DEVELOPMENT AND APPLICATIONS OF

INFRARED STRUCTURAL BIOLOGY

By

ZHOUYANG KANG

Bachelor of Science in Materials Physics
Nanjing University
Nanjing, China
2007

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2014

DEVELOPMENT AND APPLICATIONS OF

INFRARED STRUCTURAL BIOLOGY

Dissertation Approved:

Dr. Aihua Xie

Dissertation Adviser

Dr. John Mintmire

Dr. Donghua Zhou

Dr. Wouter D. Hoff

Dr. José Soulages

.

ACKNOWLEDGEMENTS

It has been quite a journey for me towards the completion of this Ph.D. thesis, which likely marks the end of my education. During the past few years, the numerous encouragement, support and love I received are the reasons that I am able to come to this far. I would like to take this opportunity to express my gratitude towards them.

I would like to first thank my family members. Thank my parents for their love, their constant support and motivation throughout my life and my education. I would like to express my deep appreciation to my wife Jing Zhu has accompanied and encouraged me all along these years. Especially I would like to thank my aunt, now in paradise, who supported and educated me for so many years like her own son. Without her I would never have this chance to pursue this Ph.D. degree in United States. May she rest in peace.

I would like to thank my friends, who have encouraged, entertained, cajoled, supported me through the dark times, celebrated with me through the good, who have been brilliant and understanding when I needed them to be, thank you.

I would like to express my thankfulness to all the colleagues and the collaborators I've been working with during these years: Anupama Thubagere and Edward Manda, Yunxing Li, Shuo Dai, Ningning Xu, Dr. Sandip Kaledhonkar, Dr. Johnny Hendriks, Dr. Lorand Kelemen, Dr. Jie Pan, Dr. Masato Kumauchi, Miwa Hara and Rachana Rathod

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

for their help in numerous lab activities and scientific discussions. Without them, there is no way I can obtain the knowledge of that many scientific skills.

I would like to thank collaborators of our groups that had supported my project. Thank Dr. Dana Brunson, Dr. R. J. Hauenstein for teaching me computational skills of using UNIX system. Thank Dr. Chris Wood and Dr. Junpeng Deng for their help when I was learning CCP4 software. Thank Dr. Richard L. Martin from Los Alamos National Lab for his advice and tips on DFT calculations.

I thank all staff members of Physics Department, for their hard work and support during my stay in OSU: Susan, Cindy, Tamara, Sandra, Melissa, Warren, Charles and Mike.

I would like to thank Dr. John Mintmire, Dr. Donghua Zhou and Dr. Jose Soulages for serving on my committee and for reviewing my dissertation. Their advice during previous committee meetings, my qualify exam and my thesis defense had helped me a lot. I'm really honored to have great scientists like them to serve on my committee.

I would like to extend my special thanks to Dr. Wouter Hoff, not only for serving on my committee, but also for all the support, advice and inspiration of science from him. I will be so grateful to have the chance of working with such a talent scientist.

Finally, I would like to express my greatest thanks to Dr. Aihua Xie for being my Ph.D. advisor. She is a thoughtful, intelligent and professional scientist. Without her complete support, guidance, advice and her encouragement, I would not complete my Ph.D. degree and obtain a broad knowledge in that many areas.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: ZHOUYANG KANG

Date of Degree: MAY, 2014

Title of Study: DEVELOPMENT AND APPLICATIONS OF INFRARED
STRUCTURAL BIOLOGY

Major Field: PHYSICS

Abstract:

Aspartic acid (Asp), Glutamic acid (Glu) and Tyrosine (Tyr) often play critical roles at the active sites of proteins. Probing the structural dynamics of functionally important Asp/Glu and Tyr provides crucial information for protein functionality. Time-resolved infrared structural biology offers strong advantages for its high structural sensitivity and broad dynamic range (picosecond to kilosecond). In order to connect the vibrational frequencies to specific structures of COO- groups and phenolic –OH groups, such as the number, type, and geometry of hydrogen bond interactions, we develop two sets of vibrational structural markers (VSM), built on the symmetric and asymmetric stretching frequencies for COO- and C-O stretching and C-O-H bending frequencies for phenolic -OH. Extensive quantum physics (density functional theory) based computational studies, combined with site-specific isotope labeling as well as site-directed mutagenesis, and experimental FTIR data on Asp/Glu in proteins, are used to establish a unique correlation between the vibrations and multiple types of hydrogen bonding interactions. Development of those vibrational structural markers significantly enhances the power of time-resolved infrared structural biology for the study of functionally important structural dynamics of COO- from Asp/Glu and phenolic –OH from Tyr residues in proteins, including rhodopsin for biological signaling, bacteriorhodopsin and PYP for proton transfer, photosystem II for energy transformation, and HIV protease for enzymatic catalysis. Furthermore, this approach can be adopted in the future development of vibrational structural markers for other functionally important amino acid residues in proteins, such as arginine (Arg), histidine (His), and serine (Ser).

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

.

CHAPTER 1

Introduction

1.1 Structural Biology of Proteins

Protein is the essential component of the biological world we are living in. Virtually all of the biological phenomenon involve the effect of protein functions. As a molecule, protein is a long chain polypeptide of amino acids folding in a particular conformation. Thus the best way to study how protein function affect biological phenomenon is to "see" the protein structures, because protein functions are determined by their structures. Scientists' effort of detecting the structure of a biological macromolecule began almost 400 years ago, when Robert Hooke and Antoni van Leeuwenhoek introduced light microscope to biological world, which made watching small biological objects such as bacteria, plant cells become possible.

But light microscope is not suitable to identify structures of the proteins since the sizes of the proteins are smaller than the diffraction limits of light microscope (~200 nm). In order to view the details of proteins, technologies with capabilities of achieving atomic level resolution need to be used. Since mid-20th century, high-resolution structural biology techniques started to become available, and were rapidly developed in recent years. As a result, huge numbers of protein structures in static state have been reported (nearly 100,000 PDB structures available in the PDB databank). Among the published PDBs, X-ray crystallography, solution NMR spectroscopy and electron microscopy contributed the most structures, with 88.4%, 10.6% and 0.7% of all structures are found using those techniques, respectively.

However, it remains challenging to use X-ray crystallography for solving protein structures at intermediate states because of its limitation for time-resolved applications. In addition, even for protein in static state, it is difficult to detect proton positions with X-ray crystallography. More importantly, proteins samples are in crystalline state, while in nature most proteins are functioning in solutions.

Comparing to X-ray crystallography, NMR spectroscopy provides much broader range of time-scale (picosecond to days, depends on techniques), and protein samples are in solution. However, analyzing large systems using NMR spectroscopy is a challenging task. Furthermore, during protein function NMR spectroscopy is more sensitive to protein conformational changes than detailed structural changes such as proton transfer, electron transfer and hydrogen-bonding interaction changes.

Therefore, in order to provide key structural details that are missing in the protein structures solved by X-ray crystallography and NMR spectroscopy, other structural biology techniques are needed.

1.2 Vibrational Spectroscopy

Atoms are oscillating. From a classical point of view, for a two-body system composed of mass m_1 and m_2 (i.e. a diatomic molecule), the restoring force F_{12} between the two masses is given by Hooke's Law:

$$F_{12} = -k\Delta x_{12} \qquad \text{(Equation 1.1)}$$

where k is the force constant and Δx_{12} is the displacement from the equilibrium position. Then the potential energy E_V of the system can be derived as:

$$E_V = \frac{1}{2}k\Delta x_{12}{}^2 \qquad \text{(Equation 1.2)}$$

And the kinetic energy of the oscillation E_T:

$$E_T = \frac{1}{2}\mu(\Delta\dot{x}_{12})^2 \qquad \text{(Equation 1.3)}$$

μ is the reduced mass of the two atoms which equals $\frac{m_1 m_2}{m_1 + m_2}$.

Total energy of the system E = E_V + E_T is conserved because of the conservation law. Therefore, the first derivative of total energy E is zero:

$$\begin{aligned}\frac{dE}{dt} &= \frac{dE_V}{dt} + \frac{dE_T}{dt} \\ &= \frac{1}{2}k\frac{d(\Delta x_{12})^2}{dt} + \frac{1}{2}\mu\frac{d(\Delta\dot{x}_{12})^2}{dt} \qquad \text{(Equation 1.4)} \\ &= 0\end{aligned}$$

And Equation 1.4 can be derived to the differential equation for a harmonic motion:

$$\frac{k}{\mu}\Delta x_{12} + \frac{d^2\Delta x_{12}}{dt^2} = 0 \qquad \text{(Equation 1.5)}$$

The solution of Equation 1.5 is

$$\Delta x_{12} = A\cos(\omega \mathrm{t} + \varphi) \qquad \text{(Equation 1.6)}$$

where A is the amplitude, ω is the circular frequency and φ is the phase, and $\omega = \sqrt{\frac{k}{\mu}}$. Thus the vibrational frequency can be presented as:

$$\upsilon = \frac{\omega}{2\pi} = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}} \qquad \text{(Equation 1.7)}$$

A non-linear molecule of N atoms such as protein has 3N-6 vibrational modes. For each mode all atoms are vibrating at a certain frequency, although for many modes only a few atoms will have large displacement while the rest atoms almost remain stationary. This frequency is not only affected by the atoms of the particular group in the molecule, but also the local environment around, meaning every molecule, and even localized functional groups within the molecule can be characterized by its unique vibrational frequencies (Siebert and Hildebrandt, 2008, Griffiths and De Haseth, 2007).

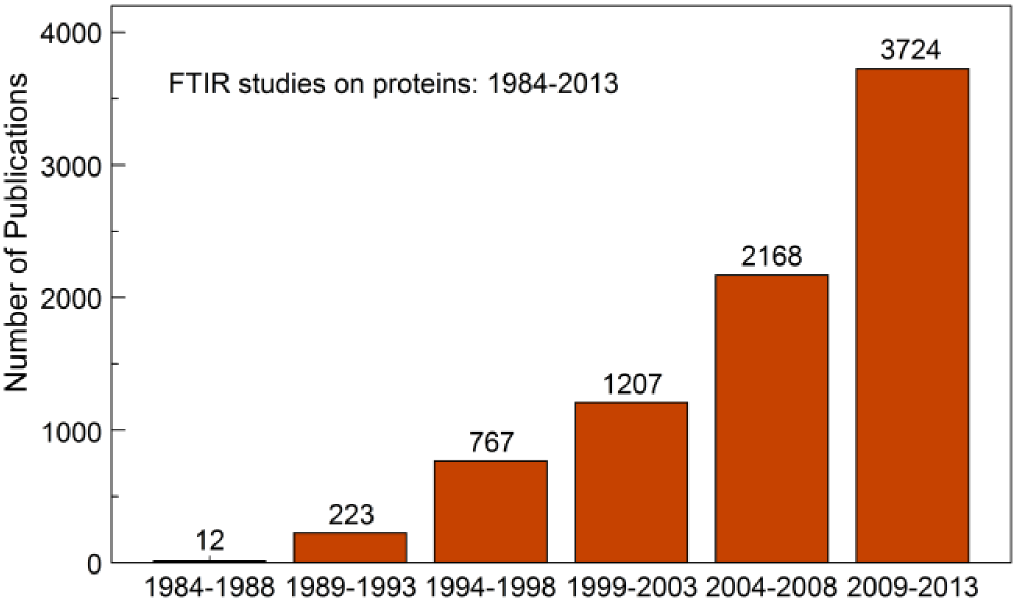


Figure 1.1 Number of publications related to “FTIR” and “Protein” by year.

Assume all oscillation are harmonic, then the vibrational energy states E_{Vi} can be described as

$$E_{Vi} = h\upsilon_i(n+\frac{1}{2}) \quad \text{(Equation 1.8)}$$

where h is the Planck constant, υ_i is the fundamental frequency of the vibration and n is the vibrational quantum number of the *i*th mode (n=0,1,2...). Figure 1.2 shows the regions of electromagnetic spectrum which are used mostly for spectroscopy purposes, wavelength ranging from 10^{-3} nm to 10^{6} nm and the corresponding frequencies ranges in terms of wavenumber from 10^{7} cm^{-1} to 10^{-2} cm^{-1}. The transition energy $\Delta E=h\nu_i$ between the ground state (υ_i=0) and the first excitation state (υ_i=1) of most vibrational modes corresponds to the vibrational frequencies in the mid-infrared region of 4000 cm^{-1} to 400 cm^{-1} (Griffiths and De Haseth, 2007), meaning the spectra of this region will give the most structural information for proteins, such as protonation state, hydrogen-bonding interactions and secondary structure formations. In fact, the number of reported infrared studies on proteins has grown exponentially in recent years, as figure xx shows. However, protein are normally large molecules. Infrared spectrum of a protein is so information rich that all groups within the protein molecule are presented in a narrow region, making it difficult to distinguish signals belong to the desired residues. And even with the signals identified, extracting detailed structural information out remains challenging due to the lack of reported guidelines.

Therefore, the development of Infrared Structural Biology is aiming at two goals. First goal is to find a way to identify infrared signals of a particular group from the infrared spectrum of a whole molecule. And the second goal is to develop the vibrational structural markers for single amino acid residues that can be used to interpret the specific infrared signals to detailed structural information.

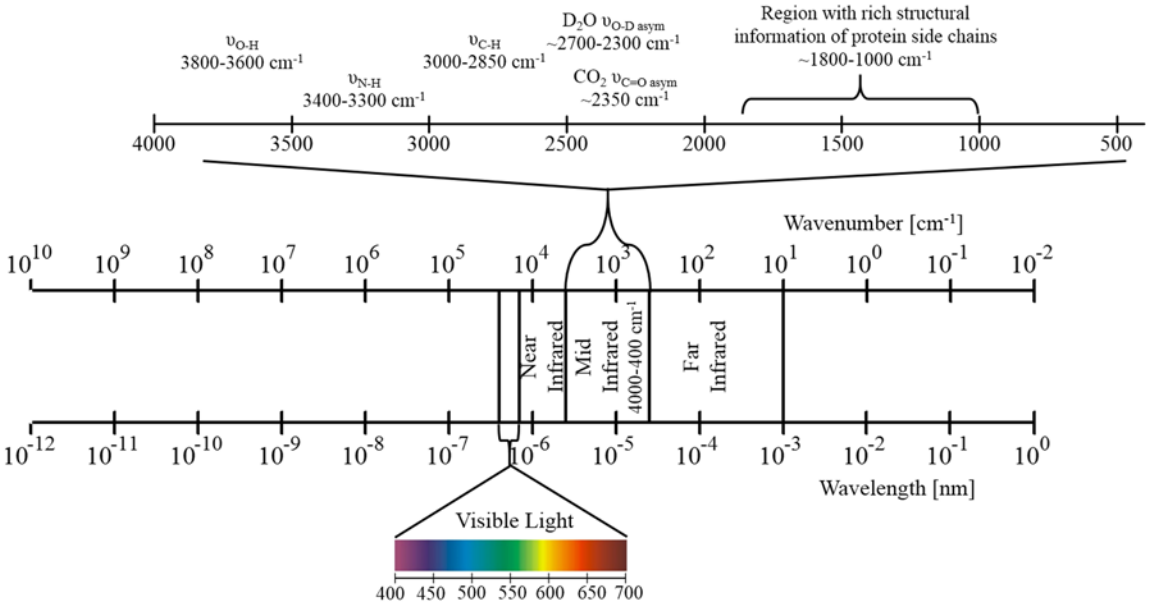


Figure 1.2 Electromagnetic spectrum with focus on infrared region

1.3 Density Functional Theory and Quantum Chemistry Calculation using Gaussian

Classical mechanics, the laws of motion for macroscopic objects, was first discovered in 17th century by Sir Isaac Newton. However, physicists found it difficult to use classical mechanics to explain the behavior of small particles such as nuclei and electrons, until the development of quantum mechanics, which provides an accurate description of laws of motion at atomic and sub-atomic scales.

The core of quantum mechanics is the Schrödinger Equation, which comes in a time-dependent form

$$E\Psi\left(\vec{r}\right)=H\Psi\left(\vec{r}\right) \qquad \text{(Equation 1.9)}$$

where E is the eigenenergy of the particles, H is the Hamiltonian operator, and Ψ is the eigenwavefunction (Shankar, 2012).

The Hamiltonian of a complex system is consisted of kinetic energy of electrons and nuclei and potential energy from their interactions, which is given by:

$$H=\hat{T}_{elec}+\hat{T}_{nuc}+\hat{V}_{Ne}+\hat{V}_{ee}+\hat{V}_{NN} \qquad \text{(Equation 1.10)}$$

where $\hat{T}_{elec}$ and $\hat{T}_{nuc}$ are kinetic energy of electrons and nuclei, and $\hat{V}_{Ne}$, $\hat{V}_{ee}$ and $\hat{V}_{NN}$ are potential energy from interactions between nuclei-electron, electron-electron and nuclei-nuclei, respectively.

For a system that has N electrons, K nuclei and carries charge Z_n, those components of its Hamiltonian can be described by

$$\hat{T}_{elec} = -\frac{\hbar^2}{2m_e}\sum_{i=1}^{N}\nabla_i^2 \qquad \text{(Equation 1.11)}$$

$$\hat{T}_{nuc} = -\frac{\hbar^2}{2M}\sum_{A=1}^{K}\nabla_A^2 \qquad \text{(Equation 1.12)}$$

$$\hat{V}_{Ne} = -\sum_{i=1}^{N}\sum_{A=1}^{K}\frac{Z_A e^2}{4\pi\varepsilon_0 r_{iA}} \qquad \text{(Equation 1.13)}$$

$$\hat{V}_{ee} = \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{e^2}{4\pi\varepsilon_0 r_{ij}} \qquad \text{(Equation 1.14)}$$

$$\hat{V}_{NN} = \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B e^2}{4\pi\varepsilon_0 R_{AB}} \qquad \text{(Equation 1.15)}$$

In the equation, index i refers to the electrons and n refers to the nuclei, m is the mass of electron and M are the mass of the different nuclei.

The solution of the Schrödinger Equation leads to the complete understanding all properties of a molecular system, as Dirac stated:

"The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble."

For complicate systems, solving exact solution for the Schrödinger Equation takes way too much computational power which is not affordable. Thus, a good approximation is necessary in order to simplify the Schrödinger Equation.

Since there is significant mass difference between electrons and nuclei, that even the lightest nucleus, ^{1}H carries over 1837 times of mass of an electron. Therefore, nucleus moves much slower than electrons. With Born-Oppenheimer approximation, electrons can be considered as moving in the field of fixed nuclei. As a result, fixed nuclei will not carry any kinetic energy and the potential energy of nuclei-nuclei interactions can be treated as a constant. The Hamiltonian then can be written as a reduced form (or called electronic Hamiltonian):

$$\begin{aligned}\hat{H}_{elec} &= \hat{T}_{elec} + \hat{V}_{Ne} + \hat{V}_{ee} \\ &= -\frac{\hbar^2}{2m_e}\sum_{i=1}^{N}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{K}\frac{Z_A e^2}{4\pi\varepsilon_0 r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{e^2}{4\pi\varepsilon_0 r_{ij}}\end{aligned} \quad \text{(Equation 1.16)}$$

And the Schrödinger Equation of the electrons is

$$\hat{H}_{elec}\Psi_{elec}\left(\vec{r}_1,\vec{r}_2...\vec{r}_N\right) = \hat{V}_{eff}(\vec{R})\Psi_{elec}\left(\vec{r}_1,\vec{r}_2...\vec{r}_N\right) \quad \text{(Equation 1.17)}$$

where Ψ_{elec} is the electronic wave function and the solution to the Schrödinger Equation, and solving this equation produces $\hat{V}_{eff}(\vec{R})$, which is the effective nuclear potential function. This potential depends on the nuclear coordinates and describes the potential energy surface of the system. Accordingly, with the use of effective nuclear potential, nuclear Hamiltonian can be written as

$$\hat{H}_{nuc} = \hat{T}_{nuc}(\vec{R}) + \hat{V}_{NN}(\vec{R}) + \hat{V}_{eff}(\vec{R}) \quad \text{(Equation 1.18)}$$

Although the Hamiltonian is much more simplified to the form as Equation 1.10, the second term is still intractable because it carries the interactions between electrons. Thus, an approximation for the eigenfunctions of the Hamiltonian is needed.

Hartree-Fock theory or Self-consistent Field method is a basic approach for this purpose. It was originally proposed by Douglas Hartree and further modified by Vladimir Fock, which help approximating the wave functions of a multi-electron atom (Bergethon, 1998).

Hartree-Fock theory is made of two approximations. The primary approximation is the central field approximation, in which Coulombic electron-electron repulsion is taken into account by integrating the repulsion term. This gives the average effect of the repulsion, but not the explicit repulsion interaction. This approximation treats the Schrödinger Equation of a multi-electron system as a combination of many single-electron Schrödinger equations, which yield to a series of single-electron wave functions (orbital) that describe the behavior of each electron individually in the net field generated by all other electrons (Young, 2001). The second approximation expresses those single-electron wave functions as a pre-defined set of basis functions which are Gaussian-type orbitals (GTO), and then linearly combine them (Bergethon, 1998, Young, 2001). The general form of Gaussian-type orbitals or Gaussian functions is:

$$g(\alpha,\vec{r}) = cx^n y^m z^l e^{-\alpha r^2} \qquad \text{(Equation 1.19)}$$

where $\vec{r}$ is composed of x, y, z, and n, l, m are integers. And constant α determines the size of the function (Foresman and Frisch, 1996).

Another approach is density functional theory (DFT)-based methods, which is a method derived from the fundamental laws of quantum mechanics. It treats the electrostatic energy term as a function of electron density instead of wave function. The origin of DFT was Hoenburg and Kohn's work published in 1964, that the ground state energy and density can be determined exactly by a unique functional (Hohenberg and Kohn, 1964), although the form such functional was not provided in their theorem.

Following their work, the approximate functionals employed by current DFT methods partition the electronic energy into four terms:

$$E = E^{T} + E^{V} + E^{J} + E^{XC} \qquad \text{(Equation 1.20)}$$

where E^T is the kinetic energy term, E^V describes the potential energy term of nuclear-electron and nuclear-nuclear interactions, E^J is the electron-electron Coulomb repulsion term, and E^{XC} is the exchange-correlation term and carries the remaining part of the electron-electron interactions, including the exchange energy arising from the antisymmetry of the quantum mechanical wave function and dynamic correlation in the motions of the individual electrons. All terms except nuclear-nuclear repulsion are functions of the electron density ρ (Foresman and Frisch, 1996).

In this thesis the method used for the calculations is B3LYP, a hybrid method of Hartree-Fock exchange and DFT exchange-correlation functionals. B3 refers to Becke's three parameters used in the exchange term E^{XC} and LYP refers to Lee, Yang and Parr's work in developing the correlation part of the functional. So far it is the most widely used hybrid method because of its superior to the other defined traditional functionals (Foresman and Frisch, 1996, Becke, 1993, Lee et al, 1988).

In this thesis all *ab initio* calculations were carried out using Gaussian 09, the latest version of the commercial computational chemistry software Gaussian series. It has the capabilities of performing *ab initio* calculations for geometry optimization, system energy and vibrational frequencies.

1.4 Photoactive Yellow Protein

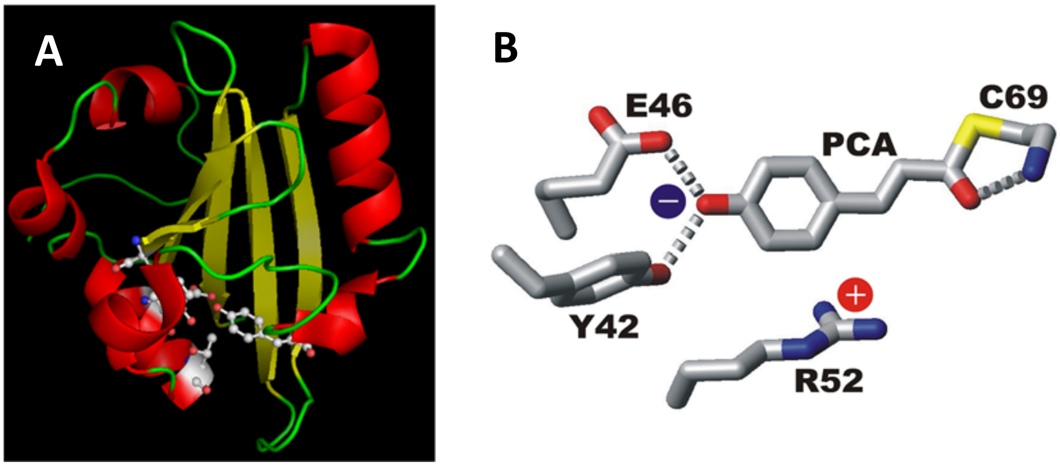


Figure 1.3 (A) crystal structure of Hhal PYP (B) active site of Hhal PYP

Photoactive yellow protein (PYP), a bacterial blue-light photoreceptor with molecular weight of 14K Da., was first discovered by Meyer in 1985 in extremely halophilic anoxygenic purple photosynthetic bacterium *Ectothiorhpdopsira halophila* (whose name was later changed to *Halorhodospira halophila* or *Hhal halophila*) (Meyer, 1985), where it functions as a photoreceptor for negative phototaxis (Sprenger et al, 1993). A yellow-colored protein, with a peak absorption at 446 nm in the blue light region as Figure 1.4 shows, the chromophore of PYP was initially proposed to be flavin (Meyer, 1985) or retinal (McRee et al, 1986). However, it was found in 1994 that the chromophore of PYP is a p-coumaric acid (pCA) covalently linked to Cys69 via a thioester bond (Hoff et al, 1994, Baca et al, 1994), shown in Figure 1.3.

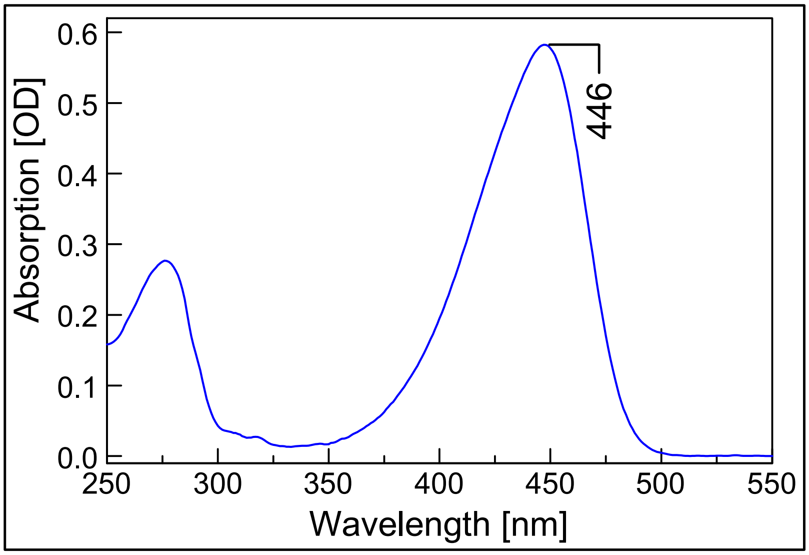


Figure 1.4 UV-Vis absorption of wild-type Hhal Photoactive Yellow Protein

The extensive crystallography studies of PYP later revealed more details of the PYP structure. First crystal structure of Hhal PYP was found in 1995 (Borgstahl et al, 1995) and the highest resolution crystal structure was reported in 2003 at 0.85 Å by Getzoff and coworkers (Getzoff et al, 2003). The obtained structure (Figure 1.3) shows that Hhal PYP, a 125-amino-acids protein, has 5 α-helix, 1 3_{10}-helix, 1 π-helix and 6 β-strands. In the active site of PYP, a negatively charged pCA chromophore is interacted with residues Glu46 and Tyr42 via ionic hydrogen bonds, with HB length found to be 2.58 Å and 2.48 Å respectively (Getzoff et al, 2003).

PYP undergoes a photocycle upon by its pCA chromophore receiving a blue-light photon (Meyer et al, 1987), as described in Figure 1.5 (Xie et al, 2001). The photocycle can be simplified to ground state pG_{446} and intermediate states pR_{465}, pB'_{355} and pB_{355}. The beginning of the photocycle is the pG_{446} to pR_{465} transition, where *trans* to *cis* photo-isomerization occurs on the negatively charged pCA chromophore. The formation of pR465 state increases the proton affinities of the phenolic oxygen on the pCA chromophore dramatically and results in absorbing a proton from the Glu46 and forms state pB'_{355} (Xie et al, 2001, Xie et al, 1996). After losing the proton, the negative charge of Glu46 in the hydrophobic PYP active site becomes energetically unstable. As a result, large conformational changes happen and eventually lead to the formation of signaling state pB_{355} (Xie et al, 2001, Sprenger et al, 1993, Xie et al, 1996, Hoff et al, 1999). Then PYP will recover from pB_{355} to pG_{446} state in 350 ms to complete the photocycle.

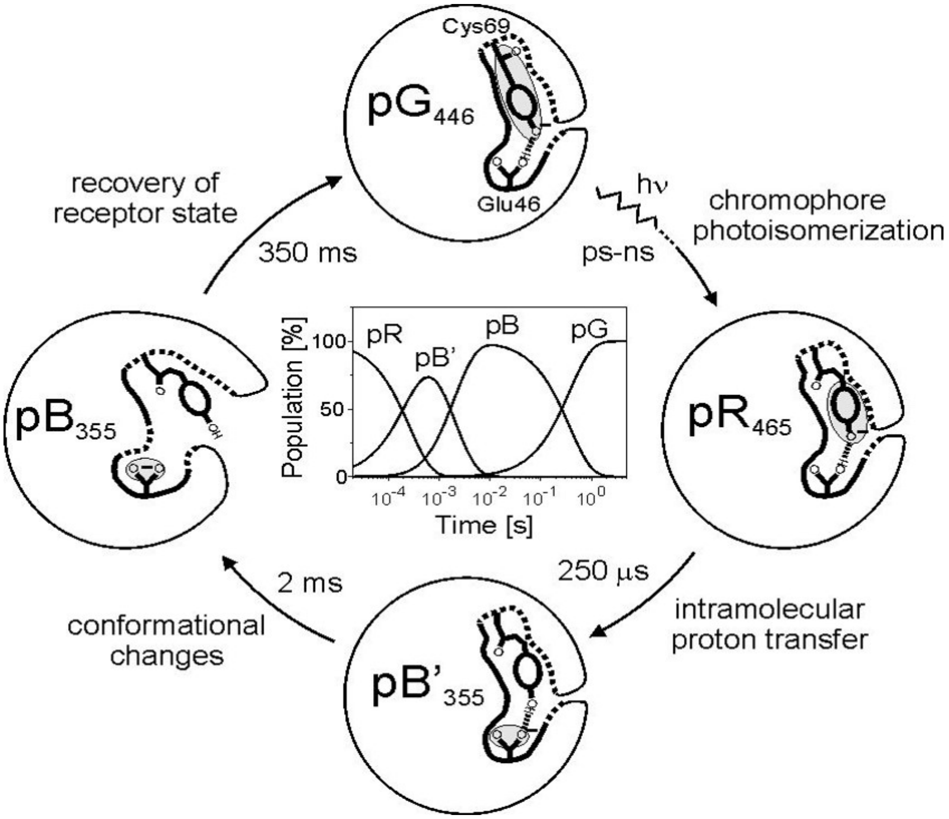


Figure 1.5 Photocycle of PYP, reported by Xie (Xie et al, 2001)

As a small highly soluble protein, PYP is an excellent model system to study key structural related biological functions such as proton transfer, hydrogen bond interactions and protein folding. Infrared spectroscopy technique has contributed significantly for solving many important functionally important structural problems for PYP. For example, with cryogenic infrared spectroscopy, it was found that Glu46 is the proton donor for the protonation of pCA chromophore during the PYP photocycle (Xie et al, 1996). And the deprotonation of Glu46 leads to the formation of energetically unstable state pB'$_{355}$ and the large-protein quake afterwards was first detected using step-scan time-resolved infrared spectroscopy with 10μs time resolution (Xie et al, 2001).

1.5 Hydrogen-Bonding Interactions

Hydrogen-bonding interaction, by short definition most recently recommended by IUPAC, is "**An attractive interaction between a hydrogen atom from a molecule or a molecular fragment X-H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation**." (Arunan et al, 2011)

As it stated, hydrogen-bond donor a molecular fragment which has a hydrogen atom covalently bonded to another electronegative atom. Hydrogen-bond acceptor is a molecular fragment carries an electronegative atom, mostly oxygen or nitrogen with a lone pair of electrons. The length hydrogen-bonding interactions is the distance between the electronegative atom from hydrogen-bond donor and acceptor. It ranges from 2.7 Å to 3.2 Å for normal neutral hydrogen-bonding interactions (Arunan et al, 2011). However, this distance can be shorter for strong ionic hydrogen-bonding interactions, or it can be longer for weak hydrogen-bonding interactions. The hydrogen-bond angle is in a range of 180°±20°.

In proteins, hydrogen-bonding interactions are important structural elements, particularly those formed between buried polar groups. (Pauling et al, 1951, Takano et al, 2003). From statistical studies Worth reported that buried polar side chain which are hydrogen-bonded are more conserved than those side chains that are not (Worth and Blundell, 2009). Depending on the microenvironment around, contributions to protein stability from each individual hydrogen-bonding interactions are different. The loss of exposed, ordinary hydrogen-bonding interactions is reported to increase an energy of approximately 2~10 kJ/mol, while the loss of some key buried hydrogen-bonding interactions could greatly increase the energy, approximately 10-40 kJ/mol (Pace, 2001, Pace et al, 2014, Pace, 2009). Meanwhile, protein folding energy is reported to be approximately 20-40 kJ/mol (Creighton, 1993). Therefore, the dissociation energy of hydrogen-bonding interactions and the protein folding energy are similar on the order that breaking of hydrogen-bonding interactions is capable of significantly destabilizing the protein folding.

Despite the critical roles hydrogen-bonding interactions play in biological world, few studies have been reported to reveal the time-resolved detection of changes on hydrogen bonding interactions during protein function. Although infrared spectroscopy has its advantage in high sensitivity to hydrogen-bonding interactions, very few guidelines which can link the peaks observed in infrared absorption to the specific hydrogen-bonding interaction. Therefore, our work of developing Vibrational Structural Markers for probing hydrogen-bonding interactions on buried deprotonated Asp/Glu residues and buried Tyr residues, as I will present in this thesis, are urgently needed for the field of Infrared Structural Biology.

CHAPTER 2

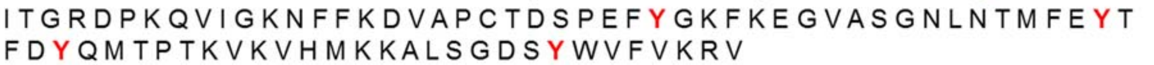
Materials and Methods

Most details of the materials and methods related to this thesis will be discussed in the individual chapters (Chapter 3-5). The additional information are summarized in this chapter.

2.1 Site-directed Mutagenesis of Tyrosine in *Halorhodospira halophila* Photoactive Yellow Protein (Hhal PYP)

PYP from *Hhal* has a low yield ~1mg per litter of *Halorhodospira halophila* cell culture (Hoff, 1995). Therefore, in order to obtain a better yield, all PYPs including wild-type and genetically engineered proteins used in this thesis were over-expressed by *Escherichia coli (E. coli)* in LB Broth for non-labeled proteins or in M9 Media for isotopic labeled proteins (Rathod et al, 2012, Kaledhonkar, 2013).

The *Hhal* PYP has 125 amino acids. Its amino acid sequence and the corresponding genetic sequence is shown in Figure 2.1. In this thesis four tyrosine residues (Codon: TAC) of PYP were targeted for mutation to phenylalanine (Codon: TTC) individually: Y76F, Y94F, Y98F and Y118F. The site-specific mutagenesis of PYP were performed following protocol B.1 in Appendix.



MEHVAFGSEDIENTLAKMDDGQLDGLAFGAIQLDGDGNILQYNAAEGD
ITGRDPKQVIGKNFFKDVAPCTDSPEFYGKFKEGVASGNLNTMFEYT
FDYQMTPTKVKVHMKKALSGDSYWVFVKRV

Figure 2.1 Wild-type PYP peptide sequence (Baca et al, 1994) . Position of Tyr42 is labeled green. Positions of Tyr76, Tyr94, Tyr98 and Tyr118 are labeled red.

The Codon switches from TAC to TTC within PYP strain were achieved by PCR overlap extension, as Figure 2.2 shows (Ho et al, 1989). For each individual mutant, primers contain wild-type and mutant codon information for both directions (5’->3’ and 3’->5’) were purchased from Invitrogen. Plasmid pET-16b with wild-type PYP sequence inserted were used as template for the PCR reaction. *GoTaq* Flexi DNA Polymerase purchased from Promega was used as the DNA polymerase for the reaction. After each step of PCR, the product was purified using QIAquick PCR purification kit from Qiagen.

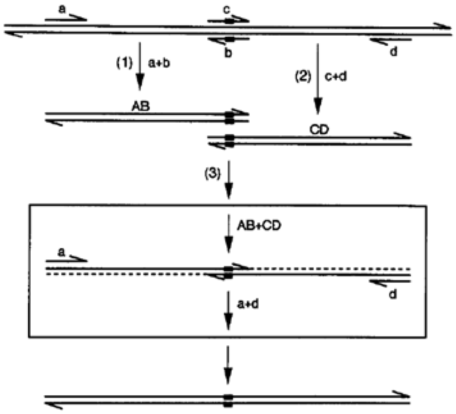


Figure 2.2 Principles of PCR overlap extension

Plasmid pET-16b was double digested at NcoI and BamHI site using NcoI-HF and BamHI-HF Enzyme purchased from New England Biolabs, then the PYP strain with mutant sequence from two-step PCR were ligated with the double digested pET-16b plasmid using T4 DNA Ligase Kit from Promega. The mixture solution of ligased plasmid was then transformed into NEB 5-alpha Competent *E. coli* purchased from New England Biolabs, incubated in SOC outgrowth medium for an hour and on agar plate overnight. Positive colonies were selected next morning for growth in 5ml individual LB Broth media with antibiotics, the plasmid was then extracted from the harvest cells using QIAprep Spin Miniprep Kit from Qiagen and submitted for DNA sequencing at OSU Biochemistry Core Facility. Plasmids with positive sequencing results were transformed to BL21 (DE3) Competent *E. coli* from New England Biolabs for protein overexpression.

2.2 Density Functional Theory Calculation

DFT calculations were carried out either using Gaussian 03 or Gaussian 09 Unix in OSU High Performance Computing Center (OSUHPCC), and most of the jobs were submitted the Cowboy cluster. Cowboy cluster is the newest and the fastest supercomputer OSU acquired. Each submitted Gaussian job occupied a 12-CPU node of the Cowboy, and used 4GB of memory for the calculation. Gaussview 5.0 was used to generate input structures and read output structures with vibrational mode frequencies. General procedure of the calculation include structural optimization, energy and frequency calculations. All DFT calculation were performed with B3LYP method, while details of the selection of the basis sets will be discussed in individual chapters. Protocols of generating input files and submitting jobs for DFT calculation can be found in Appendix.

2.3 FTIR measurement

2.3.1 Modern FTIR spectrometer

The core of modern infrared spectroscopy is a design of two-beam interferometer originally by Michelson. It is a device made of a light source, a beam splitter which divides a light beam into two paths, and a pair of mirrors with one of which is fixed and the other one is movable. Figure 2.3 shows the basic design of a Michelson Interferometer. Broadband IR light of radiation is emitted from the source, it gets divided equally by the beam splitter. Half of the light is reflected to the fixed mirror and the other half transmitted to the movable mirror without changing its direction. The two light beams after beam splitter are perpendicular to each other. Both beams will be reflected by the mirror, then combined together at the beam splitter again and go to the detector.

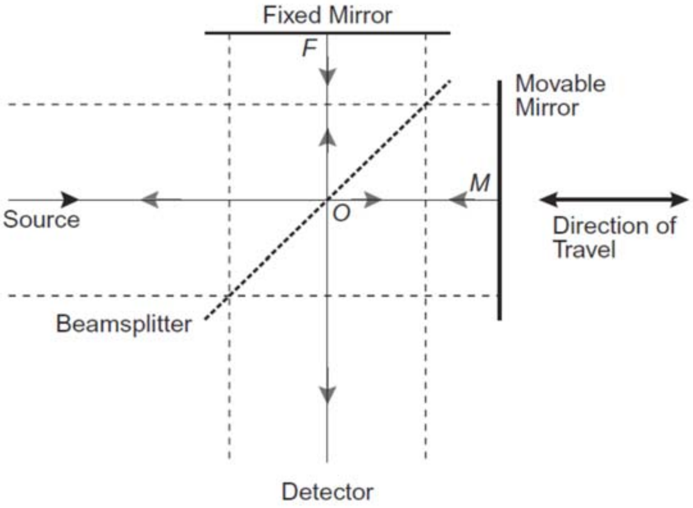


Figure 2.3 Basic design of Michelson Interferometer

Since the movable mirror is keep moving, a difference in optical path is generated. This optical path difference is named retardation with symbol δ. The intensity of the light that eventually reaches detector is retardation dependent and can be written as:

$$I'(\delta) = 0.5I(\tilde{v}_0)(1+\cos 2\pi \frac{\delta}{\lambda_0}) \qquad \text{(Equation 2.1)}$$

where *I'(δ)* is the intensity of beam measured at detector, *I(ṽ₀)* is the intensity of the source and λ_0 is the wavelength of the radiation light. Therefore, when $\delta = n\ \lambda_0$ (n is an integer, equals to 0, 1, 2...) *I'(δ)* shall have the maximum value *I(ṽ₀),* which is the same as the light intensity at source, and two beams are said to interfere constructively. If $\delta = (n+1/2)\lambda_0$ (n=0,1,2...), *I'(δ)* will be 0, meaning no light from the source reaches the detector, and two beams are said to interfere destructively.

The first term of the Intensity *I'(δ)* equals to $0.5I(\tilde{v}_0)$, a constant and is called dc component. Meanwhile the second term equals to $0.5I(\tilde{v}_0)\cos 2\pi\ \tilde{v}_0\delta$ and called ac component. This ac component is generally referred to the interferogram *I(δ)* measured from an ideal interferometer:

$$I(\delta) = 0.5I(\tilde{v}_0)\cos 2\pi\tilde{v}_0\delta \qquad \text{(Equation 2.2)}$$

In reality, a single wavenumber-dependent correction factor *H(ṽ₀)* needs to be introduced to compensate the nonideality caused by beam splitter efficiency, detector responses and amplifier characteristics. As a result, Equation 2.2 becomes

$$S(\delta) = 0.5H(\tilde{v}_0)I(\tilde{v}_0)\cos 2\pi\tilde{v}_0\delta \qquad \text{(Equation 2.3)}$$

where *S(δ)* is the real interferogram intensity. And $0.5H(\tilde{v}_0)I(\tilde{v}_0)$ can be set to be single-beam spectral intensity *B(ṽ₀)*, Equation 2.3 can be simplified as

$$S(\delta) = B(\tilde{v}_0)\cos 2\pi\tilde{v}_0\delta \qquad \text{(Equation 2.4)}$$

where interferogram *S(δ)* is the cosine Fourier transform of *B(ṽ₀) (Griffiths and De Haseth, 2007)*.

2.3.2 Bruker IFS66vs FTIR system

A Bruker IFS 66vs FTIR system Figure 2.4 made by Bruker Inc. was used for most of the infrared measurement discussed in the thesis. It is a customized FTIR spectrometer composed with three detectors: a DTGS room-temperature detector, a D316 Liquid-N_2 cooled MCT detector and a D317 Liquid-N2 cooled Photovoltaic MCT detector. The light source is a water cooled SiC glowbar emitting light in the range of 7500-370 cm^{-1}. It is equipped with a KBr beam splitter and interferometer mirror can be scanning at high speed up to 320 kHz. The optical chamber is under vacuum higher than 10 mbar during experiment, and sample chamber is under constant dry N_2 purging in order to remove water vapor and carbon dioxide. Bruker OPUS software (v5.5) were used for setting instrument parameters, performing data collection and analysis.

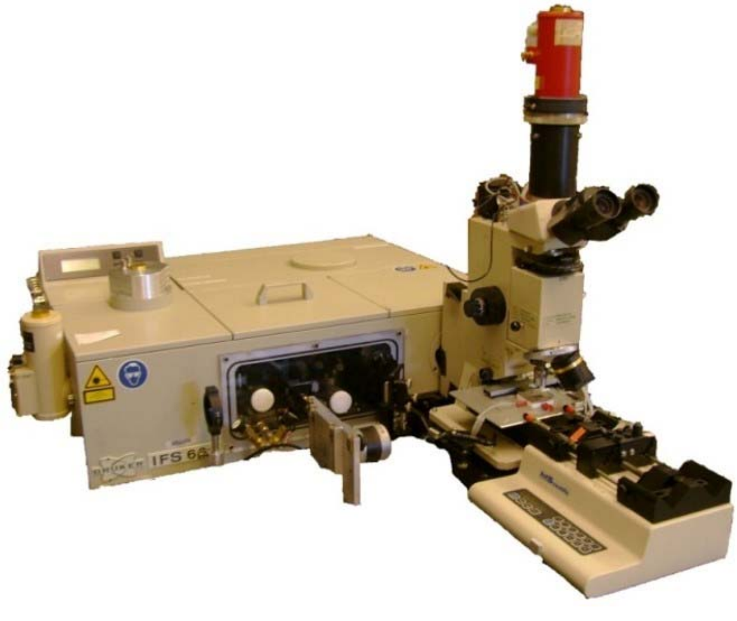


Figure 2.4 Bruker IFS 66V FTIR system

2.3.3 Infrared Sample

Unless ATR Cell is used, liquid samples with 2.7 µl to 3.0 µl volume are sandwiched with a pair of CaF_2 or BaF_2 windows (15 mm x 2 mm) separated by 6 µm or 12 µm plastic spacer. The windows are housed in a home-made copper sample holder tight with screws to prevent dry out during measurement.

2.3.4 Absorption Measurement using 3-sample Exchanger

A home-designed 3-sample exchanger made by OSU Physics machine shop was installed in the sample chamber for normal IR absorption measurement. The 3-sample exchanger is driven by a 5023-127 step motor controlled by Si3540 step motor driver (both step motor and step motor driver are purchased from Applied Motions Product Inc.). The translational motion of the 3-sample exchanger is triggered by a signal from FTIR at rapid-scan mode in order to synchronize the data collection and the sample position change. A very short time gap after the sample transition is programmed to stabilize the sample before infrared scanning. The 3-sample exchanger is made of copper and designed internal flow channels are connected to a RTE-111 (NESLAB Instruments) circulating water temperature controller in order to maintain sample temperature during measurement. Mostly the temperature was set to be 298K. Middle position of the 3-sample exchanger for absorption measurements are normally loaded a pair of empty BaF_2 or CaF_2 windows as reference. The use of 3-sample exchanger significantly reduces the noise from water vapor.

For absorption measurement, one loop is designed to perform 10 scans at each sample position with mirror scanning speed at 200 kHz and one individual measurement usually have 50 to 150 loops. To achieve high data quality, repeating 5 to 10 measurement on one sample is required. Most of the IR absorption are measured using D316 Liquid-N_2 cooled MCT detector with 2 cm^{-1} spectral resolution. IR measuring beam is set to be 5 mm in diameter.

2.3.5 Absorption Measurement using ATR Cell

For some applications, a DuraSamplIR II 9-reflection Diamond/KRS-5 Attenuated Total Reflection (ATR) cell was placed in the sample compartment under dry N_2 purge. 20 μl sample are loaded to the ATR cell diamond surface and single beam spectra is measured using D316 Liquid-N_2 cooled MCT detector with 2 cm^{-1} spectral resolution.

2.3.6 Rapid-scan Time-resolved Infrared Spectroscopy

Bruker IFS 66vs FTIR is equipped with rapid-scan option. Because of its ability to scan at a speed as high as 320 kHz, the time required for the movable mirror to travel over the required distance is sufficient for the measurement of a complete interferogram in less than 100 ms. Therefore, the system is able to synchronize sending out a signal for externally triggering and the data collection afterwards. The time-resolved measurement can reach a time resolution as high as milliseconds. And the scans measured before the triggering can be used as reference for the different absorption calculation. Normally the triggering-data collection procedure needs to be repeated many times (>100) in order to achieve high S/N ratio. Experimental setup of rapid-scan FTIR measurement is illustrated in Figure 2.5.

In this thesis most rapid-scan time-resolved infrared spectroscopy measurement are performed on PYP, with 5 mm diameter IR measuring beam, 4.5 cm^{-1} spectral resolution and mirror scanning at 200 kHz. The PYP photocycle was triggered by 4 ns laser pulses using energy of 3 mJ per pulse at the wavelength of 475 nm. The laser pulses are pumped by a Brilliant Big Sky YAG laser from Quantel and tuned to 475 nm by an OPO from Opotek.

2.3.7 Step-scan Time-resolved Infrared Spectroscopy

For some biological functions happen in the time resolution of microsecond or nanosecond, rapid-scan method won't be suitable because the time of the measuring a complete interferogram won't be able to finish within time. This limit can be overcome through the use of a step-scan interferometer. In a step-scan time-resolved FTIR measurement, the movable mirror is fixed at a desired sampling position to give a constant optical path difference, then a time-dependent change to the interferogram is measured. The movable mirror then undergoes step-wise movement to the next sampling position, perform another time-dependent measurement, and goes on until finishing measurement of the complete length of interferogram (Siebert and Hildebrandt, 2008).

The control of experimental conditions for step-scan time-resolved FTIR measurement is very crucial. The reaction of the sample must be precisely repeatable, the initiation of the reaction and data acquisition must be highly synchronized. Any variation will result in spectral artifacts (Griffiths and De Haseth, 2007). The instrument must be in extremely stable condition during the experiment, because even small vibrations on the system may cause variations on the optical path difference, which is supposed to be held constant at each step and variations will leads to spectral noise. Therefore, in our lab the Bruker IFS 66vs FTIR system is placed on a Newport RS4000 optical table (6 ft.*4 ft.*2 ft.) which is floating on N_2 during measurement for the best isolation from vibration. In addition, between the FTIR system and the vacuum pump, the vacuum tube is mounted to a vibration isolator to reduce the vibrations on the system.

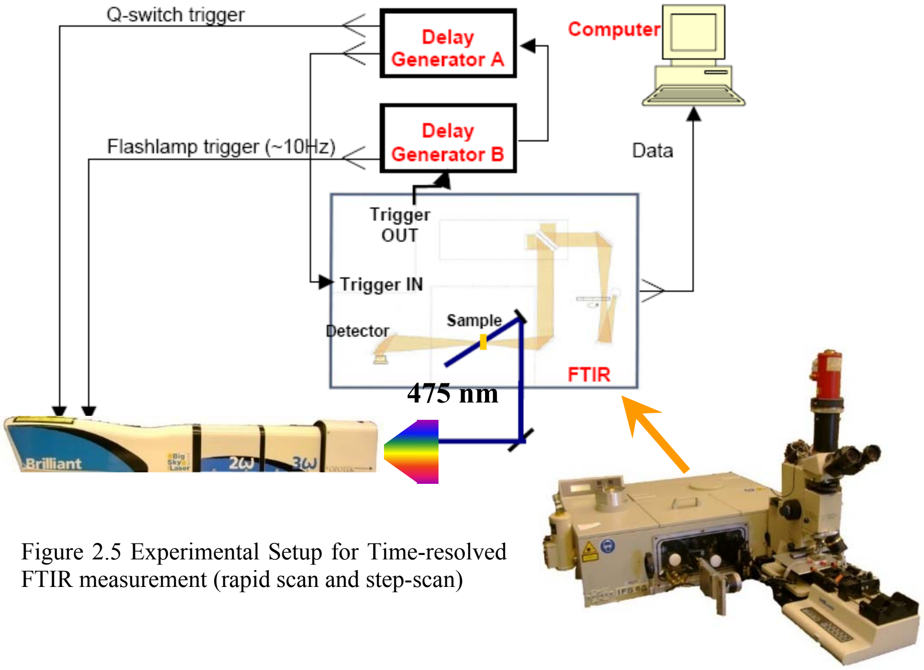


Figure 2.5 Experimental Setup for Time-resolved FTIR measurement (rapid scan and step-scan)

Set up of the optics for step-scan FTIR measurement including the FTIR instrument, the laser and the sample chamber is the same as rapid-scan FTIR, shown in Figure *2.5*. D317 Photovoltaic Liquid-N_2 cooled MCT detector is used for step-scan data collection. For step-scan measurement at microsecond time scale, signals is sent out through DC output to a 1 MHz amplifier and then to the 8-bit 200 kHz PAS82 digitizer. The AC output is connected to 16-bit 200 kHZ A/D converter within FTIR system. Data processing procedure is explained in Figure 2.6. For the step-scan FTIR measurement of PYP discussed in this thesis, spectral resolution is set to 4.5 cm^{-1} and scanning mirror velocity is 100 kHz. Phase resolution is 16 cm^{-1} with zero filling 4. A optical filter is placed in front of the detector to cutoff signals greater than 1850 cm^{-1} and only spectra in range of 1850 cm-1 to 990 cm-1 was collected and processed, therefore 505 mirror positions are used. Data was collected with quasi-logarithmic 5 μs time resolution. External delay generators are used to synchronize the laser, 3-sample exchanger and the IR scans.

