Structure-Based Design of Mutant Proteins:

I. Molecular Docking Studies of Amino Acid Binding to Wild-Type Aminoacyl-tRNA Synthetases

II. Structure-Based Design of Mutant Aminoacyl-tRNA Synthetases for Non-Natural Amino Acid Incorporation

Thesis by

Deqiang Zhang

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2003

(Submitted December 3, 2002)

Acknowledgements

My years at Caltech have been full of excitement and adventures. I am so fortunate to get to know many of the brightest people in the world here in the last several years and to learn from them. Although sometimes I wish I had taken a different course in my life here and there, I definitely would not have been who I am today without the help of many people.

First of all, I would like to thank my advisor, Prof. William A. Goddard III. I am very grateful to him for giving me the opportunity to finish my Ph.D. in his group. Bill belongs to the group of the smartest scientists and I am so amazed about his insight to almost everything across many different fields. I learned a great deal from him on how to be a good scientist.

I would also like to thank Dr. Nagarajan Vaidehi for being a mentor to me. When I first joined the group, it was Vaidehi who taught me to run molecular dynamics for the first time. I got a quick start with her help. I cannot imagine how I could have done it without her.

I thank Prof. Sunney I. Chan for supporting me for many years in his group. Sunney is a very nice person and he has always been a good advisor to all students, whether they are his own students or not. I will always remember how he helped so many of us on all sorts of problems.

I thank all the other members of my committee. I thank Prof. Richard Roberts for his advice, and I have learned a lot about critical thinking in scientific writing from him. Prof. Aron Kuppermann taught me Quantum Chemistry and his teaching style has earned a lot of respect from his students including me. Prof. Jesse (Jack) Beauchamp has been very nice to me since I just came to Caltech. I wish I had listened more carefully to him in those conversations.

My life at Caltech would not have been so fruitful without many talented colleagues and friends I have met here. Dr. Wely Floriano has taught me many useful docking techniques. Dr. Jeremy Kua helped me a lot to start my first docking study. Dr. Patrick Wintrode is a nice person to collaborate with as we did a molecular dynamics study on a family of enzymes with different thermostability. Deepshikha Datta worked together with me on methionyl-tRNA synthetase study. My other collaborators include Pin Wang of the Tirrell group at Caltech and Dr. Lei Wang of the Schultz group at Scripps. My office mate, Spencer Hall, was very helpful in many discussions we had. Jiyoung Heo showed me how to get started with Amber. I had some helpful discussions with John Wendel. Other members on the bio group, Rene Trabanino, Pete Huskey, Victor Kam, Caglar Tanrikulu, also helped me in various situations.

Thanks also go to other members in the group. Drs. Mario Blanco, Tahir Cagin, Rick Muller all offered me help on various issues. Dr. Shiang-Tai Lin helped me in the discussion of density of states calculation. I thank Darryl Willick for helping me out with computer problems.

There are other people who have helped me greatly in the last several years. Graduate secretary Dian Buchness is the first person I got to know when I first arrived here, and her help is greatly appreciated. Priscilla Boon, who is the secretary to Profs. Sunney Chan and Jack Beauchamp, has provided much help to me. Shirley Wu also deserves many thanks from me for her help.

My life here in the last six years wouldn't have been this memorable without the friendship that I have shared with my good friends, Dr. Suzie Hwang Pun, Dr. Winston Pun, Dr. Sidao (Stone) Ni, Dongdong (Dorothy) Niu, Dr. Xindong Liu, Ning Liu, Jiehui Wang and Dr. Wen Li. We have had so much great time together and so many good memories. I will always cherish our invaluable friendship. I would also like to thank the friends I have met at FIS (Fellowship of International Students) at Lake Avenue Church, especially Perlita Lim for all her caring and kindness.

I have reserved my final thanks for my families. My parents have given me all their love over the years. My sister Guifeng has supported me going through many of the decisions in my early life. My parents-in-law have loved me like their own son. My three-year-old son, George, who was my 30th birthday gift himself, has given me a lot of joy being a parent. Finally, my deepest thanks go to my wife, Dr. Lintong Li, for being supportive to me, especially the last several months when I was not around.

Abstract

Protein biosynthesis has precisely controlled accuracy, and aminoacyl-tRNA synthetases (AARSs) play an important role in charging amino acids to their cognate tRNAs with high fidelity. In some cases the misactivation of non-natural amino acids by the wild-type or mutant AARS can be utilized to incorporate these non-natural amino acids into proteins *in vivo*. Such technique has tremendous potentials in protein engineering and other applications. Therefore, it is essential to understand the amino acid recognition mechanism displayed by AARSs.

In this thesis, computational studies of the selection of natural and non-natural amino acids by AARSs at the binding stage have been conducted for methionyl-tRNA synthetase (Chapter 2), histidyl-tRNA synthetases (Chapter 3), and isoleucyl-tRNA synthetase (Chapter 4). In these chapters, molecular docking and ligand perturbation are used to elucidate the binding discrimination showed by these AARSs.

Because many non-natural amino acids carrying interesting physical and chemical properties on their side chains cannot be incorporated by using the wild-type AARSs, it is necessary to manipulate the activity of AARSs by making mutations in the binding site of amino acids. To this end, we have developed a Clash Opportunity Progressive (COP) protein design tool to redesign the binding site of AARSs. Chapter 5 describes the main steps in COP. Chapters 6 to 8 present the application of COP to different AARSs. In Chapter 6, COP has been applied to design mutant tyrosyl-tRNA synthetase (TyrRS) for

recognizing Ome-Tyr, Naph-Ala, and p-keto-Tyr. In Chapter 7, COP has been used to design mutant phenylalanyl-tRNA synthetase for p-keto-Phe. In Chapter 8, tryptophanyl-tRNA synthetase is used as a template to design mutant AARS to recognize NBD-Ala, bpy-Ala, and DAN-Ala.

The appendices are some publications and manuscripts on various other projects. Appendix I is a molecular dynamics study of laboratory-evolved pNBE enzymes with different thermostability. The findings presented here will help us to better understand the determinants in protein stability evolution. Appendix II contains experimental work I have done in the Chan group. Unfolding experiments revealed the existence of intermediates in the equilibration unfolding of RdPf. In Appendix III, femtosecond time-resolved spectroscopy was used to study the fluorescence resonance energy transfer and tryptophan solvation dynamics in RdPf.

Table of Contents

List of Figures

List of Tables

Chapter 8.

Appendix I.

Appendix II.

Appendix III.

Chapter 1

Introduction

1. The New Era of Computational Biochemistry

In the last twenty-five years the advances in the field of computational biochemistry have contributed tremendously to our understanding of complex biomolecular systems such as proteins, nucleic acids and bilayer membranes. The very first molecular dynamics simulation of a protein, the bovine pancreatic trypsin inhibitor (BPTI), was published in 1977 (1). Although the simulation is "crude" by today's standards, it was important because it introduced an important conceptual change in our view of biomolecules. The classic view of the structure of proteins and nucleic acids is static, because the protein crystal structures available at that time led to an image of "rigid" biomolecules with every atom fixed in place. It is now recognized that the atoms of which the biopolymers are composed are in a state of constant motion at ordinary temperatures. The X-ray crystal structure of a protein is merely the average atomic positions, and the atoms exhibit fluid-like motions about these averages. This work marked the beginning of modern computational biochemistry, and numerous methodological advances in computational studies of biomolecules have followed since.

Molecular dynamics (MD) is the key in the advance of computational biochemistry. There are other simulation methods, such as Langevin dynamics, Monte Carlo simulation and normal mode analysis. New techniques are being developed that treat the bulk of a biomolecule classically while applying quantum mechanics to a subset

of atoms, typically in the active site. In molecular dynamics, the motion of the system was described by Newtonian equations, and the trajectory is obtained by integrating a series of such equations.

We all know that classical mechanics agrees well with quantum mechanics when $\Delta E << k_B T$. To understand why molecular dynamics can be applied to motions like bond stretches, which has quantized energy gap much higher than the thermal energy at room temperature, we can look at an example of the motion of an O–H bond. The vibrational frequency of an O–H bond is about 100 ps^{-1}. It represents one of the highest frequency modes of vibration in a biomolecule and thus serves as a worst-case scenario for classical approximation in macromolecular simulations. One of the physical quantities of great interest is the variance in the position of atoms at equilibrium, $<(\Delta x)^2>$. An oscillator model is usually used to describe the bond stretch motion. Assume the equilibrium position is at $x = 0$, then $<(\Delta x)^2> = <x^2>$. This mean-square fluctuation about the average position is related to the B factors of crystallography and is also measurable by neutron scattering (2) and by Mössbauer spectroscopy (3). It is also one of the most important quantities in molecular dynamics simulations. For a harmonic oscillator,

$$V = \frac{1}{2} k x^2, \tag{1}$$

where V is the potential energy, k is spring constant. Considering the equal partition of energy between kinetic energy and potential energy, we can get

$$\left\langle x^2 \right\rangle = \frac{E}{k} = \frac{E}{m(2\pi f)^2}, \tag{2}$$

where E is the total energy and f is the vibrational frequency of the harmonic oscillator. Plug in the energy expressions of the harmonic oscillator from classical and quantum mechanics, assuming $f = 100\ ps^{-1}$, $T = 300\ K$ and m is the mass of a proton, we get $\langle x^2 \rangle = 5\times10^{-3}\ Å^2$ from quantum mechanics and $6\times10^{-4}\ Å^2$ from classical mechanics. The root-mean-square deviation (rmsd) is only 0.07 Å, which is modest relative to crystallographic resolutions and the equilibrium length of the O–H bond. Furthermore, when compared to motional amplitudes measured by neutron scattering, classical simulations predict too much motion (4). Thus, the reduced motion resulting from the neglect of quantum effects is overshadowed by other approximations made in simulations, such as the neglect of electronic polarizability and the assumed pairwise additivity of van der Waals forces. The overestimate of protein motion by simulations is not yet understood. To summarize, classical simulations are unable to analyze the details of bond stretching and angle bending correctly. These motions are at frequencies too high for an accurate treatment using Newton's law. However, we have observed that the errors in motional amplitude are relatively small, and errors in energy tend to cancel out in appropriately designed calculations, as when $\Delta\Delta G$'s are calculated rather than ΔG's.

It is worth mentioning that recent advances in techniques that combine quantum mechanics and classical molecular mechanics (QM/MM) now allow for an accurate and detailed understanding of processes involving bond breaking and bond making, and how enzymes catalyze those reactions. In QM/MM approach the system is partitioned into a QM region and an MM region. The QM region typically includes the substrate and the side chains of residues believed to be involved in the reaction and any cofactors. The

remainder of the protein and the solvents is included in the MM region. The applications of QM/MM method include the reactions catalyzed by triosephosphate isomerase (5, 6), bovine protein tyrosine phosphate (7) and citrate synthetase (8).

To simulate the dynamics of a macromolecule, we generally need to specify three components: The topology of the molecule (also known as the connection records), the initial coordinates or the starting structure, and a force field. There are currently more than ten force fields in use for biomolecule simulations, such as CHARMM (9), AMBER (10), OPLS (11), MMFF (12) and DREIDING (13). Some test studies on these force fields showed that they perform comparably well on proteins (14). These force fields were generally optimized using source data for small molecules from QM, electron diffraction, microware, IR and NMR spectroscopy, etc. Because of the transferability of parameters, they perform equally well on biomolecules.

Clearly computational and theoretical studies of biological molecules have advanced significantly in recent years and will progress rapidly in the future. These advances have been fueled by the ever-increasing number of available structures of proteins, nucleic acids, and carbohydrates. Computational biochemistry has many new applications in various areas. Among these new applications are molecular docking, 3-D protein structure prediction, and protein design. In this thesis, all these applications will be applied in various chapters.

2. Protein Engineering and Aminoacyl-tRNA Synthetases

Biomolecules are polymers in principle. Although many advances in synthetic polymer chemistry have been made over the last several decades to provide the polymer chemist with increasing control over the structure of macromolecules (15-18), none has provided the level of control afforded by *in vivo* methods, a level of control which is the basis of exquisite catalytic, informational, and signal transduction capabilities of proteins and nucleic acids. The Tirrell laboratories at Caltech and others have been exploring the use of *in vivo* methods for producing artificial protein polymers whose sequence, stereochemistry, and molecular weight are exactly controlled. Harnessing the control provided by *in vivo* methods in the synthesis of protein polymers should permit control of folding, functional group placement, and self-assembly at the angstrom length scale. Indeed, proteins produced by this method exhibit well-controlled chain-folded lamellar architectures (19, 20), unique smectic liquid-crystalline structures with precise layer spacings (21), and well-controlled, reversible gelation (22). The demonstrated ability of these protein polymers to form unique macromolecular architectures will be of certain importance in expanding the role of proteins as materials with interesting liquid-crystalline, crystalline, surface, electronic, and optical properties. An important continuing objective, therefore, is to expand the chemical and physical properties that can be engineered into protein polymers *in vivo*, via the incorporation of non-natural amino acids.

The *in vivo* incorporation of non-natural amino acids is controlled in large measure by the aminoacyl-tRNA synthetases (AARSs), the class of enzymes that safeguards the fidelity of amino acid incorporation into proteins. Translation is the process whereby genetic information, in the form of mRNA, is used to synthesize the corresponding sequence of amino acids found in proteins. The identity of an amino acid inserted at a particular position during protein synthesis is determined by the pairing of a codon in mRNA with a particular aminoacyl-tRNA (Figure 1). The overall fidelity of protein synthesis is dependent on the accuracy of two processes, condon-anticodon recognition and aminoacyl-tRNA synthesis. The codon-anticodon recognition is rather straightforward because of the diversity of codons. Aminoacyl-tRNAs are synthesized by the 3'-esterification of tRNAs with the appropriate amino acids. For the majority of aminoacyl-tRNAs this is accomplished by direct aminoacylation of a particular tRNA with its cognate amino acid in a two-step reaction:

$$AA + ATP + AARS \Leftrightarrow AARS\bullet AA\text{-}AMP + PPi \quad (3)$$

$$AARS\bullet AA\text{-}AMP + tRNA \Leftrightarrow AARS + AA\text{-}tRNA + AMP, \quad (4)$$

where AA is an amino acid and AARS is the corresponding aminoacyl-tRNA synthetase.

It is necessary to clarify the naming conventions used here. The specific AARSs are denoted by their three-letter amino acid designation, e.g., AlaRS for alanyl-tRNA synthetase. Alanine tRNA or $tRNA^{Ala}$ denotes uncharged tRNA specific for alanine; alanyl-tRNA or Ala-tRNA denotes tRNA aminoacylated with alanine.

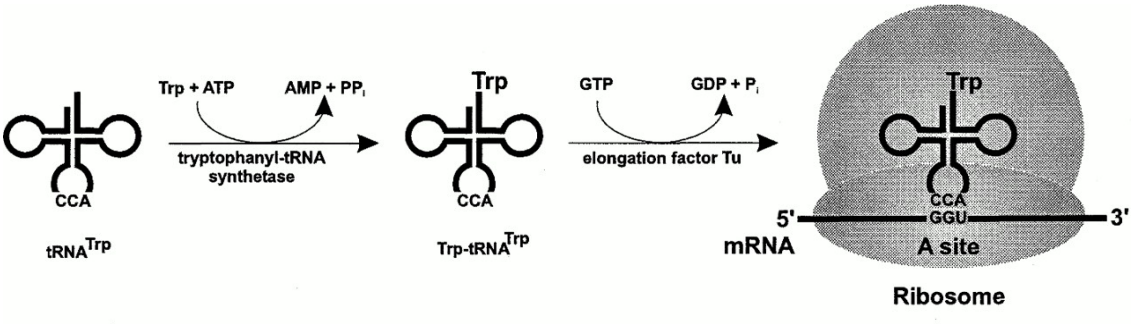


Figure 1. Role of aminoacyl-tRNA formation in the elongation phase of protein synthesis. An uncharged tRNA is first aminoacylated with the cognate amino acid to generate an aminoacyl-tRNA, which then interacts with the elongation factor. This allows delivery of the aminoacyl-tRNA to the ribosome A site, where its anticodon can interact with the cognate codon in mRNA. The example shown illustrates how this leads to the translation of the codon GGU as tryptophan. (Adapted from reference (23).)

The aminoacyl-tRNA synthetases have received much attention in recent years, and high resolution structures are now available for nearly every type of AARS (23). Many of these structures are complexes of AARS with various substrate ligands or inhibitors. Detailed biochemical and genetic characterizations have also helped our understanding of the mechanisms in various stages of the aminoacylation reaction. Collectively, these studies have now provided information on the expression, structure and function of the aminoacyl-tRNA synthetases in almost every detail. The exquisite specificity of these enzymes has been explained at a molecular level.

Concurrent with the surge in understanding of individual AARSs, the advent of whole-genome sequencing has provided a broader picture of the overall process of aminoacyl-tRNA synthesis. Over 30 complete families of AARSs are now known (24), while only the *Escherichia coli* AARSs were complete in 1991 (25). The availability of

so many sequences of AARS-encoding genes has led us to recognize the diversity of aminoacyl-tRNA synthesis mechanisms. For many years, people thought there were 20 AARSs (one for each amino acid) in each species. However, the genomic sequence of the hyperthermophilic archaeon *Methanococcus jannacshii* only contains 16 of the 20 known AARSs identified by homology technique (26). Studies have shown that this apparent shortfall can be contributed to the existence of previously uncharacterized AARSs and of novel pathways for aminoacyl-tRNA synthesis, and such pathways are found in a wide variety of organisms (27).

Numerous studies have shown that all AARSs catalyze essentially the same reaction. First, ATP and amino acid bind at the active site. The respective positioning of the α-phosphate group of the ATP and the α-carboxylate group of the amino acid allows the latter to attack the former by an inline nucleophilic displacement mechanism (Figure 2 a). This leads to the formation of an enzyme-bound mixed anhydride (aminoacyl-adenylate) and an inorganic pyrophosphate leaving group (Reaction 3 above). In the second step of the reaction, the 2'- or 3'-hydroxyl of the terminal adenosine of tRNA nucleophilically attacks the α-carbonyl of the aminoacyl adenylate (Figure 2 b). The final result is the 3'-esterification of the tRNA with the amino acid moiety and the generation of AMP as a leaving group (Reaction 4 above). The product is then released from the enzyme.

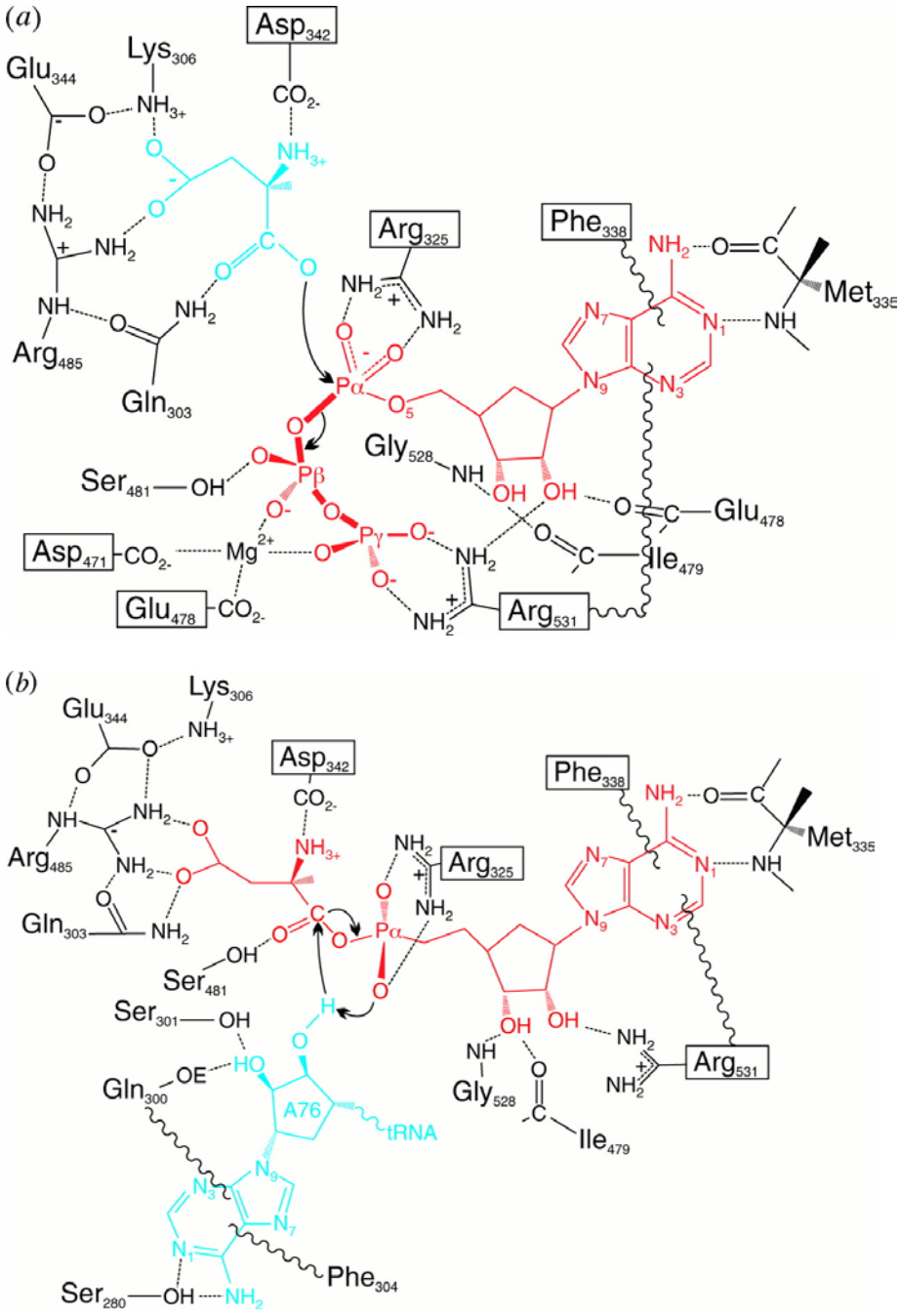


Figure 2. The two steps of the aminoacylation reaction deduced from the crystal structure of yeast AspRS. (a) Amino acid is activated by the forming of the aspartyl-adenylate and the release of pyrophosphate. The amino acid is shown in a postulated initial position. (b) Amino acid is transferred to the ribose of the 3'-end adenosine of tRNA. (Adapted from reference (23).)

Despite the conserved mechanisms of catalysis, the aminoacyl-tRNA synthetases can be divided into two unrelated classes (I and II) based on their mutually exclusive sequence motifs that reflect distinct topologies (28). In class I AARSs, the active site contains a Rossmann dinucleotide-binding domain, whereas this fold is absent from the active site of class II AARSs, which instead contains a novel antiparallel β-sheet fold. The difference results in the binding of ATP in different conformation in the active site. In class I AARSs, ATP assumes an extended conformation, while ATP is bent in class II AARSs. Inside each class, AARSs can be further divided into subclasses (29). Table 1 lists the detailed classification of all AARSs.

Table 1. The classification of aminoacyl-tRNA synthetases

Class	**AARSs**
Class I Motifs: ΦhΦGh, kmsKs	
Subclass Ia[a]	ArgRS[b], CysRS, IleRS, LeuRS, MetRS, ValRS, LysRS I[b,c]
Subclass Ib	GlnRS[b], GluRS[b]
Subclass Ic	TrpRS, TyrRS
Class II Motifs: (1)gΦxxΦxxpΦΦ (2)fRxe-h/rxxxFxxx(d/e) (3)gΦgΦgΦ(d/e)RΦΦΦΦΦΦ	
Subclass IIa	GlyRS[d], HisRS, ProRS, ThrRS, SerRS
Subclass IIb	AsnRS, AspRS, LysRS II[c]
Subclass IIc	AlaRS, GlyRS[d], PheRS

Φ: hydrophobic residue; Uppercase letter: strictly conserved; Lowercase letter: conserved but less strict; x: any residue

[a]Subclasses and motifs are defined in reference (29)

[b]ArgRS, LysRS I, GlnRS, and GluRS require the presence of tRNA for amino acid activation (30, 31)

[c]LysRS is found as both class I and class II AARS (31)

[d]GlyRS exists in two unrelated forms (32)

The other major difference between the two classes AARSs is in their binding of tRNA. From the available AARS:tRNA complex structures, class I AARSs approach the acceptor stem of tRNA from the minor groove side with variable loop facing the solvent, whereas class II AARSs approach the major groove side of the acceptor stem and the variable loop faces the synthetase. Whether this is true for all AARSs remains unclear, as there are only a handful of AARS:tRNA complex structures available.

An intrinsic requirement of aminoacyl-tRNA synthesis is that it should solely generate cognate aminoacyl-tRNA. Although there are other mechanisms guarding against the infidelity, the accuracy of aminoacyl-tRNA synthesis mostly depends on the specificity of the AARSs. The overall error of rate of aminoacyl-tRNA synthesis is about 1 in 10,000 (33). This fidelity is achieved in a number of ways: First, the AARSs make an intricate series of contacts with both their amino acid and tRNA substrates, which go a long way in ensuring that only the correct substrates are selected from the large cellular pool of similar ligands. In the case of cognate tRNA selection, the accuracy of the recognition is enhanced by the stabilization of the transition state for tRNA charging in cognate tRNA:AARS complexes and the existence if antideterminants in certain tRNAs that prevent interaction with non-cognate AARSs (34).

The discrimination of amino acids by AARSs is potentially more difficult, as some of the amino acids differ by only a methyl group. Nevertheless, some amino acids such as tyrosine and histidine have sufficiently unique side chains to prevent other amino acids in competing for the binding site. Others like valine and isoleucine do need extra

mechanism to assure the fidelity. For example, while IleRS misactivates valine at a frequency of 1 in 150, the overall error rate is only 1 in 3000 because of the proofread mechanism in IleRS (35). IleRS is able to proofread misactivated valine at two points in the aminoacylation reaction, both as a bound aminoacyl-adenylate (pre-transfer proofreading) and as a bound aminoacyl-tRNA (post-transfer proofreading). Both reactions are dependent on the presence of $tRNA^{Ile}$, as summarized below:

$$\text{IleRS} + \text{Val} + \text{ATP} \Leftrightarrow \text{IleRS}\bullet\text{Val-AMP} + \text{PPi} \quad (5)$$

$$\text{IleRS}\bullet\text{Val-AMP} + tRNA^{Ile} \Leftrightarrow \text{IleRS}\bullet\text{Val-AMP}\bullet tRNA^{Ile} \Leftrightarrow \text{IleRS}\bullet\text{Val-}tRNA^{Ile} + \text{AMP} \quad (6)$$

$$\downarrow (7) \qquad \downarrow (8)$$

$$\text{IleRS} + \text{Val} + \text{AMP} + tRNA^{Ile}$$

Thus IleRS uses a double sieve selection in its amino acid recognition. The first sieve guards against any amino acid large than isoleucine, while the second sieve hydrolyze any aminoacyl-$tRNA^{Ile}$ that has an amino acid smaller than isoleucine. Other AARSs, such as ValRS has a similar proofreading mechanism against threonine (36). The proofreading is a constant process, thus there is a significant energy cost in achieving accuracy in protein synthesis.

Because AARSs themselves are protein-based, the error in AARS itself can be propagated to affect the accuracy in the next generation protein synthesis. Therefore, there must be an error threshold in aminoacyl-tRNA synthesis. Above this threshold, the error propagation in protein synthesis will lead to an “Error Catastrophe” (37). Various

models have shown that this “Error Catastrophe” could be a genetically programmed cell death mechanism (38). So far these models have not been proved in experiments.

Because of the essential role of AARSs for cell activity, inhibition of a member of this family of enzymes is detrimental to the cell. This led to the early realization that if inhibitors of AARSs could be found that differentiate between bacterial or fungal AARSs and their human homologs, such compounds might provide a means of developing antibacterial or anti fungal agents. Over the years a number of natural products have been discovered that inhibit IleRS [furanomycin and pseudomonic acid], LeuRS [granaticin], PheRS [ochratoxin A], ProRS [cispentacin], ThrRS [borrilidin], and TrpRS [indolmycin] (see (23) for references). Currently pseudomonic acid has been developed into an antibiotic, mupirocin (39). The rapid rise of antibiotic-resistant pathogens has put considerable emphasis on the development of novel antibiotics, including search for potent AARS inhibitors. The availability of complete set of AARS from many organisms offers the potential to develop new potent AARS inhibitors.

References

1. McCammon, J. A., Gelin, B. R. & Karplus, M. (1977) *Nature* **267,** 585-590.
2. Doster, W., Cusack, S. & Petry, W. (1989) *Nature* **337,** 754-756.
3. Krupyanskii, Y. F., Parak, F., Goldanskii, V. I., Mossbauer, R. L., Gaubman, E. E., Engelmann, H. & Suzdalev, I. P. (1982) *Zeitschrift Fur Naturforschung C: Journal of Biosciences* **37,** 57-62.
4. Steinbach, P. J. & Brooks, B. R. (1994) *Chem. Phys. Lett.* **226,** 447-452.
5. Davenport, R. C., Bash, P. A., Seaton, B. A., Karplus, M., Petsko, G. A. & Ringe, D. (1991) *Biochemistry* **30,** 5821-5826.
6. Bash, P. A., Field, M. J., Davenport, R. C., Petsko, G. A., Ringe, D. & Karplus, M. (1991) *Biochemistry* **30,** 5826-5832.
7. Alhambra, C., Wu, L., Zhang, Z. Y. & Gao, J. L. (1998) *J. Am. Chem. Soc.* **120,** 3858-3866.
8. Mulholland, A. J., Lyne, P. D. & Karplus, M. (2000) *J. Am. Chem. Soc.* **122,** 534-535.
9. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998) *J. Phys. Chem. B* **102,** 3586-3616.
10. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1995) *J. Am. Chem. Soc.* **117,** 5179-5197.
11. Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. (2001) *J. Phys. Chem. B* **105,** 6474-6487.
12. Halgren, T. A. (1996) *J. Comput. Chem.* **17,** 490-519.
13. Mayo, S. L., Olafson, B. D. & Goddard, W. A. (1990) *J. Phys. Chem.* **94,** 8897-8909.
14. Becker, O. M., MacKerell, A. D., Roux, B. & Watanabe, M. (2001) (Marcel Dekker, New York).
15. Szwarc, M. (1956) *Nature* **178,** 1168-1169.
16. Faust, R. & Kennedy, J. P. (1986) *Polym. Bull.* **15,** 317-323.
17. Dias, E. L., Nguyen, S. T. & Grubbs, R. H. (1997) *J. Am. Chem. Soc.* **119,** 3887-3897.
18. Lynn, D. M., Mohr, B. & Grubbs, R. H. (1998) *J. Am. Chem. Soc.* **120,** 1627-1628.
19. McGrath, K. P., Fournier, M. J., Mason, T. L. & Tirrell, D. A. (1992) *J. Am. Chem. Soc.* **114,** 727-733.
20. Krejchi, M. T., Atkins, E. D. T., Waddon, A. J., Fournier, M. J., Mason, T. L. & Tirrell, D. A. (1994) *Science* **265,** 1427-1432.
21. Yu, S. J. M., Conticello, V. P., Zhang, G. H., Kayser, C., Fournier, M. J., Mason, T. L. & Tirrell, D. A. (1997) *Nature* **389,** 167-170.

22. Petka, W. A., Harden, J. L., McGrath, K. P., Wirtz, D. & Tirrell, D. A. (1998) *Science* **281,** 389-92.
23. Ibba, M. & Soll, D. (2000) *Annu. Rev. Biochem.* **69,** 617-650.
24. Woese, C. R., Olsen, G. J., Ibba, M. & Soll, D. (2000) *Microbiol. Mol. Biol. Rev.* **64,** 202-+.
25. Eriani, G., Dirheimer, G. & Gangloff, J. (1991) *Nucleic Acids Research* **19,** 265-269.
26. Bult, C. J., White, O., Olsen, G. J., Zhou, L. X., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H. P., Fraser, C. M., Smith, H. O., Woese, C. R. & Venter, J. C. (1996) *Science* **273,** 1058-1073.
27. Tumbula, D., Vothknecht, U. C., Kim, H. S., Ibba, M., Min, B., Li, T., Pelaschier, J., Stathopoulos, C., Becker, H. & Soll, D. (1999) *Genetics* **152,** 1269-1276.
28. Eriani, G., Delarue, M., Poch, O., Gangloff, J. & Moras, D. (1990) *Nature* **347,** 203-206.
29. Cusack, S. (1995) *Nat. Struct. Biol.* **2,** 824-831.
30. Ibba, M., Losey, H. C., Kawarabayashi, Y., Kikuchi, H., Bunjun, S. & Soll, D. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 418-423.
31. Schimmel, P. R. & Soll, D. (1979) *Annu. Rev. Biochem.* **48,** 601-648.
32. Mazauric, M. H., Reinbolt, J., Lorber, B., Ebel, C., Keith, G., Giege, R. & Kern, D. (1996) *Eur. J. Biochem.* **241,** 814-826.
33. Fersht, A. (1985) *Enzyme Structure and Mechanism* (Freeman, New York).
34. Giege, R., Sissler, M. & Florentz, C. (1998) *Nucleic Acids Research* **26,** 5017-5035.
35. Ibba, M. & Soll, D. (1999) *Science* **286,** 1893-1897.
36. Fersht, A. R. & Dingwall, C. (1979) *Biochemistry* **18,** 2627-2631.
37. Kowald, A. & Kirkwood, T. B. L. (1993) *Science* **260,** 1664-1665.
38. Freist, W., Sternbach, H., Pardowitz, I. & Cramer, F. (1998) *J. Theor. Biol.* **193,** 19-38.
39. Wuite, J., Davies, B. I., Go, M. J., Lambers, J. C., Jackson, D., Mellows, G. & Tasker, T. C. G. (1985) *J. Am. Acad. Dermat.* **12,** 1026-1031.

Chapter 2

Computational Studies of Selectivity and Specificity of Substrate Binding in Methionyl-tRNA Synthetase

Abstract

In vivo incorporation of amino acids in protein biosynthesis has a precisely controlled mechanism. The accuracy of this process is controlled to a significant extent by a class of enzymes called the aminoacyl-tRNA synthetases. Aminoacyl-tRNA synthetases achieve this control by a multi-step identification process that includes "physical" binding and "chemical" proofreading steps. However, the degree to which each synthetase uses these specificity-enhancing steps to distinguish their cognate amino acid from the non-cognate ones vary considerably. We have used HierDock computational protocol to elucidate this binding mechanism in methionyl-tRNA synthetase (MetRS) by first predicting the recognition site of Met in the apo form of methionyl-tRNA synthetase (apo-MetRS). We have developed this generalized procedure, which can be used to search for ligand binding region in globular proteins with no prior information about the binding site. We have further investigated the specificity of MetRS towards the binding of 19 other natural amino acids to both apo-MetRS and to the co-crystal structure of MetRS with Met bound to it (co-MetRS). We have established through our computed binding energies that the discrimination towards the non-cognate substrate increases in the second step of the physical binding process that is associated with a conformation change in the protein.

1. Introduction

Specific recognition of amino acids by their corresponding tRNAs and aminoacyl-tRNA synthetases (AARSs) is critical for the faithful translation of the genetic code into protein sequence information. The AARSs catalyze a two-step reaction in which amino acids are esterified to the 3′ end of their cognate tRNA substrates (1). In the first step, the amino acid and ATP are activated by the AARS to form an enzyme-bound aminoacyl-adenylate complex. In the second step, the activated amino acid is transferred to the 3′-ribose of the conserved CCA-3′ end of the cognate tRNA. The fidelity of protein synthesis depends, in most part, on the accuracy of this aminoacylation reaction. AARSs bind their cognate amino acid through a multi-step recognition process and correction mechanisms that include physical binding and a chemical proofreading (2). The four major steps involved in the transfer of aminoacyl group to the tRNA are

1. Binding of amino acid and ATP.
2. Conformational change in the AARS induced by binding and formation of the aminoacyl-adenylate complex.
3. Proofreading of misactivated non-cognate aminoacyl-adenylate complex.
4. Transfer of aminoacyl to the tRNA and proofreading.

The physical binding of the amino acid and ATP to AARSs is achieved in steps 1 and 2, which is accompanied by a conformation change in the AARSs. However, this binding event is necessary but not sufficient for the incorporation of the analog or the cognate amino acid during protein biosynthesis. Binding is followed by chemical

proofreading steps 3 and 4, which are termed as the pre-transfer and the post-transfer proofreading steps, respectively. With every step, the AARS recognizes its cognate amino acid with increased specificity, while discriminating more efficiently against the non-cognate amino acids. However, the degree to which each AARS uses the specificity enhancing steps varies considerably with regard to the 20 naturally occurring amino acids and the type of AARS. For example, tyrosyl-tRNA synthetase has the highest specificity in the first binding step whereas isoleucyl-tRNA synthetase, achieves maximum discrimination in the pre-transfer proofreading step (3-6).

Many research groups have focused on the use of *in vivo* methods for incorporating the non-natural amino acid analogs into proteins. It has been demonstrated that the wild-type translational apparatus can use non-natural amino acids with fluorinated, electroactive, unsaturated and other side chain functions (7-13). However, the number of amino acids shown conclusively to exhibit translational activity *in vivo* is small, and the chemical functionality that has been accessed by this method remains modest. Only those analogs that are able to successfully circumvent the multi-step filter mechanisms of the natural synthetases eventually get incorporated.

With an increase in efforts of incorporating artificial amino acids *in vivo*, it has become vital to enhance our understanding of the molecular level mechanism at different steps that AARSs utilize to ensure high fidelity in translation. A better understanding of this mechanism will also be very useful in allowing us to design mutants of AARS for incorporation of specific analogs and also in suggesting analogs that are more efficiently

incorporated (14-16). In this study, we have implemented a computational procedure to gain insight into the binding mechanism of methionyl-tRNA synthetase (MetRS). Computational methods are becoming increasingly important to understand the molecular level mechanisms that are not feasible with experiments and also for faster virtual screening of analogs prior to synthesis.

MetRS is a class I AARS, and undergoes a large conformational change upon substrate Met binding. It is a dimeric protein and the crystal structures of *E. coli* MetRS in its apo form and as a co-crystal with its native ligand, Met, have been solved to 1.85Å and 2.03Å resolution, respectively. [We refer to the apo form of MetRS protein structure as apo-MetRS(crystal) and the co-crystal structure of *E. coli* MetRS with Met as Met/MetRS(crystal). Note that the protein conformations in both these crystal structures are different especially in the binding site. The symbol MetRS always denotes the *E. coli* MetRS unless otherwise specified.] Both *in vivo* incorporation of Met analogs into proteins and their *in vitro* measurements of the rate of incorporation have been studied extensively and it has been demonstrated that MetRS is one of the more permissive AARSs for the incorporation of a large number of analogs (8). We are interested in computationally determining the specificity of MetRS for the natural non-cognate amino acids and Met analogs in the steps of amino acid recognition and binding. A better understanding of its binding mechanism would be useful to streamline a virtual screening approach for the incorporation of non-natural amino acids.

We have used the HierDock computational protocol (17) to first predict the binding site of Met in MetRS in the apo-MetRS (crystal). We scanned through the entire protein (except the anticodon recognition region) for predicting the preferential binding site for Met using no knowledge from the crystal structure of Met/MetRS (crystal). We refined the HierDock protocol and derived what we call as the "recognition site" which includes all the residues in the binding pocket of Met as seen in Met/MetRS (crystal), however, Met is oriented in this pocket with its side chain exposed to solvent. Our results indicate that the first step to amino acid binding is the recognition of the zwitterions part of the ligand, which is referred to as the "recognition mode." We find that apo-MetRS is able to distinguish Met from the non-cognate natural amino acids but has cysteine and serine as competitors. We also find that MetRS in the Met/MetRS (FF) protein structure has better discrimination for the twenty amino acids and once again Met has the best binding energy in this structure with Gln as a close competitor.

The calculated binding energies of Met analogs are correlated with either the *in vivo* incorporation results or the *in vitro* measurements of rate of the aminoacyl-adenylate formation. We find that in Met/MetRS (FF) protein, the analog with high incorporation rates bind better than those that do not get incorporated. In an attempt to incorporate novel Met analogs Kiick *et al.* reported that Homopropargylglycine (myag) replaces Met most efficiently utilizing the natural translation apparatus of *E. coli* while cis-crotglycine (ccg) shows almost no incorporation (18). Our calculated binding energies correlate well with the *in vivo* incorporation trends exhibited by these analogs and with the binding energies calculated by *in vitro* methods.

2. Methods

2.1 Preparation and Optimization of Structures

Ligand Structures: Both the neutral and the zwitterion forms were used for all the twenty natural amino acids as well as the five Met analogs. The ligand conformations were optimized in the extended conformation at the Hartree-Fock level of theory with a 6-31G** basis set, including Poisson-Boltzmann continuum dielectric solvation using the Jaguar computational suite (19) (Schrödinger, Portland, OR). The Mulliken charges ascertained from this calculation were retained for the subsequent molecular mechanics simulations. The conformations of the five Met analogs are shown in Figure 4 a.

Preparation and Optimization of Protein Structures: The 2.03Å *E. coli* apo-MetRS structure was obtained from PDB database (pdb code: 1QQT) that included the fully active monomer α chain of a homodimer, crystal waters, and a zinc (II) ion (20). CHARMM22 charges with the nonpolar hydrogen charges summed onto the heavy atoms were assigned to the α chain according to the parameters set forth in the DREIDING force field (21). The protein was neutralized by adding counterions (Na^+ and Cl^-) to the charged residues (Asp, Arg, Glu and Lys) and subject to a minimization of the potential energy by the conjugate gradient method using Surface Generalized Born continuum solvation method (22). The RMS in coordinates (CRMS) of all atoms after minimization is 0.68Å and this structure is referred to as apo-MetRS(FF). Using the same procedure the co-crystal structure of *E. coli* MetRS (pdb code: 1F4L; resolution 1.85Å) was minimized and the CRMS for all atoms of the minimized structure compared to the

crystal is 0.57Å (23). We refer to this structure as Met/MetRS (FF). The CRMS values for both structures are well within experimental error that demonstrates the proficiency of our FF used in present studies. The crystal waters and other bound molecules were removed for docking to maximize the searchable surface of the protein. We have used SGB continuum solvation method for all structure optimizations and energy scoring in this study with an internal protein dielectric constant of 2.5 was employed for all calculations.

2.2 HierDock Protocol

We use the HierDock procedure, which has been applied successfully to study the binding of odorants to membrane-bound olfactory receptors (17), for outer membrane protein A binding to sugars (15) and for Phe and its analogs binding to PheRS (24). The HierDock ligand screening protocol follows a hierarchical strategy for examining conformations, binding sites and binding energies. Such a hierarchical method has been shown to be necessary for docking algorithms (25). The steps in HierDock involve using coarse-grain docking methods to generate several conformations of protein/ligand complexes followed by molecular dynamics (MD) simulations including continuum solvation methods performed on a subset of good conformations generated from the coarse-grain docking. Methods combining docking and MD simulations have been tested (26) but the main drawback of these tests was that only one protein/ligand complex structure was kept from the coarse-grain docking methods for MD simulations. This is risky considering that the coarse-grain methods do not have accurate scoring functions that include solvation.

Free energy perturbation methods are generally regarded to lead to accurate free energies of binding but are computationally intensive and not readily applicable to a wide variety of ligands (27). Our goal is to derive a fast hierarchical computational protocol that uses hierarchical conformation search methods along with different levels of scoring functions, which would allow screening of amino acid analogs for AARSs. The three major steps in HierDock procedure in this paper are as follows:

- First, a coarse-grain docking procedure to generate a set of conformations for ligand binding. In this paper we used DOCK 4.0 (28, 29) to generate and score 20000 configurations, of which 10% were ranked using the DOCK scoring function.
- We then select the 20 best conformations for each ligand from DOCK and subject them to annealing molecular dynamics (MD) to further optimize the conformation in the local binding pocket, allowing the atoms of the ligand to move in the field of the protein. In this step the system was heated and cooled from 50 K to 600 K in steps of 10 K (0.05 ps at each temperature) for one cycle. At the end of annealing MD cycle, the best energy structure is retained. Annealing MD allows the ligand to readjust in the binding pocket to optimize its interaction with the protein. This fine grain optimization was performed using MPSim (30) with DREIDING force field (21) and continuum solvation methods. We use the Surface Generalized Born (SGB) continuum solvent method to obtain forces and energies resulting from the polarization of the solvent by the charges of the ligand and protein. This allows us to calculate the change in the ligand structure due to

the solvent field and hence, obtain more realistic binding energies that take into account the solvation effects on the ligand/protein structure. The annealing MD procedure generated 20 protein/ligand complexes for each ligand.

- For the 20 structures generated by annealing MD simulations for each ligand, we minimized the potential energy (conjugate gradients) of the full ligand/protein complex in aqueous solution using SGB. This step of protein/ligand-complex optimization is critical for obtaining energetically good conformations for the complex (cavity + ligand). Then we calculated binding energies as the difference between the total energy of the ligand-protein complex in solvent (ΔG(protein+ligand)) and the sum of the total energies of the protein (ΔG (protein)) and the ligand separately in solvent (ΔG(ligand)). The energies of the protein and the ligand in solvent were calculated after independent energy minimization of the protein and the ligand separately in water. Solvation energies were calculated using Poisson-Boltzmann continuum solvation method available in the software Delphi. The non-bond interaction energies were calculated exactly using all pair interactions. Thus the binding energy is given by

$$\Delta\Delta G_{calc} = \Delta G(protein + ligand) - \Delta G(protein) - \Delta G(ligand) \qquad (1)$$

Since the structure optimizations included solvation forces using the SGB continuum solvent approximation with the experimental dielectric constant, we consider that the calculated energies are free energies (31). This multi-step scanning procedure is based on docking *via* DOCK 4.0 coupled with fine-grain MM techniques. The coarse grain docked complex structures generated are scored with FF and differential solvation, which effectively filters the docked complexes

to isolate the top contenders. As demonstrated in our previous studies (unpublished result), Dock 4.0 structures vary erratically with rank, whereas filtering with MPSim optimization brings the best structures to the top of the rank list.

2.3 Scanning the Entire apo-MetRS(FF) for Predicting Binding Site of Met

For the case of apo-MetRS(FF) we wanted to test the HierDock procedure for scanning the entire protein for the favorable binding site of Met. However, it has not been tested for a case where the protein undergoes a large conformational change in the binding site after the ligand binding starting from apo-protein structure. The steps involved in the scanning procedure are as follows:

1. *Mapping of possible binding regions*. A probe of 1.4 Å radius was used to trace a 4 dots/Å negative image of the protein molecular surface, according to Connolly's method (32). The resulting data was used to generate clusters of overlapping spheres with the SPHGEN program. These spheres serve as the basis for the docking method.
2. *Definition of docking region*. The pockets of empty space of the receptor (apo-MetRS(FF)) surface represented by spheres were divided into 14 possible 10 Å wide overlapping cubes, which covered the entire protein surface. Each region was scanned to determine its suitability as a binding site.

The site that contains the greatest number of lowest energy docked conformations is designated as the putative binding region.

3. *Prediction of binding site*. Steps 1 to 3 of HierDock procedure were performed with Met as the ligand in all the 14 possible binding regions in the entire apo-MetRS(FF). The orientations of the ligand in the receptor were generated by DOCK 4.0, using flexible docking with torsional minimization of the ligand, a continuum dielectric of 1.0 and a distance cutoff of 10 Å for the evaluation of energy.

4. *Selection of the most probable binding site and best configurations*. The best conformation from each region was determined using the buried surface area cutoff criteria for the ligand along with the binding energy. Such a buried surface area cutoff is required for filtering at the coarse grain level. An average of the most buried and the least buried conformer was calculated and all conformers whose buried surface area was lower than the average were eliminated from further analysis (33). The conformations that passed the buried surface area filter were sorted by binding energies calculated using equation (1) and the conformation with the best binding energy in every region were compared between regions. All the complex energies were calculated. The region with the lowest energy binding energy calculated using equation (1) was selected as the preferential binding region.

2.4 Docking of Ligand Pool into the Binding Region and Calculating Relative Binding Energies

Steps 1 to 3 of HierDock procedure were performed for all the ligands in the ligand pool in the putative binding region and the relative binding energies for the best ligand conformations were calculated using equation (1). The ligands (20 natural amino acids and analogs of Met) were ranked according to binding affinities to determine which ligands have the highest affinity for the binding site. The best energy conformation of Met in optimized apo-MetRS(FF) structure is the predicted structure of Met in apo-MetRS(FF). We denote this predicted structure as Met/apo-MetRS(FF).

2.5 Prediction of Binding Site for Met/MetRS(FF) Co-Crystal Structure

For the case of Met/MetRS(FF) structure, the receptor was prepared by removing the Met from the Met/MetRS(FF) structure. The protein surface was mapped with spheres, as described above, and the binding regions were covered by a 12 Å × 12 Å × 12 Å box centered in the center of mass of Met ligand. Only this region was used in subsequent docking. Steps 1 to 3 of HierDock procedure were performed using the same set of control parameters but only in the known binding region. The conformation with the best energy binding energy in this region calculated using equation (1), starting from the protein structure in Met/MetRS (FF) is the predicted co-crystal structure of Met/MetRS. We denote this predicted structure of Met in MetRS co-crystal structure as Met/MetRS (HierDock).

2.6 Docking of Ligand Pool into the Binding Site and Calculating Relative Binding Energies in Met/MetRS (HierDock)

We performed steps 1 to 3 of HierDock procedure for all 20 natural amino acids and the Met analogs in the 12Å x 12Å x12Å binding region and the relative binding energies for the best ligand conformation for each ligand was calculated using equation (1). The ligands (20 natural amino acids and analogs of Met) can then be ranked according to binding affinities to determine which ligands have the highest affinity for the binding site.

2.7 Binding Energy Calculation of the 20 Natural Amino Acids and Met Analogs in the Conformation that Activates the Protein

HierDock protocol predicts the best energy conformation for each ligand (20 natural amino acids and Met analogs) in the defined 12Å x 12Å x 12Å binding region in Met/MetRS(FF) structure. These predictions give rise to different preferred binding conformation for each ligand. However, the orientation that Met adopts in the active site with all the necessary contacts required for the enzymatic activity is referred to as the "activation mode." To assess the relative binding energies of the 20 natural amino acids and their analogs in the activation mode perturbation calculations for all the ligands were performed as follows:

- An amino acid rotamer library (34) was used to generate all the conformations of each amino acid in the binding site, and a similar library was generated for the five Met analogs.

- The best rotamer was chosen by matching each rotamer k in the binding site and evaluated with the following equation using the Dreiding force field:

$$E_k = \sum_{i,j}\left(\frac{q_i q_j}{4\pi\varepsilon r_{ij}} + D_e\left(\left(\frac{r_m}{r_{ij}}\right)^{12} - 2\left(\frac{r_m}{r_{ij}}\right)^{6}\right) + D_{HB}\left(5\left(\frac{r_{HB}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{HB}}{r_{ij}}\right)^{10}\right)\cos^4\theta\right), \quad (2)$$

where i and j sum over all atoms in the ligand and protein residue residues in the binding site, q_i and q_j are partial charges on atoms i and j, respectively. r_{ij} is the distance between atoms i and j, and r_m and D_e are van der Waals distance and well depth of atoms i and j, r_{HB} and D_{HB} are hydrogen bond distance and well depth, respectively. θ is the hydrogen bond angle between atoms i, j and their bridging hydrogen atom. The hydrogen bond term is only evaluated for hydrogen bond donor and acceptor atoms. To avoid overpenalizing clash, the van der Waals radii were reduced to 90% of the standard values in the Dreiding force field.

- After the best rotamer was chosen for each ligand, the total energy was minimized in the presence of protein, and the binding energy was then calculated using equation (1) for each of the twenty natural amino acids in the "activation mode" and compared.

3. Results and Discussion

3.1 Prediction of Binding Site of Met in apo-MetRS(FF) and Met/MetRS(FF)

Figure 1 shows the location of region 14 box in apo-MetRS(FF), which was determined to be the binding region by sifting through the 14 regions in apo-MetRS. The best conformation of Met in this region shows Met to be making electrostatic interactions

with His301 and Asp52 (Figure 3 c), the two amino acids that have been shown to play a significant role in Met binding (35, 36). His301 to Ala mutation results in loss of the affinity for Met and D52A mutation reduces the K_{cat} of the adenylation reaction by four folds indicating that it has a major role on the catalytic step in the formation of methionyl adenylate. Tyr15, another key amino acid determined by mutation analysis and has been structurally observed in the co-crystal structure to form the binding pocket for Met (23, 37), is located within 5 Å of the docked Met. The main component of the binding energy in our predicted binding orientation comes from the electrostatic interactions that Met makes with Asp52 and His301 followed by its van der Waals interactions in this binding region.

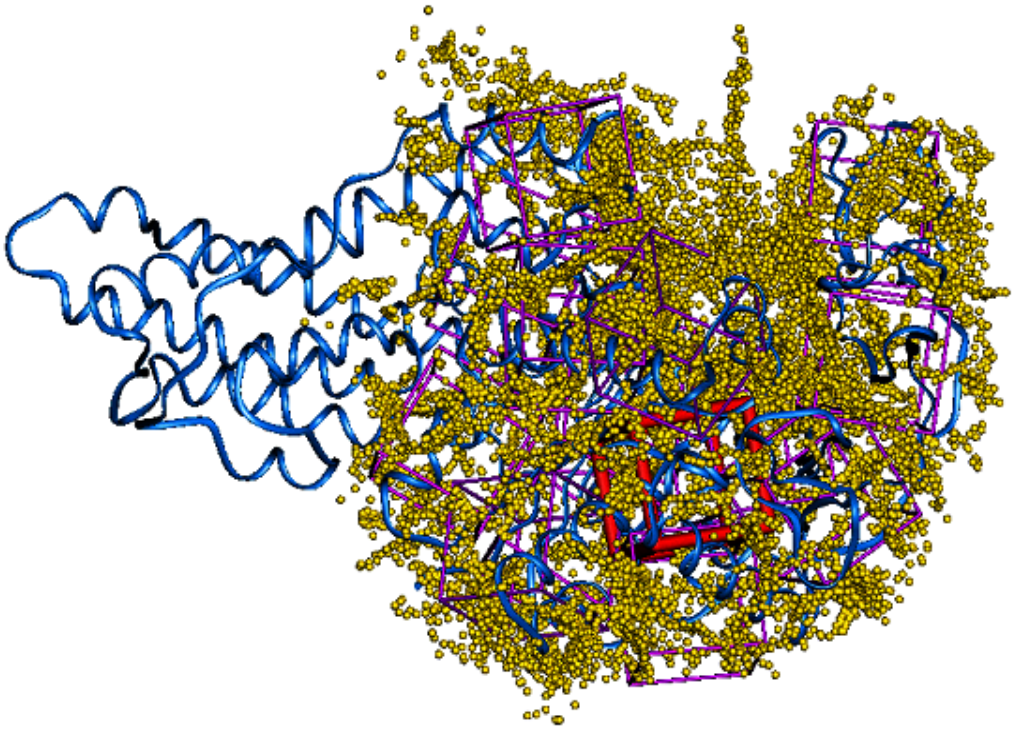
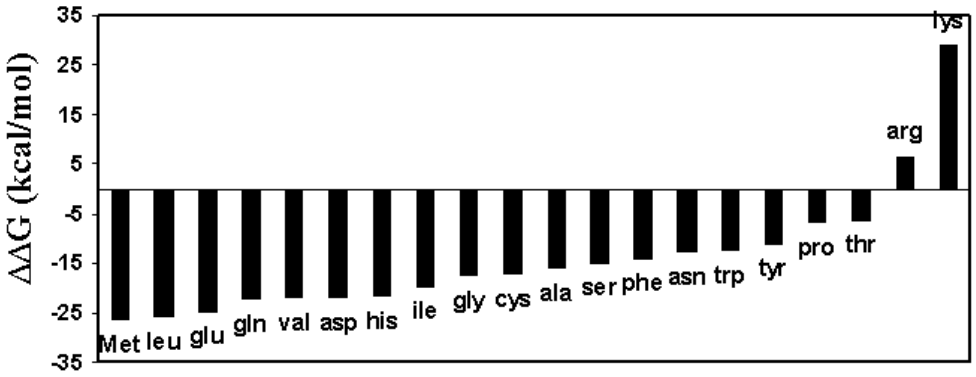


Figure 1. Sphere filled volume of MetRS representing the possible binding sites in the enzyme. The search volume was divided into 14 regions as indicated by the cubic boxes. The binding site was found in the box colored in red.

a)



b)

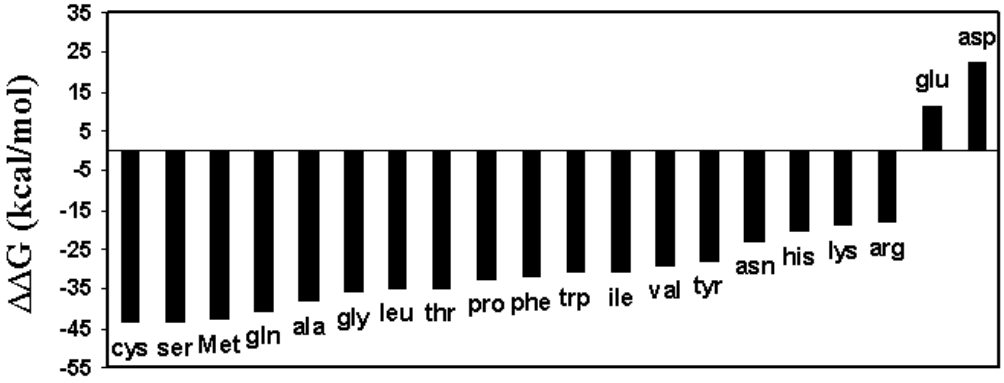
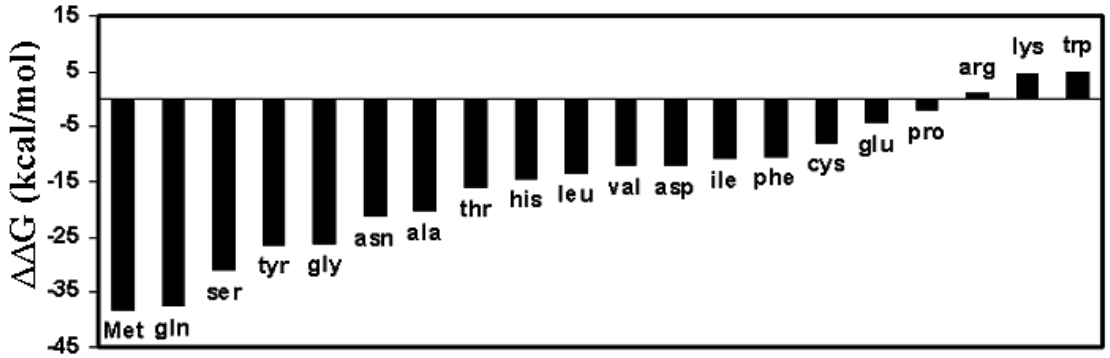


Figure 2. Binding energies of all 20 amino acids in the methionine binding site in Met/MetRS(FF) and apo-MetRS(FF). (a) Shows binding energies of the 20 amino acids when docked in the predicted methionine binding site in apo-MetRS(FF), and (b) shows the binding energies generated from perturbation analysis at the same site.

c)



d)

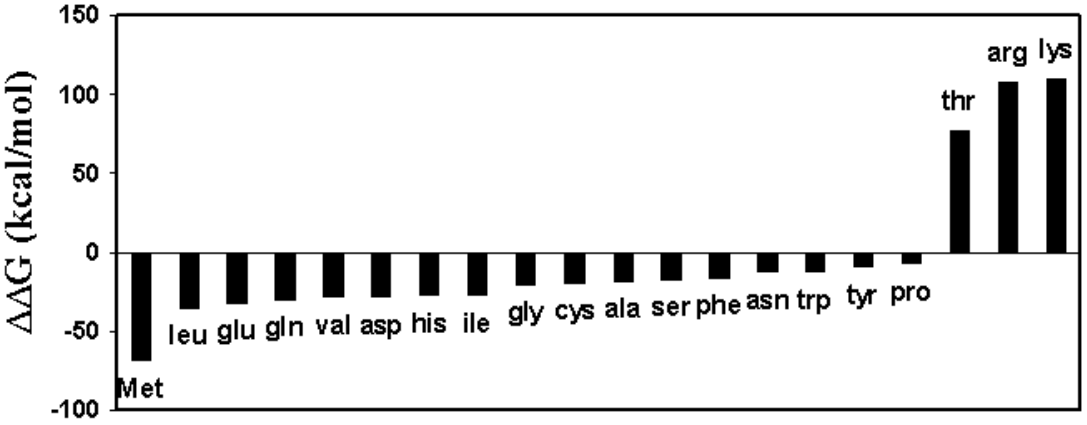


Figure 2. Binding energies of all 20 amino acids in the methionine binding site in Met/MetRS(FF) and apo-MetRS(FF). (c) Reports the binding energies generated from docking all 20 amino acids in the crystallographic methionine binding site in Met/MetRS(FF). (d) Indicates binding energies calculated from perturbation analysis at the same site.

It was speculated, and subsequently, has been observed that the terminal methyl group makes contacts with Trp305 (23). But in our model we find Met to be in exactly the opposite orientation. The terminal methyl group is solvent exposed and is 10Å away from Trp305 which is considered to stabilize the enzyme-Met complex. However, it is

suggested that Trp305 does not play a role in defining specificity. Trp305 occupies the same position in the apo enzyme and the complex, and its role has been suggested to exclude water molecules from the binding site and correct positioning of Met. In fact, it has been observed that an aromatic character of residue at this position seems to be enough to assure Met binding (36).

