Title
Correlated Motions and Allostery

Author
McClendon, Christopher

Publication Date
2011

Peer reviewed|Thesis/dissertation

Correlated Motions and Allostery

by

Christopher Lee McClendon

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Acknowledgements

I would like to first thank God for giving me the ability and the opportunity to do science, for publishable results, and for bringing people, ideas, and challenges in my journey to help me grow, and Jesus Christ for wisdom, peace, and joy as I practically try to love my neighbors as myself.

I am very thankful for my amazing parents, Raymond and Patricia McClendon, for their support, wisdom, and encouragement. They invested the big bucks for me to go to college where I wanted (Caltech), and were supportive of my career in science, pushing me to step up to the challenges of grad school rather than taking time off after being burnt out from college.

I am indebted to my advisors Matt Jacobson and Jim Wells, who made an opportunity for me to work with them as a joint student when neither lab had room for a full student. Of paramount importance in my growth as a young graduate student was the enthusiasm of my advisors Jim Wells and Matt Jacobson, and the freedom, support, and guidance they gave me away from infeasible or bad ideas and towards answering fundamental questions of high importance with techniques just beyond my previous skill level. Matt's broad and realistic perspective and timely nudges helped motivate me to focus on the most important work and curtail tangents.

I am grateful for those who have worked with me and have taught me, especially the grad students and postdocs in the Jacobson and Wells Labs. In particular, I would like to thank Jerome Nilmeier for inspiring me to do basic theoretical work, Greg Friedland for co-developing the MutInf program with me, Salma Rafi for finding a good system to apply it to, and Adam Steeves for extending it.

I would also like to thank those who taught me how to do experimental research. Initially Aaron James Link in David Tirrell's lab taught me some basics of wet lab work during a summer undergraduate program, and more recently and in the Wells lab Jack Sadowsky, Deb Datta, Julie Zorn, Dennis Wolan, and Min Zhuang have given of their time to teach me not only techniques but also different ways of thinking about and interpreting experimental results. I gratefully acknowledge funding for this work from the UCSF integrated Program in Quantitative Biology fellowship, the UCSF Cancer Research Coordinating Committee, and from a Pharmaceutical Research and Manufacturing Association Informatics fellowship. Supercomputer time was provided through NSF Teragrid and the Texas Advanced Supercomputing Center.

Acknowledgements related to specific projects:

Mutual Information:

I would like to thank my friend and colleague Jerome Nilmeier for inspiring my computational method development work on allostery, Jim Wells for posing such a grand challenge with the very interesting caspase-1 system, and Matt Jacobson for giving me the freedom to work on this challenging problem, and the patience to let me continue working on it for a long time initially without very informative results. David Mobley suggested studying error bars in analyses of correlated motions, and both Mike Gilson's and Katerina DuBay's use of mutual information inspired me to try it out – I found lower error bars than for Cartesian covariances and cross-correlations, so then moved forward extending the works of Matsuda and Killian, Kravitz, and Gilson to handle smaller

sample sizes where convergence is not guaranteed and so statistical filtering is used instead. I would also like to thank Mike Gilson, John Gross, Michelle Arkin, Hao Li, I. Kuntz, M. Kelly, J. Gross, and R. Abel for helpful discussions. I'd also like to thank M. Hutter for assistance with his Bayesian approach, and Tanja Kortemme for supporting Greg Friendland's work on this project, and Ken Dill for supporting David Mobley and Homeira Amirkhani's work on this project.

Kullback-Leibler Divergence:

This project came out of a comment from Ken Dill during my first thesis committee meeting, so thanks to Ken for inspiring this work and to Steve Presse and Mike Gilson for vetting the math and for helpful advice and comments on the manuscript.

WW domain:

I would like to thank my collaborator and friend Jesús Izaguirre for inviting me to join this project on analyzing WW domain dynamics, where much more simulation data and NMR data was available, and for helpful advice over the past few years. I would also like to thank Jeff Peng for providing NMR expertise and Faruck Morcos for being very hardworking and on-top-of-things during the long review process.

PDK1:

Jim Wells was the first to suggest that it would be possible to use Jack Sadowsky's disulfide trapping ("Tethering") data on PDK1 cysteine mutants and PIFtide with cysteine at different locations to develop a model of the PIFtide-PDK1 complex. I

would like to thank Matt Jacobson for support and useful suggestions in the practical details of actually doing this.

Caspase-1:

Jack Sadowsky taught me peptide synthesis, hydrogen-deuterium exchange, and in general how to debug experimental protocols with the help of judiciously chosen controls, and Julie Zorn, Dennis Wolan, and Deb Datta taught me soluble and insoluble protein expression and purification in bacteria, as well as how to set up some enzyme assays. I am very thankful to Deb Datta for providing protocols for caspase-1 and reagents for caspase-1, including mutant inclusion bodies, which I could then refold with other subunits to make additional mutant caspase-1 constructs. I would also like to thank Catherine Shi, who worked with me on studying caspase-1 mutants, work that will be prepared and submitted in the near future. With this training and my own practice in protocol development and in computational and analytical models to fit the data, I was able to become an enzymologist. I would also like to thank Scott Hansen and Danielle Canzio, whose assistance in performing the analytical centrifugation experiments and analyzing the data was critical in helping me directly show caspase-1's active site inhibitor-assisted dimerization. I also wish to thank colleagues in the Wells lab, Charles Biddle-Snead, William deGrado, Susan Miller, and Luhua Lai and Xiaodong Su and Peking University and labmates in the Center for Theoretical Biology at Peking University for useful interactions.

This thesis is in part derived from work published along the way:

Chapter 2 describes work published in Journal of Chemical Theory and Computation in 2009, entitled "Quantifying correlations between allosteric sites in thermodynamic ensembles", with co-authors Greg Friedland, Homeira Amirkhani, David Mobley, and Matt Jacobson.

Chapter 3 describes as-yet-unpublished work co-authored with Matt Jacobson and inspired by Ken Dill, and features simulation data mostly from published works by Jacobson group members (and former group members) Gabriela Barreiro and Lan Hua and simulation data described in Chapter 2 from Homeira Amirkhani and David Mobley.

Chapter 4 describes a self-contained part of my contribution to an article published in *PLoS Computational Biology* in Dec. 2010, "Modeling Conformational Ensembles of Slow Functional Motions in Pin1-WW" with co-authors Faruck Morcos, Santanu Chatterjee, Paul R. Brenner, Roberto López-Rendón, John Zintsmaster, Maria Ercsey-Ravasz, Christopher R. Sweet, Matt Jacobson, Jeff Peng, and Jesús A. Izaguirre.

Chapter 5 describes my contribution to work published in *PNAS* in April 12, 2011, entitled "Turning a protein kinase on or off from a single allosteric site via disulfide trapping", with co-authors Jack Sadowsky, Dennis Wolan, Mark Burlingame, and Jim Wells.

Chapter 6 describes experiments and kinetic modeling on caspase-1 that is mostly unpublished, save for some data published in the thesis of Debajyoti Datta.

Chapter 7 includes a review article written by Jim Wells and myself that was published in *Nature* in Dec. 13, 2007.

Abstract

Allostery describes altered protein function at one site due to a perturbation at another site. One mechanism of allostery involves correlated motions, which can occur even in the absence of substantial conformational change. I present two novel information-theoretic molecular dynamics simulation analysis methods, one based on the mutual information and another based on the Kullback-Leibler divergence, to identify statistically significant correlated motions from equilibrium molecular dynamics simulations and statistically significant torsional population shifts when comparing sets of simulations under different conditions. These methods have been implemented in the MutInf software package using Python and inline C code. I next describe applications of these methods, and then novel experiments and kinetic modeling regarding the enzymology of caspase-1. These experiments were performed after MutInf suggested that additional residues near the dimmer interface were important for allostery (a study to be described subsequently). I show that caspase-1's robust cooperativity requires substrate or inhibitor-assisted dimerization, and that even within the dimer, the enzyme is more active when two substrates are bound than when one substrate is bound.

Table of Contents

List of Tables

List of Figures

Chapter 2.

Chapter 3.

Chapter 7.

Compound 1(◊)—Wyeth. **TNFα**: SP403(Δ) —Sunesis-Biogen-Idec. **Survivin**: Compound 1(□), Compound 23b(▲)—Abbott. *Page 162.*

Chapter 1. Introduction

Biological organisms are made of cells, and these cells are composed of DNA, RNA, protein, water, lipids (fats), dissolved salts (ions), and organic molecules. DNA provides genetic information, RNA carries information and catalyzes some chemical reactions, lipids encapsulate the cell from its surrounding and provide boundaries for compartments within the cell, and proteins perform a much wider variety of tasks from catalyzing a plethora of reactions, sensing, carrying, and receiving signals, providing structural integrity, etc. The present work explores single proteins or complexes of proteins transmitting signals from one place on the molecule or complex to another place. Such intra-molecular or intra-molecular-complex phenomena are most generally referred to as "allostery", the topic of the present work. Suppose an enzyme has a particular active site to catalyze a reaction. A potential drug molecule that would work by inhibiting this enzyme could block the active site directly, or it could alter how the enzyme works by binding somewhere else and altering how the enzyme works or interacts with other molecules.

Allostery describes functional cooperativity between sites on a macromolecule or complex. It's Greek roots *allo* ("other") and *stereos* ("shape") are appropriate because allostery is a property of the set of shapes a macromolecule takes on – it's conformational ensemble. Allostery is of paramount importance to biological processes because it enables biomolecules to integrate physical and chemical signals such as binding and post-translational modification and convert them into changes in physiologically relevant outputs such as binding and catalysis. Furthermore, allostery is an important avenue for drug discovery, where allosteric drug binding sites can be used to modulate protein

function in a more specific fashion than competitive inhibition at an active site shared by many similar proteins, and in a more chemically feasible way than using highly charged active site inhibitors, which can present difficult ADME/Tox challenges.

Allostery describes altered protein function at one site due to a perturbation at another site. One mechanism of allostery involves correlated motions, which can occur even in the absence of substantial conformational change. I present a novel method, "MutInf", to identify statistically significant correlated motions from equilibrium molecular dynamics simulations. Our approach analyzes both backbone and sidechain motions using internal coordinates to account for the gear-like twists that can take place even in the absence of the large conformational changes typical of traditional allosteric proteins. I quantify correlated motions using a mutual information metric, which I extend to incorporate data from multiple short simulations and to filter out correlations that are not statistically significant. Applying our approach to uncover mechanisms of cooperative small molecule binding in human interleukin-2, I identify clusters of correlated residues from 50 ns of molecular dynamics simulations. Interestingly, two of the clusters with the strongest correlations highlight known cooperative small-molecule binding sites and show substantial correlations between these sites. These cooperative binding sites on interleukin-2 are correlated not only through the hydrophobic core of the protein but also through a dynamic polar network of hydrogen bonding and electrostatic interactions. Since this approach identifies correlated conformations in an unbiased, statistically robust manner, it should be a useful tool for finding novel or "orphan" allosteric sites in proteins of biological and therapeutic importance.

Next, I present a novel thermodnyamical approach to identify changes in macromoluecular structure and dynamics in response to perturbations such as mutations or ligand binding given molecular dynamics simulations of the unperturbed and perturbed constructs, using an expansion of the Kullback-Leibler Divergence that connects local population shifts in torsion angles to changes in the free energy landscape of the protein. While the Kullback-Leibler Divergence is a known formula from information theory, the novelty and power of our approach lies in its formal developments, connection to thermodynamics, built-in statistical filtering, ease of visualization of results, and extendability by adding higher-order terms. I present a formal derivation of the Kullback-Leibler Divergence expansion and then apply our method at a first-order approximation to three protein systems where ligand binding or pH titration is known from experiments to cause an effect at an allosteric site. Our results on these systems are qualitatively in agreement with experimental approaches measuring local changes in structure or dynamics such as NMR chemical shift perturbations and hydrogen-deuterium exchange mass spectrometry. As our method produces easy-to-analyze results with low backgroud, it has the potential to become a routine analysis when molecular dynamics simulations in two or more conditions are available.

Additionally, I describe a model of the kinase PDK1 bound to an allosteric activatior peptide. This model was constructed by a novel approach using homology modeling with constraints and restraints from disulfide trapping on a panel with PDK1 cysteine mutants reacted with peptide with a cysteine incorporated at various positions.

This work on developing methods to study allostery was inspired by previous experimental studies showing cooperativity in caspase-1 that was robust against a number

of single mutations. Applying our MutInf approach to caspase-1, I identified correlated motions across dimer interface as potentially playing a role in the cooperativity. In order to understand how mutations affected cooperativity, I needed to dissect the contributions of binding at one active site promoting binding at another active site from binding at one site promoting activation through promoting dimerization. Thus, I developed novel chemical kinetic models for caspase-1 and performed biophysical and enzymological experiments to test and restrain this model, and applied simplifications of the model to interpret existing data provided by Dr. Debajyoti Datta (UCSF thesis, 2010).

Targeting allosteric sites and the interfaces between proteins has huge therapeutic potential, but discovering small-molecule drugs that disrupt protein-protein interactions or work allosterically is an enormous challenge. Several success stories in the area of disrupting protein-protein interactions, however, indicate that protein-protein interfaces, and moreover allosteric sites on the surfaces of proteins, might be more tractable than has been thought. These studies discovered small molecules that bind with drug-like potencies to 'hotspots' on the contact surfaces involved in protein-protein interactions. Remarkably, these small molecules bind deeper within the contact surface of the target protein, and bind with much higher efficiencies, than do the contact atoms of the natural protein partner. Some of these small molecules are now making their way through clinical trials, so this high-hanging fruit might not be far out of reach.

Chapter 2. Quantifying correlations between allosteric sites in thermodynamic ensembles

Abstract

Allostery describes altered protein function at one site due to a perturbation at another site. One mechanism of allostery involves correlated motions, which can occur even in the absence of substantial conformational change. We present a novel method, "MutInf", to identify statistically significant correlated motions from equilibrium molecular dynamics simulations. Our approach analyzes both backbone and sidechain motions using internal coordinates to account for the gear-like twists that can take place even in the absence of the large conformational changes typical of traditional allosteric proteins. We quantify correlated motions using a mutual information metric, which we extend to incorporate data from multiple short simulations and to filter out correlations that are not statistically significant. Applying our approach to uncover mechanisms of cooperative small molecule binding in human interleukin-2, we identify clusters of correlated residues from 50 ns of molecular dynamics simulations. Interestingly, two of the clusters with the strongest correlations highlight known cooperative small-molecule binding sites and show substantial correlations between these sites. These cooperative binding sites on interleukin-2 are correlated not only through the hydrophobic core of the protein but also through a dynamic polar network of hydrogen bonding and electrostatic interactions. Since this approach identifies correlated conformations in an unbiased, statistically robust manner, it should be a useful tool for finding novel or "orphan" allosteric sites in proteins of biological and therapeutic importance.

Introduction

Originally, allosteric proteins were those where multiple subunits achieved cooperative binding through ligand-mediated shifts in conformational equilibria, for example in hemoglobin, a protein that carries oxygen in the blood, whose activity is regulated by the concentration of oxygen around it. Nowadays, allostery is broadly defined as any case in which an event at one site on a protein or complex impacts function, dynamics, or distribution of conformations of another site (for recent reviews see [1, 2]). This broader definition includes single-domain proteins as well as proteins or complexes where cooperativity occurs without substantial conformational change. Given this broader definition, it has been suggested that allostery is a property of many proteins[3-5], but is only relevant when a localized event precipitates a change in function. Recently, there has been renewed interest in uncovering allosteric mechanisms of protein regulation and in discovering new allosteric sites, which are of significant interest in biological mechanisms of protein regulation and as novel sites for drug discovery[6-8].

Typically, sites are identified as allosteric after mutational, structural, and thermodynamic characterization with allosteric protein, peptide, or small-molecule modulators, which are frequently found serendipitously[9]. As such, there has been much interest in computational approaches to identify novel allosteric sites. One of the most extensively used approaches has been the Statistical Coupling Analysis approach pioneered by Ranganathan and co-workers[10-13], where pairs of residues that tend to be mutated together in multiple sequence alignments suggest coupling between protein sites. This approach has recently been used to engineer a novel allosteric network by combining predicted allosteric pathways from a light sensor and an enzyme[14]. However,

this approach requires large multiple sequence alignments and the predicted couplings may or may not be relevant to particular proteins in the alignment[15]. Alternative methods to identify allosteric networks using sequence comparisons have also been described[16].

Other computational methods to study allosteric mechanisms and identify potential sites for allosteric regulation focus on a protein's structure and dynamics. Cooper and Dryden showed that the free energy of cooperativity could be separated into two terms: one that accounts for changes in the protein's conformational distribution (i.e. by population shifts), and one that accounts for changes in the amplitudes and/or frequencies of protein vibrational motions. One approach to studying allostery is to focus on protein vibrations[17-20] around a static structure, often by a coarse-grained normal mode analysis[21-24], in which case allosteric effects of perturbations can be calculated analytically. However, these approaches are unable to capture the anharmonicity and multi-well nature of flexible degrees of freedom in proteins. Another approach is to infer groups of residues important for a given allosteric process by analyzing structures trapped in different conformations[25-27].

Dynamical approaches to studying allostery generate an ensemble of structures and then analyze the ensemble using cross-correlations[28], contact correlations[29], principal components[29], or local unfolding correlations[30]. One widely adopted approach uses a quasi-harmonic metric for correlations that assumes an "average" structure[28, 29, 31-33]. This approximation may be appropriate for small backbone fluctuations but may not aptly describe conformational changes that involve basin-hopping, such as loop or side-chain motions. To overcome this quasi-harmonic limitation, Lange and Grubmuller[34, 35] used a mutual information method to account for both quasi-harmonic and anharmonic

correlations in atoms' motion in Cartesian space. Still other methods introduce mechanical perturbations and monitor the subsequent motions of residues[36, 37]. The latter approach can detect substantial population shifts or structural changes following the induced local perturbations, as the added energy facilitates barrier crossing.

Our MutInf approach for identifying allosteric networks quantifies correlations between the conformations of residues in different sites. We use an entropy-based approach to analyze ensembles of protein conformers, such as those from molecular dynamics or Monte Carlo simulations. The method is applicable even in cases where conformational changes are subtle, e.g., when the coupling is mostly entropic in nature[38, 39]. Unlike the approaches described above, our approach uses internal coordinates and focuses on dihedral angles, which are responsible for most low-frequency motions, in order to capture correlated changes in side chain rotamers, a highly anharmonic type of correlated motion. The most closely related previously published method is a study that examined side-chain correlations using a mutual information metric and Monte Carlo simulations of side-chains[40] on a set of fixed protein backbones.

Our MutInf method builds upon and extends previous work by 1) directly connecting correlated conformations to the molecular configurational entropy, 2) incorporating more robust entropy estimators, 3) correcting for undersampling using data from multiple simulations, 4) testing statistical significance to filter out correlated motions that are not significant, and 5) analyzing both backbone and sidechain torsions, which are frequently coupled[41, 42]. The theoretical underpinnings of our approach are described in detail in Methods. Briefly, we use second-order terms from the configurational entropy expansion, the mutual information[43], to identify pairs of residues

with correlated conformations in an equilibrium ensemble. In calculating mutual information, it does not matter whether two residues move at the same time or whether one moves, and then the other; what counts is whether these residues' conformational distributions are correlated. In this work we use the terms correlated motions and correlated conformations interchangeably.

Because we look for correlated conformations in an unbiased, statistically robust manner, we believe that MutInf will be a useful tool in the discovery of novel, "orphan" allosteric sites, where endogenous protein or small molecule allosteric modulators have yet to be discovered. As a proof-of-principle, we used our approach to identify correlations between the conformations of protein residues lining two small-molecule binding sites in human interleukin-2 (IL-2). This single-domain protein exhibits cooperative ligand binding without substantial conformational change, and to date no follow-up work has been done to uncover the mechanism for this cooperativity. We discuss the rationale behind our approach and compare its strengths and weaknesses to those of other methods and then discuss the mathematical details of our method and our novel results on IL-2.

Methods

Theoretical Underpinnings of the Model

When an equilibrium ensemble of states is altered by small perturbations, the fluctuation–dissipation theorem relates equilibrium fluctuations to the system's response, which will be proportional to equilibrium pair-correlations of the degrees of freedom and linear in the applied perturbations. This linear response theory suggests that external

forces, such as those due to ligand binding, cause the largest indirect changes in the degrees of freedom that are most correlated (at equilibrium) with those directly perturbed by the external forces. As has been previously noted[29], this also means that the response to small perturbations involves the same fluctuation pathways activated by random solvent collisions at equilibrium. Elastic network models have identified a correspondence between low-frequency normal modes and pathways used in several protein conformational changes[22, 44], suggesting that correlations observed in equilibrium simulations may propagate perturbations in structure and/or dynamics due to ligand or protein binding.

A perturbation at one site can couple to another site directly, through electrostatic or steric interactions, or indirectly, through solvent reordering, or through a network of residues with correlated conformations. When the conformation at one site depends on the conformation at another site, the sites' conformations are correlated. When the conformations are correlated, perturbations at one site can cause population shifts in conformations at other sites. Correlated conformations are then signals that can be used to identify allosteric mechanisms and predict new sites for allosteric inhibition by proteins or small molecules.

Our MutInf approach uses equilibrium molecular dynamics simulations to identify correlations in residues' conformations, from which functional coupling between sites is inferred. Approaches such as ours that infer allostery from equilibrium simulations assume that the allosteric phenomena of interest (i.e. ligand binding, protein binding, protonation state changes, etc.) make perturbations to the energy landscape that are relatively small, i.e. at most a few kT. For example, proteins and ligand binders with

fast on-rates will satisfy this assumption, while proteins and ligand binders with slow on-rates may not. Furthermore, equilibrium approaches that infer allostery assume that there are no large barriers to conformational changes required for allostery. If such barriers existed, they would prevent pairs of residues from sampling relevant correlated shifts in conformation when perturbations of interest are applied. Along these lines, these equilibrium approaches also assume that there is sufficient sampling along the degrees of freedom relevant to the allosteric phenomena, so that productive or "on-pathway" correlated motions can be observed.

To quantify correlations between residues' conformations from equilibrium simulations, we take advantage of a connection between information theory and thermodynamics. Inspired by the use of mutual information by Killian, Kravitz, and Gilson in calculating configurational entropies from conformational ensembles using internal coordinates[43], we use second-order terms from the configurational entropy expansion, i.e. the mutual information, to identify pairs of residues with correlated conformations. This approach directly and quantitatively connects correlations in conformation to configurational entropy. Using internal coordinates to calculate the mutual information has the twofold advantage of 1) capturing the rotameric, flipping, and gear-like nature of correlated side-chains; and 2) removing potentially spurious correlations that can arise due to structural alignment. The latter effect occurs because minimization of the r.m.s. error in aligning structures in Cartesian space can yield correlated displacements in many atoms' positions as some atoms are fit better than others. An overview of our approach is presented in Scheme 1.

In applying entropy and mutual information to studying allostery, we sought to obtain a measure of the statistical significance of our results and filter out noisy and artifactual correlations. To accomplish this, we extended established methods for calculating entropies and examining correlations via mutual information to handle finite sample sizes, to incorporate data from multiple simulations, to account for the variability between simulations, and to correct for the fact that multiple simulations do not, in practice, represent independent samples of the macromolecular ensemble.

Scheme 1. A schematic of the MutInf approach for identifying correlated residue conformations shows how the observed mutual information is statistically filtered and corrected before being summed over residues pairs. The resulting matrix is then clustered as in a microarray experiment in order to identify groups of residues showing similar patterns of correlations.

Calculation of Mutual Information

The configurational space of a molecule can be described in a standard Cartesian coordinate system or in an internal coordinate system of bond lengths, bond angles, and torsion angles (BAT)[43]. For proteins, key torsion angles include the φ, ψ, and ω torsion angles of the protein backbone and the χ torsion angles of the amino acid side-chains. In the present work, we consider only the φ, ψ, and heavy-atom χ torsion angles (only the first χ angle for proline) and neglect changes in bond lengths, bond angles and omega backbone torsion angles, as we believe that the dynamics of the first three are the most relevant to describing motions of biological importance[43].

Small sample sizes are notoriously challenging for entropy and mutual information-based approaches, so we use robust estimators and correct for bias using simulated data.

Configurational Entropy Expansion and Correlations Between Degrees of Freedom

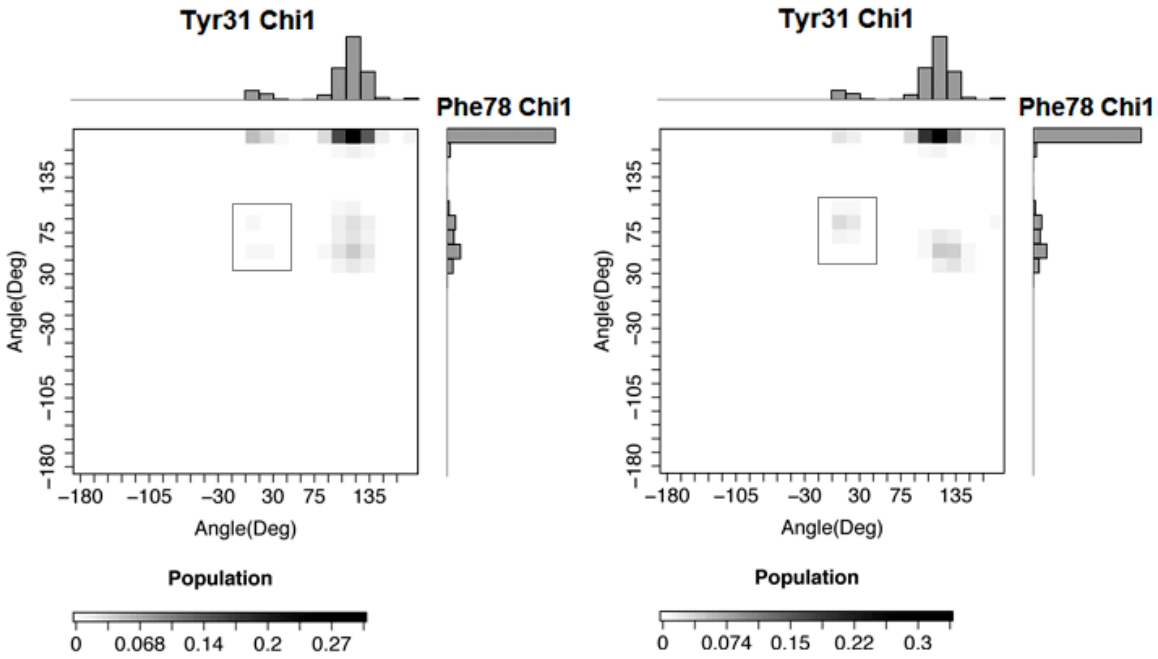
We wish to quantify correlations between residues' torsions. Following the works of Matsuda[45] and of Killian, Kravitz, and Gilson[43], we connect correlated torsions to thermodynamics using an expansion of the molecular configurational entropy into terms over single torsions, pairs of torsions, etc. The total torsional entropy is given by:

$$S_{conf} = \sum_{i}^{n} \int_{0}^{2\pi} p(\phi) \ln p(\phi) d\phi - \sum_{i}^{n} \sum_{j}^{n} \int_{0}^{2\pi} \int_{0}^{2\pi} p(\phi_1, \phi_2) \ln \frac{p(\phi_1, \phi_2)}{p(\phi_1) p(\phi_2)} d\phi_1 d\phi_2 + \ldots - \ldots \qquad (1)$$

where indices *i* and *j* are residues' torsions, and *n* is the number of torsions (φ, ψ, and χ torsion angles in the present work). The second-order term here represents a sum of the mutual information of each pair of torsions. The mutual information describes correlations between degrees of freedom, and gives a measure of how much information about one degree of freedom is gained by knowledge about another[46]. Because the mutual

information values are terms in the entropy, which is related to free energy, the mutual information in Eq. 1 is in units of kT. The mutual information has been a popular, distribution-free analysis method, and more recently has been used in the context of molecular conformational ensembles[40, 43].

As an example, consider the distributions of the χ_1 torsion angles for two side chains. For concreteness, we use an example from interleukin-2, which is described in greater detail below and involves two aromatic residues in close proximity (Figure 1). The expected joint distribution of these torsion angles under the null hypothesis (Figure 1A) of independence is merely the outer product of the marginal distributions. However, the joint distribution from the observed simulations (Figure 1B) show that these torsion angles are correlated (I = 0.203 kT), because a cross-peak (indicated by a gray box) appears in the simulations that would not be expected if these torsions were independent.



Joint Distribution under Null Hypothesis: Independence ($I = 0.00023$ kT)

Observed Joint Distribution: Correlated ($I = 0.203$ kT)

Figure 1. Joint distributions of correlated torsions are different from what would be expected if they were independent. (A) Distributions of two χ_1 torsion angles are shown along with the joint distribution expected if they were independent, i.e. the product of the marginal probabilities. (B) Distributions of the same two χ_1 torsion angles are shown along with the observed joint distribution from molecular dynamics simulations. Grey boxes highlight a cross-peak with substantial height in the observed simulations (B) but with negligible height under the null hypothesis of independence (A).

In practice, we compute the mutual information, I, between two degrees of freedom as the difference between the self-entropies and the joint entropy, using the relation, I = S(1)+S(2)−S(1,2), and a corrected histogram entropy estimate[47] over adaptive partitions[46]:

$$I = \sum_{i=1}^{r} \frac{n_i}{N}\left(\ln N - \Psi(n_i) - \frac{(-1)^{n_i}}{n_i+1}\right) + \sum_{j=1}^{s} \frac{n_j}{N}\left(\ln N - \Psi(n_j) - \frac{(-1)^{n_j}}{n_j+1}\right) - \sum_{j=1}^{rs} \frac{n_{ij}}{N}\left(\ln N - \Psi(n_{ij}) - \frac{(-1)^{n_{ij}}}{n_{ij}+1}\right) \quad (2)$$

where r and s are the number of marginal bins, n_i, n_j, and n_{ij} are the histogram counts, N is the number of data points, and Ψ is the digamma function. Adaptive partitions make

efficient use of discrete bins, preserve correlations between variables, and normalize each joint distribution to a reference distribution in which marginal counts are as uniform as discretization allows[46]. In this work we used 24 bins per dimension. Furthermore, adaptive partitions enable accurate mutual information values to be calculated whether torsional motions are large or small. Note also that we account for the two-fold symmetry in the χ_2 angle of Asp, Phe, and Tyr and in the χ_3 angle of Glu.