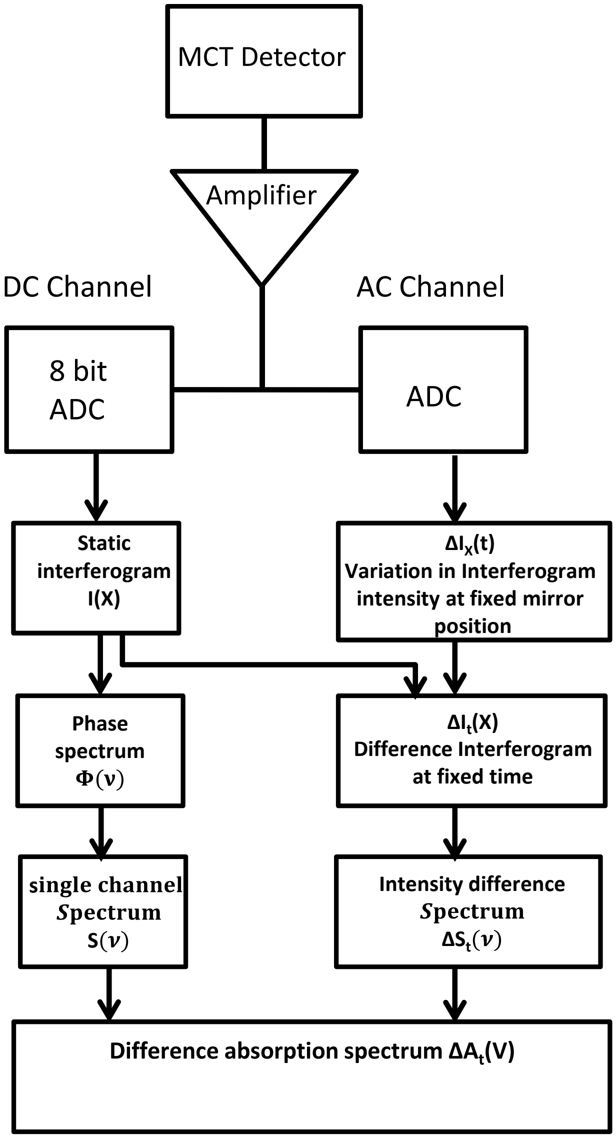

MCT Detector
Amplifier
DC Channel
AC Channel
8 bit ADC
ADC
Static interferogram I(X)
$\Delta I_X(t)$ Variation in Interferogram intensity at fixed mirror position
Phase spectrum $\Phi(\nu)$
$\Delta I_t(X)$ Difference Interferogram at fixed time
single channel Spectrum $S(\nu)$
Intensity difference Spectrum $\Delta S_t(\nu)$
Difference absorption spectrum $\Delta A_t(V)$

Figure 2.6 Data Processing Procedure for Step-scan FTIR measurement

CHAPTER 3

VSMs for Probing Hydrogen-Bonding Interactions of Buried COO(-) in Proteins

3.1 Introductions

3.1.1 Carboxylate groups in proteins

Carboxylic groups (COOH or COO(-)) are often found buried in active site of proteins as the side chain of aspartic acid (Asp pr D, Figure 3.1A) and glutamic acid (Glu or E, Figure 3.1B). Exposed side chains of Asp and Glu in solutions typically have p*K*$_a$ values around 4 (Creighton, 1993). However for Asp and Glu with side chain buried in proteins, p*K*$_a$ values are perturbed to a broad range of values, some are found to be as high as 12 (Merz, 1991, Szaraz et al, 1994, Meyer et al, 2003). As the proteins function, dramatic p*K*$_a$ value shifts are observed occurring frequently on buried carboxylic groups. This phenomenon usually leads to functionally important structural changes such as protonation/deprotonation and hydrogen-bonding interaction (Song et al, 2003, Xie et al, 2001).

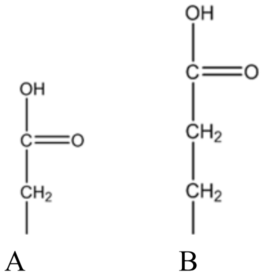


Figure 3.1 Protonated side chain of Asp and Glu (A) Asp; (B) Glu

Many buried Asp and Glu at the active site of proteins, protonated or deprotonated, are functionally important. In bacteriorhosopsin, a light-driven proton pump, 4 out of 18 Asp and Glu residues are buried: Asp-85, Asp-96, Asp-115 and Asp-212. These residues are found either inside the chromophore binding pocket or on the proton transfer pathway (Lanyi and Schobert, 2002, Engelhard et al, 1985, Krebs and Khorana, 1993, Pebay-Peyroula et al, 1997, Subramaniam and Henderson, 2000). Asp-85 and Asp-212 are crucial for the charge stabilization of the positively charged Schiff base of the retinal chromophore, as both are deprotonated at the bR state and Asp-85 will be protonated upon the deprotonation of Schiff base. Asp-96 is be the proton donor for the reprotonation of Schiff base during M to N transition (Engelhard et al, 1985, Balashov et al, 1996, Bousche et al, 1991, Braiman et al, 1988, Heberle et al, 1993, Lanyi and Schobert, 2003, Maeda et al, 2005, Rothschild et al, 1990, Rothschild et al, 1993). In Rhodopsin, upon light activation highly conserved carboxylic residue Glu-113 in the transmembrane domain is the proton acceptor for the deprotonation of Schiff base. Glu-134 functions in the proton uptake from the solvent during the regeneration of rhodopsin (Arnis et al, 1994, Mahalingam et al, 2008).

Carboxylic groups are also found widely playing significant key roles in enzymatic catalysis via all types of hydrogen-bonding interactions (Gutteridge and Thornton, 2005, Bartlett et al, 2002). In fact, carboxylate groups contributed by Asp and Glu residues, particularly in their negatively charged form carboxylate, are the most observed functioning residues in the active site of catalytic enzymes (Gutteridge and Thornton, 2005, Bartlett et al, 2002). For example, in glycosidases like cellobiohydrolase, two carboxylic groups are usually paired together to increase their pK_a values. As a result, one of the paired carboxylic residues will remain protonated and a hydrogen-bonding interaction is formed in between (Varrot et al, 2003). And in aspartic protease such as HIV-1 protease, both carboxylates function as acid-base and this proton will be donated to the substrate (Das et al, 2006). For interactions formed between carboxylate residues and positively charged groups such as arginine and lysine, charges are stabilized that neither group will lose or gain proton. This effect is crucial for functions as polarizing substrate like in adenylate kinase (Yan and Tsai, 1991) which needs a charge (either positive or negative), or for functions as performing a nucleophilic attack on the substrate like sucrose phosphorylase (Sprogoe et al, 2004), or for providing nucleophiles and acid-bases for functions. For interactions between carboxylate groups and polar groups such as histidine and threonine, the carboxylate groups raise pKa values of polar groups and turned them into acid-bases.

In Photoactive Yellow Protein (PYP), a blue-light bacterial photoreceptor, only one out of 19 carboxylate residues, Glu-46, is buried at the active site (Borgstahl et al, 1995, Getzoff et al, 2003, Xie et al, 1996), which has an abnormal high pKa of 11 or above at its receptor state (pG) (Meyer et al, 2003). Thus, Glu46 is protonated in its receptor state pG, even at neutral pH. Upon light excitation of photoactive yellow protein, Glu46 becomes deprotonated by donating a proton to the ionized chromophore during the formation of the putative signaling state (pB) upon receiving photons (Xie et al, 2001, Xie et al, 1996, Brudler et al, 2000, Imamoto et al, 1997). Previous studies showed that the deprotonation of Glu46 is functionally important. The buried COO(-) of Glu46 acts as an electrostatic epicenter that drives large protein conformational changes or “protein quake”, resulting the receptor activation of photoactive yellow protein (Xie et al, 2001).

3.1.2 Signature vibrational modes of COO(-) in infrared spectrum

Upon deprotonation of COOH, one excess electron is delocalized over carboxylate COO(-) group. Two vibrational modes, asymmetric and symmetric stretching, have been used as fingerprints for COO(-) group. The two corresponding infrared bands are typically observed with high intensities in infrared spectroscopy (Jones and McLaren, 1954). For Asp/Glu as amino acid monomers or from short peptide, reported frequencies of asymmetric and symmetric stretching of side chain COO(-) were summarized in Table 3.1.

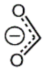
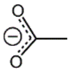
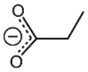
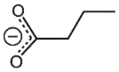
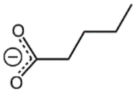
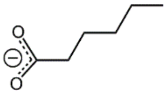
Table 3.1 Reported experimental vibrational frequencies of asymmetric stretching and symmetric stretching of COO(-) from side chain of Aspartate and Glutamate in H_2O and D_2O

	H_2O		D_2O	
	ν_{asym} (cm^{-1})	ν_{sym} (cm^{-1})	ν_{asym} (cm^{-1})	ν_{sym} (cm^{-1})
Aspartate†	1574-1579	1402	1584-1586	1404
Glutamate†	1556-1560	1404	1567-1568	1406

†Infrared data of aspartate and glutamate in H_2O were reported by Venyaminov (Venyaminov and Kalnin, 1990) and Rahmelow (Rahmelow et al, 1998); data in D_2O were reported by Chirgadze (Chirgadze et al, 1975) and Wright (Wright and Vanderkooi, 1997).

As shown in Table 3.1, there is a difference between aspartate and glutamate in the frequencies of asymmetric stretching. However, it should be noticed that the data summarized here were from the infrared absorption measurement of monomer amino acid or short peptides. Earlier infrared data (shown in Table 3.2) reported by Cabaniss (Cabaniss and McVey, 1995) suggest that for exposed COO(-) group, length of the chain attached to carboxylate group has effect on the peak positions of asymmetric and symmetric stretching. Table 3.2 show that the COO(-) stretching frequencies are sensitive to the chain length for 1-4 carbon chains and then remains stable with 4-6 carbon chains. It indicates that when the chain length further extended, its effect on the two frequencies becomes much less. Therefore, aspartate and glutamate as single amino acids in solution are expected to have different vibrational frequencies for ν_{asym} and ν_{sym}. But as residues which are part of the long peptide in proteins, COO(-) from aspartate and glutamate should result in similar infrared signals.

Table 3.2. Variations in the asymmetric and symmetric stretching frequencies of COO(-) caused by carbon chain length†.

Molecule	Structure	ν_{asym} (cm^{-1})	ν_{sym} (cm^{-1})
Formic Acid		1580	1351
Acetic Acid		1551	1416
Propanoic Acid		1545	1413
Butyric Acid		1543	1408
Pentanoic Acid		1541	1408
Hexanoic Acid		1541	1408

†Infrared absorption were measured in aqueous solution of aliphatic carboxylate by Cabaniss (Cabaniss and McVey, 1995)

3.1.3 Reported vibrational frequencies of COO(-) in proteins

In order to probe and characterize structural dynamics of protein, certain spectroscopy techniques such as UV-Vis, Circular Dichroism, Raman and infrared have been developed. Time-resolved infrared spectroscopy is particularly suitable for detecting changes of protonation state and hydrogen-bonding interactions during protein function.

Due to overlapping with infrared signals from other amino acid residues, few studies have been reported on the COO(-) signals in proteins. COO(-) of Asp and Glu in proteins are capable of being detected using infrared spectroscopy by locating either the asymmetric stretching or the symmetric stretching vibrational bands. In Sensory Rhodopsin I, a peak at 1381 cm^{-1} was assigned to symmetric stretching of COO(-) from Asp-76 in the dark state, indicating its deprotonation state (Rath et al, 1996). In Green Fluorescence Protein, 1564 cm^{-1} and 1570 cm^{-1} are assigned to Glu-222 in A1* and A2* state, respectively, indicating its deprotonation at those two states (van Thor et al, 2005).

3.1.4 Vibrational Structural Markers

A vibrational structural marker for probing hydrogen-bonding interactions of COOH group has been reported previously by our group (Nie et al, 2005). A strong correlation between the vibrational frequency of C=O stretching and the number of hydrogen-bonding interactions a COOH group forms was found: ~1759-1775 cm^{-1} for zero, ~1733-1749 cm^{-1} for one (COOH as hydrogen-bond donor or acceptor), and 1703-1710 cm^{-1} for two hydrogen-bonding interactions.

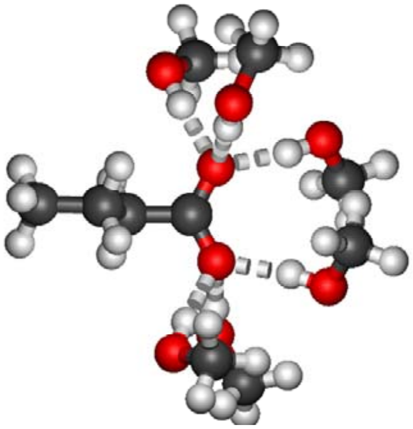


Figure 3.2 COO(-) group forms maximum of 6 hydrogen-bonding interactions

Unlike COOH, COO(-) group can form as many as 6 hydrogen-bonding interactions as Figure 3.2 shows. Hydrogen-bond complex with same number of hydrogen-bonding interactions could form in different ways. For example, for a total of two hydrogen-bonding interactions, there are four ways of arrangement. Therefore, only using one vibrational mode may not be enough to distinguish all types of hydrogen-bonding interactions of COO(-). In this project, I introduced and developed a set of two vibrational structural markers (VSM) using asymmetric and symmetric stretching frequencies. This 2D-VSM enables detection of not only the number, but also the type of hydrogen-bonding interactions formed on buried COO(-) from Asp and Glu residues in proteins.

3.2 Materials and Methods

3.2.1 Selection of protein crystal structures

Protein Data Bank (PDB), (Berman et al, 2000) stores reported protein structures from X-ray Crystallography, NMR and many other techniques. In April 2014, more than 92000 reported protein structures. In order to access the statistical occurrences of various types of hydrogen-bonding interactions of buried Asp and Glu residues, we selected and downloaded 1,182 protein crystal structures from RCSB Protein Databank. The selection criteria were: (1) Monomer proteins without binding of DNA or RNA (2) Only protein structures detected using X-ray diffraction as experimental method; (3) Structure resolution between 0.40 Å and 1.50 Å; (4) R-free value of 0.20 or better (R value is the measure of the quality of the atomic model obtained from the crystallographic data (Kleywegt and Jones, 1997)); (5) Minimum molecular weight 9000 Daltons (approximately 80 amino acid residues at least); (6) Sequence identity between each pair of selected proteins are less than 30%. Proteins are large molecules. Intrinsically, a protein molecule can have different conformation with almost the same energy. Therefore, downloaded PDB files were first treated with a UNIX script (Appendix C) which separated multiple conformations and saved to edited PDB files with only one conformation of the structure.

3.2.2 CCP4 for bioinformatics analysis of buried side chains

The AREAIMOL function of CCP4 software package, version 6.4 (Lee and Richards, 1971, Saff and Kuijlaars, 1997) was employed to calculate solvent accessible area for each individual atom in protein in order to determine its solvent accessibility. A UNIX script (Appendix C) was used in the AREAIMOL output files to collect the solvent exposure values for carboxylic oxygen of Asp and Glu residues only. The minimum solvent accessible area for the carboxylic oxygen atom in Asp or Glu is 0, indicating fully buried. The maximum accessible area for a pair of carboxylic oxygen atoms is 100 for extremely exposed carboxylate group.

3.2.3 HBPlus

HBPlus software package (v3.06) (McDonald and Thornton, 1994) was used to analyze the hydrogen-bonding interactions within each selected protein crystal structure. 3.20 Å was set to be the maximum hydrogen-bonding length during HBPlus analysis. The output data was first treated with a UNIX script (Appendix xx) to collect only hydrogen-bonding interactions involved with carboxylic oxygen of Asp and Glu. Then the data was saved and exported to spreadsheet for further statistical analysis.

3.2.4 Further data analysis using spread sheet

Statistical analysis of hydrogen-bonding interactions of buried Asp and Glu residues were performed using Microsoft Excel 2013. Among all hydrogen-bonding interactions, only those involved side chains of Asp and Glu residues were selected. Then every residue was assigned a name code in the format of "XXXXY123AAA", as "XXXX" refers to PDB ID of the protein, "Y" refers to the chain number, "123" refers to residue number, and "AAA" refers to residue name (for amino acid residue, 3-letter abbreviation was used). Each buried Asp and Glu residue from CCP4 were assigned name codes in the same format. The list of buried Asp and Glu residues was then searched by using "VLOOKUP" function of Excel, through the list of hydrogen-bonding interactions in which Asp and Glu side chain is the hydrogen-bond acceptor. A new data file is then created which contains the hydrogen-bond information of each buried Asp and Glu side chain.

3.2.5 Structural optimization and vibrational frequency calculation using Density Functional Theory

Asp and Glu residues in proteins do not carry charged amine or carboxylic groups as they are in monomer amino acid form. The COO(-) side chains of Asp and Glu residues in protein were modeled using a negatively charged butyric acid ($^{-}OOC\text{-}CH_2CH_2CH_3$). Methanol molecules were employed to serve as hydrogen-bond donors to COO(-) of buried Asp/Glu.

A significant advantage of using quantum theory, density function theory (DFT) for structural optimization and frequency calculation, is that such first principle calculations are not influenced by parameters selected as in classical calculations. All DFT calculations were carried out using Gaussian 09, Revision C.01 (Frisch, 2009) installed on Cowboy Cluster in the OSU High Performance Computing Center. GaussView 5.0 (Dennington, 2009) was used to generate input structures and visualize output structures with vibrational displacements. Procedures of calculation for structural optimization, energy and frequency calculation in Gaussian 09 are the similar to calculation in Gaussian

03 previously reported (Nie et al, 2005). A large basis set, 6-311+G(3df,2p), with 39 functions on heavy atoms and 9 functions on hydrogen atoms, is used instead of 6-31G(d). Selected computational method B3LYP/6-311+G(3df,2p) allows excellent balance of good accuracy and reasonable computational time. B3LYP/6-31G(d) gives considerably accurate results for neutral COOH. However it is not suitable for COO(-) due to the lack of diffusion functions. Since the computational power has been dramatically improved, particularly with the development of the OSU High Performance Computing center, we can perform extensive computational studies using methods with much more functions applied on charged groups within considerable time. More details of selecting computational methods will be discussed in results section. Scaling factor 0.988 was applied on the computational frequencies. This scaling factor is determined by calibrating computational frequencies for COO(-) in methanol with experimental infrared absorption data, which will be shown in Results and Discussion section.

The hydrogen-bond dissociation energy is defined as:

$$\Delta E = E_{BAC} + E_{HB-Partner} - E_{Total} \qquad \text{(Equation 3.1)}$$

where BAC stands for butyric acid. The output energy is in the atomic unit (au), which was then converted in to kJ/mol by multiplying a factor of 1 au = 2625.5 kJ/mol.

3.2.6 FTIR spectroscopy

Butyric acid (CH_3-CH_2-CH_2-COOH, >99%, purchased from Alfa Aesar) was used to prepare 600mM butyric acid solution with 1.2M NaOH in H_2O, or with 1.2M NaOD in D_2O or methanol solvent. A Bruker IFS66v FTIR spectrometer with rapid-scan option was used for infrared absorption measurement. The optical path within spectrometer was under vacuum, except the sample chamber. A 9-relection Diamond/KRS-5 ATR cell (DuraSamplIR II, SensIR) was placed in the sample compartment under dry nitrogen purge. 20 μl of butyric sample was loaded to the ATR cell. Temperature of the sample was maintained at 25°C by a recirculating temperature controller (RTE-111, NESLAB). Diameter of the infrared measurement beam was set to 5 mm. The interferometer scanning rate was 200 kHz. A liquid nitrogen cooled Mercury Cadmium Telluride (MCT) detector was used for data collection with 2 cm^{-1} spectral resolution. Single beam data in the range of 4000 -850 cm^{-1} were collected. The absorption data was averaged over 12,000 scans to achieve a high S/N ratio of 3000 with 0.2 milliOD noise. The 2nd derivative of absorbance was calculated using Bruker OPUS software (v5.5) based on Savitzky-Golay algorithm with 21 smoothing points.

3.3 Results and Discussions

3.3.1 H-Bond interactions of buried Asp & Glu in proteins

From 1,182 selected protein crystal structures, we found a total of 46875 COOH/COO(-) groups from side chain of 23,787 Asp residues and 23,088 Glu residues, representing ~40 COOH/COO(-) groups per protein. 6.2% of Asp residues (1,482 out of 23,787) are fully buried in proteins, equivalent of 1.3 Asp residues per protein. And 4.5% of Glu residues (1037 out of 23088) are fully buried in proteins, equivalent of 0.9 Glu residues per protein. The rarity of fully buried Asp and Glu reflects the costly energy required to bury charged COO(-) groups in hydrophobic interior of proteins (Xie et al, 1996, Xie et al, 2001Nie, 2006).

The fully buried carboxyl groups in proteins are often functionally important. As we have discussed, examples include Glu113 and Glu134 of rhodopsin for animal vision (Arnis et al, 1994, Mahalingam et al, 2008), Asp82, Asp96, Asp212, Asp96, and Glu204 for the light driven proton pumping process in bacteriorhodopsin (Lanyi and Schobert, 2002, Engelhard et al, 1985, Krebs and Khorana, 1993, Pebay-Peyroula et al, 1997, Subramaniam and Henderson, 2000), and Glu46 of photoactive yellow protein (Xie et al, 2001, Xie et al, 1996). Therefore in this chapter we will focus on developing vibrational structural markers for fully buried side chains of Asp and Glu.

Due to the lack of abundant hydrogen-bond donors and acceptors inside proteins, the maximum number of hydrogen-bonding interactions of buried COOH/COO(-) group is significantly less six, the number of hydrogen-bonding interactions of fully solvent exposed groups. We employed HBPlus (McDonald and Thornton, 1994) to analyze the hydrogen-bonding interactions of buried COOH/COO(-) groups from Asp and Glu residues. We exclude buried Asp/Glu residues which have hydrogen-bonding interactions with water molecules, because the flexibility of small water molecule makes its positions in most protein crystal structures inaccurate.

The results are summarized in Table 3.3. For COOH/COO(-) of Asp, 10.0% with no hydrogen-bonding interaction, 15.5% with 1 hydrogen-bonding interaction, 17.9% with 2 hydrogen-bonding interactions, 24.7% with 3 hydrogen-bonding interactions, 25.2% with 4 hydrogen-bonding interactions and 6.6% with more than 4 hydrogen-bonding interactions. And COOH/COO(-) of Glu, the percentage of forming 0, 1, 2, 3, 4 and more than 4 hydrogen-bonding interactions are 19.2%, 16.2%, 14.0%, 16.0%, 24.0% and 10.2%, respectively. As a total approximately 45.1% of buried COOH/COO(-) forms 3 or 4 hydrogen-bonding interactions, but only 8.1% forms more than 4. It means that COOH/COO(-) buried in proteins are very likely to have both carboxylate oxygen hydrogen-bonded, but less likely to satisfy its maximum hydrogen-bond potential due to the hydrophobic environment inside proteins. It should be noticed that although HBPlus do not distinguish between COOH and COOH/COO(-) groups, this does not affect our conclusion.

Table 3.3 Statistics of hydrogen-bonding interactions of fully buried and hydrophobic Asp and Glu residues in proteins

Number of HB on Carboxylate	# of Asp	Percent	# of Glu	Percent	# of Asp+Glu	Percent
0	74	10.0%	96	19.2%	170	13.7%
1	114	15.5%	81	16.2%	195	15.8%
2	132	17.9%	70	14.0%	202	16.3%
3	182	24.7%	80	16.0%	262	21.2%
4	186	25.2%	122	24.4%	308	24.9%
>4	49	6.6%	51	10.2%	100	8.1%
Total	737	100.0%	500	100.0%	1237	100.0%

Since COO(-) is a pure hydrogen-bond acceptor, the hydrogen-bond partners found paring with COO(-) in proteins are hydrogen-bond donors only. We further analysis the preferences of hydrogen-bond partners for fully buried and hydrophobic side chain of Asp and Glu residues, as the results are shown in Table 3.4. Summarized statistical data show that backbone amide is the most common hydrogen-bond donor paired with side chain of Asp/Glu. 1916 and 1416 hydrogen-bond partners are found paring with Asp and Glu side chain, respectively. That's an average of ~2.6 hydrogen-bond partners per Asp and 2.8 hydrogen-bond partners per Glu. Among these, backbone amide is found the most common hydrogen-bond donor, counting for 43.1% of the pairs with Asp and 31.1% of the pairs with Glu. Arg, which most likely carries a positive charge, counts the second most, donating 17.7% of the hydrogen-bonding interactions to Asp and 29.1% of the hydrogen-bonding interactions to Glu. A histogram corresponding to the statistical data is also plotted, shown in Figure 3.2.

Table 3.4 Statistics of hydrogen-bond partners for fully buried and hydrophobic Asp and Glu residues in proteins

HB Partner		HB Group	# of Asp	Percent	# of Glu	Percent	# of Asp+Glu	Percent
Backbone		-NH	826	43.1%	440	31.1%	1266	38.0%
Side Chain of Amino Acid Residues	Arg	-NH(+)	339	17.7%	412	29.1%	751	22.5%
	Ser	-OH	151	7.9%	139	9.8%	290	8.7%
	Tyr	-OH	119	6.2%	119	8.4%	238	7.1%
	Thr	-OH	136	7.1%	62	4.4%	198	5.9%
	His	-NH/-NH(+)	111	5.8%	71	5.0%	182	5.5%
	Lys	-NH(+)	85	4.4%	69	4.9%	154	4.6%
	Asn	-NH	59	3.1%	46	3.2%	105	3.2%
	Gln	-NH	53	2.6%	34	2.4%	87	2.6%
	Trp	-NH	36	1.9%	24	1.7%	60	1.8%
	Cys	-SH	1	0.1%	0	0.0%	1	0.0%
Total			1916	100.0%	1416	100.0%	3332	100.0%

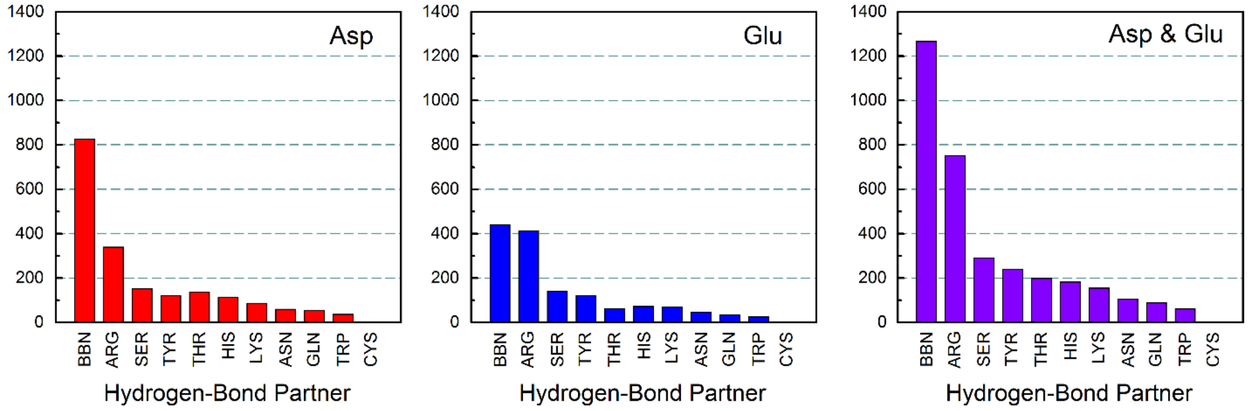


Figure 3.3 Histogram of hydrogen-bond partners to fully buried and hydrophobic Asp/Glu side chains in proteins. Graph is plotted based on the statistical results summarized in Table 3.4.

3.3.2 Criteria and Selection of Computational method

There is often a trade-off between computational accuracy and computational time. We seek a computational method that provides the best balance in between. Our criteria for selecting the optimal computational method are (1) high precision on vibrational frequency calculations (≤ 2 cm^{-1} or equivalently 0.15% of accuracy compared with the most expensive method); and (2) affordable computational time (≤ 2 days for complicated calculations). A series of computational methods using the density functional theory B3LYP with different basis sets are tested and compared for frequency calculations (Figure 3.4A & Figure 3.4B), the total number of Gaussian functions used (Figure 3.4C) and the total CPU time (Figure 3.4D). A molecular model system, with one negatively charged butyric acid molecule forming two hydrogen-bonding interactions with one water molecule is used for this test. The simplest basis set is 6-31G(d) with 123 total functions yields vibrational frequencies with overall large standard deviation (~60 cm^{-1} per vibration). An improved basis set 6-31+G(d,p), with 178 total functions dramatically improves the accuracy of vibrational frequencies with overall standard deviation reduced to below 10 cm^{-1} per vibration. The largest basis set used is 6-311++G(3df,2pd) with a total of 408 functions. The total CPU time for this calculation is 20.6 hours for this relatively small system. This method is not affordable for a large number of calculations for this work. Therefore, we search for a basis set that yield similar results as 6-311++G(3df, 2pd), but uses less computational time than this. Among 14 different basis sets tested, we found that only two basis sets, 6-311+G(3df,2p) and 6-311+G(3df,2pd) that meet the criteria on the accuracy of vibrational frequencies. Since 6-311+G(3df,2p) has fewer functions and less CPU time than 6-311+G(3df,2pd), we therefore select 6-311+G(3df,2p) as the optimal method for developing the vibrational structural markers of COO(-) for Asp/Glu in this work.

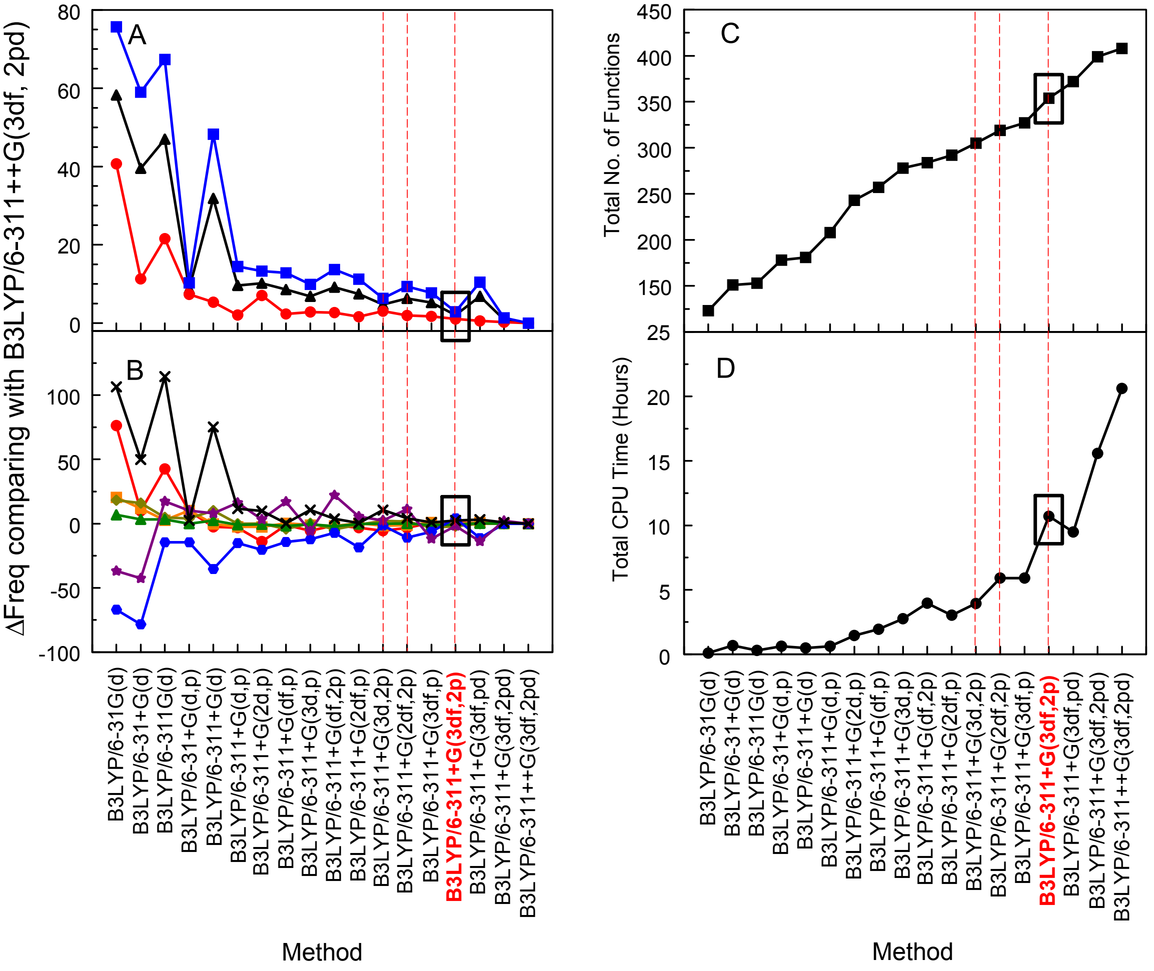


Figure 3.4 COO(-) forming 2 HB with one water molecule is used for computational method comparison (A) standard deviation of Δfreq for the vibrations of COO(-) (red), H2O (blue) and together (black) for different method, the most expensive method (B3LYP/6-311+G(3df,2pd) is used for reference. (B) Δfreq for individual vibrational modes, the most expensive method (B3LYP/6-311+G(3df,2pd) is used for reference. (C) Compare number of functions used for calculations. (D) Compare total CPU time used for calculations. List of Vibrations in (B): 1. COO(-): COO asymmetric stretching (red); 2. COO(-): COO symmetric stretching (orange); 3. COO(-): COO bending (green); 4. COO(-): C-C stretching and C-H bending (yellow green); 5. H2O: HOH bending (black); 6. H2O: O-H stretching-A (purple); 7. H2O: O-H stretching-B (blue)

3.2.3 Optimize the COO(-) structural model for the side chains of Asp and Glu.

The model molecule for frequency calculations of the side chain of Asp/Glu in proteins is optimized and selected for computational accuracy and computational speed. A small molecule, such as CH_3-COO(-), is easy to calculate but might not be a good molecular model for COO(-) in proteins. To assess how many carbons are needed to reliably model the COO(-) on the side chains of Asp and Glu, we calculated the asymmetric stretching and symmetric stretching vibrational frequencies of CH_3-$(CH_2)_n$-COO(-) interacting with one water molecule through two hydrogen-bonding interactions. The data (see Figure 3.5) show that for short side chain in the model compound, frequencies of the two vibrational modes have large dependences on the number of carbons in the chain. However, once the chain contains more than four carbons, the two vibrational frequencies are much more stable. This is due to the fact that the two vibrational modes can be strongly coupled to other vibrational modes when the side chain is short, and they are more isolated when the side chain is reasonably long, as Figure 3.5 shows. Therefore, butyrate ($CH_3CH_2CH_2COO$(-)) was selected to model the COO(-) groups from the side chains of Asp (-) and Glu (-) residues in proteins, partly to since residues in protein are always covalently linked in a long polypeptide. This result is also supported by the experimental data previously reported by Carbaniss (Cabaniss and McVey, 1995), as it has been shown in Table 3.2.

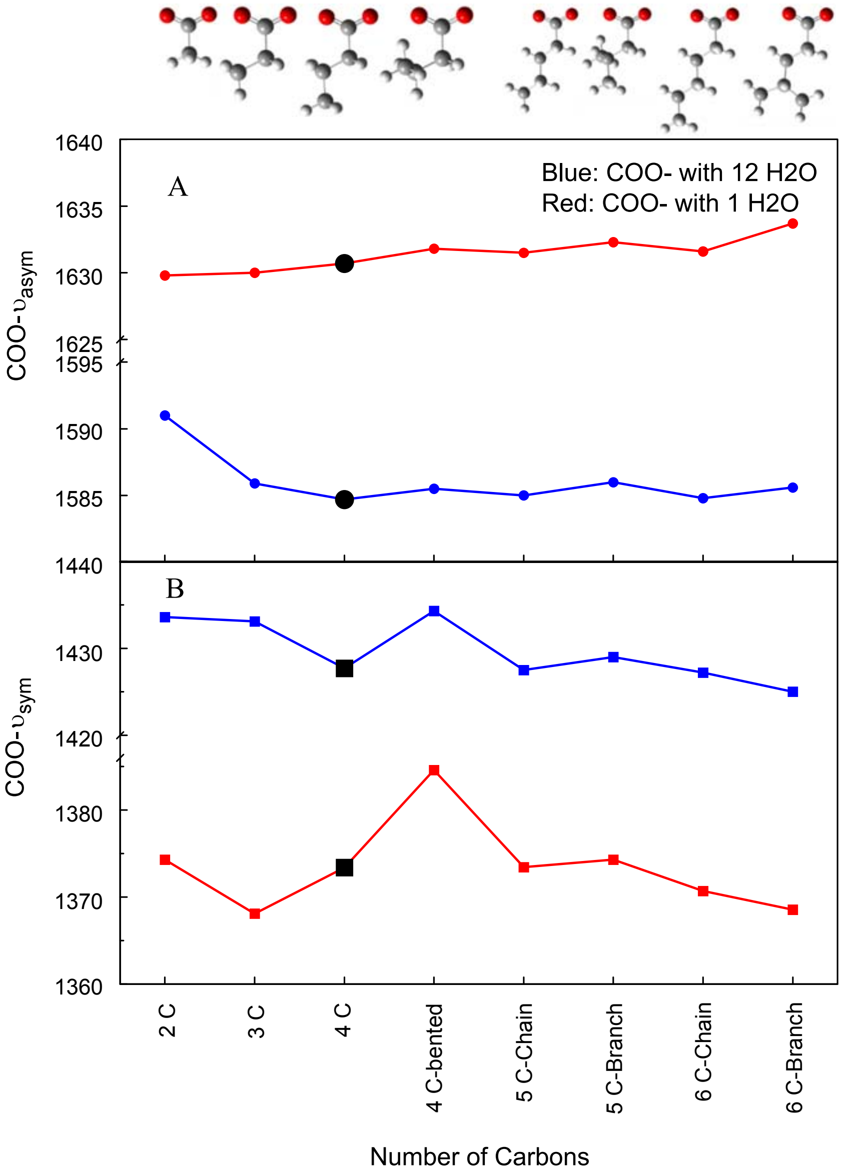


Figure 3.5 Individual model compound of COO(-) group forms 1 HB with 1 H2O (red) or forms 12 HB with 12 H2O (blue). (A) Frequencies of asymmetric stretching. (B) Frequencies of symmetric stretching.

3.2.4 Computational accuracy of COO(-) asymmetric stretching and symmetric stretching frequencies

In order to assess the accuracies of the vibrational frequencies of the symmetric and asymmetric COO(-) vibrational modes from G09 computational method, we performed a total of 16 computational studies with COO(-) interacting with methanol molecules (n=0 to 10) in different forms of hydrogen-bonding interactions. The atomic movements for the symmetric and asymmetric COO(-) vibrations are shown in Figure 3.6, while the vibrational spectra of 16 COO(-)/$(MeOH)_n$ complexes are shown in Figure 3.6. And all computational data with methanol molecule serving as hydrogen-bond donor are summarized in Table 3.9.

Without any hydrogen-bonding interaction, the asymmetric COO(-) vibration is at 1633 cm^{-1}, in the Amide I region, while the symmetric COO(-) vibrational frequency is almost 300 cm^{-1} lower, at 1344 cm^{-1}. Since the ionic COO(-) group is expected to be highly unstable without any hydrogen-bonding interaction, it is extremely rare to observe isolated COO(-) groups not hydrogen-bonded. The maximum number of direct hydrogen-bonding interactions with a COO(-) group is 6, 3 for each carboxylate oxygen. The resulting asymmetric and symmetric COO(-) frequencies are 1575 cm^{-1} and 1402 cm^{-1} respectively. In comparison, the FTIR studies of butyric acid in methanol reveal that the frequency of the asymmetric stretching appears at 1565 cm^{-1}, while the symmetric stretching is observed at 1404 cm^{-1}, indicating a 10 cm^{-1} difference (0.64%) on asymmetric stretching frequency and 2 cm^{-1} difference (0.14%) on symmetric stretching frequency. In order to more closely to better model the hydrogen-bonding interactions in solution, we then further added two more methanol molecules for the second shell of hydrogen-bonding interactions. The resulting optimized structures are shown in Figure 3.6, and the asymmetric and symmetric COO(-) stretching frequencies are further shifted to 1565 and 1404 cm-1, in excellent agreement with experimental data, suggesting excellent accuracy of vibrational frequency calculations using B3LYP/6-311+G(3d,2p). The additional extended HB interactions further improved the agreement between the computational data and the experimental data for both the symmetric and asymmetric vibrational frequencies to only 1 cm^{-1} or less in differences. And with another additional two methanol molecules added to the model system complex, the computational frequencies of the two vibrational modes remain almost identical. Therefore, this comparison demonstrates that the accuracy of the COO(-) vibrations using the B3LYP/6-311+G(3d,2p) is an excellent method, the resulting computational data are consistent with experimental FTIR data.

3.2.5 2D Vibrational Structural Markers for COO(-) group

As Figure 3.6 have shown, when the hydrogen bonds are distributed between the two oxygen, the vibrational frequencies of asymmetric stretching and symmetric stretching both have linear correlation with the number of hydrogen bonds. If no hydrogen-bonding interaction is formed, computational results show COO(-) group is having asymmetric and symmetric stretching at ~1632 cm^{-1} and ~1345 cm^{-1}. And for COO(-) with one hydrogen-bonding interaction , two types of geometry may formed as 1A and 1B in Table 3.9. The results indicate that the frequency of asymmetric stretching is almost the same as COO(-) with no hydrogen-bonding interaction, both appearing at ~1630-1640 cm^{-1}, and a ~10 cm^{-1} blue-shift is observed for the frequency of symmetric stretching.

For COO(-) group forming two hydrogen-bonding interactions, there are four possibilities of geometry of hydrogen-bond complex (2A-2D in Table 3.9). If each carboxylate oxygen in COO(-) carries one hydrogen-bond interaction (2A, 2B and 2C), then the frequencies of the two vibrational modes are quite similar to each other, appearing at ~1615-1622 cm^{-1} for asymmetric stretching and 1375 cm^{-1} for symmetric stretching. However, if two hydrogen-bonding interactions are formed on one oxygen atom, the behavior of the two vibrational frequencies are observed quite different. Frequency of asymmetric stretching is at 1635 cm^{-1}, in the same range as no hydrogen bond and one hydrogen bond, and frequency of symmetric stretching is at 1368 cm^{-1}.

When COO(-) group forms three hydrogen-bonding interactions, three kinds of geometry for hydrogen-bond complex may be formed (3A-3C in Table 3.9). Complex 3A and 3B have two hydrogen-bonding interactions on one oxygen and have the third hydrogen-bonding interaction on the other oxygen. Complex 3C, however, has all three hydrogen-bonding interactions on one oxygen while the other oxygen is left alone. The computational results indicate that such differences in the arrangement of hydrogen-bonding interactions lead to significant shift on frequencies of asymmetric and symmetric stretching. For complex 3A and 3B, asymmetric stretching are observed at 1619 cm^{-1} and 1610 cm^{-1}, and symmetric stretching are observed at 1388 cm^{-1} and 1397 cm^{-1}. Meanwhile, for complex 3C, vibrational frequencies of asymmetric and symmetric stretching appear at 1642 cm^{-1} and 1345 cm^{-1}, respectively. Despite 3 hydrogen-bonding interactions, the frequency results of complex 3C are so abnormal, almost the same as carboxylate forming no hydrogen-bonding interactions.

For COO(-) forms four hydrogen-bonding interactions, three types of geometry are found, as 4A-4C in Table 3.9. Complex 4A has four hydrogen-bonding interactions evenly distributed to two carboxylate oxygen atoms, and complex 4B and 4C each has one oxygen form three hydrogen-bonding interactions and the other oxygen form one. The frequency results implies that the even distribution as 4A results asymmetric stretching at 1602 cm^{-1} and symmetric stretching at 1400 cm^{-1}, while uneven distribution like 4B and 4C leads to 1602 cm^{-1} and 1612 cm^{-1} for asymmetric stretching, 1378 cm^{-1} and 1381 cm^{-1} for symmetric stretching.

The fifth and sixth hydrogen-bonding interaction forms on COO(-) (5, 6A-6C in Table 3.9) will cause further red-shifts on the frequency of asymmetric stretching and blue-shifts on the frequency of symmetric stretching, toward the two bands observed for exposed carboxylate group in infrared measurement. Since buried carboxylate groups in the hydrophobic interior of proteins are unlikely to form 5 or more hydrogen-bonding interactions, we will not further discuss those results here.

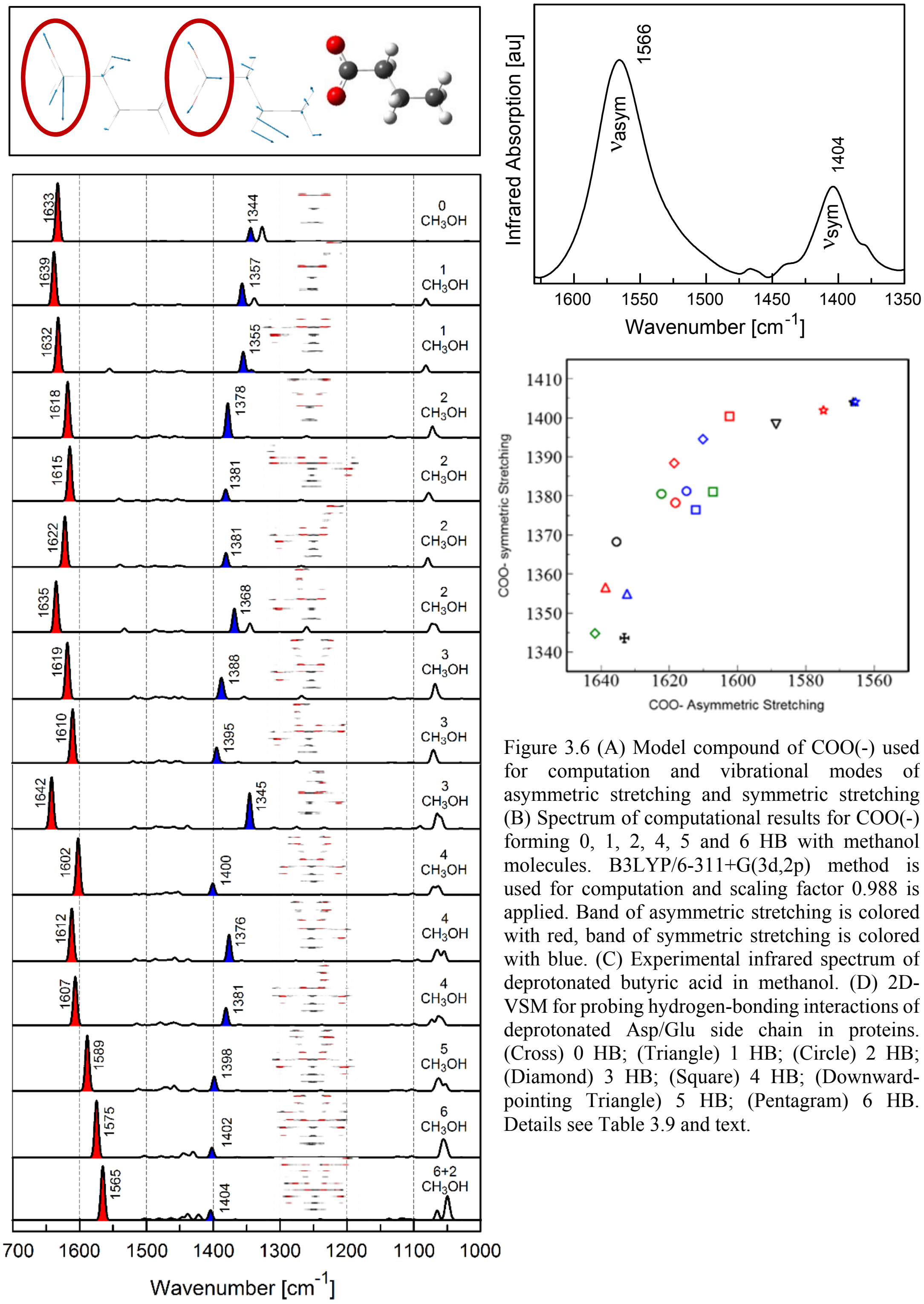


Figure 3.6 (A) Model compound of COO(-) used for computation and vibrational modes of asymmetric stretching and symmetric stretching (B) Spectrum of computational results for COO(-) forming 0, 1, 2, 4, 5 and 6 HB with methanol molecules. B3LYP/6-311+G(3d,2p) method is used for computation and scaling factor 0.988 is applied. Band of asymmetric stretching is colored with red, band of symmetric stretching is colored with blue. (C) Experimental infrared spectrum of deprotonated butyric acid in methanol. (D) 2D-VSM for probing hydrogen-bonding interactions of deprotonated Asp/Glu side chain in proteins. (Cross) 0 HB; (Triangle) 1 HB; (Circle) 2 HB; (Diamond) 3 HB; (Square) 4 HB; (Downward-pointing Triangle) 5 HB; (Pentagram) 6 HB. Details see Table 3.9 and text.

3.2.6 Hydrogen-bond strength dependence of hydrogen-bond length and hydrogen-bond angle

Hydrogen-bond dissociation energy is used to characterize the hydrogen bond strength, as the Equation 3.1 defines. The higher this dissociation energy is, the stronger a hydrogen bond will be, since this energy is required for the complete separation of a hydrogen bond donor and acceptor.

There are three factors that determine the geometry of a hydrogen bond: the length, the angle and the dihedral angle. For COOH groups, previous studies have shown that for each of three types of the one hydrogen bond formation a COOH group may have, the correlation of hydrogen bond length and the hydrogen bond energy fits well to Morse potentials $E = E_e \left\{ \left[1 - e^{-a(r-r_e)} \right]^2 - 1 \right\}$, where r_e is the optimal hydrogen bond length and E_e is the hydrogen-bond dissociation energy (Nie et al, 2005).

For COO(-) groups, however, due to the fact that the oxygen can only be hydrogen-bond acceptor, we only performed this calculation on a COO(-) hydrogen-bonded with a methanol molecule. We chose complex 1A in Table 3.9 as the model system to perform this study. The hydrogen-bond energies with hydrogen-bond length fixed at different distances were calculated in three dielectric environment beside vacuum: in benzene with ε = 2.27, in diethylether with ε = 4.24, and in chlorobenzene with ε = 5.70.

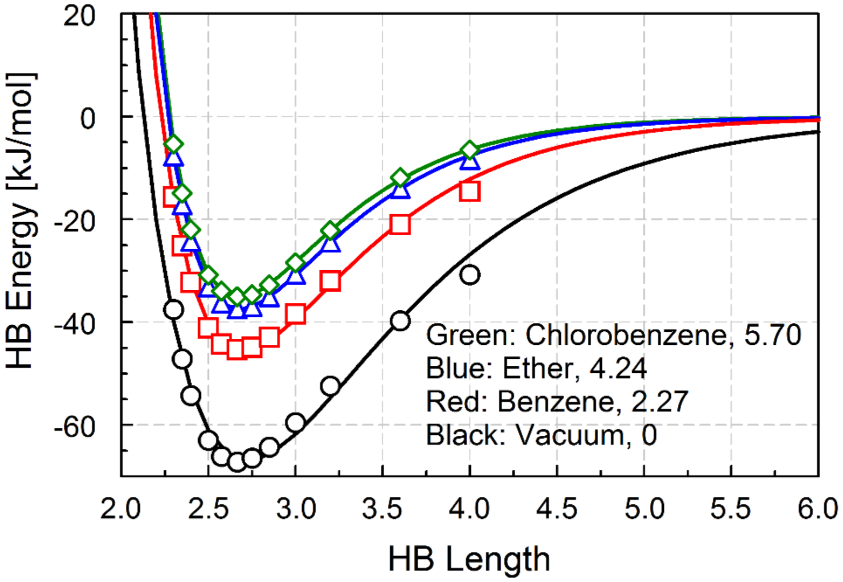


Figure 3.7 Energy landscape for hydrogen-bonding interactions between COO(-) and methanol molecule at different length in multiple dielectric environment: vacuum (ε = 0, black), benzene (ε = 2.27, red), ether (ε = 4.24, blue) and chlorobenzene (ε = 5.70, green).

With hydrogen-bond length increases from 2.30 Å to 4.0 Å, the hydrogen-bond energy output of the computation were plotted as an energy landscape, shown in Figure 3.7. The energy data points were fitted well with Morse potential. The optimal hydrogen-bond length 2.67 Å from geometry optimization of model system without any restrictions is observed with lowest energy.

