Developing an Enzyme Toolbox for Anti-Tumor Natural Product Biosynthesis

Fengbin Wang

Doctoral Degree of Philosophy: Biochemistry and Cell Biology

APPROVED, THESIS COMMITTEE

George N. Phillips, Jr., Advisor and Chair
Ralph and Dorothy Looney Professor
Department of Biosciences

Yizhi Jane Tao, Committee Chair
Associate Professor
Department of Biosciences

George N. Bennett
E. Dell Butcher Professor
Department of Biosciences

Edward P. Nikonowicz
Associate Professor
Department of Biosciences

K. C. Nicolaou
Harry C. and Olga K. Wiess Professor of Chemistry, Department of Chemistry

HOUSTON, TEXAS
April 12, 2016
ABSTRACT

Natural products, known as chemical compounds or substances produced by living organisms, are arguably the most important players in revolutionizing modern medicine. To date, natural products remain the best source of drug leads, including penicillin, streptomycin, artemisinin and others. However, one of the difficulties of natural product-based drug discovery is the structural derivatization. The traditional chemical modification methods often require multiple protection and deprotection steps for the large variety of functional groups, making the process laborious and problematic. A promising alternative approach to produce natural product derivatives is utilizing biosynthesis enzyme toolbox; for example, enzyme-based "glycol-randomization" and "alkyl-randomization" can dramatically expand diversities of important anti-cancer natural products. In my thesis work, I characterized seven biosynthesis enzymes (including sugar aminotransferase WecE, CalS13, AtmS13, glycosyltransferase SsfS6, OleD, methionine adenosyltransferase sMAT and methyltransferase DnrK) by X-ray crystallography. The structural characterizations and protein engineering of those enzymes lead to successfully expand the glyco/alkyl libraries, as well as to broaden the glyco/alkyl installation processes. These discoveries and enzyme toolbox developments are directly applicable to future drug discovery for cancer, and can be utilized as blue print to further understand the essential role of glycosylation and methylation in biology.
Acknowledgments

“To raise new questions, new possibilities, to regard old problems from a new angle, requires creative imagination and marks real advance in science.” -Albert Einstein

I am extremely thankful to many people for their support throughout my Ph.D. study. I would like to first thank my advisor, Dr. George Phillips, Jr., for giving me the opportunity to join his lab. I am deeply grateful for his mentorship, and the time and effort he devoted to me so that I may be successful. I greatly admire his enthusiasm for research and expertise guidance in crystallography, and I hope I could emulate these traits in my career. I am most appreciative of George’s support in a lot of aspects, and willingness to allow me to explore different scientific interests.

I also want to thank my thesis committee members: Dr. Yizhi Tao, Dr. George Bennett, Dr. Edward Nikonowicz and Dr. K. C. Nicolaou. They provided essential guidance and supports for my Ph.D. training. I am also indebted to my key collaborators and members in their labs, including Dr. Jon Thorson, Dr. Shanteri Singh and Tyler Huber at University of Kentucky, and Dr. John Denu, Dr. Zhangli Su, and Dr. Mike Sussman at University of Wisconsin-Madison. They have provided great supports and invaluable resources for my projects. I also want to give my best respects to the beamline staff and faculties at Argonne National Laboratory for their great support in my remote data collections, especially Dr. Spencer Anderson, Dr.
Zdzislaw Wawrzak and Dr. David Smith at LS-CAT. I also want to thank my undergraduate advisor, Dr. Chaoneng Ji at Fudan University.

I want to thank previous and present members of the Phillips lab, David Xu, Dr. Mitch Miller, Dr. Kate Helmich, Dr. Ragothaman Yennamalli, Dr. Conan Cao, Dr. Craig Bingman, Robert Smith, Dr. He Zhang, Lu Han, Jonathan Clinger, Premila Samuel, Jose Olmos, Jr., Sarah Alvarado, William Ou and Elieen Brady, for their scientific insight and great friendship. Especially I want to thank Conan, Kate and Raghu, who mentored me a lot during my rotation and first-year research, and evolved into an invaluable colleague and friend. I want to thank David, Mitch, Joey, Raghu, Eileen and Jonathan for being awesome friends and team workers on plenty of collaborative projects. I am deeply grateful to the friends I made during graduate school, in both Madison, WI and Houston, TX. You support me on various levels throughout these five years and make me your brother and arms.

Lastly I want to thank my parents for their unconditional love and support regardless of the physical distance between us. I want to thank my wife Zhangli for changing my life completely: being a ninja as scientific companion and being a hello kitty in the rest part of my life. I really look forward to our journey together, as lifetime partner, in the future.
# Contents

**Abstract** ........................................................................................................................................... i

**Acknowledgments** .......................................................................................................................... ii

**Contents** ........................................................................................................................................... iv

**List of Figures** ..................................................................................................................................... vii

**List of Tables** ....................................................................................................................................... x

**Natural Products Drug Discovery** ...................................................................................................... 1

1.1. A Brief History of Natural Product Drug Discovery ........................................................................ 2
1.2. Research of Calicheamicin and its Biosynthesis .............................................................................. 9
1.3. Glycorandomization .......................................................................................................................... 16
1.4. Alkylrandomization .......................................................................................................................... 22

**Materials, Methods and Depositions** ............................................................................................. 38

2.1. Materials ......................................................................................................................................... 38
  2.1.1. UniProt IDs of Target Genes/Proteins ......................................................................................... 38
  2.1.2. Methionine Analogs in sMAT Reactions ..................................................................................... 38
  2.1.3. Cloning, Expression and Purification ........................................................................................... 42
  2.1.4. Crystallization Kits and Equipment .............................................................................................. 43
  2.1.5. Crystal Handling and Shipping .................................................................................................... 44
  2.1.6. Biochemical Assays .................................................................................................................... 44
  2.1.7. Software Packages ...................................................................................................................... 45

2.2. Methods ......................................................................................................................................... 46
  2.2.1. Cloning Methods ........................................................................................................................ 46
  2.2.2. Expression Methods ................................................................................................................... 53
  2.2.3. Purification Methods ................................................................................................................ 56
  2.2.4. Crystallization Methods and Conditions .................................................................................... 61
  2.2.5. X-ray Diffraction Data Collection ............................................................................................... 63
  2.2.6. Structure Determination Methods ............................................................................................... 63
2.2.7. Ensemble Refinements ........................................................................................................................................64
2.2.8. Phosphate Release Assay ....................................................................................................................................64
2.2.9. Thermal Shift Assay ............................................................................................................................................65
2.2.10. Mutagenesis .....................................................................................................................................................65
2.3. PDB Depositions ..................................................................................................................................................67

**Glycorandomization: Sugar Aminotransferases and Glycosyltransferases** .........................................................72

3.1. Structural Basis for the Amino Installation between CalS13 and WecE .................................................................73
3.1.1. Aminosugars and Sugar Aminotransferases .......................................................................................................74
3.1.2. Overall Structures ..............................................................................................................................................75
3.1.3. Cofactor and Nucleotide Sugar Binding Sites .....................................................................................................81
3.1.4. A Structure-based Model for the Stereoochemical Outcome ..............................................................................84
3.2. Structural characterization of Sugar Aminotransferase AtmS13 ...........................................................................90
3.2.1. AT2433 and Sugar Aminotransferase AtmS13 .....................................................................................................91
3.2.2. Overall Structures ..............................................................................................................................................94
3.2.3 Cofactor Binding Site Comparison with CalS13 .................................................................................................97
3.3. Crystal structure of C-glycosyltransferase SsfS6 .....................................................................................................100
3.3.1. Natural Product Glycosyltransferase SsfS6 .........................................................................................................101
3.3.2. Overall Structure and Active Site Analysis of SsfS6 ........................................................................................104
3.3.3. Proposed Mechanism of SsfS6 ..........................................................................................................................108
3.4. Crystal structure of glycosyltransferase OleD ........................................................................................................111
3.4.1. Overall Structure of Glycosyltransferase OleD in the Apo Form .......................................................................112
3.4.2. Ensemble Refinements of Apo OleD and Ligand Bound OleD ........................................................................114

**Alkylrandomization: Methionine Adenosyltransferases and Methyltransferases** .............................................126

4.1. Understanding molecular recognition of promiscuity of thermophilic sMAT ..............................................127
4.1.1. Introduction of Methionine Adenosyltransferase ..............................................................................................128
4.1.2. Overall Structural Organization of sMAT ..........................................................................................................133
4.1.3. Active Site Contents of sMAT ..........................................................................................................................138
4.1.4. Unusual Product Formation during Crystallization .........................................................................................143
4.1.5. Structural Homology .......................................................................................................................................145
4.1.6. Conclusions of wt-sMAT structures .................................................................................................................149
4.1. Preliminary engineering of sMAT to expand Ado-X library ........................................ 150
4.2. Structural Studies of Carminomycin O-Methyltransferase DnrK ................................ 156
4.2.1. Functional AdoMet Isosteres Resistant to AdoMet Degradation .............................. 157
4.2.2. Complex Structure of DnrK with Unnatural Ado$^1$Hcy and CINP ............................ 162
4.2.3. Complex Structure of DnrK with Unnatural 4-methylumbeliferone .......................... 168

Future Perspectives ............................................................................................................ 178
5.1. Future Perspectives: Glycerandomization .................................................................... 179
5.2. Future Perspectives: Alkylrandomization ................................................................... 183
5.2.1. Further Engineering of sMAT to Expand Ado-X Library ........................................ 185
5.1.2. Further Engineering of DnrK .................................................................................. 186

Appendix A ......................................................................................................................... 190
A.1. Human KDM4s and Their Reader Domains ................................................................. 192
A.2. Systematic Profiling Reveals H3K23me3 Binding Specificities ................................. 194
A.3. Structural Basis for H3K23me3 Recognition ............................................................... 197
A.4. H3K23me3 Recognition Mechanism .......................................................................... 203
A.5. KDM4B-DTD as an H3K23me3 specific reader domain ............................................... 204
A.6. H3K23me3 is enriched in primary spermatocytes ....................................................... 210
A.7. H3K23me3-binding stimulates H3K36 demethylation by KDM4B ............................... 213
A.8. Conclusion highlights ................................................................................................. 218

Appendix B ......................................................................................................................... 227
B.1. Protein Dynamics of Cupin Protein from P. luminescens .............................................. 228
B.2. Structural Dynamics of CalE6 from Calicheamicin Biosynthesis ............................... 233
B.3. Structural Analysis of Enediyne Self-sacrifice Proteins CalU16 and CalU19 ............ 242
List of Figures

Figure 1-1. Natural products drug discovery before 21st century ........................................5
Figure 1-2. Challenges of natural products drug discovery after 21st century ......................7
Figure 1-3. Selected events in natural products drug discovery at 2015 ..............................10
Figure 1-4. The history events of calicheamicin research ..................................................12
Figure 1-5. Biosynthesis studies of calicheamicin gene cluster ........................................15
Figure 1-6. Natural product “glycorandomization” .........................................................18
Figure 1-7. Examples of glycorandomization .....................................................................21
Figure 1-8. Natural product “alkylrandomization” ...........................................................23
Figure 1-9. Examples of alkylrandomization .....................................................................26
Figure 2-1. Examples of Primer Design ..............................................................................47
Figure 2-2. Q5 Polymerase and Gibson Assembly .............................................................50
Figure 2-3. DNA sequencing peaks from Lone Star and SeqWright .................................52
Figure 2-4. Expression examples from different systems ..................................................55
Figure 2-5. Protein FPLC purification ..............................................................................59
Figure 2-6. General process of protein cloning, expression and purification .................60
Figure 2-7. Enzchek phosphate release assay ....................................................................66
Figure 2-8. A glance of all my PDB depositions during Ph.D. thesis ...............................69
Figure 3-1. Schemes of SATs including CalS13 and WecE ..............................................76
Figure 3-2. Overall structures and ligand maps of CalS13 and WecE .............................79
Figure 3-3. Cofactor and nucleotide sugar binding sites of CalS13 and WecE ................82
Figure 3-4. Structure based sequence alignment of all SATs with known structures .......86
Figure 3-5. Universal model for the SAT-catalyzed amine installation ............................87
Figure 3-6. AtmS13 related compounds ..........................................................................93
Figure 3-7. Overall and active architecture of protein AtmS13 vs CalS13 ..........95
Figure 3-8. Structure and putative mechanism of SsfS6 ..................................103
Figure 3-9. Dali based multiple sequence alignment of SsfS6 with other GTs .............106
Figure 3-10. Tetracycline docking results by AutoDock 4.2 ....................................110
Figure 3-11. OleD structures shown in static screenshot and dynamics form ..............113
Figure 4-1. Natural and unnatural reactions catalyzed by sMAT .............................131
Figure 4-2. The structure-based sequence alignment of MATs ................................132
Figure 4-3. The overall structure and active site contents of sMAT ..........................135
Figure 4-4. HPLC traces for representative sMAT reactions ................................136
Figure 4-5. Active site interpretation in the simulated annealing omit map ...............144
Figure 4-6. Fold and active site comparisons between sMAT, hMAT2A and eMAT .....147
Figure 4-7. Structure and preliminary assays of sMAT-N159G ...............................153
Figure 4-8. Utilization and degradation of AdoMet ..............................................160
Figure 4-9. Synthesis of AdoMet analogs by Thorson lab ..................................161
Figure 4-10. DnrK colorimetric assay and ligand-bound structures .......................163
Figure 4-11. Electron density maps of DnrK-Ado\textsuperscript{Hcy} and DnrK-CINPs ..........166
Figure 5-1. Summary of glycorandomization ......................................................180
Figure 5-2. Summary of alkylrandomization .........................................................184
Figure A-1. Distinct binding specificities of human KDM4A-C DTD domains ...........195
Figure A-2. Co-crystal structure of KDM4A-DTD with H3K23me3 part-I ...............199
Figure A-3. Co-crystal structure of KDM4A-DTD with H3K23me3 part-II ...............200
Figure A-4. H3K23me3 recognition is modulated by unique side-chain interactions ...206
Figure A-5. KDM4B-DTD is an H3K23me3-specific reader part-I ..........................207
Figure A-6. KDM4B-DTD is an H3K23me3-specific reader part-II ..........................208
Figure A-7. H3K23me3 is enriched in heterochromatin of primary spermatocytes .....211
Figure A-8. H3K23me3 binding stimulates H3K36 demethylation in KDM4B ..........216
Figure A-9. H3K23me3-mediated demethylation by KDM4B ...........................................217
Figure A-10. Sequence alignment of double-tudor domains .................................................222
Figure B-1. Structure of protein cupin plu4264 ....................................................................230
Figure B-2. Real-space correlation coefficient and B-factor plot of cupin homologs ................231
Figure B-3. Dynamics modeling and analysis of CalE6 ........................................................238
Figure B-4. Analysis of CalE6 ensemble refinement models ..................................................240
Figure B-5. Structural comparison between CalU16 and CalC .............................................245
List of Tables

Table 2-1. Methionine analogs used in chapter 4 .................................................................39
Table 2-2. Crystallization and cryoprotectant conditions .................................................62
Table 3-1. Crystallography statistics of WecE and CalS13 structures .................................78
Table 3-2. Crystallography statistics of AtmS13 structures ...............................................96
Table 3-3. Crystal Parameters, data Collection, and refinement of SsfS6 .........................105
Table 3-4. Statistics of static and dynamic OleD structures ...............................................115
Table 4-1. Data collection and refinement statistics for sMAT ...........................................137
Table 4-2. Kinetic parameters of sMAT enzyme .................................................................139
Table 4-3. Ligand-based alignment of active site residues (sMAT and eMAT) ...............140
Table 4-4. Crystallography statistics of sMAT-N159G .......................................................154
Table 4-5. Kinetic parameters and stability measurements ..............................................164
Table 4-6. Crystallography statistics of DnrK structures ..................................................167
Table 4-7. Crystallography statistics of DnrK-AdoHcy/4MU ..........................................169
Table A-1. Data collection and refinement statistics of all KDM4s .................................201
Table B-1. Refinement statistics of CalE6 ........................................................................239
Microorganisms have yielded natural products that are arguably the most important in revolutionizing modern medicine. To date, natural products remain the best source of drug leads. However, one of the shortcomings of natural product-based drug discovery program is the difficulty in structural derivatization. This chapter will first describe a brief history of natural product drug discovery, including a few milestone events, bottlenecks since the new century, recent advances, and an example of natural product drug discovery that relates to my thesis research. Then, this chapter will introduce two approaches “glycorandomization” and “alkylrandomization”, that can contribute to making new natural product derivatives. The results of those two parts will be described in chapters 3 and 4, respectively.
1.1. A brief history of natural product drug discovery

Natural products are chemical compounds or substances produced by a biological source\(^1\). In medicinal chemistry, the definition of natural products is often restricted to secondary metabolites that might not be necessary for survival, but provide an evolutionary advantage to the producing organism\(^2\text{-}^5\). Because of this, secondary metabolites natural products are usually cytotoxic drug-like molecules and have enormous chemical and structural diversity that have been selected and optimized by nature through evolution. The living organisms use them as a "chemical arsenal" against other competing organisms. In the past century, the history of medicine is full of inspiring stories of natural product discovery and their applications to human therapy\(^6\text{-}^20\). In summary, despite different definitions of a natural product in different fields, a natural product is defined as a biologically or pharmacologically active chemical compound or substance that is found in nature and produced by a living organism.

The impact of natural products on human life started thousand years ago. Since ancient times, people used a variety of plants and herbal extracts to treat diseases. One of the most inspiring medical books is called the "Compendium of Materia Medica", written by Shizhen Li in 1578 during the Ming dynasty (Figure 1-1A). It has been regarded as the most complete and comprehensive medical book ever written in the history of traditional Chinese medicine. It lists 1,892 entries containing all the items that were believed to have medicinal properties. Some of them are still
considered effective to certain diseases and are grown in herbal plantations in China (Figure 1-1A).

After entering the 20th century, herbal medicines were gradually replaced by purified compounds. The reason for this is not only the technical advances of compound isolation and purification, but also several inevitable shortcomings of herbal medicines. First herbal medicines contain a large variety of different compounds. Among those compounds, one or several of them may have biological activity, while there is a significant risk of side effects and toxicity of taking those other unknown compounds. Second, the active component usually presents in a small amount, so the pure compound is in theory more active than herbals. Last but not least, herbal medicines may be interacting with prescribed medicines and since the compound components are unknown, there is no regulations or control of this matter and their uses. Surprisingly, some countries including China still approve herbal medicines for use in all kinds of disease treatments. The side effect description in most of the commercialized Chinese herbal medicines is “unknown for now”, and this description has never been changed or updated. In my point of view, this is absolutely wrong and misleads a lot of people to believe herbal medicines don’t have side effects because they are from nature. In the past decades, several Chinese herbal medicines were reported to have severe renal toxicity or liver toxicity, and banned by other countries. Nevertheless, herbal medicines and other natural sources are still extremely important resources to discover and design new drugs.
During the 20th century, natural product drug discovery entered a golden age. Two of the most early milestone events are the discovery of penicillin and streptomycin. Penicillin antibiotics were among the first medications to be effective against variety kinds of bacterial infections and they are still widely used today. The story of penicillin discovery might be one of the best-known science stories. Briefly, penicillin was discovered by Scottish scientist Alexander Fleming in 1928 (Figure 1-1D). In the late 1930s and early 1940s, penicillin was applied to medical use after great efforts from Australian scientist Howard Florey and a team of researchers led by Ernst Chain. In 1945, Both Flory and Chain shared the Nobel Prize in Medicine with Fleming for their work\textsuperscript{21-31}. Streptomycin, a bactericidal aminoglycoside antibiotic, was the first effective treatment for tuberculosis. It was isolated from the actinobacterium \textit{Streptomyces griseus} in 1943 by Albert Schatz, a graduate student in the Selman Waksman lab (Figure 1-1E). In 1952, the Nobel Prize in Physiology or Medicine was awarded to Waksman for this discovery\textsuperscript{32-37}.

Inspired by the discovery and application benefits of penicillin and streptomycin, the intensive screening studies of unexplored bacteria strains, fungi, plants, and other resources resulted in a huge expansion of FDA-approved natural product medicines (Figure 1-1B), especially new antibiotics like tetracycline\textsuperscript{38}, rifamycin\textsuperscript{39}, lipopeptide\textsuperscript{40} and chloramphenicol\textsuperscript{41}. To date, most of FDA-approved antibiotics and antitumor drugs are natural products and their derivatives (Figure 1-1C).
Figure 1-1. Natural product drug discovery before 21st century*

(a) One page glance of the “Compendium of Materia Medica” and herbal plantation. (b) The cumulative percentage of FDA-approved natural products, separated by environmental sources42. (c) Natural products are drug leads in antibiotics and antitumor18. (d) Alexander Fleming. (e) Selman Waksman

*Figure 1-1A is adapted from http://www.cang5000.com, http://www.cnbat.cn, and http://www.shutterstock.com. The data in Figure 1-1B and 1-1C is referred from the citation [42] with the permission from Elsevier, and [18] with the permission from the Nature Publishing Group, and figures was remade. Figure 1-1D and E are from https://en.wikipedia.org.
After entering the 21st century, however, more concerns have been raised for natural product based drug discovery. First, although there are still ~23 natural product drugs in phase III experiments (Figure 1-2A), it is a fact that the discovery of new natural products is more and more difficult since 2000 and most of the 23 drugs in phase III were initially discovered between 1986-1995\textsuperscript{43}. Also, many pharmaceutical companies have abandoned or reduced their natural product based programs, and switched to high-throughput screening or fragment-based drug discovery,\textsuperscript{44} utilizing a huge library of small molecules (Figure 1-2B). Last but not the least, natural products present dramatic structural diversities that have been evolutionarily optimized. Figure 1-3C shows select historical milestones accomplished by the K.C. Nicolaou group\textsuperscript{45} in the total synthesis of several extremely important natural products, including the most potent antitumor compound calicheamicin $\gamma_1$. Due to the complexity of the natural product structures, there are historical difficulties in chemical total synthesis and making derivatives of known natural products.

In 2015, the Nobel Prize in Physiology or Medicine was awarded for natural product discoveries again. Half of the prize was awarded to William C. Campbell and Satoshi Ōmura for their discoveries of the microbial natural product avermectin to treat infections caused by roundworm parasites; the other half was awarded to Youyou Tu for her discoveries of the plant natural product artemisinin to treat Malaria (Figure 1-3A). This award has greatly inspired and brought optimism to the natural product community.
Figure 1-2. Challenges of natural product drug discovery after 21st century

(a) Worldwide approved drug compounds between 2000-2013\textsuperscript{43}. (b) Scheme of high-throughput screening and fragment-based drug discovery\textsuperscript{44}. (c) selected molecules synthesized in the K.C. Nicolaou group\textsuperscript{45}.

\textsuperscript{43}Figure 1-2A and B are adapted from citation [43] with permission from the Chinese Chemical Society (CCS). [44] with permission from the Elsevier.
Briefly, artemisinin was extracted in the early 1970s by a research group led by Youyou Tu. During the Vietnam War, the North Vietnam troops were haunted by Malaria, so they turned to China for help. China was also in a very special period called “Culture Revolution” at the time, but with a direct order from the president to help his alliance, a project named 523 was able to start gathering 60 academic and military institutions. They screened ~40,000 known compounds, compiled 200 traditional Chinese medicine recipes consisting of more than 2,000 herbs, and yet yielded nothing of practical value. Among those results, there is one plant, named *Artemisia annua L.* (Qinghao), showed promising anti-malaria activity in the mouse model, but the results cannot be reproduced. After Youyou Tu led the group, they found another note from “Handbook of Prescriptions for Emergencies” written by Hong Ge at ~300 AD. It described the use of *Artemisia annua* as "Soaking one bunch of Qinghao in 2 liters of water, squeezing, and then drinking the juice". Then Tu realized that the traditional boiling of Chinese herb medicine preparation might be deactivating the effective components. So they tried to extract at lower temperatures with diethyl ether and this finally worked out (Figure 1-3A).

Another exciting story was published in 2015 by Ling *et al* at *Nature*\(^4^6\), in which they developed several novel culture methods including “iChip” to grow soil bacteria that were uncultured previously (Figure 1-3B, adopted and edited from the original paper). They got ~10,000 extracts and screened against *S. aureus*. From their findings, the most notable is a new antibiotic termed as teixobactin, which can inhibit cell wall
biosynthesis by binding to a highly conserved region of important precursor molecules (Figure 1-3B). Together with the other recent developments in microbial genomics, studies of natural product biosynthesis, bioinformatics, and analytical technologies, and novel bacteria culture technology, natural product drug discovery certainly will continue to play extremely important roles in many kinds of therapies.

1.2. Research on calicheamicin and its biosynthesis

As discussed in chapter 1.1, a lot of discoveries of natural products changed the world dramatically. One example that is closely related to my thesis is a compound named calicheamicin. I first read the story of calicheamicin discovery in the book “Molecules that changed the world” written by K. C. Nicolaou and Tamsyn Montagnon and a paper published in JACS at 1993 by Dr. Nicolaou\textsuperscript{47}. Briefly, in 1973, a brilliant chemist Robert Bergman reported a unique reaction, which is latterly called Bergman cycloaromatization or Bergman reaction\textsuperscript{48}. In this reaction a unique group named “enediyne”, which consists two carbon-carbon triple bonds and one carbon-carbon double bond, can transfer into a di-radical form under high temperature and thus abstract hydrogen atoms from neighboring environments (Figure 1-4A). At that time, neither he, nor any other people could have imagined to find this sort of compound could exist in nature.
Figure 1-3. Selected events in natural products drug discovery at 2015

(a) The 2015 Nobel Prize in Physiology or Medicine was awarded to Campbell, Ōmura and Tu for their discoveries of the natural product avermectins and artemisinin. (b) Discovery of a new natural product antibiotic - teixobactin by a collaboration of four institutes in the US and Germany.

Figure 1-3A is adapted from official Nobel Prize site, http://www.nobelprize.org and http://www.23yy.com. Figure 1-3B is adapted and reorganized from citation [46] with permission from the Nature Publishing Group.
In late 1980s, a scientist was on vacation and driving on a highway in northwest San Antonio, Texas. Somehow, she or he decided to stop, pick up a red pebble (caliche) and send it to Lederle Laboratories. From this pebble, calicheamicin produced by *Micromonospora echinospora* ssp. was isolated. The discovery and the structure of calicheamicin (Figure 1-4B) was reported in 1987\(^\text{50}\). The total chemistry synthesis of calicheamicin is accomplished by a brilliant scientist, K.C. Nicolaou, and his group (Figure 1-4C) in 1992\(^\text{47,51}\). Even for an extremely knowledgeable chemist like Dr. Nicolaou, his first reaction when he was shown the structure of calicheamicin was that he was speechless and could not believe the structure\(^\text{47}\). Figure 1-4D shows the DNA cleavage mechanism of calicheamicin and the proposed resistance mechanism by the self-sacrifice of protein CalC\(^\text{52-55}\). For DNA cleavage, firstly the oligosaccharide part of calicheamicin binds to the DNA minor groove (sequence preference of cytosine and thymine), then the tri-sulfite part sets the stage for the Bergman reaction by a nucleophilic cleavage and brings atoms closer, and at the end the enediyne “warhead” forms the di-radical by cyclization and abstract hydrogen atoms from chromosomal DNA and thus leads to the oxidative cleavage of DNA (Figure 1-4D). To avoid suicide after making this warhead, several proteins (CalC, CalU16 and CalU19) have been identified and characterized in *M. echinospora*. Those proteins can bind to, and be cleaved by calicheamicin by a self-sacrifice mechanism (Figure 1-4D), and therefore guard different important regions of their own cells of *M. echinospora*. More details about those three suicide proteins can be found in Appendix B.3.
Figure 1-4. The discovery, total chemistry synthesis, toxicity mechanism and drug application of calicheamicin

(a) Bergman cycloaromatization. (b) Structure of calicheamicin γ1. (c) The Nicolaou group that finished the total synthesis of calicheamicin47. (d) toxicity mechanism of calicheamicin and CalC self-suicide mechanism to neutralize calicheamicin52. (e) Two antibody-conjugated calicheamicins have been developed into clinical drugs56.

*Figure 1-4C is adapted from citation [47] with permission from Dr. Nicolaou. Figure 1-4D is adapted from citation [52] with permission from Royal Society of Chemistry. Figure 1-4E is adapted from citation [56] with permission from Annual Reviews.
Due to the extremely strong cytotoxicity of calicheamicin, its clinical application was limited until scientists conjugated it onto a cancer cell specific antibody anti-CD33 (Figure 1-4E)\(^{56}\). This was the first case of antibody-drug conjugates (ADCs), which are now commonly used as targeted therapies for cancer treatment. Unlike traditional chemotherapy, ADCs in theory can target and mostly kill cancer cells over regular cells. It is more like a powerful missile targeting cancer cells with the specific antigen, rather than a bomb to your whole body. The mechanism of antibody varies, but mostly starts with ADC-antigen binding, then the ADCs are imported into the cell, antibodies are degraded in lysosome, the cytotoxic components are released to damage the cancer cells, and at the end this leads to cell death (Figure 1-4E). To date, two calicheamicin-ADCs have been developed: gemtuzumab ozogamicin (Mylotarg, GO) and inotuzumab ozogamicin. Mylotarg targets CD33\(^ {57-66}\) and was used to treat AML (acute myeloid leukemia). Mylotarg was approved by the FDA in 2001 and voluntarily withdrawn by Pfizer in 2010. However, some other regulatory authorities did not agree with the FDA decision for multiple reasons\(^ {56,67-70}\). Inotuzumab ozogamicin was also developed by Pfizer to treat non-hodgkin lymphoma, and terminated at phase III trial due to poor enrollment.

The structure of calicheamicin is so phenomenal and complicated, and even for a great chemist like Dr. Nicolaou, it took his group six years to battle with the total synthesis\(^ {47}\). Not to mention how hard it would be to make a batch of calicheamicin structural derivatives. Since scientists know that nature already presents a set of
enzymes that are able to make calicheamicin, scientists can try to study and understand their mechanism and structures, and maybe even modify them to let them make calicheamicin derivatives. Dr. Jon Thorson and his group initiated this by identification and characterization of the gene cluster of calicheamicin biosynthesis (Figure 1-5A). Briefly, 74 open reading frames were identified and named by their proposed role in the biosynthesis (Figure 1-5A). A lot of outstanding structural and engineering studies of calicheamicin biosynthesis enzymes were conducted: among them the majority parts were finished by Thorson-Phillips lab collaboration (Figure 1-5B). Briefly, the sugar assembly process of calicheamicin (CalG proteins) and the orsellinic acid (CalO) biosynthesis have been thoroughly studied. The self-resistance mechanism of M. echinospora has been revealed. Unfortunately, our knowledge of enediyne core biosynthesis is still limited. Also, it is notable that calicheamicin has four unique sugar molecules, which are known to play critical roles in the recognition of the calicheamicin-DNA binding. It is also a great general interest to understand the sugar biosynthesis (Figure 1-5C) to produce a novel sugar library and contribute to natural product diversity. So far, three enzyme structures of sugar biosynthesis have been solved by the Phillips lab, including an sugar aminotransferase CalS13, which will be discussed in chapter 3.
Figure 1-5 Biosynthesis studies of the calicheamicin gene cluster*

(a) The gene cluster of calicheamicin biosynthesis was identified by Thorson lab in 2002. (b) Timeline of enzyme structures solved in the calicheamicin gene cluster. (c) The proposed pathway for the monosaccharide's biosynthesis

*Figure 1-5A is adapted from citation [71] with permission from The American Association for the Advancement of Science. The information in Figure 1-5C was kindly provided by Thorson lab and was re-made.
1.3. Glycorandomization

In the previous chapters, I suggest that natural products and their derivatives are arguably the best drug leads and we are entering a new golden age of natural products drug discovery. In this chapter, I will focus on a major component of natural products, that is, glycosylated natural products. According to the AntiBase 2012, a key database for glycoside identification, over one fifth of 15,940 bacteria natural products are glycosides. Among those glycosides, the major components are macrolides and macrolactams. Several examples of well-known glycosylated natural products are also shown in Figure 1-6A (right), like vancomycin, rebeccamycin, avermectin and calicheamicin (previously discussed in chapter 1.2). It is well established that glycosylation of natural products can dramatically affect their drug targeting, pharmacology, and biological activity. Thus altering glycosylation patterns of essential natural products can be a very promising approach to produce novel therapeutics.

There are a few approaches to alter the glycosylation patterns of the complicated secondary metabolites. The traditional method is the chemical synthesis or semi-synthesis, which total synthesize the target compounds via fermentation. This approach in theory is only limited by the available chemistries and synthetic expertise, but a significant challenge is the uncountable structural diversity and complexity of most glycosylated natural products. There are also other in vivo methods include pathway engineering and bioconversion. This approach
can access the novel compounds by fermentation, but the diversity of the final products is clearly limited by the substrate promiscuity of many enzymes. A final route I am mostly interested in is called "glycorandomization", which are in vitro biocatalysis reactions to install randomized sugar precursors on the randomized positions of various natural product aglycons with the help of engineered enzymes (Figure 1-6B)\(^5\). This approach is first studied and developed by Thorson lab. The development of this methodology includes two aspects: (1) building a library that contains many unique sugar precursors (Figure 1-6C) and (2) engineering enzymes (enzyme toolbox development) to handle/install those unique sugars on the natural product aglycons (Figure 1-6C). In chapter 3, I will present several structural studies of sugar aminotransferases and glycosyltransferases. From the former study, I want to gain knowledge of enzyme based sugar modification, and therefore expand the diversity of the sugar library. From the latter study, I want to understand the sugar installation process for each enzyme, and therefore further engineer them based on the structural information to expand their substrate promiscuity.
Figure 1-6. Natural product “glycorandomization” *

(a) Statistics of bacterial natural products (glycosylated vs. unglycosylated) and a few examples of glycosylated natural products88,97. (b) Principles of natural product glycorandomization. (c) A glance of unique sugar precursors and enzyme toolbox development of natural product glycorandomization88. (d) A gene toolbox of glycosyltransferase and nucleotide-sugar biosynthesis96.

---

*The data and figure in Figure 1-6A is adapted from citation [88] with permission from Royal Society of Chemistry, and [97] with the permission from Elsevier. Figure 1-6C is adapted from citation [88] with permission from Royal Society of Chemistry, and Figure 1-6D is adapted from citation [96] with the permission from Elsevier.*
A sugar library used for glycorandomization was developed by Thorson lab and a comprehensive review has been published previously\(^{88}\). Now the bottleneck of glycorandomization really is the enzyme toolbox development to install those unique sugar precursors onto the essential natural products. In the meantime, there are a few types of glycosyltransferase catalysis that I can utilize (Figure 1-7A)\(^{73}\): (1) the classical glycosyltransferase catalyzed sugar transfer from nucleotide sugars to a natural product aglycon; (2) the glycosyltransferase catalyzed sugar exchange reaction to exchange native sugar appendages with alternative sugars provided by unnatural nucleotide sugars; (3) the aglycon exchange reaction wherein a sugar is excised from the native acceptor and subsequently installed on a unnatural acceptor (Figure 1-7A). In the aglycon exchange process, if the native and the unnatural acceptor are in the different compound classes, multiple glycosyltransferases are usually required.

