Mechanistic Studies on Chemical Instabilities of Recombinant Proteins

by

Bin Pan

B.Eng. in Chemical Engineering
Tsinghua University, 2003

M.S. in Chemical Engineering Practice
Massachusetts Institute of Technology, 2006

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

June 2009

Signature of Author: _____
Department of Chemical Engineering
May 10, 2009

Certified by: _____
Daniel I. C. Wang
Institute Professor of Chemical Engineering
Thesis Supervisor

Certified by: _____
Bernhardt L. Trout
Professor of Chemical Engineering
Thesis Supervisor

Accepted by: _____
William M. Deen
Professor of Chemical Engineering
Chairman, Committee for Graduate Students

Abstract

Mechanistic Studies on Chemical Instabilities of Recombinant Proteins

by

Bin Pan

Submitted to the Department of Chemical Engineering on
May 14, 2009 in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Chemical Engineering

ABSTRACT

Protein molecules are being widely used as pharmaceuticals for treating diseases ranging from diabetes and haemophilia to various types of cancers due to their great potency and specificity. However, these macromolecules are intrinsically unstable in aqueous solutions, due to the existence of various physical and chemical degradation pathways. Degraded protein molecules have much reduced biological functions, and may also have adverse effects such as immunogenicity or pharmacokinetic issues. Thus, understanding the underlying mechanisms of these degradation pathways is essential for rationally devising better ways to stabilize protein pharmaceuticals and extends their applicability. In this thesis, two important types of chemical degradation pathways, the oxidation of methionine residues and the hydrolysis of peptide bonds in monoclonal antibody molecules, are investigated from a mechanistic point of view.

In the first half of the thesis, oxidation of methionine residues in a model protein G-CSF (Granulocyte-Colony Stimulating Factor) was studied to address the issue of how protein structure affects its reactivity. Comparative oxidation studies were performed where the kinetics of oxidation of methionine residues by hydrogen peroxide (H_2O_2) in G-CSF and corresponding chemically synthesized peptides thereof were measured at different temperatures. To assess structural effects, equilibrium denaturation experiments also were conducted on G-CSF to obtain the free energy of unfolding as a function of temperature. A comparison of the relative rates of oxidation of methionine residues in short peptides with those of corresponding methionine residues in rhG-CSF yields an understanding of how protein tertiary structure affects oxidation reactions. For the temperature range studied, 4°C to 45°C, the oxidation rate constants followed an Arrhenius equation quite well, suggesting the lack of temperature-induced local structural perturbations that affect chemical degradation rates. One out of the four methionine residues, Met122, showed an activation energy significantly different than that of the corresponding peptide. Extrapolation of kinetic data predicts non-Arrhenius behavior around the melting temperature. Phenomenological modeling trying to understand the temperature dependence of rate constants was pursued. Finally, we show that the data obtained from accelerated oxidation can be used in conjunction with our models to get predictions about the long-term shelf-life oxidation comparable with experimental results.

In the latter half of this thesis, three approaches in a hierarchical order were taken in order to explain the higher rate of un-catalyzed hydrolysis of peptide bonds only in the hinge region of antibody molecules. First, *ab initio* molecular dynamic simulations were performed to understand the reaction mechanism of the hydrolysis of peptide bonds. The system solvated in explicit water molecules was modeled quantum mechanically and

dynamic transition trajectories of the chemical reaction were computed at ambient conditions. Since no unique pathway can be used to describe the reaction process due to fluctuations at finite temperature, path sampling technique was applied to obtain an ensemble of trajectories. A statistical tool, likelihood maximization, was used to extract physically important degrees of freedom by screening a large number of reaction coordinate models. The same approach was applied to the hydrolytic reactions under both acidic and neutral pH conditions, which are the most relevant to the formulation of antibody molecules. In both cases, changes in local bonding pattern close to the reaction center, as well as the solvent network, showed importance in determining the reaction dynamics of the hydrolysis of the peptide bond.

Then classical molecular dynamic simulations were performed to study the dynamics of a free hinge fragment and the hinge fragments in the antibody molecule. Important structural and dynamic differences between the two situations were revealed, especially the observation that the free hinge fragment takes on configurations much less frequently accessed by the hinge fragment when situated inside the antibody molecule.

In the third approach, a coarse-grained reaction rate model was proposed in order to explain the experimentally observed higher rate of hydrolysis of peptide bonds. A hypothesis involving a mechano-chemical mechanism was motivated by the essential constraining effect of Fab and Fc domains on the hinge region in the antibody molecule revealed in the second approach. Combining the information obtained from the previous two approaches, force was calculated along the reaction coordinate direction that was determined and verified previously. This information was integrated into a reaction rate model in order to compute the reaction rate constants. The computational results show that the mechano-chemical mechanism can yield reasonable rate constants comparable with available experimental data.

Thesis Supervisors:

Bernhardt L. Trout	Professor of Chemical Engineering
Daniel I. C. Wang	Institute Professor of Chemical Engineering

Acknowledgements

I feel deeply sorrowful that my father in not living today to see his son's achievements including the conclusion of his graduate study. It is still heart-breaking for me because my father never had an easy moment is his entire life but instead day-in-and-day-out labor work, and finally the torture and enormous pain caused by cancer in his last days. He did not even complain with a single word, but lived with a smile on his face, because of his satisfaction of the excellence my brother and I at school. My father and mother, through their hard work and struggle for a better life and through their hope for a brighter future for their next generation, have shown to me how persistent and persevering a person can be. After my father's passing, my mother continues to exemplify how a great mother can love, care, and sacrifice for her two sons dispite the huge loneliness that I imagine she often feels being so far away from both of them. These lessons from my parents will encourage me to aim high and guide me through whatever kind of difficulty and hardship I may encounter in the journey of life.

All my teachers, even from elementary school, have shown heart-warming care and selfless transfer of knowledge to this child from a poor family. I am indebted to all of them for not only their education in classroom, but also extra help in deepening my learning. I will remember their names all my life. It was through their love, only partly combined with my efforts, that I could go to Beijing for a college education.

At Tsinghua University, I met not only the brightest classmates I had ever met, but also caring teachers and professors. Professor Liu Zheng, an alumnus from the same local provincial high-school, showed me how one should diligently do research. He is a great example of continued striving for excellence and perfection. I thank him for providing me with the opportunity to learn in his lab and for guiding me and advising me about my undergraduate project. Professor Luo Guangsheng, from the first day even before I knew him, provided me with financial support after learning of my family background. I highly appreciate his care, and all the advice he gave during my studies and my visits to him. Professors Xing Xin-hui and He Xiaorong were very inspiring when they allowed me to work on some research projects in their labs.

While being able to enter Tsinghua University was a dream, being able to come to MIT was even beyond what I could have dreamed about. As my graduate study nears completion and as I look back over my years in graduate school, I find that I am so fortunate to be in such a great place, having learned so much and having interacted with so many great people who have changed my life in exciting and profound ways.

First and foremost, I would like to deeply express my gratitude to my thesis advisors, Professors Bernhardt L. Trout and Danial I.C. Wang. Professor Trout has provided great encouragement and guidance whenever any difficulty in research or life arose. He has shown not only his academic sharpness in identifying the correct reseach problems to work on, but also the patience and continuing high expectation for his student, especially a slow starter like myself. In his research group, I enjoyed a comfortable environment with full professional support. His encouragment has helped me to overcome the difficult obstacles throughout my research work. Besides being a great mentor scientifically, he is a truly great friend in life. I especially remember the time he

spent with me on a moutain-trail during our trip to Colorado for the Protein Stability Conference. The opportunities he gave for me to help recruit for the Singapore-MIT Alliance were also much appreciated.

Professor Wang is genuinely the most diligent and dedicated researcher I have ever met at MIT. He has the insightful vision for the future of biotechnology, and has an unmatched sharp and bright mind in guiding students' research work. He has been revered by many my fellow graduate students, myself in particular when I was a junior graduate student. However, as we had more interaction, his kindness and sense of humor impressed me a lot. Many thanks to Suzan Lanza for several birthday celebrations on which accasions I learned more about Professor Wang's life and the respect all of his students have for him. I just hope that he can fully recover very soon and come back to us as often as he used to.

My thesis committee members deserve sepcial acknowledgement for their contribution to this thesis work. Dr. Margaret Ricci has been remarkably helpful throughout our collaboration. No matter how busy her schedule was, she gave high priority to my thesis work and meetings. When I visitied Amgen, even weekend, I had to trouble her to come to lab to open the door for me even as she had to take care of her newborn. Her constructive suggestions and advices have added great value to my thesis. Professor Amy Keating has been always very helpful and I enjoyed every meeting at which she was present. She not only proposed alternative ways of thinking about my scientific problems and made other very valuable suggestions about my research work, but also gave very detailed advice on how to communicate more effectively with my advisors and thesis committee. Professor Charles Cooney has always be so nice and his knowledge and experience have provided very positive and valuable feedback about my work. I also greatly enjoyed serving as a teaching assistant for the Downstream Processing summer course he hosted.

Other professors at MIT, including Professors Howard Brenner, William H. Green, Jeff W. Tester, William M. Deen, Dane Wittrup, and George Stephanopoulos have all given me advice and encouragement during these years and they are thankfully acknowledged. It is certainly a privilege for me to have attended MIT, not only to listen to great lectures given by these professors, who are experts in their respective fields, but also to interact with them as friends do.

I was lucky to be able to visit Amgen twice at their Thousand Oaks, California, headquarters. Many people helped me with the experiments there. Jeff Abel deserves special thanks not only for his help, but also for his hospitality and friendship. The parties at his house with his family and the trip to Hollywood were so enjoyable and unforgettable. Dr. David Brems provided the opportunity for me to visit Amgen, and he has was hosptitable and caring during my stay there. Tom Dillon, Dr. Pavel Bondarenko and many others provided great technical help and made my life there much easier.

A number of my friends at MIT and elsewhere deserve special mention. The Trout Group is a great place to learn and to work. Everyone in the lab, both past and present, is acknowledged. I especially want to say thank-you to Drs. Jhih-Wei Chu, Brian Baynes, Baron Peters, Erik Santiso, Gregg Beckham, and Naresh Chennamsetty. Dr. Chu was the one from whom I continued the protein stabilization work. His prior work set a concrete example of how great reserch work can be done. From his computer codes I quicked got aquainted with some of the advanced simulation techniques. In addition, Dr.

Chu is a great listener and provider of advice and encouragment. Dr. Brian Baynes is the one of brightest people I ever met. He is always humorous, and is a very skilled presenter. His ability to make complex concepts simpler and to explain precisely and concisely difficult content to others are enviable. Dr. Baron Peters has admirably deep knowledge and the great mastery of what he has learned. His work always looks so beautiful to me, and he has been always happy to share his findings and knowledge. I learned a lot from discussions with him and was fortunate to be a close observer of the important contributions he made to the field. Dr. Santiso has the most exceptional programming skills I personally have seen, and I benefited so much from learning from him and the disccussions we had together. Dr. Beckham is really a great friend, a good listener, and a good helper with the work presented in this thesis. Dr. Chennamsetty has been very nice to be so helpful. Our trip to the Colorado Protein Stability Conference was so enjoyable and adventurous; it is one of happiest moments in my grdaute years. Nicholas Musolino has been an enthusiastic and willing-to-help friend and his help with manuscript corrections, discussions and other favors are kindly acknowledged.

Outside the Trout Group, I also have numerous friends who made my life at MIT so colorful and wonderful. Minglin Ma has been a long-time great friend since we attended Tsinghua together. It is fortunate to have such a close friend with whom I can share views about work and life, all the way from college to graduate school. Brian Mickus has been such a wonderful friend who really cares and helps. Gwen Wilcox is truly generous and hospitable, and I thank her for all the enjoyable dinners and parties we had at her house, when holidays would have otherwise been sad for international students like myself. Angelique Scarpa has been so nice to withstand my too-frequent requrests about signatures, forms, and in particular the great preparation for my thesis defense; for all of this I am highly appreciative. Many other friends, such as Yin Bai, Sa Xiao, Yuanli Bai, Yang Wen, Hao Huang, Sheng Jing, Xiaogeng Song, Rob Ashcraft, Curt Fischer, Gregg Thurber, Dan Pregibon,..., are all remembered for their support, friendship and encouragement.

Lastly, my wife Jie has been so understanding and caring. I feel so fortunate to have found in her a wonderful companion in the journey of life. I deeply appreciate her contributions and her sacrifice since the time of our marriage, which I deeply hope I can repay, with love and care in our future years together. Her love has been indispensible to the completion of this thesis.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Overview

1.1.1 Protein pharmaceutics and its stability

Protein molecules are linear polymers comprised of 20 different amino acids linked by peptide bonds. They are the essential macromolecules in every living organism and participate in every process within cells. They play the most diverse and dynamic roles in the human body, ranging from catalyzing biochemical reactions, regulating intracellular and intercellular signaling, and recognizing and neutralizing foreign substances, to providing structural support as cellular scaffolds, etc. As important as their correct functioning inside cells is, their malfunctioning or absence could result in severe diseases and inviability.

With the development of recombinant DNA and protein technologies, it has been made possible to produce large quantities of protein molecules from relatively simple organisms such as *E. coli* or yeast at a low cost. After purification the protein molecules can be used as therapeutics, treating diseases ranging from diabetes and haemophilia to various types of cancers. Currently more than 130 different proteins or peptides are approved for clinical use by the FDA, and many more are in the developmental stage (*1*). These protein drugs include enzyme activators or inhibitors, functional regulators, monoclonal antibodies, and various vaccines. Protein pharmaceuticals are highly specific and effective in treating diseases, and their impact on the body and the immune system is usually much less than that of small-molecule drugs.

Despite their wide applicability and great promise for curing a range of diseases, protein molecules are intrinsically unstable due to their chemical and physical properties in the aqueous solutions in which these protein drugs are usually formulated. This unstable nature poses unique difficulties in virtually every step in the drug development process, including production, purification, separation, storage and delivery of these materials. This unstable nature is in stark contrast to the small-molecule drugs. The instability of protein molecules is due to various degradation pathways, which in general can be separated into two classes, pathways involving physical instability and those involving chemical instability (*2, 3*).

Chemical instability refers to any process of modifying protein molecules via chemical reactions to yield new chemical entities, such as the deamidation of the side-chain amide linkage in a Gln or Asn residue, or oxidation of the side-chain of a Met, Tyr, or His residue, etc. In general, these changes involve bond formation or cleavage.

On the other hand, physical instability involves major changes in its higher-order structure (secondary or tertiary or quaternary); it may or may not involve any chemical modifications. These include aggregation, denaturation, adsorption to surfaces, precipitation, etc.

To stabilize the protein therapeutics against both physical and chemical degradation pathways is an essential but challenging task in the development cycle of protein pharmaceuticals. The shelf-life of a protein therapeutics in solution depends not only on its own properties, such as the amino-acid sequence, tertiary structure, unfolding melting temperature, etc., but also on various physical and chemical properties of the formulation solution, such as temperature, pH, ionic strength, additives, etc. The

objective of protein stabilization is to design a particular formulation solution to minimize the physical and chemical degradations so that an acceptable shelf-life (normally on the scale of 2 years) can be achieved. However, it is a great challenge to design such formulations because of the complexity of the interaction between a protein and its aqueous environment and its intricate and unique tertiary structure. The current industrial practice is almost entirely empirical. Thus, a detailed and mechanistic understanding of every individual degradation pathway is needed for an accurate prediction of protein stability and the rational design of formulations (*4*).

In this thesis, two aspects of chemical instabilities, namely, the oxidative degradation of methionine residues and the hydrolytic degradation of peptide bonds in antibody molecules, are studied via a combination of experimental and computational/theoretical approaches. Our focus is to elucidate the underlying mechanisms of the chemical changes at the molecular level. The detailed knowledge of the underlying mechanisms derived from this study should be useful to rationalize the formulation design and may provide novel approaches to minimize the degradation of therapeutic proteins.

1.1.2 Protein oxidative degradation

The side-chains of amino acid residues such as methionine (Met), cysteine (Cys), tryptophan (Trp), tyrosine (Tyr) and histidine (His) can all be potential sites of oxidative degradation (*5*). Among these residues, the sulfur-containing Met residue is most liable to oxidation in aqueous solutions (*3, 5*). The cause of the oxidative degradation of methionine residues during storage involves active oxygen species, such as singlet oxygen $^1O_2$, superoxide radical $\cdot O_2^-$, peroxides ROOH, and hydroxyl radicals $\cdot$OH. These

reactive oxygen species (ROS) may be introduced into the formulation from impurities in the additives or be produced by light-catalyzed or metal-catalyzed reactions (*5, 6*). The oxidation reaction of Met forms a methionine sulfoxide, after the covalent addition of an oxygen atom to the sulfur atom in the side-chain of Met. This chemical modification changes the chemical properties, and in general results in loss of biological function.

1.1.3 Stability of antibody molecules and its hydrolytic degradation

Antibodies are a special class of protein molecules, having unique structures and physical and chemical properties. Engineered monoclonal human antibodies have showed their highly effective and specific potency in treating many types of diseases, such as cancer, infectious diseases, allergies, autoimmune diseases, and inflammations (*7*). However, the special tertiary structures of antibody molecules and their properties also make it rather difficult to stabilize these therapeutic antibodies.

In general, similar to other globular proteins, antibody molecules can have physical degradation pathways such as denaturation, aggregation, and surface adsorption, and chemical degradation pathways such as deamidation, disulfide formation/exchange, non-reducible cross-linking, isomerization, oxidation, C-terminal clipping, etc. (*7*).

The hydrolytic degradation of peptide bonds in antibody molecules has been reported recently by Cordoba et al. (*8*). The hydrolysis of peptide bond results in fragmented products, a single-armed Fab and Fc fragment, and a Fab fragment. This fact is somewhat mysterious in the sense that the hydrolytic cleavage is a non-enzymatic reaction which occurs only at several peptide bond locations in the hinge region. Experimental characterization of the hydrolyzed products and quantification of hydrolytic

rates from each of these sites are rather difficult and only limited data(*8, 9*) are available for quantitative analysis.

1.2 Thesis objectives and approach

The overall goal of this thesis is to better understand the underlying mechanisms of the oxidative degradation of methionine residues and the hydrolytic degradation of antibody molecules. Specifically, this thesis seeks to accomplish five goals:

- Understand the interplay between protein structure and its chemical reaction kinetics as a function of temperature by performing chemical kinetics measurements and biophysical techniques to characterize protein structures
- Reveal the special dynamic and structural features of the hinge fragment in antibody molecules by performing classical molecular dynamics studies on the free hinge fragment and the whole antibody molecule
- Elucidate the reaction mechanism of the hydrolytic reaction of peptide bonds under both neutral and acidic pH conditions by introducing the correct reaction coordinate
- Test a hypothesis about whether the solvent accessibility or the extent of local fluctuation correlates with the hydrolytic rate for residues in the hinge region of the antibody molecule
- Test an alternative hypothesis about whether the higher rate of hydrolysis can be explained by a mechano-chemical mechanism by a coarse-grained reaction rate model

1.3 References

(1) Leader, B., Baca, Q. J., and Golan, D. E. (2008) Protein therapeutics: A summary and pharmacological classification. *Nature Reviews Drug Discovery 7*, 21-39.

(2) Manning, M. C., Patel, K., and Borchardt, R. T. (1989) Stability of Protein Pharmaceuticals. *Pharmaceutical Research 6*, 903-918.

(3) Wei, W. (1999) Instability, stabilization, and formulation of liquid protein pharmaceuticals. *International Journal of Pharmaceutics 185*, 129-188.

(4) Meyer, J. D., Ho, B., and Manning, M. C. (2002) in *Rational design of stable protein formulations* (Carpenter, F. J., and Manning, C. M., Eds.) pp 85-107.

(5) Li, S. H., Schoneich, C., and Borchardt, R. T. (1995) Chemical Pathways of Peptide Degradation.8. Oxidation of Methionine in Small Model Peptides by Prooxidant Transition-Metal Ion Systems - Influence of Selective Scavengers for Reactive Oxygen Intermediates. *Pharmaceutical Research 12*, 348-355.

(6) Nguyen, T. H. (1994) Oxidation Degradation of Protein Pharmaceuticals, in *Formulation and Delivery of Proteins and Peptides* (Cleland, J. L., and Langer, R., Eds.) pp 59-71, American Chemical Society, Washington DC.

(7) Wang, W., Singh, S., Zeng, D. L., King, K., and Nema, S. (2007) Antibody structure, instability, and formulation. *Journal of Pharmaceutical Sciences 96*, 1-26.

(8) Cordoba, A. J., Shyong, B. J., Breen, D., and Harris, R. J. (2005) Non-enzymatic hinge region fragmentation of antibodies in solution. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences 818*, 115-121.

(9) Dillon, T. M., Bondarenko, P. V., Rehder, D. S., Pipes, G. D., Kleemann, G. R., and Ricci, M. S. (2006) Optimization of a reversed-phase high-performance liquid chromatography/mass spectrometry method for characterizing recombinant antibody heterogeneity and stability. *Journal of Chromatography A 1120*, 112-120.

Chapter 2. Theoretical Background

This chapter presents the computational techniques and theoretical background employed in this thesis. The intention is not to reproduce information in textbooks and literature, but instead to summarize the relevant parts in this thesis work.

2.1 Classical molecular dynamics

Classical molecular dynamics (MD) simulations involve the numerical integration of Newton's equation of motion to generate physical trajectories, which are chronological sequences of points in phase space (the combination of configuration space and momentum space) for a system of interacting particles. Its aim is to study the detailed microscopic dynamic behavior in many-body systems which are of interest to the fields of physics, chemistry and biology. It is based on statistical mechanics, especially the ergodic hypothesis, in order to yield properties of the molecular systems both at equilibrium and at non-equilibrium (*1, 2*). The techniques of MD have been widely used in the field of biophysics to study problems involving protein systems (*3*).

2.1.1 Force fields for protein systems in aqueous solutions

In order to provide a description of the interaction between particles in a molecular system, a model or force field needs to be introduced as the input for classical MD simulations. The use of predefined function forms and the parameters fitted against experimental data or high-level quantum mechanical calculations, on one hand, greatly reduces the computational overhead; on the other hand, such use also limits their transferability to a condition which may greatly differ from the one these parameters were originally developed. Thus, it is not possible to use the force-field developed with

single molecular connectivity to study chemical reaction processes where bond breaking and formation are involved (*3*).

There are a number of force fields developed for various particular systems, such as Consistent Force Field (CFF) (*4, 5*) for organic compounds including polymers, Optimized Potential for Liquid Simulations (OPLS) (*6, 7*), Merck Molecular Force Field (MMFF) (*8*), etc. For protein systems, CHARMM (*9*) force field is widely used. The basic function forms in CHARMM include terms describing intra- and intermolecular interactions, such as the energetic penalties from the deviation of bonds and angles away from their "reference" or "equilibrium" values; the function describing how the energy changes as bonds are rotated (so called dihedral energy); and also the interaction between non-bonded parts of the system. The non-bonded interaction is calculated between all pairs of atoms. Usually this term is modeled as a Coulomb potential term for electrostatic interaction and a Lennard-Jones potential for van der Waals interactions. The force field gives the potential energy E_{total} as a function of internal coordinates (such as bond lengths, angles, dihedral angles) of the simulation system with the following mathematical form:

$$E_{total} = \sum_{bonds} K_b (b - b_0)^2 + \sum_{angles} K_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} K_\chi [1 + \cos(n\chi - \delta)] + \sum_{non-bonded} \varepsilon [(\frac{R_{\min_{ij}}}{r_{ij}})^{12} - (\frac{R_{\min_{ij}}}{r_{ij}})^6] + \frac{q_i q_j}{\varepsilon_1 r_{ij}} \quad (2.1)$$

where K_b, K_θ, and K_χ are the bond, angle, and dihedral angle force constants, respectively; b, θ, and χ are the bond length, angle, and dihedral angle, respectively, with the subscript zero denoting the equilibrium values for the individual terms. Coulomb and Lennard-Jones 6-12 terms contribute to the non-bonded interactions; ε is

the Lennard-Jones well depth and R_{min} is the distance at the Lennard-Jones minimum. The q_i is the partial charge on particle i, ε_l is the effective dielectric constant, and r_{ij} is the distance between atom i and j.

2.1.2 Numerical integration of Newton's equation of motion

After the potential energy function $V(\mathbf{r}_1,\mathbf{r}_2,...,\mathbf{r}_N)$ as a function of coordinates of the N particles in the system is defined, the motion of the system is governed deterministically by Newton's equation of motion,

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = -\nabla_{\mathbf{r}_i} V(\mathbf{r}_1,\mathbf{r}_2,...,\mathbf{r}_N) \tag{2.2}$$

where t is time, m_i is the mass, and $\mathbf{r}_i$ is the position vector of particle i.

The ordinary second-order differential equation (2.2), usually subject to some initial conditions given by such variables as the initial positions and velocities, can be numerically integrated using schemes which preserve long-time energy conservation and time reversibility such as Verlet's algorithm etc. (*1, 2*). The integration time-step needs to be matched with the fastest vibrational frequency in the simulation system in order to assure numerical stability. This requirement can severely limit its applicability to obtain trajectories long enough for interesting physical or chemical process to occur. Thus, rigid bond constraints are often applied to suppress the high frequency bond stretching modes through the SHAKE (*10*) or RATTLE algorithm (*11*).

2.1.3 Classical MD in various ensembles

Since only a limited number of particles can be dealt with in any practical computer simulation, which is far less than required to reach the thermodynamic limit,

performing the integration of Newton's equation of motion with some specified macroscopic state variables (such as the number of particles N, the chemical potential μ, the volume of the system V, the pressure P, the total energy of the system E, and the temperature T), also termed thermodynamic variables, can be advantageous in some cases. These ensembles include the microcanonical ensemble with NVE fixed, the canonical ensemble with NVT fixed, the grand-canonical ensemble with μVT fixed, and the constant temperature and pressure ensemble with NPT fixed. Different ensembles do make a difference when computing the mean-squared value of fluctuations in thermodynamic quantities.

Often, the choice to perform MD in various ensembles is to use the extended system approach. For example, to perform a canonical ensemble simulation, temperature was controlled by a thermostat which has its own dynamics. The motion of the system together with the thermostat is governed by an extended Lagrangian (*12*):

$$L_{Nose} = \sum_{i=1}^{N} \frac{m_i}{2} s^2 \dot{\mathbf{r}}_i^2 - U(\mathbf{r}^N) + \frac{Q}{2}\dot{s}^2 - \frac{L}{\beta}\ln s \tag{2.3}$$

where Q is an effective mass associate with s, the thermostat coordinate; N is the number of particles; m_i and $\mathbf{r}_i$ are the mass and the position vector of particle i, respectively; L=3N; and $\beta = 1/(k_B T)$. Similarly, a barostat can be added to a constant pressure simulation (*13*).

2.2 *Ab initio* molecular dynamics

In classical MD simulations, the system under study evolves on a single potential energy surface, which describes the ground state of the system. The electronic degrees of freedom are omitted due to the Born-Oppenheimer approximation, and only the nuclear

motion is modeled using laws of classical mechanics. This approach is usually a good approximation for molecular systems when the properties being studied are not related to the motion of light atoms (i.e., hydrogen or electrons) or to vibrations with a frequency υ such that $h\upsilon < k_B T$ where h is the Planck constant, k_B is the Boltzmann constant, and T is the temperature of the system. When electronic degrees of freedom are important, first-principle derived potential function needs to be calculated, often on-the-fly as the numerical integration of Newton's equation of motion is performed.

There are different ways of performing *ab initio* MD, mainly differing in terms of how the electronic structure was computed and how to combine the motions in the electronic degrees of freedom and in the nuclei degrees of freedom. Car-Parrinello molecular dynamics (CPMD) based on density functional theory (DFT) (*14*) was chosen in this thesis and is introduced in the following section.

2.2.1 Density functional theory

Density functional theory is an alternative formulation of quantum mechanics based on electron density rather than wave function. In a many-body molecular system, the Hamiltonian for describing the interaction between nuclei and electrons takes the following form (in atomic units, where the length unit is the Bohr radius a_0=0.5292Å, the charge unit is the charge of the electron e, and the mass unit is the mass of the electron m_e):

$$\hat{H} = -\sum_{I=1}^{P} \frac{\hbar^2}{2M_I} \nabla_I^2 - \sum_{i=1}^{N} \frac{\hbar^2}{2} \nabla_i^2 + \sum_{I<J}^{P} \frac{Z_J Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} + \sum_{i<j}^{N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} - \sum_{I}^{P} \sum_{i}^{N} \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}_i|} \quad (2.4)$$

where $R=\{\mathbf{R}_I\}$, I=1,..., P, is a set of nuclear coordinates, and $r=\{\mathbf{r}_i\}$, i=1,..., N, is a set of electron coordinates. Z_I and M_I are the nuclear charges and masses, respectively.

The energy E and the quantum state described by wave function $\Psi(\mathbf{R},\mathbf{r})$ are governed by Schrodinger's equation:

$$\hat{H}\Psi(\mathbf{R},\mathbf{r}) = E\Psi(\mathbf{R},\mathbf{r}) \tag{2.5}$$

Often, considering the large differences between the electrons and nuclei, the timescale associated with the motion of the nuclei is much slower than that associated with the motion of the electrons. Thus, the Born-Oppenheimer (BO) approximation, essentially the separation of electronic and nuclear coordinates in the many-body wave function,

$$\Psi(\mathbf{R},\mathbf{r}) = \Theta(\mathbf{R})\Phi(\mathbf{R},\mathbf{r}) \tag{2.6}$$

is often assumed. The BO approximation reduces the many-body problem to the dynamics of electrons in some frozen-in configuration of the nuclei, and often only the electronic part of the Schrodinger's equation

$$(-\sum_{i=1}^{N}\frac{\hbar^2}{2}\nabla_i^2 + \sum_{i<j}^{N}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|} - \sum_{I}^{P}\sum_{i}^{N}\frac{Z_I}{|\mathbf{R}_I-\mathbf{r}_i|})\Phi(\mathbf{R},\mathbf{r}) = E\Phi(\mathbf{R},\mathbf{r}) \tag{2.7}$$

is solved.

Even after the application of BO approximation to arrive at a simpler Equation (2.7), the many-body problem is still formidable. Kohn and Sham expressed the total ground-state energy of an interacting system of electrons in terms of electronic charge density $n(\mathbf{r})$,

$$
\begin{aligned}
&E^{KS}[\{\phi_i\}] = T_s[\{\phi_i\}] + \int d\mathbf{r}\, V_{ext}(\mathbf{r}) n(\mathbf{r}) + \frac{1}{2}\int d\mathbf{r}\, V_H(\mathbf{r}) n(\mathbf{r}) + E_{xc}[n(\mathbf{r})] \\
&n(\mathbf{r}) = \sum_i |\phi_i(\mathbf{r})|^2 \\
&T_s[\{\phi_i\}] = \sum_i \frac{\hbar^2}{2}\int d\mathbf{r}\, \phi_i^*(\mathbf{r})(-\nabla^2)\phi_i(\mathbf{r}) \\
&V_{ext}(\mathbf{r}) = -\sum_I \frac{Z_I}{|\mathbf{R}_I - \mathbf{r}|} + \sum_{I<J} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} \\
&V_H(\mathbf{r}) = \int d\mathbf{r}' \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}
\end{aligned}
\tag{2.8}
$$

where a set of orthogonal one-particle functions $\{\phi_i(\mathbf{r})\}$ is the Kohn-Sham orbitals, The $T_s[\{\phi_i\}]$ is the kinetic energy of a non-interacting reference system with the same number of electrons exposed to the same external potential as the full interacting system; $V_{ext}(\mathbf{r})$ is the fixed external potential in which the electrons move, which comprises Coulomb interactions between the electrons and nuclei; and $V_H(\mathbf{r})$ is the Hartree potential, which is classical electrostatic energy of two charge clouds. The $E_{xc}[n(\mathbf{r})]$ is the exchange-correlation energy. [] here and above denotes functional.

If the Kohn-Sham energy functional $E^{KS}[\{\phi_i\}]$ is minimized with respect to the orbitals $\{\phi_i(\mathbf{r})\}$, the resulting equation is:

$$
\{-\frac{\hbar^2}{2}\nabla_i^2 + V_{ext}(\mathbf{r}) + V_H(\mathbf{r}) + \frac{\partial E_{xc}[n(\mathbf{r})]}{\partial n(\mathbf{r})}\}\phi_i(\mathbf{r}) = \varepsilon_i \phi_i(\mathbf{r}) \tag{2.9}
$$

Hohenberg and Kohn proved that if the exact forms of $T_s[\{\phi_i\}]$ and $E_{xc}[n(\mathbf{r})]$ are known, then the exact energy would be obtained from just the density $n(\mathbf{r})$.

Much of the effort in DFT has been to find an accurate expression for the exchange-correlation functional $E_{xc}[n(\mathbf{r})]$,

$$
E_{xc}[n(\mathbf{r})] = \int d\mathbf{r}\, n(\mathbf{r}) \varepsilon_{xc}[n(\mathbf{r})] \tag{2.10}
$$

where $\varepsilon_{xc}[n(\mathbf{r})]$ is the exchange-correlation energy per particle.

Two common approximations for $E_{xc}[n(\mathbf{r})]$ are the local density approximation (LDA) and the generalized gradient approximation (GGA). In LDA, $\varepsilon_{xc}[n(\mathbf{r})]$ is simply approximated by the exchange-correlation energy associated with a homogeneous electron gas of the same density, which can be obtained precisely. If we write $\varepsilon_{xc}[n(\mathbf{r})] = \varepsilon_x[n(\mathbf{r})] + \varepsilon_c[n(\mathbf{r})]$, then

$$\varepsilon_x[n(\mathbf{r})] = -\frac{3}{4}(\frac{3}{\pi})^{1/3} n(\mathbf{r})^{1/3} \tag{2.11}$$

and $\varepsilon_c[n(\mathbf{r})]$ has been obtained from quantum Monte Carlo simulations. The LDA gives a poor description of molecular systems where the electron density undergoes rapid changes.

Therefore, in GGA, gradients of the density are added to the exchange-correlation functional,

$$E_x[n(\mathbf{r})] = \int d\mathbf{r}\, F(\mathbf{r}, \nabla\mathbf{r}) \tag{2.12}$$

Two commonly used GGA functionals are the Becke-Lee-Yang-Parr (BLYP), with the Becke exchange functional, and the Lee-Yang-Parr correlation functional, and the Perdew-Wang (PW91) functional.

2.2.2 Car-Parrinello MD (CPMD)

Car-Parrinello molecular dynamics (*15*) is an *ab initio* MD technique employing periodic boundary conditions, plane-wave basis sets, and DFT. The basic idea of the Car-Parrinello approach is to write an extended Lagrangian

$$L = \frac{1}{2}(\sum_I M_I \dot{\mathbf{R}}_I^2 + \mu \int d\mathbf{r} \sum_i |\dot{\phi}_i|^2) - E^{KS}[\{\phi_i\}, \{\mathbf{R}_i\}] \tag{2.13}$$

subject to a set of orthogonality constraints

$$\int d\mathbf{r}\, \phi_i^* \phi_j = \delta_{ij} \tag{2.14}$$

for the system which leads to a system of coupled equations of motion for both ions and electrons by explicitly introducing the electronic degrees of freedom as fictitious dynamical variables ϕ_i. In this way an explicit electronic minimization at each iteration is not needed: after an initial standard electronic minimization, the fictitious dynamics of the electrons keep them on the electronic ground state corresponding to each new ionic configuration visited along the dynamics, thus yielding accurate ionic forces. In order to maintain this adiabatic condition, it is necessary that the fictitious mass of the electrons μ is chosen small enough to avoid a significant energy transfer from the ionic to the electronic degrees of freedom.

2.3 Rare event simulation techniques

Rare events are ubiquitous in physics, chemistry, and biology, such as chemical reactions, formation of critical nuclei during crystallization processes, and the protein folding process. The common feature of these processes is the timescale separation,

$$\tau_{trans} << \tau_{stable} \tag{2.15}$$

where the transition process is an event of short duration separated by relatively longer waiting times between two transition events. The timescale separation often is the characteristic of activated processes, in which two stable states are separated by a free energy barrier.

This wide disparity of timescales can present serious computational challenges (*16*). If one were to run a long MD simulation to study the transition process, all kinetic

information one is looking for in principle can be obtained; however, it would be very impractical because most of the computational time would be spent when the system is in stable or meta-stable states because the state lifetime τ_{stable} depends exponentially on the free energy barrier height, and quite rarely will the system undergo transition. This phenomenon of rare transition is the manifestation of the timescale separation.

Another issue which makes the study of rare events more complicated is the roughness of the potential energy landscape. Unlike systems in the gas phase where there are only a few saddle points on the potential energy surface which usually can be used to sufficiently characterize the transition process, for rare events occurring in liquids, the system of interest has a rugged potential energy surface on which myriads of small energy barriers with a height of the order of k_BT, which have to be distinguished with the true potential energy barrier often larger than k_BT.

One way to get around the challenges posed by timescale separation and the roughness of the potential energy surface is to focus on the dynamic bottleneck for the rare event which is defined as the transition state surface, as the transition path sampling technique illustrates in the following section.

2.3.1 Transition path sampling

Transition path sampling (TPS) technique is basically an importance sampling algorithm in the dynamic trajectory space in order to collect an ensemble of transition trajectories. The most significant advantage of this technique is that no prior knowledge about the reaction mechanism, the reaction coordinate and the transition state is needed to begin with. Rather, the results of TPS can be used to obtain information about the transition state ensemble, the free energy profile and the reaction coordinate. The only

necessary conditions for starting a transition path sampling simulation are the definitions of free energy basins of attraction by suitable order parameters, and an initial transition trajectory, or a series of points in the phase-space which connect the two basins of attraction, which may not be physical or at the same condition as the one of interest.

The original algorithm of TPS uses two types of moves in the process of a random walk in the trajectory space, the shooting moves and the shifting moves. As in the configurational MC algorithm, these moves need to obey the detailed balance condition in which path probability needs to be used instead of the configurational probability.

A more efficient sampling technique to explore the trajectory space is the aimless-shooting algorithm (*17, 18*). It has the advantage of having more decorrelation between successive trajectories than the original version of TPS, and thus it can explore the trajectory space more quickly. As shown in Figure 2.1, the basic idea of the aimless shooting algorithm is to make shooting moves among three possible points on an old accepted trajectory, $\mathbf{x}^{(o)}_{-\Delta t}$, $x^{(o)}_{0}$, and $\mathbf{x}^{(o)}_{+\Delta t}$.

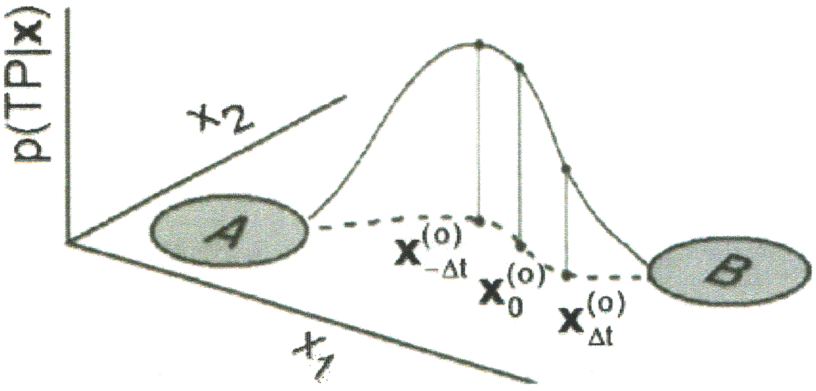


Figure 2.1: Illustration of the aimless shooting algorithm on a two-dimensional system. A and B are stable state basins, connected by a reactive trajectory. Three points ($\mathbf{x}^{(o)}_{-\Delta t}$, $x^{(o)}_{0}$, and $\mathbf{x}^{(o)}_{+\Delta t}$) are on the old trajectory. $P(\mathrm{TP}|\mathbf{x})$ is the probability of a trajectory initiated from x is a transition path.

It can derived that the condition of detailed balance translates into following acceptance criterion for any new trajectory initiated randomly from $\mathbf{x}^{(o)}_{-\Delta t}$, $\mathbf{x}^{(o)}_{0}$, and $\mathbf{x}^{(o)}_{+\Delta t}$ on the old trajectory with equal probability:

$$\rho^{o\to n}_{acc.TPS} = h_A[x^{(n)}_{-t/2}]h_B[x^{(n)}_{t/2}] \tag{2.16}$$

2.3.2 Likelihood maximization

Having collected an ensemble of trajectories, with their statistical weight in trajectory space correctly built in, there are different ways to identify the important degrees of freedom, namely, the factors in comprising the reaction coordinate. Ma et al. (*19*) developed an algorithm based on neural networks and genetic algorithm to identify reaction coordinate. In their method, neural networks are used to determine the functional dependence of the committor probability on a set of coordinates, and the genetic algorithm selects the combination of elements that yields the best fit from this set of coordinates.

An alternative and complementary technique, likelihood maximization (*17, 18*), to select physically important degrees of freedom is based on informatics theory. It screens a set of candidate collective variables for a good reaction coordinate that depends only on a few relevant variables. Selection of the candidate collective variables is crucial in both this approach and the approach developed in (*19*). The possible collective variables can be: 1) pseudo-internal coordinates including geometric quantities such as distances, angles and dihedral angles; 2) accessible surface area of, radius of gyration of, and excluded volume of the solute; 3) number of hydrogen bonds, protein native contact number, etc. The computational screening of single such collective variable is fast, thus it is feasible to perform the screening over a large number of candidate variables.

Likelihood maximization algorithm starts with a data set, which is the rejected or accepted shooting points in TPS, and it seeks a committor probability model $p_B[r(\mathbf{x})]$ and a reaction coordinate model $r(\mathbf{x})$ as a function of the shooting point configuration $\mathbf{x}$ in order to explain the realizations of $p_B[r(\mathbf{x})]$ that were obtained from TPS. The reaction coordinate model $r(\mathbf{x})$ is assumed to be a linear combination of m candidate collective variables $q_1, q_2, \ldots, q_m$

$$r(\mathrm{x}) = \alpha_0 + \sum_{k=1}^{m} \alpha_k q_k \tag{2.17}$$

where α_k (k=0, ..., m) are undetermined parameters. The model of $p_B[r(\mathbf{x})]$ can also be assumed to be

$$p_B[r(\mathrm{x})] = \frac{1}{2}[1 + \tanh(r)] \tag{2.18}$$

based on the requirement that $p_B[r(\mathbf{x})\to-\infty]\to 0$, $p_B[r(\mathbf{x})\to 0]\to 1/2$, and $p_B[r(\mathbf{x})\to+\infty]\to 1$. The models of $r(\mathbf{x})$ and $p_B[r(\mathbf{x})]$ can then be used to compute the likelihood score given the shooting data:

$$L(\boldsymbol{\alpha}) = \prod_{k=1}^{B} p_B[r(\mathbf{q}(\mathbf{x}^{(k)}))] \prod_{l=1}^{A} p_B[r(\mathbf{q}(\mathbf{x}^{(l)}))] \tag{2.19}$$

where B is the number of trajectories with end points in basin B, and A is the number of trajectories with end points in basin A. The $\mathbf{q}(\mathbf{x}^{(k)})$ are the values of collective variables at the shooting point configuration $\mathbf{x}^{(k)}$ for trajectory number k. The explicit function dependence of r on the m collective variables $\mathbf{q}$, and the function dependence of $\mathbf{q}$ on the shooting point configuration $\mathbf{x}$ was explicitly written out.

