TOWARD THERAPEUTIC NANOASSEMBLIES: THE DESIGN AND MODELING OF PROTEIN-PROTEIN INTERACTIONS

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

BRIAN RICHARD WHITE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DR. CARSTON R. WAGNER, ADVISER

NOVEMBER 2009

Acknowledgement

No number of eloquent words exist that can appropriately express my gratitude to those that have assisted me in this journey. My most sincere thanks go to my adviser, Dr. Carston R. Wagner. His leadership, education, integrity, and enthusiasm have shaped who I am as a researcher and as an individual, and if I can only reside in his shadow, I have accomplished more than I would have thought possible. I also would like to thank Dr. Patrick Hanna, whose helpful suggestions and evaluations have pushed me to succeed in his image. Many thanks also go to Dr. Donald Truhlar and Dr. Elizabeth Amin, whose limitless patience and guidance led me from a complete novice in the field of computational chemistry to the competent researcher I am today. Without their assistance, this thesis work would have been impossible. I would also like to thank Dr. Mark Distefano, whose helpful insights have improved this thesis.

I truly respect and appreciate the time spent with the former and current members of the Wagner Lab, specifically Dr. Tsui-Fen Chou, Dr. Jonathan Carlson, Dr. Phalguni Ghosh, Dr. Brahma Ghosh, Dr. Adrian Fegan, Dr. Matthew Cuellar, Dr. Brandie Kovaleski, Dr. Sid Kumarapperuma, Dr. Li Liu, Dan Drontle, Cindy Choy, Xin Zhou, Qing Li, Yan Jia, Sanaa Bardaweel, and Kevin O'Halloran. Their camaraderie "in the trenches" and helpful discourse have made my graduate work more than a memorable experience.

Lastly, I owe my deepest gratitude to my parents, Richard and Lynn; my brother, Jeffrey; my wife, Tiffany; and my closest friends. Their unconditional love and support have borne me through the times when I couldn't carry myself, and for that I can never be grateful enough.

Dedication

This thesis and the work contained within are dedicated to the birth of my daughter, Amelia May White, who has taught me more about myself in four short months than I have managed to learn in 29 years.

Abstract

Unraveling the nanoscale processes of biological pathways via the testing, replication, and visualization of the underlying mechanisms remains a persistent challenge in the study of these critical life-governing systems. Recent advances in the field of chemically induced dimerization have unlocked multiple tools for the exploration of these facets of biology, including the development of switchable signaling systems, assertion of control over protein localization in the cell, and regulation of gene expression. An additional revelation through protein complexation by chemical induction is the construction of multivalent protein-based nanostructures, capable of bearing multiple targeting agents. However, stochastic assembly of these proteins has proven unsatisfactory in generating homogeneous populations. Herein, we have taken the initial steps toward developing a protein-based biomolecular language for nanostructural assembly. Through gel filtration analysis, we have characterized the ability of interfacial point mutations to modulate the stability of a bis-methotrexate (bis-MTX) induced *E. coli* dihydrofolate reductase (DHFR) dimer over a dynamic range of 1.5 kcal/mol. Furthermore, we have employed single-molecule fluorescence assays to demonstrate the stabilization of a heterodimeric DHFR dimer, yielding 4-fold selectivity for the heterodimer over either corresponding homodimer.

In addition to our experimental characterization of the chemically induced DHFR dimer, we have also taken steps toward the construction of a tripartite computational model of dimerization in an effort to predict the effects of further mutations. We have tested a number of molecular mechanics force fields against quantum mechanical benchmarks and discovered that the MMFF94, OPLS2005, and

AMBER force fields yield the most accurate electrostatic and configurational treatment of the complex bis-MTX dimerizer. While initial attempts at calculating the binding free energy of the macromolecular complex have been unsuccessful, we have gleaned important insights into the complexities of modeling this three-body system. The advances described within the following work delineate important aspects of protein interface remodeling in a chemically induced system and provide an avenue toward the further development of both a computational model of protein interactions and the future directed assembly of protein based materials and therapeutic nanostructures.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

CHAPTER FIVE

APPENDIX ONE

LIST OF FIGURES

CHAPTER ONE

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

CHAPTER FIVE

APPENDIX ONE

LIST OF ABBREVIATIONS

13DC	DHFR-13 amino acid linker-mCherry-CVIA
13DC-CVIA	DHFR-13 amino acid linker-mCherry-CVIA
13DG	DHFR-13 amino acid linker-eGFP
13DM	DHFR-13 amino acid linker-mRFP
1DC	DHFR-Gly-mCherry-CVIA
1DC-CVIA	DHFR-Gly-mCherry-CVIA
1DC-His-CVIA	DHFR-Gly-mCherry-His6-CVIA
2,4-DAP	2,4-diaminopyrimidine
2-AMP	2-(aminomethyl)pyrazine
Aβ	amyloid β
AD	activation domain
ALEX	alternating laser excitation
AMUE	average mean unsigned error
ASC-DHFR	C85A/C152S/C162 mutated DHFR
AS-DHFR	C85A/C152S mutated DHFR
bis-MTX	bis-methotrexate
CCSD	coupled cluster theory - single and double excitations
CID	chemically induced dimerization
CIP	chemically induced proximity
CM4	charge model 4
CR	congo red
CTB	cholera toxin B pentamer
CTL	cytotoxic T-lymphocyte
CVIA	Cys-Val-Ile-Ala prenylation tag
DBD	DNA binding domain
DEAE	diethylaminoethyl
DFT	density functional theory
DHF	dihydrofolate
DHFR	dihydrofolate reductase

DNA	deoxyribonucleic acid
DNP	2,4-dinitrophenol
DTT	dithiothreitol
ecDHFR	E. coli DHFR
EDTA	ethylenediaminetetraacetic acid
eGFP	enhanced green fluorescent protein
FCS	fluorescence correlation spectroscopy
FEP	free energy perturbation
FKBP	FK506 binding protein
FM	FKBP mutant
FP	fluorescent protein
FRAP	FKBP-rapamycin associated protein
FRB	rapamycin binding fragment
FRET	Forster energy resonance transfer
GAFF	general atom force field
GB	generalized Born
GVHD	graft versus host disease
GyrB	DNA gyrase B subunit
HF	Hartree-Fock
hGH	human growth hormone
HPLC	high performance liquid chromatography
kDa	kilodalton
LB	Luria Bertani
MC	monte carlo
MD	molecular dynamics
MEP	molecular electrostatic potential
MM	molecular mechanics
MM-GBSA	molecular mechanics generalized Born surface area
MM-PBSA	molecular mechanics Poisson Boltzmann surface area
mRFP	monomeric red fluorescent protein

MS/MS	tandem mass spectrometry
mTOR	target of rapamycin
MTX	methotrexate
MTX_2C9	bis-methotrexate, nine-carbon linker
MUE	mean unsigned error
NADPH	nicotinamide adenine dinucleotide - reduced form
Ni-NTA	nickel-nitrilotriacetic acid
OD	optical density
PCR	polymerase chain reaction
PMF	potential of mean force
QM	quantum mechanics
RMUE	reduced mean unsigned error
RNA	ribonucleic acid
SAP	human serum amyloid P component
SCF	self-consistent field
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEM	standard error of the mean
SE-MO	semiempirical molecular orbital
SLF	synthetic ligand for FKBP
SSE	sum of squared errors
Stx	Shiga toxin
TCCD	two-color coincidence detection
TCR	T-cell receptor
TEM	transmission electrom microscopy
VMD	visual molecular dynamics
WFT	wave function theory
WHAM	weighted histogram analysis method
WT	wild-type

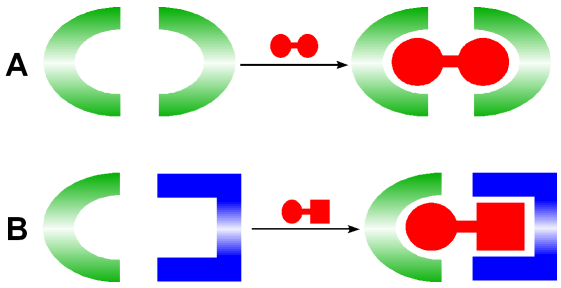
Chapter One

Chemically Controlled Protein Assembly: Techniques, Applications, and Modeling

1. Introduction

The study of biological processes has driven the efforts of modern molecular biology to unravel the microscopic capabilities of natural systems. Intrinsic to the experimental analysis of these life-governing principles is the process of testing, replicating, and visualizing the underlying biological mechanisms. As such, interacting with the nanoscale machinery of life becomes an increasingly apparent challenge.[1] The range of this pursuit spans from DNA to RNA to proteins. While the controlled assembly of nucleic acid structures is widely studied,[2,3] there are a smaller number of studies on the development of methods that investigate and exploit protein assembly and protein-protein interactions.[4,5] These ubiquitous natural phenomena form a central foundation for the regulatory choreography of life and play a critical role in the physical structure of organisms. Moreover, protein-protein interactions span a vast scale of time and size, from tiny transient interactions within the cell to the macroscopic functional arrays that make up muscle and skin. The exertion of control over protein-protein interactions represents a powerful tool in many disciplines. On the smallest scale of protein-protein assembly is induced dimerization, the stimulus driven association of a single pair of proteins. A chemical inducer of dimerization, the “dimerizer”, acts to bring the two proteins together to form a homodimer (if the proteins are the same) or heterodimer (if the proteins are different) as shown schematically in Figure 1. Chemically induced dimerization has been shown to be a powerful tool for the investigation of cellular events.

Figure 1. General principle of chemically induced dimerization (CID). In the presence of a symmetrical ligand, two proteins can be brought together to form a homodimer (A). With a non-symmetrical ligand, two different proteins can be brought together to form a heterodimer (B).

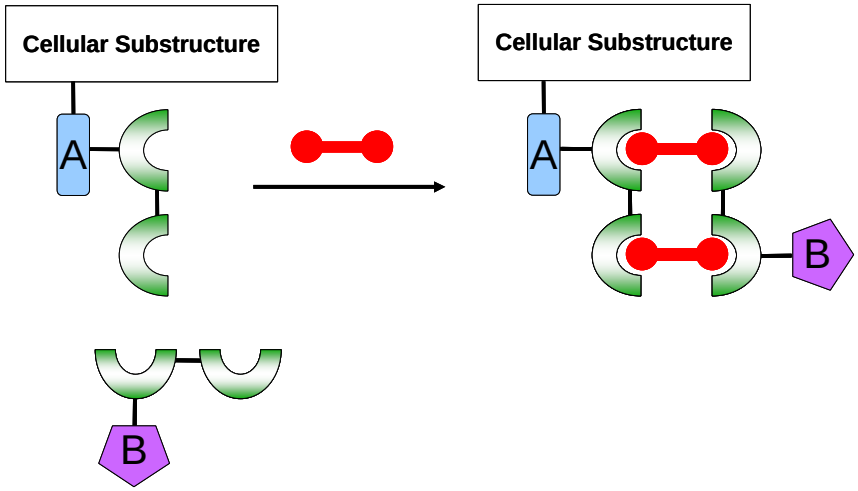


The dimerizer can be a small molecule, another protein, or even a patch of complementarity on the protein surface. A number of reviews have covered the self assembly of proteins via programmed amino acid interactions.[4,5] Herein we review the assembly of proteins under the control of small molecule chemical signals, the dramatic strides made in the refinement of synthetic CID tools, and the divergent directions this work has taken as the concept of dimerization has expanded to a larger notion of chemically induced proximity (CIP) and how CIP has been used as both an investigative and a therapeutic tool. In addition, the potential for CID to be used to direct the assembly of supramolecular protein structures will also be reviewed.

2. Chemically Induced Dimerization and Chemically Induced Proximity – Developing the Systems

Chemically induced dimerization is the controlled dimerization of a pair of proteins, via any one of a number of classes of dimerizers. The dimerizer acts to bring the two proteins together and the induced dimerization can be used to increase the effective molarity of a protein at a certain cellular substructure, thus causing chemically induced proximity of the two previously dispersed proteins (Figure 2). The increased effective concentration of the proteins may be used to activate or control a biological event. In Figure 2, the induced proximity of proteins A and B is mediated by a different protein, which has been fused to the proteins of interest. This is a common method for causing the dimerization of proteins. The result of dimerization is flexible as it is directed by the domain(s) fused to the protein which is being used to effect the

Figure 2. An example of chemically induced proximity (CIP). In the absence of the dimerizer, the effective molarity of protein B is low. While in the presence of dimerizer, protein B is recruited to the cellular location of protein A, thus controlling or initiating a biological stimulus.



dimerization: the result, for example, can be association of two dispersed cytoplasmic proteins or recruitment of a freely diffusible protein to the location held by another, such as a membrane surface,[6] a cellular compartment,[7] or a DNA/RNA binding site.[8]

Although this review focuses on small molecule induced dimerization, it is important to briefly mention other species which can induce dimerization, including protein-based dimerizers. Physiologically, induced dimerization is particularly critical in transmembrane receptor signal transduction.[9] A broad class of hormone receptors function by ligand-induced association; two copies of a receptor are brought together and thereby activated through the binding of a single hormone molecule.[10] Human growth hormone[11] and granulocyte macrophage colony-stimulating factor[12] are two examples. An added layer of complexity has been unearthed for the hormone erythropoietin and its receptor, for which hormone binding conformationally reorganizes a predimerized receptor to initiate signaling.[13,14] The human interferon-γ has also been shown to induce receptor dimerization in solution and on cell surfaces.[15] Nucleic acids are also used as inducers of dimerization in nature[16-18] and have been exploited in the formation of nanostructures *in vitro*.[19,20]

Combining a DNA binding domain (DBD) and a protein binding domain, the activation domain (AD), within one species has lead to the creation of molecules which can be used to cause the dimerization of DNA and protein complexes. These species have been used as artificial transcriptional activators to control gene expression.[21,22] The DBD recognizes and binds a particular DNA sequence allowing specific gene activation while the AD interacts with one or more components of the natural transcriptional

machinery. A number of DBDs have been used for this purpose including polyamides, peptides, and triplex forming oligonucleotides and peptide nucleic acids.[21-23] While there is good understanding of the characteristics required for designing DBDs, research continues toward discovering suitable ADs. Induced dimerization of proteins has also been used to control gene expression and is reviewed in Section 3.1.

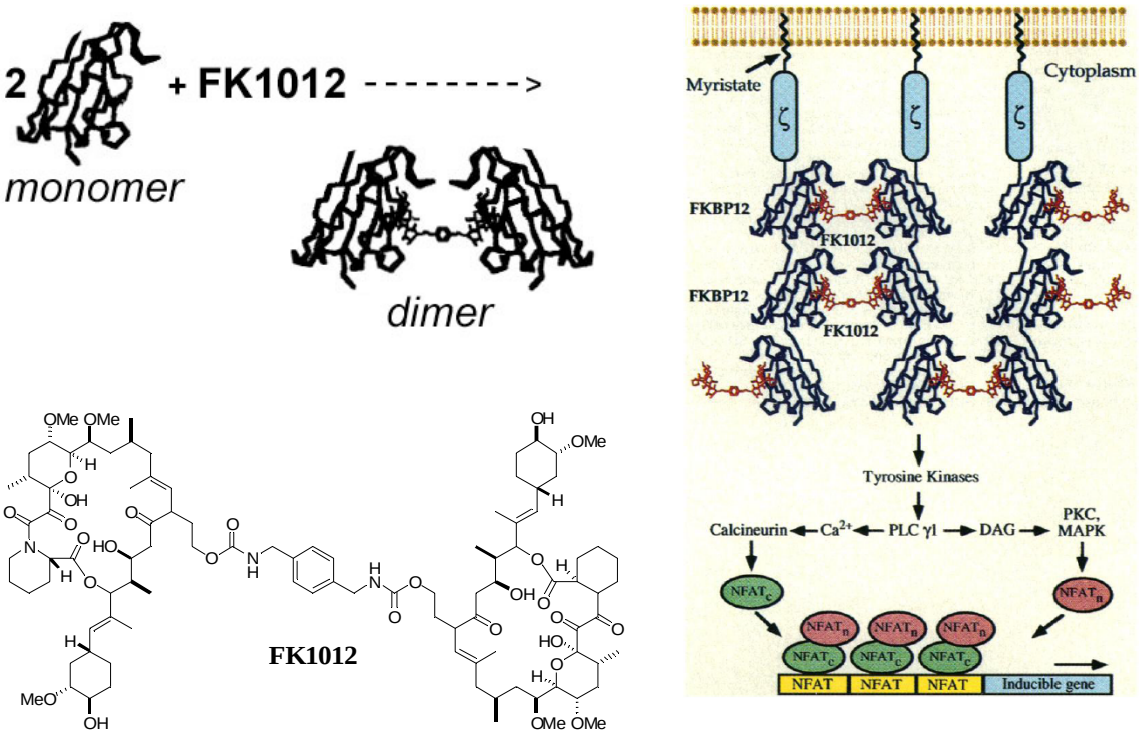
2.1. Initial Report of Small Molecule Induced Dimerization

The concept of chemically induced dimerization initiated by a small molecule rather than a hormone and its first application were introduced in a landmark paper in 1993 by Schreiber, Crabtree and coworkers.[24] They demonstrated a means by which a synthetic molecule could reproduce the ability of natural systems to use proximity as an activation switch. A bivalent derivative of the tight-binding immunosuppressive drug FK506 was shown to reversibly dimerize its protein target, FK506 Binding Protein (FKBP, Figure 3). Most importantly, the bivalent ligand, FK1012, functioning as a dimerizer, could be used to drive biological function. By fusing FKBP to the proximity regulated ζ-chain of the T-cell receptor (TCR), they produced a system by which binding of FK1012 activated the endogenous signal transduction cascade (Figure 3).

2.2 Expanding the Chemically Induced Dimerization Toolkit

The 15 years since this original work has seen a myriad of investigations that have elucidated the basic principles governing CID systems and their utility. FKBP has proven to be a particularly flexible agent for inducing proximity and has been used for a

Figure 3. The initial demonstration of the CID concept. On the left, two FKBP monomers bind the bivalent drug FK1012. On the right, FKBP-TCR fusion proteins are dimerized by FK1012, initiating intracellular signaling. Figure reprinted from Spencer et al.[24] with permission. Copyright (1993) AAAS.

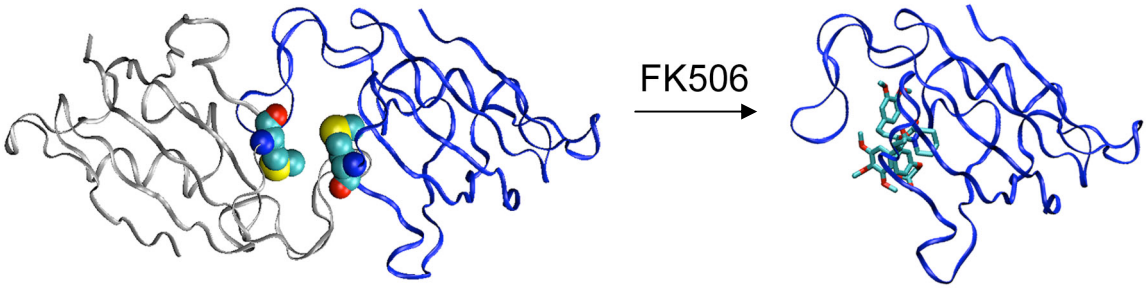


number of investigative and therapeutic applications. as well as mutated FKBPs and modified FK1012 ligands.[25,26] In one example from Clackson et al., a FK1012 analogue was synthesized with a "bump" preventing binding to the wild type FKBP by steric interference.[25] They then made an FKBP mutant (F36V) with a compensatory "hole" which allowed binding of the modified ligand with low nanomolar affinity. After synthesizing a dimerizer-based on the new ligand (AP103), they showed that the system is functional both *in vitro* and *in vivo*.

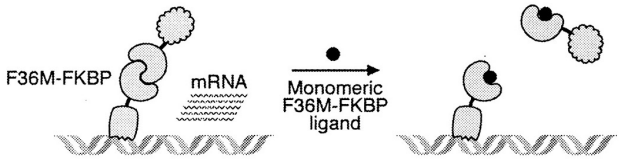
A fascinating side effect of these FKBP remodeling studies was the coincidental identification of an FKBP mutant (FM) that forms a stable dimer (K_d = 30 μM) in the absence of any ligand.[27] Moreover, synthetic nonimmunosuppressive FKBP-binding ligands were able to fully reverse self-dimerization of FM, thus creating a ligand switching system for protein aggregate disassembly (Figure 4). This phenomenon was cleverly exploited as a trigger for pharmaceutically regulated secretion.[28] Critical to this method was the observation that fusion proteins with multiple copies of FM will form not simple dimers but higher-order aggregates. Thus, by linking four copies of FM to secreted peptide hormones such as human growth hormone or insulin, the translated fusion proteins will form large aggregates retained in the endoplasmic reticulum. Addition of the FM–binding ligand disaggregated the proteins and initiated hormone secretion. In another example of ligand remodeling, Koide et al. have prepared a library of cell-permeable heterodimeric small molecules using olefin metathesis. A representative ligand library was screened for molecules which were cell-permeable and a number were shown to induce dimerization in intact cells.[29]

Figure 4. An inverse dimerization system. A) The F36M variant of FKBP exists as a constitutive dimer, left; Met36 is highlighted in a CPK model. In the presence of FK506, shown as a stick model in its binding pocket, the dimer interface is disrupted, restoring the protein to its typical monomeric state. B) Schematic representation of the inverse-dimerization system, in which FK506 binding inactivates transcription due dissociation of the F36M-FKBP dimer and concomitant breakdown of the artificial transcriptional activator. Figure 4A was rendered in VMD[30] from PDB structures 1EYM and 1BL4. Figure 4B reprinted from Rollins et al.[27] with permission. Copyright (2000) National Academy of Sciences, U.S.A.

A



B

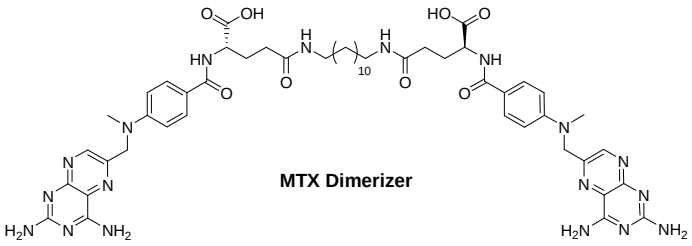


Alternative proteins (and corresponding ligands) to induce dimerization have been developed, including dihydrofolate reductase (DHFR) and a methotrexate (MTX) based ligand (Figure 5)[31,32] The natural product coumermycin (Figure 5) has also been shown to cause dimerization of a bacterial DNA gyrase B subunit (GyrB) and has been used as a dimerizer.[33,34] This expansion of the number of dimerization systems is important for the formation of heterodimeric systems. Also, the development of novel dimerization systems should lead to improved biocompatibility within natural systems by reducing/eliminating off-target effects, such as binding of the dimerizer to naturally occurring proteins.

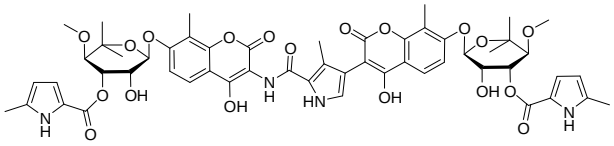
2.3. Heterodimeric Dimerization

Homodimeric dimerizers have their most elegant or economical application in switching systems that dimerize a single fusion protein, such as the original construct described by Schreiber and coworkers.[24] Such symmetric dimerizers can be used to dimerize nonequivalent fusion proteins as well. If two fusion proteins (X-A and X-B, where X represents the dimerization domain) are present in equal mixtures, and complex formation is governed by random assortment, 50% of the ligand-induced dimers should be heterodimeric (A-XX-B) with 25% of each homodimer also formed (A-XX-A, and B-XX-B). Depending on the degree of amplification available in a cellular context, this degree of activation has typically been shown to be sufficient, provided that the homodimeric species do not produce a dominant-negative effect.[35]

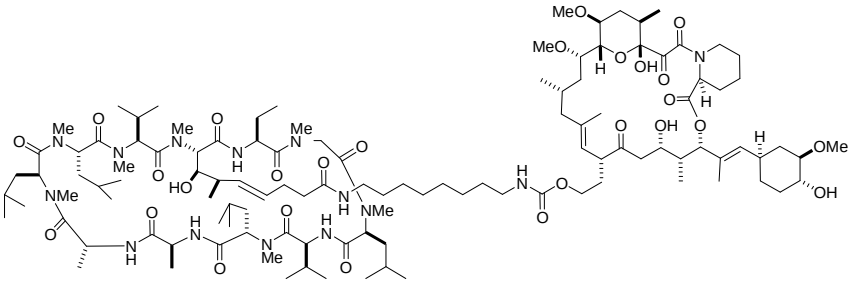
Figure 5. Structures of the methotrexate, coumermycin, and FKCsA dimerizers which cause dimerization of DHFR, GyrB and FKBP-cyclophilin, respectively.



MTX Dimerizer



Coumermycin



FKCsA

However, reliance on these probabilities is unsatisfying, and the ability to specifically produce only A-XX-B pairs is clearly the more precise and elegant route. In principle, such a system could be ligand directed, generated by a bivalent ligand with two distinct protein binding targets, or protein directed by engineering complementary mutations into the adjacent surface of the dimerized protein to promote heterodimeric pairs.

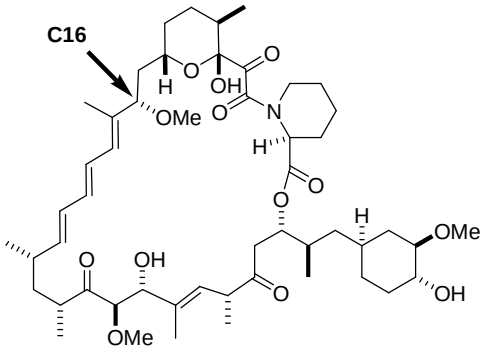
Schreiber and coworkers, choosing the former route, constructed a heterodimeric dimerizer from FK506 and cyclosporin, referred to as FKCsA (Figure 5), and demonstrated three distinct modes of selective intracellular signaling.[36] The pro-apoptotic effects of FKCsA were tested in a construct that used a membrane-localized triple-copy FKBP to recruit multiple cyclophilin-Fas fusion proteins. Dimerizer-induced localization of the Fas intracellular domain at the inner membrane surface produced a concentration dependent reduction in secreted alkaline phosphatase activity, used as a marker of cell viability. Finally, FKCsA was shown to direct nuclear localization of a GFP-cyclophilin fusion protein via dimerization with an FKBP-GAL4 fusion target.

A second method developed for generating selective heterodimers exploits the natural product rapamycin, an immunosuppressant that also binds to FKBP. Rapamycin and FK506 both exert their biological effect by an unusual mechanism: after binding to FKBP, the combination of the protein and the ligand (but neither one alone) binds to and inhibits a second cellular protein target, calcineurin, a serine-threonine phosphatase critical to T-cell receptor signaling. Rapamycin-FKBP binds to FRAP (FKBP-rapamycin associated protein, also known as mTOR, the target of rapamycin), a kinase involved in IL2/cytokine signal transduction.[37,38] FRAP is a gigantic protein, 289 kDa

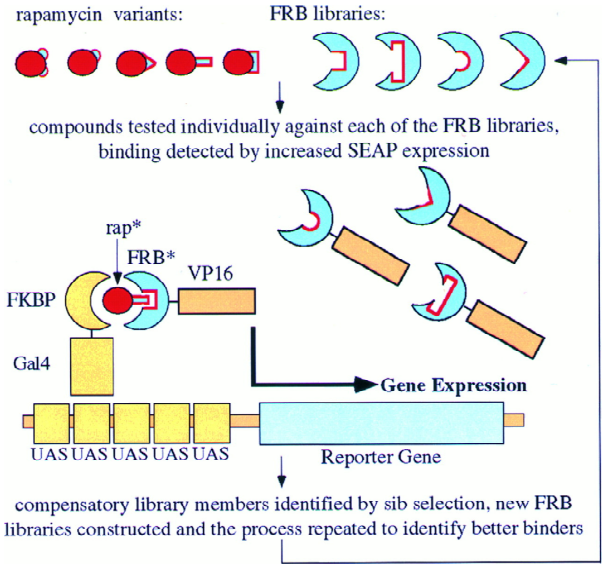
and 2549 amino acids, an undesirable size for protein engineering. However, the rapamycin binding function can be localized to a mere 90 amino acid segment, a rapamycin binding fragment (FRB, 11 kDa) that preserves the full binding affinity of the complete protein.[39] In light of these successes, the prominent role played by the immunophilins in the CID literature becomes clear: FKBP and FRB are particularly small (and therefore unobtrusive) and bind their ligands very tightly (sub nM)[40], ideal traits for protein engineering.

For derivatives of rapamycin to be maximally useful in therapeutic dimerizer systems, the immunosuppressive activity needed to be neutralized, a process that had already been undertaken in the remodeling of cyclosporin A and its binding target cyclophilin, among other examples.[25,41] Guided by the crystal structure of rapamycin in binary complex with FKBP and FRB, Schreiber and coworkers successfully designed an orthogonal rapamycin-FRB pair.[42,43] Replacement and stereochemical inversion of the C16-OMe group in a native rapamycin produced conformationally distorted analogs – "rapalogs" – that bound wild-type FRB with up to 300-fold lower affinity (Figure 6). Given the substantial conformational change in the modified rapamycin, FRB remodeling by rational design was judged impractical. Rather, an iterative genetic screen that used a three-hybrid technique – another evolving adaptation of the CID principle – was applied to select for mutations of FRB that restored binding to the altered ligand (Figure 6). The critical α-helix in FRB adjacent to the modified section of rapamycin was targeted for modification, and a triple mutant capable of binding the rapalog with nanomolar affinity identified, completing a

Figure 6. Schematic showing rapamycin with the position of modification (C16) highlighted and the experimental scheme to identify a rapalog-mutated FRB pair. The rapalogs were individually screened against a library of FRB mutants with increased expression from the reporter gene being used to identify matched rapalog-FRB mutant pairs. Figure elements reprinted from Liberles et al.[43] with permission. Copyright (1997) National Academy of Sciences, U.S.A.



Rapamycin



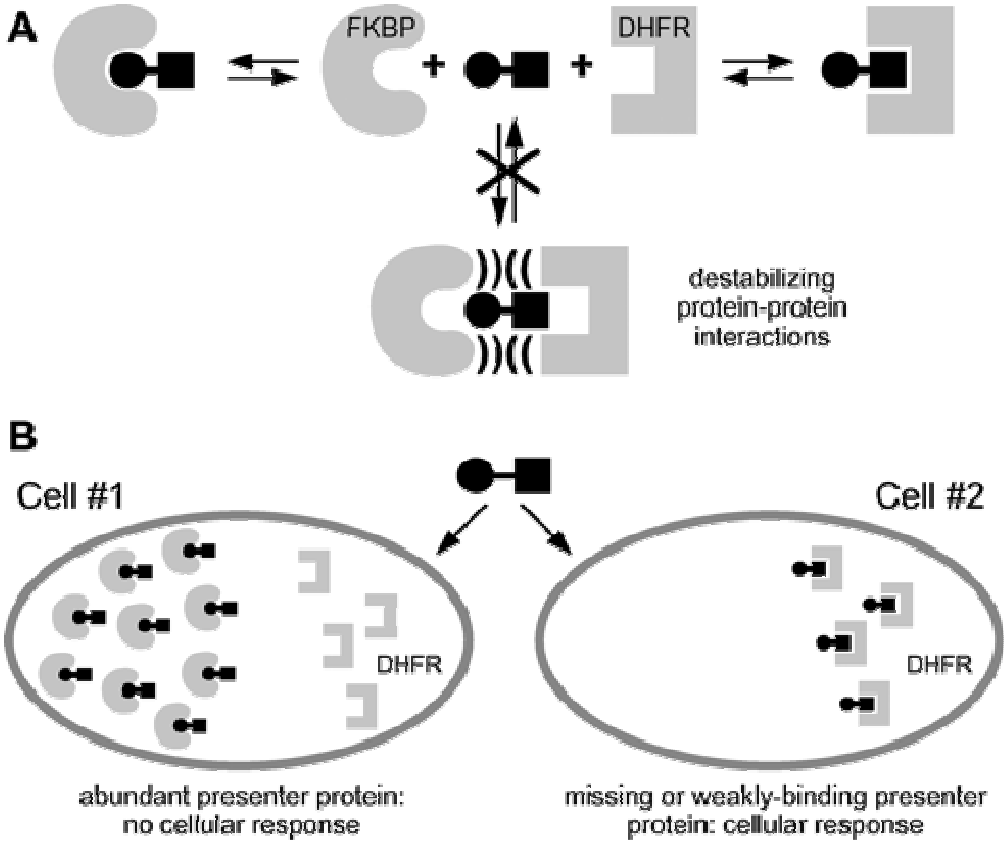
rapamycin-based orthogonal pair suitable for *in vivo* applications. Further rapalogs have been developed to permit orthogonal control of protein activity.[44]

2.4. Affinity Modulation

Chemically induced dimerization consists of two binding events – the primary event, between the monomeric protein and the dimerizer, and the secondary event, between the monomer-dimerizer complex and another monomer. This secondary binding event has been used as a means to perturb the equilibrium of the first event, a notion referred to as affinity modulation.[45] Conceptually, the protein-protein interactions introduced via dimerization, if favorable, can serve to amplify the binding strength of a ligand, akin to the mechanism by which the FK506-FKBP complex functions to inhibit calcineurin.[46] Conversely, unfavorable protein-protein interactions could reduce binding affinity, potentially desirable as a means to reduce toxicity in selective molecular contexts (Figure 7A). In the larger context, multiple-target binding has been proposed as a universal technique for engineering therapeutic selectivity by creating ligands that exert their pharmaceutical effect only in the presence of specific combinations of protein targets.[47] Crabtree and coworkers synthesized an FK506-peptide dimer and demonstrated the ability of FKBP-target interactions to enhance the affinity of an SH2 domain binding peptide by a factor of three or decrease it by a factor of six.[45,48]

The ability to harness unfavorable protein-protein interactions to generate ligand selectivity has been demonstrated *in vivo*. Wandless and coworkers exploited

Figure 7. The concept of affinity modulation. A) A bivalent drug, capable of binding to either the FKBP protein (gray semicircle) or to the DHFR target protein (gray rectangle) but not to both proteins simultaneously due to destabilizing protein-protein interactions. B) Cell-selective activity of a bivalent drug. In cells that possess an abundant, high affinity FKBP protein the bifunctional molecule will partition preferentially to bind to FKBP, leaving DHFR uninhibited. In cells lacking FKBP or with a protein which binds weakly to FKBP ligands, the bifunctional molecule will selectively partition to inhibit DHFR and elicit a cytotoxic response. Figures reprinted from Braun et al.[49] with permission. Copyright (2003) American Chemical Society.

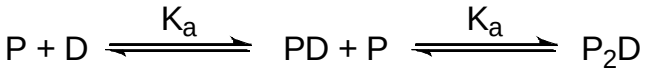


differential binding of FKBP to convert MTX into a plasmodium-selective DHFR inhibitor.[49] Alone, MTX is completely unselective, with equivalent binding affinity for both plasmodium and human DHFR. A genome search revealed that *P. falciparum* possesses a single FKBP homologue (pfFKBP), which is present in the parasite cytoplasm at a 50-fold lower concentration than the level of hFKBP in human cells. Moreover, the plasmodial enzyme was observed to bind SLF, a synthetic ligand for FKBP, with 14-fold lower affinity than did hFKBP. Wandless and coworkers thus synthesized an MTX-SLF heteroligand, reasoning that these *in vivo* differences in the prevalence and affinity of FKBP could effectively modulate the toxicity of MTX (Figure 7B). *In vivo*, MTX-SLF displayed weak cytotoxicity to MES-SA uterine cancer cells, with an IC_{50} of 25 μM. This detoxification could be reversed, lowering the IC_{50} 68-fold, by saturating the intracellular FKBP with 5 μM FK506-M, a non-toxic, monomeric FKBP-binding ligand. In contrast, when MTX-SLF was tested against live plasmodia, co-administration of FK506-M had no effect on the IC_{50} of 1.5 μM, yielding a parasite-selective therapeutic index of 16.7.[49]

2.5. Theoretical Principles Governing Dimerization

In order to better characterize the increasing number of developing practical applications for dimerizer-based systems, theoretical models describing dimerization have been derived and refined by several groups. All basic theoretical models of induced dimerization can be crudely described via the equilibrium expression shown in Scheme 1.

Scheme 1. Basic Dimer Equilibrium



Perelson and DeLisi derived one of the earliest expressions for treating a mixture of protein and dimerizer, allowing for calculation of the fraction of dimerized protein present in the solution.[50] However, this expression relied on the assumption that the total concentration of dimerizer was equal to the concentration of free dimerizer. Simply put, the dimerizer concentration must be much greater than the protein concentration. For dimerizer-based systems, where the aim is to maximize the fraction of P_2D complex, this assumption does not hold.

Hu and coworkers, while examining the DHFR-based dimerization system described in Section 2.2, initially reported that the formation of P_2D did not correspond to a simple, noncooperative binding model that would be assumed for such a complex.[31] In such a model, wherein each protein binds independently to the dimerizer, complex formation should build toward a maximum when the protein:dimerizer ratio equals 2, after which the addition of dimerizer should drive the equilibrium toward the binary complex. Although the data fit the noncooperative model up to the protein:dimerizer ratio of 2, the fraction of P_2D was generally unaffected even in the presence of a 50-fold excess of dimerizer. To explain this discrepancy, the concept of affinity modulation (cooperativity, or K_c) was introduced into the theoretical description of dimerization (Scheme 2).

Scheme 2. Dimer Equilibrium Accounting for Binding Cooperativity

$$P + D \underset{}{\overset{K_a}{\rightleftharpoons}} PD + P \overset{K_aK_c}{\rightleftharpoons} P_2D$$

Given this equilibrium expression and the mass balance Equations 1 and 2,

$$P_{tot} = [P] + [PD] + 2[P_2D] \quad (1)$$

$$D_{tot} = [D] + [PD] + [P_2D] \quad (2)$$

expressions for the concentration of singly and doubly-bound protein as well as free monomer can be derived (Equations 3-5).

$$[P] = \frac{-(1 + K_a[D]) + \sqrt{1 + K_a[D]^2 + \left(8P_{tot}K_a^2K_c[D]\right)}}{4K_a^2K_c[D]} \quad (3)$$

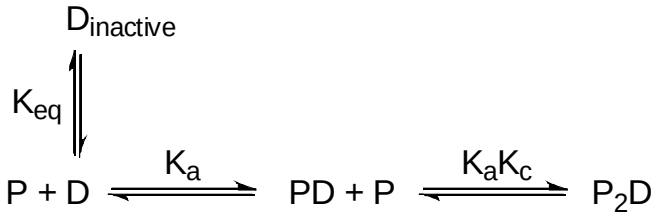
$$[PD] = K_a[P][D] \quad (4)$$

$$[P_2D] = K_a^2K_c[P]^2[D] \quad (5)$$

With the theoretical model governing dimerization now taking into account binding cooperativity, the role of putative protein-protein interactions at the newly formed protein interface in the stability (or instability) of the non-native protein dimer can be calculated. However, Carlson et al. began to examine the effects of another critical component of the DHFR dimerization event – the ligand itself. Given the flexible nature of the linker tethering the functional parts of the dimerizer, it can be envisioned that such molecules are subject to a number of intramolecular interactions, some of which may render the dimerizer unable to bind its protein target. Through

molecular modeling, gel filtration experiments, and NMR analysis, it was found that bis-MTX adopts a primarily folded state in solution, which limits the concentration of dimerizer available for binding.[32] Given this realization, the theoretical treatment of dimerization can be amended to contain the equilibrium expression between active and inactive dimerizer as shown in Scheme 3. Given this equilibrium expression and the mass balance Equations 1 and 2, a new expression for the concentration of dimer can be derived (Equation 6):

Scheme 3. Dimer Equilibrium with Cooperativity and Ligand Behavior



$$[P_2D] = \frac{K_a^2 K_c [P]^2 [D]}{(K_{eq} + 1)} \qquad (6)$$

While this treatment of the theoretical basis for dimer formation has illustrated both the concepts of protein cooperativity and ligand conformational equilibria, the model still requires an initial guess for certain concentrations followed by an iterative fit of the data to generate equilibrium constants. Whitesides and coworkers sought to simplify and generalize this model by allowing for the direct calculation of parameters of interest.[51] Revisiting the equilibrium expression found in Scheme 2, and approaching

the expressions using the monovalent ligand dissociation constant K_d rather than K_a, a new expression can be written as found in Scheme 4.

Scheme 4. Dimer Equilibrium Based on Apparent Dissociation Constants

$$P + D \underset{}{\overset{K_d'}{\rightleftharpoons}} PD + P \underset{}{\overset{K_d''}{\rightleftharpoons}} P_2D$$

In this equilibrium expression, the apparent dissociation constants can be described in terms of the monovalent ligand dissociation constant as shown in Equations 7 and 8.

$$K_d' = \frac{K_d}{2} = \frac{[P][D]}{[PD]} \quad (7)$$

$$K_d'' = \frac{2K_d}{K_c} = \frac{[P][PD]}{[P_2D]} \quad (8)$$

The statistical factors of ½ and 2 account for the different ways to form the protein-dimerizer complexes, and K_c is present to account for protein cooperativity. In this case, if protein cooperativity is less than 1, P binds less strongly to PD than it would to D alone. If $K_c = 1$, binding data will fit to a simple noncooperative binding model. If $K_c > 1$, the system is positively cooperative and the resulting dimer will be favored by stabilizing interactions. Given these equations and the mass balance Equations 1 and 2, equations describing the concentration of singly- and doubly-bound complexes can be derived (Equations 9-10).

$$[PD] = \frac{2K_d[P]D_{tot}}{K_d^2 + 2K_d[P] + K_c[P]^2} \quad (9)$$

$$[P_2D] = \frac{K_c[P]^2 D_{tot}}{K_d^2 + 2K_d[P] + K_c[P]^2} \quad (10)$$

These expressions, together with equations describing the fraction of protein in each state (free, singly-, and doubly-bound), can be transformed into two important expressions allowing for the direct calculation of K_d and K_c. The first (Equation 11) shows the total concentration of bivalent ligand at which the fraction of dimer (D_p) is at a maximum:

$$D_{tot,\max} = \frac{K_d}{2} + \frac{P_{tot}}{2} \quad (11)$$

From this equation, it can be seen that with the knowledge of the maximum value for D_p and P_{tot}, one can directly evaluate the K_d. The second expression (Equation 12) allows for the calculation of the maximum D_p.

$$D_{p,\max} = 1 + \frac{2K_d}{K_c P_{tot}} - \sqrt{\left(\frac{2K_d}{K_c P_{tot}}\right)^2 + \frac{4K_d}{K_c P_{tot}}} \quad (12)$$

Based on this equation, the maximum D_p is dependent on P_{tot}, K_d, K_c, and D_{tot}. Given the values of $D_{p,max}$, P_{tot}, and K_d, one can estimate the cooperativity, K_c.

Although this model does not take into consideration ligand conformational behavior, it remains an exact method of evaluating the ligand dissociation constant and protein cooperativity without the use of approximations or data fitting procedures. An additional consideration is that cooperativity, in this context, is a measure of all stabilizing or destabilizing interactions, which can range from protein-protein or protein-ligand effects to entropic or solvation contributions. Overall, the further

expansion of a theoretical understanding of protein dimerization and oligomerization will lead toward general and practical applications to chemical dimerizer design.

3. Application of Chemically Induced Proximity to Therapeutics

While the development of chemically induced dimerization as an investigative tool has yielded great insights into protein structure and function, the application of these tools toward pharmacological problems remains a high priority. A natural extension of the techniques explored thus far, current CID technology has focused primarily on gene expression and signal transduction. However, interesting and exciting new paths for CID-based therapies are emerging in the modification of endogenous protein-protein interactions based on small molecule induction. From a pharmaceutical standpoint, CID systems are advantageous in this regard, since the utilization of small molecules or drugs is highly desirable in many cases.

Modifications to the CID systems as described in Section 2.2 are useful in the difficult problem of devising a system that is transparent to other functional machinery at a cellular and organismic level. The challenges here differ for CID implementations in therapeutic versus investigational contexts. If the protein used as the CID switch is native to the organism, the risk of inhibiting its endogenous function must be considered. Reciprocally, the sequestration of the dimerizer by intrinsic cellular proteins could prevent the activation of the engineered switch. In contrast, non-native switching proteins, while potentially free of the dual interference problem, raise the risk of

immunogenicity in a therapeutic context. Immunological responses to a foreign protein are clearly incompatible with the long-term viability of an introduced cellular switch.

To clarify, the discussed coumermycin-based dimerization represents an example of a system utilizing the bacterial protein GyrB, raising concerns about immunogenicity, whereas the FKBP-FRB systems are less encumbered by this, as they employ proteins of human origin. However, potential targets for an immune response still exist, such as the point mutations introduced to induce specificity for bumped dimerizers (such as the F36V mutation) or the junctions created between FKBP or FRB and the protein of interest. In any case, as the refinement and development of serum-stable, protein-based treatments advances, the opportunities for CID-based therapies will only increase.

3.1. Induced Signal Transduction and Gene Expression

Over the past decades, an increase in the understanding of cellular signaling pathways and gene expression has led to an emphasis on the control of cellular machinery and the utilization of endogenous cellular defenses to combat pathogenic processes. Biomedical efforts to exploit chemical inducers of dimerization as switching systems have advanced most swiftly in this field, an arena of science with vast therapeutic promise. One major advantage of the conditionality inherent to CID systems is circumvention of the problems related to “off-target effects”, which result from the failings of non-conditional therapeutics to meet the physiologically dictated thresholds for specificity of a drug for its target and the drug target in the pathogenesis of the

treated disease. A comprehensive review of this topic has been published elsewhere.[52] While clinical implementation of cell-based medicine is widespread, spanning a range from the transfusion of erythrocytes to bone marrow organ transplantation, our scientific ability to fine-tune the effects of these therapies is limited. To further realize the vast therapeutic potential of genetic intervention, novel mechanisms that can regulate or fine-tune these complex biological tools will be required.

By 2001, just eight years after their initial description, a number of studies had demonstrated the utility of CID in regulating cell based therapies.[53] Triggered cell proliferation at the receptor level has been used to enhance growth of genetically modified cells.[54-59] In one example, skeletal myoblasts, which can be used to repair scar tissue and infarcted myocardium, yet suffer from complications in reproducibly generating large enough grafts, are transfected with modified FKBP (F36V) fused to the fibroblast growth factor receptor-1 cytoplasmic domain. AP20187, a dimeric F36V ligand related to AP1903 (Section 2.2), induces dimerization of the chimeric receptor and hence, proliferation of the transfected myoblasts in a robust, reproducible manner. Additionally, after 30 days of treatment, myoblast grafts remained stable after withdrawal of the dimerizer, and differentiated normally, showcasing the reversibility of CID based systems.[60]

Therapeutically tailored activation of gene expression by transcription-regulating CID systems has also been demonstrated.[61-64] Quintarelli et al., in an elegant combination of antitumor therapies, showed that transgenic expression of interleukins 2 and 15 by Epstein-Barr Virus-specific cytotoxic T lymphocytes (CTLs), which are

tumor-specific, increased the antitumor activity of this well-studied system. The concomitant side effects of increased interleukin levels, including systemic toxicity, were mitigated by coupling this methodology with CID-inducible expression of a caspase-9 suicide gene, which afforded elimination of transgenic CTLs and interleukin production, increasing the therapeutic feasibility of the approach.[65] In keeping with the superb amenability of dimerization systems to fundamental characterization, detailed analyses of the thermodynamic characteristics required for optimal CID-based activation and induction have been carried out.[57,66]

The competitive evaluation of dimerizer-based systems versus alternative methods for regulated gene expression has also been undertaken. In these studies, CID-based systems have shown the highest degree of regulatory selectivity, and have consistently demonstrated outstanding dynamic range and sensitivity, particularly as the systems have evolved.[35,67,68] The work by Xu et al. compared five popular systems for regulating induced gene expression from adenovirus (Ad) vectors – the tetracycline (Tet-on and T-REx), ecdysone, antiprogestin, and dimerizer-based systems.[68] The dimerizer-based system tested is an FKBP-FRB complex induced by addition of AP21967, an analog of rapamycin with decreased immunodepressive effects. The DNA binding domain (ZFHD1 – a dual zinc finger domain from human transcription factor Zif 268) is fused to FKBP, while a p65-HSF1 activation domain is fused to FRB. Using a luciferase reporter assay in three different cell lines, it was found that the FKBP-FRB system maintained the lowest basal level of expression and the highest induction factor, properties paramount to the practical usage of gene expression as a therapeutic

technique. Additionally, dimerizer-based gene expression is independent from endogenous cellular processes and exhibits a 50-fold increase in expression at inducer levels as low as 1 nM. Lastly, the dimerizer-based system benefits from the expression of binding and activation domains as individual proteins, freedom from virus-derived proteins, and the possibility that expression can be even more tightly controlled via the addition of a non-inducing competitor of AP21967. Overall, the FKBP-FRB system represents a highly feasible method for the therapeutic application of targeted gene expression.

Significant progress has also been made in the engineering of dimerization based regulatory elements, optimizing the compactness and efficiency of constructs for potential therapeutic delivery.[28,69-71] For example, the modification of rapamycin to non-immunosuppressive agents that still efficiently bind FKBP and FRAP has found success in not only the study discussed above, but also in the regulation of viral replication in antitumor therapies. Traditionally, replication-defective viral vectors are used to deliver gene-based therapies; however, use of replication-competent viral vectors represents a more efficient delivery of therapeutic genes. Unsurprisingly, there is much hesitation to use replication-competent viral vectors, since uncontrolled replication could prove disastrous to the patient. Chong et al. showed that replication-competent vectors dependant on the presence of AP21967 improve the viral spread in tumor and surrounding tissues while allowing for temporal control of replication, thereby increasing both efficiency of gene delivery and safety to the patient.[71]

Growth-arresting CID therapies can also play an important clinical role:

conditional CID-aggregation of the pro-apoptotic Fas effector has served as an effective safety switch in graft versus host disease (GVHD), one of the most pressing issues in transplantation medicine, in which T-cells are transduced with a gene expressing a chimeric Fas intracellular domain-FKBP fusion.[24,72] While the T-cells are unaffected by the addition of the gene alone, treatment with dimerizer results in the Fas-mediated apoptosis of the cells, should GVHD occur. This technique has also shown promise in large animal models.[73] Globally, steady progress has been made in validating the therapeutic viability of CID regulation in humans – defining toxicities, demonstrating efficacy in primate models, monitoring for oncogenicity or other long-term consequences.[74-77] For example, work by Richard et al. showed that although CID-dependent association of a modified murine thrombopoietin receptor allows for *in vivo* selection of genetically modified hematopoietic cells in mice and dogs with little toxicity, similar work in a nonhuman primate model yielded little expansion of genetically modified red blood cells.[78] In contrast, more recent work by Nagasawa et al. employing the same methods showed *in vivo* efficacy in human hematopoietic cells, indicating that the nonhuman primate model may not be entirely predictive of results in humans, and that although human hematopoietic cells may behave differently in humans, clinical trials are necessary to ascertain such differences.[79]

3.2. Physical Inhibition of Protein Interactions

Protein association plays a pivotal role in the pathological functions of many diseases. From combating viral capsid assembly to inhibiting infectious agent

signaling derived from a transient protein association, the search for modulators of protein interactions is an important and relevant field of pharmaceutical research. However, the development of such small molecule inhibitors of protein association has proven elusive. Not only do the relatively large surface areas influencing protein association disperse binding energy, but these same topologies are flexible, allowing for accommodation of small molecules and lack of inhibition.[80] A lean but rapidly expanding field of CID research includes circumventing this issue by utilizing induced protein dimerization as a tool for increasing the steric bulk of the underlying small molecule dimerizer. Whereas the small molecule retains its favorable pharmacological properties, the self-assembled protein complex serves as the effective pharmacophore.

This particular mode of inhibition has found success in blocking Aβ aggregation, paramount to the development of Alzheimer's disease, by utilizing a heterobifunctional dimerizer capable of binding within the aggregating Aβ fibril and concurrently recruiting the cellular chaperone FKBP, physically blocking further Aβ interactions.[80] Gestwicki et al. have demonstrated the use of dimerizer-based affinity modulation as a means to inhibit the formation of the amyloid Aβ fibrils associated with Alzheimer's disease. Small molecule inhibitors of amyloid aggregation face stern challenges common to all efforts to inhibit protein-protein interactions: they are mediated by large, often flat, surfaces that offer very little purchase for a drug to take hold.[81] Given this problem of scale, a small molecule that does bind selectively to a protein-surface target may yet be unable to disrupt aggregation. Beginning with the knowledge that the dye congo red (CR) is able to bind amyloid with reasonable affinity,

and even prevent Aβ aggregation at high concentrations (IC_{50} = 2 μM), they reasoned that a SLF-CR hybrid would recruit FKBP to the surface of Aβ monomers and small aggregates, where it would serve as a chaperone to prevent formation of amyloid fibrils (Figure 8).

In a variety of biochemical and imaging experiments, they found that SLF-CR was able to inhibit amyloidogenesis in a dose-dependent, FKBP-dependent manner, with 5-6 fold greater potency than CR alone. More impressively, they also demonstrated that SLF-CR + FKBP was able to block the toxic effects of $A\beta_{1-42}$ in neuronal tissue culture, with an EC_{50} approximately 4-fold lower than that of CR alone. Seeking to optimize the potency of this effect, the authors speculated that the local dynamics of the FKBP chaperone at the Aβ surface could influence its efficacy, and synthesized SLF-CR conjugates with variable linker length (Figure 9). The most potent compound, SLFBenz-CR, was 40-fold more potent in the presence of FKBP than CR alone in Aβ aggregation assays, with an IC_{50} of 50 nM.

A similar method of therapeutic protein association lies within the “bait and trap” capabilities of bifunctional ligands in concert with ubiquitous endogenous proteins.[40,82,83] One such protein, human serum amyloid P component (SAP), is a member of the pentraxin family of innate immune system proteins and is one of the most abundant human serum proteins. Naturally existing as a pentamer, crystallographic studies show the association of two SAP pentamers into a dimeric, face-to-face complex via a network of noncovalent pi-stacking interactions. The cholera toxin B-pentamer (CTB) adopts a similar pentameric conformation that binds with high affinity to cell

Figure 8. Inhibition of amyloid aggregation. A) Association of FKBP which is recruited by the bound SLF-CR prevents further aggregation into amyloid fibrils. B) Chemical structure of SLF-CR. Figure reprinted from Gestwicki et al.[80] with permission. Copyright (2004) AAAS.

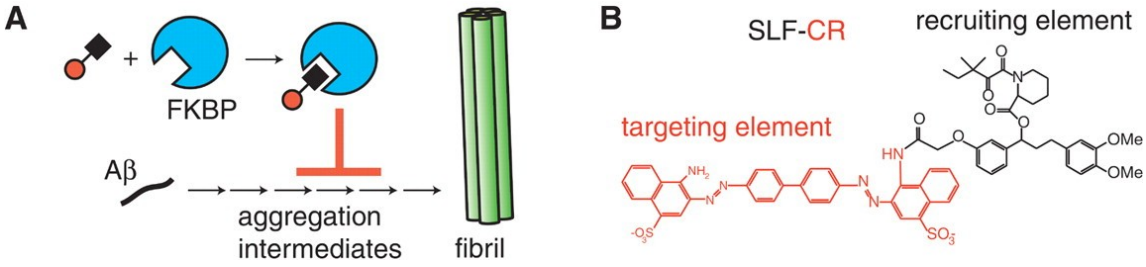
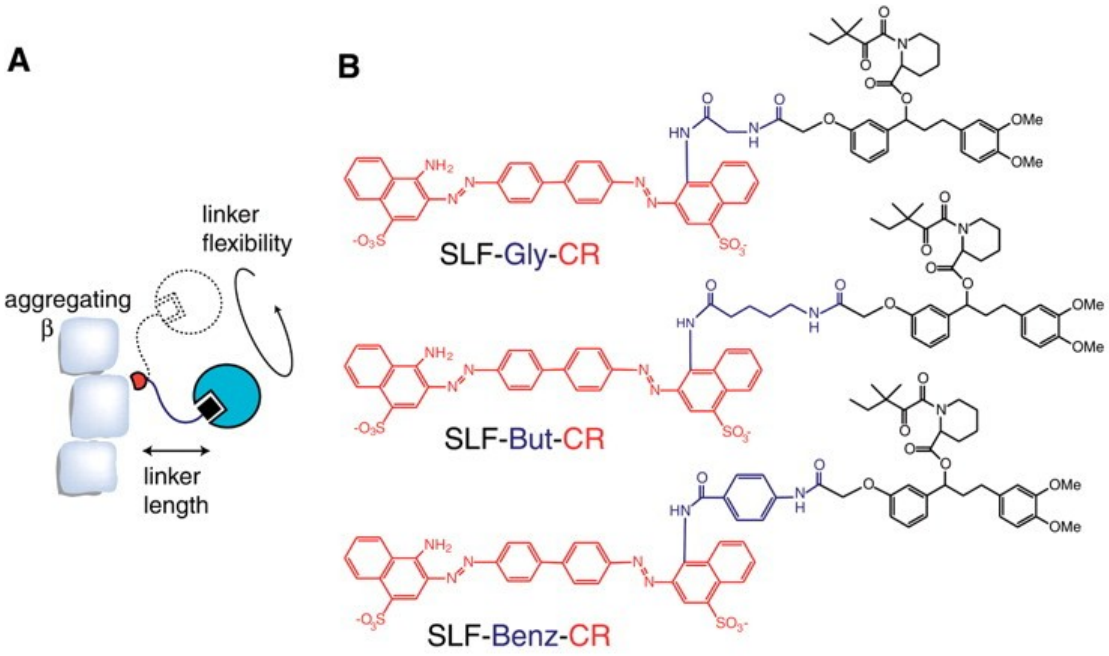


Figure 9. A) Model for potential role of linker dynamics in inhibiting Aβ aggregation. B) SLF-CR conjugates with varied inter-pharmacophore linkers. Figure reprinted from Gestwicki et al.[80] with permission. Copyright (2004) AAAS.



