Title
Backbone Flexibility in Computational Protein Design

Author
Mandell, Daniel Jonathan

Publication Date
2010

Peer reviewed|Thesis/dissertation

Backbone Flexibility in Computational Protein Design

by

Daniel J. Mandell

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological & Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Dedicated to my loving family

past, present, and future

Acknowledgements

I am extremely fortunate to have been guided by a line of mentors possessing some of the finest qualities a student could hope for. Alice Davis opened my eyes to the physical sciences, and taught me to *think* before opening my mouth; a skill I am still mastering. Sheila McIlraith, Armando Fox, and Eduardo Pelegri-Llopart taught me how to practice computer science, while Russ Altman showed me how those practices could allow me to investigate the fascinating biological questions I thought I had to abandon as a computationalist. Chris Williams showed me what quantitative rigor means, and Seth Grant taught me that if I wanted to make my work relevant to biology, I really needed to understand some biology.

As I came to appreciate at UCSF, that list extends to chemistry and physics. I am extremely grateful for the collaborative climate at UCSF, which offers academic journeymen like me the opportunity to learn the fundamental principles necessary for cross-disciplinary science. These skills are imparted through coursework, seminars, rotation projects, and above all, fellow students. I thank Mike Kim, Libusha Kelly, and Holly Atkinson for convincing me I'd be at home at UCSF, and immersing me in the culture. John Chodera, Vince Voelz, and Ilya Chorny showed me the power of molecular simulation, and reinforced the notion that a PhD in theory is a PhD in theory. I am grateful to all members of the Kortemme lab for invaluable support and criticism as we have navigated our graduate studies together. And I offer a special thanks to my friends from other UCSF labs, including Caleb Bashor,

Andrew Horwitz, Ray Nagatani, and Jason Porter, who have helped me develop both professionally and personally.

The faculty support I have received at UCSF has been unparalleled. I am grateful to my academic advisors, Andrej Sali and Patsy Babbit, who also served on my orals committee, for sound guidance on issues academic and otherwise. I especially thank my external thesis committee members, David Agard and Matt Jacobson, who directed me toward bigger questions and genuinely cared about my progress. Their feedback, starting from my qualifying exam and continuing through the process of dissertation writing, has been instrumental in producing work I can be proud to call my own. I am very fortunate for my close collaborator Vageli Coutsias, whose poetic presentation of abstract mathematical concepts almost conceals the subtle elegance of his machinations. And finally, my deepest gratitude goes to Tanja Kortemme, whose open mind, broad expertise, and overwhelming passion for science has provided the ideal environment for my doctoral studies. It was through my conversations with Tanja, whether designing a project, crafting a manuscript, staring at protein structures, or arguing about physical forces that may or may not be present, that I became a real scientist.

At the heart of it all, I thank my parents, Herb and Peggy Mandell, first on the dance floor and last to leave, who have done everything in their power to provide me with a foundation for opportunity, and my little sister, Lydia, who has shown me that some people make the world a better place just by being in it.

Chapter 1 of this dissertation contains a figure from Mandell, D.J. & Kortemme, T. Computer-aided design of functional protein interactions. *Nat Chem Biol* **5**, 797-807 (2009).

Chapter 2 is largely adapted from Mandell, D.J., Chorny, I., Groban, E.S., Levin, E., Rapp, C.S., & Jacobson, M.P. Strengths of hydrogen bonds involving phosphorylated amino acid side chains. *J Am Chem Soc* **129**, 820-7 (2007). Ilya Chorny wrote the weighted histogram analysis code, Sergio Wong performed quantum mechanics calculations, and Eli Groban performed implicit solvent Poisson-Boltzmann calculations.

Chapter 3 and Chapter 6 include material from Mandell, D.J. & Kortemme, T. Backbone flexibility in computational protein design. *Curr Opin Biotechnol* (2009).

Chapter 4 is largely adapted from Mandell, D.J., Coutsias, E.A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods* **6**, 551-2 (2009). Vageli Coutsias formulated the described approach to kinematic closure, together with the underlying techniques using polynomial resultants.

Abstract

Backbone Flexibility in Computational Protein Design

Daniel J. Mandell

Over the past two decades the field of computational protein design has produced striking successes, both by improving our understanding of the fundamental principles governing protein structure, dynamics and function, and by engineering new and modified proteins with useful properties for scientific inquiry and industrial applications. Generally, these successes have arisen from design strategies that sample amino side-chains on a polypeptide backbone with fixed atomic coordinates. Despite the well-known tendency for protein backbones to adjust in the face of sequence mutations, the fixed backbone assumption is typically maintained due to the challenge of efficiently sampling backbone and side-chain conformations, and accurately evaluating their physical favorability. This dissertation addresses these issues of sampling and evaluating protein conformations and applies the developed methods to predict proteins with new functions. The dissertation first provides a quantitative assessment of hydrogen bonding involving amino acid phosphorylation, a key post-translational modification that can alter protein function by inducing conformational

rearrangements. The dissertation then introduces a robotics-inspired method for modeling backbone flexibility in proteins, and demonstrates sub-angstrom accuracy in an application to predict the conformations of regions lacking secondary structure in protein monomers and interfaces. Finally, the dissertation describes the coupling of the developed flexible backbone method with computational sequence design to predict proteins that dimerize only in the presence of small molecule targets to act as *in vitro* or *in vivo* biosensors. Fusing split reporter fragments to the chemically induced dimer partners provides a modular approach that can, in principle, be used to detect any small molecule that has been crystallized in complex with a protein and drive a variety of enzymatic, fluorescent, and transcriptional outputs.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Perspective

The principal goal of computational protein design is to find amino acid sequences that stably adopt target protein structures. Increasingly, designs are also optimized with functional constraints in mind, such as interaction selectivity or new catalytic activity. Several ambitious engineering goals have recently been met, including designing proteins that catalyze reactions lacking natural enzymes[1,2], increasing the affinity of antibody-antigen interactions beyond *in vivo* levels[3], and engineering peptide transcription factors that oligomerize with high specificity[4]. These successes arise from concepts developed over several decades[5-8] that were highlighted by the pioneering work of Dahiyat and Mayo on the first computational design of an entire protein[9].

With some notable exceptions[10-13] most early design protocols sample amino acid side-chain conformations on a backbone template with fixed atomic coordinates. This simplifying assumption is generally made because of the difficulty of developing sampling methods and scoring functions to efficiently explore backbone flexibility. In addition, simulating the vastly enlarged conformational

space when moving from fixed to flexible backbone models requires substantial computational resources. Nevertheless, experiments have long demonstrated that protein backbones often adjust to sequence mutations[14-16]. These observations suggest that, depending on the conformational plasticity of the target fold, fixed backbone methods may neglect a significant portion of sequence space accessible to folded and functional proteins. Conformational changes to protein backbones also underlie important biological processes, such as active site gating in enzymes, switch modulation in signaling proteins, and antigen recognition in antibodies. Consequently, for computational protein design to meet the next generation of engineering challenges, as well as to improve understanding of the relationships between variations in protein sequence, structure, dynamics, and function, it is critical to develop and validate efficient methods to model varying levels of backbone flexibility.

Synopsis

This dissertation addresses the incorporation of backbone conformational variability into protein modeling and design. Enabling backbone flexibility in high-resolution protein modeling requires developments both in efficiently exploring protein conformational space ('sampling', Figure 4.1), and in evaluating the favorability of sampled conformations ('scoring', Figure 1.1). Chapter 2 focuses on the scoring issue, using multiple levels of theory to quantify the relative strengths of hydrogen bonds involving amino acid side-chains that have undergone phosphorylation, a key post-translational modification that often effects function by

inducing conformational changes. The subsequent chapters focus on high-resolution sampling. Chapter 3 begins with a history of approaches to modeling backbone flexibility that enabled early successes and illuminated challenges ahead. The chapter then describes the significant methodological advances that have occurred in recent years, and reports on newly enabled biological and engineering applications. Chapter 4 introduces a robotics-inspired technique for modeling protein conformations, and describes its application to predicting the conformations of protein regions lacking secondary structure in both monomeric proteins and protein interfaces. Chapter 5 describes the application of the developed methods to reshaping protein-protein interfaces around small molecule targets so they may function as *in vitro* or *in vivo* biosensors. Finally, Chapter 6 looks toward future challenges in flexible backbone design, and proposes how current and forthcoming approaches might be harnessed to design new functions and incorporate concepts from protein evolution.

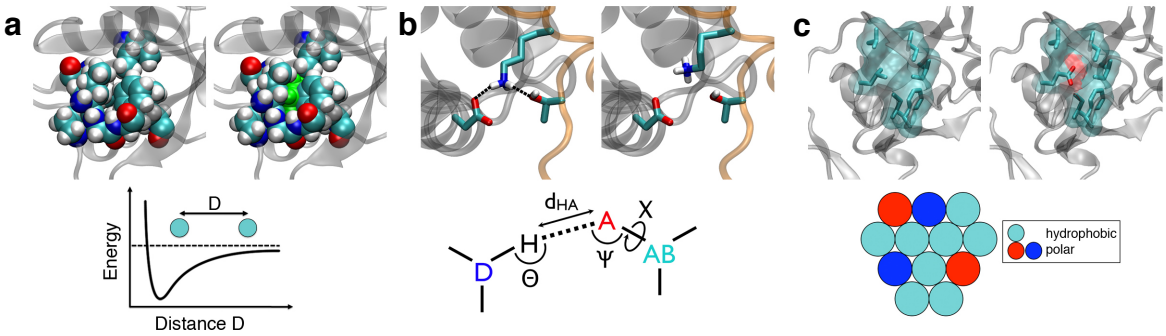

Figure 1.1: Dominant scoring terms used in computational protein design
(**a**) Atomic packing interactions. The well-packed core of a PDZ domain (top left) would be destabilized by changes in side-chain size, such as an alanine-to-phenylalanine substitution (top right, phenylalanine shown in green). Packing interactions are typically described using van der Waals potentials (bottom). The backbone is shown in transparent cartoon representation, and side-chains are shown as spheres. (**b**) Hydrogen bonding. The intricate geometry of a hydrogen bond network in the interface between the DNAse E7 (orange backbone) and its inhibitor Im7 (grey backbone) (PDB code 7cei) (left) would be disrupted by changes in side-chain donor/acceptor groups and their orientations (right). The bottom panel shows parameters used in a geometry-dependent hydrogen bonding potential[17]: the distance between the hydrogen (H) and the acceptor (A) (d_{HA}), the angle at the hydrogen (Θ), the angle at the acceptor (Ψ), and the donor (D)–hydrogen (H)–acceptor (A)–acceptor base (AB) torsion angle (χ). Side-chains are shown in stick representation. (**c**) Solvation. The hydrophobic core of a PDZ domain (left) would be destabilized by an isoleucine-to-glutamate substitution (right). Solvation potentials favor the exposure of polar groups to the solvent and associate a penalty with burial of polar groups (bottom). Core side-chains are shown as sticks with a transparent surface representation.

Chapter 2

Strengths of hydrogen bonds involving phosphorylated amino acid side-chains

Introduction

Protein phosphorylation is a key signaling mechanism in diverse cellular processes including metabolism[18], ion channel regulation[19,20], and cell cycle progression[21-23]. In eukaryotes, the main sites of phosphorylation are tyrosine, serine and threonine side-chains, while aspartate and histidine side-chains are phosphorylated in the 'two-component' signaling pathways of prokaryotes[23,24]. Addition of the phosphate group, which typically carries a -2 charge at physiological pH, perturbs the local electrostatic potential in the protein and often induces conformational changes that influence function[23] or modulate protein-protein interactions[18].

A critical property of phosphorylated residues is their propensity to accept hydrogen bonds through their phosphate oxygens, frequently with positively charged side-chains to form 'salt bridges'. Salt bridge energetics depend sensitively on the identity, proximity and orientation of the participating side-chains and their surrounding environment. Quantitative measurements of salt bridge contributions to protein stability have provided different results depending on the experimental

system and experimental design, with estimates ranging from -5.0 to -0.5 kcal/mol of stabilization in T4 lysozyme[25,26], to 2.0 to 4.0 kcal/mol of destabilization in coiled coils and Arc repressor[27,28]. There have also been a number of previous computational studies aimed at quantifying the strength of interaction of small ions[29-32]. In particular, Masunov and Lazaridis[32] used molecular dynamics methods to estimate the free energies of salt bridges between likely orientations of all charged naturally occurring amino acid side-chains.

This chapter investigates the strengths of hydrogen bonds and salt bridges involving phosphorylated amino acid side-chains using small molecule analogs for common acceptors (methyl phosphate for pSer and pThr, acetyl phosphate for pAsp) and donors (butyl ammonium for a Lys side-chain, propyl guaninidinium for Arg, and N-methylacetamide for backbone amide NH groups). Interactions of all donors with propionic acid (Glu analog) are also considered for comparison to a carboxylate receptor with -1 charge.

Multiple levels of theory are utilized, including explicit solvent molecular dynamics (MD), implicit solvent molecular mechanics (Poisson-Boltzmann), and quantum mechanics with a self-consistent reaction field treatment of solvent. This approach allows the identification of trends that are consistent across the methods, as well as to uncover the sensitivity of each method to different forces governing hydrogen bond strengths. Continuum solvent methods, primarily those based on the Poisson-Boltzmann equation[33] or more heuristic methods such as Generalized Born[34], offer substantial speed advantages relative to explicit solvent models in applications such as molecular dynamics. However, treating the solvent as a

continuum dielectric is an approximation and neglects important first-shell solvation effects related to the finite size and asymmetry of a water molecule[35-37]. The results of the explicit solvent calculations cannot be considered free of error either. Notably, molecular mechanics methods using fixed-charged force fields, as employed in the present molecular dynamics calculations, ignore the effect of electronic polarizability on hydrogen bond strengths. Even in high dielectric solvent, the strong electric field exerted by a -2 phosphate group can be expected to lead to significant polarization of the electrons on nearby molecules. To assess the potential impact of electronic polarizability on the strengths of the hydrogen bonds considered here, this work employs quantum mechanics with a large basis set, electron correlation treated at the 'local' Moller-Plesset second order perturbation theory[38] (LMP2) level, and solvent treated using a self-consistent reaction field (SCRF) method.

The central results of this chapter consist of potentials of mean force (PMFs) from the explicit solvent molecular dynamics calculations, which are one-dimensional free energy landscapes for a pair of interacting groups as a function of distance between the phosphate and hydrogen bond donors. The PMFs cannot be used trivially to predict the *absolute* free energies of association of the small molecules in solution (which requires extensive averaging over translational and rotational degrees of freedom) or the absolute strengths of hydrogen bonds in a protein environment (which depend on the local environment, such as solvent accessibility). However, the PMFs do provide insight into the *relative* intrinsic

strengths of the various types of hydrogen bonds considered, and help to address the following issues:

(1) *The conditions under which conditions Arg or Lys make stronger hydrogen bonds with a phosphorylated side-chain.* There is ample albeit indirect evidence that the ability of guanidinium ions to form bidentate hydrogen bonds with carboxylate or phosphate ions leads to particularly strong interactions[39-41]. This property of guanidinium ions has been extensively employed in the design of synthetic receptors for phosphate-containing ligands[42]. Bidentate hydrogen bonds between Arg and pSer/pThr are also commonly observed in the relatively small number of crystal structures of phosphorylated proteins[23]. However, previous computational studies have suggested that interactions of phosphorylated groups with Lys may be intrinsically stronger[43,44]. This issue is revisited here using multiple levels of theory.

(2) *The effect of phosphate protonation state on hydrogen bond strength.* The ~6 pKa of phosphate suggests that both -1 and -2 charged species may coexist at physiological pH. This work investigates the effect of phosphate protonation state on all hydrogen bonding interactions considered.

(3) *The energetic consequences of substituting a carboxylate for a phosphate.* In cases where phosphorylation of a protein leads to its activation, it is frequently useful to engineer a constitutively active mutant, e.g., for use in *in vitro* studies. Simply

substituting a Glu or Asp for the phosphorylated residue(s) is sufficient to achieve constitutive activation in many cases[45-49], but in other cases this simple strategy results in only partial activation or none at all[50,51]. The relative strengths of hydrogen bonds involving carboxylates and phosphates is also relevant to the design of inhibitors of SH2 domains[52,53], which bind phosphorylated peptides. Here, the differences in the intrinsic hydrogen bond strengths of carboxylates *vs*. phosphates with common hydrogen bond donors is examined in some detail. These results provide a foundation for understanding why Asp/Glu can sometimes substitute for phosphorylated amino acids, although other physicochemical differences will undoubtedly also play a role.

The strengths of hydrogen bonds to the phosphate backbone of RNA[54] and DNA[55] are not directly addressed here, although the results presented may have some relevance to this issue.

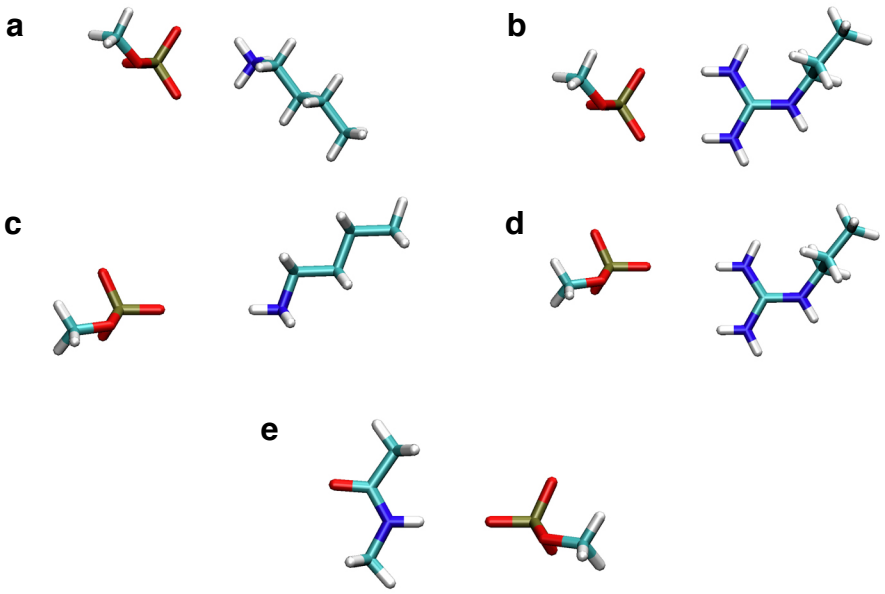

Figure 2.1: Hydrogen bond geometries considered
Only the unprotonated methyl phosphate acceptor (representing the pSer side-chain with a -2 charge) is shown. In the case of protonated phosphate groups, the hydrogen is placed on one of the oxygens not directly involved in hydrogen bonding. These geometries are referred to as (**a**) 'Lys coplanar', (**b**) 'Arg coplanar', (**c**) 'Lys collinear', (**d**) 'Arg collinear', and (**e**) 'amide NH collinear'.

Methods

MD simulation materials and parameters

MD simulations were performed using GROMACS 3.2.1[56]. The 2001 OPLS all atom force field[57] was used for stretch, bend, torsional and Lennard-Jones parameters, as well as partial charges for the glutamate, acetate, arginine, lysine, and N-methylacetamide backbone analogs. Partial charges for pSer (-1, -2) and pAsp (-1, -2) analogs were obtained from quantum mechanical calculations[58,59]. Molecules were solvated with TIP3P[60] water molecules in a 40 Å cubic box under periodic boundary conditions. The cut-off distance for the short-range neighbor list was set to 10 Å. Long-range electrostatics were calculated with particle mesh Ewald[61] with a

real-space cutoff of 9 Å for nonbonded interactions. Na^+ counterions were fixed at the corners of the solvent box as necessary to obtain electroneutrality. Prior to running molecular dynamics, the potential energy of each configuration was relaxed by steepest descent minimization, followed by 100 ps of molecular dynamics equilibration. Molecular dynamics was then run for 2.1 ns with a time step of 2 fs. Interaction energies and atomic coordinates were recorded every 500 fs. The system was propagated in time with a velocity version of the Verlet algorithm[62,63]. During and subsequent to equilibration, Nose-Hoover temperature coupling[64] and Berendsen pressure coupling[65] were used to maintain system temperature and pressure, with a reference temperature of 298 K, a reference pressure of 1.0 atm, a time constant of 1 ps, and an isothermal compressibility of 1.1×10^{-6} (kcal mol^{-1} $Å^{-3}$). The various acceptors, donors, charge states, and geometries comprised a total of 25 hydrogen bonding configurations representing more than 1 μs of molecular dynamics simulation time.

MD simulation constraints

Umbrella sampling[66] using distance and position restraints was employed to calculate a one-dimensional PMF for one of two common interaction geometries, either a coplanar approach or a collinear approach (Figure 2.1). These orientations consistently arose in simulations of the side-chain analogs without position restraints (data not shown). For collinear geometries, the donated hydrogen and its covalently bonded atom, and the accepting oxygen and its covalently bonded atom, were constrained to move on a line. For coplanar geometries, donated hydrogens

and accepting oxygens were constrained to a plane, and two additional heavy atoms from each molecule were constrained to move along a line. These constraints kept the interacting moieties facing each other, and in the case of lysine allowed the donated hydrogens to rotate slightly through the plane of interaction to sample optimal hydrogen bonding orientations. Hydrogen atom–heavy atom covalent bond lengths were constrained only in the backbone analog using the LINCS algorithm[67] to stabilize the N–H bond.

The distance between the molecules was constrained using a biasing potential at 0.5 Å intervals. A nearby heavy atom from the donor molecule was constrained to that of the acceptor using a quadratic biasing potential, $V(r) = k(r - r_i)$ where k = 143.5 (kcal/Å^2) is the biasing force constant and r_i is the point about which the molecules are constrained. In cases with guanidinium in a coplanar orientation, additional simulation with a higher force constant of 263.0, 430.4, or 860.8 (kcal/Å^2) was necessary to sample adequately around the solvation barrier.

MD analysis

The potential of mean force was calculated using probability distributions of the constrained distance, r, obtained from the umbrella sampling, as described by Souaille and Roux[68]. Trajectories from each window i were converted to biased population distributions $P_i(r)$ with a bin width of 0.1 Å. The weighted histogram analysis method (WHAM)[69] was used to merge histograms $P_i(r)$ into a single unbiased curve $P(r)$. The algorithm was considered to converge after the free energy constants for all the windows changed by less than 0.01 kcal/mol. The PMF, $\Delta G(r)$,

was then calculated using the standard relationship $\Delta G(r) = -k_b T \ln[P(r)]$ where k_b is Boltzmann's constant. Each PMF was shifted vertically so that the average potential between 10 Å and 11 Å was 0 kcal/mol. Block averaging[70] was used to calculate errors. The trajectories in each of the constrained windows were divided into N shorter trajectories. The distribution $P(r)$ for each of the N trajectories was then calculated using the methods described above. The resulting N $P(r)$ were averaged and used to calculate the standard deviation, which is reported as the error. N=20 was chosen to minimize the correlation between neighboring blocks.

Implicit solvent continuum electrostatics

Implicit solvent calculations were performed using the DelPhi program[71] to solve the linearized Poisson-Boltzman equation. In order to compare with the explicit solvent PMFs, DelPhi calculations were performed on each configuration used in the MD simulations except that the molecules were fixed at a defined distance from 2.5 Å to 11.0 Å at 0.25 Å intervals. The Coulombic and solvation (reaction field) components of the free energy from DelPhi were added to Lennard-Jones energies calculated separately to obtain the implicit solvent PMF. The nonpolar component of the solvation free energy was computed with a solvent-accessible surface area model. The partial charges and atomic radii in the explicit and implicit solvent simulations were the same; that is, the default charges and radii from Delphi are not used, in order to compare more directly with the molecular dynamics results. The Delphi calculations used 4 grid points per Å, an internal dielectric of 1, an external dielectric of 80, and an ionic strength of zero.

Quantum mechanics

Quantum mechanics calculations were performed using the Jaguar software package[72]. *Ab initio* single-point energy calculations were performed on the same coordinates employed using implicit solvent molecular mechanics. A self-consistent reaction field (SCRF) method[73] was used to mimic the condensed phase environment. The procedure starts by calculating atomic charges for the molecule in a vacuum using electrostatic fitting[74,75]. This step entails a Hartree-Fock calculation and subsequent electron correlation correction to evaluate the electrostatic potential. The response from the surrounding dielectric and corresponding surface charges is calculated. Atomic charges are then re-calculated taking into account the dielectric response. The solvation energy is calculated at each iteration until it converges. The basis set for the Hartree-Fock and electron correlation calculations was cc-pVTZ(-f). Electron correlation was treated at the level of local Moller-Plesset second order perturbation theory (LMP2)[38].

Results and discussion

Each level of theory employed in this study captures different aspects of hydrogen bonding with varying computational expense. Although not free from error, the explicit solvent molecular dynamics results include substantial averaging over conformational and rotational degrees of freedom for both solute and solvent, and comprise the core results of this chapter. The explicit solvent PMFs take a form typical to those of oppositely charged ions (e.g., Figure 2.2). The lowest free energy is generally seen when the ions are directly in contact (separation distance roughly

equal to the sum of the van der Waals radii); this minimum is referred to as the 'contact minimum'. As the separation between the ions increases, the free energy rises sharply to a solvation barrier. In many cases, as the separation between the molecules increases further the free energy reaches a second minimum, in which the solute ions are separated by approximately one water molecule, which is refered to as the 'solvent-separated minimum'. In cases with particularly well-ordered waters, a second barrier and second solvent-separated minimum may exist with further separation. At distances beyond these features the potential energy approaches zero. The free energy of the hydrogen bond or salt bridge is calculated as the difference in energy between the largest separation sampled (11 Å) and the contact minimum. The free energies of all orientations, charge states, and acceptor–donor pairs as calculated by MD in explicit solvent are summarized in Table 2.1.

Table 2.1: Hydrogen bonding free energies computed from explicit solvent MD and WHAM
Free energies with standard errors are shown in kcal/mol.

	Lys Collinear	Arg Collinear	Lys Coplanar	Arg Coplanar	Amide NH Collinear
Glu	-3.1±0.3	-3.4±0.2	-2.6±0.4	-8.5±0.1	-1.8±0.6
pSer(-1)	-3.7±0.2	-3.7±0.3	-3.5±0.1	-9.3±0.4	-1.6±0.7
pSer(-2)	-4.2±0.3	-4.7±0.3	-4.5±0.3	-10.6±0.6	-1.0±0.6
pAsp(-1)	-3.2±0.3	-3.0±0.4	-2.4±0.4	-6.3±0.4	-1.8±0.7
pAsp(-2)	-4.6±0.2	-4.5±0.3	-4.9±0.4	-7.3±0.2	-1.1±0.6

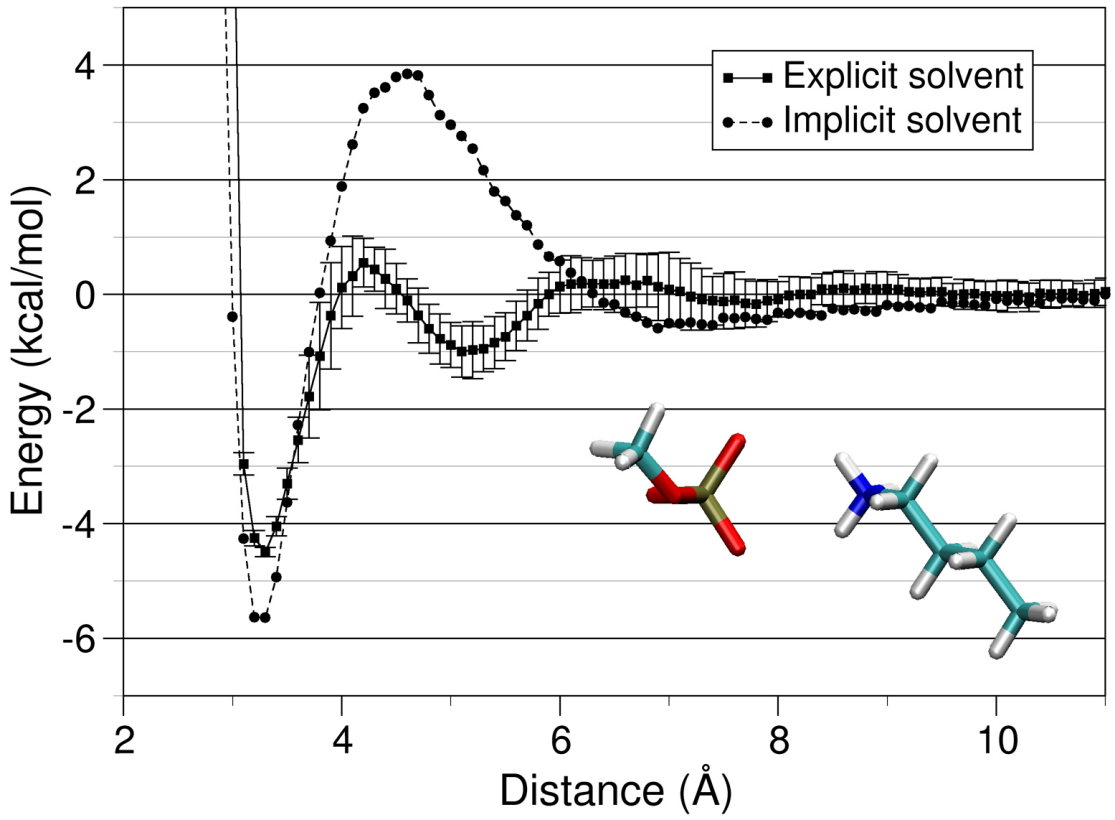

Figure 2.2: PMF for a coplanar interaction between phosphoserine and lysine
PMFs for a coplanar interaction between methyl phosphate (representing $pSer^{-2}$) and butyl ammonium (Lys) computed using MD in explicit solvent (squares with error bars, computed as described in 'Methods'), and using molecular mechanics with implicit solvent (dotted line and circles). The distance along the x-axis is measured between the phosphorus atom and the ammonium nitrogen.

Control study of NaCl

A one-dimensional PMF between Na^+ and Cl^- ions, a very well studied system, serves as a control. This experiment allows for comparisons to previous studies without the rotational degrees of freedom and multiple partial charges inherent to polyatomic systems. The calculations yielded a free energy at the contact minimum of -1.7 kcal/mol and a free energy at the top of the solvation barrier of 1.7 kcal/mol. Masunov and Lazaridis[32] found a contact minimum free energy of about -1.3

kcal/mol and a solvation barrier free energy of about 2.0 kcal/mol when using a Spherical Solvent Boundary Potential[76] for long-range electrostatics, and -2.0 kcal/mol and 1.5 kcal/mol respectively for the free energy when using Ewald summation. Other comparable studies from Smith and Deng[77], Lyubartsev and Laaksonen[29] (at 0.5 M concentration), and Martorana *et al.*[30] found comparable results, with contact minima of about -1 to -2 kcal/mol.