Similarly, we study the correlation between hydrogen-bond energy and hydrogen-angles formed by carboxylate carbon, and the two hydrogen-bonded oxygen atoms. In vacuum, a dual well landscape is formed by the energy dependence on the hydrogen-bond angle, shown in Figure 3.8. The bottoms of the two energy wells are corresponding to the two optimized angles without any restriction, which are observed at 132.80° for complex 1A and 229.46° for complex 1B in Table 3.9, respectively. The energy barrier between the two wells is found to be around 6 kJ/mol, more than twice of the thermal energy 2.5 kJ/mol. Such barrier is certainly high enough to force the hydrogen bond formed between the COO(-) and its partner to remain in the two energy favorable wells, which means the hydrogen bond angle will be like to be observed either around 132.80° or 229.46°. Based on this energy landscape in vacuum and the Boltzmann distribution, a predicted angle-dependence histogram was generated, as Figure 3.9 presents. It shows that the probability of hydrogen bond angle formed between the carboxylic and the methanol is most likely to land at the two favorite angles. In fact, since the hydrogen-bond energy at 229.5° is about 1 kJ/mol lower than the energy at 132.80°, it is predicted that 229.5° is the most likely angle for one hydrogen-bonding interaction formed on COO(-). The predicted probabilities different hydrogen-bond angle are supported by the statistical analysis of

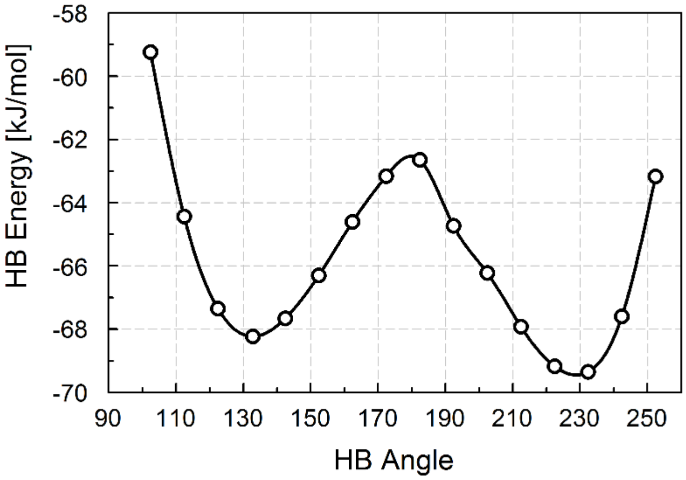


Figure 3.8 Energy landscape for hydrogen-bonding interactions between COO(-) and methanol molecule formed at different hydrogen-bond angles in vacuum.

255 X-ray and neutron-diffraction structures of small molecule carboxylates without metal ions, as previously reported by Gorbitz (Gorbitz and Etter, 1992), and the statistical analysis of 50 protein crystal structures reported by Ippolito (Ippolito et al, 1990).

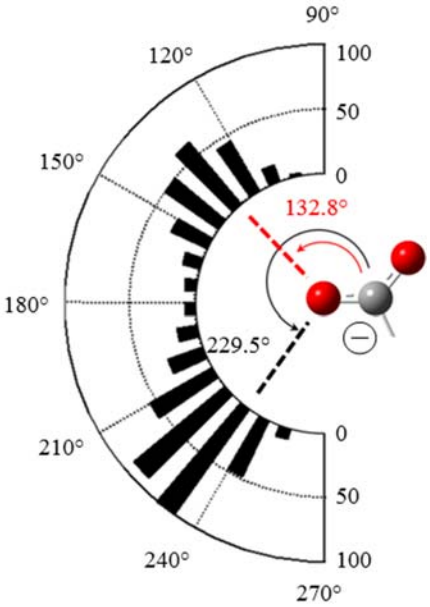


Figure 3.9 Histogram of probabilities for hydrogen-bond angle based on corresponding hydrogen-bond energy. The hydrogen-bond angle (C..O..O, 90° to 270°) between COO(-) and one methanol molecule is fixed in each energy calculation.

3.2.7 Preferences of geometry with the same number of hydrogen-bonding interactions

As data shown in Table 3.9, with same number of hydrogen-bonding interactions formed on carboxylate groups, different geometries of hydrogen-bond complex are observed. When the side chain of Asp/Glu buried in proteins has the freedom to change its position, it should move toward the state with the lowest energy. And in this case, the geometry with the highest hydrogen-bond energy will be the most stable one.

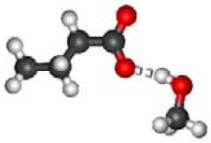
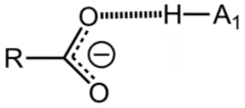
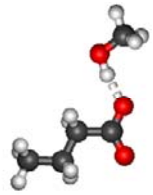
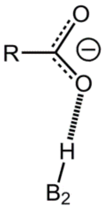
Therefore, using Boltzmann distribution, we analyzed the preferences of complex geometry when the number of hydrogen-bonding interaction is identical.

The predicted probabilities of each geometry for COO(-) forming 1, 2, 3 and 4 hydrogen-bonding interactions are shown in Table 3.5, Table 3.6, Table 3.7 and Table 3.8, respectively. When COO(-) forms one hydrogen-bonding interaction, methanol at B position resulting in 2.3 kJ/mol than at A position. The corresponding probability are 29.1% for A and 70.9% for B. For two hydrogen-bonding interactions, 2D is the most energetically favorite, showing 40.1%. However, this probability might not be accurate, as the optimized structure shows a weak interaction is formed between the carbon of methanol and one of the carboxylate oxygen. And among three geometry in which each carboxylate oxygen accepts a hydrogen-bond interaction, complex 2A has the highest chance to be formed.

For COO(-) forming three hydrogen-bonding interactions, complex 3A and 3B are more likely to be formed, as the probabilities are 35.3% and 64.2%, respectively. Complex 3B is observed with higher hydrogen-bond energy, which is likely due to the weak interaction formed between carbon of methanol and carboxylate oxygen. It should be noted that the complex 3C, which results abnormal vibrational frequencies of asymmetric and symmetric stretching, is extremely unfavorable, showing only 0.2% of probability.

And when four hydrogen-bonding interactions are formed on COO(-), the results imply that most likely the geometry 4A with evenly distributed hydrogen-bonding interactions has highest chance to be formed with 83.0% of probability, while 4B and 4C has predicted probability of 3.4% and 13.6%, respectively.

Table 3.5 Preferences of geometry when carboxylate group forms one hydrogen-bonding interaction

Optimized Structure*	Type of HB	Type of HB (figure)†	H-Bond Energy‡	Probability§
	1A		67.3	29.1%
	1B		69.6	70.9%

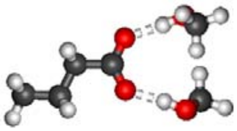
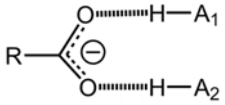
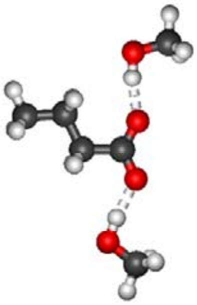
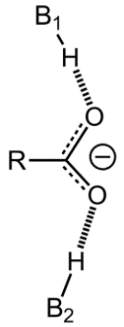
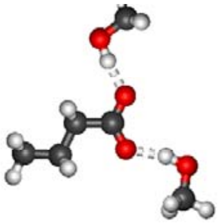
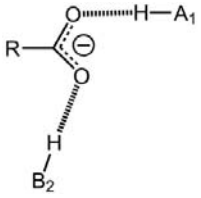
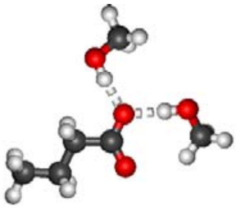
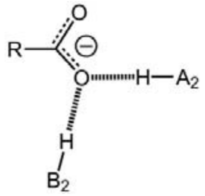
*Computational method B3LYP/6-311+G(3df, 2p) was used for geometry optimization and energy calculation. Color codes for atoms: black for carbon atoms, red for oxygen atoms, white for hydrogen atoms. White dashed lines represent hydrogen-bonding interactions or weak interactions.

†Types of hydrogen-bonding interactions are characterized by the number of hydrogen-bonding interactions formed on deprotonated butyric Asp/Glu side chain in proteins. Hydrogen-bond length is defined as the distance between two heavy atoms paired as hydrogen-bond donor and hydrogen-bond acceptor.

‡Hydrogen-bond energy is calculated in vacuum.

§Probability calculated based on Boltzmann distribution at 25 °C.

Table 3.6 Preferences of geometry when carboxylate group forms two hydrogen-bonding interactions

Optimized Structure*	Type of HB	Type of HB (figure)†	H-Bond Energy‡	Probability§
	2A		130.4	30.4%
	2B		127.8	10.4%
	2C		129.3	19.1%
	2D		131.1	40.1%

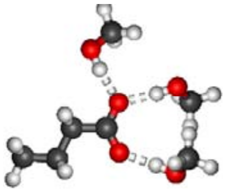
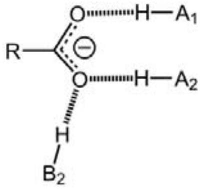
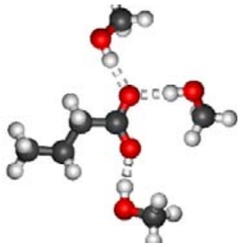
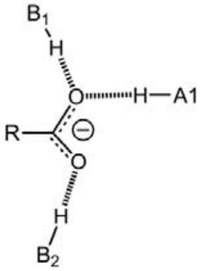
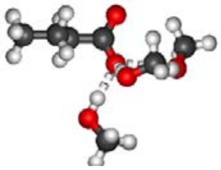
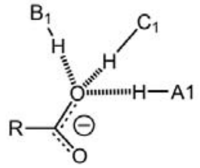
*Computational method B3LYP/6-311+G(3df, 2p) was used for geometry optimization and energy calculation. Color codes for atoms: black for carbon atoms, red for oxygen atoms, white for hydrogen atoms. White dashed lines represent hydrogen-bonding interactions or weak interactions.

†Types of hydrogen-bonding interactions are characterized by the number of hydrogen-bonding interactions formed on deprotonated butyric Asp/Glu side chain in proteins. Hydrogen-bond length is defined as the distance between two heavy atoms paired as hydrogen-bond donor and hydrogen-bond acceptor.

‡Hydrogen-bond energy is calculated in vacuum.

§Probability calculated based on Boltzmann distribution at 25 °C.

Table 3.7 Preferences of geometry when carboxylate group forms three hydrogen-bonding interactions

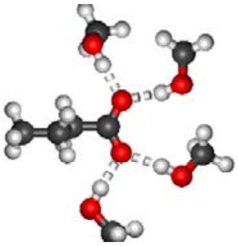
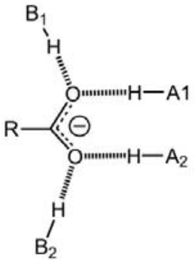
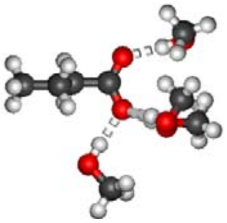
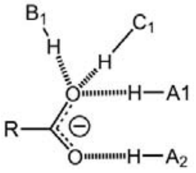
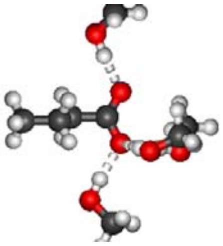
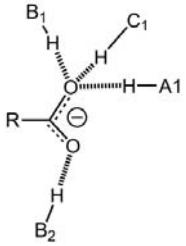
Optimized Structure*	Type of HB	Type of HB (figure)†	H-Bond Energy‡	Probability
	3A		182.8	35.3%
	3B		184.3	64.2%
	3C		169.9	0.2%

*Computational method B3LYP/6-311+G(3df, 2p) was used for geometry optimization and energy calculation. Color codes for atoms: black for carbon atoms, red for oxygen atoms, white for hydrogen atoms. White dashed lines represent hydrogen-bonding interactions or weak interactions.

†Types of hydrogen-bonding interactions are characterized by the number of hydrogen-bonding interactions formed on deprotonated butyric Asp/Glu side chain in proteins. Hydrogen-bond length is defined as the distance between two heavy atoms paired as hydrogen-bond donor and hydrogen-bond acceptor.

‡Hydrogen-bond energy is calculated in vacuum.

Table 3.8 Preferences of geometry when carboxylate group forms four hydrogen-bonding interactions

Optimized Structure*	Type of HB	Type of HB (figure)†	H-Bond Energy‡	Probability
	4A		224.7	83.0%
	4B		216.7	3.4%
	4C		220.2	13.6%

*Computational method B3LYP/6-311+G(3df, 2p) was used for geometry optimization and energy calculation. Color codes for atoms: black for carbon atoms, red for oxygen atoms, white for hydrogen atoms. White dashed lines represent hydrogen-bonding interactions or weak interactions.

†Types of hydrogen-bonding interactions are characterized by the number of hydrogen-bonding interactions formed on deprotonated butyric Asp/Glu side chain in proteins. Hydrogen-bond length is defined as the distance between two heavy atoms paired as hydrogen-bond donor and hydrogen-bond acceptor.

‡Hydrogen-bond energy is calculated in vacuum.

3.2.8 Isotopic labeling effect on the vibrational frequencies.

The vibrational frequencies of asymmetric and symmetric stretching of COO(-) group are strongly affected by isotopic labeling. We performed computational studies of ^{12}C and ^{13}C labeled deprotonated butyric acid in the environment of H_2O. Results are listed in Table 3.9. ^{13}C isotopic labeling on the carbons lead to significant red-shifts on both vibrations, ~40 cm^{-1} for asymmetric stretching and ~30 cm^{-1} for symmetric stretching. Those red-shifts are experimentally friendly, particularly useful for identifying the infrared signals corresponding to asymmetric stretching, since the frequency will shift out of the amide region upon ^{13}C labeling.

3.3 Conclusions

Our statistical analysis of 1182 protein crystal structures revealed that the chance of finding fully buried Asp and Glu residues in hydrophobic protein interior is very rare, only 6.2% of all Asp and 4.5% of all Glu residues. And our survey of hydrogen-bonding network around buried Asp and Glu residues in proteins has shown that most of the Asp and Glu residues are hydrogen-bonded (Table 3.3). In fact, more than half of buried Asp and Glu residues have at least three hydrogen-bonding interactions. Further analysis found that backbone amide is the most common hydrogen-bond donor paring with buried carboxylate groups (Table 3.4).

We performed computational studies on a model system consisting of a deprotonated butyric acid forming one hydrogen-bond interaction with a methanol molecule with different computational method. Our results have shown that B3LYP/6-311+G(3df,2p) gives the best balance between computational accuracy and efficiency.

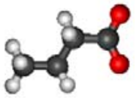
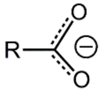
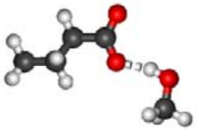
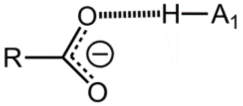
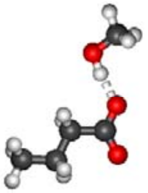
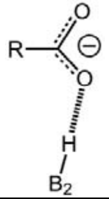
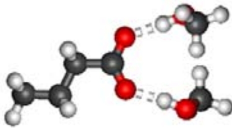
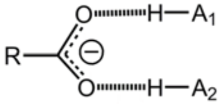
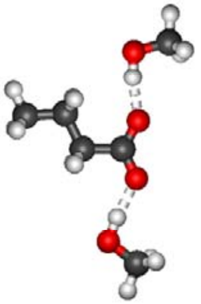
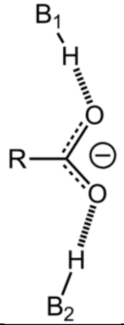
Our frequency calculations of different monocarboxylates have revealed that the vibrational frequencies of asymmetric stretching and symmetric stretching are dependent on the length of carbon chain attached. Particularly large dependence was observed when the monocarboxylate is shorter than propanoic acid (3 carbons total). And butyric acid shows similar frequency results as pentanoic acid and hexanoic acid. Therefore we concluded that butyric acid with 4 carbon atoms in the chain is the model compounds for Asp and Glu residues in protein peptides.

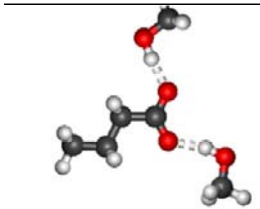
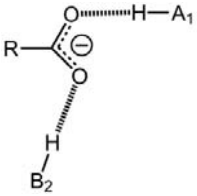
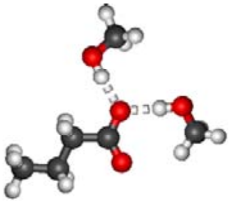
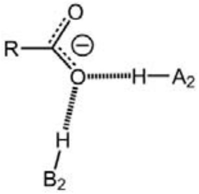
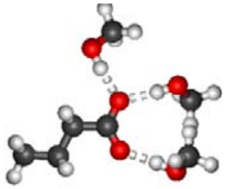
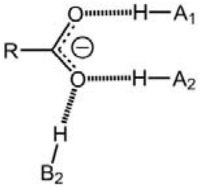
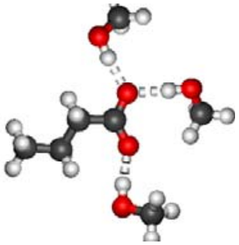
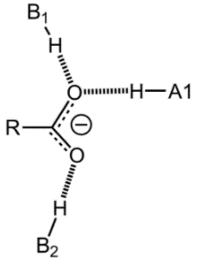
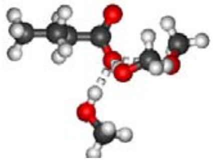
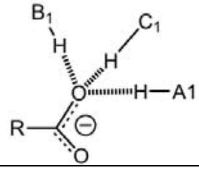
We calculated properties of hydrogen-bonding interactions formed by deprotonated butyric acid with different numbers of methanol molecules. The plotted spectrum clearly indicate that asymmetric stretching and symmetric stretching are the vibrational structural markers which can help distinguish the number and the types of hydrogen-bonding interactions. The optimized structure for model system consisting of one deprotonated butyric acid with 10 methanol molecules showed two layers of hydrogen-bonding network: carboxylate groups forms six hydrogen-bonding interactions with six methanol molecules as the inner layer, and the other four methanol molecules forms inter-methanol hydrogen-bonding interactions for the stabilization of the whole network. This geometry is believed to be a well simulating for the real hydrogen-bonding network in butyric-methanol solution. And computational frequencies fits experimental data extremely well (<0.1% error) after scaled with scaling factor 0.988.

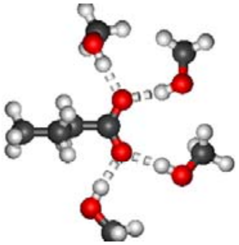
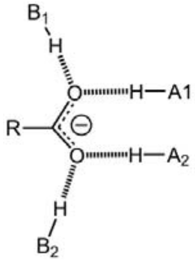
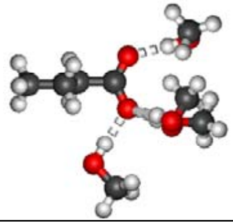
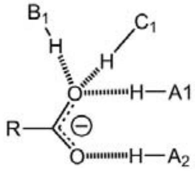
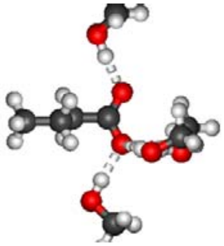
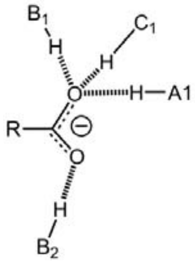
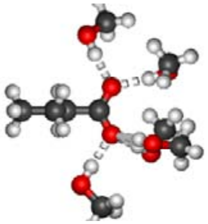
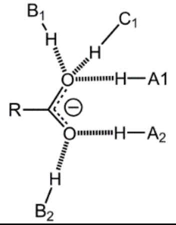
We further discussed the preference of hydrogen-bond length and hydrogen-bond angle for a carboxylate group hydrogen-bonded with a methanol using calculations with fixed parameters. We found that the geometry optimized without any restrictions always give the lowest energy, therefore resulting in the most optimal hydrogen-bond length and hydrogen-bond angle. We also discussed the probabilities of the each potential geometry when deprotonated butyric acid is forming same numbers of hydrogen-bonding interactions with methanol molecules, using the value of hydrogen-bond energy and Boltzmann distribution at 25 °C.

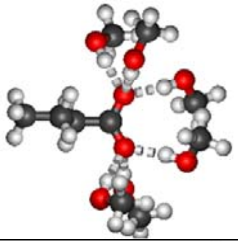
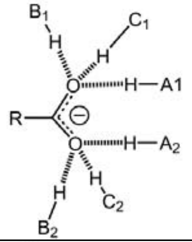
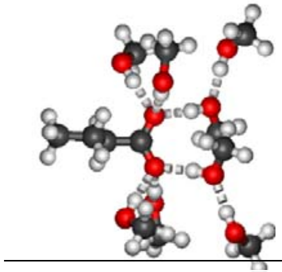
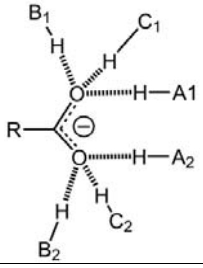
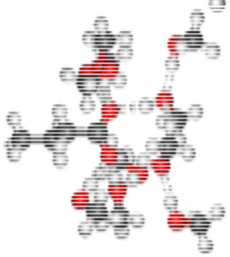
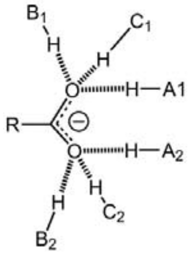
The vibrational structural markers for probing hydrogen-bonding interactions we developed are expected the enhance the power of Time-resolved Infrared Structural Biology for the studies of functionally important dynamics of Asp/Glu residues in proteins, including proton transfer in bacteriorhodopsin and PYP, and enzymatic catalysis reactions in enzyme proteins such as HIV protease.

Table 3.9 Calculated hydrogen-bond properties of depronated Asp/Glu side chain forming hydrogen-bond interactions with different number of methanol molecules.

Optimized Structure*	Type of HB/ # of Methanol†	Type of HB (figure)†	HB Length† (Å)	Total H-bond Energy‡ (kJ/mol)	Avg Energy per H-Bond‡ (kJ/mol)	ν_{asym} ^{12}C (cm^{-1}) §	ν_{sym} ^{12}C (cm^{-1}) §	ν_{asym} ^{13}C (cm^{-1}) §	ν_{sym} ^{13}C (cm^{-1}) §
	0/0		NA	NA	NA	1633.1	1343.6	1587.9	1322.5
	1A/1		2.67	67.3	67.3	1638.6	1356.6	1596.3	1332.1
	1B/1		2.63	69.6	69.6	1632.4	1354.9	1587.9	1328.1
	2A/2		A$_1$: 2.71 A$_2$: 2.70	130.4	65.2	1618.1	1378.2	1573.9	1349.0
	2B/2		B$_1$: 2.68 B2: 2.66	127.8	63.9	1614.9	1381.2	1570.3	1352.6

Optimized Structure*	Type of HB/ # of Methanol†	Type of HB (figure)†	HB Length† (Å)	Total H-bond Energy‡ (kJ/mol)	Avg Energy per H-Bond‡ (kJ/mol)	ν_{asym} ^{12}C (cm^{-1}) §	ν_{sym} ^{12}C (cm^{-1}) §	ν_{asym} ^{13}C (cm^{-1}) §	ν_{sym} ^{13}C (cm^{-1}) §
	2C/2		A_1: 2.70 B_2: 2.67	129.3	64.6	1622.2	1380.5	1579.6	1351.6
	2D/2		A_2: 2.73 B_2: 2.69	131.1	65.6	1635.4	1368.2	1592.8	1341.2
	3A/3		A_1: 2.73 A_2: 2.74 B_2: 2.73	182.8	60.9	1618.5	1388.4	1574.9	1357.5
	3B/3		A_1: 2.75 B_1: 2.72 B_2: 2.70	184.3	61.4	1610.1	1394.6	1567.5	1364.0
	3C/3		A_1: 2.78 B_1: 2.72 C_1: 2.79	169.9	56.6	1641.7	1344.8	1599.9	1317.0

Optimized Structure*	Type of HB/ # of Methanol†	Type of HB (figure)†	HB Length† (Å)	Total H-bond Energy‡ (kJ/mol)	Avg Energy per H-Bond‡ (kJ/mol)	ν_{asym} ^{12}C (cm^{-1}) §	ν_{sym} ^{12}C (cm^{-1}) §	ν_{asym} ^{13}C (cm^{-1}) §	ν_{sym} ^{13}C (cm^{-1}) §
	4A/4		A_1: 2.78 B_1: 2.75 A_2: 2.76 B_2: 2.73	224.7	56.2	1602.2	1400.4	1559.8	1370.9
	4B/4		A_1: 2.84 B_1: 2.76 C_1: 2.82 A_2: 2.74	216.7	54.2	1612.2	1376.4	1570.0	1346.7
	4C/4		A_1: 2.81 B_1: 2.76 C_1: 2.84 B_2: 2.71	220.2	55.0	1607.1	1381.0	1565.3	1351.1
	5/5	 B1 C1 H H O H—A1 R ⊖ O H—A2 H B2	A_1: 2.86 B_1: 2.78 C_1: 2.85 A_2: 2.77 B_2: 2.76	260.9	52.2	1588.7	1398.4	1546.7	1368.6

Optimized Structure*	Type of HB/ # of Methanol†	Type of HB (figure)†	HB Length† (Å)	Total H-bond Energy‡ (kJ/mol)	Avg Energy per H-Bond‡ (kJ/mol)	ν_{asym} ^{12}C (cm^{-1}) §	ν_{sym} ^{12}C (cm^{-1}) §	ν_{asym} ^{13}C (cm^{-1}) §	ν_{sym} ^{13}C (cm^{-1}) §
	6A/6	 B1 C1 H H O H—A1 R ⊖ O H—A2 H H C2 B2	A_1: 2.81 B_1: 2.83 C_1: 2.89 A_2: 2.82 B_2: 2.81 C_2: 2.89	288.4	48.1	1574.8	1401.9	1534.2	1372.4
	6B/8	 B1 C1 H H O H—A1 R ⊖ O H—A2 H H C2 B2	A_1: 2.76 B_1: 2.85 C_1: 2.91 A_2: 2.76 B_2: 2.83 C_2: 2.91	NA	NA	1565.4	1404.1	1528.1	1374.8
	6C/10	 B1 C1 H H O H—A1 R ⊖ O H—A2 H H C2 B2	A_1: 2.75 B_1: 2.83 C_1: 2.89 A_2: 2.76 B_2: 2.84 C_2: 2.84	NA	NA	1565.4	1401.8	1528.1	1372.3

*Computational method B3LYP/6-311+G(3df, 2p) was used for geometry optimization, energy and frequency calculation. Color codes for atoms: black for carbon atoms, red for oxygen atoms, white for hydrogen atoms. White dashed lines represent hydrogen-bonding interactions or weak interactions.

†Types of hydrogen-bonding interactions are characterized by the number of hydrogen-bonding interactions formed on deprotonated butyric Asp/Glu side chain in proteins. Hydrogen-bond length is defined as the distance between two heavy atoms paired as hydrogen-bond donor and hydrogen-bond acceptor.

‡Hydrogen-bond energy is calculated in vacuum.

§Vibrational frequency are scaled with scaling factor 0.988.

CHAPTER 4

VSMs for Probing Hydrogen-Bonding Interactions of Tyrosine in Proteins

4.1 Introduction

Tyrosine (abbreviated as Tyr or Y), an amino acid with a phenolic aromatic side chain, is widely found at the active sites of proteins. Tyrosine residues have been reported to play key roles directly or indirectly in fundamental biological processes, including electron transfer (Kuhne and Brudvig, 2002, Barry and Babcock, 1987), proton transfer (Nie, 2006, Rothschild et al, 1990), and phosphorylation (Hunter and Cooper, 1985), to enable a broad range of biological functions of proteins. The phenolic side chain of tyrosine has a p4.1K_a of 10.07 as an individual amino acid in solution (Lehninger et al, 2005) or ~10 as part of protein peptide exposed to solvent (Oktaviani et al, 2012). Meanwhile, buried Tyrosine side chains have been reported to have a much larger range of pK_a values, as low as ~6-7 (Schwans et al, 2013) and as high as or larger than 13 (Oktaviani et al, 2012, Baturin et al, 2011, Asher et al, 1991). Most tyrosine residues are protonated in proteins at physiological environment. The hydroxyl group (-OH) of protonated tyrosine can form hydrogen-bonding interactions as hydrogen-bond donor and hydrogen-bond acceptor. The number, type and strength of hydrogen-bonding interactions are expected to affect the biological activities of tyrosine during protein functions.

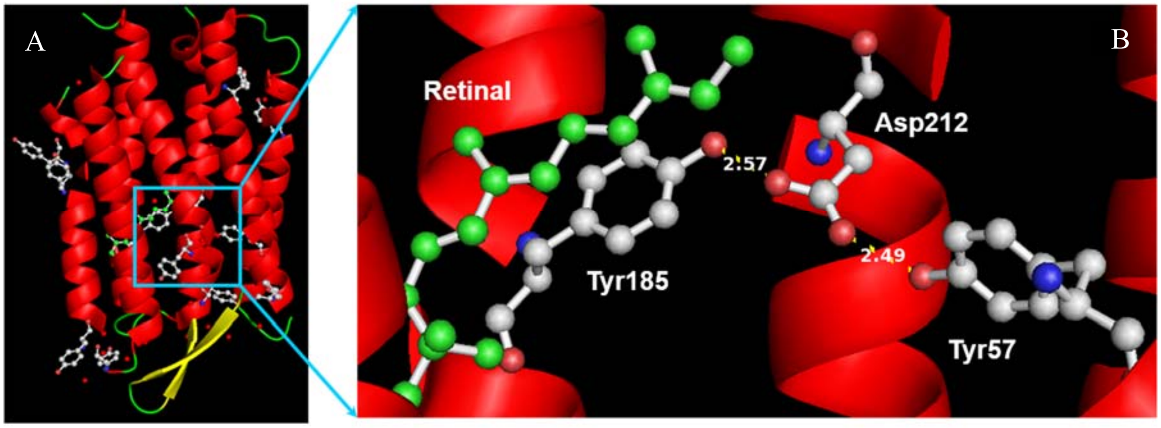


Figure 4.1 Crystal structure of bacteriorhodopsin in the bR state (PDB: 1C3W, resolution 1.55Å). (A) Crystal structure of whole protein (B) Hydrogen-bonding interactions between Tyr185, Tyr57 and Asp212.

In bacteriorhodopsin (bR), a bacterial light-driven proton pump for solar energy transduction (Oesterhe.D and Stoecken.W, 1973), 11 tyrosine residues are found out of 248 amino acids which fold into a transmembrane protein with seven transmembrane alpha helices (Luecke et al, 1999). The high resolution crystal structure of bR$_{568}$ reveals that Tyr185 and Tyr57 (shown in Figure 4.1) are structurally active, stabilizing the negatively charged carboxylic group (COO-) of Asp212 via two strong ionic hydrogen-bonding interactions at bR$_{568}$ (Luecke et al, 1998) and M$_{412}$ state (Ames et al, 1992). Mutation of Tyr185 to Phe185 in the Y185F mutant resulted in protonated carboxylic group (COOH) of Asp212 (Rothschild et al, 1990, Bousche et al, 1992, Sonar et al, 1993). And mutation of Tyr57 to Phe57 or Asn57 in the Y57F or Y57N mutant lead to no proton release, consequently no proton uptake process in bacteriorhodopsin (Govindjee et al, 1995).

In Photoactive Yellow Protein (PYP), a blue-light bacterial photoreceptor (Meyer, 1985, Meyer et al, 1987), Tyr42 interacts directly with the negatively charged phenolic group of the light sensing pCA chromophore at the active site, via strong, ionic hydrogen-bonding interaction as hydrogen-bond donor, as shown in Figure 4.2 (Borgstahl et al, 1995, Getzoff et al, 2003). Tyr42 is found largely conserved in the family of photoactive yellow proteins (Kumauchi et al, 2008). Mutation of Tyr42 to Phe or Ala resulted in the protein bleach due to the disruption of the hydrogen-bonding interactions between Tyr42

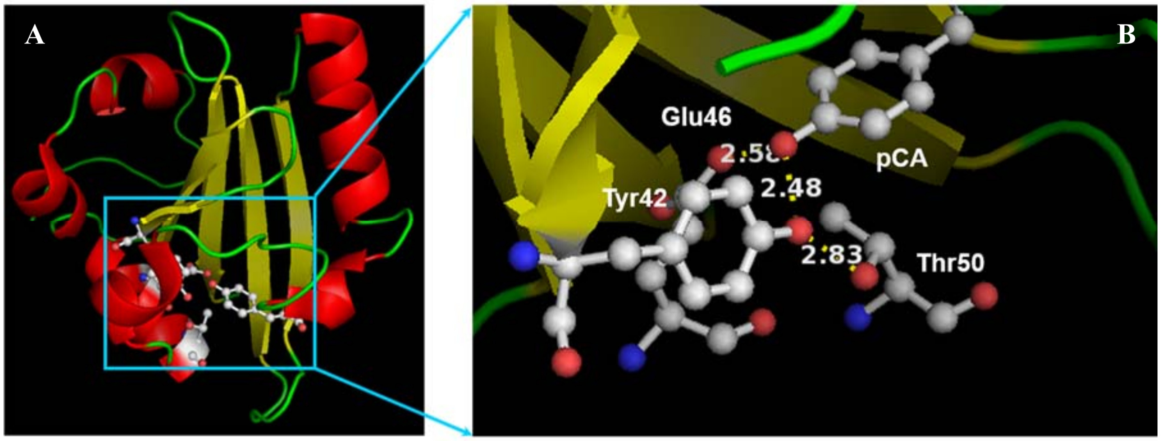


Figure 4.2 Crystal structure of Photoactive Yellow Protein in the pG state (PDB: 1NWZ, resolution 0.82Å). (A) Crystal structure of whole protein (B) Hydrogen-bonding interaction network around Tyr42 in active site.

and pCA chromophore, which stabilizes the deprotonation state of the pCA chromophore and protein conformation (Philip et al, 2010, Imamoto et al, 2001, Brudler et al, 2000). The recent studies on the proton transfer mechanism of PYP (Nie, 2006) suggest that cleavage of the Tyr42-pCA hydrogen-bonding interaction triggers proton transfer from the proton donor, Glu46, to the proton acceptor, the phenolic group of the chromophore. Direct experimental testing of this novel proton transfer mechanism in PYP will require the use of time-resolved step-scan FTIR combined with VSMs. This combination allows the detection of dynamic changes in the hydrogen-bonding interactions of Tyr42 during the receptor activation process upon absorption of a blue photon.

Photosystem II is an enormous protein complex (MW 650 kDa) widely found in the thylakoid membranes of oxygenic photosynthetic organisms including plants, algae and cyanobacteria (McEvoy and Brudvig, 2006). Each Photosystem II monomer is composed of around 20 subunits (depends on organisms) and many cofactors. The high resolution crystal structure of Photosystem II is shown in Figure 4.3 (Umena et al, 2011). Water oxidation happens in Photosystem II reaction center, which contains subunit D1, D2, CP47 and cytochrome b-559. The core of Photosystem II reaction center is the Oxygen-Evolving Complex, consisting of 4 manganese, 5 oxygen and 1 calcium. The Oxygen-

Evolving Complex acts as an electrical accumulator. In one complete cycle, it is oxidized four times by one electron at a time from P_{680}^{+}, and then the Oxygen-Evolving Complex oxidizes water to oxygen (Kok et al, 1970). Two tyrosine residues, named Tyr-Z and Tyr-D, are found in oxidizing subunit D1 and D2 in the Photosystem II reaction center (Hienerwadel et al, 1997, Berthomieu et al, 1998, Tang et al, 1993). Particularly, Tyr-Z in subunit D1 sits close to Oxygen-Evolving Complex and functions as a redox-coupled base which electronically links P_{680}^{+} to Oxygen-Evolving Complex (Gilchrist et al, 1995, de Wijn and van Gorkom, 2002, Styring et al, 2012). The recent high resolution crystal structure also revealed the hydrogen-bonding interaction between Tyr-Z and His190 in subunit D1 and hydrogen-bonding interaction between Tyr-D and His189 in subunit D2 (Umena et al, 2011). Both interactions are proposed to mediate proton shuffling with paired histidine partner, which couples electron transfer for the water oxidation process in Photosystem II (Kuhne and Brudvig, 2002, Jenson and Barry, 2009).

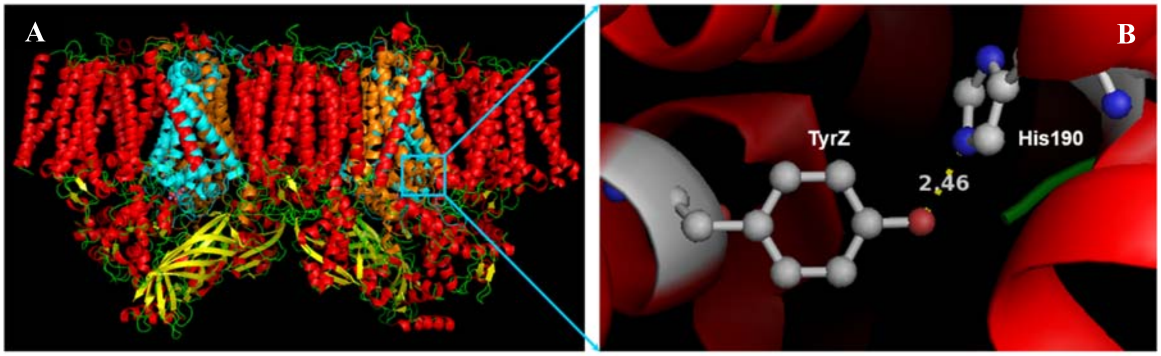


Figure 4.3 Crystal structure of Photosystem II in the S1 state (PDB: 3ARC, resolution 1.90Å). (A) Crystal structure of whole protein dimer (B) Hydrogen-bonding interaction between Tyr-Z and His190 in subunit D1.

Hydrogen-bonding interactions, particularly between buried polar groups, are important structural elements for proteins (Takano et al, 2003). It was reported that buried polar side chain which are hydrogen-bonded are more conserved than those side chains that are not (Worth and Blundell, 2009). Depending on the microenvironment around, contributions to protein stability from each individual hydrogen-bonding interactions are different. Recent studies have found that the loss of exposed, ordinary hydrogen-bonding interactions could increase an energy of approximately 2~10 kJ/mol, while the loss of some key buried hydrogen-bonding interactions could greatly increase the energy, approximately 10-40 kJ/mol (Pace, 2001, Pace, 2009, Pace et al, 2014). Meanwhile, protein folding energy is reported to be in the range of approximately 20-40 kJ/mol (Creighton, 1993). Therefore, the dissociation energy of hydrogen-bonding interactions and the protein folding energy are similar on the order that breaking of hydrogen-bonding interactions is capable of significantly destabilizing the protein folding.

Theoretically, the phenolic group of a tyrosine side chain can form 3 hydrogen-bonding interactions, 1 as donor, 2 as acceptor, as illustrated in Figure 4.4. Exposed side chain of a tyrosine residue may form as many hydrogen-bonding interactions as allowed. When side chain of tyrosine is buried inside a protein, lack of hydrogen-bond partners may significantly reduce the number of hydrogen-bonding interactions. In addition, during functionally important motions of proteins, the number, type, and strength of hydrogen-bonding interactions of tyrosine can change dramatically. That means the hydrogen-bonding status of the side chain of tyrosine may change from one intermediate state to another. Such structural changes may be crucial for protein functions.

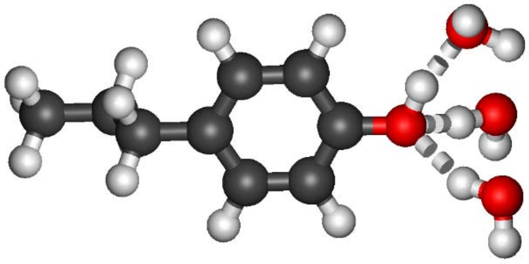


Figure 4.4 Tyrosine side chain forms maximum of three hydrogen-bonding interactions with three H2O molecules: one as hydrogen-bond donor, two as hydrogen-bond acceptor.

Time-resolved Fourier Transformed Infrared Spectroscopy (Time-resolved FTIR) on proteins is a powerful technique to detect structural information because of its excellent sensitivity to side chain protonation/deprotonation, hydrogen-bonding interactions and outstanding accessibility to dynamics. However, extracting key structural details out of infrared spectrum still remains challenging, because the structural information encoded is so rich and very few guidelines have been reported. As a result, Vibrational Structural Markers (VSM) for probing hydrogen-bonding interactions of tyrosine side chain are urgently needed for researchers to fully understand the structures of tyrosine residues in proteins based on infrared spectrum.

Such VSM for tyrosine side chain was first reported by Beining Nie (Nie, 2006). In her DFT studies, computational method B3LYP/6-31G(d) was used on 4-propylphenol, a model compound of tyrosine. The results showed that the vibrational frequencies of C-O stretching ($\nu_{7a'}$) and C-O-H bending (δ_{COH}) are most sensitive to hydrogen-bonding interactions and are assigned to be the VSM. The two vibrational modes are illustrated in Figure 4.5. However, selecting a considerably simple method B3LYP/6-31G(d) was mainly due to the limitation of computational power at the time. Although studies using this method has shown quite accurate frequency predictions on neutral groups (Nie et al, 2005), aromatic ring of tyrosine side chain makes calculations more computationally demanding because the electrons are preferably delocalized over the ring. In addition, many buried tyrosine residues in protein are hydrogen-bonded with charged groups such as COO(-) from Asp/Glu or NH(+) from Arg/Lys/His. Therefore, more advanced computational method, which includes diffusion functions, should be used to study properties of those ionic hydrogen-bonding interactions for more accurate results.

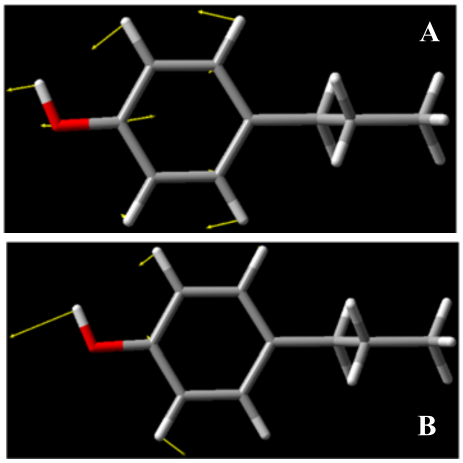


Figure 4.5 Normal vibrations of ring-H4 4-propylphenol with zero hydrogen-bonding interaction. (A) C-O stretching ($\nu_{7a'}$); (B) C-O-H bending (δ_{COH}).

Another DFT studies reported by Takahashi and Noguchi (Takahashi and Noguchi, 2007) also suggested that frequencies of C-O stretching ($\nu_{7a'}$) and C-O-H bending (δ_{COH}) are indicators of hydrogen-bonding status of tyrosine. In their studies, method B3LYP/6-31+G(d,p) was employed on p-cresol molecules with multiple types of hydrogen-bonding interactions. Their computational method is a little more advanced than B3LYP/6-31G(d), as diffusion functions are added to heavy atoms and p functions are added to hydrogen atoms. However, the hydrogen-bond partners selected in their studies are mainly organic solvent molecules. This means their results are more reliable for the hydrogen-bonding interactions of p-cresol in organic solvent, but not as convincing for probing hydrogen-bonding interactions of tyrosine side chain in proteins. Also, p-cresol only has one carbon on the chain in addition to the aromatic ring. While for real tyrosine residues in proteins, the carbon chain extends much longer with the peptide. Therefore, we employed 4-propylphenol as model compound of tyrosine side chain in our calculations.

Furthermore, both VSMs reported by Nie and Takahashi were for ring-H_4 (non-isotopic labeled) tyrosine residues. As shown in their data, when tyrosine forms one hydrogen-bonding interaction as donor and forms two hydrogen-bonding interactions, the vibrational frequencies of $\nu_{7a'}$ and δ_{COH} are not separated enough, as the differences between the two frequencies are less than 20 cm^{-1} with some hydrogen-bond partners. Such complication could make the two vibrational modes easily coupled and result in a broad band for tyrosine in real infrared spectrum of proteins. To better distinguish the two vibrational modes, especially in real infrared spectrum, we employed ring-D4 isotopic labeling of the tyrosine in this chapter, as our results will show a better separation in frequencies between the two vibrational modes, and makes it much easier to apply VSM to infrared spectrum of proteins.

In fact, ring-D4 isotopic labeling of tyrosine residues in protein has been reported previously as an efficient method to identify tyrosine signals in infrared spectrum of proteins (Hienerwadel et al, 1997, Rathod et al, 2012, Liu et al, 1995). Particularly for functionally important tyrosine residues such as Tyr185 and Tyr57 in bR, Tyr42 in PYP and Tyr-Z, Tyr-D in PSII, the site-directed mutations of those residues significantly change the biological functions of the protein, while site-specific isotopic labeling could be the perfect way to locate their infrared signals without affecting the proteins functions. However, for ring-D4 isotopic labeled tyrosine, no VSM has been reported yet. In this chapter, I will present our work of using DFT calculations to develop VSMs for probing hydrogen-bonding interactions of tyrosine in proteins. And the VSM we reported is specifically for ring-D4 isotopic labeled tyrosine, because this isotopic labeling eliminates the coupling between the C-O stretching and C-O-H bending, which leads to much clearer distinguishing of the two vibrational modes than non-isotopic labeled tyrosine. The VSM we report here can be used to identify detailed hydrogen-bonding information. Not only the number of hydrogen-bonding interactions, it could also help to characterize the types of hydrogen-bonding interactions, whether it is ordinary neutral hydrogen-bonding interaction, or it is strong ionic hydrogen-bonding interaction.

4.2 Materials and Methods

4.2.1 Frequency and Energy Calculation

All DFT calculations were carried out using Gaussian 09, Revision C.01 (Frisch, 2009) installed on Cowboy Cluster in the OSU High Performance Computing Center. GaussView 5.0 (Dennington, 2009) was used to generate input structures and visualize output structures with vibrational mode frequencies. Procedures of calculation for structural optimization, energy and frequency calculation in Gaussian 09 are the same as calculation in Gaussian 03 previously reported (Nie et al, 2005), except basis set 6-311+G(3df,2p) is selected through all computational methods for its excellent balance of good accuracy and reasonable efficiency. The detailed procedures of selecting this basis set have been discussed in Chapter 3. This basis set contains three sizes of contracted functions for each orbital type, with 3 d, 1 f functions as well as diffusion functions being added to heavy atoms and 2 p function being added to hydrogen atoms. Hybrid functional method B3LYP is used for DFT calculation.

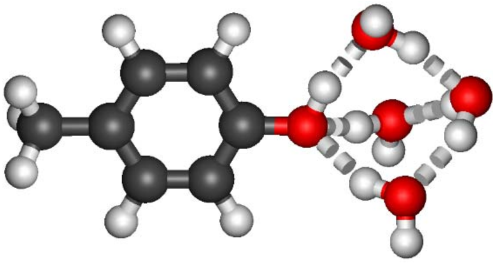


Figure 4.6 p-Cresol molecule forming 3 hydrogen-bonding interactions with 3 H2O, one extra H2O was added for stabilization of the hydrogen-bonding network.

A comparison of the three computational methods, B3LYP/6-31G(d) used by Nie (Nie, 2006), B3LYP/6-31+G(d,p) used by Takahashi (Takahashi and Noguchi, 2007) and B3LYP/6-311+G(3df,2p), is shown in Table 4.1. The comparison used the computational results of one p-cresol forming 3 hydrogen-bonding interactions with 3 H_2O molecules, and an extra H_2O was added to stabilize the three hydrogen-bonding structure. Such formation models the environment around p-cresol in H_2O solution more accurately than just placing two or three H_2O molecules, because the maximum hydrogen-bonding potential of p-Cresol in H_2O is satisfied. The arrangement of hydrogen-interactions optimized by three different computational methods are very similar, and the output geometry optimized by B3LYP/6-311+G(3df,2p) is shown in Figure 4.6.

Table 4.1 Comparison of different computational methods

Method	# of functions		Diffuse functions on 1st row atoms?	Calc $\nu_{7a'}$	Exp† $\nu_{7a'}$	Calc δ_{COH}	Exp† δ_{COH}	Calc δ_{CH}	Exp† δ_{CH}
	1st row atoms	hydrogen atoms							
B3LYP/6-31G(d)	15	2	No	1232.5	1240	1271.5	1261	1162.2	1176
B3LYP/6-31+G(d,p)	19	5	Yes	1242.6		1279.2		1171.2	
B3LYP/6-311+G(3df,2p)	39	9	Yes	1240.5		1271.5		1176.3	

†Experimental infrared spectroscopy of p-cresol solution in H_2O reported by Takahashi and Noguchi (Takahashi and Noguchi, 2007).

As summarized in Table 4.1, B3LYP/6-31G(d) is the simplest method, using the fewest number of functions, and results in the worst accuracy in frequency results comparing to experimental data. Meanwhile, B3LYP/6-311+G(3df,2p) is the most advanced method compared, as the number of functions it applied per heavy atoms or per hydrogen atoms are almost doubled from B3LYP/6-31+G(d,p). As a result, method B3LYP/6-311+G(3df,2p) generated the most accurate frequency output. Although B3LYP/6-311+G(3df,2p) is a considerably expensive method, much improved computational power of OSU's supercomputing cluster "Cowboy" allows us to finish most of our calculations within two days using this method. Therefore, we chose B3LYP/6-311+G(3df,2p) as our computational method for this chapter of thesis because of its balance between accuracy and efficiency.

We did not include the solvent effects in our calculation. Instead we explicitly employed solvent molecules as hydrogen-bond partners. The reason was discussed in Nie's previous work (Nie et al, 2005) as well as Chapter 3 of this thesis. Besides, model compounds act as the backbone along with polar and charged amino acids were utilized to serve as hydrogen-bond donor and/or acceptor for

Gaussian 09 calculations of hydrogen-bonding properties of tyrosine interacting with these residues in proteins.

4.2.2 Selection of Protein Crystal Structures

Protein Data Bank (PDB, (Berman et al, 2000)) stores all reported protein crystal structures, more than 98900 in April 2014. In order to access the statistical occurrences of various types of hydrogen-bonding interactions of buried tyrosine residues, we selected and downloaded 1182 protein crystal structures from RCSB Protein Databank. The selection criteria were: (1) Monomer proteins without binding of DNA or RNA (2) Only protein structures detected using X-ray diffraction as experimental method; (3) Structure resolution between 0.40 Å and 1.50 Å; (4) R-free value of 0.20 or better (R value is the measure of the quality of the atomic model obtained from the crystallographic data (Kleywegt and Jones, 1997)); (5) Minimum molecular weight 9000 Da (approximately 80 amino acid residues at least); (6) Sequence identity between each selected proteins are less than 30%. Since each protein crystal structure is composed of many individual molecules packed into a symmetrical arrangement (Berman et al, 2000), slight differences in the position of those individual molecules may result in multiple conformations of protein crystal structures. Therefore, downloaded PDB files were first treated with a UNIX script (Appendix C) which separated multiple conformations and saved to edited PDB files with only one conformation of the structure.

4.2.3 CCP4 for bioinformatics analysis of buried side chains

The AREAIMOL function of CCP4 software package, version 6.4 (Lee and Richards, 1971, Saff and Kuijlaars, 1997) was employed to calculate solvent accessible area for each individual atom in protein in order to determine its solvent accessibility. A UNIX script (Appendix C) was used on the AREAIMOL output files to collect the solvent exposure values for phenolic oxygen of tyrosine residues only. The minimum solvent accessible area for the oxygen atom in a phenolic group of tyrosine is 0, indicating fully buried. The maximum accessible area for a phenolic oxygen atom is 54 for extremely exposed phenolic group.

4.2.4 HBPlus

HBPlus software package (v3.06) (McDonald and Thornton, 1994) was used to analyze the hydrogen-bonding interactions within each selected protein crystal structure. 3.20 Å was set to be the maximum hydrogen-bonding length during HBPlus analysis. The output data was first treated with a UNIX script (Appendix C) to collect only hydrogen-bonding interactions involved with phenolic oxygen of tyrosine. Then the data was saved and exported to spreadsheet for statistical analysis.

4.2.5 Further data analysis using spread sheet

Statistical analysis of hydrogen-bonding interactions of buried tyrosine residues were performed using Microsoft Excel 2013. Among all hydrogen-bonding interactions, only those involved side chains of tyrosine residues were selected. Then every residue was assigned a name code in the format of "XXXXA123BCD", as "XXXX" refers to PDB ID of the protein, "A" refers to the chain number, "123" refers to residue number, and "BCD" refers to residue name (for amino acid residue, 3-letter abbreviation was used). Each buried tyrosine residues were assigned name codes in the same format. The list of buried tyrosine residues was then searched by using "VLOOKUP" function of Excel, through the list of hydrogen-bonding interactions in which tyrosine side chain is the hydrogen-bond donor, and through the list of hydrogen-bonding interactions in which tyrosine side chain is the hydrogen-bond acceptor, individually.