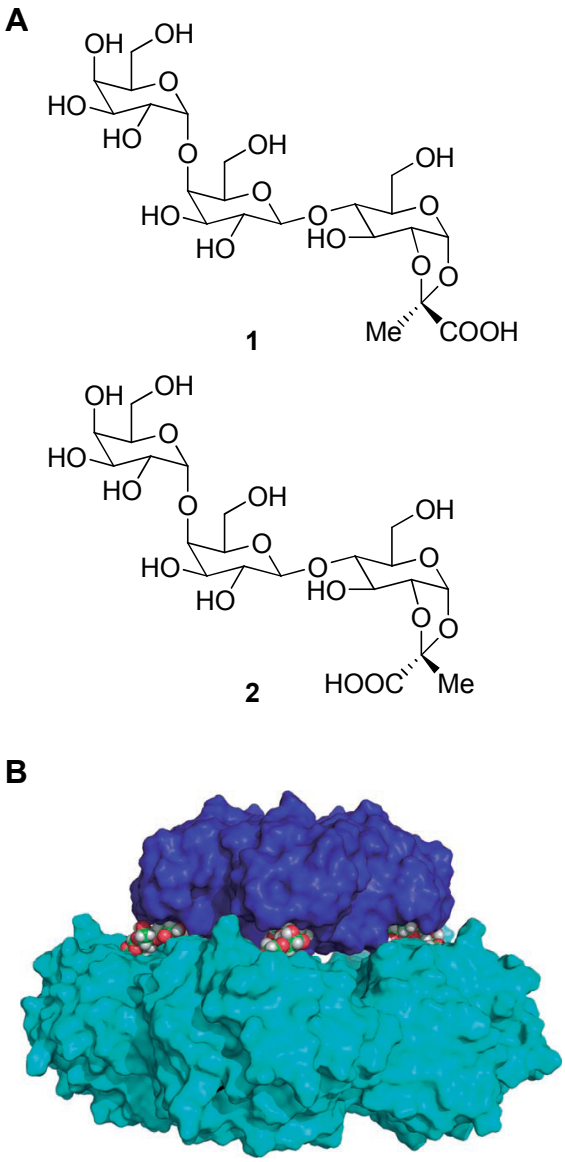
membrane carbohydrates. Inspired by the similarity between the two structures, Liu and coworkers developed a heterobifunctional inducer of dimerization exploiting the carbohydrate binding domain of CTB and the proline binding pocket of SAP linked via an ethylene spacer. Dynamic light scattering data indicated the stable formation of a ternary SAP:ligand:CTB complex, and further competition experiments noted the increased stability of the complex relative to monovalent binding, with low micromolar IC_{50} values at physiological SAP concentrations.[40]

Expanding upon this work, Soloman et al. developed a similar system in which the baited protein was the Shiga toxin (Stx) pentamer.[82,83] A bifunctional ligand was constructed featuring binding domains recognized by the proteins of interest (Figure 10). The trisaccharide moiety binds to Stx, whereas the cyclic pyruvate moiety binds to SAP. Gel permeation chromatography, in conjunction with dynamic light scattering data, indicated the formation of a Stx:ligand:SAP complex, and *in vitro* binding experiments determined that, in the presence of 0.17 μM **1**, SAP inhibits Stx with an IC_{50} of 21 nM. Furthermore, compound **1** inhibited Vero cell death via a lethal dose of Stx with an IC_{50} on the order of 15 μg/mL.[83] However, compound **1** was shown to be inactive against a human SAP transgenic mouse model due to a high degree of clearance. Currently, studies are underway to alleviate these issues.

4. Protein Nanostructural Assembly

To this point, we have considered chemically induced dimerization as a means to control the assembly of small-scale functional complexes: initiators of transcription

Figure 10. The bait-trap mechanism of toxin neutralization. A) Heterobifunctional ligands exploiting trisaccharide (Stx) and cyclic pyruvate (SAP) binding pockets. B) Molecular model of the SAP:ligand:Stx ternary complex. Stx – top, blue, SAP – bottom, turquoise. Figure reprinted from Kitov et al.[83] with permission. Copyright (2008) Wiley-VCH Verlag GmbH & Co. KGaA.



or signal transduction, inhibitors of protein aggregation, or localized active enzymes. The ability, however, to precisely connect patterned proteins is reminiscent of another important biological role for protein assemblies: structure. From the nanoscopic to macroscopic scale – from viral capsids to vertebrate muscle – proteins play a crucial role in biological structures. Moreover, many natural protein assemblies operate on a scale which is at present poorly imitated by synthetic approaches – too large for synthetic organic chemistry, too small for techniques of microfabrication.[84] This 1-100 nanometer niche, occupied so successfully by biomaterials, offers a fascinating range of scientific possibilities: from advanced protein therapeutics, to proverbial nanobots, to next-generation electronic devices.[21,85,86]

Toward this end, efforts are underway to devise techniques and unearth principles regulating the assembly of a variety of biomaterials.[62,87] Nucleic acids have been a principal building material for these efforts, as sequence recognition can both encode and construct robust structural junctions.[2,88,89] The greater structural and functional diversity of proteins offer a potentially rich opportunity for protein-based materials. In contrast to DNA or RNA, simple and elegant mechanisms for directing their assembly are lacking. In this important niche, chemically induced dimerization may play a role.

Early efforts to engineer synthetic protein assemblies have principally focused on purely protein-mediated connectivity – the association of polypeptide building blocks with intrinsic affinity, such as multimeric proteins, β-strands, coiled coils, and zinc fingers.[90,91] By fusing naturally multimeric proteins, self-assembling building

blocks that produce filaments (linear heterodimers) and cages (polyhedral vertices) have been designed.[92] A second principal strategy in successful designs of these materials has been the reorganization of bundled protein domains so that they contain intermolecularly self-complementary structures, such as coiled coils with staggered ends, or helix bundles reorganized and extended for a domain-swapping interaction.[93,94] These building blocks typically form large filaments, as the primary protein strands bundle into larger structures. Particularly intricate design efforts have been conducted and delineate several strategies to functionalize these filaments.[95,96] For example, recent work by Rele et al. has featured the development of D-periodic collagen-mimetic microfibers, in which tripeptide assembly is guided into an ordered nanostructure via electrostatic bias.[97] These methods are limited by the fact that the assembly of self-complementary elements generally occurs spontaneously *in situ* as the protein is expressed. In addition, once assembled, there is no general means to remodel or disassemble the structures.

A potential enhancement to the method of spontaneous protein self-assembly is the addition of ligand control. A few examples have appeared in the literature outlining these methods that rely on the association of multivalent ligands with multivalent proteins. If chemically induced dimerization represents the association of a bivalent ligand with monovalent binding proteins, a self-limited event, then the transformation of those binding proteins into bivalent (or multivalent) molecules creates the potential for expansion into larger scale structures.

Dotan, Freeman, and coworkers designed and observed diamond-like protein crystals constructed by the assembly of tetravalent lectins (carbohydrate binding proteins) and bivalent mannose derivatives.[98] Careful addition of the bismannopyranoside at 2:1 stoichiometry noncovalently crosslinked concanavilin A into a tetrahedral lattice, producing crystalline protein precipitates (Figure 11). This scaffold offers intriguing potential for further engineering of functional, three-dimensional protein networks. A disadvantage to the technique, however, is the inability to precisely regulate the degree of lattice assembly, as is the lack of a means to form soluble nanostructures.

Ringler and Schulz explored a more regulated route to ligand-mediated nanostructure assembly in their work with biotinylated building blocks.[99] They produced protein nanostructures based on the ligand-directed assembly of a biotin-tagged tetrameric aldolase and streptavidin (Figure 12). Biotin and bis-biotin mediated connections between the subunits enabled these constructs to be directively assembled into two dimensional lattices and varied cruciform discrete nanoassemblies. These spatially ordered, ligand-assembled constructs offer prospects for more elegant assemblies of nanoscale protein biomaterials. Primary limitations to this approach include the restriction to rectangular arrays, the unknown stability of the streptavidin-aldolase building blocks, the difficulty in forming homogeneous complexes, and the complex process required to prepare these structures.

Figure 11. Ligand-mediated assembly of a lectin nanocrystal. Tetrahedral lectin crystals: at left, the crystal structure of concanavilin A (PDB ID: 5CNA) bound to four monovalent sugars; center, the assembling complex crosslinked by the bivalent bismannopyranoside; right, a single element is outlined within a TEM image of the tetrahedral lectin crystal. Proteins are rendered in VMD from PDB coordinates 5CNA. TEM image reprinted from Dotan et al.[98] with permission. Copyright (1999) Wiley-VCH Verlag GmbH & Co. KGaA.

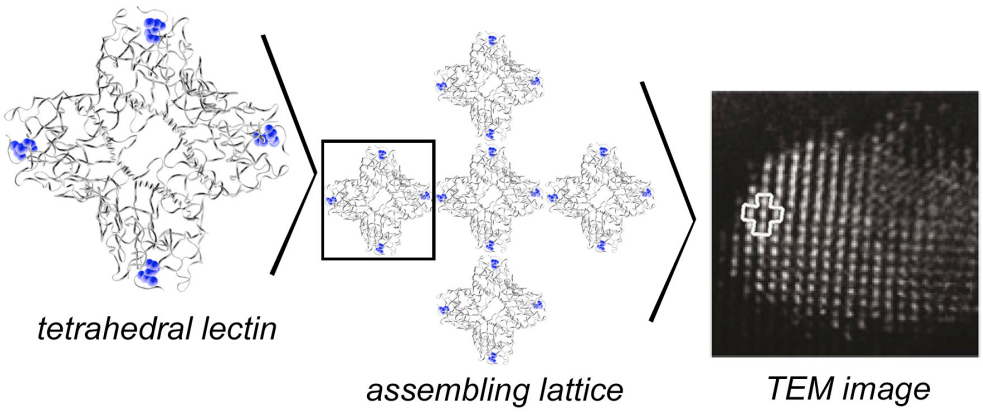
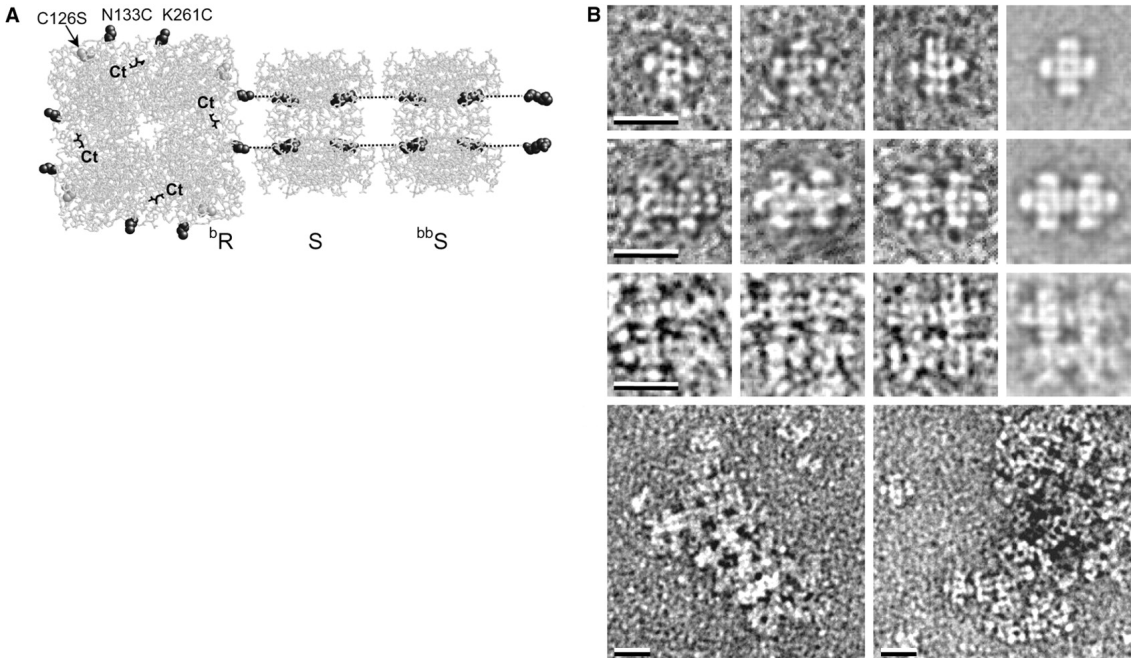


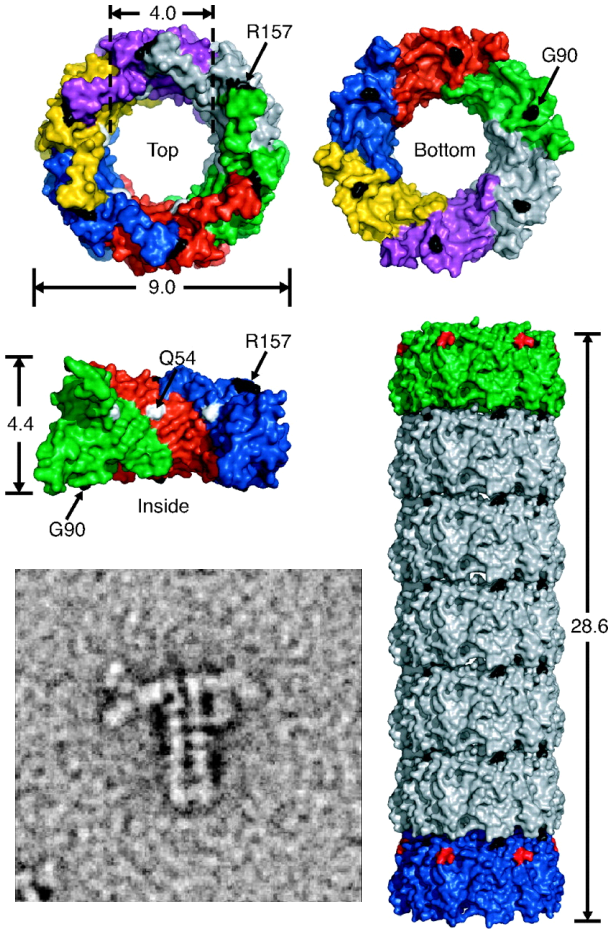
Figure 12. Biotin-linked streptavidin-aldolase nanoarrays. A) The engineered tetrameric aldolase (bR), with eight cysteine residues, is biotinylated on the right-hand face and linked to streptavidin (S). The streptavidin building block can in turn couple with the bbS unit, a streptavidin moiety pre-assembled with four bis-biotin residues. B) Transmission electron micrographs of assemblies at four discrete stages of construction. Figure reprinted from Ringler et al.[99] with permission. Copyright (2003) AAAS.



Another quickly evolving area of nanostructural assembly includes the development of nanotube structures for a wide range of uses in the broad spectrum of nanoarchitechure. Aside from carbon and inorganic nanotubes, which lack the functionality of higher-order protein structures, current research in the area focuses on naturally occurring polymeric proteins such as viral capsids, actin, tubulin, amyloid protein, and flagella. Limitations with this methodology include a lack of control over *in vivo* assembly, as well as difficulties in accessing the entire nanotubule topology.[100]

In order to work around such problems, Ballister and coworkers have established a method of developing self-assembled, tailorable nanotubes from the toroidal hexameric protein Hcp1 from *Pseudomonas aeruginosa*. Systematic cysteine replacements at strategic locations along the faces of the toroidal components allow for covalent disulfide bonding between the subunits of the tubular assembly (Figure 13). Whereas size exclusion chromatography indicated the presence of high molecular weight protein complexes in solution, repeating the experiment in an excess of reducing agent (5 mM dithiothreitol) resulted in an absence of nanotube formation. Although there is a lack of a traditional multivalent ligand in this particular system, the engineered disulfide bonds serve as the chemical switch paramount to any CID system. In addition to the controlled assembly of Hcp1 nanotubes, chain termination and tube sealing was achieved using the same methodology, with cysteine mutations at the internal face of the Hcp1 toroid. This excellent level of structural control affords the self-assembly of an Hcp1 nanocapsule, with an internal environment isolated from that of the bulk solvent.[100]

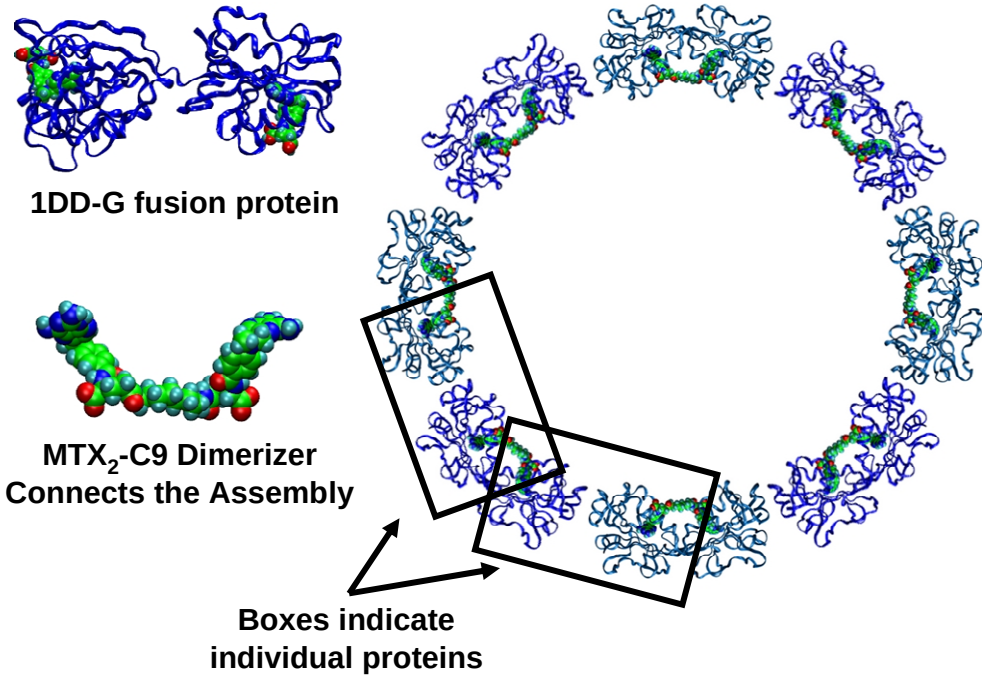
Figure 13. Protein nanotube architecture. The surface plots depict the topology of the designed nanotubes. The TEM image displays the capped nanotube. Figure elements reprinted from Ballister et al.[100] with permission. Copyright (2008) National Academy of Sciences, U.S.A.

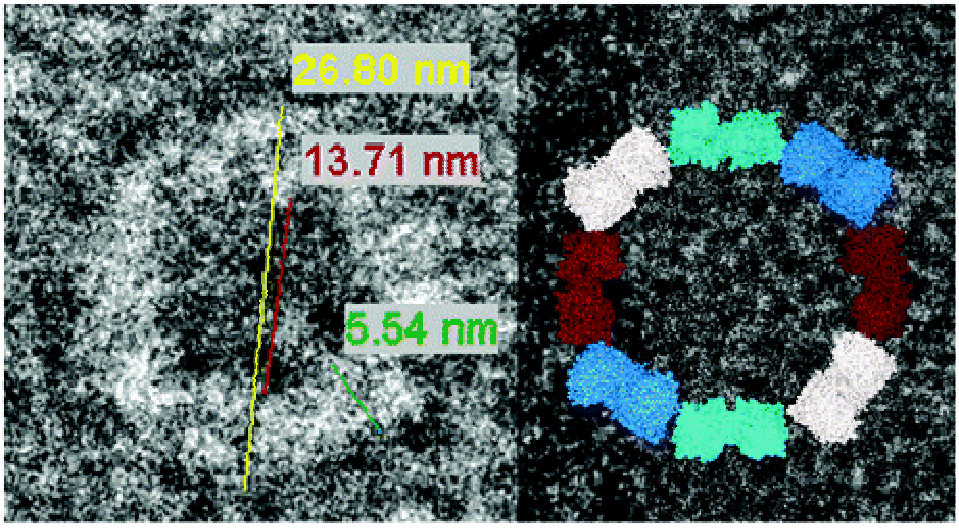


One aspect of protein nanostructure assembly given considerable attention is the increased kinetic and thermodynamic stability of multivalent complexes.[101-104] An example of this stability is described by the synthesis of a trivalent hapten that causes the aggregation of rat anti-2,4-dinitrophenol (DNP) IgG (IgG^{DNP}) into bicyclic trimers.[105] When three 2,4-DNP moieties are tethered to a central amine via a polyethylene glycol spacer, the new compound effects the formation of trimeric IgG^{DNP}s (characterized by dynamic light scattering, analytical centrifugation, and size exclusion HPLC) with the stoichiometry IgG_3:$ligand_2$. Moreover, further experimentation with a competitive inhibitor indicates the relative stability of the complex is 225-fold greater than that of a singly bound species. This increase of stability is a hallmark of multivalent protein structures, and represents an advantage in biotechnological applications.

Despite recent work advancing the development of self-assembled nanostructures, the problems of structure polydispersity and incomplete assembly still remain. From a practical therapeutic standpoint, the ideal nanostructure is both assembled and disassembled in a controlled manner. Whereas previous work has focused on assembly and the formation of stable complexes, controlled disassembly has often been overlooked. Previous work shows the ability of a bivalent methotrexate dimerizer to induce the dimerization of *E. coli* DHFR, which is further stabilized by protein-protein interactions at the DHFR-DHFR interface.[31,32] Further work by Wagner and coworkers has focused on the use of fusion proteins of DHFR (Figure 14, $DHFR_2$) to form stable, self-assembled protein nanorings where polydispersity is tunable based

Figure 14. Protein nanoring structure. Upper limb) Schematic of protein nanoring structure and assembly. Lower limb) TEM of DHFR nanorings. Figure adapted from Carlson et al.[106] with permission. Copyright (2006) American Chemical Society.

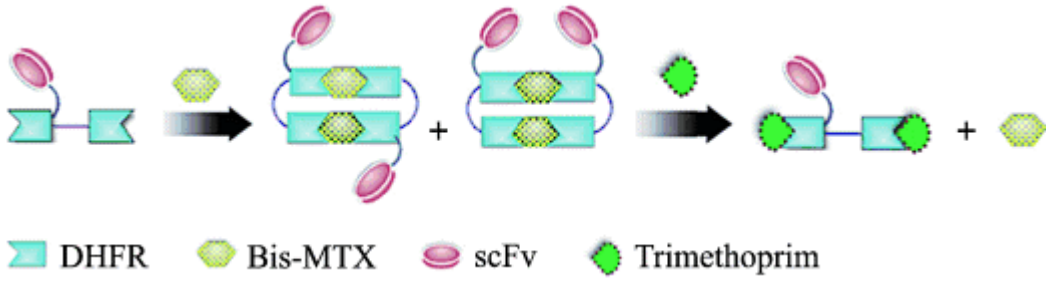




on the nature of the amino acid linker between the two proteins.[106] Light scattering analysis, in conjunction with size exclusion chromatography and TEM data, confirm the formation of toroidal species in a small window of sizes. Advantageous to this method is the pharmaceutically reversible assembly of the complex, since an excess of an inhibitor of DHFR will effect dissolution of the ring. Current limitations include the ability to exert a level of control over protein integration; however, recent work indicates the possibility of dimer interface modulation as a candidate for the development of a biomolecular language.[32]

Revisiting multivalency, further exploration from the perspective of protein rather than ligand modification has recently yielded single chain antibody (scFv) – $DHFR_2$ fusions which, when assembled into nanorings, have been used to produce divalent antibodies (Figure 15).[107] The antibody chosen binds the T-cell antigen, CD3ε, of the human T-cell receptor. The nanoring containing two copies of the single chain antibody was shown to have a comparable dissociation constant to the parent monoclonal antibody. Confocal laser microscopy was employed to show that the antibody nanorings interacted with CD3+HPB-MLT cells in a similar manner to the parental antibody. Moreover, the complexes were found to undergo controlled disassembly via the addition of the DHFR competitive inhibitor trimethoprim. The formation of larger rings in the work described in the previous paragraph leads to the vision of rings with higher valencies and avidities. Indeed, further work by Wagner and coworkers has shown that antibody nanoring sizes mimicking cellular receptor cluster topography (i.e. octomeric in nature) have led to increased avidity of fused antibodies to

Figure 15. Cartoon representation of divalent antibody nanorings. Figure reprinted from Li et al.[107] with permission. Copyright (1999) Wiley-VCH Verlag GmbH & Co. KGaA.

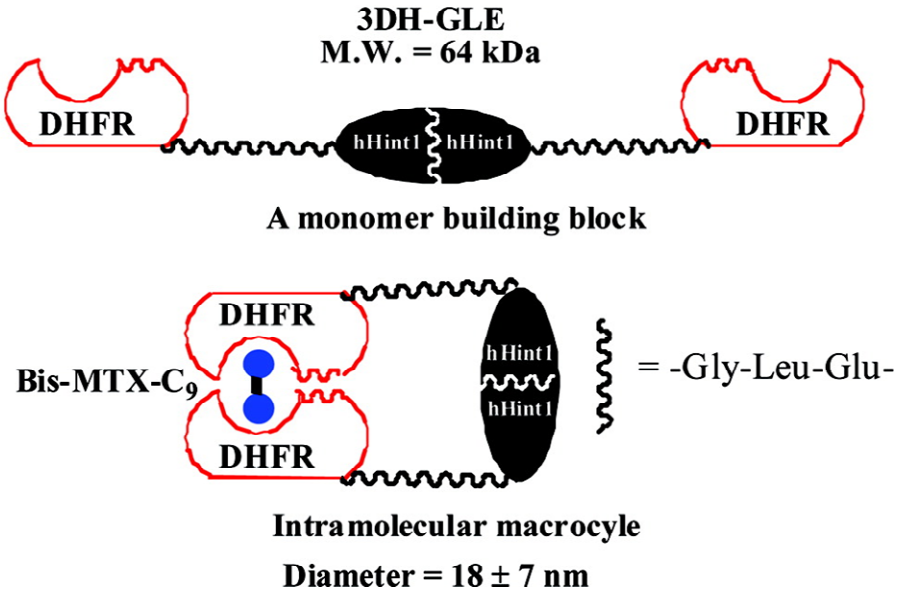


their targets (unpublished data). Additionally, modified dimerizers capable of delivering fluorophores, radionuclides and drugs to targeted tissues would expand the biotherapeutic applications of the nanorings.

Concerning the development of mixed protein nanostructures and a technique for implementing control over nanostructure self-assembly, recent work has shown it possible to form enzymatically active nanorings using DHFR-histidine triad nucleotide binding 1 (Hint1) fusion proteins.[108] Human Hint1 (hHint1) forms a stable homodimer and acts as a phosphoramidate and acyl-adenylate hydrolase. By incorporating hHint1 into the rings, functionally active, protein-based nanorings may be produced (Figure 16). The length of the linker between the DHFR and hHint1 was shown to have profound effects on the size of the rings formed. When the linker was consisted of Gly-Leu-Glu, a range of rings containing between 2 and 12 monomers was observed by size exclusion chromatography. The molecular weights of these rings ranged from 130 to 740 kDa. The largest rings (>11 monomers) were shown to have the greatest specific activity while the smaller rings were found to have activities greater than or comparable to the wild type hHint1.

In spite of this early progress, significant hurdles remain to be surmounted for protein-ligand nanoengineering to emerge from its nascent state. The ability to construct a greater geometric variety of structures is a high priority, as is the ability to exploit the ligand-reversibility of the assembly process in diverse environments. Tunable protein architectures, in which conformational assembly information can be encoded in the

Figure 16. Schematic of DHFR-hHint nanoring building blocks and macrocyclic dimer. Figure reprinted from Chou et al.[108] with permission. Copyright (2008) American Chemical Society.



primary structure, akin to the methods used for DNA and RNA, would represent another significant advance.

5. Theoretical Modeling of Intermolecular Interactions

In the 30 years since the first classical MD simulation was published,[109] computational modeling of molecular systems has increased exponentially in its application and efficacy.[110-114] Computational modeling of molecular systems can be roughly divided into two major categories – quantum mechanical (QM)[115,116] techniques and molecular mechanics (MM)[117-121] methods. The major practical difference between these approaches is the level of parameterization required to generate the potential energy surface associated with a molecular system. Whereas MM parameter sets and functional forms are optimized to reproduce an empirically observed data set and are thus limited in accuracy when applied more broadly, most QM methods do not require such parameters. This benefit comes at the price of computational cost, however, so QM methods are generally limited to small molecules or exploratory studies on larger systems.

In an effort to balance the tradeoff between accuracy and computational expense, several hybrid methods have been developed. Semiempirical molecular orbital theory (SEMO)[122-124] is one such level of QM that is rooted in the Hartree-Fock formalism, but makes a variety of approximations and obtains parameters from empirical data. Another technique for studying macromolecular systems includes QM/MM[125], in which a critical region or active site is treated with QM, while the

surroundings are treated with a suitable MM force field. Lastly, although not technically a hybrid method, the utilization of QM to find suitable parameters for an existing MM force field remains a viable option to increase the accuracy of a MM simulation, though any problems with the functional forms of the MM force field are not alleviated via this method.

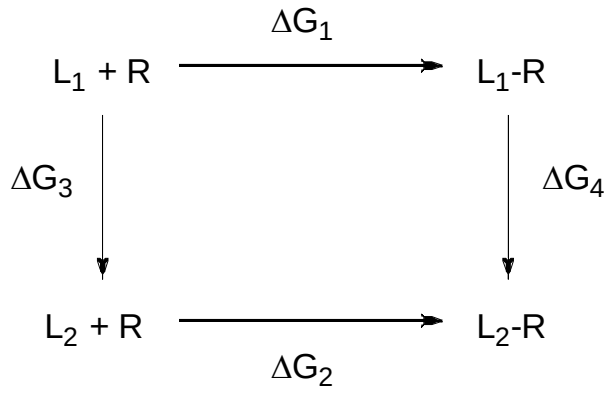
While QM treatment of a macromolecular system may become economically feasible in the future as computing power increases, the current dilemma of accurately modeling protein-ligand and protein-protein complexes remains significant, and most studies are limited to the MM environment. Given this constraint, considerable effort has been invested to increase the reliability and usefulness of MM calculations, especially in the field of biological and pharmaceutical chemistry. Developments in combinatorial chemistry and rational drug design have driven the efforts of theoretical chemists to refine MM methods describing protein-ligand interactions. Methods such as molecular docking[126] are frequently used to predict ligand binding to a protein target. Rooted in MM, molecular docking checks the fit of a ligand molecule to a particular active site. The complex can incorporate a diverse array of structures including small molecules, nucleic acids, enzymes, and peptides. Molecular docking ranks compounds by calculated binding affinities, which are prioritized by means of a scoring function. Typically, docking employs a shape-complementarity method, in which a protein structure is obtained and the active site defined.[127] The active site is then presented with a set of ligands, each of which is posed within the binding site. Scoring functions can take into account many different factors; however, the most common determinants of

ligand binding are electrostatic, hydrophobic, and van der Waals interactions between the ligand and protein.[128] Although inexpensive and rapid, molecular docking remains a coarse estimate of intermolecular interactions, due in part to a lack of phase space exploration and many local minima in the potential energy landscape.

A more elegant approach to model the behavior of intermolecular interactions is to directly estimate the free energy of the system. The equilibrium properties of a given system, ranging from phase distribution to association/dissociation constants, all rely on the free energy differences between two alternative states. Though computationally more intensive than molecular docking, Monte Carlo (MC)[129] and MD[130,131] simulations of macromolecular systems that adequately sample phase space and higher-energy states are becoming more and more feasible due to the development of newer, more advanced computing resources. Common methods for estimating free energy from MC or MD sampling include free energy perturbation[132], umbrella sampling[133], and molecular mechanics – Poisson-Boltzmann (Generalized Born) surface are (MM-PB[GB]SA) calculations.[134,135]

Free energy perturbation, while introduced by Zwanzig in 1954, was first applied to a problem of ligand binding by Lybrand, et al. in 1986.[136] The technique relies on the thermodynamic cycle for binding different ligands to a receptor as depicted in Figure 17. Since free energy is a state function, its value around a thermodynamic cycle must be zero. Therefore, in the cycle shown, $\Delta G_2 - \Delta G_1 = \Delta G_4 - \Delta G_3$, where the relative binding affinity of ligands L_1 and L_2 is $\Delta G_2 - \Delta G_1$, ΔG_3 corresponds to the free energy difference of the two ligands in solution, and ΔG_4 is the free energy difference

Figure 17. Thermodynamic cycle for ligands L_1 and L_2 binding to receptor R.



between the molecular complexes. Since free energy relies on the endpoints of these transitions and not the pathway between them, it is possible to modify the simulation (e.g. by changing atomic parameters) in any fitting way to achieve the transformation of the ligand. Lybrand, et al. were able to estimate the relative binding energies of Cl^- and Br^- ions to the synthetic macrocycle SC24 to within 0.15 kcal/mol by "mutating" the chloride ion to bromide both in explicit solvent and inside SC24 (also in solution). Bash and colleagues performed an early estimation of the relative binding free energy of several peptidomimetics to thermolysin by mutating atoms within the ligands.[137] Although their initial study yielded results in good agreement with experiment ($\Delta\Delta G_{exp}$ = 4.1 kcal/mol, $\Delta\Delta G_{calc}$ = 4.2 ± 0.5 kcal/mol), it was later shown that the results depended heavily on the charge model used to perform the calculation, highlighting the importance of proper force field parameterization in the MM environment.

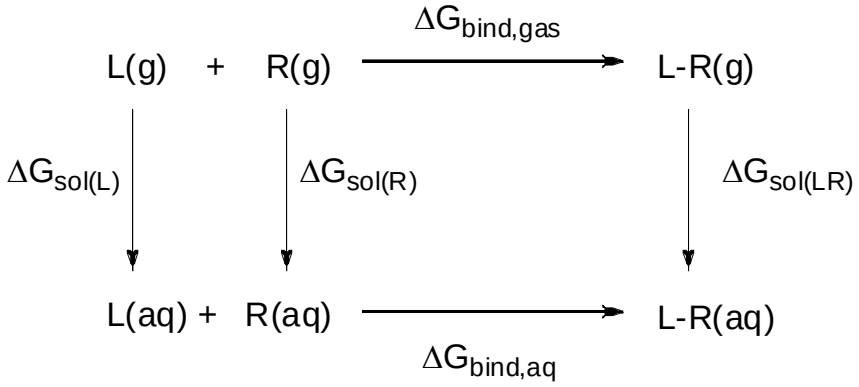
Revisiting Figure 17, the question may be raised as to why ΔG_1 and ΔG_2 cannot be simulated directly using MC or MD sampling. In short, if two molecules are simulated in solution, MC or MD will not adequately sample the phase space between the bound and unbound states, leading to poor free energy estimates due to a lack of sampling of high-energy structures. Although this is not a problem for very small systems in which minima in the free energy landscape are well defined, it remains a significant issue for more complex systems. Umbrella sampling attempts to resolve this issue by modifying the simulation so that high-energy structures are adequately sampled. By choosing a particular reaction coordinate (usually atom-atom distance) and restraining a simulation to that coordinate, high-energy structures (i.e., a ligand partially

situated in the binding pocket of a protein) may be sampled. In practice, umbrella sampling is performed in overlapping windows along a reaction coordinate, so that phase space is adequately explored and a measure of the free energy along the reaction coordinate (also known as the potential of mean force) can be constructed by accounting for the restraint using a weighting function. Early uses of umbrella sampling involved the calculation of the barrier of rotation between *trans* and *gauche* conformers of butane by Jorgensen, et al. and this technique has since also been applied to protein-ligand binding.[138,139]

While both free energy perturbation methods and umbrella sampling represent powerful methods for estimating the binding energy of two molecules, both remain time-consuming and computationally intensive due to the high number of simulations that are necessary to adequately sample phase space. The MM-PB(GB)SA method was developed to overcome this roadblock by taking advantage of another thermodynamic cycle (Figure 18) wherein the free energy of binding in the aqueous phase can be derived from the solvation and gas-phase energies of the components. MM-PBSA exists primarily as a postprocessing technique in which a single simulation of the protein complex in explicit water is obtained and analyzed by extracting the coordinates at each step in the simulation for each piece of the complex. Using MM to determine the gas-phase energies and the Poisson-Boltzmann equation (or generalized Born, in GBSA) to estimate solvation energies, an estimate of $\Delta G_{bind(aq)}$ may be obtained via equation 13:

$$\Delta G_{bind,aq} = \Delta G_{bind,g} + \Delta G_{sol(LR)} - \left(\Delta G_{sol(L)} + \Delta G_{sol(R)}\right) \qquad (13)$$

Figure 18. Thermodynamic cycle of ligand binding to a receptor in the gaseous and aqueous phases.



Due to its lower computational demand, MM-PBSA (and -GBSA, which requires even less computing power while generally achieving similar accuracy) is widely used for the study of protein-ligand and protein-protein interactions.[140-143] Despite the benefits of a single simulation, entropic contributions are expensive to calculate and introduce significant error into the estimated energies. Additionally, correlation in MD generated structures (due to the deterministic nature of the simulation) can lead to error in estimated energies, requiring significantly increased simulation time to reduce the error to an acceptable range.

While a broad selection of tools exists to estimate the binding free energy of intermolecular association, the problems of parameterization and adequate sampling remain, especially when working with a complex molecular system such as a protein-inhibitor or protein-protein complex. However, currently available computing tools are able to balance accuracy, expense, and speed to adequately estimate equilibrium properties for most users' needs, and pure QM treatment of a macromolecular system becomes increasingly feasible with continued advances in computing technology.

6. Conclusions

The field of chemically induced dimerization, while young, is rapidly expanding into its own niche in the current spectrum of biotechnological applications. The use of CID systems as investigational tools has led to the elucidation of the role of protein interactions in a number of biological events including, but not limited to, signal transduction cascades and transcriptional control. A number of studies have also shown

that it can be exploited to control the levels and post-translational structural modifications of proteins. In the future, an expanding number of ligand-protein dimerization systems which can be orthogonally activated will allow the ability for the study of increasingly complex systems.

Whereas application of CID technology to therapeutics is currently limited to controlled gene expression, signal transduction, and protein oligomerization, the recent developments in the area of protein interaction disruption utilizing endogenous cellular protein is an exciting area indeed. Continued research in this area could focus on any number of diseases requiring specific protein-protein interactions. With advances in lead-based drug design and molecular docking algorithms, the discovery of new protein targets and relevant binding domains is ever-increasing. The marriage of these phenomena holds great promise for the future of pharmaceutical development, since the use of small molecules to effect changes in protein interactions has long eluded researchers.

Similar to therapeutic applications of CID, nanostructural assembly is just beginning to see refinement in technique and application. Advancements in the understanding of the increased kinetic and thermodynamic stability of multivalent complexes have been encouraging. Perhaps most importantly, recent developments in control over protein assembly hold great promise for future endeavors in this field. Molecular recognition, while hinted at in some of the cases discussed above, still has yet to come into its own. With the goal of forming a self-assembled, switchable, and multifunctional system of limited polydispersity, further research concerning the

formation of a biomolecular language is required in order to effect the level of control necessary for such an endeavor. With such a language in place, it would be possible to use proteins as self-assembled building blocks, incorporating functionality as the researcher sees fit. Armed with such a wide array of highly refined tools, the possibilities for the application of chemical induction of protein association seem limitless.

Lastly, in the context of chemically-induced dimerization, the problem of accurately modeling intermolecular interactions in a complex system attains a new layer of complexity. Whereas accurate complexation free energy between proteins in multimeric complexes has been reported, the literature is sparse pertaining to protein—small molecule—protein modeling. While this may be a result of the relatively novel field of chemically induced dimerization, the fact remains that significant effort is still required to derive methods for adequately modeling the behavior of a CID-based complex.

Chapter Two

Toward the Simulation of the DHFR-DHFR Interface: Establishing Parameters for Methotrexate

1. Introduction

Continuing advances in molecular modeling and computational chemistry have greatly facilitated the structure-based design of small-molecule inhibitors of proteins.[110-114,144-153] Although molecular mechanics (MM) force fields[117,119-121,154] can model protein structure, they often lack parameters that accurately represent the heteroatomic groups present in pharmaceuticals.[155-157] Density functional theory[115] (DFT) and wave function theory (WFT)[116] do not require new parameters for each type of atom; however, current technology still limits the calculations to smaller molecules and exploratory studies on larger systems. Two viable approaches for simulating a protein bound to a drug-like inhibitor are to obtain MM parameters for force fields that yield accurate molecular geometries and partial charges or to find a suitable level of combined QM/MM theory[125,158,159] in which a critical or active region of the system is treated by quantum mechanics (QM) and the surroundings by MM. An economical QM level for such calculations would be semiempirical molecular orbital[122-124,160,161] (SE-MO) theory. For small enough QM regions or short simulations, one can also use more reliable QM methods such as DFT. Reliable WFT calculations are, however, affordable only for the smallest systems.

A recent application of molecular modeling is the prediction of mutation effects on protein-protein interactions.[58,106,162-167] Protein multimer stability can be modified through the introduction of interfacial residue mutations, and it would be valuable to be able to predict the relative change in stability of a mutated protein multimer compared to the wild-type species. Such calculations would aid in understanding the functional

evolution of proteins as well as the development and control of stable, self-assembled protein structures, with applications ranging from nanoscale multiprotein constructs to drug delivery. With the advent of chemically-induced dimerization (CID), our laboratory has demonstrated the ability to create self-assembled *E. coli* dihydrofolate reductase (DHFR) dimers from naturally existing DHFR monomers using a bivalent methotrexate dimerizer (MTX_2, complex – $DHFR_2MTX_2$) (Figure 1).[32] We have found that the introduction of complementary interfacial mutations putatively leads to a stabilized DHFR heterodimer, which allows for a level of control over the assembly of such constructs.

One complicating issue present in our system as well as other biological systems is the protonation state of the ligand in solution and in complex with the protein. While the DHFR inhibitor MTX is unprotonated in solution, it is protonated on N1 (Figure 2) when bound to DHFR.[168,169] This raises the question of whether it is appropriate to use a single set of MM parameters to describe MTX both in solution and bound to the enzyme. To assist the *in silico* prediction of mutation effects on dimer stability, we have undertaken a study to develop a set of MM parameters that can accurately model the $DHFR_2MTX_2$ complex. To accomplish this goal, we will first try to establish an accurate method to model the single substrate, MTX, and then attempt to extend this method to the $DHFR_2MTX_2$ complex. During this process, we have tested a large variety of methods on a set of drug-like molecules containing nitrogen heterocycles and exocyclic amino groups, and the results of these tests are presented herein because they should be of general interest for a variety of potential applications.

Figure 1. $DHFR_2MTX_2$ chemically induced dimer.

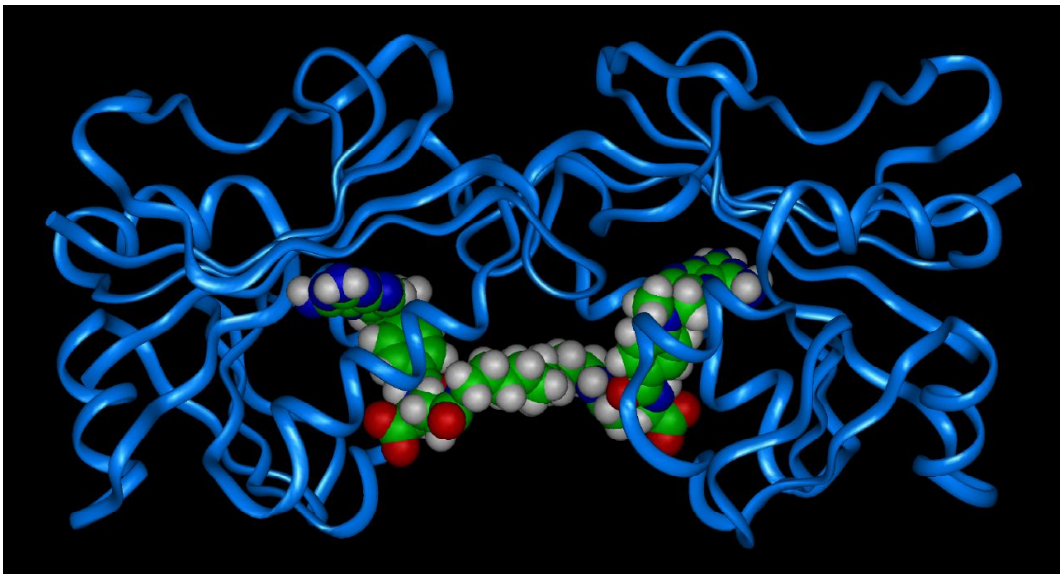
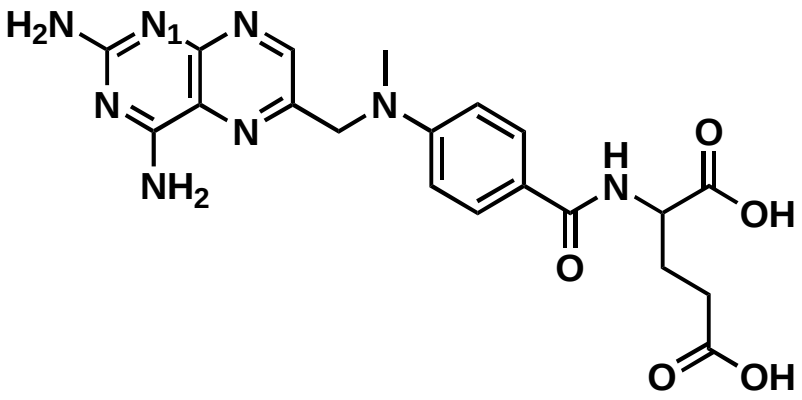


Figure 2. Chemical structure of methotrexate highlighting the position of N1.



A critical issue in simulating systems with nitrogen heterocycles is modeling the charge distributions. Since partial charge distributions are not experimental observables, we will rely on theory to establish reasonable values. For this purpose, we first require accurate geometries, and we begin by establishing benchmark values utilizing high-level WFT and DFT calculations on the MTX fragments 2-(aminomethyl)pyrazine (2-AMP), 1*H*-2-(aminomethyl)pyrazine (1*H*-2-AMP), 2,4-diaminopyrimidine (2,4-DAP), and 1*H*-2,4-diaminopyrimidine (1*H*-2,4-DAP). These fragments were chosen because of the role of the pteridine moiety in MTX binding to the DHFR active site. We then use DFT with class IV charges[170] to establish benchmark partial atomic charges. Coupled cluster theory[171,172] with single and double excitations (CCSD) and the M05-2X[55] density functional with the 6-31+G(d,p)[173] basis set are used for geometries, and charge model 4[174] (CM4) is used for partial atomic charges. The performance of widely available SE-MO and MM parameterizations is then surveyed for these four fragments to find the parameterized model that most accurately predicts the geometries, binding energies to water, and charge distribution of the unprotonated and protonated states of 2-AMP and 2,4-DAP. Additionally, we consider the selected MM methods when CM4 charges are substituted for the force field's default charges in an effort to observe if increased accuracy in partial charge distribution leads to increased performance in geometric and energetic modeling. The most accurate methods are then used to calculate partial charges and geometries for a series of pharmacophorically similar molecules containing nitrogen heterocycles and exocyclic amines, and these results are compared to DFT and CM4 benchmarks to

explore the validity of our chosen MM parameters more broadly.

2. Methods and Software

2.1. Computational Methods. For geometries, we consider three categories of QM theory plus MM. The QM categories are WFT, DFT, and SE-MO. For partial charges, we consider Mulliken population analysis,[175] MM,[154,176-179] and the CM1,[170] CM2,[180] CM3,[181] and CM4[174] charge models. DFT calculations in the aqueous phase utilize the implicit solvation model SM6,[174] while solvation in SE-MO methods[122,124] was included by using the implicit SM5.4[182] or SM5.42[183] solvation models. For aqueous-phase MM calculations, we employed explicit solvation using the respective programs' soak algorithms in conjunction with periodic boundary conditions (minimum cell size of 15 $\times$ 15 $\times$ 15 Å) to eliminate solvent-vacuum interfaces.

WFT calculations were carried out by CCSD with the 6-31+G(d,p)[173] basis set. These calculations were carried out with the *Gaussian03* computer program (Gaussian, Inc).[184] The CCSD method was chosen over the popular Møller-Plesset second order perturbation theory[185] as CCSD (or the closely related QCISD[186]) has been shown to yield more accurate geometries.[187]

The DFT methods examined are B3LYP,[188,189] mPW1PW[190,191] (which is also called mPW1PW91, mPW0, and MPW25), MPWB1K,[58,189,191] MPW1KCIS,[54] M06-L,[55] and M05-2X[55] with the 6-31+G(d,p)[173] basis set. Calculations were performed using a locally modified version of the *Gaussian03* program incorporating the MN-GSM 6.0[192] and MN-GFM 2.0.1[193] solvation and DFT modules. In order to select a

density functional to generate benchmark values, gas-phase calculations were carried out and the resulting geometries evaluated relative to CCSD. The best functional was subsequently used to perform calculations in the aqueous phase using the SM6[174] solvation model for implicit solvation.

We have tested Charge Model 4 (CM4) partial charge assignments based on gas-phase and SM6 DFT calculations. CM4 charges, the fourth generation of class IV[170] charges, have a distinct advantage over the class II[175,194-196] and III[197-199] charges used in *Gaussian03*. Whereas the reliability of class III charges depends on the wave function and basis set used, class IV charges represent an extrapolation to full configuration interaction with a complete basis set.[170,174] Furthermore, class III charges are unstable with respect to buried charges,[200-203] while class IV charges provide a reliable method for obtaining buried charges. CM4 charges, in particular, have been parameterized against a large training set (398 molecules) and are well suited for modeling aliphatic functional groups, which makes them more suitable for modeling hydrophobic effects – a primary factor in protein-protein interactions.

SE-MO methods examined in the current study include AM1,[122] PM3,[124] and PDDG/PM3.[204] Calculations were performed for AM1 and PM3 using AMSOL 7.1[205] (a derivative of AMPAC 2.1) with Mulliken, CM1, CM2, or CM3 charges obtained from gas-phase calculations. For aqueous-phase AM1 and PM3 calculations, AMSOL 7.1 was used to obtain Mulliken, CM1, or CM2 charges within the SM5.4 (for Mulliken and CM1) or the SM5.42 (for CM2) solvation models. GAMESSPLUS[192] was used to obtain a second set of CM3 charges in the gas phase in an effort to test consistency in

charge assignment across software. The notation used in this article for AM1 calculations with differing partial charge assignments is AM1, AM1-CM1, AM1-CM2, and AM1-CM3 for AM1 calculations with Mulliken, CM1, CM2, and CM3 charges, respectively. The notation for the PM3 calculations is analogous to that for AM1. It is important to note that, since they are post-self consistent field (SCF) analysis tools, CM*x* or Mulliken charges of gas-phase wave functions do not alter an optimized molecule's geometry. Slight differences may be attributed to variances in the convergences of the SCF and geometry optimizations. In solution, each SMx model uses a particular choice of charge model, and this choice, along with all the other SMx parameters, does affect the molecules' geometry. PDDG/PM3 gas-phase optimizations were performed using MOPAC 5.011mn,[122] and Mulliken partial atomic charges were obtained. In addition, PM3-Mulliken charge analyses were carried out with *Gaussian03* and MOPAC 5.011mn as part of a comparison between partial atomic charges assigned to optimized geometries calculated by PM3 and PDDG/PM3.

The MM force fields employed are AMBER,[154] AMBER*, CVFF,[176] CFF91,[206] MMFF94,[178] OPLS2005,[179] and Tripos[177]. The AMBER force field employed is the ff03 version[62] in conjunction with the general atom force field[164] (GAFF) commonly utilized for small organic systems. The AMBER* force field contains additional atomic parameters as implemented in *MacroModel*. In most cases, we elected to use nonrigid water with each force field's default parameters for water molecules. However, the AMBER* force field was locally modified to use OPLS2005 nonrigid water (in OPLS2005, the nonrigid water has the same Lennard-Jones parameters as the rigid

TIP3P water model), and the AMBER force field, via the SOLVATEOCT command, utilized the rigid TIP3P water model. Stretch, bend, Coulombic, and Lennard-Jones parameters for water models used can be found in each of the force field's descriptions (see references above), except for AMBER*, for which the modified water parameters are described by the OPLS2005 reference.

In addition to the standard force fields, we also employ local modifications of the force fields that substitute CM4 charges for their default partial charges. This combination of a force field and CM4 charges is denoted as X-CM4, where X is the name of the original force field. In contrast to gas-phase SE-MO calculations, when CM*x* charges are used with MM, they can and do alter the optimized geometry. Note that when we use CM4 charges with MM calculations, we use gas-phase M05-2X/6-31+G(d,p)/CM4 charges calculated at gas-phase M05-2X/6-31+G(d,p) geometries for gas-phase MM calculations, and we use SM6/M05-2X/6-31+G(d,p)/CM4 aqueous-phase charges calculated at aqueous-phase SM6/M05-2X/6-31+G(d,p) geometries for aqueous-phase MM calculations.

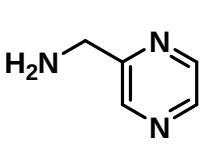
General MM optimization conditions consisted of at least 1000 steps of conjugate gradient minimization with an energy gradient cutoff of 0.01 kcal mol^{-1} $Å^{-1}$. In cases where water was included, the entire system was minimized, and although the stringent conditions for termination we have established were not typically reached, the system energy fluctuated only slightly around a stable potential energy. To test the effect of water position around the small molecules on geometries, six random orientations of 2-AMP and 2,4-DAP within the water box were used in minimizations

with OPLS2005-CM4 and MMFF94-CM4. Since this process yielded only a slight standard deviation in geometries (on the order of less than 0.006 Å in bond length and 0.7 degrees in bond angle), we used only the default water orientation for all other aqueous-phase MM optimizations. We note, however, the 2-AMP and 2,4-DAP bond lengths and angles used for OPLS2005-CM4 and MMFF94-CM4 assessment represent the average of these six results. In binding energy studies, water molecules were placed at positions on 2-AMP, 1*H*-2-AMP, 2,4-DAP, and 1*H*-2,4-DAP where the DHFR-MTX crystal structure denoted that hydrogen bonding takes place between the ligand and the enzyme.[169] The small molecule and water were optimized together, and the binding energy of the complex computed by subtracting the energies of the individual optimized molecules.

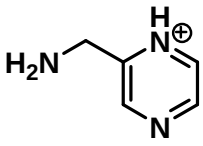
While testing the MM force fields, it was found that the CVFF and CFF91 force fields do not properly assign partial atomic charges to the 1*H*-2-AMP or 1*H*-2,4-DAP cations, as the total charge assigned to the molecule is not +1.0. We therefore calculated the gas-phase Hartree-Fock molecular electrostatic potential with the 6-31+G(d,p) basis set for the neutral and cationic species and obtained ChelpG[199] electrostatic-potential-filling charges to substitute into the force fields. Geometry optimizations with the ChelpG partial charges were then carried out in both the gaseous and aqueous phases, and the resulting geometries, denoted CVFF-HF and CFF91-HF, were used in our evaluation.

2.2. Platforms, Software, and Molecules. Quantum mechanical calculations (WFT, DFT, and SE-MO) were performed on an IBM Power4 (p690 and p655) computer system running under the AIX operating system and an SGI Altix cluster running under the Linux operating system. Molecular mechanics calculations were performed on a Silicon Graphics O2 workstation running under the IRIX 6.5 operating system. Molecules were constructed for quantal calculations using the GaussView 3.0 (Gaussian, Inc.) visualization program, and the generated Z-matrices were converted to Cartesian coordinates where appropriate. Molecules for MM calculations were constructed using *InsightII 2005* (Accelrys, Inc.) for the CFF91 and CVFF force fields, SYBYL 7.3 (Tripos, Inc.) for the MMFF94, AMBER and Tripos force fields, and *Maestro 7.5* (Schrodinger, Inc.) for the MMFF94, AMBER* and OPLS2005 force fields (utilizing the *MacroModel* and *Impact* applications, respectively). These programs were also used to set up and run the MM minimizations except for AMBER minimizations, for which SYBYL was used only to generate molecular coordinates in Protein Data Bank or Mol2 formats, and the AMBER 9[207] suite was used for the minimizations. The small molecules included in the present study are illustrated in Figure 3. Each molecule was modeled in both its neutral and protonated form. When an exocyclic amine is present, the proton was added there. Otherwise, the proton was added on a heterocyclic amine.

