

**Broad specificity of a zinc-dependent small alcohol dehydrogenase from *Thermotoga
maritima* involved in the glycerol dismutation pathway**

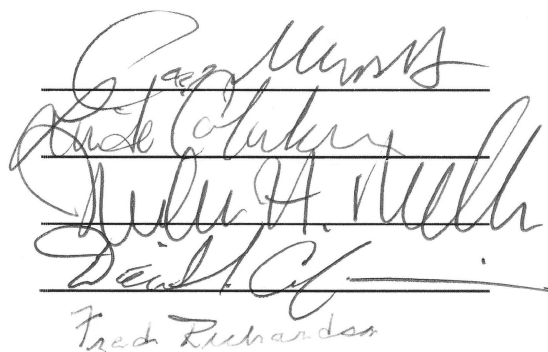
Christopher Ting-Kuang Lee
Burke, VA

B.S., B.S. University of Virginia, 2011

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Master of Science

Department of Chemistry

University of Virginia
May, 2012



The image shows four handwritten signatures stacked vertically on a background of horizontal lines. The signatures are written in dark ink. The first signature is 'Craig M. Smith', the second is 'Rick Collier', the third is 'Julie H. Smith', and the fourth is 'Fred Richardson'.

©Copyright by
Christopher Ting-Kuang Lee
All Rights Reserved
May 2013

This thesis is dedicated to second chances.

Acknowledgements

First I would like to thank my advisors, Drs. Linda Columbus and Cameron Mura for all of their faith and support in my endeavors. It is through their passion for molecular biophysics that has inspired me to choose the path I am on today; Through their constant mentoring that I am able to do a job I love.

I have been fortunate to work with a group of awesome colleagues. I would like to thank Eli Chen, Denise Field, Xiao Huang, Mingie Kang, John Allen, Hillary Bleier, Jeffrey Hatef, and Tomek Kabziński for the intellectual discussion as well as aid in experimental work. In addition, I would like to thank the members of the Mura and Columbus labs who's friendship and support have made this journey possible. In particular, thank you to Carol Price, Brett Kroncke, Dan Fox, Ryan Lo, Ryan Oliver, Ashton Brock, Jennifer Martin, Peter Randolph, and Jen Patterson.

Thank you to the members of my committee, Drs. David Cafiso, Fred Richardson, and John Bushweller for the mentoring I have received. Also special thanks to Professor Michael Shirts who gave me a second chance and has taught me so much. In addition, I would like to thank Lenny Carter, Kelley Midkiff, Dr. Baozhen Xie, and Dr. Audrey Rushin for their advice and counsel through this stage of my life.

Finally, I would like to thank my parents Tsengdar Lee, and Serena Tsai as well as my sister Gloria Lee for their support throughout the years.

Abstract

A putative glycerol dehydrogenase (TM0423) from *Thermotoga maritima* was functionally characterized using bioinformatic and biochemical techniques. Glycerol dehydrogenases (EC 1.1.1.6) catalyze the oxidation of glycerol to dihydroxyacetone (DHA) with the concomitant reduction of NAD^+ to NADH, which can be assayed at 340 nm. Enzymatic activity of TM0423 was verified at 65°C in the presence of Zn^{2+} . The use of other divalent cations, including Ni^{2+} and Co^{2+} , led to reasonable albeit varying levels of enzymatic activity. Exposure to the chelating agent EDTA resulted in complete loss of activity while subsequent reintroduction of metal co-factors to a metal-deficient reaction mixture restored function, suggesting that TM0423 is a divalent metal cation-dependent dehydrogenase. Additional substrates were screened to shed light on potential catalytic mechanisms. The thermostability and enzymatic activity of TM0423 makes it an attractive target for biofuel research.

Contents

1	The Structure Function Paradigm	1
1.1	The Basics of Enzyme Catalysis	1
1.2	Central Dogma, Enzyme Structure	2
1.3	Forces at Play	4
1.3.1	The Hydrogen Bond	5
1.3.2	The Hydrophobic Effect	6
1.3.3	Covalent Bridges	8
1.3.4	Electrostatic Interactions	9
1.4	The Protein Fold and the Sequence Paradox	11
1.5	The Folding Paradox and Energy Landscapes	13
1.6	Effects of Structural Mutations	13
1.7	Research and Dissertation Overview	16
2	Methods for Functional Prediction and Characterization	21
2.1	Sequence Homology and Phylogenetics	21
2.2	Sequence Based Methods	22

2.3	Structure Based Methods	24
2.4	Experimental Characterization	25
2.4.1	Michaelis-Menten Enzyme Kinetics	25
	Deriving the Michaelis-Menten Equation	25
	Limitations to the Michaelis-Menten Model	27
	Analyzing Experimental Data	28
2.4.2	Inhibition Models	30
	Competitive Inhibition	31
	Uncompetitive Inhibition	32
	Mixed Inhibition and Non-competitive Inhibition	33
	Substrate Inhibition	34
3	TM0423 a Putative Glycerol Dehydrogenase	41
3.1	Introduction	41
3.2	Results and Discussion	44
3.2.1	Bioinformatics	44
3.2.2	Protein Activity Dependence on the Presence of Zinc Ion	45
3.2.3	Metal Cofactor Specificity	45
3.2.4	Substrate Specificity	47
3.3	Broader Impacts	49
A	Materials and Methods	58
A.1	General materials, micro-organisms, and plasmids	58

A.2	Protein overexpression and purification	58
A.3	Enzymatic assays	59
A.4	Protein quantification and progress curves	59
A.5	Metal dependence assay	60
A.6	EDTA deactivation assay	60
A.7	Substrate analog assays	60

List of Figures

1.1	Enzyme Catalysis Reaction Profile	2
1.2	The peptide bond	4
1.3	Secondary Structure Elements	6
1.4	Lennard-Jones Potential	8
1.5	Protein Folding Energy Landscapes	12
1.6	The structural basis of phenylketonuria	16
1.7	PDB Growth Over Time	17
2.1	Ideal Michaelis-Menten Plot	29
2.2	Lineweaver-Burk Plot	30
2.3	Hanes-Woolf Plot	31
2.4	Lineweaver-Burk: Competitive inhibition	32
2.5	Lineweaver-Burk: Mixed Inhibition	33
2.6	Lineweaver-Burk: Non-Competitive Inhibition	34
3.1	Metabolic Pathways to Higher Valued Products	42

3.2	Reaction Catalyzed by Glycerol Dehydrogenase	43
3.3	Crystal Structure of TM0423	44
3.4	Structural Alignment of TmGDH and BsGDH	50
3.5	Structural Alignment of TmGDH and BsGDH	51
3.6	Specific Activity of TmGDH with Various Substrates	52
3.7	Alternate model for anaerobic fermentation of glycerol in <i>E. coli</i>	54
3.8	Hanes-Woolf plot of TmGDH with Various Substrates	55

List of Tables

3.1	Table of TmGDH Kinetic Parameters with Varying Metals	46
3.2	Table of TmGDH Kinetic Parameters with Varying Substrate	53

Chapter 1

The Structure Function Paradigm

1.1 The Basics of Enzyme Catalysis

The reaction of orotidine monophosphate (OMP) to uridine monophosphate (UMP) is the last essential step in the biosynthesis of pyrimidine, an important precursor for the synthesis of the nucleobases cytosine, thymine and uracil. If this reaction were to occur slowly, it would severely limit the rate of DNA nucleotide biosynthesis, thereby hindering life as we know it. The half-life ($t_{1/2}$) of the uncatalyzed conversion of OMP to UMP has been determined to be approximately 78 million years¹, well beyond the lifetime of modern organisms. Fortunately enzymes are a solution to this kinetics quandry. An enzyme is a protein which catalyzes a chemical reaction. In vernacular, the enzyme “increases the rate” of a particular reaction. For the OMP decarboxylation reaction, there is an OMP *decarboxylase* which reduces the reaction $t_{1/2}$ to the order of several milliseconds, which is well within the biological timescale of cellular metabolism. According to transition state theory, enzymes thermodynamically stabilize the transition state intermediate. The general energetics and reaction progress are shown in Figure 1.1. The Arrhenius relationship is given by:

$$k = A \exp(-E_a/k_B T) \quad (1.1)$$

where k is the forward rate constant, A is a pre-exponential factor, E_a is the activation energy, k_B is the Boltzmann constant and T is the absolute temperature. This equation suggests that there are two ways to increase the rate of a reaction, by reducing the activation energy (E_a), or increasing the temperature (T). Enzymes are capable of stabilizing the transition state intermediate (lowering E_a) by having specific molecular interactions with the particular ligand (i.e., specificity). Over many generations, evolution has modified the 3D structure and dynamics of enzymes to support specific interactions with a substrate.

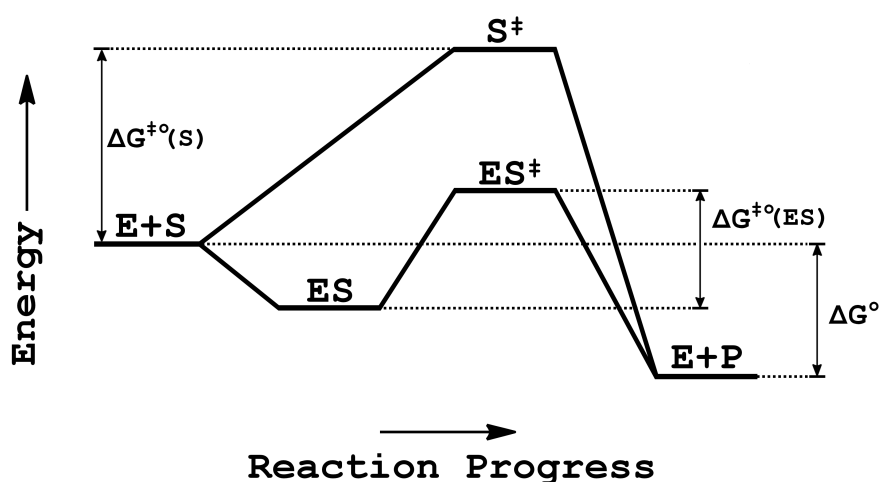


Figure 1.1: An example reaction profile for a reaction as well as its enzyme catalyzed counterpart. E represents the enzyme, S the substrate, ES the enzyme substrate complex, S^{\ddagger} the transition state of the substrate, ES^{\ddagger} the transition state of the enzyme substrate complex, and P the product. The activation free energies of both the catalyzed and uncatalyzed reactions are shown as $\Delta G^{\ddagger\circ}$. The activation energy for the uncatalyzed reaction is higher than that of the catalyzed reaction. The Arrhenius relationship (Equation 1.1), relates the activation energy to the overall rate of reaction. As a result, the catalyzed reaction will occur faster. Note that while the energy of the transition state is lowered, the energies of the substrate and product do not change; the enzyme does not change the state of equilibria, but rather only the kinetics to reach equilibrium.

1.2 Central Dogma, Enzyme Structure

In 1927 Nikolai Koltsov first proposed that genetic traits may be inherited via a “giant hereditary molecule”². He described a giant double stranded molecule where, during cell division, each strand serves as a template for the synthesis of an exact replica of itself. In 1952 Alfred

Hershey and Martha Chase determined that DNA is the genetic material of the T2 phage³. With the establishment of the fundamentals of modern genetics, the question now becomes: how is the sequence information stored in DNA transferred to molecules with biological functions? In 1958 Francis Crick first articulated the “central dogma” of molecular biology. This dogma refers to the newly developed framework, the fundamental basis for our understanding of how sequence information is transferred among the different types biopolymers (i.e., DNA, RNA, protein). Put simply, the central dogma states that DNA is transcribed to RNA, which is then translated to make proteins⁴. This final collection of proteins performs the myriad functions that are necessary to sustain life.

Proteins are composed of amino acids covalently linked through the peptide bond. While DNA can be transcribed to RNA using complementary nucleotide base pairing, it is less straightforward to understand how to translate RNA into a sequence of amino acids. Work done by Nirenberg, Khorana, and Holley, awarded the Nobel Prize in Physiology or Medicine in 1968⁵, helped to elucidate the genetic code, that maps the language of nucleic acids (A, T, G, C) to the language of proteins (20 amino acids). The genetic code is essentially a dictionary of codons (a sequence of three nucleotides), defining which of the 20 natural amino acids correspond to which codons. The “dictionary” function is performed by a complex macromolecule called the ribosome which pairs the codon with the correct amino acid and links them through a peptide condensation reaction⁶. After translation completes, the product is a polypeptide chain. The amino acid sequence of the polypeptide defines its primary sequence.

Although the polypeptide is now chemically synthesized, the protein is still unable to perform any catalytic function. Consider a thought experiment where we suppose that all of the individual amino acids of a serine protease, which has a known catalytic mechanism relying on the *catalytic triad*⁷, are placed in solution, along with a peptide of interest we wish to cleave. Quickly we realize that in this scenario the reaction will likely never occur. This is because the precise positioning and orientation of the catalytic residues are key to protein function.

Proteins typically adopt some secondary and tertiary (3D) structure through a process called protein folding⁸.

1.3 Forces at Play

Aside from the standard chemical covalent bond, which is electronic in nature, for a linear (unbranched) polypeptide there is another major force at play. The amide group of the polypeptide has multiple resonance forms, as shown in Figure 1.2a. This results in partial double bond character for the peptide bonds, which restricts the rotation across the C-N of the peptide bond. Given this restraint, each residue in the polypeptide has two (not three) rotatable bonds, as shown in Figure 1.2b. The dihedral defined by C-N-C_α-C describing the rotation about the N-C_α bond defines the ϕ torsion angle, while the dihedral defined by the atoms N-C_α-C-N describing the rotation about the C_α-C bond defines the ψ torsion angle. Upon folding of the peptide chain into a higher-order 3D structure, these (ϕ , ψ) pairs of torsion angles are not free to adopt any arbitrary value between [0, 360°] due to steric constraints within and between peptide groups.

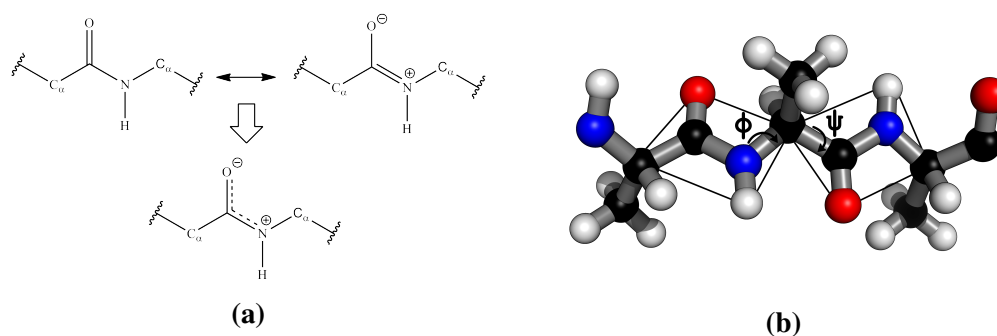


Figure 1.2: (a) **The resonance structures of the peptide bond.** The peptide amide has multiple resonance structures. This resonance causes the peptide bond to exhibit partial double bond character restricting rotational movement. The restriction of rotational movement, results in two freely rotatable bonds per amino acid residue. (b) **Torsion angles for a peptide bond.** The rotation about the N-C_α bond defines the ϕ torsion angle while the rotation about the C_α-C bond defines the ψ torsion angle. Panel (b) was generated and rendered using the PyMOL Molecular Graphics System⁹ showing an alanine tripeptide where carbons are colored black, nitrogen blue, oxygen red, and hydrogen white.

1.3.1 The Hydrogen Bond

One of the non-covalent effects that influence protein structure is the hydrogen bond, describing the attractive interaction between an electronegative atom covalently bonded to a hydrogen which interacts with another electronegative atom. While the proton typically has very little partial charge, the inductive effect of the electronegative atom covalently bonded to an (electropositive) hydrogen atom, which in turn interacts with another electronegative atom. This positive charge can then interact with the partial negative charge of a nearby atom yielding an attractive interaction. This simplistic model describes an interaction that is purely electrostatic; however, experiments using NMR and Compton scattering have shown that hydrogen bonds do contain partial covalent character^{10,11}, although the amount of covalent character predicted by experiments is still unclear¹². Depending on the identity and character of the primary sequence of a polypeptide, the backbone and amino acid side-chains may undergo extensive intramolecular hydrogen bonding; this is part of what causes the polypeptide to adopt local secondary structures.

While there are many variants of secondary structure motifs, the two most fundamental are the α -helix and the β -sheet. The α -helix is a right-handed coiled conformation where for a given amino acid i its backbone carbonyl will form a hydrogen bond with the backbone amino nitrogen of residue $i + 4$. An alanine α -helix is shown with hydrogen bonding interactions in Figure 1.3a. The β -sheet consists of β -strands that associate laterally via hydrogen bonds between every other residue, as shown in Figure 1.3b. Any pair of β -strands can be oriented parallel or anti-parallel, in terms of the N'→C' peptide backbone direction. Multiple strands can come together in this fashion to form a stable secondary structure. While secondary structure is important to the overall stability of the protein, most proteins are still unlikely to be active until these secondary structures coalesce into a fully folded, compact 3D (tertiary) structure.