For the set of trajectories generated by TPS algorithm, the following terminology is used throughout this thesis. Conclusive trajectories are the ones with both ends in

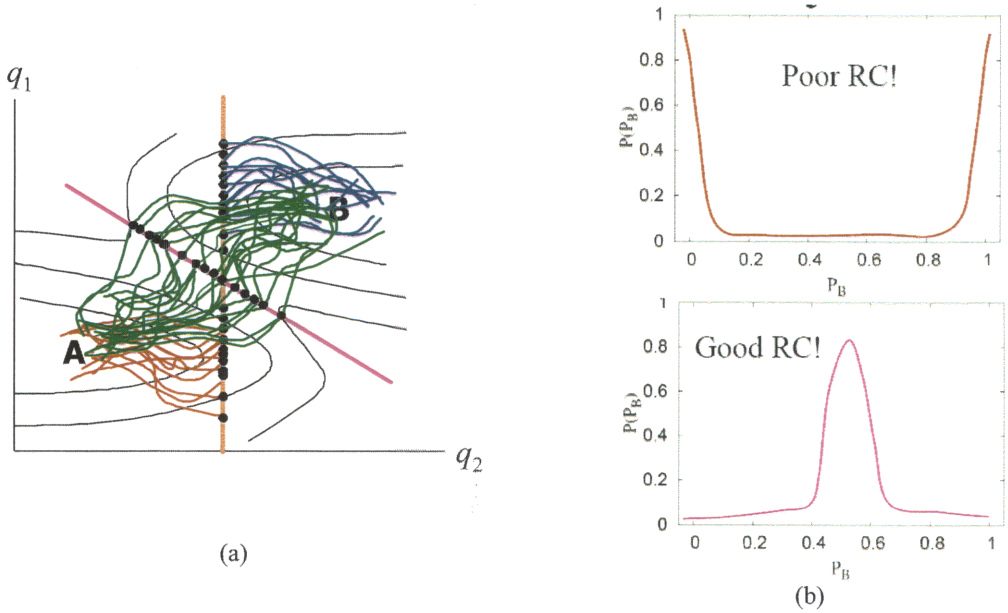
either basin, and reactive or accepted trajectories refer to the ones with both ends in different basins. Due to the possibility of running shorter trajectories, a trajectory can have either or both ends not committing to any basin. This situation has to be controlled to have a low percentage in the collected trajectories in TPS algorithm by adjusting basin definitions and the acceptance probability, and the trajectories can not be used in the likelihood maximization.

The likelihood score in Equation (2.19) is to be maximized in order to determine the unknown coefficient α_k (k=0, ..., m), the completely determine the reaction coordinate model $r(\mathbf{x})$ and the committor probability model $p_B[r(\mathbf{x})]$. In practical algorithm, the logarithm of $L(\boldsymbol{\alpha})$ is used because of its numerical ease to handle.

There is still one more degree of freedom that needs to be fixed, m, the number of the collective variables q_k (k=1, ..., m) involved in the reaction coordinate model $r(\mathbf{x})$. The Bayesian information criterion (*20*) can be used to determine whether additional variables significantly improve the reaction coordinate and thus needs to be included. Namely, when m is the same, $r(\mathbf{x})$ models with different sets of q_k (k=1, ..., m) can be compared with the log likelihood score. The larger the score is, the more likely for the model to be true given the TPS shooting data. When comparing one models of $r(\mathbf{x})$ with sets of m and n (with $n<m$) collective variables, having maximized log likelihood scores l_m and l_n respectively, if $l_m - l_n > (m-n)\times\log(N)/2$, then the reaction coordinate model $r(\mathbf{x})$ with m collective variables is statistically superior than the model with n variables, and if $l_m - l_n < (m-n)\times\log(N)/2$, then the two models are statistically indistinguishable. Here N is the number of realizations in the likelihood function, or, equivalently, the number of conclusive trajectories.

2.3.3 Reaction coordinate verification

The standard technique to verify the validity of a reaction coordinate is the committor probability p_B histogram analysis (*16, 21, 22*). The basic idea of p_B histogram analysis is illustrated in Figure 2.2. The committor describes the partitioning of short dynamic trajectories, originating from the assumed transition state region and with randomly chosen initial momenta sampled from a Maxwell distribution, into the various free energy states, in this case A and B. As shown in Figure 2.2, points from the p_B=0.5 iso-surface constructed from the assumed reaction coordinate can be chosen, and many short random trajectories can be fired from these points. The resulting shape of the distribution of the committor probability can be used to tell whether the assumed reaction coordinate is "good" or "poor". Here "good" means the reaction coordinate model includes all the dynamically relevant degrees of freedom.



(a) (b)

Figure 2.2: (a) Free energy landscape for a system with two dividing surfaces constructed from two reaction coordinate models. There are two stable states, A and B, for this free energy surface. The purple line represents the p_B=0.5 iso-surface for a "good" reaction coordinate, where most trajectories initiated from points on this line have a 0.5 probability of committing to either basin. However, if a "poor" reaction coordinate was

chosen, for example, its p_B=0.5 iso-surface can be the red line, the trajectories initiated from points on this red line have a bimodal probability distribution of committing to either basin. The probability distributions of the shooting points having different committor probability distributions for the two cases are shown in (b).

2.3.4 Reaction rate constant calculation

In the reactive flux formulism of reaction rate, the starting point is the correlation function:

$$C(t) = \frac{< h_A(0) h_B(t) >}{< h_A(t) >} \tag{2.20}$$

where $h_\Omega(t) \equiv h_\Omega(\mathbf{r}(t))$ is the characteristic function, i.e. $h_\Omega(\mathbf{r}(t)) = 1$ if $\mathbf{r}(t) \in \Omega$ and 0 elsewhere. As a consequence of the separation of timescales, for time t such that $\tau_{trans} << t << \tau_{stable}$, the rate constant and the correlation function $C(t)$ are related by

$$C(t) \approx k_f t \tag{2.21}$$

Therefore, the first derivative of $C(t)$, $k(t) \equiv \dot{C}(t)$, called the reactive flux, has a constant value equal to the forward rate constant.

Through some algebraic manipulation, one can arrive at

$$k(t) = \frac{\langle \dot{q}(0) \delta(q(0) - q^*) \theta(q(t) - q^*) \rangle}{\langle \theta(q^* - q(0)) \rangle} \tag{2.22}$$

where $q(\mathbf{r})$ is the one-dimensional reaction coordinate, <> denotes the canonical ensemble average, q^* is the reaction coordinate at the transition state, $\delta(x)$ and $\theta(x)$ are the standard Dirac delta function and Heaviside step function respectively.

2.4 Reaction rate theories

Reaction rate theory concerns with the rate at which the rare events mentioned above occur. The simplest form of the reaction rate theory is the problem of escape rate from metastable states, as illustrated in Figure 2.3. At finite temperature T, a particle may wander around in state A for a long time before it has gained enough kinetic energy from the heat bath to overcome the potential energy barrier to reach state B. Heat bath, or the specification of constant temperature T, endows thermal activation for the transition.

In a more realistic description of the reaction rate theory, the x-coordinate in Figure 2.3 is replaced by the reaction coordinate, and the dynamics along the direction of the reaction coordinate is stochastic in nature, which includes a combined effect induced by the coupling among a multitude of environmental degrees of freedom. There are two aftermaths when the dynamics only along the reaction coordinate direction is considered. One is friction, due to the reduced action of the degrees of freedom that are lost upon contraction of the complete phase-space dynamics. The other one is entropy, due to the reduction of all coupled degrees of freedom from a high-dimensional potential energy surface in full phase-space to an effective potential, i.e., the potential of mean force along the reaction coordinate direction.

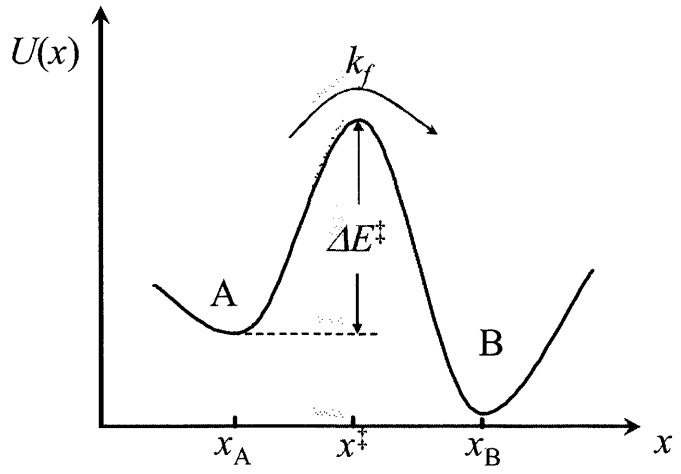


Figure 2.3: (a) The Kramers problem for a double-well potential. Two minima of the potential energy are located at x_A and x_B, and a saddle point at $x^{\ddagger}$. The energy barrier between the stable state A and the saddle point is $\Delta E^{\ddagger}$, which determines the forward escape rate k_f from A to B.

The equation of motion for the reaction coordinate can be obtained by the technique of projection operator, and the resulting "coarse-grained" dynamics is described in a form of generalized Langevin equation where the time-dependent frictional force has a memory kernel.

2.4.1 Transition state theory

Transition state theory is based on the assumptions that: 1) passage through a transition state is committed to a stable state without subsequent return for a longer period of time than the time duration of the transition dynamics; 2) thermodynamic equilibrium is maintained throughout the entire system for all degrees of freedom, including everywhere along the reaction coordinate. From the reactive flux formulism of reaction rate constant, it is straightforward to derive the transition state rate constant for the escape problem of the double-well potential in Figure 2.3. From Equation (2.22) and the assumption of "passage without return", $\theta(\dot{q}(t)-q^*)$ can be replaced by $\theta(\dot{q}(0))$, and

$$k^{TST}(t) = \frac{\langle \dot{q}(0)\delta(q(0)-q^*)\theta(\dot{q}(0))\rangle}{\langle \theta(q^*-q(0))\rangle} \tag{2.23}$$

Also using the equilibrium assumption, the ensemble average can be evaluated with equilibrium probability density, thus

$$k^{TST} = \sqrt{\frac{k_B T}{2m\pi}} \frac{\exp(-\frac{\Delta E^{\ddagger}}{k_B T})}{\int_{-\infty}^{\ddagger} \exp(-\frac{U(x)-U(x^{\ddagger})}{k_B T})dx} \tag{2.24}$$

For multidimensional cases, the expression of transition state reaction rate constant can be casted into a simple form

$$k^{TST}(t) = \frac{k_B T}{h} \exp(-\frac{\Delta G^{\ddagger}}{k_B T}) \tag{2.25}$$

where $\Delta G^{\ddagger}$ is the difference in Gibbs free energies between the transition state and the reactant state.

2.4.2 Kramers' theory

The Kramers problem is to find the rate at which a Brownian particle escapes from a potential well over a potential barrier. It is the problem depicted in Figure 2.3, with B replaced by an infinitely deep well. This problem is a well studied and widely generalized. The equation of motion for the Brownian particle is the Lagevin equation under an external potential $U(x)$,

$$m\frac{d^2x}{dt^2} = -\frac{dU}{dx} - m\zeta\frac{dx}{dt} + R(t) \tag{2.26}$$

where m is the mass of the particle, t is time, ζ is the frictional coefficient, and $R(t)$ is a time-dependent random force satisfying the fluctuation-dissipation theorem

$$< R(t)R(t') >= 2m\zeta k_B T\delta(t-t') \tag{2.27}$$

where T is the temperature, k_B is the Boltzmann's constant, $\delta(x)$ is the standard Dirac delta function.

Two key quantities determine the escape rate in Kramers problem, the vibrational frequency at the saddle point $\omega^{\ddagger}$ and the potential energy barrier $\Delta E^{\ddagger}$. Depending on two dimensionless quantities $\zeta/\omega^{\ddagger}$ and $k_B T/\Delta E^{\ddagger}$, Equation (2.26) can be simplified to give closed-form expressions for the forward reaction rate constant. At low friction and high energy barrier, the rate of escape is:

$$r = \frac{\omega_A \omega^{\ddagger}}{2\pi\zeta}\exp(-\frac{\Delta E^{\ddagger}}{k_B T}) \tag{2.28}$$

where ω_A is the vibrational frequency at well A.

While at high friction and high energy barrier, the rate of escape is:

$$r = \frac{\zeta\Delta E^{\ddagger}}{\pi k_B T}\exp(-\frac{\Delta E^{\ddagger}}{k_B T}) \tag{2.29}$$

In Figure 2.4, the dependence of the escape rate r on the frictional coefficient ζ is illustrated. It should be noted that there is a "turnover" region which connects the two limits.

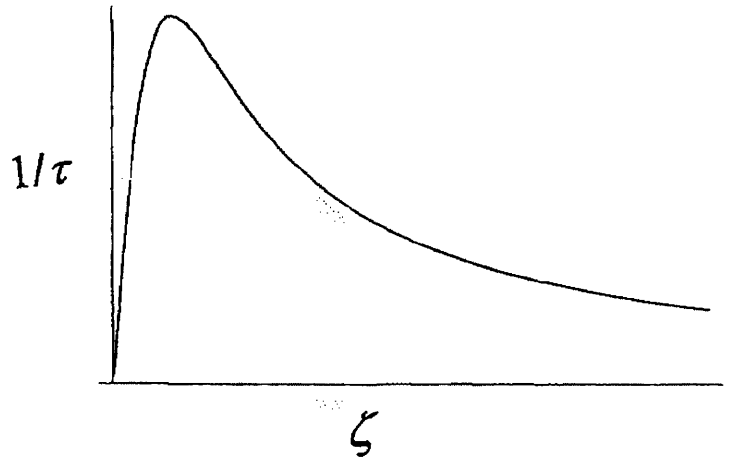


Figure 2.4: The escape rate as a function of frictional coefficient ζ. In the regime of low friction, the rate of escape is proportional to the frictional coefficient ζ. In the regime of high friction, or the energy diffusion regime, the rate is inversely proportional to friction coefficient ζ.

2.5 References

(1) Allen, M. P., and Tildesley, D. J. (1989) *Computer simulation of liquids*, Oxford University Press, Oxford [England].

(2) Frenkel, D., and Smit, B. (2002) *Understanding molecular simulation: from algorithms to applications*, 2nd ed., Academic, San Diego, Calif.; London.

(3) Tuckerman, M. E., and Martyna, G. J. (2000) Understanding modern molecular dynamics: Techniques and applications. *Journal of Physical Chemistry B 104*, 159-178.

(4) Kuczera, K., and Czerminski, R. (1983) Properties of Quadratic Force-Fields in Redundant Coordinates. *Theochem-Journal of Molecular Structure 14*, 269-280.

(5) Kuczera, K. (1987) Uniquely Defined Harmonic Force-Constants in Redundant Coordinates. *Journal of Molecular Structure 160*, 159-177.

(6) Jorgensen, W. L., and Tiradorives, J. (1988) The Opls Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin. *Journal of the American Chemical Society 110*, 1657-1666.

(7) Jorgensen, W. L., Maxwell, D. S., and TiradoRives, J. (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society 118*, 11225-11236.

(8) Halgren, T. A., and Bush, B. L. (1996) The Merck molecular force field (MMFF94). Extension and application. *Abstracts of Papers of the American Chemical Society 212*, 2-Comp.

(9) MacKerell, A. D. (1998) Developments in the CHARMM all-atom empirical energy function for biological molecules. *Abstracts of Papers of the American Chemical Society 216*, U696-U696.

(10) Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *Journal of Computational Physics 23*, 327-341.
(11) Andersen, H. C. (1983) Rattle - a Velocity Version of the Shake Algorithm for Molecular-Dynamics Calculations. *Journal of Computational Physics 52*, 24-34.
(12) Andersen, H. C. (1980) Molecular-Dynamics Simulations at Constant Pressure and-or Temperature. *Journal of Chemical Physics 72*, 2384-2393.
(13) Martyna, G. J., Tobias, D. J., and Klein, M. L. (1994) Constant-Pressure Molecular-Dynamics Algorithms. *Journal of Chemical Physics 101*, 4177-4189.
(14) Hutter, J. (2002).
(15) Car, R., and Parrinello, M. (1985) Unified Approach for Molecular-Dynamics and Density-Functional Theory. *Physical Review Letters 55*, 2471-2474.
(16) Bolhuis, P. G., Chandler, D., Dellago, C., and Geissler, P. L. (2002) Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual Review of Physical Chemistry 53*, 291-318.
(17) Peters, B., Beckham, G. T., and Trout, B. L. (2007) Extensions to the likelihood maximization approach for finding reaction coordinates. *Journal of Chemical Physics 127*, -.
(18) Peters, B., and Trout, B. L. (2006) Obtaining reaction coordinates by likelihood maximization. *Journal of Chemical Physics 125*, -.
(19) Ma, A., and Dinner, A. R. (2005) Automatic method for identifying reaction coordinates in complex systems. *Journal of Physical Chemistry B 109*, 6769-6779.
(20) Schwarz, G. (1978) Estimating Dimension of a Model. *Annals of Statistics 6*, 461-464.
(21) Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T., and Shakhnovich, E. S. (1998) On the transition coordinate for protein folding. *Journal of Chemical Physics 108*, 334-350.
(22) Bolhuis, P. G., Dellago, C., and Chandler, D. (1998) Sampling ensembles of deterministic transition pathways. *Faraday Discussions*, 421-436.

Chapter 3. Comparative Oxidation Studies of Methionine Residues in rhG-CSF

3.1 Introduction

The oxidation of methionine amino acids is an important reaction for proteins both *in vivo* and *in vitro*. *In vivo*, a number of processes of biological interest involve methionine oxidation, such as aging (*1-3*), Parkinson's diseases (*4*), and Alzheimer's disease (*5, 6*). In addition, the oxidation of a methionine residue and its reduction by methionine sulfoxide reductase is thought to be a regulatory mechanism for enzyme bioactivity (*7, 8*). *In vitro*, oxidation is important in the production of protein pharmaceuticals. An inevitable problem for these protein molecules in aqueous solution is the various physical and/or chemical degradation reactions that occur during the shelf life of the product (*9*). Oxidation is one of the common chemical degradation pathways, often resulting in structural changes (*10*) and bioactivity losses (*11*). Minimizing oxidation as well as other forms of degradation is essential in formulating pharmaceutical proteins (*12*).

rhG-CSF is a four-helix bundle protein with 175 amino acid residues (*13*), containing four methionine residues. The crystal structure (*14, 15*) is shown in Figure 3.1, and all four methionine residues are highlighted in a ball-stick representation; its main chain backbone is shown in a ribbon representation. Met 1 is at the N-terminus, encoded by the initiation codon in protein synthesis, and is therefore very flexible. Met 122 is within helix C, facing the interior and buried to a great extent. Met 127 and Met 138 both face the protein interior and are located in the B-C loop, which is more flexible than the

helix that contains Met 122. Also shown in Figure 3.1 is the location of the two tryptophan residues, Trp 59 and Trp 119, which serve as fluorescence spectral probes of the local environment. Trp 59 resides in the long loop AB between helices A and B with its side chain pointing out toward solvent, whereas Trp 119 resides within helix C with its side chain between the interface of helix C and loop CD. Trp 59 is located far away from all of the methionine residues and is much exposed despite its hydrophobic nature; while Trp 119 is close to Met 122 and is much buried. However, neither of the two tryptophan residues has a similar micro-environment as any of the methionine residues.

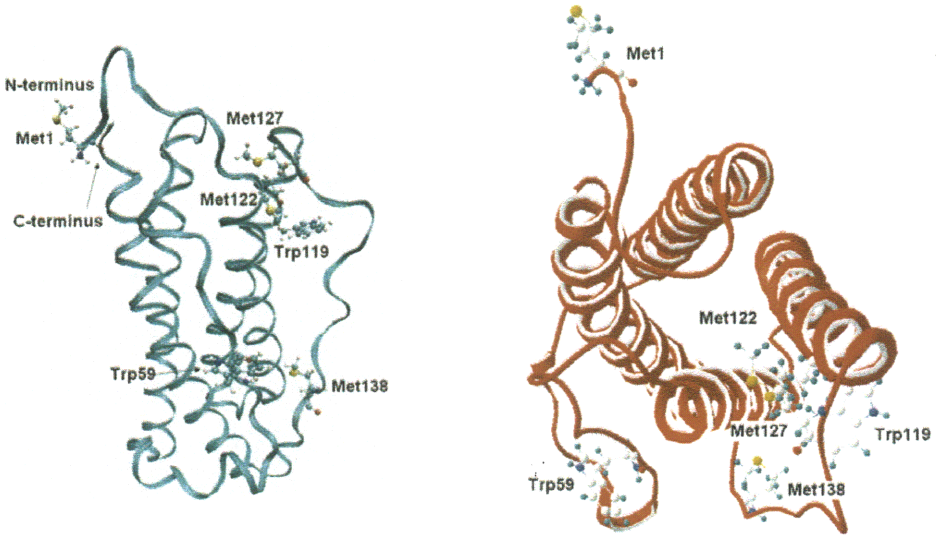


Figure 3.1: Crystal structure of rhG-CSF from (*14, 15*). Four N-terminal residues missing in the X-ray structure (PDB code: 1cd9), MTPL, were added. Their atomic position in space was determined by minimizing the potential energy in vacuum using the CHARMM force field, with the constraints that all known atomic positions from the X-ray structure were fixed.

At ambient temperature and at physiological temperature, the oxidation rate constants of different methionine residues in a given protein, such as G-CSF (*16*) and human α1-antitrypsin (*17*), can differ by an order(s)-of-magnitude. Chu *et al.* explained this observation with the introduction of a new mechanism for methionine oxidation in proteins by hydrogen peroxide (*16, 18-20*). In this mechanism, water molecules play an

important role in stabilizing the transition state. They also found that (*16, 18-20*) solvent accessible area is not sufficient to distinguish the disparate oxidation rate constants among methionine residues, as previously thought. Instead, they proposed a new structural quantity, ensemble-averaged two-shell water coordination number, derived from molecular dynamics simulations, which was found to correlate well with oxidation rate constants. An obvious next question then is how local variations in amino acids near a given methionine site affect oxidation in the absence of structural effects. There are two ways of addressing this question: through equilibrium denaturation experiments and by studying peptides that have the same sequence in the local region around each methionine residues of interest. In this paper, we report the results of both kinds of studies, performed on the very important therapeutic protein, G-CSF. Comparative kinetics studies were conducted using rhG-CSF and synthetic peptides, which were synthesized to mimic the local sequence around methionine residues in rhG-CSF. The differences in methionine oxidation rate constants as a function of temperature were obtained from experimental measurements, and a correlation with free energies of equilibrium denaturation of rhG-CSF at different temperatures was made.

3.2 Materials and methods

3.2.1 Materials.

rhG-CSF expressed and produced in *E. coli* was provided by Amgen Inc. Three short peptides with sequences, which mimic the corresponding sequences in rhG-CSF around Met 122, Met 127, and Met 138, were synthesized by SynPep Co. The peptide sequences are Acetyl-QQMEEY-$CONH_2$ (denoted pep1 and corresponding to Met 122),

Acetyl-LGMAPY-$CONH_2$ (denoted pep2 and corresponding to Met 127), Acetyl-GAMPAY-$CONH_2$ (denoted pep3 and corresponding to Met 138), all of which were acetylated at the C-terminus and tagged with a tyrosine residue at the N-terminus. Ultrapure guanidinium hydrochloride was purchased from MP Biomedicals (cat. no. 105696). All other chemicals were purchased from Sigma Aldrich.

3.2.2 Oxidation kinetics measurement for peptides.

A mixture of 0.5 mg/ml rhG-CSF with equimolar concentrations of the three short peptides was prepared in 10 mM acetate buffer at pH 4.0. Ionic strength was adjusted to 100 mM by adding NaCl. Concentrated hydrogen peroxide (30% w/w H2O2) was added into the mixture solution at different concentrations at different temperatures. The actual concentration of the stock solution of H2O2 (labeled 30% w/w) purchased from Sigma-Aldrich was determined by following the uv-vis absorbance at 240nm using an extinction coefficient 43.6. (The detailed procedure is contained in the Production Information Sheet). In all cases, hydrogen peroxide is in large excess, with molar ratios of H2O2/rhG-CSF ranging from 1:500-1:3000. Reaction mixtures were incubated at a range of temperatures, from 4 °C to 50°C in a water bath. After adding hydrogen peroxide and incubating at a specified temperature, aliquots of reaction mixture were removed at different time points and the remaining hydrogen peroxide was quenched by catalase (Sigma c9284). Samples were then centrifuged, and there were no considerable soluble aggregates in these samples up to the end of oxidation as verified by size exclusion chromatography. In addition, there was full recovery with respect to total peak area, suggesting no loss due to insoluble materials. The supernatant was divided into two

halves, one for the quantification of the oxidation kinetics of the short peptides, and the other one for Glu-C peptide map of rhG-CSF.

The sample for the quantification of the oxidation kinetics of the short peptides was analyzed by RP-HPLC using an Agilent 1100 series or Beckman Coulter Gold with a C_4 column (Phenomenex, Jupiter 5μ C_4 300Å 150×4.6 mm). The mobile-phases were 0.1% TFA in water (A) and 90% acetonitrile with 0.1% TFA in water (B). The column was equilibrated with 10% B initially. After sample injection, the separation was achieved by a linear gradient of 10% B to 40% B for 20 min, 40% B to 80% B for 30 min, a constant concentration of B at 80% for 5 min, and back to 10% for another 10 min. The column temperature was controlled at 60°C, and the flow rate was 0.8 ml/min. The absorbance signal was monitored by a visible/ultraviolet diode array detector at a wavelength of 214 nm.

3.2.3 Peptide mapping and oxidation kinetics measurement for rhG-CSF.

The sample used for rhG-CSF quantification was first digested by Glu-C endoproteinase. Glu-C endoproteinase selectively cleaves at the C-terminal side of glutamate sites in protein molecules under reducing buffer conditions. The digestion buffer was comprised of 0.5M Tris, 0.5M Tris-HCl, 0.4M methylamine hydrochloride ($CH_3NH_2 \cdot HCl$), 0.2M DTT and 6M urea. For a typical aliquot of sample with a volume of 100 μl, 200 μl digestion buffer and 10 μl 0.2 μg/μl Glu-C endoproteinase were added, and the whole mixture was incubated at room temperature for 18±1 hrs. The digested sample was then analyzed by RP-HPLC with a C_4 column (Phenomenex, Jupiter 5μ C_4 300 Å 250×2.0 mm). The mobile-phases were 0.085% TFA in water (A) and 90% acetonitrile with 0.085% TFA in water (B). The column was equilibrated with 2% B

initially. After sample injection, the separation is achieved by a linear gradient of 2% to 30% B for 30 min, 30% B to 60% B for 45 min, 60% B to 98% B for 15 min, a constant concentration of B at 98% for 15 min, and back to 2% B for 12 min. The column temperature was controlled at 40°C with a flow-rate 0.2 ml/min. The absorbance signal was monitored by a visible/ultraviolet diode array at a wavelength of 214 nm. Most of the oxidation kinetics measurements were performed in duplicate, and it was found that the deviations from fitting to the pseudo-first order reaction kinetcs were greater than the variations between two different sets of time points. Therefore, the error bars in the reported oxidation rate constants were from the variations in the linear fit to pseudo-first order reaction kinetics.

LC/MS/MS data were acquired using a model Finnigan LCQ series mass spectrometer with electro spray interface and XCaliber software from Thermo Electron Co. The data were used to identify each of the digested fragments and the oxidized methionine-containing fragments. As shown in Figure 3.2, oxidized and unoxidized forms of these methionine-containing fragments can be well separated by HPLC. Among all the major peaks identified using MS/MS, no other oxidized residues, such as trp or tyr were found. Oxidized forms I, II, III, IV, V, and unoxidized forms by L125-E163, M1-E20, and L100-E124 are indicated. The areas under these peaks were used to calculate the oxidation rate constants.

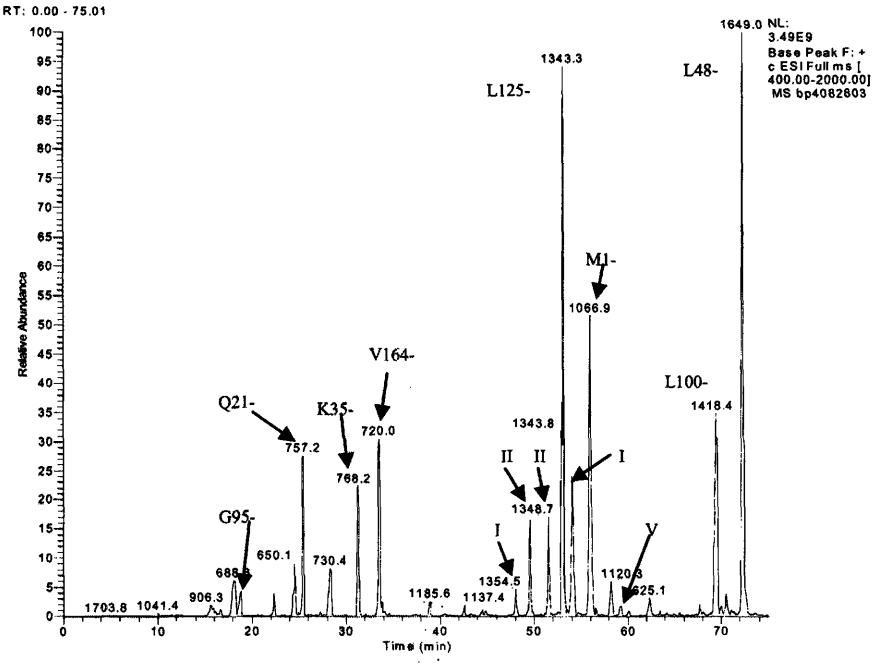


Figure 3.2: Mass spectrum of Glu-C digested peptide fragments. The number over each peak refers to the molecular weight of the fragment, with the actual fragment in rhG-CSF indicated by an arrow. I: L125-E163 with M138(O) and M127(O), II: L125-E163 with M127(O), III: L125-E163 with M138(O), IV: M1-E20 with M1(O), V: L100-E124 with M122(O).

3.2.4 Equilibrium denaturation monitored by intrinsic fluorescence.

Equilibrium denaturation by GdnHCl was performed for rhG-CSF in 10 mM acetate at pH 4.0 under a range of temperatures. An AVIV fluorometer model ATF105 with an automated titration system was used to generate fluorescence data. Two stock protein solutions at 0.5 mg/mL protein concentration were prepared with concentrated GdnHCl (approximately 8M) and with no denaturant. The titration system was comprised of a Hamilton pump, which removed a set volume of denatured protein solution from the cuvette and injected an equivalent volume of protein solution from the syringe, decreasing the concentration of GuHCl by 0.1 M for each spectral measurement. After the dilution of denaturant, the solution in the cuvette was allowed to mix and equilibrate for 3 minutes at room temperature. The fluorescence intensity was recorded

at an excitation wavelength 280 nm and emission wavelength 340 nm, per previous work (*21*).

3.2.5 Equilibrium denaturation monitored by Circular Dichoism (CD).

The same buffer condition as in fluorescence equilibrium denaturation except that rhG-CSF concentration was 0.02 mg/ml was used for CD experiments. A CD spectrophotometer model J810 from JASCO was used. A similar automated titration system was used to prepare mixtures with different denaturant concentrations. The CD signal at 222 nm was collected by SpectraManager software.

3.2.6 Data analysis of tryptophan fluorescence and CD data.

Data from equilibrium denaturation experiments were processed according to a standard method developed by Tanford (*22*). Raw CD and fluorescence data were fit to a two-state model for protein unfolding/folding.

3.3 Results

3.3.1 Temperature dependence of methionine oxidation kinetics for short peptides.

Since a large excess of H_2O_2 was used in the all of the oxidation experiments, its concentration is assumed to be a constant in the calculation of rate constants. As shown in Figure 3.3, the concentration of unoxidized peptides was normalized to their initial concentrations and plotted on a semi-logarithm scale. Linear fitting (with $R^2>0.98$) was performed on this plot to assure the validity of assumed pseudo-first order reaction and to obtain the pseudo-first order rate constants. The measured rates of methionine oxidation in the short peptides are listed in Table 3.1. Second order rate constants were obtained

from the division of pseudo-first order rate constants by the hydrogen peroxide concentration. Figure 3.4 shows that on the Arrhenius plots, the methionine oxidation rate is nearly sequence-independent in the three peptides. The oxidation rates of methionine residues in these hexa-peptides are very close to the oxidation rate of free methionine, suggesting that the primary sequence of these peptides has little effect on their oxidation rate.

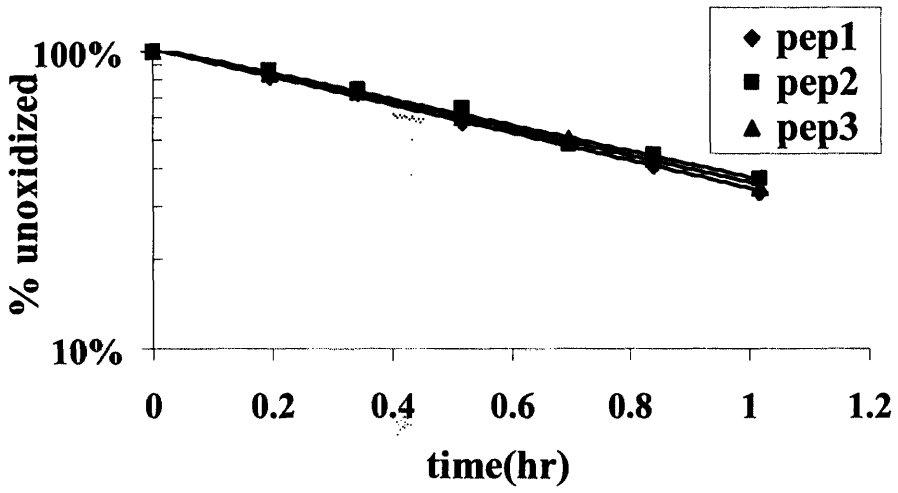


Figure 3.3: Oxidation rate constants of short peptides. This graph shows the unoxidized percentage of the three short peptides versus time, under the reaction condition in a pH4.0, 10mM sodium acetate buffer at 37°C. Linear regressions were performed to generate pseudo first order rate constants, from which second order oxidation rate constants were obtained.

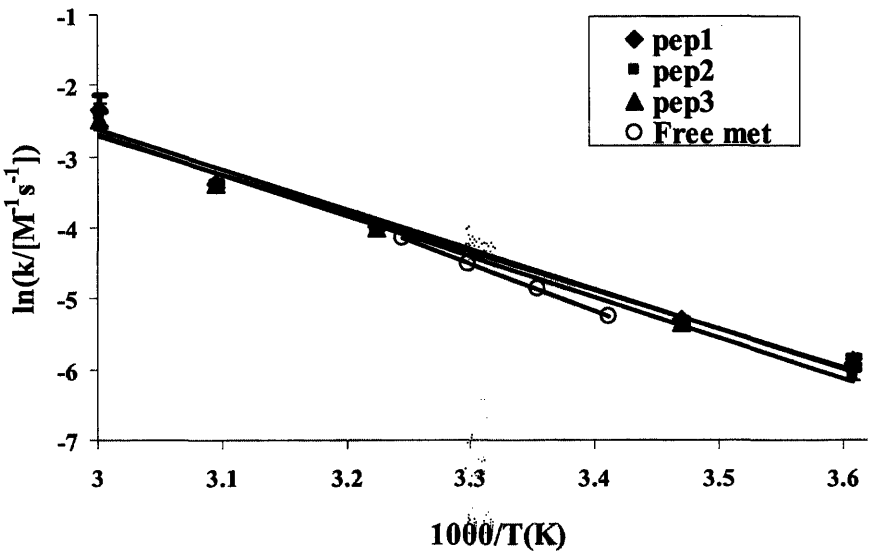


Figure 3.4: Arrhenius plots of the oxidation rate constants of methionine in short peptides as a function of temperature. A good linear relationship was obtained, and the apparent activation energies were obtained from the regression. Error bars are added from Table 3.1.

3.3.2 Temperature dependence of the oxidation kinetics of methionine residues in rhG-CSF.

Similar to the oxidation kinetics analysis for three peptides, the concentration of unoxidized methionine residues was normalized to their initial concentrations and plotted on a semi-logarithm scale, shown in Figure 3.5. Similar procedure as described for the three peptides was carried out to generate the rate constants shown in Table 3.1.

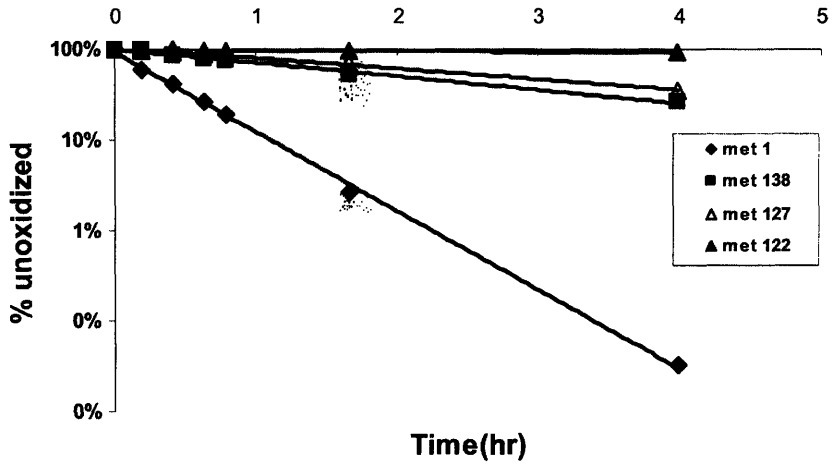


Figure 3.5: Oxidation rate constants determined for methionine residues in rhG-CSF. This graph shows the unoxidized percentage of each methionine residues in rhG-CSF, calculated from the peptide map, using the areas of the digested fragment(s). Linear regressions were performed to generate pseudo-first order rate constant, from which second order oxidation rate constants were obtained by dividing by hydrogen peroxide concentrations.

In Table 3.1, both the oxidation rate constants for the three short peptides and for the four methionine residues in rhG-CSF at different temperatures are shown. The oxidation rate constants vary by over two orders-of-magnitude amongst different methionine residues in rhG-CSF. In addition, the oxidation rate constants vary by over one order-of-magnitude in the temperature range 4-60°C for each given methionine residue.

Table 3.1: Second order oxidation rate constants of methionine residues in three short peptides and GCSF at different temperatures at pH 4.0, 10 mM NaAc buffer.

Temperature (°C)	Free met[b]	pep1	pep2	pep3	M1	M138	M127	M122
4		9.62	8.15	10.10	7.88	0.85	0.78	0.02
15		17.98	16.81	17.26	10.80	2.33	1.75	0.06
20	18.83				14.96	3.16	2.40	0.10
25[a]	28.19				16.50	6.47	2.73	0.63
30	39.33							
35	57.01							
37		67.59	63.01	65.19	40.61	6.50	5.07	0.38
45					60.38	21.61	19.19	1.83
50		123.28	114.25	122.19				

60	344.05	317.52	303.77

All data are with the unit M^{-1} hr^{-1}. [a] Results from (*16*) [b]Results from (*23*)
Corresponding methionines in peptides and in rhG-CSF: pep1-Met122, pep2-Met127, pep3-Met138

From the data in Table 3.1, Arrhenius plots were obtained and are shown in Figure 3.6. Linear regressions were performed according to Arrhenius equation $k = Ae^{-\frac{\Delta E}{RT}}$. The values of the Arrhenius parameters, activation energy ΔE and prefactor A, are listed in Table 3.2. In general, the apparent activation energies and pre-factors of three peptides are approximately equal within the experimental errors, while those of methionine resides in rhG-CSF vary significantly. Understanding these differences is essential, as explained later.

Table 3.2: Activation energies and prefactors of methionine oxidation in three short peptides and rhG-CSF at pH 4.0, 10mM NaAc buffer.

	Free met	pep1	pep2	pep3	M1	M138	M127	M122
ΔE (kcal/mol)	13.1±0.0	11.1±0.9	11.3±0.8	10.8±0.8	9.1±0.9	12.5±1.7	12.3±1.7	18.8±3.4
LN(A/(M^{-1} hr^{-1}))	25.5±0.5	14.3±1.5	14.5±1.4	13.6±1.3	10.2±1.6	14.5±2.8	13.9±2.9	22.1±5.8

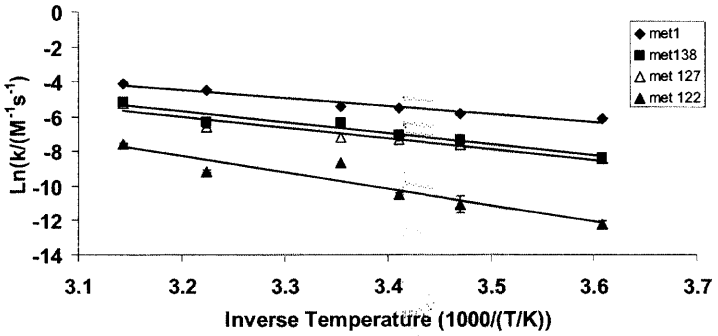


Figure 3.6: Arrhenius plots of the oxidation rate constants of methionine in short peptides as a function of temperature. A good linear relationship was obtained, and the apparent activation energies were obtained from the regression. Error bars are added from Table 3.1.

3.3.3 Equilibrium denaturation of rhG-CSF at different temperatures.

Both fluorescence and CD signals were used to track the structural changes in the equilibrium denaturation and renaturation processes. There are two tryptophan residues

in rhG-CSF, Trp 59 and Trp 119(*24*), that give rise to a fluorescence signal at 340 nm with an excitation wavelength of 280 nm. Figure 3.7 shows the raw fluorescence data that were converted to fraction of unfolded protein by fitting to a two-state model versus the concentration of GdnHCl for different temperatures. From a low temperature to 10-15°C, the curves shifted toward higher GdnHCl, and then backward to lower GdnHCl concentrations at higher temperature, therefore exhibiting a maximum stability temperature. A quantitative description of the position of each curve can be described by the midpoint denaturant concentration, C_m value (*22*) and m value.