The histogram entropy estimator above assumes that histograms are populated by a Poisson process ($n_{ij} << N$) and so is especially appropriate for sparse joint histograms. It also implicitly includes finite bin and data size corrections used in other discrete entropy estimators[48]. As a statistic for examining correlations between variables, the mutual information (with corrections discussed below) is far more robust against small sample sizes than the χ^2 statistic, which assumes $n_{ij} \geq 5$. While the nearest-neighbor approach[49] could have been used instead to compute these integrals, it can require $N \approx 50{,}000$ data points[50] or more to yield a converged estimate of the mutual information for pairs of torsions. Nearest-neighbor approaches are accurate for very large datasets but have biases under finite sample sizes that depend on the topology of conformational space sampled in simulations[50]. As our goal was to extend mutual information calculations to handle smaller sample sizes, we chose instead to use adaptive partitioning in combination with the corrected histogram mutual information estimate above (Eq. 2), so that each pair of degrees of freedom could be compared against the same empirically-generated reference distribution and evaluated for significance.

Correction for Nonzero Mutual Information in Independent Datasets

In a number of applications using mutual information, it has been found that samples of two variables that are independent can yield nonzero mutual information in calculations[46, 51-53]. We empirically observe the same in simulated data (data not shown), and this is not surprising because errors in estimates of the true mutual information are a consequence of finite samples. To correct for this, one approach is to create P permutations of the original data, so that the marginal probability distributions remain the same while correlations between the data are scrambled. One can use these permutations to establish a test of significance of the observed mutual information with a null hypothesis of independence versus a one-sided alternative. The approximate p-value is then the percentage of mutual information values from different permutations that are greater than the observed mutual information from the original data[46, 53]. Also, the average mutual information for the permuted data, the "independent information", can be subtracted from the observed mutual information to yield the "excess mutual information", a more reliable estimate of the true mutual information[51, 52].

When adaptive partitioning is not used (and hence the marginal densities are not normalized), permutation approaches are inefficient in sampling the distribution of the mutual information under the null hypothesis, because permuted values are likely to fall into bins overrepresented in the marginal densities; adaptive partitioning fixes this inefficiency by normalizing marginal densities without altering correlations between variables. One can apply the permutation approach above to nearest-neighbor estimates as well, as these also will have bias due to finite sample sizes. For example, a combined K-fold resampling/permutation test was found to be useful in conjunction with nearest-

neighbor mutual information estimates to discriminate between relevant and independent features in feature selection[53]. A major drawback to the permutation approach is that it is computationally demanding in processing time and in memory. Moreover, permutations introduce random error because not all $N!$ permutations can be made[53].

Instead, since adaptive partitioning is used in this work, we noted that the same distribution of the "independent information" is appropriate for all pairs of degrees of freedom. The distribution of the mutual information for independent variables for given data size N and number of marginal bins r and s has not yet been analytically solved, though in some cases can be empirically fit[52]. However, because an analytical, parametric approach is not available, we perform Monte Carlo sampling to obtain the reference distribution of the "independent information" for all pairs of torsions. With adaptive partitioning, the marginal counts are nearly uniform and in any case are equivalent for different pairs of torsions. Thus, all pairs of torsions will have the same distribution in histogram bin space under the null hypothesis of independence. To construct the reference distribution for a pair of independent torsions, we first make a copy of the marginal distributions for a given pair of torsions (it doesn't matter which pair we choose). Then, we choose ordered pairs of bin indices at random from these marginal distributions and place them into a 2-D histogram without replacement. The mutual information is calculated according to Eq. 2 above, and this procedure is repeated 1000 times to create a distribution of the mutual information under the null hypothesis of independence for the given number of datapoints N and number of bins r.

We use this distribution of "independent information" for a significance test of observed mutual information values, and we subtract the average "independent

information" from the observed mutual information to yield the "excess" mutual information; this filters out insignificant mutual information values and corrects for finite sample size bias. Because this analysis empirically generates a distribution under the null hypothesis, the false positive rate for keeping a nonzero mutual information value for torsions that are truly independent is α, the significance level (in our case, 0.01). This false positive rate will be further reduced by consideration of the alternative hypothesis.

Bayesian Filter to Remove False Positives

Most approaches that filter mutual information values using tests of statistical significance do so according to whether the null hypothesis of independence can be rejected using descriptive statistics. One disadvantage of these approaches is that they do not consider the distribution of the mutual information under the alternative hypothesis. In Bayesian statistics, the mutual information is a random variable with a distribution, and the probability that the mutual information is greater than a a given value can be calculated. Approximations to the distribution of the mutual information have been described that account for uncertainties in the estimates of the probability density functions[54]. The first two central moments of the distribution, the expectation E[I] and variance Var[I] of the true information given the data and prior, are given as follows:

$$E[I] = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(\Psi(n_{ij}+1) - \Psi(n_i+1) - \Psi(n_j+1) + \Psi(N+1)) \tag{3}$$

$$Var[I] = \left| \frac{K - J^2}{N+1} + \frac{M + (r-1)(s-1)(\frac{1}{2} - J) - Q}{(n+1)(n+2)} \right| + O(n^{-3}), \tag{4}$$

$$J = \sum_{ij} \frac{n_{ij}}{N} \ln \frac{n_{ij}N}{n_i n_j}, K = \sum_{ij} \frac{n_{ij}}{N} \left(\ln \frac{n_{ij}N}{n_i n_j}\right)^2, M = \sum_{ij} (\frac{1}{n_{ij}} - \frac{1}{n_i} - \frac{1}{n_j} + \frac{1}{N}) n_{ij} \ln \frac{n_{ij}N}{n_i n_j}, Q = 1 - \sum_{ij} \frac{{n_{ij}}^2}{n_i n_j},$$

where n_{ij}=n_{ij}(observed)+n_{ij}(virtual), and the virtual counts come from a noninformative Dirichlet prior (n_{ij}=1 for the uniform prior, which was used in this work). Approximations for the leading order terms for the third and fourth central moments have been reported[54], and could be used in an Edgeworth expansion to approximate the distribution, but for robustness we chose instead to simply use a Gaussian with the above mean and variance, which fit reasonably well to simulated data in a model system[54]. We then use this approximate distribution of the mutual information to calculate $P(I<E[I_{ind}])$, the probability that the true mutual information is below that expected for independent torsions (calculated using Eq. 3 averaged over 1000 simulated independent datasets). Pairs of torsions with $P(I<E[I_{ind}]) > \alpha$ are not significant and are discarded.

Corrections to the Mutual Information Accounting for Incomplete Sampling

To obtain accurate entropies and mutual information values up to second order, simulations must be run many times longer than the slowest autocorrelation and pair correlation times, and data points should represent independent observations. Due to limited computing power, this is rarely the case, and molecules in simulations carry with them some memory of their initial states. For example, consider a salt bridge. Salt bridges can form strong electrostatic interactions, and hence it can take a long time to sample their full conformational space (long autocorrelation time) and even longer to sample all populated pairwise conformations (long pair correlation times). Thus, a salt bridge may retain some memory of its initial conformation, which will fade away on the timescale of the pair decorrelation time (approximately). In practice, we decided to use data from multiple simulations to penalize this kind of undersampling in a novel way.

First, we first aggregate the counts for two degrees of freedom from a set of simulations (sample ensembles) of size n_{sims}, and calculate the mutual information for all the simulations taken together. Intuitively, two torsions in different simulations should not be correlated, as they should sample their probability distributions independently. Any non-zero (excess) mutual information between these torsions is a measure of conformational undersampling bias that we can subtract from the mutual information between the torsions for the set of simulations. To correct the mutual information for artifactual correlations due to incomplete sampling, we calculate the excess mutual information and then subtract the average excess mutual information between two degrees of freedom in different pairs of simulations (when it is positive):

$$I_{i,j}^{sims} = I_{i,j}^{sims} - < I^{ind}{}_{(N,r)} > - \binom{n_{sims}}{2}^{-1} \sum_{k=1}^{n_{sims}} \sum_{l \neq k}^{n_{sims}} I_{i,j}^{k,l} - < I^{ind}{}_{(N',r')} > \quad (5)$$

Here i and j correspond to the different torsions, l and k are the indices of the pairs of different simulations, and I^{ind} is calculated for a pair of simulations just as the independent information is calculated for a set of simulations using the Monte Carlo recipe above, except that values of $< I^{ind} >$ lower than the standard deviation of I^{ind} are zeroed to reduce noise from this term. For the mutual information between torsions in different simulations, we use half as many bins ($r' = r/2$), because the number of datapoints N' for the histograms is smaller than the total number of datapoints from all simulations, N ($N'=N/n_{sims}$). Significant mutual information values are those that have passed the significance test vs. the null hypothesis, the Bayesian filtering using the alternative hypothesis, and whose corrected excess mutual information (eq. 5) is greater than zero.

When we consider the mutual information between pairs of residues, we take the sum of the mutual information between pairs of residues' torsions:

$$I_{residues}(i,j) = \sum_{\substack{k=\phi,\varphi,\chi's \\ \text{residue(i}\neq\text{j)}}} \sum_{\substack{l=\phi,\varphi,\chi's \\ \text{residue(j)}}} I_{k,l}^{sims} \qquad (6)$$

This may overestimate the total mutual information between two residues, as we neglect the higher-order terms in Eq. 1. Inclusion of statistically significant higher order terms (which would require more data points) would further increase the accuracy of the calculated mutual information between pairs of residues. Nonetheless, our results below show that our robust use of second-order terms is a powerful means to identify residues with correlated conformations.

Molecular Dynamics Simulations

Molecular dynamics simulations on interleukin-2 (IL-2), alone and in complex with three ligands, were performed using GROMACS 3.3[55, 56] and the AMBER-99Φ forcefield[57]. Loops missing atomic coordinates, such as residues 1-5, 75-76, and 99-102 in apo IL-2, were closed using loop prediction via the Protein Local Optimization Program (PLOP[58]). Protonation states of histidine side-chains at pH 7 were given by MCCE[59, 60]: we modelled His16 as positively charged (residue name "HIP") and His55 and His79 as ε-protonated. Two ligand-bound forms of IL-2 were prepared, with either Ro26-4550 (amino(3-(2-(1-methoxy-1-oxo-3-(4-(phenylethynyl)phenyl)propan-2-ylamino)-2-oxoethyl)piperidin-1-yl)methaniminium)[61, 62] bound to the competitive IL-2Rα site (PDB:1M48), or compound 7c (1-(3,4-dihydro-1H-pyrido[3,4-b]indol-2(9H)-yl)-2-methoxyethanone) bound to the allosteric site[63]. Compound 7c was built from

PDB:1NBP by modifying the covalent compound 7t in the crystal structure to the noncovalent compound 7c in Maestro (Schrodinger, 2007), then using PLOP loop prediction to optimize the loop from residue 29 to 33, and to fill in missing residues between residues 73 to 78, in each case simultaneously optimizing side-chains within 12 Å of the given loop region. These ligands were parametrized for MD by GAFF[64] and assigned AM1-BCC charges[65, 66]. Each apo protein or complex was energy minimized in explicit solvent using GROMACS, and five copies of each of the three constructs (apo, competitive site bound, and allosteric site bound) were equilibrated at 300K (with different random seeds) using constant volume for 10 ps and using a constant pressure of 1 atm for 100 ps using the Berendsen barostat[67], with hydrogens constrained using the Lincs algorithm[68]. Equilibration of each simulation was followed by a 10 ns production run, with snapshots of the atomic coordinates recorded every 1 ps. Actual lengths of individual simulations ranged between 9.6ns and 11ns for technical reasons. RMSDs of the two compound binding sites over the simulations showed that all of the five simulations per apo or small-molecule bound construct were stable.

Ensemble Docking

We clustered MD snapshots from the IL-2 simulations with allosteric compound bound according to the coordinates of residues in the competitive site using QT clustering[69] as provided in GROMACS ("g_cluster -method gromos"). Then, to each cluster representative we docked the Roche competitive site ligand Ro26-4550[61], which binds cooperatively with the allosteric ligand. This ensemble docking was performed using the XGlide cross-docking script (Schrodinger, 2007, Script Center XGlide v. 1.1.2.6,

mmshare v. 16109, using inner and outer grid box lengths of 10Å and 18Å, respectively)[70].

Results and Discussion

We applied our MutInf approach to elucidate the mechanism of small molecule binding cooperativity in human interleukin-2. Little is known about how binding of ligand at the IL-2Rα binding site enables binding of a small molecule fragment to a cryptic allosteric site. These ligands bind at least 6.9 Å apart at their closest approach in the predicted ternary complex. Crystal structures of complexes of interleukin-2 with small molecules bound to different sites did not show substantial structural changes at the other sites (maximum Cα RMSD 0.88 Å at the allosteric site for apo PDB:1M47 and competitive-bound PDB:1M48). We therefore hypothesized that allostery and cooperativity in IL-2 arises largely from changes in dynamics and subtle population shifts rather than a major change in the preferred backbone conformation[28, 38, 39].

We used molecular dynamics to study correlated motions at the atomic level on a picosecond-nanosecond timescale, and used our MutInf approach to analyze sets of 10 nanosecond trajectories of human interleukin-2, alone and in complex with different small molecule binding partners. Our goal was to show that MutInf can identify significant correlated conformations for functionally-important residues in simulations whose lengths and recording frequencies are typical of those in the current literature.

Mutual Information From Molecular Dynamics Identifies Significant Long-Range Correlations

We first analyzed whether an unbiased, whole-protein analysis of correlations between residues in interleukin-2 would be able to identify cooperative sites and the correlations between them from the apo simulations alone. For each pair of residues, we calculate the mutual information as per (Eq. 6) between all pairs of φ, ψ, and χ torsion angles for our apo simulations of interleukin-2. The mutual information is reported in units of kT, because of the relationship between mutual information and entropy (Eq. 1).

When we plotted the statistically significant mutual information between pairs of residues' torsions in IL-2, we found that only a small subset of residue pairs are highly correlated, while many are only marginally correlated (Figure 2A). A substantial part of the present work involved incorporating more robust entropy estimators for calculating the mutual information and filtering out insignificant correlations with the help of empirical or approximated distributions under the null and alternative hypotheses. So, as a control, we plotted the unfiltered mutual information between residues' torsions in Figure 2B. Protocols with and without statistical filtering showed correlation between residues in the loops after helix 1 and between helices 2 and 3 (Figure 2A and 2B, red boxes on the diagonal and off the diagonal, respectively, and Figure 2C, red residues). Our statistical filtering, however, highlights these and removes background noise; only a subset of the pairwise correlations between residues in these regions make statistically significant contributions to the conformational entropy. Moreover, our MutInf approach identified significant correlations between residues in the loop region between helices 3 and 4 and residues in other regions, in particular residues in the floppy N-terminal tail

(Ser6, Thr7) and the beginning of the N-terminal helix (11-15), residues in the loop between helices 2 and 3, and residues in the C-terminal helix (Figure 2A, blue boxes, and Figure 2C, blue residues). The loop between helices 2 and 3 displays significant variability in the different crystal structures of IL-2, and is at least partially disordered in most structures, indicating both that it is flexible and that it can adopt at least several conformations. Residues showing significant correlations near the C-terminus include two residues in the loop before the C-helix (Glu100 and Thr101), and residues along the C-helix (Arg120, Ile128, and Leu132, proximal to IL-2's negatively-charged C-terminus).

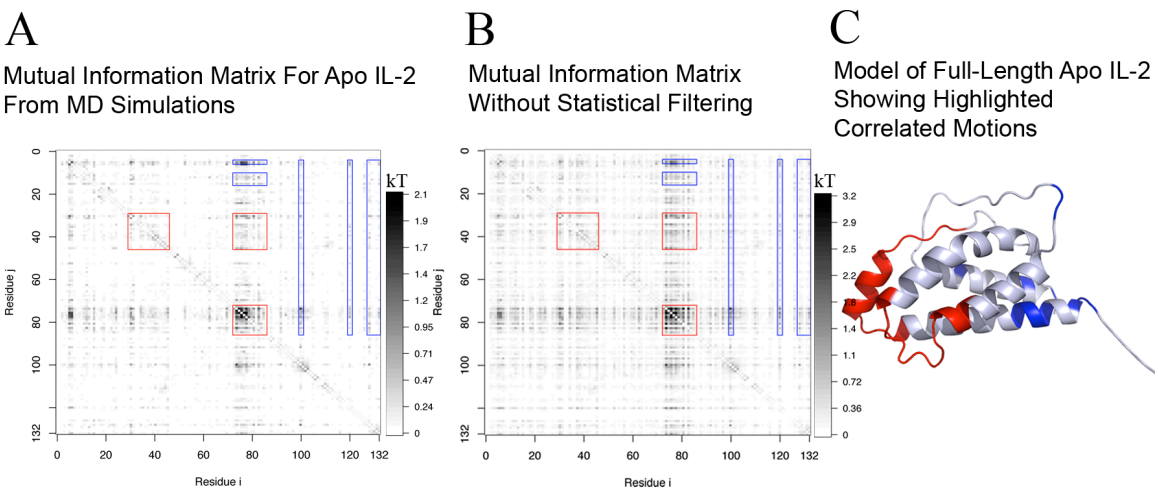


Figure 2. Mutual Information captures significant correlations between residues in human interleukin-2. (A) Mutual information between residues' torsions computed using the present approach, with statistical filtering as detailed in Methods. (B) Same as in A but without any of the aforementioned statistical corrections. (C) The model of full-length human interleukin-2 used in the apo simulations, based on the crystal structure of apo IL-2 (PDB: 1M47). Residues surrounded by red boxes in A are colored red, while residues correlated to these that are surrounded by blue boxes in A are colored blue.

We compared our MutInf method to previously-reported methods for identifying correlated motions, in particular the Gaussian Network Model (GNM) approach of Bahar and colleagues[71] and the Cartesian mutual information method of Lange and

Grubmüller[34]. Both our method and the GNM method suggested correlation between residues in the loop after helix 1 and residues in the loop between helices 2 and 3 (red boxes on the diagonal and off the diagonal, respectively, Figure 2A, 3A). We found that our approach highlighted strong correlations and gave low background noise, while the Cα cross-correlation matrix using a GNM (with 10 Å Cα-Cα cutoff) gave a noisier pattern of correlations, as did Lange and Grubmüller's mutual information method applied to the Cartesian coordinates of Cα atoms (Figure 3B).

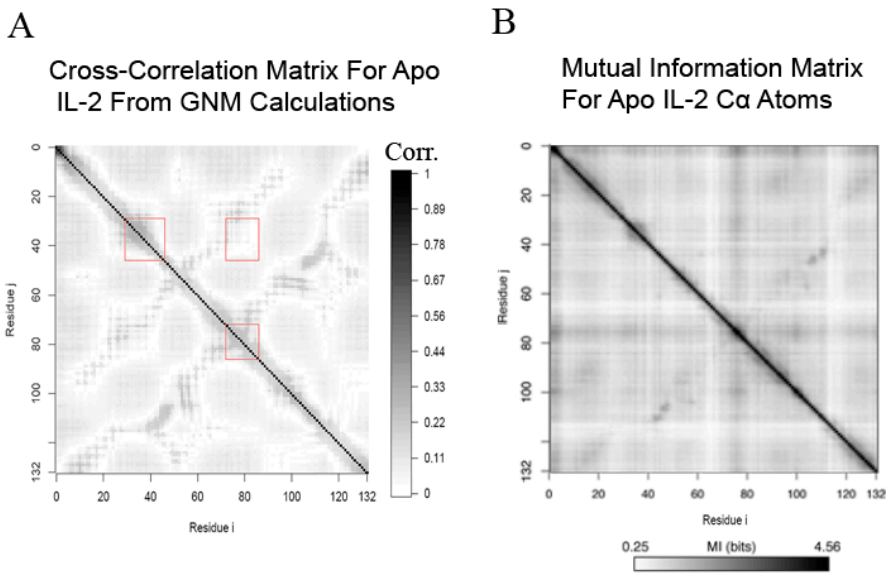


Figure 3. Comparison of pairwise, dynamical correlations between residues computed by alternative methods. (A) Absolute value of the cross-correlation matrix computed using the Gaussian Network Model. (B) Mutual Information between residues' C_α Cartesian coordinates using the approach of Lange and Grubmüller.

Our identification of significant long-range correlations led us to investigate the distribution of correlations between residue pairs as a function of distance between the residues' Cα atoms (Figure 4). As would be expected, the number of weak correlations decreases as the distance between residues increases. However, residues separated by substantial distances have correlations of 0.4 kT or more just as often as residues separated by short distances.

Looking more closely, we see that there are strong couplings between pairs of distant residues (Figure 5). Here, we highlight correlations of magnitude greater than kT between distant residues, namely those with alpha carbon separations of more than 5 Å. Again, we observe strong couplings between the N-terminus of helix A, Tyr31 at the C-terminal end of helix A, and the adaptive loop region (74, 76-78), as well as between Gln74 and Glu100, and between Glu100 and Ile128 near IL-2's C-terminus. It is not surprising that many of these residues are polar, as molecular dynamics simulations include terms for long-ranged electrostatic and ion-dipole interactions. Gaussian Network Models, on the other hand, do not model such sequence-dependent, long-range interactions, and so it is not surprising that the correlations between distant residues are typically weak. Electrostatic interactions can be both directly and indirectly responsible for long-range correlations in residues' conformations; directly, through Coulomb's law, and indirectly, through a dynamic network of charged and hydrogen-bonding polar residues, and through altering the first-shell water structure around the protein (in simulations with explicit solvent). Unlike charge-dipole or dipole-dipole interactions, where the effective range decreases through averaging over orientations[72], charge-charge interactions retain their long-range nature even when averaged over orientations.

Other factors that can give rise to correlated conformations include hydrophobic packing and rigid-body motions of semi-rigid secondary structure elements, such as α-helices. We note that our approach does not typically show strong correlations within semi-rigid elements such as α-helices or the central hydrophobic core of the four-helix bundle. Two possible reasons for this are because mutual information values are not normalized quantities and because higher-order correlations are not captured. As the

maximum mutual information between two residues is the minimum of their self-entropies (flexibilities), residues that have higher self-entropies (more flexible) can exhibit a greater magnitude of coupling with other residues. This behavior is thermodynamically appropriate because the un-normalized mutual information is related to the configurational entropy (Eq. 1), and not a normalized quantity. Furthermore, this behavior is consistent with thermodynamic considerations in intrinsically disordered proteins, where disorder in one or both domains serves to optimize allosteric coupling between the sites[73]. Such allosteric coupling does not require a network of interactions linking the sites. In the present study on interleukin-2, flexible resides at either end of a helix showed couplings > kT in some cases while intervening helical residues did not (Figure 5), because mutual information values, unlike correlation coefficients, are not normalized for residues' self-fluctuations. We note that even normalized pairwise correlations (i.e. Figure 3A) do not include the higher-order correlations that would be expected within semi-rigid elements, and while the Gaussian Normal Mode results (with default cutoffs) show weak couplings between the N-terminus of helix A and Tyr31 at the C-terminal end of helix A, and between the adaptive loop region (74, 76-78) and Glu100, these are not visibly distinct features (Figure 3A). In any case, as the goal of our approach is to identify correlations between the conformations of functional sites on a protein, it is not necessary to identify all of the residues that indirectly mediate such correlations, though this is an area for future work. We will later focus on coupling between two small molecule binding sites in interleukin-2, which are physically connected by flexible loops and sidechains, rather than semi-rigid secondary structure elements.

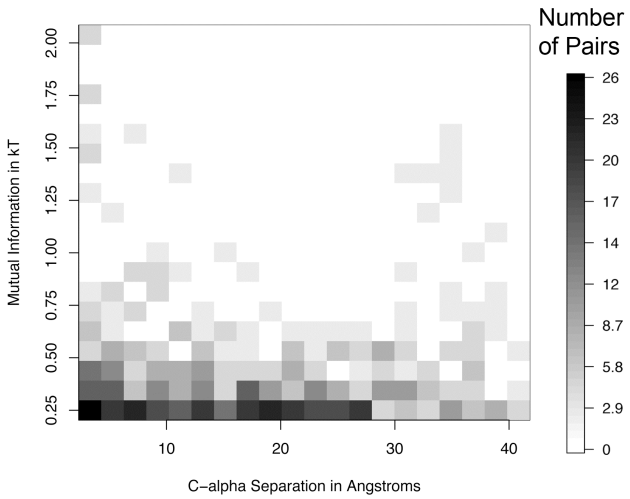


Figure 4. Most of the significant correlations are between distant residues. A 2-D histogram showing the number of pairs of correlated residues vs. Cα separation and mutual information value shows that the number and strength of correlated pairs decreases only modestly with distance. For clarity, only those pairs of correlations with a mutual information greater than 0.25 kT are shown.

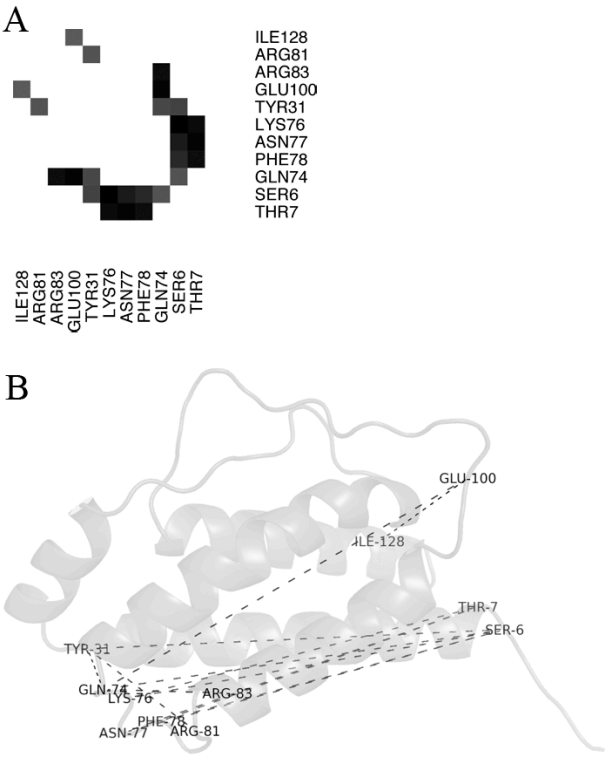


Figure 5. Several distant residues are highly correlated. (A) Correlations greater than kT are shown only for IL-2 residues whose alpha carbons are separated by more than 5 Å. (B) Dashes connecting each pair of these correlated residues show long-rage correlations across the length of the helical bundle.

Hierarchical Clustering Indentifies Dynamic "Hot-Spots" In Interleukin-2

For a site to be suitable for allosteric inhibitor design, it must be both (1) allosteric, causing shape or flexibility changes at other sites[1], and (2) druggable, having the right shape and hydrophobicity for drug-like small molecules[74]. Here, our goal was to predict which sites were most likely to alter structure or dynamics at known functional sites upon perturbation (by ligand binding, for example). To search for such sites, we wanted to identify groups of residues responsible for correlations between functional sites.

In biological networks such as protein-protein interaction networks, functional connections between various proteins are preferentially mediated by "hubs" that interact with a greater than average number of partners[75]. Similarly, functional connections between various protein sites are thought to be mediated by "hub" residues or clusters of residues[25, 27]. We hypothesized that clusters of residues correlated to many other residues, i.e. "dynamical hotspots", could be potential sites or mediators for allosteric modulation of other sites. To find such "dynamical hotspots", we performed a hierarchical clustering of the matrix of mutual information values between residues, in analogy to the analysis of microarray data, using the "heatmap" function in R (http://www.r-project.org/). We used a Euclidean distance metric so that residues showing similar patterns of correlations with other residues are clustered together. Interestingly, one cluster of residues emerged with the strongest correlations within cluster members and the strongest correlations to other residues, and was previously found to be an adaptive region that could bind a number of small-molecule fragments as measured by Tethering experiments[62]. Residues in this

cluster are colored red in Figure 6 and mostly reside in the flexible loop between helices 2 and 3, with two in the N-terminal floppy tail and one in the flexible C-terminal loop. Because the mutual information between two torsions is less than either of their self-entropies, it is not surprising that flexible residues often have high mutual information with other residues. This red cluster constitutes a "dynamic hotspot", as it is highly correlated to other clusters of residues. Furthermore, as this red cluster is correlated to the blue cluster containing the IL-2Rα inhibitor binding site, our method predicts the red cluster to be a candidate region for allosteric modulation of the IL-2Rα site. Two similar clusters can also be seen when mutual information values from subsets of the five simulations are block-averaged (for details on this method please see Chapter 4). However, our approach does not yet predict whether such a site would be druggable by small-molecule allosteric modulators or contain "hotspots" of affinity for protein-protein interactions.

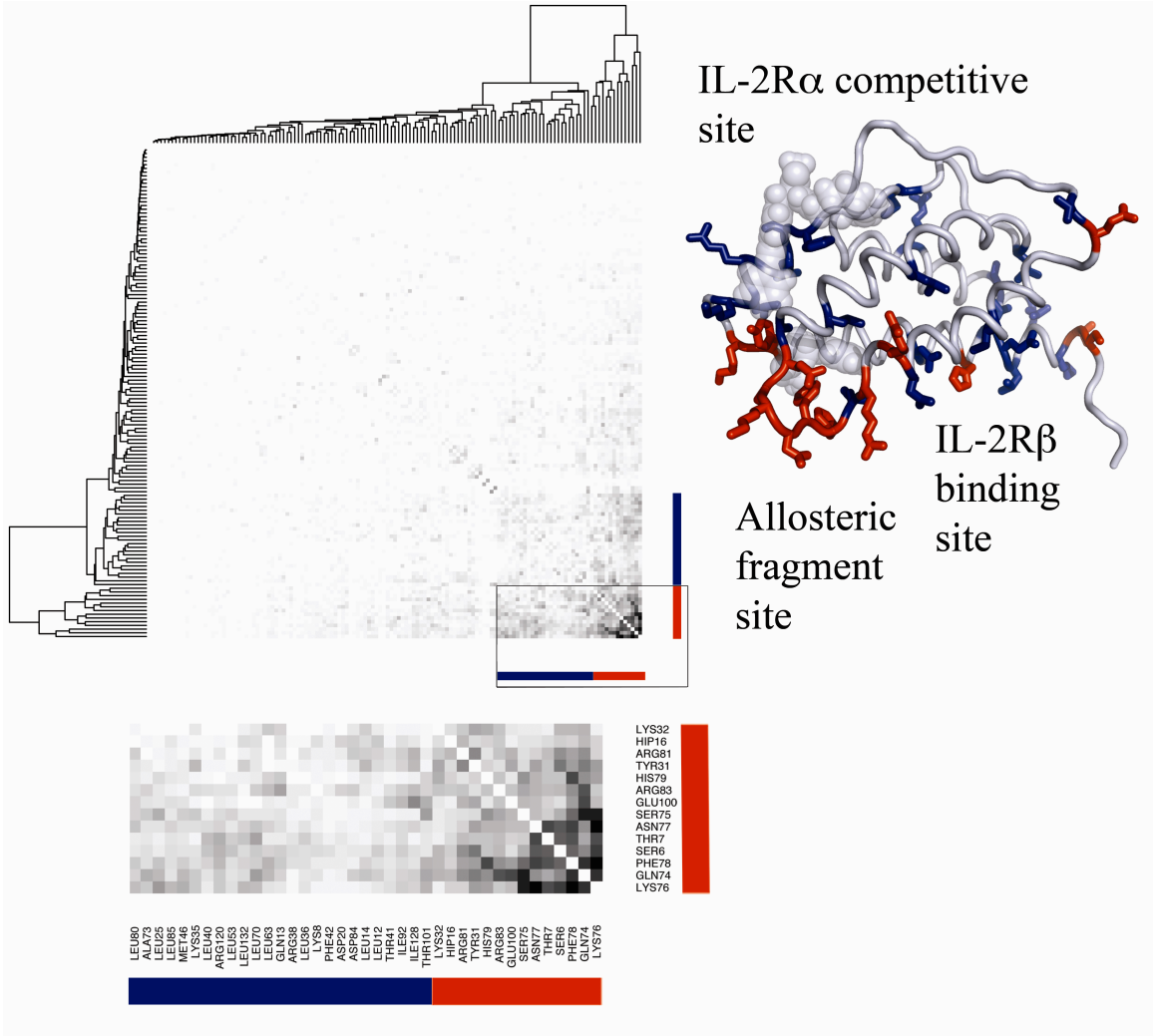


Figure 6. Hierarchical clustering of significant mutual information values identifies allosteric sites. A hierarchically-clustered heatmap shows clusters (top left) of residues with similar patterns of mutual information across IL-2 residues. A close-up view highlights numerous significant mutual information values between pairs of residues in two different clusters, red and blue. These red and blue clusters are highlighted in a model of IL-2's ternary complex (right). The strongest cluster (red sticks) chiefly involves a loop enclosing the allosteric fragment's binding site, and this cluster is correlated to a cluster (blue sticks) containing two protein binding sites, the IL-2Rα-receptor-binding/IL-2Rα-inhibitor-binding site and the IL-2Rβ-binding site. The two compound binding sites and the two protein-binding sites are directly correlated through the hydrophobic core (in the blue cluster), through a highly flexible loop (in the red cluster), and crosstalk between these elements, seen in a close-up view of the matrix (bottom).