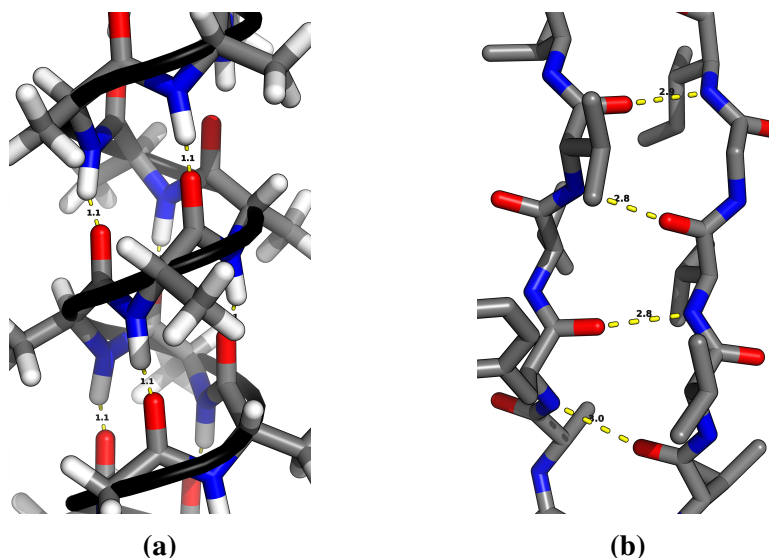


Figure 1.3: (a) A model alanine alpha helix secondary structure motif. In the α -helix we observe that the C=O of residue i is interacting with the N-H of residue $i + 4$. This creates a right-handed helix stabilized by many hydrogen bonds. (b) An example of an antiparallel beta sheet from TM0423 (PDB ID: 1KQ3). The antiparallel hydrogen bonds between the peptide N-H and adjacent strand C=O are shown. Multiple β -strands can aggregate forming a stable secondary structure. Both (a) and (b) were generated in PyMOL⁹ where carbon is shown in grey, hydrogen in white, oxygen in red, nitrogen in blue, and hydrogen bonds shown in dotted yellow.

1.3.2 The Hydrophobic Effect

The hydrophobic effect denotes the observed tendency for non-polar substances to exclude water in solution; essentially, water and oil tend to partition when brought together. While not fully understood, this effect is thought to be driven mostly by entropy as the disruption of dynamic hydrogen bonding between liquid water molecules causes them to form clathrate shells around the non-polar solutes, thus decreasing entropy^{13,14}. This entropy loss is mitigated by aggregation of the apolar moieties, minimizing the surface area-to-volume ratio¹⁵. For polypeptides in solution, this phenomenon has major implications. For soluble proteins, or proteins which are typically found in highly ionic aqueous environments, the hydrophobic residues will tend to cluster together to minimize the solvent accessible surface area in an attempt to maximize solvent entropy¹⁶. This process is termed “hydrophobic collapse”.

During hydrophobic collapse, despite having little to no classical electrostatic interactions, the non-polar residues will still “interact” with each other simply by virtue of being driven together by exclusion of aqueous solvent. As two atoms come together the overlay of their electron clouds will result in some electrostatic repulsion. Furthermore, in terms of quantum and statistical mechanics, the spin-statistics theorem states that for any system of two indistinguishable particles, the wave function of the system will remain unchanged (symmetric) for particles of integer spins (known as bosons) or inverted (antisymmetric) for particles of half-integer spins (known as fermions). For antisymmetric systems, this implies that two indistinguishable particles cannot occupy the same state¹⁷. In practice, since electrons have spin 1/2 they cannot occupy the same quantum state¹⁸. Therefore, as two atoms come together there will be some electrons forced into higher energy states. This transition requires additional energy, which is manifested as a repulsive exchange interaction.

On the other hand, there is also a set of attractive forces collectively termed the van der Waals force¹⁹. The van der Waals force term includes three different components. First the Keesom force describes the electrostatic interactions between permanent dipoles, quadrupoles, and other permanent higher-order multipoles. Second, the Debye force describes the interaction of a permanent dipole and a corresponding induced dipole. Third, the London dispersion force describes the interactions between a pair of transient induced dipoles²⁰. While the van der Waals force is relatively weak, when many such forces are present they can add up to a significant interaction. Overall, the combination of the Heisenberg exchange repulsion and the van der Waals attraction is often modeled using the Lennard-Jones potential²¹ shown in Equation 1.2 where $V(r)$ is the Lennard-Jones potential, ϵ describes the depth of the (attractive) potential well, σ is the distance where the interparticle interaction is 0, and r is the interatomic distance. The r^{-12} term is the repulsive term that reflects the effects of the Pauli repulsion while the r^{-6} term describes the attractive dispersion-based interactions.

$$V(r) = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] \quad (1.2)$$

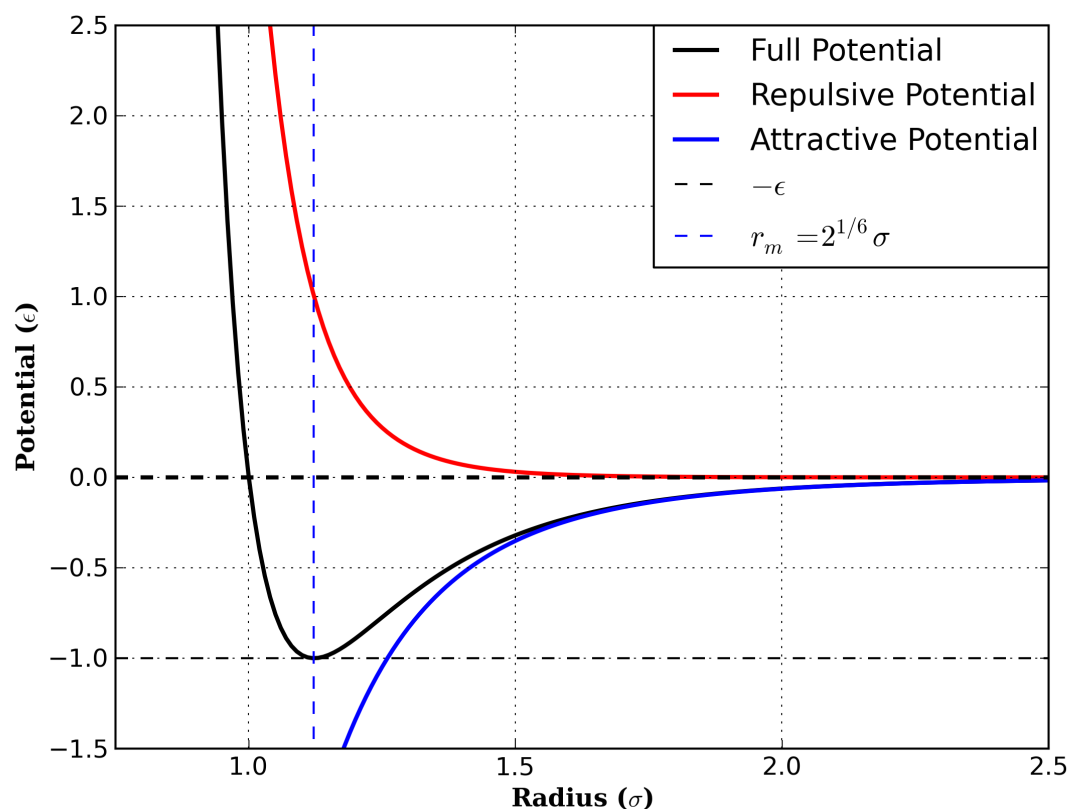


Figure 1.4: A schematic of the canonical Lennard-Jones potential where $\sigma = \epsilon = 1$. The full potential is shown in black while the attractive and repulsive contributions are shown in blue and red respectively. The attractive potential dominates at long distances up until the minima at $2^{1/6}\sigma$ where the repulsive contribution takes over. This potential is a model for Van der Waal's forces as well as the Heisenberg exchange repulsion of atoms.

1.3.3 Covalent Bridges

Another force potentially at play is the formation of covalent interactions. Native structures may contain disulfide bridges, which are covalent bonds linking two cysteine thiol groups. While in most eukaryotic cells the cytosol is a reducing environments and formation of disulfide bridges is not favorable. There exists a sulhydryl oxidase which is capable of facilitating

the formation of disulfide bonds²². These disulfide bonds stabilize proteins in many ways. The disulfide bond may bring together distant portions of the protein. While the entropy of the protein decreases, it is compensated by the formation of additional native contacts increasing enthalpic stabilization*. This may be particularly important during hydrophobic collapse as the disulfide bond constraint may form a nucleus for the condensing hydrophobic core²³. In addition, disulfide bonds may link multiple chains of an oligomeric protein together locking it in its quaternary structure. Another factor which influences quaternary structure is electrostatics.

1.3.4 Electrostatic Interactions

The electrostatic interaction between two charged particles can be described using Coulomb's law, shown in Equation 1.3, where V is the potential energy, ϵ_0 is the permittivity of free space, ϵ_r is the relative permittivity of the local environment to vacuum, q_1 and q_2 are the partial charges of two objects. The Coulombs potential has an inverse (r^{-1}) dependence on the distance and is thus very long-ranged compared to the attractive potential from van der Waals interactions, which vary as a function of r^{-6} . In addition, there is a dependence on the dielectric of the local environment. This has a potentially large effect on the strength of an electrostatic interaction. Suppose there are two residues on the surface of a soluble protein forming a salt bridge. The dielectric of the aqueous environment at 20°C will be about $\epsilon \approx 80$. For comparison, suppose there are two buried residues also forming a salt bridge. The dielectric of this hydrophobic environment will be $\epsilon \approx 2-3$. This means that the salt bridge formed in the hydrophobic environment will be $\approx 30-40\times$ stronger than that formed in the aqueous environment. While the two-body Coulomb's law can be adapted to multi-body systems by assuming linear superposition ("pairwise additivity"), Coulomb's law does not directly account for the heterogeneous dielectric of aqueous ionic solutions, and is a poorer approximation for dynamic systems featuring high charge densities.

*The concept of enthalpy-entropy compensation will be addressed more thoroughly in section 1.5.

$$V(r) = \frac{1}{4\pi\epsilon_0\epsilon_r} \frac{q_1 q_2}{r} \quad (1.3)$$

Continuum models of electrostatic interactions have been developed to help account for the heterogeneous dielectric of biomolecular systems immersed in aqueous solvent or embedded in hydrophobic environments. Classical treatment of electrostatic interactions are based on the Poisson-Boltzmann equation (PBE; Equation 1.4):

$$\vec{\nabla} \cdot \left[\epsilon(\vec{r}) \vec{\nabla} \Psi(\vec{r}) \right] = -4\pi \rho^f(\vec{r}) - 4\pi \sum_i c_i^\infty z_i \lambda(\vec{r}) \exp \left[\frac{-z_i q \Psi(\vec{r})}{k_B T} \right] \quad (1.4)$$

where $\epsilon(\vec{r})$ is the position-dependent dielectric, $\vec{\nabla} \Psi(\vec{r})$ is the gradient of the electrostatic potential (i.e., the electric field), $\rho^f(\vec{r})$ is the charge density of the solute, c_i^∞ is the bulk concentration of ion i , z_i is the valency of ionic species i , q is the elementary charge of a proton, $\lambda(\vec{r})$ is a factor that describes the position-dependent accessibility of position r to the ions in solution (shape function), k_B is the Boltzmann constant, and T is the absolute temperature^{24,25}. Two approximations are commonly employed: under conditions of low ionic strength, the PBE can be linearized to give the Debye-Hückle equation. Under conditions with no mobile ions and a homogeneous dielectric field, equation 1.4 reduces to Coulomb's law. Classical electrostatics can be used to study electrostatic potentials, pH-dependent properties of proteins, and the effects of solvation. While the PBE works well for many biological systems, it breaks down for systems with high charge density. This is because the PBE models the solvent as a dielectric continuum which responds linearly to applied fields. Under this model, strong fields such as those present around nucleic acids and highly charged proteins may result in unrealistically strong polarization of the solvent. Similarly, the PBE corresponds to only a mean field approximation of ionic solutions, meaning that ions experience the average influence of other charges in the system. For areas of high charge density, small fluctuations may be relevant to the dynamics or function of a macromolecule, but is not captured by the PB model.²⁶.

1.4 The Protein Fold and the Sequence Paradox

The final structure of a protein depends on the properties of its primary sequence⁸. But sequence space is astronomically vast: For a 100 residue protein, the number of possible sequences using the 20 natural amino acids consists of $20^{100} \approx 10^{130}$ possible sequences, which is greater than the number of atoms in the known universe! Thus, it must be impossible for proteins to have originated from random sequences. While this statement is partially correct, the probability of randomly choosing a sequence that happens to be the same as OMP decarboxylase (§1.1) is essentially zero. However, does the sequence need to be precisely the same to exhibit similar activity?

As shown before, the function of a protein ultimately depends on the particular fold of a protein, defined by the spatial arrangement of secondary structure elements. The particular fold is what stabilizes the transition state intermediate to reduce the energy barrier. Thus, the entity that is of interest is not so much a particular sequence but rather a 3D fold. While this is true, there is a huge degeneracy in sequence space. A protein can often be mutated significantly with only minor effects on its global fold (though its function may be compromised). In the extreme limit, natural amino acids can be reduced down to two categories: polar and apolar. Within each category, the amino acids are somewhat degenerate, this means that (again conceptually in the extreme limit) hydrophobic amino acids can be exchanged for one another with little effect on the overall fold, and similarly for polar residues. Using this binary alphabet, immediately the astronomical problem has been reduced to $2^{100} \approx 10^{30}$ sequences^{27,28}. Note that in practice, amino acid residues should be grouped into several, not just two, categories based upon their chemical properties.

For soluble proteins, both computer modeling and experiments have shown that the hydrophobic effect largely drives the folding process. For an “average” globular protein, 1/3 of its residues will reside in the hydrophobic core, while the other 2/3 will be polar and on the

surface²⁷. Nature needs only to conserve mostly the residues in the hydrophobic core, further reducing the problem to $2^{100/3} \approx 2^{33} \approx 10^{10}$. While this number still seems impossibly large, nature deals with these kinds of numbers all the time (i.e., a mole is 6.02×10^{23} molecules). For the OMP decarboxylase case study, assuming that in the primordial goo of life was around 100°C , the corresponding $t_{1/2}$ would be reduced to 10 years. Nature could plausibly scan sequence space for a primitive catalyst with modest affinity to drive the reaction at a reasonable rate^{1,29}.

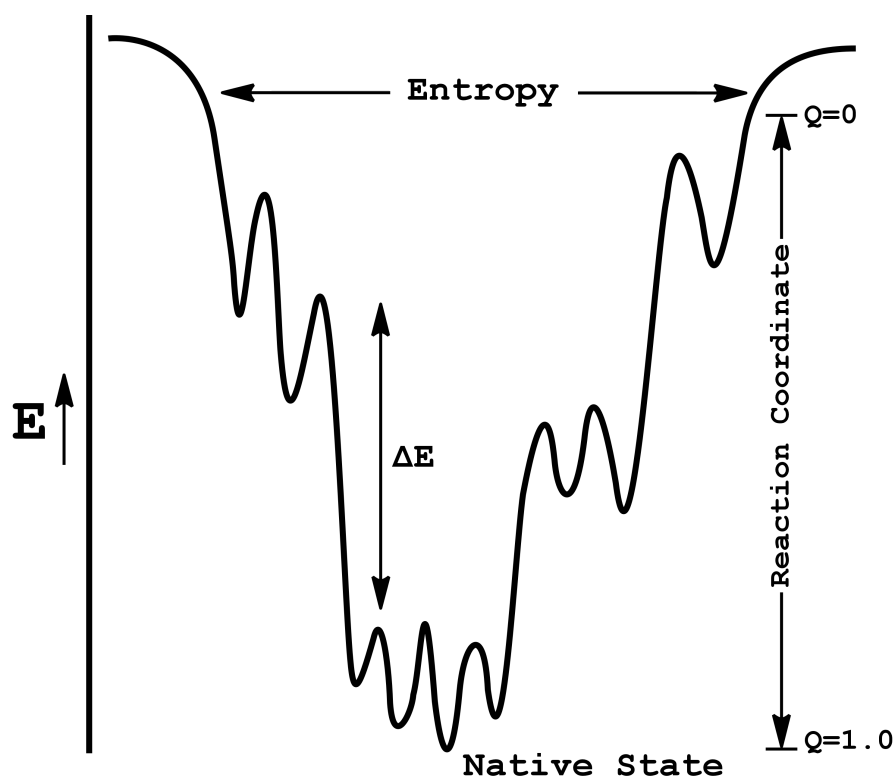


Figure 1.5: A cartoon of a multidimensional protein folding energy landscape. The width of the funnel represents the entropy while the depth of the funnel the enthalpy (or internal energy). The fraction of native contacts is denoted Q . As the protein folds, the entropy decreases but is compensated by the enthalpy of forming native contacts ($Q=1.0$). The protein folding process is not like a random, undirected search through conformational space. Instead, it is guided by the underlying free energy landscape. For funneled landscapes the folding process is akin to rolling a ball down a bumpy hill.

