of chorismate to prephenate performed using a combination of HF/4-31G and B3LYP/6-31G** electronic structure methods has been reported.[15] Most of these QM-MM calculations were performed using HF[14,20] or AM1 Hamiltonians.[10,11,17,18] For this reason, it is interesting to reanalyze the subject using a more accurate electronic structure DFT scheme. In order to do so, we have performed QM-MM calculations of the chorismate to prephenate conversion, employing a restrained energy minimization scheme to obtain the reaction energy profiles in vacuum, aqueous solution, and in the enzyme environment. In the first place, we have performed calculations to compute a potential energy profile of the reaction, neglecting the entropic contributions. This is justified, because, according to experimental results, they play a minor role in the catalysis.[4] In the second part of the chapter, we have included these effects by means of sampling the degrees of freedom orthogonal to the reaction coordinate by performing molecular dynamics simulations. We will try to shed light on the enzymatic mechanism, particularly to understand if the catalytic activity is due to a conformational compression of the reactant, a preferential stabilization of the transition state, or both and will also assess the effects of the entropic changes.

Reaction Pathway Search Calculations

All calculations were performed with the SIESTA code using DZVP basis sets, with a pseudoatomic orbital energy shift of 30 meV, a grid cutoff of 150 Ry,[21] and the generalized gradient approximation of Perdew, Burke, and Ernzerhof[22] (see Chapter 2). For the treatment of biological macromolecular environments, we used the Wang et al. force field parametrization.[23] The Lennard-Jones parameters for the QM atoms have been taken from the Wang et al. force field parametrization, and have also been applied to chorismate in an earlier work.[16] The parameters have been tested by performing QM-MM calculations in selected chorismate (QM) - water (MM) dimmers. Results for hydrogen bond distances and interaction energies were in good agreement with full QM calculations. QM-MM restrained energy minimizations have been performed to investigate the reaction path for the conversion of chorismate to prephenate. The distinguished reaction coordinate ξ was taken as the antisymmetric combination of the distances describing the breaking and forming bonds, $\xi = d_{C1-C7} - d_{C3-O1}$, as depicted in Figure 3.2. This reaction coordinate has been found to represent very closely the intrinsic reaction coordinate found in the gas phase and in solution.[10,14] The path is constructed as follows. First, an unrestricted QM-MM minimization is performed for the reactant or product to generate an initial configuration for the reaction path. The reaction path is mapped out by adding the harmonic potential term to the potential energy with ξ_0 varying in steps of 0.1 Å from about -2.0 to 2.0 Å,

and performing energy minimizations at each step. A value of k of 200 kcal mol^{-1} Å^{-2} was found to be adequate.

To consider charge transfer to the substrate and polarization effects in neighbor charged residues, which may be important for the catalytic activity of the enzyme,[12,14] we have performed calculations for two different choices of the QM subsystem, one including only the substrate moiety (Figure 3.2) and another with the substrate plus the charged side chains glu78 and arg90 (Figure 3.3). The frontier between the QM and MM portions was treated using the SPLAM[24] method (see Chapter 2).
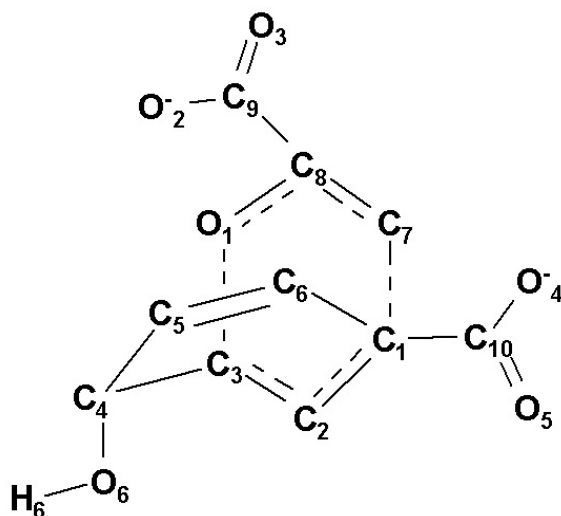
Figure 3.2. QM subsystem model of the reactant with its relevant numbering.
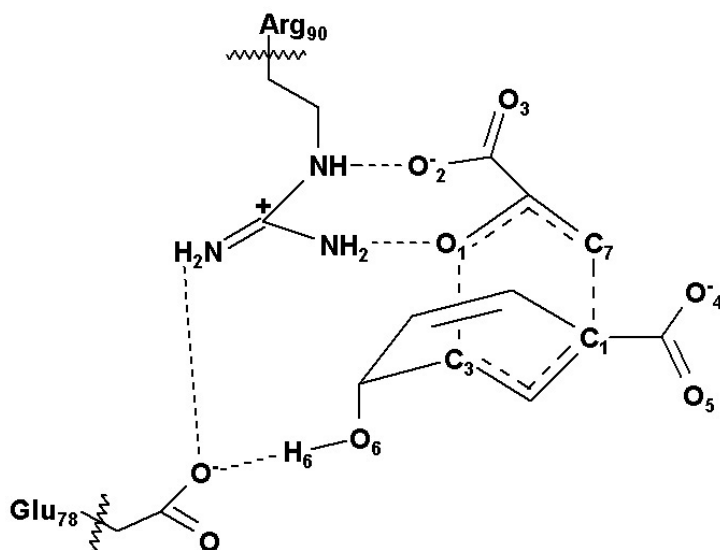
Figure 3.3. QM subsystem model including the reactant plus the charged side chains Glu78 and Arg90.

The QM-MM calculations in solution were performed by placing the chorismate in a cavity deleted from a 15 Å radius sphere of 664 water molecules. The water molecules were equilibrated by performing 100 ps of classical MD simulations at 200 K, both with chorismate and prephenate as solutes. Geometry optimizations have been performed starting with both reactant and product structures. Only the substrate atoms and the MM atoms lying on a sphere of 11 Å centered on it were allowed to move freely.

The enzyme calculations have been performed employing a starting structure obtained from *B. subtilis* in the Protein Data Bank (1COM).[25] Hydrogen atoms were added as usual. The systems were equilibrated by performing 100 ps of classical MD simulations at 300 K, both with chorismate and prephenate as solutes, to obtain correct starting geometries. We have performed calculations for two different choices of the QM subsystem. In the first case, the full system consisted of a QM region of 24 atoms, (Figure 3.2) the substrate molecule, treated at the DFT level, and a MM region, comprising the enzyme plus solvation water molecules (5627 protein atoms plus 1534 TIP3P water molecules). Only the QM atoms and the MM atoms lying on a sphere of 11 Å centered on the QM system were allowed to move freely. The model is depicted in Figure 3.4. Geometry optimizations have been performed for chorismate and prephenate in the active site. The second choice of the QM subsystem includes the substrate moiety plus the charged side chains glu78 and arg90, as shown in Figure 3.3.
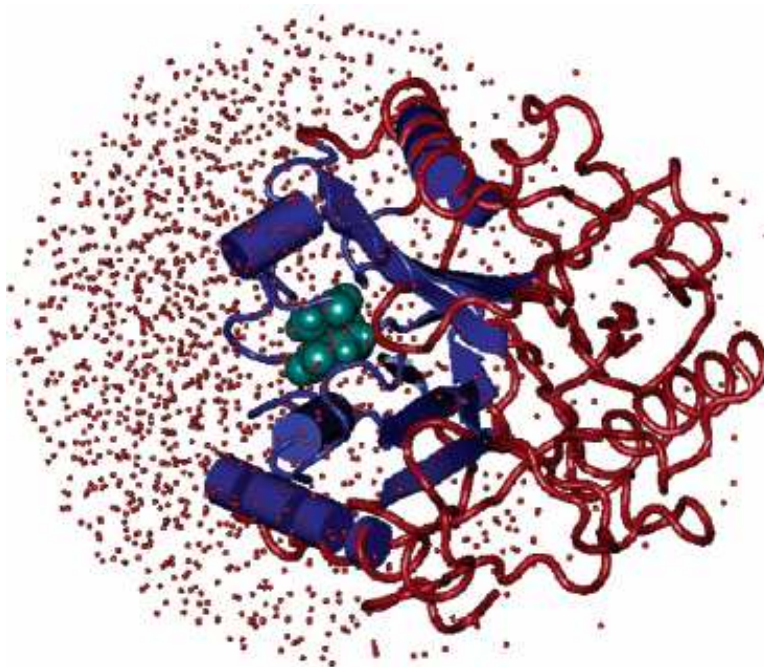
Figure 3.4. Chorismate mutase model with solvation water molecules (red dots). The substrate is depicted in cyan, the MM free atoms are depicted in blue, whereas the frozen ones are depicted in red.

Results and Discussion

To validate the SIESTA electronic structure scheme, we have performed geometry optimizations of chorismate and prephenate in vacuum. In Table 3.1 we report PBE/DZVP results computed with SIESTA for selected optimized geometrical parameters of the reactant (diequatorial conformer) and product, compared with results obtained at the B3LYP/6-31G**[26,27] level computed with the *Gaussian*98 package[28] (Figure 3.2). The computed ΔE of conversion between chorismate and prephenate of -5.8 kcal/mol at the PBE/DZVP level agrees reasonably well with that computed using B3LYP/6-31G** of -4.8 kcal/mol.

	Chorismate		Prephenate	
	PBE/DZVP	B3LYP/6-31G**	PBE/DZVP	B3LYP/6-31G**
C1-C7	4.42	4.41	1.58	1.58
C1-C3	1.45	1.45	3.83	3.80
O2-H6	1.57	1.68	4.74	4.62
C3-O1-C8	119.1	121.2	-	-
C1-C7-C8	-	-	116.5	116.2
O2-C9-C8-C7	163.7	153.2	117.0	121.7

Table 3.1. Selected geometrical parameters of chorismate diequatorial conformer and prephenate in vacuum (Å and degrees). The PBE/DZVP and B3LYP/6-31G calculations were performed using SIESTA and Gaussian98, respectively.**

Taking into account that the diequatorial to diaxial conformer preequilibrium does not contribute to the total activation free energy of the reaction in aqueous solution and in the enzyme environment,[18] we have computed the energy profile for the conversion of the diaxial chorismate conformer to prephenate. The forward and reverse reactions were examined, and the corresponding energy profiles are shown in Figure 3.5. We also include the global minimum energy of the reactant, corresponding to the diequatorial conformer.
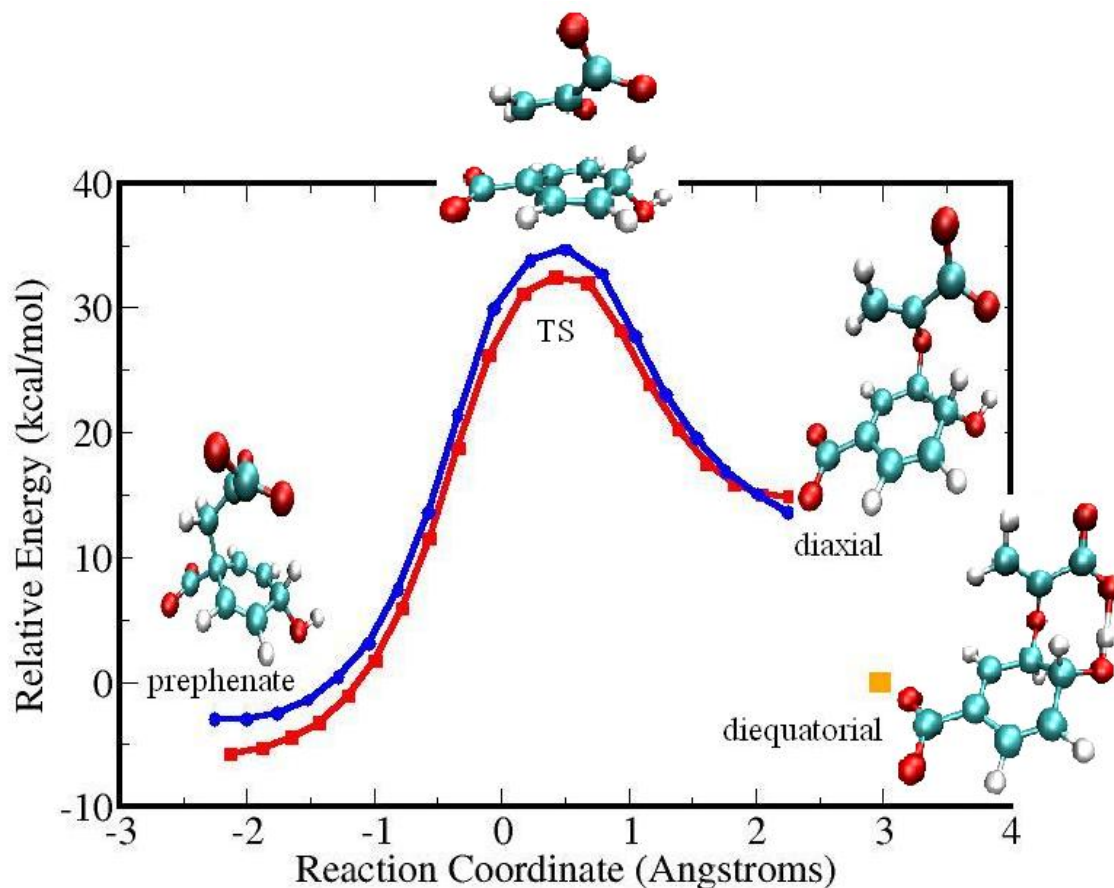
Figure 3.5. Energy profile for the forward (red, squares) and inverse (blue, circles) reaction of chorismate to prephenate in vacuum. The diequatorial conformer is also shown (orange).

As can be noted in Figure 3.5, the two profiles agree reasonably well. The computed activation energy in vacuum is 32.4 kcal/mol, in agreement with a previous value computed by Houk et al.[5] of 34.3 kcal/mol, at the BLYP/6-31G* level. These values are slightly lower than the reported B3LYP/6-31G* result of 44.0 kcal/mol,[10] consistent with the fact that DFT calculations at the gradient-corrected approximation level usually yield lower activation energies than those obtained using hybrid functionals.[29] The reaction coordinate values are 2.97, 2.25, 0.49, and -2.25 Å for the diequatorial conformer, diaxial conformer, transition state, and product, respectively. Relevant energetic parameters are presented in Table 3.2.

The computed energy difference between the active diaxial conformer and the most stable diequatorial conformer in vacuum is 14.9 kcal/mol, which compares well with previous calculations of 15.9 kcal/mol[11] and 11.7 kcal/mol,[5] at the MP2/6-31G* and BLYP/6-31G* levels, respectively. This energy difference is mainly due to an intramolecular hydrogen bond between the ring hydroxyl group and the bridged carbonyl group, and is responsible for the very high activation energy for the reaction in gas phase.

	ΔE	$\Delta E^{\neq}$	$\Delta E_{exp}^{\neq}$
Vacuum	-5.8	32.4	-
Aqueous solution	-22.0	13.8	20.7[a]
enzyme (QM subsystem Figure 3.2)	-24.9	5.3	12.7[b]
enzyme (QM subsystem Figure 3.3)	-30.6	4.3	12.7[b]

Table 3.2. Relevant Energetic Parameters for the Chorismate-to-Prephenate Conversion in Vacuum, Water Solution, and in the Chorismate mutase Enzyme (kcal/mol). [a] reference 3. [b] reference 4.

The aqueous solution chorismate to prephenate ΔE value of -22.0 kcal/mol compares reasonably well with the results of continuum model calculations of -20.5 kcal/mol,[7] at the B3LYP/6-31G* level. The energy profile is presented in Figure 3.6. Our computed activation energy is 13.8 kcal/mol, which is lower than the reported experimental value of 20.7 kcal/mol,[3,29] consistent with the computed value in vacuum.

The reaction coordinate values are 1.71, 0.43, and -1.69 Å for the reactant, transition state, and product, respectively. Relevant energetic parameters are presented in Table 3.2.
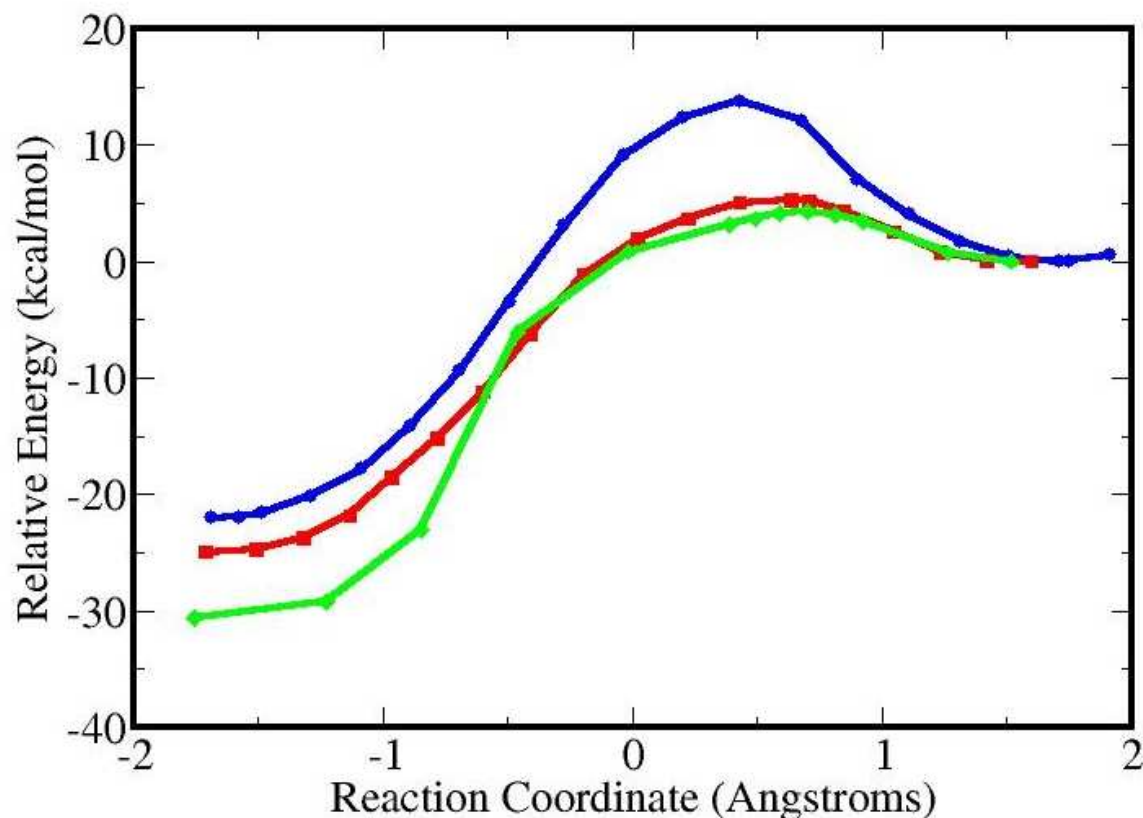
Figure 3.6. Energy profile for the reaction of chorismate to prephenate in aqueous solution (blue, circles) and in the enzyme with the two different QM subsystems: substrate (red, squares), substrate plus the charged side chains glu78 and arg90 (green, diamonds).

For the smaller QM subsystem, our predicted chorismate to prephenate ΔE in the enzyme is -24.9 kcal/mol. This lower result with respect to solution could be explained by considering that when the rearrangement is taking place an additional hydrogen bond between arg90(HH11) and glu78(OE1) is formed. This hydrogen bond is formed only for values of the reaction coordinate lower than about -0.8 Å, and consequently, it is only present in the product. This additional interaction is responsible for the more negative chorismate-to-prephenate ΔE value for the reaction in the presence of the enzyme (-24.9 kcal/mol) compared to the computed value in water (-22.0 kcal/mol) (Table 3.2 and Figure 3.7). It is important to note that the requirement for ΔE (reactant-product) to be the same in solution and in the enzyme-catalyzed reaction is valid only when both the initial and final states have the reactant and product unbound from the enzyme. It is then not surprising that we obtain different equilibrium ΔE in solution and inside chorismate mutase. The enzymatic energy profile is also presented in Figure 3.6. The reaction coordinate values are 1.60 Å, 0.71 Å and -1.71 Å for the reactant, transition state and product, respectively.
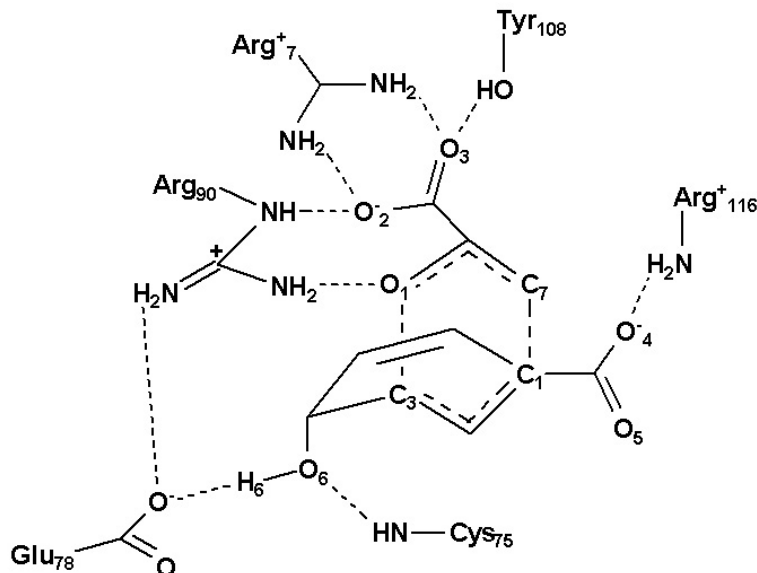
Figure 3.7. Active site model of *B. subtilis* chorismate mutase enzyme.

Energetic results are also presented in Table 3.2. It is clear that our results reproduce the catalytic effect of the enzyme, since the computed activation energy reduces from 13.8 kcal/mol to 5.3 kcal/mol when going from aqueous solution to the enzyme. The computed activation energy is lower than the reported experimental value of 12.7 kcal/mol,[2,29] but the computed difference between the activation energies of the reaction in water and in the enzyme, $\Delta(\Delta E^{\neq})$ of 8.5 kcal/mol, agrees very well with the experimental value of 8 kcal/mol. The predicted activation energy value is in the typical range of activation energies for enzymatic reactions (about 5-10 kcal/mol).

To analyze the different contributions to the activation energy, we have carried out an analysis of the different energy terms of chorismate and transition state optimized structures, as suggested by Marti et al.[10] The total QM-MM activation energies obtained in aqueous solution or in the enzyme environment can be written as the following sum:

$$\Delta E^{\neq} \approx \Delta E_{QM} + \Delta E_{MM} + \Delta E_{int} \qquad (3.1)$$

where each contribution corresponds to single point energy calculations of the transition state, relative to the reactant, computed at the structures obtained in the QM-MM calculations. The terms ΔE_{QM}, ΔE_{MM} and ΔE_{int}, correspond to the in vacuo DFT energy, the MM contribution, and the substrate-environment interaction energy, respectively. The ΔE_{QM} values are 13.5 and 11.5 kcal/mol for water and enzyme, respectively. The transition state relative energy is slightly (2 kcal/mol) more favorable in the enzyme active site than in water. This steric contribution to the

activation energy in the enzyme is attributed to a more compressed reactant structure than in solution. This is also reflected in the fact that the reaction coordinate value corresponding to chorismate in solution is 1.71 Å, whereas in the enzyme is 1.60 Å.

We can also compare the E_{QM} single point QM energies of the substrate at the optimized structures in aqueous solution and protein. The difference in chorismate E_{QM} values between solution and protein shows that the substrate is destabilized by steric compression in 2.1 kcal/mol. This effect has been reported by Jorgensen et al.[18] to be of about 9 kcal/mol. This magnitude computed for the transition state is only 0.1 kcal/mol. These facts reflect that the steric compression of the reactant by the enzyme destabilizes it and hence reduces the activation energy. Steric compression effects have also been analyzed by Khanjin et al.[9] by performing DFT calculations on a set of chorismate conformationally restricted analogues in vacuum. Our chorismate optimized C1-C7 bond distances are 4.42, 3.62, and 3.13 Å, in vacuum, water, and protein, respectively. On the other hand, the optimized C1-C7 distances in the TS span a much smaller window, with 2.61, 2.53, and 2.64 Å, in vacuum, water, and protein, respectively. Following these authors' analysis, this confirms that steric compression effects are more important in the reactant than in the TS.

The ΔE_{int} values of 5.9 kcal/mol for water and -4.7 kcal/mol in the enzyme show that the major catalytic character of the enzyme comes from an important stabilization of the transition state due to a favorable electrostatic enzyme transition state interaction. This effect is not observed in water, where the substrate is more stabilized. The electrostatic stabilization of the transition state by the enzyme is about 10.6 kcal/mol, with respect to water. The ΔE_{MM} values are -5.6 kcal/mol and -1.5 kcal/mol for water and enzyme, respectively. In both cases, the environment rearrangement stabilizes preferentially the transition state. In solution, this negative value almost compensates the unfavorable ΔE_{int}, resulting in a net activation energy of approximately 13 kcal/mol. The rearrangement energy in the enzyme is not an important contribution to the activation energy. Taking all these contributions into account, we can conclude that the catalytic activity achieved by the enzyme in comparison with the aqueous solution reaction is mainly due to both a major electrostatic stabilization of the transition state by the enzyme and a minor steric compression and hence destabilization of the substrate.

Relevant geometrical parameters are shown in Table 3.3 (Figure 3.7) for the enzyme complexed with chorismate, transition state, and prephenate.

	Chorismate	Transition State	Prephenate
C1-C7	3.13	2.64	1.59
C3-O1	1.53	1.93	3.30
H6-glu78(OE1)	1.49	1.46	1.54
O1-arg90(HH22)	1.80	1.74	1.90
O2-arg90(HE)	1.80	1.83	1.81
O2-arg7(HH21)	1.80	1.79	1.77
O3-arg7(HH12)	1.88	1.90	1.86
O3-tyr108(HH)	1.76	1.75	1.75
O4-arg116(HH12)	1.87	1.86	1.87
O6-cys75(H)	1.99	1.96	1.91
Arg90(HH11)-glu78(OE1)	3.69	3.64	1.85

Table 3.3. Relevant geometrical parameters of chorismate, transition states, and prephenate complexed with the enzyme (Å).

It can be seen that the enzyme locks the chorismate, transition state, and prephenate in its active site by making strong contacts with the relevant nearest residues. The strongest interactions are shown to be with residues glu78, arg90, arg7, tyr108, arg116, and cys75. It is important to remark that most distances between the substrate and these residues do not change significantly during the rearrangement. However, as can be seen from Table 3.3, H6-glu78(OE1) and O1-arg90(HH22) hydrogen bond distances are the ones that get shorter (from 1.49 and 1.80 Å to 1.46 and 1.74 Å, respectively) when going from the reactant to the transition state, consistently with its preferential electrostatic stabilization. Moreover, these two distances are longer in the product structures than in both reactant and transition state (1.49 and 1.80 Å in reactant and 1.54 and 1.90 Å in product, respectively). This hydrogen bond weakening is related to the formation of an additional arg90(HH11)-glu78(OE1) hydrogen bond, which is present only in the product structure, as can be noted by comparing the arg90(HH11)-glu78(OE1) distances of 3.69 and 1.85 Å in reactant and product, respectively. Finally, in Figure 3.8 we show the evolution of the reaction projected into the space spanned by the distances involved in the reaction coordinate. It is clear that the conversion is not symmetric, and also that the reaction takes place initially by increasing the C-O bond to reach the transition state with a subsequent decrease of the C-C bond to form the prephenate product.
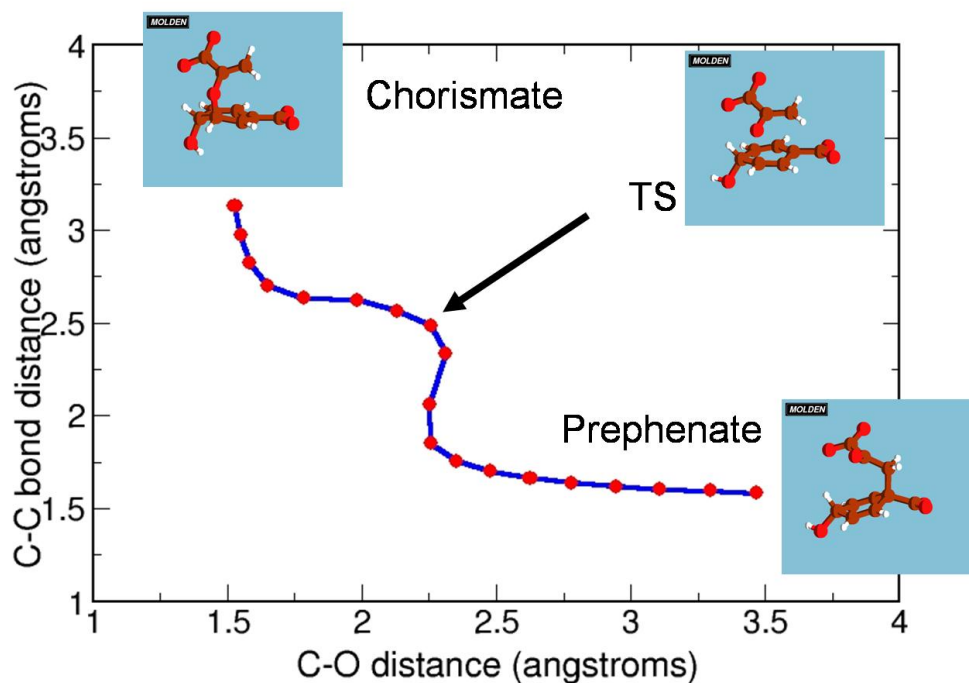
Figure 3.8. Evolution of the reaction projected into the C-O distance and C-C distance. The optimized points are depicted in red while the reaction coordinate is shown in blue.

The simulations performed with the second choice of the QM subsystem, including the substrate plus the charged side chains glu78 and arg90, yield a chorismate-to-prephenate ΔE value of -30.6 kcal/mol, compared with -24.9 kcal/mol for the smaller QM subsystem. This is probably due to the fact that the hydrogen bond between arg90(HH11) and glu78(OE1) is stronger when these residues are treated quantum mechanically. The energy profile for these calculations is presented in Figure 3.6, and energetic results are also presented in Table 3.2. The computed activation energy is 4.3 kcal/mol, which is slightly lower than the previous value of 5.3 kcal/mol predicted for the smaller QM subsystem.

A significant degree of charge transfer is observed between arg90 and O1, as can be seen by comparing the O1 Mulliken populations for the two choices of QM subsystem. The same effect is noted for the Mulliken populations of H6 and O6, due to charge transfer from glu78 (Table 3.4).

	QM subsystem (Figure 3.2)			QM subsystem (Figure 3.3)		
	Chorismate	TS	Prephenate	Chorismate	TS	Prephenate
O1	-0.274	-0.372	-0.250	-0.228	-0.292	-0.148
C3	0.236	0.330	0.184	0.276	0.332	0.198
C1	0.158	0.190	0.256	0.160	0.196	0.258
C7	0.370	0.398	0.366	0.370	0.380	0.360
O6	-0.176	-0.172	-0.148	-0.156	-0.160	-0.120
H6	0.052	0.062	0.028	-0.126	-0.128	-0.102
glu78	-1.0	-1.0	-1.0	-0.67	-0.63	-0.48
arg90	1.0	1.0	1.0	0.66	0.62	0.44
substrate	-2.0	-2.0	-2.0	-1.99	-1.99	-1.96

Table 3.4. Mulliken charges of relevant atoms and groups complexed with the enzyme.

This is consistent with previous results obtained by Lee et al.[14] Strong charge transfer effects between arg90 and glu78 are also observed, especially for the product structure where a hydrogen bond between them is present. For both QM subsystems, C3 and O1 Mulliken charges increase when going from reactant to transition state, consistently with the preferential electrostatic stabilization of the transition state. The smaller values for these charges in the product with respect to both reactant and transition state are probably related to the weakening of the hydrogen bonds with arg90(HH22) and glu78(OE1), due to the formation of the hydrogen bond between arg90(HH11) and glu78(OE1). Both charge transfer and polarization effects between the substrate and the residues arg90 and glu78 are operative when these residues are included in the QM subsystem. On the other hand, only polarization effects by the neighbor amino acids are considered when using the substrate as the QM subsystem. However, the activation energy computed with both choices of the QM charge subsystem is similar, suggesting that the catalytic activity of the enzyme is mainly related to substrate polarization effects.

Free Energy Calculations

Predictions based on the total potential energy may present flaws because motions in regions distant to the active site during pathway minimizations may significantly affect the energy profile. In this sense, calculating free energy profiles will yield much more meaningful results. However, obtaining free energy profiles require large configurational sampling, which is very expensive at the QM-MM level.

We have tested a novel scheme for obtaining free energy profiles, employing the multiple steering molecular dynamics (MSMD) approach[30] introduced in Chapter 2, for the conversion reaction of chorismate to prephenate discussed before.