Effect of interaction geometry on energy landscapes

Examples of the collinear and coplanar geometries used for the PMFs are depicted in Figure 2.1. In cases where the phosphate is protonated, the proton is placed on an oxygen not involved in the hydrogen bonding; it serves only to change the overall charge on the phosphate. In the coplanar orientation the molecules are constrained to move along the axis of the distance constraint, and the planarity constraint prevents rotation about this axis. In contrast, the collinear orientation allows for rotation about the collinear axis. The height of the solvation barriers and the stability of the solvent-separated minima depend on the relative rotation of the molecules about the collinear axis, so integrating out this degree of freedom produces a PMF with less pronounced maxima, and frequently no secondary minima, in comparison to the coplanar PMFs. Note that the free energy computed for the hydrogen bond in this type of PMF explicitly includes entropic contributions from rotations about the collinear axis.

Salt bridge interactions of Glu/Asp

The strengths of salt bridges between positively charged Lys or Arg with negatively charged carboxylate groups on Glu and Asp have been considered in several previous studies[32,78]. Similar calculations are performed here only to obtain internally consistent comparisons with the results involving phosphate groups. Taking a collinear approach, propyl guanidinium (Arg) and butyl ammonium (Lys) yield similar free energies as hydrogen bond donors to propionic acid (Glu), at -3.4 kcal/mol and -3.1 kcal/mol respectively. In a planar approach, Arg is much more stable, at -8.5 kcal/mol. The enhanced stability of Arg is largely due to the ability of its guanidinium moiety to form nearly idealized bidentate hydrogen bonds with the carboxylate moiety of Glu. Moreover, the free energy of the bidentate bonds is slightly greater than double that of the single collinear bond, suggesting some cooperativity as there is less of an entropic cost to pay in forming the second hydrogen bond. Entropic cooperativity in bidentate bonding interactions, sometimes referred to as the chelate effect, has been observed elsewhere experimentally[79].

Masunov and Lazaridis computed one-dimensional PMFs between ionizable side-chain analogs similar to the configurations used here[32]. In particular, the Glu–Arg coplanar approach used similar ionic structures and positional restraints as employed by Masunov and Lazaridis in a corresponding calculation. Their analysis yielded a contact minimum free energy and solvation barrier free energy of about -4.3 kcal/mol and 3.0 kcal/mol respectively, roughly half the magnitude obtained here. This discrepancy is likely attributable to differences in force fields and

simulation protocol. Importantly, their simulations were run with the CHARMM 19 force field[80] for side-chain analogs, while the present work used the OPLS-AA 2001 force field (see 'Methods'). Notably, the charges for the donated Arg hydrogens are 0.35*e* in CHARMM as opposed to 0.46*e* in OPLS, and the accepting Glu oxygens take a charge of -0.60*e* in CHARMM in contrast to -0.80*e* in OPLS. Further, Masunov and Lazaridis handled long-range electrostatics with a Spherical Solvent Boundary Potential[76] (SSBP) over a spherical cluster of 200 waters with an 11 Å radius, while particle mesh Ewald over a 40 Å cubic box of roughly 2150 water molecules was employed here. Figure 2 of the Masunov paper shows that SSBP produces a contact minimum about 0.7 times the depth as Ewald summation when calculating a one-dimensional PMF for Na^+ and Cl^- ions. Rodinger *et al*.[81] found that the interface between an explicit water droplet and a continuum solvent field appearing in models like SSBP polarizes the explicit waters up to 10 Å away from the solvent-vacuum boundary. Lattice summation methods like particle mesh Ewald can also introduce small artifacts related to periodicity-induced perturbations in Coulombic and solvation energies, although given the quantity and permittivity of the solvent in the present study these perturbations should nearly cancel each other[82]. Additionally, the earlier paper reports the typical use of 7 umbrella sampling windows constructed at 1 Å intervals from 3 Å to 9 Å and simulated for 200 ps. In the present study, 18 windows were used, ranging from 2.5 Å to 11 Å at 0.5 Å intervals, and each window was equilibrated for 200 ps followed by 2.1 ns of simulation.

Rozanska and Chipot[78] calculated a PMF for guanidinium and acetate using Ewald lattice summation for long-range electrostatics, which corresponds geometrically to the Glu–Arg coplanar orientation. Their simulations produced a contact minimum free energy of -2.7 kcal/mol and a solvation barrier free energy of 3.4 kcal/mol. The atomic partial charges were computed from quantum mechanics and more closely resemble those of OPLS than CHARMM, with the donated hydrogens at 0.49*e* and the accepting oxygens at -0.87*e*. Importantly, their guanidinium and acetate moieties were biased to face one another through torsional restraints. This would allow some degree of rotation through the plane of interaction, in contrast to the C-C–C-N linearity constraint imposed on the corresponding ions of the present study, which enforces nearly constant bidentate hydrogen bonding. The present work found that removing this linearity constraint reduces the stability of the salt bridge by at least 1.5 kcal/mol. Other notable differences include Rozanska and Chipot's use of 4 umbrella sampling windows rather than 18, the AMBER force field[83] for the potential energy function rather than OPLS-AA, and the TIP4P water model rather than TIP3P.

Salt bridge interactions of pSer

The central results of this chapter are quantitative assessments of the strengths of hydrogen bond interactions involving phosphorylated amino acid acceptors. As is the case with carboxylate serving as the acceptor ion, butyl ammonium (Lys) and propyl guanidinium (Arg) donors yield similar contact minima free energies in a collinear orientation with protonated methyl phosphate ($pSer^{-1}$), at -3.7 kcal/mol.

When pSer is deprotonated, the energies of the collinear orientation become more distinguishable, with $pSer^{-2}$–Arg forming a salt bridge worth -4.7 kcal/mol, compared to -4.2 kcal/mol for $pSer^{-2}$–Lys. In the planar approach, Lys and Arg produce substantially different free energy profiles. When the accepting pSer is protonated, the contact minimum reaches -9.3 kcal/mol with Arg. If pSer takes a -2 charge, the free energy for this geometry is -10.6 kcal/mol.

The depth of the pSer–Arg contact minima further demonstrate the significance of bidentate hydrogen bonding in a coplanar approach. Similar to the effects observed with Glu–Arg, a coplanar pSer–Arg salt bridge provides more than twice the stability of one that is collinear. Coplanar pSer–Lys salt bridges yield similar energies to their collinear counterparts, which was also observed with Glu–Lys. PMFs with all permutations of pSer charge states with different donors also suggest that the effects of pSer protonation are small but significant. In a collinear approach, deprotonation of pSer stabilizes a pSer–Lys salt bridge by -0.5 kcal/mol and pSer–Arg by -1.0 kcal/mol. The effect appears stronger in a coplanar approach, with deprotonation stabilizing pSer–Lys by -1.0 kcal/mol and pSer–Arg by -1.3 kcal/mol.

Mavri and Vogel[43] used PM3 semi-empirical molecular orbital calculations with the SM3 reaction field treatment of solvent[84] to investigate the strengths of interactions for methylammonium and methylguanadinium with mono- and divalent methylphosphate in several orientations. Their coplanar calculations correspond geometrically with the coplanar pSer–Lys and pSer–Arg orientations. Although the authors conclude that phosphate interactions with Lys are generally

stronger than with Arg, their PM3-SM3 calculations show interaction free energies of +6.1 kcal/mol with coplanar $pSer^{-1}$-Lys and +3.6 kcal/mol for coplanar $pSer^{-1}$–Arg. The authors also found that $phenyphosphate^{-2}$ in complex with Lys or Arg produces an interaction free energy stronger than -28.0 kcal/mol in both cases. The findings clearly contradict the present results, both those generated using molecular mechanics and using quantum mechanical methods, which are largely consistent with each other. It should be noted that the authors of this study speculated that the semi-empirical quantum mechanical model may not have been well parameterized for phosphate groups.

Luo *et al.*[31] computed strengths of salt bridge interactions for Arg and Lys with phosphate using the CHARMM 22.0 empirical force field and a generalized Born implicit solvent model. In particular, they computed a PMF between monovalent phosphate and guanidinium that corresponds geometrically to the coplanar $pSer^{-1}$–Arg orientation. The authors found a contact minimum of about -3.8 kcal/mol, less than half the depth of the contact minimum of -9.3 kcal/mol for $pSer^{-1}$–Arg obtained here. Several differences in simulation protocol may contribute to this discrepancy. The CHARMM force field places a weaker charge on the unprotonated phosphate oxygens (-0.82*e*) than the charge obtained from electrostatic potential fitting (-1.032*e*). Further, the authors used the protonated phosphate oxygen to accept one of the planar hydrogen bonds, while the present work used deprotonated oxygens to accept both hydrogen bonds. Additionally, because the authors were trying to mimic experiments involving ion pairs in high ionic strength (1 mol/L) aqueous solution, they applied an ionic shielding correction

to their ion pair calculations that weakened their interactions on the order of 1 kcal/mol. Finally, the authors used a generalized Born implicit solvent model instead of Poisson-Boltzmann and explicit solvent used here.

Some efforts have also aimed to quantify the strengths of phosphate interactions with charged side-chain analogs experimentally. Springs and Haake[85] extracted free energies of association for guanidinium–phosphate and butylamine–phosphate from pK_a shifts. The absolute free energies cannot be directly compared to the present work, because the experimental free energies are for free ions, whereas the computational PMFs represent constrained geometries, which is more appropriate to understanding hydrogen bonding in a macromolecule. In addition, the experiments were carried out in solution with 1 mol/L ionic strength, creating significant ionic shielding. However, the relative free energies can be profitably compared. Specifically, the experimentally determined free energies show a stronger interaction for guanidinium–phosphate (-0.6 kcal/mol) than butylamine–phosphate (-0.4 kcal/mol), in agreement with the results obtained here.

Salt bridge interactions of pAsp

Response regulators in bacterial 'two component' signaling systems use Asp side-chains to accept a phosphate group from a sensor histidine kinase[86]. Resonance structures involving the pi orbitals of the covalently linked carboxylate and phosphate groups suggest that the electron density on the phosphate group on pAsp

may be significantly different from that of pSer/pThr[a]. This conjecture is confirmed by the partial charges obtained by electrostatic potential fitting, as discussed in Methods. The effect of this difference on hydrogen bond strengths with the panel of hydrogen bond donors was examined next.

As with propionic acid (Glu) and methyl phosphate (pSer) acceptors, the energies for acetyl phosphate (pAsp) with propyl guanidinium (Arg) and butyl ammonium (Lys) in a collinear approach are very similar (within 1 kcal/mol). The more striking comparison arises from different charge states of pAsp. Deprotonating pAsp when accepting from collinear Lys stabilizes the interaction by -1.4 kcal/mol, and by -1.5 kcal/mol when the donor is Arg. In contrast, deprotonating pSer in a collinear salt bridge stabilizes the interaction by only -0.5 kcal/mol with Lys and -1.0 kcal/mol with Arg. The deprotonation effect increases in the coplanar approach only for lysine. Coplanar $pAsp^{-2}$–Lys shows a 2.5 kcal/mol stabilization over $pAsp^{-1}$–Lys, while coplanar $pAsp^{-2}$–Arg yields only a 1.0 kcal/mol stabilization over $pAsp^{-1}$–Arg. Bidentate hydrogen bonding continues to produce strong effects, with coplanar Arg showing a 3.3 kcal/mol stronger salt bridge than a collinear approach with $pAsp^{-1}$ and a 2.8 kcal/mol stronger interaction than collinear with $pAsp^{-2}$. Coplanar Arg also bonds stronger than coplanar Lys to pAsp in the -1 and -2 charge states by 3.9 kcal/mol and 2.4 kcal/mol respectively, as Lys cannot form bidentate bonds due to the same geometric constraints inhibiting them with Glu and pSer acceptors.

[a] A similar argument could be made regarding pTyr, *i.e.*, that there could be some conjugation between the pi electronic systems in the benzene ring and on the phosphate. However, quantum calculations on benzyl phosphate followed by electrostatic potential fitting (see Methods) suggested that the electron density on the phosphate group in pTyr is minimally different than in methyl phosphate (pSer), and this issue further was not pursued further.

The greater sensitivity to charge state of pAsp over pSer might be attributed to differences in the charge distribution for the -1 and -2 ions. The quantum mechanically calculated partial charges for the acceptor oxygens on pSer remain at -1.032*e* regardless of protonation of the remaining phosphate oxygen. In contrast, partial charges of the corresponding oxygens of pAsp decrease to -1.016*e* from -0.949*e* upon phosphate deprotonation. These differences in partial charges can be attributed to protonation effects on electron density due to resonance between the carboxylate and phosphate pi-electron systems; this effect of course does not occur in methyl phosphate (pSer).

Hydrogen bond interactions with amide NH groups

An earlier survey (data not shown) identified backbone amides as the second most common hydrogen bond partner after Arg with phosphates in phosphorylated proteins in the Protein Data Bank (PDB)[87]. N-methylacetamide (CH_3–NH–CO–CH_3) was employed as an analog of the protein backbone to investigate the free energy of backbone hydrogen bonds to the carboxylate and phosphate hydrogen bond acceptors. In all simulations the amide hydrogen was placed in a collinear geometry with its acceptor. The glutamate analog was truncated to acetate because hydrophobic interactions were observed between the aliphatic tail of Glu and the methyl caps of N-methylacetamide. In contrast to the behavior observed with Arg or Lys donors, the interaction weakens slightly when the acceptor is deprotonated. A weakened hydrogen bond arising from a stronger P–O dipole may appear counterintuitive. However, hydrogen bond formation depends on a delicate balance

between the free energy gain of bonded pairs and the loss of hydrogen bonds to surrounding waters, and the desolvation penalty is significantly lower for the protonated phosphate group. This effect was also observed by Wong *et al.*[58] in a study involving phosphate–amide interactions with phosphate acceptors possessing both -1 and -2 charges.

Implicit solvent Poisson-Boltzmann calculations

While continuum solvent models can provide substantial speed advantages relative to explicit solvent in performing free energy calculations, the merits and shortcomings of implicit solvent models remain a subject of interest and some contention. Treating the solvent as a dielectric continuum neglects important first-shell solvation effects arising from the finite size and asymmetry of a water molecule[35-37]. In particular, the very strong ionic interactions between a -2 charged phosphate with positively charged ions presents a challenging test of implicit solvent models.

Poisson-Boltzmann (PB) implicit solvent calculations were performed on all configurations (see 'Methods') to compare with the explicit solvent MD simulations. These calculations retained the same geometries, atomic radii, Lennard-Jones parameters, and partial charges as in the MD simulations. Figure 2.2 shows one example of a comparison between the explicit solvent PMF and the implicit solvent results. As has been seen in other work[32,35], the implicit solvent potentials generally contain less structure than the explicit solvent PMFs, i.e., no secondary minima, consistent with the fact that the implicit solvent model treats water as a continuum.

In addition, the implicit solvent results tend to exaggerate the energy barrier required for separating the ions from contact to infinite separation.

The primary concern here, however, is the depth of the contact minima (Table 2.2) and in this respect the implicit solvent results generally recapitulate most of the key trends observed in the explicit solvent PMFs. In particular, the implicit solvent results agree that protonating the phosphate group weakens hydrogen bonds with charged donors but strengthens interactions with the amide NH group, and that the strongest hydrogen bond of the phosphate group is the bidentate interaction with guanidinium. Overall, the implicit solvent calculations predict stronger hydrogen bonding interactions of the phosphate group than explicit solvent MD, with the largest discrepancies observed for Arg forming bidentate interactions with unprotonated phosphate. It may be possible to reduce the general over-prediction of the hydrogen bond strengths by empirically adjusting the radii used to define the dielectric surface in the implicit solvent calculation, but such an optimization has not been performed in this work. It should also be reiterated that the explicit solvent PMFs cannot be considered to be free of error either and will depend on the choice of explicit solvent model and other simulation parameters.

Table 2.2: Hydrogen bonding free energies computed from continuum electrostatics
Free energies are shown in kcal/mol.

	Lys Collinear	Arg Collinear	Lys Coplanar	Arg Coplanar	Amide NH Collinear
Glu	-5.1	-3.9	-2.7	-10.6	-2.3
pSer(-1)	-6.7	-6.1	-3.6	-13.0	-2.1
pSer(-2)	-7.7	-6.7	-5.6	-15.4	-1.2
pAsp(-1)	-5.5	-5.1	-2.4	-10.8	-2.6
pAsp(-2)	-7.2	-6.5	-5.8	-15.5	-1.3

Self-consistent reaction field quantum mechanics calculations

To investigate the possible effects of electronic polarization on the energetics of hydrogen bonding, quantum mechanics (QM) calculations using a self-consistent reaction field to mimic the condensed phase (see 'Methods') were employed. One limitation of this method with respect to the MD calculations is the use of implicit solvent. An advantage, however, is that atomic partial charges are recomputed at each distance to account for electronic polarizability. For instance, the donated Arg hydrogens in the $pSer^{-2}$–Arg coplanar configuration increase in charge from 0.54*e* and 0.51*e* at 11.0 Å separation to 0.69*e* and 0.62*e* at the contact minimum distance of 4.25 Å.

As with the PB analysis, the QM calculations were carried out on fixed orientations at 0.25 Å intervals. The total interaction energies, computed from the difference between potentials at 11.0 Å separation and the contact minima, are presented in Table 3. As observed in the explicit solvent MD results, bidentate hydrogen bonds with coplanar arginine tend to produce salt bridges about twice as strong as the monodentate collinear approach. On average, QM found 1.5 kcal/mol

stronger interactions than MD with a fixed charge force field and explicit solvent. The largest differences occur for configurations with a -2 charged receptor. The mean difference in free energy between QM and MD for -2 charged receptors is -3.1 kcal/mol, while the same figure for -1 charged receptors is -0.9 kcal/mol. It is of course reasonable that the larger charge on the -2 anions would induce larger polarization effects.

Since the quantum mechanical and PB calculations both employ an implicit solvent model, it is informative to compare trends between these methods as well. The average difference between QM and PB is -1.4 kcal/mol, similar to the difference observed when comparing to MD. As expected, the QM calculations are substantially more sensitive to polarization effects than PB. The mean free energy difference between QM and PB for -2 charged receptors is -3.0 kcal/mol, while for -1 charged receptors it is 1.0 kcal/mol.

Overall, the quantum mechanical calculations show a significant role for polarization in hydrogen bond stability, and suggest that the explicit solvent molecular dynamics simulations, using a fixed charge force field, might systematically underestimate the strengths of hydrogen bonds involving the phosphate group with a -2 charge, relative to -1 phosphate or carboxylate groups. Generally, the quantum mechanical calculations support the conclusions from the explicit solvent molecular dynamics. One key limitation of the quantum calculations, however, is that solvent is treated as a dielectric continuum, as in the implicit solvent results. Quantum mechanical simulations are possible with explicit solvent, but extensive sampling of the water is required to obtain reasonable free energies of

solvation, making this approach computationally extremely intensive. A more tractable way to assess electronic polarizability effects in explicit solvent may be to perform molecular dynamics using the new generation of polarizable force fields[88] with polarizable explicit water[89].

Table 2.3: Hydrogen bonding free energies computed from SCRF quantum mechanics
Free energies are shown in kcal/mol.

	Lys Collinear	Arg Collinear	Lys Coplanar	Arg Coplanar	Amide NH Collinear
Glu	-4.5	-3.7	-4.3	-10.2	-2.1
pSer(-1)	-4.4	-4.1	-2.5	-8.1	-0.8
pSer(-2)	-8.6	-6.8	-8.8	-13.6	-2.6
pAsp(-1)	-3.8	-3.5	-2.4	-8.2	-0.8
pAsp(-2)	-7.8	-6.4	-8.5	-12.6	-2.7

Conclusions

Calculating hydrogen bond free energies using multiple levels of theory has provided an internally consistent survey of hydrogen bond strengths for common hydrogen bonding partners, charge states, and geometries involving phosphorylated amino acid side-chains. Additionally, the results suggest relative merits and shortcomings of each level of theory for this application. The chapter concludes by returning to the issues raised in the introduction:

(1) *The conditions under which Arg or Lys make stronger hydrogen bonds with a phosphorylated side-chain.* Lys forms as strong or slightly stronger hydrogen bonds than Arg with most of the acceptors studied in the collinear approach, probably because the ε-amino group of lysine has a denser positive charge field

than the arginine guanidinium moiety. However, the results are unambiguous that the bidentate interactions available to guanidinium (Arg) with phosphate provide much stronger interactions than can be formed between ammonium ions (Lys) and phosphate in either a monodentate or bidentate geometry.

(2) *The effect of phosphate protonation state on hydrogen bond strength.* Phosphate protonation (i.e., to the -1 charge state) produces a small but significant destabilizing effect with Arg and Lys donors (~1 kcal/mol for pSer), particularly when the acceptor is pAsp (up to 2.5 kcal/mol). In contrast, the interactions with the amide NH group were mildly stabilized by acceptor protonation (~0.6 kcal/mol); this is consistent with previous work of Wong *et al*.[58]. Altogether, however, the hydrogen bond strengths of the phosphate groups in the -2 and -1 charge states are strikingly similar. This is remarkable because the hydrogen bond strength results largely from near-cancellation of two very large quantities: the strong Coulombic attraction between the ions, and the dielectric screening (and first-shell solvation effects) exerted by the water. Changing the charge state of the phosphate ion perturbs both of these quantities significantly, but apparently the changes are such that the overall strengths of the hydrogen bonds are not greatly affected.

(3) *The energetic consequences of substituting a carboxylate for a phosphate.* All of the $pSer^{-2}$ orientations with charged residue donors form stronger salt bridges than these charged residue donors do with glutamate suggesting that Asp/Glu

substitution might not mimic phosphorylation. In contrast, the strengths of the hydrogen bonds of the phosphate groups in the -1 charge state are generally closer to the corresponding hydrogen bonds of the carboxylate group, especially for $pAsp^{-1}$. Table 4 lists hydrogens bonds with phosphate acceptors that are at least 0.5 kcal/mol as strong when the acceptor is a carboxylate. Of course, in the protein microenvironment the local electrostatic field, steric restrictions, exposure to various solvent and ion concentrations, departure from ideal orientations, and other factors will significantly impact the hydrogen bond strengths computed here. Nevertheless, these calculations provide a quantitative framework for beginning to assess when substitution with Glu or Asp might mimic phosphorylation, at least when a protein structure is available, and suggest that the protonation state of the phosphate may be a critical parameter.

Table 2.4: Change in salt bridge free energy when a carboxylate is substituted for a phosphate

Charged–charged ion pairs with carboxylate acceptor substitutions worth at least 0.5 kcal/mol as with a phosphate acceptor are shown.

Ion Pair	Orientation	Predicted ΔΔG with Glu acceptor (kcal/mol)
pAsp(-1)–Lys	Collinear	0.1
pAsp(-1)–Lys	Coplanar	-0.2
pAsp(-1)–Arg	Collinear	-0.5
pAsp(-1)–Arg	Coplanar	-2.2
pAsp(-2)–Arg	Coplanar	-1.2
pSer(-1)–Arg	Collinear	0.3

Chapter 3

Approaches to modeling backbone flexibility in protein design

History of backbone flexibility in protein design

This chapter now turns from 'scoring' issues regarding the favorability of side-chain interactions to the 'sampling' problem of modeling backbone conformations in the context of protein design. Early approaches to incorporate backbone flexibility into design relied on well-characterized parameterizations of secondary structural elements[10,11]. These methods produced a striking success – the first experimentally verified designed novel protein topology, an α-helical tetramer by Harbury and co-workers[11] – but cannot easily be applied to protein structures lacking defined parametrical descriptions. Desjarlais and Handel[12] generalized the task by incorporating random phi, psi, and omega torsion moves into a flexible backbone design protocol. Their work showed the potential for flexible backbone methods to model conformational adjustments in response to mutations that significantly alter amino acid side-chain size, as measured by agreement to experimentally determined T4 lysozyme variant stabilities. In general, however, modeling backbone flexibility showed poorer correlation to the entire set of mutations than fixed

backbone calculations, consistent with observations from protein structure prediction that refinement and minimization techniques frequently move models further away from the target structure[90]. The authors noted that flexible backbone simulations improved agreement with experimental data only after applying a restraining potential based on wild-type structural features, and suggested that maintaining local structural motifs may be a key to successful designs.

Recognizing that preserving local features would provide a dramatic advantage in conformational sampling, Baker and colleagues suggested that short structural fragments encoded sufficient physical properties such that their careful assembly could significantly improve related efforts in *ab initio* protein tertiary structure prediction. Their 'fragment insertion' method, wherein stretches of successive backbone torsions are set to values from other crystal structure fragments with similar sequence and secondary structure, was automated in the program Rosetta[91] and used in early Critical Assessment of Structure Prediction (CASP) experiments with some success[92]. However, while global backbone remodeling by fragment insertion could reproduce course-grained tertiary features, higher-resolution structure prediction and sequence design required compensatory torsion adjustments to residues adjacent or distal to the insertion site to help localize backbone moves[93]. Such fragment insertion techniques with compensating torsion moves enabled the seminal design of a globular protein fold not found in naturally occurring proteins[13]. This engineering success thus highlights how advances in structure prediction[94] contribute to efforts in flexible backbone protein

design, and *vice versa*, as protein design and structure prediction can be seen as inverse problems[95].

Recent approaches to backbone flexibility in protein design

Attempts to insert fragments into polypeptide chains are frequently rejected – even after compensation – because the altered torsions can lead to significant non-local structural perturbations. To address this problem, Wang and colleagues[96] aimed to localize backbone changes by following fragment insertions with a numerical peptide closure technique termed cyclic coordinate descent (CCD)[97]. This method optimizes successive backbones torsions until the break in the peptide chain resulting from the fragment insertion falls below a set threshold. Hu and colleagues[98] iterated sequence design with structure prediction by this method and designed three tenascin loop sequences that were experimentally verified to form stable folded protein structures; one predicted loop conformation matched the determined crystal structure to sub-angstrom accuracy. Murphy and co-workers[99] combined fragment insertion with CCD closure together with active site side-chain constraints to redesign a loop in human guanine deaminase with non-native length that induced a 2.5 x 10^6-fold substrate specificity switch from guanine to ammelide. An X-ray crystallographic structure of the redesigned loop was within 1 Å Cα rmsd (root mean squared deviation) of the computational design.

The methods of Kuhlman, Hu, Murphy and colleagues highlight an unresolved issue in flexible backbone protein design: when iterating sequence design and structure prediction, it is often unclear how many iterations of each

stage are optimal. One solution is to employ methods that simultaneously optimize backbone conformations and sequences. Fung and colleagues[100] developed an integer linear programming model to calculate optimal sequences for a set of backbone templates extracted from sources such as molecular dynamics trajectories. The sequence selection model includes an energy term accounting for atomic distance statistics observed in the templates. The best scoring sequences are ranked by specificity to the target fold by generating structural ensembles of the designed sequences using a constrained simulated annealing protocol, and comparing them to an ensemble produced by the same method but applied to the native sequence. The authors employed this framework to predict sequences for human β-defensin 2 and showed that the dominant clusters of top scoring designs from different flavors of their methodology comprised a subset of the sequence variation observed in some homologs of that protein.

Georgiev and Donald[101] noted that the approximate force field and reduced side-chain representations of Fung and co-workers may result in sequences with high energies. Instead, they introduced a flexible backbone design method with all-atom side-chain detail based on dead-end elimination (DEE)[102], a search method for finding the global minimum energy sequence and conformation. Georgiev and Donald extended existing fixed-backbone DEE to consider backbone flexibility by varying torsion angles continuously over a range that confined residues to a predefined volume. The authors proved that, within this range, the DEE method exclusively prunes conformations that are not part of the global energy minimum. Applied to the β1 domain of protein G, their method found sequence mutations with

lower predicted energies than fixed backbone DEE without side-chain energy minimization, and another version of their method[103] that performs fixed backbone DEE with side-chain minimization.

In the work by Georgiev and co-workers, the implementation of backbone flexibility perturbs all residues N-terminal to an adjusted phi torsion, or C-terminal to an adjusted psi torsion. Thus, non-local effects propagate through the backbone, as described earlier for fragment insertion-based methods. In contrast, purely local backbone moves may have important advantages for sampling efficiency. Inspiration for such local perturbations came from a naturally occurring backbone move characterized by Davis and colleagues[104] when inspecting high-resolution crystal structures. This so-called 'backrub' motion applies to tripeptide segments, where the atoms between the first and third Cα atoms of the segment rotate about the axis connecting those Cα atoms. Additionally, the atoms between the first and second Cα atoms, and those between the second and third Cα atoms, rotate about their respective axes to relieve bond angle strain induced by the primary rotation. Davis and co-workers showed that this move enabled side-chain fitting into alternate electron densities observed in ultra-high-resolution crystal structures. This result suggests that the local backrub motion may describe biologically relevant correlated movements of side-chains and the corresponding backbone. Georgiev and colleagues[105] integrated an automated backrub mover into a DEE protocol for flexible backbone protein design, where backrub moves provide a small, discrete backbone ensemble for residues under design. They found that DEE with backrub applied to non-adjacent tripeptides produced more sequences and

better energies than fixed backbone DEE, although the earlier non-local flexible backbone method found a larger number of sequences with even lower energies.

In parallel to these efforts, Smith and Kortemme[106] implemented a backrub-like move in the Rosetta framework. The 'generalized backrub' applies to peptide segments (typically 2-12 residues in length) and rotates the atoms internal to the first and last Cα atoms about the virtual axis between those Cα atoms. Since the primary backbone rotation comes at the expense of N-Cα-C bond angle strain at the hinge residues, Smith and Kortemme introduced into Rosetta a molecular mechanics-based bond angle potential. Additionally, they developed quadratic equations closely approximating the molecular mechanical relationship between torsions and bond angles involving hinge Cβ and Hα atoms to optimize placement of those atoms after the backrub rotation. After demonstrating that the move successfully models many of the backrub-enabled side-chain fluctuations observed by Davis and colleagues[104], the authors predicted backbone and side-chain perturbations observed in a large dataset of point mutations (taken from[16]). In the majority of well-packed cases, the lowest energy backrub structure moved the mutated side-chain closer to the mutant crystallographic conformation than the fixed backbone prediction.