To date, there are already several successful examples of glycorandomization\(^{73,96,109-112}\) (Figure 1-7B-D). As discussed in chapter 1.2, calicheamicin antibody conjugates are very potent anti-tumor drugs and have been developed and tested in trials. Vancomycin is an essential antibiotic used to treat a number of bacterial infections, like skin, bloodstream, and bone and joint infections. In 2006, Thorson lab published a glycorandomization paper, in which they are able to generate more than 70 differentially glycosylated calicheamicin and vancomycin variants\(^{73}\). The construction of the calicheamicin library (Figure 1-7B) is accomplished by a glycosyltransferase
CalG1-based sugar exchange reactions. CalG1 is one of the four glycosyltransferases in the calicheamicin biosynthesis pathway, and its crystal structure was solved by the Phillips lab\textsuperscript{72,74}. The construction of the vancomycin library is accomplished by aglycon exchange reaction catalyzed by glycosyltransferases (Figure 1-7C). A few aglycon exchange reactions within the vancomycin class can be catalyzed by GtfD. Several other aglycon exchange reactions between vancomycin and calicheamicin can be catalyzed by GtfE\textsuperscript{113,114} and CalG1\textsuperscript{73}.

Another preliminary glycorandomization study is published by the Thorson lab about a macrolide glycosyltransferase OleD\textsuperscript{112,115-125} (Figure 1-7D). A structural and protein dynamics study of OleD will be presented in chapter 3.4. But briefly, herein in this paper, Thorson lab firstly found a fluorescent surrogate acceptor for OleD and developed a high-throughput screen based on this. Subsequently, they engineered OleD based on its structure (will be shown in chapter 3.4) and expanded the substrate promiscuity of OleD to be capable of transferring/installing a variety of nucleotide sugars. This brilliant work not only provides a general high-throughput platform for glycosyltransferase engineering, but also make OleD a very promising glycorandomization target to produce Oleandomycin derivatives.
Figure 1-7. Examples of glycorandomization*

(a) scheme of glycosyltransferase catalysis. (b) calicheamicin glycorandomization using CalG1. (c) Vancomycin glycorandomization using GtfD, GtfE and CalG1. (d) Oleandomycin glycorandomization using OleD.

*The data and figure in Figure 1-7A-C are adapted from citation [73] with the permission from the American Association for the Advancement of Science. Figure 1-7D is adapted from citation [112] with the permission from Nature Publishing Group.
1.4. Alkylrandomization

In the previous chapters, I suggest that natural products and their derivatives are arguably the best drug leads⁴⁹. A methyl/alkyl group can be found in most of those essential natural products, and some of those methyl/alkyl groups are installed by SAM (S-adenosyl methionine) stylizing enzymes. A few examples of methylated natural products are listed in Figure 1-8A, like esperamicin, evernimicin, and nogalamycin. Although methyl/alkyl groups are everywhere, scientists cannot ignore their contribution to the compound chemical properties. For example, the incorporation of alkyl chains into some natural products increases their lipophilicity, and this strategy can be used to enhance the antimicrobial activity of flavanone and chalcone¹²⁶,¹²⁷. As exemplified by natural product glycorandomization, altering alkyl-patterns on clinical essential natural products is very promising for the generation of novel therapeutics, and this approach is named as “alkylrandomization”. A platform that contains natural product methyltransferase for a broader natural product alkylation would definitely be expected to dramatically expand the potential scope of natural product chemical diversity, and using this platform will provide an optimization choice for a single function group on a specific natural product.
Figure 1-8. Natural product “alkylrandomization”

(a) A few examples of methylated natural products. (b) Principles of natural product alkylrandomization.

---

1Figure 1-8A is adapted and re-made from citation [128] with the permission from John Wiley and Sons.
The typical methyl/alkyl donor for methyltransferase\textsuperscript{129–135} (MT) catalyzed methylation/alkylation is S-adenosyl methionine (SAM, or AdoMet). Therefore, if I want to engineer methyltransferase to make it install something bigger than the methyl group, I will need the corresponding AdoMet (SAM) analogs, namely an Ado-X library (Figure 1-8B). The good news is that Thorson lab and others have demonstrated that methyltransferases of natural products, proteins and nucleic acid can transfer alternative alkyl groups in the presence of proper and suitable AdoMet analogs\textsuperscript{136–141}. The bad news is that AdoMet analogs are very hard to chemically synthesize, and are also very unstable in regular liquid solutions. So the bottleneck now for alkylrandomization is really both the generation of an AdoMet library and engineering methyltransferase. Methionine adenosyltransferase\textsuperscript{142–156} (MAT) is the family of protein to make AdoMet using ATP and methionine. So in theory if I can find a MAT and engineer it to product AdoMet analogs, I can do MAT/MT coupled engineering: supply engineered MAT/MT with ATP and proper methionine analogs (this can be chemically synthesized easily) and collect the alkylated natural products (Figure 1-8B). In chapter 4.1, a structural and engineering study of sMAT will be presented to show that an Ado-X library can be generated using sMAT\textsuperscript{157}.

A preliminary alkylrandomization example is shown in Figure 1-9. Briefly in 2006, the Thorson lab found that RebM\textsuperscript{158,159}, the \textit{O}-methyltransferase in the rebeccamycin biosynthesis, has great substrate promiscuity. And therefore using a combined RebG/RebM enzyme reaction, methyl group can be installed on different positions of rebeccamycin or similar indolocarbazoles\textsuperscript{160} (Figure 1-9A). The crystal structure, of
course, was solved by the Phillips lab to better understand the substrate promiscuity of RebM\textsuperscript{161}. The crystal structure of RebM in complex with AdoHcy enabled the interpretation of several key catalytic residues, and provides crucial information for the mechanism of this enzyme family. Also, this result is consistent with the reported ability of RebM to be able to methylate a wide range of indolocarbazole analogs, therefore this structure can be used as a blue print for future engineering\textsuperscript{161}. Furthermore, Thorson lab was able to generate a small set of differentially alkylated indolocarbazole surrogates by using a coupled MAT/MT system (hMAT plus RebM, Figure 1-9B). This is encouraging and is great preliminary data suggesting the feasibility of alkylrandomization approach to expand the natural product diversity. In chapter 4, the structural and engineering studies of sMAT and another promising methyltransferase, in terms of alkylrandomization application, will be presented and discussed.
Figure 1.9 Examples of alkylrandomization

(a) Methyltransferase RebM shows its exciting ability to install methyl group on variety positions of rebeccamycin\textsuperscript{160}. (b) Coupled hMAT/RebM reactions generate a small set of alkylated indolocarbazoles analogs\textsuperscript{128,161}.

\textsuperscript{1}Figure 1.9A is adapted from citation [160] with the permission from John Wiley and Sons. The right part of Figure 1.9B is adapted from citation [128] with the permission from John Wiley and Sons.
References


69. Castaigne, S. *et al.* Effect of gemtuzumab ozogamicin on survival of adult patients with de-novo acute myeloid leukaemia (ALFA-0701): A randomised, open-label, phase 3


Materials, Methods and Depositions

This chapter will list all the research materials (e.g. UniProt IDs of genes, synthesized compounds for biochemistry assays, commercial crystallization screens, research equipment, and biochemistry assay kits) related or used during my thesis, protein data bank depositions for solved structures or ensemble refinements, and describes methods/protocols including cloning, mutagenesis, protein expression, protein purification, protein crystallization, structural solutions, biochemical assays, and structure visualization.
2.1. Materials

2.1.1. UniProt IDs of target genes/proteins

P27833  WecE from *Escherichia coli* K-12
Q0H2X1  AtmS13 from *Actinomadura melliaura*
Q8KND8  CalS13 from *Micromonospora echinospora*
D6MSX4  SsfS6 from *Streptomyces sp. SF2575*
Q53685  OleD from *Streptomyces antibioticus*
Q98059  sMAT from *Sulfolobus solfataricus*
Q06528  DnrK from *Streptomyces peucetius*
Q8KZ94  RebM from *Lechevalieria aerocolonigenes*
O75164  KDM4A from *Homo sapiens*
O94953  KDM4B from *Homo sapiens*
Q7MZL9  Cupin plu4264 from *Photorhabdus luminescens*
Q8KNG3  CalE6 from *Micromonospora echinospora*
Q8KNE9  CalU16 from *Micromonospora echinospora*
Q8KNC7  CalU19 from *Micromonospora echinospora*

2.1.2. Methionine analogs used in sMAT reactions

This list of compounds was summarized in Table 2-1, and they have been used in the sMAT paper\(^1\). ND, not detected; NA, not applicable
Table 2-1. Methionine analogs used in Chapter 4

<table>
<thead>
<tr>
<th>Analog</th>
<th>Analog name</th>
<th>Percentage turnover</th>
<th>Theoretical mass</th>
<th>Observed mass</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Methionine</td>
<td>100</td>
<td>399.1445, 297.0896</td>
<td>399.1444, 298.0965</td>
</tr>
<tr>
<td></td>
<td>Se-Methionine</td>
<td>100</td>
<td>447.0890, 345.0340</td>
<td>447.0895, 346.0416</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(ethylsulfanyl)butanoic acid</td>
<td>84</td>
<td>413.1602, 311.1052</td>
<td>413.1599, 312.1121</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(ethylselanyl)butanoic acid</td>
<td>76</td>
<td>461.1046, 359.0497</td>
<td>416.1040, 360.0565</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(prop-2-yn-1-ylsulfanyl)butanoic acid</td>
<td>36</td>
<td>321.0896</td>
<td>322.1051</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(prop-2-yn-1-ylselanyl)butanoic acid</td>
<td>28</td>
<td>369.0340</td>
<td>370.0411</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(propa-1,2-dien-1-ylsulfanyl)butanoic acid</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(cyanomethyl) sulfanyl)butanoic acid</td>
<td>68</td>
<td>322.0848</td>
<td>323.0919</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(cyanomethyl) selanyl)butanoic acid</td>
<td>12</td>
<td>370.0293</td>
<td>371.0371</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(prop-2-en-1-ylsulfanyl)butanoic acid</td>
<td>27</td>
<td>425.1602, 323.1052</td>
<td>425.1609, 324.1129</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(propan-2-ylsulfanyl)butanoic acid</td>
<td>30</td>
<td>371.0497</td>
<td>372.0568</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(propan-2-ylselanyl)butanoic acid</td>
<td>44</td>
<td>427.1758, 325.1209</td>
<td>427.1759, 326.1281</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(propylsulfanyl)butanoic acid</td>
<td>66</td>
<td>475.1203, 373.0653</td>
<td>475.1200, 374.0726</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(propylselanyl) butanoic acid</td>
<td>53</td>
<td>427.1758, 325.1209</td>
<td>427.1762, 326.1282</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(propan-2-ylsulfanyl)butanoic acid</td>
<td>40</td>
<td>373.0653</td>
<td>374.0743</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(propan-2-ylselanyl)butanoic acid</td>
<td>30</td>
<td>437.1602, 335.1052</td>
<td>437.1599, 336.1124</td>
</tr>
<tr>
<td></td>
<td>(2S)-2-amino-4-(but-3-yn-1-ylsulfanyl)butanoic acid</td>
<td>10</td>
<td>383.0497</td>
<td>384.0566</td>
</tr>
<tr>
<td>Analog</td>
<td>Analog name</td>
<td>Percentage turnover</td>
<td>Theoretical mass</td>
<td>Observed mass</td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------------------</td>
<td>---------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((2\text{E}))-but-2-en-1-ylsulfanyl]-butanoic</td>
<td>27</td>
<td>439.1758, 337.1209</td>
<td>439.1754, 338.1275</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((2\text{methylprop-2-en-1-yl})]sulfanyl]-butanoic</td>
<td>15</td>
<td>439.1758, 337.1209</td>
<td>439.1754, 338.1282</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((2\text{methylprop-2-en-1-yl})]selanyl]-butanoic</td>
<td>46</td>
<td>385.0653</td>
<td>386.0730</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((butylsulfanyl)]-butanoic acid</td>
<td>27</td>
<td>441.1915, 339.1365</td>
<td>441.1966, 340.1477</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((butylselanyl)]-butanoic acid</td>
<td>70</td>
<td>489.1359, 387.0810</td>
<td>489.1358, 388.0882</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((2\text{methylprop-2-en-1-yl})]sulfanyl]-butanoic</td>
<td>57</td>
<td>441.1915, 339.1365</td>
<td>441.1918, 340.1439</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((3\text{amino propyl})]sulfanyl]-butanoic acid</td>
<td>ND</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((3\text{amino propyl})]selanyl]-butanoic acid</td>
<td>ND</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((2\text{E}))-penta-2,4-dien-1-ylsulfanyl]-butanoic</td>
<td>14</td>
<td>451.1758</td>
<td>451.1756</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((2\text{azido ethyl})]sulfanyl]-butanoic acid</td>
<td>50</td>
<td>454.1616, 352.1066</td>
<td>454.1615, 353.1140</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((3\text{methyl butyl})]sulfanyl]-butanoic acid</td>
<td>36</td>
<td>455.2071, 353.1522</td>
<td>455.2123, 354.1634</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((3\text{methylbut-2-en-1-yl})]sulfanyl]-butanoic acid</td>
<td>12</td>
<td>453.1915</td>
<td>453.2044</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((3\text{methylbut-2-en-1-yl})]selanyl]-butanoic acid</td>
<td>44</td>
<td>399.0810</td>
<td>400.0878</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((3\text{azido propyl})]sulfanyl]-butanoic acid</td>
<td>60</td>
<td>468.1772, 366.1223</td>
<td>468.1772, 367.1296</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((benzylsulfanyl)]-butanoic acid</td>
<td>ND</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((benzylselanyl)]-butanoic acid</td>
<td>ND</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((3\text{azido propyl})]sulfanyl]-butanoic acid</td>
<td>60</td>
<td>468.1772, 366.1223</td>
<td>468.1772, 367.1296</td>
</tr>
<tr>
<td>(\text{S})-</td>
<td>((2\text{S}))-2-amino-4-((benzylsulfanyl)]-butanoic acid</td>
<td>ND</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
Thus based on the simulated annealing Fo/Coomit map of the active site, one ADP, one PO₄, two Mg²⁺ and one ethionine (or methionine) molecule were initially built in (Fig. 5, left). However, this model does not fit the electron density perfectly, because the Fo/Coomit map does not agree with the placement of the crucial carbon atom circled in Fig. 5. Thus, it is very clear that the product has already formed and a model including PPi and AdoEth (or AdoMet) is more appropriate. The new model (Fig. 5, right) has a lower temperature factor and a better real-space correlation to electron density in the active site.

MAT-catalyzed AdoMet/AdoEth formation via ADP and Met/Eth has not been previously observed. In addition, incubation

<table>
<thead>
<tr>
<th>Analog</th>
<th>Analog name</th>
<th>Percentage turnover</th>
<th>Theoretical mass</th>
<th>Observed mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Se-NH₂</td>
<td>(2S)-2-amino-4-[(2E)-4-aminobut-2-en-1-yl]selanyl]butanoic acid</td>
<td>38</td>
<td>400.0762</td>
<td>401.0830</td>
</tr>
<tr>
<td>S</td>
<td>(2S)-2-amino-4-[(4-methylphenyl)methyl]sulfanyl]butanoic acid</td>
<td>ND</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>S</td>
<td>(2S)-2-amino-4-[(2E)-3-phenylprop-2-en-1-yl]sulfanyl]butanoic acid</td>
<td>ND</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>S</td>
<td>(2S)-2-amino-4-[(4-nitrophenyl)methyl]sulfanyl]butanoic acid</td>
<td>ND</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>S</td>
<td>(2S)-2-amino-4-[(4-nitrophenyl)methyl]selanyl]butanoic acid</td>
<td>ND</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>S</td>
<td>(2S)-2-amino-4-[(2S)-3-amino-3-carboxypropyl][sulfanyl]but-2-en-1-yl]sulfanyl]butanoic acid</td>
<td>ND</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
# 2.1.3. Cloning, expression and purification materials

<table>
<thead>
<tr>
<th>Category</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloning Vectors</td>
<td>pET28a, pET28b and pET30a from Novagen</td>
</tr>
<tr>
<td>Restriction Enzymes</td>
<td><em>NdeI, EcoRI, XhoI</em> etc. from NEB</td>
</tr>
<tr>
<td>Plasmid prep kits</td>
<td>Mini, Maxi plasmid prep kits from QIAGEN</td>
</tr>
<tr>
<td>In-Fusion Cloning</td>
<td>#639645 from Clontech Lab</td>
</tr>
<tr>
<td>DNA Polymerase</td>
<td>Q5 Hot Start High-Fidelity 2X Master Mix(^2) from NEB</td>
</tr>
<tr>
<td>PCR purification</td>
<td>DNA Clean/Concentrator columns, Genesee Scientific Corp</td>
</tr>
<tr>
<td>PCR equipment</td>
<td>Bio-Rad CFX qPCR Systems</td>
</tr>
<tr>
<td>Mutagenesis</td>
<td>Q5 Site-Directed Mutagenesis Kit</td>
</tr>
<tr>
<td>Ligation</td>
<td>Gibson Assembly Master Mix(^3) from NEB</td>
</tr>
<tr>
<td>Transformation strain</td>
<td>Turbo Competent <em>E.coli</em> from NEB</td>
</tr>
<tr>
<td>Transformation strain</td>
<td>5-alpha Competent <em>E.coli</em> from NEB</td>
</tr>
<tr>
<td>Transformation strain</td>
<td>Stellar Competent <em>E.coli</em> from Clontech Lab</td>
</tr>
<tr>
<td>Expression Medium</td>
<td>Auto-inducing MagicMedia K6803 from Life technologies echo</td>
</tr>
<tr>
<td>Expression strain</td>
<td>BL21 (DE3) competent <em>E.coli</em> from NEB</td>
</tr>
<tr>
<td>Expression strain</td>
<td>Lemo21 (DE3) Competent <em>E.coli</em> from NEB</td>
</tr>
<tr>
<td>Expression strain</td>
<td>Rosetta2 (DE3) Competent <em>E.coli</em> from EMD Millipore</td>
</tr>
<tr>
<td>Expression system</td>
<td>Wheat germ Cell free kit from CFS</td>
</tr>
<tr>
<td>Purification system</td>
<td>FPLC ÄKTAprime plus from GE Healthcare</td>
</tr>
<tr>
<td>Purification columns</td>
<td>Affinity, ion exchange and gel filtration from GE Healthcare</td>
</tr>
</tbody>
</table>
### 2.1.4. Crystallization kits and equipment

<table>
<thead>
<tr>
<th>Kit Description</th>
<th>Manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index HT</td>
<td>Hampton Research</td>
</tr>
<tr>
<td>SaltRx HT</td>
<td>Hampton Research</td>
</tr>
<tr>
<td>PegRx HT</td>
<td>Hampton Research</td>
</tr>
<tr>
<td>Peg/ion HT</td>
<td>Hampton Research</td>
</tr>
<tr>
<td>Natrix HT</td>
<td>Hampton Research</td>
</tr>
<tr>
<td>Crystal Screen HT</td>
<td>Hampton Research</td>
</tr>
<tr>
<td>JCSG-plus™</td>
<td>Molecular Dimensions</td>
</tr>
<tr>
<td>Morpheus</td>
<td>Molecular Dimensions</td>
</tr>
<tr>
<td>MIDAS™</td>
<td>Molecular Dimensions</td>
</tr>
<tr>
<td>ProPlex™</td>
<td>Molecular Dimensions</td>
</tr>
<tr>
<td>MCSG-1</td>
<td>Microlytic</td>
</tr>
<tr>
<td>MCSG-2</td>
<td>Microlytic</td>
</tr>
<tr>
<td>MCSG-3</td>
<td>Microlytic</td>
</tr>
<tr>
<td>MCSG-4</td>
<td>Microlytic</td>
</tr>
<tr>
<td>Wizard classic 1&amp;2</td>
<td>Rigaku</td>
</tr>
<tr>
<td>Wizard classic 3&amp;4</td>
<td>Rigaku</td>
</tr>
<tr>
<td>Crystallization Robot</td>
<td>Mosquito LCP from TTP labtech</td>
</tr>
<tr>
<td>Liquid handling Robot</td>
<td>Scorpion screen builder from ARI</td>
</tr>
<tr>
<td>UV Microscope</td>
<td>Protein crystal inspector from JANSi</td>
</tr>
</tbody>
</table>
2.1.5. Crystal handling and shipping

Crystal loops          Mounted CryoLoop from Hampton Research
Crystal loops          MicroLoops from MiTeGen
Cryo-protectants      CryoPro from Hampton Research
Liquid nitrogen       Chemistry stockroom at Rice University
Crystal testing       X-ray home source at Rice University
LS-CAT kit            EMBL/ESRF sample kit from Molecular Dimensions
GM/CA kit             Universal Pucks and Tools, Crystal Positioning Systems
Dry Shipper           VWR CryoPro® Vapor Shippers, V Series
Dry Shipper           CX100 Dry Shipper (GM/CA) from Taylor Wharton

2.1.6. Biochemical assays*

Phosphate release assay | EnzChek Phosphate Assay Kit\(^4\) (E-6646)
Thermal shift assay     | Sypro orange fluorescence\(^5\)
Bradford                | Bradford Protein Assay from NEB
Protein alkylation      | Reductive Alkylation\(^6\) Kit from Hampton Research
Mutagenesis             | Q5 Site-Directed Mutagenesis Kit

---
\(^*\)This section only records the biochemical assays done in the Phillips lab.
2.1.7. Software packages

Lasergene\textsuperscript{7} Developed by DNASTAR

HKL-3000\textsuperscript{8} Developed by HKL Research Inc.

XDS\textsuperscript{9} Developed by Wolfgang Kabsch Group

Phenix\textsuperscript{10} Developed by Paul Langan and Paul Adams groups

Coot\textsuperscript{11} Developed by Paul Emsley and Kevin D. Cowtan

PyMOL\textsuperscript{12} Developed by Schrödinger, Inc.

QuteMol\textsuperscript{13} Developed by ISTI - CNR

CCP4\textsuperscript{14} Developed by CCLRC Daresbury Laboratory

ChemDraw\textsuperscript{15} Developed by CambridgeSoft

Endnote\textsuperscript{16} Developed by Thomson Reuters

Mendeley\textsuperscript{17} Developed by Mendeley Ltd.

SB-Grid\textsuperscript{18} Established by Dr. Piotr Sliz

Adxv Established by Andrew Arvai

Autodock\textsuperscript{19} Developed by labs in the Scripps Research Institute

DALI server Developed by Liisa Holm's Bioinformatics Group

Molprobity\textsuperscript{20} Developed by Kinemage Group at Duke

Illustrator Developed by Adobe

Clustal W\textsuperscript{21} Developed by EMBL-EBI

4Peaks Developed by Alexander Griekspoor and Tom Groothuis

LigPlot\textsuperscript{22} Developed by European Bioinformatics Institute

ExPASy Tools\textsuperscript{23} SIB Bioinformatics Resource Portal
2.2. Methods

2.2.1. Cloning methods

Most of my primer design was accomplished with the help of Integrated DNA Technologies (IDT) online tools, and some examples are shown in Figure 2-1. Briefly, in order to PCR amplify a target DNA fragment from a vector, and then subsequently clone this fragment into a new vector (let's say between NdeI and XhoI digestion sites in pET28a), following steps are my usual procedures. (1) primer design; (2) vector digestion and purify the products; (3) PCR amplification of target gene; (4) DNA gels on step 3 products and PCR products recycles; (5) Infusion reactions of products between step 2 and step 4; (5’) Or using Gibson Assembly protocol to replace step 5; (6) Transformation to E.coli cells; (7) plasmid mini-prep for sequencing validation.

(1) Primer design, like shown in Figure 2-1, includes forward primer and reverse primer. Forward primer includes an overlap region to pET28a vector (purple and red in Figure 2-1) and then an overlap region to PCR amplified target genes (black in Figure 2-1). \( T_m \) calculation and GC\% content of the primer is only for the black region since ideally only this part will pair with PCR the templates. \( T_m \) and GC\% are purposely designed to be close to 58-62°C and 40%-60%, respectively. Similar to forward primers design, reserve primers need an extra stopping codon and complement of the sequence (Figure 2-1, right part, the bottom one of the two lines).
Figure 2-1 Examples of primer design

For PET28a Ndel Xhol

**G469 Forward**

<table>
<thead>
<tr>
<th>GC CONTENT</th>
<th>MELT TEMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.1%</td>
<td>58.9°C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>peG82a</th>
<th>G469</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCG GGC GGC AAG</td>
<td>CAT</td>
</tr>
<tr>
<td>GAT GCT TAC CCG AGA CTT AAG</td>
<td></td>
</tr>
</tbody>
</table>

**G477 Forward**

<table>
<thead>
<tr>
<th>GC CONTENT</th>
<th>MELT TEMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>55.5%</td>
<td>58.1°C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>peG82a</th>
<th>G477</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCG GGC GGC AAG</td>
<td>CAT</td>
</tr>
<tr>
<td>GAT GAT TAC CAG CTT GCT AGT GAT</td>
<td></td>
</tr>
</tbody>
</table>

**T486 Forward**

<table>
<thead>
<tr>
<th>GC CONTENT</th>
<th>MELT TEMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>52.0%</td>
<td>57.7°C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>peG82a</th>
<th>T486</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCG GGC GGC AAG</td>
<td>CAT</td>
</tr>
<tr>
<td>ACA TCA GGG TGG CTT CGA CTA T</td>
<td></td>
</tr>
</tbody>
</table>

**S500 Forward**

<table>
<thead>
<tr>
<th>GC CONTENT</th>
<th>MELT TEMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>52.2%</td>
<td>57.9°C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>peG82a</th>
<th>S500</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCG GGC GGC AAG</td>
<td>CAT</td>
</tr>
<tr>
<td>TCT GCT GGC TCQ GGG AAG A</td>
<td></td>
</tr>
</tbody>
</table>

**A513 Forward**

<table>
<thead>
<tr>
<th>GC CONTENT</th>
<th>MELT TEMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.1%</td>
<td>58.8°C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>peG82a</th>
<th>A513</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCG GGC GGC AAG</td>
<td>CAT</td>
</tr>
<tr>
<td>GGC TGG TCC CTT CGG TGA AAT</td>
<td></td>
</tr>
</tbody>
</table>

---

For PET28a Ndel Xhol

**A810 Reverse**

<table>
<thead>
<tr>
<th>GC CONTENT</th>
<th>MELT TEMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>59.9%</td>
<td>58.8°C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>peG82a</th>
<th>A810</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGA GAT GTT CTT GGG TGG TGG ATT GA4 TTT GGC</td>
<td>TAA CTC GAG GAA CAC CAC GAC</td>
</tr>
<tr>
<td>GTG GTG GTG GTG GTG GTG CTT GTA TTA CCG AAA TTC TAA GCT CCA CAG AAC ATC TCC</td>
<td></td>
</tr>
</tbody>
</table>

**A817 Reverse**

<table>
<thead>
<tr>
<th>GC CONTENT</th>
<th>MELT TEMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>51.1%</td>
<td>58.9°C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>peG82a</th>
<th>A817</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAG CTC GAG GAA AAG GCA GC</td>
<td>TAA CTC GAG GAA CAC CAC GAC</td>
</tr>
<tr>
<td>GTG GTG GTG GTG GTG GTG CTT GTA TTA TCC GCT TCC CAG CAG CAG</td>
<td></td>
</tr>
</tbody>
</table>

**E832 Reverse**

<table>
<thead>
<tr>
<th>GC CONTENT</th>
<th>MELT TEMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>54.5%</td>
<td>57.7°C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>peG82a</th>
<th>E832</th>
</tr>
</thead>
<tbody>
<tr>
<td>C GTT GAC AGC ATG ATG GAG GAG</td>
<td>TAA CTC GAG GAA CAC CAC GAC</td>
</tr>
<tr>
<td>GTG GTG GTG GTG GTG GTG CTT GTA TTA CTC ATC CAT GTG CAT GTG TCC AGC G</td>
<td></td>
</tr>
</tbody>
</table>

**D849 Reverse**

<table>
<thead>
<tr>
<th>GC CONTENT</th>
<th>MELT TEMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>60.7%</td>
<td>58.8°C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>peG82a</th>
<th>D849</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGT AAA GGA AGG AAC GAC AAG AGT TCT GAT</td>
<td>TAA CTC GAG GAA CAC CAC GAC</td>
</tr>
<tr>
<td>GTG GTG GTG GTG GTG GTG CTT GTA TTA ATG AAG ACT CTT GTC GTT CTT GTA TTA GTA</td>
<td></td>
</tr>
</tbody>
</table>

**G862 Reverse**

<table>
<thead>
<tr>
<th>GC CONTENT</th>
<th>MELT TEMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.1%</td>
<td>58.2°C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>peG82a</th>
<th>G862</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAC TCG AGG AGC AGT GGA</td>
<td>TAA CTC GAG GAA CAC CAC GAC</td>
</tr>
<tr>
<td>GTG GTG GTG GTG GTG GTG CTT GTA TTA TCC ACT GCT CCT CGA GTC</td>
<td></td>
</tr>
</tbody>
</table>

---
(2) Vector digestion and purify the products

<table>
<thead>
<tr>
<th>Volume (μL)</th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>pET28a (230 μg/μL)</td>
</tr>
<tr>
<td>4</td>
<td>Ndel (4,000 units/ml)</td>
</tr>
<tr>
<td>4</td>
<td>Xhol (4,000 units/ml)</td>
</tr>
<tr>
<td>5</td>
<td>Buffer from NEB, check which one works best for your enzymes</td>
</tr>
<tr>
<td>50</td>
<td>Total</td>
</tr>
</tbody>
</table>

Mix the pET28a vector with digestion enzymes as above, incubate at 37°C for 4 hours, and then run an agarose gel and do gel extraction.

(3) Q5 PCR amplification of target gene

I used PfuUltra DNA polymerase at the beginning, but it doesn’t work as well as Q5 from NEB. One day I used very similar conditions (new Pfu, new buffer, new dNTP) for both polymerases to amplify several genes in the same PCR plate at the same time. Q5 ends up successfully amplifying all of the genes while interestingly PfuUltra produced nothing (Figure 2-2A). Below is the reaction composition for Q5 PCR amplification.

<table>
<thead>
<tr>
<th>Volume (μL)</th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.25</td>
<td>Primer Forward (10 μM)</td>
</tr>
<tr>
<td>1.25</td>
<td>Primer Reverse (10 μM)</td>
</tr>
<tr>
<td>4</td>
<td>Template DNA (5 ng/μL)</td>
</tr>
<tr>
<td>12.5</td>
<td>Q5 2 x Master Mix</td>
</tr>
<tr>
<td>6</td>
<td>H₂O</td>
</tr>
<tr>
<td>25</td>
<td>Total</td>
</tr>
</tbody>
</table>
Typical PCR cycles are shown in below:

<table>
<thead>
<tr>
<th>Cycle</th>
<th>Temperature</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>98°C</td>
<td>30 s</td>
</tr>
<tr>
<td>2</td>
<td>98°C</td>
<td>20 s</td>
</tr>
<tr>
<td>3</td>
<td>53°C</td>
<td>20 s</td>
</tr>
<tr>
<td>4</td>
<td>72°C</td>
<td>45 s</td>
</tr>
<tr>
<td>5</td>
<td>72°C</td>
<td>2 min</td>
</tr>
<tr>
<td>6</td>
<td>4°C</td>
<td>∞</td>
</tr>
</tbody>
</table>

(4) DNA gels on step 3 products and PCR products recycles

After PCR amplification, small amount of the products were run on a DNA gel. If bands on the correct molecular weight were observed, PCR products were recycled and cleaned using a kit made by Zymo.

(5) Infusion reactions to PCR ligate amplified genes and pET28a vectors.

The protocol of this method can be found at in this reference\textsuperscript{24}. It worked at the very beginning in 2014, and the ligation enzyme expires in a few months. After 2015, however, it only worked once in about ∼300 of the PCR ligation attempts. Most bacteria plates end up with having no colonies, including their own positive controls. Interestingly, using exactly the same mixture, enzyme(s) in Gibson Assembly Kit can finish the ligation\textsuperscript{25} (Figure 2-2B) and see colonies on plate.
Figure 2-2. Q5 Polymerase and Gibson Assembly over the others

(a) DNA gels of PCR amplification results. The reaction conditions are almost identical between Q5 and PfuUltra Polymerases, except for Q5 comes with a master mix while PfuUltra to add dNTP etc. (b) Infusion ligation and Gibson assembly reactions were conducted with the same gene fragments, digested vector, and transformation cells. Gibson ended up giving a lot of colonies on the plate and they were sequencing confirmed to be correct.
(5’) Gibson Assembly protocol

As shown in Figure 2-2B, Gibson Assembly protocol is the best fit on my purpose. And the mixture is described below:

<table>
<thead>
<tr>
<th>Volume (µL)</th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2x Gibson Assemble Master Mix</td>
</tr>
<tr>
<td>0.7</td>
<td>pET28a (double enzyme digested)</td>
</tr>
<tr>
<td>1</td>
<td>PCR fragment (50-100 ng/µL)</td>
</tr>
<tr>
<td>3.3</td>
<td>H₂O</td>
</tr>
<tr>
<td>10</td>
<td>Total</td>
</tr>
</tbody>
</table>

The mixtures were incubated at 50°C for 20 min, and then transformed in Stellar cells or other competent cells like Dh5a and NEB tuber.

(6) Transformation to E.coli cells

Follow the transformation protocol come with the competent cells. Different cells might vary in their protocols. After transformation, pick single colonies on the plate, grow it in LB medium overnight with the proper antibiotic, and then extract plasmid from the cells for DNA sequencing.

(7) DNA sequencing

Send samples to Lonestar or SeqWright (GE Health) care labs for sequencing, a comparison of their sequencing qualities was shown in Figure 2-3.
Figure 2-3 DNA sequencing peaks from Lone Star and SeqWright

The sequencing peaks from a random result of SeqWright (bottom), and one of the best results I ever received from Lone Star (top). With those results, I decided to send all my samples to SeqWright, as the resolution of the spectra is better.
2.2.2. Expression methods

(1) Wheat germ cell free methods

A small scale *in vitro* cell free expression can be done using a wheat germ kit made by Cell Free Sciences (CFS) and detailed protocols can be found online. Briefly, the steps are listed as: (1) mix 2 μL DNA plasmid (1000 ng/μL) with 18 μL transcription premix and incubate at 37°C for 1 hour; (2) run a quick gel of the mRNA product and confirm that RNase does not degrade the mRNA (Figure 2-4A); (3) put 10 μL product of step 1 into the bottom of 10 μL wheat germ extract and incubate at 15°C for 20 hours.

(2) *E.coli* high throughput small-scale expression.