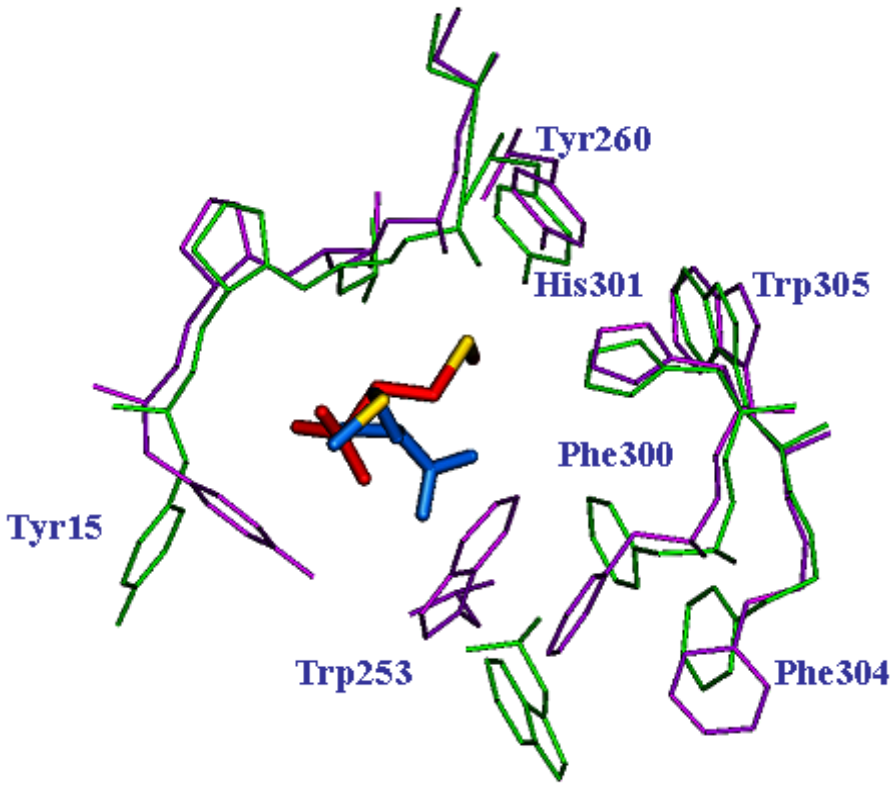


Figure 3. (a) Binding site of methionine in apo-MetRS(FF) and Met/MetRS(FF). Amino acids lining the binding pocket are shown in purple for apo-MetRS(FF) and in green for Met/MetRS(FF). Methionine orientation from perturbation analysis in Met/MetRS(FF) in shown in red and its conformation from docking in apo-MetRS(FF) is colored blue. Residues closest to methionine that undergo the largest conformation changes are labeled.

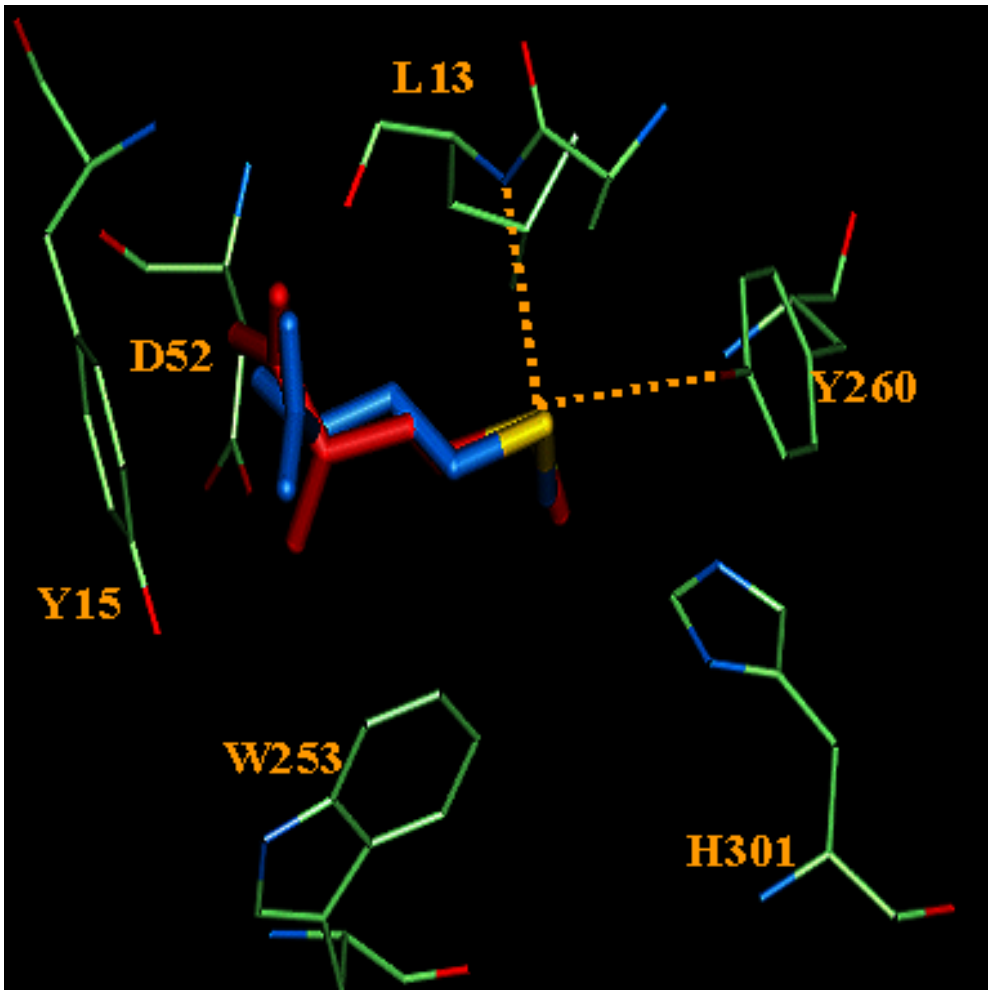


Figure 3. (b) The S^{δ} of methionine makes two hydrogen bonds – one with the terminal oxygen of Tyr260 and the other with the backbone amide of Leu13 in the docked conformation in Met/MetRS(FF). The crystal structure orientation on methionine is shown in blue. The CRMS between the two conformations is 0.55 Å.

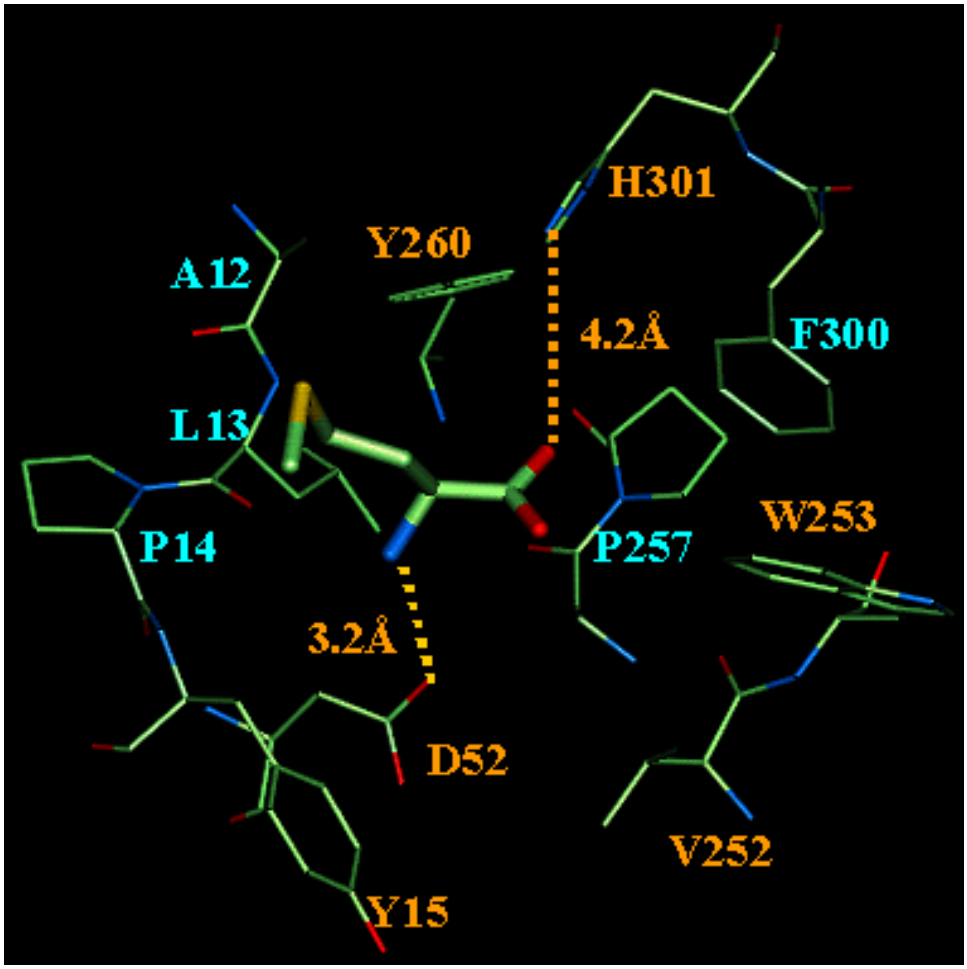


Figure 3. (c) The predicted binding site for methionine in apo-MetRS(FF). The conserved residues within 4 Å are labeled in gold and the conserved replacements are labeled in aqua.

The docked orientation in the apo enzyme occupies the identical position in the binding pocket as seen in the co-crystal structure (Figure 3 a). However the orientation of Met and the residues lining the binding pocket including parts of the protein backbone are very different in the two conformations of MetRS. Although Met seems to be making electrostatic contact with Asp52, one of the anchoring residues, the side chain of Met is not buried in the 7 Å pocket. The reason for this is that we have used the unbounded

structure of the synthetase, which, on binding to the amino acids undergoes significant conformation change. The co-MetRS structure suggests that the large solvent exposed cavity become reduced in volume as it gets partially filled with Met and Tyr15, Trp253, Phe300, Trp229, Phe304 and Tyr251. These residues are significantly displaced from their apo-enzyme orientation as they reorient to form a hydrophobic pocket for Met. In our predicted binding mode of Met in the apo-enzyme, all these residues are within 5 Å of Met ligand. We expect this to be the initial binding orientation of Met.

Another interesting observation that substantiates that the predicted orientation of Met could be the initial binding mode is that Met has one of the best binding energies of all 20 natural amino acids in this region (Figure 2 a). The specificity of this site further confirms that we have been able to find the correct binding region. Met has Ser and Cys as close competitors but they get eliminated as the protein undergoes conformation change. In an attempt to force the side chain of Met to be buried in the pocket we did annealing dynamics of the entire complex with solvation and reduced vdW radii of the ligand atoms. However, the orientation of Met did not change.

Also, a number of residues within 5 Å of Met in this region are conserved among a large number of organisms. In a sequence alignment among 59 prokaryotes we find all the amino acids within 4 Å of Met in the predicted binding region are either strictly conserved or are conserved replacements. Of the 12 residues within 4 Å of Met, 7 of them (Y15, D52, V252, W253, A256, Y260, H301) are strictly conserved and 5 (A12, L13, P14, P257, F300) are conserved replacements (Figure 3 c). This is interesting

considering that there are only 21 positions in the entire alignment that are strictly conserved and we find a third of them in our predicted binding region without any prior knowledge of the binding site. A binding search protocol for unliganded proteins followed by a sequence alignment analysis for the predicted binding region could provide more evidence on the accuracy of the predicted binding site and help in recognizing key amino acids lining binding pocket. Generally, one would expect to see conserved residues or conservative replacements in substrate binding sites in proteins across various species.

We also docked Met in the binding region of the co-MetRS(FF). This test was performed to check if we were able to predict the crystallographic binding orientation of Met in the binding pocket. This test was important to validate the accuracy of our docking protocol and the force field. Our predicted structure had a CRMS deviation of 0.55 Å from the crystal structure (Figure 3 b).

3.2 Specificity for Met in 1QQT and 1FTM

We docked all 20 amino acids and calculated their binding energies in the predicted binding region in apo-MetRS(FF**)** and the crystallographic binding site in Met/MetRS(FF). We also did perturbation studies of the natural amino acids in these two structures. The perturbation studies were done to analyze the binding energies of the non-cognate amino acids if they oriented in a similar conformation in the binding site as Met.

Perturbation analysis:

In the case of apo-MetRS(FF**)** closest competitors for Met are Ser and Cys. However, as the enzyme undergoes conformation change, its ability to discriminate against these non-cognate residues increases significantly. It has been noticed that for most synthetases there is no absolute specificity for the cognate substrate in the sense of a "lock and key" model. For example, yeast IleRS is not able to distinguish between Trp and Ile in the first step of binding because of the higher hydrophobic interactions gained by the non-cognate substrate. However, as the initial binding process is completed, the enzyme is able to discriminate against the non-cognate amino acids more easily (2, 4).

In Met/MetRS(FF), Met has the best binding energy, and it has an energy difference of more than 20 kcal/mol with its closest competitors, Asn and Arg (Figure 2 d). The closest competitors from the first binding step (Leu, Glu and Gln) are discriminated against with a very high efficiency as the structure of the protein changes.

Docking analysis:

The docking study was done predominantly to recognize possible competitors of Met. It may be possible that a non-cognate amino acid binds at the Met binding pocket but does not make the critical interactions that methionine makes in this binding pocket. In such cases, the amino acid may not be able to react with ATP and charge the tRNA. In apo-MetRS(FF), Met has the best binding energy of –26.38 kcal/mol with Leu, Gln and Glu as the closest competitors (Figure 2 a). In Met/MetRS(FF) Met again has the best energy with Gln and Ser as the closest competitors. Gln, in its preferred binding site in

Met/MetRS(FF) has its zwitterions part and the χ_1 torsional angle in the same orientation as Met at this site. Yet, its χ_2 and χ_3 angles are significantly different from that of Met. The S^{δ} of Met makes two hydrogen bonds—one with the terminal oxygen of Tyr260 and the other with the backbone amide of Leu13. However, because of the difference in its binding mode, Gln is unable to make a hydrogen bond with Tyr260 and makes only a weak hydrogen bond with the backbone amide of Leu13 (O…H-N distance of 3.9 Å).

One more observation is that the order of binding of the amino acids is identical in the docking analysis in apo-MetRS(FF) and the perturbation study in Met/MetRS(FF) (Figures 2 a, 2 d). It indicates that when the enzyme undergoes structural change, if all the amino acids were to bind in the binding mode of Met in the co-crystal structure, their order of binding would remain the same as indicated by the apo enzyme. However, the magnitudes of binding energies, which indicate the level of discrimination, would be very different.

3.3 Binding Energies of Met Analogs

To test the sensitivity of our simulation procedure, we wanted to test if we could get good correlation between the computed binding energies for the Met analogs with experimental binding energies. We tested five Met analogs of which four get incorporated into proteins with reasonable efficiency and for which the experimental binding energies are available. Ccg, which is a *cis*-form of tcg (Figure 4 a), has the lowest incorporation efficiency and hence, it was used as a negative control for which we

hoped to get the worst binding energy for this analog. Binding energy calculations of the Met analogs were carried out in the conformation that activates the protein, i.e., by perturbation analysis.

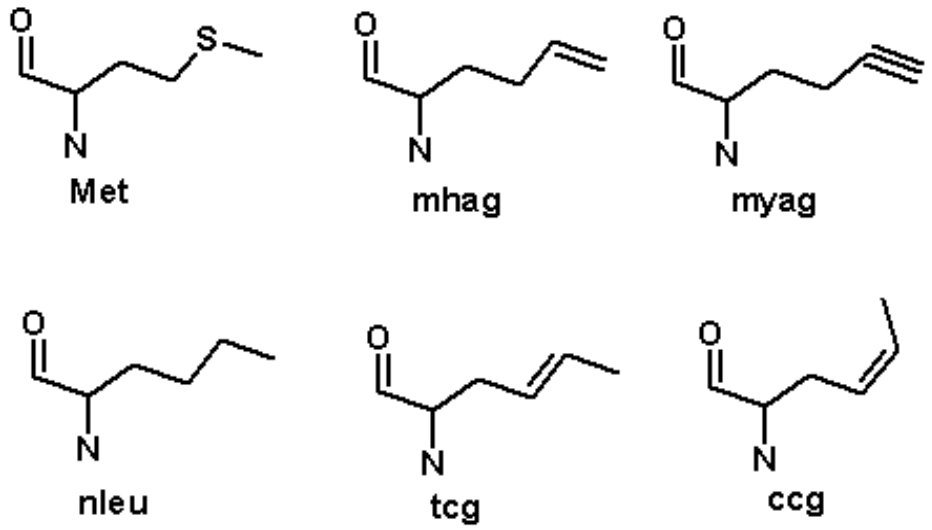
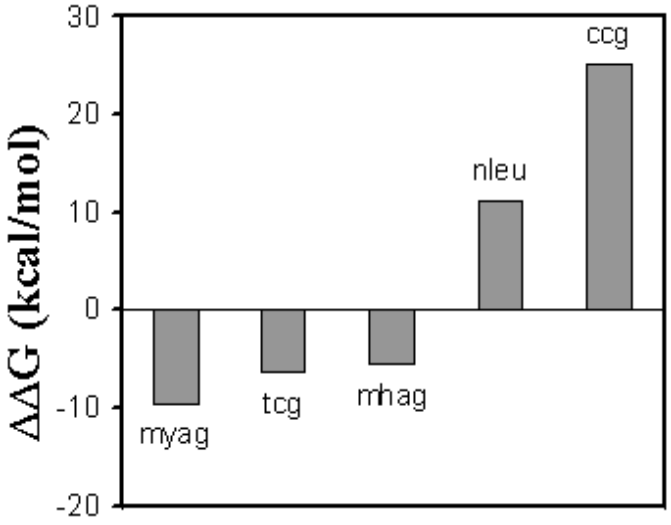


Figure 4. (a) Structures of methionine and its analogs used in this study. L-methionine (Met), homoallylglycine (mhag), homopropargylglycine (myag), norleucine (nleu), *trans*-crotglycine (tcg) and *cis*-crotglycine (ccg).

b)



c)

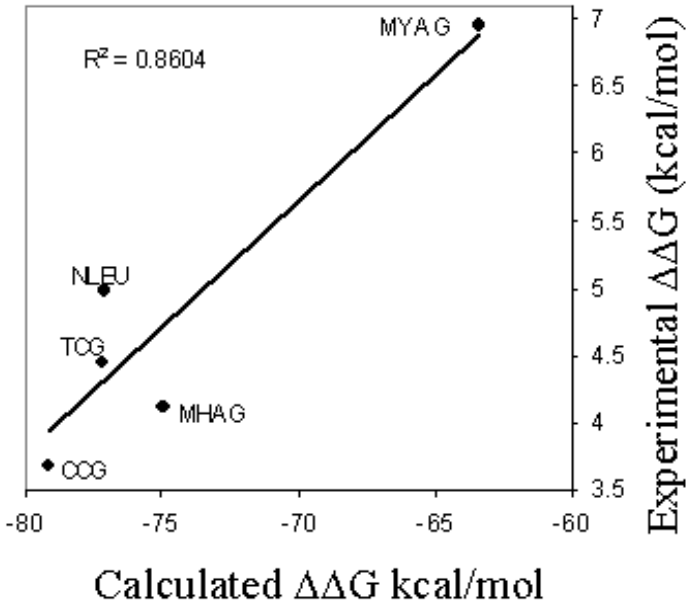


Figure 4. (b) Binding energies of the analogs in the binding site of Met/MetRS(FF) calculated using perturbation method. (c) Shows the correlation between the calculated binding energies and the experimentally observed ΔΔG with respect to methionine.

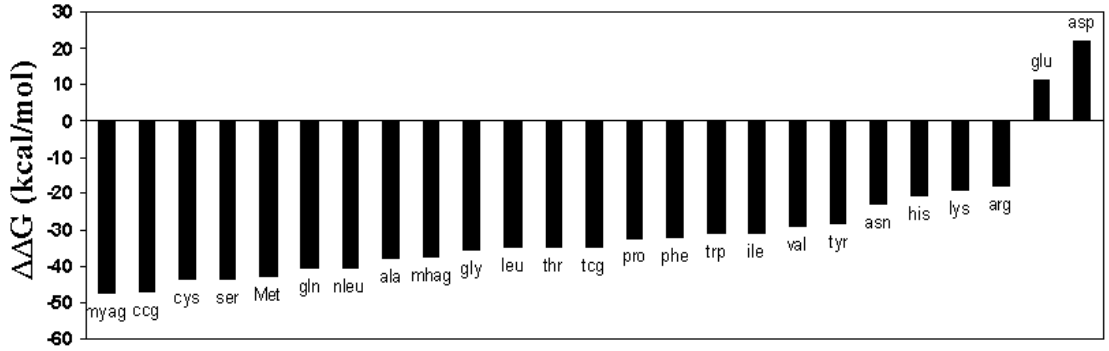


Figure 4. (d) Binding energies of analogs along with the natural amino acids in the binding site of apo-MetRS (FF).

In the case of 1QQT, the binding energies of the analogs are all in the top 50% but are interspersed with the non-cognate natural amino acids This indicates that in this conformation, MetRS has an inefficiently discrimination capability (Figure 4 d). However, in the co-crystal structure, there is a clear preference for binding the analogs. The analogs and Met have a binding energy range of -63.4 to -79.1 kcal /mol (Figure 4 b). The closest competitor from the non-cognate set of natural amino acids has a binding energy of -35.0 kcal/mol. In this conformation, we also find a good correlation between experimentally observed binding energies and computed binding energies (Figure 4 c). As we had expected, ccg has the worst binding energy and gets incorporated with the lowest efficiency whereas myag has the best calculated binding energy and has been tested to be the best Met analog. This information could be useful for initial computational scanning of the analogs before experimental testing. The binding energy of ccg in Met/MetRS(FF) could be used as a cutoff for designing new analogs and the ones that rank above the cut off could be experimentally tested for binding.

We analyzed the binding modes of ccg and tcg to understand what in particular about the cis-form of the ligand renders it to be an unfavorable ligand. We analyzed the non-bond energies of these ligands with all the residues lining the binding pocket and have tabulated our findings as pairwise interactions in Table 1. Ccg has a VDW clash with Ala12, the terminal hydroxyl group of Tyr260 and His301. At the same time, the cis orientation of terminal methyl group does not make the same favorable interactions with Ala 256 and Pro 267 as tcg (Figure 5). Since Tyr260 and His301 have an important role in the binding process as indicated by experiments, mutating them to smaller residues may be deleterious. On the other hand, it would be interesting to explore the effect of Ala to Gly mutation at position 12 on the incorporation of *cis* forms of various analogs.

Table 1. Comparison of the interactions of ccg and tcg with residues in the binding site. These energies were calculated using Equation 2.

	ccg			tcg		
Residue	vdW	Coulomb	H-bond	vdW	Coulomb	H-bond
Asp52	0.438	-21.47	-10.246	0.364	-21.427	-9.839
Leu13	-1.255	-6.045	-9.898	-1.762	-6.142	-10.19
Tyr15	-2.173	-7.773	-0.677	-3.282	-6.162	-0.122
Trp253	-3.779	-1.934	0.000	-3.779	-1.879	0.000
Ile297	-2.097	-0.454	0.000	-0.967	-1.585	0.000
Pro14	-0.973	-1.467	0.000	-1.931	-0.526	0.000
His301	-0.102	-1.189	0.000	-1.216	-1.074	0.000
Pro257	-0.670	-0.122	0.000	-1.486	-0.080	0.000
Ile293	-0.273	-0.145	0.000	-1.110	-0.232	0.000
Tyr260	-0.227	-0.116	0.000	-1.780	0.706	0.000
Ala256	-0.941	0.616	0.000	-1.414	0.601	0.000
Val252	-0.233	-0.025	0.000	-0.227	-0.047	0.000
Ala12	-0.084	0.673	0.000	-0.144	0.081	0.000

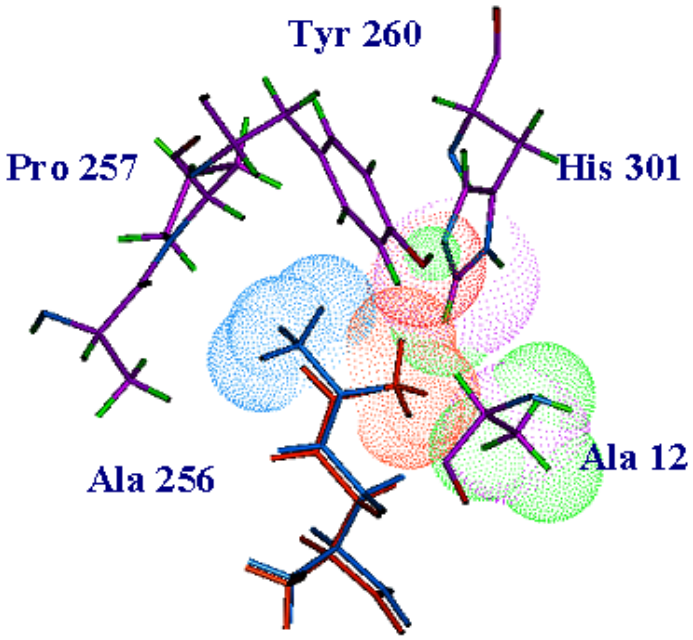


Figure 5: The binding modes or tcg and ccg, shown in the binding pocket of Met/MetRS(FF), were predicted by perturbation analysis at this site. Ccg (orange) has VDW clashes with Ala12, His301 and Tyr 260 and at the same time, *cis* orientation of the terminal methyl group created a void near Ala 256 and Pro 257. Tcg (pink) is shown to fill that void and also avoid the unfavorable VDW interactions.

MetRS has been observed to be extremely promiscuous and is able to incorporate substrates that are up to 340,000 folds poorer than Met. This could be attributed to the conformational flexibility of the active site of MetRS that has not been modeled in our simulation. The active site conformation could be different for different analogs. However, we have performed our perturbation studies only on the co-MetRS bound to the natural substrate. The active site flexibility may be important in enabling MetRS to activate Met analogs with varying side chain functionalities. One more consideration is that we are comparing our simulated binding energies to experimentally derived binding

energies that are further derived from ATP-PP_i exchange studies. ATP binding could have other structural effects on the enzyme that were not modeled in our simulations. However, it is interesting to note that we are able to get reasonably good correlation even with the limitations in the simulations. One can expect to gain more insights into the mechanism of this system with advancements in the simulation procedures.

4. Conclusions

We have studied the specificity of MetRS for Met in the first two steps of the binding process. We have demonstrated that its specificity increases in the second binding step where the enzyme undergoes a significant conformational change. We speculate that Met first anchors to residues Asp52 and His301 with its side chain and as the protein undergoes conformation change due to substrate binding (either the amino acid, ATP, or both), the cavity opens up and Met flips into the cavity. Multi-step binding mechanisms where the ligand-protein complexes display "induced fit" have been illustrated in other proteins. This has been attributed to the presence of energy gradients, or funnels, near the binding sites—the binding process initiates from a higher energy conformer and terminates in lower energy conformation (38).

When the structure to be docked is taken from the crystallized co-complex, predicting the fitted association is relatively straightforward as indicated by the docking study using Met/MetRS. Our study with the apo-MetRS illustrates that although determining the final bound conformation starting with the "free," "unbound" state of the

enzyme is extremely difficult, a refined search method can be applied to predict the correct binding region for the ligand. The predictions can be used to indicate the important residues in the binding regions that can be further tested by mutations studies. Therefore, for those enzyme crystal structures that are not co-crystallized with their substrates a powerful docking protocol, like HierDock can prove to be very useful in recognizing the binding region, even in cases where the protein is very flexible. If the molecules are relatively rigid and have smooth binding funnels with single or few minima, there is a higher likelihood that the docked conformation of the ligand in the "free," "unbound" state is the correct bound conformation since the conformational diversity of the protein is limited (39). But in the case of proteins that undergo significant conformation changes on associating with the ligand, it is unlikely that the predicted ligand plus protein complex would be the correct structure. In the case of a flexible protein, like MetRS, that has a larger conformational diversity, achieving a correct prediction bound conformation is complicated since the bound conformation could be very different from the free, unbound structure. However, the complex predicted with the apo enzyme should be regarded as an important "recognition mode" for the system, a key step in its multi-step binding process, since even at this stage of binding it could show some level of discrimination. In apo-MetRS, both docking and perturbation analysis indicate that in this conformation the enzyme is able to eliminate more than 60% of the natural amino acids. One could imagine that if the final bound complex after the change in conformation was the only filtering mechanism for an enzyme, each amino acid would first have to bind at this site, followed by the structural change in the enzyme and then get eliminated. Such a process would be both time-

consuming and energetically expensive for the enzyme. A first level of filter at the apo-enzyme conformation certainly seems to be more efficient screening mechanism adopted by flexible enzymes. It would be interesting to see how the procedure for binding site search performs in other apo-enzyme systems. We have already tested it for the predicting the binding site of Phe in *Thermus thermophillus* PheRS by scanning the entire apo-crystal structure of PheRS and have been able to find the correct binding site (unpublished results).

Binding site dynamics in enzyme brings in the question of enzyme specificity. An interesting observation about protein plasticity is that proteins displaying higher selectivity are also more rigid while those that more flexible can bind to a large number of substrates. Considering the conformational flexibility in the MetRS, as indicated by the substantial structural change in the co-crystal, it is not surprising that it is one of the more permissive aminoacyl-tRNA synthetases.

References

1. Ibba, M. & Soll, D. (2000) *Annu. Rev. Biochem.* **69,** 617-650.
2. Freist, W., Sternbach, H., Pardowitz, I. & Cramer, F. (1998) *J. Theor. Biol.* **193,** 19-38.
3. Freist, W. & Sternbach, H. (1988) *Eur. J. Biochem.* **177,** 425-433.
4. Freist, W., Sternbach, H. & Cramer, F. (1982) *Eur. J. Biochem.* **128,** 315-329.
5. Freist, W., Sternbach, H. & Cramer, F. (1987) *Eur. J. Biochem.* **169,** 33-39.
6. Freist, W., Sternbach, H. & Cramer, F. (1988) *Eur. J. Biochem.* **173,** 27-34.
7. Deming, T. J., Fournier, M. J., Mason, T. L. & Tirrell, D. A. (1997) *J. Macromol. Sci. Pure Appl. Chem.* **A34,** 2143-2150.
8. van Hest, J. C. M., Kiick, K. L. & Tirrell, D. A. (2000) *J. Am. Chem. Soc.* **122,** 1282-1288.
9. van Hest, J. C. M. & Tirrell, D. A. (1998) *FEBS Lett.* **428,** 68-70.
10. Kothakota, S., Mason, T. L., Tirrell, D. A. & Fournier, M. J. (1995) *J. Am. Chem. Soc.* **117,** 536-537.
11. Dougherty, M. J., Kothakota, S., Mason, T. L., Tirrell, D. A. & Fournier, M. J. (1993) *Macromolecules* **26,** 1779-1781.
12. Budisa, N., Steipe, B., Demange, P., Eckerskorn, C., Kellermann, J. & Huber, R. (1995) *Eur. J. Biochem.* **230,** 788-796.
13. Duewel, H., Daub, E., Robinson, V. & Honek, J. F. (1997) *Biochemistry* **36,** 3404-3416s.
14. Zhang, D. Q., Vaidehi, N., Goddard, W. A., Danzer, J. F. & Debe, D. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6579-6584.
15. Datta, D., Wang, P., Carrico, I. S., Mayo, S. L. & Tirrell, D. A. (2002) *J. Am. Chem. Soc.* **124,** 5652-5653.
16. Liu, D. R., Magliery, T. J., Pastrnak, M. & Schultz, P. G. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 10092-7.
17. Floriano, W. B., Vaidehi, N., Goddard, W. A., Singer, M. S. & Shepherd, G. M. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10712-6.
18. Kiick, K. L., Weberskirch, R. & Tirrell, D. A. (2001) *FEBS Lett.* **502,** 25-30.
19. Tannor, D. J., Marten, B., Murphy, R., Friesner, R. A., Sitkoff, D., Nicholls, A., Ringnalda, M., Goddard, W. A. I. & Honig, B. (1994) *J. Am. Chem. Soc.* **116,** 11875-11882.
20. Mechulam, Y., Schmitt, E., Maveyraud, L., Zelwer, C., Nureki, O., Yokoyama, S., Konno, M. & Blanquet, S. (1999) *J. Mol. Biol.* **294,** 1287-1297.
21. Mayo, S. L., Olafson, B. D. & Goddard, W. A., III (1990) *J. Phys. Chem.* **94,** 8897-8909.
22. Ghosh, A., Rapp, C. S. & Friesner, R. A. (1998) *J. Phys. Chem. B* **102,** 10983-10990.
23. Serre, L., Verdon, G., Choinowski, T., Hervouet, N., Risler, J. L. & Zelwer, C. (2001) *J. Mol. Biol.* **306,** 863-876.

24. Wang, P., Vaidehi, N., Tirrell, D. A. & Goddard, W. A. I. (2002) *J. Am. Chem. Soc. (in press)*.
25. Halperin, I., Ma, B. Y., Wolfson, H. & Nussinov, R. (2002) *Proteins: Struct. Funct. & Gene.* **47,** 409-443.
26. Wang, J. M., Morin, P., Wang, W. & Kollman, P. A. (2001) *J. Am. Chem. Soc.* **123,** 5221-5230.
27. Kollman, P. (1993) *Chem. Rev.* **93,** 2395-2417.
28. Ewing, T. A. & Kuntz, I. D. (1997) *J. Comput. Chem.* **18,** 1175-1189.
29. Ewing, T. J. A., Makino, S., Skillman, A. G. & Kuntz, I. D. (2001) *J.Comput. Aid. Mol. Design* **15,** 411-428.
30. Lim, K.-T., Brunett, S., Iotov, M., McClurg, R. B., Vaidehi, N., Dasgupta, S., Taylor, S. & Goddard, W. A., III (1997) *J. Comput. Chem.* **18,** 501-521.
31. Hendsch, Z. S. & Tidor, B. (1999) *Protein Sci.* **8,** 1381-92.
32. Connolly, M. L. (1983) *J. Appl. Cryst.* **16,** 548-558.
33. Stahl, M. & Bohm, H. J. (1998) *J. Mol. Graph. Model.* **16,** 121-132.
34. Bower, M. J., Cohen, F. E. & Dunbrack, R. L., Jr. (1997) *J. Mol. Biol.* **267,** 1268-82.
35. Ghosh, G., Pelka, H., Schulman, L. H. & Brunie, S. (1991) *Biochemistry* **30,** 9569-9575.
36. Fourmy, D., Mechulam, Y., Brunie, S., Blanquet, S. & Fayat, G. (1991) *FEBS Lett.* **292,** 259-263.
37. Kim, H. Y., Ghosh, G., Schulman, L. H., Brunie, S. & Jakubowski, H. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 11553-11557.
38. Zhang, C., Chen, J. & DeLisi, C. (1999) *Proteins: Struct. Funct. Gene.* **34,** 255-267.
39. Tsai, C. J., Kumar, S., Ma, B. Y. & Nussinov, R. (1999) *Protein Sci.* **8,** 1181-1190.

Chapter 3

Computational Simulations of Histidyl-tRNA Synthetases*

* This chapter is adapted from a paper to be submitted to JMB.

Abstract

Aminoacyl-tRNA synthetases play a very important role in the quality control of protein synthesis *in vivo*. This function is achieved by the unique recognition of the cognate amino acid and tRNA, in some cases with the help of proofreading against similar amino acids. Here we used the HierDock protocol to study the binding of 20 amino acids to two histidyl-tRNA synthetases from *E. coli* and *T. thermophilus* that have 3-D structures available. Ligand perturbation was also conducted to compare the binding affinity in the reaction mode. Both results show that histidyl-tRNA synthetases are able to differentiate their cognate ligand histidine from other amino acids in the binding stage. The docked conformation of histidine agreed well with the ligand binding conformation in the crystal structures.

1. Introduction

The selection of amino acids during protein biosynthesis is extremely precise. The recognition of the cognate amino acid takes place during the aminoacylation of tRNA, which is catalyzed by the aminoacyl-tRNA synthetases (AARSs). This is a two-step reaction, the amino acid first being activated by formation of aminoacyl adenylate and then transferred to the tRNA (1-3). In most cases the amino acid binds to its AARS so strong that no competitor exists in the reaction. However, there are cases that editing or proofreading mechanism is needed to guard against close competitors, such as the rejection of valine by isoleucyl-tRNA synthetases (4, 5). In every case the preferential binding of the correct amino acid has been optimized during evolution. The AARSs have binding sites that are unique to their respective amino acids. The active site generally contains hydrophilic residues that stabilize the zwitterion termini of the bound amino acid, while other residues interact with the side chain, depending on the nature of the amino acid. The specificity arises from the contacts of the side chain with the binding site conformation fitted for a given amino acid. With the sufficient positioning of the amino acid carboxylate, the facilitated reaction with a bound ATP structure is made possible, giving the anticipated aminoacyl-adenylate product.

Histidyl-tRNA synthetase (HisRS) belongs to the class II AARS (subgroup IIa) classified by their structural characteristics (6). Its substrate, histidine is one of the two standard genetically coded amino acids with heterocyclic aromatic side chain. The imidazole ring not only plays a role in stabilizing the structure of a protein by its aromatic

properties, but often has an important function in the catalytic centers of many enzymes, e.g., in acid-base catalysis, due to its unique pKa value of 6.00. Enzymatic decarboxylation of histidine generates the biologically important compound histimine. HisRS was also found to be one of the antigens in autoimmune diseases such as rheumatic arthritis (7). Hence structural study can help elucidate the epitopes that are responsible for recognition of HisRS by the autoimmune antibodies.

In this paper, we used the HierDock protocol (8) to dock all the 20 natural amino acids into the binding site of two HisRS's from *E. coli* (ecHisRS) and *T. thermophilus* (ttHisRS) that have three-dimensional structure available to date. The result showed a good correlation with the fact that the binding site is optimized for histidine binding, and the best histidine conformations from docking are 0.47 Å (ecHisRS) and 0.64 Å (ttHisRS) in rmsd from the ligands in the crystal structures. To further simulate the binding mode that histidine adopts when it is activated with ATP, we perturbed the histidine ligand into other 19 amino acids. The binding in this mode should allow us to compare the binding energies to the mischarging rate by HisRS. The result showed far better binding to histidine than any other amino acids.

2. Methods and Simulation Details

2.1 Structure Preparation

1HTT (resolution 2.6 Å) is a structure of the complex of ecHisRS and histidyl-adenylate (9) and 1ADJ (resolution 2.7 Å) is a structure of ttHisRS complexed with

histidine (10). Both structures were downloaded from the Protein Data Bank. In 1HTT, the adenylate was deleted from the structure to make the ligand as a histidine in order to compare with 1ADJ. Hydrogens were added to the structures using Biograf (MSI, San Diego, CA), and the structures were minimized with DREIDING force field (11) and Surface Generalized Born (SGB) implicit solvation (12) using MPSim (13). Conjugate gradient minimization method was employed for 2000 steps with a termination criterion of less than 0.1 kcal/mol/Å in rms force. The protein was described with CHARMM22 (14) charges, and the ligand charge was the Mulliken charge derived from Quantum Mechanical calculation.

The amino acids were build in Biograf and optimized in Jaguar 4.0 (Schrödinger, Portland, OR) with basis set 6-31G** under Poisson-Boltzmann continuum dielectric solvent (15). The electron density from molecular orbitals were fit into atom centers to obtain the Mulliken charges, which were used for the amino acids in force field calculations. Because histidine has two possible protonation states at neutral pH, they were treated as two different ligands for docking. The δ-protonated form is labeled as Hsd, and the ε-protonated form as Hse. Depending on the local environment, both forms are commonly seen in natural proteins. The third state, which is doubly protonated, is very rare at neutral pH. Therefore, we did not consider it in docking.

The binding energy for each ligand is given by

$$-\Delta\Delta G_{calc} = \Delta G(protein) + \Delta G(ligand) - \Delta G(protein + ligand). \qquad (1)$$

Since the structure optimizations included solvation forces using the SGB continuum solvent approximation with the experimental dielectric constant, we consider that the calculated energies are free energies (16).

2.2 HierDock Protocol

The HierDock protocol has been shown to efficiently dock small ligands to proteins with or without the knowledge where the binding site is (8, 17). It is based on DOCK 4.0 (18) and coupled with fine grain molecular mechanics technique. It can be divided into three steps as follows:

1. *Mapping of possible binding regions*. A probe radius of 1.4 Å is used to trace a 4 dots/Å negative image of the protein molecular surface, according to Connolly's method (19). Clusters of overlapping spheres are generated from negative image with the SPHGEN program (18). These spheres serve as the basis for the docking method.
2. *Definition of docking region*. The pockets of empty space of the receptor surface represented by spheres are divided into many 10 Å × 10 Å × 10 Å overlapping cubes, which cover the entire protein surface. Each region is scanned to determine its suitability as a binding site. The site that overwhelmingly contains the greatest number of lowest energy docked conformations is designated as the most probable binding region.
3. *Generation of docked conformations for the ligand-receptor complex*. The orientations of the ligand in the receptor are generated by DOCK 4.0, using flexible docking with torsional minimization of the ligand, a continuum

dielectric of 1.0 and a distance cutoff of 10 Å for the evaluation of energy. 1000 conformations are generated and ranked according to the DOCK 4.0 scoring function, and the top 100 structures are kept for further optimization.

4. *MPSim optimization of the complexes*. The top-ranking DOCK structures are then subjected to further optimization, using a more accurate full-atom forcefield with SGB solvation. The first stage of gas phase optimization utilizes a fully flexible ligand with a fixed protein, followed by a single point energy calculation of the solvation using dielectric of 2.0. A buried surface calculation for the ligands with a minimum threshold of 75% selects only those structures that are sufficiently buried within the protein. The 10 lowest energy conformations undergo further all-atom gas optimization with a single-point solvation energy calculation to screen for the best binding candidate.
5. *Selection of the most probable binding site and best configurations.* The conformations with the lowest energy score are selected and assumed to demonstrate preferential binding to the region.
6. *Docking of ligand pool into the binding site.* Steps 3—5 are repeated for each member of the ligand pool to obtain relative binding affinities.
7. *Ranking of ligand affinities*. The relative binding energies for the best ligand conformations are defined as the difference between the ligand in protein versus in solution. The amino acids can then be ranked according to binding affinities to determine which ligands have the highest affinity for the binding site.

Because the binding sites were well defined in both ecHisRS and ttHisRS, the binding site scanning step was skipped and only the region containing the histidine ligand in the crystal structure was used in further docking.

2.3 Binding Energy Calculation of the 20 Natural Amino Acids in the Active Conformation by Ligand Perturbation

HierDock protocol predicts the best energy conformation for each ligand (20 natural amino acids) in the defined binding region in HisRS structure. These predictions give rise to different preferred binding conformation for each ligand. But if the zwitterions part is positioned different from what is necessary for catalysis, the amino acid will not be charged to AMP even though it might bind to the protein. So it is necessary to assess the relative binding energies of the twenty natural amino acids and their analogs in the activation mode. To generate the conformation of 20 natural amino acids in the activation mode we performed the following steps:

- An amino acid rotamer library (20) was used to generate all the conformations of each amino acid in the binding site. The best rotamer was chosen by matching each rotamer k in the binding site and evaluated with the following equation using the Dreiding force field:

$$E_k = \sum_{i,j}\left(\frac{q_i q_j}{4\pi\varepsilon r_{ij}} + D_e\left(\left(\frac{r_m}{r_{ij}}\right)^{12} - 2\left(\frac{r_m}{r_{ij}}\right)^{6}\right) + D_{HB}\left(5\left(\frac{r_{HB}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{HB}}{r_{ij}}\right)^{10}\right)\cos^4\theta\right), \quad (2)$$

where i and j sum over all atoms in the ligand and protein residue residues in the binding site, q_i and q_j are partial charges on atoms i and j, respectively. r_{ij} is the distance between atoms i and j, and r_m and D_e are van der Waals distance and well

depth of atoms i and j, r_{HB} and D_{HB} are hydrogen bond distance and well depth, respectively. θ is the hydrogen bond angle between atoms i, j and their bridging hydrogen atom. The hydrogen bond term is only evaluated for hydrogen bond donor and acceptor atoms. To avoid over penalizing clash, the van der Waals radii were reduced to 90% of the standard values in the Dreiding force field.

- After the best rotamer was chosen for each ligand, the total energy was minimized in the presence of protein, and the binding energy was then calculated using equation (1) for each of the twenty natural amino acids in the "activation mode" and compared.

3. Results and Discussions

3.1 Docking of Histidine Ligands

The ecHisRS and ttHisRS proteins were minimized with SGB solvation using MPSim. The rmsd between the crystal and minimized structures were 0.71 Å for ecHisRS and 0.64 Å for ttHisRS. These values were well below the resolution of the crystal structures, which demonstrated that our choice with the DREIDING force field and CHARMM22 charges were suitable for describing proteins. Similar results were obtained in other protein simulations using the same setup (17).

Two forms of histidine, δ-protonated Hsd and ε-protonated Hse, were tried to dock into the binding site of ecHisRS and ttHisRS. It turned out that Hse was the form of choice in both cases. The hydrogen bonding network analysis on the crystal structures

using WHATCHECK (21) gave the same result. Figure 1 showed the interaction between Hse ligand and the protein in the binding sites for (a) ecHisRS and (b) ttHisRS. The rmsd for the docked Hse and the ligand in the crystal structures was 0.47 Å for ecHisRS and 0.28 Å for ttHisRS. The calculated binding energy to ecHisRS was 81.6 kcal/mol for the docked Hse, compared to 92.4 kcal/mol for the ligand in the crystal structure. For ttHisRS, it was 73.5 kcal/mol for the docked Hse versus 69.6 kcal/mol for the crystal ligand. Table 1 listed all the hydrogen bond interactions and the distances between the docked Hse and the protein in comparison with the crystal structure.

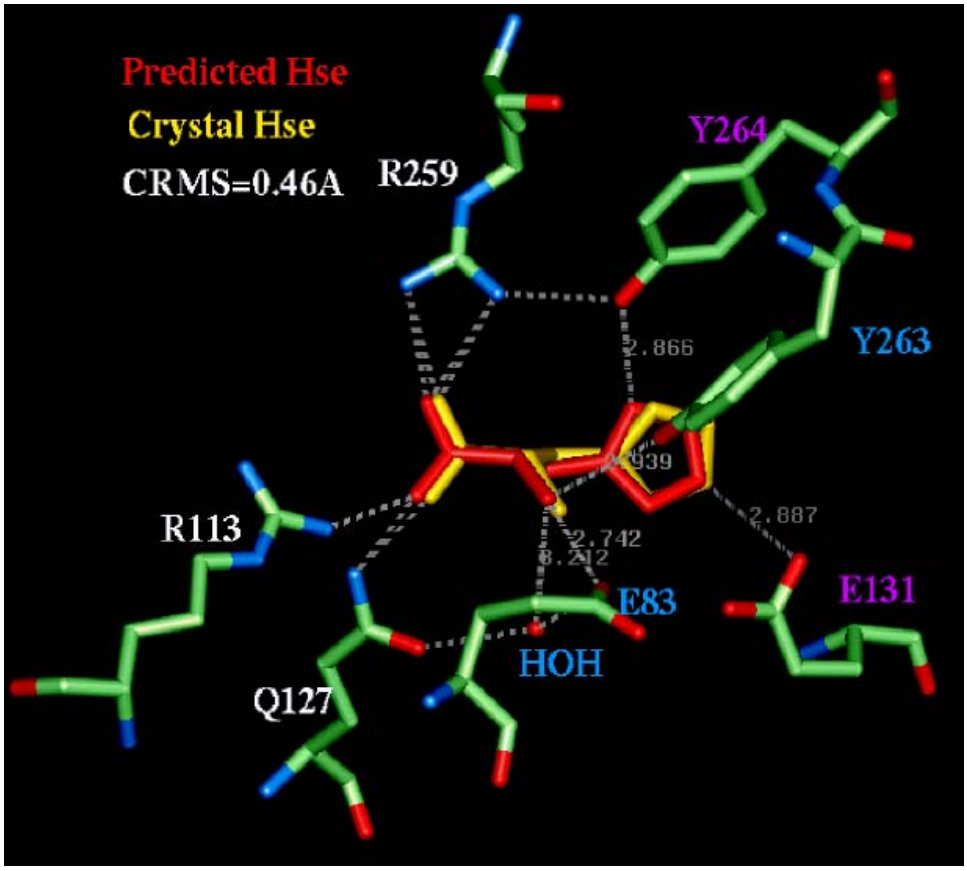


Figure 1. (a) Interactions between the ligand Hse and ecHisRS in the binding site

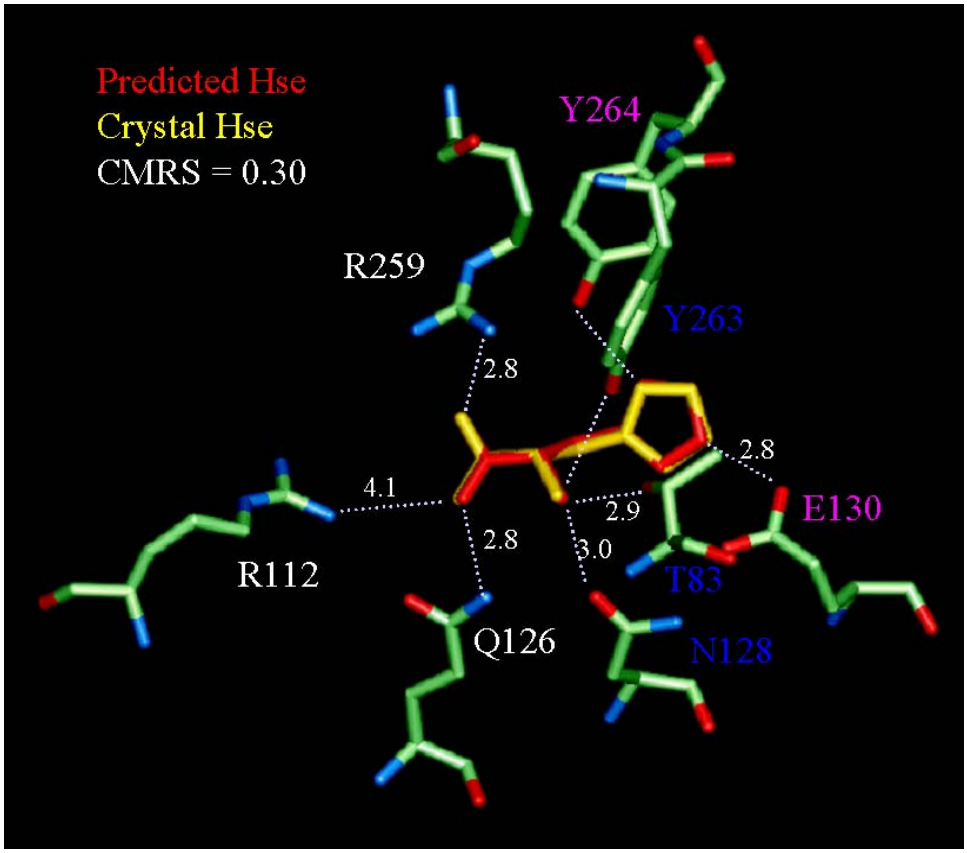


Figure 1. (b) Interactions between the ligand Hse and ttHisRS in the binding site

3.2 Docking of all other 19 natural amino acids

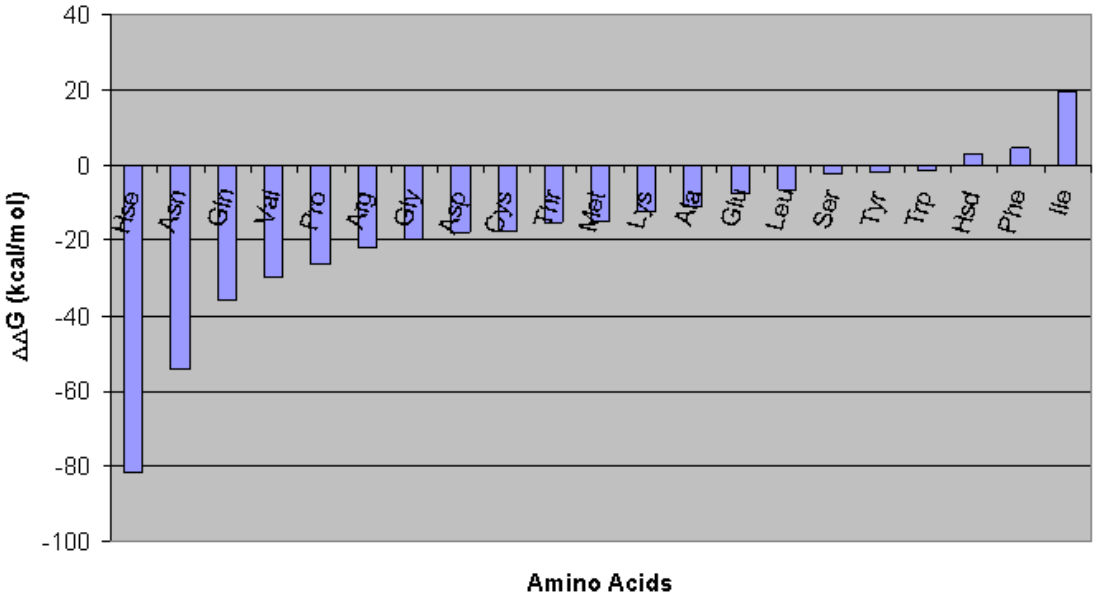
Using the same HierDock protocol, all other 19 natural amino acids were docked into the histidine binding sites of ecHisRS and ttHisRS. Figure 2 showed the binding energies of these docked amino acids along with Hse and Hsd. These results suggest that both ecHisRS and ttHisRS have a binding site optimal for Hse binding, because none of the natural amino acids has a binding energy close to Hse. This is consistent with our current understanding of HisRS's. None of the HisRS's known to date has shown any editing mechanism.

Table 1. The hydrogen bond interactions between the Hse ligand and HisRS from docking comparing to crystal structures. The distances in ( ) for the crystal structures are from the structures minimized with force field.

Ligand Atom	Protein Atom (ecHisRS)	Distance in docked structure (Å)	Distance in crystal structure (Å)
N^{δ}	Tyr32 O^{η}	2.87	2.70 (2.89)
N^{ε}	Glu131 $O^{\varepsilon 1}$	3.13	2.78 (3.31)
N	Glu83 $O^{\varepsilon 1}$	2.72	2.95 (2.64)
N	Tyr263 O^{η}	2.92	3.32 (3.09)
N	Water O	3.16	2.62 (3.64)
O	Arg113 $N^{\eta 2}$	3.15	3.84 (3.45)
O	Gln127 $N^{\varepsilon 2}$	2.88	3.43 (2.93)
O^{XT}	Arg259 $N^{\eta 2}$	2.83	3.16 (2.87)
Ligand Atom	Protein Atom (ttHisRS)	Distance in docked structure (Å)	Distance in crystal structure (Å)
N^{δ}	Tyr264 O^{η}	2.89	2.57 (2.91)
N^{ε}	Glu130 $O^{\varepsilon 1}$	2.82	2.68 (2.82)
N	Thr83 $O^{\gamma 1}$	2.91	2.63 (2.91)
N	Tyr263 O^{η}	3.69	2.67 (3.68)
N	Asn $O^{\delta 1}$	3.11	2.98 (3.00)
O	Arg112 $N^{\eta 2}$	4.12*	4.82 (3.96)*
O	Gln126 $N^{\varepsilon 2}$	2.78	3.91 (2.90)
OXT	Arg259 $N^{\eta 2}$	2.89	3.65 (2.87)

* Water mediated hydrogen bond

a)



b)

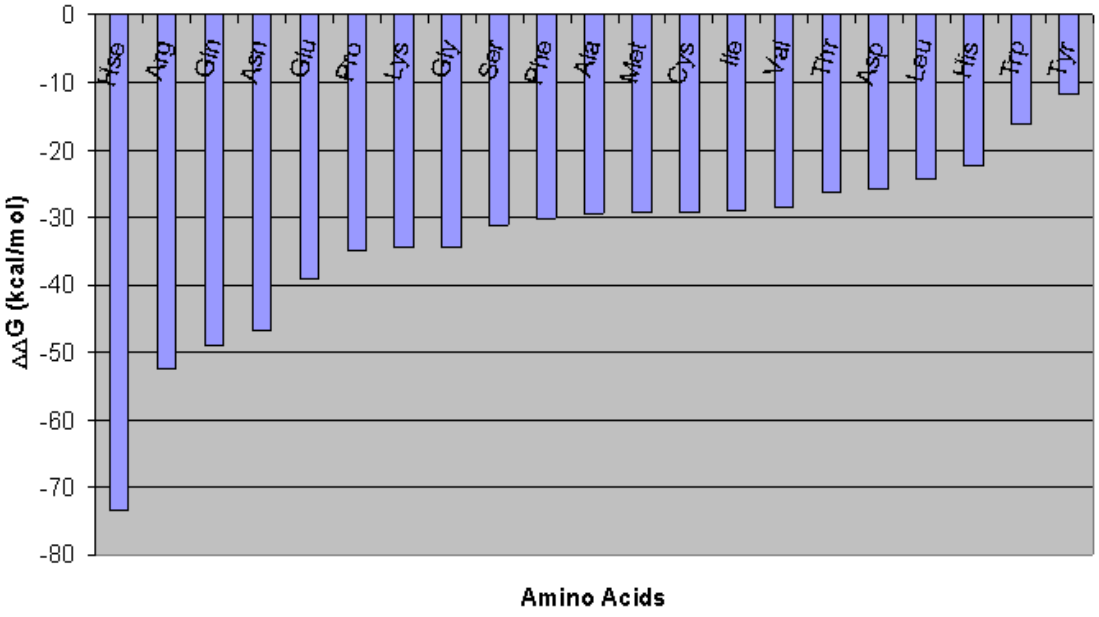
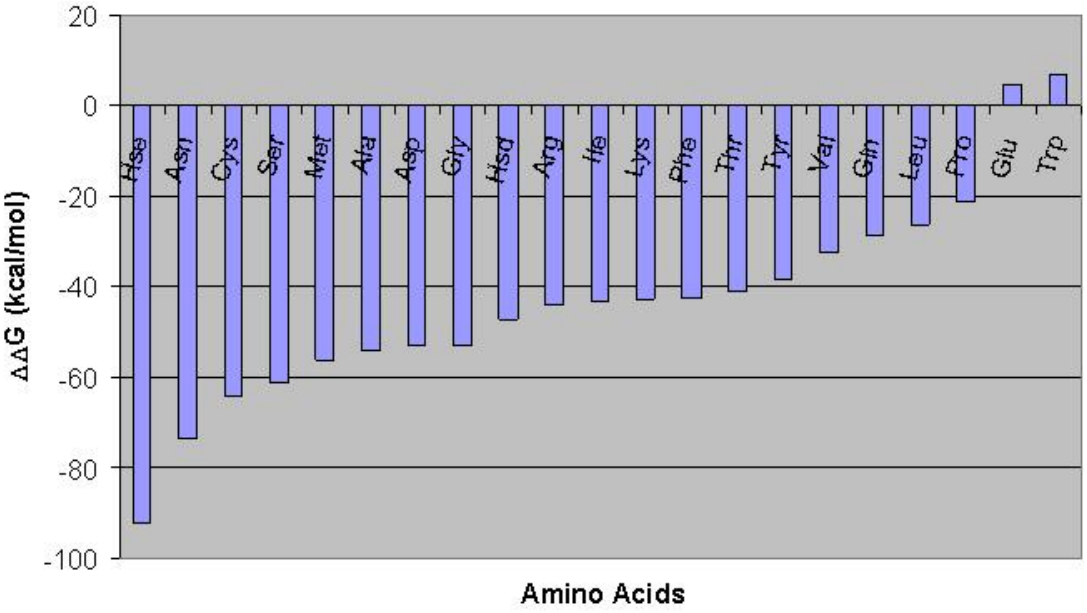


Figure 2. Binding free energies of all 20 docked amino acids to (a) ecHisRS and (b) ttHisRS.

a)



b)

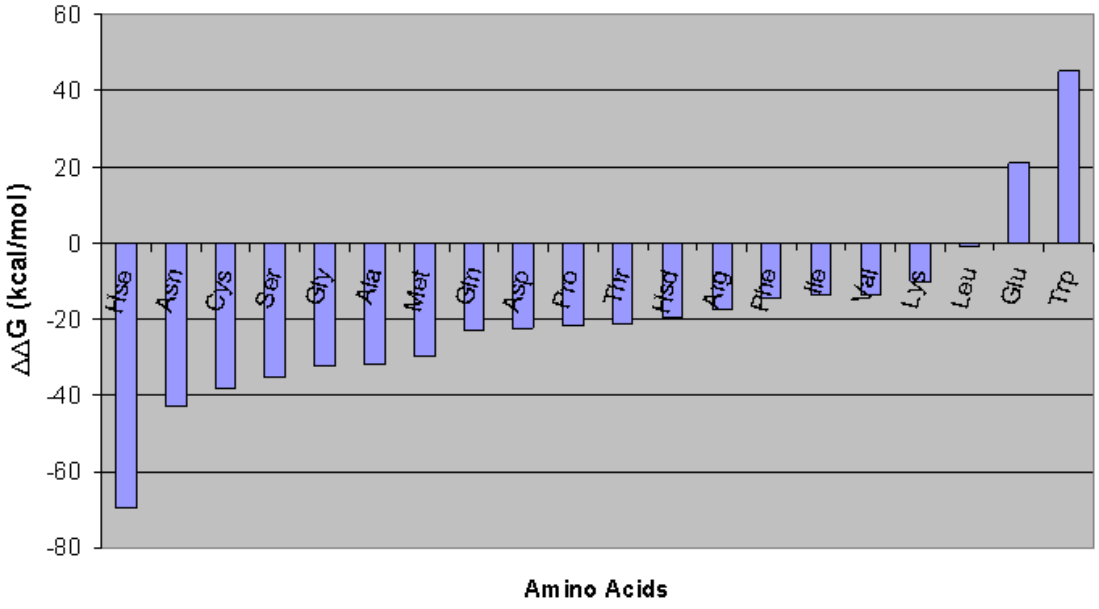


Figure 3. Binding free energies of all 20 amino acids to (a) ecHisRS and (b) ttHisRS by perturbing the crystal Hse ligand.