Applications recently enabled by flexible backbone methods

Given the significant advances in sampling and evaluating backbone conformations – and their successful integration into computational design frameworks – I now turn to biological applications enabled by recent developments in flexible backbone

design. As noted in the introduction, flexible backbone methods are needed to accurately model the structural adjustments observed in proteins in response to sequence mutations. Encouragingly, recently developed tools to predict changes in protein energetics arising from point mutations have shown improved agreement to experimentally observed changes in free energies by accounting for backbone adjustments to varying degrees. Yin and co-workers[107] showed good correlation to experimentally determined changes in free energy of folding (R = 0.75) by relaxing backbone torsions through conjugate gradient descent while modeling mutations. Benedix and colleagues [108] predicted changes in free energies of folding and binding by designing mutants against structural ensembles produced by a method that generates random structural variations satisfying experimentally observed interatomic distance constraints[109]. This approach yielded a similar correlation to experimentally observed free energies of folding as Yin and co-workers while reducing the standard deviation, although comparable results were obtained using a fixed backbone method[110].

Not all sequence mutations significantly affect protein structure and function. Instead, many protein folds 'tolerate' a considerable number of mutations[111,112]. Proteins exploit this robustness[113] to accumulate sequence changes allowing them to vary and expand their functional repertoire[114] while preserving sufficient stability. Computational methods that reproduce the tolerated sequences of protein folds might therefore improve not only our understanding of the relationship between sequence, structure, and function[115] but also our ability to engineer proteins with modified and new properties. For example, predicting a number of

low-energy sequences tolerated by a protein fold can suggest sequence mutations to increase selectivity for a particular binding partner. To this end, Fu and colleagues[116] designed helical peptides against an ensemble of backbones – generated by normal mode analysis (NMA) of helices in the PDB – targeting the protein Bcl-xL. Several designed peptides showed selective binding to Bcl-xL in pull-down assays while binding more weakly to other Bcl family members. Notably, sequences designed against conformations arising from NMA applied to the native binding peptide showed greater affinity for Bcl-xL than sequences generated on conformations from NMA on an idealized helix, highlighting challenges in *de novo* ligand design where no native template is available.

While generating large sets of designed sequences by flexible backbone methods can improve the utility of computational design, the difficulty of validating the broadened range of predictions increases commensurately. One way to assess sequences computationally designed for a protein fold is to compare the diversity obtained to the natural family of protein homologs. If computationally generated sequence libraries reasonably reproduce the tolerated sequence profiles of protein folds, statistics on residue frequencies appearing in computational simulations could be used to significantly reduce the sequence space that must be sampled in screening experiments[117]. Larson and colleagues[118] applied flexible backbone methods to explore the correlation of variability in computationally designed sequence libraries to natural protein families. The authors found that designing on an ensemble of backbones – generated by random phi/psi moves as described by Desjarlais and Handel[12] – increased the diversity of predicted sequences compared

to a fixed backbone method. PSI-BLAST searches showed similarities between designed sequences and natural family members. Subsequently, design methods were also applied to aid homology modeling. Here, designed sequences from fixed[119] and flexible[120] backbone predictions generated multiple sequence alignments to detect remote homologs, thereby providing initial template structures as starting points for comparative modeling. Following the work of Larson and co-workers, Saunders and Baker[121] used the Rosetta iterative flexible backbone design method[13] to explore the agreement of sequences generated by their fragment-based method to natural protein families. They concluded that the Rosetta protocol better recapitulated naturally occurring sequence variation than a fixed backbone approach, as well as a random phi/psi perturbation protocol. None of the methods, however, produced sequences very similar to the natural families, and the authors speculated that the lack of functional and evolutionary constraints in the simulations makes this problem particularly challenging.

Friedland and colleagues[122] extended the analysis of sequence and structure variation to solution-state dynamics. They showed that a backrub model parameterized to produce a structural ensemble of ubiquitin with good agreement to residual dipolar coupling experiments also produced sequence variation capturing a superset of the sequence diversity observed in naturally occurring ubquitin homologs. Ding and Dokholyan[123] explored the relationship between sequence identity and conformational plasticity by examining the similarity of sequences designed on structural ensembles generated by discrete molecular dynamics simulations. The authors found some correlation between the

computationally designed sequences and 3 natural families, with correlation coefficients between computational and experimental sequence entropies ranging from 0.15 to 0.46, which increased to 0.23 to 0.62 when functional residues were excluded.

The mild correlations described above underscore the challenge of validating computationally designed sequences by comparison to naturally occurring protein families that may have been subject to a range of known and unknown functional constraints during evolution. Experimental selection techniques can produce large-scale libraries of protein sequences subject to some prescribed functional constraints, such as catalytic turnover and substrate binding. Humphris and Kortemme[124] compared sequences designed against an ensemble of backbones produced by the backrub method described above[106] to sequences obtained from comprehensive phage display experiments[125] scanning the human growth hormone (hGH) interface against its receptor (hGHR). Notably, the experimentally observed hGH sequences must bind hGHR, in addition to folding into stable structures. The computational method explicitly models these constraints by designing the hGH interface in complex with hGHR. The authors found that the method was able to capture 92% of the experimentally observed sequence diversity at the 35 sequence positions in the hGH interface. Moreover, at several positions in the interface, the flexible backbone method predicted experimentally observed residues that were not found with a fixed backbone model. A computationally designed library of 6×10^9 sequences generated by this method (out of a possible 6×10^{44} sequences for the

entire 35 position interface, excluding cysteines) covered on average 50% of the experimentally observed sequence space at each position.

A role for robotics

While there are some differences in the backrub-inspired approaches described above – Georgiev and co-workers' technique works on non-overlapping tripeptides with DEE, whereas the generalized backrub can rotate larger peptide segments in a Monte Carlo protocol – in general backrub strategies are most appropriate when the desired structural perturbations are relatively small. However, functional regions in proteins can display considerably greater flexibility through variations in backbone torsion angles. For example, alternative structures of interface loops in protein switches suggest that effective modeling and design methods must capture conformational changes on the order of several angstroms. Given that backbone torsions are substantially more variable than bond angles, conformational sampling might be improved by local techniques operating on *all* backbone torsions in a peptide segment while keeping bond lengths and angles at or near ideal values.

Commonly referred to as 'loop closure', Go and Scheraga[126] first applied these concepts to biomolecules with a numerical formulation for determining allowable values for 6 torsions of peptide chains – specifically, tripeptides and cyclic peptides. Subsequent closure techniques [127-131] find conformations for longer chains using related ideas from inverse kinematics, a subfield of robotics. Finding accessible conformations of linked objects subject to constraints has been well-studied in inverse kinematics, such as determining the possible rotations of the

internal joints of a robotic arm that satisfy fixed positions for the shoulder and hand (Figure 3.1a). The 'kinematic closure' method of Coutsias and colleagues[127] analytically determines conformations for peptide chains of any length by varying 6 torsions and keeping the remaining torsions, bond lengths, and bond angles fixed. This method was extended by techniques based on polynomial resultants[132] to enable sampling of interior torsions, bond angles, and bond lengths while solving the complete ensemble of values for the 6 closure torsions (Figure 3.1b). The remaining chapters of this dissertation will describe the adoption of kinematic closure to modeling regions of proteins lacking secondary structure, and to designing proteins with new functions.

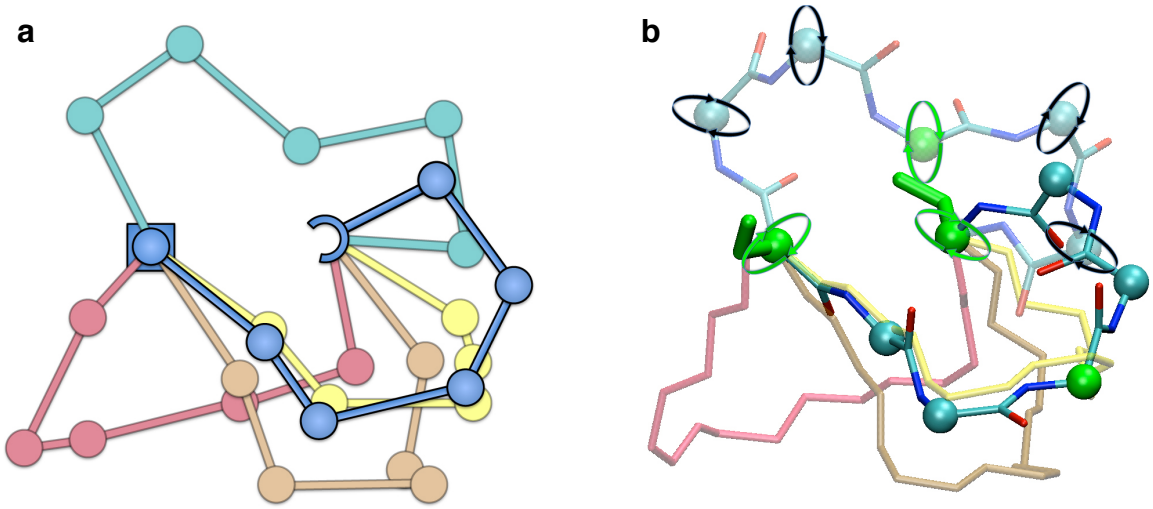

Figure 3.1: A robotics inspired method for modeling protein conformations
(**a**) Kinematic closure (KIC) on a robotic arm with seven rotatable joints (analogous to a seven-residue peptide) places the robotic hand at the same position across five conformations. The lengths of the linkages between the joints remain constant (analogous to bond lengths). The angles formed by two successive linkages are also maintained (analogous to bond angles). The conformations arise by rotating the linkages around the joints (analogous to backbone torsions). (**b**) KIC on a peptide chain (shown as transparent) following the same constraints as in **a**. Three Cα atoms are designated as pivots (green spheres). The torsions of the four non-pivot Cα atoms (black arrows) are set to values drawn probabilistically from the Ramachandran map for each residue type. KIC then finds values for the pivot torsions (green arrows) that close the chain while maintaining prescribed values for the bond lengths and bond angles. Three solutions (red, brown, and yellow) have pivot torsions in disallowed Ramachandran regions, or have steric overlap. One solution (opaque with spheres) satisfies these physical criteria. All solutions are found in a single KIC calculation.

Chapter 4

A robotics-inspired approach to modeling flexible regions of proteins and protein interfaces

Introduction

Proteins exploit the conformational variability of regions lacking secondary structure termed 'loops' to carry out diverse biological tasks including molecular recognition and signal transduction. New algorithms to engineer these functions by combining loop building and sequence design therefore have enormous practical applications, but require high-resolution *loop reconstruction*: the modeling of protein loop conformations given amino acid sequences. Loop reconstruction in protein design may be simplified conceptually by restricting changes to the functional loop regions. However, despite significant progress in loop prediction[133,134], design applications are limited by the difficulty to model purely local conformational moves and by the need for advances in sampling and evaluating loop conformations.

Here I address these challenges with a robotics-inspired local loop reconstruction method for peptide chains, called kinematic closure (KIC).

Calculating the accessible conformations of objects subject to constraints, such as determining the possible positions of the interior joints of a robot arm given fixed positions for the shoulder and fingertips, has been well-studied in inverse kinematics, a subfield of robotics. Building on the first[126] and subsequent[127-131,135-137] applications of kinematics to proteins, the KIC method presented here provides the key advantages of analytically determining all mechanically accessible conformations for 6 torsions of a peptide chain of any length, while simultaneously sampling the remaining torsions and N-Cα-C bond angles using polynomial resultants[138] (Figure 4.1a, Figure 4.4).

To enable a range of applications, I coupled KIC to the Rosetta method for protein structure modeling[95]. The loop reconstruction protocol iterates KIC calculations as Monte Carlo moves first with loop backbone minimization in a low-resolution stage, in which side-chains are represented as centroids, and then in a high-resolution all-atom stage with minimization of the loop backbone and all side-chains in the loop environment (Figure 4.3). At the beginning of each KIC simulation, I discard all native loop bond lengths, bond angles, and torsions. In addition, I perform reconstructions without knowledge of native side-chain conformations in both the loop and the protein scaffold (see 'Methods'), which makes prediction substantially more challenging, but broadens the range of applications to designing new loop conformations that may interact differently with neighboring side-chains.

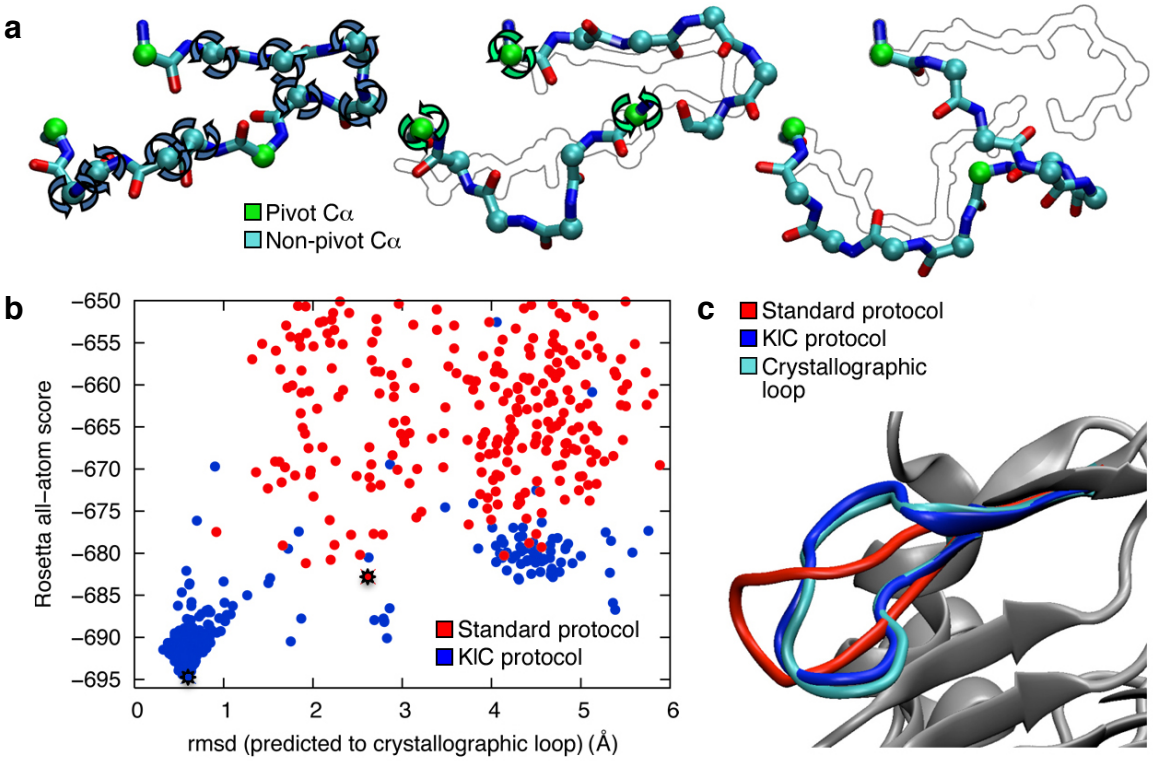

Figure 4.1: Loop reconstruction with KIC
(**a**) In the KIC move, 3 Cα atoms of an *N*-residue chain are designated as pivots (green spheres); the remaining *N* – 3 are non-pivot Cα atoms (cyan spheres; left). In a 12-residue loop, 24 torsions are modeled. Non-pivot torsions are sampled from a residue type-specific Ramachandran map, opening the chain (middle). KIC then finds all values for the pivot torsions that close the loop, if any exist, keeping the endpoints fixed (right). The previous state is shown in outline. (**b**) Performance of the Rosetta KIC protocol and standard protocols on a 12-residue loop (PDB 1srp). Only KIC densely sampled regions <1.0 Å rmsd from the crystallographic loop. Asterisks mark the lowest-scoring reconstructions from the two methods. The Rosetta all-atom score includes the enthalpy plus the solvation contribution to the entropy but not the configurational entropy. (**c**) The lowest scoring reconstructions from **b** are shown. KIC improved reconstruction accuracy to 0.6 Å from 2.6 Å using the standard protocol.

Results

I found that KIC substantially improves model accuracy over the standard loop building method in Rosetta, which combines insertion of torsion segments from homologous proteins and a numerical closure technique[136]. I generated 1,000 models by KIC, and compared the performance to the standard Rosetta method with the same number of Monte Carlo steps on 25 12-residue protein loops (dataset

1[139]). For each protein, I computed the root mean squared deviation (rmsd) of the backbone atoms of the best scoring loop model to the crystallographic loop, after superimposing the non-loop regions of the model onto the crystal structure. The KIC protocol frequently sampled regions of conformational space that were <1.0 Å from the crystallographic loop, which were not sampled by the standard protocol (Figure 4.1b). In the majority of cases (15/25), these conformations very close to the crystallographic loop could be identified as the best scoring models (Figure 4.1b,c). Over the entire 25-loop set, KIC improved the median accuracy to 0.8 Å rmsd from 2.0 Å rmsd when I applied the standard method (Figure 4.2b, Table 4.1). Since both methods use the same scoring function, these results suggest that KIC increases accuracy by improved conformational sampling (although sampling and scoring errors cannot be considered entirely independently as scoring guides the simulation trajectories; see 'Discussion' for further analysis of method performance and error sources).

To compare KIC loop reconstruction directly to the state-of-the-art molecular mechanics method[133], I applied the Rosetta KIC and standard protocols to the same 20 12-residue starting structures with perturbed loops and side-chain environments used to assess the molecular mechanics method (dataset 2[133]; Figure 4.2a). The Rosetta KIC protocol improved median accuracy to 0.9 Å from 1.2 Å using the molecular mechanics method and from 2.0 Å using the standard Rosetta method (Figure 4.2b, Table 4.2).

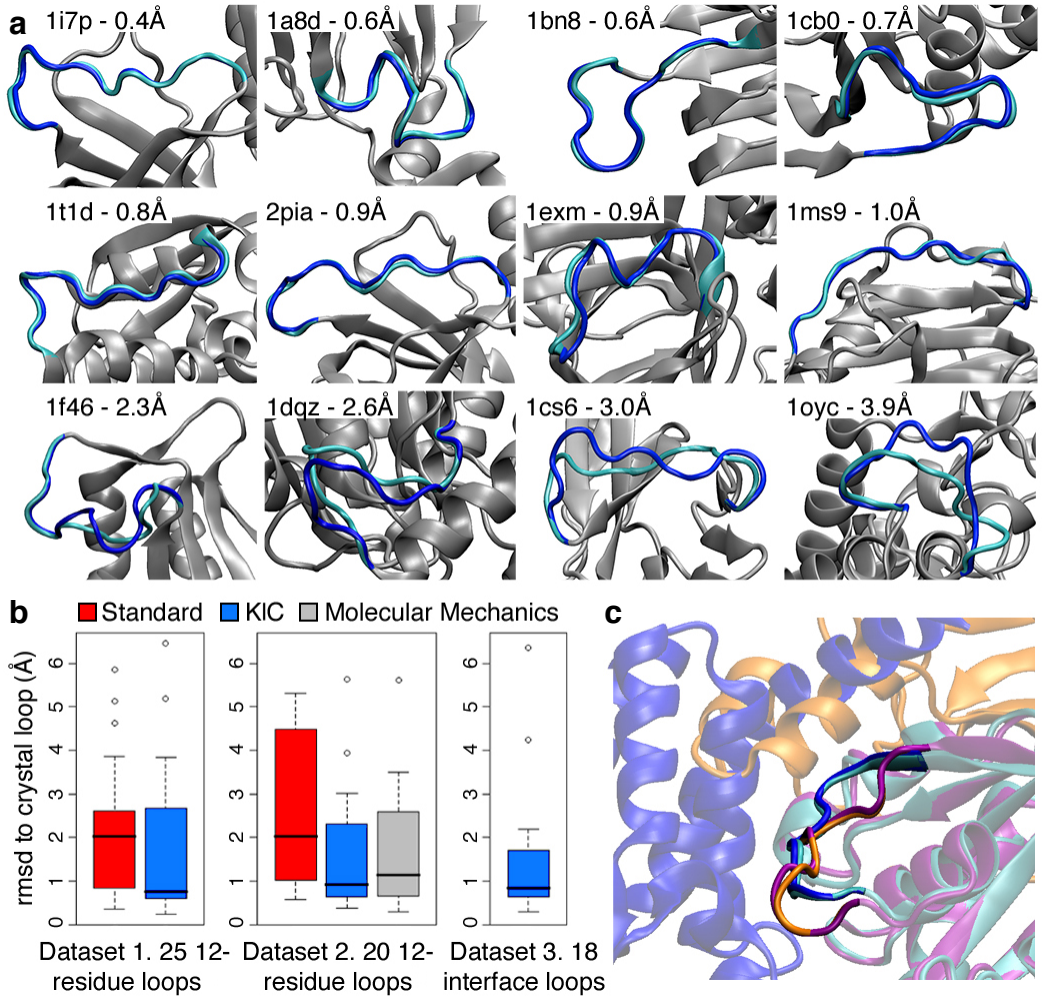

Figure 4.2: Performance of the KIC loop reconstruction protocol
(**a**) Representative set of 12-residue loop reconstructions (blue). PDB identifiers and rmsd to the crystallographic loop (cyan) are shown. (**b**) Box-plot comparison of the standard Rosetta and KIC Rosetta protocols on dataset 1 (left), both Rosetta protocols with the molecular mechanics method on dataset 2 (middle), and the KIC Rosetta protocol on dataset 3 (right). Boxes span the interquartile range (IQR, 25th–75th percentiles), black lines represent the median, whiskers extend to furthest values within 0.8 times the IQR, and open circles are outliers. (**c**) KIC reconstruction of conformational changes in the Rac switch I loop when bound to ExoS toxin (blue reconstruction on cyan crystal structure, blue partner; PDB 1he1) or Rho guanine dissociation inhibitor (orange reconstruction on purple crystal structure, orange partner; PDB 1hh4).

Functional loops in signaling proteins in complex with their partners exhibit conformational plasticity against a relatively structured core. To assess the ability of KIC to model such regions, I applied the method to interface loops from 4 proteins crystallized with 18 different partners (dataset 3). KIC reconstructed the loops to

0.8 Å median rmsd (Figure 4.2b). Notably, the KIC protocol produced high-accuracy reconstructions of the same switch protein loop adopting different conformations when bound to different partners (Figure 4.2c, Table 4.3). This result highlights the potential of KIC for modeling functional conformational changes. Sub-angstrom loop reconstructions by the local robotics-inspired sampling protocol described here could now be coupled with the Rosetta design method[95] to model and engineer protein loops precisely matching a particular binding partner, creating highly selective protein interfaces. Chapter 5 describes such an approach.

Methods

Datasets

I used two independent benchmark datasets for loops in monomeric proteins: dataset 1, a set of 40 12-residue loops originally compiled by Fiser *et al.*[140], and later studied by Rohl *et al.*[141] and Wang *et al.*[96], to facilitate comparison to previous work using the Rosetta loop modeling methodology, and dataset 2, a set of 20 12-residue loops compiled by Zhu *et al.*[142] to allow direct comparison to studies by Jacobson *et al.* [143], Zhu *et al.*[142], and Sellers *et al.*[133] The latter dataset was selected from high quality structures (resolution ≤ 2.0Å, R < 0.25) for loops with diverse sequences (<40% sequence identity), low temperature factors (<35), lack of contacts to heteroatom groups (>4Å for neutral ligands, >6.5Å for metal ions), lack of secondary structure within the loop, lack of more than 4 loop residues adjacent to either loop endpoint, and pH 6.5 – 7.5. The monomer loop datasets are shown in Table 4.1 and Table 4.2, respectively. Dataset 1 contained 15 loops with neutral ligands or charged

ions within contact distance of the loop, using the criteria specified by dataset 2, so while these loops are included in Table 4.1 they are separated from the 'filtered' dataset used for most subsequent analyses. Dataset 2 was simulated in two ways, first by the '*de novo'* method used on the Rosetta dataset, where KIC is used to place the loop into a random starting conformation, and second, by the 'perturbed' method, where the perturbed loops used in the simulations by Sellers *et al*.[133] were obtained from that group's website[144] and served as starting conformations for the Rosetta simulations. The perturbed approach was used to enable direct comparison between the Rosetta and molecular mechanics methods, since the degree of initial backbone perturbation will influence the degree to which the side-chain environment is perturbed. The '*de novo*' and 'perturbed' columns of Table 4.2 refer to this distinction.

A third independent dataset (dataset 3) was compiled to assess loop reconstruction of the same protein crystallized in complex with different partners (Table 4.3). This dataset contains 4 proteins (Rac, Ras, Cdc42, ubiquitin) crystallized with 18 different partners where the interface contains a loop that changes conformation across partners. For each of the four proteins, the loop regions to be reconstructed were defined by consecutive residues that contained any heavy atoms that were within 7Å of the binding partner in any crystal structure, and that lacked secondary structure in one or more of the crystal structures of that protein (7 residues minimum). Thus, the loop definitions were the same across complexes of the same protein, facilitating assessment of reconstruction accuracy with different

partners. Nucleotides and metal co-factors were modeled explicitly, and GDP-aluminum fluoride was modeled as GTP.

Structure preparation

Structures were prepared by first discarding all native side-chain information (including side-chain bond lengths, bond angles, and chi angles) and replacing them with rotameric conformations from the Dunbrack backbone-dependent rotamer library[145] and ideal bond lengths and angles; these rotamers were then simultaneously optimized by Metropolis Monte Carlo (MC) simulated annealing ('*repacking*') using Rosetta, as described in reference[13]. Each side-chain was then independently optimized by replacing it with the lowest energy conformation from the Dunbrack library and iterating through all positions until convergence was reached ('*rotamer trials*'). These procedures were followed by quasi-Newton all-atom energy minimization using the Davidon-Fletcher-Powell method[146] (*DFPmin*) on the loop backbone and side-chains within 10Å of the loop. The repacked, energy minimized structures served as input to the loop modeling protocol, which is depicted in Figure 4.3 and described below.

Loop modeling protocol

Loop endpoints for protein monomers were defined as in references[96,133] and shown in Table 4.1 and Table 4.2, and loop endpoints for the complexes set were defined as above (see 'Datasets') and shown in Table 4.3. The simulation proceeds through two stages of MC simulated annealing, as shown in Figure 4.3. In the first, low-resolution

stage, all side-chains are represented as centroids for coarse-grained conformational sampling. An initial KIC move is performed on the entire loop to place it into a non-native starting conformation with randomly chosen phi and psi torsion angles at non-pivot residues and phi/psi torsion angles at pivot residues determined by the kinematic closure algorithm (see 'Kinematic closure', below). During this step, native phi and psi torsions in the loop region are discarded, and bond lengths, bond angles, and omega torsions are set to ideal values. The 720 simulated annealing MC steps consist of applying KIC to a random subsegment of the loop region of length 3 to *N* (for an *N* residue loop). KIC moves are followed by line minimization of backbone torsions. The new conformation is scored and accepted or rejected by the Metropolis criterion. In the centroid stage the temperature decays exponentially from 2.0 *k*T to 1.0 *k*T, where *k* is Boltzmann's constant. The lowest energy conformation proceeds to the high-resolution all-atom stage. The repacked, minimized side-chains from the input conformation (see 'Structure preparation') are restored and those in the loop and on the surrounding scaffold with any heavy atoms within 10 Å of the new loop conformation are then repacked and subject to rotamer trials. If the loop is part of an interface (i.e., on dataset 3), side-chains from the binding partner within 10 Å of the loop are optimized as well. Relaxing the neighboring side-chains around a non-native loop conformation has the effect of starting the full-atom stage in a perturbed side-chain environment. This step makes loop reconstruction considerably more difficult, since neighboring side-chain conformations must be sampled and evaluated in addition to the loop side-chains and backbone conformations. The utility increases, however,

because in many applications (e.g., homology modeling, interface redesign) it cannot be assumed that the neighboring side-chain conformations are known *a priori*. I compared the standard and KIC Rosetta results to the method presented by Sellers *et al*.[133] that also reconstructs loops in a perturbed side-chain environment. I note that applications to comparative modeling may be even more challenging, as the loop endpoints and surrounding backbones can also be substantially perturbed, which I do not consider here. This does not preclude the application of KIC in high-resolution refinement and comparative modeling, as shown by a successful example of using the Rosetta KIC method in the most recent CASP experiment[147].

The 720 MC steps of the high-resolution stage consist of kinematic closure on random subsegments of the loop region, with one exponential simulated annealing cycle from 1.5 *k*T to 0.5 *k*T. In this high-resolution stage, KIC is followed by side-chain repacking (every 20 steps) and rotamer trials within 10 Å of the new loop conformation, and DFPmin on the loop backbone and side-chains within 10 Å of the new loop conformation. The lowest energy conformation explored during the high-resolution stage is recorded. The protocol is then iterated, and may be run over multiple processors in parallel. Reported loop reconstructions represent the lowest energy structure out of 1000 separate simulations (Figure 4.3), costing an average of ~320 CPU-hours per protein on a single 2.2 GHz Opteron processor. Each simulation trajectory is independent from the others, so they may be parallelized to dramatically speed up the protocol (up to one CPU-core per trajectory requiring less than 20 minutes per protein on average). Datasets 1 and 2 were simulated with Rosetta revision 24219, and dataset 3 was simulated with revision 27114.

Command line loop modeling options for datasets 1 and 2 were

```
-loops::kinematic -loops::nonpivot_torsion_sampling -in::file::fullatom -
out::file::fullatom -ex1aro -ex1 -ex2.
```

For dataset 3, the loop modeling options used were

```
-loops::remodel perturb_alc -loops::refine refine_alc -in::file::extra_res_fa -
in::file::extra_res_cen -in::file::fullatom -out::file::fullatom -loops::strict_loops -
ex1aro -ex1 -ex2.
```

Descriptions of all command line options used in this dissertation are given in the Appendix.