To test expression conditions for multiple gene constructs at the same time, I used an auto-inducing media named “MagicMedia” from Life technologies. Briefly, I first transformed the plasmids containing the target genes into expression competent cells, like BL21 (DE3), and then picked a single colony and grew it in 1ml LB media overnight. Next day, added 5μL overnight culture into 1ml fresh MagicMedia, and induced the protein expression following the MagicMedia protocol (K6803) at different temperatures. Cells were subsequently incubated with Bugbuster lysate, and the expression levels in both whole cells and supernatant were tested using SDS-PAGE gel (Figure 2-4B).
(3) *E.coli* large-scale expression.

For selenomethionyl protein, the plasmid construct was transformed into *E. coli* methionine auxotroph strain B834 (DE3); while for native protein production, the plasmid construct was transformed into *E. coli* strain BL21 (DE3) or Rosetta (DE3). Se-Met labeled protein overproduction was accomplished using auto-induction medium at proper temperature (mostly 16 or 25 °C) following the standard protocol\textsuperscript{27}. Native unlabeled N-terminal histag protein was overproduced in Terrific Broth media or in auto-induction MagicMedia. The expression temperature and time varies between different proteins and media. For example, the expression temperature and time for terrific broth media is 37 °C for 6 hours, 25 °C for 14 hours and 16 °C for 30 hours. After expression, the cells were harvested by centrifugation at 4,200 *g* for 30 min, and then were stored at -80 °C. Similar to small-scale expression, expression level was tested by SDS-PAGE gel before purification (Figure 2-4C).
Figure 2-4. Expression examples of wheat germ cell-free, small-scale *E.coli* and large-scale *E.coli*

(a) DNA ladder | cleaned DNA
RNA ladder | transcription
mRNA

(b) M W S

(c) M W S

M = Marker
W = Whole cells
S = Supernatant
2.2.3. Purification methods

(1) Spin column testing purification

To quickly determine whether a N-terminal histag protein can be successfully purified under given buffer conditions, I used Ni-NTA spin columns (QIAGEN 31014) and followed the supplied protocol.

(2) Batch purification for biochemical assays

To purify a few protein mutations at the same time for biochemical assays, (e.g. sMAT mutants), I used the protocol below, which is modified from a protocol published by Su et al:

1. Harvest cells at 5000 g for 20 min in a 50 ml falcon tube. Cell pellets can be stored at -80°C for weeks. Run a SDS-PAGE gel to check whether target mutants have been successfully expressed.

2. Equilibrate nickel resin (1 ml Ni resin per sMAT mutant, resin from GE) 3 times with 20 ml Starting Buffer in a 50 ml falcon tube. Mix buffer into resin by hand before centrifuging at 2000 g for 5 minutes on tabletop centrifuge. Discard buffer between washes.

3. Re-suspend bacterial pellet in Starting Buffer (20 ml per sMAT mutant).

4. Sonicate lysate with amplitude of 12% (power ~ 25 watts, 5 seconds “on” / 10 seconds “off”, total “on” time of 2 minutes). Keep lysate on ice.
5. Clarify lysate for 45 minutes at 18,000 rpm in JA-25.50 rotor, 4°C.

6. Add supernatant from clarified lysate to equilibrated resin from Step #2. Mix for 2 hours at room temperature with rotation to bind protein to resin.

7. Centrifuge resin for 5 minutes at 2000 \( g \). Discard the supernatant.

8. Wash resin once with 20 ml Starting Buffer, mixing briefly by hand inversion. Centrifuge and discard flow-through.

9. Wash resin twice with 20 ml Washing Buffer, mixing for 5 minutes at RT with rotation. Centrifuge resin and discard supernatant.

10. Elute two times with 5 ml Elution Buffer, mixing for 10 minutes at RT (transfer into a 15ml tube if necessary). Centrifuge resin and combine the elution sample together.

11. Concentrate the elution samples with centrifugal filters (10,000 NMWL) at 3500 \( g \) until the volume is around 1 ml. Discard the flow through, and add 10 ml storage buffer (3 times).

12. The final volume of protein samples should be around 0.5~1.0 ml

13. Measure the protein concentration by Nano-Drop, and freeze them by liquid nitrogen. Run a SDS-PAGE gel to check the elution sample.

*Starting Buffer (2L) = 20mM TRIS pH 8.0, 150mM NaCl, Washing Buffer (1L) = 20mM TRIS pH 8.0, 150mM NaCl, 20mM Imidazole, Elution Buffer (200 ml) = 20mM TRIS pH 8.0, 150mM NaCl, 500mM Imidazole, Storage Buffer (1L) = 20mM TRIS pH 8.0.*
(3) FPLC purification for crystallization

The first part of the FPLC purification protocol (Figure 2-6) is very similar to the batch purification protocol. Instead of using cell lysate, large quantity of *E.coli* cells was lysed via sonication on ice. Subsequently, his-tag protein was purified via a Ni-NTA chelating column (Figure 2-5AB, GE Healthcare) following a protocol with a linear imidazole (10–500 mM) elution gradient (Buffer A= 20 mM Tris/HEPES, 300 mM NaCl, pH 7.4~8.0, 0.5 mM optional reducing agent TCEP, and 0.5-5 mM optional co-factor like PLP). The his-tag was then optionally removed by thrombin protease (Figure 2-5C, Sigma) cleavage and the affinity his-tag was then removed via a second round of Ni-NTA affinity chromatography. Buffer change, if needed, was done by using PD-10 column. Afterwards, The protein solution was concentrated to ~3ml and loaded into Superdex 200 gel filtration (Figure 2-5DE) in buffer A. Finally, protein was concentrated to 10-55 mg/ml, flash frozen in liquid nitrogen, and stored at −80 °C for further crystallization. Proteins WecE, CalS13 in chapter 3, and sMAT mutants in chapter 4/5 were purified using this standard protocol.
Figure 2-5. Example and protein FPLC purification (KDM4B DTD domain)

(a) NI-NTA column. (b) SDS-PAGE gel after NI-NTA column. (c) SDS-PAGE gel after thrombin cleavage to check molecular weight “shift”. (d) Gel filtration. (e) SDS-PAGE gel after gel filtration. (f) Crystal pictures of KDM4B under regular light and UV.
Figure 2-6. General process of protein cloning, expression and purification

(1) cloning constructs. (2) E.coli expression. (3) Ni-NTA purification. (4) thrombin cleavage. (5) optional 2nd Ni-NTA column. (6) gel filtration and SDS-PAGE gels.
2.2.4. Crystallization methods and conditions

After protein FPLC purification, protein solution was adjusted to 10-20 mg/ml for the initial screening. 16 commercial 96-well screen kits were used for the sitting drop initial screening, as described in 2.1.4, with the help of Mosquito equipment. Crystallization plates were then placed in a proper temperature between 4-20°C, and were then checked under microscope several times (e.g. after 1 day, 3 days, 1 week and 3 weeks). Protein crystals can be validated from salt crystals or other contaminations by using UV microscope (Figure 2-5F) or protein dye.

Crystal optimizations, if needed, were either conducted using similar 96-well plate set-ups using the Mosquito, or using a 24-well plate with a larger reservoir and drop volume. The optimization stock buffers were purchased from the same vendor that made the initial screen to best guarantee reproducibility, and stocks were mixed using liquid handing equipment (Scorpion). Afterwards, target crystals were harvested, flash frozen into liquid nitrogen and sent out to synchrotron beamline for data collection.

A detailed list of crystallization conditions and cryoprotectant conditions for all my protein data bank depositions can be found in the Table 2-2.
**Table 2-2. Crystallization and cryoprotectant conditions**

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Solution Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>4HPV, 4KOB, 4L2Z, 4L7I</td>
<td>sMAT Solution (6.8-13.5 mg/ml protein, 25mM Tris pH 8.0) mixed in a 1:1 ratio with the well solution (1.1-1.4M Sodium phosphate monobasic monohydrate / Potassium phosphate dibasic, pH 5.6) Cryoprotected with 20% DMSO</td>
</tr>
<tr>
<td>4WS9</td>
<td>sMAT-N159G Solution (6.9 mg/ml protein, 25mM Tris pH 8.0) mixed in a 1:1 ratio with the well solution (0.1 M Sodium Acetate, 0.1 M MES pH6.5, 30% (w/v) PEG 2000 MME), cryoprotected with 27% PEG 2000 MME and 10% glycerol</td>
</tr>
<tr>
<td>4PIW</td>
<td>Se-Met WecE solution (10–25 mg/ml, HEPES pH 7.5, 150 mM NaCl) mixed in a 1:1 ratio with the well solution (10% v/v 2-Propanol, 0.1 M Sodium Citrate pH 5.0, 20% (w/v) PEG400), cryoprotected with additional 15% glycerol</td>
</tr>
<tr>
<td>4ZAH</td>
<td>WecE solution (10–25 mg/ml, 25 mM Tris pH 8, 150 mM NaCl) mixed in a 1:1 ratio with the well solution (10% v/v 2-Propanol, 0.1 M Sodium Citrate pH 5.0, 20% (w/v) PEG400), cryoprotected with additional 15% glycerol</td>
</tr>
<tr>
<td>4ZAS</td>
<td>CalS13 solution (14 mg/ml, 25 mM Tris pH 8, 150 mM NaCl) mixed in a 1:1 ratio with the well solution (200 mM ammonium sulfate, 100 mM Bis-Tris, pH 6.5, 25% w/v PEG 3350), cryoprotected with additional 15% glycerol</td>
</tr>
<tr>
<td>4XAU</td>
<td>AtmS13 solution (10-30 mg/ml, 25 mM Tris pH 8, 150 mM NaCl) mixed in a 1:1 ratio with the well solution (3.5 M sodium formate pH 7.0), cryoprotected with additional 20% glycerol</td>
</tr>
<tr>
<td>4FZR, 4G2T</td>
<td>SsfS6 Solution (11.4 mg/ml SsfS6 protein, 50mM Tris pH 8.0) mixed in a 1:1 ratio with the well solution (1.1–1.40 M Sodium phosphate monobasic monohydrate / Potassium phosphate dibasic, pH 8.2) Cryoprotected with 100% Paratone-N oil</td>
</tr>
<tr>
<td>5EEG</td>
<td>DnrK solution (20 mg/ml, 25 mM Tris pH 8, 150 mM NaCl) mixed in a 1:1 ratio with the well solution (100 mM magnesium formate dihydrate and 15% (w/v) PEG 3,350), cryoprotected with additional 10% glycerol</td>
</tr>
<tr>
<td>5EEH</td>
<td>DnrK solution (20 mg/ml, 25 mM Tris pH 8, 150 mM NaCl) mixed in a 1:1 ratio with the well solution (126M ammonium sulfate, 0.1M Tris pH 8.0, 0.2M lithium sulfate), cryoprotectant with additional 10% glycerol</td>
</tr>
<tr>
<td>4M60</td>
<td>OleO solution (10-15 mg/ml protein, 25mM Tris pH 7.5 and 150mM NaCl) mixed in a 1:1 ratio with the well solution (45% polyacrylate 2100, sodium salt, 0.1M HEPES pH 6.5) Cryoprotected with 45% polyacrylate 2100, sodium salt</td>
</tr>
<tr>
<td>5D6W</td>
<td>KDM4A-DTD solution (30mg/ml KDM4A DTD, 20mM HEPES pH 7.5, 150mM NaCl and 0.5mM TCEP) mixed in a 1:1 ratio with the well solution (1.0M ammonium tartrate dibasic and 0.1M sodium acetate trihydrate pH4.6) Cryoprotected with additional 20% glycerol</td>
</tr>
<tr>
<td>5D6X</td>
<td>KDM4A-DTD solution (30mg/ml KDM4A DTD, 20mM HEPES pH 7.5, 150mM NaCl and 0.5mM TCEP) mixed in a 1:1 ratio with the well solution (1.0M ammonium sulfate, 1% w/v peg3350 and 0.1M Bis-Tris pH5.5) Cryoprotected with additional 20% glycerol</td>
</tr>
<tr>
<td>5D6Y</td>
<td>protein-peptide solution (30mg/ml KDM4A DTD and 3.5mg/ml peptide in 20mM HEPES pH 7.5, 150mM NaCl and 0.5mM TCEP) mixed in a 1:1 ratio with the well solution (1.2M sodium phosphate monobasic, 0.8M potassium phosphate dibasic, 0.1M CAPS/sodium hydroxide pH10.5, 0.2M lithium sulfate) Cryoprotected with additional 20% glycerol</td>
</tr>
<tr>
<td>4UC4</td>
<td>KDM4B-DTD solution (30mg/ml KDM4B DTD, 20mM HEPES pH 7.5, 150mM NaCl and 0.5mM TCEP) mixed in a 1:1 ratio with the well solution (0.66M Divalent, 0.1M Buffer pH8.5 and 30% w/v EDO/PEG8K) Cryoprotected with additional 20% glycerol</td>
</tr>
</tbody>
</table>
2.2.5. X-ray Diffraction data collection

X-ray diffraction data were mostly collected remotely from the beamline station at Advanced Photon Source, Chicago. The most frequently used beamline stations are LS-CAT (Life Sciences Collaborative Access Team, 21-ID-D, F and G). I also remotely collected data in GM/CA for 2 times. In addition, SBC-CAT and LRL-CAT stations both kindly helped me collect data once.

2.2.6. Structure determination methods

First, collected datasets were indexed and scaled by HKL-3000\(^8\) or XDS\(^9\). All or only part of the frames was processed to gain a best combination of B-factors and completeness. Resolution cut was determined by a combination factors of I/sigma, completeness, R\(_{\text{merge}}\) and CC\(^{1/2}\). For structure solution of high-resolution Se-Met protein without proper molecular replacement model, like sMAT, Phenix.Hyss was used for determination of selenium atom substructure, Autosol for phasing and Phenix.autobuild for model building\(^{29}\). For the other proteins with a proper homology model, molecular replacement was utilized using one or multiple known ensemble structures as a starting model. The structures including several double conformations were manually rebuilt in several rounds by Coot\(^{11}\) and further refined by Phenix\(^{10}\). MolProbity\(^{20}\) was used to validate the quality of the coordinates. Structural figures were generated using PyMOL\(^{12}\), QuteMOL\(^{13}\), and Adobe Illustrator.
2.2.7. Ensemble refinements

Ensemble refinements of high-resolution X-ray data were done using the standard protocol of command-line based Phenix ensemble refinement\textsuperscript{30}, plus a mild global harmonic restrain of the whole macromolecular. The harmonic restrain was added as a penalty score function, to mildly punish the refinement for pushing the coordinates from its original states longer than a certain distance (e.g. 1 Å). Without this restrain, the refinement significantly increased the flexibility of a protein region with weak corresponding electron density, and this ended up with a lower $R$ but a higher $R_{free}$, and unclean difference maps. Adding this mild restrain would avoid this. Three adjustable parameters - ptls, tx and temperature offset - were also thoroughly tested and carefully picked as suggested in the original publication\textsuperscript{30}.

2.2.8. Phosphate release assay

Phosphate release assay was conducted using a commercial kit named “Enzchek” made by Life Technology\textsuperscript{4}. All the stocks were made following the kit protocol (E-6646). The only variation is I used a smaller reaction volume (100-200μL). Briefly, the steps are: (1) make an Enzchek reaction master mix; (2) make an enzyme master mix (e.g. sMAT mutant and its required reaction substrates except for methionine analogs) (3) Add methionine analogs into the 96-well plate; (4) Add enzyme master mix into the wells using channel pipette; (5) Add Enzchek reaction master mix using channel pipette and quickly mix the wells; (6) incubate at 22°C for 30 min; (7)
measure the absorbance at 360 nm using a plate reader. Notably, the ATP concentration in the enzyme master mix was calculated to have a maximum release of 150\(\mu\)M inorganic phosphate, which is corresponding to the maximum sensitivity range of the Enzchek assay. One example of using Enzchek phosphate release assay is shown in Figure 2-7. Several sMAT mutants made were tested with a relatively small methionine analog library (S1-S17 are ligands selected from Table 2-1).

2.2.9. Thermal shift assay

Several thermal shift assays were conducted using Sypro Orange dye and Bio-Rad CFX qPCR Systems following the standard protocols\(^{31}\). Before testing with any expensive ligands, I conducted a few gradient tests first in terms of protein concentration, buffer(s) pH and salt concentration to find the best suitable condition for this protein. A clear sharp peak with a higher \(T_m\) is a good indicator of suitable condition. Notably, not all proteins are suitable for this assay, and results from this assay can only serve as a preliminary screening guidance and should be validated by other further solid assays.

2.2.10. Mutagenesis

Mutagenesis is conducted using Q5 Site-Directed Mutagenesis Kit with the standard protocol available online (NEB, E0554S).
Figure 2-7. Examples of Enzchek phosphate release assay

(a) A glance of several well components on the plates; (b) Phosphate standard control; (c) Phosphate release results were converted to substrate formation percentage for sMAT mutants; (d) Comparison of EB14 with WT and N159G.
2.3. PDB Depositions

A glance of all my PDB depositions during Ph.D. thesis is shown in Figure 2-8. Below is a list of the 4-digit codes of the entries:

4ZAS  Crystal structure of CalS13 from *Micromonospora echinospora*

4PIW  Crystal structure of sugar aminotransferase WecE from *Escherichia coli K-12*

4ZAH  Crystal structure of sugar aminotransferase WecE with External Aldimine VII

4ZWV  Crystal Structure of Aminotransferase AtmS13 from *Actinomadura melliaura*

4XAU  Crystal structure of AtS13 from *Actinomadura melliaura*

4FZR  Crystal Structure of SsfS6, *Streptomyces sp. SF2575* glycosyltransferase

4G2T  Crystal Structure of *Streptomyces sp. SF2575* glycosyltransferase SsfS6, complex with thymidine diphosphate

4M60  Crystal structure of macrolide glycosyltransferases OleD

4HPV  Crystal structure of methionine adenosyltransferase from *S. solfataricus*

4K0B  Crystal structure of methionine adenosyltransferase from *Sulfolobus solfataricus* complex with AdoMet and P Pi

4L2Z  Crystal structure of methionine adenosyltransferase from *Sulfolobus solfataricus* complex with AdoEth and P Pi

4L7I  Crystal structure of methionine adenosyltransferase from *Sulfolobus solfataricus* complex with AdoMet and P Pi

4WS9  Crystal structure of sMAT N159G from *Sulfolobus solfataricus*

5EEG  Crystal structure of carminomycin-4-O-methyltransferase DnrK in complex with Ado'Hcy

5EEH  Crystal structure of carminomycin-4-O-methyltransferase DnrK in complex with AdoHcy and 2-chloro-4-nitrophenol
4UC4  Crystal structure of hybrid tudor domain of human demethylase KDM4B
5D6W  Crystal structure of double tudor domain of human demethylase KDM4A
5D6X  Crystal structure of double tudor domain of human demethylase KDM4A
5D6Y  Crystal structure of double tudor domain of human lysine demethylase KDM4A complex with histone H3K23me3
4M7P  Ensemble refinement of protein crystal structure of macrolide glycosyltransferases OleD
4M83  Ensemble refinement of protein crystal structure (2IYF) of macrolide glycosyltransferases OleD complex with UDP and Erythromycin A
4QA9  Ensemble refinement of an epoxide hydrolase from *Streptomyces carzinostaticus subsp. neocarzinostaticus*.
4XQ2  Ensemble refinement of cystathione gamma lyase (CalE6) D7G from *Micromonospora echinospora*
4Q29  Ensemble Refinement of plu4264 protein from *Photorhabdus luminescens*
3UTN  Crystal structure of Tum1 protein from *Saccharomyces cerevisiae*
4H86  Crystal structure of Ahp1 from *Saccharomyces cerevisiae* in reduced form
4KN8  Crystal structure of Bs-TpNPPase
4IFT  Crystal structure of double mutant thermostable NPPase
4IG4  Crystal structure of single mutant thermostable NPPase (N86S)
4O8C  Structure of the H170Y mutant of thermostable p-nitrophenylphosphatase
Figure 2-8. A glance of all my PDB depositions during Ph.D. thesis, grouped by chapters and topics.
Reference


20. Chen, V. B. et al. MolProbity: All-atom structure validation for macromolecular


Chapter 3

Glycorandomization: Sugar Aminotransferases and Glycosyltransferases

Natural products and their derivatives continue to play an important role in drug discovery. Many natural products are glycosylated secondary metabolites, and the sugar part(s) have been revealed to affect drug targeting, pharmacology and biological activity. So altering glycosylation patterns on clinical natural products has a very promising potential for the generation of novel therapeutics: one of those approaches is “glycorandomization”\textsuperscript{1–17}. This chapter will describe my research contributions to glycorandomization in two aspects: knowledge contribution to the randomized sugar library by studying sugar aminotransferases in sections 3.1 and 3.2; toolbox development of sugar installation enzymes (glycosyltransferases SsS6 and OleD) in section 3.3 and 3.4.
3.1. Structural basis for the stereochemical control of amine installation in nucleotide sugar aminotransferases

Sugar aminotransferases (SATs) are an important class of tailoring enzymes that catalyze the 5’-pyridoxal phosphate (PLP)-dependent stereo- and regiospecific installation of an amino group from an amino acid donor to a corresponding ketosugar nucleotide acceptor. In this section I will talk about the strategic structural study of two homologous C4 SATs (Micromonospora echinospora CalS13 and Escherichia coli WecE) that utilize identical substrates but differ in their stereochemistry of aminotransfer. This study reveals for the first time a new mode of SAT sugar nucleotide binding and, in conjunction with previously reported SAT structural studies, provides the basis from which to propose a universal model for SAT stereo- and regiochemical control of amine installation. Specifically, the universal model put forth highlights catalytic divergence to derive solely from distinctions within nucleotide sugar orientation upon binding within a relatively fixed SAT active site where the available ligand bound structures representative C4 SAT examples provide a basis for the overall model. Importantly, this study presents a new predictive model to support SAT functional annotation, biochemical study and rational engineering. A paper has been published about this work: the structural part was accomplished by me in the Phillips lab, and the biochemical experiments were conducted by a nice collaboration the Thorson lab.

*Portions of chapter 3.1 have been previously published as: Wang F, Singh S, Kim Y, Xu W, Helmich KE, Miller MD, Cao H, Bingman CA, Thorson JS, Phillips Jr GN. The structural basis for the stereochemical control of amine installation in nucleotide sugar aminotransferases. ACS Chem Biol. 2015 Sep 18;10(9):2048-56
3.1.1. Aminosugars and sugar aminotransferases (SATs)

Aminosugars are prevalent in nature, where they function as precursors in primary metabolism, building blocks within biological macromolecular structures, and key contributors to biological molecular recognition events that dictate remarkably divergent phenomena such as small molecule targeting, signaling, bacterial pathogenicity, and immunological definition.\textsuperscript{28-37} Biosynthetically, sugar amine installation is catalyzed by 5′-pyridoxal phosphate (PLP)-dependent sugar aminotransferases (SATs) belonging to the aspartate aminotransferase fold type I superfamily (AAT-I) (E.C. 2.6.1.X) that use an amino acid as the amino donor and a free ketosugar or ketosugar nucleotide as the amino acceptor\textsuperscript{38-44} (Figure 3-1A). In recent years, the crystal structures of several sugar nucleotide-dependent SATs from bacteria have been described.\textsuperscript{45-53} Yet, given the diversity of substrates employed by the SATs structurally interrogated to date (Figure 3-1B), the general features that control the regio- and/or stereospecificity of an SAT-catalyzed amine installation remain poorly understood. To address this gap in knowledge, herein I describe a structural study of two highly homologous sugar nucleotide-dependent SATs (CalS13 and WecE) that share common substrates (TDP-4-keto-4,6-dideoxy-α-d-glucose as the acceptor and L-Glu as the amino donor) but differ in the stereochemistry of the C-4 amine installation (Figure 3-1C). Of these, CalS13 was demonstrated to function as a key SAT en route to the calicheamicin aryltetrasaccharide in \textit{Micromonaspora echinospora} by comparative genomics, genetic complementation, and in vitro biochemical characterization to provide the corresponding C-4 (S) configuration\textsuperscript{54-57} (Figure 3-1C). In contrast, WecE provides the corresponding C-4 (R) configuration.
and serves as an integral SAT in lipopolysaccharide biosynthesis.\textsuperscript{56,58} The strategic structural comparison highlighted herein provides, for the first time, a uniform structural model for the molecular basis of SAT amine installation stereospecificity wherein \( C-4 \, (R) \) installation is accomplished via rotation of the sugar through two distinct strategies. Specifically, in WecE, the nucleotide portion of the substrate is bound in a nearly 180° opposed orientation to that in CalS13 wherein this “nucleotide flip” mechanism in WecE affords top face amine installation to achieve the \( C-4 \, (R) \) product. This is a novel mechanism compared to that previously put forth based upon the ligand-bound structures of ArnB and PseC,\textsuperscript{41,46,49,51} SATs which bind the nucleotide portion of the substrate in the same manner as CalS13, wherein top face amine installation is afforded via simple rotation of the sugar (i.e., a “sugar flip” mechanism). Thus, the combined “nucleotide flip” and “sugar flip” models provide a universal blueprint for nature’s stereochemical control of SAT amine installation and a potential starting point for rational SAT engineering.

3.1.2. Overall structures

The crystal structure of the CalS13/PLP/TDP-4-keto-4,6-dideoxy-\( \alpha \)-d-glucose ternary complex was determined at 2.47 Å resolution (Table 3-1, PDB 4ZAS). This ternary complex was found to belong to the \( P2 \) space group with six subunits in the asymmetric unit. All six subunits contained PLP as internal aldimines of Lys202. One subunit had sufficient electron density to model the sugar nucleotide substrate (Figure 3-2A), three had electron density for just TDP, while the remaining two lacked
Figure 3-1. Schemes of SATs including CalS13 and WecE

(A) Scheme representing a standard SAT-catalyzed sugar transamination reaction.

(B) Aminosugar nucleotide products of SATs structurally characterized to date where the amino group is colored blue (equatorial or axial designation). (C) Reactions catalyzed by CalS13 and WecE en route to calicheamicin and LPS biosynthesis, respectively.
interpretable electron density in the sugar nucleotide site. The root-mean-square deviation (r.m.s.d.) among different asymmetric unit members bound to PLP, PLP, and TDP; PLP and TDP-4-keto-4,6-dideoxy-α-d-glucose ranged between 0.21 and 0.25 Å, where only minor changes in active site side chain residues upon ligand binding were observed.

WecE cocrystallization led to WecE/PLP/TDP-4-amino-4,6-dideoxygalactose ternary complexes that diffracted at 2.70 Å (Lys181 internal aldimine-containing) and 2.24 Å resolution (sugar nucleotide external aldimine-containing), belonging to the P2 and P1 space groups, respectively (Table 3-1, PDB: 4PIW and 4ZAH, respectively), with each containing eight subunits in the asymmetric unit. The presence of the external aldimine (Figure 3-2A) is consistent with the well-established reversibility of SATs and has been observed in other SAT structures. Consistent with the comparisons among other apo and ligand-bound SATs structures, there is very little conformational change between the WecE internal aldimine and external aldimine complexes (r.m.s.d. 0.31 Å). These are reserved to a slight upward shift of the pyridinium ring into the active site and very small changes in the orientation of active site residue ligand-interacting side chains.
enzymes, multimeric state calculation by PDBePISA share many structural characteristics. Consistent with AAT-I ring into the active site and very small changes in the aldime and external aldime complexes (r.m.s.d. 0.31 Å). These are reserved to a slight upward shift of the pyridinium (sugar nucleotide external aldime-containing), belonging to (Lys181 internal aldime-containing) and 2.24 Å resolution were observed. Glucose ranged between 0.21 and 0.25 Å, where only minor dimer interface comprising an area of interface) and structures.

Table 1. Summary of Crystal Parameters, Data Collection, and Refinement

<table>
<thead>
<tr>
<th></th>
<th>Se-Met WecE</th>
<th>Native WecE</th>
<th>CalS13</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Crystal parameters</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Space group</td>
<td>P2</td>
<td>P1</td>
<td>P2</td>
</tr>
<tr>
<td>Unit-cell parameters (Å, deg)</td>
<td>125.3, 87.5, 161.6</td>
<td>88.1, 104.0, 109.4</td>
<td>71.6, 88.7, 189.8</td>
</tr>
<tr>
<td></td>
<td>90.0, 91.1, 90.0</td>
<td>72.1, 73.6, 74.1</td>
<td>90.0, 96.2, 90.0</td>
</tr>
<tr>
<td><strong>Data collection statistics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>Resolution range (Å)</td>
<td>50–2.70</td>
<td>50–2.24</td>
<td>50–2.47</td>
</tr>
<tr>
<td>No. of reflections (measured/unique)</td>
<td>275,071/90,907</td>
<td>1,781,851/233,383</td>
<td>744,102/84,933</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>89.4 (89.5)</td>
<td>99.7 (99.2)</td>
<td>100.0 (100.0)</td>
</tr>
<tr>
<td>Rmerge</td>
<td>0.112 (0.545)</td>
<td>0.253 (1.460)</td>
<td>0.117 (1.571)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>3.0 (2.9)</td>
<td>7.6 (7.5)</td>
<td>8.8 (8.1)</td>
</tr>
<tr>
<td>Mean I/σ (I)</td>
<td>10.31 (1.97)</td>
<td>5.58 (1.33)</td>
<td>11.61 (1.89)</td>
</tr>
<tr>
<td>CC1/2</td>
<td>0.99 (0.70)</td>
<td>0.99 (0.74)</td>
<td>0.99 (0.65)</td>
</tr>
<tr>
<td><strong>Refinement and model statistics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rmerge/Rfree</td>
<td>0.240/0.273</td>
<td>0.204/0.246</td>
<td>0.174/0.229</td>
</tr>
<tr>
<td>Ligands RSCC</td>
<td>0.91</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>RMSD bonds (Å)</td>
<td>0.002</td>
<td>0.010</td>
<td>0.008</td>
</tr>
<tr>
<td>RMSD angles (deg)</td>
<td>0.50</td>
<td>1.29</td>
<td>1.23</td>
</tr>
<tr>
<td>B factor - protein/ligand/solvent (Å2)</td>
<td>41.3/37.3/35.4</td>
<td>62.1/64.7/47.7</td>
<td>58.3/58.6/53.2</td>
</tr>
<tr>
<td>No. of protein atoms</td>
<td>22,501</td>
<td>23,155</td>
<td>17,112</td>
</tr>
<tr>
<td>No. of waters</td>
<td>288</td>
<td>686</td>
<td>502</td>
</tr>
<tr>
<td>No. of auxiliary molecules in the asymmetric unit'</td>
<td>8 LLP</td>
<td>8 TSK, 8 PMP</td>
<td>6 LLP, 3 TYD, 1 T46</td>
</tr>
<tr>
<td><strong>Ramachandran plot (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Favorable region</td>
<td>95.4</td>
<td>98.0</td>
<td>98.0</td>
</tr>
<tr>
<td>Additional allowed region</td>
<td>3.7</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Disallowed region</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>PDB</strong></td>
<td>4PIW</td>
<td>4ZZH</td>
<td>4ZAS</td>
</tr>
</tbody>
</table>

Values in parentheses are for the highest resolution shell. *$R_{merge} = \sum \sum |I(h)| - \langle |I(h)| \rangle / \sum \sum |I(h)|$, where $I(h)$ is the intensity of an individual measurement of the reflection and $\langle |I(h)| \rangle$ is the mean intensity of the reflection. **$R_{rfree} = \sum |F_{o} - F_{c}| / \sum |F_{o}|$, where $F_{o}$ and $F_{c}$ are the observed and calculated structure-factor amplitudes, respectively. †$R_{merge}$ was calculated as $R_{merge}$ using 5.0% of randomly selected unique reflections that were omitted from the structure refinement. ‡Ligand RSCC: ligand real-space correlation coefficient. §Ligand abbreviations: LLP (s'-pyridoxal-lysine-S'-monophosphate), TK5 (TDP-4-amino-4,6-dideoxy-α-D-galactose external aldime), PMP (pyridoxamine-S'-phosphate), TYD (thymidine-S'-diphosphate), and T46 (TDP-4-keto-4,6-dideoxy-α-D-glucose).
Figure 3-2. Overall structures and ligand maps of CalS13 and WecE

(A) The simulated annealing $m$Fo–$DFc$ omit maps highlighting ligands in the active sites of CalS13 and WecE. (B) Overall structure of CalS13 (top panel) and WecE (bottom panel) in complex with ligands.
Both CalS13 and WecE belong to the AAT-I family\textsuperscript{41,42} and share many structural characteristics. Consistent with AAT-I enzymes, multimeric state calculation by PDBe PISA\textsuperscript{60} indicated both CalS13 and WecE to be homodimers with an extensive dimer interface comprising an area of $\sim 4564 \text{ Å}^2$ (CalS13 dimer interface) and $\sim 4405 \text{ Å}^2$ (WecE dimer interface), respectively (Figure 3-2B). Each dimer assembly contains two active sites set apart by $\sim 28$–$30 \text{ Å}$ (based upon the distance between the pyridinium rings of bound PLP), with both monomers contributing critical residues to each of the active sites (Figure 3-2B). Analogous to other AAT-I enzymes, CalS13 and WecE have a conserved active site lysine (Lys202 and Lys181 in CalS13 and WecE, respectively) that forms the requisite internal aldimine with PLP, a conserved general acid (Asp173 and Asp152 in CalS13 and WecE, respectively) that is key to cofactor activation in the context of AAT-catalyzed transamination and a putative base (Gln176 and Gln155 in CalS13 and WecE, respectively). The overall architecture of each subunit is composed of a N-terminal “large domain” (colored yellow in Figure 3-2B) and a C-terminal “small domain” (colored blue in Figure 3-2B). The large domain contains a mixed $\beta$-sheet formed by seven $\beta$-strands (strand order $\beta_1$, $\beta_7$, $\beta_6$, $\beta_5$, $\beta_4$, $\beta_2$, $\beta_3$) and eight $\beta$-strands (strand order $\beta_1$, $\beta_7$, $\beta_6$, $\beta_5$, $\beta_4$, $\beta_2$, $\beta_3$, $\beta_8a$) in CalS13 and WecE, respectively, where $\beta_7$ is antiparallel to the rest of the $\beta$-strands in the $\beta$-sheet. This large $\beta$-sheet is flanked by $\alpha$-helices on both sides. The C-terminal small domain is formed by a two-stranded antiparallel sheet ($\beta$-strands $\beta_8$ and $\beta_9$) surrounded by helices. The N-terminal large and C-terminal small domains are linked by a three stranded antiparallel $\beta$-hairpin. The size of the 13 $\alpha$-helices in each subunit ranges
from 3 to 31 residues with the largest α-helix (α8, containing a kink near residues 13–14) contributing to both the large and small domains (Figure 3-2B).

