12-2014

Computer Simulations of Enzymes

Jianzhuang Yao
University of Tennessee - Knoxville, jyao6@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Jianzhuang Yao entitled "Computer Simulations of Enzymes." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Biochemistry and Cellular and Molecular Biology.

Hong Guo, Major Professor

We have read this dissertation and recommend its acceptance:

Jerome Baudry, Tongye Shen, Tamah Fridman, Elizabeth Howell, David Baker

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Computer Simulations of Enzymes

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Jianzhuang Yao

December 2014

DEDICATION

To my younger sister

Jianli Yao

To my parents

Shuzhen Wu and Guoqi Yao.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Hong Guo who has the attitude of a genius, for his continually and convincingly encouragement, patient guidance, and professional advice.

I would like to thank my committee members, Dr. Cynthia Peterson, Dr. Liz Howell, Dr. David Baker, Dr. Tongye Shen, Dr. Tamah Fridman, Dr. Chunlei Su and Dr. Jerome Baudry for their valuable suggestions on research and career development.

I want to appreciate all the members of Dr. Hong Guo's lab, Dr. Feng Chen's lab (our collaborator), Dr. Tongye Shen's lab and Dr. Xiaohan Yang's lab. Completion of this dissertation is impossible without their support.

Finally, I would like to thank my younger sister Jianli Yao, and my parents Shuzhen Wu and Guoqi Yao. They are always encouraging and supporting me with their best wishes.

ABSTRACT

Enzymes are important catalysts in living systems, and understanding catalytic mechanisms of enzymes is an important task for modern biophysics and biochemistry. Computer simulations have emerged as very useful tools for understanding how enzymes work. In this dissertation, QM/MM MD simulations were applied to study the catalytic mechanisms of several enzymes, including sedolisin, S-adenosyl-L-methionine (AdoMet)-dependent methyltransferases, and salicylic acid binding protein 2. For sedolisin, we focus on the acylation and deacylation reactions catalyzed by the enzymes. We proposed a general acid/base mechanism involving the Glu/Asp residues at the active site. MD and QM/MM free energy simulations on pro-kumamolisin show that the protonation of Asp164 would be able to trigger conformational changes and generate the functional active site for autocatalysis. The free energy simulations reported for SAMT, an AdoMet-dependent methyltransferase, showed that while the structure of the reactant complex containing salicylate, its natural substrate, is rather close to the corresponding TS structure, this is not the case for 4-hydroxybenzoate. The simulations demonstrated that additional energy is required to generate the TS-like structure for 4-hydroxybenzoate, consistent with the low activity of the enzyme toward this substrate. For protein lysine methyltransferase SET7/9, we showed that while the wild type SET7/9 may act like a mono-methylase, the Y245→A mutation could increase the ability of SET7/9 to add two more methyl groups on the target lysine. The substrate specificity of salicylic acid binding protein 2 (SABP2) has also been studied during my graduate study.

This enzyme has promiscuous esterase activity toward a series of substrates, but shows high activity toward its natural substrate methyl salicylate (MeSA). We demonstrated that SABP2 seems to represent a case in which the enzyme itself might have not been perfectly evolved and that substrate-assisted catalysis (SAC) involving its natural substrate may be used to enhance the activity and achieve substrate discrimination. In addition to enzymes, the prediction of protein-protein interactions (PPI) is also included in my dissertation. We established a robust pipeline for PPI prediction by integrating multiple classifiers using random forests algorithm. This pipeline could be very useful for predicting PPI.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

PKMTs:	protein lysine methyltransferases
AdoMet:	S-adenosyl-L-methionine
AdoHcy:	S-adenosyl-L-homocysteine
QM/MM:	quantum mechanical/molecular mechanical
MD:	molecular dynamics
PRMTs:	protein arginine methyltransferases
DFT:	density function theory
SCC-DFTB:	self-consistent charge density functional tight binding
PMF:	potential of mean force
WHAM:	weighted Histogram Analysis Method
SD:	steepest descent
ABNR:	adopted-basis Newton-Raphson
WT:	wild type
TS:	transition state
RS:	reactant state
SAC:	substrate assisted catalysis
NACs:	near attack conformers
AUC:	area under curve
BIPS:	BIANA interlog prediction server
GO:	gene ontology

BP: biological process

CC: cellular component

MF: molecular function

PPCM: protein-protein interaction classifiers merger

GO2PPI: predictor for protein-protein interactions (PPI) based on Gene Ontology (GO) annotations

PhyloProf: predictor for inferring protein-protein interactions using phylogenetic profiling (PP) methods

INTRODUCTION

Enzymes are important catalysts that are involved in most life processes.(Warshel 1991, Fersht 1999) Reaction rate enhancement of enzymes, ranging from 10^7 to 10^{20}, is significantly high.(Lad, Williams et al. 2003, Bruice 2006) Therefore, it is of fundamental and practical importance to understand the origin of the catalytic power and detailed mechanisms of enzymes. Although biochemical and structural studies (e.g. NMR and X-Ray Crystallography) shed important light on how enzymes work, the origin of enzyme's enormous catalytic power is not clearly understood, at least in many cases. For example, although Pauling proposed that catalytic power of enzymes is due to transition state stabilization,(Pauling 1946, Garcia-Viloca, Gao et al. 2004) how enzymes achieve transition state stabilization and what are the roles of the catalytic groups remain elusive for some enzymes. In the last two decades, computer simulations of enzymes have emerged as popular and effective tools to study detailed catalytic mechanism of enzymes, due in part to a dramatically increased computing power and the development of more efficient quantum mechanical/molecular mechanical (QM/MM) approaches. However, the applications of computer simulations is still lacking in elucidating the catalytic mechanism and catalytic power for many enzymes. Thus, understanding the detailed catalytic mechanism for these enzymes and determining how enzymes are able to achieve transition state stabilization are important tasks of computer simulations. In this dissertation, I seek to investigate catalytic mechanisms of several important enzymes and enzymes families including sedolisins, S-adenosyl-L-

methionine(AdoMet)-dependent methyltransferases, and salicylic acid binding protein 2 using quantum mechanical/molecular mechanical (QM/MM) calculations, MD (both classical and QM/MM) simulations and free energy (potential of mean force) simulations.

In addition to computer simulations of enzymes, I developed PPCM, a pipeline for the prediction of protein protein interactions.

Near Attack Conformers (NACs)

The idea of NACs has been extensively discussed by Bruice and coworkers.(Bruice and Lightstone 1999, Bruice and Benkovic 2000, Bruice 2002, Bruice 2006) NACs is believed to be normally formed in the ground state before the system reaches the transition state. The RS conformation can be descripted by the structures from MD simulations. In this Boltzmann distribution of conformers, NACs are the conformations resembling the transition state. As is demonstrated in Figure I-1, Bruice and his co-workiers defined the NACs as "having reacting atoms at van der Waals distance and angles ±15-20^{o} of the angle of the forming bond in TS."(Bruice 2006)

In order to determine the product specificity of PKMTs, we plotted the free energy profiles of the ground state conformers (including NACs). As discussed in previous studies of our lab(Xu 2009, Chu, Xu et al. 2010, Chu 2012, Yao, Chu et al. 2012) and in Chaptor 4, the SN2 reaction of methyl transfer from AdoMet to substrates (Lysine) should be more efficient if the methyl group of AdoMet is well aligned with the electron lone pair of N of Lysine sidechain in the reactant state with a relatively smaller θ angle and shorter C_M-N

distance. Therefore, we monitored the distributions of these angle and distance during MD simulation of reactant state (See Chapter 4 method for detail). Based on Boltzmann distribution, we are able to calculate the free energy contribution of NACs. The free energy profiles as the functions of r(C_M-N) and θ in the reactant state can be plotted from the probability densities ρ[r(C_M-N)] and ρ(θ) during MD simulations.

$$W(\xi) = - k_B T \ln \rho(\xi).$$

where k_B is the Boltzmann's constant, T is 283.15 K, and ξ is r(C_M-N) or θ.

We demonstrated that although the geometrical parameters defined by Bruce and his co-workers are very useful, they may not be the most sensitive parameters, at least in some of the cases, for understanding the efficiency of the methyl transfers and product specificity. For instance, Figure 4.3(A) shows that the distribution of NAC for the first methyl transfer in Y245A of SET7/9 may not be very significant. Nevertheless, the methyl transfer can still occur and the mutant is a trimethyltransferase.

General Acid/Base Catalysis and Substrate Assisted Catalysis

The existence of the oxyanion hole is an important property for many enzymes (such as proteinase and esterase). (Matthews, Alden et al. 1975, Blow 1976, Warshel, Naray-Szabo et al. 1989, Fuxreiter and Warshel 1998) The function of the oxyanion hole is to stabilize negative charge formation during the reactions, especially in the transition state and tetrahedral intermediate. Electrostatic effects such as hydrogen bonding interactions have been widely reported to be involved in the oxyanion hole stabilization. (Blow 1976, Warshel, Naray-Szabo et al. 1989, Fuxreiter and Warshel 1998) For instance,

the oxyanion hole of chymotrypsin is formed by two backbone amide groups (NH) of Gly193 and Ser195. The electrostatic effect provided by the hydrogen bonds from these two amide groups is used by the enzyme to stabilize the tetrahedral intermediate. The subtilisin contains a three-pronged oxyanion hole forming by side chains of Asn-155, Thr-220 and backbone amide group of Ser-221.(Matthews, Alden et al. 1975) Esterase and carboxylesterase hydrolase superfamilies also use the oxyanion hole to stabilize the tetrahedral intermediate.(Hosokawa 2008) Besides the electrostatic effect, I report the oxyanion hole stabilization through general acid/base catalysis and substrate assisted catalysis in this dissertation. In chapter 2, the mechanism of general acid/base catalysis in the oxyanion hole stabilization is discussed in detail. Briefly, we found the Asp170 of sedolisin could transfer a proton to the negative charged O atom of carbonyl group to stabilize the tetrahedral intermediate and transition states. In chapter 5, in addition to two backbone amide groups (NH) of Ala-13 and Leu-82 of SABP2, the substrate (methyl salicylate) provides a hydrogen bond from its hydroxyl group to participate in the oxyanion hole stabilization.

Computational Studies of Enzymes

Sedolisins

Sedolisins (serine-carboxyl peptidases) belong to the family S53 of clan SB of serine peptidases (MEROPS S53).(Oda, Sugitani et al. 1987, Wlodawer, Li et al. 2001, Wlodawer, Li et al. 2003) The enzymes in this family share some common properties,

including maximum activity at low pH (e.g., pH 3~5) and the presence of conserved acidic residues (aspartate and glutamate) required for the activity.(Murao, Ohkuni et al. 1993, Shibata, Dunn et al. 1998, Tsuruoka, Isono et al. 2003, Tsuruoka, Nakayama et al. 2003, Wlodawer, Li et al. 2003, Reichard, Lechenne et al. 2006, Siezen, Renckens et al. 2007) This dissertation focuses on three of the members, including Kumamolisin-As, Sedolisin, and Pro-Kumamolisin. In table I-1, the general properties of these three enzymes are listed. The members of the family also include tripeptidyl-peptidase 1 (TPP1)(Sleat, Donnelly et al. 1997) for which the loss of the activity as a result of mutations in the TPP1 gene (previously named CLN2) is believed to be the cause of a fatal neurodegenerative disease, classical late-infantile neuronal ceroid lipofuscinosis.(Sleat, Donnelly et al. 1997, Rawlings and Barrett 1999, Lin, Sohar et al. 2001, Golabek and Kida 2006)

ProKumamolisin is the zymogen of Kumamolisin, a member of the sedolisin family and is firstly crystalized in 2004.(Comellas-Bigler, Maskos et al. 2004) The pro-domain of ProKumamolisin shapes like a half-β sandwich core docking to catalytic domain, which is similar with subtilisin pro-enzyme complex. Understanding the ways by which Nature prevents unwanted activation of proteases and the mechanisms for conversion of their zymogens to active enzymes is of considerable interest. In ProKumamolisin, the uncleaved linker segment crossing the active site makes this zymogen be a good candidate to understand self-activation cleavage. In Chapter 1, we performed MD and QM/MM free energy (potential of mean force) simulations on pro-

kumamolisin.(Comellas-Bigler, Maskos et al. 2004) The predicted pK_a values are 5.22 for Asp164 in the activated Kumamolisin. Under the optimum pH 3.0 of Kumamolisin, Asp164 is protonated that is critical for conformational changes of the oxyanion hole formation. We show that the protonation of Asp164, one of the key residues for the enzyme activity, would be able to trigger conformational changes and generate the functional active site for autocatalysis. The mechanism of acylation for autocatalytic cleavage of prodomain is also derived from two-dimensional free energy simulations. The results seem to indicate that one of the reasons for sedolisins to use an aspartate (e.g., Asp164 in kumamolisin) instead of asparagine (the oxyanion-hole residue in classical serine peptidase) as a catalyst might be due in part to the requirement for the creation of a built-in switch that delays the self-activation until secretion into acidic medium.

As the first isolated and described member of the sedolisins family, Pseudomonas sp. 101 sedolisin has been extensively studied by biochemical and mutagenesis approaches along with the determination of crystal structures for several enzyme-inhibitor complexes;(Oda, Takahashi et al. 1994, Ito, Narutaki et al. 1999, Oyama, Abe et al. 1999, Wlodawer, Li et al. 2001, Wlodawer, Li et al. 2001, Wlodawer, Li et al. 2004) For example, the mutants S287A or E80A did not show any peptidase or autocatalytic activity. The D84A mutant showed 0.04% catalytic activity of compared to the wild-type enzyme. D170A and D170N mutants did not show any processing activities or any peptidase activities. A quantum mechanical/molecular mechanical (QM/MM) study

based on energy minimization approach has also been performed on this enzyme.(Bravaya, Bochenkova et al. 2006) Sedolisin has a catalytic triad Ser287−Glu80−Asp84 in place of the Ser-His-Asp triad of related (but usually smaller) classical serine peptidases. Another conserved acidic residue, Asp-170, is structurally equivalent to Asn-155 (part of the oxyanion hole) in subtilisin. The natural substrates for most of the serine carboxyl peptidases are still unknown. Some understanding of the substrate specificity for these enzymes has been achieved through inspection of the inhibitor interactions in the X-ray structures of these enzymes, and/or from the determination of their ability to process peptide libraries.(Wlodawer, Li et al. 2001, Wlodawer, Li et al. 2001, Comellas-Bigler, Fuentes-Prior et al. 2002, Wlodawer, Li et al. 2003, Comellas-Bigler, Maskos et al. 2004, Wlodawer, Li et al. 2004, Wlodawer, Li et al. 2004, Guhaniyogi, Sohar et al. 2009, Pal, Kraetzner et al. 2009) Nevertheless, the origin of differences in substrate preferences for different members of the sedolisin family is still not well understood. I have been involved in the study the detailed acylation mechanism and substrate specificities along with Dr. Xu. (Xu, Yao et al. 2011) The deacylation process was also studied by me and presented in Chapter 2 of my dissertation.

AdoMet-dependent methyltransferases: SABATH and PKMT enzymes

(1). Plants produce a large number of low molecular weight metabolites which may play critical roles in diverse biological processes such as plant growth and development as well as plant interactions with the environment. One chemical modification that has

profound impact on the biological processes mentioned above is enzymatic methylation of carboxyl moieties for some of these low molecular weight metabolites (See Figure I-2), such as salicylic acid (SA), benzoic acid (BA), indole-3-acetic acid (IAA), gibberellic acid (GA), farnesoic acid (FA), cinnamic/coumaric acid (CC) and jasmonic acid (JA), using S-adenosyl-L-methionine (AdoMet) as the methyl donor.(Taiz and Zeiger 2006) The methylated products, such as MeSA, MeBA, and MeIAA, can have biological and ecological functions that are quite different from their precursors. These enzymes that catalyze the methylation of SA, BA, IAA, GA, FA, CC and JA, for instance, are SA methyltransferase (SAMT),(Ross, Nam et al. 1999, Chen, D'Auria et al. 2003) BA methyltransferase (BAMT),(Murfitt, Kolosova et al. 2000) IAA methyltransferase (IAMT),(Qin, Gu et al. 2005, Zhao, Guan et al. 2007, Zhao, Ferrer et al. 2008) GA methyltransferase (GAMT),(Varbanova, Yamaguchi et al. 2007) FA methyltransferase (FAMT),(Yang, Yuan et al. 2006) CC methyltransferase (CCMT)(Kapteyn, Qualley et al. 2007) and JA methyltransferase (JAMT),(Seo, Song et al. 2001) respectively. These enzymes belong to the same SABATH enzyme family(D'Auria, Chen et al. 2003) and can be distinguished according to their substrate preferences. The plant SABATH family is designated based on three methyltransferases (**SA**MT, **BA**MT and **Th**eobromine synthase) which were firstly isolated and characterized. (D'Auria, Chen et al. 2003) The SABATH family also contains some other proteins, including the nitrogen-directed methyltransferases involved in caffeine biosynthesis.(Ogawa, Herai et al. 2001, McCarthy and McCarthy 2007, Yue and Guo 2014) Previous studies have shown that

small changes in primary protein sequences may lead to the functional emergence of SABATH proteins with altered substrate preferences.(Zubieta, Ross et al. 2003, Pichersky, Noel et al. 2006) The SABATH enzymes have been the subject of extensive biochemical and structural investigations.(Ross, Nam et al. 1999, Zubieta, Ross et al. 2003, Pott, Hippauf et al. 2004, Effmert, Saschenbrecker et al. 2005, Hippauf, Michalsky et al. 2010, Tieman, Zeigler et al. 2010, Zhao, Guan et al. 2010) X-Ray crystal structures are available for SAMT and IAMT. Figure I-3 shows their active site structures with substrates. The active site of SAMT shows a series of hydrophobic residues interacting with aromatic rings of substrates. Although there is not any substrate or other ligands shown in the active site of IAMT, the substrate should interact with the hydrophobic residues, such as F243, L242, F364 and/or F158, according to the position of cofactor, SAH. Based on the biochemical study of SAMT, wt SAMT shows specifically high catalytic activity on its native substrate (SA), while the mutant (Y147S/M150H/I225Q/F347Y) changes its substrate specificity to 3-Hydroxybenzoic acid. Nevertheless, the origin of their substrate specificity is still not well understood. In Chapter 3, the catalytic mechanism of SAMT is studied.

(2) The tails of histone proteins are subject to a variety of post-translational modifications, and these modifications are believed to be a part of the histone code for chromatin regulation (Strahl and Allis 2000). One of such modifications is the methylation of a range of lysine (K) residues on histones, including K4, K9, K27, K36, K79 on histone H3 and K20 on histone H4 (Marmorstein 2003). Catalyzed by protein lysine

methyltransferases (PKMTs), histone lysine methylations have been found to affect a variety of important biological processes, including heterochromatin formation, X-chromosome inactivation, transcriptional silencing and activation (Martin and Zhang 2005, Jenuwein 2006). PKMTs may be classified based on their ability to transfer one, two or three methyl groups from S-adenosyl-L-methionine (AdoMet, the methyl donor) to the ε-amino group of target lysine, a special property of the enzymes that is termed as product specificity (Strahl and Allis 2000, Martin and Zhang 2005). Since different methylation states may lead to different downstream events (Turner 2005, Lall 2007, Taverna, Li et al. 2007), it is of fundamental importance to understand the determinant of the product specificity, including the energetic and structural factors that control the product specificity. A number of experimental (Zhang, Tamaru et al. 2002, Xiao, Jing et al. 2003, Zhang, Yang et al. 2003, Collins, Tachibana et al. 2005, Couture, Collazo et al. 2005, Xiao, Jing et al. 2005, Couture, Dirk et al. 2008, Del Rizzo, Couture et al. 2010, Wu, Min et al. 2010, Xu, Wu et al. 2011) and computational studies (Hu and Zhang 2006, Guo and Guo 2007, Wang, Hu et al. 2007, Chu, Xu et al. , Hu, Wang et al. 2008, Zhang and Bruice 2008, Zhang and Bruice 2008, Xu 2009, Bai, Shen et al. 2011) have been performed to understand the origin of the product specificity for PKMTs. Structural and mutational studies have identified a Phe/Tyr switch located at the active site of many SET domain PKMTs (Xiao, Jing et al. 2003, Zhang, Yang et al. 2003, Collins, Tachibana et al. 2005, Couture, Collazo et al. 2005). It has been shown previously that the product specificity of the enzyme may depend on whether this position is occupied by a Phe or

Tyr residue. Indeed, it has been observed that this position tends to be occupied by a tyrosine residue for mono-methyltransferases (e.g. Y305 in SET7/9 (Del Rizzo, Couture et al. 2010) and Y334 in SET8 (Couture, Collazo et al. 2005, Couture, Dirk et al. 2008)) and by a phenylalanine for di- or tri-methyltransferases (e.g. F281 in DIM-5 (Zhang, Yang et al. 2003)). Moreover, the substitution at the Phe/Tyr switch position could lead to the change of the product specificity that is consistent with the observations on the wild-type enzymes. In Chapter 3, we report QM/MM free energy simulations on SET7/9 mutation (Y245A) effects on product specificity.

Salicylic acid binding protein 2* (*SABP2*): *Substrate-assisted catalysis

Substrate-assisted catalysis (SAC) is a process in which the functional groups from substrates, in addition to those from enzymes, contribute to the rate acceleration of the enzyme-catalyzed reactions.(Carter and Wells 1987, Carter, Nilsson et al. 1989, Xu, Guo et al. 2006) The removal of such functional groups from the substrates impairs the catalysis. It has been demonstrated that for engineered enzymes SAC may provide a way of drastically changing substrate specificity.(Carter and Wells 1987) The question remains, however, as to whether naturally occurring enzymes might have already used SAC as one of important strategies for the discrimination between their native substrates and promiscuous substrates that lack the groups capable of participating in SAC. Salicylic acid binding protein 2 (SABP2) studied in this work has specific esterase activity toward methyl salicylate (MeSA) and belongs to the hydrolase superfamily.(Kumar and Klessig 2003, Forouhar, Yang et al. 2005, Kumar, Park et al.

2005, Kumar, Gustafsson et al. 2006, Zhao, Guan et al. 2009) It has been shown(Kumar and Klessig 2003, Forouhar, Yang et al. 2005, Kumar, Park et al. 2005, Kumar, Gustafsson et al. 2006, Zhao, Guan et al. 2009) that Ser-81, His-238 and Asp-210 act as the catalytic triad and the main-chain amide group of Ala-13 participates in the oxyanion-hole interaction. The conversion of MeSA to SA catalyzed by SABP2 is believed to be a part of the signal transduction pathways that activate systemic acquired resistance and local defense responses to plant pathogens.(Kumar and Klessig 2003, Forouhar, Yang et al. 2005, Kumar, Park et al. 2005, Kumar, Gustafsson et al. 2006, Zhao, Guan et al. 2009) Previous studies have demonstrated that SABP2 possesses specific esterase activity toward MeSA, and this activity seems to be much stronger than the activities of this enzyme toward some other promiscuous substrates.(Kumar and Klessig 2003, Forouhar, Yang et al. 2005, Kumar, Park et al. 2005, Kumar, Gustafsson et al. 2006, Zhao, Guan et al. 2009) The exact reasons for the differences in the activities are still not clear. In Chapter 4, we report the results of computational and experimental investigations which suggest that SAC may play an important role in substrate discrimination.

Protein-Protein Interaction Prediction

In recent decades, the prediction of protein protein interaction (PPI) has become a popular research area, and many PPI prediction methods have been developed to enhance PPI prediction accuracy. Machine learning methods have shown their power in protein interaction prediction by integration of different kinds of biological data. However, to our knowledge, there is no method that focuses on enhancing PPI

prediction by integration of individual classifiers. In Chapter 6, we report a novel robust PPI prediction pipeline named PPCM (Protein-Protein Interaction Classifiers Merger) that merges all classifiers from GO2PPI and Phyloprof. As a result, our PPCM showed three advantages compared to individual classifiers. First, all tests of PPCM showed significant improvement in AUC (Area Under Curve) compared with their corresponding individual classifiers. Second, more features (classifiers) led to better improvement in the PPI prediction accuracy (AUC). Finally, all cross species PPCM achieved competitive and even better prediction accuracy compared to the SC PPCM and classifiers with highest AUC in the same category.

References

Bai, Q. F., Y. L. Shen, X. J. Yao, F. Wang, Y. P. Du, Q. Wang, N. Z. Jin, J. Hai, T. J. Hu and J. B. Yang (2011). "Modeling a New Water Channel That Allows SET9 to Dimethylate p53." Plos One **6**(5): e19856.

Blow, D. M. (1976). "Structure and mechanism of chymotrypsin." Accounts of Chemical Research **9**(4): 145-152.

Bravaya, K., A. Bochenkova, B. Grigorenko, I. Topol, S. Burt and A. Nemukhin (2006). "Molecular modeling the reaction mechanism of serine-carboxyl peptidases." Journal of Chemical Theory and Computation **2**(4): 1168-1175.

Bruice, T. C. (2002). "A view at the millennium: the efficiency of enzymatic catalysis." Accounts of chemical research **35**(3): 139-148.

Bruice, T. C. (2006). "Computational approaches: reaction trajectories, structures, and atomic motions. Enzyme reactions and proficiency." Chemical reviews **106**(8): 3119-3139.

Bruice, T. C. (2006). "Computational Approaches: Reaction Trajectories, Structures, and Atomic Motions. Enzyme Reactions and Proficiency." Chemical Reviews **106**(8): 3119-3139.

Bruice, T. C. and S. J. Benkovic (2000). "Chemical Basis for Enzyme Catalysis." Biochemistry **39**(21): 6267-6274.

Bruice, T. C. and F. C. Lightstone (1999). "Ground state and transition state contributions to the rates of intramolecular and enzymatic reactions." Accounts of chemical research **32**(2): 127-136.

Carter, P., B. Nilsson, J. P. Burnier, D. Burdick and J. A. Wells (1989). "ENGINEERING SUBTILISIN BPN' FOR SITE-SPECIFIC PROTEOLYSIS." Proteins-Structure Function and Genetics **6**(3): 240-248.

Carter, P. and J. A. Wells (1987). "ENGINEERING ENZYME SPECIFICITY BY SUBSTRATE-ASSISTED CATALYSIS." Science **237**(4813): 394-399.

Chen, F., J. C. D'Auria, D. Tholl, J. R. Ross, J. Gershenzon, J. P. Noel and E. Pichersky (2003). "An Arabidopsis thaliana gene for methylsalicylate biosynthesis, identified by a biochemical genomics approach, has a role in defense." Plant Journal **36**(5): 577-588.

Chu, Y. Z., Q. Xu and H. Guo (2008). "Understanding Energetic Origins of Product Specificity of SET8 from QM/MM Free Energy Simulations: What Causes the Stop of Methyl Addition during Histone Lysine Methylation?" J Chem Theory Comput **6**(4): 1380-1389.

Chu, Y. Z., Q. Xu and H. Guo (2010). "Understanding Energetic Origins of Product Specificity of SET8 from QM/MM Free Energy Simulations: What Causes the Stop of Methyl Addition during Histone Lysine Methylation?" Journal of Chemical Theory and Computation **6**(4): 1380-1389.

Chu, Y. Z., Yao, J.Z., Guo, H. (2012). "QM/MM MD and Free Energy Simulations of G9a-Like Protein (GLP) and Its Mutants: Understanding the Factors that Determine the Product Specificity " PLos One (accepted).

Collins, R. E., M. Tachibana, H. Tamaru, K. M. Smith, D. Jia, X. Zhang, E. U. Selker, Y. Shinkai and X. Cheng (2005). "In vitro and in vivo analyses of a Phe/Tyr switch controlling product specificity of histone lysine methyltransferases." J Biol Chem **280**(7): 5563-5570.

Comellas-Bigler, M., P. Fuentes-Prior, K. Maskos, R. Huber, H. Oyama, K. Uchida, B. M. Dunn, K. Oda and W. Bode (2002). "The 1.4 angstrom crystal structure of kumamolysin: A thermostable serine-carboxyl-type proteinase." Structure **10**(6): 865-876.

Comellas-Bigler, M., K. Maskos, R. Huber, H. Oyama, K. Oda and W. Bode (2004). "1.2 angstrom crystal structure of the serine carboxyl proteinase pro-kumamolisin: Structure of an intact pro-subtilase." Structure **12**(7): 1313-1323.

Couture, J. F., E. Collazo, J. S. Brunzelle and R. C. Trievel (2005). "Structural and functional analysis of SET8, a histone H4 Lys-20 methyltransferase." Genes & Development **19**(12): 1455-1465.

Couture, J. F., L. M. A. Dirk, J. S. Brunzelle, R. L. Houtz and R. C. Trievel (2008). "Structural origins for the product specificity of SET domain protein methyltransferases." Proc Natl Acad Sci U S A **105**(52): 20659-20664.

D'Auria, J. C., F. Chen and E. Pichersky (2003). The SABATH family of MTs in Arabidopsis thaliana and other plant species. Recent Adv Phytochem. **37:** 95–125.

Del Rizzo, P. A., J. F. Couture, L. M. A. Dirk, B. S. Strunk, M. S. Roiko, J. S. Brunzelle, R. L. Houtz and R. C. Trievel (2010). "SET7/9 Catalytic Mutants Reveal the Role of Active Site Water Molecules in Lysine Multiple Methylation." J Bio Chem **285**(41): 31849-31858.

Effmert, U., S. Saschenbrecker, J. Ross, F. Negre, C. M. Fraser, J. P. Noel, N. Dudareva and B. Piechulla (2005). "Floral benzenoid carboxyl methyltransferases: From in vitro to in planta function." Phytochemistry **66**(11): 1211-1230.

Fersht, A. (1999). Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding, Macmillan.

Forouhar, F., Y. Yang, D. Kumar, Y. Chen, E. Fridman, S. W. Park, Y. Chiang, T. B. Acton, G. T. Montelione, E. Pichersky, D. F. Klessig and L. Tong (2005). "Structural and biochemical studies identify tobacco SABP2 as a methyl salicylate esterase and implicate it in plant innate immunity." Proceedings of the National Academy of Sciences of the United States of America **102**(5): 1773-1778.

Fuxreiter, M. and A. Warshel (1998). "Origin of the Catalytic Power of Acetylcholinesterase: Computer Simulation Studies." Journal of the American Chemical Society **120**(1): 183-194.

Garcia-Viloca, M., J. Gao, M. Karplus and D. G. Truhlar (2004). "How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations." Science **303**(5655): 186-195.

Golabek, A. A. and E. Kida (2006). "Tripeptidyl-peptidase I in health and disease." Biological Chemistry **387**(8): 1091-1099.

Guhaniyogi, J., I. Sohar, K. Das, A. M. Stock and P. Lobel (2009). "Crystal Structure and Autoactivation Pathway of the Precursor Form of Human Tripeptidyl-peptidase 1, the Enzyme Deficient in Late Infantile Ceroid Lipofuscinosis." Journal of Biological Chemistry **284**(6): 3985-3997.

Guo, H. B. and H. Guo (2007). "Mechanism of histone methylation catalyzed by protein lysine methyltransferase SET7/9 and origin of product specificity." Proc Natl Acad Sci U S A **104**(21): 8797-8802.

Hippauf, F., E. Michalsky, R. Q. Huang, R. Preissner, T. J. Barkman and B. Piechulla (2010). "Enzymatic, expression and structural divergences among carboxyl O-methyltransferases after gene duplication and speciation in Nicotiana." Plant Molecular Biology **72**(3): 311-330.

Hosokawa, M. (2008). "Structure and Catalytic Properties of Carboxylesterase Isozymes Involved in Metabolic Activation of Prodrugs." Molecules **13**(2): 412-431.

Hu, P., S. Wang and Y. Zhang (2008). "How do SET-domain protein lysine methyltransferases achieve the methylation state specificity? Revisited by ab initio QM/MM molecular dynamics simulations." Journal of the American Chemical Society **130**(12): 3806-3813.

Hu, P. and Y. K. Zhang (2006). "Catalytic mechanism and product specificity of the histone lysine methyltransferase SET7/9: An ab initio QM/MM-FE study with multiple initial structures." J Am Chem Soc **128**(4): 1272-1278.

Ito, M., S. Narutaki, K. Uchida and K. Oda (1999). "Identification of carboxyl residues in pepstatin-insensitive carboxyl proteinase from Pseudomonas sp. 101 that participate in catalysis and substrate binding." Journal of Biochemistry **125**(1): 210-216.

Jenuwein, T. (2006). "The epigenetic magic of histone lysine methylation." FEBS J **273**(14): 3121-3135.

Kapteyn, J., A. V. Qualley, Z. Z. Xie, E. Fridman, N. Dudareva and D. R. Gang (2007). "Evolution of cinnamate/p-coumarate carboxyl Methyltransferases and their role in the biosynthesis of methylcinnamate." Plant Cell **19**(10): 3212-3229.

Kumar, D., C. Gustafsson and D. F. Klessig (2006). "Validation of RNAi silencing specificity using synthetic genes: salicylic acid-binding protein 2 is required for innate immunity in plants." Plant Journal **45**(5): 863-868.

Kumar, D. and D. F. Klessig (2003). "High-affinity salicylic acid-binding protein 2 is required for plant innate immunity and has salicylic acid-stimulated lipase activity." Proceedings of the National Academy of Sciences of the United States of America **100**(26): 16101-16106.

Kumar, D., S. Park, F. Forouhar, Y. Yang, A. Vlot, E. Fridman, Y. Chiang, T. Acton, V. Shulaev, G. Montelione, E. Pichersky, L. Tong and D. Klessig (2005). "Identification of SA-binding protein 2 (SABP2) as a critical component of plant innate immunity." Phytopathology **95**(6): S55-S55.

Lad, C., N. H. Williams and R. Wolfenden (2003). "The rate of hydrolysis of phosphomonoester dianions and the exceptional catalytic proficiencies of protein and

inositol phosphatases." Proceedings of the National Academy of Sciences **100**(10): 5607-5610.

Lall, S. (2007). "Primers on chromatin." Nat Struc Mol Biol **14**(11): 1110-1115.

Lin, L., I. Sohar, H. Lackland and P. Lobel (2001). "The human CLN2 protein/tripeptidyl-peptidase I is a serine protease that autoactivates at acidic pH." Journal of Biological Chemistry **276**(3): 2249-2255.

Marmorstein, R. (2003). "Structure of SET domain proteins: a new twist on histone methylation." Trends Biochem Sci **28**(2): 59-62.

Martin, C. and Y. Zhang (2005). "The diverse functions of histone lysine methylation." Nat Rev Mol Cell Biol **6**(11): 838-849.

Matthews, D. A., R. A. Alden, J. J. Birktoft, S. T. Freer and J. Kraut (1975). "X-ray crystallographic study of boronic acid adducts with subtilisin BPN' (Novo). A model for the catalytic transition state." Journal of Biological Chemistry **250**(18): 7120-7126.

McCarthy, A. A. and J. G. McCarthy (2007). "The structure of two N-methyltransferases from the caffeine biosynthetic pathway." Plant Physiology **144**(2): 879-889.

Murao, S., K. Ohkuni, M. Nagao, K. Hirayama, K. Fukuhara, K. Oda, H. Oyama and T. Shin (1993). "PURIFICATION AND CHARACTERIZATION OF KUMAMOLYSIN, A NOVEL THERMOSTABLE PEPSTATIN-INSENSITIVE CARBOXYL PROTEINASE FROM BACILLUS NOVOSP MN-32." Journal of Biological Chemistry **268**(1): 349-355.

Murfitt, L. M., N. Kolosova, C. J. Mann and N. Dudareva (2000). "Purification and characterization of S-adenosyl-L-methionine: Benzoic acid carboxyl methyltransferase,

the enzyme responsible for biosynthesis of the volatile ester methyl benzoate in flowers of Antirrhinum majus." Archives of Biochemistry and Biophysics **382**(1): 145-151.

Oda, K., M. Sugitani, K. Fukuhara and S. Murao (1987). "PURIFICATION AND PROPERTIES OF A PEPSTATIN-INSENSITIVE CARBOXYL PROTEINASE FROM A GRAM-NEGATIVE BACTERIUM." Biochimica Et Biophysica Acta **923**(3): 463-469.

Oda, K., T. Takahashi, Y. Tokuda, Y. Shibano and S. Takahashi (1994). "CLONING, NUCLEOTIDE-SEQUENCE, AND EXPRESSION OF AN ISOVALERYL PEPSTATIN-INSENSITIVE CARBOXYL PROTEINASE GENE FROM PSEUDOMONAS SP-101." Journal of Biological Chemistry **269**(42): 26518-26524.

Ogawa, M., Y. Herai, N. Koizumi, T. Kusano and H. Sano (2001). "7-Methylxanthine methyltransferase of coffee plants - Gene isolation and enzymatic properties." Journal of Biological Chemistry **276**(11): 8213-8218.

Oyama, H., S. Abe, S. Ushiyama, S. Takahashi and K. Oda (1999). "Identification of catalytic residues of pepstatin-insensitive carboxyl proteinases from prokaryotes by site-directed mutagenesis." Journal of Biological Chemistry **274**(39): 27815-27822.

Pal, A., R. Kraetzner, T. Gruene, M. Grapp, K. Schreiber, M. Gronborg, H. Urlaub, S. Becker, A. R. Asif, J. Gartner, G. M. Sheldrick and R. Steinfeld (2009). "Structure of Tripeptidyl-peptidase I Provides Insight into the Molecular Basis of Late Infantile Neuronal Ceroid Lipofuscinosis." Journal of Biological Chemistry **284**(6): 3976-3984.

Pauling, L. (1946). "Molecular architecture and biological reactions." Chemical and engineering news **24**(10): 1375-1377.

Pichersky, E., J. P. Noel and N. Dudareva (2006). "Biosynthesis of plant volatiles: Nature's diversity and ingenuity." Science **311**(5762): 808-811.

Pott, M. B., F. Hippauf, S. Saschenbrecker, F. Chen, J. Ross, I. Kiefer, A. Slusarenko, J. P. Noel, E. Pichersky, U. Effmert and B. Piechulla (2004). "Biochemical and structural characterization of benzenoid carboxyl methyltransferases involved in floral scent production in Stephanotis floribunda and Nicotiana suaveolens." Plant Physiology **135**(4): 1946-1955.

Qin, G. J., H. Y. Gu, Y. D. Zhao, Z. Q. Ma, G. L. Shi, Y. Yang, E. Pichersky, H. D. Chen, M. H. Liu, Z. L. Chen and L. J. Qu (2005). "An indole-3-acetic acid carboxyl methyltransferase regulates Arabidopsis leaf development." Plant Cell **17**(10): 2693-2704.

Rawlings, N. D. and A. J. Barrett (1999). "Tripeptidyl-peptidase I is apparently the CLN2 protein absent in classical late-infantile neuronal ceroid lipofuscinosis." Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology **1429**(2): 496-500.

Reichard, U., B. Lechenne, A. R. Asif, F. Streit, E. Grouzmann, O. Jousson and M. Monod (2006). "Sedolisins, a new class of secreted proteases from aspergillus fumigatus with endoprotease or tripeptidyl-peptidase activity at acidic pHs." Applied and Environmental Microbiology **72**(3): 1739-1748.

Ross, J. R., K. H. Nam, J. C. D'Auria and E. Pichersky (1999). "S-adenosyl-L-methionine : salicylic acid carboxyl methyltransferase, an enzyme involved in floral scent production and plant defense, represents a new class of plant methyltransferases." Archives of Biochemistry and Biophysics **367**(1): 9-16.

Seo, H. S., J. T. Song, J. J. Cheong, Y. H. Lee, Y. W. Lee, I. Hwang, J. S. Lee and Y. D. Choi (2001). "Jasmonic acid carboxyl methyltransferase: A key enzyme for jasmonate-regulated plant responses." Proceedings of the National Academy of Sciences of the United States of America **98**(8): 4788-4793.

Shibata, M., B. M. Dunn and K. Oda (1998). "Substrate specificity of pepstatin-insensitive carboxyl proteinase from Bacillus coagulans J-4." Journal of Biochemistry **124**(3): 642-647.

Siezen, R. J., B. Renckens and J. Boekhorst (2007). "Evolution of prokaryotic subtilases: Genome-wide analysis reveals novel subfamilies with different catalytic residues." Proteins-Structure Function and Bioinformatics **67**(3): 681-694.

Sleat, D. E., R. J. Donnelly, H. Lackland, C. G. Liu, I. Sohar, R. K. Pullarkat and P. Lobel (1997). "Association of mutations in a lysosomal protein with classical late-infantile neuronal ceroid lipofuscinosis." Science **277**(5333): 1802-1805.

Strahl, B. D. and C. D. Allis (2000). "The language of covalent histone modifications." Nature **403**(6765): 41-45.

Taiz, L. and E. Zeiger (2006). Plantphysiology. Sunderland, MA, Sinauer Associates.

Taverna, S. D., H. Li, A. J. Ruthenburg, C. D. Allis and D. J. Patel (2007). "How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers." Nat Struc Mol Biol **14**(11): 1025-1040.

Tieman, D., M. Zeigler, E. Schmelz, M. G. Taylor, S. Rushing, J. B. Jones and H. J. Klee (2010). "Functional analysis of a tomato salicylic acid methyl transferase and its role in synthesis of the flavor volatile methyl salicylate." Plant Journal **62**(1): 113-123.

Tsuruoka, N., Y. Isono, O. Shida, H. Hemmi, T. Nakayama and T. Nishino (2003). "Alicyclobacillus sendaiensis sp nov., a novel acidophilic, slightly thermophilic species isolated from soil in Sendai, Japan." International Journal of Systematic and Evolutionary Microbiology **53**: 1081-1084.

Tsuruoka, N., T. Nakayama, M. Ashida, H. Hemmi, M. Nakao, H. Minakata, H. Oyama, K. Oda and T. Nishino (2003). "Collagenolytic serine-carboxyl proteinase from Alicyclobacillus sendaiensis strain NTAP-1: Purification, characterization, gene cloning, and heterologous expression." Applied and Environmental Microbiology **69**(1): 162-169.

Turner, B. M. (2005). "Reading signals on the nucleosome with a new nomenclature for modified histones." Nat Struc Mol Biol **12**(2): 110-112.

Varbanova, M., S. Yamaguchi, Y. Yang, K. McKelvey, A. Hanada, R. Borochov, F. Yu, Y. Jikumaru, J. Ross, D. Cortes, C. J. Ma, J. P. Noel, L. Mander, V. Shulaev, Y. Kamiya, S. Rodermel, D. Weiss and E. Pichersky (2007). "Methylation of gibberellins by Arabidopsis GAMT1 and GAMT2." Plant Cell **19**(1): 32-45.

Wang, S. L., P. Hu and Y. K. Zhang (2007). "Ab initio quantum mechanical/molecular mechanical molecular dynamics simulation of enzyme catalysis: The case of histone lysine methyltransferase SET7/9." J Phys Chem B **111**(14): 3758-3764.

Warshel, A. (1991). Computer modeling of chemical reactions in enzymes and solutions, Wiley New York.

Warshel, A., G. Naray-Szabo, F. Sussman and J. K. Hwang (1989). "How do serine proteases really work?" Biochemistry **28**(9): 3629-3637.

Wlodawer, A., M. Li, Z. Dauter, A. Gustchina, K. Uchida, H. Oyama, B. M. Dunn and K. Oda (2001). "Carboxyl proteinase from Pseudomonas defines a novel family of subtilisin-like enzymes." Nature Structural Biology **8**(5): 442-446.

Wlodawer, A., M. Li, A. Gustchina, Z. Dauter, K. Uchida, H. Oyama, N. E. Goldfarb, B. M. Dunn and K. Oda (2001). "Inhibitor complexes of the Pseudomonas serine-carboxyl proteinase." Biochemistry **40**(51): 15602-15611.

Wlodawer, A., M. Li, A. Gustchina, H. Oyama, B. M. Dunn and K. Oda (2003). "Structural and enzymatic properties of the sedolisin family of serine-carboxyl peptidases." Acta biochimica Polonica **50**(1): 81-102.

Wlodawer, A., M. Li, A. Gustchina, H. Oyama, K. Oda, B. B. Beyer, J. Clemente and B. M. Dunn (2004). "Two inhibitor molecules bound in the active site of Pseudomonas sedolisin: a model for the bi-product complex following cleavage of a peptide substrate." Biochemical and Biophysical Research Communications **314**(2): 638-645.

Wlodawer, A., M. Li, A. Gustchina, N. Tsuruoka, M. Ashida, H. Minakata, H. Oyama, K. Oda, T. Nishino and T. Nakayama (2004). "Crystallographic and biochemical investigations of kumamolisin-As, a serine-carboxyl peptidase with collagenase activity." Journal of Biological Chemistry **279**(20): 21500-21510.

Wu, H., J. R. Min, V. V. Lunin, T. Antoshenko, L. Dombrovski, H. Zeng, A. Allali-Hassani, V. Campagna-Slater, M. Vedadi, C. H. Arrowsmith, A. N. Plotnikov and M. Schapira (2010). "Structural Biology of Human H3K9 Methyltransferases." Plos One **5**(1): e8570.

Xiao, B., C. Jing, G. Kelly, P. A. Walker, F. W. Muskett, T. A. Frenkiel, S. R. Martin, K. Sarma, D. Reinberg, S. J. Gamblin and J. R. Wilson (2005). "Specificity and mechanism of the histone methyltransferase Pr-Set7." Genes & Development **19**(12): 1444-1454.

Xiao, B., C. Jing, J. R. Wilson, P. A. Walker, N. Vasisht, G. Kelly, S. Howell, I. A. Taylor, G. M. Blackburn and S. J. Gamblin (2003). "Structure and catalytic mechanism of the human histone methyltransferase SET7/9." Nature **421**(6923): 652-656.

Xu, Q., Chu, Y-Z, Guo, H-B, Smith, J.C., Guo, H. (2009). "Energy Triplets for Writing Epigenetic Marks: Insights from QM/MM Free-Energy Simulations of Protein Lysine Methyltransferases." Chemistry - A European Journal **15**: 12596 – 12599.