1.5 The Folding Paradox and Energy Landscapes

Another combinatorial explosion, Levinthal's Paradox³⁰, arises when considering how a given protein sequence folds to a particular 3D structure. The folding process cannot simply occur by random search, as the conformational space is far too vast. Returning to our hypothetical 100 residue protein, each peptide bond has 4 preferred $\phi\psi$ angles, corresponding to alpha helix, beta sheet, and two others, which leads to $4^{100} \approx 10^{60}$ conformations.

The resolution to the above "paradox" is the fact that the folding process does not consist of randomly scanning across every degree of freedom. Rather folding to the correct structure is really a process guided by the interatomic forces outlined above. Thus, thermodynamics can help us here^{27,31,32}. If a system has n degrees of freedom $\chi = [\chi_1, \chi_2, \dots, \chi_n]$, the stable states of the system $\chi^* = [\chi_1^*, \chi_2^*, \dots, \chi_n^*]$ correspond to the minima of the free energy function $E(\chi) = E(\chi_1^*, \chi_2^*, \dots, \chi_n^*)$ that mathematically describes the energy landscape of protein folding. If the protein folding energy landscape is plotted, shown in Figure 1.5, the shape of the landscape is funnel like. Suggesting that the process of folding is really guided by the forces and interactions between the elements of the system. Initially, the system will be at high entropy. As the protein begins to fold, the entropy decreases; however, the overall free energy, G° , is compensated by the favorable enthalpy changes accompanying formation of native contacts in a process known as enthalpy/entropy compensation. Due to the enthalpy/entropy compensation the net ΔG_{fold}° is small; Thus, proteins are only marginally stable. The protein folding process is akin to rolling a ball down a bumpy hill rather than randomly rolling a ball on a flat plane (hole-in-one)³¹.

1.6 Effects of Structural Mutations

Charles Darwin first proposed his theory of evolution in 1859, outlining the principle of natural selection³³. Natural selection describes the process in which nature enforces the "survival of

the fittest”. From a biochemical perspective, the phenomenon of life exists due to the molecular machinery of life, the synergy of an overwhelmingly large number of metabolic and catabolic chemical as well as associated feedback and regulation pathways. The analogy between the molecular machinery of life and a complex factory holds true in many regards. The removal of the keystone gear in the factory will result in complete loss of production; in an organism the deactivation of such a protein will result in organism death. Similarly, the addition of an optimized component may boost productivity; while, the enhancement of a protein’s activity may result in improved survival. Over many generations, gradual mutations in various proteins may strengthen or weaken an organism. A weak organism may die before reproduction, thus halting future generations for inheriting its poor genes; this is the basis for evolution defined as survival of the fittest.

In humans, many disorders arising from mutations in a particular protein have been discovered. One such disorder is the autosomal recessive metabolic genetic disorder phenylketonuria (PKU), which is characterized by mutations in the phenylalanine hydroxylase (PheOH) protein rendering it nonfunctional. PheOH normally catalyzes the reaction of phenylalanine to tyrosine with the concomitant hydroxylation of tetrahydrobiopterin (BH₄) all by breaking up molecular oxygen. While there is a plethora of mutations and subsequent disease phenotypes, the following discussion focuses on three specific mutations; F254I, E280K, Δ L364.

Although the catalytic mechanism of wild type PheOH is still not fully understood, some insight has been gained through the use of structural biology as well as kinetics studies. One scheme for the catalytic mechanism involves first the binding of both the phenylalanine as well as the tetrahydrobiopterin in the active site near the nonheme iron. Molecular oxygen will be cleaved hydroxylating the tetrahydrobiopterin and oxidizing the FeII to FeIV. The FeIV will then deposit the additional oxygen onto the phenylalanine³⁴. After some carbocation rearrangement tyrosine is formed.

In wild type PheOH phenylalanine 254 is speculated to π -stack with the pterin ring of

the BH₄. The π -stacking interaction creates specificity for the BH₄ as well as orients it in the proper pose. The structure of PheOH highlighting F254 is shown in Figure 1.6. Mutating Phe254 to isoleucine removes the capability to π -stack likely interfering with the binding of the BH₄ substrate. Due to the poor binding of the required BH₄ cofactor, mutation F254I results in a mild to severe PKU phenotype³⁵. This is an example of an active site residue mutation where the activity of the protein is significantly affected, but the overall fold is only marginally changed. These critical active site residues are typically extremely conserved, as evolutionary mutations often lead to a diseased phenotype and thus natural selection takes its effect.

Another well-studied mutation is E280K which causes mild to severe PKU. In PheOH structures, glutamate 280 forms a hydrogen bond to histidine 146 as well as an important salt bridge to arginine 158 (Figure 1.6). Furthermore, there are two charged glutamic acids in the active-site of PheOH. The substitution of either of these amino acids results in a large perturbation to the electrostatic potential of the active site which affects substrate affinity³⁵. While this mutation is similar to the F254I in that they both affect substrate affinity, E280K also affects the overall fold of PheOH. The salt bridge between E280 and R158 holds the helix and loop in the proper orientation. Disruption of the salt bridge may allow the helix and loop to shift positions. This mutation results in PheOH having only 1% of the wild-type specific activity in biochemical assays.

Finally, an extreme case of a mutation which disrupts the fold is Δ L364. Leucine 364 is positioned between two rigid prolines (362 and 366) and appears to be a necessary spacer between the prolines; Without L364, the improper positioning of the prolines leads to a catastrophic perturbation of the global fold signified by loss of secondary structure in the area³⁵. PheOH with the Δ L364 mutation exhibits no activity in biochemical assays³⁸. Thus, while leucine 364 does not directly participate in the catalytic mechanism of PheOH, its deletion can still have a large effect on the activity of the protein.

One can attempt to generalize that a particular residue is either *structural* (one that is nec-

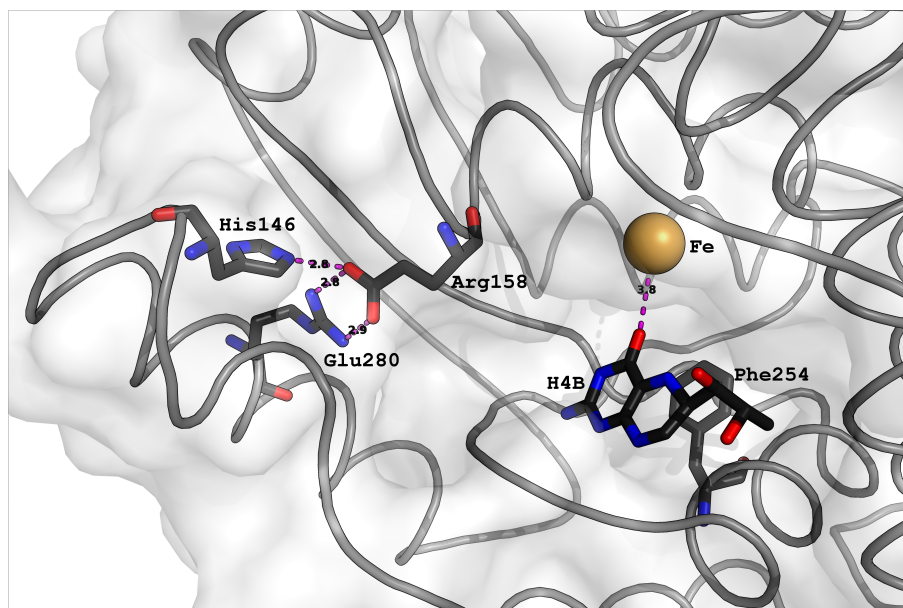


Figure 1.6: The structural basis of phenylketonuria. The structure of phenylalanine hydroxylase is shown with residues H146, E280, R158, and F254 shown with oxygen in red, carbon in black and nitrogen in blue. The observed mutation F254I likely interferes with the proper binding of the pterin and phenylalanine substrates³⁵. While the mutation of E280K causes mild to severe phenylketonuria. E280 is shown to hydrogen-bond to H146 and R158 maintaining the fold. In addition, there are two glutamate residues in the active site, mutations in either results in a large change in electrostatic potential³⁵. This structure has been generated from the phenylalanine hydroxylase structure PDB: 1PAH³⁶. Hydrogen bonding networks were optimized using H++³⁷ and the figure was rendered in PyMOL⁹.

essary to preserve structure), or catalytic (one that directly participates in stabilizing the transition state). However, due to the highly cooperative nature of protein dynamics and energetics, it is difficult to separate and classify individual catalytic contributions into independent additive components³⁹. Thus, many computational and experimental methods have been developed to aid in surveying the energetics of protein catalysis. Several computational and experimental techniques will be addressed in this thesis.

1.7 Research and Dissertation Overview

Over the past decade, NIH-funded Protein Structure Initiative (PSI) centers have developed several high-throughput structural genomics (SG) pipelines. These efforts have generated large numbers of new protein 3D structures, distributed across all three domains of life (Figure 1.7).

Yet, despite these technological achievements, much remains unknown about the natural (biochemical) function and catalytic profiles of many of these now structurally characterized proteins. Chapter 3 focuses on the determination of biological function of TM0423, a putative metal-dependent glycerol dehydrogenase. This protein was previously structurally characterized via x-ray crystallography by JCSG (PDB: 1kq3)⁴⁰. Characterization of the metal dependence as well as the substrate specificity of TM0423 are presented in this thesis. Materials and methods are included in the final appendix.

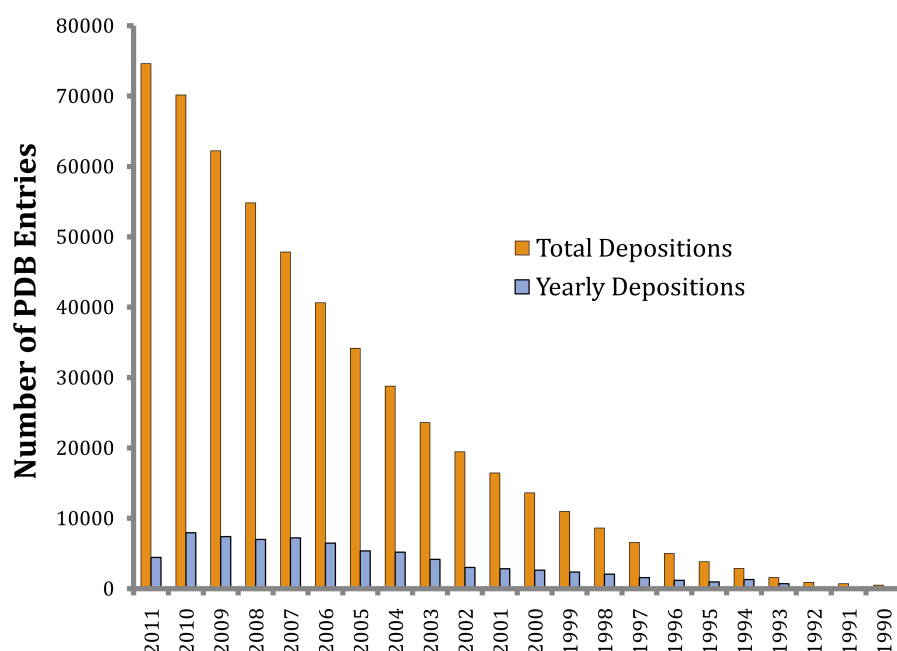


Figure 1.7: PDB growth over time. The number of structures in the protein data bank (PDB) yearly since 1990 is shown. The total number of structures is shown in orange while the number of deposited structures for a particular year is shown in blue. (Data from PDB.org)

Bibliography

- (1) Radzicka, A.; Wolfenden, R. *Science* **Jan. 1995**, 267, 90–93.
- (2) Soyfer, V. N. en *Nature reviews. Genetics* **Sept. 2001**, 2, 723–9.
- (3) Hershey, A. D.; Chase, M. *The Journal of general physiology* **May 1952**, 36, 39–56.
- (4) Crick, F. *Nature* **Aug. 1970**, 227, 561–3.
- (5) Holley, R.; Khorana, H. G.; Nirenberg, M. W. The Nobel Prize in Physiology or Medicine 1968.
- (6) Rodnina, M. V.; Beringer, M.; Wintermeyer, W. *Trends in biochemical sciences* **Jan. 2007**, 32, 20–6.
- (7) Polgár, L. *Cellular and molecular life sciences : CMLS* **Oct. 2005**, 62, 2161–72.
- (8) Anfinsen, C. B. *Science* **July 1973**, 181, 223–230.
- (9) Schrödinger, LLC The PyMOL Molecular Graphics System, Version 1.3r1., Aug. 2010.
- (10) Cordier, F.; Rogowski, M.; Grzesiek, S.; Bax, A. *Journal of magnetic resonance (San Diego, Calif. : 1997)* **Oct. 1999**, 140, 510–2.
- (11) Isaacs, E.; Shukla, A.; Platzman, P.; Hamann, D.; Barbiellini, B.; Tulk, C. *Physical Review Letters* **Jan. 1999**, 82, 600–603.
- (12) Romero, A. H.; Silvestrelli, P. L.; Parrinello, M. en *The Journal of Chemical Physics* **July 2001**, 115, 115.

-
- (13) Pauling, L., *The Nature of the Chemical Bond, An Introduction to Modern Structural Chemistry*, EN, 3rd ed.; Cornell University Press: 1960.
- (14) Tanford, C., *The hydrophobic effect: Formation of micelles and biological membranes*; John Wiley & Sons Inc.: New York, 1973.
- (15) Tanford, C. *Proceedings of the National Academy of Sciences of the United States of America* **Sept. 1979**, 76, 4175–6.
- (16) Kauzmann, W., *Advances in Protein Chemistry Volume 14*; Advances in Protein Chemistry, Vol. 14; Elsevier: 1959.
- (17) Slater, J. *Physical Review* **Nov. 1929**, 34, 1293–1322.
- (18) Thomas, L. H. *Nature* **Apr. 1926**, 117, 514–514.
- (19) Waals, J. D. van der On the continuity of the gaseous and liquid states., Ph.D. Thesis, Leiden University, 1873.
- (20) London, F. *Transactions Of The Faraday Society* **1937**, 33, 8–26.
- (21) Lennard-Jones, J. E. *Proceedings of the Physical Society* **Sept. 1931**, 43, 461–482.
- (22) Hatahet, F.; Nguyen, V. D.; Salo, K. E. H.; Ruddock, L. W. *Microbial cell factories* **Jan. 2010**, 9, 67.
- (23) Zhou, N. E.; Kay, C. M.; Hodges, R. S. *Biochemistry* **Mar. 1993**, 32, 3178–87.
- (24) Fogolari, F.; Brigo, A.; Molinari, H. *Journal of molecular recognition : JMR*, 15, 377–92.
- (25) Honig, B.; Nicholls, A. *Science* **May 1995**, 268, 1144–9.
- (26) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proceedings of the National Academy of Sciences of the United States of America* **Aug. 2001**, 98, 10037–41.
- (27) Dill, K. A. *Protein science : a publication of the Protein Society* **June 1999**, 8, 1166–80.

-
- (28) Bryngelson, J. D. *Proceedings of the National Academy of Sciences* **Nov. 1987**, 84, 7524–7528.
- (29) Eigen, M. *Naturwissenschaften* **Oct. 1971**, 58, 465–523.
- (30) Levinthal, C. *Journal de Chimie Physique et de PhysicoChimie Biologique* **1968**, 65, 44–45.
- (31) Wolynes, P. G. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* **Feb. 2005**, 363, 453–64, 453–64.
- (32) Sali, A.; Shakhnovich, E.; Karplus, M. *Nature* **May 1994**, 369, 248–51.
- (33) Darwin, C., *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*; John Murray: London, 1859.
- (34) Pavon, J. A.; Fitzpatrick, P. F. *Biochemistry* **Sept. 2006**, 45, 11030–7.
- (35) Erlandsen, H.; Stevens, R. C. *Molecular genetics and metabolism* **Oct. 1999**, 68, 103–25.
- (36) Erlandsen, H.; Fusetti, F.; Martinez, A.; Hough, E.; Flatmark, T.; Stevens, R. C. *Nature structural biology* **Dec. 1997**, 4, 995–1000.
- (37) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. *Nucleic acids research* **July 2005**, 33, W368–71.
- (38) Svensson, E.; Döbeln, U. von; Eisensmith, R. C.; Hagenfeldt, L.; Woo, S. L. *European journal of pediatrics* **Feb. 1993**, 152, 132–9.
- (39) Kraut, D. A.; Carroll, K. S.; Herschlag, D. *Annual review of biochemistry* **Jan. 2003**, 72, 517–71.
- (40) Brinen, L. S. et al. *Proteins* **Feb. 2003**, 50, 371–4.