3.1.3. Cofactor and nucleotide sugar binding sites

The PLP-binding site is situated within a deep cleft of the active site of each monomer. The cofactor is oriented such that the pyridinium ring is positioned above the central β-sheet of each monomer unit (Figure 3-2B), while the PLP phosphate moiety points toward the dimer interface. The pyridinium ring and C6 methyl of PLP are positioned for π–π stacking and hydrophobic interactions with the side chains of Phe96 and Val141 in CalS13 (Figure 3-3A) and Phe81 and Val126 in WecE, respectively (Figure 3-3B). The pyridoxal C5 hydroxyl also forms a hydrogen bond with the side chain amine of Gln176 in CalS13 and Gln155 and Tyr321 in WecE, respectively, while the cofactor N1 ring nitrogen is anchored by hydrogen bonds with the side chain of the previously discussed general acid active residue (Asp173, CalS13; Asp152, WecE) and backbone amide of Thr99 and Thr84 in CalS13 and WecE, respectively (Figure 3-3). The PLP phosphate is held in place by several direct and indirect water-mediated intrasubunit (Gly62, Thr63, and Ser197 in CalS13; Cys55, Thr56, and Ser176 in WecE) and intersubunit (Arg250 in CalS13; Ser232 in WecE) hydrogen bonding interactions (Figure 3-3A-B).
Figure 3-3. Cofactor and nucleotide sugar binding sites of CalS13 and WecE

Cofactor binding site of (A) CalS13 and (B) WecE. Nucleotide sugar binding site of (C) CalS13 and (D) WecE. (E) Overlay of the orientation of binding of TDP-4-keto-4,6-dideoxy-α-d-glucose from CalS13 and TDP-4-amino-4,6-dideoxy-α-d-galactose as external aldimine from WecE in the active site of CalS13.
The thymidine binding pockets in CalS13 and WecE are remarkably opposed to one another by nearly 180° and this distinguishing feature defines the stereochemical outcome of aminotransfer by virtue of the sugar orientation (bottom face installation in CalS13 versus top face installation in WecE). Specifically, these distinct nucleotide interaction loops are L0 (residues Ser13-Ala16) and L1 (residues Gly32-Gly37) in CalS13 and L7 (residues Arg213-Thr225) and L9 (residues Val318-Ile322) in WecE, respectively (Figure 3-3C-E). Key thymidine hydrogen bonding interactions include the thymidine exocyclic O4 with the L1 backbone amide of Leu34 and Glu35 in CalS13 and the L7 backbone amide of Tyr224 in WecE, respectively, and the thymidine endocyclic N3 with the L1 backbone carbonyls of Gly32 and L7 Tyr224 in CalS13 and WecE, respectively. Additional stability is provided by thymine aromatic stacking contributions from the L1 side chain of Arg33 and the L7 side chain of Tyr224 of CalS13 and WecE, respectively. The nucleotide ribose moiety in both CalS13 and WecE is solvent exposed and does not make any direct contacts with the enzyme. The pyrophosphate of the nucleotide sugar extends away from the ribose moiety through a deep protein cavity that leads to the cofactor binding site of the protein and is anchored to the enzyme by direct hydrogen bonding interactions with the guanidinium side chains of L7 of Arg233 and Arg237 in CalS13 (Figure 3-3C) and the β9 Arg352 L7, Arg213, and L7 side chain hydroxyl of Tyr224 of WecE (Figure 3-3D), respectively. Interestingly, CalS13 also contains a structurally equivalent L7 tyrosine (Tyr242), but this residue does not interact with the substrate pyrophosphate. Direct sugar contacts are apparent in WecE (Figure 3-3D) and include hydrogen bonds of the pyranose O2 and O3 with the imidazole ring of His320 (loop L9) (Figure 3-3D). In
contrast, only water-mediated contacts are observed in CalS13 between the pyranose \( O2 \) and the L1 side chain carboxylate of Glu35 (Figure 3-3C). Cumulatively, the “base flip” observed in WecE as compared to CalS13 dictates the divergence in the stereochemical outcome of aminotransfer among these two highly homologous enzymes (Figure 3-3E), where the key pyrophosphate interactions may direct the divergence in sugar nucleotide orientation.

3.1.4. A structure-based model for the stereochemical outcome

To date, crystal structures of nine other sugar nucleotide-dependent SATs have been determined. These include the following: DesI and DesV from \( S. \) venezuelae involved in TDP-desosamine biosynthesis;\(^\text{45,46}^\) Per from \( C. \) crescentus CB15 involved in the GDP-perosamine biosynthesis;\(^\text{47}^\) QdtB from \( T. \) thermosaccharolyticum involved in the biosynthesis of TDP-mycaminose;\(^\text{52}^\) ArnB from \( S. \) typhimurium involved in the biosynthesis of UDP-4-amino-l-arabinose;\(^\text{49}^\) PseC from \( H. \) pylori involved in the biosynthesis of UDP-pseudaminic acid;\(^\text{51}^\) AtS13 from \( A. \) melliaura, a putative C4-aminotransferase involved in the biosynthesis of TDP-4-alkylamino-2,4-dideoxypentose of AT2433 (PDB 4RXK and 4XAU);\(^\text{61}^\) WbpE from \( P. \) aeruginosa PAO1 involved in the biosynthesis of UDP-ManNAc-(3NAc)A;\(^\text{48}^\) and PglE from \( C. \) jejuni (PDB: 1061, 1062, 1069) involved in the biosynthesis of UDP-4-amino-4,6-dideoxy-\( \alpha \)-d-GlcNAc. Of these, CalS13, DesI, Per, and PglE catalyze C4 equatorial amine installation to provide the corresponding C4-S stereocenter, while WecE, PseC, and ArnB catalyze C4 axial amine installation to provide the corresponding C4-R (WecE and PseC) or C4-
S (ArnB) stereocenter. The structure elucidation of two highly homologous C4 SATs that act upon an identical substrate but differ solely in their stereochemistry of amine installation adds to this growing set of SAT structures and presents critical new information from which a universal SAR model can be proposed as highlighted below. Importantly, an analysis of all nine SATs reveals structural conservation of the key residues involved in catalysis (Figure 3-4); specifically, the active site lysine (Lys202 and Lys181 in CalS13 and WecE, respectively) that forms the requisite PLP internal aldimine, the conserved general acid (Asp173 and Asp152 in CalS13 and WecE, respectively), and a conserved putative active site base (Gln176 and Gln155 in CalS13 and WecE, respectively). This apparent conservation of a fixed active site scaffold across SATs that catalyze diverse stereo- and regiochemically amine installation reactions suggests subtle alterations in the sugar substrate orientation to drive stereochemical divergence. Within this context, a comparison of ligand-bound SAT structures reveals two distinct strategies to control the sugar orientation. Specifically, compared to CalS13, DesI, Per, or PglE, the inverted C4 amine installation catalyzed by PseC and ArnB is accomplished via substrate pyrophosphate rotation to ultimately achieve inverted active site entry of the sugar (i.e., a “sugar flip” mechanism, Figure 3-5). This strategy was first noted by Holden and co-workers via a comparison of PseC and DesI.46 Surprisingly, the current study reveals WecE to accomplish the requisite sugar reorientation via an alternative unprecedented mode of nucleotide binding (i.e., a “base flip” mechanism, Figure 3-5) where WecE stands as the only structurally characterized SAT to date to adopt this alternative nucleotide-binding orientation.
Figure 3-4. Structure based sequence alignment of all SATs with known structures in the PDB. Secondary structural features of CalS13 are illustrated at the top. The numbering of the amino acids in the figure corresponds to CalS13. The residues designated by lowercase letters correspond to residues that are invisible in the structures.
Figure 3-5. Universal model for the stereo- and regiochemical control of an SAT-catalyzed amine installation

For each of the four possible amine installation reactions (C4 equatorial, C4 axial, C3 equatorial, and C3 axial), the fixed active-site orientation is highlighted in relation to distinct sugar orientations.
This model can be further extended to C3 SATs. Specifically, analysis of WdpE, QdtB, and DesV ligand bound structures reveals each to utilize the “sugar flip” mechanism reminiscent of PseC and ArnB to accomplish inverted sugar orientation to that observed for CalS13-like catalysts (Figure 3-5). This is consistent with the corresponding observed C3 equatorial amine stereochemical outcome. Furthermore, C3 SATs adopt a ∼30° twist of the sugar as compared to C4 SAT comparators to enable amine installation chemistry at C3 rather than C4. Based upon this trend, I anticipate the C3 SATs that catalyze C3 axial amine installation to bind their sugar nucleotides in an orientation reminiscent of CalS13-like catalysts but with the required a ∼30° twist of the sugar to enable C3 modification. While many such putative C3 axial SATs exist (DnrJ,62 Snogl,63 and AknZ64), their three-dimensional structures have yet to be reported.

In summary, the strategic selection and structural study of two homologous C4 SATs (CalS13 and WecE) that utilize identical substrates but differ in their stereochemistry of aminotransfer reveals for the first time a new mode of SAT sugar nucleotide binding. Global analysis of the structures deriving from the current study and those previously reported provide the basis from which to propose a universal model for SAT stereo- and regiochemical control of amine installation. Specifically, this universal model highlights catalytic divergence to derive solely from distinctions within nucleotide sugar orientation upon binding within a relatively fixed SAT active site where the available ligand bound structures of the three out of four representative C3 and C4 SAT examples provide a basis for the overall model. Cumulatively, this study and the corresponding universal model provide a blueprint
for future SAT engineering and biochemical study. This work also highlights the importance of structure elucidation for one or more representative members of the $C_3$ axial set to test the SAR hypothesis put forth.
3.2. Structural characterization of AtmS13, a putative sugar aminotransferase involved in indolocarbazole AT2433 biosynthesis

AT2433 from *Actinomadura melliaura* is an indolocarbazole antitumor antibiotic structurally distinguished by its unique aminodideoxypentose-containing disaccharide moiety. The corresponding sugar nucleotide-based biosynthetic pathway for this unusual sugar derives from comparative genomics where AtmS13 has been suggested as the contributing sugar aminotransferase (SAT). Determination of the AtmS13 X-ray structure at 1.50 Å resolution reveals it as a member of the aspartate aminotransferase fold type I (AAT-I). Structural comparisons of AtmS13 with homologous SATs that act upon similar substrates implicate potential active site residues that contribute to distinctions in sugar C5 (hexose vs. pentose) and/or sugar C2 (deoxy vs. hydroxyl) substrate specificity. A paper has been published about this work, and I am an author of this paper. Two crystal structures were solved in this project: the higher resolution one was solved by Youngchang Kim in MCSG and re-refined by me in the Phillips lab, and the lower resolution one was solved by me in the Phillips lab. Moreover, Thorson lab conducted the biochemical assays.

---

3.2.1. AT2433 and sugar aminotransferase AtmS13

Functionalized deoxysugars appended to natural products are often critical to the given metabolite's specificity on a tissue-, cellular-, and/or molecular-level where the corresponding functionalized deoxyhexoses are among the most diverse and best studied to date.\textsuperscript{29,43,69–72} In contrast, while functionalized deoxypentoses serve a similar purpose, the fundamentals of their biosynthesis remain largely uncharacterized. AT2433 (Figure 3-6A) from \textit{Actinomadura melliaura} is a model aminodeoxypentose bearing member of the indolocarbazoles (exemplified also by 1; rebeccamycin, 2; and staurosporine, 3; Figure 3-6A). These actinomycete-derived alkaloids and potent inhibitors of topoisomerase I and kinases are relevant to anticancer, antitubercular, antimalarial, and antiviral drug development.\textsuperscript{73–76} The AT2433 aminodideoxypentose uniquely influences the metabolite's molecular mechanism and solubility.\textsuperscript{77,78} It is also notably found as part of structurally distinct naturally occurring enediynes (calicheamicin, 4; and esperamicin, 5; Figure 3-6B). Comparative genomics of gene loci responsible for biosynthesis of the AT2433,\textsuperscript{55} calicheamicin,\textsuperscript{54} and rebeccamycin\textsuperscript{79} provided the basis for a proposed aminodideoxypentose biosynthetic pathway (Figure 3-6C).\textsuperscript{55,77–79} The gene cluster composition is consistent with previous esperamicin metabolic labeling studies and biosynthetic studies of Gram-negative exopolysaccharide aminopentoses. All these metabolites are involved in antibiotic resistance, biofilm formation and evasion of innate immunity.\textsuperscript{49,80–83} Subsequent biochemical characterization identified CalS8 as a TDP-\(\alpha\)-d-glucose dehydrogenase\textsuperscript{84} and CalS9 as a TDP-\(\alpha\)-d-glucuronic acid decarboxylase\textsuperscript{57} providing further support for this pathway. Of the remaining
putative enzymes encoded by the AT2433 gene cluster, AtmS13 is the only apparent sugar aminotransferase and shares high sequence homology with CalS13 (62% identity and 76% similarity), the requisite calicheamicin TDP-4-keto-6-deoxy-α-d-glucose C4-aminotransferase.\textsuperscript{56,57} The putative AtmS13 substrate (TDP-4-keto-2,6-dideoxy-α-d-xylose) differs from the established CalS13 substrate (TDP-4-keto-6-deoxy-α-d-glucose) at C2 (H vs. OH) and C5 (H vs. CH2OH). Moreover, replacement of CalS13 with AtmS13 in standard CalS13 biochemical assays failed to provide turnover (unpublished data). In an effort to elucidate the molecular determinants that potentially contribute to the anticipated distinct sugar aminotransferase (SAT) TDP-hexose (CalS13) versus TDP-pentose (AtmS13) specificity, herein I describe the X-ray structure determination at 1.50 Å resolution of AtmS13 with covalently attached cofactor, pyridoxal-5’-phosphate (PLP). These studies reveal AtmS13 as a member of aspartate aminotransferase fold type I superfamily (AAT-I). A comparison of the AtmS13 with structurally similar C4-SATs that act upon related substrates highlights potential active site features that may explain substrate preferences at C2 (H vs. OH) and/or C5 (H vs. CH2OH) as a basis for further SAT biochemical and/or engineering studies.
Figure 3-6. Schemes of AtmS13 related compounds

(A) Indolocarbazoles AT2433 (1, X = H or Cl), rebeccamycin 2 and staurosporine 3. (B) Aminopentose-containing enediynes calicheamicin 4, and esperamicin 5 where R1 = methyl, ethyl or isopropyl. (C) Proposed biosynthetic pathway of aminodideoxypentose shared by 1, 4, and 5.
3.2.2. Overall structure

Crystal structures of Se-Met labeled and native AtmS13 with covalent PLP internal aldimine were obtained in two different crystal forms at 1.50 and 3.00 Å resolution, respectively, by co-crystallizing AtmS13 in the presence of PLP (Table 3-2). The PLP adduct of Se-Met and native AtmS13 (PDB entries 4ZWV and 4XAU) belonged to $C_2$ and $R32$:h space groups, respectively. The asymmetric unit contained two subunits bearing a large interaction surface area of ~4,600 Å$^2$ representing the formation of a dimer. Because the root mean square deviation (r.m.s.d.) between the structures of two forms is 0.20 Å, the rest of the discussion in the text refers to the high-resolution structure of Se-Met AtmS13. Both subunits contain a covalent PLP internal aldimine with the side-chain amine of the active site lysine residue Lys187 but adopt alternative poses (PLP1 and PLP2, Figure 3-7A). The overall fold of AtmS13 belongs to the aspartate aminotransferase fold type I superfamily and shares many structural characteristics with AAT-I members. Each dimer of AtmS13 contains two active sites spaced ~28 Å apart, with each monomer contributing critical residues to each active site (Figure 3-7A). Analogous to other AAT-I enzymes, AtmS13 has a conserved lysine (Lys187) that contributes to the PLP cofactor internal aldimine and a conserved general acid (Asp158) that is key to cofactor activation via PLP pyridinium nitrogen protonation. The overall architecture of each subunit comprises a N-terminal ‘large domain’ and a C-terminal ‘small domain’ (Figure 3-7A). The large domain contains a mixed β-sheet formed by seven β-strands (strand order $\beta_1$, $\beta_7$, $\beta_6$, $\beta_5$, $\beta_4$, $\beta_2$, $\beta_3$) where $\beta_7$ is antiparallel.
**Figure 3-7. Overall and active architecture of protein AtmS13 vs CalS13**

(A) Homodimer structure of AtmS13 (PDB 4ZWV). (B) Residues involved in cofactor binding. The alternative PLP orientations (PLP1 and PLP2) are colored green and yellow, respectively. (C) Comparison of active site residues of AtmS13 with CalS13. PLP binding site where the internal aldimine belonging to AtmS13 and CalS13 are colored blue and green, respectively. AtmS13 and CalS13 residues are colored marine blue and yellow, respectively. (D) Sugar nucleotide binding site (AtmS13 residues, marine blue; CalS13 residues, yellow; sugar nucleotide, green with pyrophosphate linkage in orange).
Table 3-2. Summary of Crystal Parameters, Data Collection, and Refinement

<table>
<thead>
<tr>
<th>Crystal parameters</th>
<th>Se-Met AtmS13</th>
<th>Native AtmS13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>C2</td>
<td>R32_h</td>
</tr>
<tr>
<td>Unit-cell parameters (Å, °)</td>
<td>204.1, 73.0, 58.6, 90.0, 103.7, 90.0</td>
<td>223.1, 233.1, 460.9, 90.0, 90.0, 120.0</td>
</tr>
</tbody>
</table>

Data collection statistics

| Wavelength (Å) | 0.97929 | 0.98600 |
| Resolution range (Å) | 29.4–1.50 | 30.0–3.00 |
| No. of reflections (measured/unique) | 560,589/130,016 | 978,914/95,972 |
| Completeness (%) | 97.8 (82.0) | 100.0 (99.0) |
| Redundancy | 4.3 (3.1) | 10.2 (9.5) |
| Mean | 20.6 (3.7) | 7.6 (1.2) |
| CC<sup>1/2</sup> | 0.99 (0.84) | 0.99 (0.51) |

Refinement and model statistics

| R<sub>cryst</sub> | 0.145/0.186 | 0.167/0.220 |
| Ligands RSCC | 0.99 | 0.87 |
| RMSD bonds (Å) | 0.006 | 0.014 |
| RMSD angles (°) | 1.10 | 1.200 |
| B factor—protein/ligand/solvent (Å<sup>2</sup>) | 16.8/14.3/35.3 | 109.5/73.2/84.1 |
| No. of protein atoms | 2,721 | 20,446 |
| No. of waters | 1,155 | 252 |
| No. of auxiliary molecules in the asymmetric unit | 2 LLP<sup>*</sup> | 1 PLP |

Ramachandran plot (%)

| Favorable region | 99.0 | 95.0 |
| Additional allowed region | 1.0 | 4.7 |
| Outliers region | 0.0 | 0.3 |

Values in parentheses are for the highest resolution shell.

<sup>*</sup>R<sub>cryst</sub> = <sup>2</sup> | | \(| I(h) - <I(h)> | / \sum | I(h) | |

<sup>*</sup>R<sub>merge</sub> = \( \sum_i \sum_h | I(h) - <I(h)> | / \sum_i \sum_h | I(h) | 

<sup>*</sup>Mean I/sigma = \| I(h) | / \sum | I(h) | |

<sup>**</sup>R<sub>free</sub> was calculated as R<sub>cryst</sub> using 5.0% of randomly selected unique reflections that were omitted from the structure refinement.

<sup>*RSCC</sup> is the real-space correlation to electron density calculated by Phenix.

<sup>*LLP is the Lys187 PLP internal aldimine.**
3.2.3. Cofactor binding site and comparison with CalS13

The cofactor-binding site of AtmS13 is within a deep cleft in the interior of the active site of each monomer at the two far ends of the length of the dimer interface with two distinct orientation of PLP per monomer observed; both as an internal aldimine with Lys187 (PLP1 colored green and PLP2 colored yellow in Figure 3-7B). The PLP1 and PLP2 sites have 50% occupancy and similar B-factors (14.2 and 14.3) where the PLP phosphate moieties of both occupy the same position but the respective PLP pyridinium ring orientations differ by roughly 120° (Figure 3-7B). Of these two orientations in AtmS13, PLP2 is consistent with the PLP orientation in other structurally characterized SATs containing bound PLP. The former involve π-stacking of the pyridinium ring with Phe81. For the latter, the PLP C5 hydroxyl hydrogen bonds with His161 and the cofactor N1 pyridinium ring nitrogen is anchored by a salt bridge with the conserved general acid (Asp158) (Figure 3-7B). The phosphate group of the cofactor orients toward to the dimer interface and is held in place by several direct hydrogen bonding interactions. These involve the backbone amides of Gly54 and Thr55, the side chain hydroxyls Thr55 and Ser182 and the side chain hydroxyl of Tyr214 from the adjacent subunit (Figure 3-7B). The phosphate group of the cofactor also makes several water mediated interactions. These include the side chain carboxylates of Asp193 and side chain hydroxyl of Tyr214 from the adjacent subunit.
The structure of the biochemically characterized CalS13 involved in calicheamicin aminohexose biosynthesis was described in chapter 3-1. An overlay of the AtmS13 and CalS13 structures resulted in a root mean square deviation of 0.83 Å between the two structures. Upon analysis of the active site of AtmS13 and CalS13, the key PLP interacting residues are largely conserved, a minor exception being the PLP C5 hydroxyl hydrogen bond partner AtmS13 His161 is replaced by CalS13 Gln176 (Figure 3-7C). Comparison of an AtmS13 TDP-4-keto-6-deoxy-α-d-glucose docked model with the CalS13 ternary complex reveals similar conservation of the arginine residues (Arg218/Arg222 and Arg233/Arg237 in AtmS13 and CalS13, respectively) responsible for binding the nucleotide sugar pyrophosphate but a slight variation in the residue that stacks over the thymidine ring (Arg33 in CalS13, Tyr25 in AtmS13) (Figure 3-7D).

Interestingly, few ketosugar-specific interactions are observed in the CalS13 ternary complex or the AtmS13 docked model. Specifically, the single water-mediated hydrogen bonding interaction of the CalS13 side chain carboxylate of Glu35 with the substrate sugar hydroxyl is absent in AtmS13 (Gly27) (Figure 3-7D). The same position is occupied by Asn49 in S. venezuelae DesI (PDB entry 2PO3), a SAT that also utilizes TDP-4-keto-6-deoxy-α-d-glucose as a substrate, suggesting that specific sugar C2 contacts may be required for substrates containing a C2 hydroxyl. Of the other SATs that have been structurally studied, ArnB is the only other pentose C4-SAT. However, compared to AtmS13, ArnB-catalyzed transamination leads to
inversion of $C_4$-stereochemistry where the sugar substrate orientation in ArnB is flipped in the active site to enable stereochemical inversion. Therefore the putative $C_2$ and/or $C_5$ contacts in ArnB do not shed further light on the putative AtmS13/CalS13 substrate divergence. In conclusion, this AtmS13 structural study adds an additional blueprint to the growing body of structural information regarding factors that define substrate specificity within $C_4$-SATs. As such, this work sets the stage for future SAT biochemical and/or engineering studies designed to further probe SAT SAR.
3.3. Crystal structure of SsfS6, the putative C-glycosyltransferase involved in SF2575 biosynthesis\textsuperscript{85}\textsuperscript{*}

The molecule known as SF2575\textsuperscript{85-89} from Streptomyces sp. is a tetracycline polyketide natural product that displays antitumor activity against murine leukemia \textit{P388 in vivo}. In the SF2575 biosynthetic pathway, SsfS6 has been implicated as the crucial C-glycosyltransferase (C-GT) that forms the C-C glycosidic bond between the sugar and the SF2575 tetracycline-like scaffold. Here, I describe the crystal structure of SsfS6 in the free form and in complex with TDP, both at 2.4 Å resolution. The structures reveal SsfS6 to adopt a GT-B fold wherein the TDP and docked putative aglycon are consistent with the overall C-glycosylation reaction. As one of only a few existing structures for C-glycosyltransferases, the structures described herein may serve as a guide to better understand and engineer C-glycosylation. A paper has been published about this work, and I am an author of this paper.\textsuperscript{85} Two crystal structures were solved by me in the Phillips lab, and protein sample was prepared by Maoquan Zhou in the Thorson lab.

\textsuperscript{*}Portions of chapter 3.3 have been previously published as: Wang F, Zhou M, Singh S, Yennamalli RM, Bingman CA, Thorson JS, Phillips GN Jr. Crystal structure of SsfS6, the putative C-glycosyltransferase involved in SF2575 biosynthesis. \textit{Proteins}. 2013 Jul;81(7):1277-82
3.3.1. Natural product glycosyltransferase SsfS6

Natural products and their derivatives are an important resource for the discovery of antibiotics and anticancer drugs. The bioactivity of such natural products can be influenced by enzyme-catalyzed glycosylation and include natural products bearing O-, N-, S- or even C-glycosides. An example of the latter is SF2575, a tetracycline polyketide isolated by Hatsu et al. from Streptomyces sp. SF2575 that displays potent cytotoxicity against P388 leukemia cells in vitro (IC50 = 7.5 ng/mL) and increases the life span of P388 leukemia cell ip-implanted mice in vivo. The biosynthesis of SF2575 has received considerable attention culminating in the work by Tang et al. on the cloning, genetics and annotation of the SF2575 biosynthetic gene cluster and subsequent biochemical study of key enzymes encoded by this newly identified biosynthetic locus. Specifically, several SF2575 intermediates including 3 (Figure 3-8A) have been identified via proton NMR and HRMS and sequence analysis of ssf biosynthetic gene cluster detected only one putative glycosyltransferase, SsfS6, which has high sequence homology to a C-GT HedJ from hedamycin biosynthetic pathway. Based upon this impressive body of work, SsfS6 has been annotated as the unique C-C-bond forming C-GT that catalyzes the transfer of d-olivose from TDP-d-olivose to C9 of the tetracycline-like aglycon (Figure 3-8A).

Glycosyltransferases comprise 94 diverse sequence-based families of enzymes, but they are known to adopt only three different folds (GT-A, GT-B, and predicted transmembrane fold GT-C). The few C-GT structures elucidated to date adopt a GT-B
fold,\textsuperscript{97,98} which consists of two spatially distant $\beta/\alpha/\beta$ Rossmann-like domains. As the novel C-glycoside products synthesized by C-GTs are chemically unique because of their remarkable stability to chemical and enzymatic hydrolysis,\textsuperscript{99} there remains notable interest in the synthesis and application of C-glycoside-containing O-glycoside surrogates.\textsuperscript{93–96} In contrast to O-GTs, mechanistic, in vitro biochemical and/or structural studies of C-GTs are still limited, which severely restricts both our level of understanding, and ability to engineer and/or exploit, these remarkable catalysts.\textsuperscript{97,98,100–102}

Here I describe the crystal structures of the putative C-GT SsfS6 in both apo and thymidine diphosphate (TDP) bound forms. The structure of TDP-bound SsfS6 elucidates the TDP-enzyme interactions and reveals that the loop residue N196 impedes binding of UDP-sugars by sterically occluding the ribose 2'-OH group. Also two potential active site bases implicated for C-glycosyl catalysis were identified. They are consistent with docking studies that highlight potential binding modes of the putative tetracycline-like aglycon. Cumulatively, this study is an additional structural blueprint for the further general study of C-glycosylation and specifically, SF2575.
Figure 3-8. Structure and putative mechanism of SsfS6

(A) Putative biosynthetic pathway of SF2575 (B) Quaternary structure of the SsfS6. (C) TDP in the binding site of SsfS6 with the simulated annealing omit map. (D) Active site comparison between SsfS6 bound to TDP and free form. (E) The proposed mechanism of C-glycosylation catalyzed by SsfS6.
### 3.3.2. Overall structure and active site analysis of SsfS6

X-ray diffraction data were collected to a resolution of 2.4 Å for both apo SsfS6 and the SsfS6-TDP complex (Table 3-3). Although only one SsfS6 monomer was detected in the asymmetric unit, the chain forms a C2-symmetric dimer with another SsfS6 chain in the unit cell (Figure 3-8B), with a buried interface area of 1574 Å² as calculated by PISA. Other members of the GT1 sequence family, such as UrdGT2, CalG3 and SpnG, adopt similar homodimer architectures. The dimeric interface comprises residues within both the N-terminal and the C-terminal domains, including intertwined hydrophobic interactions with a contact center generated by two stacking tryptophan residues Trp22. Thus members of the Thorson and Phillips groups propose that SsfS6 is a dimer in solution. The SsfS6 monomer consists of two β/α/β Rossmann-like domains that face each other and form an active site cleft. The N-terminal domain (residues 1-198) comprises seven parallel β-sheets surrounded by six α-helices while the C-terminal domain (residues 214-383) consists of six parallel β-sheets surrounded by seven α-helices. A long loop (resides 184-213) containing a small conserved α-helix connects the two domains. Several loops are not assignable in the electron density maps of either apo SsfS6 or SsfS6-TDP. A number of SsfS6 structurally related proteins were identified by Dali from which the top five hits (27%~31% sequence identity with SsfS6) are listed as aligned sequences (Figure 3-9). The best of these was the C-GT UrdGT2 (PDB ID 2P6P, Z = 46.1), while the remaining four were O-GTs. Superposition of SsfS6 and these five comparators confirms common folding features of GT-B family proteins.
Table 3-3. Crystal Parameters, Data Collection, and Refinement of SsfS6

<table>
<thead>
<tr>
<th></th>
<th>SsfS6</th>
<th>SsfS6-TDP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data collection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution range (Å)</td>
<td>50-2.40</td>
<td>50-2.40</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>1.13</td>
<td>1.13</td>
</tr>
<tr>
<td>Space group</td>
<td>P61,22</td>
<td>P61,22</td>
</tr>
<tr>
<td>a, b, c (Å)</td>
<td>82.61, 82.61, 230.52</td>
<td>81.14, 81.14, 229.91</td>
</tr>
<tr>
<td>No. of molecules per</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>asymmetric unit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measured reflections</td>
<td>218,101</td>
<td>288,940</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>18,943</td>
<td>18,077</td>
</tr>
<tr>
<td>$R_{merge}$(^{a})</td>
<td>0.094</td>
<td>0.20</td>
</tr>
<tr>
<td>Completeness</td>
<td>99.04 (96.18)</td>
<td>98.94 (93.87)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>11.5 (11.8)</td>
<td>15.9 (16.0)</td>
</tr>
<tr>
<td>Mean I/sigma(I)</td>
<td>11.24 (1.54)</td>
<td>8.24 (2.41)</td>
</tr>
<tr>
<td><strong>Refinement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R_{cryst}$/R(_{free})(^{c})</td>
<td>0.224/0.262</td>
<td>0.244/0.269</td>
</tr>
<tr>
<td>No. of protein atoms</td>
<td>2,642</td>
<td>2,687</td>
</tr>
<tr>
<td>No. of ligand atoms</td>
<td>-</td>
<td>25</td>
</tr>
<tr>
<td>No. of solvent atoms</td>
<td>110</td>
<td>67</td>
</tr>
<tr>
<td>Average B factor (Å(^2))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>39.7</td>
<td>32.3</td>
</tr>
<tr>
<td>Water</td>
<td>36.5</td>
<td>26.3</td>
</tr>
<tr>
<td>TDP</td>
<td>-</td>
<td>26.7</td>
</tr>
<tr>
<td>RMSD from ideal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond length (Å)</td>
<td>0.004</td>
<td>0.005</td>
</tr>
<tr>
<td>Bond angles (deg)</td>
<td>0.861</td>
<td>0.780</td>
</tr>
<tr>
<td>Ramachandran plot (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Favored regions</td>
<td>96.8</td>
<td>99.1</td>
</tr>
<tr>
<td>Allowed regions</td>
<td>2.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Outliers</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>PDB ID</td>
<td>4FZR</td>
<td>4G2T</td>
</tr>
</tbody>
</table>

Values in parentheses are for the highest resolution shell.

\(^a\) $R_{merge} = \frac{\Sigma_h \sum_i |I(h)| - \langle |I(h)| \rangle}{\Sigma_h \sum_i |I(h)|}$, where $I(h)$ is the intensity of an individual measurement of the reflection and $\langle |I(h)| \rangle$ is the mean intensity of the reflection.

\(^b\) $R_{cryst} = \frac{\sum(h)\sum_i |F_{obs}(h)| - |F_{calc}(h)|}{\sum(h)\sum_i |F_{obs}(h)|}$, where $F_{obs}$ and $F_{calc}$ are the observed and calculated structure-factor amplitudes.

\(^c\) $R_{free}$ was calculated as $R_{cryst}$ using 5.0% of randomly selected unique reflections (in thin resolution shells) that were omitted from the structure refinement.
Figure 3-9. Dali based multiple sequence alignment of SsfS6 with other GTs

Every 20th residue of SsfS6 is labeled and the secondary structure of SsfS6 is displayed on top of the sequence. Conservation of residues by 100%, 80% and 60% are colored red, orange and yellow, respectively. The plots show sequence similarities between different GTs. Residues involved in the SsfS6-TDP binding are marked by asterisk. While the GT-B fold is in common among at these proteins, the active site residues are somewhat varied.
The SsfS6-TDP complex structure reveals TDP bound to the C-terminal domain of SsfS6 with its pyrophosphate pointing toward the active site (Figure 3-8C). Interactions between TDP and SsfS6 are multi-faceted and include interactions with thymine, deoxyribose, and the pyrophosphate of TDP. Thymine interactions include hydrogen bonding between thymine $O_2$ and Leu279 main chain NH, thymine $N_3$, and Phe277 main chain $NH$, and thymine $O_4$ and Val256 main chain $NH$. For the deoxyribose, the $NH_2$ of Asn196 forms a hydrogen bond with the TDP $3'\cdot OH$ group and also likely excludes UDP due to anticipated steric infringement with the $2'\cdot OH$ group. TDP specificity of SsfS6 may also derive from the substitution of T300 for a conserved glutamate within UDP-utilizing enzymes (such as UGT71G1, 2ACW),$^{103}$ which, in the latter, is important for hydrogen bonding interactions with the $2'\cdot OH$ and $3'\cdot OH$ groups of UDP ribose, further stabilized by a conserved glutamine. In SsfS6, Thr300 also coordinates Asn196 to help select for the TDP-sugar and Leu279 substitutes for the above-mentioned glutamine.

In most GT-B enzymes, the NDP-sugar pyrophosphate is stabilized by a neighboring “H-X3-G-T loop”$^{107}$ which corresponds to SsfS6 292HGGHGT297. Important contributors to pyrophosphate hydrogen bonding include the side chain group of His292 and the main chain $NH$ groups of His295, Gly296, and Thr297. A comparison of the apo and TDP-bound forms of SsfS6 reveals a shift of several residues (255AVS257 and 276QFP278) near the active site upon ligand binding. In addition, the loop 220GTRVPL225 became ordered upon ligand binding wherein main chain
NH group of Thr221 engaged hydrogen bonding interactions with the TDP pyrophosphate (Figure 3-8D).