3.3 Perturbation of Hse in the binding site

Because only the binding mode that places the zwitterions of the other 19 amino acids the same position as histidine binding can lead to misactivation, it is useful to look at the competitive binding of all amino acids in that binding mode. This binding can be simulated with amino acid perturbation in the binding site, in which the zwitterions are fixed, and the side chain is mutated into other resides. Figure 3 showed the binding energies of all 20 amino acids to (a) ecHisRS and (b) ttHisRS from perturbation. In the case of ecHisRS, there is no real competition from other amino acids, while Asn and Thr stand out from the rest. For ttHisRS, Asn seems to be a strong competitor. Figure 4 shows the overlaid Asn and Hse in the activation binding mode. The COO^- moiety of Asn is slightly further (0.4 Å) into the binding site than Hse. Because the aminoacylation reaction requires the COO^- moiety positioned exactly, Asn will be less effective in misactivation.

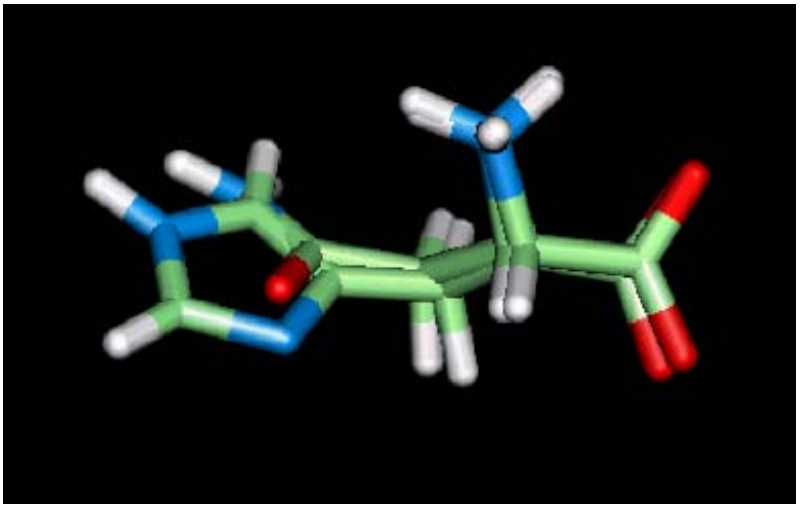


Figure 4. Comparison between Hse and Asn in the activation binding mode.

4. Conclusions

We have used the HierDock protocol to successfully predict the binding site and binding modes of histidine in ecHisRS and ttHisRS. The rmsd of between the docked Hse ligand and the crystal structure Hse ligand is 0.47 Å for ecHisRS and 0.64 Å for ttHisRS. Other 19 natural amino acids have also been docked into the binding site of HisRS's, and the result shows that Hse binds to HisRS significantly tighter than any other natural occurring amino acids in both ecHisRS and ttHisRS.

Ligand perturbation from Hse has also been performed to calculate the binding energies of 20 amino acids in the activation mode for aminoacylation. These results also show that there is almost no competition from other amino acids in binding to HisRS's. These results are consistent with our current understanding that HisRS's use ligand binding as the sole mechanism in amino acid selection.

References

1. Ibba, M. & Soll, D. (1999) *Science* **286,** 1893-1897.
2. Ibba, M. & Soll, D. (2000) *Annu. Rev. Biochem.* **69,** 617-650.
3. Freist, W., Sternbach, H., Pardowitz, I. & Cramer, F. (1998) *J. Theor. Biol.* **193,** 19-38.
4. Fersht, A. R. & Dingwall, C. (1979) *Biochemistry* **18,** 2627-2631.
5. Nureki, O., Vassylyev, D. G., Tateno, M., Shimada, A., Nakama, T., Fukai, S., Konno, M., Hendrickson, T. L., Schimmel, P. & Yokoyama, S. (1998) *Science* **280,** 578-582.
6. Eriani, G., Delarue, M., Poch, O., Gangloff, J. & Moras, D. (1990) *Nature* **347,** 203-206.
7. Mathews, M. B. & Bernstein, R. M. (1983) *Nature* **304,** 177-179.
8. Floriano, W. B., Vaidehi, N., Goddard, W. A., Singer, M. S. & Shepherd, G. M. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10712-6.
9. Arnez, J. G., Harris, D. C., Mitschler, A., Rees, B., Francklyn, C. S. & Moras, D. (1995) *EMBO J.* **14,** 4143-4155.
10. Aberg, A., Yaremchuk, A., Tukalo, M., Rasmussen, B. & Cusack, S. (1997) *Biochemistry* **36,** 3084-3094.
11. Mayo, S. L., Olafson, B. D. & Goddard, W. A., III (1990) *J. Phys. Chem.* **94,** 8897-8909.
12. Ghosh, A., Rapp, C. S. & Friesner, R. A. (1998) *J. Phys. Chem. B* **102,** 10983-10990.
13. Lim, K.-T., Brunett, S., Iotov, M., McClurg, R. B., Vaidehi, N., Dasgupta, S., Taylor, S. & Goddard, W. A., III (1997) *J. Comput. Chem.* **18,** 501-521.
14. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998) *J. Phys. Chem. B* **102,** 3586-3616.
15. Tannor, D. J., Marten, B., Murphy, R., Friesner, R. A., Sitkoff, D., Nicholls, A., Ringnalda, M., Goddard, W. A. I. & Honig, B. (1994) *J. Am. Chem. Soc.* **116,** 11875-11882.
16. Baker, N. A., Sept, D., Holst, M. J. & McCammon, J. A. (2001) *Abstracts of Papers of the American Chemical Society* **221,** U437-U437.
17. Wang, P., Vaidehi, N., Tirrell, D. A. & Goddard, W. A. I. (2002) *J. Am. Chem. Soc. (in press)*.
18. Ewing, T. J. A., Makino, S., Skillman, A. G. & Kuntz, I. D. (2001) *J.Comput. Aid. Mol. Design* **15,** 411-428.
19. Connolly, M. L. (1983) *J. Appl. Cryst.* **16,** 548-558.
20. Bower, M. J., Cohen, F. E. & Dunbrack, R. L., Jr. (1997) *J. Mol. Biol.* **267,** 1268-82.

21. Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996) *Nature* **381,** 272-272.

Chapter 4

Computational Studies of the Recognition of Isoleucine and Fluorinated Analogs by Isoleucyl-tRNA Synthetase

Abstract

Protein biosynthesis has an unmatched accuracy and aminoacyl-tRNA synthetases are responsible for the recognition fidelity to large extent. It has been demonstrated in lab that some non-natural amino acids can be incorporated using the wild-type apparatus of *in vivo* protein synthesis. Recently, two tri-fluorinated isoleucines were tested using the wild-type $tRNA^{Ile}$:IleRS apparatus to make proteins containing fluorinated non-natural amino acids. But only one tri-fluorinated isoleucine was incorporated successfully.

In this paper, we simulated the binding of isoleucine and three fluorinated isoleucine analogs to isoleucyl-tRNA synthetase under surface generalized Born continuum electrostatic solvent. We found that while all three analogs have van der Waals clash with the binding site, the clash is less severe for the γ-tri-fluorinated Ile. The result showed that the δ_1-tri-fluorinated Ile analog binds IleRS 7.4 kcal/mol less than Ile, while the γ_2-tri-fluorinated Ile analog binds 1.8 kcal/mol less than Ile. This is consistent with the experimental result. A third analog, the hexa-flurinated Ile, which has not been experimentally tested, showed slightly better binding affinity than the γ_2-tri-fluorinated Ile. We performed a component analysis to show that the γ_2-methyl binding region in IleRS is slightly bigger and more adaptive to bigger binding group than the δ_1-methyl binding region. Furthermore, it facilitates a bigger binding region for the δ_1-methyl group. Solvation is another factor against the binding of the δ_1-tri-fluorinated Ile analog.

The δ_1-methyl binding region is more polar than the γ_2-methyl binding region. As a result, the more hydrophobic CF_3 group goes to the γ_2-methyl binding region easier than the δ-methyl binding region.

1. Introduction

Protein biosynthesis is a precisely controlled mechanism. It has been known for a long time that aminoacyl-tRNA synthetases (AARSs) are responsible for the accurate recognition of cognate amino acid-tRNA pairs. There is also a recognition process for codon-anticodon match up in the ribosome, but the mechanism is rather straightforward. On the other hand, AARSs achieve the selection by a multiple step process, including the binding selection and pre- and post-transfer "proofreading" mechanisms (1). There are up to four steps involved in the selection (2):

(1) Binding of amino acid and ATP
(2) Conformational change in the AARS induced by binding and formation of aminoacyl-adenylate complex.
(3) Proofreading of misactivated non-cognate aminoacyl adenylate complex
(4) Transfer of aminoacyl to the 3' end of the tRNA and proofreading

However, not every AARS use all four steps for amino acid selection.

Isoleucyl-tRNA synthetase (IleRS) is one of the first AARSs that have been known to have a proofreading or editing mechanism in guarding against misactivation of non-cognate amino acids (3). The term "double sieving" illustrates the way that IleRS selects isoleucine (Ile) over other natural amino acids (4). In the binding step 1 IleRS rejects any amino acid that has a larger side chain than Ile. In the proofreading step 3 IleRS hydrolyzes any aminoacyl adenylate complex with a side chain smaller than Ile,

such as Val, Ala, etc. (5). The proofreading requires the presence of $tRNA^{Ile}$ in the binding site in order to transfer the intermediate product to the proofreading site (6).

Protein biosynthesis has many advantages over the traditional polymer synthesis, such as well-defined chain length, sequence and fold. There have been a lot of efforts trying to use *in vivo* protein synthesis to make proteins containing non-natural amino acids (7-20). Recently the Tirrell lab at Caltech tried to incorporate two tri-fluorinated Ile analogs to protein *in vivo* using IleRS. By analyzing the products they found that only one of the analogs was incorporated while the other was not. Because the proofreading mechanism is to hydrolyze aminoacyl adenylate that has a smaller side chain than Ile, it is not expected to hydrolyze any fluorinated aminoacyl adenylate formed due to the bulkier size of the CF_3 group than CH_3. There is much evidence that showed that once a non-natural amino acid is charged into an tRNA, it almost certainly will be put into a protein in the ribosome (21). Therefore, the initial binding step seems to be the only step to govern the selection here.

Computer simulation offers a great opportunity in understanding the selection of ligand binding in the molecular level. The crystal structure of *E. coli* IleRS co-crystallized with Ile in the binding site has been solved at a resolution of 2.7 Å (Cusack, personal communications). This can be used as a starting structure in simulation and the analogs can be perturbed from Ile easily. Molecular dynamics simulations can provide a good binding free energy change when the change in ligands is small. Also models based on continuum electrostatics can often provide semi-quantitative binding free energies

(22). Another important advantage computer simulations offer is that the overall binding free energy can be easily decomposed into different components. This can be valuable information when manipulating the activity of AARSs by mutation or protein engineering as well as understanding the selection mechanism (23).

In this paper, we present the results of simulating the binding of Ile and three fluorinated Ile analogs to IleRS using the surface generalized Born (SGB) continuum electrostatic solvation model (24). SGB is a good approximation to the Poisson-Boltzmann (PB) continuum solvation model, and is significantly faster than PB. The result showed that the δ_1-tri-fluorinated Ile analog (Idf) binds IleRS 0.3 kcal/mol less than Ile, while the γ_2-tri-fluorinated Ile analog (Igf) binds 4.5 kcal/mol less than Ile. This is consistent with the experimental result that Idf was incorporated, while Igf was not. We performed a component analysis to show that the δ1-methyl binding region in IleRS is slightly bigger and more adaptive to bigger binding group than the γ2-methyl binding region. Val, known to be misactivated by IleRS and subsequently edited out by the proofreading mechanism, and hexa-fluorinated Ile analog (Ihf) have also been simulated.

2. Methods and Simulation Details

The crystal structure of IleRS from *E. coli* co-crystallized with Ile in the binding site was obtained from Prof. Cusack. Hydrogens were added to the structure using Biograf (Molecular Simulations, San Diego, CA). The structure was first annealed with heavy atoms fixed to optimize the hydrogen bonds of the structure. The structure was

then minimized using conjugate gradient method for 2000 steps under SGB continuum solvation using MPSim (25). Cell multiple method (26) was used to calculate the nonbond interactions. Dreiding force field (27) was used in energy expression. Protein was described with CHARMM22 (28) charges. The charges for Ile and Ile analogs were Mulliken charges by fitting the molecular orbitals to atom centers in quantum mechanics. Jaguar 4.0 (Schrödinger, Portland, OR) was used to run the calculation with 6-31G** basis set under Poisson-Boltzmann continuum dielectric solvent (29). The optimized IleRS-Ile complex showed a rmsd of 0.30 Å from the original crystal structure.

The Ile ligand was taken out and mutated into three fluorinated Ile analogs (Figure 1). An annealing dynamics was performed on each analog to find the best conformation of the side chain in the binding site of IleRS. The free energy of each ligand complexed with IleRS was calculated by minimizing the complex under SGB continuum electrostatic solvent. A dielectric constant of 2 for protein, and 78.2 for solvent was used in the simulations. The probe radius of solvent was 1.4 Å. The free energy of the ligand and IleRS alone were also calculated this way, and the binding free energy was calculated using

$$-\Delta\Delta G_{binding} = \Delta G(protein) + \Delta G(ligand) - \Delta G(protein + ligand). \qquad (1)$$

Generally the free energy calculated this way omits the entropy contributions from translations and rotations, so they can only be compared relative to different ligands (22, 30). Nonetheless, the differential value between these binding free energies can be compared to experimental values.

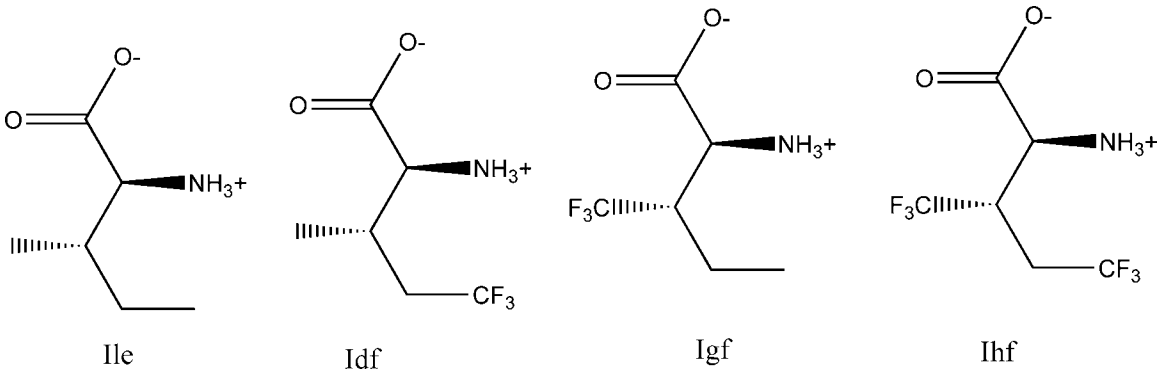


Figure 1. Structure of Ile and three fluorinated analogs used in this simulation.

A component analysis was performed as previously described (23). It is the same technique in principle as in reference (22). The following equation was used to calculate the interaction energy between the ligand and residue *k* in the binding site:

$$E_k = \sum_{i,j}\left(\frac{q_i q_j}{4\pi\varepsilon r_{ij}} + D_e\left(\left(\frac{r_m}{r_{ij}}\right)^{12} - 2\left(\frac{r_m}{r_{ij}}\right)^{6}\right) + D_{HB}\left(5\left(\frac{r_{HB}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{HB}}{r_{ij}}\right)^{10}\right)\cos^4\theta\right), \quad (2)$$

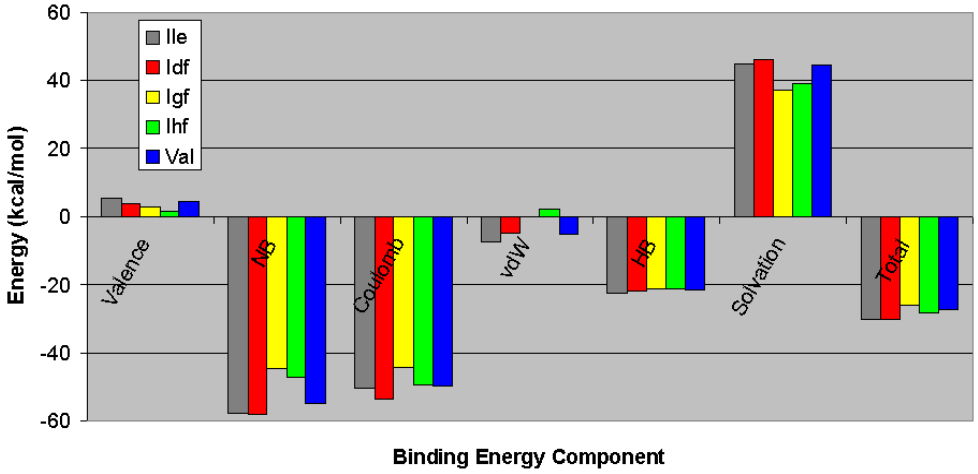
where *i* and *j* sum over all atoms in the ligand and protein residue *k* in the binding site, q_i and q_j are partial charges on atoms *i* and *j*, respectively. r_{ij} is the distance between atoms *i* and *j*, and r_m and D_e are van der Waals distance and well depth of atoms *i* and *j*, r_{HB} and D_{HB} are hydrogen bond distance and well depth, respectively. θ is the hydrogen bond angle between atoms *i*, *j* and their bridging hydrogen atom. The hydrogen bond term is only evaluated for hydrogen bond donor and acceptor atoms.

3. Results and Discussions

The binding free energies of Ile, Val, and three fluorinated Ile analogs (Idf, Igf, and Ihf, see Figure 1) to IleRS were calculated under SGB continuum electrostatic solvent. The contributions from each free energy component (valence, vdW, coulomb, hydrogen bond and solvation) were also calculated using Equation 1. These results were plotted in Figure 2. Please note that the nonbond (NB) energy is the sum of vdW and Coulomb energy.

It is apparent from Figure 2 b that there are three major differences in binding free energy components: Coulomb, vdW and solvation energies. In the case of Idf, vdW and Coulomb interactions play different roles. Coulomb energy favors the binding of Idf, while vdW disfavors it. These two interactions cancel out, as a result there is no net contribution from NB energy. Other energies seem to cancel out as well for Idf. Therefore, the total binding energy of Idf is almost the same as Ile. For Igf, both Coulomb and vdW interactions disfavor Igf, leading to a NB energy of 13 kcal/mol worse than Ile. Although solvation favors Igf binding, it is not enough to offset the NB energy. As a result, Igf is the worst binding ligand among the five ligands we studied here.

a)



b)

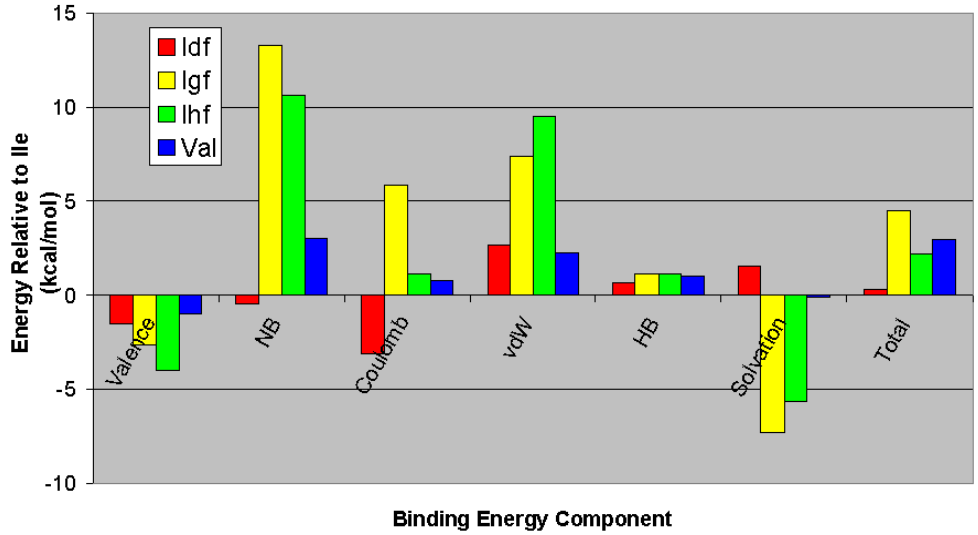


Figure 2. (a) The binding energies of Ile, Val, and three fluorinated Ile analogs to IleRS decomposed into components. (b) Same as in (a), but relative to Ile.

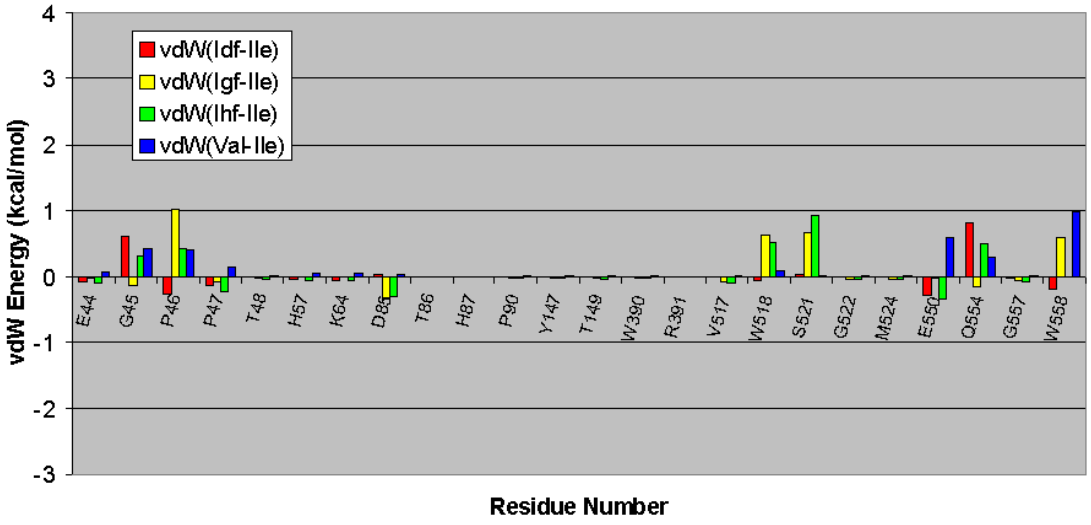
A third hexa-fluorinated Ile analog Ihf, which has not been tested experimentally due to difficulty in synthesis, shows a binding energy between Idf and Igf. Although the vdW energy strongly disfavors Ihf, the Coulomb energy only slightly disfavors it. The valence part in the binding, which is the response to the vdW clash in the binding site,

favors Ihf binding by 4 kcal/mol compared to Ile. As expected, Ihf is more hydrophobic than Ile, leading to a favorable solvation contribution than Ile by 5.5 kcal/mol. Ihf is a case we are trying to predict, and it is useful to compare it with Val. It has been known that Val binds IleRS 200 times weaker than Ile, and it translates into 3 kcal/mol binding energy difference between Ile and Val (31). This is in agreement with our result here. But Val is also known to be misactivated by IleRS, and Ihf shows 1 kcal/mol better than Val, therefore we predict that Ihf WILL be incorporated by IleRS *in vivo*.

A component analysis was also performed using Equation 2 to compare the contributions from each individual residue in the binding site (Figure 3a-d). Figure 3 a shows the vdW contributions from each residue with 5 Å of Ile in binding each ligand compared to Ile. Residues G45 and Q554 show strong clash with Idf, but other residues such as P46, P47, E550 and W558 actually compensates part of the clash by having favorable interaction with Idf. Please note that these interactions are with the ligand directly and it does not consider the propagation of the clash into protein-protein interaction, thus they don't add up to the value in Figure 2 b. But they should be proportional to each other. For Igf, the clashes are with different residues, because the CF_3 is on different binding region from Idf. There are four residues that show severe clash with Igf, and they are P46, W518, S521, and W558. G45, D85 and Q554 have favorable vdW interaction with Igf. It is an interesting observation that most of the clash and favorable interaction residues are complimentary between Idf and Igf (Figure 3 a). For most residues, the vdW interaction with Ihf is just the addition of Igf and Idf. For

Val, the unfavorable vdW is from not making some of the contacts with protein due to the lack of a CH_2 group compared to Ile.

a)



b)

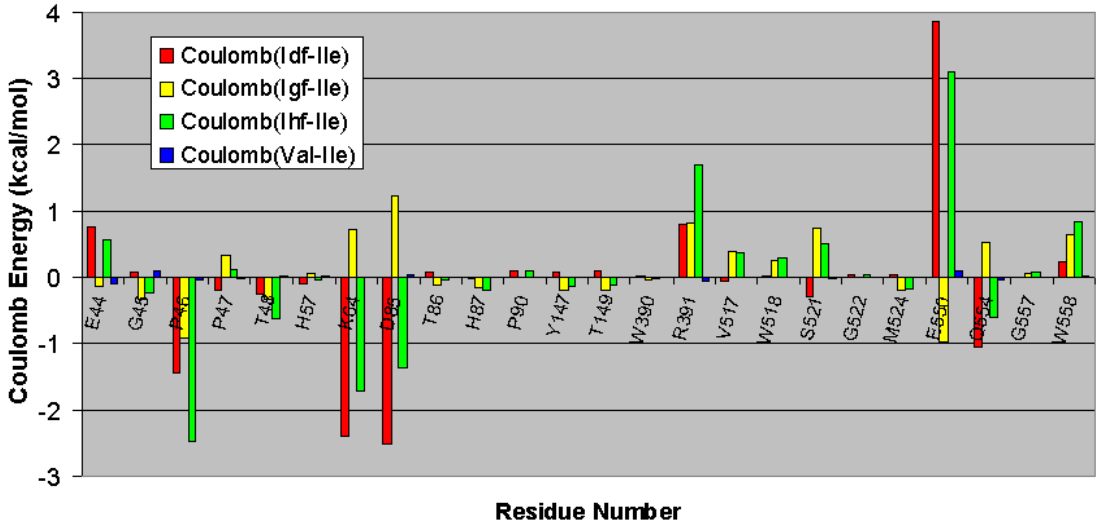
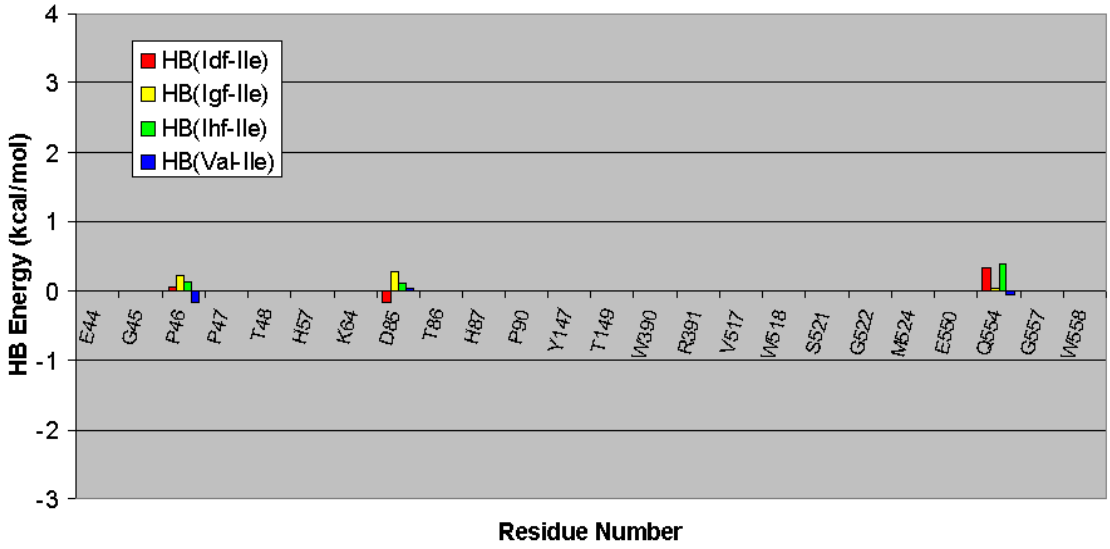


Figure 3. The component analysis for each residue in the binding site with ligands compared to Ile. (a) vdW interaction. (b) Coulomb interaction.

For Coulomb interactions with Idf, E550 strongly disfavors Idf by almost 4 kcal/mol compared to Ile, but it is compensated by favorable interactions with P46, K64, D85 and Q554. The net contribution from Coulomb interaction is favorable for Idf binding. This is consistent with the conclusion we drew from Figure 2 b. For Igf, there is no dominating Coulomb interaction from a single residues, but it seems that there are more residues with unfavorable interactions (K64, D85, R391, S521, Q554 and W558) than residues with favorable interactions (P46 and E550 only). Again, the Coulomb energies for Ihf can be obtained by the addition of Idf and Igf. And the complementary rule holds for most residues except P46, which has favorable interactions for both Idf and Igf. As a result of the addition, Ihf is strongly favored by the Coulomb interaction with P46. This can be explained by the fact that P46 is located close to the C^{β} atom of the ligands. The electron withdrawing effect of CF_3 leaves the C^{β} with more positive charges in all three cases, and is favored by the oxygen atom of the main chain carbonyl group of P46. For Val, there is very little change in the Coulomb interactions compared to Ile.

The hydrogen bonding interactions show very little change in all four ligands compared to Ile (Figure 3 c). This implies that the zwitterions of all different ligands position equally well, and the interactions with protein are strong (Figure 2 a). It also eases the concern that the binding mode might be different for some of the ligands, therefore these ligands might not be activated even though they have a good binding energy with protein.

c)



d)

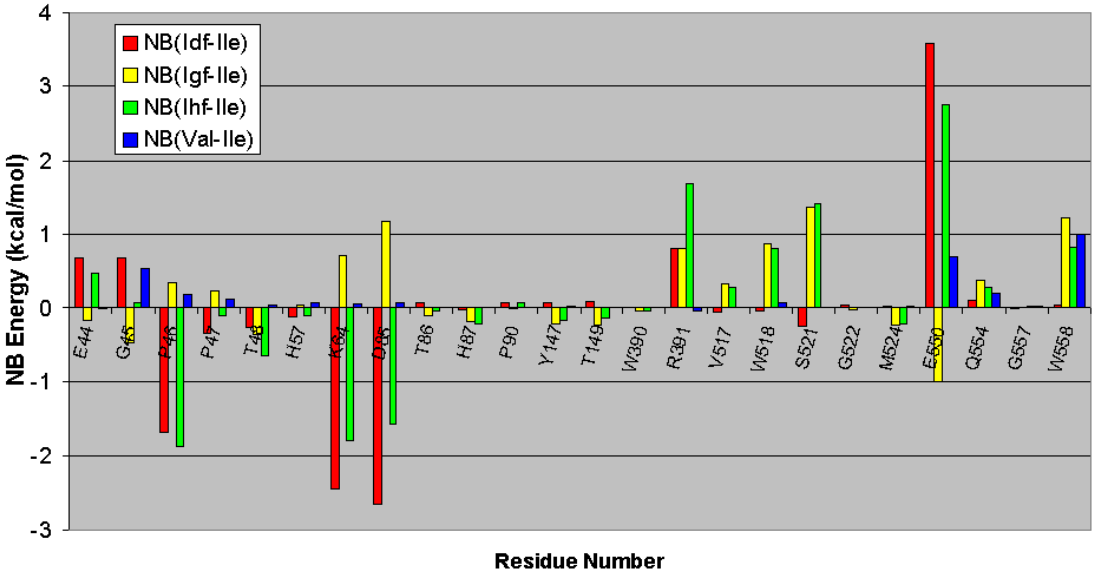


Figure 3. The component analysis for each residue in the binding site with ligands compared to Ile. (c) Hydrogen bond (HB) interactions. (d) Nonbond (NB) interactions. Here NB interactions include vdW, Coulomb and HB interactions.

Figure 3(d) show the total nonbond interactions as a whole for each residue in the binding site of IleRS with all ligands compared to Ile. The purpose is to find which residues are important in differentiating the binding of each ligands. And this information is very useful when it comes to design of an optimal mutant IleRS for non-cognate ligand binding. Coulomb energies dominate the nonbond interactions for most important residues in the binding site. This is true for E44, K64, D85, R391, and E550, as they are all charged residues. Other important residues, such as G45, P46, V517, W518, S521, Q554, and W558 have comparable contributions from both vdW and Coulomb energies. Comparing between Idf and Igf, P46, P47, K64, D85, S521, and W558 are the residues favoring Idf over Igf in binding to IleRS, while E44, G45, and E550 are the residues favoring Igf. The overall effect is that there are more residues favoring Idf binding and larger favorable energy interactions for Idf than Igf. As a result, Idf can be incorporated into proteins using IleRS *in vivo*, and Igf cannot. Again, for Ihf, four residues P46, T48, K64, and D85 favor Ihf binding, and E44, R391, W518, S521, E550, and W558 disfavor Ihf. The overall effect for Ihf is that it is much less favored than Ile and Idf, but slightly more favorable than Val, and more favorable than Igf. Using Val as a reference, which is known to be able to bind to IleRS, Ihf seems to be able to bind to IleRS as well.

4. Conclusions

We have simulated the binding of Ile, Val, and three fluorinated Ile analogs to IleRS. Component analysis was performed on each system to elucidate the contributions

from each residue in the binding site. The overall binding order is Ile > Idf > Ihf > Val > Igf. Using Val as a reference, both Idf and Ihf are better binders, and Igf is a worse binder than Val. Component analysis shows that Coulomb, VdW, and solvation energies are the main energy components in the difference. And Coulomb interaction seems to dominate the overall energy interactions with some charged residues in the binding site.

References

1. Ibba, M. & Soll, D. (2000) *Annu. Rev. Biochem.* **69,** 617-650.
2. Freist, W., Sternbach, H. & Cramer, F. (1992) *Eur. J. Biochem.* **204,** 1015-1023.
3. Ibba, M. & Soll, D. (1999) *Science* **286,** 1893-1897.
4. Schmidt, E. & Schimmel, P. (1994) *Science* **264,** 265-267.
5. Nureki, O., Vassylyev, D. G., Tateno, M., Shimada, A., Nakama, T., Fukai, S., Konno, M., Hendrickson, T. L., Schimmel, P. & Yokoyama, S. (1998) *Science* **280,** 578-582.
6. Nomanbhoy, T. K., Hendrickson, T. L. & Schimmel, P. (1999) *Molecular Cell* **4,** 519-528.
7. Budisa, N., Steipe, B., Demange, P., Eckerskorn, C., Kellermann, J. & Huber, R. (1995) *Eur. J. Biochem.* **230,** 788-796.
8. Cowie, D. B. & Cohen, G. N. (1957) *Biochim. Biophys. Acta* **26,** 252-261.
9. Deming, T. J., Fournier, M. J., Mason, T. L. & Tirrell, D. A. (1997) *J. Macromol. Sci. Pure Appl. Chem.* **A34,** 2143-2150.
10. Duewel, H., Daub, E., Robinson, V. & Honek, J. F. (1997) *Biochemistry* **36,** 3404-3416s.
11. Kiick, K. L. & Tirrell, D. A. (2000) *Tetrahedron* **56,** 9487-9493.
12. Kiick, K. L., van Hest, J. C. M. & Tirrell, D. A. (2000) *Angew. Chem. Int. Ed. Engl.* **39,** 2148-2152.
13. Kothakota, S., Mason, T. L., Tirrell, D. A. & Fournier, M. J. (1995) *J. Am. Chem. Soc.* **117,** 536-537.
14. Richmond, M. H. (1963) *J. Mol. Biol.* **6,** 284-294.
15. Sharma, N., Furter, R., Kast, P. & Tirrell, D. A. (2000) *FEBS Lett.* **467,** 37-40.
16. Tang, Y., Ghirlanda, G., Vaidehi, N., Kua, J., Mainz, D. T., Goddard, W. A., DeGrado, W. F. & Tirrell, D. A. (2001) *Biochemistry* **40,** 2790-2796.
17. Tang, Y., Ghirlanda, G., Petka, W. A., Nakajima, T., DeGrado, W. F. & Tirrell, D. A. (2001) *Angewandte Chemie-International Edition* **40,** 1494-1498.
18. van Hest, J. C. M. & Tirrell, D. A. (1998) *FEBS Lett.* **428,** 68-70.
19. van Hest, J. C. M., Kiick, K. L. & Tirrell, D. A. (2000) *J. Am. Chem. Soc.* **122,** 1282-1288.
20. Yoshikawa, E., Fournier, M. J., Mason, T. L., & Tirrell, D. A. (1994) *Macromolecules* **27,** 5471-5475.
21. Mendel, D., Ellman, J. A. & Schultz, P. G. (1991) *J. Am. Chem. Soc.* **113,** 2758-2760.
22. Archontis, G., Simonson, T. & Karplus, M. (2001) *J. Mol. Biol.* **306,** 307-327.
23. Zhang, D. Q., Vaidehi, N., Goddard, W. A., Danzer, J. F. & Debe, D. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6579-6584.
24. Ghosh, A., Rapp, C. S. & Friesner, R. A. (1998) *J. Phys. Chem. B* **102,** 10983-10990.
25. Lim, K.-T., Brunett, S., Iotov, M., McClurg, R. B., Vaidehi, N., Dasgupta, S., Taylor, S. & Goddard, W. A., III (1997) *J. Comput. Chem.* **18,** 501-521.
26. Ding, H. Q., Karasawa, N. & Goddard, W. A., III (1992) *J. Chem. Phys.* **97,** 4309-4315.

27. Mayo, S. L., Olafson, B. D. & Goddard, W. A., III (1990) *J. Phys. Chem.* **94,** 8897-8909.
28. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998) *J. Phys. Chem. B* **102,** 3586-3616.
29. Tannor, D. J., Marten, B., Murphy, R., Friesner, R. A., Sitkoff, D., Nicholls, A., Ringnalda, M., Goddard, W. A. I. & Honig, B. (1994) *J. Am. Chem. Soc.* **116,** 11875-11882.
30. Lin, J. H., Baker, N. A. & McCammon, J. A. (2002) *Biophys. J.* **83,** 1374-1379.
31. Stryer, L. (1995) *Biochemistry* (W. H. Freeman and Company, New York).
32. Madura, J. D., Briggs, J. M., Wade, R. C., Davis, M. E., Luty, B. A., Ilin, A., Antosiewicz, J., Gilson, M. K., Bagheri, B., Scott, L. R. & Mccammon, J. A. (1995) *Computer Physics Communications* **91,** 57-95.

Chapter 5

The COP Protein Design Tool*

* This chapter is adapted from the provisional patent application "The COP Protein Design Tool," filed at Caltech on April 12, 2002.

1. Introduction

Proteins are synthesized with precise control over sequence, leading to the vast range of specific structures and functional properties observed in nature. Even so the monomer pool for proteins is limited to the 20 natural amino acids. Increasing the monomer pool by incorporating new amino acid analogs would allow development of fascinating new bioderived polymers exhibiting novel but well-controlled architectures (1, 2). This could lead to many interesting applications including incorporating a fluorescence probe to elucidate specifics of protein structure and function (3) and incorporating selenium-substituted serine to facilitate crystallization processes in proteins (4).

The *in vivo* incorporation of amino acid analogs into proteins is controlled in large measure by aminoacyl-tRNA synthetases (AARSs), the class of enzymes that safeguards the fidelity of amino acid incorporation into proteins. It has been demonstrated that the wild-type translational apparatus can be used to incorporate some amino acid analogs into protein (5-11). However, the number of amino acid analogs incorporated in proteins in vivo is small, and the functionalities of these analogs have been limited. To expand the range of amino acid analogs that can be incorporated in vivo, it is desirable to manipulate the activity of the AARSs (12, 13). There has been steady progress in developing the twenty-first AARS-suppressor tRNA pairs *in vivo* (14, 15). The biggest success is the design of a novel orthogonal tRNA and tyrosyl-tRNA synthetase (TyrRS) [from

Methanococcus jannaschii tyrosyl tRNA synthetase (hereafter denoted as M.jann-TyrRS)] that incorporates O-methyl L-tyrosine (OMe-Tyr) site-specifically in protein in response to an amber nonsense codon (16). Such procedures have tremendous potential to expand the genetic codes in living cells, but the current combinatorial experiments, which considered 5^{20} mutation trials on five residues expected to be at the binding site of the tyrosine ligand, can become cumbersome.

In this chapter, we describe the Clash Opportunity Progressive (COP) design that can computationally design a mutant protein that would preferentially bind an analog ligand over the natural ligand occurring in the wild type protein binding. The method has been applied to design a series of mutant tyrosyl-tRNA synthetase (TyrRS), phenylalanyl-tRNA synthetase (PheRS), and tryptophanyl-tRNA synthetase (TrpRS) for various non-natural amino acids. Chapters 6, 7 and 8 will present these results.

2. Methods

The Clash Opportunity Progressive (COP) procedure is a structure-based rational redesign of a binding site. Given a protein structure with its binding site for the wild-type ligand, the redesigned protein or mutant will specifically bind an analog of the wild-type ligand. This procedure is useful in predicting which mutations in the binding site are essential for preferential binding to a specific ligand. We demonstrate the design strategy by designing AARS mutants that activate a specific amino acid analog preferentially compared to all natural amino acids. Our design goal is for the mutant protein to

preferentially bind the target amino acid analog *versus* the wild type ligand (and any other natural amino acid). To do this we calculate the differential binding energy of the desired analog against any other potential competitor ligand that might bind selectively to the mutant. For example, in redesigning TyrRS, we calculate the differential binding energy of the analog against Tyr and Phe. For cases in which the analog is much larger than Tyr, we might consider Trp as a potential competitor for the redesigned mutant AARS. The COP design procedure for designing mutant AARS comprises the progressive sequence of steps (Figure 1):

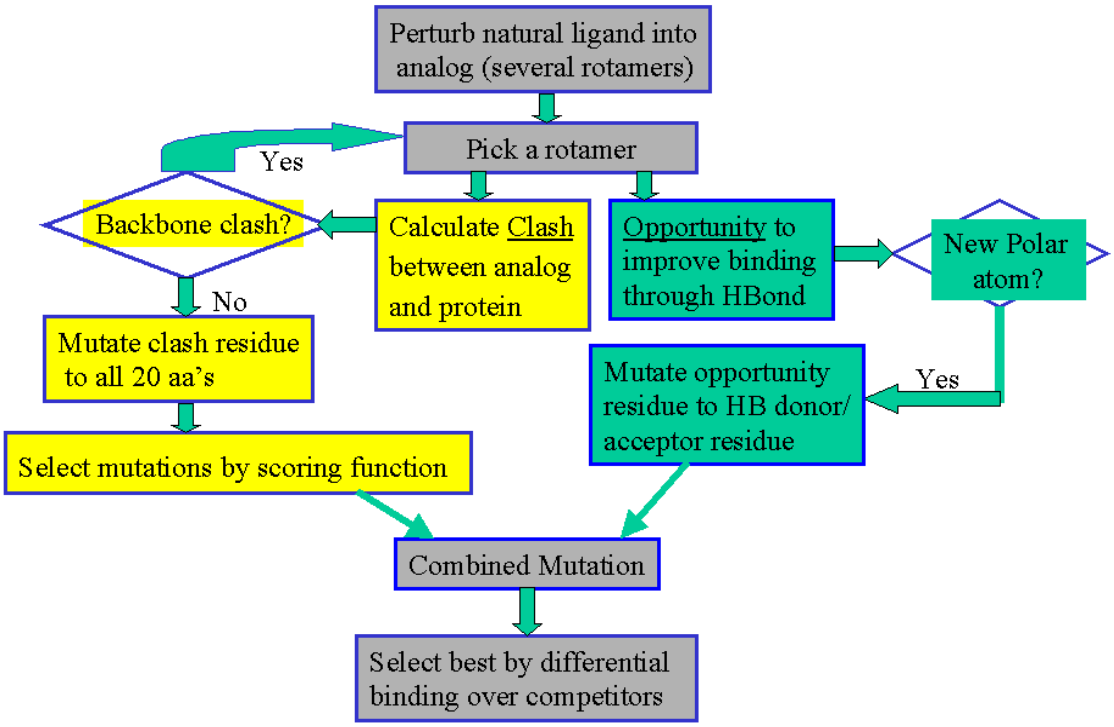


Figure 1. The flowchart of COP. Each step is color-coded.

Step 0: Conformation determination.

We first determine the favorable conformations of the analog. These can be generated the various rotamers of the ligand over a grid of dihedral angles and calculated their energies in solution using QM (alternatively, this can be carried out using a force field with molecular mechanics).

In building the analogs, we want to preserve the binding site as much as possible, which means the zwitterions for non-natural amino acids will always be conserved. Only binding site for the side chain is redesigned. This is important because the AARS binds not only the amino acid, but also ATP and the cognate tRNA (17). These binding events need to be in the exact position in order for the AARS to catalyze the reaction (18). Sometimes the analog is much bigger than the wild-type ligand, which is true for many non-natural amino acids with interesting properties, we might need to find an optimal orientation for the side chain of the analogs. In Chapter 8, we will describe how to generate such conformations for the analogs.

Step 1: Clash identification.

The low energy rotamers from Step 0 are then docked into the binding site. To do this the natural amino acid in the binding pocket is replaced with the energetically favorable rotamers of the analog while keeping the backbone of the ligand fixed (in order that the reaction center for the formation of the aminoacyl–AMP complex would be

retained for the analog). Then the analog rotamer is matched onto the binding site and the non-bond energy contributions (E_k) are calculated for each residue k in the binding pocket. These calculations can use any reliable force field or can use quantum mechanics. In our illustrations, we use Equation 1 [the functional forms for these Coulomb, van der Waals, and hydrogen bond non-bond interactions are from the DREIDING force field (19)]:

$$E_k = \sum_{i,j}\left(\frac{q_i q_j}{4\pi\varepsilon r_{ij}} + D_e\left(\left(\frac{r_m}{r_{ij}}\right)^{12} - 2\left(\frac{r_m}{r_{ij}}\right)^{6}\right) + D_{HB}\left(5\left(\frac{r_{HB}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{HB}}{r_{ij}}\right)^{10}\right)\cos^4\theta\right), \quad (1)$$

where i and j sum over all atoms in the ligand and protein residue k, of interest, q_i and q_j are partial charges on atoms i and j, respectively. r_{ij} is the distance between atoms i and j, and r_m and D_e are van der Waals distance and well depth of atoms i and j, r_{HB} and D_{HB} are hydrogen bond distance and well depth, respectively. θ is the hydrogen bond angle between atoms i, j and their bridging hydrogen atom. Please note that the hydrogen bond term is only evaluated for hydrogen bond pair atoms. When there is no bridging hydrogen atom for i and j, the hydrogen bond term is turned off. In principle any functional form from any forcefield can be used for this component analysis.

Those residues in the wild-type protein having bad clashes with the analog are marked for mutation. Because the protein backbone is fixed in Step 2, we require that the analog rotamer should *not* clash with the backbone of the protein. Analog rotamers having a severe clash with the protein backbone are discarded. Table 3 in Chapter 6 shows an example of the bad clashes of the analog in the wild-type protein.

Step 2. Relieving clashes with point mutations.

Those residues having bad clashes with the analog are mutated sequentially to all the other 19 amino acids. These point mutations use conformations that can be generated over a grid of dihedral angles or by using side chain rotamer libraries (20). The backbone of the protein is held fixed in this stage. After each point mutation the side chain alone is optimized while keeping the rest of the protein fixed. This optimization can be by energy minimization (used in our example) and can be using Molecular Dynamics or Monte Carlo techniques. This optimization is important because the initial side chain placement may not be optimal, leading to local bad contacts that might give misleading indications of the viability of this specific conformation. Then we calculate the contribution from this mutated side chain to the binding energy of both the analog and the wild-type amino acid using, for example, Equation 2. Also the clashes of this mutated residue with neighboring residues in the protein are calculated. The best mutations are selected using a scoring energy function consisting of a weighted sum of the differential non-bond interaction energy of the mutated residue with the ligand and the analog, and the non-bond interaction energy of the mutated residue with the rest of the residues in the protein. We find that weights of 0.75 to 1.0 for ligand-protein interaction, and 0.0 to 0.25 for protein-protein interactions are useful. Since the energy cost of desolvating an amino acid to place it in the binding pocket of a protein can be important in some systems, we also add desolvation penalty to the energy of the mutated residue. The desolvation penalty can be calculated using any of a variety of methods (including SGB (21), AVGB (Zamanakos et al., unpublished result), Poisson-Boltzmann (22) solvation method). Here

we calculate the differential binding energies for all twenty possible amino acid mutations, as shown in Table 4 in Chapter 6. This procedure is repeated for all the residues showing a clash in the binding pocket of the analog (Step 1). The mutation candidates for further consideration are selected based on the scoring energy between the analog and the natural ligand. Here we consider both mutations that favor binding of the analog ligand and ones that disfavor the natural ligand. Based on this differential scoring we select a subset of amino acids for each residue that will be used later for simultaneous combinatorial mutations at each clash site (from Step 1).

Step 3. Stabilizing point mutations (opportunities).

After identifying candidate mutation for relieving clashes, we look for *opportunities* for mutations in the binding pocket that would stabilize the analog ligand or disrupt the bonding with the natural ligand. Thus we consider residues near the ligand (in our example we used a cutoff of 6 Å) and look for residues that might take advantage, for example, of hydrogen bond donor or acceptor atoms that are different between the analog and the natural ligand. Another strategy for doing this is to calculate the void space in the binding pocket after making the clash mutations and consider any residue close to a large void as a candidate for a stabilizing point mutation. These opportunity mutations are selected with the same procedure as for clash mutations. It is important to include the original choice for the opportunity mutations to compare the effect of opportunity part in the final mutants generated by combined mutations in Step 4.

Step 4. Combined mutations.

Steps 2 and 3 lead to a subset of mutation candidates that are expected to either relieve clashes or provide stabilization opportunities to the binding of the analog in preference to the wild type ligand. In Step 4 we generate simultaneous mutations from each of the chosen subsets of mutations. For example, if the clash analysis (Step 2) leads to 3 residues with 2, 3 and 4 candidates and the opportunity analysis leads to one residue with 5 possible mutation candidates (say for making hydrogen bonds with the analog), then we would consider $2 \times 3 \times 4 \times 5=120$ possible protein mutants. In Step 4 we generate the best possible rotamer combination for each of these 120 mutant proteins (here we optimize the side chains, for example, using conjugate gradient minimization). Then after selecting optimum side chains for all 120 cases, the structure of the whole mutant protein is optimized both with the natural ligand and with the analog. This optimization can be by energy minimization (used in our example) and can be using Molecular Dynamics or Monte Carlo techniques. Before doing the optimization we first re-examine the residues near any of the mutated residues to determine the optimum side chain conformation. Finally the differential binding energy of the analog to the natural amino acid in the mutant is calculated. These calculations can use any reliable force field or can use quantum mechanics. In our illustrations we use Equation 2 with DREIDING force field and including SGB solvation.

$$-\Delta\Delta G = \Delta G(protein) + \Delta G(ligand) - \Delta G(protein + ligand) \qquad (2)$$

Step 5. Relaxation of the free mutant protein without the ligand.

In Step 4 we considered mutations and side chain conformations that enhance binding of the protein to the analog. However, it is possible that some mutations would disrupt the folding of the free protein when the ligand is not present. Thus for the best mutants selected from Step 4, we re-optimize the side chains *without* the ligand in the binding site. This allows the side chain of each mutation to go into the part of the binding site normally occupied by the ligand. In this step we first reselect the side chain conformation from the side chain library and then optimize the structure of the full protein using the including solvation (for example, SGB continuum solvent procedure). In these calculations we might include explicit water in the active site to better represent the stability of the active site without ligand. Once the mutant structure is optimized without ligand, the ligand is then matched on to the binding site and the potential energy of the resulting structure is minimized using for example SGB solvation. This is done for both the analog and the natural ligand (and any other ligands that might bind to the mutated site). For the analog ligand this will generally lead to a weaker binding energy than Step 4, because we now include the penalty paid to push the side chains away from the binding site as the ligand binds. However, the natural ligand may have a stronger binding energy for Step 5. Thus the differential binding energy in Step 5 will generally be smaller than in Step 4. We denote this differential binding energy as the "relaxed protein binding energy," since the mutants were optimized with no ligand in the binding pocket. The binding energy is defined the same as in Equation 2.

Step 6. Mutant Selection.

From Steps 4 and 5 we select candidate mutants with good binding energies to the analog both for the relaxed protein (Step 5) and from Step 4 and which have high differential binding energies to the natural amino acid. While redesigning AARS for binding to a specific analog, it is important that the mutant AARS activates only the analog and not any other natural amino acid. Thus the best candidate mutants are tested further for binding to other natural amino acids. To do this we dock likely natural amino acid competitors into both the relaxed and optimized binding sites, using the procedures described in Step 1. The binding energy is calculated for each ligand/mutant pair. The mutants are finally ranked by the difference in binding energies between the analog and its competitors. The better binding energy is taken either from the relaxed or the optimized mutant cases.

There is a possibility that the designed mutant protein might not be able to fold correctly. For example, if there is charged residue placed in the protein core without favorable local stabilizing interactions, it is a strong destabilizing force. In order to detect such cases in post design, we use a consensus method to evaluate the interactions for each residue involved in the design. The consensus is from all the AARS structures our group has worked on in the last couple of years. These AARSs include TyrRS (PDB: 2ts1, 3ts1, 4ts1), PheRS (PDB: 1b70), SerRS (PDB: 1ses, 1set, 1sry, 1fyf), ArgRS (PDB: 1bs2), MetRS(PDB: 1f4l), HisRS (PDB: 1adj, 1hht). Table 1 lists the energies for each amino acid from the consensus:

Table 1. Interactions energies of each amino acid in crystal structures of AARSs. These values are used to decide if an amino acid is in an unfavorable position.

Residue	Average Energy (kcal/mol)	Standard Deviation (kcal/mol)
Ala	-1.893	3.654
Arg	-108.953	40.459
Asn	-28.324	9.948
Asp	-48.155	14.464
Cys	-4.297	3.145
Gln	-23.980	7.071
Glu	-44.910	11.809
Gly	-2.857	3.652
His	-6.088	6.505
Ile	3.522	5.380
Leu	1.613	5.037
Lys	-43.560	10.184
Met	-2.067	4.624
Phe	6.536	6.326
Pro	8.183	6.653
Ser	-6.136	5.444
Thr	-6.364	5.733
Trp	16.568	7.371
Tyr	0.960	5.725
Val	1.578	4.564

In a case where a residue has lower interaction energy with its neighboring residues, a warning message will be given with a stability score of the residue. The score is defined by the energy difference divided by the standard deviation. A score higher

than 2 generally means the residue is in a very unfavorable position, i.e., it is not making enough interactions with other residues to stabilize the fold.

Steps 1 to 6 are repeated for other low energy rotamers of the analog from Step 0.

3. Advantages and Improvements over Existing Methods

Existing computational protein design methods focus mainly on design of protein core, i.e., the packing effect of various protein side chains combinatorially. Such methods use different algorithms to tackle the combinatorial nature of side chain packing, such as dead-end-elimination, branch-and-bound, Monte Carlo and genetic algorithms. Some methods were also extended to include surface residues. Nonetheless, these methods almost exclusively design for a certain protein fold, i.e., replacing many residues to achieve better protein stability while maintaining the original fold. Since these methods are focused on design of a whole protein, computational efficiency required very crude energy evaluators. In addition, these existing design methods have one or more of the following drawbacks:

1. Partial force field, which cannot describe proteins accurately.
2. Energy function is not complete.
3. Solvation is often empirical, if present at all.
4. Backbone fixed all the time.

On the other hand, the purpose of the COP method is to make the minimum number of rational structure based mutations (we understand that 5 is practicallly a maximum) in a protein so that the protein binds preferentially to the desired ligand compared to the native ligand. The COP method uses the principle minimum change design, which focuses on mutation of residues in the binding site. There may be circumstances in which it is appropriate to modify residues outside of the active site. This simplifies the problem greatly, because the number of residues involved is typically much less than the number of residues involved in a protein core design. In addition, the residues required to do mutation are often distant, hence no combinatorial side chain placement problem exists. Other advantages in the COP design methods include the following:

- COP can use any force field valid for both protein and ligand (particularly valuable here are generic force fields such as DREIDING or UFF that are valid for a large part of the periodic table) and it can use quantum mechanics for the region of the active site.

- COP uses a complete nonbond energy function such as in Equation 1. This function includes (Coulomb) electrostatic interactions (which may be describe as point charges as in Equation 1 or maybe described as distributed charges as in Qeq (23) or ReaxFF (24)), nonelectrostatic nonbond interaction (referred to as van der Waals) which may be described by a 6-12 Lennard-Jones potential (or Morse form or exponential six), and an explicit hydrogen bond potential

(which may be described by a radial potential (e.g., 10-12 Lennard-Jones) and may use a three-body cosine angle term as in Equation 1.

- Solvation is explicitly included in the COP design procedure. Continuum implicit solvation methods such as SGB or AVGB can be used to describe the role of solvation in the structure and energies of protein and ligand in water (or other) solvent. This greatly decreases the computation effort over the use of explicit solvent. However, explicit solvent molecules can be included in the evaluation of the best cases for final selection in the design.

- The protein backbone can be allowed to move (distort in response to the mutations, solvent, and ligand) at any part of the algorithm. COP allows the protein backbone to be fully movable in any part of the optimization. The designed protein can be better optimized with backbone flexibility.

- COP design adds the functionality of recognizing a new analog ligand to a mutant AARS, and it selects against any natural amino acids. This ensures that the designed AARS binds the analog amino acid exclusively, therefore can be used as an orthogonal tRNA-synthetase tRNA pair that corresponds to the twenty-first amino acid (25).

4. Possible variations and modifications

The basic COP design methodology can be modified in many ways that are various manifestations of the same idea.

- Force field used.

Although the DREIDING force field is chosen in this procedure, other force fields, such as AMBER(26), CHARMM (27), OPLS (28), etc., can also be used to calculate clashes and binding energies. The functional forms of the nonbond energy in Equation 1 can have different forms: The dielectric constant in the Coulomb term can be distance dependent. The charges for both protein and ligand can be varied. This includes charges from experiment, or charge based on various models (such as but not limited to QEq, Del Re, Gasteiger). The van der Waals term can have different forms, such as a Morse potential instead of a Leonard-Jones potential. The Leonard-Jones potential can be made 6-10 or even softer to allow closer contact. Finally the hydrogen bond term can have several different variations. Here we use three-body form, but two-body or four-body form is common, too. In some force fields, hydrogen bond is treated implicitly as part of the Coulomb term.

- Solvation methods.

Solvation is an important factor in determining biomolecular stability and binding properties. As an integrated part in the COP design, implicit solvent is used to minimize the structures and calculate binding energies. This implicit solvent model includes, but not limited to, Surface Generalized Born (SGB) model, Solvent Accessible Surface Area

(SASA)/ Analytical Volume Generalized Born (AVGB), and Poisson Boltzmann (PB) model. In addition, explicit solvent can be easily added to the calculation of binding energy here. What is important is that the solvation model must be accurate enough to account for the solvation effect in the ligand binding to the protein.

- Methods of binding energy calculation.

It is always important to get the correct relative binding energies for different ligands. Here we can use minimization in the Potential of Mean Force (PMF) from implicit continuum solvent model to calculate binding (free) energies. Alternatively we can use the average dynamic free energy instead of free energy from a single conformation. Other methods can be used to calculate binding free energies, but they may require significantly longer computation time. These methods include Free Energy Perturbation (FEP) (29), and methods based on thermodynamic cycles. For the case of amino acids binding to AARS, we found that the binding energy with PMF is very close to experimental numbers when the average dynamic free energy is used.

- Input protein/ligand structures.

The COP design program requires a protein and a ligand structure as input. Generally speaking, the protein structure should be in high quality from either X-ray or NMR study. However, protein structures of high quality from theoretical modeling can also be used. Indeed for the example used here of *M. Jannaschii*, the protein structure was obtained computationally by combining the STRUCTFAST structure alignment predicting with molecular dynamics using a force field. In some cases a homologous

protein can be used in the design, and mutations can be translated back into the protein of interest according to the sequence alignment between the two proteins. The structures of the ligand can be obtained from crystallographic databases or can be predicted using quantum mechanics or a force field. The binding site can be determined from crystal structure containing a ligand bound to the target protein, or if no ligand is present in the protein structure it can be modeled using various docking techniques. An example of such docking techniques is HierDock (30), which has been used extensively to predict and verify the binding of amino acids to AARS.

- Protein side chain modeling.

There are several side chain modeling methods that can be incorporated into the COP procedure. These side chain modeling methods include scwrl (20), scap(31), and methods based on branch-and-bound, dead-end-elimination algorithms.

- Different type of protein function design.

COP is a generic method that can be applied to any protein for recognizing a desired ligand. The ligand type can be any molecule that is a binding target to a protein and has some sort of anchoring point as the reaction center. More generally, the binding can be between two proteins. With one of the proteins changes in mutation, COP can be applied to design a complimentary counterpart for binding.

5. Features Believed to Be New

The new feature in this invention is that we use a full force field in the design with hydrogen bond capability and solvation effects. In addition, we allow protein fully movable in the stage of binding calculation. This algorithm has been designed to make few mutations to recognize a desired ligand. Hence the energy function is more accurate and also biased towards recognizing the new ligand compared to its competitors. The conformational search of side chain rotamers can be exhaustive along with an all-atom energy function that allows COP to be unique.

6. The Graphical User Interface for COP

To make COP user-friendlier, we have written a graphic user interface (GUI) for COP using Glade, a GTK-based free user interface builder. Figure 2 shows the screen snapshot when COP is started using the graphical interface.