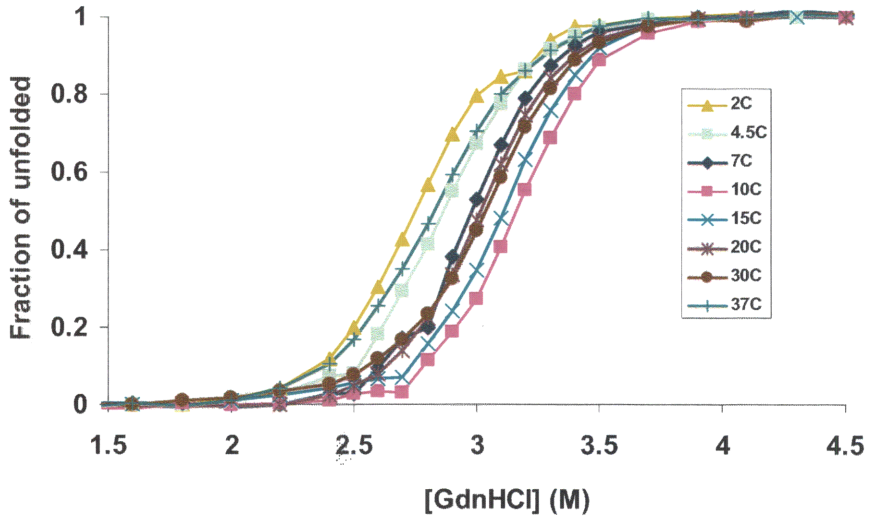


Figure 3.7: Fraction of unfolded versus denaturant GdnHCl concentration at different temperatures as indicated, converted and fitted from the CD signal at 222nm.

The results from CD and tryptophan fluorescence are listed in Table 3.3. The definitions of m value and C_m follow those of Tanford (*22*).

Table 3.3: Data from CD and tryptophan fluorescence following equilibrium denaturation of rhG-CSF at different temperatures in pH 4.0, 10 mM NaAc buffer.

Temperature (℃)	Tryptophan fluorescence[a]			CD results[b]		
	delta G (kcal/mol)	m value[c] (kcal/(M mol)	C_m value[c] (M)	delta G (kcal/mol)	m value[c] (kcal/(M mol)	C_m value[c] (M)

2	9.74	3.21	3.02	11.52	3.85	3.00
4	11.58	3.79	3.06	11.58	3.79	3.06
7	11.16	3.61	3.10	10.86	3.51	3.10
10	11.18	3.57	3.15	11.64	3.75	3.12
15	11.08	3.50	3.18	11.08	3.50	3.18
20[d]	9.00	2.85	3.15	11.19	3.19	3.50
25	9.06	2.94	3.09	9.06	2.94	3.09
30	9.25	3.05	3.05	8.63	2.87	3.01
34	7.42	2.55	2.92			
37	5.84	2.25	2.61	7.02	2.56	2.74

[a]Fluorescence data was recorded for a sample containing 0.5 mg/ml rhG-CSF in 10 mM NaAc at pH 4.0. [b]CD data were reorded for a sample containing 0.02 mg/ml rhG-CSF in 10 mM NaAc at pH 4.0. [c]Values follow from (*22*). D Compare with results from (*21*) with a delta G unfolding 9.0±0.3 kcal/mol under similar condition.

3.3.4 Fit to Gibbs-Helmholtz equation.

From a well-known thermodynamic relation, one has

$$\frac{\partial \frac{\Delta G}{T}}{\partial \frac{1}{T}})_P = \Delta G - T(\frac{\partial \Delta G}{\partial T})_P = \Delta G - T\Delta S = \Delta H \tag{3.1}$$

While for small molecules, it is quite safe to assume ΔH=constant

$$\frac{\partial \frac{\Delta G}{T}}{\partial \frac{1}{T}})_P = -R(\frac{\partial \log(K)}{\partial \frac{1}{T}})_P = \Delta H \tag{3.2}$$

which is the van't Hoff relationship. (An exothermic reaction has a positive slope on the $\log(K)$ versus $1/T$ plot because ΔH_{rxn}<0.)

When applying this to the protein unfolding process N↔U:

$$\frac{\partial \frac{\Delta G_{(unf)}}{T}}{\partial \frac{1}{T}})_P = \Delta H_{(unf)} \tag{3.3}$$

or equivalently

$$\frac{\partial \frac{\Delta G_{(unf)}}{T}}{\partial T})_P = -\frac{\Delta H_{(unf)}}{T^2} \tag{3.4}$$

Integrating this expression from T^* to T and assuming that the temperature dependence of $\Delta H_{(unf)}$ is $\Delta H_{(unf)} = \Delta H^*_{(unf)} + \Delta C_p(T - T^*)$ and $\Delta H^*_{(unf)}$, where the ΔC_p's are all constants over this temperature range,

$$\frac{\Delta G_{(unf)}}{T} - \frac{\Delta G^*_{(unf)}}{T^*} = \frac{\Delta H^*_{(unf)} - \Delta C_P T^*}{T} - \frac{\Delta H^*_{(unf)} - \Delta C_P T^*}{T^*} - \Delta C_P \log(\frac{T}{T^*}) \tag{3.5}$$

If taking T^* to be T_m, the melting temperature, then $\Delta G^*_{(unf)}$=0, so

$$\Delta G_{(unf)} = \Delta H^0_{(unf)} - T\frac{\Delta H^0_{(unf)}}{T_m} + \Delta C_P[T - T_m - T\log(\frac{T}{T_m})] \tag{3.6}$$

or for folding process U↔N

$$\Delta G_{(folding)} = \Delta H^0_{(folding)} - T\frac{\Delta H^0_{(folding)}}{T_m} + \Delta C_P[T - T_m - T\log(\frac{T}{T_m})] \tag{3.7}$$

The individual Gibbs free energy changes of folding at each temperature are plotted versus temperature according to the Gibbs-Helmholtz equation, as shown in Figure 3.8. In Table 3.4, the parameters obtained from the fitting to both CD and fluorescence data are listed. The difference is the manifestation that the two techniques track different events during the protein folding/unfolding process.

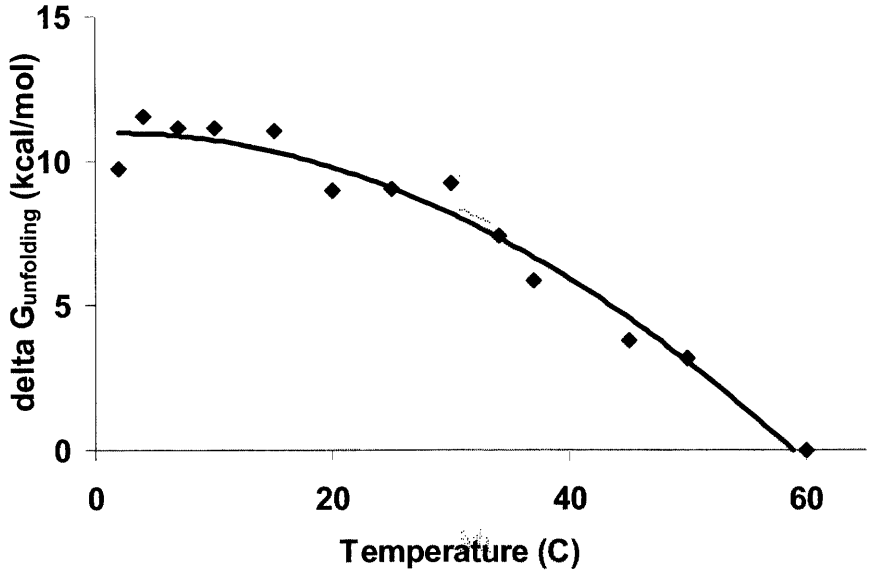


Figure 3.8: Fit of Gibbs free energy change versus temperature according to the Gibbs-Helmholtz equation. Results obtained from fluorescence equilibrium denaturation experiments. Results were also obtained from CD experiments. All equilibrium denaturation experiments were conducted in 10 mM sodium acetate buffer at pH 4.0.

Table 3.4: Parameters fitted from CD and tryptophan fluorescence data according to Gibbs-Helmholtz equation.

	$\Delta H^0_{(folding)}$ (kcal/mol)	T_m (K)	ΔC_p (kcal/(mol K))
Fluorescence	-1111.47	58.94	1.85
CD	-2037.95	60.63	3.28

3.4 Discussion

3.4.1 Comparative kinetic analysis on the oxidation of methionine residues.

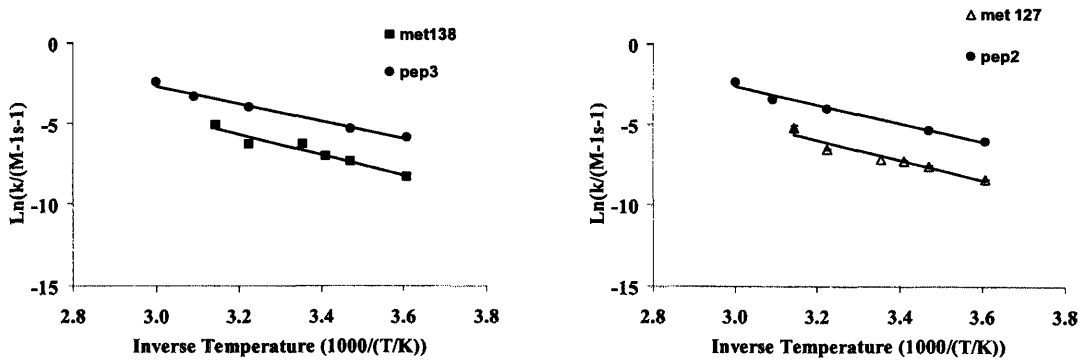
Short peptides each containing one methionine residue were designed to have the same amino acid sequences corresponding to the sequences around each methionine residue in rhG-CSF. Comparison between the oxidation kinetics of corresponding methionine residues in peptides and in the protein can provide information about how

protein tertiary structure influences oxidation reactions. Figure 3.9 shows Arrhenius plots of the oxidation rates of specific methionine residues within the intact protein versus the corresponding peptide. The oxidation rates within the intact protein were generally slower than within the corresponding peptides, presumably due to steric interference, reduced flexibility, and limited diffusion of reactive oxidation species to the methionine site. In addition, the extent of structural influence depends on the specific location of such residue, possibly determined by the micro-environment around it. Observing these differences, the questions we want to address are:

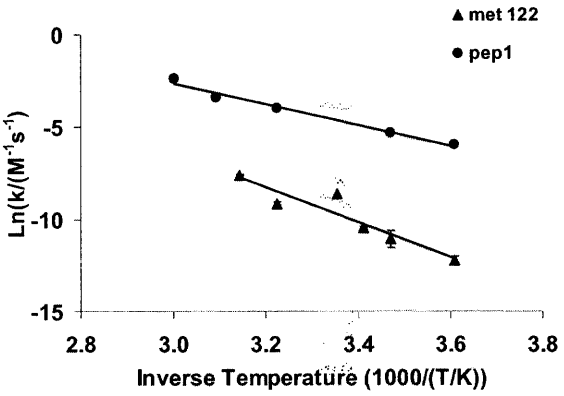
1. How can we understand the structural effects on oxidation kinetics?

2. Why does the additional complexity of structural effects that supposedly would result in non-Arrhneius oxidation kinetics still yield Arrhenius temperature dependence in the temperature ranged examined?

3. Can a simple phenomenological model(s) be developed to account for the structural effects on oxidation rate constants, including its temperature dependence? The model needs to agree with the experimental kinetic data for both peptides and protein. Moreover, a reasonable model needs to have the ability to be reduced back to the case of peptides when the protein loses its structure.



(a) (b)



(c)

Figure 3.9: Comparisons of the oxidation of methionine residues in rhG-CSF and those in corresponding peptides on the Arrehnius plot. (a) met 138 and pep3 (b) met 127 and pep2 (c) met 122 and pep1. Lines across each set of data points represent the fit to the Arrhenius equation.

3.4.2 Relationship between structure and oxidation kinetics at different temperatures.

Unlike reactions involving only small molecules, reactions of macromolecules such as proteins are complicated by their tertiary structure. A reaction involving a buried residue in the hydrophobic core is obviously much slower than one exposed freely in solvent, where another reactant can access it to form a reaction complex. The effect of temperature on the reaction kinetics is also different for reactions involving small molecules versus those involving macromolecules. Specifically, reactions involving macromolecules could involve significant changes in their conformation that in turn could influence the reactivity of a specific reaction group. The structural effect on macromolecular reactions has not been fully elucidated, in part because the structural

changes in macromolecules such as proteins are difficult to characterize. The structural changes can be classified into categories according to their relative magnitude:

1. Thermal motion — The energy change on the order of kT that results from finite temperature fluctuations due to the fact that protein molecules can assume a large number of nearly isoenergetic conformations, so called "conformational substates"(*25*).

2. Loss of partial/local structure — The energy changes greater than kT in which local structure undergoes a significant change, such as loss of helical structure or local denaturation.

3. Denaturation — A global structure change with much larger energy change where many interactions contribute to the energy difference.

All these changes will have different effects on reaction kinetics depending on where the specific reaction group is located.

Figure 3.10 is a hypothetic schematic showing the free energy versus some reaction coordinate in the course of methionine oxidation. Depending on whether or not a stable intermediate can form after the oxidant molecule accesses the reaction site, two situations are described. The origin of structural effects observed in experiments is explained by a reaction barrier, the height of which depends on the location and micro-environment within the protein structure. In the "oxidant-bound intermediate" model depicted in Figure 3.10 (a), a stable intermediate, [P-S···O] is formed between protein P with its thioether sulfur S in a methionine residue and oxidant molecule O. The protein structure poses a free energy barrier to the formation of the intermediate [P-S···O]. There might be multiple such barriers (not drawn in the figure) between the initial state, where oxidant molecules are free in solution and the final state, and where the stable

intermediate is formed and ready for the reaction to take place. Blue curves describe cases for methionine residue in short peptides, while black and red for the case when the methionine residue is more or less buried, respectively. Alternatively, in the "non oxidant-bound intermediate" model shown in Figure 3.10 (b), no stable intermediate is required for the intrinsic reaction to occur for oxidation of methionine residues under the influence of protein structure. Situations represented by (a) the "oxidant-bound intermediate" model and (b) the "non oxidant-bound intermediate" model in Figure 3.10 correspond to the phenomenological models, (a) and (b), respectively, in Figure 3.11.

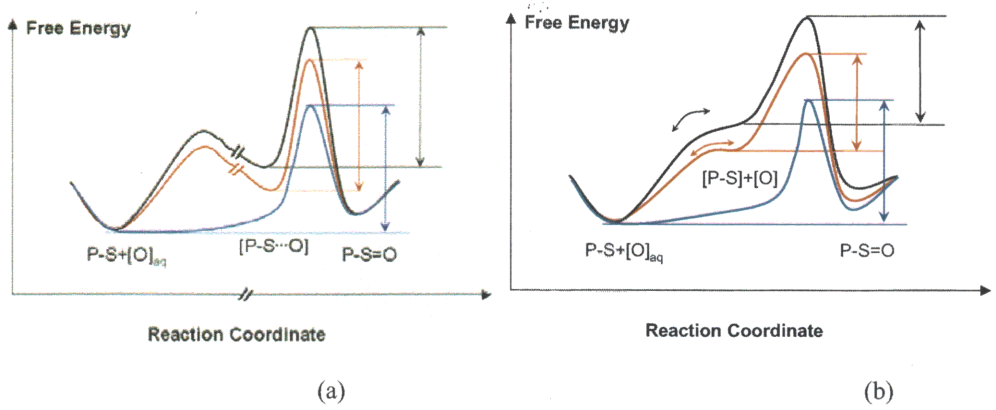


Figure 3.10: Free energy diagram of oxidation of methionine residues. (a) The "oxidant-bound intermediate" model (b) The "non oxidant-bound intermediate" model.

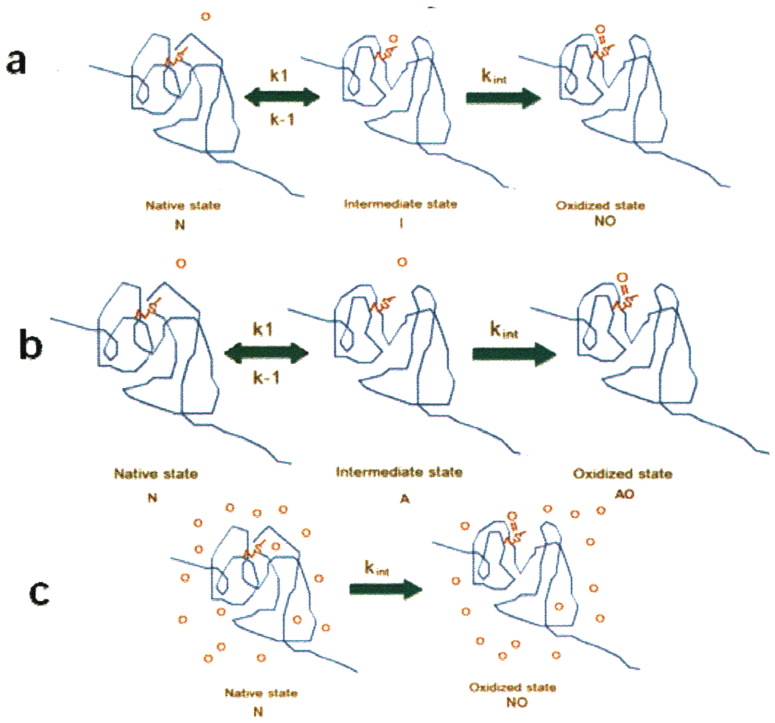


Figure 3.11: Phenomenological models that account for the influence of protein structure on oxidation kinetics. S represents the sulfur site in methionine residues, O represents a small molecule oxidizing reagent such as hydrogen peroxide, and S=O represents the methionine sulfoxide bond formed in the oxidation process. (a) "oxidant-bound intermediate" model (b) "non oxidant-bound intermediate" model (c) "effective oxidant concentration" model.

3.4.3 Phenomenological models for the relationship between protein structure and oxidation kinetics.

As mentioned before, the presence of conformational features for different methionine residues results in the disparate oxidation rates, which are also quite different from those in short peptides at a given temperature. Such phenomenological models can account for the conformational influence on oxidation kinetics and therefore result in a mechanistic description of the process. Three models are shown in Figure 3.11. In the "oxidant-bound intermediate" model shown in Figure 3.11 (a), the oxidation proceeds through an intermediate state which is a complex formed by the binding of oxidant O and

protein after the oxidant molecule gets close to the sulfur site. Alternatively, the "non oxidant-bound intermediate" model in (b) depicts the oxidation of sequestered methionine residues in a protein with a complex structure requiring the local structural changes. In the "effective oxidant concentration" model shown in (c), the oxidant concentration near the methionine site is not equal to its bulk concentration, but rather an effective concentration $[O]_{eff}$. Here, the equilibrium distribution of oxidant in- and outside the local region of the protein is described by a Gibbs free energy, ΔG, much like the preferential binding/exclusion Gibbs free energy (25).

For the models above, one can obtain the apparent second-order rate constants. (See the Appendix for the detailed derivations.)

For the "oxidant-bound intermediate" model (a),

$$r_{overall} = \frac{k_{\text{int}}}{[O]_0} \frac{1}{1+\frac{c^{\varnothing}}{[O]_0} e^{\frac{\Delta G}{RT}}} [O]_0([I]+[N]) \tag{3.8}$$

For the "non oxidant-bound intermediate" model (b),

$$r_{overall} = \frac{1}{1+e^{\frac{\Delta G^{(I)}_{unf}}{RT}}} k_{\text{int}}[O]_0([A]+[N]) \tag{3.9}$$

For the "effective oxidant concentration" model (c),

$$r_{overall} = (k_{\text{int}} e^{\frac{\Delta G_{exclusion}}{RT}})[O]_0[N] \tag{3.10}$$

Here, $[X]$ denotes the concentration of species X with the subscript meaning initial concentration (t=0), where X connotes O (oxidant), I (bound intermediate), N (native protein), or A (unbound intermediate), k_{int} is the intrinsic oxidation rate constant of methionine when there is no structural effect, or more specifically, the oxidation rate constant of methionine residue in the short peptide, $c^{\varnothing}$ is the concentration of the

reference solution, taken to be 1 M, and ΔG is the standard state Gibbs free energy change, with subscripts or superscript indicating different conditions, defined in the Appendix.

Therefore, under the condition that the oxidant is in large excess, the reaction can be considered to be pseudo-first order. Including structural effects decreases the apparent rate constant, consistent with the experimental observation above (Table 3.1). In addition to these observations, non-linear least-square fitting to three phenomenological models was performed. The objective function is defined as

$$\min_{\Delta H, \Delta C_p} (\sum_{T_i} (\ln(k_j(T_i)) - \ln(k_{j,T_i}))^2 \tag{3.11}$$

where k_{j,T_i} is the experimental rate constant for methionine residue j at temperature T_i, and $k_j(T_i)$ is the rate constant calculated from any model by substituting the Gibbs-Helmholtz equation for the Gibbs energy of structure effect, i.e.,

$$\Delta G = \Delta H - \frac{T}{T_m}\Delta H + \Delta C_p[T - T_m - T\log(\frac{T}{T_m})] \tag{3.12}$$

In the fitting, T_m is fixed at the melting temperature, 60°C, for rhG-CSF under the buffer condition under which the oxidation experiments were performed. The results of the least-square non-linear fit are shown in Figure 3.12 for each methionine residue. All three models fit the rate constants measured reasonably well.

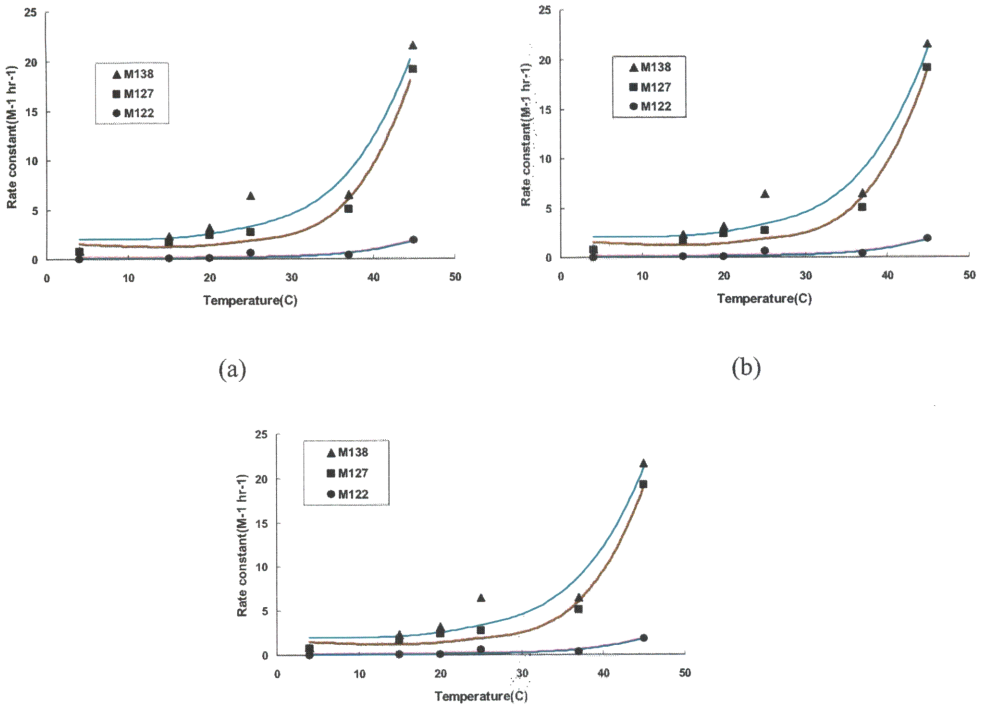


Figure 3.12: Results of least-square-fit to the phenomenological models. (a) the "oxidant-bound intermediate" model, (b) the "non oxidant-bound intermediate" model, and (c) the "effective oxidant concentration" model correspond to those presented in Figure 3.11 respectively. Lines across each set of data points represent the fit to the three models.

To distinguish the models based on experimental kinetics data, the rate-controlling process governing methionine oxidation needs to be identified. One difficulty in doing so is the lack of known functional form for the changes in Gibbs free energy associated with the structural effects. By fitting the experimental data, one can choose ΔC_p as any order of polynomial function of temperature to get an optimal degree of agreement (in Table 3.5 and Figure 3.12, ΔC_p was chosen as a linear function of

temperature). However, this approach does not provide mechanistic insight. Such insight could be obtained from molecular simulation, as in previous work (*16, 20*).

Table 3.5: Least square non-linear fit of model parameters

		$\Delta H^{(local)}_{(folding)}$ (kcal/mol)	$\Delta C_p{}^{(local)}_{(folding)}$ (kcal/ ° C mol)
Model a	met138	-22.7	-0.5
	met127	-28.1	-0.7
	met122	-52.7	-1.2
Model b	met138	-20.3	-0.5
	met127	-26.7	-0.7
	met122	-52.3	-1.1
Model c	met138	-22.7	-0.5
	met127	-28.1	-0.7
	met122	-52.7	-1.2

As shown in Figure 3.13, two different ways of plotting the data are used to reveal the structural effect on the oxidation of methionine residues in rhG-CSF, based on, for example, the "non oxidant-bound intermediate" model in Figure 3.11. We assume the intrinsic oxidation rate constant k_{int} in equation (9) to be that obtained from peptide oxidation, that is

$$k_{\text{int}} = k_{peptides} \tag{3.13}$$

The following expression can be easily derived

$$\ln\left(\frac{k_{peptide}}{k_{apparent}} - 1\right) = \frac{\Delta G^{(l)}_{unf}}{RT} \tag{3.14}$$

Therefore, by plotting $\ln\left(\frac{k_{peptide}}{k_{apparent}} - 1\right)$ versus $1/T$, the temperature dependence of $\Delta G^{(l)}_{unf}$ can be obtained, i.e. $\Delta G^{(l)}_{unf}$ as a function of T, shown in the left panel of Figure 3.13. In Figure 3.13 (b), the oxidation rate constant ratios of methionine residue in rhG-CSF with its corresponding peptide are plotted with a scaled Gibbs free energy of

denaturation. As the temperature increases, the oxidation rate ratios of the intact protein and the corresponding peptides also increase and approach the ideal value of 1 (although still much less than 1). Concomitantly, as the Gibbs free energy drops, the oxidation rate ratio approaches zero, which is its value at the melting temperature.

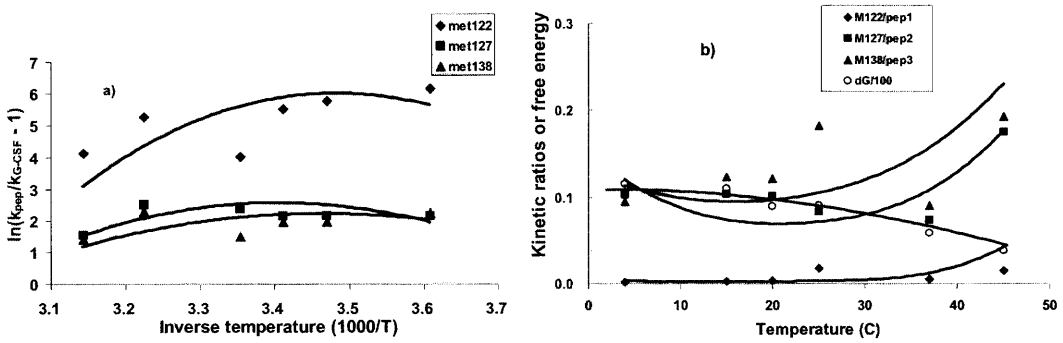


Figure 3.13: Comparative kinetic data between three short peptides and methionine residues in rhG-CSF. a) $\ln(\frac{k_{peptide}}{k_{apparent}} - 1)$ versus $1/T$ for three methionine residues in rhG-CSF. Solid lines represent fittings by equation (9) using parameters from Table 3.5. b) Ratios of oxidation rate constants of methionine residue in rhG-CSF with its corresponding peptide are plotted with a scaled Gibbs free energy of denaturation. Solid lines represent fittings by equation (9) using parameters from Table 3.5 and fitting by equation (7).

As temperature increases, the difference in the rate constants for oxidation of corresponding methionine residues in peptides and in protein becomes smaller, since the protein gradually loses its compact structure. At the point of thermal unfolding, the methionine residue has essentially the same micro-environment as that in the peptide, and the oxidation rate constants should become similar. Thus, we may consider the convergence of the lines in Figure 3.9 at some higher temperature. However, clearly the Arrhenius lines of Met 138 and Met 127 are almost parallel to the lines of their corresponding peptides, and the line of Met 122 will intersect with that of pep1 at a

temperature (~220°C) much higher than the melting temperature of rhG-CSF (~60°C). Therefore, we hypothesize that near the melting temperature of rhG-CSF, the structural effects are eliminated or at least minimized, and methionine oxidation kinetics will exhibit non-Arrhenius behavior, deviating from these Arrhenius lines significantly. Unfortunately, this is difficult to test experimentally for rhG-CSF due to the fact that rhG-CSF has a great propensity to aggregate at temperatures approaching the thermal melting temperature of 60°C.

The "non oxidant-bound intermediate" phenomenological model (b) can illustrate the expectations described above. Parameters fitted from oxidation rate constants in the experimental temperature range were used to extrapolate the Arrhenius lines to a temperature range near T_m and even beyond, as shown in Figure 3.14. Even though the "non oxidant-bound intermediate" model can produce the expected behavior near T_m and beyond, there is some sacrifice in the fitting of experimental data, as quantified in Table 3.6.

Table 3.6: Comparisons of standard deviations calculated from Arrhenius equation versus experimental rate constants and those calculated from model (b) versus those from experiments

	met138	met127	met122
std of ln(k)-ln(k_arrhenius)	0.28	0.28	0.58
std of ln(k)-ln(k_model b)	0.41	0.30	0.96

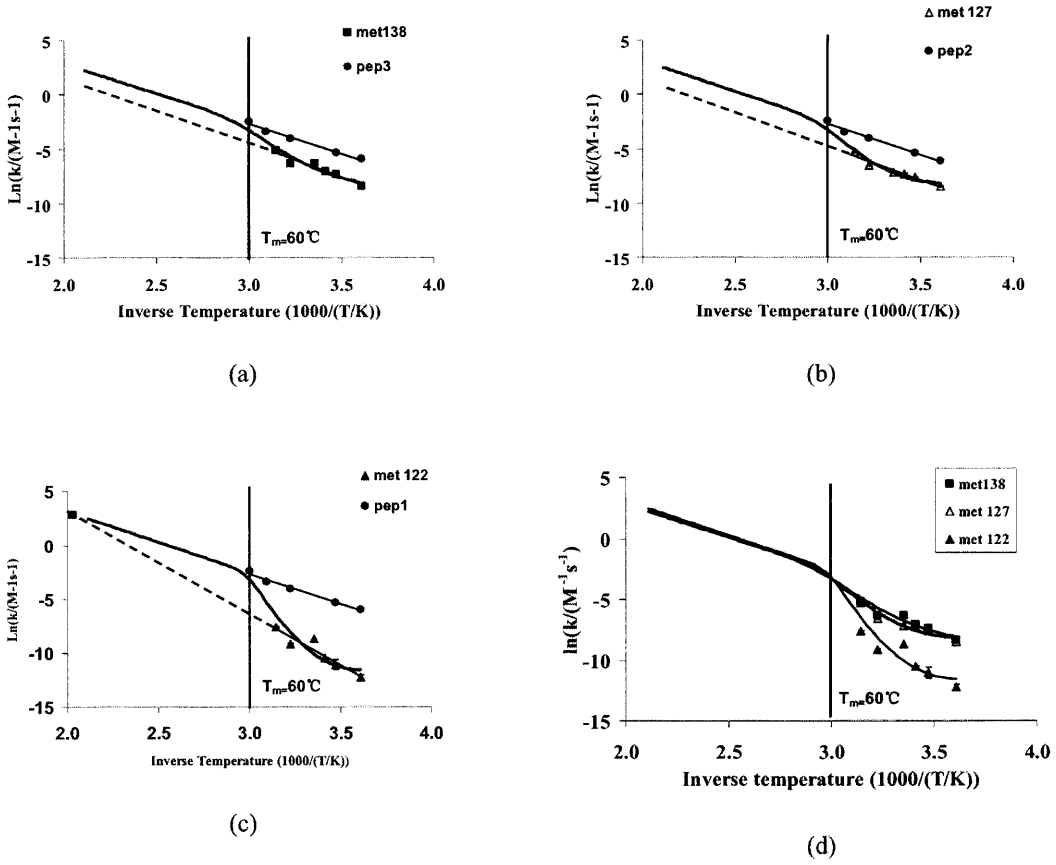


Figure 3.14: Comparison between Arrhenius fit versus the "non oxidant-bound intermediate" model fit. (a) Met 138 and pep3 (b) Met 127 and pep2 (c) Met 122 and pep1 (d) altogether. Dashed lines in (a), (b) and (c) represent the direct extrapolations of Arrhenius lines for methionine residues in rhG-CSF. Curved lines represent the predicted behavior by the "non oxidant-bound intermediate" model.

3.4.4 Analysis based on activation energy differences.

As shown in Table 3.2, there are statistically significant differences among the activation energies. These differences are hypothesized to be the manifestation of structural effects based on previous studies (*16*) in which it was concluded that without structural hindrance on the accessibility of water molecules, there would be equal activation barrier energies for all four methionines in rhG-CSF when oxidized by hydrogen peroxide. The "non oxidant-bound intermediate" model in Figure 3.11 is now

taken, as an example, to account for the differences in activation energies among methionine residues in rhG-CSF. Similar analyses can be done for the other models.

If one assumes that k_{int} follows the Arrhenius equation, $k_{\text{int}} = Ae^{\frac{\Delta E^{\dagger}}{RT}}$, where the constant A has the dimension s^{-1}, then:

$$k_{apparent} = \frac{Ae^{\frac{\Delta E^{\dagger}}{RT}}}{1+e^{\frac{\Delta G_{unf}^{(I)}}{RT}}} \tag{3.15}$$

Notice that there are 2 limits for this apparent rate constant expression.

1. when $-\Delta G_{unf}^{(I)} << RT$, then

$$k_{apparent} = Ae^{\frac{\Delta E^{\dagger}}{RT}} \tag{3.16}$$

2. when $-\Delta G_{unf}^{(I)} >> RT$

$$k_{apparent} = Ae^{\frac{\Delta E^{\dagger} + \Delta G_{unf}^{(I)}}{RT}} = Ae^{\frac{\Delta E_{apparent}}{RT}} \tag{3.17}$$

In both cases, the apparent reaction rate can be simplified to an Arrhenius equation. In case 2, the apparent activation energy contains the contribution from the local structural change. The activation energy difference between each methionine in rhG-CSF and that in the corresponding peptide can be treated as the local folding free energy. Actually this activation energy difference can be used to ascertain whether or not the methionine oxidation is locally or globally dependent on structure with respect to temperature, as follows:

$$\begin{cases} \text{No structural dependence, when } -\Delta G_{unf}^{(1)} << RT \\ \text{Local structural dependence, when } -\Delta G_{unf}^{(1)} \sim RT \\ \text{Global structural dependence, when } -\Delta G_{unf}^{(1)} >> RT \end{cases}$$

As for methionine oxidation in rhG-CSF, it can be seen (in Table 3.2) that the activation energy difference in Met 122 and Met 1 in rhG-CSF is at least 3 kcal/mol, which is much larger than *RT* (0.549 kcal/mol ~ 0.659 kcal/mol) over the temperature range tested (4 ~ 60 °C). Therefore, the latter extreme scenario is applicable in that the oxidation of Met 122 in rhG-CSF has global structural dependence. However, Met 127 and Met 138 display intermediate values of activation energy difference, which may imply that they have a local structural dependence on the oxidation reaction.

The analysis above can not account for the unexpected smaller activation energy of Met 1 in rhG-CSF. Perhaps the negative charge on the C-terminus impacts methionine oxidation, based on the difference between Met 1 and free methionine and the other methionine residues in rhG-CSF or peptides, where methionine residues are either negatively charged or bonded to neighboring residues. The protonation states of both N-terminus and C-terminus of methionine residue could affect the organization of water network, which plays an essential role in oxidation kinetics (*18, 26*), and give rise to differences in apparent Arrhenius parameters.

3.4.5 Implications for biochemistry.

To our knowledge, the dependence of chemical kinetics on temperature of the general type of reaction studied in this work has not been previously published. There is similar methionine oxidation work on the protein IL1-RA (*27*). Others studied the accessibility of cysteine residues in IL-1RA (*28*) and G-CSF and other proteins (*29, 30*).

Still in general there has been difficulty in the characterization of the complex structure of protein molecule and the structural effects on reaction rate. However, there are many examples where the complex structure of protein molecules plays a similar and essential role in chemical reactivity. Start & Stein (*31*) studied the alkylation of methionine residues in ribonuclease A by iodoacetate or iodoacetamide and found that only at low pH or in the presence of denaturant, rapid alkylation could occur. They also found that at the condition when the protein is unfolded, the alkylation rates for different methionine residues are nearly the same. This can be explained by the free energy barrier posed by protein structure being too large to overcome unless the structure of protein molecules is perturbed. An equivalent rate of reactivity at various sites within the unfolded protein indicates that the intrinsic reaction barrier is similar for different methionines. As another example, t-butyl hydrogen peroxide oxidizes methionine residues much slower than hydrogen peroxide (*32*). Despite possible differences in oxidizing capability, t-butyl hydrogen peroxide is larger in terms of molecular volume. This effectively increases the energy barrier posed by protein structure and also destabilizes the intermediate state. Liu et al. (*33*) found that testicular cytochrome c in the ferrous state has a slower oxidation rate by hydrogen peroxide than its counterpart in somatic cells due to changes in protein structure, as indicated by differing water patterns inside the protein structure and the interactions between the heme group and its surrounding residues. The origin of this difference is also expected to be explained in the framework of this study. In addition, the model and theory developed here were also tested on another protein IL1-RA (*27*), with which the only data of methionine oxidation rate constants as a function of temperature could be found, and were confirmed for their applicability (data not shown).

Experimentally measured rate constants can be fit with several plausible phenomenological models, such as presented in this paper. However, it is difficult to ascertain which model is correct due to the lack of knowledge about the structure dynamic or equilibrium properties of a local region around methionine residues in the complex structure of a protein molecule. A possible direction to obtain such information can be intrinsic fluorescence by a substituting tryptophan residue at the methionine site. The dynamics of local structure likely has a large influence on the reactivity and reaction rate. The understanding of local structural motions will not only be crucial to studies like methionine oxidation, but also can be applied to protein aggregation, ligand transport and binding, and enzymatic catalysis.

The models developed here were aimed at the understanding the "dual" role that temperature plays in methionine oxidation in protein molecules, that is, how temperature affects chemical kinetics intrinsically and affects protein conformation secondarily. These models have some minimal requirements, such as a reasonable explanation of the apparent Arrhenius behavior of oxidation of methionine residues in rhG-CSF over the temperature range examined, and restoration of the simple Arrhenius behavior when protein structural effects become minimal at temperatures much higher than melting temperature. Taking into account all of these requirements, the models are expected to cover a wider range of temperature. In essence, the modeling presented in the paper puts into proper context both the kinetics of oxidation and our understanding of protein conformational changes by presenting mathematically the physical basis of the oxidation of methionine residues in proteins. Because our models have a physical basis, we expect it to be generally applicable.

3.4.6 Implications for pharmaceutical shelf-life prediction.

An important criterion for protein pharmaceutical formulations is stabilization using an appropriate buffer, isoosmotic additives and other excipients under optimal temperature and pH conditions. Since the desired shelf-life is typically 18-24 months and degradation pathways such as oxidation usually occur very slowly over the course of that timeframe, it is highly desirable to have a rapid shelf-life prediction method for screening formulations. One way of accelerating oxidation is to elevate the temperature. Therefore, development of a model to predict oxidation at low temperature using data at high temperatures will facilitate predicting protein shelf-life. Quantifying the effect of protein structure on chemical kinetics is a prerequisite for such prediction. One major assumption with the model developed here is that the oxidation pathway using the oxidizing chemical species (H_2O_2) among reactive oxygen species is similar to oxidation in the absence of H_2O_2. Likewise, accelerated oxidation at higher temperature should be predictive of oxidation at lower temperatures. In addition, the model does not capture any other significant competing degradation pathways, such as aggregation, which is inevitably problematic in the shelf-life studies at elevated temperatures above storage temperature.

A critical unknown parameter in oxidation studies is the level of peroxides in the formulation buffer. In Table 3.7, the amount of rhG-CSF degradation is shown at various hydrogen peroxide levels (from micromolar to millimolar) and at different time lengths (from 6-month to 24-month) in storage buffer. The estimation is based on the experimental rate constants at a temperature 29°C from forced oxidation at which some long-term experimental data is available. Peroxide levels of several micromolar matches

experimental data reasonably well. This suggests that an accurate estimate of shelf-life can be made using appropriate concentrations of hydrogen peroxide and active oxygen species experimentally measured as inputs in addition to the kinetic data of accelerated oxidation at high temperatures. Thus, forced oxidation studies at elevated storage temperatures can guide rational formulation of protein pharmaceuticals against oxidative degradation.

Table 3.7: Degradation of rhG-CSF (in percentage) estimated from kinetic data at 29°C, 10mM acetate buffer at pH 4.0

time			6-month				12-month				24-month			
			M1	M138	M127	M122	M1	M138	M127	M122	M1	M138	M127	M122
Prediction from measured rate constant	hydrogen peroxide levels	0.4μM	4%	1%	1%	0%	8%	2%	2%	0%	16%	4%	3%	0%
		1μM	10%	3%	2%	0%	20%	5%	4%	0%	35%	10%	7%	1%
		10μM	66%	23%	18%	1%	89%	41%	32%	3%	99%	65%	54%	6%
		0.1mM	100%	93%	85%	13%	100%	99%	98%	25%	100%	100%	100%	44%
		1mM	100%	100%	100%	76%	100%	100%	100%	94%	100%	100%	100%	100%
Measured[b]		0.19μM active oxygen 0.38μM H_2O_2		2.5±1.0%[a]		2.6±1.4%		2.9±0.8%[a]		2.8±1.0%				

[a]Sum of percentage of oxidation of both M127 and Met138

[b]Long-term experimental data provided by AMGEN, Inc., performed using a commercial formulation buffer

3.5 Conclusions

In this work, we studied the temperature dependence of oxidation rate constants of methionine residues in several chemically synthesized peptides and in recombinant human granulocyte-colony stimulating factor (rhG-CSF) by hydrogen peroxide (H2O2). Experiments of the equilibrium denaturation of rhG-CSF also were conducted, and Gibbs free energies of unfolding as a function of temperature were calculated based on experimental data. We found significant variation among the oxidation rate constants for different methionine residues in rhG-CSF and as a function of temperatures. The rate constants for each methionine residue can be fit reasonably well according to the Arrhenius equation. This suggests that degradation is governed by the intrinsic oxidation reaction rather than a local conformational event having a complex temperature dependence. We also found that if we assume the existence of an additional activation

free energy barrier due to the transport of the oxidant molecule H2O2, a more complicated, non-Arrhenius equation ensues. However, this equation simplifies to the Arrhenius equation under certain circumstances. We classified the methionine residues in rhG-CSF according to the degree to which the protein structure affects oxidation kinetics, i.e., no structural dependence, local structural dependence and global structural dependence. Additionally, three models were developed to consider the structural effect of protein molecules on the oxidation of methionine residues. We found these models can fit the experimental data equally well. The "non oxidant-bound intermediate" model, in particular, can produce the anticipated temperature dependence of rate constants near the melting temperature and even beyond, when the influence of protein structure becomes diminishingly small. However, the phenomenological models we considered can not be distinguished based purely on phenomenological rate constant data. Trusted local dynamical information such as that which could be determined from molecular simulations would be needed. An example was shown of the shelf-life prediction of protein pharmaceuticals using the temperature dependence of oxidation rate and model prediction matched stability data well.