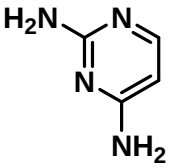
Figure 3. Chemical structures of small molecules used in the current study.



2-(aminomethyl)pyrazine

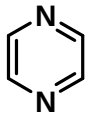


1*H*-2-(aminomethyl)pyrazine

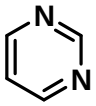


2,4-diaminopyrimidine

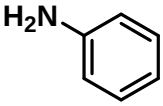
1*H*-2,4-diaminopyrimidine



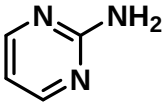
pyrazine



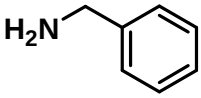
pyrimidine



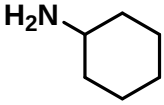
aniline



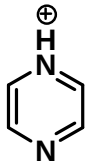
2-aminopyrimidine



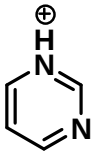
benzylamine



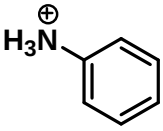
cyclohexylamine



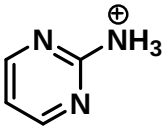
pyrizinium



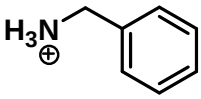
pyrimidinium



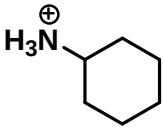
phenylammonium



pyrimidin-2-yl-ammonium



benzylammonium



cyclohexylammonium

3. Results and Discussion

3.1. Error Analysis. In order to rank the methods we chose to test, we calculate the unsigned residual between a calculated value with method *m* and the corresponding benchmark value:

$$R_i^{x,y,m} = \left| x_i^{\text{calc},m}(y) - x_i^{\text{benchmark}}(y) \right| \qquad (1)$$

where *y* is the phase (gas phase or aqueous phase) and $x_i(y)$ signifies case *i* of molecular property *x* in phase *y*, e.g., when *x* is bond length (*r*), $x_1(y)$ is the first bond length. Alternatively, *x* could stand for partial charge (*q*), bond angle (θ), or binding energy (E_b). The overall error in a particular molecular property *x* for a particular method *m* and phase *y* is quantified by the mean unsigned error (MUE):

$$MUE_{x,y,m} = \frac{\sum_{i=1}^{n_{x,y}} R_i^{x,y,m}}{n_{x,y}} \qquad (2)$$

where $n_{x,y}$ is the number of combinations of x_i and *y* for which $R_i^{x,y}$ is evaluated.

We also calculate the average mean unsigned error (AMUE) for each property across *N* methods:

$$AMUE = \frac{\sum_{m=1}^{N} MUE_{x,y,m}}{N} \qquad (3)$$

Dividing this value by a method's MUE yields the reduced (unitless) mean unsigned error (RMUE) for the method for partial charge, bond length, or bond angle.

$$RMUE_{x,y,m} = \frac{MUE_{x,y,m}}{AMUE_{x,y}} \qquad (4)$$

The RMUE is a measure of each method's performance relative to the mean of the others for calculating a particular molecular property. A value of 1.0 indicates that the method is average. Lower values indicate better methods, while higher values indicate worse methods. The reason for introducing these unitless reduced quantities is so that we can combine errors for *r* and θ (which have different units) in order to make an overall assessment for combined geometric performance.

We define a reduced deviance ($D_{y,m}$) of a SE-MO or MM method from the average performance in either the gas or aqueous phase by averaging the RMUE for *r* and θ in each method for both 2-AMP, 1*H*-2-AMP:

$$D_{y,m} = \frac{1}{4}\left(RMUE_{r,y,m}^{2-AMP} + RMUE_{\theta,y,m}^{2-AMP} + RMUE_{r,y,m}^{1H-2-AMP} + RMUE_{\theta,y,m}^{1H-2-AMP}\right) \qquad (5)$$

The reduced deviance for 2,4-DAP and 1*H*-2,4-DAP is averaged with the reduced deviance for 2-AMP and 1*H*-2-AMP to yield an overall performance. Reduced deviance in partial charge (*q*) assignment is calculated similarly. Reduced deviance for the validation of the most accurate methods is also calculated similarly, with all molecules taken into account.

3.2 Establishing Geometry and Partial Charge Benchmark Sets. The MUE of the gas-phase geometries calculated by DFT with respect to the CCSD-calculated geometries is given in Table 1. All DFT methods perform well, with mean unsigned errors within ~0.01 Å for bond length and less than one degree for bond angles. Based on its high degree of accuracy for both bond length and angle, we chose to proceed with the M05-2X level of theory to generate our geometric benchmark set.

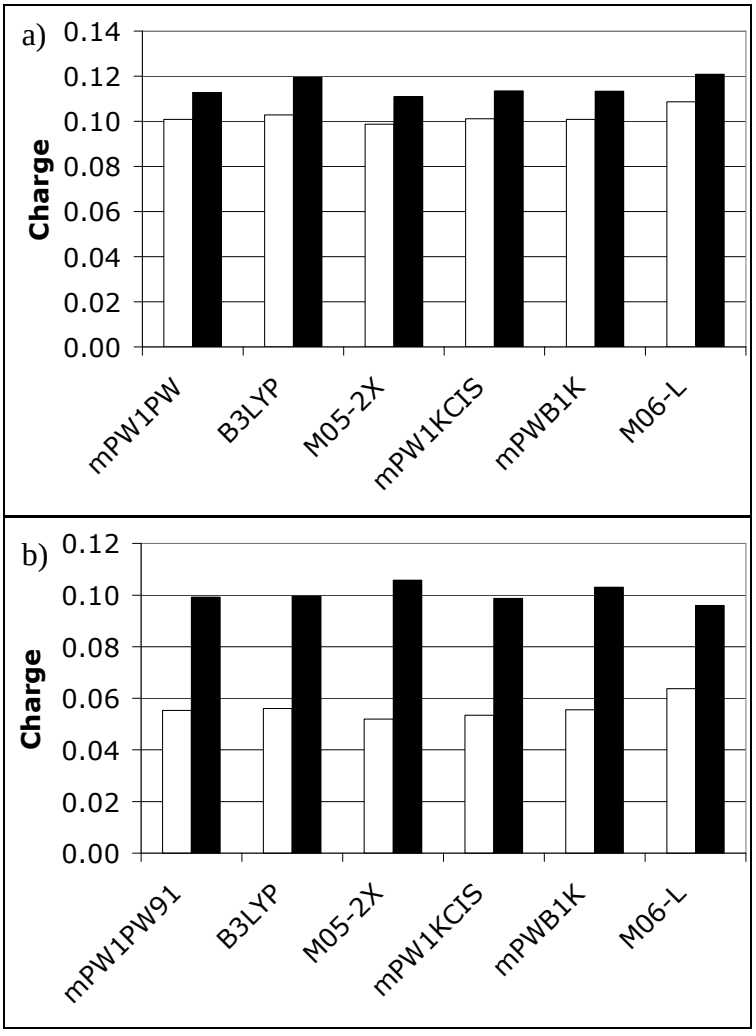
Table 1. Mean Unsigned Error (Å and deg) for Gas-Phase Bond Lengths and Angles Between Those Calculated with CCSD and Each Functional.

Method	2-AMP		1*H*-2-AMP		2,4-DAP		1*H*-2,4-DAP	
	Bond Length	Bond Angle	Bond Length	Bond Angle	Bond Length	Bond Angle	Bond Length	Bond Angle
B3LYP	0.003	0.73	0.003	0.37	0.003	0.82	0.003	0.27
MPW1PW	0.005	0.44	0.005	0.49	0.005	0.87	0.004	0.26
mPWB1K	0.011	0.52	0.010	0.42	0.012	0.99	0.010	0.26
mPW1KCIS	0.005	0.78	0.005	0.50	0.004	0.90	0.004	0.28
M06-L	0.006	0.52	0.005	0.33	0.004	0.54	0.004	0.22
M05-2X	0.004	0.28	0.004	0.26	0.005	0.94	0.003	0.22
AMUE[a]	0.006	0.54	0.005	0.40	0.006	0.84	0.005	0.25

[a]Mean of entire column

We selected the well-validated[174,181,208,209] CM4 charge model to obtain benchmark partial atomic charges. In order to examine whether the CM4 charge model would assign partial charge similarly for each functional used, we compared the gas-phase CCSD partial charges generated by Mulliken population analysis (a class II[170] charge method) to CM4 partial charges assigned by each density functional tested (note that the CM4 charges are probably more accurate). The results are summarized in Figure 4. In this figure and in this whole article, charges are given in atomic units in which the charge on a bare proton is 1.0. Overall, the mean unsigned deviations between the CCSD Mulliken analysis and the CM4-assigned partial charges vary by $\leq$ 0.01 charge units regardless of the functional or phase tested. We therefore applied the geometrically accurate M05-2X functional in conjunction with M05-2X/CM4 partial charges to obtain our benchmark set.

Figure 4. Mean unsigned deviation of DFT/CM4 charges relative to CCSD/Mulliken charges for a) 2-AMP and its cation (White – 2-AMP, black – 1H-2-AMP) and b) DAP and it's cation (White – DAP, black 1H-DAP).



3.3. Exploration of CVFF and CFF91 Atom Typing and Charge Distribution. As mentioned in Section 2.2, a deficiency in the CVFF and CFF91 force fields is that they do not assign a total charge of +1.0 to the pyrazinium or pyridinium cations. Upon further exploration using 2-AMP, the default atom type assigned to N1 in 2-AMP by both force fields is "np" (an sp^2 nitrogen in a 5- or 6-membered ring), and the partial charge on this unprotonated nitrogen is –0.22 in CVFF and –0.48 in CFF91. Upon protonation and automated reassignment of atom types by *InsightII 2005*, the CVFF nitrogen atom type remains unchanged, and the CFF91 atom type changes to "nh+" (a protonated nitrogen in a 6-membered ring). The protons added have partial charges of +0.28 (CVFF) and +0.33 (CFF91), and the partial charges on the nitrogens change to –0.50 and –0.81 charge units, respectively. The partial charge on all other atoms in the molecule are unaffected by the addition of the proton. This charge balancing yields a total charge on the molecule of 0, which is incorrect for a cation.

Some of the partial atomic charges in the CVFF and CFF91 force fields are derived from fits to the Hartree-Fock molecular electrostatic potential,[210-212] and we took this as a cue for how to correct the problem in a way consistent with these force fields. In particular, we carried out single-point, gas-phase Hartree-Fock (6-31G(d,p) basis set) calculations on the optimized 2-AMP and 1*H*-2-AMP geometries and obtained partial atomic charges by electrostatic potential fitting with the ChelpG[199] algorithm. The resulting partial atomic charges were used in the CVFF and CFF91 force fields for geometry optimizations in both the gas and aqueous phases. We used the gas-phase partial charges in both phases since the original CVFF and CFF91 partial

charges are based on gas-phase calculations (we note that all MM force fields considered in this article use partial charges that do not depend on the phase). To denote the new Hartree-Fock partial charges in the force fields, we name the force fields that use these newly assigned charges as CVFF-HF and CFF91-HF.

3.4. Evaluation of SE-MO and MM Calculations. We tested available SE-MO and MM parameter sets in an effort to select the most accurate method for modeling our small molecules both in the gas phase and in solvent. Additionally, we substituted CM4 charges into the MM force fields tested to observe effects on geometric accuracy. A summary of the results of these tests is given in Figures 5 and 6 (gas phase) and 7 and 8 (aqueous phase). In the gas phase, the AM1 and PM3 SE-MO methods utilizing the CM1, CM2, and CM3 charge models as well as the CVFF-HF, CFF91-HF, AMBER*, and MMFF94 force fields all predict the partial charges of the atoms in both sets of molecules to within 0.15 (Figs. 5a and 6a). In most of the methods tested, the partial charge assignment becomes less accurate in the cationic species – the most notable examples, for which the average error increases by a factor of 2, are the OPLS2005 force field for 1*H*-2-AMP, the PM3-CM1, -CM2, and -CM3 methods for 1*H*-2-AMP, and the SE-MO methods utilizing the CM1, CM2, and CM3 charge models for 1*H*-2,4-DAP (with the exception of PM3-CM3). In contrast, the MMFF94 force field becomes more accurate for 1*H*-2-AMP and the AMBER force field becomes more almost 2-fold more accurate for 1*H*-2,4-DAP, although AMBER's overall charge assignment is somewhat inaccurate (the mean MUE for the neutral and cationic species

Figure 5. MUE in a) partial charge, b) bond length, and c) bond angle for selected SE-MO and MM methods in the gas phase relative to M05-2X/CM4. White – 2-AMP, black – 1H-2-AMP. AMS denotes AMSOL, and GMSP denotes GAMESSPLUS.

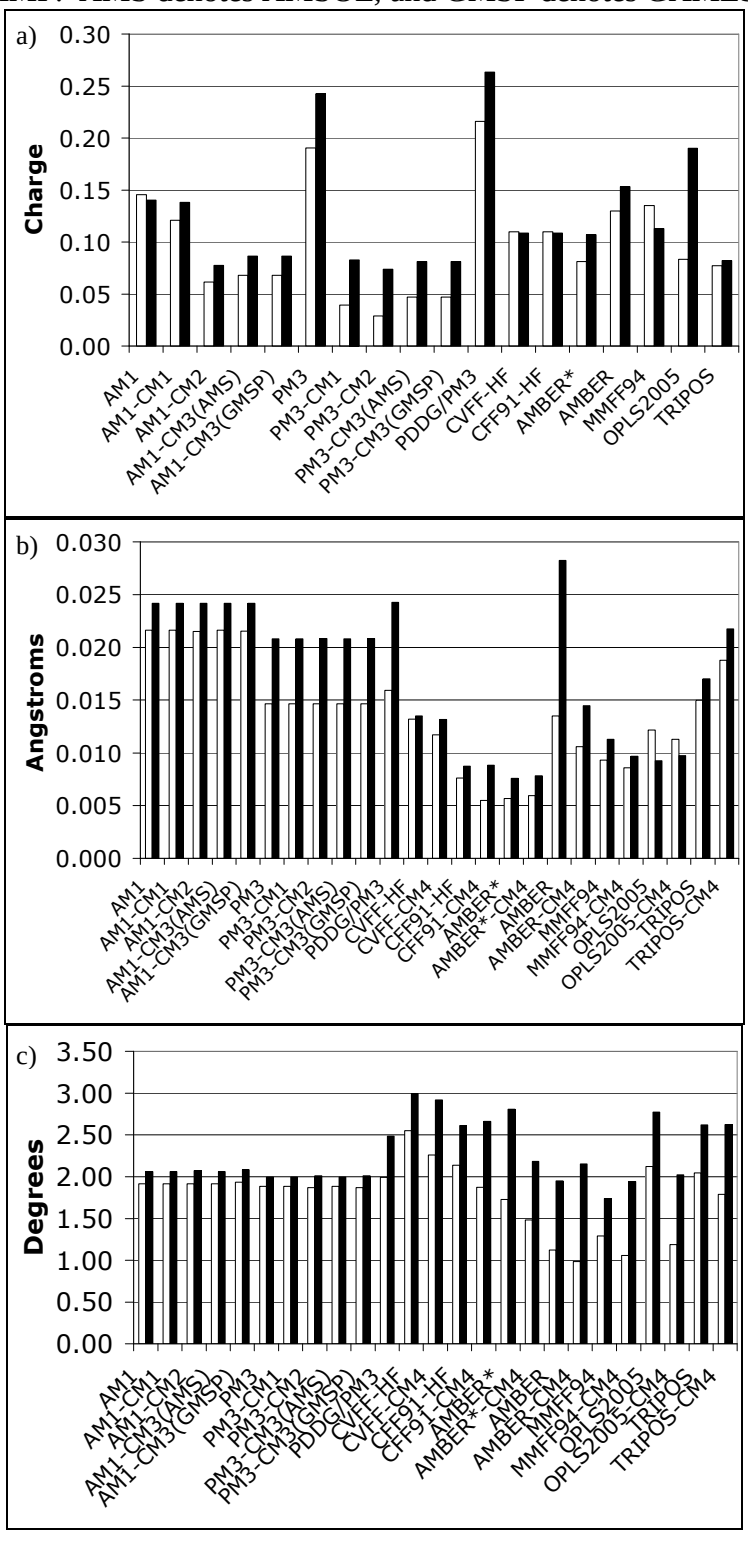


Figure 6. MUE in a) partial charge, b) bond length, and c) bond angle for selected SE-MO and MM methods in the gas phase relative to M05-2X/CM4. White – 2,4-DAP, black – 1H-2,4-DAP.

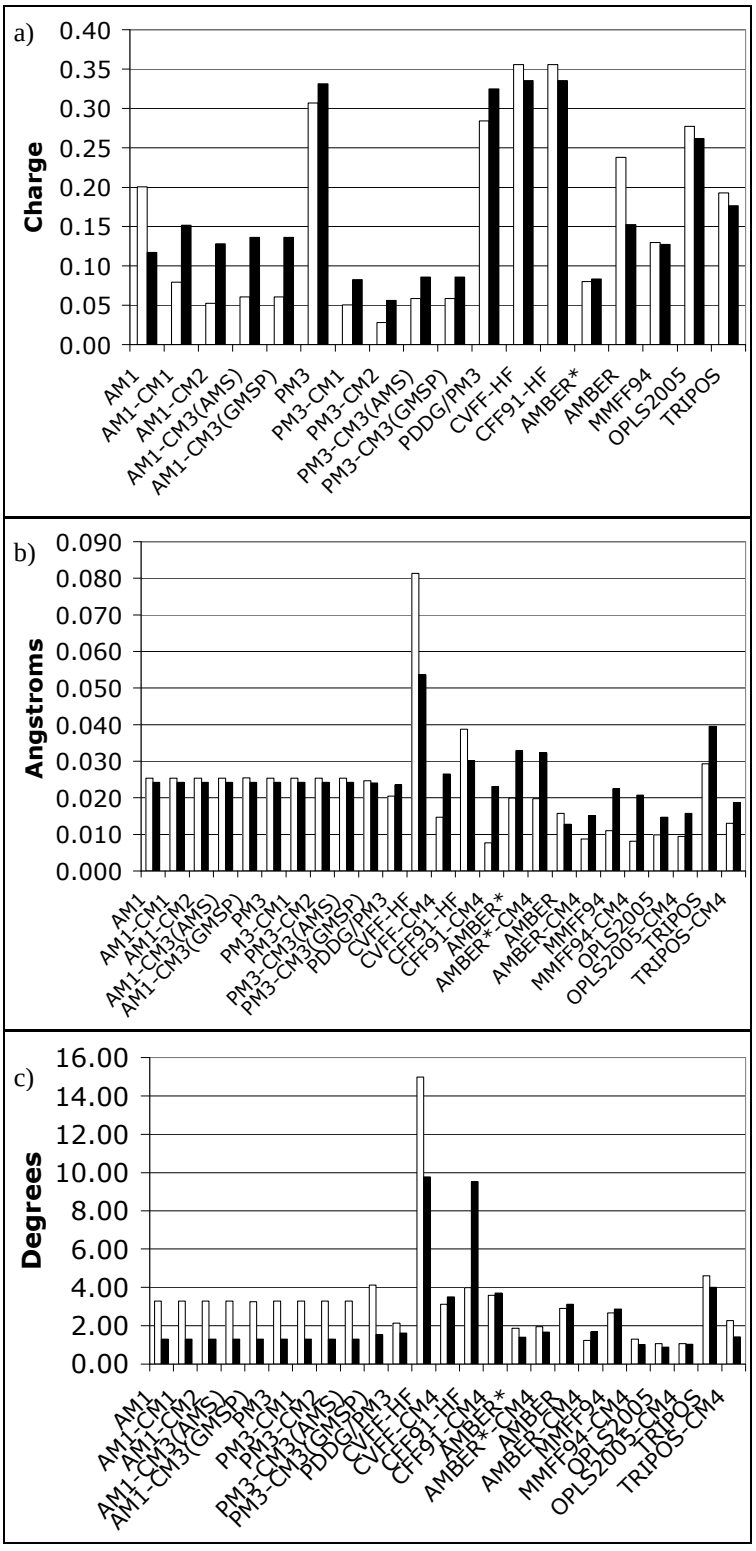


Figure 7. MUE in a) partial charge, b) bond length, and c) bond angle for selected SE-MO and MM methods in solution relative to M05-2X/CM4. White – 2-AMP, black – 1H-2-AMP.

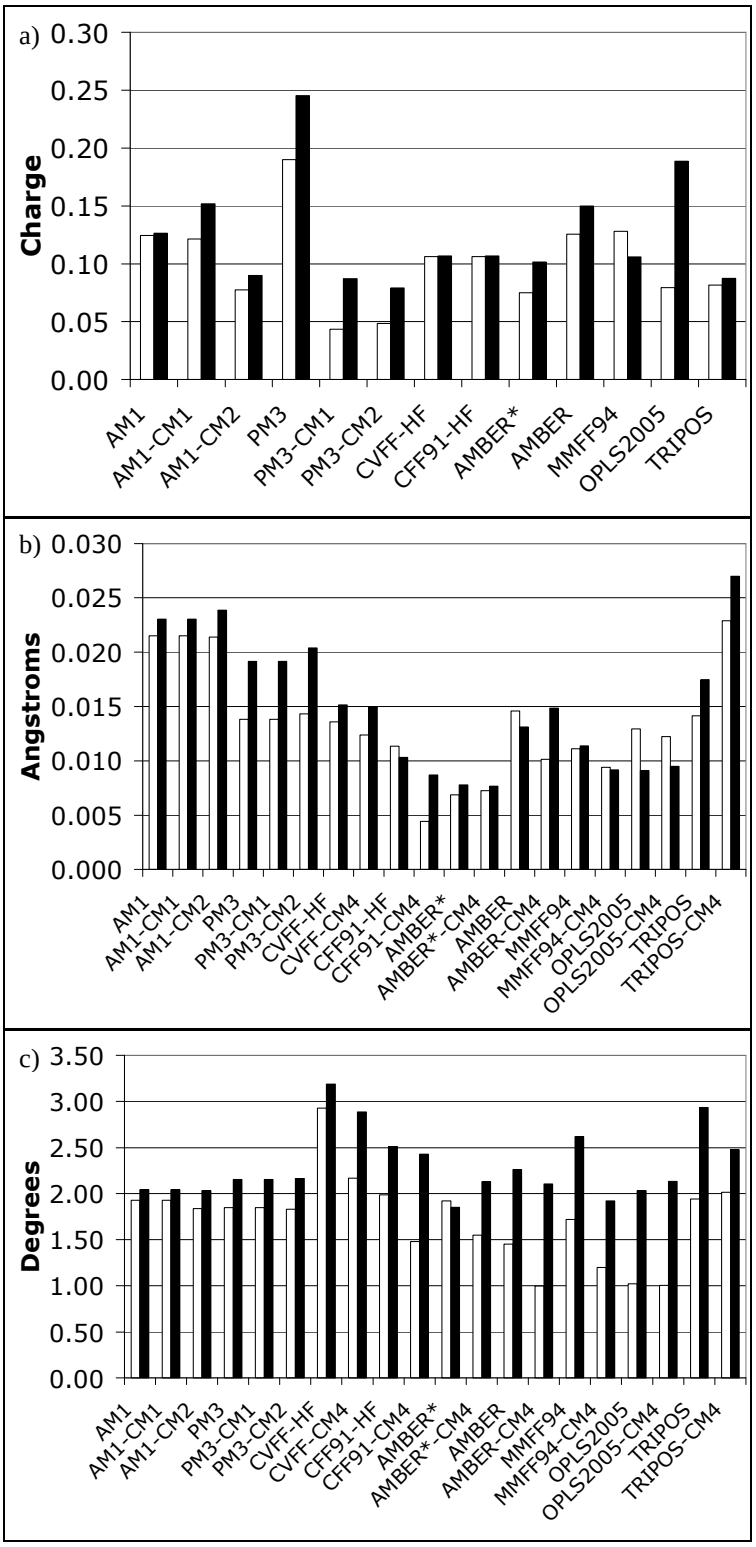
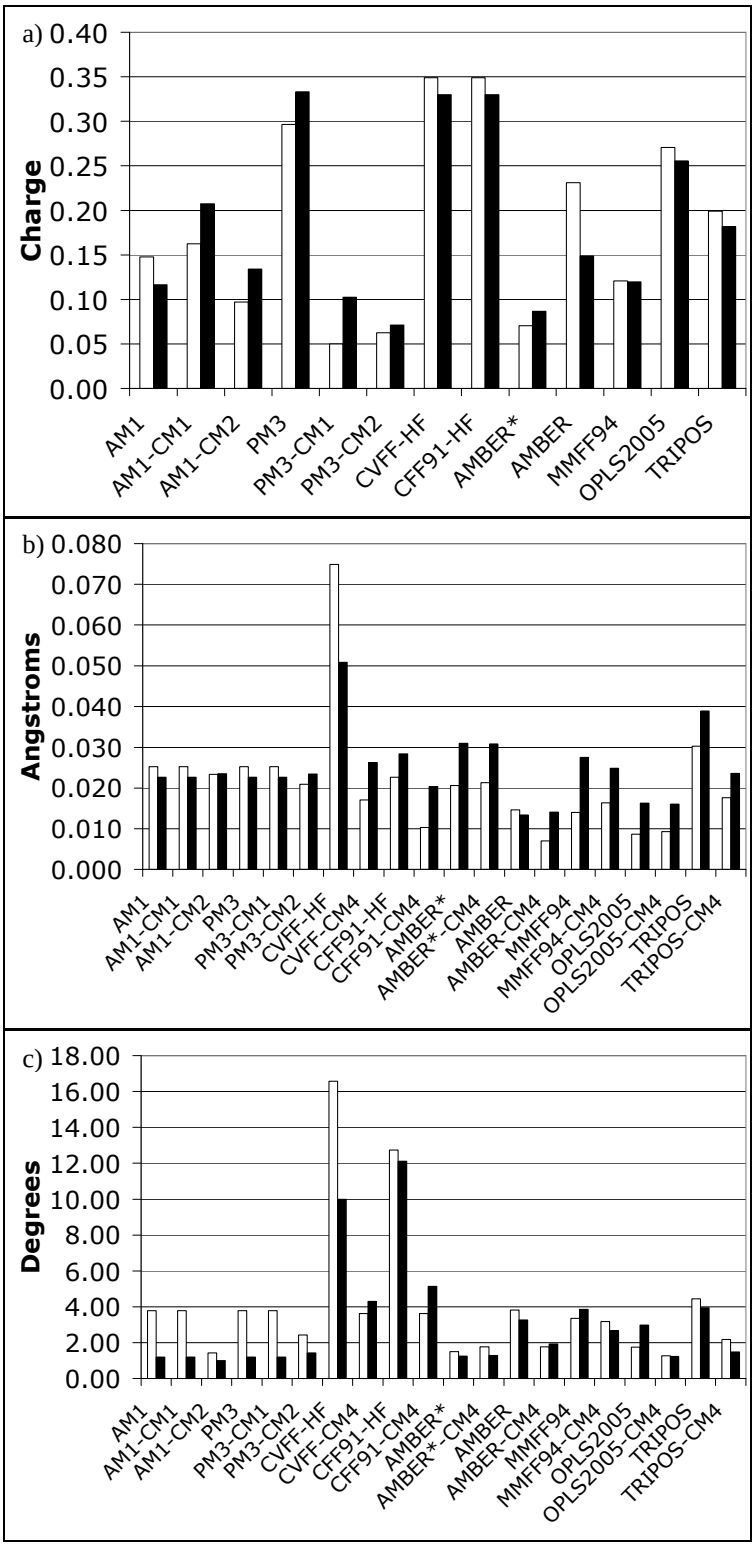


Figure 8. MUE in a) partial charge, b) bond length, and c) bond angle for selected SE-MO and MM methods in solution relative to M05-2X/CM4. White – 2,4-DAP, black – 1H-2,4-DAP.



together is 0.20). The force fields with CM4 charges substituted are not included in the partial charge analysis since their mean unsigned error in partial charge is always zero.

The gas-phase partial charges assigned by the PDDG/PM3 method have a mean unsigned error of 0.22 and 0.28 for the neutral species and 0.26 and 0.32 for the cationic species of 2-AMP and 2,4-DAP, respectively. Table 2 summarizes our comparison of the PDDG/PM3 charges to Mulliken population analysis of the PM3-optimized structure of 2-AMP as calculated by both *Gaussian03* and MOPAC 5.011mn (the numbering system is given in Figure A1 of Appendix One); this comparison verifies the consistency of the two programs (as a check). Figures 5 and 6 show that PDDG/PM3 is less accurate than PM3 for gas-phase partial charges, bond lengths, and bond angles for both 2-AMP and 1*H*-2-AMP and for gas-phase bond angles for both 2,4-DAP and 1*H*-2,4-DAP. Because the PDDG reparameterization of PM3 deteriorated the performance for these molecules, PDDG/PM3 was not considered in the aqueous-phase calculations that follow.

Table 2. Gas-Phase Partial Charges of 2-AMP Calculated by PM3 (Mulliken Population Analysis) and PDDG/PM3.

Atom	PM3(G03)	PM3(MOPAC)	PDDG/PM3(MOPAC)
C1	-0.1310	-0.1309	-0.1316
C2	-0.1067	-0.1067	-0.1647
N3	-0.0382	-0.0382	-0.0274
C4	-0.1095	-0.1094	-0.1664
C5	-0.1085	-0.1085	-0.1581
N6	-0.0239	-0.0239	-0.0211
H7	0.1344	0.1344	0.1814
H8	0.1319	0.1319	0.1798
H9	0.1316	0.1316	0.1789
C10	-0.0569	-0.0569	-0.1491
H11	0.0677	0.0677	0.1151
H12	0.0830	0.0830	0.1296
N13	-0.0282	-0.0282	-0.0928
H14	0.0301	0.0301	0.0665
H15	0.0242	0.0242	0.0598

For 2-AMP and 1*H*-2-AMP, gas-phase bond lengths (Fig. 5b) are predicted to within ~0.01 Å only by the CFF91-HF, AMBER*, MMFF94, and OPLS2005 force fields. All methods except OPLS2005 become less accurate when modeling the cationic species. Notably, the AMBER force field becomes 2-fold less accurate when modeling the cation, and it clearly benefits from revised charges, as the AMBER-CM4 force field yields a twofold performance improvement when modeling bond lengths. On average, force fields with CM4 charges substituted tend to perform slightly better than those with default charges, except in the case of TRIPOS-CM4. Gas-phase bond angles (Fig. 5c) are predicted to within ~2.0 degrees by all the SE-MO methods, while the majority of the MM force fields are ~0.5-1 degree less accurate. Exceptions include the MMFF94 and AMBER force fields, which predict them to within ~1.5 degrees. All methods are slightly more inaccurate dealing with the cation. As with bond lengths, when CM4 charges are substituted, bond angle predictivity is increased slightly in most methods. OPLS2005-CM4 is notable in that it has a twofold improvement in performance over OPLS2005 when modeling the neutral species and a smaller, yet significant (~0.5°), increase in accuracy for the cationic species.

For 2,4-DAP and 1*H*-2,4-DAP, bond lengths (Fig. 6b) are predicted to an accuracy of ~0.025 Å by all methods except CVFF-HF, CFF91-HF, and TRIPOS. The MMFF94 and OPLS2005 force fields both display accuracy less than 0.01 Å for 2,4-DAP, while AMBER and OPLS2005 are accurate to better than 0.015 Å for 1*H*-2,4-DAP. MMFF94 is two-fold less accurate for the 1*H*-2,4-DAP cation. With regard to 2,4-DAP and 1*H*-2,4-DAP gas-phase bond angles, the best performing methods include

the AMBER*, AMBER, MMFF94, and OPLS2005 force fields, predicting bond angle to within 1,9, 3.1, 2.9, and 1.1 degrees for both 2,4-DAP and 1*H*-2,4-DAP, respectively. With respect to force fields with CM4 charges substituted, the clearest example of a force field benefiting from revised charges is the CVFF-HF and CFF91-HF force fields which, as previously discussed, have major partial charge assignment problems. When CM4 charges are substituted into these force fields, both methods show a dramatic improvement in performance for modeling bond lengths, with CVFF-CM4 yielding a fourfold more accurate bond length prediction overall, and CFF91-CM4 yielding the same improvement in accuracy when modeling the neutral species. The TRIPOS force field also benefits, with TRIPOS-CM4 displaying a twofold improvement in accuracy for both molecules.

All methods except CVFF-HF, CFF91-HF, and TRIPOS predict bond angles to within ~4.0 degrees for both species (Fig. 6c). Both CVFF-HF and CFF91-HF presented a challenge when attempting to minimize structure with respect to the fact that 1*H*-2,4-DAP could not be assigned proper atomic partial charges. As described earlier, we substituted the default charges with ChelpG charges derived from HF calculations on 2,4-DAP and 1*H*-2,4-DAP. When these charges were substituted into the force field, the charges on the protons on the exocyclic amines were made largely more positive than the default charges. This effect, in conjunction with the largely negatively charged heterocyclic amines, caused the minimization to distort the sp^3 structure of the exocyclic amines and pull the protons toward the pyrimidine ring. The result is a large error in the assignment of bond angle. When the minimization is

performed with the default charges, no distortion of bond angle takes place; however, these charges correspond to a non-physical total molecular charge. As observed in tests already described, CM4 charge substitution greatly improves geometric modeling performance. The CVFF-CM4 and CFF91-CM4 force fields essentially eliminate the problems seen with these force fields, causing both to model bond angles with accuracy on par with the rest of the methods used. In addition, the AMBER-CM4, MMFF94-CM4, and TRIPOS-CM4 force fields all yield accuracies at least twofold greater than their counterparts.

For 2-AMP and 1*H*-2-AMP in the aqueous phase, the AM1-CM2, PM3-CM1, and PM3-CM2 methods as well as the CVFF-HF, CFF91-HF, AMBER*, MMFF94, and TRIPOS force fields retain a high degree of accuracy for partial charge prediction (MUE $\leq$ 0.13, Fig. 7a). Among these methods, the MMFF94 force field is again the only method to become more accurate when modeling the protonated species. On the other hand, the OPLS2005 force field becomes 2-fold less accurate when predicting the charge of the cation. Bond lengths (Fig. 7b) are calculated with an error similar to that of the gas phase, with the CFF91-HF, AMBER*, MMFF94, and OPLS2005 force fields maintaining a mean unsigned error of ~0.01 Å. Interestingly, while it becomes much less accurate when assigning partial charge to the ionic species, the OPLS2005 force field becomes more accurate in bond length prediction. As with 2-AMP and 1*H*-2-AMP in the gas phase, CM4 partial charge substitution generally produces a modest increase in bond length accuracy. A notable exception is the CFF91-CM4 force field, with an MUE <0.005 Å. When compared to calculations in the gas phase, all methods

either retain their accuracy or become slightly less accurate when predicting bond angle in solution except for the OPLS2005 force field. Except for the CVFF-HF, AMBER*, and TRIPOS force fields, all methods are still accurate to ~2.5 degrees or less (Fig 7c). Again, on average, CM4 charge substitution slightly increases bond angle accuracy for both species.

In the aqueous-phase treatment of 2,4-DAP and 1*H*-2,4-DAP, partial charge performance is quite varied (Fig. 8a). The best performing methods are the PM3-CM1 and -CM2 SE-MO methods and the AMBER* and MMFF94 force fields, which all model the partial charges to an accuracy of 0.12 or better. In the treatment of these molecules, all the force fields except AMBER* become more accurate to varying degrees when dealing with the cationic species. The quality of modeling the bond lengths (Fig. 8b) is similar for most methods; however, the CVFF-HF force field is particularly inaccurate, with a MUE in bond length for 2,4-DAP of 0.075 Å. The AMBER and OPLS2005 force fields perform particularly well for both species, with MUEs of $\leq$ 0.016 Å overall. However, the OPLS2005 force field becomes twofold less accurate when modeling the cationic species. MMFF94 also performs relatively well, modeling the neutral species with a MUE $\leq$ 0.014 Å, but, like OPLS2005, becomes twofold less accurate when modeling the cation. CVFF-CM4, CFF91-CM4, AMBER-CM4 (for the neutral species) and TRIPOS-CM4 all yield about a twofold increase in accuracy. Other force fields with CM4 charges in place show little to no difference. Bond angles (Fig. 8c) are treated with accuracy similar to that in the gas phase, with the CVFF-HF and CFF91-HF force fields performing very poorly (MUE $\geq$ 10 degrees). All

methods except the TRIPOS, CVFF-CM4, and CFF91-CM4 force fields model bond angles to within ~4.0 degrees, although CVFF-CM4 and CFF91-CM4 again represent a great improvement over their parent force fields. Again, CM4 treatment of the force fields generally leads to an almost twofold increase in accuracy for at least one, if not both, of the molecules modeled by each method.

Comparison of Figs. 5 and 6 to Figs. 7 and 8 shows that the SE-MO methods and force fields are about equally accurate when comparing gas-phase to aqueous-phase results, so no one method in particular stands out as extremely ill-suited to work in either the gaseous or aqueous phase.

3.5. Overall Geometric Assessment. In order to make an overall geometric assessment, we consider the reduced deviances ($D_{y,m}$) in partial charge assignment (Table 3) and geometric modeling (Table 4). Reduced deviance in partial charge shows that seven of the eleven SE-MO methods tested (the exceptions being AM1, AM1-CM1, PM3, and PDDG/PM3) perform better than the average method. In fact, PM3-CM2 is a factor of 2.5 better than average. The only MM methods that predict partial charge better than average are the AMBER*, MMFF94, and TRIPOS force fields. This assessment, however, also must take into account that the PM3-CM*x* and AM1-CM*x* methods utilize charge methods that serve as precursors to charge model 4. Since the training sets for the various CM*x* models share some molecules, it is perhaps not surprising that the partial atomic charges of the various CM*x* models show some agreement.

Table 3. Reduced Deviance ($D_{y,m}$) in Partial Charge for Each SE-MO and MM Method Tested.

	2-AMP and 1*H*-2-AMP			2,4-DAP and 1*H*-2,4-DAP			All Molecules		
Method	Gas	Aqueous	Mean	Gas	Aqueous	Mean	Gas	Aqueous	Mean
PM3-CM2	0.45	0.56	0.50	0.25	0.36	0.31	0.35	0.46	0.40
PM3-CM1	0.54	0.56	0.55	0.40	0.41	0.40	0.47	0.49	0.48
AM1-CM2	0.63	0.75	0.69	0.54	0.62	0.58	0.58	0.68	0.63
AMBER*	0.85	0.78	0.81	0.49	0.42	0.46	0.67	0.60	0.64
MMFF94	1.15	1.06	1.11	0.77	0.65	0.71	0.96	0.85	0.91
TRIPOS	0.73	0.75	0.74	1.12	1.03	1.07	0.92	0.89	0.91
AM1-CM1	1.18	1.21	1.19	0.69	1.00	0.84	0.93	1.10	1.02
AM1	1.32	1.12	1.22	0.97	0.71	0.84	1.14	0.92	1.03
AMBER	1.29	1.22	1.25	1.19	1.02	1.11	1.24	1.12	1.18
OPLS2005	1.20	1.15	1.17	1.63	1.42	1.52	1.41	1.28	1.35
CVFF-HF	1.00	0.96	0.98	2.09	1.83	1.96	1.54	1.39	1.47
CFF91-HF	1.00	0.96	0.98	2.09	1.83	1.96	1.54	1.39	1.47
PM3	1.96	1.92	1.94	1.92	1.70	1.81	1.94	1.81	1.88
PM3-CM3(GMSP)†,[a]	0.57	ND	ND	0.43	ND	ND	0.50	ND	ND
PM3-CM3(AMS)†,[a]	0.57	ND	ND	0.43	ND	ND	0.50	ND	ND
AM1-CM3(GMSP)†,[a]	0.70	ND	ND	0.58	ND	ND	0.64	ND	ND
AM1-CM3(AMS)†,[a]	0.70	ND	ND	0.58	ND	ND	0.64	ND	ND
PDDG/PM3[a]	2.17	ND	ND	1.83	ND	ND	2.00	ND	ND

[a]ND = not determined. †AMS = calculated using AMSOL7.1, GMSP = calculated using GAMESPLUSS

Table 4. Reduced Deviance ($D_{y,m}$) in Combined Geometry for Each SE-MO and MM Method Tested.

	2-AMP and 1*H*-2-AMP			2,4-DAP and 1*H*-2,4-DAP			All Molecules		
Method	Gas	Aqueous	Mean	Gas	Aqueous	Mean	Gas	Aqueous	Mean
OPLS2005-CM4	0.73	0.76	0.75	0.45	0.45	0.45	0.59	0.60	0.60
AMBER-CM4	0.78	0.80	0.79	0.52	0.49	0.50	0.65	0.65	0.65
MMFF94-CM4	0.66	0.71	0.68	0.50	0.85	0.68	0.58	0.78	0.68
OPLS2005	0.96	0.76	0.86	0.43	0.61	0.52	0.69	0.68	0.69
AMBER*-CM4	0.67	0.71	0.69	0.86	0.77	0.82	0.77	0.74	0.75
AMBER*	0.76	0.73	0.75	0.84	0.74	0.79	0.80	0.74	0.77
CFF91-CM4	0.78	0.70	0.74	0.97	0.96	0.97	0.88	0.83	0.85
MMFF94	0.70	0.93	0.81	0.85	0.95	0.90	0.78	0.94	0.86
AMBER	1.03	0.94	0.98	0.85	0.80	0.82	0.94	0.87	0.90
PM3-CM2	1.05	1.10	1.08	0.91	0.74	0.83	0.98	0.92	0.95
PM3	1.05	1.07	1.06	0.91	0.85	0.88	0.98	0.96	0.97
PM3-CM1	1.05	1.07	1.06	0.91	0.85	0.88	0.98	0.96	0.97
TRIPOS-CM4	1.19	1.43	1.31	0.65	0.69	0.67	0.92	1.06	0.99
AM1-CM2	1.24	1.27	1.25	0.91	0.68	0.79	1.07	0.98	1.02
CVFF-CM4	1.04	1.10	1.07	1.03	1.03	1.03	1.04	1.06	1.05
AM1-CM1	1.24	1.28	1.26	0.91	0.85	0.88	1.07	1.06	1.07
AM1	1.24	1.28	1.26	0.91	0.85	0.88	1.07	1.06	1.07
TRIPOS	1.09	1.15	1.12	1.49	1.33	1.41	1.29	1.24	1.26
CFF91-HF	0.85	0.94	0.90	2.02	2.30	2.16	1.44	1.62	1.53
CVFF-HF	1.12	1.27	1.20	3.60	3.20	3.40	2.36	2.24	2.30
PM3-CM3(AMS)†,[a]	1.05	ND	ND	0.91	ND	ND	0.98	ND	ND
PDDG/PM3[a]	1.19	ND	ND	0.79	ND	ND	0.99	ND	ND
PM3-CM3(GMSP)†,[a]	1.05	ND	ND	0.99	ND	ND	1.02	ND	ND
AM1-CM3(GMSP)†,[a]	1.24	ND	ND	0.91	ND	ND	1.07	ND	ND
AM1-CM3(AMS)†,[a]	1.24	ND	ND	0.91	ND	ND	1.07	ND	ND

[a]ND = not determined. †AMS = calculated using AMSOL7.1, GMSP = calculated using GAMESPLUSS

Across the molecules, reduced deviance is fairly consistent when the large errors in CVFF-HF and CFF91-HF for 2,4-DAP and 1*H*-2,4-DAP are taken into account. A slight skewing of the data may be occurring due to the aforementioned error since CVFF-HF and CFF91-HF actually perform as well as the average in partial charge assignment for 2-AMP and 1*H*-2-AMP, but perform so poorly in the other cases that their overall reduced deviance is high. However, if we were to carry on with either the CVFF-HF or CFF91-HF force fields by using the original method[211] of partial charge assignment for these force fields, we would be forced to perform electrostatic potential fitting on each molecule we study, and it is known that one of the weaknesses of ChelpG is fitting to the molecular electrostatic potential (MEP) of larger systems, where buried atoms are screened from the points where the MEP is evaluated, and changes of the partial charges on these atoms have only small effects on the MEP. Attempting to fit these charges, then, often produces non-physical results.[199-202]

Whereas the SE-MO charges and CM*x* charges are different in the gas phase and the aqueous phase, the MM charges (with the exception of those models using CM4 charges) are the same. Table 3 allows for further examination of an issue discussed briefly at the end of Section 3.4, namely the question of whether standard MM charges are more appropriate for the gas phase or for liquid-phase solution. Some force fields are explicit on this issue, for example, OPLS is explicitly named for its use in liquid-phase simulations. Others are implicitly designed for use in liquid phases simply because that is where the greatest number of applications of MM force fields occur. Table 3 shows that MMFF94, AMBER, OPLS2005, CVFF-HF, and CFF91-HF all

perform significantly better for partial atomic charges in water than in the gas phase, and AMBER* and TRIPOS are slightly better in the aqueous phase than in the gas phase. The finding that the partial charges are more indicative of the charge distribution in the aqueous phase than the gas phase in all seven cases is quite remarkable and is encouraging.

The reduced deviance in geometric modeling is given in Table 4. Of particular interest are the excellent performance of the MM-CM4 methods and the repair of the CFF91-HF force field. Prior to partial charge replacement, the CFF91-HF force field has a geometric $D_{y,m}$ equal to 1.54. After CM4 treatment, the $D_{y,m}$ for CVFF-CM4 is 0.86, representing a significant increase in the geometric modeling capabilities of the force field by fixing partial charge assignment problems. Other MM-CM4 methods also perform very well, with OPLS2005-CM4, AMBER-CM4, and MMFF94-CM performing 39%, 36%, and 32% better than the average, respectively. As far as non-CM4 treated force fields are concerned, the OPLS2005 force field performs the best ($D_{y,m}$ = 0.68), with AMBER*, MMFF94, and AMBER also modeling geometries better than average. The tested SE-MO methods all perform very similarly around a $D_{y,m}$ of ~1.0.

Another interesting point in Table 4 is the comparison of performance in the gas phase with that in aqueous solution. Of the fourteen MM rows in Table 4, six show smaller errors for gas-phase geometries, and eight show smaller errors for aqueous geometries (of the eight other methods for which such comparison is possible, all show better agreement in the aqueous phase). The MMFF94 method is particularly

noteworthy in showing (with either MMFF94 charges or CM4 charges) much better accuracy for gas-phase geometries than for aqueous ones.

Based on a combined assessment of partial charge assignment and geometric modeling performance, we find that the AMBER, AMBER*, MMFF94, and OPLS2005 force fields, along with their CM4-treated counterparts, would be suitable for carrying on into a validation step against a larger set of molecules. Each force field has some particular advantages. The AMBER force field has traditionally been highly regarded for its use in modeling biopolymers, and, in this case, is used with the incorporated parameters of the General Amber Force Field (GAFF). Atom types in GAFF are designed to be more general than those of traditional AMBER force fields, in an effort to cover a larger portion of organic space. The parameterization of GAFF was developed in an effort to reproduce restrained electrostatic potential (RESP) charges[200,204] at the MP2/6-31G(d) level and reproduce MP2/6-31G(d) and crystallographic geometries.[164] The AMBER* force field has been implemented in *MacroModel* and been modified to better reproduce HF/6-31+G(d) data on peptides as well as small organic molecules, especially those with nitrogen as a component.[213,214] The MMFF94 force field has been parameterized against a large (ca. 2800 structures), high-quality *ab inito* training set for use with both organic molecules and biopolymers, specifically in solution.[215]

3.6. Comparison of Binding Energies. We carried out gas-phase binding energy calculations between 2-AMP, 1*H*-2-AMP, 2,4-DAP, and 1*H*-2,4-DAP and water molecules placed at locations where hydrogen bonding has been described in the DHFR-MTX crystal structure (Figure 9). The results of this study are presented in Figure 10, wherein the MUE is the mean unsigned error for a method in binding energy for all neutral molecules, all charged molecules, or both sets together. Compared to the M05-2X benchmark binding energies, all methods except for CVFF-HF and CFF91-HF predict binding energies in both sets of species with an MUE of less than 7.0 kcal/mol. Fourteen of the 16 methods tested predict binding energy for the charged species less accurately than for the neutral species, with some methods – such as AMBER and AMBER* – displaying a fourfold increase in MUE. Two methods, AM1 and CVFF-CM4, predict binding energy for protonated species more accurately than for neutral species, with MUE differences between the two species of 1.0 and 2.0 kcal/mol, respectively.

Figure 9. Molecular systems used in binding energy calculations. Position of the water molecule reflects the optimized complex.

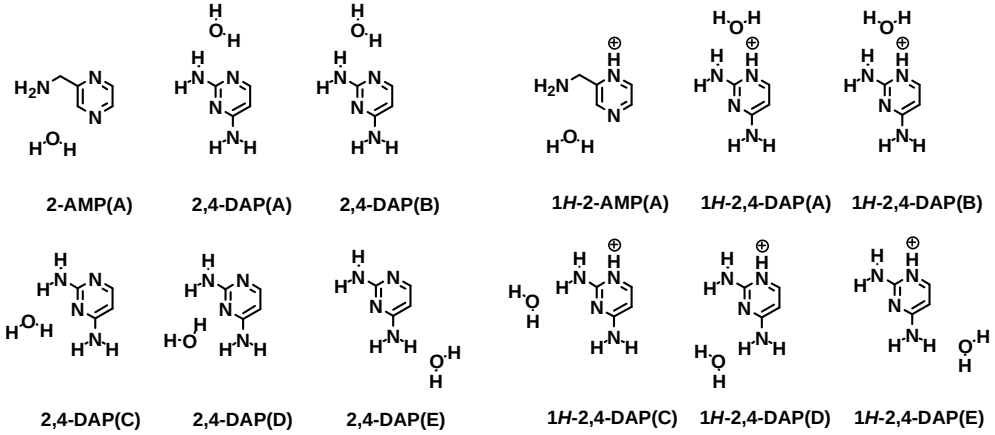
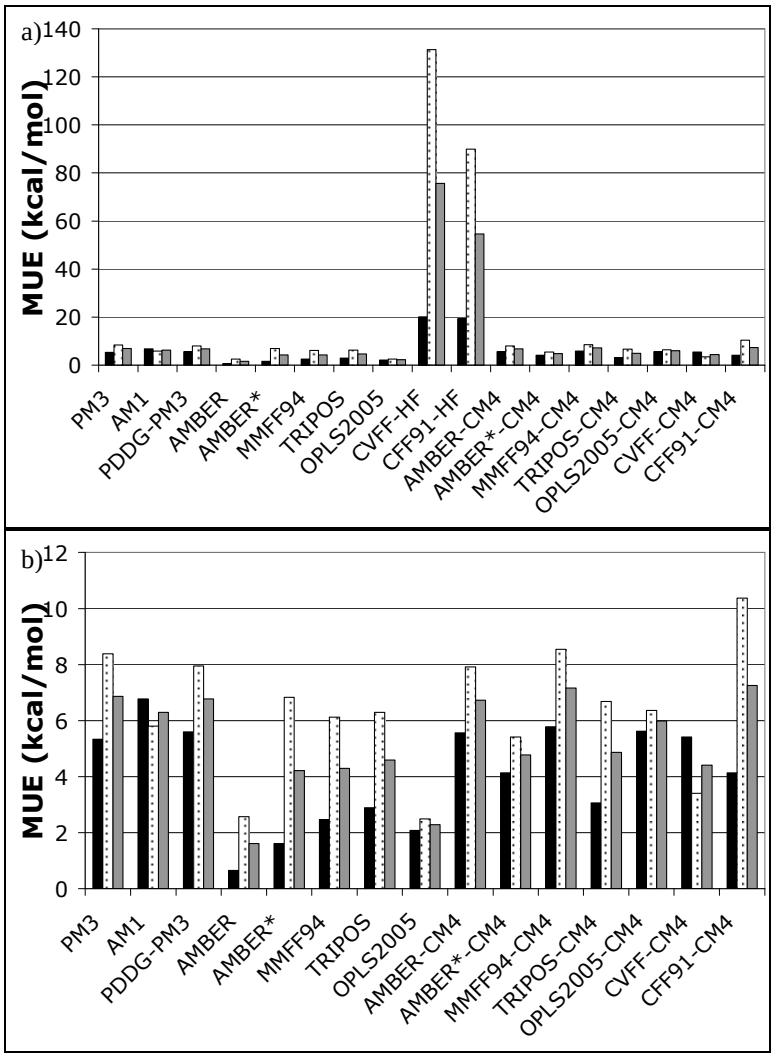


Figure 10. MUE in prediction of binding energy between 2-AMP, 1H-2-AMP, 2,4-DAP, and 1H-2,4-DAP and water molecules place at hydrogen bonding locations denoted in the DHFR-MTX crystal structure. A) includes all methods tested and b) omits CVFF-HF and CFF91-HF for easier viewing. Black – neutral species, dotted – charged species, gray – both sets of species.



The CVFF-HF and CFF91-HF force fields perform quite poorly, likely as a result of the necessary substitution of CHELPG charges for the force fields' default charges. Notably, the substitution of CM4 charges into the CVFF and CFF91 force fields again improves their modeling accuracy; however, this is not the case with the remaining MM force fields. Out of all the methods tested, the most accurate are the OPLS2005 and AMBER force fields, which predict binding energy for both sets of molecules to an MUE of 2.28 and 1.61 kcal/mol, respectively. In addition to its excellent accuracy, the OPLS2005 force field shows consistency between the neutral and charged species, calculating binding energy to an MUE of 2.07 and 2.49, respectively. The TRIPOS, MMFF94, and AMBER* force fields all predict binding energies about two-fold less accurately, with MUEs of 4.59, 4.29, and 4.22 kcal/mol, respectively.

3.7. Validation of the AMBER, AMBER*, MMFF94, and OPLS2005 Force Fields.

Tables 5 and 6 summarize our validation of the AMBER, AMBER*, MMFF94, and OPLS2005 force fields (as well as their MM-CM4 counterparts) in the gaseous and aqueous phases against both the molecules already included in the study as well as an additional set of pharmacophore-containing molecules and their cations shown in Figure 3. In contrast to its relatively poor performance assigning partial charges to 2-AMP, 1*H*-2-AMP, 2,4-DAP, and 1*H*-2,4,-DAP, the AMBER force field performs better in this case than OPLS2005, although the degree to which this increased performance is relevant is debatable. All three force fields excel at assigning partial charge to different

types of molecules (data not shown), so it is not clear which is consistently more accurate.

Table 6 summarizes the reduced deviance for each method in modeling molecular geometries. Whereas Fig. 22 showed large increases in performance when CM4 charges are substituted for the original charges of several MM methods, here we see a much more modest effect, although incorporating CM4 charges does improve the geometric modeling capabilities of three of the force fields. The MMFF94 and MMFF94-CM4 force fields are clearly superior when modeling the molecules included in the validation. With this in mind, we consider the MMFF94-CM4 force field as the most accurate force field for our system. The performance of the unmodified MMFF94 force field is also excellent, and it merits consideration because it is already well defined.

Table 5. Reduced Deviance ($D_{y,m}$) in Partial Charge for Force Field Validation of All Neutral and Charged Species in Fig. 3.

Method	Gas Phase	Aqueous Phase	Both Phases
AMBER*	0.68	0.69	0.68
MMFF94	1.04	1.04	1.04
AMBER	1.05	1.07	1.06
OPLS2005	1.18	1.19	1.18

Table 6. Reduced Deviance ($D_{y,m}$) in Combined Geometry for Force Field Validation of All Neutral and Charged Species in Fig. 3.

Method	Gas Phase	Aqueous Phase	Both Phases
MMFF94-CM4	0.73	0.79	0.76
MMFF94	0.80	0.85	0.82
OPLS2005-CM4	0.82	0.91	0.86
OPL2005	0.95	0.88	0.92
AMBER-CM4	0.90	0.93	0.92
AMBER	1.03	0.97	1.00
AMBER*	1.38	1.32	1.34
AMBER*-CM4	1.39	1.35	1.37

3.8. Comparison of CM4 and MMFF94 Charge Distribution Calculated for Methotrexate's Neutral and Cationic Forms. Table 7 contains the partial charge distribution of the gas-phase optimized structure of MTX (neutral and cationic) as calculated by M05-2X/6-31+G(d,p)/CM4 and MMFF94. The numbering systems for the molecules are given in Appendix One. Both the ionic and nonionic species' charge distributions are calculated to a MUE of 0.17 by MMFF94, validating this force field's good performance for such a large molecule.

Table 7. Partial Charge Distribution of Gas-Phase Neutral and Cationic MTX as Calculated by M05-2X/CM4 and MMFF94.