4.3 Results and Discussions:

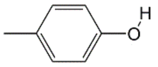
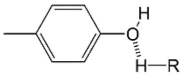
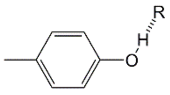
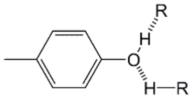
4.3.1 Buried tyrosine in proteins

From 1182 selected protein crystal structures, we found a total of ~14549 tyrosine residues, approximately ~12 tyrosine per protein. 26.1% of these tyrosine residues (3790 out of 14549) have fully buried phenolic group, equivalent of 3 tyrosine residues per protein. Consistent with the low solubility of phenolic ring, our results show that tyrosine phenolic group is quite hydrophobic, comparing to other polar amino acid residues such as Asp (6.2% fully buried) and Glu (4.5% fully buried), as discussed in Chapter 3.

4.3.2 Hydrogen-bonding statistics of fully buried tyrosine residues in proteins

We employed HBPlus software package (v3.06) (McDonald and Thornton, 1994) to analyze the hydrogen-bonding interactions of fully buried phenolic group of tyrosine, as the results are summarized in Table 4.2. The results show that fully buried tyrosine residues in proteins are most likely be a hydrogen-bond donor, as 45.8% of all fully buried tyrosine residues may form one hydrogen-bonding interaction as donor only, and 49.1% may form hydrogen-bonding interactions as donor or as acceptor. It means that overall ~94.9% of all fully buried tyrosine residues may donate a hydrogen-bonding interaction. Meanwhile, 2.3% of the fully buried tyrosine residues may be hydrogen-bond acceptor only, with 49.1% that may form donor or acceptor, resulting overall 51.4% of all fully buried tyrosine residues are found capable of accepting at least one hydrogen-bonding interaction. Only 2.9% are completely isolated from environment, forming no hydrogen-bonding interaction at all. Such results are similar to previous data reported by McDonald and Thornton, that tyrosine is a hydrogen-bond donor almost twice as often as it is a hydrogen-bond acceptor (McDonald and Thornton, 1994). However, our studies are more extensive as our data base has 1182 high-resolution (higher than 1.50 Å) protein structures comparing to the previous report (57 protein structures with resolution higher than 2.00 Å).

Table 4.2 Statistics of hydrogen-bonding interactions of fully buried tyrosine residues in proteins

Role of Tyr in HB		# of Tyr	Percent
No hydrogen-bond		109	2.9%
Acceptor		87	2.3%
Donor		1734	45.8%
Donor/Acceptor/Donor & Acceptor		1860	49.1%
Total		3790	100%

4.3.3 Hydrogen-bonding partners for fully buried tyrosine residues in proteins

As data in Table 4.2 have shown, 97.1% (3681 out of 3790) of the tyrosine residues with phenolic group fully buried are hydrogen-bonded. We also analyzed the preferences of hydrogen-bond partners for the side chain of fully buried tyrosine residues, as the results are listed in Table 4.3. A corresponding histogram was plotted based on the data of Table 4.3, shown in Figure 4.7. Due to missing of proton positions in most of the protein crystal structures, it is difficult to determine tyrosine's role, whether as donor or as acceptor, in a hydrogen-bonding interaction, if the corresponding hydrogen-bond partner may serve both as hydrogen-bond donor and as hydrogen-bond acceptor.

Results summarized in Table 4.3 indicate that total of 5065 potential hydrogen-bond pairs are found between tyrosine side chain and a hydrogen-bond partner. That's approximately 1.4 hydrogen-bonding interactions per tyrosine residue. Protein backbone C=O is found to be the most common hydrogen-bond partner for tyrosine, as tyrosine only serves as donor, counting for 27.0% of all pairs. Water molecules are found to be the second favorite, as it counts for 17.0% of all hydrogen-bonding interactions, while tyrosine can be either donor or acceptor as it is hydrogen-bonded with H_2O. Protein backbone –NH is another favorite hydrogen-bonding for tyrosine to pair with, while tyrosine only serves as hydrogen-bond acceptor, counting for 9.3% of all pairs found. Meanwhile, Asp and Glu residues are the most common hydrogen-bond partners from side chain of amino acid residues, counting for 10.8% and 8.2% of all hydrogen-bonding interactions. Because of the side chain group COO(-), tyrosine only serves as hydrogen-bond donor when pairs with Asp/Glu.

Table 4.3 Statistics of hydrogen-bond partners for fully buried tyrosine residues in proteins

HB Partner of Tyr		HB Group	Role of Tyr in HB	# of Pairs	Percent
Backbone		C=O	Donor	1366	27.0%
Backbone		-NH	Acceptor	472	9.3%
H_2O		-OH	Donor/Acceptor	859	17.0%
Side chain of Amino Acid Residues	Asp	COO(-)	Donor	549	10.8%
	Glu	COO(-)	Donor	417	8.2%
	Asn	C=O/-NH	Donor/Acceptor	197	3.9%
	Gln	C=O/-NH	Donor/Acceptor	173	3.4%
	His	-NH/-NH(+)	Donor/Acceptor	169	3.3%
	Arg	-NH(+)	Acceptor	160	3.2%
	Lys	-NH(+)	Acceptor	59	1.2%
	Trp	-NH	Acceptor	41	0.8%
	Ser	-OH	Donor/Acceptor	176	3.5%
	Thr	-OH	Donor/Acceptor	154	3.0%
	Tyr	Phenol -OH	Donor/Acceptor	97	1.9%
	Cys	-SH	Donor/Acceptor	4	0.1%
Other Residues		various	Donor/Acceptor	172	3.4%
Total				5065	100%

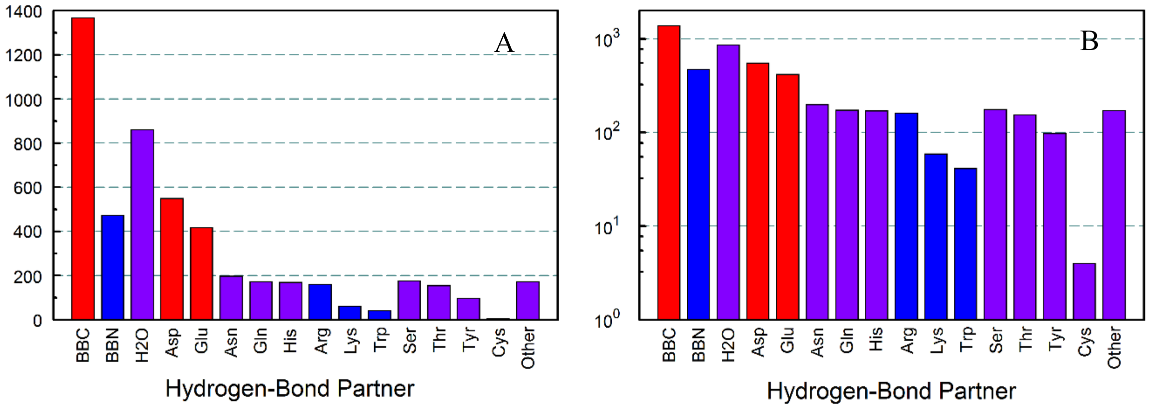


Figure 4.7 Histogram showing statistics of hydrogen-bond partners for buried tyrosine side chain found in 1182 protein crystal structures.(shown in Table 4.3) (A)Y-axis in linear scale (B) Y-axis in logarithmic scale; Color code: Red-tyrosine as hydrogen-bond donor; Blue-tyrosine as hydrogen-bond acceptor; Purple-tyrosine as hydrogen-bond donor or hydrogen-bond acceptor; BBC=Backbone Carbonyl (C=O), BBN=Backbone Amide (-NH)

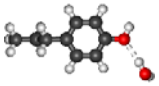
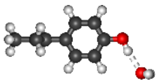
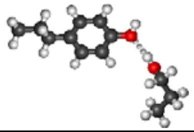
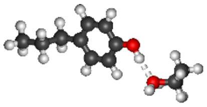
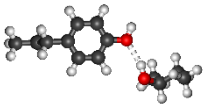
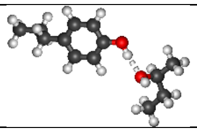
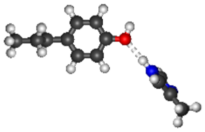
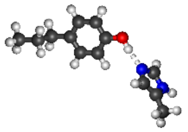
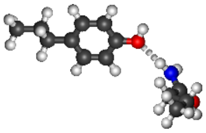
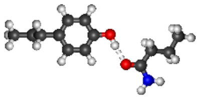
4.3.4 Tyrosine is a preferred hydrogen-bond donor

As statistical results in Table 4.3 shows, in almost 40% of all hydrogen-bonding interactions found, tyrosine can be either hydrogen-bond donor or hydrogen-bond acceptor. The limitation of capability to detect proton positions with x-ray crystallography has generated a great challenging for the bioinformatics studies on the hydrogen-bonding interactions of those side chains who are capable of either serving as hydrogen-bond donor or hydrogen-bond acceptor or both.

Here we analyzed this problem from an energetic perspective. For each pair of tyrosine and its hydrogen-bond partners (H_2O, Ser, Thr, His and Asn/Gln), the hydroxyl group of tyrosine can serve as either a hydrogen-bond donor or acceptor. Our survey of experimental data on hydrogen-bonding interactions of tyrosine has revealed that tyrosine as hydrogen-bond donor is almost twice as often as hydrogen-bond acceptor. We performed computational studies on the structure and hydrogen-bond energy of tyrosine with a set of partners which can serve both as hydrogen-bond donor and as hydrogen-bond acceptor. The results are summarized in Table 4.4. For a water molecule, the hydrogen-bonding energy with tyrosine as donor is about 8.8 kJ/mol stronger than with tyrosine as acceptor. The energy difference is almost the same when calculated in vacuum or in dielectric environment of 4.24, which resembles the hydrophobic environment of protein interior. According to Boltzmann distribution, the data implies that 97.5% of tyrosine-H_2O will exist with tyrosine as the hydrogen-bond donor. For amino acids of Ser, Thr, His and Asn/Gln, the hydrogen-bonding energy with tyrosine as donor is 12-22 kJ/mol stronger than tyrosine as acceptor, indicating that practically 100% of tyrosine are hydrogen-bond donor. Therefore, our computational results show that indeed tyrosine is predominant hydrogen-bond donor when interacting with H_2O, and side chains of Ser, Thr, His and $CONH_2$ of Asn/Gln, if only one hydrogen-bonding interaction exists between tyrosine and the hydrogen-bond partner.

This result has an important application in hydrogen-bond analysis of protein structures that lack the structural information of hydrogen atoms. Currently, majority of reported protein structures do not have hydrogen atoms due to its weak electron density. When adding hydrogen atoms to those protein crystal structures, our results can be used to guide where the hydrogen atom should be positioned so that tyrosine will be the hydrogen-bond donor for relative stronger hydrogen-bonding interactions.

Table 4.4 Comparison of calculated hydrogen-bond properties of tyrosine forming one hydrogen-bond as acceptor or as donor with the same hydrogen-bond partner.

Partner	Structure Acceptor	HB Length (Å)	HB Angle (°)	Structure Donor	HB Length (Å)	HB Angle (°)	ΔE_{HB} (kJ/mol)†		P_D‡
							ε=0	ε=4.24	
H_2O		2.96	122.9		2.86	116.7	8.8	9.2	97.08%
Ser		2.97	122.1		2.83	116.1	11.9	11.8	99.14%
Thr		3.01	114.9		2.82	116.5	15.2	15.5	99.77%
His		3.05	130.3		2.83	116.7	21.7	21.7	99.98%
Asn		3.12	131.2		2.77	115.4	22.3	19.7	99.99%

Computational method B3LYP/6-311+G(3df,2p) was used for geometry optimization and energy calculation.
† ΔE_{HB} was calculated by hydrogen-bond energy when tyrosine is donor subtracting hydrogen-bond energy when tyrosine is acceptor. ΔE_{HB} was calculated at two different environment: vacuum (ε=0) and ether (ε=4.24).
‡ P_D refers to the probability of tyrosine's role is hydrogen-bond donor when paired with the given hydrogen-bond partner. Probabilities were calculated using Boltzmann Distribution based on the hydrogen-bond energy difference in vacuum. See more details in text.

4.3.5 VSM of Tyr from computational studies

The phenolic group of tyrosine is able to form multiple types of hydrogen-bonding interactions. Our computational studies show that it can form one hydrogen-bonding interactions as donor and two hydrogen-bonding interactions as acceptor. Table 4.5, Table 4.6, Table 4.7, Table 4.8 and Table 4.9 list the results from the computational studies of model systems for tyrosine forming different types of hydrogen-bonding interactions.

Determining the vibrational modes sensitive to hydrogen-bonding interactions

To identify the vibrational modes that are sensitive to hydrogen-bonding interactions on side chain of a ring-D4 tyrosine residue, we performed a series of calculations on tyrosine side chain forming hydrogen-bonding interactions with 0, 1, 2 and 4 water molecules, whereas tyrosine is acting as hydrogen-bond donor, hydrogen-bond acceptor, or both. The output geometry as well as corresponding computational spectrum are shown in Figure 4.8.

The results indicate that there are three bands sensitive to hydrogen-bonding interactions. Same as previous reported for ring-H4 tyrosine, C-O stretching ($\nu_{7a'}$ or ν_{CO}, bands filled with red in Figure 4.8) contributes one of the modes. The other two modes, meanwhile both have C-O-H bending partially involved, coupling with some tyrosine aromatic ring vibrations. Therefore we named the one with lower frequency δ_{COH}-α (bands filled with blue in Figure 4.8), and the one with higher frequency δ_{COH}-β (bands filled with green in Figure 4.8). Those three vibrational modes are also illustrated by vibrational vectors, shown in Figure 4.9.

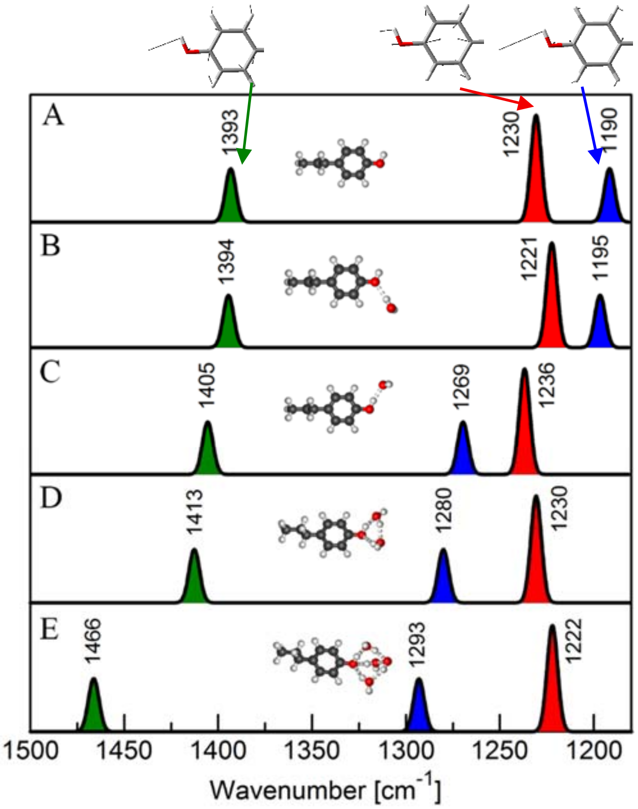


Figure 4.8 Computational spectrum showing 3 vibrational modes sensitive to different types of hydrogen-bonding interactions on tyrosine side chain. (A) 0 hydrogen-bonding interaction; (B) 1 hydrogen-bonding interaction with H_2O as hydrogen-bond acceptor; (C) 1 hydrogen-bonding interaction with H_2O as hydrogen-bond donor; (D) 2 hydrogen-bonding interactions with 2 H_2O molecules as hydrogen-bond donor and acceptor; (E) total of 3 hydrogen-bonding interactions with 3 H2O molecules, 1 as hydrogen-bond donor and 2 as hydrogen-bond acceptor. Additional H_2O was added to stabilize hydrogen-bond network. B3LYP/6-311+G(3df,2p) method was used, and computational frequencies were scaled using scaling factor 0.988. Computational intensities were normalized. Red: C-O stretching ($\nu_{7a'}$); Blue: C-O-H bending-α ($\delta_{COH-\alpha}$); Green: C-O-H bending-β ($\delta_{COH-\beta}$).

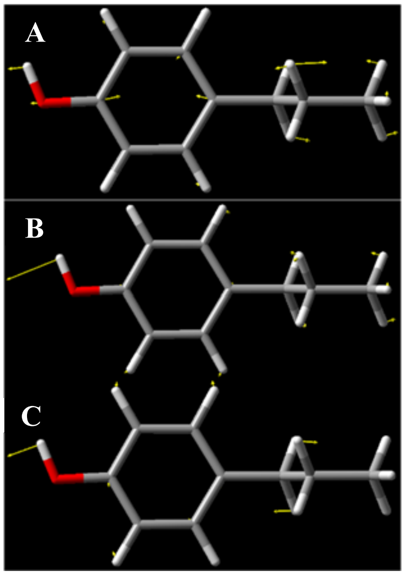


Figure 4.9 Normal vibrations of ring-H4 4-propylphenol with zero hydrogen-bonding interaction. (A) C-O stretching ($\nu_{7a'}$); (B) C-O-H bending-α ($\delta_{COH-\alpha}$); (C) C-O-H bending-β ($\delta_{COH-\beta}$)..

No hydrogen-bonding interaction and weak interactions

As indicated in Table 4.5, without hydrogen-bonding interactions or forming weak interactions with partners such as CH_3CH_3, isolated tyrosine residue has $\nu_{7a'}$ at ~1229 cm^{-1}. And δ_{COH}-α is observed at 1190 cm^{-1}, while band δ_{COH}-β is located at ~1393 cm^{-1}.

One hydrogen-bonding interaction as acceptor

Phenolic oxygen may form one hydrogen-bonding interaction as an acceptor with a hydrogen-bond donor. When such hydrogen-bonding interaction is formed, it is observed from the computational results that the frequencies of $\nu_{7a'}$red-shift to ~1226 cm^{-1} for weak hydrogen-bond donor such as Cys with –SH group, to ~1220-1223 cm^{-1} for normal neutral hydrogen-bond donor such as Ser or His with –OH or –NH group, to ~1217 cm-1 for strong neutral hydrogen-bond donor such as Asp with –COOH group, and to ~1187-1195 cm^{-1} for positively charged hydrogen-bond donor such as His(+) and Lys(+) with –NH(+) group. For the frequencies of δ_{COH}-α, blue-shifts are observed as to ~1191 for weak hydrogen-bond donor, to ~1194-1197 cm^{-1} for normal neutral hydrogen-bond donor, to ~1204 cm-1 for strong neutral hydrogen-bond donor and to ~1207-1211 cm^{-1} for positively charged hydrogen-bond donor. Meanwhile, the frequencies of δ_{COH}-β are found to be stable in the region of ~1393-1395 cm^{-1} after tyrosine forms a hydrogen-bonding interaction as an acceptor. The hydrogen-bond energies in vacuum (ε=0) are also calculated, as summarized in Table 4.6.

Because of the variations of frequencies shift based on the hydrogen-bond donors to tyrosine, we plotted the correlation between frequencies of $\nu_{7a'}$ and hydrogen-bond energy, as well as the correlation between frequencies of δ_{COH}-α and $\nu_{7a'}$, in Figure 4.10.

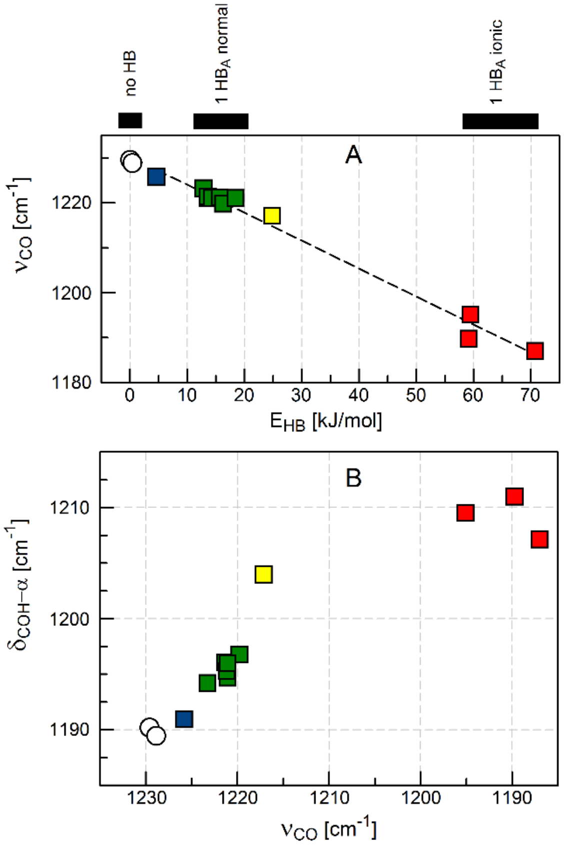


Figure 4.10 Hollow circle: tyrosine forms no hydrogen-bonding interaction; Square: tyrosine as acceptor forms one hydrogen-bonding interaction with weak donor (blue), normal neutral donor (green), strong neutral donor (yellow) and positively charged donor (red). (A) Correlation between hydrogen-bond energy and frequencies of $\nu_{7a'}$; (B) Correlation between frequencies of δ_{COH}-α and $\nu_{7a'}$.

Figure 4.10 (A) clearly indicates a linear dependence of C-O stretching frequencies on the energy of the hydrogen-bonding interactions formed on tyrosine. We linearly fitted the correlation, as the dashed line shows. It should be noted that when tyrosine is an acceptor, the energies of hydrogen-bonding interactions are generally low, only ~13-19 kJ/mol for normal neutral hydrogen-bond donors with –OH or –NH as the donor group. However, for donor groups that are willing to lose hydrogen such as strong neutral or positively charged hydrogen-bond donors with –COOH and –NH(+) as the donor group, hydrogen-bond energies are much higher, ~25 kJ/mol for strong neutral hydrogen-bonding interactions and ~59-71 kJ/mol for ionic (positively charged) hydrogen-bonding interactions, respectively.

One hydrogen-bonding interaction as donor

When the phenolic group of tyrosine donates a hydrogen-bonding interaction to a hydrogen-bond partner, from the results summarized in Table 4.7 we observe blue-shifts on the computational frequencies of all three vibrational modes: $\nu_{7a'}$, δ_{COH}-α and δ_{COH}-β. When the hydrogen-bond partner is a neutral group, our calculations show hydrogen-bonding energy varies from ~17-38 kJ/mol. Correspondingly, the frequencies of $\nu_{7a'}$are found in the range of ~1222-1243 cm^{-1}, while huge blue-shifts are observed on the frequencies of δ_{COH}-α, as they appear at ~1260-1282 cm^{-1}, and frequencies of δ_{COH}-β also blue-shift to ~1398-1420 cm^{-1}. Meanwhile, if hydrogen-bonding interaction is ionic, tyrosine is forming hydrogen-bonding interactions as donor with a negatively charged group such as COO(-) of Asp/Glu or a H_2PO_3(-). Much more significant blue-shifts are found on the frequencies of three vibrational modes, that $\nu_{7a'}$bands are now at ~1255-1270 cm^{-1}, while δ_{COH}-α bands are located at ~1287-1290 cm-1 and δ_{COH}-β are found with frequencies higher than ~1477 cm^{-1}. The much higher hydrogen-bond energy caused by ionic hydrogen-bonding interactions should be the reason for these further blue-shifts. Similar to tyrosine forming hydrogen-bonding interactions as acceptor, we plotted correlations between frequencies of $\nu_{7a'}$ and hydrogen-bond energy, as well as correlations between frequencies of two vibrational modes, δ_{COH}-α and $\nu_{7a'}$.

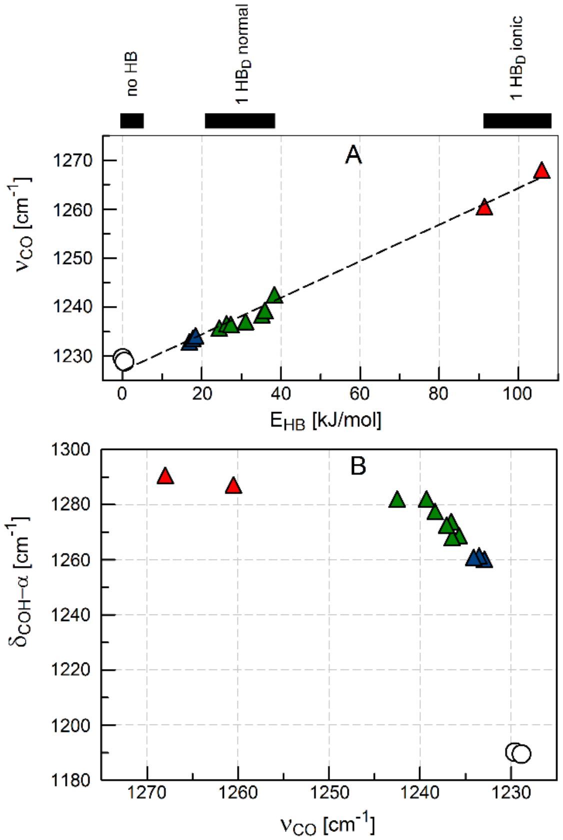


Figure 4.11 Hollow circle: tyrosine forms no hydrogen-bonding interaction; Triangle: tyrosine as donor forms one hydrogen-bonding interaction with weak donor (blue), normal neutral donor (green), and negatively charged donor (red). (A) Correlation between hydrogen-bond energy and frequencies of $\nu_{7a'}$; (B) Correlation between frequencies of δ_{COH}-α and $\nu_{7a'}$.

As Figure 4.11implies, there is a linear relationship between the frequencies of $\nu_{7a'}$and hydrogen-bond energy. However, in contrast to the relationship found for tyrosine as an acceptor, the linear fit (dashed line) in A indicates a positive slope. Stronger hydrogen-bonding interaction leads to higher $\nu_{7a'}$frequencies. It is also observed that for δ_{COH}-α, there is a huge frequency jump (larger than 65 cm-1) from tyrosine forming no hydrogen-bonding interaction to forming one hydrogen-bonding interaction as donor. This indicates the frequencies of C-O-H bending is affected much more by the hydrogen-bonding interactions when the phenolic hydrogen of tyrosine is directly involved.

Two hydrogen-bonding interactions as acceptor and donor

Phenolic group of tyrosine may also form two hydrogen-bonding interactions, one as donor and one as acceptor at the same time. As our calculations of tyrosine forming one hydrogen-bonding interaction as donor or forming one hydrogen-bonding interaction as acceptor have shown, neutral and ionic hydrogen-bonding interactions will have different effects on the frequencies of the three vibrational modes which are sensitive to hydrogen-bonding interactions. Therefore we analyzed tyrosine with two hydrogen-bonding interactions in three categories:

(1) Tyrosine is hydrogen-bond donor and acceptor to two neutral groups.

The frequencies of $\nu_{7a'}$ have a blue-shift to ~1230-1235 cm^{-1}, but this shift (~0-5 cm^{-1}) is much smaller than the blue-shift of ~3-13 cm^{-1} caused by forming single hydrogen-bonding interaction as donor only. However, considering the red-shift of ~7 cm^{-1} caused by forming one hydrogen-bonding interaction as acceptor, the overall small blue-shift could be explained as the mutual effect of two hydrogen-bonding interactions.

Similarly, since both types of hydrogen-bonding interactions, when tyrosine as donor or as acceptor, result in blue-shifts on the frequencies of the two C-O-H bending modes δ_{COH}-α and δ_{COH}-β, our computational results show that when tyrosine forms two hydrogen-bonding interactions as donor and acceptor, those two modes both have bigger blue-shifts on vibrational frequencies. As the Table 4.8 shows, δ_{COH}-α appears around ~1280-1290 cm^{-1}, and δ_{COH}-β is at ~1413-1465 cm^{-1}. It should be noted that the frequencies of δ_{COH}-β have a large range of variation. In fact, it is observed that when the total energy of two hydrogen-bonding interactions increased from 40 kJ/mol to 51 kJ/mol, δ_{COH}-β frequency has a separation of >30 cm^{-1}.

(2) **Tyrosine is hydrogen-bond donor to a negatively charged group and is hydrogen-bond acceptor to a neutral group.**

We observed $\nu_{7a'}$band at ~1243 cm^{-1}. The further blue-shifts comparing with two neutral hydrogen-bond partners is likely the result of ionic hydrogen-bonding interaction, which is similar to what we observed when tyrosine as donor only forms one hydrogen-bond interaction with a negatively charged group.

Meanwhile, δ_{COH}-α is at ~1290 cm^{-1}, δ_{COH}-β is at ~1477 cm^{-1}. Comparing with two neutral hydrogen-bond partners, δ_{COH}-α band seems to not have much blue-shifts, while δ_{COH}-β does have a blue-shift of >10 cm^{-1}.

(3) **Tyrosine is hydrogen-bond donor to a neutral group and is hydrogen-bond acceptor to a positively charged group.**

In this case, we found that the $\nu_{7a'}$frequency red-shifts to ~1206 cm^{-1}. The low C-O stretching frequency could be explained by the large red-shifts on $\nu_{7a'}$frequency when tyrosine forms one hydrogen-bonding interaction as acceptor with a positively charged group.

δ_{COH}-α and δ_{COH}-β are located at ~1297 cm^{-1} and ~1453 cm^{-1}, respectively. This result is expected since for tyrosine, forming hydrogen-bonding interaction either as donor or acceptor leads to blue-shifts on C-O-H bending frequencies. Also, unlike tyrosine forming two hydrogen-bonding interactions with a neutral group and a negatively charged group, the δ_{COH}-β does not have further blue-shift comparing to the results when tyrosine is hydrogen-bonded with neutral partners only. Such phenomenon is also observed when tyrosine forms one hydrogen-bonding interaction as acceptor, as large hydrogen-bond energy does not have much effect on the δ_{COH}-β frequency.

Three hydrogen-bonding interactions as two acceptor and one donor

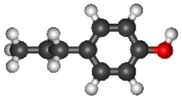
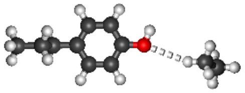
It was previously reported that tyrosine residues are capable of forming two hydrogen-bonding interactions at maximum based on the searching of protein crystal structures (McDonald and Thornton, 1994) and they suggested that the lone pairs on phenolic oxygen of tyrosine are more likely to be partially delocalized within the aromatic ring. However, with our extensive research on 1189 protein crystal structures, several tyrosine residues that potentially form three hydrogen-bonding interactions are located. For example, in the crystal structure of PYP (PDB ID: 1NWZ), it is found that the phenolic group of Tyr94 is in such triple hydrogen-bond complex. As Y94 is the hydrogen-bond donor to the backbone C=O of Cys69, and accepts two hydrogen-bonding interactions from a water molecule and –OH group of Ser72 (Getzoff et al, 2003).

Also, Takahashi and Noguchi have shown in their work that three hydrogen-bonding interactions on tyrosine is very unlikely based on their DFT calculations of p-cresol with three water molecules (Takahashi and Noguchi, 2007). However, the computational method (B3LYP/6-31+G(d,p)) they used is limited with number of functions applied, and their model system does not have enough water molecules to simulate the true environment around tyrosine in solution, which will help stabilizing the triple hydrogen-bond interactions.

We performed DFT calculations on tyrosine forming three hydrogen-bonding interactions with three water molecules, we also add an additional water molecule which provides hydrogen-bonding interactions to those three water molecules. The optimized geometry shows that such model system consisting total of four water molecules stabilizes a triple hydrogen-bond complex as described, with corresponding hydrogen-bond distances (O..O) 2.67 Å for tyrosine as donor and 2.90 Å, 2.91 Å for tyrosine as acceptor. We also performed calculations on a model system of tyrosine hydrogen-bonded with backbone C=O, serine –OH and a water molecule, the same as Y94 in PYP, the geometry structure after optimization also shows a stabilized triple hydrogen-bond complex, suggesting the triple hydrogen-bond complex observed in PYP crystal structure is possible. Hydrogen-bond properties of those two systems are shown in Table 4.9.

Meanwhile, the frequency calculations of the two triple hydrogen-bond complex shows $\nu_{7a'}$ frequencies red-shift to the range of ~1222-1226 cm^{-1} and frequencies of δ_{COH}-α and δ_{COH}-β blue-shift to ~1293 cm^{-1} and ~1466-1468 cm^{-1}. The red-shifts of $\nu_{7a'}$frequencies could be explained as the results of tyrosine forming hydrogen-bonding interactions as acceptor to two donors, as comparing to the circumstances of forming one hydrogen-bonding interaction as donor and one hydrogen-bonding interaction as acceptor with neutral groups, the additional hydrogen-bonding interaction, in which tyrosine is the acceptor causes the red-shift of $\nu_{7a'}$ frequency. Similarly the further blue-shifts observed on δ_{COH}-α and δ_{COH}-β are also the results of forming that additional hydrogen-bond interaction.

Table 4.5 Calculated hydrogen-bond properties of ring-D4 tyrosine side chain forming no hydrogen-bonding interaction

Optimized Structure*	HB Partner*	Type of HB[†]	HB Length[†] (Å)	HB Angle[†] (deg)	HB Dihedral Angle[†] (deg)	H-bond Energy[‡] (kJ/mol)	$\nu_{C\text{-}O}$[§] (cm^{-1})	δ_{COH}-α[§] (cm^{-1})	δ_{COH}-β[§] (cm^{-1})
	NA	0	NA	NA	NA	NA	1229.6	1190.2	1393.2
	Ethane	0	4.01	166.0	19.9	0.4	1228.9	1189.5	1392.8

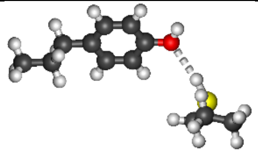
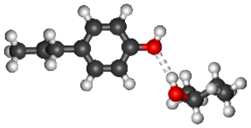
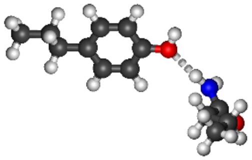
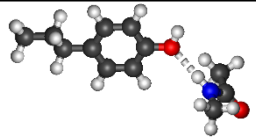
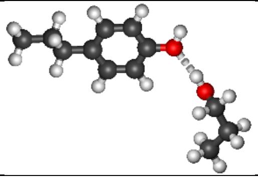
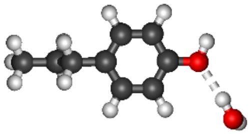
*Computational method B3LYP/6-311+G(3df, 2p) was used for geometry optimization, energy and frequency calculation. Color codes for atoms: black for carbon atoms, red for oxygen atoms, blue for nitrogen atoms, yellow for sulful atoms, white for hydrogen atoms, orange for phosphate atoms. White dashed lines represent hydrogen-bonding interactions or weak interactions.

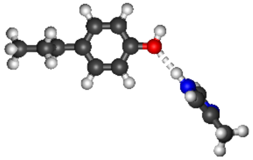
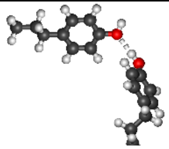
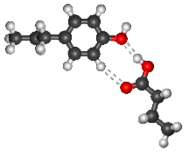
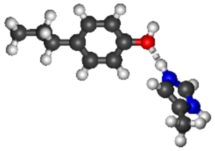
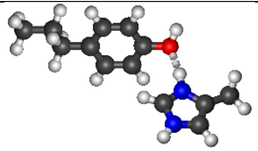
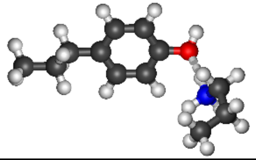
†Types of hydrogen-bonding interactions are characterized by the number of hydrogen-bonding interactions formed on a tyrosine residue and the role of tyrosine residue plays in that interaction. D refers to tyrosine as hydrogen-bond donor; D refers to tyrosine as hydrogen-bond acceptor. Hydrogen-bond length is defined as the distance between two heavy atoms paired as hydrogen-bond donor and hydrogen-bond acceptor. If two or more hydrogen-bonding interactions are formed, the top value is for the hydrogen-bonding interaction in which tyrosine serves as hydrogen-bond donor.

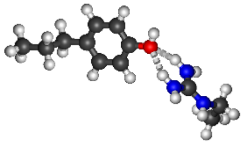
‡Hydrogen-bond energy is calculated in vacuum.

§Vibrational frequency are scaled with scaling factor 0.988.

Table 4.6 Calculated hydrogen-bond properties of ring-D_4 tyrosine side chain forming hydrogen-bonding interaction as hydrogen-bond acceptor

Optimized Structure*	HB Partner*	Type of HB[†]	HB Length[†] (Å)	HB Angle[†] (deg)	HB Dihedral Angle[†] (deg)	H-bond Energy[‡] (kJ/mol)	ν_{C-O}[§] (cm^{-1})	δ_{COH}-α[§] (cm^{-1})	δ_{COH}-β[§] (cm^{-1})
	Cys	1A weak	3.78	120.8	21.8	4.6	1225.8	1190.9	1394.2
	Thr	1A normal	3.01	114.9	3.5	12.6	1220.8	1194.6	1394.6
	Asn	1A normal	3.12	131.0	8.5	13.6	1221.1	1196.0	1395.0
	Backbone	1A normal	3.13	131.4	13.0	13.6	1221.4	1196.1	1395.3
	Ser	1A normal	2.97	122.1	1.3	14.4	1221.1	1194.7	1394.3
	H_2O	1A normal	2.96	122.9	4.4	15.7	1221.2	1195.3	1394.4

Optimized Structure*	HB Partner*	Type of HB[†]	HB Length[†] (Å)	HB Angle[†] (deg)	HB Dihedral Angle[†] (deg)	H-bond Energy[‡] (kJ/mol)	$\nu_{C\text{-}O}$[§] (cm^{-1})	δ_{COH}-α[§] (cm^{-1})	δ_{COH}-β[§] (cm^{-1})
	His	1A normal	3.05	130.3	5.7	16.6	1219.9	1198.0	1395.2
	Tyr	1A normal	2.90	126.9	9.9	18.4	1221.1	1196.0	1395.2
	Asp/Glu	1A strong	2.85	135.8	0.2	24.8	1217.1	1204.0	1395.8
	His(+)_A	1A ionic	2.76	120.5	48.1	59.4	1195.8	1209.5	1392.7
	His(+)_B	1A ionic	2.78	107.7	63.5	59.2	1189.7	1211.0	1387.6
	Lys(+)	1A ionic	2.74	110.1	61.3	70.7	1187.0	1207.1	1384.5

Optimized Structure*	HB Partner*	Type of HB†	HB Length† (Å)	HB Angle† (deg)	HB Dihedral Angle† (deg)	H-bond Energy‡ (kJ/mol)	$\nu_{C\text{-}O}$§ (cm^{-1})	δ_{COH}-α§ (cm^{-1})	δ_{COH}-β§ (cm^{-1})
	Arg(+)	2A ionic	2.96/2.94	127.5/ 125.0	-22.6/34.8	59.7	1192.4	1210.0	1397.0

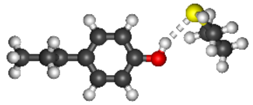
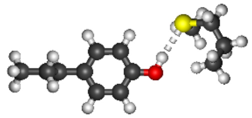
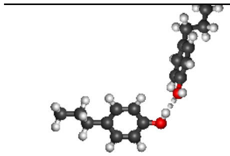
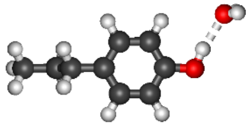
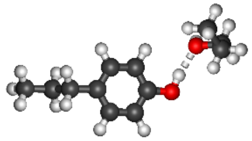
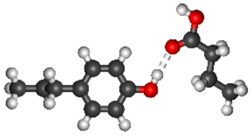
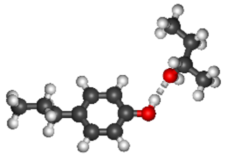
*Computational method B3LYP/6-311+G(3df, 2p) was used for geometry optimization, energy and frequency calculation. Color codes for atoms: black for carbon atoms, red for oxygen atoms, blue for nitrogen atoms, yellow for sulful atoms, white for hydrogen atoms, orange for phosphate atoms. White dashed lines represent hydrogen-bonding interactions or weak interactions.

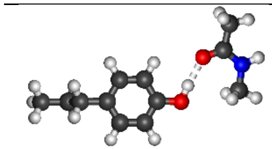
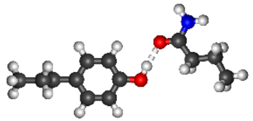
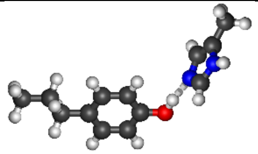
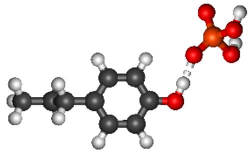
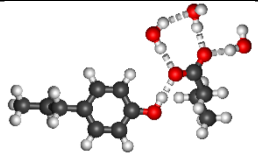
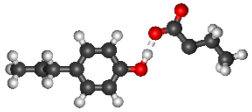
†Types of hydrogen-bonding interactions are characterized by the number of hydrogen-bonding interactions formed on a tyrosine residue and the role of tyrosine residue plays in that interaction. D refers to tyrosine as hydrogen-bond donor; D refers to tyrosine as hydrogen-bond acceptor. Hydrogen-bond length is defined as the distance between two heavy atoms paired as hydrogen-bond donor and hydrogen-bond acceptor. If two or more hydrogen-bonding interactions are formed, the top value is for the hydrogen-bonding interaction in which tyrosine serves as hydrogen-bond donor.

‡Hydrogen-bond energy is calculated in vacuum.

§Vibrational frequency are scaled with scaling factor 0.988.

Table 4.7 Calculated hydrogen-bond properties of ring-D_4 tyrosine side chain forming hydrogen-bonding interaction as hydrogen-bond donor

Optimized Structure*	HB Partner*	Type of HB[†]	HB Length[†] (Å)	HB Angle[†] (deg)	HB Dihedral Angle[†] (deg)	H-bond Energy[‡] (kJ/mol)	$\nu_{C\text{-}O}$[§] (cm^{-1})	δ_{COH}-α[§] (cm^{-1})	δ_{COH}-β[§] (cm^{-1})
	Cys	1D weak	3.35	125.6	5.7	16.9	1232.9	1260.2	1398.3
	Met	1D weak	3.37	116.4	7.3	17.8	1233.5	1261.4	1397.5
	Tyr	1D weak	2.90	116.0	2.1	18.4	1234.1	1260.8	1400.1
	H_2O	1D normal	2.86	116.3	0.0	24.4	1235.7	1268.7	1405.4
	Ser	1D normal	2.83	116.0	0.0	26.2	1236.6	1273.9	1410.1
	Asp/Glu	1D normal	2.87	114.3	1.0	27.3	1236.5	1268.1	1408.0
	Thr	1D normal	2.81	116.5	0.0	27.8	1237.1	1272.5	1411.1

Optimized Structure*	HB Partner*	Type of HB†	HB Length† (Å)	HB Angle† (deg)	HB Dihedral Angle† (deg)	H-bond Energy‡ (kJ/mol)	$\nu_{C\text{-}O}$§ (cm^{-1})	δ_{COH}-α§ (cm^{-1})	δ_{COH}-β§ (cm^{-1})
	Backbone	1D normal	2.78	114.6	1.3	35.2	1238.4	1277.5	1414.6
	Asn	1D normal	2.77	115.4	0.4	35.9	1239.3	1282.1	1419.7
	His	1D normal	2.82	116.7	1.8	38.3	1242.5	1282.1	1419.7
	Phosphate(-)	1D ionic	2.58	110.0	2.4	91.4	1260.5	1287.2	1476.7
	Asp/Glu(-)_A	1D ionic	2.63	114.8	2.9	NA	1254.1	1288.7	1483.7
	Asp/Glu(-)_B	1D ionic	2.51	110.5	0.5	105.9	1268.0	1290.7	1651.6

*Computational method B3LYP/6-311+G(3df, 2p) was used for geometry optimization, energy and frequency calculation. Color codes for atoms: black for carbon atoms, red for oxygen atoms, blue for nitrogen atoms, yellow for sulful atoms, white for hydrogen atoms, orange for phosphate atoms. White dashed lines represent hydrogen-bonding interactions or weak interactions.

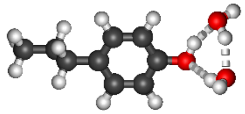
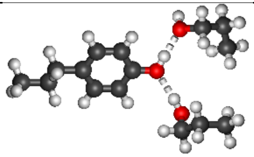
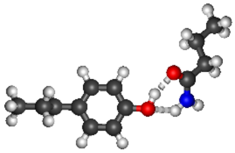
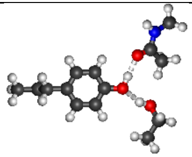
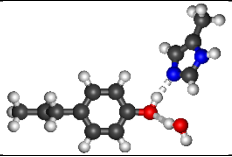
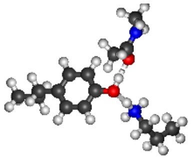
†Types of hydrogen-bonding interactions are characterized by the number of hydrogen-bonding interactions formed on a tyrosine residue and the role of tyrosine residue plays in that interaction. D refers to tyrosine as hydrogen-bond donor; D refers to tyrosine as hydrogen-bond acceptor. Hydrogen-bond length is defined as the distance between two

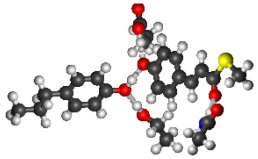
heavy atoms paired as hydrogen-bond donor and hydrogen-bond acceptor. If two or more hydrogen-bonding interactions are formed, the top value is for the hydrogen-bonding interaction in which tyrosine serves as hydrogen-bond donor.

‡Hydrogen-bond energy is calculated in vacuum.

§Vibrational frequency are scaled with scaling factor 0.988.

Table 4.8 Calculated hydrogen-bond properties of ring-D4 tyrosine side chain forming two hydrogen bonding interactions, one as hydrogen-bond donor and one as hydrogen-bond acceptor

Optimized Structure*	HB Partner*	Type of HB†	HB Length† (Å)	HB Angle† (deg)	HB Dihedral Angle† (deg)	H-bond Energy‡ (kJ/mol)	$\nu_{C\text{-}O}$§ (cm^{-1})	δ_{COH}-α§ (cm^{-1})	δ_{COH}-β§ (cm^{-1})
	$2H_2O$	1D1A normal	2.77 2.86	122.0 133.4	11.9 88.5	NA	1230.3	1279.8	1412.5
	2Ser	1D1A normal	2.78 2.93	114.1 123.0	10.2 14.3	43.7	1230.8	1278.8	1415.6
	Asn	1D1A normal	2.71 3.00	123.3 142.6	11.9 71.7	40.4	1234.4	1285.8	1421.0
	Backbone Ser	1D1A normal	2.72 2.85	113.3 139.7	2.0 45.3	51.2	1233.8	1287.6	1454.7
	His H_2O	1D1A normal	2.75 2.86	119.6 128.4	4.0 71.4	69.5	1234.7	1290.3	1464.6
	Backbone Lys(+)	1D(0)1A(+) ionic	2.58 2.64	115.7 108.9	19.7 71.9	NA	1206.0	1297.4	1453.0

Optimized Structure*	HB Partner*	Type of HB†	HB Length† (Å)	HB Angle† (deg)	HB Dihedral Angle† (deg)	H-bond Energy‡ (kJ/mol)	$\nu_{C\text{-}O}$§ (cm^{-1})	δ_{COH}-α§ (cm^{-1})	δ_{COH}-β§ (cm^{-1})
	pCA(-) Thr	1D(-)1A(0) ionic	2.59 2.86	114.3 123.0	0.0 0.0	NA	1243.4	1289.8	1477.4

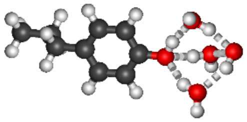
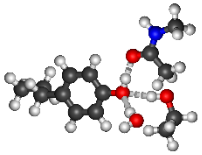
*Computational method B3LYP/6-311+G(3df, 2p) was used for geometry optimization, energy and frequency calculation. Color codes for atoms: black for carbon atoms, red for oxygen atoms, blue for nitrogen atoms, yellow for sulful atoms, white for hydrogen atoms, orange for phosphate atoms. White dashed lines represent hydrogen-bonding interactions or weak interactions.

†Types of hydrogen-bonding interactions are characterized by the number of hydrogen-bonding interactions formed on a tyrosine residue and the role of tyrosine residue plays in that interaction. D refers to tyrosine as hydrogen-bond donor; D refers to tyrosine as hydrogen-bond acceptor. Hydrogen-bond length is defined as the distance between two heavy atoms paired as hydrogen-bond donor and hydrogen-bond acceptor. If two or more hydrogen-bonding interactions are formed, the top value is for the hydrogen-bonding interaction in which tyrosine serves as hydrogen-bond donor.

‡Hydrogen-bond energy is calculated in vacuum.

§Vibrational frequency are scaled with scaling factor 0.988.

Table 4.9 Calculated hydrogen-bond properties of ring-D4 tyrosine side chain forming three hydrogen bonding interactions, one as hydrogen-bond donor and two as hydrogen-bond acceptor

Optimized Structure*	HB Partner*	Type of HB†	HB Length† (Å)	HB Angle† (deg)	HB Dihedral Angle† (deg)	H-bond Energy‡ (kJ/mol)	$\nu_{C\text{-}O}$§ (cm^{-1})	δ_{COH}-α§ (cm^{-1})	δ_{COH}-β§ (cm^{-1})
	3H$_2$O	1D2A normal	2.67 2.91 2.90	120.8 129.2 136.8	6.2 76.8 -59.0	NA	1222.1	1293.2	1466.3
	Backbone Ser H$_2$O	1D2A normal	2.65 2.96 2.89	114.9 130.7 127.5	2.4 68.6 -57.0	73.5	1226.1	1293.5	1468.1

*Computational method B3LYP/6-311+G(3df, 2p) was used for geometry optimization, energy and frequency calculation. Color codes for atoms: black for carbon atoms, red for oxygen atoms, blue for nitrogen atoms, yellow for sulful atoms, white for hydrogen atoms, orange for phosphate atoms. White dashed lines represent hydrogen-bonding interactions or weak interactions.

†Types of hydrogen-bonding interactions are characterized by the number of hydrogen-bonding interactions formed on a tyrosine residue and the role of tyrosine residue plays in that interaction. D refers to tyrosine as hydrogen-bond donor; D refers to tyrosine as hydrogen-bond acceptor. Hydrogen-bond length is defined as the distance between two heavy atoms paired as hydrogen-bond donor and hydrogen-bond acceptor. If two or more hydrogen-bonding interactions are formed, the top value is for the hydrogen-bonding interaction in which tyrosine serves as hydrogen-bond donor.

‡Hydrogen-bond energy is calculated in vacuum.

§Vibrational frequency are scaled with scaling factor 0.988.

2D VSM for Tyrosine Hydrogen-Bonding Interactions

Here we are able to introduce a 2-Dimenstional Vibrational Structural Marker (2D VSM) for identifying hydrogen-bonding interactions of ring-D4 tyrosine. Similar to the previous studies reported by Nie (Nie, 2006) and Takahashi (Takahashi and Noguchi, 2007), we also use the frequencies of $\nu_{7a'}$as horizontal axis. For vertical axis, however, upon ring-D4 labeling, C-O-H bending vibrational mode is coupled with ring vibration at two positions as they are named δ_{COH}-α and δ_{COH}-β, we plot the 2D VSM in two ways: one use δ_{COH}-α and one use δ_{COH}-β, in Figure 4.12 and Figure 4.13.