For this well studied reaction, free energy profiles have been reported. Jorgensen and coworkers studied the environment effects of the reaction by using a combined (AM1) QM-MM MC/FEP method.[18] Marti et al.[10] calculated the free energy profile for the reaction in aqueous solution and in the enzyme, using an umbrella sampling AM1 scheme. Most of these studies agree that the enzyme reduces the activation free energy of the reaction by providing an environment which preferentially stabilizes the transition state. We present predictions of free energy profiles in an enzymatic system using the MSMD ideas proposed by Jarzynski, employing a QM-MM DFT Hamiltonian.

The calculations have been performed employing again a solvated structure of the *B. subtilis* chorismate mutase (496 TIP3P water molecules were added centered on the active site). The total system included the 24 substrate atoms plus 7115 MM atoms. In order to obtain correct starting geometries, the system was equilibrated performing classical MD simulations using the Amber8 package,[31] taking both chorismate and prephenate as solutes. The equilibration scheme was performed as follows:

(1) Classical structural optimization in 3 procedure steps, only hydrogens, only waters and the whole system.

(2) Classical thermalization to 300K; 100 ps with temperature gradient from 150 to 250 K and 100 ps with temperature gradient from 250 to 300 K (dt = 1 fs), only the substrate plus a sphere of 15 Å from it were allowed to move freely.

(3) Classical MD simulation at 300 K using the Berendsen[32] algorithm for 2.5 ns.

From the last 2 ns of the classical simulation of chorismate and prephenate plus chorismate mutase, we collected 20 starting structures of each (40 total), for our multiple steering QM-MM simulations. Each of them was relaxed at the QM-MM level for 0.5 ps at 300 K using the Berendsen algorithm (10 fs of bath coupling constant) with a time step of 0.5 fs (deuterium was used for all hydrogen atoms). Only atoms within a sphere of 15 Å from the QM structure were allowed to move.

All QM-MM calculations have been performed using the same basis set, functional, solute *LJ* parameters, and reaction coordinate as in the previous chorismate to prephenate calculations (Figure 3.2). These 20+20 QM-MM relaxed structures were used as starting points for the MSMD runs. The reaction coordinate has been changed from ξ = 1.8 Å to ξ = -1.8 Å for a set of 15+15 QM-MM structures at a constant pulling speed of 2 Å/ps, and for another set of 5+5 QM-MM structures at constant speed of 1 Å/ps.[33] A force constant of 200 kcal mol^{-1} Å^{-1} was used in all cases. In addition, the potential of mean force has been computed using an umbrella sampling scheme.[34] A total of 12 windows simulations of 5 ps each have been employed, using as starting structures the snapshots of the constrained energy minimizations performed previously.

The total simulation time corresponding to the MSMD simulations is 5500 hours for the set of 15+15 QM-MM structures and pulling speed of of 2 Å/ps (30 runs of 4000 steps each of about 2.75 minutes). The total simulation time of the umbrella sampling simulation is about the same (12

windows of 10000 steps each), corresponding to a total of 60 ps for both calculations. The MSMD simulations take the advantage that the initial configurations are statistically unrelated, since they were taken from a long classical MD simulation of the initial state. All calculations were performed on Pentium 4 IBM workstations.

Results and Discussion

In Figure 3.9 we show the values of accumulated work vs ξ for chorismate to prephenate conversion for the 20 trajectories. Also shown is the standard deviation of the work values. These data should be trusted from $\xi = 1.8$ to -0.5 Å, at which point $\sigma_w > 3k_B T$.
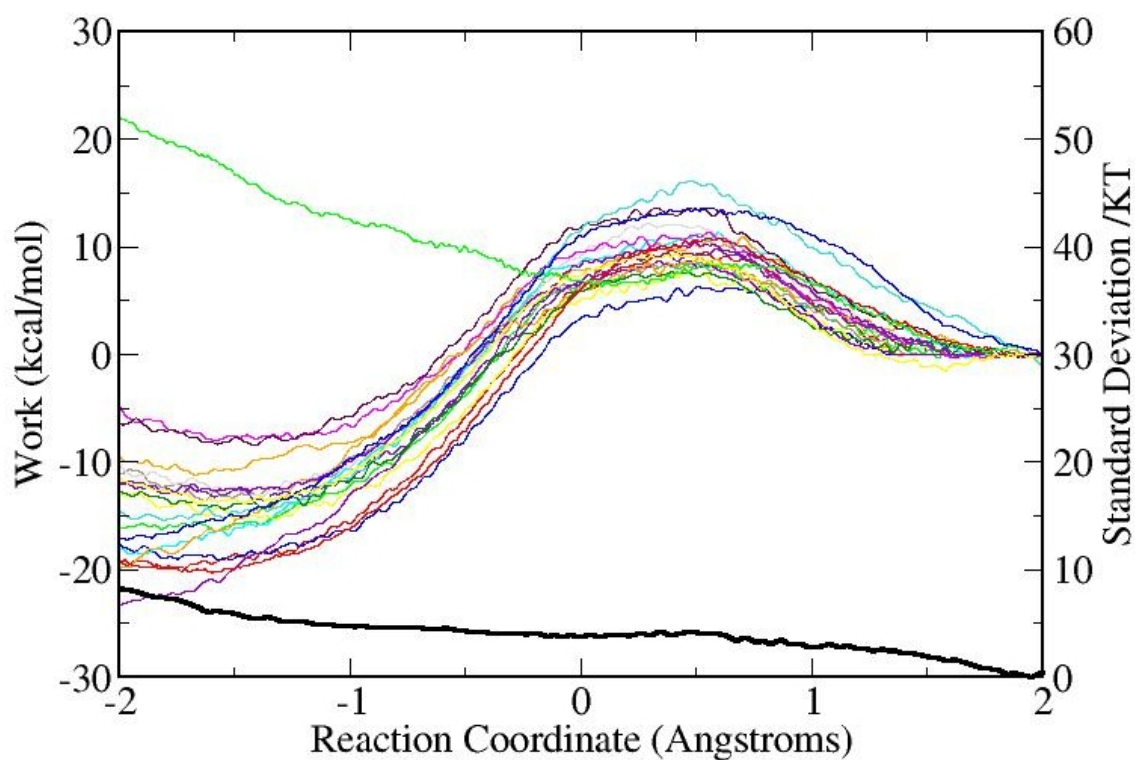
Figure 3.9. Chorismate to prephenate work for the 20 runs (colored) and the standard deviation (thick black line).

Figure 3.10 has the same data starting at the prephenate side of the reaction. These data are good only from $\xi = -1.8$ to +0.5 Å.
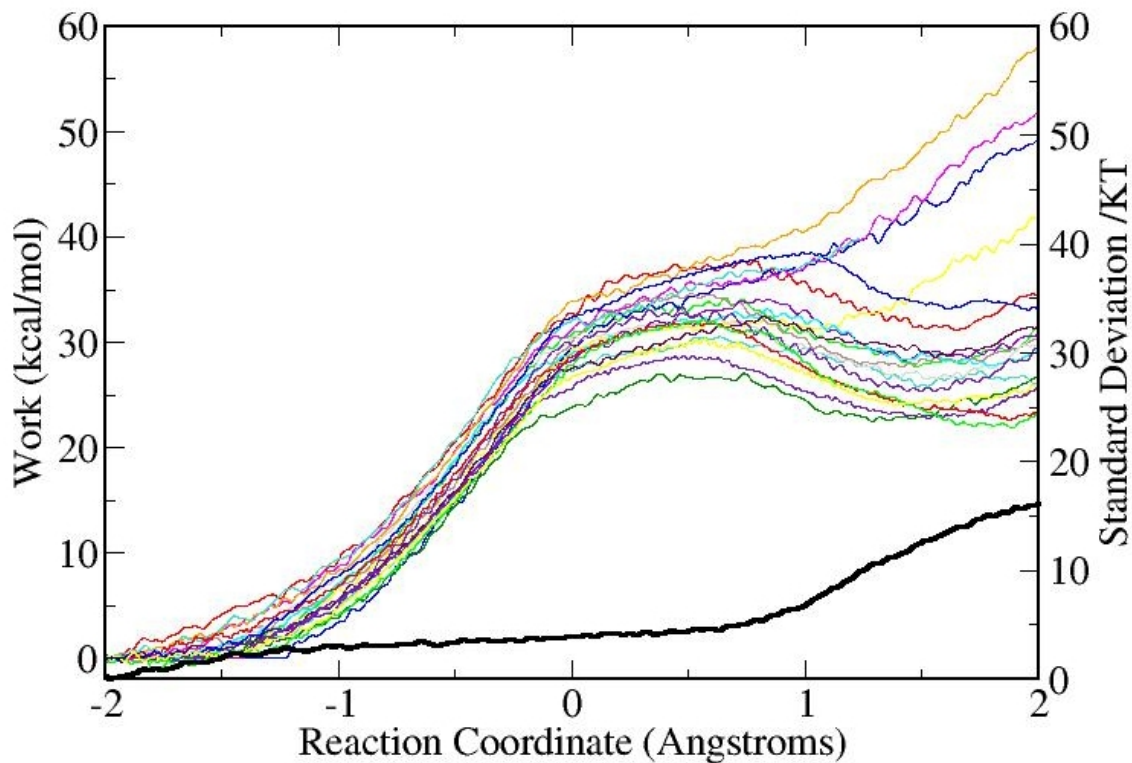
Figure 3.10. Prephenate to chorismate work for the 20 runs (colored) and the standard deviation (thick black line).

Figure 3.11, in red, shows the Jarzynski estimator for the free energy of set 1 (15+15 structures, pulling speed of 2 Å/ps) according to the Jarzynski estimator (see Chapter 2). In green, we present the same results for set 2 (5+5 structures, pulling speed of 1 Å/ps). They both have been obtained by joining the forward and reverse free energies curves obtained by exponential averaging of work from Figure 3.9 and Figure 3.10, respectively.
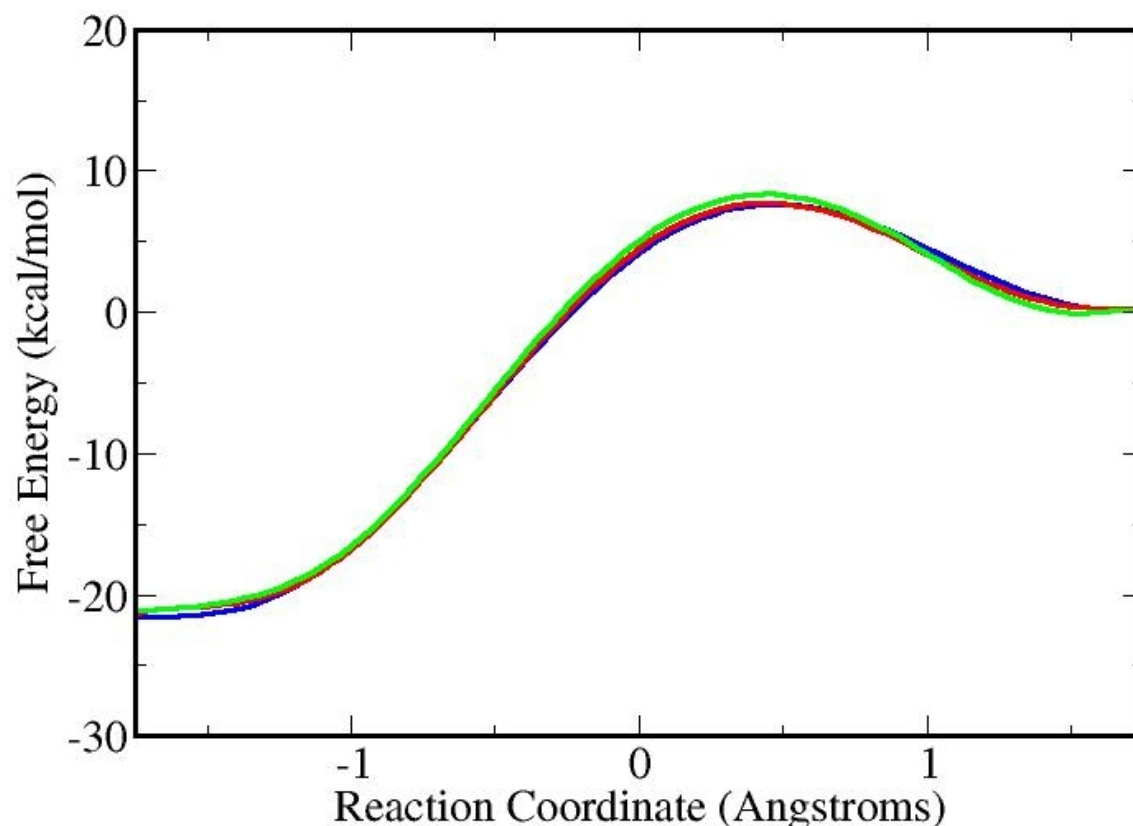
Figure 3.11. Free energy profile from chorismate ($\xi \approx$ 1.75 Å) to prephenate ($\xi \approx$ -1.75 Å), calculated using Jarzynski's equality both forward and reverse data are used) for set 1 (red), for set 2 (green), and for umbrella sampling scheme (blue).

We can conclude that the different pulling velocities do not change quantitatively the $\Delta G^{\neq}$ values obtained. Although our $\Delta G^{\neq}$ values are lower compared to the experimental ones, (7.7 kcal/mol vs. 15 kcal/mol) due to flaws in our DFT description,[29] the calculated entropic effect is negative, in agreement with the experimental value (-9.1 eu).[7] Previous calculation of the entropic effect for this reaction in reference 10 computes the wrong sign. In addition, the potential of mean force has been computed using an umbrella sampling scheme, for comparison. The blue curve in Figure 3.11 presents the umbrella sampling data, in perfect agreement with the MSMD estimator, but computed in a much less straightforward way.

In Figure 3.12 we show the evolution of the runs from chorismate to prephenate (left) projected into the C-O distance and C-C distance and from prephenate to chorismate (right). As it can be seen, the dynamics samples mainly around minimum energy path (also shown). However, it is evident in Figure 3.12 that there are also a significant number of trajectories that deviate strongly from the minimum energy path and do not form product. Is there something wrong with these

trajectories? By analyzing these trajectories, it can be seen that they correspond to an alternative reaction, yielding enol-pyruvate and para hydroxy-benzoic acid (Figure 3.13). It is interesting to remark that this reaction is catalyzed by a related enzyme, the chorismate lyase.[35]
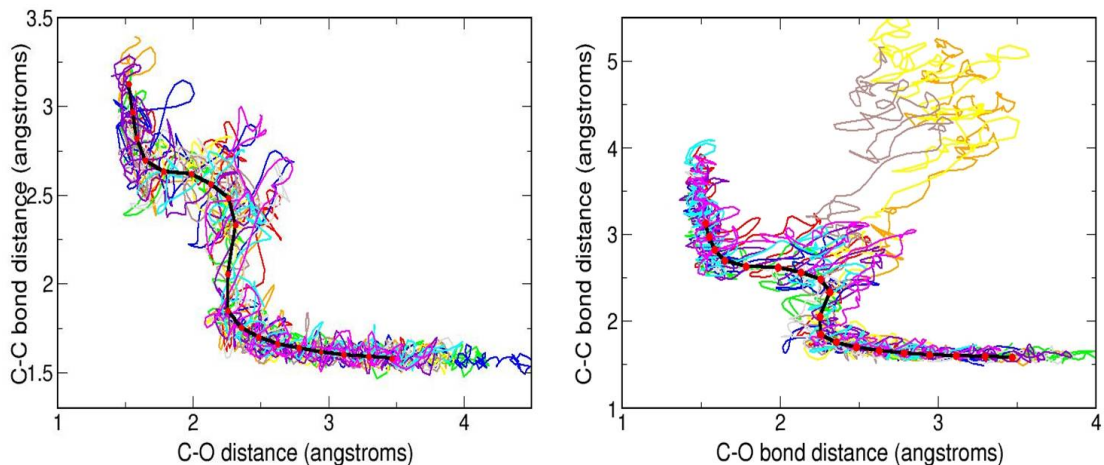

Figure 3.12. Evolution of the 20 runs from chorismate to prephenate projected into the C-O distance and C-C distance (left). Evolution of the 20 runs from prephenate to chorismate projected into the C-O distance and C-C distance (right). The minimum energy path projection is also shown in black.
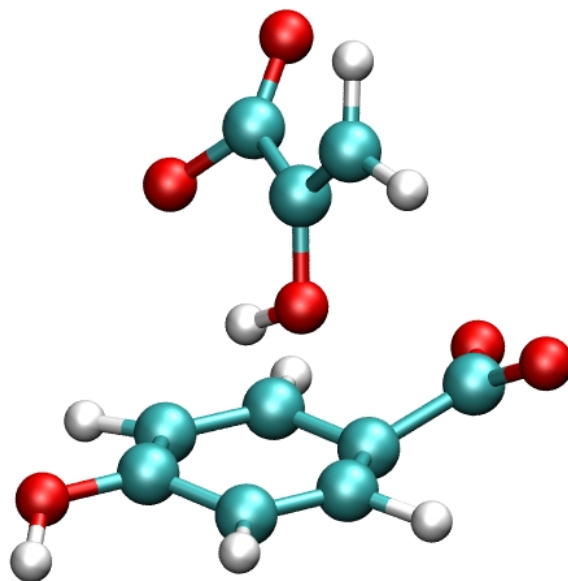
Figure 3.13. Structure of the enol piruvate plus para-hydroxy-benzoic acid, product formed in some of the trajectories.

Conclusions

In this chapter we have validated our own QM-MM scheme by performing calculations of the conversion of chorismate to prephenate in aqueous solution and in the active site of *B. subtilis* chorismate mutase enzyme. We have predicted a catalytic activity in the enzyme environment with respect to the uncatalyzed solution reaction, reflected in an activation energy difference $\Delta(\Delta E^{\neq})$ of 8.5 kcal/mol ($k_{cat}/k_{uncat} = 1.56\mathbf{x}10^{6}$). This value agrees very well with the experimental value of 8 kcal/mol ($k_{cat}/k_{uncat} = 6.73\mathbf{x}10^{5}$). The catalytic activity achieved by the enzyme in comparison with the aqueous solution reaction is mainly due to both a major electrostatic stabilization of the transition state by the enzyme and a minor steric compression and hence destabilization of the substrate, reducing the activation energy of the reaction.

Moreover, we have computed the free energy profile for this reaction employing the MSMD approach. Our predicted $\Delta G^{\neq}$ value is approximately of 7.7 kcal/mol and the calculated entropic effect is negative (-9.0 e.u.), in agreement with the experimental value (-9.1 e.u.).

We can conclude that our QM-MM implementation is capable of providing insight in chemical processes in complex environments. Moreover, due to the computational efficiency of our implementation, we have been able to obtain for the first time QM-MM free energy profiles at the DFT level.[36]

[1] Bartlett, P. A.; Nakagawa, Y.; Johnson, C. R.; Reich, S. H.; Luis, A. J. J. Org. Chem. 1988, 53, 3195.

[2] Haslam, E. Shikimic Acid: Metabolism and Metabolites; John Wiley & Sons: New York, 1993.

[3] Andrews, P. R.; Smith, G. D.; Young, I. G. Biochemistry 1973, 18, 3492.

[4] Galopin, C. C.; Zhang, S.; Wilson, D. B.; Ganem, B. Tetrahedron Lett. 1996, 37, 8675.

[5] Wiest, O.; Houk, K. N. J. Am. Chem. Soc. 1995, 117, 11628.

[6] Wiest, O.; Montiel, D. C.; Houk, K. N. J. Phys. Chem. A 1997,101, 8378.

[7] Kast, P.; Tewari, Y. B.; Wiest, O.; Hilvert, D.; Houk, K. N.; Goldberg, R. N. J. Phys. Chem. B 1997, 101, 10976.

[8] Davidson, M. M.; Guest, J. M.; Craw, J. S.; Hillier, I. H.; Vincent, M. A. J. Chem. Soc., Perkin Trans. 2 1997, 1395.

[9] Khanjin, N. A.; Snyder, J. P.; Menger, F. M. J. Am. Chem. Soc. 1999, 121, 11831.

[10] Martí, S.; Andrés, J.; Moliner, V.; Silla, E.; Tuñon, I.; Bertrán, J.; Field, M. J. J. Am. Chem. Soc. 2001, 123, 1709.

[11] Martí, S.; Andrés, J.; Moliner, V.; Silla, E.; Tuñon, I.; Bertrán, J. J. Phys. Chem. B 2000, 104, 11308.

[12] Worthington, S. E.; Roitberg, A. E.; Krauss, M. J. Phys. Chem. B 2001, 103, 7087.

[13] Roitberg, A. E.; Worthington, S. E.; Holden, M. J.; Mayhew, M. P.; Krauss, M. J. Am. Chem. Soc. 2000, 122, 7312.

[14] Lee, Y. S.; Worthington, S. E.; Krauss, M.; Brooks, B. R. J. Phys. Chem. B 2002, 106, 12059.

[15] Woodcock, H. L.; Hodoscek, M.; Sherwood, P.; Lee, Y. S.; Schaefer, H. F., III; Brooks, B. R. Theor. Chem. Acc. 2003, 109, 140.

[16] Carlson, H. A.; Jorgensen, W. L. J. Am. Chem. Soc. 1996, 118, 8475.

[17] Repasky, M. P.; Guimaráes, C. R. W.; Chandrasekhar, J.; Tirado-Rives, J.; Jorgensen, W. L. J. Am. Chem. Soc. 2003, 125, 6663.

[18] Guimaráes, C. R. W.; Repasky, M. P.; Chandrasekhar, J.; Tirado-Rives, J.; Jorgensen, W. L. J. Am. Chem. Soc. 2003, 125, 6892.

[19] Gou, H.; Cui, Q.; Lipscomb, W. N.; Karplus, M. Biochemistry 2001, 98, 9032.

[20] Ranaghan, K. E.; Ridder, L.; Szefczyk, B.; Sokalski, A. W.; Hermann, J. C.; Mulholland, A. J. Org. Biomol. Chem. 2004, 7, 968.

[21] Soler, J. M.; Artacho, E.; Gale, J. D.; García, A.; Junquera, J.; Ordejón, P.; Sánchez-Portal, D. J. Phys.: Condens. Matter 2002, 14, 2745.

[22] Perdew, J. P.; Burke, K.; Ernzerhof, M. Phys. Rev. Lett. 1996, 77, 3865.

[23] Wang, J.; Cieplak, P.; Kollman, P. A. J. Comput. Chem. 2000, 21, 1049.

[24] Eichinger, M.; Tavan, P.; Hutter, J.; Parrinello, M. J. Chem. Phys. 1999, 110, 10452.

[25] Chook, Y. M.; Gray, J. V.; Ke, H.; Lipscomb, W. N. J. Mol. Biol. 1994, 240, 476.

[26] Becke, A. J. Chem. Phys. 1993, 98, 5648.

[27] Lee, C.; Yang, W.; Parr, R. Phys. Rev. B 1988, 37, 785.

[28] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R.; Burant, J.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. Gaussian 98, Rev. A7, Gaussian, Inc.: Pittsburgh, PA, 1998.

[29] Baker, J.; Muir, M.; Andzelm, J. J. Chem. Phys. 1995, 102, 2063.

[30] Jarzynski, C. Phys. Rev Lett. 1997, 78, 2690.

[31] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J. Am. Chem. Soc. 1995, 117, 5179.

[32] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W.; DiNola, A.; Haak, J. R. J. Chem. Phys. 1984, 81, 3684.

[33] Cascella, M.; Raugei, S.; Carloni, P. J. Phys. Chem. B 2004, 108, 369.

[34] Torrie, G. M.; Valleau, J. P. J. Comp. Phys. 1977, 23, 187.

[35] Gallagher, D. T.; Mayhew, M.; Holden, M. J.; Howard, A.; Kim, K. J.; Vilker, V. L. Proteins. 2001, 44, 304.

[36] Crespo, A.; Martí, M. A.; Estrin, D.A.; Roitberg, A. J. Am. Chem. Soc. Comm. 2005, 127, 6940.

Chapter 4: NO Detoxification Reaction Catalyzed by Truncated Hemoglobin N

Introduction

Mycobacterium tuberculosis is the causative agent of human tuberculosis, a disease which infects about one-third of the human population and which causes more than a million deaths per year. Most of those who are infected live in third world countries, where the chances to survive are slim.[1] In healthy individuals, the infection is contained by the immune system, which forces the bacteria into dormancy. It has been reported that nitric oxide and related species are produced in the macrophages during the initial growth infection stage, and may be involved in restricting the bacteria in latency.[2] Both endogenous and exogenous NO have been shown to inhibit the development of parasites, such as Trypanosome, Leishmania, Plasmodium, and Toxoplasma,[3,4] probably due to the inactivation of parasite cystein proteases, and/or other key enzymes.

The toxic effects of NO can be reduced or even eliminated by the development of resistance mechanisms in microorganisms. One of such mechanisms consists in the oxidation of nitric oxide with heme bound O_2 to yield the innocuous nitrate ion, equation (4.1). This mechanism has been found to be operative in bacteria and yeast containing flavohemoglobins.[5,6] The same reaction has been implicated in vertebrate myoglobin and hemoglobin[7,8] and has been proposed to play an important role in the physiology of the recently discovered neuroglobin and cytoglobin heme proteins.[9,10]

$$\text{Fe(II)O}_2 + \text{NO} \;\rightarrow\; \text{Fe(III)} + \text{NO}_3^- \tag{4.1}$$

M. tuberculosis has been found to encode small heme proteins, referred to as truncated hemoglobins (trHb), which have been proposed to provide protection of bacilli against NO.[11] These proteins are widely distributed in bacteria, unicellular eukaryotes and higher plants, and constitute a distinct group within the hemoglobin superfamily.[12] Thus, they all show very low amino acid sequence homology to vertebrate and non-vertebrate Hbs, with sequence identities <15%.[13] The typical tertiary structure of trHb is based on a two-over-two helical sandwich,[14,15] compared with the three-over-three helical sandwich of the classical Hb fold. The proximal HisF8 heme linked residue is conserved throughout the Hb and trHb families. The main stabilizing residue in the distal cavity of trHb is typically TyrB10.[14] Two different truncated hemoglobins, known as trHbO and trHBN,[11] are produced by *M. tuberculosis*. While trHbO is expressed during the growth phase, trHbN expression is enhanced during the stationary phase.[11] Recent results from the Bolognesi and Guertin's groups show that trHbN appears to be better suited for performing NO/O_2 chemistry, and that it actively detoxifies NO, yielding the innocuous nitrate as

product.[14-19] X-ray trHbs structures reveal that trHbN host a tunnel/cavity system connecting the heme moiety with the exterior,[11,18] whose function is probably associated to controlling the ligand diffusion/storage properties. Moreover, it has been shown that these proteins can bind Xe atoms in the crystalline state, and that the Xe atoms map along the tunnel cavity system.[16]

In order to shed light on the molecular basis for the NO scavenging reaction of oxygenated trHbN from *M. tuberculosis,* we have performed classical molecular dynamics (MD) and hybrid quantum-classical (QM-MM) calculations. To this end, attention has been paid to both the intrinsic fluctuations of the protein and to the details of the heme-controlled oxidation of NO to gain insight into several fundamental issues on the enzyme mechanism. First, how is the channel designed to ensure fast entry of NO into the distal cavity, and how do thermal fluctuations affect its structure? Second, which are the key structural determinants for achieving the large oxygen affinity necessary to guarantee efficient detoxification in hypoxic environments? Thirdly, which are the main determinants for protein catalysis in the reaction of noncovalently bound NO with the Fe(II)O$_2$ moiety to yield nitrate? Fourthly, how can we characterize the factors controlling diatomic ligand diffusion through the tunnel system in trHbN? and finally, which is the mechanism for selective/distinct migration of O$_2$ and NO to the heme to achieve the most efficient NO detoxification?

Methods

Classical Molecular Dynamics

Simulations were performed starting from the crystal structure of wild-type oxy-trHbN, at 1.9 Å resolution (PDB entry 1idr).[15] The reported structure is homodimeric, but taking into account the fact that the interface area between the two trHb chains is small,[14] we have performed simulations only for the A subunit. The system was immersed in a preequilibrated box of TIP3P water molecules of about 60.0 x 50.0 x 70.0 Å.[20] Water molecules located at less than 2.4 Å from any atom of the protein were removed. Three Na$^+$ atoms were added to neutralize the system. The final system contains the model protein, 7176 water molecules and the added ions, leading to a total of 23518 atoms. Simulations were performed in the NPT ensemble. The system was simulated employing periodic boundary conditions and Ewald sums[21] for treating long range electrostatic interactions. The optimized system was then heated and equilibrated in three steps:

    (i)   100 ps of MD heating the whole system from 100 to 200 K.

    (ii)   heating of the entire system from 200 to 298 K for 100 ps.

    (iii)   equilibration of the entire system for 400 ps at 298 K.

The equilibrated structure was the starting point of a 10 ns MD simulation at constant temperature (298 K). SHAKE was used to keep bonds involving H atoms at their equilibrium length.[22] This allowed us to employ a 2 fs time step for the integration of Newton's equations. Amber99[23] and TIP3P[20] force fields were used to describe the protein and water, respectively. The oxygenated heme model system charges were determined using RESP[24] charges and HF/6-31G(d) wave functions following the protocol recommended in the Amber web page. The van der Waals parameters were taken from Amber99. Even if the presence of the transition metal introduces an additional challenge in the simulation, MD simulations are now commonly applied for the investigation of metalloproteins.[25] Frames were collected at 1 ps intervals, giving rise to a total of 10^4 frames, which were subsequently used to analyze the trajectory.

Finally, a 10 ns MD simulation at constant temperature (298 K) was performed for the deoxy form of the trHbN, employing the same protocol as described before, where the O_2 molecule was removed from the initial structure.

Cavity Calculations

The cavity of the ligand diffusion tunnel was determined using the SURFNET protocol as developed by Laskowski.[26] The shape of the cavity was determined by defining a regular grid of 0.2 Å spacing, where each grid point was assigned a value of 1 (inside the cavity) or 0 (outside the cavity), depending on its accessibility. The 3D matrices defined by the grids were then used to define the accessible volume.

Classical Molecular Interaction Potential (cMIP)

cMIP calculations were used to locate nitric oxide molecules in hidden regions of the protein, as well as to determine the ability of the protein to interact with NO. For this purpose, cMIP[27] calculations were carried out using NO as probe. The origin of the grid was located at the center of the active site (heme group), and a spacing of 0.35 Å was used. The grid was extended long enough to cover the whole channel (approximately 30 Å). A polar grid of 20 degrees was used to study the different orientations of the probe molecule in a given grid element.

Essential Dynamics

The set of structures collected along the 10 ns MD simulation were used to explore the dynamical behavior of the protein. To this end, residues 1-15, which form the N-terminal domain, were excluded, since this region was found to be quite flexible (see below), but does not contribute to the structure of the trHb fold nor to the definition of the ligand diffusion tunnel. Then, all the backbone atoms in the rest of the residues were retained and used to superpose the sampled structures. These structures were subsequently used to determine the essential motions, that is,

the structural changes that explain more conformational variability in the dynamics of the protein skeleton along the trajectory. To this end, covariance matrices were built up and diagonalized. The eigenvectors define the type of essential motions, and the associated eigenvalues determine how much of the positional variance in the trajectory is explained by each eigenvector.[28-31]

QM-MM Calculations

QM-MM calculations[32-37] were carried out to explore the effect of protein (and solvent) environment on O_2 binding and in the NO oxidation. There are several QM-MM schemes available, differing mainly in the electronic structure level treatment of the QM subsystem and the force field implemented. Here we employ our own QM-MM implementation where the QM subsystem is treated at the DFT level using the SIESTA implementation.[38] The SIESTA method has shown an excellent performance for medium and large systems, and has also proved to be appropriate for biomolecules, and specifically for heme models.[39] The use of standard norm-conserving pseudopotentials[40] avoids the computation of core electrons, smoothing at the same time the valence charge density, as described in Chapter 2. In the present study, the nonlinear partial-core correction[41] is applied to the iron atom. For all atoms, basis sets of double plus polarization quality were employed, with a pseudoatomic orbital energy shift of 30 meV and a grid cutoff of 150 Ry.[38,39] Calculations were performed using the generalized gradient approximation functional proposed by Perdew, Burke and Ernzerhof.[42] This combination of functional, basis sets, and grid parameters has been already validated for heme models.[39] The classical subsystem was treated using the Amber99 force field parametrization.[23]

The initial structure for QM-MM calculations was taken from a selected snapshot collected along the classical MD simulation. The NO molecule, which was not considered in the classical MD system, was added at high affinity sites determined from CMIP calculations, close to the heme active site. Since our QM-MM code cannot manage periodic boundary conditions, we solvated this structure with a cap of 7490 water molecules. This structure was thermalized at 300 K. Then the system was cooled down slowly to 0 K, and subsequently optimized using Amber99. We have selected the iron porphyrinate plus the NO and O_2 ligands and the axial histidine as the quantum subsystem. The rest of the protein and the water molecules were treated classically. The final system consisted of 50 QM atoms and 24411 MM atoms. We allowed free motion for QM atoms and for the 809 MM atoms located inside a sphere of 13.5 Å from the QM subsystem center of mass. The frontier between the QM and MM portions of the system was treated by the SPLAM method as described in Chapter 2.[43,44] Link atoms were used in the proximal histidine to separate the QM treated imidazole ring from the peptidic backbone. The ferrous unbound pentacoordinated heme group isolated or in the protein was treated as a high spin (HS) quintuplet state, which is known to be the ground state for this system.[45,46] The ferrous O_2 complex was treated as a low

spin singlet state, which is known to be the ground state.[45,46] All species involved in reaction (4.1) were treated in the low spin doublet and high spin quartet states. More technical details about the QM-MM implementation can be found elsewhere.[47]

Ligand affinities (ΔE_L) were calculated using equation (4.2), where E_{Enz-L} is the energy of the ligand bound enzyme, E_{Enz} is the energy of the ligand free enzyme and E_L is the energy of the isolated ligand:

$$\Delta E_L = E_{Enz-L} - E_L - E_{Enz} \tag{4.2}$$

The reaction intermediates formed in the oxidation process of NO to nitrate anion were fully optimized without imposing any constraints. For the model systems used in calculations *in vacuum*, the minimum energy nature of the stationary points was subsequently verified by normal mode analysis. Obtaining accurate free energy profiles requires an extensive sampling, which is computationally very expensive and difficult to achieve at the DFT QM-MM level. For these reasons potential energy profiles were determined using restrained energy minimizations along the reaction path that connects reactant, intermediate and product states, as described in Chapter 2. This procedure, which has provided valuable information in previous QM-MM reaction paths calculations,[47] is used here to obtain qualitative insight into the effect of the protein environment, and particularly on the role played by Tyr33 by comparing the energy profiles obtained for the wild type enzyme and the Tyr33 → Phe mutant. For this purpose, an additional term was added to the potential energy according to $V(\xi) = k\ (\xi - \xi_0)^2$, where k is an adjustable force constant, 200 kcal/(molÅ^2), ξ is the value of the reaction coordinate in the system particular configuration and ξ_0 is the reference value of the reaction coordinate (see below for the choice of the reaction coordinate in the different reactions). Varying ξ_0, the system is forced to follow the minimum reaction path along the given coordinate.