	MTX			MTX+				MTX			MTX+		
Atom	CM4	MMFF94	Res.	CM4	MMFF94	Res.	Atom	CM4	MMFF94	Res.	CM4	MMFF94	Res.
C1	0.44	0.72	0.28	0.51	0.77	0.26	C29	-0.12	0.09	0.21	-0.08	0.09	0.17
C2	0.29	0.41	0.12	0.27	0.41	0.14	H30	0.07	0.15	0.08	0.07	0.15	0.08
C3	0.07	0.31	0.24	0.09	0.31	0.22	H31	0.09	0.15	0.06	0.10	0.15	0.05
C4	0.29	0.62	0.33	0.35	0.67	0.32	C32	0.00	0.37	0.37	-0.01	0.37	0.38
C5	0.14	0.17	0.03	0.18	0.17	0.02	H33	0.06	0.00	0.06	0.07	0.00	0.07
C6	0.14	0.16	0.02	0.21	0.16	0.05	H34	0.06	0.00	0.06	0.06	0.00	0.06
H7	0.07	0.15	0.08	0.09	0.15	0.06	H35	0.06	0.00	0.06	0.06	0.00	0.06
N8	-0.59	-0.90	0.31	-0.52	-0.90	0.38	C36	0.36	0.54	0.19	0.33	0.54	0.22
H9	0.31	0.40	0.09	0.33	0.40	0.07	O37	-0.38	-0.57	0.19	-0.36	-0.57	0.22
H10	0.32	0.40	0.08	0.34	0.40	0.06	N38	-0.43	-0.73	0.30	-0.43	-0.73	0.30
N11	-0.61	-0.90	0.29	-0.55	-0.90	0.35	H39	0.26	0.37	0.11	0.26	0.37	0.11
H12	0.32	0.40	0.08	0.35	0.40	0.05	C40	0.11	0.36	0.26	0.11	0.36	0.25
H13	0.32	0.40	0.08	0.34	0.40	0.07	H41	0.07	0.00	0.07	0.07	0.00	0.07
N14	-0.44	-0.62	0.18	-0.41	-0.62	0.21	C42	-0.08	0.00	0.08	-0.07	0.00	0.07
N15	-0.44	-0.62	0.18	-0.40	-0.18	0.22	H43	0.07	0.00	0.07	0.07	0.00	0.07
N16	-0.32	-0.62	0.30	-0.31	-0.62	0.31	H44	0.07	0.00	0.07	0.07	0.00	0.07
N17	-0.29	-0.62	0.33	-0.28	-0.62	0.34	C45	-0.10	0.06	0.16	-0.10	0.06	0.16
C18	0.03	0.51	0.49	0.03	0.51	0.48	H46	0.09	0.00	0.09	0.09	0.00	0.09
H19	0.05	0.00	0.05	0.06	0.00	0.06	H47	0.09	0.00	0.09	0.09	0.00	0.09
H20	0.06	0.00	0.06	0.07	0.00	0.07	C48	0.29	0.66	0.37	0.29	0.66	0.37
N21	-0.33	-0.84	0.51	-0.34	-0.84	0.50	C49	0.29	0.66	0.37	0.29	0.66	0.37
C22	0.18	0.10	0.08	0.16	0.10	0.06	O50	-0.35	-0.57	0.22	-0.35	-0.57	0.22
C23	-0.11	-0.15	0.04	-0.12	-0.15	0.03	O51	-0.34	-0.57	0.23	-0.33	-0.57	0.24
C24	-0.11	-0.15	0.04	-0.10	-0.15	0.05	O52	-0.33	-0.65	0.33	-0.32	-0.65	0.33
C25	-0.05	-0.15	0.10	-0.04	-0.15	0.11	H53	0.32	0.50	0.18	0.33	0.50	0.18
H26	0.06	0.15	0.09	0.05	0.15	0.10	O54	-0.36	-0.65	0.29	-0.36	-0.65	0.29
C27	-0.02	-0.15	0.13	-0.01	-0.15	0.14	H55	0.32	0.5	0.18	0.32	0.50	0.18
H28	0.07	0.15	0.08	0.07	0.15	0.08	H56	N/A*	N/A*	N/A	0.33	0.46	0.13
							MUE			0.17			0.17

*Atom not present in the protonated molecule

4. Conclusions

We have studied 30 systems, each of which is studied with up to 31 methods in one or two phases (gaseous and aqueous). We found that the M05-2X density functional with the 6-31+G(d,p) basis set yields geometries very close to those obtained with coupled-cluster calculations. M05-2X is therefore useful in obtaining benchmark values for larger molecules involved in drug design.

The assignment of appropriate partial atomic charges is critical to accurate modeling of molecules by molecular mechanics. We found that substitution of CM4 charges for the original charge parameters of a given MM model improved the geometric accuracy of all seven force fields for which this substitution was tested, with some errors in geometry decreasing by factors of 3.5 and 4 in the two most dramatic cases. With the improved partial charge assignment, four of the MM methods come very close to reproducing coupled-cluster calculations.

Although the substitution of CM4 charges into our MM force fields improved geometric accuracy, it had the opposite effect on predicting binding energies for the majority of the force fields tested. Thus, the overall improvement of a MM force field does not lie solely in the improvement of one aspect of that method, and charge substitution should be used with care.

We have found that the MMFF94-CM4 force field, in which CM4 charges are substituted for the MMFF94 default charges, yields the most accurate geometries for representative fragments of methotrexate, as well for as an additional set of drug-like molecules. Furthermore, the MMFF94 force field without charge substitution exhibits

the second best geometric performance among the sixteen methods tested. However, in our binding energy studies, we find that excellent performance modeling geometries and charge distributions does not necessarily correlate directly to the prediction of energetics. Therefore, when the combined charge distribution, geometric, and energetic results are taken into account, we consider the MMFF94, AMBER/GAFF, AMBER*, and OPLS2005 force fields to be the most accurate and economical methods available for modeling small molecules containing nitrogen heterocycles and exocyclic amines. We expect these methods to be suitable for use in modeling more general nitrogen-containing small molecules as well as larger systems including the bis-methotrexate chemical inducer of dimerization, the protein-ligand complex, and the residues contained at the DHFR-DHFR interface.

Chapter Three

Protein Interface Remodeling in a Chemically Induced DHFR Dimer

1. Introduction

Molecular recognition plays a key role in the definition of the protein and nucleic acid interactions that create and operate living systems. The myriad of known protein-protein interactions have been extensively studied in an effort to uncover their diverse mechanisms and role in biological processes as well as develop therapeutic tools that can exploit such interactions for the treatment of disease.[216-218] Much effort has been put forth toward delineating the composition and role of interfacial amino acids,[219,220] the contribution of thermodynamic forces to protein complex stability,[221,222] and the mechanisms by which two proteins in a veritable sea of macromolecules go about recognizing each other.[223-225] Recent advances in bioinformatics have lead to ever-increasing mining of structural complementarity and sequence similarity data; however, the need for a model system to validate the hypotheses derived from such work has remained a constant and challenging priority.[226-228]

Protein interfaces have been well-studied in the past, and it is now understood that the composition of the protein interface is very similar to the rest of the protein surface.[229] While charged residues are somewhat less common at interfaces due to the net loss of energy due to desolvation, hydrophobic residues such as Phe, Trp, and Met are conserved in many cases.[162,223] The sidechain packing at protein interfaces is relatively compact, and this defining characteristic can often be used to differentiate true protein interfaces from crystallographic packing artifacts.[223,225] This packing density relates closely to the common incorporation of hydrophobic residues at interfaces as the residue sidechains orient in the densest possible manner to minimize contact with solvent.[230-232]

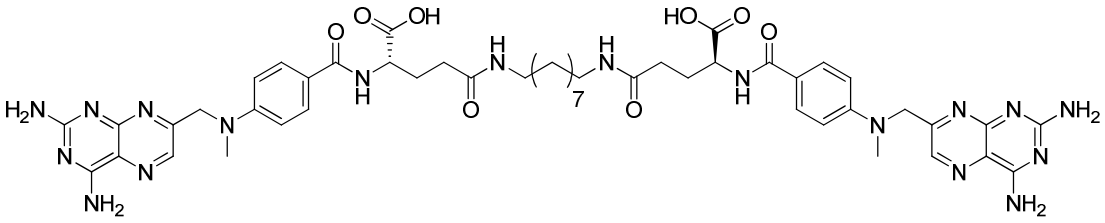
It has been shown that protein complex binding energy is frequently localized in interfacial hotspots, where several key residue interactions contribute the majority of the binding energy.[233-235] Much effort has been directed toward investigating the nature of these hotspots – notably by methods such as alanine scanning mutagenesis, in which residues are replaced serially with alanine in an effort to quantitate their contributions to overall binding energy.[235,236] On the theoretical front, computational alanine scanning[237] and free energy decomposition[165] have become effective methods for the determination of individual residue contributions to binding energy, even resolved to the net contributions from the backbone and sidechain independently. The realization that binding energies are localized in this manner is an important discovery that has the potential to drive research toward the design of tailored protein interfaces wherein practical structural and functional modifications may be achieved.

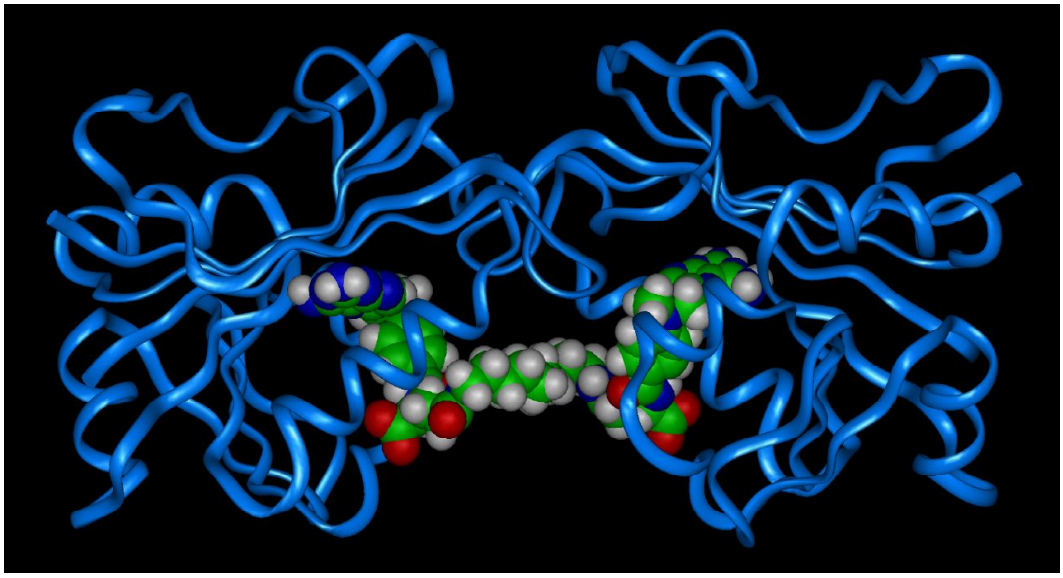
Toward the end of tailored protein interfaces, research has not kept pace with the analysis of individual residue energetics. Early work by Cunningham and Wells showed that replacement of key residues at the recognition site of human growth hormone (hGH) with alanine modified its affinity for its biological targets – the hGH and prolactin receptors.[238,239] While the wild-type enzyme binds each with equivalent affinity, replacement of the key residues yielded a 34,000-fold selectivity for the hGH receptor. Other investigations have utilized the mass reconfiguration of electrostatic interactions as well as the replacement of buried polar residues with nonpolar isosteres.[240-242] Further work assessing the interplay of electrostatic and hydrophobic interactions has also shown the stabilizing and destabilizing effects of these modifications.

In the context of therapeutics, protein interface remodeling is an attractive pursuit as the selective inhibition of protein complexation represents the ability to control a wide array of biological mechanisms. Small molecule inhibition would achieve this goal, in effect reshaping the interface to render it non-complementary.[243,244] However, this has remained an elusive target due to the topologically bland surfaces present at protein interfaces that often preclude selectivity. Protein surfaces simply lack the key conserved features amenable to rational design that are found in enzyme binding pockets.[81] However, avenues to circumvent these difficulties may become apparent given a proper model system.

With the advent of chemically induced dimerization, it has become possible to control the formation of a protein complex. Since the landmark work of Schrieber et al.[24], several dimerization systems have been well-studied. One such system is the chemically induced DHFR dimer, first described by Hu and coworkers.[31] The DHFR dimer is selectively assembled via the addition of a bivalent inhibitor of DHFR – bis-MTX-C9 (Figure 1). Our characterization of this system has uncovered several key aspects that highlight its suitability as a model system for investigating the effects of interfacial point mutations on dimer stability. First, as noted, dimerization only occurs in the presence of a specific ligand; second, the complex can only be disassembled via the addition of a competitive inhibitor of the dimerizer; third, the thermodynamics of

Figure 1. Structures of bis-MTX-C9 (MTX_2C9, top) and the chemically induced DHFR dimer (bottom).





complex assembly have been well-characterized.[31] These factors combine to yield a system well-suited to the study of the weak intermolecular interactions that dominate transient protein complexation.[245,246] Additionally, the relatively small surface area of the DHFR dimer interface (520 $Å^2$) benefits practical experimentation such that a lesser number of individual mutations should yield more pertinent observations in the course of modified dimer characterization.

In our laboratory, we have previously analyzed the importance of ligand conformational equilibria in the context of chemically induced protein dimerization.[32] Tangent to this discussion and the purpose of the current work is to characterize the role of inter-residue cooperativity present in the newly formed interface. Such interactions are exploited in an effort to perturb the stability of the novel dimer. Herein we present a method for the modulation and quantitation of interfacial cooperativity via point mutations at the chemically induced DHFR dimer interface.

2. Results and Discussion

2.1. Mutation Scheme Selection and K_d Analysis

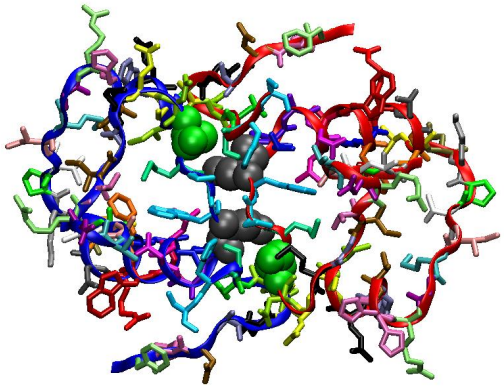
Examination of the DHFR crystal structure (PDB ID: 4DFR) reveals several candidates for interfacial mutations where residue sidechains interact across the C_2-symmetric interface. The set of interactions that characterize the dimer interface are summarized in Table 1. Inter-backbone interactions that make up a large part of the interface are not ready targets for mutagenesis. Three pairwise interactions, characterized by primarily sidechain interactions, are presented in Figure 2. Of these three, the Ala19 – Asn23 pair is an attractive target for initial experiments due to the

Table 1. Sites of interfacial contacts in the DHFR dimer interface. Residue pairs are classified by the type of contact: sidechain-sidechain (S-S), sidechain-backbone (S-B), and backbone-backbone (B-B).

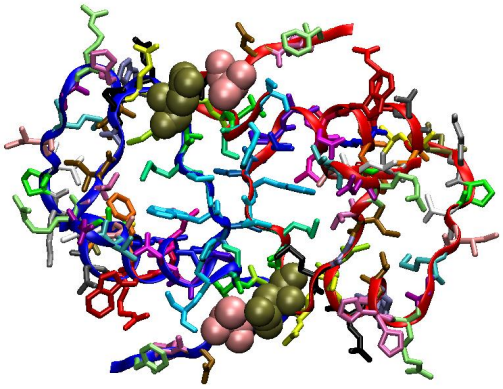
Residue A	Residue B	Type of Contact
Asn18	Ala143 Asp144	S-S S-B
Ala19	Asn23	S-S
Ala19	Asp144	B-B
Asn23	Asn23	S-S
Ala145	Glu48 Ser49	S-S
Gln146	Gly51 Glu48 Ser49	S-S
Ser148	Pro21	S-B, S-S

Figure 2. Views of key contacts at the ecDHFR dimer interface. The truncated protein shown represents the interfacial portion of the dimer, with residues of interest rendered as VDW. Different DHFR monomers are represented by the blue and red ribbons. The entire structure is rotated 90º about the longitudinal axis as seen in Figure 1. A) Ala19 (green) and Asn23 (gray). B) Asn18 (tan) and Ala143 (pink). C) Glu48 (purple), Ser49 (blue), and Ala145 (orange). D) All residues in a-c.

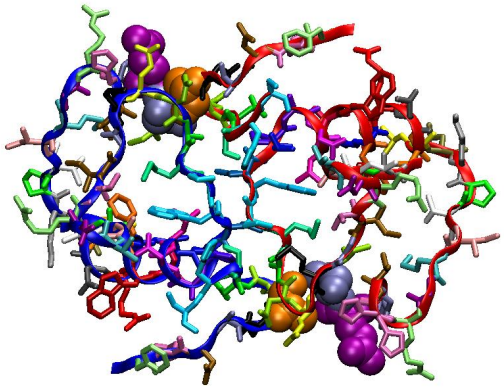
a.



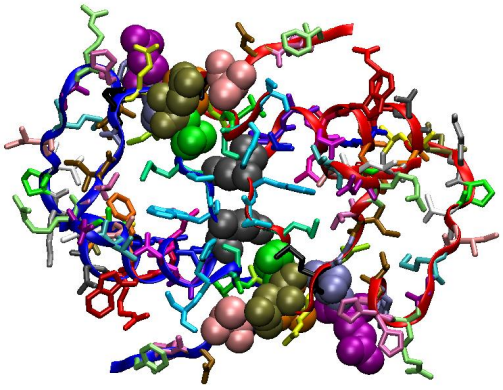
b.



c.



d.



close proximity of their sidechains (3-4 Å), their central location in the dimer interface, and the lack of a hydrogen bonding present in the pair. Other targets remain viable for future experiments. These design considerations lead us to hypothesize that mutations using group-representative amino acids at these positions would allow for the role of steric, hydrophobic, and electrostatic interactions in the thermodynamic stability of the dimer to be defined.

The residues selected for mutation are unique insomuch as they reside in a loop in the protein known in the literature as the M20 loop, named for the central methionine. This particular loop is significant as it plays a central role in ecDHFR catalysis.[247] Previous work evaluating the catalytic mechanism of ecDHFR has demonstrated the ability of the enzyme to tolerate substitutions of these two residues without compromising its catalytic activity.[248] This evidence further supported our design plans; however, we decided to verify these findings independently.

To more carefully ascertain the effects of mutations in M20 loop of DHFR, we performed an analysis of the binding affinity for DHFR to MTX via a fluorescence quenching assay (see Methods section). The relatively tight binding constant for DHFR (590 pM)[249] leads to a high degree of uncertainty in the estimate of K_d due to the narrow window of concentrations leading to a well-fitted binding curve. However, for all mutants except N23F, the K_d remains statistically unaltered (Table 2). In the case of N23F, the binding affinity shows an apparent 4-fold decrease, to approximately 2.9 nM. Reasons for this change in binding affinity may be attributed to long-range inter-residue interactions affecting the MTX binding pocket, decreased mobility of the M20 loop,

Table 2. WT and Mutant K_d data.[a]

Protein	K_d (pM)	Protein	K_d (pM)
WT	695 ± 253	A19Y	402 ± 106
N23Y	894 ± 267	A19S	606 ± 283
N23S	462 ± 269	A19Q	462 ± 166
N23Q	766 ± 326	A19L	583 ± 377
N23L	1141 ± 311	A19H	319 ± 176
N23F[b]	2972 ± 966	A19F	327 ± 136

[a]Data are for mutations developed by B.R. White. Mutants prepared by Jonathan C.T. Carlson are not shown.
[b]Result of two independent experiments

interaction of the hydrophobic residue with the pteridine ring of MTX, or a combination of all three. While this perturbation in binding affinity may confound competition assay results at lower concentrations, the relatively high concentration of enzyme at which the assay is performed renders this small change irrelevant in the context of our results.

2.2. Competition Experiments

In order to quantitate the degree to which point mutations stabilize or destabilize the chemically induced DHFR dimer (DHFR CID), we have developed a competition assay wherein a pre-equilibrated DHFR CID is denatured with increasing equivalents of MTX, leading to a curve which can be fit to the following equation:

$$[E_2D] = \frac{K_c K_{a1} K_{a2} (0.5 - [E_2D])(1-2[E_2D])^2}{K_{eq} K_{aMTX}{}^2 (M_t - E_t + 2[E_2D])^2} \quad (1)$$

In this equation (derived in Appendix Two), K_{a1} and K_{a2} are the binding affinities for the first and second bis-MTX binding events, and are assumed to be equal to K_{aMTX}. K_c and K_{eq} are the cooperativity and dimerizer equilibrium constants, respectively; M_t is the total added MTX concentration; E_t is the total enzyme concentration; and $[E_2D]$ represents the experimentally observed dimer concentration. The relative stability of the mutant dimer complexes represents a modification of the value of K_c, since the K_{eq} for the dimerizer remains constant over the analysis. Therefore, comparison of the K_{eq}/K_c ratio found for each mutant to the wild-type dimer represents the relative effects of point mutations at the interface on cooperativity. These effects can be quantified in terms of energy by utilizing the equation:

$$\Delta\Delta G = -RT \ln\left(\frac{K_{eq}/K_{c,mut}}{K_{eq}/K_{c,wt}}\right) \qquad (2)$$

A typical competition curve, including model denaturation curves based on several K_{eq}/K_c ratios appears in Figure 3. It is apparent that the lower the value of K_c (and hence the larger the value of the ratio), the less stable the induced dimer, and the easier it is to denature the complex.

2.3. Interfacial Mutations Modulate Dimer Stability

The results of competition experiments probing the effects of point mutations on the cooperativity in the DHFR CID are tabulated in Table 3 and shown graphically in Figure 4. Ratios of the mutant:WT K_{eq}/K_c values and the associated energy perturbations are shown in Table 4. Globally, the data spans a dynamic range of cooperativity from 0.35 – 4.44 fold destabilization. The data reveal several trends. First, all mutations of Ala19 are destabilizing in nature, the least being A19Y, with a mutant:WT ratio of 1.35. Examination of the DHFR crystal structure, which is isomorphous the dimerized DHFR crystal structure (obtained from Dr. Vivian Cody, University at Buffalo, Hauptman Woodward Institute, not yet deposited in the protein data bank) shows that the conformations of Ala19 and Asn23 are oriented such that Ala19 is buried within the protein interface, and as such, is likely less tolerant of modification (Figure 1). In fact, it is the introduction of charge-charge repulsion that affects dimerization more severely than steric bulk (A19E and A19K, ratios of 2.22 and 3.52, respectively), likely due to the forced close proximity of the charges and the inability of Ala19 to shift to a more stable

Figure 3. Typical competition denaturation curve highlighting K_{eq}/K_c ratios of 0.1, 1.4, and 10. Data points for wild-type denaturation represent the effect of adding 0.5 to 5.0 equivalents of MTX to a MTX_2C9-induced dimer. Error bars are derived from the standard deviation in three independent experiments. Fit lines display the effects of varying the K_{eq}/K_c ratio in Equation 1.

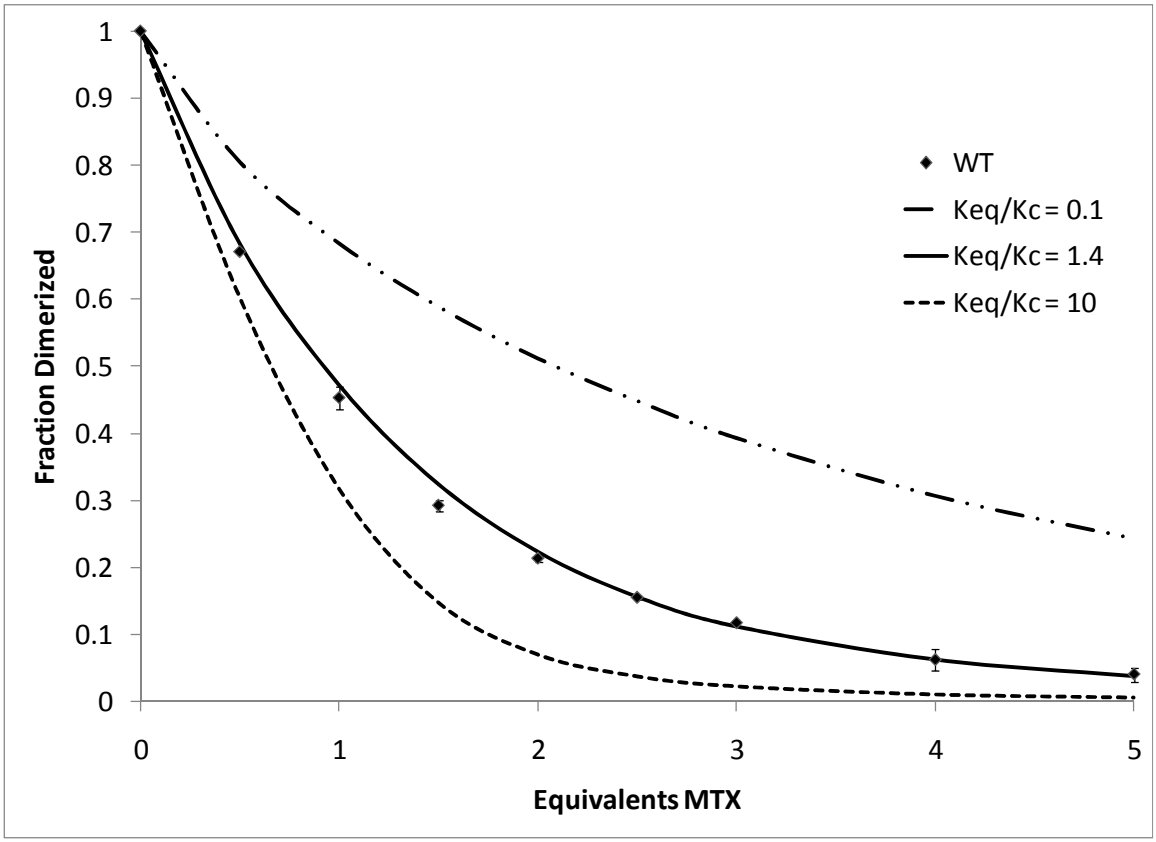


Table 3. Compiled K_{eq}/K_c ratios for WT and mutant DHFR dimers. See text for a discussion of estimation of error.

Protein	K_{eq}/K_c Ratio	Error Estimate – Range	% Error
WT	1.35	1.20 – 1.60	14.8
A19E	3.00	2.40 – 3.80	23.3
A19F	2.09	1.70 – 2.30	14.3
A19H	2.82	2.40 – 3.30	16.0
A19K	4.75	4.00 – 6.00	21.1
A19L	2.41	2.20 – 2.75	11.4
A19N	4.50	4.15 – 4.75	6.7
A19Q	2.46	2.00 – 3.05	21.4
A19S	2.17	1.55 – 3.00	33.5
A19Y	1.82	1.65 – 2.20	15.1
N23A	1.20	1.05 – 1.35	12.5
N23E	5.50	4.50 – 6.50	18.2
N23F	0.47	0.40 – 0.55	15.9
N23H	2.75	2.45 – 3.30	15.5
N23K	6.00	4.75 – 7.75	25.0
N23L	0.77	0.70 – 0.85	9.8
N23Q	0.51	0.35 – 0.60	24.8
N23S	1.24	1.05 – 1.35	12.1
N23Y	0.74	0.65 – 0.85	13.5

Figure 4. Results of competition assays for each DHFR variant studied. Error bars represent the range of ratios determined from error analysis (see text).

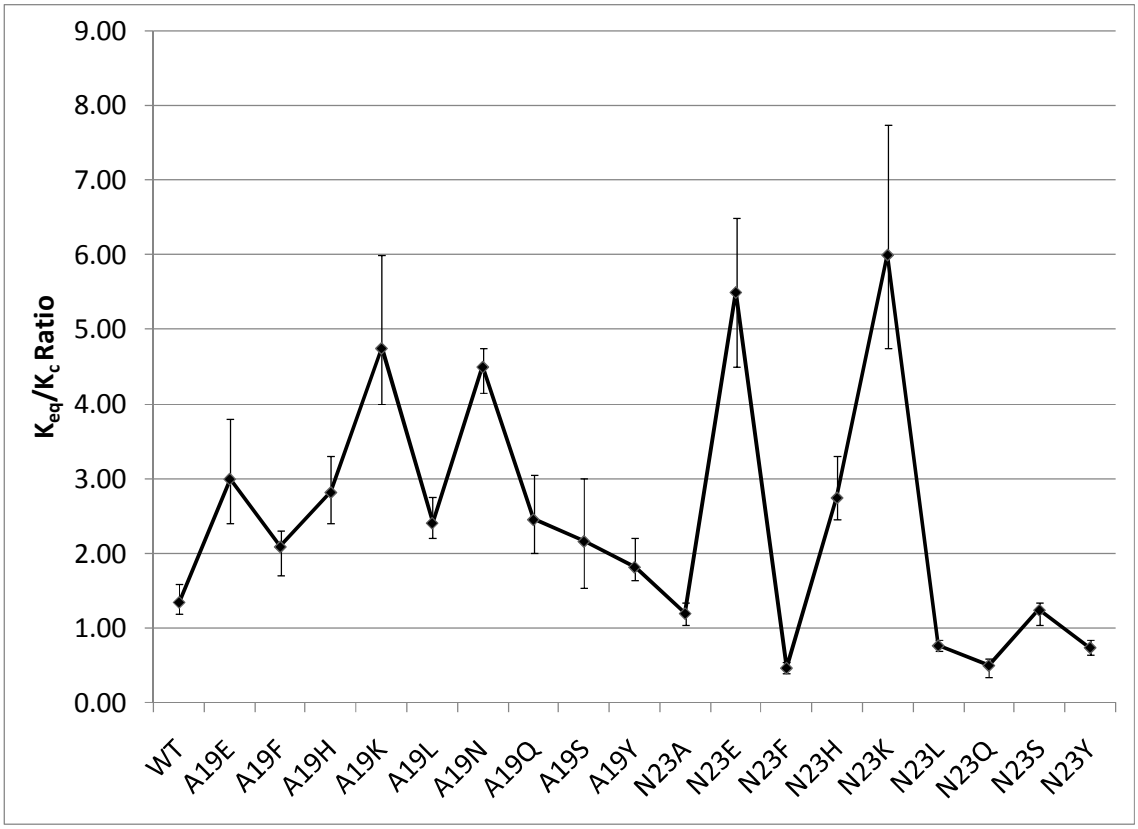


Table 4. Ratios of mutant:WT K_{eq}/K_c values and associated ΔΔG values.

Protein	Mutant/WT K_{eq}/K_c Ratio	ΔΔG (kcal/mol)	Protein	Mutant/WT K_{eq}/K_c Ratio	ΔΔG (kcal/mol)
A19E	2.22	0.47 ± 0.14	N23A	0.89	-0.07 ± 0.07
A19F	1.55	0.26 ± 0.09	N23E	4.07	0.83 ± 0.11
A19H	2.09	0.44 ± 0.09	N23F	0.35	-0.62 ± 0.09
A19K	3.52	0.74 ± 0.12	N23H	2.04	0.42 ± 0.09
A19L	1.79	0.34 ± 0.07	N23K	4.44	0.88 ± 0.14
A19N	3.33	0.71 ± 0.04	N23L	0.57	-0.33 ± 0.06
A19Q	1.82	0.35 ± 0.12	N23Q	0.37	-0.58 ± 0.16
A19S	1.60	0.28 ± 0.20	N23S	0.92	-0.05 ± 0.07
A19Y	1.35	0.18 ± 0.09	N23Y	0.55	-0.35 ± 0.08

conformation. Other mutations at Ala19 have a pronounced effect, primarily A19N and A19H. The polar character of these mutations appears to be a major contributor to the destabilization found in these cases, and could be attributed to an increased desolvation penalty associated with the formation of the interface.

Perturbation of Asn23 yields similar results in terms of charge-charge repulsion (see N23E and N23K); however, most mutations at this position are relatively stabilizing. Examination of the crystal structure indicates that Asn23 is more spatially accommodating than Ala19, as it is capable of reorganizing into the solvent-occupied area surrounding the MTX binding pocket. The implications of this are twofold and are supported by the data. First, while a lower-energy dimer may stably form, the reorientation of Lys23 or Glu23 back into the interior of the interface to escape the desolvation penalty associated with the presence of ionic residues in a solvent-accessible area results in charge-charge repulsion. Second, the introduction of hydrophobic residues helps to stabilize the interface as hydrophobic interactions close to solvent potentially assist in restricting the conformational freedom of the interface, reinforcing interactions in the local area.

In both cases, the case of the histidine mutation raises interesting questions. At pH 7.0, near the intrinsic pKa of histidine (6.0 – 6.5), the charge state of this residue will be highly dependent on the environment, and could serve as either a hydrogen bond acceptor in the unprotonated state or a charge center if protonated. At pH 7.0, our experiments indicate that the level of destabilization associated with the His mutation (approximately twofold), while only barely approaching that of charge repulsion as observed in ionic

mutations (average 3.6-fold), is not stabilizing as seen in other hydrophobic residues. Competition experiments at pH 6.0 show a higher level of destabilization ($\Delta\Delta G$ = 0.85 kcal/mol, data not shown) when His is much more likely to be unprotonated, indicating that charge repulsion is likely not the cause of destabilization at pH 7.0. From this it can be reasoned that the His residue at the interface is neutral at pH 7.0, and the net destabilizing effects must be attributed to some other factor.

While the effects of charge repulsion at the interface are pronounced, the effects of steric bulk and hydrophobicity are more subtle and require in-depth examination. To correlate the general trends associated with steric bulk and hydrophobicity with change in cooperativity, we assessed total residue 19/23 sidechain volume using the data tables from Tsai, et al.[250] Surprisingly, in terms of steric bulk, no correlation exists between sidechain volume and $\Delta\Delta G$ (Figure 5). While it is apparent from the crystal structure that mutations at Asn23 have the ability to move relatively freely and are likely less sensitive to sidechain bulk, this is particularly surprising for Ala19, given its location deep within the interface. This finding is a testament to the conformational flexibility apparently inherent to this locale in the DHFR interface, and it can be expected that further attempts to destabilize the interface through point mutations should rely primarily on other forms of interaction such as ionic pairing or polarity.

For exploring correlation between mean 19/23 sidechain hydrophobicity and K_{eq}/K_c, we referenced the quantitative measure of hydrophobicity given by Carugo, et al.[251] A perhaps unsurprising weakly negative (stabilizing) correlation is present in this comparison (Figure 6). Many protein interfaces show a high hydrophobic character, with

Figure 5. Total residue 19/23 sidechain volumes (top) and their relationship to ΔΔG (bottom).

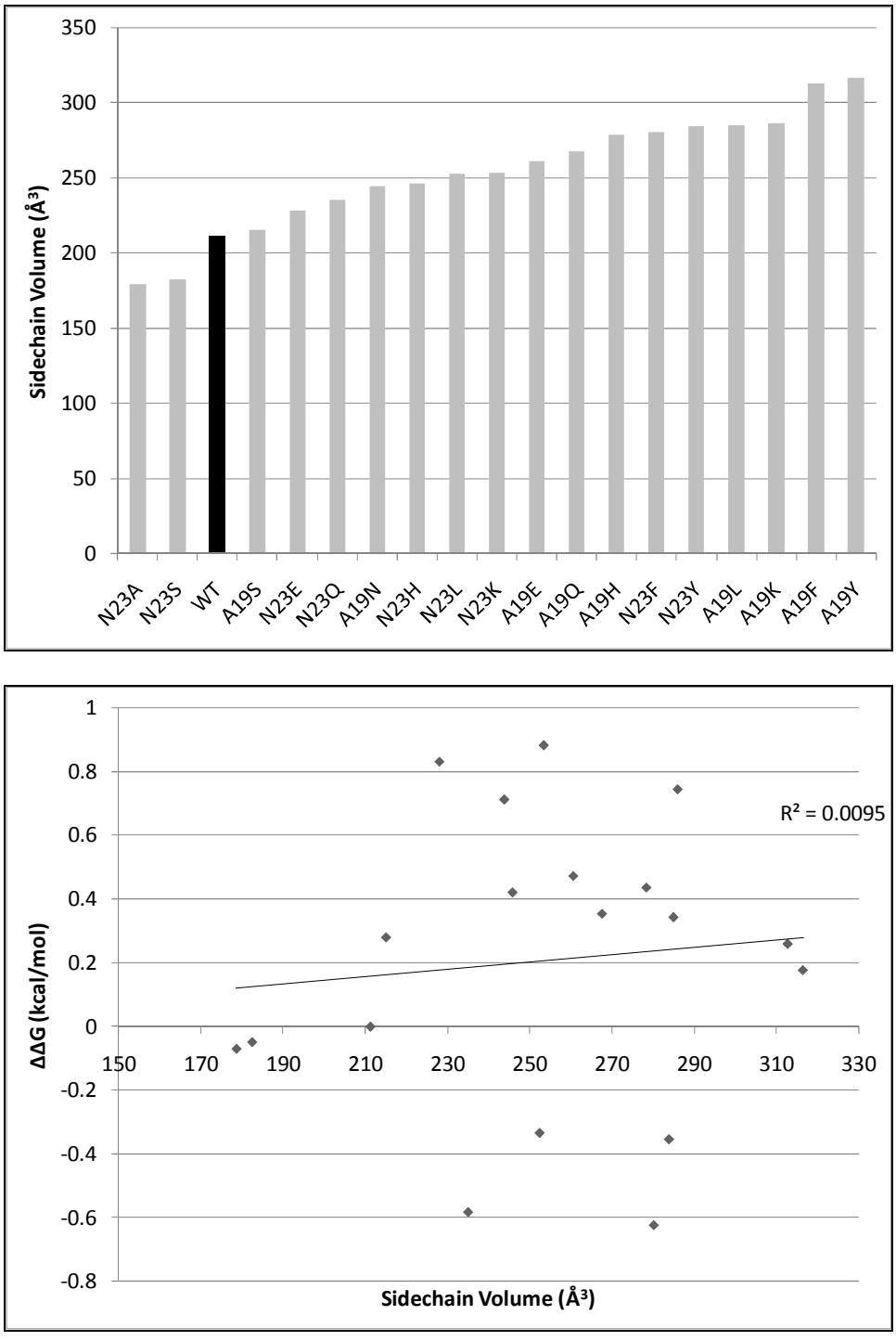
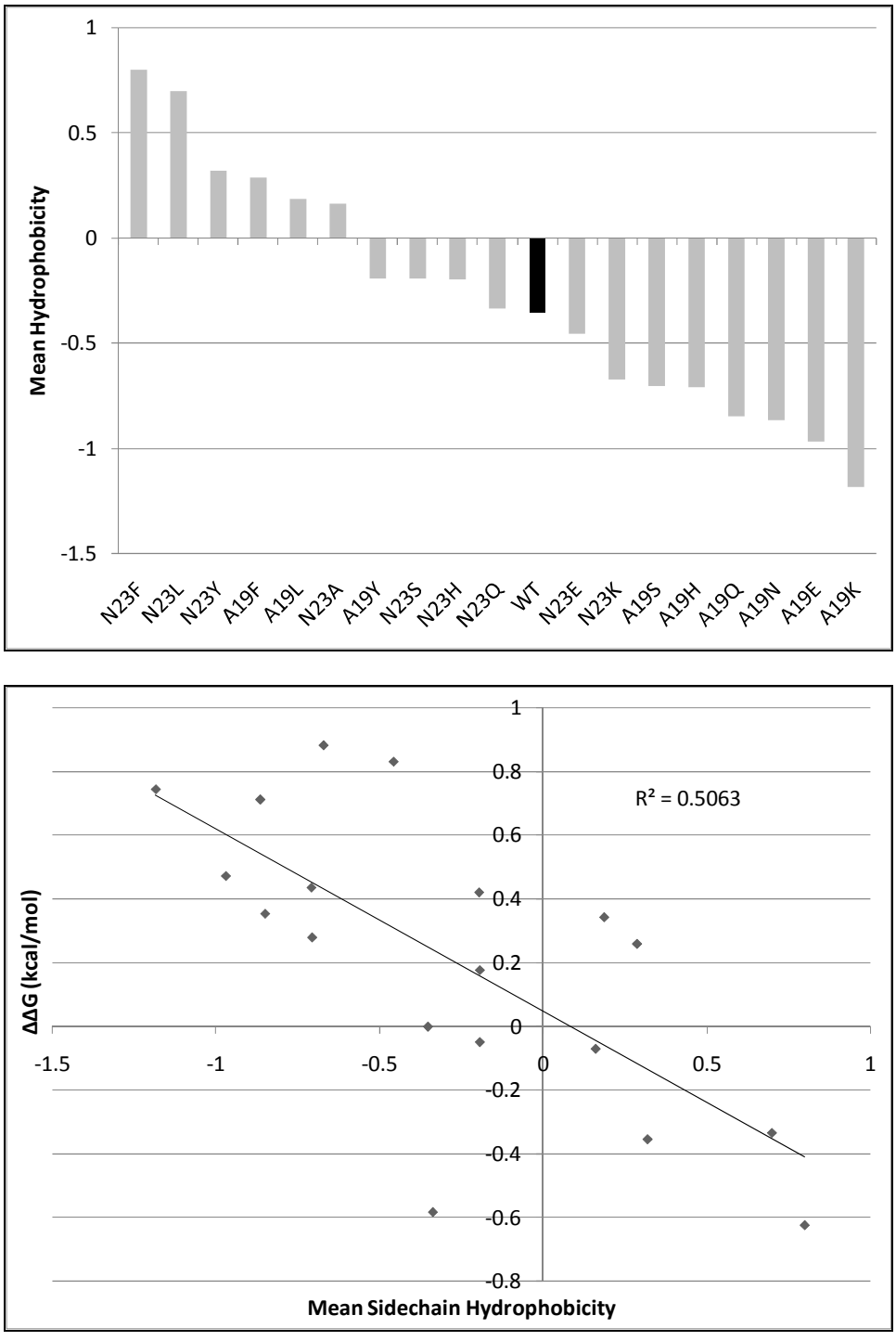


Figure 6. Residue 19/23 mean hydrophobicity (top) and negative (stabilizing) relationship to ΔΔG (bottom).



some of the most highly conserved residues at protein interfaces being Trp, Phe, and Met.[162] The low desolvation penalty for a hydrophobic patch on the surface of the protein combined with the tendency to segregate and stabilize away from solvent support the stabilizing effect of introducing hydrophobic residues into the DHFR interface. In contrast, introducing highly charged or polar residues will achieve interface destabilization and serve to decrease K_c.

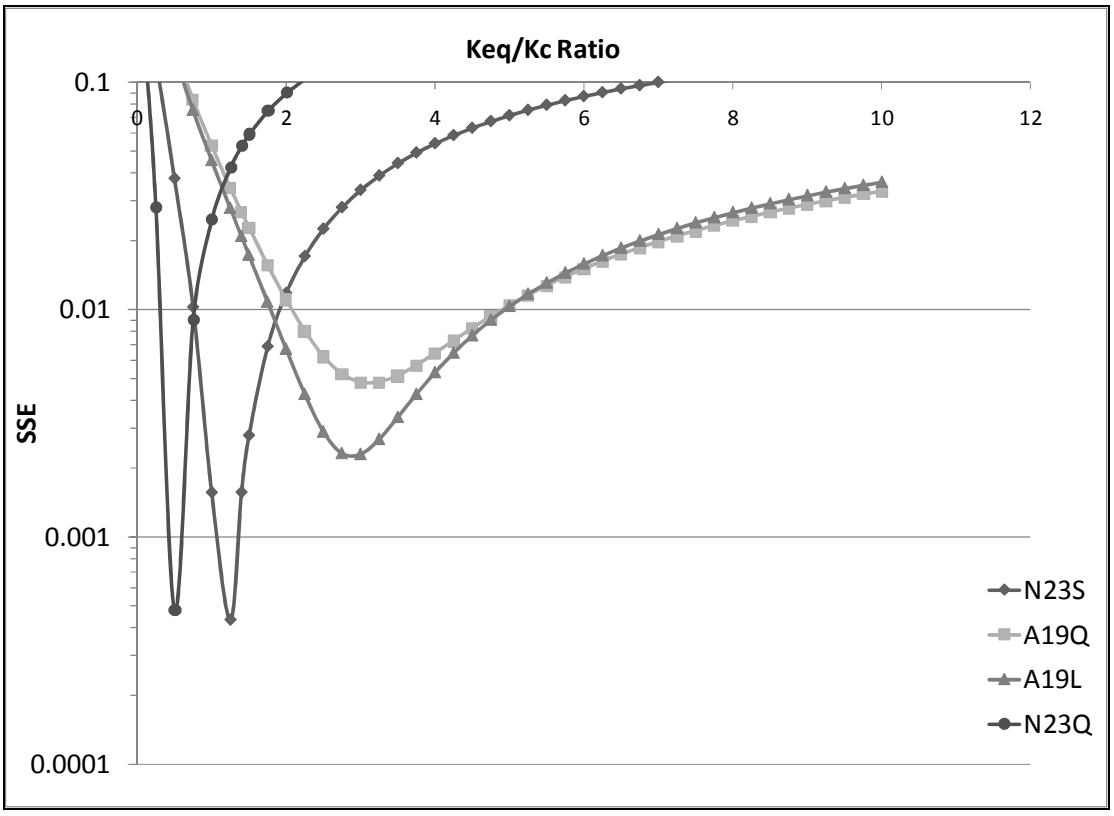
2.4. Data Fitting and Error Analysis

Throughout the development of the competition assay, it has become apparent that there are four main contributors to error introduced in the estimation of K_{eq}/K_c – sample preparation, chromatographic separation, peak area integration, and model data fitting. Of these sources, the first and the last are the most significant. While careful sample preparation yields reproducible results, any human error (i.e. the miscalculation of MTX added) has the potential to introduce large errors in the final ratio estimate due to the sensitivity of the dimer to small changes in MTX concentration. Repeated gel filtration and peak integration of a single sample yields variations of less than 2 percent. Due to the complexity of the cubic equation necessary to fit the equilibrium data, the final source of error necessitates an unusual fitting procedure. Since we were unable to derive an analytic solution for Equation 1, Mathematica (Wolfram Research) was used to generate a series of model data based on various K_{eq}/K_c ratios. This model data was used to construct a reference table of possible denaturation curves, and the solver module in Microsoft Excel was used to fit the data by selecting the value for K_{eq}/K_c that yielded the lowest sum of squared errors (SSE). The SSE value provides a measurement by which

the quality of fit can be quantitated; however, it does not readily provide an estimate of the error in the fitted parameter. In order to generate an estimate of this error, the fit quality is evaluated graphically, by plotting SSE as a function of the K_{eq}/K_c ratio (see Figure 7 for representative error estimates). The error estimate is then yielded by calculating the range of K_{eq}/K_c values that produced a SSE within 25% of the best-fit value. This 25% range was chosen arbitrarily and was based on both visual inspection of the curve behavior and the range of standard deviations observed for the wild-type data.

Upon inspection of the graphical error analysis, it becomes apparent that the error estimate is not symmetrical, with the lower bound being estimated more precisely than the upper. Additionally, the sensitivity of the experiment to detect small changes in K_{eq}/K_c, especially when this value is small, is highlighted. When the K_{eq}/K_c value is large, the equilibrium approaches the asymptote of stoichiometric dissociation, and the estimate of the value becomes poorer. While this may pose a problem for future experiments in which K_c is all but abolished as a contributor to dimerization, this assay will be highly sensitive to more stable dimers formed as interface remodeling is developed. In cases where minimal values of K_c are required, however, the use of an inhibitor such as trimethoprim, with a lower binding affinity than MTX, may increase the resolving power of the assay.

Figure 7. Competition data error analysis. For each experiment, fit quality (SSE) is plotted against K_{eq}/K_c value and graphed on a logarithmic scale and the shape of the curve used to evaluate the precision of the fit. A deep well indicates an accurate fit; a narrow well indicates a sensitive value for K_{eq}/K_c, increasing precision; and a broad, shallow well indicates a poorer fit and less precise data.



3. Conclusions

The series of mutations we have engineered at the DHFR interface have demonstrated an ability to modulate dimer stability over a range of at least a 1.5 kcal/mol range or nearly an order of magnitude change in the cooperativity equilibrium constant. Although this modulation falls short of the energy differences obtained for the hGH receptor (6.1 kcal/mol)[239], when considered relative to the comparatively small scale of DHFR-DHFR interface cooperativity, estimated at $\Delta G \leq -3.1$ kcal/mol, this represents a significant change of ~50%.[32]

In the course of our study, we have characterized the effects of representative amino acid point mutations at the DHFR dimer interface. The modest correlation between sidechain hydrophobicity and increased stability reinforces previous findings that hydrophobic hotspots tend to be conserved among protein interfaces, likely due to a lessened desolvation penalty and a gain in enthalpy associated with tighter binding. However, this affinity is likely to be bounded to an extent by the entropic penalties associated with an increase in the rigidity of the interface.[252] In contrast, destabilization of the interface can be best achieved by introducing electrostatic repulsion. Interestingly, the introduction of steric bulk seems to be overshadowed by hydrophobic and electrostatic effects, and can be dismissed as an effective means of interface modulation.

In terms of the development of a model system, we have demonstrated the utility of the chemically induced DHFR dimer as a platform for testing the effects of mutations on protein cooperativity. The competition assay represents a highly sensitive method of quantitating mutation effects, especially if the desired outcome is a highly stable interface

and increased values for K_c. If characterization of a highly destabilized interface is required, although sensitivity is only moderately decreased in the current model, employing a tighter-binding dimerizer (i.e. trimethoprim-based) would notably increase precision. Due to the favorable energetics associated with chemically induced dimer formation, even in the absence of protein cooperativity, our model represents an advantage insomuch as highly destabilizing interactions can still be quantitated. In other native dimer systems, excessive destabilization can yield a completely disrupted complex, precluding high-resolution study of the interface.

While the development of this model system assists in the challenge of establishing a sensitive platform for cooperativity testing, a number of intriguing possibilities exist for practical application. In the context of a dimer energy landscape, there exists three ways to assemble a dimer from two monomers – AA, AB, and BB. When examining homodimerization, the individual components are the same and there is no selectivity for AB formation. However, if the end product desired is a heterodimer, there are two methods for achieving selectivity – relative stabilization of the AB complex, resulting in a minimum energy associated with AB formation; or destabilization of either AA or BB, resulting in a net increase in homodimer free energy.[241,253,254] Stabilization of a heterodimeric species paves the way for the development of a biomolecular language of protein self-assembly, leading to user-directed control over the assembly of protein nanostructures and bispecific targeting platforms.[106] While literature methods currently rely on a primarily *ligand*-directed methods of achieving heterodimerization (i.e. rapamycin, which intrinsically targets two different

proteins)[36,42,43], a method relying on *protein*-directed heterodimerization would represent an elegant, conceptually attractive avenue for increased control over protein interactions and assembly.

4. Materials and Methods

General

Kits for site-directed mutagenesis and plasmid DNA preparation were obtained from Stratagene and Invitrogen, respectively. Oligonucleotides used as primers in the mutagenesis reaction were obtained from Integrated DNA Technologies through the University of Minnesota BioMedical Genomics Center. Methotrexate, NADPH, and MTX-agarose were purchased from Sigma-Aldrich. Anion exchange chromatography was performed using DE52 DEAE cellulose purchased from Whatman. Competent JM-109 *E. coli* cells were purchased from Promega (Madison, WI). Salts for buffer preparation were of reagent grade and purchased from Mallinckrodt, Fisher, or Sigma-Aldrich. C9-bis-MTX was synthesized as previously described and purified to ≥99% purity.[32] DHF was prepared fresh as previously described and stored under argon at -80ºC.[255] All other reagents were of reagent grade or better and purchased from Sigma-Aldrich.

4.1. Protein Expression, Purification, and Characterization

Site-Directed Mutagenesis

To generate mutant *ec*DHFR plasmids, the QuickChange protocol from Strategene was utilized. In short, complementary primer oligonucleotides bearing the mutations of interest are bound to the parent plasmid and PCR cycling achieves exponential generation of the mutated plasmid. Mutated plasmid DNA is recovered from transformed XL1-Blue *E. coli* via the PureLink HiPure Plasmid Miniprep Kit from Invitrogen. Sequencing of the mutated plasmid by the University of Minnesota Microchemical Facility is used to verify the presence of the mutation. Constructs were generated from the pTZwt1-3 plasmid, a gift from the lab of Virginia F. Smith, Pennsylvania State University, Department of Chemistry. Oligonucleotides (reverse primer sequence is complementary to forward) used to introduce the mutations are listed below:

A19H; 5’-C GTT ATC GGC ATG GAA AAC CAC ATG CCA TGG-3’

A19F; 5’-C GTT ATC GGC ATG GAA AAC TTC ATG CCA TGG-3’

A19Y; 5’-C GTT ATC GGC ATG GAA AAC TAC ATG CCA TGG-3’

A19S; 5’-C GTT ATC GGC ATG GAA AAC TCC ATG CCA TGG-3’

A19L; 5’-C GTT ATC GGC ATG GAA AAC CTC ATG CCA TGG-3’

A19Q; 5’-C GTT ATC GGC ATG GAA AAC CAG ATG CCA TGG-3’

A19K; 5’-CGC GTT ATC GGC ATG GAA AAC AAG ATG CCA TGG-3’

A19E; 5’-C GTT ATC GGC ATG GAA AAC GAG ATG CCA TGG-3’

N23F; 5’-G CCA TGG TTC CTG CCT GCA GAT CTC GCC TGG-3’

N23Y; 5’-G CCA TGG TAC CTG CCT GCA GAT CTC GCC TGG-3’

N23S; 5’-G CCA TGG AGC CTG CCT GCA GAT CTC GCC TGG-3’

N23L; 5’-G CCA TGG CTC CTG CCT GCA GAT CTC GCC TGG-3’

N23Q; 5’-G CCA TGG CAG CTG CCT GCA GAT CTC GCC TGG-3’

N23H; 5’-G CCA TGG CAC CTG CCT GCA GAT CTC GCC TGG-3’

N23K; 5’-G CCA TGG AAG CTG CCT GCA GAT CTC GCC TGG-3’

N23E; 5’-G CCA TGG GAG CTG CCT GCA GAT CTC GCC TGG-3’

Protein Expression and Purification

Mutant plasmid DNA was transformed into the JM109 *E. coli* expression line. Resulting colonies were inoculated into 4 mL LB broth containing 100 µg/mL ampicillin and grown at 37˚C overnight with shaking at 250 rpm. Glycerol was added to cell cultures to a final concentration of 15% (v/v) and stocks were frozen at -80˚C until use.

For protein expression, starter cultures were prepared using 4 mL LB broth containing 100 µg/mL ampicillin, 20 µg/mL trimethoprim, and a 40 µL inoculation of JM-109 cells bearing the plasmid of interest. These cultures were grown for a minimum of 8 hours at 37˚C with shaking at 250 rpm before a 500 µL aliquot was transferred to 50 mL LB containing the same antibiotics and grown for a minimum of 8 hours under the same conditions. 1 L LB broth containing 100 µg/mL ampicillin was inoculated with 10 mL of the 50 mL culture and grown for a minimum of 12 hours under the same growth conditions. The cell OD600 typically reaches >1.3 during this period.

Cells were recovered via centrifugation at 7500g for 15 minutes, then the cells lysed via a 30 min incubation in lysis buffer (10 mM KH_2PO_4, 100 µM EDTA, 1 mM DTT, 1

mg/mL Lysozyme, pH 8.0) containing 1 Complete© protease inhibitor tablet (Roche) and 8 x 15 seconds sonication. The crude lysate was centrifuged at 40,000g for 40 min at 4˚C and the soluble fraction subjected to the addition of 30% (w/v) $(NH_4)_2SO_4$ over 30 min at 4˚C with vigorous stirring. Soluble protein was recovered via centrifugation at 40,000g for 20 min at 4˚C. The lysate was dialyzed against 4 L equilibration buffer (10 mM KH_2PO_4, 0.1 mM EDTA, 0.5 M KCl, 0.5 mM DTT, pH 6.0) for a minimum of 4 hours at 4˚C, then loaded onto a methotrexate agarose column. The bound protein was washed with a high salt buffer (50 mM KH_2PO_4, 1 mM EDTA, 1 M KCl, 1 mM DTT, pH 6.0) until the A280 and A260 of the eluate is ≤0.1, at which time the protein was eluted with folate elution buffer (50 mM KH_2PO_4, 1 mM EDTA, 1 M KCl, 3 mM folic acid, 1 mM DTT, pH 9.0).