Chemical Shift Perturbations Upon Binding Corroborate Predicted Correlated Motions

While direct experimental methods to identify correlated motions by NMR are limited, one can use chemical shift perturbations and changes in side-chain order parameters of residues outside a binding pocket to identify population shifts in residues within or proximal to allosteric sites that accompany ligand binding. A study by scientists at Roche identified IL-2 residues showing backbone and side-chain amide chemical shift perturbations upon binding IL-2Rα receptor or IL-2Rα-competitive small molecule[61]. For example, Tyr31, Gln74 and Ser75 (in the strong cluster colored "red" in Figure 6) show strong $^{15}N/^{1}H$ chemical shift perturbations following competitive ligand binding, though these residues are not in the competitive binding site. While ring current effects from the ligand's biphenyl group could contribute to these shift perturbations, they are qualitatively consistent with our prediction that these residues' conformations are correlated to the conformation of the IL-2Rα binding site. Furthermore, several residues or regions distal from the IL-2/IL-2Rα interface identified by our approach as highly correlated to the "blue" cluster (encompassing many residues in the IL-2/IL-2Rα interface, PDB: 1Z92[76]) showed substantial chemical shift perturbations upon IL-2Rα binding (Figure 7). Unfortunately, resonance overlap restricted the analysis of chemical shift perturbations, notably in most of the flexible loop in the red cluster, so we cannot test our prediction that one would see many perturbations in the flexible loop. It is important to note that none of the nine residues showing insignificant chemical shift perturbations (Asn26, Thr37, Met104, Cys105, Tyr107, Thr113, Ile122)[61] appeared in the red or blue highly correlated clusters.

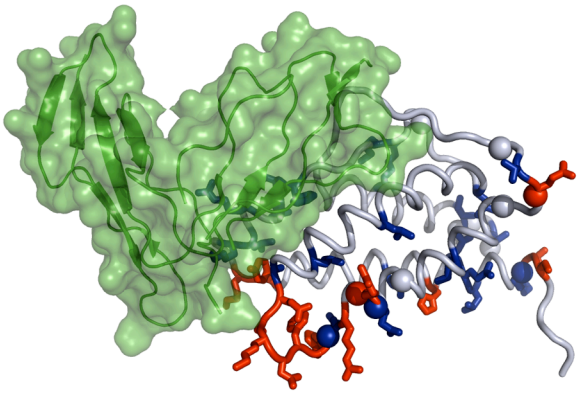


Figure 7. Predicted couplings are consistent with regions perturbed upon IL-2Rα binding. Regions distant from the IL-2Rα receptor binding site that show substantial backbone chemical shift perturbations upon IL-2Rα binding[61] roughly correspond to regions with residues whose conformations are correlated with residues in the IL-2Rα binding site (predominantly residues in the "blue" cluster in Figure 6). Amides on IL-2 whose resonances shifted by more than three linewidths upon IL-2Rα binding or fell below 7% of the original intensity are shown as spheres. Residues from the "red and "blue" clusters shown in Figure 6 are colored accordingly. IL-2 is as shown in cartoon and sticks as in Figure 6, while IL-2Rα is shown in green.

Communication Between Cooperative Compound Binding Sites Involves a Polar, Solvent-Exposed Network and a Greasy Core

. In the previous sections we used a global description of pairwise couplings to identify putative allosteric sites from correlations between clusters of residues. Presently, we apply our method to study the mechanism of coupling between two given allosteric sites. From our matrix of pairwise correlations between residues in apo-IL2 and representative structures from the conformational ensemble, we could infer a structural mechanism by which the two small molecule binding sites might be coupled. In Figure 8 we show matrix elements corresponding to residues in the IL-2Rα-competitive site and residues in the allosteric site, along with representative conformations of these residues. These representative conformers were picked by clustering the MD snapshots according

to the RMSD of residues in the "red" cluster that belonged to the highly flexible loop or were proximal to the N-terminus.

Thermodynamically, the two sites are coupled directly by the off-diagonal gray matrix elements in Figure 8 in the box denoting "Correlations Between Sites". These two sites may also be coupled by higher-order terms involving other residues, which our pairwise analysis does not address (save for the hierarchical clustering which uses patterns of correlation rather than the correlations themselves). From the representative conformations for these residues in Figure 8, the two sites appear to be coupled via a polar network on the protein surface and a greasy core. Two residues are common to both binding sites, namely Lys35 and Met39. The side-chain of Met39, for example, can directly interact with both of the cooperative small molecules, and will be discussed in more detail later. A number of polar side-chains pointing toward the solvent form a network of hydrogen-bonding and electrostatic interactions. In particular, Gln74 (dark green lines) samples a wide swath of conformations, sometimes hydrogen bonding with basic residues in the competitive site (Lys35, dark blue, and Arg38, light blue), and other times hydrogen bonding with basic Arg81 (gray) near the allosteric site. Also, correlations between residues in the greasy core connect hydrophobic surfaces of both compound binding sites; when bound, the allosteric or competitive small molecule would be contiguous with this hydrophobic network. Notably, the matrix elements showing "Correlations Between Sites" indicate that the conformations of residues in the polar network and greasy core are coupled. Thus, a more accurate mechanism for the coupling would be that the two sites are coupled via a polar network, a greasy core, and crosstalk between these elements.

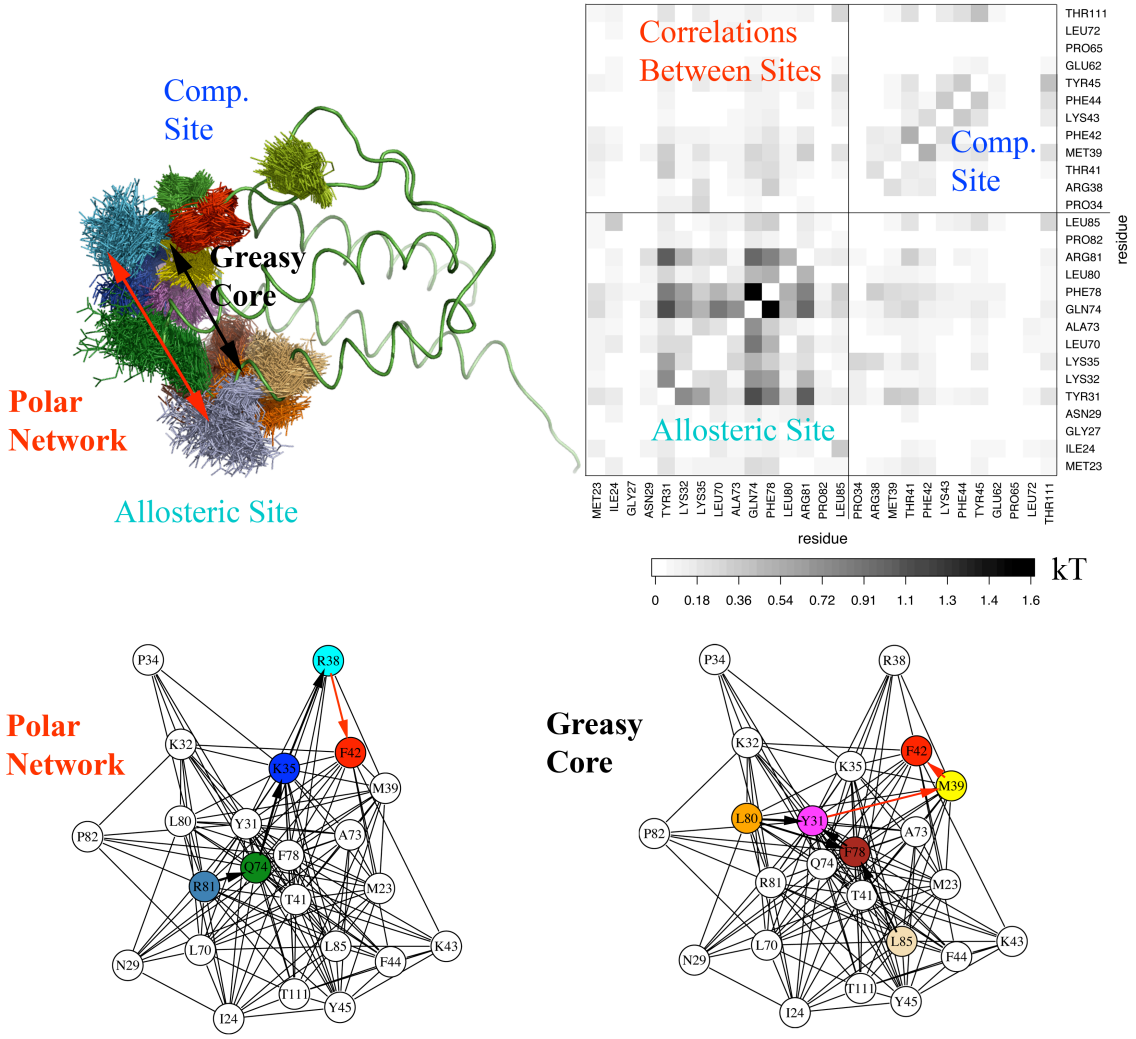


Figure 8. Direct, pairwise correlations couple residues in the IL-2Rα-competitive site (at the IL-2:IL-2Rα interface) to residues in the allosteric fragment-binding site (near the IL-2:IL-2Rβ interface). (Top left) Apo IL-2 is shown in green ribbon while representative conformations of residues showing strong correlations within and between these sites are shown with lines. Overlap between clouds of residues' conformations suggest steric coupling, particularly in the greasy core, from Leu80 (orange) and Ler85 (tan) to Phe78 (brown), to Tyr31 (magenta), to Met39 (yellow), and to Phe42 (red). (Top right) A subset of the full matrix of pairwise correlations reveals direct correlations between residues in the two sites, with the labeled boxes showing correlations within the allosteric site, within the competitive site, and between these two sites. (Bottom) A force-directed network diagram[77] for residues in these sites filtered for correlations of at least 0.05kT shows Phe78, Tyr31, Gln74, and Arg81 as central "hub" residues mediating correlations between the sites.

Ligand Binding At Allosteric Site Causes Rotamer Population Shifts That Promote Competitive Site Inhibitor Binding

The preceding analysis suggests that the two sites are connected via correlated motions, and that this could explain the observed allostery. To directly test our hypothesis that the experimentally observed cooperativity between the sites involves subtle population shifts in residues exhibiting correlated motions, we performed additional simulations with a competitive or an allosteric inhibitor bound, and asked whether simulations with the allosteric inhibitor bound would cause population shifts in the competitive site similar to those observed in simulations with the competitive inhibitor. Comparing the crystal structures of apo IL-2 (PDB:1M47) to competitive-site-inhibited IL-2 (PDB:1M48), we note that the motion of two side chains, Met39 and Phe42, opened up a binding groove for the ligand that was closed in the apo structure.

We then asked whether population shifts caused by allosteric ligand binding would help open up the competitive site for competitive ligand binding. To address this question, we examined side-chain torsion angle distributions for Phe42 and Met39 in simulations with compound at either site and compared these to distributions from the apo simulations (using histogram bins of 6 degrees and 10 ps time intervals). The populations of side chain dihedrals angles in Phe42 and Met39 show substantial differences between the apo protein and the protein bound with competitive and allosteric inhibitors (Figure 9, Figure 10). Interestingly, the populations observed for the allosteric

compound-bound protein are intermediate between apo IL-2 and competitive-site-bound IL-2. Phe42, a hot-spot residue critical for ligand binding and protein binding, adopts a different χ_1 rotamer for ligand binding than it does for protein binding, which is more similar to the apo rotamer (Figure 9). The χ_1 rotamer selected by ligand in competitive-site-bound simulations (100% population) was more populated (89%) in allosteric-bound simulations than in apo simulations (39%), showing a population shift caused by allosteric compound binding.

We predict that Met39 is an important mediator of binding cooperativity because its conformation is correlated to that of Phe42 and because it shows χ_1 and χ_2 population shifts upon allosteric compound binding in the same direction as population shifts from the apo to competitive inhibitor-bound distributions. In crystal structures, Met39 adopts similar conformations in complexes with competitive inhibitor or allosteric inhibitor that both differ from the conformation in the apo structure. Met39 is in an "up" conformation in the apo structure, packed against hot-spot residue Phe42. In competitive inhibitor-bound structures, the side-chain of this Met moves down to slightly enlarge the pocket for a ligand aromatic ring, while in the allosteric-bound structure, the Met sidechain moves down to interact weakly with the ligand and fill in part of the hydrophobic pocket opened to accommodate the ligand (Figure 10). Interestingly, this Met39 is not critical for a high-affinity competitive ligand to bind at the competitive site[78], presumably because mutating it to alanine would simply make that hydrophobic pocket a little larger. However, it is currently unknown whether this residue is required for allosteric ligand binding or for the binding cooperativity, as we predict. Our calculations indicate that cross-talk contributing to cooperativity involves not only the greasy core (of which Met39 and Phe42 are a part),

but also the loop between helices 3 and 4 and a polar network involving a number of basic residues on the protein surface. This hypotheses could be further tested by conservative mutations of residues such as Gln74, which is not part of either ligand's binding site; Lys35, whose alkyl tail but not polar head contacts compound 1 in the crystal structure; or Met39, whose alanine mutation only shows a slight effect on competitive ligand binding[79]

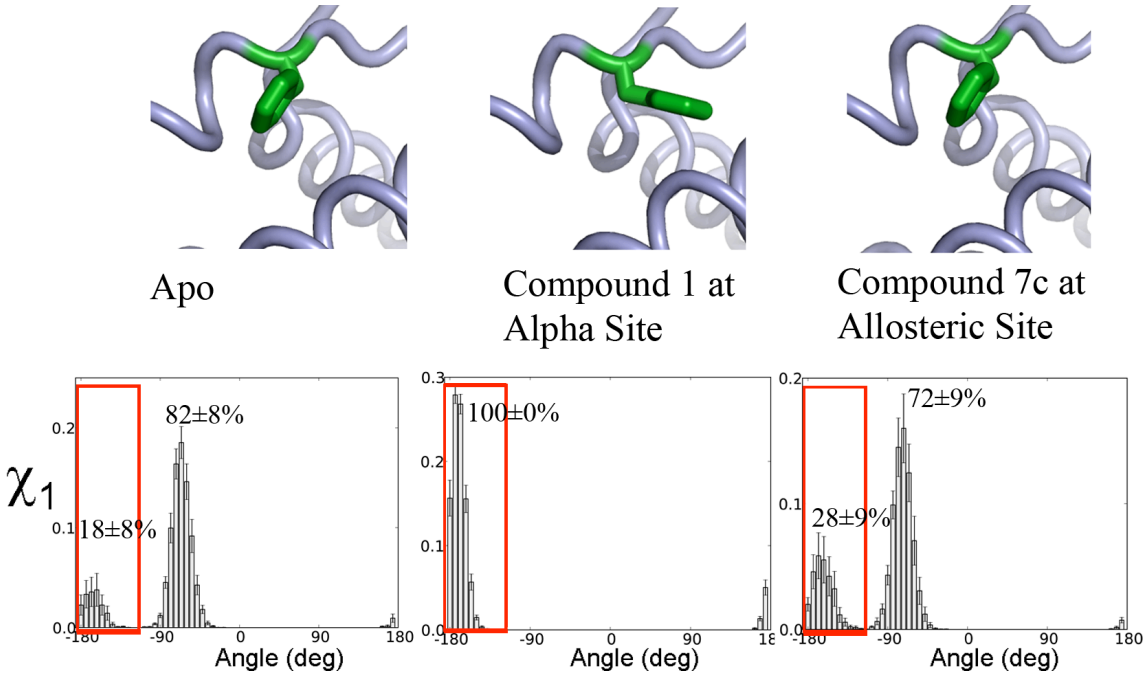


Figure 9. Compound binding to the allosteric site causes a population shift in the conformation of hot-spot residue Phe42 that favors binding compound at the IL-2Rα-competitive site. (Top) Conformations of Phe42 in apo and compound-bound crystal structures (PDB IDs 1M47, 1M48, and 1NBP, resp.). (Bottom) Histograms of Phe42's χ_1 angle from MD simulations. Red boxes highlight the χ_1 population selected by ligand binding at the competitive site.

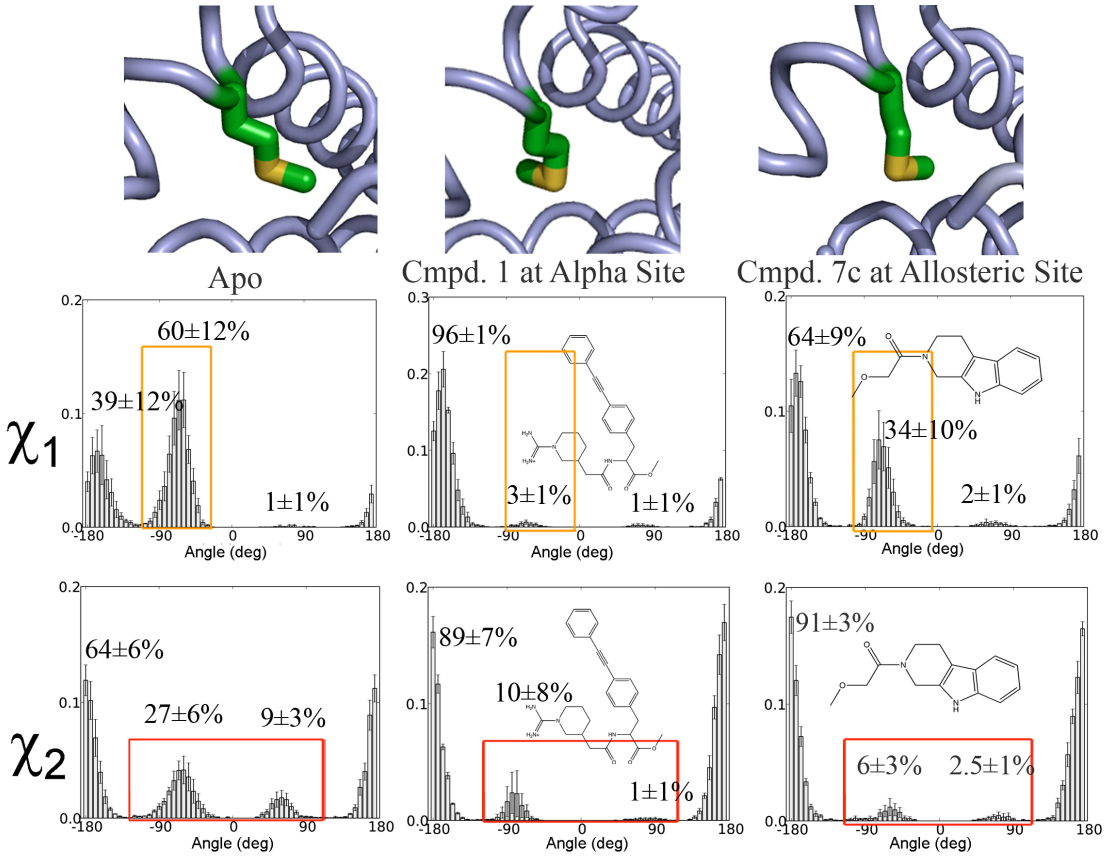


Figure 10. Compound binding to the IL-2Rα site or to the allosteric site selects conformations of Met39 favorable for binding compound at the other site. (Top) Conformations of Met39 in apo and compound-bound crystal structures (PDB IDs 1M47, 1M48, and 1NBP, resp.). (Bottom) Histograms of Met39's χ_1 and χ_2 angles from MD simulations. Orange boxes in χ1 and red boxes in χ2 highlight populations suppressed in ligand-bound simulations.

Conformational Selection *In Silico* by Allosteric Ligand

The preceding analysis suggests that binding at the allosteric site can positively modulate binding at the competitive site through changing the distribution of side chain rotamers. To more directly assess the relationship between these relatively subtle changes and ligand binding, we have performed small-molecule docking against snapshots from the MD simulations (Figure 11). Although the scores from docking to MD snapshots cannot be interpreted as accurate binding affinities, we use them as way of

qualitatively assessing whether the conformation of the site is appropriate for binding the competitive inhibitor. Because this site consists of many flexible side-chains, RMSD of the competitive binding site residues to the complexed crystal structure was not an appropriate measure of whether the competitive site's conformation was favorable for competitive ligand binding. We found that the best-scoring docked pose roughly superimposes with the crystal ligand (2.8 Å RMSD without fitting, 1.4 Å RMSD with rotational/translational fitting, Figure 11C). The cluster of MD snapshots with the best-scoring docked ligand represented 0.033% of total snapshots. Thus, our relatively short molecular dynamics simulations sampled conformers suitable for binding competitive site ligand at 300K in the presence of allosteric ligand, enabling us to create a model of the ternary complex (Figure 11C).

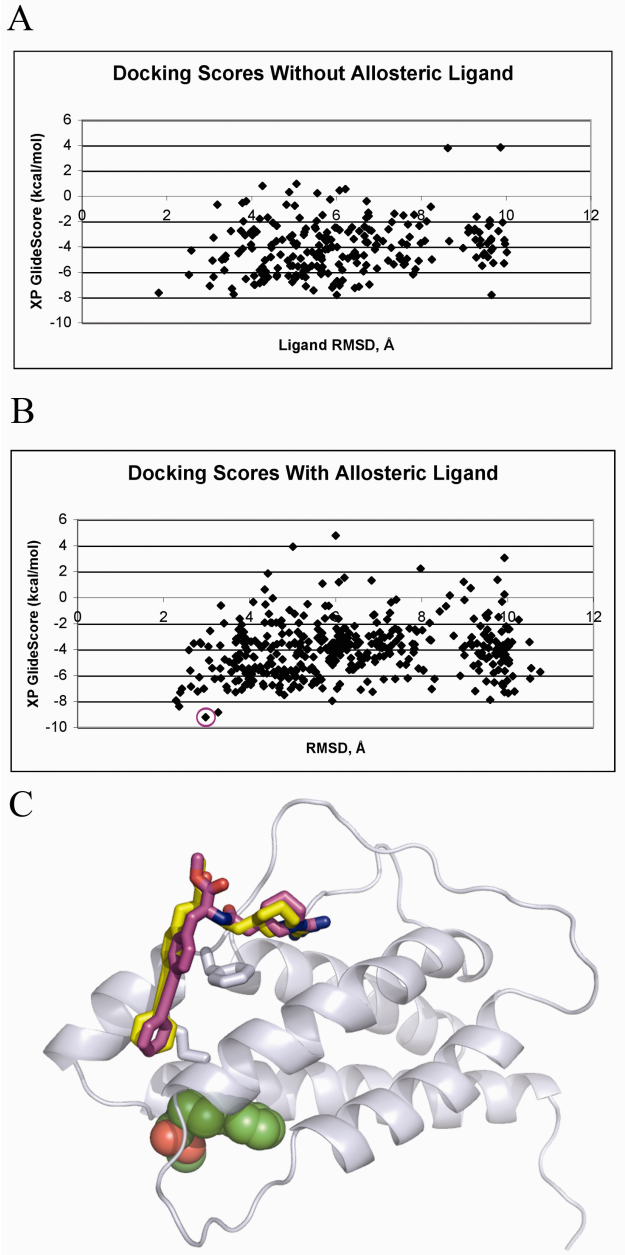


Figure 11. Docking using Glide XP selects a holo-like conformation from an MD ensemble. (A, B) Plots of docking score vs. ligand RMSD to the forcefield-minimized co-crystal conformation show that the best-scoring docked poses from simulations with (B) but not without (A) allosteric compound bound had relatively low RMSD values. The best-scoring pose is circled. (C) Molecular dynamics snapshot from a simulation of IL-2 with bound allosteric fragment is shown with docked (yellow) vs. superimposed X-ray (magenta) conformations of a micromolar small molecule inhibitor of IL-2Rα binding. Though the absolute RMSD for this ligand is 2.9Å (1.6Å RMSD after fitting), it has a binding mode very similar to that of the crystal ligand.

Conclusions

We have reported novel improvements to mutual information calculations that make them robust enough for relatively short molecular dynamics simulations and have applied our MutInf method to interrogate the mechanism of small molecule binding cooperativity in human interleukin-2. We found better separation of signal from noise in

our matrix of correlations between pairs of torsions than in similar matrices that examined backbone C_α correlations. We identified not only local correlations in sequence and in distance space but also long-range correlations. Clustering the matrix of mutual information between residues, we identified a few clusters whose residues showed strong patterns of correlations. Two of these clusters highlighted key functional sites, namely the IL-2Rα-competitive protein interface/inhibitor binding site and a highly flexible loop that has to move to reveal a cryptic binding pocket for the allosteric ligand. Furthermore, we found that the conformations of a number of pairs of residues in these two functional sites were strongly correlated.

As MutInf identified known cooperative binding or functional sites within the top clusters and correlations between them, we believe that this approach will be useful in identifying novel allosteric sites for proteins or for small molecules. For example, we predict potential allostery between the flexible loop surrounding the allosteric compound binding site and the N-terminus of helix 1, the loop region around Glu100, and the C-terminus. Our prediction is further supported by the observation that all of these regions showed significant backbone NMR chemical shift perturbations upon binding of the IL-2Rα receptor[61]. However, the biological roles of these regions are not clear. The C-terminus of IL-2 interacts weakly with the γ_c receptor[75, 80, 81] (K_D ~0.7 mM) and independently of IL-2Rα binding. NMR chemical shift perturbations and isoelectric point changes upon addition of methionine to IL-2's N-terminus suggested a potential interaction between the N- and C-termini of apo IL-2 in solution[82]. Intriguingly, Thr3 is a site on human IL-2 that is variably glycosylated[83]; in mice, the N-terminus is longer and displays substantial sequence and glycosylation pattern variability, which in turn impacts

IL-2's function in Type I diabetes in mouse models[84, 85]. An important caveat is that correlated motions are necessary but not sufficient for biologically-relevant allostery.

Though our statistical filtering enables us to find significant correlations in relatively short simulations, in applications where accurate total conformational entropy is desired, multiple longer simulations may be required to obtain absolute total entropy values that all converge to the same value, and higher-order terms might be needed. Nonetheless, our approach is useful in determining which residues or groups of residues show correlated conformations and which residues may mediate crosstalk between functional sites. One caveat is that MutInf focuses on coupled residue conformations rather than vibrations, and so we may not efficiently capture the role of semi-rigid elements in mediating correlations between more flexible sites. Though we did not look at motions faster than 1 ps, these are likely not as critical for ligand binding cooperativity in interleukin-2, where the residues linking the sites are primarily in flexible loops and have flexible side-chains. In general, however, such faster-timescale motions can help mediate cooperativity between more flexible sites, and so future work is needed to properly account for these in our approach.

Our calculations suggest that small molecule binding cooperativity in human interleukin-2 involves subtle population shifts and correlated conformations of two binding pockets coupled through a greasy core and a solvent-exposed polar network. New biophysical techniques to directly measure correlated motions by NMR would be useful in testing our predictions about correlated motions that couple allosteric sites.

References

1. Tsai, C.-J.; del Sol, A.; Nussinov, R. Allostery: Absence of a Change in Shape Does Not Imply that Allostery Is Not at Play. *J. Mol. Biol.* **2008,** *378* (1), 1-11.

2. Cui, Q.; Karplus, M. Allostery and cooperativity revisited. *Protein Sci.* **2008,** *17* (8), 1295-1307.

3. Lindsley, J. E.; Rutter, J. Whence cometh the allosterome? *Proc. Natl. Acad. Sci. U. S. A.* **2006,** *103* (28), 10533-10535.

4. Gunasekaran, K.; Ma, B.; Nussinov, R. Is allostery an intrinsic property of all dynamic proteins? *Proteins: Struct., Funct., Bioinf.* **2004,** *57* (3), 433-443.

5. Kuriyan, J.; Eisenberg, D. The origin of protein interactions and allostery in colocalization. *Nature.* **2007,** *450* (7172), 983-990.

6. Hardy, J. A.; Wells, J. A. Searching for new allosteric sites in enzymes. *Curr. Opin. Struct. Biol.* **2004,** *14* (6), 706-715.

7. Jeffrey Conn, P.; Christopoulos, A.; Lindsley, C. W. Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nat. Rev. Drug Discov.* **2009,** *8* (1), 41-54.

8. Zhang, J.; Yang, P. L.; Gray, N. S. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer.* **2009,** *9* (1), 28-39.

9. Hardy, J. A.; Wells, J. A. Searching for new allosteric sites in enzymes. *Curr. Opin. Struct. Biol.* **2004,** *14* (6), 706-715.

10. Shulman, A. I.; Larson, C.; Mangelsdorf, D. J.; Ranganathan, R. Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell.* **2004,** *116* (3), 417-429.

11. Suel, G. M.; Lockless, S. W.; Wall, M. A.; Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* **2003,** *10* (1), 59-69.

12. Lockless, S. W.; Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science.* **1999,** *286* (5438), 295-299.

13. Hatley, M. E.; Lockless, S. W.; Gibson, S. K.; Gilman, A. G.; Ranganathan, R. Allosteric determinants in guanine nucleotide-binding proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2003,** *100* (24), 14445-14450.