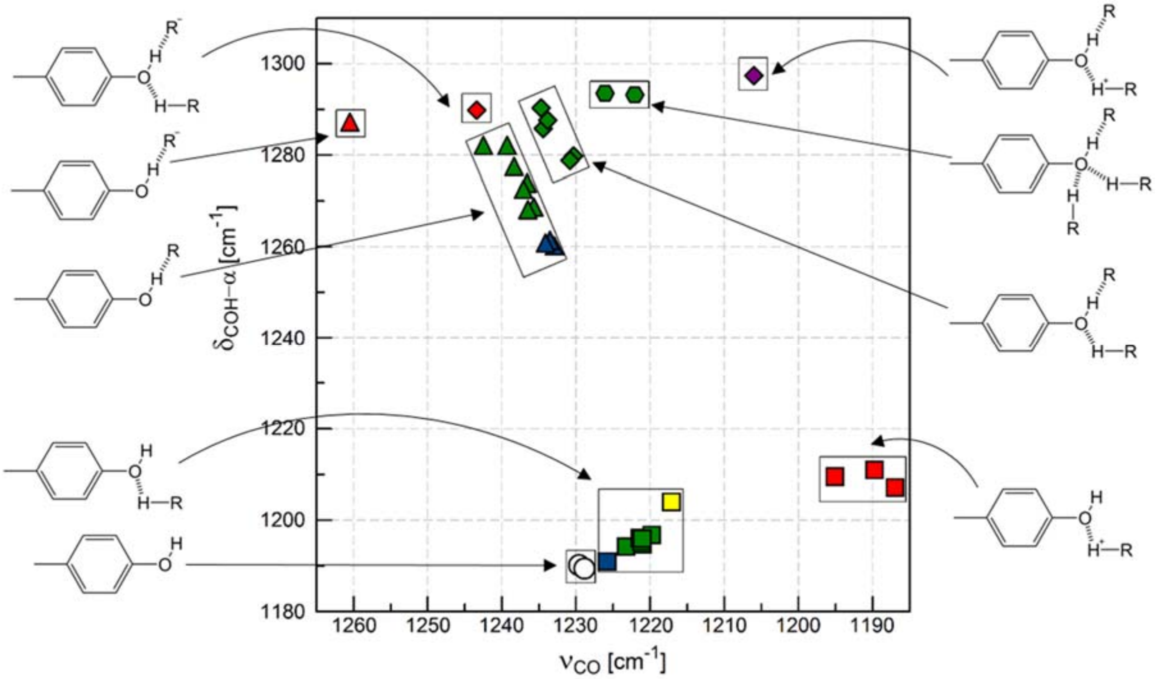


Figure 4.12 2D VSM of tyrosine side chain using frequencies of C-O stretching (x-axis) and C-O-H bending-α (y-axis)

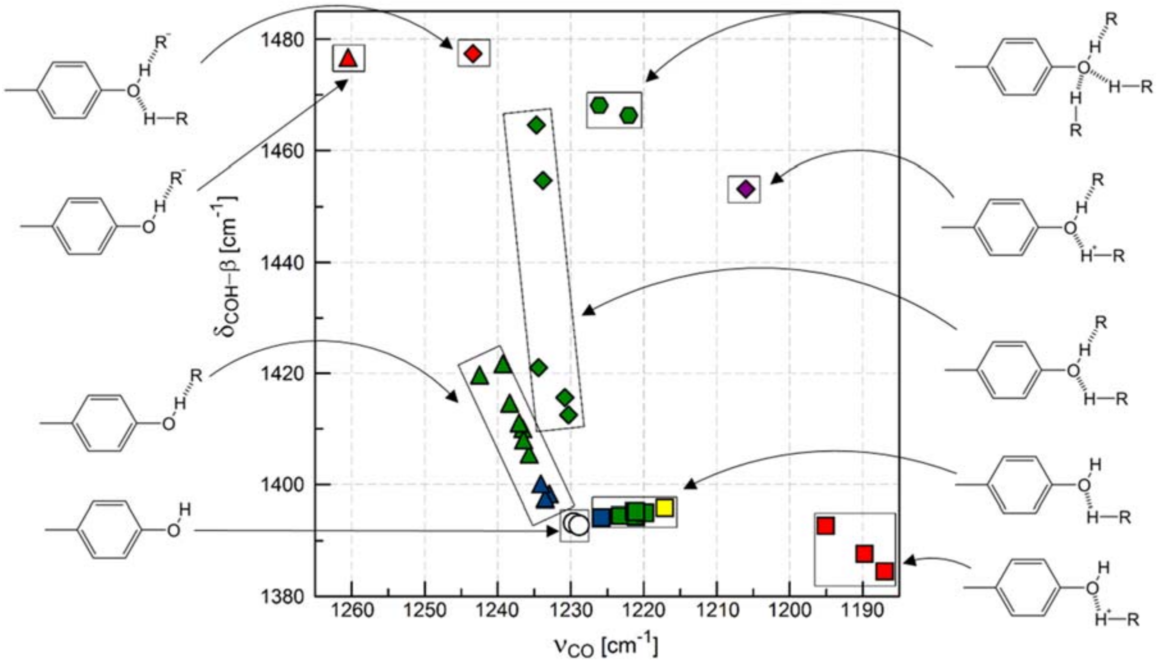


Figure 4.13 2D VSM of tyrosine side chain using frequencies of C-O stretching (x-axis) and C-O-H bending-β (y-axis)

It can be seen that tyrosine forms no hydrogen-bonding interaction or one hydrogen-bonding interaction with neutral and positively charged partners as acceptor, the 2D VSM using $\nu_{7a'}$and δ_{COH}-α is sufficient to identify the hydrogen-bonding status of tyrosine. However, for tyrosine forms one hydrogen-bonding interaction as donor and forms two or three hydrogen-bonding interactions as donor and acceptor with neutral or negatively charged groups, the corresponding area in Figure 4.12 is crowded with multiple types of hydrogen-bonding interactions. To help clarify those hydrogen-bonding interactions, 2D VSM using $\nu_{7a'}$ and δ_{COH}-β can be applied because of its better separation in different types of multiple hydrogen-bonding interactions formed on tyrosine side chain. In fact, from the computational results we observed that the presence of C-O-H bending in δ_{COH}-α is getting weaker with the hydrogen-bonding energy on tyrosine side chain increases. At the same time, presence of C-O-H bending in δ_{COH}-β is getting stronger. Particularly when charged hydrogen-bond partners are involved in the multiple hydrogen-bonding interactions tyrosine side chain forms, we found that δ_{COH}-β mode is almost dominated by C-O-H bending only.

The most obvious results revealed by the two 2-D vibrational structural markers are the unique frequency fingerprints for ionic hydrogen-bonding interactions. Only when tyrosine side chain is accepting a hydrogen-bonding interaction from a positively charged hydrogen-bond donor, frequency of $\nu_{7a'}$ is observed below 1210 cm^{-1}. And only when tyrosine side chain is donating a hydrogen-bonding interaction to a negatively charged hydrogen-bond acceptor, frequency of δ_{COH}-β is above 1470 cm^{-1}.

4.4 Conclusions

The results from our calculations of hydrogen-bonding energies between tyrosine and a hydrogen-bond partner have shown that such energy is higher when tyrosine serves as hydrogen-bond donor than the energy when it serves as hydrogen-bond acceptor, meaning tyrosine side chain is a preferred hydrogen-bond donor Table 4.4. This results are supported by our statistical survey of hydrogen-bond analysis of buried tyrosine side chain from 1182 high resolution protein crystal structures shown in Table 4.2 and Table 4.3.

Our computational approach revealed the possibility of tyrosine phenolic group forming three hydrogen-bonding interactions for the first time. To simulate hydrogen-bonding network around tyrosine phenolic oxygen as it is exposed to water, we add 4 water molecules to the model system: three forming triple hydrogen-bond complex with tyrosine side chain, and the other one forms additional hydrogen-bonding interactions between water molecules for the stabilization. The optimized geometry shows the sp^3 hybridization is energetic favorable for tyrosine side chain exposed in water.

Our computational studies on side chain of ring-D4 tyrosine residue forming different hydrogen bonding interactions indicate that three vibrational modes, which are corresponding to C-O stretching ($\nu_{7a'}$ or ν_{CO}), and C-O-H bending coupling with asymmetric ring vibration (δ_{COH}-α and δ_{COH}-β), are the vibrational structural markers for probing hydrogen-bonding interactions of tyrosine side chain. We found that the correlation between hydrogen-bond energy and the vibrational frequencies of $\nu_{7a'}$ is almost linear, when phenolic group of tyrosine side chain forms single hydrogen-bonding interactions with a hydrogen-bond partner. It is observed that when tyrosine is hydrogen-bond donor, the stronger the hydrogen-bonding interaction, the higher the $\nu_{7a'}$ frequency; and when the tyrosine is hydrogen-bond acceptor, the stronger the hydrogen-bonding interaction, the lower the $\nu_{7a'}$ frequency.

We plot two 2-dimensional Vibrational Structural Markers (2D-VSM) to help identify hydrogen-bonding interactions of tyrosine side chain. One use the frequencies of δ_{COH}-α and $\nu_{7a'}$, and the other use the frequencies of δ_{COH}-β and $\nu_{7a'}$. Our results show that when tyrosine is hydrogen-bond acceptor interacting with a positively charged hydrogen-bond partner, frequencies of $\nu_{7a'}$ will be observed below 1210 cm^{-1}, and when tyrosine is hydrogen-bond donor interacting with a negatively charged hydrogen-bond partner, δ_{COH}-β mode will be dominated by C-O-H bending, and is found to be higher than 1470 cm^{-1}. Meanwhile, for neutral polar hydrogen-bond partners paring with tyrosine side chain, types of hydrogen-bonding interactions can be identified using the combination of two 2D-VSM we reported.

The vibrational structural markers for probing hydrogen-bonding interactions we developed are expected to enhance the power of Time-resolved Infrared Structural Biology for the studies of functionally important dynamics of tyrosine residues in proteins, including proton transfer in bacteriorhodopsin and PYP and electron transfer in Photosystem II.

CHAPTER 5

Side-chain specific isotopic labeling of proteins for infrared structural biology: the case of ring-D_4-tyrosine isotope labeling of photoactive yellow protein

Rachana Rathod[1], **Zhouyang Kang[2]**, Steven D. Hartson[3], Masato Kumauchi[1], Aihua Xie[2], and Wouter D. Hoff[1,4]

Departments of Microbiology and Molecular Genetics[1], Physics[2], Biochemistry and Molecular Biology[3], and Chemistry[4], Oklahoma State University, Stillwater, Oklahoma 74078, USA

Abstract

An important bottleneck in the use of infrared spectroscopy as a powerful tool for obtaining detailed information on protein structure is the assignment of vibrational modes to specific amino acid residues. Side-chain specific isotopic labeling is a general approach towards obtaining such assignments. We report a method for high yield isotope editing of the bacterial blue light sensor photoactive yellow protein (PYP) containing ring-D_4-Tyr. PYP was heterologously overproduced in *E. coli* in minimal media containing ring-D_4-Tyr in the presence of glyphosate, which inhibits endogenous biosynthesis of aromatic amino acids (Phe, Trp, and Tyr). Mass spectrometry of the intact protein and of tryptic peptides unambiguously demonstrated highly specific labeling of all five Tyr residues in PYP with 98% incorporation and undetectable isotopic scrambling. FTIR spectroscopy of the protein reveals a characteristic Tyr ring vibrational mode at 1515 cm^{-1} that is shifted to 1436 cm^{-1}, consistent with that from *ab initio* calculations. PYP is a model system for protein structural dynamics and for receptor activation in biological signaling. The results described here open the way to the analysis of PYP using isotope-edited FTIR spectroscopy with side-chain specific labeling.

Key words: isotope editing; FTIR spectroscopy; protein mass spectrometry; photoactive yellow protein; glyphosate

5.1 Introduction

Fourier transform infrared (FTIR) spectroscopy is a promising technique for obtaining highly detailed structural information for protein functional structural dynamics, particularly with respect to hydrogen bonding (Nie et al, 2005, Nie, 2006, Thubagere, 2008) and proton transfer (Berthomieu and Hienerwadel, 2009, Xie and Hoff, 2009, Kotting and Gerwert, 2005, Vogel and Siebert, 2000, Xie et al, 2001, Rothschild et al, 1990). Infrared structural biology is an emerging technology that detects protein structures and structural dynamics using infrared spectroscopy. A variety of time-resolved and temperature resolved infrared spectroscopic techniques have been developed, including rapid-scan FTIR (Braiman et al, 1987, van der Horst et al, 2009, Barth et al, 1996), step-scan FTIR with microsecond and nanosecond time resolutions, picosecond pump probe infrared spectroscopy (Xie et al, 2001, Hessling et al, 1997, Hu et al, 1996, Hage et al, 1996, Hastings, 2001, Sun and Frei, 1997, Dioumaev and Braiman, 1997), and cryogenic FTIR for cold-trapped intermediate states. A Vibrational Structural Marker (VSM) database library is being developed to translate infrared signals into quantitative structural information(Nie et al, 2005). However, the development of VSM for quantitative structural characterization and applications of VSM for protein structure-function studies have been hampered by the challenges of identification and assignment of key vibrational modes in a protein to a specific side-chain. This important bottleneck limits the application of FTIR spectroscopy as a powerful tool for time-resolved protein structural biology.

Important tools in the assignment of signals to specific side chains are site directed mutagenesis (Gerwert, 1999) and isotope editing (Arkin, 2006, Das et al, 1999, Li et al, 1997, Tatulian, 2010, Warscheid et al, 2008, Zhang et al, 1994). The methods for site-directed mutagenesis are well-established, but are currently limited to the relatively slow process of residue-per-residue attempts to confirm suspected assignments. In the case of isotope editing a strategy for the specific isotopic labeling of residues is needed. One approach is homogeneous ^{13}C or ^{15}N labeling. While this can contribute to band assignment and can help resolve overlap between different signals, it is generally not sufficiently specific to yield unique assignments. Another powerful method is the use of H/D exchange. This can reveal different types of secondary structure (deJongh et al, 1997) and changes in solvent exposure during protein function (Hoff et al, 1999). However, the large number of exchanging sites in a protein usually precludes the assignment of vibrational signals to a specific residue. Approaches to isotopically label selected regions of the protein involve the overexpression of proteins in the presence of specific isotopically labeled amino acids (Muchmore et al, 1989), intein-based segmental labeling (Muona et al, 2010, Xu et al, 1999, Yamazaki et al, 1998), and *in vitro* translation systems (Sonar et al, 1994, Yabuki et al, 1998). A final approach is the complete chemical synthesis of the protein under study (Dawson and Kent, 2000). While this is highly useful for the study of small peptides (Decatur and Antonic, 1999, Decatur, 2006), it is feasible but technically challenging for entire proteins.

Intein-based techniques, *in vitro* translation, and total chemical synthesis of proteins are powerful but technically demanding approaches. The growth of *E. coli* in defined media in the presence of isotopically labeled amino acids offers an attractive strategy to side-chain specific isotopic labeling. Two possible problems associated with this approach are isotope dilution, in which endogenous synthesis of the labeled amino acid that is added to the growth medium results in a reduced level of isotopic labeling of the side chain under study, and isotope scrambling, where the cellular amino acid metabolism will cause labeling of other side chains. Such problems can be reduced by the use of auxotrophic *E. coli* strains that are unable to synthesize specific amino acids (Warscheid et al, 2008, Muchmore et al, 1989) or the use of inhibitors that block specific pathways for amino acid biosynthesis. In the case of the biosynthesis of aromatic amino acids the inhibitor glyphosate is available (Kim et al, 1990).

Recent developments in technology have made mass spectrometry a highly attractive tool to determine the degree to which problems in side-chain specific isotopic labeling occur. Significant interest exists in using isotopic labeling to aid in the interpretation of FTIR signals of proteins. The use of mass spectrometry to quantitatively determine the pattern of isotope incorporation has recently been initiated (Warscheid et al, 2008). Here we use mass spectrometry to show that specific and high-level Tyr labeling is readily achieved by overexpression of the protein under study in *E. coli*, grown in defined media in the presence of glyphosate. These studies were performed using photoactive yellow protein (PYP), a bacterial photoreceptor (Meyer, 1985, Meyer et al, 1987, Sprenger et al, 1993) and model system for functional protein dynamics (Cusanovich and Meyer, 2003, Hellingwerf et al, 2003, Kumauchi et al, 2008). The results indicate that FTIR spectroscopic studies of PYP (Xie et al, 2001, Brudler et al, 2001, Imamoto et al, 1997, Xie et al, 1996) can greatly benefit from side-chain specific isotopic labeling.

5.2 Materials and methods

Chemicals

Ring-D_4 L-tyrosine (DLM-451-1 with 98 % isotope enrichment) was purchased from Cambridge Isotope Laboratories. All other amino acids and chemicals were from Sigma-Aldrich. DEAE Sepharose^{TM} Fast flow for Anion Exchange chromatography and Sephadex^{TM} G-50 Superfine for size exclusion chromatography were purchased from GE Healthcare. Ring-D_4 L-tyrosine will be abbreviated as Tyr-D_4 throughout the text below.

Expression strain and plasmid

E.coli strain BL21 (DE3) and the plasmid pET-16b were obtained from New England Biolabs and Novagen respectively and the plasmid was constructed as reported previously (Sambrook and Gething, 1989).

Composition of the minimal medium

Growth media for specific Tyr-D_4 isotopic labeling were prepared based on (Sambrook et al, 1989) with modifications as described in (Muchmore et al, 1989). First, a 10X stock of M9 salts plus all 20 amino acids (including either Tyr or Tyr-D_4) was prepared as described in Table 5.1. Following the addition of amino acids, the medium was autoclaved for 20 minutes. The remaining ingredients (Table 5.2) were filter sterilized (except $MgSO_4$, which was autoclaved) and were added to the M9 medium containing the amino acids after it had cooled down to room temperature. Finally, 50 mg/mL ampicillin and 1 g of the aromatic amino acid biosynthesis inhibitor glyphosate (dissolved in 2 mL 10M NaOH) were added just before inoculation of the medium with *E. coli*.

PYP color check of glycerol stocks of pET16b-E.*coli* BL21(DE3)

Before using the isotopically labeled medium, a test culture was done using the natural/ unlabeled $_{L}$-tyrosine to check the viability and protein expression of the transformed *E.coli* BL21(DE3) cells maintained as glycerol stocks in -20°C. The cells were inoculated in 10 mL M9 medium with ampicillin and were incubated overnight on a shaker at 37°C. After 16 hours, the expression of *apo*-PYP was induced by 0.01 mL 1M IPTG and was reincubated in the shaker. After 4 hours, the cells were harvested at 3750 rpm for 10 min. after which they were resuspended in 1 mL 8M Urea and stirred for 30 min on ice. The lysed cells were spun down at 15000 rpm for 20 min and the supernatant was diluted with equal volume of 10 mM Tris-HCl, pH 7.5 followed by the addition of 100 µL/1L culture of the anhydride of *p*-coumaric acid, which was synthesized as previously described (Imamoto et al, 1995). The reaction mixture was incubated for 30 min in 4°C and to check proper reconstitution of PYP with its *p*-coumaric acid chromophore, the mixture was spun down at 15,000 rpm for 5 min and the Abs_{446} was recorded before preparing the isotopically labeled medium.

Purification of Tyr-D_4 PYP

Growth medium containing Tyr-D_4 was inoculated with *E. coli* expressing high levels of PYP as determined by the above color check procedure. After reconstitution with the chromophore, the reaction mixture was dialyzed overnight against 10 mM Tris-HCl, pH 7.5 at 4°C in a dark room. The mixture was then applied to a DEAE Sepharose column prewashed and equilibrated with 10 mM Tris-HCl, pH 7.5. After washing the column with the same buffer, Tyr-D_4 PYP was eluted in fractions in the same buffer with 100 mM NaCl and its optical purity index (PI) (Abs_{278}/Abs_{446}) was measured. When needed, a more refined elution was performed by using a gradient of NaCl (50-200 mM), in which PYP eluted around 80-120 mM NaCl. The purification process was repeated and combined with size exclusion chromatography until a PI of 0.43 was obtained. Before applying the Tyr-D_4 PYP sample on size exclusion column, it was pre-equilibrated with 100 mM NaCl. The sample used for application was concentrated using an ultrafiltration membrane (Amicon, Centriprep-10) to 1/100th volume of the column. Tyr-D_4 PYP was eluted in fractions with 100 mM NaCl and the PI for each was determined. Tyr-D_4 PYP with a PI of 0.43 was used for further analysis.

Mass spectrometric analysis of isotopic labeling on PYP

To determine the MW of intact protein, the protein sample was diluted in 60%:40%:10% acetonitrile/water/methanol mixture containing 0.1% formic acid. Samples were ionized by infusion using a New Objective ion source and metal coated infusion tips (Econotips) from New Objective. The positive ion signals were recorded manually using the FT mass analyzer of an LTQ Orbitrap LX mass spectrometer, using a resolution setting of 7,500 to collect a mass spectrum showing the protein's average mass without resolved isotopes and using a resolution setting of 100,000 to resolve individual isotopes. The highly resolved mass spectra were used to estimate charge states, and these estimates were used for manual deconvolution to calculate charge and MW.

To determine the efficiency of protein labeling, the labeled protein was denatured in 8M urea, reduced with TCEP to disrupt disulfide bonds, alkylated with iodoacetamide to prevent reoxidation of Cys residues, and digested overnight with 8 μg/ml trypsin, using 100 mM Tris pH 8.5 to buffer all solutions. Trypsinolytic peptides were purified using C18 affinity tips (OMIX), and analyzed by on a hybrid LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific) coupled to a New Objectives PV-550 nanoelectrospray ion source and an Eksigent NanoLC-2D chromatography system. Peptides were analyzed by trapping on a 2.5 cm pre-column and analytical separation on a 75 μm ID fused silica column, using a vented column configuration packed in house with 10 cm of Magic C18 AQ and terminating with an integral fused silica emitter pulled in house. Peptides were eluted using a 5-40% ACN/0.1% formic acid gradient performed over 40 min at a flow rate of 300 nL/min. During each one-second full-range FT-MS scan (nominal resolution of 60,000 FWHM, 300 to 2000 m/z), the three most intense ions were analyzed via MS/MS in the linear ion trap. MS/MS settings used a trigger threshold of 8,000 counts, mono-isotopic precursor selection (MIPS), and rejection of parent ions that had unassigned charge states, were previously identified as contaminants on blank gradient runs, or were previously selected for MS/MS (data dependent acquisition using a dynamic exclusion for 150% of the observed chromatographic peak width). Centroided ion masses were extracted using the extract_msn.exe utility from Bioworks 3.3.1 and were used for database searching with Mascot v2.2.04 (Matrix Science) and X! Tandem v2007.01.01.1 (www.thegpm.org).

Searches of the whole SwissProt database (downloaded 04/10/10 and containing 516,081 authentic sequences and an equal number of reversed decoy sequences) were conducted using the following search parameters: 15 ppm parent ion mass tolerance, 0.8 Da fragment ion tolerance, cleavage with trypsin (allowing one missed cleavage), and the variable modifications pyroglutamate modification of N-terminal Q, oxidization of M, alkylation of C with iodoacetamide, and formylation or acetylation of the proteins' N termini.

Peptide and protein identifications were validated using Scaffold v2.2.00 (Proteome Software) and the PeptideProphet algorithm (Keller et al, 2002). Probability thresholds were greater than 95% probability for protein identifications, based upon at least 2 peptides identified with 80% certainty, providing an experiment-specific protein false discovery rate of 0.1%. Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony.

Using the m/z ratios of the peptides thus identified, chromatograms for each PYP-derived peptide ion were manually extracted, smoothed, and peak areas calculated using Xcalibur. Peak areas were compared for the Tyr-D_4 labeled peptides with the corresponding unlabeled peptides to calculate the labeling percentage.

FTIR spectroscopy

Tyr-D_4 PYP and Hh PYP samples for FTIR spectroscopy were prepared by concentrating and then washing with 50 mM phosphate buffer in D_2O, using a Microcon (YM-10, Millipore) centrifuge filter. As is commonly done for setting the pH of protein samples in D_2O for NMR spectroscopy, the pH of the phosphate buffer in D_2O used here for FTIR spectroscopy was set at 7.1 using a pH electrode that was calibrated using pH standard solutions in H_2O, referred to as a pH* 7.1. Each FTIR sample consisted of 2.7 μl of PYP at 8mM protein concentration sandwiched between two 15 mm diameter BaF_2 windows and separated using a 12 μm spacer. A pair of 15 mm diameter BaF_2 windows without any sample was used as reference. The temperature of samples during the IR measurements was maintained at 300 K using a water circulation system (NESLAB, RTE-111). PYP samples were measured at steady state with a Bruker IFS 66v FTIR spectrometer in the range of 4000 – 600 cm^{-1} at 2 cm^{-1} spectral resolution and 40 kHz scanning velocity, using a liquid nitrogen-cooled Mercury Cadmium Telluride (MCT) detector. During the measurement, the water vapor in the optic path was vacuumed out, and the sample chamber was purging with dry nitrogen gas (2.2 liter per minute). The absorption data was averaged over 160 scans. The 2nd derivative of absorbance was calculated using Bruker OPUS software (version 5.5), based on Savitzky-Golay algorithm with 13 smoothing points.

Ab *initio* vibrational calculations

A 4-propylphenol molecule forming two hydrogen bonds as acceptor and donor with 2 isopropyl alcohol molecules was employed to model the side chain of Tyr residues in protein. The preliminary structural optimization of the input structure was performed usingPM3 method, a semi-empirical method. Density functional theory, an *ab initio* method, was employed for highly accurate structural optimization and vibrational frequency calculation using B3LYP/6-31G(d). All structural and frequency calculations were carried out using Gaussian03 (Frisch et al, 2003). A scaling factor of 0.9668 was used to compensate the overestimated force constants in order to calibrate computational and experimental results.

5.3 Results and discussion

Overexpression of PYP in minimal media.

Current overexpression protocols for PYP (Kumauchi et al, 2002) use Luria broth or 2X YT to grow *E. coli* BL21 (DE3) containing the *pyp* gene from *Halorhodospira halophila* in pET or pQE expression vectors. To allow side-chain specific isotopic labeling, growth in defined media is needed. We found that growth in M9 medium supplemented with all 20 amino acids and glyphosate (Table 5.1) yielded approximately 50 mg of PYP per liter culture upon reconstitution and dialysis. Under similar conditions we obtained essentially indistinguishable yields of PYP per liter culture after growth in LB.

While the use of His-tagged PYP for convenient purification by Ni-NTA chromatography has been effectively used, for a range of biophysical measurements it is preferable to use untagged, native PYP. Here we use anion exchange and size exclusion chromatography to obtain untagged Tyr-D_4 PYP of very high purity as indicated by its purity index (ratio of absorbance at 280 nm and 446 nm) of 0.43, which is widely used as the standard for pure *H. halophila* PYP (Kumauchi et al, 2002). Typically 40 to 50% of the PYP observed immediately following sample reconstitution and dialysis was recovered in the final purified protein sample.

To achieve side-chain specific labeling with the Tyr-D_4 derivative, the same conditions were used, but the Tyr in the growth medium was replaced by Tyr-D_4 at 50 mg per liter. The ring-D_4-Tyr used in these experiments was 98% Ring-D_4 labeled. Generally, before performing growth in the presence of isotopically labeled compounds, we performed a check of the quality of the overexpression strain by small scale incubation (color-check procedure; details in Materials and Methods section). The yield and UV/vis absorption maxima of both labeled and unlabeled PYP reconstituted from *E.coli* grown in either LB or M9 growth medium were comparable.

Mass spectrometric determination of the isotopic labeling pattern.

To determine the labeling efficiency of the five Tyr residues in PYP, and the possibility of isotope scrambling to Phe (or other) side chains, we performed mass spectrometry of full-length unlabeled and Tyr-D_4-labeled PYP. We also analyzed labeling efficiency by measuring the abundance of labeled and unlabeled peptides produced upon trypsin digestion of Tyr-D_4 PYP.

First, we measured the mass spectrum of the unlabeled and Tyr-D_4 labeled samples at lower resolution, in which individual isotope effects were not resolved. The m/z ratios of the five species with +9 to +13 charges yielded experimental molecular weights of 14019.09 ± 0.12 Da for unlabeled PYP. Based on the predicted molecular weight of unlabeled PYP of 14019.12 Da calculated based on its covalent structure, the experimentally observed value is in agreement with the observed mass within 0.04 Da. The experimentally determined mass for Tyr-D_4 PYP was 14038.56 ± 0.06 (Figure 5.1A). This is 19.44 ± 0.14 heavier than unlabeled PYP. Since the expected increase in mass of the five Tyr-D_4 residues in the labeled PYP is 20.16, this measurement provides an estimate of the labeling percentage (19.44/20.16) of 96.4 ± 0.7 %.

In the isotope-resolved high-resolution spectra (Figure 5.1B,C) of unlabeled and Tyr-D_4 labeled PYP a series of peaks is detected, which is caused by the natural variation in the number of ^{13}C nitrogen atoms (1.1% natural abundance) in PYP. As a result, series of adjacent peaks that differ by the mass of one neutron are observed in these spectra. The analysis of these data requires that for the unlabeled and Tyr-D_4 labeled samples peaks corresponding to the same level of ^{13}C incorporation are selected. When the central peaks with m/z values of 1169.24 (unlabeled) and 1170.91 (Tyr-D_4) in Figure 5.1 are selected, this corresponds to an increase in mass of 20.04 Da, very close to the 20.16 Da expected for the increase in mass due to Tyr-D_4 labeling. However, this selection of corresponding peaks is not fully unambiguous. This complication does not occur in the analysis of the mass spectra measured at lower resolution, as described above.

More detailed insights into the labeling pattern and labeling efficiency of the Tyr-D_4 PYP sample was obtained through LC-MS/MS analysis of tryptic digests. A total of 12 different trypsinolytic PYP-derived peptides were detected. Together, these peptides covered 91% of all residues in PYP (Figure 5.2), including all five Tyr residues. All of the residues not detected were present in sequences in which two Lys residues were in close proximity (within 5 residues or fewer); the small size of the resulting peptides can explain their absence in the LC-MS/MS data.

Each of the 12 PYP-derived peptides was unambiguously identified by its MS/MS fragmentation pattern (Figure 5.3), using statistically based proteomics scoring algorithms (see Methods). For Tyr-containing peptides the identification was also inspected manually to confirm that the major ions in the MS/MS spectra represented a contiguous series of b and y fragments derived from the parent peptide. High-quality MS/MS data were obtained, as judged from (i) the high percentage of MS/MS fragments that could be assigned to the trypsinolytic PYP-derived mother peptide, and (ii) the presence of a continuous ladder of fragments (see Figure 5.3). In these data the majority of the detected were generated through single fragmentation events that cleave the mother peptide into two fragments (b and y ions). Figure 5.3 provides an example of the mapping of the observed fragments onto the amino acid sequence for the peptide KALSGDSYWVFVK that contains Tyr118. This analysis unambiguously demonstrated the incorporation of Tyr-D_4 in all five of the Tyr side chains in PYP (Table 5.3).

To further characterize the pattern of incorporation of Tyr-D_4 into PYP, we analyzed two issues that are of general concern for site-specific isotope editing of proteins obtained by *in vivo* expression: isotope dilution resulting in a reduced percentage of labeling of the target sites (in this case Tyr), and isotope scrambling to other side chains. To examine isotope dilution we use the observation that for three different PYP-derived peptides both the unlabeled and the Tyr-D_4 containing species were detected (Table 5.3). While the elution times of the corresponding labeled and unlabeled peptides were essentially identical, they could be distinguished based on their 4 Da increase in mass (Figure 5.4). Comparison of the areas of chromatographic elution peaks representing each labeled and unlabeled peptide quantitatively revealed very low amounts of unlabeled Tyr-containing peptides in the tryptic digest of Tyr-D_4 labeled PYP. Thus, the high coverage percentage of the amino acid sequence of PYP provides detailed information on the pattern of isotope labeling of PYP. Figure 4 depicts the elution profiles for a pair of Tyr-D_4-labeled and unlabeled peptides (for peptide KALSGDSYWVFVK), with the amplitude of the elution profile of the unlabeled peptide multiplied by a factor 45, illustrating that the unlabeled peptide is present at ~50-fold lower concentration.

Quantitative analysis of the peak areas in the elution pattern of the observed unlabeled and Tyr-D_4-labeled peptides yielded values of 2.0%, 2.3%, and 2.4% unlabeled side chains at Tyr76 and Tyr118 (Table 5.3). For the other three Tyr side chains these small contributions to the chromatogram were not detected, presumably due their intensities falling below the noise threshold. Since all five Tyr residues would be expected to be exhibit an identical labeling percentage, and since the major species in the mass spectra of intact Tyr-D_4-labeled PYP exhibited a 20 Da mass increase compared to unlabeled PYP, we conclude that the conditions used here yield the specific isotopic labeling of all Tyr residues in PYP with an efficiency of 97.7%. This corresponds well with the values of 96.4% derived above from mass spectrometry of intact PYP. The Tyr-D_4 that was used to grow the *E. coli* cultures contained a ring-D_4 labeling level of 98%, in excellent agreement with the experimentally detected level of Tyr-D_4 incorporation in PYP. These results therefore demonstrate the quantitative incorporation of the added Tyr derivative into PYP, and thus the absence of isotope dilution.

With respect to isotope scrambling, we considered the possibility of incorporation of Phe-D_4 in PYP (possibly through prephenate as an intermediate). The inclusion of Phe-D_4 in the list of possible derivatives used during the analysis of the LC-MS/MS data did not yield any PYP-derived Phe-D_4 containing peptides. Thus, isotope scrambling to Phe was undetectably low.

FTIR spectroscopy of Tyr-D_4 labeled PYP.

Unlabeled and Tyr-D_4-labeled PYP samples were spectroscopically studied using static Fourier transform infrared (FTIR) spectroscopy. To facilitate the identification of shifts in peak positions, second derivative spectra were calculated for the unlabeled and Tyr-D_4-labeled PYP. This revealed a series of characteristic peak shifts (Figure 5.5). Comparison of these data with calculated vibrational spectra for free Tyr and Tyr-D_4 illustrates that the pattern of peak shifts observed in the protein samples corresponds to that expected for ring-D_4 labeling of the Tyr side chains in PYP. The strong peak at

1515 cm^{-1} has essentially disappeared in ring-D$_4$ labeled, and has shifted to ~1435 cm^{-1}. Thus, clear signals from Tyr side chains can be detected in FTIR spectra from intact PYP. This opens the way to future time-resolved FTIR difference spectroscopic studies of these signals.

Isotopic labeling is widely used in NMR spectroscopy. An important difference with the application of isotopic labeling to FTIR spectroscopy is that both the labeled and unlabeled variants contribute to the vibrational spectrum. Because isotopic labeling usually causes only fairly small shifts in vibrational frequency, partially isotopic labeling will result in the splitting of the bands under study. This effect therefore results in additional complications of already crowded FTIR spectra. In addition, if the isotopic label is scrambled to other side chains, the vibrational signals of these groups will also be split, hampering the assignment of vibrational modes to specific types of side chains and the unambiguous interpretation of the spectra. Thus, it is greatly preferable to perform vibrational measurements on samples with a very high labeling percentage. The results reported here show that this is entirely feasible for Tyr side chains.

In the case that isotope scrambling is difficult to avoid, such as for the Asp side chain (Warscheid et al, 2008), detailed knowledge of the level of labeling of the group under study and the pattern of isotope scrambling is of great value for interpreting the FTIR spectra obtained using the samples. Such quantitative information on the pattern of isotopic labeling can be readily obtained using the approach reported here based on LC-MS/MS analysis of tryptic digests of the isotopically labeled protein. The approach reported here for PYP is applicable to studies of Tyr side chains in any protein for which overproduction and purification has been achieved and that is amenable to mass spectrometry. Since glyphosate also inhibits the synthesis of phenylalanine and tryptophan in cells, the same method is expected to work for side-chain specific isotope editing of these two residues in proteins. In addition, in combination with the use of VSM of Tyrosine (Nie, 2006), FTIR spectroscopy can provide valuable information on the hydrogen bonding status, protonation state, and redox state of Tyr side chains (Xie et al, 2001, Hienerwadel et al, 1997, Takahashi and Noguchi, 2007). Since structural and/or chemical changes of Tyr side chains frequently are essential events at the catalytic site of proteins (see for example (Hienerwadel et al, 1997, Gutteridge and Thornton, 2008)), the Tyr labeling and quantification procedure reported here is expected to be of general use.

Acknowledgements

This work was funded by NSF grant MCB-1051590 to W.D.H, OCAST grant HR10-078 to A.X., and NSF grant DBI-0722494 to S.D.H. Mass spectrometry analyses were performed in the DNA/Protein Resource Facility at Oklahoma State University, using resources supported by the NSF MRI and EPSCoR programs.

Table 5.1. Composition of minimal growth medium for Tyr-D_4 labeling of proteins

10X stock of M9 salts	
128 g	Na_2HPO_4 x $7H_2O$
30 g	KH_2PO_4
5 g	NaCl
10 g	NH_4Cl
Dilute to 1 Liter.	
The following amino acids were added after 100 mL of 10X M9 salts were diluted to 900 mL.	
For 1 Liter culture	
0.5 g	L-alanine
0.4 g	L-arginine
0.4 g	L-asparagine
0.4 g	L-aspartic acid
0.0725 g	L-cystine-HCl x H_2O
0.4 g	L-glutamine
1.413 g	Na-glutamate x 8.3 H_2O
0.55 g	L-glycine
0.143 g	L-histidine-HCl x H_2O
0.23 g	L-isoleucine
0.23 g	L-leucine
0.42 g	L-lysine-HCl
0.25 g	L-methionine
0.13 g	L-phenylalanine
0.1 g	L-proline
0.21 g	L-serine
0.23 g	L-threonine
0.17 g	L-tryptophan
0.23 g	L-valine
0.050g	L-tyrosine or L-tyrosine-ring-D_4

This growth medium is autoclaved.

Table 5.2 Additional chemicals to the minimal medium for Tyr-D_4 labeling of proteins.

For 1 Liter medium

a.	4 mL		1M $MgSO_4$
b.	1 mL		0.01M $FeCl_3$
c.	50 mL		40% glucose
d.	10 mL of solution containing		
	i	2 mg	$CaCl_2$ x $2H_2O$
	ii	2 mg	$ZnSO_4$ x $7H_2O$
	iii	2 mg	$MnSO_4$ x H_2O
	iv	2 mg	H_3BO_4
	v	2 mg	$CuSO_4$
e.	50 mg	Thiamine (dissolved in 5 mL dH_2O)	
f.	1 mg	Biotin (dissolved in 1 mL dH_2O)	
g.	50 mg	Niacin (dissolved in 1-2 mL dH_2O)	
h.	50 mg	Ampicillin (dissolved in 1 mL dH_2O)	
i.	1 g	Glyphosate (dissolved in 2 mL 10M NaOH)- *added just before inoculation.*	

Table 5.3 Summary of LC-MS/MS analysis of Tyr-containing peptides observed in the tryptic digest of PYP obtained after growth in Tyr-D$_4$ labeling medium.

Peptide	Tyr in peptide	D$_4$Y peptide m/z	Unlabeled peptide m/z	D$_4$Y peptide peak area	Unlabeled peptides peak area	% unlabel-ed	Charge on the peptide	Mascot Ion Score	X! Tandem Peptide Score
18-52	Tyr42	1209.91	1208.58	1.9 x 10^5	NM	N/A	3	81.3	6.55
18-55	Tyr42	1323.31	1321.97	2.6 x 10^5	NM	N/A	3	141.7	6.42
65-78	Tyr76	795.36	793.35	275703418	5720529	2.0	2	79.7	7.37
79-104	Tyr94, Tyr98	1009.82	1007.14	255052	NM	N/A	3	78.6	5.44
81-104	Tyr94, Tyr98	918.1	915.42	182082	NM	N/A	3	33.7	1.72
111-123	Tyr118	752.41	750.40	6055989	146122	2.4	2	96.6	6.96
112-123	Tyr118	688.37	686.35	13806710	327650	2.3	2	82.6	4.89

NM: not measurable. N/A: not applicable

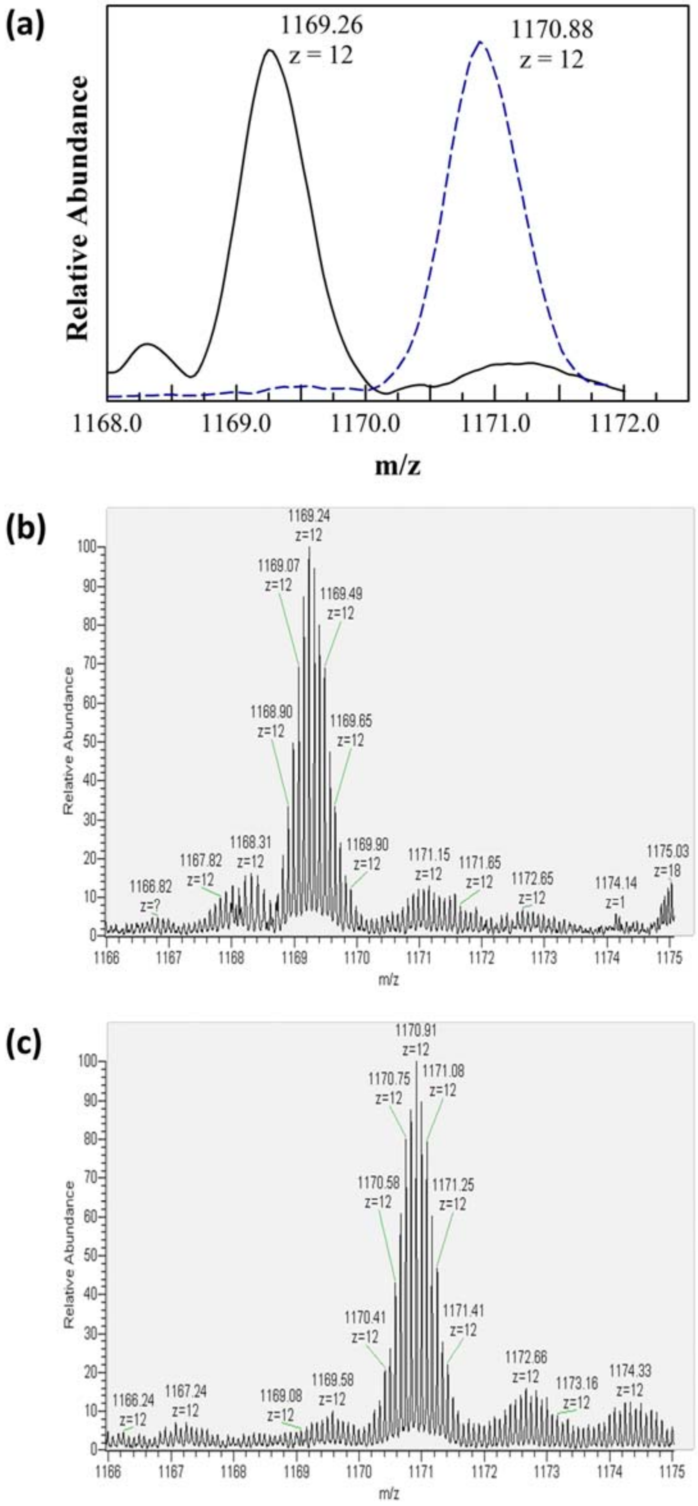


Figure 5.1 Mass spectrometry of Tyr-D_4-labeling of PYP. (A) Low-resolution mass spectra of intact unblabeled (solid black line) and Tyr-D_4-labeled (dashed blue line) PYP. The observed difference in mass caused by the Ring-D_4 labeling of the five Tyr side chains in PYP is 19.44 Da. Isotope-resolved high-resolution mass spectrum of intact unlabeled (B) and Tyr-D_4-labeled (C) PYP. Comparison of the spectra in B and C reveals a 20.04 Da increase in mass.

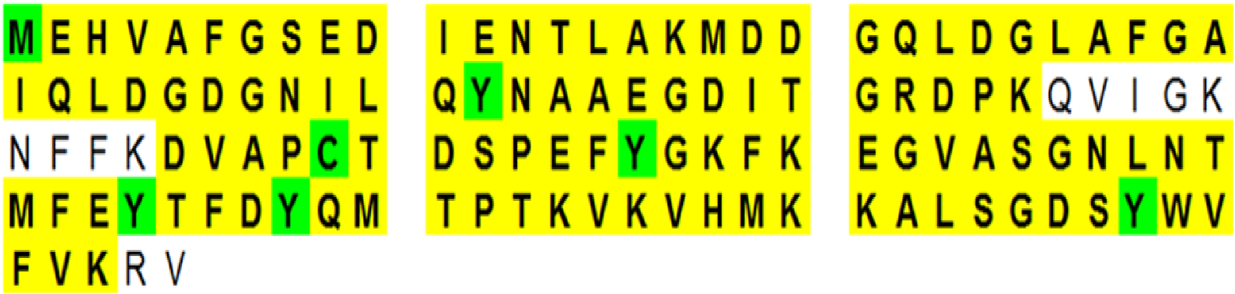


Figure 5.2 Coverage map of PYP after tryptic digestion. Residues identified in peptides through MS-MS analysis are indicated in color. The modified residues are highlighted in green: M was oxidized, C was carboxyamidomethylated using iodoacetamide, Y was Ring-D_4 labeled. All 5 Tyr side chains of PYP were detected as Ring-D_4 labeled.

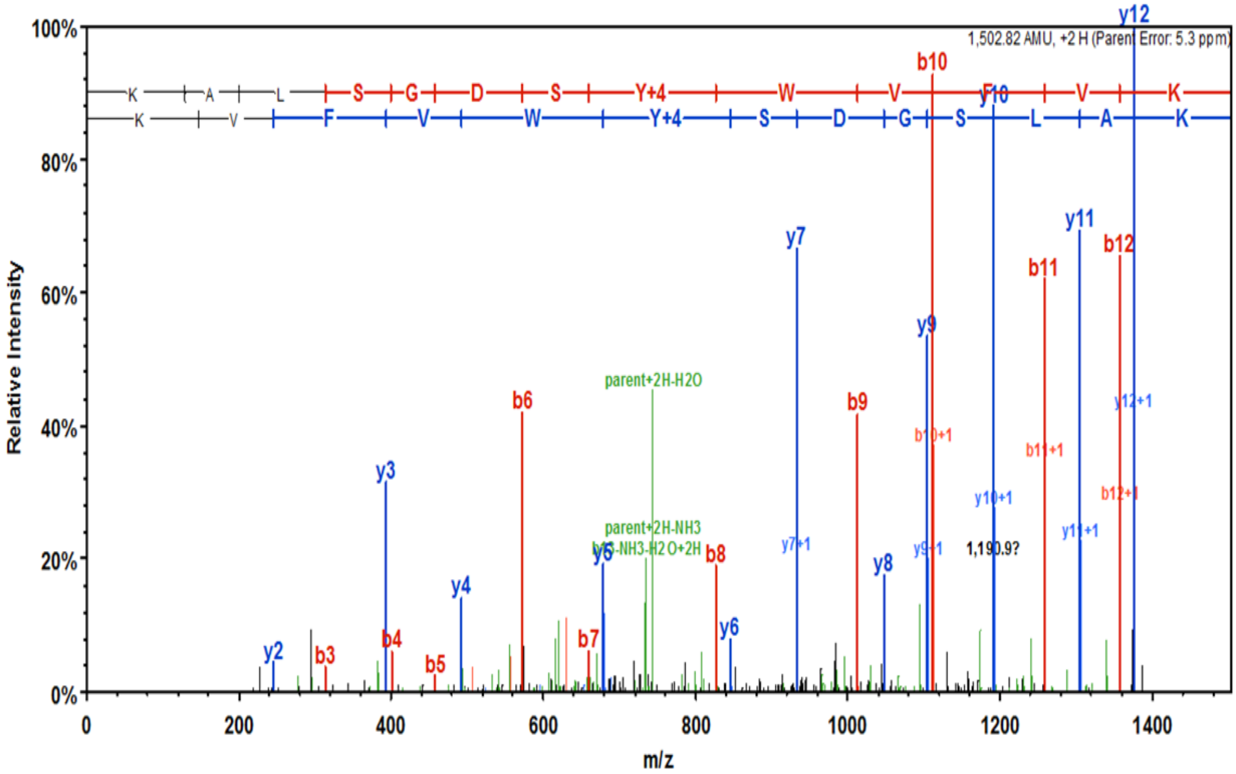


Figure 5.3 Example of MS/MS identification of Tyr118-ring-D₄ labeled peptide in the tryptic digest of PYP. The fragmentation pattern is annotated to allow identification of the amino acid sequence, with the top row of text representing the peptide sequence in an N- to C-terminal orientation to visualize the b ion fragments containing the N terminus, and the second row of text representing the peptide sequence in the C- to N-terminal orientation to visualize y ion fragments containing the C-terminus. All assigned ions (except when indicated) carry a single charge and are the result of a single fragmentation event at the peptide bond. The vertical bars in the indicated amino acid sequence line up with the m/z of the ion origination from cleavage at that site.

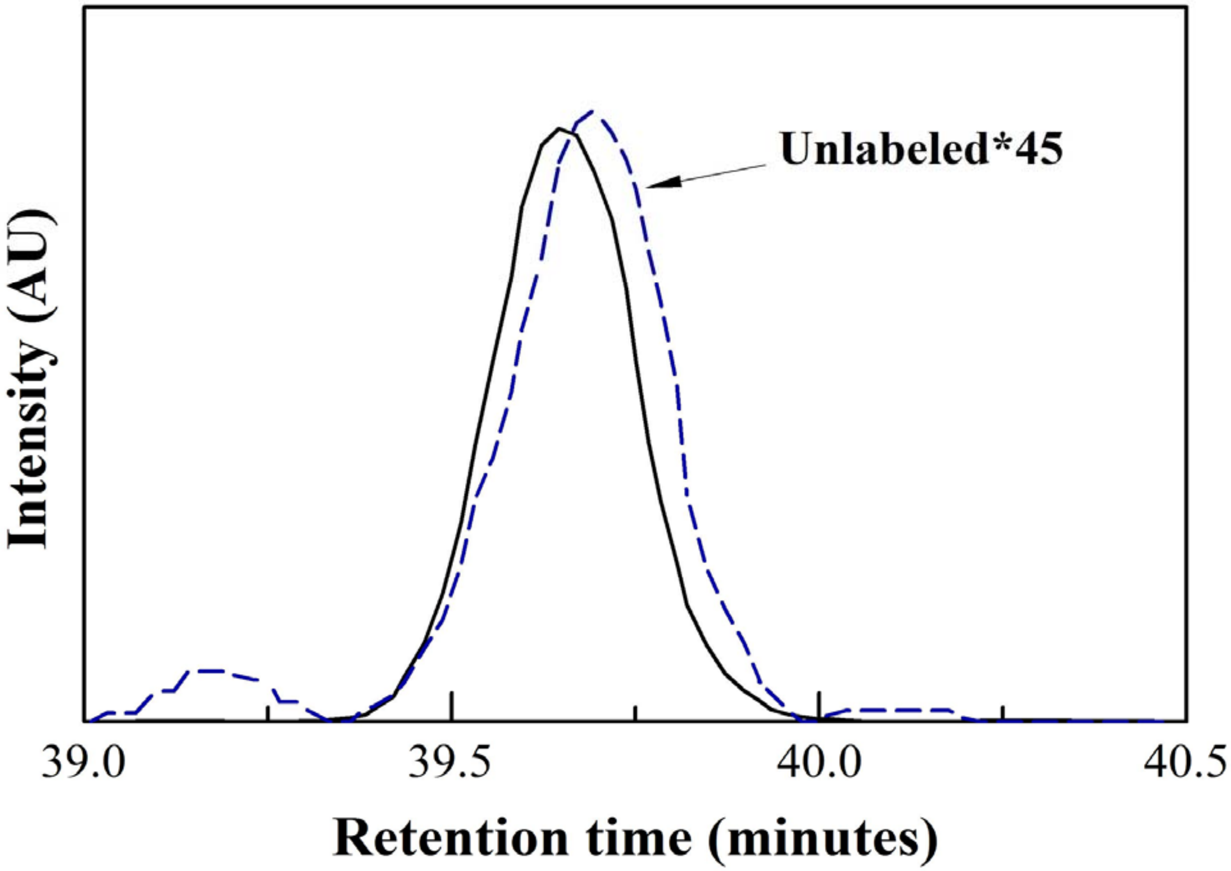


Figure 5.4 Example of the mass-resolved detection of the elution of a Tyr-D_4 peptide (black line) and the corresponding unlabeled (dashed blue line) peptide. In this figure the intensity of the mass signal of the unlabeled peptide KALSGDSYWVFVK was multiplied by a factor 45, demonstrating a ~98% labeling efficiency at Tyr118 (see Table 3 for details).

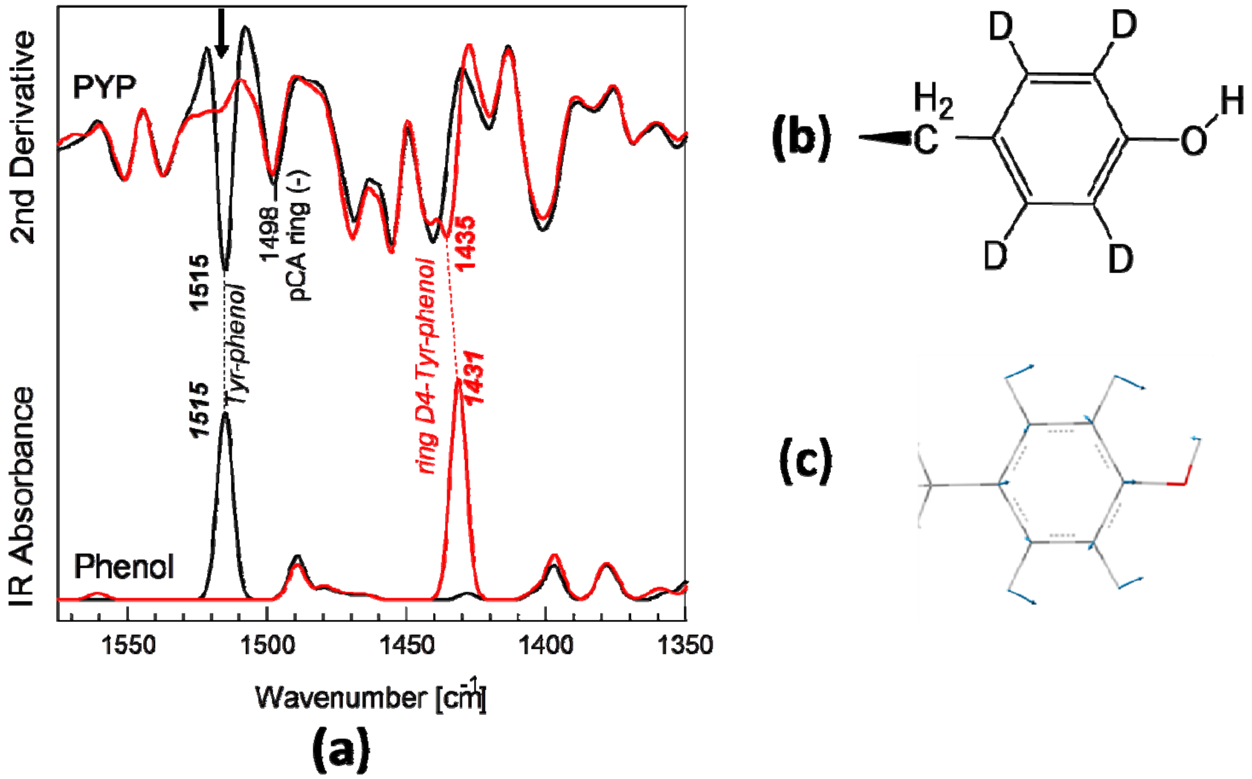


Figure 5.5 Infrared signals from ring-D$_4$-Tyr. (A) Top spectra are the second derivative of infrared absorption of hHal PYP (black) and ring-D$_4$ isotope edited hHal PYP (red). The characteristic ring mode at 1515 cm^{-1} is indicated by an arrow. The bottom two are computational data from the phenolic ring resembling the side chain of tyrosine (black) and ring-D$_4$ isotope edited side chain (red) of tyrosine. (B) The side chain structure of ring-D$_4$-Tyr. (C) The 1515 cm^{-1} vibrational mode of tyrosine with predominately CH bending coupled ring deformation. Upon ring-D$_4$ labeling, this vibrational frequency is shifted from 1515 cm^{-1} to 1435 cm^{-1} (A top), highly consistent with the band shift from ab initio computational data (A bottom).