### 3.3.3. Proposed mechanism of SsfS6

Several research groups have proposed that inverting C-glycosylation is achieved by a single nucleophilic displacement at the anomeric carbon, reminiscent of an inverting O-glycosyltransferase mechanism.\textsuperscript{100,108} In C-glycosylation, activation of the aromatic carbon nucleophile is believed to derive from deprotonation of an adjacent hydroxyl (i.e., the C-10 hydroxyl of the SsfS6 D ring). Indirect evidence to support this mechanism includes the ability of the C-GT UrdGT2 to glycosylate surrogate substrates which contain a heteroatom nucleophile at the C-glycosylation site (e.g., analogous to C-9 of the SsfS6 D ring) and recent engineering studies which demonstrated the swapping of glycosidic bond-type specificity between C-GTs and O-GTs.\textsuperscript{108}

Based upon the determined structure of the SsfS6-TDP complex, members of the Thorson and Phillips groups postulate either D58 or E316 may serve as the active-site base and the side chain of H292 may be used to stabilize the β-phosphate during catalysis. While the nature of the actual SsfS6 aglycon remains to be elucidated, and co-crystallization experiments with a range of tetracycline analogs were unsuccessful, molecular docking was employed in an attempt to glean additional information.\textsuperscript{60}
Four docked tetracycline-like aglycon orientations were identified as potential poses in which the protein SsfS6 could productively interact (Figure 3-10). These four poses were given to three independent scoring functions\textsuperscript{109,110} for validation of the docked poses. This analysis was consistent with the putative role of Asp58 or Glu316 as the active site base and also revealed the side chain of Arg222 to occupy the proposed space for olivose and thereby be involved in product release. In summary, while the unknown nature of the SsfS6 aglycon prohibits biochemical study of this intriguing catalyst, the structural studies presented herein are consistent with a putative canonical C-GT mechanism (Figure 3-8E) and highlight key residues that may impact upon catalysis and/or substrate specificity. As such, this work may advance the study of SF2575 biosynthesis and also set the stage for future protein engineering work on SsfS6.
Figure 3-10. Tetracycline docking results by AutoDock 4.2

Glu316 was used as docking box center and seven residues around the cavity were made flexible in order to increase the size of cavity. A, B and C are potential poses for Glu316 as the active base, while D is for Asp58. Docked d-TDP-olivose and tetracycline are represented as sticks. SsfS6 is displayed as both ribbon and surface to show the backbone and cavity, while the potential catalytic bases, Asp58 and Glu316, are highlighted as sticks. The distances between the glycosylated carbon of aglycone and anomerlic carbon of sugar, and the distances between the oxygen of putative base Glu316 (or Asp58) and D ring hydroxyl of aglycone are also shown.
3.4. Crystal structure of OleD, an excellent model GT for studying substrate promiscuity and making novel therapeutics

As described in the previous chapters, natural product pool has served as drug leads and many of which are glycosylated secondary metabolites. So altering glycosylation patterns on clinical natural products has a very promising potential for the generation of novel therapeutics, and this process is called “glycorandomization”. Therefore, glycosyltransferases that have novel substrate activities are very important targets for this purpose. In this chapter, a crystal structure of macrolide glycosyltransferase OleD in the apo form is reported, and protein dynamics investigation of both apo and ligand bound forms are conducted. Preliminary function assays from Thorson lab also show that OleD is a very promising target to be applied in the glycorandomization. The structural studies of this section were done by me and Joey in the Phillips lab and all the biochemistry assays were designed and finished by the Thorson lab.

*Portions of chapter 3.4 have not been published when my thesis is written. The corresponding published PDB entries are 4M60, 4M7P and 4M83.
3.4.1. Crystal structure of apo glycosyltransferase OleD

Glycosyltransferases comprise 94 diverse sequence-based families of enzymes, but they are known to adopt only three different folds (GT-A, GT-B and predicted transmembrane fold GT-C)\textsuperscript{118-123}. The mechanistic, in vitro biochemical and/or structural studies of O-GTs are well studied, so it is great time now to exploit our understanding, and ability to engineer these remarkable catalysts. In this case, O-GT OleD, the X-ray diffraction data were collected to a resolution of 1.7 Å (Table 3-4). OleD is monomeric in crystal packing as calculated by PISA\textsuperscript{60} as well as in gel filtration. OleD adopts previous discussed GT-B fold as shown in Figure 3-11AB, and share the common features of GT-B fold proteins. A ligand-bound structure of OleD was previously solved by Bolam et al\textsuperscript{124}. The kinetic parameters, the UDP, and donor sugar binding sites of OleD have also been nicely characterized in the same paper\textsuperscript{124}. Recently, the Thorson lab and others have demonstrated that the scope of glycosyltransferase OleD-catalyzed sugar installation can be broaden\textsuperscript{125-137}. Preliminary results from Thorson lab also revealed a new mutation of OleD, after using “HT-glycosyl-scanning” screen to test a FDA-approved drug library, might lead to some new glycosides. This brilliant future possible work will be discussed in Chapter 5.
Figure 3-11. OleD structures shown in static screenshot and dynamics forms

Apo structure of OleD (PDB: 4M60) is shown in cartoon (a) and surface (b) mode. (c) Ensemble refinement result of apo OleD is deposited as 4M7P. (d) Ensemble refinement of ligand bound OleD structure (2IYF) is deposited as 4M83.
3.4.2. Ensemble refinements of apo OleD and ligand bound OleD

A single static structure is usually used to represent a protein crystal structure. However, even packed as crystals, protein still has, to some extent, dynamics, and those dynamics could be missed by only showing the structure in a single conformation. Ensemble refinements have been a great approach to extract dynamics information for high-resolution crystal data, and to date, multiple approaches have been developed to accomplish this\textsuperscript{138,139}. Application of ensemble refinement method developed by PHENIX group\textsuperscript{138} to both apo and ligand-bound OleD structures reduces the $R_{\text{cryst}}$ and $R_{\text{free}}$ compared to regular PHENIX refinements (Table 4-4)\textsuperscript{140}. TLS groups were selected based on TLSMD, global harmonic restraints were applied on the whole macromolecular with a weight of $10^{-5}$. The three main optimized empirical parameters for these ensemble refinements are: wxray\_coupled\_tbath\_ offset=2.0, ptls=0.7, and tx=0.8. To avoid any deposit issues with functional molecule occupancies, I defined the output model number to be 20. For those ensemble refinement results, I was able to visualize the aglycon-gating loop, which is not built in or partially invisible in the static X-ray structure and therefore guide future engineering work. To date, the newest mutation identified by the Thorson lab with novel functions has a mutation on the gating loop\textsuperscript{141–148}, which is only visible in the ensemble refinement.
Table 3-4. Statistics of static and dynamic OleD structures

<table>
<thead>
<tr>
<th></th>
<th>OleD-apo static</th>
<th>OleD-apo ensemble</th>
<th>OleD-ligand ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Crystal parameters</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Space group</td>
<td>$R3\text{;}H$</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Unit-cell parameters (Å, °)</td>
<td>124.1, 124.1, 67.6</td>
<td>90.0, 90.0, 120.0</td>
<td>---</td>
</tr>
<tr>
<td><strong>Data collection statistics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.98</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Resolution range (Å)</td>
<td>50-1.77</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>No. of reflections (measured / unique)</td>
<td>441,118 / 37,890</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>100.0 (100.0)</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>$R_{\text{merge}}$</td>
<td>0.082 (0.650)</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Redundancy</td>
<td>11.6 (11.0)</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Mean $I / \sigma(I)$</td>
<td>11.1 (1.45)</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>CC$^{1/2}$</td>
<td>0.99 (0.73)</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Refinement and model statistics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R_{\text{cryst}} / R_{\text{free}}$</td>
<td>0.183 / 0.208</td>
<td>0.150 / 0.187</td>
<td>0.160 / 0.201</td>
</tr>
<tr>
<td>$\Delta R_{\text{cryst}} / \Delta R_{\text{free}}$</td>
<td>N.A.</td>
<td>-0.033 / -0.021</td>
<td>-0.040 / -0.031</td>
</tr>
<tr>
<td>RMSD bonds (Å)</td>
<td>0.008</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>RMSD angles (°)</td>
<td>1.16</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>B factor - protein/solvent (Å$^2$)</td>
<td>36.1 / 42.8</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>No. of protein atoms</td>
<td>2,914</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>No. of waters</td>
<td>245</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>No. of ensemble models</td>
<td>N.A.</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td><strong>Ramachandran plot (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Favorable region</td>
<td>96.0</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Additional allowed region</td>
<td>4.0</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Disallowed region</td>
<td>0.0</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>PDB</strong></td>
<td>4M60</td>
<td>4M7P</td>
<td>4M83</td>
</tr>
</tbody>
</table>

$R_{\text{merge}} = \sum_n \sum_i |I_i(h) - <I(h)>| / \sum_n \sum_i I_i(h)$, where $I_i(h)$ is the intensity of an individual measurement of the reflection and $<I(h)>$ is the mean intensity of the reflection.
References


78. Facompre, M. et al. DNA binding and topoisomerase I poisoning activities of novel


Chapter 4

Alkylrandomization: Methionine Adenosyltransferases and Methyltransferases

Natural products and their derivatives continue to play an important role in drug discovery. Most natural product structures contain methyl group(s) or alkyl group(s). So as exemplified by natural product glycorandomization, altering alkyl patterns on clinical essential natural products is very promising for the generation of novel therapeutics, and this approach is named “alkylrandomization”. In this chapter I will present my research contributions to alkylrandomization in two aspects: expending “Ado-X” alkyl library by structural and engineering studies of methionine adenosyltransferase (sMAT) in sections 4.1; toolbox development of methyl-/alkyl-installation enzyme methyltransferase DnrK in section 4.2.
4.1. Understanding molecular recognition of promiscuity of thermophilic methionine adenosyltransferase from *Sulfolobus solfataricus*

Methionine adenosyltransferase (MAT) is a family of enzymes that utilizes ATP and methionine to produce S-adenosylmethionine (AdoMet), the most crucial methyl donor in the biological methylation of biomolecules and bioactive natural products. Here in this chapter, I will describe the MAT from *Sulfolobus solfataricus* (sMAT), an enzyme from a poorly explored class of the MAT family, has the ability to produce a range of differentially alkylated AdoMet analogs in the presence of non-native methionine analogs and ATP. To investigate the molecular basis for AdoMet analog production, I have crystallized the sMAT in the AdoMet bound, S-adenosylethionine (AdoEth) bound and unbound forms. Notably, among these structures, the AdoEth bound form offers the first MAT structure containing a non-native product, and cumulatively these structures add new structural insight into the MAT family and allow for detailed active site comparison with its homologs in *Escherichia coli* and human. As a thermostable MAT structure from archaea, the structures herein also provide a basis for future engineering to potentially broaden AdoMet analog production as reagents for methyltransferase-catalyzed ‘alkylrandomization’ and/or the study of methylation in the context of biological processes. The structural part was accomplished by me in Phillips lab while the biochemistry part, unless specially mentioned, is done by Thorson lab.

4.1.1. Introduction

S-adenosylmethionine (AdoMet) is the most widely used methyl donor in biological systems, and the only known family of enzymes that synthesize AdoMet is methionine adenosyltransferase (MAT), which utilizes methionine and ATP as substrates (Figure 4-1A)\(^2\)\(^{-11}\). The reaction occurs in an unusual two-step mechanism, in which the adenosyl group is transferred from ATP to the sulfur atom of methionine cleaving the triphosphate and the triphosphate is subsequently hydrolyzed to diphosphate (PPi) and phosphate (Pi)\(^12\). MAT is present in all living organisms from bacteria to mammals. To date, MAT structures from *Escherichia coli*\(^13\),\(^14\)*, *Campylobacter jejuni*\(^15\), *Burkholderia pseudomallei* (PDB code 3IML), *Entamoeba histolytica* (PDB code 3SO4), *Mycobacterium marinum* (PDB code 3RV2), *Mycobacterium avium* (PDB code 3S82), *Mycobacterium tuberculosis* (PDB code 3TDE), *Thermococcus kodakarensis*\(^16\), *Rattus norvegicus*\(^17\) and *Homo sapiens*\(^18\) have been solved. They are more than 50% identical to one other and share several common features: (a) they usually appear as oligomeric proteins with a highly conserved three-domain fold\(^19\); (b) divalent cations such as Mg\(^{2+}\) are required for activity and monovalent cations such as K\(^+\) are able to enhance the reaction rate\(^20\); (c) a flexible loop suspended above the active site serves as a ‘gate’ and is involved in the catalytic activity of the enzyme\(^21\). In addition, some members of the MAT family have recently been demonstrated to produce non-native AdoMet analogs from non-native methionine analogs\(^22\)\(^{-24}\). And in the context of coupled reactions containing permissive methyltransferases, those MATs enable the generation of natural product ‘alkylrandomizations’\(^25\). Thus, there is renewed interest
in assessing the substrate tolerance of additional MATs and extending the molecular level understanding of MAT–substrate interactions.

In contrast to well-characterized MATs in bacteria and eukaryotes, studies of a sequence divergent class of MATs in archaea has been limited. The first archaeal MAT was identified from *Sulfolobus solfataricus* by Porcelli et al.\(^2\) in 1988, having only ~20% sequence identity with *E. coli* and human MATs (Figure 4-2). Later, another archaeal MAT from *Methanococcus jannaschii* (mjMAT) was characterized in detail in terms of kinetics parameters, substrate specificity and folding\(^2\). More recently, an apo crystal structure of MAT from *Thermococcus kodakarensis* has been reported, which provides new structural insights on archaeal MAT\(^1\). The improved thermostability but comparable kinetics parameters with MATs in bacteria and eukaryotes show archaeal MATs to be of great interest for enzyme engineering applications. However, structural information of active site contents and broad substrate specificity assessment for archaeal MATs has been lacking. Here I describe *Sulfolobus solfataricus* MAT (sMAT) to enable the cumulative synthesis of a broad panel of unnatural AdoMet analogs (31 analogs detected) starting from synthetic S/Se-alkylated Met analogs (42 analogs) or commercial sources (three analogs). In addition, this study highlights the crystal structures of a thermostable MAT (sMAT), in three different forms: AdoMet bound, a non-native product S-adenosylethionine (AdoEth) bound, and the unbound form. Interestingly, in contrast to its low sequence similarity to other MATs, sMAT displays the typical three-domain fold and partly
conserved active site architecture. Unlike other known MAT structures, the activity of sMAT cannot be stimulated by ionic potassium\textsuperscript{26}. This can be supported structurally by the presence of a lysine side chain (Lys63) in sMAT, which probably has a similar function to potassium ion in other MATs. Further, the capture of the first atypical ligand bound structure of MAT provides insights on the nature of sMAT broad substrate specificity and a potential template for future engineering toward expanding the substrate scope. Cumulatively, the results in this study provide the first atomic view of the poorly explored class of MATs from archaea and expose sMAT as an efficient catalyst for AdoMet analog production that is amenable to downstream AdoMet-utilizing processes. Combined with the another work from the Thorson lab on other MATs\textsuperscript{25}, a “Ado-X” library is generated for the coupled alkylrandomization reaction with downstream methyltransferases.
Figure 4-1. Natural and unnatural reactions catalyzed by sMAT

(A) The reaction catalyzed by methionine adenosyltransferase. (B) Turnover of S/Se-Met analogs to the corresponding AdoMet analogs catalyzed by sMAT based upon RP-HPLC (average percentage error ≤ 4%). Bars colored green (referring to branched L-alkyl-substituted analogs) denote analogs for which only trace turnover was observed with MATs studied to date. As controls, no product formation was observed in the absence of sMAT, S/Se-Met analogs or ATP.
The structure-based sequence alignment of MAT from *Sulfolobus solfataricus* (sMAT), MAT from *E. coli* (eMAT) and human MAT (hMAT2A). Secondary structural features of sMAT are shown at the bottom.

![Figure 2. The structure-based sequence alignment of MAT from *Sulfolobus solfataricus* (sMAT), MAT from *E. coli* (eMAT) and human MAT (hMAT2A). Secondary structural features of sMAT are shown at the bottom.](image-url)
4.1.2. Overall structural organization

The crystal structures of sMAT have been determined successfully at 2.19 Å or 2.39 Å resolution for the AdoMet-ligated form, 2.49 Å for the AdoEth-ligated form and 2.21 Å for the unliganded form (Table 4-1). Similar to E. coli MAT (eMAT)\textsuperscript{28}, rat liver MAT (rlMAT)\textsuperscript{29} and human MATs (hMAT1A and hMAT2A)\textsuperscript{18}, sMAT packs as a tetramer (Figure 4-3A). All four sMAT structures contain two subunits, A and B, in the asymmetric unit and the tetramer is formed by a two-fold crystallographic symmetry axis. The buried surface interfaces between the two subunits A and B and the two dimers AB and CD are calculated to be 2570 and 1870 Å\textsuperscript{2} respectively\textsuperscript{30}. These areas are similar to those from other bacterial MATs\textsuperscript{13–15}, but larger than rlMAT\textsuperscript{17} and slightly smaller than hMAT\textsuperscript{18}. In the tetramer, there are four potential ligand binding sites: two sites sandwiched by A and B and the other two sites between C and D. Compared with other known MAT structures, the sMAT monomer adopts a similar three-domain architecture with secondary structure variants (Figure 4-3A). Interestingly, unlike other MATs, structural variations were observed between subunits A and B in sMAT with an average rmsd of 0.51 Å in both the liganded and unliganded structures. In addition, the maximum rmsd between all the A subunits and the maximum rmsd between all the B subunits in all sMAT structures are calculated as 0.29 Å and 0.21 Å, respectively. A similar observation was also reported in a recently solved archaeal MAT structure\textsuperscript{16}. As a result, half of the active sites within sMAT have a more open conformation than the others. Consistently, in all the ligand bound sMAT structures, only half of the active sites within the sMAT tetramer
are occupied while the other half are unoccupied. In addition, only the gating loops outside the occupied active site become ordered (Figure 4-3CD).

The substrate specificity for sMAT based upon RP-HPLC is illustrated in Figure 4-1B wherein observed 5’-methyl-thio(seleno)-5’-deoxyadenosine (MSeA) production (via RP-HPLC, Figure 4-4) was interpreted as product based upon the well-established AdoMet decay pathways indicating MSeA to directly derive from AdoMet (not ATP)\textsuperscript{31-33}. The putative substrates tested were those recently reported to profile the substrate specificity of a range of MATs and these analogs were specifically designed to interrogate both steric and electronic contributions to turnover\textsuperscript{25}. Of the 45 putative substrates (See detail compound structures in Chapter 2) tested with sMAT, 11 led to appreciable (> 50%) AdoMet analog production, an additional 15 led to moderate (25%–50%) conversion, while five offered detectable product (< 25%) under the conditions described. In general, smaller alkyl- substitutions were better tolerated, suggesting steric infringement to possibly prohibit larger substitutions. Interestingly, in the case where direct comparisons could be made, the degree of unsaturation correlated with a reduction in turnover (e.g. propyl > allyl > propargyl). Importantly, notable turnover was observed with branched analogs (Figure 4-1B, highlighted in green) that previously led to only trace product with MATs studied to date.
Figure 4-3. The overall molecular structure and active site contents of sMAT

(A) Tetramer assembly of sMAT in the crystal structure as calculated by pisa. (B) Stereoview of sMAT–ligand interactions. Side view of sMAT dimer with AdoEth bound (C) and apo sMAT dimer (D). The gating loop region is highlighted in red.
HPLC traces for representative sMAT reactions illustrating the production of AdoMet analogs and/or 5'-methyl-thio(seleno)-5'-deoxyadenosine (MSeA) in the presence of a select set of l-Met analogs. Starting material (ATP) is designated by a dot.
Table 4-1. Data collection and refinement statistics for sMAT

Values in parentheses are for the highest resolution shell. $R_{\text{free}}$ was calculated as $R_{\text{crys}}$ using 5.0% of randomly selected unique reflections (in thin resolution shells) that were omitted from the structure refinement. RSCC is the real-space correlation to electron density calculated by PHENIX34.

<table>
<thead>
<tr>
<th>Added in crystallization</th>
<th>sMAT</th>
<th>sMAT+AdoMet</th>
<th>sMAT+AdoMet</th>
<th>sMAT+AdoEth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measured reflections</td>
<td>1 478 624</td>
<td>1 541 606</td>
<td>1 854 076</td>
<td>1 101 136</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>74 368</td>
<td>78 996</td>
<td>59 745</td>
<td>53 034</td>
</tr>
<tr>
<td>$R_{\text{merge}}$</td>
<td>0.118 (0.909)</td>
<td>0.142 (0.918)</td>
<td>0.165 (0.606)</td>
<td>0.184 (0.225)</td>
</tr>
<tr>
<td>Completeness</td>
<td>99.61 (96.38)</td>
<td>99.75 (97.50)</td>
<td>100.0 (100.0)</td>
<td>98.31 (82.97)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>19.9 (9.2)</td>
<td>19.5 (14.3)</td>
<td>26.0 (22.2)</td>
<td>20.7 (21.0)</td>
</tr>
<tr>
<td>Mean (I/</td>
<td>$\sigma$</td>
<td>14.55 (1.41)</td>
<td>14.94 (3.08)</td>
<td>27.01 (5.14)</td>
</tr>
<tr>
<td>Refinement</td>
<td>$R_{\text{crys}}/R_{\text{free}}$</td>
<td>0.1990/0.2201</td>
<td>0.1639/0.1888</td>
<td>0.1710/0.1978</td>
</tr>
<tr>
<td>No. of protein atoms</td>
<td>6301</td>
<td>6396</td>
<td>6396</td>
<td>6410</td>
</tr>
<tr>
<td>No. of ligand atoms</td>
<td>-</td>
<td>48</td>
<td>43</td>
<td>44</td>
</tr>
<tr>
<td>No. of solvent atoms</td>
<td>304</td>
<td>707</td>
<td>596</td>
<td>589</td>
</tr>
<tr>
<td>Average B factor (Å²)</td>
<td>38.20</td>
<td>31.6</td>
<td>37.5</td>
<td>40.1</td>
</tr>
<tr>
<td>Protein</td>
<td>-</td>
<td>28.7</td>
<td>30.3</td>
<td>38.2</td>
</tr>
<tr>
<td>Ligands RSCC</td>
<td>0.39</td>
<td>0.98</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>Bond length (Å)</td>
<td>0.002</td>
<td>0.007</td>
<td>0.006</td>
<td>0.007</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>0.69</td>
<td>1.33</td>
<td>0.70</td>
<td>0.97</td>
</tr>
<tr>
<td>Ramachandran plot (%)</td>
<td>99</td>
<td>98</td>
<td>98</td>
<td>99</td>
</tr>
<tr>
<td>Favored regions</td>
<td>0.25</td>
<td>0.12</td>
<td>0.12</td>
<td>0.0</td>
</tr>
<tr>
<td>Outliers</td>
<td>4HPV</td>
<td>4L71</td>
<td>4K0B</td>
<td>4L2Z</td>
</tr>
</tbody>
</table>
Table 4-2 highlights a comparison of the kinetic parameters for L-methionine and the non-native substrate for which a ligand bound structure is available (L-ethionine). The changes in kinetic parameters of the sMAT for both the substrates are moderate from 37 °C to 65 °C. Compared with *Thermococcus kodakarensis* MAT and mjMAT, sMAT appears to be a somewhat better enzyme because sMAT has a 100 times smaller Km for methionine, a slightly smaller Km for ATP, and a similar $k_{cat}$. More importantly, the data reflect that L-ethionine is kinetically competent and comparable to the native substrate L-methionine. At either temperature, the $k_{cat}$ values for the sMAT reaction with L-methionine or L-ethionine are similar and the reduced proficiency with L-ethionine compared with the native substrate L-methionine derives from a combination of higher $K_m$ values for both L-ethionine and ATP.

**4.1.3. Active site contents**

The MAT catalyzed AdoMet formation, occurs via a sequential two-step mechanism. In the first step, AdoMet is formed by a direct $S_N2$ reaction, in which the sulfur atom of methionine attacks the C5’ position of ATP and thus cleaves the polyphosphate chain from ATP. In the second step, the triphosphate is further hydrolyzed to PPi and Pi$^{35-61}$. Komoto et al.$^{14}$ identified two critical residues, Lys165 and His14, in eMAT for this proposed $S_N2$ reaction based on their ligand bound structures. Interestingly, even with significant sequential variations to eMAT and other MATs, several conserved residues were observed in sMAT, mainly located around the active site, including the two crucial residues Lys201 and His29 for the proposed $S_N2$ reaction (Table 4-3).
Table 4-2. Kinetic parameters of sMAT enzyme

<table>
<thead>
<tr>
<th>Temperature (°C)</th>
<th>Constant substrate</th>
<th>Varied substrate</th>
<th>$k_{cat}$ (min$^{-1}$)</th>
<th>$K_m$ (mM)</th>
<th>$k_{cat}/K_m$ (mM$^{-1}$·min$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>ATP</td>
<td>L-methionine</td>
<td>2.06 ± 0.03</td>
<td>0.0023 ± 0.0003</td>
<td>895 ± 100</td>
</tr>
<tr>
<td></td>
<td>ATP</td>
<td>L-ethionine</td>
<td>2.34 ± 0.07</td>
<td>0.0066 ± 0.0015</td>
<td>417 ± 47</td>
</tr>
<tr>
<td></td>
<td>L-methionine</td>
<td>ATP</td>
<td>1.97 ± 0.08</td>
<td>0.0686 ± 0.0110</td>
<td>29 ± 7</td>
</tr>
<tr>
<td></td>
<td>L-ethionine</td>
<td>ATP</td>
<td>2.48 ± 0.09</td>
<td>0.0969 ± 0.0130</td>
<td>26 ± 7</td>
</tr>
<tr>
<td>65</td>
<td>ATP</td>
<td>L-methionine</td>
<td>2.92 ± 0.07</td>
<td>0.0028 ± 0.0007</td>
<td>1042 ± 100</td>
</tr>
<tr>
<td></td>
<td>ATP</td>
<td>L-ethionine</td>
<td>2.99 ± 0.11</td>
<td>0.0070 ± 0.0016</td>
<td>427 ± 69</td>
</tr>
<tr>
<td></td>
<td>L-methionine</td>
<td>ATP</td>
<td>2.81 ± 0.08</td>
<td>0.0145 ± 0.0025</td>
<td>194 ± 32</td>
</tr>
<tr>
<td></td>
<td>L-ethionine</td>
<td>ATP</td>
<td>3.10 ± 0.14</td>
<td>0.0692 ± 0.0108</td>
<td>52 ± 13</td>
</tr>
</tbody>
</table>
Table 4-3. Ligand-based alignment of active site residues (sMAT and eMAT)

Residues for proposed Sn2 reaction are highlighted in bold. The atom numbers used for interaction analysis here are the numbers from the AdoEth bound structure (PDB code 4L2Z).

<table>
<thead>
<tr>
<th>Interaction partner</th>
<th>sMAT</th>
<th>eMAT</th>
<th>DALI alignment</th>
<th>Chain</th>
</tr>
</thead>
<tbody>
<tr>
<td>O₁, O₂ in PPi, O₂ in PO₄</td>
<td>K25</td>
<td>K245</td>
<td>Not detected</td>
<td>B</td>
</tr>
<tr>
<td>O₃ in PPi</td>
<td>H29</td>
<td>H14</td>
<td>Detected</td>
<td>B</td>
</tr>
<tr>
<td>Mg²⁺</td>
<td>D31</td>
<td>D16</td>
<td>Detected</td>
<td>B</td>
</tr>
<tr>
<td>O₇, O₈ in AdoEth</td>
<td>H58</td>
<td>Q98</td>
<td>Not detected</td>
<td>A</td>
</tr>
<tr>
<td>O₇, O₈ in AdoEth</td>
<td>N60</td>
<td>K269</td>
<td>Not detected</td>
<td>A</td>
</tr>
<tr>
<td>K63 (K⁺ in eMAT)</td>
<td>D62</td>
<td>E42</td>
<td>Detected</td>
<td>A</td>
</tr>
<tr>
<td>O₅, O₇ in PPi</td>
<td>K63</td>
<td>K⁺</td>
<td>–</td>
<td>A</td>
</tr>
<tr>
<td>Ethyl group in AdoEth</td>
<td>L145</td>
<td>I102</td>
<td>Detected</td>
<td>A</td>
</tr>
<tr>
<td>O₂₆ in AdoEth</td>
<td>D199</td>
<td>D163</td>
<td>Detected</td>
<td>B</td>
</tr>
<tr>
<td>O₃ in PPi</td>
<td>K201</td>
<td>K165</td>
<td>Detected</td>
<td>B</td>
</tr>
<tr>
<td>Stacked with adenine ring</td>
<td>Y270</td>
<td>F230</td>
<td>Detected</td>
<td>B</td>
</tr>
<tr>
<td>O₂₇ in AdoEth</td>
<td>D282</td>
<td>D238</td>
<td>Detected</td>
<td>B</td>
</tr>
<tr>
<td>O₁, O₃ in PO₄</td>
<td>R288</td>
<td>R244</td>
<td>Detected</td>
<td>B</td>
</tr>
<tr>
<td>Mg²⁺</td>
<td>E305</td>
<td>D271</td>
<td>Not detected</td>
<td>A</td>
</tr>
<tr>
<td>O₆ in PPi, O₁ in PO₄</td>
<td>K310</td>
<td>K265</td>
<td>Detected</td>
<td>A</td>
</tr>
<tr>
<td>O₅ in PPi</td>
<td>H315</td>
<td>–</td>
<td>–</td>
<td>A</td>
</tr>
<tr>
<td>Ethyl group in AdoEth</td>
<td>I349</td>
<td>I302</td>
<td>Detected</td>
<td>A</td>
</tr>
</tbody>
</table>
The interactions between sMAT and products in the active site are multifaceted as illustrated in Figure 4-3B. The adenine ring of AdoMet is recognized by a hydrogen bond with the side chain of aspartate 144 and a stacking interaction with the aromatic ring of tyrosine 270. Several water molecules surrounding the adenine ring also form a hydrogen bond network to the enzyme. The 2′-OH and 3′-OH of the adenosine ribose interact with the side chains of aspartate 199, aspartate 282 and serine 277. Similar interactions involving aspartic acids have been seen in other MAT structures bound with AdoMet, but not for serine\textsuperscript{14,18}. The methionine/ethionine moiety (of AdoMet and AdoEth, respectively) forms hydrogen bonds with four residues, in which the amino group interacts with the side chain of aspartate 282 and the carboxylate group interacts with the side chain of histidine 58, asparagine 60 and asparagine 159. The methyl or ethyl group is buried in a slightly hydrophobic pocket surrounded by asparagine 159, aspartate 160, isoleucine 349, leucine 145 and the adenine ring. Like similar observations in eMAT\textsuperscript{14}, the PPI and Pi form a U-shaped conformation with two magnesium ions closely stacked on both sides. Further, the two magnesium sites are formed with the side chains of aspartate 31, glutamate 305 and three water molecules. The phosphate groups are surrounded and stabilized by the side chains of several basic amino acids: lysine 25, histidine 29, lysine 201, arginine 288 and lysine 310.

As described, the ligands are solvent inaccessible and thus the entrance of the active site requires a dynamic and flexible region. A flexible loop region was previously
identified as the gate for the active site in MATs. This gate loop feature has been confirmed by crystal structures of eMAT and hMAT2A, in which the loop becomes ordered when ligands are bound and it becomes disordered when the active site is empty. A similar gating loop (residues 141–155) region was identified in sMAT (Figure 4-3), which has a similar pose to eMAT and hMAT2A and interacts with the adenine ring and the methionine or ethionine moiety. When the active site is empty, residues 141–144 form a β-sheet with residues 95–96 while the rest of the loop is poorly seen or unidentifiable in electron density. When ligands are bound in the active site, the loop region becomes ordered and is anchored above the active site with residues 141–144 shifting from the origin position and with residues 145–149 forming a small α-helix.

In various structures of eMAT and hMAT2A, the active site ligand occupation is correlated with an ordered active site gating loop. These studies are consistent with our observations in sMAT structures. In addition to this, ligand orientations in the eMAT and human MAT studies are similar to sMAT’s. Different ligand orientations with an unordered gating loop have been reported before in rlMAT structures and an earlier set of eMAT structures, but have some unusual features: (a) in rlMAT the temperature factors of most of the ligand coordinates are above 70 Å², and in earlier eMAT structures the temperature factors of ADP (or ADP mimic) are above 114 Å²; (b) the flexible loop above the active site was undefined even though the active sites were reported to be occupied; (c) even though rlMAT and eMAT have a 59%
sequence identity and all critical residues are conserved, they show completely
different ligand orientations in the active site; (d) X-ray data for rlMAT and earlier
eMAT structures are not available online, and thus ligand real-space correlation
coefficients cannot be calculated. These unusual features suggest that the active site
contents of rlMAT and earlier eMAT are questionable. In contrast to rlMAT and the
earlier eMAT structures, all the hMAT2A, later eMAT and sMAT structures have
reasonable temperature factors and a good ligand real-space correlation to electron
density at the active site.

4.1.4. Unusual product formation during crystallization

As described in 'Chapter 2', two sets of ligand bound crystals were obtained in the
presence of 5 mM ADP, 10 mM ethionine (or methionine), 10 mm Mg$^{2+}$ ion and 1.4 M
NaHPO$_4$/K$_2$HPO$_4$. Thus based on the simulated annealing $F_o - F_c$ omit map of the
active site, one ADP, one PO$_4$, two Mg$^{2+}$ and one ethionine (or methionine) molecule
were initially built in (Figure 4-5, left). However, this model does not fit the electron
density perfectly, because the $F_o - F_c$ omit map does not agree with the placement of
the crucial carbon atom circled in Figure 4-5. Thus, it is very clear that the product
has already formed and a model including P Pi and AdoEth (or AdoMet) is more
appropriate. The new model (Figure 4-5, right) has a lower temperature factor and a
better real-space correlation to electron density in the active site. MAT-catalyzed
AdoMet/AdoEth formation via ADP and Met/Eth has not been previously observed.
In addition, incubation of sMAT in the presence of ADP and methionine at 65 °C for 90 min under standard assay conditions led to < 2% AdoMet formation. Thus, two
explanations for this unusual product formation have been proposed: (a) the ADP stock solution is contaminated by a sufficient amount of ATP; (b) the unusual reaction catalyzed by sMAT can actually occur in vitro, but may take as long as 1 month to complete, which corresponds to the time of crystal growth in this experiment.

4.1.5. Structure homology

A DALI search\textsuperscript{64} for structures similar to the sMAT monomer returned several hits, all of which are previously solved MAT structures with Z-scores between 23 and 29. These MAT structures share a very high level of overall sequence identity (> 50%) and a high level of conservation among residues associated with substrate binding. Interestingly, sMAT only has a maximum sequence identity of 19% with these known MATs, but shares a similar three-domain (Figure 4-6A).