14. Lee, J.; Natarajan, M.; Nashine, V. C.; Socolich, M.; Vo, T.; Russ, W. P.; Benkovic, S. J.; Ranganathan, R. Surface sites for engineering allosteric control in proteins. *Science.* **2008,** *322* (5900), 438-442.

15. Fuentes, E. J.; Gilmore, S. A.; Mauldin, R. V.; Lee, A. L. Evaluation of Energetic and Dynamic Coupling Networks in a PDZ Domain Protein. *J. Mol. Biol.* **2006,** *364* (3), 337-351.

16. Page, M. J.; Carrell, C. J.; Di Cera, E. Engineering Protein Allostery: 1.05 Å Resolution Structure and Enzymatic Properties of a Na+-activated Trypsin. *J. Mol. Biol.* **2008,** *378* (3), 666-672.

17. Ming, D.; Wall, M. E. Quantifying allosteric effects in proteins. *Proteins: Struct., Funct., Bioinf.* **2005,** *59* (4), 697-707.

18. Hawkins, R. J.; McLeish, T. C. B. Coarse-Grained Model Of Entropic Allostery. *Phys. Rev. Lett.* **2004,** *93* (9), 098104.

19. Ming, D.; Wall, M. E. Allostery in a Coarse-Grained Model of Protein Dynamics. *Phys. Rev. Lett.* **2005,** *95* (19), 198103.

20. Hawkins, R. J.; McLeish, T. C. B. Dynamic allostery of protein alpha helical coiled-coils. *J. Royal Soc. Interface.* **2006,** *3* (6), 125-138.

21. Zhang, D.; McCammon, J. A. The Association of Tetrameric Acetylcholinesterase with ColQ Tail: A Block Normal Mode Analysis. *PLoS Comput. Biol.* **2005,** *1* (6), e62.

22. Chennubhotla, C.; Yang, Z.; Bahar, I. Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL. *Mol. BioSyst.* **2008,** *4* (4), 287-292.

23. Dengming Ming, M. E. W. Quantifying allosteric effects in proteins. *Proteins: Struct., Funct., Bioinf.* **2005,** *59* (4), 697-707.

24. Chennubhotla, C.; Bahar, I. Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. *Mol. Syst. Biol.* **2006,** *2*.

25. Daily, M. D.; Upadhyaya, T. J.; Gray, J. J. Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins: Struct., Funct., Bioinf.* **2008,** *71* (1), 455-466.

26. Daily, M. D.; Gray, J. J. Local motions in a benchmark of allosteric proteins. *Proteins: Struct., Funct., Bioinf.* **2007,** *67* (2), 385-399.

27. Daily, M. D.; Gray, J. J. Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput. Biol.* **2009,** *5* (2), e1000293.

28. Li, L.; Uversky, V. N.; Dunker, A. K.; Meroueh, S. O. A Computational Investigation of Allostery in the Catabolite Activator Protein. *J. Am. Chem. Soc.* **2007,** *129* (50), 15668-15676.

29. Bradley, M. J.; Chivers, P. T.; Baker, N. A. Molecular Dynamics Simulation of the Escherichia coli NikR Protein: Equilibrium Conformational Fluctuations Reveal

Interdomain Allosteric Communication Pathways. *J. Mol. Biol.* **2008,** *378* (5), 1155-1173.

30. Sayar, K.; Ugur, O.; Liu, T.; Hilser, V.; Onaran, O. Exploring allosteric coupling in the alpha-subunit of Heterotrimeric G proteins using evolutionary and ensemble-based approaches. *BMC Struct. Biol.* **2008,** *8* (1), 23.

31. Horstink, L. M.; Abseher, R.; Nilges, M.; Hilbers, C. W. Functionally important correlated motions in the single-stranded DNA-binding protein encoded by filamentous phage Pf3. *J. Mol. Biol.* **1999,** *287* (3), 569-577.

32. Liu, J.; Nussinov, R. Allosteric effects in the marginally stable von Hippel-Lindau tumor suppressor protein and allostery-based rescue mutant design. *Proc. Natl. Acad. Sci. U. S. A.* **2008,** *105* (3), 901-906.

33. Watney, J. B.; Hammes-Schiffer, S. Comparison of Coupled Motions in Escherichia coli and Bacillus subtilis Dihydrofolate Reductase. *J. Phys. Chem. B.* **2006,** *110* (20), 10130-10138.

34. Lange, O. F.; Grubmüller, H., Generalized correlation for biomolecular dynamics. *Proteins: Struct., Funct., Bioinf.* **2006,** *62* (4), 1053-1061.

35. Lange, O. F.; Grubmüller, H.; de Groot, B. L. Molecular Dynamics Simulations of Protein G Challenge NMR-Derived Correlated Backbone Motions. *Angew. Chem. Intl. Ed.* **2005,** *44* (22), 3394-3399.

36. Ming, D.; Cohn, J.; Wall, M. Fast dynamics perturbation analysis for prediction of protein functional sites. *BMC Struct. Biol.* **2008,** *8* (1), 5.

37. Ho, B. K.; Agard, D. A. Probing the Flexibility of Large Conformational Changes in Protein Structures through Local Perturbations. *PLoS Comput. Biol.* **2009,** *5* (4), e1000343.

38. Cooper, A.; Dryden, D. T. F. Allostery without conformational change. *Eur. Biophys. J.* **1984,** *11* (2), 103-109.

39. Popovych, N.; Sun, S.; Ebright, R. H.; Kalodimos, C. G. Dynamically driven protein allostery. *Nat. Struct. Mol. Biol.* **2006,** *13* (9), 831-838.

40. Lenaerts, T.; Ferkinghoff-Borg, J.; Stricher, F.; Serrano, L.; Schymkowitz, J.; Rousseau, F. Quantifying information transfer by protein domains: Analysis of the Fyn SH2 domain structure. *BMC Struct. Biol.* **2008,** *8* (1), 43.

41. Smith, C. A.; Kortemme, T. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. *J. Mol. Biol.* **2008,** *380* (4), 742-756.

42. Friedland, G. D.; Linares, A. J.; Smith, C. A.; Kortemme, T. A Simple Model of Backbone Flexibility Improves Modeling of Side-chain Conformational Variability. *J. Mol. Biol.* **2008,** *380* (4), 757-774.

43. Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. Extraction of configurational entropy from molecular simulations via an expansion approximation. *J. Chem. Phys.* **2007,** *127* (2), 024107-024116.

44. Zheng, W.; Brooks, B. R.; Thirumalai, D. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Natl. Acad. Sci. U. S. A.* **2006,** *103* (20), 7664-7669.

45. Matsuda, H. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Phys. Rev. E.* **2000,** *62* (3), 3096.

46. Steuer, R.; Kurths, J.; Daub, C. O.; Weise, J.; Selbig, J. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics.* **2002,** *18* (suppl_2), S231-240.

47. Grassberger, P. Finite sample corrections to entropy and dimension estimates. *Phys. Lett. A.* **1988,** *128* (6-7), 369-373.

48. Roulston, M. S. Estimating the errors on measured entropy and mutual information. *Physica D.* **1999,** *125* (3-4), 285-294.

49. Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods. *J. Comput. Chem.* **2008,** *29* (10), 1605-1614.

50. Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. *J. Comput. Chem.* **2007,** *28* (3), 655-668.

51. Karchin, R.; Kelly, L.; Sali, A. Improving functional annotation of non-synonomous SNPs with information theory. *Pac. Symp. Biocomput.* **2005**, 397-408.

52. George Shackelford, K. K., Contact prediction using mutual information and neural nets. *Proteins: Struct., Funct., Bioinf.* **2007,** *69* (S8), 159-164.

53. Francois, D.; Rossi, F.; Wertz, V.; Verleysen, M. Resampling methods for parameter-free and robust feature selection with mutual information. *Neurocomputing* **2007,** *70* (7-9), 1276-1288.

54. Hutter, M.; Zaffalon, M. Distribution of mutual information from complete and incomplete data. *Comp. Stat. & Data Anal.* **2005,** *48* (3), 633-657.

55. Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **2001,** *7* (8), 306-317.

56. Spoel, D. V. D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005,** *26* (16), 1701-1718.

57. Sorin, E. J.; Pande, V. S. Exploring the Helix-Coil Transition via All-Atom Equilibrium Ensemble Simulations. *Biophys. J.* **2005,** *88* (4), 2472-2493.

58. Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. A hierarchical approach to all-atom protein loop prediction. *Proteins: Struct., Funct., and Bioinf.* **2004,** *55* (2), 351-367.

59. Alexov, E. G.; Gunner, M. R., Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys. J.* **1997,** *72* (5), 2075-2093.

60. Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. Combining Conformational Flexibility and Continuum Electrostatics for Calculating pKas in Proteins. *Biophys. J.* **2002,** *83* (4), 1731-1748.

61. Emerson, S. D.; Palermo, R.; Liu, C.-M.; Tilley, J. W.; Chen, L.; Danho, W.; Madison, V. S.; Greeley, D. N.; Ju, G.; Fry, D. C. NMR characterization of interleukin-2 in complexes with the IL-2Rα receptor component, and with low molecular weight compounds that inhibit the IL-2/IL-Rα interaction. *Protein Sci.* **2003,** *12* (4), 811-822.

62. Arkin, M. R.; Randal, M.; DeLano, W. L.; Hyde, J.; Luong, T. N.; Oslob, J. D.; Raphael, D. R.; Taylor, L.; Wang, J.; McDowell, R. S.; Wells, J. A.; Braisted, A. C.

Binding of small molecules to an adaptive protein-protein interface. *Proc. Natl. Acad. Sci. U. S. A.* **2003,** *100* (4), 1603-1608.

63. Hyde, J.; Braisted, A. C.; Randal, M.; Arkin, M. R. Discovery and characterization of cooperative ligand binding in the adaptive region of interleukin-2. *Biochemistry (Mosc).* **2003,** *42* (21), 6475-6483.

64. Junmei, W.; Romain, M. W.; James, W. C.; Peter, A. K.; David, A. C. Development and testing of a general amber force field. *J. Comput. Chem.* **2004,** *25* (9), 1157-1174.

65. Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000,** *21*, 132.

66. Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002,** *23*, 1623.

67. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984,** *81* (8), 3684-3690.

68. Berk, H.; Henk, B.; Herman, J. C. B.; Johannes, G. E. M. F., LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997,** *18* (12), 1463-1472.

69. Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; Gunsteren, W. F. v.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew. Chem. Intl. Ed.* **1999,** *38* (1-2), 236-240.

70. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006,** *49* (21), 6177-6196.

71. Yang, L.-W.; Rader, A. J.; Liu, X.; Jursa, C. J.; Chen, S. C.; Karimi, H. A.; Bahar, I. oGNM: online computation of structural dynamics using the Gaussian Network Model. *Nucl. Acids Res.* **2006,** *34* (suppl_2), W24-31.

72. Dill, K.; Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Chemistry & Biology*. Garland Science: 2002.

73. Hilser, V. J.; Thompson, E. B. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2007,** *104* (20), 8311-8315.

74. Rekharsky, M. V.; Mori, T.; Yang, C.; Ko, Y. H.; Selvapalam, N.; Kim, H.; Sobransingh, D.; Kaifer, A. E.; Liu, S.; Isaacs, L.; Chen, W.; Moghaddam, S.; Gilson, M. K.; Kim, K.; Inoue, Y. A synthetic host-guest system achieves avidin-biotin affinity by overcoming enthalpy entropy compensation. *Proc. Natl. Acad. Sci. U.S.A.* **2007,** *104* (52), 20737-20742.

75. Wang, H.; Kakaradov, B.; Collins, S. R.; Karotki, L.; Fiedler, D.; Shales, M.; Shokat, K. M.; Walther, T.; Krogan, N. J.; Koller, D. A complex-based reconstruction of the S. cerevisiae interactome. *Mol. Cell. Proteomics.* **2009**, M800490-MCP800200.

76. Rickert, M.; Wang, X.; Boulanger, M. J.; Goriatcheva, N.; Garcia, K. C. The structure of interleukin-2 complexed with its alpha receptor. *Science.* **2005,** *308* (5727), 1477-1480.

77. Fruchterman, T. M. J.; Reingold, E. M. Graph drawing by force-directed placement. *Softw. Pract. Exper.* **1991,** *21* (11), 1129-1164.

78. Thanos, C. D.; DeLano, W. L.; Wells, J. A. Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc. Natl. Acad. Sci. U.S.A.* **2006,** *103* (42), 15422-15427.

79. Thanos, C. D.; DeLano, W. L.; Wells, J. A. Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc. Natl. Acad. Sci. U.S.A.* **2006,** *103* (42), 15422-15427.

80. Liparoto, S. F.; Myszka, D. G.; Wu, Z.; Goldstein, B.; Laue, T. M.; Ciardelli, T. L. Analysis of the Role of the Interleukin-2 Receptor Gamma Chain in Ligand Binding. *Biochemistry.* **2002,** *41* (8), 2543-2551.

81. Stauber, D. J.; Debler, E. W.; Horton, P. A.; Smith, K. A.; Wilson, I. A. Crystal structure of the IL-2 signaling complex: paradigm for a heterotrimeric cytokine receptor. *Proc. Natl. Acad. Sci. U. S. A.* **2006,** *103* (8), 2788-2793.

82. Endo, S.; Yamamoto, Y.; Sugawara, T.; Nishimura, O.; Fujino, M. The Additional Methionine Residue at the N-Terminus of Bacterially Expressed Human Interleukin-2 Affects the Interaction between the N- and C-Termini. *Biochemistry.* **2001,** *40* (4), 914-919.

83. Robb, R. J.; Smith, K. A. Heterogeneity of human T-cell growth factor(s) due to variable glycosylation. *Mol. Immunol.* **1981,** *18* (12), 1087-1094.

84. Podolin, P. L.; Wilusz, M. B.; Cubbon, R. M.; Pajvani, U.; Lord, C. J.; Todd, J. A.; Peterson, L. B.; Wicker, L. S.; Lyons, P. A. Differential glycosylation of interleukin 2, the molecular basis for the NOD Idd3 type 1 diabetes gene? *Cytokine.* **2000,** *12* (5), 477-482.

85. Sgouroudis, E.; Albanese, A.; Piccirillo, C. A. Impact of Protective IL-2 Allelic Variants on CD4+Foxp3+ Regulatory T Cell Function In Situ and Resistance to Autoimmune Diabetes in NOD Mice. *J. Immunol.* **2008,** *181* (9), 6283-6292.

Chapter 3. The Kullback-Leibler divergence expansion reveals effects of allosteric perturbations

E-mail:

Abstract

We present a novel thermodnyamical approach to identify changes in macromoluecular structure and dynamics in response to perturbations such as mutations or ligand binding given molecular dynamics simulations of the unperturbed and perturbed constructs, using an expansion of the Kullback-Leibler Divergence that connects local population shifts in torsion angles to changes in the free energy landscape of the protein. While the Kullback-Leibler Divergence is a known formula from information theory, the novelty and power of our approach lies in its formal developments, connection to thermodynamics, built-in statistical filtering, ease of visualization of results, and extendability by adding higher-order terms.

We present a formal derivation of the Kullback-Leibler Divergence expansion and then apply our method at a first-order approximation to three protein systems where ligand binding, post-translational modification, or pH titration is known from experiments to cause an effect at an allosteric site. Our results on these systems are qualitatively in agreement with experimental approaches measuring local changes in structure

*To whom correspondence should be addressed

or dynamics such as NMR chemical shift perturbations and hydrogen-deuterium exchange mass spectrometry. As our method produces easy-to-analyze results with low backgroud, it has the potential to become a routine analysis when molecular dynamics simulations in two or more conditions are available.

1 Introduction

We describe a computational approach that can be used to quantify our intuition about effects of ligand binding on a conformational ensemble in a thermodynamically meaningful way and formulate hypotheses about differences in conformational ensembles at the residue-level that can be tested by alanine scanning or other mutagenesis techniques. Moreover, the results are easy-to-visualize and filtered for statistical significance so that the eye is immediately drawn to residues showing significant perturbations.

With recent advancements in software and hardware for molecular dynamics simulations[1,2],[3] and the increased presentation of conformational ensembles produced with guidance or restraints from experimental data[4],[5] there is an increased need for analysis techniques that can make statistical inferences regarding these conformational ensembles. One important challenge is to determine, in an unbiased way, how perturbations such as ligand or protein binding or post-translational modification alter conformational ensembles. Residues that show the largest response to particular perturbations could be important for sending or receiving biochemical signals.

In a protein or protein complex, we would like to know where significant changes in structure and dynamics occur following a perturbation such as ligand binding, mutation, or post-translational modification. Also, we would like to know whether two different ligands binding to the same site cause similar or different effects at distant sites. All of these phenomena can involve both conformational changes and flexibility changes and can be gleaned from rigorous thermodynamic conformational ensembles such as those generated by methods such as molecular dynamics. A large number of methods have been previously

presented to quantify changes in structure in dynamics. Here, we present a method for comparing two conformational ensembles – capturing both coformational changes and entroy changes – based on a thermodynamic framework, and with built-in statistical significance filters. Furthermore, Identifying these differential effects of ligand binding would be useful in allosteric activator or inhibitor design, where binding alone is not sufficient for a change in activity. To answer these quenstions, we calculate shifts in the populations of residues' conformational distributions at near and distant sites following a given perturbation. We analyze residues' conformational distributions in torsion space, as torsions provide an apt local description of biologically-relevant functional motions and do not have frame-fitting issues inherent to Cartesian analysis.[6] Especially for protein side chains, notions about "average" positions and changes to "average" position are of limited use in making predictions from equilibrium simulations, as their distributions are often multimodal, in torsional or Cartesian space.

To identify protein residues exhibiting changes in structure and dynamics following ligand binding, mutation, or post-translational modification, we compare conformational ensembles (i.e. from molecular simulations) in a thermodynamically meaningful way. How to do this comparison is a distinct question from merely comparing structures, or comparing flexibilities or entropies of various degrees of freedom. Due to the high-dimensionality of a molecule's conformational space, approaches to compare molecular conformational ensembles typically focus on global phenomena or localized phenomena. Approaches for capturing global differences between conformational ensembles typically reduce the dimensionality by discretizing conformational space over a subset of degrees of freedom (i.e. Cα atoms) into rapidly-converting "microstates" and slowly-converting "macrostates",[5] by changing basis into a subset of the most significant collective coordinates by performing some flavor of principal coordinates analysis[7,6].[8] Approaches for capturing localized differences between conformational ensembles, however, typically focus on average structural changes, average flexibility changes, contact maps,[9] or correlated motions[10].[11] In this

work, we develop a method for quantifying and visualizing localized differences between conformational ensembles that can capture subtle changes in structure and/or flexibility, filtered through statistical corrections. Our method is not intended to replace these other methods, but rather to provide a tool to identify key players and hypotheses to test by mutating one or more residues. The key distinctions of our method are its connection to thermodynamics and ease of visualization that is on par with standard RMSD and r.m.s. fluctuation analyses.

To quantify "population shifts" in residues' conformational distributions, we use the Kullback-Leibler divergence, a measure of the free energy difference between two equilibrium ensembles, where one ensemble is the "reference" ensemble and the other is the "perturbed" ensemble[12].[13] The Kullback-Leibler divergence or relative entropy is a fundamental quantity in information theory, and its differential version is given by:

$$KL(x_1,...,x_m||x_1^*,...,x_m^*) = \int J(x_1,...,x_m)\rho(x_1,...,x_m)\ln\frac{\rho(x_1,...,x_m)}{\rho^*(x_1^*,...,x_m^*)}dx_1,...,dx_m, \quad (1)$$

where ρ^* is the probability density function (p.d.f.) of the reference ensemble, and ρ is the p.d.f. of the perturbed ensemble, and J indicates the Jacobian determinant. Torsion angles (with fixed bond lengths and angles) or orthogonal Cartesian basis sets have a Jacobian of unity, facilitating analysis.

The Kullback-Leibler divergence was previously derived to second order for a harmonic Hamiltonian and applied to normal-modes models of proteins.[14] It was applied to trypsinogen to not only refine a normal-mode model against atomistic simulation data, but also to quantify coupling between trypsinogen's active and regulatory sites. In a different study, this measure was applied to identify functional sites in a large test set of proteins. This approach to identify functional sites was based on the observation that functional sites tended to co-localize with surface sites where addition of a spherical probe and harmonic couplings to nearby protein atoms caused a large change in the total Kullback-Leibler di-

vergence for the protein.[15] d The Kullback-Leibler divergence was also applied at second order in a perturbational formulation of principal components analysis (PCA) to identify effective perturbations deriving from arbitrary linearly independent perturbation functions that contribute to differences between conformational ensembles.[6] In PCA of Cα atoms, a common practice in molecular dynamics simulations, these perturbation functions are identity operators on the Cartesian coordinates. Here, the eigenvalues of the perturbation functions' covariance matrix (or that of the Cα coordinates in typical applications of PCA) are related to the Kullback-Leibler divergence between ensembles.

To calculate the Kullback-Leibler divergence for macromolecular conformational ensembles containing many degrees of freedom, we propose an expansion over increasing numbers of degrees of freedom. We derive a novel expansion of the K-L divergence over single degrees of freedom, pairs of degrees of freedom, etc, utilizing the Generalized Kirkwood Superposition Approximation (GKSA), which has been previously used by Matsuda[16] and by Killian et. al. for a configurational entropy expansion.[17]

The most immediate application in our work is use of first-order terms to calculate the Kullback-Leibler divergence for protein residues from sums of the Kullback-Leibler divergences of their constituent torsions, which could be readily refined by use of second-order terms within residues. We expect that second- and higher-order terms will also be useful in future applications. Importantly, our expansion connects such "local" Kullback-Leibler divergences to the global Kullback-Leibler divergence for the conformational ensemble, which has connections to the free energy; other measures of comparison such as r.m.s. deviation or chi-squared analysis lack this strong connection to thermodynamics. Our method scales linearly with the number of residues in the protein (neglecting inter-residue second-order and higher-order terms) and is thus applicable to large proteins, protein-protein complexes, and DNA/RNA.

The novelty of our work lies in: (1) providing a thermodynamics-based local comparison between conformational ensembles that accounts for both changes in structure and

dynamics, in contrast with commonly-used methods such as root-mean-squared deviation (RMSD) or root-mean-squared fluctuation (RMSF or B-factor analysis) (2) deriving an expansion that proscribes a way to compare distributions of multiple degrees of freedom (ex. the multiple torsions of a protein residue or those of a group of residues), and a systematic way to improve the accuracy of such comparisons (3) our discovery of a useful quantity, what we call the "mutual divergence"–analogous to mutual information (except it uses relative entropy instead of entropy)–and it's higher-order analog, (4) providing a discretization correction to the Kullback-Leibler divergence, and (5) using bootstrap resampling on the Kullback-Leibler divergence for statistical filtering and correction for sampling bias.

1.1 Marginal probability distributions of configurational space and the Generalized Kirkwood Superposition Approximation

A protein's geometry is most commonly described in Cartesian coordinates or in internal bond-angle-torsion (BAT) coordinates. We use BAT coordinates, and focus our analysis on ϕ, ψ, and χ torsion angles, as we believe these capture motions of most biophysical relevance. The distribution of the m torsion angles $(x_1, ..., x_m)$ of a protein's "perturbed" equilibrium conformational ensemble (perturbed by mutation, ligand binding, post-translational modification, etc.) give rise to a probability distribution $\rho(x_1, ..., x_m)$ over m degrees of freedom; these are compared with a "reference" conformational ensemble having probability distribution $\rho*$.

As the number of snapshots of a protein's geometry required to adequately approximate this m-dimensional probability distribution function (p.d.f.) grows exponentially with increasing m, we wish to approximate the m-dimensional p.d.f. using marginal distributions of $\rho(x_1, ..., x_m)$ involving only one and two variables. Such marginal distributions of order

n are defined as follows:

$$\rho_1(x_j) = \int J(x_1,...,x_m)\rho(x_1,...,x_m)\prod_{i\neq j} dx_i = \rho_{1,\mathbf{s}=\{j\}}, \tag{2}$$

$$\rho_2(x_j,x_k) = \int J(x_1,...,x_m)\rho(x_1,...,x_m)\prod_{i\neq j,k} dx_i = \rho_{2,\mathbf{s}=\{j,k\}}, \tag{3}$$

$$\rho_n(x_{j1},...,x_{jn}) = \int J(x_1,...,x_m)\rho(x_1,...,x_m)\prod_{i\neq\{j_n\}} dx_i = \rho_{n,\mathbf{s}=\{x_1,...,x_{n-1}\}} \tag{4}$$

Where **s** denotes a set of degrees of freedom. In what follows, the subscript of probability densities $\rho_{n,k}$ will either have one index indicating the number of degrees of freedom and an argument list, or be expressed in shortened notation using two indices: the first, n, indicating the number of degrees of freedom in the probability density function, and the second, $\{k\}$, indicating a set of indices of degrees of freedom comprising the p.d.f.

The Generalized Kirkwood Superposition Approximation (GKSA) is of key importance for the foundation of the present work. The GKSA at order $m-1$ approximates a probability distribution with m degrees of freedom using lower-order probability density functions consisting of a subset of the degrees of freedom, up to $m-1$ degrees of freedom for an order $m-1$ GKSA, and is perhaps easiest to express in log form:

$$\ln\hat{\rho}_m^{m-1}(x_1,...,x_m) = \sum_{n=1}^{m-1}(-1)^{m-n+1}\ln\prod_{\mathbf{k}}^{C_n^m}\rho_{n,\mathbf{k}} \tag{5}$$

where C_n^m indicates all $\binom{m}{n}$ combinations of nth-order marginal probability density functions of ρ, and $\hat{\rho}^{m-1}$ indicates the order m-1 GKSA approximation of ρ.

2 Methods

2.1 Kullback-Leibler Divergence Expansion for three variables

To motivate the expansion of the Kullback-Leibler divergence, consider a probabilitiy distribution $\rho(x_1,...,x_m) = \rho(\phi,\psi,\chi)$ that is a function of a set τ of three variables,$\tau = \{\phi,\psi,\chi\}$. Suppose for example that these three variables denote the backbone and first sidechain torsion angles of an amino acid in a peptide or protein. The Kirkwood expansion for ρ is then:

$$\rho_3(\phi,\psi,\chi) \approx \frac{\rho_2(\phi,\psi)\rho_2(\phi,\chi)\rho_2(\psi,\chi)}{\rho_1(\phi)\rho_1(\psi)\rho_1(\chi)} = \frac{\prod_{\mathbf{g}}^{C_2^3}\rho_{2,\mathbf{g}}}{\prod_{\mathbf{k}}^{C_1^3}\rho_{1,\mathbf{k}}} \tag{6}$$

Where the notation C_p^q denotes all q-choose-p combinations of order-p marginal distributions, and **g** and **k** denote two-member and one-member sets of degrees of freedom comprising a particular combination of these C_p^q order-p marginals. Consider probability distributions ρ and ρ^* over m degrees of freedom.