Classical Free Energy Profiles

In order to study in detail the properties of the tunnel cavity systems in trHb N from *M. tuberculosis*, we have calculated the diffusion free energy profiles for a diatomic neutral ligand along the tunnels for the oxy and free protein. The free energy profile was constructed by performing constant velocity MSMD runs, using Jarzynski's expression, as described in Chapter 2.[48] This technique has already been successfully applied to the study of enzyme catalysis studies (see Chapter 3).[49]

In the present study, the chosen reaction coordinate λ was chosen as the iron-ligand distance. The force constant used was 200 kcal mol^{-1} Å^{-1}. The pulling velocities used were 0.05 Å/ps and 0.1 Å/ps. To reconstruct the free energy profile of diffusion of the ligand along the tunnels, the following sets of MSMD runs were performed: starting from equilibrated MD structures with $\lambda(t=0)$

corresponding with the ligand in distal pocket, the crystallographic Xe-binding sites and the ligand outside the tunnels, 10 MSMD runs were performed in each direction (forward/exit and backward/entry) for each of the two pulling velocities. In cases in which two overlapping profiles were obtained (from entry and exit sets), we confirmed that both of them matched.

In order to asses the difference in the ligand entry/exit into the channel due to Phe62 conformations, the free energy profile for the opening/closing transition of Phe62 was calculated using umbrella-sampling techniques[50] with and without oxygen, using the Phe62 C_α-C_β dihedral angle as distinguished coordinate and a set of 12 windows of 1 ns each (see Chapter 2).

Results and Discussion

Structural flexibility of the protein

The trHbN protein exhibits the characteristic structural features of the trHb fold, mainly based on the B, E, G and H α-helices of the classical globin fold (see Figure 4.1).[14] However, comparison with the sperm whale myoglobin indicates that only 54 C_α pairs can be matched to trHbN, with a root-mean square deviation of 2.1 Å,[51] which indicates the existence of notable structural deviations. Among the features of the trHb fold that deviate from the classical globin fold, it can be noticed the presence of an extended loop substituting for most of the F-helix, the deletion of the A-helix, which is reduced to a one-turn A-helix tying the N-terminal region to the protein core, and a 12-residue insertion (pre-A region) at the trHbN terminus, which protrudes from the compact protein fold (see Figure 4.1).[14,15]
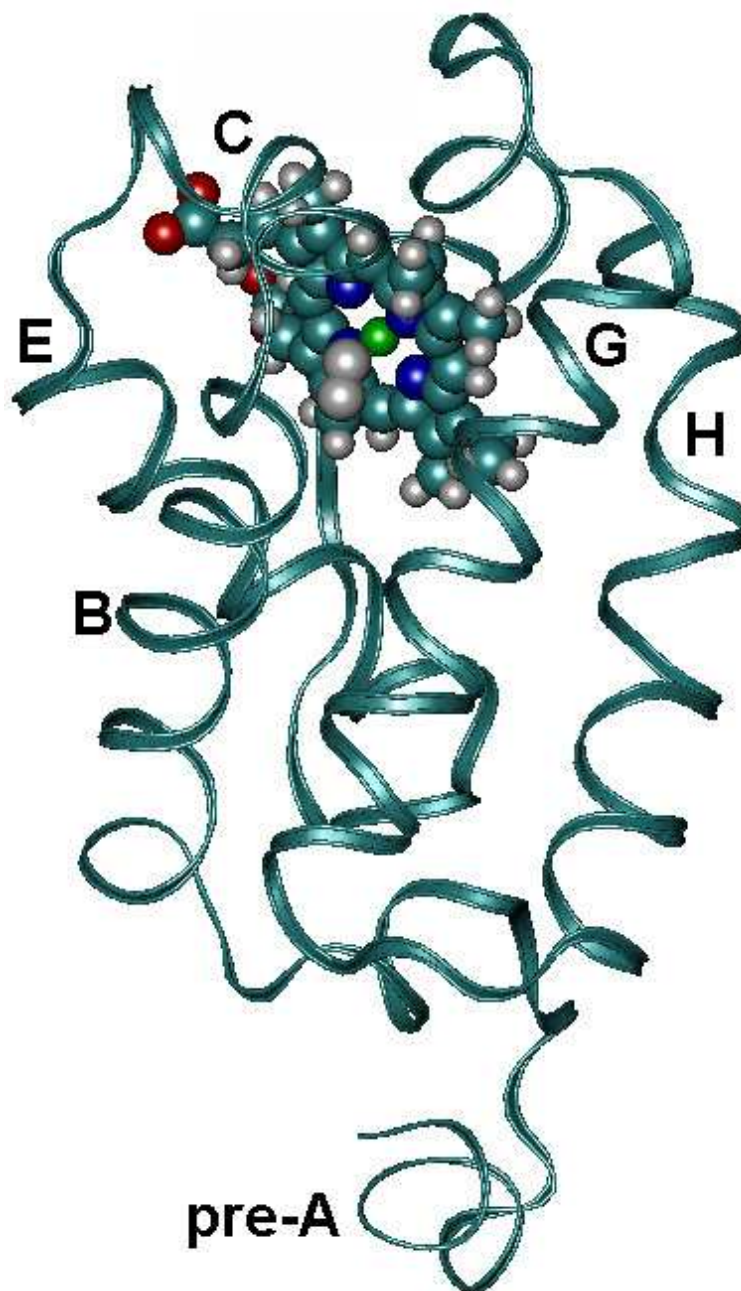
Figure 4.1. Ribbon representation of the structural elements that form the trHb fold.

MD calculations for the native protein in aqueous solution provide a stable trajectory for the 10 ns simulation time, as can be seen by inspecting the fluctuations in both the potential energy and the positional root-mean square deviation (RMSD; computed using the subunit A of the crystal structure of 1idr as the reference structure) determined for the heavy atoms in the protein along the simulation (Figure 4.2). It is worth noting that a large fraction of the structural flexibility stems from the N-terminal (residues 1-15) region, whose RMSD (around 2.6 Å) accounts for most of the

total protein RMSD (see Figure 4.2). This enhanced flexibility can be realized from the lack of the elongated α-helix expected in a conventional Hb fold, the protrusion of the N-terminal region from the protein core, and the favorable hydration of the highly polar sequence motif in the 6-11 (Arg-Leu-Arg-Lys-Arg-Glu) pre-A region. In contrast, the main fold of the protein core (residues 16-127), which comprises the heme active site, is well preserved and exhibits smaller RMSD values (around 1.6 Å).
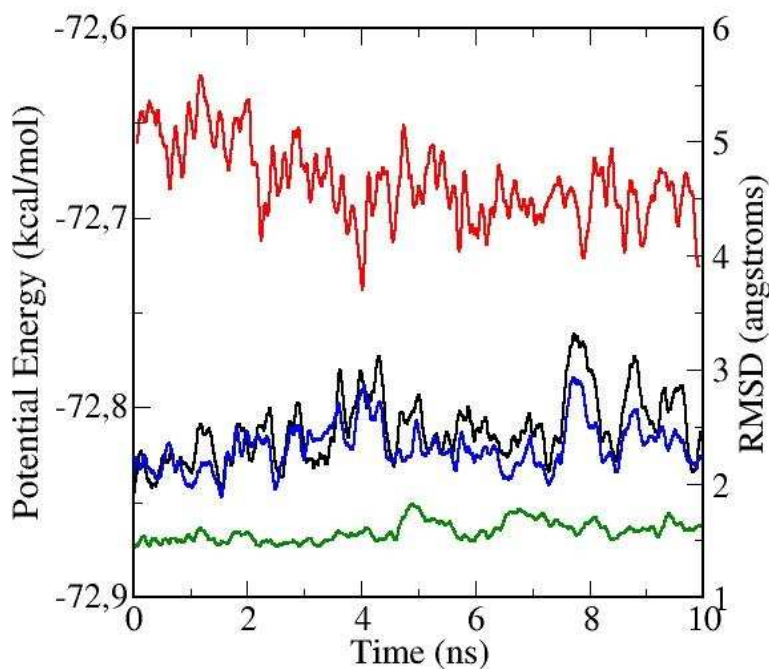
Figure 4.2. Time evolution of the RMSD (Å) of the trHbN protein (blue), the N-terminal region (residues 1-15; black) and the protein core (green), which includes the heme active site, as well as the total potential energy of the entire system (kcal/mol; red).

To gain insight into the dynamical behavior of trHbN, the essential motions of the protein core were examined by diagonalizing the positional covariance matrix determined for the backbone atoms. The calculated eigenvalues decrease in magnitude smoothly, thus indicating that a variety of motions determine the structural flexibility of the protein backbone, as noted previously for other proteins.[51-55] Thus, near 44% of the backbone conformational flexibility is accounted for when the ten principal components are considered, and only the first three normal motions contribute separately more than 5% to the structural variance of the backbone. The first essential movement (8% of structural variance) mainly involves the relative motion of the B and E helices (see Figure 4.3). The second principal motion (7.3%) mostly affects helix C and the hinge regions between E-F and F-G helices. Finally, the third component (6.4%) involves fluctuations in the hinge region between B and C helices, as well as displacements in B, G and H helices. Overall,

these results indicate that the backbone fluctuations in trHbN arise from a complex interplay of movements which involve the different secondary structure elements.
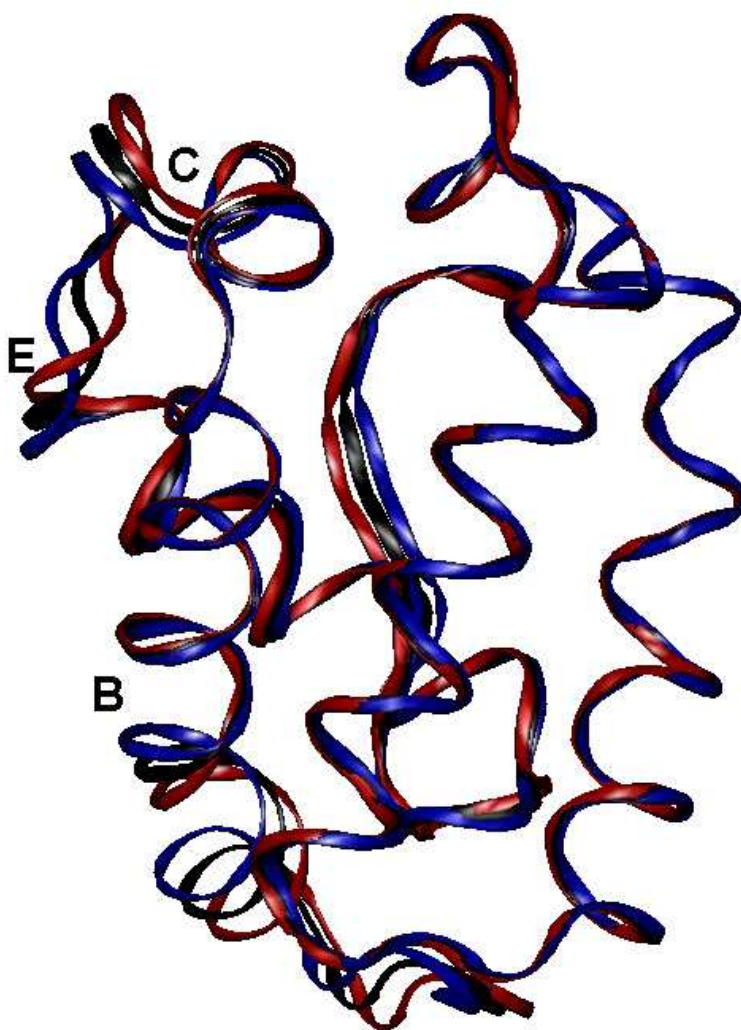

Figure 4.3. Ribbon representation of the structural displacement of B and E helices in trHbN originated from the first essential motion of the protein core backbone. The figure represents the average backbone structure of the protein core (black), as well as those structures (red and blue) corresponding to the maximum displacements along the first essential movement.

As noted by Bolognesi and co-workers,[14,15] one of the most striking features in trHbN is the presence of an almost continuous tunnel through the protein matrix that connects the heme distal pocket to the protein surface at two distinct sites. The tunnel is composed by two orthogonal branches (Figure 4.4), stretching around 20 and 8 Å, respectively, from the access sites to the heme ligand site. The inner surface of the largest tunnel branch is mainly defined by apolar residues pertaining to the B (Ile19, Ala24, Ile25, Val28, Val29, Phe32) and E (Phe62, Ala63,

Leu66) helices. For the shortest branch, it also involves apolar residues in G (Ala95, Leu98) and H (Leu116, Ile119) helices.
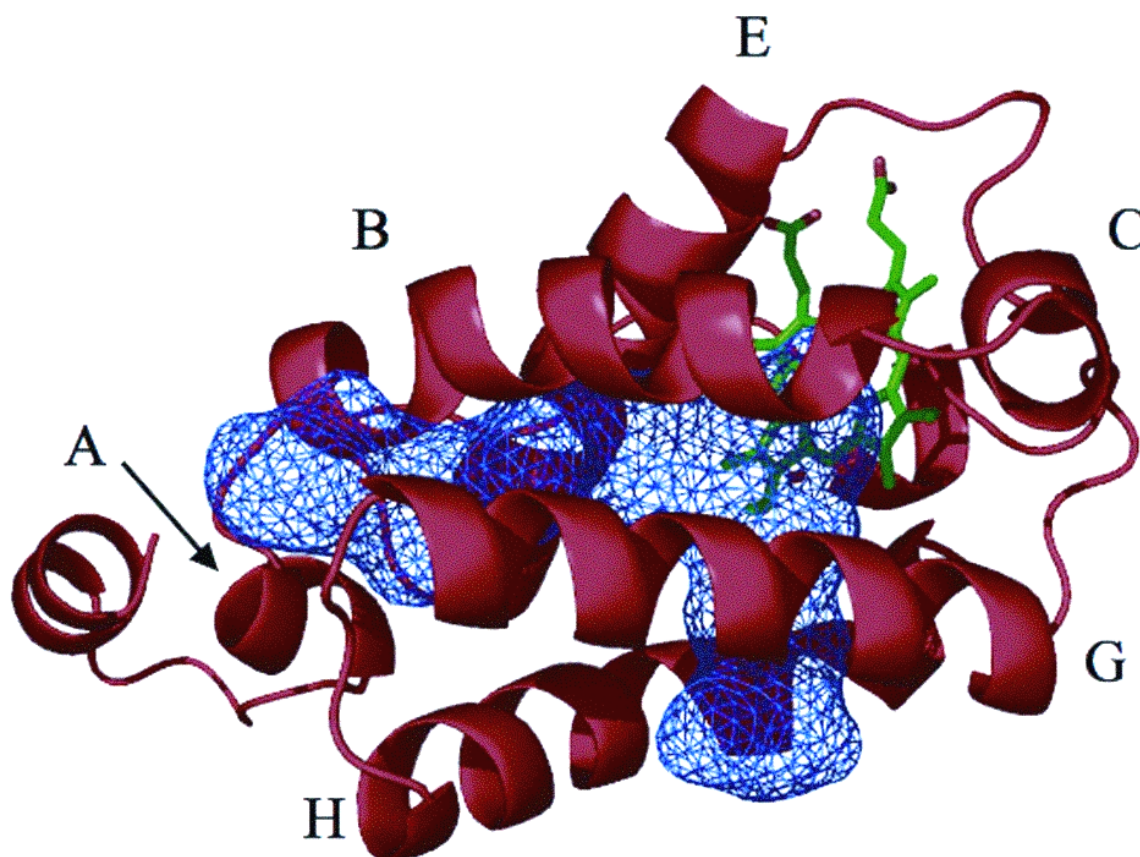
Figure 4.4. X-ray crystal structure of oxy-trHbN used in simulations, showing the labels of the helices, the location of the heme group (green) and the two branches of the tunnel system (blue mesh). The long branch (≈20 Å long) is horizontal in the figure, and can be seen as roughly perpendicular to the heme distal face. The short branch (≈8 Å long) lies roughly parallel to the heme plane, accessing the heme distal site pocket from the inner part of the heme crevice.

Inspection of the fluctuations experienced by those residues along the simulation revealed that the opening of the passage through the tunnel is mainly controlled by Phe62, a residue which lies close to Phe32 and Leu98. In fact, the analysis of the MD simulation showed that Phe62 populates two main conformations, which are characterized by average torsional angles around the C_α-C_β bond of about 40 and -50 degrees (Figure 4.5). Interestingly, Milani et al.[15] have pointed out that this residue is also observed in two alternate conformations differing by ≈63 degrees by rotation around the C_α-C_β bond in the X-ray crystallographic structure. When the C_α-C_β torsional angle is close to 40 degrees, which is the most populated conformation sampled along the trajectory, there is no apparent restriction to the conformational flexibility around the C_β -

C_γ bond (see Figure 4.5). However, when the C_α-C_β dihedral angle is close to -50 degrees, the conformational flexibility around the C_β -C_γ bond is more restricted, which can be attributed to unfavorable contacts with Ile25 and Ile119.
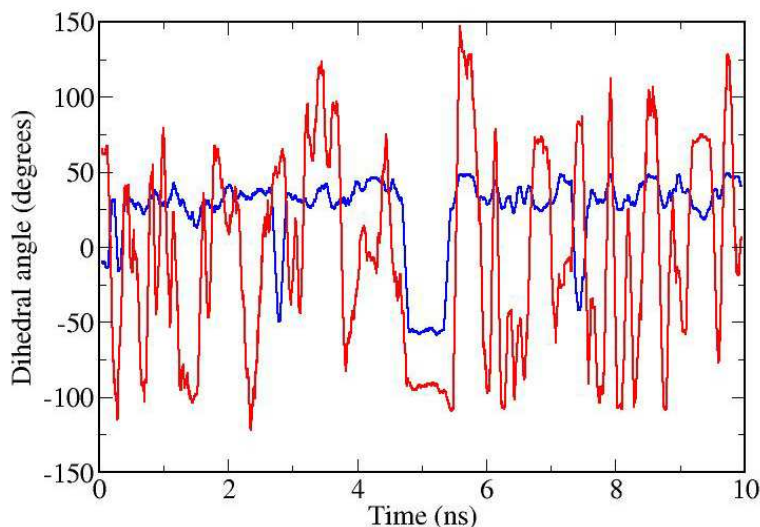

Figure 4.5. Time evolution of the Phe62 C_α-C_β (blue line) and C_β-C_γ (red line) dihedral angles.

The analysis of the MD trajectory also revealed that the adoption of two different conformations around the C_α-C_β torsional angle of Phe62 is crucial for the opening of the channel leading to the active site. Thus, when the torsion around the C_α-C_β bond is ≈40 degrees, the phenyl ring of Phe62 enters the ligand diffusion tunnel and impedes the access of the ligand to the active site. However, when the dihedral angle around the C_α-C_β bond changes from ≈40 degrees to ≈-50 degrees, the benzene ring roughly lies parallel to the axis of the tunnel and there is a drastic reduction in the steric hindrance for the ligand diffusion through the tunnel. This is reflected in the distance from Phe62 to Phe32, which is generally in the range 3-5 Å in the former conformation, but increases up to 9 Å when such a conformational change takes place (see Figure 4.6). These two conformational states might, therefore, be associated to *closed* and *open* states of the channel (see Figure 4.7). In the closed conformation the tunnel volume amounts, on average, to ≈350 Å^3, which is similar to the value obtained for the X-ray crystallographic structure.[15] The transition from the closed state to the open one is associated to an increase in the tunnel volume, which amounts on average to ≈78 Å^3, a value somewhat smaller than the range of van der Waals volumes typically assigned to methylbenzene (from 95 to 102 Å^3).[56] On the other hand, such a conformational transition increases the width of the tunnel, which is found to reach values up to ≈ 3.4 Å as determined from the shortest contact distance between the van der Waals surfaces of the residues that define the bottleneck of the tunnel.
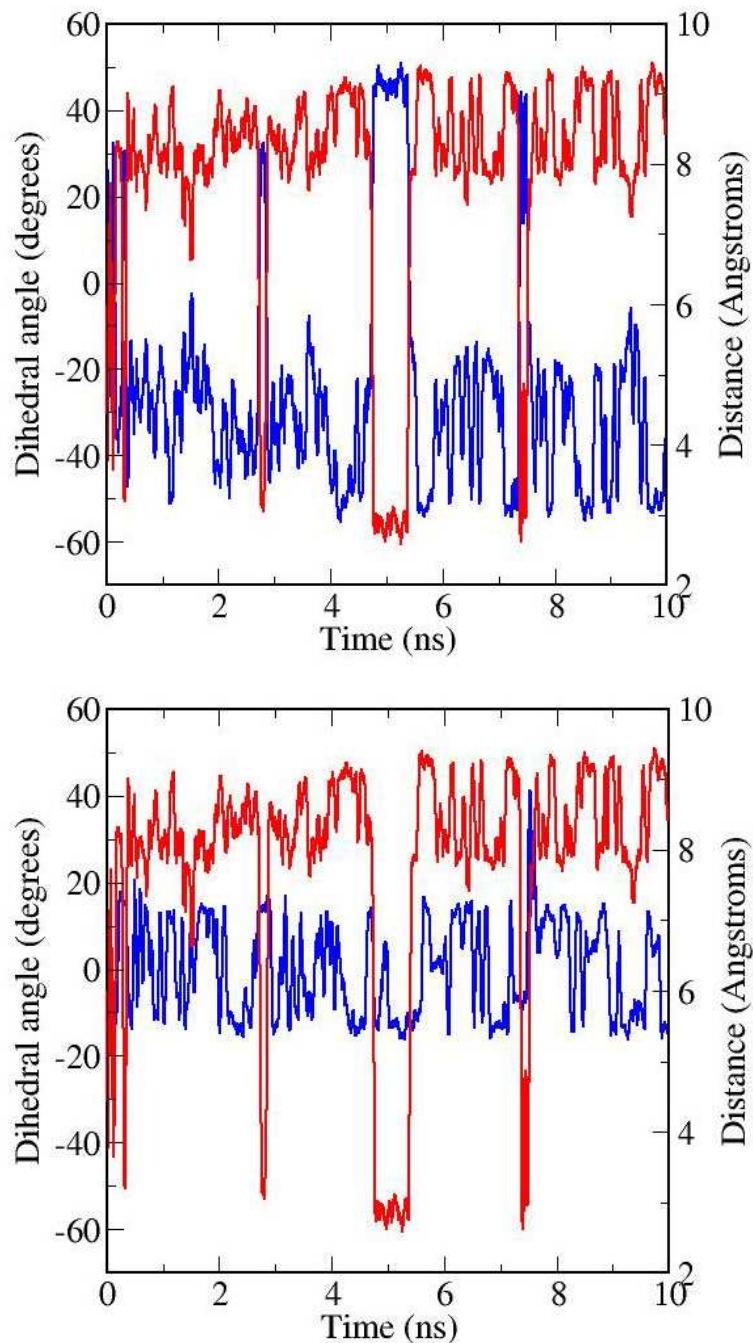
Figure 4.6. Time evolution of the Phe62 C_{α}-C_{β} dihedral angle (red line), and of the distances from (top) Phe62 and Phe32 (blue line; measured from atom HZ in Phe62 to atom HE21 in Phe32) and (bottom) Phe62 and Leu98 (blue line; measured from atom HZ in Phe62 to atom HD21 in Leu98).
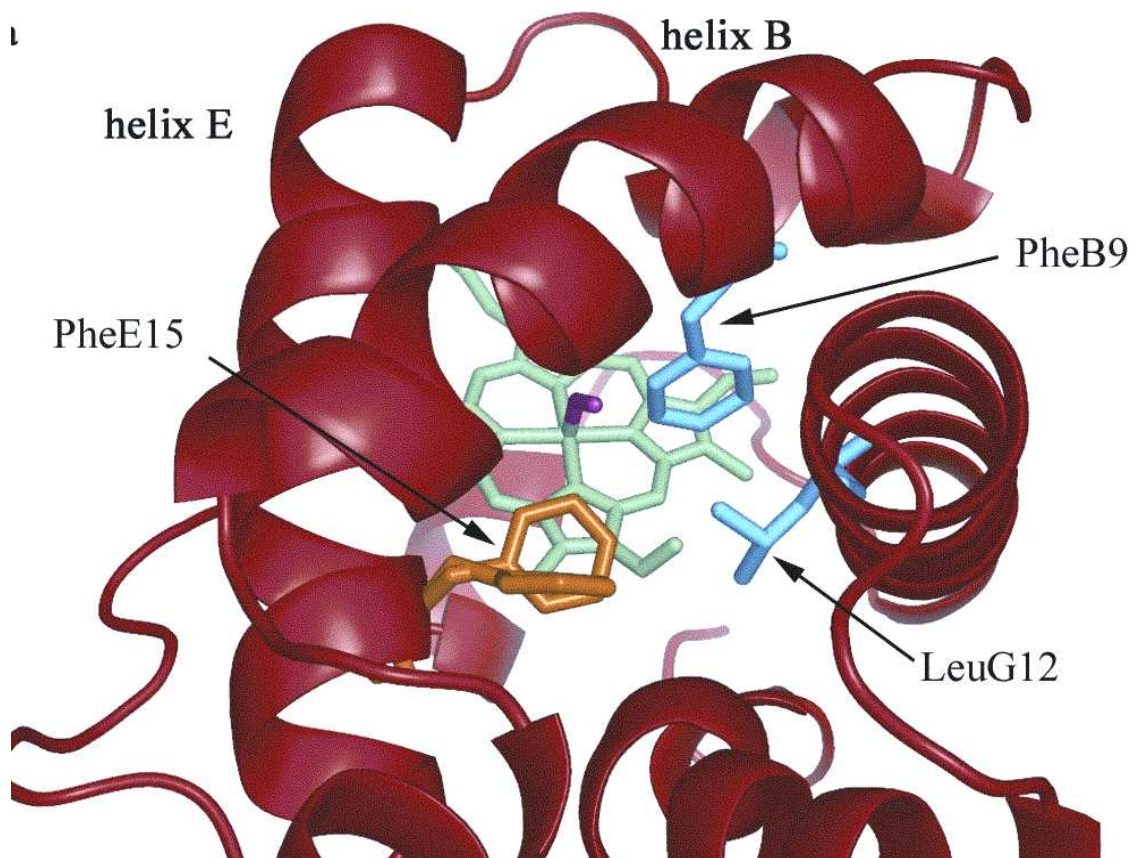
Figure 4.7. Diatomic ligand access to the heme group in trHbN. Conformational change of Phe62 (orange) side chain along the C_α-C_β dihedral angle leading to *open* and *closed* states viewed along the tunnel long branch leading to the heme (green) in oxy-trHbN. The residues that define the bottleneck to ligand migration (Phe32, Leu98; blue) in the *closed* state are also shown.

The finding that the adoption of the two conformations around the C_α-C_β bond leads to open and closed configurations was further verified by delineating the channel by using 3D grids defined according to Laskowski's method (Figure 4.8).[26] For the open conformation, the surface of the channel leading from the heme group to the outer space is well defined, without discontinuities. Moreover, it clearly shows that the adoption of the open conformation leads to opening of the two branches of the tunnel leading to the heme binding site. In contrast, the walls of the cavity are broken in those snapshots where Phe62 adopts a closed conformation (Figure 4.8).
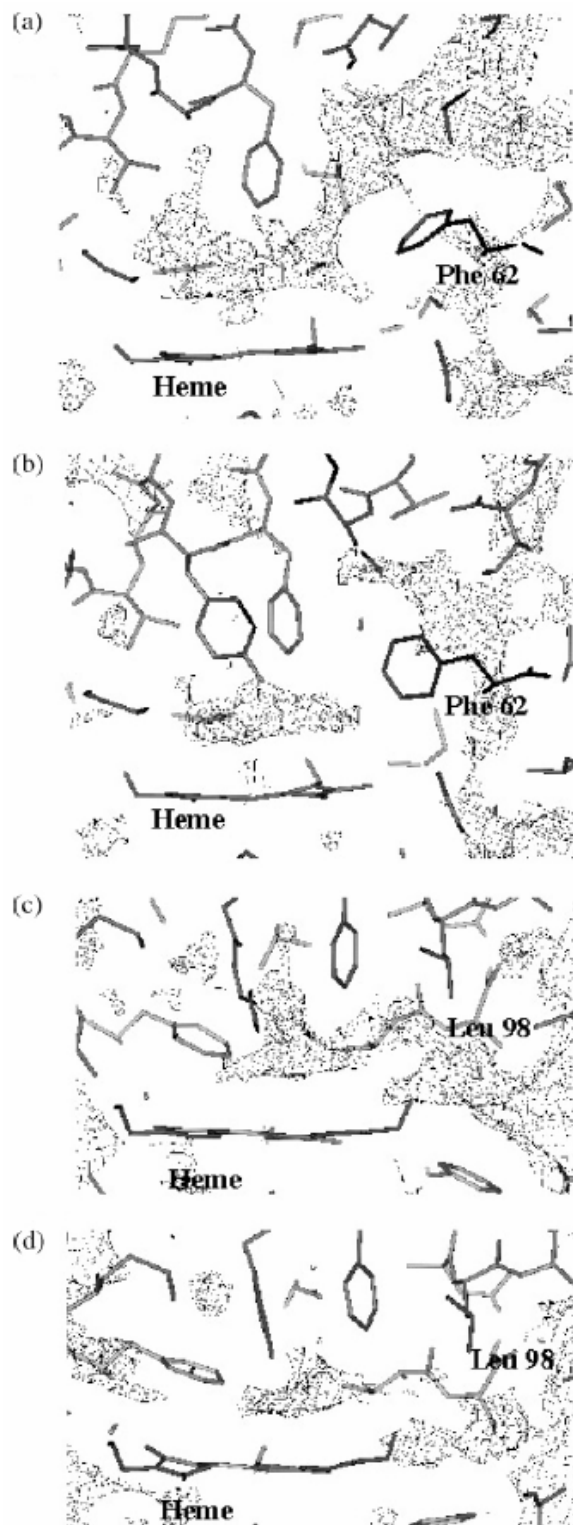
Figure 4.8. Representation of the ligand diffusion tunnel in trHbN. (a) and (b) correspond to views of the largest branch of the tunnel in the open and closed channel conformations, respectively, whereas (c) and (d) show the corresponding views for the shortest branch of the tunnel in the open and closed conformations.