Chapter 2

Methods for Functional Prediction and Characterization

2.1 Sequence Homology and Phylogenetics

Homology forms a basis for comparative biology. Two organisms are said to be homologous if they are descendant from a common ancestor. As with species comparison, homology for genetic or protein sequences is also defined in terms of shared ancestry. Shared ancestry for a DNA sequence can arise from primarily two mechanisms: speciation, or gene duplication. Given some protein P_x from species X , if species X speciates into two species Y , Z both containing mutated P_x : P_y , P_z respectively; Proteins P_y and P_z are said to be *orthologs*, or sequences separated by a speciation event, in this case species X would be the shared ancestor. On the other hand, during a gene duplication event, where a gene is duplicated to occupy two different positions in a genome, the resultant copies are *paralogous*. Another case of homology can result from horizontal gene transfer between multiple species, termed xenologous genes. Approximately 24% of the *T. maritima* genome is thought to be related to an archaeal xenolog¹. Furthermore even among various *Thermogota* strains there is a high degree of lateral gene transfer². By considering the homology of various genes or proteins, it is possible to obtain unique insight into the way genomes evolve.

This study of the evolutionary relationships of genes and proteins is called molecular phylogenetics. Given the branching nature of evolution, the natural visualization technique is the phylogenetic tree. By using a well inferred phylogenetic tree, hidden relationships such as a possible lateral gene transfer event maybe discovered. Typically the clustering of proteins or clades are expected to generally obey trends such as Carl Woese's three-domain system³. Sometimes when exceptions are identified, there may be some significant evolutionary event which occurred. The most popular methods to generate a phylogenetic tree fall under the category of cladistics. These methods are dependent on some mathematical model describing the probability of relationship. For genes and proteins, the most common approach to identifying similarity for phylogenetic study is sequence alignments.

2.2 Sequence Based Methods

During the 1970's both the Sanger⁴ and Maxam-Gilbert⁵ methods for sequencing DNA were published independently. With the invention of these methods, it first became possible to sequence entire genomes of complex organisms. Shortly after in 1987, Applied Biosystems began marketing the model ABI 370, the first automated sequencing machine. Since then, projects such as the Human Genome Project⁶ have driven low-cost and high-throughput technologies. This vast improvement in sequencing technology has allowed for the generation of an unprecedented amount of sequence information to be studied. Various sequence based methods of analyzing this data have been and are being developed to this day.

The important question now becomes: what kinds of hidden information does this data contain, and what may such information reveal about a putative enzyme and its potential catalytic activity and mechanism? As presented in §1.1, the activity of an enzyme is determined by how its 3D structure and dynamics are able to stabilize a transition state intermediate. While just knowing the sequence current methods cannot perfectly predict the 3D structure. However using techniques from the field of informatics, it is possible to extract information statistically

using methods of differing rigor. The most straight simplistic method is the sequence alignment. A sequence alignment is performed by arranging the sequence of DNA, RNA, or protein to identify regions of similarity. Sequences with greater degrees of conservation are expected to exhibit similar biochemical characteristics. While exhaustive brute force calculation of the sequence alignment is tedious and slow, software tools such as BLAST⁷ and Clustal[8] have been optimized using computer science techniques such as dynamic programming. Blast and Clustal can be used to create alignments of hundreds of sequences to create multiple sequence alignments (MSA). By determining degree of amino acid conservation by comparing more sequences, the statistical confidence is increased.

When simple visual comparison of general conservation is insufficient, it is possible to use additional techniques from informatics such as hidden markov models (HMM)^{9,10} to statistically analyze the probabilities of sequence similarity. Analysis by HMM of large databases of sequences allows for the generation of conserved domain fingerprints. Essentially each fingerprint serves as a best guess for some identifying amino acid sequence for enzyme family. Databases of these domain fingerprints such as Pfam¹¹ have been created to assist in identifying conserved domains in sequences of unknown function.

Instead of relying on an informatics model to generate predictions, it is also possible to perform so called *ab initio* protein structure prediction. The basic idea is to use an all-atom model with a full molecular dynamics (MD) force field to simulate the natural folding process^{12–14}. Unfortunately due to the amount of simulation time to survey the vastness of fold space, such a simulation is only possible for the smallest protein systems¹⁵. The amount of computation can be reduced by reducing the number of degrees of freedom. One method to reduce the degrees of freedom is the use of a coarse grain model¹⁶. Additional novel methods such as the use of an online multiplayer game to harness human intellect to predict structures have also been developed¹⁷. Further methods which reduce the amount of fold space through the use of experimental constraints or statistical predictions can be done.

Protein threading or homology modeling can greatly reduce the amount of simulation time by eliminating the initial conformational search. As stated by Anfinsen's dogma, the 3D structure of a protein is dependent on its amino acid sequence¹⁸. Suppose we have a protein of known structure (x) which has a similar sequence to a protein (y) with an unknown structure. Because of the properties outlined in Anfinsen's dogma, we can use protein x as a template to thread the sequence of protein y through to generate a basic model. This model can then be enhanced through the application of different molecular mechanics (MM) force fields to minimize the energy of the predicted structure. The accuracy of homology models is highly dependent on the initial sequence similarity of the template and unknown. If the difference is too vast, the resulting model is likely to contain significant error.

2.3 Structure Based Methods

While sequence comparisons can provide valuable information, ultimately it is the 3D structure and dynamics which determine whether or not the transition state is stabilized or not. As a result similarity in structure is more significant than sequence similarity (§1.4). The basic method of structural comparison is by visual identification. While visually comparing a few proteins it is possible for a normal human however, annotation of an entire library becomes difficult. SCOP a popular visual based categorization relies on the protein knowledge and photographic memory of Alexey Murzin¹⁹.

A basic computational structural alignment algorithm can be performed by simply calculating the average RMSD between matching amino acids of two protein structures. Unfortunately such an algorithm has difficulty accounting for the additivity of structural deviations. For example, if there is a protein family with a large hinge and two domains, in one structure if the hinge is open and in the other the hinge is closed; then the RMSD based comparison will fail at identifying the similarity. Fortunately additional computational structural alignment algorithms have been devised such as DALI²⁰, and VAST²¹ to calculate the probability of structural

similarity taking into account some of the additivity of structural deviations.

While structural similarity can serve as a great guide to prediction of protein function, structural similarity studies are still missing out on the entire dynamics picture. Fundamentally the crystal structures used are at 0K and do not contain any dynamics information. This yields significant problems in predicting actual kinetics where the hinge motion or fold opening can greatly hinder or increase enzymatic turnover.

2.4 Experimental Characterization

2.4.1 Michaelis-Menten Enzyme Kinetics

Individual proteins may play key roles in a metabolic pathway. Information regarding the specific reaction, substrate specificity, turnover, and efficiency of the protein provides one pixel of insight towards what occurs in the mass synergy of chemical compounds that is life. Because many of these parameters are difficult to computationally predict, to characterize the function of a protein of interest, often the protein is isolated and an attempt is made to biophysically reconstitute the reaction *in vitro* to obtain kinetics information. The determined kinetics information can be processed and fit to an enzyme kinetics model to extract the kinetics parameters. The simplest model for enzyme kinetics is the Michaelis-Menten model.

Deriving the Michaelis-Menten Equation

The Michaelis-Menten model relies on two major assumptions: the rapid equilibrium approximation, and the quasi-steady-state approximation. Given the general enzyme substrate reaction scheme:



where E is the free enzyme, S is the free substrate, ES is the enzyme-substrate complex, and P is the product. The rapid equilibrium approximation states that:

$$k_1[E][S] = k_{-1}[ES] \quad (2.2)$$

basically the enzyme and substrate and the enzyme-substrate complex are in rapid equilibrium. This assumption guarantees that the overall reaction is dependent on only one step; thus, the reaction rate can be denoted as:

$$v = k_2[ES] \quad (2.3)$$

where v is the rate of the reaction. Meanwhile, the quasi-steady-state approximation states that:

$$\frac{d[ES]}{dt} = k_1[E][S] - k_{-1}[ES] - k_2[ES] = 0 \quad (2.4)$$

or that the concentration of the enzyme-substrate complex ($[ES]$) is constant with respect to time. A simple rearrangement and defining the Michaelis constant: $K_m = \frac{k_{-1}+k_2}{k_1}$ leads to:

$$[ES] = \frac{k_1[E][S]}{(k_{-1} + k_2)} = \frac{[E][S]}{K_m} \quad (2.5)$$

Because the instantaneous free enzyme concentration ($[E]$) is difficult to measure experimentally, the total enzyme concentration (E_T) which can be calculated as the sum of all terms

containing enzyme: $E_T = [E] + [ES]$, is introduced.

$$[ES] = \frac{(E_T - [ES])[S]}{K_m} \quad (2.6)$$

$$[ES] = \frac{E_T[S] - [ES][S]}{K_m} \quad (2.7)$$

$$K_m[ES] = E_T[S] - [ES][S] \quad (2.8)$$

$$[ES](K_m + [S]) = E_T[S] \quad (2.9)$$

$$[ES] = \frac{E_T[S]}{K_m + [S]} \quad (2.10)$$

Substituting the result for $[ES]$ into the rate equation (2.3) yields:

$$v = \frac{k_2 E_T [S]}{K_m + [S]} \quad (2.11)$$

Substituting in the definition of the maximal rate of reaction $v_{max} \equiv k_2 E_T$ results in the Michaelis-Menten Equation^{22,23}:

$$v = \frac{v_{max} [S]}{K_m + [S]} \quad (2.12)$$

Limitations to the Michaelis-Menten Model

While the Michaelis-Menten model is applicable to many problems, the basic assumptions impose some limitations on the set of systems it models correctly. The rapid equilibrium assumption is based upon the law of mass action²⁴, which only applies in conditions of free diffusion. If the solvent is very viscous then the equilibration rate may be slowed greatly violating this assumption. In cells the cytoplasm is more like a gel than a freely diffusing liquid. Therefore kinetics in the cytoplasm cannot be modeled using the Michaelis-Menten model. On the other hand, the quasi-steady-state assumption requires that $[ES]$ is constant with respect to time. This assumption is only valid if:

$$\frac{E_T}{S_T + K_m} \ll 1 \quad (2.13)$$

where E_T is the total enzyme concentration, S_T is the total substrate concentration, and K_m is the Michaelis constant. The total enzyme concentration must be much less than the substrate concentration. Thus, for weakly binding systems where K_m is small, the Michaelis-Menten model may not apply²³.

Analyzing Experimental Data

When conducting a Michaelis-Menten kinetics experiment the reaction rate v is measured with respect to varying substrate concentration $[S]$. Subsequently a Michaelis-Menten plot can be generated by plotting the $[S]$ against the v , shown in Figure 2.1. To determine the maximal reaction rate v_{max} , and Michaelis constant K_m (corresponding to the substrate concentration resulting in half the maximal rate), non-linear fitting methods are then used to fit the data. However, using non-linear methods was not always possible. In the past, non-linear fitting was computationally inaccessible. As a result, multiple linearized representations of the Michaelis-Menten equation were used to analyze data; These representations include the Lineweaver-Burk²⁵, Hanes-Woolf²⁶, and Eadie-Hofstee²⁷ plots.

The Lineweaver-Burk²⁵ representation, shown in equation 2.14, can be derived by taking the reciprocal of the Michaelis-Menten equation 2.12.

$$\frac{1}{v} = \frac{K_M}{v_{max}} \frac{1}{[S]} + \frac{1}{v_{max}} \quad (2.14)$$

While popular in older works, the Lineweaver-Burk plot is error prone and should not be used to calculate kinetic parameters. Because the axes are hyperbolic, the error structure of the original data is not properly preserved. Essentially the reciprocal of the data yields a non-uniform distribution of data across the plot affecting the accuracy of the linear regression. Aside from statistical errors for this model, there are also experimental constraints. Because the independent variable (x-axis) is $1/[S]$; To obtain small values of $1/[S]$ a large value of $[S]$ must be used. Due to solubility and cost constraints, often the high $[S]$ data cannot be obtained.

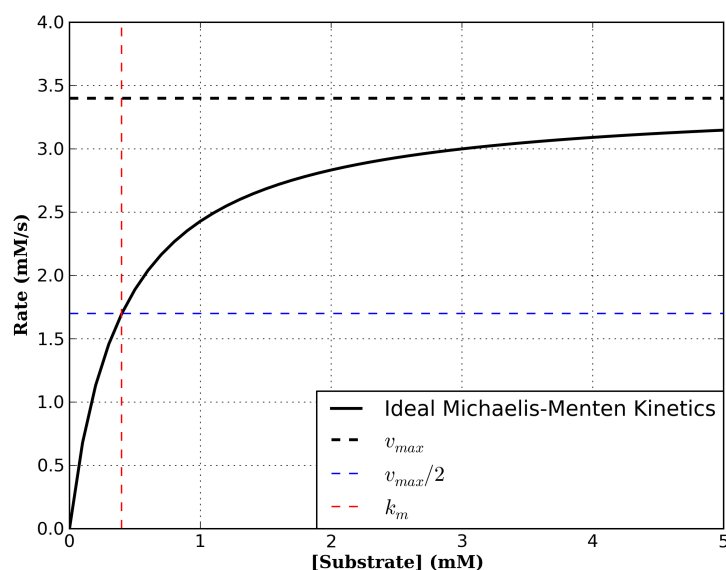


Figure 2.1: A ideal Michaelis-Menten plot. The substrate concentration $[S]$ is plotted against the experimentally determined reaction rate, often measure spectrophotometrically. In this ideal example, the v_{max} and K_m were set to be 3.4mM and 0.4mM respectively. With respect to this representation of the kinetics data, the maximal reaction rate, v_{max} , corresponds to the horizontal asymptote. The Michaelis constant, K_m , can be calculated by finding the substrate concentration corresponding to the half maximal rate $v_{max}/2$. In the past, this style plot was used mostly for visual confirmation of Michaelis-Menten kinetics as non-linear fitting was computationally inaccessible. As a result, many linear representations of the Michaelis-Menten were developed. With the advent of faster computers and algorithms, non-linear fitting has become commonplace and is the more accurate method for calculation of kinetic parameters.

As a result, the extrapolation to obtain the x- and y-intercepts is large and error prone. The Lineweaver-Burk plot's primary utility is for fast visual inspection of kinetics data.

Additional linear representations have also been developed to ameliorate some of the statistical issues faced by the Lineweaver-Burk plot. One such representation is the Hanes-Woolf which is simply the Lineweaver-Burk representation multiplied by $[S]$, as shown in equation 2.15.

$$\frac{[S]}{v} = \frac{1}{v_{max}}[S] + \frac{K_m}{v_{max}} \quad (2.15)$$

The Hanes-Woolf plot (Figure 2.3) does not suffer from the same non-uniform distribution of the dependent variable. The linear regression of the Hanes-Woolf theoretically gives $-K_m$ at

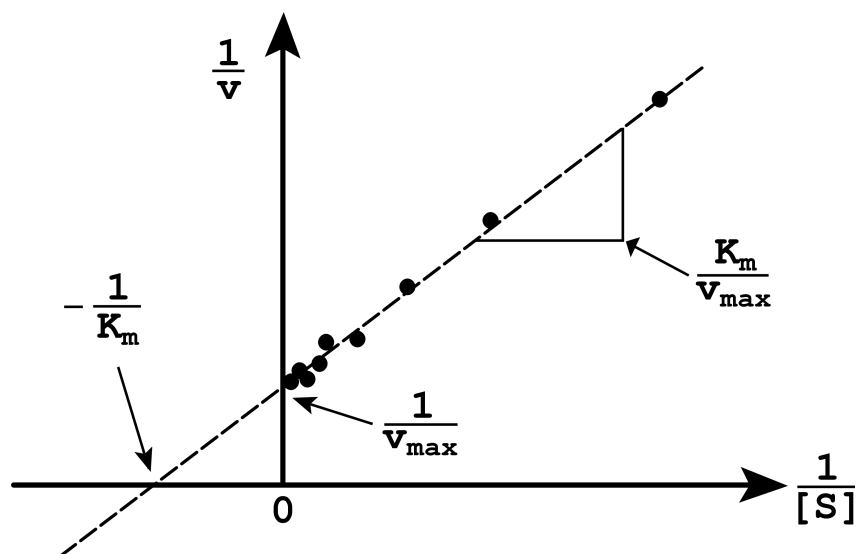


Figure 2.2: A model Lineweaver-Burk plot, where K_m is the Michalis constant, v_{\max} maximal reaction rate, $[S]$ concentration of substrate, and v the reaction rate. The Lineweaver-Burk plot is a linearized representation of the Michaelis-Menten equation (2.12). The y-intercept represents $1/v_{\max}$ while the x-intercept represents $-1/K_m$, and the slope gives the K_m/v_{\max} . While the Lineweaver-Burk plot is generally useful for quick visual inspection of kinetics data, the error model is not preserved. Furthermore, because of the hyperbolic axes, the data is not uniformly distributed which is another source of error.

the x-axis, K_m/v_{\max} at the y-axis, and $1/v_{\max}$ for the slope. Unfortunately, the linear regression of the linear transformation is still not correct because it generates the fit based upon the observed $1/v$ and calculated $1/v_{\max}$ instead of v . Furthermore, neither axis of the representation is an independent variable and thus the correlation coefficient R is also not applicable. The Hanes-Woolf like the Lineweaver-Burk plot was historically used for determination of kinetic parameters but both have since been superseded by nonlinear regression methods. Although, while nonlinear regression methods work well for systems exhibiting Michaelis-Menten kinetics, additional factors affect some systems.