Xu, Q., H. Guo and A. Wlodawer (2006). "The importance of dynamics in substrate-assisted catalysis and specificity." Journal of the American Chemical Society **128**(18): 5994-5995.

Xu, S. T., J. Wu, B. F. Sun, C. Zhong and J. P. Ding (2011). "Structural and biochemical studies of human lysine methyltransferase Smyd3 reveal the important functional roles of its post-SET and TPR domains and the regulation of its activity by DNA binding." Nucleic Acids Res **39**(10): 4438-4449.

Yang, Y., J. S. Yuan, J. Ross, J. P. Noel, E. Pichersky and F. Chen (2006). "An Arabidopsis thaliana methyltransferase capable of methylating farnesoic acid." Archives of Biochemistry and Biophysics **448**(1-2): 123-132.

Yao, J. Z., Y. Z. Chu, R. An and H. Guo (2012). "Understanding Product Specificity of Protein Lysine Methyltransferases from QM/MM Molecular Dynamics and Free Energy Simulations: The Effects of Mutation on SET7/9 beyond the Tyr/Phe Switch." Journal of Chemical Information and Modeling **52**(2): 449-456.

Yue, Y. and H. Guo (2014). "Quantum Mechanical/Molecular Mechanical Study of Catalytic Mechanism and Role of Key Residues in Methylation Reactions Catalyzed by Dimethylxanthine Methyltransferase in Caffeine Biosynthesis." Journal of Chemical Information and Modeling **54**(2): 593-600.

Zhang, X. and T. C. Bruice (2008). "Enzymatic mechanism and product specificity of SET-domain protein lysine methyltransferases." Proc Natl Acad Sci U S A **105**(15): 5728-5732.

Zhang, X., H. Tamaru, S. I. Khan, J. R. Horton, L. J. Keefe, E. U. Selker and X. Cheng (2002). "Structure of the Neurospora SET domain protein DIM-5, a histone H3 lysine methyltransferase." Cell **111**(1): 117-127.

Zhang, X., Z. Yang, S. I. Khan, J. R. Horton, H. Tamaru, E. U. Selker and X. Cheng (2003). "Structural basis for the product specificity of histone lysine methyltransferases." Mol Cell **12**(1): 177-185.

Zhang, X. D. and T. C. Bruice (2008). "Product specificity and mechanism of protein lysine methyltransferases: Insights from the histone lysine methyltransferase SET8." Biochemistry **47**(25): 6671-6677.

Zhao, N., J. L. Ferrer, J. Ross, J. Guan, Y. Yang, E. Pichersky, J. P. Noel and F. Chen (2008). "Structural, biochemical, and phylogenetic analyses suggest that indole-3-acetic acid methyltransferase is an evolutionarily ancient member of the SABATH family." Plant Physiology **146**(2): 455-467.

Zhao, N., J. Guan, J. L. Ferrer, N. Engle, M. Chern, P. Ronald, T. J. Tschaplinski and F. Chen (2010). "Biosynthesis and emission of insect-induced methyl salicylate and methyl benzoate from rice." Plant Physiology and Biochemistry **48**(4): 279-287.

Zhao, N., J. Guan, F. Forouhar, T. J. Tschaplinski, Z. M. Cheng, L. Tong and F. Chen (2009). "Two poplar methyl salicylate esterases display comparable biochemical properties but divergent expression patterns." Phytochemistry **70**(1): 32-39.

Zhao, N., J. Guan, H. Lin and F. Chen (2007). "Molecular cloning and biochemical characterization of indole-3-acetic acid methyltransferase from poplar." Phytochemistry **68**(11): 1537-1544.

Zubieta, C., J. R. Ross, P. Koscheski, Y. Yang, E. Pichersky and J. P. Noel (2003). "Structural basis for substrate recognition in the salicylic acid carboxyl methyltransferase family." Plant Cell **15**(8): 1704-1716.

Appendix

Table I-1. Serine-carboxyl peptidases studies by author's lab.

Optimum pH, catalytic triad, and oxyanion hole residue are summarized below.

Properties	Sedolisin	Pro-Kumamolisin	Kumamolisin-As
Opt. pH	3.0	3.0	3.9
Catalytic triad	Ser287 Glu80 Asp84	Ser278 Glu78 Asp82	Ser278 Glu78 Asp82
Oxy. residue	Asp170	Asp164	Asp164
References	Oda et al 1987	Comellas et al 2004	Tsuruoka et al 2003

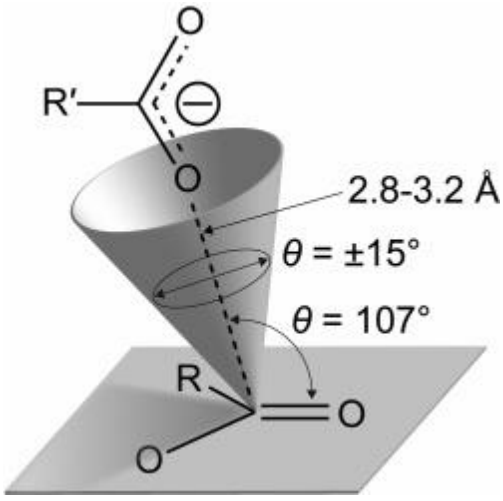


Figure I-1. Near Attach Conformers (NACs).

NACs is a useful tool, but not a state during reaction. (From Bruice 2006)

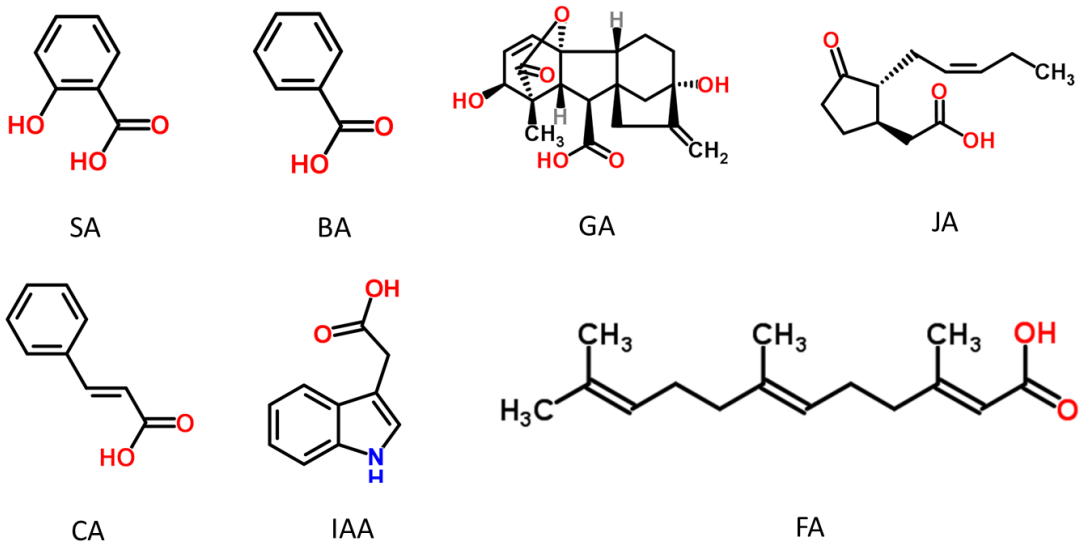


Figure I-2.Carboxyl moieties catalyzed by SABATH family enzymes.

SA, Salicylic Acid; BA, Benzoic Acid; GA Gibberellic Acid; JA, Jasmonic Acid; CA, Cinnamic/Coumaric Acid; JA, Jasmonic Acid; FA, Farnesoic Acid.

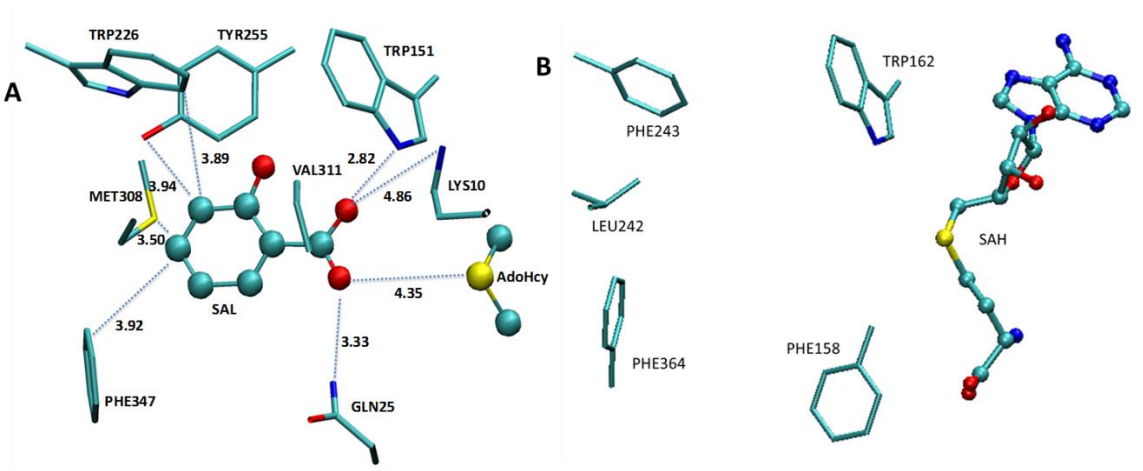


Figure I-3. Active Sites of SAMT (A) and IAMT (B).

The figure is plotted based on X-Ray crystal structures, SAMT (PDB ID 1M6E) and IAMT (PDB ID 3B5I).

COMPUTATIONAL METHODOLOGY

Understanding the origin of the catalytic power and detailed mechanisms of enzymes is a major task for current computational enzymology. Enzymes behavior can be described by computational models based on the fundamental laws of chemistry and physics, such as quantum mechanics, molecular mechanics and statistical mechanics.(Brooks, Bruccoleri et al. 1983, Brooks, Brooks et al. 2009) Computational modeling may provide the information that is difficult to be determined by experiments.(Yang, Gao et al. 2003) In this dissertation, we applied a series of computational methods to describe how enzymes work, including molecular dynamics (MD) simulations, quantum mechanical/molecular mechanical (QM/MM) MD and free energy simulations.

Molecular Dynamic Simulations.

MD simulations are applied for studying the structural dynamic and thermodynamic properties of enzymes. By integration of Newton's equations of motion, MD simulation is able to determine the coordinates of the enzymes as a function of time. Although quantum corrections to the atomic dynamics are not applied, MD simulation should be a valid tool for sampling configuration change of enzymes such as Pro-Kumamolisin self-activation discussed in Chapter 1.

The molecular mechanics (MM) force fields, also called the potential energy function, are used to calculate the potential energy of the system such as an enzyme. So far, several popular MM force fields have been widely used for MD simulations, such as CHARMM force field and AMBER force field. In this dissertation, We applied the all-atom

CHARMM force field(MacKerell, Bashford et al. 1998) to calculate the potential energy for enzymes and ions. Water molecules were described by an explicit water model called TIP3P model.(Jorgensen, Chandrasekhar et al. 1983)

The MD simulations are run with periodic boundary conditions in a water box. Langevin thermostat (Adelman and Doll 1976) is used to control the system temperature. The electrostatic interaction is handled by Particle Mesh Ewald method.

Quantum Mechanical/Molecular Mechanical Molecular Dynamic and Free Energy (Potential of Mean Force) Simulations.

Treating a whole biological macromolecule like an enzyme by the quantum mechanics requires extremely large amounts of computing time. In order to shrink the computer time to a reasonable level, the hybrid QM/MM potentials are introduced to combine the MM potentials and the QM potentials. The QM/MM potentials are widely used to study enzymatic reactions involving bond making and breaking processes. In the QM/MM simulations, only a small amount of atoms are treated by the QM potentials and those atoms are called as the QM region. The remainder, so called MM region, is treated by MM force fields. Typically, QM region includes the parts of the enzyme that are involved in the reaction.

The QM potentials consist of *ab initio* and semi-empirical approaches. Many *ab initio* approaches have been developed, such as Hartree-Fock (HF) theory, MP2 theory, and density functional theory (DFT).(Gao and Truhlar 2002) A widely used and efficient hybrid-functional version of DFT, B3LYP (Becke, three-parameter, Lee-Yang-Parr) that

implemented in GAMESS-US program was used in this dissertation for benchmark calculations (See Chapter 5). However, *ab initio* QM approaches are very expensive in term of computer time. Therefore, some semi-empirical methods have been developed and applied to the quantum mechanical calculations, such as MNDO, AM1, PM3, and SCC-DFTB.(Cui, Elstner et al. 2001) Among these semi-empirical approaches, the SCC-DFTB method, derived from density functional theory with a second-order expansion of the DFT, is as fast as AM1 and PM3, but shows more reliable results of energies and vibration frequencies in some cases.(Elstner, Porezag et al. 1998) By using our benchmark calculations (See Chapter 5), the energies and atomic geometry determined by SCC-DFTB were found to be reasonably comparable to the *ab initio* methods, such as B3LYP with MM potentials. Therefore, in this dissertation we applied SCC-DFTB in the quantum mechanical calculations.

The treatment of the covalent boundary atoms in the combined QM/MM method is one difficult issue. Several methods have been developed to deal with this issue, including hydrogen link atom,(Field, Bash et al. 1990) delocalized Gaussian MM (DGMM) charges, and generalized hybrid orbital (GHO) method.(Gao, Amara et al. 1998) The hydrogen link atom method, that seems to work well with the SCC-DFTB/MM method, was adopted in this dissertation.

Stochastic Boundary (SB) condition(Brooks, Brunger et al. 1985) was applied in our QM/MM MD free energy simulations of the catalytic mechanisms of enzymes. A system under the typical SB condition is divided into three parts, including reaction region,

buffer region, and reservoir zone. First, an atom in the active site is selected as the reference center, which is close to reactive center. Second, the atoms in the reservoir region are deleted, while the remainder part is in the reaction zone. Third, the reaction zone can be further separated into the reaction region and the buffer region. The atoms in the buffer region are treated by Langevin dynamics, while the atoms in the reaction region are treated by molecular dynamics. QM atoms are in the reaction region.

The umbrella sampling method(Torrie and Valleau 1974) implemented in the CHARMM program along with the Weighted Histogram Analysis Method (WHAM)(Kumar, Bouzida et al. 1992) are applied to determine the change of the free energy (potential of mean force) as a function of the reaction coordinate. The reaction coordinates are chosen depending on the catalytic mechanisms of the enzymes. The umbrella sampling methods are able to sample the states far from the equilibrium by adding a force constrain on the reaction coordinates. So, these states of coordinate space such as transition state can be explored using a harmonic potential function. This harmonic function has been defined as

$$U = \frac{\mathbf{1}}{E} \times K \times (r - r_{\mathbf{0}})^{E}$$

where the default value of E is 2, K is the force constant, and r is the reaction coordinate, and r_0 is the origin of the harmonic function on the selected reaction coordinate. The relative free energy change as a function of the reaction coordinate can be determined by the WHAM program through statistically translation of the data sampled by umbrella sampling.

References

Adelman, S. and J. Doll (1976). "Generalized Langevin equation approach for atom/solid-surface scattering: General formulation for classical scattering off harmonic solids." The Journal of Chemical Physics **64**: 2375.

Brooks, B. R., C. L. Brooks, A. D. MacKerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels and S. Boresch (2009). "CHARMM: the biomolecular simulation program." Journal of computational chemistry **30**(10): 1545-1614.

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus (1983). "CHARMM - A PROGRAM FOR MACROMOLECULAR ENERGY, MINIMIZATION, AND DYNAMICS CALCULATIONS." Journal of Computational Chemistry **4**(2): 187-217.

Brooks, C. L., A. Brunger and M. Karplus (1985). "ACTIVE-SITE DYNAMICS IN PROTEIN MOLECULES - A STOCHASTIC BOUNDARY MOLECULAR-DYNAMICS APPROACH." Biopolymers **24**(5): 843-865.

Cui, Q., M. Elstner, E. Kaxiras, T. Frauenheim and M. Karplus (2001). "A QM/MM implementation of the self-consistent charge density functional tight binding (SCC-DFTB) method." Journal of Physical Chemistry B **105**(2): 569-585.

Elstner, M., D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai and G. Seifert (1998). "Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties." Physical Review B **58**(11): 7260-7268.

Field, M. J., P. A. Bash and M. Karplus (1990). "A COMBINED QUANTUM-MECHANICAL AND MOLECULAR MECHANICAL POTENTIAL FOR MOLECULAR-DYNAMICS SIMULATIONS." Journal of Computational Chemistry **11**(6): 700-733.

Gao, J., P. Amara, C. Alhambra and M. J. Field (1998). "A Generalized Hybrid Orbital (GHO) Method for the Treatment of Boundary Atoms in Combined QM/MM Calculations." The Journal of Physical Chemistry A **102**(24): 4714-4721.

Gao, J. and D. G. Truhlar (2002). "QUANTUM MECHANICAL METHODS FOR ENZYME KINETICS." Annual Review of Physical Chemistry **53**(1): 467-505.

Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein (1983). "COMPARISON OF SIMPLE POTENTIAL FUNCTIONS FOR SIMULATING LIQUID WATER." Journal of Chemical Physics **79**(2): 926-935.

Kumar, S., D. Bouzida, R. H. Swendsen, P. A. Kollman and J. M. Rosenberg (1992). "THE WEIGHTED HISTOGRAM ANALYSIS METHOD FOR FREE-ENERGY CALCULATIONS ON BIOMOLECULES .1. THE METHOD." Journal of Computational Chemistry **13**(8): 1011-1021.

MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin and M. Karplus (1998). "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†." The Journal of Physical Chemistry B **102**(18): 3586-3616.

Torrie, G. M. and J. P. Valleau (1974). "MONTE-CARLO FREE-ENERGY ESTIMATES USING NON-BOLTZMANN SAMPLING - APPLICATION TO SUBCRITICAL LENNARD-JONES FLUID." Chemical Physics Letters **28**(4): 578-581.

Yang, W., Y. Q. Gao, Q. Cui, J. Ma and M. Karplus (2003). "The missing link between thermodynamics and structure in F1-ATPase." Proceedings of the National Academy of Sciences **100**(3): 874-879.

CHAPTER I

UNDERSTANDING THE AUTOCATALYTIC PROCESS OF PRO-KUMAMOLISIN ACTIVATION FROM MOLECULAR DYNAMICS AND QM/MM FREE ENERGY SIMULATIONS

A version of this chapter was originally published by Jianzhuang Yao, Alexander Wlodawer, and Hong Guo:

Jianzhuang Yao, Alexander Wlodawer, and Hong Guo "Understanding the Autocatalytic Process of Pro-kumamolisin Activation from Molecular Dynamics and Quantum Mechanical/Molecular Mechanical (QM/MM) Free-Energy Simulations." *Chemistry-A European Journal* 19 (2013): 10849-10852.

Abstract

Understanding the ways by which Nature prevents unwanted activation of proteases and the mechanisms for conversion of their zymogens to active enzymes is of considerable interest. We performed MD and QM/MM free energy (potential of mean force) simulations on pro-kumamolisin, a member of the sedolisin family. We show that the protonation of Asp164, one of the key residues for the enzyme activity, would be able to trigger conformational changes and generate the functional active site for autocatalysis. The mechanism of acylation for autocatalytic cleavage of prodomain is also derived from two-dimensional free energy simulations. The results seem to indicate that one of the reasons for sedolisins to use an aspartate (e.g., Asp164 in kumamolisin) instead of asparagine (the oxyanion-hole residue in classical serine peptidase) as a catalyst might be due in part to the requirement for the creation of a built-in switch that delays the self-activation until secretion into acidic medium.

Introduction

Proteolytic enzymes are synthesized as inactive precursors or zymogens to promote folding, prevent unwanted protein degradation, and provide a mechanism for regulating protein function through proteolytic activation. The availability of the three-dimensional structures of zymogens along with those of the active enzymes has provided considerable insights into the ways by which Nature prevents unwanted activation, as well as mechanisms for the conversion of zymogens to active enzymes.(Khan and James 1998) Nevertheless, detailed understanding of the activation processes and energetics involved is still lacking, and a variety of hypotheses concerning the activation mechanisms need to be confirmed.

Kumamolsin belongs to the recently characterized family of serine-carboxyl peptidases (sedolisins)(Wlodawer, Li et al. 2003) that were originally described by Murao, Oda, and co-workers about 25 years ago.(Oda, Sugitani et al. 1987, Oda 2012) Sedolisins are present in a wide variety of organisms, including archaea, bacteria, molds, slime molds (mixomycetes), amoebas, fishes, and mammals, and they are most active at low pH and some of them also at high temperature. The members of the family also include tripeptidyl-peptidase 1 (TPP1) for which the loss of the activity as a result of mutations in the *TPP1* gene (previously named *CLN2*) is believed to be the cause of a fatal neurodegenerative disease, classical late-infantile neuronal ceroid lipofuscinosis.(Sleat, Donnelly et al. 1997) Like other proteolytic enzymes, sedolisins are synthesized as inactive precursors to prevent unwanted activation and proteolysis. The inactive precursors for

sedolisins include propeptide of ~200 amino acids in length. These propeptides may play a role in the folding of the proteins as well as protecting them from being prematurely activated. The activation cleavage of the prodomains was found to occur after the zymogens release into the acidic environment, leading to the production of the active enzymes.(Oda 2012)

Crystal structures have been determined for the inactive Ser278Ala pro-kumamolisin mutant(Comellas-Bigler, Maskos et al. 2004) as well as for the active enzyme.(Comellas-Bigler, Fuentes-Prior et al. 2002) The superposition of the structure of the Ser278Ala pro-kumamolisin mutant (determined at the neutral pH of the crystallization buffer) with kumamolisin showed that the catalytic domain of the mutant exhibits a virtually identical structure compared to the active enzyme, with an rms deviation of only about 0.5 Å for 350 C_α atoms.(Comellas-Bigler, Maskos et al. 2004) The pro-kumamolisin structure has therefore proved that the catalytic domain has basically adopted a mature-like conformation already in the zymogen form. One of the key structural features of the proenzyme is the presence of a salt bridge between the P_3-Arg169p linker residue (here the suffix 'p' designates the residues in the pro-domain) and Asp164 from the catalytic domain.(Comellas-Bigler, Maskos et al. 2004) The existence of this salt bridge seems to prevent the formation of the functional configuration for Asp164, one of the key residues in the oxyanion hole formation.(Guo, Wlodawer et al. 2005, Guo, Wlodawer et al. 2006, Xu, Guo et al. 2006, Xu, Guo et al. 2007) An interesting question is whether the protonation of Asp164 and subsequently breaking of this salt

bridge during secretion into an acidic medium would trigger conformational changes that would lead to generation of the well-positioned general acid/base catalyst (Asp164) and the functional active site (as observed in the mature enzyme), in the preparation of the self-activation.(Comellas-Bigler, Maskos et al. 2004)

To answer these questions, we applied molecular dynamics (MD) and two-dimensional free energy (potential of mean force) simulations with quantum mechanical/molecular mechanical (QM/MM) potential to map the detailed process and energetics of the conversion of the kumamolysin zymogen to the active enzyme. MD simulations and QM/MM approaches have been extensively used for providing understanding of many biological processes, but, to the best of our knowledge, they have not been used for studying zymogen activation. The CHARMM35b program(Brooks, Bruccoleri et al. 1983) was used in this study along with the Self-Consistent Charge Density Functional Tight Binding (SCC-DFTB) method for the QM treatment.(Cui, Elstner et al. 2001) The QM/MM approach used here has been applied previously in the studies of the catalytic mechanisms for a number of enzyme systems, including the enzymatically active forms of kumamolisin-As and sedolisin, members of the same family.(Guo, Wlodawer et al. 2005, Guo, Wlodawer et al. 2006, Xu, Guo et al. 2006, Xu, Guo et al. 2007, Xu, Li et al. 2010)

Method

The initial coordinates for pro-kumamolisin were based on the crystallographic structure (PDB code: 1T1E, resolution 1.18Å) of the S278A mutant of pro-kumamolisin(Comellas-

Bigler, Maskos et al. 2004). Using MOE, the A287 residue was converted back to Ser278 by a point mutation, followed by optimizations. The structure was constructed with different protonation states of the titratable residues under acidic (pH = 3.0) conditions, based on pKa values calculated with MOE or local environments. Asp164 is deprotonated under neutral condition and forms an ion pair with P3-Arg169p in the crystal structure of 1T1E. On the other hand, it was presumed to be protonated under acidic condition to act as a general acid/base catalyst.

The whole system was solvated in a 85 x 106 x 93 Å size water box containing TIP3P water(Jorgensen, Chandrasekhar et al. 1983) using the NAMD2.9 program (Phillips, Braun et al. 2005). 53 chloride ions (Cl-) were added to neutralize the positive charged system (pH=3). The resulting system contained 78919 atoms, including 23678 bulk water molecules, a calcium ion and 74 water molecules form the crystal crystal structure. MD simulation was run with periodic boundary conditions with CHARMM force field (PARAM27) (MacKerell, Bashford et al. 1998). Simulation was run with time step 0.002 ps and the SHAKE algorithm (Ryckaert, Ciccotti et al. 1977) was used to constrain all bonds to hydrogens. Coordinates were saved every 500 steps. After 200 steps minimization was performed with protein atoms fixed, 2000 steps unconstrained minimization was performed. Then system was gradually heated to 300 K from 100 K within a time period of 100 ps. Finally, 10 ns equilibration and 10 ns product constant pressure (NPT) runs was followed at 1 atm. Langevin thermostat (Adelman and Doll 1976) was used to control the system temperature. The electrostatic interaction was handled by Particle Mesh Ewald

method. A typical structure from 15ns MD simulation was selected to perform QM/MM MD simulation.

In QM/MM MD simulation, the solvated system was partitioned with stochastic boundary condition (Brooks, Brunger et al. 1985): with the carbonyl carbon atom of P1-His171p as the reference center, the reservoir regions with radius R >22Å away were deleted, the buffer regions with 20 Å reference center, the reservoir regions with radius R > faLangevin Dynamics (LD), and the reaction region with radius R < 20 Å was applied to QM/MM MD simulations. The resulting system after the partition contained around 3400 atoms, including 74 water molecules from the X-ray structure. In the reaction region, the QM part included the side chains of Glu78, Asp82, Asp164, and Ser287, as well as part of the linker-peptide including the carbonyl of P2-Pro170p, the backbone of P1-His171p and the Cα and amine group of P1'-Phe172p. The atoms in the QM region was simulated with the Self-Consistent Charge Density Functional Tight Binding (SCC-DFTB) (Cui, Elstner et al. 2001)method in the CHARMM program. The rest of the reaction region was treated with the MM method using the all-hydrogen potential function (PARAM27). The covalent bonds on the boundary between the QM and MM regions were treated with link-atom approach (Field, Bash et al. 1990), which is available in the CHARMM program.

After generating the effective potential energy surfaces (PES), the two-dimensional free energy changes (Potential of mean force, or PMF) from the substrate complex to the acyl-enzyme were determined with the umbrella sampling method (Torrie and Valleau 1974) implemented in the CHARMM program and two-dimensional Weighted Histogram

Analysis method (2d-WHAM) (Kumar, Bouzida et al. 1992). Time step for QM/MM MD simulation is 0.001 ps per step. The first reaction coordinate, ξ1 = R(C-N) – R(C…Oγ), is the difference between the distance of the scissile peptide bond R[C(His171p)-N(Phe172p)] and the distance of the nucleophilic attack R[C(His171p)…Oγ(Ser278)]. The second reaction coordinate, ξ2 = r(Oδ2-Ha)-r(O-Ha), is the difference between the distance of the scissile hydrogen covalent bond R(Oδ2 (ASP164)-Ha (ASP164)) and the distance of the hydrogen bond R(O(His171p)-Ha (ASP164)). 1296 windows were used in the 2D PMF calculations and free energy mesh-contour plot were created with a bin size of 0.05 x 0.05 Å2. For each window, we performed 50 ps heating, 50 ps equilibration 50 ps production run with a force constants of 1000kcal·Mol-1·Å-2. Time step for QM/MM MD simulation is 0.001 ps per step. The first reaction coordinate, ξ1 = R(C-N) – R(C…Oγ), is the difference between the distance of the scissile peptide bond R[C(His171p)-N(Phe172p)] and the distance of the nucleophilic attack R[C(His171p)…Oγ(Ser278)]. The second reaction coordinate, ξ2 = r(Oδ2-Ha)-r(O-Ha), is the difference between the distance of the scissile hydrogen covalent bond R(Oδ2 (ASP164)-Ha (ASP164)) and the distance of the hydrogen bond R(O(His171p)-Ha (ASP164)). 1296 windows were used in the 2D PMF calculations and free energy mesh-contour plot were created with a bin size of 0.05 x 0.05 2equilibration 50 ps production run with a force constants of 1000kcal·Mol·Å.

Results and Discussion

The X-ray structure of the Ser278Ala pro-kumamolisin mutant (PDB ID: 1T1E) has been used to generate the model for the wild-type pro-kumamolisin through manually changing Ala278 to Ser278 (Figure 1.1A). As can be seen from Figure 1.1A, the catalytic Ser residue, along with Glu78 and Asp82, seems to be located at the correct position and would attack the carbonyl carbon of P_1-His171p during the autocatalytic cleavage. This agrees with the earlier suggestion that the catalytic triad in the pro-kumamolisin structure has already adopted a mature-like conformation.(Comellas-Bigler, Maskos et al. 2004) By contrast, Asp164 is located at a different position compared to that observed in the active enzyme,(Comellas-Bigler, Fuentes-Prior et al. 2002) due to forming a salt bridge with the P_3-Arg169p (Figure 1.1A). Asp164 is therefore unable to participate in the stabilization of the TI during the autocatalytic cleavage of the peptide bond between P_1-His171p and P_1'-Phe172p. We aimed to address the question as to whether the protonation of Asp164 would be able to trigger conformational changes that would position this residue so that it could act as a general acid/base catalyst. The mechanism for the acylation reaction of the autocatalytic process was also studied using the two-dimensional QM/MM free energy simulations. The first reaction coordinate, ξ_1 = R(C-N) – R(C…O_γ), is the distance difference between the scissile peptide bond, R[C(His171p)-N(Phe172p)], and the nucleophilic attack distance from Ser278, R[C(His171p)…O_γ(Ser278)]. The second reaction coordinate, ξ_2 = r($O_{\delta 2}$-H_a)-r(O-H_a), is the one for the proton transfer (H_a) between Asp164 and the carboxyl group oxygen ($O_{\delta 2}$) of His171p.

The average active-site structure for pro-kumamolisin after the protonation of Asp164 and 15*ns* MD simulations is given in Figure 1.1B. It is evident that the salt bridge between the P_3-Arg169p linker residue and Asp164 from the catalytic domain has already been broken; Arg169p moved away from its position in pro-kumamolisin and interacts instead with solvent molecules. Interestingly, after the MD simulations the active site has changed to the functional configuration with Asp164 well aligned to act as a general acid/base catalyst. As mentioned earlier, the catalytic triad in the pro-kumamolisin structure has already adopted a mature-like conformation. Figure 1.1B shows that this catalytic machinery has not been significantly altered by the MD simulations and seems to remain functional. Thus, our simulations have confirmed the hypothesis that the protonation of Asp164 would be able to trigger conformational changes to generate the well-positioned general acid/base catalyst (Asp164) and a functional active site for the next step of the autocatalytic pro-kumamolisin activation. We have also studied the mechanism of the acylation reaction for the autocatalytic cleavage of the prodomain. The two-dimensional free energy (potential of mean force) map for the acylation step of the autocatalytic cleavage is given in Figure 1.2A. The map was obtained based on the QM(SCC-DFTB)/MM potential and a typical active structure that is quite close to the average structure obtained from the MD simulations (see above). The free energy curve along the minimum free energy path in Figure 1.2A is shown in Figure 1.2B. A comparison of the curve in Figure 1.2B with the one for the acylation reaction involving kumamolisin_As(Xu, Guo et al. 2007) shows that the two

curves are rather similar, even though the barriers in Figure 1.2B are slightly lower (by about 2 kcal/mol). The calculated barrier for the acylation is about 13 kcal/mol for activation. We also determined the free energy curve without the protonation of Asp164 and breaking of the salt bridge, and the barrier is considerably higher than the one in Figure 1.2B (by more than 10 kcal/mol). This result suggests that the protonation of Asp164 and breaking of the salt bridge are necessary for autolysis, consistent with experimental observations. This result suggests that the protonation of Asp164 and breaking of the salt bridge are necessary for autolysis, consistent with experimental observations.

The average active site structures for the autolysis obtained from the QM/MM free energy simulations are given in Figure 1.3A-C; the locations of H_a, H_b and H_c are also shown on the right in each case. Ser278 is the nucleophile that attacks the carbonyl carbon atom C(His171p) , whereas Glu78 and Asp164 act as the general base and acid catalysts, respectively. The motions of the three protons that may be important for catalysis [i.e., those between Asp164 and the carboxyl group oxygen ($O_{\delta 2}$) of His171p, Ser278 and Glu78, and Glu78 and Asp82] are monitored based on the trajectories from the corresponding windows of the free energy simulations and plotted on the right of Figure 1.3A-C. Figure 1.3A shows that, for the initial state before the chemical step, these protons are located on Asp164, Ser278, and Asp82, respectively, consistent with the average structure shown on the left. The average active-site structure near the transition state for the nucleophilic attack (TS_{a1}) is given in Figure 1.3B. Figure 1.3B (left)

shows that there is a low-barrier hydrogen bond formed at TS_{a1}, with H_a basically located in the middle of the two oxygen atoms, and this is consistent with the plot in the top panel of Figure 1. 3B. The same seems to be true for H_b as well. Figure 1. 3C shows that at TI the protons are located on P_1-His171p, Glu78, and Asp82, respectively.

We have also studied the mechanism of the acylation reaction for the autocatalytic cleavage of the prodomain. The two-dimensional free energy (potential of mean force) map for the acylation step of the autocatalytic cleavage is given in Figure 1. 2A. The map was obtained based on the QM(SCC-DFTB)/MM potential and a typical active structure that is quite close to the average structure obtained from the MD simulations (see above). The free energy curve along the minimum free energy path in Figure 1.2A is shown in Figure 1.2B. A comparison of the curve in Figure 1.2B with the one for the acylation reaction involving kumamolisin_As(Xu, Guo et al. 2007) shows that the two curves are rather similar, even though the barriers in Figure 1.2B are slightly lower (by about 2 kcal/mol). The calculated barrier for the acylation is about 13 kcal/mol for activation. We also determined the free energy curve without the protonation of Asp164 and breaking of the salt bridge, and the barrier is considerably higher than the one in Figure 1.2B (by more than 10 kcal/mol). This result suggests that the protonation of Asp164 and breaking of the salt bridge are necessary for autolysis, consistent with experimental observations.

Conclusion

We were able to support the hypothesis that the protonation of Asp164 would be able to trigger conformational changes to position this residues as a general acid/base catalyst, thus reconstructing the functional active site. We have also studied the mechanism of the acylation reaction for the autocatalytic cleavage of the prodomain and obtained the free energy map and catalytic mechanism for the acylation reaction. The results show that the autolysis uses the same catalytic mechanism that was indicated in our earlier studies involving the catalytic domains of kumamolisin-As and sedolisin.(Guo, Wlodawer et al. 2005, Guo, Wlodawer et al. 2006, Xu, Guo et al. 2006, Xu, Guo et al. 2007, Xu, Li et al. 2010) One of the interesting questions posed by sedolisins is why this family uses an aspartic acid residue (Asp164) instead of Asn (the residue that creates the oxyanion hole in the classical serine peptidases). The results reported here seem to indicate that one of the reasons might be due to the requirement for the creation of a built-in switch that delays the self-activation of sedolisins until secretion into the acidic medium. Additional investigations are necessary to confirm this hypothesis.

Acknowledgement

We thank Drs. Toru Nakayama and Qin Xu for helpful discussions and Professor Martin Karplus for a gift of the CHARMM program. This work was supported by the National Science Foundation (Grant number: 0817940 to H.G.) and the NSF TeraGrid resources

provided by the University of Texas at Austin, as well as by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

References

Adelman, S. and J. Doll (1976). "Generalized Langevin equation approach for atom/solid-surface scattering: General formulation for classical scattering off harmonic solids." The Journal of Chemical Physics **64**: 2375.

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus (1983). "CHARMM - A PROGRAM FOR MACROMOLECULAR ENERGY, MINIMIZATION, AND DYNAMICS CALCULATIONS." Journal of Computational Chemistry **4**(2): 187-217.

Brooks, C. L., A. Brunger and M. Karplus (1985). "ACTIVE-SITE DYNAMICS IN PROTEIN MOLECULES - A STOCHASTIC BOUNDARY MOLECULAR-DYNAMICS APPROACH." Biopolymers **24**(5): 843-865.

Comellas-Bigler, M., P. Fuentes-Prior, K. Maskos, R. Huber, H. Oyama, K. Uchida, B. M. Dunn, K. Oda and W. Bode (2002). "The 1.4 angstrom crystal structure of kumamolysin: A thermostable serine-carboxyl-type proteinase." Structure **10**(6): 865-876.

Comellas-Bigler, M., K. Maskos, R. Huber, H. Oyama, K. Oda and W. Bode (2004). "1.2 Å Crystal Structure of the Serine Carboxyl Proteinase Pro-Kumamolisin: Structure of an Intact Pro-Subtilase." Structure **12**(7): 1313-1323.

Comellas-Bigler, M., K. Maskos, R. Huber, H. Oyama, K. Oda and W. Bode (2004). "1.2 angstrom crystal structure of the serine carboxyl proteinase pro-kumamolisin: Structure of an intact pro-subtilase." Structure **12**(7): 1313-1323.

Cui, Q., M. Elstner, E. Kaxiras, T. Frauenheim and M. Karplus (2001). "A QM/MM implementation of the self-consistent charge density functional tight binding (SCC-DFTB) method." Journal of Physical Chemistry B **105**(2): 569-585.

Field, M. J., P. A. Bash and M. Karplus (1990). "A COMBINED QUANTUM-MECHANICAL AND MOLECULAR MECHANICAL POTENTIAL FOR MOLECULAR-DYNAMICS SIMULATIONS." Journal of Computational Chemistry **11**(6): 700-733.

Guo, H. B., A. Wlodawer and H. Guo (2005). "A general acid-base mechanism for the stabilization of a tetrahedral adduct in a serine-carboxyl peptidase: A computational study." Journal of the American Chemical Society **127**(45): 15662-15663.

Guo, H. B., A. Wlodawer, T. Nakayama, Q. Xu and H. Guo (2006). "Catalytic role of proton transfers in the formation of a tetrahedral adduct in a serine carboxyl peptidase." Biochemistry **45**(30): 9129-9137.

Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein (1983). "COMPARISON OF SIMPLE POTENTIAL FUNCTIONS FOR SIMULATING LIQUID WATER." Journal of Chemical Physics **79**(2): 926-935.

Khan, A. R. and M. N. G. James (1998). "Molecular mechanisms for the conversion of zymogens to active proteolytic enzymes." Protein Science **7**(4): 815-836.

Kumar, S., D. Bouzida, R. H. Swendsen, P. A. Kollman and J. M. Rosenberg (1992). "THE WEIGHTED HISTOGRAM ANALYSIS METHOD FOR FREE-ENERGY CALCULATIONS ON BIOMOLECULES .1. THE METHOD." Journal of Computational Chemistry **13**(8): 1011-1021.

MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus (1998). "All-atom empirical potential for molecular modeling and dynamics studies of proteins." Journal of Physical Chemistry B **102**(18): 3586-3616.

Oda, K. (2012). "New families of carboxyl peptidases: serine-carboxyl peptidases and glutamic peptidases." Journal of Biochemistry **151**(1): 13-25.

Oda, K., M. Sugitani, K. Fukuhara and S. Murao (1987). "PURIFICATION AND PROPERTIES OF A PEPSTATIN-INSENSITIVE CARBOXYL PROTEINASE FROM A GRAM-NEGATIVE BACTERIUM." Biochimica Et Biophysica Acta **923**(3): 463-469.

Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé and K. Schulten (2005). "Scalable molecular dynamics with NAMD." Journal of Computational Chemistry **26**(16): 1781-1802.

Ryckaert, J.-P., G. Ciccotti and H. J. C. Berendsen (1977). "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes." Journal of Computational Physics **23**(3): 327-341.

Sleat, D. E., R. J. Donnelly, H. Lackland, C. G. Liu, I. Sohar, R. K. Pullarkat and P. Lobel (1997). "Association of mutations in a lysosomal protein with classical late-infantile neuronal ceroid lipofuscinosis." Science **277**(5333): 1802-1805.

Torrie, G. M. and J. P. Valleau (1974). "MONTE-CARLO FREE-ENERGY ESTIMATES USING NON-BOLTZMANN SAMPLING - APPLICATION TO SUBCRITICAL LENNARD-JONES FLUID." Chemical Physics Letters **28**(4): 578-581.

Wlodawer, A., M. Li, A. Gustchina, H. Oyama, B. M. Dunn and K. Oda (2003). "Structural and enzymatic properties of the sedolisin family of serine-carboxyl peptidases." Acta biochimica Polonica **50**(1): 81-102.

Xu, Q., H. Guo and A. Wlodawer (2006). "The importance of dynamics in substrate-assisted catalysis and specificity." Journal of the American Chemical Society **128**(18): 5994-5995.

Xu, Q., H. B. Guo, A. Wlodawer, T. Nakayama and H. Guo (2007). "The QM/MM molecular dynamics and free energy simulations of the acylation reaction catalyzed by the serine-carboxyl peptidase kumamolisin-As." Biochemistry **46**(12): 3784-3792.

Xu, Q., L. Li and H. Guo (2010). "Understanding the Mechanism of Deacylation Reaction Catalyzed by the Serine Carboxyl Peptidase Kumamolisin-As: Insights from QM/MM Free Energy Simulations." J. Phys. Chem. B, **114**(32): 10594–10600.

Appendix

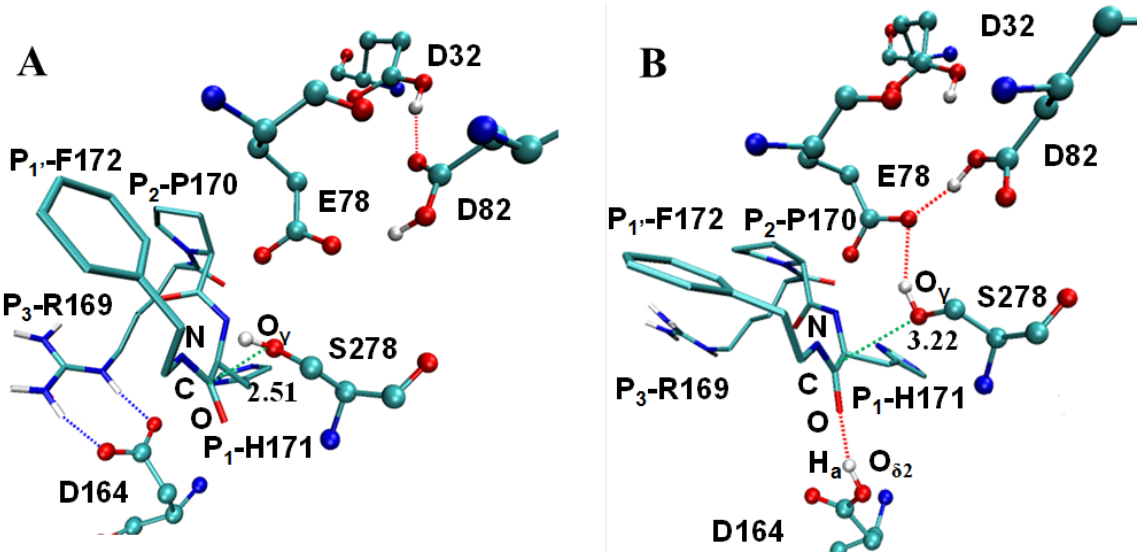


Figure 1.1. The active site structures of pro-kumamolisin.

(A) The active site of the crystal structure of pro-kumamolisin with Ala278 changed to Ser (wild-type) before the protonation of Asp164. The residues from the catalytic domain are shown in ball-and-stick and those from the pro-domain are in stick. For clarity, most of hydrogen atoms are not added. (B) The average structure after the protonation of Oδ2 and 15ns MD simulations. A similar result was obtained with the protonation of the other oxygen of Asp164.

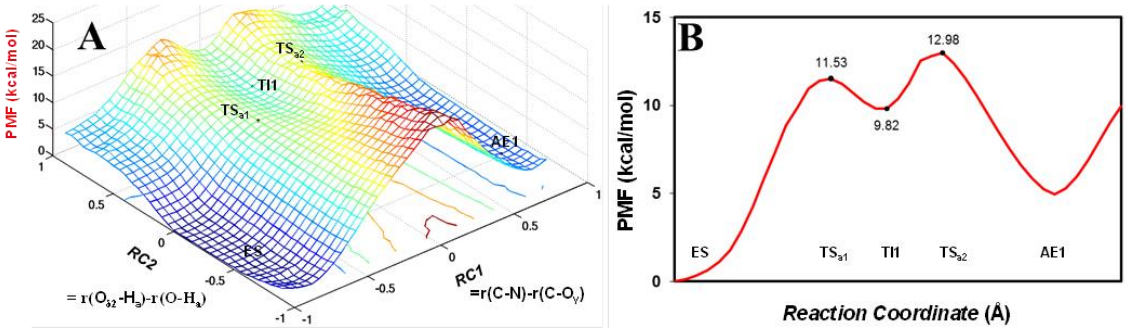


Figure 1.2. The free energy profiles of pro-kumamolisin.