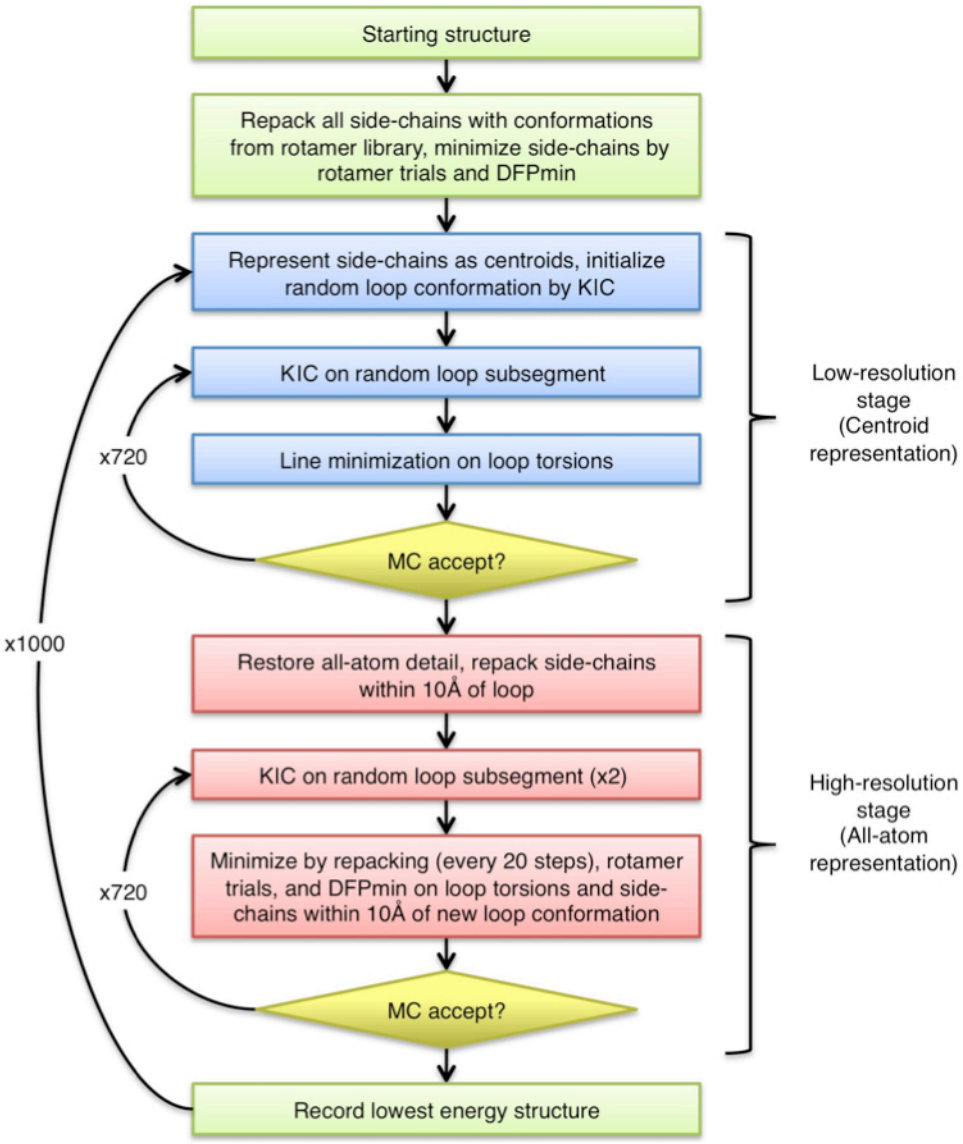

Figure 4.3: The Rosetta KIC loop reconstruction protocol

Kinematic closure

The atomic coordinates of the backbone atoms (N, Cα, C) of a random loop sub-segment of length 3 to *N* (for a loop of *N* residues) are supplied to the kinematic solver. The Cα atoms of the first, middle, and last residues are designated as pivots, and the remaining *N*-3 Cα atoms are designated as non-pivots. Torsions for each non-pivot Cα are sampled according to the Ramachandran probabilities for the residue type, and N-Cα-C bond angles are set to random values within one-half the standard deviation (σ = 2.48˚) above and below the mean (110.86˚) observed in ultra-high-resolution crystal structures (<1.0 Å resolution) in the PDB. This step effectively *opens* the loop segment at the pivots, breaking the continuity of the peptide chain. To close the loop, the kinematic solver finds values of the six pivot torsions for which the perturbed segments may be rejoined to form a new closed loop. As discussed in the next section (Polynomial resultants), there may be up to sixteen sets of such solutions, or none. Solutions are randomly applied to the loop segment until two filters are passed. The first filter computes the Rosetta Ramachandran score, which is a statistical potential derived from a smoothed, highly flattened version of the residue- and secondary structure-specific frequency with which a given (phi/psi) pair occurs in a set of high-resolution crystal structures[93], and accepts or rejects the conformation by the Metropolis criterion. The second filter is a backbone steric screen that ensures the distance between loop backbone atoms (N, Cα, C, O, and Cβ if not glycine) and all other backbone atoms is greater than the sum of the Lennard-Jones radii of the atoms times an overlap factor

(set to 0.7). The accepted solution is returned to the protocol for minimization and scoring. If no solution passes the filters, new values for the non-pivot torsions and N-Cα-C bond angles are drawn and closure is attempted again. Closure calculations execute 2,000 times per second on a 1.8 GHz Opteron processor.

Kinematic techniques were first applied to proteins[126] by calculating the accessible torsion angles of tripeptides with fixed bond angles, bond lengths, and endpoints. Other kinematics-inspired approaches have since been used in protein modeling[97,127-131,135,137]. Applications have included calculating conformations of cyclic peptides[126], exploring loop motions in one protein test case[127,128], and correlating loop models with spectroscopic observables from nuclear magnetic resonance experiments like order parameters and residual dipolar couplings in two proteins[137]. These methods have not been tested on large datasets on the problem of loop reconstruction, and each of these methods has lacked an analytical solution[97,126,129,137], has been applicable only to tripeptides or required consecutive pivot residues[126,127,135], or has not been coupled to a full-atom energy function[97,126,128-131,135].

Polynomial resultants

The details of the geometric steps taken by the algorithm are given in reference[132]. The construction proceeds by identifying 3 atoms before the N-terminus of the loop, and 3 atoms after the C-terminus. These two triads are assumed to have known positions in space. Together, they constitute the anchoring hinges for the two ends of the loop. They are denoted h_1 and h_2 (Figure 4.4a). The loop atoms are augmented

by the hinge atoms. Together they form the extended loop, which on the outset is considered to be in an extended conformation with all bond lengths and bond angles set to canonical values and all torsions set to 180.0 degrees. Three nonconsecutive atoms (not on the hinges), indexed p_1, p_2, p_3 with $p_1 + 2 \leq p_2 \leq p_3 - 2$ are chosen as the pivots for loop closure, and the loop is partitioned into four fragments: (1) $F_{3,b}$ including atoms from $h_{1,1}$ (the first atom of h_1) to p_1; (2) F_1 including atoms from p_1 to p_2; (3) F_2 including atoms from p_2 to p_3; and (4) $F_{3,a}$ including atoms from p_3 to $h_{2,3}$ (the third atom of h_3) (Figure 4.4b). Next, the four fragments thus defined are constructed using prescribed values for all their internal degrees of freedom (bond lengths, bond angles, and torsions). Arbitrary values can be chosen. At this stage, the bond angles at the three pivot atoms and the torsions about the bonds adjacent to pivot atoms (i.e., the 'pivot bond angles' and the 'pivot torsions') are not defined. Since the two hinges are anchored to the (known) rest of the molecule and thus have known absolute positions in space, the fragments $F_{3,a}, F_{3,b}$ are thus constructed with known positions in space for all their atoms relative to the hinges (and thus to the rest of the molecule). Their end atoms ($p_3, p_3 + 1, p_3 + 2, p_1 - 2, p_1 - 1, p_1$) are now fixed in space (Figure 4.4c).

The other two fragments, F_1 and F_2 are completely determined in their own body frames (Figure 4.4d), but their placement relative to the molecule is still to be determined. Each fragment is characterized by certain geometrical quantities that will enter as parameters in the loop closure equations. Referring to Figure 4.4e these are: (1) ξ_i, the angle formed by atoms $(p_i + 1, p_i, p_{i+1})$; (2) η_i, the angle

$(p_i, p_{i+1}, p_{i+1}-1)$; (3) d_i, the virtual bond length (p_i, p_{i+1}); and (4) δ_i, the dihedral angle $(p_i+1, p_i, p_{i+1}, p_{i+1}-1)$. The Rosetta implementation uses virtual segments composed from only the first and last triads of atoms in each segment, avoiding unnecessary reconstructions. These virtual segments must be assembled into a closed triangle (Figure 4.4f), provided the three lengths d_1, d_2, d_3 satisfy the triangle inequalities. If the triangle can be constructed, the three exterior angles $\alpha_1, \alpha_2, \alpha_3$, are among the parameters defining the loop closure equations below. An additional requirement for the proper assembly of the loop is that the pivot bond angles $\theta_i, i = 1,2,3$ must assume their prescribed values. That may be possible to accomplish by rotating segment F_i about the virtual bond joining pivots (p_i, p_{i+1}) by angle τ_i (Figure 4.4g). The additional atoms, $p_i + 2, p_{i+1} - 2$ that are included in each virtual segment allow the calculation of the six pivot torsions, once the virtual segments have been rotated to their correct positions, so that the angle $(p_i - 1, p_i, p_i + 1) = \theta_i$. Note that the loop closure equations are formulated in the body frame of the three pivot atoms. To convert to the space frame of the rest of the molecule, the fragment F_3 is assumed fixed, and the rest of the loop (fragments F_1, F_2) is rotated about (p_3, p_1) by the angle $-\tau_3$. Determining the pivot torsions completes the specification of all internal degrees of freedom for the missing loop, which can now be constructed, closing the gap (Figure 4.4h).

The bond angle constraints lead to the loop closure equations[127]. These are a system of three polynomials that are quadratic in each of the variables:

$$L_2(u_3)u_1^2 + L_1(u_3)u_1 + L_0(u_3) = 0,$$
$$\left(M_{22}u_2^2 + M_{21}u_2 + M_{20}\right)u_1^2 + \left(M_{12}u_2^2 + M_{11}u_2 + M_{10}\right)u_1 + \left(M_{02}u_2 + M_{01}u_2 + M_{00}\right) = 0,$$
$$N_2(u_3)u_2^2 + N_1(u_3)u_2 + N_0(u_3) = 0.$$

The variables are $u_i = \tan\left(\tau_i/2\right), i = 1,2,3$. The $L_i, N_i, i = 0,1,2$ are quadratic polynomials in u_3, while the $M_{ij}, i, j = 0,1,2$ are constants. Each polynomial depends on only two of the u_i. Throughout, the notation of Coutsias *et al.*[127] is followed, and the reader is referred to that reference for the values of the polynomial coefficients. The code encodes the atomic coordinates of each virtual segment as a set of triaxial parameters as in Coutsias *et al.*[127] These parameters are used to populate a matrix $R(u_3)$ called the Dixon Resultant (DR) that results from eliminating the variables u_1, u_2 (any two of the variables could have been eliminated in favor of the remaining one). The necessary and sufficient condition that the above system of three polynomials in the three variables u_1, u_2, u_3 has a common solution is expressed by the equation[132]

$$R(u_3)V(u_1,u_2) := \begin{bmatrix} 0 & A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\ A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 & 0 \\ 0 & B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\ B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \\ 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 & 0 \\ 0 & D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\ D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ u_1 \\ u_1^2 \\ u_1^3 \\ u_2 \\ u_1u_2 \\ u_1^2u_2 \\ u_1^3u_2 \end{bmatrix} = 0$$

where

$$A_i := M_{i1}N_0 - M_{i0}N_1,$$
$$B_i := M_{i2}N_0 - M_{i0}N_2,$$
$$C_i := M_{i2}N_1 - M_{i1}N_2,$$
$$D_i := L_i.$$

Since its coefficients are quadratic polynomials in u_3 the DR can be written as a matrix polynomial

$$R(u_3) = R_2u_3^2 + R_1u_3 + R_0.$$

The above matrix equation can be recast as a generalized eigenvalue problem

$$\left(\begin{bmatrix} I & 0 \\ 0 & R_2 \end{bmatrix} u_3 - \begin{bmatrix} 0 & I \\ -R_0 & -R_1 \end{bmatrix}\right)\begin{bmatrix} V \\ u_3V \end{bmatrix} = 0.$$

This eigenproblem can be solved directly using the QZ factorization algorithm. An attractive feature of this approach is that the remaining variables u_1, u_2 are also found directly from the solution of this generalized eigenproblem, since they appear explicitly as particular components (resp. V_2, V_5) of the corresponding generalized eigenvector while u_3 is the generalized eigenvalue[132]. Sixteen solutions are always found, but some or all may be complex. To have geometrical meaning the solutions must be real, so complex solutions are discarded. The eigenproblem has the advantage of robustness and conceptual simplicity, but it can be computationally expensive, as each step of the iterative QZ algorithm scales with the cube of matrix size. As an alternative, we can get u_3 from the condition that the determinant of the DR must vanish. Having found values for u_3 for which $R(u_3)$ becomes singular, we can determine the desired components of its null-vector V by Cramer's rule. Since the coefficients of R are quadratic polynomials in u_3, its determinant is a polynomial

of degree 16 in u_3, and by examining the existence of real solutions only, a substantial speedup can be accomplished. The polynomial conversion has been carried out optimally by careful regrouping of the terms and employing Lagrange expansions in complementary minors. By a rearrangement of rows, we have the equivalent form

$$\det(R)=\begin{vmatrix} D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 & 0 \\ A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 & 0 \\ B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 & 0 \\ 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 & 0 \\ 0 & D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\ 0 & A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\ 0 & B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\ 0 & 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \end{vmatrix}.$$

In this form we get a compact expansion in terms of 4×4 minors

$$\det(R)=-P_{1235}P_{3567}+P_{1256}P_{2367}-P_{1257}P_{2357}-P_{1356}P_{1367}+P_{1357}^2-P_{1567}P_{1237}$$

where P_{ijkl} is the determinant of the minor formed by rows 1,...,4 and columns *i*, *j*, *k* and *l*. We have

$$P_{1235}P_{3567}:=\begin{vmatrix} D_0 & D_1 & D_2 & 0 \\ A_0 & A_1 & A_2 & B_0 \\ B_0 & B_1 & B_2 & C_0 \\ 0 & 0 & 0 & D_0 \end{vmatrix}\begin{vmatrix} D_2 & 0 & 0 & 0 \\ A_2 & B_0 & B_1 & B_2 \\ B_2 & C_0 & C_1 & C_2 \\ 0 & D_0 & D_1 & D_2 \end{vmatrix}=D_0D_2\begin{vmatrix} D_0 & D_1 & D_2 \\ A_0 & A_1 & A_2 \\ B_0 & B_1 & B_2 \end{vmatrix}\begin{vmatrix} B_0 & B_1 & B_2 \\ C_0 & C_1 & C_2 \\ D_0 & D_1 & D_2 \end{vmatrix}=$$

$$D_0D_2\left(C_0\begin{vmatrix} B_1 & B_2 \\ D_1 & D_2 \end{vmatrix}-C_1\begin{vmatrix} B_0 & B_2 \\ D_0 & D_2 \end{vmatrix}+C_2\begin{vmatrix} B_0 & B_1 \\ D_0 & D_1 \end{vmatrix}\right)\left(A_0\begin{vmatrix} D_1 & D_2 \\ B_1 & B_2 \end{vmatrix}-A_1\begin{vmatrix} D_0 & D_2 \\ B_0 & B_2 \end{vmatrix}+A_2\begin{vmatrix} D_0 & D_1 \\ B_0 & B_1 \end{vmatrix}\right)$$

$$P_{1256}P_{2367} := \begin{vmatrix} D_0 & D_1 & 0 & 0 \\ A_0 & A_1 & B_0 & B_1 \\ B_0 & B_1 & C_0 & C_1 \\ 0 & 0 & D_0 & D_1 \end{vmatrix} \begin{vmatrix} D_1 & D_2 & 0 & 0 \\ A_1 & A_2 & B_1 & B_2 \\ B_1 & B_2 & C_1 & C_2 \\ 0 & 0 & D_1 & D_2 \end{vmatrix} =$$

$$\left(\begin{vmatrix} D_0 & D_1 \\ A_0 & A_1 \end{vmatrix} \begin{vmatrix} C_0 & C_1 \\ D_0 & D_1 \end{vmatrix} - \begin{vmatrix} D_0 & D_1 \\ B_0 & B_1 \end{vmatrix} \begin{vmatrix} B_0 & B_1 \\ D_0 & D_1 \end{vmatrix} \right) \left(\begin{vmatrix} D_1 & D_2 \\ A_1 & A_2 \end{vmatrix} \begin{vmatrix} C_1 & C_2 \\ D_1 & D_2 \end{vmatrix} - \begin{vmatrix} D_1 & D_2 \\ B_1 & B_2 \end{vmatrix} \begin{vmatrix} B_1 & B_2 \\ D_1 & D_2 \end{vmatrix} \right)$$

$$P_{1257}P_{2357} := \begin{vmatrix} D_0 & D_1 & 0 & 0 \\ A_0 & A_1 & B_0 & B_2 \\ B_0 & B_1 & C_0 & C_2 \\ 0 & 0 & D_0 & D_2 \end{vmatrix} \begin{vmatrix} D_1 & D_2 & 0 & 0 \\ A_1 & A_2 & B_0 & B_2 \\ B_1 & B_2 & C_0 & C_2 \\ 0 & 0 & D_0 & D_2 \end{vmatrix} =$$

$$\left(\begin{vmatrix} D_0 & D_1 \\ A_0 & A_1 \end{vmatrix} \begin{vmatrix} C_0 & C_2 \\ D_0 & D_2 \end{vmatrix} - \begin{vmatrix} D_0 & D_1 \\ B_0 & B_1 \end{vmatrix} \begin{vmatrix} B_0 & B_2 \\ D_0 & D_2 \end{vmatrix} \right) \left(\begin{vmatrix} D_1 & D_2 \\ A_1 & A_2 \end{vmatrix} \begin{vmatrix} C_0 & C_2 \\ D_0 & D_2 \end{vmatrix} - \begin{vmatrix} D_1 & D_2 \\ B_1 & B_2 \end{vmatrix} \begin{vmatrix} B_0 & B_2 \\ D_0 & D_2 \end{vmatrix} \right)$$

$$P_{1356}P_{1367} := \begin{vmatrix} D_0 & D_2 & 0 & 0 \\ A_0 & A_2 & B_0 & B_1 \\ B_0 & B_2 & C_0 & C_1 \\ 0 & 0 & D_0 & D_1 \end{vmatrix} \begin{vmatrix} D_0 & D_2 & 0 & 0 \\ A_0 & A_2 & B_1 & B_2 \\ B_0 & B_2 & C_1 & C_2 \\ 0 & 0 & D_1 & D_2 \end{vmatrix} =$$

$$\left(\begin{vmatrix} D_0 & D_2 \\ A_0 & A_2 \end{vmatrix} \begin{vmatrix} C_0 & C_1 \\ D_0 & D_1 \end{vmatrix} - \begin{vmatrix} D_0 & D_2 \\ B_0 & B_2 \end{vmatrix} \begin{vmatrix} B_0 & B_1 \\ D_0 & D_1 \end{vmatrix} \right) \left(\begin{vmatrix} D_0 & D_2 \\ A_0 & A_2 \end{vmatrix} \begin{vmatrix} C_1 & C_2 \\ D_1 & D_2 \end{vmatrix} - \begin{vmatrix} D_0 & D_2 \\ B_0 & B_2 \end{vmatrix} \begin{vmatrix} B_1 & B_2 \\ D_1 & D_2 \end{vmatrix} \right)$$

$$P_{1357}^2 := \begin{vmatrix} D_0 & D_2 & 0 & 0 \\ A_0 & A_2 & B_0 & B_2 \\ B_0 & B_2 & C_0 & C_2 \\ 0 & 0 & D_0 & D_2 \end{vmatrix}^2 = \left(\begin{vmatrix} D_0 & D_2 \\ A_0 & A_2 \end{vmatrix} \begin{vmatrix} C_0 & C_2 \\ D_0 & D_2 \end{vmatrix} - \begin{vmatrix} D_0 & D_2 \\ B_0 & B_2 \end{vmatrix} \begin{vmatrix} B_0 & B_2 \\ D_0 & D_2 \end{vmatrix} \right)^2$$

$$P_{1567}P_{1237} := \begin{vmatrix} D_2 & 0 & 0 & 0 \\ A_2 & B_0 & B_1 & B_2 \\ B_2 & C_0 & C_1 & C_2 \\ 0 & D_0 & D_1 & D_2 \end{vmatrix} \begin{vmatrix} D_0 & D_1 & D_2 & 0 \\ A_0 & A_1 & A_2 & B_0 \\ B_0 & B_1 & B_2 & C_0 \\ 0 & 0 & 0 & D_0 \end{vmatrix} = D_2 D_0 \begin{vmatrix} B_0 & B_1 & B_2 \\ C_0 & C_1 & C_2 \\ D_0 & D_1 & D_2 \end{vmatrix} \begin{vmatrix} D_0 & D_1 & D_2 \\ A_0 & A_1 & A_2 \\ B_0 & B_1 & B_2 \end{vmatrix}$$

$$= P_{3567}P_{1235}.$$

There are only 9 different 2×2 determinants involved in these calculations, each resulting in a quartic polynomial in u_3. The computation of the coefficients of the characteristic polynomial can be carried out in under 1800 flops. The

polynomial is solved efficiently by the method of Sturm chains[148] and each solution results in a set of torsions for the pivot residues that close the loop. The determination of the variables u_1, u_2 now requires additional calculation. Briefly, the generalized eigenvectors are null vectors of the DR matrix R. Once u_3 has been found for which $\det(R)$ vanishes, we may determine specific components of the null vectors of $R(u_3)$ by using Cramer's rule. Most of the determinant minors involved in this computation are already known from the calculation of the characteristic polynomial.

$$\begin{bmatrix} D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 & 0 \\ A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 & 0 \\ B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 & 0 \\ 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 & 0 \\ 0 & D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\ 0 & A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\ 0 & B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\ 0 & 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \end{bmatrix} \begin{bmatrix} 1 \\ u_1 \\ u_1^2 \\ u_1^3 \\ u_2 \\ u_1 u_2 \\ u_1^2 u_2 \\ u_1^3 u_2 \end{bmatrix} = 0 \Rightarrow$$

$$\begin{bmatrix} D_1 & D_2 & 0 & 0 & 0 & 0 & 0 \\ B_1 & B_2 & 0 & C_0 & C_1 & C_2 & 0 \\ 0 & 0 & 0 & D_0 & D_1 & D_2 & 0 \\ D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\ A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\ B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\ 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_1^2 \\ u_1^3 \\ u_2 \\ u_1 u_2 \\ u_1^2 u_2 \\ u_1^3 u_2 \end{bmatrix} = \begin{bmatrix} D_0 \\ B_0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

where we have omitted one of the equations since it is dependent on the others and moved the first column to the RHS of the resulting system of 7 equations in the 7 unknowns. We only need to solve for the first and fourth components; in a generic situation one would apply LU factorization. However here Cramer's rule can be used

effectively, since most of the necessary determinantal computations have already been done for finding the characteristic polynomial. We have

$$d_0 := \begin{vmatrix} D_1 & D_2 & 0 & 0 & 0 & 0 & 0 \\ B_1 & B_2 & 0 & C_0 & C_1 & C_2 & 0 \\ 0 & 0 & 0 & D_0 & D_1 & D_2 & 0 \\ D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\ A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\ B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\ 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \end{vmatrix}$$

$$u_1 = \frac{d_1}{d_0} := \frac{1}{d_0} \begin{vmatrix} D_0 & D_2 & 0 & 0 & 0 & 0 & 0 \\ B_0 & B_2 & 0 & C_0 & C_1 & C_2 & 0 \\ 0 & 0 & 0 & D_0 & D_1 & D_2 & 0 \\ 0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\ 0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\ 0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\ 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \end{vmatrix}$$

$$u_2 = \frac{d_2}{d_0} := \frac{1}{d_0} \begin{vmatrix} D_1 & D_2 & 0 & D_0 & 0 & 0 & 0 \\ B_1 & B_2 & 0 & B_0 & C_1 & C_2 & 0 \\ 0 & 0 & 0 & 0 & D_1 & D_2 & 0 \\ D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\ A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\ B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\ 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \end{vmatrix}.$$

These are expanded as follows

$$d_0 = D_0 \begin{vmatrix} D_1 & D_2 \\ B_1 & B_2 \end{vmatrix} P_{3567} - D_1 \left(\begin{vmatrix} C_0 & C_1 \\ D_0 & D_1 \end{vmatrix} P_{2367} - \begin{vmatrix} C_0 & C_2 \\ D_0 & D_2 \end{vmatrix} P_{2357} \right) + D_2 \left(\begin{vmatrix} C_0 & C_1 \\ D_0 & D_1 \end{vmatrix} P_{1367} - \begin{vmatrix} C_0 & C_2 \\ D_0 & D_2 \end{vmatrix} P_{1357} \right)$$

$$d_1 = -D_0 \left(\begin{vmatrix} D_0 & D_2 \\ B_0 & B_2 \end{vmatrix} P_{3567} + \begin{vmatrix} C_0 & C_1 \\ D_0 & D_1 \end{vmatrix} P_{2367} + \begin{vmatrix} C_0 & C_2 \\ D_0 & D_2 \end{vmatrix} P_{2357} \right)$$

$$d_2 = \left(D_1 P_{2367} - D_2 P_{2357} \right) \begin{vmatrix} D_0 & D_1 \\ B_0 & B_1 \end{vmatrix} - \left(D_1 P_{1367} - D_2 P_{1357} \right) \begin{vmatrix} D_0 & D_2 \\ B_0 & B_2 \end{vmatrix} - D_0 \begin{vmatrix} C_1 & C_2 \\ D_1 & D_2 \end{vmatrix} P_{1237}.$$

As the P_{ijkl} and the various 2×2 determinants in these expressions have already been computed in the calculation of the characteristic polynomial, the computation of the $d_i, i = 1,2,3$ can be accomplished with an additional cost of under 400 flops.

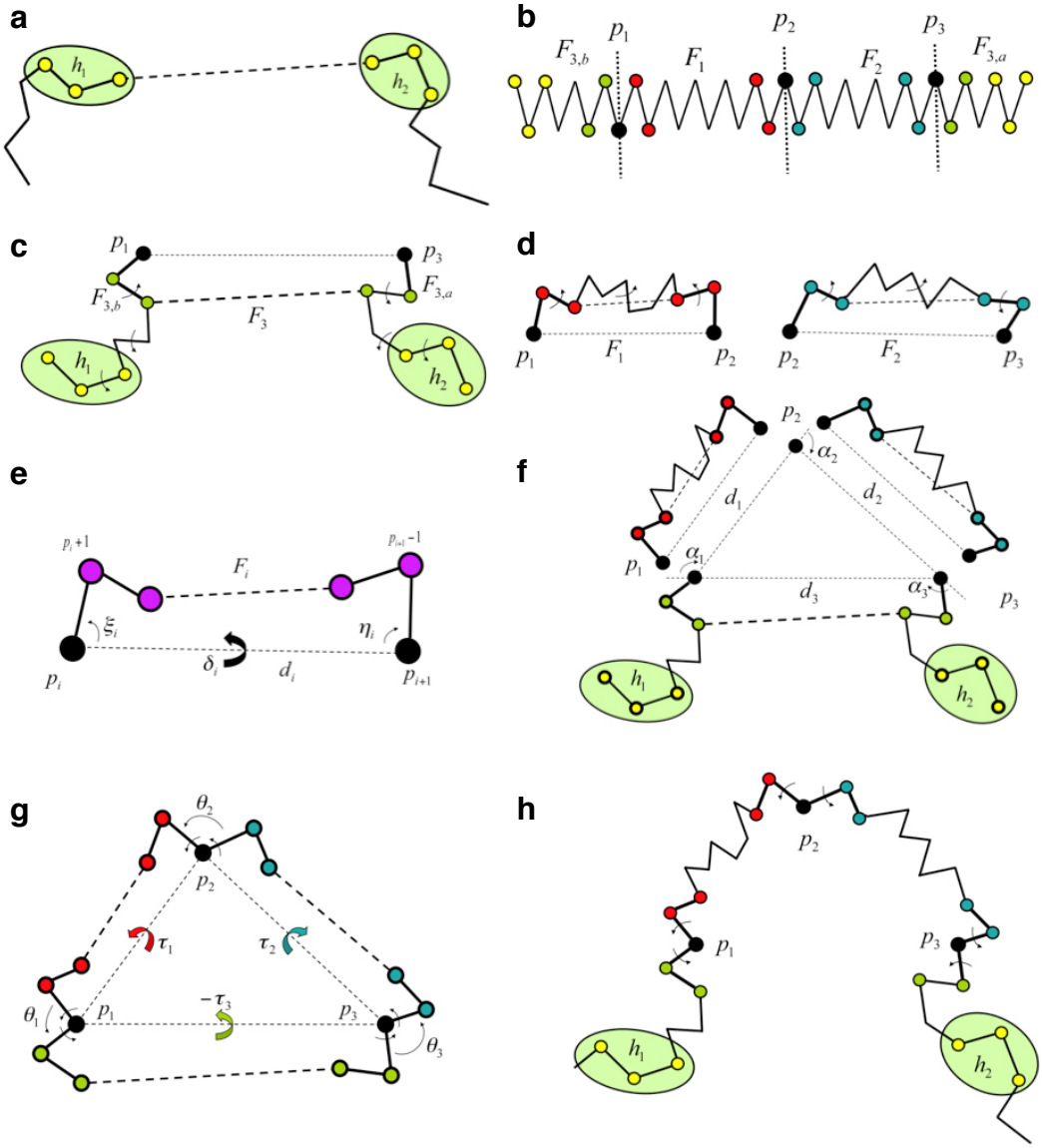

Figure 4.4: Geometric steps taken by the kinematic closure solver
(**a**) Hinge N-Cα-C triads h_1, h_2 are defined flanking an arbitrary peptide chain. (**b**) The chain is partitioned into four fragments F_1, F_2, $F_{3,a}$, and $F_{3,b}$, defined by the three pivot Cα atoms p_1, p_2 and p_3. (**c**) The hinges are fixed in space, and the fragments $F_{3,a}$ and $F_{3,b}$ are constructed from the hinges using prescribed geometry. (**d**) The other two fragments, F_1 and F_2, are determined in their body frame with prescribed internal bond lengths, bond angles, and torsions, but are yet to be positioned with respect to $F_{3,a}$ and $F_{3,b}$. (**e**) Geometrical parameters for the kinematic closure equations are defined for the 4 fragments. (**f**) The fragments are assembled into a triangle such that three lengths d_1, d_2, and d_3 satisfy the triangle inequality. The resulting exterior angles α_1, α_2, and α_3 form additional parameters for the loop closure equations. (**g**) The atoms of the 3 segments connecting two adjacent pivot atoms are rotated about the axis between the two pivots by an angle τ_i so that the prescribed pivot bond angles θ_i are satisfied. (**h**) The chain is converted from the body frame of the pivots to the space frame of the hinges by assuming the fragment F_3 is fixed and rotating the remaining fragments by angle $-\tau_3$.