2.4.2 Inhibition Models

Moderation and balance is key to life! In nature, the myriad of complex metabolic pathways must be regulated in some fashion. Without regulation, living organisms could not respond

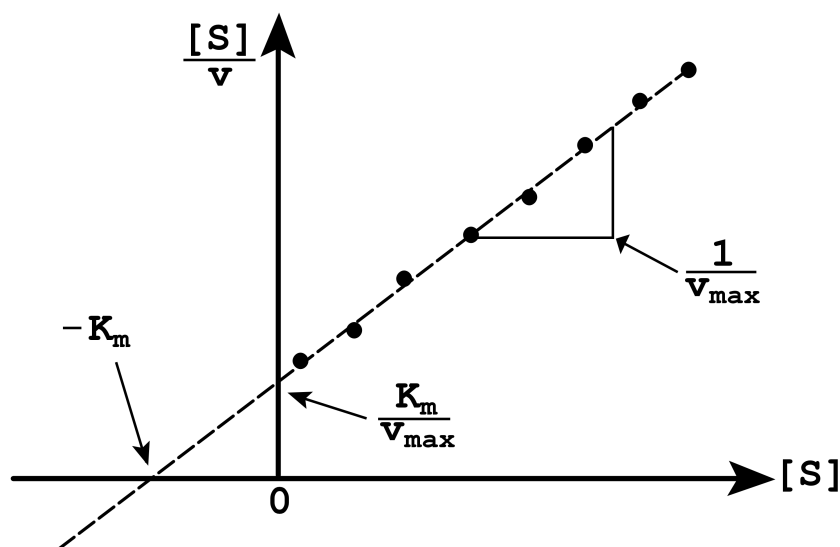


Figure 2.3: A model Hanes-Woolf plot, where K_m is the Michalis constant, v_{max} maximal reaction rate, $[S]$ concentration of substrate, and v the reaction rate. The Hanes-Woolf plot is an improvement upon the Lineweaver-Burk plot, as it does not overemphasize the data obtained at low $[S]$. However, given neither the x- and y-axis represent independent variables, the correlation coefficient R is not applicable. This figure is a modified version of the Hanes-Woolf Plot graphic from Wikipedia.

to stimuli and would die. An enzyme inhibitor is a molecule which interacts with a enzyme and decreases its activity. While enzyme inhibition can occur irreversibly, this discussion with focus on reversible inhibitors.

Reversible inhibitors bind to enzymes using the non-covalent interactions presented in §1.3. Because there is no chemical reaction between a reversible inhibitor and the enzyme, the inhibitor can be removed using dilution or dialysis. There are four mechanisms in which a reversible inhibitor interacts with the enzyme: competitive, non-competitive, mixed, and uncompetitive.

Competitive Inhibition

In competitive inhibition, the substrate and inhibitor cannot occupy the active site at the same time. Competitive inhibitors are often substrate analogs with some specificity for the active site. Because the competitive inhibitor will reduce the effective concentration of free enzyme $[E]$, the apparent K_m for the substrate will increase. At high concentrations of substrate relative

to inhibitor, the substrate can out compete the inhibitor and the reaction will proceed as normal (v is not affected)^{28,29}. The model Lineweaver-Burk plot demonstrating competitive inhibition is shown in Figure 2.4. It is clear that the v_{max} of the reaction is unchanged while the apparent K_m increases with increasing inhibitor concentration.

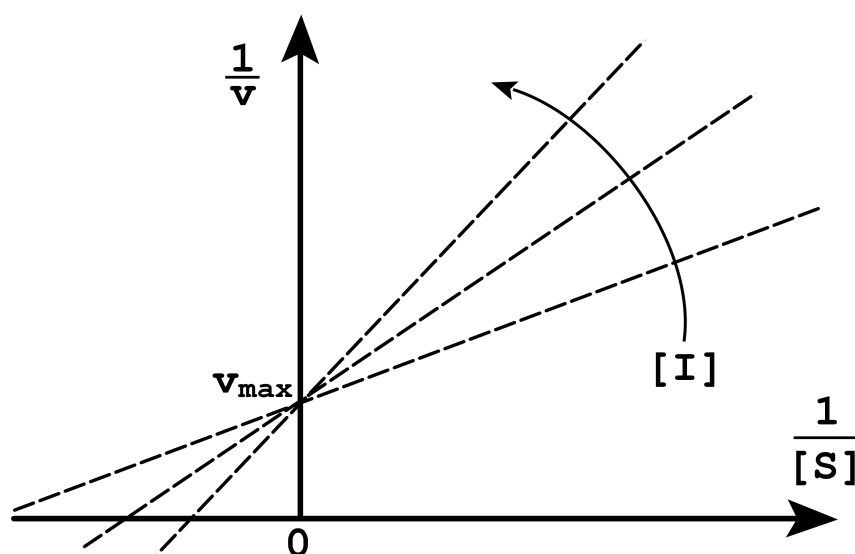


Figure 2.4: Competitive inhibition as portrayed by the Lineweaver-Burk plot. Note that the v_{max} of the overall reaction does not change. This is because only the inhibitor or the substrate can be bound to the enzyme at a given time. Once the substrate is bound, the rate of reaction is not affected. As inhibitor concentration $[I]$ is increased, the apparent K_m given by the x-intercept ($-1/K_m$) decreases.

Uncompetitive Inhibition

An uncompetitive inhibitor binds only to the enzyme substrate complex ($ES + I \rightleftharpoons ESI$). The enzyme-substrate-inhibitor complex is not active and effective activated enzyme-substrate complex $[ES]$ concentration is decreased. When factored into equation 2.3, the rate of reaction v will decrease. Given that the inhibitor is reducing the effective $[ES]$, the K_m will decrease based upon Le Chatelier's principle. Uncompetitive inhibition is not to be confused with non-competitive inhibition which is a form of mixed inhibition!

Mixed Inhibition and Non-competitive Inhibition

A mixed inhibitor differs from an uncompetitive inhibitor in that the inhibitor can bind to an enzyme with or without the substrate bound. While it can bind both E or ES , the affinities for each are different. Often mixed type inhibitors are allosteric inhibitors. The binding of the inhibitor will perturb the conformation of the enzyme leading to varying affinity and turnover of the substrate. Mixed inhibitors can interfere with substrate binding (increasing K_m) and/or hinder catalysis (decrease v_{max}). The model Lineweaver-Burk plot demonstrating mixed inhibition is shown in Figure 2.5.

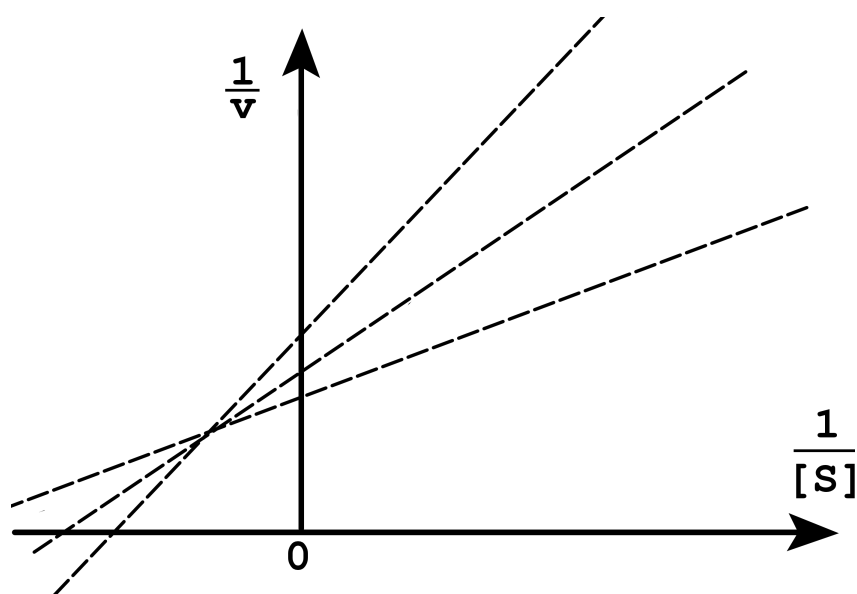


Figure 2.5: Mixed inhibition shown using a Lineweaver-Burk plot. In a mixed type inhibited system, the binding of the inhibitor will typically confer some conformational change in the enzyme. This conformational change may hinder catalysis, shown by a decrease in v_{max} . Furthermore, through the allosteric effect, the affinity for the substrate may also be affected, shown by the increased K_m .

One particular form of mixed inhibition is non-competitive inhibition. In this system, the binding of the inhibitor affects the activity but does not affect the binding of the substrate. As a result, the apparent v_{max} will decrease however the K_m will remain the same. An example of a non-competitively inhibited system is shown using a Lineweaver-Burk plot in Figure 2.6.

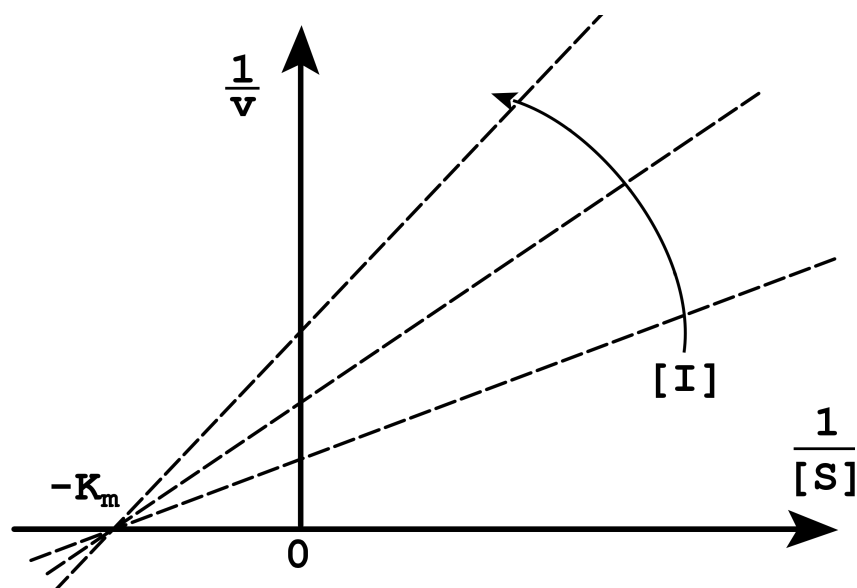


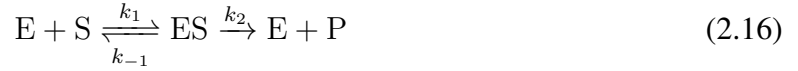
Figure 2.6: Non-competitive inhibition shown using a Lineweaver-Burk plot. In a non-competitively inhibited system, the binding of the inhibitor reduces the activity but does not affect the binding of the substrate. As the inhibitor concentration $[I]$ is increased the apparently v_{max} will decrease. This is reflected in the plot in the y-intercept which shows $1/v_{max}$.

Substrate Inhibition

Substrate inhibition is a special case where there is a progressive decline of activity as high substrate conditions. This type of inhibition was originally thought to be caused by artificially high substrate laboratory conditions, and thus not physiologically relevant. Since the first discovery of substrate inhibition, many systems such as tyrosine hydroxylase, acetylcholinesterase, phosphofructokinase, DNA methyl transferase, among many others have been identified to be substrate inhibited³⁰. Recent estimates state that approximately 20% of all known enzymes are substrate inhibited³¹; Furthermore, that substrate inhibition is a biologically relevant method of regulation. This is evidenced by three primary reasons: First, many enzymes operate under substrate concentrations higher than that of v_{max} . Second, structural studies have identified secondary allosteric substrate binding sites. Third, increasing evidence supports that substrate inhibition plays a key role in regulation of some metabolic pathways³⁰.

There are primarily two simple mechanisms for substrate inhibition. For any substrate

inhibited system, the active enzyme complex ES can further bind another substrate S to form SES . In one model, the SES intermediate can still turnover to form product. This process can be described by the three reactions:



Where E is the free enzyme, S is the free substrate, ES is the enzyme-substrate complex, P is the product, and SES is the inhibited enzyme-substrate complex. A rate equation can be derived for this system:

$$\frac{v}{E_T} = \frac{k_2[S] + k_4 \frac{S^2}{K_i}}{K_a + [S] + \frac{[S]}{k_1 K_i} + \frac{[S]^2}{K_i}} \quad (2.19)$$

Where E_T is the total enzyme concentration, $[S]$ the substrate concentration, K_a is the equilibrium constant $\frac{k_{-1}}{k_1}$, and K_i is the inhibition equilibrium constant constant $\frac{k_{-3}}{k_3}$.

For this thesis, a second model where the SES intermediate is catalytically inactive as proposed by Haldane³². This type of system can be described by the following reaction schemes:



A rate equation can be derived by employing similar assumptions analogous to the Michaelis-Menten derivation.

$$v = k_2[ES] \quad (2.22)$$

The total concentration of enzyme must be conserved.

$$E_0 = E + ES + SES \quad (2.23)$$

The equilibrium approximation becomes slightly more complex to account for the additional ESS complex. We define the following variables K_a and K_i describing the equilibrium constants as follows:

$$K_a = \frac{k_{-1}}{k_1} = \frac{[E][S]}{[ES]} \quad (2.24)$$

$$K_i = \frac{k_{-3}}{k_3} = \frac{[ES][S]}{[SES]} \quad (2.25)$$

Subsequently it follows that:

$$[ES] = \frac{[E][S]}{K_a} \quad (2.26)$$

$$[SES] = \frac{[ES][S]}{K_i} = \frac{[E][S]^2}{K_a K_i} \quad (2.27)$$

The relationship between $[ES]$ and E_T can be established as:

$$\frac{[ES]}{E_T} = \frac{[ES]}{[E] + [ES] + [SES]} \quad (2.28)$$

$$= \frac{\frac{[E][S]}{K_a}}{[E] + \frac{[E][S]}{K_a} + \frac{[E][S]^2}{K_a K_i}} \quad (2.29)$$

$$= \frac{[S]}{K_a + [S] + \frac{[S]^2}{K_i}} \quad (2.30)$$

$$[ES] = \frac{E_T [S]}{K_m + [S] + \frac{[S]^2}{K_i}} \quad (2.31)$$

Substituting back into equation 2.22 and substituting $v_{max} = k_2 E_T$:

$$v = \frac{k_2 E_T [S]}{K_a + [S] + \frac{[S]^2}{K_i}} = \frac{v_{max} [S]}{K_a + [S] + \frac{[S]^2}{K_i}} \quad (2.32)$$

J.B.S. Haldane's rate equation for substrate inhibited kinetics results³². Using l'Hôpital's rule the limit of equation 2.32 is determined to be $\frac{1}{S}$. Thus, as substrate concentration approaches infinite, the reaction rate will drop to zero. Intuitively, this is because as substrate concentration becomes infinite, Le Chatelier's principle will drive all of the enzyme towards the inactive ternary complex SES .

Bibliography

- (1) Nelson, K. E. et al. *Nature* **May 1999**, 399, 323–9.
- (2) Mongodin, E. F.; Hance, I. R.; Deboy, R. T.; Gill, S. R.; Daugherty, S.; Huber, R.; Fraser, C. M.; Stetter, K.; Nelson, K. E. *Journal of bacteriology* **July 2005**, 187, 4935–44.
- (3) Woese, C. R.; Fox, G. E. *Proceedings of the National Academy of Sciences of the United States of America* **Nov. 1977**, 74, 5088–90.
- (4) Sanger, F.; Coulson, A. R. *Journal of molecular biology* **May 1975**, 94, 441–8.
- (5) Maxam, A. M.; Gilbert, W. *Proceedings of the National Academy of Sciences of the United States of America* **Feb. 1977**, 74, 560–4.
- (6) Venter, J. C. et al. *Science* **Feb. 2001**, 291, 1304–51.
- (7) Altschul, S. *Nucleic Acids Research* **Sept. 1997**, 25, 3389–3402.
- (8) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. *Molecular systems biology* **Jan. 2011**, 7, 539.
- (9) Eddy, S. R. *Current Opinion in Structural Biology* **June 1996**, 6, 361–365.
- (10) Eddy, S. R. *Nature biotechnology* **Oct. 2004**, 22, 1315–6.