(A) Two-dimensional free energy (potential of mean force) map for the acylation step of self-activation obtained from the QM/MM simulations. (B) The free energy curve along the minimum free energy path in A. Here ES is used for the initial state of pro-Kum before acylation (i.e., after the conformational changes to generate the active configuration), TSa1 is the first TS, TI1 is the tetrahedral intermediate for acylation, TSa2 is the second TS, and AE1 is the state corresponding to the acyl-enzyme in the enzyme-catalyzed reaction.

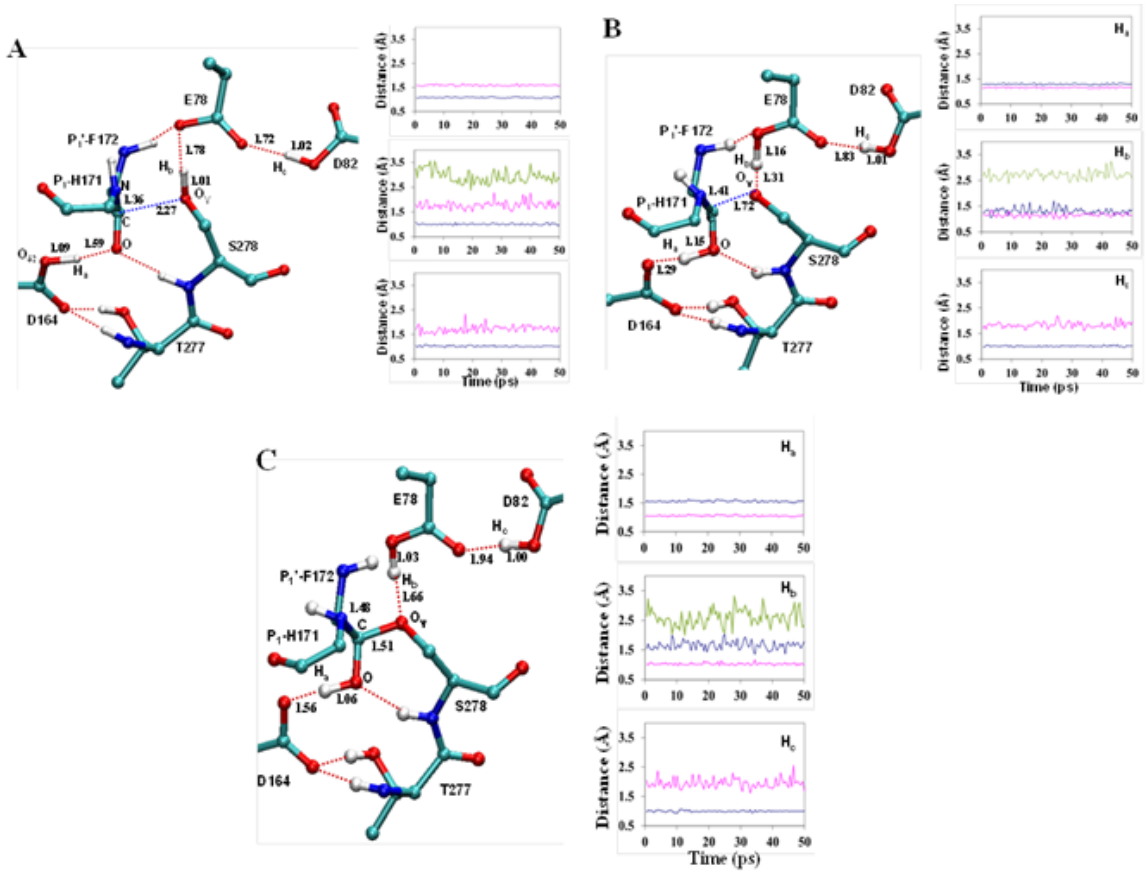


Figure 1.3. Certain active site structures along the reactant path (see Figure 1.2B); the positions of H_a, H_b, and H_c are also given on the right.

Left: Initial structure before acylation. Right: Top panel: the distances from O (D164) and O (carbonyl group of P_1-His) to Ha (colored in blue and magenta, respectively); Middle panel: the distances from O (S278), O (E78) and N (amino group of P_1-His) to H_b (colored in blue, magenta and green, respectively). The locations of the hydrogen atoms, H_a, H_b and H_c, are shown in the structure on the left; Bottom panel: the distances from O (E78) and O (D82) to H_c (colored in magenta and blue, respectively). (B) TS_{a1} (C) TI.

CHAPTER 2

QM/MM AND FREE ENERGY SIMULATIONS OF DEACYLATION REACTION CATALYZED BY SEDOLISIN, A SERINE CARBOXYL PEPTIDASE

A version of this chapter was originally published by Jianzhuang Yao, Qin Xu and Hong Guo:

Jianzhuang Yao, Qin Xu and Hong Guo "QM/MM and free-energy simulations of deacylation reaction catalysed by sedolisin, a serine-carboxyl peptidase." *Molecular Simulation* 39 (2013): 206-213.

Abstract

Quantum mechanical/molecular mechanical (QM/MM) free energy simulations were performed to understand the deacylation reaction catalyzed by sedolisin (a serine-carboxyl peptidase) and to elucidate the catalytic mechanism and role of the active-site residues during the process. The results given here demonstrate that Asp170 may act as a general acid/base catalyst for the deacylation reaction. It is also shown that the electrostatic oxyanion hole interactions involving Asp170 may be less effective in transition state stabilization for the deacylation step in the sedolisin-catalyzed reaction compared to the general acid/base mechanism. The proton transfer processes during the enzyme-catalyzed process were examined and their role in the catalysis are discussed.

Introduction

Sedolisin is the first isolated and described member of the sedolisin family of enzymes (MEROPS S53) that has been extensively studied by biochemical and mutagenesis approaches; for a recent review of the enzymes in this family, see Ref. 1.(Oda 2012)

Sedolisin and other enzymes in the sedolisin family normally exhibit maximum activity at low pH and often high temperature with a fold resembling that of subtilisin (even though they are significantly larger).(Wlodawer, Li et al. 2003) Thus, it is of considerable interest to understand the similarities and differences between sedolisins and classical serine peptidase (e.g., subtilisin) in their catalytic mechanisms and substrate specificities. The importance of the enzymes in the sedolisin family has been demonstrated by the fact that a fatal neurodegenerative disease, classical late-infantile neuronal ceroid lipofuscinosis, is caused by the loss of the activity of the enzyme tripeptidyl-peptidase 1 (TPP1) belonging to this family as a result of mutations in the *TPP1* gene (previously named *CLN2*).(Sleat, Donnelly et al. 1997) The crystal structures have shown that the defining features for the enzymes in the sedolisin family are a unique catalytic triad, Ser-Glu-Asp (e.g., Ser-287−Glu-80−Asp-84 for sedolisin) in place of the Ser-His-Asp triad of related classical serine peptidases.(Wlodawer, Li et al. 2001, Wlodawer, Li et al. 2001, Wlodawer, Li et al. 2003, Wlodawer, Li et al. 2004) In addition, there is an aspartic acid residue (e.g., Asp170 for sedolisin) that occupies the position of Asn155 in subtilisin (a residue that is involved in creating the oxyanion hole in the serine peptidase) that is also of importance for the enzyme's activity.

Computational studies have been performed previously for determining the similarities and differences between the catalytic mechanisms of "classical" serine peptidases and sedolisins as well as inhibitor binding process.(Guo, Beahm et al. 2004, Guo, Wlodawer et al. 2005, Guo, Wlodawer et al. 2006, Xu, Guo et al. 2006, Sun, Li et al. 2007, Xu, Li et al.

2010) The simulations were also performed for the acylation step of the enzyme-catalyzed reaction involving sedolisin. (Bravaya, Bochenkova et al. 2006, Xu, Yao et al. 2011) Nevertheless, to the best of our knowledge the deacylation reaction for sedolisin has not been examined. One conclusion from the earlier quantum mechanical/molecular mechanical (QM/MM) molecular dynamics (MD) and free energy simulations on the acylation reaction of sedolisin(Xu, Yao et al. 2011) is that an aspartic acid residue (i.e., Asp170 that replaces Asn155 in subtilisin) might act as a general acid catalyst to protonate the substrate and stabilize the tetrahedral intermediate (TI) during the nucleophilic attack. Moreover, Asp170 was found to act as a general base during the formation of the acyl-enzyme from the tetrahedral intermediate based on the computer simulations and may therefore play multiple roles during the catalysis.(Xu, Yao et al. 2011) It was proposed that the general acid/base mechanism observed for the acylation reactions involving sedolisin and kumamolisin-As may be applied to other members of the sedolisin family as well. However, the question remains as to whether the general acid/base mechanism involving Asp170 may also be used for the deacylation process during the catalysis of sedolisin. Therefore, it would be of considerable interest to examine the nature of the general acid/base catalysis as well as the detailed mechanism for the deacylation reaction catalyzed by sedolisin.

Here we examine this deacylation step based on the structure of the acyl-enzyme from the previous investigations using QM/MM MD and free energy (potential of mean force) simulations. It is shown that Asp170 indeed acts as a general acid/base catalyst in the

deacylation step. Moreover, the detailed mechanism of deacylation is provided through the free energy simulations.

Method

A fast semi-empirical density-functional method (SCC-DFTB)(Cui, Elstner et al. 2001, Elstner, Cui et al. 2003) implemented in the CHARMM program(Brooks, Bruccoleri et al. 1983) was used in the QM/MM molecular dynamics (MD) and free energy (potential of mean force or PMF) simulations of the deacylation process. The efficiency of the semi-empirical QM methods (such as SCC-DFTB) makes possible to sample millions of structures and conformations for enzyme systems and to determine the free energy profiles of the enzyme-catalyzed reactions. These important tasks for understanding enzymes are often not feasible with the high-level first-principle *ab initio* methods without use of considerable computational resources. However, comparisons of the performance of the SCC-DFTB method with the results from high-level *ab intio* calculations on certain models as well as on simplified enzyme systems are of importance. Such tests have been done for kumamolsin-As in our previous investigations; for more detailed discussions concerning the use of the SCC-DFTB method and comparison with high level *ab initio* methods for kumamolisin-As, see Refs (Xu and Guo 2004, Guo, Rao et al. 2005, Guo, Wlodawer et al. 2006, Xu, Guo et al. 2007).

The initial coordinates for studying the deacylation process were prepared from the structures at the end of the free energy simulation of the acylation process with the LLE*FL substrate that were published earlier.(Xu, Yao et al. 2011) To generate the reactant

structure for acylation, the X-ray structure of sedolisin complexed with the inhibitor pseudoiodotyrostatin (PDB ID: 1NLU) and the X-ray structure of the S287A mutant of prokumamolisin (PDB ID: 1T1E) were superposed to produce the initial structure. In the structures used for studying deacylation, the acyl-enzyme was formed through the carbonyl group of the scissile peptide bond and Ser287. For this complex (designated as AE in Figure 2.1), the first product from the peptide bond cleavage was still retained in the active site. In order to remove this peptide fragment, a constraint with the force constant of 100 kcal·Mol^{-1}·$Å^{-2}$ was applied to a reaction coordinate that is the average value of the two distances from the carbonyl carbon of the P_1 residue to the nitrogen of the $P_{1'}$ amino group and to the $P_{1'}$ C_β atom, respectively. This reaction coordinate was gradually increased from 4.00 Å to 11.50 Å in 16 windows with 40ps for each window. The first product was removed from the site of the bond breaking making. The resulting structure (AE2) was then treated with two cycles of minimization (see below), heating (from 50 K to 300 K in 20 ps) and equilibration (at 300 K for 80 ps). A water molecule (W1) that is close to the P_1 carbonyl carbon was changed from the MM region into QM region (see below), and this water molecule was used later for the nucleophilic attack in the deacylation process. The system was solvated by a modified TIP3P water model(Jorgensen, Chandrasekhar et al. 1983, Neria, Fischer et al. 1996) using the standard procedure in the CHARMM program. The stochastic boundary MD method(Brooks, Brunger et al. 1985) was applied to the solvated system. To be consistent with the previous study of the acylation reaction, the reaction region was the part of the system

with radius $R < 16$ Å, and the buffer region had R equal to 16 Å $\leq R \leq$ 18 Å. The reference center for this partitioning was chosen to be the carbonyl carbon atom (C) of the residue at the P_1 site of the substrate. The Langevin Dynamics (LD) in the buffer region had frictional constants as 250 ps^{-1} for the protein atoms and 62 ps^{-1} for the water molecules. In the reaction region, the side chains of Glu80, Asp84, Asp170, and Ser287 as well as a part of the substrate involved in the reaction [the carbonyl of P_2 (C=O), the whole P_1 residue, and a part of $P_{1'}$ (C_α-NH)] were treated by the quantum mechanical approach, and the rest of the system by molecular mechanical (MM) method. The atoms in the QM region were treated with SCC-DFTB method as described above. The all-hydrogen potential function (PARAM22)(MacKerell, Bashford et al. 1998) was used for the MM method. The link-atom approach(Field, Bash et al. 1990) available in the CHARMM program was used to separate the QM and MM regions. The resulting system contains 3430 atoms (2544 enzyme atoms, 85 substrate atoms, and 801 atoms for 267 water molecules, including 79 crystal water molecules). The same procedure was used for the study of the deacylation processes in the D170N mutant for which the acyl-enzyme structure was generated simply by replacing Asp170 by Asn. This allows a better comparison of the deacylation process in wild-type and mutated enzyme with the similar initial condition. The simulations were also performed for the cases in which the proton on Asp170 was removed (i.e., with deprontonated Asp170) or fixed by the Shake algorithm(Ryckaert, Ciccotti et al. 1977) (to prevent Asp164 acting as the general acid/base catalyst for the nucleophilic attacks).

The entire stochastic boundary system was first optimized by adopted basis Newton-Rhaphson (ABNR) method. Then the system was gradually heated from 50 to 300 K in 20 ps, equilibrated at 300 K for 80 ps and then simulated at 300K for at least 1 ns. A 1-fs time step was used for integration of the equations of motion, and the coordinates were saved every 50 fs for analyses. As discussed previously,(Xu, Yao et al. 2011) for the acylation process the reaction coordinate is defined as R(C-N) – R(C…O_γ), the difference between the bond length of the scissile peptide bond R(C-N) and the distance involving O_γ(Ser287) during the nucleophilic attack. For the deacylation process, the reaction coordinate ξ is defined as ξ = R(C-O_γ) – R(C…O_{w1}), where R[C…O_{W1}] is the distance involving W1 (which makes nucleophilic attack on P_1 carbonyl carbon). The umbrella sampling method(Torrie and Valleau 1974) implemented in the CHARMM program with Weighted Histogram Analysis method (WHAM)(Kumar, Bouzida et al. 1992) was applied to determine the change of the free energy (Potential of mean force, or PMF) for the deacylation reaction in wild-type sedolisin under different conditions (see above) and its mutant. Sixteen to nineteen windows were used to generate the free energy profiles for changing AE2 to the product complex (EP), and for each window 100 ps simulations were formed (50 equilibration and 50ps production run). The force constants used in these free energy simulations were 100 ~ 800 kcal·Mol^{-1}·$Å^{-2}$.

Results and Discussion

Structural and dynamic properties of the acyl-enzyme (AE2) after the first product is removed. The average active-site structure for the acyl-enzyme obtained from the

QM/MM MD simulations is given in Figure 2.2A. As is evident from Figure 2.2A, the nucleophile W1 is well positioned to attack the carbonyl carbon of the substrate during the MD simulations. Glu80 and Asp170 are located at the right positions to act as the general base and acid catalysts, respectively, during the nucleophilic attack, consistent with the previous investigations.(Xu, Yao et al. 2011) Figure 2.2A also shows that Asp84 forms a strong hydrogen bond with Glu80 in the acyl-enzyme. The similar observation was made for kumamolisin-As.(Sun, Li et al. 2007) The top panel on the right in Figure 2.2A shows that Asp170 interacts with the carbonyl group of the P_1 residue with H_a mainly located on Asp170. This is consistent with the expectation that the role of this residue is to stabilize the tetrahedral intermediate either through the general acid/base mechanism (i.e., similar to its role in the acylation step of the sedolisin catalyzed reaction) or by the strengthening of the interaction during the charge formation (i.e., as in the case of classical serine peptidases); the role of Asp170 for deacylation obtained from the simulations will be discussed later. The hydrogen bonding interactions from the backbone amide group of Thr286 as well as its side chain seem to be crucial in stabilizing the position of Asp170. Given the important function of Asp170, the stabilization of Asp170 to the correct position should be of importance. In addition to Asp170, the amide group of the serine residue (Ser287) interacts with the carbonyl oxygen as in the case of classical serine peptidases.

Comparison of different free energy profiles of the deacylation reaction. The free energy (potential of mean force) profiles as a function of the reaction coordinate (ξ) for the

deacylation reaction in wild-type (red solid line) and D170N (green dashed line) are compared in Figure 2.2B; the free energy profile for the case in which the proton on Asp170 was removed (i.e., with deprontonated Asp170. Orange dotted-dash line) or fixed by the Shake algorithm(Ryckaert, Ciccotti et al. 1977) (to prevent Asp164 acting as the general acid/base catalyst for the nucleophilic attacks. Blue dotted line) are also given. It should be pointed out that the free energy profiles in Figure 2.2B were based on the one-dimensional free-energy simulations (as it would be very time consuming to determine multi-dimensional free energy profiles). For the step-wise bond breaking and making events, the use of one-dimensional free energy simulations may not pose a serious problem. However, for the concerted processes some information that would be contained in the multi-dimensional free-energy profiles might be absent in the one-dimensional profiles. Indeed, there are certain proton transfers during the deacylation reaction which may be coupled to the rearrangement of the system (i.e., the nucleophilic attacks and the formation of the product). For instance, our earlier two-dimensional free energy simulations(Guo, Wlodawer et al. 2005, Guo, Wlodawer et al. 2006) on the nucleophilic attack of Ser-278 on the AcIPF inhibitor in kumamolisin-As showed that there are important free-energy relationships between the nucleophilic attack and the proton transfers. Such free energy relationships between the nucleophilic attack and proton transfers are absent in the one-dimensional free energy profiles such as those in Figure 2.2B. However, the previous study(Guo, Wlodawer et al. 2006) also indicated that one-dimensional free energy simulations with the selection of the suitable reaction coordinate

(as we did here) should be sufficient to determine the key energetic properties for the reactions.

Figure 2.2B shows that TI2 is considerably more stable in wild-type than in D170N, leading to the relatively low free energy barrier for the formation of TI2. The free energy profile of D170N is rather similar to the one for which the proton on Asp170 was fixed by the Shake algorithm, confirming the suggestion that the barrier lowering is due, at least in part, to the proton transfer. These results suggest that for the deacylation process in sedolisin the general acid/base mechanism is probably more effective in lowering the energy barrier than the stabilization of the charge formation on the carbonyl oxygen through the hydrogen bonding interaction (e.g., by Asn170 in the D170N mutant), consistent with the results of the previous investigations of the acylation process.(Xu, Yao et al. 2011) Although the barrier for the formation of TI2 is higher for D170N compared to those for wild-type, it is still considerably lower than the barrier for the reaction involving the deprotonated Asp170 in wild-type. This indicates that the enzyme cannot use the deprotonated Asp170 during the reaction, and this is consistent with the fact that sedolisin is active at low pH.(Wlodawer, Li et al. 2003, Oda 2012) The results of the simulations reported here extend the earlier proposal(Xu, Yao et al. 2011) concerning the importance of the general acid/base mechanism involving Asp170 to the case of the stabilization of TI2 in the deacylation process; the results are consistent with the available mutagenesis studies on the enzymes in this family. For instance, for the D164N mutant of

kumamolisin-As, a low but appreciable level of catalytic activity (1.3%) was found,(Okubo, Li et al. 2006) corresponding to an increase of the activation barrier of about 2.5 kcal/mol.

Structures of TS_{d1}, TI2, TS_{d2} and EP and Proton Transfer processes. The average active-site structures of TS_{d1} (near the transition state for the nucleophilic attack by W1), TI2 (the tetrahedral intermediate of deacylation), TS_{d2} (near the transition state for the formation of the product complex) and EP (product complex) are given in Figure 2.3 along with the fluctuations of some distances involved in the proton transfers which are coupled with the deacylation reaction. The average structure in Figure 2.3A shows that a strong hydrogen bond is formed between Asp170 and the carbonyl group of the P_1 residue, and this is supported by the fluctuations of the distances of H_a to O (D170) and O from the carbonyl group, respectively, in the top panel on the right in Figure 2.3A. Moreover, the proton on W1 (H_c) has already moved to Glu80 at TS_{d1}, whereas H_b is now mainly located on Asp84 as expected (i.e., due to the protonation of Glu80). Sedolisin was found to have relatively higher specificity for the substrates with Glu at the P_1 site.(Wlodawer, Li et al. 2004) It was found in our earlier simulations(Xu, Yao et al. 2011) that a proton transfer/shift occurred during the formation of the acyl-enzyme involving Asp-170 and the sidechain of P_1-Glu. Such proton transfer or partial proton transfer was, however, not observed in the deacylation process (see Figure 2.3).

The locations of the three protons (H_a, H_b and H_c) at different stages of the reaction in wild-type are plotted in Figure 2.4. Figure 2.4A shows that H_a is located on Asp170 in both AE2 (green) and EP (magenta) states with the H_a-O (Asp170) distance as about 1 Å.

However, this proton moves to the carbonyl oxygen of the His residue at P_1 of the substrate when the tetrahedral intermediate (TI2) is formed (blue). Figure 2.4B shows that while H_b is mainly located on Asp84 in TI2 (blue), it is involved in a low-barrier hydrogen bond in AE2 and EP and spends a lot of time on both Asp84 and Glu80. In Figure 2.4C, H_c moves to Glu80 as a result of the formation of TI2 (from green to blue) and then moves to S287 (magenta) during the formation of the product.

Conclusion

The deacylation reaction catalyzed by sedolisin has been studied by the QM/MM MD and free energy simulations. The results given here suggested that Asp170 acts as a general acid/base catalyst not only in the acylation step, but also in the deacylation step. It is also shown that the electrostatic oxyanion hole interactions involving Asp170 may be less effective in transition state stabilization for the deacylation step in the sedolisin-catalyzed reaction compared to the general acid/base mechanism. The proton transfer processes during the enzyme-catalyzed process are examined and their role in the catalysis is discussed.

Acknowledgement

We thank Drs. Wlodawer, Nakayama and Haobo Guo for useful discussions and Professor Martin Karplus for a gift of the CHARMM program. This work was supported by National Science Foundation (Grant number: 0817940 to HG) and NSF TeraGrid resources provided by University of Texas at Austin.

References

Bravaya, K., A. Bochenkova, B. Grigorenko, I. Topol, S. Burt and A. Nemukhin (2006). "Molecular modeling the reaction mechanism of serine-carboxyl peptidases." Journal of Chemical Theory and Computation **2**(4): 1168-1175.

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus (1983). "CHARMM - A PROGRAM FOR MACROMOLECULAR ENERGY, MINIMIZATION, AND DYNAMICS CALCULATIONS." Journal of Computational Chemistry **4**(2): 187-217.

Brooks, C. L., A. Brunger and M. Karplus (1985). "ACTIVE-SITE DYNAMICS IN PROTEIN MOLECULES - A STOCHASTIC BOUNDARY MOLECULAR-DYNAMICS APPROACH." Biopolymers **24**(5): 843-865.

Cui, Q., M. Elstner, E. Kaxiras, T. Frauenheim and M. Karplus (2001). "A QM/MM implementation of the self-consistent charge density functional tight binding (SCC-DFTB) method." Journal of Physical Chemistry B **105**(2): 569-585.

Elstner, M., Q. Cui, P. Munih, E. Kaxiras, T. Frauenheim and M. Karplus (2003). "Modeling zinc in biomolecules with the self consistent charge-density functional tight binding (SCC-DFTB) method: Applications to structural and energetic analysis." Journal of Computational Chemistry **24**(5): 565-581.

Field, M. J., P. A. Bash and M. Karplus (1990). "A COMBINED QUANTUM-MECHANICAL AND MOLECULAR MECHANICAL POTENTIAL FOR MOLECULAR-DYNAMICS SIMULATIONS." Journal of Computational Chemistry **11**(6): 700-733.

Guo, H. B., R. F. Beahm and H. Guo (2004). "Stabilization and destabilization of the C-delta-h center dot center dot center dot O=C hydrogen bonds involving proline residues in helices." Journal of Physical Chemistry B **108**(46): 18065-18072.

Guo, H. B., N. Rao, Q. Xu and H. Guo (2005). "Origin of tight binding of a near-perfect transition-state analogue by cytidine deaminase: Implications for enzyme catalysis." Journal of the American Chemical Society **127**(9): 3191-3197.

Guo, H. B., A. Wlodawer and H. Guo (2005). "A general acid-base mechanism for the stabilization of a tetrahedral adduct in a serine-carboxyl peptidase: A computational study." Journal of the American Chemical Society **127**(45): 15662-15663.

Guo, H. B., A. Wlodawer, T. Nakayama, Q. Xu and H. Guo (2006). "Catalytic role of proton transfers in the formation of a tetrahedral adduct in a serine carboxyl peptidase." Biochemistry **45**(30): 9129-9137.

Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein (1983). "COMPARISON OF SIMPLE POTENTIAL FUNCTIONS FOR SIMULATING LIQUID WATER." Journal of Chemical Physics **79**(2): 926-935.

Kumar, S., D. Bouzida, R. H. Swendsen, P. A. Kollman and J. M. Rosenberg (1992). "THE WEIGHTED HISTOGRAM ANALYSIS METHOD FOR FREE-ENERGY CALCULATIONS ON BIOMOLECULES .1. THE METHOD." Journal of Computational Chemistry **13**(8): 1011-1021.

MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C.

Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus (1998). "All-atom empirical potential for molecular modeling and dynamics studies of proteins." Journal of Physical Chemistry B **102**(18): 3586-3616.

Neria, E., S. Fischer and M. Karplus (1996). "Simulation of activation free energies in molecular systems." Journal of Chemical Physics **105**(5): 1902-1921.

Oda, K. (2012). "New families of carboxyl peptidases: serine-carboxyl peptidases and glutamic peptidases." J. Biochem. **151** (1): 13-25.

Okubo, A., M. Li, M. Ashida, H. Oyama, A. Gustchina, K. Oda, B. M. Dunn, A. Wlodawer and T. Nakayama (2006). "Processing, catalytic activity and crystal structures of kumamolisin-As with an engineered active site." Febs Journal **273**(11): 2563-2576.

Ryckaert, J. P., G. Ciccotti and H. J. C. Berendsen (1977). "NUMERICAL-INTEGRATION OF CARTESIAN EQUATIONS OF MOTION OF A SYSTEM WITH CONSTRAINTS - MOLECULAR-DYNAMICS OF N-ALKANES." Journal of Computational Physics **23**(3): 327-341.

Sleat, D. E., R. J. Donnelly, H. Lackland, C. G. Liu, I. Sohar, R. K. Pullarkat and P. Lobel (1997). "Association of mutations in a lysosomal protein with classical late-infantile neuronal ceroid lipofuscinosis." Science **277**(5333): 1802-1805.

Sun, J., Y. Li and Z. Zhao (2007). "Phylogenetic profiles for the prediction of protein–protein interactions: How to select reference organisms?" Biochemical and Biophysical Research Communications **353**(4): 985-991.

Torrie, G. M. and J. P. Valleau (1974). "MONTE-CARLO FREE-ENERGY ESTIMATES USING NON-BOLTZMANN SAMPLING - APPLICATION TO SUBCRITICAL LENNARD-JONES FLUID." Chemical Physics Letters **28**(4): 578-581.

Wlodawer, A., M. Li, Z. Dauter, A. Gustchina, K. Uchida, H. Oyama, B. M. Dunn and K. Oda (2001). "Carboxyl proteinase from Pseudomonas defines a novel family of subtilisin-like enzymes." Nature Structural Biology **8**(5): 442-446.

Wlodawer, A., M. Li, A. Gustchina, Z. Dauter, K. Uchida, H. Oyama, N. E. Goldfarb, B. M. Dunn and K. Oda (2001). "Inhibitor complexes of the Pseudomonas serine-carboxyl proteinase." Biochemistry **40**(51): 15602-15611.

Wlodawer, A., M. Li, A. Gustchina, H. Oyama, B. M. Dunn and K. Oda (2003). "Structural and enzymatic properties of the sedolisin family of serine-carboxyl peptidases." Acta biochimica Polonica **50**(1): 81-102.

Wlodawer, A., M. Li, A. Gustchina, H. Oyama, K. Oda, B. B. Beyer, J. Clemente and B. M. Dunn (2004). "Two inhibitor molecules bound in the active site of Pseudomonas sedolisin: a model for the bi-product complex following cleavage of a peptide substrate." Biochemical and Biophysical Research Communications **314**(2): 638-645.

Xu, Q. and H. Guo (2004). "Quantum mechanical/molecular mechanical molecular dynamics simulations of cytidine deaminase: From stabilization of transition state analogues to catalytic mechanisms." Journal of Physical Chemistry B **108**(7): 2477-2483.

Xu, Q., H. Guo and A. Wlodawer (2006). "The importance of dynamics in substrate-assisted catalysis and specificity." Journal of the American Chemical Society **128**(18): 5994-5995.

Xu, Q., H. B. Guo, A. Gorin and H. Guo (2007). "Stabilization of a transition-state analogue at the active site of yeast cytosine deaminase: Importance of proton transfers." Journal of Physical Chemistry B **111**(23): 6501-6506.

Xu, Q., L. Y. Li and H. Guo (2010). "Understanding the Mechanism of Deacylation Reaction Catalyzed by the Serine Carboxyl Peptidase Kumamolisin-As: Insights from QM/MM Free Energy Simulations." Journal of Physical Chemistry B **114**(32): 10594-10600.

Xu, Q., J. Z. Yao, A. Wlodawer and H. Guo (2011). "Clarification of the Mechanism of Acylation Reaction and Origin of Substrate Specificity of the Serine-Carboxyl Peptidase Sedolisin through QM/MM Free Energy Simulations." Journal of Physical Chemistry B **115**(10): 2470-2476.

Appendix

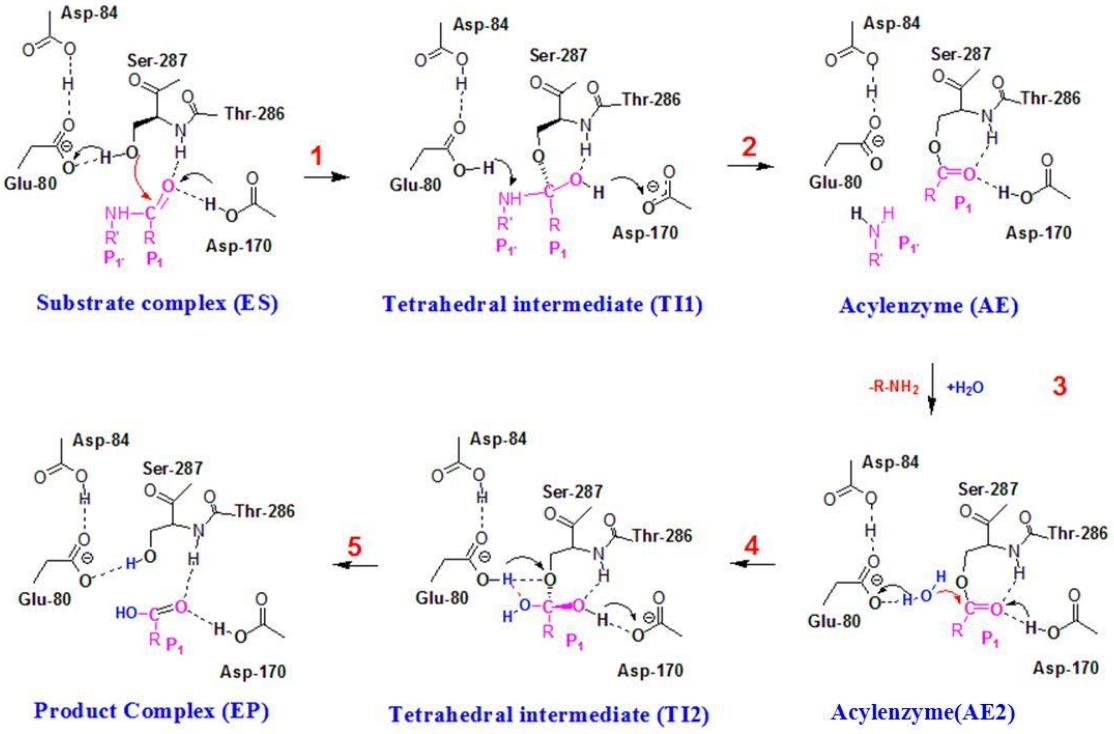


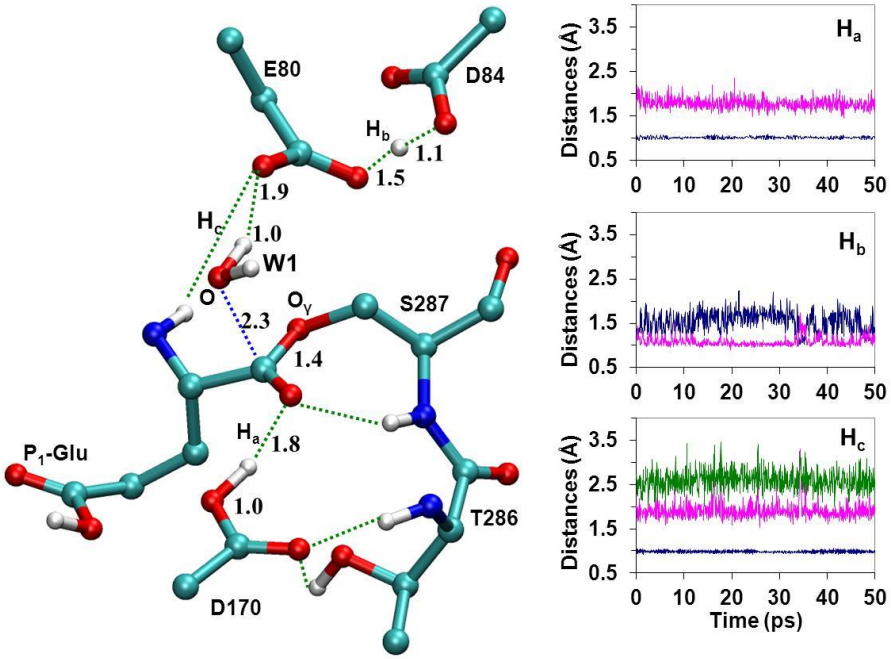
Figure 2.1. The catalytic mechanism and the role of the key active-site residues for the both acylation and deacylation reactions based on this study as well as the previous investigation.

The acyl-enzyme is termed as AE2 to distinguish it with the one (AE) for which the N-terminal region of the substrate still remains at the active site after the cleavage.

Figure 2.2. The average structure of the active site of the acyl-enzyme and the free energy profiles for the deacylation reaction.

(A) Left: The average structure of the active site of the acyl-enzyme that was formed after the acylation step of the reaction involving the LLE*FL substrate as well as release of the cleaved first product of the substrate from the active site (see text). Most of the hydrogen atoms are not shown here for clarity. Right: The distances (Å) from certain oxygen atoms to three hydrogen atoms, H_a, H_b and H_c, as functions of time (ps). These atoms may be involved in proton transfer processes during the deacylation step (see below). Top panel: the distances from O (Asp170) and O (carbonyl group of the P1 residue) to H_a (colored in blue and magenta, respectively); Middle panel: the distances from O (E80) and O (Asp84) to H_b (colored in blue and magenta, respectively); Bottom panel: the distances from O (W1), O (Glu80) and O (Ser287) to H_c (colored in blue, magenta and green, respectively). The locations of the hydrogen atoms, H_a, H_b and H_c, are shown in the structure on the left. (B) The free energy profiles for the deacylation reaction of this work. The reaction coordinate for deacylation is ξ=R(C-Oγ)-R(C-O); see Figure 2.2A for the definition of the atoms. TI2 is the tetrahedral intermediate for deacylation. TSd1 and TSd2 are the first and second transition states, respectively. Wild-type sedolisin: red and solid line; D170N: green dashed line. In addition, the free energy profile for the wild type enzyme with the proton fixed on Asp170 with the SHAKE algorithm (blue dotted line) or with deprotonated Asp170 is also given (orange dotted dash line).

A


E80
D84
H_b
1.1
1.5
1.9
H_c
1.0
W1
O
O_γ
S287
2.3
1.4
H_a
1.8
P_1-Glu
1.0
T286
D170
Distances (Å)
H_a
H_b
H_c
Time (ps)

B

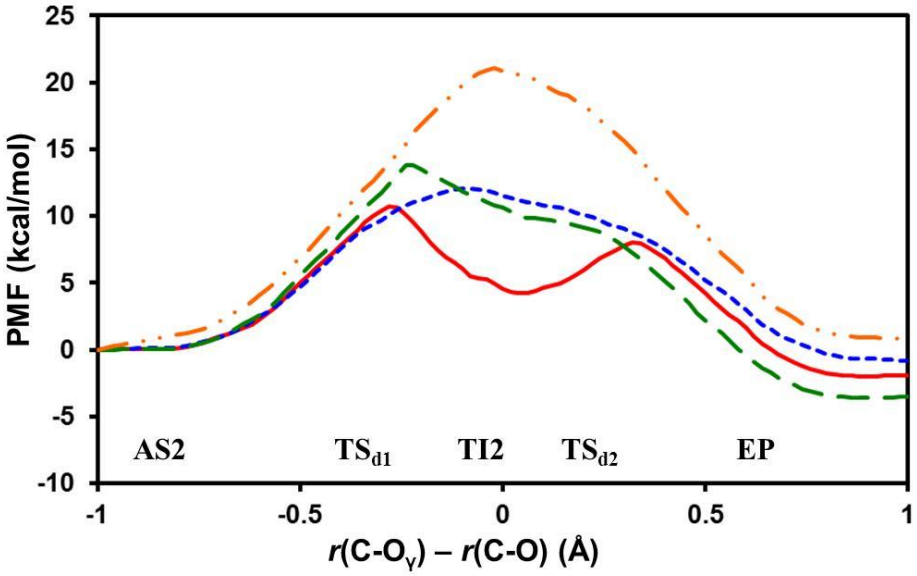
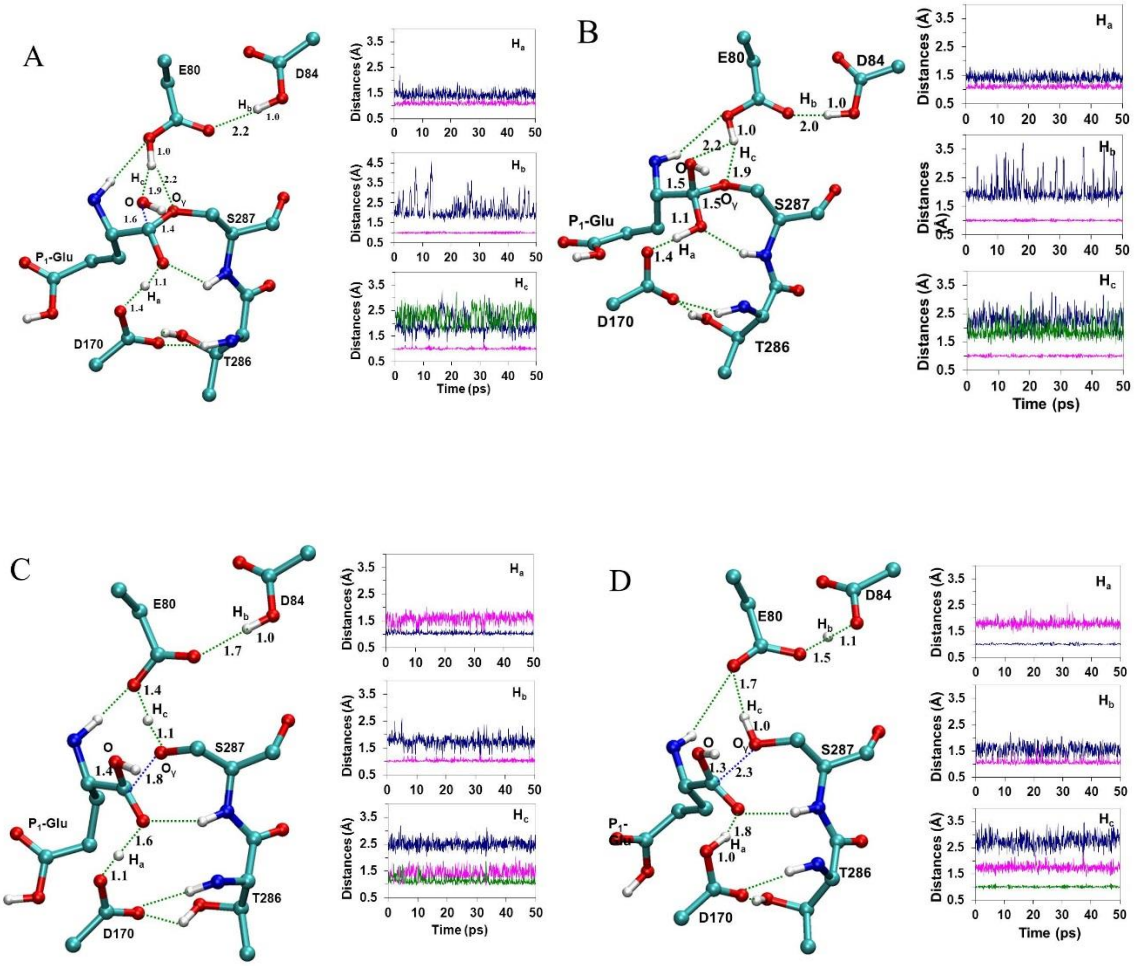

PMF (kcal/mol)
AS2
TS_{d1}
TI2
TS_{d2}
EP
$r(C\text{-}O_\gamma) - r(C\text{-}O)$ (Å)

Figure 2.3 The active-site structures and the distances (Å) to the hydrogen atoms (protons) as functions of time at different stages of the deacylation reaction (see the caption of Figure 2.2A).

(A) The first transition state of the deacylation reaction (TS_{d1}). Top panel: the distances from O (Asp170) and O (carbonyl group of the P_1 residue) to H_a colored in blue and magenta, respectively. Middle panel: the distances from O (Glu80) and O (Asp84) to H_b colored in blue and magenta, respectively. Bottom panel: the distances from O (W1), O (Glu80) and O (Ser287) to H_c colored in blue, magenta and green, respectively. (B) The tetrahedral intermediate (TI2). The color schemes for the distances are the same as in (A). (C) The second transition state (TS_{d2}). The color schemes for the distances are the same as in (A). (D) Product. The color schemes for the distances are the same as in (A).


A
E80
D84
S287
T286
D170
P1-Glu
B
C
D
Distances (Å)
Time (ps)

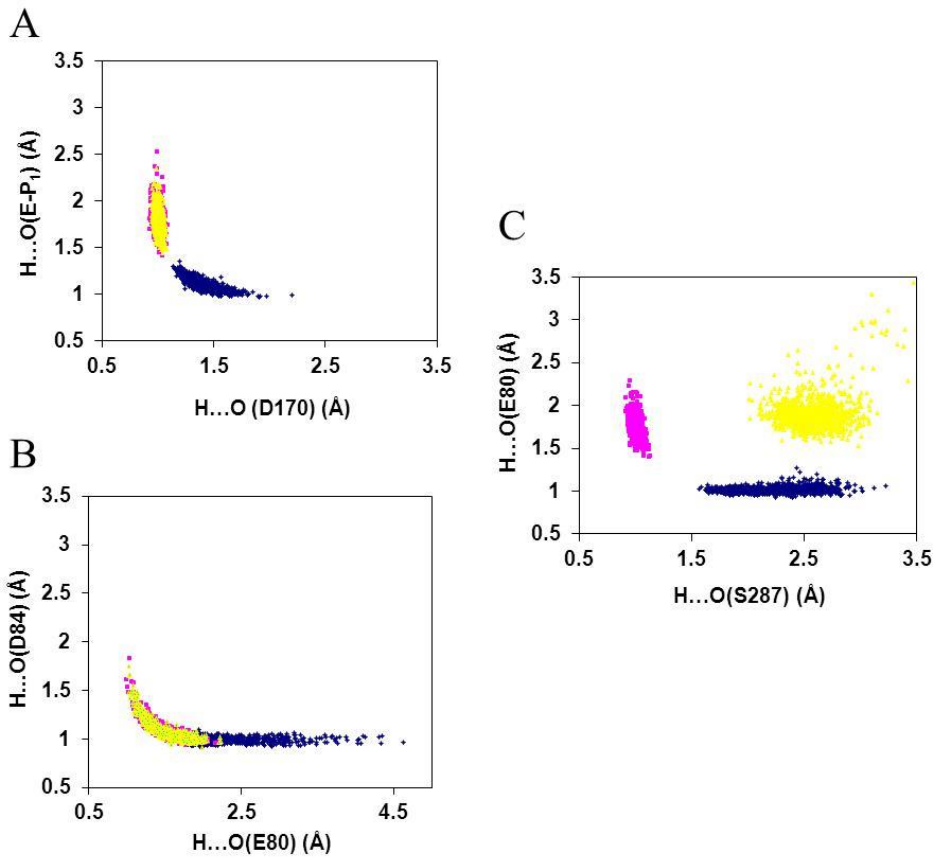


Figure 2.4 Locations and transfers of the protons among AE2 (yellow), TI2 (blue) and EP (magenta) in wild-type.

(A) H_a. H_a is located on Asp170 in both AE2 (yellow) and EP (magenta) states with the H…O (Asp170) distance as about 1 Å. However, this proton moves to the carbonyl oxygen of the residue at P_1 of the substrate when the tetrahedral intermediate (TI2) is formed (blue). (B) H_b. While H_b spends most of the time on Asp84 in TI2 (blue), it is involved in a strong hydrogen bond in AE2 (yellow) and EP (magenta). (C) H_c. H_c on W1 moves to Glu80 as a result of the formation of TI2 (from yellow to blue) and then moves to Ser287 (magenta) during the formation of the product.