Similar findings can be obtained from inspection of the isocontour plots determined from CMIP calculations for the interaction of a NO probe with the protein. Calculations performed for both open and closed states of the protein reveal the existence of a favorable interaction energy region in the distal side of the heme group, located between the iron atom and Tyr33 (Figure 4.9). Such a region also protrudes towards the intersection of the two tunnel branches, thus occupying a space in the cavity that coincides with one of the secondary Xe binding sites recently found by Bolognesi´s group upon treatment of trHbN crystals under Xe pressure.[16] There is, however, a fundamental difference in the isocontour plots determined for the open and closed conformations of the protein. When the protein adopts an open conformation (see Figure 4.9, top), there is a continuous energy isocontour that delineates the channel leading from the tunnel entrance to the heme binding site. In contrast, for those structures corresponding to the closed state, such an energy isocontour is broken due to the phenyl ring of Phe62, which enters the long branch of the tunnel (see Figure 4.9, bottom). In this latter case, it is worth noting the existence of a favorable interaction energy region located in the tunnel long branch facing the Phe62 benzene ring, which is also in agreement with the main Xe binding site recently identified by Bolognesi et al.[16]
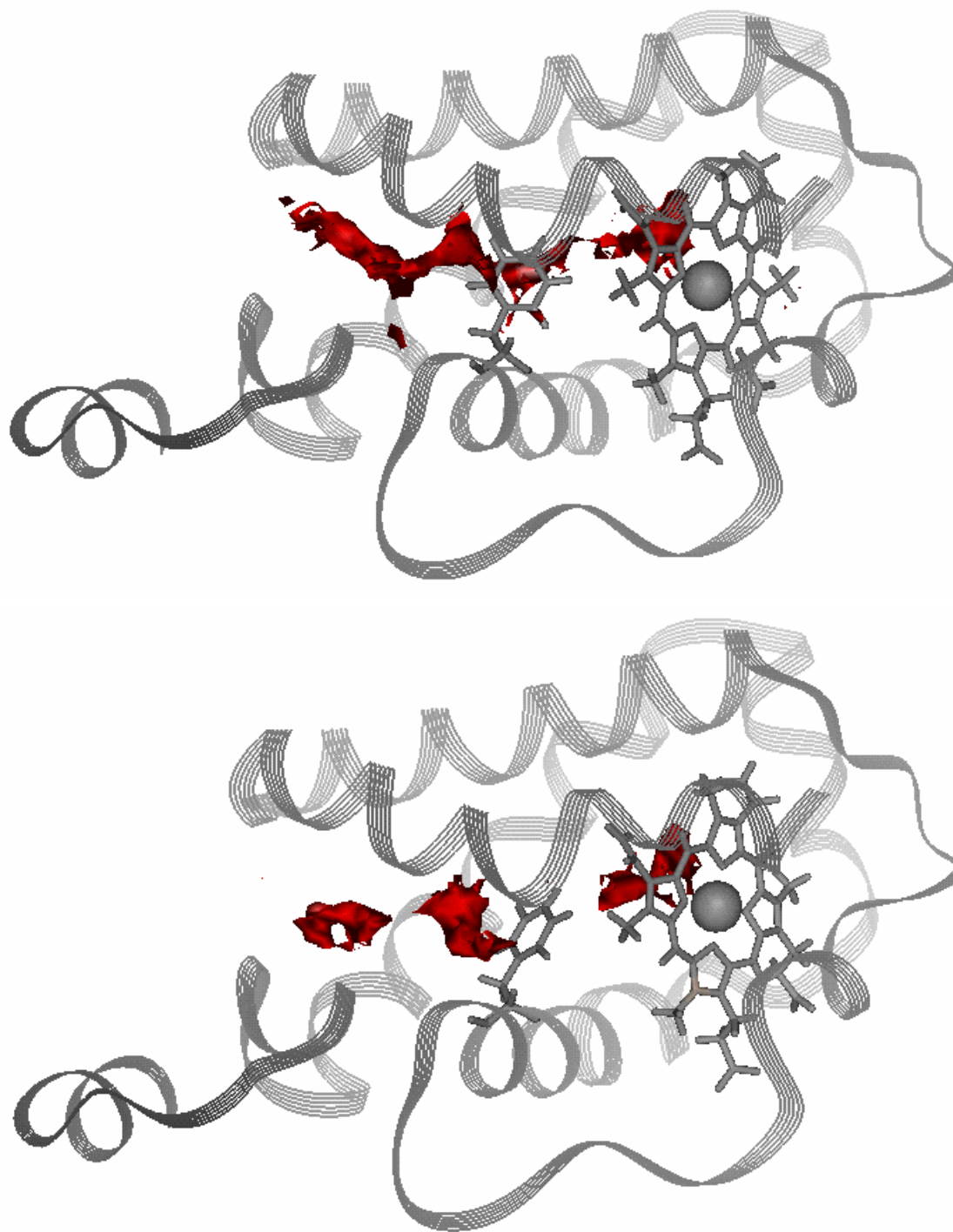
Figure 4.9. Isocontour plots corresponding to the interaction energy between the NO probe and the protein determined from classical molecular interaction potential (CMIP) calculations for selected snapshots representative of the *open* (top) and *closed* (bottom) conformations, which are shown in comparable orientations. Calculations were carried out in the absence of the O_2 molecule. Both Phe62 and the heme group (with the iron atom represented by a sphere) are explicitly displayed. Contour plots correspond to -2 kcal/mol.

The preceding findings point out that the main bottleneck to the passage of ligands through the channel comes from the relative arrangement between the side chains of Phe32, Phe62 and Leu98, which pertain to E, B and G helices, respectively. Moreover, Phe62 can be considered to be the gate of the main channel leading to the heme group in the active site. Even though the limited length of the simulation precludes to perform a statistical analysis of the ratio between the population of open and closed conformations of the channel, inspection of Figure 4.6 clearly points out that the conformational change around the C_α-C_β bond of Phe62 is not a rare event. In fact, several conformational transitions occurred spontaneously along the simulation, and in one case the main channel remained in the open configuration for around 0.8 ns. Those transitions can be expected to originate from the fluctuations in the protein structure, which can be characterized in terms of the separate essential motions that determine the intrinsic flexibility of the protein. It is therefore reasonable to expect that when those essential motions act in a concerted way and have the appropriate magnitude, the collective structural fluctuation experienced by the protein backbone might trigger the conformational change in Phe62, thus leading to the transition between open and closed states.

To explore the connection between the local motions of the side chain of Phe62 and the essential motions that modulate the structural fluctuations of the protein backbone, we determined the correlation between the displacements of the backbone C_α atoms and the fluctuation in the distance from Phe62 to either Phe32 or Leu98. This was accomplished by using equation (4.3),[57] where $r_{\tau i}$ is the correlation coefficient between the $\tau(t)$ (τ = x, y, or z) coordinates of the C_α atom of residue i at time t, and $\rho(t)$ denotes the distance between residues.

$$r_{\tau i} = \frac{\left\langle (\tau_i(t) - <\tau_i>_t)(\rho_i(t) - <\rho_i>_t) \right\rangle}{\sqrt{\left\langle (\tau_i(t) - <\tau_i>_t)^2 \right\rangle_t \left\langle (\rho_i(t) - <\rho_i>_t)^2 \right\rangle_t}} \qquad (4.3)$$

No significant correlation was found when the interresidue distance parameters, $\rho(t)$, were compared with the position of the C_α atoms projected along each of the main principal components that define the essential dynamics of the protein. This indicates that the fluctuations between the residues that define the bottleneck in the ligand diffusion tunnel cannot be attributed to a specific motion between structural elements. A clear correlation was, however, found between the fluctuations in the Phe62-Leu98 distance and the positional displacements of the C_α atoms for around 50% of the residues in the protein, mainly pertaining to helices B, C and F, and segments of helices E and H. This finding, therefore, suggests that the opening of the channel between residues Phe62 and Leu98 stems from a concerted motion of a large fraction of the protein. Finally, no clear correlation was found between the Phe62-Phe32 distance and the positional fluctuations of the C_α atoms. This finding suggests that the mechanism leading to opening of the passage through the ligand diffusion tunnel stems from the interplay between

collective motions of the protein backbone and changes in dihedral angles of specific residues, particularly Phe62, which would eventually facilitate the transition from the closed conformation to the open one.

Finally, we have analyzed the structural fluctuations in the vicinity of the heme active site. The O_2 molecule bound to the Fe atom interacts with Tyr33 through a hydrogen-bond along the whole trajectory (average OH-O1 and OH-O2 distances of 3.27 and 2.76 Å, respectively). This interaction has been proposed to play an important role in the stabilization of the heme bound oxygen.[17] On the other hand, a much weaker interaction is found between O_2 and Gln58 (average NE-O1 and NE-O2 of 3.67 and 3.82 Å, respectively), which is in agreement with the experimental evidence.[15] In fact, the analysis of the MD trajectory revealed a hydrogen-bond contact between Gln58 and Tyr33 (average NE-O distance of 3.00 Å). These results, therefore, give support to the suggestion[14,15] that Gln58 is crucial in the detoxification reaction, because it might interact with the NO molecule and contribute to position correctly NO to react with O_2. At this point, CMIP calculations performed for selected snapshots in both open and closed conformations reveal the existence of a region favorable for the interaction with the NO probe in the proximity of the bound O_2 (see Figure 4.10). In fact, such a region corresponds to the most favorable interaction site for the NO probe in the whole ligand diffusion tunnel.
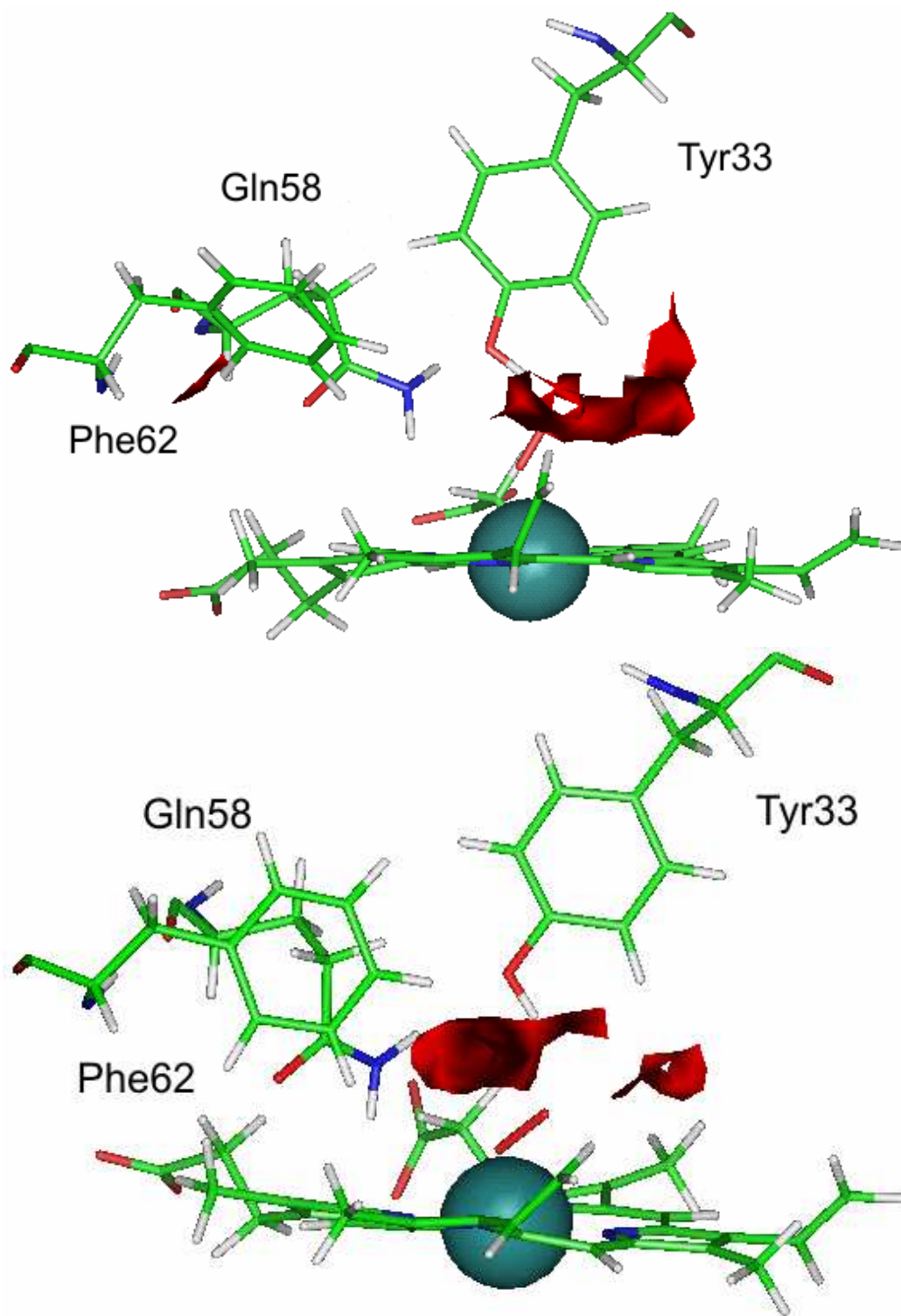
Figure 4.10. Isocontour plots corresponding to the interaction energy between the NO probe and the O_2-bound form of the TrHb N enzyme determined from classical molecular interaction potential (CMIP) calculations for selected snapshots representative of the open (top) and closed (bottom) conformations. Besides Tyr33, Gln58 and Phe62, the plots display the oxygen molecule (in red) and the heme group with the iron atom represented by a sphere. Contour plots correspond to -3 (open) and -2 (closed) kcal/mol.

Effect of the environment on the O₂ affinity of the heme group

A necessary requirement to grant survival of *M. tuberculosis* under hypoxic conditions when the bacilli enter latency is the high O_2 affinity of the enzyme (P_{50} ≈0.01 mm Hg), which has been attributed to the interaction of the bound oxygen with Tyr33.[11] Following previous studies,[58] our main interest here is to explore by means of electronic structure calculations the influence of the protein environment in modulating the O_2 binding to the trHbN enzyme. To this end, the geometrical and energetic features of the interaction between oxygen and the heme group in vacuum, the wild type (wt) protein, and the Tyr33 → Phe mutant have been determined. Table 4.1 reports selected properties for the bound and unbound states of the heme group in the preceding systems.

	Vacuum		Wt TrHb		Tyr33→Phe33 (mut)	
	Oxy	Free	Oxy	Free	Oxy	Free
Fe-O1	1.78	-	1.84 (2.19, 2.06)	-	1.86	-
O1-O2	1.29	-	1.31 (1.20, 1.20)	-	1.31	-
Fe -O1-O2	121.7	-	120.7 (122.4, 114.2)	-	120.4	-
O2-Tyr33$_{OH}$	-	-	2.75 (3.17, 3.09)	-	-	-
O1-Gln58$_{NE}$	-	-	3.52 (3.86, 3.77)	-	3.37	-
Tyr33$_{OH}$-Gln58$_{NE}$	-	-	2.98 (3.13, 2.83)	2.99	-	-
Fe-His81$_{NE}$	2.12	2.16	2.06 (2.10, 2.11)	2.13	2.09	2.10
Fe-Hem:N$_{plane}$	0.04	-0.27	0.02 (0.02, 0.02)	-0.25	0.02	-0.29
His81: tilt angle	88.1-90.4	97.9-98.5	87.6-91.9 (88.1-91.5, 89.2-92.0)	94.9-101.4	87.8-92.2	94.3-102.7
His81: rotation angle	12.4	13.7	33.9 (29.0, 29.0)	33.6	28.7	27.7
Δq(O₂)	-0.20	-	-0.36	-	-0.35	-
Δq(His81)	0.16	-	0.21	-	0.22	-
ΔE	-21.4		-37.2		-34.3	

Table 4.1. Relevant geometrical parameters (distances in Å, and angles in degrees), changes in Mulliken charges (in units of electron), and energetical changes (ΔE; in kcal/mol) for the binding of O₂ to the heme group (Hem + O₂ → Hem-O₂) in vacuum, in the wild protein and in the Tyr33 → Phe mutated protein determined from QM-MM calculations. Experimental values for the wt oxygenated protein (A and B chains) reported in Ref. 15 are given in parenthesis.

For the isolated heme group the binding energy amounts to -21.4 kcal/mol, which is somewhat lower than the binding energy reported by Rovira et al. from BLYP computations.[46] In trHbN calculations predict a substantial increase (by around 16 kcal/mol) of the O_2 binding in trHbN, though the distance from the Fe atom to the O_2 is enlarged by 0.06 Å (a distance of 1.84 Å is found for the wt enzyme; see Table 4.1)[59] compared to the model system in vacuum. Despite the

uncertainties that might arise from the computational approach used to estimate the binding energies, the preceding result support the crucial role played by the protein environment in modulating the O_2 affinity. Particularly, the hydroxyl group of Tyr33 is hydrogen bonded to O_2, with a distance between the hydroxyl oxygen and the O_2 molecule of 2.75 Å, thus supporting the implication of Tyr33 in assisting the binding of O_2. To further examine the specific role played by Tyr33, we have determined the binding of O_2 for the Tyr33 → Phe mutant. Though the Fe··O_2 distance remains little affected, the binding energy in the mutant is around 3 kcal/mol smaller that for the wt protein, which agrees qualitatively with the ≈150-fold increase in the O_2 dissociation rate in the mutant.[11,17]

Several studies in model heme compounds[60,61] have pointed out that the O_2 binding might be modulated through another mechanism which exploits the hydrogen-bond formed between the proximal His81 and the carbonyl group of Met77 (1.77 Å in the optimized structure for the wt enzyme; see Figure 4.11). Such a hydrogen bond augments the donation capacity of His81, increasing the Fe-His bond strength and in turn reinforcing the Fe-O_2 bond. This mechanism is supported by the results given in Table 4.1, since the optimized His-Fe coordination bond decreases by 0.06 Å in trHbN compared to the model system in vacuum. Indeed, the net Mulliken charge on the oxygen is also consistent with the strength of the bond: -0.20 and -0.36 e in vacuum and in the protein, respectively, showing a larger degree of back-donation from the metal in the latter case. Finally, the net positive charge on the proximal histidine is significantly larger in the protein compared with the isolated model system, confirming the enhanced donor capacity of this residue arising from hydrogen-bonding to Met77. In order to quantify this effect we performed an additional QM-MM calculation in which proximal electrostatic interactions were annihilated, thus eliminating the C=O Met77-His81 hydrogen-bond interaction. The computed oxygen affinity was reduced by 1.8 kcal/mol, showing that the hydrogen-bond effect is small but significant.
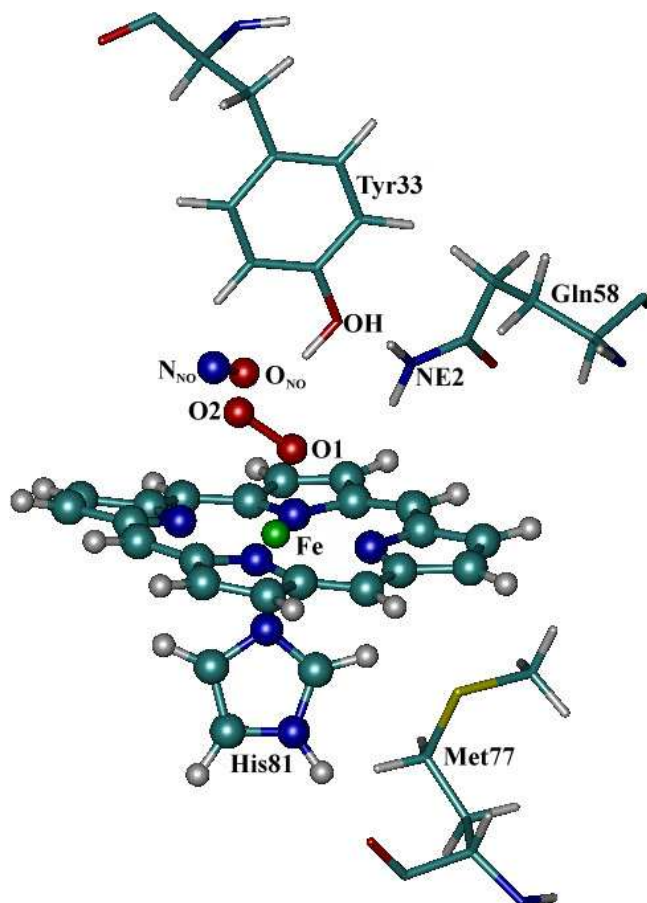
Figure 4.11. Relevant numbering representation of the trHbN active site (QM atoms are depicted using balls and sticks motif).

NO reaction with coordinated O_2

The interaction of NO with the O_2-bound heme group was examined by using selected structures obtained from CMIP calculations (see above and Figure 4.10) as starting configurations for the QM-MM calculations. After QM-MM optimization the NO molecule was best placed in the active site close to O_2, and interacting with Tyr33 and Gln58 residues (N_{NO}-O2: 2.99 Å, N_{NO}-Tyr33HH: 3.17 Å and O_{NO}-Gln58HE21: 2.90 Å). The oxidation reaction of NO was investigated considering a mechanism in which a peroxynitrite ion is formed from the attack of NO to the coordinated O_2. Subsequently, the metal center catalyzes the isomerization of peroxynitrite to nitrate. This latter process has been reported to proceed rapidly in both model systems[62,63] as well as in myoglobin and hemoglobin.[64] Calculations were performed for the isolated model system (vacuum), for the same model system solvated in aqueous solution (by using a cluster of 1061 water molecules), in the wt protein, and in the Tyr33 → Phe mutant.

According to the preceding mechanism, we first investigated the NO addition to the coordinated O_2 to yield peroxynitrite, equation (4.4). The proposed concerted mechanism was analyzed using restrained energy minimizations along the selected reaction coordinate, the distance between the nitrogen of NO and one oxygen atom ion $\xi_1 = d(O_2-N_{NO})$ (Figure 4.11).

$$Fe(II)\text{-}O_2 \ + NO \ \rightarrow \ Fe(III) \ [^-OONO] \tag{4.4}$$

The reaction proceeds barrierless in all cases considered in the QM-MM calculations (see Figure 4.12). The computed ΔE values were found to be -14.6, -20.0, -23.0 and -21.5 kcal/mol for vacuum, water, wt protein and mutant, respectively. The reaction involves the formation of a bond between O1 and the N atom of NO, and the weakening of both O1-O2 and N-O bonds, as can be seen from the structural parameters shown in Table 4.2. There are no significant differences between the geometrical features of reactants and products in going from vacuum to protein. It is interesting to note that the O1-O2 bond distance in the resulting coordinated peroxynitrite is longer than the one corresponding to the isolated species (1.38 Å). This is consistent with the fact that coordination to heme weakens this bond, and therefore favors isomerization to nitrate anion. However, it is clear that the protein environment around the heme group does not play a significant role in catalyzing the formation of nitrate.
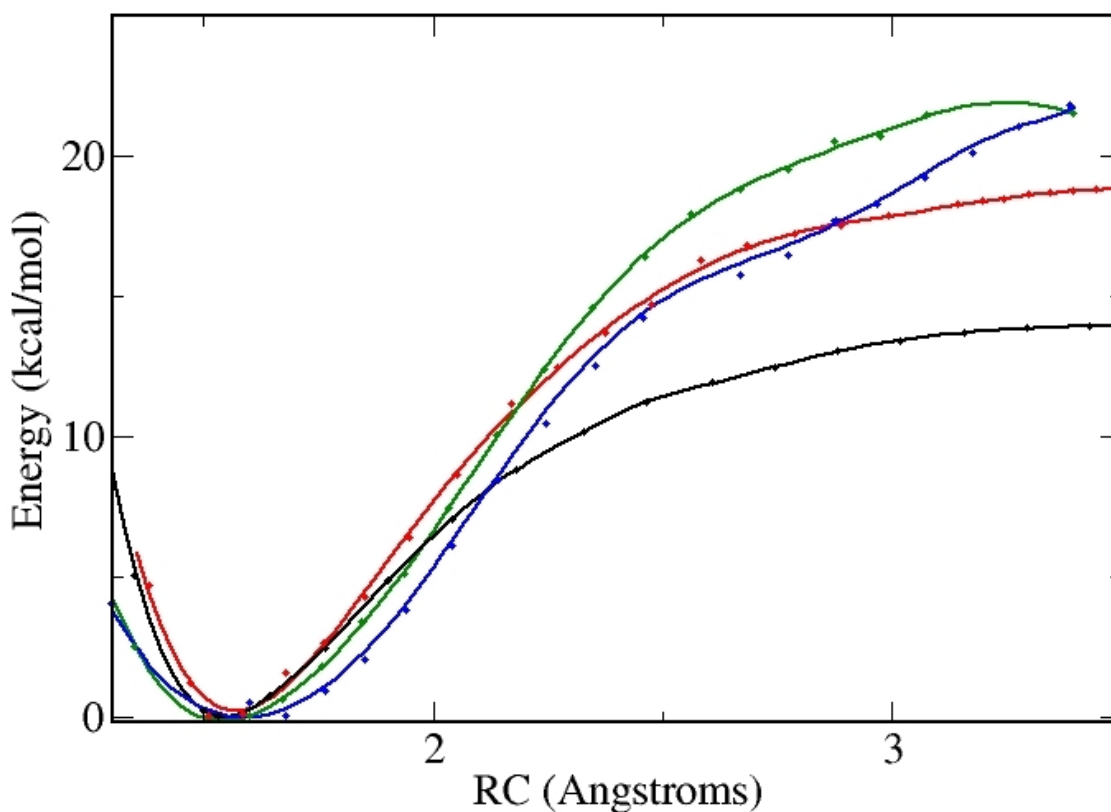

Figure 4.12. Energy profile of reaction described by equation (4.4). (black line) vacuum, (red line) water, (blue line) wt protein, and (green line) Tyr33 → Phe mutant.

	Vacuum		Water		Protein	
	Reactive	Product	Reactive	Product	Reactive	Product
Fe-O$_1$	1.83	1.78	1.82	1.80	1.83 (1.84)	1.87 (1.87)
O$_1$-O$_2$	1.28	1.45	1.28	1.48	1.32 (1.30)	1.41 (1.40)
O$_2$-N$_{NO}$	5.00	1.53	4.45	1.51	3.40 (3.35)	1.60 (1.60)
N$_{NO}$-O$_{NO}$	1.18	1.20	1.18	1.19	1.17 (1.18)	1.18 (1.19)
ΔE	-14.6		-20.0		- 23.0 (-21.5)	

Table 4.2. Relevant geometrical (distances in Å) and energetical (ΔE: energy difference between products and reactants; in kcal/mol) parameters for the NO addition to the coordinated O$_2$ to yield peroxynitrite, equation (4.4). For the reactions in protein, values in parenthesis correspond to the Tyr33 → Phe mutant.

The spin populations and Mulliken charges for selected groups for the reaction in the protein environment are given in Table 4.3. As expected, in the reactants the unpaired electron is mainly located on the NO moiety. The electronic structure of the Fe-O$_2$ complex shows the anti-ferromagnetic coupling of two regions of opposite spins, which is in agreement with previous calculations.[46] The peroxynitrite coordinated product exhibits the unpaired electron almost totally on the Fe atom, and a significant amount of charge donation (0.57 *e*) from peroxynitrite to the metal. In both reactants and products, the quartet state is significantly higher in energy than the doublet state.

	Reactive		Approximate Transition Structure		Product	
	spin pop.	q	spin pop.	q	spin pop.	q
NO	0.92	0.07	-	-	-	-
O$_2$	-0.75	-0.39	-	-	-	-
OONO	-	-	-	-	0.10	-0.43
Fe	0.88	0.76	-	-	0.95	0.80
Imidazole	0.00	0.20	-	-	0.00	0.26
Porfirine	-0.05	-0.64	-	-	-0.05	-0.63
ΔE$_{HS-LS}$	13.2		-		23.5	

Table 4.3. Spin energy gaps (kcal/mol), and spin and Mulliken populations *(e)* of selected groups for relevant species of equation (4.4) in protein.

The isomerization process leading to nitrate has been proposed to proceed in two steps according to the reactions indicated in equation (4.5) and equation (4.6). The reaction coordinates

chosen to examine the two processes were $\xi_2 = d(O_1-O_2)$ and $\xi_3 = d(O_1-N_{NO})$, respectively (Figure 4.11).

$$Fe(III) \, [^-OONO] \; \rightarrow \; Fe(IV)=O \; + NO_2 \tag{4.5}$$

$$Fe(IV)=O + NO_2 \; \rightarrow \; Fe(III)[NO_3^-] \tag{4.6}$$
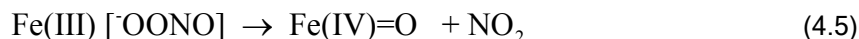
Although the reaction indicated in equation (4.5) involves the breaking of the O_1-O_2 bond, it is still exergonic (ΔE values of -15.7, -18.0, -8.1 and -7.7 kcal/mol for vacuum, water, wt protein and mutant, respectively) and almost barrierless (energy barrier = 1.2 kcal/mol; see Figure 4.13). This can be understood by considering that breaking of this bond yields a very stable Fe(IV)=O species, as can be observed by the short Fe-O1 bond distance, which amounts to 1.65 and 1.67 Å in vacuum and in the protein, respectively (Table 4.4). It can also be noted that the O2-N_{NO} bond becomes significantly shorter in the product compared to the reactant, 1.24 and 1.53 Å, respectively, consistently with an important stabilization of the product. The analysis of spin and Mulliken populations (Table 4.5) shows that the peroxynitrite complex breaks homolytically. The doublet species is the ground state. However, the spin energy gap with the quartet state is small. Similar results for spin populations of the oxyferryl compound have been reported for thiolate porphyrin complexes.[65,66]
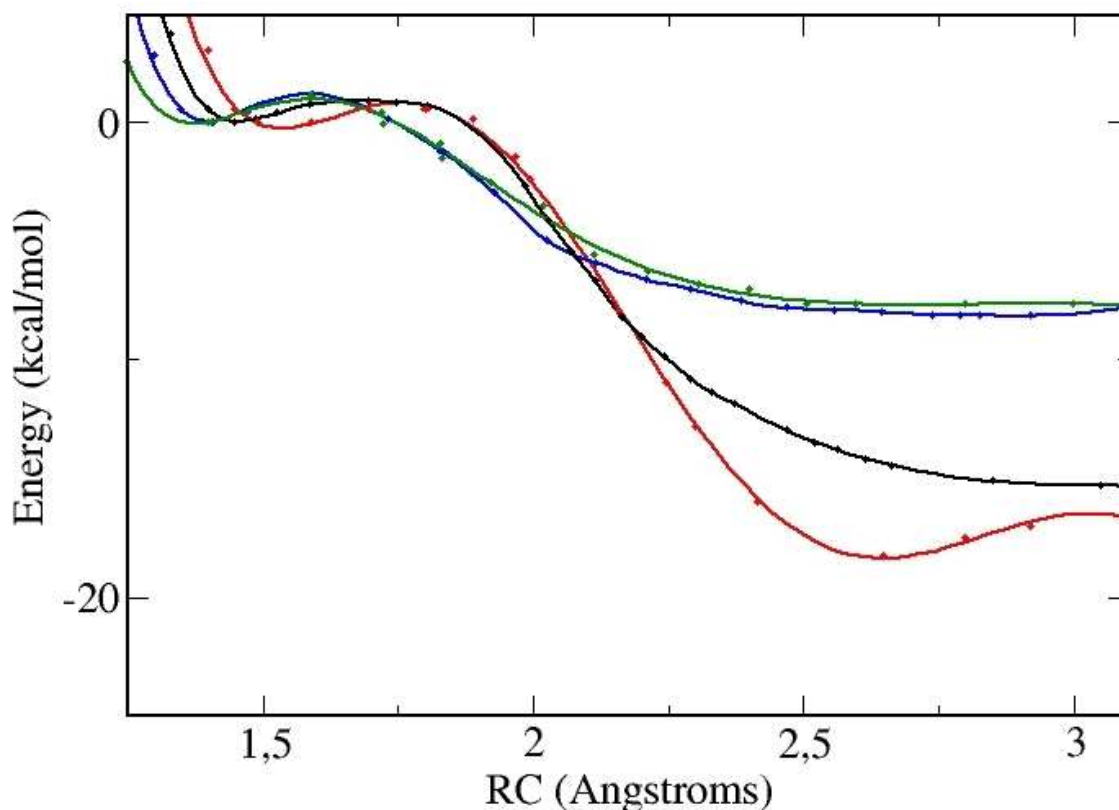

Figure 4.13. Energy profile of reaction described by equation (4.5). (black line) vacuum, (red line) water, (blue line) wt protein, and (green line) Tyr33 → Phe mutant.

	Vacuum		Water		Protein	
	Reactive	Product	Reactive	Product	Reactive	Product
$Fe-O_1$	1.78	1.65	1.80	1.67	1.87 (1.87)	1.67 (1.67)
O_1-O_2	1.45	3.40	1.48	2.65	1.41 (1.40)	2.79 (2.72)
O_2-N_{NO}	1.53	1.24	1.51	1.23	1.60 (1.60)	1.24 (1.24)
$N_{NO}-O_{NO}$	1.20	1.24	1.19	1.23	1.18 (1.19)	1.24 (1.24)
ΔE	-15.7		-18.0		-8.1 (-7.7)	

Table 4.4. Relevant geometrical (distances in Å) and energetical (ΔE: energy difference between products and reactants; in kcal/mol) parameters for reaction described by equation (4.5). For the reactions in protein, values in parenthesis correspond to the Tyr33 → Phe mutant.