Fractions containing DHFR activity as assayed by previous methods were pooled and dialyzed against 4 x 4 L of Tris dialysis buffer (50 mM Tris, 1 M NaCl, 1 mM EDTA, 0.5 mM DTT, pH 7.2) for a minimum of 4 hours at 4˚C each. A final dialysis against 4 L of DEAE equilibration buffer (10 mM Tris, 1 mM EDTA, 1 mM DTT, pH 7.2) for a minimum of 4 hours at 4˚C prepared the protein for loading onto a DEAE anion exchange column. The protein was eluted with a gradient of 0 – 40% buffer B over 300 minutes, then 40 – 100% buffer B over the next 420 minutes. Buffer A is the equilibration buffer described above, and buffer B is DEAE elution buffer (10 mM Tris, 1 mM EDTA, 0.5 M KCl, 1 mM DTT, pH 7.2). Collected fractions were measured for A280, A260, and DHFR activity. Purified DHFR was concentrated to ~1 mg/mL via Amicon centrifugal ultrafiltration devices and stored at 4˚C until use. Typical protein

yield is 5-15 mg per liter of LB culture. The purity of the proteins was assayed by gel filtration and SDS-PAGE electrophoresis and the mass of the wild-type and mutant proteins verified via LC-MS.

DHFR Activity Assay

MTEN buffer (50 mM MED, 25 nM Tris, 0.1 M NaCl, 25 mM Ethanolamine, pH 7.0) NADPH stockl solution to a final concentration of 100 µM, and an enzyme sample are mixed to a final volume of 1 mL minus the necessary volume of DHF addition. After a 2 minute incubation, a baseline reading at 340 nm is taken to verify zero activity. DHF is then added from a concentrated master stock to a final concentration of 50 µM, the sample is mixed, and the absorbance is read at 340 nm for 1 minute. The rate of absorbance decline corresponds to V_o in µM/min and is calculated with the known extinction coefficient for the DHFR catalyzed reaction, 11,300 $M^{-1}cm^{-1}$.[256] DHF and NADPH master stock concentrations are estimated spectrophotometrically using the reagents' extinction coefficients at 280 and 340 nm, repectively.

Mutant DHFR K_d Assay

To assay the affinity for MTX to mutant DHFR, a fluorescence assay measuring the quenching of DHFR fluorescence upon MTX binding was employed. DHFR was diluted to a final concentration of 50 nM in 4 mL MTEN buffer and a baseline fluorescence reading taken, scanning emission from 300-400 nm with excitation at 290nm. Serial additions of MTX were performed, and the emission at 340 nm recorded. Data were fit using JMP-IN 4.0 (SAS Institute). The K_d for all but one mutant (N23F) was statistically unaltered from that of wild-type DHFR (0.590 nM).[249]

Protein Gel Filtration

Gel filtration samples were prepared as a 5 µM final DHFR concentration in P500 buffer (0.5M NaCl, 50 mM KH_2PO_4, 1 mM EDTA, pH7.0) with 5% (v/v) glycerol. The samples were loaded on to a Sephadex G-75 column (GE Biosciences) on a Beckman System Gold HPLC and eluted at 0.5 mL/min with P500 buffer. The relative peak intensities were quantitated by absorbance at 280 nm.

4.2. Protein Concentration Assays

Three methods were employed to obtain accurate protein concentration. First the Bradford assay was used to estimate the concentration of the protein sample. Second, the A280 of diluted DHFR samples was measured and the extinction coefficient reported by Taira, et al. was used to calculate protein concentration.[256] While this extinction coefficient (31,000 $M^{-1}cm^{-1}$) may not accurately represent that of mutant proteins with Tyr mutations, it was found that this error did not introduce significant uncertainty into the concentration estimate. Since the purified DHFR samples contained additional small molecules absorbing at 280 nm, gel filtration of the sample yielded a correction for the optical purity (percentage of the total A280 area under the curve). This correction factor was applied to the A280 concentration estimate. Lastly, DHFR activity was titrated with a known concentration of MTX. The MTX concentration was determined spectrophotometrically using the extinction coefficient at 302 nm (22,100 $M^{-1}cm^{-1}$) in 0.1 N NaOH.[257]

4.3. Competition Experiments

(Performed in collaboration with Jonathan Carlson)

The concentrations of both monovalent and bivalent MTX were assayed spectrophotometrically. The extinction coefficient for bis-MTX was estimated at 47,400 $M^{-1}cm^{-1}$, based on the value reported by Rosowsky and coworkers for a MTX γ-amide.[258] Stock samples of DHFR:bis-MTX were mixed at a stoichiometry of 2:1.05 and incubated in P500buffer containing 5% glycerol (v/v) for a minimum of 3 hours. The five percent excess of dimerizer was added in order to ensure complete initial dimerization, and was shown to perturb the data far less than other sources of experimental error (data not shown). Complete dimerization of this initial stock was verified by gel filtration chromatography as described. The stock sample was then split and a range of MTX equivalents added (0.5 to 2.5x). Samples were incubated at room temperature for a minimum of 3 hours and assayed via gel filtration. The fraction of dimer present was obtained from corrected integration of the absorbance of the trace at 280 nm. The denaturation curve was then fit as described above to yield the K_{eq}/K_c ratio with Mathematica (Wolfram Research) and Microsoft Excel.

Chapter Four

Computational Modeling of the Chemically Induced DHFR Dimer Interface

1. Introduction

While the use of computational techniques to model molecular behavior is not a novel field, the steady increase in computing capabilities over the past 20 years has resulted in the broad-scale application of theoretical chemistry to experimental design. Increased interest in the binding of protein-ligand and protein-protein complexes has driven the development of theoretical methods for the *in silico* prediction of macromolecular behavior in solution. While most available computational techniques require some knowledge of the nature of the binding affinity to properly calibrate parameters for a target of study, more elegant free energy approaches have been developed that allow for the calculation of binding affinities without a priori experimental information.[259-261] Studies of a broad range of problems ranging from biomolecular recognition[262] to computational drug design[263] can all benefit from accurate calculations of the absolute binding free energies of protein-ligand and protein-protein interactions.

One straightforward method for the calculation of absolute interaction free energies is to simply pull one binding partner away from another and calculate the change in free energy. This approach, however, is limited by the inherently inadequate ability of Monte Carlo or molecular dynamics (MD) simulations to sample all of conformational space. Indeed, simulations in a thermal ensemble preferentially sample lower-energy structures. Since higher-energy structures that may contribute a great deal to the free energy of a system are not well-sampled, the resulting free energy will be poorly converged and probably inaccurate. A better method of evaluating absolute free energies is to choose a pathway to remove the binding partner and use specialized algorithms

called rare-event sampling to adequately sample higher energy conformations along the way. Examples of such methods are considered next.

Since free energy is a state function, any pathway can, in principle, be chosen to remove one binding partner from another and calculate the resulting energy differences. Potential of mean force (PMF) approaches such as free energy perturbation (FEP) and umbrella sampling are two methods for achieving such a goal.[136,138,139] In the FEP approach, the ligand is transformed into another species by changing its parameters from those corresponding to the first species to those corresponding to the second. In the case of absolute binding free energy estimation, the ligand is annihilated and the potential of mean force evaluated. Care must be taken in this method to properly treat the changing atomic parameters, and particular issues arise during the modification of charged atoms. Umbrella sampling represents a non-Boltzmann approach to increase conformational sampling efficiency by restraining a reaction coordinate during MD simulation. Umbrella sampling has the advantage of employing an at least feasible physical pathway, rather than the nonphysical FEP technique. Examples of both techniques have been applied to the estimation of binding free energies of protein-ligand[139,264] and protein-protein[265] complexes.

The rigor and accuracy of the PMF techniques comes at the cost of computational resources. In an effort to balance accuracy with efficiency, an approximate endpoint approach has been developed. This so-called Molecular Mechanics – Poisson Boltzmann and Surface Area[135] (MM-PBSA) method is a popular method that relies on a mixed scheme combining conformations sampled from MD simulations in explicit solvent with

solvation free energy estimates from a Poisson-Boltzmann implicit solvent model, solute surface areas, and gas-phase energetic estimates.[266,267] Recently, the MM-GBSA (where GB denotes "generalized Born") method has become more popular, because the GBSA implicit solvent algorithm is faster than the Poisson-Boltzmann treatment of electrostatics.[165] The MM-PBSA technique relies on post-processing of an MD simulation to obtain the desired energetic and entropic contributions. Typically, a single-state approach is used in which only the complex of interest is simulated and the binding partners are extracted for analysis from the resulting trajectory. An increasing number of groups have used this endpoint approach to estimate protein-protein[143,165] and protein-ligand interactions.[268,269] Despite the usefulness of the MM-PB(GB)SA method, however, it can be limited in its accuracy and present difficulties in improving results since it does not offer a rigorous computational route to the desired free energies.[139]

In terms of modeling the behavior of protein complexation, few studies have dealt with more than one binding event at a time. Often, in such cases, constraints are placed on bond length or angles,[270] simulations are restricted to a local region of interest[271] (omitting or replacing the surroundings with an average reaction potential), or symmetry is forced onto the MD simulation of a multimer by calculating forces on a monomer and replicating them in the other monomers.[272] Such approximations often result in nonphysical treatment of the system, neglect of long-range interactions, and underestimates of the entropic contributions. The dihydrofolate reductase/bis-MTX model system for dimerization and examination of mutation effects at the chemically induced protein interface is an optimal candidate for comprehensive computational

exploration, given that there exists a glut of data concerning the theoretical and experimental behavior of DHFR. To our knowledge, no one has yet attempted computational modeling of a tripartite chemically induced protein dimer, and it would be highly beneficial to develop a model system in which to test interfacial and ligand modifications *in silico*.

Surveying the methods described above, it is apparent that the FEP method of estimating absolute binding energy will be problematic for our particular needs, since removal of the entire protein partner would prove far too complicated. Thus, we have elected to utilize the two remaining methods – umbrella sampling and MM-GBSA – to estimate the binding free energies associated with DHFR binding to MTX_2C9 and DHFR binding to the intermediate DHFR-MTX_2C9 complex. Given accurate calculation of these binding energies, estimation of the effects of mutations at the DHFR dimer interface would be possible and represent an extremely useful tool for future protein engineering in our laboratory.

2. Methods and Software

2.1. Theory

The formation of the chemically induced DHFR complex can be described using the equilibrium expression found in Equation 1,

$$2E + L \xleftrightarrow{K_{a1}} EL + E \xleftrightarrow{K_{a2}} E_2L \qquad (1)$$

where E is the DHFR enzyme and L is the bis-MTX ligand. Using the expression for free energy and the assumptions that (i) K_{a1} is equivalent to the association constant for MTX

binding to DHFR and (ii) $K_{a2} = K_{a1} * K_c$, we can derive an equation for the calculation of K_c, the cooperativity constant between the interacting proteins, and its corresponding free energy (Equations 2-5).

$$\Delta G_M = -RT \ln K_{a1} \quad (2)$$

$$\Delta G_D = -RT \ln K_{a2} \quad (3)$$

$$K_c = \frac{e^{\frac{\Delta G_D}{-RT}}}{e^{\frac{\Delta G_M}{-RT}}} \quad (4)$$

$$\Delta G_c = -RT \ln K_c \quad (5)$$

The difference between the mutant and wild-type ΔG_c values results in $\Delta\Delta G$, a measure of the changes in cooperativity associated with interfacial mutations, which is then correlated to the experimental results presented in Chapter 3.

The theoretical derivation of expressions used for calculating absolute binding energies of protein-ligand complexes has been thoroughly discussed elsewhere[261,273], and will not be recapitulated in detail here, but rather the main points will be summarized. For umbrella sampling, the PMF – called *W(r)* – as a function of the reaction coordinate, taken here as the distance *r* from the protein to the ligand (defined more precisely below), can be obtained from the probability, $\rho(r)$, of finding a ligand a given value of *r*.[268]

$$W(r) = -k_B T \ln[\rho(r)] \quad (6)$$

The absolute binding free energy is the difference between *W(r)* for *r* corresponding to bound complex and *r* corresponding to unbound ligand and protein. Given an arbitrary

path along which the ligand is pulled out of the binding site, this reaction coordinate can be expressed as

$$r_j = r_o + j\Delta r \qquad (7)$$

where r_o is the initial bound position, Δr is a step size, and j is a step number; the steps being called windows. The PMF is calculated along the reaction coordinate via umbrella sampling by the use of a biasing potential of the form

$$V_j(r) = \frac{1}{2} k_j \left(r - r_j\right)^2 \qquad (8)$$

which restrains the ligand at window j to have r close to r_j using force constant k_j. The distributions in three-dimensional space are then merged using the weighted histogram analysis method (WHAM) to form an unbiased distribution, $\rho(r)$, and the PMF curve is calculated by Equation 6.

For MM-GBSA calculations, the formula for binding free energy involves the thermodynamic cycle of solvation and binding as shown in Figure 18 (see Introduction) and can be expressed as

$$\Delta G_{\text{bind(aq)}} = \Delta G_{\text{bind(g)}} + \Delta G_{\text{sol(complex)}} - \left(\Delta G_{\text{sol(lig)}} + \Delta G_{\text{sol(rec)}}\right) \quad (9)$$

In the MM-GBSA model, solvation energies are calculated for each state using a generalized Born solvation model for electrostatics and a surface area calculation and empirical term for nonpolar contributions:

$$\Delta G_{\text{sol}} = \Delta G_{\text{GB}} + \Delta G_{\text{np}} \qquad (10)$$

where

$$\Delta G_{\text{np}} = \gamma \text{SASA} + b \qquad (11)$$

in which SASA is the solvent-accessible surface area, γ is a parameterized value for surface tension, and b is a parameterized offset value (typically zero). The gas phase binding energy is obtained by calculating the differences in internal, electrostatic, and van der Waals terms for each structure in the gas phase via

$$\Delta G_{\text{bind(g)}} = \Delta G_{\text{gas(complex)}} - \left(\Delta G_{\text{gas(lig)}} + \Delta G_{\text{gas(rec)}}\right) \qquad (12)$$

$$\Delta G_{\text{gas}} = \Delta G_{\text{internal}} + \Delta G_{\text{electrostatic}} + \Delta G_{\text{vdW}} - T\Delta S \qquad (13)$$

In a typical one-state MM-GBSA approach, a single simulation of the complex is performed and the structures corresponding to the ligand and receptor are extracted from the coordinates of the complex in the trajectory (this is called the single-trajectory method later in the chapter). Normal mode or harmonic analyses may be performed to yield entropic contributions, however, the high computational cost of these calculations deters them from being used, as the (possibly severe) approximation is generally made that entropy contributions will cancel out in the course of analyzing similar macromolecular systems. In a two-state MM-GBSA approach, all parts of the complex are simulated; however, their energies are calculated in the same manner (this is called the full-trajectory method later in the chapter).

2.2. Platforms and Software

Umbrella sampling simulations were performed on an IBM BladeCenter Linux cluster (IBM LS21 compute nodes) at the University of Minnesota Supercomputing Institute. Simulations and calculations for MM-GBSA were performed on an HP DL185 cluster under the Red Hat Linux operating system (Chinook) at the Environmental

Molecular Sciences Laboratory at Pacific Northwest National Laboratory. All simulations were performed using the AMBER10 suite of programs.[274] All proton addition, protein solvation/neutralization, and parameter generation was performed using AmberTools v1.2. Automated MM-GBSA calculations were performed using the mm_pbsa Perl script contained in the AMBER9 distribution. Structure preparation involving bond building and charge modification was performed using Maestro (Schrodinger, Inc.).

2.3. Umbrella Sampling

Structures were taken from the B chain of the x-ray crystal structure of *E. coli* DHFR (PDB ID: 4DFR) and modified using Maestro (Schrodinger, Inc.). Crystallographic water molecules were deleted except for those within 10 Å of the protein surface. Protons were placed using xLEaP (AMBER10) and optimized via conjugate gradient minimization (rms gradient <0.0001 kcal*mol^{-1}Å^{-1}). Force field parameters for MTX were generated using Antechamber (AMBER10). The protein was solvated using an octahedral TIP3P water box extending 15 Å from the surface of the protein, and the system was neutralized by adding 10 sodium counterions. To remove initial bad contacts, 300 steps of conjugate gradient minimization were performed. The protein was then held fixed and the water molecules heated from 10 to 310 K over 50 ps, followed by 50 ps of constant volume equilibration and 300 ps of constant pressure equilibration. The restraints on the protein were then removed and the entire system heated from 10 to 310 K over 50 ps, followed by 50 ps of constant volume equilibration and 500 ps of constant pressure equilibration. 2 ns of NPT production-phase MD at 310

K were collected in 200 ps increments, with coordinates saved every 0.5 ps. All simulations were performed with a non-bonded cutoff of 10 Å.

The final structure from the unrestrained simulation was used as a starting point for the umbrella sampling. The reaction coordinate *r* is defined as the distance from the proton on N1 (see Chapter 2, Figure 2) to the carboxyl carbon of Asp27. To reduce the sizes of the perturbation of the system, windows were generated serially. Distance restraints for the beginning of sampling were determined via measurement of the distance in the final snapshot of the unrestrained structure. In particular, the final structure from the non-restrained simulation was restrained to a MTX N1 proton – Asp27 carboxyl carbon distance of 2.8 Å and equilibrated for 100 ps with constant pressure. The structure resulting from this calculation was then used as the starting point for a simulation restrained at 2.95 Å and that final structure used for a simulation restrained at 3.1 Å and so on in both directions (Δr = 0.15 Å). After equilibration, data was typically collected for 1 ns at 310 K and 1 atm. The force constant k_j for the restraint was 50 kcal/mol*Å^2 applied parabolically about the chosen distance. Nonbonded interactions were first cut off at 10 Å, and then 14 Å; then a full Ewald treatment of the nonbonded interaction was performed. Simulations were run in a truncated octahedral water bath containing TIP3P water extending 15 Å from the surface of the protein. Measurements of the restrained distance at were written to an output file every 20 ps during the simulation, and a custom FORTRAN script was used to generate a plot of distance probabilities (see Appendix Three). The Weighted Histogram Analysis Method (WHAM) was used to merge the probabilities from the simulation windows and calculate the PMF.[275]

2.4. MM-GBSA

Structure preparation

Structures were taken from the x-ray crystal structure of *E. coli* DHFR (PDB ID: 4DFR) and modified using Maestro (Schrodinger, Inc.). Monomeric DHFR bound to MTX was generated using the A chain from the PDB file. Unbound DHFR was generated from the B chain of the PDB file. Mutant species were generated using the mutate residue function in Maestro. Parameters for C9 were generated using Antechamber as implemented in AMBER10. C9 was built for dimeric systems by building in the methylene linker region and adjusting the formal charge on N1 of the molecule from 0 to +1to account for protonation when bound to the protein. C9 for the monomeric systems was built in the same manner. Systems were solvated with TIP3P water in a truncated octahedron extending 12 Å from the surface of the protein and neutralized via the addition of an appropriate number of sodium counterions.

Simulation protocol

Initially, bad contacts in the structure were removed via 2000 steps of steepest descents minimization followed by 2000 steps of conjugate gradient minimization. The system was then heated from 10 to 300 K over 10 ps with a protein backbone restraint of 1.0 kcal/mol. The density was then equilibrated over 20 ps with the same restraint, followed by a three step pressure equilibration consisting of 20 ps with a 1.0 kcal/mol restraint on the protein backbone, 50 ps with a 0.5 kcal/mol restraint, and 1 ns of unrestrained equilibration. SHAKE was used to constrain bonds between hydrogen and nonhydrogen atoms, and long-range electrostatic interactions were cut off at 14 Å. Data

was collected at 300 K and 1 atm for 4 ns, with snapshots of the trajectory saved every 10 ps.

Free Energy Calculations

For free energy calculations, the MM-GBSA method was used as implemented in AMBER9. We use the MM-GBSA method to calculate ΔG for

$$L + R \xrightarrow{\Delta G_{bind}} C$$

Where R, L, and C denote receptor, ligand, and complex. In the "monomer" calculations, R is E, L is C_9, and ΔG is ΔG_{a1}, whereas in the "dimer" calculations, R is E-C_9, L is E, and ΔG is ΔG_{a2}. For each choice of E, two kinds of simulations can be run: single-trajectory or full-trajectory simulations. In the former, one runs dynamics only for the solvated C system, but in the latter, one also runs dynamics for the solvated R and L systems (see Section 2.1). First, snapshots of C, R, and L were generated from the dimer complex (E_2L) and the monomer complex (EL) for the single-trajectory analysis or all simulations for the full-trajectory analysis and stripped of water and counterions. The gas-phase energy and implicit solvation free energy are then calculated by programs module *Sander* for each structure and averaged over all the snapshots. The three GB methods available[276-278] were used to calculate solvation free energies for each structure in the single-trajectory analysis, and Onufriev's GB model[279] was used for the full-trajectory analysis (see Results for discussion). The ionic strength used for the GB calculation is 0.5 M, reflecting the NaCl concentration of P500 buffer typically used for dimerization experiments (see Chapter 3). The dielectric constant of water was set to 80, and the dielectric constant for the interior of the protein was set to 1.0. The solvent

accessible surface area was calculated using the LCPO method.[280] The surface tension (γ) used to calculate the nonpolar contribution to the desolvation free energy was 0.0072 kcal*A^{-2}.

3. Results and Discussion

3.1. Umbrella Sampling

In order to validate our hypothesis that umbrella sampling could be used to calculate the binding free energy associated with protein interface cooperativity, we first tested our methods on a smaller system, namely the DHFR monomer bound to monovalent MTX. It stands to reason that if accurate calculations and an appropriate reaction coordinate could be obtained using a model system, scale-up of our system to contain two monomeric species should be feasible. Beginning with an equilibrated structure of DHFR bound to MTX, the distance between the proton on N1 of MTX and the Asp27 sidechain carboxyl carbon was selected as the reaction coordinate. By sequentially increasing this distance, we reasoned that this would be similar to MTX undergoing dissociation from the DHFR binding pocket. Initial restrained simulations were carried out and the results are shown in Figure 1.

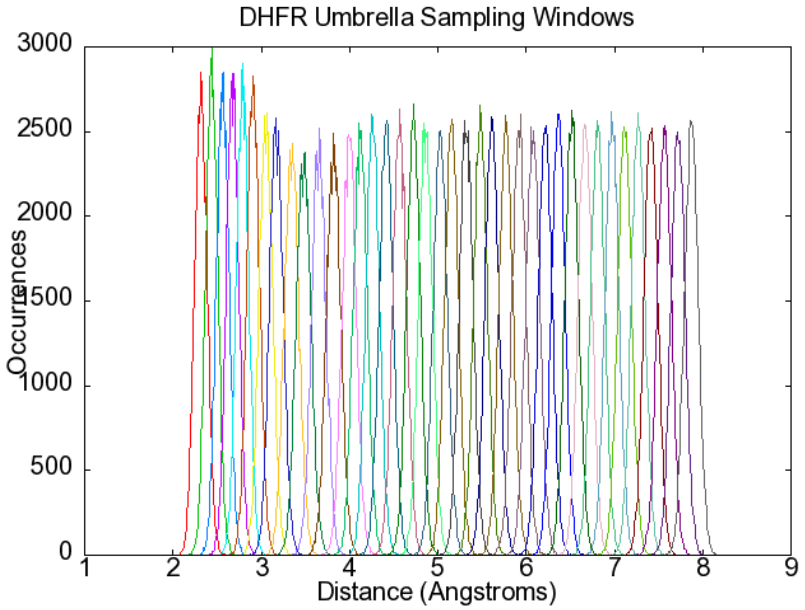
To assure adequate sampling of phase space, it is crucial that care is taken in selecting both the interval between window distances r_j along the reaction coordinate and the force constant chosen to enforce the resulting high-energy conformation. Weak force constants will broaden the window, leading to poor sampling of high-energy structures and an imprecise calculation of the PMF. Large force constants will cause peak width to

narrow significantly, requiring an excessive number of simulations. In the case of exceedingly large force constants, the simulation may fail due to the magnitude of the forces applied to maintain the restrained distance. Examination of the sampling windows in Figure 1A reveals that each window chosen overlaps several adjacent windows, implying thorough sampling of phase space. The peak width observed in the probability histogram supports our choice of a 50 kcal/mol*Å^2 force constant, though this value could likely be reduced and the distance between simulations increased without sampling penalties.

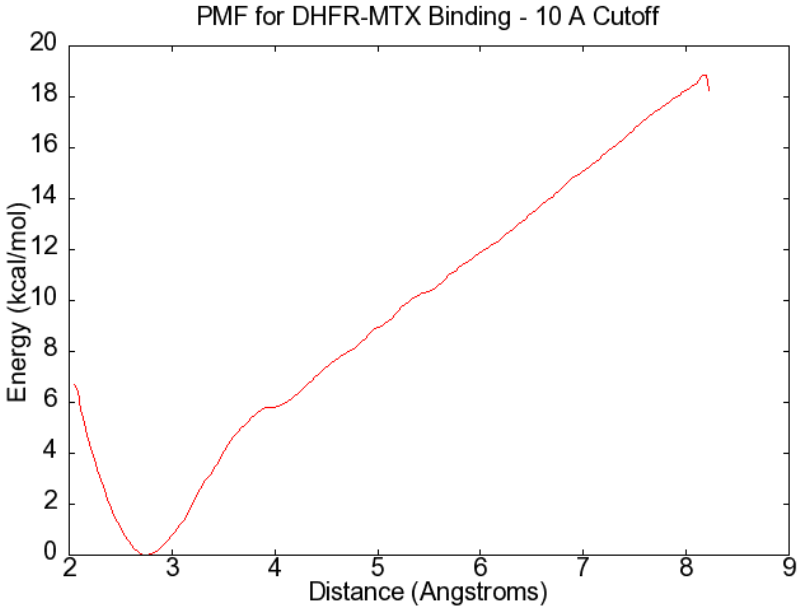
Inspection of the PMF curve generated from our initial umbrella sampling simulations (Figure 1B) yields disappointing results. While the energy minimum centers around the equilibrium distance we measured from the unrestrained simulation as expected, as MTX is drawn out of the binding pocket, the PMF fails to converge to a stable energy. To explain this error, we hypothesized that the 10 Å nonbonded truncation used in the simulations was too short. During MD simulations, calculation of long-range interactions dominates the computing time necessary for each time step. Truncation of these interactions by a number of methods is common practice to reduce the computing cost of a simulation.[281,282] However, studies by Schreiber and Steinhauser have shown that cutoff distance affects the behavior of solvated polypeptides, to the extent that the helical nature of their model system was strongly dependent upon the cutoff distance chosen.[283]

Figure 1. Results of umbrella sampling utilizing a 10 Å nonbonded cutoff. A) Sampling windows each represent 1 ns production phase MD with the reaction coordinate restrained about the Asp27-H(N1) distance corresponding approximately to the peak maximum. B) PMF resulting from this series of simulations.

A.



B.



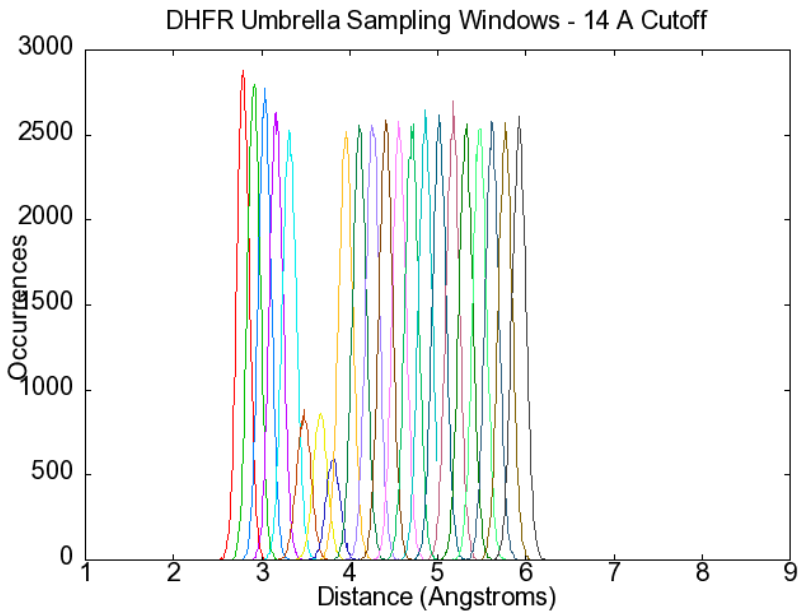
Truncation of nonbonded interactions results in discontinuity in the potential energy function. While this discontinuity has no impact on the evolution of an MD trajectory, it has been shown to have effects on the relative free energies of differing conformations.[284] Therefore, the simulations were repeated using a larger nonbonded cutoff of 14 Å.

The results of our simulations with the longer-range nonbonded cutoff are presented in Figure 2. Fewer simulations were performed in order to test our hypothesis. Analysis of the sampling windows (Figure 2A) shows similar results to the 10 Å cutoff, indicated good sampling and force constant selection. Examination of the PMF, however, yields results similar to the previous simulations. Despite the increased computing cost, we elected to pursue MD simulations with a full Ewald treatment of the electrostatics, which would circumvent the poor convergence properties of Coulombic interactions by splitting them into two rapidly converging series in real and reciprocal space.[285] This treatment of long-range interactions avoids the problem of choosing a suitable nonbonded cutoff.

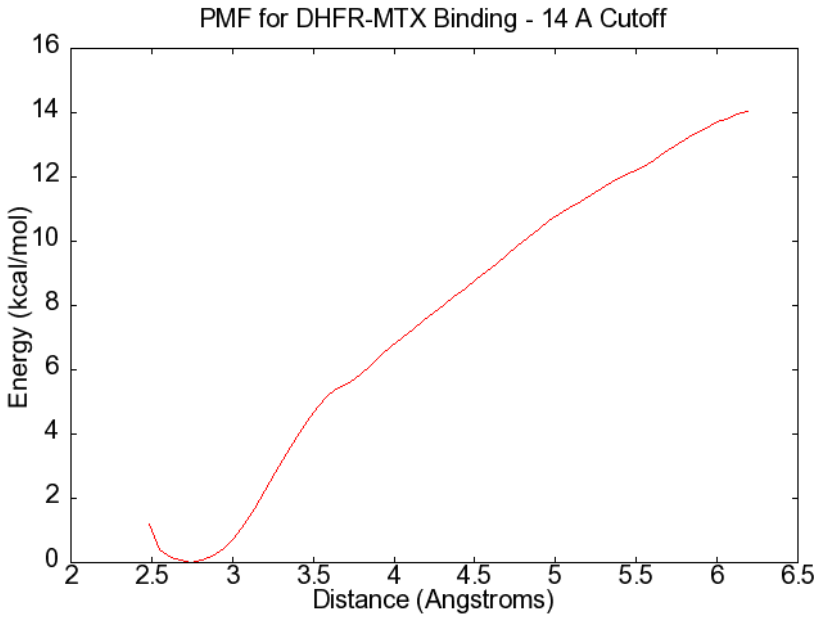
Simulation time was shortened to 250 ps and the distance between reaction coordinate points (Δr) was increased to 0.25 Å in our Ewald-treated systems in an effort to test the validity of our hypothesis as well as conserve computing time. Additionally, examination of the structure generated during the 7.9 Å restrained simulation (14 Å cutoff) showed that although the pteridine moiety of MTX was being forced out of the

Figure 2. Results of umbrella sampling utilizing a 14 Å nonbonded cutoff. A) Sampling windows each representing 1 ns production phase MD with the reaction coordinate restrained about the Asp27-H(N1) distance approximately corresponding to the peak maximum. B) PMF resulting from this series of simulations.

A.



B.



binding pocket of DHFR, the γ-carboxylate tail still appeared to interact with the protein (Figure 3). To explore whether we had chosen a sub-optimal reaction coordinate, we performed an additional series of simulations restraining the distance between the backbone oxygen of Leu94 and the centrally-located N10 (PABA amine nitrogen) on MTX. Results from these simulations are shown in Figures 4 and 5. The sampling windows in both sets of simulations display adequate overlap (Figs. 4a and 5a), however, as before, the PMF curves fail to converge to a stable energy.

While these discouraging results could be due to surmountable errors in the treatment of long-range interactions or force field parameterization issues, it became evident upon inspection of the resulting structures (Figures 3 and 6) that full extradition of MTX from the binding pocket of DHFR would require an inordinate number of serial simulations. The flexibility and length (~15 Å when extended) of MTX, especially when additional ligand-protein interactions (specifically with Arg57 and His28)[286] that tether MTX to the protein are taken into account, appear to preclude the estimation of the PMF of MTX binding by umbrella sampling. It is possible that enforcing rigidity upon the protein and ligand, as performed by Lee and Olson[268] when using umbrella sampling to estimate FK506 binding to FKBP, may assist in the expedient removal of the ligand from the protein, but it is unclear how this approach could be applied to the larger tripartite system without making extremely severe approximations about dimer behavior. Given these considerations and their implications for calculating the PMF of protein-protein binding we elected to abandon this approach and focus on methods more suited to estimating the binding free energy associated with protein complexation.

Figure 3. Extradition of MTX from the binding pocket of DHFR. A, B, and C all show rotated views of the same simulation (7.9 Å restraint). Arrows point to the carboxylate carbon of Asp27 and H(N1) of MTX. The interaction of the γ-carboxylate tail of MTX with the protein is circled.

A.

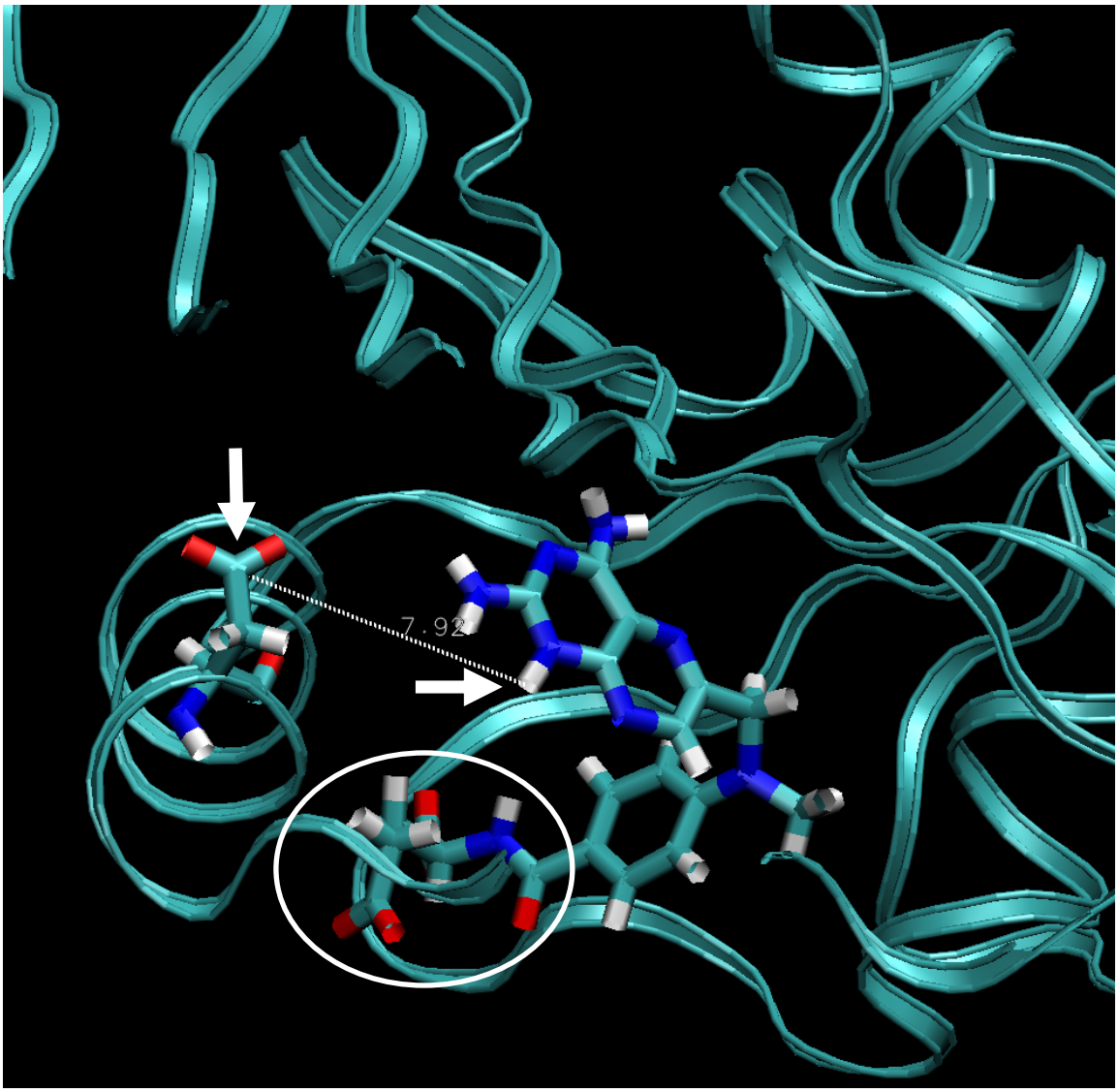


Figure 3, continued.

B.

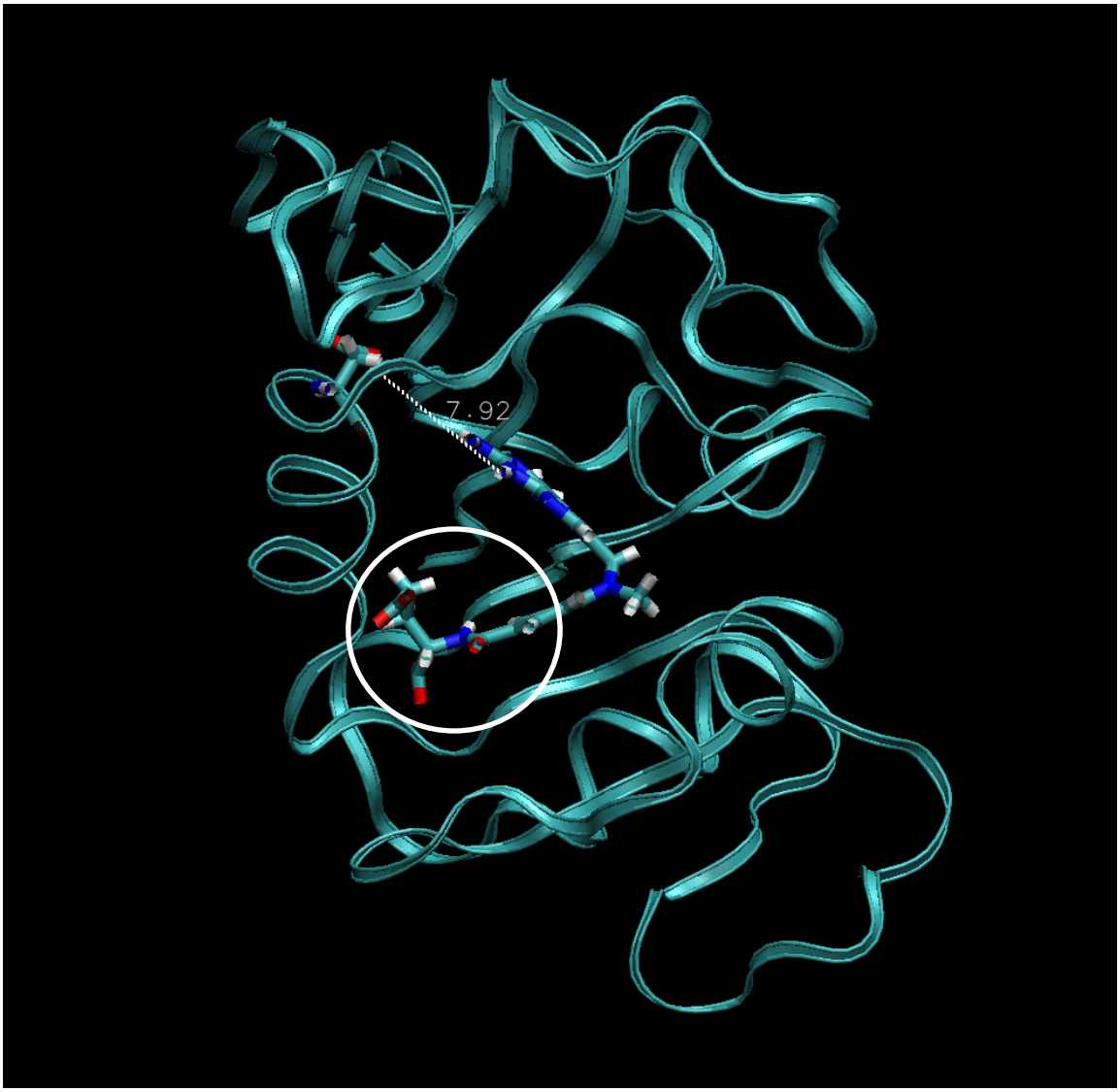


Figure 3, continued.

C.

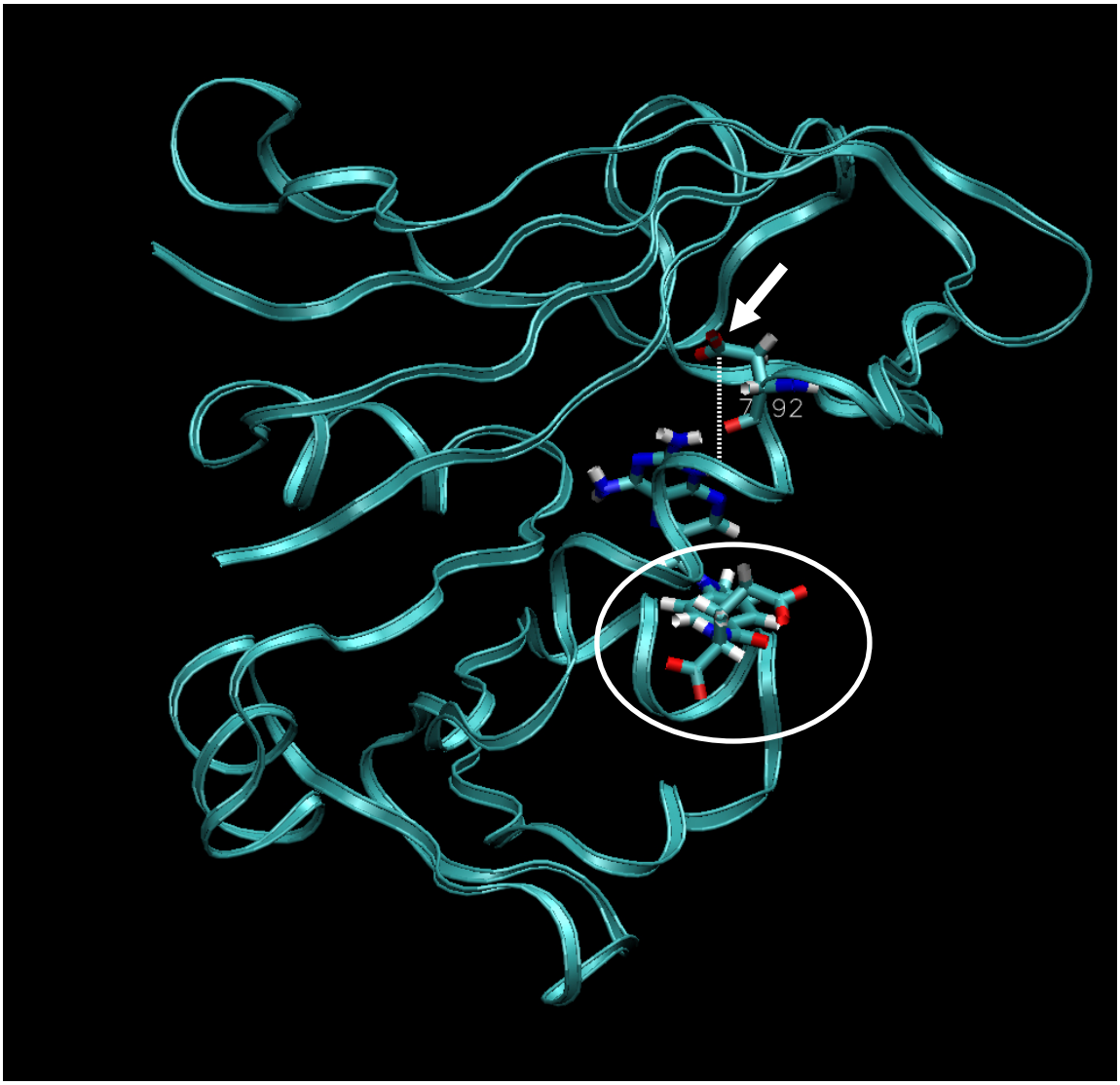
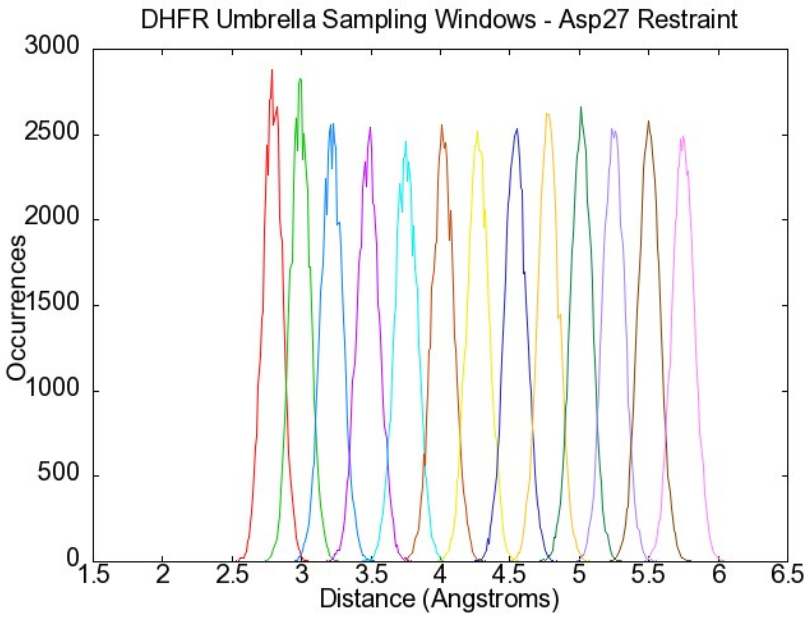


Figure 4. Results of umbrella sampling utilizing full Ewald treatment of long-range interactions in the Asp27-H(N1) restrained system. A) Sampling windows each representing 1 ns production phase MD with the reaction coordinate restrained about the Asp27-H(N1) distance approximately corresponding to the peak maximum. B) PMF resulting from this series of simulations.

A.



B.

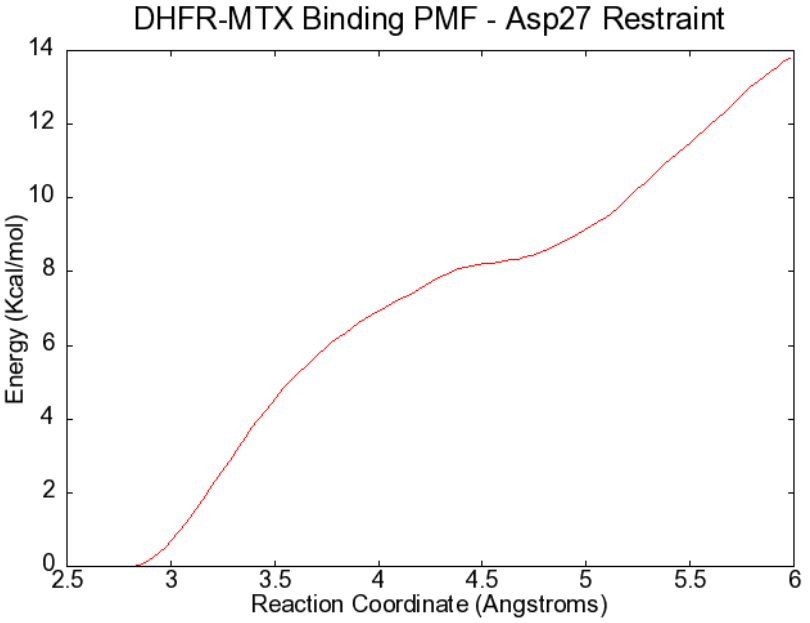
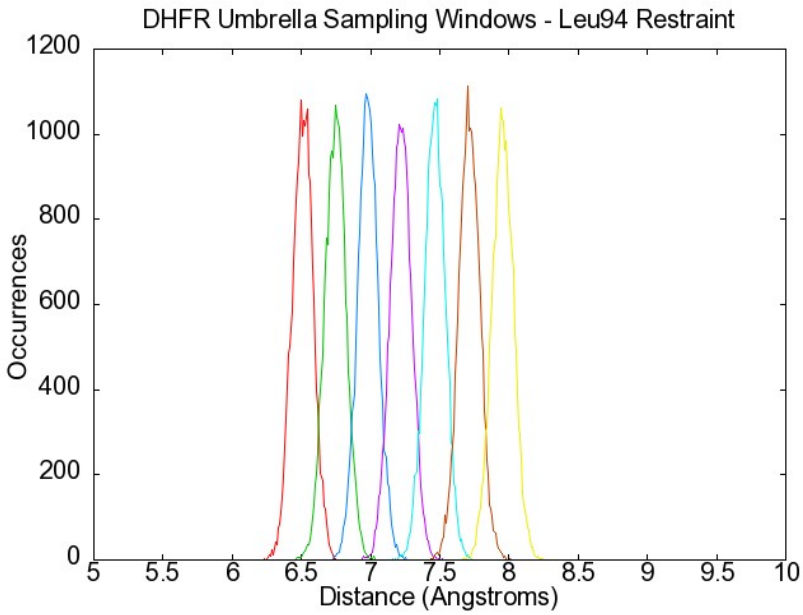


Figure 5. Results of umbrella sampling utilizing full Ewald treatment of long-range interactions in the Leu94-N10 restrained system. A) Sampling windows each representing 1 ns production phase MD with the reaction coordinate restrained about the Leu94-N10 distance approximately corresponding to the peak maximum. B) PMF resulting from this series of simulations.

A.



B.

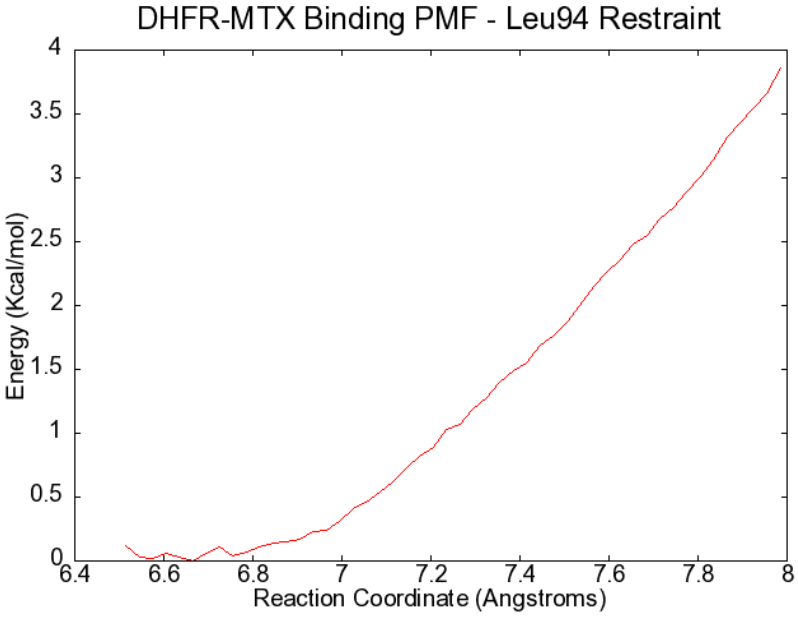


Figure 6. Progression of MTX out of the binding pocket of DHFR. A) 6.5 Å restraint. B) 8.75 Å restraint.

A.

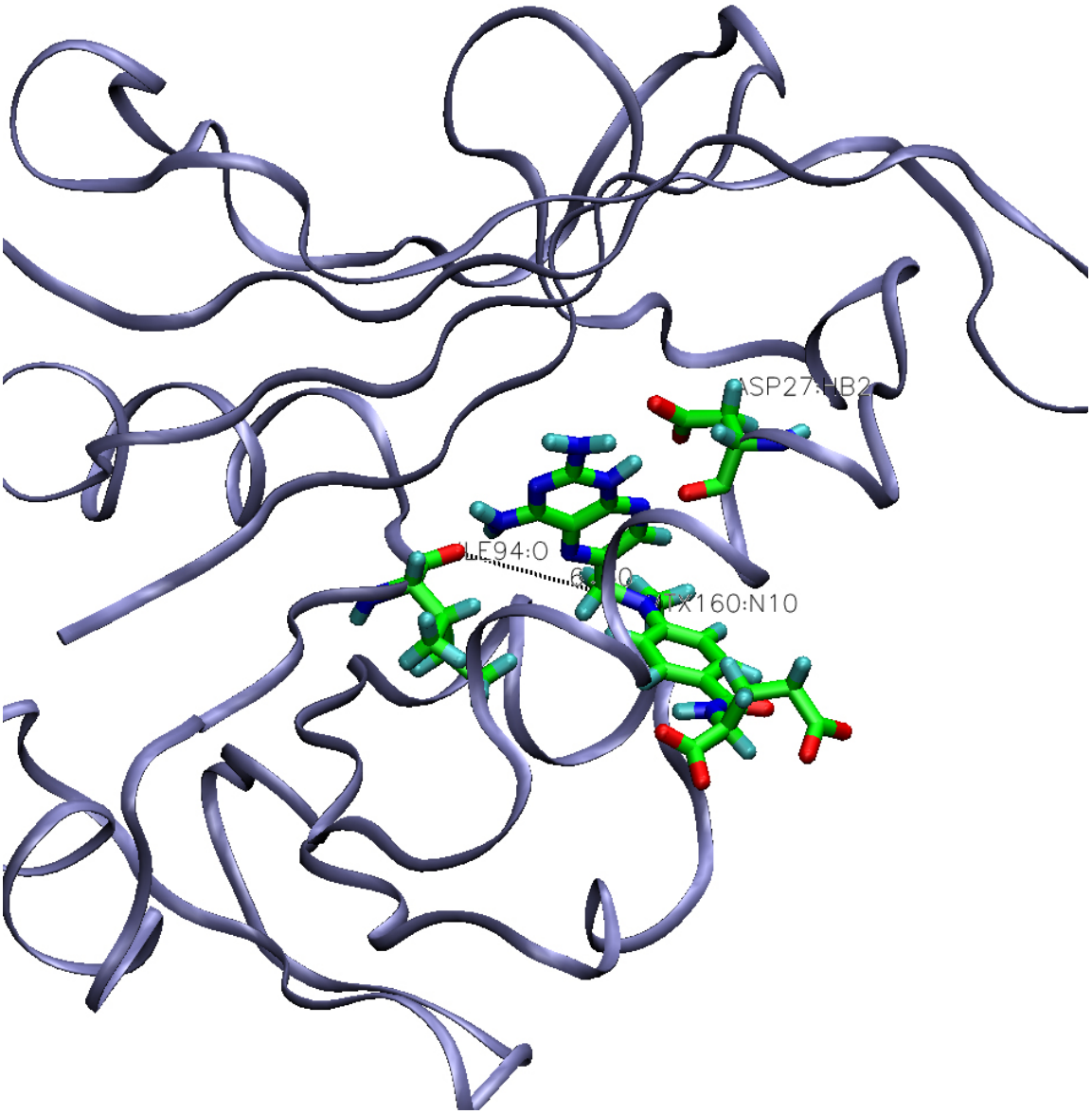
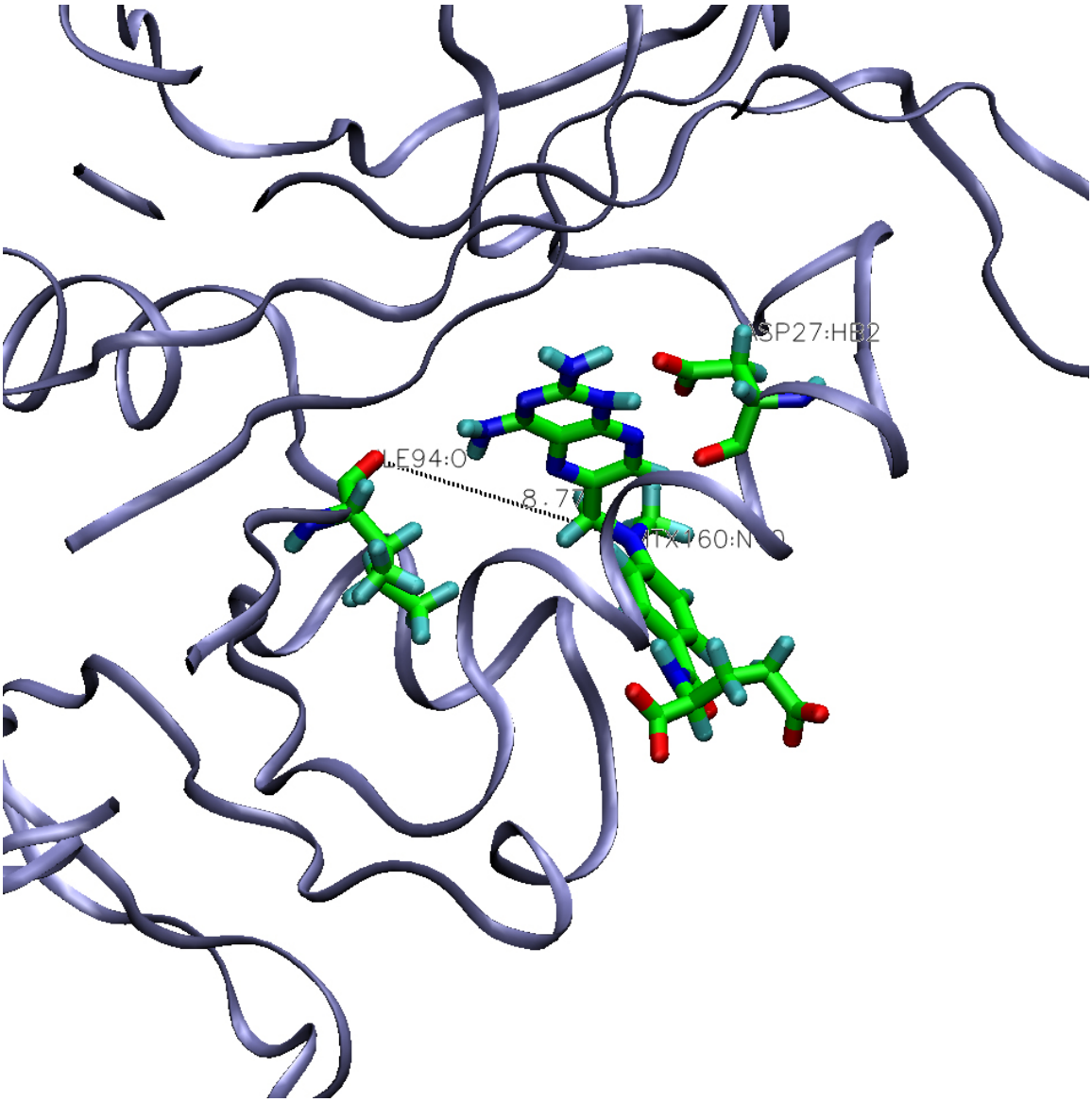


Figure 6, continued.