3.6 Appendix

For the three phenomenological models shown in Figure 3.11, the overall rate expressions can be obtained as follows.

In the "oxidant-bound intermediate" model (a), the elementary steps hypothesized in the model are as follows:

$$O + N \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} I$$

$$I \xrightarrow{k_{int}} NO$$

in which the formation of the intermediate complex of the protein and oxidant is assumed to be reversible. Furthermore the equilibrium of the first step can be expressed as

$$\frac{[I]}{[N][O]} = \frac{k_1}{k_{-1}} = \frac{K}{c^{\varnothing}} = \frac{e^{\frac{\Delta G}{RT}}}{c^{\varnothing}} \tag{3.18}$$

where $c^{\varnothing}$ is the concentration of the reference solution, taken to be 1 M, and ΔG is the standard state Gibbs free energy change.

Mass conservation:

$$[I]+[N]+[NO]=[N]_0 \tag{3.19}$$

Rate expression:

$$r_{overall} = \frac{d[NO]}{dt} = k_{int}[I] \tag{3.20}$$

So that

$$[I] = \frac{[N]_0 - [NO]}{1 + \frac{c^{\varnothing} e^{\frac{\Delta G}{RT}}}{[O]}} \tag{3.21}$$

And

$$r_{overall} = \frac{d[NO]}{dt} = k_{int}[I] = k_{int} \frac{[N]_0 - [NO]}{1 + \frac{c^{\varnothing} e^{\frac{\Delta G}{RT}}}{[O]}} \tag{3.22}$$

with the initial condition $[NO]_0$=0, and when oxidant is in much large excess $[N]_0 << [O] \approx [O]_0$. Integration yields

$$[NO] = [N]_0 \left(1 - e^{-\frac{k_{int}}{1+\frac{c^{\varnothing} e^{\frac{\Delta G}{RT}}}{[O]_0}} t}\right) \tag{3.23}$$

$$[N]+[I]=[N]_0 e^{\frac{k_{int}}{1+\frac{c^{\varnothing} e^{\frac{\Delta G}{RT}}}{[O]_0}} t} \tag{3.24}$$

$$r_{overall}=\frac{d[NO]}{dt}=k_{int}\frac{1}{1+\frac{c^{\varnothing} e^{\frac{\Delta G}{RT}}}{[O]_0}}([I]+[N])=\frac{k_{int}}{[O]_0}\frac{1}{1+\frac{c^{\varnothing}}{[O]_0}e^{\frac{\Delta G}{RT}}}[O]_0([I]+[N]) \tag{3.25}$$

For the "non oxidant-bound intermediate" model (b), the assumed elementary steps hypothesized in the model are as follows:

$$N \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} A$$

$$A+O \xrightarrow{k_{int}} AO$$

If the structural change is very fast and in equilibrium, the intrinsic oxidation will be the rate determining step, leading to:

$$\frac{[A]}{[N]}=\frac{k_1}{k_{-1}}=K=e^{-\frac{\Delta G^{(I)}_{unf}}{RT}} \tag{3.26}$$

$$[A]+[N]+[AO]=[P]_0 \tag{3.27}$$

One also has:

$$r_{overall}=\frac{d[AO]}{dt}=k_{int}[A][O] \tag{3.28}$$

$$[A]=\frac{k_1([P]_0-[AO])}{k_1+k_{-1}}=\frac{[P]_0-[AO]}{1+e^{\frac{\Delta G^{(I)}_{unf}}{RT}}} \tag{3.29}$$

Therefore

$$r_{overall}=\frac{d[AO]}{dt}=\frac{k_{int}}{1+e^{\frac{\Delta G^{(I)}_{unf}}{RT}}}([P]_0-[AO])[O]=\frac{k_{int}}{1+e^{\frac{\Delta G^{(I)}_{unf}}{RT}}}([A]+[N])[O] \tag{3.30}$$

It can be seen that the apparent rate constant in this model is

$$k_{apparent} = \frac{k_{int}}{1 + e^{\frac{\Delta G_{unf}^{(I)}}{RT}}} \tag{3.31}$$

For the "effective oxidant concentration" model (c), the oxidant concentration near the methionine site is not equal to its bulk concentration, but rather an effective concentration $[O]_{eff}$. The equilibrium distributions of oxidant inside and outside the protein are described by a Gibbs free energy ΔG, much like the preferential binding/exclusion Gibbs free energy, but used to describe the different distributions of co-solute near protein surface and in bulk solvent.

Therefore, concentrations of oxidant in bulk solvent $[O]_0$ and that near methionine site have the equilibrium relationship

$$\frac{[O]_{eff}}{[O]_0} = e^{-\frac{\Delta G_{exclusion}}{RT}} \tag{3.32}$$

Thus, the phenomenological oxidation rate becomes

$$r_{overall} = k_{int}[O]_{eff}[N] = (k_{int}[O]_0 e^{-\frac{\Delta G_{exclusion}}{RT}})[N] \tag{3.33}$$

Here ΔG contains the information of protein complex structure. Therefore, it can be a complex function of temperature.

3.7 References

(1) Chao, C. C., Ma, Y. S., and Stadtman, E. R. (1997) Modification of protein surface hydrophobicity and methionine oxidation by oxidative systems. *Proceedings of the National Academy of Sciences of the United States of America 94*, 2969-2974.

(2) Stadtman, E. R. (2001) Protein oxidation in aging and age-related diseases. *Healthy Aging for Functional Longevity 928*, 22-38.

(3) Trifunovic, A., Hansson, A., Wredenberg, A., Rovio, A. T., Dufour, E., Khvorostov, I., Spelbrink, J. N., Wibom, R., Jacobs, H. T., and Larsson, N. G. (2005) Somatic mtDNA mutations cause aging phenotypes without affecting reactive oxygen species production. *Proc Natl Acad Sci U S A 102*, 17993-8.

(4) Hokenson, M. J., Uversky, V. N., Goers, J., Yamin, G., Munishkina, L. A., and Fink, A. L. (2004) Role of individual methionines in the fibrillation of methionine-oxidized alpha-synuclein. *Biochemistry 43*, 4621-4633.

(5) Markesbery, W. R. (1997) Oxidative stress hypothesis in Alzheimer's disease. *Free Radical Biology and Medicine 23*, 134-147.

(6) Watson, A. A., Fairlie, D. P., and Craik, D. J. (1998) Solution structure of methionine-oxidized amyloid beta-peptide (1-40). Does oxidation affect conformational switching? *Biochemistry 37*, 12700-12706.

(7) Vogt, W. (1995) Oxidation of Methionyl Residues in Proteins - Tools, Targets, and Reversal. *Free Radical Biology and Medicine 18*, 93-105.

(8) Ciorba, M. A., Heinemann, S. H., Weissbach, H., Brot, N., and Hoshi, T. (1997) Modulation of potassium channel function by methionine oxidation and reduction. *Proceedings of the National Academy of Sciences of the United States of America 94*, 9932-9937.

(9) Wei, W. (1999) Instability, stabilization, and formulation of liquid protein pharmaceuticals. *International Journal of Pharmaceutics 185*, 129-188.

(10) Anbanandam, A., Urbauer, R. J. B., Bartlett, R. K., Smallwood, H. S., Squier, T. C., and Urbauer, J. L. (2005) Mediating molecular recognition by methionine oxidation: Conformational switching by oxidation of methionine in the carboxyl-terminal domain of calmodulin. *Biochemistry 44*, 9486-9496.

(11) Kim, Y. H., Berry, A. H., Spencer, D. S., and Stites, W. E. (2001) Comparing the effect on protein stability of methionine oxidation versus mutagenesis: steps toward engineering oxidative resistance in proteins. *Protein Engineering 14*, 343-347.

(12) Nguyen, T. H. (1994) Oxidation Degradation of Protein Pharmaceuticals, in *Formulation and Delivery of Proteins and Peptides* (Cleland, J. L., and Langer, R., Eds.) pp 59-71, American Chemical Society, Washington DC.

(13) Souza, L. M., Boone, T. C., Gabrilove, J., Lai, P. H., Zsebo, K. M., Murdock, D. C., Chazin, V. R., Bruszewski, J., Lu, H., Chen, K. K., Barendt, J., Platzer, E., Moore, M. A. S., Mertelsmann, R., and Welte, K. (1986) Recombinant Human Granulocyte Colony-Stimulating Factor - Effects on Normal and Leukemic Myeloid Cells. *Science 232*, 61-65.

(14) Aritomi, M., Kunishima, N., Okamoto, T., Kuroki, R., Ota, Y., and Morikawa, K. (1999) Atomic structure of the GCSF-receptor complex showing a new cytokine-receptor recognition scheme. *Nature 401*, 713-718.

(15) Hill, C. P., Osslund, T. D., and Eisenberg, D. (1993) The Structure of Granulocyte-Colony-Stimulating Factor and Its Relationship to Other Growth-Factors. *Proceedings of the National Academy of Sciences of the United States of America 90*, 5167-5171.

(16) Chu, J. W., Yin, J., Wang, D. I. C., and Trout, B. L. (2004) Molecular dynamics simulations and oxidation rates of methionine residues of granulocyte colony-stimulating factor at different pH values. *Biochemistry 43*, 1019-1029.

(17) Griffiths, S. W., and Cooney, C. L. (2002) Relationship between protein structure and methionine oxidation in recombinant human alpha 1-antitrypsin. *Biochemistry 41*, 6245-6252.

(18) Chu, J. W., and Trout, B. L. (2004) On the mechanisms of oxidation of organic sulfides by H2O2 in aqueous solutions. *Journal of the American Chemical Society 126*, 900-908.

(19) Chu, J. W., Yin, J., Brooks, B. R., Wang, D. I. C., Ricci, M. S., Brems, D. N., and Trout, B. L. (2004) A comprehensive picture of non-site specific oxidation of methionine residues by peroxides in protein pharmaceuticals. *Journal of Pharmaceutical Sciences 93*, 3096-3102.

(20) Chu, J. W., Yin, J., Wang, D. I. C., and Trout, B. L. (2004) A structural and mechanistic study of the oxidation of methionine residues in hPTH(1-34) via experiments and simulations. *Biochemistry 43*, 14139-14148.

(21) Brems, D. N. (2002) The kinetics of G-CSF folding. *Protein Science 11*, 2504-2511.

(22) Tanford, C. (1970) Protein denaturation. C. Theoretical models for the mechanism of denaturation. *Adv Protein Chem 24*, 1-95.

(23) Yin, J., Chu, J. W., Ricci, M. S., Brems, D. N., Wang, D. I. C., and Trout, B. L. (2005) Effects of excipients on the hydrogen peroxide-induced oxidation of methionine residues in granulocyte colony-stimulating factor. *Pharmaceutical Research 22*, 141-147.

(24) Lu, H. S., Klein, M. L., and Lai, P. H. (1988) Narrow-Bore High-Performance Liquid-Chromatography of Phenylthiocarbamyl Amino-Acids and Carboxypeptidase-P Digestion for Protein C-Terminal Sequence-Analysis. *Journal of Chromatography 447*, 351-364.

(25) Frauenfelder, H., Sligar, S. G., and Wolynes, P. G. (1991) The Energy Landscapes and Motions of Proteins. *Science 254*, 1598-1603.

(26) Chu, J. W., Brooks, B. R., and Trout, B. L. (2004) Oxidation of methionine residues in aqueous solutions: Free methionine and methionine in granulocyte colony-stimulating factor. *Journal of the American Chemical Society 126*, 16601-16607.

(27) Thirumangalathu, R., and Carpenter, J. *private communication and to be published.*

(28) Roy, S., Katayama, D., Dong, A. C., Kerwin, B. A., Randolph, T. W., and Carpenter, J. F. (2006) Temperature dependence of benzyl alcohol- and 8-anilinonaphthalene-1-sulfonate-induced aggregation of recombinant human interleukin-1 receptor antagonist. *Biochemistry 45*, 3898-3911.

(29) Pipes, G. D., Kosky, A. A., Abel, J., Zhang, Y., Treuheit, M. J., and Kleemann, G. R. (2005) Optimization and applications of CDAP labeling for the assignment of cysteines. *Pharmaceutical Research 22*, 1059-1068.

(30) Raso, S. W., Abel, J., Barnes, J. M., Maloney, K. M., Pipes, G., Treuheit, M. J., King, J., and Brems, D. N. (2005) Aggregation of granulocyte-colony stimulating factor in vitro involves a conformationally altered monomeric state. *Protein Science 14*, 2246-2257.

(31) Stark, G. R., and Stein, W. H. (1964) Alkylation Of The Methionine Residues Of Ribonuclease In 8 M Urea. *J Biol Chem 239*, 3755-61.

(32) Lu, H. S., Fausset, P. R., Narhi, L. O., Horan, T., Shinagawa, K., Shimamoto, G., and Boone, T. C. (1999) Chemical modification and site-directed mutagenesis of methionine residues in recombinant human granulocyte colony-stimulating factor:

Effect on stability and biological activity. *Archives of Biochemistry and Biophysics 362*, 1-11.

(33) Liu, Z., Lin, H., Ye, S., Liu, Q. Y., Meng, Z., Zhang, C. M., Xia, Y., Margoliash, E., Rao, Z., and Liu, X. J. (2006) Remarkably high activities of testicular cytochrome c in destroying reactive oxygen species and in triggering apoptosis. *Proc Natl Acad Sci U S A 103*, 8965-70.

Chapter 4. Molecular Mechanism of Hydrolysis of Peptide Bonds at Neutral pH

4.1 Introduction

Covalent degradation of proteins can occur by a multitude of different mechanisms, such as oxidation(*1, 2*), deamidation(*3, 4*), hydrolysis of peptide bonds on the polypeptide backbone(*5*), and other degradations(*6*). Monoclonal antibodies, a primary modality for biopharmaceuticals(*7*), have been reported by several different groups to undergo non-enzymatic fragmentation in the hinge region(*8-10*). This finding was identified mainly as the hydrolysis of several peptide bonds in the hinge. We believe that understanding the underlying reaction mechanism, particularly from a molecular perspective, will help to direct formulation development and antibody engineering efforts to minimize the extent of degradation in a rational and more efficient way. An example of this approach is the discovery of water interactions as a key parameter in the oxidation mechanism, which led to the formulation strategy of manipulating solvent accessibility to control methionine oxidation(*1*).

There have been numerous experimental and computational studies on the hydrolysis of peptide bonds in aqueous solutions due to its important biological relevance. To elucidate enzyme proficiencies in catalyzing peptide bond hydrolysis, a number of groups(*5, 11-13*) extensively characterized the reaction rates of hydrolysis of peptide bonds. It was found that in general, the reaction rate at neutral pH is extremely slow, with a half-time corresponding to hundreds of years. However, this reaction occurs much faster in both acidic and basic conditions, as studied by Smith et al(*14*). They extensively

mapped the pH and concentration dependences of reaction rates for the hydrolytic reaction of a peptide bond in N-(phenylacetyl)glycyl-D-valine (PAGV) and identified three regimes of reaction mechanisms: an acid-catalyzed mechanism for pH<4 with a first-order reaction rate in concentrations of both PAGV and H^+, a base-catalyzed mechanism for pH>10 with a first-order reaction rate in concentrations of both PAGV and OH^-, and a water-mediated mechanism for pH values in between with a first-order reaction rate only in the concentration of PAGV.

A number of computational efforts(*15-17*) were devoted to understanding the energetics of reactants, products, and intermediate species involved in the hydrolytic reaction. However, these studies showed only a static picture of the reaction process. Many other analyses included dynamic simulations. Stranton et al.(*18*) used combined quantum mechanical and classical mechanical calculations to study the hydrolytic reaction of peptide bonds catalyzed by trypsin in solution; free energies of the reaction were reported. Zahn et al.(*19-23*) studied the hydrolytic reaction of N-MAA in various solution conditions, namely acidic, basic and neutral pHs. In these *ab initio* calculations of the potential of mean force, reaction coordinates were assumed and the projection of the free energy hyper-surface onto these coordinates was performed using constrained molecular dynamics. However, the assumed reaction coordinates were never tested to determine if they were the correct ones.

Even though the hydrolytic reaction occurs slowly at pH 4-10, hydrolysis of therapeutic antibodies can be significant in physiological conditions during circulation or within the formulation solution conditions over shelf-life. Cordoba et al.(*9*) reported a several-percent hydrolytic degradation of residues in the hinge region of IgG1 antibody

molecules over an incubation period of three months. Their study showed that the hydrolysis of the polypeptide backbone could occur at multiple sites in the hinge to varying extents. They also showed that the reaction was non-enzymatic around neutral pH. Using optimized reversed-phase methods coupled with mass spectrometry, Dillon et al.(*10*) reached similar conclusions about the location and extent of hydrolytic cleavage of peptide bonds in the hinge region for IgG2 antibodies. Furthermore, the resulting fragments subsequently associated to form high molecular weight species via a clip-mediated aggregation mechanism(*24, 25*), which can be the primary degradation pathway at elevated temperatures for IgG2 antibodies. Thus, from a scientific and practical standpoint, elucidation of the underlying mechanism of the hydrolytic reaction is essential.

Distinct mechanisms of the hydrolytic reaction have been proposed in the literature for different solution pHs. In both acidic and basic conditions, the hydrolytic reaction is accelerated by the protonation of the carbonyl oxygen or the addition of a hydroxyl group onto the carbonyl carbon in the peptide bond, respectively. However, under neutral pH conditions (with a pH range 4-10, for example, on a di-peptide model), the hydrolytic reaction is extremely slow, having an exceptionally high reaction barrier of 27-30 kcal/mol, extrapolated from the experimental data(*9, 12, 14*). We are interested in this "neutral pH" range (pH 4-10) for therapeutic proteins, since it represents the most relevant pH conditions for formulation applications and physiological environment.

In this work, we study the hydrolytic reaction of a peptide bond using N-methyl acetyl acrylamide (N-MAA) as a model compound, in which two methyl groups are the minimal but computationally tractable constituents on the peptide bond –NH-CO-. *Ab*

initio molecular dynamics simulations at finite temperature equipped with transition path sampling (TPS)(*26-28*), likelihood maximization(*29-31*) and p_B histogram analysis(*28*) techniques were utilized to gain an understanding of the reaction mechanism, including identification and verification of the reaction coordinate.

What we mean by the reaction coordinate here is a descriptor of the real physical progress of the transition from the reactant state to the product state. The reaction coordinate as defined has to be contrasted with order parameters, which serve mainly to distinguish the two ending states. An order parameter (OP) is a quantity whose value can be used to distinguish different thermodynamic states, for example, reactant or product states, and crystalline, amorphous or liquid states. A reaction coordinate has to be an order parameter, while the contrary is not necessarily true. Identification of the "correct" reaction coordinate from a collection of order parameters is a very challenging problem, especially when the transition is complex. However, the information about the "correct" reaction coordinate is necessary when computing quantities of interest, such as free energy barriers and reaction rate constants, from molecular simulations. Also, we believe that the knowledge of the "correct" reaction coordinate, and hence the reaction mechanism, can provide essential information on the molecular level for judicious engineering of complex systems.

The committor probability $p_B(\mathbf{x})$, which can be interpreted as the probability that a trajectory initiated with Maxwell-Boltzmann distributed momenta at the configurational vector $\mathbf{x}$ reaches the product state B before reaching the reactant state A, can be used to best describe the mechanism of the transition during an activated process, thus serving as the "true" reaction coordinate(*32*). As illustrated by Metzner et al.(*33*) and E et al.(*34*),

various quantities involving $p_B(\mathbf{x})$, such as the probability current of reactive trajectories and the average frequency of reactive trajectories, allow one to fully characterize the statistical properties of the transition trajectories in the trajectory space, and thus to compute quantities of interest such as reaction rate constants. However, in general, the $p_B(\mathbf{x})$ is costly to compute and is a function of the highly dimensional configurational vector $\mathbf{x}$; it provides no insight into the physical characteristics about the transition dynamics. Therefore, approximations of $p_B(\mathbf{x})$ with a lower dimensional (preferably a one-dimensional) descriptor, involving physical quantities such as bond lengths, dihedral angles, bonding number(*35*), density fluctuation(*36*), etc., are required to describe the transition process in providing essential physical insights.

To our knowledge, we are the first to conduct this very detailed analysis of complex chemical reactions in a condensed-matter system using path sampling techniques combined with techniques for the determination of the reaction coordinate. The approach can be directly applied to other types of transitions in which direct transition dynamics are infrequent and the transition state is short-lived.

4.2 Overview

At finite temperatures, no single transition trajectory, a series of points in the phase-space connecting the reactant and product states, can be used to describe the whole reaction process due to thermal fluctuations. In trajectory space, each transition path is associated with a statistical weight(*28*), contributing to the experimentally observed transition process. It is therefore necessary to collect an ensemble of reactive trajectories and calculate the probability distribution according to their statistical weights in order to calculate quantities of interest that can be compared with experiments. The technique

that implements this as a Monte Carlo procedure in trajectory space is transition path sampling (*26-28*)(TPS).

Transition path sampling is well-suited for studying transition processes that have time-scale separation and exhibit a rough potential energy surface. Time-scale separation is often the characteristic of activated processes, where two stable states, the reactant state A and the product state B, are separated by a large free energy barrier. In rare events, such as chemical reactions, formation of critical nuclei during crystallization processes, and the protein-folding process, the time scale for the system to wander in the valley of the stable states is much longer than the time scale in which the transition dynamics occur. Therefore, a direct molecular simulation starting from a stable basin is very computationally inefficient in collecting reactive trajectories. Another issue that makes the study of rare events more complicated is the roughness of the potential energy landscape. For systems in the gas phase, the potential energy surface has only a few saddle points, which usually can be used to sufficiently characterize the transition process. In contrast, for rare events occurring in solution, the system of interest has a rugged potential energy surface on which myriads of small energy barriers with heights of the order of k_BT, must be distinguished, with the true potential energy barrier often larger than k_BT. One way to circumvent the challenges posed by the timescale separation and the roughness of the potential energy surface is to focus, as the TPS technique does, on the dynamic bottleneck for the rare event which is defined as the transition state surface. TPS starts with a pre-determined transition path connecting two stable states, and then by making shooting and shifting moves in the trajectory space using a Monte Carlo procedure, TPS can eventually lead to the true dynamics at transition states and an

ensemble of physically meaningful pathways. Eventually, complete reaction profiles, transition states, free energies of reaction barrier, and reaction rates can be obtained.

The prerequisites to perform a transition path sampling are two: 1) definitions of two stable states and 2) an initial trajectory, or a series of points in phase space, which connect the two stable states. This initial trajectory may not be physical or at the same condition as the one of interest. The most appealing feature of TPS is that no prior knowledge about the reaction mechanism, the reaction coordinate and the transition state is needed to begin with.

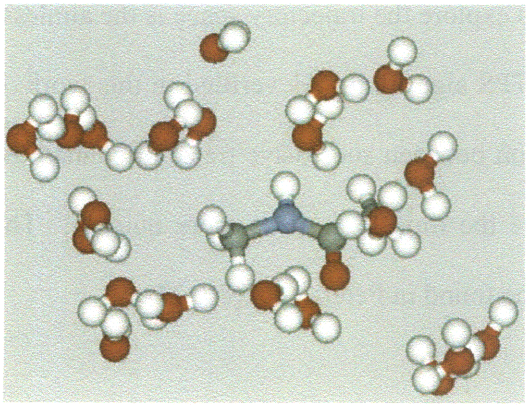
A more efficient sampling technique to explore the trajectory space is the aimless-shooting algorithm (*30, 31*), a variant of the TPS algorithm. As verified in this work, it has the advantage of having more decorrelation between successive trajectories than the original version of TPS and thus can explore the trajectory space more quickly. The detailed implementation of the algorithm can be found in Peters et al.(*30, 31*).

4.3 Methodology

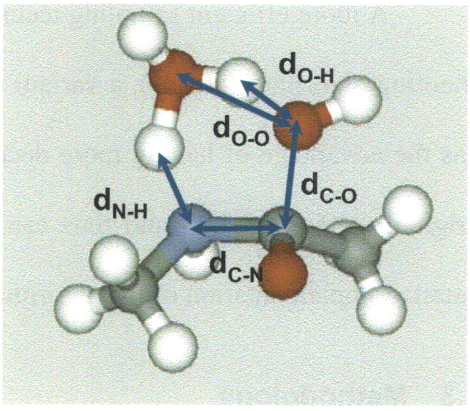
4.3.1 System description.

A simulation setup very similar to Zahn's potential-of-mean-force calculations[22] was chosen for our path sampling and analysis, but with a few important differences. First we performed an equilibration MD simulation by running long trajectories using the classical CHARMM force field[37] under the ambient temperature and pressure. This simulation was done using the CHARMM package[38] under the NPT ensemble. The force field parameters for N-MAA were taken from similar structures and the geometry of N-MAA was fixed. The TIP3P force field parameters were used for the water molecules in

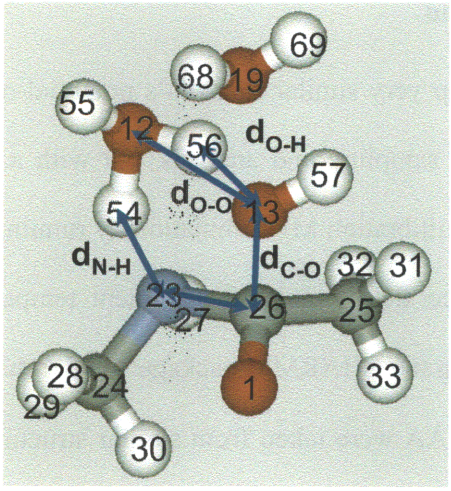
the system. The final equilibrated system had one N-MAA and twenty-one water molecules in a 12.27Å×8Å×8Å simulation cell, as Figure 1 shows. We found that the value of the water density in Zahn's studies was too high, while the water density values used in these studies had correct value at the ambient temperature and pressure. Next, long (~50 ps) constant-temperature MD trajectories were run with the C-O, the O-H, and the N-H distances as shown in Figure 4.1 constrained in the Car-Parrinello molecular dynamics (CPMD)[39] simulation using the CPMD package[40] in order to equilibrate other degrees of freedom.



(a)



(b)



(c)

Figure 4.1: Simulation box (a) together with bond distances used to define basins of stable states (b). In (c), atom labels used in the system are shown, to be used to refer to the order parameters defined in this study.

4.3.2 Stable basin definitions.

Here three distances, the C-O distance, the O-H distance, and the N-H distance as shown in Figure 4.1 are used to tell which stable state a particular configuration corresponds to. Long molecular dynamics trajectories without any constraints/restraints were run using CPMD to obtain the fluctuations in these bond distances, and then an appropriate range was taken by computing the variances of bond distances in the stable bond regions with adjustment such that in aimless shooting procedure, no returning back to indeterminate region once the system reaches one stable basin (shown in Figure 4.2). The mean values of these bond distances and their fluctuations are listed in Table 4.1, and the choice of basin definitions is also given. The values of these bond distances are listed as in Table 4.1.

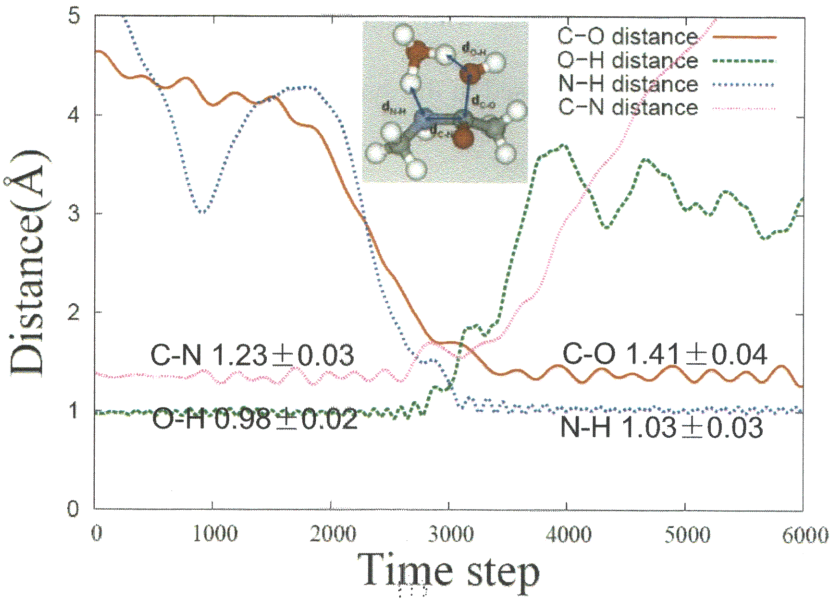


Figure 4.2: Transition trajectory with the associated changes in the OP's of bond distances.

Table 4.1: Ranges of bond distances in Figure 4.2 used for definitions of basins of stable states. A configuration corresponds to a particular stable state (either reactant or product) only when three bond distances are simultaneously within the specified ranges.

	fluctuation				definition	
	reactant		product		reactant	product
	mean	std	mean	std		
C-O dist(Å)			1.40	0.04	(2.07, +∞)	[1.27, 1.58]
O-H dist(Å)	0.98	0.05			[0.82, 1.14]	(1.68,+∞)
H-N dist(Å)			1.03	0.08	(1.79, +∞)	[0.79, 1.26]

During our path sampling procedure, we did not observe any basin overlapping situation, in which a particular configuration satisfied both the reactant and product definitions.

4.3.3 Order parameters.

Order parameters were selected based on quite extensive screening combined with physical intuition into the reaction mechanism. More specifically, a systematic procedure for including candidate OP's was used as follows. The set of candidate OP's includes distances of all possible pairs within the system, cosine values of angles and dihedral angles for all possible triplets and quadruplets, respectively, selected from all atoms around the midpoint of C-N bond within 6.5Å cutoff. The reason for using cosine values of these angles and dihedral angles instead of their absolute values is to avoid possible discontinuity when they take on values at the boundaries of their range. This procedure generated up to a total of 3,241,875 candidate order parameters. Other types of OP's used in previous studies, such as bonding number (*35*), density fluctuation (*36*), were not considered here. Our consideration is that the exact reaction coordinate $p_B(r)$ is a function of configuration vector only, and therefore, geometric quantities are enough to describe the reaction dynamics if an suitable ensemble is chosen at first. It should also be noted that while coordination numbers may be better collective variables than specific inter-atomic distances, the committment time for this reaction is so short that there is no permutation of the active water molecules during the reaction events. Exhaustive one OP variable screening in likelihood maximization was done for this vast set of candidate order parameters. Sooner a combinatorial problem arises even when going to two OP variables situation. The workaround used in this studied will be discussed shortly.

4.3.4 Aimless shooting.

The aimless shooting algorithm, a modified version of transition path sampling, as described by Peters and Trout (*31, 39*), was applied to harvest an ensemble of

independent trajectories according to their statistical weight. As with the transition path sampling method (*27, 28, 40*), aimless shooting requires 1) accurate definitions of the basins of stable states and 2) an initial trajectory that connects the stable basins.

The first reactive trajectory was obtained by guessing a high potential energy configuration with particular values of C-O, O-H and N-H bond distances. Then constrained MD with ~ 2ps was carried out to relax all the other degrees of freedom in the system to remove potential artifacts introduced when fixing these three distances. Both forward and backward shooting trajectories from the equilibrated configuration were then obtained with assigned velocities drawn from Maxwell-Boltzmann distribution. Repeated shootings were done until a reactive initial trajectory was obtained since basins of stable states were already defined. This resulting initial trajectory also provides some information on how fast the transition dynamics takes place, based on which the appropriate overall length of MD steps (~600 with a time step of 4 a.u.) can be derived.

Two-point version of aimless shooting has the following procedure. Two configurations close to hypothesized transition state were selected from the initial reactive trajectory and one of the two is chosen randomly from which forward and backward half-trajectories were shot. Momenta for forward shooting are generated from a Maxwell-Boltzmann distribution with no net linear and angular momenta for the whole system. Momenta for backward shooting were the reverse of those for forward shooting. The two configurations have a time displacement Δt which is an adjustable parameter and needs to be carefully set. If the forward and backward half-trajectories combine to give a reactive trajectory, this new trajectory is accepted and the two configurations with a time

displacement Δt to the previous shooting point was recorded as a new two-point from which the shooting procedure is repeated.

As described by Beckham et al. (*39*), time displacement Δt has to be chosen appropriately to yield an acceptance ratio between 40%-60%. If it is too large, the algorithm tends to go too far away from the transition-state region leading to a low acceptance rate with many consecutive unaccepted trajectories; if it is too small, the aimless shooting algorithm will be very inefficient in exploring shooting point configuration space and therefore more trajectories and needed to obtain a good approximation to the reaction coordinate. For chemical reaction in which bond breaking and forming steps are involved, Δt is expected to be smaller than more diffusive systems, since transitions driven by strong interactions are short in terms of transition duration.

Dynamic trajectories were collected using the CPMD package (*38*) in the NVT ensemble. A time step of 4 a.u. (~ 0.1 fs) and an electron fictitious mass of 400 a.u. were used. A chain of four Nose-Hoover thermostats were used to control temperature at 300 K. The molecular orbitals were described by a plane wave basis with an energy cutoff of 70 Ry. Vanderbuilt pseudopotentials and BLYP density functionals were used. We found selecting from two points, $\mathbf{x}_{-\Delta t}$, or $\mathbf{x}_{+\Delta t}$ is sufficient to sample the transition state ensemble and that is what we did in this study.

In the aimless shooting procedure, the trajectory length is set to be as short as possible in order to save computational time, resulting in the possibility of generating inconclusive trajectories, which have at least one end point in forward and backward half-trajectories does not lie in any basin of stable state. A half-trajectory step of 600 was found to maintain the level of inconclusive trajectories at or below 10%. A time

displacement, Δt, of 15 a.u. is chosen to yield an acceptance rate of 44.7%. 1660 trajectories were collected for later analysis.

In order to know how efficiently the trajectory space was sampled, autocorrelation function was computed to describe how dependent successive trajectories are. The configurations of shooting points in the aimless shooting procedure form an ordered series, which can be treated as a time series. Each configuration which leads to a reactive trajectory (meaning the next shooting step was a shifting) was aligned with a reference configuration by a best-fit procedure on the reaction core part (N-MAA and the two attaching water molecules). Then C-O distance (no need to align) and the root-mean-square-deviation of the reaction core part were calculated, followed with the calculation of normalized autocorrelation function calculation. As shown in Figure 4.3, both descriptors tell that essentially no memory exists in the following shooting trajectory. The fact that aimless shooting has much more decorrelation was presumably because of the complete renewal of momenta in each MC move in exploring the trajectory space.

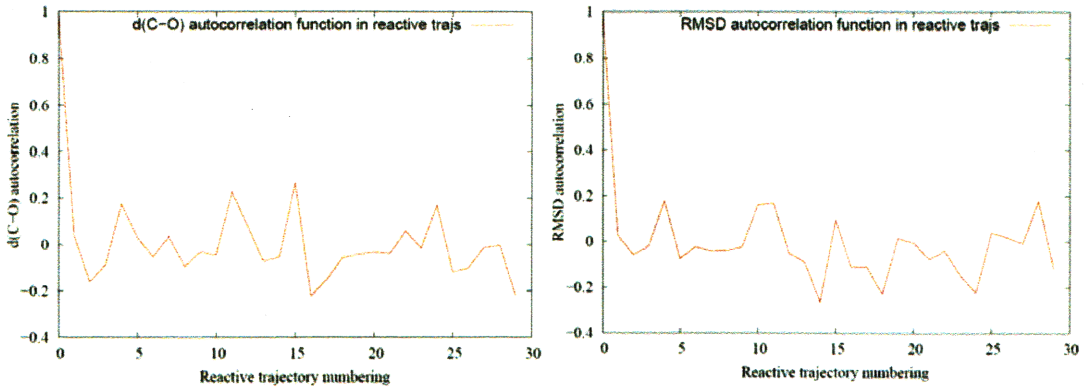


Figure 4.3: Autocorrelation function to describe the dependence of successive shooting moves in aimless shooting.

4.3.5 Likelihood maximization.

As described in Peters and Trout (*30*), the reaction coordinate, r, is modeled as a linear combination of candidate OP's, denoted as **q**, with α_0 through α_m as adjustable coefficients:

$$r(\mathbf{q}) = \alpha_0 + \sum_{k=1}^{m} \alpha_k q_k \tag{4.1}$$

It is noted that the choice of linear combination is for convenience purpose only, and a non-linear reaction coordinate expression could be chosen.

The model for the committor probability $p_B(r)$ was chosen to be

$$p_B(r) = [1 + \tanh(r)]/2 \tag{4.2}$$

This committor probability model was used to maximize the likelihood function with respect to the set of coefficients α_i's (i=0, …, m)

$$L(\vec{\alpha}) = \prod_{\vec{x}_k \to B} p_B(\vec{x}_k) \cdot \prod_{\vec{x}_k \to A} [1 - p_B(\vec{x}_k)] \tag{4.3}$$

only using outcomes of forward half-trajectories. In principle, if an ensemble of candidate OP's are proposed, the maximization of likelihood (3) should be performed over all combinations of OP's to determine the best reaction coordinate according to the models of Equations 1 and 2. However, when the set of candidate OP's is large, combinatorial problem arises for exhaustive screening of the best reaction coordinate. One is forced to reduce the size of the set of candidate OP's when the number of OP variables m increases. One choice for doing this is to do one OP variable exhaustive screening and use the best OP's for higher m searching, as was used in this study. For the best approximate reaction coordinate, the approximate transition state iso-surface can be

obtained by setting $p_B(r) = 1/2$. This occurs at $r = 0$, so setting $r(\mathbf{q}) = 0$ defines the approximate transition state iso-surface.

As informed by other examples such as alanine-dipeptide (*36*), very complex reaction coordinate might be involved in our system. In order not be biased by any assumption about which OP's are important, an exhaustive but systematic approach was taken to find the best reaction coordinate model in likelihood maximization. Over 3 million candidate OP's were screened individually first. In order to tackle the algorithmic complexity problem in higher dimensional (d>1) likelihood maximization, the following systematic approach was adopted. Basically, it assumes that important OP's previously screened based on likelihood scores will also be important in comprising reaction coordinate models in high dimensions. It starts with best m one-dimensional OP's. Then in each round for d dimensional optimization, every best ranked d-1 dimensional result is supplemented with every best m one-dimensional OP to give a d dimensional model. Then only n best d dimensional results are retained for d+1 dimensional screening. This way, each round roughly has $m*n$ optimization problems to solve.

4.3.6 Uncertainty analysis in likelihood maximization.

The criterion to discriminate different reaction coordinate models $r(\mathbf{q})$ is the Bayesian information criterion (BIC) (*31*), which equals log(N)/2 where N is the number of accepted trajectories in aimless shooting. If the difference in two log-likelihood scores is greater than BIC, the reaction coordinate model having a higher log-likelihood score is superior in describing the transition process; while on the other hand, if the difference in two log-likelihood scores is within BIC, the two reaction coordinate models are

indistinguishable, at least from the data collected. However, due to finite sampling, there is statistical uncertainty present in the estimate of likelihood score in (3) (*30*).

$$\sigma^2(\ln L) = \sum_{x_k} p_B(\vec{x}_k)[1 - p_B(\vec{x}_k)]\{\ln p_B(\vec{x}_k) - \ln[1 - p_B(\vec{x}_k)]\}^2 \qquad (4.4)$$

where the sum is over all shooting points each of which is a p_B-realization.

4.3.7 Reaction coordinate validation.

After the likelihood maximization to generate an approximation to reaction coordinate, its correctness must be checked. This can be done by computing the estimate of the probability of reaching product basin (p_B) from the predicted transition state region obtained in likelihood maximization as commonly referred as a committor distribution analysis, or p_B histogram computation[27]. In this procedure, independent configurations are generated which all satisfy the predicted transition state. Then a number of trajectories are initiated with the momenta drawn from a Maxwell-Boltzmann distribution from these configurations and the estimate of p_B values for these configuration can thus be obtained. Then a histogram of the number of configurations versus p_B values can be constructed.

For complex reaction coordinate like the ones used in this study, generating independent configurations for p_B histogram computation can be done efficiently by the BOLAS algorithm (*41*). First, shooting points were examined and several of them were selected close to the predicted transition state region, as defined by $r(\mathbf{q}) = 0$ in Equation 1. Very short trajectories are fired randomly from each initial configuration and the endpoints are evaluated to determine if they are within a narrow window on the transition state iso-surface. If so, this configuration is accepted and becomes the next shooting

point. This process is repeated until an adequate number of configurations is generated from which to shoot reactive trajectories to build a p_B histogram.

To construct the histogram, trajectories are shot from each configuration with a length corresponding to half the length of a reactive trajectory. The endpoints of the trajectories are evaluated and a histogram is constructed of the probability of reaching basin B from the predicted transition state isosurface. The basin definitions for constructing the p_B histograms correspond to the same basin definitions used for the reactant and product basins in the aimless shooting simulations. An adequate approximation to the true reaction coordinate will yield a histogram that is sharply peaked at $p_B = 0.5$ (*28*). Additionally, one can make a quantitative comparison of the histogram to the binomial distribution, which will have a mean value, $\mu = 0.50$ with a standard deviation, $\sigma = 0.050$.

The trajectories for the generation of new configurations are ~10 fs or 100 MD steps long, and the endpoint window width at $r = 0$ is constrained within a range of ±1% of the total configuration space sampled, as measured by Δr. For each histogram assembled in this study, 20 shooting points are collected. From each configuration collected, 20 trajectories are shot, corresponding to approximately 400 trajectories for each histogram. The trajectory length for calculating p_B values is ~60 fs or 600 MD steps, which is half the length of the reactive trajectories in the aimless shooting simulations, again resulting in a low rate of inconclusive paths.