CHAPTER 6

Summary and Concluding Remarks

6.1 Development of Infrared Structural Biology

Time-resolved Fourier transformed infrared spectroscopy is an powerful tool to study structural information of protein molecules. Not only because of its broad application in time-resolution (nano-second to second) which provides outstanding accessibility to protein dynamics, but also because of its excellent sensitivity to side chain protonation/deprotonation, hydrogen-bonding interactions.

Hydrogen-bonding interactions, particularly of those formed between buried residue side chains inside proteins, are important structural elements. Hydrogen-bond dissociation energy is suggested to be in the range of 2-40 kJ/mol for neutral ordinary hydrogen-bonding interactions, and this value could be much higher for ionic hydrogen-bonding interactions (Pace, 2009, Pace et al, 2014). Since protein folding energy is around 40 kJ/mol (Creighton, 1993), the changes of hydrogen-bonding interactions is capable of destabilizing protein folding. In fact, reviews of many important biological functions of proteins, including proton transfer, electron transfer and enzymatic catalysis all show involvement of hydrogen-bonding interactions (Chapter 3-4).

The survey of solvent exposure accessibilities for Asp/Glu residues (Chapter 3) and Tyr residues (Chapter 4) in 1182 protein crystal structures revealed that most of Asp/Glu residues are somewhat exposed to either the solvent environment outside proteins, or to water molecules inside proteins. Only 6.2% of Asp and 4.5% of Glu residues are found fully buried in a hydrophobic interior inside proteins. Meanwhile, despite of its polar –OH group, 26.1% of Tyr residues are found with their phenolic group fully buried.

Further statistical analysis of hydrogen-bonding interactions on those fully buried Asp/Glu and Tyr residues from same dataset of protein crystal structures summarized the preference of their hydrogen-bond partners. Backbone is found to be the most common hydrogen-bond partners for Asp/Glu and Tyr side chains, which is likely due to the fact that backbone is the most common structure in protein peptides. Amon side chain of amino acids, Arg residue, which mostly carries a positive charge in proteins, is by far the most favorite hydrogen-bond partner for buried carboxylate groups from Asp/Glu. While for Tyr, we see most of the hydrogen-bonding interactions are formed with Asp/Glu residues which likely carry negative charges. In addition, we found buried Tyr residues without any hydrogen-bonding interaction is very rare, only 2.9%. And from statistical results we found the chance of Tyr being a hydrogen-bond donor is almost twice as its chance of being a hydrogen-bond acceptor. In later part of Chapter 4 we predicted that Tyr is a much more preferred hydrogen-bond donor than hydrogen-bond acceptor, by applying Boltzmann distribution on calculated hydrogen-bond energies. The statistical data support our prediction.

(1) For carboxylate groups from Asp/Glu, we developed 2D VSMs using frequencies of asymmetric stretching and symmetric stretching. Our data indicate that generally as the total number of hydrogen-bonding interactions increases, red-shifts are observed on frequency of asymmetric stretching and blue-shifts are observed on symmetric stretching (Table 3.9). It should be noticed that when the number of hydrogen-bonding interactions are the same, the geometry of hydrogen-bond complex may be different. Therefore, based on the calculated hydrogen-bond energies, we

predicted the probabilities of each geometry being formed using Boltzmann distribution. Our data implies that geometries with uneven hydrogen-bond distribution are unlikely to be formed (Table 3.5-Table3.8).

(2) For phenolic group from Tyr, we developed 2D VSMs for ring-D_4 Tyr, using frequencies of C-O stretching (ν_{CO}) and COH bending (δ_{COH}-α and δ_{COH}-β). We chose ring-D_4 Tyr instead of ring-H4 Tyr mainly because the C-O stretching and C-O-H bending for ring-H_4 Tyr are coupled when strong hydrogen-bonding interactions are formed, making it difficult to characterize the corresponding signals in real infrared spectrum of proteins. For ring-D_4 Tyr, however, we found the frequencies of C-O stretching and C-O-H bending are separated enough to be identified individually. Although observed that C-O-H bending mode at two positions for ring-D_4 Tyr, δ_{COH}-α at 1300 cm^{-1} and δ_{COH}-β at ~1400 cm^{-1} or higher. We presented two 2D VSMs, one uses δ_{COH}-α vs. ν_{CO} (Figure 4.12), and one uses δ_{COH}-β vs. ν_{CO} (Figure 4.13). Our results have shown that δ_{COH}-β vs. ν_{CO} is particularly useful to identify ionic hydrogen-bonding interactions, as positively charged hydrogen-bond donor paring with Tyr results frequency of ν_{CO} lower than 1210 cm^{-1}, and negatively charged hydrogen-bond acceptor paring with Tyr results frequency of δ_{COH}-β higher than 1470 cm^{-1}. And the other types of hydrogen-bonding interactions can be solved using the combination of the two 2D VSMs.

6.2 Applications of Infrared Structural Biology

In order to apply infrared structural biology for protein crystal structures, we need to first identify signals of specific residues from infrared spectrum of whole protein. The combination of two experimental techniques are very useful for this purpose: site-directed mutagenesis and site-specific isotopic labeling. I presented our work of labeling Tyr residues in PYP with ring-D_4 Tyr (Chapter 5). PYP was heterologously overproduced in *E. coli* in minimal media containing ring-D_4 Tyr in the presence of glyphosate, which inhibits endogenous biosynthesis of aromatic amino acids, including Tyr. Mass spectrometry of the intact protein and of tryptic peptides unambiguously demonstrated highly specific labeling of all five Tyr residues in PYP with 98% incorporation and undetectable isotopic scrambling. FTIR spectroscopy of the protein reveals a characteristic Tyr ring vibrational mode at 1515 cm-1 that is shifted to 1436 cm-1, consistent with that from ab initio calculations.

We have also engineered site-directed mutagenesis of Tyr to Phe in PYP for Tyr76, Tyr94, Tyr98 and Tyr118. The only Tyr residue not mutated is Tyr42, mainly because the mutation product Y42F PYP have been found not functional (Philip et al, 2010, Imamoto et al, 2001, Brudler et al, 2000). We then prepared samples of PYP mutant with ring-D_4 Tyr. The Step-scan FTIR data measured on ring-H_4 Tyr PYP, ring-D_4 PYP and static FTIR data measured on ring-H_4 Tyr PYP, ring-D_4 PYP as well as four mutant Y76F PYP, Y94F PYP, Y98F PYP and Y118F PYP with ring-D_4 PYP were compared with each other and analyzed. The results suggest potential evidences of Tyr42 breaking hydrogen-bonding interaction during pR to pB' transition, which supports the proposed proton transfer mechanism. However, due to the reason that these data are still under further investigation, I did not include them in this thesis.

6.3 Future Outlook

Many interesting questions still remain unanswered. The core of developing Infrared Structural Biology, is to develop a Vibrational Structural Marker Library for amino acid residues in proteins. With the establishment of such library researchers will be able to use VSM stored in the library to characterize structural details of important amino acid residues in proteins. So far, we have developed VSM for protonated Asp/Glu as reported by Beining Nie (Nie et al, 2005), and VSMs for deprotonated Asp/Glu and protonated Tyr reported in this thesis. There are many other VSMs for polar side chains of amino acid waiting to be developed, including Ser/Thr, His(0)/His(+), Lys(+), Arg(+), etc. The completion of this VSM library will significantly enhance the power of infrared spectroscopy as a tool for the investigation of protein structures, and is expected to provide a huge boost to the field of Infrared Structural Biology.

As it has been discussed in this thesis, an important application of the VSM we have developed is to study the changes of hydrogen-bonding network around pCA chromophore, Tyr42 and Glu46 at the active site of PYP during photocycle. The first step is to locate the infrared signals belong to Tyr42 and extract corresponding information of hydrogen-bonding interaction changes of Tyr42. The manuscript reports these results is currently under preparation. However, this study will not reflect the details of hydrogen-bond partners paring with Tyr42 during photocycle. In order to complete story, future works of isotopic labeling of Glu residues and pCA chromophore in PYP are needed. Therefore their infrared signals during photocycle will be able to identified and the encoded structural details could be extracted, using the data stored in VSM library.

REFERENCES

1. Looking at Structures: Dealing with Coordinates.
2. Ames, J. B., M. Ros, J. Raap, J. Lugtenburg, and R. A. Mathies. 1992. Time-resolved ultraviolet resonance Raman studies of protein structure: application to bacteriorhodopsin. Biochemistry 31:5328-5334.
3. Arkin, I. T. 2006. Isotope-edited IR spectroscopy for the study of membrane proteins. Curr Opin Chem Biol 10:394-401.
4. Arnis, S., K. Fahmy, K. P. Hofmann, and T. P. Sakmar. 1994. A conserved carboxylic acid group mediates light-dependent proton uptake and signaling by rhodopsin. The Journal of biological chemistry 269:23879-23881.
5. Arunan, E., G. R. Desiraju, R. A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D. C. Clary, R. H. Crabtree, J. J. Dannenberg, P. Hobza, H. G. Kjaergaard, A. C. Legon, B. Mennucci, and D. J. Nesbitt. 2011. Defining the hydrogen bond: An account (IUPAC Technical Report). Pure and Applied Chemistry 83:1619-1636.
6. Asher, S. A., P. J. Larkin, and J. Teraoka. 1991. Ultraviolet resonance Raman and absorption difference spectroscopy of myoglobins: titration behavior of individual tyrosine residues. Biochemistry 30:5944-5954.
7. Baca, M., G. E. Borgstahl, M. Boissinot, P. M. Burke, D. R. Williams, K. A. Slater, and E. D. Getzoff. 1994. Complete chemical structure of photoactive yellow protein: novel thioester-linked 4-hydroxycinnamyl chromophore and photocycle chemistry. Biochemistry 33:14369-14377.
8. Balashov, S. P., E. S. Imasheva, R. Govindjee, M. Sheves, and T. G. Ebrey. 1996. Evidence that aspartate-85 has a higher pK(a) in all-trans than in 13-cisBacteriorhodopsin. Biophysical journal 71:1973-1984.
9. Barry, B. A., and G. T. Babcock. 1987. Tyrosine radicals are involved in the photosynthetic oxygen-evolving system. Proc Natl Acad Sci USA 84:7099-7103.
10. Barth, A., F. vonGermar, W. Kreutz, and W. Mantele. 1996. Time-resolved infrared spectroscopy of the Ca2+-ATPase - The enzyme at work. The Journal of biological chemistry 271:30637-30646.
11. Bartlett, G. J., C. T. Porter, N. Borkakoti, and J. M. Thornton. 2002. Analysis of catalytic residues in enzyme active sites. Journal of molecular biology 324:105-121.
12. Baturin, S. J., M. Okon, and L. P. McIntosh. 2011. Structure, dynamics, and ionization equilibria of the tyrosine residues in Bacillus circulans xylanase. Journal of biomolecular NMR 51:379-394.
13. Becke, A. D. 1993. Density-functional thermochemistry. III. The role of exact exchange. J. Chem. Phys. 98:5648-5652.
14. Bergethon, P. R. 1998. The physical basis of biochemistry. Springer-Verlag New York, Inc., New York.
15. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. Nucleic Acids Research 28:235-242.

16. Berthomieu, C., and R. Hienerwadel. 2009. Fourier transform infrared (FTIR) spectroscopy. Photosynth Res 101:157-170.

17. Berthomieu, C., R. Hienerwadel, A. Boussac, J. Breton, and B. A. Diner. 1998. Hydrogen bonding of redox-active tyrosine Z of photosystem II probed by FTIR difference spectroscopy. Biochemistry 37:10547-10554.

18. Borgstahl, G. E., D. R. Williams, and E. D. Getzoff. 1995. 1.4 A structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore. Biochemistry 34:6278-6287.

19. Bousche, O., M. Braiman, Y. W. He, T. Marti, H. G. Khorana, and K. J. Rothschild. 1991. Vibrational Spectroscopy of Bacteriorhodopsin Mutants - Evidence That Asp-96 Deprotonates during the M-]N Transition. The Journal of biological chemistry 266:11063-11067.

20. Bousche, O., S. Sonar, M. P. Krebs, H. G. Khorana, and K. J. Rothschild. 1992. Time-resolved Fourier transform infrared spectroscopy of the bacteriorhodopsin mutant Tyr-185-->Phe: Asp-96 reprotonates during O formation; Asp-85 and Asp-212 deprotonate during O decay. Photochem Photobiol 56:1085-1095.

21. Braiman, M. S., P. L. Ahl, and K. J. Rothschild. 1987. Millisecond Fourier-transform infrared difference spectra of bacteriorhodopsin's M412 photoproduct. Proc Natl Acad Sci USA 84:5221-5225.

22. Braiman, M. S., T. Mogi, T. Marti, L. J. Stern, H. G. Khorana, and K. J. Rothschild. 1988. Vibrational Spectroscopy of Bacteriorhodopsin Mutants - Light-Driven Proton Transport Involves Protonation Changes of Aspartic-Acid Residue-85, Residue-96, and Residue-212. Biochemistry 27:8516-8520.

23. Brudler, R., T. E. Meyer, U. K. Genick, S. Devanathan, T. T. Woo, D. P. Millar, K. Gerwert, M. A. Cusanovich, G. Tollin, and E. D. Getzoff. 2000. Coupling of hydrogen bonding to chromophore conformation and function in photoactive yellow protein. Biochemistry 39:13478-13486.

24. Brudler, R., R. Rammelsberg, T. T. Woo, E. D. Getzoff, and K. Gerwert. 2001. Structure of the I1 early intermediate of photoactive yellow protein by FTIR spectroscopy. Nat Struct Biol 8:265-270.

25. Cabaniss, S. E., and I. F. McVey. 1995. Aqueous infrared carboxylate absorbances: Aliphatic monocarboxylates. Spectrochim Acta A 51:2385-2395.

26. Chirgadze, Y. N., O. V. Fedorov, and N. P. Trushina. 1975. Estimation of amino acid residue side-chain absorption in the infrared spectra of protein solutions in heavy water. Biopolymers 14:679-694.

27. Creighton, T. E. 1993. Proteins: Structures and Molecular Properties. W. H. Freeman.

28. Cusanovich, M. A., and T. E. Meyer. 2003. Photoactive yellow protein: a prototypic PAS domain sensory protein and development of a common signaling mechanism. Biochemistry 42:4759-4770.

29. Das, A., V. Prashar, S. Mahale, L. Serre, J. L. Ferrer, and M. V. Hosur. 2006. Crystal structure of HIV-1 protease in situ product complex and observation of a low-barrier hydrogen bond between catalytic aspartates. Proc Natl Acad Sci USA 103:18464-18469.

30. Das, K. P., L. P. Choo-Smith, J. M. Petrash, and W. K. Surewicz. 1999. Insight into the secondary structure of non-native proteins bound to a molecular chaperone alpha-crystallin - An isotope-edited infrared spectroscopic study. The Journal of biological chemistry 274:33209-33212.

31. Dawson, P. E., and S. B. Kent. 2000. Synthesis of native proteins by chemical ligation. Annual review of biochemistry 69:923-960.

32. de Wijn, R., and H. J. van Gorkom. 2002. S-state dependence of the miss probability in Photosystem II. Photosynth Res 72:217-222.

33. Decatur, S. M. 2006. Elucidation of residue-level structure and dynamics of polypeptides via isotope-edited infrared spectroscopy. Acc Chem Res 39:169-175.

34. Decatur, S. M., and J. Antonic. 1999. Isotope-edited infrared spectroscopy of helical peptides. J Am Chem Soc 121:11914-11915.

35. deJongh, H. H. J., E. Goormaghtigh, and J. M. Ruysschaert. 1997. Monitoring structural stability of trypsin inhibitor at the submolecular level by amide-proton exchange using Fourier transform infrared spectroscopy: A test case for more general application. Biochemistry 36:13593-13602.

36. Dennington, R. K., T.; Millam, J. 2009. GaussView. J. Semichem Inc., Shawnee Mission KS.

37. Dioumaev, A. K., and M. S. Braiman. 1997. Two bathointermediates of the bacteriorhodopsin photocycle, distinguished by nanosecond time-resolved FTIR spectroscopy at room temperature. The journal of physical chemistry. B 101:1655-1662.

38. Engelhard, M., K. Gerwert, B. Hess, W. Kreutz, and F. Siebert. 1985. Light-driven protonation changes of internal aspartic acids of bacteriorhodopsin: an investigation by static and time-resolved infrared difference spectroscopy using [4-13C] aspartic acid labeled purple membrane. Biochemistry 24:400-407.

39. Foresman, J. B., and Æ. Frisch. 1996. Exploring chemistry with electronic structure methods. Gaussian Inc., Pittsburgh.

40. Frisch, M. J., G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, J. Vreven, T., K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. 2003. Gaussian 03, Revision A.1. Gaussian, Inc., Pittsburgh, PA.

41. Frisch, M. J. T., G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J;. 2009. Gaussian 09. Gaussian, Inc., Wallingford CT.

42. Gerwert, K. 1999. Molecular reaction mechanisms of proteins monitored by time-resolved FTIR-spectroscopy. Biol Chem 380:931-935.

43. Getzoff, E. D., K. N. Gutwin, and U. K. Genick. 2003. Anticipatory active-site motions and chromophore distortion prime photoreceptor PYP for light activation. Nat Struct Biol 10:663-668.

44. Gilchrist, M. L., J. A. Ball, D. W. Randall, and R. D. Britt. 1995. Proximity of the Manganese Cluster of Photosystem-Ii to the Redox-Active Tyrosine Y-Z. Proc Natl Acad Sci USA 92:9545-9549.

45. Gorbitz, C. H., and M. C. Etter. 1992. Hydrogen-Bonds to Carboxylate Groups - the Question of 3-Center Interactions. J Chem Soc - Perkin Transactions 2:131-135.

46. Govindjee, R., M. Kono, S. P. Balashov, E. Imasheva, M. Sheves, and T. G. Ebrey. 1995. Effects of Substitution of Tyrosine-57 with Asparagine and Phenylalanine on the Properties of Bacteriorhodopsin. Biochemistry 34:4828-4838.

47. Griffiths, P., and J. A. De Haseth. 2007. Fourier Transform Infrared Spectrometry. Wiley.

48. Gutteridge, A., and J. M. Thornton. 2005. Understanding nature's catalytic toolkit. Trends in biochemical sciences 30:622-629.

49. Gutteridge, A., and J. M. Thornton. 2008. Understanding Nature's Catalytic Toolkit. Trends Biochem. Sci. 30:1153-1165.

50. Hage, W., M. Kim, H. Frei, and R. A. Mathies. 1996. Protein dynamics in the bacteriorhodopsin photocycle: A nanosecond step-scan FTIR investigation of the KL to L transition. J Phys Chem-Us 100:16026-16033.

51. Hastings, G. 2001. Time-resolved step-scan Fourier transform infrared and visible absorption difference spectroscopy for the study of photosystem I. Applied Spectroscopy 55:894-900.

52. Heberle, J., D. Oesterhelt, and N. A. Dencher. 1993. Decoupling of Photo-Cycle and Proton Cycle in the Asp85->Glu Mutant of Bacteriorhodopsin. Embo J 12:3721-3727.

53. Hellingwerf, K. J., J. Hendriks, and T. Gensch. 2003. Photoactive Yellow Protein, A New Type of Photoreceptor Protein: Will This “Yellow Lab” Bring Us Where We Want to Go?‖. J Phys Chem A 107:1082-1094.

54. Hessling, B., J. Herbst, R. Rammelsberg, and K. Gerwert. 1997. Fourier transform infrared double-flash experiments resolve bacteriorhodopsin's M-1 to M-2 transition. Biophysical journal 73:2071-2080.

55. Hienerwadel, R., A. Boussac, J. Breton, B. A. Diner, and C. Berthomieu. 1997. Fourier transform infrared difference spectroscopy of photosystem II tyrosine D using site-directed mutagenesis and specific isotope labeling. Biochemistry 36:14712-14723.

56. Ho, S. N., H. D. Hunt, R. M. Horton, J. K. Pullen, and L. R. Pease. 1989. Site-Directed Mutagenesis by Overlap Extension Using the Polymerase Chain-Reaction. Gene 77:51-59.

57. Hoff, W. D. 1995. Photoactive yellow protein. A new family of eubacterial blue-light photoreceptors. University of Amsterdam.

58. Hoff, W. D., P. Dux, K. Hard, B. Devreese, I. M. Nugterenroodzant, W. Crielaard, R. Boelens, R. Kaptein, J. Vanbeeumen, and K. J. Hellingwerf. 1994. Thiol Ester-Linked P-Coumaric Acid as a New Photoactive Prosthetic Group in a Protein with Rhodopsin-Like Photochemistry. Biochemistry 33:13959-13962.

59. Hoff, W. D., A. Xie, I. H. M. Van Stokkum, X. J. Tang, J. Gural, A. R. Kroon, and K. J. Hellingwerf. 1999. Global conformational changes upon receptor stimulation in photoactive yellow protein. Biochemistry 38:1009-1017.

60. Hohenberg, P., and W. Kohn. 1964. Inhomogeneous Electron Gas. Phys Rev B 136:B864-+.

61. Hu, X., H. Frei, and T. G. Spiro. 1996. Nanosecond step-scan FTIR spectroscopy of hemoglobin: ligand recombination and protein conformational changes. Biochemistry 35:13001-13005.

62. Hunter, T., and J. A. Cooper. 1985. Protein-tyrosine kinases. Annual review of biochemistry 54:897-930.

63. Imamoto, Y., T. Ito, M. Kataoka, and F. Tokunaga. 1995. Reconstitution photoactive yellow protein from apoprotein and p-coumaric acid derivatives. Febs Lett 374:157-160.

64. Imamoto, Y., H. Koshimizu, K. Mihara, O. Hisatomi, T. Mizukami, K. Tsujimoto, M. Kataoka, and F. Tokunaga. 2001. Roles of amino acid residues near the chromophore of photoactive yellow protein. Biochemistry 40:4679-4685.

65. Imamoto, Y., K. Mihara, O. Hisatomi, M. Kataoka, F. Tokunaga, N. Bojkova, and K. Yoshihara. 1997. Evidence for proton transfer from Glu-46 to the chromophore during the photocycle of photoactive yellow protein. The Journal of biological chemistry 272:12905-12908.

66. Ippolito, J. A., R. S. Alexander, and D. W. Christianson. 1990. Hydrogen bond stereochemistry in protein structure and function. Journal of molecular biology 215:457-471.

67. Jenson, D. L., and B. A. Barry. 2009. Proton-coupled electron transfer in photosystem II: proton inventory of a redox active tyrosine. J Am Chem Soc 131:10567-10573.

68. Jones, L. H., and E. Mclaren. 1954. Infrared Spectra of Ch3coona and Cd3coona and Assignments of Vibrational Frequencies. J Chem Phys 22:1796-1800.

69. Kaledhonkar, S. 2013. Strucutral Dynamics of Photoactive Yellow Protein. In Physics. Oklahoma State University.

70. Karthikeyan, K. S., A. Parameswaran, and B. P. Rajan. 1986. Mercury toxicity in dental personnel. Journal of the Indian Dental Association 58:215-220.

71. Keller, A., A. I. Nesvizhskii, E. Kolker, and R. Aebersold. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74:5383-5392.

72. Kim, H. W., J. A. Perez, S. J. Ferguson, and I. D. Campbell. 1990. The Specific Incorporation of Labeled Aromatic-Amino-Acids into Proteins through Growth of Bacteria in the Presence of Glyphosate - Application to Fluorotryptophan Labeling to the H+-Atpase of Escherichia-Coli and Nmr-Studies. Febs Lett 272:34-36.

73. Kleywegt, G. J., and T. A. Jones. 1997. Model building and refinement practice. In Macromolecular Crystallography, Pt B. C. W. Carter, and R. M. Sweet, editors. 208-230.

74. Kok, B., B. Forbush, and M. Mcgloin. 1970. Cooperation of Charges in Photosynthetic O2 Evolution .1. A Linear 4step Mechanism. Photochemistry and Photobiology 11:457-&.

75. Kotting, C., and K. Gerwert. 2005. Proteins in action monitored by time-resolved FTIR spectroscopy. Chemphyschem 6:881-888.

76. Krebs, M. P., and H. G. Khorana. 1993. Mechanism of light-dependent proton translocation by bacteriorhodopsin. J Bacteriol 175:1555-1560.

77. Kuhne, H., and G. W. Brudvig. 2002. Proton-coupled electron transfer involving Tyrosine Z in photosystem II. The journal of physical chemistry. B 106:8189-8196.

78. Kumauchi, M., N. Hamada, J. Sasaki, and F. Tokunaga. 2002. A role of methionine100 in facilitating PYPM-decay process in the photocycle of photoactive yellow protein. J Biochem 132:205-210.

79. Kumauchi, M., M. T. Hara, P. Stalcup, A. Xie, and W. D. Hoff. 2008. Identification of six new photoactive yellow proteins--diversity and structure-function relationships in a bacterial blue light photoreceptor. Photochem Photobiol 84:956-969.

80. Lanyi, J., and B. Schobert. 2002. Crystallographic structure of the retinal and the protein after deprotonation of the Schiff base: the switch in the bacteriorhodopsin photocycle. Journal of molecular biology 321:727-737.

81. Lanyi, J. K., and B. Schobert. 2003. Mechanism of proton transport in bacteriorhodopsin from crystallographic structures of the K, L, M1, M2, and M2' intermediates of the photocycle. Journal of molecular biology 328:439-450.

82. Lee, B., and F. M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. Journal of molecular biology 55:379-400.

83. Lee, C. T., W. T. Yang, and R. G. Parr. 1988. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density. Phys Rev B 37:785-789.

84. Lehninger, A. L., D. L. Nelson, and M. M. Cox. 2005. Lehninger Principles of Biochemistry. W. H. Freeman.

85. Li, T. S., T. Horan, T. Osslund, G. Stearns, and T. Arakawa. 1997. Conformational changes in G-CSF/receptor complex as investigated by isotope-edited FTIR spectroscopy. Biochemistry 36:8849-8857.

86. Liu, X.-M., S. Sonar, C.-P. Lee, M. Coleman, U. L. RajBhandary, and K. J. Rothschild. 1995. Site-directed isotope labeling and FTIR spectroscopy: assignment of tyrosine bands in the bR → M difference spectrum of bacteriorhodopsin. Biophys Chem 56:63-70.

87. Luecke, H., H. T. Richter, and J. K. Lanyi. 1998. Proton transfer pathways in bacteriorhodopsin at 2.3 Angstrom resolution. Science 280:1934-1937.

88. Luecke, H., B. Schobert, H. T. Richter, J. P. Cartailler, and J. K. Lanyi. 1999. Structure of bacteriorhodopsin at 1.55 A resolution. Journal of molecular biology 291:899-911.

89. Maeda, A., R. B. Gennis, S. P. Balashov, and T. G. Ebrey. 2005. Relocation of water molecules between the Schiff base and the Thr46-Asp96 region during light-driven unidirectional proton transport by bacteriorhodopsin: an FTIR study of the N intermediate. Biochemistry 44:5960-5968.

90. Mahalingam, M., K. Martinez-Mayorga, M. F. Brown, and R. Vogel. 2008. Two protonation switches control rhodopsin activation in membranes. Proc Natl Acad Sci USA 105:17795-17800.

91. McDonald, I. K., and J. M. Thornton. 1994. Satisfying hydrogen bonding potential in proteins. Journal of molecular biology 238:777-793.

92. McEvoy, J. P., and G. W. Brudvig. 2006. Water-splitting chemistry of photosystem II. Chemical reviews 106:4455-4483.

93. McRee, D. E., T. E. Meyer, M. A. Cusanovich, H. E. Parge, and E. D. Getzoff. 1986. Crystallographic characterization of a photoactive yellow protein with photochemistry similar to sensory rhodopsin. The Journal of biological chemistry 261:13850-13851.

94. Merz, K. M. 1991. Determination of Pkas of Ionizable Groups in Proteins - the Pka of Glu-7 and Glu-35 in Hen Egg-White Lysozyme and Glu-106 in Human Carbonic Anhydrase-Ii. J Am Chem Soc 113:3572-3575.

95. Meyer, T. E. 1985. Isolation and characterization of soluble cytochromes, ferredoxins and other chromophoric proteins from the halophilic phototrophic bacterium Ectothiorhodospira halophila. Biochim Biophys Acta 806:175-183.

96. Meyer, T. E., S. Devanathan, T. T. Woo, E. D. Getzoff, G. Tollin, and M. A. Cusanovich. 2003. Site-specific mutations provide new insights into the origin of pH effects and alternative spectral forms in the photoactive yellow protein from halorhodospira halophilia. Biochemistry 42:3319-3325.

97. Meyer, T. E., E. Yakali, M. A. Cusanovich, and G. Tollin. 1987. Properties of a Water-Soluble, Yellow Protein Isolated from a Halophilic Phototrophic Bacterium That Has Photochemical Activity Analogous to Sensory Rhodopsin. Biochemistry 26:418-423.

98. Muchmore, D. C., L. P. McIntosh, C. B. Russell, D. E. Anderson, and F. W. Dahlquist. 1989. Expression and nitrogen-15 labeling of proteins for proton and nitrogen-15 nuclear magnetic resonance. Methods Enzymol 177:44-73.

99. Muona, M., A. S. Aranko, V. Raulinaitis, and H. Iwai. 2010. Segmental isotopic labeling of multi-domain and fusion proteins by protein trans-splicing in vivo and in vitro. Nat Protoc 5:574-587.

100. Nie, B. 2006. Probing hydrogen bonding interactions and proton transfer in proteins. Oklahoma State University, Stillwater, OK. 243.

101. Nie, B., J. Stutzman, and A. Xie. 2005. A vibrational spectral marker for probing the hydrogen-bonding status of protonated Asp and Glu residues. Biophysical journal 88:2833-2847.

102. Oesterhe.D, and Stoecken.W. 1973. Functions of a New Photoreceptor Membrane. Proc Natl Acad Sci USA 70:2853-2857.

103. Oktaviani, N. A., T. J. Pool, H. Kamikubo, J. Slager, R. M. Scheek, M. Kataoka, and F. A. Mulder. 2012. Comprehensive determination of protein tyrosine pKa values for photoactive yellow protein using indirect 13C NMR spectroscopy. Biophysical journal 102:579-586.

104. Pace, C. N. 2001. Polar Group Burial Contributes More to Protein Stability than Nonpolar Group Burial†. Biochemistry 40:310-313.

105. Pace, C. N. 2009. Energetics of protein hydrogen bonds. Nat Struct Mol Biol 16:681-682.

106. Pace, C. N., H. Fu, K. Lee Fryar, J. Landua, S. R. Trevino, D. Schell, R. L. Thurlkill, S. Imura, J. M. Scholtz, K. Gajiwala, J. Sevcik, L. Urbanikova, J. K. Myers, K. Takano, E. J. Hebert, B. A. Shirley, and G. R. Grimsley. 2014. Contribution of hydrogen bonds to protein stability. Protein science : a publication of the Protein Society:n/a-n/a.

107. Pebay-Peyroula, E., G. Rummel, J. P. Rosenbusch, and E. M. Landau. 1997. X-ray structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipidic cubic phases. Science 277:1676-1681.

108. Philip, A. F., M. Kumauchi, and W. D. Hoff. 2010. Robustness and evolvability in the functional anatomy of a PER-ARNT-SIM (PAS) domain. Proc Natl Acad Sci USA 107:17986-17991.

109. Rahmelow, K., W. Hubner, and T. Ackermann. 1998. Infrared absorbances of protein side chains. Anal Biochem 257:1-11.

110. Rath, P., E. Spudich, D. D. Neal, J. L. Spudich, and K. J. Rothschild. 1996. Asp76 is the Schiff base counterion and proton acceptor in the proton-translocating form of sensory rhodopsin I. Biochemistry 35:6690-6696.

111. Rathod, R., Z. Kang, S. D. Hartson, M. Kumauchi, A. Xie, and W. D. Hoff. 2012. Side-chain specific isotopic labeling of proteins for infrared structural biology: the case of ring-D4-tyrosine isotope labeling of photoactive yellow protein. Protein expression and purification 85:125-132.

112. Rothschild, K. J., M. S. Braiman, Y. W. He, T. Marti, and H. G. Khorana. 1990. Vibrational spectroscopy of bacteriorhodopsin mutants. Evidence for the interaction of aspartic acid 212 with tyrosine 185 and possible role in the proton pump mechanism. The Journal of biological chemistry 265:16985-16991.

113. Rothschild, K. J., T. Marti, S. Sonar, Y. W. He, P. Rath, W. Fischer, and H. G. Khorana. 1993. Asp(96) Deprotonation and Transmembrane Alpha-Helical Structural-Changes in Bacteriorhodopsin. The Journal of biological chemistry 268:27046-27052.

114. Saff, E. B., and A. B. J. Kuijlaars. 1997. Distributing many points on a sphere. Math Intell 19:5-11.

115. Sambrook, J., E. F. Fritsch, and T. Maniatis. 1989. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, New York.

116. Sambrook, J., and M. J. Gething. 1989. Protein structure. Chaperones, paperones. Nature 342:224-225.

117. Schwans, J. P., F. Sunden, A. Gonzalez, Y. Tsai, and D. Herschlag. 2013. Uncovering the determinants of a highly perturbed tyrosine pKa in the active site of ketosteroid isomerase. Biochemistry 52:7840-7855.

118. Shankar, R. 2012. Principles of Quantum Mechanics. Springer London, Limited.

119. Siebert, F., and P. Hildebrandt. 2008a. Vibrational spectroscopy in life science. In Tutorials in biophysics. Wiley-VCH, Weinheim. x, 310 p.

120. Siebert, F., and P. Hildebrandt. 2008b. Instrumentation. In Vibrational Spectroscopy in Life Science. Wiley-VCH Verlag GmbH & Co. KGaA. 63-97.

121. Sonar, S., M. P. Krebs, H. G. Khorana, and K. J. Rothschild. 1993. Static and time-resolved absorption spectroscopy of the bacteriorhodopsin mutant Tyr-185 .fwdarw. Phe: Evidence for an equilibrium between bR570 and an O-like species. Biochemistry 32:2263-2271.

122. Sonar, S., C. P. Lee, M. Coleman, N. Patel, X. M. Liu, T. Marti, H. G. Khorana, U. L. Rajbhandary, and K. J. Rothschild. 1994. Site-Directed Isotope Labeling and Ftir Spectroscopy of Bacteriorhodopsin. Nat Struct Biol 1:512-517.

123. Song, Y., J. Mao, and M. R. Gunner. 2003. Calculation of proton transfers in Bacteriorhodopsin bR and M intermediates. Biochemistry 42:9875-9888.

124. Sprenger, W. W., W. D. Hoff, J. P. Armitage, and K. J. Hellingwerf. 1993. The Eubacterium Ectothiorhodospira-Halophila Is Negatively Phototactic, with a Wavelength Dependence That Fits the Absorption-Spectrum of the Photoactive Yellow Protein. J Bacteriol 175:3096-3104.

125. Sprogoe, D., L. A. van den Broek, O. Mirza, J. S. Kastrup, A. G. Voragen, M. Gajhede, and L. K. Skov. 2004. Crystal structure of sucrose phosphorylase from Bifidobacterium adolescentis. Biochemistry 43:1156-1162.

126. Styring, S., J. Sjoholm, and F. Mamedov. 2012. Two tyrosines that changed the world: Interfacing the oxidizing power of photochemistry to water splitting in photosystem II. Biochim Biophys Acta 1817:76-87.

127. Subramaniam, S., and R. Henderson. 2000. Molecular mechanism of vectorial proton translocation by bacteriorhodopsin. Nature 406:653-657.

128. Sun, H., and H. Frei. 1997. Time-resolved step-scan Fourier transform infrared spectroscopy of triplet excited duroquinone in a zeolite. The journal of physical chemistry. B 101:205-209.

129. Szaraz, S., D. Oesterhelt, and P. Ormos. 1994. pH-induced structural changes in bacteriorhodopsin studied by Fourier transform infrared spectroscopy. Biophysical journal 67:1706-1712.

130. Takahashi, R., and T. Noguchi. 2007. Criteria for determining the hydrogen-bond structures of a tyrosine side chain by fourier transform infrared spectroscopy: density functional theory analyses of model hydrogen-bonded complexes of p-cresol. The journal of physical chemistry. B 111:13833-13844.

131. Takano, K., J. M. Scholtz, J. C. Sacchettini, and C. N. Pace. 2003. The contribution of polar group burial to protein stability is strongly context-dependent. The Journal of biological chemistry 278:31790-31795.

132. Tang, X. S., D. A. Chisholm, G. C. Dismukes, G. W. Brudvig, and B. A. Diner. 1993. Spectroscopic Evidence from Site-Directed Mutants of Synechocystis Pcc6803 in Favor of a Close Interaction between Histidine-189 and Redox-Active Tyrosine-1608 Both of Polypeptide-D2 of the Photosystem-Ii Reaction-Center. Biochemistry 32:13742-13748.

133. Tatulian, S. A. 2010. Structural analysis of proteins by isotope-edited FTIR spectroscopy. Spectrosc-Int J 24:37-43.

134. Thubagere, A. J. 2008. Advanced FTIR spectroscopy of protein structural dynamics. Oklahoma State University, Stillwater, OK. 142.

135. Umena, Y., K. Kawakami, J. R. Shen, and N. Kamiya. 2011. Crystal structure of oxygen-evolving photosystem II at a resolution of 1.9 A. Nature 473:55-60.

136. van der Horst, M. A., T. P. Stalcup, S. Kaledhonkar, M. Kumauchi, M. Hara, A. Xie, K. J. Hellingwerf, and W. D. Hoff. 2009. Locked chromophore analogs reveal that photoactive yellow protein regulates biofilm formation in the deep sea bacterium Idiomarina loihiensis. J Am Chem Soc 131:17443-17451.

137. van Thor, J. J., G. Zanetti, K. L. Ronayne, and M. Towrie. 2005. Structural events in the photocycle of green fluorescent protein. The journal of physical chemistry. B 109:16099-16108.

138. Varrot, A., T. P. Frandsen, I. von Ossowski, V. Boyer, S. Cottaz, H. Driguez, M. Schulein, and G. J. Davies. 2003. Structural basis for ligand binding and processivity in cellobiohydrolase Cel6A from Humicola insolens. Structure 11:855-864.

139. Venyaminov, S., and N. N. Kalnin. 1990. Quantitative IR spectrophotometry of peptide compounds in water (H2O) solutions. I. Spectral parameters of amino acid residue absorption bands. Biopolymers 30:1243-1257.

140. Vogel, R., and F. Siebert. 2000. Vibrational spectroscopy as a tool for probing protein function. Current Opinion in Chemical Biology 4:518-523.

141. Warscheid, B., S. Brucker, A. Kallenbach, H. E. Meyer, K. Gerwert, and C. Kotting. 2008. Systematic approach to group-specific isotopic labeling of proteins for vibrational spectroscopy. Vib Spectrosc 48:28-36.

142. Worth, C. L., and T. L. Blundell. 2009. Satisfaction of hydrogen-bonding potential influences the conservation of polar sidechains. Proteins 75:413-429.

143. Wright, W. W., and J. M. Vanderkooi. 1997. Use of IR absorption of the carboxyl group of amino acids and their metabolites to determine pKs, to study proteins, and to monitor enzymatic activity. Biospectroscopy 3:457-467.

144. Xie, A., and W. D. Hoff. 2009. Advanced infrared spectroscopic techniques for structural biology and functional proteomics. BioOptics 32:18-21.

145. Xie, A., W. D. Hoff, A. R. Kroon, and K. J. Hellingwerf. 1996. Glu46 donates a proton to the 4-hydroxycinnamate anion chromophore during the photocycle of photoactive yellow protein. Biochemistry 35:14671-14678.

146. Xie, A., L. Kelemen, J. Hendriks, B. J. White, K. J. Hellingwerf, and W. D. Hoff. 2001. Formation of a new buried charge drives a large-amplitude protein quake in photoreceptor activation. Biochemistry 40:1510-1517.

147. Xu, R., B. Ayers, D. Cowburn, and T. W. Muir. 1999. Chemical ligation of folded recombinant proteins: segmental isotopic labeling of domains for NMR studies. Proc Natl Acad Sci USA 96:388-393.

148. Yabuki, T., T. Kigawa, N. Dohmae, K. Takio, T. Terada, Y. Ito, E. D. Laue, J. A. Cooper, M. Kainosho, and S. Yokoyama. 1998. Dual amino acid-selective and site-directed stable-isotope labeling of the human c-Ha-Ras protein by cell-free synthesis. Journal of biomolecular NMR 11:295-306.

149. Yamazaki, T., T. Otomo, N. Oda, Y. Kyogoku, K. Uegaki, N. Ito, Y. Ishino, and H. Nakamura. 1998. Segmental isotope labeling for protein NMR using peptide splicing. J Am Chem Soc 120:5591-5592.

150. Yan, H. G., and M. D. Tsai. 1991. Mechanism of Adenylate Kinase - Demonstration of a Functional-Relationship between Aspartate-93 and Mg2+ by Site-Directed Mutagenesis and Proton, P-31, and Magnesium-25 Nmr. Biochemistry 30:5539-5546.

151. Young, D. 2001. Computational chemistry: a practical guide for applying techniques to real world problems. John Wiley & Sons, Inc., New York.

152. Zhang, M., H. Fabian, H. H. Mantsch, and H. J. Vogel. 1994. Isotope-edited Fourier transform infrared spectroscopy studies of calmodulin's interaction with its target peptides. Biochemistry 33:10883-10888.

APPENDICES

Appendix A.1

Protocol G1 How to Generate Input Files

Zhouyang Kang, June 2013

- **Overview**

Gaussian is a program developed for the computation of electronic structures. It can predict the energies, molecular structures, vibrational frequencies and molecular properties of molecules and reactions in many chemical environments. The current version of Gaussian series is Gaussian 09 (released in 2009). For complex systems, Gaussian calculation may take hours to days or even weeks to finish, depends on the complexity of input structure, the method of calculation and the computational power. Although the algorithms of Gaussian have been optimized a lot comparing with its original version in terms of computational speed and accuracy, a well-planned strategy is highly valuable in order to complete a computational project in short time by optimizing the successful rate of the calculation and minimizing the computational time.

General procedure of a Gaussian calculation can be described as below:

- Prepare input files (including input structures and computational commands)
- Submit to Gaussian software for calculation, the procedure of computation should follow (defined by route sections of input files):
  1. Pre geometry optimization using classical mechanics method (PM3)
  2. Geometry optimization using quantum mechanics method (B3LYP, etc.) and the energy calculation
  3. Other functioning calculations (Vibrational frequencies calculation, energy calculation in other chemical environment, etc.)
- Monitor computational process, if needed, stop the computation and make adjustment.
- Analyze final output files
- Organize files, including log document, input, output and checkpoint files.

For understanding the principles of Density Functional Theory and the details of Gaussian software, the following references are useful:

- *Exploring Chemistry with Electronic Structure Methods. 2nd Edition*. James B. Foresman and AEleen Frisch
- *A Chemist's Guide to Density Functional Theory.* Wolfram Koch, Max C. Holthausen
- *Gaussian 09 User's Reference (Online):* http://www.gaussian.com/g_tech/g_ur/g09help.htm

The example input files used in this protocol can be found at

Dropbox\Xie Group\Xie Lab Protocols\Gaussian Calculations\Input Examples

- **Understanding the structure of input files**

Input files for Gaussian calculation consists of several sections (see example 1):

- ***Link 0 Commands*** (line start with %): Locate and name checkpoint files, require memory and CPU. (not terminated by blank line)
- ***Route sections*** (line start with #): Specify desired calculation, including method, basis set, model chemistry, and other functions. (terminated by a blank line)
- ***Title sections***: Brief describe the job, researcher name, date, computer used and the software used. This section will appear in output file for reference purpose and will not be read by Gaussian for calculation.
- ***Molecule specification***: Specify charge, spin multiplicity, geometry coordinates and other molecular information such as hydrogen bond interactions or fixed geometry constants of the input structure. (terminated by a blank line)
- ***Optional additional sections***: Additional input needed for specific job types such as further geometry optimization with more accurate method, isotopic labeling, energy calculation in different solvent, etc.

Example 1: Input file for structure optimization of a two water molecule system in vacuum

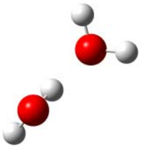
```
%chk=2H2O_Cartesian                                   /# Link 0 Commands name checkpoint file#/
%mem=4000MB                                           /# Link 0 Commands require memory#/
#PM3 Scf=Direct Opt=ModRedundant Test                 /# Route section pre geometry optimization#/
                                                      /# Blank line terminate route section#/
2 H2O forms a hydrogen bond in Cartesian Coordinates, Zhouyang, 06/24/13, HPCC, G09   /# Title section#/
                                                      /# Blank line terminate title section#/
0   1                                             /# charge, spin multiplicity (M=2S+1) #/
 O          0.00000000   0.00000000   0.00000000                                 ┐
 O          2.75612569   0.03585373  -0.72455897 ┐                               │
 H          0.00000000   0.00000000   0.96096300 │  /# geometry in cartesian     │
 H          0.93525958   0.00000000  -0.22143073 ├      coordinates#/            ├/#molecule specification#/
 H          2.92276993   0.79180018  -1.29187942 │                               │
 H          2.89741618  -0.72201730  -1.29692645 ┘                               │
1   2                                             /# atom 1 and 2 formhydrogen bond#/ ┘

--Link1--                                   /#additional step for further geometry optimization calculation#/
%chk=2H2O_Cartesian                                   /# Link 0 Commands locate checkpoint file#/
%Mem=4000MB                                           /# Link 0 Commands require memory#/
#B3LYP/6-311+G(3df,2p) Scf=Direct Geom=Allcheck Opt=ModRedundant Test  /# Route section geometry
                                                                          optimization#/
/# Additional steps can be added following this input if needed.#/
```

A. Geometry Optimization

Geometry optimization is used in nearly all Gaussian calculations. The input structure can be in format of Cartesian (x, y, z) or Internal Coordinates (Z-matrix) (bond length, bond angle, and dihedral angle) as illustrated below.

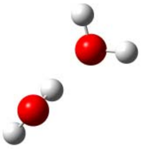
Cartesian Coordinates

O	0.00000000	0.00000000	0.00000000
O	2.75612569	0.03585373	-0.72455897
H	0.00000000	0.00000000	0.96096300
H	0.93525958	0.00000000	-0.22143073
H	2.92276993	0.79180018	-1.29187942
H	2.89741618	-0.72201730	-1.29692645



Internal Coordinates (Z-matrix): (with O..O distance fixed at 2.85 Å)

O							
O	1	B1					
H	1	B2	2	A1			
H	1	B3	2	A2	3	D1	0
H	2	B4	1	A3	3	D2	0
H	2	B5	1	A4	3	D3	0



Variables:

~~B1~~	~~2.85000000~~
B2	0.96096300
B3	0.96111500
B4	0.95972800
B5	0.96017500
A1	104.72803596
A2	1.58284334
A3	109.15406664
A4	106.49450157
D1	-27.27375494
D2	-123.76059619
D3	124.40054779

/# Fixed constant is cut from coordinates and pasted at the bottom#/

Constants:

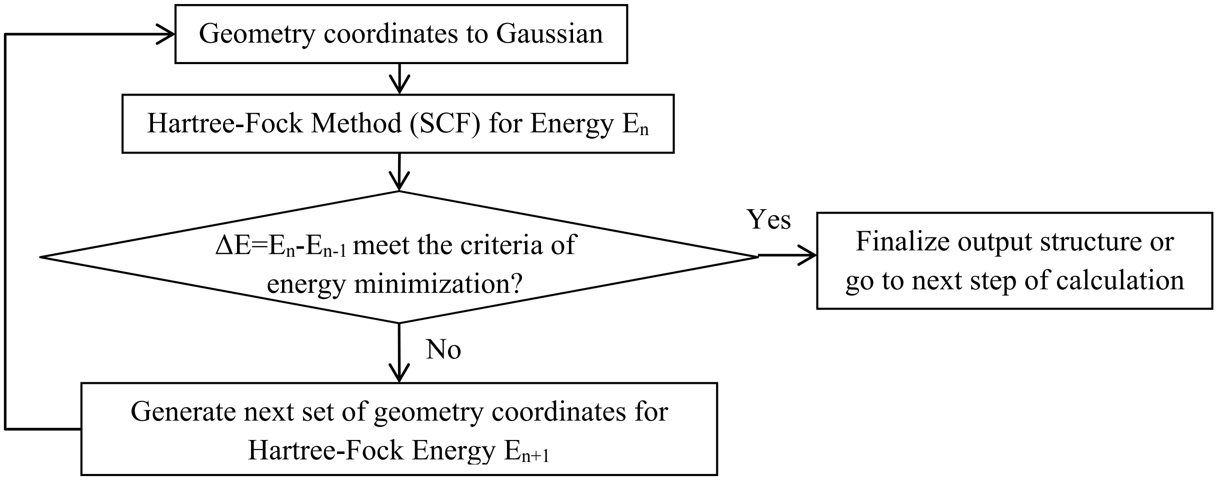
B1	2.85000000

/# Fixed geometry constant#/

- *Q: Which coordinates should be used?*

  A: **Cartesian coordinates is strongly recommended for most of Gaussian calculations. For geometry contain fixed parameters, internal coordinates should be used.** This is because with internal coordinates, each interaction needs to be specified. Particularly at the hydrogen bond site, the angle which formed by two heavy atoms and one hydrogen atom might fall to a negative value during the geometry optimization, which will terminate the calculation with error.

The process for geometry optimization can be summarized by the flow chart below:


Geometry coordinates to Gaussian
Hartree-Fock Method (SCF) for Energy E_n
$\Delta E=E_n-E_{n-1}$ meet the criteria of energy minimization?
Yes
Finalize output structure or go to next step of calculation
No
Generate next set of geometry coordinates for Hartree-Fock Energy E_{n+1}

Among all types of Gaussian calculations, geometry optimization is one of the most time-consuming steps. Thus, we normally applied a two-step procedure:

Step 1: Pre geometry optimization with classical mechanics method (PM3)

Step 2: Geometry optimization with DFT method (B3LYP, etc.)

Example ***route section*** for Step 1:

(For Cartesian coordinates)

```
#PM3 Scf=Direct Opt=Modredundant Test
```

Or (For Internal coordinates)

```
#PM3 Scf=Direct Opt=Z-matrix Test
```

Example ***route section*** for Step 2:

(For Cartesian coordinates)

```
#B3LYP/6-311+G(3df,2p) Scf=Direct Geom=Allcheck Opt=ModRedundant Test
```

Or (For Internal coordinates)

```
#B3LYP/6-311+G(3df,2p) Scf=Direct Geom=Allcheck Opt=Z-matrix Test
```

- **Useful Keywords and Options in Route Sections**

  - **Scf**

Direct Requests a direct SCF calculation, in which the two-electron integrals are recomputed as need. Default SCF procedure in Gaussian.

QC Calls for the use of quadratically convergent SCF procedure. Use linear searches when far from convergence and Newton-Raphson steps when close (unless the energy goes up). Slower method than regular SCF but more reliable. Available for unrestricted open shell calculations only.

XQC Add an extra *Scf=QC* step in case the first-order SCF has not converged.

Conver=N Sets the SCF convergence criterion to 10^{-N}.

MaxCycle=N Changes the maximum number of SCF cycles permitted to N. Default is 64 (or 512 for *Scf=QC*).