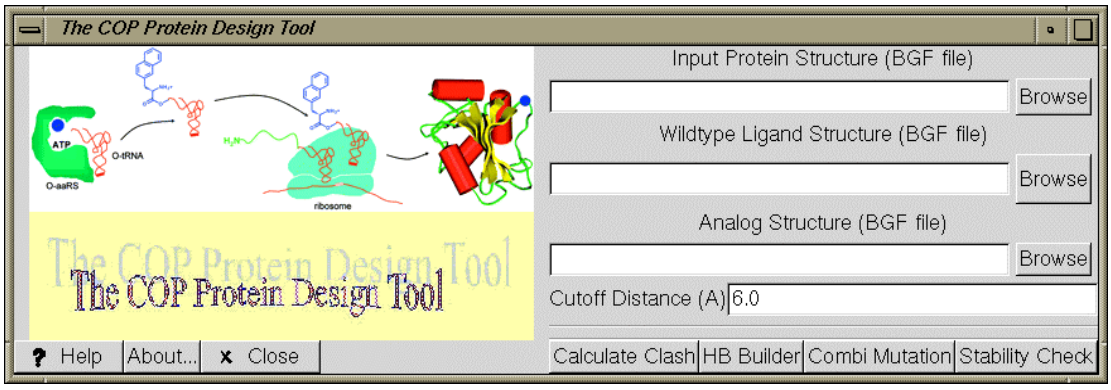


Figure 2. The graphical interface of COP.

Clicking a button brings out a popup window with the corresponding information. For example, clicking the "About…" button will open a window showing the version and copyright information of COP. The help window is designed to let user know the conventions used in the COP program. The four buttons in the bottom right of the window are for carrying four functions in COP, with each corresponding to a different program. "Calculate Clash" will run the clash identification to find mutation residues and their mutation targets to relieve clash. "HB Builder" uses a rotamer library to build possible hydrogen bond donor or acceptor residues in the binding site to stabilize new polar atoms in the analog ligand, if there is any. "Combi Mutation" carries out the combined mutation step in COP, and calculates the binding energies of each ligand including competing natural amino acids to generated mutants. A list of top candidates will be given at the end. Finally the "Stability Check" step will eliminate any mutant that potentially cannot fold correctly.

References

1. Petka, W. A., Harden, J. L., McGrath, K. P., Wirtz, D. & Tirrell, D. A. (1998) *Science* **281,** 389-92.
2. Tang, Y., Ghirlanda, G., Vaidehi, N., Kua, J., Mainz, D. T., Goddard, W. A., DeGrado, W. F. & Tirrell, D. A. (2001) *Biochemistry* **40,** 2790-2796.
3. Noren, C. J., Anthony-Cahill, S. J., Griffith, M. C. & Schultz, P. G. (1989) *Science* **244,** 182-8.
4. Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990) *EMBO J.* **9,** 1665-72.
5. Richmond, M. H. (1963) *J. Mol. Biol.* **6,** 284-294.
6. Yoshikawa, E., Fournier, M. J., Mason, T. L., & Tirrell, D. A. (1994) *Macromolecules* **27,** 5471-5475.
7. van Hest, J. C. M., Kiick, K. L. & Tirrell, D. A. (2000) *J. Am. Chem. Soc.* **122,** 1282-1288.
8. Kothakota, S., Mason, T. L., Tirrell, D. A. & Fournier, M. J. (1995) *J. Am. Chem. Soc.* **117,** 536-537.
9. Cowie, D. B. & Cohen, G. N. (1957) *Biochim. Biophys. Acta* **26,** 252-261.
10. Budisa, N., Steipe, B., Demange, P., Eckerskorn, C., Kellermann, J. & Huber, R. (1995) *Eur. J. Biochem.* **230,** 788-796.
11. Duewel, H., Daub, E., Robinson, V. & Honek, J. F. (1997) *Biochemistry* **36,** 3404-3416s.
12. Kast, P. & Hennecke, H. (1991) *J. Mol. Biol.* **222,** 99-124.
13. Ibba, M. & Hennecke, H. (1995) *FEBS Lett.* **364,** 272-275.
14. Liu, D. R., Magliery, T. J., Pastrnak, M. & Schultz, P. G. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 10092-7.
15. Kowal, A. K., Kohrer, C. & RajBhandary, U. L. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 2268-73.
16. Wang, L., Brock, A., Herberich, B. & Schultz, P. G. (2001) *Science* **292,** 498-500.
17. Ibba, M. & Soll, D. (1999) *Science* **286,** 1893-1897.
18. Wang, P., Vaidehi, N., Tirrell, D. A. & Goddard, W. A. I. (2002) *J. Am. Chem. Soc. (in press)*.
19. Mayo, S. L., Olafson, B. D. & Goddard, W. A., III (1990) *J. Phys. Chem.* **94,** 8897-8909.
20. Bower, M. J., Cohen, F. E. & Dunbrack, R. L., Jr. (1997) *J. Mol. Biol.* **267,** 1268-82.
21. Ghosh, A., Rapp, C. S. & Friesner, R. A. (1998) *J. Phys. Chem. B* **102,** 10983-10990.
22. Tannor, D. J., Marten, B., Murphy, R., Friesner, R. A., Sitkoff, D., Nicholls, A., Ringnalda, M., Goddard, W. A. I. & Honig, B. (1994) *J. Am. Chem. Soc.* **116,** 11875-11882.
23. Rappé, A. K. & Goddard, W. A., III (1991) *J. Phys. Chem.* **95,** 3358.

24. van Duin, A. C. T., Dasgupta, S., Lorant, F. & Goddard, W. A., III (2001) *J. Phys. Chem. A* **105,** 9396-9409.
25. Wang, L. & Schultz, P. G. (2002) *Chem. Commun.***,** 1-11.
26. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1995) *J. Am. Chem. Soc.* **117,** 5179-5197.
27. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998) *J. Phys. Chem. B* **102,** 3586-3616.
28. Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. (2001) *J. Phys. Chem. B* **105,** 6474-6487.
29. Becker, O. M., MacKerell, A. D., Roux, B. & Watanabe, M. (2001) (Marcel Dekker, New York).
30. Floriano, W. B., Vaidehi, N., Goddard, W. A., Singer, M. S. & Shepherd, G. M. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10712-6.
31. Xiang, Z. X. & Honig, B. (2001) *J. Mol. Biol.* **311,** 421-430.

Chapter 6

Application of COP to Design Mutant Tyrosyl-tRNA Synthetases from *Methanococcus jannacshii*[*]

[*] Part of this chapter has been published in PNAS, 99, 6579-84

1. Introduction

Proteins are synthesized with precise control over sequence, leading to the vast range of specific structures and functional properties observed in nature. Even so the monomer pool for proteins is limited to the 20 natural amino acids. Increasing the monomer pool by incorporating new amino acid analogs would allow development of fascinating new bioderived polymers exhibiting novel but well-controlled architectures (1, 2). This could lead to many interesting applications ranging from incorporating a fluorescence probe to elucidate specifics of protein structure and function (3), to incorporating selenium-substituted serine to facilitate crystallization processes in proteins (4).

The *in vivo* incorporation of amino acid analogs into proteins is controlled in large measure by aminoacyl-tRNA synthetases (AARSs), the class of enzymes that safeguards the fidelity of amino acid incorporation into proteins. It has been demonstrated that the wild-type translational apparatus can be used to incorporate some amino acid analogs into protein (5-11). However, the number of amino acid analogs incorporated in proteins *in vivo* is small, and the functionalities of these analogs have been limited. To expand the range of amino acid analogs that can be incorporated *in vivo*, it is desirable to manipulate the activity of the AARSs (12, 13). There has been steady progress in developing the twenty-first AARS-suppressor tRNA pairs *in vivo* (12, 13). The biggest success is the design of a novel orthogonal tRNA and tyrosyl-tRNA synthetase (TyrRS) from *Methanococcus jannacshii* tyrosyl-tRNA synthetase (*mj*-TyrRS) that incorporates O-

methyl L-tyrosine (OMe-Tyr) site-specifically in protein in response to an amber nonsense codon (14). A few other non-natural amino acids were incorporated by the same group using the same apparatus since then. Such procedures have tremendous potential to expand the genetic codes in living cells, but the current combinatorial experiments, which considered 5^{20} mutation trials on five residues expected to be at the binding site of the tyrosine ligand, can become cumbersome. In this chapter we summarize the result of using the Clash-Opportunity Progressive Design algorithm (denoted as COP) to redesign the binding site of *mj*-TyrRS for the preferential binding of OMe-Tyr, Naphthyl-Ala and *p*-keto-Phe over natural amino acids. The design for OMe-Tyr leads to three mutants, of which the best mutant [Y32Q, D158A] is expected to bind OMe-Tyr strongly while discriminating against Tyr. This mutant is similar to the one [Y32Q, D158A, E107T, L162P] designed by Wang et al using combinatorial experiments. We predict that the new mutant will have much greater activity while retaining significant discrimination between OMe-Tyr and Tyr.

Since there is no crystal structure available for *mj*-TyrRS, we predicted the three-dimensional structure for wild-type *mj*-TyrRS, based on a combination of the STRUCTFAST sequence alignment and structure prediction algorithm with molecular dynamics (MD) including continuum solvent forces. [To select the 5 residues to modify in their experiments, Wang et al. (14) used a sequence alignment between *mj*-TyrRS and *Bacillus stearothermophillus* tyrosyl-tRNA synthetase (*bs*-TyrRS).] To validate the predicted structure for *mj*-TyrRS, we use MD plus continuum solvent energies to

demonstrate that tyrosine (Tyr) is the preferred ligand over the 19 other natural amino acids.

2. Methods

2.1 Structure Prediction for TyrRS from Methanococcus jannacshii

Because the crystal structure of TyrRS from *Methanococcus jannacshii* was not available, we predicted the structure using STRUCTFAST homology technique. There are three TyrRS crystal structures in the Protein Data Bank. They are all from *Bacilus stearothermophilus*, with different ligands in the structures. Structure 2ts1 has no ligand (15), 3ts1 with Tyr-AMP bound (15), and 4ts1 has a Tyr in the binding site (16). By using the sequence of the wild-type mj-TyrRS from Genbank (accession number: Q57834), the three-dimensional structure of the main chain of mj-TyrRS was predicted with STRUCTFAST homology modeling technique (Debe & Goddard, unpublished result). The structure of 4ts1 was used as the template in the prediction. The sequence identity between the two sequences is 32.1%. The main chain atoms of the initial predicted mf-TyrRS structure agree with the corresponding residues of 4ts1 structure to 0.64 Å in root mean square difference (rmsd) in coordinates after aligning the two structures using DALI (17).

To place the Tyr ligand in the predicted structure, we matched the side chain conformation of the five strictly conserved residues (Tyr32, Tyr151, Gln155, Asp158 and Gln173) in the binding site of mj-TyrRS with those conformations from the 4ts1 crystal

structure. The rest of the side chains for the predicted mj-TyrRS structure were added by using side chain modeling program SCWRL version 2.7 (18, 19) while keeping the conformations in the binding site fixed. Here we used the backbone-dependent side chain rotamer library with SCWRL to optimize the side chain conformations. The potential energy of the resulting structure was then minimized using conjugate gradient method with MPSIM (20), which allowed all the side chains to move but kept the main chain fixed. In MPSIM the Cell Multiple Method (CMM) (21) was used to rapidly yet accurately calculate the nonbond interactions. The protein was described with the DREIDING force field (22) with CHARMM22 (23) charges.

For the Tyr ligand and other amino acids in the simulation, we used Mulliken charges derived from the molecular orbitals in quantum mechanics (QM). The QM calculations were done at the HF level with the 6-31G** basis set. The geometry of the molecules was optimized under forces from Poisson Boltzmann continuum solvation (24) inside of QM package Jaguar 4.0 (Schrodinger, Portland, OR).

After optimizing the side chain conformations in the protein, the potential energy of the whole protein was minimized, with all atoms movable but with distance constraint on the hydrogen bonds between the phenolic OH group of the Tyr ligand and the Tyr32 and Asp158 side chains. This minimized structure was then used as a starting structure for annealing molecular dynamics (MD), where all constraints were relaxed. Each cycle of annealing MD involved heating the system from 50 to 600 K and cooling from 600 to 50 K in steps of 20 K for 0.5 ps. These annealing MD calculations included solvent

forces from the Surface Generalized Born (SGB) continuum solvation method (25) with a dielectric constant of 80 for bulk solvent, 2 for protein and a solvent probe radius of 1.4 Å. The final structure, shown in Figure 1, was used to predict the binding of all 20 natural amino acids and to design mutant TyrRS for non-natural amino acids.

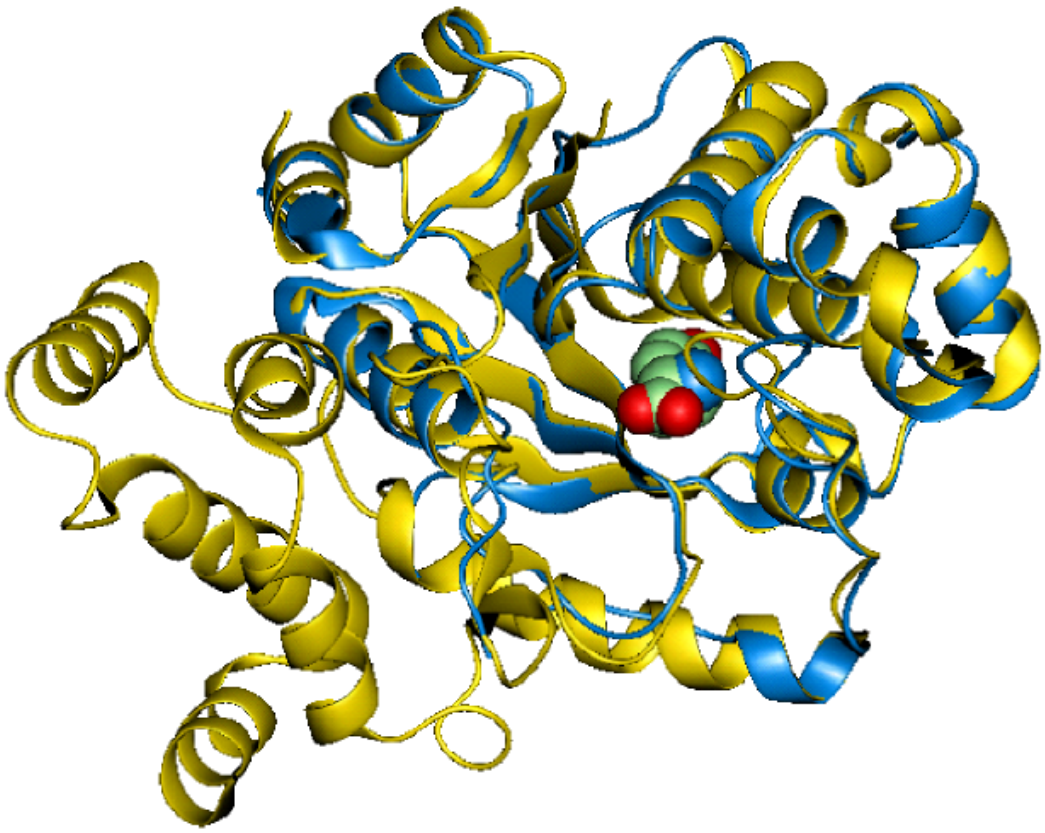


Figure 1. Comparison of predicted mj-TyrRS structure (in blue) and the crystal structure for *B. stearothermophilus* TyrRS (4ts1) (in yellow). The Tyr ligand is shown as a ball model.

2.2 Docking All 20 Natural Amino Acids to the Predicted mj-TyrRS

To validate our predicted *mj*-TyrRS structure, we docked all 20 natural amino acids to the Tyr binding site of the TyrRS. To guard against misactivation noncognate amino acids, AARSs must be able to charge the correct amino acid to its corresponding tRNA. The activation step consists of the bound amino acid forming the aminoacyl adenylate complex and subsequent transferring the aminoacyl group to the 3'-end of the bound tRNA. While there are extra proofreading mechanisms to ensure the fidelity for some AARSs, many AARSs recognize their substrate with very high specificity. And TyrRS is one of them (26, 27). It has been also shown that PheRS, another AARS recognizing their substrates with high specificity, has calculated bonding energies correlating well with the translational activity measured *in vivo* (28).

In order to get a binding conformation for other amino acids, the Tyr ligand obtained in mj-TyrRS structure optimization was used to build 19 other amino acids. To preserve the reaction center for activating the amino acids, the contact between the zwitterions of the amino acid and the appropriate residues in the binding site was fixed. SCWRL was used to mutate the side chain into 19 other amino acids. Each of the resulting amino acids was minimized in the binding site of the protein using conjugate gradient method.

The binding energy of each amino acid is calculated as

$$-\Delta\Delta G_{binding} = \Delta G(protein) + \Delta G(ligand) - \Delta G(protein+ligand), \quad (1)$$

where ΔG(protein+ligand) is the free energy for the protein-ligand complex, while ΔG(protein) and ΔG(ligand) are the free energy for the protein and ligand alone, respectively. The structure optimization was always done with SGB continuum solvation. Such continuum solvation model is optimized with the potential mean force (PMF) from bulk solvent, the total energies are very close to the free energy of the system (29). This is true especially for tight bound complexes.

2.3 Mutant mj-TyrRS Design for Recognizing Non-Natural Amino Acids

The clash opportunity progressive (COP) design algorithm (30) was used to design mutant mj-TyrRS for recognizing non-natural amino acids. The non-natural amino acids used in this chapter are listed in Figure 2.

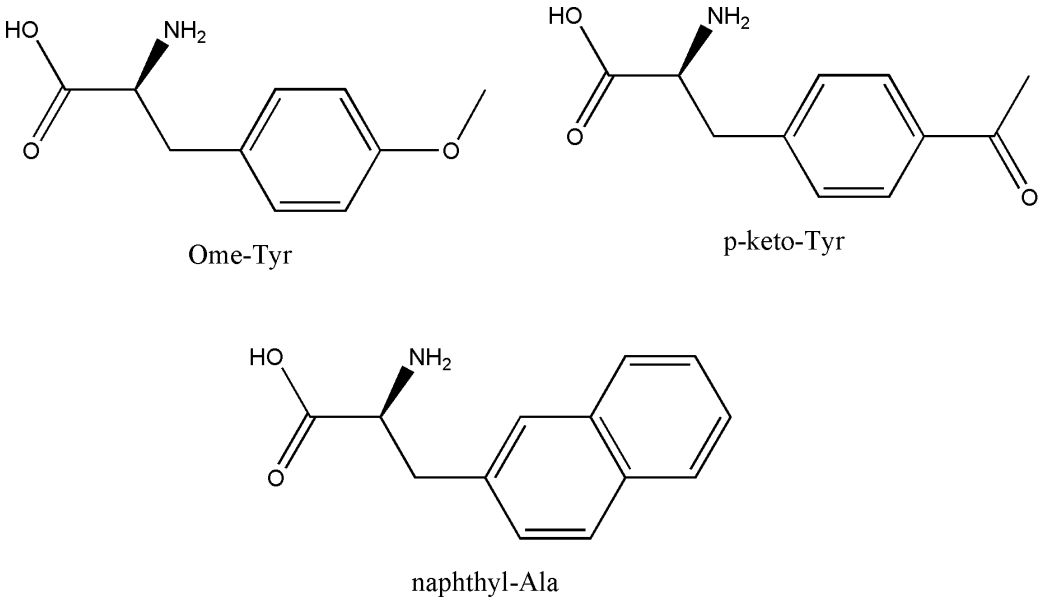


Figure 2. Non-natural amino acids used in mutant mj-TyrRS design in this chapter.

In determining the low-energy rotamers for each non-natural amino acid, several rotamers over a grid were generated and the geometry was then minimized in QM using Jaguar 4.0. Only the ones with low energy were used for subsequent COP design. The Mulliken charges were also obtained from these calculations and used for these amino acids in the design.

These low-energy conformation structures are then built in Biograf using the Tyr structure as a starting structure. The atoms that are identical in the non-natural amino acids are left as where they are. This means the zwitterions of the non-natural amino acids are always in the same position as in the Tyr ligand. These structures were used as input to the COP along with the Tyr ligand and mj-TyrRS structure. COP requires that the new atoms in the non-natural amino acids be labeled as HETATM in order to cut the binding site. The cutoff distance was 6 Å for the binding site. Residues outside the cutoff distance are not included in the clash calculation, but are included in the binding energy calculation.

The following equation is used in calculating the nonbond interaction energies between the ligand and residues k in the binding site:

$$E_k = \sum_{i,j}\left(\frac{q_i q_j}{4\pi\varepsilon r_{ij}} + D_e\left(\left(\frac{r_m}{r_{ij}}\right)^{12} - 2\left(\frac{r_m}{r_{ij}}\right)^{6}\right) + D_{HB}\left(5\left(\frac{r_{HB}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{HB}}{r_{ij}}\right)^{10}\right)\cos^4\theta\right), \quad (2)$$

where i and j sum over all atoms in the ligand and protein residue k, of interest, q_i and q_j are partial charges on atoms i and j, respectively. r_{ij} is the distance between atoms i and j, and r_m and D_e are van der Waals distance and well depth of atoms i and j, r_{HB} and D_{HB} are

hydrogen bond distance and well depth, respectively. θ is the hydrogen bond angle between atoms *i*, *j* and their bridging hydrogen atom. Please note that the hydrogen bond term is only evaluated for hydrogen bond pair atoms. When there is no bridging hydrogen atom for *i* and *j*, the hydrogen bond term is turned off.

3. Results and Discussions

3.1 Assessment of the Quality of the Predicted mj-TyrRS Structure

Figure 1 shows the predicted structure of mj-TyrRS superimposed with the crystal structure of the *B. stearothermophilus* TyrRS (4ts1). The general folds are very similar between the two structures. The main chain structure, predicted by STRUCTFAST homology technique, led to an initial rmsd of 0.64 Å for all aligned main chain atoms between the two structures before minimization. After full minimization, the main chain rmsd increases to 1.75 Å for the 139 structurally aligned residues. However, the conserved five residues (Tyr32, Tyr151, Gln155, Asp158, and Gln173) in the binding site have a rmsd of 0.62 Å for all heavy atoms. The structural alignment was done using DALI (17).

The main difference between the two structures is that mj-TyrRS lacks the α-helical domain present in 4ts1 (residue 246–317), which is consistent with the observation that mj-TyrRS has only a minimalist tRNA anticodon loop-binding domain (31).

For mj-TyrRS, the tyrosine ligand binds in the deep cleft formed by several α-helices and β-strands in the α/β domain. The phenolic hydroxyl group of the Tyr ligand makes hydrogen bonds with Tyr32-O^{η} and Asp158-$O^{\delta 1}$, both with a hydrogen bond distance of 2.87 Å. The amino group of the Tyr ligand make three hydrogen bonds with Tyr151-O^{η}, Gln155-$O^{\varepsilon 1}$, and Gln173-$O^{\varepsilon 1}$. Table 1 lists all the hydrogen bonds and their distances that the Tyr ligand makes in the binding site. Comparison of these distances (see Table 1) shows that these hydrogen bonds made by Tyr ligand are very similar.

Table 1. Hydrogen bonds in the binding site of the predicted *mj*-TyrRS structure, compared with the hydrogen bonds in *B. stearothermophilus* TyrRS (*bs*-TyrRS) crystal structure (PDB ID 4ts1).

	mj-TyrRS		*bs*-TyrRS (4ts1)	
ligand atoms	protein atom	HB distance (Å)	protein atom	HB distance (Å) *
O^{η}	Tyr32 O^{η}	2.80	Tyr34 O^{η}	2.93 (2.87)
O^{η}	Asp158 $O^{\delta 1}$	3.02	Asp176 $O^{\delta 1}$	2.27 (2.83)
N	Gln75 $O^{\varepsilon 1}$	3.14	Asp78 $O^{\delta 1}$	2.91 (2.87)
N	Tyr151 O^{η}	2.83	Tyr169 O^{η}	2.78 (2.94)
N	Gln173 $O^{\varepsilon 1}$	3.12	Gln173 $O^{\varepsilon 1}$	3.13 (3.28)
O^{XT}	Gln75 $N^{\varepsilon 2}$	5.56**	Lys82 N^{ε}	4.83 (4.97)**

* The values quoted here are from the crystal structure for *bs*-TyrRS (pdb code: 4ts1). The values in parentheses are after minimization using the DREIDING FF.

** Water mediated hydrogen bonds (HB)

3.2 Docking of the Natural Amino Acids to mj-TyrRS

As described in the methods section, all 20 natural amino acids were docked into the binding site of Tyr in *mj*-TyrRS. The binding conformation was first minimized with protein fixed and followed by relaxing the binding site residues. The final optimization was done without any constraint on the protein. All structure optimizations were done with implicit SGB continuum solvent. Table 2 shows the binding energies of these amino acids to *mj*-TyrRS. As expected, the wild-type ligand Tyr has a much higher binding energy than any other natural amino acids. The closest binding competitors are Ala, Asn and His, but all bind at least 16 kcal/mol less favorably than Tyr.

Table 2. Binding energies (including solvation) for the 20 natural amino acids docked to the binding site of the predicted structure for *mj*-TyrRS

Amino Acid	Binding Energy (kcal/mol)	Amino Acid	Binding Energy (kcal/mol)
Tyr	43.8	Val	16.2
Ala	27.2	Ile	14.1
Asn	27.2	Leu	12.0
His	27.1	Gln	9.7
Thr	26.8	Arg	2.3
Phe	26.6	Pro	1.3
Ser	25.6	Glu	-3.5
Gly	24.1	Met	-13.8
Cys	22.9	Trp	-20.6
Asp	16.4	Lys	-56.9

Although there are several steps involved in the selection specificity in AARSs, TyrRS has been known for a long time to be able to differentiate its cognate amino acid

in the initial binding stage (26). Hence the binding profile validates our predicted *mj*-TyrRS structure.

3.3 Design of Mutant mj-TyrRS for OMe-Tyr

Starting with the predicted *mj*-TyrRS structure, we used the COP algorithm to design mutant *mj*-TyrRS for selective binding of OMe-Tyr.

OMe-Tyr has two equally favorable rotamers with the carbon of the methyl group in the same plane as the aromatic ring. Both rotamers were matched in the binding site of Tyr in wild-type *mj*-TyrRS, keeping the zwitterions fixed in the structure. Component analysis of the energy contributions of each residue in the binding site to the binding of OMe-Tyr using Equation 2. The binding site is defined as the entire residue for all atoms within 6 Å cutoff distance of the ligand. Twenty-six residues are found in the binding site. The nonbond interaction energies between OMe-Tyr and these residues in the binding site are summarized in Table 3. Because rotamer 2 clashed with protein backbone at position Gly34 and rotamer 1 did not, we considered only rotamer 1 further.

From Table 3, we can see that Asp158 has a very severe clash with OMe-Tyr, and hence this residue was selected for mutation to relieve the clash. In addition, Tyr32 has a strong contribution to Tyr binding over OMe-Tyr, and it represented an opportunity to mutate to some other residue to disfavor Tyr.

Table 3. Interaction energies of OMe-Tyr ligand (both rotamers) and of Tyr ligand with the predicted wild-type structure of *mj*-TyrRS. The interactions are shown for all ligands with any atom within 6Å of the Tyr (referred to as the binding site)

Residue	OMe-Tyr (rotamer 1)	OMe-Tyr (rotamer 2)	Tyr	Difference[#]
Gln 155	-11.15	-10.65	-10.66	-0.48
Met 154	-0.93	-0.60	-0.60	-0.32
Ala 67	-1.69	-1.43	-1.42	-0.27
Gln 109	-0.99	-0.91	-0.91	-0.08
Leu 66	-0.21	-0.13	-0.13	-0.08
Asn 157	0.13	0.20	0.20	-0.08
Val 156	-0.24	-0.17	-0.17	-0.07
Phe 108	-0.16	-0.10	-0.10	-0.06
Leu 65	-1.28	265.94[*]	-1.23	-0.05
His 160	-0.25	-0.21	-0.21	-0.04
Gly 105	-0.08	-0.03	-0.03	-0.04
Phe 35	-1.52	-1.50	-1.49	-0.04
Pro 152	-0.32	-0.28	-0.28	-0.04
Ile 159	-0.05	-0.02	-0.02	-0.03
Ile 33	-0.23	-0.21	-0.20	-0.03
His 70	-2.89	-2.87	-2.88	-0.01
Tyr 161	-0.08	-0.06	-0.07	-0.01
His 177	-0.38	-0.39	-0.37	-0.01
Leu 69	0.02	0.02	0.02	0.00
Gly 34	-2.04	302.84[**]	-2.06	0.02
Gln 173	-11.15	-11.20	-11.17	0.03
Asp 68	-0.94	-0.98	-0.97	0.03
Tyr 151	-9.44	-9.52	-9.66	0.21
Glu 36	-1.14	-1.27	-1.36	0.22
Tyr 32	-13.45	12745.3[*]	-15.69	2.25
Asp 158	2450.97[*]	-0.80	-15.44	2466.41

Difference between OMe-Tyr rotamer 1 and Tyr

*Large van der Waals energy showing steric clashes of protein side chain with OMe-Tyr ligand.

**Steric clash with main chain

We considered all 20 amino acids as possible mutations for both Asp158 and Tyr32, and selected a subset of them that favor the binding of OMe-Tyr over Tyr for

further consideration. Table 4 shows those subsets: for Tyr32 Glu, Asp, Gln, Phe and Met; for Asp158: Ala.

Table 4. Binding scores of the best 6 mutations for Tyr32 and Asp158 in *mj*-TyrRS. Scores (in kcal/mol) are nonbond interaction energies of the mutated residue with the OMe-Tyr or Tyr. Based on these results, we selected the 5 mutations with negative difference for Tyr32, and one case for Asp158.

Tyr32	Tyr	OMe-Tyr	difference	Asp158	Tyr	OMe-Tyr	difference
Glu	0.13	-0.28	-0.41	Ala	-0.41	-0.92	-0.51
Asp	-0.14	-0.37	-0.23	Gly	-0.26	-0.08	0.18
Gln	-0.10	-0.28	-0.18	Ser	-0.52	2.68	3.20
Met	-0.32	-0.37	-0.05	Cys	-0.99	4.88	5.87
Phe	-0.45	-0.49	-0.04	Asp	-1.70	4.36	6.06
Ser	-0.08	-0.07	0.01	Asn	-0.64	10.54	11.18

In stage 2, we generate all mutants by combining the mutations from each site we identified in the previous stage. SCWRL was used to make these mutants. With 5 choices for Tyr32 and one choice for Asp158, a total number of 5 mutants were generated. These side chains were optimized separately for the OMe-Tyr and Tyr in the binding site. We then carried out energy minimization of the mutant structures. First only the mutated residues were allowed to move, followed by full minimization with all atoms movable. The binding energies for both Tyr and OMe-Tyr were then calculated for all mutants generated. The results are shown in the top half of Table 5.

Table 5. Binding energies of OMe-Tyr and Tyr to the wild-type *mj*-TyrRS and designed mutants.

Y32	D158	E107	L162	OMe-Tyr (kcal/mol)	Tyr (kcal/mol)	differential (OMe-Tyr – Tyr)
Y	D	E	L	-12.34	43.83	-56.17*
E	A			37.11	37.64	-0.54
D	A			43.85	38.24	5.60**
Q	A			48.93	42.30	6.62**
F	A			39.06	39.81	-0.76
M	A			44.17	39.00	5.16**
E	A	T	P	27.48	34.98	-7.51
D	A	T	P	31.65	25.58	6.06
Q	A	T	P	27.06	17.72	9.33***
F	A	T	P	31.54	34.06	-2.53
M	A	T	P	27.20	28.69	-1.50

The first row is for the wild type, which binds Tyr well but not OMe-Tyr. The next five rows (boxed) consider the mutations for Y32 and D158 identified in Table 4. The three cases denoted as ** are considered to be promising cases worth testing. The last five rows consider these same five mutations, but with the E107T and L162P mutations observed in the experiments. The case denoted as *** is the one determined experimentally.

* Wild-type *mj*-TyrRS

** Chosen designed mutant *mj*-TyrRS

*** Mutant *mj*-TyrRS found experimentally

We also calculated the binding energy of Tyr and OMe-Tyr to the wild-type *mj*-TyrRS, which is 44 kcal/mol for Tyr but –12 kcal/mol for OMe-Tyr. The result is shown in the top row in Table 5. All five mutants bind Tyr less strongly (range from 38—42

kcal/mol) whereas these five mutants bind OMe-Tyr by 37—49 kcal/mol. Of the five mutants, three favor binding of OMe-Tyr over Tyr by at least 5 kcal/mol. These are [Y32Q, D158A], [Y32M, D158A], and [Y32D, D158A], which have binding energies of 49, 44, and 44 kcal/mol foe OMe-Tyr, and differential binding energies of 7, 5, and 6 kcal/mol between OMe-Tyr and Tyr. The other two cases both lead to weak binding and favor Tyr over OMe-Tyr, hence we will ignore them.

Figure 3 shows the predicted binding site for OMe-Tyr in the best case, [Y32Q, D158A]. We can see that residues Ala67, Ala158 and Leu65 form a hydrophobic pocket for the methyl group of OMe-Tyr. The amide $N^{\varepsilon 2}$ of Gln32 has close contact with the oxygen atom of the OMe group (3.79 Å), whereas the $O^{\varepsilon 1}$ atom is stabilized by forming a weak hydrogen bond (3.58 Å) with the main chain NH of Leu65. These hydrogen bonds might be stabilized by an intervening water. The mutant [Y32M, D158A] is also a favorable candidate. However, for [Y32D, D158A], the charged group of the Asp does not seem to have a favorable stabilization of the charged group, which may lead to folding problem.

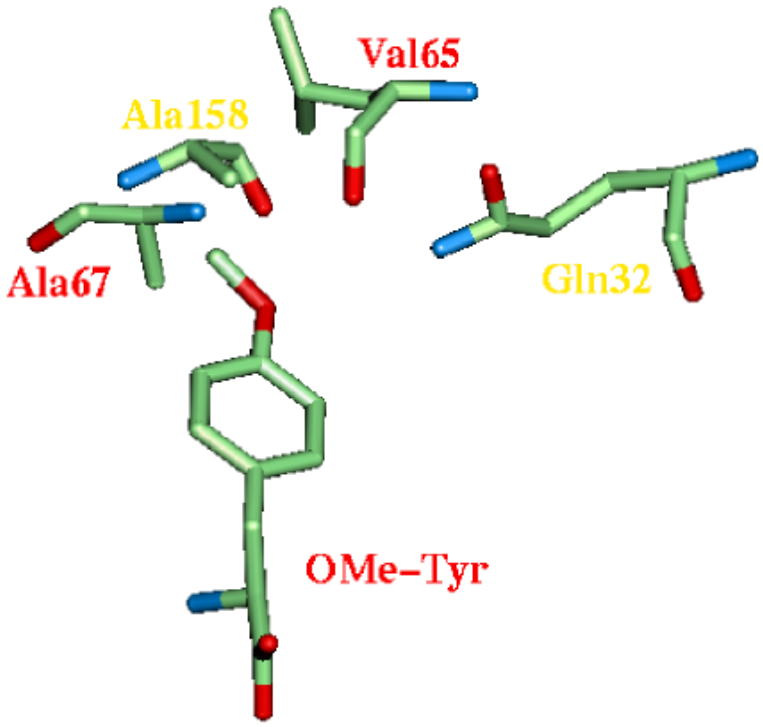


Figure 3. The predicted binding site surrounding the OMe-Tyr in the COP designed mutant [Y32Q, D158A] *mj*-TyrRS. The mutated residues (Gln32 and Ala158) are labeled in yellow. Ala67, Ala158, and Leu65 form a hydrophobic pocket for the methyl group. The amide $N^{\varepsilon 2}$ of Gln32 has close contact with the oxygen of the OMe group, whereas the $O^{\varepsilon 1}$ of Gln32 is stabilized by forming a weak hydrogen bond with the main chain NH of Leu65. (Both may have intervening water molecules.)

We can then compare our prediction with experiment. Wang et al. carried out a combinatorial experiment to find a mutant optimal for OMe-Tyr binding. Because there was no crystal structure available, they used a sequence alignment with the 4ts1 structure. Their alignment suggested five residues (Y32, D158, E107, L62, and I159) are in the binding site. They then screened a library containing 5^{20} mutants. The selection was carried out by first screening binding for any amino acid in the presence of OMe-Tyr in

the media, followed by screening against natural amino acids without OMe-Tyr in the media to find the one least able to bind Tyr and any other amino acid. Their study led to a mutant [Y32Q, D158A, E107T, L162P]. I159 did not change in this mutant.

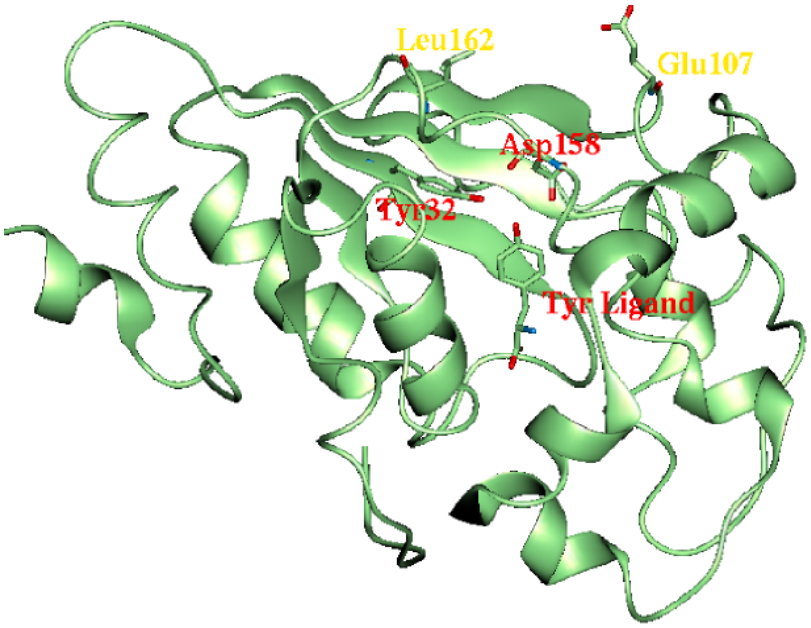


Figure 4. The predicted *mj*-TyrRS with explicit side chains for two residues Tyr32 and Asp158 involved in the design. The Tyr ligand is also shown (labeled in red). The positions of two other residues Leu162 and Glu107 are also shown (labeled in yellow).

Thus the experiments identified the [Y32Q, D158A] mutations designed by COP to be the best. However, the experimental mutant also included E107T and L162P. COP did not consider these two residues because both residues are far from the binding site in our predicted *mj*-TyrRS structure (Figure 4). Glu107 is on the surface and is 12.9 Å from the binding site (from the C^{α} of Glu107 to the O^{η} of the Tyr ligand). Leu162 is 14.5 Å

away from the binding site. In Wang et al.'s alignment, Leu162 and Glu107 in *mj*-TyrRS correspond to residues Leu180 and Asn123 in the 4ts1 structure. In 4ts1, Leu162 is in the middle of a β-strand on the bottom of the binding site, and a mutation Leu→Pro will certainly disrupt the secondary structure, thus destabilize the protein. Asn123 is in the core of the protein, it thus seems unlikely that a charged Glu could fold into this structure. In our predicted structure, both these two residues are well outside the binding site, thus COP did not find these residues as mutation targets.

To understand why the combinatorial experiments led to a different selection than the COP design, we made mutants including the [E107T, L162P] mutations along with the five cases from COP design. The L162P mutation requires a change in the main chain conformation, and therefore we carried out annealing MD to optimize the backbone structure. The resulting best-energy structure was selected to calculate binding energies. These numbers are also included in Table 5.

We find that the experiment selected mutant [Y32Q, D158A, E107T, L162P] leads to a dramatically weak binding (17 kcal/mol) toward Tyr. Because the experiments conducted several rounds of negative selection against any natural amino acids, our favored mutant [Y32Q, D158A] would have been screened out due to its higher affinity to Tyr (42.3 kcal/mol). We also find that the experimental mutant leads to a differential binding energy of 9 kcal/mol for OMe-Tyr over Tyr. This is by far the best differential binding among all the mutants in our design. However, the net binding of OMe-Tyr to the mutant is only 27 kcal/mol. It could explain the observation that the mutant led to an

incorporation rate much poorer than for the natural amino acid. Thus, the calculations do seem to be consistent with the experiment and the way it carried out.

Because our predicted mutants all would have been in the mutant library in the experiment, it would be interesting to reexamine the three cases predicted by COP to determine how effective they are and how the screening can be further improved to make better mutants. We suspect the predicted differentials of 5-7 kcal/mol may be sufficient to obtain specificity. In addition, the total binding energies of 44-49 kcal/mol for OMe-Tyr suggest that these new mutants would be much more active.

3.4 Design of Mutant mj-TyrRS for Naphthyl-Alanine

Our next design target non-natural amino acid was L-3-(2-naphthyl)alanine (naph-Ala). The same predicted *mj*-TyrRS was used. Two rotamers of the naphthyl-Ala were built from the Tyr ligand. Mulliken charges from QM calculation were assigned to naphthyl-Ala. Each of the two rotamers were matched into the binding site of Tyr, and clashes were calculated between the ligand and proteins. The result was listed in Table 6.

From Table 6, Q155 has a main chain clash with rotamer 1 of naph-Ala. This eliminates rotamer 1 from further consideration. Rotamer 2 does not have any main chain clash, therefore the following design steps are only applied to rotamer 2. Using a cutoff of 0.5 kcal/mol, two residues Y32 and D158 are selected as mutation sites. Each of the 20 amino acids is tried on the two positions one at a time. The mutated residue is

minimized, and the interaction energies with naph-Ala and Tyr are evaluated. Finally a score is calculated for each mutation. These results are listed in Table 7.

Table 6. Interaction energies of naph-Ala (both rotamers) and of Tyr with *mj*-TyrRS. The interactions are shown for all ligands with any atom within 6Å of the Tyr (referred to as the binding site). All energies are in kcal/mol.

Residue	Naph-Ala (rotamer 1)	Naph-Ala (rotamer 2)	Tyr	Difference[#]
H177	-0.87	-1.49	-0.78	-0.71
L65	-2.34	-2.36	-1.78	-0.58
G34	-1.60	-2.00	-1.56	-0.44
I33	-0.54	-0.78	-0.40	-0.38
Q155	-5.48 (6.01)*	-14.38	-14.06	-0.31
Q173	-2.50	-3.08	-2.81	-0.27
A167	0.03	-0.28	-0.08	-0.20
V168	-0.26	-0.44	-0.25	-0.19
A67	-1.72	-1.49	-1.36	-0.13
G169	-0.26	-0.31	-0.19	-0.12
A180	-0.08	-0.18	-0.12	-0.06
H160	0.03	-0.14	-0.58	0.43
D158	35128.41	10.99	-16.06	27.05
Y32	1350.52	4100030.92	-15.33	4100046.24

* Q155 has main chain clash 6.01 kcal/mol.

Difference is between naph-Ala rotamer 1 and Tyr.

Table 7. Interaction energies of each mutation on two mutation sites (Y32 and D158). All 20 amino acids are tried on each site. The final score is calculated as described in the method section using 95% of the nonbond interaction with ligands plus 5% of the constraint energy for the mutated residue with neighboring residues within protein.

Tyr32	Tyr	Naph-Ala	Score	Asp158	Tyr	Naph-Ala	Score
I	-0.47	-1.63	-1.59	G	0.02	-0.50	-0.61
Q	-0.08	-1.97	-1.29	A	-0.35	-0.35	-0.24
V	-0.28	-1.03	-1.18	C	-1.12	0.52	1.33
N	-0.34	-1.82	-1.15	I	-1.22	0.64	2.01
M	-0.51	-1.17	-1.04	V	-1.16	0.80	2.17
T	-0.07	-1.03	-0.95	H	-1.50	0.68	2.26
L	-0.58	-1.71	-0.92	M	-1.11	1.37	2.29
P	-0.23	-0.72	-0.86	F	1.71	3.45	3.04
C	-0.19	-0.63	-0.68	N	-2.95	-0.28	3.23
A	-0.16	-0.50	-0.58	S	-3.57	0.18	3.75
G	-0.12	-0.38	-0.37	Q	-0.24	3.07	3.76
S	-0.15	-0.56	-0.17	T	-4.55	-0.39	4.17
E	0.52	-5.23	2.35	Y	15.60	19.64	7.21
D	-0.23	-2.07	6.21	D	-5.25	-4.61	7.37
H	-0.41	5.88	6.71	K	2.21	5.91	8.88
F	-0.61	10.32	10.49	E	-3.75	-0.89	9.74
R	-1.48	5.87	11.87	W	1.75	8.82	10.61
Y	-0.77	11.80	12.46	P	-0.66	9.13	11.40
K	-2.52	8.43	14.86	L	3.62	28.29	25.91
W	-1.56	62.16	66.50	R	1129.31	1168.98	116.71

From Table 7, there are 12 choices for Tyr32 (I, Q, V, N, M, T, L, P, C, A, G, S), and two choices for Asp58 (G and A). A cutoff value of 0 is used to choose mutations that favor naph-Ala over Tyr. In the next step, $12 \times 2 = 24$ mutants are generated. The mutants are first minimized with only the mutation residues movable and the rest of the protein and ligand fixed, followed by a full optimization. Equation 1 is then used to score each mutants for binding with naph-Ala and Tyr. Two possible competitors from natural amino acids, Trp and Phe, are also scored for binding to the mutants designed. Normally

this procedure is only performed for mutants with high affinity to the designed analog, however, every mutants are evaluated with competitors because there are only 24 mutants designed. Table 8 lists the mutants we have designed using COP.

Table 8. Designed mutant TyrRS for binding naph-Ala using COP. The binding energies are in kcal/mol. The difference is between naph-Tyr and the competitor with the best binding energy.

Y32	D158	naph-Tyr	Tyr	Phe	Trp	Difference	Rank	Stability
M	A	37.53	30.08	28.46	-15.38	7.45	1	OK
M	G	36.47	29.59	19.68	-16.55	6.88	2	OK
Q	G	34.50	28.80	22.57	-19.66	5.70	3	OK
I	G	34.54	28.95	22.25	-13.48	5.59	4	OK
L	G	33.15	27.75	24.62	-20.90	5.40	5	OK
V	G	34.01	28.90	25.43	-16.02	5.11	6	OK
N	G	34.50	29.45	28.41	-15.64	5.05	7	OK
T	G	33.83	28.86	22.08	-14.24	4.97		
P	G	33.24	28.31	22.21	-7.38	4.93		
C	G	32.71	27.81	19.55	-21.73	4.90		
A	G	34.17	29.45	20.22	-20.36	4.72		
S	G	34.14	29.42	19.35	-19.92	4.72		
G	G	34.16	29.48	22.30	-16.70	4.68		
Q	A	33.80	29.30	23.06	-19.38	4.50		
I	A	33.76	29.43	22.51	-13.35	4.33		
L	A	33.64	29.53	26.05	-19.56	4.11		
N	A	33.43	29.55	21.95	-22.38	3.88		
V	A	33.21	29.36	27.61	-16.01	3.85		
C	A	33.71	29.94	19.81	-20.90	3.77		
T	A	33.04	29.33	22.41	-13.91	3.71		
P	A	32.41	28.78	22.45	-7.30	3.63		
A	A	33.46	29.85	19.85	-22.43	3.61		
S	A	33.46	29.89	19.87	-18.58	3.57		
G	A	33.35	29.82	19.84	-18.40	3.53		

From Table 8, we see that all these COP designed mutants have good binding energies to naph-Tyr and better binding energy than any of the competitors. Among these mutants, there are seven of them having binding energies at least 5 kcal/mol better

than any of the competitors. And the stability checks performed on them all indicate they can fold into the native state without any problem.

Further, we compare these mutants with the experimental mutant selected from a library of 5^{20} mutants with five positions each replaced by one of the 20 natural amino acids (32). These five positions are Y32, D158, I159, L162, and A167. And the experimental mutant is Y32L-D158P-I159A-L162Q-A167V. Compared with COP designed mutants, the first two mutation sites are the same as what COP found. However, the other three residues I159, L162 and A167 are not in contact with the analog naph-Ala, thus COP did not identify them as mutation residues. The mutation Y32L also appears in two designed mutants with good binding affinity toward naph-Ala. On the other hand, P was not a choice for D158 in COP design. The reason is that it requires main chain conformational change in the mutation D158P. The phi/psi angles for D158 in mj-TyrRS is $-58^{o}/113^{o}$, while for P it should be either $-57^{o}/-38^{o}$ or $-63^{o}/139^{o}$. We performed a simulation by making a mutant with all the mutations found in experiment. Due to the main chain conformational change in the D158P mutation, an annealing dynamics was carried out on the mutant to allow the back bone of the mutant to adjust to an optimal position. Simulation showed the phi/psi angles were $-53^{o}/125^{o}$ in the optimized mutant. The mutation V159A facilitated the main chain conformational change by allowing the backbone move further away from the ligand (Figure 5).

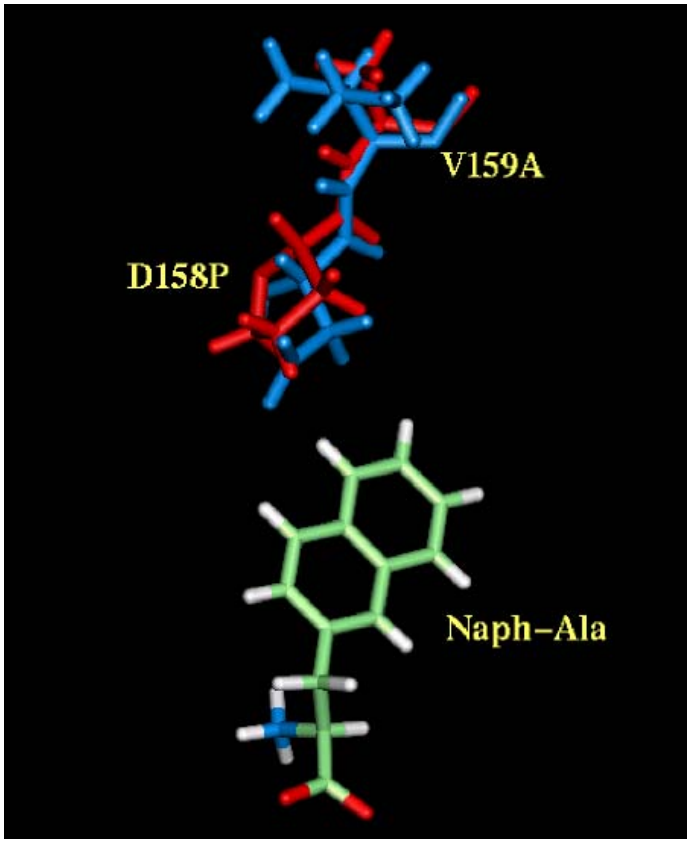


Figure 5. The main chain change in the mutant selected in experiment. The phi/psi angle changed from $-58^{o}/113^{o}$ in the wild-type mj-TyrRS (Blue) to $-53^{o}/125^{o}$ in the mutant (Red). The mutation V159A facilitated the change by allowing the main chain move further from the ligand naph-Ala.

We also calculated the binding energies of naph-Ala and its competitors (Tyr, Phe, Trp) and the results are listed in Table 9. As a comparison, the wild-type mj-TyrRS and a few COP designed mutants are also listed.

From Table 9, it is seen that the experimentally selected mutant has the worst binding energy to Tyr, which is also the best binding natural amino acid among the competitors to naph-Ala. This is consistent with the experimental procedure, in which

several rounds of negative selection against natural amino acids were performed. Our simulation shows that that procedure selected the mutant with the worst binding affinity to natural amino acids, however, neither the affinity to naph-Ala nor the differential binding affinity to naph-Ala was optimized in the experiment. Such selected mutants usually only have moderate binding affinity to the analogs intended to design for. In this case, four of COP designed mutants have better binding affinity to naph-Ala with at least the same differential binding affinity to naph-Ala.

Table 9. The binding energies (in kcal/mol) of the wild-type mj-TyrRS (first row), experimentally selected mutant (second row), and some COP designed mutants for binding naph-Ala and some natural amino acid competitors.

Y32	D158	I159	L162	A167	Naph-Ala	Tyr	Phe	Trp	Diff
					-11.41	50.37	21.73	-21.72	-61.78
L	P	A	Q	V	32.63	27.23	21.25	-18.17	5.40
M	A				37.53	30.08	28.46	-15.38	7.45
M	G				36.47	29.59	19.68	-16.55	6.88
Q	G				34.50	28.80	22.57	-19.66	5.70
I	G				34.54	28.95	22.25	-13.48	5.59
L	G				33.15	27.75	24.62	-20.90	5.40

Figure 6 shows naph-Ala in the binding site of the best COP designed mutant Y32M-D158A. Q155, D158A, Y32M, and A167 together form the binding site for the extra aromatic ring in naph-Ala. In the experimentally selected mutant, A167V occupies larger space such that Y32 can only be one of the residues smaller than M, in this case it is V.

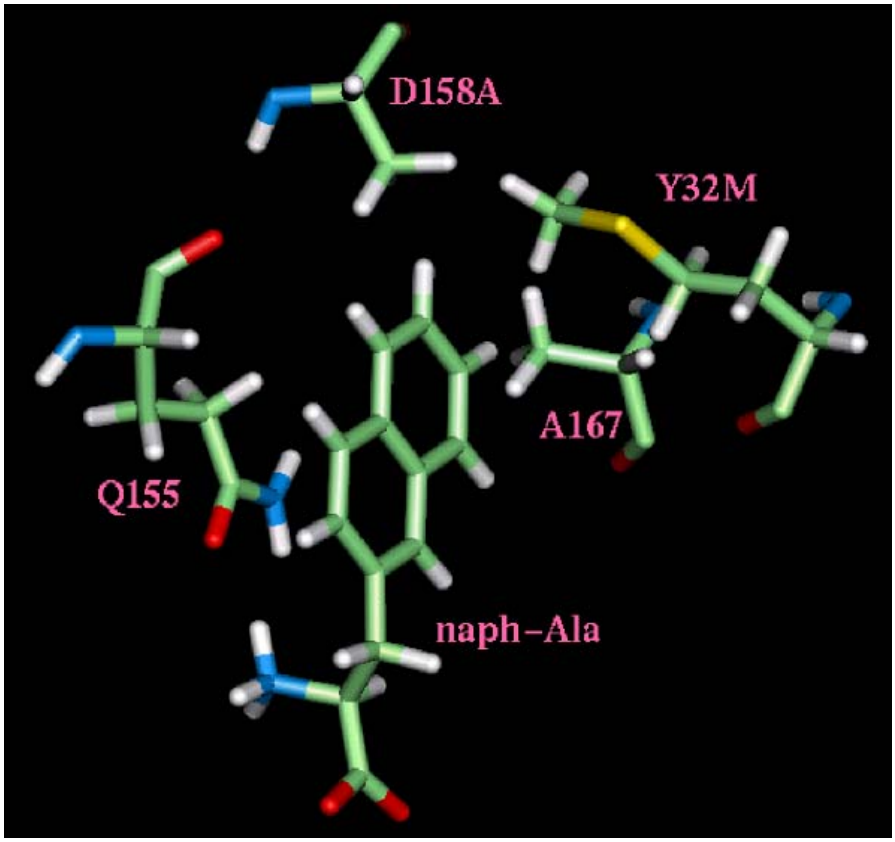


Figure 6. naph-Ala in the binding site of the best COP designed mutant. The extra aromatic ring in naph-Ala is surrounded by Q155, D158A, Y32M and A167.

3.5 Design of Mutant mj-TyrRS for p-keto-Tyr

The next Tyr analog for design is p-keto-Tyr. Two low energy rotamers of keto-Tyr were built in Biograf. The carbonyl group is conjugate with the aromatic ring in these two rotamers. Again Mulliken charges from quantum mechanics were used for ligands.

Both rotamers of keto-Tyr were matched into the binding site of Tyr in mj-TyrRS. Clashes were calculated using COP. The interactions between keto-Tyr, Tyr and residues in the binding site of mj-TyrRS were listed in Table 10.

Table 10. The interaction energies between two rotamers of keto-Tyr, Tyr and all residues in the binding site (defined as within 6 Å of keto-Tyr). The differential is between rotamer 1 and Tyr. Energies are in kcal/mol.

Residue	Keto-Tyr (Rotamer 1)	Keto-Tyr (Rotamer 2)	Tyr	Difference
N157	-0.61	0.18	0.27	-0.88
L65	-2.41	-1.32	-1.78	-0.63
G34	-1.81	-1.65	-1.56	-0.25
H177	-0.99	-0.81	-0.78	-0.21
V156	-0.32	-0.50	-0.17	-0.14
I159	-0.11	-0.10	-0.01	-0.09
A67	-1.45	-1.44	-1.36	-0.09
H160	-0.50	-1.12	-0.58	0.07
A180	-0.05	-0.25	-0.12	0.08
A167	0.03	-0.22	-0.08	0.11
Q155	-12.35	-7.88(4.82)*	-14.06	1.71
D158	441.75	7388.24	-16.06	457.81
Y32	10734317.94	520822.01	-15.33	10734333.27

* note: Numbers in ( ) are main chain clashes.

From Table 10, rotamer 2 has main chain clash with Q155, and rotamer 1 has no main chain clash, therefore rotamer 2 was not used in further steps and only rotamer 1 was further considered.

There are two residues having severe clash with keto-Tyr, and they are Y32 and D158. A third residue, Q155, has a less favorable interaction with keto-Tyr than Tyr. However, this residue is not considered as a clash residue, because the interaction with

keto-Tyr is negative. Further Q155 is an essential residue anchoring the amino acid binding by forming a hydrogen bond with the zwitterions of the amino acid ligand. Hence, Q155 was not included in the mutation list. The hydrogen bond design procedure indicates no optimal hydrogen bond residue can be found for the carbonyl group.

Next step is to try all 20 amino acids in position Y32 and D158 one at a time. The mutated residue is minimized with everything else fixed first, followed by calculating the interaction energy of the mutation with the ligands and the rest of the protein. A preferential score for each amino acid is then calculated (Table 11).

Table 11. Score for each mutation into all 20 amino acids at position Y32 and D158. The score was calculated as 95% of the differential interaction energy with keto-Tyr and with Tyr plus 5% of the constraint energy of the mutated residue with its neighbor residues in protein.

Tyr32	Tyr	keto-Tyr	Score	Asp158	Tyr	keto-Tyr	Score
M	-0.58	-1.27	-1.00	G	0.02	-1.75	-1.95
Q	-0.08	-1.73	-0.95	A	-0.38	1.84	1.84
N	-0.34	-1.56	-0.77	N	-0.87	0.90	2.02
I	-0.47	-0.90	-0.74	S	-0.58	1.66	2.24
V	-0.28	-0.61	-0.64	T	-1.09	3.01	4.14
L	-0.58	-1.17	-0.48	C	-1.00	4.01	4.54
T	-0.07	-0.69	-0.47	I	-1.28	5.59	6.83
P	-0.23	-0.45	-0.44	V	-1.19	6.26	7.48
C	-0.19	-0.34	-0.29	M	-1.16	6.99	7.74
A	-0.16	-0.30	-0.29	Q	-0.57	7.43	8.27
G	-0.12	-0.22	-0.14	H	-3.29	6.82	10.02
S	-0.15	-0.26	0.24	K	1.80	7.95	11.42
F	-0.83	0.13	0.50	Y	14.04	30.19	18.71
H	-0.63	0.04	1.06	E	-3.90	8.65	18.92
Y	-0.99	0.41	1.22	D	-4.59	8.38	19.06
E	0.53	-4.94	2.60	F	4.43	22.51	19.27
W	1.74	5.06	3.06	W	0.55	23.42	26.02
D	-0.23	-1.76	6.51	P	-1.00	23.16	31.79
R	-1.47	6.39	12.60	L	4.21	83.34	81.30
K	-2.52	7.93	14.57	R	92.14	444.31	395.32

From Table 11, there are 11 choices (M, Q, N, I, V, L, T, P, C, A, G) for Y32, and one choice (G) for D158, using a cutoff of 0 kcal/mol in score. However, it is always safe to add the next choice to the list when there is only one choice. Therefore (G, A) was chosen for D158. The combined mutation generates 11×2 = 22 mutants. The binding energies of each mutant for keto-Tyr and its competitors are calculated using Equation 1. Table 12 lists the binding energies for all 22 mutants designed here.

Table 12. The binding energies (in kcal/mol) of COP designed mutants for binding with keto-Tyr and its competitors.

Y32	D158	Keto-Tyr	Phe	Tyr	Trp	Difference	Rank	Stability check
N	G	31.41	29.54	22.32	-23.02	1.87		No
P	G	28.34	26.48	22.25	-11.58	1.86		No
G	G	30.85	29.48	22.23	-19.79	1.37	1	OK
M	G	29.33	28.44	20.2	-21.8	0.89	2	OK
A	G	30.14	29.49	22.01	-21.73	0.65	3	OK
Q	G	28.09	28.6	19.64	-21.84	-0.51		
V	G	28.18	29.44	19.44	-24.28	-1.26		
T	G	28.1	29.44	19.46	-23.7	-1.34		
C	G	28	29.45	19.43	-21.48	-1.45		
L	A	28.56	30.09	27.69	-18.36	-1.53		
I	G	27.34	29.54	19.46	-23.24	-2.2		
L	G	26.78	29.6	27.54	-18.05	-2.82		
P	A	23.56	26.95	22.68	-13.75	-3.39		
I	A	26.6	30.06	19.73	-23.36	-3.46		
N	A	23.36	30.02	22.53	-23.1	-6.66		
M	A	23.41	30.24	22.73	-21.84	-6.83		
A	A	22.97	29.98	22.24	-21.88	-7.01		
G	A	22.72	29.96	22.44	-19.91	-7.24		
Q	A	20.07	29.11	19.89	-21.89	-9.04		
V	A	20.14	29.92	19.67	-24.39	-9.78		
T	A	20.02	29.93	19.7	-23.8	-9.91		
C	A	19.84	29.93	19.67	-21.45	-10.09		

From Table 12, there are only three mutants that have better binding affinity to keto-Tyr than its competitors from natural amino acids and have no problem folding into

the native fold. Phe seems to be the main competitor in the design here, which makes sense because both polar residues recognizing Tyr over Phe are being mutated to less polar residues. The best mutant is Y32G-D158G with a 1.37 kcal/mol binding energy better than Phe, and the next two mutants Y32M-D158G and Y32A-D158G both have less than 1 kcal/mol binding energy better than Phe. These deferential binding energies are probably not big enough to exclude the misactivation of Phe *in vivo*. Therefore, no good mutants emerged from the design here.