Continuing with our example, inserting the Kirkwood expansion for ρ and ρ^* into the equation above yields:

$$KL(x_1,...,x_m||x_1^*,...,x_m^*) = \int \rho(x_1,...,x_m) \ln \frac{\frac{\prod_{\mathbf{g}}^{C_2^3}\rho_{2,\mathbf{g}}}{\prod_{\mathbf{k}}^{C_1^3}\rho_{1}}}{\frac{\prod_{\mathbf{g}}^{C_2^3}\rho^*_{2,\mathbf{g}}}{\prod_{\mathbf{k}}^{C_1^3}\rho^*_{1,\mathbf{k}}}} dx_1,...,dx_m \tag{7}$$

Converting the log of a product into a sum of logs:

$$KL(x_1,...,x_m||x_1^*,...,x_m^*) = \int \rho(x_1,...,x_m)(\sum_{\mathbf{g}}^{C_2^3} \ln \frac{\rho_{2,\mathbf{g}}}{\rho^*_{2,\mathbf{g}}} - (\sum_{\mathbf{k}}^{C_1^3} \ln \frac{\rho_{1,\mathbf{k}}}{\rho^*_{1,\mathbf{k}}}) dx_1,...,dx_m \tag{8}$$

Due to linearity,

$$KL(x_1,...,x_m||x_1^*,...,x_m^*) = \sum_{\mathbf{g}}^{C_2^3} \int \rho(x_1,...,x_m) \ln \frac{\rho_{2,\mathbf{g}}}{\rho_{2,\mathbf{g}}^*} dx_1,...,dx_m \quad - \sum_{\mathbf{k}}^{C_1^3} \int \rho(x_1,...,x_m) \ln \frac{\rho_{1,\mathbf{k}}}{\rho_{1,\mathbf{k}}^*} dx_1,...,dx_m$$

For each of these sums of C_m^n combinations of log terms, we can integrate out the $m-n$ degrees of freedom that are not part of each log term, and define $d\tau^n$ as the differential volume element over the remaining n variables in each term.

$$KL(x_1,...,x_m||x_1^*,...,x_m^*) = ((\sum_{\mathbf{g}}^{C_2^3} \int \rho_{2,\mathbf{g}} \ln \frac{\rho_{2,\mathbf{g}}}{\rho_{2,\mathbf{g}}^*} d\tau^2) - (\sum_{\mathbf{k}}^{C_1^3} \int \rho_{1,\mathbf{k}} \ln \frac{\rho_{1,\mathbf{k}}}{\rho_{1,\mathbf{k}}^*} d\tau^1) \tag{9}$$

Expanding, we see that this is merely the sum of Kullback-Leibler divergences of pairwise p.d.f.'s (with respect to their equilibrium values) minus the Kullback-Leibler divergences of individual p.d.f.'s:

$$\begin{aligned} KL(x_1,...,x_m||x_1^*,...,x_m^*) = &- \int \rho_1(\phi) \ln \frac{\rho_2(\phi)}{\rho_2^* \phi} d\phi - \int \rho_1(\psi) \ln \frac{\rho_2(\psi)}{\rho_2^* \psi} d\psi \\ &- \int \rho_1(\chi) \ln \frac{\rho_2(\chi)}{\rho_2^* \chi} d\chi + \int \rho_2(\phi,\psi) \ln \frac{\rho_2(\phi,\psi)}{\rho_2^* \phi,\psi} d\phi d\psi \\ &+ \int \rho_2(\phi,\chi) \ln \frac{\rho_2(\phi,\chi)}{\rho_2^* \phi,\chi} d\phi d\chi + \int \rho_2(\psi,\chi) \ln \frac{\rho_2(\psi,\chi)}{\rho_2^* \psi,\chi} d\psi d\chi \end{aligned} \tag{10}$$

2.2 General derivation of Kullback-Leibler Divergence Expansion

Now that we have illustrated the Kullback-Leibler divergence Expansion for three degrees of freedom, we next provide a general derivation of the expansion to m degrees of freedom, following similar procedures used in the entropy expansion in Killian et. al.[17] and in

Matsuda.[16] Applying the GKSA approximation to p and p* inside the logarithm of Eq. (1),

$$KL^{m-1} = \int J(x_1,...,x_m) * \rho_m(x_1,...,x_m) \sum_{n=1}^{m-1} (-1)^{m-n+1} \ln \prod_{\mathbf{k}}^{C_n^m} \frac{\rho_{n,\mathbf{k}}}{\rho^*_{n,\mathbf{k}}} dx_1,...,dx_m \qquad (11)$$

The superscript above KL denotes the order of the approximation to the Kullback-Leibler divergence. Again, C_n^m indicates all $\binom{m}{n}$ combinations of nth-order marginal probability density functions of ρ, and $\hat{\rho}^{m-1}$ indicates the order m-1 GKSA approximation of ρ. As before, **k** denotes n-member sets of degrees of freedom comprising a particular combination of these C_n^m order-n marginals. Converting the log of the product into a sum over logs and taking the sum outside the integral,

$$KL^{m-1} = \sum_{n=1}^{m-1} (-1)^{m-n+1} \sum_{\mathbf{k}}^{C_n^m} \int J(x_1,...,x_n) \rho_m(x_1,...,x_m) \ln \frac{\rho_{n,\mathbf{k}}}{\rho^*_{n,\mathbf{k}}} dx_1,...,dx_m \qquad (12)$$

We then integrate over the m-n dimensions that are independent of the log terms; these each integrate to unity. Next, define $d\tau^n$as the differential element of volume corresponding to the n dimensions that remain (including the remaining portions of the Jacobian determinant):

$$KL^{m-1} = \sum_{n=1}^{m-1} (-1)^{m-n+1} \{ \sum_{\mathbf{k}}^{C_n^m} \int \rho_{n,\mathbf{k}} \ln \frac{\rho_{n,\mathbf{k}}}{\rho^*_{n,\mathbf{k}}} d\tau^n \} \qquad (13)$$

As the term in braces in just an n-th order joint K-L divergence associated with each subset **k** of n degrees of freedom chosen from $(x_1,..,x_m)$, this simplifies to:

$$KL^{m-1} = \sum_{n=1}^{m-1} (-1)^{m-n+1} \{ \sum_{\mathbf{k}}^{C_n^m} KL_{n,\mathbf{k}} \} \qquad (14)$$

To calculate the requisite integrals, we can partition m-dimensional continuous torsional space into a discrete space of histogram bins. Each degree of freedom's marginal p.d.f. is discretized into hisogram bin probabilities p_i (with reference counts p_i^* , and joint histograms for marginal p.d.f.'s involving pairs of degrees of freedom are given by bin proba-

bilities p_{ij} (with refernece counts p^*_{ij}. These probabilities each must sum to unity: $\sum_i p_i = 1$, $\sum_{ij} p_{ij} = 1$. This partitioning leads to the following expansion over contributions from single degrees of freedom, pairs, triples, etc. for m degrees of freedom:

$$KL = (-1)^m \sum_{\mathbf{k}}^{C_1^m} \sum_{i}^{nbins} p_i \ln \frac{p_i}{p_i^*} + (-1)^{m-1} \sum_{\mathbf{g}}^{C_2^m} \sum_{ij}^{nbins^2} p_{ij} \ln \frac{p_{ij}}{p_{ij}^*} + ... \tag{15}$$

Now, we note that the signs of the terms depends on the number of terms: this is not ideal for an expansion. Thus, we need to introduce a term for the Kullback-Leibler divergence that will play the same role as the mutual information in the entropy expansion of Matsuda.[16] We call this the Mutual divergence, M, between two degrees of freedom, with marginal p.d.f.'s specified by p_i and p_j and joint histogram p_{ij} in one ensemble (the "target" ensemble), and with marginal p.d.f.'s specified by p_i and p_j and joint histogram p_{ij} in another ensemble (the "reference" ensemble):

$$M_2 = \sum_i p_i \ln \frac{p_i}{p_i^*} + \sum_j p_j \ln \frac{p_j}{p_j^*} - \sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{p_{ij}^*} \tag{16}$$

This can be equivalently expressed by combining terms into a single arument in the logarithm:

$$M_2 = \sum_i \sum_j p_{ij} \ln \frac{p_i p_j p_{ij}^*}{p_i^* p_j^* p_{ij}} \tag{17}$$

Here, the sums over i, j, and ij refer to one- and two-dimensional p.d.f.'s from a given pair g of degrees of freedom.

Alternatively, we can view this mutual divergence M_2 as a cross-information minus the mutual information of the reference state:

$$M_2 = \sum_i \sum_j p_{ij} \ln \frac{p_{ij}^*}{p_i^* p_j^*} - I(p_i^*, p_j^*) \tag{18}$$

Where $I(p_i^*, p_j^*)$ indicates the mutual information between these p.d.f.'s. Mutual diver-

gence can be generalized to higher order:

$$M_n(x_1,...,x_m||x_1^*,...,x_m^*) = \sum_{k=1}^{m}(-1)^{k+1}\sum_{i_1<...<i_k} D(x_{i_1},...,x_{i_k}||x_{i_1}^*,...,x_{i_k}^*) \tag{19}$$

Moreover, the mutual divergence satisfies a recursion relation analogous to Matsuda's recursion relation for higher-order mutual information:[16]

$$\begin{aligned} M_n(x_1,...,x_m||x_1^*,...,x_m^*) = {} & M_{m-1}(x_1,...,x_{m-2},x_{m-1}||x_1^*,...,x_{m-2}^*,x_{m-1}^*) \\ & + M_{m-1}(x_1,...,x_{m-2},x_m||x_1^*,...,x_{m-2}^*,x_m^*) \\ & - M_{m-1}(x_1,...,x_{m-2},x_{m-1}x_m||x_1^*,...,x_{m-2}^*,x_{m-1}^*x_m^*) \end{aligned} \tag{20}$$

Here, $x_{m-1}x_m$ indicates the joint distribution of these degrees of freedom, as in the third term of Eq. (16). In terms of the probability density, the higher-order mutual divergence is given by:

$$M_n(x_1,...,x_m||x_1^*,...,x_m^*) = (-1)^m \sum_{x_1,...,x_m} p_m(x_1,...,x_m)\ln\left(\frac{p_m^*(x_1,...,x_m)\hat{p}_{m-1}(x_1,...,x_m)}{\hat{p}_{m-1}^*(x_1,...,x_m)p_n(x_1,...,x_m)}\right) \tag{21}$$

where p_m is the target distribution, p_m^* is the reference distribution, and $\hat{p}_{m-1}$ and $\hat{p}_{m-1}^*$ are their Generalized Kirkwood Superposition Approximations, which consist of up to order $m-1$ probability densities. Applying this relation to the Kullback-Leibler divergence, we obtain the desired expansion:

$$KL = \sum_{n=1}^{m}\sum_{i}^{nbins} p_i \ln\frac{p_i}{p_i^*} - \sum_{n=1}^{m}\sum_{n'\neq n}\sum_{ij}^{nbins^2} p_{ij}\ln\frac{p_i p_j p_{ij}^*}{p_i^* p_j^* p_{ij}} + ... \tag{22}$$

2.3 Local Kullback-Leibler divergence

We are interested in population shifts caused by perturbations that reflect subtle changes in structure and/or dynamics in particular protein residues. We can visualize these most

readily using the first-order terms from our expansion. Consider the terms in the Kullback-Leibler divergence arising from a particular degree of freedom. These we will denote the "local" Kullback-Leibler divergence and provide an information-theoretic, quantitative measure of the extent to which the p.d.f. for a given degree of freedom deviates from the equilibrium p.d.f. This provides a much less-biased measure of changes in probability density than the chi-squared statistic.

$$KL_1 = \sum_{i}^{nbins} p_i \ln \frac{p_i}{p_i^*} \tag{23}$$

To calculate the Kullback-Leibler divergence for a single protein residue, we simply sum the Kullback-Leibler divergences between the reference and target ensemble for each of the residue's ϕ, ψ, and χ torsion angles:

$$KL_{res_n} = \sum_{\phi,\psi,\chi's} \sum_{i}^{nbins} p_i \ln \frac{p_i}{p_i^*} \tag{24}$$

While this expressions is very similar to the well-known $p \ln p$ expression for entropy, with the non-uniform reference state p_i^* making this the relative entropy, it is thermodynamically distinct – it is a measure of similarity of two probability density functions, rather than the disorder of a particular probability density function. While presently we focus on applications of this first-order term, which has been used to compare Markov models of conformational ensembles[5] from molecular dynamics simulations but has not been widely applied on a per-residue level, the full derivation presented here establishes a systematic approach to improve our method. At the per-residue level or at the groups-of-residues level , we could improve our method by considering pairs of torsions within a residue or set of residues, etc. Furthermore, application of our method at the pairs-of-residues level or at higher order could identify changes in correlated motions, which could be a promising direction for future research, but is beyond the scope of the present work.

For the local Kullback-Leibler divergence over a set of d.o.f such as a protein residue,

we might want to include second-order mutual divergence terms to obtain more accurate (though not necessarily as stable) results; however, we note that second-order terms require substantially more sampling than first-order terms.[18] The most impact to our visualization of local divergence values would be made by calculating these second-order terms within a single residue's torsions to improve the estimate of the Kullback-Leibler divergence for that residue. Currently, however, we focus on first-order terms, as these are most readily, rapidly, and robustly calculated, and the computational cost scales lineraly with system size.

2.4 Statistical corrections to the Kullback-Leibler divergence

If the "target" ensemble is the same as the equilibrium ensemble, the Kullback-Leibler Divergence will be zero. However, this is not often the case due to sample variability. Furthermore, if applied naively, it might be difficult to extract meaningful population shifts due to different simulation conditions versus artefactual popualtion shifts due to sample variability. In order to improve the signal-to-noise ratio in our calculation the Kullback-Leibler divergence and thereby capture meaningful differences between conformational ensembles, we will calculate the K-L divergence expected from sample variability in the "reference" ensemble and use it for a significance test and to correct the calculated values.

To generate a realistic measure of sample variability, we use a statistical bootstrapping approach. We split the full reference ensembles into *nsims* blocks (usually corresponding to clones of the same system with different random number seeds, or large continuous blocks from long simulations), and take half of the blocks at a time as a surrogate target ensemble and the complementary half as a surrogate reference ensemble. We aggregate the counts for the torsions to construct probability distributions and calculate the K-L divergence between all combinations of surrogate distributions. Any non-zero average K-L divergence between these distributions is a measure of average bias that we can later subtract from the total K-L divergence between the full "reference" ensemble and the full "target" ensemble, when it is

significant. The K-L divergence under the null hypothesis that the average K-L divergence is no greater than that expected from sample variability in the reference ensemble is then given by:

$$KL_1^{H_0} = \binom{nsims}{nsims/2}^{-1} \sum_{blocks}^{\frac{nsims}{2}} \sum_{i}^{nbins} p_i \ln \frac{p_i^S}{p_{i^{S^C}}} \tag{25}$$

where S denotes subsamples and S^C are their complements. To test for statistical significance of the obseved Kullback-Leibler divergence, we use the distribution of these surrogate Kullback-Leibler divergence values to obtain a p-value for the null hypothesis that the average Kullback-Leibler divergence is no greater than that expected from sample variability in the reference ensemble. If this p-value for a particular torsion is less than the significance level (in this case, set at a permissive $\alpha = 0.1$), then the Kullback-Leibler divergence is set to zero; if not, then the average Kullback-Leibler divergence between the surrogate distributions described above is subtracted from the total, in a manner similar to corrections to mutual information[19]:[11]

$$\hat{KL}_1 = KL_1 - KL_1^{H_0} \tag{26}$$

2.5 Truncation of Kullback-Leibler divergence

Given our expansion in Eq. (22), one may wonder why truncation at a particular order might be appropriate, especially as the number of terms at each order increases combinatorially before contracting towards the tail of the expansion. In the analogous configurational entropy expansion,[17] small molecule systems achieved remarkable agreement with entropies from rigorous free energy calculations by only including first and second-order terms in the expansion, with the highly-correlated cyclohexane requiring up to third-order terms. We note that the pairwise mutual divergence between two degrees of freedom is less than or equal to the sum of the corresponding first-order Kullback-Leibler divergence terms. Thus,

for the mutual divergence to be significantly greater than zero, at least one of the constituent degrees of freedom must be statistically significant.

It is important to note that higher-order terms in Eq. (22) capture only changes in distributions missed by lower-order terms. For example, the mutual divergence captures population shifts in pairs of degrees of freedom that are missed by the first-order Kullback-Leibler divergence. The key parameter governing the maximal order needed for convergence of the expansion is the number of coupled independent components or modes (i.e. effective dimensionality) in the most collective motions in the system. A recent study used a novel approach to partition molecular dynamics trajectories into independent subspaces of coupled modes,[20] and found a block-like pattern where groups of pairwise correlated modes had minimal couplings with other blocks of correlated modes in a 100ns simulation of lysozyme. In lysozyme, there was a maximum of 6 modes per block, with most blocks only containing a few modes. Thus, the maximum number of coupled independent components in this study was six, so the Kullback-Leibler Divgence expansion should only require terms up to sixth order. Even though the number of terms at each order might increase, the sparsity of the matrix of mode couplings at second order suggests that a lower fraction of higher-order terms would have significant values. Other studies have also taken advantage of the sparsity of second-order couplings to more efficiently diagonalize the Hamiltonian for the protein [cite Izaguirre2010 and another paper involving block normal modes].

Practically, higher-order mutual divergences will require exponentially more data points sampled to give a robust estimate, as the volume of space increases exponentially with the number of degrees of freedom. Currently, only calculations up to third order might be practical with microseconds of simulation data.[18] We are working on new approaches to push this boundary. Even still, by neglecting high-order terms, we would not be missing relevant phenomena that could be significant given our limited amount of data and current algorithm.

2.6 Jensen-Shannon divergence

The Jensen-Shannon divergence is a slight variation of the Kullback-Leibler divergence, and has the added benefit of treating both "reference" and "target" ensembles symmetrically, albeit at a cost of possibly providing lower signal-to-noise. Furthermore, the Jensen-Shannon divergence is related to thermodynamic length, an asymptotic bound on energy dissipated in a finite-time transformation from one state to another.[21] Since the Kullback-Leibler divergence expansion is general for any "reference" distribution, we can take the new reference distribution to be merely the superposition of the former "reference" distributions, and calculate the Jensen-Shannon divergence as the mean of the Kullback-Leibler divergences between either ensemble and this new refrence distribution:

$$\begin{aligned} JS(x_1,...,x_m||y_1,...,y_m) = &\frac{1}{2}KL(x_1,...,x_m||(x_1,...,x_m)+(y_1,...,y_m)) \\ &+\frac{1}{2}KL(y_1,...,y_m||(x_1,...,x_m)+(y_1,...,y_m)) \end{aligned} \tag{27}$$

2.7 Molecular dynamics simulations

Now that we have detailed the mathematical basis for the Kullback-Leibler divergence expansion and its connection to thermodynamics, we would like to next illustrate the method using examples of previously-published molecular dynamics studies on human interleukin-2 and Talin[11,22] and new molecular dynamics trajectories on a kinase, PDK1. We focus on examples of allosteric communication involving ligand binding and protonation-state changes, where static structures alone were unable to explain distant effects due to a perturbation.

For the new molecular dynamics simulations of PDK1, we prepared the protein and ligand with Maestro's Protein Preparation Wizard (Schrodinger, 2009), with protonation states of histidine and Asn/Gln flips assigned by ProtAssign (Schrodinger, 2009) in the

preparation wizard. Each model was solvated in SPC water [ref] in a cubic simulation box, and Na+ and Cl- ions were added to neutralize the system and then an additional 0.1 M NaCl was added.

The full simulation system was energy-minimized using Desmond[1] in five stages with the following atoms held : 1) all heavy atoms; 2) all backbone (N-Cα-C-O) heavy atoms and side-chain heavy atoms; 3) all heavy atoms; 4) all backbone atoms; 5) no restraints. Minimizations were performed with no less than 100 steps of Steepest Descent minimization followed by L-BFGS optimization after a gradient of 10.0 kcal mol^{-1} Å^{-1} is reached up to a total of 10,000 steps or a gradient of 0.1 kcal mol^{-1} Å^{-1}.

After full minimization of the system, an equilibration was performed. First, the systems were annealed to a temperature of 300K using Langevin dynamics at constant temperature and volume over 50 ps with all heavy atoms restrained. Subsequently, Langevin dynamics at constant temperature and pressure with a target temperature and pressure of 300 K and 1 atm were performed in stages: 1) 50 ps with all heavy atoms restrained with 50 kcal mol^{-1} Å^{-1} force constants; 2) 50 ps with all backbone and side-chain heavy atoms restrained with 50 kcal mol^{-1} Å^{-1} force constants; 3) 150 ps with all heavy atoms restrained with force constants reduced over the course of the simulation from 25 to 5 kcal mol^{-1} Å^{-1}; 4) 100 ps of simulation restraining only the backbone heavy atoms, over which the force constants of the restraints were reduced from 5.0 to 0.0 kcal mol^{-1} Å^{-1}; 5) 100 ps of the unrestrained system. All Langevin dynamics simulations were performed with a 100 ps^{-}1 damping constant.

Then, production runs of 10 ns were performed on each system using the Martyna-Tobias-Klein integrator [ref] with a reference temperature of 300 K and a reference pressure of 1 atm. Snapshots were output every 1.002 ps. The thermostat featured an equilibrium temperature of 300K, a relaxation time of 1 ps, chain length of 3, and update frequency of 2 steps for the system and for the barostat. The barostat featured a relaxation time of 2 ps, a reference pressure of 1 atm, isotropic coupling, and a compressibility of 4.5x10^{-}5 bar^{-}1.

Both the Langevin dynamics and standard molecular dynamics simulations were performed with all bonds involving hydrogens constrained, a 2 fs time step for the bonded and short-range nonbonded interactions and updating of long-range nonbonded interactions every 6 fs using the RESPA multiple time step approach [ref]. Short-range coulombic and van der Waals nonbonded interactions were cutoff at 9.0, and long-range electrostatics were computed using the smooth particule mesh Ewald method. Pairlists were constructed using a distance of 10.5 and a migration interval of 12 ps.

3 Results

3.1 An allosteric small molecule activator of PDK1

PDK1 is a member of the AGC family of kinases, including protein kinases A (PKA), B (AKT), and C (multiple isozymes). Kinases in this family have an allosteric site for activation on the beta-rich N-lobe, where binding of a usually-phosphorylated hydrophobic peptide activates the enzyme. In recent years, small molecules have been discovered that also bind to this same site and promote or inhibit activity. Precisely how these small moleules alter PDK1's activity is not known. Stabilization of the unique active conformation of protein kinase catalytic machinery is one likely mechanism. The mechanism of one previously-reported noncovalent small molecule activator, PS48, was studied using hydrogen-deuterium exchange mass spectrometry experiments to determine which peptide regions of the kinase have amide protons that are protected from exchange with solvent dueterons. In these experiments, amide protons both near the binding site and distant from the binding site (Figure 1) showed protection from solvent exchange, indicating more stable backbone hydrogen bonds and hence less large-scale slow-motion flexibility. Interestingly, some of these protected regions include the DFG-loop (cyan) and activation loop (teal), whose proper positioning is essential for activity. Mutation of Thr226, adjacent to the DFG sequence, to alanine abolishes the ability of PS48 to activate the kinase.

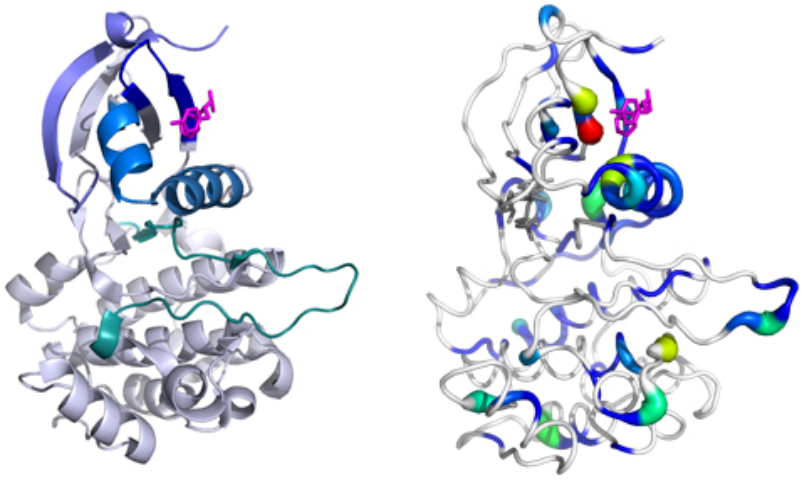


Figure 1: **Kullback-Leibler Divergence highlights PDK1 regions that show protection in hydrogen-deuterium exchange experiments upon addition of an allosteric small molecule activator.** All panels are colored on the same scale. White indicates statistically insignificant divergence, and significant divergence increase from blue to red.(Left) A small molecule activator of PDK1 was previously shown to protect various peptide regions (each shown in a different color) from hydrogen-deuterium exchange. (Right) Kullback-Leibler Divergence values are mapped onto the structure using PyMOL's "b-factor putty" preset. White indicates statistically insignificant divergence, and significant divergence increase from blue to red. Statistically significant Kullback-Leibler divergence values from apo to noncovalent allosteric activator-bound conformational ensembles show population shifts in most of the regions showing protection upon ligand binding. Note that the resolution of the HDX experiments is at the peptide-level, and reflect both fast and slow motions, up to the minute time-scale.

To demonstrate our Kullback-Leibler Divergence method on this pharmaceutically-interesting target where some aspects of the allosteric activation mechanism were previously known, we performed a series of 10ns molecular dynamics simulations on PDK1 with and without PS48 bound (PDB: 3HRF), taking snapshots every 1 ps. We then calculated the Kullback-Leibler Divergence between apo and PS48-bound conformational ensembles from the MD simulations. Our calculations indicated that the compound caused significant population shifts in the torsion angles of residues around the compound's binding site: the αC-helix, the beta strands 145-149 and 154-159, and the αB-helix. Furthermore, there were significant population shifts in torsion angle populations distant from the PS48 binding site, for example the "G" in the DFG-loop, in the activation loop, and

the F-helix. Though the timescale of the simulations is short and the timescale of the hydrogen-deuterium exchange experiments is long, the Kullback-Leibler Divergence values and hydrogen-deuterium exchange results show compound-induced changes in similar regions, except the HDX experiments showed protection an N-terminal peptide containing the β1-strand and Gly-rich loop, while the Kullback-Leibler Divergence only showed a population shift in one residue in the Gly-rich loop (Ser94). Nonetheless, these results serve as a powerful demonstration of how our method can identify potential allosteric effects of ligand binding or mutation.

3.2 Allosteric inhibition by lysine acetylation in mitochondrial 3-hydroxy-3-methylglutaryl CoA synthase 2

When eukaryotes transition from the fed to fasted state, carbohydrate utilization and fatty acid synthesis cease and fatty acid oxidation and ketogenesis are induced; in diabetes, similar but more pronounced metabolic changes occur.[23] Acetyl-CoA generated from fatty acid oxidation is converted to β-hydroxybutyrate in the mitchondria. Mitochondrial 3-hydroxy-3-methylglutaryl CoA synthase 2 (HMGCS2) is the rate-limiting enzyme in the synthesis of β-hydroxybutyrate and is normally acetylated at K310, K447, and K473, which inhibit its activity. The sirtuin SIRT3 deacetylates HMGCS2 to increase its activity and the cell's production of β-hydroxybutyrate.

Molecular dynamics simulations (over 11-20ns each) on HMGCS2 in the deacetylated, activated form and with acetylations at various lysine residues each showed that acetylation of specific lysines and not nearby bystander controls produced significant conformational and dynamical changes in HMGCS2.[24] Specifically, K310 near the acetyl-CoA binding site forms ion pairs with both of the phosphates of the acetyl-CoA in the deacetylated form of HMGC2, but these interactions are broken by acetylation of K310, which removes the positive charge. Large changes were observed in the conformation of the helix containing residues 350-367, in which this helix moves away from the acetyl-CoA, while minimal

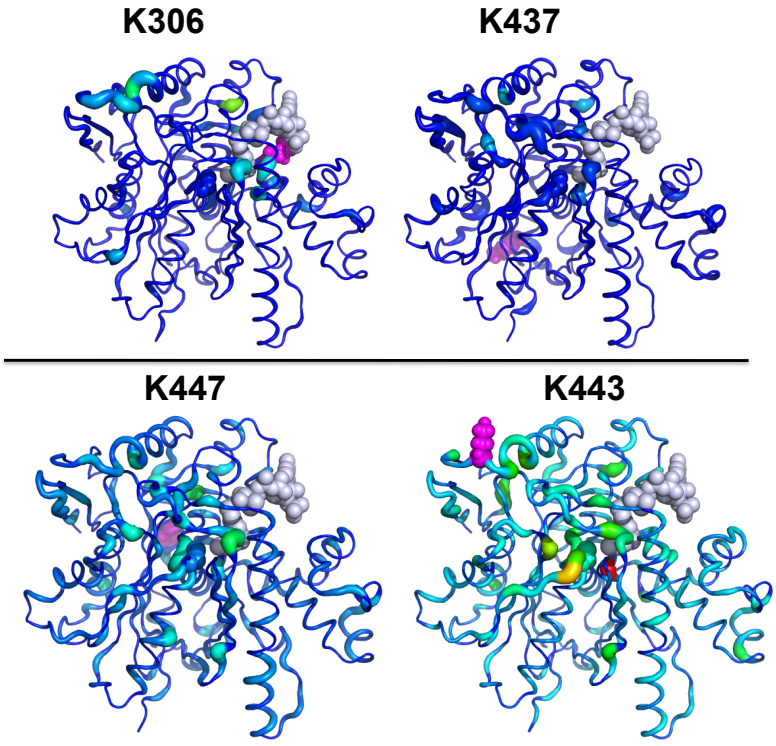


Figure 2: **Kullback-Leibler divergences show position-specific effects of lysine acetylation in HMGCS2.** Kullback-Leibler divergences between deacetylated and acetylated HMGCS2 conformational ensembles for different lysine acetylations are shown relative to one another (all are on the same scale). The lysine acetylated in each case is shown in purple spheres. (Top) Though lysines 306 and 437 are proximal to lysines that are deacetylated to activate this enzyme, acetylation of lysine 306 or 437 does not yield significant changes in structure and dynamics near the acetyl-CoA binding site (gray spheres) as assessed by the Kullback-Leibler Divergence. (Bottom) In contrast to these negative controls, acetylation at lysine 447 or 443 causes substantial divergences proximal to the active site and the tail of the acetyl-CoA, and some background of divergences across the whole protein.

changes were observed in the helix containing K310. Importantly, both K447 and 4K73 are distant from the acetyl-CoA bound at the active site (Figure 2 , gray spheres). Thus, the effects of these acetylations at K447 and K443 are allosteric in nature since they inhibit activity over a distance.

Interestingly, some distant lysines but not others also showed significant changes in structure and dynamics near the active site upon acetylation. Acetylation at K473 or K447

showed larger fluctuations in loop 1 (residues 242-251) and a shift in the average position of loop 2 (residues 131-140), effects that propagated to the end of the active site where the acetyl-CoA binds, and altered the positions of catalytic residues His301 and Cys166.

As negative controls, two lysine residues whose acetylations do not inhibit activity were studied. In contrast to the other lysines, these did not show similar marked changes in structure and dynamics at the active site.

Presently, we wondered what significant population shifts upon lysine acetylation would be identified by our Kullback-Leibler divergence method. The Kullback-Leibler divergence results showed a marked difference between the control lysine acetylations and those that inhibited enzyme activity (acetylations at lysines K447 and K473) (Figure 2). The control lysine acetlyations did not show the pronounced divergences seen with the natural inhibitory lysine acetylations at positions 447 and 473. Importantly, acetylation at K447 or K443 causes significant population shifts in the catalytic residues: in the loop from position 163 to 168 containing the active site cysteine (C166), and in His301. Furthermore, these acetylations both cause substantial population shifts in a turn (239-241) near the acetyl-CoA tail, in Lys83 at the other end of the acetyl-CoA near the nucleotide ring, and in a helix-turn containing residues 380-385, which buttress the loop containing the active site cysteine (residues 163-168).

In summary, the Kullback-Leibler divergence highlighted new residues showing significant perturbations upon acetylation that were missed by previous structural and r.m.s. fluctuation analyses.

3.3 Communication between small molecule binding sites in in interleukin-2

Interleukin-2 (IL-2) is a small cytokine that has been studied extensively as a model system for small molecules inhibiting protein-protein interactions. Additionally, we found it to be a useful model system for studying small-molecule cooperativity.[11] Binding of ligand

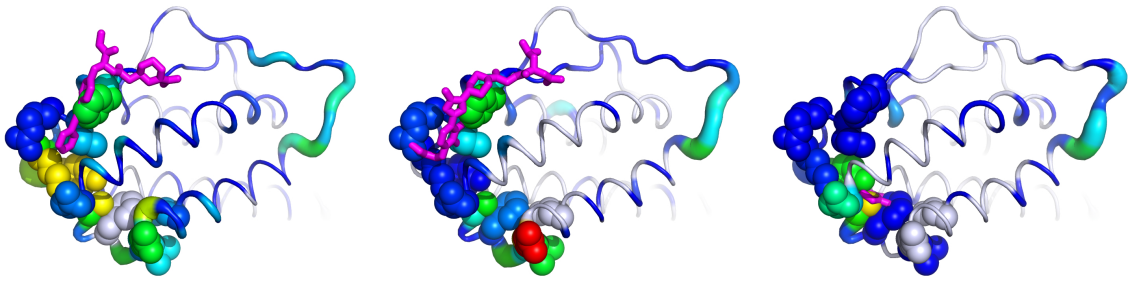


Figure 3: **Kullback-Leibler divergences Between Apo and Ligand-Bound IL-2 Ensembles Show Differential Allosteric Effects** The Kullback-Leibler divergence between the apo IL-2 ensemble and various ligand-bound ensembles was calculated and mapped onto the apo structure. All panels are on the same scale, and ligands are superimposed for reference. Residues shown in spheres indicate a previously-reported allosteric network that is thought to couple the two small-molecule binding sites through significant correlated motions[11] (Left) IL-2 with a micromolar ligand at the IL-2Rα site. (Center) IL-2 with an optimized nanomolar inhibitor at the IL-2Rα site featuring receptor-mimicking electrostatics.[25] (Right) IL-2 with an allosteric small molecule fragment at a cryptic site. Both of the IL-2Rα-competitive ligands at left and center cause substantial population shifts in the long, flexbile loop proximal to the allosteric fragment at right. However, the smaller inhibitor (left) but not the larger one (center) shows a substantial population shift on the helix-turn-helix at the backside of the fragment's binding site. Furthermore, the allosteric fragment (right) gives population shifts not only at its binding site but also in the helix behind the IL-2Rα site ligands, as would be expected from thermodynamic linkage – indicating that the sampling was sufficient to observe some allosteric effect.

to one site in IL-2 facilitated binding of a small molecule fragment to a cryptic, transient pocket, which is gated by a loop on the opposite face from the four-helix bundle.[26] While X-ray structures were unable to show how binding to one side affected binding at the other side, our previously published molecular dynamics study offered novel insights. We identified putative allosteric network residues using statistically-significant correlated motions quantified by mutual information in torsion space.[11] Our mutual information analysis suggested that a solvent-exposed polar network and a greasy hydrophobic core couple the two sites, so that binding of one inhibitor causes a change in dynamics at the other site.