For the comparison of active site residues, crystal structures of sMAT, eMAT and hMAT2A were aligned by ligands as described in Table 4-3. Surprisingly, 16 of 17 active site residues detected in sMAT have an identical or similar residue in eMAT and hMAT2A. The only extra residue sMAT has is His315, which forms a hydrogen bond with O5 in diphosphate. DALI-based sequence alignment was able to identify 11 pairs of residues (Figure 4-2). Eight of them are conserved among sMAT and other MATs, including the crucial residues histidine and lysine for the proposed S\textsubscript{N}2-like mechanism\textsuperscript{14,18}. The other three pairs are very similar residues at the same spot: for
example, in sMAT tyrosine 270 forms stacking interactions with the adenine ring of AdoEth/AdoMet, while in eMAT it is phenylalanine 230. Intriguingly, there are another five pairs of residues that are not detectable via DALI search: the side chain of sMAT lysine 25 (eMAT lysine 245) helps stabilize the triphosphate group; the side chains of sMAT histidine 58 and asparagine 60 (eMAT glutamine 98 and lysine 269) form hydrogen bonds with the carboxyl group of methionine or ethionine; the side chain of sMAT glutamate 305 (eMAT aspartate 271) forms ionic bonds with the magnesium ion; the side chain of sMAT lysine 63 occupies the same spot as the eMAT potassium ion and helps stabilize the diphosphate ligand (Figure 4-6B).

Unlike other known MATs, it has been previously reported that the activity of sMAT cannot be enhanced by K\textsuperscript+\textsuperscript{26}. In the present study, all the crystals of sMAT were obtained from the crystallization condition containing more than 150 mM potassium, but electron density suitable for K\textsuperscript+ was not observed in any data sets. In addition, potassium dependence has been previously reported in a close MAT homolog from *Methanococcus jannaschii*, which shares the same active site residues with sMAT except for the lysine\textsuperscript{24}. Combined with the active sites alignment evidence above (Figure 4-6B), it is very likely that the catalytic activity of sMAT is not affected by K\textsuperscript+, because the lysine in sMAT serves to present the requisite cation properties.
Figure 4-6. Fold and active site comparisons between sMAT, hMAT2A and eMAT

(A) Monomers of sMAT, hMAT2A and eMAT show a similar three-domain fold. (B) Substrate-based alignment of sMAT and eMAT shows a space overlap between lysine in sMAT and K⁺ in eMAT. (C) Substrate-based alignment of sMAT, hMAT2A and eMAT displays the clashes around the sMAT ethyl group.
Interestingly, eMAT and hMAT2A also have some ability to incorporate ethionine. The ethionine turns over with sMAT and hMAT2A is near 100% whereas with eMAT is just 10%\(^\text{25}\). A ligand-based alignment (Figure 4-6C) shows that sMAT has a larger cavity around the ethyl/methyl group than either hMAT2A or eMAT. Placement of the ethyl group in eMAT will cause serious clashes with isoleucine 102 and 302, while in hMAT2A the ethionine causes moderate clashes with isoleucine 139 and 344. In sMAT a leucine (L145) is substituted for one of the conserved isoleucines in other MATs (isoleucine 102 in eMAT), which provides more active site flexibility for ethyl group binding. The ethyl group in sMAT only has minor clashes with leucine 145 and isoleucine 349. Therefore, it is very likely that the better proficiency of sMAT is caused by a larger cavity adjacent to the methyl/ethyl group. Also, branched analogs highlighted in Figure 4-1B turn over significantly better with sMAT, comparing with eMAT, hMAT2A and mjMAT\(^\text{25}\). Interestingly, mjMAT is a thermophilic archaeal MAT that has all active site residues conserved with sMAT. However, further comparison between their active site cavities cannot be conducted because mjMAT structure remains unknown. The current structural information suggests that the better turnover rate of branched AdoMet analogs with sMAT is possibly mediated by some general orientation/dynamics of the gating loop and/or secondary shell variations. The specific residues contributing to this are currently unknown.
4.1.6. Conclusions of wt-sMAT structures

The crystal structures herein provide the atomic view of a clearly divergent class of MATs from archaea and add new active site architecture understanding to the MAT family. The sMAT has the characteristic fold and the typical tetramer assembly of known MATs, but it is the first structure in the Protein Data Bank for a thermostable MAT. In addition, the slightly expanded substrate scope of sMAT over other MATs studied to date highlights sMAT as a useful tool for the production of AdoMet analogs. In conjunction with the recent demonstration of coupled MAT-methyltransferase systems for differential alkylation, this chemoenzymatic strategy circumvents a major liability in the use of synthetic AdoMet analogs, namely, the dramatic instability of the AdoMet analogs. Thus, the structural insights regarding sMAT provided herein offer another blueprint from which to pursue future engineering to further broaden the catalyst promiscuity toward AdoMet analog production.

Further, the elucidation of the active site architecture with the atypical product bound, the characterization of the gating loop region, as well as the sMAT turnover reactions with different AdoMet analogs provides a blueprint for future AdoMet analog production. Since the two-step AdoMet synthesis process catalyzed by sMAT does not involve large conformational changes of ATP and methionine/ethionine, the cavity around the methyl/ethyl group is most probably the limiting factor for installing more function groups on the deoxyribose sugar. Since the cavity is formed by the adenine ring, asparagine 159, aspartate 160, isoleucine 349 and leucine 145 in
the gating loop, mutations of these amino acids, especially on isoleucine 349 and leucine 145, or/and mutations on the gating loop residues can potentially increase the size of the cavity or change the local electrostatic field to accept more functional groups, such as the compounds with low or zero turnover rate described in Chapter 2.

4.1.7. Preliminary engineering of sMAT to expand Ado-X library

Despite its divergent sequence, sMAT displays a three-domain fold and tetrameric oligomerization, characteristic of other known MATs. However, the cavity of sMAT active site is larger than eMAT, displaying fewer steric clashes with the ethyl group of AdoEth. This may explain its intrinsic ability to use ethionine as an alternative substrate (Figure 4-7A). At 65°C, sMAT displays substrate promiscuity towards a variety of methionine analogues. However, other enzymes used in synthetic schemes, including methyltransferases, are degraded at this temperature. Additionally, while wildtype sMAT does possess limited promiscuity towards methionine analogues, it still is unable to catalyze reactions with many potentially useful substrates. Therefore, modifying the active site of sMAT to further broaden substrate acceptance and improve catalytic activity at lower temperatures can accelerate the production of diverse Ado-X analogues.

---

*This part of work has not been published, and cannot be accomplished without the help of Elileen Brady and Jonathan Clinger*
Members of the Phillips groups first identified a mutation in N159G that significantly improves catalytic activity of sMAT towards three methionine analogues with extended alkyl chains. The crystal structure of sMAT-N159G has been determined at a resolution of 2.80 Å. The asymmetric unit included 12 molecules consisting of 3 tetramers. The tetramer packing of sMAT-N159G is similar to the previously characterized WT sMAT, composed of two homo-dimers related by two-fold symmetry (Figure 4-7B). $R$ and $R_{free}$ for the final model were refined to be 22.1% and 25.6% respectively. Data collection and refinement statistics are detailed in Table 4-4.

The previously identified class of MATs from archaea, including sMAT$^{14}$ and tMAT$^{16}$, share divergence at key residues compared to E. coli$^{14}$ and human MAT$^{18}$. Leucine 145 in sMAT is conserved in tMAT, but replaced by isoleucine in eMAT and hMAT2A. Similarly, asparagine 159 is conserved in sMAT and tMAT, but substituted for glycine in eMAT and hMAT2A. The presence of glycine at this position in eMAT and hMAT2A suggests that substituting glycine for asparagine in sMAT will preserve the integrity of the active site while allowing for greater room in the pocket, and correspondingly improved ability to accept larger substrates. The crystal structure of N159G confirms that the single amino acid substitution N159G did not disrupt the overall active site architecture (Figure 4-7C). Similar to the previously reported wildtype sMAT structure, the four active sites in the tetramer are enclosed in the hydrophobic surface formed at the interface of the dimers. As is observed in the WT structure, half of the active sites assume a more open conformation, with the other two confined in a
partially closed arrangement. The correspondence of these data, which were obtained from crystals grown in different conditions, suggests that the alternate active site conformations of sMAT are not simply an artifact of lattice packing.

When tested against an abbreviated library of methionine analogues sent from Thorson lab, sMAT-N159G had increased ability to utilize extended substrates under phosphate release assay\(^6^5\) (Figure 4-7D). After 1-hour incubation at 37°C, WT-sMAT displayed negligible <2% conversion of 3 methionine analogues, while sMAT-N159G showed 20%~30% conversion of the same 3 analogues. At the same time, sMAT-N159G showed slightly lower activity with either methionine or ethionine, which is likely to be a side effect of increasing space in the active site cavity. The correspondence of structural analysis and functional characterization of sMAT N159G mutant confirms that expanding the active site cavity of sMAT is a viable strategy for broadening acceptance of extended substrates. The introduction of a single amino acid substitution in sMAT preserved active site organization while increasing promiscuity. While sMAT N159G was able to use substrates with expanded alkyl chains, it also maintained activity towards methionine and ethionine.
Figure 4-7. Structure and preliminary assays of sMAT-N159G

(a) Sequence alignments of MATs from different organisms showing relatively conserved active site architecture. (b) Tetrameric packing of sMAT-N159G. (c) Comparison of active site residues between wt-sMAT and sMAT-N159G. Preliminary phosphate release assay result of sMAT-N159G (d) and sMAT-L145I-N159G-I349F (e).
Table 4-4. Crystallography statistics of sMAT-N159G

<table>
<thead>
<tr>
<th>Data collection</th>
<th>sMAT-N159G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution range (Å)</td>
<td>30-2.80</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.98</td>
</tr>
<tr>
<td>Space group</td>
<td>P 2₁</td>
</tr>
<tr>
<td>Unit cell</td>
<td>74.7, 289.2, 174.7, 90, 99.85, 90</td>
</tr>
<tr>
<td>No. of molecules per asymmetric unit</td>
<td>12</td>
</tr>
<tr>
<td>Measured reflections</td>
<td>527,284</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>134,765</td>
</tr>
<tr>
<td>( R_{merge} )</td>
<td>0.144 (0.879)</td>
</tr>
<tr>
<td>Completeness</td>
<td>96.9 (95.5)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>3.9 (2.2)</td>
</tr>
<tr>
<td>Mean I/sigma</td>
<td>4.68 (1.33)</td>
</tr>
<tr>
<td>CC (1/2)</td>
<td>30.6</td>
</tr>
<tr>
<td><strong>Refinement</strong></td>
<td></td>
</tr>
<tr>
<td>( R_{cryst}/R_{free} )</td>
<td>0.221/0.256</td>
</tr>
<tr>
<td>No. of protein atoms</td>
<td>36,534</td>
</tr>
<tr>
<td>No. of solvent atoms</td>
<td>942</td>
</tr>
<tr>
<td>Average B factor (Å²)</td>
<td>44.3</td>
</tr>
<tr>
<td>RMSD from ideal</td>
<td></td>
</tr>
<tr>
<td>Bond length (Å)</td>
<td>0.016</td>
</tr>
<tr>
<td>Bond angles (deg)</td>
<td>1.62</td>
</tr>
<tr>
<td>Ramachandran plot (%)</td>
<td></td>
</tr>
<tr>
<td>Favored regions</td>
<td>96</td>
</tr>
<tr>
<td>Outliers</td>
<td>0.4</td>
</tr>
<tr>
<td>PDB ID</td>
<td>4WS9</td>
</tr>
</tbody>
</table>

Values in parentheses are for the highest resolution shell.
After finding N159G, Eileen and I continued to screen interesting mutations of sMAT by randomized PCR and batch purification of small-scale sMAT mutant proteins in the Phillips lab. And Jonathan, Eileen and I found a triple mutation of sMAT that displays phenomenal activities towards a few methionine analogues (Figure 4-8E) under phosphate release assay. This protein will be followed up by Jonathan Clinger in the Phillips lab and sent to the Thorson lab for HPLC and HRMS validation. So, the development of a promiscuous sMAT introduces a powerful tool for the production of diverse Ado-X analogues, even at a low temperature. This strategy represents a method for alkylrandomization of natural product libraries, when coupled with a permissive methyltransferase like RebM or DnrK.
4.2. Structural studies of the carminomycin O-Methyltransferase

DnrK and exploiting its ability as methyl/alkyl installation tool

S-adenosyl-L-methionine (AdoMet) is an essential enzyme co-substrate in fundamental biology with an expanding range of biocatalytic and therapeutic applications. As described in chapter 4.1, AdoMet has been expanded into an Ado-X library by structural studies of methionine adenosyltransferase sMAT from *Sulfolobus solfataricus*. In order to accomplish alkylrandomization, the next part is the find methyltransferases that can utilize this Ado-X library. The structural and biochemical studies of carminomycin O-methyltransferase DnrK will be discussed in this chapter. Thorson and Phillips lab chose DnrK as target for two reasons, (1) its ability to use a stable, functional AdoMet isosteres that are resistant to the major physiological AdoMet degradative pathways, and (2) its substrate promiscuity to use multiple substrates as methyl acceptors including several anti-tumor natural products. As usual, the structural part was accomplished by me in Phillips lab while the biochemistry part, unless specially mentioned, is done by Thorson lab.

---

*Portions of chapter 4.2 have not been published by the time my thesis is written*
4.2.1. Functional AdoMet isosteres resistant to AdoMet degradation pathways

Methyltransferase (MT)-catalyzed S-adenosyl-L-methionine (AdoMet)-dependent methylation is essential to all walks of life where it contributes to modulating the function of a vast array of biomolecules ranging from small metabolites to macromolecules\textsuperscript{66–70}. Consistent with this, alterations in methylation-dependent processes are associated with cancer, neuropsychiatric disorders, inflammation, metabolic disorders, fundamental development medicine, susceptibility to disease drug response and drug resistance\textsuperscript{68,70}. While there have been great advances in methylation-dependent bioinformatics and disease-associated biomarkers, the study of intracellular MT spatial/temporal resolution, specificity and function remains a challenge. Toward this end, the pioneering demonstration of AdoMet analogs, bearing alternative alkyl donor substituents, as co-substrates for DNA\textsuperscript{67} or natural product MTs have inspired new tools and strategies to study natural product\textsuperscript{25,71–73}, protein\textsuperscript{23,31}, and nucleic acid\textsuperscript{74,75} methylation where the recent advent of enzyme-based strategies for the synthesis of differentially alkylated AdoMet analogs has simplified access to these unique cosubstrates\textsuperscript{23,31,74,75}. However, the stability of AdoMet or corresponding differentially alkylated analogs under physiological conditions limits their utility as reagents or therapeutics by virtue of three fundamental degradative processes: i) intramolecular cyclization to homoserine lactone and 5’-deoxy-5’-(alkylthio)adenosine (Figure 4-8, pathway a); ii) depurination (Figure 4-8, pathway b); and iii) sulfonium racemization\textsuperscript{32,33,74}. To address this deficiency, herein I describe that the chemoenzymatic synthesis of AdoMet isosteres designed to circumvent the major AdoMet degradative pathways
and demonstrate, via both biochemical and structural studies, these analogs to serve as productive MT co-substrates for the prototypical class I MT DnrK. This work highlights a new conceptual strategy to modulate the fundamental properties of AdoMet and corresponding analogs as reagents in epigenetic, proteomic, biocatalytic and therapeutic applications and also demonstrates a novel high throughput colorimetric assay for DnrK based upon the non-native surrogate substrate 2-chloro-4-nitrophenol (ClNP).

Adenine isosteres, such as 7-deazaadenine (7dzA), have been successfully employed to reduce the propensity for nucleotide depurination (Figure 4-8A, pathway b). Thus, Thorson lab focused their attention on addressing the AdoMet degradative intramolecular cyclization to homoserine lactone and 5’-deoxy-5’-(alkylthio) adenosine (Figure 4-8, pathway a) by replacing the AdoMet carboxylate with a sterically constrained and/or less nucleophilic isosteric functional group. As potential carboxylate isosteres, tetrazolates display a similar pKa, are less nucleophilic and offer improved membrane permeability76. In addition, MM2 calculations predict the formation of a sterically strained tetrazole-containing bicyclic ring structure via intramolecular cyclization (Figure 4-8, pathway a) as energetically unfavorable. To test this concept, the tetrazole-L-methionine isostere (L-1Met) was synthesized from commercially available N-Fmoc-L-methionine in four simple steps (Figure 4-9, 41% overall yield) via a slight modification of previously reported methods and subsequently evaluated as a substrate for the permissive methionine adenosyltransferase hMAT2A25. Importantly, this study revealed the $k_{cat}$ of hMAT2A-catalyzed reactions using native substrate L-methionine (L-Met) or isosteric L-1Met as
nearly identical with an 18-fold observed reduction in $K_m$ and enabled ~50% conversion to desired Ado$^t$Met in 20 hours in a standard non-optimized small scale assay. Using these parameters as a guide, a subsequent large scale chemoenzymatic reaction followed by preparative HPLC provided 19.9 mg of Ado$^t$Met, the purity and composition of which was confirmed via HPLC, NMR and HRMS.

The subsequent comparison of AdoMet and Ado$^t$Met revealed Ado$^t$Met as > 5-fold improved in overall stability. Consistent with our hypothesis, similar levels of AdoMet and Ado$^t$Met depurination (Figure 4-8, pathway b) were observed while Ado$^t$Met enabled complete inhibition of the observed AdoMet degradative intramolecular cyclization reaction (Figure 4-8, pathway a).
Figure 4-8. Utilization and degradation of AdoMet

General scheme highlights methyltransferase-catalyzed use of AdoMet in methylation reactions (upper) and major chemical degradation pathways of AdoMet (lower). AdoMet is S-adenosyl-L-methionine (also referred to as SAM); AdoHcy is S-adenosyl-L-homocysteine (also referred to as SAH).
Figure 4-9. Synthesis of AdoMet analogs and corresponding relevant degradation pathways by Thorson lab

(A) Synthesis of tMet. (B) General scheme for hMAT2A-catalyzed synthesis of AdoMet, Ado\textsuperscript{1}Met, 7\textsuperscript{dz}AdoMet and 7\textsuperscript{dz}Ado\textsuperscript{1}Met. (C) Degradative pathways for Ado\textsuperscript{1}Met, 7\textsuperscript{dz}AdoMet and 7\textsuperscript{dz}Ado\textsuperscript{1}Met.
4.2.2. Complex structure of DnrK with unnatural Ado^Hcy and CINP

To assess the utility of the stabilized Ado^Met as a co-substrate for MTs, I selected the prototypical class I MT DnrK as a model. DnrK is the carminomycin 4-O-MT that catalyzes a culminating step in the biosynthesis of the anticancer agent daunorubicin in *Streptomyces peucetius* and has also been reported to methylate various modified anthracyclines and flavonoids as surrogate substrates. Capitalizing on the permissive nature of DnrK, Thorson lab developed the first real-time colorimetric assay for DnrK-catalyzed alkylation using CINP as a surrogate acceptor substrate where alkylation abolishes the classical nitrophenolate color (A410) observed under basic conditions (Figure 4-10A). While the kinetic parameters for the native DnrK substrate carminomycin have not been reported, the determined CINP kinetic parameters are consistent with CINP as a sufficient non-native surrogate where clear dose- dependency was observed in real-time assays (Figure 4-10B). A comparison of the DnrK kinetic parameters for AdoMet (k\textsubscript{cat} = 0.017 ± 0.001 min\textsuperscript{-1}, K\textsubscript{m} = 138 ± 7 μM) and AdotMet (k\textsubscript{cat} = 0.031 ± 0.001 min\textsuperscript{-1}, K\textsubscript{m} = 335 ± 40 μM) revealed nearly equivalent specificity constants (k\textsubscript{cat}/K\textsubscript{m}) (Table 4-5). As one of the first high throughput colorimetric assay for prototypical class I model MTs, this convenient assay presents an enabling platform for future class I MT engineering, biochemical studies and/or tool development.
Figure 4-10. DnrK colorimetric assay and ligand-bound structures

(A) General scheme for a high-throughput DnrK colorimetric assay enabled by the surrogate acceptor 2-chloro-4- nitrophenol (CINP). (B) Representative assay results demonstrating DnrK-catalyzed alkylation of CINP diminishes classical CINP color (A410) over time. (C) DnrK-Ado'Hcy (green)/AdoHcy (grey) ligand-bound structures. (D) DnrK-CINP (purple)/carminomycin (green) ligand-bound structures with ligands represented as stick models.
### Table 4-5. Kinetic parameters and stability measurements

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Substrate</th>
<th>$k_{cat}$ min$^{-1}$</th>
<th>$K_m$ µmol L$^{-1}$</th>
<th>$k_{cat}/K_m$ Relative</th>
<th>$t_{1/2}$ min$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hMAT2A</td>
<td>L-Met$^b$</td>
<td>8.0 ± 0.1</td>
<td>397 ± 29</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>L'-Met$^b$</td>
<td>6.1 ± 0.6</td>
<td>7.212 ± 1.250</td>
<td>0.04</td>
<td>-</td>
</tr>
<tr>
<td>DnrK</td>
<td>ClNP$^d$</td>
<td>0.019 ± 0.001</td>
<td>106 ± 10</td>
<td>N.A.$^e$</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>AdoMet$^f$</td>
<td>0.017 ± 0.001</td>
<td>138 ± 7</td>
<td>1</td>
<td>942 ± 7</td>
</tr>
<tr>
<td></td>
<td>Ado'Met$^f$</td>
<td>0.031 ± 0.001</td>
<td>335 ± 40</td>
<td>0.8</td>
<td>5,007 ± 6</td>
</tr>
<tr>
<td></td>
<td>$^{7\text{dz}}$AdoMet</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1,243 ± 1</td>
</tr>
<tr>
<td></td>
<td>$^{7\text{dz}}$Ado'Met</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>no degradation$^g$</td>
</tr>
</tbody>
</table>

$^a$Stability studies conducted at 37 °C, pH 8.0. $^b$hMAT2A substrates were assayed at an initial concentration of ATP = 2,000 µM. $^c$Not determined. $^d$DnrK substrate ClNP was assayed at an initial concentration of AdoMet = 1,600 µM. $^e$Not applicable. $^f$DnrK substrates AdoMet and Ado'Met were assayed at an initial concentration of ClNP = 1,500 µM. $^g$No detectable degradation after 3,300 min.
To better understand the molecular recognition of surrogate substrates Ado^4Met and ClNP by DnrK, the three-dimensional structures for the DnrK/S-adenosyl-tetrazole-L-homocysteine (Ado^4Hcy, the byproduct after methyl transfer) and DnrK-ClNP-AdoHcy were determined at 2.25 and 1.82 Å resolution in space groups C2 and P2_1, respectively (Table 4-6). The average real-space correlation coefficient of all ligands built into these two DnrK structures is over 0.96, and the electron density maps can be found in Figure 4-11. In both crystal forms, DnrK is identified as a tightly packed dimer with a buried surface area of ~3100 Å^2, which is consistent with previously reported DnrK structures. As expected, the DnrK-Ado^4Hcy structure reveals Ado^4Hcy and AdoHcy to bind in a nearly identical manner where the Ado^4Hcy tetrazolate forms extra hydrogen bonds to water molecules and the corresponding sulfur resides slightly closer to the donor nucleophile (Figure 4-10C). This observation is also in agreement with reports of reduced donor-acceptor distances in O-MTs correlating with higher turnover rates, as the $k_{cat}$ of DnrK with Ado^4Met is nearly double the $k_{cat}$ of DnrK with AdoMet. Consistent with the biochemical data, the orientation of one molecule of ClNP in the DnrK-ClNP ligand-bound structure is poised for alkylation reminiscent to that of the native acceptor carminomycin. Surprisingly, five additional ordered ClNPs were also observed in the same structure (Figure 4-10D and Figure 4-11), suggesting a wide array of non-productive ClNP binding modes by virtue of key hydrogen bonds and p-stacking interactions.
**Figure 4-11. Electron density maps of DnrK-Ado^Hcy and DnrK-CINPs**

(A) $2mFo-DFc$ omit map of Ado$^Hcy$ was contoured at 1.3σ. (B) $2mFo-DFc$ omit map of AdoHcy and CINPs was contoured at 1.3σ. (C) $2mFo-DFc$ omit map of AdoHcy and 4MU was contoured at 1.3σ. (D) $2mFo-DFc$ omit map of an unsuccessful example of ligand soaking, which is actually the majority of data collection.
Table 4-6. Crystal parameters, data collection, and refinement statistics

<table>
<thead>
<tr>
<th></th>
<th>DnrK - Ado(^{1})Hcy</th>
<th>DnrK - AdoHcy/CINPs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Crystal parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Space group</td>
<td>( P \ 2_1 )</td>
<td>( C \ 2 )</td>
</tr>
<tr>
<td>Unit-cell parameters (Å, °)</td>
<td>60.6, 104.1, 63.0</td>
<td>123.3, 110.7, 116.9</td>
</tr>
<tr>
<td></td>
<td>90.0, 105.2, 90.0</td>
<td>90.0, 120.2, 90.0</td>
</tr>
<tr>
<td><strong>Data collection statistics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>Resolution range (Å)</td>
<td>50-2.25</td>
<td>50-1.82</td>
</tr>
<tr>
<td>No. of reflections (measured / unique)</td>
<td>152,746 / 35,358</td>
<td>580,623 / 120,915</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.4 (96.7)</td>
<td>98.5 (96.5)</td>
</tr>
<tr>
<td>( R_{\text{merge}} )</td>
<td>0.134 (0.346)</td>
<td>0.072 (0.392)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>4.3 (4.2)</td>
<td>4.8 (4.3)</td>
</tr>
<tr>
<td>Mean I / sigma (I)</td>
<td>7.9 (3.1)</td>
<td>10.5 (2.8)</td>
</tr>
<tr>
<td>( \text{CC}^{1/2} )</td>
<td>0.99 (0.88)</td>
<td>0.99 (0.89)</td>
</tr>
<tr>
<td><strong>Refinement and model statistics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( R_{\text{cryst}} / R_{\text{free}} )</td>
<td>0.184 / 0.246</td>
<td>0.154 / 0.186</td>
</tr>
<tr>
<td>Ligands RSCC(^$)</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>RMSD bonds (Å)</td>
<td>0.011</td>
<td>0.007</td>
</tr>
<tr>
<td>RMSD angles (°)</td>
<td>1.31</td>
<td>1.12</td>
</tr>
<tr>
<td>( B ) factor - protein/ligand/solvent (Å(^2))</td>
<td>23.1 / 13.5 / 31.4</td>
<td>31.1 / 29.9 / 41.6</td>
</tr>
<tr>
<td>No. of protein atoms</td>
<td>5,137</td>
<td>8,138</td>
</tr>
<tr>
<td>No. of waters</td>
<td>491</td>
<td>913</td>
</tr>
<tr>
<td>No. of auxiliary molecules in 1 chain</td>
<td>1 Ado(^{1})Hcy</td>
<td>1 AdoHcy, 6 CINPs</td>
</tr>
<tr>
<td><strong>Ramachandran plot (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Favorable region</td>
<td>98.0</td>
<td>98.2</td>
</tr>
<tr>
<td>Additional allowed region</td>
<td>2.0</td>
<td>1.8</td>
</tr>
<tr>
<td>Disallowed region</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>PDB</strong></td>
<td>5EEG</td>
<td>5EEH</td>
</tr>
</tbody>
</table>

\( R_{\text{merge}} = \sum_h \sum_i |I_i(h) - \langle I(h)\rangle| / \sum_h \sum_i |I_i(h)|, \) where \( I_i(h) \) is the intensity of an individual measurement of the reflection and \( \langle I(h)\rangle \) is the mean intensity of the reflection.

\( \text{ligand RSCC} \) is ligand real-space correlation coefficient, which provides an objective measure of the fit of atom coordinates to electron density.
4.2.3. Complex structure of DnrK with unnatural 4MU

To continue study DnrK after chapter 4.2.2, a few crystallization attempts were made to get new structures of DnrK with new unnatural ligands. Those ligands include resazurin, resorufin and 4-methylumbeliferone (4MU). So far, soaking and co-crystallization experiments of resazurin and resorufin are not successful after trying more 10 crystal forms. DnrK-4MU structure has been solved after screening about 10 crystal forms (Table 4-7, Figure 4-11).

In conclusion, this chapter highlights the efficient chemoenzymatic synthesis of the first enzymatically competent AdoMet isosteres that are considerably less prone to the two primary AdoMet degradation pathways, where the tetrazole isosteric replacement also restricts the potential for AdoMet carboxylate participation in sulphonium racemization. Based on the DnrK-Ado'Met biochemical and structural studies, Ado'Met is able to substitute for AdoMet as a functional enzyme co-substrate and implicates the isosteric modification as potentially advantageous in the context of AdoMet-dependent enzymatic processes, as exemplified by AdoMet decarboxylase catalyzed AdoMet decarboxylation. Notably, AdoMet decarboxylase, a central catalyst in polyamine biosynthesis, is a validated anticancer and antiprotozoal drug target where substrate surrogates lacking the requisite AdoMet carboxylate serve as potent inhibitors. Thus, the conceptual framework for engineering AdoMet chemical and metabolic stability put forth presents unique opportunities to advance AdoMet therapeutic applications including, but not limited to, those involving MTs.
Table 4-7. Crystal parameters, data collection, and refinement statistics

<table>
<thead>
<tr>
<th>Crystal parameters</th>
<th>DnrK - AdoHcy/4MU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>C 2</td>
</tr>
<tr>
<td>Unit-cell parameters (Å, °)</td>
<td>122.2, 111.1, 116.2</td>
</tr>
<tr>
<td></td>
<td>90.0, 120.1, 90.0</td>
</tr>
<tr>
<td>Data collection statistics</td>
<td></td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.98</td>
</tr>
<tr>
<td>Resolution range (Å)</td>
<td>50-1.84</td>
</tr>
<tr>
<td>No. of reflections (measured / unique)</td>
<td>527,978 / 116,128</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>98.5 (99.9)</td>
</tr>
<tr>
<td>$R_{merge}$</td>
<td>0.077 (0.450)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>4.5 (4.5)</td>
</tr>
<tr>
<td>Mean I / sigma (I)</td>
<td>12.6 (3.3)</td>
</tr>
<tr>
<td>$CC^{1/2}$</td>
<td>0.99 (0.89)</td>
</tr>
<tr>
<td>Refinement and model statistics</td>
<td></td>
</tr>
<tr>
<td>$R_{cryst}$ / $R_{free}$</td>
<td>0.160 / 0.187</td>
</tr>
<tr>
<td>Ligands RSCC§</td>
<td>1.16</td>
</tr>
<tr>
<td>RMSD bonds (Å)</td>
<td>0.008</td>
</tr>
<tr>
<td>RMSD angles (°)</td>
<td>1.12</td>
</tr>
<tr>
<td>B factor - protein/ligand/solvent (Å²)</td>
<td>32.1 / 39.9 / 41.8</td>
</tr>
<tr>
<td>No. of protein atoms</td>
<td>8,110</td>
</tr>
<tr>
<td>No. of waters</td>
<td>916</td>
</tr>
<tr>
<td>No. of auxiliary molecules in 1 chain</td>
<td>6 x 4MU</td>
</tr>
<tr>
<td>Ramachandran plot (%)</td>
<td></td>
</tr>
<tr>
<td>Favorable region</td>
<td>98.0</td>
</tr>
<tr>
<td>Additional allowed region</td>
<td>1.8</td>
</tr>
<tr>
<td>Disallowed region</td>
<td>0.2</td>
</tr>
</tbody>
</table>

PDB undeposited

$R_{merge} = \sum_h \sum_i |I_i(h) - <I(h)>| / \sum_h \sum_i I_i(h)$, where $I_i(h)$ is the intensity of an individual measurement of the reflection and $<I(h)>$ is the mean intensity of the reflection.

$§$ligand RSCC is ligand real-space correlation coefficient, which provides an objective measure of the fit of atom coordinates to electron density.
Reference


28. Takusagawa, F., Kamitori, S., Misaki, S. & Markham, G. D. Crystal structure of S-


57. Sakata, S. F. *et al.* Effect of fasting on methionine adenosyltransferase expression and


100. Bitonti, A. J., Dumont, J. a & McCann, P. P. Characterization of Trypanosoma brucei brucei S-adenosyl-L-methionine decarboxylase and its inhibition by Berenil,


Chapter 5

Future Perspectives: Glycorandomization and Alkylrandomization*

As discussed in Chapter 3 and 4, glycorandomization and alkylrandomization are essential and powerful approaches to exploit nature’s chemical arsenal and generate new natural product derivatives. In this chapter I will first briefly summarize the results from chapter 3 and 4 and how those results contribute to the processes of glycorandomization and alkylrandomization. Then I will list what else could be investigated as follow-up projects to accomplish the long-term goal, that is, making new pharmaceutical compounds.

*All contents in chapter 5 are not published and include preliminary results that might be published in future. Most structural work was initiated by me and will be passed to other members in Phillips lab after my graduation. Some of the biochemistry data is from Thorson lab, and some simple assays were conducted in Phillips lab. Methionine library are friendly sent from Thorson lab as well.
5.1. Future Perspectives: Glycorandomization

In glycorandomization (Chapter 3) part, I first solved several sugar aminotransferase (SAT) structures (WecE, CalS13 and AtmS13\textsuperscript{1,2}) to gain knowledge of amino group installation on nucleotide sugar. The universal model that our SAT team built in this study provides a blueprint for future SAT engineering and the biochemical study. Also since the aminoglycosides are functionally and structurally known to participate in a lot of biological activities, like inhibition of protein synthesis, glycosidase inhibition and DNA recognition, this study also helps to better understand the amino-sugar and thus have knowledge contribution to the randomized sugar library\textsuperscript{1,2} (Figure 5-1).

In Chapter 3 I also presented several solved glycosyltransferases structures (SsfS6\textsuperscript{3} and OleD) that can be potentially engineered and developed as a randomized sugar installation tool This is the final and the most exciting step to generate novel pharmaceutical compounds\textsuperscript{4-20}. In SsfS6, while the unknown nature of the SsfS6 aglycon prohibits biochemical study of this intriguing catalyst, our work will advance the study of SF2575 biosynthesis and also set the stage for future protein engineering work on SsfS6. While in OleD case, Thorson lab used the HT "glycosyl-scanning" screen to test a FDA- approved drug library\textsuperscript{21-26}. As expected, this screen lead to some new glycosides (Figure 5-1 bottom right, compound structure not shown).
Figure 5-1. Summary of glycorandomization

The scheme of glycorandomization is shown at top. Crystal structures of sugar aminotransferases and glycosyltransferases are shown at bottom left and bottom middle, respectively. New glycosylated compounds detected are shown at bottom right, and the image is blurred due to confidential reasons.
Briefly, there are two future directions that particularly interest us. First there is a set of putative substrates that are very interesting as they contain no apparent nucleophile but HPLC and HRMS clearly support that a glycoside is enzymatically formed. Ryan Hughes of the Thorson lab is currently working on scaling these reactions up for structure elucidation, and those glycosides are likely to be carbon-glycosides. They present potentially very interesting ligands to consider for co-crystallization experiments or ligand soaking experiments. Specially, chlorpromazine, clomipramine, 4MU are the three acceptors to try each pair with UDP-Glc (or UDP) as the donor. The first two compounds are the substrates for quaternary aminoglycoside formation, which probably contains novel chemistry, but 4MU is also interesting for engineering.