4. Conclusions

In this chapter, we have applied the COP protein design algorithm to design mutant TyrRS that would bind optimally and preferentially to new Tyr analogs over Tyr and all other natural amino acids.

Because there was no experimental three-dimensional structure available for mj-TyrRS, we used STRUCFAST to predict the alignment and backbone fold. We then used a series of energy minimization and annealing dynamics to optimize the predicted structure. We found that this predicted structure binds Tyr much stronger than any other natural amino acids, which we consider a validation of the predicted structure. In addition, the success in predicting mutants that compare consistently with experiment provides additional evidence in favor of the predicted structure.

References

1. Petka, W. A., Harden, J. L., McGrath, K. P., Wirtz, D. & Tirrell, D. A. (1998) *Science* **281,** 389-92.
2. Tang, Y., Ghirlanda, G., Petka, W. A., Nakajima, T., DeGrado, W. F. & Tirrell, D. A. (2001) *Angewandte Chemie-International Edition* **40,** 1494-1498.
3. Noren, C. J., Anthony-Cahill, S. J., Griffith, M. C. & Schultz, P. G. (1989) *Science* **244,** 182-8.
4. Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990) *EMBO J.* **9,** 1665-72.
5. Richmond, M. H. (1963) *J. Mol. Biol.* **6,** 284-294.
6. Yoshikawa, E., Fournier, M. J., Mason, T. L., & Tirrell, D. A. (1994) *Macromolecules* **27,** 5471-5475.
7. van Hest, J. C. M., Kiick, K. L. & Tirrell, D. A. (2000) *J. Am. Chem. Soc.* **122,** 1282-1288.
8. Kothakota, S., Mason, T. L., Tirrell, D. A. & Fournier, M. J. (1995) *J. Am. Chem. Soc.* **117,** 536-537.
9. Cowie, D. B. & Cohen, G. N. (1957) *Biochim. Biophys. Acta* **26,** 252-261.
10. Budisa, N., Steipe, B., Demange, P., Eckerskorn, C., Kellermann, J. & Huber, R. (1995) *Eur. J. Biochem.* **230,** 788-796.
11. Duewel, H., Daub, E., Robinson, V. & Honek, J. F. (1997) *Biochemistry* **36,** 3404-3416s.
12. Liu, D. R., Magliery, T. J., Pastrnak, M. & Schultz, P. G. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 10092-7.
13. Kowal, A. K., Kohrer, C. & RajBhandary, U. L. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 2268-73.
14. Wang, L., Brock, A., Herberich, B. & Schultz, P. G. (2001) *Science* **292,** 498-500.
15. Brick, P. & Blow, D. M. (1987) *J. Mol. Biol.* **194,** 287-97.
16. Brick, P., Bhat, T. N. & Blow, D. M. (1989) *J. Mol. Biol.* **208,** 83-98.
17. Holm, L. & Sander, C. (1993) *J. Mol. Biol.* **233,** 123-38.
18. Bower, M. J., Cohen, F. E. & Dunbrack, R. L., Jr. (1997) *J. Mol. Biol.* **267,** 1268-82.
19. Dunbrack, R. L., Jr. (1999) *Proteins* **Suppl,** 81-7.
20. Lim, K.-T., Brunett, S., Iotov, M., McClurg, R. B., Vaidehi, N., Dasgupta, S., Taylor, S. & Goddard, W. A., III (1997) *J. Comput. Chem.* **18,** 501-521.
21. Ding, H. Q., Karasawa, N. & Goddard, W. A., III (1992) *J. Chem. Phys.* **97,** 4309-4315.
22. Mayo, S. L., Olafson, B. D. & Goddard, W. A., III (1990) *J. Phys. Chem.* **94,** 8897-8909.
23. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R.,

Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998) *J. Phys. Chem. B* **102,** 3586-3616.

24. Tannor, D. J., Marten, B., Murphy, R., Friesner, R. A., Sitkoff, D., Nicholls, A., Ringnalda, M., Goddard, W. A., III & Honig, B. (1994) *J. Am. Chem. Soc.* **116,** 11875-11882.
25. Ghosh, A., Rapp, C. S. & Friesner, R. A. (1998) *J. Phys. Chem. B* **102,** 10983-10990.
26. Fersht, A. R., Shindler, J. S. & Tsui, W. C. (1980) *Biochemistry* **19,** 5520-4.
27. Freist, W., Sternbach, H., Pardowitz, I. & Cramer, F. (1998) *J. Theor. Biol.* **193,** 19-38.
28. Wang, P., Vaidehi, N., Tirrell, D. A. & Goddard, W. A., III (2002) *J. Am. Chem. Soc. (in press)*.
29. Hendsch, Z. S. & Tidor, B. (1999) *Protein Sci.* **8,** 1381-92.
30. Zhang, D. Q., Vaidehi, N., Goddard, W. A., Danzer, J. F. & Debe, D. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6579-6584.
31. Steer, B. A. & Schimmel, P. (1999) *J. Biol. Chem.* **274,** 35601-6.
32. Wang, L., Brock, A. & Schultz, P. G. (2002) *J. Am. Chem. Soc.* **124,** 1836-1837.

Chapter 7

Structure-Based Design of Mutant Phenylalanyl-tRNA Synthetase for Incorporation of p-Keto-Phenylalanine

Abstract

Aminoacyl-tRNA synthetases are a class of enzymes to guard the fidelity in protein biosynthesis. It has been shown that some non-natural amino acids can be incorporated into protein *in vivo* using the wild-type apparatus. However, the number of such non-natural amino acids is rather limited, and the functionalities carried by these non-natural amino acids are minimal. In order to incorporate non-natural amino acids with more interesting properties, it is necessary to manipulate the activity of the aminoacyl-tRNA synthetase.

In this paper we report the result of a structure-based mutant design of phenylalanyl-tRNA synthetase for incorporation of *p*-keto-phenylalanine using our previously published Clash Opportunity Progressive (COP) protein design tool. There have been some improvements on COP since then. The designed mutants have been shown experimentally to be able to incorporate *p*-keto-phenylalanine *in vivo*. We have also been able to show why some of our previously designed mutants did not work experimentally. The improved COP should be more accurate in mutant designs.

1. Introduction

The aminoacyl-tRNA synthetases (AARSs) catalyze the esterification of amino acids to their cognate tRNAs (1). The accuracy of the reaction is essential due to its nature of protein biosynthesis fidelity. On the other hand, protein biosynthesis is a great tool to make biomaterials with precise control over sequence, structure and function. In natural the monomer pool is limited to the 20 natural amino acids. It has been shown that the monomer pool of amino acids can be increased by incorporating some non-natural amino acids using the wild-type AARS apparatus (2-5). These bioderived polymers are controlled by a genetic sequence, but have novel yet well controlled architectures (2, 6-14). However, the number of non-natural amino acids incorporated into protein using wild-type AARSs is small, and the functionalities carried by these non-natural amino acids are very limited. Typically, these non-natural amino acids are analogs of natural amino acids with little difference in the side chain. There are many non-natural amino acids that have desired chemical or physical properties cannot be incorporated this way. The most important reason is that these amino acid analogs are very different from their natural amino acid counterpart. Therefore, they are rejected by the AARSs in the esterification to tRNAs.

In order to overcome the limitations, it is desirable to manipulate the activity of AARSs to make them recognize non-natural amino acids (15). Another reason for doing AARS activity manipulation is the promise of expanding the genetic codes by developing novel tRNA:AARS pairs orthogonal to existing such pairs in cells (16). It is typically

done by evolving the suppressor tRNA with nonsense codon to pair with a cross-species mutant AARS, which recognizes a non-natural amino acid instead of one of the natural amino acids. The design of mutant AARS has been the bottleneck in this process due to the lack of an effective mutant screening method. It is typically done by screening a library of AARS mutants, in which several positions are replaced by all 20 amino acids. Five such positions will generate 5^{20} mutants. There has been some success, but it is very time-consuming and cumbersome.

We have previously developed a Clash Opportunity Progressive (COP) protein design tool for the purpose of mutant AARS design (15). There have been some improvements to COP since we last reported the procedure. Most significantly the opportunity part of the design uses a protein side chain rotamer library to design possible hydrogen bonds for new polar atoms. The scoring function was also updated to use a new Analytical Volume Generalized Born (AVGB) continuum electrostatic solvation model (17). Our improved COP has been applied to design a mutant phenylalanyl-tRNA synthetase, and the mutant has been experimentally tested to be able to recognize the target analog *p*-keto-phenylalanine (keto-Phe). We were also able to show why some of our previously designed mutant using the old COP did not work in experiment.

2. Methods

The experiment of keto-Phe incorporation was done in *E. coli*. However, there is no crystal structure available for *E. coli* phenylalanyl-tRNA synthetase (PheRS) in the

PDB. Instead PheRS from *Thermus thermophilus* has been crystallized and solved under different conditions previously (18-21). Because the homology between PheRSs from *E. coli* and *T. Thermophilus* is very high (46.2% identical residues, with only a few deletions), we used PheRS from *T. thermophilus* as the modeling system. The structure of PheRS complexed with Phe (PDB ID: 1B70 resolution 2.7 Å) was downloaded from the Protein Data Bank, and hydrogens were added using Biograf (Accelrys, San Diego, CA). The structure was minimized with conjugate gradient method to an rms force of 0.1 kcal/mol/Å or maximum of 5000 steps. Dreiding force field (22) was used for energy expression. The protein was described using CHARMM22 (23) charges, while charges for the ligands were Mulliken charges derived from molecular orbitals in quantum mechanics using Jaguar 4.5 (Schrödinger, Portland, OR). The minimized structure was used in the design.

The keto-Phe analog was built from the Phe ligand in the minimized structure. There are two rotamers with equally favorable energies (Figure 2). The Clash Opportunity Progressive (COP) design algorithm has been described previously (15). There have been some new improvements to COP since it was published. There is no change in identifying clashes, but for the opportunity part, we have implemented a rotamer library based procedure to build potential hydrogen bonds between the protein and the new ligand. The library was based on the protein side chain rotamer library used by SCAP (24). First the new analog ligand is compared with the wild-type ligand to see if there is extra polar atom for the analog ligand. If no such atom found for the analog, no hydrogen bond building is necessary. Depending on the polarity of the polar atom in the

analog, all residues within 6 Å of the polar atom are considered to build a new hydrogen bond donor/acceptor residue. Each rotamer from the side chain rotamer library is tried to see if there is a hydrogen bond can be formed between that residue and the polar atom in the analog. Rotamers that clash with the backbone of the protein and the analog will be eliminated. Once a hydrogen bond forming residue is found in this way, other residues that have side chain clash with it will be mutated to eliminate the clash. These mutations will be added to the list of mutations from clash identification and van der Waals interaction opportunity optimization. Previously this part was done by visualization and there is a great uncertainty which residue is a good target. Another change we have implemented is that for mutations from opportunity, we now always include the original choice of residue type, because these mutations are not absolutely necessary.

In the scoring method, we now have an option to use Analytical Volume Generalized Born (AVGB) solvation to better account for the solvation effect for protein/ligands. The binding energy for each protein/ligand pair is calculated as

$$-\Delta\Delta G = \Delta G(protein) + \Delta G(ligand) - \Delta G(protein + ligand), \quad (1)$$

where ΔG(protein) and ΔG(ligand) are the free energies of the protein and ligand alone, respectively, ΔG(protein+ligand) is the free energy for the complex.

Finally, some of our designed mutants have been proved to be unable to fold correctly due to unfavorable interactions caused by the mutation. This is especially true when we put a charged residue in the protein core. To solve this problem, we now use an individual residue interaction energy test, which was based on the statistics of the

interaction energy of residues in AARSs. The interaction energy of each residue in the binding site with its neighbors is calculated. A score s(n) for residue n is calculated by

$$s(n) = \frac{E(n) - E(AA)}{\sigma(AA)}, \qquad (2)$$

where E(n) is the interaction energy of residue n with its neighboring residues, E(AA) and σ(AA) are the interaction energy and the standard deviation of residue type AA from statistics. Table 1 shows the values we used in our COP procedure. A score higher than 2 usually indicates a high possibility that the designed mutant has folding problem.

Table 1. The interaction energies and standard deviations of each amino acid type from all the AARSs structures known so far.

AA	E(AA) (kcal/mol)	σ(AA) (kcal/mol)
Ala	-1.893	3.654
Arg	-108.953	40.459
Asn	-28.324	9.948
Asp	-48.155	14.464
Cys	-4.297	3.145
Gln	-23.980	7.071
Glu	-44.910	11.809
Gly	-2.857	3.652
His	-6.088	6.505
Ile	3.522	5.380
Leu	1.613	5.037
Lys	-43.560	10.184
Met	-2.067	4.624
Phe	6.536	6.326
Pro	8.183	6.653
Ser	-6.136	5.444
Thr	-6.364	5.733
Trp	16.568	7.371
Tyr	0.960	5.725
Val	1.578	4.564

3. Results and Discussions

The PheRS/Phe complex was minimized as described in the methods section. The rmsd for the complex between and after minimization was 0.20 Å. This showed that our force field was compatible with the parameters used in the original structure determination. Figure 1 is the minimized PheRS in ribbon representation with Phe shown as ball model.

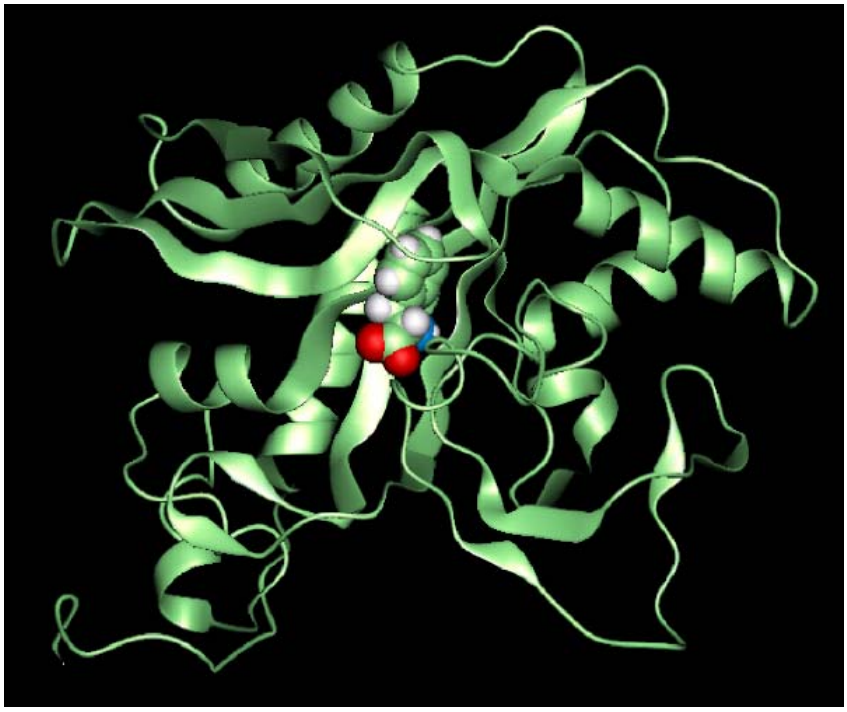


Figure 1. The ribbon representation of PheRS with Phe bound in the active site.

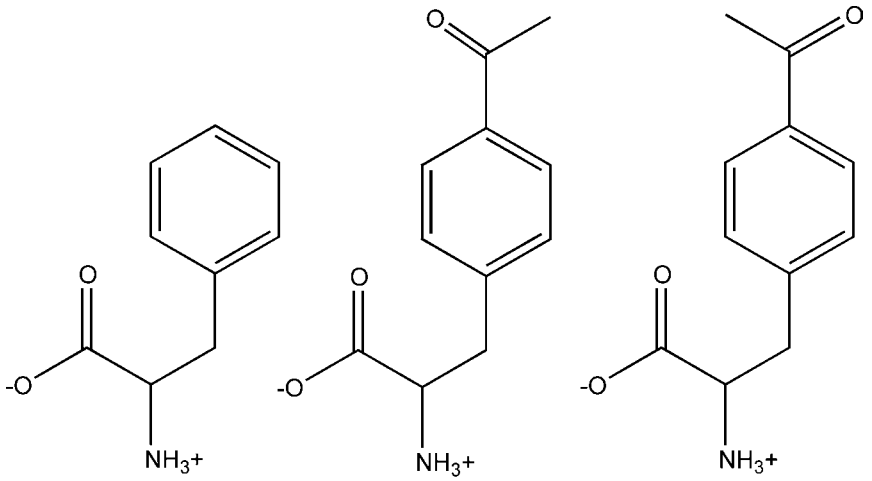


Figure 2. The wild-type ligand Phe and two rotamers of the p-keto-Phe used in the design.

The two rotamers of keto-Phe were built from the Phe ligand in the minimized PheRS/Phe complex structure. Figure 2 shows the structure of Phe and two rotamers of p-keto-Phe used in the design. Each rotamers of keto-Phe was matched into the binding site of Phe in PheRS, and clashes were calculated for each rotamer. Table 2 shows the clash result for both rotamers.

It is seen that rotamer 1 of keto-Phe clashes with the protein backbone with G284 and A283, while rotamer 2 does not clash with any backbone atoms. Therefore, only rotamer 2 was used in the following steps of design. Two residues V261 and A314 were identified as mutation target residues. Each of them was mutated into all 20 amino acids using scwrl. A backbone-dependent rotamer library was used to place the side chain conformation with the lowest constraint energy in the mutation site. Table 3 shows the result of the interaction of Phe and keto-Phe with these mutations.

Table 2. Clashes calculated for each rotamers of keto-Phe in the binding site of PheRS. The binding site is defined as within 6 Å of the keto group in the analog.

Residue	keto-Phe (r1)	Phe	Difference
E220	-6.54	-2.80	-3.75
F260	-4.31	-3.34	-0.97
F258	-4.87	-3.90	-0.97
G282	-2.32	-1.66	-0.66
M285	-0.70	-0.12	-0.58
A265	-0.56	-0.18	-0.38
V286	-0.40	-0.07	-0.33
L222	-0.26	-0.07	-0.18
G264	0.11	-0.00	0.11
G316	-1.69	-1.86	0.17
F315	-1.33	-1.69	0.35
G284	4.17*	-0.52	4.69
A283	25.29*	-1.74	27.03
A314	60.12	-1.31	61.43
V261	4705.49	-1.27	4706.76

* Indicates main chain clash.

Residue	keto-Phe (r2)	Phe	Difference
E220	-8.60	-2.80	-5.80
G282	-2.79	-1.66	-1.13
F260	-4.19	-3.34	-0.85
A283	-2.38	-1.74	-0.64
G284	-1.08	-0.52	-0.56
F315	-2.23	-1.69	-0.55
F258	-4.39	-3.90	-0.48
G264	-0.40	-0.00	-0.40
S180	-8.94	-8.57	-0.36
G221	-0.44	-0.15	-0.29
A265	-0.39	-0.18	-0.21
V286	-0.25	-0.07	-0.17
M285	-0.09	-0.12	0.03
G316	-1.77	-1.86	0.09
E262	0.21	-0.39	0.60
V261	161.40	-1.27	162.66
A314	931.41	-1.31	932.72

Table 3. The interactions between mutated residues with Phe and keto-Phe (r2)

V261	Interaction with Phe	Interaction with keto-Phe	Difference
G	-0.16	-0.92	-0.75
A	-0.51	0.22	0.54
S	-0.61	0.57	1.19
C	-1.04	3.68	4.41
N	-1.25	6.46	8.24
T	-1.35	8.39	9.69
P	-1.17	6.93	10.59
V	-1.41	10.14	11.13
D	-2.11	3.64	14.87
I	-1.57	25.68	29.09
H	-1.89	28.76	31.90
L	-1.22	115.67	118.74
E	-3.01	124.58	141.29
M	62.13	228.14	166.30
Y	57.30	289.95	249.12
W	203.24	423.05	255.14
K	418.60	733.53	331.06
F	28.16	412.26	393.18
R	204.46	2379.7	2195.66
Q	0.72	7311.2	7312.93

A314	Interaction with Phe	Interaction with keto-Phe	Difference
G	-0.60	-1.45	-0.85
D	-1.51	25.33	26.84
C	-1.27	30.68	31.95
S	-1.54	31.00	32.54
A	-1.07	31.75	32.82
T	-0.83	34.45	35.29
E	-1.91	37.54	39.45
H	-1.24	38.49	39.74
L	-1.50	38.29	39.79
Q	-1.47	39.26	40.73
I	-0.30	41.36	41.67
M	-1.53	40.28	41.81
V	-0.17	42.13	42.3
N	8.40	51.23	42.83
W	-1.66	42.74	44.41
F	-1.54	51.99	53.53
Y	-1.51	52.89	54.4
P	-1.39	85.06	86.45
R	-1.60	110.25	111.85
K	-0.08	269.75	269.83

Using a cutoff of 0 kcal/mol, only one choice for both V261 and A314 were selected, and both were Gly. There is a polar oxygen atom in the keto group. However, the hydrogen bond design algorithm in COP did not find optimal hydrogen bonds for the keto group. Thus COP designed a V261G-A314G mutant only. Previously we did the hydrogen bond design part by visualization and decided to build a hydrogen bond donor residue on V286. We also tried to make room for V286 mutation by make L222 to smaller residues. Table 4 lists the binding energies to the double Gly mutant and some mutants we previously designed. Both rotamers of keto-Phe and some competitors (Phe and Tyr in this case) were used as binding ligands. As a test, the binding energy to the wild-type PheRS and an A314G mutant were also calculated. The A314G mutant has been previously shown to be able to bind p-Br-Phe.

Table 4. Binding energies of both rotamers of keto-Phe and competitors (Phe and Tyr) to wild-type and mutant PheRS designed using COP. The first row is for wild-type PheRS, the second row is the double Gly mutant PheRS designed by COP, and rows 3-6 are mutant PheRS designed by the old COP using visualization in hydrogen bond building. The last row is a mutant known to bind p-Br-Phe. Energies are in kcal/mol. A question mark in stability check column denotes that the mutation is questionable in stability.

V261	A314	V286	L222	Phe	Tyr	keto-Phe	Stability check
				24.10	9.50	3.61	Y
G	G			4.35	9.61	17.06	Y
G	G	R		11.47	26.51	17.91	N
G	G	Q	I	5.45	11.95	12.28	?
G	G	Q		8.41	13.64	15.53	?
G	G	H	V	4.32	10.67	16.26	Y
	G			6.58	9.68	4.15	Y

V261G-A314G-V286R was a mutant we designed previously using visualization as a procedure to build hydrogen bond between the protein and the analog. Using the new hydrogen bond builder with a side chain rotamer library, COP did not choose any residue to build hydrogen bond donors. Also using the new scoring method, this mutant now shows a less favorable binding energy than Tyr, thus it will also be rejected. The stability check also failed to give a stable protein fold.

The V261G-A314G mutant shows a good differential binding energy between keto-Phe and its competitors, Phe and Tyr in this case. It favors keto-Phe binding by 7.45 kcal/mol better than Tyr, the closest competitor. Figure 3 shows the binding site of keto-Phe in the designed V261G-A314G mutant.

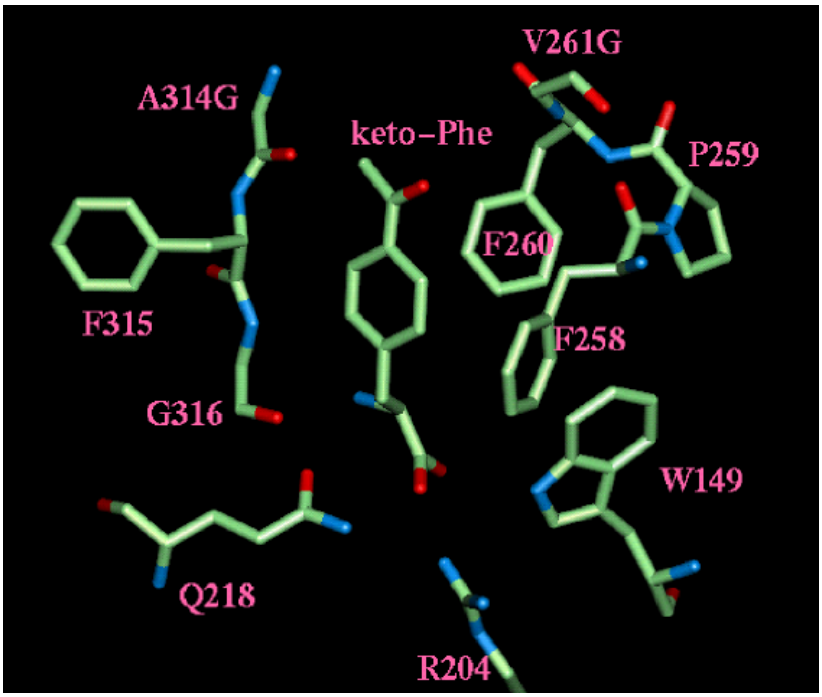


Figure 3. Keto-Phe in the binding site of the V261G-A314G mutant TrpRS. No hydrogen bond is found for the carbonyl group in the side chain of keto-Phe.

There is no specific polar interaction with the carbonyl group of the side chain of keto-Phe from the protein, i.e., no hydrogen bond is formed. The two mutations V261G and A314G enable the binding of keto-Phe by making the binding site larger to accommodate the extra acetyl group in the side chain of keto-Phe. Other interactions remain the same as seen in the wild-type Phe-PheRS complex. Resides 258-261, 282-284, and 314-316 form the binding pocket for the side chain of keto-Phe. On the zwitterions part, E220, S180 and Q218 form hydrogen bonds with the N-terminus, and W149, H178 and R204 form hydrogen bonds with the C-terminus. These interactions anchors the amino acid ligand specifically for activation with ATP to form aminoacyl adenylate complex.

4. Conclusions

We have applied the COP protein design tool to design mutant PheRS for the *in vivo* incorporation of p-keto-Phe. A mutant V261G-A314G was designed, and showed good binding affinity to keto-Phe and good differential binding to keto-Phe than its competitors from the natural amino acids.

Using a protein mutation stability check based on amino acid interaction energy from statistics, we have shown that some of our previously designed mutants including a V261G-A314G-V286R mutant cannot fold correctly due to the lack of stabilizing interactions from other neighboring residues.

References

1. Ibba, M. & Soll, D. (2000) *Annu. Rev. Biochem.* **69,** 617-650.
2. Kiick, K. L. & Tirrell, D. A. (2000) *Tetrahedron* **56,** 9487-9493.
3. Apostol, I., Levine, J., Lippincott, J., Leach, J., Hess, E., Glascock, C. B., Weickert, M. J. & Blackmore, R. J. (1997) *J. Biol. Chem.* **272,** 28980-28988.
4. Budisa, N., Steipe, B., Demange, P., Eckerskorn, C., Kellermann, J. & Huber, R. (1995) *Eur. J. Biochem.* **230,** 788-796.
5. Deming, T. J., Fournier, M. J., Mason, T. L. & Tirrell, D. A. (1997) *J. Macromol. Sci. Pure Appl. Chem.* **A34,** 2143-2150.
6. Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990) *EMBO J.* **9,** 1665-72.
7. Kiick, K. L., van Hest, J. C. M. & Tirrell, D. A. (2000) *Angew. Chem. Int. Ed. Engl.* **39,** 2148-2152.
8. Kothakota, S., Mason, T. L., Tirrell, D. A. & Fournier, M. J. (1995) *J. Am. Chem. Soc.* **117,** 536-537.
9. Tang, Y., Ghirlanda, G., Petka, W. A., Nakajima, T., DeGrado, W. F. & Tirrell, D. A. (2001) *Angewandte Chemie-International Edition* **40,** 1494-1498.
10. Tang, Y., Ghirlanda, G., Vaidehi, N., Kua, J., Mainz, D. T., Goddard, W. A., DeGrado, W. F. & Tirrell, D. A. (2001) *Biochemistry* **40,** 2790-2796.
11. van Hest, J. C. M. & Tirrell, D. A. (1998) *FEBS Lett.* **428,** 68-70.
12. van Hest, J. C. M., Kiick, K. L. & Tirrell, D. A. (2000) *J. Am. Chem. Soc.* **122,** 1282-1288.
13. Woese, C. R., Olsen, G. J., Ibba, M. & Soll, D. (2000) *Microbiol. Mol. Biol. Rev.* **64,** 202-+.
14. Yoshikawa, E., Fournier, M. J., Mason, T. L., & Tirrell, D. A. (1994) *Macromolecules* **27,** 5471-5475.
15. Zhang, D. Q., Vaidehi, N., Goddard, W. A., Danzer, J. F. & Debe, D. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6579-6584.
16. Wang, L., Brock, A., Herberich, B. & Schultz, P. G. (2001) *Science* **292,** 498-500.
17. Zamanokas, G. & Goddard, W. A. I. (2002) *To be published.*
18. Fishman, R., Ankilova, V., Moor, N. & Safro, M. (2001) *Acta Cryst. Section D-Biol. Cryst.* **57,** 1534-1544.
19. Goldgur, Y., Mosyak, L., Reshetnikova, L., Ankilova, V., Lavrik, O., Khodyreva, S. & Safro, M. (1997) *Structure* **5,** 59-68.
20. Mosyak, L., Reshetnikova, L., Goldgur, Y., Delarue, M. & Safro, M. G. (1995) *Nat. Struc. Biol.* **2,** 537-547.
21. Reshetnikova, L., Moor, N., Lavrik, O. & Vassylyev, D. G. (1999) *J. Mol. Biol.* **287,** 555-568.
22. Mayo, S. L., Olafson, B. D. & Goddard, W. A., III (1990) *J. Phys. Chem.* **94,** 8897-8909.
23. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir,

L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998) *J. Phys. Chem. B* **102,** 3586-3616.

24. Xiang, Z. X. & Honig, B. (2001) *J. Mol. Biol.* **311,** 421-430.

Chapter 8

Design of Mutant Tryptophanyl-tRNA Synthetases for Non-Natural Amino Acid Incorporation

Abstract

Protein biosynthesis accuracy is mainly determined by the fidelity of recognition of its cognate amino acid by aminoacyl-tRNA synthetase. It has been shown that the wild-type aminoacyl-tRNA synthetase can be used to incorporate some non-natural amino acids. To further increase the number of non-natural amino acids that can be incorporated, it is necessary to modify the synthetase to recognize these analogs. We have previously used tyrosyl-tRNA synthetase and phenylalanyl-tRNA synthetase as templates to design mutant aminoacyl-tRNA synthetases to recognize non-natural amino acids successfully. However, many of the amino acid analogs carrying interesting properties are significantly larger than tyrosine and phenylalanine, and cause our design algorithm to fail due to main chain clashes. Here we used tryptophanyl-tRNA synthetase as a template to design mutant synthetases. Because the size of the side chain in tryptophan is larger, amino acid analogs with larger side chain can be designed without causing main chain clash. The results of designing for three ligands NBD-Ala, bpy-Ala, and DAN-Ala were presented here.

1. Introduction

Protein biosynthesis has excellent control over sequence and accuracy, leading to the enormously vast varieties of protein folds and functionalities. This control is unmatched by any of the modern polymer synthesis techniques after many years of advance. It is thus desirable to use protein biosynthesis as a tool to synthesize protein polymers with controlled sequence, stereochemistry, and molecular weight. However, the monomer pool in protein biosynthesis is limited to the 20 natural occurring amino acids. The monomer pool needs to be increased in order to add more monomers into protein materials.

The *in vivo* incorporation of non-natural amino acids into proteins is controlled in large measure by the aminoacyl-tRNA synthetases (AARSs), the class of enzymes that safeguards the fidelity of amino acid incorporation into proteins. It has been demonstrated that it is possible to use the wild-type translational apparatus to incorporate non-natural amino acids with fluorinated (1, 2), unsaturated (3-5), electroactive (6), and other side chain functions (7-10). Nevertheless, the number of amino acids shown conclusively to exhibit translational activity *in vivo* is small, and the chemical functionality that has been assessed by this method remains modest. These non-natural amino acids are typically the close analogs of the wild-type natural amino acid. In order to expand the chemical and physical properties that can be engineered into proteins, it is necessary to manipulate the activity of the AARS to further expand the range of non-natural amino acids that can be incorporated.

We have previously developed a Clash Opportunity Progressive (COP) protein design algorithm (11) to modify an AARS to recognize an amino acid analog carrying different side chain and not recognized by the wild-type AARS. The algorithm has been applied to design mutant TyrRS and PheRS to selectively bind OMe-Tyr, p-keto-Phe, and naphthyl-Ala. However, there are other non-natural amino acids carrying more interesting chemical and physical functionalities cannot be designed successfully because the large side chain clashes with the protein backbone. It is therefore desirable to use TrpRS as a template to design mutant AARSs. The binding site of Trp is significantly larger than Tyr and Phe, as a result the chance of backbone clash with the protein is smaller. Here we have used COP to design mutant TrpRS for recognizing 2-amino-3-(7-nitro-benzo[1,2,5]oxadiazol-4-ylamino)-propionic acid (NBD-Ala), 2-amino-3-[2,2'] bipyridinyl-5-yl-propionic acid (bpy-Ala), and 2-Amino-4-(6-dimethylamino-naphthalen-2-yl)-4-oxo-butyric acid (DAN-Ala). Some good mutants that have good differential binding energy to these analogs over natural amino acids have been designed.

2. Methods

2.1 Structure Preparation

The crystal structure of TrpRS from *B. stearothermophilus* with Trp bound in the binding site (PDB: 1I6M, resolution: 1.72 Å) (12) was downloaded from the Protein Data Bank. Hydrogens were added to the structure using Biograf (Accelrys, San Diego, CA), followed by annealing on the hydrogens to optimize the hydrogen bond network. The

heavy atoms were fixed during the annealing. The structure was subject to further optimization using conjugate gradient minimization with all atoms movable for 2000 steps and with a convergence criterion of rms force reaching less than 0.1 kcal/mol/Å. The simulation program MPSim (13) was used along DREIDING force field (14). Surface Generalized Born (SGB) continuum solvation (15) was included in the optimization to account for the solvation effect. The protein was described with CHARMM22 charges (16), while the ligand Trp was with Mulliken charges derived from molecular orbitals in quantum mechanics (QM). The QM calculation was carried out at HF level using 6-31G** basis set in Jaguar 4.5 (Schrödinger, Portland, OR). The geometry optimization was performed with forces calculated from Poisson-Boltzmann continuum dielectric solvent (17). The same Mulliken charges were used for the two analogs and other competing natural amino acids. The rms deviation between the optimized complex structure and the original crystal structure was 0.32 Å. And the protein TrpRS and ligand Trp were split from the complex for use in the mutant TrpRS design.

2.2 The COP Protein Design Algorithm

The COP protein design algorithm has been previously described (11). The procedure will be described here briefly. The first step is analog structure preparation. Several low energy rotamers of the analog is built from the wild-type amino acid. The zwitterions part of the analog is the same as that of the ligand. Then each of the rotamers of the analog is put into the binding site of the protein, and the following equation is used to calculate the nonbond interaction between the analog and residue k in the binding site:

$$E_k = \sum_{i,j}\left(\frac{q_i q_j}{4\pi\varepsilon r_{ij}} + D_e\left(\left(\frac{r_m}{r_{ij}}\right)^{12} - 2\left(\frac{r_m}{r_{ij}}\right)^{6}\right) + D_{HB}\left(5\left(\frac{r_{HB}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{HB}}{r_{ij}}\right)^{10}\right)\cos^4\theta\right), \quad (1)$$

where i and j sum over all atoms in the ligand and protein residue k, of interest, q_i and q_j are partial charges on atoms i and j, respectively. r_{ij} is the distance between atoms i and j, and r_m and D_e are van der Waals distance and well depth of atoms i and j, r_{HB} and D_{HB} are hydrogen bond distance and well depth, respectively. θ is the hydrogen bond angle between atoms i, j and their bridging hydrogen atom. Please note that the hydrogen bond term is only evaluated for hydrogen bond pair atoms. When there is no bridging hydrogen atom for i and j, the hydrogen bond term is turned off.

The same is done for the wild-type amino acid ligand, and the difference of the interaction energy with the analog and the wild-type amino acid is calculated for each residue. Those residues that have an unfavorable binding contribution to the analog will show positive differential interaction energies. These residues either clash with the analog or do not make enough interactions with the analog, thus they represent the opportunity to improve the binding to the analog. These residues will be mutated into all 20 natural amino acids to find choices that can wither relieve the clash or improve the interaction with the analog. A score is calculated by considering the differential interaction energy of the mutated residue with the analog and the wild-type amino acid, the constraint energy of the mutated residue with the rest of the protein. Mutations with positive scores are selected for use in making mutants in the combined stage.

The mutant proteins are then generated by combining the choices for each mutation site, and optimized by minimization. The binding energy of the analog to the mutant is calculated as

$$-\Delta\Delta G_{binding} = \Delta G(protein) + \Delta G(ligand) - \Delta G(protein + ligand), \qquad (2)$$

where ΔG(protein), ΔG(ligand) and ΔG(protein+ligand) are the free energies of the protein, the ligand, and the protein ligand complex, respectively.

The binding energies of competitors from natural amino acids are also calculated, and the differential binding energy between the analog and the best competitor is used as a criterion to select mutants. These mutants will be checked by stability of each mutated residues to make sure that they make enough interactions with the rest of the protein so that the fold is stable.

3. Results and Discussions

3.1 Design for NBD-Ala

Two rotamers of NBD-Ala with low energy were built in Biograf using the same coordinates for zwitterions as in wild-type Trp ligand. These two rotamers along with Trp are shown in Figure 1. There are other rotamers for NBD-Ala, however, they have less overlap with the binding site of Trp and almost certainly they will clash with the protein backbone in the design using COP. Hence, only the two rotamers shown here were used.

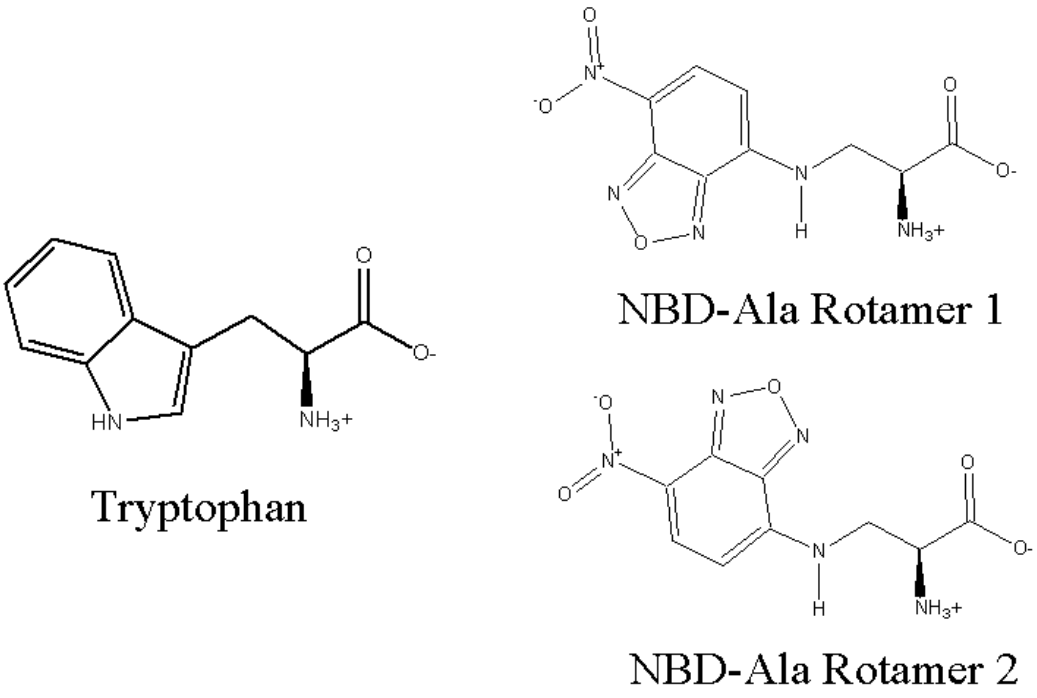


Figure 1. Tryptophan and NBD-Ala (two rotamers) used in the design of mutant TrpRS.

The two rotamers of NBD-Ala were placed into the binding site of TrpRS, and the interaction energies of each residue with NBD-Ala and Trp ligand were calculated using Equation 1. The difference of the interaction between NBD-Ala and Trp was then calculated. Table 1 shows the result of these calculations.

Table 1. Interaction energies (in kcal/mol) of each residue in the binding site with NBD-Ala and Trp. The binding site is defined as with 6 Å of the side chain of NBD-Ala. Residues labeled with * have at least 1 kcal/mol less favorable interactions with NBD-Ala compared to Trp.

Residue	NBD-Ala (r1)	Trp	Difference
G7	-4.53	1.18	-5.70
Q147	-20.16	-18.07	-2.09
S6	-3.81	-2.11	-1.70
C38	-1.07	0.13	-1.20
Q80	-0.35	0.65	-1.00
P142	-0.03	0.69	-0.72
Y125	-17.34	-16.72	-0.61
T138	-0.44	0.03	-0.47
Q9	-12.56	-12.19	-0.37
G144	-0.40	-0.11	-0.29
V40	-1.49	-1.36	-0.13
A174	-0.28	-0.27	-0.01
I151	-0.83	-0.85	0.02
I8	-2.34	-2.38	0.04
H43	-2.60	-2.81	0.21
I140	0.13	-0.16	0.29
F26	-0.17	-0.57	0.40
A22	0.04	-0.58	0.62
I133	-0.65	-1.29	0.63
A130	0.50	-0.38	0.87
L128	0.76	-0.13	0.88
F5	-0.37	-1.46	1.09*
M129	-4.37	-7.03	2.65*
D132	-2.65	-26.57	23.93*
V143	127.41	-2.09	129.51*
V141	982.50	-1.02	983.52*

Table 1. (continued)

Residue	NBD-Ala (r2)	Trp	Difference
Q147	-22.84	-18.07	-4.76
G7	-1.73	1.18	-2.90
S6	-3.50	-2.11	-1.40
D146	7.43	8.67	-1.23
Y125	-17.88	-16.72	-1.16
Q9	-13.01	-12.19	-0.82
I8	-3.18	-2.38	-0.80
I151	-1.24	-0.85	-0.39
A174	-0.51	-0.27	-0.24
G144	-0.35	-0.11	-0.24
C38	-0.08	0.13	-0.21
T138	-0.16	0.03	-0.19
Q80	0.59	0.65	-0.06
F26	-0.63	-0.57	-0.06
I133	-1.29	-1.29	-0.00
I140	-0.13	-0.16	0.03
A22	-0.53	-0.58	0.04
P142	0.86	0.69	0.16
L128	0.06	-0.13	0.19
V40	-1.07	-1.36	0.29
H43	-2.24	-2.81	0.57
M129	-5.48	-7.03	1.55*
F5	0.84	-1.46	2.30*
D132	-12.96	-26.57	13.61*
V143	115.18	-2.09	117.28*
V141	333.89	-1.02	334.91*

From Table 1, five residues (V141, V143, D132, M129 and F5) were found to have less favorable interactions with rotamer 1 of NBD-Ala compared to Trp. The same five residues had less favorable interactions with rotamer 2 of NBD-Ala. Here a cutoff value of 1 kcal/mol was used to select residues to mutate. We have been using the cutoff value of 0.5 kcal/mol, however, it would give us more than five residues to mutate. In our design, we limit the number of residues for mutation to 5. There were no main chain

clash between either rotamer of NBD-Ala and the protein. Therefore, both rotamers were used in further design steps.

Each residue was mutated into 20 natural amino acids one by one, and the interaction energies between the mutated residue and ligands (NBD-Ala and Trp) again were calculated using Equation 1. A score was assigned to each mutation by combining 95% of the differential interaction energy with the mutated residue between NBD-Ala and Trp, plus 5% of the constraint energy of the mutated residue with the rest of the protein. A score cutoff of 0 kcal/mol was used to select good mutations favoring NBD-Ala binding over Trp. Note that the interaction energy between the mutated residue and NBD-Ala had to be favorable for the mutation to be chosen. Table 2 lists the mutations that were chosen for each rotamers of NBD-Ala. Because the interaction energy of these residues with NBD-Ala were all positive except F5 as seen in Table 1, the original choice of each residue was also included in the mutation lists. The reason for doing this is that some mutants might need only three or less mutations to achieve optimal selection.

Table 2. Selected mutations for residues identified in clash calculation

Residues	Mutations for rotamer 1	Mutations for rotamer 2
V141	G A Q	A G Q
V143	A T N G C	T N C A S G
D132	I M N G S A H C T P V	Q W L I M G
M129	M L I	M L I N S
F5	F W N	Y F

Finally these mutations were combined to make mutant TrpRS. A total number of $3 \times 5 \times 11 \times 3 \times 3 = 1485$ mutants were generated for rotamer 1 of NBD-Ala, and for rotamer 2 the number is $3 \times 6 \times 6 \times 5 \times 2 = 1080$. These mutants were each scored by its binding energy to NBD-Ala calculated using Equation 2. A cutoff binding energy of 25 kcal/mol was used to select mutants with good binding energy to NBD-Ala. These selected mutants were further scored with binding energy to Trp, Tyr and Phe, because they were thought to be the main competitors from natural amino acids due to their similarity to NBD-Ala. The mutants were ranked by the differential binding energy between NBD-Ala and the best competitor among Trp, Tyr and Phe. Finally a stability check for each mutation was performed for each mutant. Those mutants with mutations making unfavorable protein-protein interactions were discarded due to their possible problem with folding. Table 3 lists the top mutants designed by COP for NBD-Ala.

Table 3. Binding energies of designed mutants with better binding energy to NBD-Ala than any competitors. The difference is between NBD-Ala and the best competitor. Binding energies are in kcal/mol. (a) Rotamer 1 of NBD-Ala, (b) Rotamer 2 of NBD-Ala.

(a)

V141	V143	D132	M129	F5	NBD-Ala	Phe	Tyr	Trp	Difference
G	T	M	M	F	41.27	32.30	30.32	-159.00	8.97
G	C	M	M	F	52.40	44.10	41.75	-97.39	8.30
A	T	M	M	F	53.27	45.21	43.13	-103.92	8.06
A	G	M	M	F	52.14	44.18	41.93	-81.31	7.96
G	C	I	I	F	33.56	25.61	22.66	-78.39	7.95
G	T	I	I	F	35.07	27.16	24.82	-179.72	7.91
A	A	M	M	F	51.62	44.08	41.95	-116.13	7.54
A	G	I	I	F	33.11	25.74	23.28	-58.68	7.37
A	T	I	I	F	34.02	26.88	24.59	-186.32	7.14
A	C	I	I	F	32.50	25.54	22.77	-199.37	6.96
A	A	I	I	F	32.59	25.69	23.20	-58.50	6.90

Table 3. (continued)

(b)

V141	V143	D132	M129	F5	NBD-Ala	Phe	Tyr	Trp	Difference
G	T	I	L	F	45.79	32.74	29.59	-238.22	13.05
G	T	I	N	F	46.44	32.13	33.71	-80.73	12.73
G	T	L	S	F	40.79	32.11	26.48	-230.24	8.68
A	T	L	S	F	40.01	31.93	26.51	-313.87	8.08
G	T	M	I	F	39.69	31.70	29.93	-119.66	7.99
A	T	M	I	F	38.89	31.68	29.93	-114.50	7.21
G	T	M	L	F	35.50	28.69	22.28	-124.12	6.81
G	C	L	N	F	29.11	22.33	15.97	-264.66	6.78
G	T	I	M	F	40.84	34.18	29.90	-78.08	6.66
G	S	I	I	F	32.17	25.93	23.90	-59.58	6.24
A	T	M	L	F	34.65	28.61	22.32	-166.98	6.04
G	C	I	I	F	31.58	25.61	22.66	-78.39	5.97
A	S	I	I	F	31.44	25.83	23.60	-130.75	5.61

Most of the mutants generated have been shown to not have enough favorable interactions between the mutated residues and the rest of the protein; therefore they might not fold correctly. These mutants were eliminated from further consideration.

Among the mutants designed based on rotamer 1 of NBD-Ala, F5 remained F for all of them. Interestingly V141 and M129 were always mutated to the same residue. The V141 mutation, which had severe clash with NBD-Ala before mutation, can be either G or A. This mutation is presumably for relieving the clash between the nitro group of NBD-Ala and the protein. M129 seemed to be less critical as it could be either M or I. Both seem to have the same size. The V143T mutation formed a hydrogen bond with the nitro group (Figure 2). D132 recognizes the side chain NH of the Trp ligand in wild-type TrpRS, and mutation D132I blocks the normal binding of Trp. As a result, Trp is hardly a competitor in the competitive binding (Table 3).

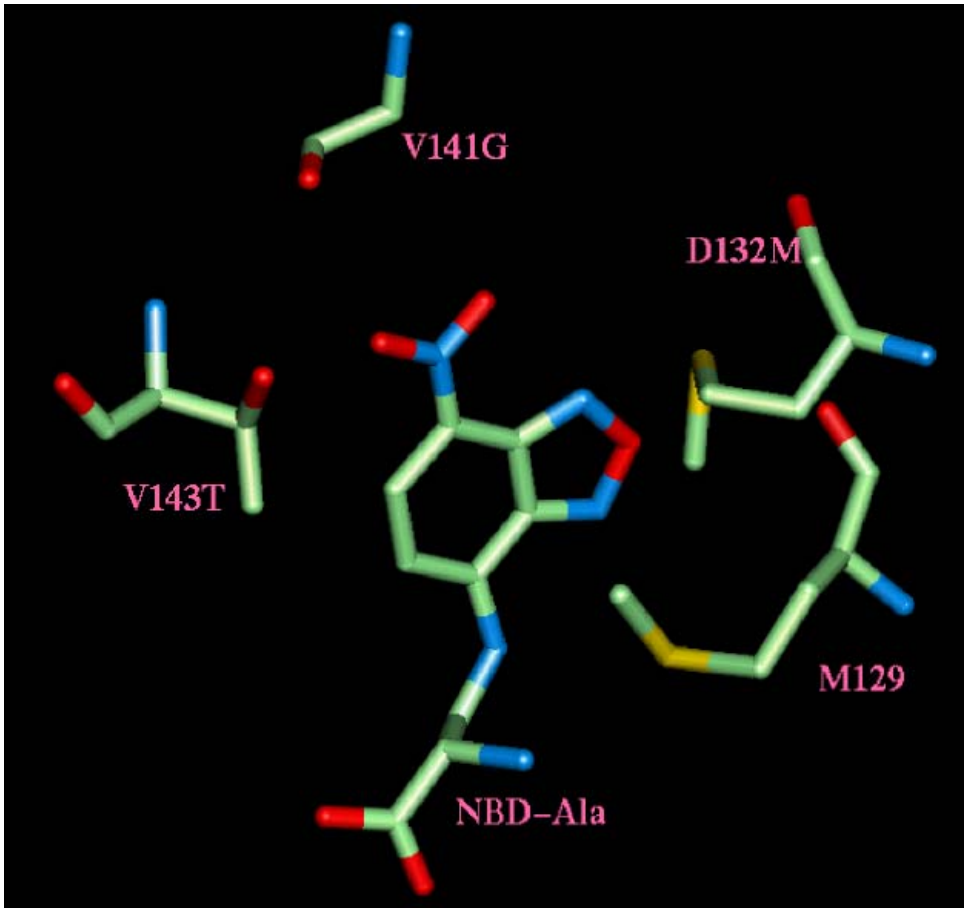


Figure 2. The binding site of NBD-Ala in one of the mutants designed by COP. V143T forms a hydrogen bond with the nitro group.

Thirteen mutants were designed based on rotamer 2 of NBD-Ala. Similar to the case of rotamer 1, mutation V141 to G or A allows the nitro group to go in further in the binding site, and V143T forms a hydrogen bond with the five-member ring of the NBD. The choice for D132 was among M, L and I. Again, this mutation blocks the binding of Trp, and Trp indeed binds poorly to these mutants. The position M129 seems to be less discriminative, and can be any of L, M, S, I or N. F5 did not change in these mutants.

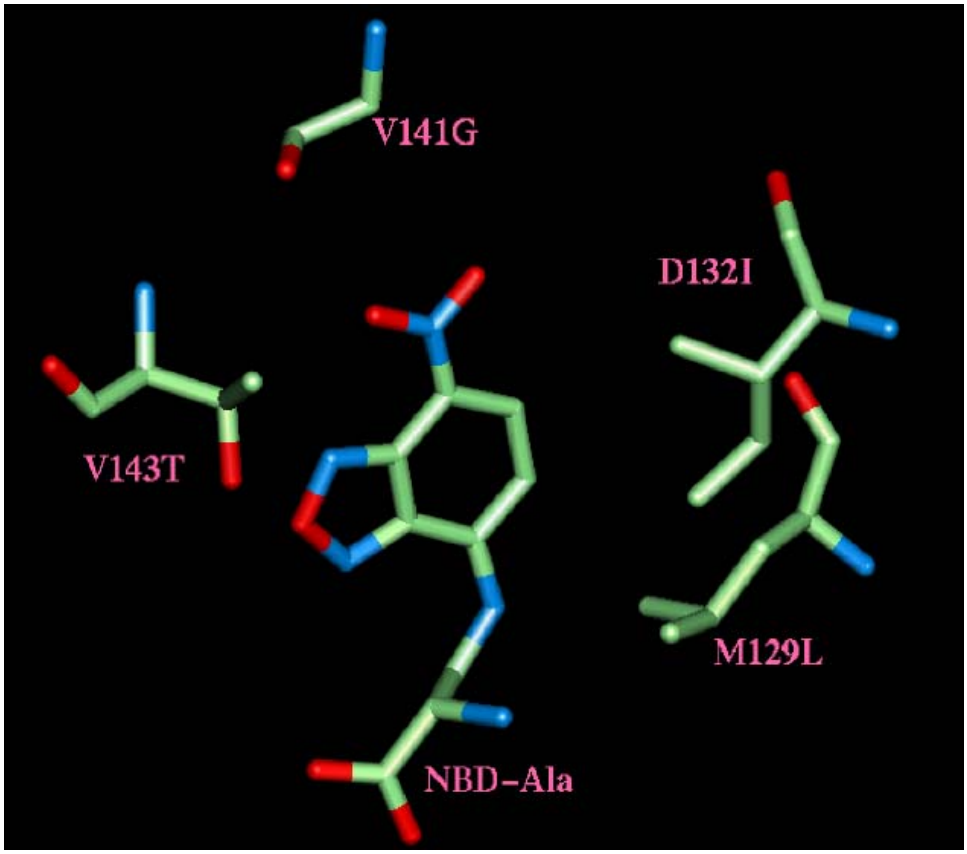


Figure 3. The mutations in mutant V141G-V143T-D132I-M129L-F5F designed by COP based on rotamer 2 of NBD-Ala.

3.2 Design for bpy-Ala

The second design case was for bpy-Ala. Quantum mechanics calculation using Jaguar showed that the trans-conformation was 8.39 kcal/mol lower in energy than the cis-conformation. Bpy-Ala is a good binding agent to transitional metal ions. In free solution it is usually in trans-conformation, and upon binding it switches to cis-conformation to form coordinated binding from both nitrogen atoms. Hence only the trans-conformation was used in preparing the rotamers. Two low energy rotamers in

trans-conformation were built in Biograf. These two rotamers were shown in Figure 4. To ensure that the binding mode was the same as for Trp, the coordinates of the zwitterions were borrowed from Trp.

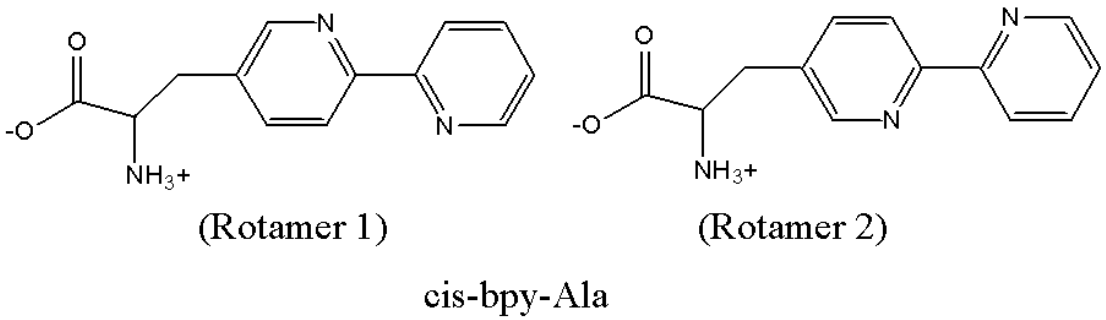


Figure 4. Two rotamers of bpy-Ala used in COP design of mutant TrpRS.

When these rotamers were simply placed in the binding site of TrpRS, they had very bad clash with the backbone of TrpRS. The reason was that these rotamers were not aligned optimally with the binding site for Trp. Figure 5 showed the alignment of bpy-Ala and Trp before and after optimization, and it is seen that the alignment is significantly better after optimization.

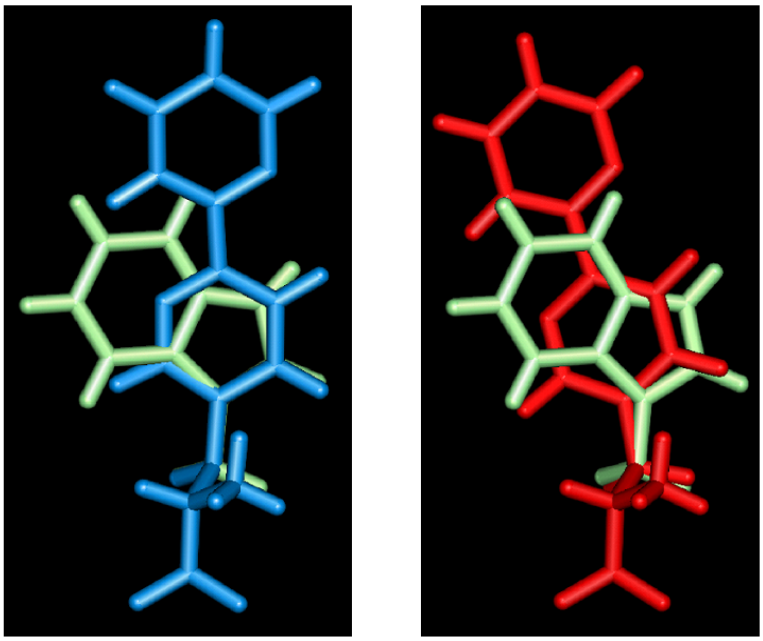


Figure 5. Alignment between Trp and bpy-Ala before and after optimization. Trp is in green, bpy-Ala is in blue (before optimization) and red (after optimization).

In order to optimize the orientation of the side chain of bpy-Ala, the following procedure was adopted to get the conformation with lowest backbone clash energy. First a mutant TrpRS with all Gly was generated, i.e., all the side chains of the protein were taken out. A grid of conformations for bpy-Ala was then generated by changing the $\chi 1$ and $\chi 2$ angles ($\chi 1$ was defined as the dihedral angle of N-Cα-Cβ-Cγ, and $\chi 2$ was defined as the dihedral angle of Cα-Cβ-Cγ-Cδ, see Figure 4). Finally the bpy-Ala analog was put into the binding site of the all Gly TrpRS, and the energy of the complex was calculated after 10 steps of steepest descents minimization of the analog with the protein fixed. The energy was plotted in a two-dimensional energy surface plot (Figure 6).

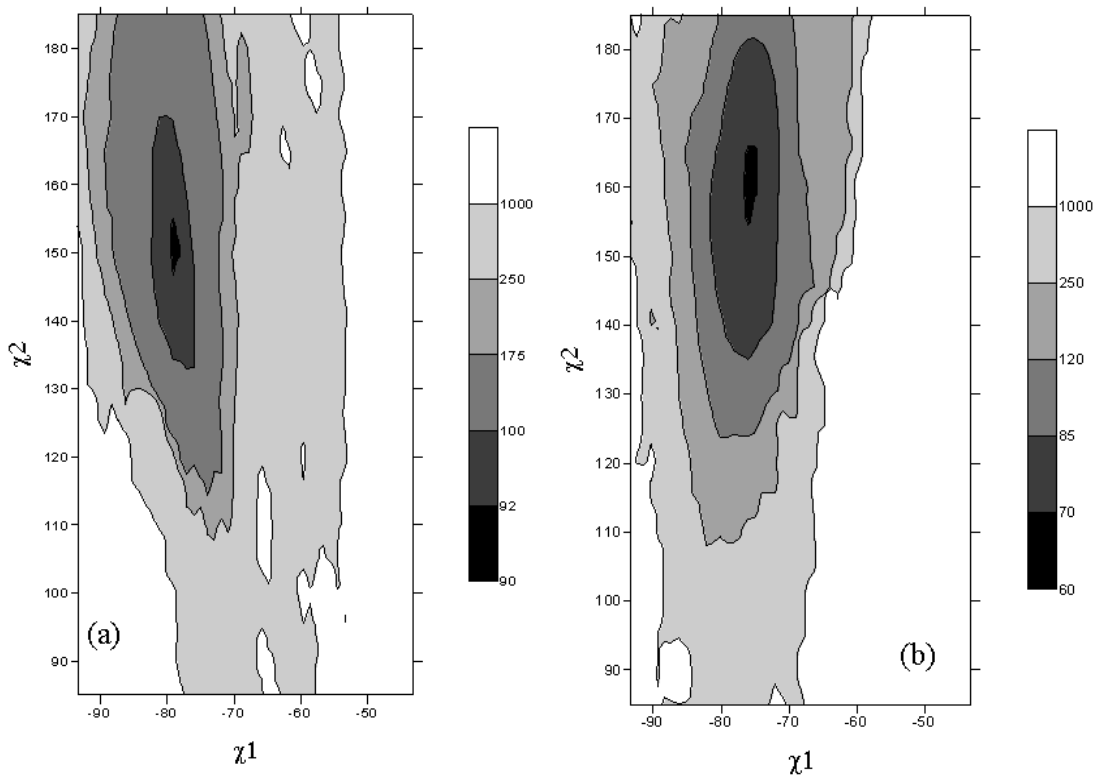


Figure 6. The energy surface of interaction between the all Gly TrpRS (backbone) with bpy-Ala by changing the χ1 and χ2 angles of bpy-Ala for (a) rotamer 1 and (b) rotamer 2. Ten steps of steepest descents minimization were performed before the energy evaluation.