-
- (11) Finn, R. D.; Mistry, J.; Tate, J.; Coghill, P.; Heger, A.; Pollington, J. E.; Gavin, O. L.; Gunasekaran, P.; Ceric, G.; Forslund, K.; Holm, L.; Sonnhammer, E. L. L.; Eddy, S. R.; Bateman, A. *Nucleic acids research* **Jan. 2010**, 38, D211–22.
- (12) Bonneau, R.; Baker, D. en *Annual review of biophysics and biomolecular structure* **Jan. 2001**, 30, 173–89.
- (13) Piana, S.; Sarkar, K.; Lindorff-Larsen, K.; Guo, M.; Gruebele, M.; Shaw, D. E. *Journal of molecular biology* **Jan. 2011**, 405, 43–8.
- (14) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Proceedings of the National Academy of Sciences of the United States of America* **July 2012**, 1201811109–.
- (15) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *Journal of the American Chemical Society* **Feb. 2010**, 132, 1526–8.
- (16) Tozzini, V. *Current opinion in structural biology* **Apr. 2005**, 15, 144–50.
- (17) Cooper, S.; Khatib, F.; Treuille, A.; Barbero, J.; Lee, J.; Beenen, M.; Leaver-Fay, A.; Baker, D.; Popović, Z.; Players, F. *Nature* **Aug. 2010**, 466, 756–60.
- (18) Anfinsen, C. B. *Science* **July 1973**, 181, 223–230.
- (19) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. *Journal of molecular biology* **Apr. 1995**, 247, 536–40.
- (20) Holm, L.; Rosenström, P. *Nucleic acids research* **July 2010**, 38, W545–9.
- (21) Gibrat, J. F.; Madej, T.; Bryant, S. H. *Current opinion in structural biology* **June 1996**, 6, 377–85.
- (22) Michaelis, L.; Menten, M. L. *Biochem. Z* **1913**, 49.
- (23) Briggs, G. E.; Haldane, J. B. S. *The Biochemical journal* **Jan. 1925**, 19, 338–9.
- (24) Guldberg, C. M.; Waage, P. *Forhandlinger: Videnskabs-Selskabet i Christiania* **1864**, 35.
- (25) Lineweaver, H.; Burk, D. *J. Am. Chem. Soc.* **1934**, 56, 658–666.

-
- (26) Haldane, J. B. S. *Nature* **Apr. 1957**, 179, 832–832.
- (27) Hofstee, B. H. J. *Science* **Sept. 1952**, 116, 329–331.
- (28) Voet, D.; Voet, J. G., *Biochemistry*, 3rd ed.; John Wiley & Sons Inc.: 2004.
- (29) Garrett, R. H.; Grisham, C. M., *Biochemistry*; Brooks Cole: 2012, p 1280.
- (30) Reed, M. C.; Lieb, A.; Nijhout, H. F. *BioEssays : news and reviews in molecular, cellular and developmental biology* **May 2010**, 32, 422–9.
- (31) Chaplin, M. F.; Bucke, C., *Enzyme Technology*; CUP Archive: 1990, p 280.
- (32) Haldane, J. B. S., *Enzymes*; M. I. T. Press: Cambridge, 1965.

Chapter 3

TM0423 a Putative Glycerol Dehydrogenase

3.1 Introduction

Over the past century, issues such as global warming and sustainability of fossil fuels have become societal concerns. As a result, the search for sustainable and environmentally conscious fuels has become increasingly heated. One solution to the fossil fuel problem consists of using biomass to produce biofuels and bioethanol. The traditional approaches for the production of bioethanol involve the fermentation of either starch or cellulose¹. However, as global demand for ethanol fuels increases, other methods for the production of ethanol are becoming necessary.

During the production of biodiesel, fatty acids are reacted with alcohols in a transesterification reaction to yield esters of fatty acids and glycerol. This crude glycerol byproduct can be refined and marketed. However, with the dramatic increase in biodiesel production, the amount of crude glycerol produced has also increased. This surplus crude glycerol has caused glycerol prices to drop to the extent that many now consider glycerol a waste product². Fortunately, through anaerobic fermentative glycolytic pathways, higher value products such as ethanol and butanol can be produced (Figure 3.1). These products represent an opportunity to bring more economic value and sustainability to the biofuel production industry.

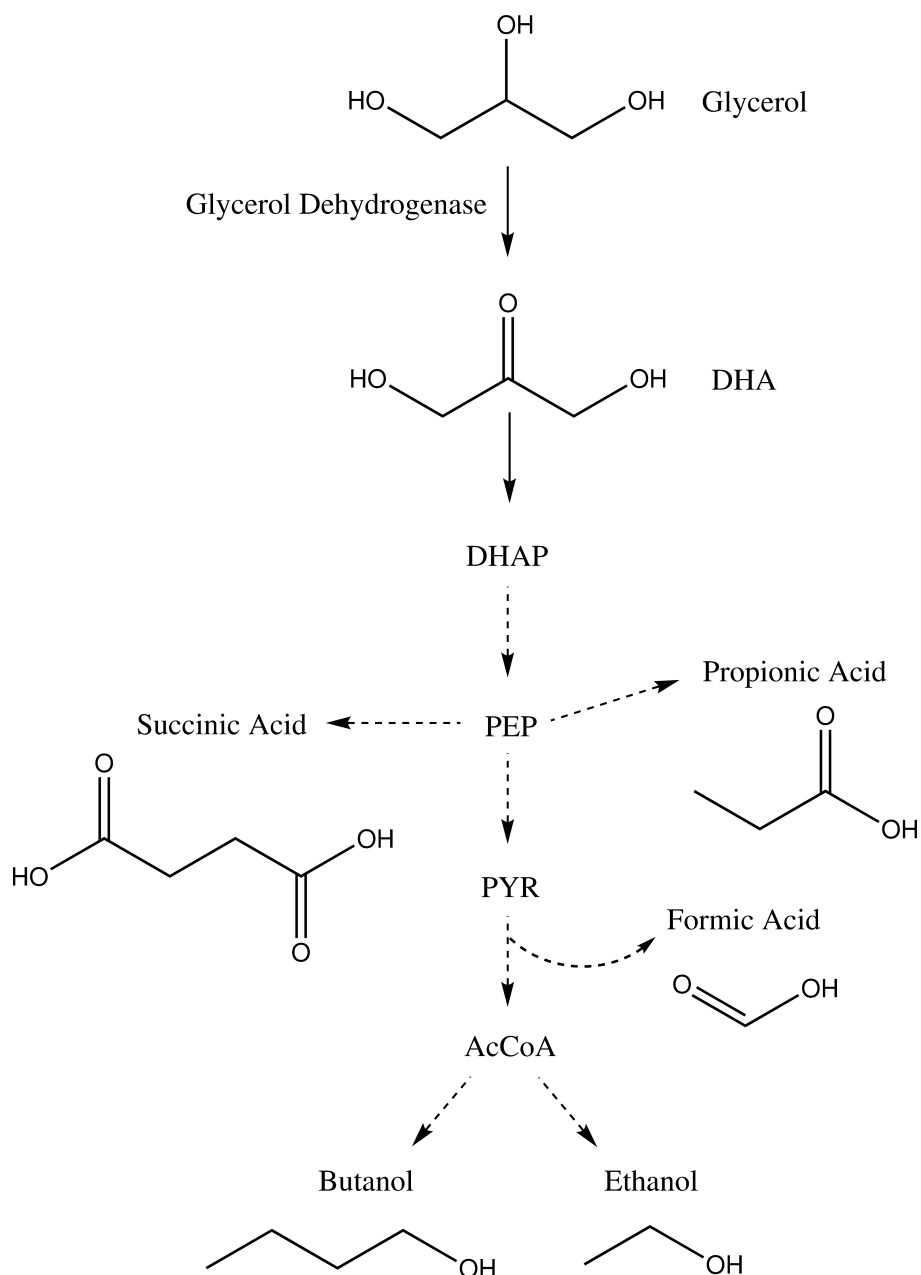


Figure 3.1: Potential metabolic pathways towards the formation of higher value products. The production of biodiesel results in large streams of waste glycerol. It is possible to refine this waste glycerol then through the coupling of many metabolic enzymes, produce higher value products. Broken lines represent pathways composed of several reactions².

The first step of the proposed fermentative glycolytic pathway involves the oxidation of glycerol to dihydroxyacetone (DHA) with the concomitant reduction of NAD^+ to NADH. This reaction is catalyzed by glycerol dehydrogenase (GDH) (Figure 3.2). GDH's are members of

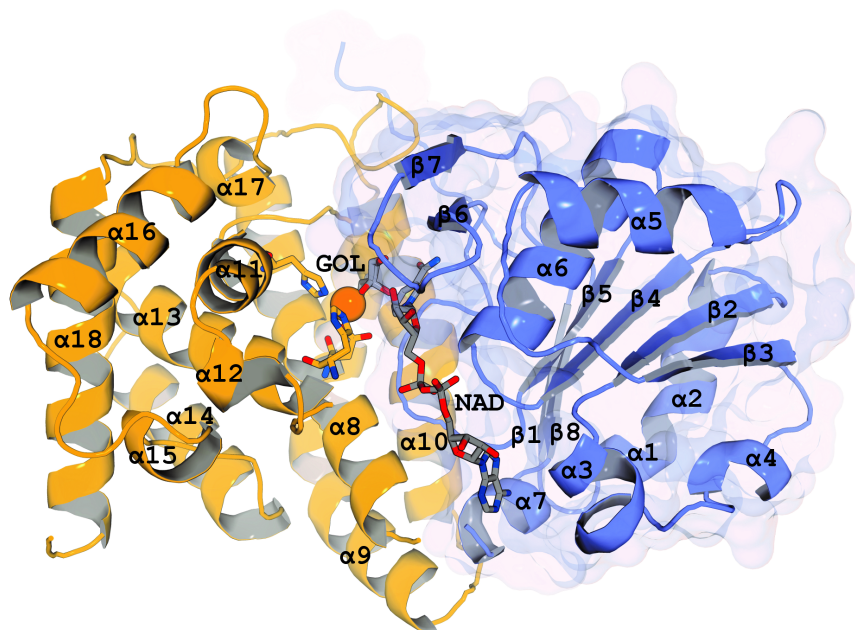


Figure 3.3: The crystal structure of TM0423 (deposited in the Protein Data Bank with accession code 1KQ3). The monomer consists of 18 α -helices ($\alpha 1$ - $\alpha 18$) and 8 β -sheets ($\beta 1$ - $\beta 8$). TM0423 consists of a Rossman fold domain and a multihelical domain, colored in slate blue and bright orange, respectively. The van der Waals surface of the Rossman fold domain is presented as a transparent surface. The ligands NAD⁺ and glycerol (colored in grey with red oxygens and blue nitrogens) are shown in the active site. The zinc cation is shown in orange and the Asp169, His252, and His269 chelating residues are shown.

bioinformatic predictions were then validated experimentally. We show that while TM0423 is a putative glycerol dehydrogenase it exhibits broad substrate specificity and is likely involved in alternate dehydrogenase activity along the glycerol dismutation pathway. The intrinsic hyperthermostability of TM0423 is ideal for industrial applications such as biofuel production.

3.2 Results and Discussion

3.2.1 Bioinformatics

TM0423 is a member of the iron-containing alcohol dehydrogenase (Fe-ADH) family (Pfam: PF00465) with two Fe-ADH Prosite signatures (Prosite motifs PS0060 and PS00913). A protein sequence BLAST against the set of proteins in the PDB revealed high sequence similarity to

another member of the Fe-ADH family, a glycerol dehydrogenase from *B. stearrowthermophilus* (BsGDH). The sequence alignment showed 49% sequence identity with an expectation value of 3×10^{-89} , suggesting distant homology. Additional structure guided searches using the DALI¹⁵ server was performed using the crystal structure of TM0423 (PDB ID 1KQ3). As previously noted BsGDH, exhibited high structural similarity to TM0423. The structural alignment between a BsGDH and TM0423 shows that many of the annotated catalytic residues are spatially conserved between BsGDH and TM0423 (Figure 3.4). Based upon the sequence and structural comparison, TM0423 was proposed to be a glycerol dehydrogenase (TmGDH).

3.2.2 Protein Activity Dependence on the Presence of Zinc Ion

The dependence of protein activity on the presence of Zn^{2+} was examined by adding increasing concentrations of EDTA to reaction cocktails containing a constant ZnCl_2 concentration. Assuming that EDTA is a stronger Zn^{2+} chelator than TmGDH, the increase in EDTA concentration would result in a net reduction of the total amount of Zn^{2+} ions available to TmGDH, ultimately resulting in the complete elimination of Zn^{2+} ions in complex with TmGDH. As EDTA concentration was increased, the specific activity of the protein was drastically reduced (Figure 3.5). At high EDTA concentrations, TmGDH activity was not detected. Similarly, when no zinc was added to the reaction, activity was not observed suggesting that TmGDH is a divalent metal cation dependent glycerol dehydrogenase.

3.2.3 Metal Cofactor Specificity

TmGDH has a catalytic metal chelation site formed by residues His-269, His-252, and Asp-169 as well as by either water or a glycerol- like ligand in an approximately tetrahedral geometry. This forms a high affinity metal binding site with broad metal specificity. To investigate the specificity of the metal required for TmGDH activity, the kinetic assay was conducted using a variety of divalent metals. TM0423 was found to be active with all metals tested (Zn^{2+} , Ni^{2+} ,

Co^{2+} , Fe^{2+} , and Mn^{2+}). Hanes-Woolf plots (Figure 3.5) were used to calculate the apparent K_m , V_{\max} , and catalytic efficiency (k_{cat}/K_m) for Zn^{2+} , Ni^{2+} , and Co^{2+} (Table 3.1). Of these three metals, TM0423 had the highest catalytic efficiency with Zn^{2+} as the metal cofactor ($115.0 \text{ s}^{-1} \text{ M}^{-1}$), with the least activity observed with Co^{2+} .

Metal	V_{\max} ($\mu\text{M/s}$)	K_m (mM)	k_{cat} (s^{-1})	k_{eff} ($\text{s}^{-1} \text{ M}^{-1}$)
Zn^{2+}	1.08 ± 0.02	24 ± 4	2.71 ± 0.06	115 ± 18
Ni^{2+}	1.8 ± 0.2	79 ± 18	4.4 ± 0.5	55 ± 6
Co^{2+}	0.6 ± 0.1	38 ± 19	1.5 ± 0.3	38 ± 9

Table 3.1: TmGDH Kinetic Parameters with Varying Metals

Although Fe^{2+} and Mn^{2+} were also tested, precise determination of the kinetic parameters was hampered by the independent background reaction rate observed with these metals (Figures 3.5c,d). During the assay, a drastic shift in solution color was observed possibly due to metal autoxidation with buffer conditions¹⁶. This color shift resulted in largely distorted kinetics data associated with large error of measurement, which could not be modeled using a standard Michaelis-Menten model; protein activity was observed for both the Fe^{2+} and Mn^{2+} systems, however.

The relative catalytic efficiency observed ($\text{Zn}^{2+} > \text{Ni}^{2+} > \text{Co}^{2+} > \text{Fe}^{2+} > \text{Mn}^{2+}$) is consistent with the Irving-Williams series describing the relative stabilities of complexes formed by a metal ligand complex ($\text{Mn}^{2+} < \text{Fe}^{2+} < \text{Co}^{2+} < \text{Ni}^{2+} < \text{Cu}^{2+} > \text{Zn}^{2+}$). Given that the Irving-Williams series describes general stability of complexes and not reactivity, it is reasonable to suggest that the stability of the metal-protein complex has a large bearing on the overall reactivity of the protein. Furthermore the high enzymatic activity observed with zinc can be attributed to the high Lewis acidity of Zn^{2+} . The zinc ion is stable to oxidative and reductive effects due to its filled d orbital. This 3d orbital also hybridizes with the 4s orbital to increase Lewis acidity, better stabilizing negative charges during catalysis¹⁷. Overall this behavior is consistent

with the predicted catalytic mechanism of the close structural homolog BsGDH. The increased Lewis acidity of the metal allows for increased stabilization of the alkoxide intermediate, allowing for deprotonation of the alcohol substrate by a proton relay mechanism¹⁸.

In addition to its ability to act as a Lewis acid catalyst, the coordination geometry of the zinc ion is important. The filled d orbital of zinc results in a ligand-field stabilization energy of zero in all directions. This allows for varying coordination numbers and thus complex geometries. TmGDH may take advantage of this as the catalytic metal binding residues form a short linker region of 18 residues between His-269, His-252 and a long linker region between His-252 and Asp-169. This long linker region may provide plasticity for the protein to adopt an entatic conformation conferring substrate specificity.