We hypothesized that the Kullback-Leibler divergence analysis would show significant population shifts in torsion angle distributions of residues implicated in the allosteric network by our previous mutual information analysis. We caluclated the local, residue-

by-residue Kullback-Leibler divergence between apo and ligand-bound conformational ensembles from five ten-ns molecular dynamics simulations (Figure 3), using an additional finite system correction given in the Appendix. Ligand-bound conformational ensembles shown here include a micromolar IL-2Rα-competitive inhibitor, a nanomolar IL-2Rα-competitive inhibitor, and a weak fragment that only would bind in the presence of the micromolar inhibitor at the IL-2Rα-competitive site; cooperative binding of this fragment with the nanomolar inhibitor was not tested.

Globally, all ligands cause some population shifts in the flexible loop between helices three and four. As this loop is somewhat removed from the small-molecule binding sites, yet structurally connected through helix two, it is not clear whether population shifts here are functionally relevant, especially since residues that are more flexible are, in general, more likely to show substantial population shifts, even after the corrections above, simply because they have more entropy. Likewise, the population shifts right before helix one merely represent "wagging" of the tail of IL-2, and it is not surprising that this effect is large in the left and right panels where there are greater population shifts in the turn after helix one than in the center panel, since any torque on one end of a helix could easily be propagated to the other end due to the semi-rigid nature of the helix.

To test our hypothesis that residues implicated in the allosteric network show population shifts upon binding ligand at either site, we show part of the allosteric network in spheres (Figure 3). As can be seen, population shifts are seen along this structurally-contiguous network of residues upon binding of ligand at either site. In particular, Tyr31 seems to be an important mediator of allostery, as it is highlighted in both the left and right panels, whose respective ligands bind with positive cooperativity. Although this tyrosine is not directly contacting the IL-2Rα-competitive inhibitor, the methionine in cyan located above it does contact the IL-2Rα-competitive inhibitor, and also contacts the allosteric fragment in the simulations. There are a number of polar residues proximal to Tyr31 that also show population shifts upon binding of allosteric fragment but that do contact the competitive in-

hibitor. Thus, nearly all the residues implicated in the allosteric network linking compound binding sites in our previous study[11] show statistically significant population shifts upon binding either IL-2Rα-competitive inhibitor or allosteric small-molecule fragment.

3.4 Talin's pH Sensor

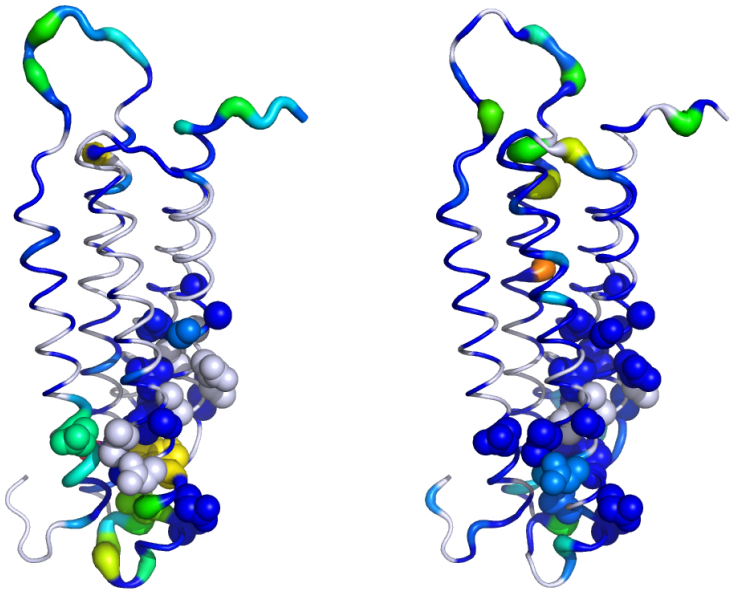


Figure 4: **Wildtype and pH-sensor mutant Talin show different population shifts upon pH change.** Kullback-Leibler divergences between the pH 8.0 ensemble and the pH 6.0 ensemble for Talin are shown for wildtype (left) and H2148F Talin (right). Actin-binding site residues shown in[22] are represented as spheres. These divergences highlight the region proximal to the pH sensor (at the top of the structure, with His2418 shown in sticks) and the actin-binding site (spheres). Some significant divergences along the helices coupling these two sites suggest that subtle motions of these helical residues – either motions of the side-chains or rigid-body motions of the helices themselves – might contribute to coupling between the two sites. Wildtype and H2418F Talin show different patterns of population shifts in these actin binding site residues. In particular, the 2386-2389 region and Glu2308 show strikingly larger population shifts in wildtype than in H2418F Talin, and as such might be prime residues to target for mutagenesis, where differential effects in wildtype and H2418F backgrounds on pH-dependent F-actin binding would be predicted.

Talin is an integrin-associated focal adhesion protein that plays a key role in focal adhesion remodeling, the rate-limiting step in migration of adherent cells. Talin's binding to actin filaments is thought to act as a "clutch" to regulate focal adhesion turnover; this bind-

ing is regulated by pH, so that higher-affinity actin binding occurs at lower pH, while lower-affinity actin binding occurs at higher pH(8-10). In motile cells, which have an increased intracellular pH with respect to nonmotile cells, focal adhesion turnover is increased due to the lower affinity of talin for actin. To investigate the mechanism by which pH change alters talin's structure, dynamics, and actin-binding ability, constant-pH molecular dynamics simulations of the I/LWEQ domain of talin1 (the likely focal-adhesion-associated talin, RCSB PDB code 2JSW) without the C-terminal dimerization domain (PDB 2JSW) were performed at pH 8.0 and pH 6.0 for 10ns.[22] As the I/LWEQ domain of talin has a single histidine residue, this histidine and nearby acidic residues with upshifted predicted pKa values were hypothesized to constitute the pH sensor, and their protonation states were sampled during the constant-pH simulation. At pH 6.0, this His2418 showed 73% protonation, while at pH 8.0, it only showed 2% protonation. Significant differences in structure and dynamics were observed near the actin binding site on the other side of the five-helix bundle from the pH sensor, especially in the loop connecting helices 2 and 3 and the helical regions near this loop. These predictions were qualitatively confirmed by NMR amide chemical shift perturbation analysis comparing talin at pH 6.0 to talin at pH 8.0.

On the basis of these calculations, this His2418 was mutated to a Phe, and in constant-pH MD simulations showed structural and dynamical differences at pH 6.0 and pH 8.0 that were different from wildtype, and in NMR showed altered chemical shift perturbations in NMR titrations. Moreover, this mutant showed decreased F-actin binding *in vitro* and altered focal adhesion turnover in migrating cells.

In this work, we apply our Kullback-Leibler divergence method to compare the conformational ensembles at these two different pH values for both wildtype at H2418F Talin, using the pH 8.0 as the reference ensemble since it is generally more flexible. We found (Figure 4) that pH change caused substantial population shifts distant from the pH sensor in both cases, although these were more pronounced in the wildtype than in the mutant. Furthermore, the wildtype typically showed larger population shifts than the mutant in

residues in the actin-binding site (shown in spheres). Interestingly, the bottom of helices 1 and 3 in the wildtype show substantial Kullback-Leibler divergences; in the NMR titration experiments,[22] both these regions showed either chemical shift changes or line broadening.

Given the population shifts in the putative pH sensor and actin binding site upon pH change in wildtype and H2418F Talin, we wondered how protonation state changes in the pH sensor are propagated to the actin binding site. We observed subtle yet significant population shifts in the helices connecting the sites, which are qualitiatively consistent with NMR chemical shift perturbations which generally did not show large chemical shift perturbations in these residues, except in amides proximal to the pH sensor. We suspect that a combination of direct electrostatics and subtle rigid-body motions of the helices are responsible for coupling the pH sensor to the actin binding site.

4 Discussion

To address the challenge of comparing conformational ensembles, we have developed a novel approach grounded in thermodynamics, information theory, and statistics. We use the Kullback-Leibler divergence to quantify changes in torsion angle probability distributions, which reflect biologically-relevant processes such as rotamer flips, changes in local secondary structure, etc. Inspired by previous work, we developed the Kullback-Leibler divergence Expansion, which provides an approximation the Kullback-Leibler divergence of whole molecules (proteins in this work) in terms of marginal probability density functions involving far fewer degrees of freedom; in this work, we have found that even the first-order terms can give considerable insight into which residues are most affected by perturbations such as ligand binding or pH change (i.e. proton binding).

The fact that we observe many significant divergences in surface polar residues after correcting for sample variability suggests that these residues may play a role in propagating binding at one site to a change in structure and/or dynamics at another site. As these surface

polar side-chains are often not part of evolutionarily-conserved networks,[27] their ability to propagate these kinds of perturbations lies in the sum effects of multiple residues working in a parallel fashion and is less related to the amino acid identity and more related to their physical properties of containing a charge or strong, sticky dipole that can flop around with the help of a flexible linker. A similar role for correlated protein side chain motions in mediating long-range couplings was suggested by DuBay and Geissler.[28]

There are several algorithmic improvements that could be made to our approach. Multiple calculations at different histogram bin sizes could be used, and an optimal histogram size chosen for each degree of freedom; such an approach has been shown to lead to more accurate entropy calculations.[29] A k-Nearest Neighbor (KNN) approach could also be used to calculate the Kullback-Leibler divergence.[30] Our residue-level analysis of the "local" could be augmented by including second-order terms (i.e. the mutual divergence) within residues, which could benefit from adaptive partitioning, as in our previous work on mutual infromation.[11]

Acknowledgement

We would like to thank Ken Dill for inspiring this work, and to Steve Presse and Michael Gilson for helpful advice and comments on the manuscript. This work was supported by NSF Teragrid Allocation ..... and the Texas Advanced Computing Center (TACC). CLM was supported by a PhRMA Foundation Informatics Fellowship. This project was supported by NIH grants R01(caspase-1) and R01#.

5 Appendix

5.1 Robust Histogram Estimate of Kullback-Leibler divergence using Renyi Generalized divergence

To obtain a finite-sample size correction to the Kullback-Leibler divergence, we will adapt the derivation presented by Grassberger.[31] Though this was only used in the interleukin-2 system presented here, it is provided as an option in the program, and is given here for completeness and for possible inclusion in other code packages. We will consider the Kullback-Leibler divergence as a limit of the Renyi Generalized Divegence,

$$KL_i = \lim_{\alpha \to 1} D_\alpha(P||P^*) \tag{28}$$

where

$$D_\alpha(P||P^*) = \frac{1}{(\alpha - 1)} \ln(\sum_{i=1}^{n} \frac{p_i^\alpha}{(p_i^*)^{\alpha-1}}) \tag{29}$$

For finite sample sizes there will be some uncertainty in the p_i. Considering the actual histogram counts, we write:

$$p_i^\alpha = \left(\frac{< n_i >}{N}\right)^\alpha, (p_i^*)^\alpha = \left(\frac{< n_i^* >}{N}\right)^\alpha \tag{30}$$

To obtain $< n >^\alpha$, we assume a Poisson distribution for n_i in successive realizations (i.e. assuming we are using a fine enough discretization such that $p_i << 1$). For a positive integer α, we would then have

$$< n >^\alpha = \left\langle \frac{n!}{(n-\alpha)!} \right\rangle \tag{31}$$

However, as we consider the limit as α approaches 1, we need a continuous analog using Γ functions. Grassberger found an asymptotic expansion for $< n_i >^\alpha$ and showed

that two terms gave numerically robust results for Shannon etropies.

$$< n >^{\alpha} = \frac{\Gamma(n+1)}{\Gamma(n-a+1)} - \frac{(-1)^n \Gamma(a+1)\sin(\pi a)}{\pi(n+1)} \tag{32}$$

This same approximation is used in our previously-published MutInf method.[11] Then, we use this expression for $< n >$ and evaluate the Renyi Generalized divergence in the $\alpha \to 1$ limit to give us the Kullback-Leibler divergence. Invoking L'Hopital's Rule, we obtain:

$$\lim_{\alpha \to 1} D_{\alpha}(P||Q) = \lim_{\alpha \to 1} \Big(\sum_{i=1}^{nbins} \frac{Nf\Big(n_i^*, \alpha - 1\Big)}{f\Big(n_i, \alpha\Big)} \frac{\partial}{\partial \alpha} \sum_{i=1}^{nbins} \frac{f\Big(n_i, \alpha\Big)}{Nf\Big(n_i^*, \alpha - 1\Big)} \Big) \tag{33}$$

$$\lim_{\alpha \to 1} D_{\alpha}(P||Q) = \sum_{i}^{nbins} \frac{\Psi(n_i) n_i n_i^* + (-1)^{n_i} n_i^* + (-1)^{n_i^*} n_i n_i^* - \Psi(n_i^*) n_i n_i^* - n_i}{n_i^*} \tag{34}$$

However, this expression is not numerically robust in practice, so we truncate the expression for $< n >^{alpha}$ at the first term:

$$< n >^{\alpha} = \frac{\Gamma(n+1)}{\Gamma(n-a+1)} \tag{35}$$

which then provides a more robust estimate for $D_{\alpha}(P||Q)$:

$$KL_1 = \lim_{\alpha \to 1} D_{\alpha}(P||Q) = \sum_{i}^{nbins} \frac{n_i}{N} \Big(\Psi(n_i) - \Psi(n_i^*) - \frac{1}{n_i^*} \Big) \tag{36}$$

Using a series approximation of the digamma function, $\Psi(x) \approx \ln(x) - \frac{1}{2x}$, it can be readily seen that the regular Kullback-Leibler divergence is recovered along with a correction term that decreases in size as histogram counts increase.

References

(1) Bowers, K. J.; Chow, E.; Huafeng, X.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Yibing, S.; Shaw, D. E. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *SC 2006 Conference, Proceedings of the ACM/IEEE*; 2006.

(2) Shaw, D. E. *et al. SIGARCH Comput. Archit. News* **2007,** *35,* 1–12.

(3) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. *Journal of Computational Chemistry* **2009,** *30,* 864-872 1096-987X.

(4) Lange, O. F.; Lakomek, N.-A.; FarÃĺs, C.; SchrÃűder, G. F.; Walter, K. F. A.; Becker, S.; Meiler, J.; GrubmÃijller, H.; Griesinger, C.; de Groot, B. L. *Science* **2008,** *320,* 1471-1475.

(5) Morcos, F.; Chatterjee, S.; McClendon, C. L.; Brenner, P. R.; LÃşpez-RendÃşn, R.; Zintsmaster, J.; Ercsey-Ravasz, M.; Sweet, C. R.; Jacobson, M. P.; Peng, J. W.; Izaguirre, J. A. *PLoS Comput Biol* **2010,** *6,* e1001015.

(6) Koyama, Y. M.; Kobayashi, T. J.; Tomoda, S.; Ueda, H. R. *Physical Review E* **2008,** *78,* 046702 Copyright (C) 2011 The American Physical Society Please report any problems to prola@aps.org PRE.

(7) Ramanathan, A.; Savol, A. J.; Langmead, C. J.; Agarwal, P. K.; Chennubhotla, C. S. *PLoS ONE* **2010,** *6,* e15827.

(8) Lange, O. F.; GrubmÃijller, H. *Proteins: Structure, Function, and Bioinformatics* **2008,** *70,* 1294-1312 1097-0134.

(9) Bradley, M. J.; Chivers, P. T.; Baker, N. A. *Journal of Molecular Biology* **2008,** *378,* 1155-1173 0022-2836 doi: DOI: 10.1016/j.jmb.2008.03.010.

(10) Lange, O. F.; GrubmÃijller, H. *Proteins: Structure, Function, and Bioinformatics* **2006,** *62,* 1053-1061 10.1002/prot.20784.

(11) McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P. *Journal of Chemical Theory and Computation* **2009,** *5,* 2486-2502 doi: 10.1021/ct9001812 1549-9618 doi: 10.1021/ct9001812.

(12) Qian, H. *Physical Review E* **2001,** *63,* 042103 Copyright (C) 2011 The American Physical Society Please report any problems to prola@aps.org PRE.

(13) Wall, M. E. *AIP Conference Proceedings* **2006,** *851,* 16-33.

(14) Ming, D.; Wall, M. E. *Proteins: Structure, Function, and Bioinformatics* **2005,** *59,* 697-707 10.1002/prot.20440.

(15) Ming, D.; Cohn, J.; Wall, M. *BMC Structural Biology* **2008,** *8,* 5.

(16) Matsuda, H. *Physical Review E* **2000,** *62,* 3096 Copyright (C) 2009 The American Physical Society Please report any problems to prola@aps.org PRE.

(17) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. *The Journal of Chemical Physics* **2007,** *127,* 024107-16.

(18) Killian, B. J.; Kravitz, J. Y.; Somani, S.; Dasgupta, P.; Pang, Y.-P.; Gilson, M. K. *Journal of Molecular Biology* **2009,** *389,* 315-335 0022-2836 doi: DOI: 10.1016/j.jmb.2009.04.003.

(19) Karchin, R.; Kelly, L.; Sali, A. *Pac Symp Biocomput.* **2005,** 397-408.

(20) Sakuraba, S.; Joti, Y.; Kitao, A. *The Journal of Chemical Physics* **2010,** *133,* 185102.

(21) Crooks, G. E. *Physical Review Letters* **2007,** *99,* 100602 Copyright (C) 2011 The American Physical Society Please report any problems to prola@aps.org PRL.

(22) Srivastava, J.; Barreiro, G.; Groscurth, S.; Gingras, A. R.; Goult, B. T.; Critchley, D. R.; Kelly, M. J. S.; Jacobson, M. P.; Barber, D. L. "Structural model and functional significance of pH-dependent talinÃćâĆňâĂIJactin binding for focal adhesion remodeling", 2008 10.1073/pnas.0805163105.

(23) McGarry, J. D.; Foster, D. W. *Annual Review of Biochemistry* **1980,** *49,* 395-420.

(24) Shimazu, T.; Hirschey, M. D.; Hua, L.; Dittenhafer-Reed, K. E.; Schwer, B.; Lombard, D. B.; Li, Y.; Bunkenborg, J.; Alt, F. W.; Denu, J. M.; Jacobson, M. P.; Verdin, E. *Cell Metabolism* **2010,** *12,* 654 - 661.

(25) Thanos, C. D.; DeLano, W. L.; Wells, J. A. *Proceedings of the National Academy of Sciences* **2006,** *103,* 15422-15427 10.1073/pnas.0607058103.

(26) Hyde, J.; Braisted, A. C.; Randal, M.; Arkin, M. R. *Biochemistry* **2003,** *42,* 6475-83 0006-2960 (Print) Journal Article.

(27) Halabi, N.; Rivoire, O.; Leibler, S.; Ranganathan, R. *Cell* **2009,** *138,* 774-86 1097-4172 (Electronic) 0092-8674 (Linking) Journal Article Research Support, Non-U.S. Gov't.

(28) DuBay, K. H.; Geissler, P. L. *Journal of Molecular Biology* **2009,** *391,* 484-497 0022-2836 doi: DOI: 10.1016/j.jmb.2009.05.068.

(29) Baron, R.; HuÃŇËĘnenberger, P. H.; McCammon, J. A. *Journal of Chemical Theory and Computation* **2009,** *5,* 3150-3160 doi: 10.1021/ct900373z 1549-9618 doi: 10.1021/ct900373z.

(30) Piro, P.; Anthoine, S.; Debreuve, E.; Barlaud, M. Image retrieval via Kullback-Leibler divergence of patches of multiscale coefficients in the KNN framework. In

Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop on; 2008.

(31) Grassberger, P. *Physics Letters A* **1988,** *128,* 369-373.

Chapter 4. Correlated motions in the Pin1 WW-domain couple substrate docking at the WW domain to the catalytic domain interface

Introduction

Proteins in their native state can adopt a plethora of shapes, or conformations; this conformational plasticity is critical for regulation and function in many systems. However, it has remained difficult to determine what these different conformations look like at the atomic level. Protein-protein interactions are often mediated by flexible loops that experience conformational dynamics on the microsecond to millisecond time scales. NMR relaxation studies can map these dynamics. We apply our MutInf method and molecular dynamics simulations to study correlated motions in the apo Pin1-WW domain at an atomistic level, for which NMR has revealed conformational dynamics of a flexible loop in the millisecond range.

Methods

We use the "MutInf" method[1] to quantify correlations between residues' conformations from equilibrium molecular dynamics simulations performed on the Pin-1 WW domain[2], representing over a microsecond of data, referred to as the "APO Extended 1" ensemble in the paper. Briefly, this method calculates the mutual information between pairs of residues, applies statistical corrections and tests of significance for the mutual information values, and then clusters the matrix of mutual information between residues to identify groups of residues showing similar patterns of correlations.

We followed the same protocol as the previously published method[1], with the following modifica- tions. We split the APO Extended 1 ensemble[2] into six equal-sized

segments, after removing the 10% of snapshots where the WW domain's heavy atoms were within 5Å of those of a periodic image. Also, we added a statistical bootstrapping approach to the protocol as an additional statistical filter to require the reproducibility of a correlation between a pair of torsions. We split the full trajectory into six time segments, and take four out of six segments at a time as a sample ensemble, or "block", from which we aggregate the histogram counts for the two torsions and calculate the mutual information for each sample ensemble. This bootstrapping approach is similar to block-averaging; our "blocks" are composed of multiple, not-necessarily contiguous, and large time segments. The Wilcoxon signed-rank test is used to test the null hypothesis that the average (corrected) mutual information is zero against a one-sided alternative. If the p-value is less than or equal to $\alpha = 0.01$, the average of the mutual information values for the "blocks" is reported, otherwise it is zeroed.

To calculate the mutual information between each pair of residues, we take the sum of the mutual information over all pairs of φ, ψ, and χ dihedral angles, each pair comprising one angle from each residue. We then clustered the mutual information between pairs of residues using the "heatmap" function in the R statistical package with a Euclidean distance metric.

Additionally, we calculated the mutual information between residues' C-α coordinates using the same procedure as above, using x, y, and z coordinates in place of φ, ψ, and χ. Rotational and translational motion was removed prior to analysis by a rotational/translational fit involving only C-α atoms.

We analyze simulation data using a thermodynamics-based mutual information metric to find pairs of residues with correlated conformations in the conformational

ensemble. In a conformational ensemble, it does not matter whether one residue moves, then another, so we can use correlated conformations and correlated motions interchangeably, as no time offsets are used. This approach provides an analysis of correlated motions that is complementary to NMR Rex measurements. We find that Loop 1 residues form a cluster that is correlated with key residues that lie in the catalytic domain interface[3]. These correlations are mediated by some residues in the β2-β3 loop (Loop 2), providing mechanistic insight into how Loop 1 dynamics may affect function of Pin1.

To identify correlated motions beyond those that were studied by NMR relaxation[4] and a comparison with chemical shifts calculated from the conformational ensemble[2], we use the "MutInf" method to quantify correlations between residues' conformations from equilibrium simulations. Briefly, this method calculates the mutual information between pairs of residues using backbone and side chain torsions and applies statistical corrections and tests of significance for the mutual information values. It then clusters the matrix of mutual information between residues to identify groups of residues showing similar patterns of correlations. We followed the same protocol as the previously published method[1], with modifications described above. Most notably, we filtered out snapshots in which the WW domain's heavy atoms were within 5Å of those a periodic image. This was needed because our simulation box was rather small.

Results

Correlated Motions

Correlated protein motions are of great interest as a possible mechanism for intra-protein communication. The NMR studies[4] examined the motions of backbone NHs of

Loop 1. The NH motions are only a subset of the Loop 1 degrees of freedom. Thus, while the NMR data may reflect correlated motion, it may not supply enough information for their characterization. Computational approaches can bridge these information gaps. Accordingly, we investigated the possibility of correlated motions between the Loop 1 residues and other residues that would be invisible to the NMR experiments focused on μs-ms motions. We used a previously reported mutual information method, "MutInf", to look for statistically significant correlated torsional motions in an unbiased way, independently of the MSM analysis. This entailed generating a conformational ensemble of the apo Pin1-WW domain via molecular dynamics simulations, and then identifying pairs of residues showing statistically significant correlated motions. Critically, this approach: (i) makes no quasi-harmonic assumptions about motions relative to an "average" structure; (ii) filters out insignificant correlations; (iii) and quantifies correlated motions in thermodynamic units. Additionally, we applied our approach to calculate the mutual information between Pin1-WW domain's Cα Cartesian coordinates.

Substrate binding in Pin1 WW results in information relay from Loop 1 to the catalytic site of Pin1 via domain interface residues in Loop 2

To identify groups of residues showing similar magnitudes of correlation with other residues, we hierarchically clustered our matrix of mutual information between residues' torsions. The cluster with the strongest correlated motions (shown in red in Figure 1B) consists chiefly of Loop 2 residues. In full-length Pin1, these residues lie at the interface between the WW domain and its flexibly tethered isomerase domain. Figure 1A further shows substantial correlation between residues in this red cluster, a blue cluster containing four residues within the substrate-binding Loop 1, a yellow cluster

consisting of mostly hydrophobic core residues proximal to Loop 2, and a fourth magenta cluster containing mostly residues within Loop 1 (Figure 2). Notably, the magenta cluster contains many basic residues that form salt bridges with the phosphorylated substrate in a holo structure. Thus, substrate binding would not only perturb motions of substrate binding Loop 1, but also those of the WW-catalytic domain interface Loop 2. Focusing on the two tryptophans in the WW domain (from which the WW domain derives its name), we see that Trp29's statistically significant coupling with Trp6 does not appear to be mediated by any particular proximal shared residue (i.e. not through Gln28); rather, these two functional residues are coupled indirectly through the intervening Loop 1 (red cluster). This is most clearly seen by comparing the representative structures of macrostates 21 and 22 (Video S1). Combining these results with previous NMR studies suggests that Loop 1 can relay information about substrate binding to the catalytic site via the domain interface residues in Loop 2. We also analyzed the mutual information between Cα Cartesian coordinates after removing rotational/translational motions, and found the C-terminal part of Loop 2 highly correlated to the rest of the protein (Figure 1C). This Cartesian analysis complements the torsion-space analysis in Figre 1A. NMR studies implicated methyl-bearing residues in Loop 2 (Ile-23 and Thr-24 in the red cluster) in a dynamic network of residues that show perturbed dynamics upon substrate binding[5].

Figure 1. Correlated motions couple the catalytic domain interface to the substrate-binding loop of Pin1's WW domain. The WW domain is shown in cartoon and sticks, the catalytic domain as a surface, and the substrate in spheres. The structure shown is from PDB entry 1F8A. Only the WW domain was simulated; the catalytic domain is only shown for reference. (A) Hierarchical clustering of the mutual information between residues' torsions identifies several functionally important groups of residues. (B) Most residues in the red cluster lie in the catalytic domain interface and are correlated with residues in magenta cluster, which includes a number of key substrate-binding residues.

All residues exhibiting slow motions in NMR experiments are in either the red or magenta clusters. (C) Mutual information between Cα atoms complements torsional analysis and importantly captures correlated motions of secondary structure elements, highlighting correlated motions between the first β-strand (residues 7–9) and Loop 1 (residues 10–16), between the first β-strand and the second β-strand (residues 17–21), and between the C-terminal part of Loop 2 and the beginning of the third β-strand (residues 23–26) and the rest of the protein.

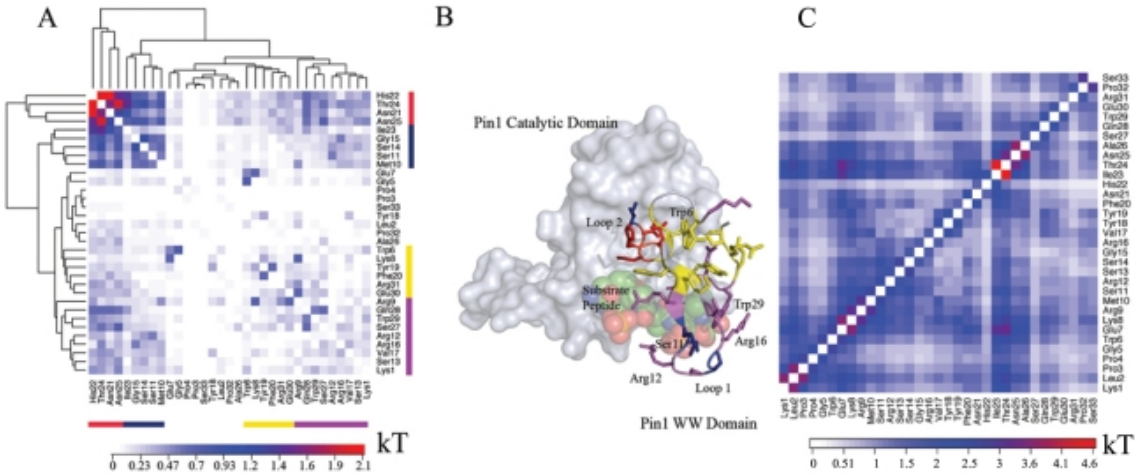


Figure 2. Superposition of representative structures for all 40 macrostates shows diverse conformations of Loop 1.

Other NMR studies showed coupled rotational tumbling of the two Pin1 domains in the presence but not the absence of substrate peptides of particular sequences[6]. Recently,

peptides with two Pin1 binding sites separated by rigid linkers were used to ask whether Pin1 displays cooperative binding[7]. These studies found that while binding at one site facilitated binding at the other through bivalency, no significant cooperativity was observed. However, these studies did not rule out a role for substrate binding to the WW domain in substrate turnover at the active site. As correlated motions are necessary but not sufficient for allosteric crosstalk between distant sites, the functional role of this dynamic network that connects Pin1's active site to its WW-domain's substrate-binding site remains unclear and merits further study.

References

1. McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P., Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles. *J Chem Theory Comput* **2009,** *5* (9), 2486-2502.

2. Morcos, F.; Chatterjee, S.; McClendon, C. L.; Brenner, P. R.; Lopez-Rendon, R.; Zintsmaster, J.; Ercsey-Ravasz, M.; Sweet, C. R.; Jacobson, M. P.; Peng, J. W.; Izaguirre, J. A., Modeling conformational ensembles of slow functional motions in Pin1-WW. *PLoS Comput Biol* **2010,** *6* (12), e1001015.

3. Verdecia, M. A.; Bowman, M. E.; Lu, K. P.; Hunter, T.; Noel, J. P., Structural basis for phosphoserine-proline recognition by group IV WW domains. *Nat Struct Mol Biol* **2000,** *7* (8), 639-643.