The minimum in the energy surface plot for rotamer 1 of bpy-Ala had $\chi1 = -78.5^{o}$ and $\chi2 = 150^{o}$, and for rotamer 2 the minimum was at $\chi1 = -76^{o}$ and $\chi2 = 160^{o}$. The conformation of the two rotamers was adjusted to these values.

These two rotamers were then placed in the binding site of the wild-type TrpRS, and the binding contribution from each residue in the binding site was calculated using Equation 1. The same was done for the Trp ligand, and the difference was calculated

between bpy-Ala (both rotamers) and Trp for each residue. These results were listed in Table 4.

Table 4. The interaction energies (in kcal/mol) of each residue in the binding site of TrpRS with bpy-Ala (both rotamers) and Trp. The residues were sorted by the interaction energy difference with bpy-Ala and Trp. The numbers in ( ) indicate clash with the backbone of TrpRS.

Residue	bpy-Ala (rotamer 1)	Trp	Difference
Q9	-12.76	-11.48	-1.28
Q147	-17.74	-16.61	-1.13
C38	-0.77	0.02	-0.79
G7	-0.22	0.29	-0.51
S6	-2.26	-2.06	-0.19
I8	-2.20	-2.19	-0.01
I151	-0.50	-0.76	0.26
V143	-1.07	-1.88	0.81
M129	-5.60	-6.75	1.15
H150	1.21	-0.44	1.65
V40	3.24	-1.48	4.71
D132	-5.29	-22.82	17.53
V141	30.36	-1.03	31.40
I133	241.02(12.84)	-1.36	242.39
F5	6251.94	-1.55	6253.48

Residue	bpy-Ala (rotamer 2)	Trp	Difference
S6	-3.93	-2.06	-1.87
G7	-1.05	0.29	-1.34
Q9	-12.78	-11.48	-1.30
Q147	-17.67	-16.61	-1.06
C38	-1.00	0.02	-1.02
Q80	-0.30	0.45	-0.75
H43	-2.78	-2.65	-0.13
I8	-2.15	-2.19	0.04
M129	-4.98	-6.75	1.77
V40	8.79	-1.48	10.27
I133	35.93(9.76)	-1.36	37.30
D132	29.24	-22.82	52.06
F5	468827.98	-1.55	468829.53

From Table 4, it is obviously that both rotamers still had clash with the backbone of the protein. However, the backbone clash was not very severe, and usually several steps of minimization could relieve these clashes significantly. However, rotamer 1 had 8 residues that need to be mutated, if we use our normal cutoff of 0.5 kcal/mol. When there is more than 5 resides involved in the mutation, it usually means too much change for the binding site. Therefore, we chose to design for rotamer 2, which had less main chain clash and only 5 mutations to do. The 5 mutation residues were F5, D132, I133, V40, and M129. A cutoff of 0.5 kcal/mol was used here.

We then tried to mutate each of the five residues to all 20 natural amino acids one by one. The mutated residue conformation was chosen from a rotamer library with the lowest energy rotamer being selected. A score was calculated as 95% of the differential interaction energy of the mutated residue with bpy-Ala and Trp, plus 5% of the constraint energy between the mutated residue and the rest of the protein. Mutations with positive score were chosen, and the choices were listed in Table 5.

Table 5. Selected mutations for residues identified in clash calculation

Residues	Mutations
F5	G
D132	G A
I133	T S M A G C
V40	A G
M129	M L I F H C V

The mutants were generated by combining the mutations from each residue, and a total number of $1 \times 2 \times 6 \times 2 \times 7 = 168$ mutants were generated. Each mutant was scored by the binding energy to bpy-Ala using Equation 2. A cutoff of 25 kcal/mol was used to select good binding mutants to calculate binding energies to competitors from natural amino acids. Here we assumed Trp, Tyr and Phe as competitors because their similar size. The difference of the binding energy between bpy-Ala and the competitor with the best binding energy was calculated and used to select best mutants. Finally a stability check for each mutation was performed for each mutant. Those mutants with mutations making unfavorable protein-protein interactions were discarded due to their possible problem with folding. Table 6 listed all 14 of those mutants with binding energy to bpy-Ala at least 5 kcal/mol better than any of the competitors.

Table 6. Binding energies of designed mutants with better binding energy to bpy-Ala than any competitors. The difference is between bpy-Ala and the best competitor. Binding energies are in kcal/mol.

F5	D132	I133	V40	M129	bpy-Ala	Phe	Tyr	Trp	Difference
G	A	T	A	C	38.81	30.28	27.10	27.11	8.53
G	A	A	A	C	37.94	30.28	27.03	26.59	7.66
G	A	S	A	C	37.92	30.31	27.02	27.21	7.61
G	G	T	A	C	36.53	29.74	27.75	29.60	6.79
G	G	A	A	C	35.98	29.69	27.75	23.28	6.29
G	G	S	A	C	35.71	29.76	27.70	29.77	5.94
G	A	M	A	C	37.37	31.67	28.37	28.15	5.70

Among these seven mutants designed by COP, three residues have the same mutation in all of them. These mutations are F5G, V40A and M129C. D132 is mutated to either G or A. I133, which has slight clash with bpy-Ala in main chain, can be mutated into A, S, T, or M. Figure 7 shows the binding site of bpy-Ala formed by these mutations in mutant F5G-D132A-I133T-V40A-M129C. It is seen in Figure 7 that the extra six-member ring takes the space opened by mutation F5G. D132A also opens some space for the extra six-member ring in bpy-Ala. Other mutations contribute to the binding of bpy-Ala by shaping the binding site according to the orientation assumed by bpy-Ala upon its binding. V40A makes the orientation of the Cβ-Cγ bond possible. I133T seems to form a weak hydrogen bond with the nitrogen atom in the second six-member ring in bpy-Ala. The distance between the nitrogen atom and the Oγ1 in T133 is 3.2 Å.

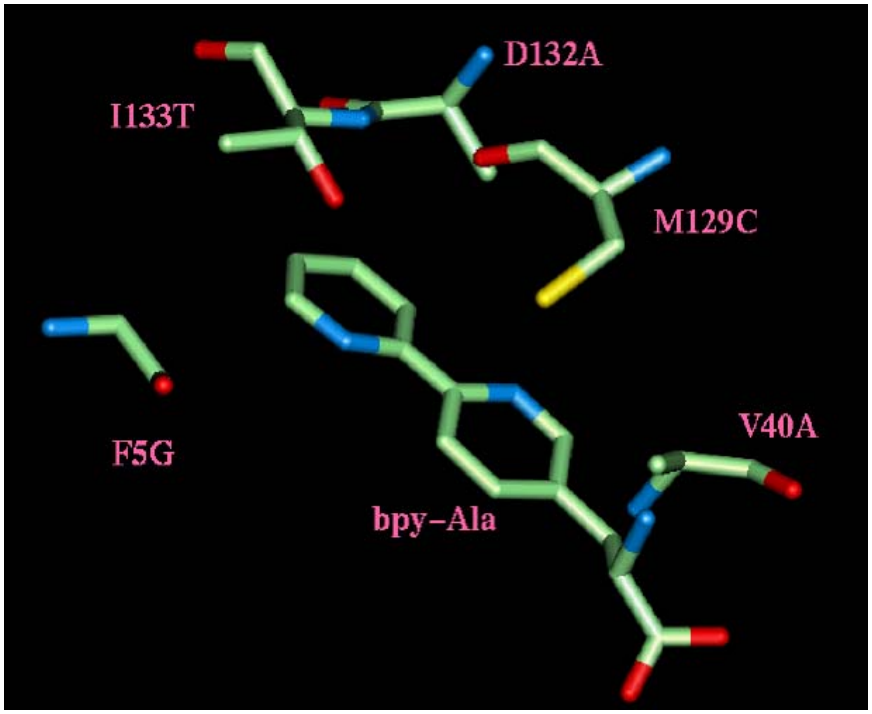


Figure 7. The mutations in the binding site of bpy-Ala in mutant F5G-D132A-I133T-V40A-M129C.

3.3 Efforts in designing for DAN-Ala

2-Amino-4-(6-dimethylamino-naphthalen-2-yl)-4-oxo-butyric acid (DAN-Ala) is a synthetic fluorescent amino acid, and it was reported earlier this year that it could be used as an internal probe for protein electrostatics (18). The fluorophore it carries is 6-dimethylamino-2-naphthalene (DAN), and DAN undergoes large charge redistribution upon excitation and has nearly ideal environment sensor property (19). DAN-Ala has been incorporated into the B1 domain of streptococcal protein G and the Kir2.1 and Shaker potassium ion channels using chemical synthesis method (18). Due to its potential as a protein electrostatic probe, it would be very interesting to see if COP can be used to design a mutant AARS to recognize DAN-Ala.

Figure 8 shows the structure of DAN-Ala along with Trp. Compared with Trp, DAN-Ala side chain is about 3 to 4 C—C bond length longer than Trp. And indeed we saw very bad main chain clash with DAN-Ala on residues V141, V143 and P142 (Table 7). Therefore, our effect on design for DAN-Ala failed.

Table 7. Clash calculation with DAN-Ala in the binding site of TrpRS

Residue	DAN-Ala	Trp	Difference
S6	-0.36	-2.05	1.69
M129	-5.20	-7.26	2.06
Q147	-14.62	-19.03	4.41
F5	4.68 (6.31)*	-1.42	6.10
D132	2.55	-28.41	30.97
V143	113649.80 (339.54)*	-1.92	113651.72
V141	668295.71 (211526.31)*	-1.03	668296.74
P142	1463046.56 (1463044.84)*	0.59	1463045.97

Note: * indicates main chain clash

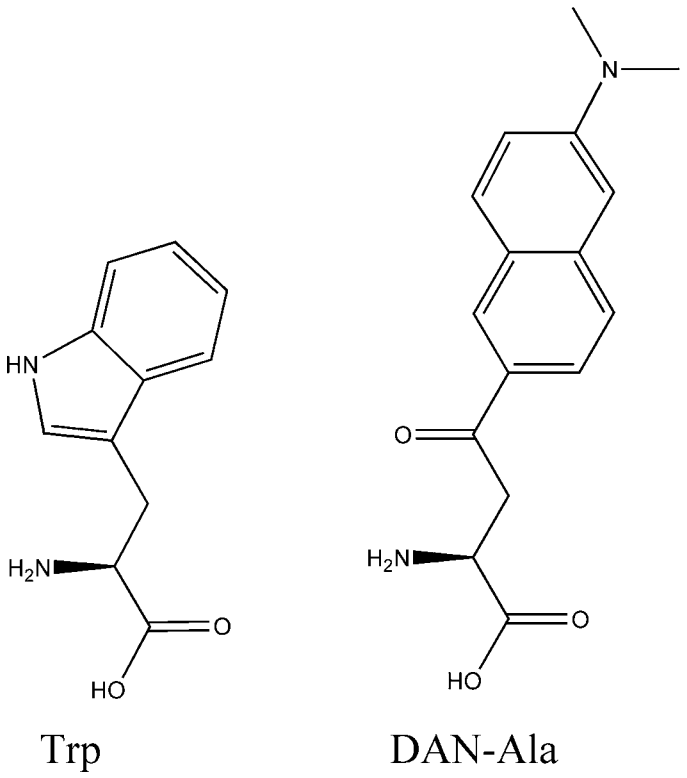


Figure 8. The structure of DAN-Ala along with Trp

Conclusions

We have used the COP protein design tool to design mutant AARSs for recognizing NBD-Ala, bpy-Ala using TrpRS. Some mutants showing good preferential binding energy to these non-natural amino acids were designed. Currently mutant TrpRS:tRNA pairs are being evolved experimentally, and it would be interesting to test the efficiency of these mutants designed here.

A third non-natural amino acid, DAN-Ala, has large main chain clash with several residues in the binding site, therefore COP failed to design any mutant.

References

1. Richmond, M. H. (1963) *J. Mol. Biol.* **6,** 284-294.
2. Yoshikawa, E., Fournier, M. J., Mason, T. L., & Tirrell, D. A. (1994) *Macromolecules* **27,** 5471-5475.
3. Deming, T. J., Fournier, M. J., Mason, T. L. & Tirrell, D. A. (1997) *J. Macromol. Sci. Pure Appl. Chem.* **A34,** 2143-2150.
4. van Hest, J. C. M. & Tirrell, D. A. (1998) *FEBS Lett.* **428,** 68-70.
5. van Hest, J. C. M., Kiick, K. L. & Tirrell, D. A. (2000) *J. Am. Chem. Soc.* **122,** 1282-1288.
6. Kothakota, S., Mason, T. L., Tirrell, D. A. & Fournier, M. J. (1995) *J. Am. Chem. Soc.* **117,** 536-537.
7. Cowie, D. B. & Cohen, G. N. (1957) *Biochim. Biophys. Acta* **26,** 252-261.
8. Dougherty, M. J., Kothakota, S., Mason, T. L., Tirrell, D. A. & Fournier, M. J. (1993) *Macromolecules* **26,** 1779-1781.
9. Budisa, N., Steipe, B., Demange, P., Eckerskorn, C., Kellermann, J. & Huber, R. (1995) *Eur. J. Biochem.* **230,** 788-796.
10. Duewel, H., Daub, E., Robinson, V. & Honek, J. F. (1997) *Biochemistry* **36,** 3404-3416s.
11. Zhang, D. Q., Vaidehi, N., Goddard, W. A., Danzer, J. F. & Debe, D. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6579-6584.
12. Retailleau, P., Yin, Y. H., Hu, M., Roach, J., Bricogne, G., Vonrhein, C., Roversi, P., Blanc, E., Sweet, R. M. & Carter, C. W. (2001) *Acta Crystallogr.D* **57,** 1595-1608.
13. Lim, K.-T., Brunett, S., Iotov, M., McClurg, R. B., Vaidehi, N., Dasgupta, S., Taylor, S. & Goddard, W. A., III (1997) *J. Comput. Chem.* **18,** 501-521.
14. Mayo, S. L., Olafson, B. D. & Goddard, W. A., III (1990) *J. Phys. Chem.* **94,** 8897-8909.
15. Ghosh, A., Rapp, C. S. & Friesner, R. A. (1998) *J. Phys. Chem. B* **102,** 10983-10990.
16. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998) *J. Phys. Chem. B* **102,** 3586-3616.
17. Tannor, D. J., Marten, B., Murphy, R., Friesner, R. A., Sitkoff, D., Nicholls, A., Ringnalda, M., Goddard, W. A. I. & Honig, B. (1994) *J. Am. Chem. Soc.* **116,** 11875-11882.
18. Cohen, B. E., McAnaney, T. B., Park, E. S., Jan, Y. N., Boxer, S. G. & Jan, L. Y. (2002) *Science* **296,** 1700-1703.
19. Macgregor, R. B. & Weber, G. (1986) *Nature* **319,** 70-73.

Appendix I

Protein Dynamics in a Family of Laboratory Evolved Thermophilic Enzymes*

* This appendix is adapted from a paper to JMB (in press) and coauthored with Patrick L. Wintrode, Nagarajan Vaidehi, Frances H. Arnold and William A. Goddard, III

Summary

Molecular dynamics simulations were employed to study how protein solution structure and dynamics are affected by adaptation to high temperature. Simulations were carried out on a *para*-nitrobenzyl esterase (~450 residues) and two thermostable variants that were generated by laboratory evolution. Although these variants display much higher melting temperatures than wild-type (up to 18 ^{o}C higher) they are both >97% identical in sequence to the wild-type. In simulations at 300 K the thermostable variants remain closer to their crystal structures than wild-type. However, they also display increased fluctuations about their time-averaged structures. Additionally, both variants show a small but significant increase in radius of gyration relative to wild-type. The vibrational density of states was calculated for each of the esterases. While the density of states profiles are similar overall, both thermostable mutants show increased populations of the very lowest frequency modes (< 10 cm^{-1}), with the more stable mutant showing the larger increase. This indicates that the thermally stable variants experience increased concerted motions relative to wild-type. Taken together, these data suggest that adaptation for high temperature stability has resulted in a restriction of large deviations from the native state and a corresponding increase in smaller scale fluctuations about the native state. These fluctuations contribute to entropy and hence to the stability of the native state. The largest changes in localized dynamics occur in surface loops, while other regions, particularly the active site residues, remain essentially unchanged. Several mutations, most notably L313F and H322Y in variant 8G8, are in the region showing the largest increase in

fluctuations, suggesting that these mutations confer more flexibility to the loops. As a validation of our simulations, the fluctuations of Trp102 were examined in detail, and compared with Trp102 phosphorescence lifetimes that were previously measured. Consistent with expectations from the theory of phosphorescence, an inverse correlation between out of plane fluctuations on the picosecond time scale and phosphorescence lifetime was observed.

Introduction

The physical basis for the remarkable stability of enzymes isolated from thermophilic organisms has been the subject of intensive research [1; 2]. There are now numerous studies comparing the sequences and structures of thermophilic enzymes with those of homologous enzymes from mesophilic organisms (ibid). These studies have found many types of stabilizing interactions in thermophilic enzymes, and there does not appear to be a single preferred mechanism for stabilization. In general, it appears that thermophilic enzymes have adapted to high temperature through the accumulation of numerous mildly stabilizing interactions, including salt bridges, hydrogen bonds, and van der Waals contacts.

Researchers have also focused on the dynamic properties of thermophilic enzymes. The conformational flexibility of homologous thermophilic and mesophilic enzymes has been probed using different techniques, such as fluorescence quenching [3], hydrogen/deuterium exchange [4; 5; 6], molecular dynamics (MD) simulations [7], and neutron scattering [8; 9]. Many of these studies have found that the conformational flexibility of thermophilic enzymes at room temperature is considerably reduced compared to mesophilic enzymes. At the same time, the flexibility of thermophilic enzymes near their physiological (high) temperatures is often comparable to that of mesophilic enzymes at their physiological (moderate) temperatures [3; 4]. These findings concluded that reduced flexibility is a necessary consequence of thermostabilization, i.e., more stable proteins are less prone to have their structures perturbed by thermal fluctuations and therefore appear less flexible. Some researchers have further concluded that these differences in

conformational flexibility have functional consequences and can explain observed differences in the temperature-activity profiles of thermophilic and mesophilic enzymes (for example, the fact that thermophilic enzymes generally display poor activity at moderate temperatures) [1]. Briefly, the argument states that conformational fluctuations in enzymes play an important role in their function as catalysts, but these fluctuations can also lead to the loss of structure and function if they become too large. Since the magnitude of the fluctuations experienced by an enzyme will depend on the available thermal energy, k_BT, evolution has modified the strength and number of stabilizing interactions in enzymes to achieve the optimal balance of stability and flexibility at a given temperature. As a result, large changes in temperature will disrupt this balance, causing cold adapted enzymes to become unstable (at high temperatures), and thermophilic enzymes to become too rigid to function effectively (at low temperatures).

The results of several recent studies, however, are in marked contrast to those cited above. From the millisecond timescale dynamics of the hyperthermophilic rubredoxin from *Pyrococcus furiosus* investigated by NMR-monitored hydrogen/deuterium exchange, it was concluded that the protein's conformational flexibility at room temperature is indistinguishable from that of mesophilic proteins on this time scale [6]. The room temperature dynamics of a pair of mesophilic and thermophilic α-amylases were probed using both hydrogen exchange and inelastic neutron scattering [8]. This study also found no discernable difference in dynamics as monitored by hydrogen exchange, and found *increased* mobility on the picosecond time scale in the thermophilic protein, as measured by neutron scattering.

The confusion regarding the relationship of conformational dynamics to stability and function in proteins stems from several sources. Firstly, different studies have monitored flexibility using different techniques. While all of these techniques are sensitive to protein conformational fluctuations, they often monitor very different aspects of these fluctuations. Fluorescence quenching relies on the quenching of fluorescing tryptophan residues by acrylamide, and is therefore only sensitive to those motions that allow acrylamide molecules to penetrate into the core of proteins and interact with buried tryptophans. Hydrogen/deuterium (H/D) exchange is sensitive to both local and global unfolding motions that expose buried amide hydrogens to water. However, under EX2 conditions, where most studies have been performed, H/D exchange rates are proportional to the equilibrium constant for the conformational change(s) that result in the exposure of a given hydrogen [10]. H/D exchange under these conditions is therefore a static measure of flexibility: it reflects the equilibrium populations of different conformations [11; 12]. Inelastic neutron scattering is sensitive to the motions of individual hydrogens on the picosecond time scale [9], but it provides no spatial resolution. One can only measure the distribution of amplitudes for an entire protein and thus cannot assign a given amplitude of motion or relaxation time to a particular hydrogen atom.

In addition to the fact that protein mobility was probed using different techniques, the various studies of dynamics in thermally stable enzymes have employed different proteins, often with quite distinct native topologies. It is possible that different protein structures have had their dynamics altered in different ways by adaptation to high temperature. Finally, a single pair of homologous thermophilic and mesophilic enzymes

will typically differ at many (often >100) amino acid positions[13]. While some of these amino acid differences will be related to high temperature adaptation, many others will be neutral [14]—the result of genetic drift—or will reflect adaptation of other enzyme properties. Although not all amino acid differences between a mesophilic enzyme and its thermophilic cousin will be directly related to temperature adaptation, they all have the potential to affect dynamics. It is therefore not straightforward to determine which observed differences in protein dynamics are related to high temperature adaptation and which are the result of neutral drift or adaptation to unrelated properties, a difficulty analogous to that of interpreting differences in amino acid sequences [15].

The ambiguities introduced by the presence of non-adaptive mutations can be avoided by studying a family of extremophilic enzymes evolved in the laboratory [15; 16]. A *para*-nitrobenzyl esterase (~450 residues) from *B. subtilus* was evolved for increased thermostability while its activity at room temperature was retained [17; 18]. The final, eighth generation mutant 8G8 had a melting temperature 18^{o}C higher than wild-type and had room temperature activity twice that of wild-type. In spite of these large functional differences, 8G8 differs from wild-type at only 13 out of 490 amino acid positions. Subsequently, the structures of wild-type, an intermediate mutant in the evolutionary pathway (referred to as 56C8) and the thermophilic mutant, 8G8, were determined by x-ray crystallography[19]. The three-dimensional structure of the thermostable esterase 8G8, including the locations of the thermostabilizing mutations, is shown in Figure 1.

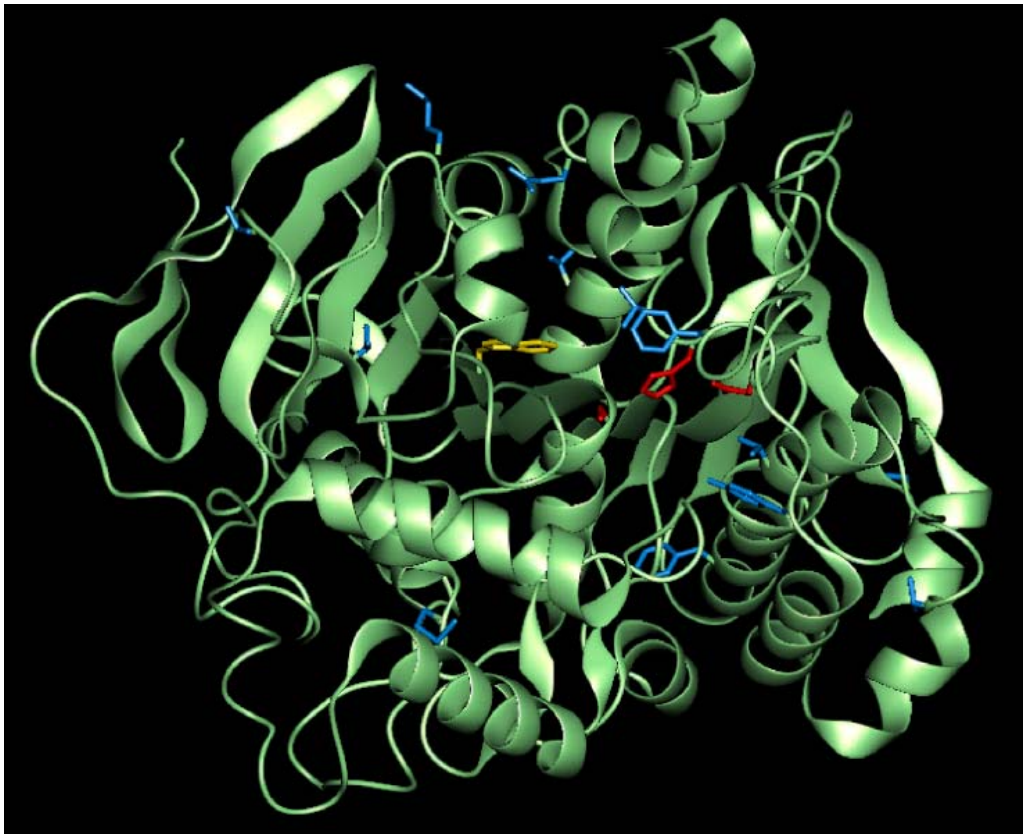


Figure 1. Three-dimensional structure of the *p*NB esterase mutant 8G8. The catalytic triad is shown in red and the 13 mutations from wild-type are shown in blue. Trp 102 is shown is yellow.

In order to investigate the relationship between stability, dynamics, and evolution, we have used the crystal structures of wild-type, 56C8, and 8G8 as the starting point for MD simulations including solvation. These MD simulations show the differences in dynamics between a set of proteins with a wide range of thermal stabilities, evolved under known selection pressures and differing by only a small number of functional mutations.

Results

Validation of the simulations

The overall calculated root-mean-square deviation (CRMS) of all atoms of the minimized wild-type, 8G8 and 56C8 structures from their respective crystal structures are 0.39 Å, 0.72 Å and 0.79 Å. This shows that the forcefield and the surface generalized Born solvation method [20] are suitable for describing the dynamics of the system. Figure 2 a shows the overall CRMS from the crystal structure for all atoms during the MD simulations. The large CRMS change in the initial several picoseconds is the result of heating the system from 0K to 300 K, and a simulation of the wild-type at 77 K shows a much smaller CRMS of 1.3 Å after equilibrium (data not shown). From Figure 2 b, it is clear that the system has equilibrated after 100 ps of simulations (within ~0.1 Å CRMS of the average). The overall fluctuations from the crystal structure are smaller for the thermostabilized mutants 8G8 and 56C8 than for wild-type. Figure 3 a, b and c compare the calculated CRMS for all the residues to the temperature factor (B factor) reported in the crystal structure for wild-type, 56C8 and 8G8, respectively. The fluctuations in CRMS by residue correlate well with the temperature factors from crystallographic data. For example in Figure 3c, residues 148 to 153 and 312 to 322 in 8G8 have high CRMS and also show large temperature factors. Both the high B factors and the large fluctuations during the simulations are expected for these regions, as they are composed of surface loops. This correlation between experimental and calculated fluctuations provides validation that our simulations accurately describe the dynamics of the system.

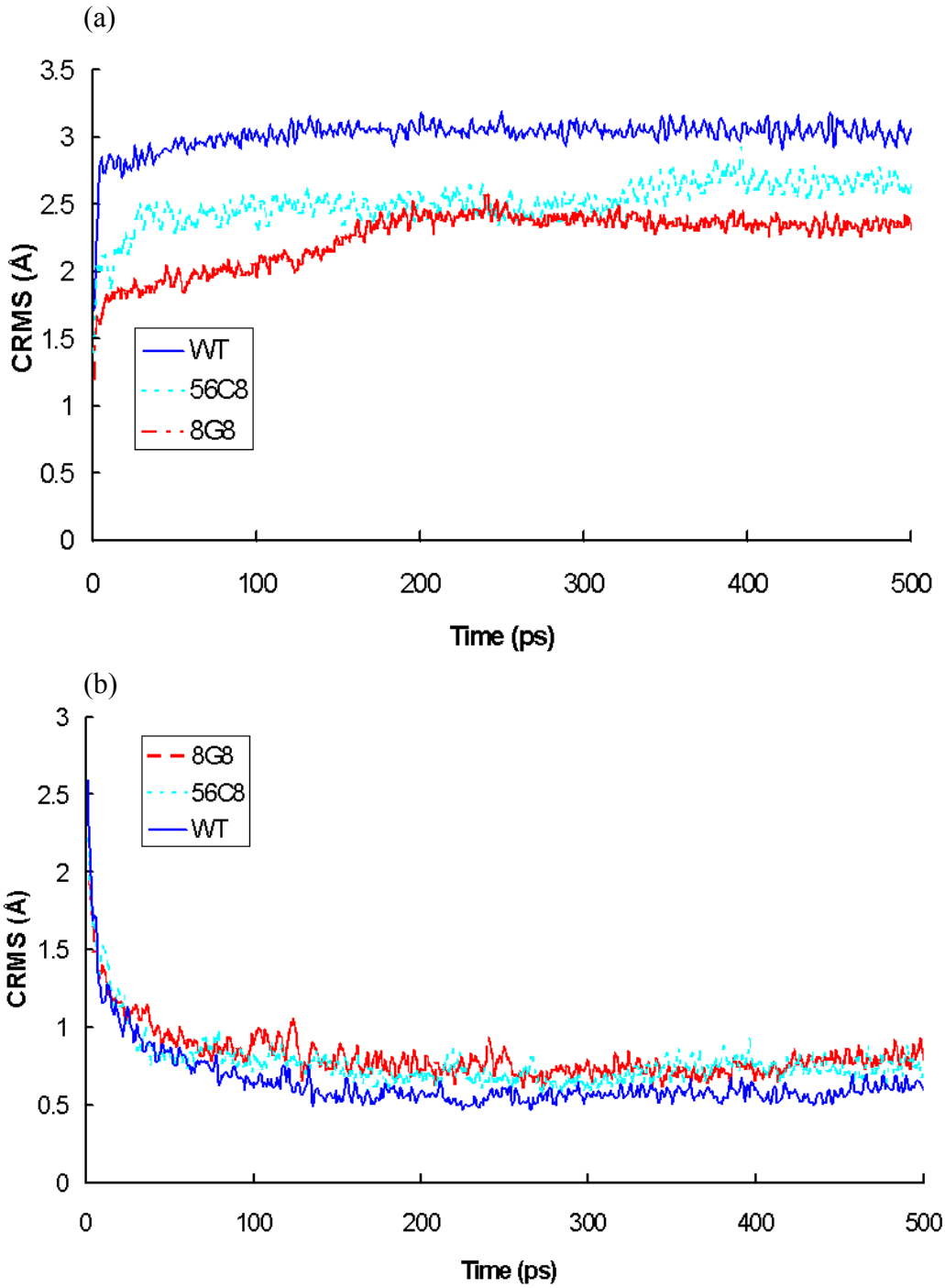
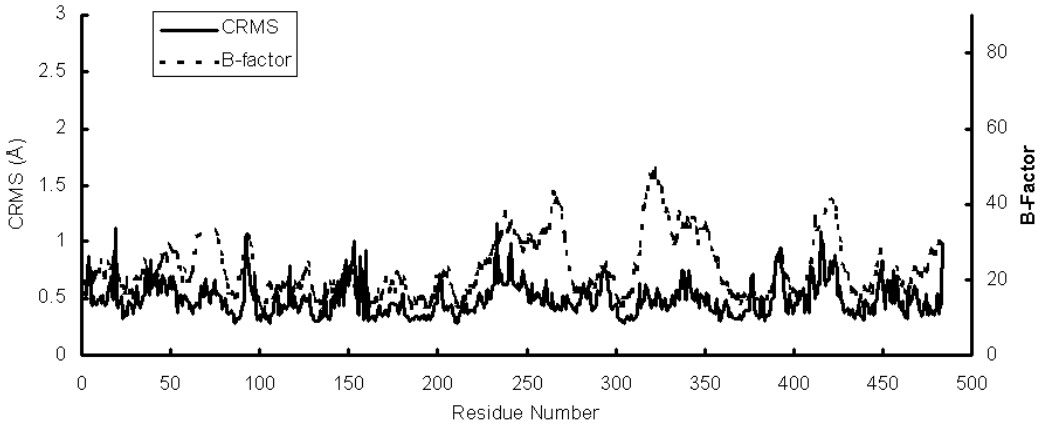
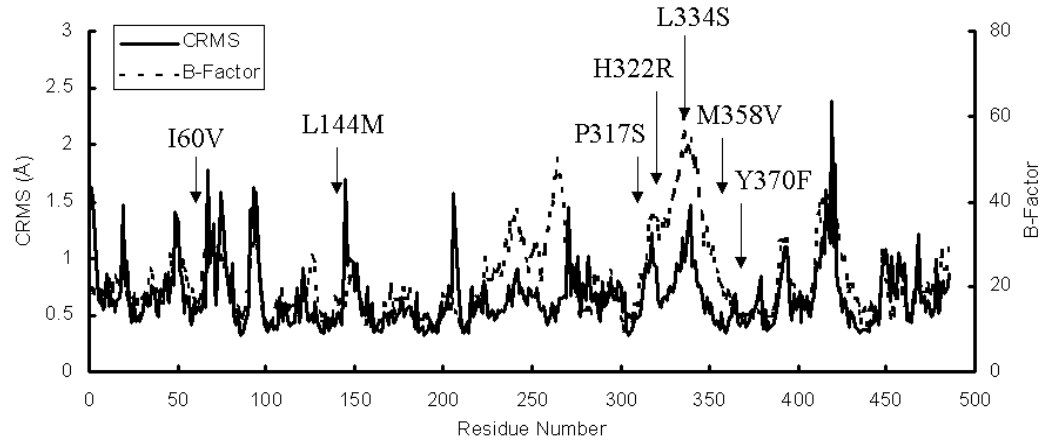


Figure 2. Overall all-atom CRMS from (a) the minimized crystal structure, and (b) the time-averaged dynamic structure for wild-type *p*NBE and the two mutants as functions of time. (Color coding: WT-blue, 56C8-cyan, 8G8-red)

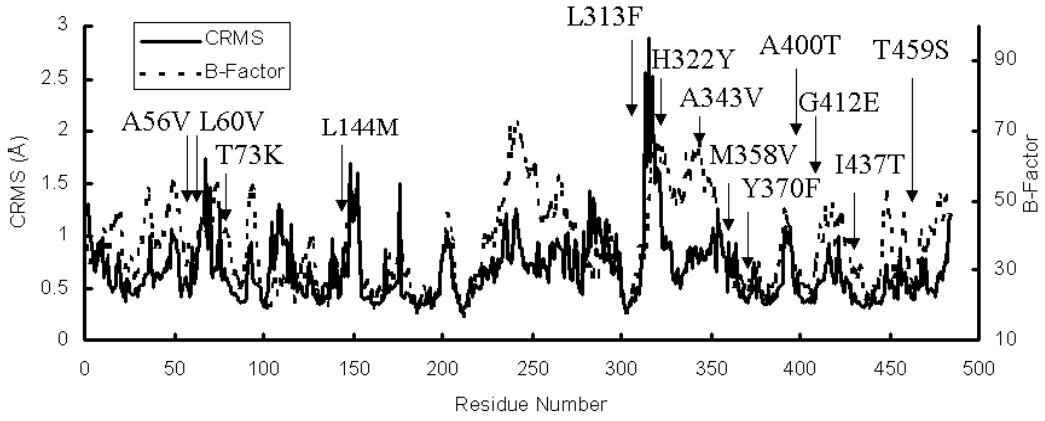
(a)


CRMS
B-factor
CRMS (Å)
B-Factor
Residue Number

(b)


CRMS
B-Factor
I60V
L144M
P317S
H322R
L334S
M358V
Y370F
CRMS (Å)
B-Factor
Residue Number

(c)


CRMS
B-Factor
A56V
L60V
T73K
L144M
L313F
H322Y
A343V
M358V
Y370F
A400T
G412E
I437T
T459S
CRMS (Å)
B-Factor
Residue Number

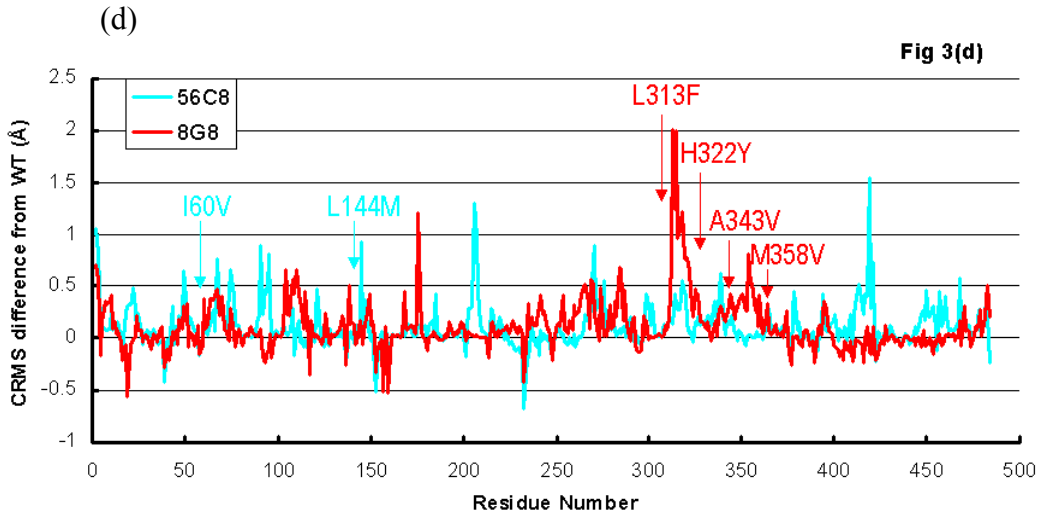


Figure 3. Averaged CRMS from the time-averaged structure during the last 400 picoseconds of the simulation *vs*. residue compared with experimental temperature factors for (a) wild-type, (b) 56C8 and (c) 8G8. All mutations in the two variants were labeled in (b) and (c). Figure 3 d shows the difference in CRMS from the time-averaged structure between wild-type and 56C8 (cyan) and between wild-type and 8G8 (red).

Differences in flexibility

The CRMS from (a) the crystal structure and (b) the time-averaged structure for wild-type and both mutants as a function of time are shown in Figures 2 a and b, respectively. These CRMS values are averaged over all the residues. The time-averaged structure is based on coordinates constructed from the average of all snapshots from MD simulations from 100 ps to 500 ps at interval of 1 ps. The time-averaged structure represents the average solution structure for a given protein. Relative to wild-type, the time-averaged structures of the thermostable mutants (particularly the most stable mutant

8G8) are closer to the crystal structures. However, the thermostable mutants actually show increased fluctuations about their time-averaged structures relative to wild-type.

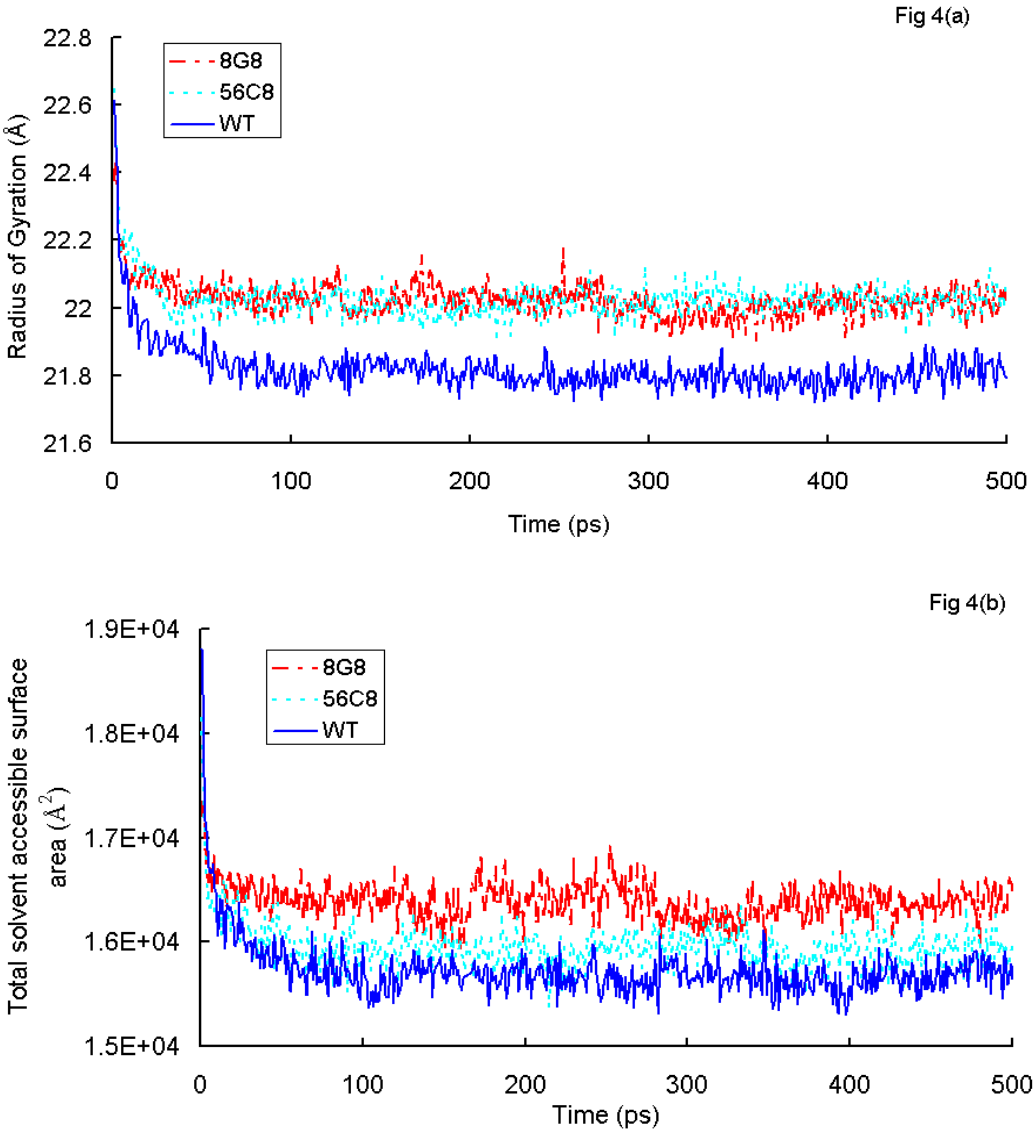


Figure 4. (a) Radius of gyration calculated using Equation 2 and (b) Total solvent accessible surface areas for wild-type (blue), 56C8 (cyan) and 8G8 (red) as functions of time.

Figure 4a shows the calculated radius of gyration for each of the esterases as a function of time. Both 56C8 and 8G8 show a small but statistically significant increase (about 1%) in radius of gyration relative to wild-type. This indicates a more expanded average structure, possibly as a result of larger or more frequent conformational "breathing" motions. Figure 4b shows the total solvent accessible surface area as a function of time for each of the three esterases. Here, 8G8 shows a small but significant increase (about 5%) relative to wild-type. This again indicates a more "open" average structure for 8G8. In contrast, 56C8 shows only a very small difference from wild-type.

Density of states

For each of the three esterases, the vibrational density of states was calculated from the velocity autocorrelation function generated from the simulations. While the densities of states for the three structures are quite similar, there are discernible differences. Figure 5a shows the power spectra (vibrational density of states) of wild-type, 56C8 and 8G8. The overall shape of the spectra are similar to the power spectra of other proteins, including those calculated from MD simulations (as was done in this work) and from normal mode analysis[21]. It is also similar to neutron scattering spectra obtained for globular proteins [21]. This demonstrates that our sampling rate is sufficient to capture the salient features of the dynamics, including the peak at ~3000 cm^{-1} which corresponds to hydrogen vibrations. Significant differences can be discerned in the DOS profiles at the lower wavenumbers. Figure 5b shows an expanded view of the DOS profiles from 2.0 to 25 cm^{-1}. Relative to wild-type, thermostable variants have substantially increased their populations of the lowest frequency modes (below 10 cm^{-1}).

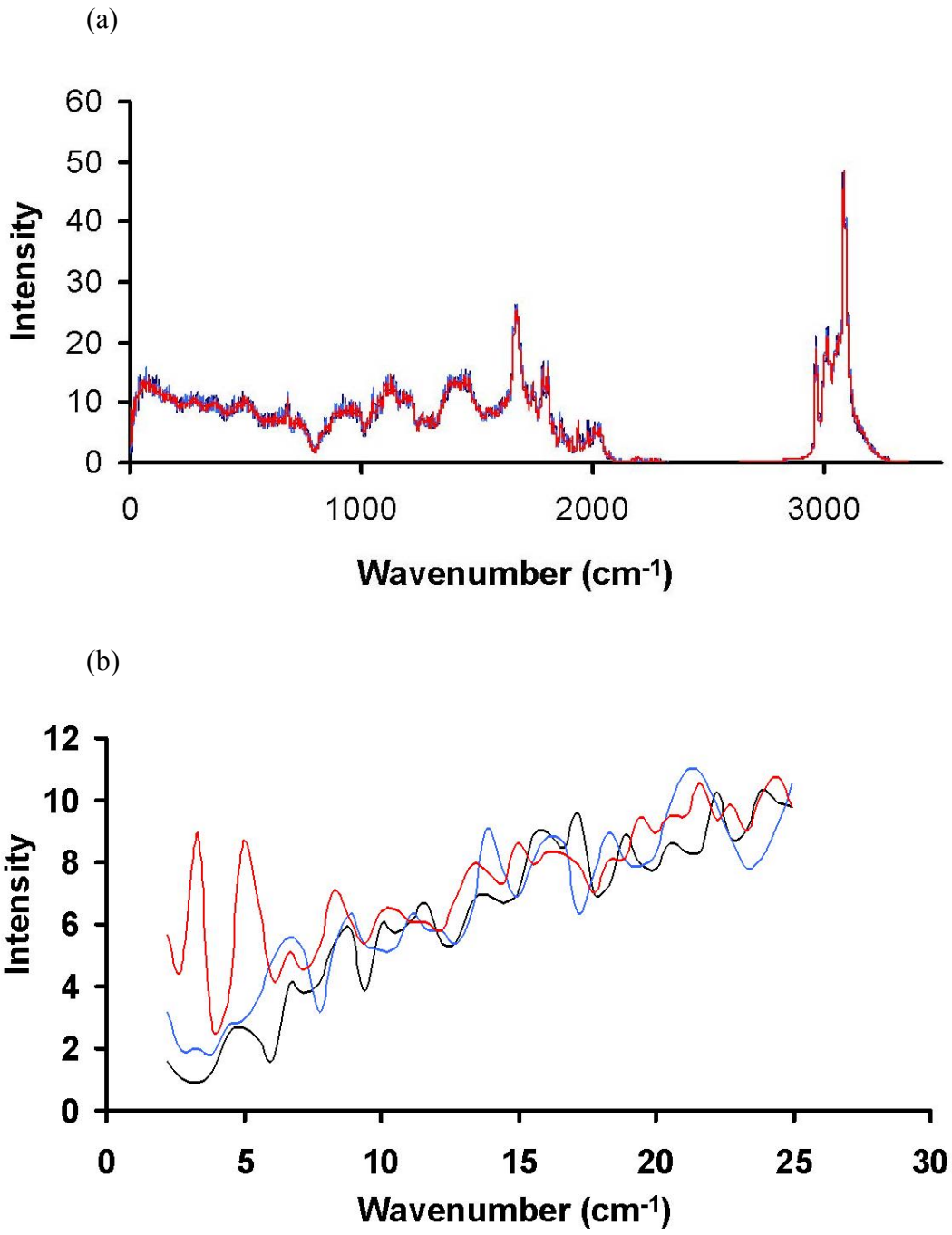


Figure 5. (a) Vibrational density of states (DOS) calculated using Equation 1 for wild-type (black line), 56C8 (blue line), and 8G8 (red line). (b) Expanded view of the DOS from 2.0 to 25 cm^{-1}. Intensity of DOS has the unit of number of modes per cm^{-1}.

Theoretical studies [22] have established that the modes in this region correspond to collective motions, in which many atoms in a given protein region move in a concerted manner. The thermostable esterases thus appear to undergo concerted motions more often than wild-type, in contrast to what would be expected if higher stability were always accompanied by reduced mobility. The degree to which these low frequency modes have been increased varies among the mutants, with 8G8 showing a larger increase than 56C8. The rank order of the degree of increase in the low frequency modes thus matches the rank order of thermostability for the 8G8 and 56C8 variants, with the more stabilized mutant showing larger increase. That 56C8 and 8G8 experience concerted motions more frequently than wild-type is very consistent with the observation that both of these mutants have a larger radius of gyration than wild-type.

Localized changes in mobility

As thermostability is increased, both the N and C terminal domains of pNBE show significant reductions in CRMS from the crystal structure. At the same time, CRMS from the time-averaged structure is found to increase in many regions. Figure 3d shows the difference in CRMS fluctuations about the time-averaged structure between wild-type and the two mutants calculated for each residue. This is calculated by subtracting the CRMS given for wild-type in Figure 3 a from the CRMS given for each of the two mutants in Figures 3 b and c. Some important mutations are labeled in the plot. It is seen that 8G8 shows larger fluctuations about the time-averaged structure than wild-type, especially for those residues around the mutations. Also these fluctuations are over longer stretches of consecutive residues in 8G8 than wild-type. The most dramatic changes

between the thermophilic mutant 8G8 and wild-type are seen in the regions comprised of residues 240-290, 312-360, and 410-420 (Figure 3 d). Coincidentally, there are several mutations in this region, most notably L313F and H322Y, which suggests that these mutations confer more flexibility in thermophilic 8G8. Residues 310-320 form a part of the active site, and Glu310 itself is part of the catalytic triad. For all three residues that make up the catalytic triad—Ser189, Glu310, and His399—the CRMS from the average structure are very close for wild-type and 8G8 (Table 1). This indicates that the dynamic properties of the critical active site residues have been conserved in spite of significant changes in other regions. 56C8 shows larger differences from wild-type at the active site. This may be a reflection of the fact that, experimentally, 56C8 is less catalytically active than wild-type in aqueous solution (56C8 was originally evolved for activity for catalyzing hydrolysis of *para*-nitrobenzyl ester in aqueous organic solvents, see [17] for details). A region that shows negative differences in CRMS between wild-type and the mutants is comprised of residues 414-420. This is interesting because in most other regions 8G8 shows large CRMS than wild-type. These residues form a short turn of α-helix in 56C8 and 8G8. In wild-type, however, they do not appear in the electron density, indicating that this regions is highly dynamic or disordered [19]. For purposes of the simulation, these residues were built into the wild-type structure based on the coordinates of 8G8. The fact that this region has high mobility during our simulation of wild-type is consistent with the observation that this region does not appear in the electron density of the wild-type crystal structure, and further confirms that our simulations have correctly described the structural dynamics of the system.

Table 1. Fluctuations of the catalytic triad (Å) during the last 400 ps of MD simulation

Residue	Wild-type	56C8	8G8
Ser 189	0.35	0.32	0.34
His 399	0.39	0.60	0.42
Glu 310	0.32	0.44	0.39

It is of interest to identify the nature of the low frequency modes of motion whose populations are increased in the thermophilic mutants relative to wild-type (see the discussion of the vibrational density of states above). Aligning structures taken at various times during the simulation allows us to identify those regions that show the largest differences in mobility between wild-type and the mutants. The aligned snapshots of 8G8 show larger average RMS deviations than those of wild-type; consistent with the fact that 8G8 shows larger fluctuations about its time-averaged structure. From the alignment, the largest displacements are located in the regions comprised of residues ~240-295 and ~310-360 (Figure 6 a, b). These regions contain several helices (residues 252-266, 287-294, 326-333, 337-345, and 350-362) whose positions fluctuate up to 6Å in 8G8, while they remain essentially superimposable in wild-type. It is interesting to note that these helices shift their positions essentially as rigid bodies, with little evidence of unraveling or deformation. Such motions involve a large number of atoms moving in a concerted manner, and are thus good candidates for the low frequency motions that are increased in the thermostable mutants. Further, the surface loop comprised of residues ~315-323 undergoes fluctuations of up to 6Å in 8G8, while it shows little movement in wild-type.

These motions each involve up to 14 residues moving in a concerted manner, and may thus contribute to the increased density of states seen at wavenumbers below ~10 cm^{-1}. We note that in this same region the wild-type structure exhibits large deviations from the crystal structure (Figure 7 a, b). In particular, the helices from residues 326-333 and 337-345 in wild-type shift significantly from their initial positions, while they remain close in 8G8.

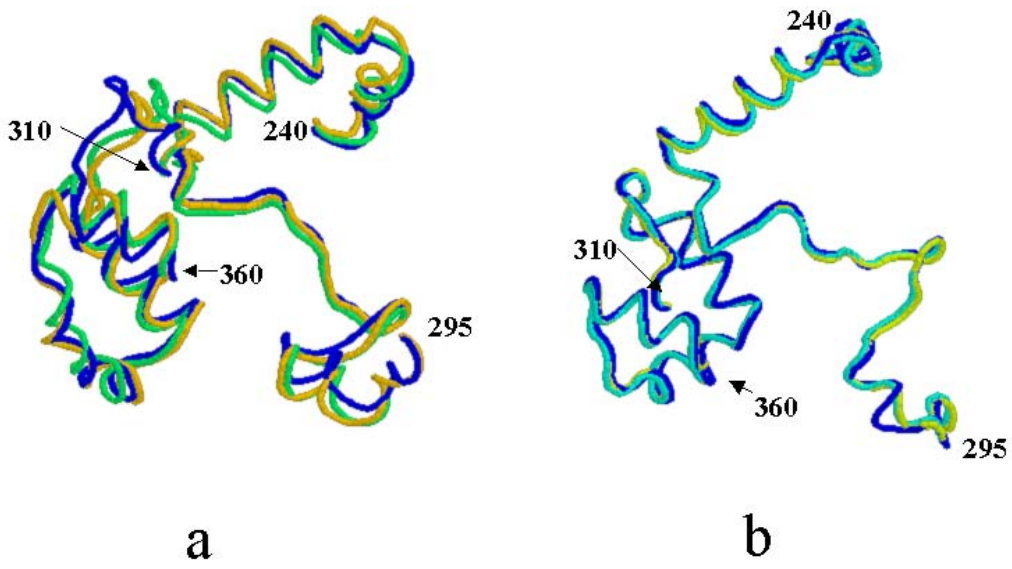


Figure 6. Superimposed snapshots of (a) 8G8 and (b) wild-type taken at 100 (blue), 300 (green), and 500 ps (yellow). Residues 240-300 and 310-360 are shown.

The largest difference in RMSD about the time-averaged structure between wild-type and 8G8 is seen in the loop comprised of residues 315-323 (Figure 3d). This region is flanked by two stabilizing mutations, L313F and H322Y. L313F forms an edge-face interaction with Phe 314, while H322Y forms interactions with a number of residues

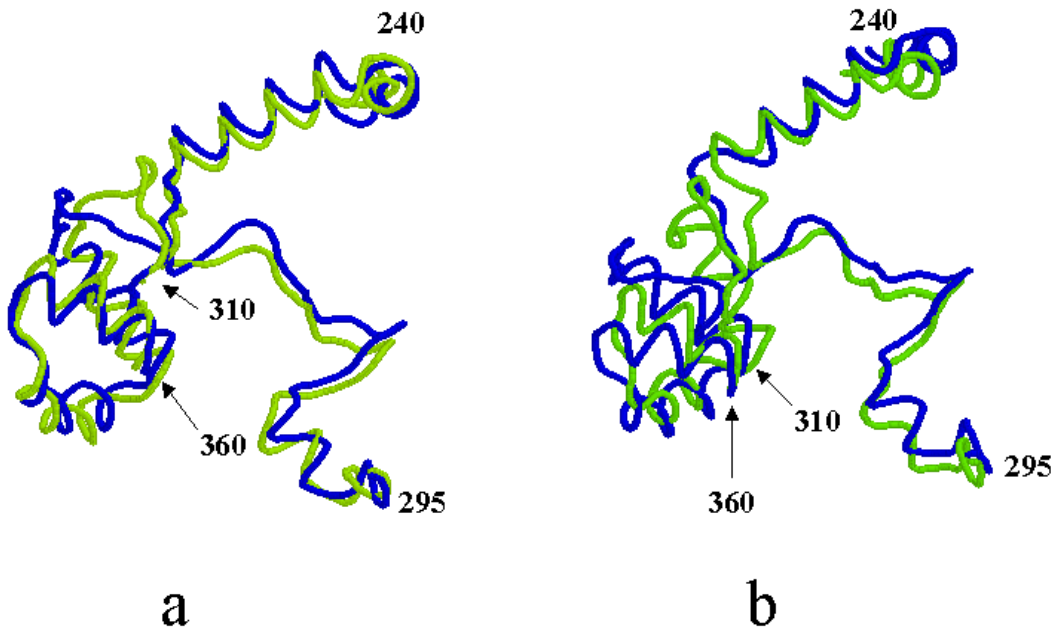


Figure 7. Superimposed energy minimized crystal structure (blue) and structure at 300 ps (green) for (a) 8G8 and (b) wild-type.

including Ile 270 and Val 358 (resulting from the stabilizing mutation M358V). These contacts are maintained during the simulation, and appear to prevent the 315-323 loop from shifting substantially away from its position in the crystal structure (as occurs in wild-type). While the two ends of the loop are thus fixed in place, the region between them experiences larger fluctuations during the simulation than in wild-type. The RMSD from the time-averaged structure for 8G8 (Figure 3 c) reveals several "spikes", i.e., stretches of residues in which the RMSD is markedly greater than in the surrounding regions. Examining the distribution of mutations in Figure 3 c, it can be seen that mutations are found in the "valleys" immediately adjacent to these areas. These observations suggest that the stabilizing mutations in 8G8 act as local anchors, locking down specific regions and preventing them from deviating far from their conformation in

the crystal structure. Regions adjacent to these anchor points experience increased fluctuations, though these fluctuations occur about a mean conformation that is closer to the crystal structure than would be the case if the stabilizing mutations were absent. The case of the 215-223 loop flanked by mutations L313F and H322Y illustrate this particularly well, and provides a clearer physical picture of what is meant by the statement that the thermophilic mutants remain closer to their crystal structures while exhibiting increased fluctuations about their time-averaged structures.