B.



3.2. MM-GBSA

In light of the problems faced using umbrella sampling to evaluate the binding free energy between two proteins, we elected to switch to the MM-GBSA method, which has come into heavy use recently as a technique for estimating binding free energies for protein-ligand and protein-protein complexes. Our initial attempts to estimate DHFR dimer cooperativity were focused on a single-trajectory approach, in which only the two structures corresponding to the *complexes* formed during dimerization (Equation 21, EL and E_2L) are simulated. Structures corresponding to the ligand and receptor in each case (E and L or EL and E) are individually extracted from the atomic coordinates recorded in each snapshot during the simulation and are processed along with the entire complex.

A source of error commonly present in MD simulation is the correlation of structures over the course of the simulation due to the nonrandom movement inherent to MD. Structural correlation leads to an underestimate of the error in the calculation. Although this problem is difficult to alleviate completely, analysis of structures far enough away from each other temporally can lessen the impact of correlation. To obtain a reasonable estimate of error, we split the trajectory into sets of 100 or 40 structures, as seen in Table 1. For a truly uncorrelated simulation, the standard error of the mean (SEM) over all 400 structures should equal that of the subdivided trajectory. As can be observed in Table 1, this is not the case, and there is a degree of correlation present in our simulation. Division of the trajectory into four sets of 100 structures yields a SEM of 0.98 kcal/mol, and division into 10 sets of 40 structures yields a SEM of 0.67 kcal/mol. While it is attractive to use the larger error estimate as a measure of statistical error, this SEM

effectively represents only four sets of structures, and as such is not a comprehensive picture of the overall simulation. Therefore, we decided to use the error in 10 sets of 40 structures, as this represents a good balance between correlation and accurate error estimate. In future simulations, the impact of correlation could be reduced by increasing the simulation time and widening the time step between recorded structures.

To observe the effects of the generalized Born model selected for the calculation of solvation energies, we tested all three models available in the mmpbsa module in the AMBER9 package. Results from these three separate calculations and their comparison to experimental $\Delta\Delta G$ values are shown in Tables 2-4 and Figure 7. The Cramer and Truhlar GB model, modified by Tsui and Case (IGB=1),[276,277] as well as the Onufriev GB model (IGB=2)[279] are the best performing models, each displaying similar correlations to the experimental data. The modified Onufriev model (IGB=5)[287] performs very poorly for our system, despite earlier work that indicated this model agrees well with the Poisson-Boltzmann treatment of the electrostatic component of the solvation energy. It can be argued that this enhancement in electrostatic treatment does not represent an advantage when evaluating the primarily hydrophobic nature of the protein interface.

Several areas of concern arise upon inspection of the data. First, the errors in the calculation of binding free energy, though typically around 1–2 kcal/mol, are too great to yield statistically meaningful results when compared to the experimental data range (~1.5 kcal/mol). In order to reduce the error to a value (~0.1 kcal/mol) that would allow meaningful comparison to experiment, simulations of ~100 ns would be required, as a ten-fold reduction in error requires a 100-fold increase in simulation time.

Table 1. Representative error and correlation analysis in MM-GBSA simulations. All values are in kcal/mol.[a]

	WT - Monomer			WT - Dimer		
Snapshots	**Mean**	**Std. Dev.**	**Std. Err.**	**Mean**	**Std. Dev.**	**Std. Err.**
1-400	-69.52	4.12	0.21	-128.64	5.63	0.28
1-100	-66.91			-130.08		
101-200	-70.24			-130.74		
201-300	-69.38			-124.94		
301-400	-71.57	1.96	0.98	-128.79	2.59	1.30
1--40	-66.22			-130.67		
41-80	-67.00			-128.41		
81-120	-68.46			-131.53		
121-160	-70.64			-132.63		
161-200	-70.57			-128.83		
201-240	-67.93			-125.64		
241-280	-70.25			-124.16		
281-320	-71.40			-125.95		
321-360	-73.29			-127.28		
361-400	-69.50	2.15	0.67	-131.27	2.86	0.90

[a]The quantity tabulated as the mean is ΔG_M. The next column gives its standard deviation σ. The next column is standard error of the mean (SEM) given by SEM = $\sigma/n^{½}$ where n is the number of subdivisions used to calculate the standard deviation.

Table 2. MM-GBSA data collected using IGB=1 (Tsui's GB model).[277] All values are in kcal/mol. ΔG_M and ΔG_D are obtained from simulations of the DHFR-C9 and the $DHFR_2$-C9 complexes, respectively. Values for $\Delta\Delta G_{exp}$ are presented in Chapter 3.

Protein	ΔG_M	**Error**	ΔG_D	**Error**	ΔG_C	**Error**	$\Delta\Delta G_{cak}$	$\Delta\Delta G_{exp}$
WT	-69.52	0.67	-128.64	0.90	-59.12	1.12	0	N/A
A19E	-70.05	1.36	-121.80	0.73	-51.75	1.54	7.37	0.47
A19F	-76.97	0.71	-138.85	0.90	-61.88	1.15	-2.76	0.26
A19H	-86.68	1.52	-136.54	1.63	-49.86	2.23	9.26	0.44
A19K	-82.24	1.44	-130.43	0.42	-48.19	1.50	10.93	0.74
A19L	-70.06	1.85	-140.72	0.71	-70.66	1.98	-11.54	0.34
A19Q	-65.38	1.19	-141.32	0.96	-75.94	1.53	-16.82	0.35
A19S	-78.45	0.80	-122.48	1.15	-44.03	1.40	15.09	0.28
A19Y	-63.26	1.11	-132.91	1.76	-69.65	2.07	-10.53	0.18
N23E	-72.14	2.76	-130.94	1.02	-58.80	2.94	0.32	0.83
N23F	-68.79	0.87	-135.07	1.36	-66.28	1.61	-7.16	-0.62
N23H	-72.61	2.54	-129.99	0.58	-57.38	2.60	1.74	0.42
N23K	-73.63	1.97	-124.97	1.36	-51.34	2.40	7.78	0.88
N23L	-75.95	1.49	-137.52	0.65	-61.57	1.62	-2.45	-0.33
N23Q	-74.13	0.90	-139.72	0.76	-65.59	1.18	-6.47	-0.58
N23S	-70.47	1.96	-127.37	0.96	-56.90	2.18	2.22	-0.05
N23Y	-71.03	1.18	-129.40	0.70	-58.37	1.37	0.75	-0.35

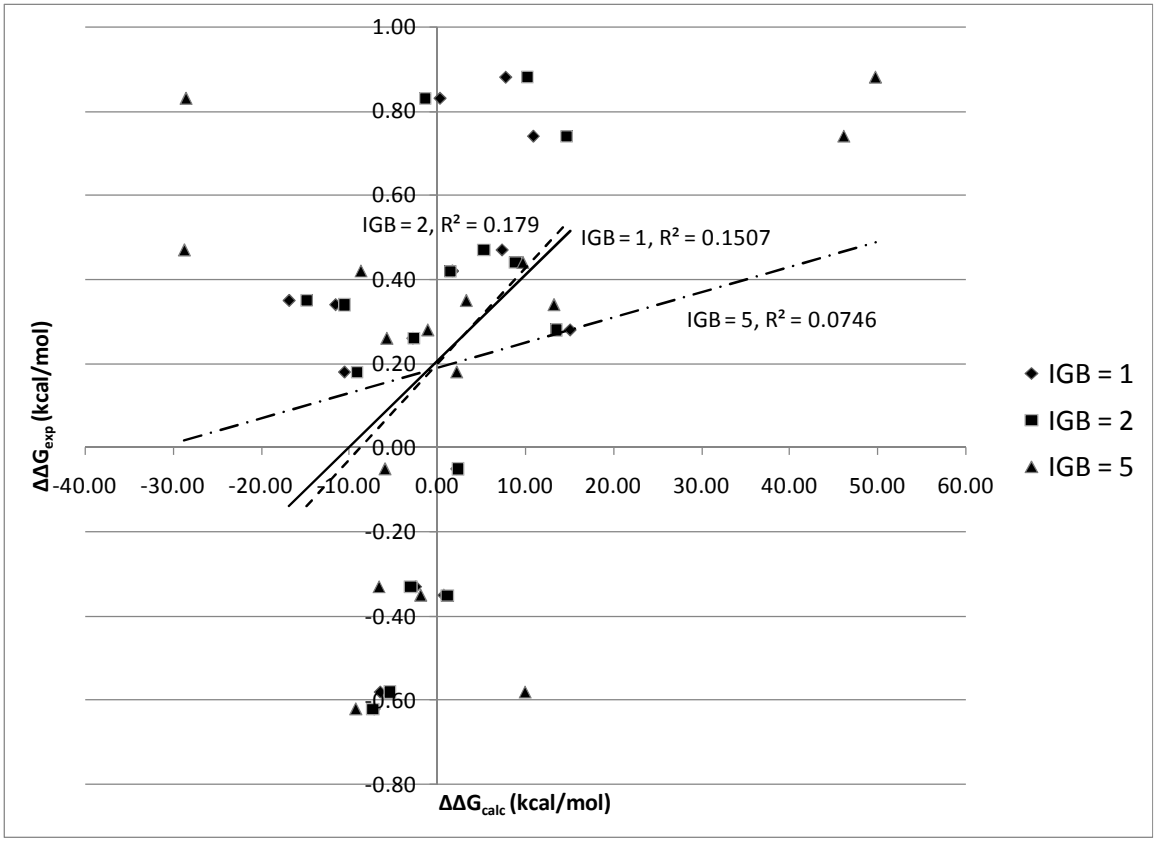
Table 3. MM-GBSA data collected using IGB=2 (Onufriev's GB model). All values are in kcal/mol. ΔG_M and ΔG_D are obtained from simulations of the DHFR-C9 and the $DHFR_2$-C9 complexes, respectively. Values for $\Delta\Delta G_{exp}$ are presented in Chapter 3.

Protein	ΔG_M	Error	ΔG_D	Error	ΔG_C	Error	$\Delta\Delta G_{cak}$	$\Delta\Delta G_{exp}$
WT	-63.63	0.54	-117.95	0.81	-54.32	0.98	0	N/A
A19E	-63.23	1.17	-112.20	0.60	-48.97	1.31	5.35	0.47
A19F	-70.46	0.71	-127.44	0.82	-56.98	1.08	-2.66	0.26
A19H	-79.62	1.90	-125.02	1.50	-45.40	2.42	8.92	0.44
A19K	-76.00	1.39	-115.62	0.39	-39.62	1.45	14.70	0.74
A19L	-63.67	1.67	-128.50	0.69	-64.83	1.81	-10.51	0.34
A19Q	-59.72	1.11	-128.82	0.95	-69.10	1.46	-14.78	0.35
A19S	-70.93	0.78	-111.68	1.04	-40.75	1.30	13.57	0.28
A19Y	-57.67	0.99	-121.08	1.64	-63.41	1.92	-9.09	0.18
N23E	-64.87	2.41	-120.49	0.97	-55.62	2.59	-1.30	0.83
N23F	-62.83	0.83	-124.41	1.24	-61.58	1.49	-7.26	-0.62
N23H	-66.42	2.30	-119.16	0.56	-52.74	2.36	1.58	0.42
N23K	-67.16	1.76	-111.19	1.32	-44.03	2.20	10.29	0.88
N23L	-69.02	1.29	-126.33	0.60	-57.31	1.42	-2.99	-0.33
N23Q	-68.10	0.81	-127.73	0.68	-59.63	1.06	-5.31	-0.58
N23S	-64.18	1.79	-116.14	0.86	-51.96	1.99	2.36	-0.05
N23Y	-65.43	1.10	-118.58	0.69	-53.15	1.30	1.17	-0.35

Table 4. MM-GBSA data collected using IGB=5 (Modified Onufriev GB model)[287]. All values are in kcal/mol. ΔG_M and ΔG_D are obtained from simulations of the DHFR-C9 and the $DHFR_2$-C9 complexes, respectively. Values for $\Delta\Delta G_{exp}$ are presented in Chapter 3.

Protein	**ΔG_M**	**Error**	**ΔG_D**	**Error**	**ΔG_C**	**Error**	**$\Delta\Delta G_{cak}$**	**$\Delta\Delta G_{exp}$**
WT	-79.89	1.13	34.84	1.70	114.73	2.03	0	N/A
A19E	-72.81	1.13	13.22	1.26	86.03	1.70	-28.70	0.47
A19F	-82.73	0.96	26.30	1.65	109.03	1.91	-5.70	0.26
A19H	-88.97	2.02	35.52	1.84	124.49	2.73	9.76	0.44
A19K	-94.54	1.84	66.41	1.48	160.95	2.36	46.22	0.74
A19L	-83.09	3.75	44.93	1.12	128.02	3.92	13.29	0.34
A19Q	-74.82	1.92	43.24	1.47	118.06	2.41	3.33	0.35
A19S	-86.06	0.84	27.62	1.59	113.68	1.80	-1.05	0.28
A19Y	-66.21	1.81	50.76	1.85	116.97	2.58	2.24	0.18
N23E	-71.40	1.91	14.82	1.71	86.22	2.56	-28.51	0.83
N23F	-80.05	2.03	25.45	2.37	105.50	3.12	-9.23	-0.62
N23H	-75.22	1.98	30.87	1.42	106.09	2.44	-8.64	0.42
N23K	-78.44	0.88	86.10	4.14	164.54	4.23	49.81	0.88
N23L	-81.97	0.79	26.18	1.14	108.15	1.39	-6.58	-0.33
N23Q	-86.81	1.76	37.93	1.83	124.74	2.54	10.01	-0.58
N23S	-77.81	1.51	31.01	0.98	108.82	1.80	-5.91	-0.05
N23Y	-81.35	1.13	31.53	1.41	112.88	1.80	-1.85	-0.35

Figure 7. MM-GBSA data correlated with experimental ΔΔG values for each GB model tested.



Second, the calculated energies are highly inaccurate, and although this could be expected to a degree since, except for those contained in the implicit solvation calculation, entropic contributions are not considered in this technique, the extremely weak correlation between the data indicates almost random results (R^2 = 0.18 and 0.15 for IGB=2 and 1, respectively). It is possible that normal mode analysis of the structures could estimate the contribution of vibrational entropy and increase the quality of the correlation, but this would introduce additional error into the calculation, further exacerbating the issue of simulation time.

Examination of the results for each individual mutant (Table 5) reveals the inconsistencies in calculating binding energies for each of the mutants. Monomeric binding energies should all approximate each other given that the dissociation constant for MTX binding to DHFR remains essentially constant despite mutations to the Ala19 and Asn23 positions (see Chapter 3). The mutants A19H, A19K, and A19Q all fall outside one standard deviation from the mean value for C9 binding energy among all the proteins. Errors in the calculation of binding between the singly bound monomer and the second DHFR yield seven of the seventeen proteins falling outside one standard deviation. Although this represents a Gaussian distribution of error, the standard deviations of 5.48 and 6.08 kcal/mol for the first and second binding events, respectively, represents too large an error to make statistically significant conclusions about the relative differences in cooperativity between the wild-type and mutated DHFRs. Analysis of the individual energetic contributions in the mutants yields inconclusive results (Table 6). Innocuous mutations such as A19L appear to involve large changes (~50 kcal/mol) in

Table 5. Examination of energies in MM-GBSA calculations (IGB=2). All values are in kcal/mol.

Protein	ΔG_M	ΔG_D	ΔG_C
WT	-63.63	-117.95	-54.32
A19E	-63.23	-112.20	-48.97
A19F	-70.46	-127.44	-56.98
A19H	-79.62	-125.02	-45.40
A19K	-76.00	-115.62	-39.62
A19L	-63.67	-128.50	-64.83
A19Q	-59.72	-128.82	-69.10
A19S	-70.93	-111.68	-40.75
A19Y	-57.67	-121.08	-63.41
N23E	-64.87	-120.49	-55.62
N23F	-62.83	-124.41	-61.58
N23H	-66.42	-119.16	-52.74
N23K	-67.16	-111.19	-44.03
N23L	-69.02	-126.33	-57.31
N23Q	-68.10	-127.73	-59.63
N23S	-64.18	-116.14	-51.96
N23Y	-65.43	-118.58	-53.15
Average	-66.64	-120.73	-54.08
Std. Dev.	5.48	6.08	8.41

Table 6. Selected examples of variance in individual contributions to energy from MM-GBSA calculations (IGB=1 in this case) on the monomeric DHFR-C9 complex. All values are in kcal/mol. Note that the three selected free energies do not encompass all the terms in equations 10-13 that yield the total binding free energy.

Protein	ΔG_{ele}	ΔG_{vdW}	ΔG_{sol}	ΔG_{tot}
WT	-121.37	-53.94	105.79	-69.52
A19L	-168.11	-55.16	153.21	-70.06
A19H	-149.40	-72.37	135.10	-86.67
A19K	-147.98	-66.41	132.16	-82.23
A19F	-145.46	-63.70	132.19	-76.97
N23L	-160.80	-57.12	141.97	-75.95
N23H	-149.09	-58.01	134.49	-72.61
N23K	-138.86	-60.02	125.25	-73.63
N23F	-143.37	-54.36	128.95	-68.78

electrostatic and solvation contributions, while yielding similar overall binding energies. Additionally, mutations exploring two very different amino acid properties (i.e. A19K and A19F) appear to have more in common in terms of individual energetics than would be expected. Further decomposition of these energies into the contributions from the complex, receptor, and ligand reveals that treatment of the ligand (C_9) is consistent, and the differences lie primarily in the treatment of the protein (data not shown). Attempts to rectify this issue by simulating each of the structures in the analysis separately (full-trajectory MM-GBSA approach) were unsuccessful (data not shown). This analysis demonstrates the difficulty in determining the exact cause of the differences in binding free energies calculated, and a more rigorous method for calculating binding free energy may be required.

4. Conclusions

We have tested two methods – umbrella sampling and MM-GBSA – for estimating the free energy of protein cooperativity associated with chemically induced DHFR dimerization. In our studies, neither method has yielded acceptable results. For umbrella sampling, the flexibility and length of MTX hindered our attempts to calculate the free energy of MTX association with DHFR. Protein-ligand interactions that are unrestrained in this particular simulation cause the ligand to remain associated with the protein despite movement of the main binding moiety out of the active site and thus, non-convergence in our PMF. Although this particular issue could possibly be resolved by fixing the geometries of the protein and ligand, it is not clear how this particular

technique could be properly applied to a three-body system, though this does not preclude future attempts.

The endpoint-based MM-GBSA method also generated highly discouraging results. As a qualitative measure of our performance, we can compare our results to those collected by Sgobba and coworkers, who performed MM-GBSA calculations on a series of *P. falciparum* DHFR inhibitors[269] including the anitfolate pyrimethamine, which has a similar binding free energy for pfDHFR as MTX for ecDHFR (-12.7 and -12.6 kcal/mol, respectively). Our calculated binding energies without vibrational entropic considerations are reasonably close to the Sgobba group's estimates, and if we were to include entropic contributions into our calculations, we may increase our correlation to experimental data to some degree. However, this highly qualitative analysis only pertains to the first MTX_2C9 binding event, and given the inconsistent energetics associated with the proteins, reliable results may be difficult to obtain.

Additionally, the issue of error in the calculations still remains. The best errors reported using the MM-PB(GB)SA method tend to be on the order of 1-2 kcal/mol, which is insufficient to evaluate the small differences in binding energy observed in our experimental work. Increased simulation time may reduce these errors enough for results to be statistically relevant, but unless a better treatment of our system is developed (i.e. reparameterization of the force field or approximations yielding smaller errors for shorter simulations), the simulation time required is currently too great. Therefore, the best we can currently hope for out of theoretical modeling of this system is a qualitative screening process in which mutations yield the correct trends, so that we may include at least a

small degree of rational design in our future engineering. Alternatively, design of highly destabilizing interactions may lead to ΔΔG values that span a wider dynamic range, allowing for larger errors in our calculations without sacrificing statistical significance. Overall, it is clear that much more effort is required to establish an accurate model of chemically induced dimerization, and such work may require computing resources that exceed the capabilities of current technology and theory.

Chapter Five

Toward a Biomolecular Language of Self-Assembly: Stabilization of a DHFR Heterodimer

1. Introduction

Chemically induced dimerization has one of its most powerful applications in the regulation of protein proximity. Many cellular signaling pathways involve proximity as a means of transmitting information.[288] In proximity-based pathways, the subcellular localization of a signaling molecule is altered to favor interaction with the recipient. By creating a high concentration of one molecule in the vicinity of another, the rate of chemical reactions (i.e. phosphate transfer or nucleotide release) is increased. The induction of this proximity as regulated by a synthetic dimerizer has been used in a number of instances (see Chapter 1). Another application related to induced proximity is the formation of self-assembled protein nanostructures capable of bearing additional protein domains such as antibodies.[106,107] The induced proximity of the underlying protein architecture bolsters the stability of the structure by increasing the apparent concentration of the monomers in the proximity of the dimerizer.

While practical applications of chemically induced homodimerization such as these have been well-studied and exploited, the selective formation of heterodimeric complexes has been a difficult issue to address. In terms of both cellular signaling and structural engineering, it is possible to produce a functional heterodimer via fusion of two different proteins to the dimerization domains, purification of these heterodimers from a stochastic mixtures has a maximum possible yield of 50%. Clearly the formation of a heterodimeric dimerization domain represents a more elegant and precise approach.

Two practical techniques exist to achieve this goal. The ligand-directed approach involves asymmetric chemical synthesis of the dimerizer so it binds to two different

targets. Schreiber and coworkers have used this method to develop an asymmetric dimerizer targeting FKBP and CsA.[289] Fusion of FKBP to a membrane targeting domain and CsA to the Fas tail effected apoptosis upon addition of the dimerizer in cells transfected with the fusion proteins. Additionally, transcriptional activation was observed in cells transfected with FKBP fused to the GAL4 DNA binding domain and CsA fused to the VP16 activation domain. Other work by Lin et al. has generated a dexamethasone-MTX heterobifunctional ligand capable of recruiting a glucocorticoid receptor-B42 fusion to a DHFR-LexA fusion.[290]In the area of therapeutic development, Portoghese and coworkers have developed so-called MDAN ligands (μ-δ agonist-antagonist) that are composed of an agonist of the μ opioid receptor (oxymorphone) and an antagonist of the δ receptor (naltrindole). One such dimerizer, MDAN-21 (containing a 21-atom linker between the pharmacophores), produces neither tolerance nor dependence upon administration to mice. The authors conclude that this effect is a result of the heterodimer formed between the two opioid receptors.[291]

Though the ligand-directed approach has received much attention, it is not without its drawbacks.[292] Often the molecules of interest are already known to bind to the proteins of interest, so it is rarely necessary to devise a synthesis from scratch, but many synthetic variables come into play. For example, the site of tethering the molecules must be carefully considered. Pharmacophores may become disrupted in the process of joining them together. Ligand conformational equilibria will affect the behavior of the dimerizer. The typically large size and hydrophobic nature of the compounds often has deleterious effects on their solubility as well as therapeutic and *in vivo* usefulness. All these synthetic

considerations, among others, must be evaluated before a dimerization strategy can be developed.

Another avenue of inducing heterodimerization that has not received as much attention is the protein-directed method. In such a method, the contacts between the induced dimer are modified to achieve selective heterodimer formation. A survey of the literature reveals only sparse attempts at remodeling protein interfaces to induce heterodimer formation. Typically, the most efficient route to obtaining selective heterodimerization has been rooted in directional disulfide formation, where disulfides at engineered cysteine residues confer specificity.[100,293] Other methods have relied upon bump-hole pairing of complementary electrostatic or steric mutations.[294-297] To our knowledge, these strategies have not yet been applied to a chemically induced dimerization system, though switchable assembly of heterodimers represents a novel, intriguing field in which interface remodeling strategies can be studied in a highly controlled environment. Selective assembly of multiple heterodimeric complexes with a single dimerizer would be a highly advantageous, particularly in the field of protein nanostructural assembly.

Previous work in our laboratory has shown that stable, self-assembled protein nanorings can be generated via a $DHFR_2$ fusion protein and MTX_2C9.[106] Fusion protein linker length and composition has been shown to affect oligomerization of the protein subunits insomuch as longer linkers lead to smaller ring sizes due to the close proximity of the binding sites and the flexibility of the linker. Expanding upon this work, we have shown that protein nanorings bearing eight copies of anti-CD3 single chain antibodies

(DHFR-DHFR-antiCD3 fusions) increase the affinity of the antibody for its target and can be assembled and disassembled at will.[107] Furthermore, nanorings containing DHFR-human Hint fusion proteins form oligomers and retain their catalytic activity.[108]

While these self-assembled complexes are vastly interesting in their own right, the attractive possibility of a multivalent nanostructure remains a high priority. Given control over protein assembly, it becomes feasible to generate protein structures targeting multiple binding partners. One example would be a protein nanoring bearing both the anti-CD3 and anti-CD22 single-chain antibodies. Such a structure could, in principle, target both the T-cell receptor and CD22 present on the surface of B-cell tumors, bringing the cells into close proximity and representing a novel treatment for B-cell leukemia.

We have shown in Chapter 3 that we are able to modulate the stability of a chemically induced DHFR homodimer by introducing point mutations at the newly formed protein interface. Herein we report the stabilization of a chemically induced DHFR heterodimer, and the first steps toward the development of a biomolecular language for protein nanostructural assembly.

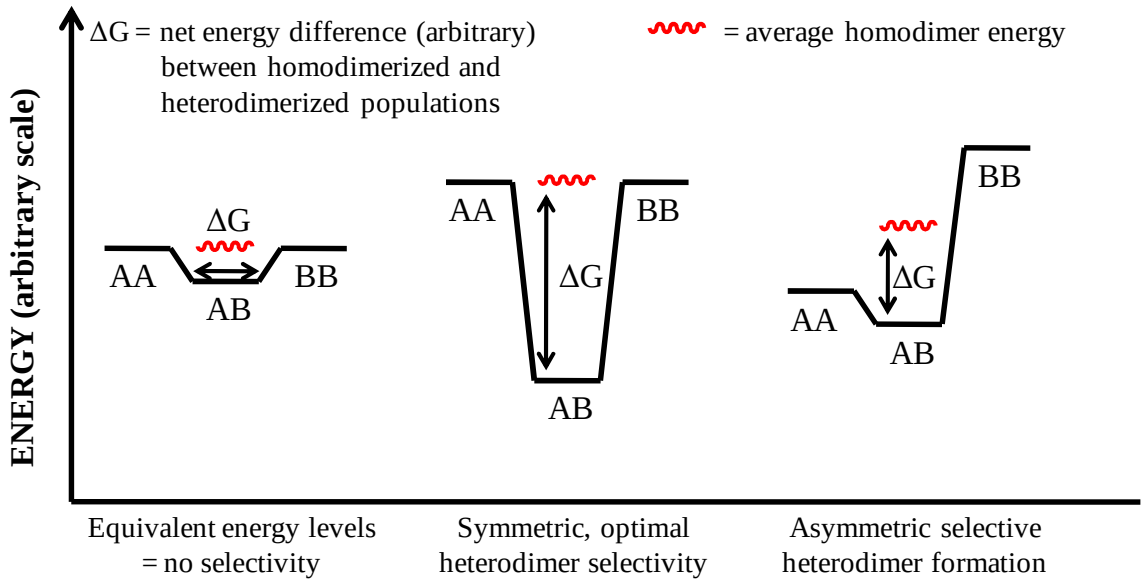
2. Experimental Design and Rationale

Based on simple statistical analysis, if there is no selectivity for homo- or heterodimer formation, the three species will distribute into a 1:2:1 ratio of AA:AB:BB. Modification of the energetics of any of these three species will perturb this distribution. Ideal selectivity arises from a heterodimer that is substantially lower in energy than either homodimer. Analysis of the dimerization energetics, however, reveals that

heterodimerization can be significantly favored if just one of the homodimers can be destabilized. This scenario exists naturally for the Jun-Fos transcription factor pair,[253,254] and has been demonstrated previously in the engineered version of the Arc repressor designed by Tidor and Sauer.[241] This somewhat counterintuitive principle is illustrated in Figure 1.

Given this consideration, we hypothesized that by introducing destabilizing mutations at the interface of homodimeric self-assembled DHFR, we could effect stabilization of the heterodimer. The most obvious way to perform this is to use the A19E/K or N23E/K mutations as complementary binding partners. Both show excellent destabilization of the homodimeric complex (see Chapter 3), and aside from the favorable electrostatic interactions that could assist in deepening the energy well of the heterodimer, these particular mutations are analogous to the bump-hole Ala19-Asn23 pairing present in the wild-type interface. In the context of our experimental design, we expect the 1:2:1 ratio to shift toward heterodimer formation. In addition to testing favorable electrostatic pairings, we also hypothesize that the other homodimer-destabilizing interaction discussed in Chapter 3 will favor formation of a heterodimeric species. Given this framework for experimental design, the key issue of detection must then be addressed.

Figure 1. Dimerization energy landscape – the energetics of heterodimer selectivity.

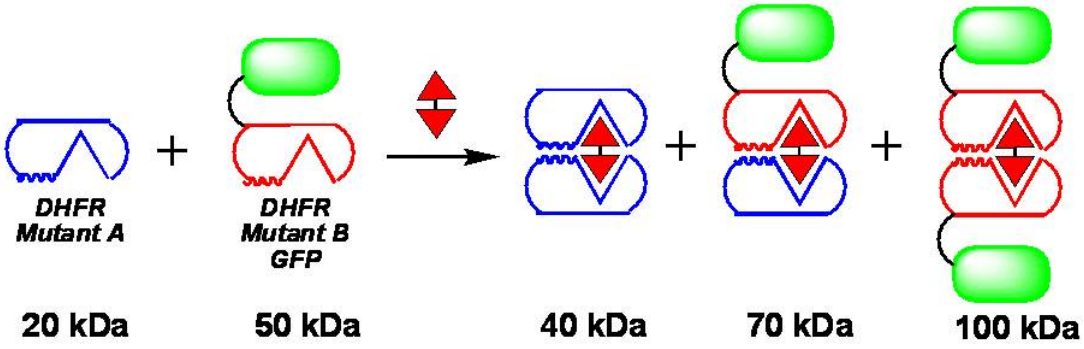


2.1. Progression from Gel Filtration to Fluorescence Detection

In our initial attempts to assay the formation of heterodimeric DHFR pairs, we decided to pursue a route with which we were previously familiar – the gel filtration assay. However, since the resolving power of the assay is based on the hydrodynamic radius (R_H) of the different species in solution, it is apparent that this technique will not resolve heterodimeric from homodimeric species if using proteins of similar R_H. Therefore, we decided to approach the problem using a DHFR fusion protein wherein the fused protein acts as a molecular weight tag, allowing for resolution of the different species when a non-tagged and a tagged mutant are assembled into the dimer (Figure 2). The selection of the size of the molecular weight tag must be made carefully, as the resolution of size exclusion chromatography is not a linear function; rather it is logarithmic, akin to gel electrophoresis. Additionally, it is important that the tag chosen be as inert as possible, since uncontrolled protein-protein interaction either with DHFR or with another tag will alter the observed R_H of the species being analyzed. We reasoned that the use of the monomeric fluorescent proteins eGFP, mRFP, and mCherry should yield the proper distribution of R_H while maintaining a low degree of protein-protein interactions within the test solution.

During the course of our preparation of the DHFR-FP fusion proteins, we encountered several roadblocks concerning protein expression that proved fatal to this approach (see section 3.1 for a discussion of these problems). In light of this, it was necessary to devise a new method for detection of heterodimeric complexes. Recent work in the fluorescent detection of protein complexes spurred our interest in this direction.

Figure 2. Rationale behind the fusion protein approach to resolving heterodimeric species.



Two-photon fluorescence correlation spectroscopy[298,299] (FCS) is one such technique that allows for the resolution of several protein species in a mixed solution. FCS is based on fluctuations in the fluorescence intensity observed in a small (< 1 fL) volume that are due to individual proteins entering or leaving the volume.[300] Correlation functions are then used to estimate the concentration of the protein complexes, and in the context of dual-color FCS, protein-protein interactions can be analyzed. Brightness analysis, related to the average fluorescence intensity of a single particle, provides a means for a quantitative characterization of protein-protein interactions, as it encodes the stoichiometry of the complex.[301] Conceptually, if a heterodimeric complex bears two different fluorophores, the brightness of the molecule in two different detection channels (typically a red and a green channel) can be quantified. If a homodimeric complex enters the observation volume, the brightness will correspond to twice the brightness of just one fluorophore in the proper channel. In this manner, the stoichiometric composition of an unknown mixture of complexes can be quantitated. In practice, however, we found that the method relied very heavily on the photophysics of the selected fluorophores, and that unless the brightness of the two fluorophores was approximately equal, extraction of stoichiometric data became impossible (*vida infra*).

Another fluorescence technique, however, presented itself as a viable replacement for FCS. Two-color coincidence detection[302,303] (TCCD) and alternating laser excitation[304] (ALEX) are methods similar to FCS in that they rely upon a femtoliter observation volume and two separate detection channels to resolve proteins or complexes bearing two different fluorophores. However, instead of a single excitation beam, two

distinct and overlapping lasers are used to excite the fluorophores. In TCCD, continuous excitation of the sample volume by both lasers is employed, whereas in ALEX, the lasers alternate excitation fast enough to excite a molecule with both wavelengths as it passes through the observation window (Figure 3). Both methods yield similar results despite this subtle difference. These single-molecule techniques represent powerful methods for observing the properties of ensembles of billions of molecules by observing just a few single molecules due to ergodicity. When ergodicity breaks down, observation of single molecules yields far more information than ensemble averages, and the small scale of the experiments lends itself toward a significant reduction in the amount of material necessary to draw experimental conclusions. These methods presented a highly attractive means of detecting heterodimerization in our chemically induced complexes, and herein we report the trials of optimizing a detection method and the results obtained from ALEX analysis of our DHFR complexes.

3. Results and Discussion

3.1. Fusion Protein Generation and Dimerization Assays

Our attempts to develop a gel filtration assay for heterodimerization focused on the use of fusion proteins of DHFR wherein the fused protein acted as a molecular weight tag to allow for the resolution of the three different species. A general scheme of our proposed fusion protein development is shown in Figure 4. The first attempt at a fusion we pursued was the development of a DHFR-eGFP construct. Repeated attempts at

Figure 3. Schematic representation of the ALEX assay. A) shows the ALEX microscope setup, utilizing two independent lasers with dichroic mirrors to direct the beams. Electrooptical modulators combined with polarizers result in an alternating laser excitation. After filtering, emission is detected on photodiodes. B) shows a typical data trace from an ALEX assay, noting the coincident bursts corresponding to protein complexes. Figure reproduced from Kapanidis et al.[304]

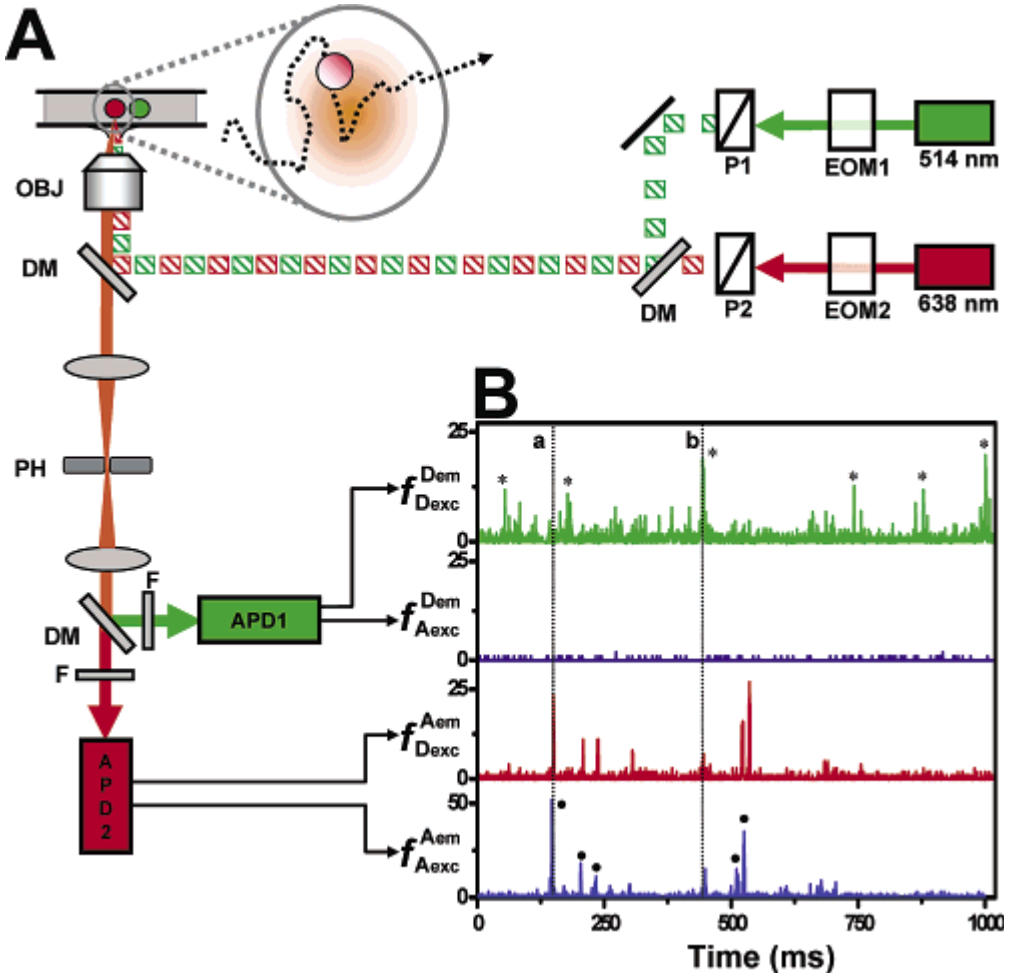
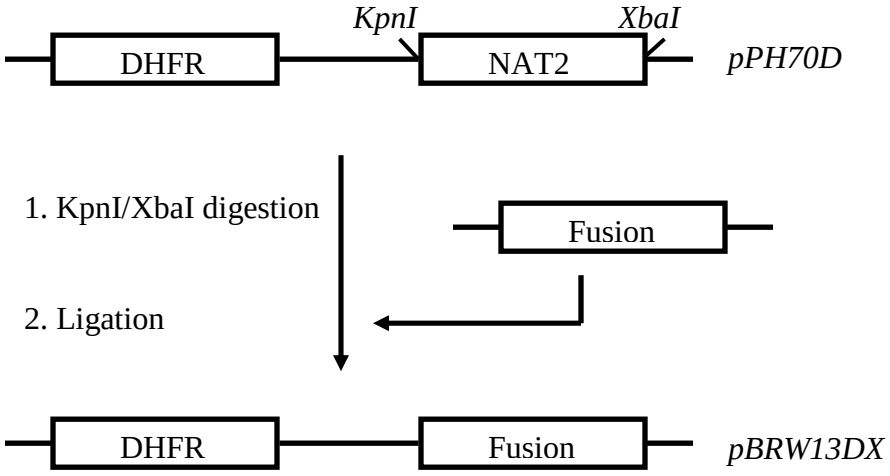


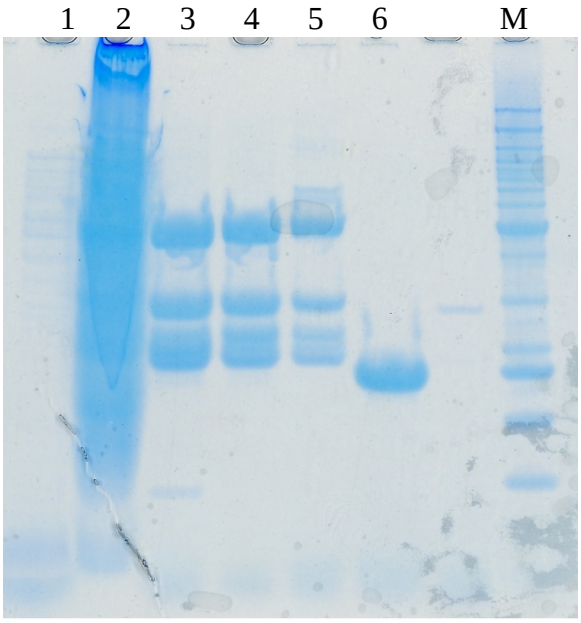
Figure 4. General DHFR-FP fusion protein generation scheme. The 'X' in the fusion protein plasmid name corresponds to the fusion protein installed (G – eGFP, M – mRFP, C – mCherry).



cloning and expressing the plasmid yielded unsatisfying results (data not shown), and the finding that eGFP occasionally forms a weak dimer[305] lead us to change the fluorescent protein to the monomeric red fluorescent protein, mRFP. Cloning of the mRFP plasmid was successfully achieved (as verified by plasmid sequencing) and expression and putification efforts commenced. SDS-PAGE analysis of this expression is shown in Figure 5. Though the protein contained contaminating bands at 26, 30, and 37 kDa, we elected to attempt dimerization analysis and optimize the purification later. Results from our preliminary gel filtration experiments can be found in Figure 6. The data collected presented several unexpected issues. First, the 13DM fusion protein eluted later in the analysis than the wtDHFR homodimer (Figure 6A and B), indicating a smaller hydrodynamic radius despite a two-fold increase in mass. Even more astonishing, the homodimer of the fusion protein appeared to adopt an even smaller hydrodynamic radius compared to that of its monomeric cognate. Heterodimerization yielded two peaks, one corresponding to a wtDHFR dimer and another that apparently corresponded to a wtDHFR:13DM heterodimer (Figure 6D). However, resolution of the two peaks was poor and, when the unexpected R_H values of the fusion protein are taken into account, this peak cannot reliably indicate a measure of heterodimerization.

As can be seen in Figure 4, the first fusion constructs expressed contained a 13 amino acid linker between the two proteins. We reasoned that the apparent decreases in R_H were due to the flexibility of this linker and interactions of the fluorescent protein with its DHFR partner. If the proteins were to interact strongly, the resulting globular structure could, in principle, yield a compacted species with a smaller relative R_H despite

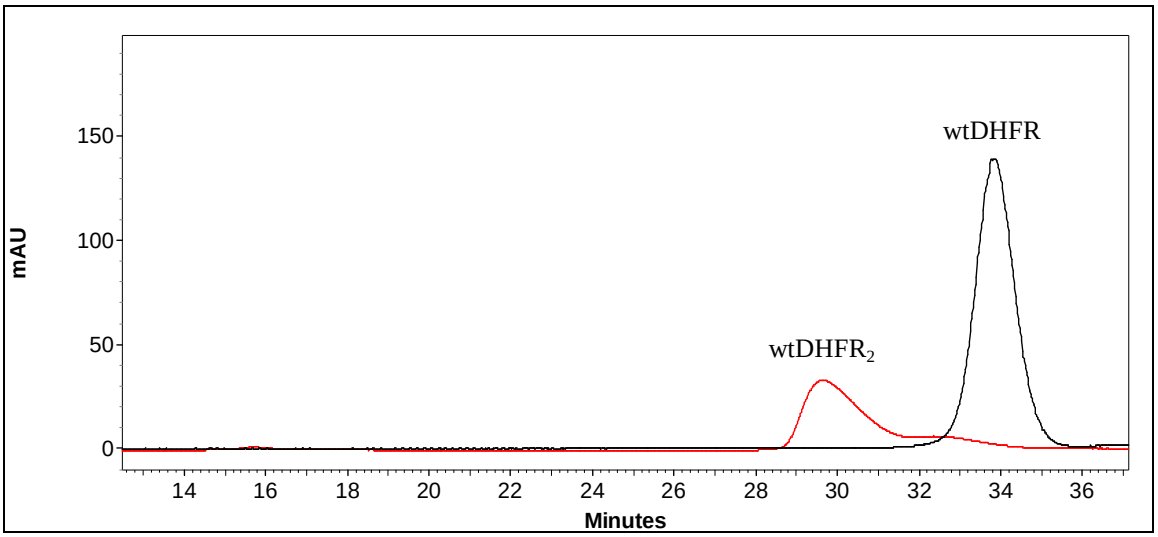
Figure 5. SDS-PAGE of expression and purification of DHFR-mRFP with a 13 amino acid linker (13DM).



M –Bechmark Ladder
1 – Uninduced
2 – Sonication supernatant
3 – post-MTX
4 – 13DM
5 – 13DM, 10x dilution
6 – wtDHFR

Figure 6. Gel filtration results of attempted DHFR:DHFR-mRFP heterodimerization. A) wtDHFR and wtDHFR treated with 2 equivalents of MTX_2C9. B) 13DM and 13DM treated with 2 equivalents MTX_2C9. C) wtDHFR and 13DM, no dimerizer. D) All species listed in A-C and 1:1 13DM:wtDHFR with 2equivalents MTX_2C9. Expected molecular weights: wtDHFR, 18 kDa; 13DM, 46 kDa; $wtDHFR_2$, 37 kDa; $13DM_2$, 93 kDa; wtDHFR-13DM heterodimer, 65 kDa.

A.



B.

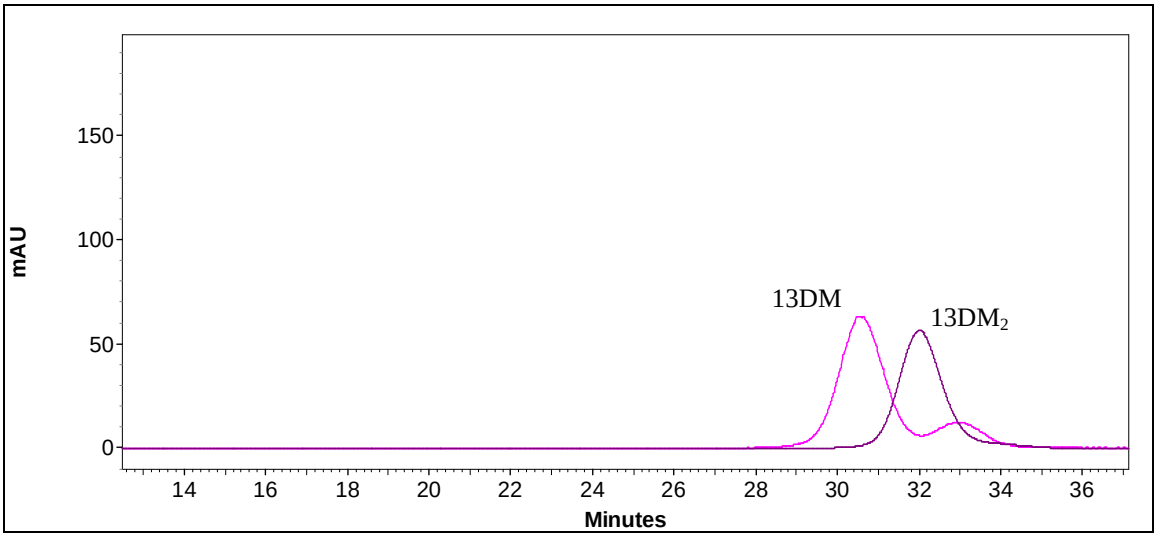
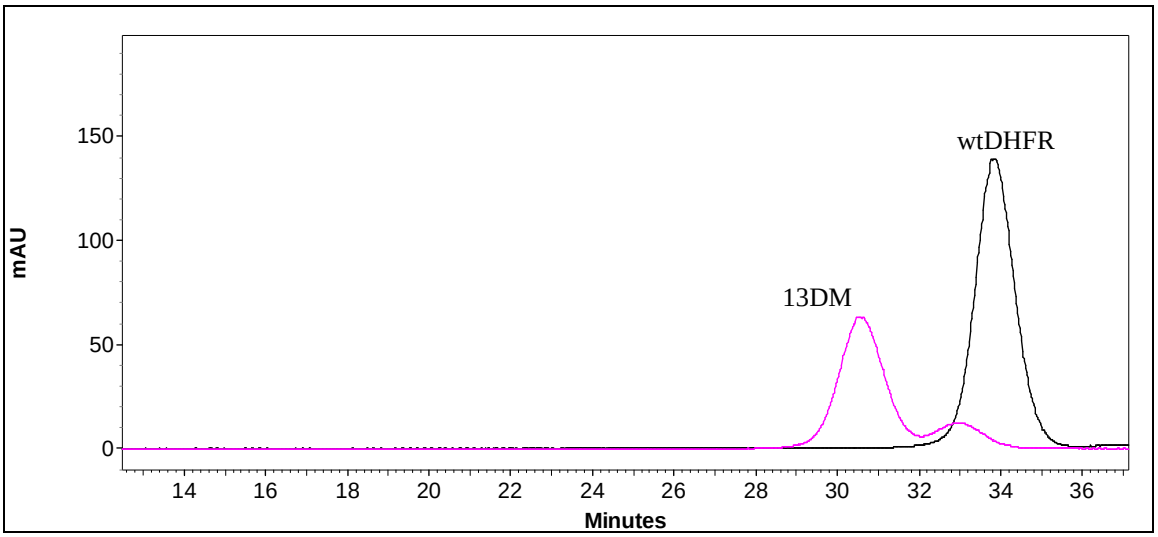
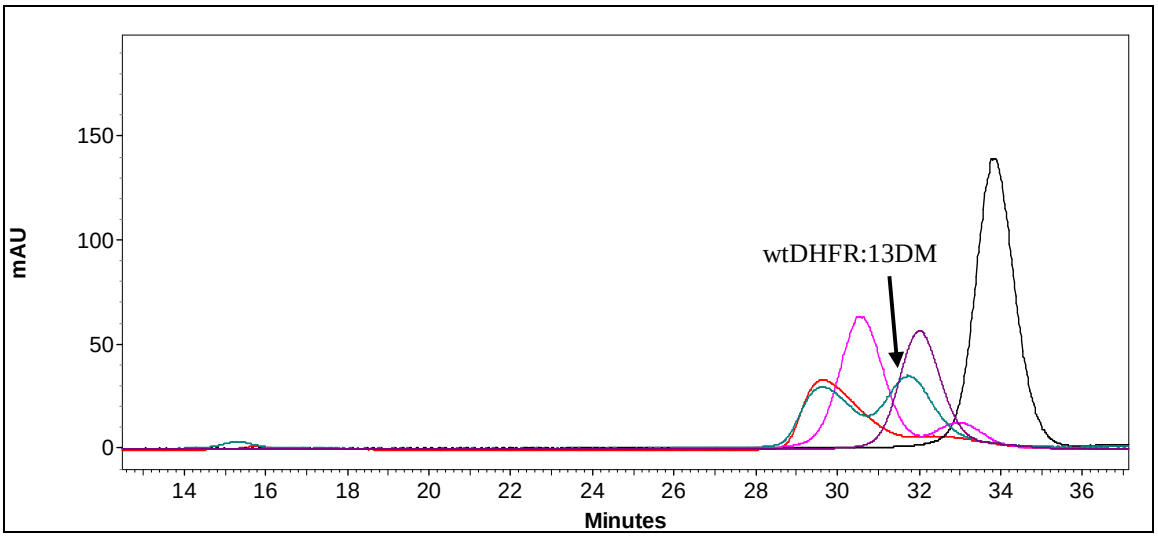


Figure 6, continued.

C.



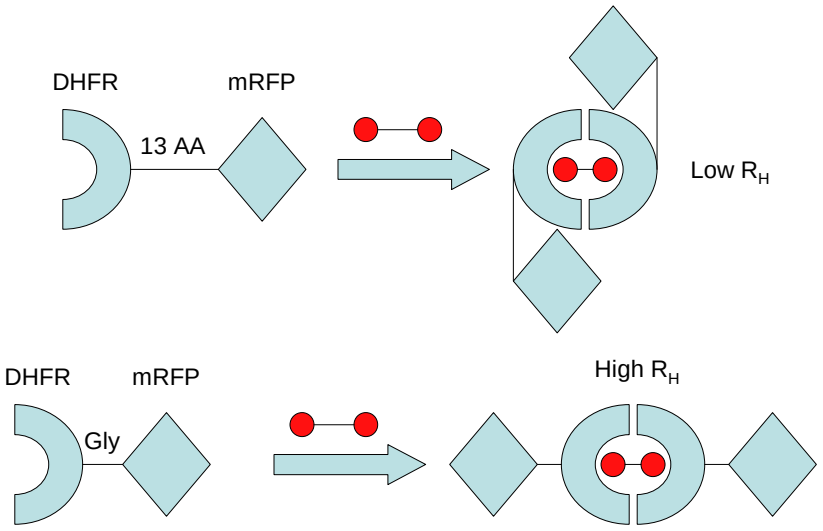
D.



increased mass. Additionally, the presence of the smaller protein contaminants was troubling. If these proteins were, in fact, able to bind the MTX column used during purification, it is possible they are fusion protein truncations that retain DHFR activity and, hence, affinity for the dimerizer, which would greatly confound the gel filtration results. To attempt to circumvent both of these problems, we elected to shorten the linker between the proteins to a single glycine, performing two main functions. First, glycine-rich linker regions (particularly the commonly used Gly_3Ser motif) used in fusion protein production can have deleterious effects on protein expression, since linkers of this nature adopt a highly extended, hydrated conformation that is susceptible to proteolysis.[306] Our 13 amino acid linker sequence contained the sequence N–GLGGGGGLVPRGT–C. Secondly, we hypothesized that the flexibility of the linker region allowed for interaction between the fluorescent protein and its DHFR cognate. If the two proteins interact strongly enough, they may form a compact globular structure with a smaller relative R_H, despite the increase in mass. This principle is illustrated conceptually in Figure 7.

Around this time, our laboratory became interested in collaboration with the Distefano group at the University of MN Department of Chemistry concerning a DHFR-mCherry-CVIA fusion protein and the possibility for the generation of DHFR-oliogomerized DHFR-DNA nanostructures. The basis for this structural assembly has been described by Duckworth et al.[307] Since we were encountering issues with the DHFR-mRFP fusion protein and desired to work more closely with procedures already established by the Distefano group, we elected to switch to a DHFR-mCherry-CVIA fusion construct, using the mCherry-CVIA sequence gifted to us from the Distefano lab.

Figure 7. Cartoon representation of the putative effects of reducing fusion protein linker length.