CHAPTER 3

QM/MM FREE ENERGY SIMULATIONS OF SALICYLIC ACID METHYLTRANSFERASE: THE EFFECTS OF STABILIZATION OF TS-LIKE STRUCTURES ON SUBSTRATE SPECIFICITY

A version of this chapter was originally published by Jianzhuang Yao, Qin Xu, Feng Chen and Hong Guo:

Jianzhuang Yao, Qin Xu, Feng Chen, and Hong Guo "QM/MM Free Energy Simulations of Salicylic Acid Methyltransferase: the Effects of Stabilization of TS-Like Structures on Substrate Specificity" *The Journal of Physical Chemistry B* 115 (2010): 389-396.

Abstract

Salicylic acid methyltransferases (SAMTs) synthesize methyl salicylate (MeSA) using salicylate as the substrate. MeSA synthesized in plants may function as an airborne signal to activate the expression of defense-related genes and could also be a critical mobile signaling molecule that travels from the site of plant infection to establish systemic immunity in the induction of disease resistance. Here the results of QM/MM free energy simulations for the methyl transfer process in *Clarkia breweri* SAMT (CbSAMT) are reported to determine the origin of the substrate specificity of SAMTs. The free energy barrier for the methyl transfer from S-adenosyl-L-methionine (AdoMet) to 4-hydroxybenzoate in CbSAMT is found to be about 5 kcal/mol higher than that from AdoMet to salicylate, consistent with the experimental observations. It is suggested that the relatively high efficiency for the methylation of salicylate compared to 4-hydroxybenzoate is due, at least in part, to the reason that a part of the stabilization of the transition state (TS) configuration is already reflected in the reactant complex, presumably, through the binding. The results seem to indicate that the creation of the

substrate complex (e.g., through mutagenesis and substrate modifications) with its structure closely resembles TS might be fruitful for improving the catalytic efficiency for some enzymes. The results show that the computer simulations may provide important insights into the origin of the substrate specificity for the SABATH family and could be used to help experimental efforts in generating engineered enzymes with altered substrate specificity.

Introduction

Plants produce a large number of low molecular weight metabolites, and some of these metabolites play critical roles in diverse biological processes including plant growth and development as well as plant interactions with the environment. Chemical modifications of such metabolites by different enzymes may have a profound impact on these activities. One of such modifications is enzymatic methylation of carboxyl moieties of certain small molecules, such as salicylic acid (SA), benzoic acid (BA), indole-3-acetic acid (IAA), gibberellic acid (GA), farnesoic acid (FA), cinnamic/coumaric acid (CC) and jasmonic acid (JA), using S-adenosyl-L-methionine (AdoMet) as the methyl donor.(Taiz and Zeiger 2006) The methylation of carboxyl group of these compounds may affect the concentrations of the free acids in plant tissues, and the corresponding methylated products can have biological and ecological functions that are quite different from their precursors. The enzymes that catalyze the methylation of SA, BA, IAA, GA, FA, CC and JA, for instance, are SA methyltransferase (SAMT),(Ross, Nam et al. 1999, Chen, D'Auria et al. 2003) BA methyltransferase (BAMT),(Murfitt, Kolosova et al. 2000) IAA methyltransferase

(IAMT),(Qin, Gu et al. 2005, Zhao, Guan et al. 2007, Zhao, Ferrer et al. 2008) GA methyltransferase (GAMT),(Varbanova, Yamaguchi et al. 2007) FA methyltransferase (SAMT),(Yang, Yuan et al. 2006) CC methyltransferase (CCMT)(Kapteyn, Qualley et al. 2007) and JA methyltransferase (JAMT),(Seo, Song et al. 2001) respectively. These enzymes belong to the same SABATH enzyme family(D'Auria, Chen et al. 2003) (which was named based on the first three identified members of the family, SAMT, BAMT and theobromine synthase). They can be distinguished according to their substrate preferences and are homologous at the protein sequence level. The SABATH family also contains some other proteins, including the nitrogen-directed methyltransferases involved in caffeine biosynthesis.(Ogawa, Herai et al. 2001, McCarthy and McCarthy 2007) It has been demonstrated that the emergence of novel SABATH MT activity can occur rapidly and small changes in primary protein sequences can lead to the functional emergence of SABATH proteins with altered substrate preferences.(Zubieta, Ross et al. 2003, Pichersky, Noel et al. 2006) Although the SABATH enzymes have been a subject of extensive biochemical and structural investigations,(Ross, Nam et al. 1999, Zubieta, Ross et al. 2003, Pott, Hippauf et al. 2004, Effmert, Saschenbrecker et al. 2005, Hippauf, Michalsky et al. 2010, Tieman, Zeigler et al. 2010, Zhao, Guan et al. 2010) the factors that determine the substrate specificity and the origin for the differences in the breadth of substrate preference of individual enzymes are still not well understood.

SA methyltransferases (SAMTs) are among the best studied members of the SABATH family that synthesize methyl salicylate (MeSA) using SA as the substrate. The closely

related enzymes of the family include BA methyltransferases (BAMTs) which have a higher catalytic efficiency and preference for benzoic acid (BA) lacking the 2-hydroxyl group of SA. It is generally believed that the reactive carboxyl groups of the substrates are fully or predominantly deprotonated at cellular pH values,(Zubieta, Ross et al. 2003) and the methyl transfers should occur without enzyme-mediated general acid/base catalysis (Figure 3.1A). SA is an important plant signal/hormone and can influence a number of processes, including seed germination, seedling establishment, cell growth, respiration, senescence-associated gene expression, responses to abiotic stresses, and fruit yield.(Loake and Grant 2007, Santner and Estelle 2009, Vlot, Dempsey et al. 2009) Moreover, it has been demonstrated that SA may signal the activation of disease resistance following pathogen infection.(Loake and Grant 2007, Santner and Estelle 2009, Vlot, Dempsey et al. 2009) MeSA synthesized by SAMT is a volatile ester and is normally absent in plants. However, MeSA can be dramatically induced upon pathogen infection.(Seskar, Shulaev et al. 1998, Huang, Cardoza et al. 2003) It has been suggested that MeSA may function as an airborne signal to activate the expression of defense-related genes for disease resistance in non-damaged organs of the same or neighboring plants.(Shulaev, Silverman et al. 1997) Recent studies have also shown that MeSA may be a critical mobile signaling molecule that travels from the site of infection to establish systemic immunity in the induction of a long-lasting and broad-spectrum disease resistance,(Park, Kaimoyo et al. 2007) the so-called systemic acquired resistance (SAR).(Sticher, MauchMani et al. 1997) Furthermore, MeSA is a flavor ingredient for taste

and scent of many fruits and flowers and is often produced synthetically and used for making candy, food and medicine. Given the importance of SA and MeSA in plant biology and food industry, understanding the mechanism of the SA methylation catalyzed by SAMT as well as the origin of its substrate specificity is of considerable interest. Such understanding may also provide important insights into the substrate specificity of other members of the SAMT family whose activities may influence diverse biological processes in plants.

Previous experimental studies(Zubieta, Ross et al. 2003, Effmert, Saschenbrecker et al. 2005) have shown that the substrate recognition and catalytic efficiency of SAMTs and other members of the family appear to have subtle biochemical characteristics and their dependences on the active-site and other residues may be complicated. Therefore, applications of different experimental and computational approaches may be required for quantitatively assessing the structural features for substrate recognition and for identifying the amino-acid residues important for the enzyme's function. Computer simulations can provide important insights into the energetic origins of substrate specificity and help to predict the effects of mutations quantitatively, as demonstrated from numerous previous investigations. We have applied molecular dynamics (MD) and free energy (potential of mean force) simulations with hybrid quantum mechanical/molecular mechanical (QM/MM) potentials to study several enzyme-catalyzed reactions earlier, including protein lysine methyltransferases (PKMTs)(Guo and Guo 2007, Xu, Chu et al. 2009, Chu, Xu et al. 2010) which belong to another family of

AdoMet-dependent methyltransferases and may catalyze the methyl transfer processes from AdoMet to the lysine residues on the tails of histone proteins. It was found that there is a good correlation between the activities of the enzymes observed experimentally and the results of the computer simulations,(Guo and Guo 2007, Xu, Chu et al. 2009, Chu, Xu et al. 2010) suggesting the computational approaches are well suited for the investigations of enzyme-catalyzed methyl transfer processes.

Here, we report the results of QM/MM MD and free energy simulations for the methyl transfers in *Clarkia breweri* SAMT (CbSAMT). One of the interesting experimental results is the observation that the catalytic efficiency of SAMTs is considerably lower for 4-hydroxybenzoic acid (4HA) than for SA.(Ross, Nam et al. 1999, Zubieta, Ross et al. 2003, Effmert, Saschenbrecker et al. 2005, Hippauf, Michalsky et al. 2010, Zhao, Guan et al. 2010) This experimental result may also be used as an important testing ground in the determination of usefulness of the computational approaches for understanding substrate specificity for the SAMT family. It is still not clear as to why the change of the hydroxyl group from the 2-position (as in SA) to 4-position (as in 4HA) can lead to such large decrease in the activity. For instance, for CbSAMT the activity on 4HA is only 0-0.8% of that on SA.(Ross, Nam et al. 1999, Zubieta, Ross et al. 2003) The free energy simulations reported in this work show that the free energy barrier for the methyl transfer from AdoMet to 4-hydroxybenzoate in CbSAMT is about 5 kcal/mol higher than that from AdoMet to salicylate, consistent with the experimental observations.(Ross, Nam et al. 1999, Zubieta, Ross et al. 2003) The results show that the structure of the

reactant complex containing salicylate resembles the transition state (TS) of the reaction with the carboxylate of salicylate well aligned with the methyl group of AdoMet. Thus, a part of the stabilization of the TS configuration is already reflected in the reactant complex, presumably, through the binding. This is, however, not the case for 4-hydroxybenzoate. The existence of the 4-hydroxyl group and the increase of the length of the molecule appear to weaken some of the key hydrogen bonds and produce new repulsions between 4-hydroxybenzoate and some of the neighboring residues (e.g., Phe347). This leads to a poor alignment of the reactive groups in the 4-hydroxybenzoate reactant complex that is considerably different from those of the transition state of the 4-hydroxybenzoate methylation reaction and the reactant complex containing salicylate mentioned above. Thus, additional energy is needed for generating the TS configuration during the methyl transfer, leading to a higher free energy barrier (lower efficiency). To the best of our knowledge, this is the first time that the QM/MM MD and free energy simulations are used for studying the SASBTH family.

Methods

QM/MM free energy (potential of mean force) simulations were applied to determine free energy profiles for the methyl transfer from AdoMet to the carboxyl group of salicylate and 4-hydroxybenzoate, respectively, and to characterize the active-site dynamics of the reactant complexes using the CHARMM program.(Brooks, Bruccoleri et al. 1983) AdoMet/S-adenosyl-L-homocysteine (AdoHcy) and the substrates were treated by QM and the rest of the system by MM. The link-atom approach(Field, Bash et al. 1990)

as implemented in CHARMM was applied to separate the QM and MM regions. A modified TIP3P water model(Jorgensen, Chandrasekhar et al. 1983) was employed for the solvent. The stochastic boundary molecular dynamics method(Brooks, Brunger et al. 1985) was used for the QM/MM MD and free energy simulations. The system was separated into a reaction zone and a reservoir region, and the reaction zone was further divided into a reaction region and a buffer region. The reaction region was a sphere with radius *r* of 20 Å, and the buffer region extended over 20 Å ≤ *r* ≤ 22 Å. The reference center for partitioning the system was chosen to be the carboxyl oxygen atom (O) of the substrate where the methyl group is transferred to (see Figure 3.1A and 1B). The resulting systems contained around 5600 atoms, including about 400-500 water molecules.

The SCC-DFTB method(Elstner, Porezag et al. 1998, Cui, Elstner et al. 2001) implemented in CHARMM was used for the QM atoms and the all-hydrogen CHARMM potential function (PARAM27)(MacKerell, Bashford et al. 1998) was used for the MM atoms. High level ab initio methods (e.g., B3LYP and MP2) are too time-consuming to be used for MD and free energy simulations. To correct the errors due to the deficiency of SCC-DFTB method, an empirical correction approach introduced in the earlier study of PKMT SET7/9(Guo and Guo 2007) was applied to the free energy curves obtained from the potential of mean force simulations in the present work. This approach has been applied to the studies of PKMT DIM-5(Xu, Chu et al. 2009) and SET8(Chu, Xu et al. 2010) previously. In this approach, the results of the SCC-DFTB and B3LYP/6-31G** methods for the description of the methyl transfer in a small model system involving AdoMet and

salicylate were compared using an energy minimization-based approach. This comparison allowed us to understand the performance of the semi-empirical method in the description of the bond breaking and making for the system under investigation. Some systematic deviations of the SCC-DFTB method in the description of the energetics of the methyl transfer were found from such comparison in the small model system containing AdoMet and salicylate. A simple linear function $\Delta E_{corr} = k_c \times R + E_c$ was derived based on these results and applied to correct the possible errors in the free energy profiles, as was done in the earlier studies of PKMTs.(Guo and Guo 2007, Xu, Chu et al. 2009, Chu, Xu et al. 2010) Here, k_c = 4.4 kcal·mol^{-1}·Å^{-1} and E_c = -5.7 kcal/mol and *R* is the reaction coordinate for the methyl transfer from AdoMet to salicylate (see below).

The initial coordinates for the reactant complexes were based on the crystallographic structure (PDB code: 1M6E) of CbSAMT complexed with salicylate and AdoHcy.(Zubieta, Ross et al. 2003) A methyl group was manually added to AdoHcy to form AdoMet. In addition, for the reactant complex containing 4-hydroxybenzoate, the hydroxyl group at the 2-position of salicylate was removed and a hydroxyl group was manually added to the 4-position of the ring to generate the substrate. The initial structures for the entire stochastic boundary systems were optimized using the steepest descent (SD) and adopted-basis Newton-Raphson (ABNR) methods. The systems were gradually heated from 50.0 to 310.15 K in 50 ps. A 1-fs time step was used for integration of the equation of motion, and the coordinates were saved every 50 fs for analyses. 1.5 ns QM/MM MD simulations were carried out for each of the reactant complexes. The umbrella sampling

method(Torrie and Valleau 1974) implemented in the CHARMM program along with the Weighted Histogram Analysis Method (WHAM)(Kumar, Bouzida et al. 1992) was then applied to determine the changes of the free energy (potential of mean force) as a function of the reaction coordinate for the methyl transfers from AdoMet to the carboxylate groups of salicylate and 4-hydroxybenzoate, respectively. The reaction coordinate was defined as a linear combination of $r(C_M\text{-}O)$ and $r(C_M\text{-}S_\delta)$ [$R = r(C_M\text{-}S_\delta) - r(C_M\text{-}O)$] (see Figure 3.1B for the definitions of the atoms). The determination of multidimensional free energy maps would be too time-consuming. Our previous study(Guo, Wlodawer et al. 2006) indicated that the one dimensional free energy simulations with the selection of a suitable reaction coordinate reflecting the key bond-breaking making events may be able to capture the key energetic properties for the reaction (e.g., the free energy barrier). For each methyl transfer process, twenty windows were used, and for each window 100 ps simulations were performed with 50 ps equilibration. The force constants of the harmonic biasing potentials used in the PMF simulations were 50 to 500 kcal mol^{-1} $Å^{-2}$. The statistical errors for the free energy profiles were also estimated and were found to be quite small (see the Supporting Information).

Results and Discussion

The active-site structures of the reactant complexes for the methyl transfers from AdoMet to salicylate (Figure 3.2B) and from AdoMet to 4-hydroxybenzoate (Figure 3.2C) in CbSAMT obtained from the simulations are compared with that of the X-ray structure

(Figure 3.2A). Figure 3.2B shows that the active-site structure for the salicylate complex generally agrees with the experimental structure shown in Figure 3.2A. For instance, the position of the reactive carboxylate group of salicylate is fixed by the hydrogen bonds from Trp-151 and Gln-25, and these interactions are expected to be important for an efficient methyl transfer. Comparison of Figure 3.2A and 2B shows that there are some deviations in the positions of the residues and ligand in the two structures. These differences seem to be reasonable due in part to the fact that the experimental structure is not the reactant complex and contains AdoHcy rather than AdoMet. The existence of the methyl group as well as the positive charge on AdoMet is expected to produce some changes in the interactions which could lead to certain modifications in the active site structure. It is interesting to note from Figure 3.2B that the conserved Lys-10 residue interacts with the carboxylate of salicylate instead of the carbonyl group of Met-9 as observed in the X-ray structure.(Zubieta, Ross et al. 2003) This seems to suggest that Lys-10 might play a role in the substrate binding, although a detailed investigation is still necessary. The free energy simulations (see below) suggest that Lys-10 is likely to change to the position observed in the X-ray structure and form a hydrogen bond with the carbonyl group of Met-9 during the methyl transfer from AdoMet to salicylate. The result from the simulations is therefore not inconsistent with the X-ray structure. Figure 3.2B also shows that the methyl group of AdoMet is well aligned with the lone pair of electrons of the carboxylate of salicylate (see also the top of Figure 3.2D), in agreement with the suggestion(Takusagawa, Fujioka et al. 1998, Zubieta, Ross et al. 2003) that SAMT relies on

the proper positioning of the SA substrate's ionized and desolvated carboxyl group near the reactive methyl group of SAM to facilitate methyl transfer. Moreover, it has been shown from the previous computational studies(Guo and Guo 2007, Xu, Chu et al. 2009, Chu, Xu et al. 2010) on PKMTs that a good alignment between the transferable methyl group and the lone pair of electrons (where the methyl group is transferred to) is an important factor for an efficient methyl transfer for the PKMT enzymes as well. The results of the simulations thus suggest that the requirement for a good alignment of the reactive groups seems to be satisfied for the reactant SAMT complex containing salicylate, consistent with the substrate specificity of this enzyme. As will be discussed below, this reactant structure with the well aligned reactive groups resembles the transition state configuration for the methyl transfer so that less free energy would be required for the system to generate such configuration and to reach the transition state.

It should be pointed out that a proper balance of different interactions (e.g., the hydrogen bonds involving Trp-151 and Gln-25 and hydrophobic interactions involving Phe-347, Tyr-255, Trp226 and Met-308) is likely to be crucial for producing the good alignment between the reactive groups from the methyl donor and acceptor in the salicylate complex. The change in the substrate structure or replacement of some key residues may destroy such balance and lead to a poor alignment. The structures in Figure 3.2A and 2B show that the active site of the enzyme seems to be rather crowded and does not have enough space to accommodate a hydroxyl group at 4-position of the benzoate ring without significant structural modifications. Consistent with this observation, the move

of the hydroxyl group from 2- to 4-position in the substrate in going from salicylate to 4-hydroxybenzoate perturbs the active site interactions and changes the active site structure significantly (Figure 3.2C). For instance, 4-hydroxybenzoate in the reactant complex seems to be pushed towards AdoMet, presumably, by the newly formed interactions of the 4-hydroxyl group with the active site residues (e.g., Phe-347 and Met-308). This movement of the substrate is likely to affect the alignment of the reactive groups. Figure 3.2D compares the relative orientations of AdoMet and the carboxylate for some typical structures of the reactant complexes containing salicylate (Top) and 4-hydroxybenzoate (Bottom) obtained from the MD simulations. As is evident from Figure 3.2D, while a good alignment can be made in the salicylate complex, this is not the case for the 4-hydroxybenzoate complex. The methyl group of AdoMet is pushed away considerably by 4-hydroxybenzoate. The distance between the carboxylate oxygen of 4-hydroxybenzoate and the S_δ atom of AdoMet can reach as small as 2.6Å. As a result, it is less likely that the methyl group would be able to be located between these two atoms for achieving a good alignment with the lone pair of electrons of the carboxylate and generating the TS-like structure (see below for a more detailed discussion). The lack of the proper balance of the interactions as a result of the change of the position of the hydroxyl group is presumably the key reason for the deviation of the reactant structure of the 4-hydroxybenzoate complex from the TS-like structure observed for the salicylate complex.

The free-energy profiles for the methyl transfers from AdeMot to salicylate and 4-hydroxybenzoate in CbSAMT are plotted in Figure 3.3 as a function of the reaction coordinate (*R*). The free energy barriers were calculated to be about 16 and 21 kcal/mol for the salicylate and 4-hydroxybenzoate complexes, respectively. The increase of the free energy barrier from 16 kcal to 21 kcal/mol in going from salicylate to 4-hydroxybenzoate is in a good agreement with the experimental finding that the activity of the enzyme on 4HA is only 0-0.8% of that on SA.(Ross, Nam et al. 1999, Zubieta, Ross et al. 2003) The k_{cat} value for the methylation of SA in CbSAMT was estimated to be 0.092 s^{-1} corresponding to an activation barrier of 18.9 kcal/mol based on the transition state theory. The calculated free energy barrier (16 kcal/mol) is therefore about 3 kcal/mol lower than this estimated experimental activation barrier. The difference is probably related to the use of density functional approaches in this study that could underestimate the reaction barriers. It should be pointed out, however, that the relative free energy barriers from the simulations, as opposed to the absolute barriers, are expected to be more important for the comparison of substrate specificities for different molecules. The relative free energy barriers are less sensitive to the choice of the QM method due to the cancellation of the errors and are presumably more reliable. Figure 3.3 also shows that the increase of the free energy along the free energy profile for the 4-hydroxybenzoate complex occurs well before the system reaches the transition state. Indeed, the free energy difference is already 5 kcal/mol at *R* = -0.3Å and remains almost the same until the

system passes the transition state (*R* = 0.3Å). The detailed explanation for this result will be given below.

The r(O… S_δ) distance as a function of time for Windows 1, 3, and 5 of the free energy simulations of the methyl transfer involving salicylate and 4-hydroxybenzoate are monitored in Figure 3.4A and 4B, respectively. This distance should be at least 4.0-4.5 Å in order to have the transferable methyl group located between O (from the methyl acceptor) and S_δ (from the methyl donor) as required during the methyl transfer (note that because the lone pairs of electrons are on the same plane as the carboxylate group, the methyl group should be preferably located on this plane as well). Figure 3.4A shows that this condition on r(O… S_δ) is satisfied for the salicylate complex for the three windows. It can also be shown that this is generally true for almost all the windows (including those near the transition state) during the methyl transfer involving salicylate; the only exceptions are the last few windows near the product state. By contrast, the r(O… S_δ) distance in the 4-hydroxybenzoate complex remains rather small in general (between 2.5 and 4 Å) in the first two plots (Windows 1 and 3), consistent with the structure in Figure 3.2C. The distance is able to reach relatively large values (e.g., 3.5 to 4 Å) in some instances in the second plot, suggesting 4-hydroxybenzoate and AdoMet move apart in order to accommodate the methyl group between them as the reaction proceeds. It is of interest to note that the two molecules are able to be 4.0-4.5 Å apart (see the third plot in Figure 3.4B, corresponding to Window 5 in the free energy simulations with *R* = -0.3Å) well before the system reaches the transition state of the methyl transfer (~*R* =

0.3Å). It should be pointed out that additional energy is expected to be required for achieving this separation of 4-hydroxybenzoate and AdoMet within the enzyme complex during much of the methyl transfer process. Such energy costs could contribute to the increase of the free energy observed in the free energy profiles in going from salicylate to 4-hydroxybenzoate that covers much of the free energy profile (not just at the transition state), leading to the higher free energy barrier for the methyl transfer involving 4-hydroxybenzoate (see above).

Figure 3.5 plots the structures at *R* = -0.3Å and near the transition state (*R* ~ 0.3Å) for the methyl transfer processes involving salicylate (Figure 3.5A and 5B) and 4-hydroxybenzoate (Figure 3.5C and 5D), respectively. Comparison of Figure 3.2C and Figure 3.5C shows that the local active-site structure seems to have been perturbed during the formation of the TS-like structure, as 4-hydroxybenzoate and AdoMet have to move apart in order to accommodate the methyl group between them (see above). This movement seems to lead to the formation of the hydrogen bond between the carboxylate and Gln25 (Trp151) at *R* = -0.3Å (see Figure 3.5C). These hydrogen bonds are broken again near TS (Figure 3.5D), presumably, due to the neutralization of the negative charge of the carboxylate by the methylene group. It is of interest to note that the relative orientations of the reactive groups (the transferable methyl group and carboxylate) are rather similar for the structures in Figure 3.5 containing salicylate and 4-hydroxybenzoate; e.g., r(C_M…O) and r(S_δ… C_M) are 2.0 and 2.4 Å near the transition state, respectively, in the both Figure 3.5B and 5D. As discussed earlier, the structural features for the corresponding reactant

complexes, on the other hand, are significantly different. Indeed, the relative orientations of the reactive groups in the reactant complex of the salicylate complex (see Figure 3.2B and the top of Figure 3.2D) are rather close to the corresponding TS-like and TS structures in Figure 3.5A and 5B with an average r(O… S_δ) distance of about 4.3 Å. Thus, a part of the TS stabilization is probably already reflected on the reactant state by generation of such TS-like configuration through the binding. By contrast, the relative orientations of the reactive groups in the reactant complex of the 4-hydroxybenzoate complex (Figure 3.2C and the bottom of Figure 3.2D) are significantly distorted from those of the corresponding TS-like and TS structures (Figure 3.5C and 5D), and the average r(O… S_δ) distance is only around 2.8 Å. As discussed earlier, the additional energy for generating the TS-like structure for the 4-hydroxybenzoate complex could contribute to the increase of the free energy along the free energy profile (e.g., ~5 kcal/mol at R = -0.3 to 0.3Å) in going from the salicylate to 4-hydroxybenzoate complex, and this leads to the increase of the free energy barrier for the methyl transfer. It would be of interest to replace certain residues and remove some of the interactions that lead to the distorted reactant structure for the 4-hydroxybenzoate complex. The potential residues to be replaced by smaller residues may include Phe-347, Trp284 and Trp226 that are located near the hydroxyl group of 4-hydroxybenzoate (see Figure 3.5C and 5D). Additional combined experimental and computational investigations are under way to study the activity of wild-type and mutated enzyme on 4-hydroxybenzoate.

Conclusions

One of the interesting observations from the previous experimental investigations(Ross, Nam et al. 1999, Zubieta, Ross et al. 2003, Effmert, Saschenbrecker et al. 2005, Hippauf, Michalsky et al. 2010, Zhao, Guan et al. 2010) is that the catalytic efficiency of SAMTs is considerable lower for 4-hydroxybenzoic acid (4HA) than for salicylic acid (SA). It was not clear as to why the change of the hydroxyl group from the 2-position (as in SA) to 4-position (as in 4HA) could lead to such large decrease in the activity. The free energy simulations reported in this work showed that the free energy barrier for the methyl transfer from AdoMet to 4-hydroxybenzoate in CbSAMT is about 5 kcal/mol higher than that from AdoMet to salicylate, consistent with the experimental observations.(Ross, Nam et al. 1999, Zubieta, Ross et al. 2003) It was found that the structure of the reactant complex containing salicylate is rather close to the corresponding TS structure. Thus, a part of the TS stabilization is probably already reflected in the reactant state by generation of such TS-like configuration through the binding. By contrast, the structure of the reactant complex containing 4-hydroxybenzoate is significantly distorted from the corresponding TS structure. Thus, additional energy seems to be required to generate the TS-like structure during the methyl transfer process. This leads to an up-shift of the free energy profile and a higher free energy barrier (lower efficiency) for the methyl transfer. The results seem to indicate that for a given substrate the efficiency of the catalysis might depend on the energy cost for generating the TS-like structure at the active site during the reaction, although other factors may be involved as well. Thus, the creation of the

reactant structure (e.g., through mutagenesis and substrate modifications) that requires less energy to generate the TS-like structure in the active site might be fruitful for improving the catalytic efficiency for some enzymes, although stabilization of the charge formation may be important as well (e.g., as observed in the case of cytidine deaminase(Guo, Rao et al. 2005, Xu, Guo et al. 2007)). The results presented here suggest that the computer simulations may provide important insights into the origin of the substrate specificity for the SABATH family and could be used to help experimental efforts in generating engineered enzymes with altered substrate specificity.

Acknowledgment

We thank Professor Martin Karplus for a gift of the CHARMM program. This work was supported by the National Science Foundation (Grant number: 0817940 to HG), the DOE Office of Biological and Environmental Research - Genome to Life Program through the BioEnergy Science Center (BESC) (to FC) and the Sun Grant Initiative (to FC). We are also grateful for the computer resources (Newton) from University of Tennessee Knoxville.

References

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus (1983). "CHARMM - A PROGRAM FOR MACROMOLECULAR ENERGY, MINIMIZATION, AND DYNAMICS CALCULATIONS." Journal of Computational Chemistry **4**(2): 187-217.

Brooks, C. L., A. Brunger and M. Karplus (1985). "ACTIVE-SITE DYNAMICS IN PROTEIN MOLECULES - A STOCHASTIC BOUNDARY MOLECULAR-DYNAMICS APPROACH." Biopolymers **24**(5): 843-865.

Chen, F., J. C. D'Auria, D. Tholl, J. R. Ross, J. Gershenzon, J. P. Noel and E. Pichersky (2003). "An Arabidopsis thaliana gene for methylsalicylate biosynthesis, identified by a biochemical genomics approach, has a role in defense." Plant Journal **36**(5): 577-588.

Chu, Y. Z., Q. Xu and H. Guo (2010). "Understanding Energetic Origins of Product Specificity of SET8 from QM/MM Free Energy Simulations: What Causes the Stop of Methyl Addition during Histone Lysine Methylation?" Journal of Chemical Theory and Computation **6**(4): 1380-1389.

Cui, Q., M. Elstner, E. Kaxiras, T. Frauenheim and M. Karplus (2001). "A QM/MM implementation of the self-consistent charge density functional tight binding (SCC-DFTB) method." Journal of Physical Chemistry B **105**(2): 569-585.

D'Auria, J. C., F. Chen and E. Pichersky (2003). The SABATH family of MTs in Arabidopsis thaliana and other plant species. Recent Adv Phytochem. **37:** 95–125.

Effmert, U., S. Saschenbrecker, J. Ross, F. Negre, C. M. Fraser, J. P. Noel, N. Dudareva and B. Piechulla (2005). "Floral benzenoid carboxyl methyltransferases: From in vitro to in planta function." Phytochemistry **66**(11): 1211-1230.

Elstner, M., D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai and G. Seifert (1998). "Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties." Physical Review B **58**(11): 7260-7268.

Field, M. J., P. A. Bash and M. Karplus (1990). "A COMBINED QUANTUM-MECHANICAL AND MOLECULAR MECHANICAL POTENTIAL FOR MOLECULAR-DYNAMICS SIMULATIONS." Journal of Computational Chemistry **11**(6): 700-733.

Guo, H. B. and H. Guo (2007). "Mechanism of histone methylation catalyzed by protein lysine methyltransferase SET7/9 and origin of product specificity." Proceedings of the National Academy of Sciences of the United States of America **104**(21): 8797-8802.

Guo, H. B., N. Rao, Q. Xu and H. Guo (2005). "Origin of tight binding of a near-perfect transition-state analogue by cytidine deaminase: Implications for enzyme catalysis." Journal of the American Chemical Society **127**(9): 3191-3197.

Guo, H. B., A. Wlodawer, T. Nakayama, Q. Xu and H. Guo (2006). "Catalytic role of proton transfers in the formation of a tetrahedral adduct in a serine carboxyl peptidase." Biochemistry **45**(30): 9129-9137.

Hippauf, F., E. Michalsky, R. Q. Huang, R. Preissner, T. J. Barkman and B. Piechulla (2010). "Enzymatic, expression and structural divergences among carboxyl O-

methyltransferases after gene duplication and speciation in Nicotiana." Plant Molecular Biology **72**(3): 311-330.

Huang, J., Y. J. Cardoza, E. A. Schmelz, R. Raina, J. Engelberth and J. H. Tumlinson (2003). "Differential volatile emissions and salicylic acid levels from tobacco plants in response to different strains of Pseudomonas syringae." Planta **217**(5): 767-775.

Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein (1983). "COMPARISON OF SIMPLE POTENTIAL FUNCTIONS FOR SIMULATING LIQUID WATER." Journal of Chemical Physics **79**(2): 926-935.

Kapteyn, J., A. V. Qualley, Z. Z. Xie, E. Fridman, N. Dudareva and D. R. Gang (2007). "Evolution of cinnamate/p-coumarate carboxyl Methyltransferases and their role in the biosynthesis of methylcinnamate." Plant Cell **19**(10): 3212-3229.

Kumar, S., D. Bouzida, R. H. Swendsen, P. A. Kollman and J. M. Rosenberg (1992). "THE WEIGHTED HISTOGRAM ANALYSIS METHOD FOR FREE-ENERGY CALCULATIONS ON BIOMOLECULES .1. THE METHOD." Journal of Computational Chemistry **13**(8): 1011-1021.

Loake, G. and M. Grant (2007). "Salicylic acid in plant defence-the players and protagonists." Current Opinion in Plant Biology **10**(5): 466-472.

MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin

and M. Karplus (1998). "All-atom empirical potential for molecular modeling and dynamics studies of proteins." Journal of Physical Chemistry B **102**(18): 3586-3616.

McCarthy, A. A. and J. G. McCarthy (2007). "The structure of two N-methyltransferases from the caffeine biosynthetic pathway." Plant Physiology **144**(2): 879-889.

Murfitt, L. M., N. Kolosova, C. J. Mann and N. Dudareva (2000). "Purification and characterization of S-adenosyl-L-methionine: Benzoic acid carboxyl methyltransferase, the enzyme responsible for biosynthesis of the volatile ester methyl benzoate in flowers of Antirrhinum majus." Archives of Biochemistry and Biophysics **382**(1): 145-151.

Ogawa, M., Y. Herai, N. Koizumi, T. Kusano and H. Sano (2001). "7-Methylxanthine methyltransferase of coffee plants - Gene isolation and enzymatic properties." Journal of Biological Chemistry **276**(11): 8213-8218.

Park, S. W., E. Kaimoyo, D. Kumar, S. Mosher and D. F. Klessig (2007). "Methyl salicylate is a critical mobile signal for plant systemic acquired resistance." Science **318**(5847): 113-116.

Pichersky, E., J. P. Noel and N. Dudareva (2006). "Biosynthesis of plant volatiles: Nature's diversity and ingenuity." Science **311**(5762): 808-811.

Pott, M. B., F. Hippauf, S. Saschenbrecker, F. Chen, J. Ross, I. Kiefer, A. Slusarenko, J. P. Noel, E. Pichersky, U. Effmert and B. Piechulla (2004). "Biochemical and structural characterization of benzenoid carboxyl methyltransferases involved in floral scent production in Stephanotis floribunda and Nicotiana suaveolens." Plant Physiology **135**(4): 1946-1955.

Qin, G. J., H. Y. Gu, Y. D. Zhao, Z. Q. Ma, G. L. Shi, Y. Yang, E. Pichersky, H. D. Chen, M. H. Liu, Z. L. Chen and L. J. Qu (2005). "An indole-3-acetic acid carboxyl methyltransferase regulates Arabidopsis leaf development." Plant Cell **17**(10): 2693-2704.

Ross, J. R., K. H. Nam, J. C. D'Auria and E. Pichersky (1999). "S-adenosyl-L-methionine : salicylic acid carboxyl methyltransferase, an enzyme involved in floral scent production and plant defense, represents a new class of plant methyltransferases." Archives of Biochemistry and Biophysics **367**(1): 9-16.

Santner, A. and M. Estelle (2009). "Recent advances and emerging trends in plant hormone signalling." Nature **459**(7250): 1071-1078.

Seo, H. S., J. T. Song, J. J. Cheong, Y. H. Lee, Y. W. Lee, I. Hwang, J. S. Lee and Y. D. Choi (2001). "Jasmonic acid carboxyl methyltransferase: A key enzyme for jasmonate-regulated plant responses." Proceedings of the National Academy of Sciences of the United States of America **98**(8): 4788-4793.

Seskar, M., V. Shulaev and I. Raskin (1998). "Endogenous methyl salicylate in pathogen-inoculated tobacco plants." Plant Physiology **116**(1): 387-392.

Shulaev, V., P. Silverman and I. Raskin (1997). "Airborne signalling by methyl salicylate in plant pathogen resistance." Nature **385**(6618): 718-721.

Sticher, L., B. MauchMani and J. P. Metraux (1997). "Systemic acquired resistance." Annual Review of Phytopathology **35**: 235-270.

Taiz, L. and E. Zeiger (2006). Plantphysiology. Sunderland, MA, Sinauer Associates.

Takusagawa, F., M. Fujioka, A. Spies and R. L. Schowen (1998). S-Adenosylmethionine (AdoMet)-Dependent Methyltransferases. Comprehensive Biological Catalysis: A Mechanistic Reference. M. Sinnott, London: Academic Press. **1:** 1-30.

Tieman, D., M. Zeigler, E. Schmelz, M. G. Taylor, S. Rushing, J. B. Jones and H. J. Klee (2010). "Functional analysis of a tomato salicylic acid methyl transferase and its role in synthesis of the flavor volatile methyl salicylate." Plant Journal **62**(1): 113-123.

Torrie, G. M. and J. P. Valleau (1974). "MONTE-CARLO FREE-ENERGY ESTIMATES USING NON-BOLTZMANN SAMPLING - APPLICATION TO SUBCRITICAL LENNARD-JONES FLUID." Chemical Physics Letters **28**(4): 578-581.

Varbanova, M., S. Yamaguchi, Y. Yang, K. McKelvey, A. Hanada, R. Borochov, F. Yu, Y. Jikumaru, J. Ross, D. Cortes, C. J. Ma, J. P. Noel, L. Mander, V. Shulaev, Y. Kamiya, S. Rodermel, D. Weiss and E. Pichersky (2007). "Methylation of gibberellins by Arabidopsis GAMT1 and GAMT2." Plant Cell **19**(1): 32-45.

Vlot, A. C., D. A. Dempsey and D. F. Klessig (2009). "Salicylic Acid, a Multifaceted Hormone to Combat Disease." Annual Review of Phytopathology **47**: 177-206.

Xu, Q., Y. Z. Chu, H. B. Guo, J. C. Smith and H. Guo (2009). "Energy Triplets for Writing Epigenetic Marks: Insights from QM/MM Free-Energy Simulations of Protein Lysine Methyltransferases." Chemistry-a European Journal **15**(46): 12596-12599.

Xu, Q., H. B. Guo, A. Gorin and H. Guo (2007). "Stabilization of a transition-state analogue at the active site of yeast cytosine deaminase: Importance of proton transfers." Journal of Physical Chemistry B **111**(23): 6501-6506.

Yang, Y., J. S. Yuan, J. Ross, J. P. Noel, E. Pichersky and F. Chen (2006). "An Arabidopsis thaliana methyltransferase capable of methylating farnesoic acid." Archives of Biochemistry and Biophysics **448**(1-2): 123-132.

Zhao, N., J. L. Ferrer, J. Ross, J. Guan, Y. Yang, E. Pichersky, J. P. Noel and F. Chen (2008). "Structural, biochemical, and phylogenetic analyses suggest that indole-3-acetic acid methyltransferase is an evolutionarily ancient member of the SABATH family." Plant Physiology **146**(2): 455-467.

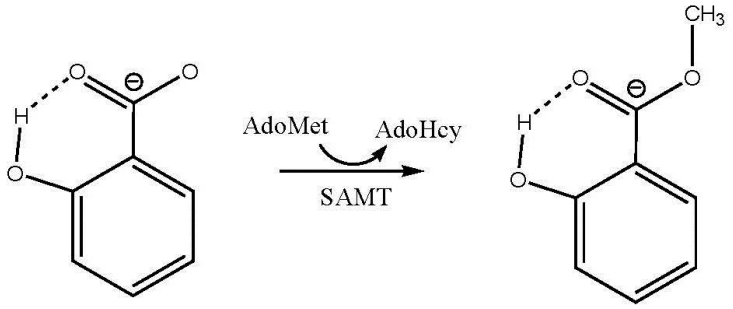
Zhao, N., J. Guan, J. L. Ferrer, N. Engle, M. Chern, P. Ronald, T. J. Tschaplinski and F. Chen (2010). "Biosynthesis and emission of insect-induced methyl salicylate and methyl benzoate from rice." Plant Physiology and Biochemistry **48**(4): 279-287.

Zhao, N., J. Guan, H. Lin and F. Chen (2007). "Molecular cloning and biochemical characterization of indole-3-acetic acid methyltransferase from poplar." Phytochemistry **68**(11): 1537-1544.

Zubieta, C., J. R. Ross, P. Koscheski, Y. Yang, E. Pichersky and J. P. Noel (2003). "Structural basis for substrate recognition in the salicylic acid carboxyl methyltransferase family." Plant Cell **15**(8): 1704-1716.

Appendix

(A)



(B)

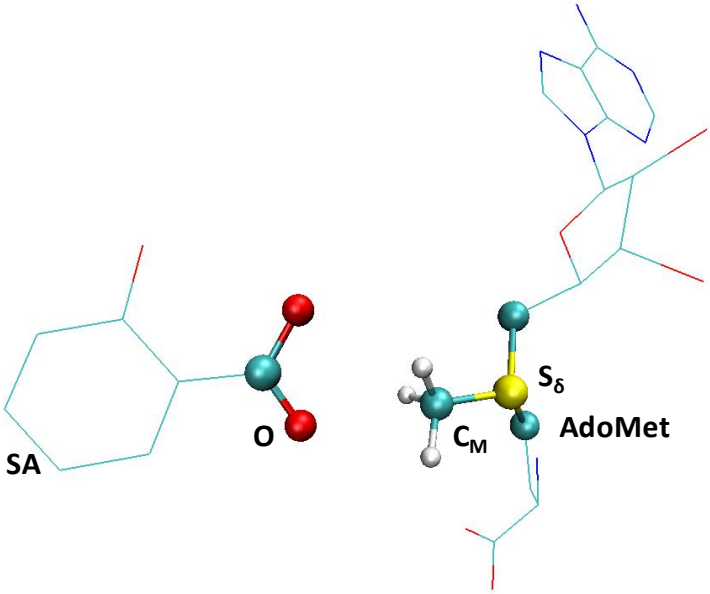


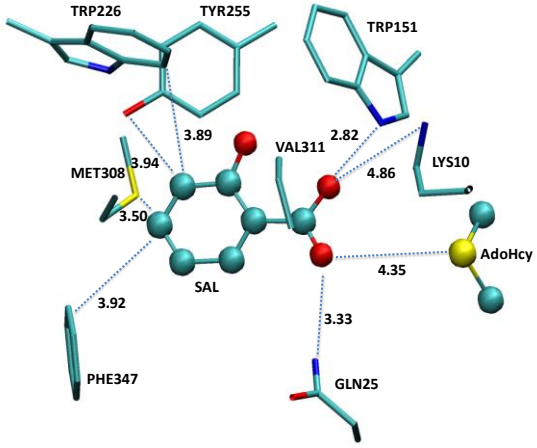
Figure 3.1 Catalytic mechanism and corresponding reaction coordinate.

(**A**) The reaction catalyzed by SAMT. (B) The atoms used for the definition of the reaction coordinate in calculating the free energy profiles of the methyl transfers. The reaction coordinate is: $R = r(C_M \ldots S_\delta) - r(C_M \ldots O)$.

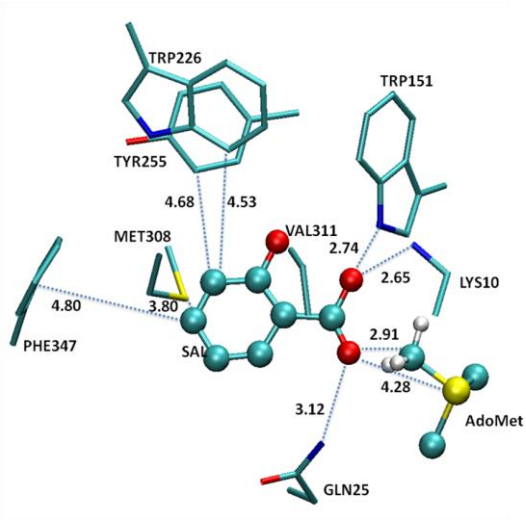
Figure 3.2 Key structures of CbSAMT complex.

(A) The X-ray structure of the active site of CbSAMT complexed with AdoHcy and salicylate. SAMT is shown in sticks, and AdoHcy and salicylate are in balls and sticks. Only the three atoms from AdoHcy and the residues that are close to salicylate are shown for clarity. Some distances are given. **(B)** The structure of the active site of the reactant complex containing AdoMet and salicylate. Some average distances from the simulations are given. **(C)** The structure of the active site of the reactant complex containing AdoMet and 4-hydroxybenzoate along with some average distances obtained from the simulations. **(D)** The relative positions of AdoMet and the substrates in typical structures of the reactant complexes containing salicylate (SAL) (based on Figure 3.2B) and 4-hydroxybenzoate (4BA) (based on Figure 3.2C), respectively. Top: salicylate complex; Bottom: 4-hydroxybenzoate complex.