Mobility of Trp 102

The primary mechanism for non-radiative decay in phosphorescing Trp residues is vibrational coupling between the triplet state and the ground state due to out-of-plane distortions of the aromatic ring [23; 24]. Thus, longer phosphorescence lifetimes indicate reduced local fluctuations. Tryptophan phosphorescence lifetimes have been used to probe local mobility/rigidity in proteins [25]. The phosphorescence lifetimes of Trp102 were measured previously for wild-type and several thermostable mutants, including 8G8 [18]. Lifetimes were not experimentally measured for 56C8, but were determined for the closely related mutant 1A5D1 (56C8 has all of the stabilizing mutations present in 1A5D1 plus two additional mutations, L334S and P317S, which do not lie near Trp102). It was found that all mutants show increased phosphorescence lifetimes relative to wild-type, with 8G8 showing the largest increase (1.9 times that of wild-type). The lifetime measurements for Trp102 provide an ideal means for assessing the quality of our simulations, since the motions responsible for phosphorescence decay are believed to occur on the picosecond timescale. Figure 8 a shows the all-atom CRMS from the time-

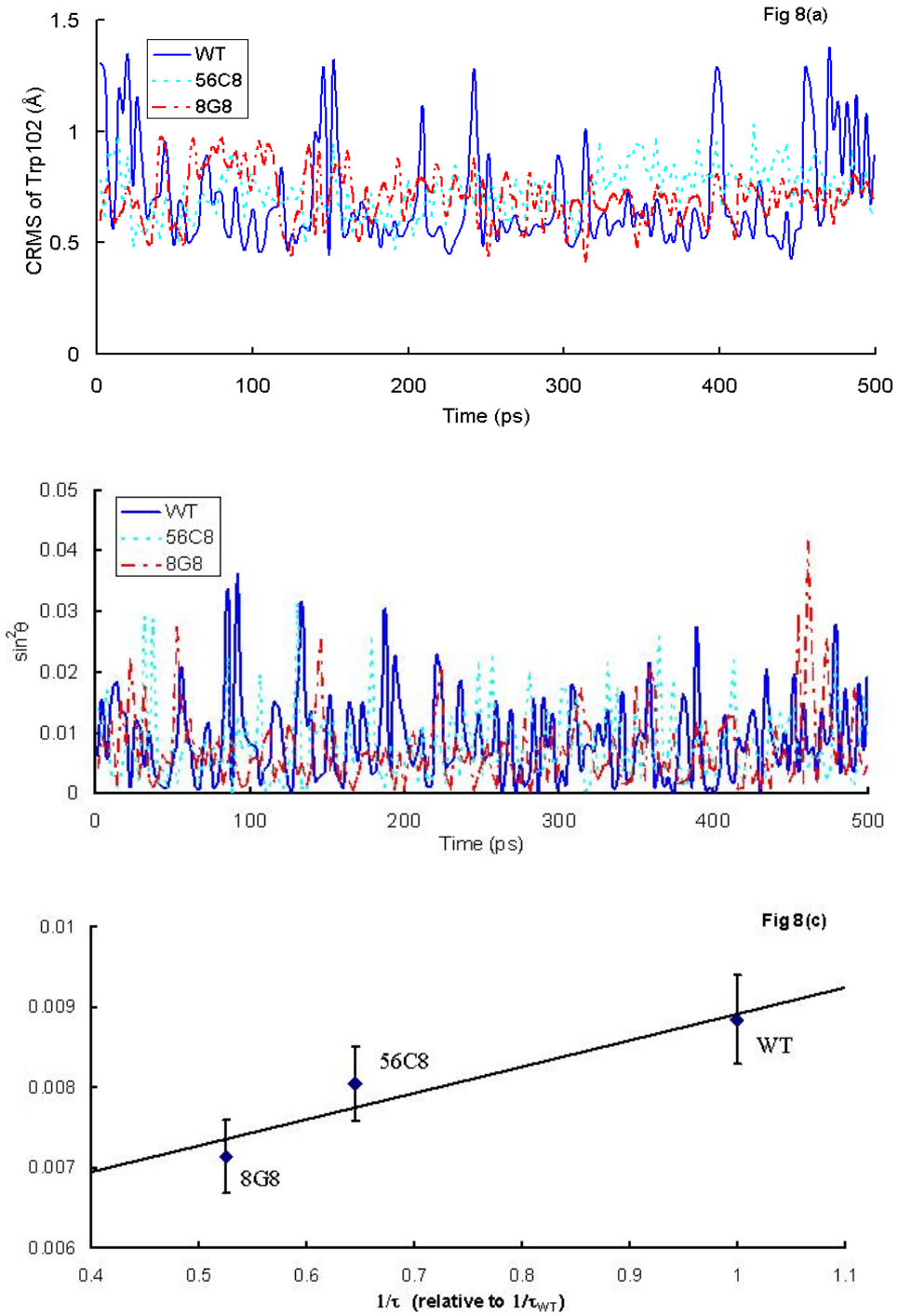


Figure 8. (a) All-atom CRMS from the time-averaged dynamic structure for Trp102 as a function of time. (b) Out-of-plane bending motions (see text) for Trp102 as a function of time. (c) Time-averaged $\sin^2\theta$ versus the inverse of phosphorescence lifetime.

averaged structure of Trp102 for wild-type, 56C8, and 8G8. The average CRMS in all three enzymes is ~ 0.5Å. However, the frequency and magnitude of larger deviations is clearly greater for wild-type than for 8G8. Figure 8 b shows the out-of-plane bending motion angle for Trp102 as a function of time. From Figure 8a and 8b it is clear that Trp102 in WT shows larger instantaneous fluctuations in CRMS from the crystal structure and also higher values for $\sin^2\theta$ in the out-of-plane bending motion than either 56C8 or 8G8. If the time-averaged $\sin^2\theta$ value is plotted versus the inverse phosphorescence lifetime, the data can be fit into a straight line (Figure 8 c). This is consistent with the theory that phosphorescence lifetime inversely correlates with $\sin^2\theta$ [23]. These motions occur at the picosecond timescale and thus are in accordance with the experimental findings. The degree of larger deviations for 56C8 falls between wild-type and 8G8. Thus the degree of Trp102 motion on the picosecond timescale ranks as wild-type > 56C8 > 8G8. This is in excellent correspondence with what is expected from the observed phosphorescence lifetimes. As expected, an inverse correlation is seen between the degree of Trp102 motions during the simulation and the measured phosphorescence lifetimes.

Discussions

The most striking result from this study is the apparent discrepancy between different measures of flexibility. Based the CRMS from the crystal structure it appears that the thermophilic mutant 8G8 is indeed more rigid than its mesophilic parent. However, the CRMS from the time-averaged structure, the radius of gyration, and the

population of low frequency modes all indicate that 8G8 is in fact *more mobile* than wild-type. These observations are not necessarily contradictory. Proteins experience a wide array of motions spanning a vast range of amplitudes and timescales. There is no reason to expect that all of these different modes of motion will change in the same way as a given protein evolves to become more stable.

Our results indicate a need to refine the term "flexibility" as it is currently used in discussions of temperature adaptation in proteins. The timescales, amplitudes, and locations of the motions presumed to contribute to flexibility must be specified. Clearly certain modes of motion, particularly those that may initiate unfolding, must be reduced if a protein is to achieve increased stability at elevated temperatures. In the case of *p*NBE, this is perhaps being reflected in the reduced deviations of surface loops from their positions in the crystal structure.

For many other types of motions, however, there is little evidence that they are detrimental to global stability. In fact, there is increasing evidence that various modes of motion contribute to the stability of the native state. In general, work has focused on small amplitude local fluctuations. Recent neutron scattering studies of a mesophilic and thermophilic α-amylase found that the thermostable enzyme displayed increased mobility relative to the mesophilic one on the 0.3 to 6.0 picosecond timescale [8; 9]. These same studies, however, found no difference in mobility between the two proteins in the unfolded state, suggesting that increased conformational entropy in the native state may contribute to the stability of the thermophilic enzyme. Recent NMR relaxation studies of

bond vector motion in proteins have led to estimates of the contribution of native-state conformational entropy to stability [26]. These studies conclude that local internal protein motions may increase stability by increasing the entropy of the native state. Such motions can also raise the melting temperature of a protein by increasing the heat capacity of the native state and thus decreasing the heat capacity difference between the native and denatured states (ΔC_p^{unfold}) [27]. Such a decrease in ΔC_p will expand the temperature range in which a protein remains folded by both raising the melting temperature and decreasing the cold-denaturation temperature. The idea that protein motions contribute to stability is supported by the opposing trends we observed for the CRMS deviations from the esterase crystal structures and the CRMS fluctuations about the time-averaged structures during the simulations. Relative to wild-type, the thermophilic esterase 8G8 maintains its average structure closer to the crystal structure even while it experiences greater fluctuations about this average. This idea is also supported by the fact that the thermophilic variants show increases in the population of the low frequency modes. Further, calculation of thermodynamic parameters from the velocity autocorrelation function does indicate that thermostabilization is accompanied by increases in the absolute native state entropy and heat capacity. Due to the limited simulation time it is unlikely that these calculated values would agree exactly with measured values (excepting the heat capacity, these absolute values are not practically amenable to experimental measurement in any case). However, they do indicate that, on this timescale, the thermostable mutants explore more degrees of freedom than wild-type.

The dynamics of folded proteins are often described in terms of a high dimensional energy landscape consisting of many conformational substates (local energy minima) separated by energy barriers [28]. In this picture, transitions may be expected to occur frequently between conformations of similar energy separated by low barriers. In contrast, transitions between conformations separated by large energy barriers will be observed only rarely. Our results indicate that different aspects of the energy landscape can respond to the demands of high temperature adaptation in different ways. On a global scale, the fact that 56C8 and 8G8 are more stable than wild-type indicates that the distance between the native state ensemble and the denatured state ensemble has increased. Both mutants remain closer to their minimum energy structures during the simulation. An equivalent statement is that conformational substates, which depart substantially from the minimum energy structure, are being sampled less frequently. This suggests that the energetic barriers separating such substates from the minimum energy state have been increased. At the same time, the increased fluctuations about the time-averaged structure indicate that the barriers between some conformational substates have been reduced.

In addition to these differences in overall mobility, it is also clear that different regions of the protein have manifested different changes in mobility in response to temperature adaptation. In general, surface loops show the largest differences in mobility (as measured by both CRMS from the crystal structure and the time-averaged structure) in going from wild-type to 8G8. Large displacement of surface loops from their native conformation has been identified in other studies as an important factor in thermostability

[29; 30]. Such large loop displacements may expose the hydrophobic core of a protein to water penetration, leading to unfolding. It is not unexpected, therefore, to find that alterations in loop mobility accompany the thermostabilization of *p*NBE. At the same time other regions of the protein display much smaller changes in mobility as a result of temperature adaptation. In particular, the RMS fluctuations of the three catalytic residues are remarkably conserved between wild-type and 8G8 (Table 1). While fluctuations about the time-averaged structure increased in many regions of the thermostabilized pNBE variants, they decreased in the immediate vicinity of Trp 102, consistent with the experimental observation that its phosphorescence lifetime is longer in more stable mutants.

It is useful to compare our results with two recent simulation studies of thermophilic and mesophilic proteins. Colombo and Merz [30] performed extensive simulations of wild-type subtilisin E and a homology model of its thermophilic counterpart 5-3H5, which was generated by directed evolution. They also found a difference in the flexibilities as measured by CRMS from the crystal structure and by CRMS from the time-averaged structure. While the thermostable enzyme showed smaller deviations from the crystal structure, it showed larger fluctuations about the time-averaged structure. The simulations reported here allow for comparisons with experimental data in the form of crystallographic temperature factors and tryptophan phosphorescence lifetimes. The observed correlation of our results with the available experimental data increases our confidence that the increased flexibility seen in the thermostable mutants is real and not simply an artifact of the simulation.

In another study, Lazaridis *et al.* [7] performed simulations on mesophilic and hyperthermophilic rubredoxins. In contrast to the results of this study and that of Colombo and Merz, Lazaridis *et al.* found that, by all measures, the hyperthermophile showed slightly reduced mobility relative to the mesophile at room temperature. The hyperthermophilic rubredoxin was natural rather than laboratory evolved, and the results of Lazaridis *et al.* may indicate that natural thermophilic enzymes are altered in their conformational mobility in fundamentally different ways than laboratory evolved thermophilic enzymes (the experimental results from the α-amylases argue against this, however). It is also possible that the low mobility of the hyperthermophilic rubredoxin is the result of functional considerations. In contrast to enzymes, which are thought to require a considerable degree of mobility in order to catalyze reactions efficiently, the rubredoxins are thought to undergo only very minor conformational changes during oxidation/reduction [31].

Given the variety of protein topologies and functions, it is not surprising to find that different proteins have responded to the challenge of high temperature adaptation in different ways. The family of laboratory-evolved pNB esterases has provided us with a unique opportunity to examine how protein dynamics change in response to adaptive evolution without the complications and ambiguities introduced by neutral mutations and unknown selective pressures. Our study has found that adaptation to high temperature results in a rich variety of alterations in dynamic behavior, including reductions in certain types of motions and increases in others. This may help explain the apparently

contradictory experimental results that have been reported in studies of thermophilic enzymes. Some techniques, such as H/D exchange of the slowly exchanging hydrogens, will be sensitive primarily to motions that involve significant departures from the native state. We have seen a reduction of such displacements from the minimum energy structure in our simulation of the thermophilic esterases. Other techniques, such as neutron scattering and nuclear magnetic relaxation, can detect small amplitude local fluctuations, which we have seen increased in the thermophilic esterases. Our study suggests that the apparently contradictory results reported from these diverse experimental techniques might reflect different aspects of dynamic adaptation to high temperatures.

Materials and Methods

The three crystal structures WT, 8G8 and 56C8 were used as the starting structure for all the molecular dynamics (MD) simulations (PDB entries 1QE3, 1C7J and 1C7I). All the hydrogen atoms were added explicitly using Polygraf, and counterions Na^+ and Cl^- were added to neutralize the side chains of Asp, Glu, Arg and Lys [32]. These counterions are allowed full freedom to move in the dynamics, but stay close to the original positions.

We used Dreiding forcefield [33] with the charges from CHARMM22. [34] The nonbond (Coulomb and van der Waals) interactions were calculated using the Cell Multipole Method [35] (CMM) for fast and accurate calculations of nonbond interactions. CMM scales linearly with the number of atoms, so that no cutoffs are required.

Since inclusion of explicit water in the MD simulations is computationally intensive, we have used the Surface Generalized Born (SGB) continuum solvation model [20] to calculate both the energy *and the forces* due to the solvent acting on the protein structure. SGB accounts for the response in the protein conformation due to electrostatic interactions with the solvent, which is assumed to extend beyond the solvent accessible surface of the protein. The only properties of solvent required are the dielectric constant (78.32) and the solvent radius (1.4 Å). We assume an ionic strength of 0.1. We assumed that the internal dielectric constant of the protein is 2.0. The SGB method is an approximation to the Poisson-Boltzmann continuum solvation model [36] for calculating energies and forces but is much faster. Of course both are faster than explicitly including the water solvent.

The total potential energy of each of the three structures was minimized using conjugate gradients. The minimization was performed for 1000 steps with a termination criterion that the RMS force is less than 0.1 kcal/mol/Å.

Constant temperature Hoover MD simulations [37] were carried out on all three structures for 500 ps. The temperature of the simulations was set to 300 K. The parallel MPSIM MD program [38; 39] was used for all simulations.

Calculation of RMS difference in coordinates:

As a simple measure of similarity between two structures we calculate the root mean square of the differences in coordinates between all corresponding atoms in the two structures. First an optimal translation and rotation is performed to superpose the center of mass and the moments of inertia and then the coordinates of equivalent positions are compared. This is referred to as the CRMS (coordinates root mean square) difference. In calculating the CRMS (and radius of gyration) we do *not* include the counterions.

We first validated our forcefield by calculating the CRMS differences between the experimental crystal structure and the energy-minimized structure. Generally a CRMS lower than 0.6 A is considered to indicates that the forcefield is sufficiently accuracy.

We also report CRMS differences between the snapshots of the MD structures at various time intervals and the starting minimized energy structure. The CRMS with respect to the minimized structure shows the temperature factors for various regions of the protein indicating which parts are more flexible than others (in solution).

The average MD structure was calculated by averaging the coordinates of the various MD snapshots from 100 ps to 500 ps at 1 ps time intervals. This average structure represents the structure of the protein equilibrated in salt and solvent.

Calculation of density of states:

The vibrational density of states (power spectrum) was calculated from the Fourier transform of the velocity auto-correlation function:

$$S(\nu) = 2\beta \sum_{j=1}^{3N} m_j \widetilde{C}_{vv}(\nu), \qquad (1)$$

where $\widetilde{C}_{vv}(\nu)$ is the Fourier transform of the velocity auto-correlation function $C_{vv}(t)$, m_j is the mass of atom j, $\beta = 1/(k_B T)$ and k_B is Boltzmann's constant, T is the absolute temperature of the system. Sampling of the velocities was done as low as every 1 fs to obtain the high frequency and less often to obtain the low frequency modes.

Calculation of solvent accessible surface area:

The solvent accessible surface area is calculated using Connolly's molecular surface calculation program in which a probe is rolled along the surface of the protein [40]. The probe size used was 1.4 Å.

Calculation of radius of gyration:

The radius of gyration is calculated using the following definition:

$$R_G = \left(\frac{\sum m_i r_i^2}{\sum m_i} \right)^{1/2}, \qquad (2)$$

in which r_i is the distance of the atom i from the center of mass of the protein molecule, and m_i is the mass of the i^{th} atom.

Calculation of Trp out-of-plane bending modes:

In the absence of oxygen, the phosphorescence lifetime of a buried Trp is primarily determined by the out-of-plane motion. In the present calculations we have defined out-of-plane bending as involving the nitrogen atom in the indole ring [23; 24]. The out-of-plane bending is measured as the angle, θ, between the p orbital of the N atom and the normal of the six-member ring. The phosphorescence lifetime of the indole ring decreases with increasing $\sin^2\theta$ value [23].

Acknowledgements:

We thank Dr. Anne Gershenson (Brandeis University) for her insight and helpful discussions. This work was supported in part by the Army Research Office and the Center for Science and Engineering of Materials at Caltech (NSF-MRSEC). The facilities of the Materials and Process Simulation Center used in this project are supported also by DOE (ASCI ASAP), NSF (CTS and MRI), NIH, ARO-MURI, Chevron Corp., MMM, Seiko-Epson, Dow Chemical, Avery-Dennison Corp., Kellogg's, General Motors, Asahi Kasei, the Beckman Institute, and ONR.

References

1. Jaenicke, R. & Bohm, G. (1998). The stability of proteins in extreme environments. *Curr. Opin. Struct. Biol.* 8, 738-748.
2. Fields, P. A. (2001). Protein function at thermal extremes: balancing stability and flexibility. *Comp. Biochem. Phys. A* 129, 417-431.
3. Varley, P. G. & Pain, R. H. (1991). Relation between stability, dynamics and enzyme activity in 3-phosphoglycerate kinases from yeast and Thermus thermophilus. *J. Mol. Biol.* 220, 531-8.
4. Zavodszky, P., Kardos, J., Svingor & Petsko, G. A. (1998). Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc. Natl. Acad. Sci. USA* 95, 7406-11.
5. Hollien, J. & Marqusee, S. (1999). Structural distribution of thermostability in a thermophilic enzyme. *Proc. Natl. Acad. Sci. USA* 96, 13674-13678.
6. Hernandez, G., Jenney, F. E., Adams, M. W. & LeMaster, D. M. (2000). Millisecond time scale conformational flexibility in a hyperthermophilic protein at ambient temperature. *Proc. Natl. Acad. Sci. USA* 97, 3166-3179.
7. Lazaridis, T., Lee, I. & Karplus, M. (1997). Dynamics and unfolding pathways of a hyperthermophilic and a mesophilic rubredoxin. *Protein Sci.* 6, 2589-2605.
8. Fitter, J. & Heberle, J. (2000). Structural equilibrium fluctuations in mesophilic and thermophilic alpha-amylase. *Biophys. J.* 79, 1629-1636.
9. Fitter, J., Herrman, R., Haub, T., Lechner, R. E. & Dencher, N. A. (2001). Dynamical properties of alpha-amylase in the folded and unfolded state: the role of thermal equilibrium fluctuations for the conformational entropy and protein stabilization. *Physica B* 301, 1-7.
10. Englander, S. W. & Kallenbach, N. R. (1983). Hydrogen-exchange and structural dynamics of proteins and nucleic acids. *Quarterly Rev. Biophys.* 16, 521-655.
11. Hilser, V. J. & Freire, E. (1996). Structure based calculation of the equilibrium folding pathway of proteins: correlation with hydrogen exchange protection factors. *J. Mol. Biol.* 262, 756-772.
12. Tang, K. E. & Dill, K. A. (1998). Native protein fluctuations: the conformational-motion temperature and the inverse correlation of protein flexibility with protein stability. *J. Biomol. Struct. Dyn.* 16, 397-411.
13. Adams, M. & Kelly, R. (1998). Finding and using hyperthermophilic enzymes. *Trands Biotech.* 16, 329-332.
14. Kimura, M. (1983). *The neutral theory of molecular evolution*, Cambridge University Press, Cambridge.
15. Wintrode, P. L. & Arnold, F. H. (2000). Temperature adaptation of enzymes: lessons from laboratory evolution. *Adv. Protein Chem.* 55, 161-225.
16. Wintrode, P. L., Miyazaki, K. & Arnold, F. H. (2000). Cold adaptation of a mesophilic subtilisin-like protease by laboratory evolution. *J. Biol. Chem.* 275, 31635-40.
17. Giver, L., Gershenson, A., Freskgard, P. O. & Arnold, F. H. (1998). Directed evolution of a thermostable esterase. *Proc. Natl. Acad. Sci. USA* 95, 12809-12813.
18. Gershenson, A., Schauerte, J. A., Giver, L. & Arnold, F. H. (2000). Tryptophan phosphorescence study of enzyme flexibility and unfolding in laboratory-evolved thermostable esterases. *Biochemistry* 39, 4658-4665.
19. Spiller, B., Gershenson, A., Arnold, F. H. & Stevens, R. C. (1999). A structural view of evolutionary divergence. *Proc. Natl. Acad. Sci. USA* 96, 12305-10.
20. Ghosh, A., Rapp, C. S. & Friesner, R. A. (1998). Generalized Born model based on a surface integral formulation. *J. Phys. Chem. B* 102, 10983-10990.
21. Hinsen, K. & Kneller, G. R. (2000). Projection methods for the analysis of complex motions in macromolecules. *Molecular Simulations* 23, 275-292.
22. Levitt, M., Sander, C. & Stern, P. S. (1985). Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.* 181, 423-447.
23. McGlynn, S. P. (1969). *Molecular Spectroscopy of the Triplet State*, Prentice-Hall, Englewood Cliffs, NJ.
24. Lower, S. K. & El-Sayed, M. A. (1966). The triplet state and molecular electronic processes in organic molecules. *Chem. Rev.* 66, 199-241.

25. Schauerte, J. A., Steel, D. G. & FGafni, A. (1997). Time-resolved room temperature tryptophan phosphorescence in proteins. *Method Enzymol.* 278, 49-71.
26. Stone, M. J. (2001). NMR relaxation studies of the role of conformational entropy in protein stability and ligand binding. *Acc. Chem. Res.* 34, 379-88.
27. Seewald, M. J., Pichumani, K., Stowell, C., Tibbals, B. V., Regan, L. & Stone, M. J. (2000). The role of backbone conformational heat capacity in protein stability: temperature dependent dynamics of the B1 domain of Streptococcal protein G. *Protein Sci.* 9, 1177-93.
28. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. (1991). The energy landscapes and motions of proteins. *Science* 254, 1598-1603.
29. Caflisch, A. & Karplus, M. (1994). Molecular dynamics simulation of protein denaturation: solvation of the hydrophobic cores and secondary structure of barnase. *Proc. Natl. Acad. Sci. USA* 91, 1746-1750.
30. Colombo, G. & Merz, K. M. (1999). Stability and activity of mesophilic subtilisin E and its mesophilic homolog: insights from molecular dynamics simulations. *J. Am. Chem. Soc.* 121, 6895-6903.
31. Yelle, R. B., Park, N. S. & Ichiye, T. (1995). Molecular dynamics simulations of rubredoxin from Clostridium pasteurianum: changes in structure and electrostatic potential during redox reactions. *Proteins* 22, 154-67.
32. Vaidehi, N. & Goddard, W. A., III. (1997). The pentamer channel stiffening model for drug action on human rhinovirus HRV-1A. *Proc. Natl. Acad. Sci. USA* 94, 2466-71.
33. Mayo, S. L., Olafson, B. D. & Goddard, W. A., III. (1990). DREIDING - a generic force field for molecular simulations. *J. Phys. Chem.* 94, 8897-8909.
34. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* 102, 3586-3616.
35. Ding, H. Q., Karasawa, N. & Goddard, W. A., III. (1992). Atomic level simulations on a million particles: the cell multiple method for Coulomb and London nonbond interactions. *J. Chem. Phys.* 97, 4309-4315.
36. Tannor, D. J., Marten, B., Murphy, R., Friesner, R. A., Sitkoff, D., Nicholls, A., Ringnalda, M., Goddard, W. A., III & Honig, B. (1994). Accurate first principles calculation of molecular charge distributions and solvation energies from ab initio Quantum Mechanics and continuum dielectric theory. *J. Am. Chem. Soc.* 116, 11875-11882.
37. Hoover, W. G. (1984). Computer-simulation of many-body dynamics. *Physics Today* 37, 44-50.
38. Lim, K.-T., Brunett, S., Iotov, M., McClurg, R. B., Vaidehi, N., Dasgupta, S., Taylor, S. & Goddard, W. A., III. (1997). Molecular dynamics fot very large systems on massively parallel computers: the MPSim program. *J. Comput. Chem.* 18, 501-521.
39. Vaidehi, N. & Goddard, W. A., III. (2000). Hierarchical NEIMO simulations for large scale domain motions in phosphoglycerate kinase. *J. Phys. Chem. A* 104, 2375-83.
40. Connolly, M. L. (1983). Analytical molecular surface calculation. *J. Appl. Cryst.* 16, 548-558.

Appendix II

Guanidine Hydrochloride-Induced Unfolding of Rubredoxin from *Pyrococcus furiosus*: Evidence for "Unfolding" Intermediates*

* This appendix is based on a manuscript for Biochemistry coauthored with Michael W. W. Adams, and Sunney I. Chan

Abbreviations

RdPf, rubredoxin from *Pyrococcus furiosus*; GuHCl, guanidine hydrochloride; DTT, dithiothreitol; ANS, 1-anilino-naphthalene-8-sulfonate; CD, circular dichroism; F, folded state; U, unfolded state; I_1, the first intermediate state; I_2, the second intermediate state; K, equilibrium constant; f_F, fraction of folded; f_{I1}, fraction of the first intermediate; f_{I2}, fraction of the second intermediate; f_U, fraction of unfolded; z_i, the fractional change in the spectral property from the folded state to the i-th intermediate state; ΔG_D^{H2O}, global unfolding free energy; $\Delta G_{I,i}$, the free energy difference between the folded state and the i-th intermediate state; m_D, the dependence between free energy change and denaturant concentration; R, universal gas constant; T, absolute temperature.

Abstract

Rubredoxin from *Pyrococcus furiosus* (RdPf) is a hyperthermophilic protein that does not undergo thermal melting even at temperatures well above 100 ^{o}C at neutral pH. However, it is possible to denature the protein at pH 2 by disrupting all the salt bridges. We report here denaturant-induced unfolding of RdPf in the presence of guanidine hydrochloride at pH 2 as followed by UV-visible absorption, fluorescence and CD spectroscopic techniques. The results obtained suggest that at least two different intermediates are present in the equilibrium unfolding pathway. The two intermediates have maximum population at 2.4 and 3.3 M guanidine hydrochloride, respectively. The global unfolding free energy ($\Delta G_D^{H_2O}$) for RdPf at pH 2 is estimated to be 13.6 kcal/mol. However, if the disrupted salt bridges are taken into account, this value will be larger for RdPf at neutral pH, suggesting that RdPf is thermodynamically more stable than most other single-domain proteins. The factors contributing to the stability of RdPf are estimated, and the results suggest that the enhanced protein stability does not require strong stabilizing forces, merely tuning of the forces to accomplish a redistribution of thermally accessible conformational states.

Introduction

Hyperthermophiles are microorganisms that have optimal growth temperatures of at least 80 ^{o}C. Several of them have optimal temperatures over 100 ^{o}C, which is the highest temperature that life is known to exist to date. Most of the proteins isolated from such organisms exhibit correspondingly enhanced thermostability. This property is useful both for the investigation of fundamental biological questions of protein stability, and for the development of biotechnological applications that require protein stability at high temperatures. Examples of such applications are detergent manufacturing, production of high-fructose corn syrup, the polymerase chain reaction (PCR), etc. (*1*).

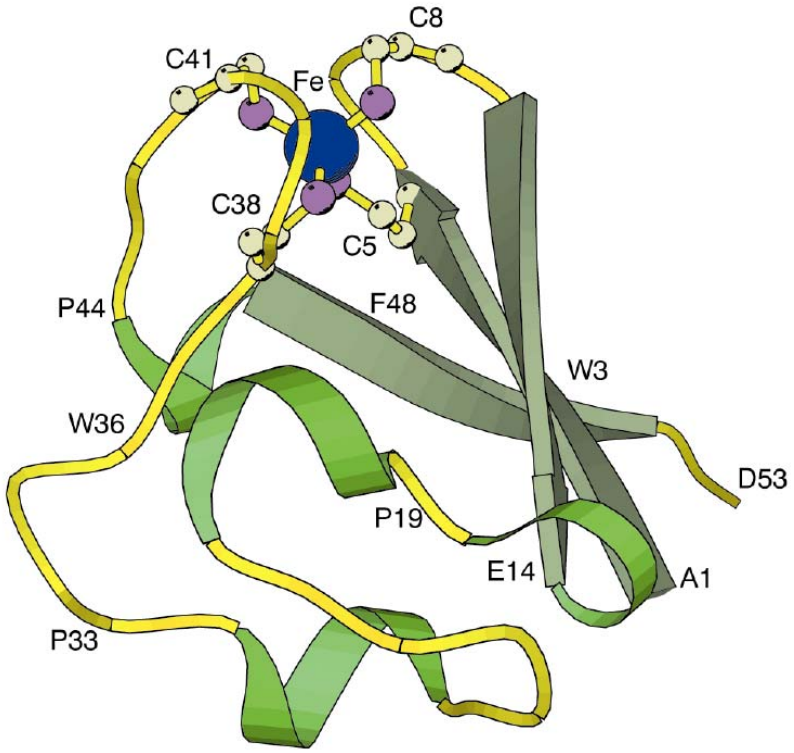
The origin of the thermostability of enzymes from hyperthermophiles has not been fully understood, despite that many primary sequences of hyperthermophilic proteins and their mesophilic counterparts are available for comparison, and high-resolution structures of hyperthermophilic proteins have become available for many proteins, e.g., rubredoxin (*2, 3*), GAPDH (*4*), aldehyde ferredoxin oxidoreductase (*5*), histone HMfB (*6*) and glutamate dehydrogenase (*7*). Surprisingly, the structures of homologous proteins from hyperthermophilic sources are strikingly very similar to their mesophilic counterparts. For example, the RMS difference between rubredoxin from the hyperthermophile *Pyrococcus furiosus* (RdPf) and the mesophile *Clostridium pasteurianum* (RdCp) is only 0.47 Å for the main chain (*3*), presumably because of the high sequence identity and very few additions or deletions of amino acid residues. It is not clear that hyperthermophilic proteins are intrinsically more stable thermodynamically. Preliminary investigations with RdPf have suggested that such proteins could be

kinetically trapped in some local energy minimum (Cavagnero et al., unpublished result). To resolve this issue, thermodynamic studies on hyperthermophilic proteins are greatly needed.

An ideal model system to address the issue of protein hyperthermostability is the rubredoxin from *Pyrococcus furiosus* (RdPf), a non-heme iron-protein with 53 amino acids. Its biological function is not known, although it is believed to take part in electron transfer processes in the cytoplasm (*8*). Both X-ray crystallographic and NMR solution structures of RdPf are available (*2, 3*). Its amino acid sequence and three-dimensional structure made from Molscript (*9*) are shown in Figure 1. Calorimetric studies on RdPf suggest that the protein has a melting temperature of 113 ^{o}C (*10*). Results of hydrogen-exchange studies on the native protein have been used to infer that the melting temperature of this protein could be as high as 170 ^{o}C (*11*). The role of salt bridges and the β-sheets in determining the hyperthermostability of RdPf has been evaluated (*12, 13*). Recently a RdPf variant without the Fe-$(Cys)_4$ center has been designed and used as a model to evaluate the contribution of surface salt bridges to protein stability (*14, 15*). The conformational flexibility of RdPf has been discovered to be in millisecond time scale using hydrogen exchange technique (*16*). In addition, several theoretical studies have appeared on the protein dynamics and thermal unfolding simulation of RdPf (*17-19*).

In the present study, we report the results of guanidine hydrochloride-induced RdPf unfolding experiments and the unfolding transition determined by UV-visible absorption, fluorescence and circular dichroism spectroscopic measurements. Evidence

is obtained for unfolding intermediates in the denaturant-induced unfolding process. The implications of these findings will be discussed.




```
1         5         10        15        20        25        30
A K W V C K I C G Y I Y D E D A G D P D N G I S P G T K F E

          35        40        45        50    53
E L P D D W V C P I C G A P K S E F E K L E D
```

Figure 1. Molscript representation of RdPf structure. The iron-4cys coordination is shown in ball-and-stick model. The amino acid sequence is shown below the structure.

Materials and Methods

Materials.

RdPf was isolated and purified from *Pyrococcus furiosus* as described previously(*20*). The purity of the protein was adjudged by the ratio of extinction coefficients between 280 nm and 380 nm. This ratio was 0.40 for the samples used in this study compared to 0.42 reported in literature(*20*). Protein concentrations were determined by measuring their absorption at 280 nm. The protein was stored in 50 mM Tris buffer (pH 8.0) with 0.3 M NaCl at –20 ^{o}C. It was concentrated and buffer-exchanged with a Microcon 3 device (Amicon) before use.

Ultra-pure guanidine hydrochloride (GuHCl) was purchased from Baker and used without further purification. Dithiothreitol (DTT, electrophoresis grade) was purchased from ICN Biomedical.

Spectroscopic measurements of the unfolding transition.

8 M GuHCl stock solution was prepared before the denaturation experiments. The stock solution was then diluted to different concentrations in attempts to denature the protein. 20 mM phosphate buffer was added to the protein solution to maintain the pH at either 2.0 or 7.0. A pHM93 pH meter from Radiometer Copenhagen with a general-purpose pHC2406 electrode was used to measure the pH values. The samples were incubated typically for 60 hours at room temperature to reach equilibrium.

UV-visible absorption spectra were recorded with a Hewlett-Packard 8452A single-beam diode array spectrophotometer. A 1 cm quartz cuvette was used in spectra acquisition. All the experiments were performed at room temperature. Buffers containing appropriate concentrations of GuHCl were used as the blank. Published values for the extinction coefficients of RdPf at the wavelengths of maximum absorption were used (*20*). The UV-visible absorption spectra were recorded and the absorbance at 380 nm was used to follow the unfolding process.

The tryptophan fluorescence was recorded on a Hitachi F-5000 Fluorescence Spectrophotometer. The excitation wavelength was set at 280 nm and the emission spectrum was recorded at the wavelength range from 290 nm to 450 nm. Since the buffer and GuHCl do not fluoresce at this region, no blank was subtracted. The spectra were plotted against wavenumber or the reciprocal of wavelength and then simulated by a three-peak Gaussian function centered at 310 nm, 335 nm and 355 nm, respectively. The fluorescence intensity at 353.8 nm was used to analyze the data, because the ratio of the fluorescence intensity of the unfolded protein to the folded protein shows a maximum value of about 5 at this wavelength.

Far-UV circular dichroism spectra were acquired with a JASCO 600 spectropolarimeter using a 0.1 cm path quartz cell from Hellma. RdPf concentrations were typically from 5 to 10 μM. The CD spectra were recorded over the range of 200-300 nm. The scan step was set at 0.2 nm and the scan speed was 50 nm per min. The sensitivity of the CD signal was 50 mdeg. Each spectrum was an average of three scans

to improve the signal-to-noise ratio. The ellipticity at 225 nm was used as a measure of the secondary structure (β-sheet). The buffer and GuHCl show no CD in this region.

Analysis of the unfolding curves.

Since more than one technique has been used to follow the unfolding process, we shall refer to the various physical or spectroscopic parameters as "*y*" in the present generic analysis of the data.

For a two-state folding process, we may write (*21*)

$$\text{Folded (F)} \Leftrightarrow \text{Unfolded (U)}. \qquad (1)$$

Here, it is assumed that only the folded and unfolded conformations are present at significant concentrations. Consequently, $f_F + f_U = 1$, where f_F and f_U are the fraction of protein present in the folded and unfolded conformations, respectively. Thus, the observed value of the spectroscopic parameter y at any denaturant concentration will be $y = y_F f_F + y_U f_U$, where y_F and y_U represent the values of y characteristic of the folded and unfolded states, respectively, under the conditions where y is being measured. Combining these equations yields

$$f_U = \frac{y_F - y}{y_F - y_U} \quad . \qquad (2)$$

The equilibrium constant, K, and the free energy change, ΔG, can be calculated using

$$K = \frac{f_U}{f_F} = \frac{f_U}{1 - f_U} = \frac{y_F - y}{y - y_U} \qquad (3)$$

and

$$\Delta G = -RT \ln K = -RT \ln \frac{y_F - y}{y - y_U}, \qquad (4)$$

where R is the gas constant (1.987 cal/mol/K) and T is the absolute temperature. Values of y_F and y_U in the transition region are obtained by extrapolating from the pre- and post-transition regions. When y remains unchanged in the pre- or post-transition region, y_F and y_U can be simply the value of y at 0 M and 6 M denaturant concentration, respectively. Values of K can be measured most accurately near the midpoints of the denaturation curves, the errors become substantial for K values outside the range 0.1 — 10. Consequently, only ΔG values within ±1.5 kcal/mol are generally used.

In a scenario where more than one step is observed in the unfolding transition, intermediates have to be included in the equilibrium(*21*):

$$\text{Folded (F)} \Leftrightarrow \text{Intermediate 1 } (I_1) \Leftrightarrow \bullet\bullet\bullet\bullet\bullet\bullet$$

$$\Leftrightarrow \text{Intermediate n } (I_n) \Leftrightarrow \text{Unfolded (U)} \qquad (5)$$

Whether or not the intermediates are present directly on the unfolding pathway from F to U is not important, because the distribution is the same in the equilibrium. When stable intermediate states, I_i are present, and each characterized by the property y_i and fraction f_i, the observed extent of unfolding, e.g. Equation (2), becomes

$$f_{obs} = f_U + \sum_i f_i z_i \ , \qquad (6)$$

where

$$z_i = \frac{y_i - y_F}{y_U - y_F}. \qquad (7)$$

Thus, f_{obs} will differ from f_U by an amount that depends on the concentration of the intermediates weighted by their z_i values. The z_i value for an intermediate is the fractional change in y in going from F to I_i and is likely to be between 0 and 1, because y_i will generally fall between y_F and y_U (*21*).

The linear extrapolation model is used in the free energy analysis. The relationship between ΔG_D and the denaturant concentration is (*21*):

$$\Delta G_D = \Delta G_D^{H_2O} - m[denaturant] \qquad (8)$$

It is assumed that both ΔG_D and $\Delta G_{I,i}$ obey this model independently, i.e., they have different m values, where $\Delta G_{I,i}$ is the free energy difference between the fully folded state and the i-th intermediate state. Combining Equations (2)-(4) and (6)-(8), we can obtain

$$f_{obs} = f_U + \sum_i z_i f_i = \frac{e^{-\frac{\Delta G_D^{H2O} - m_D[D]}{RT}} + \sum_i z_i e^{-\frac{\Delta G_{I,i}^{H2O} - m_i[D]}{RT}}}{1 + e^{-\frac{\Delta G_D^{H2O} - m_D[D]}{RT}} + \sum_i e^{-\frac{\Delta G_{I,i}^{H2O} - m_i[D]}{RT}}}, \qquad (9)$$

where [D] is the concentration of denaturant, m_D and m_i are the m values for the fully unfolded state and the i-th intermediate state, respectively. All the other parameters are the same as defined before. Equation (9) has been used to fit the experimental data in all cases.

The data analysis was done using Microcal Origin (version 5.0). The built-in multi-Gaussian function was used to simulate the fluorescence spectra. User-defined functions were constructed for the two-state, three-state and four-state transitions and used to simulate the unfolding transition curve.

Results

GuHCl-induced unfolding of RdPf at pH 2

UV-visible absorption.

The guanidine hydrochloride-induced unfolding of RdPf in 20 mM potassium phosphate buffer at pH 2 and 25 ^{o}C, was monitored by UV-visible absorption spectroscopy. RdPf exhibits four absorption bands in the UV-visible region. The absorption band at 280 nm is contributed by tryptophan and tyrosine residues (*22*). The other three bands at 380 nm, 490 nm and 570 nm are due to the charge-transfer transitions between the iron and the four sulfurs (*23*). The UV-visible absorption spectrum of RdPf revealed no changes upon lowering the pH from 7.0 to 2.0 (*12*). The 380 nm band was chosen to follow the unfolding process because the differential absorption between the fully unfolded protein and the folded native protein is the largest.

The observed fraction of unfolded protein was obtained from Equation (2) by substituting A_{380} for y and plotted as a function of GuHCl concentration (Fig. 2). The absorbance of the fully folded RdPf was simply taken from the spectrum at 0 M GuHCl because A_{380} shows no change with GuHCl concentration in the pre-transition region (0 — 1 M). The protein was fully unfolded under 6.0 M GuHCl so that the absorbance is zero.

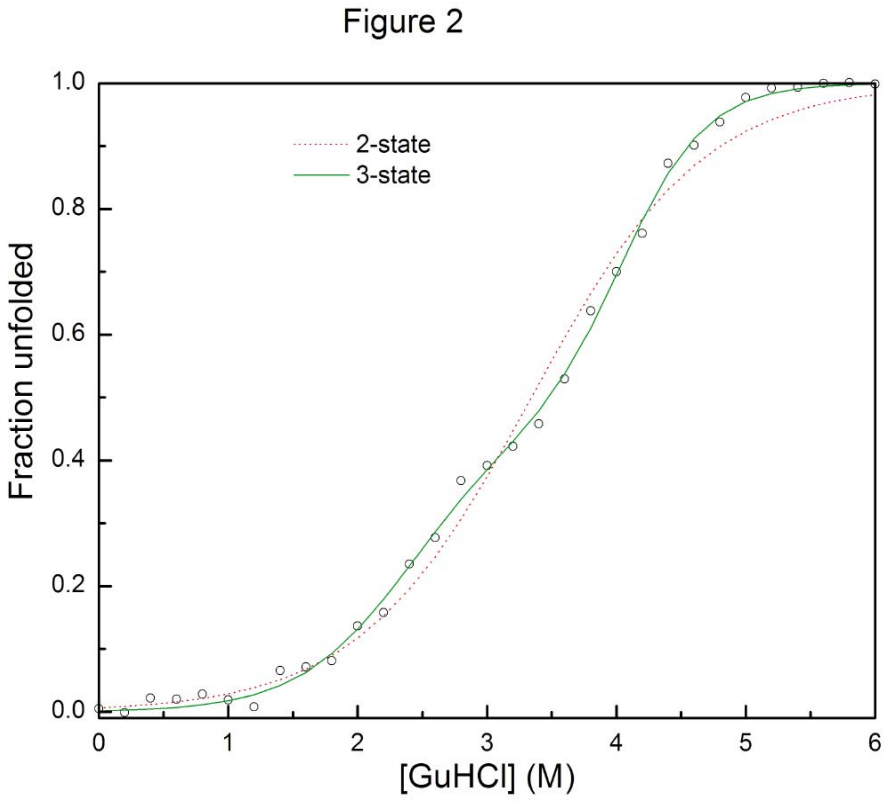


Figure 2. GuHCl-induced unfolding of RdPf at pH 2 as followed by UV-visible absorption spectra. The samples have been incubated at room temperature for 40 hours before recording. The unfolding curve is analyzed by taking the absorbance at 380 nm. Two-state and three-state models have been used to fit the transition curve.

It is clear that at least one intermediate exists in the unfolding transition, because there is a shoulder at about 3.2M GuHCl (Figure 2). A two-state transition has a smooth curve. The presence of intermediates in the unfolding is very important in determining the unfolding parameters. The presence of intermediates can greatly increase the unfolding free energy of a protein. For example, a two-state analysis of the α-subunit of

tryptophan synthase led to ΔG_D = 3.6 kcal/mol, but when the intermediate is taken into account, ΔG_D =11.0 kcal/mol (*24*).

Here, in the case of RdPf, the ΔG_D was 2.98 kcal/mol with a *m* value of 0.89 kcal/mol/M for a two-state model, while for a three-state model, the parameters were ΔG_D = 10.60 kcal/mol, m_D = 3.17 kcal/mol/M, ΔG_I = 3.27 kcal/mol, m_I = 1.37 kcal/mol/M, z_I = 0.46. The simulated two-state and three-state transition curves are also shown in Figure 2. It is very clear that the fit to a three-state model is much better than that to a two-state model. The results of fitting the data to a four-state model are included in Table 1.

Table 1. Parameters for two- and multi-state fits to the observed transition curves (units: ΔG/ kcal/mol, m/ kcal/mol/M)

Model	Spec	$\Delta G_{I1}^{H_2O}$	$\Delta G_{I2}^{H_2O}$	$\Delta G_D^{H_2O}$	m_{I1}	m_{I2}	m	z_1	z_2
	UV			2.98			0.89		
Two-	Fluo			3.82			1.18		
state	CD			3.24			1.08		
	UV	3.27		10.60	1.37		3.18	0.46	
Three-	Fluo	3.14		10.61	0.97		2.86	0.88	
state	CD	2.63		9.39	0.76		2.43	1.26	
	UV	2.10	5.33	13.16	1.19	2.45	4.35	0.10	0.51
Four-	Fluo	2.73	5.58	13.55	0.69	1.94	4.05	0.48	0.83
state	CD	1.72	4.16	13.09	0.15	1.64	4.05	0.56	0.78

Tryptophan fluorescence.

Figure 3 shows the GuHCl-induced unfolding of RdPf at pH 2.0 at room temperature as followed by tryptophan fluorescence spectroscopy. The fluorescence maximum shifts from about 335 nm at lower GuHCl concentration to about 355 nm at higher GuHCl concentration. The fluorescence intensity at 353.8 nm was used to analyze the data, because the ratio of the fluorescence intensity of the unfolded protein to the folded protein has a maximum value of about 5 at this wavelength. In determining the intensity of the folded and unfolded protein, again we found little dependence on the

Figure 3

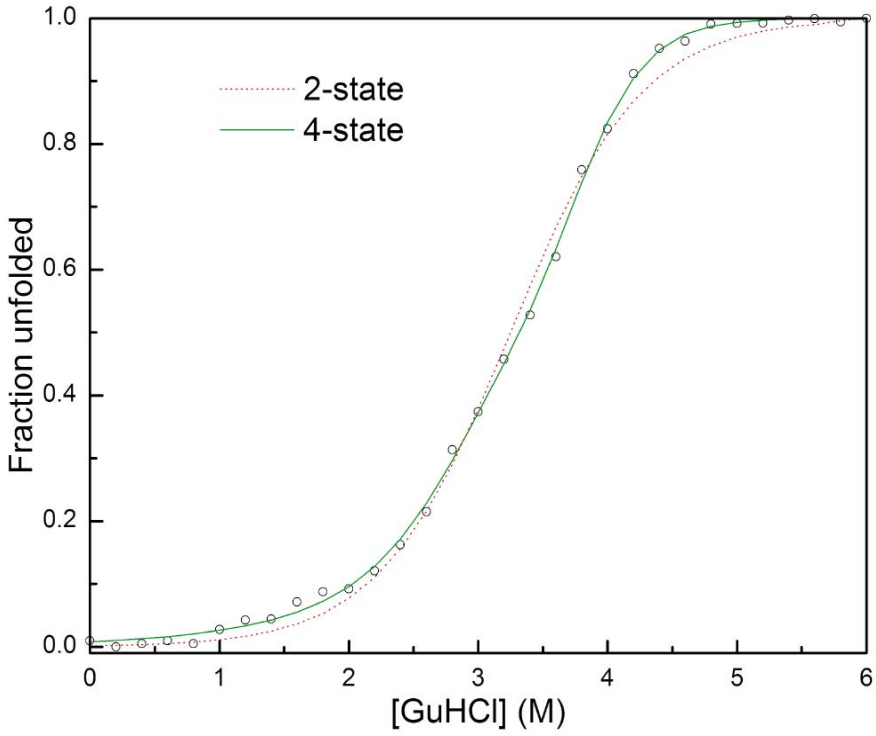


Figure 3. GuHCl-induced unfolding of RdPf at pH 2 and room temperature derived from tryptophan fluorescence spectra. The fluorescence intensity at 353.8 nm was used to calculate the fraction unfolded. Three-state and four-state models have been applied to fit the unfolding curve.

denaturant concentration. Surprisingly, another intermediate appears at 1.8 M GuHCl, and it is apparent that this intermediate is different from the one identified by UV-visible absorption spectroscopy. One possible reason that we could not observe this intermediate by the UV-visible absorption spectroscopy might be its similar extinction coefficient as the native state. In fact, the non-coincidence of different techniques can reassure the presence of intermediates (*25*). The data from tryptophan fluorescence were fitted to two-state, three-state and four-state models using equation (9) and the parameters are presented in Table 1.

Figure 4

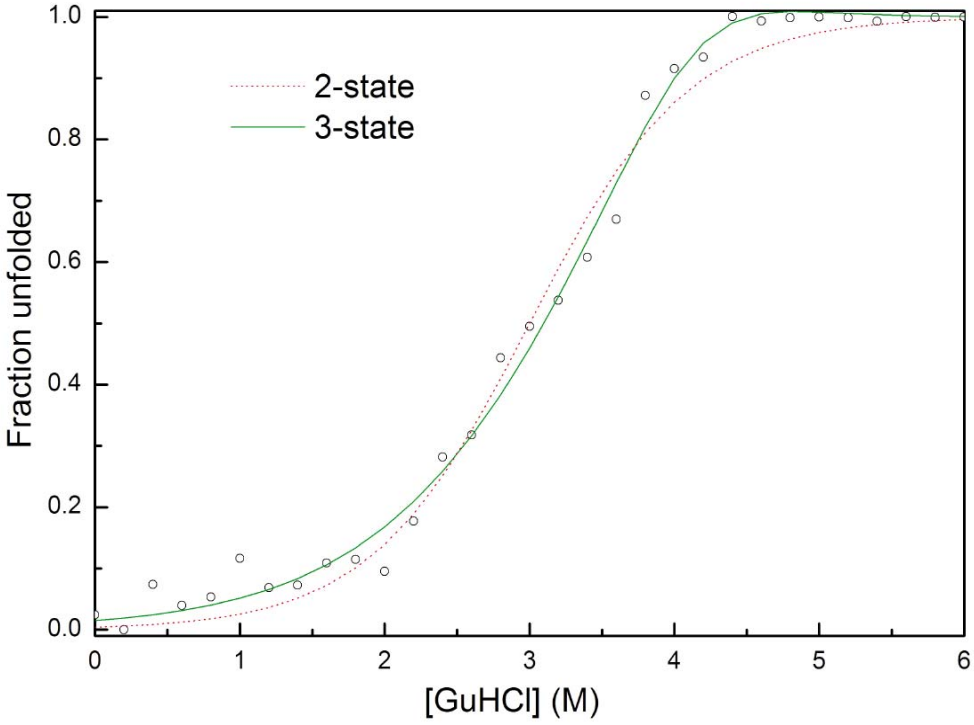


Figure 4. The unfolding transition curve of RdPf at pH 2 and room temperature as followed by circular dichroism. The fraction unfolded is calculated by taking the intensity at 225 nm. The curve has been fitted to a two-state and a three-state model.

Circular dichroism.

The secondary structure of RdPf includes a three-stranded β-sheet and several short α-helical turns (*3*). It is believed that the negative band at 225 nm represents the β-sheet content of RdPf(*12*). The ellipticity at 225 nm has been transformed to fraction unfolded by using equation (2). The fraction unfolded is shown in Figure 4. The fraction of unfolded changes irregularly at lower concentrations of GuHCl. However, in the transition region where the ellipticity is actually decreasing, the pattern is much smoother. To complicate matters, the CD band at 225 nm shows double negative peaks in several RdPf samples. This suggests that the change of the secondary structure of RdPf with GuHCl concentration is highly complex. Our attempts to fit the unfolding data to two-, three- and four-state models are summarized in Table 1.

Populational distribution of species at pH 2.

From the parameters in Table 1, the fractions of the folded forms(f_F), two intermediates (f_{I1} and f_{I2}) and the totally unfolded forms (f_U) can be calculated as a function of GuHCl concentration. The appropriate equations are

$$\frac{f_{I_1}}{f_F} = e^{-\frac{\Delta G_{I,1}^{H2O} - m_1[D]}{RT}},$$

$$\frac{f_{I_2}}{f_F} = e^{-\frac{\Delta G_{I,2}^{H2O} - m_2[D]}{RT}},$$

$$\frac{f_U}{f_F} = e^{-\frac{\Delta G_D^{H2O} - m[D]}{RT}},$$

and

$$f_F + f_{I_1} + f_{I_2} + f_U = 1.$$

Figure 5 shows the fraction distribution of these states for a four-state model with the parameter set deduced from the tryptophan fluorescence experiment. According to this model, the first intermediate state is already 5% populated at very low GuHCl concentration and accumulated to a maximum of 30% at 2.4 M GuHCl. The second intermediate reaches a maximum population of 70% at 3.3 M GuHCl. The fully unfolded form does not appear until about 2.5 M GuHCl.

Figure 5

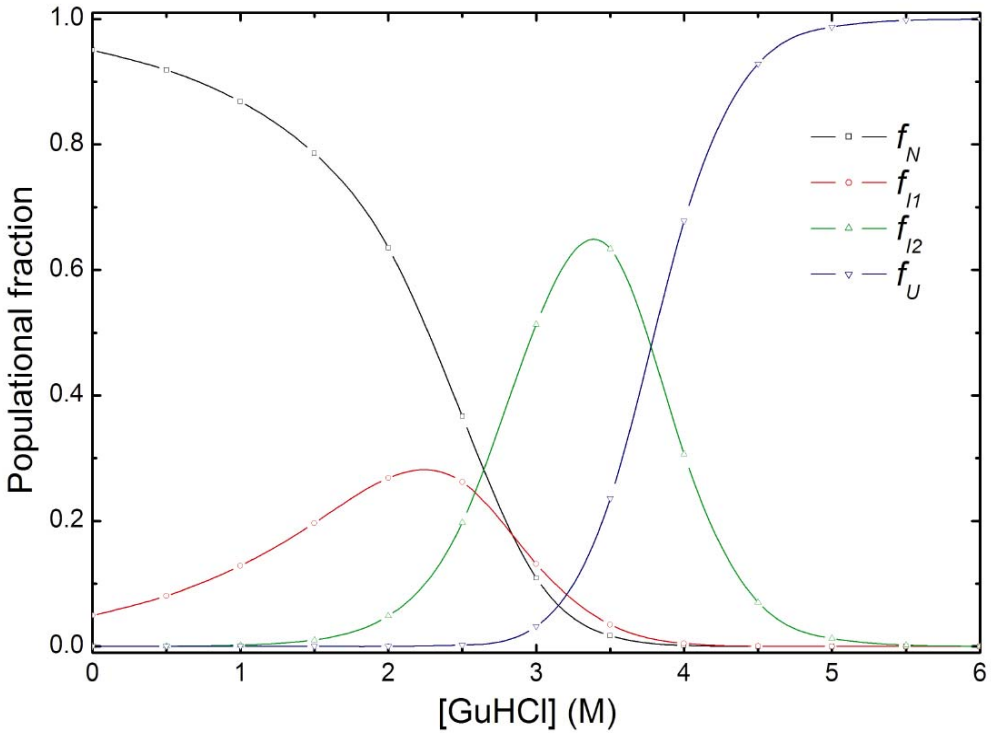


Figure 5. The population distribution of the fully folded, the two intermediates and the fully unfolded state as a function of GuHCl concentration. The parameters are from the four-state model fit with the tryptophan fluorescence experiment.

GuHCl-induced unfolding of RdPf at pH 7.0

GuHCl-induced unfolding of RdPf at pH 7.0 revealed that RdPf was partially denatured in 8 M GuHCl at pH 7. In Figure 6 a, b, c, we show the UV-visible, tryptophan fluorescence and UV-visible CD spectra of RdPf in 8 M GuHCl at 25 ^{o}C after 60 hrs.

Figure 6 (a)

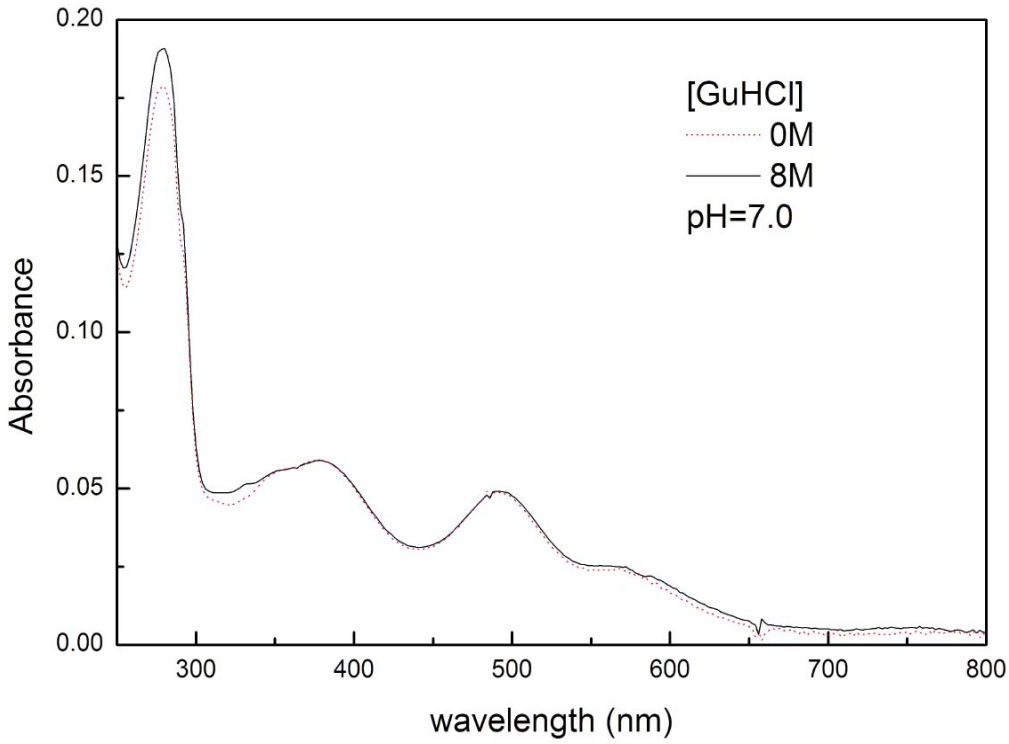


Figure 6. (a) GuHCl-induced unfolding of RdPf at pH 7.0 as monitored by UV-visible absorption spectra. The protein was incubated at 0 and 8M GuHCl at room temperature for 60 hours.

UV-visible absorption.

First, the visible absorption of RdPf shows a very little change. This indicates that iron-sulfur ligation is not perturbed in the presence of 8 M GuHCl. However, the

intensity of the peak at 280 nm decreases about 10%. The absorption band at 280 nm is mainly due to the absorption by tryptophan residues (*22*). RdPf has two tryptophan residues (residue 3 and residue 36). One possibility is that one of the tryptophan residues becomes exposed to the solvent at 8 M GuHCl.

Tryptophan fluorescence.

Tryptophan fluorescence is more sensitive than UV absorption as an indicator of its environment. As expected, the fluorescence of the two tryptophans reveals both an intensity increase and a shift of the maximum to the red. The fluorescence maximum of RdPf appears at 335 nm in 0 M GuHCl and 350 nm in 8 M GuHCl. 335 nm is the wavelength maximum for a buried tryptophan (*22*). On the other hand, 350 nm is not exactly the fluorescence wavelength maximum for an exposed tryptophan. However, the data are consistent with one buried tryptophan residue, and one exposed and solvated at 8 M GuHCl. Given that the fluorescence of native RdPf arises from two buried tryptophan residues, half of the fluorescence of RdPf in 0 M GuHCl can be subtracted from the fluorescence of RdPf in 8 M GuHCl. The residue spectrum was then fitted as a function of wavenumber by a single Gaussian function centering at 355 nm (Figure 7). The resultant fluorescence spectrum is consistent with an exposed tryptophan. Molecular dynamics simulation studies on the unfolding pathway of RdPf has predicted that Trp 36 is among the most labile residues along the unfolding pathway (*19*). A site-directed mutation Trp 36 $\rightarrow$ Phe could resolve this question.

Figure 6 (b)

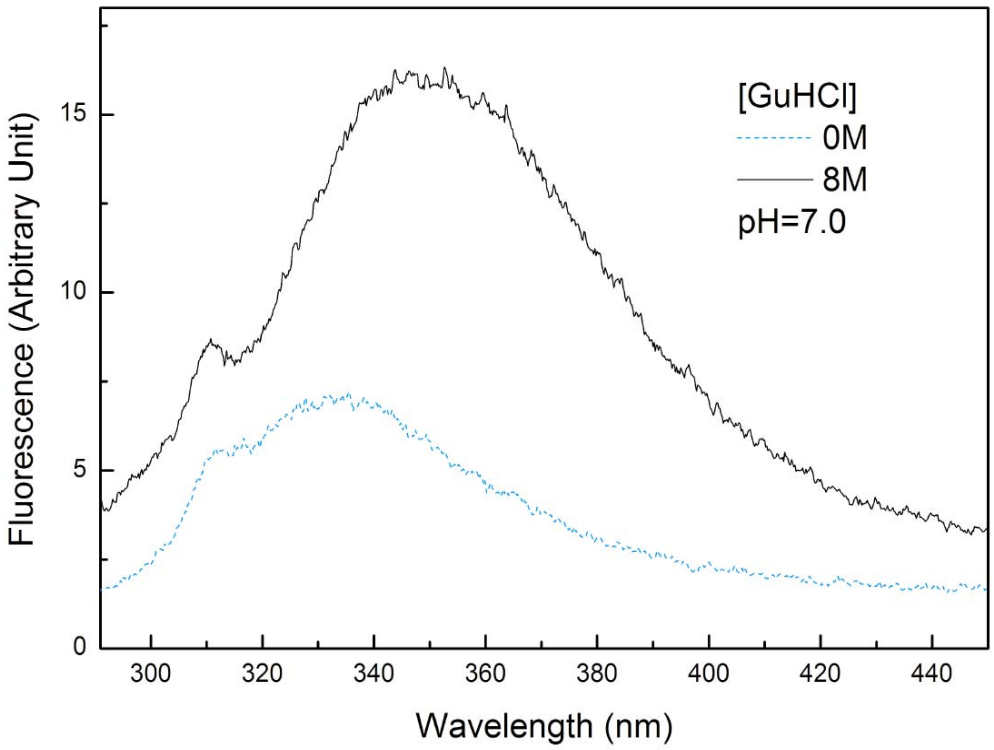


Figure 6. (b) GuHCl-induced unfolding of RdPf at pH 7.0 as monitored by tryptophan fluorescence spectra. The protein was incubated at 0 and 8M GuHCl at room temperature for 60 hours.

Circular dichroism.

From the CD spectra (Figure 6 c), we conclude that the secondary structure content decreases about 15% at 8 M GuHCl compared to 0 M GuHCl. This could be due to the partial disruption of the β-sheet. Another change noted occurs in the aromatic group region around 500 nm. The negative bands are contributed by some rigid aromatic groups (*22*). The decrease in the 470 nm peak indicates that a number of aromatic groups have become more flexible, possibly exposed and solvated, in the presence of 8 M

GuHCl, which is in accordance with the observations by UV and fluorescence spectroscopy.