Second, Ryan Hughes of the Thorson lab has concrete evidence of identified putative inhibitors. Of these, one of the most potent to date is silver sulfadiazine. While the IC$_{50}$ is only 1.5 µM, and this is actually among the most potent for a glycosyltransferase inhibitor and few, if any, non-nucleotide based glycosyltransferase inhibitors exist. Thus, this also is a very interesting part to consider for OleD co-crystallization. Except for silver sulfadiazine, mometasone furoate, adapalene and auranofin are also established inhibitors (with IC$_{50s}$ in µM range) and so all those compounds can presumably be tried in co-crystallization without substrates or other ligands. Silver sulfadiazine is the most active inhibitor but all are distinct structural scaffolds and so
there is an interest to understand how these diverse ranges of ligands potentially interact with the enzyme.

Notably, all the biochemistry assay were conducted by the Thorson lab using the newest generation of OleD mutant “loki”\textsuperscript{27-30}, which has been considered to have better substrate promiscuity. Based on this, wt OleD has not been tested on the same assays, and this makes the co-crystallization experiments a little bit tricky. As wt-OleD has been crystallized easily, and so far loki-OleD has not been crystallized after a few attempts. Using wt OleD as protein scaffold may be easier to grow crystals, but the substrates and inhibitors may not bind onto the protein. Using loki-OleD, apparently, has the opposite problems.
5.2. Future Perspectives: Alkylrandomization

In Chapter 4 alkylrandomization part, I first solved structures of methionine adenosyltransferase from *Sulfolobus solfataricus* (sMAT) to expand compound Ado-X library (randomized alkyl- library). The slightly expanded substrate scope of sMAT over other MATs studied to date highlights sMAT as a useful tool for the production of AdoMet analogs. In conjunction with the demonstration of coupled MAT-methyltransferase systems for differential alkylation, this chemoenzymatic strategy circumvents a major liability in the use of synthetic AdoMet analogs - namely, the dramatic instability of the AdoMet analogs. Thus, the structural insights provided herein offer another blueprint from which to pursue future engineering to further broaden the catalyst promiscuity toward AdoMet analog production.

Second, I solved structures of methyltransferase DnrK, which can be potentially engineered and developed as randomized alkyl- installation tool, which is the final and exciting step to generate novel pharmaceutical compounds. There are also examples of methyltransferases that have been applied for this purpose, for example RebM31-34. DnrK is excellent target for alkylrandomization for two reasons: (1) its ability to utilize unnatural AdoMet to prevent AdoMet degradation and (2) its preliminary substrate promiscuity before any mutation engineering.
Figure 5-2. Summary of alkyrlrandomization

The scheme of alkyrlrandomization is shown at top. Crystal structures of methionine adenosyltransferase and methyltransferase are shown at bottom left and bottom middle, respectively. New methylated/alkylated compounds detected are shown at bottom right. RebM is a protein target previously solved by Jason McCoy of the Phillips lab.$^{31}$
5.2.1. Further engineering of sMAT to expand Ado-X library*

A common step in many biosynthetic schemes is the transfer of a methyl group to a substrate by a methyltransferase. The discovery of intrinsically permissive O-methyltransferases that are able to transfer a variety of alkyl groups from AdoMet analogues has enabled the possibility of alkylrandomization for natural product libraries. The feasibility of such an approach has been demonstrated with the published biosynthetic schemes for Rebeccamycin and Daunomycin. However, the feasibility of methyltransferase-facilitated alkylrandomization is currently limited by the lack of methods to produce AdoMet analogues to serve as alkyl donors. Methionine Adenosyltransferases (MATs) catalyze the synthesis of AdoMet from methionine and ATP, allowing them to be exploited in the development of a biosynthetic pathway to generate AdoMet-analogues. An example from this class of enzymes is sMAT. Based on its divergent sequence, sMAT has been hypothesized to have topology and folding distinct from bacterial and eukaryotic MATs that have been thus far characterized. The thermostability of sMAT and its pre-existing ability to use ethionine equivalently with methionine makes it an appealing target for engineering. After finding N159G and the triple mutation of sMAT, this project will be followed up by Jonathan Clinger in the Phillips lab. Phillips lab plans to send those targets to the Thorson lab for HPLC and HRMS validation. The development of a promiscuous sMAT introduces a powerful tool for the production of diverse Ado-X analogues, even at a low temperature. The sMAT team plans to screen more sMAT mutations to complement out sMAT toolbox for the generation of Ado-X library. This
strategy also represents a method for alkylation randomization of natural product libraries, when coupled with a permissive methyltransferase like RebM or DnrK in the future.

5.2.2. Further engineering of DnrK

As discussed in Chapter 4 and Chapter 5.2.1, with the help of sMAT and its engineered version, I now have an Ado-X library in hand. The next step of alkylation randomization will be developing DnrK into a permissive methyltransferase that can install the methyl/alkyl groups on the natural product targets. The previous experience of RebM enzyme engineering could serve as a guideline for this process31–34. The development of permissive methyltransferase DnrK will include, but not limit to, (1) co-crystallization experiments of DnrK with new chemical compounds (ongoing compounds, resazurin and resorufin), (2) high throughput alkyl-scanning screen development to quickly identify whether the alkyl product is formed, (3) protein engineering and dynamics study of DnrK to expand its substrate promiscuity (for both alkyl donors and acceptors), and (4) coupled sMAT and DnrK assay to generate novel alkylated natural products or even novel therapeutics.
References


31. Singh, S. *et al.* Structure and mechanism of the rebeccamycin sugar 4′-O-


The KDM4 histone demethylases are conserved epigenetic regulators linked to development, spermatogenesis and tumorigenesis. However, how the KDM4 family targets specific chromatin is largely unknown. Here, an extensive histone peptide microarray analysis uncovered unexpected trimethyl-lysine histone-binding preferences among the closely related KDM4 double-tudor domains (DTDs). KDM4A/B DTDs bound strongly to H3K23me3, a poorly understood histone modification recently shown to be enriched in meiotic chromatin of ciliates and nematodes. The 2.28-Å co-crystal structure of KDM4A-DTD in complex with H3K23me3-peptide revealed key intermolecular interactions for H3K23me3 recognition. Furthermore, analysis of the 2.56-Å KDM4B-DTD crystal structure pinpointed the underlying residues required for exclusive H3K23me3 specificity, an interaction supported by in vivo co-localization of KDM4B and H3K23me3 at
heterochromatin in mammalian spermatocytes during meiosis. Functionally, H3K23me3 binding by KDM4B stimulated H3K36 demethylation within the same histone tail. Together, these results support a novel epigenetic mechanism whereby H3K23me3-binding by KDM4B directs localized H3K36 demethylation during meiosis and spermatogenesis.

Z.S., F.W., J.H.L., S.D.T., G.N.P and J.M.D. conceived and designed the study. F.W., M.D.M and G.N.P obtained crystals and solved the structures. Z.S. carried out most of the molecular and biochemical experiments. F.W. and Z.S. performed structural modeling and analysis. J.H.L. purified MLA histones and performed demethylase assays. K.E.S., R.P., E.V. and S.D.T. generated H3K23me3 antibody and performed immunofluorescence on rodent testes. K.A.K. performed histone PTM mass spec analysis. J.J.T. purified histones from Tetrahymena. M.D.B. and V.K. synthesized and purified peptides. Z.S. and K.A.K. extracted histones from mouse tissue and cell lines. Z.S., F.W., S.D.T. and J.M.D. drafted the manuscript. All authors discussed the results and commented on the manuscript.

*Portions of appendix A will be published as: Su Z, Wang F, Lee JH, Stephens KE, Papazyan R, Voronina E, Kraykramer KA, Thorpe JJ, Boersma MD, Kuznetsov V, Miller MD, Taverna SD, Phillips Jr GH, Denu JM. Reader Domain Specificity Directs Lysine Demethylase-4 Family Function. The paper has been submitted but not published when this thesis is written. Su Z and Wang F are co-first authors.
A.1. Human lysine femethylase-4 (KDM4) and their reader domains

Histone lysine methylation regulates gene expression by recruiting or displacing chromatin-binding proteins\(^1\text{–}^4\). KDM4 (JMJD2) is a conserved iron-dependent jumonji-domain demethylase family that is essential during development\(^5\text{–}^8\). Disrupting the only KDM4 enzyme in *C. elegans* induced germ cell apoptosis and DNA replication defects\(^9\). Overexpression of individual mammalian KDM4 proteins has been associated with oncogenesis, cancer growth and metastasis in various cancer types and other conditions including cardiac failure and autism\(^10\text{–}^{12}\). In vertebrates, KDM4A, KDM4B and KDM4C share similar domain organization\(^13\) (Appendix Figure A-1). The N-terminal catalytic domains of KDM4A-C display demethylase activity that can convert di-/tri-methylated lysines to lower methylated states at H3K9 and H3K36 with comparable kinetics\(^13\). Despite similar catalytic activities, individual KDM4 members exhibit varied chromatin associations and biological functions\(^14\text{–}^{16}\). These observations suggest an uncharacterized mechanism controls KDM4 protein functions on chromatin.

Vertebrate KDM4A-C proteins contain a conserved double-tudor domain (DTD) and a potential zinc-finger domain (ZnF) at the C-terminus (Appendix Figure A-1a). These types of chromatin-interacting modules (also known as reader domains) often mediate binding to specific histone modification states\(^17\text{–}^{19}\). Tudor domains are part of the “Royal Family” reader domains, which usually recognize methylated lysine residues\(^20,21\). In particular, double tudor domain from KDM4A (KDM4A-DTD) was
shown to form an unique integral structural unit and recognize methylated lysines. Deletion of the C-terminal domain in KDM4 proteins resulted in a change of subcellular localization, changed demethylase activity and disruption of other KDM4 functions, suggesting functional roles for the C-terminal DTDs. However, there has been no comprehensive investigation of the histone binding properties for KDM4B and KDM4C DTDs.

To better understand how reader domains regulate the overall chromatin-acting functions among the closely related KDM4 family members, the Denu lab determined and compared histone interactomes of the C-terminal double-tudor domains in human KDM4A-C proteins (Appendix Figure A-1a). From our biochemical and structural profiling, Denu lab found KDM4A, KDM4B and KDM4C DTDs display different histone-binding preferences. Denu lab showed that these double-tudor domains utilize an aromatic cage as a general mechanism to coordinate trimethyllysine and most importantly, the sequence specificity is largely determined by side-chain interactions with surrounding residues. Specifically, Denu lab describe the unique interaction between KDM4-DTDs and H3K23me3, a histone modification enriched in heterochromatin during meiosis in primary spermatocytes. Our new crystal structures and homology models explained the origins of H3K23me3 specificity by KDM4B, and these biochemical and structural data were supported by the co-localization of full-length KDM4B with H3K23me3 in vivo. Denu lab demonstrate that H3K23me3 binding stimulates cis-histone H3K36 demethylation by
KDM4B. These results suggest H3K23me3 as a novel link between KDM4 family and mammalian germ cell development, potentially to eliminate methylation at H3K36 at heterochromatic regions. The molecular determinants of histone-binding specificity by double tudor domains serves as a general framework to understand KDM4 family functions.

A.2. Systematic profiling reveals H3K23me3 binding specificities among KDM4A-C double-tudor domains

Denu lab cloned, expressed and purified human KDM4A-C double tudor domains (DTDs) as recombinant proteins with N-terminal affinity tags from E. coli. KDM4A-C DTDs were probed with our recently developed combinatorial histone peptide microarray covering 746 histone PTM states\textsuperscript{29}. Analysis of the peptide microarray assay revealed surprising discrimination for H3K23me3 binding among the three double tudor domains (Appendix Figure A-1b). Particularly, KDM4B-DTD displayed exclusive binding to H3K23me3 (Appendix Figure A-1b). KDM4A-DTD, which displayed strong binding to H3K23me3 as well, also bound H3K4me3 and H4K20me3 (Appendix Figure A-1b), consistent with previous findings\textsuperscript{23,30}. In contrast, KDM4C-DTD bound specifically to H3K4me3 (Appendix Figure A-1b).
Appendix Figure A-1. Distinct binding specificities of human KDM4A-C double-tudor domains reveal H3K23me3 discrimination

(a) Domain organization of human KDM4 family members. (b) Systematic profiling of histone binding preferences of KDM4A-C DTDs on histone peptide microarray. (c) Quantitative measurement of H3 (17-32) K23me3 binding constant with KDM4A-C DTDs by fluorescence polarization. (d) Differential histone interactome of KDM4A-C DTDs validated by fluorescence polarization assays. (e) Sequence alignment of KDM4A-C DTDs.
Denu lab further quantified the methylated histone binding of KDM4-DTDs by employing a solution-based binding assay (fluorescence polarization, Appendix Figure A-1cd). Overall, the derived binding constants were consistent with the peptide array analysis. KDM4A-DTD binds H3K23me3 peptide at low micromolar affinity ($K_d = 2.2 \mu M$), slightly stronger than its interaction with H3K4me3 and H4K20me3 (Appendix Figure A-1d), and KDM4C-DTD displayed highly specific binding to H3K4me3 ($K_d = 6.8 \mu M$) (Appendix Figure A-1d). In dramatic contrast to KDM4A-DTD and KDM4C-DTD, KDM4B-DTD only displayed striking specificity for H3K23me3 ($K_d = 10.3 \mu M$), with no significant binding to H3K4me3 (Appendix Figure A-1d). Collectively, this systematic analysis indicates that KDM4A-C DTDs have distinct and surprising histone binding preferences that could not have been predicted based on the high similarity in their overall amino acid sequences (>60% identity, >80% similarity, see Appendix Figure A-1e). Nevertheless, how KDM4-DTDs achieve sequence specificity for residues proximal to the trimethyl-lysine remained unclear. To define the mechanism of such sequence selectivity for KDM4-DTD binding, I proceeded to structurally and biochemically detail these interactions.
A.3. X-ray crystallography uncovers structural basis for H3K23me3 recognition

To elucidate the molecular basis of recognizing H3K23me3, purified KDM4A-DTD was co-crystallized with an H3 (19-28) K23me3 peptide, and the peptide-bound structure was determined at 2.3-Å with a space group of P 32 (PDB ID: 5D6Y) (Appendix Figure A-2a-c, Appendix Table A-1, Appendix Figure A-3). From the crystallization screening, I also determined two apo structures of KDM4A-DTD at 1.99-Å and 2.15-Å (PDB ID: 5D6W, 5D6X; Appendix Table A-1). In all the structures, KDM4A-DTD forms two lobes (HTD-1 and HTD-2, HTD = hybrid tudor domain) with interweaving β-strands. The co-crystal structure suggests the H3K23me3 peptide forms extensive interaction with KDM4A-DTD (~540 Å² by PISA[31]) and such binding is mostly mediated through residues in HTD-2 of KDM4A-DTD.

The structure of the H3K23me3-bound KDM4A-DTD shows that the trimethyl-lysine group of H3K23 is coordinated via cation-π and hydrophobic interactions by an aromatic cage formed by F932, W967 and Y973 of the reader domain. To examine the selectivity of methylation state on H3K23, Denu lab determined binding constants of KDM4A-DTD for mono-/di-/tri-methylated H3K23 peptides. The binding results suggest the KDM4A-DTD aromatic cage favors trimethyl-lysine (Kd = 2.20 μM) over dimethyl-lysine (Kd = 9.06 μM), or monomethyl-lysine (Kd = 98.55 μM) at H3K23 (Appendix Figure A-2e). Correspondingly, H3K23me3 binding is greatly decreased (Kd = 284.90 μM) with the aromatic cage substitution Y973A. The conserved role of this aromatic cage in directing PTM selectivity (Appendix Figure A-3) towards
trimethyl-lysine suggests specific recognition with H3K23me3 is influenced by neighboring residues besides the trimethyl-lysine moiety on histones. The orientation of H3K23me3 peptide in KDM4A is more similar to H3K4me3 than H4K20me3 peptide (Appendix Figure A-3b). However, other H3K4me3 reader domains like ING2-PHD could not bind H3K23me3 (Appendix Figure A-3c), because ING2-PHD recognizes the free N-terminus (H3A1) of H3K4me3 peptide and does not accommodate the longer N-terminus of H3K23me3. In contrast, the more open channel between the two lobes in double-tudor domains can accommodate the longer polypeptide chain N-terminal to H3K23me3.

H3T22, the neighboring residue of H3K23, plays an important role in the H3K23me3 recognition by KDM4A-DTD. H3T22 side-chain forms a hydrogen bond with N940, while its main-chain forms hydrogen bonds with N940 and S938 side-chains (Appendix Figure A-2f). The importance of H3T22 in H3K23me3-binding is illustrated by the decreased H3K23me3 binding with phosphorylation on H3T22 ($K_d = 83.69$ μM) or when N940 is changed into Ala in KDM4A ($K_d = 57.41$ μM) (Appendix Figure A-2g), both of which would interfere with the hydrogen bond of H3T22 side-chain. The effect of H3T22ph on H3K23me3-binding is reminiscent of the relationship between H3T3ph and H3K4me3. Indeed, a similar loss of KDM4A binding to H3K4me3 was observed with H3T3ph (Appendix Figure A-3d).
Appendix Figure A-2. Co-crystal structure of KDM4A-DTD with H3K23me3 (1)

(a-c) Co-crystal structure of KDM4A-DTD and H3 (19-27) K23me3 peptide at 2.28-Å resolution. (d-j) Biochemical validation of intermolecular interactions observed in KDM4A-H3K23me3 structure. (k) Schematic representation of the molecular interactions between KDM4A-DTD and H3(19-27)K23me3 peptide as identified in the co-crystal structure.
Appendix Figure A-3. Co-crystal structure of KDM4A-DTD with H3K23me3 (2)

(a) Packing view of six subunits in the KDM4A-H3K23me3 structure. (b) Comparison of histone peptide orientation in the KDM4A-DTD co-crystal structures. (c) H3K4me3-reader ING2-PHD does not interact with H3K23me3 as determined by fluorescence polarization. (d) KDM4A-N940 and H3T22 interaction pair is important for H3K4me3 binding. (e) Surface electronic potential of H3K23me3 binding by KDM4A-DTD. The zoomed in box highlights H3R26 binding site. (f) KDM4A-DTD N931D specifically increases H3K23me3 binding. (g) Sequence of surrounding residues for H3K14, H3K23, H3K4 and H4K20.
### Appendix Table A-1. Data collection and refinement statistics of all KDM4s

<table>
<thead>
<tr>
<th></th>
<th>KDM4B-DTD apo form</th>
<th>KDM4A-DTD H3K23me3</th>
<th>KDM4A-DTD apo form 1</th>
<th>KDM4A-DTD apo form 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data collection</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Space group</td>
<td>$P 3 1$</td>
<td>$P 3 2$</td>
<td>$R 3 2 : h$</td>
<td>$I 2 3$</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td>$a, b, c$ (Å)</td>
<td>$77.4, 77.4, 50.8$</td>
<td>$106.2, 106.2, 79.2$</td>
<td>$190.6, 190.6, 117.0$</td>
</tr>
<tr>
<td></td>
<td>$a, b, g$ (°)</td>
<td>$90.0, 90.0, 120.0$</td>
<td>$90.0, 90.0, 90.0$</td>
<td>$90.0, 90.0, 90.0$</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>30-2.56 (2.68-2.56)</td>
<td>30-2.28 (2.32-2.28)</td>
<td>36-1.99 (2.01-1.99)</td>
<td>42-2.15 (2.28-2.15)</td>
</tr>
<tr>
<td>$R_{merge}$</td>
<td>0.120 (0.797)</td>
<td>0.225 (1.885)</td>
<td>0.130 (1.158)</td>
<td>0.076 (0.523)</td>
</tr>
<tr>
<td>$I / sI$</td>
<td>6.73 (2.28)</td>
<td>12.95 (1.42)</td>
<td>12.6 (2.35)</td>
<td>25.1 (5.74)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>100.0 (100.0)</td>
<td>100.0 (100.0)</td>
<td>100.0 (100.0)</td>
<td>100.0 (100.0)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>6.6 (6.3)</td>
<td>11.3 (8.7)</td>
<td>9.8 (9.7)</td>
<td>13.8 (13.8)</td>
</tr>
<tr>
<td>$CC_{1/2}$</td>
<td>0.99 (0.74)</td>
<td>0.99 (0.46)</td>
<td>0.99 (0.67)</td>
<td>0.99 (0.95)</td>
</tr>
<tr>
<td><strong>Refinement</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>30-2.56</td>
<td>30-2.28</td>
<td>36-1.99</td>
<td>42-2.15</td>
</tr>
<tr>
<td>No. reflections</td>
<td>71,716</td>
<td>466,599</td>
<td>542,229</td>
<td>306,479</td>
</tr>
<tr>
<td>$R_{work} / R_{free}$</td>
<td>0.197 / 0.254</td>
<td>0.264 / 0.295</td>
<td>0.168 / 0.195</td>
<td>0.159 / 0.199</td>
</tr>
<tr>
<td>Peptide RSCC*</td>
<td>---</td>
<td>0.92</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>No. atoms (non-H)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>1,716</td>
<td>5,069</td>
<td>3,795</td>
<td>1,974</td>
</tr>
<tr>
<td>Peptide</td>
<td>---</td>
<td>248</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Water</td>
<td>66</td>
<td>216</td>
<td>515</td>
<td>242</td>
</tr>
<tr>
<td>B-factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>48.2</td>
<td>42.0$^8$ / 99.1$^9$</td>
<td>39.6</td>
<td>34.9</td>
</tr>
<tr>
<td>Peptide</td>
<td>---</td>
<td>46.0</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Water</td>
<td>44.7</td>
<td>48.3</td>
<td>53.0</td>
<td>50.2</td>
</tr>
<tr>
<td>R.m.s. deviations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.009</td>
<td>0.011</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>1.12</td>
<td>1.39</td>
<td>1.10</td>
<td>1.11</td>
</tr>
<tr>
<td>PDB</td>
<td>4UC4</td>
<td>5D6Y</td>
<td>5D6W</td>
<td>5D6X</td>
</tr>
</tbody>
</table>

---

*RSCC is the real-space correlation to electron density calculated by phenix.*

$^8$Average B-factor of ordered chains A-D in asymmetric unit.

$^9$Average B-factor of disordered chains E-F in asymmetric unit.
Unique contacts with H3R26 also contribute to H3K23me3 recognition by KDM4A. The H3R26 side-chain forms a hydrogen bond with KDM4A N931 (2.9-Å), while the main-chain of H3R26 forms a hydrogen bond with the KDM4A-DTD G935 main-chain (Appendix Figure A-2h). H3R26 fits nicely into a negatively charged cleft formed by residues spanning from E929 to D933. To provide direct biochemical evidence of the H3R26-N931 interaction pair in H3K23me3 recognition, Denu lab modified H3R26 in the context of H3K23me3 peptide and observed decreased binding (32.3 μM) between wild-type KDM4A-DTD and H3K23me3R26A peptide (Appendix Figure A-2i). Interestingly, modifying H3R26 with asymmetric dimethylation dropped H3K23me3 binding mildly to 8.7 μM, suggesting KDM4A-DTD might read H3K23me3 and H3R26me2a in combinatorial fashion. Zhangli in the Denu lab also observed a dramatic decrease in H3K23me3 binding (73.9 μM) with N931R substitution (Appendix Figure A-2j), which could form electrostatic repulsion with H3R26. Conversely, changing N931 into aspartic acid (N931D) increased H3K23me3 binding by about 3-fold ($K_d = 0.85$ μM), potentially by providing ionic interaction with H3R26. Interestingly, N931D also increased H3K14me3 binding ($K_d = 12.24$ μM), likely due to the enhanced ionic interaction between D931 and H3R17 at the +3 site of H3K14me3, reminiscent of the distance between H3K23me3 and H3R26 at the +3 position (Appendix Figure A-3g). Overall the H3R26-N931 pair mediates H3K23me3 recognition by KDM4A-DTD and this contact is unique to H3K23me3 binding. Besides main-chain hydrogen bonds between H3R26 and G935, other main-chain hydrogen bonds are found between H3A24 and F937, together forming three β sheet-like main-chain hydrogen bonds. The intermolecular interactions determined from the KDM4A-
H3K23me3 co-crystal structure involve all residues from H3L20 to H3R26 (Appendix Figure A-2k), consistent with the clear electron density of this region in the structure. The molecular arrangement of the H3K23me3-bound structure defines an H3K23me3-specific interaction network and suggests that modifications on H3T22 and H3R26 could play regulatory roles in modulating DTD binding.

**A.4. H3K23me3 recognition is mediated by unique side-chain interactions**

Previously, H3K4me3 and H4K20me3 were characterized as histone PTMs recognized by KDM4A-DTD\(^{23,30}\). D945 of KDM4A was identified as the residue important for H3K4me3 binding by forming a salt bridge with H3R2, whereas D939 was reported critical for H4K20me3 binding due to its interaction with H4R19. I noted that neither D945 nor D939 contacts the H3K23me3 peptide in the KDM4A-H3K23me3 structure. Using this observation, I generated a series of D945 substitutions (D945A/L/S) and found that they did not affect H3K23me3 binding, although such changes did decrease H3K4me3 binding. Similarly, D939A only weakens H4K20me3 binding but not H3K23me3 binding or H3K4me3 binding. The fact that individual KDM4-DTD substitutions at different residues could exclusively disrupt H3K23me3, H3K4me3 and H4K20me3 binding respectively, without affecting the other two interactions, demonstrates that H3K23me3 recognition is distinct from either H3K4me3 or H4K20me3 interactions.
To further confirm the specific H3K23me3 recognition is mediated by unique side chain interactions, Denu lab combined amino acid substitutions that reduce H3K4me3 (D945R) and H4K20me3 (D939R) affinity while increasing H3K23me3 affinity (N931D) as a triple mutant (“DRR”) in KDM4A-DTD (Appendix Figure A-4ab). The DRR mutant retained a low micromolar Kd value for H3K23me3 binding while discriminating against other PTM states, leading to overall enhanced H3K23me3 specificity (Appendix Figure A-4c). The engineered “DRR” reader domain specifically recognized H3K23me3 full-length recombinant histone (Appendix Figure A-4d). The successful engineering to selectively recognize H3K23me3-containing histones highlights the detailed molecular knowledge gleaned from this study.

A.5. KDM4B-DTD is an H3K23me3 specific reader domain

The above biochemical analysis validated the intermolecular interactions between KDM4A-DTD and H3K23me3 observed from the co-crystal structure. Because KDM4B-DTD also interacts with H3K23me3 with high discrimination, I investigated whether KDM4B-DTD engages H3K23me3 using a molecular network similar to KDM4A-DTD. Indeed, KDM4B-DTD binds H3K23me3 with sensitivity to PTM states on H3K23 and neighboring residues in similar fashion to that observed for KDM4A-DTD (Appendix Figure A-5ab), suggesting that the interaction network with H3K23me3 is shared between KDM4A-DTD and KDM4B-DTD. However, our initial analysis showed a unique specificity of KDM4B-DTD: unlike KDM4A-DTD and KDM4C-DTD, KDM4B-DTD displayed dramatically poor interaction with H3K4me3.
Such poor interaction with H3K4me3 was also confirmed with more physiologically relevant chromatin substrates, native nucleosomes purified from human cell line MCF-7, and KDM4B-DTD showed no enrichment of H3K4me3-containing chromatin (Appendix Figure A-6a). To better understand the molecular basis of the specific H3K23me3 binding and the disfavored H3K4me3 binding by KDM4B-DTD, I crystalized KDM4B-DTD and successfully determined the structure of the apo-form at a resolution of 2.56-Å (PDB ID: 4UC4, Appendix Table A-1). Analysis of the first KDM4B-DTD structure reveals a highly similar overall tertiary structure when compared to KDM4A-DTD and KDM4C-DTD (Appendix Figure A-6b). Next I modeled KDM4B-DTD with the H3(19-27)K23me3 (Appendix Figure A-5c) or H3(1-7) K4me3 (Appendix Figure A5d-g) peptide, based on the highly conserved structural fold of HTD-2 from the corresponding KDM4A co-crystal structures (PDB ID: 5D6Y and 2GFA). Indeed, the KDM4B-H3K23me3 model displayed a similar interaction network (Appendix Figure A-6c) as KDM4A-H3K23me3, including the hydrogen bonds with H3T22 and H3R26. Specific KDM4B mutants that impaired its H3K23me3 binding also confirm the KDM4B-H3K23me3 model (Appendix Figure A-6d).
Appendix Figure A-4. H3K23me3 recognition is modulated by unique side-chain interactions

(a-b) Structural representation and a table summary about the side-chain interactions for trimethyl-lysine histone recognition by KDM4A-DTD. (c) Triple mutant (N931D/D939R/D945R, or DRR) reader is specific for H3K23me3. (d) DRR mutant fails to recognize H3Kc4me3 methyl-lysine analog (MLA) but retains binding to H3Kc23me3 MLA histone in a far western setting.
Appendix Figure A-5. KDM4B-DTD is an H3K23me3-specific reader (1)

(a-b) Conserved interaction between KDM4B-DTD and H3K23me3. (c) Based on the structural alignment of HTD-2 domains, H3(19-27)K23me3 peptide from KDM4A structures is modeled into KDM4B-DTD apo structure. (d-g) Key residues contributing to decreased H3K4me3 binding identified from KDM4B-H3K4me3 model. (f-g). Corresponding residues in KDM4A are labeled. Enhanced H3K4me3 binding by KDM4B mutants N960A and S965D.
**Appendix Figure A-6. KDM4B-DTD is an H3K23me3-specific reader (2)**

(a) MARCC confirms enrichment of H3K4me3-containing chromatin by KDM4A and KDM4C but not KDM4B. (b) Ribbon representation of KDM4A-C DTD overall structures aligned by HTD-2 domains. (c) Interaction network in the modeled KDM4B-H3K23me3 structure. (d) H3K23me3 binding of KDM4B-DTD mutants. (e) Clash observed between H3R26 and M911 in modeled KDM4C-H3K23me3 structure. (f) Effects of H3K23 methylation state and secondary modifications on H3T22/R26 for H3K23me3 binding with KDM4B-DTD and KDM4C-DTD.
To pinpoint the residues that account for the poor H3K4me3 affinity of KDM4B, I looked for apparent clashes or missing molecular interactions in the modeled KDM4B-H3K4me3 structure (Appendix Figure A-5d-g). I first noticed that N960 in KDM4B-DTD showed a significant steric clash with H3A1 and H3R2 in the model. To test if N960 hinders the interaction between H3K4me3 and KDM4B-DTD, Denu lab made the Asn-to-Ala substitution, which should alleviate the steric clash. Indeed, with the N960A variant, H3K4me3 binding was rescued to $K_d = 303 \, \mu M$; whereas the corresponding substitutions in KDM4A (N940A) dramatically decreased H3K4me3 binding most likely due to disrupting the hydrogen bond with H3T3. S965 in KDM4B might also contribute to the poor affinity for H3K4me3, likely due to the lack of strong ionic interaction between H3R2 and S965; such ionic interaction is fulfilled between KDM4A-D945 and H3R2 (Appendix Figure A-5f). Additionally, KDM4B-S965 forms a weaker hydrogen bond with H3R2 in the modeled structure than KDM4A-D945. Consistent with this hypothesis, H3K4me3 binding in KDM4B-S965D ($K_d \sim 200 \, \mu M$) is tighter than that in wildtype KDM4B. At the same time, corresponding substitutions of KDM4A-D945S displayed lower H3K4me3 binding than the corresponding wild-type proteins. Similar models of KDM4C-DTD with H3K4me3 and H3K23me3 suggest general conserved interactions as with KDM4A-DTD, except for the H3R26 clash in KDM4C-H3K23me3 model (Appendix Figure A-6e-f), which might explain KDM4C’s lower affinity to H3K23me3. Together these results suggest that despite overall sequence and structure homology, specific differences of key residues result in unique H3K23me3 specificity of KDM4B-DTD. This also highlights the multiple side-chain interactions that define the sequence specificity of KDM4 reader domains.
A.6. H3K23me3 is enriched in primary spermatocytes that undergo meiosis

The specific H3K23me3 interaction by human KDM4 reader domains in vitro prompted us to investigate further the physiological context of such interaction. Only recently has a biological context for the histone PTM H3K23me3 been described in Tetrahymena and C. elegans\textsuperscript{32,33}. However, direct evidence of its relevance in mammals has not been documented. To search for a physiologically relevant source of H3K23me3 in mammalian systems, Denu lab first acid-extracted endogenous histones from a panel of mammalian cell lines and mouse tissues. Denu lab also purified histones from germline micronuclei and non-germline macronuclei in Tetrahymena, as H3K23me3 was previously found exclusively enriched in the micronuclei\textsuperscript{32}. Using an H3K23me3-specific antibody produced in-house, Denu lab found specific enrichment of H3K23me3 in mouse testes compared to other non-germline sources (Appendix Figure A-7a). Higher H3K23me3 levels in germline (micronuclei) histone were detected in Tetrahymena, consistent with previous observations. Using the recombinant H3K23me3-specific reader as a probe (DRR mutant), a similar enrichment pattern was observed in testes histones (Appendix Figure A-7b).
Appendix Figure A-7. H3K23me3 is enriched in heterochromatin of primary spermatocytes that undergo meiosis

(a) Specific expression of H3K23me3 in different mouse tissues and mammalian cell lines. (b) Germline-enriched H3K23me3 detected by H3K23me3 antibody and DRR reader. (c) Immunofluorescence identifies H3K23me3 enrichment in meiotic cells. (d) Mass spectrometry quantification of modifications on H3K23-containing peptide. (e) KDM4B (red) shows co-localization with H3K23me3 (green) by indirect immunofluorescence in rat testes. (f-g) H3K23me3 (green) is localized at heterochromatin rather than active chromatin as suggested by co-localization with H3K27me3 (d) but not H3K9ac (e) (red).
To test if H3K23me3 detected in mouse testes also displays association with meiotic chromatin as found previously in *Tetrahymena* and *C. elegans*, Denu lab performed indirect immunofluorescence of H3K23me3 in mouse testes (Appendix Figure A-7c). Taverna lab detected an enrichment of H3K23me3 in mouse primary spermatocytes undergoing meiosis, as indicated by the appearance of the meiosis-marker SCP3 (synaptonemal complex protein 3), which coincides with meiotic entry and leptotene. H3K23me3 signals were also detected in spermatids that finished meiosis, but much weaker in spermatogonia that have not entered meiosis. The specificity of the H3K23me3 antibody was validated by peptide competition. To further validate the higher methylation levels of H3K23 in mouse testes as detected by antibody and the recombinant reader probe, Denu lab also quantified different modification levels on H3K23 by mass spectrometry. Indeed, Denu lab observed relatively higher global levels of H3K23 methylation (~10% using the quantified levels of monomethylation) in testes (Appendix Figure A-7d) compared with levels (<1%) observed in other mammalian tissues or cell lines\(^{34-36}\) (Levels of di- and tri-methylation at H3K23 are difficult to quantify by mass spec due to the low global abundance in mammals). This observation provides new evidence on H3K23me3 being a conserved histone PTM associated with meiosis in mammalian systems.