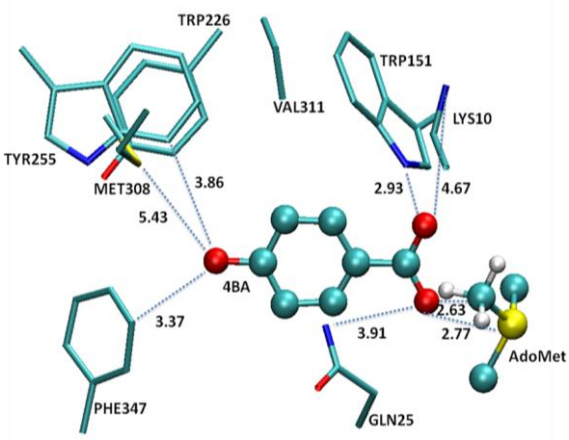
(A)


TRP226
TYR255
TRP151
3.89
VAL311
2.82
MET308
3.94
LYS10
4.86
3.50
AdoHcy
4.35
SAL
3.92
3.33
PHE347
GLN25

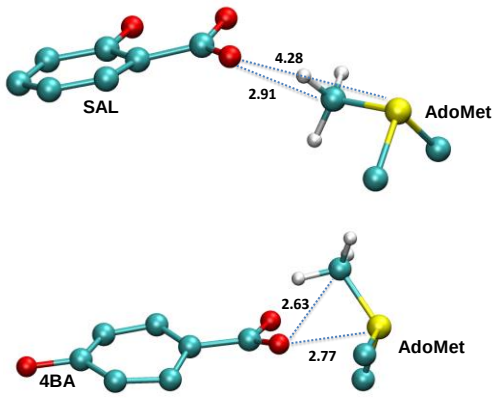
(B)


TRP226
TRP151
TYR255
4.68
4.53
MET308
VAL311
2.74
LYS10
2.65
4.80
3.80
PHE347
SAL
2.91
4.28
AdoMet
3.12
GLN25

(C)


TRP226
TRP151
VAL311
LYS10
TYR255
MET308
3.86
2.93
4.67
5.43
4BA
2.63
3.37
3.91
2.77
AdoMet
PHE347
GLN25

(D)


4.28
SAL
2.91
AdoMet
2.63
AdoMet
2.77
4BA

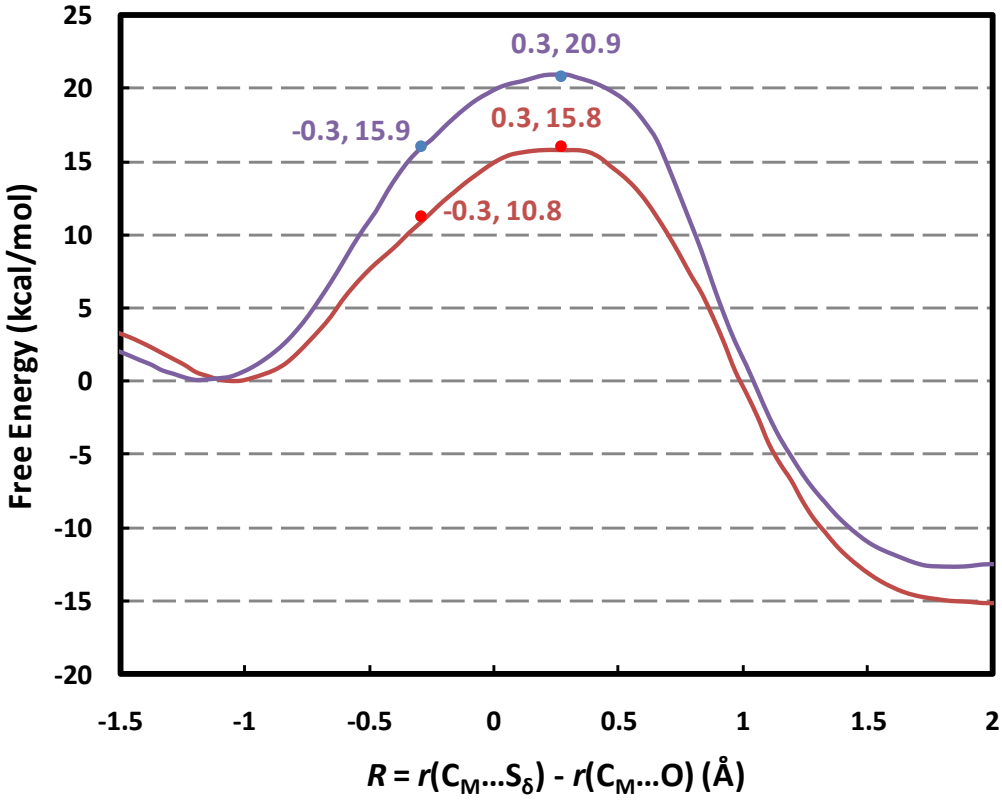


Figure 3.3. Free energy profiles of salicylate and 4-hydroxybenzoate methylations.

Free energy (potential of mean force) changes for the methyl transfers from AdoMet to salicylate and 4-hydroxybenzoate, respectively, as a function of the reaction coordinate [$R = r(C_M...S_\delta) - r(C_M...O)$] in CbSAMT. The methyl transfer involving salicylate: red; The methyl transfer involving 4-hydroxybenzoate: blue. The relative free energies at R = -0.3Å and 0.3Å (near the transition state) are also given.

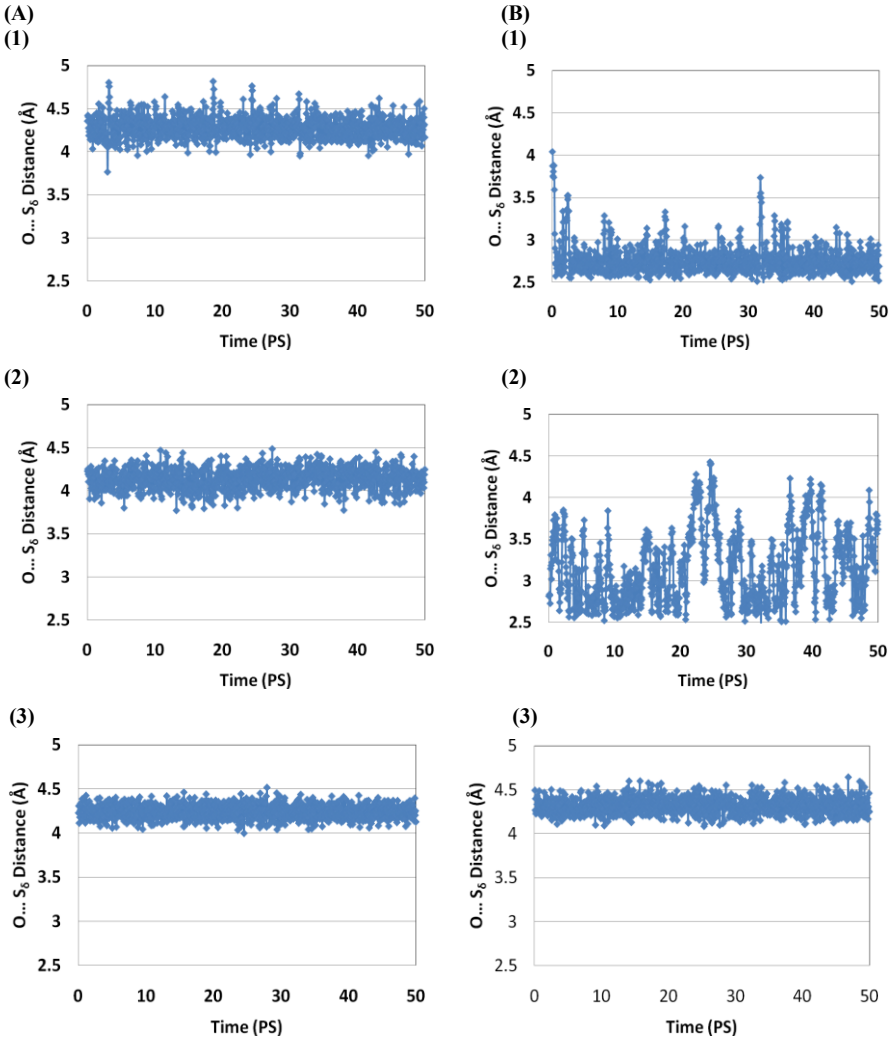


Figure 3.4. Key distance distributions along the reaction coordinate.

(A) The r(O… S_δ) distance as a function of time for some of the windows during the free energy simulations of the methyl transfer involving salicylate. (1) Window 1; (2) Window 3; (3) Window 5. **(B)** The r(O… S_δ) distance as a function of time during the free energy simulations of the methyl transfer involving 4-hydroxybenzoate. (1) Window 1; (2) Window 3; (3) Window 5.

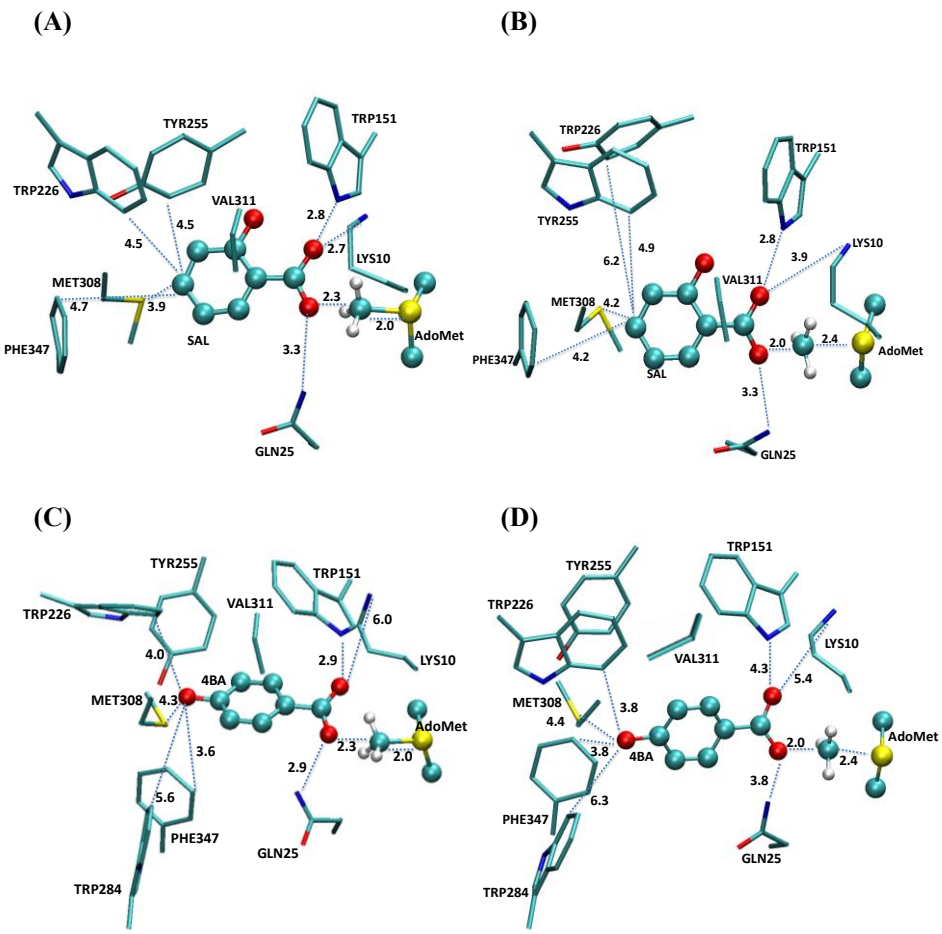


Figure 3.5. The key average structures.

(A) The average structure obtained from Window 5 of the free energy simulations for the methyl transfer involving salicylate (see Figure 3.4A). **(B)** The average structure near transition state for the methyl transfer involving salicylate. **(C)** The average structure obtained from Window 5 of the free energy simulations for the methyl transfer involving 4-hydroxybenzoate (see Figure 3.4B). **(D)** The average structure near transition state for the methyl transfer involving 4-hydroxybenzoate.

CHAPTER 4

UNDERSTANDING PRODUCT SPECIFICITY OF PROTEIN LYSINE METHYLTRANSFERASES FROM QM/MM MD AND FREE ENERGY SIMULATIONS: THE EFFECTS OF MUTATION ON SET7/9 BEYOND THE TYR/PHE SWITCH

A version of this chapter was originally published by Jianzhuang Yao, Yuzhuo Chu, Ran An, and Hong Guo:

Jianzhuang Yao, Yuzhuo Chu, Ran An, and Hong Guo "Understanding Product Specificity of Protein Lysine Methyltransferases from QM/MM MD and Free Energy Simulations: the Effects of Mutation on SET7/9 beyond the Tyr/Phe Switch" *Journal of chemical information and modeling* 52 (2012): 449-456.

Abstract

The results of hybrid quantum mechanical/molecular mechanical (QM/MM) free energy (potential of mean force) simulations for methyl transfer processes in SET7/9 and its Y245A mutant are compared to address the question concerning the change of the product specificity as well as catalytic efficiency due to the mutation. One of the key questions is whether or not the free energy profiles of methyl transfers may be used to predict the change of the product specificity as a result of the mutations for the residues that are not located at the Tyr/Phe switch position. The simulations show that while the wild type SET7/9 is a mono-methylase, the Y245→A mutation increases the ability of the enzyme to add more methyl groups on the target lysine (i.e., acting as a tri-methylase). However, the first methyl transfer process seems to become less efficient in the mutant compared to that in wild-type. All these results are consistent with experimental observations concerning the effects of the mutation on the product specificity and catalytic efficiency. Thus, the previous suggestion that the energetics of the methyl transfer reactions may determine the product specificity, at least in some cases, is

confirmed by the present work. Moreover, the dynamic information of the reactant complexes obtained from the QM/MM MD simulations shows that the ability of the reactant complexes to form the reactive TS-like configurations may be used as an important indicator for the prediction of the product specificity of PKMTs, consistent with previous computational studies.

Introduction

Lysine residues on the tails of histone proteins may be methylated by protein lysine methyltransferases (PKMTs) using S-adenosyl-l-methionine (AdoMet) as the methyl donor.(Martin and Zhang 2005, Jenuwein 2006) This and some other post-translational modifications generate important epigenetic marks of the histone code for regulation of chromatin.(Strahl and Allis 2000) One of the key properties of PKMTs is their product specificities that determine how many methyl groups that the enzyme may "write" on the target lysine.(Xiao, Wilson et al. 2003, Cheng, Collins et al. 2005, Martin and Zhang 2005) Different methylation states produced by different PKMTs can lead to specific patterns of histone modifications and networks of interacting proteins for the regulation of chromatin structure and gene expression.(Turner 2005, Lall 2007, Taverna, Li et al. 2007) Determining the factors that control the product specificity of the enzymes and understanding the reasons that stop the continuation of methylation process (up to a maximum of three methyl groups) are therefore of considerable importance.

The product specificity of PKMTs has been a subject of extensive experimental(Wu, Min et al. , Tachibana, Sugimoto et al. 2001, Zhang, Tamaru et al. 2002, Xiao, Jing et al. 2003,

Zhang, Yang et al. 2003, Collins, Tachibana et al. 2005, Couture, Collazo et al. 2005, Xiao, Jing et al. 2005, Couture, Collazo et al. 2006, Couture, Dirk et al. 2008, Del Rizzo, Couture et al. 2010, Xu, Wu et al. 2011) and computational investigations(Hu and Zhang 2006, Guo and Guo 2007, Wang, Hu et al. 2007, Hu, Wang et al. 2008, Zhang and Bruice 2008, Zhang and Bruice 2008, Xu 2009, Chu, Xu et al. 2010, Georgieva and Himo 2010, Bai, Shen et al. 2011) in the last several years. Previous structural and biochemical studies[5] have identified some active-site structural features that seem to be of importance in controlling the product specificity of the enzymes.(Zhang, Yang et al. 2003, Collins, Tachibana et al. 2005, Couture, Dirk et al. 2008) One important feature is the general occurrence of Tyr/Phe residues at certain positions of the active site.(Xiao, Jing et al. 2003, Zhang, Yang et al. 2003, Collins, Tachibana et al. 2005, Couture, Collazo et al. 2005) It has been found that replacement of some of the Try/Phe residues could lead to changes of product specificity and/or enzyme's catalytic efficiency. For instance, a single Tyr or Phe residue occupies a structurally similar position in the active sites of several PKMTs with this residue normally being Phe for di- and tri-methylases (e.g., Phe-281 in DIM-5(Zhang, Yang et al. 2003)) and Tyr for mono-methylases (e.g., Y305 in SET7/9(Del Rizzo, Couture et al. 2010) and Y334 in SET8(Couture, Collazo et al. 2005, Couture, Dirk et al. 2008)). Interestingly, the tri-methylase DIM-5 can be converted into a mono-/di-methylase by the F281→Y mutation, whereas the mono-methylases SET7/9 and SET8 may be changed to di-methylases through the corresponding Tyr→Phe mutation (e.g., Y305→F and Y334→F mutations for SET7/9 and SET8, respectively).(Zhang, Yang et al. 2003, Couture, Collazo

et al. 2005, Couture, Dirk et al. 2008) In other words, the Tyr→Phe mutation at this position (termed as Tyr/Phe switch position) of the active site tends to increase the ability of the enzyme to add more methyl group(s) to the target lysine (if the residue is Tyr in wild-type), while the Phe→Tyr mutation tends to decrease this ability (if the residue is Phe in wild-type). The similar observations have been made for some other PKMTs.(Wu, Min et al. , Collins, Tachibana et al. 2005) Different suggestions have been proposed to explain how this Tyr/Phe switch residue could affect the product specificity of PKMTs,(Wu, Min et al. , Couture, Dirk et al. 2008, Del Rizzo, Couture et al. 2010) including the proposal that the Tyr/Phe switch may regulate the product specificity through changing the size of the active site or altering the affinity of water molecule(s) at the active site.

Previous investigations have demonstrated that mutations of residues at some other positions of the active sites of PKMTs may also lead to changes of product specificity and/or catalytic efficiency.(Wu, Min et al. , Zhang, Yang et al. 2003, Del Rizzo, Couture et al. 2010) However, unlike the Tyr/Phe switch mentioned above, only limited studies have been performed for determining the role of these residues on the product specificity and for understanding the origin for the effects of these mutations. The enzymes for which such changes of product specificity have been observed include, but not limited to, SET7/9,(Xiao, Jing et al. 2003, Del Rizzo, Couture et al. 2010) G9a methyltransferase and its close relative G9a-like-protein (GLP).(Wu, Min et al. , Collins, Tachibana et al. 2005) For instance, there is a Tyr residue (Y245) hydrogen-bonding to the methyl accepting nitrogen

of the target lysine in the X-ray structures of SET7/9 complexed with substrates.(Xiao, Jing et al. 2003, Del Rizzo, Couture et al. 2010) It has been found that the Y245→A mutation is able to increase the enzyme's ability to add two additional methyl groups to the target lysine (even though Y245 is not really located at the Tyr/Phe switch position), and the Y245A mutant is therefore a tri-methylase.(Xiao, Jing et al. 2003, Del Rizzo, Couture et al. 2010) Nevertheless, the mono-methylation process seems to become significantly less efficient for the Y245A mutant compared to that in wild-type, probably due in part to the reduction of the transition state stabilization for the methyl transfer as a result of this mutation(Guo and Guo 2007) and/or less optimum orientations between the methyl donor and acceptor.(Hu and Zhang 2006)

We and some other groups(Hu and Zhang 2006, Guo and Guo 2007, Wang, Hu et al. 2007, Hu, Wang et al. 2008, Zhang and Bruice 2008, Zhang and Bruice 2008, Xu 2009, Chu, Xu et al. 2010, Georgieva and Himo 2010, Bai, Shen et al. 2011) have studied PKMTs by computer simulations and tried to understand their product specificities as well as changes of the product specificities as a result of certain mutations. The energetic origin for the existence of specific product specificity was examined for some PKMTs and their mutants. It was proposed based on systematic investigations of the energetics for the first, second and third methyl transfers in the enzymes that the reason for the stop of methylation process at certain stage (i.e., before the completion of all the three methyl additions) was probably due to a significant increase of the activation barrier for one of the methyl transfer reactions.(Guo and Guo 2007, Hu, Wang

et al. 2008, Xu 2009, Chu, Xu et al. 2010) Therefore, the relative efficiencies of the chemical steps involving the three methyl transfers from S-adenosyl-l-methionine (AdoMet) to the ε-amino group of the target lysine/methyl lysine in PKMTs may control, at least in some cases, how the epigenetic marks of lysine methylation are written.(Guo and Guo 2007, Xu 2009, Chu, Xu et al. 2010) For instance, for SET8 the simulations showed that the Y334→F mutation changed the first occurrence of an inefficient methyl transfer (i.e., with a relatively high free-energy barrier) from the second methyl transfer to the third methyl transfer.(Chu, Xu et al. 2010) Thus, in contrast to the wild-type enzyme that may not be able to catalyze the second methyl transfer (i.e., a mono-methylase), the Y334F mutant becomes a di-methylase. Interestingly, it has been shown that this principle for understanding product specificity might even be applied to some of the inactive mutants (e.g., the F281W mutant of DIM-5) that have a relatively high activation-barrier for the first methyl transfer.(Xu 2009) However, the question remains as to whether the previous conclusions(Xu 2009, Chu, Xu et al. 2010) could be generalized to the cases involving residues other than that occupying the Tyr/Phe switch position. As a first step of answering this question, the QM/MM free energy simulations are applied to SET7/9 and its Y245A mutant with the TAF10 substrate in this work. The free energy barriers for the methyl transfers obtained from the simulations are found to be well correlated with the experimental observations on the product specificities for these cases, supporting our earlier suggestions(Guo and Guo 2007, Xu 2009, Chu, Xu et al. 2010). Specifically, unlike wild-type the Y245A mutant is found to act like a tri-

methylase based on the results of the simulations. However, the first methyl transfer process is less efficient in Y245A than in wild-type, consistent with the results of previous experimental studies(Xiao, Jing et al. 2003),(Del Rizzo, Couture et al. 2010) and MD simulations.(Hu and Zhang 2006) Furthermore, the QM/MM MD simulations are performed on the reactant complexes of the first, second, and third methyl transfers for each of the systems systematically, and a good correlation is obtained between the ability of the reactant complexes to form the reactive TS-like configurations for the methyl transfers and product specificity.

Methods

QM/MM free energy (potential of mean force) simulations were applied in the determination of the free energy profiles for the first, second, and third methyl transfers in SET7/9 and its Y245A mutant using the models generated from the newly determined X-ray structures containing AdoHcy and unmodified or methylated TAF10 peptide (see below);(Couture, Collazo et al. 2006, Del Rizzo, Couture et al. 2010) the MD simulations were also performed on the reactant complexes of the methyl transfers for each system to study their dynamic properties. AdoMet/AdoHcy and the sidechains of the target lysine/methyl lysine were treated by QM and the rest of the systems were treated by MM. The link-atom approach(Field, Bash et al. 1990) implemented with CHARMM(Brooks, Bruccoleri et al. 1983) was used to separate the QM and MM regions. The reason for this QM and MM partition is to be consistent with our previous studies(Xu 2009, Chu, Xu et al. 2010) which seem to be able to reproduce experimental results. This

may also make it easier for comparison. The stochastic boundary molecular dynamics method(Brooks, Brunger et al. 1985) was used for the QM/MM MD and free energy simulations. The system was separated into a reaction zone and a reservoir zone, and the reaction zone was further divided into a reaction region and a buffer region. The reaction region was a sphere with radius *r* of 22 Å, and the buffer region had *r* equal to 20 Å ≤ *r* ≤ 22 Å. The reference center for partitioning the system was chosen to be the C_δ atom of the target lysine/methyl lysine. The resulting systems contained around 5500 atoms, including about 900 water molecules.

The SCC-DFTB method(Cui, Elstner et al. 2001) implemented in CHARMM(Brooks, Bruccoleri et al. 1983) was used for the QM atoms and the all-hydrogen CHARMM potential function (PARAM27)(MacKerell, Bashford et al. 1998) was used for the MM atoms. A modified TIP3P water model(Jorgensen, Chandrasekhar et al. 1983, Neria, Fischer et al. 1996) was employed for the solvent. In our earlier studies,(Guo and Guo 2007, Xu 2009, Chu, Xu et al. 2010) the results of the SCC-DFTB and B3LYP/6-31G** methods for the description of the methyl transfer in a small model system were compared using an energy minimization-based approach. This comparison allowed us to understand the performance of the semi-empirical method in the description of the bond breaking and making for the methyl transfer and to derive an empirical formula for the correction of the errors in the free energy curves from potential of mean force simulations. It was shown that that the energy curves from the corrected SCC-DFTB and B3LYP/6-31G** were very close,(Guo and Guo 2007, Xu 2009, Chu, Xu et al. 2010)

supporting the use of this approach to make the first-order correction to the bond breaking and making events involving the simple and similar S_N2 methyl transfer processes. The similar approach of this first-order correction to the free energy curves was also adopted in this work.

The initial coordinates of the reactant complexes for the first, second and third methyl transfers were based on the crystallographic complexes of SET7/9 and its Y245A mutant containing AdoHcy and unmodified or methylated TAF10 peptide (PDB codes: 3M53, 3M57, 3M58, 3M59 and 2F69)(Couture, Collazo et al. 2006, Del Rizzo, Couture et al. 2010). In each case, a methyl group was added to AdoHcy to form AdoMet, the methyl donor. For the reactant complexes of the first, second and third methyl transfers in Y245A, the X-ray structures containing K189 (3M57), K189me (3M58) and K189Me2 (3M59) were used, respectively, to generate the corresponding reactant complexes. For the second and third methyl transfers in wild-type, the reactant complexes were generated by replacing one and two hydrogen atoms on the target lysine in the X-ray structure (3M53) by one and two methyl groups, respectively, to generate the corresponding reactant complexes. The initial structures for the entire stochastic boundary system were optimized using the steepest descent (SD) and adopted-basis Newton-Raphson (ABNR) methods. The systems were gradually heated from 50.0 to 298.15 K in 50 ps and equilibrated at 298.15 K for 500 ps. A 1-fs time step was used for integration of the equation of motion, and the coordinates were saved every 50 fs for analyses. After 500-ps of equilibration were performed, 1-ns QM/MM MD simulations were carried out for

each of the reactant complexes of the first, second and third methyl transfers. As discussed in the previous studies,(Guo and Guo 2007, Xu 2009, Chu, Xu et al. 2010) the S_N2 methyl transfer from AdoMet to lysine/methyl lysine is presumably more efficient if the S-CH_3 group of AdoMet is well aligned with the lone pair of electrons on N_ξ in the reactant complex; i.e., with a small ϑ angle and relatively short C_M-N_ξ distance (about 3Å). Therefore, we determined the distributions of $r(C_M\text{-}N_\xi)$ and ϑ from the MD trajectories; ϑ was defined as the angle between the direction of the C_M-S_δ bond ($\boldsymbol{r}_2$) and the direction of the electron lone pair ($\boldsymbol{r}_1$) (see Figure 4.1). Moreover, the histogram method was used in the calculation of the probability density distributions for $r(C_M\text{-}N_\xi)$ and ϑ as well as the relative free energies as a function of $r(C_M\text{-}N_\xi)$.

The umbrella sampling method(Torrie and Valleau 1974) implemented in the CHARMM program along with the Weighted Histogram Analysis Method (WHAM)(Kumar, Bouzida et al. 1992) was applied to determine the change of the free energy (potential of mean force) as a function of the reaction coordinate for the methyl transfer from AdoMet to lysine/methyl lysine in SET7/9 and Y245A. The reaction coordinate was defined as a linear combination of $r(C_M\text{-}N_\xi)$ and $r(C_M\text{-}S_\delta)$ [$R = r(C_M\text{-}S_\delta) - r(C_M\text{-}N_\xi)$] (see Figure 4.1). For each methyl transfer process, twenty windows were used, and for each window 50-ps production runs were performed after 50-ps equilibration. The force constants of the harmonic biasing potentials used in the PMF simulations were from 50 to 400 kcal mol^{-1} $Å^{-2}$.

Results and Discussion

The average active-site structure of the reactant complex and the structure near the transition state for the first methyl transfer in SET7/9 are shown in Figure 4.2A and 2C, respectively. Figure 4.2A shows that the average active-site structure of the reactant complex has the lone pair of electrons on N_ξ of the target lysine well aligned with the methyl group of AdoMet, in agreement with previous MD studies.(Hu and Zhang 2006, Guo and Guo 2007, Wang, Hu et al. 2007, Hu, Wang et al. 2008, Xu 2009, Chu, Xu et al. 2010) These previous simulations already showed that there seemed to be a good correlation between the alignment of the methyl group with the lone pair of electrons and the efficiency of the methyl transfer. Y245 and an active-site water molecule (W1) accept stable hydrogen bonds from the ε-amino group of target lysine, and these interactions may help to orientate the electron lone pair towards the methyl group of AdoMet in the reactant complex. It has been observed in our earlier study(Guo and Guo 2007) that these hydrogen bonds became generally stronger as the system changed from the reactant complex to transition state. The similar observation was also made in the present simulations. Indeed, the average distances for the two hydrogen bonds decrease by 0.2-0.6Å in going from the reactant complex to the transition state. Such strengthening of the hydrogen bonding interactions may play a role in the transition state stabilization and increase the efficiency of the methyl transfer, as proposed previously.(Guo and Guo 2007)

Consistent with the average structure in Figure 4.2A and previous investigations,(Hu and Zhang 2006, Guo and Guo 2007, Wang, Hu et al. 2007, Hu, Wang et al. 2008, Xu 2009, Chu, Xu et al. 2010) the distribution plot (Figure 4.2B, Left) shows that there is a large population of the structures with relative short $r(C_M\text{-}N_\xi)$ distances and small values of the ϑ angle; the average $r(C_M\text{-}N_\xi)$ distance in the reactant complex is around 2.9 Å and the average ϑ angle is less than 30°. Figure 4.2C shows that the $r(C_M\text{-}N_\xi)$ and $r(C_M\text{-}S_\delta)$ distances change to around 2.1 Å and 2.3 Å, respectively, as the system reaches the structure near the transition state for the methyl transfer. The similar structural parameters have also been obtained for the transition states of other methyl transfers, including those for the second and third methyl transfers in SET7/9 and all the three methyl transfers in the Y245A mutant of SET7/9 (see below).

It has been proposed that the CH…O hydrogen bonding between the AdoMet methyl group and oxygen atoms within the SET domain active site may play an important role in AdoMet binding and catalysis.(Couture, Hauk et al. 2006, Horowitz, Yesselman et al. 2011) The interactions involving the C-H group has attracted considerable attention recently in the structural biology and computational chemistry communities due in part to the frequent occurrence of short C-H…O contacts in the structures of proteins, peptides, and DNA (For a review of the C-H…O interactions and the results of high-level ab initio quantum mechanical calculations based on protein structures as well as possible polarization effects,(Guo and Karplus 1994) see Ref. 17).(Guo, Beahm et al. 2004, Guo, Gorin et al. 2009) For SET7/9, one of the possible interactions involves Y335 (see Figure

4.2), a conserved residue.(Couture, Hauk et al. 2006, Horowitz, Yesselman et al. 2011) The results of the simulations show that Y335 is indeed within the accepted carbon-oxygen hydrogen bond distances;(Guo, Beahm et al. 2004) for the first methyl transfer the average distance from oxygen atom of Y335 to carbon atom of the transferable methyl group is about 3.3 Å in the both structures of the reactant complex and transition state.

The average active-site structure of the reactant complex for the second methyl transfer in SET7/9 is shown in Figure 4.2D; the structure near the transition state is plotted in Figure 4.2F. In contrast to the case of the first methyl transfer, the lone pair of electrons on N_ξ of the methylated lysine cannot be well aligned with the methyl group of AdoMet in the reactant complex; the average $r(C_M\text{-}N_\xi)$ distance is 4.7 Å. This fact is further demonstrated by the distribution plot in Figure 4.2E (Left) which shows the $r(C_M\text{-}N_\xi)$ distance is generally long during the MD simulations with a broad distribution of ϑ; the free energy cost for generating the structures with shorter $r(C_M\text{-}N_\xi)$ distance (e.g., 3 Å) more suitable for the reaction is expected to be quite high based on the free energy plot (Figure 4.2E, Right). These results are consistent with previous investigations.(Hu and Zhang 2006, Guo and Guo 2007, Wang, Hu et al. 2007, Hu, Wang et al. 2008, Xu 2009) Figure 4.2D and 2E show that W1 is present at the active site in the both reactant complex and transition state, and previous structural studies seem to suggest that the failure for W1 to dissociate may inhibitor the further methylation of the mono-methyl lysine by wild-type SET7/9 and SET8 (an enzyme that has the similar product specificity as SET7/9).(Couture, Dirk et al. 2008, Del Rizzo, Couture et al. 2010) Consistent with this

suggestion, our earlier QM/MM MD and free energy simulations on SET8(Chu, Xu et al. 2010) showed that while W1 was present at the active site of wild-type SET8 during the second methyl transfer, it moved away from its original position in the Y334F mutant (the Tyr/Phe switch mutation) of SET8 during the second methyl transfer. This seems to make the second methyl transfer more efficient in Y334F and prevent the stop of the methylation process after the first methyl transfer. Comparison of the structures in Figure 4.2A and 2C shows that the structure of the reactant complex for the first methyl transfer in wild-type is rather similar to the corresponding structure near TS. This is in contrast to the case for the second methyl transfer where the reactant structure (Figure 4.2D) is significantly distorted from the corresponding TS structure (Figure 4.2F).The results support the earlier suggestion that one of the reasons for the existence of the relatively low free-energy barrier for the first methyl transfer compared to those for the second and third methyl transfers (see below) is likely owed to the fact that a part of TS stabilization is already reflected on the reactant state of the first methyl transfer through the generation of the TS-like conformation.(Xu 2009, Chu, Xu et al. 2010) For the second methyl transfer in wild-type, the average distance from oxygen atom of Y335 to carbon atom of the transferable methyl group is significantly longer than that for the first methyl transfer (about 4.3 and 3.8 Å for the reactant complex and transition state of the second methyl transfer, respectively).

Figure 4.3A, 3C and 3E show the average structures of the reactant complexes for the first, second and third methyl transfers, respectively, in the Y245A mutant of SET7/9 along with

the distribution plots and free energy analyses; the corresponding structures near the transition states are given in Figure 4.3B, 3D and 3F, respectively. The lone pair of electrons on N_ξ of the target lysine/mono-methyl lysine/di-methyl lysine seems to be able to aligned with the methyl group of AdoMet reasonably well for the first, second and third methyl transfers; the average $r(C_M\text{-}N_\xi)$ distance are around 3.1 Å. The distribution is broad for the first methyl transfer. Nevertheless, the population for the structures with the well aligned lysine and methyl group of AdoMet is still quite high from our simulations, as indicated by the free energy plot in Figure 4.3C (Right). Indeed, the free energy cost for generating structures with $r(C_M\text{-}N_\xi) \sim 3$ Å is negligible. It appears that additional space at the active site generated by the substitution of Y245 by Ala would make it easier for the addition of more methyl groups (see below).

The free energy profiles for the first, second and third methyl transfers in SET7/9 and Y245A are shown in Figure 4.4A and 4B, respectively. As is evident from Figure 4.4A, for wild-type the free energy barriers for the second and third methyl transfers are significantly higher than that for the first methyl transfer (about 3 and 6 kcal/mol higher, respectively). Therefore, addition of the second methyl would probably become more difficult and the methylation process would stop after the first methyl transfer, consistent with the observation that the enzyme is a mono-methylase. It is of interest to note that the difference (about 3 kcal/mol) for the free energy barriers of the first and second methyl transfers is quite similar to an earlier result (about 3.7 kcal/mol) from a quite different QM/MM approach;(Hu, Wang et al. 2008) some of the difference might be due

to the use of different X-rays structures with different substrates. For Y245A, the free energy barriers for all the three methyl transfer steps are not very significantly different. Figure 4.4 shows that the free energy barrier for the first methyl transfer is about 2 kcal/mol higher in Y245A than that in wild-type, and as a result, the methyl transfer process probably becomes slower in the mutant. This is consistent with previous experimental results(Xiao, Jing et al. 2003, Del Rizzo, Couture et al. 2010) which showed that the kcat value for the mono-methylation in Y245A was about 30-fold lower than that in wild-type. The increase of the free energy barrier for the first methyl transfer in going from wild-type to Y245A could be related to the loss of the transition state stabilization involving Y245 (see above), although other factors may be involved as well.(Hu and Zhang 2006, Del Rizzo, Couture et al. 2010) Interestingly, the experimental study also suggested that for Y245A the kcat values for di-methylation and tri-methylation are higher than that for mono-methylation. Our simulations results given in Figure 4.4B show that the barriers for the second and third methyl transfers are indeed 0.7 and 1.3 kcal/mol lower than that for the first methyl transfer. In our earlier studies,(Xu 2009, Chu, Xu et al. 2010) we proposed the use of two free energy triplets, (0, Δ2-1W, Δ3-1W) and (ΔM-W, Δ2-1M, Δ3-1M), to describe the energetics of methyl transfers in PKMTs and their mutants (see Figure 4.4 for explanation). The energy triplets for wild-type and the Y245A mutant can be written as (0, 3, 5.9) and (2.1, -0.7, -1.3), respectively. It should be pointed out that the relative free energy barriers, as opposed to the absolute barriers, are expected to be more reliable and important in the determination of the product specificity. The relative free

energy barriers are expected to be less sensitive to the choice of the QM method and experimental structural differences due to the cancellation of the errors. This is one of the reasons that we proposed the use of two free energy triplets to describe the energetics of methyl transfers in PKMTs and their mutants that are based on the relative free energy barriers.

Conclusions

In this study, the QM/MM MD and free energy simulations have been performed on different methyl transfer steps from AdoMet to target lysine/methyl-lysine in SET7/9 and its Y245A mutant. As far as the product specificity is concerned, one of the key questions is what causes the stop of further methyl addition during histone lysine methylation (which is different from the question concerning the rate-limiting step of the enzyme-catalyzed methyl-transfer process). In the earlier publications,(Guo and Guo 2007, Xu 2009, Chu, Xu et al. 2010) DIM-5, SET7/9 and SET8 as well as their mutants at the Tyr/Phe switch position were studied, and it was shown that the stop of methylation (product specificity) of PKMTs could be identified based on the relative free energy barriers for the three methyl transfers (the energy triplets) as well as the dynamic properties of the reactant complexes. The next question that needs to be addressed is whether or not the free energy profiles and the energetic triplets may be used to understand and predict the changes of the product specificity as a result of the mutations at other positions. This question was addressed in the present study. The free energy barriers for the methyl transfers obtained here were found to be well correlated with the experimental

observations on the change of product specificity due to the Y245→A mutation. Indeed, the simulations showed that while the wild type SET7/9 may act like a mono-methylase, the Y245→A mutation could increase the ability of SET7/9 to add two more methyl groups on the target lysine and lead to a tri-methylase. The simulations also suggested that the first methyl transfer process in Y245A might become less efficient compared to the first methyl transfer in wild-type, consistent with experimental observations and previous MD simulations (see above). The QM/MM MD simulations were also performed on the reactant complexes of the first, second, and third methyl transfers for each of the systems, and a good correlation was obtained between the ability of the reactant complexes to form the reactive configurations for the methyl transfers and product specificity.

Acknowledgement

This work is supported in part by the National Science Foundation Award (Grant number: 0817940 to H. Guo) and the NSF TeraGrid resources provided by University of Texas at Austin. YZC is supported by a NSF NIMBioS fellowship at University of Tennessee Knoxville. We thank Prof. Martin Karplus for a gift of CHARMM program and Prof. Xiaodong Cheng and Drs. Qin Xu and Haobo Guo for useful discussions.

Reference

Bai, Q. F., Y. L. Shen, X. J. Yao, F. Wang, Y. P. Du, Q. Wang, N. Z. Jin, J. Hai, T. J. Hu and J. B. Yang (2011). "Modeling a New Water Channel That Allows SET9 to Dimethylate p53." Plos One **6**(5).

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus (1983). "CHARMM - A PROGRAM FOR MACROMOLECULAR ENERGY, MINIMIZATION, AND DYNAMICS CALCULATIONS." Journal of Computational Chemistry **4**(2): 187-217.

Brooks, C. L., A. Brunger and M. Karplus (1985). "ACTIVE-SITE DYNAMICS IN PROTEIN MOLECULES - A STOCHASTIC BOUNDARY MOLECULAR-DYNAMICS APPROACH." Biopolymers **24**(5): 843-865.

Cheng, X., R. E. Collins and X. Zhang (2005). "Structural and sequence motifs of protein (histone) methylation enzymes." Annu Rev Biophys Biomol Struct **34**: 267-294.

Chu, Y. Z., Q. Xu and H. Guo (2010). "Understanding Energetic Origins of Product Specificity of SET8 from QM/MM Free Energy Simulations: What Causes the Stop of Methyl Addition during Histone Lysine Methylation?" Journal of Chemical Theory and Computation **6**(4): 1380-1389.

Collins, R. E., M. Tachibana, H. Tamaru, K. M. Smith, D. Jia, X. Zhang, E. U. Selker, Y. Shinkai and X. Cheng (2005). "In vitro and in vivo analyses of a Phe/Tyr switch controlling product specificity of histone lysine methyltransferases." J Biol Chem **280**(7): 5563-5570.

Couture, J. F., E. Collazo, J. S. Brunzelle and R. C. Trievel (2005). "Structural and functional analysis of SET8, a histone H4 Lys-20 methyltransferase." Genes & Development **19**(12): 1455-1465.

Couture, J. F., E. Collazo, G. Hauk and R. C. Trievel (2006). "Structural basis for the methylation site specificity of SET7/9." Nature Structural & Molecular Biology **13**(2): 140-146.

Couture, J. F., L. M. A. Dirk, J. S. Brunzelle, R. L. Houtz and R. C. Trievel (2008). "Structural origins for the product specificity of SET domain protein methyltransferases." Proceedings of the National Academy of Sciences of the United States of America **105**(52): 20659-20664.

Couture, J. F., G. Hauk, M. J. Thompson, G. M. Blackburn and R. C. Trievel (2006). "Catalytic roles for carbon-oxygen hydrogen bonding in SET domain lysine methyltransferases." Journal of Biological Chemistry **281**(28): 19280-19287.

Cui, Q., M. Elstner, E. Kaxiras, T. Frauenheim and M. Karplus (2001). "A QM/MM implementation of the self-consistent charge density functional tight binding (SCC-DFTB) method." Journal of Physical Chemistry B **105**(2): 569-585.

Del Rizzo, P. A., J. F. Couture, L. M. A. Dirk, B. S. Strunk, M. S. Roiko, J. S. Brunzelle, R. L. Houtz and R. C. Trievel (2010). "SET7/9 Catalytic Mutants Reveal the Role of Active Site Water Molecules in Lysine Multiple Methylation." Journal of Biological Chemistry **285**(41): 31849-31858.

Field, M. J., P. A. Bash and M. Karplus (1990). "A COMBINED QUANTUM-MECHANICAL AND MOLECULAR MECHANICAL POTENTIAL FOR MOLECULAR-DYNAMICS SIMULATIONS." Journal of Computational Chemistry **11**(6): 700-733.

Georgieva, P. and F. Himo (2010). "Quantum Chemical Modeling of Enzymatic Reactions: The Case of Histone Lysine Methyltransferase." Journal of Computational Chemistry **31**(8): 1707-1714.

Guo, H. and M. Karplus (1994). "SOLVENT INFLUENCE ON THE STABILITY OF THE PEPTIDE HYDROGEN-BOND - A SUPRAMOLECULAR COOPERATIVE EFFECT." Journal of Physical Chemistry **98**(29): 7104-7105.

Guo, H. B., R. F. Beahm and H. Guo (2004). "Stabilization and destabilization of the C-delta-h center dot center dot center dot O=C hydrogen bonds involving proline residues in helices." Journal of Physical Chemistry B **108**(46): 18065-18072.

Guo, H. B., A. Gorin and H. Guo (2009). "A peptide-linkage deletion procedure for estimate of energetic contributions of individual peptide groups in a complex environment: Application to parallel β-Sheets " Interdisciplinary Sciences: Computational Life Sciences **1**(1): 12-20.

Guo, H. B. and H. Guo (2007). "Mechanism of histone methylation catalyzed by protein lysine methyltransferase SET7/9 and origin of product specificity." Proc Natl Acad Sci U S A **104**(21): 8797-8802.

Horowitz, S., J. D. Yesselman, H. M. Al-Hashimi and R. C. Trievel (2011). "Direct Evidence for Methyl Group Coordination by Carbon-Oxygen Hydrogen Bonds in the Lysine Methyltransferase SET7/9." Journal of Biological Chemistry **286**(21): 18658-18663.

Hu, P., S. Wang and Y. Zhang (2008). "How do SET-domain protein lysine methyltransferases achieve the methylation state specificity? Revisited by ab initio QM/MM molecular dynamics simulations." Journal of the American Chemical Society **130**(12): 3806-3813.

Hu, P. and Y. K. Zhang (2006). "Catalytic mechanism and product specificity of the histone lysine methyltransferase SET7/9: An ab initio QM/MM-FE study with multiple initial structures." Journal of the American Chemical Society **128**(4): 1272-1278.

Jenuwein, T. (2006). "The epigenetic magic of histone lysine methylation." FEBS J **273**(14): 3121-3135.

Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein (1983). "COMPARISON OF SIMPLE POTENTIAL FUNCTIONS FOR SIMULATING LIQUID WATER." Journal of Chemical Physics **79**(2): 926-935.

Kumar, S., D. Bouzida, R. H. Swendsen, P. A. Kollman and J. M. Rosenberg (1992). "THE WEIGHTED HISTOGRAM ANALYSIS METHOD FOR FREE-ENERGY CALCULATIONS ON BIOMOLECULES .1. THE METHOD." Journal of Computational Chemistry **13**(8): 1011-1021.

Lall, S. (2007). "Primers on chromatin." Nature Structural & Molecular Biology **14**(11): 1110-1115.

MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus (1998). "All-atom empirical potential for molecular modeling and dynamics studies of proteins." Journal of Physical Chemistry B **102**(18): 3586-3616.

Martin, C. and Y. Zhang (2005). "The diverse functions of histone lysine methylation." Nat Rev Mol Cell Biol **6**(11): 838-849.

Neria, E., S. Fischer and M. Karplus (1996). "Simulation of activation free energies in molecular systems." Journal of Chemical Physics **105**(5): 1902-1921.

Strahl, B. D. and C. D. Allis (2000). "The language of covalent histone modifications." Nature **403**(6765): 41-45.