4.4 Results and Discussions

4.4.1 Initial trajectory.

As mentioned previously, an initial reactive trajectory was obtained by first running long equilibration MD when fixing postulated bond distances of C-O, O-H and N-H labeled in Figure 4.1 in order to remove potential artifacts when using constraints. Then repeated forward and backward shootings were tried until the two half-trajectories combined to yield a reactive trajectory. Figure 4.2 shows the how bond distances change in this particular trajectory, together with extended sampling in the stable states in order to see how these bond distances fluctuate. It is clear that the transition dynamics has a time scale ~ 1000 MD steps. This provides a basis for choosing the MD steps in aimless shooting procedure to reach the compromise of efficiency and minimal number of inconclusive trajectories.

4.4.2 Trajectory characteristics.

Snapshots from a typical reactive trajectory shows what happens during the transition as shown for the reaction core part in Figure 4.4, and for the whole system together with the changes in the fleeting hydrogen bond network in Figure 4.5. As shown in Figure 4.4, most of the trajectories collected showed the hydrolysis proceed in a concerted fashion, i.e., two proton transfer process and C-O bond formation occur simultaneously instead of forming any long-lived intermediate species (compared with Zahn's (*21*)). In all of the reactive trajectories collected, C-N bond breaking was observed to produce the final hydrolyzed product, a methylamine molecule and an acetic acid molecule, both in their non-ionized forms due to neutral pH condition. In Figure 4.5,

significant change in the hydrogen-bond pattern close to the reaction core can be seen, indicating the effects of solvent degrees of freedom.

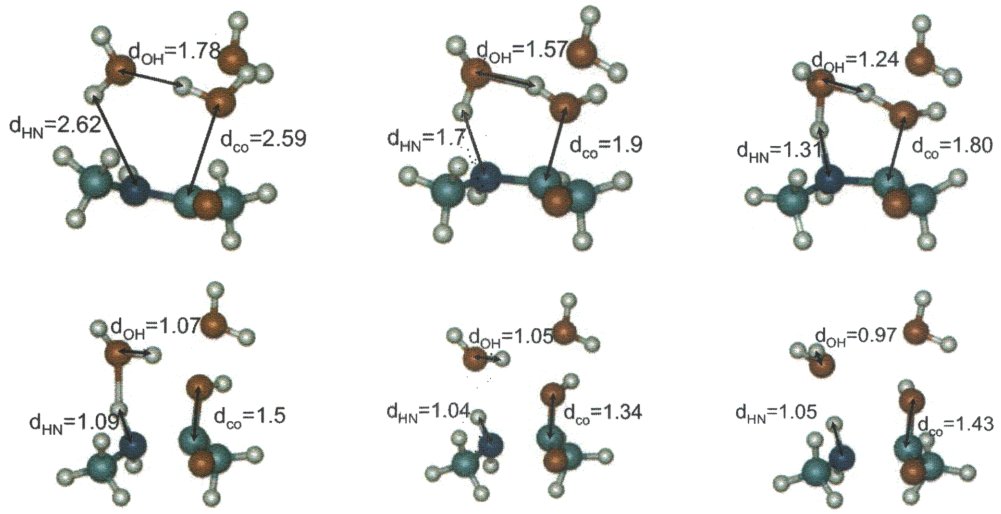


Figure 4.4: Key snapshots describing the reaction process. Only three water were shown for clarity. The overall trajectory is 1200 MD steps, during which complete hydrolysis reaction occurs. C-O bond formation and two proton transfer steps are concerted.

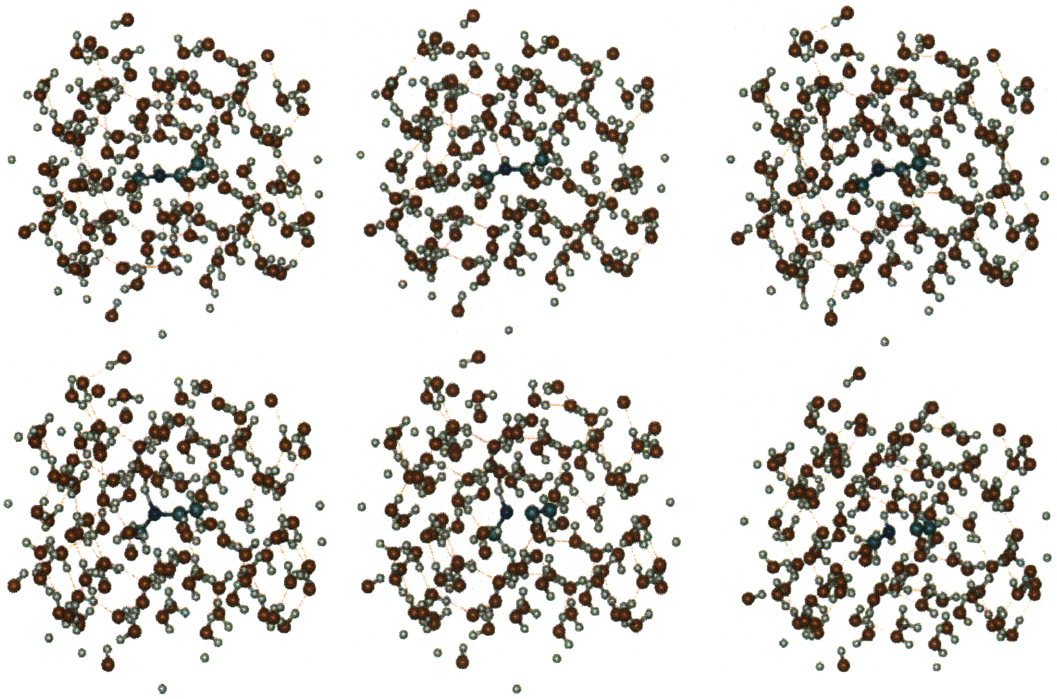


Figure 4.5: Snapshots of the fleeting hydrogen bonding network of solvent water.

4.4.3 Likelihood maximization.

The results of likelihood maximization are shown in Table 4.2. Few best reaction coordinate models in each dimensions up to six-dimension are listed. In addition, one OP variable best reaction coordinate models are also pictorially shown in Figure 4.7. Based on likelihood scores, it is much better in describing the hydrolysis reaction in the statistical sense of likelihood maximization than the order parameter of the distance between O(13)-C(26), which was used in the calculation of potential-of-mean-force for the rate-limiting step of the hydrolysis of N-MAA at neutral pH. As more OP's used in the linear combination expression of reaction coordinate model in (1), higher and higher log-likelihood scores were obtained, suggesting a complex reaction mechanism involving many physical degrees of freedom. In our approximation scheme for likelihood

maximization procedure when including more OP variables in the reaction coordinate models, log-likelihood score achieved to be within BIC criterion up to five-dimension, implying a convergent result.

Table 4.2: Likelihood maximization results for N=1650 aimless shooting paths, with a BIC=log(N/2)=3.704. The order parameters (OP's) have the following meaning: d(n1,n2) is the distance between atom number n1 and n2, a(n1,n2, n3) is the angle comprised of atom number n1, n2 and n3, phi(n1,n2, n3,n4) is the dihedral angle comprised of atom number n1, n2, n3 and n4. The column α's gives the vector $\alpha=(\alpha_0, \alpha_1, \ldots, \alpha_n)$ corresponding to reduced and normalized OP $q_i \in [0, 1]$.

Number of OP variables in RC model	OP's in best ranked RC models	Likelihood Score(ln(L))	α's
1	phi(O13-C26-H35-H56)	-913.613	1.470,-2.588
	phi(O16-O4-O6-H56)	-917.579	-0.973, 3.033
	phi(O16-O6-C23-H56)	-919.482	1.744,-2.759
	phi(O13-O6-H27-H56)	-923.036	2.156,-3.236
	phi(C26-H37-H42-H56)	-925.344	-1.218, 3.361
	d(O13-C26)[a]	-1111.496	0.264,-0.146
2[b]	phi(C25-O1-C26-H56), phi(H31-H56-H75-H43)	-846.562	0.880,-2.711, 1.976
	a(O13-H29-H56), d(C26-O15-H75-H54)	-849.811	0.495,-2.509, 1.856
	phi(C25-O1-C26-H56), d(O13-H56)	-850.537	0.622,-2.506, 2.051
	d(H54-H55), d(O12-H56)	-852.926	0.182, 2.405,-2.613
	a(O13-H29-H56), d(H54-H55)	-853.336	0.179,-2.344, 2.178
3[b]	phi(C25-O1-C26-H56), phi(H31-H56-H75-H43), phi(H54-O19-H56-H55)	-819.826	1.423,-2.201, 1.708,-1.223
	d(O13-H56), d(H54-H55), phi(C25-O1-C26-H56)	-823.649	-0.584, 1.960, 1.417,-1.785
	a(O13-H29-H56), d(H54-H55), phi(H45-O19-H39-H56)	-824.471	-0.244,-2.137, 1.758, 1.367
	a(O13-H29-H56), d(H54-H55), phi(H45-O6-H39-H56)	-824.819	0.996,-1.949, 1.780,-1.497
	phi(C25-O1-C26-H56), d(H31-H56-H75-H43), phi(O13-H28-H42-H56)	-825.253	0.426,-2.391, 1.503, 1.312
	phi(C25-O1-C26-H56), d(O12-H56), d(H54-H55)	-826.484	1.337,-1.657,-2.101, 1.654
4[b]	phi(C25-O1-C26-H56), phi(H52-O6-H74-H56), d(H54-H55), d(O13-H56)	-810.862	-0.940,-1.663, 1.445, 1.230, 1.437
	phi(C25-O1-C26-H56), phi(O13-O6-N23-H56), phi(H31-H56-H75-H43), d(H54-H55)	-812.480	0.836,-1.862,-1.259, 1.362, 1.154
	phi(C25-O1-C26-H56), phi(H31-H56-H75-H43), phi(O13-H28-H42-H56), d(H54-H55)	-813.420	-0.283,-1.983, 1.383, 1.197, 0.980
	phi(C25-O1-C26-H56), phi(H52-O6-H74-H56), phi(H54-H55), phi(O12-H56)	-813.633	0.443,-1.540, 1.460, 1.381,-1.515
	a(O13-H29-H56), d(H54-H55), phi(O5-O22-	-813.873	0.313,-1.578,

	H54-H56), phi(C25-O1-C26-H56)		1.313, 1.345,-1.421
	a(O13-H29-H56), d(H54-H55), phi(C25-O1-C26-H56), phi(H45-O6-H39-H56)	-814.044	1.586,-1.667, 1.416,-1.088,-1.256
5[b]	d(O13-H56), d(H54-H55), phi(C25-O1-C26-H56), a(O13-O3-H56), phi(H52-O6-H74-H56)	-803.535	-2.891, 2.991, 1.547,-1.695, 1.693, 1.593
	d(O13-H56), d(H54-H55), phi(C25-O1-C26-H56), a(O13-O3-H56), phi(O13-H42-H56-C26)	-803.849	-2.547, 3.704, 1.665,-1.848, 2.804,-1.397
	d(O13-H56), d(H54-H55), phi(C25-O1-C26-H56), a(O13-O3-H56), phi(O11-O6-H74-H56)	-804.279	-4.160, 3.982, 1.672,-1.766, 2.896, 1.48
	phi(C25-O1-C26-H56), phi(H31-H56-H75-H43), phi(H54-O19-H56-H55), phi(N23-O5-H52-H54), phi(O11-H56-H71-H27)	-804.847	1.347,-2.342, 1.966,-1.327, 1.277,-1.201
	d(O13-H56), d(H54-H55), phi(C25-O1-C26-H56), a(O13-O3-H56), phi(H27-H37-H56-H42)	-805.091	-2.901, 3.186, 1.575,-1.756, 1.937, 1.101
6[b]	phi(C25-O1-C26-H56), phi(H31-H56-H75-H43), phi(H54-O19-H56-H55), phi(C23-O5-H52-H54), phi(O11-H56-H71-H27), phi(H47-O6-H62-H56)	-800.130	2.125,-2.284, 2.024,-1.308, 1.280,-1.754,-0.922
	d(O13-H56), d(H54-H55), phi(C25-O1-C26-H56), a(O13-O3-H56), phi(O11-O6-H74-H56), phi(H41-O22-H54-H56)	-800.176	-4.310, 3.512, 1.541,-1.784, 2.830, 1.498, 0.942
	phi(C25-O1-C26-H56), phi(H31-H56-H75-H43), phi(H54-O19-H56-H55), phi(C23-O5-H52-H54), phi(O11-H56-H71-H27), phi(O13-H28-H42-H56)	-800.222	1.013,-2.307, 1.787,-1.038, 1.210,-1.240, 0.750
	d(O13-H56), d(H54-H55), phi(C25-O1-C26-H56), a(O13-O3-H56), phi(O13-H42-H56-C26), phi(C26-O20-H33-H56)	-800.317	-2.814, 3.930, 1.761,-1.489, 3.042,-1.160,-0.695
	a(O13-H29-H56), d(H54-H55), phi(C25-O1-C26-H56), phi(O13-H42-H56-C26), phi(H45-O19-H39-H56), phi(O7-H39-H56-O19)	-800.362	0.292,-1.320, 1.738,-1.528,-1.391, 1.262, 0.956
	d(O13-H56), d(H54-H55), phi(C25-O1-C26-H56), a(O13-O3-H56), phi(O11-O6-H74-H56), phi(O5-H22-H54-H56)	-800.397	-4.511, 3.786, 1.592,-1.784, 3.067, 1.403, 0.768

[a]This order parameter was used in previous potential-of-mean-force calculation (*21*)
[b]Results in higher-dimensional likelihood maximization, with m=100, n=100. Convergence was achieved at dimension d=5.

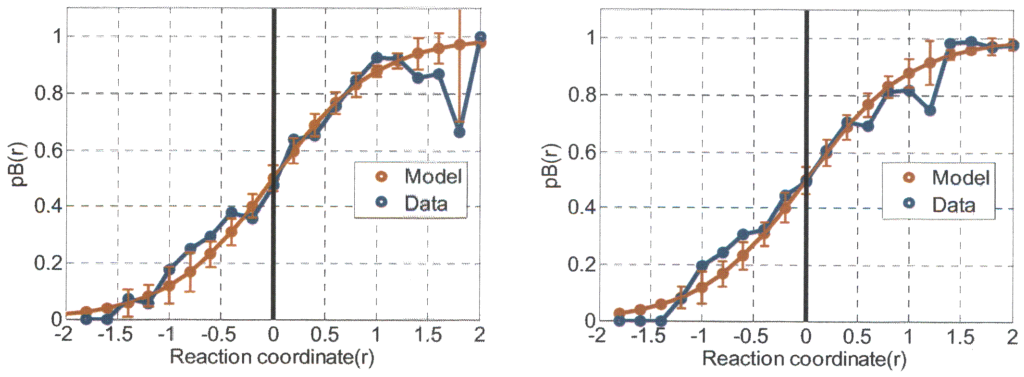
The $p_B(r)$ model given in Equation (4.2) is used to calculate the likelihood function as shown in Equation (4.3). The corresponding models for the 1-dimensional reaction coordinate approximations are shown for both systems. The OPs in the expression for r are provided on a normalized basis such that $q_i \in [0, 1]$.

The reaction coordinate models for both the best three-dimension and the five-dimension were checked against the aimless shooting data, as shown in Figure 4.6. All

accepted shooting points for which the forward and backward shootings led to a conclusive trajectory were used to construct two histograms of the reaction coordinate determined from likelihood maximization. The two histograms were based on only on the half-trajectories of the forward shooting from each accepted shooting point whether it ends in reactant basin or in product basin. Then $p_B(r)$ data values in Figure 4.6 were computed as the ratio

$$p_B(r)\,|_{data,i\text{ th bin}} = \frac{N_{fB,i}}{N_{fA,i} + N_{fB,i}} \tag{4.5}$$

where $N_{fA,i}$ and $N_{fB,i}$ stands for the number of shooting points giving the reaction coordinate model value in i th bin that led to forward shooting half-trajectory ends in A, or B respectively. Thus, the comparison of model vs. data provides a measure of how well aimless shooting collects information about transition paths and shooting points, as well as how good likelihood maximization is calculated. In Figure 4.6, one can see that both reaction coordinate models give satisfactory fit to the aimless shooting data.



(a) (b)

Figure 4.6: Comparison of $p_B(r)$ model vs. aimless shooting data. Here half trajectory pB(r) model was used, i.e. $p_B(r)$=[1+tanh(r)]/2. Note that the error bars appear on the model, not the data. The error bars show how far shooting point data should deviate from

the probabilities $p_B(r)$ for a perfect reaction coordinate model. (a) 3-OP variable reaction coordinate model (b) 5-OP variable reaction coordinate model.

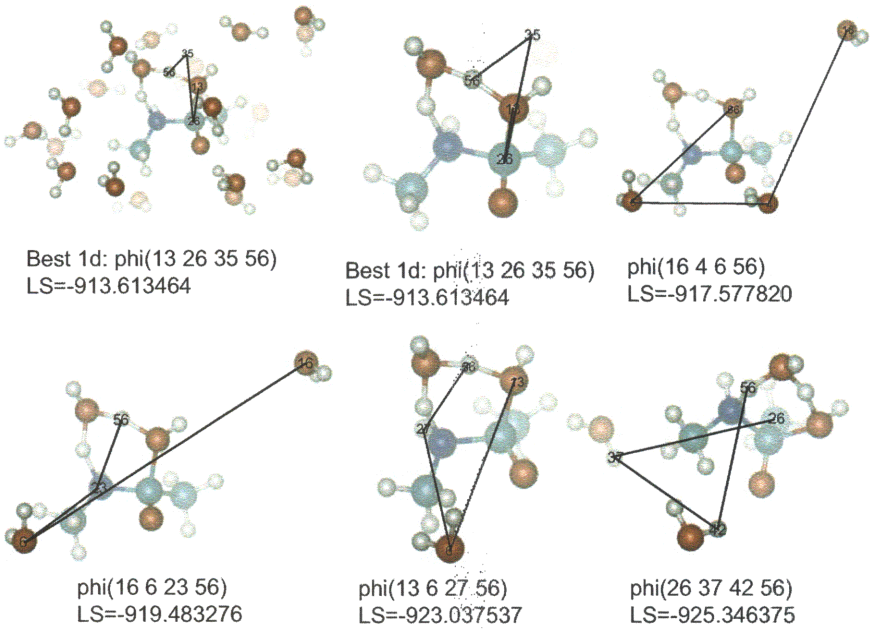


Figure 4.7: Illustration of best ranked one-OP variable OP's in the likelihood maximization procedure. The OP is given as a dihedral among quadruple atoms (denoted as phi(atom_index_1, atom_index_2, atom_index_3, atom_index_4), or as an angle among triple atoms (denoted as a(atom_index_1, atom_index_2, atom_index_3), or as a bond distance between pair atoms (denoted as d(atom_index_1, atom_index_2). The associated likelihood scores (LS) are also given. One observation is that almost all these best ranked involves the hydrogen atom indexed as 56:H.

Using equation shown in (4), the statistical uncertainty in the log-likelihood score was computed with the collection of accepted shooting points to be $\sigma^2(\ln L)=14.86$ and $\sigma^2(\ln L)= 12.60$ for best three-dimensional and five-dimensional reaction coordinate

model respectively. These values provide reasonable confidence in our log-likelihood score value (*42*).

Figure 4.8 showed the OP's which comprising the best reaction coordinate model with five OP variables. These include the local bonding pattern changes, such as proton H56 being transferred between the two attacking water molecules, and the newly formed N23-H54 bond. They also reflect some influence of the solute N-MAA itself on the reaction dynamics, such as the dihedral angle between C25-O1-C26-H56. Solvent degrees of freedom in affecting reaction dynamics are also needed, as seen by the presence of an angle O13-O3-H56, and a dihedral angle H52-O6-H75-H56. The inclusion of both local OP's close to the reaction center and the OP's describing the solvent networks suggests the importance of solvent in determining reaction dynamics.

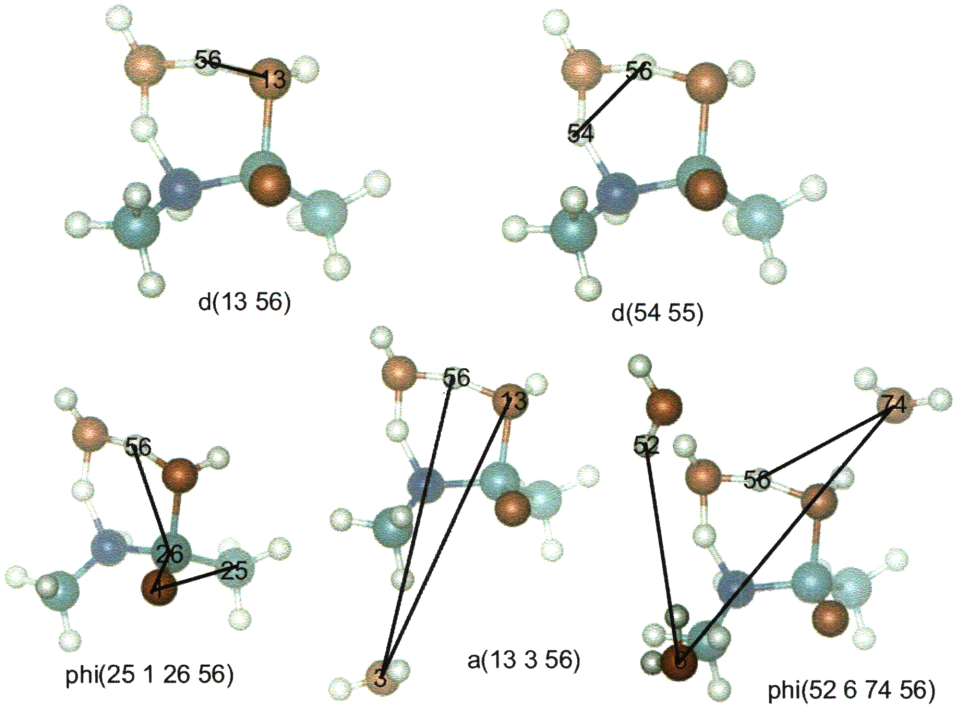


Figure 4.8: Illustration of constituent OP's in the best 5-OP variable reaction coordinate model. Naming of these OP's is the same as in Figure 4.7.

4.4.4 Reaction coordinate validation.

Four p_B histograms were computed using the method described above. These include using the C-O distance reaction coordinate model, the best one-dimensional reaction coordinate model, the best two-dimensional reaction coordinate model, and the best five-dimensional reaction coordinate model. The results are shown in Figure 4.9. The poor description of the reaction process by the C-O distance reaction coordinate model can be seen in this p_B histogram, since its distribution is very skewed. Both the best one-dimensional reaction coordinate model and the best two-dimensional model show inferior histogram than the best five-dimensional reaction coordinate model.

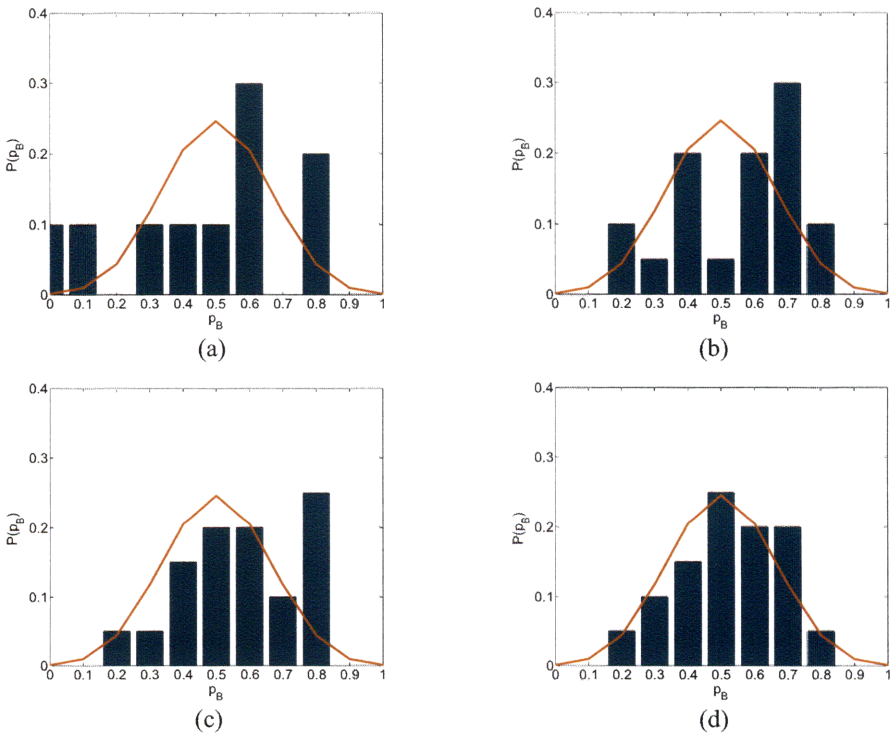


Figure 4.9: Committor probability histogram using C(26)-O(13) as reaction coordinate model (a), best one-OP variable reaction coordinate model (b), best two-OP variable reaction coordinate model (c) and best five-OP variable reaction coordinate model (d),

compared with binomial distribution (red line). Quantification of means and standard deviations for these histograms following the procedure in Peters (*42*) is shown in Table 4.3.

Table 4.3: Maximal likelihood estimates for means and standard deviations in the p_B histograms shown in Figure 4.8. The procedure was used as in Peters (*42*).

reaction coordinate model/distribution	μ_h	σ_h
C(26)-O(13)	0.470	0.066
best one-OP variable	0.550	0.034
best two-OP variable	0.575	0.031
best five-OP variable	0.525	0.024
binomial distribution	0.500	0.025
ideal $P(P_B)$	0.500	0.000

4.5 Summary and conclusions

In this study, the mechanism of hydrolysis reaction of peptide bond at neutral pH was studied using a model compound N-MAA. Due to fluctuations at finite temperature, path sampling method was used to generate an ensemble of trajectories according to their statistical weight in trajectory space. *Ab initio* molecular dynamics technique was applied to advance the time evolution of the reaction and collect trajectories. Likelihood maximization and its modification were used in extracting physically important degrees of freedom in the system and approximations of the reaction coordinate were compared. It was found that this hydrolysis reaction is very complex in nature, and involves many degrees of freedom. The specific conclusions obtained in our study are:

- Hydrolysis of N-MAA at neutral pH occurs in a concerted fashion; no stable or long-lived intermediate was found in our path sampling simulations.
- Likelihood maximization procedure was extended to screen higher dimensional reaction coordinate models, and within BIC, a reaction

coordinate with five constituent geometric variables was found to be the best in describing the path ensemble we generated.

- In the best reaction coordinate model, both geometric quantities which reflect bond making and breaking dynamics, and those which reflect the solvent network changes, are included, suggesting a complicated reaction involving many degrees of freedom.
- Several p_B histograms were computed to verify the results of likelihood maximization, and the quantified goodness of these best-ranked reaction coordinate models is in accord with the their respective likelihood score.

The technique of likelihood maximization is a very powerful statistical tool in determining a reaction mechanism, especially when it is complex. However, new problems arise when the set of candidate order parameters is large, such as the combinatorial problem in exhaustive screening for the best reaction coordinate model. This study provides an approach to carrying out a curtailed optimization procedure to determine the reaction coordinate. Even though it may be difficult to relate a concise physical picture with the likelihood maximization using many OP's, as far as p_B histogram and the calculation of quantities of interest such as free energy profile and reaction rates concern, this improvement has its advantage.

4.6 References

(1) Chu, J. W., Yin, J., Brooks, B. R., Wang, D. I. C., Ricci, M. S., Brems, D. N., and Trout, B. L. (2004) A comprehensive picture of non-site specific oxidation of methionine residues by peroxides in protein pharmaceuticals. *Journal of Pharmaceutical Sciences 93*, 3096-3102.

(2) Pan, B., Abel, J., Ricci, M. S., Brems, D. N., Wang, D. I. C., and Trout, B. L. (2006) Comparative oxidation studies of methionine residues reflect a structural effect on chemical kinetics in rhG-CSF. *Biochemistry 45*, 15430-15443.

(3) Liu, D. T. Y. (1992) Deamidation - a Source of Microheterogeneity in Pharmaceutical Proteins. *Trends in Biotechnology 10*, 364-369.

(4) Kosky, A. A., Razzaq, U. O., Treuheit, M. J., and Brems, D. N. (1999) The effects of alpha-helix on the stability of Asn residues: Deamidation rates in peptides of varying helicity. *Protein Science 8*, 2519-2523.

(5) Kahne, D., and Still, W. C. (1988) Hydrolysis of a Peptide-Bond in Neutral Water. *Journal of the American Chemical Society 110*, 7529-7534.

(6) Wei, W. (1999) Instability, stabilization, and formulation of liquid protein pharmaceuticals. *International Journal of Pharmaceutics 185*, 129-188.

(7) Daugherty, A. L., and Mrsny, R. J. (2006) Formulation and delivery issues for monoclonal antibody therapeutics. *Advanced Drug Delivery Reviews 58*, 686-706.

(8) Cohen, S. L., Price, C., and Vlasak, J. (2007) beta-elimination and peptide bond hydrolysis: Two distinct mechanisms of human IgG1 hinge fragmentation upon storage. *Journal of the American Chemical Society 129*, 6976-+.

(9) Cordoba, A. J., Shyong, B. J., Breen, D., and Harris, R. J. (2005) Non-enzymatic hinge region fragmentation of antibodies in solution. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences 818*, 115-121.

(10) Dillon, T. M., Bondarenko, P. V., Rehder, D. S., Pipes, G. D., Kleemann, G. R., and Ricci, M. S. (2006) Optimization of a reversed-phase high-performance liquid chromatography/mass spectrometry method for characterizing recombinant antibody heterogeneity and stability. *Journal of Chromatography A 1120*, 112-120.

(11) Brown, R. S., Bennet, A. J., and Slebockatilk, H. (1992) Recent Perspectives Concerning the Mechanism of H3O+-Promoted and Oh--Promoted Amide Hydrolysis. *Accounts of Chemical Research 25*, 481-488.

(12) Bryant, R. A. R., and Hansen, D. E. (1996) Direct measurement of the uncatalyzed rate of hydrolysis of a peptide bond. *Journal of the American Chemical Society 118*, 5498-5499.

(13) Radzicka, A., and Wolfenden, R. (1996) Rates of uncatalyzed peptide bond hydrolysis in neutral solution and the transition state affinities of proteases. *Journal of the American Chemical Society 118*, 6105-6109.

(14) Smith, R. M., and Hansen, D. E. (1998) The pH-rate profile for the hydrolysis of a peptide bond. *Journal of the American Chemical Society 120*, 8910-8913.

(15) Krug, J. P., Popelier, P. L. A., and Bader, R. F. W. (1992) Theoretical-Study of Neutral and of Acid and Base Promoted Hydrolysis of Formamide. *Journal of Physical Chemistry 96*, 7604-7616.

(16) Antonczak, S., Ruizlopez, M. F., and Rivail, J. L. (1994) Ab-Initio Analysis of Water-Assisted Reaction-Mechanisms in Amide Hydrolysis. *Journal of the American Chemical Society 116*, 3912-3921.

(17) Bakowies, D., and Kollman, P. A. (1999) Theoretical study of base-catalyzed amide hydrolysis: Gas- and aqueous-phase hydrolysis of formamide. *Journal of the American Chemical Society 121*, 5712-5726.

(18) Stanton, R. V., Perakyla, M., Bakowies, D., and Kollman, P. A. (1998) Combined ab initio and free energy calculations to study reactions in enzymes and solution: Amide hydrolysis in trypsin and aqueous solution. *Journal of the American Chemical Society 120*, 3448-3457.

(19) Zahn, D. (2004) Car-Parrinello molecular dynamics simulation of base-catalyzed amide hydrolysis in aqueous solution. *Chemical Physics Letters 383*, 134-137.

(20) Zahn, D. (2004) Investigation of the complex catalyzed amide hydrolysis from reaction coordinate of acid molecular dynamics simulations. *Chemical Physics 300*, 79-83.

(21) Zahn, D. (2004) On the role of water in amide hydrolysis. *European Journal of Organic Chemistry*, 4020-4023.

(22) Zahn, D. (2003) Theoretical study of the mechanisms of acid-catalyzed amide hydrolysis in aqueous solution. *Journal of Physical Chemistry B 107*, 12303-12306.

(23) Zahn, D., Schmidt, K. F., Kast, S. M., and Brickmann, J. (2002) Quantum/classical investigation of amide protonation in aqueous solution. *Journal of Physical Chemistry A 106*, 7807-7812.

(24) Buren, N. V., Rehder, D., Matsumura, H. G. M., and Jacob, J. (in press) Elucidation of two major aggregation pathways in an IgG2 antibody. *Journal of Pharmaceutical Sciences*.

(25) N. Perico, J. P., T.M. Dillon, M.S. Ricci. (in press) Conformational implications of an inversed pH-dependent antibody aggregation. *Journal of Pharmaceutical Sciences*.

(26) Bolhuis, P. G., Chandler, D., Dellago, C., and Geissler, P. L. (2002) Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual Review of Physical Chemistry 53*, 291-318.

(27) Dellago, C., Bolhuis, P. G., Csajka, F. S., and Chandler, D. (1998) Transition path sampling and the calculation of rate constants. *Journal of Chemical Physics 108*, 1964-1977.

(28) Dellago, C., Bolhuis, P. G., and Geissler, P. L. (2002) Transition path sampling. *Advances in Chemical Physics, Vol 123 123*, 1-78.

(29) Camargo, A. C. M., Gomes, M. D., Reichl, A. P., Ferro, E. S., Jacchieri, S., Hirata, I. Y., and Juliano, L. (1997) Structural features that make oligopeptides susceptible substrates for hydrolysis by recombinant thimet oligopeptidase. *Biochemical Journal 324*, 517-522.

(30) Peters, B., Beckham, G. T., and Trout, B. L. (2007) Extensions to the likelihood maximization approach for finding reaction coordinates. *Journal of Chemical Physics 127*, -.

(31) Peters, B., and Trout, B. L. (2006) Obtaining reaction coordinates by likelihood maximization. *Journal of Chemical Physics 125*, -.

(32) Weinan, E., Ren, W. Q., and Vanden-Eijnden, E. (2005) Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes. *Chemical Physics Letters 413*, 242-247.

(33) Metzner, P., Schutte, C., and Vanden-Eijnden, E. (2006) Illustration of transition path theory on a collection of simple examples. *Journal of Chemical Physics 125*, -.

(34) E, W., and Vanden-Eijnden, E. (2006) Towards a theory of transition paths. *Journal of Statistical Physics 123*, 503-523.

(35) Geissler, P. L., Dellago, C., Chandler, D., Hutter, J., and Parrinello, M. (2001) Autoionization in liquid water. *Science 291*, 2121-2124.

(36) Ma, A., and Dinner, A. R. (2005) Automatic method for identifying reaction coordinates in complex systems. *Journal of Physical Chemistry B 109*, 6769-6779.

(37) Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983) Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry 4*, 187-217.

(38) CPMD, C. I. C.-. Copyright MPI für Festkörperforschung Stuttgart 1997-2001.

(39) Beckham, G. T., Peters, B., Starbuck, C., Variankaval, N., and Trout, B. L. (2007) Surface-mediated nucleation in the solid-state polymorph transformation of terephthalic acid. *Journal of the American Chemical Society 129*, 4714-4723.

(40) Bolhuis, P. G., Dellago, C., and Chandler, D. (1998) Sampling ensembles of deterministic transition pathways. *Faraday Discussions*, 421-436.

(41) Radhakrishnan, R., and Schlick, T. (2004) Biomolecular free energy profiles by a shooting/umbrella sampling protocol, "BOLAS". *Journal of Chemical Physics 121*, 2436-2444.

(42) Peters, B. (2006) Using the histogram test to quantify reaction coordinate error. *Journal of Chemical Physics 125*, -.

Chapter 5. Molecular Mechanism of Acid-Catalyzed Hydrolysis of N-MAA

5.1 Introduction

Chemical stability of peptide bond in protein molecules is essential for life, as well as for the applications of protein pharmaceutics. Recombinant protein molecules can have chemical degradation pathways such as oxidation (*1, 2*), deamidation, hydrolysis of peptide bond on the protein backbone and etc (*3*). Monoclonal antibody molecules (*4*), as a very promising class of bio-pharmaceutics, have been reported by several different groups to undergo non-enzymatic fragmentation in the hinge region (*5-8*). This was identified mainly as the hydrolysis of several peptide bonds in the hinge. We believe that understanding of the underlying reaction mechanism particularly from a molecular perspective will help to minimize the extent of degradation in a rational and more efficient way. Examples of this approach include manipulating solvent accessibility to control methionine oxidation after identifying that water exposure is a key parameter in the oxidation process (*1*).

There has been a good deal of experimental and computational studies on the hydrolysis of peptide bond in aqueous solution due to its important biological relevance. Spurred by the interest in understanding the enzyme proficiencies in catalyzing peptide bond hydrolysis, a number of groups (*9-12*) extensively measured the reaction rates of hydrolysis of peptide bonds. It was found that in general, the reaction rate at neutral pH is extremely slow, with a half-time corresponding to hundreds of years. However, this reaction occurs much faster in both acidic and basic conditions, as carefully studied by

Smith et al. (*13*). They extensively studied the pH and concentration dependences of reaction rates for the hydrolysis reaction of a peptide bond in N-(phenylacetyl)glycyl-D-valine (PAGV), and identified three regimes of reaction mechanisms: acid catalyzed mechanism for pH<4 with a reaction rate first-order in both concentrations of PAGV and of H^+, base catalyzed mechanism for pH>10 with a reaction rate first-order both in concentrations of PAGV and of OH^-, and water mediated mechanism for pH values in between with a reaction rate first-order only in the concentration of PAGV. It was found that in general, the reaction rate under acidic pH condition is tens of thousands times faster when compared with the situation under neutral pH, reaction rate.

A number of computational efforts (*14-16*) were devoted to understanding the energetics of both stable and intermediate species involved in the hydrolysis reaction. However, these studies showed only a static picture of the reaction process. Stranton et al. (*17*) used combined quantum mechanical and classical mechanical calculations to study the hydrolysis reaction of peptide bonds catalyzed by trypsin in solution and free energies of reaction were reported. Zahn et al. and Zahn (*18-22*) studied the hydrolysis reaction of N-MAA in various solution conditions, namely acidic, basic and neutral pH. In these *ab initio* calculations of the potential of mean force, reaction coordinates were assumed and the projection of the free energy hyper-surface onto these coordinates was performed using constrained molecular dynamics.

Even though the hydrolysis reaction occurs slowly, the importance of its occurrence became evident when it was found in the formulation studies of antibody pharmaceutics. Cordoba et al. (*6*) reported several percent hydrolytic degradation of residues in the hinge region of IGG-1 antibody molecules over an incubation period of

three months. Their study showed the hydrolysis could occur at different peptide bonds in the hinge with variable extent of degradation. They also showed that the reaction was un-catalyzed around neutral pH or slightly acidic pH condition. Dillon et al. (*7, 8*) developed new analytical techniques, and reached similar conclusions about the location and extent of hydrolytic cleavage of peptide bonds in hinge region. Thus it is practically interesting to understand the underlying mechanism of hydrolysis reaction in both acidic and neutral pH conditions, since most antibody molecules are found to attain their most stability under these conditions (*23*).

Different mechanisms of the hydrolysis reaction have been proposed in the literature for different solution pH conditions. In acidic condition, as shown in Figure 5.1, the hydrolysis reaction is accelerated by the protonation of carbonyl oxygen (*10*), followed by a water-assisted H_2O attack on the resulting O-protonated amide to yield a tetrahedral intermediate I1. This is the rate-limiting step as confirmed by careful experiments of solvent kinetic isotope effects. (*10*) Following the formation of the tetrahedral intermediate, a proton is installed by another water molecule on the amide nitrogen to form another intermediate I2, and C-N cleavage ensues with formation of un-protonated carboxylic acid and an amine.

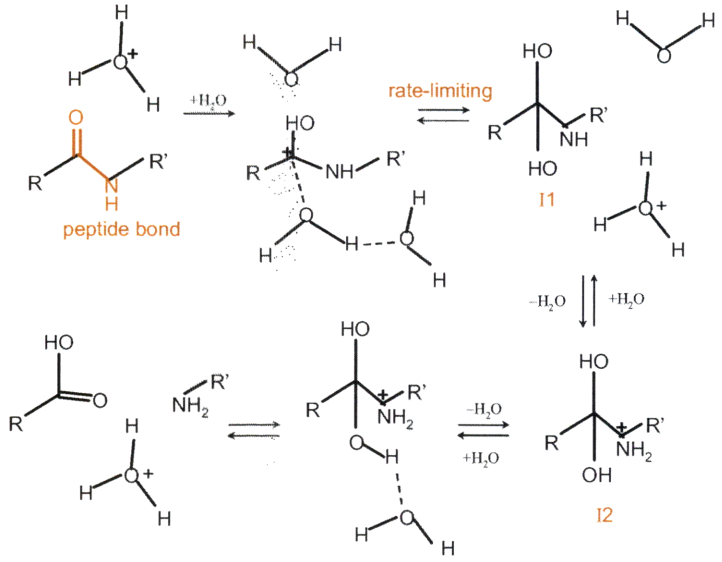


Figure 5.1: The acid-catalyzed pathway of hydrolysis reaction of peptide bond (*10*). Only the rate limiting step is studied in this paper. I1 and I2 are the two meta-stable intermediates.

In Chapter 4 and this chapter, we studied the hydrolysis reaction of peptide bond under both acidic and neutral pH conditions by using a model compound N-methyl acetyl acrylamide (N-MAA), in which two methyl groups are the minimal but computationally reasonable constituents on the peptide bond –NH-CO-. *Ab initio* molecular dynamics simulations at finite temperature equipped together with transition path sampling (TPS) (*24-26*), likelihood maximization (*27-29*) and p_B histogram analysis (*26*) techniques were carried out to gain understanding of the reaction mechanism, including identification and verification of reaction coordinate. This work aims at addressing the identification of reaction coordinate in the rate-determining step, i.e., the formation of intermediate I1 in the acid-catalyzed pathway of hydrolysis reaction.

What we mean by reaction coordinate here is a descriptor of the real physical progress of the transition from the reactant state to the product state. This has to be

contrasted with order parameters, which serve mainly to distinguish the two ending states. An order parameter is a quantity whose value can be used to distinguish different thermodynamic states, examples like reactant or product states, and crystalline, amorphous or liquid states. A reaction coordinate has to be an order parameter; while the contrary is not necessarily true. Identification of the "correct" reaction coordinate from a collection of potential order parameters is a very challenging problem, especially when the transition is complex. However, the information of "correct" reaction coordinate is necessary when computing quantities of most of our interest, such as free energy barriers and reaction rate constants from molecular simulations. Also we believe that the knowledge of the "correct" reaction coordinate, and hence the reaction mechanism, can provide essential molecular level insight for judicious engineering of complex systems.