4. Namanja, A. T.; Peng, T.; Zintsmaster, J. S.; Elson, A. C.; Shakour, M. G.; Peng, J. W., Substrate recognition reduces side-chain flexibility for conserved hydrophobic residues in human Pin1. *Structure* **2007,** *15* (3), 313-27.

5. Namanja, A. T.; Peng, T.; Zintsmaster, J. S.; Elson, A. C.; Shakour, M. G.; Peng, J. W., Substrate Recognition Reduces Side-Chain Flexibility for Conserved Hydrophobic Residues in Human Pin1. **2007,** *15* (3), 313-327.

6. Jacobs, D. M.; Saxena, K.; Vogtherr, M.; Bernado, P.; Pons, M.; Fiebig, K. M., Peptide Binding Induces Large Scale Changes in Inter-domain Mobility in Human Pin1. *J. Biol. Chem.* **2003,** *278* (28), 26174-26182.

7. Daum, S.; Lucke, C.; Wildemann, D.; Schiene-Fischer, C., On the benefit of bivalency in peptide ligand/pin1 interactions. *J Mol Biol* **2007,** *374* (1), 147-61.

Chapter 5. Disulfide trapping data provides a refined model of the PIFtide-PDK1 complex

Introduction

I developed a novel procedure of using disulfide trapping data on cysteine-mutant peptides and a cysteine-mutant kinase, PDK1, to identify constraints and restraints in the complex of PDK1 with a natural activator, PIFtide. I then used these to construct a model of the wildtype PIFtide-PDK1 complex that was consistent with the disulfide trapping data.

Results

Homology models of PIFtide bound to PDK1, with and without disulfide crosslinking-based restraints

To construct models of PIFtide bound to PDK1, a structure of PDK1 with a noncovalent small molecule bound in the hydrophobic motif pocket on the N-lobe was used to represent an active, "holo" conformation of the hydrophobic motif pocket (PDB: 3HRF). Then, a PIFtide variant bound to AKT (PDB: 1O6L) was superimposed by alignment in PyMOL using PDK1 residues 75-161 (3HRF numbering). Hydrogen bonds were optimized using the Protein Preparation Wizard in Maestro (Schrodinger, LLC).

The initial model of the PIFtide-PDK1 complex was refined in a series of stages using loop and side chain optimization in the Protein Local Optimization Program (PLOP)(1, 2). Implicit solvation with the variable-dielectric SGB model(3) was used along with an ionic strength of 0.025 (i.e., corresponding to 25mM NaCl). Two different

protocols were used: a restrained protocol, with Cβ-Cβ restraints between each cysteine mutant position on PDK1 and the PIFtide residue(s) with the highest % conjugation, and another without restraints. To obtain optimal parameters for the restrained protocol, conformational search distance cut-off constraints were varied from 4.0Å to 9.0Å at 0.5Å intervals, and steric screening overlap factors were varied to be 0.60, 0.65, or 0.70. During minimization steps, a distance of 5.0Å and a force constant of 1.0 kcal/mol/Å^2 was used for all restraints. The lowest distance and then most permissive overlap factor where the constraints in all steps were satisfied was 7.0Å, with an overlap factor of 0.60. In the protocol without restraints, the overlap factor was raised to 0.70 to avoid combinatorial explosion at low sampling resolution.

Model refinement was performed in stages (see table below) with residue ranges chosen to give broad sampling while avoiding combinatorial explosion. Backbones of residues with flexible side-chains were also movable during minimization steps in each optimization. The lowest energy model from unrestrained and restrained protocols were then taken forward into further refinement using unrestrained molecular dynamics simulations.

Step	Restrained	Unrestrained
1	Side chain optimization of Lys76 and Arg131	
2	Loop optimization*, Tyr472-Trp479	
3	Loop optimization, Glu465-Phe470	Loop optimization, Glu465-Phe473
4	Loop optimization*, Gln467-Asp474	Loop optimization*, Met469-Phe475
5	Loop optimization*, Asp472-Asp478	Loop optimization*, Asp471-Asp478

*Loop optimizations that also optimized side chains within 7.5 Ang of the given segment.

Molecular dynamics simulations

To examine the stability of each of the models and to further refine each of the PIFtide-PDK1 complex models in an unbiased, physics-based manner, duplicate unrestrained molecular dynamics simulations in explicit solvent were performed using the Desmond (4) software package, with the OPLS-AA/SPC forcefield(5, 6). An orthorhombic simulation box was used with a buffer of 10Å on each side. Na^+ and Cl^- ions were added to neutralize the system, then 0.025 M NaCl was added. After minimization and equilibration, duplicate (different random seed) production runs of 5 ns were performed on each system using the Martyna-Tobias-Klein integrator(7) at 300 K (Nose-Hoover thermostat(8)) and 1 atm. Snapshots were output every 1.002 ps. All bonds involving hydrogens were constrained, a 2 fs time step for the bonded and short-range nonbonded interactions was used, and long-range nonbonded interactions were updated every 6 fs using the RESPA multiple time step approach. Short-range coulombic and van der Waals nonbonded interactions were cut-off at 9.0Å, and long-range electrostatics were computed using the smooth particle-mesh Ewald method. Pairlists were constructed using a distance of 10.5 Å and a migration interval of 12 ps. To determine the most favorable conformer and an "envelope" of favorable conformations, clustering of PIFtide and PDK1 residues in Table I was performed using GROMACS 4.0.7(9). PIFtide-PDK1 Cβ-Cβ distances were measured using a Tcl script and VMD 1.8.7(10).

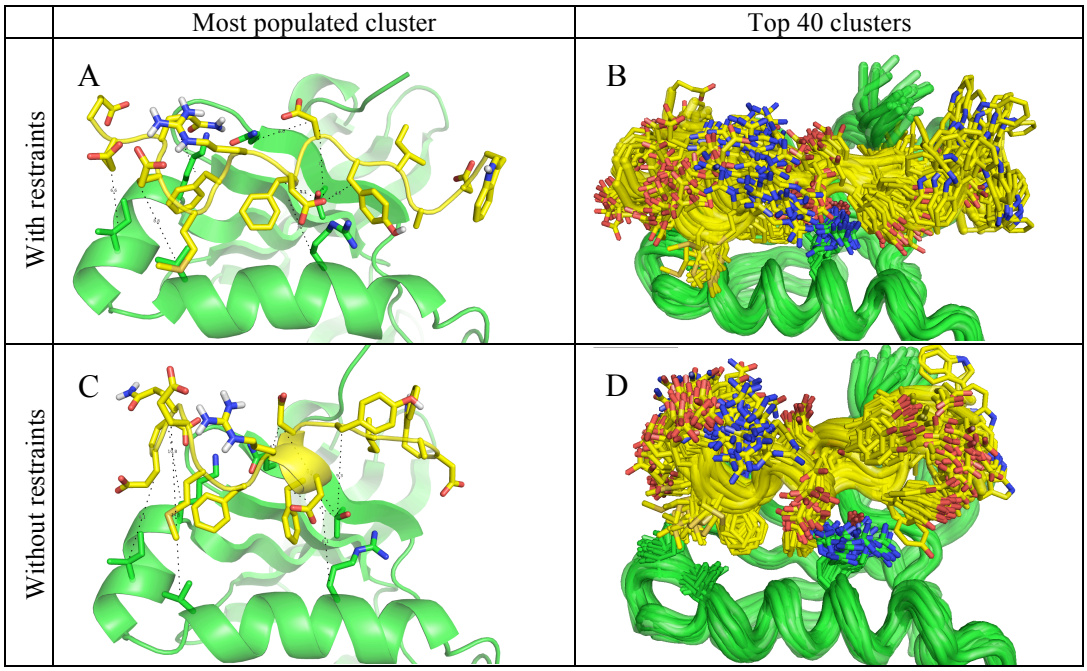


Figure 1. Molecular dynamics simulations of a model of the PIFtide-PDK1 complex created using restraints derived from disulfide crosslinking yielded lower PIFtide-PDK1 Cβ-Cβ distances and a narrower conformational distribution than the those of a model created without restraints. (A,C) Cluster representative with the greatest population from the simulations of models created with and without restraints, respectively. Dotted lines indicate Cβ-Cβ distances given in Table I. (B,D) Superposition of top 40 clusters' (by population) representative structures from the simulations of models created with and without restraints, respectively.

Distance(Å) between ...		Protocol	
PIFtide Cβ	PDK1 Cβ	Restrained	Unrestrained
Glu 466	Ile 119	**6.6**	8.1
Gln 467	Lys 115	**9.6**	10.8
Glu 468	Val 124	**8.8**	14.5
Glu 472	Arg 131	**7.5**	8.1
Phe 473	Thr 148	**4.7**	5.4
Asp 474	Thr 148	**7.5**	9.8
Asp 474	Gln 150	7.0	**6.3**
Tyr 475	Thr 148	**4.1**	9.3

Table I. Distances between PIFtide and PDK1 Cβ atoms used to derive distance restraints show that the MD ensemble from the restrained homology model generally has lower Cβ- Cβ distances than from the ensemble from the unrestrained homology model.

Distances are shown for the cluster representative with the highest population and averaged over the whole ensemble. Note that the average distances are lower especially in the N-terminus of PIFtide and between PDK1's Thr148 and PIFtide's Asp474 and Tyr475, the most C-terminal residues for which restraints were available.

References

1. Jacobson MP*, et al.* (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins* 55(2):351-367.
2. Zhu K, Shirts MR, Friesner RA, & Jacobson MP (2007) Multiscale Optimization of a Truncated Newton Minimization Algorithm and Application to Proteins and Protein-Ligand Complexes. *Journal of Chemical Theory and Computation* 3(2):640-648.
3. Zhu K, Shirts MR, & Friesner RA (2007) Improved Methods for Side Chain and Loop Predictions via the Protein Local Optimization Program: Variable Dielectric Model for Implicitly Improving the Treatment of Polarization Effects. *Journal of Chemical Theory and Computation* 3(6):2108-2119.
4. Bowers KJ*, et al.* (2006) Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *SC 2006 Conference, Proceedings of the ACM/IEEE*, pp 43-43.
5. Kaminski GA, Friesner RA, Tirado-Rives J, & Jorgensen WL (2001) Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptidesâ€ *The Journal of Physical Chemistry B* 105(28):6474-6487.
6. Berendsen HJC, Postma JPM, van Gunsteren WF, & Hermans J (1981) *Intermolecular Forces* (D. Reidel Publishing Company, Dordrecht) p 331−342.

7. Martyna GJ, Tobias DJ, & Klein ML (1994) Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics* 101(5):4177-4189.

8. Hoover WG (1985) Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A* 31:1695–1697.

9. Hess B, Kutzner C, van der Spoel D, & Lindahl E (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation* 4(3):435-447.

10. Humphrey W, Dalke A, & Schulten K (1996) VMD -- Visual Molecular Dynamics. *Journal of Molecular Graphics* 14:33-38.

Chapter 6. Substrate and inhibitor-induced dimerization and cooperativity in caspases-1 and -3

Introduction

Many oligomeric enzymes show positive cooperativity where substrate binding induces a transition from a less active to more active state (for recent reviews, see [1]). The structural basis for this phenomenon could derive from either substrate binding inducing oligomerization and/or inducing an allosteric change in the oligomer. For example, B-Raf, a kinase important in cell proliferation has recently been shown to undergo a substrate-induced monomer to dimer transition that leads to a highly active form of the enzyme[2]. By contrast, glycogen phosphorylase is known to show positive cooperativity through binding of substrate leading to a conformational switch that induces a more active dimer[3].

Caspases, a family of aspartate-specific cysteine proteases important in inflammation and apoptosis, are also known to undergo allosteric transitions (for review see XX). The most dramatic is caspase-1 which shows a Hill coefficient of 1.4 for substrate activation[4]. Crystal structures have been solved of caspase-1 in the presence and absence of active site inhibitors showing both to be dimeric structures but with substantial changes in the active site region (Figure 1)[5]. The apo-like structure has also been trapped with allosteric inhibitors at the dimer interface some 15 Å from either active site confirming that the apo-state is indeed inactive[5]. These studies suggest that caspase-1 is capable of substrate-induced activation through allosteric transition within the dimer.

However, caspase-1 is generated from proteolysis of pro-caspase-1, an inactive and monomeric precursor. Thus, caspase-1 is subject to dimerization after proteolytic maturation. Moreover, cross-linking studies suggest that caspase-1 at low concentration can exist in a monomeric form[6]. We wished to quantify the dimerization constant of the mature enzyme in the presence and absence of substrate or active-site inhibitors to understand how binding influences dimerization, and to determine the relative importance of these processes to substrate activation and cooperativity.

Here, we show using biophysical and kinetic measurements that active-site binding of inhibitor or substrate shifts the monomer-dimer equilibrium constant in caspase-1 a hundred-fold to favor dimer formation under physiologically relevant conditions. We also show that binding at one site enhances by ten-fold the catalytic efficiency at the second site. We present an integrated model suggesting that at physiological concentrations caspase-1 exists predominantly as a monomer, that undergoes both substrate-induced dimerization as well as substrate or inhibitor-induced activation that account for the positive cooperativity observed. In contrast we find that caspase-3, which is a constitutive dimer at physiologic conditions, lacks positive cooperativity and shows a very weak inhibitor-induced activation. These data suggest that the changes in oligomer state upon substrate binding in caspase-1 versus the lack thereof for caspase-3 can account for the differences in positive cooperativity and has important implications for different biological functions of these two enzymes.

Results

To examine how binding at the active site influences the monomer-dimer equilibrium constant in caspase-1 we employed various biophysical methods (**Figure 2**). Analytical ultracentrifugation was used to determine the dimerization constant in the absence of substrate or inhibitor, taking advantage of an active site C285A mutant that is incapable of self-cleavage and is less prone to aggregation. As shown in **Figure 2A**, the dissociation constant (K_D) calculated for apo caspase-1 (C285A) is 109μM. In sharp contrast, caspase-1 that is fully inhibited with the covalent inhibitor (z-VAD-fmk) forms a 20-fold tighter complex, with a calculated K_D of 5μM (90% confidence interval, [2.4μM, 10μM] (**Figure 2B**).

To examine the dependence of enzyme activity on enzyme concentration and obtain data to fit a kinetic model, we next evaluated the initial rates for hydrolysis of the Ac-WEHD-afc substrate by changing both enzyme and substrate concentrations. We varied the enzyme concentration starting at 10 nM which is far below the dimerization constant measured in the presence of inhibitor. The concentration was raised up to 500nM, thus approaching the low micromolar dissociation constant measured in the presence of inhibitor and importantly ten-fold above the "effective" dimerization constant from the enzymatic data. Studies at higher enzyme concentrations were challenging because they depleted the substrate too quickly for accurate steady-state rates to be determined. As seen in **Figure 3A**, the activity per enzyme subunit undergoes a dramatic rise as the enzyme concentration is increased from below its inhibitor-induced dimerization constant (K_D ~5μM). These data can be fit to the minimal kinetic model that describes the two conformational on- and off-states and the monomer-to-dimer transitions **(Figure 3B)**. Substrate can be hydrolyzed from the monomer or dimer states either when one site is occupied or when both sites are occupied. Using the AUC data to provide restraints on the dimerization affinities, the kinetic constants were fitted to the experimentally determined steady-state activity per enzyme values using a quasi-steady-state solution of the chemical kinetics equations (see Methods for details), and summarized in Table I. This model assumes that substrate (Ac-WEHD-afc) provides the same enhancement in dimerization affinity as inhibitor (z-VAD-fmk), and that the

dimerization affinity with one molecule of z-VAD-fmk bound is close to that with two molecules of z-VAD-fmk bound. Under these assumptions, we obtained an excellent fit to the measured steady-state activities, where the measured dimerization constants (**Figure 2**) are in reasonable agreement with the ratios of the on- and off-rates fit to the steady-state kinetic data in **Figure 3B** (5μM for caspase-1 with two Ac-WEHD-afc bound, 5.6μM for caspase-1 with one Ac-WEHD-afc bound, and 100μM with no Ac-WEHD-afc bound).

We next wished to understand how binding at one site alone affected the kinetic activity of caspase-1. To study this, we created a homogeneous preparation of hemi-labeled caspase-1 in which we labeled only one of the two active sites in the caspase-1 dimer with the covalent inhibitor, z-VAD-fmk. This was achieved by creating two tagged versions of the caspase-1 in which the p20 subunit, containing the active site cysteine, was fused to either an N-terminal Strep-tag or His_6-tag (**Figure 4**). We refolded each p20 in the presence of the p10 subunit and purified each. Control experiments showed the tags enzymes had virtually the same kinetic properties as the wild-type enzyme (Table III). The His-tagged caspase-1 was then fully labeled with the covalent active site inhibitor z-VAD-fmk. Complete labeling was confirmed by full inactivation of enzyme activity and quantitative labeling seen by mass spectrometry. The labeled His_6-tagged caspase-1 was denatured with guanidine-HCl in the presence of an excess of Strep-tagged caspase-1. The mixture was dialized and refolded allowing scrambling of the p20 subunits and generation of hemi-labeled caspase-1 marked by the presence of both the His_6-tag and Strep-tag. The hemi-labeled dual tagged enzyme could be purified away from the homo-tagged enzymes by a dual affinity column: first a nickel column to recover the His_6-tagged enzymes, and then an avidin column to recover the His_6/Strep-tagged enzyme. Mass spectrometry confirmed the dual tagged caspase-1 contained a single z-VAD-fmk label on the His_6-tagged-p20 subunit .

The kinetic constants determined by Michaelis-Menten analysis is shown in **Table II**. When corrected for having half-sites available, the hemi-labeled enzyme shows roughly an 18-fold increase in k_{cat} and a two-fold increase in K_M. The resulting catalytic efficiency (k_{cat}/K_M) for the hemi-labeled enzyme is nine times higher than the unlabeled

control. Because these data are collected under conditions some of the wild-type and hemi-labeled enzyme are dimerized, a substantial portion of the rate enhancement enzyme reflects how binding of inhibitor to one site promotes dimerization, which promotes catalysis at the second site presumably through stabilizing the active conformation at both sites. To compare the hemi-labeled experiments to our previous kinetic model on wild-type caspase-1, we solved a simplified kinetic model given in **Figure 4B**. For the parameters, we took the 5μM dimerization affinity in the presence of z-VAD-fmk from the AUC data and the catalytic rate from parameter "k_9" in Figure 3 and Table I, and then fit the K_M to the steady-state kinetic data; the 44nM value for K_M obtained is in very good agreement with the 30nM value from the wild-type kinetic model (K_M = (k8r + k9)/k8f.

If binding of inhibitor at one site enhances activity at the other site, it is possible that a direct titration of wild-type caspase-1 with z-VAD-fmk could activate at sub-stoichiometric equivalents of inhibitor if we generated hemi-labeled intermediates. Indeed, titration of wild-type caspase-1 lead to systematic activation up to 1.4 fold that plateaus between 0.2 and 0.4 equivalents of inhibitor to active site. Further titration leads to steep fall off reaching zero at about one equivalent of inhibitor per active site. In contrast, the hemi-labeled enzyme shows no activation but undergoes linear inactivation to zero when one reaches about one equivalent per available active site.

We hypothesized the 9-fold enhancement in catalytic efficiency between pure hemi-labeled enzyme and wild-type, compared to the 1.4 fold enhancement seen for the spontaneous titration of wild-type, results because we have only a small amount of hemi-labeled intermediate from z-VAD-fmk enhancing labeling of the second site. In fact, if we assume that hemi-labeled material generated from the titration of the wild-type enzyme should be 9-times more active we can calculate at any point in the titration the ratio of bis- versus hemi-labeled enzyme prior to substrate addition, using our kinetic model in **Figure 3** but assuming all inhibitor is covalently bound and there is no inhibitor coming on/off. Then, assuming the hemi-labeled enzyme is nine times more active than the wild-type (whether dimerized or not), we can calculate the relative activity compared to the uninhibited wild-type caspase-1. Remarkably, we obtained good qualitative

agreement with the experimental data, reproducing the "roller-coaster" phenomenon of activation then inhibition at increasing inhibitor concentrations (Fig. 5), though the agreement is not ideal as our model is greatly simplified (a full model would require a total of nine states and several new parameters and so was deemed too complex for the given data). As a control, we examined the active site titration of the hemi-labeled enzyme and found no marked increase in activity upon addition of inhibitor.

In contrast to caspase-1, caspase-3 is a constitutive dimer and has a Hill coefficient close to 1.0 (ref to Datta or Scheer). We wished to test how hemi-labeling or spontaneous labeling of caspase-3 affected its kinetics. We generated the hemi-labeled construct for caspase-3 as we did for caspase-1 and determined the Michaelis-Menten constants (**Table II**). As shown, once corrected for half of the active sites, the hemi-labeled caspase-3 shows only a 2.5-fold enhancement in k_{cat}, a 2-fold increase in K_M, and overall a very slight 1.3-fold enhancement in k_{cat}/K_M. Next we evaluated how spontaneous titration of caspase-3 with z-VAD-fmk affects its kinetics (**Figure 6**). As shown, titration with z-VAD-fmk caused a linear decrease in activity for both the wild-type caspase-3 as well as the hemi-labeled caspase-3. Thus in sharp contrast, caspase-3 shows dramatically lower propensity for substrate or inhibitor-induced activation.

Discussion

We have presented both biophysical and enzymological evidence that substrate- and inhibitor-assisted dimerization is a major contributor to the cooperativity seen in caspase-1. Importantly, we expressed and purified a half-labeled form of the enzyme that exhibited substantially greater activity than the wild-type, unlabeled enzyme. Furthermore, we observed the unusual phenomenon where sub-stochiometric amounts of inhibitor activated the enzyme; while such an effect is not desired in a drug discovery application, this experiment and the ability of our model to reproduce this phenomenon gave further support for our model that substrate or inhibitor activates caspase-1 through driving formation of a dimeric caspase-1 that has greater activity than the monomeric form.

Does dimerization promote better catalytic activity or better substrate binding? The kinetic parameters suggest that the apo, inactive dimer binds substrate weakly (k_4f = 153 $M^{-1}s^{-1}$, k_4r = 1.9 s^{-1}), while the dimer with one substrate bound ("hemi-labeled") binds substrate more rapidly ($k_8f = 1.2x10^8$ $M^{-1}s^{-1}$, k_8r = 24.4 s^{-1}), suggesting that binding of inhibitor to one site does increase the affinity for inhibitor at the second site on the dimer. In contrast, the catalytic rate for dimeric caspase-1 does not significantly depend on whether one or both sites are bound (k_7 = 1.596 s^{-1} vs. k_9 = 1.74 s^{-1}), indicating the cooperativity lies in substrate binding rather than in catalysis. Thus, we suggest that in the absence of substrate or inhibitor, caspase-1 is in an inactive conformation, while when substrate or inhibitor is bound to one subunit, the other subunit is more likely to be in the active conformation.

Previous work on the caspase-1 enzyme used mutational and structural studies to uncover a linear circuit of functional residues running between the two active sites through the allosteric site. Enzymatic activity is strongly affected by perturbations of this circuit, suggesting that the interactions are important for stabilizing the active conformation of caspase-1. Kinetic analysis also demonstrated robust positive cooperativity, which is unique among caspases [1]. This study extends those results through the use of site-directed mutagenesis and chemical ligands to probe the conformational state of caspases.

As prototypical members of the inflammatory and executioner caspase families, caspase-1 and caspase-3 provide an interesting contrast in enzymatic properties. Previous work had demonstrated that while caspase-1 demonstrated positive cooperativity, caspase-3 did not. In addition, a comparison of baseline enzymatic activity of wild type enzyme shows that caspase-3 is a more active protease than caspase-1 (almost an order of magnitude greater catalytic turnover, as measured by k_{cat}). We have now demonstrated that caspase-1 can be strongly activated by using the "half-labeling" technique to generate a hybrid caspase-1 heterodimer, whereas caspase-3 shows only a slight degree of activation in the same setting.

Taken together, these data suggest a model of caspase conformational states shown in Figure 2A. The free caspase homodimer is in equilibrium between an inactive and active conformation. The observations of lower intrinsic catalytic turnover, positive cooperativity, and hybrid activation all point to caspase-1 residing primarily in the inactive conformation. In contrast, the higher intrinsic activity of caspase-3, lack of positive cooperativity, and inability to be activated by half-labeling all suggest that caspase-3 is sitting in a predominantly active conformation even in the absence of ligand.

The use of a "heterodimer" technique to probe questions of allostery in caspases specifically, and proteases in general, is not without precedent. Recent work by Denault and colleagues used engineered caspase-7 heterodimers where only one of the subunits was rendered catalytically inactive by mutagenesis. This study observed that the activity of the resulting caspase-7 heterodimer was half that of wild type, suggesting that the unaltered catalytic site maintained full activity and that the two catalytic domains in the caspase-7 dimer are equal and independent [5]. Based on our results using the hybrid caspase-3 heterodimer, we would predict that caspase-7, also an executioner caspase, would reside predominantly in the active conformation and not show coupling between active sites. The work by Denault and colleagues seems to support that prediction.

HIV protease is another protease that like caspases exists as a homodimer in its active form. In a study presented by Rozzelle and colleagues, HIV protease was engineered with three point mutations to create a dominant-negative inhibitor of wild type HIV protease. This inhibitor acts by forming a heterodimer with wild type HIV protease and inhibiting its activity [6]. Because of the functional coupling we see between the two subunits of the caspase-1 dimer, it would be interesting to test whether a caspase-1 heterodimer could be generated where an inactive subunit could inhibit the activity of the active subunit. This would stand in marked contrast to the result in caspase-7 described above.

In fact, previous reports have described a dominant-negative effect of a mutated caspase-1 construct. Friedlander and colleagues created an active-site caspase-1 mutant by knocking out the catalytic cysteine with a C285G mutation. This variant was then expressed in a transgenic mouse model under the control of a neuron-specific promoter.

They were able to show that expression of this caspase-1 variant protected neurons from apoptosis following trophic factor withdrawal and reduced brain injury following ischemic events [7]. A subsequent study showed that expression of the same caspase-1 dominant-negative mutant in a mouse model of Huntington's disease delayed symptoms and prolonged survival [8]. The molecular mechanism by which this caspase-1 mutant is able to exert a dominant-negative effect has not been elucidated, but it would be interesting to test this mutant using our method for engineering a hybrid heterodimer.

The elucidation of the differences in conformational state and allosteric regulation between inflammatory and executioner caspases raises questions as to what the biological significance of these observations could be. The inflammatory and apoptotic pathways share very similar molecular elements. Both can be activated by extracellular and intracellular receptors that sense danger signals, both involve scaffolding proteins and the formation of macromolecular complexes, and both pathways are crucially dependent on proteolytic events. However, the end result of the two pathways is very different. In the case of inflammatory signaling and caspase-1 activation, this pathway leads to the release of inflammatory cytokines from the cell and subsequent activation of immune cells. In contrast, the apoptotic pathway and caspase-3 activation leads to the programmed cell death of the cell in which the pathway is activated.

Recent studies have taken a proteomic approach to globally profile proteolytic cleavage products following induction of various caspase pathways. These studies have shown that the number of potential caspase substrates during inflammatory signaling is small, on the order of thirty [9]. Our observation that the naked caspase-1 catalytic domain resides in an inactive conformation suggests that positive cooperativity may provide an additional selectivity filter for cleaving pro-inflammatory substrates only when they are concentrated in cells. It has been recently suggested that caspase-1 cleaves pro-IL-1β that has been concentrated at membranes, in inflammasomes, or in vesicles [10]. However, in apoptotic signaling events, the potential number of caspase substrates is on the order of a thousand [11]. Given these observations, it seems to make sense that the executioner caspase-3 would reside in a more active conformation in comparison to caspase-1. The apoptotic pathway, which results in a vast number of substrates being cleaved and the

dismantling of a cell requires a very active protease, whereas the inflammatory pathway, with few substrates cleaved in a specific cytokine signaling event, seems to call for a protease under much tighter regulation. Our model provides a mechanism by which the very conformational state of the various caspase family members recapitulates their role in specific biologic pathways and provides an additional level of regulation.

Materials and Methods

Expression and purification of caspase-1

Recombinant caspase-1 was prepared by expression in *Escherichia coli (E. coli)* as insoluble inclusion bodies followed by refolding [3]. The p20 (residues 120 – 297) and p10 (residues 317 – 404) subunits of wild type human caspase-1 were cloned into NdeI and EcoRI restriction endonuclease sites of the pRSET plasmid (Invitrogen, Carlsbad, CA). Site-directed mutagenesis was performed to construct the C285A active-site-null mutant.

Caspase-1 subunits were expressed separately in *E. coli* BL21(DE3) Star cells (Invitrogen). Cells were harvested following induction of a log phase culture with 1mM IPTG for 4 h at 37°C and then disrupted with a microfluidizer. The inclusion body pellets were isolated by centrifugation of lysate for 20 min at 4°C. Pellets were washed once with 50 mM HEPES (pH 8.0), 300 mM NaCl, 1 M guanidine-HCl, 5 mM DTT, and 1% Triton X-100, and washed two more times with the same buffer without the detergent. The washed inclusion body pellets were solubilized in 6 M guanidine-HCl and 20 mM DTT, and stored frozen at -80°C.

Refolding of caspase-1 was done by combining guanidine-HCl-solubilized large and small subunits (10mg of large subunit and 20mg of small subunit) in a 250 mL beaker, followed by rapid dilution with 100 mL of 50 mM HEPES (pH 8.0), 100 mM NaCl, 10% sucrose, 1 M nondetergent sulfobetaine 201 (NDSB-201), and 10 mM DTT. Renaturation proceeded at room temperature for 6 h. Samples were centrifuged at 16,000 g for 10 minutes to remove precipitate, and then dialyzed overnight at 4°C against 50 mM sodium acetate (pH 5.9), 25 mM NaCl, 5% glycerol, and 4 mM DTT. Dialyzed protein was purified by cation exchange chromatography using a pre-packed 5 mL

HiTrap SP HP column (GE Healthcare Bio-sciences Corp, Piscataway, NJ). Protein was eluted using a linear gradient of 0-1.0 M NaCl over 20 min in a buffer containing 50 mM sodium acetate (pH 5.9) and 5% glycerol. Peak fractions were pooled and β-ME was added to a concentration of 1mM before samples were stored frozen at -80°C.

For the kinetic analyses at varying enzyme and substrate concentrations for wildtype caspase-1, the cation exchange peak fractions were concentrated using Millipore Ultrafree-15 devices with a MWCO of 10,000 Da and further purified using size exclusion chromatography using a Superdex 200 16/60 column in 25 mM Tris (pH 8.0), 50 mM NaCl, 5% glycerol, and 1mM DTT. For the AUC experiments, cation exchange peak fractions were concentrated and further purified using size exclusion chromatography using a Superdex 200 16/60 column in a buffer containing 50 mM HEPES (pH 8.0), 50 mM KCl, and 200 mM NaCl, and then frozen at -80°C.