	Reactive		Approximate Transition Structure		Product	
	spin pop.	q	spin pop.	q	spin pop.	q
OONO	0.10	-0.43	0.13	-0.47	-	-
NO_2	-	-	-0.03	-0.15	-0.82	-0.11
O1	-	-	0.16	-0.32	0.70	-0.38
Fe	0.95	0.80	0.93	0.81	1.21	0.80
Imidazole	0.00	0.26	0.00	0.23	0.00	0.18
Porfirine	-0.05	-0.63	-0.06	-0.57	-0.09	-0.49
ΔE_{HS-LS}	23.5		29.8		3.3	

Table 4.5. Spin energy gaps (kcal/mol), and spin and Mulliken populations *(e)* of selected groups for relevant species of equation (4.5) in protein.

The reaction described by equation (4.6) can be analyzed as the addition of a negatively charged O atom coordinated to the heme to the NO_2 radical. The driving force of this reaction is probably related to the formation of a strong O-N bond, as can be seen from Table 4.6. Inspection of the energy profiles (see Figure 4.14) reveals that in all cases the reaction is exergonic (computed ΔE values of -12.2, -11.7, -18.0 and -19.3 kcal/mol for vacuum, water, wt protein and mutant, respectively) and proceeds with very small barriers (less than 2.6 kcal/mol). As expected, the product consists on a nitrate ion coordinated to the Fe. The unpaired electron is almost completely located on the metal (Table 4.7). Experimental results obtained for heme model systems in water solution show that the yield of this reaction is very dependent of the reaction conditions. However, this is due to the fact that the NO_2 radical may escape and react with other

species in solution.[63] On the other hand, in the protein the situation is completely different, since NO_2 is trapped in the pocket formed by Phe32, Tyr33, Val36, Phe46, and Val94.
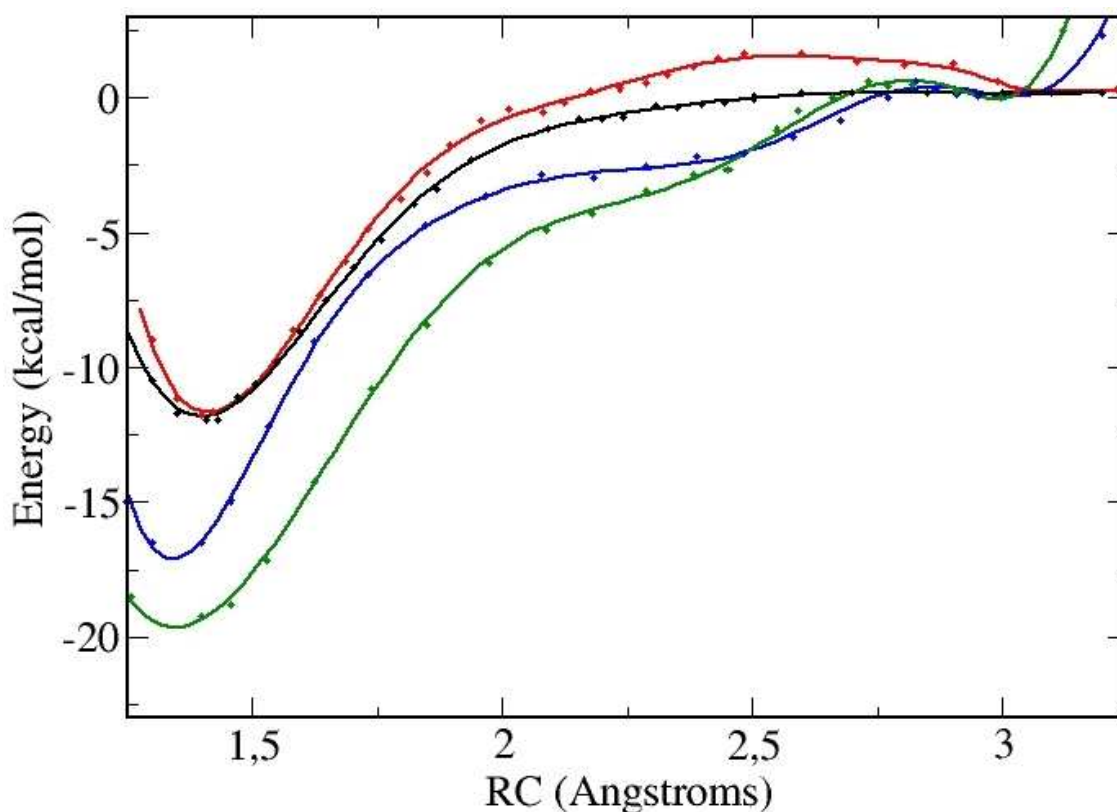

Figure 4.14. Energy profile of reaction described by equation (4.6). (black line) vacuum, (red line) water, (blue line) wt protein, and (green line) Tyr33 → Phe mutant.

	Vacuum		Water		Protein	
	Reactive	Product	Reactive	Product	Reactive	Product
Fe-O₁	1.65	1.84	1.67	1.93	1.67 (1.67)	1.94 (1.92)
O₁-N_{NO}	2.52	1.40	3.25	1.46	3.02 (2.99)	1.40 (1.40)
O₂-N_{NO}	1.24	1.25	1.23	1.25	1.24 (1.24)	1.25 (1.25)
N_{NO}-O_{NO}	1.24	1.25	1.23	1.23	1.24 (1.24)	1.25 (1.25)
ΔE	-12.2		-11.7		-18.0 (-19.3)	

Table 4.6. Relevant geometrical (distances in Å) and energetical (ΔE: energy difference between products and reactants; in kcal/mol) parameters for reaction described by equation (4.6). For the reactions in protein, values in parenthesis correspond to the Tyr33 → Phe mutant.

	Reactive		Approximate Transition Structure		Product	
	spin pop.	q	spin pop.	q	spin pop.	q
NO$_2$	-0.82	-0.11	-0.76	-0.11	-	-
O1	0.70	-0.38	0.66	-0.37	-	-
NO$_3$	-	-	-	-	0.10	-0.54
Fe	1.21	0.80	1.18	0.80	0.94	0.81
Imidazole	0.00	0.18	0.00	0.19	0.00	0.25
Porfirine	-0.09	-0.49	-0.08	-0.50	-0.04	-0.52
ΔE_{HS-LS}	3.3		4.5		16.5	

Table 4.7. Spin energy gaps (kcal/mol), and spin and Mulliken populations *(e)* of selected groups for relevant species of equation (4.6) in protein.

On the basis of the preceding results, it can be concluded that the NO oxidation is a favorable process for the different environments investigated here. Experimental results show that bimolecular rate constants for NO-induced oxidation are similar to those of NO binding to the deoxygenated proteins, showing that the rate limiting process is ligand diffusion and not the chemical reaction itself.[7] The experimental reaction rate (1 10^5 M^{-1}s^{-1}) of peroxynitrite to NO$_2$ conversion catalyzed by a model porphyrin (FeTMPS) in aqueous solution corresponds to peroxynitrite coordination and subsequent O-O breaking to form the oxyferryl species and NO$_2$.[63] The peroxynitrite coordinated intermediate has not been detected, suggesting that peroxynitrite coordination is rate limiting. Therefore, the bond breaking process should be much faster, indicating a shallow barrier. Our computed results for reaction (4.5), which corresponds to O-O bond breaking of the peroxynitrite coordinated complex show no energetic barrier and are consistent with the experimental results. The experimental reaction rate for the reaction of an oxyferryl complex with NO$_2$ has not been measured.[63] However, it has been claimed that it should be faster than NO$_2$ dimerization, which is a very fast process (k=5 10^7 M^{-1}s^{-1}). Consistently with this, our results show a small barrier for this process. Finally, the results obtained for the Tyr33 $\rightarrow$ Phe mutant are qualitatively similar to those obtained for the wt protein, thus suggesting that the protein environment, and particularly Tyr33, does not play a significant role in assisting the reaction.

Implications into the functional role of trHbN

Similarly to the dioxygenase activity of (flavo) hemoglobins and myoglobin,[67-69] the trHbN enzyme has been suggested to play an essential role in the O$_2$-sustained NO detoxification.[11] Clearly,

understanding of the molecular factors that modulate the efficiency of trHbN is crucial to design therapeutic strategies against diseases such as human tuberculosis.

One of the factors that have been hypothesized to play a decisive contribution in trHbN is Tyr33, which was argued to modulate not only the O_2 affinity to the heme group, but also to develop a catalytic role in the enzyme reaction. This latter role, however, is not supported by the preceding results, as the reaction profile determined for the wt protein is very similar to that obtained for the Tyr33 $\rightarrow$ Phe mutant. Our results point out that Tyr33 is mainly involved in enhancing the oxygen affinity of the protein, which is an important ingredient for determining the NO detoxification ability of the protein, as noted in the enhanced interaction energy of the wt enzyme compared to the Tyr33 $\rightarrow$ Phe form. However, besides the energetic contribution due to hydrogen bonding to Tyr33, other factors modulate the O_2 affinity of trHbN. In particular, our results suggest that the donation capacity of His81, enhanced through hydrogen-bonding to Met77, might play a relevant contribution.

The analysis of the MD trajectory of oxy-trHbN, as well as the results derived from QM-MM calculations, point out that Tyr33 forms a hydrogen-bond with the nitrogen atom of Gln58. The results also indicate that this latter residue does not interact with the O_2 molecule, in agreement with X-ray crystallographic data[15] and resonance Raman spectra.[17] In fact, it must be noted that the occurrence of the Tyr33-Gln58 pair does not necessarily lead to strong stabilization of the heme-bound O_2 in different Hbs.[15,70-72] Nevertheless, present results give support to the suggestion[14,15] that such a hydrogen-bonded interaction is crucial to fix and orient properly the NO molecule with regard to the heme-bound O_2. In this sense, the hydrogen-bonded network formed by the Tyr33-Gln58 pair might be relevant for the enzyme reaction because of its role in mediating ligand binding stabilization and in assisting the mutual orientation of the reactants in the heme active site, as will be discussed later.

A more subtle effect might also be envisaged for Tyr33 from the analysis of the MD simulation. Formation of the hydrogen bond between Tyr33 and Gln58, an interaction that was found to be maintained along the whole trajectory, permits to anchor the B helix to the E helix (Tyr33 and Gln58 occupy the B10 and E11 positions in those helices). The results derived from the analysis of the MD trajectory point out the occurrence of conformational transitions leading to the opening of the channel, and that those transitions stem from the interplay between collective "breathing" motions of the protein backbone and local changes in dihedral angles of Phe62, which would eventually facilitate the transition from the closed conformation to the open one. It might, therefore, be speculated that the Tyr33-Gln58 pair-assisted O_2 binding to the heme group modulate the natural dynamics of the protein and regulate in this way the passage of NO through the ligand diffusion tunnel. This hypothesis would be in agreement with experimental data derived

from comparison of the resonance Raman spectra in the wt trHbN and the Tyr33 $\rightarrow$ Phe mutant, which indicate that i) the Tyr33$\rightarrow$ Phe mutation does not alter the conformational flexibility of the free enzyme, and that ii) binding to the distal site of the heme group triggers a large-scale conformational change in the enzyme.[19] This hypothetical model would, then, provide a mechanistic model valuable to link the intrinsic dynamical features of the protein with the ligand binding properties, and ultimately with the physiological role played by the trHbN enzyme. In order to shed light on this issue, we have performed calculations of free energy profiles for both O_2 and NO migration inside the enzyme.

Free energy profiles and protein dynamics

The oxygenated protein free energy profiles for ligand migration, obtained through steered MD simulations indicated that in the open state NO diffusion is almost free of barriers, and access to the heme cavity is favored by 3-4 kcal/mol (Figure 4.15). On the contrary, in the closed state the steric hindrance of Phe62 phenyl ring undermines ligand access, leading to a steady increase in the free energy profile (Figure 4.15). Interestingly, the free energy minimum found at about 11 Å from the heme matches one of the Xe binding sites observed experimentally in *M. tuberculosis* trHbN.[16] Steered MD simulations also revealed that migration of NO through the tunnel short branch is not favored, related to a progressive rise in the free energy (Figure 4.16). Taken together, the above results suggest that NO access to the heme distal cavity should preferentially occur through the long tunnel branch in oxy-trHbN.
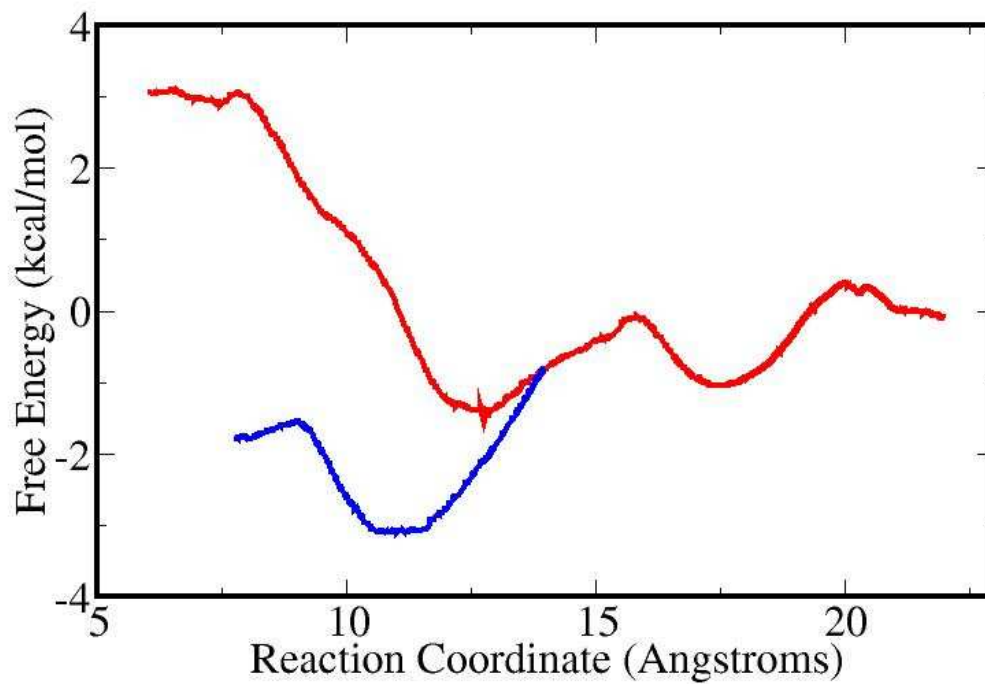
Figure 4.15. Free energy profile for NO migration along the long tunnel branch in both *open* (blue) and *closed* (red) states of oxy-trHbN; the distance from the Fe atom to the nitrogen in NO is used as the driven coordinate.
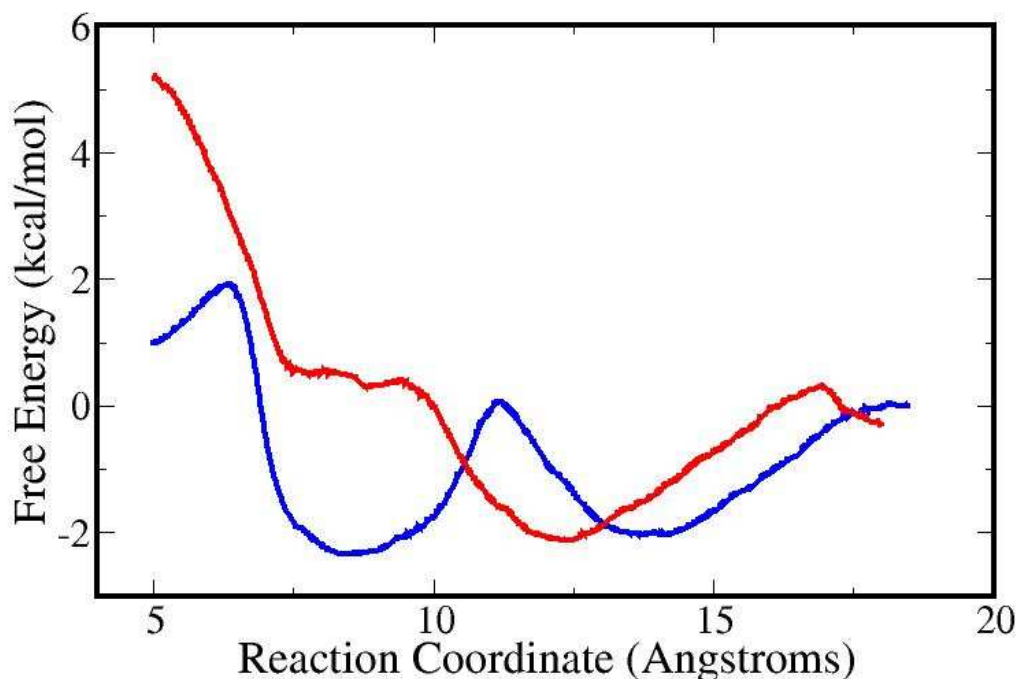
Figure 4.16. Free energy profile for diatomic ligand (NO; O_2) migration along the short tunnel branch for oxy-trHbN (NO, red) and deoxy-trHbN (O_2, blue); the distance from the Fe atom to either nitrogen (NO) or oxygen (O_2) atoms is used as the driven coordinate.

Strikingly, the ligand migration pathway described above changed drastically in deoxy-trHbN. The analysis of the MD trajectory for deoxy-trHbN showed that the long branch of the tunnel remains closed during the whole simulation time. In fact, the Phe62 torsional free energy profile, obtained using umbrella sampling calculations, reveals that in deoxy-trHbN the conformational transition between open and closed states of the tunnel gating residue Phe62 must overcome a free energy barrier higher than 6 kcal/mol (compared with the 3 kcal/mol barrier of oxy-trHbN; see Figure 4.17). Moreover, the open state (less stable by ≈2 kcal/mol relative to the closed state in oxy-trHbN) is further destabilized by 1.5 kcal/mol in deoxy-trHbN. In contrast, ligand migration through the short branch of the tunnel is now allowed with a small free energy barrier (≈2 kcal/mol; Figure 4.16) before reaching a flat minimum, which corresponds to a cavity formed by residues PheG5, AlaG9, and IleH11. In fact, this cavity is one of the Xe binding pockets observed by X-ray crystallography in the short tunnel branch.[16] Access to the Fe atom from this cavity requires surpassing a free energy barrier of ≈4 kcal/mol (Figure 4.16), reflecting a narrow passage through the heme and ValG8.
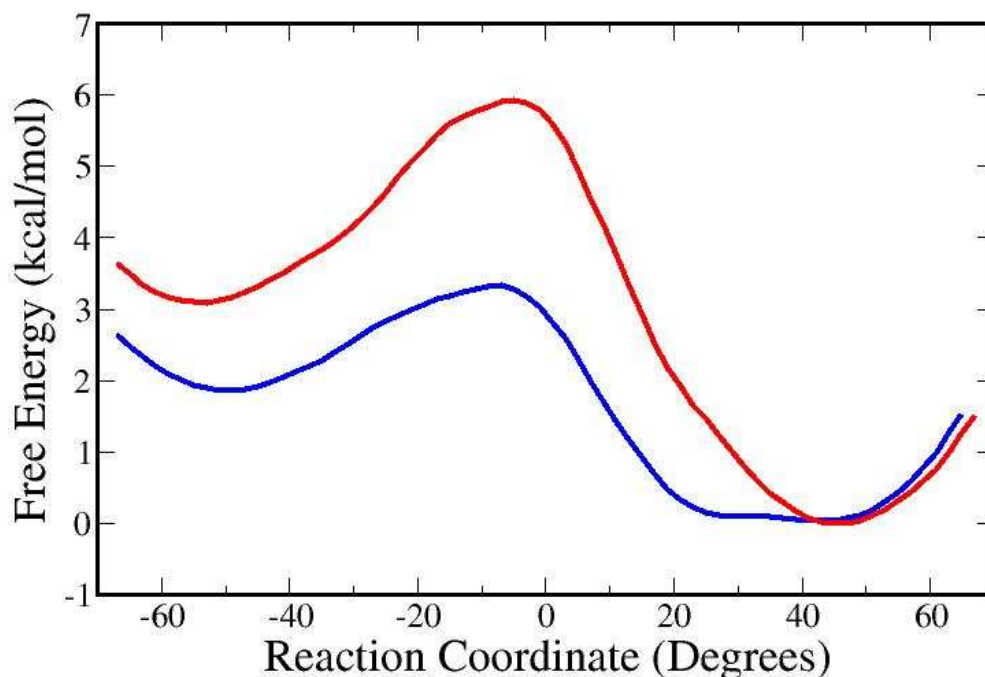
Figure 4.17. Phe62 torsional free energy profile for the deoxy-trHbN (red) and oxy-trHbN (blue) forms. The reaction coordinate has been chosen as the C_α-C_β dihedral angle (40 and -50 degrees for the *closed* and *open* configurations, respectively).

The preceding results suggest that *M. tuberculosis* trHbN has evolved a molecular mechanism for selective ligand diffusion, where O_2 accesses the heme through the tunnel short branch and, upon binding to the Fe atom, facilitates migration of NO through the tunnel long branch. Such first set of conclusions raises the question of how does O_2 binding to the heme group regulate the ligand access pathway. Our simulation results indicate that such regulation may be achieved through alteration of the heme distal site hydrogen-bonded network, a structural feature previously recognized to play a key role in trHbN heme/ligand stabilization.[15] In the deoxy-trHbN simulation, the side chain of Gln58 adopts an extended all-trans conformation, allowing the terminal amino group to be hydrogen-bonded to the hydroxyl group of Tyr33. In addition, the simulation shows a dynamical fluctuation between (Tyr33)O-H⋯O=C(Gln58) and (Tyr33)O⋯H-N(Gln58) hydrogen bonds, such that Tyr33 hydroxyl group acts both as hydrogen-bond donor and acceptor during the simulation (Figure 4.18a). In contrast, in oxy-trHbN the heme-Fe coordinated O_2 fixes Tyr33 through formation of a hydrogen-bond, thus allowing the phenolic hydroxyl group to act exclusively as hydrogen-bond acceptor for the amide NH_2 group of Gln58. Under these conditions Gln58 side chain is forced to fluctuate between two staggered conformations (Figure

117

4.18b). Under the latter conformational states the distance between the side chains of Gln58 and Phe62 is shortened by ~1.5 Å due to different residue packing. Thus, Gln58 might be viewed as an internal "switch" that would regulate the open/closed state of the Phe62 gate, within the tunnel long branch through a conformational compression mechanism.
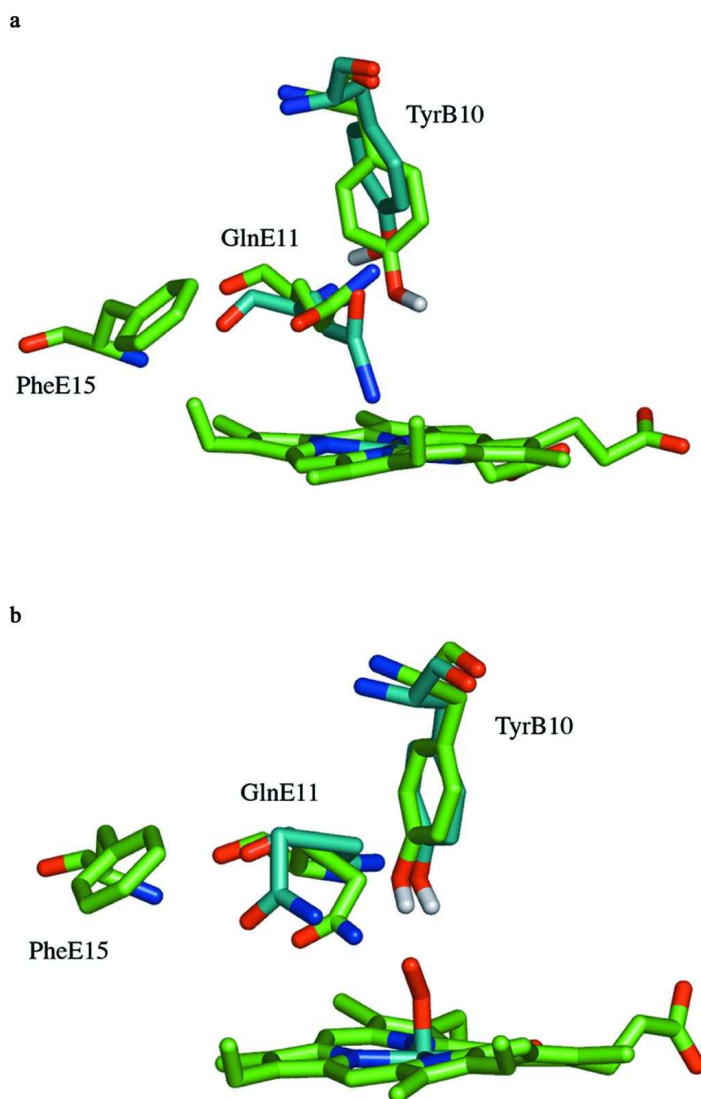

Figure 4.18. Interactions between Tyr33 and Gln58 in deoxy-trHbN and oxy-trHbN. (a) In deoxy-trHbN the side chain of Gln58 adopts an *all-trans* conformation (average torsions around C_α-C_β, C_β-C_γ, and C_γ-C_δ dihedrals of 40.2, 167.1 and 93.7 degrees), though the relative orientation of the amide group of Gln55 and hydroxyl group of Tyr33 fluctuate along the trajectory. (b) In oxy-trHbN the side chain of Gln58 populates two *staggered* conformation characterized by torsional angles around C_α-C_β, C_β-C_γ, and C_γ-C_δ bonds of (+45.0, -74.8, -99.3 degrees) and (-50.2, +67.2, +48.5 degrees), respectively. Transition of the former conformation to the latter one makes Gln58 to approach Phe32 by ~1.5 Å.

Conclusions

The efficiency of the NO oxidation to nitrate by oxygenated trHbN depends on the entry of the reactants NO and O_2 to the distal cavity and on the heme capacity to catalyze the reaction. The results allowed us to identify the opening of the ligand diffusion tunnel, a process controlled by Phe62, which can be considered to be the gate of the channel leading to the heme group in the active site. Moreover, the opening of the tunnel stems from a complex network of conformational changes, which involve both local changes in side chains of specific residues, particularly Phe62, and collective motions of the protein backbone. The results derived from QM-MM calculations indicate that the protein catalyzes the chemical reactions leading to the formation of nitrate mainly by means of the heme group, with no significant contributions of the protein environment. However, our results suggest that the residues that form the binding site, especially the Tyr33-Gln58 pair, might be relevant to the physiological function played by trHbN at different levels. First, they contribute to modulate the affinity of O_2 and NO ligands. Second, by isolating the reaction intermediates formed transiently in the active site, they prevent other non desired reactions observed in model porphyrins in aqueous solution to proceed. Third, the stabilization of the Tyr33-Gln58 pair promoted upon O_2 binding alter the essential dynamics of the protein, leading in turn to a mechanism for ligand-induced regulation, and to the modulation of the NO detoxification process operative in *M. tuberculosis*, where the protein dynamics and the protein matrix tunnel system have evolved to allow the selective access of O_2 and NO ligands to the heme through distinct migration paths. The key feature in such mechanism is the change of contacts formed by Tyr33 and Gln58, where Gln58 acts as a switch regulating the ligand access pathway by altering the essential movements of the protein backbone. Thus, O_2 binding to the heme should induce compression of Phe62 by the side chain of Gln58, and promote the relative displacement of helices B and E, which combined with local rearrangements in the side chains that surround Phe62, should open the tunnel long branch, facilitating capture of NO from the aqueous phase. This mechanism would contribute to enhance the efficiency of NO detoxification, while minimizing the fraction of NO that directly binds to the deoxy-heme. It also provides a basis to rationalize why NO conversion to nitrate by oxy-trHbN occurs at a rate faster than O_2 binding to the deoxy protein,[15] an explanation that cannot be extracted from the X-ray structures and from the kinetic data alone.

As a final consideration, we note that other heme proteins have been proposed to perform multiligand chemistry. Among these, oxy-hemoglobin, oxy-myoglobin, and the recently discovered neuroglobin and cytoglobin.[8,7,73] Untangling the structural determinants of the interplay between ligand binding, protein dynamics and ligand migration will be essential in order to shed light on the emerging field of multiligand chemistry performed by heme proteins.

---

[1] Bloom, B. R. (ed.) Tuberculosis: pathogenesis, protection and control. ASM Press, Washington D.C. 1994.

[2] MacMicking, J.D.; North, R.J.; La Course, R.; Mudgett, J.S.; Shah, S. K.; Nathan, C.F. Proc. Natl. Acad. Sci. USA, 1997, 94, 5243.

[3] Clark, I.A.; Rocket, K. A. Adv. Parasitol. 1996, 37, 1.

[4] Colasanti, M.; Gradoni, L.; Mattu, M.; Persichini, T.; Salvati, L.; Venturini, G.; Ascenzi, P. Int. J. Mol. Med. 2002, 9, 131.

[5] Gardner, P. R.; Gardner, A. M.; Martin, A. M.; Dou, Y.; Li, T.; Olson, J. S.; Zhu, H.; Riggs, A. F. J. Biol. Chem. 2000, 275, 12581.

[6] Liu, L.; Zeng, M.; Hausladen, A.; Heitman, J.; Stamler, J. S. Proc. Natl. Acad. Sci. USA, 2000, 97, 4672.

[7] Eich, R. F.; Li, T.; Lemon, D. D.; Doherty, D. H.; Curry, S. R.; Aitken, J. F.; Mathews, A. J.; Johnson, K. A.; Smith, R. D.; Phillips, G. N.; Olson, J. S. Biochemistry, 1996, 35, 6976.

[8] Ascenzi, P.; Slavati, L.; Brunori, M. FEBS Lett. 2001, 501, 103.

[9] Pesce, A.; Bolognesi, M.; Bocedi, A.; Ascenzi, P.; Dewilde, S.; Moens, L.; Hankeln, T.; Burmester, T. EMBO Reports 2002, 3, 1146.

[10] Pesce, A.; Dewilde, S.; Nardini, M.; Moens, L.; Ascenzi, P.; Hankeln, T.; Burmester, T.; Bolognesi, M. Structure 2003, 11, 1087.

[11] Couture, M.; Yeh, S.; Wittenberg, B. A.; Wittenberg, J. B.; Ouellet, Y.; Rousseau, D. L.; Guertin, M. Proc. Natl. Acad. Sci. USA, 1999, 96, 11223.

[12] Moens, L.; Vanfleteren, J.; Van der Peer, Y.; Peeters, K.; Kapp, O.; Czeluzniak, J.; Goodman, M.; Blaxter, M.; Vinogradov, S. Mol. Biol. Evol. 1996, 13, 324.

[13] Couture, M.; Chamberland, H.; St-Pierre, B.; Lafontaine, J.; Guertin, M. Mol. Gen. Genet. 1994, 243, 185-197.

[14] Pesce, A.; Couture, M.; Dewilde, S.; Guertin, M.; Yamauchi, K.; Ascenzi, P.; Moens, L.; Bolognesi, M. EMBO J. 2000, 19, 2424.

[15] Milani, M.; Pesce, A.; Ouellet, Y.; Ascenzi, P.; Guertin, M.; Bolognesi, M. EMBO J. 2001, 20, 3902.

[16] Milani, M.; Pesce, A.; Ouellet, Y.; Dewilde, S.; Friedman, J.; Ascenzi, P.; Guertin, M.; Bolognesi, M. J. Biol. Chem. 2004, 279, 21520.

[17] Yeh, S.R.; Couture, M.; Ouellet, Y.; Guertin, M.; Rousseau, D. L. J. Biol. Chem. 2000, 275, 1679.

[18] Ouellet, H.; Ouellet, Y.; Richard, C.; Labarre, M.; Wittemberg, B.; Wittemberg, J.; Guertin, M. Proc. Natl. Acad. Sci. USA, 2002, 99, 5902.

[19] Mukai, M. ; Ouellet, Y. ; Ouellet, H. ; Guertin, M. ; Yeh, S. Biochemistry, 2004, 43, 2764.

[20] Jorgensen, W. L.; Chandrasekar, J.; Madura, J.; Impey, R. W.; Klein, M. L. J. Chem. Phys. 1983, 79, 926.

[21] Luty, B. A.; Tironi, I. G.; van Gunsteren, W. F. J. Chem. Phys. 1995, 103, 3014.

[22] Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. J. Comput. Phys. 1977, 23, 327.

[23] Wang, J.; Cieplak, P.; Kollman, P. A. J. Comput. Chem. 2000, 21, 1049.

[24] Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. J. Phys. Chem. 1993, 97, 10269.

[25] Banci, L.; Curr. Op. Chem. Biol. 2003, 7, 143.

[26] Laskowski, R. A. J. Mol. Graph. 1995, 13, 323.

[27] Gelpí, J. L.; Kalko, S.; de la Cruz, X.; Barril, X.; Cirera, J.; Luque, F. J.; Orozco, M. Proteins, 2001, 45, 428.

[28] García, A. E. Phys. Rev. Lett. 1992, 68, 2696.

[29] Rueda, M.; Kalko, S. G.; Luque, F. J.; Orozco, M. J. Am. Chem. Soc. 2003, 125, 8007.

[30] Fradera, X.; Márquez, M.; Smith, B.; Orozco, M.; Luque, F. J. J. Org. Chem. 2003, 68, 4663.