Elimination of native bias

For loop reconstruction to have broad applicability it is important to carry out predictions with minimal knowledge of the native side-chain environment. The Rosetta KIC protocol first discards all native side-chain chi angles, bond angles, and bond lengths and repacks the side-chains using conformations from a rotamer library[145]. This initial repacking (without the presence of any native side-chains at any position) is carried out against the native backbone (dataset 1) or on the perturbed backbone in dataset 2 obtained from Sellers *et al*[133]. Subsequently, an initial kinematic closure discards the native loop backbone torsions, bond angles, and bond lengths, and places the loop into a non-native starting conformation with idealized bond lengths and bond angles (except for N-Cα-C bond angles, which have been sampled without knowledge of the native values). After the protocol completes the centroid stage, all side-chains within 10 Å of the predicted loop conformation are discarded and repacked. This step entails that at the beginning of the full-atom stage, side-chains within 10 Å of the loop have been optimized against a predicted non-native backbone in datasets 1 and 3, and all side-chains have been optimized against an initially perturbed backbone in dataset 2.

Discussion

Both conformational sampling and accurate scoring are significant challenges for high-resolution protein modeling. In the following sections I discuss examples of successes and failures of loop reconstruction arising in both areas, and evaluate the

sensitivity of the KIC method to modified sampling parameters, together with the required computational cost.

Conformational sampling

Accurate loop reconstruction requires substantial conformational sampling, owing to the extensive conformational space accessible to protein loops. If conformations near the crystallographic loop are not sampled, reconstruction accuracy will be poor. Even if near-native conformations are sampled, the scoring function must discriminate them from the ensemble of conformations explored during the course of the simulation (insofar as the crystallographic structure represents the lowest free energy conformation of the protein). I sought to determine which failure cases (reconstruction accuracy ≥1.0 Å) were attributable to insufficient sampling, and which suffered from incorrect scoring for both Rosetta methods on datasets 1^{96} and 2^{133} (accuracy is measured as global loop rmsd to the native backbone N, Cα, C, O atoms throughout this chapter). To do so, I compared the scores of the lowest-scoring reconstructions to the scores of the crystallographic loops. If the crystallographic loop scored lower (better) than the lowest-scoring model, the failure resulted at least in part from insufficient conformational sampling, because the scoring function would have discriminated very-near crystallographic conformations had they been sampled. Conversely, if the lowest-scoring reconstruction was lower in score than the crystallographic loop, the failure was attributable to the scoring function, since near-crystallographic conformations scored worse than conformations ≥1.0 Å away. The scores of the crystallographic

loops were obtained by relaxing the repacked, minimized input structures through 100 independent trajectories of the full-atom stage of the KIC protocol fixed at a temperature of 0.5 *k*T and recording the lowest scoring conformations within 0.5 Å of each crystallographic loop. On the 25 loops in the filtered dataset 1, I found that 16 out of 18 failure cases were due to poor conformational sampling using the standard protocol, compared to 5 out of 10 such cases using KIC (Table 4.4). On dataset 2, all 15 failures were attributable to insufficient sampling using the standard protocol, compared to 6 out of 10 using the KIC protocol (Table 4.5). Contributions from scoring and sampling cannot be completely decoupled since Metropolis Monte Carlo simulations accept or reject conformations with a probability dependent on the score. Since both protocols use the same number of steps over identical simulated annealing schedules with the same scoring function, however, these results suggest that the KIC protocol, while imperfect, substantially improves conformational sampling compared to the standard protocol. Additionally, the results show that the enhanced torsion sampling enabled by KIC can reveal scoring errors by finding low scoring structures distant from the crystallographic loop. Cases 4i1b and 1tgh in Table 4.4 and 1my7, 2pia, 1m3s and 1oyc in Table 4.5 are examples where scoring errors become apparent when sampling is enhanced with the KIC protocol.

Dataset 3 provides an additional perspective on conformational sampling with KIC because all the loops were crystallized in multiple conformations bound to different protein partners. Since protein modeling and design methods frequently transplant existing structures into new contexts as templates, it is useful to know

how often the predicted loop more closely resembles the crystallographic loop than the same loop crystallized with different partner proteins. I pairwise-superimposed the cores of all loop proteins in dataset 3 and computed the global backbone N, Cα, C, O rmsds between the conformations of the loops bound to different partners. These rmsds, which show that even shorter 7-residue loops are capable of assuming significantly different conformations across complexes, are reported in Table 4.3. In 57 of 68 cases, the predicted loop was closer to the crystallographic loop than the same loop crystallized with another partner (shown in bold). This result highlights the potential of the method in refinement applications (predicting a conformation closer than the template structure) and also for modeling loop changes in important conformational switch proteins.

Factors not modeled

Errors in loop reconstruction can result from structural features that are not explicitly considered by the modeling method. For the Rosetta KIC protocol, such factors include crystallization conditions at pH values outside the neutral range, amino acid residues with shifted ionization constants, and residues with *cis* peptide bonds (since the protocol currently does not sample *cis* peptide bonds). Other errors could result from interactions between loop residues and neighboring protein copies in the crystal lattice, since the simulations are not performed within the crystallographic unit cell. To check for possible crystal packing effects, I reconstructed the crystal lattice using Pymol[149] and computed the changes in solvent accessible surface area (SASA) with and without the crystal context using

Surface Racer[150] (1.2 Å probe radius, using Richards 1977[151] van der Waals radii) for all loop residues in datasets 1 and 2. Cases where the delta SASA with and without the crystal context was >200 $Å^2$ were considered to have significant crystal packing. Table 4.6 and Table 4.7 show which failure cases had significant crystal packing by this measure, *cis* peptide bonds, or pH values well outside the neutral range.

Energy function simplifications and errors

The most significant scoring function failure in Table 4.5 involves a protein loop with specific interactions with a buried water molecule (Old Yellow Enzyme, PDB 1oyc, 2.0 Å resolution). The crystal structure suggests that this water molecule (H_2O 609), which has a B-factor (~24) that is lower than the average B-factor for waters in this structure (~29), forms a hydrogen bonding network with the backbone carbonyl of loop residue Ser 206, the side-chain hydroxyl group of Ser 136, and the backbone amide of Ser 138 (Figure 4.5). The loop reconstruction deviates substantially from the crystallographic loop in the region where the buried water molecule interacts with the loop backbone. Interactions with water molecules are a common source of error associated with the use of an implicit solvent model such as the one implemented in Rosetta (see next paragraph) that ignores effects resultant from the discrete size and asymmetry of water molecules and the geometric constraints of water-mediated hydrogen bonding interactions.

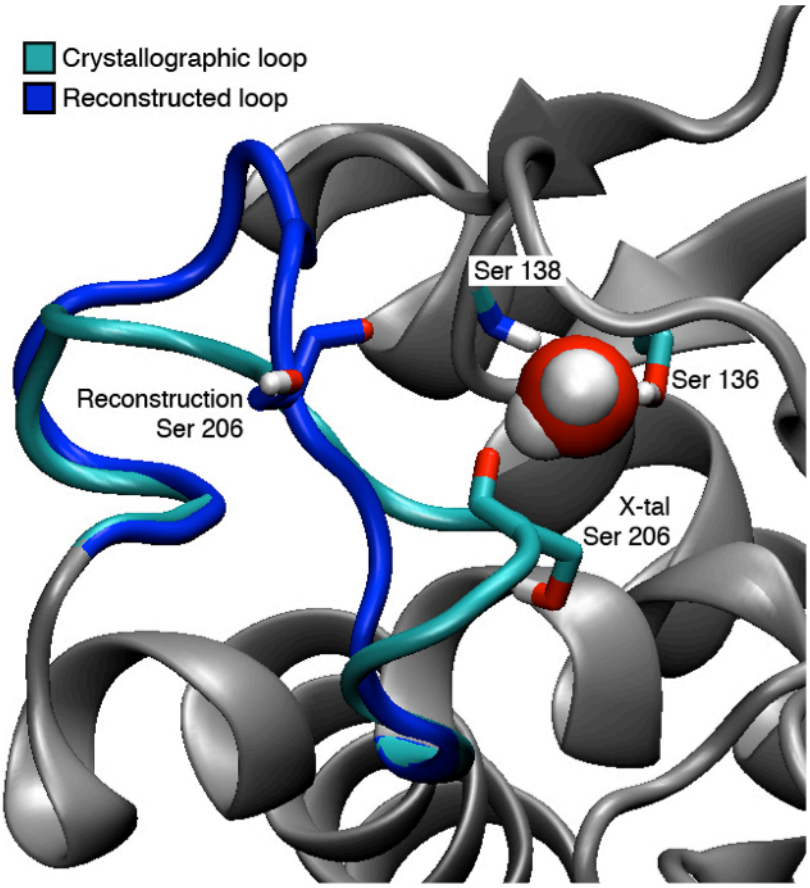

Figure 4.5: Effect of specific interactions of loop atoms with a buried water molecule
The loop residues subject to reconstruction from Old Yellow Enzyme (PDB 1oyc) are colored cyan in the crystal structure conformation, and the reconstruction is shown in blue. The backbone carbonyl of loop residue Ser 206 is shown in sticks, along with the side-chains of Ser 206, Ser 136, and the backbone amide of Ser 138 in the crystal structure. The backbone carbonyl and side-chain of Ser 206 on the reconstructed loop are also shown in sticks. Hydrogen atoms are included in the crystal structure. Explicit water molecules are not included in the loop reconstruction simulations, and this protein produces the most significant scoring error with the KIC protocol on dataset 2.

Even when all atoms are explicitly represented, evaluating the energetic contribution of charged and polar interactions is a significant challenge for any scoring function. The Rosetta all-atom scoring function uses a combination of an orientation-dependent hydrogen bond term[17] with an implicit solvation model[152] to assess hydrogen bonding in protein structures. As Chapter 2 describes in detail, it can be difficult to reconstruct the complex hydrogen bonding networks observed in

some protein structures, due to the delicate energetic balance between forming inter-residue hydrogen bonds and losing hydrogen bonds to solvent (in addition to the absence of polarization effects that are ignored by most methods). Figure 4.6a shows an example of two loop residues that participate in a hydrogen bonding and polar interaction network with two other residues in human 5'-deoxy-5'-methylthioadenosine phosphorylase (PDB 1cb0). A loop side-chain (Asp 43) accepts hydrogen bonds from the backbone and side-chain of Arg 63, which in turn interacts with Glu 31 and another loop side-chain, Tyr 33. In the KIC reconstruction of this loop and the surrounding side-chain environment (Figure 4.6b, 0.6 Å rmsd to the crystallographic loop), the hydrogen bonds and polar interactions between loop residue Asp 43, and neighbors Arg 63 and Glu 31 are recovered, suggesting that Rosetta sufficiently samples these side-chain conformations and that the hydrogen bonding terms can successfully evaluate the hydrogen bonding interactions in the presence of a perturbed backbone. Nevertheless, the reconstruction orients the side-chain of Tyr 33 out into bulk solvent, demonstrating that some electrostatic effects are too subtle for the Rosetta hydrogen bonding and solvation terms to model accurately.

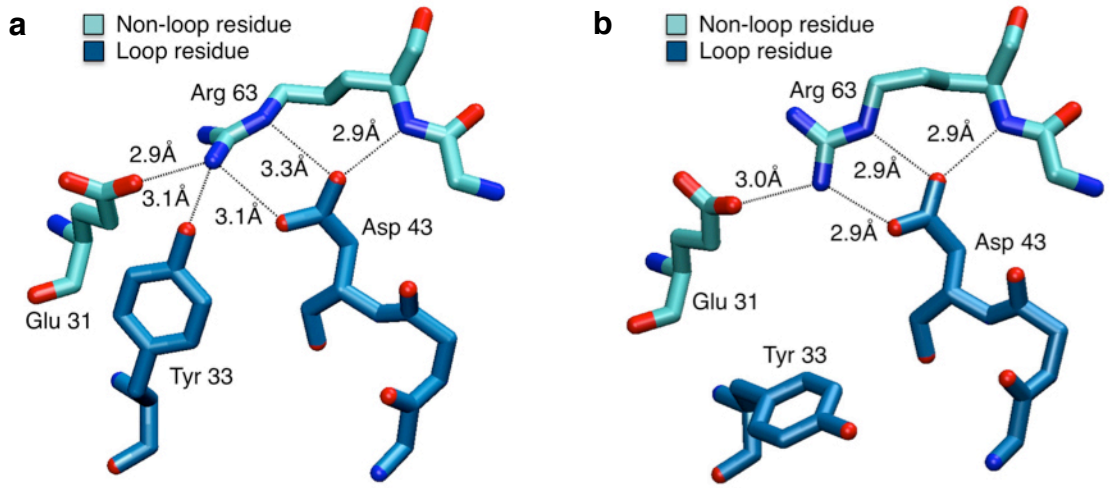

Figure 4.6: Loop reconstruction with a complex hydrogen bonding and polar network
(**a**) Loop residues Asp 43 and Tyr 33 form a hydrogen bonding and polar network with the backbone amide of Arg 63 and the side-chains of Arg 63 and Glu 31 in the crystal structure of human 5'-deoxy-5'-methylthioadenosine phosphorylase (PDB 1cb0). (**b**) The loop reconstruction of 1cb0 recovers the hydrogen bonds and polar interactions between loop residue Asp 43 with the amide backbone of Arg 63 and the side-chains of Arg 63 and Glu 31, but orients the side-chain of Tyr 33 toward bulk solvent. The loop was reconstructed to 0.6 Å accuracy.

Sensitivity to simulation parameters

As described above in 'Methods', the Rosetta KIC protocol samples N-Cα-C bond angles. To assess the importance of bond angle sampling to reconstruction accuracy, I re-ran the simulations on the combined 45 loops from datasets 1 and 2 using the same KIC protocol except I fixed the loop N-Cα-C bond angles at canonical values (110.86˚). The fixed bond angle protocol achieved similar performance (1.0 Å median rmsd) as the protocol that sampled bond angles (0.9 Å rmsd), suggesting that the contribution of bond angle sampling is small (Table 4.8). This result of the relatively small effect of bond angle sampling is consistent with the overwhelmingly greater variability of backbone torsions compared to bond angles, and Laskowski *et al.*[153] have shown that the observed variability of bond angles decreases at very high

crystallographic resolution. Additionally, Coutsias *et al.*[127] showed in an earlier analysis of their KIC method that while N-Cα-C bond angle sampling increases the number of closable loops, it does not produce more native-like conformations. These results also suggest that bond angle sampling is not a significant bottleneck to the performance of the standard Rosetta method. In general, high-resolution structure prediction may not require sampling far from canonical bond angles, although it may be important in cases of experimentally observed bond angle strain, as in small cyclic peptides.

I also considered the effect of the simulated annealing schedule – originally performed on a fairly narrow range – on reconstruction performance. Rather than varying the temperature as described above in 'Loop modeling protocol', I fixed the temperature for both centroid and full-atom stages at 1.0 *k*T. Again, the modified protocol performed nearly as well as the original protocol with simulated annealing (Table 4.8), achieving a median accuracy of 1.0 Å with fixed temperature compared to 0.9 Å with annealed temperature. Taken together, these results show that the protocol is quite robust to changes in some simulation parameters, and suggest that the most important feature is the enhanced torsion sampling provided by KIC.

Computational cost

As noted above in 'Loop modeling protocol', the KIC protocol requires ~320 CPU-hours on a single 2.2 GHz Opteron processor to generate 1,000 models. The standard protocol requires ~280 CPU-hours to generate the same number of models on the same processor. To assess the performance of both Rosetta methods

as a function of CPU time, I performed shorter constant-time simulations on datasets 1 and 2. Each protocol was run for 120 CPU-hours on each protein using the same parameters as in the longer simulations. The rmsd of the best-scoring reconstruction to the crystallographic loop was computed in the same manner as the longer simulations. I found that using equal computational time, KIC improved the median reconstruction accuracy to 0.9 Å from 1.9 Å using the standard protocol on dataset 1 and improved median accuracy to 1.2 Å from 1.9 Å using the standard protocol on dataset 2. When both protocols were started from the perturbed loops from ref[133] on dataset 2, KIC improved median accuracy to 1.2 Å from the standard protocol value of 2.2 Å.

The molecular mechanics method required ~260 CPU-hours for each 12-residue loop simulation (B. Sellers, personal communication). As noted by Sellers *et al.*[133], the reported results employ side-chain optimization in a 7.5 Å shell around the reconstructed loops. The Rosetta KIC and standard protocols optimize side-chains within 10.0 Å of the loops. As additionally noted in Figure 3 in reference[133], the molecular mechanics method requires roughly twice the computational time to optimize side-chains within 10.0 Å of the loop compared to 7.5 Å on 8-residue loops. It can thus be expected that the molecular mechanics method will require at least as much computational time as the KIC protocol when optimizing side-chains within 10.0 Å of 12-residue loops.

Longer loops

The number of geometrically accessible loop conformations grows exponentially as loop length increases. Modeling loops much longer than 12-residues might require sufficient sampling resources as to render the problem currently intractable. To investigate this issue I performed *de novo* loop reconstructions on 10 18-residue RT loops from the SH3 protein domain family. The dataset, summarized in Table 4.9, contains 5 cases with RT loops bound to different peptides, and 5 unbound cases. The average pairwise sequence identity across the proteins is 34%, and the modeled loops are even more divergent, with an average pairwise sequence identity of 27%. The average pairwise backbone rmsd of the loops after superposition onto one of the cases (PDB 1abq) is 1.42 Å. For each case I generated ~4000 structures (instead of 1000 structures) with Rosetta revision 34279, and otherwise applied the same loop modeling protocol used for the protein complexes dataset. Thus, the peptide backbones were fixed, but their interfacial side-chain conformations were reconstructed together with the loop and neighboring side-chains.

KIC reconstructed the 18-residue loops to 0.9 Å median rmsd to the crystallographic loops across the set, similar to the performance observed on the other datasets. A representative set of reconstructions is shown in Figure 4.7. The only case modeled to >1.5 Å rmsd from native (PDB 1i1j, 3.5 Å rmsd accuracy) includes a disulfide bond that was not explicitly modeled. These results demonstrate that KIC can model at least one family of 18-residue loops to sub-angstrom accuracy using a tractable amount of sampling (roughly 4-fold as compared to 12-residue loops). Success on these cases may arise from energy landscapes that funnel Monte

Carlo trajectories toward near-native conformations. Other 18-residue cases, like some 12-residue cases, may have more rugged energy landscapes that require substantially more sampling to consistently find native-like conformations, or may have energetic features that are not accurately modeled by the Rosetta scoring function.

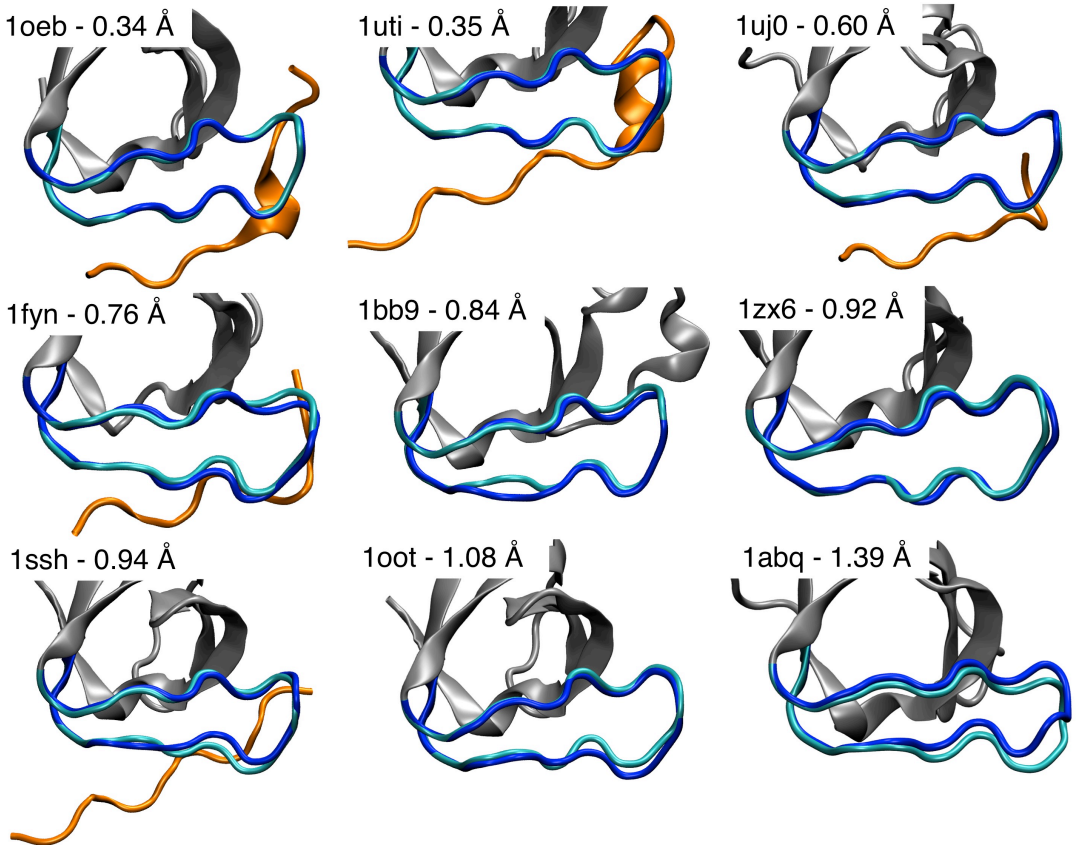

Figure 4.7: Representative set of 18-residue SH3 domain loop reconstructions
Backbone RMSD for each reconstruction (cyan) to the crystallographic loop (blue) is shown together with the PDB identifier. Peptide partners are shown in orange when present. One case not shown (PDB 1i1j; 3.5 Å accuracy) contains a disulfide bond from the crystallographic loop to the protein scaffold that was not explicitly modeled.

Table 4.1: KIC and standard protocol loop reconstruction accuracy on dataset 1

PDB	Loop residues	Standard protocol (Å rmsd from native)	KIC protocol (Å rmsd from native)	Pass ligand/ion filter?
1541	153-164	1.6	3.3	No
1arp	201-212	2.3	0.5	No
1ctm	9-20	5.4	2.9	No
1cyo	12-23	0.8	5.2	Yes
1dts	41-52	5.8	6.4	Yes
1eco	35-46	0.6	0.4	Yes
1ede	150-161	1.2	0.7	Yes
1ezm	122-133	2.4	2.7	Yes
1hfc	165-176	8.5	8.2	No
1ivd	365-376	7.4	2.1	No
1msc	9-20	3.7	3.2	Yes
1onc	23-34	3.8	0.5	Yes
1pbe	129-140	2.0	0.6	Yes
1pmy	77-88	2.6	2.6	No
1prn	15-26	7.0	6.6	No
1rcf	88-99	5.0	0.6	No
1rro	17-28	2.2	0.4	Yes
1scs	199-210	2.3	2.9	No
1srp	311-322	2.6	0.6	Yes
1tca	305-316	2.6	0.6	Yes
1thg	127-138	1.6	1.1	Yes
1thw	178-189	2.4	2.7	Yes
1tib	99-110	0.7	1.2	Yes
1tml	243-254	0.7	0.4	Yes
1xif	203-214	1.8	0.7	Yes
2cpl	145-156	0.4	0.2	Yes
2cyp	191-202	0.8	0.5	No
2ebn	136-147	3.9	2.1	Yes
2exo	293-304	1.2	0.8	Yes
2pgd	361-372	3.3	5.1	No
2rn2	90-101	1.1	0.8	Yes
2sil	255-266	2.0	1.0	Yes
2sns	111-122	3.3	3.6	No
2tgi	48-59	4.6	3.1	Yes
3cla	176-187	0.7	1.0	Yes
3cox	478-489	1.0	1.1	No
3hsc	72-93	0.5	0.5	Yes
451c	16-27	4.7	5.8	No
4enl	372-383	2.7	3.6	No
4ilb	46-57	5.1	3.8	Yes
	Mean	2.8	2.3	
	Median	2.4	1.2	
	Filtered mean[a]	2.2	1.6	
	Filtered median[a]	2.0	0.8	

[a] For cases that pass the ligand/ion filter (loop heavy atoms are ≥4.0Å from ligand heavy atoms and ≥6.5Å from charged ions).

Table 4.2: Performance of standard Rosetta, KIC Rosetta, and molecular mechanics protocols on dataset 2

PDB	Loop residues	Standard protocol *de novo* rmsd (Å)	KIC protocol *de novo* rmsd (Å)	Standard protocol perturbed[a] rmsd (Å)	KIC protocol perturbed rmsd (Å)	Molecular mechanics perturbed rmsd (Å)[b]
1a8d	155-166	5.4	6.9	5.3	0.6	2.8
1arb	182-193	1.6	1.0	5.1	1.4	2.6
1bhe	121-132	7.1	0.8	4.9	0.7	0.7
1bn8	298-309	2.5	0.7	1.7	0.6	2.6
1c5e	82-93	0.8	0.5	5.1	0.4	1.7
1cb0	33-44	1.0	0.6	1.1	0.7	0.3
1cnv	188-199	2.3	1.4	2.8	2.1	3.3
1cs6	145-156	2.5	3.0	4.0	3.0	3.5
1dqz	209-220	1.9	0.7	1.8	2.6	0.6
1exm	291-302	0.6	0.9	2.8	0.9	0.5
1f46	64-75	2.1	2.5	0.7	2.3	1.1
1i7p	63-74	0.7	2.7	0.8	0.4	0.3
1m3s	68-79	3.6	6.3	2.2	5.6	5.6
1ms9	529-540	2.5	0.4	2.8	1.0	2.5
1my7	254-265	2.0	2.3	0.6	2.3	0.9
1oth	69-80	0.6	0.6	1.9	0.6	0.7
1oyc	203-214	3.2	4.0	1.7	3.9	1.2
1qlw	31-42	3.3	1.0	5.0	0.9	1.4
1t1d	127-138	0.5	0.8	0.6	0.8	1.0
2pia	30-41	1.1	1.0	1.0	0.9	0.5
	Mean	2.3	1.9	2.6	1.6	1.7
	Median	2.1	1.0	2.0	0.9	1.2

[a] 'Perturbed' means simulations began with starting structures used in Sellers *et al.*, 2008.
[b] Values taken directly from Table S4 in Sellers *et al.*, 2008.

Table 4.3: Performance of the KIC protocol on dataset 3

PDB	Loop protein	Partner	Chains	Rmsd (Å)	Loop rmsds to the loop conformation in the other complex structures (in order listed)[a]	PDB loop	Length	Ligand	Cofactor
1doa	Cdc42	Rho GDI	A,B	2.2	**3.7**, **2.9**, 1.7, **4.1**	30-40	11	GDP	Mg2+
1grn	Cdc42	CDC42 GAP	A,B	0.9	**3.7, 5.0, 3.6, 1.1**	30-40	11	GDP/AF3	Mg2+
1gzs	Cdc42	SOP-E (Toxin)	A,B	2.1	**2.9, 5.0, 2.3, 6.2**	30-40	11	none	none
1ki1	Cdc42	Intersectin	A,B	4.3	1.7, 3.6, 2.3, **5.4**	30-40	11	none	none
1nf3	Cdc42	Par	A,C	1.5	**4.1**, 1.1, **6.2**, **5.4**	30-40	11	GNP/MG	Mg2+
1g4u	Rac	GAP SPTP	R,S	0.7	**0.8, 4.6**	30-39	10	GDP/AF3	Mg2+
1he1	Rac	Toxin	C,A	0.4	**0.8, 4.6**	30-39	10	GDP/AF3	Mg2+
1hh4	Rac	Rho GDI	A,D	0.8	**4.6, 4.6**	30-39	10	GDP	Mg2+
1bkd	Ras	Son of Sevenless-I	R,S	6.4	**9.9, 9.8, 9.6**	28-37	10	none	none
1he8	Ras	PI-3 Kinase	B,A	1.7	**9.9**, 0.5, 1.0	28-37	10	GNP	Mg2+
1k8r	Ras	BRY-2RBD	A,B	1.5	**9.8**, 0.5, 0.9	28-37	10	GNP	Mg2+
1wq1	Ras	Ras-GAP	R,G	0.6	**9.6, 1.0, 0.9**	28-37	10	GDP/AF3	Mg2+
1cmx	Ubiquitin	Modified Ubiquitin	B,A	0.3	**3.3, 2.3, 1.3, 1.0, 1.1**	306-312	7	none	none
1fxt	Ubiquitin	Conjugating Enzyme	B,A	0.7	**3.3, 2.3, 2.5, 3.1, 1.5**	6-12	7	none	none
1nbf	Ubiquitin	Deubiquitinating Enzyme	D,A	1.0	**2.3, 2.3, 2.4, 3.0, 2.7**	306-312	7	none	none
1wr6	Ubiquitin	GGA3-GAT	E,A	0.5	**1.3, 2.5, 2.4, 4.3, 0.9**	6-12	7	none	none
1wrd	Ubiquitin	TOM-GAT	B,A	0.6	**1.0, 3.1, 3.0, 4.3**, 0.6	6-12	7	none	none
2d3g	Ubiquitin	HRS-UIM	A,B+P	0.6	**1.1, 1.5, 2.7, 0.9**, 0.6	6-12	7	none	none
			Mean	1.5					
			Median	0.8					

[a] The core of the loop protein was pairwise-superimposed onto the structures of the loop protein bound to other partners. Global loop rmsds to the loop protein in the other structures are shown in the order listed in the table (descending from top). Cases where the predicted loop rmsd is less than the rmsd to the loop bound to another partner are shown in bold (57 / 68 cases).

Table 4.4: KIC and standard protocol sampling and scoring errors on dataset 1
Cases where reconstruction accuracy is ≥1.0 Å are shown. Gray boxes are primarily scoring errors, white boxes are primarily due to insufficient sampling.

KIC Protocol		Standard Protocol	
PDB	**best scoring model score – crystallographic loop score**	**PDB**	**best scoring model score – crystallographic loop score**
1ezm	9.47	1tca	21.74
2ebn	4.94	1ezm	15.75
2tgi	4.41	1srp	15.32
1thw	2.03	2exo	12.77
1tib	1.63	1pbe	10.73
4i1b	-0.76	2ebn	8.28
1thg	-1.17	2tgi	7.19
1cyo	-1.58	2rn2	7.18
1dts	-8.64	1thw	6.87
1msc	-39.53	1thg	6.64
		1ede	6.38
		1rro	3.09
		1xif	3.05
		2sil	2.55
		4i1b	2.52
		1onc	1.46
		1dts	-3.50
		1msc	-36.09

Table 4.5: KIC and standard protocol sampling and scoring errors on dataset 2
Cases where reconstruction accuracy is ≥1.0 Å are shown. Gray boxes are primarily scoring errors, white boxes are primarily due to insufficient sampling.