  - **Opt**

ModRedundant Add, delete or modify redundant internal coordinate definitions (including scan and constraint information) before performing the calculation. This option requires a separate input section following the geometry specification (i.e. atoms that form hydrogen bonds)

Z-matrix Perform the optimization with the Berny algorithm using internal coordinates. It will request a POpt (partial optimization), which the geometry constants defined will be held fixed and the other variables will be optimized. And if no variable is fixed, then a FOpt (full optimization) will be applied on all variables.

Loose Sets the optimization convergence criteria to a maximum step size of 0.01 au and an RMS force of 0.0017 au.

Loose Convergence Criteria	Normal Convergence Criteria
Max Force: 0.002500	Max Force: 0.000450
RMS Force: 0.001667	RMS Force: 0.000300
Max Displacement: 0.010000	Max Displacement: 0.001800
RMS Displacement: 0.006667	RMS Displacement: 0.001200

ReadFC Extract force constants form a checkpoint file. For difficult optimization jobs, it is recommended to use lower level method to optimize geometry and calculate frequency, then optimize geometry with higher level method and use *ReadFC* option to read force constants from the previous frequency calculation.

CalcFC For difficult calculations, calculate force constants before geometry optimization.

MaxCycles=N Sets the maximum number of optimization steps to N. The default is the maximum of 20 and twice the number of redundant internal coordinates in use (for the default procedure) or twice the number of variables to be optimized (for other procedures). **There is no need to set any N number larger than the maximum cycle number allowed by Gaussian.**

MaxStep=N Sets the maximum size for an optimization step (the initial trust radius) to 0.01N Bohr or radius. The default value for N is 30. This option is useful for situations that energies are already close to local minimum but optimizations still have difficulties to converge.

Restart Restarts a geometry optimization from the checkpoint file. In this case, the entire route section will consist of the *Opt* keyword and the same options to it as specified for the original job (along with *Restart*). No other input is needed.

- **Geom**

Checkpoint Reads molecule specification (including variables) from checkpoint file, but reads charge and multiplicity from input stream.

Allcheck Reads whole molecule specification (including variables, charge and multiplicity) and title section from checkpoint file.

Step=N Retrieves the structure produced by the N^th^ step of a failed or partial geometry optimization. However, it will not read step from successful geometry optimization.

Table G1-1: Input files for pre geometry optimization (PM3) then DFT optimization (6-311+G(3df,2p))

Molecule	Charge	Spin Multiplicity	Input file Name	Comments
H2O	0	1	G1-1A	
Butyric acid	-1	1	G1-1B	

Table G1-2: Input files for pre geometry optimization (PM3) then DFT optimization (6-311+G(3df,2p))

Molecule	Charge	Spin Multiplicity	Input file Name	Comments
2H2O	0	1	G1-2A	1 Hydrogen Bond
3H2O	0	1	G1-2B	3 Hydrogen Bonds
BAC-methanol	-1	1	G1-2C	1 Hydrogen Bond

Table G1-3 Input files for pre geometry optimization (PM3) then DFT optimization (6-311+G(3df,2p))

Molecule	Charge	Spin Multiplicity	Input file Name	Comments
BAC-methanol	-1	1	G1-3A	Fixed HB Distance
BAC-methanol	-1	1	G1-3B	Fixed HB Angle
BAC-methanol	-1	1	G1-3C	Fixed Dihedral Angle

B. A Special Case I: Gaussian calculation of Transition Metals

See protocol G2: Protocol for Gaussian Calculation on Transition Metal Compounds (Gaussian09)

Table G1-4 Input files for Non zero spin calculation

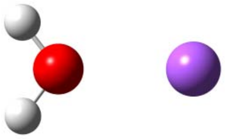
Molecule	Charge	Spin Multiplicity	Input file Name	Comments
Mn_2O_2(2+)	+2	1	G1-4A	Basis set TBD
		3	G1-4B	
		5	G1-4C	
		7	G1-4D	
		9	G1-4E	

C. A Special Case II: Gaussian calculation using different basis sets on different atoms

The basis sets we normally use for atoms in the first three rows of periodic table may not be suitable for atoms with more orbitals, particularly heavy metal atoms. Thus, it is necessary to use different basis sets on those atoms even though they are in one input structure. (See Example 2)

Example 2: How to use different basis sets (one for heavy atoms, one for normal)

```
%chk=Na_1H2O
%mem=4000MB
#B3LYP/GenECP Scf=Direct Opt=Z-matrix Freq Test   /# Route section, GenECP is used#/

Na cation forms 1 HB with H2O, Different basis sets applied, Zhouyang, 06/24/13, HPCC, G09

1 1
 Na
 O          1       B1
 H          2       B2   1       A1
 H          2       B3   1       A2   3       D1   0

   B1         2.49634400
   B2         0.96000000
   B3         0.96000000
   A1       112.35610528
   A2       120.70000000
   D1      -131.60270131

H O 0
6-31G(d)                                          /# 6-31G(d) on all hydrogen and oxygen#/
****                                              /# Separator: terminates the center definition block#/
Na 0
LANL2DZ                                          /# LANL2DZ on all sodium#/
****                                             /# Separator: terminates the center definition block#/
```



- **GenECP**

GenECP is equivalent to *Gen Pseudo=Read*, in which *Gen* allows a user-specified basis set to be used in a Gaussian calculation, and *Pseudo=Read* requests that a model potential be substituted for the core electrons.

Table G1-5 Input files for heavy atoms using different basis sets

Molecule	Charge	Spin Multiplicity	Input file Name	Comments
Na-H2O	+1	1	G1-5A	

D. Energy Calculation in different solvent

The ideal way of calculating the minimum energy of a model compound in solvent is optimizing its geometry with solvent effect. However, such function is not optimal yet in current version of Gaussian. An alternative way is to optimize the geometry of input structure in vacuum, then perform single point energy calculation for that stationary point in different solvent.

Example ***route section*** for single point energy calculation in solvent after geometry is optimized:

```
#B3LYP/6-311+G(3df,2p) Geom=AllCheck SCRF=(PCM, Solvent=Benzene) Test
```

- **Useful Keywords and Options in Route Sections**

  - **SCRF**

This keyword requests that a calculation be performed in the presence of a solvent by placing the solute in a cavity within the solvent reaction field

PCM The Polarizable Continuum Model (PCM) performs a reaction field calculation using the integral equation formalism model (IEFPCM).

Solvent=item Selects the solvent in which the calculation is to be performed. If unspecified, the solvent defaults to water. Item is a solvent name chosen from the list of solvent that Gaussian recognizes.

Frequently used solvents and dielectric constants for calculation in Xie lab

Benzene: ε=2.2706
DiethylEther: ε=4.2400
ChloroBenzene: ε=5.6968
Water: ε=78.3553

A full list of recognized solvents and dielectric constants can be found in document “List of Defined Solvents.docx”

Dielectric=val Sets the value for the dielectric constant of the solvent. This option overrides *Solvent* if both are specified.

Table G1-6 Input files for electronic energy calculation in solvents (including geometry optimization as first step)

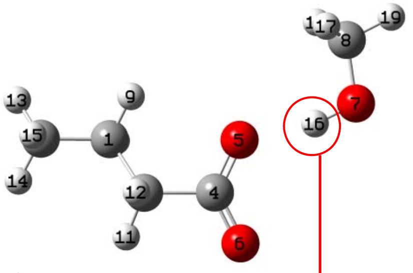
Molecule	Charge	Spin Multiplicity	Input file Name	Comments
BAC-methanol	-1	1	G1-6A	In Vacuum and Benzene

E. Vibrational Frequency Calculations (with Isotopic Labeling)

Vibrational frequencies are computed by determining the second derivatives of the energy with respect to the Cartesian nuclear coordinates and then transforming to mass-weighed coordinates. *This transformation is only valid at a stationary point!* Thus, frequency calculation can only be performed after the geometry is optimized with the local minimum energy is reached.

Example 3: Vibrational frequency calculations

```
%chk=BAC-_Methanol_D2O
%mem=1000MB
#PM3 Scf=Direct Opt=ModRedundant Test

Deprotonated BAC forms 1 hydrogen bond with methanol molecule, Zhouyang, 06/24/13, HPCC, G09

0 1
 C              0.00000000    0.00000000    0.00000000
Continued geometry Cartesian coordinates...

6   7

--Link1--                      /# additional step for further geometry optimization and frequency calculation#/
%chk=BAC-_Methanol_D2O
%Mem=4000MB
#B3LYP/6-311+G(3df,2p) Scf=Direct Geom=Allcheck Opt=ModRedundant Freq Test

--Link1--                      /# additional step for frequency calculation with isotopic labeling#/
%chk=BAC-_Methanol_D2O
%Mem=4000MB
#B3LYP/6-311+G(3df,2p) Scf=Direct Geom=AllCheck Freq(ReadFC, ReadIsotopes) Test

300.0  1.0                    /# Specify Temperature (K) and Pressure (atmospheres)#/
12
12
12
12
16
16
16
12
1
1                             /# Specify isotope mass of each atom#/
1
1
1
1
1
2
1
1
1
```



- **Useful Keywords and Options in Route Sections**

  - **Freq**

Raman Computes Raman intensities in addition to IR intensities.

ReadIsotopes Allows specifying alternatives to the default temperature, pressure, frequency scale facor and/or isotopes – 298.15K, 1 atmosphere, no scaling, and the most abundant isotopes.

Table G1-7 Input files for frequency calculations.

Molecule	Charge	Spin Multiplicity	Input file Name	Comments
BAC-methanol	-1	1	G1-7A	Basic Freq
BAC-methanol	-1	1	G1-7B	Fixed geometry
BAC-methanol	-1	1	G1-7C	Isotopic Labeled

Gaussian 03/09 Command List

1. Pre Optimization using PM3

   (i) For Cartesian Coordinates

   ```
   #PM3 Scf=Direct Opt=ModRedundant Test
   ```

   (i) For Internal Coordinates

   ```
   #PM3 Scf=Direct Opt=Z-matrix Test
   ```

2. Additional step for DFT geometry optimization and energy calculation in vacuum

   (i) Without fixed geometry constants

   ```
   #B3LYP/6-31G(d) Geom=Allcheck Scf=Direct Opt=ModRedundant Test
   #B3LYP/6-311+G(2d,p) Geom=Allcheck Scf=Direct Opt=ModRedundant Test
   #B3LYP/6-311+G(3df,2p) Geom=Allcheck Scf=Direct Opt=ModRedundant Test
   ```

   (ii) With fixed geometry constants

   ```
   #B3LYP/6-31G(d) Geom=Allcheck Scf=Direct Opt=Z-Matrix Test
   #B3LYP/6-311+G(2d,p) Geom=Allcheck Scf=Direct Opt=Z-Matrix Test
   #B3LYP/6-311+G(3df,2p) Geom=Allcheck Scf=Direct Opt=Z-Matrix  Test
   ```

   (iii) Use loose convergence (usually applied with fixed geometry constants)

   ```
   #B3LYP/6-31G(d) Geom=Allcheck Scf=Direct Opt=(loose, Z-Matrix) Test
   #B3LYP/6-311+G(2d,p) Geom=Allcheck Scf=Direct Opt=(loose, Z-Matrix) Test
   #B3LYP/6-311+G(3df,2p) Geom=Allcheck Scf=Direct Opt=(loose, Z-Matrix) Test
   ```

   Note: For difficult calculations, Opt=CalcFC option may be used

3. Additional step for more accurate energy calculation after structure is optimized

   ```
   #B3LYP/6-311+G(2d,p) Geom=Allcheck Scf=Direct Test
   #B3LYP/6-311+G(3df,2p) Geom=Allcheck Scf=Direct Test
   ```

4. (i) Additional step for energy calculation in solvent other than vacuum (i.e. DMSO)

   ```
   #B3LYP/6-311+G(2d,p) Geom=AllCheck SCRF=(PCM,Solvent=DMSO) Test
   #B3LYP/6-311+G(3df,2p) Geom=AllCheck SCRF=(PCM,Solvent=DMSO) Test
   ```

   (ii) Additional step for energy calculation in solvent other than vacuum, and place and individual sphere on particular hydrogen atom (i.e. H no. 15, and solvent DMSO)

   ```
   #B3LYP/6-311+G(2d,p) Geom=AllCheck SCRF=(PCM,Solvent=DMSO)
   SPHEREONH=15 Test
   #B3LYP/6-311+G(3df,2p) Geom=AllCheck SCRF=(PCM,Solvent=DMSO)
   SPHEREONH=15 Test
   ```

5. Additional step for frequency calculation

   (i) Without isotopic Labeling

   ```
   #B3LYP/6-311+G(2d,p) Geom=AllCheck Scf=Direct Freq Test
   #B3LYP/6-311+G(3df,2p) Geom=AllCheck Scf=Direct Freq Test
   ```

   (ii)With isotopic labeling

   ```
   #B3LYP/6-311+G(2d,p) Geom=AllCheck Scf=Direct Freq(ReadFC,ReadIsotopes) Test
   #B3LYP/6-311+G(3df,2p) Geom=AllCheck Scf=Direct Freq(ReadFC,ReadIsotopes) Test
   ```

   Note: Frequency calculation must use same method as geometry optimization!

Appendix A.2

Protocol G2 Gaussian 09 Calculation on Transition Metal Compounds

Zhouyang Kang, Shuo Dai, Ningning Xu, Aihua Xie

06/21/2013

Transition metal elements such as Fe ($3d^6 4s^2$) and Mn ($4s^2 3d^5$) carry unpaired electrons. For the same charge state, the metal elements could have different spins. In order to find the geometry of the compound, one should test and compare the energy at different spin states.

Electron structure of Mn, Mn^{2+}, Mn^{3+} and Mn^{4+}, Mn^{7+}

Mn:	**1s2 2s2 2p6 3s2 3p6 3d5 4s2**
Mn^{2+}:	**1s2 2s2 2p6 3s2 3p6 3d5**
Mn^{3+}:	**1s2 2s2 2p6 3s2 3p6 3d4**
Mn^{4+}:	**1s2 2s2 2p6 3s2 3p6 3d3**
Mn^{7+}:	**1s2 2s2 2p6 3s2 3p6**

Take $(Mn_2O_2)^{2+}$ as an example, each Mn carries charge 3+. Since Mn^{3+} have 4 unpaired electrons on the 3d orbital (see appendix), then the total spin of $(Mn^2O^2)^{2+}$ could be 4, 3, 2, 1, 0. The spin multiplicity is calculated by 2S+1, which means 2*4+1=9 for total spin 4 case. And pin multiplicity 1, 3, 5, 7 are also possible for the system.

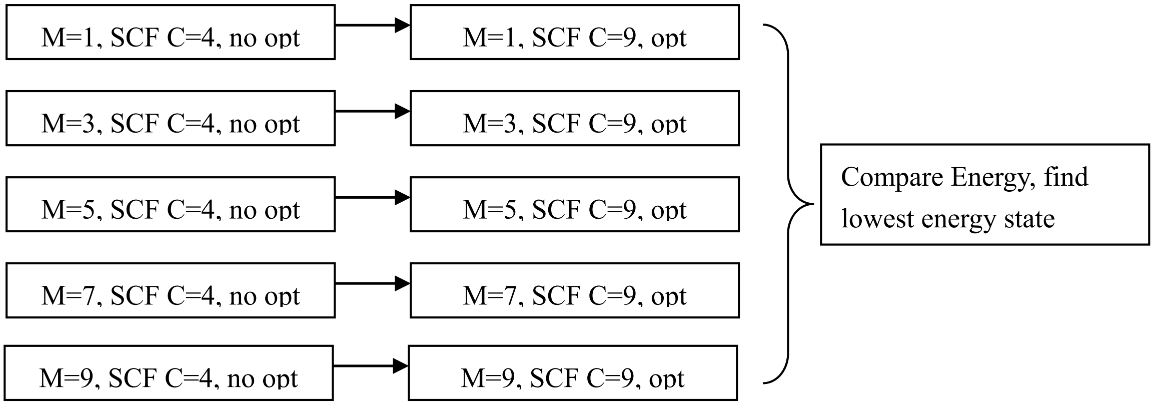
The process of calculation is illustrated as below:

Step1: No geometry optimization, calculate a solution for SCF with low convergence criteria.

(C=4, ΔE=10^{-4}au)

Step2: Geometry optimization, calculate SCF solution with high convergence criteria. (C=9, ΔE=10^{-9}au)

Unrestricted open-shell calculation method is used because of high spin multiplicity.



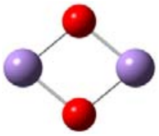
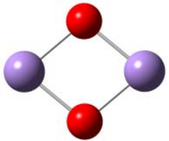
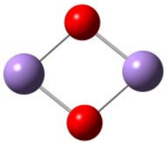
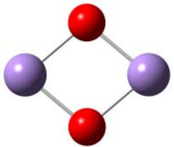
M: Spin Multiplicity

C: Convergence criteria for SCF calculation

Table G2-1 Energy comparison after optimized geometry of $(Mn_2O_2)^{2+}$ using different basis sets

Basis Set	1	3	5	7	9
6-31G(d)	-2451.265	-2451.429	-2451.407	-2451.456	-2451.544
6-311+G(d)	-2451.512	-2451.641	-2451.605	-2451.662	-2451.752
6-311+G(3df)	-2451.530		-2451.606	-2451.683	-2451.770
6-31G(d)	-2451.311	-2451.394	-2451.461	-2451.456	-2451.544

Table G2-2 Geometry optimization of $(Mn_2O_2)^{2+}$ using different basis sets

Basis set	Spin Multiplicity	Ground Energy	Output	Distance Mn..Mn	Angle O..Mn..O
6-31G(d)	9	-2451.544		2.74	80.0
6-311+G(d)	9	-2451.752		2.78	79.3
6-311+G(3df)	9	-2451.770		2.75	80.2
6-31G(d)/G03	9	-2451.544		2.73	80.0

In the example input files, 6-31G* was selected as the basis set. This is just for the purpose of testing the sequence of calculation with relatively less functions. However, in reality this basis set is not suitable for calculating Mn and Fe. The basis set which should be used to treat the transition metal elements is still to be determined.

Keywords and Options

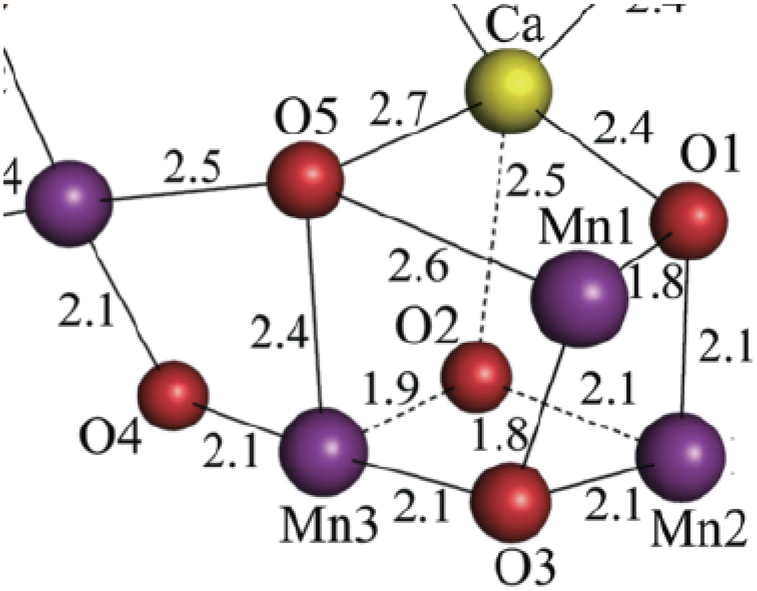
- **Guess**

Read Requests that the initial guess be read from the checkpoint file.

Mix Requests that the HOMO and LUMO be mixed so as to destroy α-β and spatial symmetries. This is useful in producing UHF wave functions for singlet states.

Keywords SCF and OPT see Protocol G1

Structure of Oxygen Evolving Center from Photosystem II (3ARC.PDB)



Appendix A.3

Protocol G3 Running Gaussian 09 on Cowboy Cluster of HPCC-OSU

Zhouyang Kang

Department of Physics, Oklahoma State University

1. Introduction to Cowboy Cluster

Cowboy is the newest supercomputing cluster at Oklahoma State University High Performance Computing Center (OSU-HPCC), which was assembled by Advanced Clustering Technologies, Inc. It contains 252 nodes, within each there are two Intel Xeon E5-2620 "Sandy Bridge" hex core 2.0G CPUs, which means every node comes with 12 CPUs. The capability of parallel computing, which means using multiple CPUs at the same time allows jobs to be computed at a speed that normal desktop cannot compete with. For example, a Gaussian09 job takes 12 hours for a single-CPU desktop will only take around 1 hour for Cowboy Cluster to give the same result.

2. Apply an account on HPC Center

Send an E-mail to hpcc@okstate.edu includes the following information:

**First Name:*

**Last Name:*

**Name you prefer to be called:*

**Desired login name (lowercase letters or numbers - max 8 characters):*

**a group/project name (same rules as for login name):*

**Status (faculty member, graduate or undergraduate student, staff member):*

**Advisor or name of Oklahoma faculty or staff member whose project you're on:*

(Please cc your advisor when you send in your account request)

**Department:*

**Institution and campus location (e.g. OSU-Stillwater):*

**Email address:*

**Telephone number:*

**Number to send password via text message:*

**Short description of your project:*

**Do you agree to abide by the user policies User_Policies?*

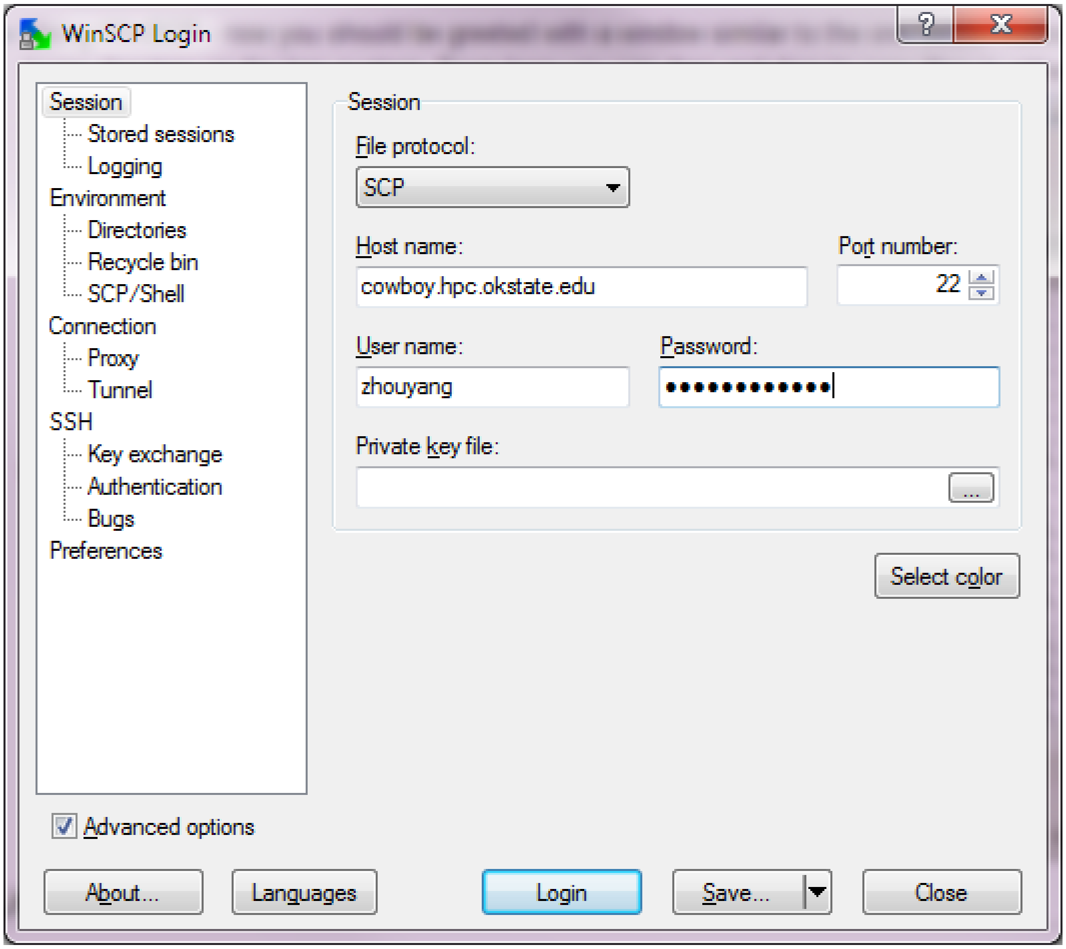
A response from hpcc regarding the request is expected within two business days.

Also include what software you need to use in the email (in case of Gaussian09), you will be asked to confirm your agreement on the Legal Terms provided by Gaussian Inc.

3. Submit Gaussian jobs on HPC

For Windows system, WinSCP is the recommended software package to manage files and submitting jobs on HPC. The installation file can be downloaded from http://winscp.net/eng/download.php

After installation, run the WinSCP software and a window will be opened as below:



WinSCP Login

Session

File protocol: SCP

Host name: cowboy.hpc.okstate.edu

Port number: 22

User name: zhouyang

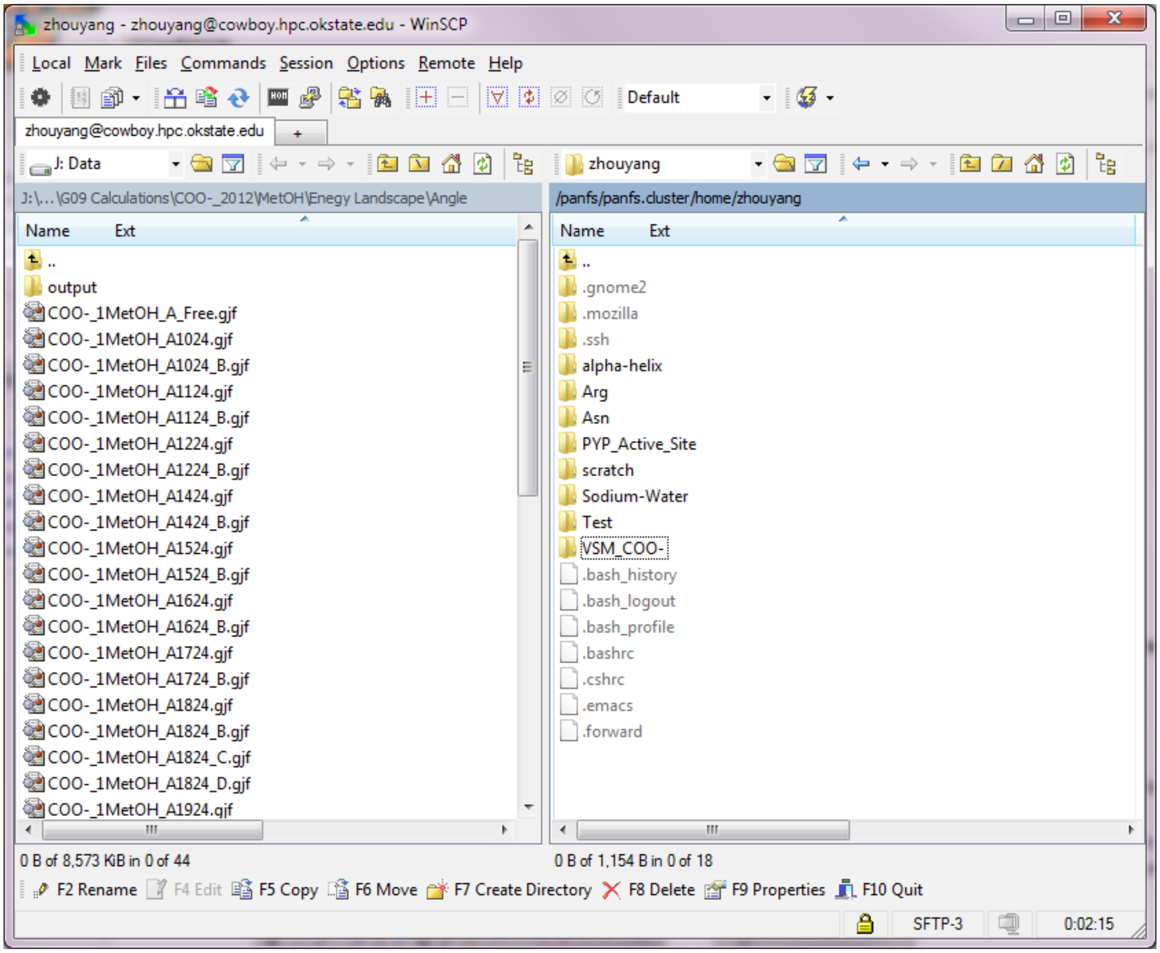
Password: ●●●●●●●●●●●●

Private key file:

☑ Advanced options

Choose **SCP** as the File protocol, and **cowboy.hpc.okstate.edu** as the port number, fill the username and password then click "Login".

After login, a new window will be opened as below, notice on the left is the local disk and on the right is the home directory on HPC. In order to copy files from either direction, simply drag and drop.


zhouyang - zhouyang@cowboy.hpc.okstate.edu - WinSCP
Local Mark Files Commands Session Options Remote Help
Default
zhouyang@cowboy.hpc.okstate.edu
J: Data
zhouyang
Name Ext
F2 Rename
F4 Edit
F5 Copy
F6 Move
F7 Create Directory
F8 Delete
F9 Properties
F10 Quit
SFTP-3
0:02:15

```
J:\...\G09 Calculations\COO-_2012\MetOH\Enegy Landscape\Angle
..
output
COO-_1MetOH_A_Free.gjf
COO-_1MetOH_A1024.gjf
COO-_1MetOH_A1024_B.gjf
COO-_1MetOH_A1124.gjf
COO-_1MetOH_A1124_B.gjf
COO-_1MetOH_A1224.gjf
COO-_1MetOH_A1224_B.gjf
COO-_1MetOH_A1424.gjf
COO-_1MetOH_A1424_B.gjf
COO-_1MetOH_A1524.gjf
COO-_1MetOH_A1524_B.gjf
COO-_1MetOH_A1624.gjf
COO-_1MetOH_A1624_B.gjf
COO-_1MetOH_A1724.gjf
COO-_1MetOH_A1724_B.gjf
COO-_1MetOH_A1824.gjf
COO-_1MetOH_A1824_B.gjf
COO-_1MetOH_A1824_C.gjf
COO-_1MetOH_A1824_D.gjf
COO-_1MetOH_A1924.gjf
0 B of 8,573 KiB in 0 of 44
```

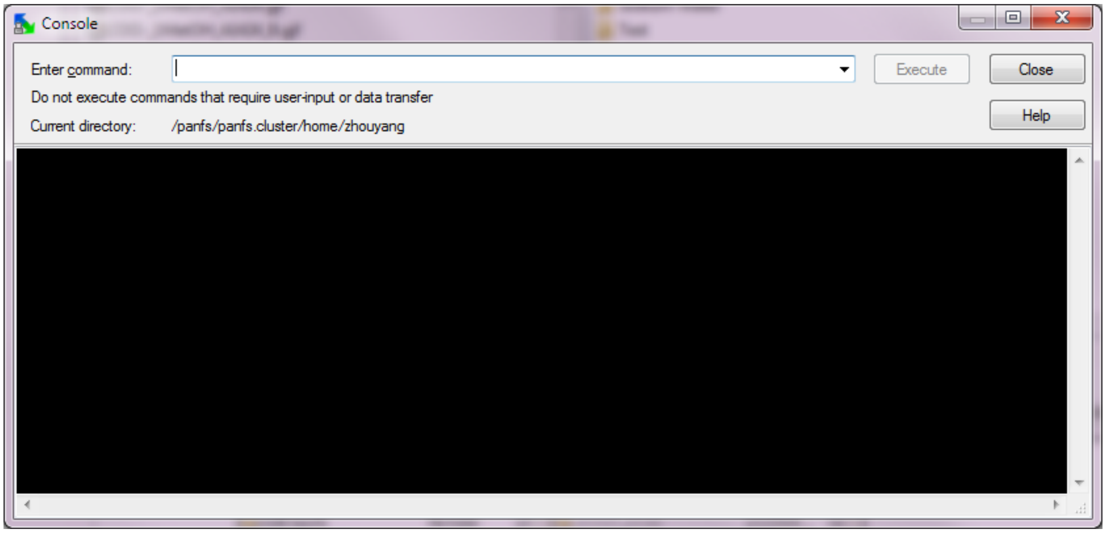
```
/panfs/panfs.cluster/home/zhouyang
..
.gnome2
.mozilla
.ssh
alpha-helix
Arg
Asn
PYP_Active_Site
scratch
Sodium-Water
Test
VSM_COO-
.bash_history
.bash_logout
.bash_profile
.bashrc
.cshrc
.emacs
.forward
0 B of 1,154 B in 0 of 18
```

To run Gaussian 09, a job script and an input file are required. To create the job script, simply right click in the targetted folder, then select new->file, you will be asked to type the name, “job1.pbs” can be used here for example, and “.pbs” is the extension for the job script files. In the new file, copy and paste the following paragraph, except the comment sentences. (labeled with /#...#/)

```
#!/bin/bash

#PBS -q batch
#PBS -l nodes=1:ppn=12 /#determine the number of nodes and the number of CPUs in the node will be used, a Gaussian09 job is able to use 12 CPUs for the parallel computing, which will increase effieciency significantly#/
#PBS -l walltime=120:00:00 /#determine the max running time, 120:00:00 means 120 hours and it is the wall time limit on HPC#/
#PBS -j oe

mkdir -p /home/zhouyang/scratch/g09/$PBS_JOBID
export GAUSS_SCRDIR=/scratch/zhouyang/g09/$PBS_JOBID
module load gaussian/g09
cd $PBS_O_WORKDIR

g09 <COO-_only_B_04122013.gjf>COO-_only_B_04122013.out /#name the input file name and the desired output file name#/
date
```

The input file of Gaussian09 running on HPC will have the same format as it for the windows version of Gaussian03 or Gaussian09. File extension ".gjf" is accepted by Gaussian 09 on HPC. An input file can be created by using GaussView in windows and edited in Notepad, then be dragged and copied to the home directory on HPC.

Click **Commands** button on the top bar and choose "**Open Terminal**". A new window named "console" will be popped out:



Enter command:
Do not execute commands that require user-input or data transfer
Current directory: /panfs/panfs.cluster/home/zhouyang

To submit the job script, type “qsub job1.pbs” then enter. “job1.pbs” refers to the job script which was just created in the protocol. After the job is submitted, a job ID number will be given.

To check all job status, type “qstat” then enter, the jobs are currently running on HPC will be shown. To kill a particular job, type “qdel JOBID” using the job ID number and certain job will be ceased.

To view the real-time output, go to the folder then double click the output file, it will be opened in text. To view the structure, right click on the file selected, choose open, then the it will be opend by GaussView.

After the calculation is finished, download the output file and the checkpoint file to the local computer.

4. Technical Information of OSU-HPCC (Courtesy of OSU-HPCC website)

http://hpcwiki.it.okstate.edu/index.php/Cowboy

OSUHPCC's newest supercomputer, Cowboy, was funded by NSF MRI grant “Acquisition of a High Performance Compute Cluster for Multidisciplinary Research,” OCI-1126330, 9/1/11-8/31/14, $908,812, PI Brunson.

This cluster, Cowboy, from Advanced Clustering Technologies consists of: 252 standard compute nodes, each with dual Intel Xeon E5-2620 “Sandy Bridge” hex core 2.0 GHz CPUs, with 32 GB of 1333 MHz RAM and two “fat nodes” each with 256 GB RAM and an NVIDIA Tesla C2075 card. The aggregate peak speed is 48.8 TFLOPs, with 3048 cores, 8576 GB of RAM.

Cowboy also includes 92 TB of globally accessible high performance disk provided by three shelves of Panasas ActivStor12, this includes 20x 2TB drives and peak speed of 1500MB/s read and 1600MB/s write per shelf. The total solution provides an aggregate of 4.5GB/s read and 4.8GB/s write.

The interconnect networks are Infiniband for message passing, Gigabit Ethernet for I/O, and an ethernet management network. The Infiniband for message passing is Mellanox Connect-X 3 QDR in a 2:1 oversubscription. There are a total of 15x MIS5025Q switches providing both the leaf and spine components. Each leaf is connects to 24 compute nodes, and 12x 40Gb QDR links to the spine. Point-to-point latency is approx 1 microsecond. The ethernet network includes 11 leaf gigabit switches that connect to 24 compute nodes. Each leaf is uplinked via 2x 10G network ports to the spine 64 port Mellanox MSX1016 10 Gigabit switch. The network configuration provides a 1.2:1 oversubscription.

5. Useful Links

1. http://www.gaussian.com/g_tech/g_ur/g09help.htm Gaussian09 User's Reference

2. http://hpcwiki.it.okstate.edu/index.php/Main_Page OSU HPC wiki

3. http://cccbdb.nist.gov/ Experimental and Computational thermochemical data for a selected set of 1420 gas-phase atoms and molecules.

Appendix A.4

Gaussian FAQ with Trouble Shooting

Zhouyang Kang

Department of Physics, Oklahoma State University

Q1: Why my job died?

There are many types of reason for a job to be terminated with error. Open the output file and go to the last line to read the error message.

Q2: What if my job died because convergence cannot be achieved?

This error may happen during geometry optimization with error code 9999 at the end of the output file.

If the job is for frequency calculation, the normal criteria for convergence have to be achieved in order to obtain force constant for accurate frequency results. Open the output file, check the convergence of each step by searching the word "energy=-". Find the step with best convergence then delete the rest steps, save it as a separated output file. Use Gaussview to open and edit it then save it as a new input. Restart the calculation and keep monitoring.

If the job is for energy calculation, particularly with fixed geometry constants, the normal criteria for convergence are not necessary to be achieved. Open the output file, check the convergence of each step and find the predicted changes between each step by searching the word "Predicted change in Energy=-". The value followed is in the unit of AU (atomic unit). Multiply that value by 2619.6 to convert it into kJ/mol. Since our requirement for energy landscape is in the precision of 0.02kJ/mol, any step with a predicted change of energy below 7.0D-05 can be accepted as the job may be terminated early.

Q3: What if my job died because of error “Conversion from Z-matrix to cartesian coordinates failed:”

This error may happen when the input geometry is in Z-matrix format. When a certain angle value is around 0, during the calculation Gaussian may have the angle change to negative, and such error will happen because of the legal range for an angle is between 0 to 180 degrees.

To solve the problem, first open the output file in notebook to the end and locate the “negative angle”. Then open the output in Gaussview, edit the “negative angle” to a positive value around 10 degrees. Save it as a new input and restart the calculation.

Appendix A.5

LIST OF DEFINED SOLVENTS

The following solvent keywords are accepted with the **SCRF=Solvent** option. We list the ε values here for convenience, but be aware it is only one of many internal parameters used to define solvents. Thus, simply changing the ε value will not define a new solvent properly.

- **Water**: ε=78.3553
- **Acetonitrile**: ε=35.688
- **Methanol**: ε=32.613
- **Ethanol**: ε=24.852
- **IsoQuinoline**: ε=11.00
- **Quinoline**: ε=9.16
- **Chloroform**: ε=4.7113
- **DiethylEther**: ε=4.2400
- **Dichloromethane**: ε=8.93
- **DiChloroEthane**: ε=10.125
- **CarbonTetraChloride**: ε=2.2280
- **Benzene**: ε=2.2706
- **Toluene**: ε=2.3741
- **ChloroBenzene**: ε=5.6968
- **NitroMethane**: ε=36.562
- **Heptane**: ε=1.9113
- **CycloHexane**: ε=2.0165
- **Aniline**: ε=6.8882
- **Acetone**: ε=20.493
- **TetraHydroFuran**: ε=7.4257
- **DiMethylSulfoxide**: ε=46.826
- **Argon**: ε=1.430
- **Krypton**: ε=1.519
- **Xenon**: ε=1.706
- **n-Octanol**: ε=9.8629
- **1,1,1-TriChloroEthane**: ε=7.0826
- **1,1,2-TriChloroEthane**: ε=7.1937
- **1,2,4-TriMethylBenzene**: ε=2.3653
- **1,2-DiBromoEthane**: ε=4.9313
- **1,2-EthaneDiol**: ε=40.245
- **1,4-Dioxane**: ε=2.2099
- **1-Bromo-2-MethylPropane**: ε=7.7792
- **1-BromoOctane**: ε=5.0244
- **1-BromoPentane**: ε=6.269
- **1-BromoPropane**: ε=8.0496
- **1-Butanol**: ε=17.332
- **1-ChloroHexane**: ε=5.9491
- **1-ChloroPentane**: ε=6.5022

- **1-ChloroPropane**: ε=8.3548
- **1-Decanol**: ε=7.5305
- **1-FluoroOctane**: ε=3.89
- **1-Heptanol**: ε=11.321
- **1-Hexanol**: ε=12.51
- **1-Hexene**: ε=2.0717
- **1-Hexyne**: ε=2.615
- **1-IodoButane**: ε=6.173
- **1-IodoHexaDecane**: ε=3.5338
- **1-IodoPentane**: ε=5.6973
- **1-IodoPropane**: ε=6.9626
- **1-NitroPropane**: ε=23.73
- **1-Nonanol**: ε=8.5991
- **1-Pentanol**: ε=15.13
- **1-Pentene**: ε=1.9905
- **1-Propanol**: ε=20.524
- **2,2,2-TriFluoroEthanol**: ε=26.726
- **2,2,4-TriMethylPentane**: ε=1.9358
- **2,4-DiMethylPentane**: ε=1.8939
- **2,4-DiMethylPyridine**: ε=9.4176
- **2,6-DiMethylPyridine**: ε=7.1735
- **2-BromoPropane**: ε=9.3610
- **2-Butanol**: ε=15.944
- **2-ChloroButane**: ε=8.3930
- **2-Heptanone**: ε=11.658
- **2-Hexanone**: ε=14.136
- **2-MethoxyEthanol**: ε=17.2
- **2-Methyl-1-Propanol**: ε=16.777
- **2-Methyl-2-Propanol**: ε=12.47
- **2-MethylPentane**: ε=1.89
- **2-MethylPyridine**: ε=9.9533
- **2-NitroPropane**: ε=25.654
- **2-Octanone**: ε=9.4678
- **2-Pentanone**: ε=15.200
- **2-Propanol**: ε=19.264
- **2-Propen-1-ol**: ε=19.011
- **3-MethylPyridine**: ε=11.645
- **3-Pentanone**: ε=16.78
- **4-Heptanone**: ε=12.257
- **4-Methyl-2-Pentanone**: ε=12.887
- **4-MethylPyridine**: ε=11.957
- **5-Nonanone**: ε=10.6
- **AceticAcid**: ε=6.2528
- **AcetoPhenone**: ε=17.44
- **a-ChloroToluene**: ε=6.7175
- **Anisole**: ε=4.2247

- **Benzaldehyde**: ε=18.220
- **BenzoNitrile**: ε=25.592
- **BenzylAlcohol**: ε=12.457
- **BromoBenzene**: ε=5.3954
- **BromoEthane**: ε=9.01
- **Bromoform**: ε=4.2488
- **Butanal**: ε=13.45
- **ButanoicAcid**: ε=2.9931
- **Butanone**: ε=18.246
- **ButanoNitrile**: ε=24.291
- **ButylAmine**: ε=4.6178
- **ButylEthanoate**: ε=4.9941
- **CarbonDiSulfide**: ε=2.6105
- **Cis-1,2-DiMethylCycloHexane**: ε=2.06
- **Cis-Decalin**: ε=2.2139
- **CycloHexanone**: ε=15.619
- **CycloPentane**: ε=1.9608
- **CycloPentanol**: ε=16.989
- **CycloPentanone**: ε=13.58
- **Decalin-mixture**: ε=2.196
- **DiBromomEthane**: ε=7.2273
- **DiButylEther**: ε=3.0473
- **DiEthylAmine**: ε=3.5766
- **DiEthylSulfide**: ε=5.723
- **DiIodoMethane**: ε=5.32
- **DiIsoPropylEther**: ε=3.38
- **DiMethylDiSulfide**: ε=9.6
- **DiPhenylEther**: ε=3.73
- **DiPropylAmine**: ε=2.9112
- **e-1,2-DiChloroEthene**: ε=2.14
- **e-2-Pentene**: ε=2.051
- **EthaneThiol**: ε=6.667
- **EthylBenzene**: ε=2.4339
- **EthylEthanoate**: ε=5.9867
- **EthylMethanoate**: ε=8.3310
- **EthylPhenylEther**: ε=4.1797
- **FluoroBenzene**: ε=5.42
- **Formamide**: ε=108.94
- **FormicAcid**: ε=51.1
- **HexanoicAcid**: ε=2.6
- **IodoBenzene**: ε=4.5470
- **IodoEthane**: ε=7.6177
- **IodoMethane**: ε=6.8650
- **IsoPropylBenzene**: ε=2.3712
- **m-Cresol**: ε=12.44
- **Mesitylene**: ε=2.2650

- **MethylBenzoate**: ε=6.7367
- **MethylButanoate**: ε=5.5607
- **MethylCycloHexane**: ε=2.024
- **MethylEthanoate**: ε=6.8615
- **MethylMethanoate**: ε=8.8377
- **MethylPropanoate**: ε=6.0777
- **m-Xylene**: ε=2.3478
- **n-ButylBenzene**: ε=2.36
- **n-Decane**: ε=1.9846
- **n-Dodecane**: ε=2.0060
- **n-Hexadecane**: ε=2.0402
- **n-Hexane**: ε=1.8819
- **NitroBenzene**: ε=34.809
- **NitroEthane**: ε=28.29
- **n-MethylAniline**: ε=5.9600
- **n-MethylFormamide-mixture**: ε=181.56
- **n,n-DiMethylAcetamide**: ε=37.781
- **n,n-DiMethylFormamide**: ε=37.219
- **n-Nonane**: ε=1.9605
- **n-Octane**: ε=1.9406
- **n-Pentadecane**: ε=2.0333
- **n-Pentane**: ε=1.8371
- **n-Undecane**: ε=1.9910
- **o-ChloroToluene**: ε=4.6331
- **o-Cresol**: ε=6.76
- **o-DiChloroBenzene**: ε=9.9949
- **o-NitroToluene**: ε=25.669
- **o-Xylene**: ε=2.5454
- **Pentanal**: ε=10.0
- **PentanoicAcid**: ε=2.6924
- **PentylAmine**: ε=4.2010
- **PentylEthanoate**: ε=4.7297
- **PerFluoroBenzene**: ε=2.029
- **p-IsoPropylToluene**: ε=2.2322
- **Propanal**: ε=18.5
- **PropanoicAcid**: ε=3.44
- **PropanoNitrile**: ε=29.324
- **PropylAmine**: ε=4.9912
- **PropylEthanoate**: ε=5.5205
- **p-Xylene**: ε=2.2705
- **Pyridine**: ε=12.978
- **sec-ButylBenzene**: ε=2.3446
- **tert-ButylBenzene**: ε=2.3447
- **TetraChloroEthene**: ε=2.268
- **TetraHydroThiophene-s,s-dioxide**: ε=43.962
- **Tetralin**: ε=2.771

- **Thiophene**: ε=2.7270
- **Thiophenol**: ε=4.2728
- **trans-Decalin**: ε=2.1781
- **TriButylPhosphate**: ε=8.1781
- **TriChloroEthene**: ε=3.422
- **TriEthylAmine**: ε=2.3832
- **Xylene-mixture**: ε=2.3879
- **z-1,2-DiChloroEthene**: ε=9.2

Appendix B.1

Protocol of Molecular Mutagenesis of Hhal PYP

Zhouyang Kang

Department of Physics, Oklahoma State University

Introduction

This protocol is developed during the process of preparing site-directed mutagenesis for Hhal PYP in 2010.

Content of Index

Protocol 1: Primer Preparation

Method

1) Received primers are dried powder and amount of primers were labeled on the tube in the unit of nmole. Calculate the amount of H_2O needs to be added in order to make 100µM stock solution.

*Y (µl)= 10*X (nmole)*

2) Add Y µl H2O to each corresponding primer tube, dissolve by tapping the bottom of the tube, store in deep freezer at -80 Celsius.

3) Prepare 1.5ml Eppendorf centrifuge tube, add 36 µl H2O, and mix with 4 µl 100µM stock primer solution to make 10µM primer solution for regular use, stock in -20 Celsius.

Protocol 2: Overlap Extension Technique: 1st PCR

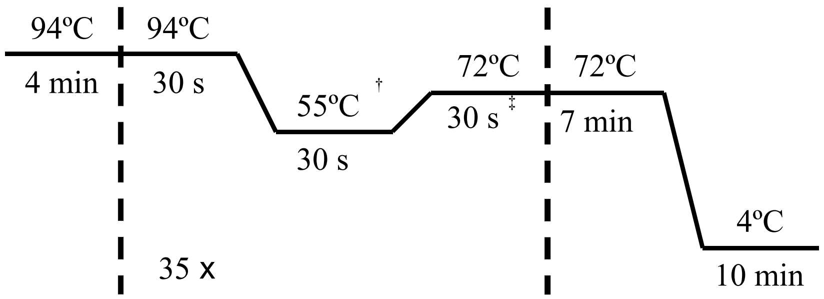
Method

1) Prepare 50 µl sample reaction according to the table below in a 0.2 ml PCR tube. (*GoTaq* Flexi DNA Polymerase pack purchased from Core Facility)

Component	amount (µl)	Remarks
Nuclease-free H_2O (autoclaved H2O)	*x*	add enough to obtain a total reaction volume to 50 µl.
5× *GoTaq* Flexi Reaction Buffer	10.00	Use colorless buffer
25mM $MgCl_2$	3	1.5mM final concentration in PCR mix
DNA template	*x*	add ~15ng plasmid DNA (5~6kbp)
FW Flanking Primer (10 µM)[*]	1.00	
RV Internal Primer (10 µM)[*]	1.00	
dNTP (10 mM each NTP)	1.50	
GoTaq DNA Polymerase[#]	1.00	

[#] *Note*: ***Taq DNA Polymerase*** *should be the last component added to minimize primer degradation. It is a temperature sensitive enzyme, keep in freezer or ice before use.*

2) Run a temperature cycling program on the reaction mix. For the temperature cycling parameters use the scheme and notes below as a guide.



† Annealing temperature should be optimized to 10 Celsius below melting temperature (T_m) of primer.

3) Prepare 50 µl sample reaction according to the table below and run the same temperature cycling program on the reaction mix as step 2. This step should be done parallel with step 1-2.

Component	amount (µl)	Remarks
Nuclease-free H_2O (autoclaved H2O)	*x*	add enough to obtain a total reaction volume to 50 µl.
5× *GoTaq* Flexi Reaction Buffer	10.00	Use colorless buffer
25mM $MgCl_2$	3	1.5mM final concentration in PCR mix
DNA template	*x*	add ~15ng plasmid DNA (5~6kbp)
FW Internal Primer (10 µM)[*]	1.00	
RV Flanking Primer (10 µM)[*]	1.00	
dNTP (10 mM each NTP)	1.50	
GoTaq DNA Polymerase[#]	1.00	

#Note: ***Taq DNA Polymerase*** *should be the last component added to minimize primer degradation. It is a temperature sensitive enzyme, keep in freezer or ice before use.*

* Flanking Primer: Primers that hybridize at one end of the target sequence (Ho, 1989)

* Internal Primer: Primers that hybridizes at the site of the mutation and contains the mismatched bases (Ho, 1989)

Protocol 3: Overlap Extension Technique: 2nd PCR

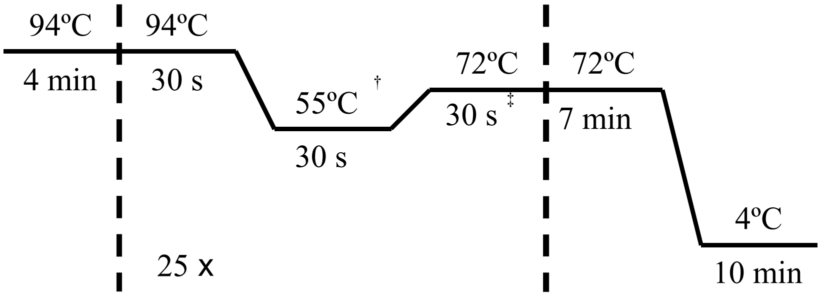
Method

1) Prepare 50 µl sample reaction according to the table below in a 0.2 ml PCR tube.

Component	amount (µl)	Remarks
Nuclease-free H_2O (autoclaved H2O)	*x*	add enough to obtain a total reaction volume to 50 µl.
5× *GoTaq* Flexi Reaction Buffer	10.00	Use colorless buffer
25mM $MgCl_2$	3	1.5mM final concentration in PCR mix
FW Flanking Primer (10 µM)*	1.5	
RV Flanking Primer (10 µM)*	1.5	
1st PCR Product (FW)	1.25	
1st PCR Product (RV)	1.25	
dNTP (10 mM each NTP)	1.50	
GoTaq DNA Polymerase#	1.00	

#*Note*: ***Taq DNA Polymerase*** *should be the last component added to minimize primer degradation. It is a temperature sensitive enzyme, keep in freezer or ice before use.*

2) Run a temperature cycling program on the reaction mix. For the temperature cycling parameters use the scheme and notes below as a guide.



† Annealing temperature should be optimized to 10 Celsius below melting temperature (T_m) of primer.