Having established the physiological link of H3K23me3 with meiosis in mammalian spermatogenesis, Denu lab next investigate the in vivo connection between H3K23me3 and its putative reader, KDM4. Because KDM4B harbors the unique ability
to read the H3K23me3 modification, Denu lab probed rat testes tissue to determine the extent of overlapping nuclear localization between KDM4B full-length protein and H3K23me3. A high degree of overlap was found between KDM4B and H3K23me3 staining in spermatocytes (Appendix Figure A-7e). The temporal and sub-nuclear overlap of both H3K23me3 and KDM4B in the same cell population suggests a relevant connection during meiosis. Another heterochromatic histone PTM, H3K27me3, was more broadly distributed across nuclei of germline cells; however, Denu lab found that in meiotic cells, H3K27me3 staining was highly correlated with the punctate distribution of H3K23me3 (Appendix Figure A-7f). In contrast, the euchromatic histone PTM H3K9ac is more diffused inside the spermatocyte nucleus, and does not overlap with H3K23me3 (Appendix Figure A-7g). These immunofluorescence results map H3K23me3 to heterochromatin, but not euchromatin. Altogether, our results suggest that H3K23me3 helps target KDM4B within heterochromatin of mammalian meiotic spermatocytes.

A.7. H3K23me3-binding stimulates H3K36 demethylation by KDM4B

The co-localization of H3K23me3 and KDM4B in vivo, and the specific ability of KDM4B to bind H3K23me3 in vitro prompted us to investigate the functional role of H3K23me3 binding by KDM4 proteins. KDM4A-C proteins are conserved histone lysine demethylases that have been linked to removal of H3K9me and H3K36me3. More recently, the catalytic domains of KDM4A-C were shown to demethylate H3K27 in vitro as well, although at a much slower rate. Having determined binding
specificity of the C-terminal double-tudor domain in KDM4B for H3K23me3, Zhangli and I reasoned that this interaction might enhance the activity of full-length KDM4B against certain histone methylations that co-occur with H3K23me3 in cis (Appendix Figure A-8a).

To investigate the functional consequences of KDM4 proteins binding to H3K23me3, I built several structural models that incorporate both the catalytic and the double tudor domain. One structural model included KDM4A and dually modified peptide H3K4me3K9me3 (Appendix Figure A-9a), and demonstrated the simultaneous engagement of H3K4me3 and H3K9me3. However, when the C-terminal reader domain binds H3K23me3, this binding mode requires the opposite orientation (head to toe) between histone H3 and KDM4 protein, suggesting that H3K27 and H3K36, but not H3K9, are more likely to be the demethylase target when H3K23me3 is bound to the double-tudor domain. Our structural analysis shows no molecular interaction between KDM4B double-tudor domain and the H3K27 or H3K36 residue, suggesting H3K23me3 binding should not directly impede H3K27 or H3K36 demethylation. I then built a structural model of KDM4B and H3K23me3K27me3 (Appendix Figure A-9b), which provided evidence that either H3K27 or H3K36 is accessible for demethylation by the KDM4B catalytic domain.
To experimentally test the ability of KDM4B to demethylate H3K27me3 or H3K36me3 upon H3K23me3-binding, Zhangli compared demethylation activity toward H3K27me3 or H3K36me3 by full-length KDM4B in the presence or absence of H3K23me3 on the same histone peptide (Appendix Figure A-8b and Appendix Figure A-9c). Using tandem mass spectrometry, Denu lab did not detect significant demethylation of H3K27me3 peptide by KDM4B, even after prolonged incubation, consistent with previous reports of H3K27me3 being a poor substrate for the KDM4 family\textsuperscript{5,6,8,37}. The lack of measurable demethylation on the dually modified H3K23me3K27me3 peptide indicated that the negligible demethylation of H3K27me3 was not be stimulated by H3K23me3 binding. In stark contrast, Denu lab observed consistent H3K36me3 demethylation by KDM4B, and the catalytic efficiency ($V_{\text{max}}/K_M$) of H3K36me3 demethylation increased by 4-fold when the doubly modified H3K23me3K36me3 peptide was the substrate (Appendix Figure A-8c). This enhanced demethylation was the primary result of the lower $K_M$ value for H3K23me3-containing peptide ($K_M = 0.3$ μM) versus the singly modified H3K36me3 peptide ($K_M = 1.3$ μM), and is consistent with the intrinsic ability of the KDM4B-DTD to bind H3K23me3 with low micromolar affinity. Using a mass spectrometry-based demethylation assay, Denu lab corroborated the findings: the co-existence of H3K23me3 and H3K36me3 on the same peptide yielded faster rates of H3K36me2 and H3K36me1 product formation and a corresponding faster depletion of substrate (H3K23me3) compared to peptide with only H3K36me3 present (Appendix Figure A-8d). The mass spec analysis also confirms that increased activity of KDM4B in the
presence of H3K23me3 is indeed due to demethylation of H3K36me3, and not the result of altered demethylation at H3K23me3.

**Appendix Figure A-8. H3K23me3 binding stimulates H3K36 demethylation in KDM4B**

(a) Summary of histone PTM targets of reader domain and demethylase domain for KDM4B. (b) Mass spectrometry analysis of demethylation reaction by KDM4B full-length protein. (c) Comparison of Vmax/KM derived from Michaelis-Menten curves of demethylation of H3K36me3 and H3K23me3K36me3 peptides by full-length KDM4B enzyme. (d) Co-existing H3K23me3 increases H3K36me3 conversion to H3K36me1/2 by KDM4B. (e) KDM4B (red) shows distinct localization with H3K36me3 (green) by indirect immunofluorescence in rat primary spermatocytes. (f) Proposed model of H3K23me3-mediated H3K36me3 demethylation by KDM4B during meiosis.
Appendix Figure A-9. H3K23me3-mediated demethylation by KDM4B

(a-b) Structural models of simultaneous coordination by catalytic domain and reader domain in KDM4A/B. (a) H3(1-13)K4me3K9me3 peptide was modeled with KDM4A. (b) H3(19-28)K23me3K27me3 peptide is modeled with KDM4B. (c) Mass spectrometry analysis of demethylation reaction by KDM4B full-length protein.
Based on the observation that full-length KDM4B displays enhanced demethylase activity at H3K36me3 in the context of doubly modified H3K23me3K36me3, Denu lab determined the sub-nuclear distribution of KDM4B and H3K36me3 in rat primary spermatocytes (Appendix Figure A-8e). Consistent with the in vitro kinetic results, H3K23me3-specific KDM4B and H3K36me3 staining did not overlap and supports the efficient demethylation of H3K36me3 at these loci. Together the results suggest that the interaction between H3K23me3 and KDM4B-DTD endows the demethylase with enhanced catalytic efficiency for removing H3K36me3 when these chromatin marks exist on the same histone tail.

**A.8. Conclusions highlights**

In this study, Zhangli and I described the molecular basis and functional relevance of H3K23me3 recognition by the KDM4A and KDM4B double-tudor domains. Biochemical analysis revealed the intricate molecular interactions that govern PTM specificity and sequence specificity of KDM4-DTDs for H3K23me3. In particular, I demonstrated the significance of side-chain interactions with residues surrounding the trimethyl-lysine in mediating KDM4B-DTD's binding preference of H3K23me3 over H3K4me3. This study suggests a novel role for KDM4B in heterochromatin maintenance during meiosis, mediated by H3K23me3-dependent binding and demethylation of H3K36me3 (Appendix Figure A-8f). Lastly, our results support a conserved role of H3K23me3 during germ cell development.
KDM4 demethylases are conserved epigenetic regulators of histone methylation that are required for normal development\textsuperscript{5-10}. The C-terminal DTDs in the multiple KDM4 homologs are only present in vertebrates (Appendix Figure A-10), suggesting functional diversification as a result of genome duplication during evolution. The sophisticated molecular network present in KDM4-DTDs revealed by this study underlies the histone interactomes of human KDM4A-C DTDs, and sheds light on distinct functions of KDM4 members\textsuperscript{11-16}. The mild phenotype of individual KDM4 knockout mice\textsuperscript{40,41} implicates some compensating functions among KDM4 members, which is consistent with overlapping histone PTM interactomes among KDM4A-C DTDs. However, the H3K23me3-specific reader KDM4B-DTD displays poor H3K4me3 binding, while KDM4A-DTD and KDM4C-DTD have low-micromolar binding affinity with H3K4me3. This observation is in line with previous findings that KDM4C but not KDM4B overlaps with H3K4me3 genomic distribution, and the fact that targeting of KDM4C is dependent on its double tudor domain\textsuperscript{41}. Altogether the data suggests a model in which chromatin-association of KDM4 members is mediated by both their divergent and common reader binding targets to ensure correct epigenetic programming during development.

Germ cell chromatin is subject to epigenetic regulation\textsuperscript{42,43}. Here our KDM4 team proposes that H3K23me3 and H3K36me3 might be involved in mammalian germ cell development, especially during the mitosis-to-meiosis transition, through crosstalk between the KDM4 reader and catalytic domains. Taverna lab revealed H3K23me3
enrichment in mouse and rat testes, specifically in primary spermatocytes undergoing meiosis. This provides new evidence about H3K23me3 being a conserved histone modification associated with meiotic chromatin, as previously identified in Tetrahymena and C. elegans. The enrichment of H3K23me3 is also consistent with reduced immuno-staining of H3K23ac during preleptotene to pachytene stage in mouse testes tissues. The function and mechanism of this unique association with meiotic chromatin still needs further investigation. It remains to be experimentally tested if H3K23me3 plays similar roles in meiotic DNA damage in both mammalian systems and Tetrahymena.

Previously, the KDM4 family and H3K36me3 were independently implicated in germ cell development, however the underlying relationship was not clear. Based on our detailed structural and biochemical analysis, I would postulate that the reader domain of KDM4B has evolved to discriminate for exclusive binding to H3K23me3 with possible regulation by PTMs on H3T22 and H3R26. The H3K23me3 binding by KDM4A and KDM4B is likely to be conserved across species, given the conservation of the H3K23me3-discriminating residue (N931 in human KDM4A and N951 in human KDM4B) as shown in Appendix Figure A-10. The functional analyses reported here demonstrate that the distinct mode of H3K23me3 binding stimulates the demethyalse activity of KDM4B toward H3K36me3 on the same histone tail. These results are in accord with such “cross-talk” between a histone methylation “reader” and “eraser”,

leading to cis-histone demethylation reported for PHF8, KDM7A\textsuperscript{46,47}, KDM4A and KDM4C\textsuperscript{48}.

Interestingly, Denu lab found H3K23me3 also overlaps significantly with H3K27me3 by immunofluorescence in rat spermatocytes. This finding is consistent with previous mass spectrometry results showing H3K23me3 physically co-exists on a subset of H3K27me3 modified histones in Tetrahymena and C. elegans, suggesting that similar positive “crosstalk” between H3K23me3/H3K27me3 occurs in mammals as well. Furthermore, KDM4B binding to H3K23me3 was not disrupted in the presence of H3K27me3. Thus, the combined heterochromatic H3K23me3/H3K27me3 signature may be used to recruit KDM4B to remove H3K36me3 at heterochromatin during the initial stages of meiosis. Indeed, H3K27me3 and H3K36me3 are located in mutually exclusive regions of germline chromatin in C. elegans. Whether H3K23me3-dependent down-regulation of H3K36me3 serves to reduce gene expression or increase meiotic DNA damage repair at heterochromatin needs further investigation. Collectively, these observations suggest enrichment of H3K23me3 during meiosis is conserved across multiple species, and our molecular characterization provides a mechanistic basis for the future investigation of the relationship between KDM4B and H3K23me3 in the germline.
Appendix Figure A-10. Sequence alignment of double tudor domains from KDM4 members across multiple species

Sequence alignment is generated by Jalview\textsuperscript{49,50}. (bt = Bos Taurus, ce = Caenorhabditis elegans, cl = Canis lupus familiaris, dr = Danir rerio, gg = Gallus gallus, hs = Homo sapiens, mm = Mus musculus, pt = Pan troglodytes, rn = Rattus norvegicus, xt = Xenopus tropicalis)
References


Appendix B

Structural Dynamics: Cupin protein and calicheamicin biosynthesis enzymes CalE6, CalU16 and CalU19

Natural products and their derivatives continue to play an important role in drug discovery. As described in Chapter 3.4, ensemble refinement as a protein dynamics approach, can extract important information from high-resolution X-ray structures. This appendix chapter includes the structural dynamics and bioinformatics studies during my Ph.D. thesis on several very important natural product biosynthesis enzymes: Cupin from *Photorhabdus luminescens*, and enzymes CalE6, CalU16 and CalU19 from calicheamicin biosynthesis pathway. Publications corresponding to different sections will be footnoted separately.
B.1. Protein dynamics of Cupin protein from *Photorhabdus luminescens*\(^{1}\)

The cupin protein superfamily was established to distinguish proteins by their distinctive β-barrel fold yet they possess diverse functionality\(^2,3\). To date, both catalytic and noncatalytic cupin proteins have been identified, such as sugar binding effector proteins, dioxygenases, and decarboxylases. Cupin proteins are identified by their overall structure and through the presence of two conserved motifs \(G(X)_5HXXH(X)_3E(X)_6G\) and \(G(X)_3PXG(X)_2H(X)_3N\), where the letters in bold represent the residues that can be responsible for metal binding. In some cases, one of the two histidinyl residues in motif 1 can be replaced by glutaminyl, aspartyl, or glutamyl residues. Currently, the range of metals known to be coordinated by these four residues are cadmium, cobalt, copper, iron, manganese, nickel, and zinc\(^2,3\). The function of a cupin protein is frequently dictated by the bound metal. For example, cupin dioxygenases bind to iron or nickel, while the cupin decarboxylases are manganese dependent. Interestingly, there are cases where these residues do not bind metal, rather they are involved in positioning the substrate in the active site as observed for some sugar isomerasers\(^2,3\).

---

Phillips and Thomas lab have been interested in a group of putative cupin proteins that have been identified using a bioinformatics approach to natural product discovery. The genes coding for these cupin homologs are always associated with two other genes that code for a methionyl-tRNA synthetase (MetRS) homolog and a lysine/ornithine N-hydroxylase homolog in more than 40 bacterial species. Interestingly, in some species the cupin is fused to the N-terminus of the MetRS homolog, suggesting a coordinated function between these enzymes. These gene clusters have been noticed by others\textsuperscript{4,5}, but neither a metabolite nor a function has been associated with them yet. To aid in deciphering the function of these gene clusters in these various bacterial species, Phillips and Thomas lab took a structural genomics approach. A crystal screen using several cupin proteins encoded by the targeted gene clusters identified conditions for the crystallization of plu4264, a 14.5 kDa protein of unknown function from \textit{Photorhabdus luminiscens} TTO1. Protein purification, crystallization, and structure elucidation were performed by the Midwest Center for Structural Genomics. The crystal structure of plu4264 was determined using SAD phasing to a resolution of 1.35 Å as a homodimer (Appendix Figure B-1a). Sequence alignments and secondary structure prediction suggest that the amphipathic helices are present in the majority of the putative cupins the Thomas lab identified, suggesting that the helices may be important to the function of these proteins. Finally, there is a large hydrophobic interface between the β-barrels of each monomer, forming a surface area of \(\sim 1900 \, \text{Å}^2\) (calculated using PISA\textsuperscript{6}) that creates the main interactions between the two monomers.
Appendix Figure B-1. Structure of protein cupin plu4264

(a) Cartoon representation of plu4264. (b) Cartoon representation of chain A Ni$^{2+}$ coordination site. (c) Loop flexibility of plu4264: all conformations of the loops are represented as separate lines. (d) Real-space correlation coefficient and B-factor plot of plu4264 regular structure.
Appendix Figure B-2. Real-space correlation coefficient and B-factor plot of cupin plu4264 homologs

(a) The same three loop regions in plu4264 homologs were detected by DALI alignment and highlighted in the figure with arrows. Depending on whether flexibility was observed, arrows are colored by black (Flexible) or grey (Not flexible). In all plots the red traces are B-factors and the blue traces are RSCC. The x-axis notes the amino acid number, the left y-axis is the RSCC value and the right y-axis is the B-factor.
Plu4264 has the two characteristic motifs of cupin proteins that include the residues that compose the nickel-binding site (Appendix Figure B-1b). Interestingly, several flexible loop regions (residues 46–55, 68–75, and 84–92) close to the metal were observed from analysis of the backbone residues B-factors and the real-space correlation coefficient (Appendix Figure B-1d). To further investigate the loop flexibility, I ran a Phenix ensemble refinement on the structure (Appendix Figure B-1c). The ensemble refinement results validate the proposed loop flexibility, reduce the R/R_free, and increase the overall map correlation coefficient (F-model v.s. 2mFo-dFc). Details of the ensemble refinement were included in the Chapter 2, and the corresponding coordinates were deposited in the Protein Data Bank 4Q29. Several DALI-based close structural homologs of cupin were also analyzed on those loop regions: they all have some similar loop flexibilities, but not necessarily in all the three regions (Appendix Figure B-2). In brief, taking into consideration of other results from the Thomas lab, the function of Cupin plu4264 could be akin to that of Tm1287 but with a larger, less charged substrate.
B.2. Structural dynamics of a methionine γ-lyase for calicheamicin biosynthesis: rotation of the conserved tyrosine stacking with pyridoxal phosphate*

CalE6 from *Micromonospora echinospora* is a (pyridoxal 5’ phosphate) PLP-dependent methionine γ-lyase involved in the biosynthesis of calicheamicins. This chapter reports the crystal structure of a CalE6-MES complex showing ligand-induced rotation of Tyr100, which stacks with PLP, resembling the corresponding tyrosine rotation of true catalytic intermediates of CalE6 homologs. Elastic network modeling and crystallographic ensemble refinement reveals mobility of the N-terminal loop, which involves both tetrameric assembly and PLP binding. Modeling and comparative structural analysis of PLP-dependent enzymes involved in Cys/Met metabolism shine light on the functional implications of the intrinsic dynamic properties of CalE6 in catalysis and holoenzyme maturation.

*Portions of Appendix B.2 have been submitted to Structural Dynamics as: Cao H, Tan K, Wang F, Bigelow L, Yennamalli RM, Jedrzejczak R, Babnigg G, Bingman CA, Joachimiak A, Kharel MK, Singh S, Thorson JS, Phillips Jr GN. Structural dynamics of a methionine γ-lyase for calicheamicin biosynthesis: rotation of the conserved tyrosine stacking with pyridoxal phosphate. I only contribute to the structural dynamics study part, and the paper was not published by the time my thesis is written.*
Structural dynamics on various time and length scales are inherent properties of biological macromolecules and are often related to their functions. For example, enzymes, which are chosen by nature to lower the energy barrier of transition state conformations ultimately require finite structural flexibility to advance the catalytic cycle and avoid being trapped in either substrate bound or product bound states. Protein dynamics can be probed not only by solution approaches like nucleic magnetic resonance (NMR) and other spectroscopies, but also by combination of X-ray crystallography and computational modeling. Recent advances in time-resolved serial femtosecond crystallography (TR-SFX) using pulsed ultra-bright X-ray free electron lasers (XFEL) show promise in capturing photoactive proteins in action at atomic level by taking data within a very short time courses (on the order of 100 fs) during which light-induced conformational changes have been previously initiated. Among many computational structural biology methods, molecular dynamics (MD) simulation, elastic network models (ENM) and ensemble refinement offer ways complementary to experiments to identify motions that occur at a wide range of temporal and spatial scale relevant to biomolecule function. Unlike MD simulation, which yields a time course of conformational trajectory, ENM performs low-frequency normal mode analysis (NMA) based on an harmonic approximation around the starting reference model, which is assumed as a minimum energy conformation. Ensemble refinement describes protein dynamics by fitting ensemble models to diffraction data which accounts for both anisotropic and anharmonic distributions. I have applied ENM and ensemble refinement approaches to extract the dynamic information from the experimental structure of
CalE6, a PLP-dependent methionine γ-lyase from *Micromonospora echinospora*, which is encoded by the calicheamicin biosynthetic gene cluster^{21,22}.

Calicheamicin γ1 is a prototype of the 10-membered enediyne family of antibiotics that contains a characteristic bicyclo tridecadiynene core^{23–27}. Like all 9- and 10-membered enediyynes, calicheamicin-induced oxidative DNA strand scission is enabled by cycloaromatization of the enediyne core to form a highly reactive diradical species^{27}. In calicheamicin, this cycloaromatization event is initiated via reductive activation of a unique allylic trisulfide ‘trigger’ and the corresponding reactive diradical is exquisitely positioned by the calicheamicin aryltetrasaccharide, which contains multiple distinctly functionalized sugars that contribute to DNA minor groove recognition and affinity. Given its incredible potency, calicheamicin also served as the warhead of the very first clinically approved monoclonal antibody (mAb) drug conjugate for targeted cancer therapies^{28–32}. To date, the putative enzymes responsible for sulfur mobilization and installation during the formation of enediyne trisulfide and thiosugar remain unknown. Protein BLAST-based functional annotation indicates that the calicheamicin biosynthetic cluster contains at least three candidate genes that may encode for enzymes involved in the requisite sulfur biochemistry [calE4 (putative cysteine desulfurase), calS4 (putative selenocysteine lyase/cysteine desulfurase) and calE6 (methionine γ-lyase)]^{33}. It cannot be ruled out that participation of sulfur-transfer genes from primary-metabolism as reported for the biosynthesis of 2-thiosugar-containing natural product also occurs during...
biosynthesis of the 4-thiosugar moiety of calicheamicin. Among the putative genes responsible for sulfur transfer during calicheamicin biosynthesis, the corresponding gene product CalE6 was recently confirmed to show methionine γ lyase activity\(^{21}\). However, the contribution of this enzymatic activity to calicheamicin thiosugar and/or trisulfide formation remains unknown. CalE6 structure was first determined in the MES bound form by our group as part of the Protein Structure Initiative (PDB code: 4Q31), which showed the same homo-tetrameric overall structure.

I next examined dynamic properties of the CalE6 through two orthogonal approaches, normal mode analysis and crystallographic ensemble refinement (Appendix Table B-1). The methods revealed similar mobility distribution as analyzed with Mobi Server\(^{34}\) based on scaled RMSD of atomic position in the ensemble models and with B-factor based mobility distributions of the single model crystal structure (Appendix Figure B-3a). The N-terminal extended loop for tetrameric assembly and PLP phosphate binding of CalE6 consistently showed relatively high mobility from all the analyses above, which agrees well with the conformational change of the corresponding N-terminal region from disordered to structured upon PLP recruitment into human cystathionine γ-lyase (PDB entries 2NMP and 3ELP, 41% sequence identity with CalE6)\(^{35}\). A brief survey of crystallographic B-factor distribution of homologs of CalE6 available in PDB (sequence identity 27-45%) suggests nontrivial conservation of intrinsic mobility of the same N-terminal region that links PLP binding to tetramer assembly, which is consistent with experimental
observation by Sun et al. that apo-hCSE exists as a weaker tetramer compared with
PLP-hCSE complex in solution. Elastic network model analysis suggests a similar
mobility distribution of the slowest normal mode of CalE6 to the combined ensemble
of all top 5 slow modes, both showing additional spatially clustered mobile elements
near the C-terminus that are absent in the crystallographic ensemble models
(Appendix Figure B-3a-c). This discrepancy between local mobility from normal
mode analysis and crystallographic modeling can likely be explained by crystal
packing effect where the otherwise mobile elements aforementioned were restricted
by interactions between neighbor molecules in the crystal lattice. While normal mode
analysis simulates overall protein mobility at the backbone level, experimentally
based ensemble refinement proves to be useful to extract both major and minor
conformations at the side chain level. Ensemble refinement was able to identify the
different rotamers of Tyr 100 of CalE6 covering the conformational space as observed
in the static crystal structures of different ligand complexes (Appendix Figure B-4). As
expected, the pyridine ring of PLP is restricted to a limited conformational space by
the internal aldimine linkage with Lys 197 that is distinct from the conformer
observed for external aldimine intermediates (Appendix Figure B-4). This, in essence,
reflects the presence of a minor population of alternative conformers per residue in
100K cryogenic crystal conditions. In agreement, there is clear variation in side-chain
mobility for different residues in the vicinity of PLP cofactor and the active site, like
the flexible Tyr 100 vs. barely mobile Arg 48 (Appendix Figure B-4).
Appendix Figure B-3. Dynamics modeling and analysis of CalE6

(a) Combination of all top 5 low frequency normal modes simulated by ElNemo server. (b) The top slowest normal mode from ElNemo simulation analyzed by Mobi. (c) Crystallographic ensemble refinement models analyzed by Mobi (PDB entry 4XQ2). (d) Single model crystal structure shown as B-factor putty using PyMOL preset.
Appendix Table B-1. Refinement statistics of CalE6

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Single Model Refinement</th>
<th>Ensemble Refinement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Data Bank ID code</td>
<td>4Q31</td>
<td>4XQ2</td>
</tr>
<tr>
<td>Spacegroup</td>
<td>1222</td>
<td></td>
</tr>
<tr>
<td>Cell dimensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a$, $b$, $c$ (Å)</td>
<td>146.9, 147.0, 349.9</td>
<td></td>
</tr>
<tr>
<td>$a$, $\beta$, $\gamma$ (°)</td>
<td>90.0, 90.0, 90.0</td>
<td></td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.97915</td>
<td></td>
</tr>
<tr>
<td>Resolution of data collection (Å)</td>
<td>34.0 – 2.10 (2.14-2.10)</td>
<td></td>
</tr>
<tr>
<td>No. of unique reflections</td>
<td>218967 (21668)</td>
<td></td>
</tr>
<tr>
<td>Completeness % (Å)</td>
<td>99.9 (99.0)</td>
<td></td>
</tr>
<tr>
<td>Redundancy</td>
<td>7.4 (7.3)</td>
<td></td>
</tr>
<tr>
<td>$R_{\text{sym}}$</td>
<td>0.14 (0.72)</td>
<td></td>
</tr>
<tr>
<td>$I/X^b$</td>
<td>16.6 (2.9)</td>
<td></td>
</tr>
<tr>
<td>Resolution range in refinement (Å)</td>
<td>34.0 – 2.10 (2.12-2.10)</td>
<td></td>
</tr>
<tr>
<td>No. of unique reflections (work/test)</td>
<td>218929/10997</td>
<td></td>
</tr>
<tr>
<td>$R_{\text{cryst}}$</td>
<td>15.3 (19.5)</td>
<td>14.5 (17.7)</td>
</tr>
<tr>
<td>$R_{\text{free}}$</td>
<td>19.1 (26.2)</td>
<td>18.3 (24.4)</td>
</tr>
<tr>
<td>Mean coordinate error $^d$ (Å)</td>
<td>0.19</td>
<td>0.14</td>
</tr>
<tr>
<td>Rmsd bond length (Å)</td>
<td>0.007</td>
<td>0.009</td>
</tr>
<tr>
<td>Rmsd bond angles (°)</td>
<td>1.07</td>
<td>1.34</td>
</tr>
<tr>
<td>Average B value (Å$^2$)</td>
<td>29.7/29.1/33.5/48.3</td>
<td>25.3/24.8/24.7/71.6</td>
</tr>
<tr>
<td>(overall/protein/waters/ligand)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of non-hydrogen atoms</td>
<td>24719</td>
<td>243453</td>
</tr>
<tr>
<td>No. of protein atoms</td>
<td>22606</td>
<td>226060</td>
</tr>
<tr>
<td>No. of waters</td>
<td>1823</td>
<td>14543</td>
</tr>
<tr>
<td>No. of ligands and sugars</td>
<td>8 MES, 28 glycerol,</td>
<td>80 MES, 280 glycerol,</td>
</tr>
<tr>
<td></td>
<td>5 formic acid, 12 Cl$^-$</td>
<td>30 formic acid, 120 Cl$^-$</td>
</tr>
<tr>
<td>Ramachandran Statistics $^f$ (%)</td>
<td>97, 2.5, 0.5</td>
<td>92, 6.3, 1.7</td>
</tr>
</tbody>
</table>

Values in parenthesis are for the highest resolution shell.

$^aR_{\text{sym}} = \frac{\sum_{hkl} |I(hkl)| - \langle |I(hkl)| \rangle}{\sum_{hkl} |I(hkl)|}$, where $I(hkl)$ is the intensity of an individual measurement of the symmetry related reflection and $\langle |I(hkl)| \rangle$ is the mean intensity of the symmetry related reflections.

$^bI/X$ is defined as the ratio of averaged value of the intensity to its standard deviation.

$^cR_{\text{cryst}} = \frac{\sum_{hkl} |F_{\text{obs}}| - |F_{\text{calc}}|}{\sum_{hkl} |F_{\text{obs}}|}$, where $F_{\text{obs}}$ and $F_{\text{calc}}$ are the observed and calculated structure-factor amplitudes.

$^dR_{\text{free}}$ was calculated as $R_{\text{cryst}}$ using randomly selected small fractions (5%) of the unique reflections that were omitted from the structure refinement.

$^e$Mean coordinate error was calculated based on maximum likelihood.

$^f$Ramachandran statistics indicate the percentage of residues in the most favored, additionally allowed and outlier regions of the Ramachandran diagram as defined by MolProbity$^{as}$. 
Appendix Figure B-4. Analysis of ensemble refinement models with reference to two different conformations of CalE6

Structures were aligned using PyMOL based on the tetrameric form (chains A-D). All four active site residues were shown, with Lys 197-PLP and Tyr 100 highlighted as sticks and the other side chains as lines. Structures were differently colored based on carbon atoms, with ensemble models (teal), MES bound form (cyan), and sulfate bound form (magenta). Non-covalent ligands were omitted and backbone traces were shown as transparent ribbon for clarity.
Conan and I demonstrated through comparative structural analysis of CalE6 and its homologs that the conserved tyrosine residue stacking with PLP is subject to ligand-induced rotation. The same type of concerted motion of PLP and tyrosine pair of PLP-dependent enzymes involved in Cys/Met metabolism is found in multiple external aldimine intermediate structures with respect to their native states, emphasizing a possible dynamic role of this functional side chain during catalysis. The overall dynamics of CalE6, revealed consistently through elastic network model analysis and ensemble refinement, resembles the conformational flexibility of human cystathionine γ-lyase required for holoenzyme maturation, which involves both structuring of the otherwise disordered N-terminal loop responsible for PLP phosphate binding and the swing motion of the active site loop to bring the conserved tyrosine closer to PLP for stacking with the pyridine ring. In summary, the intrinsic dynamics of Tyr 100 of CalE6 can be possibly understood in the context of PLP enzyme function at two levels - catalytic role as an indirect electrostatic modulator or direct proton donor; structural role contributing to cofactor affinity during holoenzyme maturation. Future work of direct observation of these transient dynamic events via pulse-triggered time-resolved experiments can further elucidate the sequence and time scale of individual steps and their functional relevance.

Calicheamicin γ1 (CAL) is an enediyne antitumor compound produced by Micromonospora echinospora. Despite extensive analysis and biochemical study, several genes in the biosynthetic gene cluster of CAL remain functionally unassigned. Using a structural genomics approach and biochemical characterization, two proteins encoded by genes from the CAL biosynthetic gene cluster assigned as “unknowns”, CalU16 and CalU19, were characterized. Structure analysis revealed that they possess the enediyne self-sacrifice mechanism like protein CalC. Subsequent study by Thorson lab revealed CalU16 and CalU19 to confer resistance to CAL, and reminiscent of the prototype CalC, both CalU16 and CalU19 were cleaved by CAL in vitro. My role in this project is only providing structural analysis and CalU16 and CalU19 from multiple approaches and double-checking the structural annotation of CalU16. So this chapter will be described very briefly.

The calicheamicins are a prototype of the naturally occurring 10-membered enediyne antitumor antibiotic family and were first reported in 1987 as metabolites of *Micromonospora echinospora*. Members of this family share a structurally conserved enediyne core, also often referred to as the “warhead” as this structural unit is central to the fundamental enediyne mechanism of action\(^4^4-^4^8\). In all family members, the enediyne core is strategically decorated with a bioorthogonal “triggering” system and specific appendages that enhance affinity to the metabolite’s target (DNA/RNA). Calicheamicin’s high affinity for the minor groove of DNA insures that the diradical species is quenched via hydrogen abstraction from the backbone of opposing strands of dsDNA to form DNA radicals that, in the presence of oxygen, result in facile double-strand scission\(^4^3-^4^8\). The gene cluster encoding for calicheamicin biosynthesis was cloned from *M. echinospora* and sequenced nearly a decade ago, there remain a number of genes (~30%) within this locus annotated as “unknowns” due to a lack of homologues and/or biochemical characterization of corresponding gene products\(^2^2\).

In this project, Thorson and Phillips lab describe the application of structural genomics as a basis for the functional characterization of two proteins encoded by such “unknowns” - CalU16 and CalU19. Specifically, structure elucidation (via both NMR and X-ray crystallography from structural genomic centers) revealed CalU16 to be a structural homologue of CalC, a protein previously characterized as among the first reported enediyne “self-sacrifice” resistance proteins (Appendix Figure B-5)\(^4^9\). Prompted by this structure-based revelation, subsequent biochemical
characterization of CalU16 and its homologue CalU19 revealed both to serve in a similar capacity wherein CalU19 also displayed the unprecedented ability to trigger enediyne cycloaromatization in the absence of endogenous reducing agents.

While there exist many natural product biosynthetic loci that also encode for more than one resistance mechanism for the encoded natural product\textsuperscript{50,51}, the encoded redundancy highlighted by the current study is uncommon and may suggest discrete self-sacrifice proteins to contribute to subtle distinctions in their localization and/or function. For example, the predicted isoelectric point of CalC (10.16) is dramatically different from CalU16 (4.24) or CalU19 (4.69), and this is consistent with CalC's demonstrated ability to bind DNA (the target of enediyne) under physiological pH\textsuperscript{52}. Under identical conditions, CalU16/U19 are predicted to possess an overall negative charge, which may contribute to distinct intracellular localization and/or unique protein–protein interactions possibly including those proteins/enzymes involved in CAL biosynthesis. Cumulatively, this project extends the body of work focused upon understanding how bacteria construct and control highly reactive and lethal metabolites and expands the number of known “self-sacrifice” enediyne resistance proteins (Appendix Figure B-5).
Appendix Figure B-5. Structural comparison between CalU16 and CalC

Structures representations of CalC (a) and CalU16 (b) are shown. CalC was shown as DNA-binding model, X-ray structure, and CAL docking model. CalU16 was shown as b-factor plot, X-ray structure, and CAL docking model.

(a) CalC

(b) CalU16
References


25. Nicolaou, K. C. The battle of calicheamicin γI1