KIC Protocol		Standard Protocol	
PDB	**best scoring model score – crystallographic loop score**	**PDB**	**best scoring model score – crystallographic loop score**
1a8d	9.36	1a8d	21.98
1f46	7.91	1cnv	18.88
1cnv	5.25	1qlw	16.86
1i7p	4.69	1dqz	16.66
1qlw	4.18	1bhe	14.89
1cs6	2.41	1f46	9.63
1my7	-0.56	1arb	6.32
2pia	-1.63	1bn8	6.12
1m3s	-3.89	1cs6	5.36
1oyc	-5.58	1cb0	5.33
		1m3s	2.82
		1ms9	1.39
		1oyc	1.13
		1my7	0.61
		2pia	0.61

Table 4.6: Potential error sources from benchmark dataset 1
Cases where reconstruction accuracy is ≥1.0 Å using the KIC protocol are shown. Scoring errors are shaded gray as defined in Table 4.4.

PDB	Non-modeled factor(s)	Reconstruction rmsd (Å)
1dts	Crystal packing	6.4
1cyo	Crystal packing	5.2
4i1b		3.8
1msc	Crystal packing	3.2
2tgi	Crystal packing, low pH (4.2)	3.1
1ezm		2.7
1thw		2.7
2ebn	*Cis* proline	2.1
1tib	low pH (4.0)	1.2
1thg		1.1

Table 4.7: Potential sources of error from benchmark dataset 2
Cases where reconstruction accuracy is ≥1.0 Å using the KIC protocol are shown. Scoring errors are shaded gray as defined in Table 4.5.

PDB	Non-modeled factor(s)	Reconstruction rmsd (Å)
1a8d		6.9
1m3s	Crystal packing	6.3
1oyc		4.0
1cs6	*Cis* proline	3.0
1i7p		2.7
1f46	Crystal packing, *Cis* proline	2.5
1my7		2.3
1cnv	low pH (3.0-5.0)	1.4
1qlw		1.0
2pia	Crystal packing	1.0

Table 4.8: Sensitivity of reconstruction accuracy to simulation parameters
Mean and median rmsds are shown for the 3 protocols over all 45 loops from dataset 1 (filtered) and dataset 2. The input structures for dataset 2 are the perturbed starting structures used in ref[133].

	KIC Protocol	KIC Protocol with fixed N-Cα-C bond angles	KIC Protocol with temperature fixed at 1.0 *k*T
Mean (Å)	1.6	1.7	1.6
Median (Å)	0.9	1.0	1.0

Table 4.9: Performance of KIC protocol on 18-residue SH3 domain loops

PDB	Chains	PDB loop	Peptide ligand sequence	Accuracy (Å rmsd from native)
1abq	A	68-85	none	1.4
1bb9	A	28-45	none	0.8
1fyn	A,B	89-106	PPAYPPPPVP	0.8
1i1j	B	26-43	none	3.5
1oeb	A,D	6-23	PAPSIDRSTKPPL	0.3
1oot	A	6-23	none	1.1
1ssh	A,B	8-25	EGPPPAMPARPT	0.9
1uj0	A,B	209-226	TPMVNRENKPP	0.6
1uti	A,D	6-23	GQPPLVPPRKEKMRGK	0.4
1zx6	A	7-24	none	0.9
			Mean	1.1
			Median	0.9

Chapter 5

A method for flexible backbone design of protein-based small molecule biosensors

Concept and rationale

As a challenging test of engineering new functions by flexible backbone design, I have coupled KIC conformational sampling with protein sequence design to predict protein-based constructs to serve as small molecule biosensors. Here, the goal is to reshape the interfaces of existing heterodimeric protein complexes so that association of the protein partners becomes dependent on the presence of a small molecule target, similar to the rapamycin-induced association of FKBP12 and FRB[154]. This process is termed 'chemically induced dimerization', and the constituent proteins are called CIDs. Each CID partner is linked to one component of a split reporter (e.g., split-GFP, split-DHFR, yeast two-hybrid) so that small molecule-induced association of the partners is detectable by optical, enzymatic or transcriptional readouts (Figure 5.1). These constructs have two key features: First, the CID pairs can in principle be coupled to any split reporter in a modular fashion to drive a variety of outputs, and second, the sensors may be transcriptionally linked to pathways regulating cell survival or biosynthesis of the small molecule target.

These features are particularly advantageous for engineering cells to efficiently synthesize products such as biofuels and other value-added chemicals. Enzymatic or fluorescent outputs enable facile screening of many cell strains for high target yield. Further, detection of small molecule outputs of biosynthetic pathways *in vivo* can actuate downstream synthesis pathways, or regulate cell survival as a selection mechanism for directed evolution to optimize target production or conversion.

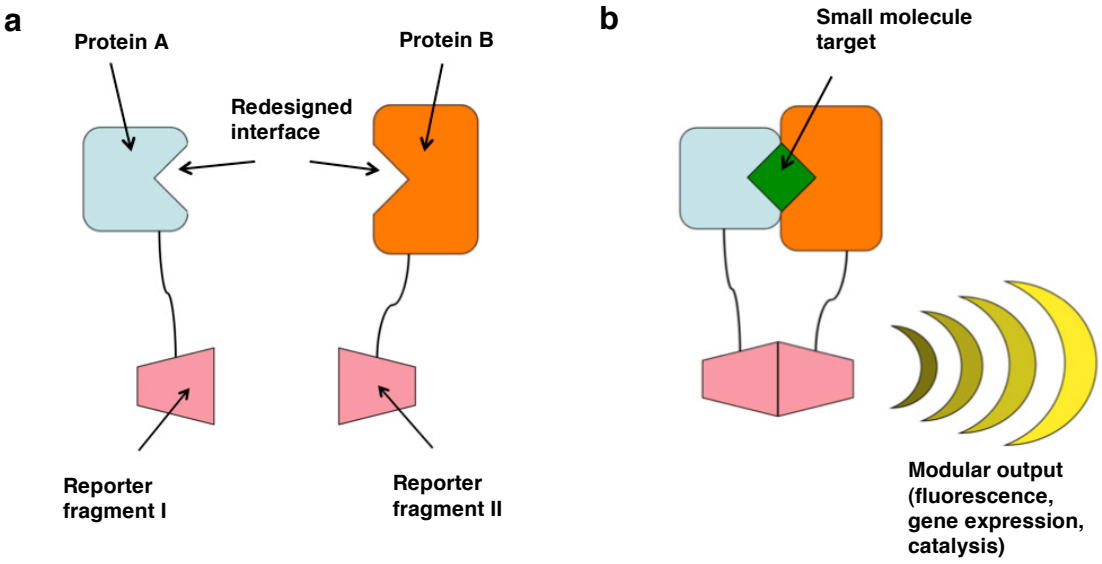

Figure 5.1: Schematic of a modular protein-based small molecule biosensor
(**a**) The interface of an existing heterodimeric protein complex is redesigned around a small molecule target. Each protein partner is linked to a fragment of a split reporter. In the absence of the target, the redesigned surfaces are unmatched, so the proteins do not associate. (**b**) Introduction of the small molecule target drives association of the redesigned protein partners, reconstituting the split reporter and generating an output signal. The split reporter component is modular and could produce fluorescent, enzymatic, or transcriptional readouts.

Several naturally occurring examples of CIDs exist. Besides rapamycin, Brefeldin A is a fungal toxin that stabilizes an inactive complex between ARF1 and Sec7 to repress vesicle transport[155,156]. Fusicoccin, another fungal toxin, increases the affinity of the activated C-terminus of the plasma membrane proton pump to a 14-3-3 protein by 90-fold[157], leading to permanent activation of the H^+ ATPase[158].

Aside from natural examples, some groups have produced synthetic small molecules to control CIDs based on natural products such as FK506[159], cyclosporin[160], and coumermycin[161]. The synthetic compounds typically serve as bivalent dimerizers of CIDs fused to proteins regulating processes ranging from apoptosis and gene transcription[162] to cell proliferation[163]. Reverse dimerization (chemically induced dissociation) has also been demonstrated as another potential regulatory tool[164]. Additionally, RNA aptamers have been selected that recognize the shared surface of a protein–small molecule complex[165].

Despite the existence of naturally and synthetically controlled CIDs, engineering chemical dependency into existing heterodimers remains challenging. Protein interfaces are notoriously resistant to small molecule interference due in large part to their typically broad, flat character[166]. Much can be learned, however, from the many proteins that have evolved to associate with small molecules. In order to benefit from the rich structural data on naturally occurring small molecule binding sites, I have pursued a 'motif directed' strategy toward the design of modular small molecule biosensors. In this approach the side-chain geometries of structurally characterized small molecule binding sites are 'transplanted' onto existing protein-protein interfaces, termed 'scaffolds'. Similar strategies have been used to introduce a copper binding site into *E. coli* thioredoxin[167], to graft the interleukin-4 binding epitope onto the coiled-coil domain of the yeast transcription factor GCN4[168], to transplant 'hot spot' residues from erythropoietin for its receptor onto a PH domain[169], and to replace residues in TEM1 β-lactamase that bind its inhibitor with a core module from another protein yielding wild-type affinity[170].

Notably, a motif-based approach similar to the methods described below was recently employed to create enzyme active sites that catalyze reactions not observed in nature[1,2]. The key distinctions of my approach are that the motif residues are transplanted across protein interfaces rather than into monomeric proteins, and that the interactions are optimized for binding rather than catalysis. The process is depicted graphically in Figure 5.2.

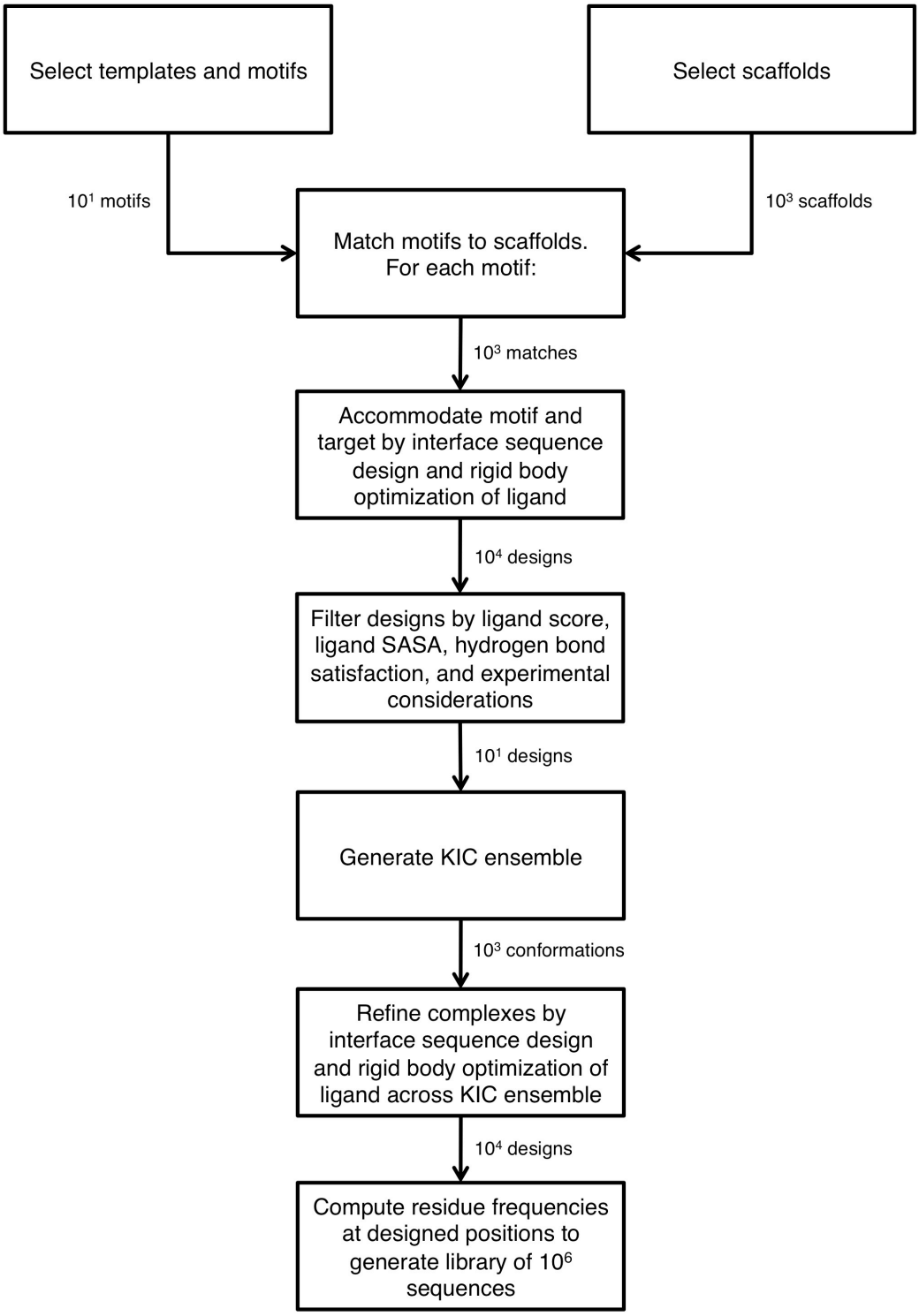

Figure 5.2: Process model for generating small molecule biosensors from existing protein complexes

I chose to pursue farnesyl pyrophosphate (FPP) as a small molecule target. As a member of the terpene class of isoprenoids, FPP is a metabolic intermediate in a number of key biosynthesis pathways[171]. Since the pathway products serve as commodity chemicals and biofuels, but the intermediates are inconspicuous (i.e., are not chromophoric, fluorescent, or essential for growth), there is a growing need for *in vivo* biosensors that enable high-throughput screening and directed evolution of pathways producing and converting isoprenoid intermediates like FPP. These practical considerations together with the availability of rich crystallographic data on FPP complexed with several different proteins (29 X-ray structures with <2.8 Å resolution in the PDB) suggest that FPP is a strong target for a proof of principle.

Methods

Selection of scaffolds

I searched the PDB for protein interface scaffolds suitable for accommodating small molecule binding sites. I focused on heterodimeric complexes to avoid issues of solubility arising from splitting homodimeric protein partners. The search criteria specified X-ray structures of heterodimeric protein complexes with ≤95% sequence identity solved at ≤2.8 Å resolution between chains of 75 to 300 residues that were expressed in *E. coli*. The resulting 612 structures were filtered to remove HETATM records and multiple densities (only the first densities listed were kept). Selenomethionines were converted to methionines. While the scaffold set is matched against the FPP binding site in the present work, in general it can be searched against any small molecule binding motif.

Selection of templates

Binding motifs are transplanted from existing small molecule–protein complexes termed 'templates'. The PDB contains 29 X-ray structures of FPP–protein complexes at <2.8 Å resolution to serve as templates. I visually inspected each interface to identify cases where 4 residues could define an encompassing portion of the FPP binding surface. 18 potential templates were discarded because FPP bound in complex with an inhibitor or other small molecule, forming a binding site that cannot easily be reproduced by amino acid side-chains. Other cases were discarded because the binding site was formed by small contributions from too many residues to define a motif. Ultimately, 4 template motifs were selected for subsequent matching and design (Table 5.1). Note that for PDB templates 1kzo and 3dpy, one of the motif residues comes from a co-associated peptide substrate, and the motif from 1t0a contains residues from a homotrimeric interface. In all cases non-polar hydrogens were added to FPP, and a single polar hydrogen was placed on the O5 oxygen.

Matching of motif residues

I scanned the scaffold set for backbones that might accommodate the small molecule target and binding motif using a geometric matching procedure[172]. For each template, the relationship between the motif side-chains and the target is uniquely defined by 6 geometric constraints, shown in Figure 5.3. The matching algorithm scans the first motif residue constraints across a set of scaffold positions (here, all positions with Cα atoms within 15 Å of the other chain). At each position, the motif

residue is placed into rotameric conformations from the Dunbrack backbone-dependent rotamer library[145]. For each side-chain conformation, the small molecule target is placed relative to the motif residue using the geometric constraints defined from the template. Conformations that place the target without introducing steric clashes between the motif side-chain, the target, and the scaffold backbone are recorded as 'hits'. The process is iterated for the remaining motif residues, comparing the clash-free target positions to those from the previous motif residues using an efficient geometric hashing technique. After all selected scaffold positions have been screened against all motif conformations, cases where hits for each motif residue place the target into the same geometric bin are recorded as 'matches'. Only matches where at least one motif side-chain is placed on a different chain than the other motif side-chains (i.e., across the scaffold interface) are considered further. Many matches may be found for a set of scaffold interface residue positions, corresponding to highly similar target placements. One such match is randomly selected for design, which is termed a 'unique match'. The numbers of unique matches arising from each template motif for FPP are shown in Table 5.1.

The quality and quantity of matching results are tuned by a number of parameters. To slightly relax the angle and torsion constraints, I sampled 5 degrees above and below the values computed from the template. There are also Euclidean and Euler parameters that determine the bin size for geometric hashing, which I set to 2.0 Å and 20.0 degrees, respectively. A bump tolerance parameter allows for some steric overlap – to be resolved in the design stage – which I set to 0.6 Å within the van der Waals radii of two contacting atoms. I also allow motif residues to be

matched by other residue types with similar side-chain moieties. The following groups of residues may be matched by any residue in the group: 'DE', 'LVI', 'FYW', and 'ST'. I set these parameters to produce a reasonable number of unique matches for design (on the order of several hundred). Evaluating matches directly is not necessarily informative, since a good match (one that faithfully reproduces the template motif) may yield poor designs due to an inability of the 'second shell' residues to accommodate the motif and target, while less precise matches may yield more promising designs after being subjected to rigid body optimization of the ligand and backbone relaxation of the scaffold by KIC. Thus, all unique matches are passed on to design.

A number of parameters control Rosetta-specific matching options including the number of side-chain conformations sampled. Command line options used for Rosetta revision 35441 follow:

```
match.linuxgccrelease -database minirosetta_database –s 1SVX.pdb -match:lig_name LG1 -
match:grid_boundary 1SVX.gridlig -match:scaffold_active_site_residues 1SVX.pos -
match:geometric_constraint_file 3bnx.cst -extra_res_fa 3bnx_LG.fa.params -
output_matches_per_group 10 -ex1 -ex2 -extrachi_cutoff 0 -euclid_bin_size 2.0 -
euler_bin_size 20.0 -bump_tolerance 0.6 -match:output_format PDB -
match:consolidate_matches -match:output_matchres_only
```

Descriptions of all Rosetta command line options appearing in this dissertation are provided in the Appendix.

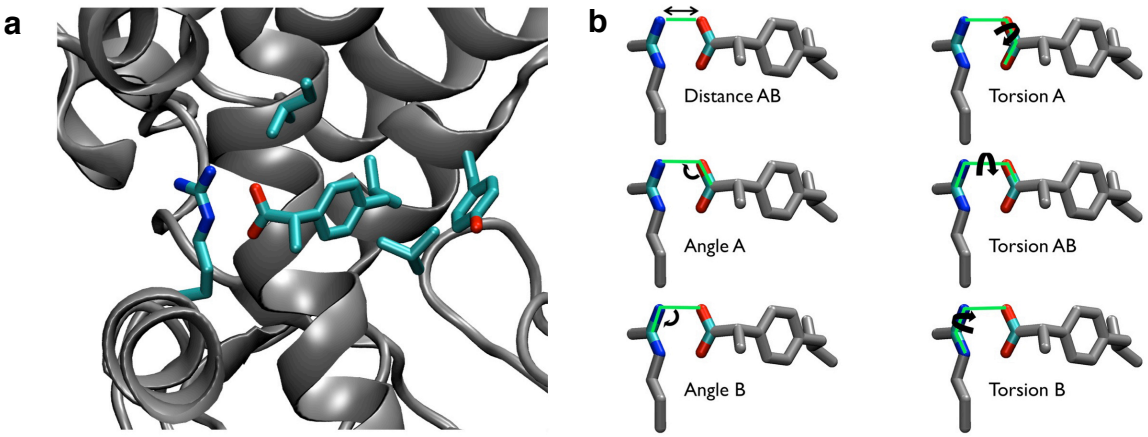

Figure 5.3: Geometric constraints employed by the matching algorithm
(**a**) A small molecule target (ibuprofen) is shown in complex with a protein monomer (Ovine COX-1, PDB 1eqg). The target (center) and four motif residues comprising the dominant binding surface are shown as sticks. (**b**) For each motif residue (here, the arginine), six geometric constraints are encoded (one distance, two angles and three torsions) that uniquely place the target given the position of the motif residue. The atoms involved in the constraint calculations are shown in polar colors with oxygen in red, nitrogen in blue and carbon in cyan; the remaining atoms are in silver.

Initial designs and analysis

The matching algorithm places the motif residues and target into a scaffold interface while avoiding clashes with the backbone (Figure 5.4a). However, the procedure will likely introduce unfavorable interactions with the residues surrounding the motif, or 'second shell' residues (Figure 5.4b). In order to accommodate the target and binding motif, I applied a protocol that iterates between rigid body optimization of the target[173] and sequence design of the second shell residues[1,2]. In the design step, all residues with a Cα atom within 6.0 Å of any ligand heavy atom are designable (they may change residue type), as well as any residue with a Cα atom within 8.0 Å of any ligand heavy atom that also has a Cβ atom closer to the ligand than the Cα atom. Additionally, all residues with a Cα atom within 10.0 Å of any ligand heavy atom are subject to repacking (simulated annealing Metropolis Monte

Carlo optimization of side-chain conformations) together with any residue with a Cα atom within 12.0 Å of the ligand with a Cβ atom that is closer to the ligand than the Cα atom. The protocol iterates rigid body optimization and sequence design 3 times, producing interfaces with more favorable interactions between the motif, target, and second shell residues (Figure 5.4c). Command line options used for Rosetta revision 36129 follow:

```
EnzdesFixBB.linuxgccrelease -database minirosetta_database -s 1BH9_R33Y94L120F121.pdb -
extra_res_fa 3bnx_LG.fa.params enzdes:detect_design_interface -enzdes:cut1 6.0 -
enzdes:cut2 8.0 -enzdes:cut3 10.0 -enzdes:cut4 12.0 -enzdes:cst_opt -enzdes:cst_design -
enzdes:cst_min -enzdes:cstfile 3bnx.cst -enzdes:bb_min -enzdes:chi_min -
enzdes:design_min_cycles 3 -ex1 -ex2 -use_input_sc  -nstruct 999 -
enzdes:start_from_random_rb_conf
```

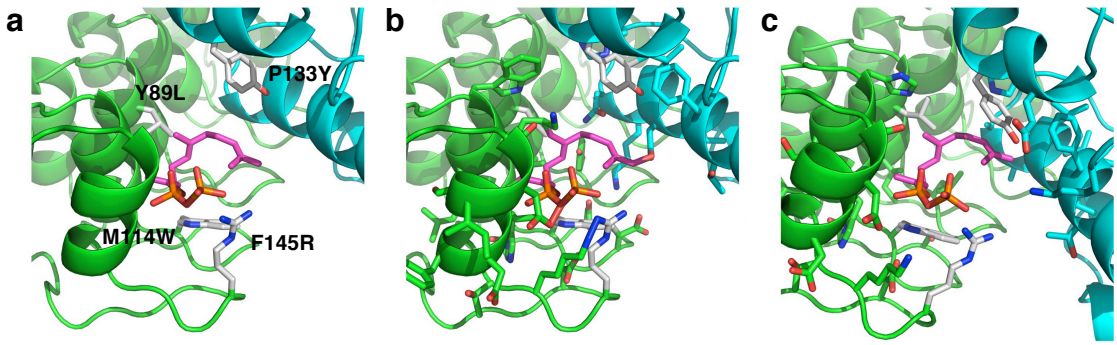

Figure 5.4: Accommodating a target and motif after geometric matching
(**a**) FPP (center in sticks, farnesyl tail in magenta) and its binding motif from aristolochene synthase (PDB 3bnx; silver sticks, labeled mutations) is transplanted into a complex (PDB 1svx) between an engineered ankyrin repeat protein (green) and maltose binding protein (cyan). (**b**) After matching, the motif and target clash with nearby side-chains in the scaffold crystal structure. (**c**) An iterative rigid body optimization and sequence design protocol reshapes the interface to accommodate the target and motif residues while maintaining the relationship between the motif and target.

For each template I produced on the order of 10^4 designs and created distributions over computed structure quality measures. Four of these distributions

from the 3bnx template are shown in Figure 5.5. The ligand score corresponds to the predicted binding energy between the ligand and scaffold interface (lower is better, Figure 5.5a). The ligand solvent accessible surface area (SASA) score (Figure 5.5b) measures the burial of the ligand from 0.0 (completely solvent exposed) to 1.0 (completely buried), and is calculated from the fraction of surface area accessible to a probe with a 1.4 Å radius, roughly the size of a water molecule. The number of hydrogen bonds between the scaffold and the ligand (Figure 5.5c) and the number of buried unsatisfied hydrogen bonds on the ligand (Figure 5.5d) are also shown. For FPP, visual inspection of representative members of the distributions across all templates suggested the following filter for selecting designs for further refinement: ligand score < -6.0, ligand SASA > 0.6, ligand hydrogen bonds > 1, unsatisfied buried ligand hydrogen bonds = 0. The number of designs passing the filter for all FPP templates is shown in Table 5.1. Passing designs were then further filtered to remove scaffolds imposing additional challenges such as cases with small molecules crystallized at the predicted target binding site that stabilize nearby structural features, like GTPases with bound nucleotides. Complexes that were purified from inclusion bodies or were expressed in the *E. coli* periplasm for crystallization were also discarded.

I also repeated the motif-directed design process using only 3 motif residues from 3bnx (R314, W308 and L184). Since each additional motif residue reduces the number of well matched scaffolds, motifs of 3 residues could match more scaffolds with greater fidelity to the template. As a disadvantage, however, the matcher transplants less of the template binding site to the scaffolds, and so maintains fewer

binding site features, including ligand burial. This issue dominated the designs resulting from matching the 3-residue motif from 3bnx. None of the designs passed the filters at the thresholds described above. The design closest to passing the filters matched the 3 residues to a colicin–immunity protein interface (PDB 1v74) with a ligand score of -3.9 and a ligand SASA score of 0.4, suggesting insufficient ligand burial and less favorable ligand–scaffold interactions. Figure 5.6 compares the surfaces of this 3-residue motif design and the 4-residue motif design that passed the filters with the best ligand score.

Ultimately, a design passing all filters with the best ligand score on a complex (PDB 1svx) between an engineered ankyrin repeat protein and maltose binding protein (MBP) with a transplanted 4-residue motif from template 3bnx was selected for refinement by generation of KIC conformational ensembles and additional sequence design.

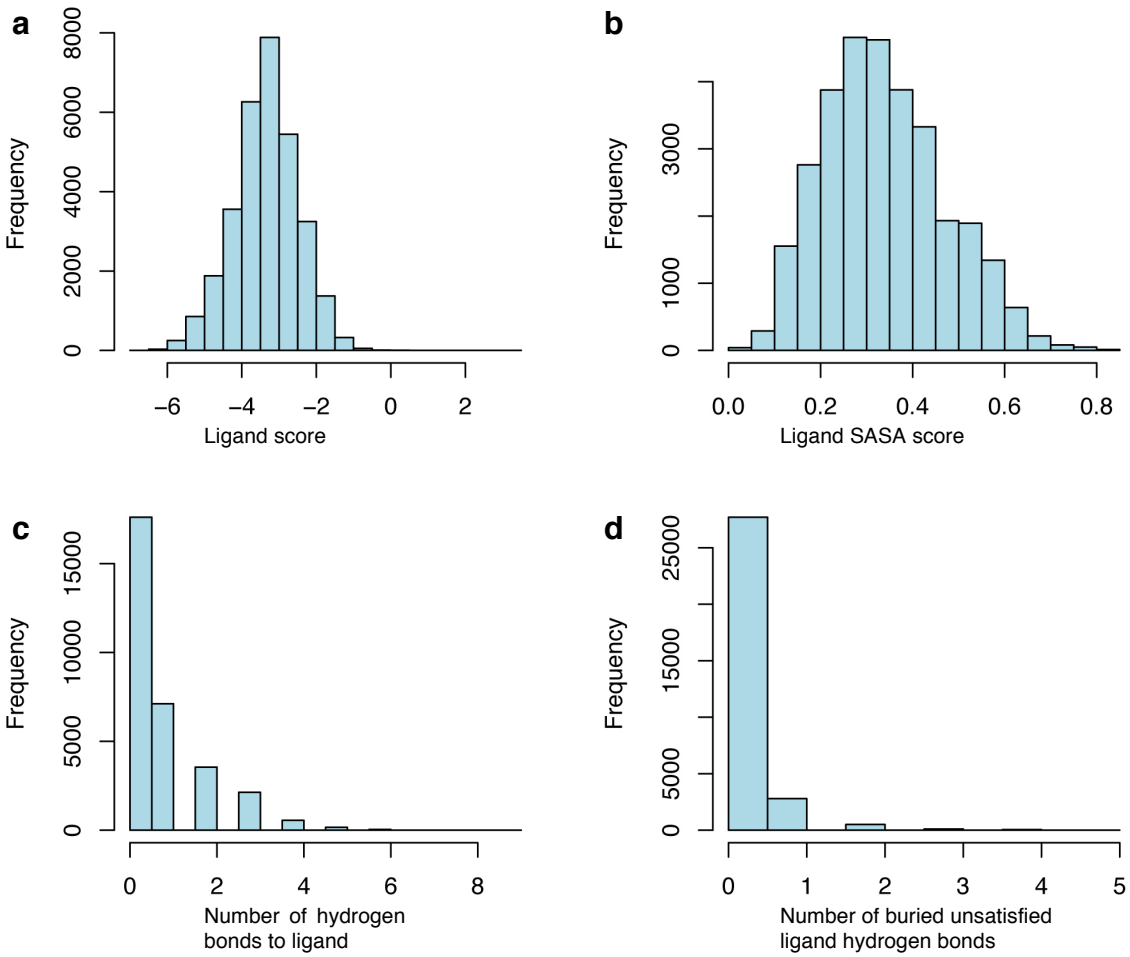

Figure 5.5: Score distributions for designs on matched scaffolds
(**a**) The ligand score measures the predicted binding energy between the target and the scaffold interface (lower score is more favorable predicted binding). (**b**) The solvent accessible surface area (SASA) score measures burial of the ligand from 0.0 (completely solvent exposed) to 1.0 (completely buried). (**c**) The number of hydrogen bonds formed between the scaffold and the target. (**d**) The number of buried unsatisfied hydrogen bond donors and acceptors on the target.