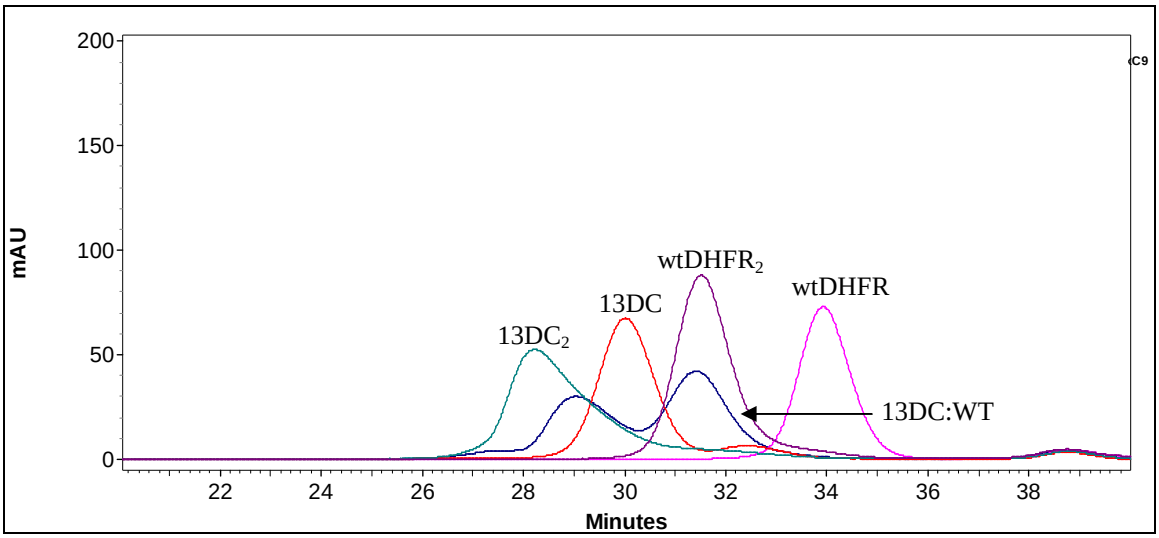


Expression of the first mCherry-CVIA fusion protein was attempted with a direct replacement of the mRFP section of the plasmid construct, yielding a DHFR-13 amino acid-mCherry-CVIA fusion (13DC-CVIA). While the contaminating bands were still present in SDS-PAGE analysis of the purified protein, we decided to test dimerization of the protein. The results from our analysis can be found in Figure 8. Interestingly, the distribution of the dimerized and monomeric species improved dramatically, insomuch as the hydrodynamic radii of the species were brought into alignment with the relative masses, unlike our DHFR-mRFP fusion protein. However, the heterodimeric species remained divided among two species rather than three, and resolution between them remained poor. We reasoned that the contaminating bands must be removed in order to assure deconvoluted analysis of the gel filtration results.

Initial attempts at expression of the DHFR-Gly-mCherry-CVIA (1DC-CVIA) fusion protein resulted in the persistent presence of the contaminating bands observed in Figure 5 (*vida infra*), refuting the hypothesis that our poly-Gly linker was the culprit behind our contaminants. Exhaustive analysis of the possible problems associated with fusion protein expression yielded several possible issues that could be arising during our work. First, rare codon bias pertains to the limited availability of particular tRNA pools for the expression of non-bacterial proteins in bacterial expression lines.[308] To address this problem, we changed our expression cell line from BL21 *E. coli* to the *Rosetta 2* expression line (Stratagene). This cell line contains additional plasmids coding for the limited tRNAs. Expression in this cell line did not reduce the formation of the truncated products. Secondly, the protein expression rate in cells depends directly on the

Figure 8. Gel filtration analysis of wtDHFR and 13DC-CVIA. A) Note that the distribution of the R_H of each of the protein species becomes more aligned with molecular weight than previous experiments with DHFR-mRFP.

A.



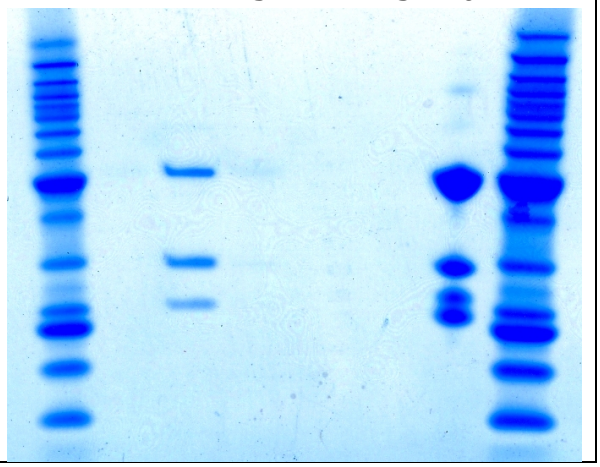
stability of the mRNA transcript, and *in vivo* degradation of this transcript can yield truncated protein expression.[309] To explore this possibility, we again changed our expression line to the BL21 Star cell line (Invitrogen), which is RNAseE deficient, putatively improving mRNA stability. Unfortunately, this workaround was also met with failure. Lastly, excessive stress to the bacteria during overexpression may yield shortened products, so we examined the effects of low-temperature growth, shorter induction times, and richer media on expression.[309] Again, the contaminating protein bands persisted in SDS-PAGE analysis. We hypothesized that the truncated products were a result of stalled translation since the proteins apparently still retained DHFR activity, as evidenced by their affinity for the MTX column. Therefore, we reasoned that introduction of a C-terminal His_6 tag would allow for the isolation of full-length fusion protein.

Expression of our newly constructed DHFR-Gly-mCherry-His_6-CVIA (1DC-His-CVIA) again yielded disappointing results. The protein was first purified on a Ni-NTA column yielding the SDS-PAGE results shown in Figure 9A. Although one contaminating band has apparently been removed via this addition, two still remain. We now hypothesized that the protein contaminants were actually not truncations at all, but highly stable conformers of the full-length fusion protein that were resistant to heat and SDS denaturation. This was tested by exposing the protein to the chemical denaturants Urea and Guanidine-HCl (Figure 9B). As evidenced by the SDS-PAGE gel, no significant denaturation took place. Due to the strongly denaturing conditions the protein mixture was exposed to, our hypothesis was thus unsupported. In an effort to discover the identity of these contaminating proteins, we submitted

Figure 9. A) Purification of the His-tagged DHFR-mCherry fusion protein via Ni-NTA chromatography. B) Denaturation assay of 1DC-His-CVIA.

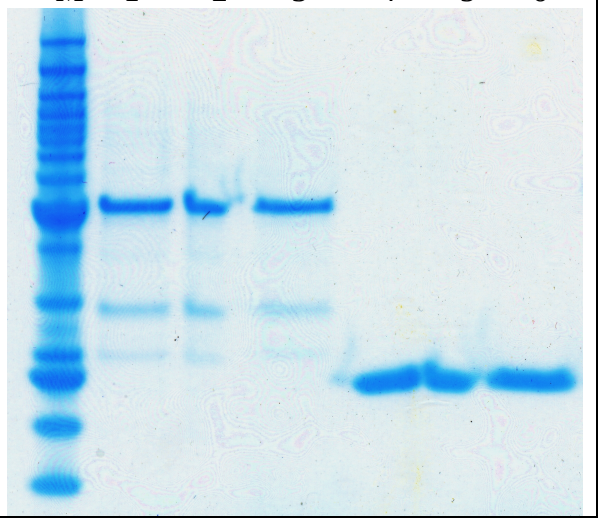
A.

M – Protein benchmark ladder
1 – Fxn 1from Ni-NTA column
2 – Fxn 2
3 – Fxn 3
4 – Fxn 4
5 – Fxn 5
6 – f13DC-CVIA (purified From MTX colum)



B.

M – Protein benchmark ladder
1 – 1DC (untreated)
2 – 1DC (8M Urea)
3 – 1DC (6M Guanidine-HCl)
4 – wtDHFR (untreated)
5 – wtDHFR (8M Urea)
6 – wtDHFR (6M Guanidine-HCl)



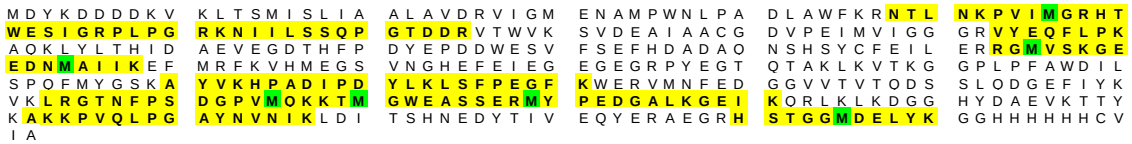
the excised bands from the gel in Figure 9A (lane 2) to the University of Minnesota Mass Spectrometry facility for tryptic digest and tandem mass spectrometry analysis. Results of this analysis are presented in Figure 10. Examination of the data reveals that peptides corresponding to the full-length 1DC-His_6-CVIA protein were found in both the top and middle bands (36% and 29% sequence coverage, respectively), and peptides contained in DHFR were present in the bottom band (16% coverage). These data supported the assertion that *all three* bands contained at least DHFR and the His tag (due to Ni-NTA affinity). Consequently, it was decided to discontinue pursuit of the fusion protein tag method of heterodimer characterization due to these pervading, complex expression issues and develop a different method of dimer analysis utilizing the FCS or single-molecule TCCD/ALEX techniques.

3.2. Fluorescence Correlation Spectroscopy

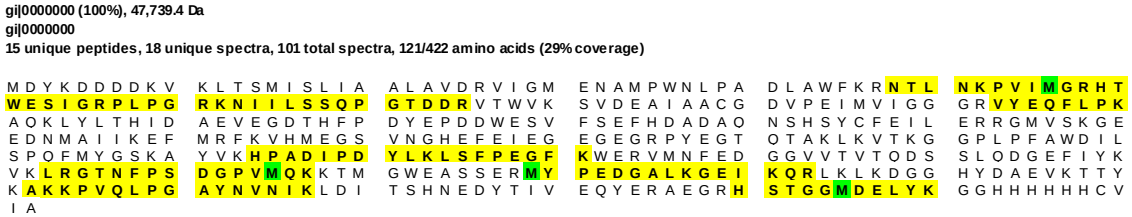
Our research into alternative methods for stoichiometric analysis of dimeric complexes led us first to FCS and collaboration with the lab of Dr. Joachim Muller in the Department of Physics at the University of Minnesota. In order to prepare proteins for analysis with this method, we first optimized a protein labeling by generating a cysteine-free variant of DHFR known as AS-DHFR. Then, a short glycine linker and a cysteine were introduced at the C-terminal end of the protein to provide a known attachment point for maleimide-functionalized dyes (ASC-DHFR). Utilizing the solid-state labeling methods described by Weiss and coworkers[310] in conjunction with

Figure 10. Results from tryptic digest and MS/MS analysis of excised bands corresponding to 1DC-His$_6$-CVIA expression. A) upper band, B) middle band, C) lower band.

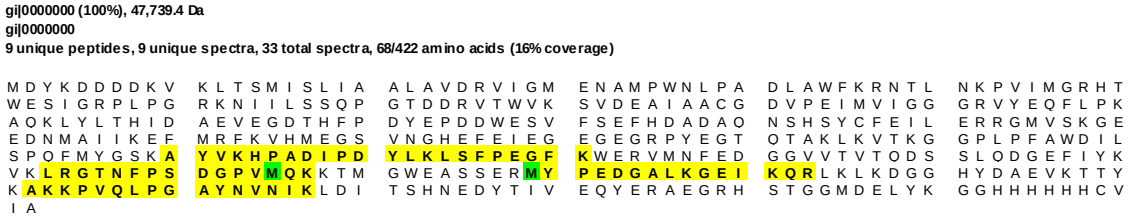
A.



B.



C.



purification via gel filtration chromatography, we were able to isolate the pure, labeled protein in excellent yield (≥ 90% labeled). We verified the stability of the protein conjugates using gel filtration and found the protein conjugates to be stable for at least one week at room temperature (data not shown). Additionally, for our application, FRET between the two dyes is undesirable, since our aim was to acquire two discreet emission signals that must be generated from the emission of each dye separately, rather than from the secondary emission of fluorescence due to FRET. Although comparison of the excitation and emission spectra of Alexa-488 and Alexa-647 indicates minimal excitation of Alexa-647 at the emission wavelength of Alexa-488 (520 nm), we tested the dimer complex for FRET and found no significant response (data not shown). Lastly, the dimerization behavior of the labeled proteins was verified by gel filtration chromatography.

In order to validate the FCS method for use in characterizing heterodimers, we first explored the behavior of our homodimeric system. The results of this analysis can be found in Table 1. The data indicates that the Alexa-488 dimeric complex is only 1.5x brighter than the monomer, and that the Alexa-647 dimer is equal in brightness to the monomeric species. These data present a puzzling problem. While an excessively low concentration of the complex would result in dissociation of the dimer, this assay is performed at 5 nM, an order of magnitude higher than the K_d of MTX for DHFR, suggesting that concentration issues do not enter into the analysis. Attempts to adjust the wavelength of excitation did not remedy the problem, nor did modification of the sample buffer. It was reasoned that the photophysical properties of the labels (i.e. label-protein

Table 1. FCS data for the Alexa-488 and -647 homodimers. Values represent the average of at least three independent trials.

Sample	Average Brightness	Brightness Ratio
Alexa-488 Monomer	16.624	1.52
Alexa-488 Dimer	25.226	
Alexa-647 Monomer	6.189	0.99
Alexa-647 Dimer	6.148	

interactions, spatial orientation of the labels, or label quenching affecting brightness correlation analysis) was the source of the error. Problems such as these are exceedingly difficult to unravel, as uncovering the exact source of the problem requires extensive testing of a multitude of available labels. Additionally, the global fitting procedure required to obtain a statistical analysis of dimer stoichiometry requires that the brightness of the two labels be approximately equal. Even without proper changes in brightness between monomer and dimer, it is apparent that the brightness of the Alexa488 label is ~twofold greater than Alexa647, which would lead to an overrepresentation of the Alexa488 conjugate. With these problems in mind, we elected to adjust our focus to a technique less dependent on the photophysical properties of the labels, ALEX.

3.3. ALEX Assays of Dimer Stoichiometry

The ALEX assay represents an advantage over the two-photon FCS method since it uses two distinct excitation lasers as opposed to the single, high-wavelength laser. Therefore, discrete excitation and subsequent quantitation of different fluorophores and their emission is easier to achieve. To validate that we could use the ALEX assay to determine the stoichiometry of our protein complexes, we first explored the use of the assay to measure a normal distribution of protein complex assembly. When two populations of the same DHFR, each labeled with a different fluorophore, are mixed in equal parts, they will sort into a random distribution of “pseudo” homodimers and heterodimers in the 1:2:1 ratio. This, of course, assumes there is no selectivity present in the distribution caused by the addition of the labels. Results from our analysis of these

populations are shown in Table 2. The data is presented in terms of the total number of fluorescent bursts and the manner in which the bursts occurred (in either one or the other or both detection channels). Since bursts occurring in both channels at the same time can only correspond to dimerized protein, a ratio of pseudohomodimer to pseudoheterodimer can be obtained. Given a stochastic dimerization, the total fraction of dimer present in the sample will be equal to twice the fraction of coincident bursts detected in the course of the acquisition, since half of the total coincident bursts will correspond to the number of homodimers present in each single channel (i.e. 1:2:1 ratio of A-XX-A to A-XX-B to B-XX-B).

Since the total protein concentration in the experiment (50 pM) is at a concentration below that of the K_d of DHFR complex formation (~250 pM), a homogeneous solution of fully dimerized species in not achievable. However, mathematical modeling of the behavior of DHFR dimers at low concentration reveals several fortuitous implications. At lower concentrations, changes in K_c become much more sensitive to changes in the fraction of dimer (Figure 11). Therefore, given a lower concentration, changes in K_c are directly proportional to the fraction of dimer (δ) present in the mixture, and therefore, relative ΔΔG values may be obtained from a series of mutations by the Equation 1.

$$\Delta\Delta G = -RT \ln \frac{\delta_{\text{mutant}}}{\delta_{\text{wt}}} \qquad (1)$$

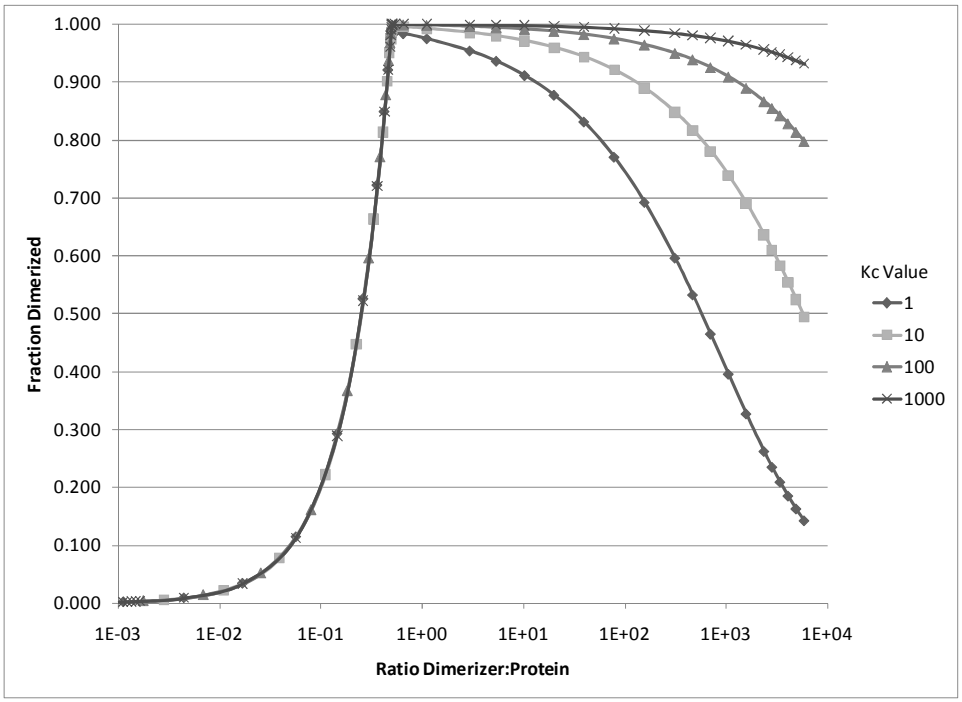
These values, along with a comparison to the ΔΔG values obtained from competition experiments, can be found in Table 3 and Figure 12.

Table 2. Results of the ALEX assay for equimolar mixtures of DHFR labeled with either Alexa488 or Alexa647. The total number of fluorescent bursts, the bursts in each channel and in both channels, the ratio of bursts in each channel, and the fraction of dimer corresponding to the bursts.

Protein	Total Bursts	Ch 1	Ch 1 & 2	Ch 2	Fraction Dimer
WT	2113	799	378	936	0.358
N23Q	2324	907	621	796	0.534
N23S	2100	1049	322	732	0.307
N23L	1874	780	155	939	0.165
N23Y	1168	573	95	500	0.163
A19Q	1689	680	116	894	0.137
N23H	1835	1002	106	727	0.116
N23K	2493	1215	119	1159	0.095
N23E	1686	936	76	675	0.090
A19F	1910	1025	85	800	0.089
A19Y	1881	968	59	855	0.063
N23F	1931	1164	57	710	0.059
A19L	2145	1050	62	1033	0.058
A19H	2120	1172	56	893	0.053
A19K	1762	868	37	857	0.042
A19E	2382	1188	45	1149	0.038
A19S	1490	707	27	756	0.036

Figure 11. Model dimerization data for a total protein concentration of A) 5 µM and B) 50 pM.

A.



B.

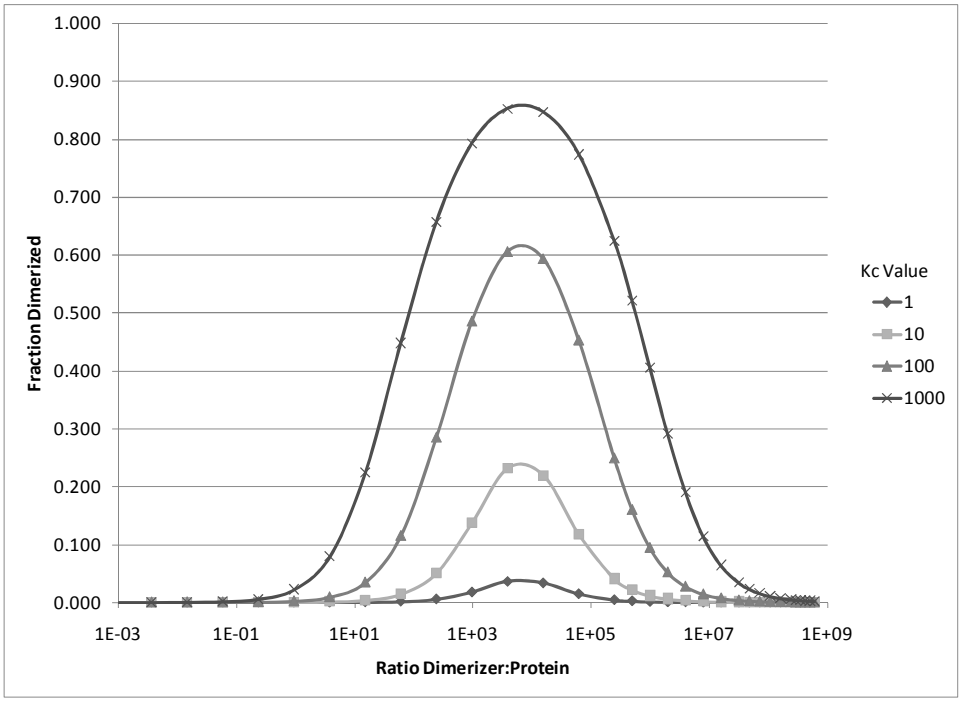
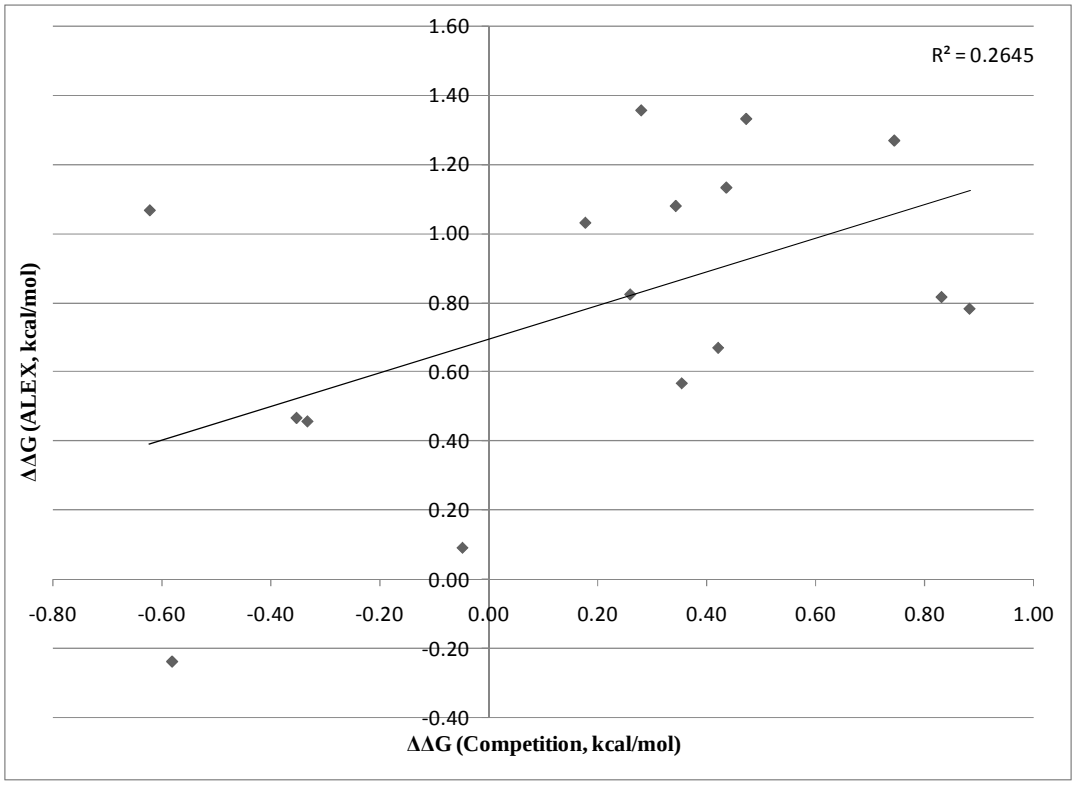


Table 3. Comparison of ΔΔG values obtained from competition and ALEX assays.

Protein	ΔΔG (comp)	ΔΔG (ALEX)	Protein	ΔΔG (comp)	ΔΔG (ALEX)
A19E	0.47	1.33	N23E	0.83	0.82
A19F	0.26	0.82	N23F	-0.62	1.07
A19H	0.44	1.13	N23H	0.42	0.67
A19K	0.74	1.27	N23K	0.88	0.78
A19L	0.34	1.08	N23L	-0.33	0.46
A19Q	0.35	0.57	N23Q	-0.58	-0.24
A19S	0.28	1.35	N23S	-0.05	0.09
A19Y	0.18	1.03	N23Y	-0.35	0.47

Figure 12. Correlation of ΔΔG values obtained from competition and ALEX assays.



Although the correlation to the competition data is weak, we must address several issues before we can discount the comparison. First, it is important to note that, in the context of this experiment, careful sample preparation for the ALEX assay is of the utmost importance. Due to the sensitivity of fraction dimer observed upon the protein concentration, small errors in concentration estimates of the protein will result in large errors in K_c estimate. Secondly, in contrast to the MTX competition assays, the thermodynamics of dimerizer binding to DHFR becomes an issue. Whereas in the competition assay, the fully dimerized complex is dissociated by addition of the competitive inhibitor and there will be no unbound DHFR present, in the ALEX assay, interaction between the unbound half of C9 and the surface of DHFR will hinder the second binding event. This issue presents a significantly challenging problem to test; however, computational study of the relative binding energies between DHFR mutants and C9 may lend some insight. A summary of our analysis, gleaned from results presented in Chapter 4, indicates that for over 60% of the proteins exhibiting a higher (more destabilized) $\Delta\Delta G$ in the ALEX assay, relatively favorable binding energy between C9 and DHFR exists on the order of -0.3 kcal/mol (data not shown). This indicates that hindrance of the second binding event due to surface interactions of the protein with the free end of the dimerizer may lead to an overall higher K_d of complex formation, and thus a lower fraction dimer observed in the experiment. Upon exploring these issues, it is perhaps unsurprising to notice a difference in observed $\Delta\Delta G$ values. Given that the ALEX assay does not contain the extra factor of MTX, and is a more direct observation

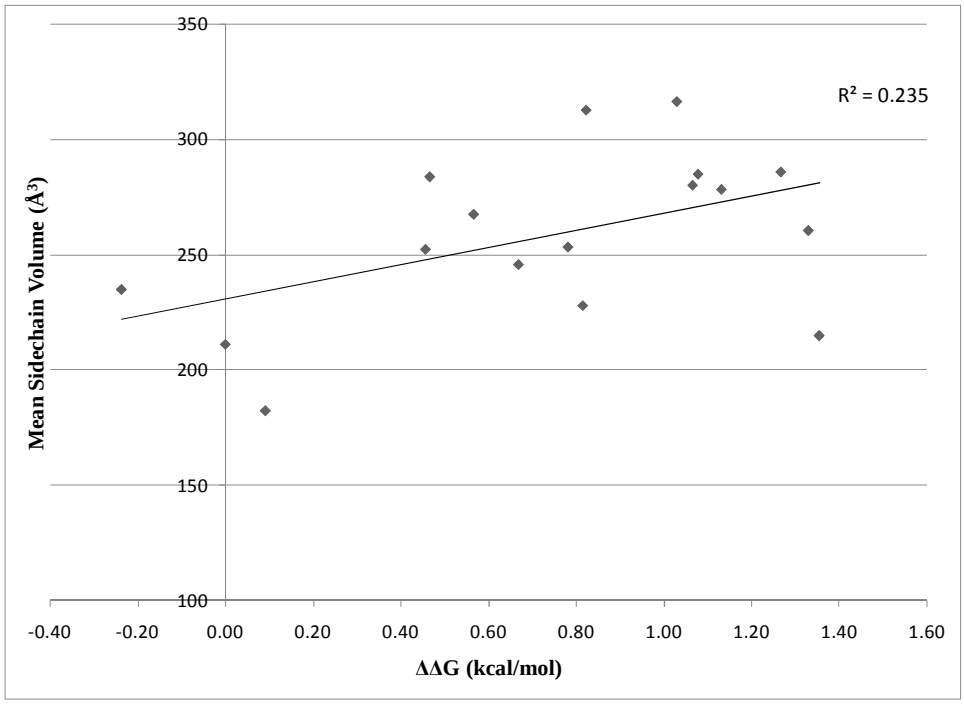
of dimerization behavior, we believe the results from the ALEX assay to be superior to those obtained from the competition assays.

Performing an analysis similar to that presented in Chapter 3, we examined the relationships between sidechain bulk and hydrophobicity to the stability of the interface. The results of this query can be found in Figure 13. Interestingly, while the destabilizing relationship between steric bulk and $\Delta\Delta G$ becomes slightly more pronounced in the ALEX experiments relative to the competition experiments (see Chapter 3), the moderate stabilizing correlation between sidechain hydrophobicity and $\Delta\Delta G$ is significantly reduced. It is possible this observation is due to the effects of lowered concentration on the ensemble. In the context of the competition experiments, the protein complex is already formed and dissociated via the addition of inhibitor, reducing observable effects on the kinetic parameters of the system. As reviewed above, in the ALEX assay, concentration is the only perturbing factor. The addition of steric hindrance may lead to a reduction of k_{on} for dimer formation. With no coupled change to k_{off} occurring, as evidenced by relationship between sterics and $\Delta\Delta G$ in the competition experiments, a lower amount of fraction dimer will be observed, leading to a reduced K_c value and a destabilized dimer.

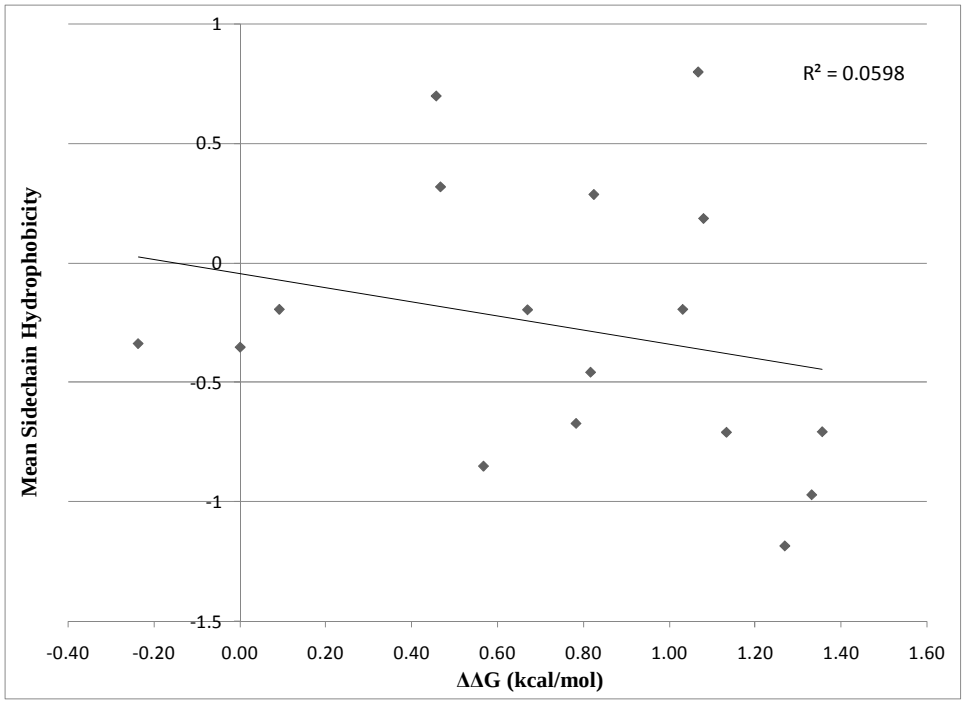
Concerning the observed change in stability associated with the introduction of hydrophobic residues, it has been shown that increased protein concentration is linked to higher thermodynamic stability of protein complexes.[311,312] This effect is related to changes in translational and conformational entropy associated with macromolecular crowding, insomuch as it is more favorable entropically to form a desolvated cavity in

Figure 13. Relationship between ΔΔG and A) sidechain steric bulk and B) sidechain hydrophobicity.

A.



B.



which to form the dimer when proteins are in close proximity. Increased hydrophobicity similarly leads to the sequestration of proteins into entropically favored cavities. Whereas experimental conditions with high concentrations of protein would favor the rapid formation of a “hydrophobic dimer cavity” in aqueous solution, those with low protein concentrations would have to pay a penalty in terms of dissolution of the “monomer cavity” and subsequent reestablishment of the dimer cavity. This penalty would be less for proteins with more hydrophilic character, as entropy has less of an effect on solutes able to interact electrostatically with the surrounding water molecules, hence the decrease in observed correlation between dimer stability and hydrophobicity.

With the hypothesis that a destabilized homodimer should favor heterodimer formation, we selected several of the most destabilized homodimers and paired them with other destabilized species in an effort to selectively generate a heterodimeric species. The results of this study are presented in Table 4. Based on analysis of the number of coincident bursts, which result only from heterodimer passage through the detection volume, we obtain an estimate of the amount of heterodimer in the solution. Background occurrence of coincident bursts was shown to be minimal for the wild-type homodimer (data not shown), and this was judged sufficient for all species since no change was introduced to significantly affect protein diffusion rates. In this experiment, unlike the fraction dimer calculated in the homodimerization experiment, the fraction of heterodimer is calculated only from the coincident bursts.

To estimate the amount of homodimeric pairs present in each of the singly-labeled populations (which contain both homodimers and monomeric DHFR), we apply the

Table 4. Results of ALEX assays of heterodimeric pairs.

Protein Pair	Bursts	Ch 1	Ch 1 & 2	Ch 2	Fraction Heterodimer
N23H/N23E	997	350	171	476	0.172
N23E/N23K	986	328	101	557	0.102
N23H/A19S	545	264	22	259	0.040
A19S/A19Q	550	242	18	290	0.033
N23F/A19Y	497	216	17	264	0.034
A19Y/A19L	797	348	27	422	0.034
A19Y/A19S	591	292	14	285	0.024
N23H/N23K	1140	518	59	563	0.052
A19Y/A19S	1060	494	24	542	0.023
A19F/A19Y	880	369	23	488	0.026
N23H/A19K	678	352	14	312	0.021
A19H/A19S	868	377	13	478	0.015
A19Y/A19L	941	434	19	488	0.020
A19E/A19K	1052	555	14	484	0.013
N23H/A19E	613	311	11	291	0.018
A19E/N23K	1087	467	17	604	0.016
A19H/A19K	1072	590	14	468	0.013
A19H/N23E	604	306	16	282	0.026
A19K/N23E	650	228	6	416	0.009
A19H/N23K	644	324	8	312	0.012
A19F/A19L	809	315	9	486	0.011
A19H/A19E	1077	460	7	610	0.006
N23F/A19L	428	194	2	232	0.005

fraction dimer observed from the appropriate homodimerization experiment (see Table 2) as a correction factor by multiplying this value by the total number of noncoincident bursts seen in each channel. This represents the maximum homodimer that could be present in the channel, since lowering the effective concentration of the monomers during heterodimerization will yield a decrease in homodimerization that is not observable using this technique. To further characterize the homodimers present in solution, it would be possible to employ a three-color ALEX assay with a label tethered to MTX_2C9, and this method is currently under development in our laboratory. With this corrected interpretation of the burst profile, we can now measure the relative selectivity for the heterodimer. The corrected heterodimerization data and a measure of heterodimer selectivity are shown in Tables 5 and 6.

Two pairs in our analysis, N23H-N23E and N23E-N23K, yield selectivity for the heterodimeric species of 4.1 and 2.7, respectively. Interestingly, the mutant leading to the least stable homodimer (A19S) did not generate selectivity for the heterodimeric species. This is not completely unexpected, however, since although destabilization of the homodimer should deepen the relative energy well for the heterodimer, if the heterodimer interface represents a more destabilized conformation as well, the net energy gains will be unrealized. It appears that creation of a favorable electrostatic pairing is an effective method of generating stabilized heterodimers, but the positioning of the charge is important (N23E-N23K vs. A19E-A19K). As discussed in Chapter 3, Asn23 has more freedom to move around in the interface, and this appears to have a large effect on interface remodeling, as although the most destabilizing mutations are all present

Table 5. Corrected burst data corresponding to homodimeric pairs in the ALEX heterodimerization assay.

Protein Pair (A/B)	Ch 1 Bursts	Fraction A:A	A:A Homodimers	Ch 2 Bursts	Fraction B:B	B:B Homodimers
N23H/N23E	350	0.116	41	476	0.090	43
N23E/N23K	328	0.09	30	557	0.095	53
N23H/A19S	264	0.116	31	259	0.036	9
A19S/A19Q	242	0.036	9	290	0.137	40
N23F/A19Y	216	0.059	13	264	0.059	17
A19Y/A19L	348	0.063	22	422	0.058	24
A19Y/A19S	292	0.063	18	285	0.036	10
N23H/N23K	518	0.116	60	563	0.095	53
A19Y/A19S	494	0.063	31	542	0.036	20

Table 6. Ratio of homodimers to heterodimers and average selectivity for heterodimerization.

Protein Pair (A/B)	A:A	A:B	B:B	Ratio AA:AB:BB	Average Heterodimer Selectivity
N23H/N23E	41	171	43	1.0:4.2:1.1	4.10
N23E/N23K	30	101	53	1.0:3.4:1.8	2.67
N23H/A19S	31	22	9	3.3:2.4:1.0	1.54
A19S/A19Q	9	18	40	1.0:2.1:4.6	1.26
N23F/A19Y	13	17	17	1.0:1.3:1.3	1.18
A19Y/A19L	22	27	24	1.0:1.2:1.1	1.17
A19Y/A19S	18	14	10	1.8:1.4:1.0	1.06
N23H/N23K	60	59	53	1.1:1.1:1.0	1.04
A19Y/A19S	31	24	20	1.6:1.2:1.0	1.00

at the less-tolerant Ala19 position, the most successfully stabilized heterodimers contain the Asn23 mutant. Overall, the results clearly illustrate that the development of a stable heterodimer is indeed a problem of subtlety.

4. Conclusions

In the preceding work, we have described the difficulties associated with optimizing a detection method for chemically induced dimeric systems utilizing the traditionally homodimeric proteins. However, the application of a dual labeling scheme and analysis via alternating laser excitation yields a mildly quantitative estimate of protein stability. In order to obtain a rigorously quantitative estimate of protein cooperativity, the experiment could be broadened in scope to include multiple ratios of dimerizer to protein, and a set of data could be fit to the corresponding model data to yield a more accurate estimate of K_c.

We have shown that stabilization of a heterodimeric CID system can be achieved using a single point mutation at the protein interface. While further refinement of our computational model will greatly aid in the selection of further mutations to test, studies are already underway using doubly-mutated variants of DHFR. While the current work has shown that a subtle approach to interface remodeling is necessary to enhance heterodimer selectivity, the potential of double mutations is an important area to explore.

Developments in the nascent field of protein interface remodeling represent a wide array of practical applications for the study of intracellular signaling, protein-protein interactions, or protein nanostructural assembly. To our knowledge, the preceding work

represents the first example of a stabilized, chemically induced heterodimer. The first steps have been taken toward the development of a biomolecular language for controlled protein self-assembly. As such, this heterodimer will serve as the basis for exploration toward other, more stable heterodimeric pairs as well as development of stable, self-assembled, bivalent protein nanorings.

5. Materials and Methods

5.1. Fusion Protein Generation and Purification

Generation of Fusion Protein Vectors

To generate fusion protein constructs, plasmids bearing the sequence for the fluorescent protein to be inserted were generous gifts from the Dr. Mark Distefano (eGFP and mCherry-CVIA, University of Minnesota Department of Chemistry) or Dr. Stephen Ekker (mRFP, University of Minnesota Department of Genetics, Cell Biology and Development). First, primer sequences were developed to insert a 5' KpnI and 3' XbaI restriction sites via PCR. PCR amplification of the plasmids was performed using the PCR Platinum Supermix kit (Invitrogen) and associated protocol. Once purified, plasmids containing the fluorescent protein and the pPH70D plasmid were digested with the proper restriction enzymes (Promega, Madison, WI), the insert and parent plasmid DNA isolated via gel electrophoresis on a 0.7% agarose gel, and ligated. The ligation reactions were performed at a 3:1 ratio of insert to vector DNA with T4 DNA ligase for 12 hours at 16ºC. The ligase was heat-inactivated and the mixture transformed into DH5α *E. coli*

(Invitrogen). Fluorescent colonies were picked, the plasmid containing the fusion protein sequence purified, and the vector transformed into BL21 *E. coli* for expression.

His-tagged fusion proteins were generated using the QuickChange site-directed mutagenesis kit (Stratagene). The codon insertions (for GGHHHHHH) were inserted 4 codons at a time by splitting the insertion into two separate mutagenesis reactions. Primers for the two insertions are listed below (reverse primers are the complement of the forward primers listed):

BRW/6hisInsS1; 5'-CGA GCT GTA CAA GCA TCA CCA TCA CTG CGT CAT CGC-3'

BRW/6hisInsS2; 5'-C GAG CTG TAC AAG GGG GGG CAT CAC CAT CAC CAT CAC TGC GTC ATC GCA-3'

Fusion Protein Expression and Purification

Initially, fusion proteins were overexpressed in LB media at 37ºC by induction with IPTG at a final concentration of 0.35 mM. Cell lysis and protein purification via MTX-agarose affinity chromatography was performed as described in Chapter 3.

For 1DC-His6-CVIA, the parent vector was transformed into BL21(DE3) Star *E. coli*. Colonies were picked and cultured in 10 mL LB-Amp for at least 8 hours at 37˚C. 10 mL of culture was added to 500 mL LB-Amp and the culture incubated with shaking and aeration at 37˚C for 2.5 hours, and the culture was induced with 0.3 mM IPTG and allowed to grow for another 7 hours. Cells were recovered via centrifugation at 7,500g for 30 min at 4˚C. Pellets were combined and resuspended in 8 mL lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM Imidazole, 1 mM DTT, 1 mg/mL Lysozyme, pH 8.0)

and incubated at room temperature with shaking for 30 min. Cells were lysed via 8 x 30 second rounds of sonication at 4ºC. The lysate was centrifuged at 40,000g for 30 min at 4˚C, and the supernatant loaded directly onto a Ni-NTA column (10 mL Ni-NTA slurry) washed with 100 mL elution buffer (50 mM NaH_2PO_4, 300 mM NaCl, 250 mM Imidazole, 1 mM DTT, pH 8.0) and equilibrated with 100 mL equilibration buffer (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM Imidazole, 1 mM DTT, pH 8.0). The column is washed with 500 mL equilibration buffer, and the protein eluted with 240 mL elution buffer, collected in 8mL fractions.

5.2. ASC-DHFR Generation and Purification

Site-Directed Mutagenesis

To generate mutant *ec*DHFR plasmids, the QuickChange protocol from Strategene was utilized. In short, complementary primer oligonucleotides bearing the mutations of interest are bound to the parent plasmid and PCR cycling achieves exponential generation of the mutated plasmid. Mutated plasmid DNA is recovered from transformed XL1-Blue *E. coli* via the PureLink HiPure Plasmid Miniprep Kit from Invitrogen. Sequencing of the mutated plasmids by the University of Minnesota Microchemical Facility was used to verify the presence of the mutation. Oligonucleotides (reverse sequence is complementary to forward) used to introduce the mutations are listed below:

C85A; 5’-GAA GCC ATC GCG GCG GCT GGT GAC GTA CCA G-3’

C152S; 5’-G CAG AAC TCG CAT AGC TAT AGT TTC GAA ATC CTC GAG C-3’

C162; 5’-C CTC GAG CGT CGT GGA GGA TGC TAA TTA ATT AAT TCA CTG GCC GTC-3’

In the first attempt, all three sets of primers were used in the QuickChange multi-site-directed mutagenesis kit. Several plasmids containing two of the three mutations were used in the regular QuickChange site-directed mutagenesis kit to introduce the final mutation. This plasmid was then used as the parent for the same mutations described in Chapter 3.

Protein Expression and Purification

All ASC-DHFR proteins were expressed and purified identical to the method described in Chapter 3.

5.3. Solid-State Labeling of ASC-DHFR

In order to efficiently label ASC-DHFR, the method of Kim, et al. was used.[310] An aliquot of ASC-DHFR is reduced with 15 mM DTT for 2 hours at 4˚C. Finely ground 90% ammonium sulfate (w/v, 613.7 g/L) is added and the slurry stirred at 4˚C for one hour. The concentration of the protein slurry is estimated via the Bradford assay. 20 nmol of protein slurry is spun down at 13,000g for 5 min at 4˚C. The supernatant is discarded and the pellet gently washed with ice cold P500 buffer containing 90% (w/v) $(NH_4)_2SO_4$. The solution is centrifuged as before, and the wash step repeated twice. The pellet is then dissolved in 400 µL P500 buffer containing a 200 nmol of Alexa Fluor 488 C_5- or 647 C_2-maleimide. The labeling reaction is allowed to proceed for 1 hour before 1 mM DTT is added quench the reaction. The reaction mixture is loaded directly onto a Sephadex G-75 column (GE Biosciences) and the protein eluted at 0.75 mL/min. The protein peak corresponding to monomeric DHFR was collected and degree of labeling quantitated via

A280/A494 or A280/A650 for Alexa488 or -647, respectively. Typical labeling efficiency was ≥ 90%.

5.4. Gel Filtration

Gel filtration samples for dimerization were prepared as a 5 µM final DHFR concentration in P500 buffer (0.5M NaCl, 50 mM KH_2PO_4, 1 mM EDTA, pH 7.0) with 5% (v/v) glycerol. The samples were loaded on to a Sephadex G-75 or G-200 column (GE Biosciences) on a Beckman System Gold HPLC and eluted at 0.5 mL/min with P500 buffer. For homodimerization experiments, 5 µM protein was treated with 5.5 µM MTX_2C9. For heterodimerization experiments, 2.5 µM of each protein was mixed and then treated with 5.5 µM MTX_2C9.

For purification of labeled DHFR, 250 µL of the labeling reaction was directly loaded on the G-75 column and the mixture eluted with P500 at 0.75 mL/min. The peak in the A280/A488 (Alexa488 conjugates) or A280/A594 (Alexa647 conjugates) traces corresponding to the retention time of unlabeled DHFR was collected to yield pure, labeled product.

5.5. FRET Analysis

Alexa488 and -647 ASC-DHFR homo- and heterodimeric samples were prepared analogously to gel filtration samples and diluted to 1 µM in a final volume of 500 µL. First, Alexa488 homodimers were excited at 470 nm and emission scanned from 500 to 800 nm. Next, Alexa647 homodimers were excited at 470 nm and emission scanned to

verify zero emission. Alexa647 homodimers were then excited at 520 nm (peak Alexa488 emission) and emission scanned to verify Alexa647 excitation at the wavelength of Alexa488 emission. Lastly, Alexa488-Alexa647 heterodimers (maximum 50% of the protein mixture) were excited at 470 nm and emission scanned to obtain an emission trace associated with the FRET response. If FRET response were present, an emission peak corresponding to Alexa647 emission would be present, and the response could be quantified by subtracting the integrated area of the Alexa488 homodimer emission trace from that of the heterodimer emission trace. In our investigation, no such difference was observed.

5.6. Fluorescence Correlation Spectroscopy

(Performed in collaboration with Yan Chen)

For monomeric samples, a solution of 5 µM Alexa-labeled protein in P500 buffer with 0.05% Tween-20 was mixed in a final volume of 200 µL. Homodimeric samples consisted of 5 µM protein and 5 µM MTX_2C9 in P500 with 0.05% Tween-20 in a final volume of 200 µL. Heterodimeric samples consisted of equal parts each protein conjugate in a final concentration of 5 µM plus 5 µM MTX_2C9 in a final volume of 200 µL. Samples were diluted to yield appropriate fluorescence intensity (typically to ~50 nM) and the brightness analyzed via confocal fluorescence microscopy at varying powers to determine the appropriate power settings to analyze sample brightness change. The sample was excited at 1000 nm and data collected for 60 s. Emission was detected in two channels using a dichroic mirror to filter signal with wavelength shorter than 580 nm

(green channel) from that with longer wavelength (red channel). Brightness is derived from intensity and fluorescence fluctuation amplitude.

5.7. Alternating Laser Excitation Assays

(Performed in collaboration with Younggyu Kim)

For either homodimeric or heterodimeric complexes, equal parts Alexa488-labeled DHFR and Alexa647-labled DHFR were mixed and 1.1 equivalents to MTX_2C9 added. In homodimeric samples, only one protein variant is used; but in heterodimeric samples, the Alexa488 conjugate is one ASC-DHFR mutant and the Alexa647 conjugate is a different mutant. The sample is diluted to 50 pM prior to analysis in an ALEX setup as pictured in Figure 3. In short, the sample is analyzed via confocal microscopy wherein a small observation volume is excited alternately with lasers at 514 and 648 nm. Resulting fluorescent emission is filtered and divided between two detection channels by a dichroic mirror and fluorescent bursts recorded during an analysis time of at least 5 minutes. The total number of separate and coincident fluorescent bursts is counted, and the number in each channel expressed as a ratio of the total number of events. Mathematical modeling of the dimerization data was performed using the expression for free enzyme concentration, from which the concentrations of all remaining species can be derived:

$$[E] = \frac{-\left(1 + K_{a1}[D_a]\right) + \sqrt{\left(1 + K_{a1}[D_a]\right)^2 + 8K_{a1}K_{a2}K_c[D_a]E_t}}{4K_{a1}K_{a2}K_c[D_a]}$$

BIBLIOGRAPHY

(1) Brent, R. *Nat. Biotechnol.* **2004**, *22*, 1211.
(2) Seeman, N. C. *Chem. Biol.* **2003**, *10*, 1151.
(3) Seeman, N. C. *Mol. Biotechnol.* **2007**, *37*, 246.
(4) Bromley, E. H. C.; Channon, K.; Moutevelis, E.; Woolfson, D. N. *ACS Chem. Biol.* **2008**, *3*, 38.
(5) Ulijn, R. V.; Smith, A. M. *Chem. Soc. Rev.* **2008**, *37*, 664.
(6) Diver, S. T.; Schreiber, S. L. *J. Am. Chem. Soc.* **1997**, *119*, 5106.
(7) de Graffenried, C. L.; Laughlin, S. T.; Kohler, J. J.; Bertozzi, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 16715.
(8) Gendreizig, S.; Kindermann, M.; Johnsson, K. *J. Am. Chem. Soc.* **2003**, *125*, 14970.
(9) Klemm, J. D.; Schreiber, S. L.; Crabtree, G. R. *Annu. Rev. Immunol.* **1998**, *16*, 569.
(10) Boger, D. L.; Goldberg, J. *Bioorg. Med. Chem.* **2001**, *9*, 557.
(11) Cunningham, B. C.; Ultsch, M.; De Vos, A. M.; Mulkerrin, M. G.; Clauser, K. R.; Wells, J. A. *Science* **1991**, *254*, 821.
(12) Carr, P. D.; Gustin, S. E.; Church, A. P.; Murphy, J. M.; Ford, S. C.; Mann, D. A.; Woltring, D. M.; Walker, I.; Ollis, D. L.; Young, I. G. *Cell* **2001**, *104*, 291.
(13) Livnah, O.; Stura, E. A.; Middleton, S. A.; Johnson, D. L.; Jolliffe, L. K.; Wilson, I. A. *Science* **1999**, *283*, 987.
(14) Remy, I.; Wilson, I. A.; Michnick, S. W. *Science* **1999**, *283*, 990.
(15) Greenlund, A. C.; Schreiber, R. D.; Goeddel, D. V.; Pennica, D. *J. Biol. Chem.* **1993**, *268*, 18103.
(16) Pion, E.; Ullmann, G. M.; Amé, J.-C.; Gérard, D.; de Murcia, G.; Bombarda, E. *Biochemistry* **2005**, *44*, 14670.
(17) Wong, I.; Chao, K. L.; Lohman, T. M. *J. Biol. Chem.* **1992**, *267*, 7596.
(18) Linger, B. R.; Kunovska, L.; Kuhn, R. J.; Golden, B. L. *RNA* **2004**, *10*, 128.
(19) Yossi, W.; Braunschweig, A. B.; Wilner, O. I.; Cheglakov, Z.; Willner, I. *Chem. Commun.* **2008**, 4888.
(20) Cheglakov, Z.; Weizmann, Y.; Braunschweig, A. B.; Wilner, O. I.; Willner, I. *Angew. Chem. Int. Ed.* **2008**, *47*, 126.
(21) Ferrari, M. *Nat. Rev. Cancer* **2005**, *5*, 161.
(22) Mapp, A. K.; Aseem Z, A. *ACS Chem. Biol.* **2007**, *2*, 62.
(23) Stafford, R. L.; Dervan, P. B. *J. Am. Chem. Soc.* **2007**, *129*, 14026.
(24) Spencer, D. M.; Wandless, T. J.; Schreiber, S. L.; Crabtree, G. R. *Science* **1993**, *262*, 1019.
(25) Clackson, T. *Curr. Opin. Struct. Biol.* **1998**, *8*, 451.
(26) Amara, J. F.; Clackson, T.; Rivera, V. M.; Guo, T.; Keenan, T.; Natesan, S.; Pollock, R.; Yang, W.; Courage, N. L.; Holt, D. A.; Gilman, M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10618.

(27) Rollins, C. T.; Rivera, V. M.; Woolfson, D. N.; Keenan, T.; Hatada, M.; Adams, S. E.; Andrade, L. J.; Yaeger, D.; van Shravendijk, M. R.; Holt, D. A.; Gilman, M.; Clackson, T. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 7096.
(28) Pollock, R.; Issner, R.; Zoller, K.; Natesan, S.; Rivera, V. M.; Clackson, T. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 13221.
(29) Koide, K.; Finkelstein, J. M.; Ball, Z.; Verdine, G. L. *J. Am. Chem. Soc.* **2001**, *123*, 398.
(30) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33.
(31) Kopytek, S. J.; Standaert, R. F.; Dyer, J. C. D.; Hu, J. C. *Chem. Biol.* **2000**, *7*, 313.
(32) Carlson, J. C. T.; Kanter, A.; Thudappathy, G. R.; Cody, V.; Pineda, P. E.; McIvor, R. S.; Wagner, C. R. *J. Am. Chem. Soc.* **2003**, *125*, 1501.
(33) Ali, J. A.; Jackson, A. P.; Howells, A. J.; Maxwell, A. *Biochemistry* **1993**, *32*, 2717.
(34) Farrar, M. A.; Olson, S. H.; Perlmutter, R. M. *Methods Enzymol.* **2000**, *327*, 421.
(35) Clackson, T. *Gene Ther.* **2000**, *7*, 120.
(36) Belshaw, P. J.; Ho, S. N.; Crabtree, G. R.; Schreiber, S. L. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 4604.
(37) Brown, E. J.; Albers, M. W.; Shin, T. B.; Ichikawa, K.; Keith, C. T.; Lane, W. S.; Schreiber, S. L. *Nature* **1994**, *369*, 756.
(38) Sabatini, D. M.; Erdjument-Bromage, H.; Lui, M.; Tempst, P.; Snyder, S. H. *Cell* **1994**, *78*, 35.
(39) Chen, J.; Zheng, X. F.; Brown, E. J.; Schreiber, S. L. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 4947.
(40) Banaszynski, L. A.; Liu, C. W.; Wandless, T. J. *J. Am. Chem. Soc.* **2005**, *127*, 4715.
(41) Belshaw, P. J.; Schoepfer, J. G.; Liu, K.-Q.; Morrison, K. L.; Schreiber, S. L. *Angew. Chem. Int. Ed.* **1995**, *34*, 2129.
(42) Choi, J.; Chen, J.; Schreiber, S. L.; Clardy, J. *Science* **1996**, *273*, 239.
(43) Liberles, S. D.; Diver, S. T.; Austin, D. J.; Schreiber, S. L. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 7825.
(44) Bayle, J. H.; Grimley, J. S.; Stankunas, K.; Gestwicki, J. E.; Wandless, T. J.; Crabtree, G. R. *Chem. Biol.* **2006**, *13*, 99.
(45) Briesewitz, R.; Ray, G. T.; Wandless, T. J.; Crabtree, G. R. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 1953.
(46) Vogel, K. W.; Briesewitz, R.; Wandless, T. J.; Crabtree, G. R. *Adv. Protein Chem.* **2001**, *56*, 253.
(47) Varshavsky, A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 2094.
(48) Rosen, M. K.; Amos, C. D.; Wandless, T. J. *J. Am. Chem. Soc.* **2000**, *122*, 11979.
(49) Braun, P. D.; Barglow, K. T.; Lin, Y. M.; Akompong, T.; Briesewitz, R.; Ray, G. T.; Haldar, K.; Wandless, T. J. *J. Am. Chem. Soc.* **2003**, *125*, 7575.
(50) Perelson, A. S.; DeLisi, C. *Math. Biosci.* **1998**, *48*, 71.
(51) Mack, E. T.; Perez-Castillejos, R.; Suo, Z.; Whitesides, G. M. *Anal. Chem.* **2008**, *80*, 5550.