5.2 Overview

At finite temperature, no single transition trajectory connecting the reactant and product states can be used to describe the whole reaction process due to thermal fluctuation. In trajectory space, each transition path is associated with a statistical weight (*26*), contributing to the experimentally observed transition process. It is therefore necessary to collect an ensemble of reactive trajectories and calculate the probability distribution according to their statistical weight in order to calculate quantities of interest that can be compared with experiments. The technique that implements this Monte Carlo procedure in trajectory space is transition path sampling (*24-26*).

Transition path sampling is well-suitable for studying transition process which has time-scale separation. In rare events, the time scale for the system to wander in the valley of stable states is much longer than the time in which the transition dynamics take place.

Therefore, direct molecular simulation starting from a stable basin is very inefficient in collecting reactive trajectories. TPS starts with a pre-determined transition path connecting two stable states, and then by doing shooting and shifting moves in trajectory space using a monte carlo procedure, can eventually lead to the true dynamics at transition states and an ensemble of physically meaningful pathways. Eventually, complete reaction profiles, transition states, barrier free energies, and reaction rates can be obtained.

The prerequisites to perform a transition path sampling are 1) an initial trajectory in phase space (no need to be physical) and 2) definitions of two stable states. The most appealing feature of TPS is that it does not need any prejudgment about the reaction mechanism.

5.3 Methodology

5.3.1 System description.

A very similar simulation setup to Zahn's potential of mean force calculations (*21*) was chosen for our path sampling and analysis, but with a few important differences. Firstly we performed equilibration MD by running long trajectories using classical CHARMM force field under ambient temperature and pressure. This was done using the CHARMM package (*30*) under the NPT ensemble. Force field parameters for N-MAA were taken from similar structures and its geometry was fixed. TIP3P force field parameters were used for water in the system and ChARMM force field was used for HCl. In the final equilibrated system, our 12.49Å×9Å×9Å simulation cell is comprised of 31 water molecules, one N-MAA, and one HCl in order to provide an acidic solution

condition, as shown in Figure 5.2. We found the value of water density in Zahn's studies is too high, while ours has the correct value at ambient temperature and pressure. Next, long (~50 ps) constant-temperature MD steps were performed with the O1-H1, the C-O1, and the H1-O2 distances as shown in as shown in Figure 5.2 constrained in Car-Parrinello molecular dynamics (CPMD) using CPMD package (*31*) in order to equilibrate other degrees of freedom.

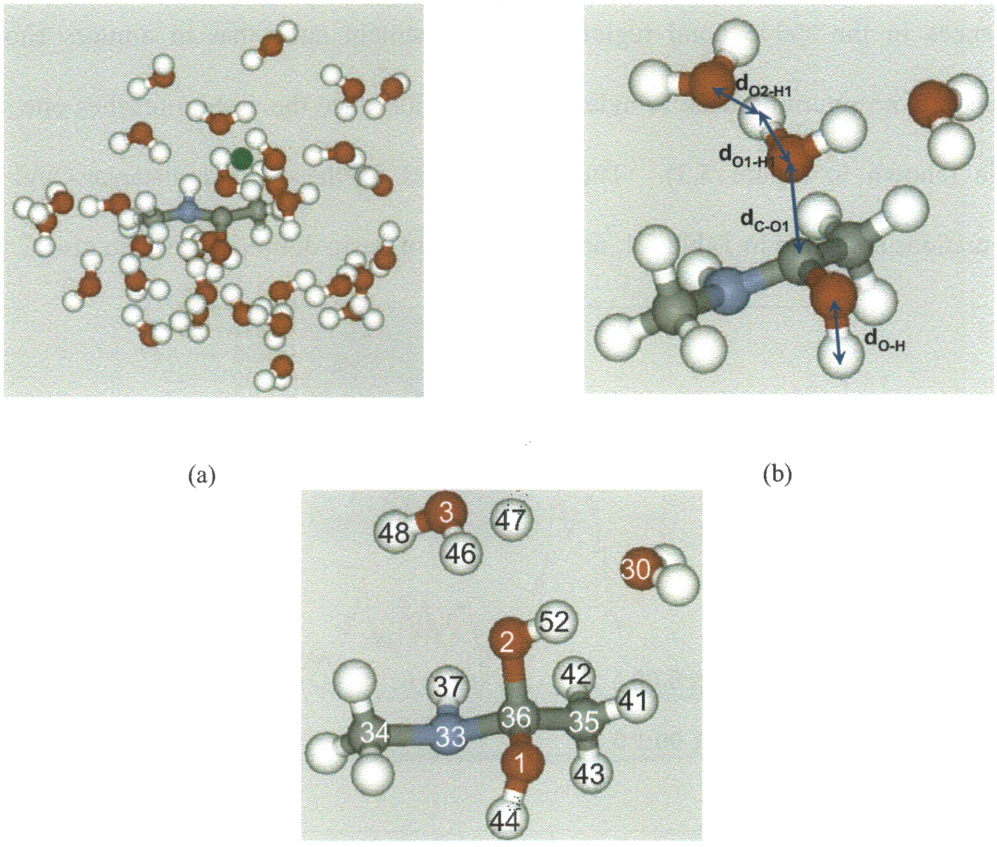


Figure 5.2: Simulation box (a) together with bond distances used to define basins of stable states (b). In (c), atom labels used in the system are shown, to be used to refer to the order parameters defined in this study.

5.3.2 Stable basin definitions.

Here three distances, the O1-H1 distance, the C-O1 distance, and the H1-O2 distance as shown in as shown in Figure 5.2 are used to tell which stable state a particular configuration corresponds to. Long molecular dynamics trajectories (~ 5ps) without any constraints/restraints were run using CPMD to obtain the fluctuations in these bond distances, and then an appropriate range was taken by computing the variances of bond distances in the stable bond regions with adjustment such that in aimless shooting procedure, no returning back to indeterminate region once the system reaches one stable basin (shown in Figure 5.3). The mean values of these bond distances and their fluctuations are listed in Table 5.1, and the choice of basin definitions is also given.

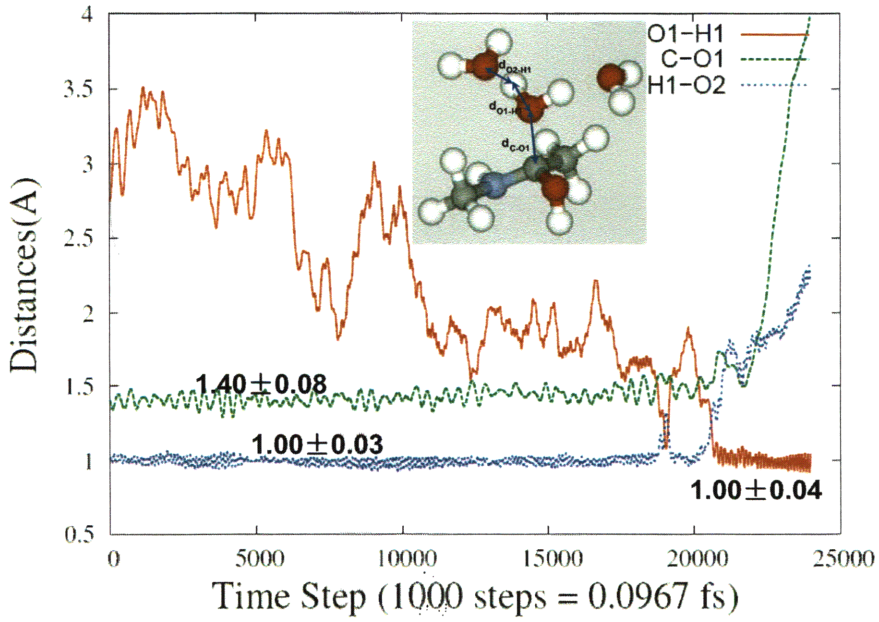


Figure 5.3: Transition trajectory with the associated changes in the OP's of bond distances. The distances are d(O1-H1), d(C-O1), and d(H1-O2), corresponding to d(O2-H46), d(C36-O1), and d(H46-O3), respectively.

Table 5.1: Ranges of bond distances in Figure 5.2 used for definitions of basins of stable states. A configuration corresponds to a particular stable state (either reactant or product) only when three bond distances are simultaneously within the specified ranges.

	fluctuation				definition	
	reactant		product		reactant	product
	mean	std	mean	std		
O1-H1 dist(Å)	1.00	0.04			[0.88, 1.12]	(1.62, +∞)
O1-C dist(Å)			1.40	0.08	(2.14,+∞)	[1.28, 1.64]
H1-O2 dist(Å)			1.00	0.03	(1.62, +∞)	[0.88, 1.12]

During our path sampling procedure, we did not observe any basin overlapping situation, in which a particular configuration satisfied both the reactant and product definitions.

5.3.3 Order parameters.

Order parameters were selected based on quite extensive screening combined with physical intuition into the reaction mechanism. More specifically, a systematic procedure for including candidate OP's was used as follows. The set of candidate OP's includes distances of all possible pairs within the system, cosine values of angles and dihedral angles for all possible triplets and quadruplets, respectively, selected from all atoms around the midpoint of C-N bond within 7.5Å cutoff. The reason for using cosine values of these angles and dihedral angles instead of their absolute values is to avoid possible discontinuity when they take on values at the boundaries of their range. This procedure generated up to a total of 5,349,859 candidate order parameters. Other types of OP's used in previous studies, such as bonding number (*32*), density fluctuation (*33*), were not considered here. Our consideration is that the exact reaction coordinate $p_B(r)$ is a function of configuration vector only, and therefore, geometric quantities are enough to describe the reaction dynamics if an suitable ensemble is chosen at first. It should also be noted that while coordination numbers may be better collective variables than specific inter-atomic distances, the committment time for this reaction is so short that there is no permutation of the active water molecules during the reaction events. Exhaustive one-OP

variable screening in likelihood maximization was done for this vast set of candidate order parameters. Sooner a combinatorial problem arises even when going to two OP variables situation. The workaround used in this studied will be discussed shortly.

5.3.4 Aimless shooting.

The aimless shooting algorithm, a modified version of transition path sampling, as described by Peters and Trout (*28, 29*), was applied to harvest an ensemble of independent trajectories according to their statistical weight. As with the transition path sampling method (*24-26*), aimless shooting requires 1) accurate definitions of the basins of stable states and 2) an initial trajectory that connects the stable basins.

The first reactive trajectory was obtained by guessing a high potential energy configuration with particular values of O2-H46, C36-O1, and H46-O3 bond distances. Then constrained MD with ~ 5ps was carried out to relax all the other degrees of freedom in the system to remove potential artifacts introduced when fixing these three distances. Both forward and backward shooting trajectories from the equilibrated configuration were then obtained with assigned velocities drawn from Maxwell-Boltzmann distribution. Repeated shootings were done until a reactive initial trajectory was obtained since basins of stable states were already defined. This resulting initial trajectory also provides some information on how fast the transition dynamics takes place, based on which the appropriate overall length of MD steps (~2000 with a time step of 4 a.u.) can be derived.

Two-point version of aimless shooting has the following procedure. Two configurations close to hypothesized transition state were selected from the initial reactive trajectory and one of the two is chosen randomly from which forward and backward half-trajectories were shot. Momenta for forward shooting are generated from

a Maxwell-Boltzmann distribution with no net linear and angular momenta for the whole system. Momenta for backward shooting were the reverse of those for forward shooting. The two configurations have a time displacement Δt which is an adjustable parameter and needs to be carefully set. If the forward and backward half-trajectories combine to give a reactive trajectory, this new trajectory is accepted and the two configurations with a time displacement Δt to the previous shooting point was recorded as a new two-point from which the shooting procedure is repeated.

As described by Beckham et al. (*34*), time displacement Δt has to be chosen appropriately to yield an acceptance ratio between 40%-60%. If it is too large, the algorithm tends to go too far away from the transition-state region leading to a low acceptance rate with many consecutive unaccepted trajectories; if it is too small, the aimless shooting algorithm will be very inefficient in exploring shooting point configuration space and therefore more trajectories and needed to obtain a good approximation to the reaction coordinate. For chemical reaction in which bond breaking and forming steps are involved, Δt is expected to be smaller than more diffusive systems, since transitions driven by strong interactions are short in terms of transition duration.

Dynamic trajectories were collected using the CPMD package (*31*) in the NVT ensemble. A time step of 4 a.u. (~ 0.1 fs) and an electron fictitious mass of 400 a.u. were used. A chain of four Nose-Hoover thermostats were used to control temperature at 298 K. The molecular orbitals were described by a plane wave basis with an energy cutoff of 70 Ry. Vanderbuilt pseudopotentials and BLYP density functionals were used. We found selecting from two points, $\mathbf{x}_{-\Delta t}$, or $\mathbf{x}_{+\Delta t}$ is sufficient to sample the transition state ensemble and that is what we did in this study.

In the aimless shooting procedure, the trajectory length is set to be as short as possible in order to save computational time, resulting in the possibility of generating inconclusive trajectories, which have at least one end point in forward and backward half-trajectories does not lie in any basin of stable state. A half-trajectory step of 2000 was found to maintain the level of inconclusive trajectories at or below 10%. A time displacement, Δt, of 25 a.u. is chosen to yield an acceptance rate of 48.2%. 1836 trajectories were collected for later analysis.

As proved in our companion work on the hydrolysis reaction of peptide bond in neutral pH, we expect the trajectories collected by aimless shooting algorithm have much de-correlation, and therefore can explore the trajectory space quite efficiently.

5.3.5 Likelihood maximization.

As described in Peters and Trout (*28*), the reaction coordinate, r, is modeled as a linear combination of candidate OP's, denoted as **q**, with α_0 through α_m as adjustable coefficients:

$$r(\mathbf{q}) = \alpha_0 + \sum_{k=1}^{m} \alpha_k q_k \tag{5.1}$$

It is noted that the choice of linear combination is for convenience purpose only, and a non-linear reaction coordinate expression could be chosen.

The model for the committor probability $p_B(r)$ was chosen to be

$$p_B(r) = [1 + \tanh(r)]/2 \tag{5.2}$$

This committor probability model was used to maximize the likelihood function with respect to the set of coefficients α_i's (i=0, …, m)

$$L(\vec{\alpha}) = \prod_{\vec{x}_k \to B} p_B(\vec{x}_k) \cdot \prod_{\vec{x}_k \to A} [1 - p_B(\vec{x}_k)] \tag{5.3}$$

only using outcomes of forward half-trajectories. In principle, if an ensemble of candidate OP's are proposed, the maximization of likelihood (3) should be performed over all combinations of OP's to determine the best reaction coordinate according to the models of Equations 1 and 2. However, when the set of candidate OP's is large, combinatorial problem arises for exhaustive screening of the best reaction coordinate. One is forced to reduce the size of the set of candidate OP's when the number of OP variables *m* increases. One choice for doing this is to do one OP variable exhaustive screening and use the best OP's for higher *m* searching, as was used in this study. For the best approximate reaction coordinate, the approximate transition state iso-surface can be obtained by setting $p_B(r) = 1/2$. This occurs at $r = 0$, so setting $r(\mathbf{q}) = 0$ defines the approximate transition state iso-surface.

As informed by other examples such as alanine-dipeptide (*33*), very complex reaction coordinate might be involved in our system. In order not be biased by any assumption about which OP's are important, an exhaustive but systematic approach was taken to find the best reaction coordinate model in likelihood maximization. Over 5 million candidate OP's were screened individually first. In order to tackle the algorithmic complexity problem in higher dimensional (d>1) likelihood maximization, the following systematic approach was adopted. Basically, it assumes that important OP's previously screened based on likelihood scores will also be important in comprising reaction coordinate models in high dimensions. As shown in Figure 5.6, it starts with best *m* one-dimensional OP's. Then in each round for *d* dimensional optimization, every best ranked *d-1* dimensional result is supplemented with every best *m* one-dimensional OP to

give a *d* dimensional model. Then only *m* best d dimensional results are retained for *d+1* dimensional screening. This way, each round roughly has *m*m* optimization problems to solve.

5.3.6 Reaction coordinate validation.

After the likelihood maximization to generate an approximation to reaction coordinate, its correctness must be checked. This can be done by computing the estimate of the probability of reaching product basin (p_B) from the predicted transition state region obtained in likelihood maximization as commonly referred as a committor distribution analysis, or p_B histogram computation[27]. In this procedure, independent configurations are generated which all satisfy the predicted transition state. Then a number of trajectories are initiated with the momenta drawn from a Maxwell-Boltzmann distribution from these configurations and the estimate of p_B values for these configuration can thus be obtained. Then a histogram of the number of configurations versus p_B values can be constructed.

For complex reaction coordinate like the ones used in this study, generating independent configurations for p_B histogram computation can be done efficiently by the BOLAS algorithm (*35*). First, shooting points were examined and several of them were selected close to the predicted transition state region, as defined by $r(\mathbf{q}) = 0$ in Equation 1. Very short trajectories are fired randomly from each initial configuration and the endpoints are evaluated to determine if they are within a narrow window on the transition state iso-surface. If so, this configuration is accepted and becomes the next shooting point. This process is repeated until an adequate number of configurations is generated from which to shoot reactive trajectories to build a p_B histogram.

To construct the histogram, trajectories are shot from each configuration with a length corresponding to half the length of a reactive trajectory. The endpoints of the trajectories are evaluated and a histogram is constructed of the probability of reaching basin B from the predicted transition state isosurface. The basin definitions for constructing the p_B histograms correspond to the same basin definitions used for the reactant and product basins in the aimless shooting simulations. An adequate approximation to the true reaction coordinate will yield a histogram that is sharply peaked at $p_B = 0.5$ (*26*). Additionally, one can make a quantitative comparison of the histogram to the binomial distribution, which will have a mean value, $\mu = 0.50$ with a standard deviation, $\sigma = 0.050$.

The trajectories for the generation of new configurations are ~5 fs or 50 MD steps long, and the endpoint window width at $r = 0$ is constrained within a range of ±1% of the total configuration space sampled, as measured by Δr. For each histogram assembled in this study, 20 shooting points are collected. From each configuration collected, 20 trajectories are shot, corresponding to approximately 400 trajectories for each histogram. The trajectory length for calculating p_B values is ~200 fs or 2000 MD steps, which is half the length of the reactive trajectories in the aimless shooting simulations, again resulting in a low rate of inconclusive paths.

5.4 Results and Discussions

5.4.1 Initial trajectory.

As mentioned previously, an initial reactive trajectory was obtained by first running long equilibration MD when fixing postulated bond distances of O2-H46, C36-

O1, and H46-O3 labeled in Figure 5.2 in order to remove potential artifacts when using constraints. Then repeated forward and backward shootings were tried until the two half-trajectories combined to yield a reactive trajectory. Figure 5.3 shows the how bond distances change in this particular trajectory, together with extended sampling in the stable states in order to see how these bond distances fluctuate. It is clear that the bond distances in stable states compare well with known experimental values, and that the transition dynamics has a time scale ~ 4000 MD steps. This provides a basis for choosing the MD steps in aimless shooting procedure to reach the compromise of efficiency and minimal number of inconclusive trajectories.

5.4.2 Trajectory characteristics.

Snapshots from a typical reactive trajectory shows what happens during the transition as shown for the reaction core part in Figure 5.4. Most of the trajectories collected showed the rate-determining step of hydrolysis proceeds in a concerted fashion, i.e., one proton transfer process between two nearby water molecules and C-O bond formation occur simultaneously. The initial step of protonation of carbonyl oxygen of the peptide bond was also observed in some trajectories, in agreement with the conclusion that this initial step does not lead to a stable intermediate.

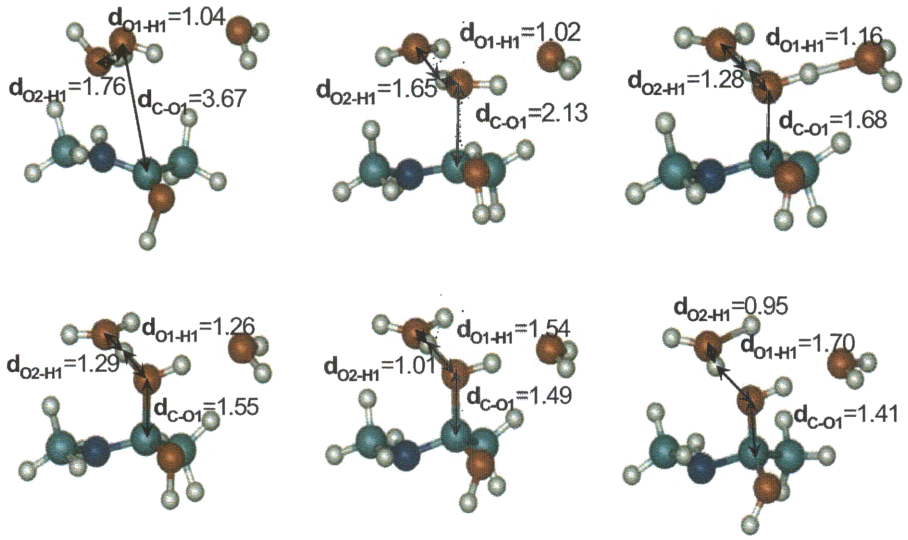


Figure 5.4: Key snapshots describing the rate-determining step in acid-catalyzed hydrolysis pathway. Only three water molecules are shown for clarity. The overall trajectory is 4000 MD steps, during which the formation of intermediate I1 occurs. C-O bond formation and a proton transfer step are concerted.

5.4.3 Statistics of trajectory ensemble.

In Figure 5.5, the total number of four different types of trajectories is plotted against trajectory index recorded in the aimless shooting procedure. Type 1 and 2 trajectories are reactive, since they connect both reactant and product states; while type 3 and type 4 are rejected shooting moves. As required in configurational Monte Carlo sampling, the acceptance ratio, defined as the ratio of the number of reactive trajectories to that of total, should be around 40%-50% for both efficiency and sampling speed in exploring trajectory space. As just stated, our choice of Δt yielded 44.8% acceptance ratio, which is rather good. It is expected in the aimless shooting algorithm, the number of trajectories of type 1 at any point in the accumulation of trajectory ensemble should be

approximately equal to type 2 due to random nature of assigning initial shooting velocities, as is case in Figure 5.5. One also expects the number of trajectories of type 3 should be approximately equal to 4 if the trajectory space was sufficiently explored during the aimless shooting procedure. Here, more product side of transition state region was explored, because there are more 4 type trajectories than 3 type trajectories.

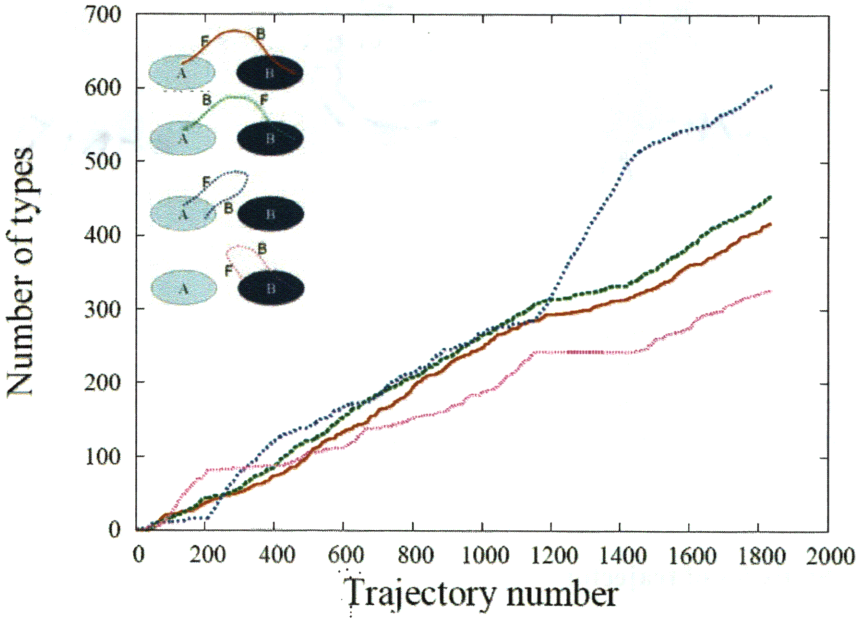


Figure 5.5: Accumulation of four different types of trajectories in the procedure of aimless shooting for acid-catalyzed hydrolysis of peptide bond. These types are: forward half trajectory having reached (reactant) basin A and backward half trajectory having reached (product) basin B (type 1), forward half trajectory having reached basin B and backward half trajectory having reached basin A (type 2), both forward and backward half trajectories having reached basin A (type 3), both forward and backward half trajectories having reached basin B (type 4).

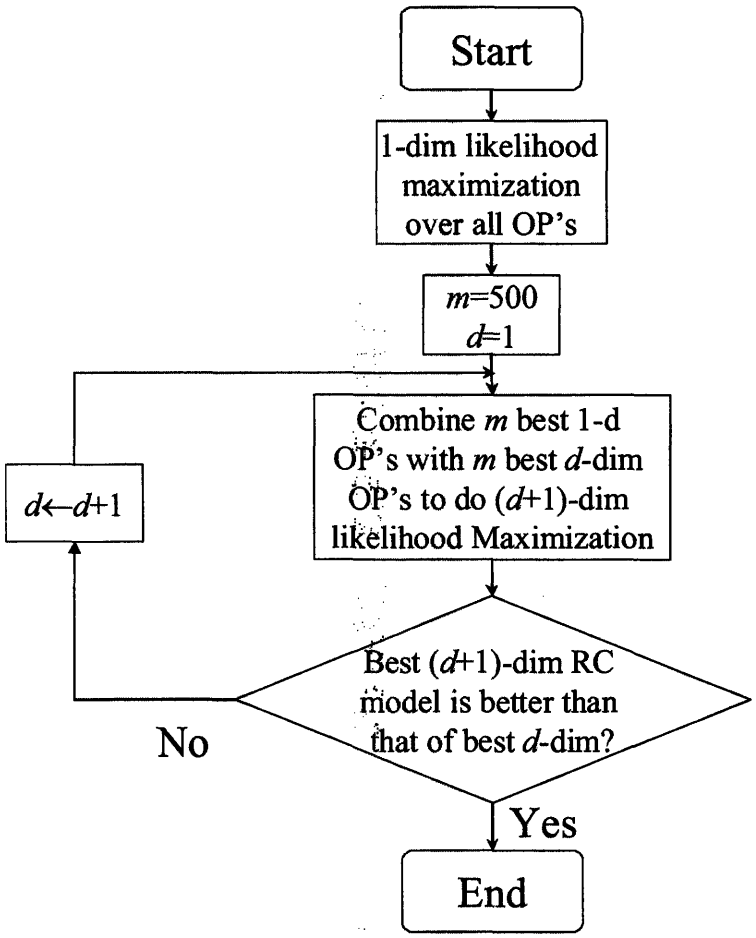


Figure 5.6: Flowchart showing how likelihood maximization was carried out when more OP variables are included in comprising a RC model. Essentially every iteration for more than two OP variable RC model screened approximately m*m models.

5.4.4 Likelihood maximization.

The results of likelihood maximization are shown in Table 5.2. A few best reaction coordinate models with different number of OP variables up to five are listed. In addition, one-OP variable best reaction coordinate models are also pictorially shown in Figure 5.8. Based on likelihood scores, it is much better in describing the hydrolysis reaction in the statistical sense of likelihood maximization than the order parameter of the distance between O(2)-C(36), which was used in the calculation of potential-of-mean-

force for the rate-limiting step of the hydrolysis of N-MAA under acidic pH condition (*21*). As more OP's used in the linear combination expression of reaction coordinate model in (1), higher and higher log-likelihood scores were obtained, suggesting a complex reaction mechanism involving many physical degrees of freedom. In our approximation scheme for likelihood maximization procedure when including more OP variables in the reaction coordinate models, log-likelihood score achieved to be within BIC criterion up to three-OP variable, implying a convergent result.

Table 5.2: Likelihood maximization results for N=1836 aimless shooting paths, with a BIC=log(N/2)=3.758. The order parameters (OP's) have the following meaning: d(n1,n2) is the distance between atom number n1 and n2, a(n1,n2, n3) is the angle comprised of atom number n1, n2 and n3, phi(n1,n2, n3,n4) is the dihedral angle comprised of atom number n1, n2, n3 and n4. The column α's gives the vector $\alpha=(\alpha_0, \alpha_1, \ldots, \alpha_n)$ corresponding to reduced and normalized OP $q_i \in [0, 1]$.

Number of OP variables in RC model	OP's in best ranked RC models	Likelihood Score(ln(L))	α's
1	phi(O22-O15-O20-H46)	-1062.013	0.867,-1.907
	phi(H46-O25-H75-H93)	-1065.467	-1.175, 2.039
	phi(O22-O25-C34-H46)	-1066.497	0.896,-2.003
	phi(O22-O17-O24-H46)	-1066.692	0.794,-1.842
	phi(O22-O6-H70-H46)	-1068.820	0.802,-1.857
	d(O2-C36)[a]	-1178.543	0.216,-0.745
2	a(O22-O24-H46), phi(O2-O20-82-H46)	-1041.097	-0.395, 1.501,-1.149
	phi(O2-O20-H82-H46), d(O22-H46)	-1041.910	1.176,-1.191,-1.470
	a(O22-H89-H46), phi(O2-O20-H82-H46)	-1042.756	-0.308, 1.397,-1.143
	phi(O2-O20-H82-H46), phi(O22-H68-H89-H46)	-1043.599	1.103,-1.176,-1.340
	a(O22-O15-H46), phi(O15-H46-H105-H55)	-1046.120	-1.377, 1.358, 1.349
3	phi(O15-H46-H105-H55), d(O22-H46), phi(O2-O20-H82-H46)	-1035.572	0.537, 0.798,-1.189,-0.877
	a(O22-H106-H46), d(O22-H46), phi(O2-O20-H82-H46)	-1035.881	0.527, 0.728,-1.115,-0.941
	a(O22-O24-H46), phi(O15-H46-H105-H55), phi(O2-O20-H82-H46)	-1036.748	-0.659, 1.208, 0.696,-0.899
	a(O22-O32-H46), d(O22-H46), phi(O2-O20-H82-H46)	-1037.456	0.627, 0.582,-1.198,-0.910
	a(O22-O24-H46), phi(O2-O20-H82-H46), phi(O25-H46-H106-H103)	-1037.498	0.011, 1.161,-0.990,-0.552

	phi(O15-H46-H105-H55), d(O22-H46), phi(O2-O20-H82-H46)	-1035.572	0.537, 0.798,-1.189,-0.877
4	a(O22-O24-H46), phi(O2-O20-H82-H46), phi(H80-O25-H46-H106), phi(H46-O25-H106-H80)	-1032.876	-5.897, 1.147,-1.053, 5.913, 5.490
	a(O22-H106-H46), d(O22-H46), phi(O2-O20-H82-H46), phi(O3-O25-H75-H46)	-1033.518	0.196, 1.120,-1.398,-1.071, 0.676
	phi(O15-H46-H105-H55), d(O22-H46) phi(O2-O20-H82-H46), phi(O3-O20-O25-H46)	-1033.568	0.863, 1.138,-1.478,-0.994,-0.560
	phi(O15-H46-H105-H55), d(O22-H46), phi(O2-O20-H82-H46), phi(O22-O15-H75-H46)	-1033.700	0.441, 0.939,-1.720,-0.990, 0.737
	phi(O15-H46-H105-H55), d(O22-H46), phi(O2-O20-H82-H46), phi(O22-O15-O17-H46)	-1033.739	0.408, 0.950,-1.901,-0.954, 0.910
	a(O22-O24-H46), phi(O2-O20-H82-H46), phi(H80-O25-H46-H106), phi(H46-O25-H106-H80)	-1032.876	-5.897, 1.147,-1.053, 5.913, 5.490

[a]This order parameter was used in previous potential-of-mean-force calculation(*21*).

The reaction coordinate models for both the RC models with best two-OP variable and the three-OP variable were checked against the aimless shooting data, as shown in Figure 5.7. All accepted shooting points for which the forward and backward shootings led to a conclusive trajectory were used to construct two histograms of the reaction coordinate determined from likelihood maximization. The two histograms were based on only on the half-trajectories of the forward shooting from each accepted shooting point whether it ends in reactant basin or in product basin. Then $p_B(r)$ data values in Figure 5.6 were computed as the ratio

$$p_B(r)\,|_{data,i\text{ th bin}} = \frac{N_{fB,i}}{N_{fA,i} + N_{fB,i}} \tag{5.4}$$

where $N_{fA,i}$ and $N_{fB,i}$ stands for the number of shooting points giving the reaction coordinate model value in i th bin that led to forward shooting half-trajectory ends in A, or B respectively. Thus, the comparison of model vs. data provides a measure of how well aimless shooting collects information about transition paths and shooting points, as

well as how good likelihood maximization is calculated. In Figure 5.7, one can see that both reaction coordinate models give satisfactory fit to the aimless shooting data.

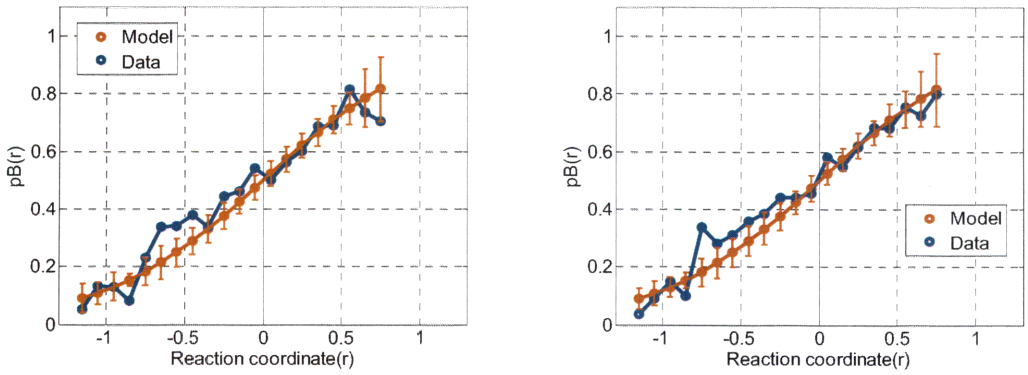


Figure 5.7: Comparison of $p_B(r)$ model vs. aimless shooting data. Here half trajectory $p_B(r)$ model(*28*) was used, i.e. $p_B(r)=[1+\tanh(r)]/2$. Note that the error bars appear on the model, not the data. The error bars show how far shooting point data should deviate from the probabilities $p_B(r)$ for a perfect reaction coordinate model. (a) two-OP variable reaction coordinate model (b) three-OP variable reaction coordinate model.

Using equation shown in (4), the statistical uncertainty in the log-likelihood score was computed with the collection of accepted shooting points to be $\sigma^2(\ln L)=15.33$ and $\sigma^2(\ln L)= 14.60$ for best RC models with two- and three-OP variable respectively. These values provide reasonable confidence in our log-likelihood score value (*36*).

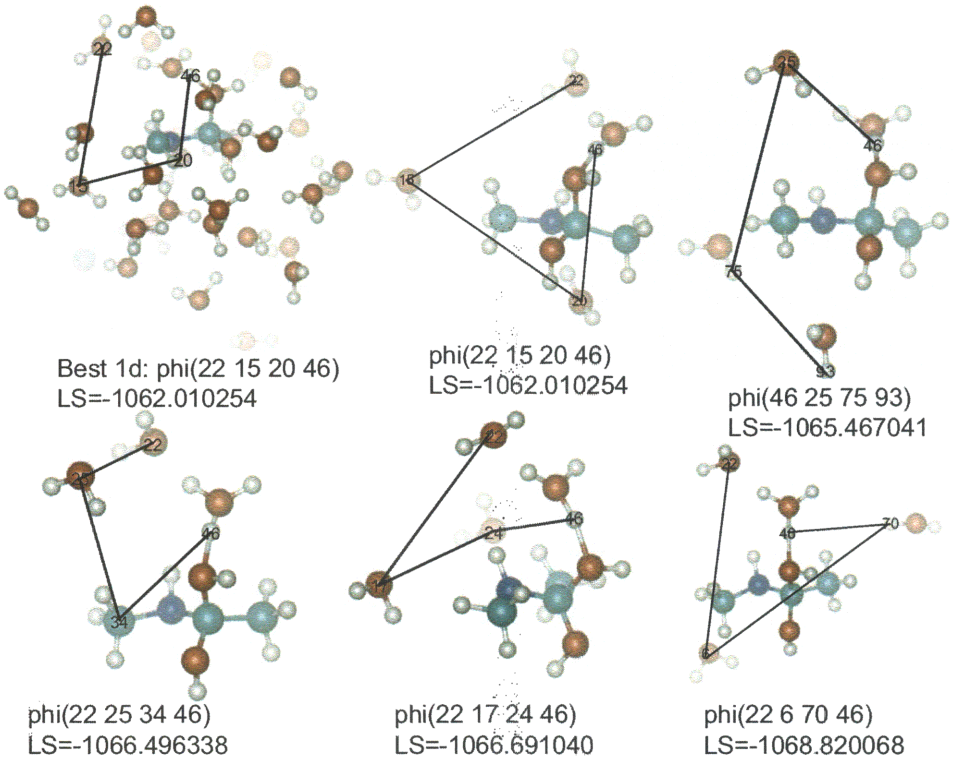


Figure 5.8: Illustration of best ranked one-dimensional OP's in the likelihood maximization procedure. The OP is given as a dihedral among quadruple atoms (denoted as phi(atom_index_1, atom_index_2, atom_index_3, atom_index_4), or as an angle among triple atoms (denoted as a(atom_index_1, atom_index_2, atom_index_3), or as a bond distance between pair atoms (denoted as d(atom_index_1, atom_index_2). The associated likelihood scores (LS) are also given. One observation is that almost all these best ranked involves the hydrogen atom indexed as 46:H.

Figure 5.9 showed the OP's which comprising the best reaction coordinate model with three OP variables. Similar to our study on the hydrolysis reaction under neutral pH, these OP's include the local bonding pattern changes, such as proton H46 being transferred between the two water molecules, namely the O22-H46 distance. Solvent degrees of freedom in affecting reaction dynamics are also needed, as seen by the presence of two dihedral angles phi(O15-H46-H105-H55) and phi(O2-O20-H82-H46). The inclusion of both local OP's close to the reaction center and the OP's describing the solvent networks suggests the importance of solvent in determining reaction dynamics.

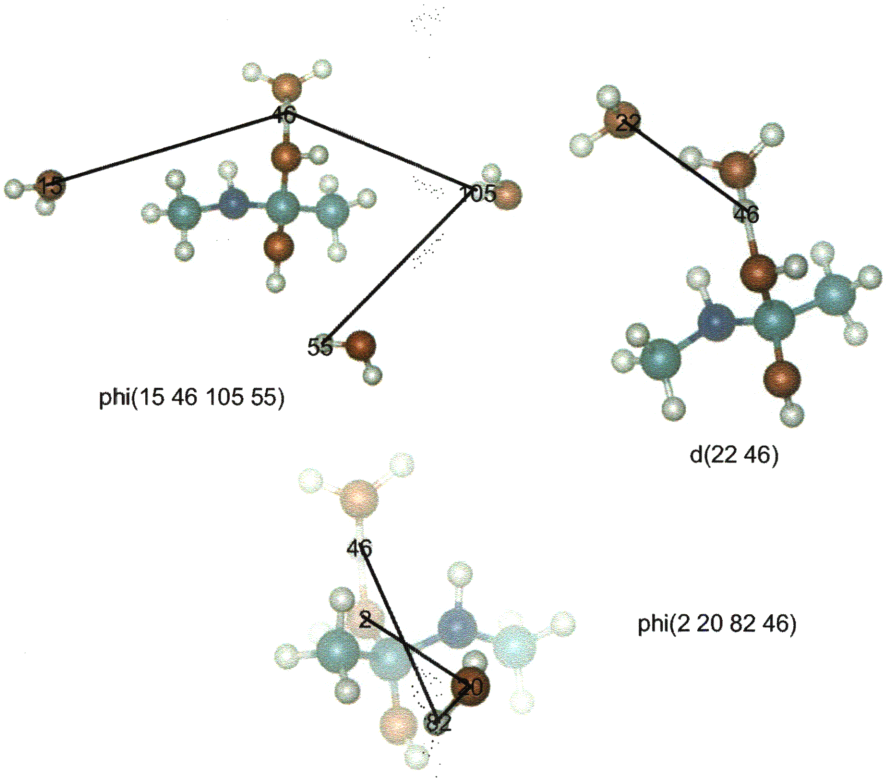


Figure 5.9: Illustration of constituent OP's in the best 3-OP variable RC model. Naming of these OP's is the same as in Figure 5.8.

5.4.5 Reaction coordinate validation.

Four p_B histograms were computed using the method described above. These include using the C36-O2 distance reaction coordinate model, the best one-, two-, and three-OP variable RC model. The results are shown in Figure 5.10. The poor description of the reaction process by the C-O distance RC model can be seen in this p_B histogram, since its distribution is very skewed. Both the best one- and the best two-OP variable RC models show inferior histogram than the best three-OP variable RC model.