Table 5.1: Matching and design performance for FPP binding motifs
The filter requires that designs have a ligand score < -6.0, SASA score > 0.6, at least 2 hydrogen bonds with the target, and no buried unsatisfied target hydrogen bonds.

Template PDB	Protein partner	Motif residues	Resolution (Å)	Number of unique matches	Number of designs passing filter
1kzo	Protein farnesyltransferase	chain B: R291, Y251, W303. chain C: I10	2.2	79	31
1t0a	2C-Methyl-D-Erythritol-2,4-cyclodiphosphate Synthase	chain A: I101, F9. chain B: F9. chain C: F9	1.6	371	0 (1 if SASA filter relaxed to 0.5)
3bnx	Aristolochene synthase	chain A: R314, W308, L184, F153	2.1	370	81
3dpy	Protein farnesyltransferase	chain B: R291, Y251, W303. chain C: I2008	2.7	43	0 (2 if SASA filter relaxed to 0.5)

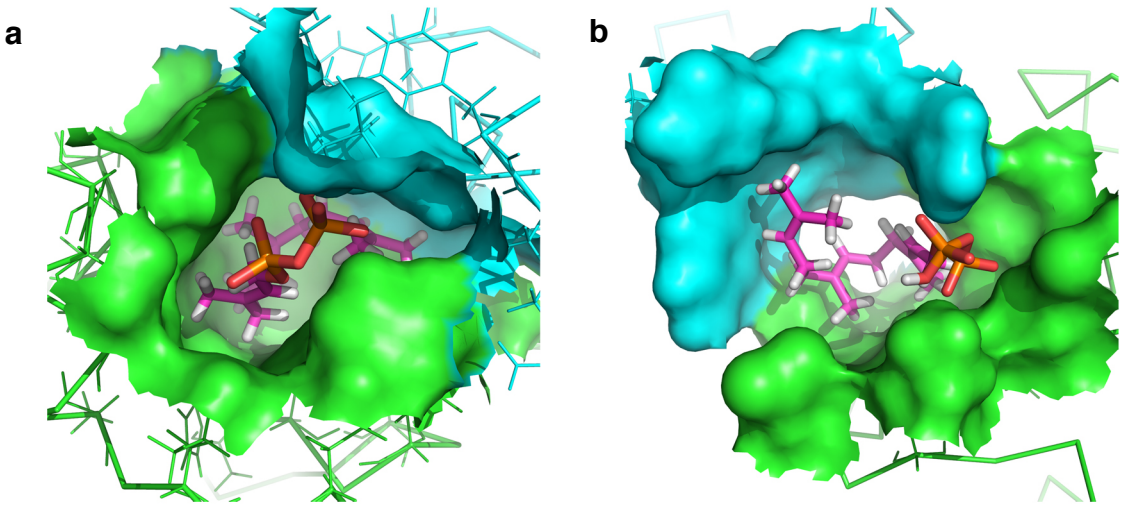

Figure 5.6: Comparison of designs resulting from 4- and 3-residue motif matches
FPP is shown in sticks with a magenta farnesyl tail in both panels. (**a**) Surface representation of the best ligand score design after matching a 4-residue FPP-binding motif from aristolochene synthase and filtering for features such as hydrogen bond satisfaction. The scaffold (PDB 1svx) consists of an engineered ankyrin repeat protein (green) and maltose binding protein (cyan). (**b**) Surface representation of the best ligand score design (also after filtering for features such as hydrogen bond satisfaction) from matching 3 of the 4 FPP-binding residues used in **a**. The farnesyl tail is less buried and makes fewer favorable interactions with the scaffold (colicin D, green and colicin D immunity protein, cyan; PDB 1v74) in comparison to **a**.

KIC ensemble design

To model the conformational adjustments that may occur in concert with sequence mutations[14-16], matched scaffold designs are subject to KIC over their entire backbones producing conformational ensembles. A second round of Rosetta sequence design is then applied across the ensembles to the side-chains surrounding the binding site in order to accommodate the transplanted motif residues and the small molecule target. Designing across a conformational ensemble, rather than a single backbone, can improve agreement between the transplanted motif and the template, and generates a diversity of predicted low-energy sequences.

I generated near-native conformational ensembles with KIC (200 conformations with 0.9 Å average rmsd to the X-ray structure) using a modified protocol compared to *de novo* loop reconstruction. The ensemble generation protocol skips the low-resolution centroid stage and sets the starting temperature at 1.2 *k*T. Instead of modeling only loop regions, KIC moves are applied to any segment of 3-12 residues in the protein. Further, to focus sampling on near-native conformations, non-pivot torsions are sampled within a vicinity of 3 degrees of their input values before each kinematic move, instead of sampling from the allowable Ramachandran space. A KIC ensemble for the ankryin repeat–MBP complex described in 'Initial designs and analysis' is shown in Figure 5.7. The command line used to generate the ensemble with Rosetta revision 36129 follows:

```
loopmodel.linuxgccrelease -database minirosetta_database -loops:refine refine_kic -
loops:input_pdb  1SVX_R134W103L78Y286__DE_19.pdb -loops:loop_file
1SVX_R134W103L78Y286__DE_19.loop -extra_res_fa 1SVX_R134W103L78Y286__DE_19_LG.fa.params -
in:file:extra_res_cen 1SVX_R134W103L78Y286__DE_19_LG.cen.params -in:file:native
1SVX_R134W103L78Y286__DE_19.pdb -loops:kic_max_seglen 12 -loops:outer_cycles 1 -
loops:refine_init_temp 1.2 -loops:vicinity_sampling -loops:vicinity_degree 3 -ex1 -ex2
```

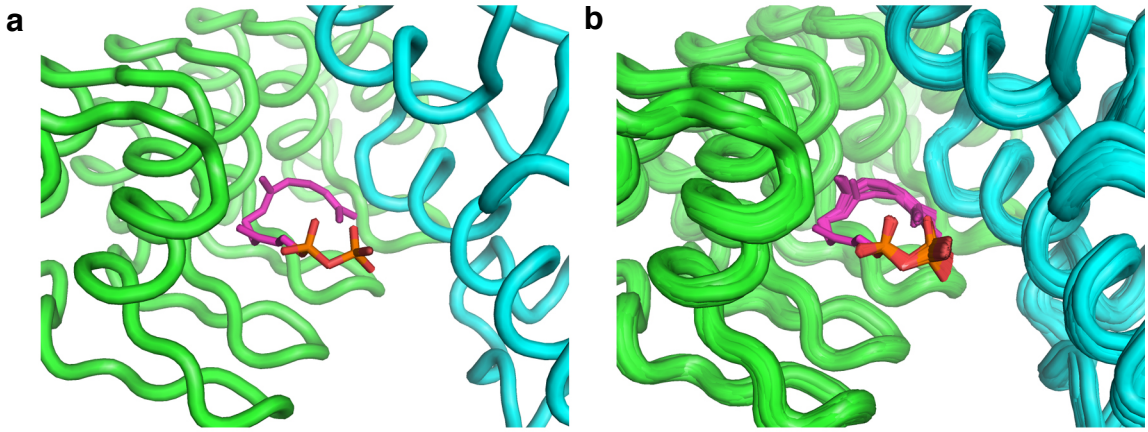

Figure 5.7: KIC conformational ensemble for a designed small molecule-binding complex (**a**) A single backbone conformation after matching FPP (colored sticks) into a scaffold complex between an engineered ankyrin repeat protein (green ribbons) and maltose binding protein (blue ribbons) is shown. (**b**) A KIC ensemble of the designed complex consisting of 200 conformations is shown.

A similar protocol for small molecule rigid body optimization and scaffold sequence design from 'Initial designs and analysis' is applied to refine every member of the KIC ensemble. The selection of residues to be redesigned, fixed, or modeled as wild-type is performed manually rather than selected by distance cutoffs to focus side-chain optimization around the motif and target while minimizing perturbations to nearby networks of scaffold side-chain interactions (redesigned positions for the 1svx scaffold are shown in Figure 5.9). Designs resulting from this step, e.g., Figure 5.8d, are predicted to improve small molecule binding as constraint satisfaction and ligand scores improve. Moreover, designing across a conformational ensemble, rather than a single backbone, produces a diverse sequence library[124] that can be assayed *in vivo* for biosensor activity.

A sample command line for this step used with Rosetta revision 36129 follows:

```
EnzdesFixBB.linuxgccrelease -database minirosetta_database -s
1_1SVX_R134W103L78Y286__DE_19_0001.pdb -resfile resfile -in:file:extra_res_fa
3bnx_LG.fa.params -enzdes:cst_opt -enzdes:cst_design -enzdes:cst_min -enzdes:cstfile
3bnx.cst -enzdes:bb_min -enzdes:chi_min -enzdes:design_min_cycles 3 -ex1 -ex2 -ex3 -ex4 -
use_input_sc -score:hbond_His_Phil_fix -score:no_his_his_pairE -nstruct 50 -
enzdes:start_from_random_rb_conf
```

Results

The process of matching, design, ensemble generation, and refinement described in 'Methods' produces dimeric protein complexes with predicted interfacial small molecule binding sites transplanted from high-resolution X-ray structures. In this work, I redesigned the interface between an engineered ankyrin repeat protein and maltose binding protein (MBP) to depend on FPP for association. Key individual steps are depicted in Figure 5.8. Of note, Figure 5.8d shows improved recapitulation of the template binding motif after designing on a KIC ensemble compared to the initial fixed backbone design, and Figure 5.8f shows the surface of a designed interfacial binding site with similar features to the surface of the template binding site (Figure 5.8e).

The success of motif-directed design is predicated largely on the fidelity of transplanted residue conformations to the original binding site. It is also critical to engineer favorable interactions between the transplanted motif and target with the surrounding 'second shell' residues. There is frequently, though not always, a tradeoff between designs that faithfully reproduce the binding motif and those predicted to have more favorable energies. For this reason, as well as due to

imperfections in computational scoring functions, it is useful to generate a broad set of sequences to assay experimentally. By computing sequence profiles from designs performed on KIC ensembles (Figure 5.9, top panel), I produced a sequence library for the ankyrin repeat–MBP complex predicted to act as a CID for FPP. Note that sequence profiles are computed from the lowest energy sequence and conformation from simulated annealing Metropolis Monte Carlo simulations in the design step.

For comparison, I also generated an equal number of sequences using a fixed backbone approach on the 3bnx crystal structure with the motif residues and target transplanted from the initial design used to generate the KIC ensemble. Sequence profiles for the fixed backbone designs are shown in Figure 5.9, bottom panel. The profiles show a greater amount of variation than might be expected from fixed backbone design due to the translational and rotational freedom of the target. At several key positions, however, the KIC ensemble designs show clear advantages. The predicted structural basis for these distinctions is shown in Figure 5.10. V152 and Y122 are well represented only in the KIC designs (Figure 5.9). Using fixed backbone design, N156 orients toward the target, similar to the wild-type conformation, where it orients toward the interface and hydrogen bonds with K122. In the KIC ensemble, small backbone adjustments allow N156 to flip out from the interface, as designed positions V152 and Y122 form the side and top of the FPP binding pocket. Although Y122 appears in the fixed backbone design with the best ligand score, V152 is absent, and a sequence library built only from the dominant residues in the fixed backbone designs would omit these potentially favorable interactions.

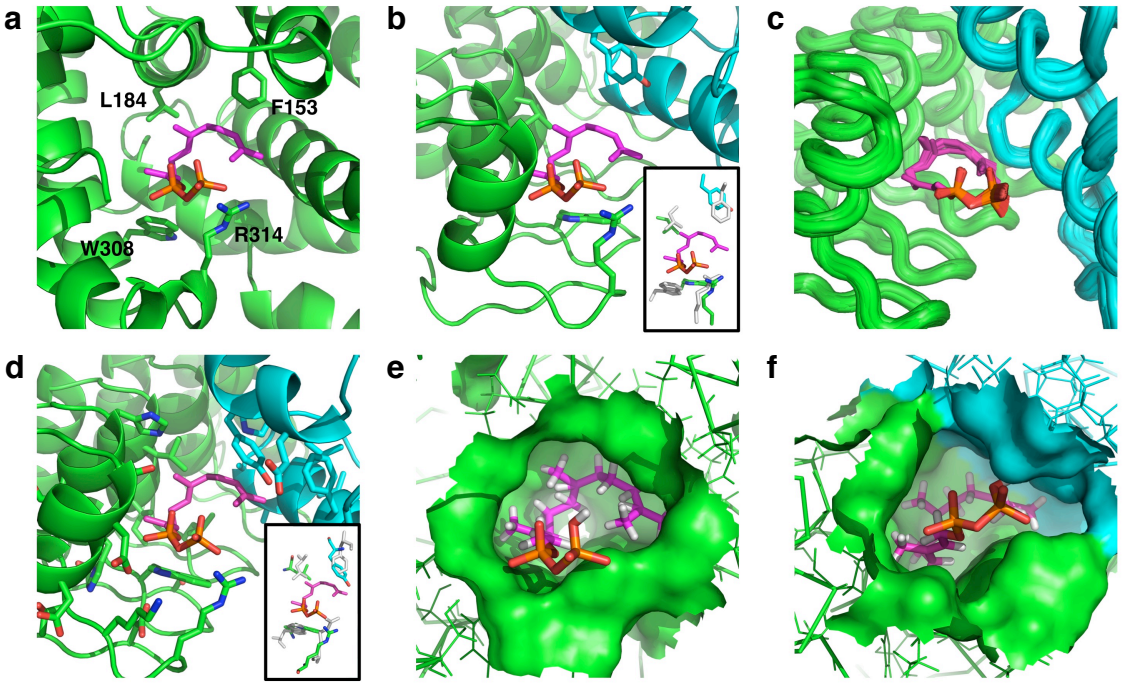

Figure 5.8: Design of protein-based biosensors for farnesyl pyrophosphate (FPP)
Panels **a** and **e** are X-ray structures, remaining panels are models. (**a**) Key side-chains forming a binding motif for FPP (shown as sticks in all panels with the farnesyl tail in magenta, hydrogens shown in **e** and **f**) are identified in a template complex with aristolochene synthase (PDB 3bnx). (**b**) Geometric constraints relating the key side-chains to FPP are encoded and matched against a scaffold dataset of heterodimeric protein interfaces. Here, the motif residues are transplanted into an existing complex (PDB 1svx) between an engineered ankyrin repeat protein (green) and maltose binding protein (cyan). The matched motif residues are shown as sticks (F153 is matched by a tyrosine for more favorable predicted solvation energy). An overlay of the matched side-chains and the original binding motif (silver) after superimposing the FPP atoms is shown inset. (**c**) Backbone flexibility is modeled by a KIC ensemble to better accommodate the motif residues and target. Ribbon representations of 200 backbones generated by KIC are shown. (**d**) Sequence design is applied to the shell of residues surrounding the motif residues and the target across all backbones in the KIC ensemble. A design that closely recapitulates the original binding motif is shown with designed side-chains in sticks. An overlay of the matched side-chains and the original binding motif (silver) after superimposing the FPP atoms is shown inset. Designing on the KIC ensemble improves overall agreement to the known motif in comparison to **b** (the arginine reorients slightly in concert with a small adjustment to the pyrophosphate moiety arising from energy minimization of FPP). (**e**) Surface representation of the wild-type FPP binding site in aristolochene synthase from **a**. (**f**) Surface representation of a designed FPP-dependent heterodimer modeled from the KIC ensemble reproduces features of the aristolochene synthase binding site in **e**.

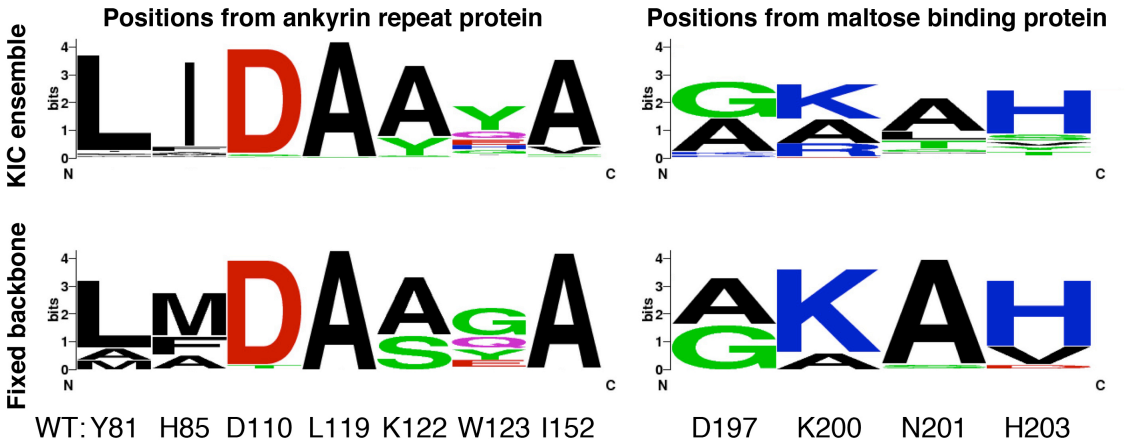

Figure 5.9: Sequence profiles for flexible and fixed backbone design of FPP biosensors
Sequence profiles for residues observed in 12 positions of a redesigned CID pair for FPP are shown for flexible (top panel) and fixed (bottom panel) backbone designs. Wild-type scaffold residues (PDB 1svx) are shown at the bottom. Profiles are computed over 10^4 designs for each protocol. KIC sequences were produced in equal numbers across the members of the KIC ensemble, and fixed backbone design was applied to the 3bnx crystal structure with the motif residues and FPP transplanted from the same initial design used to generate the KIC ensemble. Valine at position 152 and tyrosine at position 122 are predicted only in the KIC ensemble profiles.

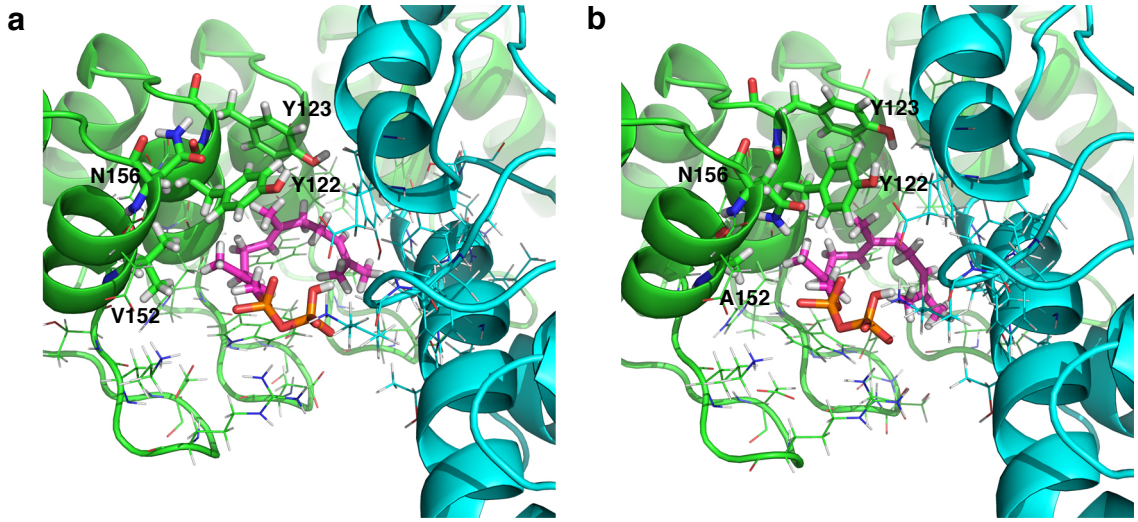

Figure 5.10: Comparison of best scoring models from flexible and fixed backbone design
(**a**) The KIC ensemble design with the best ligand score is shown. N156 flips away from the target, making room for V152 and Y122 to form the side and top of the FPP binding pocket. V152 and Y122 are well represented only in the KIC designs. (**b**) The fixed backbone design with the best ligand score is shown. N156 stays oriented toward the interface, similar to the wild-type crystal structure (PDB 1svx). This configuration partially buries the N156 amide group in a hydrophobic environment and leaves room only for alanine or glycine at position 152.

The dominant residues in the KIC sequence profiles can be used to design a sequence library. I am currently pursuing this approach to produce FPP biosensors in collaboration with Jay Keasling's lab. Fuzhong Zheng from the Keasling lab and I have designed a library of size 10^6 using degenerate codons that well covers the observed sequence profiles. Note that the library includes most of the dominant residues, but does not reproduce their relative frequencies, which is difficult to control. The library also does not take into account covariation among the residues that may appear in the designs. An advantage, however, is that the CID library can be genetically fused to split-DHFR as a life-or-death experimental selection for complexes with maximal dependence on FPP for association. Fuzhong Zheng is currently testing the designed library *in vivo*, together with *in vitro* characterization of several of the top scoring designs by FPP-dependent pull-down assays. The Keasling lab plans to use the most sensitive DHFR–activating biosensors to optimize *E. coli* cells for high yield of FPP conversion by directed evolution.

Chapter 6

Conclusion

This dissertation has addressed fundamental issues in high-resolution sampling of backbone conformations and evaluating the favorability of inter- and intramolecular interactions to extend computational protein design techniques toward flexible backbone models. Chapter 2 assessed the strengths of hydrogen bonds involving residues that have been phosphorylated, a key post-translational modification that frequently affects protein function by inducing conformational perturbations. The chapter compared calculations of free energies using multiple levels of theory, consisting of explicit solvent molecular dynamics, implicit solvent molecular mechanics, and quantum mechanics, and discussed the strengths and shortcomings of each approach in computing hydrogen bonding strengths across a panel of donors, acceptors and orientations. Chapter 3 introduced the sampling problem in flexible backbone design, and summarized current and prior approaches together with recently enabled applications. Chapter 4 introduced a robotics-inspired method for modeling protein conformations, and demonstrated sub-angstrom accuracy in reconstructing loops in monomeric proteins and protein interfaces, with the latter result suggesting that the method might be used to model functional

conformational plasticity. Chapter 5 coupled the flexible backbone methods from Chapter 4 with computational sequence design to predict modular, protein-based small molecule biosensors, which are currently being validated experimentally. The modular nature of the described sensors enables the constituent protein pairs to drive association of any split reporter, in principle, allowing for a diversity of fluorescent, enzymatic, and transcriptional output behaviors.

While the biosensors described in Chapter 5 are designed with outputs directed toward efficient cellular production of their targets, future work could focus on new approaches to drive cellular processes by chemical induction of dimerization. For example, it may be useful to enable spatiotemporal control of a transcriptional activator, or to localize two components of a signaling pathway. The known natural chemicals that induce dimerization events frequently have side-effects (as noted in Chapter 5 many derive from natural toxins), and light-inducible interactions cannot be used in animals like mice. A new set of rationally designed protein interactions induced by non-toxic chemicals would thus be invaluable to deconstructing cellular processes.

In general, flexible backbone design provides powerful tools to predict vast numbers of sequences consistent with a given protein fold, subject to stability and binding constraints. It is now critical to consider techniques to validate these developments in ways that improve understanding of current shortcomings. While functional characterization of folding or binding activities of single or a few designs demonstrates outstanding achievements in engineering, such validation provides only case studies. More comprehensive comparisons of medium- to large-scale

libraries computationally designed against functional constraints to experimentally screened sequence libraries are needed. By examining over- and under-represented portions of experimentally observed sequence space, such studies can reveal fundamental flaws in computational scoring functions and sampling techniques. Only when methods improve to consistently identify functional molecules as top predictions will researchers likely adopt these approaches more broadly for engineering new and modified protein functions. In addition, difficult engineering goals can potentially be met by bringing flexible backbone methods to the powerful combination of computational design and experimental selection[117,174].

Presently, the major design successes using flexible backbone approaches have focused on creating new structures[11,13] rather than functional conformations. Recent flexible backbone methods have produced sequences that fold into predicted stable loop structures[98,99] or adopt multiple distinct conformations[175]. With the development of improved high-resolution methods to represent structural variability in design frameworks, the field is now poised to make significant contributions to broadening the functional repertoire by engineering new functional conformations into proteins. For instance, the current seminal successes in enzyme design[1,2] involve fixed backbone models with catalytic activity toward a single substrate. Flexible backbone methods could be used to find backbone conformations providing a range of selectivity toward multiple related substrates, yielding modular enzymes that may be linked into pathways to perform sequential partial reactions. Other functional conformations might involve interface loops in signaling proteins. Flexible elements in such switch proteins (Figure 4.2c) could be redesigned[176] to

vary signaling activity in response to posttranslational modifications or to alter binding to downstream effectors. Further, flexible backbone methods could reshape or extend protein interfaces to increase selectivity and affinity.

There is also much to be gained by further exploring the relationship between protein evolution and computational design. It has been suggested that related protein folds evolve through small numbers of obligate transitional sequences[177]. Computationally, as Kuhlman and colleagues have demonstrated[13], it is possible to produce entirely new topologies without progressive transitions through sequence space. Nevertheless, protein design methodologies might benefit from following an evolutionary model[178] wherein new sequences and backbone conformations are obtained through a progression of functional design checkpoints (e.g., designs to carry out successive partial reactions). Such an approach might demonstrate particular utility in producing enzymes that bind new substrates or perform new reactions[179], since the step-wise accumulation of partial reaction mechanisms through evolutionarily accessible trajectories would be explicitly modeled.

Bibliography

1. Rothlisberger, D. et al. Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190-5 (2008).
2. Jiang, L. et al. De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-91 (2008).
3. Lippow, S.M., Wittrup, K.D. & Tidor, B. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat Biotechnol* **25**, 1171-6 (2007).
4. Grigoryan, G., Reinke, A.W. & Keating, A.E. Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature* **458**, 859-864 (2009).
5. Ponder, J.W. & Richards, F.M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* **193**, 775-91 (1987).
6. Richardson, J.S. & Richardson, D.C. The de novo design of protein structures. *Trends Biochem Sci* **14**, 304-9 (1989).
7. DeGrado, W.F., Wasserman, Z.R. & Lear, J.D. Protein design, a minimalist approach. *Science* **243**, 622-8 (1989).
8. Hellinga, H.W. & Richards, F.M. Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry. *J Mol Biol* **222**, 763-85 (1991).

9. Dahiyat, B.I. & Mayo, S.L. De novo protein design: fully automated sequence selection. *Science* **278**, 82-7 (1997).

10. Su, A. & Mayo, S.L. Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci* **6**, 1701-7 (1997).

11. Harbury, P.B., Plecs, J.J., Tidor, B., Alber, T. & Kim, P.S. High-resolution protein design with backbone freedom. *Science* **282**, 1462-7 (1998).

12. Desjarlais, J.R. & Handel, T.M. Side-chain and backbone flexibility in protein core design. *J Mol Biol* **290**, 305-18 (1999).

13. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-8 (2003).

14. Baldwin, E.P., Hajiseyedjavadi, O., Baase, W.A. & Matthews, B.W. The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. *Science* **262**, 1715-8 (1993).

15. Lim, W.A., Hodel, A., Sauer, R.T. & Richards, F.M. The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proc Natl Acad Sci U S A* **91**, 423-7 (1994).

16. Bordner, A.J. & Abagyan, R.A. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins* **57**, 400-13 (2004).

17. Kortemme, T., Morozov, A.V. & Baker, D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol* **326**, 1239-59 (2003).

18. Audette, G.F. et al. The 1.9 A resolution structure of phospho-serine 46 HPr from Enterococcus faecalis. *J Mol Biol* **303**, 545-53 (2000).

19. Patel, A.J. & Honore, E. Properties and modulation of mammalian 2P domain K+ channels. *Trends Neurosci* **24**, 339-46 (2001).

20. Martens, J.R., Kwak, Y.G. & Tamkun, M.M. Modulation of Kv channel alpha/beta subunit interactions. *Trends Cardiovasc Med* **9**, 253-8 (1999).

21. Vermeulen, K., Van Bockstaele, D.R. & Berneman, Z.N. The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer. *Cell Prolif* **36**, 131-49 (2003).

22. Feng, M.H., Philippopoulos, M., MacKerell, A.D. & Lim, C. Structural characterization of the phosphotyrosine binding region of a high-affinity SH2 domain-phosphopeptide complex by molecular dynamics simulation and chemical shift calculations. *Journal of the American Chemical Society* **118**, 11265-11277 (1996).

23. Johnson, L.N. & Lewis, R.J. Structural basis for control by phosphorylation. *Chemical Reviews* **101**, 2209-2242 (2001).

24. Johnson, L.N. & Oreilly, M. Control by phosphorylation. *Current Opinion in Structural Biology* **6**, 762-769 (1996).

25. Anderson, D.E., Becktel, W.J. & Dahlquist, F.W. Ph-Induced Denaturation of Proteins - a Single Salt Bridge Contributes 3-5 Kcal Mol to the Free-Energy of Folding of T4-Lysozyme. *Biochemistry* **29**, 2403-2408 (1990).

26. Sun, D.P., Sauer, U., Nicholson, H. & Matthews, B.W. Contributions of engineered surface salt bridges to the stability of T4 lysozyme determined by directed mutagenesis. *Biochemistry* **30**, 7142-53 (1991).

27. Schneider, J.P., Lear, J.D. & DeGrado, W.F. A designed buried salt bridge in a heterodimeric coiled coil. *Journal of the American Chemical Society* **119**, 5742-5743 (1997).

28. Waldburger, C.D., Schildbach, J.F. & Sauer, R.T. Are Buried Salt Bridges Important for Protein Stability and Conformational Specificity. *Nature Structural Biology* **2**, 122-128 (1995).

29. Lyubartsev, A.P. & Laaksonen, A. Osmotic and activity coefficients from effective potentials for hydrated ions. *Physical Review E* **55**, 5689-5696 (1997).

30. Martorana, V., La Fata, L., Bulone, D. & San Biagio, P.L. Potential of mean force between two ions in a sucrose rich aqueous solution. *Chemical Physics Letters* **329**, 221-227 (2000).

31. Luo, R., David, L., Hung, H., Devaney, J. & Gilson, M.K. Strength of Solvent-Exposed Salt-Bridges. *Journal of Physical Chemistry B* **103**, 727-736 (1999).

32. Masunov, A. & Lazaridis, T. Potentials of Mean Force between Ionizable Amino Acid Side Chains in Water. *Journal of the American Chemical Society* **125**, 1722-1730 (2003).

33. Gilson, M.K. Theory of Electrostatic Interactions in Macromolecules. *Current Opinion in Structural Biology* **5**, 216-223 (1995).

34. Ghosh, A., Rapp, C.S. & Friesner, R.A. Generalized born model based on a surface integral formulation. *Journal of Physical Chemistry B* **102**, 10983-10990 (1998).