3.2.4 Substrate Specificity

Alcohol dehydrogenases typically exhibit broad substrate specificity or moonlighting functions and are often involved in multiple, related metabolic pathways. The substrate specificity of TmGDH was assayed by substituting glycerol in the standard assay with a variety of molecules structurally similar to glycerol (Figure 3.6). The highest specific activity was observed for 1,2-propanediol, a result previously observed for glycerol dehydrogenases from *Cellulomonas* sp. NT3060, and *E. coli*^{7,19}. Additional substrates, including ethylene glycol, Tris, 2-propanol, 1-propanol, tert-butanol, 1-butanol, bis-tris propane, and 1,3-propanol, were also screened with moderate to negligible activity.

Top substrates for TmGDH all contain a vicinal diol, a shared structural characteristic. The substrate specificity of structural homolog BsGDH has also been studied, yielding a similar conclusion: vicinal diols are key to GDH activity. Given that the ligand field stabilization energy of the zinc ion is zero in all directions, it is possible that the coordination number of the zinc ion can shift to stabilize multiple ligands. In the crystal structure of TmGDH, a Tris buffer molecule was seen interacting with the zinc ion at both the amino and alcohol sites in

the active site²⁰. Similarly, the crystal structure of *B. stearrowthermophilus* shows a glycerol molecule interacting with the zinc ion at both the O1 and O2 sites¹³. This suggests that the substrate orientation in the active site may be stabilized by the zinc ion. Comparably, although Tris and bis-Tris propane both contain vicinal diols, they lack a removable hydride at the usual site of oxidation. This suggests that while Tris and bis-Tris propane likely have some affinity to TmGDH, they cannot be turned over due to mechanistic complications.

Given the previously examined structural similarities to the well characterized glycerol dehydrogenase BsGDH, which exhibits 77% activity with 1,2-propanediol compared to glycerol, we expected that the specific activity for glycerol would be higher than that of other substrates. However, 1,2-propanediol exhibited about 27% more activity than the glycerol assays. To investigate the basis for this, additional kinetic parameters were determined for substrates 1,2-propanediol, glycerol, and ethylene glycol (Table 3.2). 1,2-propanediol appears to have a K_m of 4.3 mM, which is less than that of glycerol (calculated to be 28 mM). It was previously observed for a GDH from *Cellulomonas* sp. NT3060 that bulky functional groups are sterically unfavorable at the third position¹⁹. A similar steric bias may be present in TmGDH, actively selecting for 1,2-propanediol which does not have an alcohol present at the third position.

Recently a new model for the anaerobic fermentation of glycerol was discovered in *E. coli*. In this model methylglyoxal is converted to hydroxyacetone by an aldo-keto reductase. The produced hydroxyacetone is then reduced to 1,2-propanediol by GDH (*gldA*) (Figure 3.7). Given the observed high activity of TmGDH with 1,2-propanediol, in addition to the bioinformatic or previous experimental identification of enzymes catalyzing intermediate reactions, the reverse reaction from hydroxyacetone to 1,2-propanediol with the concomitant oxidation of NADH was also surveyed. Substrate inhibited kinetics was observed for hydroxyacetone activity (Figure 3.8). The experimental data was fit with a basic model for substrate inhibition:

$$v_0 = \frac{v_{max}[S]}{K_a + [S] + \frac{[S]^2}{K_i}} \quad (3.1)$$

where v_0 is the initial rate, V_{max} is the maximum turnover rate, K_A is the affinity constant, K_i is the inhibition constant, and S is the substrate concentration. In addition a plot of the initial rate relative to the $\log([S])$ shows a clear bell-shape describing substrate inhibited kinetics. This result yields little information regarding the product of 1,2-propanediol oxidation for TmGDH. Instead it is reasonable to suggest that the hydroxyacetone intermediate to 1,2-propanediol evolution is not likely to be the catalyzed by TmGDH.

3.3 Broader Impacts

After over a century of metabolome mapping, it is still unknown exactly how all of the proteins in an organismal system interact. Determination of the natural function of these SG proteins will lead to a more comprehensive understanding of the metabolome and how the molecular machinery of cells work. As our understanding increases, we can rely less and less upon differential equation based kinetics models and more on accurate atomistic models of interactions; eventually leading to the simulation of an entire cell! Such simulations can allow us to understand the plethora of effects a minor change on a system component will have on the system whole. This has applications ranging from predicting the secondary effects of drugs to allowing studies on better engineering organisms to be more efficient.

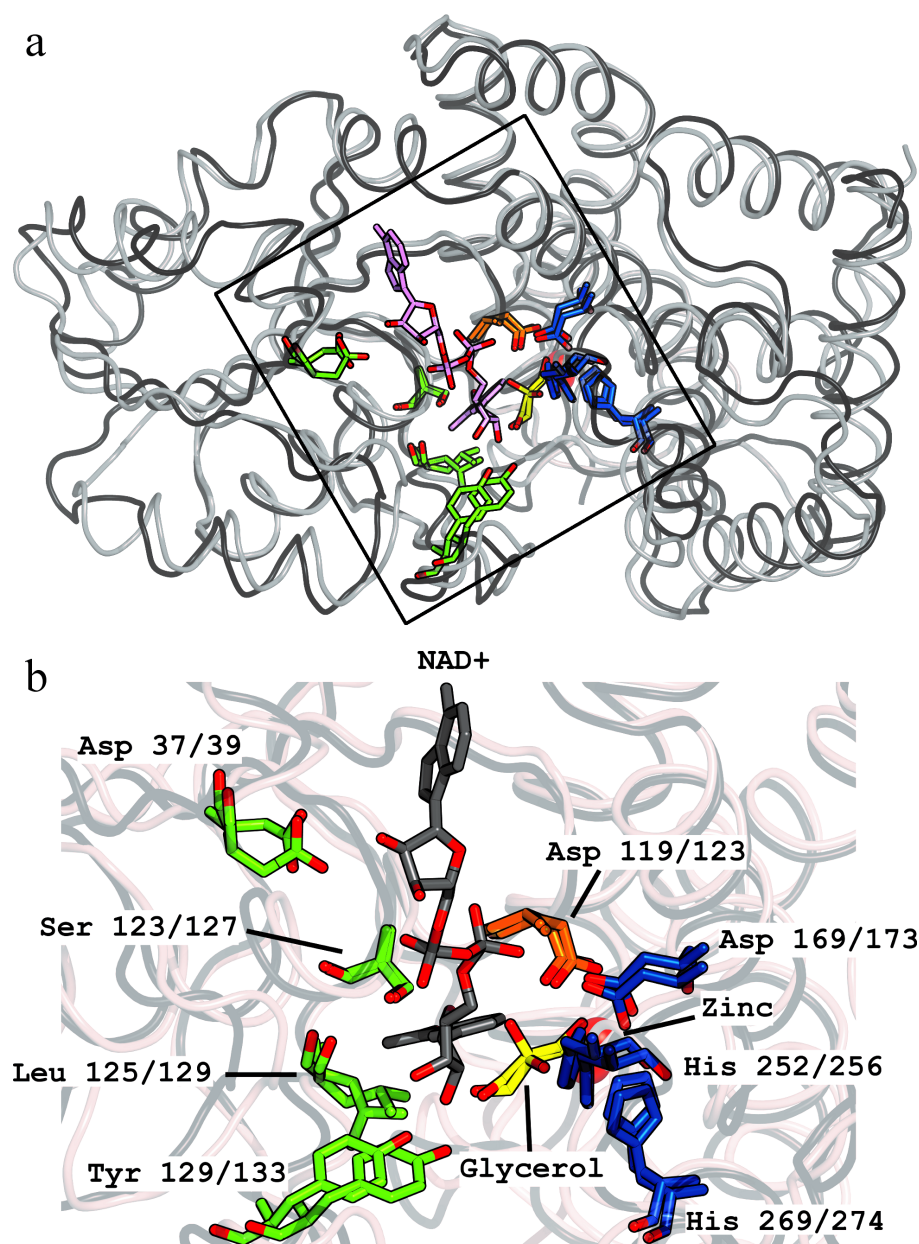


Figure 3.4: (a) Structural alignment of TM0423 and GDH from *Bacillus stearothermophilus* (PDB ID 1JPU, 1JQ5, 1JPA) generated using RMSD based structural alignment provided by Py-MOL. The catalytic residues are shown in color with substrates glycerol and NAD⁺ in yellow and pink, respectively. (b) A zoomed in representation of the active site alignment. Residues colored in light green are NAD⁺ binding; orange are substrate binding, and blue are metal binding. NAD⁺ is shown in grey and substrate glycerol in yellow. Oxygens are highlighted in red. The RMSD of TM0423 to GDH from *B. stearothermophilus* was 0.907Å and the RMSD of the catalytic residues was 0.751Å.

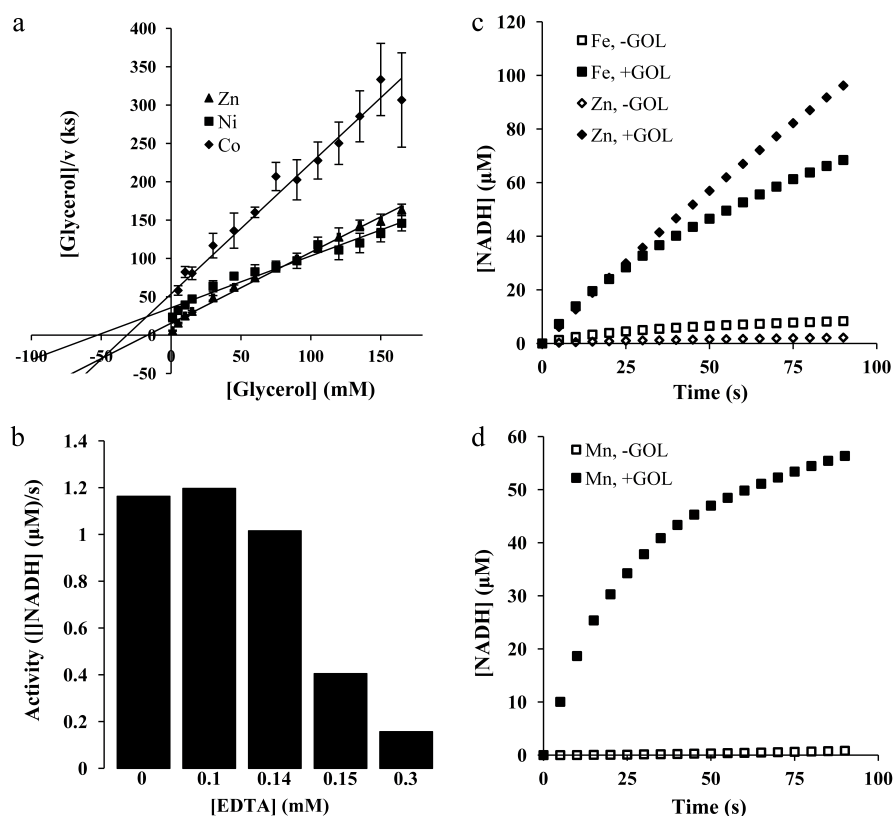


Figure 3.5: (a) Hanes-Woolf representation of TM0423 activity with metals Zn^{2+} , Co^{2+} , and Ni^{2+} . The divalent metal cation dependence of TM0423 was examined at pH 8.0, and $65^{\circ}C$. Extracted kinetic parameters suggest that Zn^{2+} is the primary cation by having the highest catalytic efficiency. (b) Specific activity of TM0423 with glycerol in the presence of varying concentrations of EDTA. As EDTA concentration is increased, there is a sharp inflection point where protein activity drops, thus suggesting that TM0423 is metal dependent. (c) Activity of TM0423 with Fe. The concentration of NADH relative to the time of the reaction is shown. The reaction was run at 400 nM TM0423, 0.3 mM Fe or Zn, 165 mM glycerol, 0.375 mM NAD^{+} , 150 mM NaCl in 50 mM phosphate buffer. Blank controls were run without glycerol. (d) Activity of TM0423 with Mn. The concentration of NADH relative to the time of the reaction is shown. The reaction was run at 400 nM TM0423, 0.3 mM Mn^{2+} , 90 mM glycerol, 0.375 mM NAD^{+} . The blank control was run without glycerol.

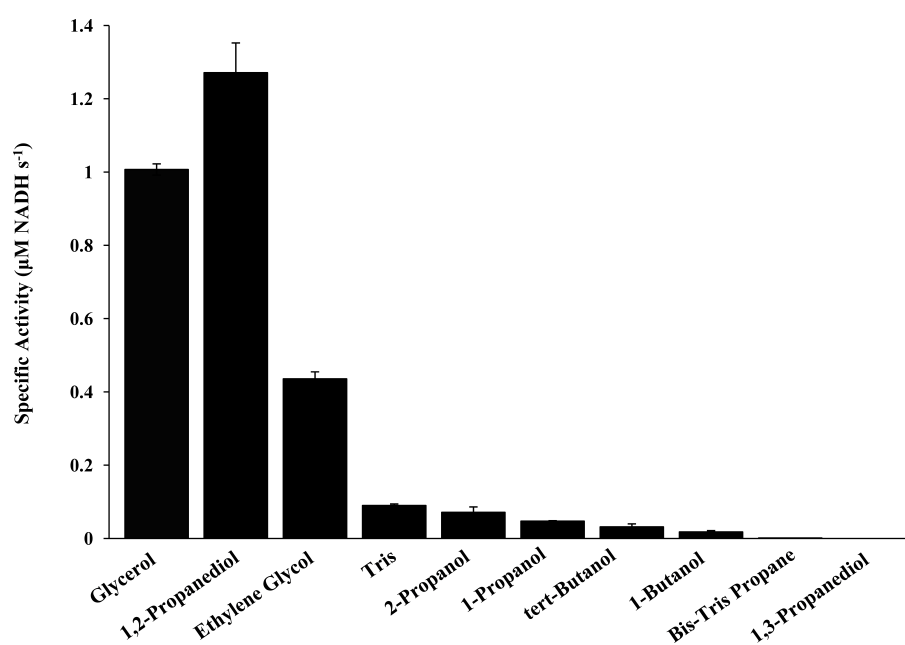


Figure 3.6: Specific activity of TM0423 with multiple substrates with a Zn^{2+} co-factor. TM0423 activity was assessed with alternate substrates, 1,2-propanediol, ethylene glycol, tris, 2-propanol, 1-propanol, tert-butanol, 1-butanol, bis-tris propane, and 1,3-propanediol to investigate the catalytic mechanisms. 1,2-propanediol exhibits the highest specific activity, followed by glycerol then ethylene glycol.