[31] Wlodek, S. T.; Clark, T. W.; Scott, L.R.; McCammon, J. A. J. Am. Chem. Soc. 1997, 119, 9513.

[32] Warshel, A.; Levitt, M. J. Mol. Biol. 1976, 103, 227.

[33] Zhang, X.; Harrison, D. H. T.; Cui, Q.; J. Am. Chem. Soc. 2002, 124, 14871.

[34] Ryde, U. Curr. Op. Chem. Biol.; 2003, 7, 136.

[35] Ridder, L.; Harvey, J. N.; Rietjens, I. M. C. M.; Vervoort, J.; Mulholland, A. J.; J. Phys. Chem. B, 2003, 107, 2118.

[36] Schöneboom, J. C.; Cohen, S.; Lin, H.; Shaik, S.; Thiel, W. J. Am. Chem. Soc. 2004, 126, 4017.

[37] Devi-Kesavan, L. S.; Gao, J. J. Am. Chem. Soc. 2003, 125, 1532.

[38] Soler, J. M.; Artacho, E.; Gale, J.; García A.; Junquera, J. Ordejón, P.; Sánchez-Portal D. J. Phys: Cond. Matt. 2002, 14, 2745.

[39] Martí, M. A.; Scherlis, D. A.; Doctorovich, F. A.; Ordejón, P.; Estrin, D. A. J. Biol. Inorg. Chem. 2003, 6, 595.

[40] Troullier, N.; Martins, J. L. Phys. Rev. B 1991, 43, 1993.

[41] Louie, S. G.; Froyen, S.; Cohen, M. L. Phys. Rev. B 1982, 26, 1738.

[42] Perdew, J. P.; Burke, K.; Ernzerhof, M. Phys. Rev. Lett. 1996, 77, 3865.

[43] Eichinger, M.; Tavan, P.; Hutter, J.; Parrnello, M. J. Chem. Phys. 1999, 110, 10452.

[44] Rovira, C.; Schultze, B.; Eichinger, M.; Evanseck, J. D.; Parrinello, M. Biophys J. 2001, 81, 435.

[45] Scherlis, D. A.; Martí, M. A.; Ordejón, P.; Estrin, D. A. Int. J. Quant. Chem. 2002, 90, 1505.

[46] Rovira, C.; Kunc, K.; Hutter, J.; Ballone, P.; Parrinello, M. J. Phys. Chem. A, 1997, 101, 8914.

[47] Crespo, A.; Scherlis, D. A.; Martí, M. A.; Ordejón, P.; Roitberg, A. E.; Estrin, D. A. J. Phys. Chem. B 2003, 107, 13728.

[48] Jarzynsky, C. Phys. Rev. Lett. 1997, 78, 2690.

[49] Crespo, A.; Martí, M.; Estrin, D.; Roitberg, A. J. Am. Chem. Soc. Comm. 2005, 127, 6940.

[50] Leach, A. R. Molecular Modelling, 2nd ed.; Prentice Hall, New Jersey, 2001.

[51] Kachalova, G. S.; Popov, A. N.; Bartunik, H. D. Science, 1999, 284, 473.

[52] Englander, S. W.; Kallenbach, N. R. Q. Rev. Biophys. 1983, 16, 521.

[53] Wuthrich, K. Q. Rev. Biophys. 1987, 19, 3.

[54] Amadei, A.; Linssen, A. B.; Berendsen, H. J. Proteins 1993, 17, 412.

[55] McCammon, J. A.; Harvey, S. C. Dynamics of Proteins and Nucleic Acids; Cambridge University Press: New York, 1987.

[56] Gogonea, V.; Osawa, E. J. Comput. Chem. 1995, 16, 817.

[57] Shen, T.; Tai, K.; Henchman, R. H.; McCammon, J. A. Acc. Chem. Res. 2002, 35, 332.

[58] Scherlis, D. A.; Estrin, D. A. J. Am. Chem. Soc. 2001, 123, 8436.

[59] The experimental X-ray structure was solved at 1.9 Å resolution, which implies a mean error in atomic coordinates about 0.1 Å. However, differences between 0.1-0.3 Å for the selected geometrical parameters shown in Table 4.1 are found in the two subunits of the X-ray crystallographic structure. Particularly, the larger value found for the Fe-O1 distance in 1IDR (2.06 and 2.19 Å) relative to other oxygenated heme proteins (myoglobin 1MBO: 1.83 Å; ascaris hemoglobin: 1.90 Å) might reflect, at least in part, differences in the restrains used in the crystallographic refinement (M. Bolognesi, personal communication).

[60] Kundu, S.; Snyder, B.; Das, K.; Chowdhury, P.; Park, J.; Petrich, J. W.; Hargrove, M. S. Struct. Funct. Gen. 2002, 46, 268.

[61] Jain, R.; Chan, M. K. J. Biol. Inorg. Chem. 2003, 8, 1.

[62] Jensen, M. P.; Riley, D.P. Inorg. Chem. 2002, 41, 4788.

[63] Shimanovich, R.; Groves, J. T. Arch. Biochem. Biophys. 2001, 387, 307.

[64] Herold, S.; Shivashankar, K. Biochemistry, 2003, 42, 14036.

[65] Guallar, V.; Friesner, R. A. J. Am. Chem. Soc.; 2004, 126, 8501.

[66] Schoneboom, J. C.; Lin, H.; Reuter, N.; Thiel, W.; Cohen, S.; Ogliaro, F.; Shaik, S. J. Am. Chem. Soc. 2002, 124, 8142.

[67] Poole, R. K.; Hughes, M. N. Mol. Microbiol. 2000, 36, 775.

[68] Flögel, U.; Merx, M. W.; Gödecke, A.; Decking, U. K.; Schrader, J. Proc. Natl. Acad. Sci. USA, 2001, 98, 735.

[69] Frauenfelder, H.; McMahan, B. H.; Austin, R. H.; Chu, K.; Groves, J. T. Proc. Natl. Acad. Sci. USA, 2001, 98, 2370.

[70] Couture, M.; Das, T. K.; Lee, H. C.; Peisach, J. ; Rousseau, D. L. ; Wittenberg, B. A.; Wittenberg, J. B.; Guertin, M. J. Biol. Chem. 1999, 274, 6898.

[71] Das, T. K.; Weber, R. E.; Dewilde, S.; Wittenberg, J. B.; Wittenberg, B. A.; Yamauchi, K.; van Hauwaert, M. L.; Moens, L.; Rousseau, D. L. Biochemistry, 2000, 39, 14330.

[72] Das, T. K.; Couture, M.; Ouellet, Y.; Guertin, M. ; Rousseau, D. L. Proc. Natl. Acad. Sci. USA 2001, 98, 479-484.

[73] Brunori, M. Proc. Natl. Acad. Sci. USA 2005, 102, 8483.

Chapter 5: Hydroxylation Reaction Catalyzed by Peptidylglycine *α*-Hydroxylating Monooxygenase

Introduction

Cu proteins are common in biology and play important roles in O_2 activation and reduction.[1-3] Many peptide hormones and neuropeptides require amidation of their C terminus for biological activity.[4-6] Peptidylglycine α-amidating monooxygenase (PAM) uses separate enzymatic domains to catalyze this amidation reaction in two steps: (i) hydroxylation of the CA of a C-terminal glycine and (ii) disproportionation of the α-hydroxyglycine.[7,8] The first step is catalyzed by peptidylglycine α-hydroxylating monooxygenase (PHM), a two-copper, ascorbate-dependent enzyme that is the more extensively studied of the two PAM domains (see Figure 5.1). Kinetic and mutagenesis studies,[9,10] detailed kinetic isotope effect measurements,[11-14] and extended x-ray absorption fine structure measurements,[10,15-17] as well as determination of the structures of several forms of the enzyme,[18,19] have been used in studying PHM. Despite these intense efforts, key aspects of the PHM mechanism remain elusive.[9,20,21] Perhaps the most puzzling aspect of the PHM reaction is the mechanism of electron transfer between the active-site coppers and the subsequent reduction of molecular oxygen. Crystallographic, spectroscopic, and kinetic studies have provided conflicting information about how electrons are transferred between the copper atoms, which copper atom is the binding site for molecular oxygen, and where the chemical steps of the PHM reaction take place. Depending on the coordination of the dioxygen to Cu_B, the copper-bound oxygen may also be a part of the electron transfer path, coupling electron transfer to oxygen reduction and hydrogen abstraction.
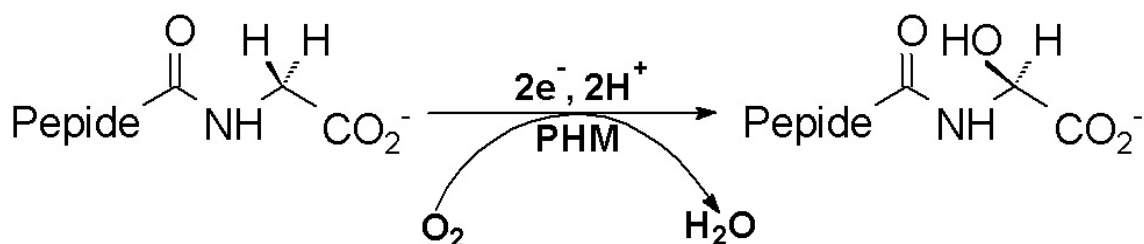

Figure 5.1. Stoichiometry of the global reaction catalyzed by PHM.

Another well characterized enzyme related to PHM is the dopamine β-monooxygenase (DβM), which catalyzes the dopamine benzylic C-H bond hydroxylation using molecular O_2 in a stereo-

and regiospecific fashion in the same way as PHM does.[3] The active sites of these two proteins consist of two inequivalent Cu centers largely separated in space (≈11 Å in PHM)[18] with no direct bridging ligands and no observable magnetic interactions.[22] The crystal structure of PHM indicates that the Cu_B site, where the substrate hydroxylation occurs, is coordinated by two histidine and one methionine ligands to the protein backbone, and the other Cu_A site, which provides an additional electron through long-range electron transfer (ET) to the Cu_B site, has three histidine ligands from the protein.[3,18,19] As mentioned before, since the two Cu sites in PHM and DβM are far away, the mechanism for this intramolecular long-range ET is not clear. A superoxide channeling mechanism[20] and a substrate-facilitated ET mechanism (either through the substrate[19] or through protein residues brought closer upon substrate binding[9]) have been proposed to account for this inter-Cu intramolecular ET process.

Previous kinetic and mechanistic studies have shown that the reaction mechanism for the substrate hydroxylation in PHM and DβM are very similar.[3] The enzymatic cycle starts with both the Cu_A and Cu_B sites at the Cu^I oxidation state (reduction by ascorbate in physiological conditions). When a substrate is present, O_2 reacts with the reduced protein forming a reactive Cu/O_2 intermediate, which then cleaves the substrate C-H bond via an H-atom abstraction mechanism generating a substrate radical.[23-26] Significant ^2H and ^{18}O isotope effects have been observed on the C-H bond cleavage reaction leading to the final hydroxylated product.[3,12-14,25,27] The reactive Cu/O_2 intermediate has been widely proposed to be an as-yet unobserved Cu^{II}-OOH species that would either abstract the substrate hydrogen directly or go through a Cu^{II}-oxyl intermediate first before abstracting the substrate H-atom.[3] A $Cu^{II}O_2^{2-}/Cu^I$-O_2^- species has also been proposed by Prigge as a possible reactive intermediate for the H-atom abstraction reaction.[21]

Chen and Solomon[28] have studied the oxygen activation and the reaction mechanism in this enzyme by applying DFT computational tools in model systems. They calculated the reaction thermodynamics and potential energy surfaces to investigate possible reactive Cu/O_2 species for H-atom abstraction in PHM. They have evaluated two possible mononuclear Cu/O_2 species, the 2-electron reduced Cu^{II}-OOH and the 1-electron reduced side-on Cu^{II}-superoxo intermediates. The substrate H-atom abstraction reaction by Cu^{II}-OOH, was found to be thermodynamically feasible, but with a very high activation barrier (37 kcal/mol). In contrast, H-atom abstraction from the substrate by the second intermediate was found to be a nearly isoenergetic process with a low reaction barrier (14 kcal/mol), suggesting that this species is the reactive one. They have also proposed a reasonable pathway for posterior substrate hydroxylation involving a "water-assisted" direct OH transfer to the substrate radical, generating a high-energy Cu^{II}-oxyl species, which provides the driving force for the ET to complete the reaction in PHM.

Kamachi et al.[29] have investigated the dopamine hydroxylation by the Copper–superoxo, –hydroperoxo, and –oxo species of DβM in order to identify the active species in the reaction and to reveal the key functions of the surrounding amino acid residues in substrate binding (see Figure 5.2). The authors performed DFT/QM-MM optimizations and subsequently evaluated the reactivity of the three oxidants with small model systems extracting the QM structures from the QM-MM simulations of the whole-enzyme model, by computing DFT/QM energy profiles. The activation energy of the H-abstraction from the substrate by the Cu^{II}-superoxo moiety and the posterior O-O cleavage and OH group transfer to the substrate radical were found to be ≈17 kcal/mol and ≈10 kcal/mol respectively, suggesting that Cu^{II}-superoxo species can promote the H-abstraction in DβM. They also rule out the Cu^{II}-hydroperoxo species as an active oxidant, because the activation barrier of the H-abstraction was found to be more than 40 kcal/mol. Finally, they concluded that the Cu^{III}-oxo species reasonably promotes the H-abstraction because the activation energy was found to by only 3.8 kcal/mol, and 6.5 kcal/mol for the final rebound step.
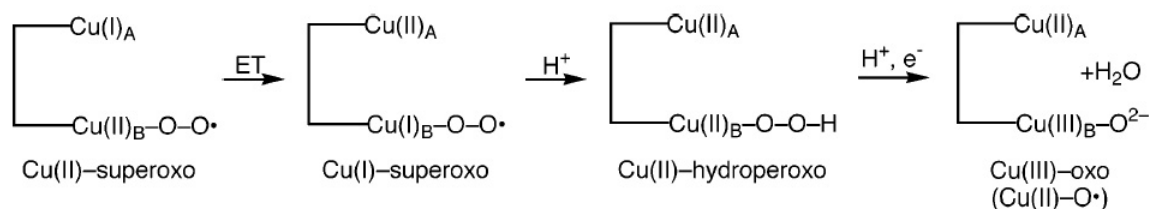
Figure 5.2. Reaction mechanism studied by Kamachi et al.[29] and proposed active species responsible for H-abstraction.

Tian et al.[30] proposed an oxo-mediated radical mechanism that is consistent with the results of the kinetic studies. In their mechanism the Cu^{II}-hydroperoxo species abstracts a hydrogen atom from a nearby tyrosine residue to produce a Cu^{III}-oxo species, a tyrosyl radical, and a water molecule. The produced Cu^{III}-oxo species abstracts a hydrogen atom from the benzylic position of dopamine with the formation of a radical intermediate in DβM. However, recent mutational studies[14] demonstrated that the reactivity of a Y318F mutant of PHM remains unchanged from the wild-type enzyme, which implies that the Tyr318 residue in PHM is not essential.

The Cu^{II}-superoxo species was also proposed to be responsible for the C-H bond activation in the DβM reaction by Evans.[31] The Cu^{II}-superoxo species is produced by dioxygen binding to the Cu_B center in the resting Cu^{I} state. This mechanism involves an initial hydrogen-atom abstraction from the benzylic position of dopamine, followed by the recombination between the substrate radical and the resultant Cu^{II}-hydroperoxo species to form norepinephrine. In this mechanism electron transfer from the Cu_A center occurs after the C-H bond activation in contrast to other mechanisms.

They have also demonstrated[31] that a molar ratio of O_2 consumption to product formation is 1:1 not only for highly reactive substrates but also for weakly reactive substrates. A mechanism in which the reductive activation of dioxygen precedes the C-H bond cleavage cannot explain such a tight coupling of oxygen and substrate consumption because either peroxide or superoxide leaks into the solvent from the reactive oxygen intermediate with extended lifetime.

Recently, the Amzel's laboratory[32] solved the x-ray structure of the precatalytic complex of PHM with bound peptide and oxygen, determined to 1.85 Å resolution. In this work they provided insight into several geometric and mechanistic issues of the PHM catalytic mechanism. Firstly, dioxygen binds to Cu_B with an end-on η^1 geometry in the precatalytic complex. The Cu_B-O-O angle is 110°, and the O-O distance refined to a value of 1.23 Å. This geometry is compatible with dioxygen or superoxide bound to copper, but not with Cu-peroxo species. They also proposed that an ET pathway between Cu_A and Cu_B would form when peptide substrate binds to reduced PHM,[19] and this pathway linked His_{108} (a Cu_A ligand), a water molecule, and the C terminus of the substrate through hydrogen bonds. Moreover, they concluded than once the precatalytic complex of reduced PHM-peptide-O_2 is formed, ET and H-abstraction occur. ET from Cu_A to the Cu_B-dioxygen species (Cu_B^{II}-O_2^-) takes place using the pathway described above. The resulting reduced oxygen species (Cu_B^{I}-O_2^-) abstracts the glycine *pro-S* hydrogen of the peptide substrate. The proposed mechanism is shown in Figure 5.3. H-abstraction requires that one of the oxygen atoms of O_2 comes in contact with the glycine *pro-S* hydrogen. A rotation of the distal oxygen atom by ≈110° around the Cu_B-O bond places the distal oxygen atom about 2.2 Å from the hydrogen atom—an ideal position to facilitate the hydrogen abstraction step. Inspection of the structure indicates that the barrier for this rotation is probably small and that thermal rotation of the distal oxygen atom could place it a short distance from the glycine *pro-S* hydrogen. As a final consideration, the authors propose that because the Cu-O_2 complex was not observed in structures of oxidized PHM or in structures of reduced PHM without substrate, the substrate binds to the reduced enzyme before oxygen will bind, which prevents the binding and activation of oxygen unless catalysis is primed to proceed.
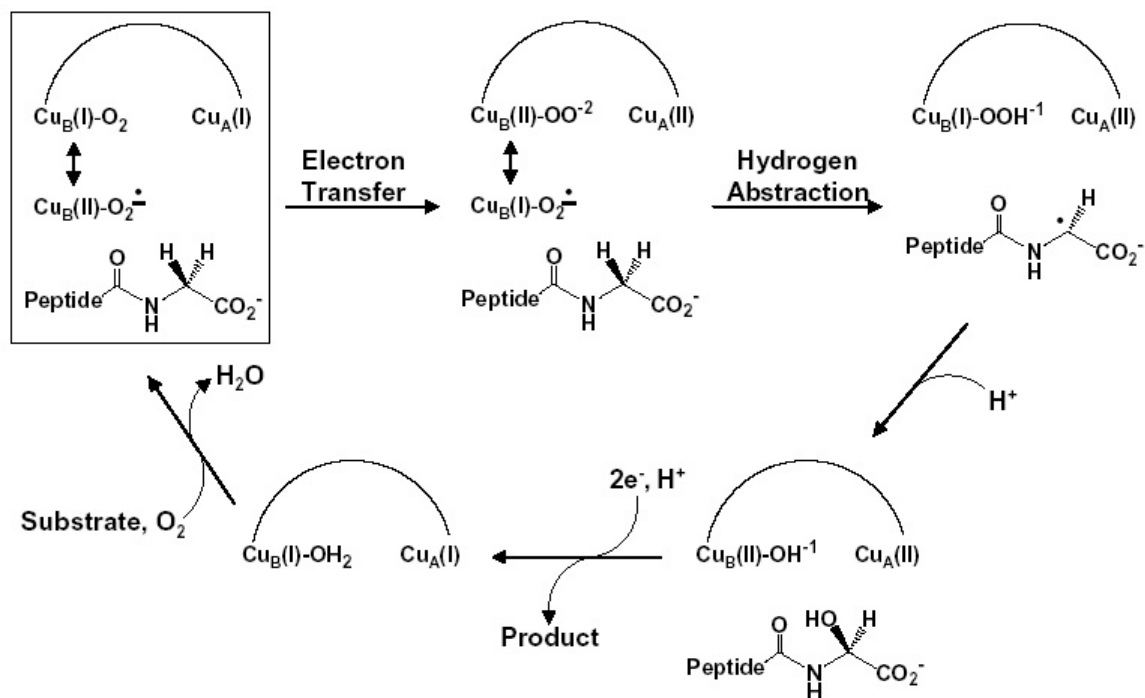
Figure 5.3. Reaction mechanism proposed by the Amzel's group.[32]

Taking all these contributions into account, it is clear that there is no consensus on the mechanism and/or the identity of the active species responsible for the H-abstraction of the PHM reaction. In this chapter of the thesis we will shed light on the molecular basis of the H-abstraction mechanism active in PHM by means of QM-MM calculations. We will start by calculating the O_2 binding energy to the Cu_B site in both oxidation states, before and after ET, and also in presence/absence of the substrate in order to answer whether if ET between Cu_A and Cu_B occurs when peptide substrate binds to reduced PHM and/or if the substrate binds to the reduced enzyme before oxygen binding. We will also compute the rotational barrier of the distal oxygen atom around the Cu_B-O bond in order to asses the role of this rotational motion in the catalytic mechanism. Finally, we will try to identify the most reactive oxygenated species active in PHM, paying particular attention to how this species is formed, in order to shed light on the molecular basis of the H-abstraction mechanism in PHM. All these insights may apply to other copper oxygen-activating enzymes, such as DβM, and to the design of biomimetic complexes.

Methods

Simulations were performed starting from the crystal structure of PHM with bound peptide and coordinated dioxygen, at 1.85 Å resolution (PDB entry 1sdw).[32] Hydrogen atoms were added as usual and the system was solvated using a 30 Å water cap. The system was equilibrated by

performing a 1 ns classical MD simulations at 300 K to obtain correct QM-MM starting geometries by using the Amber8 package.[33] Amber99[34] and TIP3P force fields were used to describe the protein/substrate and water, respectively. Both Cu_A and Cu_B model system charges were determined using RESP[35] charges and HF/6-31G(d) wave functions following the protocol recommended in the Amber web page. The van der Waals parameters were taken from Amber99. Even if the presence of the transition metal introduces an additional challenge in the simulation, MD simulations are now commonly applied for the investigation of metalloproteins.[36]

QM-MM calculations[37-42] were carried out to explore the effect of the environment on O_2 binding and rotational energy and in the H-abstraction reaction. We employed our own QM-MM implementation where the QM subsystem is treated at the DFT level using the SIESTA implementation.[43] The use of standard norm-conserving pseudopotentials[44] avoids the computation of core electrons, smoothing at the same time the valence charge density, as described in Chapter 2. In the present study, the nonlinear partial-core correction[45] is applied to the Cu atom. For all atoms, basis sets of double plus polarization quality were employed, with a pseudoatomic orbital energy shift of 30 meV and a grid cutoff of 150 Ry.[43] Calculations were performed using the generalized gradient approximation functional proposed by Perdew, Burke and Ernzerhof.[46] This combination of functional, basis sets, and grid parameters has been validated for both Cu_A and Cu_B isolated model systems. The classical subsystem was treated using the Amber99 force field parametrization.[34]

The initial structure for QM-MM calculations was taken from a selected snapshot of the classical MD simulation. We have selected the Cu_B plus the coordinated side-chains of residues His_{242}, His_{244} and Met_{314} and the O_2 ligand and a glycine substrate as the quantum subsystem. The rest of the protein, including the Cu_A center, and the water molecules were treated classically. We allowed free motion for QM atoms and for the MM atoms located inside a sphere of 13.5 Å from the QM subsystem center of mass. The frontier between the QM and MM portions of the system was treated by the SPLAM method as described in Chapter 2.

Ligand affinities (ΔE_L) were calculated using equation (5.1), where E_{Enz-L} is the energy of the ligand bound enzyme, E_{Enz} is the energy of the ligand free enzyme and E_L is the energy of the isolated ligand:

$$\Delta E_L = E_{Enz-L} - E_L - E_{Enz} \qquad (5.1)$$

Obtaining accurate free energy profiles requires an extensive sampling, which is computationally very expensive and difficult to achieve at the DFT QM-MM level. For these reasons potential energy profiles were determined using restrained energy minimizations along the reaction path that connects reactant, intermediate and product states, as described in Chapter 2.[47] For this purpose, an additional term was added to the potential energy according to $V(\xi) = k (\xi - \xi_0)^2$, where k is an adjustable force constant, ξ is the value of the reaction coordinate in the system particular configuration and ξ_0 is the reference value of the reaction coordinate (see below for the choice of the reaction coordinate in the different cases). Varying ξ_0, the system is forced to follow

the minimum reaction path along the given coordinate. The force constant for reaction coordinates involving distances was 200 kcal/(molÅ^2) and for reaction coordinates involving angles of 100 kcal/(molθ^2).

For all our calculations, the spin-unrestricted approximation was used. In each case two spin states were taken into consideration as will be shown later for each calculated species.

Results and Discussion

Oxygen Binding Energies

The oxygen affinities were calculated using equation (5.1), for both oxidation states and also in presence/absence of the substrate in order to determine if the ET occurs when peptide substrate binds to reduced PHM and/or if the substrate binds to reduced enzyme before oxygen binding. The PHM oxygenated resting state, $Cu_B^{II}-O_2^-$, has been evaluated in both the singlet and triplet spin states, being the more stable the triplet one, with and without substrate bound. However, without the coordinated oxygen, the spin ground state happens to be the singlet. After, ET, the oxygenated state $Cu_B^I-O_2^-$ has also been evaluated in both the doublet and quartet spin states, being the former one the spin ground state in all cases. The calculated oxygen binding energies in all cases is shown in Table 5.1, together with the high-spin/low-spin energy differences and the relevant Mulliken population analysis (*e*) and spin densities. Relevant numbering of the QM subsystem of all calculations is shown in Figure 5.4.

	$Cu_B^{II}-O_2^-$				$Cu_B^I-O_2^-$			
	with substrate		without substrate		with substrate		without substrate	
	oxy	free	oxy	free	Oxy	free	oxy	free
Spin State	*triplet*	*singlet*	*triplet*	*singlet*	*doublet*	*doublet*	*doublet*	*doublet*
$\Delta E_{HS\text{-}LS}$	-4.4	67.5	-4.4	63.2	79.5	53.6	30.0	70.3
Cu	0.562 (0.416)	0.322 (0.0)	0.534 (0.398)	0.342 (0.0)	0.537 (0.189)	0.038 (0.578)	0.533 (0.199)	0.320 (0.022)
O1	-0.200 (0.692)	-	-0.175 (0.719)	-	-0.426 (0.348)	-	-0.435 (0.339)	-
O2	-0.204 (0.724)	-	-0.214 (0.730)	-	-0.459 (0.431)	-	-0.508 (0.434)	-
ΔE_L	19.5		21.7		68.8		52.8	

Table 5.1. Oxygen binding energies (kcal/mol), high-spin/low-spin energy differences (kcal/mol), Mulliken charges (*e*) and spin densities (in parenthesis) for systems with and without coordinated substrate before and after ET.
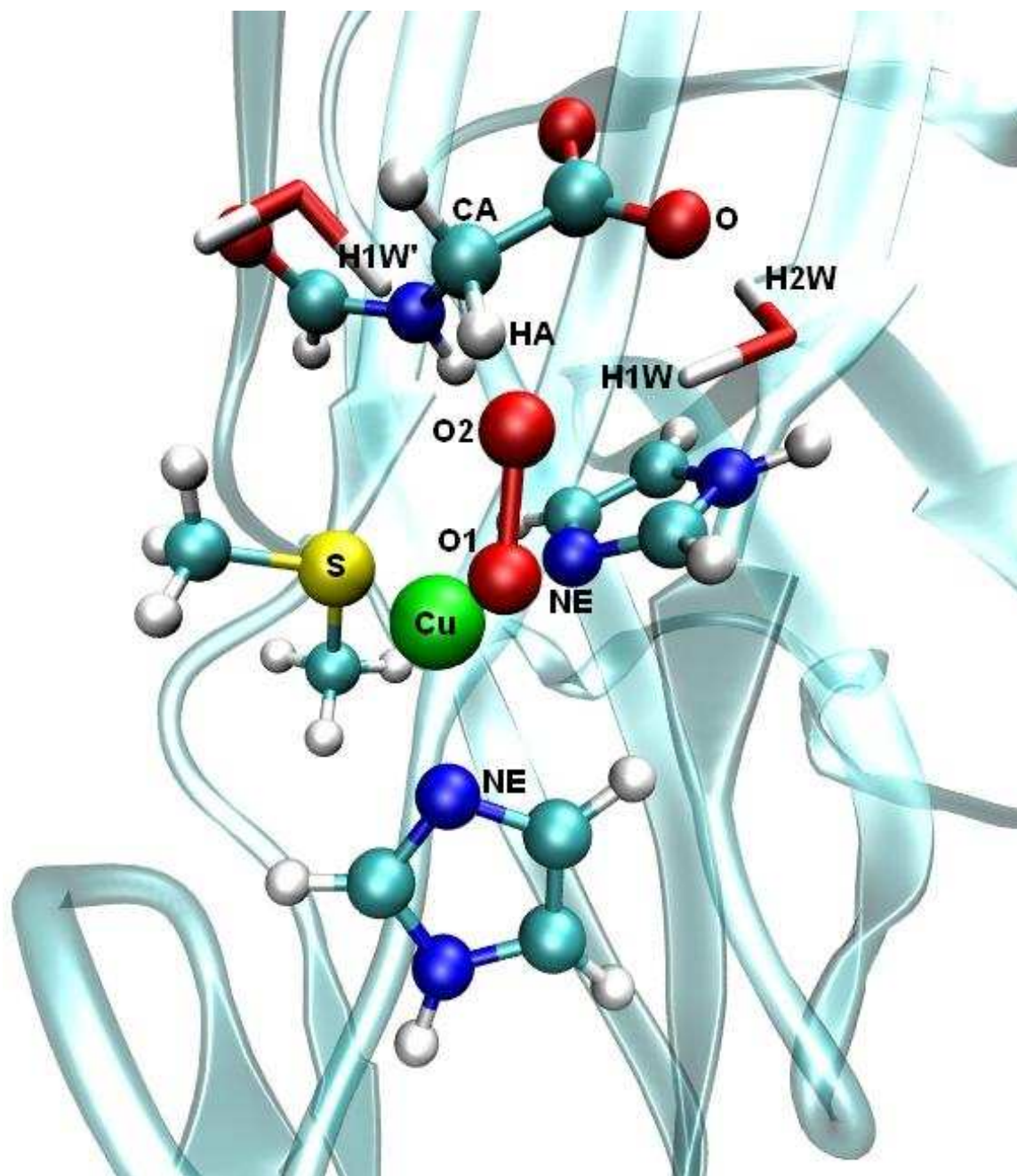
Figure 5.4. Relevant numbering of the QM subsystem employed in all calculations.

Table 5.2 shows some relevant geometrical parameters of all species involved in the calculation of the oxygen affinities. From both tables, it can be concluded that molecular oxygen binds to copper to obtain the Cu_B^{II}-O_2^- species but the substrate does not make any significant difference, as can be seen in the values for the oxygen binding energy of about 20 kcal/mol in both cases. However, once oxygen in coordinated, ET should occur to form the Cu_B^I-O_2^- species because the coordinated oxygen has a much higher binding energy in this case, of about 69 and 53 kcal/mol

with and without substrate, respectively. The difference in the oxygen affinity values between Cu_B^{II}-O_2^- and Cu_B^{I}-O_2^- can be attributed to the higher π-backdonation of the last one, which strengthens the bond. This effect is stronger for the metal in a low oxidation number, since in that case more electron density is available, as happens in the Cu_B^{I}-O_2^- species, and is reflected in a longer O1-O2 bond (1.31 Å and 1.40 Å in Cu_B^{II}-O_2^- and Cu_B^{I}-O_2^-, respectively) and an increase of the negative Mulliken population over coordinated O_2 (-0.404e and -0.885e in Cu_B^{II}-O_2^- and Cu_B^{I}-O_2^-, respectively) due to the larger population of the O_2 π* anti-ligand orbital. Moreover, as can also be seen the substrate considerably increases the binding energy in Cu_B^{I}-O_2^-, suggesting that O_2 is prone to coordination once the substrate is already placed in the enzymatic active site, as reported by Prigge et al.[32] This increase in the binding energy can be rationalized from Table 5.2. In the case with the substrate present, there is a strong H-bond interaction between O2 atom and the surrounding waters (O2-H1W distance of 1.63 Å and 1.81 Å with and without substrate, respectively). Moreover, this water, which is thought to be involved in the ET pathway, has a close contact with the substrate carboxylate (H2W-O_{subs} distance of 1.66 Å), which can also make the water position more constrained in order to produce an increase in the O_2 binding energy when the substrate is present. The O2 atom also strongly interacts with the substrate HA (O2-HA distance of 2.20 Å and HA atom Mulliken charge of -0.176e), further stabilizing the coordinated O_2 in presence of the substrate. Taken all this information into account, we can conclude that once the O_2 binds to PHM, ET occurs in order to form the Cu_B^{I}-O_2^- species which has a higher binding energy and also that O_2 binds to the reduced enzyme after substrate binding.