Tachibana, M., K. Sugimoto, T. Fukushima and Y. Shinkai (2001). "SET domain-containing protein, G9a, is a novel lysine-preferring mammalian histone methyltransferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3." Journal of Biological Chemistry **276**(27): 25309-25317.

Taverna, S. D., H. Li, A. J. Ruthenburg, C. D. Allis and D. J. Patel (2007). "How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers." Nature Structural & Molecular Biology **14**(11): 1025-1040.

Torrie, G. M. and J. P. Valleau (1974). "MONTE-CARLO FREE-ENERGY ESTIMATES USING NON-BOLTZMANN SAMPLING - APPLICATION TO SUBCRITICAL LENNARD-JONES FLUID." Chemical Physics Letters **28**(4): 578-581.

Turner, B. M. (2005). "Reading signals on the nucleosome with a new nomenclature for modified histones." Nat Struct Mol Biol **12**(2): 110-112.

Wang, S. L., P. Hu and Y. K. Zhang (2007). "Ab initio quantum mechanical/molecular mechanical molecular dynamics simulation of enzyme catalysis: The case of histone lysine methyltransferase SET7/9." Journal of Physical Chemistry B **111**(14): 3758-3764.

Wu, H., J. R. Min, V. V. Lunin, T. Antoshenko, L. Dombrovski, H. Zeng, A. Allali-Hassani, V. Campagna-Slater, M. Vedadi, C. H. Arrowsmith, A. N. Plotnikov and M. Schapira "Structural Biology of Human H3K9 Methyltransferases." Plos One **5**(1).

Xiao, B., C. Jing, G. Kelly, P. A. Walker, F. W. Muskett, T. A. Frenkiel, S. R. Martin, K. Sarma, D. Reinberg, S. J. Gamblin and J. R. Wilson (2005). "Specificity and mechanism of the histone methyltransferase Pr-Set7." Genes & Development **19**(12): 1444-1454.

Xiao, B., C. Jing, J. R. Wilson, P. A. Walker, N. Vasisht, G. Kelly, S. Howell, I. A. Taylor, G. M. Blackburn and S. J. Gamblin (2003). "Structure and catalytic mechanism of the human histone methyltransferase SET7/9." Nature **421**(6923): 652-656.

Xiao, B., J. R. Wilson and S. J. Gamblin (2003). "SET domains and histone methylation." Current Opinion in Structural Biology **13**(6): 699-705.

Xu, Q., Chu, Y-Z, Guo, H-B, Smith, J.C., Guo, H. (2009). "Energy Triplets for Writing Epigenetic Marks: Insights from QM/MM Free-Energy Simulations of Protein Lysine Methyltransferases." Chemistry - A European Journal **15**: 12596 – 12599.

Xu, S. T., J. Wu, B. F. Sun, C. Zhong and J. P. Ding (2011). "Structural and biochemical studies of human lysine methyltransferase Smyd3 reveal the important functional roles of its post-SET and TPR domains and the regulation of its activity by DNA binding." Nucleic Acids Research **39**(10): 4438-4449.

Zhang, X. and T. C. Bruice (2008). "Enzymatic mechanism and product specificity of SET-domain protein lysine methyltransferases." Proceedings of the National Academy of Sciences of the United States of America **105**(15): 5728-5732.

Zhang, X., H. Tamaru, S. I. Khan, J. R. Horton, L. J. Keefe, E. U. Selker and X. Cheng (2002). "Structure of the Neurospora SET domain protein DIM-5, a histone H3 lysine methyltransferase." Cell **111**(1): 117-127.

Zhang, X., Z. Yang, S. I. Khan, J. R. Horton, H. Tamaru, E. U. Selker and X. Cheng (2003). "Structural basis for the product specificity of histone lysine methyltransferases." Mol Cell **12**(1): 177-185.

Zhang, X. D. and T. C. Bruice (2008). "Product specificity and mechanism of protein lysine methyltransferases: Insights from the histone lysine methyltransferase SET8." Biochemistry **47**(25): 6671-6677.

Appendix

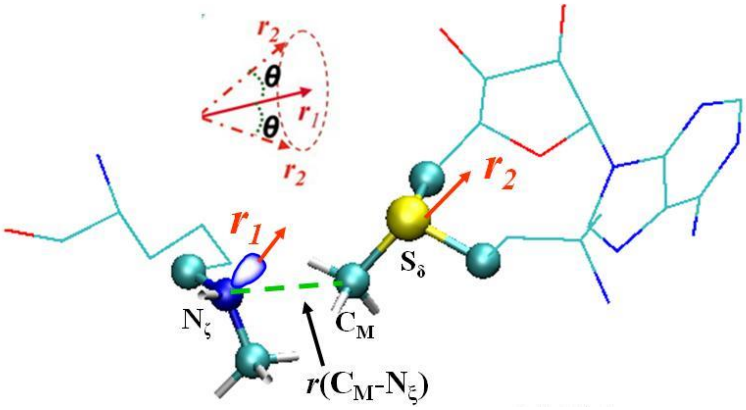


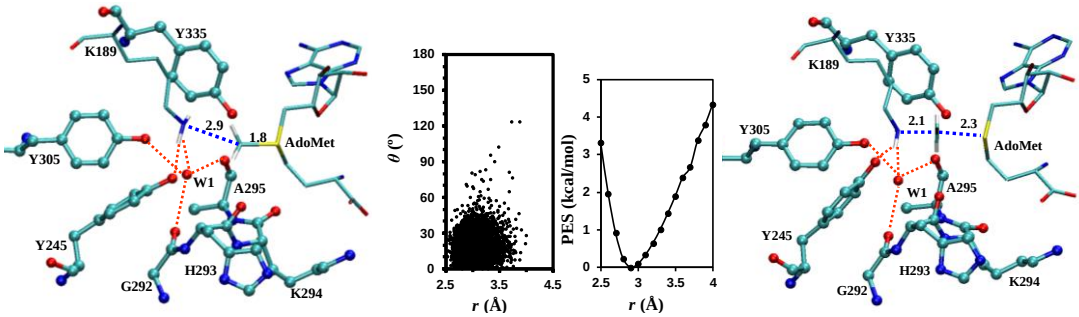
Figure 4.1. The definition of the structural parameters for monitoring the relative orientation of AdoMet and K189me1 [K189 and K189(me)2] in the reactant complex.

The efficiency of the methyl transfer may be related to the distributions of $r(C_M...N_\zeta)$ and θ in the reactant complexes. θ is defined as the angle between the two vectors r1 and r2. Here r1 is the direction of the lone pair of electrons on N_ζ and r2 is the vector pointing from C_M to S_δ. The reaction coordinate for calculating the free energy profiles for the methyl transfers is $R = r(C_M...S_\delta) - r(C_M...N_\zeta)$.

Figure 4.2. MD results for the wild-type enzyme (SET7/9) with TAF10-K189 peptide as the substrate.

(A) The average active-site structure of the reactant complex for the first methyl transfer. SET7/9 is shown in balls and sticks, and AdoMet and TAF10-K189 are in sticks. Hydrogen atoms are not shown for clarity except for those on N_ζ and transferable methyl group. Hydrogen bonds are indicated by red dotted lines, and the distances related to the reactant coordinate are also shown. (B) Two-dimensional plot of $r(C_M…N_\zeta)$ and θ distributions based on the 1-ns simulations of the reactant complex for the first methyl transfer along with the corresponding free energy change as a function of $r(C_M...N_\xi)$ obtained from the distributions. (C) The average structure near the transition state for the first methyl transfer obtained from the free energy (potential of mean force) simulations. (D) The average active-site structure of the reactant complex for the second methyl transfer. (E) The two-dimensional plot of $r(C_M…N_\zeta)$ and θ distributions based on the 1 ns simulations of the reactant complex for the second methyl transfer along with the corresponding free energy change as a function of $r(C_M...N_\xi)$. (F) The average structure near the transition state for the second methyl transfer obtained from the free energy simulations.

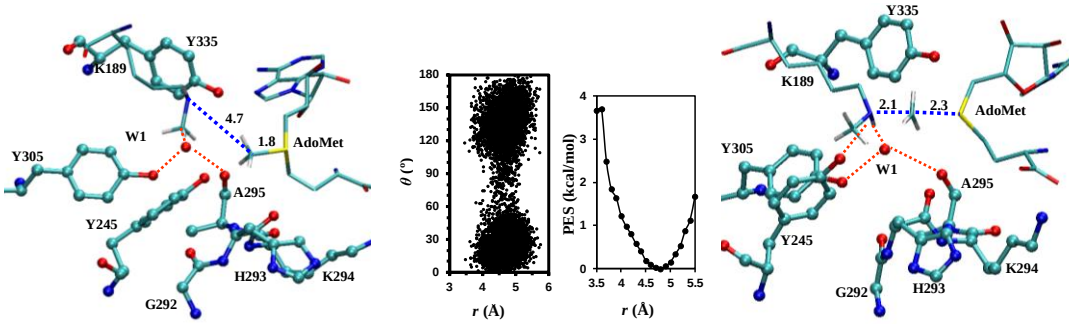
(A) **(B)** **(C)**



Reactant, 1st methyl transfer

TS, 1st methyl transfer

(D) **(E)** **(F)**



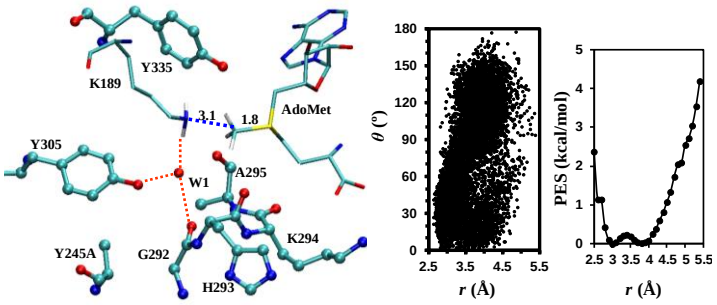
Reactant, 2nd methyl transfer

TS, 2nd methyl transfer

Figure 4.3. MD results for Y245A.

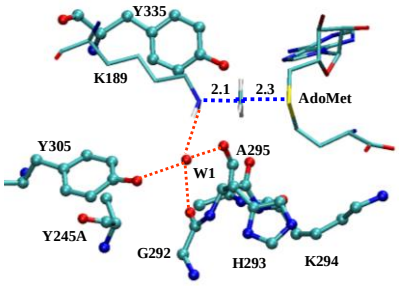
(A) Left: The average active-site structure of the reactant complex for the first methyl transfer in the mutant. Right: the two-dimensional plot of r(C_M…N_ζ) and θ distributions based on the 1 ns simulations of the reactant complex for the first methyl transfer along with the corresponding free energy change as a function of r(C_M...N_ξ) obtained from the distributions. (B) The average structure near the transition state for the first methyl transfer obtained from the free energy simulations. (C) Left: The average active-site structure of the reactant complex for the second methyl transfer. Right: the two-dimensional plot of r(C_M…N_ζ) and θ distributions based on the 1.5-ns simulations of the reactant complex for the second methyl transfer along with the corresponding free energy change as a function of r(C_M...N_ξ). (D) The average structure near the transition state for the second methyl transfer obtained from the free energy simulations. . (E) Left: The average active-site structure of the reactant complex for the third methyl transfer. Right: the two-dimensional plot of r(C_M…N_ζ) and θ distributions of the reactant complex for the third methyl transfer along with the corresponding free energy change as a function of r(C_M...N_ξ). (F) The average structure near the transition state for the third methyl transfer obtained from the free energy simulations.

(A)



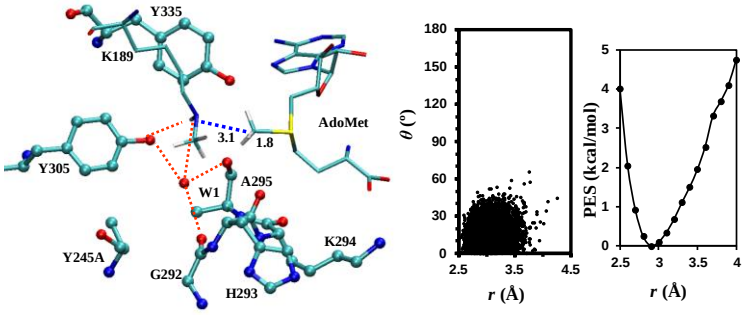
Reactant, 1st methyl transfer

(B)



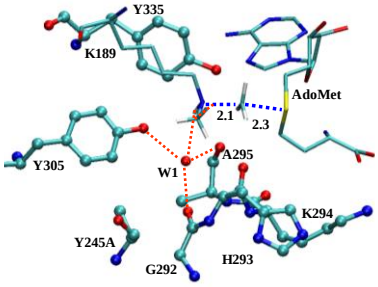
TS, 1st methyl transfer

(C)



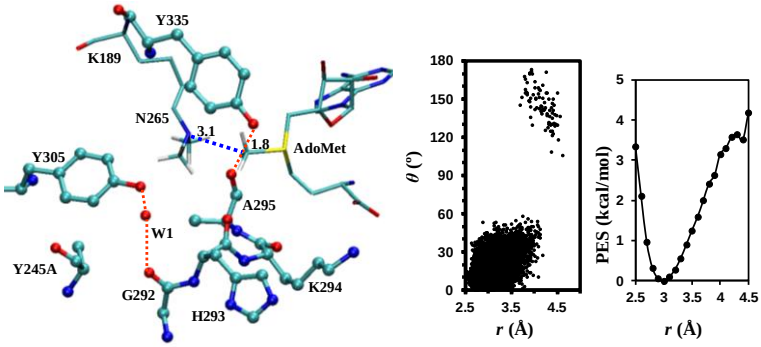
Reactant, 2nd methyl transfer

(D)



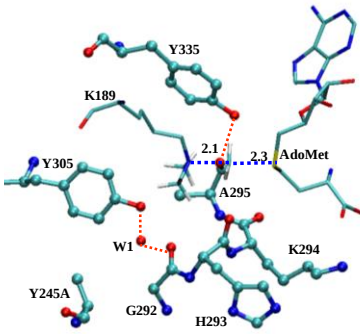
TS, 2nd methyl transfer

(E)



Reactant, 3rd methyl transfer

(F)

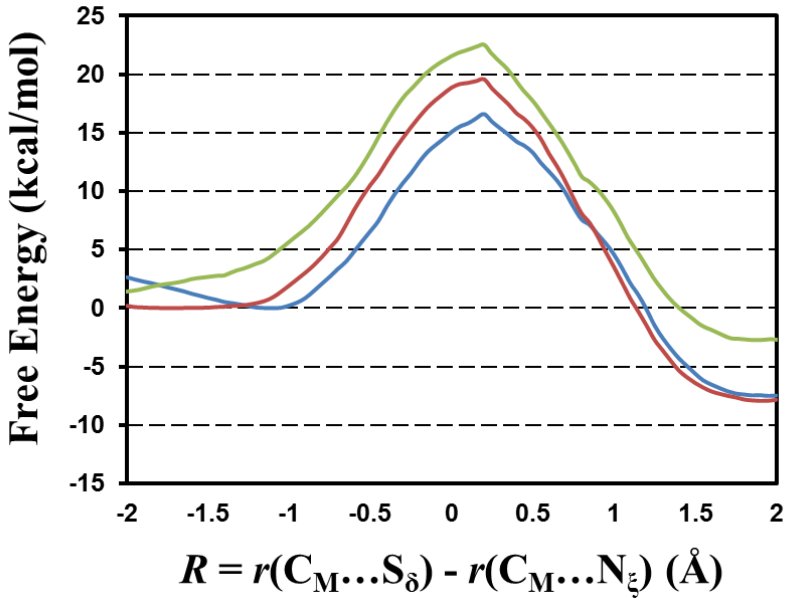


TS, 3rd methyl transfer

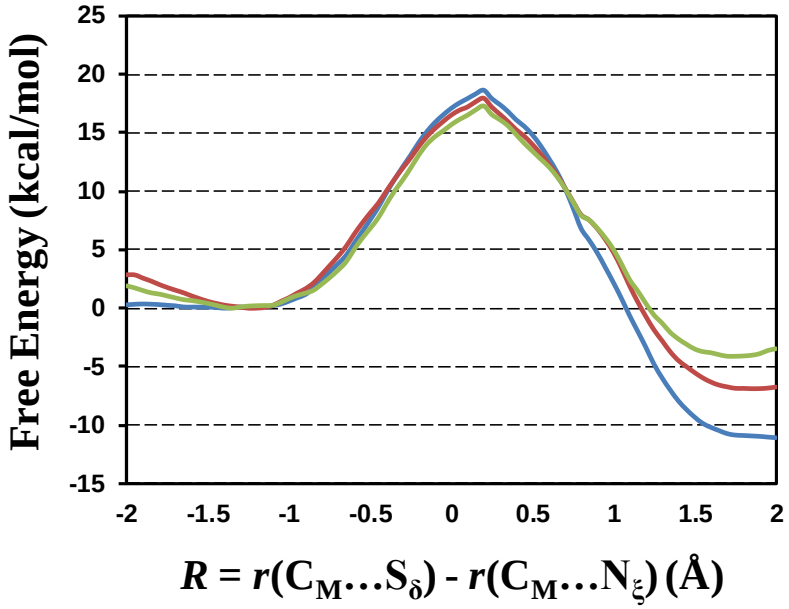
Figure 4.4. Free energy profiles for SET7/9 and its Y245A mutant.

(A) Free energy (potential of mean force) changes for the first, second and third methyl transfers, respectively, as a function of the reaction coordinate [R = $r(C_M...S_\delta) - r(C_M...N_\zeta)$] in the wild-type SET7/9 with TAF10-K189 peptide as the substrate. The first methyl transfer: blue line with a free energy barrier of 16.6 kcal/mol; the second methyl transfer: red line with a free energy barrier of 19.6 kcal/mol; the third methyl transfer: green line with a free energy barrier of 22.5 kcal/mol. Two free energy triplets, (0, Δ2-1W, Δ3-1W) and (ΔM-W, Δ2-1M, Δ3-1M) for wild-type and mutated enzyme, respectively, have been proposed to describe the product specificity of PKMTs and their mutants. For wild-type enzyme the second (Δ2-1W) and third (Δ3-1W) parameters are the differences in the free-energy barriers between the second and first and between the third and first methyl transfers, respectively. For the mutated enzyme, the first parameter (ΔM-W) is the difference in the free-energy barriers for the first methyl transfer in the wild-type and mutant. The second (Δ2-1M) and third (Δ3-1M) parameters are the differences in the free-energy barriers between the second and first and between the third and first methyl transfers, respectively. The energy triplet for SET7/9 with TAF10-K189 substrate was therefore calculated to be (0, 3, 5.9). (B) The free energy changes for the first, second and third methyl transfers as a function of the reaction coordinate in the Y245A mutant. The first methyl transfer: blue line with a free energy barrier of 18.7 kcal/mol; the second methyl transfer: red line; the third methyl transfer: green line. The energy triplet for Y245A is (2.1, -0.7, -1.3).

(A)


Free Energy (kcal/mol)
25
20
15
10
5
0
-5
-10
-15
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
R = r(C_M…S_δ) - r(C_M…N_ξ) (Å)

(B)


Free Energy (kcal/mol)
25
20
15
10
5
0
-5
-10
-15
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
R = r(C_M…S_δ) - r(C_M…N_ξ) (Å)

CHAPTER 5

QM/MM AND EXPERIMENTAL STUDY OF SUBSTRATE-ASSISTED CATALYSIS, A NOVEL SUBSTRATE DISCRIMINATION MECHANISM IN A PROMISCUOUS ENZYME

Jianzhuang Yao, Haobo Guo, Minta Chaiprasongsuk, Nan Zhao, Feng Chen, and Hong Guo. “QM/MM and Experimental Study of Substrate-Assisted Catalysis: A Novel Substrate Discrimination Mechanism in a Promiscuous Enzyme” *To be submitted.*

Abstract

Although one of enzyme’s hallmarks is their highly catalytic specificity to natural substrates, substrate promiscuity has been increasingly reported recently. Promiscuous Enzymes generally show different catalytic efficiencies to different substrates. This substrate discrimination may be achieved by diversity of binding affinity to active site and/or different positioning relative to catalytic residues. It is believed from protein engineering perspective that substrate-assisted catalysis (SAC) may provide a strategy for drastically changing enzyme substrate specificity. The question remains as to whether naturally occurring promiscuous enzymes might have used SAC to achieve substrate discrimination. Here, we report the results of quantum mechanical/molecular mechanical (QM/MM) calculations and experimental study on salicylic acid binding protein 2 (SABP2) that has promiscuous esterase activity toward a series of substrates, but shows specific highly activity toward its natural substrate methyl salicylate (MeSA). It is demonstrated that SABP2 may represent a case in which the enzyme itself might have not been perfectly evolved and that SAC involving its natural substrate (MeSA) may be used to enhance the activity specifically on this substrate and therefore achieve substrate discrimination. The theoretical prediction is subsequently confirmed

experimentally by comparing the results for MeSA and methyl benzoate (MeBA) which lacks this ability of participating in SAC.

Introduction

Traditionally, enzymes are considered to specifically catalyze reaction for a single substrate. However, catalytic promiscuity of enzymes is not a new concept(Hult and Berglund 2007, Nath and Atkins 2007) and has been reported frequently.(Theodossis, Walden et al. 2004, Faridmoayer, Fentabil et al. 2008, Huang, Horváth et al. 2009, Cohen-Rosenzweig, Guan et al. 2014) The mechanism of promiscuity has been discussed in several reviews.(Aharoni, Gaidukov et al. 2005, Khersonsky, Roodveldt et al. 2006, Babtie, Tokuriki et al. 2010, Khersonsky and Tawfik 2010, Gatti-Lafranconi and Hollfelder 2013) Promiscuous enzymes generally show different catalytic efficiencies for different substrates. (k_{cat}/K_M).(Kopycki, Rauh et al. 2008, Sabini, Hazra et al. 2008) For promiscuous substrates, weakly binding to active site results in high K_M values, while inappropriate positioning relative to catalytic residues of active site leads to low k_{cat}.(Babtie, Tokuriki et al. 2010, Khersonsky and Tawfik 2010) Although the factors that may affect the catalytic efficiencies for different substrates for promiscuous enzymes have been discussed, to the best of our knowledge the role of substrate-assisted catalysis (SAC) has not been studied in detail. SAC is a process in which the functional groups from substrates, in addition to those from enzymes, contribute to the rate acceleration of the enzyme-catalyzed reactions.(Carter and Wells 1987, Carter, Nilsson et al. 1989, Xu, Guo et al. 2006) The removal of such functional groups from the substrates impairs the catalysis. It has been

demonstrated that for engineered enzymes SAC may provide a way of drastically changing substrate specificity. Question remains, however, as to whether naturally occurring enzymes might have already used SAC as one of important strategies for the discrimination between their natural substrates and promiscuous substrates that lack the groups capable of participating in SAC. Therefore, understanding the relationship between substrate promiscuity and SAC is of considerable importance in enzymology and might be helpful in protein engineering(Bornscheuer and Kazlauskas 2004, Lairson, Watts et al. 2006) and drug design.(Fernández, Tawfik et al. 2005, Lewinson, Adler et al. 2006)

Salicylic acid binding protein 2 (SABP2) studied in this work has specific esterase activity toward methyl salicylate (MeSA) and belongs to the *α/β* fold carboxylesterase hydrolase superfamily (EC 3.1.1).(Kumar and Klessig 2003, Forouhar, Yang et al. 2005, Kumar, Park et al. 2005, Kumar, Gustafsson et al. 2006, Zhao, Guan et al. 2009) Previous studies have demonstrated that SABP2 possesses specific esterase activity toward MeSA.(Kumar and Klessig 2003) MeSA is a compound found in plants and produced from salicylic acid (SA) by methylation activity of SA methyltransferase (SAMT).(Yao, Xu et al. 2010) The conversion of MeSA to salicylic acid (SA) catalyzed by SABP2 is believed to be a part of the signal transduction pathways that activate systemic acquired resistance and local defense responses to plant pathogens.(Kumar and Klessig 2003, Forouhar, Yang et al. 2005, Kumar, Park et al. 2005, Kumar, Gustafsson et al. 2006, Zhao, Guan et al. 2009) It has been shown(Forouhar, Yang et al. 2005) that Ser-81, His-238 and Asp-210 act as the catalytic triad (Figure 5.1a); the esterase activity was mostly or totally abrogated by Ser81Ala

mutation.(Forouhar, Yang et al. 2005) The SABP2 catalyzed reaction consists of acylation and deacylation processes. In Figure 5.1a, the proposed acylation reaction mechanism shows that Ser-81 serves as nucleophile and attacks the carbon atom of the ester bond. His-238, with help of Asp-210, functions as the general acid and base catalyst during the catalysis. It has been shown (Figure 5.1b) that SABP2 can catalyze reactions on a series of substrates.(Kumar and Klessig 2003, Forouhar, Yang et al. 2005, Kumar, Park et al. 2005, Kumar, Gustafsson et al. 2006, Zhao, Guan et al. 2009, Padhi, Fujii et al. 2010, Tripathi, Jiang et al. 2010) However, the structure basis of SABP2 substrate promiscuity is still unclear. Interestingly, SABP2 natural activity seems to be much stronger than the activities of this enzyme toward its promiscuous substrates.(Forouhar, Yang et al. 2005, Zhao, Guan et al. 2009, Padhi, Fujii et al. 2010) One possible explanation is that the active site of SABP2 is too crowded to be occupied by some relatively large substrates.(Forouhar, Yang et al. 2005) Nonetheless, the exact reasons for different catalytic efficiencies involving different substrates are still not clear.

The X-Ray crystal structure of tobacco SABP2 complexed with salicylic acid (SA)(Forouhar, Yang et al. 2005) reveals that only backbone amide group of Ala-13 in the enzyme seems to participate in the oxyanion hole interaction through the hydrogen bonding interaction with the SA carbonyl oxygen atom, as shown in Figure 5.2. Another possible Hydrogen bond involved in the oxyanion hole stabilization is from the backbone amide group of Leu-82. Nevertheless, this seems to be a weak hydrogen bond, as demonstrated by the fact that the corresponding hydrogen-bond distance is 3.55Å in the product complex

(between non-hydrogen atoms). The third hydrogen-bond donor for the oxyanion hole interaction is the hydroxyl group from the substrate, the hydrogen-bond distance is 2.66Å between two oxygen atoms (Figure 5.2). An interesting question is whether SABP2 would use the SAC mechanism involving the oxyanion-hole interaction to achieve substrate discrimination.

To elucidate the role of the oxyanion hole interaction involving the substrate in substrate promiscuity and discrimination, we have studied MeSA and MeBA acylation reaction catalyzed by SABP2 using computer simulations (with SCC-DFTB/MM Hamiltonian(Cui, Elstner et al. 2000)). The experimental study was also performed to confirm theoretical prediction. The results of computational and experimental investigations suggest that the substrate-assisted three-pronged oxyanion hole may results in the substrate discrimination between natural and promiscuous substrates.

Methods

Models. The initial coordinates for the reactant complexes were based on the crystallographic structure (PDB code: 1Y7I, 2.1 Å resolution) of tobacco SABP2 in complex with SA.(Forouhar, Yang et al. 2005) A methyl group was manually added to SA to form MeSA. The side chain of Ser81 was manually placed between His238 and the carbonyl carbon of MeSA to complete the catalytic triad (See below). Hydrogen atoms of protein were added by the HBUILD module(Brünger and Karplus 1988) implemented in CHARMM,(Brooks, Bruccoleri et al. 1983, Brooks, Brooks et al. 2009) but substrates hydrogen atoms were added manually. All acidic and basic amino acids protonation states

were determined by surrounding environment under physiological pH condition. The reactant complex was solvated by a water droplet with 22 Å radius by the standard superimposing protocol at the center of Ser-81 oxygen atom (O_γ labeled in Figure 5.2), and solvent water molecules within 2.8 Å of any crystal atoms were removed. A modified TIP3P water model(Jorgensen 1981, Neria, Fischer et al. 1996) was employed for the solvent. The substrates and the side chains of Ser-81, His-238 and Asp-210 were treated by QM and the rest of the system by MM. The link-atom approach(Field, Bash et al. 1990) implemented in the CHARMM program(Brooks, Brooks et al. 2009) was applied to separate the QM and MM regions. The DIV scheme(König, Hoffmann et al. 2005) implemented in CHARMM was used to treat the QM/MM frontier. The SCC-DFTB method(Cui, Elstner et al. 2001) with empirical dispersion corrections(Elstner, Hobza et al. 2001) implemented in CHARMM was used for the QM atoms and the all-hydrogen CHARMM potential function (PARAM27)(MacKerell, Bashford et al. 1998) was used for the MM atoms. For the reactant complex containing MeBA, the hydroxyl group at the 2-position of MeSA was removed to generate the substrate. The resulting systems contained around 5400 atoms, including about 600 water molecules.

Stochastic boundary molecular dynamics (MD) and potential mean force (PMF) free energy simulations. Stochastic boundary molecular dynamics method(Brooks, Brunger et al. 1985) was used with the oxygen atom (O_γ) of the Ser-81 as the reference center. The reaction region was a sphere with radius r of 20 Å, and the buffer region extended over 20 Å ≤ r ≤ 22 Å. All bonds involving hydrogen atoms except the ones involved in proton

transfers in the system were constrained by the SHAKE algorithm.(Ryckaert, Ciccotti et al. 1977) The initial structures for the entire stochastic boundary systems were optimized using the steepest descent (SD) and adopted-basis Newton-Raphson (ABNR) methods. The systems were gradually heated from 50.0 to 298.15K in 100 ps. A 1-fs time step was used for integration of the equation of motion. 500 ps QM/MM MD simulations were carried out for each of the reactant complexes.

The umbrella sampling method(Torrie and Valleau 1974) implemented in the CHARMM program along with the Weighted Histogram Analysis Method (WHAM)(Kumar, Bouzida et al. 1992) was then applied to determine the change of the free energy (potential of mean force) as a function of the reaction coordinate. The reaction coordinate (RC) was defined as a linear combination of r(C…O) and r(C...O_Y) [RC = r(C…O)-r(C...O_Y)], as shown in Figure 5.2. The determination of multidimensional free energy maps would be too time-consuming. Several previous *ab initio* studies(Zhang, Kua et al. 2002, Ishida and Kato 2003) on catalytic triad indicated that the one dimensional free energy simulations with the selection of a suitable reaction coordinate reflecting the key bond-breaking and bond-making events may be able to capture the key energetic properties for the acylation reaction (e.g., the free energy barrier). For each acylation process, more than forty windows were used, and for each window 100 ps simulations were performed with 50 ps equilibration. The force constants of the harmonic biasing potentials used in the PMF simulations were 50 to 500 kcal·mol^{-1}·Å^{-2}. The key average structure properties for important windows were provided in Table 5.1.

Benchmark calculations. Benchmark calculations with high level QM/MM calculations have been successfully used by several previous SCC-DFTBPR studies(Hou and Cui 2011, Hou and Cui 2013) to prove the applicability of SCC-DFTBPR/MM on different enzyme systems. To our knowledge, this is the first study applying SCC-DFTB to study the classic catalytic triad (Ser, His, and Asp). So, benchmark calculations were also carried out based on our enzyme complex model to explicitly demonstrate the reliability of the SCC-DFTB method on simulating the bond breaking and making process of the SABP2 acylation process. First, we compared the optimized structures of reactant complexes obtained based on the B3LYP(Becke 1988, Lee, Yang et al. 1988, Becke 1993)/MM and SCC-DFTB/MM methods The QM(B3LYP/6-31G(d,p))/MM calculations were performed for the acylation processes in SABP2 using GAMESS-US method(Schmidt, Baldridge et al. 1993) implemented in the CHARMM program.(Brooks, Bruccoleri et al. 1983, Brooks, Brooks et al. 2009) The basis set (6-31G(d,p)) has been successfully used on *ab initio* QM/MM simulations of Acetylcholinesterase.(Zhou, Wang et al. 2010, Sirin, Zhou et al. 2012)

In addition to the comparison of the reactant complex structures, we made the similar comparisons on the structures near the transition states as well as the tetrahedral intermediates on potential energy surfaces. The adiabatic mapping calculations were carried out based on the following protocol. The reaction coordinate is the same as that used in the PMF calculations. A harmonic constraint with force constant of 2000 kcal·mol^{-1}·Å^{-2} was added to the RC to guide the SABP2 acylation reaction. The RC value was

increased stepwise from reactant complex to product complex, with the step size of 0.2 Å. An adopted basis Newton Raphson (ABNR) minimization was carried out under the constraint. The RC value was then decreased from the product complex to the reactant complex. The forward and backward progresses were repeated many times to obtain a smooth and more reliable potential energy surface (PES) profile along the reaction path.

The B3LYP/MM correction of the SCC-DFTB/MM PMF barrier. Table 5.2 shows the energetic data from our benchmark calculations. The reaction barriers for the SABP2 acylation reactions with MeSA and MeBA tend to be underestimated by SCC-DFTB/MM compared with B3LYP/MM. Therefore, a barrier correction is necessary for the SCC-DFTB results.(Hou and Cui 2011, Hou and Cui 2013) The correction, to be added to the PMF free energy barrier, is the difference in potential energy barriers from the adiabatic mapping calculations based on the SCC-DFTB/MM and B3LYP/MM methods, respectively.

Kinetic Data. Vmax for PtSABP2-1 and PtSABP2-2 using MeSA as substrate have been experimentally determined in our previous study.(Zhao, Guan et al. 2009) It was used to calculate the kcat of this enzyme with MeSA. Vmax of PtSABP2-1 and PtSABP2-2 using MeBA as substrate was experimentally determined using the same protocol as that for determining the Vmax of PtSABP2-1 and PtSABP2-1 with MeSA.(Zhao, Guan et al. 2009) PtSABP2-1 and PtSABP2-2 recombinant protein expressed in E. coli was purified, then used in a two-step radiochemical esterase assay to determine the Vmax using MeBA as substrate. The obtained Vmax values were used to calculate the kcat of PtSABP2-1 and PtSABP2-2 using MeBA as substrate, as shown in Table 5.2.

Results and Discussion

Before comparing our theoretical calculations to the experimental results, it is worth pointing out that the SABP2 esterase activity consists of two steps (acylation and deacylation). The B3LYP/MM potential energy functions for the both acylation and deacylation have been determined for MeSA and MeBA, and these barriers suggest that the acylation step is rate-limiting step (see below).

The active site structure of SABP2 complexed with the product SA was plotted in Figure 5.2 based on the X-ray structure of the SABP2-SA complex.(Forouhar, Yang et al. 2005) The catalytic triad (Ser-81, His-238, and Asp-210) is clearly presented in the active site. Although there is a hydrogen bonding between His-248 and Asp-210, the hydroxyl group of Ser-81 side chain does not form a hydrogen-bond with His-238. In order to performing the nucleophilic attack by Ser-81 on the carbon atom (C) of substrates, the hydrogen-bonding network of catalytic triad was generated by changing the torsion angle of the side chain of Ser-81, consistently with a computer model from the previous investigation.(Forouhar, Yang et al. 2005) As a result, the residues of the catalytic triad are well aligned for the nucleophilic attack by Ser-81 on the carbon atom (C) of MeSA as required for acylation. In the active site of crystal structure, the oxyanion hole is formed by two hydrogen bonds from SA and Ala-13, respectively. It is of interest to note that the amide group of Leu-82 forms a rather week interaction with the carbonyl group O of the substrate with a distance of 3.55 Å. This oxyanion hole construction indicates that the two hydrogen bonds (O…O 2.64 Å and N…O 2.79 Å) may be of considerable importance for

the oxyanion hole stabilization and the weak hydrogen bond from the amide group of Leu-82 (N…O 3.55 Å) may be enhanced as a result of the developing of negative charge along the reaction coordinate of the acylation process.

MeSA acylation in SABP2. MeSA is the natural substrate of SABP2. The active-site structures of SABP2 complexed with MeSA obtained from the QM(B3LYP and SCC-DFTB)/MM geometry optimizations are shown in Figure 5.3a; the initial structures for the optimization calculations were from different snapshots based on the X-ray structure of the SABP2-SA complex and QM(SCC-DFTB)/MM molecular dynamics simulations. The optimized structure of the reactant complex by SCC-DFTB/MM shows generally good agreement with that obtained from B3LYP/MM. The hydrogen-bonding networks of the catalytic triad are aligned very well from the both QM/MM methods. The specific three-pronged oxyanion hole, consisting of two relatively strong hydrogen bonds and a weak one, also agrees well with the crystal structure. Although reactant state optimizations from the two QM/MM methods were started from two different snapshots of QM(SCC-DFTB)/MM MD simulation, most of the distances from the two QM/MM methods are similar. However, there are still some relatively large differences (more than 0.25 Å) from the two different QM/MM methods. The proton has been transferred from the side chain of His238 to the acylation intermediate, which is different with PMF calculation (Figure 5.4f). At this point, SCC-DFTB/MM and B3LYP/MM calculations show the same result, which is the most important for this benchmark calculation. The comparison indicates that the non-covalent bonds such as hydrogen bonds may still not be well described by

SCC-DFTB.(Yang, Yu et al. 2007, Gaus, Cui et al. 2011) Overall, SCC-DFTB/MM Hamiltonian seems to be a reliable method for describing the structure features of SABP2 complexed with MeSA.

For the most structural features in Figure 5.3b-d, the SCC-DFTB/MM and B3LYP/MM results show the similar trends from adiabatic mapping calculations. For example, the hydrogen bond from Leu-82 to the substrate is enhanced with developing the negative charge on the oxygen in going from RS to TI. The shapes of potential energy surfaces for the two QM/MM methods also agree well with each other for the two energy ridges (TSs) and a basin (TI), as shown in Table 5.2. The second TS was calculated to have a higher barrier. The barrier from SCC-DFTB/MM is lower than that calculated with B3LYP/MM (12.2 vs 17.7 kcal/mol). The ~5.5 kcal/mol underestimation by SCC-DFTB/MM in the energy barrier compared to the high level QM/MM calculation is consistent with a previous SCC-DFTBPR study.(Hou and Cui 2013) Although SCC-DFTB still has room for further improvement, it seems to be an efficient and reasonably reliable method for determining the relative barriers(Hou and Cui 2011, Hou, Zhu et al. 2012, Hou and Cui 2013) for computational simulation of acylation reaction of MeSA in SABP2.

The PMF free energy profile along the reaction coordinate of the acylation reaction of MeSA obtained from the SCC-DFTB/MM calculations is shown in Figure 5.4a. As can be seen from Figure 5.4e, there exists a stable tetrahedral intermediate (TI) during the acylation step (the structure has been shown in Figure 5.1b). The PMF profile with the SCC-DFTB/MM method (Figure 5.4a) shows the existence of two TSs (Figure 5.4b and

Figure 5.4f). The highest barrier is located at the second TS with a reaction coordinate near at 0.4 Å. The free energy barrier is 9.5 kcal/mol, which is 6 kcal/mol lower than the experimental kinetic estimate(Forouhar, Yang et al. 2005) based on k_{cat}, *as shown* in Table 5.2. This seems not surprising since the adiabatic mapping benchmark (discussed above) finds that SCC-DFTB/MM underestimates the potential energy barrier compared to B3LYP/MM by ~5.5 kcal/mol (Table 5.2). With an energy correction, our corrected PMF free energy barrier (14.9 kcal/mol) shows a good agreement with the experimental estimation (15.5 kcal/mol) based on the transition state theory. The potential energy function for the deacylation step was obtained as well; the energy barrier was calculated to be 15.4 kcal/mol from the QM(B3LYP/6-31G(d,p))/MM method. This suggests that the acylation process is the rate-limiting step. In Figure 5.4b, the bond lengths of C-O and C-O_γ were plotted as functions of RC during PMF calculation. The bond lengths of r(C…O) and r(C-O_γ) change smoothly and intersect at RC around 0 Å. Several important bond lengths from RS, TSs, and TI has been summarized at Table 5.1.

Figure 5.4c-f show that the average structures along the reaction path share the similar properties as those found in the optimized structures (Figure 5.3a-d) in terms of patterns of hydrogen bonding network and oxyanion hole construction. The role of the catalytic triad in acylation process seems to be the same as in the classic serine protease.(Blow 1976, Zhang, Kua et al. 2002) For instance, the carboxylate of Asp-208 works with the incipient imidazolium cation to stabilize the TS and the TI by the electrostatic interaction.(Blow 1976, Warshel, Naray-Szabo et al. 1989)

We also investigated the changes of the hydrogen bonds formed in the three-pronged oxyanion hole (Ala-13, Leu-82, and MeSA hydroxyl group). The changes in the heavy atom distances (N-O, O-O) along the reaction coordinate are shown in Figure 5.7. In RS (Figure 5.4a), two hydrogen bonds are formed between the MeSA carbonyl O_1 and the amide group of Ala-13 and hydroxyl group of MeSA, with the N-O and O-O distances of 1.90 and 1.80 Å, respectively. It is of interest to note that the amide group of Leu-82 forms a rather week interaction with O_1 of the substrate with a distance of 2.96 Å in RS. Thus, the enzyme seems to have been evolved with a less effective oxyanion hole for the stabilization of the tetrahedral intermediates (TIs) during the catalysis. Interestingly, the 2-OH group of MeSA, in addition to the amide group of Ala-13, forms an intra-molecular hydrogen bond with O_1 of the substrate with a hydrogen-bond distance of 1.67 Å. Moreover, this hydrogen bond seems to strengthen significantly as the reaction proceeds from the reactant complex to the tetrahedral intermediate (see Figure 5.4c-f Figure 5.7) based on the PMF calculations. Thus, the 2-OH group from the MeSA substrate may help to generate a more effective oxyanion hole and lower the activation barrier through SAC. This may contribute to the relatively high specificity of SABP2 toward MeSA compared to that toward some other compounds (Figure 5.1b) which lack such or similar interactions.

MeBA acylation in SABP2. To understand the substrate promiscuity of SABP2 and test the hypothesis concerning the importance of SAC in substrate discrimination, the catalytic mechanism of acylation involving methyl benzoate (MeBA) (which is one of the promiscuous substrates of SABP2 and lacks the 2-OH hydroxyl group) was also studied.

There is no available experimental data (such as k_{cat}) for the MeBA hydrolysis in tobacco SAPB2.

Similar with the MeSA benchmark comparisons (See above), the SCC-DFTB/MM minimizations for MeBA in SABP2 give overall similar results to B3LYP/MM calculations (Figure 5.5a-b). In the optimized RS (Figure 5.5a), SABP2 complexed with MeBA shows identical patterns of the hydrogen bonding network with the MeSA complex. Furthermore, MeBA shows the similar oxyanion hole configuration with MeSA (Figure 5.3a). Without the OH group in the substrate, MeBA complex forms a two-pronged oxyanion hole instead of three-pronged one. As a result of the missing hydrogen bond in the oxyanion hole, the amide group of Leu-82 moves closer to O_1 of the MeBA carbonyl group compared with that in the optimized RS of the MeSA complex from the both SCC-DFTB/MM and B3LYP/MM calculations. For the results of adiabatic mapping calculations, SCC-DFTB/MM underestimates the barrier compared to B3LYP/MM by a few kcal/mol (14.9 vs 19.5 kcal/mol), which is similar to the situation for MeSA. Moreover, the structures near TS from the two QM/MM methods are consistent with each other (Figure 5.5b). The proton has been transferred from side chain of His238 to the acylation product. This phenomenon has been observed and discussed in the MeSA benchmark calculations. Compared with the MeSA acylation process (Figure 5.3), a stable TI was not observed for the reaction involving MeBA from the adiabatic mapping calculation. Therefore, the comparison of the RS active sites for the MeSA and MeBA complexes does not support the hypotheses that the inappropriate positioning of MeBA compared to MeSA leads to

low k_{cat}. The results seem to support our hypotheses that SAC may contribute to substrate discrimination.

The calculated PMF (Figure 5.6) with the SCC-DFTB/MM method indicates an exothermic reaction with the plateau TS with the highest point at RC ~ 0.1 Å. This PMF profile supports our hypothesis that SABP2 forms a sub-optimal oxyanion hole during evolution. In Figure 5.6b, the distances of C-O and C-O_γ were plotted as functions of RC during the PMF calculation. Similar to the case involving MeSA, the distances of r(C…O) and r(C-O_γ) change smoothly and intersect at RC around 0 Å. The calculated free energy barrier is 11.4 kcal/mol, which is 1.9 kcal/mol higher than the PMF free energy barrier involving MeSA. As is shown in Table 5.2, this barrier is likely underestimated by about 4.9 kcal/mol because of the systematic error of SCC-DFTB. After the barrier correction based on adiabatic mapping calculations with the different QM/MM methods, the corrected estimate for the free energy barrier is about 16.1 kcal/mol (Table 5.2), which is 1.2 kcal/mol higher than corrected MeSA free energy barrier. To confirm the theoretical prediction concerning the involvement of the hydroxyl group of MeSA in SAC and substrate discrimination, we measured the esterase activities of PtSABP2-1 and PtSABP2-2 toward MeSA and MeBA, respectively (Table 5.2). PtSABP2-1 and PtSABP2-2(Zhao, Guan et al. 2009) are 98% identical to each other and has over 77% sequence identity with tobacco SABP2(Forouhar, Yang et al. 2005). Their activity profiles for different substrates was found to be basically the same as that of tobacco SABP2.(Forouhar, Yang et al. 2005) Consistent with the results of the theoretical calculations, the activation barriers based

on the k_{cat} values for MeSA (0.120 s^{-1}) and MeBA (0.007 s^{-1}) catalyzed by PtSABP2-1 were found to be 16.3 and 18.0 kcal/mol, respectively (0.45 s^{-1} for tobacco SABP2 with MeSA(Forouhar, Yang et al. 2005)). Correspondingly, the K_M values were found to be 68.2μM and 39.6μM for MeSA and MeBA, respectively. Similar to the case of PtSABP2-1, the activation barriers based on k_{cat} values for MeSA (0.187 s^{-1}) and MeBA (0.038 s^{-1}) catalyzed by PtSABP2-2 are 16.0 and 17.0 kcal/mol. Again, the K_M values were found to be 24.6μM and 42.1μM for MeSA and MeBA, respectively. Thus, k_{cat} is the major determinant for the specificity. The measured activation energy difference between MeSA and MeBA in both PtSABP2-1 and PtSABP2-2 based on transition state theory is about 1.0-1.7 kcal/mol, which agrees well with our PMF free energy calculation (1.2 kcal/mol).