TmGDH				
	V_{\max} ($\mu\text{M/s}$)	K_m (mM)	k_{cat} (s^{-1})	k_{eff} ($\text{s}^{-1}\text{M}^{-1}$)
Glycerol	1.46 ± 0.14	28 ± 10	3.65 ± 0.34	141 ± 44
1,2-propanediol	1.71 ± 0.05	4.3 ± 0.9	4.27 ± 0.13	1044 ± 245
Ethylene Glycol	0.92 ± 0.03	42 ± 4	2.29 ± 0.08	54 ± 4

BsGDH pH 7.4				
	V_{\max} ($\mu\text{M/s}$)	K_m (mM)	k_{cat} (s^{-1})	k_{eff} ($\text{s}^{-1}\text{M}^{-1}$)
Glycerol	-	50	-	-
1,2-propanediol	-	-	-	-
Ethylene Glycol	-	-	-	-

<i>E. coli</i> GldA				
	V_{\max} ($\mu\text{M/s}$)	K_m (mM)	k_{cat} (s^{-1})	k_{eff} ($\text{s}^{-1}\text{M}^{-1}$)
Glycerol	-	56	22.4	400
1,2-propanediol	-	-	-	-
Ethylene Glycol	-	-	-	-

FucO 1,2-propanediol oxidoreductase				
	V_{\max} ($\mu\text{M/s}$)	K_m (mM)	k_{cat} (s^{-1})	k_{eff} ($\text{s}^{-1}\text{M}^{-1}$)
Glycerol	-	-	-	-
1,2-propanediol	-	5.4 ± 0.1	3.8 ± 0.04	710 ± 0.01
Ethylene Glycol	-	51 ± 2	4.0 ± 0.06	80 ± 2

<i>S. pombe</i> GDH2 pH 10				
	V_{\max} ($\mu\text{M/s}$)	K_m (mM)	k_{cat} (s^{-1})	k_{eff} ($\text{s}^{-1}\text{M}^{-1}$)
Glycerol	288	0.5	-	-
1,2-propanediol	288	0.07	-	-
Ethylene Glycol	-	-	-	-

Table 3.2: TmGDH Kinetic Parameters with Varying Substrates

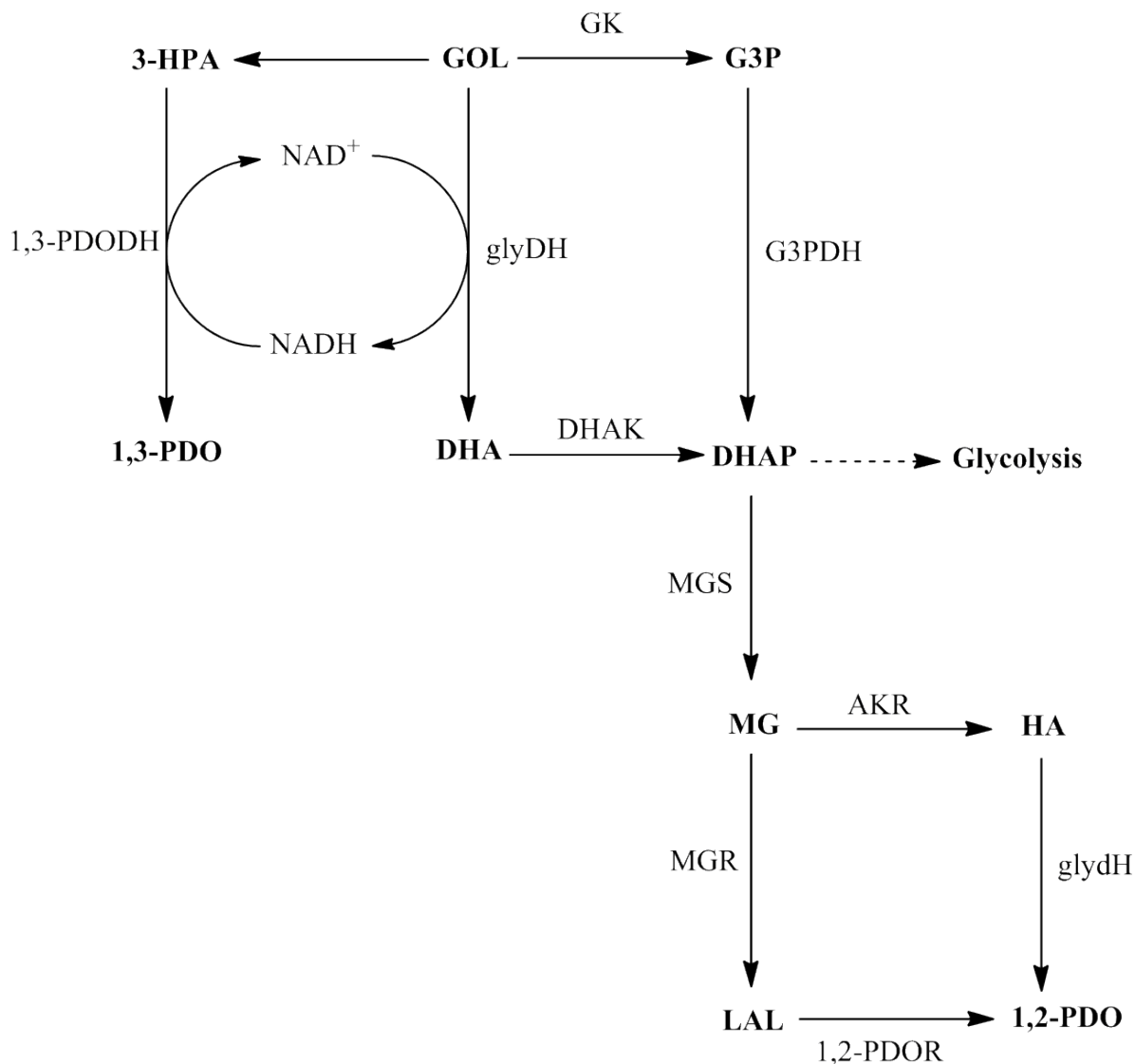


Figure 3.7: Alternate model for anaerobic fermentation of glycerol in *E. coli* GlyDH catalyzes the conversion of glycerol (GOL) to dihydroxyacetone (DHA) with the concomitant reduction of NAD⁺. To redox-balance this reaction, a 1,3-propanediol dehydrogenase (1,3-PDODH) converts 3-hydroxypropanal (3-HPA) to 1,3-propanediol (1,3-PDO). The DHA product can then be phosphorylated by dihydroxyacetone kinase (DHAK) to form dihydroxyacetone phosphate (DHAP). Alternatively GOL can be converted to glycerol-3-phosphate (G3P) by glycerol kinase (GK). Then the G3P to DHAP by a glycerol-3-phosphate dehydrogenase (G3PDH). DHAP can then be used in glycolysis or alternatively converted to methylglyoxal (MG) by a methylglyoxal synthase (MGS). MG has then been observed to either be converted to lactaldehyde (LAL) by a methylglyoxal reductase (MGR) or to hydroxyacetone (HA) by an aldo-keto reductase (AKR) in *E. coli*²¹. Finally either HA or LAL can be catalyzed to 1,2-propanediol (1,2-PDO) a terminal waste product.

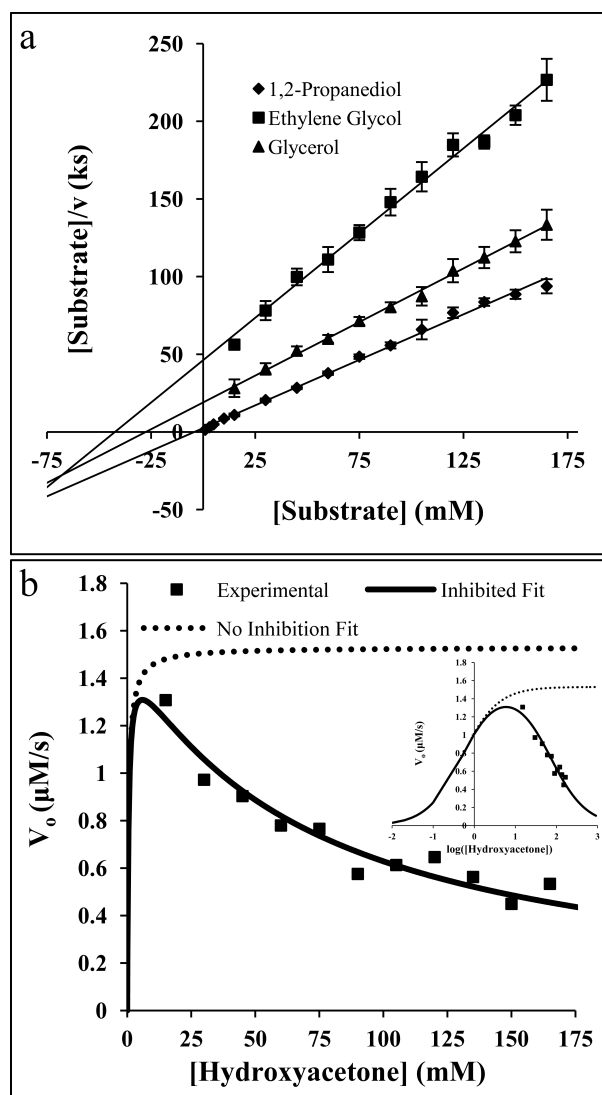


Figure 3.8: (a) Hanes-Woolf plot of TM0423 with substrates glycerol, ethylene glycol, and 1,2-propanediol. Reaction conditions were 400 nM TM0423, 0.3 mM Zn, 1-165 mM substrate, 0.375 mM NAD^+ , 150 mM NaCl in 50 mM phosphate buffer. 1,2-Propanediol was observed to be the best substrate, followed by glycerol and ethylene glycol. (b) Michaelis-Menten plot of the substrate concentration relative to the initial rate of the reaction for TM0423 with hydroxyacetone and NADH as substrates. A substrate inhibited Michaelis-Menten model was used to fit the experimental data (shown as a solid line). The parameters derived from the inhibited model were used to plot the ideal Michaelis-Menten kinetics (shown as a dotted line). The inset is a plot of the logarithmic substrate concentration compared to the initial rate. The observed kinetics fits the bell-shaped curve of the inhibited model shown in the solid line, while the uninhibited form is shown using the dotted line.

Bibliography

- (1) Marchetti, J.; Miguel, V.; Errazu, A. *Renewable and Sustainable Energy Reviews* **Aug. 2007**, *11*, 1300–1311.
- (2) Yazdani, S. S.; Gonzalez, R. *Current opinion in biotechnology* **June 2007**, *18*, 213–9.
- (3) Al-Karadaghi, S.; Cedergren-Zeppezauer, E. S.; Hövmoller, S. *Acta crystallographica. Section D, Biological crystallography* **Nov. 1994**, *50*, 793–807.
- (4) Joernvall, H.; Persson, B.; Krook, M.; Atrian, S.; Gonzalez-Duarte, R.; Jeffery, J.; Ghosh, D. *Biochemistry* **May 1995**, *34*, 6003–6013.
- (5) Niehaus, F.; Bertoldo, C.; Kähler, M.; Antranikian, G. *Applied microbiology and biotechnology* **June 1999**, *51*, 711–29.
- (6) Shams Yazdani, S.; Gonzalez, R. *Metabolic engineering* **Nov. 2008**, *10*, 340–51.
- (7) Tang, C. T.; Ruch, F. E.; Lin, C. C. *Journal of bacteriology* **Oct. 1979**, *140*, 182–7.
- (8) Daniel, R.; Stuert, K.; Gottschalk, G. *Journal of bacteriology* **Aug. 1995**, *177*, 4392–401.
- (9) LIN, E. C.; MAGASANIK, B. *The Journal of biological chemistry* **June 1960**, *235*, 1820–3.
- (10) Scharschmidt, M.; Pfeleiderer, G.; Metz, H.; Brümmer, W. *Hoppe-Seyler's Zeitschrift für physiologische Chemie* **July 1983**, *364*, 911–21.

-
- (11) Yamada, H.; Nagao, A.; Nishise, H.; Tani, Y. *Agricultural and Biological Chemistry* **1982**, *46*, 2333–2339.
- (12) Chen, H.; Nie, J.; Chen, G.; Fang, B. *Sheng wu gong cheng xue bao = Chinese journal of biotechnology* **Feb. 2010**, *26*, 177–82.
- (13) Ruzheinikov, S. N.; Burke, J.; Sedelnikova, S.; Baker, P. J.; Taylor, R.; Bullough, P. a.; Muir, N. M.; Gore, M. G.; Rice, D. W. *Structure (London, England : 1993)* **Sept. 2001**, *9*, 789–802.
- (14) Rao, S.; Rossman, M. *Journal of Molecular Biology* **May 1973**, *76*, 241–250.
- (15) Holm, L.; Rosenström, P. *Nucleic acids research* **July 2010**, *38*, W545–9.
- (16) Welch, K. D.; Davis, T. Z.; Aust, S. D. *Archives of biochemistry and biophysics* **Jan. 2002**, *397*, 360–9.
- (17) Dudev, T.; Lim, C. en *Annual review of biophysics* **Jan. 2008**, *37*, 97–116.
- (18) Hammes-Schiffer, S.; Benkovic, S. J. en *Annual review of biochemistry* **Jan. 2006**, *75*, 519–41.
- (19) Leichus, B. N.; Blanchard, J. S. *Biochemistry* **Dec. 1994**, *33*, 14642–9.
- (20) Brinen, L. S. et al. *Proteins* **Feb. 2003**, *50*, 371–4.
- (21) Gonzalez, R.; Murarka, A.; Dharmadi, Y.; Yazdani, S. S. *Metabolic engineering* **Sept. 2008**, *10*, 234–45.

Appendix A

Materials and Methods

A.1 General materials, micro-organisms, and plasmids

The plasmid construct containing TM0423 (JCSG Clone: TmCD00084978; GenBank: AAD35508) cloned from *T. maritima* strain MSB8, was kindly provided by JCSG¹. Protein was expressed in an *E. coli* HK100 expression strain (JCSG, derived from DH10B Genehogs by Invitrogen). Unless otherwise specified, all other reagents were from Sigma-Aldrich.

A.2 Protein overexpression and purification

The plasmid was transformed into chemically competent *E. coli* HK100 cells using heat shock². Cultures of transformed *E. coli* were grown in Luria-Bertani (LB) medium containing 100 mg/L ampicillin at 37°C; protein expression was induced by addition of arabinose to 0.02% during log phase growth ($OD_{600} = 0.8$). Following four hours of induction, cells were harvested by centrifugation at 10,000 x g at 4°C for 20 minutes. The cell pellet was resuspended in lysis buffer (50 mM phosphate buffer pH 8.0, 150 mM NaCl, 2 mM $MgCl_2 \cdot 6H_2O$) and lysed using an M-110L pneumatic microfluidizer (Microfluidics).

The whole cell lysate was clarified by centrifugation at 10,000 x g at 4°C for 20 minutes. The cleared supernatant was collected, heated to 60°C for 15 minutes to thermally precipitate

host cell proteins, and cleared again by centrifugation. The resulting supernatant containing TM0423 protein was loaded into a Ni-NTA agarose resin (Bio-Rad) column (2.67 mL of 50% resin slurry per liter of cell culture) which was previously washed with at least 10 column volumes dH₂O and equilibrated with 10 column volumes lysis buffer. The column with bound protein was washed with 15 column volumes wash buffer (50 mM phosphate buffer (pH 8.0), 150 mM NaCl, 60 mM imidazole); protein was eluted with 5 column volumes of elution buffer (50 mM phosphate buffer pH 8.0, 150 mM NaCl, 600 mM imidazole). The successful over-expression and purification of the target protein was monitored by sodium dodecyl sulfate-polyacrylamide gel (Bio-Rad) electrophoresis (SDS-PAGE)³. To remove excess metal chelated from the column resin, the sample was dialyzed against dialysis buffer 1 (50 mM phosphate buffer (pH 8.0), 150 mM NaCl, and 10 mM EDTA) overnight. A second dialysis against dialysis buffer 2 (50 mM phosphate pH 8.0, 150 mM NaCl) was then performed to remove EDTA.

A.3 Enzymatic assays

For all kinetic assays, the change in NADH absorbance was monitored at 340 nm using a PerkinElmer LAMBDA 25 UV/Vis spectrometer equipped with a PTP-1+1 Peltier System (Perkin Elmer, Waltham, MA). The extinction coefficient of $6220\text{ M}^{-1}\text{cm}^{-1}$ for NADH was used to calculate changes in NADH concentration. The Michaelis-Menten enzyme kinetics model was used to determine the apparent values of V_{max} , k_{cat} , and K_m . Unless otherwise specified all assays were run at 65°C. under standard assay conditions of 50 mM phosphate buffer (pH 8.0), 1-165 mM substrate, 0.375 mM NAD⁺, and 400 nM TM0423.

A.4 Protein quantification and progress curves

Protein concentration was determined by absorbance at 280 nm in a ND-1000 NanoDrop® spectrophotometer using the molar extinction coefficient of $22,265\text{ M}^{-1}\text{cm}^{-1}$ for TM0423 (calculated by ExPASy

ProtParam⁴). Activity progress curves were generated by monitoring the increase in absorbance at 340 nm resulting from the TM0423-dependent conversion of NAD⁺ to NADH. The optimal TM0423 concentration of 400 nM was used for all assays.

A.5 Metal dependence assay

Metal dependence was monitored in the standard reaction with 15-165 mM Glycerol. Five different divalent metals were tested at 0.3 mM to determine the metal dependence of the enzyme: ZnCl₂, NiCl₂·6 H₂O, CoCl₂·6H₂O, FeCl₂·4H₂O, and MnCl₂.

A.6 EDTA deactivation assay

To assay the dependence of TM0423 on divalent metals, EDTA was titrated into a reaction mixture and activity was observed. ZnCl₂ at 0.3 mM was added to the standard reaction mixture and the concentration of EDTA was varied between 0-400 μ M.

A.7 Substrate analog assays

The specific activity of TM0423 was tested with various glycerol analogs including 1,2-propanediol, ethylene glycol, tris, 2-propanol, 1-propanol, 2-methyl-2-propanol, 1-butanol, bis-tris propane, and 1,3-propanediol. Each substrate was tested independently in the standard reaction with the substrate analogs at 120 mM.

Additional kinetic assays were run under standard condition for analogs 1,2-propanediol and ethylene glycol to obtain kinetic parameters. In addition, a kinetic assay was run to survey the reduction of hydroxyacetone to 1,2-propanediol by TM0423 with the concomitant oxidation of NADH to NAD⁺. The reaction conditions for this assay were 50 mM phosphate buffer (pH 8.0), 15-165 mM hydroxyacetone, 0.375 mM NADH, and 400 nM TM0423.

Bibliography

- (1) Lesley, S. a. et al. *Proceedings of the National Academy of Sciences of the United States of America* **Sept. 2002**, 99, 11664–9.
- (2) Hanahan, D. *Methods* **1983**, 166, 557–580.
- (3) Laemmli, U. K. *Nature* **Aug. 1970**, 227, 680–685.
- (4) Gasteiger, E.; Gattiker, A.; Hoogland, C.; Ivanyi, I.; Appel, R. D.; Bairoch, A. *Nucleic acids research* **July 2003**, 31, 3784–8.