35. Chorny, I., Dill, K.A. & Jacobson, M.P. Surfaces Affect Ion Pairing. *Journal of Physical Chemistry B* **2005**, 24056-24060 (2005).

36. Asthagiri, D., Schure, M.R. & Lenhoff, A.M. Calculation of Hydration Effects in the Binding of Anionic Ligands to Basic Proteins. *Journal of Physical Chemistry B* **104**, 8753-8761 (2000).

37. Yu, Z., Jacobson, M.P., Rapp, C.S. & Friesner, R.A. First-Shell Solvation of Ion Pairs: Correction of Systematic Errors in Implicit Solvent Models. *Journal of Chemical Physics* **108**, 6643-6654 (2004).

38. Saebo, S. & Pulay, P. Local Treatment of Electron Correlation. *Annual Review of Physical Chemistry* **44**, 213-236 (1993).

39. Singh, J., Thornton, J.M., Snarey, M. & Campbell, S.F. The geometries of interacting arginine-carboxyls in proteins. *FEBS Letters* **224**, 161-171 (1987).

40. Saenger, W. & Wagner, K.G. An X-ray Study of the Hydrogen Bonding in the Crystalline L-Arginine Phosphate Monohydrate Complex. *Acta Crystallographica* **B28**, 2237-2244 (1972).

41. Lewin, S. Ionic Linkages in Protein Interactions. *Journal of Theoretical Biology* **23**, 279-284 (1969).

42. Schug, K.A. & Lindner, W. Noncovalent binding between guanidinium and anionic groups: Focus on biological- and synthetic-based

arginine/guanidinium interactions with phosph[on]ate and sulf[on]ate residues. *Chemical Reviews* **105**, 67-113 (2005).

43. Mavri, J. & Vogel, H.J. Ion pair formation of phosphorylated amino acids and lysine and arginine side chains: A theoretical study. *Proteins-Structure Function and Genetics* **24**, 495-501 (1996).

44. Deerfield, D.W., Nicholas, H.B., Hiskey, R.G. & Pedersen, L.G. Salt or Ion Bridges in Biological-Systems - a Study Employing Quantum and Molecular Mechanics. *Proteins-Structure Function and Genetics* **6**, 168-192 (1989).

45. Charbon, G., Breunig, K.D., Wattiez, R., Vandenhaute, J. & Noel-Georis, I. Key role of Ser562/661 in Snf1-dependent regulation of Cat8p in Saccharomyces cerevisiae and Kluyveromyces lactis. *Mol Cell Biol* **24**, 4083-91 (2004).

46. Kassenbrock, C.K. & Anderson, S.M. Regulation of ubiquitin protein ligase activity in c-Cbl by phosphorylation-induced conformational change and constitutive activation by tyrosine to glutamate point mutations. *J Biol Chem* **279**, 28017-27 (2004).

47. Huang, W. & Erikson, R.L. Constitutive activation of Mek1 by mutation of serine phosphorylation sites. *Proc Natl Acad Sci U S A* **91**, 8960-3 (1994).

48. Klose, K.E., Weiss, D.S. & Kustu, S. Glutamate at the site of phosphorylation of nitrogen-regulatory protein NTRC mimics aspartyl-phosphate and activates the protein. *J Mol Biol* **232**, 67-78 (1993).

49. McCabe, T.J., Fulton, D., Roman, L.J. & Sessa, W.C. Enhanced electron flux and reduced calmodulin dissociation may explain "calcium-independent" eNOS activation by phosphorylation. *J Biol Chem* **275**, 6123-8 (2000).

50. Zhang, J., Zhang, F., Ebert, D., Cobb, M.H. & Goldsmith, E.J. Activity of the MAP kinase ERK2 is controlled by a flexible surface loop. *Structure* **3**, 299-307 (1995).

51. Mansour, S.J., Candia, J.M., Matsuura, J.E., Manning, M.C. & Ahn, N.G. Interdependent domains controlling the enzymatic activity of mitogen-activated protein kinase kinase 1. *Biochemistry* **35**, 15529-36 (1996).

52. Garcia-Echeverria, C. Antagonists of the Src homology 2 (SH2) domains of Grb2, Src, Lck and ZAP-70. *Current Medicinal Chemistry* **8**, 1589-1604 (2001).

53. Cody, W.L., Lin, Z.W., Panek, R.L., Rose, D.W. & Rubin, J.R. Progress in the development of inhibitors of SH2 domains. *Current Pharmaceutical Design* **6**, 59-98 (2000).

54. Calnan, B.J., Tidor, B., Biancalana, S., Hudson, D. & Frankel, A.D. Arginine-mediated RNA recognition: the arginine fork. *Science* **252**, 1167-71 (1991).

55. Frigyes, D., Alber, F., Pongor, S. & Carloni, P. Arginine-phosphate salt bridges in protein–DNA complexes: a Car–Parrinello study. *Journal of Molecular Structure (Theochem)* **574**, 39-45 (2001).

56. Van Der Spoel, D. et al. GROMACS: fast, flexible, and free. *J Comput Chem* **26**, 1701-18 (2005).

57. G.A. Kaminski, R.A.F., J. Tirado-Rives, and W.L. Jorgensen. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with acccurate quatum checmical calculations on peptides. *Journal of Physical Chemistry B* **105**, 6474-6487 (2001).

58. Wong, S.E., Bernacki, K. & Jacobson, M. Competition between intramolecular hydrogen bonds and solvation in phosphorylated peptides: Simulations with explicit and implicit solvent. *Journal of Physical Chemistry B* **109**, 5249-5258 (2005).

59. Groban, E.S., Narayanan, A. & Jacobson, M.P. Conformational changes in protein loops and helices induced by post-translational phosphorylation. *PLoS Comput Biol* **2**, e32 (2006).

60. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. & Klein, M.L. Comparison of Simple Potential Functions for Simulating Liquid Water. *Journal of Chemical Physics* **79**, 926-935 (1983).

61. Darden, T., York, D. & Pedersen, L. Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *Journal of Chemical Physics* **98**, 10089-10092 (1993).

62. Swope, W.C., Andersen, H.C., Berens, P.H. & Wilson, K.R. A Computer-Simulation Method for the Calculation of Equilibrium-Constants for the Formation of Physical Clusters of Molecules - Application to Small Water Clusters. *Journal of Chemical Physics* **76**, 637-649 (1982).

63. Verlet, L. Computer Experiments on Classical Fluids .I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review* **159**, 98 (1967).

64. Hoover, W.G. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review. A* **31**, 1695-1697 (1985).

65. Berendsen, H.J.C., Postma, J.P.M., Vangunsteren, W.F., Dinola, a. & Haak, J.R. Molecular-Dynamics with Coupling to an External Bath. *Journal of Chemical Physics* **81**, 3684-3690 (1984).

66. Torrie, G.M. & Valleau, J.P. Nonphysical sampling distributions in Monte Carlo free-energy estimation - Umbrella Sampling. *Journal of Computational Physics* **23**, 187-199 (1977).

67. Hess, B., Bekker, H., Berendsen, H.J.C. & Fraaije, J.G.E.M. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**, 1463-1472 (1997).

68. Souaille, M. & Roux, B. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput. Phys. Commun* **135**, 40-57 (2001).

69. Kumar, S., Bouzida, D., Swendsen, R.H., Kollman, P. & Rosenberg, J.M. The weighted histogram analysis method for free-energy calculations on biomolecules. I: The method. *Journal of Computational Chemistry* **13**, 1011-1021 (1992).

70. Allen, M.P. & Tildesley, D.J. *Computer Simulation of Liquids*, (Oxford University Press, 1989).

71. Nicholls, a. & Honig, B. A Rapid Finite-Difference Algorithm, Utilizing Successive over-Relaxation to Solve the Poisson-Boltzmann Equation. *Journal of Computational Chemistry* **12**, 435-445 (1991).

72. Jaguar. 5.0 edn (Schrodinger, L.L.C., Portland, OR, 1991-2003).

73. Tannor, D.J. et al. Accurate First Principles Calculation of Molecular Charge-Distributions and Solvation Energies from Ab-Initio Quantum-Mechanics and Continuum Dielectric Theory. *Journal of the American Chemical Society* **116**, 11875-11882 (1994).

74. Chirlian, L.E. & Francl, M.M. Atomic Charges Derived from Electrostatic Potentials - a Detailed Study. *Journal of Computational Chemistry* **8**, 894-905 (1987).

75. Woods, R.J., Khalil, M., Pell, W., Moffat, S.H. & Smith, V.H. Derivation of Net Atomic Charges from Molecular Electrostatic Potentials. *Journal of Computational Chemistry* **11**, 297-310 (1990).

76. Beglov, D. & Roux, B. Finite Representation of an Infinite Bulk System - Solvent Boundary Potential for Computer-Simulations. *Journal of Chemical Physics* **100**, 9050-9063 (1994).

77. Smith, D.E. & Dang, L.X. Computer-Simulations of Nacl Association in Polarizable Water. *Journal of Chemical Physics* **100**, 3757-3766 (1994).

78. Rozanska, X. & Chipot, C. Modeling ion-ion interaction in proteins: A molecular dynamics free energy calculation of the guanidinium-acetate association. *Journal of Chemical Physics* **112**, 9691-9694 (2000).

79. Breslow, R., Belvedere, S., Gershell, L. & Leung, D. The chelate effect in binding, catalysis, and chemotherapy. *Pure Appl. Chem.* **72**, 333-342 (2000).

80. Brooks, B.R. et al. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry* **4**, 187-217 (1983).

81. Rodinger, T., Howell, P.L. & Pomes, R. Absolute free energy calculations by thermodynamic integration in four spatial dimensions. *Journal of Chemical Physics* **123**(2005).

82. Hunenberger, P.H. & McCammon, J.A. Effect of artificial periodicity in simulations of biomolecules under Ewald boundary conditions: a continuum electrostatics study. *Biophysical Chemistry* **78**, 69-88 (1999).

83. Cornell, W.D. et al. A 2Nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *Journal of the American Chemical Society* **117**, 5179-5197 (1995).

84. Cramer, C.J. & Truhlar, D.G. Am1-Sm2 and Pm3-Sm3 Parameterized Scf Solvation Models for Free-Energies in Aqueous-Solution. *Journal of Computer-Aided Molecular Design* **6**, 629-666 (1992).

85. Springs, B. & Haake, P. Equilibrium-Constants for Association of Guanidinium and Ammonium-Ions with Oxyanions - Effect of Changing Basicity of Oxyanion. *Bioorganic Chemistry* **6**, 181-190 (1977).

86. Hoch, J.A. & Silhavy, T.J. (eds.). *Two-Component Signal Transduction*, (ASM Press, Washington, DC, 1995).

87. Berman, H.M. et al. The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42 (2000).

88. Halgren, T.A. & Damm, W. Polarizable Force Fields. *Current Opinion in Structural Biology* **11**, 236-242 (2001).

89. Yu, H. & van Gunsteren, W.F. Charge-on-spring polarizable water models revisited: From water clusters to liquid water to ice. *Journal of Chemical Physics* **121**, 9549-9564 (2004).

90. Jagielska, A., Wroblewska, L. & Skolnick, J. Protein model refinement using an optimized physics-based all-atom force field. *Proc Natl Acad Sci U S A* **105**, 8268-73 (2008).

91. Simons, K.T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* **268**, 209-25 (1997).

92. Simons, K.T., Bonneau, R., Ruczinski, I. & Baker, D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **Suppl 3**, 171-6 (1999).

93. Rohl, C.A., Strauss, C.E.M., Misura, K.M.S. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66--93 (2004).

94. Bradley, P., Misura, K.M.S. & Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868-71 (2005).

95. Schueler-Furman, O., Wang, C., Bradley, P., Misura, K. & Baker, D. Progress in Modeling of Protein Structures and Interactions. *Science* **310**, 638-642 (2005).

96. Wang, C., Bradley, P. & Baker, D. Protein-protein docking with backbone flexibility. *J Mol Biol* **373**, 503--519 (2007).

97. Canutescu, A.A. & Dunbrack, R.L.J. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci* **12**, 963--972 (2003).

98. Hu, X., Wang, H., Ke, H. & Kuhlman, B. High-resolution design of a protein loop. *Proc Natl Acad Sci U S A* **104**, 17668--17673 (2007).

99. Murphy, P.M., Bolduc, J.M., Gallaher, J.L., Stoddard, B.L. & Baker, D. Alteration of enzyme specificity by computational loop remodeling and design. *Proc Natl Acad Sci U S A* **106**, 9215-20 (2009).

100. Fung, H.K., Floudas, C.A., Taylor, M.S., Zhang, L. & Morikis, D. Toward full-sequence de novo protein design with flexible templates for human beta-defensin-2. *Biophys J* **94**, 584-99 (2008).

101. Georgiev, I. & Donald, B.R. Dead-end elimination with backbone flexibility. *Bioinformatics* **23**, i185-94 (2007).

102. Desmet, J., Demaeyer, M., Hazes, B. & Lasters, I. The Dead-End Elimination Theorem and Its Use in Protein Side-Chain Positioning. *Nature* **356**, 539-542 (1992).

103. Georgiev, I., Lilien, R.H. & Donald, B.R. A novel minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. *Research in Computational Molecular Biology, Proceedings* **3909**, 530-545 (2006).

104. Davis, I.W., Arendall, W.B.r., Richardson, D.C. & Richardson, J.S. The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure* **14**, 265-74 (2006).

105. Georgiev, I., Keedy, D., Richardson, J.S., Richardson, D.C. & Donald, B.R. Algorithm for backrub motions in protein design. *Bioinformatics* **24**, i196-204 (2008).

106. Smith, C.A. & Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J Mol Biol* **380**, 742-56 (2008).

107. Yin, S., Ding, F. & Dokholyan, N.V. Eris: an automated estimator of protein stability. *Nat Methods* **4**, 466-7 (2007).

108. Benedix, A., Becker, C.M., de Groot, B.L., Caflisch, A. & Bockmann, R.A. Predicting free energy changes using structural ensembles. *Nat Methods* **6**, 3-4 (2009).

109. de Groot, B.L. et al. Prediction of protein conformational freedom from distance constraints. *Proteins* **29**, 240-51 (1997).

110. Guerois, R., Nielsen, J.E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* **320**, 369-87 (2002).

111. Koehl, P. & Levitt, M. Protein topology and stability define the space of allowed sequences. *Proc Natl Acad Sci U S A* **99**, 1280-5 (2002).

112. Taverna, D.M. & Goldstein, R.A. Why are proteins so robust to site mutations? *J Mol Biol* **315**, 479-84 (2002).

113. Wagner, A. Robustness, evolvability, and neutrality. *FEBS Lett* **579**, 1772-8 (2005).

114. Gupta, R.D. & Tawfik, D.S. Directed enzyme evolution via small and effective neutral drift libraries. *Nat Methods* **5**, 939-42 (2008).

115. Xia, Y. & Levitt, M. Simulating protein evolution in sequence and structure space. *Curr Opin Struct Biol* **14**, 202-7 (2004).

116. Fu, X., Apgar, J.R. & Keating, A.E. Modeling backbone flexibility to achieve sequence diversity: the design of novel alpha-helical ligands for Bcl-xL. *J Mol Biol* **371**, 1099-117 (2007).

117. Hayes, R.J. et al. Combining computational and experimental screening for rapid optimization of protein properties. *Proc Natl Acad Sci U S A* **99**, 15926-31 (2002).

118. Larson, S.M., England, J.L., Desjarlais, J.R. & Pande, V.S. Thoroughly sampling sequence space: large-scale protein design of structural ensembles. *Protein Sci* **11**, 2804-13 (2002).

119. Pei, J., Dokholyan, N.V., Shakhnovich, E.I. & Grishin, N.V. Using protein design for homology detection and active site searches. *Proc Natl Acad Sci U S A* **100**, 11361-6 (2003).

120. Larson, S.M., Garg, A., Desjarlais, J.R. & Pande, V.S. Increased detection of structural templates using alignments of designed sequences. *Proteins* **51**, 390-6 (2003).

121. Saunders, C.T. & Baker, D. Recapitulation of protein family divergence using flexible backbone protein design. *J Mol Biol* **346**, 631-44 (2005).

122. Friedland, G.D., Lakomek, N.A., Griesinger, C., Meiler, J. & Kortemme, T. A Correspondence Between Solution-State Dynamics of an Individual Protein

and the Sequence and Conformational Diversity of its Family. *Plos Computational Biology* **5**, - (2009).

123. Ding, F. & Dokholyan, N.V. Emergence of protein fold families through rational design. *PLoS Comput Biol* **2**, e85 (2006).

124. Humphris, E.L. & Kortemme, T. Prediction of protein-protein interface sequence diversity using flexible backbone computational protein design. *Structure* **16**, 1777-88 (2008).

125. Pal, G., Kouadio, J.-L.K., Artis, D.R., Kossiakoff, A.A. & Sidhu, S.S. Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. *J Biol Chem* **281**, 22378-85 (2006).

126. Go, N. & Scheraga, H.A. Ring Closure and Local Conformational Deformations of Chain Molecules. *Macromolecules* **3**, 178-187 (1970).

127. Coutsias, E.A., Seok, C., Jacobson, M.P. & Dill, K.A. A kinematic view of loop closure. *J Comput Chem* **25**, 510-28 (2004).

128. Cortes, J., Simeon, T., Remaud-Simeon, M. & Tran, V. Geometric algorithms for the conformational analysis of long protein loops. *J Comput Chem* **25**, 956-67 (2004).

129. Lee, A., Streinu, I. & Brock, O. A methodology for efficiently sampling the conformation space of molecular structures. *Phys Biol* **2**, S108-15 (2005).

130. Noonan, K., O'Brien, D. & Snoeyink, J. Probik: Protein Backbone Motion by Inverse Kinematics. *The International Journal of Robotics Research* **24**, 971-982 (2005).

131. Milgram, R.J., Liu, G. & Latombe, J.C. On the structure of the inverse kinematics map of a fragment of protein backbone. *J Comput Chem* **29**, 50-68 (2008).

132. Coutsias, E.A., Seok, C., Wester, M.J. & Dill, K.A. Resultants and Loop Closure. *International Journal of Quantum Chemistry* **106**, 176-189 (2005).

133. Sellers, B.D., Zhu, K., Zhao, S., Friesner, R.A. & Jacobson, M.P. Toward better refinement of comparative models: predicting loops in inexact environments. *Proteins* **72**, 959-971 (2008).

134. Felts, A.K. et al. Prediction of Protein Loop Conformations Using the AGBNP Implicit Solvent Model and Torsion Angle Sampling. *J. Chem. Theory Comput.* **4**, 855-868 (2008).

135. Wedemeyer, W.J. & Scheraga, H.A. Exact analytical loop closure in proteins using polynomial equations. *Journal of Computational Chemistry* **20**, 819-844 (1999).

136. Canutescu, A.A. & Dunbrack, R.L., Jr. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci* **12**, 963-72 (2003).

137. Shehu, A., Clementi, C. & Kavraki, L.E. Modeling protein conformational ensembles: from missing loops to equilibrium fluctuations. *Proteins* **65**, 164--179 (2006).

138. Coutsias, E.A., Seok, C., Wester, M.J. & Dill, K.A. Resultants and Loop Closure. *Int. J. Quant. Chem.* **106**, 176-189 (2006).

139. Wang, C., Bradley, P. & Baker, D. Protein-protein docking with backbone flexibility. *J Mol Biol* **373**, 503-519 (2007).

140. Fiser, A., Do, R.K. & Sali, A. Modeling of loops in protein structures. *Protein Sci* **9**, 1753--1773 (2000).

141. Rohl, C.A., Strauss, C.E.M., Chivian, D. & Baker, D. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55**, 656--677 (2004).

142. Zhu, K., Pincus, D.L., Zhao, S. & Friesner, R.A. Long loop prediction using the protein local optimization program. *Proteins* **65**, 438--452 (2006).

143. Jacobson, M.P. et al. A hierarchical approach to all-atom protein loop prediction. *Proteins* **55**, 351--367 (2004).

144. Jacobson, M.P. Loop decoy sets. (2008).

145. Dunbrack, R.L.J. & Cohen, F.E. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* **6**, 1661--1681 (1997).

146. Press, W., Teukolsky, S. & Vetterling, W. *Numerical Recipes: The Art of Scientific Computing, Third Edition*, (Cambridge University Press, Cambridge, 2007).

147. Raman, S. et al. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **77 Suppl 9**, 89-99 (2009).

148. Hook, D.G. & McAree, P.R. Using Sturm Sequences to Bracket Real Roots of Polynomial Equations in *Graphics gems* (Academic Press, New York, 1990).

149. Delano, W.L. The PyMOL Molecular Graphics System. (DeLano Scientific LLC, San Carlos, CA, USA).

150. Tsodikov, O.V., Record, M.T. & Sergeev, Y.V. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *Journal of Computational Chemistry* **23**, 600-609 (2002).

151. Richards, F.M. Areas, volumes, packing and protein structure. *Annu Rev Biophys Bioeng* **6**, 151-76 (1977).

152. Lazaridis, T. & Karplus, M. Effective energy function for proteins in solution. *Proteins* **35**, 133-52 (1999).

153. Laskowski, R.A., Moss, D.S. & Thornton, J.M. Main-chain bond lengths and bond angles in protein structures. *J Mol Biol* **231**, 1049-67 (1993).

154. Banaszynski, L.A., Liu, C.W. & Wandless, T.J. Characterization of the FKBP.rapamycin.FRB ternary complex. *J Am Chem Soc* **127**, 4715-21 (2005).

155. Peyroche, A. et al. Brefeldin A acts to stabilize an abortive ARF-GDP-Sec7 domain protein complex: involvement of specific residues of the Sec7 domain. *Mol Cell* **3**, 275-85 (1999).

156. Mossessova, E., Corpina, R.A. & Goldberg, J. Crystal structure of ARF1*Sec7 complexed with Brefeldin A and its implications for the guanine nucleotide exchange mechanism. *Mol Cell* **12**, 1403-11 (2003).

157. Wurtele, M., Jelich-Ottmann, C., Wittinghofer, A. & Oecking, C. Structural view of a fungal toxin acting on a 14-3-3 regulatory complex. *EMBO J* **22**, 987-94 (2003).

158. Jahn, T. et al. The 14-3-3 protein interacts directly with the C-terminal region of the plant plasma membrane H(+)-ATPase. *Plant Cell* **9**, 1805-14 (1997).

159. Spencer, D.M., Wandless, T.J., Schreiber, S.L. & Crabtree, G.R. Controlling signal transduction with synthetic ligands. *Science* **262**, 1019-24 (1993).

160. Belshaw, P.J., Ho, S.N., Crabtree, G.R. & Schreiber, S.L. Controlling protein association and subcellular localization with a synthetic ligand that induces heterodimerization of proteins. *Proc Natl Acad Sci U S A* **93**, 4604-7 (1996).

161. Farrar, M.A., Alberol, I. & Perlmutter, R.M. Activation of the Raf-1 kinase cascade by coumermycin-induced dimerization. *Nature* **383**, 178-81 (1996).

162. Amara, J.F. et al. A versatile synthetic dimerizer for the regulation of protein-protein interactions. *Proc Natl Acad Sci U S A* **94**, 10618-23 (1997).

163. Whitney, M.L., Otto, K.G., Blau, C.A., Reinecke, H. & Murry, C.E. Control of myoblast proliferation with a synthetic ligand. *J Biol Chem* **276**, 41191-6 (2001).

164. Rollins, C.T. et al. A ligand-reversible dimerization system for controlling protein-protein interactions. *Proc Natl Acad Sci U S A* **97**, 7096-101 (2000).

165. Plummer, K.A., Carothers, J.M., Yoshimura, M., Szostak, J.W. & Verdine, G.L. In vitro selection of RNA aptamers against a composite small molecule-protein surface. *Nucleic Acids Res* **33**, 5602-10 (2005).

166. Wells, J.A. & McClendon, C.L. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **450**, 1001-9 (2007).

167. Hellinga, H.W., Caradonna, J.P. & Richards, F.M. Construction of new ligand binding sites in proteins of known structure. II. Grafting of a buried transition metal binding site into Escherichia coli thioredoxin. *J Mol Biol* **222**, 787-803 (1991).

168. Domingues, H., Cregut, D., Sebald, W., Oschkinat, H. & Serrano, L. Rational design of a GCN4-derived mimetic of interleukin-4. *Nat Struct Biol* **6**, 652-6 (1999).

169. Liu, S. et al. Nonnatural protein-protein interaction-pair design by key residues grafting. *Proc Natl Acad Sci U S A* **104**, 5330-5 (2007).

170. Potapov, V. et al. Computational redesign of a protein-protein interface for high affinity and binding specificity using modular architecture and naturally occurring template fragments. *J Mol Biol* **384**, 109-19 (2008).

171. Kirby, J. & Keasling, J.D. Biosynthesis of plant isoprenoids: perspectives for microbial engineering. *Annu Rev Plant Biol* **60**, 335-55 (2009).

172. Zanghellini, A. et al. New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci* **15**, 2785-94 (2006).

173. Davis, I.W. & Baker, D. RosettaLigand docking with full ligand and receptor flexibility. *J Mol Biol* **385**, 381-92 (2009).

174. Treynor, T.P., Vizcarra, C.L., Nedelcu, D. & Mayo, S.L. Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proc Natl Acad Sci U S A* **104**, 48-53 (2007).

175. Ambroggio, X.I. & Kuhlman, B. Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J Am Chem Soc* **128**, 1154-61 (2006).

176. Ambroggio, X.I. & Kuhlman, B. Design of protein conformational switches. *Curr Opin Struct Biol* **16**, 525-30 (2006).

177. Meyerguz, L., Kleinberg, J. & Elber, R. The network of sequence flow between protein structures. *Proc Natl Acad Sci U S A* **104**, 11627-32 (2007).

178. Weinreich, D.M., Delaney, N.F., Depristo, M.A. & Hartl, D.L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111-4 (2006).

179. Yoshikuni, Y. & Keasling, J.D. Pathway engineering by designed divergent evolution. *Curr Opin Chem Biol* **11**, 233-9 (2007).

Appendix

Descriptions of Rosetta command line options

Descriptions for all command line options appearing in this dissertation are shown. Parentheticals note options that have been deprecated or activated by default in the current release of Rosetta, version 3.1.

Revision 24219 options	Description
loops:kinematic	Activates KIC loop modeling (deprecated in favor of loops:remodel perturb_kic and loops:refine refine_kic)
loops:nonpivot_torsion_sampling	Activates sampling of non-pivot torsions from Ramachandran space (active by default)

Revision 27114 options	Description
loops:remodel perturb_alc	Activates centroid stage of KIC loop modeling protocol (deprecated in favor of loops:remodel perturb_kic)
loops:refine refine_alc	Activates full atom stage of KIC loop modeling protocol (deprecated in favor of loops:refine refine_kic)
loops:strict_loops	Prevents random growing of loops (active by default in KIC loop modeling)

Revision 35441 options	Description
match:lig_name	Name of target used for matching
match:grid_boundary	Location of file with Cartesian bounding box defining scaffold regions available to matcher
match:scaffold_active_site_residues	Location of file listing scaffold positions available for matching
match:geometric_constraint_file	Location of file describing the geometry of the target relative to the motif residues
match:euclid_bin_size	The bin width for the 3-dimensional coordinate hasher, in angstroms
match:euler_bin_size	The bin width for the euler angle hasher, in degrees
match:bump_tolerance	The permitted level of spherical overlap between any two atoms, in angstroms
match:output_format PDB	Tells the matcher to output PDB files
match:consolidate_matches	Instead of outputting all matches, group matches by matched motif sequence and similarity of target placement, and then record only the top match:output_matches_per_group from each group
match:output_matches_per_group	Number of matches to output from each consolidated group
match:output_matchres_only	Only output the matched residues and target, rather than the whole pose, for every match (active by default)

Revision 36129 options	Description
s	Starting structure for design
enzdes:detect_design_interface	Automatically detect design/repack region around target based on distance cutoffs
enzdes:cut1	Design any residue with a Cα within this distance of a target heavy atom
enzdes:cut2	Design any residue with a Cα within this distance of a target heavy atom and a Cβ closer to that target atom
enzdes:cut3	Repack any residue with a Cα within this distance of a target heavy atom
enzdes:cut4	Repack any residue with a Cα within this distance of a target heavy atom and a Cβ closer to that target atom
enzdes:cst_opt	Minimize motif–target interactions before design. All designable non-motif residues are mutated to alanine and a reduced energy function that does not contain vdW-attractive or solvation terms is used for minimization
enzdes:bb_min	Allows the backbone to be slightly flexible during minimization
enzdes:chi_min	Allows the dihedrals of the motif residues to move during minimization
enzdes:cst_design	Activates the iterative minimization / design protocol
enzdes:design_min_cycles	Number of minimization / design iterations
enzdes:start_from_random_rb_conf	Start with a random target conformation if a multi-model PDB file is supplied
score:hbond_His_Phil_fix	Alters the hydrogen bond angular dependence for histidines
score:no_his_his_pairE	Sets the pair term for histidine–histidine to zero
loops:loop_file	Path to loop definition file
loops:kic_max_seglen	Maximum number of residues in KIC move segments (12 by default)
loops:outer_cycles	Number of outer cycles in KIC Monte Carlo protocols
loops:refine_init_temp	Initial temperature for the KIC full-atom Monte Carlo protocol

General options shared across revisions	Description
database	Location of Rosetta database
in:file:fullatom	Enables full atom input of PDB or centroid structures (including side-chain conformations)
resfile	Path to file specifying the residues to design and repack
ex1	Includes extra chi1 sub-rotamers (+/- one standard deviation, 3 samples)
ex2	Includes extra chi2 sub-rotamers (+/- one standard deviation, 3 samples)
ex3	Includes extra chi3 sub-rotamers (+/- one standard deviation, 3 samples)
ex4	Includes extra chi4 sub-rotamers (+/- one standard deviation, 3 samples)
ex1aro	Includes extra chi1 sub-rotamers (+/- one standard deviation, 3 samples) for aromatic residues (implied by ex1)
extrachi_cutoff	Number of neighbors a residue must have before extra rotamers are used
in:file:extra_res_fa	Specifies path to full atom parameter file for non-protein molecules
in:file:extra_res_cen	Specifies path to centroid parameter file for non-protein molecules
out:file:fullatom	Enables full atom output of PDB or centroid structures (active by default in KIC loop modeling)
use_input_sc	Use rotamers from input structure in packing
nstruct	Number of models to produce

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Daniel J. Mandell — June 7, 2010

Daniel J. Mandell — Date