The average structurefor the reactant complex and that at R = -0.1 (i.e., the structure that is close to the TI in the reaction involving MeSA) are given in Figures 4.6c and 6d, respectively. Figure 5.6a shows the free energy function along the reaction coordinate of the acylation reaction involving MeBA obtained from the SCC-DFTB/MM PMF calculations. As can be seen from Figure 5.6a, the energy barrier for acylation is not located at a ridge as shown in MeSA TS2 (Figure 5.4a), but at a plateau. A stable TI was not obtained for the reaction involving MeBA from the calculation; this suggests that the enzyme SABP2 may not perfectly evolved for MeBA hydrolysis due to the insufficient oxyanion hole stabilization. The energy barriers from experimental and theoretical studies are about 1.0-1.7 kcal/mol higher than the barrier involving MeSA (see above).

This difference may be attributed to the absence and existence of the interaction involving the 2-OH group in MeBA and MeSA, respectively. The existence of this group can presumably increase the effectiveness of the oxyanion-hole interaction and provide additional stabilization of TS with the MeSA substrate. Nevertheless, several previous studies on the oxyanion hole showed that the disturbing electrostatic interaction of a hydrogen bond or destroying the oxyanion hole may cause an increase of the barrier by 1.5-3.0 kcal/mol in Thrombin and Trypsin(Bobofchak, Pineda et al. 2005) and abrogation of the catalytic activities of subtilisin(Bryan, Pantoliano et al. 1986) and serine-carboxyl peptidase.(Xu, Yao et al. 2011, Yao, Xu et al. 2012) Interestingly, the hydrogen bond formed by Leu-82 in the oxyanion hole shows an enhanced strength in the MeBA complex compared to the MeSA complex(Figure 5.4c-e, Figure 5.6c,d and Figure 5.7). This mechanism may provide an explanation to the substrate promiscuity of SABP2 that has catalytic activities toward a series of substrates (Figure 5.1b).

Conclusion

Many enzymes have been found to possess remarkable ability to discriminate between their substrate and closely related molecules.(Fersht 1998, Hedstrom 2002) However, our understanding of this ability of enzymes is still lacking. For SABP2, the enzyme seems to have been evolved with a less perfect oxyanion hole with the additional contribution coming from the natural substrate MeSA. For other compounds without such well-positioned functional group, the catalysis would be less efficient and the discrimination between MeSA and these compounds would then be achieved. Our results therefore

support the suggestion that some naturally occurring enzymes might have already used SAC as one of important strategies for substrate discrimination. The experimental data seem to support our theoretical prediction that the participation of the 2-OH group for the catalysis may contribute, at least in part, to the relatively strong esterase activity of SABP2 towards MeSA. It would be of considerable interest to examine whether other naturally occurring enzymes would use SAC for substrate discrimination as well.

Previous experimental studies have shown that SABP2 is highly specific to MeSA at physiologically relevant concentrations.(Forouhar, Yang et al. 2005, Zhao, Guan et al. 2009) Our study suggests the high specificity to MeSA may root in SAC. As discussed above, the hydrogen bonding in oxyanion hole may contributes about 1.5-3.0 kcal/mol(Bobofchak, Pineda et al. 2005) or even more to favor the reaction. However, SABP2 shows substrate promiscuity toward a series of substrates (Table 5.1b). By losing a substrate-assisted hydrogen bond, the activation energy barrier of MeBA hydrolysis is only increased by 1.0-1.7 kcal/mol. The difference of catalytic activity between natural substrate and promiscuous ones seems too small to be explained based on the missing OH group of promiscuous substrates compared to natural substrate MeSA. Interestingly, we found two-pronged oxyanion hole provide a hydrogen bonding enhancement mechanism by the amide group of Leu-82 which may partly compensate the loss of the TS stabilization from missing one hydrogen bonding. Nevertheless, the promiscuous substrates of SABP2 previously used in the study (such as pNP acetate, methyl jasmonate and methyl indole-acetic acid) have structures that are quite different from that of MeSA

(Table 5.1b). Thus, the relatively weak activities for these compounds may have different origins (e.g., different binding modes) besides of the lack of the contribution from the 2-OH group for the oxyanion-hole interactions.

Acknowledgement

This work was supported by grants 0817940 from the National Science Foundation (H.G.) and by the Department of Energy Office of Biological and Environmental Research - Genome to Life Program through the BioEnergy Science Center (BESC) (to FC).

References

Aharoni, A., L. Gaidukov, O. Khersonsky, S. M. Gould, C. Roodveldt and D. S. Tawfik (2005). "The 'evolvability' of promiscuous protein functions." Nature Genetics **37**(1): 73-76.

Babtie, A., N. Tokuriki and F. Hollfelder (2010). "What makes an enzyme promiscuous?" Current Opinion in Chemical Biology **14**(2): 200-207.

Becke, A. D. (1988). "Density-functional exchange-energy approximation with correct asymptotic behavior." Physical Review A **38**(6): 3098-3100.

Becke, A. D. (1993). "Density-functional thermochemistry. III. The role of exact exchange." The Journal of Chemical Physics **98**(7): 5648-5652.

Blow, D. M. (1976). "Structure and mechanism of chymotrypsin." Accounts of Chemical Research **9**(4): 145-152.

Bobofchak, K. M., A. O. Pineda, F. S. Mathews and E. Di Cera (2005). "Energetic and Structural Consequences of Perturbing Gly-193 in the Oxyanion Hole of Serine Proteases." Journal of Biological Chemistry **280**(27): 25644-25650.

Bornscheuer, U. T. and R. J. Kazlauskas (2004). "Catalytic Promiscuity in Biocatalysis: Using Old Enzymes to Form New Bonds and Follow New Pathways." Angewandte Chemie International Edition **43**(45): 6032-6040.

Brooks, B. R., C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E.

Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus (2009). "CHARMM: The Biomolecular Simulation Program." Journal of Computational Chemistry **30**(10): 1545-1614.

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus (1983). "CHARMM - A PROGRAM FOR MACROMOLECULAR ENERGY, MINIMIZATION, AND DYNAMICS CALCULATIONS." Journal of Computational Chemistry **4**(2): 187-217.

Brooks, C. L., A. Brunger and M. Karplus (1985). "ACTIVE-SITE DYNAMICS IN PROTEIN MOLECULES - A STOCHASTIC BOUNDARY MOLECULAR-DYNAMICS APPROACH." Biopolymers **24**(5): 843-865.

Brünger, A. T. and M. Karplus (1988). "Polar hydrogen positions in proteins: Empirical energy placement and neutron diffraction comparison." Proteins: Structure, Function, and Bioinformatics **4**(2): 148-156.

Bryan, P., M. W. Pantoliano, S. G. Quill, H. Y. Hsiao and T. Poulos (1986). "Site-directed mutagenesis and the role of the oxyanion hole in subtilisin." Proceedings of the National Academy of Sciences **83**(11): 3743-3745.

Carter, P., B. Nilsson, J. P. Burnier, D. Burdick and J. A. Wells (1989). "ENGINEERING SUBTILISIN BPN' FOR SITE-SPECIFIC PROTEOLYSIS." Proteins-Structure Function and Genetics **6**(3): 240-248.

Carter, P. and J. A. Wells (1987). "ENGINEERING ENZYME SPECIFICITY BY SUBSTRATE-ASSISTED CATALYSIS." Science **237**(4813): 394-399.

Cohen-Rosenzweig, C., Z. Guan, B. Shaanan and J. Eichler (2014). "Substrate Promiscuity: AglB, the Archaeal Oligosaccharyltransferase, Can Process a Variety of Lipid-Linked Glycans." Applied and Environmental Microbiology **80**(2): 486-496.

Cui, Q., M. Elstner, E. Kaxiras, T. Frauenheim and M. Karplus (2000). "A QM/MM Implementation of the Self-Consistent Charge Density Functional Tight Binding (SCC-DFTB) Method." The Journal of Physical Chemistry B **105**(2): 569-585.

Cui, Q., M. Elstner, E. Kaxiras, T. Frauenheim and M. Karplus (2001). "A QM/MM implementation of the self-consistent charge density functional tight binding (SCC-DFTB) method." Journal of Physical Chemistry B **105**(2): 569-585.

Elstner, M., P. Hobza, T. Frauenheim, S. Suhai and E. Kaxiras (2001). "Hydrogen bonding and stacking interactions of nucleic acid base pairs: A density-functional-theory based treatment." The Journal of Chemical Physics **114**(12): 5149-5155.

Faridmoayer, A., M. A. Fentabil, M. F. Haurat, W. Yi, R. Woodward, P. G. Wang and M. F. Feldman (2008). "Extreme Substrate Promiscuity of the Neisseria Oligosaccharyl Transferase Involved in Protein O-Glycosylation." Journal of Biological Chemistry **283**(50): 34596-34604.

Fernández, A., D. S. Tawfik, B. Berkhout, R. Sanders, A. Kloczkowski, T. Sen, B. Jernigan, D. S. Tawfik, B. Berkhout, R. W. Sanders, A. Kloczkowski, T. Z. Sen and R. L. Jernigan (2005). "Protein Promiscuity: Drug Resistance and Native Functions—HIV-1 Case." Journal of Biomolecular Structure and Dynamics **22**(6): 615-624.

Fersht, A. (1998). Structure and Mechanism in Protein Science, W.H. Freeman and Company.

Field, M. J., P. A. Bash and M. Karplus (1990). "A COMBINED QUANTUM-MECHANICAL AND MOLECULAR MECHANICAL POTENTIAL FOR MOLECULAR-DYNAMICS SIMULATIONS." Journal of Computational Chemistry **11**(6): 700-733.

Forouhar, F., Y. Yang, D. Kumar, Y. Chen, E. Fridman, S. W. Park, Y. Chiang, T. B. Acton, G. T. Montelione, E. Pichersky, D. F. Klessig and L. Tong (2005). "Structural and biochemical studies identify tobacco SABP2 as a methyl salicylate esterase and implicate it in plant innate immunity." Proceedings of the National Academy of Sciences of the United States of America **102**(5): 1773-1778.

Gatti-Lafranconi, P. and F. Hollfelder (2013). "Flexibility and Reactivity in Promiscuous Enzymes." ChemBioChem **14**(3): 285-292.

Gaus, M., Q. Cui and M. Elstner (2011). "DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB)." Journal of Chemical Theory and Computation **7**(4): 931-948.

Hedstrom, L. (2002). "Serine protease mechanism and specificity." Chemical Reviews **102**(12): 4501-4523.

Hou, G. and Q. Cui (2011). "QM/MM Analysis Suggests That Alkaline Phosphatase (AP) and Nucleotide Pyrophosphatase/Phosphodiesterase Slightly Tighten the Transition State for Phosphate Diester Hydrolysis Relative to Solution: Implication for Catalytic

Promiscuity in the AP Superfamily." Journal of the American Chemical Society **134**(1): 229-246.

Hou, G. and Q. Cui (2013). "Stabilization of Different Types of Transition States in a Single Enzyme Active Site: QM/MM Analysis of Enzymes in the Alkaline Phosphatase Superfamily." Journal of the American Chemical Society **135**(28): 10457-10469.

Hou, G., X. Zhu, M. Elstner and Q. Cui (2012). "A Modified QM/MM Hamiltonian with the Self-Consistent-Charge Density-Functional-Tight-Binding Theory for Highly Charged QM Regions." Journal of Chemical Theory and Computation **8**(11): 4293-4304.

Huang, F.-C., G. Horváth, P. Molnár, E. Turcsi, J. Deli, J. Schrader, G. Sandmann, H. Schmidt and W. Schwab (2009). "Substrate promiscuity of RdCCD1, a carotenoid cleavage oxygenase from Rosa damascena." Phytochemistry **70**(4): 457-464.

Hult, K. and P. Berglund (2007). "Enzyme promiscuity: mechanism and applications." Trends in Biotechnology **25**(5): 231-238.

Ishida, T. and S. Kato (2003). "Theoretical Perspectives on the Reaction Mechanism of Serine Proteases: The Reaction Free Energy Profiles of the Acylation Process." Journal of the American Chemical Society **125**(39): 12035-12048.

Jorgensen, W. L. (1981). "QUANTUM AND STATISTICAL MECHANICAL STUDIES OF LIQUIDS .10. TRANSFERABLE INTERMOLECULAR POTENTIAL FUNCTIONS FOR WATER, ALCOHOLS, AND ETHERS - APPLICATION TO LIQUID WATER." Journal of the American Chemical Society **103**(2): 335-340.

Khersonsky, O., C. Roodveldt and D. S. Tawfik (2006). "Enzyme promiscuity: evolutionary and mechanistic aspects." Current Opinion in Chemical Biology **10**(5): 498-508.

Khersonsky, O. and D. S. Tawfik (2010). Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. Annual Review of Biochemistry, Vol 79. R. D. Kornberg, C. R. H. Raetz, J. E. Rothman and J. W. Thorner. **79:** 471-505.

König, P. H., M. Hoffmann, T. Frauenheim and Q. Cui (2005). "A Critical Evaluation of Different QM/MM Frontier Treatments with SCC-DFTB as the QM Method." The Journal of Physical Chemistry B **109**(18): 9082-9095.

Kopycki, J. G., D. Rauh, A. A. Chumanevich, P. Neumann, T. Vogt and M. T. Stubbs (2008). "Biochemical and Structural Analysis of Substrate Promiscuity in Plant Mg2+-Dependent O-Methyltransferases." Journal of Molecular Biology **378**(1): 154-164.

Kumar, D., C. Gustafsson and D. F. Klessig (2006). "Validation of RNAi silencing specificity using synthetic genes: salicylic acid-binding protein 2 is required for innate immunity in plants." Plant Journal **45**(5): 863-868.

Kumar, D. and D. F. Klessig (2003). "High-affinity salicylic acid-binding protein 2 is required for plant innate immunity and has salicylic acid-stimulated lipase activity." Proceedings of the National Academy of Sciences of the United States of America **100**(26): 16101-16106.

Kumar, D., S. Park, F. Forouhar, Y. Yang, A. Vlot, E. Fridman, Y. Chiang, T. Acton, V. Shulaev, G. Montelione, E. Pichersky, L. Tong and D. Klessig (2005). "Identification of SA-

binding protein 2 (SABP2) as a critical component of plant innate immunity." Phytopathology **95**(6): S55-S55.

Kumar, S., D. Bouzida, R. H. Swendsen, P. A. Kollman and J. M. Rosenberg (1992). "THE WEIGHTED HISTOGRAM ANALYSIS METHOD FOR FREE-ENERGY CALCULATIONS ON BIOMOLECULES .1. THE METHOD." Journal of Computational Chemistry **13**(8): 1011-1021.

Lairson, L. L., A. G. Watts, W. W. Wakarchuk and S. G. Withers (2006). "Using substrate engineering to harness enzymatic promiscuity and expand biological catalysis." Nat Chem Biol **2**(12): 724-728.

Lee, C., W. Yang and R. G. Parr (1988). "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density." Physical Review B **37**(2): 785-789.

Lewinson, O., J. Adler, N. Sigal and E. Bibi (2006). "Promiscuity in multidrug recognition and transport: the bacterial MFS Mdr transporters." Molecular Microbiology **61**(2): 277-284.

MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus (1998). "All-atom empirical potential for molecular modeling and dynamics studies of proteins." Journal of Physical Chemistry B **102**(18): 3586-3616.

Nath, A. and W. M. Atkins (2007). "A Quantitative Index of Substrate Promiscuity†." Biochemistry **47**(1): 157-166.

Neria, E., S. Fischer and M. Karplus (1996). "Simulation of activation free energies in molecular systems." Journal of Chemical Physics **105**(5): 1902-1921.

Padhi, S. K., R. Fujii, G. A. Legatt, S. L. Fossum, R. Berchtold and R. J. Kazlauskas (2010). "Switching from an Esterase to a Hydroxynitrile Lyase Mechanism Requires Only Two Amino Acid Substitutions." Chemistry & Biology **17**(8): 863-871.

Ryckaert, J. P., G. Ciccotti and H. J. C. Berendsen (1977). "NUMERICAL-INTEGRATION OF CARTESIAN EQUATIONS OF MOTION OF A SYSTEM WITH CONSTRAINTS - MOLECULAR-DYNAMICS OF N-ALKANES." Journal of Computational Physics **23**(3): 327-341.

Sabini, E., S. Hazra, S. Ort, M. Konrad and A. Lavie (2008). "Structural Basis for Substrate Promiscuity of dCK." Journal of Molecular Biology **378**(3): 607-621.

Schmidt, M. W., K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis and J. A. Montgomery (1993). "GENERAL ATOMIC AND MOLECULAR ELECTRONIC-STRUCTURE SYSTEM." Journal of Computational Chemistry **14**(11): 1347-1363.

Sirin, G. S., Y. Zhou, L. Lior-Hoffmann, S. Wang and Y. Zhang (2012). "Aging Mechanism of Soman Inhibited Acetylcholinesterase." The Journal of Physical Chemistry B **116**(40): 12199-12207.

Theodossis, A., H. Walden, E. J. Westwick, H. Connaris, H. J. Lamble, D. W. Hough, M. J. Danson and G. L. Taylor (2004). "The Structural Basis for Substrate Promiscuity in 2-

Keto-3-deoxygluconate Aldolase from the Entner-Doudoroff Pathway in Sulfolobus solfataricus." Journal of Biological Chemistry **279**(42): 43886-43892.

Torrie, G. M. and J. P. Valleau (1974). "MONTE-CARLO FREE-ENERGY ESTIMATES USING NON-BOLTZMANN SAMPLING - APPLICATION TO SUBCRITICAL LENNARD-JONES FLUID." Chemical Physics Letters **28**(4): 578-581.

Tripathi, D., Y.-L. Jiang and D. Kumar (2010). "SABP2, a methyl salicylate esterase is required for the systemic acquired resistance induced by acibenzolar-S-methyl in plants." FEBS Letters **584**(15): 3458-3463.

Warshel, A., G. Naray-Szabo, F. Sussman and J. K. Hwang (1989). "How do serine proteases really work?" Biochemistry **28**(9): 3629-3637.

Xu, Q., H. Guo and A. Wlodawer (2006). "The importance of dynamics in substrate-assisted catalysis and specificity." Journal of the American Chemical Society **128**(18): 5994-5995.

Xu, Q., J. Yao, A. Wlodawer and H. Guo (2011). "Clarification of the Mechanism of Acylation Reaction and Origin of Substrate Specificity of the Serine-Carboxyl Peptidase Sedolisin through QM/MM Free Energy Simulations." The Journal of Physical Chemistry B **115**(10): 2470-2476.

Yang, H. Yu, D. York, Q. Cui and M. Elstner (2007). "Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method: Third-Order Expansion of the Density Functional Theory Total Energy and Introduction of a Modified Effective Coulomb Interaction." The Journal of Physical Chemistry A **111**(42): 10861-10873.

Yao, J., Q. Xu, F. Chen and H. Guo (2010). "QM/MM Free Energy Simulations of Salicylic Acid Methyltransferase: Effects of Stabilization of TS-like Structures on Substrate Specificity." The Journal of Physical Chemistry B **115**(2): 389-396.

Yao, J., Q. Xu and H. Guo (2012). "QM/MM and free-energy simulations of deacylation reaction catalysed by sedolisin, a serine-carboxyl peptidase." Molecular Simulation **39**(3): 206-213.

Zhang, Y., J. Kua and J. A. McCammon (2002). "Role of the Catalytic Triad and Oxyanion Hole in Acetylcholinesterase Catalysis: An ab initio QM/MM Study." Journal of the American Chemical Society **124**(35): 10572-10577.

Zhao, N., J. Guan, F. Forouhar, T. J. Tschaplinski, Z. M. Cheng, L. Tong and F. Chen (2009). "Two poplar methyl salicylate esterases display comparable biochemical properties but divergent expression patterns." Phytochemistry **70**(1): 32-39.

Zhou, Y., S. Wang and Y. Zhang (2010). "Catalytic Reaction Mechanism of Acetylcholinesterase Determined by Born−Oppenheimer Ab Initio QM/MM Molecular Dynamics Simulations." The Journal of Physical Chemistry B **114**(26): 8817-8825.

Appendix

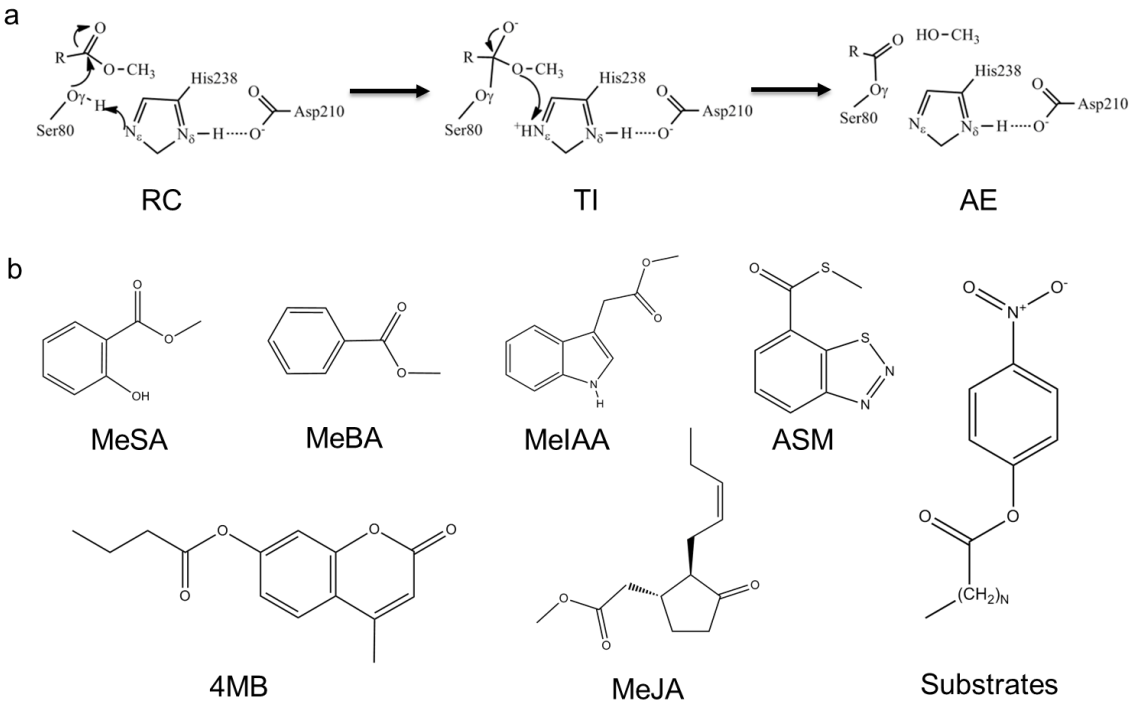


Figure 5.1. Acylation catalytic mechanism of SABP2.

(a) Acylation process: Reactant complex, Tetrahedral intermediate, and Acyl-enzyme complex (AE). (b) Promiscuous substrates catalyzed by SABP2. MeSA, methyl salicylate;(Forouhar, Yang et al. 2005, Zhao, Guan et al. 2009) MeBA, methyl benzoate; MeJA, methyl jasmonate;(Forouhar, Yang et al. 2005, Zhao, Guan et al. 2009) MeIAA, methyl indole acetic acid;(Forouhar, Yang et al. 2005, Zhao, Guan et al. 2009) ASM, acibenzolar-S-methyl;(Tripathi, Jiang et al. 2010) 4MB, 4-methylumbelliferone butyrate;(Kumar and Klessig 2003) Substrate* includes para-nitrophenyl (pNP) palmitate, pNP myristate, pNP butyrate,(Kumar and Klessig 2003) and pNP acetate.(Padhi, Fujii et al. 2010)

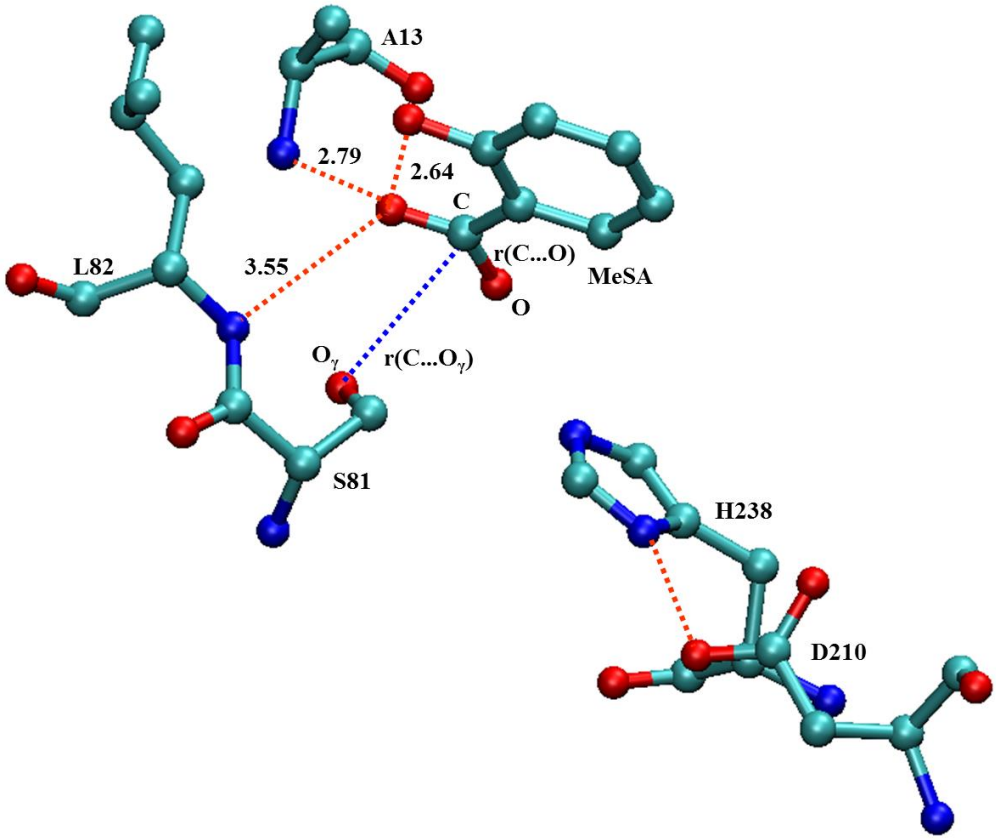


Figure 5.2. Active site of the X-Ray crystal structure of SABP2 (PDB code 1Y7I).

Reaction coordinate, RC = r(C…O)-r(C...Oγ), is labeled. All distances are in Å.

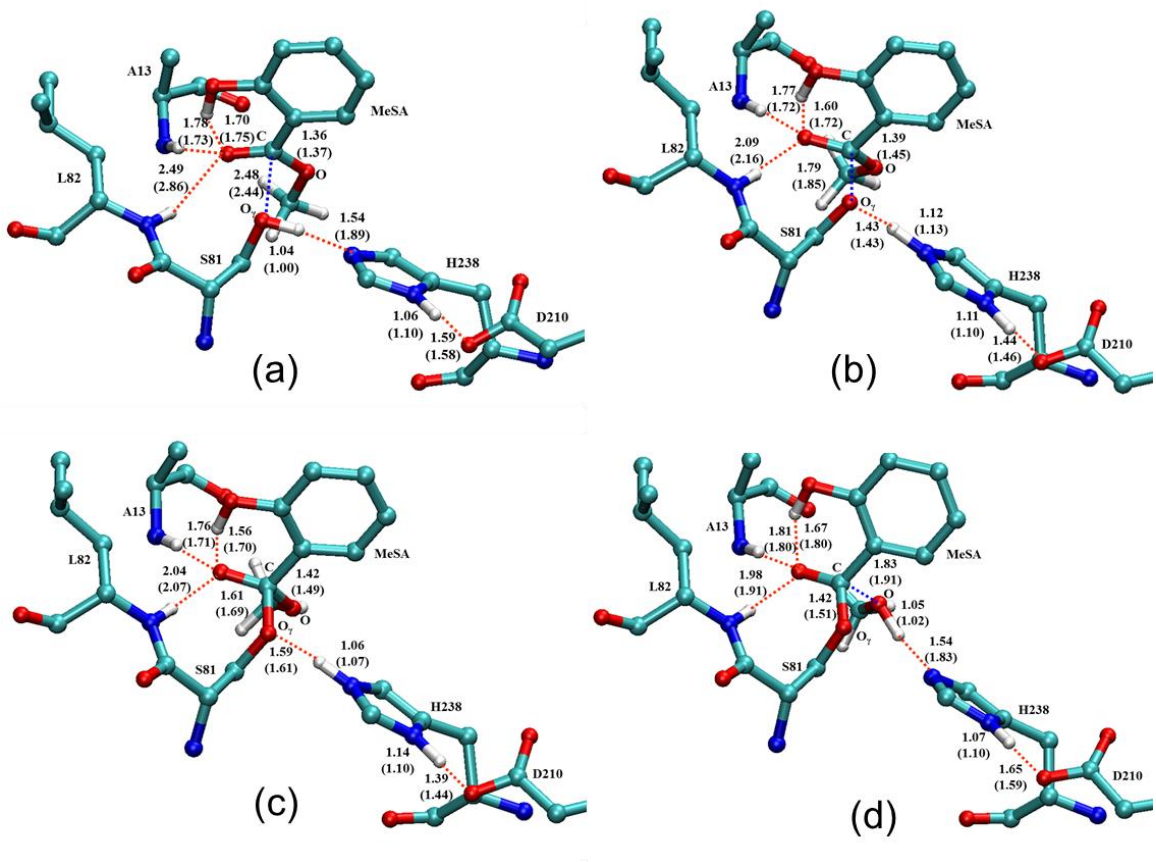


Figure 5.3. Benchmark calculations for SABP2 complexed with MeSA.

Numbers within parentheses are obtained by SCC-DFTB/MM optimization with dispersion correction; those without parentheses are obtained by B3LYP/6-31G(d,p)/MM optimization. Unimportant hydrogens are omitted for clarity. (a) Reactant state; (b) Transition state 1 by adiabatic mapping; (c) Tetrahedral intermediate by adiabatic mapping; (d) Transition state 2 by adiabatic mapping. All distances are in Å.

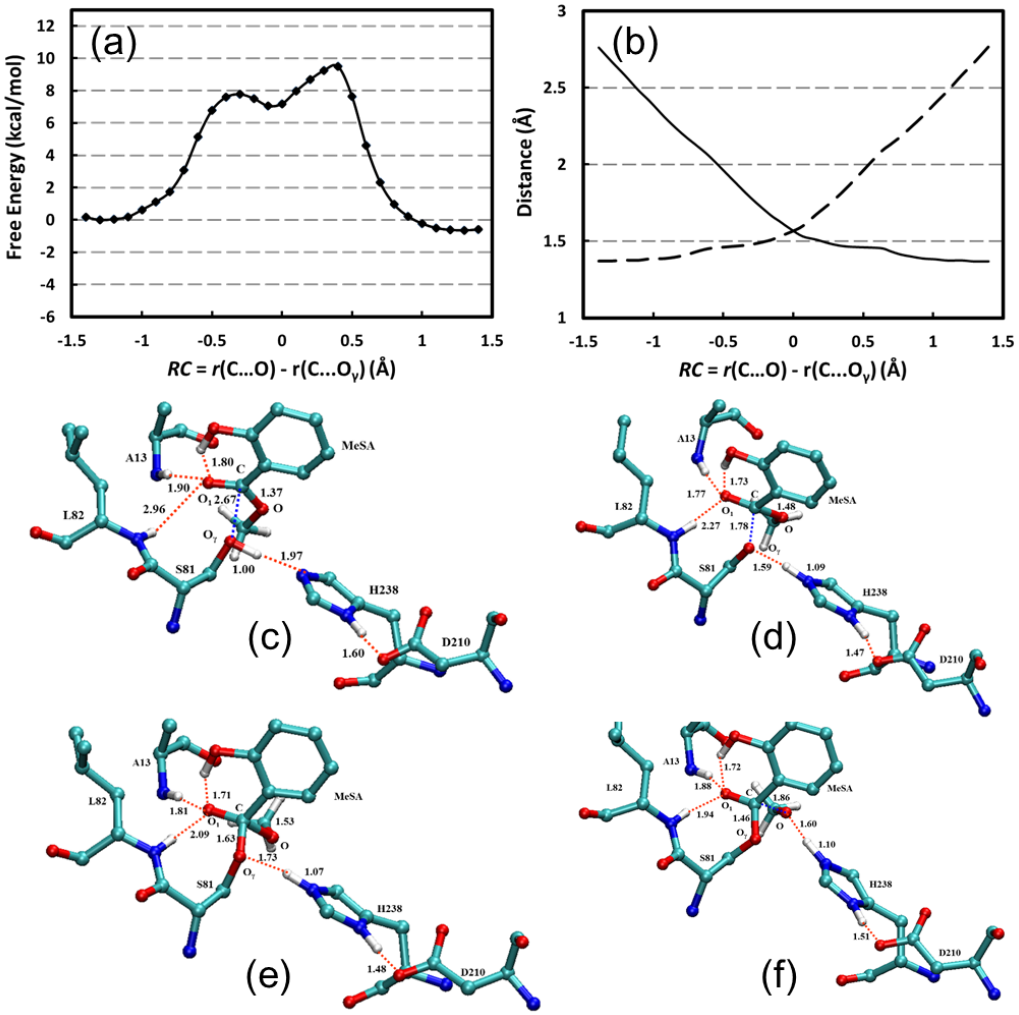


Figure 5.4 PMF calculation results for SABP2 complexed with MeSA with SCC-DFTB/MM.

(a) PMF along the reaction coordinate. (b) Changes of average distances along the reaction coordinate. Solid line stands for r(C…Oγ); dash line stands for r(C…O). (c) Reactant state. (d) Transition state 1; (e) Tetrahedral intermediate; (f) Transition state 2. Unimportant hydrogens are omitted for clarity. Distances are labeled in Å, and energies are in kcal/mol.

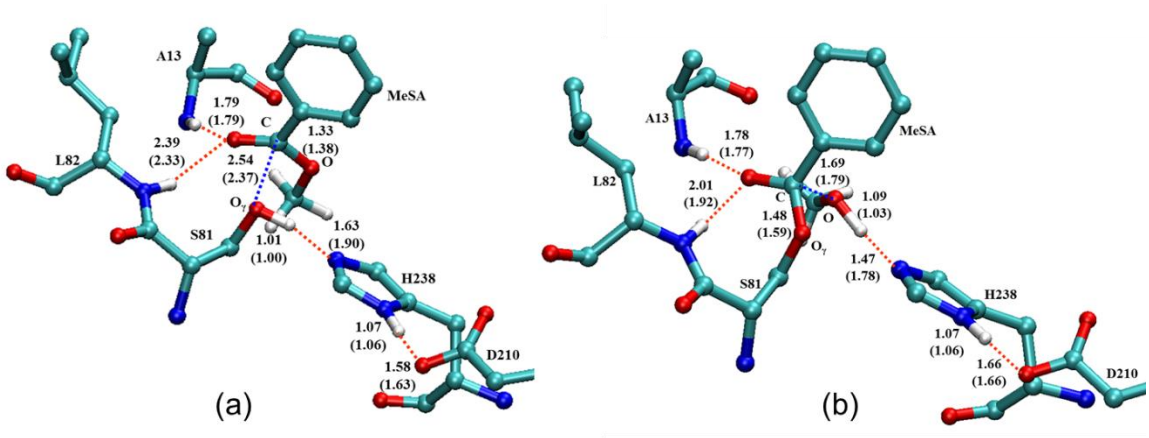


Figure 5.5. Benchmark calculations for SABP2 complexed with MeBA.

Numbers within parentheses are obtained by SCC-DFTB/MM optimization with dispersion correction; those without parentheses are obtained by B3LYP/6-31G(d,p)/MM optimization. Unimportant hydrogens are omitted for clarity. (a) Reactant state; (b) Transition state (TS) by adiabatic mapping. All distances are labeled in Å.

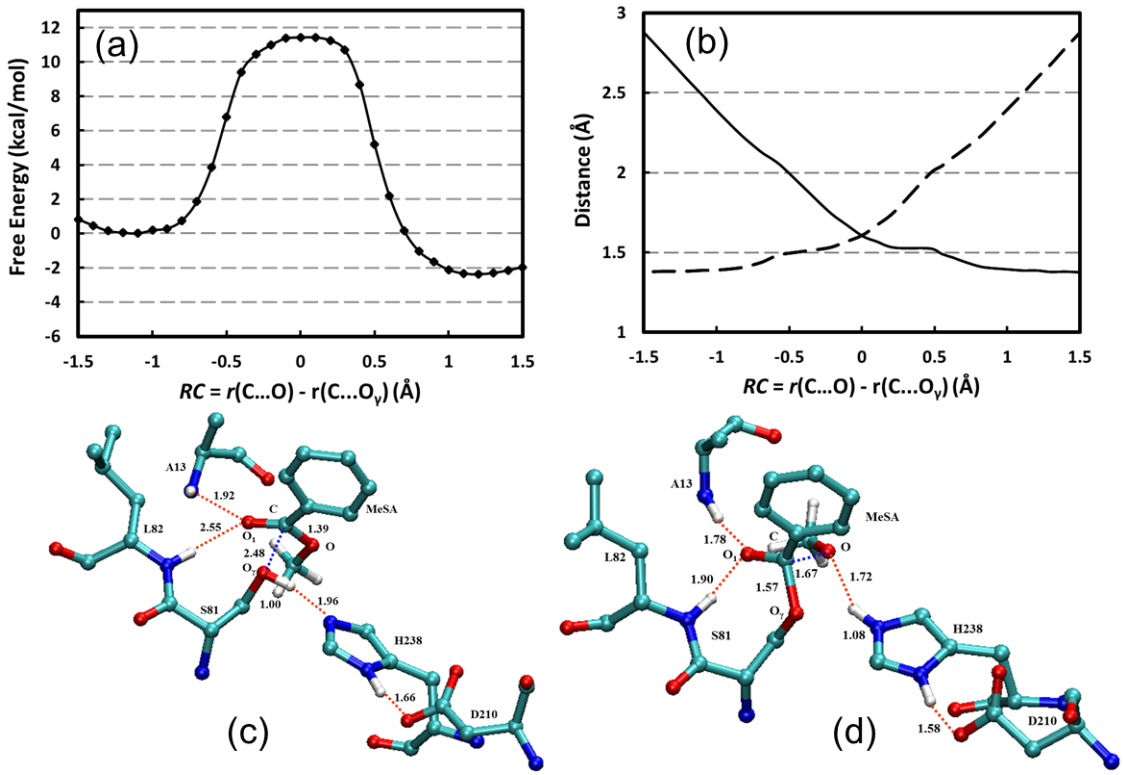


Figure 5.6. PMF calculation results for SABP2 complexed with MeBA with SCC-DFTB/MM.

(a) PMF along the reaction coordinate. (b) Changes of average distances along the reaction coordinate. Solid line stands for r(C…Oγ); dash line stands for r(C…O). (c) Reactant state. (d) Transition state. Unimportant hydrogens are omitted for clarity. Distances are labeled in Å, and energies are in kcal/mol.

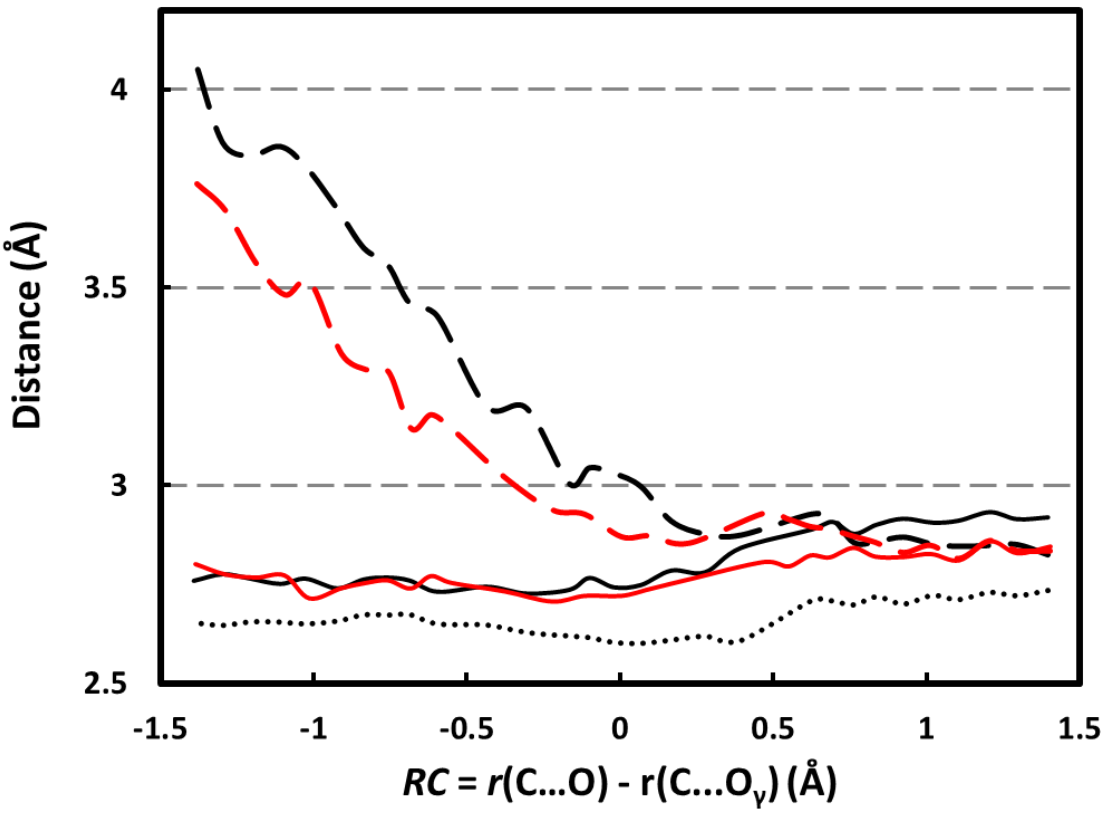


Figure 5.7 Calculated heavy atoms distances (N-O and O-O) between the carbonyl oxygen of substrates and the groups (NH and OH) of oxyanion hole (Ala-13, LEU-81, and MeSA) along the reaction coordinate for the acylation reaction.

Black lines for MeSA complex and red lines for MeBA complex. Solid lines for NH of Ala13, dash lines for NH of Leu-82, and dot line for OH of MeSA.

Table 5.1. Bond Lengths of Reaction Coordinates.

RC related average bond lengths for the RSs, TIs, and TIs of MeSA and MeBA acylation in tobacco SABP2 from QM(SCC-DFTB)/mm free energy simulations (All in Å).

Substrates	Acylation Process	RC[a]	C-O	O_γ-C
MeSA	RS	-1.30 ± 0.05	1.37 ± 0.03	2.67 ± 0.06
	TS1	-0.30 ± 0.06	1.78 ± 0.06	1.48 ± 0.03
	TI	-0.10 ± 0.06	1.63 ± 0.06	1.53 ± 0.04
	TS2	0.40 ± 0.04	1.46 ± 0.03	1.86 ± 0.05
MeBA	RS	-1.10 ± 0.05	1.39 ± 0.03	2.48 ± 0.06
	TS	0.10 ± 0.05	1.67 ± 0.05	1.57 ± 0.04

[a]RC is defined as the difference between C-O and O_γ-C.

Table 5.2. Acylation Free Energy Barriers of SABP2.

MeSA and MeBA acylation barriers in SABP2 from experiments and calculations.

	PtSABP2-1[a]	PtSABP2-2[a]	tbSABP2[a]	TS1[b]	TI[b]	TS2[b]	B3LYP correction[c]
MeSA	16.3	16.0	15.5	7.8 (10.6/4.1)	7.1 (10.3/3.9)	9.5 (17.7/12.2)	14.9
				TS			
MeBA	18.0	17.0	N/A	11.4 (19.5/14.9)			16.1

[a]Activation barriers (kcal/mol) calculated by transition state theory at 298.15K based on K_{cat} value; K_{cat} value are 0.120 and 0.187 for PtSABP2-1 and PtSABP2-2, respectively. K_{cat} value of tbSABP2 is 0.45 S^{-1} from Forouhar *et al*, 2005. [b]Values (without parentheses) are from SCC-DFTB/MM PMF simulations; Values (in parentheses before slash) are from adiabatic mapping calculations with B3LYP/6-31G*/MM; Values (in parentheses after slash) are from adiabatic mapping calculations with SCC-DFTB/MM. [c]Values are PMF results plus a barrier energy correction which is the corresponding barriers difference between the B3LYP/6-31G*/MM and SCC-DFTB/MM levels from adiabatic mapping calculations.

CHAPTER 6

PPCM: COMBINING MULTIPLE CLASSIFIERS TO IMPROVE PROTEIN-PROTEIN INTERACTION PREDICTION

Jianzhuang Yao, Hong Guo, and Xiaohan Yang. “PPCM: Combining multiple classifiers to improve protein-protein interaction prediction.”*To be submitted.*

ABSTRACT

In recent decades, protein-protein interaction (PPI) prediction has become a popular research area and many classifiers have been developed for PPI prediction. However, no single classifier has been able to predict PPI with high-confidence. We postulate that combining individual classifiers could improve PPI prediction accuracy over individual classifiers alone. We developed protein-protein interaction classifiers merger (PPCM) that combines output from two PPI prediction tools -- GO2PPI and Phyloprof using random forests algorithm. The performance of PPCM was tested by area under the curve (AUC) using an assembled gold standard PPI database that contains positive PPI pairs as well as negative PPI pairs. Our AUC test showed that PPCM significantly improved PPI prediction accuracy over corresponding individual classifiers. We found that more classifiers incorporated into PPCM led to better improvement in the PPI prediction accuracy. Furthermore, cross species PPCM achieved competitive and even better prediction accuracy compared to the single species PPCM. This study established a robust pipeline for PPI prediction by integrating multiple classifiers using random forests algorithm. This pipeline will be useful for predicting PPI in non-model species.