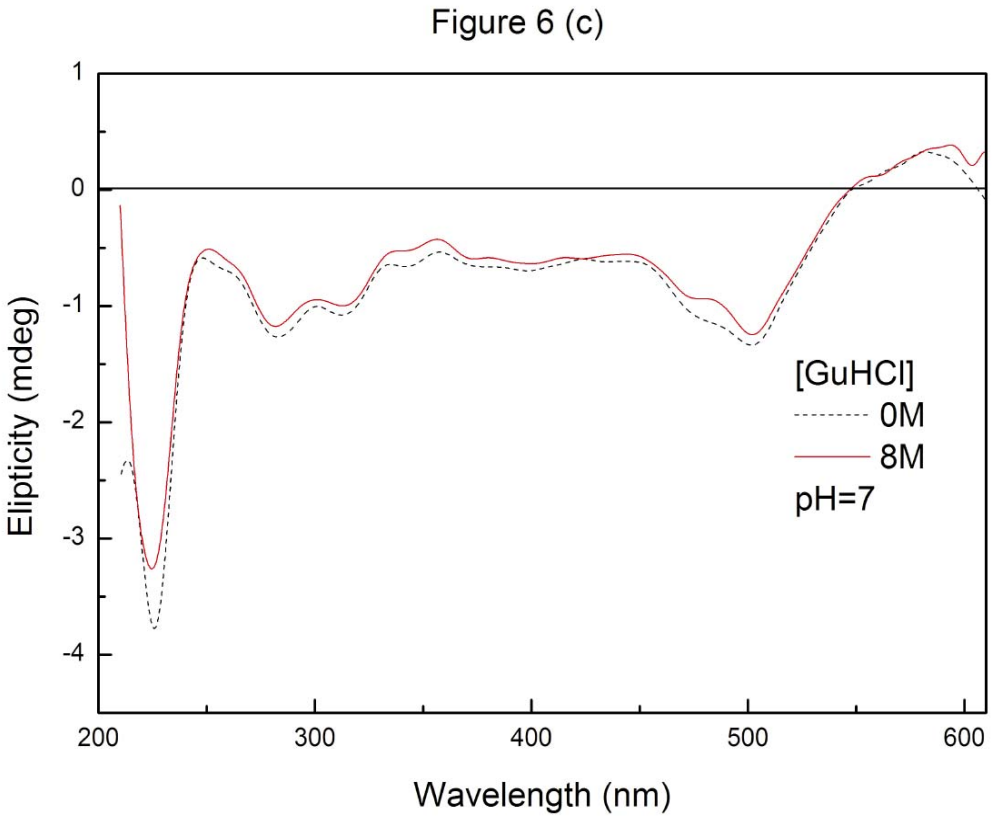


Figure 6. (c) GuHCl-induced unfolding of RdPf at pH 7.0 as monitored by UV-visible CD. The protein was incubated at 0 and 8M GuHCl at room temperature for 60 hours.

Figure 7

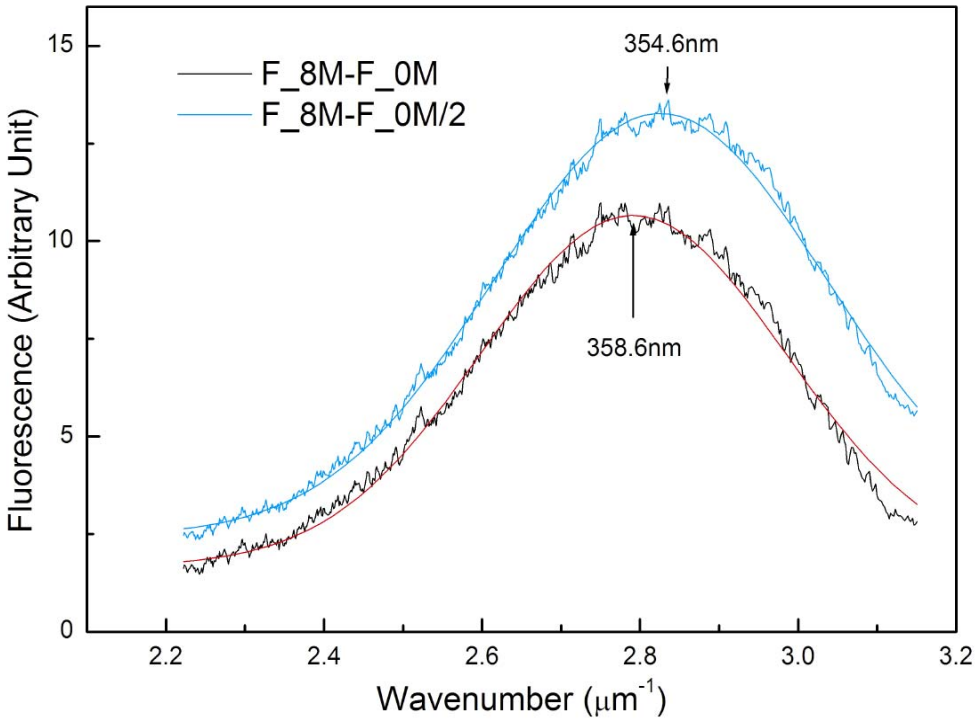


Figure 7. The difference fluorescence spectra of RdPf at 8M GuHCl (F_{8M}) and 0M GuHCl (F_{0M}). The spectra are plotted *vs*. wavenumber μm^{-1}. The whole and half of F_{0M} were subtracted from F_{8M} in F_{8M}-F_{0M} and F_{8M}-½F_{0M}, respectively. The two spectra have been simulated by a single Gaussian function (peak wavelength shown in curves).

Discussions

The equilibrium has been reached before measurement

An important issue in the equilibrium unfolding experiments here is that we have allowed the protein to reach equilibrium before measurements are taken. Thermal denaturation of RdPf has been reported to be irreversible (*10, 26*). The denaturant-

induced unfolding studied here indicates that we are observing the unfolding under reversible conditions. On the other hand, we must remind ourselves that RdPf is strongly resistant toward denaturation. The protein does not denature at 6 M GuHCl at neutral pH and room temperature even after 24 hours of incubation (*20*). In our experiments, we have incubated the solutions for 40 to 80 hours to ensure that the equilibrium has been reached. We took pains to ascertain that there were no spectroscopic changes between 40 and 80 hours of incubation (data not shown). Finally, we have observed the kinetics of the unfolding of RdPf by monitoring the absorption at 380 nm. Several transient phases were observed, with the final phase occurring on a time constant of about an hour. The absorption value matched the observed equilibrium value when the kinetic curve was extrapolated to infinite time. When we tried to reverse the process by diluting the solution to lower the concentration of GuHCl from 6 M to 3 M, only a small part of the absorbance at 380 nm was restored after 40 hours. This suggests that either the reversed reaction is extremely slow, or the incorporation of iron into the apo-protein requires additional factors.

The properties of the intermediate states

It is possible to derive some characteristics of the unfolding intermediates and the unfolded state from the spectroscopic properties that we have observed here. First, the fully unfolded state of RdPf has no iron-sulfur coordination as the visible absorption bands are absent in 6 M GuHCl at pH 2. The secondary structure is disrupted as monitored by CD spectroscopy. The hydrophobic core is exposed because the tryptophan fluorescence has an intensity maximum at 355 nm, which is the characteristic feature of

solvated tryptophan residues (*22*). The fluorescence intensity is 5 times stronger than that of the folded RdPf, which is an indication of the loss of iron. Iron strongly quenches the tryptophan fluorescence but does not shift the wavelength (the fluorescence of apo-RdPf has a maximum at 335 nm). On the other hand, solvation decreases tryptophan fluorescence and shifts the maximum intensity from 335 nm to 355 nm. The ratio of the intensity is about 20 for the former effect, but only 4 for the latter. These factors account for the observed 5-fold increase in the fluorescence intensity of RdPf from the folded to unfold state.

Second, there remains the question as to whether or not in the unfolded RdPf disulfide bonds are formed when the iron atom has been displaced. We have attempted to address this issue by adding 50 mM DTT to the solution in the unfolding experiments of RdPf at pH 2. If disulfide bonds were formed in the unfolded protein, the free energy difference between the native state and the unfolded state should be different. In our experiments, no difference was observed in ΔG_D. Also, the expected CD bands around 250 nm for the disulfide bond did not appear. We thus conclude that there is no oxidation of the sulfhydryls upon unfolding of the RdPf. The reason may be the low pH value used in the present study. When the iron-sulfur bonds are broken, there are two possible reactions: (i) oxidation by di-oxygen; and (ii) the protonation. The reactions compete with each other. The high concentration of protons at pH 2 would favor the latter. At neutral pH, there is at least one disulfide bond formed upon displacement of the iron from the protein (Zhang et al., unpublished results).

Generally a *z* value close to 0 implies that the intermediate is more similar to the folded form than the unfolded form, while a *z* value close to 1 implies a structure that is closer to the unfolded form. The *z* values of the first intermediate of RdPf observed at pH 2 are 0.10, 0.48 and 0.56 as monitored by UV-visible absorption, fluorescence and CD, respectively. These properties are expected from the partially unfolded form of RdPf at pH 7, and thus the same intermediate may have been formed. The partially unfolded form at pH 7 has native-like iron-sulfur ligation. When the visible absorption by the iron-sulfur cluster is used as a criterion, this partially unfolded state is very close to the folded state. That is, a small *z* value close to 0 is noted. One of the tryptophan residues is exposed in the partially unfolded form of RdPf. The fluorescence gains some intensity because the tryptophan residue is further away from the iron. On the other hand, it loses some intensity because of solvation. A *z* value of 0.48 for fluorescence is a reasonable value for such a structure. A *z* value of 0.56 is higher than expected for the CD property of the intermediate, because only 15% of the CD change was observed at 8 M GuHCl at pH 7. However, this may be peculiar to the irregular CD change in the low GuHCl concentration region.

A *z* value of 0.83 for the second intermediate at pH 2 from the fluorescence is almost twice the *z* value of the first intermediate. Thus, the two tryptophan residues are both exposed in the case of the second intermediate. A *z* value of 0.51 for the absorption at 380 nm suggests that the intermediate has a distorted iron-sulfur coordination. A *z* value of 0.78 from CD means that the intermediate is almost devoid of secondary structure.

From Figure 5, the first intermediate has a population about 5% in H_2O with no denaturant at pH 2. This result is in agreement with the previous observation that RdPf can bind ANS at lower ionic strengths (*12*). Since the intermediate has a tryptophan residue exposed, ANS binding is expected and can shift the equilibrium to the ANS-bound form. The other tryptophan residue is still buried, thus only a portion of the hydrophobic core is open, and the observed 1:1 binding is reasonable. At higher ionic strengths, the fully unfolded state is stabilized and ANS binding is impossible. This suggests that the fully folded state is destabilized by electrostatic interactions at pH 2. This conclusion is borne out by the distribution of surface charges under these conditions; at pH 2, the protein surface of RdPf is positively charged, especially in the region of the closely spaced A1, K2, K28 and K50 (*12*).

Comparison with other studies on RdPf unfolding

It is interesting to compare this unfolding pathway with that deduced from earlier studies. The unfolding pathway has been suggested from thermal unfolding (*27*): The native protein first loses some secondary structure, then releases the iron; more secondary structure is then lost, and this is finally followed by the formation of unfolded state. While our CD data do support the loss of the secondary structure, we observe the release of the iron only in the final step of the GuHCl denaturation examined here. Another theoretical study (*19*) on the thermal unfolding pathways of RdPf concluded that RdPf unfolds first by opening of the loop region to expose the hydrophobic core, followed by the unzipping of the β-sheet. Although the disruption of the iron is not reported by this

study, the simulation time was probably too short to account for the iron loss. Because the fully unfolded state exists in a high free energy state, it will require rather long time to observe the event in simulation.

The unfolding free energy has been given in the native state hydrogen exchange experiment on RdPf (*11*). A value of about 17 kcal/mol was obtained from the free energy — temperature curve. This is in very good agreement with our current study if we consider the difference in our experiment conditions. The unfolding free energy obtained in our study is 3.4 kcal/mol less. It is presumably due to the disruption of the salt bridges in RdPf by the low pH condition.

We conclude this study with some thoughts on the contributions of different factors to the hyperthermostability of RdPf. These factors include the secondary structures, electrostatic interactions, hydrophobic core and the iron-sulfur ligation. The contribution of the iron-sulfur ligation can be obtained from the free energy difference of the fully unfolded state and the partially unfolded state in 3.3 M GuHCl at pH 2, which is supposed to be essentially devoid of tertiary packing. The result is 13.6 - 5.6 = 8.0 kcal/mol. Thus, we can conclude that this is the most important factor that contributes to the hyperthermostability of RdPf. However, this factor alone cannot account for the hyperthermostability of RdPf, presumably because it makes a similar contribution to the stability of mesophilic rubredoxins. The unfolding study on of a RdPf variant without the Fe-$(Cys)_4$ center gives an unfolding free energy 3.2 kcal/mol at 1 ^{o}C (*14*). It seems that the only difference between this variant and native RdPf is the Fe-$(Cys)_4$ center.

However, this gives a total unfolding free energy of 11.2 kcal/mol for native RdPf. The difference might be partly due to the temperature difference in these studies. It is also possible that the small change in the structure of the variant actually has a bigger impact on the overall stability of the protein than we expect. The contributions from secondary structures and hydrophobic packing can be estimated from the free energy difference between the two intermediates, which is 5.6-2.7 = 2.9 kcal/mol. The contribution from the electrostatic interactions can be estimated by comparing the stability of RdPf at pH 2 and pH 7. At pH 7, only the first unfolded state is reached under 8M GuHCl, while at pH 2 the protein is fully unfolded. If the free energy difference between the folded state and the partially unfolded state at pH 7 is known, we can estimate the contribution of the salt bridges by comparing it to the free energy difference between the folded state and the first partially unfolded state at pH 2. On the other hand, the 3.4 kcal/mol unfolding free energy difference between the hydrogen exchange experiment (*11*) and our study here gives a first-order approximation for the contribution of the electrostatic interactions. In any case, we know that these salt bridges play an important role in stabilizing RdPf. The salt bridges might be an even more important stabilizing factor at high temperatures, if we take into account the fact that the dielectric constant of water decreases from 82.20 at 20^{o}C to 55.51 at 100 ^{o}C.

To summarize, even though the contributions from the different factors to the hyperthermostability cannot be dissected clearly, we conclude that the folding and unfolding of this protein is not highly co-operative. Single-domain proteins are normally only marginally stable, and the distribution of protein conformational states is such that

the protein folds or unfolds essentially cooperatively. It is evident that proteins can be produced more stable thermodynamically, though the judicious addition of salt bridges and hydrophobic bonds, and this increased stability must necessarily come at the expense of decreased overall conformational flexibility, which is reflected in the cooperativity of the folding and unfolding of the protein. The presence of intermediates is consistent with a dispersion of conformational states, each set responding differentially to denaturing forces, either thermal or chemical denaturants. The protein merely shifts its structure among these different subsets of accessible states under varying environmental conditions. Thus, enhanced protein stability does not require strong stabilizing forces, merely tuning of the forces to accomplish a redistribution of thermally accessible conformational states.

References

1. Adams, M. W. W. & Kelly, R.M. (1995) *Chem. Eng. News 73*, 32-42.
2. Blake, P. R., Park, J.-B., Zhou, Z.H., Hare, D.R., Adams, M.W.W. & Summers, M.F. (1992) *Protein Sci. 1*, 1508-1521.
3. Day, M. W., Hsu, B. T., Joshua-Tor, L., Park, J.-B., Zhou, Z. H., Adams, M. W. W., & Rees, D. C. (1992) *Protein Sci. 1*, 1494-1507.
4. Korndorfer, I., Steipe, B., Huber, R., Tomschy, A., Jaenicke, R. (1995) *J. Mol. Biol. 246*, 511-521.
5. Chan, M. K., Mukund, S., Kletzin, A., Adams, M.W.W., & Rees, D.C. (1995) *Science 267*, 1463-1469.
6. Starich, M. R., Sandman, K., Reeve, J.N., & Summers, M.F. (1996) *J. Mol. Biol. 255*, 187-203.
7. Yip, K. S. P., Stillman, T.J., Britton, K.L., Artymiuk, P.J., Bocker, P.J., Sedelnikova, S.E., Engel, P.C., Pasquo, A., Chiavaluce, P., Consalvi, V., Scandurra, R., & Rice, D.W. (1995) *Structure 3*, 1147-1158.
8. Chen, L., Liu, M.-Y., Le Gall, J., Fareleira, P., Santos, H. & Xavier, A.V. (1993) *Biochem. Biophys. Res. Commun. 193*, 100-105.
9. Kraulis, P. J. (1991) *J. Appl. Crystallogr. 24*, 946-950.
10. Klump, H. H., Adams, M.W.W. & Robb, F.T. (1994) *Pure & Appl. Chem. 66*, 485-489.
11. Hiller, R., Zhou, Z.H., Adams, M.W.W. & Englander, S.W. (1997) *Proc. Natl. Acad. Sci. USA 94*, 11329-11332.
12. Cavagnero, S., Zhou, Z.H., Adams, M.W.W. & Chan, S.I. (1995) *Biochemistry 34*, 9865-9873.
13. Eidsness, M. K., Richie, K.A., Burden, A.E., Kurtz, D.M.J., & Scott, R.A. (1997) *Biochemistry 36*, 10406-10413.
14. Strop, P. a. M., S. L. (1999) *J. Am. Chem. Soc. 121*, 2341-2345.
15. Strop, P. a. M., S. L. (2000) *Biochemistry 39*, 1251-1255.
16. Hernandez, G., Jenney, F. E. Jr, Adams, M. W. W., & Le Master, D. M. (2000) *Proc. Natl. Acad. Sci. USA 97*, 3166-3170.
17. Bradley, E. A., Stewart, D.E., Adams, M.W.W., & Wampler, J.E. (1993) *Protein Sci. 2*, 650-665.
18. Jung, D. H., Kang, N.S., & Jhon, M.S. (1997) *J. Phys. Chem. A 101*, 466-471.
19. Lazaridis, T., Lee, I. & Karplus, M. (1997) *Protein Sci. 6*, 2589-2605.
20. Blake, P. R., Park, J.-B., Bryant, F. O., et al., Adams, M.W.W. (1991) *Biochemistry 30*, 10885-10895.
21. Pace, C. N. (1986) *Methods in Enzymology 131*, 266-280.
22. Schimid, F. X. (1990) in *Protein Structure - A Practical Approach* (Creighton, T. E., Ed.) pp 253, IRL Press at Oxford University Press, Oxford, England.
23. Yachadra, V. K., Hare, J., Moura, I., & Spiro, T.G. (1983) *J. Am. Chem. Soc. 105*, 6455-6461.
24. Matthews, C. R. C., M.M. (1981) *Biochemistry 20*, 784-792.
25. Saito, Y. W., A. (1983) *Biopolymers 22*, 2105-2122.

26. Cavagnero, S., Zhou, Z.H., Adams, M.W.W. & Chan, S.I. (1998) *Biochemistry 37*, 3377-3385.
27. Cavagnero, S., Debe, D., Zhou, Z.H., Adams, M.W.W. & Chan, S.I. (1998) *Biochemistry 37*, 3369-3376.

Appendix III

Femtosecond Dynamics of Rubredoxin: Tryptophan Solvation and Resonance Energy Transfer in the Protein*

* This appendix is a paper published in PNAS coauthored with Dongping Zhong, Samir Kumar Pal, Sunney I. Chan, and Ahmed H. Zewail.

Femtosecond dynamics of rubredoxin: Tryptophan solvation and resonance energy transfer in the protein

Dongping Zhong, Samir Kumar Pal, Deqiang Zhang, Sunney I. Chan, and Ahmed H. Zewail*

Laboratory for Molecular Sciences, Arthur Amos Noyes Laboratory of Chemical Physics, California Institute of Technology, Pasadena, CA 91125

Contributed by Ahmed H. Zewail, October 31, 2001

We report here studies of tryptophan (Trp) solvation dynamics in water and in the *Pyrococcus furiosus* rubredoxin protein, including the native and its apo and denatured forms. We also report results on energy transfer from Trp to the iron-sulfur [Fe-S] cluster. Trp fluorescence decay with the onset of solvation dynamics of the chromophore in water was observed with femtosecond resolution (≈160 fs; 65% component), but the emission extended to the picosecond range (1.1 ps; 35% component). In contrast, the decay is much slower in the native rubredoxin; the Trp fluorescence decay extends to 10 ps and longer, reflecting the local rigidity imposed by residues and by the surface water layer. The dynamics of resonance energy transfer from the two Trps to the [Fe-S] cluster in the protein was observed to follow a temporal behavior characterized by a single exponential (15–20 ps) decay. This unusual observation in a protein indicates that the resonance transfer is to an acceptor of a well-defined orientation and separation. From studies of the mutant protein, we show that the two Trp residues have similar energy-transfer rates. The critical distance for transfer (R_0) was determined, by using the known x-ray data, to be 19.5 Å for Trp-36 and 25.2 Å for Trp-3, respectively. The orientation factor (κ^2) was deduced to be 0.13 for Trp-36, clearly indicating that molecular orientation of chromophores in the protein cannot be isotropic with κ^2 value of 2/3. These studies of solvation and energy-transfer dynamics, and of the rotational anisotropy, of the wild-type protein, the (W3Y, I23V, L32I) mutant, and the fmetPfRd variant at various pH values reveal a dynamically rigid protein structure, which is probably related to the hyperthermophilicity of the protein.

Tryptophan (Trp) is the most important fluorophore among amino acid residues for optical probing of proteins. However, Trp fluorescence is complex because of different rotamers in the ground state and the two nearly degenerate electronic states (^1L_a, ^1L_b) with perpendicular transition moments. Accordingly, numerous studies (1–10) have focused on the lifetime, quantum yield, Stokes shift, and fluorescence anisotropy. Most of these studies were made with picosecond or nanosecond time resolution (3, 6–10). To probe the local protein dynamics, Trp solvation by neighboring soft/rigid water molecules, or by other polar amino acid residues, must be resolved on the femtosecond time scale. Moreover, such studies are important for examining the nature of resonance energy transfer (RET) that is used for deducing distances and orientations between the Trp and quenchers in the protein (e.g., see refs. 3 and 10, and references therein).

We choose the hyperthermophilic iron-sulfur protein, *Pyrococcus furiosus* rubredoxin (PfRd), as a prototype system (Fig. 1). The high-resolution x-ray crystallographic structures of PfRd at 0.95 Å and its formylmethionine variant (fmetPfRd) at 1.2 Å, have been recently reported (11). PfRd is a small protein of 53 amino acid residues with an iron atom coordinated by the sulfur atoms of four cysteine side chains and functions as an electron-transfer protein (12). It is approximately ellipsoidal in overall shape with a hydrophilic tail protruding into the solvent at the C terminus region. The structure consists of a three-stranded antiparallel β-sheet with a hydrophobic core containing six aromatic residues (Fig. 1); two of them are Trp-3 and Trp-36. The [Fe-S] cluster has a strong charge-transfer absorption band at 380 nm (13, 14), which overlaps with the Trp emission. Thus, RET between Trp and the cluster is expected and has been observed in other iron-sulfur proteins (15).

Experimental Methods

All experimental measurements were carried out by using the femtosecond-resolved fluorescence up-conversion technique (16). All protein samples were generously provided by the group of Michael W. W. Adams at the University of Georgia. L-Trp, *N*-acetyl-L-tryptophanamide (NATA) and *N*-acetyl-L-Trp ethyl ester (NATEE) were purchased from Sigma. Ultrapure guanidine hydrochloride (GdnHCl) was obtained from Baker, and trichloroacetic acid from Fisher. All chemicals were used as received.

The iron-sulfur protein rubredoxin we obtained was isolated and purified from *Pyrococcus furiosus*. The protein was concentrated and stored in 50 mM Tris buffer (pH 8) with 0.3 M NaCl at −20°C. Its purity was checked routinely by measuring the ratio of the extinction coefficient at 280 nm and 380 nm (17). For most experiments, the formylmethionine variant of PfRd (fmetPfRd) was used with a concentration of 0.3 mM in a 20 mM phosphate buffer at pH 7. Aside from the wild-type PfRd, we also examined the (W3Y, I23V, L32I) mutant of fmetPfRd in which the Trp-3 was replaced by Tyr, leaving only one Trp (Trp-36) in the protein.

Apo-fmetPfRd was prepared (17) first by denaturing the holo-protein in 10% trichloroacetic acid. After a few hours, the precipitated apo-fmetPfRd was suspended twice in 6% trichloroacetic acid and then dissolved in 3 M GdnHCl. After renaturation of the protein by dialysis against 50 mM Tris buffer at pH 8, the apo-PfRd solution was clear. Circular dichroism showed that the apo protein has a similar secondary structure to the holo form.

The steady–state absorption and fluorescence spectra (265-nm excitation) are shown in Fig. 2. The distinctive absorption spectrum has an intense band at 280 nm (Trp absorption). Other bands at 380 nm, 490 nm, and 570 nm (not shown) arise from charge transfer within the [Fe-S] cluster, from the cysteinyl thiolates to Fe(III) (13, 14). The mutant and the variant of PfRd show the same absorption as the wild-type PfRd, but their thermostabilities are slightly different (18). The fluorescence emission of apo-fmetPfRd peaks at 337 nm. For the native protein, the Trp emission at the red side strongly overlaps with the iron-sulfur charge-transfer absorption and thus it is

Abbreviations: Trp, tryptophan; fmetPfRd, formylmethionine variant of *Pyrococcus furiosus* rubredoxin; RET, resonance energy transfer; NATA, *N*-acetyl-L-tryptophanamide; NATEE, *N*-acetyl-L-Trp ethyl ester.

*To whom reprint requests should be addressed. E-mail: zewail@caltech.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

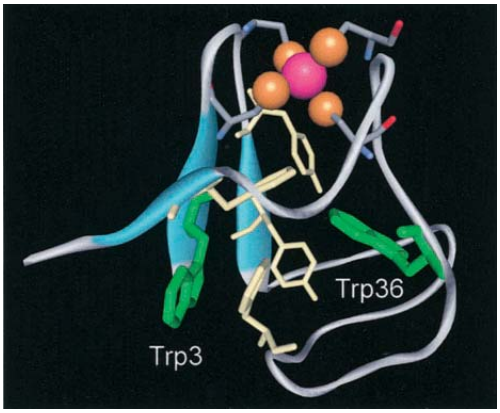


Fig. 1. Ribbon presentation of high-resolution x-ray structure of hyperthermophilic *Pyrococcus furiosus* rubredoxin (11). Six aromatic amino acid residues in the hydrophobic core are, along the primary sequence, Trp-3, Tyr-10, Tyr-12, Phe-29, Trp-36, and Phe-48. A tail at the C terminus protrudes into the solvent.

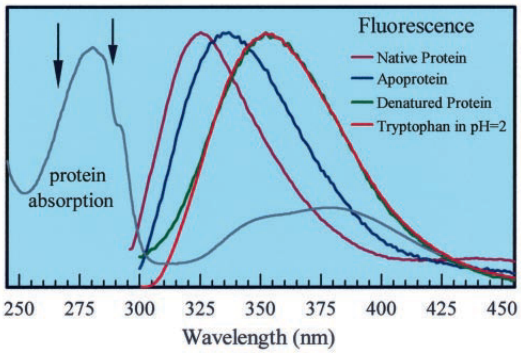


Fig. 2. Absorption of the PfRd protein and normalized fluorescence emission for different systems. Note that the spectral overlap between the Trp emission in apo-PfRd and the [Fe-S] cluster absorption in PfRd. The fluorescence intensity in PfRd is actually much weaker than that in apo-PfRd. The arrows mark two excitation wavelengths used in this study, 265 nm and 288 nm.

quenched through RET by several orders of magnitude. Accordingly, the resulting spectrum is blue-shifted, peaking at 325 nm (Fig. 2) because of the comparable time scales of energy transfer and solvation (see below).

The fluorescence emission of Trp in buffer solution shifts to the red, 352 nm, because of solvation. Clearly, the Trp residues buried in apo-fmetPfRd (337 nm) are in a more hydrophobic environment than in the buffer solution. The denatured form of fmetPfRd has a similar emission spectrum to that of Trp in buffer solution, except for a tail at the blue side because of the tyrosine emission of the protein. To mimic the peptide bond in protein, we also studied the model systems: NATA in a phosphate buffer solution (30 mM) at pH 2 and NATEE in a solvent of *p*-dioxane. The emission spectrum of NATA is similar to that of Trp with a slight shift to the red side, by 5 nm; for NATEE it shifts to the blue side and peaks at 330 nm.

Results and Discussion

Excited States of Trp. The two electronic excited states, ^1L_a and ^1L_b, are both involved in the absorption and fluorescence emission. Two transition dipoles are nearly perpendicular to each other, and the direction of the ^1L_b transition dipole is along the side chain (19). The ^1L_a state has a larger static dipole moment than its ground state so it is more sensitive to solvation. In polar solvents, the ^1L_a state is red-shifted and becomes lower in energy than the ^1L_b state. The observed steady–state fluorescence, especially at wavelengths longer than 340 nm, is mainly from the ^1L_a state (8). For each electronic state, there are three rotamers of the alanyl side chain of Trp (1). The reported two principal lifetimes, ≈500 ps and ≈3 ns in bulk solution, were attributed to different conformers, and, as discussed below, our observed long-time component (>500 ps) at different wavelengths is an average value of the two lifetimes.

Solvation Dynamics of Trp. Fig. 3*A* shows femtosecond-resolved fluorescence transients of Trp in phosphate buffer at pH 2 with a systematic series of detection wavelengths. All transients have three distinct time scales. At the blue end of 310 nm, the signal decays with time constants of 700 fs (78%), 3.13 ps (8.6%), and ≈518 ps (13.4%); at 340 nm, it first rises in 200 fs and then decays in 1.8 ps (20%) and ≈865 ps (80%); at 370 nm, it rises in 330 fs (86%) and 1.9 ps (14%) and then decays in ≈1.12 ns; and at the red end of 440 nm, the transient rises in 410 fs (57%) and 2.21 ps (43%) and finally decays in ≈1.45 ns. We also studied NATA in the same buffer and similar transients were observed. For NATEE in *p*-dioxane, we observed slower temporal behaviors, e.g., at 310 nm the signal decays in 4.7 ps (58%) and ≈500 ps (42%) and at 340 nm it rises in 420 fs (72%) and 11 ps (28%) and then decays in ≈3.2 ns. These results are indicative of solvation.

The initial femtosecond decay at the blue side and the rise at the red side dominantly result from solvation processes, and are not due to the electronic relaxation (^1L_a and ^1L_b coupling) and vibrational cooling. This conclusion was based on the following observations: First, the internal conversion between ^1L_a and ^1L_b occurs in <100 fs as measured by ultrafast anisotropy decay (see below). Recent studies deduced a time constant of 10–40 fs for the internal conversion of 5-methoxyindole in hexadecane (20). Second, Trp emission strongly depends on solvent polarity. From *p*-dioxane (or in the protein) to buffer solution, the emission peak shifts from 330 nm (337 nm) to 357 nm. Third, the initial Trp dynamics also show different temporal behaviors in different solvents. In *p*-dioxane it occurs in several picoseconds (21) whereas in water it is on the femtosecond time scale, as expected for water solvation (22, 23). Finally, all transients gated at various emission wavelengths are nearly independent of excitation wavelength (265 nm and 288 nm), ruling out a large contribution from vibrational cooling.

By following the time-resolved emission (Stokes shift with time), we constructed the correlation function (solvent response function) to obtain the solvation time: $c(t) = [\nu(t) - \nu(\infty)]/[\nu(0) - \nu(\infty)]$, where $\nu(t)$, $\nu(0)$, and $\nu(\infty)$ are time-resolved emission maxima in cm^{-1}, respectively. The c(t) function, shown in Fig. 3*A Inset* gives an apparent biexponential behavior: 160 fs (65%) and 1.1 ps (35%). These two solvation times are close to the reported values (126 fs and 880 fs) in bulk water (23). The former reflects the librational motion of water molecules and the latter represents their diffusive motion.

After establishing the Trp solvation dynamics in bulk water, we studied Trp solvation in the apo-fmetPfRd and the denatured form of the protein. The results are given in Fig. 3 *B* and *C*, respectively. In apo-fmetPfRd, the transient at 310 nm decays in 1.2 ps (17%), 12 ps (26%), and ≈320 ps (57%); for 340 nm, it first rises in 200 fs and then decays ≈530 ps and for 370 nm, it rises in 200 fs (89%), 5.6 ps (11%) and then decays ≈640 ps. Clearly, Trp buried in apo-fmetPfRd shows multiexponential temporal behavior and the solvation processes become slower. According to the x-ray structure (11), both Trp-3 and Trp-36 face the

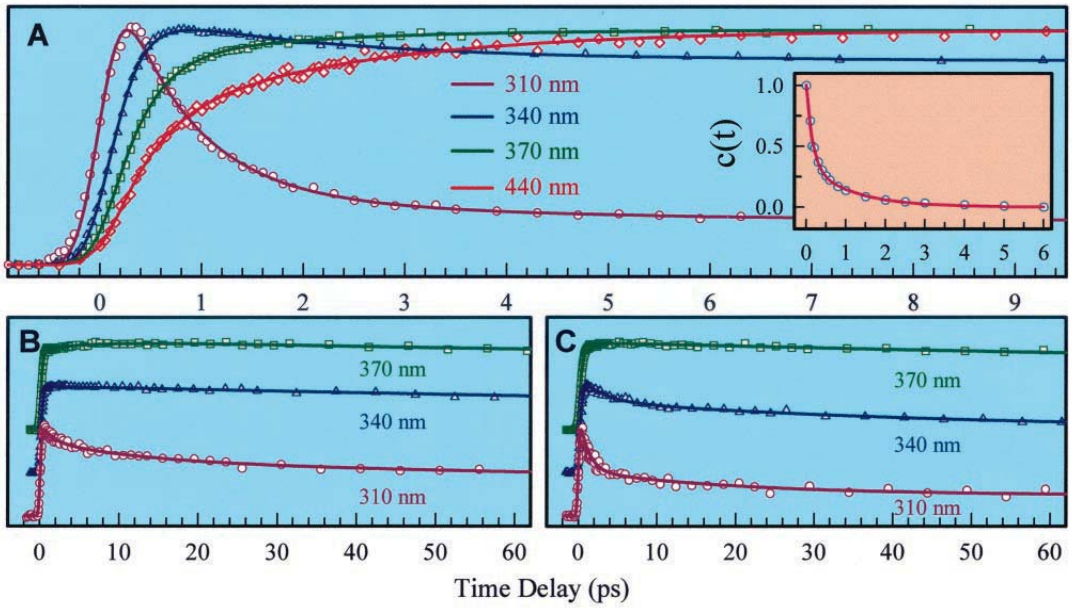


Fig. 3. (*A*) Normalized, femtosecond-resolved fluorescence decay of Trp in buffer solution at pH 2 with a series of wavelength detection. (*Inset*) The constructed solvent response c(t). Normalized fluorescence decay of Trp in the apo-fmetPfRd protein (*B*) and in its denatured state (*C*). Note that the transient at 340-nm emission in *B* decays only with a long lifetime of ≈530 ps.

hydrophobic core, but they are partially exposed to the protein surface and also interact with neighboring polar (Tyr) and negatively charged (Asp and Glu) residues. Thus, the observed multiple time scales represent dynamical motions of both water molecules in the rigid water layer around the protein surface and the interacting residues.

Very recently, Vivian and Callis (24) have carried out extensive theoretical studies of the interactions of Trp with water and other amino acid residues in proteins for interpretation of the fluorescence shifts. The simulation of a single Trp in proteins under the partial exposure to water gave a time scale of several picoseconds for the decay of the shifted spectral components because of the large-amplitude motions of the protein backbone and side chains and/or wholesale rearrangement of nearby hydrogen-bonded water clusters, consistent with our observation of solvation occurring up to 10 ps or longer. By placing an extrinsic dye probe in a protein pocket, solvation dynamics of polar amino acid residues or rigid water molecules on slower time scales also have been recently reported (25, 26). Ultrafast solvation dynamics in protein by using the intrinsic probe, Trp, were not reported before.

Fig. 3*C* shows the dynamics of Trp residues in the denatured fmetPfRd obtained by addition of 6 M GdnHCl into the protein solution at pH 2. At 310 nm, the transient decays in 1.2 ps (56%), 19 ps (22%), and ≈818 ps (22%); at 340 nm, it first rises in 280 fs and then decays 2.83 ps (25%), 40 ps (25%), and ≈1 ns (50%); and at 370 nm, it initially rises in 300 fs (91%), 1.52 ps (9%) and then decays with long components (≥1 ns). These results show a faster solvation process in the denatured fmetPfRd than in its apo form, consistent with the fact that both Trp-3 and Trp-36 are exposed to more water molecules in its denatured state. This is also evident from the emission spectra shifting from 337 nm in the apo form to 352 nm in the denatured form; see Fig. 2. However, the solvation process in the denatured state is much slower than that of Trp in water. Under 6 M GdnHCl, the ratio of water molecules to GdnHCl is ≈5:1. Thus, the observed slow dynamics results from the increased "viscosity" because of the high cationic and anionic concentrations (27) as well as the random-coiled polypeptide chain. But, the polarity of water molecules leads predominantly to the same steady–state emission spectra of Trp both in water and in the denatured protein.

Resonance Energy Transfer. Fig. 4 shows the fluorescence transients of Trps in fmetPfRd for a series of fluorescence detection wavelengths after 288-nm excitation. Four striking results were observed: (*i*) The transient decay time systematically increases, from the blue side to the red side, and the decay because of RET follows a single-exponential behavior at all wavelengths (Fig. 4*A*) (*ii*); all transients decay to zero in 100 ps and no longer components are observed; (*iii*) an initial constant signal within 3 ps is observed at wavelengths longer than 340 nm (Fig. 4*B*); and (*iv*) transients obtained at 265-nm excitation (not shown) show similar temporal behavior. In the transients in which we observed the energy transfer, the effect of solvation is reflected on the shorter time scale: femtosecond rise at the red side and a small picosecond decay component (≈2 ps and 13%) at the blue side (310 nm) as observed in the apoprotein.

Specifically, at 310 nm the transient dominantly decays in 15.6 ps. At 320 nm, the transient follows a single-exponential decay in 16.4 ps. From 340 nm, we start to observe a constant signal in 2–3 ps and then it decays in 20.3 ps at 340 nm and 22.6 ps at 360 nm, respectively. The observed gradual increase of time constants (15–23 ps) toward longer wavelengths is from the influence of picosecond solvation of Trp in the protein. Thus, the decay time of ≈20 ps obtained at 340 nm best represents the dynamics of RET between Trps and the [Fe-S] cluster because we only observed a long component (>500 ps) of the solvated state in apo-fmetPfRd at this wavelength (Fig. 3*B*).

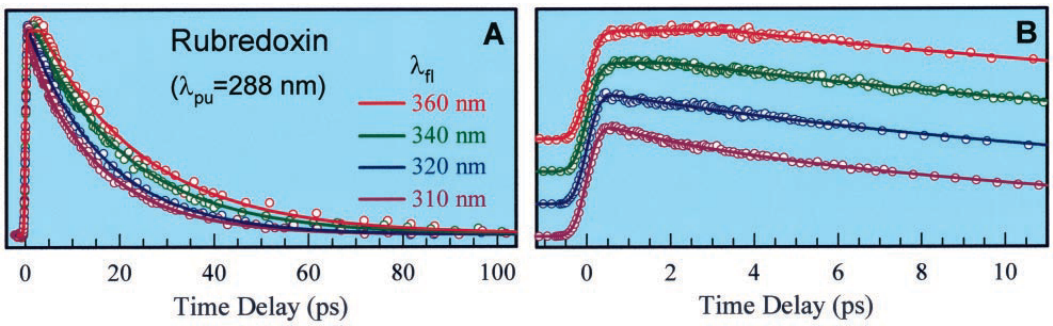


Fig. 4. Normalized, femtosecond-resolved fluorescence decay of Trp in the fmetPfRd protein at 288-nm excitation with a series of wavelength detection at long time scale (*A*) and for the initial part (*B*). Note that at wavelengths longer than 340 nm, a constant signal was observed up to 3 ps.

According to the theory of Förster energy transfer (28), the RET rate from Trp to the [Fe-S] cluster depends on the relative position (r) and orientations of donor (Trp) and acceptor ([Fe-S] cluster), and the rate of transfer k_{RET} can be expressed as follows: $k_{RET} = (R_0/r)^6/\tau_D$, $R_0 = 9.78 \times 10^2(\kappa^2 \cdot n^{-4} \cdot Q_D \cdot J)^{1/6}$. R_0 (in nm), the critical transfer distance, is defined as the donor-acceptor distance at which the transfer efficiency is 50%. κ^2 is the orientation factor, n is the refractive index of the medium (≈1.4), τ_D and Q_D are the donor excited-state lifetime and quantum yield in the absence of the acceptor (here in apoprotein), respectively, and J is the spectral overlap integral (in unit of cm^3/M) between donor-emission and acceptor-absorption.

It is striking that for RET each transient decays to zero with only a single exponential temporal behavior although the fmet-PfRd protein has two Trps (Trp-3 and Trp-36). According to the x-ray structure, the distance between Trp-36 (the middle point of the C–C bridge in the indole ring) and the center of [Fe-S] cluster is ≈9.6 Å and it is ≈13.2 Å for Trp-3. Both Trps in apo-fmetPfRd have a similar lifetime τ_D and quantum yield Q_D, and if they have the same R_0 we should observe two distinct RET rates differing by a factor of 6.8, obviously inconsistent with our result of a single exponential decay. Thus, our observations indicate that either the two Trps have similar energy-transfer rates, but with different R_0 (or κ^2), or one of Trps has $\kappa^2 = 0$; i.e., no energy transfer.

During RET, Trps in the protein do not undergo significant tumbling motions (see below); the anisotropy studies indicate that they are actually rigid. In such a short time of 20 ps, each Trp has a certain value of the orientation factor κ^2. Here, we must also consider Trp–Trp RET. The critical distance for Trp–Trp is in the range of 5–12 Å (29, 30), and the distance between Trp-3 and Trp-36 is ≈10.6 Å. The energy transfer between Trp-3 and Trp-36 takes >100 ps. If one Trp doesn't transfer energy to the [Fe-S] cluster ($\kappa^2 = 0$), but it transfers energy to the other Trp, we would observe a longer component (>100 ps) in the transients, again inconsistent with our observation. Thus, the case for $\kappa^2 = 0$ is excluded.

The observed single-exponential decay in 20 ps indicates that the two Trps have similar RET rates but with different κ^2 values. We further carried out site-directed mutagenesis studies by replacing Trp-3 with tyrosine. At 265-nm excitation, we observed for RET a single exponential decay time of 13 ps for the mutant with only one energy donor Trp-36, and 17.5 ps for fmetPfRd with two energy donors Trp-36 and Trp-3 at 340-nm detection (Fig. 5). Therefore, the decay time for Trp-3 because of RET is estimated to be ≈20 ps by simulations of our transients.

The lifetimes of Trps in apo-fmetPfRd were measured to be 290 ps (53%), 1.44 ns (37%), and 4.03 ns (10%). The average lifetime (τ_D) is ≈1.1 ns. Thus, the deduced average critical distance R_0 for RET between Trp-36 and the [Fe-S] cluster is ≈19.5 Å and ≈25.2 Å for Trp-3. The overall quantum yield (Q_D) was estimated to be 0.15, and the spectral overlap integral was evaluated as 1.2×10^{-14} cm^3/M. We obtained the orientation factor κ^2 to be 0.13 for Trp-36 and 0.62 for Trp-3. The orientation

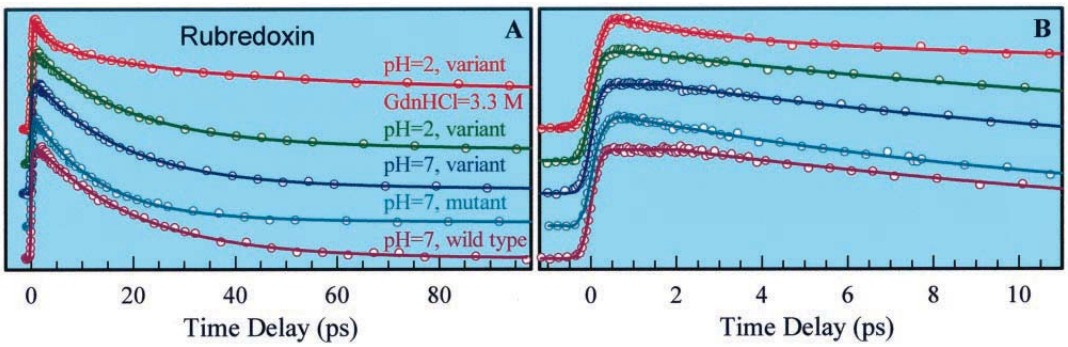


Fig. 5. (*A*) Normalized, femtosecond-resolved fluorescence transients of Trp in various systems and under different conditions for 340-nm detection at 265-nm excitation. The corresponding initial parts are shown in *B*. Note that only the wild-type PfRd and the variant fmetPfRd at pH 7 have an initial constant signal for 2–3 ps; see text for detail.

Table 1. Orientation factors based on the high-resolution x-ray crystallographic structure

	Trp-3–Trp-36				Tyr-12–Trp-36		Tyr-10–Trp-36		Tyr-12–Trp-3		Tyr-10–Trp-3	
	L_a–L_a	L_a–L_b	L_b–L_a	L_b–L_b	L_b–L_a	L_b–L_b	L_b–L_a	L_b–L_b	L_b–L_a	L_b–L_b	L_b–L_a	L_b–L_b
κ^2	0.50	0.22	0.66	0.61	0.35	1.62	0.32	1.23	0.14	0.17	0.25	2.41

factor for Trp-36 former is far away from 0.67, the value for assumption of isotropic orientation distributions for both donor and acceptor molecules.

On the other hand, if we use the isotropic κ^2 value in our case of PfRd, we would obtain a critical distance (R_0) of 25.5 Å for Trp. The resulting distance, *r*, from Trp-36 to the cluster center is deduced to be 12.5 Å. Compared to the x-ray structural distance of 9.6 Å for Trp-36, a 30% error in this case is introduced for use of the isotropic value of 0.67 for κ^2. Thus, care must be taken when the isotropic value is used to calculate the distance in protein. The consequence of the new κ^2 values of Trp for more theoretical studies of electron transfer in the [Fe-S] cluster of the rubredoxin is clear.

The distinct initial flatness for a time of 2–3 ps, observed at wavelengths longer than 340 nm before the decay by RET, may indicate the presence of an early build-up process. One possibility is RET between Tyr and Trp and this process must take place in several picoseconds. In the hydrophobic core of the protein, there are another four aromatic residues, Tyr-10 and Tyr-12, Phe-29, and Phe-48. At 265-nm excitation, the absorption is distributed by ≈75% for Trp, ≈22% for Tyr and ≈3% for Phe, and the distribution becomes ≈90% for Trp and ≈10% for Tyr at 288-nm excitation (31). RET may occur between these aromatic residues, as observed in other proteins (32, 33), especially from Tyr to Trp. The calculated κ^2 values for different energy-transfer pairs based on the x-ray structure are given in Table 1. The estimated RET time constants are in the range of 2–6 ps for Tyr-12–Trp-36 (^1L_b) and 12–27 ps for Tyr-12–Trp-36 (^1L_a) with a separation of 5.9 Å by using a critical distance of 14–17 Å (29, 34) and a lifetime (τ_D) of 1.7 ns for Tyr residues (35). The energy transfer of all other pairs takes much longer time. Further studies can be made by tuning the excitation wavelength to the red side at 295–300 nm to excite Trp only and eliminate the Tyr contribution.

To examine the influence of the structure on the observed rates of RET, we compared the transients for all different forms. Fig. 5 shows fluorescence transients gated at 340-nm emission, under 265-nm excitation, for the different systems and conditions reported here. Except for the partial denatured protein at 3.3 M GdnHCl, all transients show for RET a single exponential decay time: 20 ps for the wild-type PfRd, 13 ps for the mutant, 17.5 ps for the variant (fmetPfRd) at pH 7, and 18 ps for the variant at pH 2. We also observed a long component (>250 ps; 15%) in the case of the variant PfRd under the condition of pH 2, indicating that Trp residues in some protein conformers with unique orientations escape the quenching by RET. This observation shows a more flexible structure at pH 2 than those at pH 7 and in the wild-type form.

PfRd is a hyperthermophilic protein and its melting temperature is as high as ≈200°C (36). It is not fully denatured at pH 7, but it easily unfolds at pH 2 when using denaturants to disrupt all of the salt bridges (37, 38). Experimental results (38) at pH 2 have shown two different intermediates occurring in the unfolding pathway at 2.4 M and 3.3 M GdnHCl. Our studies of fmetPfRd at the latter concentration show a triple-exponential temporal behavior: 2.6 ps (25%), 25 ps (32%) and ≈432 ps (43%); the initial decay reflects solvation processes as observed in the denatured protein (Fig. 3*C*). The decay time of 25 ps is for RET between Trp residues and the [Fe-S] cluster. The last long component is the lifetime of unquenched Trp residues of the denatured protein. Thus, the ratio of the folded to the unfolded species is about 1:1.3.

The initial temporal behaviors are shown in Fig. 5*B*. Only the wild-type and the variant proteins at pH 7 show a constant signal for ≈3 ps. This observation indicates that both the wild-type and the variant (at pH 7) have a more rigid structure and favorable RET between Trp-36 and Tyr-12. On the other hand, the variant at pH 2 and the mutant must have a more flexible structure, consistent with their lower thermostability (11).

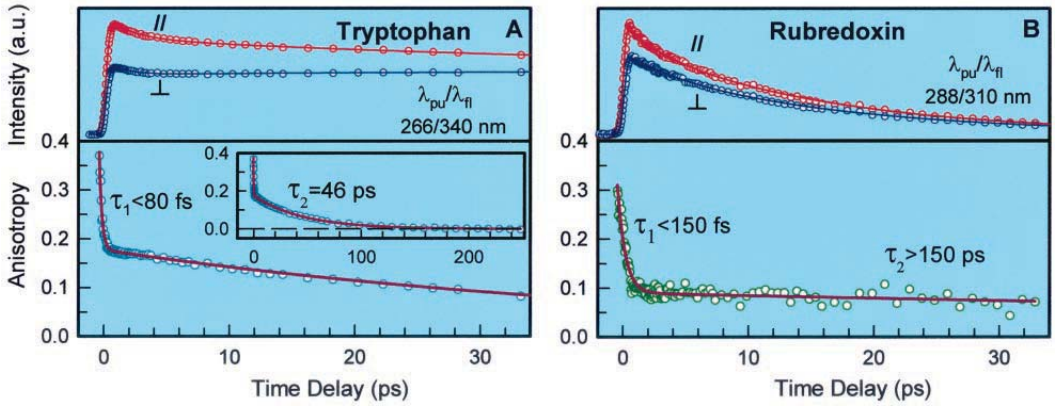


Fig. 6. (*A*) Femtosecond-resolved fluorescence transients of Trp in water at parallel and perpendicular conditions (*Upper*) for 340-nm emission at 265-nm excitation and the corresponding anisotropy decay (*Lower*). (*Inset*) The anisotropy decay for the long time scale. (*B*) Femtosecond-resolved fluorescence transients of Trps in the fmetPfRd protein at parallel and perpendicular conditions (*Upper*) for 310-nm emission at 288-nm excitation and the corresponding anisotropy decay. The anisotropy at 340-nm emission shows an identical temporal behavior (not shown). Note that the anisotropy of Trps in the protein stays constant on the picosecond time scale.

Anisotropy and Internal Conversion. The initial anisotropy of Trp, "r_0", has been extensively studied and discussed (3, 4, 8, 39, 40). In brief, it shows strong dependence on excitation wavelength. Even under subpicosecond resolution, the measured value is still far from the ideal value of 0.4. For example, it is ≈0.2 at 330-nm emission under 295-nm excitation (4). The observed smaller value r_0 is due to ultrafast internal conversion between two nearly degenerate electronic states of ^1L_a and ^1L_b with mutually perpendicular transition dipole moments, and is the average one after the internal mixing. The dependence of r_0 on excitation wavelength is attributed to the different contributions of the two states at different wavelength.

Fig. 6 shows our measured anisotropy of Trp in buffer solution and in the fmetPfRd protein. At 265-nm excitation (Fig. 6*A*), the anisotropy gated at 340 nm promptly drops to 0.19 in <200 fs and then decays with a time constant of 46 ps because of the orientation relaxation. Thus, internal conversion is ultrafast and the time constant is estimated to be <80 fs. Here, we observed high initial anisotropy (≈0.37) at negative time ($t < 0$) because of the broad experimental response function (≈300 fs) (41, 42). The observed apparent r_0 (0.19) here is consistent with the reported value of ≈0.2 (4).

The anisotropy of Trp in the protein at 288-nm excitation shows a similar behavior. It initially drops to 0.1 in <450 fs but then stays at this value during RET. The measured anisotropy of r_0 (0.1) is also close to the reported value of ≈0.12 (4). The deduced time constant for internal conversion is <150 fs. In contrast with Trp in water, the Trp residue is rigid in the protein (43, 44). Thus, during RET from Trp to the [Fe-S] cluster, their orientations are relatively frozen and the orientation factor κ^2 is uniquely determined, consistent with the single-exponential decay observed in RET. This result also reveals a rigid structure of the hydrophobic core in PfRd.

Conclusions

The reported studies with femtosecond time resolution of Trp elucidate its solvation dynamics in different environments and its resonance energy transfer in *Pyrococcus furiosus* rubredoxin. The solvation process of Trp in water was observed to be ultrafast, 160 fs and 1.1 ps, but in the protein it covers a wide range of a much longer time scale. This slow solvation process, which is evident in the time-dependent spectral relaxation, is the origin for nonexponential subnanosecond decays of Trp fluorescence in proteins. The internal conversion between ^1L_a and ^1L_b states occurs in <100 fs, and the frequently reported r_0 value is actually the average anisotropy after the internal state mixing. These results are significant for future studies of local protein structures and dynamics by using Trp as an intrinsic probe. Resonance energy transfer between Trps and the [Fe-S] cluster in the protein was observed to follow a single-exponential temporal behavior on the picosecond time scale. The two Trp residues have similar rates. The critical distance and the corresponding orientation factor for each Trp are uniquely determined. Studies involving measurements of both the population decay and the anisotropy for the wild-type, the mutant and the variant at different pH values reveal a dynamically rigid protein structure. This inflexible structure is probably related to its thermostability (45).

The reported studies indicate that energy transfer occurs on a much faster time scale than the local orientation relaxation ($\tau_{RET} \ll \tau_{orien}$) and that the transfer efficiency is as high as 100%. This finding contrasts the other limit where $\tau_{orien} \ll \tau_{RET}$; in this limit, a mobile energy donor results in multiple energy-transfer rates with relatively low efficiency, contrary to our observation. Solvation in the protein occurs on a similar time scale to that of energy transfer ($\tau_{solv} \approx \tau_{RET}$), resulting in wavelength-dependent transfer rates, as observed in this study. With $\tau_{orien} \gg \tau_{solv} \approx \tau_{RET}$, energy transfer is separated from solvation by femtosecond-resolved fluorescence gating of the relaxed state as observed here. If $\tau_{orien} \approx \tau_{solv} \approx \tau_{RET}$, energy transfer is convoluted with both orientational relaxation and solvation. Thus, the elucidation of the time scales, in this case τ_{orien}, τ_{solv}, and τ_{RET}, is crucial to the understanding of protein dynamics.

Note. In the process of writing this work, we learned of a study of Trp solvation in water (46). The ultrafast solvation in 160 fs was not resolved and the reported ≈1.2 ps solvation time is consistent with our observed long-time component (1.1 ps) reported here.

We like to thank Prof. Michael W. W. Adams and Dr. Francis E. Jenney, Jr. (University of Georgia) for the generous gift of all proteins reported here. We acknowledge the assistance of Dr. Spencer Baskin for the measuring of the lifetime of the apoprotein and for helpful discussion. This work was supported by the National Science Foundation.

1. Szabo, A. G. & Rayner, D. M. (1980) *J. Am. Chem. Soc.* **102,** 554–563.
2. Petrich, J. W., Chang, M. C., McDonald, D. B. & Fleming, G. R. (1983) *J. Am. Chem. Soc.* **105,** 3824–3832.
3. Beechem, J. M. & Brand, L. (1985) *Annu. Rev. Biochem.* **54,** 43–71.
4. Ruggiero, A. J., Todd, D. C. & Fleming, G. R. (1990) *J. Am. Chem. Soc.* **112,** 1003–1014.
5. Callis, R. R. (1997) *Methods Enzymol.* **278,** 113–150.
6. Hochstrasser, R. M. & Negus, D. K. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 4399–4403.
7. Eftink, M. R. (1991) in *Methods of Biochemical Analysis: Protein Structure Determination*, ed. Suelter, C. H. (Wiley, New York), Vol. 35.
8. Hansen, J. E., Rosenthal, S. J. & Fleming, G. R. (1992) *J. Phys. Chem.* **96,** 3034–3040.
9. Lakowicz, J. R. (2000) *Photochem. Photobiol.* **72,** 421–437.
10. Lakowicz, J. R., ed. (2000) *Topics in Fluorescence Spectroscopy: Protein Fluorescence* (Kluwer Academic/Plenum, New York), Vol. 6.
11. Bau, R., Rees, D. C., Kurtz, D. M., Jr., Scott, R. A., Huang, H., Adams, M. W. W. & Eidsness, M. K. (1998) *J. Biol. Inorg. Chem.* **3,** 484–493.
12. Sieker, L. C., Stenkamp, R. E. & Legall, J. (1994) *Methods Enzymol.* **243,** 203–216.
13. Lovenberg, W., ed. (1973) *Iron Sulfur Proteins* (Academic, New York), Vol. 1.
14. Lowery, M. D., Guckert, J. A., Gebhard, M. S. & Solomon, E. I. (1993) *J. Am. Chem. Soc.* **115,** 3012–3013.
15. Dorovska-Taran, V., van Hoek, A., Link, T. A., Visser, A. J. W. G. & Hagen, W. R. (1994) *FEBS Lett.* **348,** 305–310.
16. Fiebig, T., Chachisvilis, M., Manger, M., Zewail, A. H., Douhal, A., Garcia-Ochoa, I. & Ayuso, A. D. H. (1999) *J. Phys. Chem. A* **103,** 7408–7418.
17. Blake, P. R., Park, J. B., Bryant, F. O., Aono, S., Magnuson, J. K., Eccleston, E., Howard, J. B., Sunmers, M. F. & Adams, M. W. W. (1991) *Biochemistry* **30,** 10885–10895.
18. Eidsness, M. K., Richie, K. A., Burden, A. E., Kurtz, D. M., Jr., & Scott, R. A. (1997) *Biochemistry* **36,** 10406–10413.
19. Albinsson, B., Kubista, M., Norden, B. & Thulstrup, E. W. (1989) *J. Phys. Chem.* **93,** 6646–6654.
20. Shen, X. & Knutson, J. R. (2001) *Chem. Phys. Lett.* **339,** 191–196.
21. Reynolds, L., Gardecki, J. A., Frankland, S. J. V., Horng, M. L. & Maroncelli, M. (1996) *J. Phys. Chem.* **100,** 10337–10354.
22. Jarzeba, W., Walker, G. C., Johnson, A. E., Kahlow, M. & Barbara, P. F. (1988) *J. Phys. Chem.* **92,** 7039–7041.
23. Jimenez, R., Fleming, G. R., Kumar, P. V. & Maroncelli, M. (1994) *Nature (London)* **369,** 471–473.
24. Vivian, J. T. & Callis, P. R. (2001) *Biophys. J.* **80,** 2093–2109.
25. Jordanides, X. J., Lang, M. J., Song, X. & Fleming, G. R. (1999) *J. Phys. Chem. B* **102,** 3044–3052.
26. Changenet-Barret, P., Choma, C. T., Gooding, E. F., DeGrado, W. F. & Hochstrasser, R. M. (2000) *J. Phys. Chem. B* **104,** 9322–9329.
27. Nandi, N., Bhattacharyya, K. & Bagchi, B. (2000) *Chem. Rev.* **100,** 2013–2045.
28. Förster, Th. (1965) in *Modern Quantum Chemistry*, ed. Sinanoglu, O. (Academic, New York), Vol. 3, pp. 93–137.
29. Steinberg, I. Z. (1971) *Annu. Rev. Biochem.* **40,** 83–114.
30. Griep, M. A. & McHenry, C. S. (1990) *J. Biol. Chem.* **265,** 20356–20363.
31. Rava, R. P. & Spiro, T. G. (1985) *J. Phys. Chem.* **89,** 1856–1861.
32. Wu, P. G., James, E. & Brand, L. (1993) *Biophys. Chem.* **48,** 123–133.
33. Andrews, D. L. & Demidov, A. A., eds. (1999) *Resonance Energy Transfer* (Wiley, New York).
34. Weber, G. (1960) *Biochem. J.* **75,** 335–345.
35. Ferreira, S. T., Stella, L. & Gratton, E. (1994) *Biophys. J.* **66,** 1185–1196.
36. Hiller, R., Zhou, Z. H., Adams, M. W. W. & Englander, S. W. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 11329–11332.
37. Cavagnero, S., Zhou, Z. H., Adams, M. W. W. & Chan, S. I. (1998) *Biochemistry* **37,** 3377–3385.
38. Zhang, D., Ramakrishnan, V. & Chan, S. I. (1999) *J. Inorg. Biochem.* **74,** 348–348.
39. Valeur, B. & Weber, G. (1977) *Photochem. Photobiol.* **25,** 441–444.
40. Hu, Y. & Fleming, G. R. (1991) *J. Chem. Phys.* **94,** 3857–3866.
41. Myers, A. B., Holt, P. L., Pereira, M. A. & Hochstrasser, R. M. (1986) *Chem. Phys. Lett.* **132,** 585–590.
42. Baskin, J. S. & Zewail, A. H. (1994) *J. Phys. Chem.* **98,** 3337–3351.
43. Munro, I., Pecht, I. & Stryer, L. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 56–60.
44. Ichiye, T. & Karplus, M. (1983) *Biochemistry* **22,** 2884–2893.
45. Lazaridis, T., Lee, I. & Karplus, M. (1997) *Protein Sci.* **6,** 2589–2605.
46. Shen, X. & Knutson, J. R. (2001) *J. Phys. Chem. B* **105,** 6260–6265.