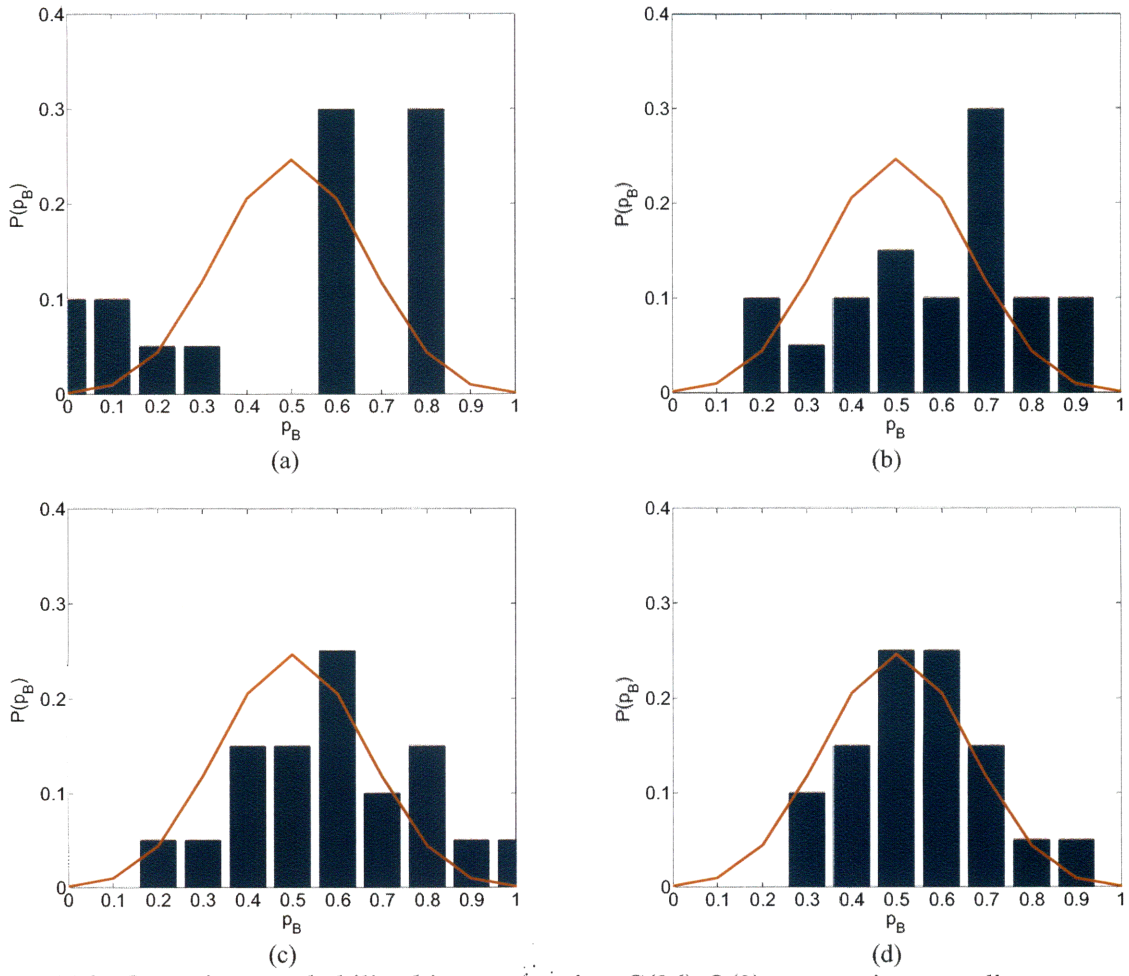


Figure 5.10: Committor probability histogram using C(36)-O(2) as reaction coordinate model (a), best one-OP variable reaction coordinate model (b), best two-OP variable reaction coordinate model (c) and best three-OP variable reaction coordinate model (d), compared with binomial distribution (red line). Quantification of means and standard deviations for these histograms following the procedure in Peters(*36*) is shown in Table 5.3.

5.5 Summary and conclusions

In this study, the mechanism of the rate-determining step of hydrolysis reaction of peptide bond under acidic pH condition was studied using a model compound N-MAA. Due to fluctuations at finite temperature, path sampling method was used to generate an ensemble of trajectories according to their statistical weight in trajectory space. *Ab initio*

molecular dynamics technique was applied to advance the time evolution of the reaction and collect trajectories. Likelihood maximization and its modification were used in extracting physically important degrees of freedom in the system and approximations of the reaction coordinate were compared. It was found that this hydrolysis reaction is very complex in nature, and involves many degrees of freedom. The specific conclusions obtained in our study are:

- Rate-determining step of the hydrolysis reaction of N-MAA under acidic pH occurs in a concerted fashion; a stable intermediate was found to be the final state in our path sampling simulations.
- Likelihood maximization procedure was extended to screen RC models with more OP variable, and within BIC, a reaction coordinate with three constituent geometric variables was found to be the best in describing the path ensemble we generated.
- In the best RC model, both geometric quantities which reflect bonding pattern changes, and those which reflect the solvent network changes, are included, suggesting a complicated reaction involving many degrees of freedom.
- Several p_B histograms were computed to verify the results of likelihood maximization, and the quantified goodness of these best-ranked reaction coordinate models is in accord with the their respective likelihood score.

Table 5.3: Maximal likelihood estimates for means and standard deviations in the p_B histograms shown in Figure 5.8. The procedure was used as in (*36*).

reaction coordinate model/distribution	μ_h	σ_h
C(36)-O(2)	0.565	0.109
best one-OP variable	0.590	0.042
best two-OP variable	0.595	0.040

best three-OP variable	**0.550**	**0.023**
binomial distribution	**0.500**	**0.025**
ideal $P(P_B)$	**0.500**	**0.000**

5.6 References

(1) Chu, J. W., Yin, J., Brooks, B. R., Wang, D. I. C., Ricci, M. S., Brems, D. N., and Trout, B. L. (2004) A comprehensive picture of non-site specific oxidation of methionine residues by peroxides in protein pharmaceuticals. *Journal of Pharmaceutical Sciences 93*, 3096-3102.

(2) Pan, B., Abel, J., Ricci, M. S., Brems, D. N., Wang, D. I. C., and Trout, B. L. (2006) Comparative oxidation studies of methionine residues reflect a structural effect on chemical kinetics in rhG-CSF. *Biochemistry 45*, 15430-15443.

(3) Wei, W. (1999) Instability, stabilization, and formulation of liquid protein pharmaceuticals. *International Journal of Pharmaceutics 185*, 129-188.

(4) Daugherty, A. L., and Mrsny, R. J. (2006) Formulation and delivery issues for monoclonal antibody therapeutics. *Advanced Drug Delivery Reviews 58*, 686-706.

(5) Cohen, S. L., Price, C., and Vlasak, J. (2007) beta-elimination and peptide bond hydrolysis: Two distinct mechanisms of human IgG1 hinge fragmentation upon storage. *Journal of the American Chemical Society 129*, 6976-+.

(6) Cordoba, A. J., Shyong, B. J., Breen, D., and Harris, R. J. (2005) Non-enzymatic hinge region fragmentation of antibodies in solution. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences 818*, 115-121.

(7) Dillon, T. M., Bondarenko, P. V., Rehder, D. S., Pipes, G. D., Kleemann, G. R., and Ricci, M. S. (2006) Optimization of a reversed-phase high-performance liquid chromatography/mass spectrometry method for characterizing recombinant antibody heterogeneity and stability. *Journal of Chromatography A 1120*, 112-120.

(8) Dillon, T. M., Ricci, M. S., Rehder, D. S., Flynn, G., Liu, Y. D., and Bondarenko, P. V. (2007) Discovery and characterization of conformational isoforms of human monoclonal IgG2 antibodies. *Nature*.

(9) Kahne, D., and Still, W. C. (1988) Hydrolysis of a Peptide-Bond in Neutral Water. *Journal of the American Chemical Society 110*, 7529-7534.

(10) Brown, R. S., Bennet, A. J., and Slebockatilk, H. (1992) Recent Perspectives Concerning the Mechanism of H3O+-Promoted and Oh--Promoted Amide Hydrolysis. *Accounts of Chemical Research 25*, 481-488.

(11) Bryant, R. A. R., and Hansen, D. E. (1996) Direct measurement of the uncatalyzed rate of hydrolysis of a peptide bond. *Journal of the American Chemical Society 118*, 5498-5499.

(12) Radzicka, A., and Wolfenden, R. (1996) Rates of uncatalyzed peptide bond hydrolysis in neutral solution and the transition state affinities of proteases. *Journal of the American Chemical Society 118*, 6105-6109.

(13) Smith, R. M., and Hansen, D. E. (1998) The pH-rate profile for the hydrolysis of a peptide bond. *Journal of the American Chemical Society 120*, 8910-8913.

(14) Krug, J. P., Popelier, P. L. A., and Bader, R. F. W. (1992) Theoretical-Study of Neutral and of Acid and Base Promoted Hydrolysis of Formamide. *Journal of Physical Chemistry 96*, 7604-7616.

(15) Antonczak, S., Ruizlopez, M. F., and Rivail, J. L. (1994) Ab-Initio Analysis of Water-Assisted Reaction-Mechanisms in Amide Hydrolysis. *Journal of the American Chemical Society 116*, 3912-3921.

(16) Bakowies, D., and Kollman, P. A. (1999) Theoretical study of base-catalyzed amide hydrolysis: Gas- and aqueous-phase hydrolysis of formamide. *Journal of the American Chemical Society 121*, 5712-5726.

(17) Stanton, R. V., Perakyla, M., Bakowies, D., and Kollman, P. A. (1998) Combined ab initio and free energy calculations to study reactions in enzymes and solution: Amide hydrolysis in trypsin and aqueous solution. *Journal of the American Chemical Society 120*, 3448-3457.

(18) Zahn, D. (2004) Car-Parrinello molecular dynamics simulation of base-catalyzed amide hydrolysis in aqueous solution. *Chemical Physics Letters 383*, 134-137.

(19) Zahn, D. (2004) Investigation of the complex catalyzed amide hydrolysis from reaction coordinate of acid molecular dynamics simulations. *Chemical Physics 300*, 79-83.

(20) Zahn, D. (2004) On the role of water in amide hydrolysis. *European Journal of Organic Chemistry*, 4020-4023.

(21) Zahn, D. (2003) Theoretical study of the mechanisms of acid-catalyzed amide hydrolysis in aqueous solution. *Journal of Physical Chemistry B 107*, 12303-12306.

(22) Zahn, D., Schmidt, K. F., Kast, S. M., and Brickmann, J. (2002) Quantum/classical investigation of amide protonation in aqueous solution. *Journal of Physical Chemistry A 106*, 7807-7812.

(23) Wang, W., Singh, S., Zeng, D. L., King, K., and Nema, S. (2007) Antibody structure, instability, and formulation. *Journal of Pharmaceutical Sciences 96*, 1-26.

(24) Bolhuis, P. G., Chandler, D., Dellago, C., and Geissler, P. L. (2002) Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual Review of Physical Chemistry 53*, 291-318.

(25) Dellago, C., Bolhuis, P. G., Csajka, F. S., and Chandler, D. (1998) Transition path sampling and the calculation of rate constants. *Journal of Chemical Physics 108*, 1964-1977.

(26) Dellago, C., Bolhuis, P. G., and Geissler, P. L. (2002) Transition path sampling. *Advances in Chemical Physics, Vol 123 123*, 1-78.

(27) Camargo, A. C. M., Gomes, M. D., Reichl, A. P., Ferro, E. S., Jacchieri, S., Hirata, I. Y., and Juliano, L. (1997) Structural features that make oligopeptides susceptible substrates for hydrolysis by recombinant thimet oligopeptidase. *Biochemical Journal 324*, 517-522.

(28) Peters, B., Beckham, G. T., and Trout, B. L. (2007) Extensions to the likelihood maximization approach for finding reaction coordinates. *Journal of Chemical Physics 127*, -.

(29) Peters, B., and Trout, B. L. (2006) Obtaining reaction coordinates by likelihood maximization. *Journal of Chemical Physics 125*, -.

(30) Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983) Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry 4*, 187-217.

(31) CPMD, C. I. C.-. Copyright MPI für Festkörperforschung Stuttgart 1997-2001.

(32) Geissler, P. L., Dellago, C., Chandler, D., Hutter, J., and Parrinello, M. (2001) Autoionization in liquid water. *Science 291*, 2121-2124.

(33) Ma, A., and Dinner, A. R. (2005) Automatic method for identifying reaction coordinates in complex systems. *Journal of Physical Chemistry B 109*, 6769-6779.

(34) Beckham, G. T., Peters, B., Starbuck, C., Variankaval, N., and Trout, B. L. (2007) Surface-mediated nucleation in the solid-state polymorph transformation of terephthalic acid. *Journal of the American Chemical Society 129*, 4714-4723.

(35) Radhakrishnan, R., and Schlick, T. (2004) Biomolecular free energy profiles by a shooting/umbrella sampling protocol, "BOLAS". *Journal of Chemical Physics 121*, 2436-2444.

(36) Peters, B. (2006) Using the histogram test to quantify reaction coordinate error. *Journal of Chemical Physics 125*, -.

Chapter 6. A Coarse-grained Model of Peptide Bond Hydrolysis in Antibody

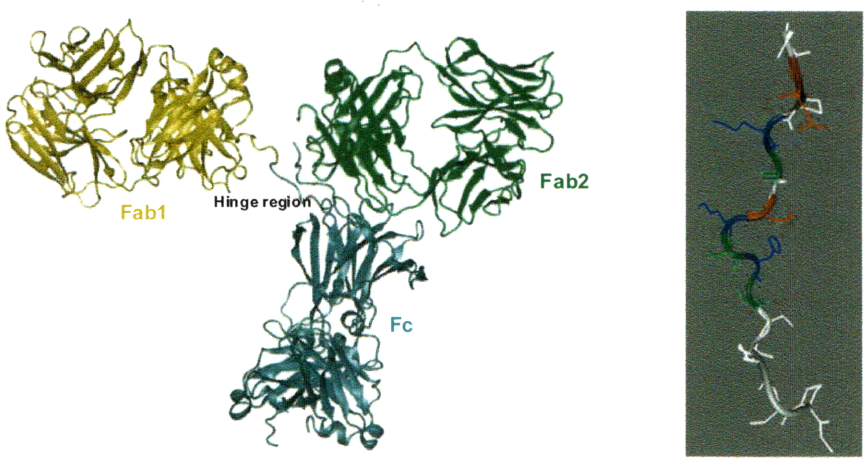
6.1 Introduction

The chemical stability of peptide bonds in protein molecules is essential for life, as well as for the applications of protein pharmaceutics. Recombinant protein molecules can have chemical degradation pathways such as oxidation (*1, 2*), deamidation, hydrolysis of peptide bond on the protein backbone, etc (*3*). Monoclonal antibody molecules (*4*), as a very promising class of bio-pharmaceutics, have been reported by several different groups to undergo non-enzymatic fragmentation in the hinge region (*5-8*). This was identified mainly as the hydrolysis of several peptide bonds in the hinge region. We believe that understanding of the underlying reaction mechanism particularly from a molecular perspective will help to minimize the extent of degradation in a rational and more efficient way. Examples of this approach include manipulating solvent accessibility to control methionine oxidation after identifying that water exposure is a key parameter in the oxidation process (*1*).

Under normal conditions without enzymatic activity, the hydrolysis reaction of peptide bonds is very slow, with a half-life on the order of hundreds of years (*9*). However, the rate of this reaction increases by several orders of magnitude when solution pH changes, either in lower or higher pH's (*10*). Cordoba et al. (*6*) observed several percent hydrolytic degradation of residues in the hinge region of IgG-1 antibody molecules over an incubation period of several months. Their study showed the hydrolysis could occur at different peptide bonds in the hinge with a variable extent of

degradation. They also showed that the reaction was un-catalyzed around neutral pH or slightly acidic pH conditions. Dillon et al. (*7, 8*) developed new analytical techniques, and reached similar conclusions about the location and extent of hydrolytic cleavage of peptide bonds in hinge region. Thus it is practically interesting to understand the underlying mechanism of enhanced hydrolysis reaction in both acidic and neutral pH conditions, since most antibody therapeutics are formulated under these conditions (*11*).

Genetically modified human antibody molecules are often produced in simpler microorganisms such as *Escherichia coli* or *Saccharomyces cerevisiae* by recombinant protein technologies, and are formulated in particular solutions in order to maximize their chemical and physical stability. Antibody molecules have a special structure, which is crucial for their biological function. Figure 6.1 shows a crystal structure of human IgG1 b12, obtained by Saphire et. al. (*12*). An antibody molecule is comprised of two heavy chains with identical amino acid sequences, as well as two light chains that also have identical amino acid sequences. Even though in terms of sequence the two heavy and two light chains are the same, their atomic coordinates in the crystal structure are highly asymmetric. A key part of the structure of the antibody molecule is the peptide segment between the C_H1 and the C_H2 domains, the hinge region, which connects both the Fab fragments and the Fc fragment. The hinge region is highly susceptible to attack by proteases, giving rise to Fab, $F(ab')_2$, and Fc fragments. A number of studies (*11, 13-19*) have shown the importance of the hinge region in terms of antibody biological function and its stability. For example, Tan et al. (*18*) and Dall'Acqua et al. (*19*) studied the effects of length and composition of the hinge region on the segmental flexibility, and its various biological activities.



(a) (b)

Figure 6.1: The crystal structure of human IgG-1 b12 (*12*) (a) and one of its hinge fragments (b).

In explaining their finding of the occurrence of un-catalyzed hydrolysis of peptide bonds in the hinge region, Cordoba et al. (*6*) speculated that the local flexibility of the hinge region may lower the activation energy needed for hydrolytic cleavage. But exactly how this is achieved was not detailed in their study. In this work, two approaches with detailed molecular and coarse-grained modeling were pursued with the goal of identifying the cause of the higher rate of hydrolysis in the hinge region, based on two hypotheses. One hypothesis is that the local solvent accessibility may be higher for the more flexible hinge region, resulting in its higher rate of hydrolytic cleavage. Our approach to test this hypothesis is to find whether the dynamics and structural features of the hinge fragment in the antibody are significant compared with a free hinge in order to have a correlation with the higher hydrolysis rates. The other hypothesis is that the hinge region as a weak link connecting the denser and more modular Fab and Fc domains is constrained to assume different configurations than a free hinge fragment, and it may experience net pulling forces over a time scale which is much longer than the intrinsic hydrolysis reaction time, ultimately resulting in the higher rate of hydrolysis. We tested

our hypothesis by obtaining the forces in the reaction coordinate direction from a classical molecular dynamics trajectory, and computing the resulting reaction rate constant when subjecting to these forces to see whether it is larger than without the forces.

Considering the special arrangement of globular and bulkier Fc and Fab parts with respect to the hinge region, the second hypothesis to explain the higher rate of hydrolysis of peptide bonds in the hinge region is proposed based on mechano-chemical phenomena. Mechano-chemistry, as reviewed by Beyer (*20*), involves the activation of covalent bonds by the presence of an external mechanical force. Recently, mechanical means was also used to bias the reaction pathways to generate products not obtainable from thermal or light-induced reactions (*21*). Our hypothesis states that due to the random movement of the Fc and Fab parts of antibody molecule, pulling forces exerted on the hinge fragment by such concerted motion result in the higher rate of breakage of peptide bonds in the hinge region.

In a companion work, we studied the hydrolysis reaction of peptide bonds under neutral pH conditions by using a model compound N-methyl acetyl acrylamide (N-MAA). *Ab initio* molecular dynamic simulations at finite temperature were carried out to sample the transition trajectory space and an approximation to the reaction coordinate was obtained. The reaction coordinate is the dynamic characterization of the reaction progress from the reactant state to the product state, a physically relevant order parameter which condenses the information on presumably a huge number of degrees of freedom into one. The projection of free energy surface onto the reaction coordinate direction gives the free energy profile, and thus the dynamic relevant free energy barrier, from which the reaction rate constant is made easier to compute. The approximate reaction

coordinate is used in our second approach to obtain the forces projected onto this one-dimensional reaction coordinate to see whether it has any effect on the rate of the hydrolysis reaction.

6.2 Overview

In order to reveal the difference in the dynamic features of hinge fragments free in solution and situated inside the antibody molecule, molecular dynamic (MD) simulations of both systems in explicit solvent water molecules were performed. The whole antibody MD simulation was performed by Chennamsetty et. al, and the dynamic trajectory was used directly for the analysis presented here. A separate MD simulation for the free hinge in explicit solvent was performed with the initial structure taken from the crystal structure (*12*). For the test of our first hypothesis, several structural and dynamic quantities were calculated from both MD trajectories. These included the root-mean-square-deviation (RMSD), root-mean-square-fluctuation (RMSF), end-to-end distance of the hinge fragment, dynamic solvent accessible area, and water coordinate number for each peptide bond in the hinge fragment.

In order to test the second hypothesis, we assumed the projected forces along the reaction coordinate direction have the same distribution from the reactant state to the product state; therefore, the projected forces only in the reactant basin are needed. Also, because the mathematical expression of the one-dimensional reaction coordinate is based on near-transition-state configurations in our *ab initio* work and exclusively involves geometric quantities such as distances, angles, and dihedral angles, one correspondence rule needs to be developed in identifying the atoms involved in comprising the RC

expression. The force vector in each frame of the classical MD trajectory was calculated before the projection was computed.

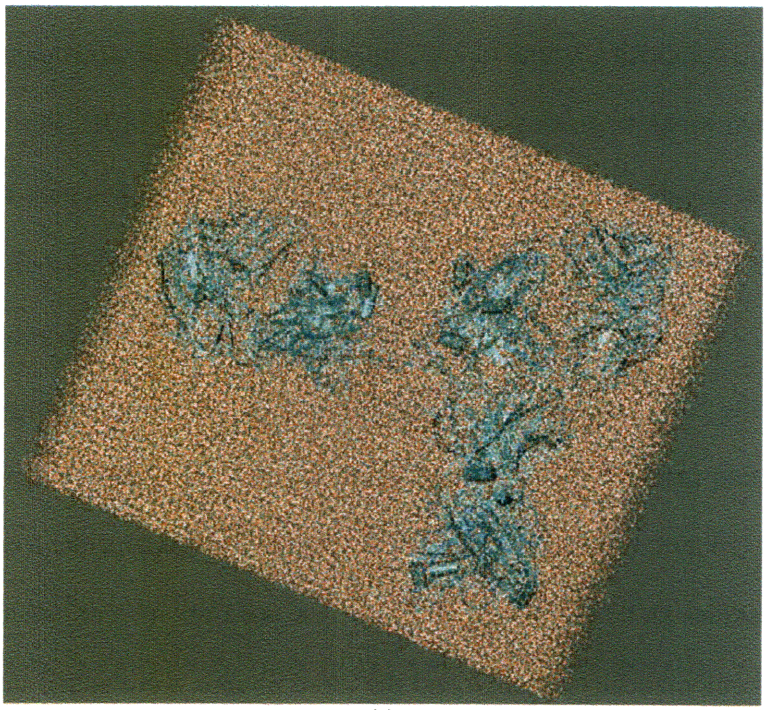
Reaction rates were computed according to Bennett-Chandler procedure (*22-25*). Langevin dynamics was used to compute the corrections to the transition state theory of reaction rate constant.

6.3 Methodology

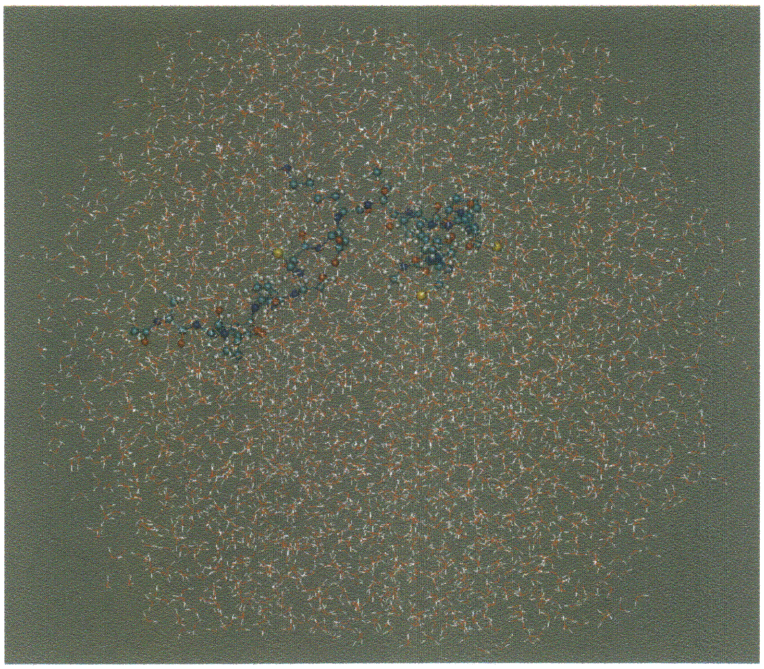
6.3.1 System description.

The details of the MD simulation of the whole antibody can be found in Chennamsetty et al. It consists of a full glycosylated antibody molecule solvated in explicit solvent, comprising a total of 202130 atoms contained in a rectangular simulation box. It has two hinge fragments, denoted hinge fragment I and II. For the free hinge segment, initial structure was taken from the crystal structure of human IgG1 b12 (*12*) (PDB code 1HZH) with residue number from Ala225 to Pro240 (the complete sequence is: Ala225-Glu-Pro-Lys-Ser-Cys-Asp-Lys-Thr-His-Thr-Cys-Pro-Pro-Cys-Pro240). According to the classification of the hinge region (*15*), here only the upper and middle hinge amino acid residues were included because hydrolysis was only observed for the peptide bonds in these residues. CHARMM (*26*) force field parameters were used for the amino acid residues and TIP3P water potential was used for solvent molecules. Charge neutrality was automatically satisfied and therefore no ions were added into the system. A solvation shell of at least 8 Å in determining a truncated octahedron was chosen to minimize the interaction between the hinge fragment and its periodic images. Periodic boundary condition was applied and Ewald summation was used in treating the long-

range electrostatic interactions (*22, 27*). The final system has 3,481 water molecules and an N-acetylated and C-methylated 16-residue hinge fragment, with a total 10,633 atoms. Covalent bonds involving hydrogen atoms were constrained using the SHAKE algorithm, allowing the use of a larger integration time step of 2 fs. The cubic box containing the truncated octahedron has a side length of 53.10 Å after performing constant pressure and temperature dynamics to equilibrate the system for a nanosecond. The lengths of the MD trajectories for the whole antibody and for the free hinge fragment are 50 ns and 3 ns, respectively. The two trajectories were saved every 100 ps and 20 ps, respectively, for data analysis. Both simulations were performed using the CHARMM software package (*26*).



(a)

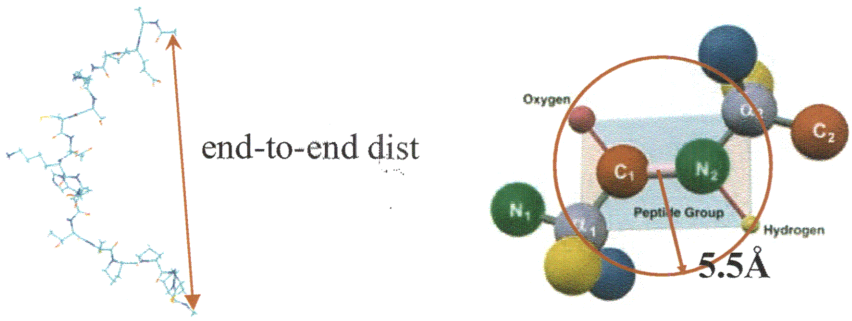


(b)

Figure 6.2: Simulation box for the whole antibody system (a) and for the hinge fragment (b). Both systems contain explicit solvent water molecules, shown in a line representation.

The RMSD from the original X-ray crystal structure was calculated for each saved frame in both trajectories. The RMSF was calculated as the root-mean-square of the standard deviations in the coordinates for each residue. For the whole MD antibody trajectory, this was done after the Fab and Fc domains were aligned to their respective crystal structures using a best-fit algorithm. These domains were aligned separately because there was larger relative motion between these domains than within each domain and that RMSF should reflect the local flexibility of instead of the overall translational and rotational displacements. The dynamic solvent accessible area for each residue was calculated according to the standard technique (*28*), with a radius of the solvent probe sphere 1.4 Å. The end-to-end distance of the hinge fragment and the water coordination number for each peptide bond in the hinge fragment was defined as the distance between

the nitrogen atom on Ala225 and the alpha carbon on Pro240, and as the number of water molecules within a cutoff of 5.5 Å from the midpoint of each C-N bond, respectively, as shown in Figure 6.3. A cutoff of 5.5 Å was chosen in order to include a two solvation shell.



(a) (b)

Figure 6.3: Schematics showing for the definitions of end-to-end distance (a) and water coordination number (b).

6.3.2 Determination of reaction coordinate.

The reaction coordinate for the hydrolysis reaction of a model compound N-MAA was determined in our previous study, and a reaction coordinate model involving 5 geometric order parameters was used here directly. As discussed in details in our previous work, the reaction coordinate was obtained from the likelihood maximization algorithm after screening a huge number of reaction coordinate models on an ensemble of trajectories sampled by the "aimless shooting" algorithm (*29*), which is a variant version of transition path sampling (*30, 31*). It is a good one-dimensional approximation to the "true" reaction coordinate, the committor probability p_B (*29, 32*), and its goodness of choice was verified by the standard committor probability p_B histogram analysis (*31, 33*).

The justification for the use of this reaction coordinate is that hydrolysis reaction is localized around the peptide bond. In our coarse-grained model presented here for the reaction rate constant calculation, the dynamics along the reaction coordinate direction were assumed to be the same as the one in N-MAA system, and all other factors which may affect the reaction dynamics in the antibody molecule, including some of the solvent effect and the effect of Fab and Fc domains connecting to the hinge fragment, were modeled as external influences, specifically represented by the forces exerted by the part of the system outside of the reaction center along the reaction coordinate direction. This is the way to link with the mechano-chemical mechanism.

6.3.3 Calculation of force along the reaction coordinate direction.

For each coordinate frame in the two MD trajectories, atoms corresponding to the ones that comprise of the one-dimensional reaction coordinate determined in our previous work need to be identified first. A procedure of best-fit alignment was used in finding the correspondence. A configuration from the transition state ensemble in the QM system was chosen. The whole antibody system or the hinge fragment system was then rotated and translated in order to best-align six atoms (the alpha carbon in the N-terminal, N, HN, C, O, and the alpha carbon in the C-terminal) close to the peptide bond. The choice of these six atoms was motivated by their ease of identification since in both QM and MD systems they show up as nearly constant relative configurations. The choices with less or more atoms than these six atoms for alignment were verified to make little difference, but would require more care of identifying their correspondence. In this newly generated set of coordinates for the whole antibody system or the hinge fragment system, the atoms closest to the QM atoms were identified as in correspondence. In this

way, all atoms (denoting this collection of atoms as N) involved in the reaction coordinate expression determined by likelihood maximization could be identified. The peptide bond with the largest extent (*6*, *7*) of hydrolysis in the antibody molecule, Asp222-Lys223, and the five-OP variable best reaction coordinate model were chosen as our focus in this study. Thus the collection N=10 atoms. Forces that represent the interactions between atoms in the collection N and all the rest in the system were calculated using the CHARMM (*26*) command INTER. This force vector $\vec{F}^{3N}$ with $3N$=3×10=30 components for each coordinate frame in the MD trajectories was then used to calculate its component F_{RC} along the direction of the five-OP variable best reaction coordinate model q according to the following expression:

$$F_{RC} = \vec{F}^{3N} \cdot \frac{\vec{\nabla}^{3N} q}{\left|\vec{\nabla}^{3N} q\right|} \tag{6.1}$$

6.3.4 Computation of reaction rate.

An estimate of the free energy barrier for the hydrolytic reaction was performed based on the experimental data (*6*), and the free energy barrier was chosen to be 32 kcal/mol. A free energy profile as a function of the reaction coordinate was chosen based on two points: 1) the reactant state is relatively less stable than the product state, and 2) the free energy barrier is 32 kcal/mol. Its expression was determined to be:

$$U(x) = \frac{1}{x^6} + 10[1 - e^{-(x-1)}]^2 + 30e^{-(\frac{x-3}{2})^2} - 10 \tag{6.2}$$

The shape of the free energy profile and its derivative, i.e. the force, are shown in Figure 6.4.

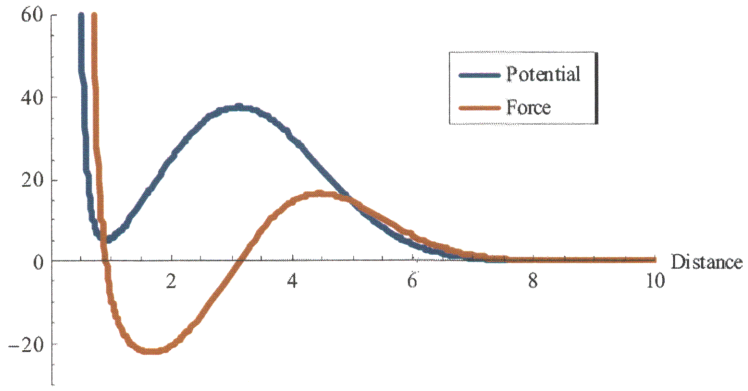


Figure 6.4: The potential energy from an estimate of the free energy barrier for the hydrolytic reaction and the force for the calculation of the reaction rate constant in the coarse grained model.

After analysis of the saddle point and the two minima, the reactant and product states were determined to be x<3.139 and x>3.139 respectively. The rate constant for the transition from the reactant state to the product state given the interaction shown in Equation (6.2) was calculated using the Bennett-Chandler procedure (*22-25*). The forward reaction rate constant is expressed as:

$$k(t) = \frac{\langle \dot{q}(0)\delta(q(0)-q^*)\theta(q(t)-q^*)\rangle}{\langle \theta(q^*-q(0))\rangle} = \frac{\langle \dot{q}(0)\delta(q(0)-q^*)\theta(q(t)-q^*)\rangle}{\langle \delta(q^*-q(0))\rangle} \frac{\langle \delta(q^*-q(0))\rangle}{\langle \theta(q^*-q(0))\rangle} \tag{6.3}$$

Here, $k(t)$ is the reactive flux, $q(t)$ is the reaction coordinate at time t in a dynamic trajectory, $\dot{q}(t)$ is the velocity projected onto the reaction coordinate direction at time t, $\langle\rangle$ denotes the canonical ensemble average, q^* is the reaction coordinate at the transition state, and $\delta(x)$ and $\theta(x)$ are the standard Dirac delta function and the Heaviside step function, respectively. The forward reaction rate constant k_f is related to the reactive flux

by $k(t)=k_f$ for time t such that $\tau_{trans} << t << \tau_{stable}$, where τ_{trans} and τ_{stable} are the timescales for the system to spend during the transition process and in stable states, respectively.

It can be factored into two terms:

$$k(t) = \frac{\langle \dot{q}(0)\delta(q(0)-q^*)\theta(q(t)-q^*)\rangle}{\langle \delta(q^*-q(0))\rangle} \frac{\langle \delta(q^*-q(0))\rangle}{\langle \theta(q^*-q(0))\rangle} \tag{6.4}$$

In the first step, a constrained MD was performed to calculate $A(t)$:

$$A(t) = \frac{\langle \dot{q}(0)\delta(q(0)-q^*)\theta(q(t)-q^*)\rangle}{\langle \delta(q^*-q(0))\rangle} \tag{6.5}$$

The second step is to calculate $P(q^*)$:

$$P(q^*) = \frac{\langle \delta(q^*-q(0))\rangle}{\langle \theta(q^*-q(0))\rangle} \tag{6.6}$$

This calculation was more conveniently done using numerical integration

$$\ln \frac{P(q^*)}{P(q_A)} = \int_{q_A}^{q^*} dq' \frac{\left\langle |H|^{-\frac{1}{2}} \partial(\ln|J| - \beta U)/\partial p' \right\rangle_c}{\left\langle |H|^{-\frac{1}{2}} \right\rangle_c} \tag{6.7}$$

Here $\langle\rangle_c$ denotes the canonical average in the constrained ensemble where $q(t)=q'$, $H = \sum_{i=1}^{N} \frac{1}{m_i} \frac{\partial q}{\partial \vec{r}_i} \cdot \frac{\partial q}{\partial \vec{r}_i}$, $|J|$ is the Jacobian of the transformation from the Cartesian coordinates $\vec{r}_i$ to the generalized ones containing q, and $\beta = 1/(k_B T)$.

An over-damped Langevin dynamics

$$-\frac{\partial U}{\partial r} + \zeta \frac{dr}{dt} + F_{RC} = 0 \tag{6.8}$$

was used to propagate the time evolution of the dynamic trajectory. Here t is time, r is the reaction coordinate, U is the potential energy model, ζ is the frictional coefficient, and F_{RC} is the force projected along the reaction coordinate direction.

The only parameter necessary to calibrate in Equation (6.8) is the frictional coefficient. A calculation of forces exerted by the solvent molecules on the hinge fragment was carried out for each coordinate frame in the MD trajectories. The probability distribution of the forces obtained was verified to be Guassian, and was calibrated with a Gaussian distribution, whose variance was calculated. From the fluctuation-dissipation theorem, the frictional coefficient can be identified as (*34*)

$$\zeta = \frac{\mathrm{var}(F_{\mathrm{solvent}}(t))}{2k_B T} \tag{6.9}$$

Using the MD trajectories, ζ =0.055 (kcal·sec)/(mol·Å) and 0.059 (kcal·sec)/(mol·Å) were obtained for the free hinge fragment and the whole antibody system, respectively.

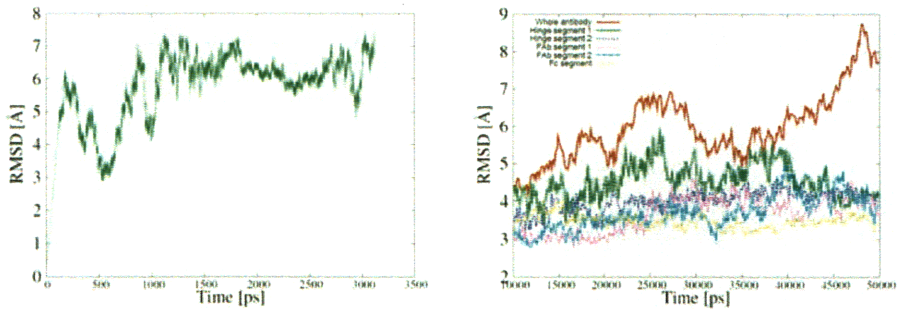
6.4 Results and discussion

6.4.1 Structural and dynamic differences for antibody hinge vs. free hinge.

As expected, the structure and dynamics of a free hinge fragment and those of the two hinge fragments in the antibody molecule can be quite different due to the special arrangement of the antibody structure. These differences were characterized in terms of the RMSD, RMSF, solvent accessible area, water coordination number and end-to-end distance.

As shown in Figure 6.5, the RMSD of the free hinge fragment rises quickly to 6 Å after 200 ps in the dynamics run, and then fluctuates in the 6-7 Å window for most of the

3 ns MD trajectory. However, in the whole antibody system, the globular and compact Fab and Fc domains have a low RMSD less than 4 Å while the two hinge fragments exhibit relatively higher but still different RMSD values. The hinge fragment I is the one which has the same initial structure as the free hinge fragment before the MD run, and it has a higher RMSD than the hinge fragment II but lower values than the free hinge fragment. These results confirm wide recognition of the flexibility of the hinge region in antibody molecules, but also show the constraining effect of Fab and Fc domains on the dynamics of hinge fragments in the antibody molecule.



(a) (b)

Figure 6.5: RMSD calculations for hinge fragment system (a) and the whole antibody system (b).

In Figure 6.6, the RMSFs of all atoms, only backbone atoms, and only side-chain atoms for each residue in the hinge fragment were calculated separately in order to determine how much the backbone of the fragment fluctuates, even though all of these three values show the same trend. In the free hinge fragment, because both ends are free in solution, their RMSFs showed much higher values then the residues in the middle. The two hinge fragments in antibody molecule exhibit quite different RMSF behavior. Hinge fragment I shows much higher RMSF than the hinge fragment II, while it is only partially comparable with the free hinge fragment. In the N-terminal side and the middle of the upper hinge region, the hinge fragment I and the free hinge have more similarity

than for the residues on the C-terminal side. On the C-terminal side, fluctuation of the hinge fragment I in the antibody molecule is severely limited, similar to all the residues in the hinge fragment II. These results again show the constraining effect of Fab and Fc domains on the fluctuation of hinge fragments in the antibody molecule.

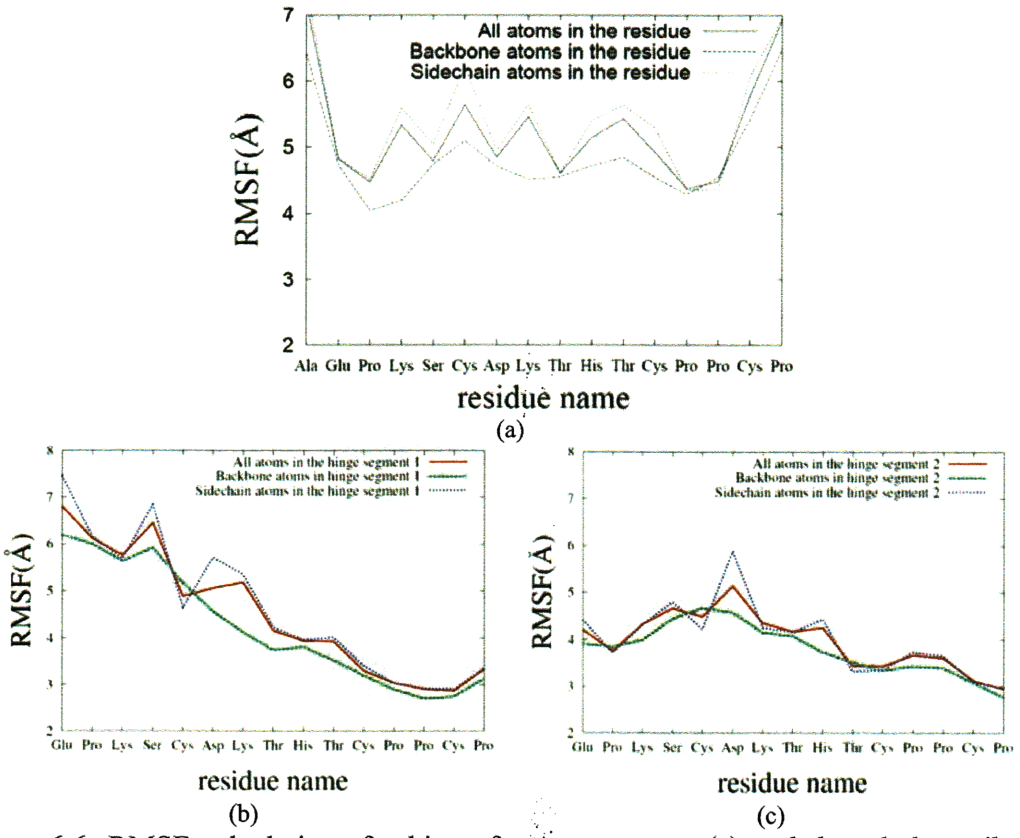
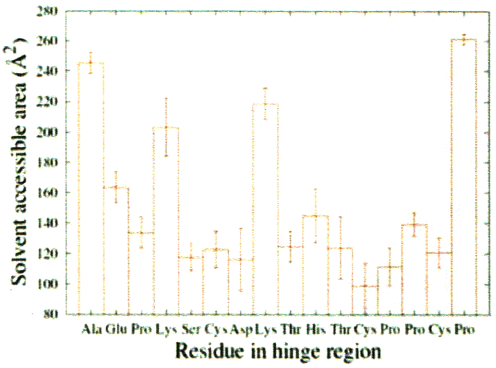


Figure 6.6: RMSF calculations for hinge fragment system (a) and the whole antibody system (b) and (c).