Introduction

Protein-protein interaction (PPI) networks play important roles in many cellular activities, including complex formation and metabolic pathways (Gavin, Bosche et al. 2002), and identification of PPI pairs provides insight into the molecular basis of cellular processes (Alberts 1998). Several high-throughput experimental approaches have been developed for PPI identification, including two-hybrid assays (Devos and Russell 2007), tandem affinity purification followed by Mass Spectrometry (Gavin, Aloy et al. 2006), and protein microarrays (Kumar and Snyder 2002). These high-throughput methods have produced a large amount PPI data, which have been accumulated in the public PPI databases, such as DIP (Xenarios, Rice et al. 2000) and STRING (Franceschini, Szklarczyk et al. 2013). However, the results generated by these high-throughput methods lack reliability (von Mering, Krause et al. 2002), and have limited coverage of PPIs in any given organism (Planas-Iglesias, Bonet et al. 2013). Additional experimental information for PPI includes X-ray structures of protein complexes in the PDB databank (Sussman, Lin et al. 1998). Although significant experimental progress has led to a rapid accumulation of protein structural information, PDB data is only available for a small number of proteins. In particular, the information concerning protein complex structures is limited compared with the large volume of protein sequences available in the public databases (Hart, Ramani et al. 2006). Therefore, protein complex information in the PDB databank is not sufficient for large-scale PPI prediction.

To overcome the limitations in PPI identification using experimental methods, computational approaches have been developed to achieve large-scale PPI prediction in various organisms (Chen and Liu 2005, Qi, Klein-Seetharaman et al. 2005, Liu, Kim et al. 2008, Garcia-Garcia, Guney et al. 2010, Yu, Chou et al. 2010, Gallone, Simpson et al. 2011). Traditional input features for PPI prediction are mainly from biological data sources, which can be divided into four categories: Gene Ontology (GO)-based, structure-based, network topology-based, and sequenced-based features (Theofilatos, Dimitrakopoulos et al. 2011). Each individual computational PPI prediction method utilizes only one or a few input sources for PPI prediction. For example, BIPS only takes protein sequences as input for Interolog searching (Garcia-Garcia, Schleker et al. 2012). Bio::Homology::InterologWalk takes protein sequences and well-known PPI networks as input (Gallone, Simpson et al. 2011). Although these methods with single or several features as input can generate fairly accurate results, they are unable to take advantage of other input features that could be helpful for PPI prediction. Thus, machine learning methods (e.g., Bayesian classifiers (Jansen, Yu et al. 2003), Artificial Neural Networks (ANN) (Chen, Liu et al. 2008), Support Vector Machines (SVM) (Gomez, Noble et al. 2003), and Random Forests (Strobl, Boulesteix et al. 2007)) have been developed to integrate multiple features as inputs. Machine learning approaches achieve higher performance than other methods; among them, Random Forests shows the best performance (Qi, Bar-Joseph et al. 2006).

In addition to the progress in identification of informative features for PPI prediction, a variety of algorithms have been developed to improve the PPI prediction accuracy (Theofilatos, Dimitrakopoulos et al. 2011). For instance, Phylogenetic Profiling (PP) uses genome-scale and network-based features as inputs for PPI prediction founded on the assumption that the co-occurrence of two proteins across taxa indicates a good chance for them to function together (GAASTERLAND and RAGAN 1998, Pellegrini, Marcotte et al. 1999). Although PPI prediction by PP has shown good performance in prokaryotes, it is has poor performance in PPI prediction in eukaryotes, because of modularity of eukaryotic proteins, biased diversity of available genomes, and large evolutionary distances (Snitkin, Gustafson et al. 2006, Jothi, Przytycka et al. 2007). Several studies indicate that the accuracy of PPI prediction by PP can be improved by selecting the appropriate reference taxa and matching the reference taxa to the known PPI network (Jothi, Przytycka et al. 2007, Sun, Li et al. 2007, Herman, Ochoa et al. 2011). Recently, Simonsen *et al*. (2012) developed a PPI prediction software PhyloProf (Simonsen, Maetschke et al. 2012) that integrates four PPI prediction methods including the original PP method (Pellegrini, Marcotte et al. 1999), mutual information (MI) method (Date and Marcotte 2003), hypergeometric distribution based method (Wu, Kasif et al. 2003), and the extension of the hypergeometric distribution (RUN) method (Cokus, Mizutani et al. 2007). Also, PhyloProf provides six reference taxa optimization methods including Tree Level Filtering, Iterative Taxaon Selection, Genetic Algorithm, and Tree based search (Singh, Wall et al. 2008, Simonsen, Maetschke et al. 2012). Furthermore, there are four

PPI networks available in PhyloProf, including the networks from *Escherichia coli* (EC)*, Saccharomyces cerevisiae* (hereafter referred to as SC)*, Drosophila melanogaster* (DM)*,* and *Arabidopsis thaliana* (AT). In short, PhyloProf provides a series of PPI prediction classifiers as a result of various combinations of PPI prediction methods, reference taxa optimization methods, and networks from different species.

Another sophisticated PPI prediction software called GO2PPI has been developed to use Gene Ontology and PPI networks as input (Maetschke, Simonsen et al. 2012). By introducing a concept called inducer to combine machine learning and semantic similarity techniques, GO2PPI can provide a series of PPI prediction classifiers that are combinations of machine learning methods (i.e. Naïve Bayes (NB) and Random Forests), GO categories (i.e. biological process (BP), cellular component (CC), and molecular function (MF)), and networks from eight species (*Homo sapiens* (HS)*, Mus musculus* (MM)*, S. pombe* (SP)*, Chlamydia trachomatis* (CT)*,* SC, AT, EC, and DM)*.*

To build off of this research, we developed a pipeline PPCM (i.e., PPI prediction classifiers merger) to enhance the PPI prediction accuracy by merging multiple PPI prediction classifiers using the random forests algorithm. To the best of our knowledge, this study is the first effort to merge multiple classifiers by machine learning for PPI prediction.

Methods

We created a training and a test dataset containing direct interacted protein pairs of yeast for protein-protein interaction (PPI) prediction using a method described Y. Qi *et*

al. (Qi, Bar-Joseph et al. 2006). Briefly, 2865 positive PPI pairs were obtained from the DIP database (Xenarios, Rice et al. 2000). These direct interaction protein pairs were tested to be highly confident PPI pairs by small-scale experiments. Since there was insufficient high-confident negative data (Smialowski, Pagel et al. 2010), negative PPI pairs were generated by randomly pairing proteins followed by removing the positive PPI pairs (Zhang, Wong et al. 2004). Finally, the positive PPI pairs and the negative PPI pairs were combined by a ratio of 1 to 100 into a "gold standard" dataset.

The results of PPI prediction classifiers were used as features of PPCM. Specifically, PhyloProf has three kinds of input parameters, including four PPI prediction methods, eight Reference Taxa Optimization methods, and four PPI networks. Without the time-consuming PPI prediction method 'run', there were 96 different classifiers based on different combinations of parameters provided by PhyloProf (Table 6.2). As mentioned above, GO2PPI has three kinds of input parameters as well, including two machine learning methods, seven GO terms or terms combinations (BP, CC, MF, BPCC, BPMF, CCMF, and BPCCMF), and seven PPI networks. In the same way, there were 98 different combinations of classifiers provided by GO2PPI (Table 6.3). We used combined GO terms in this study, because the best accuracy was achieved by the integration of three GO terms in the GO2PPI paper (Maetschke, Simonsen et al. 2012).

The PPCM pipeline, as illustrated in Figure 6.1, was developed to combine multiple classifiers for enhancing PPI prediction accuracy. Specifically, a protein pair is first evaluated by classifiers provided by PPI prediction software, such as GO2PPI (add

reference here) and Phyloprof (add reference here). Then, the classification scores from individual classifiers are used as input features to generate the final PPI prediction score using random forests algorithm, implemented in the Berkeley Random Forests package (Breiman 2001). GO2PPI has 98 PPI prediction classifiers, among which 14 are SC-related and 84 are not SC-related (cross species) classifiers (Table 6.2). Phyloprof has 98 PPI prediction classifiers, among which 24 are SC-related and 72 not SC-related (cross species) classifiers (Table 6.3).

The aforementioned gold standard PPI database that contains 30,000 PPI pairs with a positive-to-negative PPI ratio of 1:100 was used to evaluate the PPI prediction accuracy. The following measures were used to evaluate PPI prediction results: the true positive rate (TPR, also called sensitivity) -- defined as the ratio of correctly predicted positive PPI pairs among all positive PPI pairs, the true negative rate (TNR; also called specificity) -- defined as the ratio of correctly predicted negative PPI pairs among all negative PPI pairs and the false positive rate (FPR; also called Type I error) -- defined as the ratio of incorrectly predicted PPI pairs among all negative PPI pairs. FPR is one minus TNR. The receiver operating characteristic (ROC) curves were created by plotting TPR versus FPR. The Area Under the Curve (AUC) was used as a measure of the prediction accuracy. The AUC value was calculated using the following equation:

$$AUC = \frac{1}{2}\sum_{k=1}^{n}\big((X_k - X_{k-1})(Y_k + Y_{k-1})\big), \qquad (1)$$

Where, X_k is the FPR at k pair, Y_k is the TPR at k pair in the ranked PPI pair list. The prediction process was repeated 25 times, and the average AUC value was reported. We evaluated the PPI prediction accuracy of PPCMs as well as the classifiers in GO2PPI and Phyloprof using AUC. We introduced three categories of PPCM, including GO2PPI, Phyloprof, and GO2PPI+Phyloprof, with each further divided to three sub-categories: SC, cross-species and all species (i.e., SC plus cross species) (Table 6.1).

Results and Discussion

Using our Gold Standard dataset, the average AUC of the 14 SC-related classifiers in GO2PPI (Table 6.1) was 0.63 and rf|bpcc|SC was the most accurate classifier, with an AUC of 0.64, among these 14 classifiers (Figure 6.2A). The average AUC of the 84 cross-species related classifiers in GO2PPI (Table 6.2) was 0.57 and rf|bpcc|HS was the most accurate classifier, with an AUC of 0.61, among these 84 classifiers (Figure 6.2B). The average AUC of all the 96 (all-species) classifiers in GO2PPI (Table 6.2) was 0.58 and rf|bpcc|SC was the most accurate classifier, with an AUC of 0.64, among these 98 classifiers (Figure 6.2C). The AUCs of PPCMs are 0.70, 0.68, and 0.70 for SC, cross-species and all-species PPCM, respectively (Figure 6.2). These results indicate that PPCM significantly improved PPI prediction accuracy compared with their corresponding classifiers in GO2PPI category.

Compared with the most accurate classifier in GO2PPI category, the cross species PPCM improves AUC by 11%. The improvement of PPCM in SC PPCM was only 9% (Figure 6.2), indicating that the cross species PPCM had better performance than the SC classifier.

The better performance of cross-species PPCM (containing 84 features) than SC PPCM (containing 14 features) suggests that the larger number of features incorporated into PPCM enhanced PPI prediction accuracy in GO2PPI category.

Performance of PPCM in the Phyloprof category

Again, using our Gold Standard dataset, the average AUC of the 24 SC-related classifiers in Phyloprof (Table 6.3) was 0.64 and mi|et|SC was the most accurate classifier, with an AUC of 0.71, among these 24 classifiers (Figure 6.3A). The average AUC of the 72 cross-species related classifiers in Phyloprof (Table 6.3) was 0.61 and mi|et|EC was the most accurate classifier, with an AUC of 0.72, among these 84 classifiers (Figure 6.3B),. The average AUC of all the 96 (all-species) classifiers in Phyloprof (Table 6.3) was 0.62 and mi|et|EC was the most accurate classifier, with an AUC of 0.72, among these 96 classifiers (Figure 6.3C). The AUCs of PPCMs are 0.72, 0.76, and 0.77 for SC, cross-species and all-species PPCM, respectively (Figure 6.3). These results indicate that PPCM significantly improved PPI prediction accuracy compared with their corresponding classifiers in the Phyloprof category. Compared with the most accurate classifier in the Phyloprof category, the cross-species PPCM improves AUC by 6%, while the improvement by SC PPCM is only 1% (Figure 6.2), indicating that the cross species PPCM had better performance in AUC improvement. The better performance of cross-species PPCM (containing 72 features) than SC PPCM (containing 24 features) suggests that more features incorporated into PPCM could enhance PPI prediction accuracy in the Phyloprof category.

After separate evaluation of PPCM in the GO2PPI and Phyloprof categories, we assessed the performance of PPCM in the GO2PPI+Phyloprof category that combined all the classifiers in both GO2PPI and Phyloprof. The AUCs of PPCMs in the GO2PPI+Phyloprof category were 0.83, 0.85, and 0.86 for SC, cross-species and all-species PPCM, respectively (Figure 6.4), significantly higher than those of PPCMs in either GO2PPI or Phyloprof category separately (Figs. 6.2-6.3). Compared with the highest AUCs of individual classifiers in GO2PPI and Phyloprof category, the cross species PPCM improves AUC by 18% and the improvement by SC PPCM was 17% (Figs. 6.2-6.4). These results indicate that PPCM based on all the 194 classifiers from both GO2PPI and Phyloprof could generate more accurate PPI prediction than PPCM based on a fewer number of classifiers in GO2PPI or Phyloprof individually, further supporting the aforementioned premise that more features incorporated into PPCM would enhance PPI prediction accuracy. In summation, based on our combinatorial approach, our cross species PPCM results yield informative predictions that will help build high-quality PPI networks for non-model organisms. Such prediction will be valuable for non-model organisms that lack biological data and PPI prediction software for non-model organisms (Theofilatos, Dimitrakopoulos et al. 2011).

Acknowledgement

Authors wish to thank G. A. Tuskan and T.J. Tschaplinski for providing edits and constructive comments. This research was supported by the Department of Energy, Office

of Science, Genomic Science Program (under award number DESC0008834). Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract Number DE–AC05–00OR22725.

References

Alberts, B. (1998). "The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists." Cell **92**(3): 291-294.

Breiman, L. (2001). "Random Forests." Mach. Learn. **45**(1): 5-32.

Chen, X.-w., M. Liu and Y. Hu (2008). Integrative Neural Network Approach for Protein Interaction Prediction from Heterogeneous Data. Advanced Data Mining and Applications. C. Tang, C. Ling, X. Zhou, N. Cercone and X. Li, Springer Berlin Heidelberg. **5139:** 532-539.

Chen, X. W. and M. Liu (2005). "Prediction of protein-protein interactions using random decision forest framework." Bioinformatics **21**(24): 4394-4400.

Cokus, S., S. Mizutani and M. Pellegrini (2007). "An improved method for identifying functionally linked proteins using phylogenetic profiles." BMC Bioinformatics **8**(Suppl 4): S7.

Date, S. V. and E. M. Marcotte (2003). "Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages." Nat Biotech **21**(9): 1055-1062.

Devos, D. and R. B. Russell (2007). "A more complete, complexed and structured interactome." Current Opinion in Structural Biology **17**(3): 370-377.

Franceschini, A., D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering and L. J. Jensen (2013). "STRING v9.1: protein-protein interaction networks, with increased coverage and integration." Nucleic Acids Research **41**(D1): D808-D815.

GAASTERLAND, T. and M. A. RAGAN (1998). "Constructing multigenome views of whole microbial genomes." Microbial & comparative genomics **3**(3): 177-192.

Gallone, G., T. I. Simpson, J. D. Armstrong and A. P. Jarman (2011). "Bio::Homology::InterologWalk - A Perl module to build putative protein-protein interaction networks through interolog mapping." Bmc Bioinformatics **12**.

Garcia-Garcia, J., E. Guney, R. Aragues, J. Planas-Iglesias and B. Oliva (2010). "Biana: a software framework for compiling biological interactions and analyzing networks." Bmc Bioinformatics **11**(1): 56.

Garcia-Garcia, J., S. Schleker, J. Klein-Seetharaman and B. Oliva (2012). "BIPS: BIANA Interolog Prediction Server. A tool for protein–protein interaction inference." Nucleic Acids Research **40**(W1): W147-W151.

Gavin, A.-C., P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M.-A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.-M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell and G. Superti-Furga (2006). "Proteome survey reveals modularity of the yeast cell machinery." Nature **440**(7084): 631-636.

Gavin, A.-C., M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin,

G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer and G. Superti-Furga (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature **415**(6868): 141-147.

Gomez, S. M., W. S. Noble and A. Rzhetsky (2003). "Learning to predict protein–protein interactions from protein sequences." Bioinformatics **19**(15): 1875-1881.

Hart, G. T., A. Ramani and E. Marcotte (2006). "How complete are current yeast and human protein-interaction networks?" Genome Biology **7**(11): 120.

Herman, D., D. Ochoa, D. Juan, D. Lopez, A. Valencia and F. Pazos (2011). "Selection of organisms for the co-evolution-based study of protein interactions." BMC Bioinformatics **12**(1): 363.

Jansen, R., H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt and M. Gerstein (2003). "A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data." Science **302**(5644): 449-453.

Jothi, R., T. Przytycka and L. Aravind (2007). "Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment." BMC Bioinformatics **8**(1): 1-17.

Kumar, A. and M. Snyder (2002). "Proteomics: Protein complexes take the bait." Nature **415**(6868): 123-124.

Liu, Y., I. Kim and H. Zhao (2008). "Protein interaction predictions from diverse sources." Drug Discovery Today **13**(9-10): 409-416.

Maetschke, S. R., M. Simonsen, M. J. Davis and M. A. Ragan (2012). "Gene Ontology-driven inference of protein–protein interactions using inducers." Bioinformatics **28**(1): 69-75.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates (1999). "Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles." Proceedings of the National Academy of Sciences **96**(8): 4285-4288.

Planas-Iglesias, J., J. Bonet, J. García-García, M. A. Marín-López, E. Feliu and B. Oliva (2013). "Understanding Protein–Protein Interactions Using Local Structural Features." Journal of Molecular Biology **425**(7): 1210-1224.

Qi, Y., Z. Bar-Joseph and J. Klein-Seetharaman (2006). "Evaluation of different biological data and computational classification methods for use in protein interaction prediction." Proteins: Structure, Function, and Bioinformatics **63**(3): 490-500.

Qi, Y., J. Klein-Seetharaman and Z. Bar-Joseph (2005). "Random forest similarity for protein-protein interaction prediction from multiple sources." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing: 531-542.

Qi, Y. J., Z. Bar-Joseph and J. Klein-Seetharaman (2006). "Evaluation of different biological data and computational classification methods for use in protein interaction prediction." Proteins-Structure Function and Bioinformatics **63**(3): 490-500.

Simonsen, M., S. Maetschke and M. A. Ragan (2012). "Automatic Selection of Reference Taxa for Protein-Protein Interaction Prediction with Phylogenetic Profiling." Bioinformatics.

Singh, S., D. P. Wall, S. Singh and D. P. Wall (2008). "Testing the Accuracy of Eukaryotic Phylogenetic Profiles for Prediction of Biological Function." Evolutionary Bioinformatics **4**(EBO-4-Wall-et-al-Supplementary): 0-0.

Smialowski, P., P. Pagel, P. Wong, B. Brauner, I. Dunger, G. Fobo, G. Frishman, C. Montrone, T. Rattei, D. Frishman and A. Ruepp (2010). "The Negatome database: a reference set of non-interacting protein pairs." Nucleic Acids Research **38**(suppl 1): D540-D544.

Snitkin, E., A. Gustafson, J. Mellor, J. Wu and C. DeLisi (2006). "Comparative assessment of performance and genome dependence among phylogenetic profiling methods." BMC Bioinformatics **7**(1): 420.

Strobl, C., A.-L. Boulesteix, A. Zeileis and T. Hothorn (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution." BMC Bioinformatics **8**(1): 25.

Sun, J., Y. Li and Z. Zhao (2007). "Phylogenetic profiles for the prediction of protein–protein interactions: How to select reference organisms?" Biochemical and Biophysical Research Communications **353**(4): 985-991.

Sussman, J. L., D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter and E. E. Abola (1998). "Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules." Acta Crystallographica Section D **54**(6 Part 1): 1078-1084.

Theofilatos, K. A., C. M. Dimitrakopoulos, A. K. Tsakalidis, S. D. Likothanassis, S. T. Papadimitriou and S. P. Mavroudi (2011). "Computational Approaches for the Prediction of Protein-Protein Interactions: A Survey." Current Bioinformatics **6**(4): 398-414.

von Mering, C., R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields and P. Bork (2002). "Comparative assessment of large-scale data sets of protein-protein interactions." Nature **417**(6887): 399-403.

Wu, J., S. Kasif and C. DeLisi (2003). "Identification of functional links between genes using phylogenetic profiles." Bioinformatics **19**(12): 1524-1530.

Xenarios, I., D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte and D. Eisenberg (2000). "DIP: the Database of Interacting Proteins." Nucleic Acids Research **28**(1): 289-291.

Yu, C. Y., L. C. Chou and D. T. H. Chang (2010). "Predicting protein-protein interactions in unbalanced data using the primary structure of proteins." Bmc Bioinformatics **11**.

Zhang, L., S. Wong, O. King and F. Roth (2004). "Predicting co-complexed protein pairs using genomic and proteomic data integration." BMC Bioinformatics **5**(1): 38.

Appendix

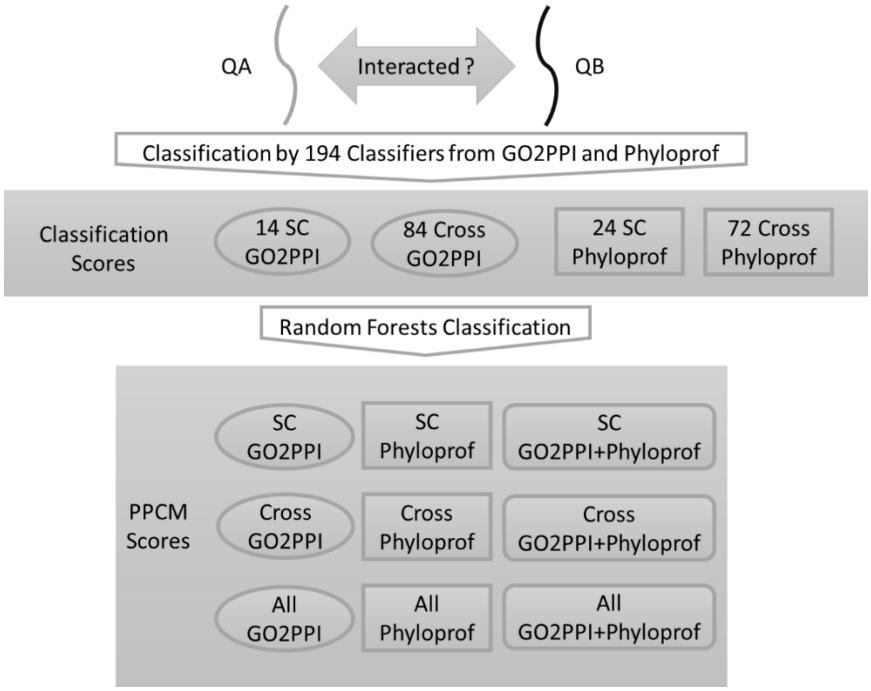


Figure 6.1. The PPCM pipeline.

The PPCM pipeline for protein-protein interaction prediction. Given a pair of query proteins QA and QB, their interaction possibility was first predicted by each of the 194 classifiers from GO2PPI and Phyloprof. Then, the classification scores were merged using random forests algorithm to generate the final PPI prediction score. Nine PPI classification scores were provided by PPCM. “SC” represents PPI networks in *Saccharomyces cerevisiae*. “Cross” represents all PPI networks except SC. “All” represents all the PPI networks in both SC and cross-species.

Figure 6.2. Comparison of PPI prediction accuracy in the GO2PPI category.

(A) PPI prediction based on classifiers related to SC; (B) PPI prediction based on classifiers related to cross-species; and (C) PPI prediction based on classifiers related to all-species. “Average” represents the mean AUC of all the classifiers in each category. “Highest” represents the classifier with highest AUC among all the classifiers in each category. Error bars show standard deviation. “*” indicates that AUC of PPCM was significantly (P-value < 0.05; T-test) higher than that of the most accurate classifier in each category.

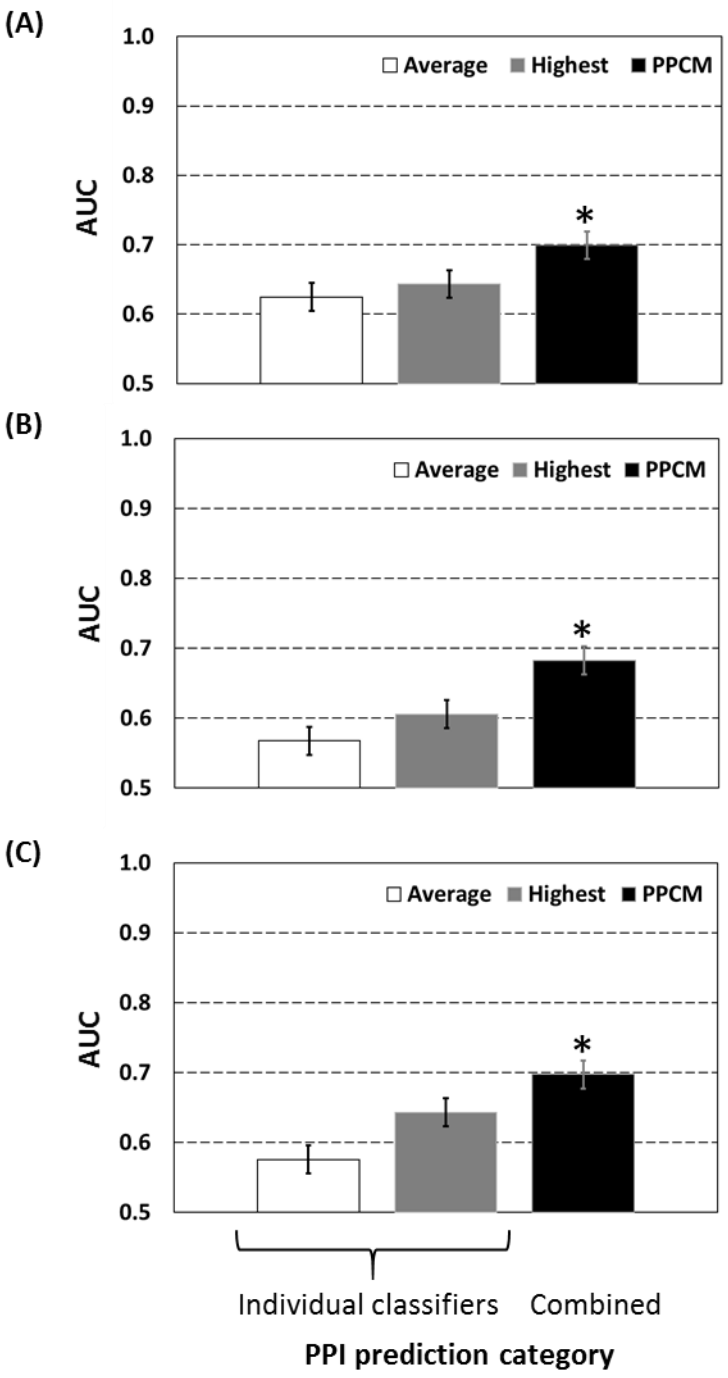
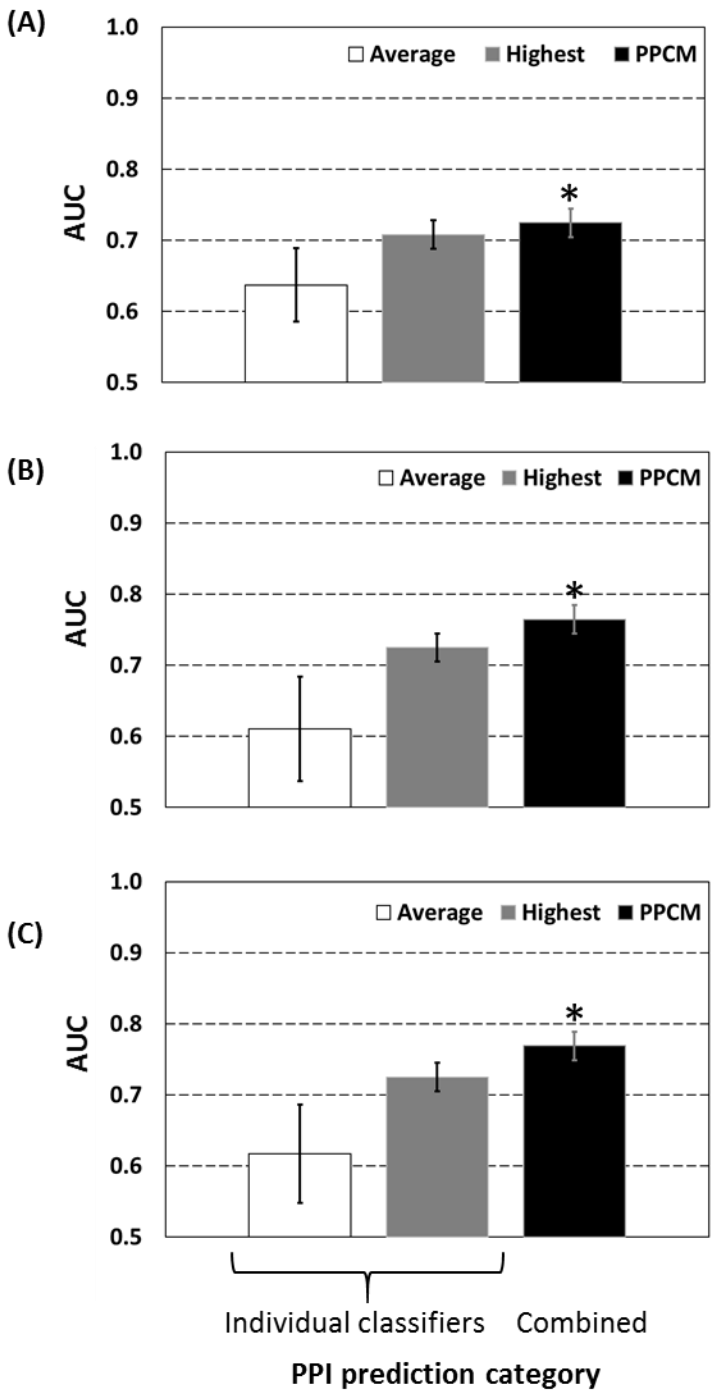

(A)
(B)
(C)
Average
Highest
PPCM
AUC
1.0
0.9
0.8
0.7
0.6
0.5
*
Individual classifiers
Combined
PPI prediction category

Figure 6.3. Comparison of PPI prediction accuracy in the Phyloprof category.

(A) PPI prediction based on classifiers related to SC; (B) PPI prediction based on classifiers related to cross-species; and (C) PPI prediction based on classifiers related to all-species. “Average” represents the mean AUC of all the classifiers in each category. “Highest” represents the classifier with highest AUC among all the classifiers in each category. Error bars show standard deviation. “*” indicates that AUC of PPCM was significantly (P-value < 0.05; T-test) higher than that of the most accurate classifier in each category.


(A)
1.0
0.9
0.8
0.7
0.6
0.5
AUC
Average
Highest
PPCM
*
(B)
1.0
0.9
0.8
0.7
0.6
0.5
AUC
Average
Highest
PPCM
*
(C)
1.0
0.9
0.8
0.7
0.6
0.5
AUC
Average
Highest
PPCM
*
Individual classifiers
Combined
PPI prediction category

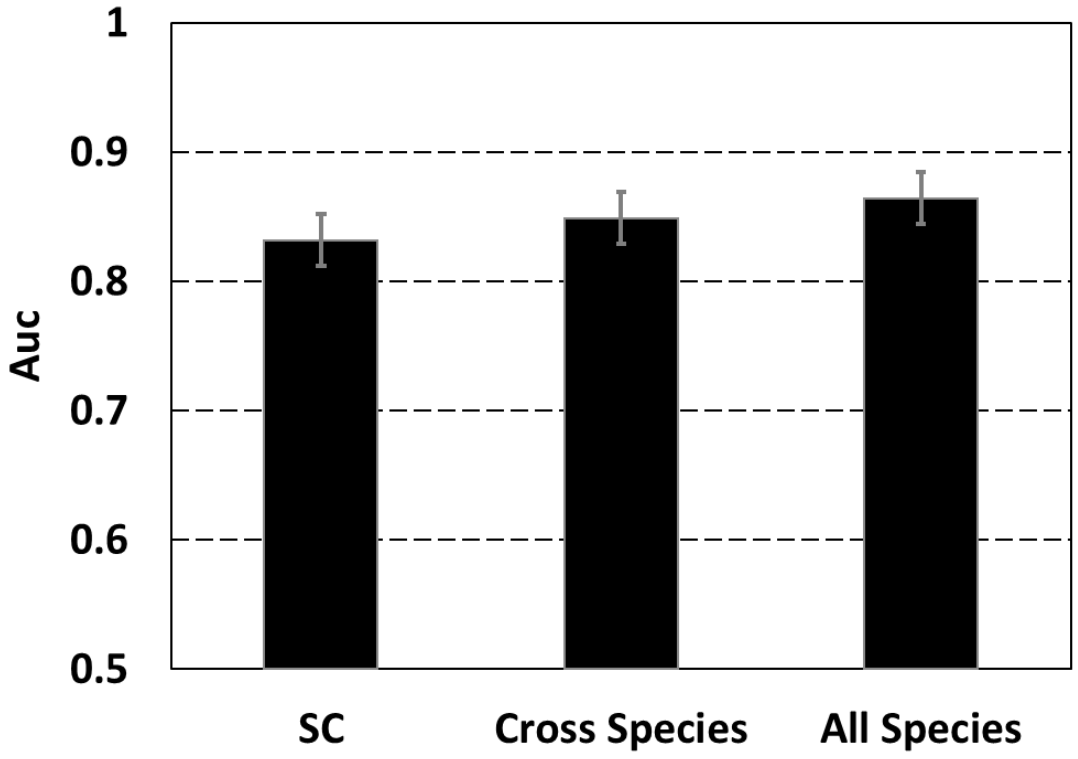


Figure 6.4. Comparison of PPI prediction accuracy in the GO2PPI+Phyloprof category. Error bars show standard deviation.

Table 6.1. Summary of PPCM classifiers.

Random Forests are applied to integrate all of individual classifiers from the following categories.

Software categories	Species categories*	Number of classifiers
GO2PPI	SC	14
	Cross-species	84
	All-species	98
Phyloprof	SC-species	24
	Cross-species	72
	All-species	96
GO2PPI+Phyloprof	SC-species	38
	Cross-species	156
	All-species	194

*SC represents PPI networks in Saccharomyces cerevisiae; cross-species represents all PPI networks except SC; and “all-species” represents all the PPI networks in both SC and cross-species.

Table 6.2. The individual classifiers of GO2PPI merged by PPCM

GO2PPI provides three parameters including machine learning algorithm, gene ontology, and network.

NO.	Machine learning algorithm|GO|NetWork
1	nb|bp|AT
2	nb|bp|DM
3	nb|bp|EC
4	nb|bp|HS
5	nb|bp|MM
6	nb|bp|SC
7	nb|bp|SP
8	nb|bpcc|AT
9	nb|bpcc|DM
10	nb|bpcc|EC
11	nb|bpcc|HS
12	nb|bpcc|MM
13	nb|bpcc|SC
14	nb|bpcc|SP
15	nb|bpccmf|AT
16	nb|bpccmf|DM
17	nb|bpccmf|EC
18	nb|bpccmf|HS
19	nb|bpccmf|MM
20	nb|bpccmf|SC
21	nb|bpccmf|SP
22	nb|bpmf|AT
23	nb|bpmf|DM
24	nb|bpmf|EC
25	nb|bpmf|HS
26	nb|bpmf|MM
27	nb|bpmf|SC
28	nb|bpmf|SP
29	nb|cc|AT
30	nb|cc|DM
31	nb|cc|EC
32	nb|cc|HS
33	nb|cc|MM

Table 6.2. Continued.

NO.	Machine learning algorithm|GO|NetWork
34	nb|cc|SC
35	nb|cc|SP
36	nb|ccmf|AT
37	nb|ccmf|DM
38	nb|ccmf|EC
39	nb|ccmf|HS
40	nb|ccmf|MM
41	nb|ccmf|SC
42	nb|ccmf|SP
43	nb|mf|AT
44	nb|mf|DM
45	nb|mf|EC
46	nb|mf|HS
47	nb|mf|MM
48	nb|mf|SC
49	nb|mf|SP
50	rf|bp|AT
51	rf|bp|DM
52	rf|bp|EC
53	rf|bp|HS
54	rf|bp|MM
55	rf|bp|SC
56	rf|bp|SP
57	rf|bpcc|AT
58	rf|bpcc|DM
59	rf|bpcc|EC
60	rf|bpcc|HS
61	rf|bpcc|MM
62	rf|bpcc|SC
63	rf|bpcc|SP
64	rf|bpccmf|AT
65	rf|bpccmf|DM
66	rf|bpccmf|EC
67	rf|bpccmf|HS
68	rf|bpccmf|MM
69	rf|bpccmf|SC
70	rf|bpccmf|SP
71	rf|bpmf|AT

Table 6.2. Continued.

NO.	Machine learning algorithm|GO|NetWork
72	rf|bpmf|DM
73	rf|bpmf|EC
74	rf|bpmf|HS
75	rf|bpmf|MM
76	rf|bpmf|SC
77	rf|bpmf|SP
78	rf|cc|AT
79	rf|cc|DM
80	rf|cc|EC
81	rf|cc|HS
82	rf|cc|MM
83	rf|cc|SC
84	rf|cc|SP
85	rf|ccmf|AT
86	rf|ccmf|DM
87	rf|ccmf|EC
88	rf|ccmf|HS
89	rf|ccmf|MM
90	rf|ccmf|SC
91	rf|ccmf|SP
92	rf|mf|AT
93	rf|mf|DM
94	rf|mf|EC
95	rf|mf|HS
96	rf|mf|MM
97	rf|mf|SC
98	rf|mf|SP

Table 6.3. The individual classifiers of Phyloprof merged by PPCM

Phyloprof provides three parameters including network, taxa optimization method, and PPI prediction method.

NO.	Network|taxa optimization|PPI prediction
1	AT|hg|et
2	AT|hg|ga
3	AT|hg|gapd
4	AT|hg|gatc
5	AT|hg|ite
6	AT|hg|ot
7	AT|hg|rt
8	AT|hg|to
9	AT|mi|et
10	AT|mi|ga
11	AT|mi|gapd
12	AT|mi|gatc
13	AT|mi|ite
14	AT|mi|ot
15	AT|mi|rt
16	AT|mi|to
17	AT|sb|et
18	AT|sb|ga
19	AT|sb|gapd
20	AT|sb|gatc
21	AT|sb|ite
22	AT|sb|ot
23	AT|sb|rt
24	AT|sb|to
25	SC|hg|et
26	SC|hg|ga
27	SC|hg|gapd
28	SC|hg|gatc
29	SC|hg|ite
30	SC|hg|ot
31	SC|hg|rt
32	SC|hg|to
33	SC|mi|et
34	SC|mi|ga
35	SC|mi|gapd

Table 6.3. Continued.

NO.	Network|taxa optimization|PPI prediction
36	SC|mi|gatc
37	SC|mi|ite
38	SC|mi|ot
39	SC|mi|rt
40	SC|mi|to
41	SC|sb|et
42	SC|sb|ga
43	SC|sb|gapd
44	SC|sb|gatc
45	SC|sb|ite
46	SC|sb|ot
47	SC|sb|rt
48	SC|sb|to
49	HS|hg|et
50	HS|hg|ga
51	HS|hg|gapd
52	HS|hg|gatc
53	HS|hg|ite
54	HS|hg|ot
55	HS|hg|rt
56	HS|hg|to
57	HS|mi|et
58	HS|mi|ga
59	HS|mi|gapd
60	HS|mi|gatc
61	HS|mi|ite
62	HS|mi|ot
63	HS|mi|rt
64	HS|mi|to
65	HS|sb|et
66	HS|sb|ga
67	HS|sb|gapd
68	HS|sb|gatc
69	HS|sb|ite
70	HS|sb|ot
71	HS|sb|rt
72	HS|sb|to
73	EC|hg|et

Table 6.3. Continued.

NO.	Network|taxa optimization|PPI prediction
74	EC|hg|ga
75	EC|hg|gapd
76	EC|hg|gatc
77	EC|hg|ite
78	EC|hg|ot
79	EC|hg|rt
80	EC|hg|to
81	EC|mi|et
82	EC|mi|ga
83	EC|mi|gapd
84	EC|mi|gatc
85	EC|mi|ite
86	EC|mi|ot
87	EC|mi|rt
88	EC|mi|to
89	EC|sb|et
90	EC|sb|ga
91	EC|sb|gapd
92	EC|sb|gatc
93	EC|sb|ite
94	EC|sb|ot
95	EC|sb|rt
96	EC|sb|to

CONCLUSIONS

Because of the importance of enzymatic reactions in life processes, understanding the catalytic mechanism of enzymes has fundamental and practical importance. In addition to biomedical and structural studies (e.g. NMR and X-Ray Crystallography), computer modeling of enzymatic reactions has shown its advantages in understanding the origin of enzymes catalytic power in the recent two decades. However, the applications of computer simulations for enzymes is still lacking in elucidating the catalytic mechanism and catalytic power for many enzymes.

In this dissertation, I studied catalytic mechanisms of several important enzymes including Sedolisins, S-adenosyl-L-methionine(AdoMet)-dependent methyltransferases, Salicylic acid binding protein 2 using QM/MM calculations, MD (both classical and QM/MM) and free energy (potential of mean force) simulations.

In simulations of sedolisin, the deacylation reaction of sedolisin was studied through the QM/MM MD and free energy simulations. The proton transfer processes during the enzyme-catalyzed process were examined and their role in the catalysis were discussed. The simulation results showed that the general acid/base mechanism is more effective than the electrostatic oxyanion hole stabilization in both acylation and deacylation reactions of sedolisin. Asp170 functions as a general acid/base in oxyanion hole stabilization. For pro-kumomalisin, we were able to confirm the hypothesis that the protonation of Asp164 would be able to trigger conformational changes to position this residue as a general acid/base catalyst, thus reconstructing the functional active site.

We have also studied the mechanism of the acylation reaction for the autocatalytic cleavage of the prodomain and obtained the free energy map and catalytic mechanism for the acylation reaction. The results show that the autolysis uses the same catalytic mechanism that was indicated in our earlier studies involving the catalytic domains of kumamolisin-As and sedolisin. One of the interesting questions posed by sedolisins is why this family uses an aspartic acid residue (Asp164) instead of Asn (the residue that creates the oxyanion hole in the classical serine peptidases). The results reported here seem to indicate that one of the reasons might be due to the requirement for the creation of a built-in switch that delays the self-activation of sedolisins until secretion into the acidic medium.

In the simulations of SAMT with the SA and 4HA substrates, we reported that change of the hydroxyl group from the 2-position (as in SA) to 4-position (as in 4HA) could lead to large activity reduction. The free energy simulations reported in this work showed that the free energy barrier difference between SA and 4HA is about 5kcal/mol, which agrees well with experimental study. Simulations showed that the reactant complex with salicylate is close to the corresponding TS structure, while the reactant complex containing 4-hydroxybenzoate is not. The active site binding configuration of the reactant complex containing 4-hydroxybenzoate is significantly distorted from the corresponding TS structure. Thus, additional energy seems to be required to generate the TS-like structure during the methyl transfer process.

In the simulations of SET7/9, different methyl transfer steps from AdoMet to target lysine/methyl-lysine in SET7/9 and its Y245A mutant have been performed. Our results showed that free energy profiles and the energetic triplets could be used to understand and predict the changes of the product specificity as a result of the mutations. The free energy barriers for the methyl transfers obtained here were found to be well correlated with the experimental observations on the change of product specificity due to the Y245→A mutation.

In the simulations of SABP2, we proposed a novel substrate discrimination mechanism in the substrate promiscuous enzyme. SABP2 seems to have been evolved with a less perfect oxyanion hole with the additional contribution coming from the natural substrate MeSA. Our results supported the suggestion that some promiscuous enzymes might have used SAC as one of important strategies for substrate discrimination. Interestingly, we also found two-pronged oxyanion hole provides a hydrogen bonding enhancement mechanism by the amide group of Leu-82 which may partly compensate the energy loss by missing one hydrogen bond.

In addition to catalytic mechanism of enzymes, I developed protein-protein interaction classifiers merger (PPCM) that merged classifiers from two PPI prediction tools called GO2PPI and Phyloprof using random forests algorithm. First, PPCM significantly improved PPI prediction accuracy over corresponding individual classifiers. Second, more classifiers incorporated into PPCM led to better improvement in the PPI prediction

accuracy. Third, cross species PPCM achieved competitive and even better prediction accuracy compared to the single species PPCM.

VITA

Jianzhuang Yao was born in Cangzhou, Hebei, China at 1982. He attended Dalian University, China in 2002 and got his bachelor's degree majoring in Bioengineering in 2006. He then entered Capital Normal University and got his Master's degree majoring in Cell Biology in 2009. He then joined the department of Biochemistry, Cellular and Molecular Biology at the University of Tennessee, Knoxville for his Ph.D. study. He is defending his dissertation in December 2014.