Reference (Protocol 2-3)

Adapted from a protocol obtained from Dr. Masato Kumauchi
Site-directed Mutagenesis using PCR overlap extension techniques: (Ho, 1989)

Protocol 4: QIAquick PCR Purification Kit Protocol (100 bp - 10 kb)

This protocol purifies DNA fragments ranging from 100 bp to 10 kb from primers, nucleotides, polymerases, restriction enzymes, and salts using QIAquick spin columns in a microcentrifuge.

Method

All centrifuge steps are at 10,000 rcf

3) Add 5 volumes of Buffer PB to 1 volume of reaction mix containing the DNA to be purified.

> *Note: Removal of mineral oil or kerosene is not necessary, but should not be considered as part of the reaction mix when determining the volume of Buffer PB to be added.)*

Place a QIAquick spin column in a provided 2ml collection tube.

To bind DNA, apply the sample to the QIAquick column and centrifuge for 1 minute.

Discard flow-through and place the QIAquick column back into the tube.

To wash, add 0.75 ml Buffer PE to the QIAquick column and centrifuge for 1 minute.

Discard the flow-through, and centrifuge an additional 1 minute at 17,900 rcf to remove residual wash buffer.

Place the QIAprep Spin Column in a clean 1.5 ml microcentrifuge tube. Repeat centrifuge at 17,900 rcf for 1 minute multiple times until no extra liquid comes out from the column. (This step is critical because Buffer **PE** contains 70% ethanol, which could damage enzymes that will be used in the reaction afterwards).

Place the QIAquick column in a clean 1.5 ml microcentrifuge tube.

To elute DNA, add 50 l Buffer EB (or 10 mM Tris·HCl, pH 8.5 or H_2O pH 7-8.5) to the center of the QIAquick membrane, let the column stand for 20 minutes and centrifuge the column for 1 minute at 17,900 rcf.

> *Note: The provided Buffer **EB** may have a much lower pH than 8.5, even though according to the label the pH should be 8.5. This should be checked if you are going to use this buffer.*

Addapted from QIAquick® Spin Handbook (March 2008) Qiagen®.

> *Note: no specifics are supplied by Qiagen on the composition of Buffers PB and PE.*

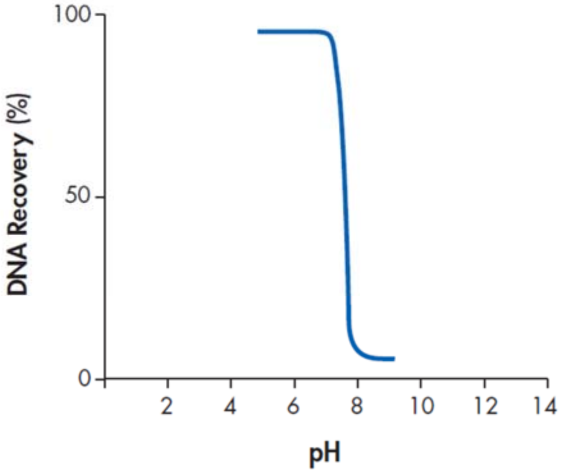


Figure 2. pH dependence of DNA adsorption to QIAquick membranes. 1 µg of a 2.9 kb DNA fragment was adsorbed at different pHs and eluted with Buffer EB (10 mM Tris·Cl, pH 8.5). The graph shows the percentage of DNA recovery, reflecting the relative adsorption efficiency, versus pH of adsorption.

Protocol 5: QIAprep® Spin Miniprep Kit Protocol

This protocol should result in a plasmid isolate suitable for sequencing.

Method

1) Prepare an overnight culture (of a strain containing the wanted plasmid) in LB with an appropriate antibiotic (usually Ampicillin).

> *Note 1: Best results are obtained if the plasmid is contained within an* E. coli *DH5α strain.*
>
> *Note 2: In order to obtain high plasmid yields, use LB medium containing 10 g·l^{-1} NaCl.*
>
> *Note 3: Try to not grow the overnight culture for longer than 16 hours as this may negatively influence the plasmid yield.*

Centrifuge 5-8 ml overnight culture for 10 min at 2250 rcf at 4 Celsius. (Use Allegra X-12R centrifuge in Dr. Hoff's lab, select SX4750 rotor, set speed 3750 rpm)

> *Note: Collect supernatant(s) separately and pull in the liquid waste bottle (contain bleach in the bottle, will be autoclaved before discard to the drain)!*

Resuspend the pelleted bacterial cells in 250 µl Buffer **P1**.

Add 250 µl Buffer **P2** and gently invert the tube 4-6 times. Do not allow the resulting lysis reaction to proceed for more than **5 minutes**.

Add 350 µl Buffer **N3** and invert the tube immediately and gently 4-6 times. A precipitate should be formed now.

Centrifuge for 10 minutes at 17,900 rcf. (Use Eppendorf 5417C microcentrifuge in Dr. Hoff's lab, set speed 13000 rpm)

Apply the supernatant from step 6 to the QIAprep Spin Column using a pipette.

> *Note : The pellet is usually on the side of the eppendorf cup and not on the bottom.*

Centrifuge for 1 minute at 17,900 rcf, and discard the flow-trough.

> *Note : If you opted to prepare multiple bacterial pellets of the same culture you can repeat step 7 and 8 until all have been applied to the same QIAprep Spin Column.*

Wash the QIAprep Spin Column by adding 750 µl Buffer **PE** and centrifuging for 1 minute at 17,900 rcf.

Discard the flow-through, and centrifuge an additional 1 minute at 17,900 rcf to remove residual wash buffer.

Place the QIAprep Spin Column in a clean 1.5 ml eppendorf cup. Repeat centrifuge at 17,900 rcf for 1 min multiple times until no extra liquid come out from the column. (This step is critical because Buffer **PE** contains 70% ethanol, which could damage enzymes that will be used in the reaction afterwards)

Place the QIAprep Spin Column in a clean 1.5 ml eppendorf cup. To elute DNA, apply 50 µl 10 mM Tris·HCl pH 8.4 to the center of the QIAprep Spin Column, let it stand for 20 minutes and centrifuge for 1 minute at 17,900 rcf.

> *Note: The provided Buffer **EB** may have a much lower pH than 8.5, even though according to the label the pH should be 8.5. This should be checked if you are going to use this buffer.*

Materials

Buffer / Medium	Composition	Remarks
LB	10 g·l^{-1} (Bacto-)Trypton 5 g·l^{-1} Yeast Extract 10 g·l^{-1} NaCl	The use of 10 g·l^{-1} NaCl, instead of 5 g·l^{-1}, improves plasmid yield.
P1 (resuspension buffer)	50 mM Tris·HCl pH 8.0 10 mM EDTA 100 µg·ml^{-1} RNase A	store at 2-8°C, after addition of RNase A
P2 (Lysis buffer)	200 mM NaOH 1% SDS (w/v)	store at room temperature
N3 (Neutralization buffer)	Guanidium chloride 25-50% Acetic acid 10-25% and other nonhazardous additions	store at room temperature
10 mM Tris·HCl pH 8.4	prepare from 1 M stock via 100x dilution. 1 M Tris (12.11 g per 100 ml) pH 8.4 (adjust pH with HCl)	store at room temperature

References

Adapted from 'QIAquick® Spin Handbook', March 2008, Qiagen®.

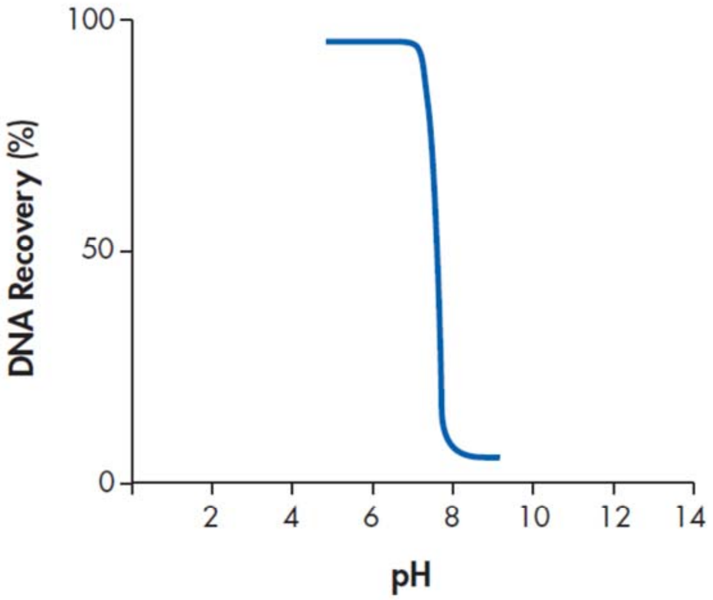


Figure 2. pH dependence of DNA adsorption to QIAquick membranes. 1 µg of a 2.9 kb DNA fragment was adsorbed at different pHs and eluted with Buffer EB (10 mM Tris·Cl, pH 8.5). The graph shows the percentage of DNA recovery, reflecting the relative adsorption efficiency, versus pH of adsorption.

Protocol 6: Agarose gel electrophoresis

Method

1) Based on the number of samples, choose gel tray with right size (use small tray for 1-11 samples, use large tray for 12-24 samples). Place gel tray in the white gel caster and put the comb in place.

For a small gel dissolve 0.5-2% w/v high melting-temperature agarose in 50 ml TAE buffer (prepare from 50x stock) in an 250ml flask by boiling the mixture (use microwave oven). The exact amount of agarose depends on the size of the fragments you want to separate. See table below.

Agarose concentration (% w/v)	DNA Size (kb)
0.50 (0.25 g per 50 ml)	1-30
0.75 (0.375 g per 50 ml)	0.8-12
1.00 (0.5 g per 50 ml)	0.5-10
1.25 (0.625 g per 50 ml)	0.4-7
1.50 (0.75 g per 50 ml)	0.2-3
2-5* (1-2.5 g per 50 ml)	0.01-0.5
* Sieving Agarose	

2) Poor the molten agarose into the gel tray. Allow 20-40 minutes for the gel to solidify. (Lower concentration Agarose solution needs longer solidifying time)

After the gel has solidified, take gel tray out, use Kimwipe carefully wipe out all the small gel fragments attached on the outside of the tray (those small fragments will cause electrophoresis failure). Then place the tray with gel into the Electrophoresis Cell filled with TAE buffer (1x). The gel should only be slightly submerged in buffer (2-6 mm).

Prepare loading samples:

DNA samples: 1 µl 6x Loading Dye + 2 µl DNA + 3 µl 1x TAE buffer = 6 µl total loading sample
Reference: 1 µl 6x Loading Dye + 1 µl DNA ladder + 3 µl 1x TAE buffer =6 µl total loading sample

Carefully mix the loading sample by pipetting, load 6 µl samples to each well. Make sure you have one well for DNA ladder reference. You may need to add more DNA if concentration is too low, then you can reduce the amount of TAE buffer to make total volume of loading sample stay at 6 µl.

Run the gel with 100V. When the blue front runs 3 quarters way of the gel (15~20 min), stop electrophoresis. Take the gel out (without the tray!), carefully put it in the box that contains Ethidium Bromide (EB) TAE buffer, and shake for 25-30 min.

Wear gloves! Take gel out of EB TAE buffer, use AlphaImager (Dr. Hoff's lab) to take UV image of the gel. Save the image, then take gel out and use Kimwipe tissue with 70% Ethanol to clean the gel stand in the machine, discard the gel, tissue and gloves afterwards.

Materials

Buffer	Composition	Remarks
Electrophoresis buffer (TAE buffer)	40 mM Tris 20 mM Acetic acid 1 mM EDTA pH 8.3	Store at room temperature. *Prepare from 50× stock solution.*
50× TAE buffer	per 0.5 liter:	Store at room temperature.

	121.14 g Tris Base 28.6 ml glacial acetic acid (30.03 g) 9.31 g $Na_2EDTA \cdot 2H_2O$ (Titriplex III)	
Sample loading dye (10x stock)	50% (v/v) glycerol 0.25% (/v) bromophenol blue 0.25% (/v) xylene cyanole FF in 1x TAE buffer	store at room temperature. *No more than 1-10 ml should be prepared of the 10x stock.* *You can also use only one of the dyes,* ***i.e****. only bromophenol blue, or only xylene cyanole FF*

Protocol 7: Double Digestion (Dr. Masato Kumauchi)

Different plasmid/insert strain may need different restriction enzymes for digestion based on what restriction site plasmid/insert strain has in the sequence. In this protocol the pET16b and PYP insert strain are specifically designed for *NcoI* and *BamHI* enzymes double digestion. For other restriction site and corresponding enzymes, you may check at http://en.wikipedia.org/wiki/List_of_restriction_enzyme_cutting_sites or use Enzyme finder provided by *New England Biolabs*: http://www.neb.com/nebecomm/EnzymeFinder.asp

Method

3) Prepare 60 µl double digestion mix sample, see table below.

Component	amount (µl)	Remarks
Nuclease-free H_2O (autoclaved H2O)	11.40	add enough to obtain a total reaction volume to 60 µl.
10× NEBuffer 3	6.00	
DNA solution	40.00	
BSA	0.60	Prevent BamHI sticky to tube wall
BamHI#	1.00	5'…G^GATCC…3'
NcoI/NdeI#*	1.00	5'…C^CATGG…3'/5'…CA^TATG…3'

#*Note*: *BamHI, NcoI and NdeI are temperature sensitive enzymes that should be stored at -20 Celsius all the time. Make sure you take them out of freezer just before you are ready to add them and keep them in ice. Put enzymes back to freezer immediately when you finish.*

**Different mutant may need different restriction site and different plasmid. For PYP mutation that replacing one of the amino acids in PYP sequence, NcoI enzyme and pET-16b plasmid are used, for PYP mutation to add extra tags at the beginning of the sequence, NdeI enzyme and pET-26b(+) plasmid are used. Plasmid maps of pET-16b and pET-26b(+) are available in Appendix.*

4) Incubate double digestion mixture at 37 Celsius for at lease 2 hours (*NcoI-BamHI*) or 4 hours (*NdeI-BamHI*).

Note: Both BamHI and NcoI will loose most of activities after 8 hour, NdeI will not loose activities even after 8 hours, however no more than 4 hours of digestion time is recommended in order to prevent star activity.

5) After digestion, use Protocol 6 to purify the product and check gel electrophoresis.

Materials

Buffer	Composition	Remarks
1x NEBuffer 3	50mM Tris-HCl 100mM NaCl 10mM $MgCl_2$ 1mM Dithiothreitol	

References

Adapted from a protocol obtained from Dr. Masato Kumauchi

Protocol 8: Double Digestion (Dr. Junpeng Deng)

This protocol is for insert Tyr mutant PYP sequence into pET16b plasmid (Ampicillin resistance), for insert Extra Tag PYP sequence into pET26b plasmid (Kanamycin resistance), go to Protocol 9

Different plasmid/insert strain may need different restriction enzymes for digestion based on what restriction site plasmid/insert strain has in the sequence. In this protocol the pET16b and PYP insert strain are specifically designed for *NcoI* and *BamHI* enzymes double digestion. For other restriction site and corresponding enzymes, you may check at http://en.wikipedia.org/wiki/List_of_restriction_enzyme_cutting_sites or use Enzyme finder provided by *New England Biolabs*: http://www.neb.com/nebecomm/EnzymeFinder.asp

Method

1) Prepare 20 µl double digestion mix sample in 0.2 ml PCR tube, see table below.

Component	amount (µl)	Remarks
Nuclease-free H_2O (autoclaved H2O)	7	add enough to obtain a total reaction volume to 20 µl.
10× NEBuffer 4	2.00	Final concentration 1X
wtPYP in pET16b solution	10.00	
BamHI-HF[#*]	1.00	5'…G^GATCC…3'
NcoI-HF[#*]	1.00	5'…C^CATGG…3'
10× Antarctic Phosphatase Reaction Buffer[$]	2.00	Final concentration 1X
Antarctic Phosphatase[$]	1.00	Prevent self ligation

[#]*Note*: *BamHI, NcoI and NdeI are temperature sensitive enzymes that should be stored at -20 Celsius all the time. Make sure you take them out of freezer just before you are ready to add them and keep them in ice. Put enzymes back to freezer immediately when you finish.*

**Different mutant may need different restriction site and different plasmid. For PYP mutation that replacing one of the amino acids in PYP sequence, NcoI enzyme and pET-16b plasmid are used, for PYP mutation to add extra tags at the beginning of the sequence, NdeI enzyme and pET-26b(+) plasmid are used. Plasmid maps of pET-16b and pET-26b(+) are available in Appendix.*

[$]*Add to plasmid double digestion reaction mixture only. For insert DNA, instead of Phosphatase buffer and Phosphatase, add H_2O to make final volume 20 µl.*

2) Incubate double digestion mixture at 37 Celsius for 1 hour, do not exceed 2 hours!!!

Note: Both BamHI-HF and NcoI-HF will loose most of activities after 8 hour, NdeI will not loose activities even after 8 hours, however no more than 4 hours of digestion time is recommended in order to prevent star activity.

3) After digestion, use Protocol 6 to purify the product and check gel electrophoresis.

Materials

Buffer	Composition	Remarks
1x NEBuffer 3	50mM Tris-HCl 100mM NaCl 10mM $MgCl_2$ 1mM Dithiothreitol	

References

Adapted from a protocol obtained from Dr. Junpeng Deng

Protocol 9: Ligation (use QuickLigase)

This protocol will result in the ligation of digested insert strain with digested vector

Method

1) Use UV-Vis to check the amount of vector and insert. (Check OD@260 nm, use OD@330nm as baseline)

*Concentration of DNA (ng/µl) = (Abs_{260}-Abs_{330}) * dilution factor * 500*

2) Combine 25 ng of vector with a 3-fold molar excess of insert. Adjust volume to 5 µl with autoclaved H_2O

*Note: pET16b NcoI-BamHI vector size~5.7kbp, PYP insert size~0.4kbp. Thus 25 ng pET16b vector should mix with 3*25*0.4/5.7 = 5.3ng of PYP insert*

3) Add 5 µl 2x Quick Ligation Buffer and mix

4) Add 0.5 µl of Quick T4 DNA Ligase and mix thoroughly

5) Centrifuge briefly and incubate at room temperature (25 Celsius) for 5 minutes

6) Chill on ice, then transform or store at -20 Celsius

Note: Do not heat inactivate. Heat inactivation dramatically reduces transformation efficiency.

Materials

Buffer	Composition	Remarks
2x Quick Ligation Buffer	132 mM Tris-HCl 20mM MgCl2 2mM dithiothreitol 2mM ATP 15% Polyethylene glycol (PEG 6000)	

References

Adapted from 'Quick Ligation Protocol', New England Biolabs

http://www.neb.com/nebecomm/products/protocol2.asp

Notes

Previous experiences shows QuickLigase has low ligation efficiency and high error rate

Protocol 10: Ligation (use T4 DNA Ligase)

This protocol will result in the ligation of digested insert strain with digested vector

Method

7) Use *NanoDrop* (core facility) to check the amount of vector and insert. (Check OD@260 nm, use OD@330nm as baseline)

8) For 5 µl reaction volume, mix:

a. 0.5 µl 10X T4 DNA Ligase Reaction Buffer

b. 0.5 µl T4 DNA Ligase

c. 3.5 µl Double-digested DNA insert

d. 0.5 µl Double-digested vector

9) Centrifuge briefly and incubate at room temperature (25 Celsius) for 2 hours.

10) Chill on ice, then transform or store at -20 Celsius

Note: Do not heat inactivate. Heat inactivation dramatically reduces transformation efficiency.

Materials

Buffer	Composition	Remarks
10x T4 DNA Ligase Reaction Buffer	50 mM Tris-HCl 10mM MgCl2 10mM DTT 1mM ATP pH 7.5@25 Celsius	

References

Adapted from 'T4 DNA Ligation Protocol', New England Biolabs and Dr. Junpeng Deng's protocol

http://www.neb.com/nebecomm/products/protocol2.asp

Protocol 11: Transformation (for NEB 5-alpha Competent *E. coli*)

This protocol will result in the transformation of quick ligation products to competent cell.

Method

Before start, make sure water bath has reached temperature of 42 Celsius. Then perform steps 1-7 in the tube provided by NEB (with 50 µl competent cell inside).

1) Thaw a tube of NEB 5-alpha Competent *E.coli* cells on ice for 10 minutes.

2) Add 1-5 µl containing 1 pg- 100 ng of plasmid DNA to the cell mixture. Carefully flick the tube 4-5 times to mix cells and DNA. **DO not vortex.**

3) Place the mixture on ice for 30 minutes. Do not mix.

4) Heat shock at exactly 42 Celsius for exactly 30 seconds. Do not mix.

Note: different competent E.coli have different requirement for timing of heat shock. NEB 5-alpha requires exactly 30s, while BL21 (DE3) requires exactly 10s. Check user manuals of each competent E.coli.

5) Place on ice for 5 minutes. Do not mix.

6) Pipette 950 µl of room temperature SOC into the mixture.

7) Place at 37 Celsius for 60 minutes. Shake vigorously (250 rpm) or rotate.

8) Warm selection plates to 37 Celsius.

9) Mix the cells thoroughly by flicking the tube and inverting, then perform several 10-fold serial dilutions in SOC.

10) Spread 50-100 µl of each dilution onto a selection plate and incubate overnight at 37 Celsius.

11) Add 100 µl of original dilution mixture to 900 µl LB with Ampicillin, incubate overnight.

Note: Competent cells with plasmid successfully transformed can be survived in LB with Ampicillin added because there is ampicillin site in plasmid. This medium with living cells can be used for further plating.

Materials

Medium	Composition	Remarks
SOC	2% Vegetable peptone (or Tryptone) 0.5% Yease Extract 10mM NaCl 2.5mM KCl 10mM $MgCl_2$ 10mM $MgSO_4$ 20mM Glucose	No antibiotics added, carefully operate, work near the flame.

References

Adapted from 'Certificate of Analysis for NEB 5-alpha Competent E.coli (High Efficiency)', New England Biolabs

Protocol 12: Electroporation Transformation (for home-made DH5α Competent *E. coli*)

This protocol will result in the transformation of ligation products to competent cell.

Method

1) Clean 1mm electroporation cuvettes with 90% ethanol then rinse de-ionized H_2O several times. Try to dry cuvettes as much as you can after the rinse. Then pre-chill cuvettes on ice.

2) Prepare enough amount of SOC recovery medium or LB w/o antibiotics in several 1.5 ml eppendorf microcentrifuge tubes. (minimum 250 µl medium in one tube for one transformation)

Note: SOC stored at 4 Celsius in refrigerator, LB stored in glass bottle at room temperature. Make sure you are working near the flame at this step.

3) Thaw a tube of home-made DH5α competent *E.coli* on ice for 10min.

4) Collect things below on a cart then move to the small room near entrance of Dr. Hoff's lab in shared facility.

0.5-10 µl Eppendorf Pipette and tips for DNA purpose

20-200 µl Eppendorf Pipette and tips for General purpose

100-1000 µl Eppendorf Pipette and tips for General Purpose

Kimwipe tissues

1 dispensing bottle of de-ionized water

1 dispensing bottle of 90% ethanol

Ice basket with competent cells, ligased DNA and cuvettes on ice

Tube rack carried labeled empty microcentrifuge tubes (for incubation of competent cells after transformation) and enough tubes with room temperature SOC or LB w/o antibiotics

5) Add 1-2 µl containing 1 pg- 100 ng of ligased DNA or plasmid to the 50µl cell mixture. Carefully flick the tube 4-5 times to mix cells and DNA. **DO not vortex.**

Note: Make sure you do not introduce more than ~0.25mM final ion concentration to DNA-Competent cell mix (add less than 2 µl ligased mix to 50 µl competent cells), which will cause low efficiency of transformation or even cause arc current to kill all competent cells.

6) Heat shock at exactly 42 Celsius for exactly 30 seconds. Do not mix.

Note: different competent E.coli have different requirement for timing of heat shock. NEB 5-alpha requires exactly 30s, while BL21 (DE3) requires exactly 10s. Check user manuals of each competent E.coli.

7) Place on ice for 5 minutes. Do not mix.

8) Pipette 950 µl of room temperature SOC into the mixture.

9) Place at 37 Celsius for 60 minutes. Shake vigorously (250 rpm) or rotate.

10) Warm selection plates to 37 Celsius.

11) Mix the cells thoroughly by flicking the tube and inverting, then perform several 10-fold serial dilutions in SOC.

12) Spread 50-100 µl of each dilution onto a selection plate and incubate overnight at 37 Celsius.

13) Add 100 µl of original dilution mixture to 900 µl LB with Ampicillin, incubate overnight.

Note: Competent cells with plasmid successfully transformed can be survived in LB with Ampicillin added because there is ampicillin site in plasmid. This medium with living cells can be used for further plating.

Materials

Medium	Composition	Remarks
SOC	2% Vegetable peptone (or Tryptone) 0.5% Yease Extract 10mM NaCl 2.5mM KCl 10mM $MgCl_2$ 10mM $MgSO_4$ 20mM Glucose	No antibiotics added, carefully operate, work near the flame.

References

Adapted from 'Certificate of Analysis for NEB 5-alpha Competent E.coli (High Efficiency)', New England Biolabs

Protocol 13: Colony PCR (Dr. Masato Kumauchi)

This protocol is to search for the candidate colonies that can be incubate to extract plasmid for DNA sequencing.

Method

1) Prepare several 600 µl tubes, add 10 µl LB w/ Amp to each tube.

Use 0.5-10 µl pipette, pick single colonies on the plate, and transformed each colony to one tube with 10 µl LB w/ Amp by pipetting. Recommended 8-12 single colonies each plate if possible.

Incubate micro cultures with picked single colonies at 37 Celsius for 2-3 hour, until cultures have become turbid.

Prepare PCR mixture, see the table below (210 µl mixture for 15 tubes, scale up for more tubes)

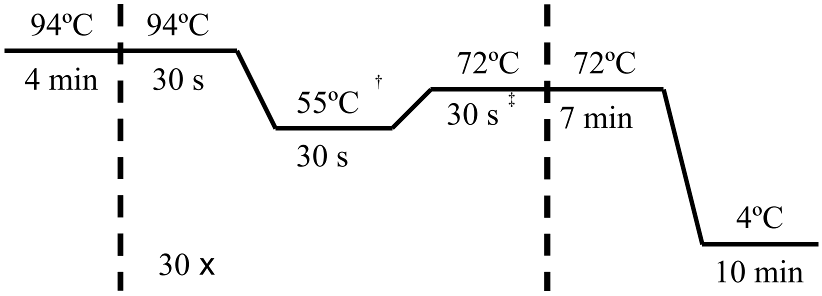
Component	amount (µl)	Remarks
Nuclease-free H_2O (autoclaved H2O)	*180.50*	
10× Standard *Taq* Reaction Buffer	22.50	
FW Flanking Primer	2.00	
RV Flanking Primer	2.00	
dNTP (10 mM each NTP)	2.00	More than general protocol for large target
Taq DNA Polymerase[#]	1.00	

[#]*Note*: ***Taq DNA Polymerase*** *should be the last component added to minimize primer degradation. It is a temperature sensitive enzyme, keep in freezer or ice before use.*

Divide PCR mixture to several tubes, each tube should have 14 µl mixture.

Take 1 µl from white pellet located at the bottom of incubated micro culture, add to PCR tube with 14 µl PCR mixture as template.

Run a temperature cycling program on the reaction mix. For the temperature cycling parameters use the scheme and notes below as a guide.



Prepare 50 ml 2% Agarose in 1x TAE buffer gel ~20 minutes before the completion of PCR. (Similar to Step 1-4 of Protocol 6)

When PCR finished, add 3 µl Loading Dye (6x) to each PCR tube. Load samples and run gel electrophoresis and take UV image following Step 6-7 of Protocol 6.

If band shows around 400 bp, take 5 µl micro culture out of corresponding tubes, incubate in 10 ml LB w/ Amp culture overnight (do not extent to more than 16 hours!).

Next morning, following Protocol 5 to extract plasmid, submit 15-20 µl plasmid solution to Core Facility for DNA sequencing.

> *Note: Core facility run DNA sequencing twice a week on Tuesday and Thursday. Deadline for submitting samples are 5pm on Monday and Wednesday. In urgent cases, you may submit samples by 8am on Tuesday and Thursday Morning.*

References

Adapted from a protocol obtained from Dr. Masato Kumauchi

Protocol 14: Colony PCR (Dr. Junpeng Deng)

This protocol is to search for the candidate colonies that can be incubate to extract plasmid for DNA sequencing.

Method

1) Prepare several Eppendorf 1.5ml centrifuge tube, add 1ml LB w/ antibiotics to each tube.

2) Use sterilized tooth stick, pick the entire single colony, then drop stick tip to each tube. Pick 12~24 colonies from each plate.

3) Incubate 1ml culture w/ colony at 37 Celsius for 2 hrs, shake at 250 rpm.

4) Prepare PCR mixture, see the table below (210 µl mixture for 15 tubes, scale up for more tubes)

Component	amount (µl)	Remarks
Nuclease-free H_2O (autoclaved H2O)	*180.50*	
10× Standard *Taq* Reaction Buffer	22.50	
FW Flanking Primer	2.00	
RV Flanking Primer	2.00	
dNTP (10 mM each NTP)	2.00	More than general protocol for large target
Taq DNA Polymerase[#]	1.00	

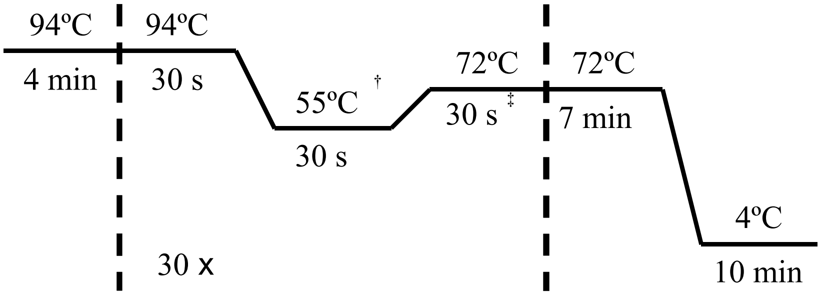
[#]*Note*: ***Taq DNA Polymerase*** *should be the last component added to minimize primer degradation. It is a temperature sensitive enzyme, keep in freezer or ice before use.*

5) Divide PCR mixture to several tubes, each tube should have 14 µl mixture.

6) Take 10 µl growth cell culture, then mix with 15 µl H_2O, boil at 95 Celsius for 15 min.

7) Spin down at ~14,000g for 10 min, then take 1 µl supernatant as template, add to 14 µl Colony PCR mixture.

8) Run a temperature cycling program on the reaction mix. For the temperature cycling parameters use the scheme and notes below as a guide.



9) Prepare 50 ml 2% Agarose in 1x TAE buffer gel ~20 minutes before the completion of PCR. (Similar to Step 1-4 of Protocol 6)

10) When PCR finished, add 3 µl Loading Dye (6x) to each PCR tube. Load samples and run gel electrophoresis and take UV image following Step 6-7 of Protocol 6.

11) If band shows around 400 bp, take 5 µl micro culture out of corresponding tubes, incubate in 10 ml LB w/ Amp culture overnight (do not extent to more than 16 hours!).

12) Next morning, following Protocol 5 to extract plasmid, submit 15-20 µl plasmid solution to Core Facility for DNA sequencing.

Note: Core facility run DNA sequencing twice a week on Tuesday and Thursday. Deadline for submitting samples are 5pm on Monday and Wednesday. In urgent cases, you may submit samples by 8am on Tuesday and Thursday Morning.

References

Adapted from a protocol obtained from Dr. Junpeng Deng

Appendix B.2

Preparation of Electro-competent *E. Coli* Cells

What is electro-poration?

“Poration”---to form pores. “Electro-poration”---to generate pores in the cell membranes using electric pulses.

In molecular biology, electroporation is used to open pores through the cell membranes using a strong pulsed electric field so that the plasmids can enter the cells.

If the E-field is too weak, the cell membranes will remain intact (no transformation). If the E-field is too strong, the cell membranes will be destroyed irreversibly and thus cells die. *So it is important to apply the E-field pulse with the right strength and the right duration!*

Three critical considerations:

(1) Keep cells in excellent health so that they can endure and survive the electric shock
(2) Remove ions from cell culture to prevent sparks during electric shock (electroporation)
(3) Store the electro-competent cells safely to keep their vitality
(4) Since no anti-biotics is used, handle bacterial growth and transfer with care to prevent contamination

Good References:

(1) Adapted from ‘Bacterial Electro-transformation and Pulse Controller Instruction Manual’ from BIO-RAD (Catalog Number 165-2098; Version 2-89).
(2) http://en.wikipedia.org/wiki/Electroporation (see p. 3)

Getting Started:

Read the protocol carefully. Find all the materials and equipment you need. Make a reservation of the centrifuge and the centrifuge rotor.

Preparation of Electro-competent Cells

Check list 1 (harvesting)

(1) Reserve the centrifuge (Sorvall) for harvesting the cells (Dr. Burnap)
(2) Reserve the rotor (Sorvall SLC-6000) (Dr. Hoff) (1 L bottles)
(3) Reserve the SS-34 rotor (50 ml) (Dr. Hoff)

Check List 2 (autoclave)

Prepare and auto-clave the following items:

(1) 20ml LB culture (no anti-biotics) in 200 ml glass flask with a cotton plug for overnight growth
(2) 1L LB culture (no anti-biotics) in 2.8L flask
(3) 2L nanopure water in 2.8 L flask
(4) 40 ml of 10% (v/v) glycerol (90% DDW) in 50 ml glass bottle
(5) 4 1L centrifuge bottles (to be autoclaved together with liquids)
(6) 1 50 ml centrifuge tube

Check if there are 50 of the autoclaved 1.5ml eppendorf centrifuge tubes with o-ring cap.

Check List 3 (pre-chill)

(1) Pre-chill two trays of water for cooling the cell culture before harvesting
(2) Pre-chill 2L nanopure water

(3) Pre-chill 50 of 1.5 ml autoclaved tubes
(4) Pre-chill 4 of 1L harvesting bottles
(5) Pre-chill the centrifuge
(6) Pre-chill the two rotors (for 1L and 50 ml tubes)

***E. coli* cells for transformation will be used to prepare electro-competent cells.**

(1) Inoculate *20 ml* ***LB medium*** (no antibiotics) in 200 ml flask with the *E. coli* competent cells. Grow overnight at 37°C (~12 hrs).

(2) Inoculate 10 ml over-night culture to 1 liter **LB medium** (no antibiotics) and grow the culture at **37°C** with vigorous shaking at 220 rpm.

(3) Monitor the OD_{600} of the culture until it reaches 0.35 to 0.45 OD. (The best electroporation results are obtained with cells that grow vigorously).

(4) Chill the cells rapidly (in 10 to 20 min) to preserve their vitality. Bring the cell culture to the cold room, chill in icy water (plastic tray) for 5 minutes, then move the second tray, and chill another 5 minutes).

(5) Distribute 1L chilled cell culture to 4 centrifuge bottles (4ºC) (250ml culture per bottle), centrifuge them culture at 4200 rpm (3900-g) for 15 minutes using a Sorval Evolution RC refrigerated centrifuge at 4°C and a pre-chilled Sorvall® SLC-6000 rotor.

Note 1: Collect all supernatants in a glass bottle and sterilize before discarding it down the drain! this also goes for the supernatants in subsequent centrifuge steps.

Since this step, the cells should be on ice-water mix all the time in order to keep the activity of competent cells.

(6) Resuspend pellets in a total of 1 liter (250ml for each bottle) of cold sterile NANOpure water (4ºC). Centrifuge as in step 4.

(7) Resuspend pellets in a total of 0.5 liter (125ml for each bottle) of cold sterile NANOpure water (4ºC). Centrifuge as in step 4.

(8) Resuspend pellets in a total of 36 ml (8 ml for each bottle) of cold sterile 10% (v/v) Glycerol. Transfer all the cell suspension to a 50 ml antoclaved centrifuge tube. Centrifuge in SS-34 rotor for 15 minutes at ?? rpm (3900 g).

(9) Resuspend pellet with cold sterile 10% (v/v) Glycerol to a final volume of 2-3 ml. The cell concentration should be about 1-3 x10^{10} cells/ml.

(10) Distribute the cells in 50 µl aliquots over 1.5 ml eppendorf tubes, freeze each one immediately in liquid nitrogen. Store them at –80°C (Donghua Zhou's lab). Cells should be good for at least 6 months under these conditions.

Materials

Medium	Composition	Remarks
LB medium	10 g·l^{-1} (Bacto-)Trypton 5 g·l^{-1} Yeast extract 5 g·l^{-1} NaCl (or 10 g·l^{-1} NaCl)	

Reference 2

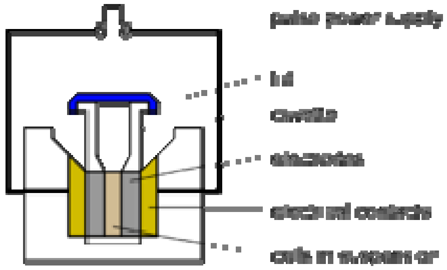


Diagram of the major components of an electroporator with cuvette loaded.

Dr.Eberhard Neumann - Electroporation Founder

Electroporation, or **electropermeabilization**, is a significant increase in the electrical conductivity and permeability of the cell plasma membrane caused by an externally applied electrical field. It is usually used in molecular biology as a way of introducing some substance into a cell, such as loading it with a molecular probe, a drug that can change the cell's function, or a piece of coding DNA.[1]

Electroporation is a dynamic phenomenon that depends on the local transmembrane voltage at each point on the cell membrane. It is generally accepted that for a given pulse duration and shape, a specific transmembrane voltage threshold exists for the manifestation of the electroporation phenomenon (from 0.5 V to 1 V). This leads to the definition of an electric field magnitude threshold for electroporation (E_{th}). That is, only the cells within areas where $E \geqq E_{th}$ are electroporated. If a second threshold (E_{ir}) is reached or surpassed, electroporation will compromise the viability of the cells, i.e., irreversible electroporation.[2]

In molecular biology, the process of electroporation is often used for the *transformation* of bacteria, yeast, and plant protoplasts. In addition to the lipid membranes, bacteria also have cell walls which are different from the lipid membranes and are made of peptidoglycan and its derivatives. However, the walls are naturally porous and only act as stiff shells that protect bacteria from severe environmental impacts. If bacteria and plasmids are mixed together, the plasmids can be transferred into the cell after electroporation. Several hundred volts across a distance of several millimeters are typically used in this process. Afterwards, the cells have to be handled carefully until they have had a chance to divide producing new cells that contain reproduced plasmids. This process is approximately ten times as effective as *chemical transformation*.[1][3]

This procedure is also highly efficient for the introduction of foreign genes in tissue culture cells, especially mammalian cells. For example, it is used in the process of producing knockout mice, as well as in tumor treatment, gene therapy, and cell-based therapy. The process of introducing foreign DNAs into eukaryotic cells is known as transfection.

Appendix C

Unix/Linux Script for CCP4 and HBPlus

Zhouyang Kang

Script 1a: Separate multiple structural conformations in PDB

```
#!/bin/bash


#It seems the native pdb files won't run properly through areaimol.
#This script will take each pdb file in the folder it is located in and write just the lines
#that contain CRYST1, SCALE, and ATOM to a new file that has the pdb name with .trunc added.
#It will then make a folder called pdb_trunc_files and move the new pdb.trunc files into it.
#Run this script in the pdb folder that contains your pdb files.


mkdir ./pdb_trunc_files


for file in *.pdb
do
    grep '^HEADER' $file >> $file.trunc
    grep '^TITLE' $file >> $file.trunc
        grep '^CRYST1' $file >> $file.trunc
        grep '^SCALE' $file >> $file.trunc
        grep '^ATOM' $file >> $file.trunc
        grep '^HETATM' $file >> $file.trunc
        sed "s/\([ABCDX]\)\([0-9]\{4\}\)/\1 \2/g" $file.trunc >> $file.trunca
        rm -f $file.trunc
        mv $file.trunca $file.trunc


# extract the PDB ID from *.pdb
PDB_ID=`sed -n '1p' $file | awk '{print $NF}'`
echo "$PDB_ID " >> pdb_truncator_log
```

```
# output the AASP, BASP and ASP to three files which will be used in step 4.
# identify if there are A,B conformers. And split the pdb file into two A_* and B_*


# If there are 'Axxx A' (where xxx is AA's name, this is to exclude 'ATOM' character) then generate a B
file and do the following if clause. (the original file will be modified into A file)
        grep "A... [ABCDX]" $file.trunc && cat $file.trunc >> B_$file.trunc


        if (($? == 0))
        then


                # delete the Bxxx line
                sed -e'/[BC]... [ABCDX]/{d}'  $file.trunc > tmp_A
                # change Axxx into xxx
                sed  's/A\(...\) \([ABCDX]\)/ \1 \2/g' tmp_A > A_$file.trunc

                # delete the Axxx line
                sed -e'/A... [ABCDX]/{d}'  B_$file.trunc > tmp_B
                # change Bxxx into xxx
                sed  's/[BC]\(...\) \([ABCDX]\)/ \1 \2/g' tmp_B > B_$file.trunc



                rm -f tmp_A
                rm -f tmp_B
                rm -f $file.trunc



        fi

# change done to the HOHs
# Is B_$file.trunc exist?
```

```
if [ -f B_$file.trunc ]
then
    grep "AHOH" $file.trunc

    if (($? == 0))
    then

        # delete the BHOH line
        sed -e'/BHOH/{d}'  A_$file.trunc > tmp_A
        # change AHOH into HOH
    sed -e 's/AHOH/ HOH/g' tmp_A > A_$file.trunc



        # delete the AHOH line
        sed -e'/AHOH/{d}'  B_$file.trunc > tmp_B
        # change BHOH into HOH
        sed -e 's/BHOH/ HOH/g' tmp_B > B_$file.trunc

        rm -f tmp_A
        rm -f tmp_B
        rm -f $file.trunc

    fi

else
    grep "AHOH" $file.trunc && cat $file.trunc >> B_$file.trunc

    if (($? == 0))
    then
```

```
            # delete the BHOH line
            sed -e'/BHOH/{d}'  $file.trunc > tmp_A
            # change AHOH into HOH
        sed -e 's/AHOH/ HOH/g' tmp_A > A_$file.trunc


            # delete the AHOH line
            sed -e'/AHOH/{d}'  B_$file.trunc > tmp_B
            # change BHOH into HOH
            sed -e 's/BHOH/ HOH/g' tmp_B > B_$file.trunc

            rm -f tmp_A
            rm -f tmp_B
        fi

    fi


done

mv ./*.pdb.trunc ./pdb_trunc_files/
mv ./*.ent.trunc ./pdb_trunc_files/

echo

exit 0
```

Script 1b: Submission script for script 1a

```
#!/bin/bash
#PBS -V
#	pass environment variables to job

#PBS -q express

#PBS -N areaimol_script
#	name you want to give your job on the queue (change to what you want)

#PBS -l nodes=1:ppn=1
#	request 1 nodes w/1 processors per node (please use 1 ppn unless you have a good reason not to.)

#PBS -l walltime=1:00:00
#	give the walltime your code will need.  your job will be killed if it goes over.
#	Your job will start sooner, the shorter the walltime.
#PBS -j oe
#	join the output and error files into one file

cd $PBS_O_WORKDIR
#	change to current working directory
$HOME/pdb_scripts/pdb_truncator.scpt
date
```

Script 2a: Run Areaimol of CCP4 on UNIX

```
#!/bin/sh


#
# From " Example scripts for areaimol using toxd data"
#


# used to distinguish different runs in html logfile
CCP_PROGRAM_ID=run1
export CCP_PROGRAM_ID
####################################################
#
# Simple area calculation (verbose output)
#
####################################################
# This script will loop through and read the pdb.trunc files in the folder it is in and
# run them through areaimol. This creates a .brk file for each pdb file.
#It then makes a folder called brk_files and puts the new pdb.trunc.brk files in it.


for file in *.trunc
do
areaimol XYZIN $file \
        XYZOUT $file.brk <<eof-area
VERB      ! Verbose output
OUTPUT    ! Output pseudo-pdb file
END
eof-area
done
mkdir ./brk_files
mv ./*.brk ./brk_files/
```

Script 2b: Submission script for script 2a

```
#!/bin/bash

#PBS -V
#       pass environment variables to job

#PBS -q express

#PBS -N areaimol_script
#       name you want to give your job on the queue (change to what you want)

#PBS -l nodes=1:ppn=1
#       request 1 nodes w/1 processors per node (please use 1 ppn unless you have a good reason not to.)

#PBS -l walltime=1:00:00
#       give the walltime your code will need.  your job will be killed if it goes over.
#       Your job will start sooner, the shorter the walltime.
#PBS -j oe
#       join the output and error files into one file

source /opt/ccp4/6.4.0/gcc/bin/ccp4.setup-sh

cd $PBS_O_WORKDIR
$HOME/pdb_scripts/areaimol.scpt
Date
```

Script 3: Analyze areaimol output for collecting Asp and Glu data only

```
#!/bin/sh
#
# tagger1  -  Script to tag (via custom suffix) PAIRS of lines in modified
#             PDB file, based on numerical values in column #11.
#
#  ver 0.1  2014-02-12  rjh
#


for file in *.brk
do
       grep '\<OD.*ASP\>' ./$file | sed 's/ \([ABCDX]\)\([0-9]\)/ \1 \2/g' >> $file.DE
            grep '\<OE.*GLU\>' ./$file | sed 's/ \([ABCDX]\)\([0-9]\)/ \1 \2/g' >> $file.DE
done

for trunc_file in *.pdb.trunc *.ent.trunc
do

            echo "Processing ${trunc_file}.. "

# extract the PDB ID from *.pdb.trunc
PDB_ID=`sed -n '1p' $trunc_file | awk '{print $NF}'`
echo "PDB ID is $PDB_ID "

# replace the ATOM with the PDB_ID in the *.pdb.trunc.brk.DE
file=${trunc_file}.brk.DE
sed -i.bak "s/ATOM/${PDB_ID}/" $file

done
```

```
for file2 in *.DE

do tagger ./$file2 >> $file2.dat

done

cat *.dat >finalDE.dat
```

Sub-script of Script 3: Analyze areaimol output for collecting Asp and Glu data only

```
#!/bin/sh
#
# tagger1  -  Script to tag (via custom suffix) PAIRS of lines in modified
#             PDB file, based on numerical values in column #11.
#
#  ver 0.1  2014-02-12  rjh
#

cat "$@" | awk ' BEGIN {

        str1 = "1 0 0"
        str2 = "0 1 0"
        str3 = "0 0 1"
}

{
        line1=$0
        coo1=$11
        getline
        line2=$0
        coo2=$11

        if (coo1+coo2<8) {

          strx=str1
          printf "%s %s  %.1f\n", line1,strx,coo1+coo2
          print line2

        } else {
          if (coo1>15 && coo2>15) {

          strx=str2
          printf "%s %s  %.1f\n", line1,strx,coo1+coo2
          print line2
          }
          else {

          strx=str3
          printf "%s %s  %.1f\n", line1,strx,coo1+coo2
          print line2
          }
        }

}'
```

Script 4: Modify original PDB files for HBPlus

```
#!/bin/bash


for file in *.pdb
do
        grep '^HEADER' $file >> $file.trhb
        grep '^TITLE' $file >> $file.trhb
        grep '^CRYST1' $file >> $file.trhb
        grep '^SCALE' $file >> $file.trhb
        grep '^ATOM' $file >> $file.trhb
        grep '^HETATM' $file >> $file.trhb
        grep '^CONECT' $file >> $file.trhb



# If there are 'Axxx A' (where xxx is AA's name, this is to exclude 'ATOM' character) then generate a B file and do the following if clause. (the original file will be modified into A file)
        grep "A... [ABCDX]" $file.trhb && cat $file.trhb >> B_$file.trhb


        if (($? == 0))
        then


                # delete the Bxxx line
                sed -e'/[BCD]... [ABCDX]/{d}'  $file.trhb > tmp_A
                # change Axxx into xxx
                sed  's/A\(...\) \([ABCDX]\)/ \1 \2/g' tmp_A > A_$file.trhb


                # delete the Axxx line
                sed -e'/A... [ABCDX]/{d}'  B_$file.trhb > tmp_B
                # change Bxxx into xxx
                sed  's/[BCD]\(...\) \([ABCDX]\)/ \1 \2/g' tmp_B > B_$file.trhb
                rm -f tmp_A
                rm -f tmp_B
                rm -f $file.trhb

        fi

# change done to the HOHs
# Is B_$file.trhb exist?
        if [ -f B_$file.trhb ]
        then
                grep "AHOH" $file.trhb

                if (($? == 0))
                then

                        # delete the BHOH line
                        sed -e'/BHOH/{d}'  A_$file.trhb > tmp_A
                        # change AHOH into HOH
                sed -e 's/AHOH/ HOH/g' tmp_A > A_$file.trhb

                        # delete the AHOH line
                        sed -e'/AHOH/{d}'  B_$file.trhb > tmp_B
                        # change BHOH into HOH
                        sed -e 's/BHOH/ HOH/g' tmp_B > B_$file.trhb

                        rm -f tmp_A
                        rm -f tmp_B
                        rm -f $file.trhb
                fi

        else
                grep "AHOH" $file.trhb && cat $file.trhb >> B_$file.trhb

                if (($? == 0))
                then

                        # delete the BHOH line
                        sed -e'/BHOH/{d}'  $file.trhb > tmp_A
                        # change AHOH into HOH
                sed -e 's/AHOH/ HOH/g' tmp_A > A_$file.trhb

                        # delete the AHOH line
                        sed -e'/AHOH/{d}'  B_$file.trhb > tmp_B
                        # change BHOH into HOH
                        sed -e 's/BHOH/ HOH/g' tmp_B > B_$file.trhb

                        rm -f tmp_A
                        rm -f tmp_B
                fi

        fi

done

echo

exit 0
```

Script 5: Run HBPlus and Collect Hydrogen-bond information

```
#!/bin/bash

mkdir All_HBond
mkdir Asp_Glu_HBond
mkdir Asp_HBond
mkdir Glu_HBond
mkdir trhb_files
mkdir clean_files
mkdir alt_files

for file0 in *.trhb

do echo $file0 | clean

done

for file1 in *.new

do

file2=`echo $file1 | sed "s/new/trhb/g"`

hbplus $file1 $file2 -d 3.2

done

for file3 in *.hb2
do
        grep '\<ASP.OD*.\>' ./$file3 >> $file3.hbDE
        grep '\<GLU.OE*.\>' ./$file3 >> $file3.hbDE
        grep '\<ASP.OD*.\>' ./$file3 >> $file3.hbD
        grep '\<GLU.OE*.\>' ./$file3 >> $file3.hbE

done

mv *.hb2 ./All_HBond
mv *.hbDE ./Asp_Glu_HBond
mv *.hbD ./Asp_HBond
mv *.hbE ./Glu_HBond
cp *.trhb ./All_HBond
cp *.trhb ./Asp_Glu_HBond
cp *.trhb ./Asp_HBond
cp *.trhb ./Glu_HBond

mv *.trhb ./trhb_files
mv *.new ./clean_files
mv *.alt ./alt_files

echo
```

Script 5: Convert hydrogen-bond information to correct format for statistical analysis

```
#!/bin/sh
#
# sort hb2 (HBPlus output) files and add PDB ID to the first column
# Zhouyang Kang v1.0 2014/02/18

mkdir bak_files
mkdir trhb_copy
mkdir HB_D_A
mkdir HB_D_B
mkdir HB_D_S
mkdir hbD_files

for file in *.hbD
do

sort $file -k 3  >> $file.sort

done

for trhb_file in *.pdb.trhb *.ent.trhb

do

echo "Processing ${trhb_file}.. "

# extract the PDB ID from *.pdb.trhb & *.ent.trhb
PDB_ID=`sed -n '1p' $trhb_file | awk '{print $NF}'`
echo "PDB ID is $PDB_ID "
```

```
hb_file=`echo $trhb_file | sed "s/trhb/hb2/g"`

# Add PDB ID to the first column of *.hbD

file2=${hb_file}.hbD.sort
sed -i.bak "s/\([A-Z]*\)/${PDB_ID} \1/" $file2



done

echo "Organizing files..."

mv *.bak ./bak_files
mv *.trhb ./trhb_copy
mv A*.sort ./HB_D_A
mv B*.sort ./HB_D_B
mv *.sort ./HB_D_S
mv *.hbD ./hbD_files

echo "Finished!"
```

VITA

Zhouyang Kang

Candidate for the Degree of

Doctor of Philosophy

Thesis: DEVELOPMENT AND APPLICATIONS OF INFRARED STRUCTURAL BIOLOGY

Major Field: PHYSICS

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Physics at Oklahoma State University, Stillwater, Oklahoma in May, 2014

Completed the requirements for the Bachelor of Science in Materials Physics at Nanjing University, Nanjing, China in 2007.

Experience:

Teaching Assistant in Department of Physics, Oklahoma State University, August, 2007 – May, 2014
Research Assistant in Department of Physics, Oklahoma State University, August, 2007 – May, 2014

Awards:

Outstanding Graduate Research Assistant, Department of Physics, Oklahoma State University, April, 2014

Professional Memberships:

American Physical Society since 2009