	Cu_B^{II}-O_2^-				Cu_B^{I}-O_2^-			
	with substrate		without substrate		with substrate		without substrate	
	oxy	free	oxy	free	oxy	free	oxy	free
O1-O2	1.31	-	1.30	-	1.40	-	1.38	-
Cu-O1	1.99	-	2.01	-	1.97	-	1.96	-
Cu-S$_{Met314}$	2.33	2.27	2.33	2.30	2.29	2.25	2.31	2.30
Cu-Nε$_{His242}$	2.04	1.99	2.04	1.99	2.17	1.93	2.12	1.95
Cu-Nε$_{His244}$	1.99	1.94	1.98	1.93	2.07	1.96	2.03	1.97
O2-HA	2.41	-	-	-	2.20	-	-	-
O2-H1W	1.84	-	2.04	-	1.63	-	1.81	-
O2-H1W'	2.05	-	2.14	-	1.77	-	1.75	-
H2W-O_{subs}	1.64	1.62	-	-	1.66	1.69	-	-
Cu-O1-O2	120.77	-	119.41	-	118.02	-	123.77	-
S$_{Met314}$-Cu-O1-O2	-65.01	-	-58.15	-	-29.54	-	-93.29	-

Table 5.2. Relevant geometrical parameters of all species involved in the calculation of the oxygen affinities (angstroms and degrees).

Oxygen Rotational Motion

In order to asses the role of the molecular motion involved in the PHM reaction and following the conclusions presented by Prigge et al.[32] we have calculated the potential energy profile along the S_{Met314}-Cu-O1-O2 dihedral angle in both Cu_B^{II}-O_2^- and Cu_B^{I}-O_2^- species in presence of substrate. Our results are shown in Figure 5.5. We have evaluated the dihedral angle in the direction where the O_2 molecule points towards the substrate HA atom. As can be seen in the figure, the Cu_B^{II}-O_2^- crystallographic angle of 40° corresponds to a local minimum along the surface.[32] Moreover, the rotation of the O_2 along this surface to reach a lower minimum, where the O2 atom points towards the substrate HA atom with distance lower than 2.4 Å, can be achieved thermally as described by Prigge et al.[32] because the energy barrier is of only about 2 kcal/mol. The surface has also two more minima corresponding to dihedral angles of -30° and -65° (this last corresponds to the rotation of about 110° proposed by Prigge et al.[32]), respectively (see also Table 5.2). Furthermore, if we take a look at the surface corresponding to the Cu_B^{I}-O_2^- species, the reaction coordinate also shows the presence of two minima at -30° and -85°, respectively and also a thermally accessible rotational barrier of about 1 kcal/mol. However, the stable crystallographic angle of 40° is not a minimum on this surface. This fact leads us to the conclusion that once the O_2 binds, leading to a Cu_B^{II}-O_2^- minimum conformation of 40°, ET occurs to reach the Cu_B^{I}-O_2^- species, but in an energy maximum. This conformation would spontaneously relax to one of its minimums, namely the -30° one, where the O2 atom is located directly towards the substrate HA at a distance of about 2.2 Å (see Table 5.2). This fact also supports the conclusions obtained in the preceding section. Finally, it is interesting to note that in both surfaces, the O2-HA distance corresponding to a reaction coordinate minimum, is of less than 2.4 Å in all cases.
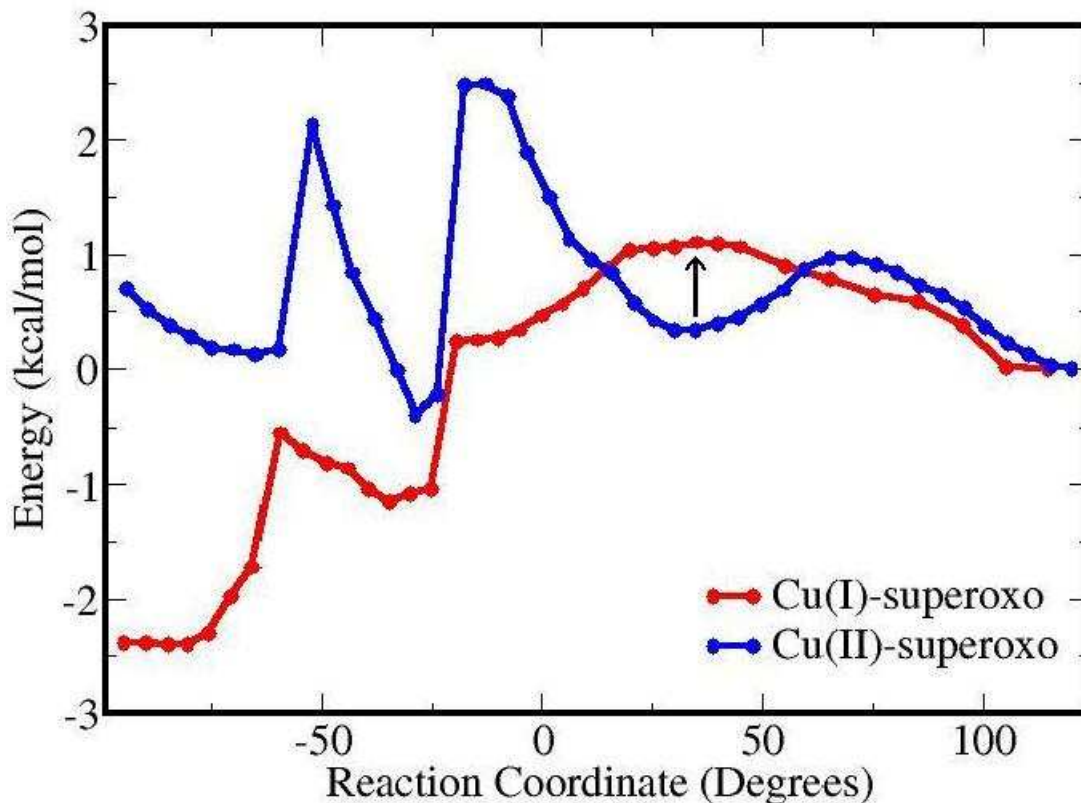
Figure 5.5. Potential energy profile (kcal/mol) along the S_{Met314}-Cu-O1-O2 dihedral angle (degrees) in both Cu_B^{II}-O_2^- and Cu_B^I-O_2^- species.

Hydrogen Abstraction Mechanism

In this section we will identify the most reactive oxygenated species active in PHM. We will pay attention particularly to how this species is formed, in order to shed light on the molecular basis of the H-abstraction mechanism operative in PHM.

We will start by considering the HA atom abstraction by coordinated O_2 in both Cu_B^{II}-O_2^- and Cu_B^I-O_2^- species, even though we have already determined that upon O_2 coordination, ET should probably occur. The H-abstraction by Cu_B^{II}-O_2^- can be represented by the following reaction:

$$Cu_B^{II}\text{-}O_2^- + HA\text{-}CA\text{-}subs \rightarrow Cu_B^{II}\text{-}OOHA + \bullet CA\text{-}subs \qquad (5.2)$$
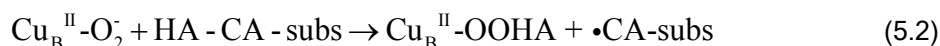
We have calculated the energy profile along the reaction coordinate defined as $\xi = d_{CA\text{-}HA}\text{-}d_{O2\text{-}HA}$ in both the singlet and triplet spin states. The results are shown in Figure 5.6. As can be seen, the ground state reactant is Cu_B^{II}-O_2^- in the triplet state (see Table 5.1), whereas the product has almost the same energy in both spin states and a change in the spin surface occurs upon reaction. The activation parameters are 25 (20) kcal/mol for the triplet (singlet) surface (the triplet is ≈5 kcal/mol more stable). This value is higher than the one obtained by Chen and Solomon[28] of

14 kcal/mol or Kamachi and co-workers[29] of 17 kcal/mol for isolated model systems. We think that in our case, where the whole protein is included, this imposes position constraints over the reactants, which may be responsible for the higher activation energy. The reaction coordinate changes from -1.3 Å in the reactant Cu_B^{II}-O_2^- to 1.3 Å in the Cu_B^{II}-hydroperoxo formed, and the O1-O2, O2-HA and CA-HA distances are 1.31 Å, 2.42 Å and 1.12 Å; and 1.46 Å, 1.00 Å and 2.29 Å, in reactants and products, respectively. Although we believe that this species is probably not the responsible for H-abstraction as discussed above, the activation energy of this reaction is also very high compared to the experimental estimation of $\Delta H^{\ddagger}$ for the C-H cleavage step in PHM of ≈13 kcal/mol obtained by Francisco et al.[13] in a temperature dependence intrinsic primary isotope effect determination.
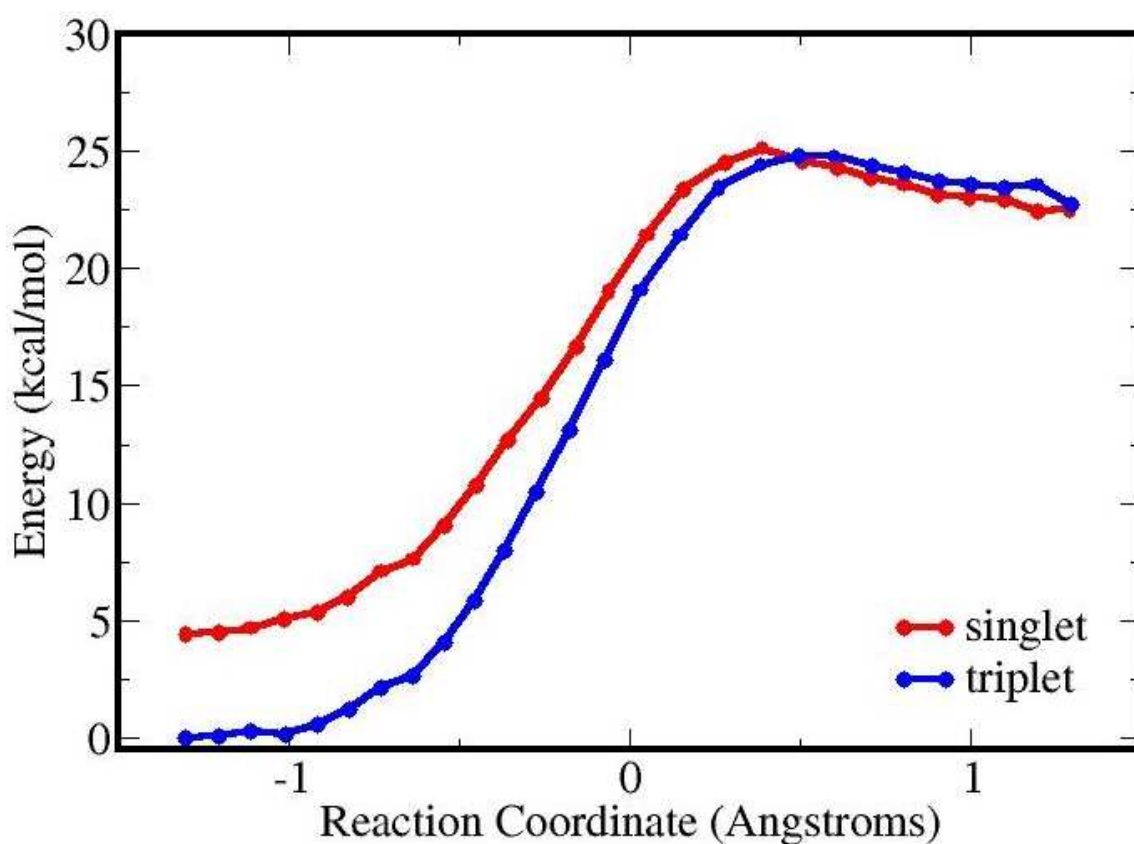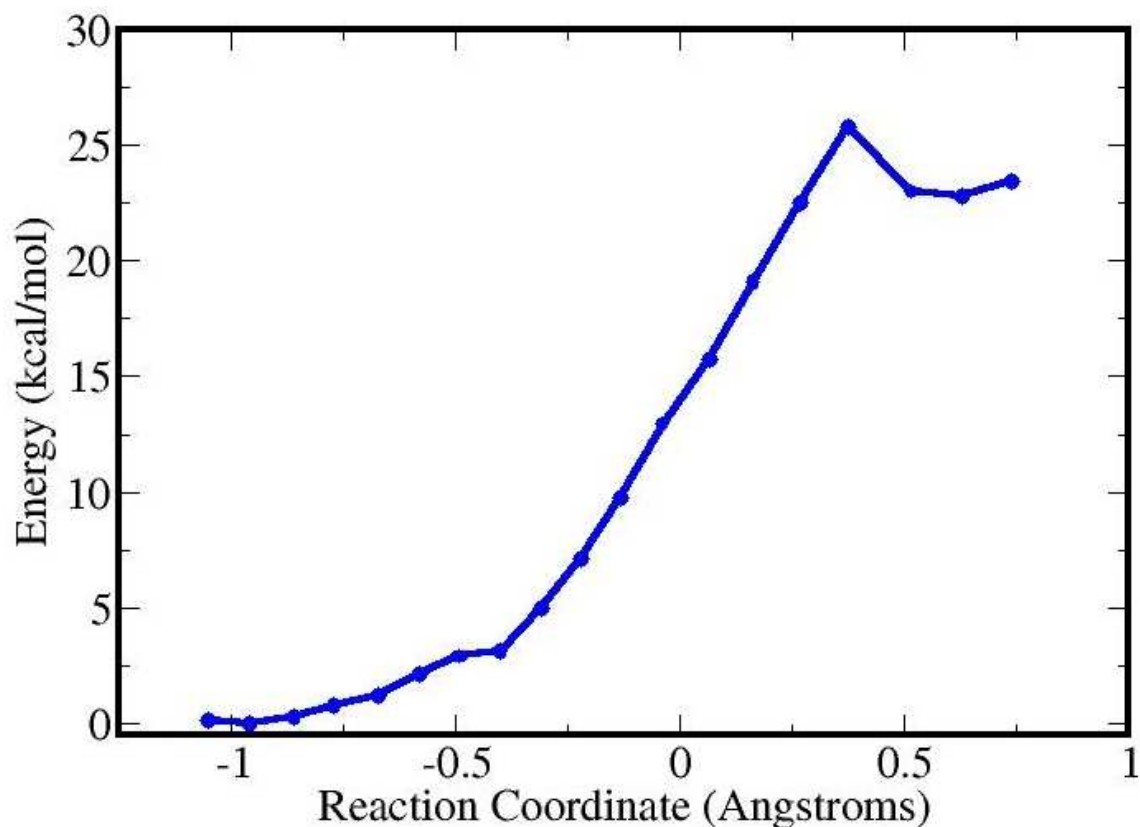

Figure 5.6. Potential energy profile (kcal/mol) for HA-abstraction reaction catalyzed by Cu_B^{II}-O_2^-, reaction (5.2) in both singlet (red) and triplet (blue) spin surfaces. The reaction coordinate is -1.3 Å in the reactant and 1.3 Å in the product.

The H-abstraction catalyzed by Cu_B^{I}-O_2^- can be described by a similar reaction:

$$Cu_B^{I}\text{-}O_2^- + HA\text{-}CA\text{-}subs \rightarrow Cu_B^{I}\text{-}OOHA + \cdot CA\text{-}subs \qquad (5.3)$$
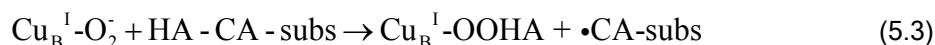
The reaction coordinate was defined also as $\xi = d_{CA-HA}-d_{O2-HA}$, and we have calculated both doublet and quartet spin energy profiles. However, as the doublet lays 80 kcal/mol lower in energy from the quartet (see Table 5.1) and both surfaces are similar in shape, we only show the results of the doublet surface in Figure 5.7. The activation energy is this case is of about 25 kcal/mol in both spin states. The reaction coordinate goes form -1.0 Å in the reactant Cu_B^I-O_2^- to 0.7 Å in the Cu_B^I-hydroperoxo formed, and the O1-O2, O2-HA and CA-HA distances are 1.40 Å, 2.17 Å and 1.12 Å; and 1.49 Å, 1.04 Å and 1.78 Å, in reactants and products, respectively. Although Cu_B^I-O_2^- was proposed by Prigge et al.[32] to be the active species responsible for HA-abstraction in PHM, our high activation energy value of 25 kcal/mol compared to the much lower experimental value,[13] leads us to think that Cu_B^I-O_2^- is not also the active H-abstracting oxygenated species in PHM.


Figure 5.7. Potential energy profile (kcal/mol) for HA-abstraction reaction catalyzed by Cu_B^I-O_2^-, reaction (5.3) in the doublet spin surface. The reaction coordinate is -1.0 Å in the reactant and 0.7 Å in the product.

At this point is interesting to find a new oxygenated species that can be responsible for the H-abstraction process and to further rationalize the mechanism active in PHM. Chen and Solomon[28] and Kamachi and co-workers[29] proposed the copper-hydroperoxo, Cu_B^{II}-OOH, formed upon H^+ abstraction from water by Cu_B^I-O_2^-. However, they obtained a very high activation energy (≈40

kcal/mol) for H-abstraction, arguing against this species as the reactive one. Nevertheless, it is interesting to evaluate how this species is formed and also to confirm that it is not the reactive oxygenated H-abstracting species active in PHM. In order to do so, we have evaluated both the formation of the Cu_B^{II}-OOH species and the further H-abstraction reaction. The first reaction can be summarized as follows:

$$Cu_B^{I}\text{-}O_2^- + H_3O^+ \rightarrow Cu_B^{II}\text{-}OOH_W + H_2O \tag{5.4}$$
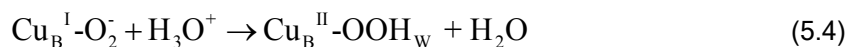
Here, we have added three O_2 coordination water molecules and a proton into the QM subsystem in order to perform our calculations. The reaction coordinate is defined as follows: $\xi = d_{O2\text{-}HW} - d_{HW\text{-}OW}$, and we have calculated the reaction energy profile only for the doublet spin state, because the quartet lays 14.5 kcal/mol higher. Our results are shown in Figure 5.8. As it can be seen the reaction occurs spontaneously with no activation energy, and with an energy difference of -17.5 kcal/mol. We can conclude then, that the Cu_B^{I}-O_2^- species spontaneously abstracts a proton from the surrounding water molecules to form Cu_B^{II}-OOH. The reaction coordinate goes form -0.6 Å in the reactant Cu_B^{I}-O_2^- to 0.55 Å in the Cu_B^{II}-hydroperoxo formed, and the O1-O2, O2-HW and HW-OW distances are 1.45 Å, 1.64 Å and 1.03 Å; and 1.43 Å, 1.03 Å and 1.58 Å, in reactants and products, respectively. The substrate H-abstraction reaction catalyzed by Cu_B^{II}-OOH can be written as:

$$Cu_B^{II}\text{-}OOH_W + HA\text{-}CA\text{-}subs \rightarrow Cu_B^{II}\text{-}O\text{-}(OH_W HA) + \bullet CA\text{-}subs \tag{5.5}$$
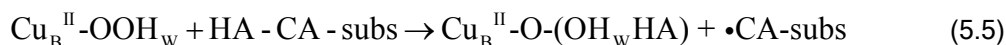
The QM subsystem is also the same as the reaction before. The reaction coordinate is now defined as follows: $\xi = d_{CA\text{-}HA} - d_{O2\text{-}HA}$, and we also have calculated the reaction energy profile only for the doublet spin state. The corresponding energy profile is shown also in Figure 5.8. We also can conclude that the Cu_B^{II}-OOH species can not abstract the hydrogen from the substrate given the fact that our activation energy is 42.8 kcal/mol, and compares well with the 40 kcal/mol obtained by Chen and Solomon[28] and Kamachi and co-workers.[29] The reaction coordinate varies form -1.7 Å in reactants to 0.55 Å in products, and the O1-O2, O2-HA and CA-HA distances are 1.47 Å, 2.82 Å and 1.12 Å; and 1.55 Å, 1.10 Å and 1.65 Å, in reactants and products, respectively.
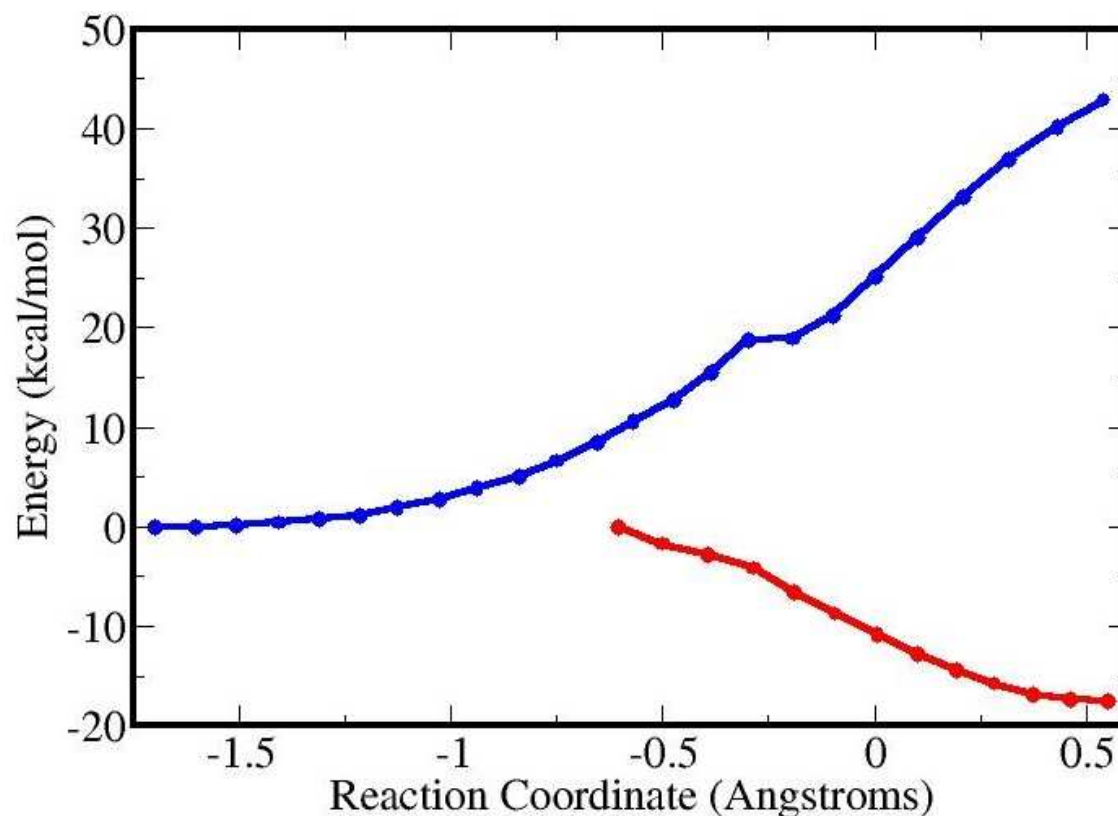
Figure 5.8. Potential energy profile (kcal/mol) for proton abstraction from water by Cu_B^I-O_2^- to form Cu_B^{II}-OOH (red) equation (5.4), and HA-abstraction reaction catalyzed by Cu_B^{II}-OOH (blue) reaction (5.5), both in the doublet spin surface. The reaction coordinate is -0.6 Å in the reactant and 0.55 Å in the product for equation (5.4) and -1.7 Å in the reactant and 0.55 Å in the product for reaction (5.5).

Now it is interesting to think how reaction can continue if we assume that the Cu_B^{II}-OOH species is formed as concluded before. Following the reaction stoichiometry, in which two protons are consumed and a water molecule is produced, it is not surprising that the Cu_B^{II}-OOH molecule formed will abstract a second proton from a surrounding water. If we add another proton to Cu_B^{II}-OOH, and optimize the resulting structure, we obtain the following species: $[Cu_BO^{.....}HOH]^{+2}$. The new water molecule coordinated to O1 (HW-O2-HW') spontaneously twists and coordinates to the O1 atom by one of its hydrogen atoms. The optimized Cu-O1 and O1-HW distances are 1.80 Å and 2.21 Å, respectively. This fact suggests that once the Cu_B^I-O_2^- is formed, spontaneously abstracts one proton to yield the Cu_B^{II}-OOH species and subsequently abstracts another proton to produce the $[Cu_BO]^{+2}$ species with the release of a water molecule. Taking all this information into account we are now interested in checking if these copper-monooxygenated species can abstract the substrate hydrogen. Kamachi and co-workers[29] proposed that the species responsible for H-abstraction in DβM is $[Cu_BO]^{+1}$, arising from ascorbate reduction of the $[Cu_BO]^{+2}$ formed, before substrate H-abstraction. In order to correctly characterize the active oxygenated H-abstracting

species in PHM, we have performed the potential energy profiles for H-abstraction catalyzed by both $[Cu_BO]^{+2}$ and $[Cu_BO]^{+1}$ species.

The H-abstraction catalyzed by $[Cu_BO]^{+2}$ can be written as follows:

$$[Cu_BO]^{+2} + HA\text{-}CA\text{-}subs \rightarrow [Cu_BOHA]^{+2} + \bullet CA\text{-}subs \qquad (5.6)$$

The reaction coordinate is as follows: $\xi = d_{CA-HA}\text{-}d_{O1-HA}$, and we have calculated the reaction energy profile for both the doublet and quartet spin states. The results are shown in Figure 5.9. As can be seen, the quartet spin state is the ground state in reactants and is stabilized by 3.8 kcal/mol with respect to the doublet state. Following the energy profile in the quarter state gives an intermediate state where the abstracted hydrogen forms an OH coordinated to the metal. The activation energy is 0.15 kcal/mol and the intermediate energy is 9.4 kcal/mol lower than the reactant. The reaction coordinate varies form -1.0 Å in reactants to 1.0 Å in products, and the Cu-O1, O1-HA, CA-HA and O1-CA distances are 1.79 Å, 2.10 Å, 1.12 Å and 3.14 Å; and 1.86 Å, 1.00 Å, 2.00 Å and 2.99 Å, in reactants and products, respectively. The optimized reactive specie $[Cu_BO]^{+2}$, accommodates the O1 atom in order to abstract the HA, as can be seen in the short O1-HA optimized distance of 2.10 Å. The second half of the hydroxylation reaction, the OH rebound step, will be considered later. However, if we follow the energy profile for the doublet spin state, our results are quite different and surprising. In fact, the obtained product is the overall reaction product, the hydroxylated substrate. The activation energy is only 0.16 kcal/mol and the overall energy difference -69.9 kcal/mol. This product is ≈60 kcal/mol more stable than the previous intermediate formed on the quartet surface and ≈83 kcal/mol more stable than the corresponding hydroxylated product in the quartet spin state. The reaction coordinate goes form -1.1 Å in reactants to 0.9 Å in products, and the Cu-O1, O1-HA, CA-HA and O1-CA distances are 1.80 Å, 2.22 Å, 1.12 Å and 3.30 Å; and 3.74 Å, 1.00 Å, 1.90 Å and 1.43 Å, in reactants and products, respectively, which reflects the spontaneous hydroxylation of the substrate, as can be seen in the shortening of the O1-CA bond distance (1.43 Å) in the product. Thus, the correct reaction on the doublet surface must be considered as:

$$[Cu_BO]^{+2} + HA\text{-}CA\text{-}subs \rightarrow [Cu_B]^{+2} + HA\text{-}O1\text{-}CA\text{-}subs \qquad (5.7)$$
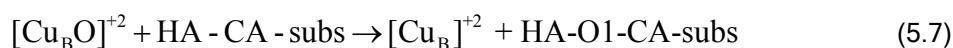
Taking this information into account, we can conclude that the reaction pathway should involve a spin-inversion electronic process near a crossing region between the quartet and doublet potential energy surfaces, thus the stable $[Cu_BO]^{+2}$ quartet state should change to the doublet one upon H-abstraction, in order to form the hydroxylated product. To complete the reaction, a water molecule should probably coordinate to $[Cu_B]^{+2}$ to yield $[Cu_B\text{-}OH_2]^{+2}$, because Cu complexes typically exhibit tetrahedral coordination. Inspection of the final optimized structure indicates that there are two water molecules (shown in Figure 5.4) close to the Cu atom with distances from about 4.6 to 5.7 Å. One of these water molecules may be prone to coordinate the $[Cu_B]^{+2}$. Finally, the $[Cu_B\text{-}OH_2]^{+2}$ should be reduced by ascorbate to yield $[Cu_B\text{-}OH_2]^{+1}$ and start another catalytic cycle upon product release.
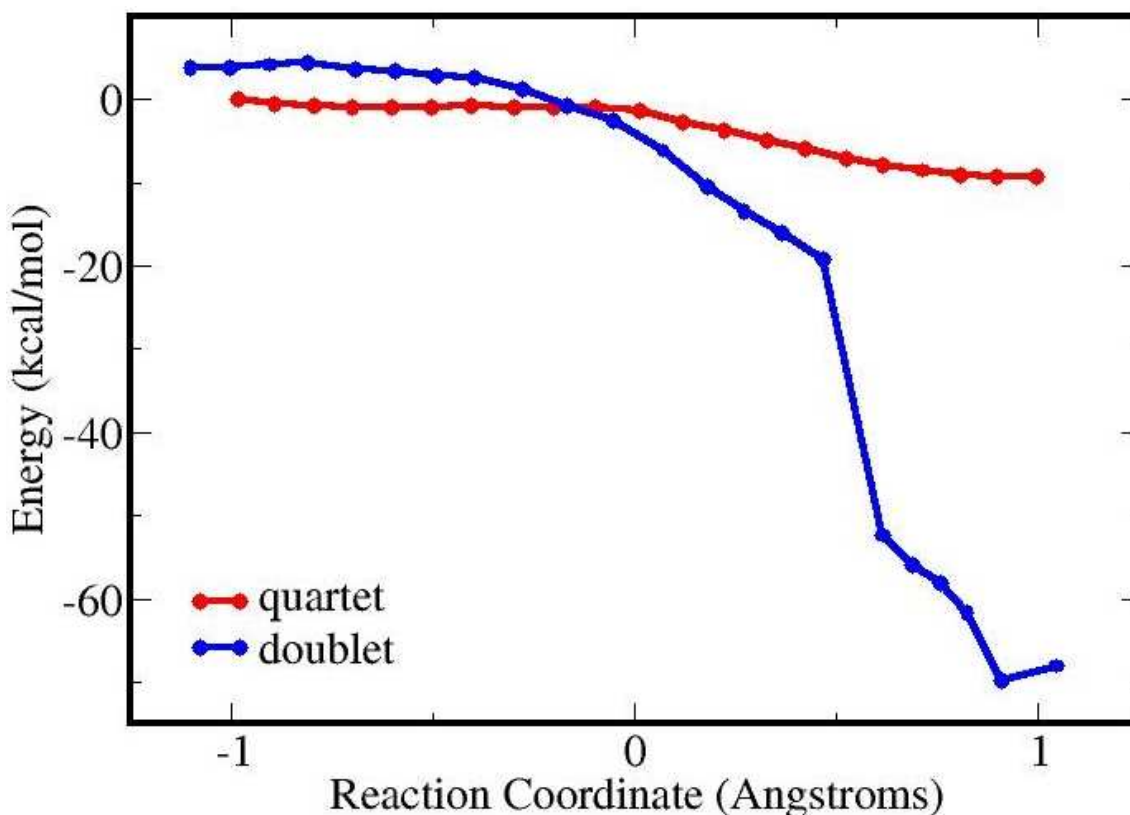
Figure 5.9. Potential energy profile (kcal/mol) for HA-abstraction reaction catalyzed by $[Cu_BO]^{+2}$ in both quartet (red), reaction (5.6) and doublet (blue), reaction (5.7) spin surfaces. The reaction coordinate goes from ≈-1.0 Å in reactants to ≈1.0 Å in products in both cases.