(52) Miller, C. P.; Blau, C. A. *Gene Ther.* **2008**, *15*, 759.
(53) Neff, T.; Blau, C. A. *Blood* **2001**, *97*, 2535.
(54) Zhao, Y.; Gonzalez-Garcia, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012.
(55) Koh, J.-T.; Ge, C.; Zhao, M.; Wang, Z.; Krebsbach, P. H.; Zhao, Z.; Franceschi, R. T. *Mol. Ther.* **2006**, *14*, 684.
(56) Blau, C. A.; Peterson, K. R.; Drachman, J. G.; Spencer, D. M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 3076.
(57) Abida, W. M.; Carter, B. T.; Althoff, E. A.; Lin, H.; Cornish, V. W. *Chembiochem* **2002**, *3*, 887.
(58) Zhao, H. X. *Curr. Med. Chem.* **2004**, *11*, 539.
(59) Stevens, K. R.; Rolle, M. W.; Minami, E.; Ueno, S.; Nourse, M. B.; Virag, J.; Reinecke, H.; Murry, C. E. *Human Gene Ther.* **2007**, *18*, 401.
(60) Whitney, M. L.; Otto, K. G.; Blau, C. A.; Reinecke, H.; Murry, C. E. *J. Biol. Chem.* **2001**, *276*, 41191.
(61) Carlotti, F.; Zaldumbide, A.; Martin, P.; Boulukos, K. E.; Hoeben, R. C.; Pognonec, P. *Cancer Gene Ther.* **2005**, *12*, 627.
(62) Duan, Y.; Wu, C.; Choudhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A. *J. Comp. Chem.* **2003**, *24*, 1999.
(63) Rivera, V. M.; Clackson, T.; Natesan, S.; Pollock, R.; Amara, J. F.; Keenan, T.; Magari, S. R.; Phillips, T.; Courage, N. L.; Cerasoli, F., Jr.; Holt, D. A.; Gilman, M. *Nat. Med.* **1996**, *2*, 1028.
(64) Athavankar, S.; Peterson, B. R. *Chem. Biol.* **2003**, *10*, 1245.
(65) Quintarelli, C.; Vera, J. F.; Savoldo, B.; Giordano Attanese, G. M. P.; Pule, M.; Foster, A. E.; Heslop, H. E.; Rooney, C. M.; Brenner, M. K.; Dotti, G. *Blood* **2007**, *110*, 2793.
(66) de Felipe, K. S.; Carter, B. T.; Althoff, E. A.; Cornish, V. W. *Biochemistry* **2004**, *43*, 10353.
(67) Senner, V.; Sotoodeh, A.; Paulus, W. *Neurochem. Res.* **2001**, *26*, 521.
(68) Xu, Z. L.; Mizuguchi, H.; Mayumi, T.; Hayakawa, T. *Gene* **2003**, *309*, 145.
(69) Rivera, V. M.; Gao, G. P.; Grant, R. L.; Schnell, M. A.; Zoltick, P. J.; Rozamus, L. W.; Clackson, T.; Wilson, J. M. *Blood* **2005**, *105*, 1424.
(70) Banaszynski, L.; Chen, L.; Maynard-Smith, L.; Ooi, A.; Wandless, T. *Cell* **2006**, *126*, 995.
(71) Chong, H.; Ruchatz, A.; Clackson, T.; Rivera, V. M.; Vile, R. G. *Mol. Ther.* **2002**, *5*, 195.
(72) Thomis, D. C.; Marktel, S.; Bonini, C.; Traversari, C.; Gilman, M.; Bordignon, C.; Clackson, T. *Blood* **2001**, *97*, 1249.
(73) Berger, C.; Blau, C. A.; Huang, M. L.; Iuliucci, J. D.; Dalgarno, D. C.; Gaschet, J.; Heimfeld, S.; Clackson, T.; Riddell, S. R. *Blood* **2004**, *103*, 1261.
(74) Kiem, H. P.; Sellers, S.; Thomasson, B.; Morris, J. C.; Tisdale, J. F.; Horn, P. A.; Hematti, P.; Adler, R.; Kuramoto, K.; Calmels, B.; Bonifacino, A.; Hu, J.; von Kalle, C.; Schmidt, M.; Sorrentino, B.; Nienhuis, A.; Blau, C. A.; Andrews, R. G.; Donahue, R. E.; Dunbar, C. E. *Mol. Ther.* **2004**, *9*, 389.

(75) Kohn, D. B.; Sadelain, M.; Dunbar, C.; Bodine, D.; Kiem, H. P.; Candotti, F.; Tisdale, J.; Riviere, I.; Blau, C. A.; Richard, R. E.; Sorrentino, B.; Nolta, J.; Malech, H.; Brenner, M.; Cornetta, K.; Cavagnaro, J.; High, K.; Glorioso, J. *Mol. Ther.* **2003**, *8*, 180.
(76) Neff, T.; Horn, P. A.; Valli, V. E.; Gown, A. M.; Wardwell, S.; Wood, B. L.; von Kalle, C.; Schmidt, M.; Peterson, L. J.; Morris, J. C.; Richard, R. E.; Clackson, T.; Kiem, H. P.; Blau, C. A. *Blood* **2002**, *100*, 2026.
(77) Iuliucci, J. D.; Oliver, S. D.; Morley, S.; Ward, C.; Ward, J.; Dalgarno, D.; Clackson, T.; Berger, H. J. *J. Clin. Pharmacol.* **2001**, *41*, 870.
(78) Richard, R. E.; De Claro, R. A.; Yan, J.; Chien, S.; Von Recum, H.; Morris, J.; Kiem, H. P.; Dalgarno, D. C.; Heimfeld, S.; Clackson, T.; Andrews, R.; Blau, C. A. *Mol. Ther.* **2004**, *10*, 730.
(79) Nagasawa, Y.; Wood, B. L.; Wang, L.; Lintmaer, I.; Guo, W.; Papayannopoulou, T.; Harkey, M. A.; Nourigat, C.; Blau, C. A. *Stem Cells* **2006**, *24*, 908.
(80) Gestwicki, J. E.; Crabtree, G. R.; Graef, I. A. *Science* **2004**, *306*, 865.
(81) Cochran, A. G. *Chem. Biol.* **2000**, *7*, R85.
(82) Soloman, D.; Kitov, P. I.; Paszkiewicz, E.; Grant, G. A.; Sadowska, J. M.; Bundle, D. R. *Org. Lett.* **2005**, *7*, 4369.
(83) Kitov, P. I.; Lipinski, T.; Paszkiewicz, E.; Soloman, D.; Sadowska, J. M.; Grant, G. A.; Mulvey, G. L.; Kitova, E. N.; Klassen, J. S.; Armstrong, G. D.; Bundle, D. R. *Angew. Chem. Int. Ed.* **2008**, *47*, 672.
(84) Whitesides, G. M.; Mathias, J. P.; Seto, C. T. *Science* **1991**, *254*, 1312.
(85) Whitesides, G. M.; Boncheva, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4769.
(86) Sarikaya, M.; Tamerler, C.; Jen, A. K.; Schulten, K.; Baneyx, F. *Nat. Mater.* **2003**, *2*, 577.
(87) Clark, J.; Singer, E. M.; Korns, D. R.; Smith, S. S. *Biotechniques* **2004**, *36*, 992.
(88) Seeman, N. C. *Trends Biotechnol.* **1999**, *17*, 437.
(89) Seeman, N. C. *Nano. Lett.* **2001**, *1*, 22.
(90) Giesecke, A. V.; Fang, R.; Juong, J. K. *Mol. Syst. Biol.* **2006**, *2*, 1.
(91) Yeates, T. O.; Padilla, J. E. *Curr. Opin. Struct. Biol.* **2002**, *12*, 464.
(92) Padilla, J. E.; Colovos, C.; Yeates, T. O. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 2217.
(93) Pandya, M. J.; Spooner, G. M.; Sunde, M.; Thorpe, J. R.; Rodger, A.; Woolfson, D. N. *Biochemistry* **2000**, *39*, 8728.
(94) Ogihara, N. L.; Ghirlanda, G.; Bryson, J. W.; Gingery, M.; Degrado, W. F. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 1404.
(95) Ryadnov, M. G.; Woolfson, D. N. *Angew. Chem. Int. Ed.* **2003**, *42*, 3021.
(96) Ryadnov, M. G.; Woolfson, D. N. *J. Am. Chem. Soc.* **2004**, *126*, 7454.
(97) Rele, S.; Song, Y.; Apkarian, R. P.; Qu, Z.; Conticello, V. P.; Chaikof, E. L. *J. Am. Chem. Soc.* **2007**, *129*, 14780.
(98) Dotan, N.; Arad, D.; Frolow, F.; Freeman, A. *Angew. Chem. Int. Ed.* **1999**, *38*, 2363.
(99) Ringler, P.; Schulz, G. E. *Science* **2003**, *302*, 106.

(100) Ballister, E. R.; Lai, A. H.; Zuckerman, R. N.; Cheng, Y.; Mougous, J. D. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 3733.
(101) O'Reilly, M. K.; Collins, B. E.; Han, S.; Liao, L.; Rillahan, C.; Kitov, P. I.; Bundle, D. R.; Paulson, J. C. *J. Am. Chem. Soc.* **2008**, *130*, 7736.
(102) Rao, J. H.; Lahiri, J.; Weis, R. M.; Whitesides, G. M. *J. Am. Chem. Soc.* **2000**, *122*, 2698.
(103) Rao, J. H.; Lahiri, J.; Isaacs, L.; Weis, R. M.; Whitesides, G. M. *Science* **1998**, *280*, 708.
(104) Rao, J. H.; Whitesides, G. M. *J. Am. Chem. Soc.* **1997**, *119*, 10286.
(105) Bilgicer, B.; Moustakas, D. T.; Whitesides, G. M. *J. Am. Chem. Soc.* **2007**, *129*, 3722.
(106) Carlson, J. C. T.; Jena, S. S.; Flenniken, M.; Chou, T.-F.; Siegel, R. A.; Wagner, C. R. *J. Am. Chem. Soc.* **2006**, *128*, 7630.
(107) Li, Q.; Hapka, D.; Chen, H.; Vallera, D. A.; Wagner, C. R. *Angew. Chem. Int. Ed.* **2008**, *47*, 10179.
(108) Chou, T.-F.; So, C.; White, B. R.; Carlson, J. C. T.; Sarikaya, M.; Wagner, C. R. *ACS Nano* **2008**, *2*, 2519.
(109) McCammon, J. A.; Gelin, J. B.; Karplus, M. *Nature* **1977**, *267*, 585.
(110) Gohlke, H.; Klebe, G. *Agnew. Chem. Int. Ed.* **2002**, *41*, 2644.
(111) Hessler, G.; Klabunde, T. *ChemBioChem* **2002**, *3*, 928.
(112) Lugovsky, A. A.; Degterev, A. I.; Fahmy, A. F.; Zhou, P.; Gross, J. D.; Yuan, J.; Wagner, G. *J. Am. Chem. Soc.* **2002**, *124*, 1234.
(113) Perez, J. J.; Concho, F.; Llorens, O. *Curr. Med. Chem.* **2002**, *24*, 2209.
(114) Rao, G. S.; Bhatnagar, S.; Ahuja, V. *J. Biomol. Struct. Dyn.* **2002**, *20*, 31.
(115) Kohn, W.; Becke, A. D.; Parr, R. G. *J. Phys. Chem.* **1996**, *100*, 12974.
(116) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. *J. Chem. Phys.* **1987**, *87*, 5968.
(117) Bowen, J. P.; Allinger, N. L. *Rev. Comp. Chem.* **1991**, *2*, 81.
(118) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. J.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
(119) Dinur, U.; Hagler, A. T. *Rev. Comp. Chem.* **1991**, *2*, 99.
(120) MacKerell Jr., A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr., R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher III, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.
(121) Petersson, I.; Liljefors, T. *Rev. Comp. Chem.* **1996**, *9*.
(122) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(123) Pople, J. A.; Beveridge, D. L. *Approximate Molecular Orbital Methods*; McGraw-Hill, New York, 1970.
(124) Stewart, J. J. P. *J. Comp. Chem.* **1989**, *10*, 209.
(125) Gao, J.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467.

(126) Lengauer, T.; Rarey, M. *Curr. Opin. Struct. Biol.* **1996**, *6*, 402.
(127) Shoichet, B. K.; Kuntz, I. D.; Bodian, D. L. *J. Comp. Chem.* **2004**, *13*, 380.
(128) Jain, A. N. *Curr. Protein Pept. Sci.* **2006**, *7*, 407.
(129) Metropolis, N.; Ulam, S. *J. Am. Stat. Assoc.* **1949**, *44*, 335.
(130) Alder, B. J.; Wainwright, T. E. *J. Chem. Phys.* **1959**, *31*, 459.
(131) Rahman, A. *Phys. Rev.* **1964**, *136*, A405.
(132) Zwanzig, R. W. J. *J. Chem. Phys.* **1954**, *22*, 1420.
(133) Torrie, G. M.; Valleau, J. P. *J. Comp. Phys.* **1977**, *23*, 187.
(134) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L. *Accts. Chem. Res.* **2000**, *33*, 889.
(135) Srinivasan, J.; Cheatham III, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. *J. Am. Chem. Soc.* **1998**, *120*, 9401.
(136) Lybrand, T. P.; McCammon, J. A.; Wipff, G. *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 833.
(137) Bash, P. A.; Singh, U. C.; Brown, F. K.; Langridge, R.; Kollman, P. A. *Science* **1987**, *235*, 574.
(138) Jorgensen, W. L.; Gao, J.; Ravimohan, C. *J. Phys. Chem.* **1985**, *89*, 3470.
(139) Woo, H.-J.; Roux, B. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6825.
(140) Kuhn, B.; Kollman, P. A. *J. Am. Chem. Soc.* **2000**, *122*, 3909.
(141) Lee, T.; Kollman, P. A. *J. Am. Chem. Soc.* **2000**, *122*, 4385.
(142) Gohlke, H.; Kiel, C.; Case, D. A. *J. Mol. Bio.* **2003**, *330*, 891.
(143) Nesmelova, I. V.; Sham, Y.; Gao, J.; Mayo, K. H. *J. Biol. Chem.* **2008**, *283*, 24155.
(144) Shahripour, A. B.; Plummer, M. S.; Lunny, E. A.; Albrecht, H. P.; Hays, S. J.; Kostlan, C. R.; Sawyer, T. K.; Walker, N. P. C.; Brady, K. D.; Allen, H. J.; Talanian, R. V.; Wong, W. W.; Humblet, C. *Bioog. Med. Chem.* **2002**, *10*, 31.
(145) Anderson, A. *Chem. Biol.* **2003**, *10*, 787.
(146) Berlicki, L.; Kafarski, P. *Curr. Org. Chem.* **2005**, *9*, 1829.
(147) Ikejiri, M.; Bernardo, M. M.; Meroueh, S. O.; Brown, S.; Chang, M.; Fridman, R.; Mobashery, S. *J. Org. Chem.* **2005**, *70*, 5709.
(148) Liu, J.; Zhang, Z.; Tan, X.; Hol, W. G. J.; Verlinde, C. L. M. J.; Fan, E. *J. Am. Chem. Soc.* **2005**, *127*, 2044.
(149) Armstrong, K. A.; Tidor, B.; Cheng, A. C. *J. Med. Chem.* **2006**, *49*, 2470.
(150) Ortiz, A. R.; Gomez-Puentas, P.; Leo-Macias, A.; Lopez-Romero, P.; Lopez-Vinas, E.; Morreale, A.; Murcia, M.; Wang, K. *Curr. Top. Med. Chem.* **2006**, *6*, 41.
(151) Rao, G. S.; Ramachandran, M. V.; Bajaj, J. S. *J. Biomol. Struct. Dyn.* **2006**, *23*, 377.
(152) Ragno, R.; Simeoni, S.; Castellano, S.; Vicidomini, C.; Mai, A.; Caroli, A.; Tramontano, A.; Bonaccini, C.; Trojer, P.; Bauer, I.; Brosch, G.; Sbardella, G. *J. Med. Chem.* **2007**, *50*, 1241.
(153) Strockbine, B.; Rizzo, R. C. *Proteins: Structure, Function, and Genetics* **2007**, *67*, 630.
(154) Cornell, W. D.; Cieplak, P. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

(155) Boyd, D. B.; Snoddy, J. D.; Lin, H. S. *J. Comp. Chem.* **1991**, *12*, 635.
(156) Jalaie, M.; Lipkowitz, K. *Rev. Comp. Chem.* **1999**, *14*.
(157) Khandelwal, A.; Lukacova, V.; Comez, D.; Kroll, D. M.; Raha, S.; Balaz, S. *J. Med. Chem.* **2005**, *48*, 5437.
(158) Mulholland, A. J. *Theor. Comp. Chem.* **2001**, *9*, 597.
(159) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185.
(160) Richards, N. J. *Molecular Orbital Calculations for Biological Systems*; Oxford University Press: New York, 1998.
(161) McKercher, S. R.; Lombardo, C. R.; Bobkov, A.; Jia, X.; Assa-Muut, N. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 511.
(162) Elcock, A. H.; McCammon, J. A. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 2990.
(163) Prazulj, N.; Wigle, D. A.; Jurisica, I. *Bioinformatics* **2004**, *20*, 340.
(164) Gao, Y.; Wang, R.; Lai, L. *J. Mol. Model.* **2004**, *10*, 44.
(165) Gohlke, H.; Case, D. A. *J. Comp. Chem.* **2004**, *25*, 238.
(166) Ababou, A.; van der Vaant, A.; Gogonea, V.; Merz, K. M. J. *Biophys. Chem.* **2007**, *125*, 221.
(167) Aslan, F. M.; Yu, Y.; Vajda, S.; Mohr, S. C.; Cantor, C. R. *J. Biotech.* **2007**, *128*, 213.
(168) Sutton, P. A.; Cody, V.; Smith, D. *J. Am. Chem. Soc.* **1986**, *108*, 4155.
(169) Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. *J. Biol. Chem.* **1982**, *257*, 13650.
(170) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput-Aid Mol. Des.* **1995**, *9*, 87.
(171) Cizek, J. *J. Chem. Phys.* **1966**, *45*, 4256.
(172) Purvis, G. D.; Bartlett, R. J. *J. Chem. Phys.* **1982**, *76*, 1910.
(173) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.
(174) Chamberlin, A. C.; Pu, J.; Kelly, C. P.; Thompson, J. D.; Xidos, J. D.; Li, J.; Zhu, T.; Hawkins, G. D.; Chuang, Y.-Y.; Fast, P. L.; Lynch, B. J.; Liotard, D. A.; Rinaldi, D.; Gao, J.; Cramer, C. J.; Truhlar, D. G.
(175) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833.
(176) Dauber-Osguthorpe, P.; Roberts, V. A.; Osguthorpe, D. J.; Wolff, J.; Genest, M.; Hagler, A. T. *Proteins: Structure, Function, and Genetics* **1988**, *4*, 31.
(177) Clark, M.; Cramer III, R. D.; van Opdenhosch, N. *J. Comp. Chem.* **1989**, *10*, 982.
(178) Halgren, T. A. *J. Comp. Chem.* **1996**, *17*, 490.
(179) Damm, W.; Frontera, A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Comp. Chem.* **1997**, *18*, 1955.
(180) Clackson, T.; Yang, W.; Rozamus, L. W.; Hatada, M.; Amara, J. F.; Rollins, C. T.; Stevenson, L. F.; Magari, S. R.; Wood, S. A.; Courage, N. L.; Lu, X.; Cerasoli, F.; Gilman, M.; Holt, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 10437.
(181) Brom, J. M.; Schmitz, B. J.; Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*.
(182) Chambers, C. C.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *1000*, 16385.
(183) Li, J.; Zhu, J.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1998**, *102*, 1820.

(184) Helgaker, T.; Gauss, J.; Jorgensen, P.; Olsen, J. *J. Chem. Phys.* **1997**, *106*, 6430.
(185) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Phys. A* **2003**, *107*, 1384.
(186) Wong, M. W.; Radom, L. *J. Phys. Chem. A* **1998**, *102*, 2237.
(187) Byrd, E. F. C.; Sherrill, C. D.; Head-Gordon, M. *J. Phys. Chem. A* **2001**, *105*, 9736.
(188) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
(189) Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040.
(190) Perdew, J. P.; Burke, K.; Wang, Y. *Phys. Rev. B* **1996**, *54*, 16533.
(191) Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664.
(192) Chamberlin, A. C.; Kelly, C. P.; Thompson, J. D.; Xidos, J. D.; Li, J.; Hawkins, W. P. D.; Zhu, T.; Rinaldi, D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G.; Frisch, M. J., 2006.
(193) Zhao, Y.; Truhlar, D. G., 2006.
(194) Mulliken, R. S. *J. Chem. Phys.* **1935**, *3*, 564.
(195) Lowdin, P. O. *J. Chem. Phys.* **1950**, *18*, 365.
(196) Mulliken, R. S. *J. Chem. Phys.* **1962**, *36*, 3428.
(197) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comp. Chem.* **1986**, *7*, 230.
(198) Besler, B. H.; Merz, K. M.; Kollman, P. A. *J. Comp. Chem.* **1990**, *11*, 431.
(199) Breneman, C. M.; Wiberg, K. B. *J. Comp. Chem.* **1990**, *11*, 361.
(200) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269.
(201) Francl, M. M.; Carey, C.; Chirlian, L. E.; Gange, D. M. *J. Comp. Chem.* **1996**, *17*, 367.
(202) Singh, U. C.; Kollman, P. A. *J. Comp. Chem.* **1984**, *5*, 129.
(203) Chirlian, L. E.; Francl, M. M. *J. Comp. Chem.* **1987**, *1987*, 894.
(204) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comp. Chem.* **2002**, *23*, 1601.
(205) Hawkins, G. D.; Giesen, D.; Lynch, G.; Chambers, C.; Rossi, I.; Storer, J.; Li, J.; Zhu, T.; Thompson, J.; Winget, P.; Lynch, B. J.; Rinaldi, D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. **2004**.
(206) Maple, J. R.; Hwang, M. J.; Stockfish, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. *J. Comp. Chem.* **1994**, *15*, 162.
(207) Case, D. A.; Darden, T. A.; Cheatham III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Peasani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. University of California, San Francisco, 2006.
(208) Winget, P.; Thompson, J. D.; Xidos, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2002**, *106*.
(209) Kalinowski, J. A.; Lesyng, B.; Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*.
(210) Momany, F. J. *J. Phys. Chem.* **1978**, *85*, 592.

(211) Dinur, U.; Hagler, A. T. *J. Am. Chem. Soc.* **1989**, *111*, 5149.
(212) Hobza, P.; Kabelec, M.; Sponer, J.; Mejzlik, P.; Vondrasek, J. *J. Comp. Chem.* **1996**, *118*, 1136.
(213) McDonald, D. Q.; Still, W. C. *Tetrahedron Lett.* **1992**, *33*, 7747.
(214) Chakravorty, S.; Reynolds, C. H. *J. Mol. Graph. Model.* **1999**, *17*, 315.
(215) Halgren, T. A. *J. Comp. Chem.* **1999**, *20*, 730.
(216) Reilly, M. T.; Cunningham, K. A.; Natarajan, A. *Neuropsychopharmacology* **2009**, *34*, 247.
(217) Salwinski, L.; Miller, C. S.; Smith, A. J.; Pettit, F. K.; Bowie, J. U.; Eisenberg, D. *Nucleic Acids Res.* **2004**, *32*, D449.
(218) Keskin, O.; Gursoy, A.; Ma, B.; Nussinov, R. *Chem. Rev.* **2008**, *108*, 1225.
(219) Glaser, F.; Steinberg, D. M.; Vakser, I. A.; Ben-Tal, N. *Proteins: Structure, Function, and Genetics* **2001**, *43*, 89.
(220) Ofran, Y.; Rost, B. *J. Mol. Biol.* **2003**, *325*, 377.
(221) Sheinerman, F. B.; Norel, R.; Honig, B. *J. Mol. Biol.* **2002**, *318*, 161.
(222) Hendsch, Z. S.; Tidor, B. *Protein Sci.* **1999**, *8*, 1381.
(223) Lo Conte, L.; Chothia, C.; Janin, J. *J. Mol. Biol.* **1999**, *285*, 2177.
(224) Kobe, B.; Kajava, A. V. *Curr. Opin. Struct. Biol.* **2001**, *11*, 725.
(225) Young, L.; Jernigan, R. L.; Covell, D. G. *Protein Sci.* **1994**, *3*, 717.
(226) Marcotte, E. M.; Pellegrini, M.; Ng, H. L.; Rice, D. W.; Yeates, T. O.; Eisenberg, D. *Science* **1999**, *285*, 751.
(227) Uetz, P.; Goit, L.; Cagney, G.; Mansfield, T. A.; Judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S.; Rothberg, J. M. *Nature* **1999**, *403*, 623.
(228) Walls, P. H.; Sternberg, M. J. E. *J. Mol. Biol.* **1992**, *228*, 277.
(229) Jones, S.; Thornton, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13.
(230) Vaskar, I. A.; Aflalo, C. *Proteins* **1994**, *20*, 320.
(231) Crowley, P. B.; Otting, G.; Schlard-Ridley, B. G.; Canters, G. W.; Ubbink, M. *J. Am. Chem. Soc.* **2001**, *123*, 10444.
(232) Tsai, C. J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. *Protein Sci.* **1997**, *6*, 53.
(233) Reichmann, D.; Rahat, O.; Cohen, M.; Neuvirth, H.; Schreiber, G. *Curr. Opin. Struct. Biol.* **2007**, *17*, 67.
(234) Clackson, T.; Wells, J. A. *Science* **1995**, *267*, 383.
(235) Bogan, A. A.; Thorn, K. S. *J. Mol. Biol.* **1998**, *280*, 1.
(236) Thorn, K. S.; Bogan, A. A. *Bioinformatics* **2001**, *17*, 284.
(237) Massova, I.; Kollman, P. A. *J. Am. Chem. Soc.* **1999**, *121*, 8133.
(238) Lowman, H. B.; Cunningham, B. C.; Wells, J. A. *J. Biol. Chem.* **1991**, *266*, 10982.
(239) Cunningham, B. C.; Wells, J. A. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 3407.
(240) Nohaile, M. J.; Hendsch, Z. S.; Tidor, B.; Sauer, R. T. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 3109.
(241) Hendsch, Z. S.; Nohaile, M. J.; Sauer, R. T.; Tidor, B. *J. Am. Chem. Soc.* **2001**, *123*, 1264.

(242) Hendsch, Z. S.; Jonsson, T.; Sauer, R. T.; Tidor, B. *Biochemistry* **1996**, *35*, 7621.
(243) Arkin, M. R.; Wells, J. A. *Nat. Rev. Drug Discovery* **2004**, *3*, 301.
(244) Sillerud, L. O.; Larson, R. S. *Curr. Protein Pept. Sci.* **2005**, *6*, 151.
(245) Morell, M.; Espargaro, A.; Aviles, F. X.; Ventura, S. *Proteomics* **2007**, *7*, 1023.
(246) Tang, C.; Iwahara, J.; Clore, G. M. *Nature* **2006**, *444*, 383.
(247) Sawaya, M. R.; Kraut, J. *Biochemistry* **1997**, *36*, 586.
(248) Nakamura, T.; Iwakura, M. *J. Biol. Chem.* **1999**, *274*, 19041.
(249) Appleman, J. R.; Howell, E. E.; Kraut, J.; Kuhl, M.; Blakley, R. L. *J. Biol. Chem.* **1988**, *263*, 9187.
(250) Tsai, C. J.; Taylor, R.; Chothia, C.; Gerstein, M. *J. Mol. Biol.* **1999**, *290*, 253.
(251) Carugo, O. *In Silico Biol.* **2003**, *3*, 417.
(252) Brooijmans, N.; Sharp, K. A.; Kuntz, I. D. *Proteins: Structure, Function, and Genetics* **2002**, *48*, 645.
(253) O'Shea, E. K.; Rutkowski, R.; Kim, P. S. *Cell* **1992**, *68*, 699.
(254) O'Shea, E. K.; Rutkowski, R.; Stafford, W. F. I.; Kim, P. S. *Science* **1989**, *245*, 646.
(255) Blakley, R. L. *Nature* **1960**, *188*, 231.
(256) Taira, K.; Benkovic, S. J. *J. Med. Chem.* **1988**, *31*, 129.
(257) Seeger, D. R.; Cosulich, D. B.; Smith, J. M.; Hultquist, M. E. *J. Am. Chem. Soc.* **1949**, *71*, 1751.
(258) Rosowsky, A.; Forsch, R. A.; Freisheim, J. H.; Galivan, J.; Wick, M. J. *J. Med. Chem.* **1984**, *1984*, 888.
(259) Aqvist, J.; Luzhkov, V. B.; Brandsdal, B. O. *Accts. Chem. Res.* **2002**, *35*, 358.
(260) Hermans, J.; Wang, L. *J. Am. Chem. Soc.* **1997**, *119*, 2707.
(261) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047.
(262) McCammon, J. A. *Curr. Opin. Struct. Biol.* **1998**, *8*, 245.
(263) Jorgensen, W. L. *Science* **2004**, *303*, 1813.
(264) Fukunishi, Y.; Mikami, Y.; Nakamura, H. *J. Phys. Chem. B.* **2003**, *107*, 13201.
(265) Izrailev, S.; Stepaniants, S.; Balsera, M.; Oono, Y.; Schulten, K. *Biophys. J.* **1997**, *72*, 1568.
(266) Massova, I.; Kollman, P. A. *Perspect. Drug Discovery* **2000**, *18*, 113.
(267) Gilson, M. K.; Rashin, A.; Fine, R.; Honig, B. *J. Mol. Bio.* **1985**, *184*, 503.
(268) Lee, M. S.; Olson, M. A. *Biophys. J.* **2006**, *90*, 864.
(269) Rastelli, G.; Del Rio, A.; Degliesposti, G.; Sgobba, M. *J. Comp. Chem.* **2009**, *Early View - Published Online June 30, 2009*.
(270) Levitt, M.; Sander, C.; Stern, P. *Int. J. Quantum Chem.* **1983**, *10*, 181.
(271) Brooks, C. A.; Brunger, A.; Karplus, M. *Biopolymers* **1985**, *24*, 843.
(272) Simonson, T.; Perahia, D. *Biophys. J.* **1992**, *61*, 410.
(273) Luo, H.; Sharp, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 10399.
(274) Case, D. A.; Darden, T. A.; Cheatham III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.;

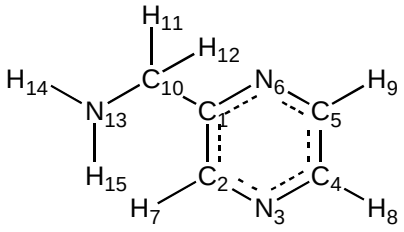
Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. University of California, San Francisco, 2008.
(275) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011.
(276) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824.
(277) Tsui, V.; Case, D. A. *Biopolymers* **2001**, *56*, 257.
(278) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 265.
(279) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712.
(280) Weiser, J.; Shenkin, P. S.; Still, W. C. *J. Comput. Chem.* **1999**, *20*, 217.
(281) Loncharich, R. J.; Brooks, B. R. *Proteins: Structure, Function, and Genetics* **1989**, *6*, 32.
(282) Brooks, C. L. I.; Pettit, B. M.; Karplus, M. *J. Chem. Phys.* **1985**, *83*, 5897.
(283) Schreiber, H.; Steinhauser, O. *Biochemistry* **1992**, *31*, 5856.
(284) Steinbach, P. J.; Brooks, B. R. *J. Comput. Chem.* **1994**, *15*, 667.
(285) Ewald, P. *Ann. Phys.* **1921**, *64*, 253.
(286) Schnell, J. R.; Dyson, H. J.; Wright, P. E. *Ann. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 119.
(287) Feig, M.; Onufriev, A.; M.S., L.; Im, W.; Case, D. A.; III, C. L. B. *J. Comput. Chem.* **2004**, *25*, 265.
(288) Austin, D. J.; Crabtree, G. R.; Schreiber, S. L. *Chem. Biol.* **1994**, *1*, 131.
(289) Belshaw, P. J.; Ho, S. N.; Crabtree, G. R.; Schreiber, S. L. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 4604.
(290) Lin, H.; Abida, W. M.; Sauer, R. T.; Cornish, V. W. *J. Am. Chem. Soc.* **2000**, *122*, 4247.
(291) Daniels, D. J.; Lenard, N. R.; Etienne, C. L.; Law, P. Y.; Roerig, S. C.; Portoghese, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 19208.
(292) Corson, T. W.; Aberle, N.; Crews, C. M. *ACS Chem. Biol.* **2008**, *3*, 11.
(293) Zhu, Z.; Presta, L. G.; Zapata, G.; Carter, P. *Protein Sci.* **1997**, *6*, 781.
(294) Ridgeway, J. B. B.; Presta, L. G.; Carter, P. *Protein Eng.* **1996**, *9*, 617.
(295) Atwell, S.; Ridgeway, J. B. B.; Wells, J. A.; Carter, P. *J. Mol. Bio.* **1997**, *270*, 26.
(296) Ward, W. H. J.; Jones, D. H.; Fersht, A. R. *Biochemistry* **1987**, *26*, 4131.
(297) Grueninger, D.; Treiber, N.; Ziegler, M. O. P.; Koettner, J. W. A.; Schulze, M.-S.; Schulz, G. E. *Science* **2008**, *319*, 206.
(298) Chen, Y.; Muller, J. D. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 3147.
(299) Schwille, P.; Meyer-Almes, F. J.; Rigler, R. *Biophys. J.* **1997**, *72*, 1878.
(300) Elson, E. L.; Magde, D. *Biopolymers* **1974**, *13*, 1.
(301) Chen, Y.; Muller, J. D.; Ruan, Q.; Gratton, E. *Biophys. J.* **2002**, *82*, 133.
(302) Orte, A.; Clarke, R.; Balasubramanian, S.; Klenerman, D. *Anal. Chem.* **2006**, *78*, 7707.
(303) Ren, X.; Li, H.; Clarke, R. W.; Alves, D. A.; Ying, L.; Klenerman, D.; Balasubramanian, S. *J. Am. Chem. Soc.* **2006**, *128*, 4992.

(304) Kapanidis, A. N.; Laurence, T. A.; Lee, N. K.; Margeat, E.; Kong, X.; Weiss, S. *Accts. Chem. Res.* **2005**, *38*, 523.
(305) Shaner, N. C.; Campbell, R. E.; Steinbach, P. A.; Giepmans, B. N. G.; Palmer, A. E.; Tsien, R. Y. *Nat. Biotechnol.* **2004**, *22*, 1567.
(306) Kavoosi, M.; Creagh, A. L.; Kilburn, D. G.; Haynes, C. A. *Biotechnology & Bioengineering* **2007**, *98*, 599.
(307) Duckworth, B. P.; Chen, Y.; Sham, Y.; Muller, J. D.; Taton, T. A.; Distefano, M. D. *Angew. Chem. Int. Ed.* **2007**, *46*, 8819.
(308) Wakagi, T.; Oshima, T.; Imamura, H.; Matsuzawa, H. *Biosci. Biotechnol. Biochem.* **1998**, *62*, 2408.
(309) Sorensen, H. P.; Mortensen, K. K. *J. Biotech.* **2005**, *115*, 113.
(310) Kim, Y.; Ho, S. O.; Gassman, N. R.; Korlann, Y.; Landorf, E. V.; Collart, F. R.; Weiss, S. *Bioconjugate. Chem.* **2008**, *19*, 786.
(311) Tamura, A.; Privalov, P. L. *J. Mol. Bio.* **1997**, *273*, 1048.
(312) Wang, W.; Xu, W.-X.; Levy, Y.; Trizac, E.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 5517.

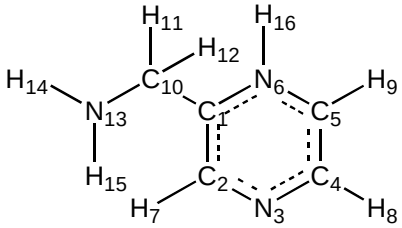
Appendix One

Molecular Numbering Systems and Benchmark Data for MTX Parameter Establishment

Figure S1. The numbering system utilized in our study for 2-AMP and 1*H*-2-AMP.

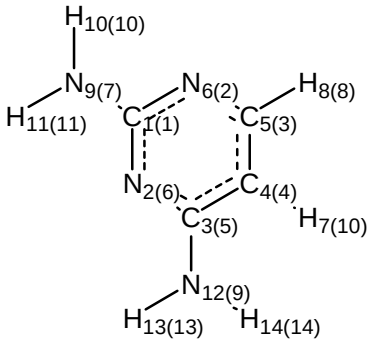


2-(aminomethylpyrazine)

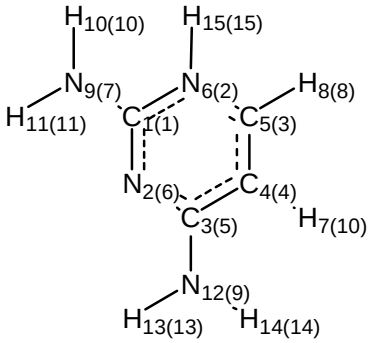


1*H*-2-(aminomethylpyrazine)

Figure S2. The numbering system utilized in our study for 2,4-DAP and 1*H*-2,4-DAP. The numbering schemes in parentheses are for those molecules used in the water binding study, which were constructed after the optimization studies.



2,4-diaminopyrimidine



1*H*-2,4-diaminopyrimidine

Figure S3. The numbering system utilized in our study for methotrexate (MTX) and the methotrexate cation (MTX+).

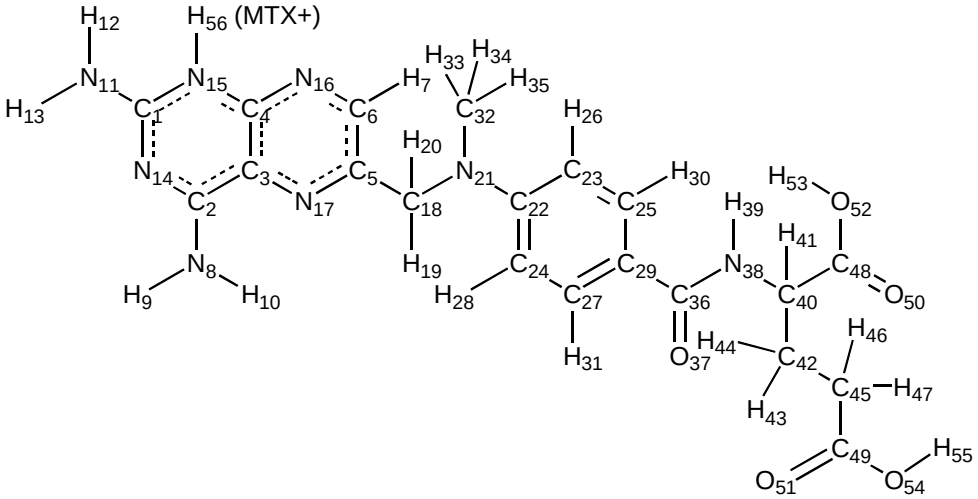


Table S1. Benchmark Calculations of Partial Atomic Charge on 2-AMP and 1*H*-2-AMP in the Gas Phase by CCSD/6-31+G(d,p) and M05-2X/6-31+G(d,p). Note that Mulliken Charges are Shown for CCSD, and CM4 Charges are Shown for M05-2X Calculations.

	2-AMP		1*H*-2-AMP	
Atom	CCSD/Mulliken	M05-2X/CM4	CCSD/Mulliken	M05-2X/CM4
C1	0.17	0.11	0.04	0.20
C2	-0.10	0.10	-0.04	0.13
N3	-0.21	-0.33	-0.11	-0.26
C4	-0.10	0.08	-0.06	0.14
C5	0.11	0.08	0.14	0.14
N6	-0.24	-0.32	-0.26	-0.27
H7	0.17	0.08	0.23	0.11
H8	0.17	0.08	0.24	0.12
H9	0.17	0.08	0.25	0.13
C10	-0.33	0.04	-0.17	0.06
H11	0.13	0.05	0.20	0.10
H12	0.12	0.06	0.20	0.10
N13	-0.66	-0.68	-0.78	-0.66
H14	0.31	0.28	0.34	0.31
H15	0.29	0.28	0.34	0.31
H16	N/A*	N/A*	0.45	0.32

*Not present in the molecule

Table S2. Benchmark Calculations of Partial Atomic Charge on 2-AMP and 1*H*-2-AMP in the Aqueous Phase by M05-2X/6-31+G(d,p). Note That Both Sets of Data are CM4 Charges.

Atom	2-AMP	1*H*-2-AMP
C1	0.10	0.22
C2	0.09	0.13
N3	-0.36	-0.31
C4	0.09	0.12
C5	0.09	0.16
N6	-0.35	-0.25
H7	0.08	0.11
H8	0.10	0.11
H9	0.10	0.14
C10	0.05	0.07
H11	0.06	0.11
H12	0.06	0.11
N13	-0.69	-0.66
H14	0.28	0.31
H15	0.30	0.31
H16	N/A*	0.33

Table S3. Benchmark Calculations of Bond Lengths in 2-AMP and 1*H*-2-AMP in the Gas Phase by CCSD/6-31+G(d,p) and M05-2X/6-31+G(d,p).

	2-AMP		1*H*-2-AMP	
Bond	CCSD	M05-2X	CCSD	M05-2X
C1-C2	1.40	1.40	1.40	1.40
C1-N6	1.34	1.33	1.34	1.33
C1-C10	1.51	1.51	1.52	1.52
C2-N3	1.34	1.33	1.33	1.33
C2-H7	1.08	1.08	1.08	1.08
N3-C4	1.34	1.33	1.35	1.34
C4-C5	1.40	1.39	1.39	1.38
C4-H8	1.08	1.08	1.08	1.08
C5-N6	1.34	1.33	1.35	1.34
C5-H9	1.08	1.08	1.08	1.08
N6-H16	N/A*	N/A*	1.03	1.03
C10-H11	1.10	1.10	1.09	1.09
C10-H12	1.09	1.09	1.09	1.09
C10-N13	1.47	1.46	1.46	1.45
N13-H14	1.02	1.01	1.01	1.01
N13-H15	1.01	1.01	1.01	1.01

Table S4. Benchmark Calculations of Bond Lengths in 2-AMP and 1*H*-2-AMP in the Aqueous Phase by M05-2X/6-31+G(d,p).

Bond	2-AMP	1*H*-2-AMP
C1-C2	1.40	1.40
C1-N6	1.33	1.33
C1-C10	1.51	1.52
C2-N3	1.33	1.33
C2-H7	1.08	1.08
N3-C4	1.33	1.34
C4-C5	1.39	1.38
C4-H8	1.08	1.08
C5-N6	1.33	1.34
C5-H9	1.08	1.08
N6-H16	N/A*	1.03
C10-H11	1.10	1.09
C10-H12	1.09	1.09
C10-N13	1.46	1.45
N13-H14	1.01	1.01
N13-H15	1.01	1.01

Table S5. Benchmark Calculations of Bond Angles on 2-AMP and 1*H*-2-AMP in the Gas Phase by CCSD/6-31+G(d,p) and M05-2X/6-31+G(d,p).

	2-AMP		1*H*-2-AMP	
Angle	CCSD	M05-2X	CCSD	M05-2X
C2-C1-N6	121.09	120.94	116.74	116.82
C2-C1-C10	122.59	122.32	126.61	126.79
N6-C1-C10	116.32	116.74	116.65	116.39
C1-C2-N3	122.69	122.54	122.42	122.14
C1-C2-H7	120.60	120.41	120.32	120.33
N3-C2-H7	116.71	117.05	117.26	117.53
C2-N3-C4	115.76	116.08	118.23	118.63
N3-C4-C5	121.95	121.78	122.00	121.84
N3-C4-H8	117.09	117.32	117.30	117.51
C5-C4-H8	120.96	120.91	120.70	120.66
C4-C5-N6	122.10	121.96	117.32	117.31
C4-C5-H9	120.91	120.84	124.79	124.54
N6-C5-H9	116.99	117.20	117.89	118.15
C1-N6-C5	116.40	116.70	123.29	123.25
C1-N6-H16	N/A*	N/A*	113.06	112.47
C5-N6-H16	N/A*	N/A*	123.66	124.28
C1-C10-H11	108.09	107.54	108.12	108.00
C1-C10-H12	109.45	109.20	108.12	108.00
C1-C10-N13	109.10	109.36	109.15	109.38
H11-C10-H12	107.76	107.75	107.02	106.92
H11-C10-N13	113.80	113.72	112.14	112.18
H12-C10-N13	108.59	109.18	112.14	112.18
C10-N13-H14	109.03	109.23	111.91	112.64
C10-N13-H15	110.16	110.82	111.91	112.64
H14-N13-H15	107.36	108.30	107.47	108.24

Table S6. Benchmark Calculations of Bond Angles on 2-AMP and 1*H*-2-AMP in the Aqueous Phase by M05-2X/6-31+G(d,p).

Angle	2-AMP	1*H*-2-AMP
C2-C1-N6	120.79	116.67
C2-C1-C10	121.70	125.95
N6-C1-C10	117.51	117.39
C1-C2-N3	122.86	122.75
C1-C2-H7	120.25	119.74
N3-C2-H7	116.89	117.51
C2-N3-C4	115.81	117.78
N3-C4-C5	121.82	122.23
N3-C4-H8	117.44	117.52
C5-C4-H8	120.74	120.25
C4-C5-N6	122.09	117.47
C4-C5-H9	120.59	124.26
N6-C5-H9	117.31	118.28
C1-N6-C5	116.63	123.11
C1-N6-H16	N/A*	113.64
C5-N6-H16	N/A*	123.26
C1-C10-H11	107.56	107.62
C1-C10-H12	108.83	107.70
C1-C10-N13	110.68	110.07
H11-C10-H12	107.52	106.66
H11-C10-N13	113.41	112.17
H12-C10-N13	108.71	112.38
C10-N13-H14	108.84	111.77
C10-N13-H15	109.03	111.76
H14-N13-H15	106.36	108.29

Table S7. Benchmark Calculations of Partial Atomic Charge on 2,4-DAP and 1*H*-2,4-DAP in the Gas Phase by CCSD/6-31+G(d,p) and M05-2X/6-31+G(d,p). Note that Mulliken Charges are Shown for CCSD, and CM4 Charges are Shown for M05-2X Calculations.

	2,4-DAP		1*H*-2,4-DAP	
Atom	CCSD/Mulliken	M05-2X/CM4	CCSD/Mulliken	M05-2X/CM4
C1	0.44	0.44	0.67	0.51
N2	-0.51	-0.45	-0.52	-0.41
C3	0.19	0.33	0.22	0.39
C4	-0.06	-0.20	0.10	-0.12
C5	-0.01	0.15	-0.03	0.19
N6	-0.44	-0.44	-0.52	-0.39
H7	0.17	0.09	0.23	0.13
H8	0.16	0.08	0.24	0.12
N9	-0.61	-0.63	-0.65	-0.55
H10	0.32	0.31	0.36	0.33
H11	0.32	0.31	0.38	0.34
N12	-0.61	-0.62	-0.61	-0.53
H13	0.33	0.31	0.38	0.34
H14	0.30	0.31	0.37	0.33
H15	N/A*	N/A*	0.39	0.33

Table S8. Benchmark Calculations of Partial Atomic Charge on 2,4-DAP and 1*H*-2,4-DAP in the Aqueous Phase by M05-2X/6-31+G(d,p). Note That Both Sets of Data are CM4 Charges.

Atom	2,4-DAP	1*H*-2,4-DAP
C1	0.42	0.51
N2	-0.49	-0.45
C3	0.32	0.37
C4	-0.19	-0.13
C5	0.14	0.20
N6	-0.50	-0.38
H7	0.11	0.13
H8	0.08	0.13
N9	-0.62	-0.55
H10	0.32	0.35
H11	0.32	0.34
N12	-0.58	-0.55
H13	0.33	0.33
H14	0.34	0.34
H15	N/A*	0.35

Table S9. Benchmark Calculations of Bond Lengths in 2,4-DAP and 1*H*-2,4-DAP in the Gas Phase by CCSD/6-31+G(d,p) and M05-2X/6-31+G(d,p).

	2,4-DAP		1*H*-2,4-DAP	
Bond	CCSD	M05-2X	CCSD	M05-2X
C1-N2	1.35	1.34	1.33	1.32
C1-N6	1.34	1.34	1.37	1.37
C1-N9	1.38	1.36	1.34	1.33
N2-C3	1.34	1.33	1.34	1.34
C3-C4	1.41	1.41	1.44	1.44
C3-N12	1.38	1.36	1.33	1.33
C4-C5	1.38	1.38	1.35	1.35
C4-H7	1.08	1.08	1.08	1.08
C5-N6	1.35	1.34	1.38	1.37
C5-H8	1.08	1.09	1.08	1.08
N6-H15	N/A*	N/A*	1.01	1.01
N9-H10	1.01	1.00	1.01	1.01
N9-H11	1.01	1.00	1.01	1.01
N12-H13	1.01	1.01	1.01	1.01
N12-H14	1.01	1.00	1.01	1.01

Table S10. Benchmark Calculations of Bond Lengths in 2,4-DAP and 1*H*-2,4-DAP in the Aqueous Phase by M05-2X/6-31+G(d,p).

Bond	2,4-DAP	1*H*-2,4-DAP
C1-N2	1.34	1.33
C1-N6	1.34	1.36
C1-N9	1.36	1.33
N2-C3	1.34	1.34
C3-C4	1.41	1.43
C3-N12	1.34	1.33
C4-C5	1.37	1.35
C4-H7	1.09	1.08
C5-N6	1.34	1.37
C5-H8	1.09	1.08
N6-H15	N/A*	1.02
N9-H10	1.01	1.01
N9-H11	1.01	1.01
N12-H13	1.01	1.01
N12-H14	1.01	1.01

Table S11. Benchmark Calculations of Bond Angles on 2,4-DAP and 1*H*-2,4-DAP in the Gas Phase by CCSD/6-31+G(d,p) and M05-2X/6-31+G(d,p).

	2,4-DAP		1*H*-2,4-DAP	
Angle	CCSD	M05-2X	CCSD	M05-2X
N2-C1-N6	127.23	127.03	122.28	121.97
N2-C1-N9	116.02	116.36	118.75	119.18
N6-C1-N9	116.70	116.60	118.95	118.85
C1-N2-C3	116.28	116.49	118.39	118.86
N2-C3-C4	121.94	121.86	122.14	121.93
N2-C3-N12	116.37	116.46	117.00	117.17
C4-C3-N12	121.62	121.66	120.86	120.90
C3-C4-C5	115.90	115.65	116.89	116.76
C3-C4-H7	122.07	122.21	121.91	122.04
C5-C4-H7	122.02	122.14	121.20	121.20
C4-C5-N6	123.92	124.09	120.03	120.27
C4-C5-H8	120.43	120.13	123.94	123.58
N6-C5-H8	115.65	115.77	116.03	116.15
C1-N6-C5	114.73	114.88	120.26	120.21
C1-N6-H15	N/A*	N/A*	120.44	120.45
C5-N6-H15	N/A*	N/A*	119.29	119.34
C1-N9-H10	115.38	117.85	122.94	123.68
C1-N9-H11	114.94	117.26	117.42	117.46
H10-N9-H11	116.12	119.67	118.29	118.86
C3-N12-H13	114.45	116.57	119.30	119.02
C3-N12-H14	116.83	119.54	121.92	121.95
H13-N12-H14	114.79	118.02	118.78	119.03

Table S12. Benchmark Calculations of Bond Angles on 2,4-DAP and 1*H*-2,4-DAP in the Aqueous Phase by M05-2X/6-31+G(d,p).

Angle	2,4-DAP	1*H*-2,4-DAP
N2-C1-N6	126.96	122.06
N2-C1-N9	116.37	119.52
N6-C1-N9	116.65	118.42
C1-N2-C3	116.89	118.15
N2-C3-C4	121.23	122.39
N2-C3-N12	117.44	117.47
C4-C3-N12	121.33	120.14
C3-C4-C5	115.94	116.81
C3-C4-H7	122.22	121.80
C5-C4-H7	121.83	121.39
C4-C5-N6	124.27	120.08
C4-C5-H8	120.10	123.64
N6-C5-H8	115.63	116.28
C1-N6-C5	114.69	120.52
C1-N6-H15	N/A*	120.36
C5-N6-H15	N/A*	119.12
C1-N9-H10	118.09	122.39
C1-N9-H11	118.62	119.05
H10-N9-H11	117.84	118.56
C3-N12-H13	120.37	120.29
C3-N12-H14	121.00	121.10
H13-N12-H14	118.60	118.60

Table S13. Benchmark Binding Energies Calculated in the Gaseous Phase by M05-2X/6-31+G(d,p).

System	Binding Energy (kcal/mol)
2,4-DAP(A)	-11.05
2,4-DAP(B)	-11.05
2,4-DAP(C)	-9.94
2,4-DAP(D)	-10.48
2,4-DAP(E)	-6.46
2-AMP	-7.15
1*H*-2,4-DAP(A)	-18.80
1*H*-2,4-DAP(B)	-18.80
1*H*-2,4-DAP(C)	-12.32
1*H*-2,4-DAP(D)	-11.78
1*H*-2,4-DAP(E)	-14.24
1*H*-2-AMP	-7.94

Appendix Two

Derivation of Basic Competition Model Equation

The following derivation makes the assumption that K_{eq} is large, implying that a second molecule of MTX (in the context of the dimerization experiment) displaces the dimerizer, rather than dissolving another dimer. We also assume that K_{a1} and K_{a2} are approximately equivalent to K_{aMTX}. This derivation is reported here as performed by Dr. Jonathan C.T. Carlson.

Relevant Equilibrium Expressions

$$PDP \rightleftharpoons PD + P \qquad K = 1/K_c K_{a2}$$

$$PD \rightleftharpoons P + D_o \qquad K = 1/K_{a1}$$

$$D_o \rightleftharpoons D_c \qquad K = 1/K_{eq}$$

$$2P + 2M \rightleftharpoons 2PM \qquad K = K_{aMTX}^2$$

$$PDP + 2M \rightleftharpoons 2PM + D_c \qquad K = (K_{aMTX}^2 K_{eq})/(K_{a1} K_{a2} K_c) \text{ or } \sim K_{eq}/K_c$$

Derivation

Based on the expressions above, the apparent equilibrium constant can be written as:

$$K = \frac{[PM]^2 D_c}{[PDP][M]^2} \qquad (1)$$

where [PDP] is the concentration of the dimer complex, [PM] is the concentration of the DHFR-MTX complex, and [M] is the concentration of MTX. Rearrangement and substitution yields:

$$[PDP] = \frac{[PM]^2 D_c}{K[M]^2} = \frac{K_c D_c [PM]^2}{K_{eq}[M]^2} \qquad (2)$$

To express the equation in terms of experimentally observable values, we express the species as their total concentrations:

$$P_t = 2[PDP] + [PD] + [P] + [PM] \qquad (3)$$

$$M_t = [M] + [PM] \qquad (4)$$

$$D_t = D_o + D_c + [PD] + [PDP] \qquad (5)$$

If K_{eq} is large, the values for [PD], [P], and D_o are zero, and we can rewrite the equations as:

$$[PM] = P_t - 2[PDP] \quad (6)$$

$$[M] = M - P_t + 2[PDP] \quad (7)$$

$$D_c = D_t - [PDP] \quad (8)$$

Substitution into equation 2 yields the following:

$$[PDP] = \frac{K_c(D_t - [PDP])(P_t - 2[PDP])^2}{K_{eq}(M_t - P_t + 2[PDP])^2} \quad (9)$$

In terms of the competition experiments, the observed fraction dimer ([PDP*]) can be substituted for [PDP], the equivalents dimerizer for D_t, and the equivalents MTX for M_t, yielding the finalized equation used for data fitting:

$$[PDP^*] = \frac{K_c(0.5 - [PDP^*])(P_t - 2[PDP^*])^2}{K_{eq}(eq.MTX - P_t + 2[PDP^*])^2} \quad (9)$$

Note that this equation serves functionally the same purpose as Equation 8, but is only modified to facilitate data analysis in the context of the competition experiment.

Appendix Three

Custom FORTRAN Script for Umbrella Sampling Probability Distributions

```
      PROGRAM READDIST
      IMPLICIT REAL*8 (A-H,O-Z)
      PARAMETER (MAXGRD=3500)
      DIMENSION GRM(MAXGRD),GRP(MAXGRD)
      DIMENSION GN(MAXGRD)
C
C
      NGRID=3500
      GWD=0.01D+00
      GINI=1.0D+00
C
      NSTPI=0
      NSTEP=25000
C
C
      GINI=GINI-GWD
C
      DO 10 IG=1,NGRID
        GRM(IG)=GINI+GWD*DBLE(IG)-0.5D+00*GWD
        GRP(IG)=GINI+GWD*DBLE(IG)+0.5D+00*GWD
        GN(IG)=0.0D+00
 10   CONTINUE
C
C
      DO 100 ISTPI=1,NSTPI
        READ(*,*) TMP,R
 100  CONTINUE
C
      DO 1000 ISTEP=1,NSTEP
C
        READ(*,*) TMP,R
C
        DO 200 IG=1,NGRID
          IF(R.GT.GRM(IG).AND.R.LE.GRP(IG)) THEN
            GN(IG)=GN(IG)+1.0D+00
            GOTO 1000
          END IF
 200      CONTINUE
C
 1000 CONTINUE
C
C
      CONST=1.0D+00/GWD/DBLE(NSTEP)
```

```
C
      WRITE(*,9999)
 9999 FORMAT("#    DISTANCE   NUMBER    PROBABILITY")
C
      DO 2000 IG=1,NGRID
        RG=0.5D+00*(GRM(IG)+GRP(IG))
        WRITE(*,'(1X,F12.6,F10.0,F12.6)') RG,GN(IG),CONST*GN(IG)
 2000 CONTINUE
C
      STOP
      END
```