In assessing the solvent exposure of each residue in the hinge region, two descriptors were calculated. The solvent accessable area, averaged over the MD trajectories, is one way to quantify the exposure of each residue to the solvent molecules relative to the protein interior. As shown in Figure 6.7, these values in general correlate with their hydrophobicity. Residues at both ends of the free hinge fragment have larger values, as expected. However, when comparing to the values of free hinge fragment,

much smaller values were observed for the hinge fragment I. This finding is because, to be discussed shortly, that the free hinge assumed much more configurations which bestow higher solvent accessable area to the residues than the hinge fragment in the antibody molecule.

Another descriptor used for characterizing the solvent exposure is the water coordination number, which is the average number of water molecules within the vicinity of each peptide bond, as depicted in Figure 6.3. To define the water coordination number in this way was motivated by the fact that the hydrolytic reaction of a peptide bond needs nearby water molecules physically present. Thus the coordination number is expected to be a more accurate descriptor for the hydrolytic reaction if there were such a correlation. In Figure 6.8, the values of the water coordination number for each peptide bond in the hinge fragments are shown. The water coordination number showed less variability in the free hinge fragment than the hinge fragments in the antibody molecule, probably also because more exposed configurations can be assumed when the hinge fragment is free in solution, as discussed in the end-to-end distance results. The more variable results for the hinge fragments I and II do not show particular correlation with the extent of hydrolytic reaction for hinge residues (*6, 7*), suggesting that a more complicated mechanism could be involved, resulting in the enhanced reaction rate.



(a)

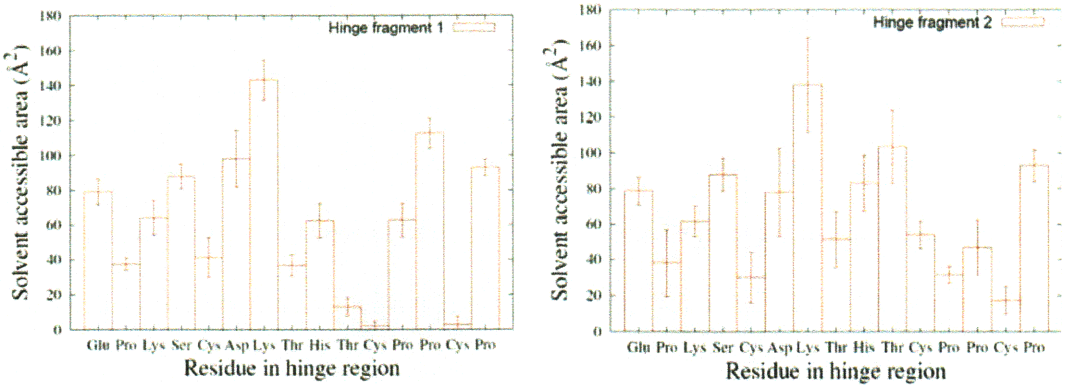


Figure 6.7: Solvent accessible area calculations for the hinge fragment system (a) and the whole antibody system (b) and (c).

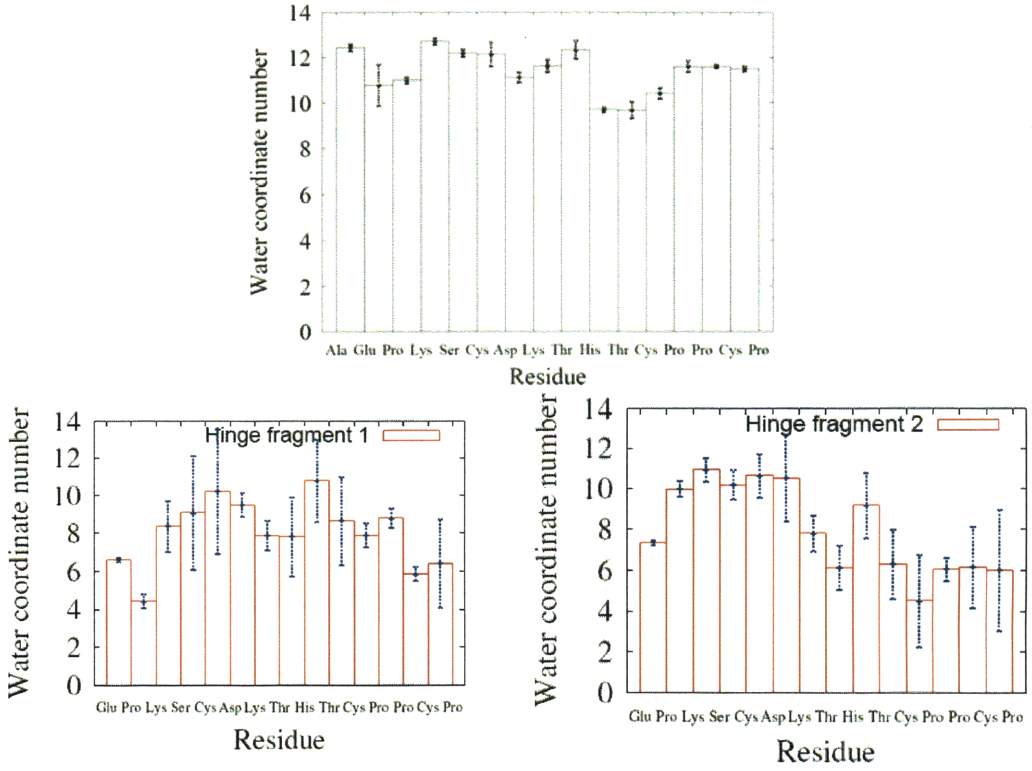


Figure 6.8: Water coordination number calculations for the hinge fragment system (a) and the whole antibody system (b) and (c).

In order to analyze in more detail what differences there were between the free hinge fragment and the fragments in the antibody molecule, the end-to-end distance was calculated and the results are shown in Figure 6.9. The free hinge fragment has a rapidly changing end-to-end distances over a time period of only 3 ns, while the end-to-end distances of the hinge fragment I and II in the antibody molecule stay relatively constant over a period of 40 ns. Figure 6.9 (a) shows several particular configurations of the hinge corresponding to extremely large or small values of end-to-end distances. These configurations either are more bent or more extended, which was never observed in the much longer whole antibody simulation. This behavior of the free hinge fragment taking on much less accessible configurations of the whole antibody hinge fragments may explain the differences in all structural and dynamic quantities discussed so far, and may suggest the important constraining effects of the globular Fc and Fab domains on the hinge region mentioned earlier.

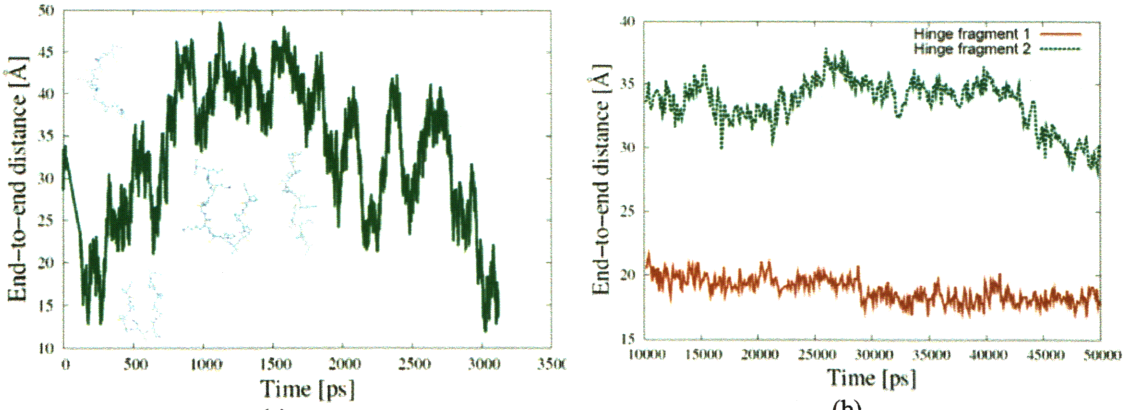
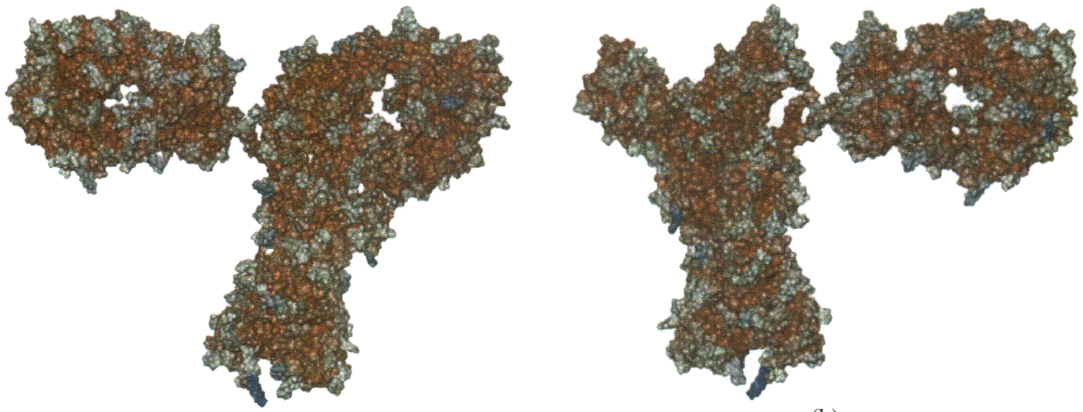


Figure 6.9: End-to-end distance calculations for the hinge fragment system (a) and the whole antibody system (b).

In order to compare the differences in the solvent accessibility and the local fluctuations between the hinge fragments and other parts of the antibody molecule, Figure 6.10 shows a color coded solvent accessible area averaged over the 40 ns MD trajectory. The values of the solvent accessible area of the residues are inversely proportional to their hydrophobicity, but they also show correlation with their particular positions relative to the protein surface. There are several patches on the surface which show very large solvent accessible areas; however, there are more patches even on the surface that show much smaller solvent accessible areas, indicating their hydrophobic nature.

In Figure 6.11, the color coded RMSF is shown. The C-terminal of one of the heavy chains has the largest RMSF, due to its free end nature. There is also one surface patch close to the hinge fragment in one of the heavy chains that exhibits high RMSFs. However, even though the two hinge fragments do show somewhat higher values of RMSF, they do not have the extreme values.

In terms of solvent accessibility and local fluctuation, no conclusive evidence shown in Figure 6.10 and Figure 6.11 confirms the hypothesis that largest solvent accessibility or the largest flexibility results in the higher rate of hydrolysis.



(a) (b)

Figure 6.10: The solvent accessible area calculated for each residue in the whole antibody, the front view (a) and the back view (b). The redder the color code, the smaller the solvent accessible area; the bluer the color code, the larger the solvent accessible area.

(a) (b)

Figure 6.11: RMSF calculated for each residue in the whole antibody, a front view (a) and the back view (b). The redder the color code, the smaller the RMSF; the more blue the color code, the larger the RMSF.

6.4.2 Force projection onto the direction of the reaction coordinate.

The reaction coordinate determined by the likelihood maximization technique in our previous work involves five order parameters, which are two distances, one angle, and two dihedral angles. These order parameters describe not only the changes of the local bonding pattern, but also the changes in the solvent network in influencing the hydrolytic reaction. The force vector $\vec{F}^{3N}$ for the N atoms involved in the reaction coordinate expression q was first calculated, followed by the calculation of projection of $\vec{F}^{3N}$ onto the q direction F_{RC} according to Equation (6.1).

In Figure 6.12 and Figure 6.13, the projected forces as a function of time are plotted. The range for these forces is about 80 kcal/mol/Å, which is equivalent to 5.56 nN on an individual bond. The projected forces can be negative (pushing in the direction

of the reaction coordinate) or positive (pulling in the direction of the reaction coordinate), with some simple statistics shown in Table 6.1. In Table 6.1, Figure 6.11, and Figure 6.12, the projected forces in the free hinge are smaller in absolute values than those in the hinge fragment I.

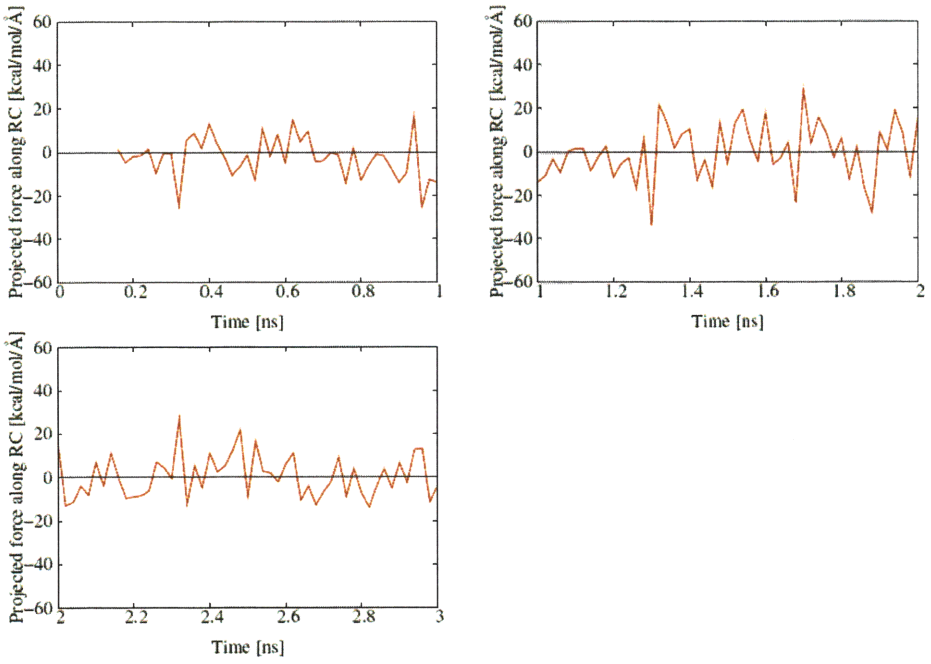


Figure 6.12: Force along the five-OP variable reaction coordinate for the free hinge fragment during the 3 ns MD trajectory.

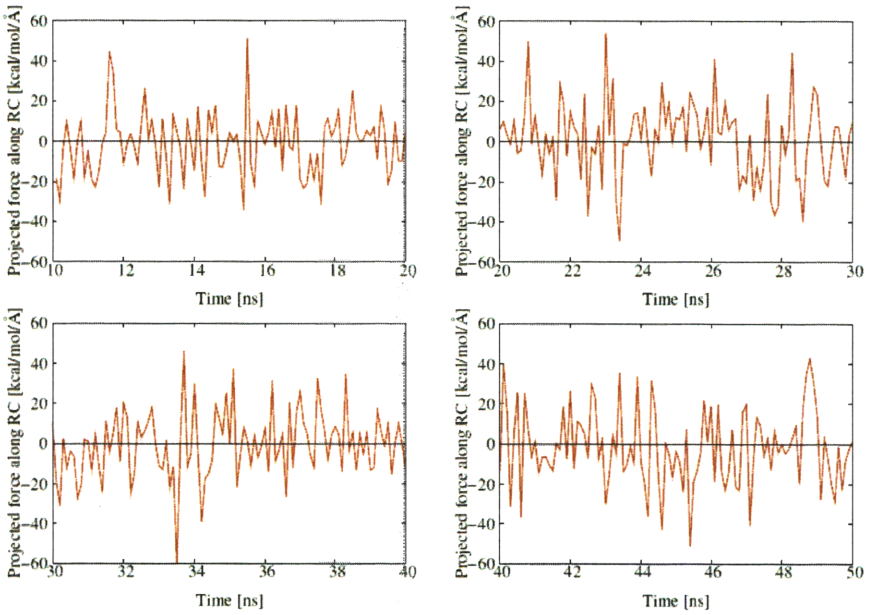


Figure 6.13: Force along the five-OP variable reaction coordinate for the hinge fragment I of antibody molecule during the 50 ns MD trajectory.

Table 6.1: Statistics of the projected forces along the direction of the reaction coordinate in the free hinge fragment and in the hinge fragment I in the antibody molecule. All numerical numbers are in units of kcal/mol/Å.

	Free hinge fragment	Hinge fragment I
Min	-35.42	-60.84
1st quantile	-8.45	-12.34
Median	-1.83	-0.23
Mean	-0.76	-0.50
3rd quantile	6.57	10.07
Max	28.26	54.67

Even though the mean values of these forces are negative for both the free hinge fragment and the hinge fragment I, due to the much shorter time scale of the reaction dynamics than the time scale of calculating these force projections, it is likely that the accumulated pulling force within some time period is enough to affect the reaction dynamics.

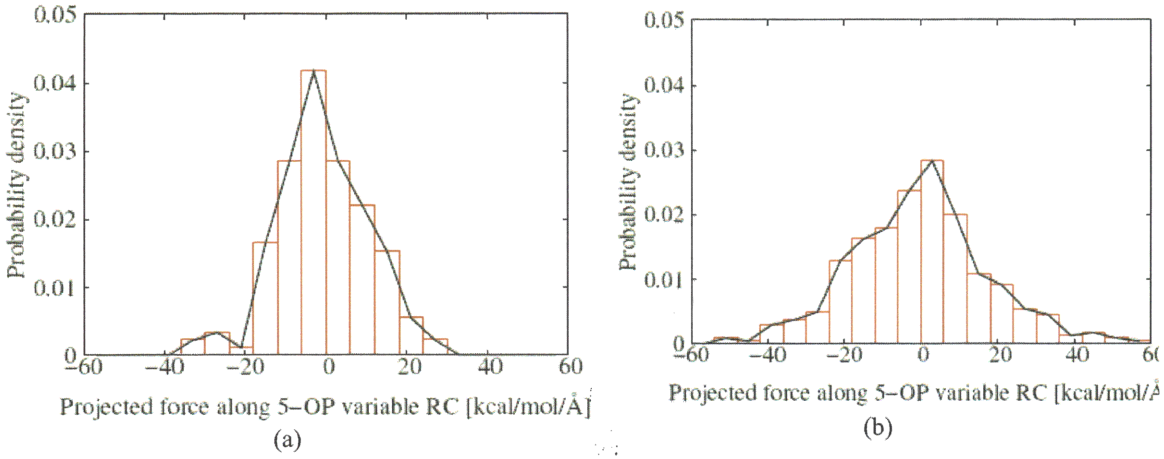


Figure 6.14: Probability distribution (normalized) of the calculated forces in the direction of the reaction coordinate for the free hinge fragment (a) and the hinge fragment I in the antibody molecule (b).

Figure 6.14 shows the histograms of the projected forces along the reaction coordinate for the free hinge fragment and the hinge fragment I in the antibody. The p-values of the Anderson–Darling test (*35*) on normality for the distribution of the projected forces are 0.32 and 0.08 for the free hinge fragment and the hinge fragment I, respectively. These p-values suggest that the hinge fragment I has more likely non-Gaussian distributed projected forces along the direction of the reaction coordinate.

6.4.3 Reaction rate calculation.

As outlined above, the reaction rate constant can be calculated following one constrained MD simulation for Equation (6.5) and a series of constrained MD simulations for Equation (6.7). The results for these calculations are shown in Figure 6.15.

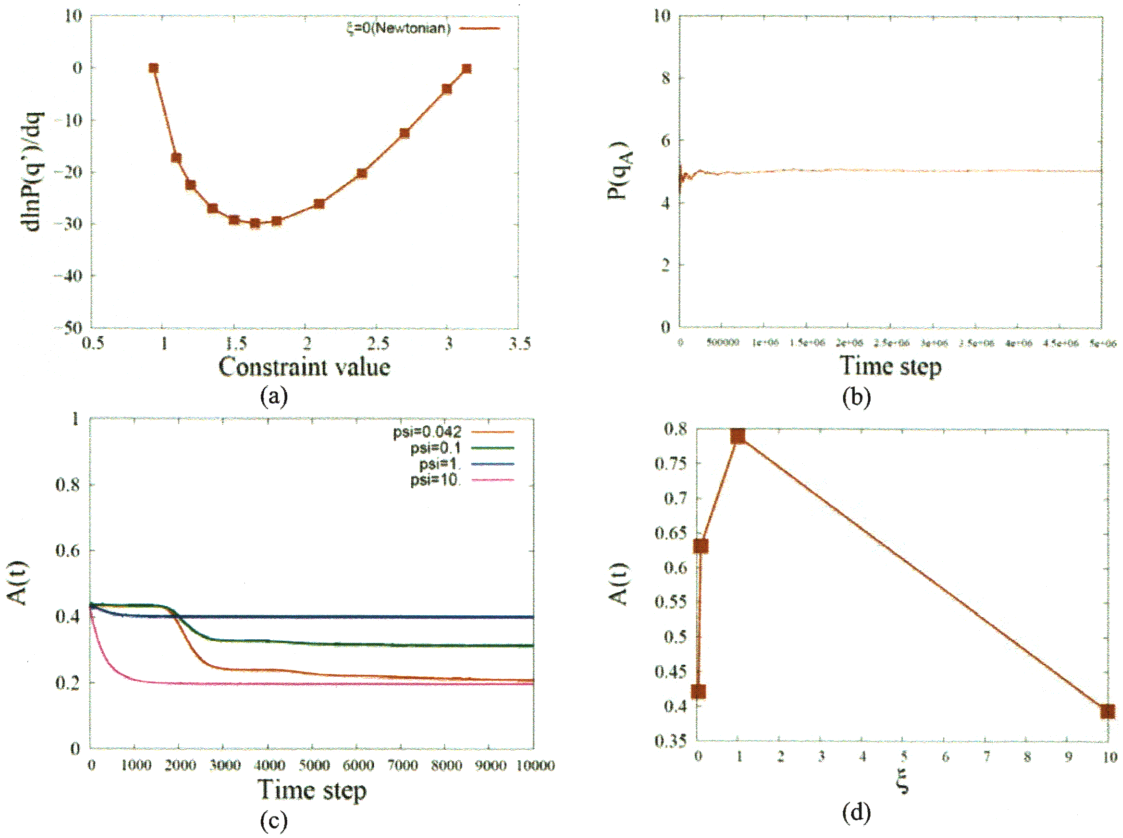


Figure 6.15: Calculation of the reaction rate constant for the simplified model. (a) Constrained MD simulations to calculate the numerical integration as shown in Equation (6.7). (b) The probability defined in Equation (6.6) when the system is in state A. (c) The factor A(t) defined in Equation (6.5) as a function of time for different values of frictional coefficients. (d) The plateau values of A(t) shown in (c) are plotted against the frictional coefficients.

In Figure 6.15 (a), 12 constrained Langevin dynamics run at different reaction coordinate values were performed in order to numerically integrate the right hand side of Equation (6.7). In (b), an normal Langevin dynamic run was performed to obtain $P(q_A)$, which the system can access easily since it is the potential energy well. In (c), the term $A(t)$ in Equation (6.5) was calculated in a constrained ensemble when the reaction coordinate was fixed at the transition state value for different values of the frictional coefficients in the Langevin equation Figure 6.8. This term $A(t)$ contains the dynamic correction to the transition state reaction rate (TST). Then in (d), the overall forward

reaction rate at different values of frictional coefficients is shown. Interestingly, the rates exhibit a high frictional coefficient region, a low frictional coefficient region, and a turn-over region that has been much discussed in the literature, for example in the review by Hanggi et al. (*36*). From this calculation, one can also observe that the frictional coefficient did not show any significant effect on the reaction rate constant. Thus the choice of ζ=0.059 (kcal·sec)/(mol·Å) from fitting the projected forces against a normal distribution exerted by solvent molecules was used in subsequence calculations.

Having a constant pulling force along the direction of the reaction coordinate has a two-fold effect. On one hand, the transmission coefficient is lowered, because from Equation (6.5), those trajectories with large negative $\dot{q}(0)$ would still commit to the product basin, i.e., $\dot{q}(T) > q^*$ with the total length of the trajectory denoted as T. The portion of contribution by these trajectories to the $A(t)$ in Equation (6.5), which is directly proportional to the transmission coefficient, will cancel some of the original positive contributions to $A(t)$. However, on the other hand, as shown in Figure 6.16, the constant pulling force will also decrease the energy barrier for the reaction, more than the decrease it causes in the transmission coefficient. Therefore, the net effect of having a constant pulling force is an increase in the reaction rate constant, as simply accounted for in Table 6.2. In Table 6.2, the calculated reaction rate constants were tabulated against the assumed constant pulling forces.

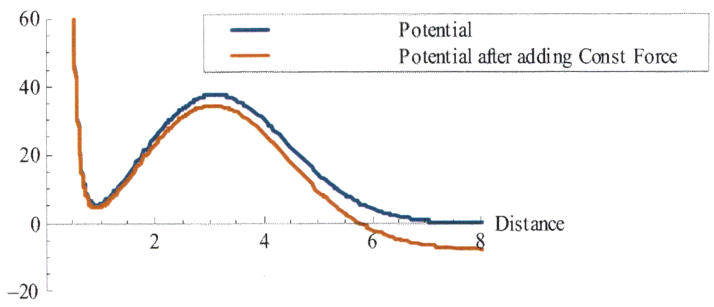


Figure 6.16: The effect of a constant pulling force on the potential energy profile.

Table 6.2: When applying various constant pulling forces, the reaction rate for the simplified reaction rate model at 37°C.

Constant Force along reaction coordinate direction [kcal/mol/Å]	Reaction rate constant [$\times 10^{-10}$ sec^{-1}]	Times of enhancement comparing with no net pulling force
0	0.830	1
0.2	1.730	2.08
0.4	3.595	4.33
0.6	7.451	8.98
0.8	15.372	18.5
1.0	32.131	38.7
1.2	70.323	84.7
1.4	140.23	169

On one hand, even though the mean values of force distributions as calculated previously are zero for both the free hinge fragment and the hinge fragment I in the antibody molecule, due to the short timescale for the reaction dynamics as discussed in our previous *ab initio* MD work, it is still likely that the net pulling forces which are in effect over a longer period of time may influence the reaction dynamics. A simplistic model to understand the timescale difference is to obtain a time-averaged pulling force. In this procedure, the negative forces along the direction of the reaction coordinate are discarded, since they do not contribute to the reaction dynamics over a longer timescale than the timescale of the reaction dynamics itself. Only positive forces along the

direction of the reaction coordinate are averaged. For this simplistic model, averaged forces 2.02 and 3.32 kcal/mol/Å were obtained for the free hinge fragment and the hinge fragment I in the antibody molecule, respectively.

On the other hand, using the experimental kinetics data obtained by Cordoba et al. (*6*) and by Smith et al. (*10*), one can obtain the ratio of the rate constants for the hydrolytic reaction in the hinge region of the antibody molecule and in the simple di-peptide. At the same temperature, 37 °C, this ratio is within the range of 100-150.

From both aspects of the analysis, one can therefore see from Table 6.2 that an extra 1.3 kcal/mol/Å of time-averaged forces between the free hinge fragment and the hinge fragment I in the antibody molecule would result in similar amount of rate enhancement. Thus this mechano-chemical model provides a reasonable explanation for the 100-150 times faster rate of hydrolysis observed in the hinge region of the antibody molecule.

6.5 Summary and conclusions

In this study, two approaches were undertaken in understanding the cause of the enhanced rate of hydrolysis of peptide bonds in the antibody molecule. Both approaches were driven by our two hypotheses, i.e., 1) there may be a correlation between solvent exposure and the rate of hydrolysis, and 2) there may be a mechano-chemical mechanism involved in the hydrolytic reaction of peptide bonds in the hinge fragment of the antibody molecule. Classical MD simulations for free hinge fragment and the whole antibody molecule were performed, and the trajectories thereby obtained were analyzed. Structural and dynamic differences between the free hinge fragment and the hinge

fragments in the antibody molecule were revealed, especially that the free hinge fragment takes on configurations less likely to be accessible to the hinge fragment when situated inside the antibody molecule. These specific differences include the following:

- Both the RMSD and RMSF of the free hinge fragment are mostly larger than the corresponding hinge fragment in the antibody, suggesting less fluctuation and movement when the hinge is connecting the Fab and Fc domains.
- The solvent accessible area and water coordination numbers revealed different aspects of the solvent exposure for residues and the peptide bonds in the hinge region respectively. There seems to be no direct correlation with the rate of hydrolysis. This suggests that other factors may have significant influence on the dynamics of the hydrolytic reaction.
- The end-to-end distance calculation revealed why the differences mentioned above occur; the free hinge fragment assumed more both bent and extended configurations than the hinge fragment situated inside antibody molecule, revealing a constraining effect from the Fab and Fc domains.

Due to the complexity of the antibody structure and the prohibitive computational cost of *ab initio* molecular simulations required for studying chemical reactions, a coarse grained approach was taken in testing our second hypothesis. This approach was also motivated by the findings in our test of the first hypothesis, mainly from the fact that hinge fragments in the antibody molecule are constrained and cannot adopt configurations as freely as the "free" hinge fragment. Forces along the direction of the reaction coordinate were projected using the classical MD trajectories and were used in a simplified reaction rate model in order to compute reaction rate constants. Our

calculations suggest that a mechano-chemical mechanism is possible to enhance the reaction rate of hydrolysis for peptide bonds in the hinge region of antibody molecules by lowering the energy barrier.

6.6 References

(1) Chu, J. W., Yin, J., Brooks, B. R., Wang, D. I. C., Ricci, M. S., Brems, D. N., and Trout, B. L. (2004) A comprehensive picture of non-site specific oxidation of methionine residues by peroxides in protein pharmaceuticals. *Journal of Pharmaceutical Sciences 93*, 3096-3102.

(2) Pan, B., Abel, J., Ricci, M. S., Brems, D. N., Wang, D. I. C., and Trout, B. L. (2006) Comparative oxidation studies of methionine residues reflect a structural effect on chemical kinetics in rhG-CSF. *Biochemistry 45*, 15430-15443.

(3) Wei, W. (1999) Instability, stabilization, and formulation of liquid protein pharmaceuticals. *International Journal of Pharmaceutics 185*, 129-188.

(4) Daugherty, A. L., and Mrsny, R. J. (2006) Formulation and delivery issues for monoclonal antibody therapeutics. *Advanced Drug Delivery Reviews 58*, 686-706.

(5) Cohen, S. L., Price, C., and Vlasak, J. (2007) beta-elimination and peptide bond hydrolysis: Two distinct mechanisms of human IgG1 hinge fragmentation upon storage. *Journal of the American Chemical Society 129*, 6976-+.

(6) Cordoba, A. J., Shyong, B. J., Breen, D., and Harris, R. J. (2005) Non-enzymatic hinge region fragmentation of antibodies in solution. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences 818*, 115-121.

(7) Dillon, T. M., Bondarenko, P. V., Rehder, D. S., Pipes, G. D., Kleemann, G. R., and Ricci, M. S. (2006) Optimization of a reversed-phase high-performance liquid chromatography/mass spectrometry method for characterizing recombinant antibody heterogeneity and stability. *Journal of Chromatography A 1120*, 112-120.

(8) Dillon, T. M., Ricci, M. S., Rehder, D. S., Flynn, G., Liu, Y. D., and Bondarenko, P. V. (2007) Discovery and characterization of conformational isoforms of human monoclonal IgG2 antibodies. *Nature*.

(9) Bryant, R. A. R., and Hansen, D. E. (1996) Direct measurement of the uncatalyzed rate of hydrolysis of a peptide bond. *Journal of the American Chemical Society 118*, 5498-5499.

(10) Smith, R. M., and Hansen, D. E. (1998) The pH-rate profile for the hydrolysis of a peptide bond. *Journal of the American Chemical Society 120*, 8910-8913.

(11) Wang, W., Singh, S., Zeng, D. L., King, K., and Nema, S. (2007) Antibody structure, instability, and formulation. *Journal of Pharmaceutical Sciences 96*, 1-26.

(12) Saphire, E. O., Parren, P. W. H. I., Pantophlet, R., Zwick, M. B., Morris, G. M., Rudd, P. M., Dwek, R. A., Stanfield, R. L., Burton, D. R., and Wilson, I. A. (2001) Crystal structure of a neutralizing human IgG against HIV-1: A template for vaccine design. *Science 293*, 1155-1159.

(13) Ito, W., and Arata, Y. (1985) Proton Nuclear Magnetic-Resonance Study on the Dynamics of the Conformation of the Hinge Segment of Human G1-Immunoglobulin. *Biochemistry 24*, 6467-6474.

(14) Kim, H., Matsunaga, C., Yoshino, A., Kato, K., and Arata, Y. (1994) Dynamical Structure of the Hinge Region of Immunoglobulin-G as Studied by C-13 Nuclear-Magnetic-Resonance Spectroscopy. *Journal of Molecular Biology 236*, 300-309.

(15) Roux, K. H., Strelets, L., Brekke, O. H., Sandlie, I., and Michaelsen, T. E. (1998) Comparisons of the ability of human IgG3 hinge mutants, IgM, IgE, and IgA2, to form small immune complexes: A role for flexibility and geometry. *Journal of Immunology 161*, 4083-4090.

(16) Roux, K. H., Strelets, L., and Michaelsen, T. E. (1997) Flexibility of human IgG subclasses. *Journal of Immunology 159*, 3372-3382.

(17) Schneider, W. P., Wensel, T. G., Stryer, L., and Oi, V. T. (1988) Genetically Engineered Immunoglobulins Reveal Structural Features Controlling Segmental Flexibility. *Proceedings of the National Academy of Sciences of the United States of America 85*, 2509-2513.

(18) Tan, L. K., Shopes, R. J., Oi, V. T., and Morrison, S. L. (1990) Influence of the Hinge Region on Complement Activation, C1Q Binding, and Segmental Flexibility in Chimeric Human-Immunoglobulins. *Proceedings of the National Academy of Sciences of the United States of America 87*, 162-166.

(19) Dall'Acqua, W. F., Cook, K. E., Damschroder, M. M., Woods, R. M., and Wu, H. (2006) Modulation of the effector functions of a human IgG1 through engineering of its hinge region. *Journal of Immunology 177*, 1129-1138.

(20) Beyer, M. K., and Clausen-Schaumann, H. (2005) Mechanochemistry: The mechanical activation of covalent bonds. *Chemical Reviews 105*, 2921-2948.

(21) Hickenboth, C. R., Moore, J. S., White, S. R., Sottos, N. R., Baudry, J., and Wilson, S. R. (2007) Biasing reaction pathways with mechanical force. *Nature 446*, 423-427.

(22) Frenkel, D., and Smit, B. (2002) *Understanding molecular simulation: from algorithms to applications*, 2nd ed., Academic, San Diego, Calif.; London.

(23) Nowick, A. S., and Burton, J. J. (1975) *Diffusion in solids: recent developments*, Academic Press, New York.

(24) Chandler, D. (1978) Statistical-Mechanics of Isomerization Dynamics in Liquids and Transition-State Approximation. *Journal of Chemical Physics 68*, 2959-2970.

(25) RuizMontero, M. J., Frenkel, D., and Brey, J. J. (1997) Efficient schemes to compute diffusive barrier crossing rates. *Molecular Physics 90*, 925-941.

(26) Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983) Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry 4*, 187-217.

(27) Allen, M. P., and Tildesley, D. J. (1989) *Computer simulation of liquids*, Oxford University Press, Oxford [England].

(28) Wodak, S. J., and Janin, J. (1980) Analytical Approximation to the Accessible Surface-Area of Proteins. *Proceedings of the National Academy of Sciences of the United States of America-Physical Sciences 77*, 1736-1740.

(29) Peters, B., and Trout, B. L. (2006) Obtaining reaction coordinates by likelihood maximization. *Journal of Chemical Physics 125*, -.

(30) Dellago, C., Bolhuis, P. G., and Geissler, P. L. (2002) Transition path sampling. *Advances in Chemical Physics, Vol 123 123*, 1-78.

(31) Bolhuis, P. G., Chandler, D., Dellago, C., and Geissler, P. L. (2002) Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual Review of Physical Chemistry 53*, 291-318.

(32) Hummer, G. (2004) From transition paths to transition states and rate coefficients. *Journal of Chemical Physics 120*, 516-523.

(33) Du, R., Pande, V. S., Grosberg, A. Y., Tanaka, T., and Shakhnovich, E. S. (1998) On the transition coordinate for protein folding. *Journal of Chemical Physics 108*, 334-350.

(34) Zwanzig, R. (2001) *Nonequilibrium statistical mechanics*, Oxford University Press, Oxford; New York.

(35) Anderson, T. W., and Darling, D. A. (1952) Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics 23*, 193-212.

(36) Hanggi, P., Talkner, P., and Borkovec, M. (1990) Reaction-Rate Theory - 50 Years after Kramers. *Reviews of Modern Physics 62*, 251-341.

Chapter 7. Overall Conclusions and Recommendations

7.1 Conclusions

The overall thesis goal was to better understand chemical instabilities of protein molecules from a mechanistic perspective. Two common types of chemical degradation pathways, the oxidation of methionine residues and the hydrolysis of peptide bonds in antibody molecules, were studied by a combination of experiments, theories, and molecular simulations.

In the first part of the methionine oxidation work, we conclude specifically that:

1. There is significant variation among the oxidation rate constants for different methionine residues in rhG-CSF and as a function of temperatures.
2. The rate constants for each methionine residue can be fit reasonably well according to the Arrhenius equation. This finding suggests that oxidative degradation of methionine residues is governed by the intrinsic oxidation reaction rather than a local conformational event having a complex temperature dependence.
3. Assuming the existence of an additional activation free energy barrier due to the transport of the oxidant molecule H_2O_2, a more complicated, non-Arrhenius equation ensues. However, this equation simplifies to the Arrhenius equation under certain circumstances.
4. Methionine residues in rhG-CSF can be classified into three categories according to the degree to which the protein structure affects oxidation kinetics, i.e., no structural dependence, local structural dependence and global

structural dependence. Additionally, three models were developed to consider the structural effect of protein molecules on the oxidation of methionine residues. We found that these models can fit the experimental data equally well. The "non oxidant-bound intermediate" model, in particular, can produce the anticipated temperature dependence of rate constants near the melting temperature and even beyond, when the influence of the protein structure becomes diminishingly small. However, the phenomenological models we considered cannot be distinguished based purely on phenomenological rate constant data. Trusted local dynamic information such as that which could be determined from molecular simulations would be needed.

5. An example was shown of the shelf-life prediction of protein pharmaceuticals using the temperature dependence of oxidation rate and the model prediction matched the stability data well.

In the second part of the work on the hydrolytic reaction of peptide bonds, we conclude specifically that:

1. Hydrolysis of N-MAA at neutral pH occurs in a concerted fashion; no stable or long-lived intermediate was found in our path-sampling simulations; the rate-determining step of the hydrolytic reaction of N-MAA under acidic pH occurs in a concerted fashion; a stable intermediate was found to be the final state in our path-sampling simulations.
2. Likelihood maximization procedure was extended to screen reaction coordinate models with more than one constituent order parameters, and within BIC, a reaction coordinate with five constituent geometric variables

was found to be the best in describing the path ensemble we generated for the hydrolytic reaction of N-MAA under neutral pH condition and a reaction coordinate with three constituent geometric variables was found for the rate limiting step of the hydrolysis reaction of N-MAA under acidic pH.

3. In both of the best reaction coordinate models, both geometric quantities which reflect bond making and breaking dynamics, and those which reflect the solvent network changes, are included, suggesting a complicated reaction involving many degrees of freedom for the hydrolysis reaction under neutral and acidic pH's.
4. Several p_B histograms were computed to verify the results of likelihood maximization, and the quantified goodness of these best-ranked reaction coordinate models is in accord with the their respective likelihood score.
5. Both RMSD and RMSF of the free hinge fragment is mostly larger than the corresponding hinge fragment in the antibody, suggesting less fluctuation and movement when the hinge is connecting the Fab and Fc domains.
6. Solvent accessible area and water coordination numbers revealed different aspects of the solvent exposure for residues and the peptide bonds in the hinge region respectively. There seems to be no direct correlation with the rate of hydrolysis. This suggests that there maybe other factors which have significant influence on the dynamics of the hydrolysis reaction.
7. End-to-end distance calculation revealed as to why the differences mentioned above occur; free hinge fragment assumed more both bent and extended

configurations than the hinge fragment situated inside antibody molecule, revealing a constraining effect from the Fab and Fc domains

7.2 Recommendations

7.2.1 Other types of oxidative instabilities

Oxidation of methionine residues in protein molecules can be caused by several other types of oxidants, and needs to be understood better in order to be controlled in a formulation, for example, photo-induced oxidation by singlet oxygen; oxidation by free radicals, such as superoxide radical $\cdot O_2^-$, peroxides ROOH, and hydroxyl radicals ·OH; metal-catalyzed oxidation; etc. Some of these oxidation pathways can be rather complicated, and experimental and computational means need to be combined in order to reveal the underlying mechanisms. The chemistry occurring in protein molecules is also complicated by the structure of protein molecules, as studied in detail in this thesis. The need to reveal the relationship between structure and chemical kinetics in protein chemistry is clear in order to obtain a full-scale picture about the reaction process.

7.2.2 Further test of the mechano-chemical mechanism of elevated hydrolysis of peptide bonds in antibody molecules

In this thesis, due to the challenges presented in understanding the hydrolytic reaction of peptide bonds in antibody molecules, such as the huge size of antibody molecules which is difficult to computationally handle, the facts that chemical reaction needs to be dealt with quantum mechanically and that a large collection of reaction pathways rather than a single trajectory are needed to characterize the transition process, a hierarchical approach was taken. However, with the increasing speed and capacity of

computer hardware and the fast development of more efficient parallel algorithms, a better approach would be to perform simulations using combined quantum-mechanical and molecular-mechanical (QM/MM) technique. In this approach, there is no need to separately consider the constraining effects on the hinge region by other parts of antibody molecules; the QM/MM technique gives a smooth and more elegant way to handle the problem of chemical reactivities in biomolecules like proteins. Also, the large spatial scale and longer timescale motion of the antibody molecule on the hydrolysis reaction in the hinge fragment need to be considered as well.

7.2.3 Rational and integrated design of formulation of protein therapeutics

Knowledge gained in this thesis can be used to guide formulation of protein therapeutics in a more rational and integrated way. More specifically,

1. A shelf-life prediction method can be developed similar to the approach proposed in this thesis. Care needs to be taken when the degradation kinetics can deviate from simple kinetics models, in particular because of the distinct structural influence in reactions of protein molecules than small molecules. Once the underlying degradation kinetics is determined, accelerated degradation studies can be performed, and various formulation designs can be screened faster and their shelf-life performance information can be obtained.
2. Oxidation of methionine residues by hydrogen peroxide is closely related to solvent accessibility and local structural fluctuations in the protein molecule. It would be interesting to test this finding in formulation designs by introducing excipients which may keep protein molecules more compact or less solvent exposed by protecting its surface.

3. In this thesis, it was found that the mechano-chemical mechanism is a reasonable explanation for the observed higher rate of hydrolysis in antibody molecules. If this hypothesis withholds in further tests, it may be an inherent instability for antibody molecules, due to the special arrangement in the structure of antibody molecules is the same. However, there are still possible ways to control the fluctuations and movement of different parts of antibody molecules, such as by adding excipients so as to adjust the formulation solution viscosity.