Let us now consider the H-abstraction catalyzed by $[Cu_BO]^{+1}$, obtained after ascorbate reduction, in order to correct identify the abstracting species as proposed by Kamachi and co-workers.[29] The H-abstraction catalyzed by $[Cu_BO]^{+1}$ reaction is:

$$[Cu_BO]^{+1} + HA\text{-}CA\text{-}subs \rightarrow [Cu_BOHA]^{+1} + \bullet CA\text{-}subs \qquad (5.8)$$

The reaction coordinate is again: $\xi = d_{CA\text{-}HA}\text{-}d_{O1\text{-}HA}$, and we have calculated the reaction energy profile for both the singlet and triplet spin states. The results are shown in Figure 5.10. As can be seen, the triplet spin state is the ground state in reactants and is stabilized by 4.9 kcal/mol with respect to the singlet state, which is almost the same as found by Kamachi and co-workers[29] (6.5 kcal/mol). Following the energy profile in the stable triplet state gives an intermediate where the abstracted hydrogen forms an OH coordinated to the metal, with activation energy of 4.1 kcal/mol and energy difference of -7.1 kcal/mol. These values compare reasonably well with the ones obtained by Kamachi and co-workers[29] (activation energy 3.8 kcal/mol and energy difference -20 kcal/mol). The energy profile for the excited singlet spin state is almost the same than the triplet one, but the activation energy is reduced to 1.8 kcal/mol and the intermediate formed lays only

0.3 kcal/mol higher than the triplet one, as can be seen in the figure. The reaction coordinate varies form -1.05 (-1.05) Å in reactants to 1.04 (1.05) Å in products on the triplet (singlet) surface, and the Cu-O1, O1-HA, CA-HA and O1-CA distances are 1.80 (1.80) Å, 2.18 (2.17) Å, 1.13 (1.12) Å and 3.28 (3.27) Å; and 1.89 (1.90) Å, 0.99 (0.99) Å, 2.03 (2.04) Å and 2.98 (2.97) Å in reactants and products on the triplet (singlet) surface, respectively.
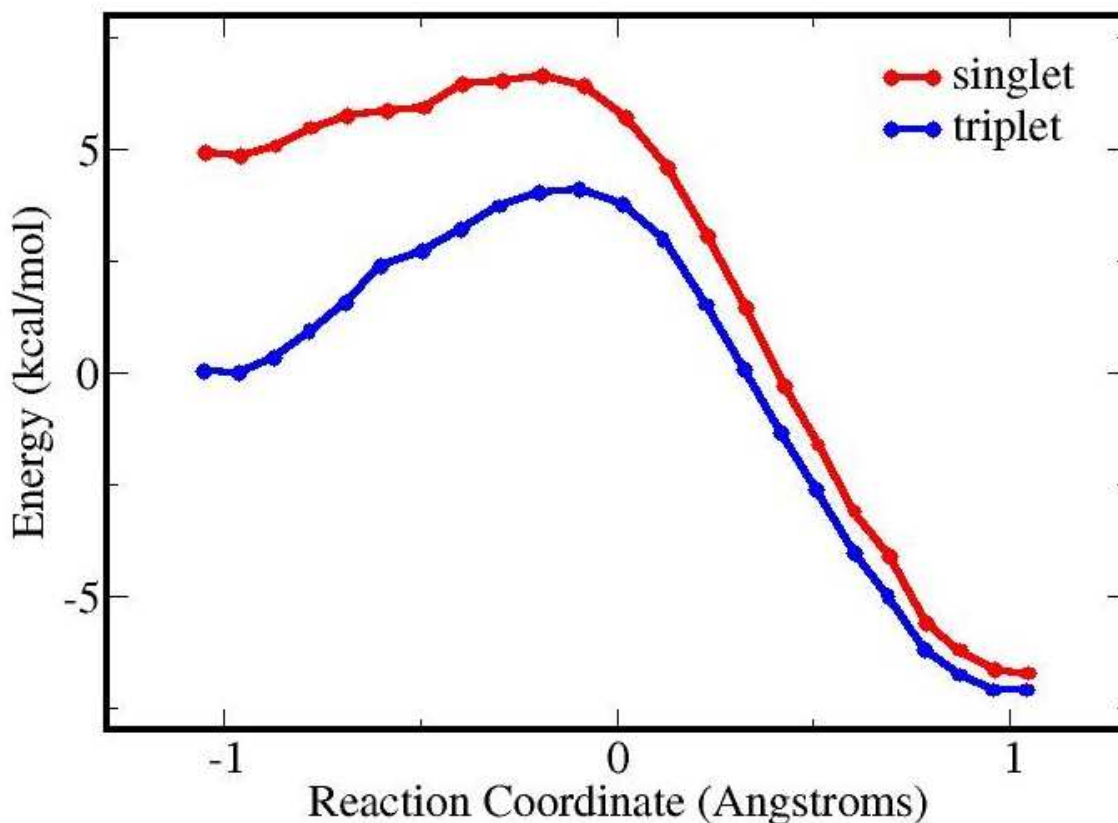

Figure 5.10. Potential energy profile (kcal/mol) for HA-abstraction reaction catalyzed by [Cu$_B$O]$^{+1}$ in both singlet (red) and triplet (blue) spin surfaces, corresponding to reaction (5.8). The reaction coordinate goes from ≈-1.0 Å in reactants to ≈1.0 Å in products in both cases.

Finally, we will consider the OH rebound step, in order to correctly complete the operative mechanism in PHM and decide whether if the [Cu$_B$O]$^{+1}$ is also an active oxygenated species. We will study this reaction in all the cases that the OH coordinated intermediate has been formed, namely: [Cu$_B$O]$^{+2}$ in the quartet spin surface and [Cu$_B$O]$^{+1}$ in both the singlet and triplet ones. The rebound reaction can be summarized as follows:

$$[Cu_BOHA]^{+2(1)} + \cdot CA\text{-subs} \rightarrow [Cu_B]^{+2(1)} + HA\text{-O1-CA-subs} \qquad (5.9)$$
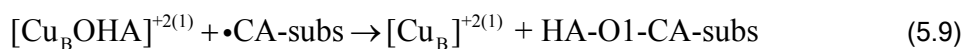
In all cases the reaction coordinate chosen was: $\xi = d_{CA\text{-}O1}$. All the results are shown in Figure 5.11. Let us start by considering the rebound step for the [Cu$_B$OHA]$^{+2}$ species in the quartet spin

surface. As can be seen, the energy difference for this process is 37.8 kcal/mol, suggesting that this reaction is not thermodynamically spontaneous. This fact is in agreement with our previous results, which state there must be a spin-inversion electronic process between the quartet and doublet potential energy surfaces, being the last one that completes the overall hydroxylation reaction for the $[Cu_BO]^{+2}$ species in one step; and with the fact that the hydroxylated formed product is 83 kcal/mol less stable in the quartet spin surface. The reaction coordinate varies from 2.99 Å in reactants to 1.39 Å in products, and the Cu-O1, O1-HA, CA-HA and O1-CA distances are 1.86 Å, 1.00 Å, 2.00 Å and 2.99 Å; and 3.37 Å, 0.99 Å, 1.92 Å and 1.39 Å, in reactants and products, respectively. However, if we consider the rebound reaction for the $[Cu_BOHA]^{+1}$ species the results are quite different. The triplet reactant is only 0.3 kcal/mol more stable, as said before. However, the triplet surface has an activation energy of 37.3 kcal/mol, whereas the singlet only of 5.4 kcal/mol. This fact suggests that in order to form the hydroxylated product, there must also be a spin-inversion electronic process between the triplet and singlet potential energy surfaces. These results are in agreement with those obtained by Kamachi and co-workers[29] (51.4 and 6.5 kcal/mol for the triplet and singlet surfaces, respectively and presence of a spin-inversion process). The energy difference to form the hydroxylated product is of -40.9 kcal/mol in the singlet surface. The reaction coordinate varies form 3.00 (3.00) Å in reactants to 1.42 (1.42) Å in products for the singlet (triplet) surface, and the Cu-O1, O1-HA, CA-HA and O1-CA distances are 1.90 (1.89) Å, 0.99 (0.99) Å, 2.06 (2.03) Å and 3.00 (2.98) Å; and 3.29 (3.29) Å, 0.99 (1.02) Å, 1.98 (2.05) Å and 1.42 (1.42) Å, in reactants and products, respectively for the singlet (triplet) surface. These results allow us to conclude that for the rebound step for the $[Cu_BO]^{+1}$ species only the singlet state can be mechanistically correct.

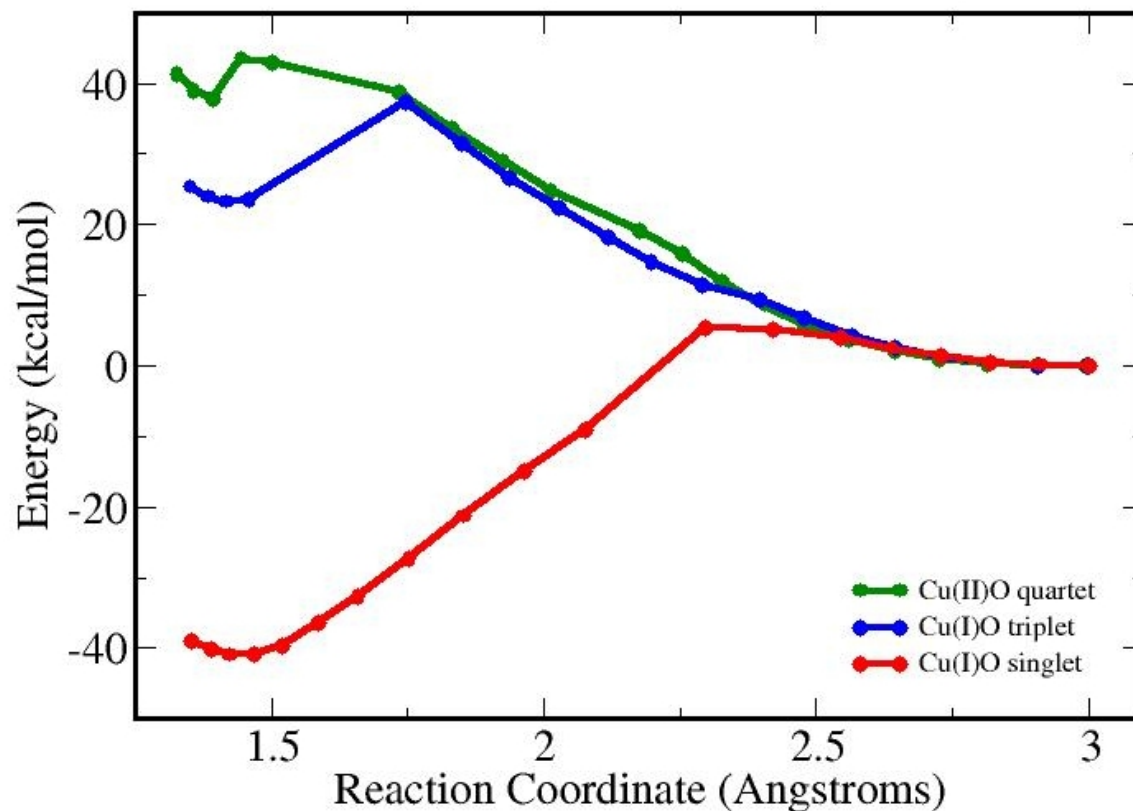**Figure 5.11. Potential energy profile (kcal/mol) for the rebound reaction catalyzed by [Cu$_B$O]$^{+1}$ in both singlet (red) and triplet (blue) spin surfaces and by [Cu$_B$O]$^{+2}$ in the quartet spin surface (green), corresponding to reaction (5.9). The reaction coordinate goes from ≈3.0 Å in reactants to ≈1.4 Å in products in all cases.**

Taking all the results described above, we can conclude that the active species responsible for the H-abstraction in PHM is probably [Cu$_B$O]$^{+2}$. This species is formed exothermically and can abstract the substrate HA atom with almost zero activation energy yielding the hydroxylated product. Finally, a water molecule coordinates to [Cu$_B$]$^{+2}$ and ascorbate reduces [Cu$_B$-OH$_2$]$^{+2}$ to [Cu$_B$-OH$_2$]$^{+1}$ which will start another catalytic cycle upon product release. These results allow us to suggest the following mechanism involved in PHM, which is schematically shown in Figure 5.12.
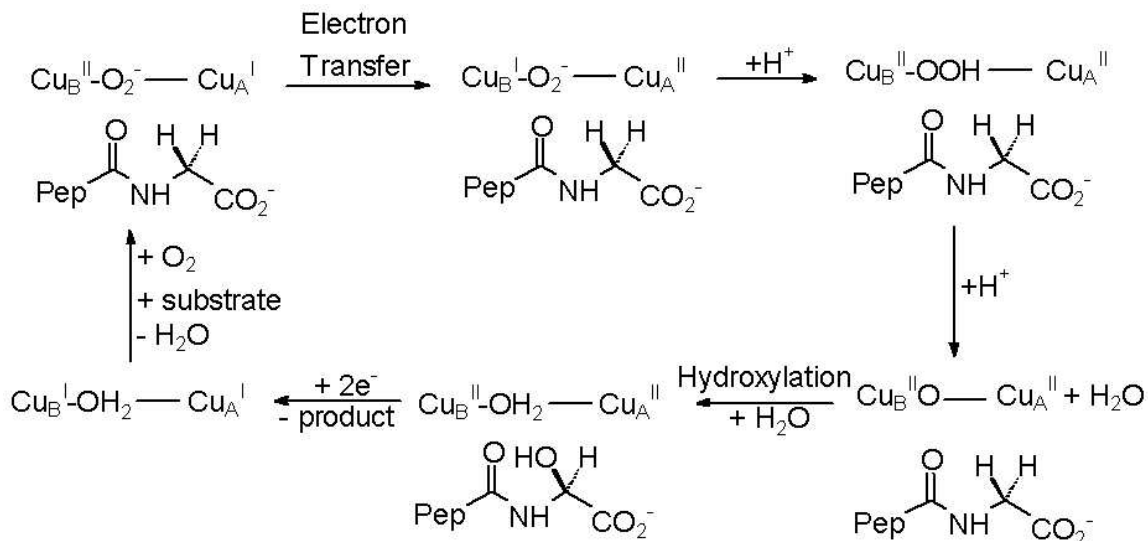
Figure 5.12. Proposed mechanism operative in PHM.

Conclusions

In this work we have studied the hydroxylation reaction of the $C\alpha$ of a C-terminal glycine catalyzed by PHM, by means of QM-MM computational techniques. Our results regarding the calculation of the O_2 binding energy to the Cu_B suggest that molecular oxygen binds to copper to obtain the $Cu_B{}^{II}\text{-}O_2{}^-$ species and then ET occurs to form the $Cu_B{}^{I}\text{-}O_2{}^-$ species because the coordinated oxygen has a much higher binding energy in this case, due to a greater π-backdonation effect. Moreover, the substrate considerably increases the binding energy in $Cu_B{}^{I}\text{-}O_2{}^-$, due to closer interactions and geometric constraints, suggesting that O_2 is prone to coordination once the substrate is already placed in the enzymatic active site. These results are also in agreement with the rotational barrier calculation of the distal oxygen atom around the Cu_B-O bond. Once the O_2 binds, leading to a $Cu_B{}^{II}\text{-}O_2{}^-$ minimum conformation of 40°, ET occurs to form the $Cu_B{}^{I}\text{-}O_2{}^-$ species but in an energy maximum conformation, which will subsequently relax to one of its minimums, where the oxygen atom is located directly towards the substrate HA. We have also identified the most reactive oxygenated species active in PHM, taking particular attention to how this species was formed and have shown new insights into the molecular basis of the H-abstraction mechanism operative in PHM. We have computed potential energy profiles for the H-abstraction reaction for both $Cu_B{}^{II}\text{-}O_2{}^-$ and $Cu_B{}^{I}\text{-}O_2{}^-$ species, concluding that none of them is the responsible for the reaction due to their high activation energies. We have also found, by inspecting the potential energy surfaces, that the $Cu_B{}^{II}\text{-}OOH$ species is spontaneously formed by abstracting a neighbor water proton, but cannot abstract the hydrogen atom from the substrate. However, it can spontaneously abstract another water proton to form $[Cu_BO]^{+2}$ with water release.

Finally, we have evaluated if this monooxygenated species and the one obtained by prior reduction with ascorbate, $[Cu_BO]^{+1}$, were able to promote the substrate H-abstraction. Our results show that both species are able to produce hydrogen abstraction. Whereas the $[Cu_BO]^{+2}$ abstracts the HA atom concertedly with almost no activation energy and involving a spin-inversion electronic process between the quartet reactant ground state and doublet product ground state potential energy surfaces, the $[Cu_BO]^{+1}$ abstracts the hydrogen to form an intermediate with OH coordinated in the triplet surface with activation energy of 4.1 kcal/mol. However, this intermediate continues the reaction by a rebound step, which has activation energy of 5.4 kcal/mol, and also involves a spin-inversion between the triplet reactant ground state and singlet product ground state potential energy surfaces.

Taking all these contributions into account, we believe that the active species responsible for the H-abstraction in PHM, which is consistent with the experimental results and with the lower overall activation energy is $[Cu_BO]^{+2}$. This species is formed exothermically and can abstract the substrate HA atom without activation energy forming the hydroxylated product. Finally, a neighbor water molecule coordinates and ascorbate reduces $[Cu_B\text{-}OH_2]^{+2}$ to $[Cu_B\text{-}OH_2]^{+1}$ which will start another catalytic cycle upon product release.

These results allow us to suggest that the mechanism active in PHM is the one shown in Figure 5.12. All these insights regarding understanding the catalytic mechanism may be a general feature of oxygen-activating metalloenzymes and may apply to other copper oxygen-activating enzymes, such as DβM, and to the design of biomimetic complex.

[1] Holm, R. H.; Kennepohl, P.; Solomon, E. I. Chem. Rev. 1996, 96, 2239.

[2] Solomon, E. I.; Sundaram, U. M.; Machonkin, T. E. Chem. Rev. 1996, 96, 2563.

[3] Klinman, J. P. Chem. Rev. 1996, 96, 2541.

[4] Cuttitta, F. Anat. Rec. 1993, 236, 87.

[5] Merkler, D. J. Enzyme Microb. Technol. 1994, 16, 450.

[6] Eipper, B. A.; Stoffers, D. A.; Mains, R. E. Annu. Rev. Neurosci. 1992, 15, 57.

[7] Bradbury, A. F.; Finnie, M. D.; Smyth, D. G. Nature 1992, 298, 686.

[8] Eipper, B. A.; Milgram, S. L.; Husten, E. J.; Yun, H. Y.; Mains, R. E. Protein Sci. 1993, 2, 489.

[9] Bell, J.; El Meskini, R.; D'Amato, D.; Mains, R. E.; Eipper, B. A. Biochemistry 2003, 42, 7133.

[10] Eipper, B. A.; Quon, A. S.; Mains, R. E.; Boswell, J. S.; Blackburn, N. J. Biochemistry 1995, 34, 2857.

[11] Takahashi, K.; Onami, T.; Noguchi, M. Biochem. J. 1998, 336, 131.

[12] Francisco, W. A.; Merkler, D. J.; Blackburn, N. J.; Klinman, J. P. Biochemistry 1998, 37, 8244.

[13] Francisco, W. A.; Knapp, M. J.; Blackburn, N. J.; Klinman, J. P. J. Am. Chem. Soc. 2002, 124, 8194.

[14] Francisco, W. A.; Blackburn, N. J.; Klinman, J. P. Biochemistry 2003, 42, 1813.

[15] Blackburn, N. J.; Rhames, F. C.; Ralle, M.; Jaron, S. J. Biol. Inorg. Chem. 2000, 5, 341.

[16] Boswell, J. S.; Reedy, B. J.; Kulathila, R.; Merkler, D.; Blackburn, N. J. Biochemistry 1996, 35, 12241.

[17] Jaron, S.; Blackburn, N. J. Biochemistry 2001, 40, 6867.

[18] Prigge, S. T.; Kolhekar, A. S.; Eipper, B. A.; Mains, R. E.; Amzel, L. M. Science 1997, 278, 1300.

[19] Prigge, S. T.; Kolhekar, A. S.; Eipper, B. A.; Mains, R. E.; Amzel, L. M. Nature Struct. Biol. 1999, 6, 976.

[20] Jaron, S.; Blackburn, N. J. Biochemistry 1999, 38, 15086.

[21] Prigge, S. T.; Mains, R. E.; Eipper, B. A.; Amzel, L. M. Cell. Mol. Life Sci. 2000, 57, 1236.

[22] Ljones, T.; Skotland, T. In Copper Proteins and Copper Enzymes; Lontie, R.; Ed.; CRC Press: Boca Raton, FL, 1984; p 131.

[23] Fitzpatrick, P. F.; Flory, D. R.; Villafranca, J. J. Biochemistry 1985, 24, 2108.

[24] Fitzpatrick, P. F.; Villafranca, J. J. J. Am. Chem. Soc. 1985, 107, 5022.

[25] Miller, S. M.; Klinman, J. P. Biochemistry 1985, 24, 2114.

[26] Wimalasena, K.; May, S. W. J. Am. Chem. Soc. 1989, 111, 2729.

[27] Stewart, L. C.; Klinman, J. P. Annu. Rev. Biochem. 1988, 57, 551.

[28] Chen, P.; Solomon, E. I. J. Am. Chem. Soc. 2004, 126, 4991.

[29] Kamachi, T.; Kihara, N.; Shiota, Y.; Yoshizawa, K. Inorg. Chem. 2005, 44, 4226.

[30] Tian, G.; Berry, J. A.; Klinman, J. P. Biochemistry 1994, 33, 226.

[31] Evans, J. P.; Ahn, K.; Klinman, J. P. J. Biol. Chem. 2003, 278, 49691.

[32] Prigge, S. T.; Eipper, B. A.; Mains, R. E.; Amzel, L. M. Science 2004, 304, 864.

[33] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. Jr.; Ferguson, D. M. Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J. Am. Chem. Soc. 1995, 117, 5179.

[34] Wang, J.; Cieplak, P.; Kollman, P. A. J. Comput. Chem. 2000, 21, 1049.

[35] Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. J. Phys. Chem. 1993, 97, 10269.

[36] Banci, L.; Curr. Op. Chem. Biol. 2003, 7; 143.

[37] Warshel, A.; Levitt, M. J. Mol. Biol. 1976, 103, 227.

[38] Zhang, X.; Harrison, D. H. T.; Cui, Q.; J. Am. Chem. Soc. 2002, 124, 14871.

[39] Ryde, U. Curr. Op. Chem. Biol. 2003, 7, 136.

[40] Ridder, L.; Harvey, J. N.; Rietjens, I. M. C. M.; Vervoort, J.; Mulholland, A. J.; J. Phys. Chem. B 2003, 107, 2118.

[41] Schöneboom, J. C.; Cohen, S.; Lin, H.; Shaik, S.; Thiel, W. J. Am. Chem. Soc. 2004, 126, 4017.

[42] Devi-Kesavan, L. S.; Gao, J. J. Am. Chem. Soc. 2003, 125, 1532.

[43] Soler, J. M.; Artacho, E.; Gale, J.; García A.; Junquera, J. Ordejón, P.; Sánchez-Portal D. J. Phys: Cond. Matt. 2002, 14, 2745.

[44] Troullier, N.; Martins, J. L. Phys. Rev. B 1991, 43, 1993.

[45] Louie, S. G.; Froyen, S.; Cohen, M. L. Phys. Rev. B 1982, 26, 1738.

[46] Perdew, J. P.; Burke, K.; Ernzerhof, M. Phys. Rev. Lett. 1996, 77, 3865.

[47] Crespo, A.; Scherlis, D. A.; Martí, M. A.; Ordejón, P.; Roitberg, A. E.; Estrin, D. A. J. Phys. Chem. B 2003, 107, 13728.

Chapter 6: Overview, Conclusions and Perspectives

In this chapter we would like to make a brief overview of the preceding chapters of the thesis, as well as a summary of the conclusions obtained for each of the biochemical reactions studied in each of them. Finally, we will present some general conclusions regarding the computer simulation of chemical reactivity of biological systems, as well as some future perspectives in this interesting field.

The aims of our work can by summarized as follows:

1) Developing a hybrid QM-MM methodology suited for describing chemical reactions in complex environments. In order to do so, we have coupled the efficient DFT based SIESTA code for describing the electronic structure of the active site together with the Amber classical force field parameterization for the rest of the environment. This particular objective was in fact quite ambitious, but as can be seen in our results it was successfully achieved. In Chapter 2 of this work, we have introduced the theoretical foundations of these kinds of methodologies, paying attention to QM and MM schemes and providing a complete description of hybrid QM-MM methodologies, with emphasis in the implementation of our SIESTA code.

2) Applying our QM-MM technique to study relevant biological systems. In order to validate our QM-MM code, we have started by performing calculations on the well-known Chorismate Murase system and in a second place; we have applied our scheme to the investigation of open problems in metalloproteins.

In Chapter 3 we have discussed the chorismate to prephenate conversion catalyzed by the *Bacillus subtilis* chorismate mutase. Our results suggests a correct validation of our hybrid code, because we have correctly predicted a catalytic activity in the enzyme environment reflected in an activation energy difference $\Delta(\Delta E^{\neq})$ of 8.5 kcal/mol with respect to the uncatalyzed reaction, which compares well with the 8 kcal/mol obtained experimentally. Furthermore, we have computed the free energy profile for this reaction obtaining a $\Delta G^{\neq}$ value of 7.7 kcal/mol and an entropic effect of -9.0 e.u. also in agreement with the experimental value of -9.1 e.u.

In Chapter 4 we have presented our results for the NO detoxification process performed by *Mycobacterium tuberculosis* truncated hemoglobin N (trHbN), which acts as a NO-dioxygenase defense mechanism. Our results suggest that the efficiency of the NO oxidation to nitrate depends on the entry of the reactants and on the heme capacity to catalyze the reaction. We have identified the opening of a ligand diffusion tunnel that is controlled by the Phe62 gate, forming a

complex network of conformational changes. Our results also suggest that the Tyr33-Gln58 pair is relevant to the physiological function because it contributes to modulate the O_2 and NO affinity and alters the essential dynamics of the protein, leading to a mechanism for ligand-induced regulation. Thus, O_2 binding to the heme should induce compression of Phe62 by the side chain of Gln58, in order to open the tunnel long branch, facilitating the capture of NO. This mechanism provides a basis to rationalize why NO conversion by oxy-trHbN occurs at a rate faster than O_2 binding to the deoxy protein. Finally, our QM-MM results indicate that the protein catalyzes the chemical reactions mainly by means of the heme group, with no significant contributions of the protein environment.

Finally, we have investigated the hydroxylation reaction catalyzed by peptidylglycine α-hydroxylating monooxygenase (PHM) in Chapter 5. Our results suggest that upon O_2 binding electron transfer occurs to form the Cu_B^I-O_2^- specie because of the more favorable binding energy value. Moreover, the substrate considerably increases the binding energy suggesting that O_2 is prone to coordination once the substrate is already placed in the enzymatic active site. We have also shown new insights into the molecular basis of mechanism active in PHM. We have also computed potential energy profiles for the H-abstraction reaction in order to identify the most reactive oxygenated specie active in PHM. We have concluded that neither Cu_B^{II}-O_2^- and Cu_B^I-O_2^- species, nor Cu_B^{II}-OOH can abstract the hydrogen atom from the substrate. Finally, we have evaluated if $[Cu_BO]^{+2}$ or the reduced $[Cu_BO]^{+1}$ species were able to promote the substrate H-abstraction, concluding that whereas the $[Cu_BO]^{+2}$ abstracts the HA atom concertedly with almost no activation energy, $[Cu_BO]^{+1}$ does it in a two-step manner with activation energies of about 5 kcal/mol each. These facts, allow us to suggest that the active species in PHM is $[Cu_BO]^{+2}$.

At this point it is interesting to comment some general issues regarding the application of computational methodologies to study chemical reactivity in biological processes. In the recent years, this approach has become a fundamental tool for elucidating enzymatic reactivity and for providing microscopic information hardly accessible by other techniques. In this context, quantum mechanics and molecular mechanics techniques have emerged as essential tools in chemical research and specifically simulations based in DFT schemes have been widely adopted by the scientific community. The methods of choice to describe chemical process that take place in complex and large systems are based on hybrid QM-MM techniques. The prime reason for this is that enzymatic active sites are limited to a small number of atoms in all enzymes. Thus, the region where the chemical events take place can be described by means of an expensive but efficient QM scheme, while the rest of the protein can be adequately represented by a MM force field parametrization. The use of these QM-MM methodologies for the investigation of chemical reactivity in solution and in enzymes has increased enormously in the past years and has shown to be a powerful tool as has been discussed in the present work. Under these considerations, it is

logical to think that computer simulation schemes, primarily QM-MM methods, are now in a position where they can significantly contribute to the elucidation and interpretation of the enzymatic chemical reactivity and allow us to describe protein behavior at a molecular level.

What will happen with computer simulation techniques in the near future? Nowadays, they are usually employed to complement experimental research, but it is not unlikely to guess that with improvements in methodology and faster computational tools, computational methods will become a fundamental technique for accessing to chemical mechanisms and microscopic detailed information that would not be possible to obtain by any other method. Finally, it is interesting to show the following parallelism: quantum mechanics was born at the beginnings of the last century, only a hundred years ago. The HF method was developed in the 30's and DFT in the 60's. However, computer simulations have only become an essential tool in chemical research in the last 20 years, due to the development of new and powerful computers. We believe that in the next decades there will be a breakthrough in the scientific community, which will move their goals to study more complex problems and to develop new and more accurate theoretical methods. It is not surprising that these models, called "multi-scale" models,[1-3] which are being used for bridging across disparate length and time scales in large scale numerical modeling and simulation of complex physical systems, will become the methods of choice in the near future. Thus, a scenario in which computer simulation techniques based in these models are applied to describe the behavior of an entire organelle or cell is, in our view, perfectly possible.

....

Darío Estrín

....

Alejandro Crespo

[1] Adam, J. A.; Bellomo, N. eds. Modelling and Simulation in Science, Engineering, and Technology, Birkhäuser, Boston, 1997.

[2] Araujo, R.P.; McElwain, D. L. S. Bull. Math. Biol., 2004, 66, 1039.

[3] Bentz, D. P.; Quenard, D. A.; Baroghel-Bouny, V.; Garboczi, E. J.; Jennings, H. M. Materials and Structures 1995, 28, 450.

Agradecimentos

Bueno, al fin llegué a los agradecimientos...parece mentira, pero fue lo primero que pensé cuando empecé a escribir la tesis, allá por Septiembre del 2005. Siendo hoy, 1° de Febrero, veré si puedo plasmar aqui y ahora lo que por tanto tiempo tuve en mi cabeza.

Primero empezaré por los agradeciemientos científicos para luego pasar a los personales... espero no olvidarme de nadie...

A Darío, por ser un "groso" como dicen por ahí...por ser una de las mejores personas para trabajar y un gran jefe, como pocos los hay...si lo habré hecho sufrir...gracias por todo, estuvo muy bueno haber trabajado con vos...

A Adrián, porque trabajar con él en USA es como hacerlo acá en Buenos Aires, gracias por bancarme y aguantarme las veces que estuve por alla....también fue muy lindo haber trabajado con vos todo éste tiempo...lástima que seas bostero, pero bueno....

A Marce, la verdad es que fue y es muy divertido trabajar con vos, hemos hecho una buena dupla todos estos años programando a los gritos...por algo nos llamaron "los hermanos macana"...

A Damián S., por haberme enseñado a recorrer los primeros pasos con el Siesta...

A Ojitos, que hoy no está, gracias por haberme traído al grupo allá por Junio del 99, y por haber compartido conmigo muchas horas de trabajo, pero más de boludeo...

A todos los chicos del grupo....a los más viejos...Nano, que siempre está para arreglar las problemas de todos y Damián B. siempre trabajando mucho y de buen humor...a los más jóvenes....Lula, Vero, Laura, Pablo, M. Laura, Lucas, Cristian, Leo, Ezequiel y Matias....y a las que ya se fueron, Vale y Dolores...

A los que estuvieron acá de pasada, Pablo D. (el uruguayo), Santi (el tucumano) y Susana (la catalana)...

A mi consejero de tesis, Ernesto Marceca con el cual nos divertimos durante toda la tesis..."a esa temperatura" y a sus muchachos...Marcos, que lo conozco desde ya hace mucho, a Gustavo, Germán y Alvaro...

A las chicas de Fabio: Flor, Nati y cia...y a Fabio....

A todos en en DQIAQF, en particular Pablo H, Andi, Leo, Maru, Pocha y cia...y a Pablo E.

A "los chicos", que me bancaron en su casa siempre que fui a USA...la verdad es que no hubiera sido lo mismo sin uds, y además también los conozco desde hace mucho, cuando empezamos la carrera...gracias...

 A Javíer Luque y Pablo Ordejón en Barcelona por haberme tratado muy bien cuando fui...estuvo muy copado haber ido a trabajar con uds...

A nuestros colegas experimentales Mario Amzel y Martino Bolognesi por habernos sugerido problemas interesantes y por sus fructíferas discusiones.

A los chicos de "Conexión" por haberme divertido mucho estos 2 últimos años con uds en las comisiones de la facu...

Al DQIAQF, CONICET, Agencia y F. Antorchas por haber aportado la plata todos estos años....

A la UBA porque sigue siendo una de las mejores universidades de América y encima es pública y gratuita...

y ahora quiero pasar a los agradeciemtos personales...

A Agus, en este año tan importante para nosotros, que me bancó estos 4 años de tesis...y se aguantó mis charlas del Siesta....gracias...espero estar con vos mucho años más...

A mis viejos, esta tesis también es para uds, gracias por todo...

A toda mi familia, y a mis amigos: Pablo Crespo (hay muchos Pablos), el enano, Fede, Taco, Lucho, Rodri, Leo, Rafa, Juanma, Massi, Maxi, el Tano, quiqui...y a todos los demás....

Al Gato, por haber ganado Roland Garros en el 2004 y al Rojo, por que es una pasión...

Gracias a todos...Gracias...Totales....