Analytical Ultracentifrugation

Sedimentation equilibrium experiments were performed on a Beckman XL-I analytical ultracentrifuge at 20°C, at rotor speeds of 10,000, 14,000, and 20,000 r.p.m. For the C285A active-site-null construct, samples were centrifuged at 90,000 r.p.m. for 10 minutes to remove remaining aggregate prior to measurement. To obtain initial absorbance values of between 0.2 and 0.8 AU, loading concentrations were 16μM, 12μM, and 8μM for C285A caspase-1, and 19.5μM for doubly-labeled wildtype caspase-1. The lower equilibrium concentrations observed in the AUC experiment indicate that some aggregate formed and was subsequently spun to the bottom of the cell. Data analysis was performed using SEDFIT[12] and SEDPHAT[13]. For global fitting, mass conservation was employed and the meniscus, bottom, local concentrations, and log association constant were floated as parameters.

Active site titrations

Functional protein concentration for enzyme kinetic analysis was determined by active-site titration [14]; caspase was incubated in assay buffer for 2 hrs at room temperature with a titration from a zero- to 2-fold stoichiometric ratio using the irreversible active-site inhibitor z-VAD-FMK. The protein was diluted to an enzyme

concentration of 50 nM and activity was determined using fluorogenic tetrapeptide substrate (Enzo Life Sciences) at 25 μM for caspase-1 or 50 μM for caspase-3. The substrates used were Ac-WEHD-afc for caspase-1 constructs and Ac-DEVD-afc for caspase-3 constructs [15].

Expression and purification of half-labeled caspase constructs

The generation of hybrid caspase-1 and caspase-3 constructs follows the protocol described above. Caspase-1 (residues 120-297) and caspase-3 (residues 29-175) large subunits containing an N-terminal His_6- or Strep-affinity tags were generated by designing 5'- primers with the appropriate sequence for the affinity tag and using polymerase chain reaction (PCR) to generate dsDNA inserts. The affinity-tag amino acid sequences were "MRGSHHHHHHSAG-" for the His_6-tagged construct and "MWSHPQFEKSAG-" for the Strep-tagged construct. The Strep-tag is an eight amino acid peptide that binds with high selectivity to the streptavidin variant Strep-Tactin (IBA GmbH, Germany). These inserts were sub-cloned into NdeI and EcoRI restriction endonuclease sites of the pRSET plasmid (Invitrogen, Carlsbad, CA). The p10 small subunit of caspase-1 (residues 317 – 404) and the p17 small subunit of caspase-3 (residues 176 – 277) were also cloned into the pRSET plasmid. These constructs were then expressed and purified as inclusion body pellets as described above. In addition, full-length caspase-3 constructs (residues 1-277) containing N-terminal affinity tags were generated using the above 5'- primer and then sub-cloned into NdeI and EcoRI restriction endonuclease sites of the pET-23b plasmid (Novagen). This construct was used for soluble expression of caspase-3.

Expression and purification of half-labeled caspase-1

To generate the hemi-labeled caspase-1 construct, active caspase-1 was first generated by refolding a His_6- or Strep-tagged large subunit with the small subunit to generate the active caspase dimer, as described above. Following refolding, samples were centrifuged at 16,000 g for 10 minutes to remove precipitate, diluted two-fold into 50 mM sodium acetate (pH 5.9) buffer, and then centrifuged once more at 16,000 g for 10 minutes. Samples were filtered and purified by cation exchange chromatography.

The tagged caspase-1 construct was then labeled with the irreversible active site inhibitor z-VAD-fmk in labeling buffer containing 50 mM HEPES (pH 7.4), 50 mM KCl, 200 mM NaCl, and 10 mM DTT overnight at 4°C. Complete labeling of the tagged p20 subunit was verified by liquid chromatography-mass spectrometry (LC-MS; Waters, Milford, MA) and complete inhibition of catalytic activity. Excess inhibitor was removed using a Superdex 200 10/300 gel filtration column (GE Amersham) in buffer containing 25 mM Tris (pH 8.0), 50 mM NaCl, 5% glycerol, and 1mM DTT. The VAD-fmk-labeled tagged caspase-1 was then concentrated using Millipore Ultrafree-15 devices with a MWCO of 10,000 Da and then denatured in 6M guanidine. This sample was then refolded in the presence of the other tagged p20 subunit and excess p10 subunit. Refolding and purification by cation exchange chromatography were done as described above. The three affinity-tagged caspase-1 species were then separated using sequential 1mL HisTrap and 1mL StrepTrap columns for affinity purification (GE Healthcare), and the final half-labeled caspase-1 construct was purified by size exclusion chromatography using a Superdex 200 16/60 column in 25 mM Tris (pH 8.0), 50 mM NaCl, 5% glycerol, and 1mM DTT. Final verification of sample purity as being properly half-labeled was performed by LC-MS.

The control heterodimer caspase-1 with both His_6- or Strep-affinity tags but no labeling with the active-site inhibitor z-VAD-fmk was generated in a similar fashion as above. Three caspase-1 constructs, the His_6-tagged p20, Strep-tagged p20, and p10 subunits from inclusion bodies were refolded together, then purified as above using cation exchange chromatography, sequential His_6- and Strep-tag affinity based purification, and finally size exclusion chromatography.

Expression and purification of half-labeled caspase-3

To generate the half-labeled caspase-3 construct, active caspase-3 was first generated by soluble expression in *E. coli* BL21(DE3)pLysS cells (Stratagene). Cells were grown in 2xYT media containing 200 μg/ml ampicillin and 50 μg/ml chloramphenicol at 37 °C to an OD600nm of 0.8-1.0. Overexpression of caspase-3 was induced with 200 μM IPTG at 37°C for three hours. Cells were harvested and resuspended in 100 mM Tris (pH 8.0) and 100 mM NaCl for lysis by microfluidization

(Microfluidics). The cell lysate was spun at 45,000xg for 30 minutes at 4°C. Caspase-3 with an N-terminal His_6-affiniy tag was isolated using a 1 ml HisTrap HP Ni-NTA affinity column (GE Amersham) eluted with buffer containing 200 mM imidazole. The eluted protein was diluted two-fold with buffer containing 20 mM Tris, pH 8.0 and then purified by anion-exchange chromatography (HiTrap Q HP, GE Amersham) with 30-column volume gradient from 0-0.5 M NaCl.

The His_6-affiniy tagged caspase-3 construct was then labeled and purified with the irreversible active site inhibitor z-VAD-fmk as described above for caspase-1. Following refolding with Strep-affinity tagged p17 large subunit and p12 small subunit, samples were dialyzed overnight at 4°C against 20 mM Tris (pH 5.9), 1 mM DTT. Dialyzed protein was purified by anion-exchange chromatography as described above. The three affinity-tagged caspase-3 species were then separated using sequential 1mL HisTrap and 1mL StrepTrap affinity purification (GE Healthcare), and the final half-labeled caspase-3 construct was purified using a Superdex 200 16/60 gel filtration in 20 mM Tris (pH 8.0), 50 mM NaCl, and half-labeling was verified by LC-MS.

The control heterodimer caspase-3 with both His_6- or Strep-affinity tags but no labeling with the active-site inhibitor z-VAD-fmk was generated in a similar fashion as above. Three caspase-3 constructs, the His_6-tagged p17, Strep-tagged p17, and p12 subunits from inclusion bodies were refolded together, and then purified as above using anion exchange chromatography, sequential His_6- and Strep-tag affinity based purification, and finally size exclusion chromatography.

Enzyme kinetic analysis

Kinetic analysis of caspase-1 was performed in a buffer containing 50 mM HEPES (pH 8.0), 50 mM KCl, 200 mM NaCl, 10 mM DTT, 0.1% 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS), and NaOH was added dropwise to correct the pH to 8.0. Kinetic analysis of caspase-3 was performed in buffer containing 50 mM HEPES (pH 7.4), 50 mM KCl, 0.1 mM EDTA, 1 mM DTT and 0.1% CHAPS. Steady-state kinetic analysis was done by titrating enzyme with fluorogenic tetrapeptide substrate (Ac-WEHD-AFC for caspase-1 constructs and Ac-

DEVD-AFC for caspase-3 constructs, Enzo Life Sciences). Kinetic data was collected for a 10 min time course using a Spectramax M5 microplate reader (Molecular Devices, Sunnyvale, CA) with excitation, emission, and cutoff filters set to 365 nm, 495 nm, and 435 nm, respectively.

For the tagged caspse-1 and caspase-3 constructs, Kinetic constants V_{max}, K_M, and the Hill coefficient (n_{Hill}) were calculated using GraphPad PRISM. The initial velocity (*v)*, measured in relative fluorescence units per unit time, was plotted versus the logarithm of substrate concentration. The model used to fit the data is a sigmoidal does-response curve with variable slope, and from this model all three kinetic constants were derived. The general equation of this model is Y = Bottom + (Top-Bottom)/(1+10^((LogEC$_{50}$-X)*HillSlope)), where Y is the initial velocity, X is the logarithm of the substrate concentration, and Top, Bottom, EC_{50} (K_M), and Hill Slope are free parameters fit to the data. A standard curve using pure afc product was used to convert relative fluorescence units to units of concentration (μM). In determining kinetic constants for caspases we observed that at saturating substrate concentrations, the enzyme exhibited decreasing activity as substrate concentration increased, most likely due to product inhibition. In order to correctly fit our data using non-linear regression, data points exhibiting product inhibition were excluded.

To test our proposed kinetic model for wildtype caspase-1, enzyme and substrate where both varied, using the above assay buffer of 50 mM HEPES (pH 8.0), 50 mM KCl, 200 mM NaCl, 10 mM DTT, 0.1% 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS). Caspase-1 was varied from 500nM to 5.8nM in a serial 1.5-fold dilution, and substrate peptide Ac-WEHD-AFC was varied from 0.34uM to 53.2uM in a 1.4-fold dilution in 96-well format, at 20% DMSO so that the final concentration of DMSO was 2% by volume. A linear region of the data within the first two minutes was selected for the steady-state activity measurement. Enzyme concentrations of 333nM and 500nM and substrate concentrations below 1.31uM were excluded from analysis due to the fact that they consumed substrate too rapidly. To convert between relative fluorescence units and product concentration, a steady-state a curve was constructed using the relative fluorescence units at the highest enzyme concentration after substrate

was fully consumed, to internally correct for the effects of caspase-1 on the relative fluorescence.

Steady-state kinetic modeling

Chemical kinetic equations were derived from the scheme in Fig. 3C. These equations were solved assuming quasi-steady state conditions. The derivatives with respect to time of the concentration of species A,B,D, and E was set to zero. It was assumed that substrate concentrations at steady-state were equal to their intial values; this typical approximation was deemed appropriate as the data-points used for fitting were under a linear product production per unit time phase. Enzyme mass conservation was used. To solve the equations, we made temporary approximations that were then updated over ten iterations, these approximations used in our iterative analytic method gave results that were consistent with fully-numerical solutions for the parameters reported here. Parameters were fit using the Matlab optimization toolkit using an objective function consisting of steady-state rates from the assay data and restraints for ratios of dimerization rates based on the AUC data, assuming the hemilabeled dimerization affinity is close to the doubly-labeled dimerization affinity. Parameter optimization was performed in two stages. An intial unconstrained minimization ("fminsearch" function) was performed and then followed by a constrained nonlinear reression ("lsqnonlin") with physically-reasonable constraints (see supplemental data), for example to keep on-rates less than $1x10^{9}s^{-1}$. The fitted parameters are given in Table I.

Acknowledgments:

We wish to thank colleagues in the lab, William deGrado, and Susan Miller for useful interactions. This work was supported by the National Institute of Health (ROI-AI070292 to J.A.W. and #GMO7618 to D.D.), and by a PhRMA Foundation fellowship to C.L.M.

Figures and Legends:

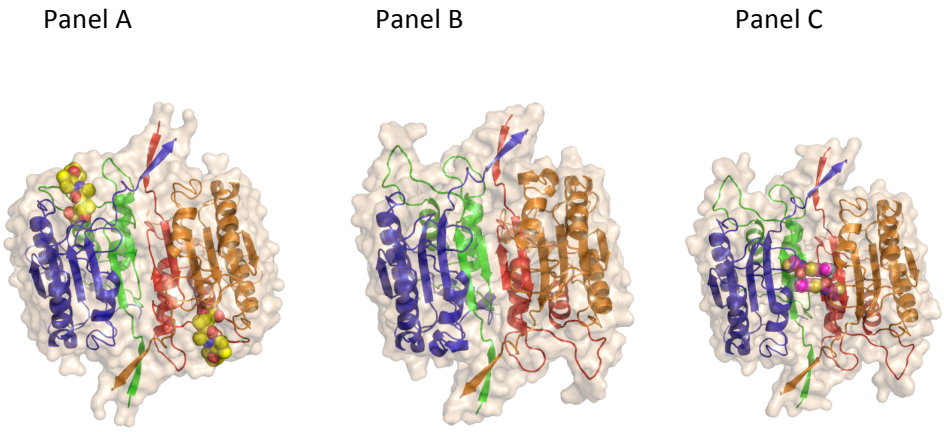
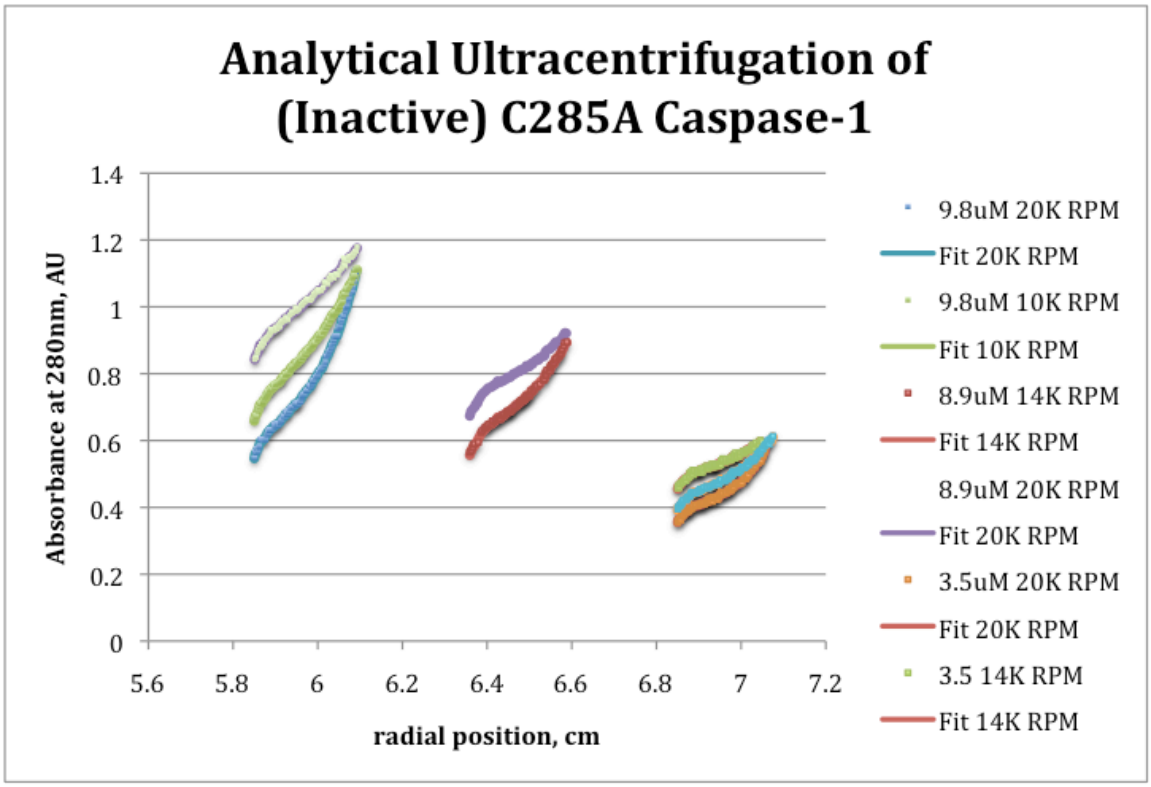
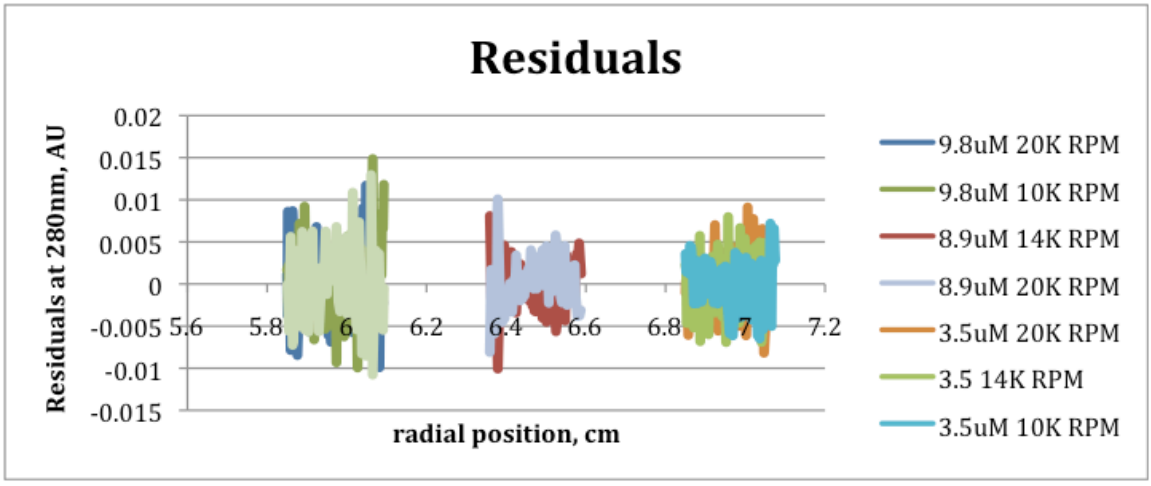


Figure 1. Structures of Caspase-1 with the active site inhibitor z-VAD-fmk (PDB coordinates 2HBQ; Thornberry et al) (**Panel A**), with no ligand (PDB coordinates 1SC1 Scheer et al) (**Panel B**) or with the allosteric inhibitor (PDB coordinates 2FQQ; Romanowski et al) (**Panel C**). Ligands are shown as space-filled models. The large subunits (p20) are colored blue and orange and the small subunits (p10) are colored green and red. As shown, the blue and green chains comprise the left half of the caspase-1 dimer, while the red and orange chains form the right half. The loops near the active sites (top left and bottom right of structures) undergo marked changes between conformations.

Panel A.


Analytical Ultracentrifugation of (Inactive) C285A Caspase-1
Absorbance at 280nm, AU
1.4
1.2
1
0.8
0.6
0.4
0.2
0
5.6
5.8
6
6.2
6.4
6.6
6.8
7
7.2
radial position, cm
9.8uM 20K RPM
Fit 20K RPM
9.8uM 10K RPM
Fit 10K RPM
8.9uM 14K RPM
Fit 14K RPM
8.9uM 20K RPM
Fit 20K RPM
3.5uM 20K RPM
Fit 20K RPM
3.5 14K RPM
Fit 14K RPM


Residuals
Residuals at 280nm, AU
0.02
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
5.6
5.8
6
6.2
6.4
6.6
6.8
7
7.2
radial position, cm
9.8uM 20K RPM
9.8uM 10K RPM
8.9uM 14K RPM
8.9uM 20K RPM
3.5uM 20K RPM
3.5 14K RPM
3.5uM 10K RPM

Panel B.

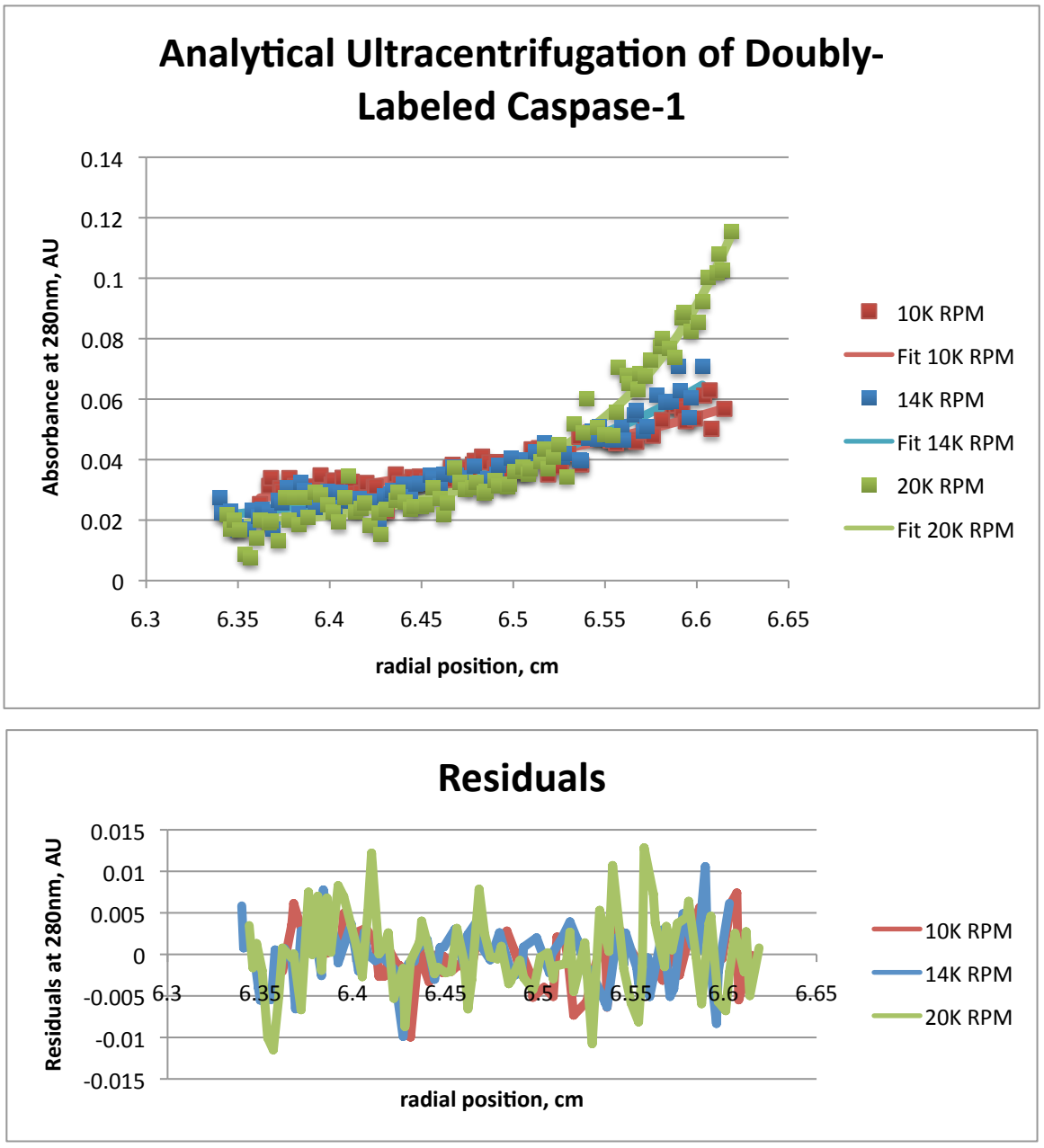
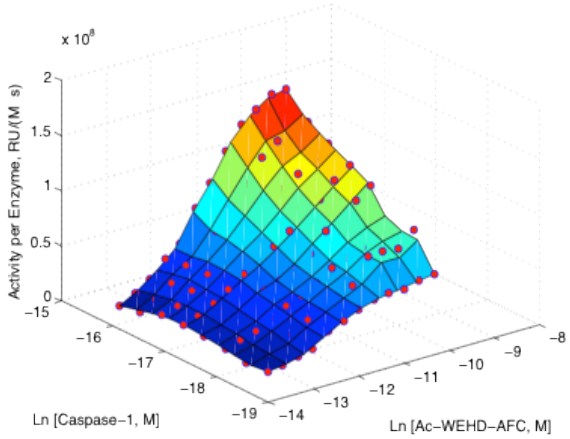
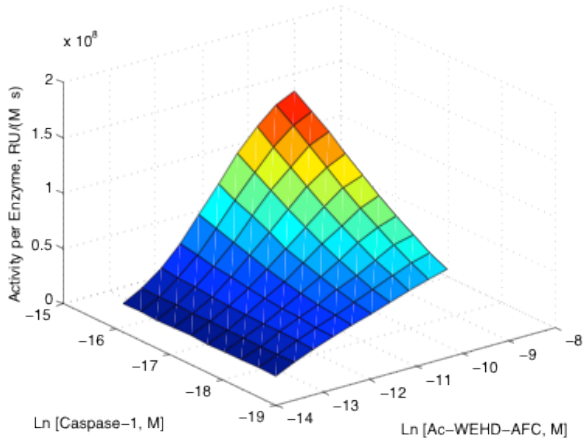


Figure 2. Solution measurements of the dimerization constants for caspase-1 yielded a K_D of 109μM for apo caspase-1 and a K_D of 5μM for active site inhibitor-bound caspase-1. **Panel A.** Analytical ultracentrifugation of catalytically-inactive C285A caspase-1 in the absence of ligands at 10,000, 14,000, and 20,000 r.p.m., at three different equilibrium concentrations: 9.8μM, 8.9μM, and 3.5 uM. **Panel B.** Analytical ultracentrifugation of caspase-1 fully inhibited with the covalent active site inhibitor z-VAD-fmk at 10,000, 14,000, and 20,000 r.p.m., at an equilibrium concentration of 1μM.

Panel A.



Panel B.



Panel C.

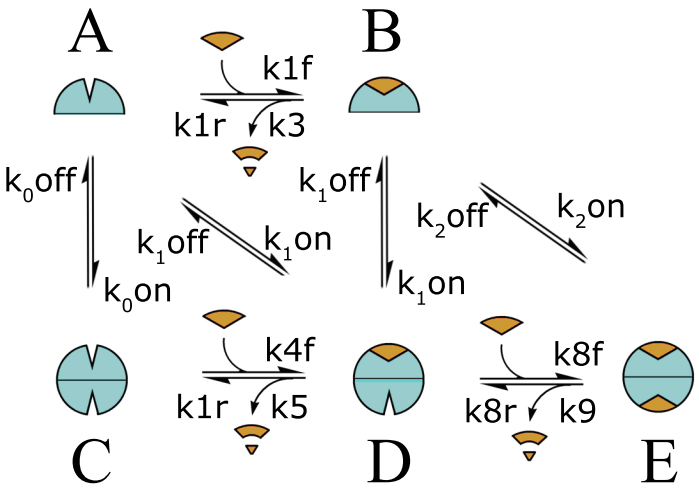
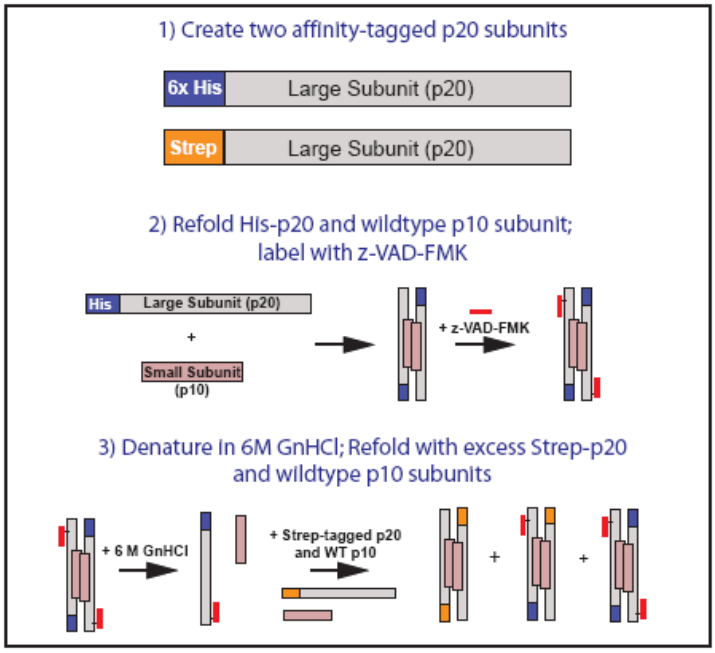
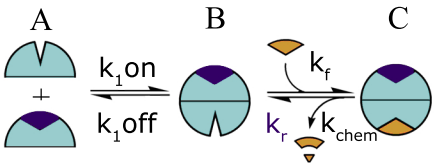


Figure 3. Steady-state kinetics for cleavage of Ac-WEHD-afc by caspase-1. **Panel A.** Experimentally-determined steady-state substrate cleavage kinetics varying [caspase-1] from 8.67 to 148 nM, and [substrate] from 1.31 to 532 μM. **Panel B.** Fitted steady-state cleavage rates from our kinetic model with parameters shown in **Panel C**. **Panel C.** Steady-state kinetic model describing on- and off-states for caspase-1, in the presence and absence of substrate (open and closed shapes, respectively), undergoing monomer to dimer transitions.

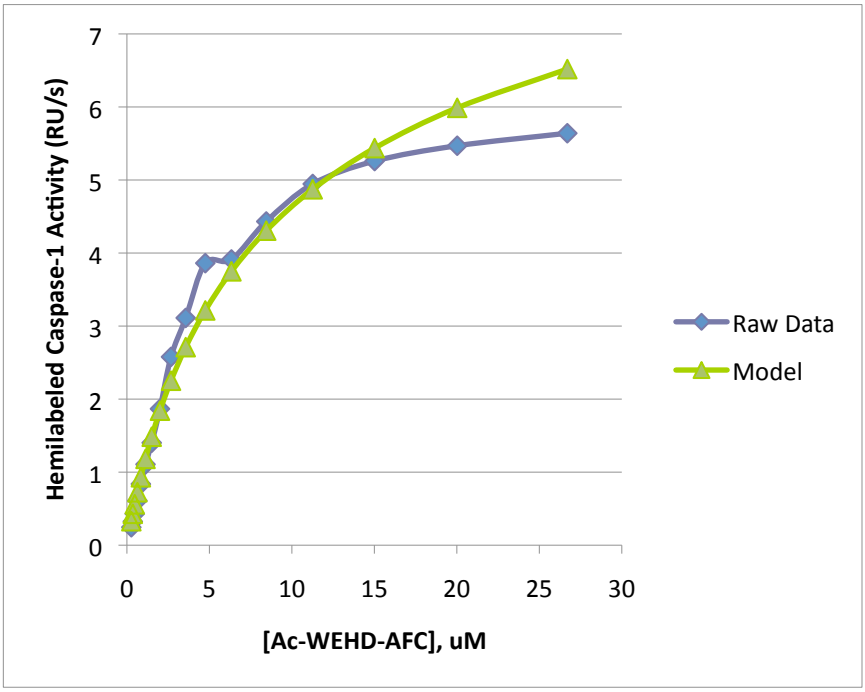
Panel A.



Panel B.



Panel C.

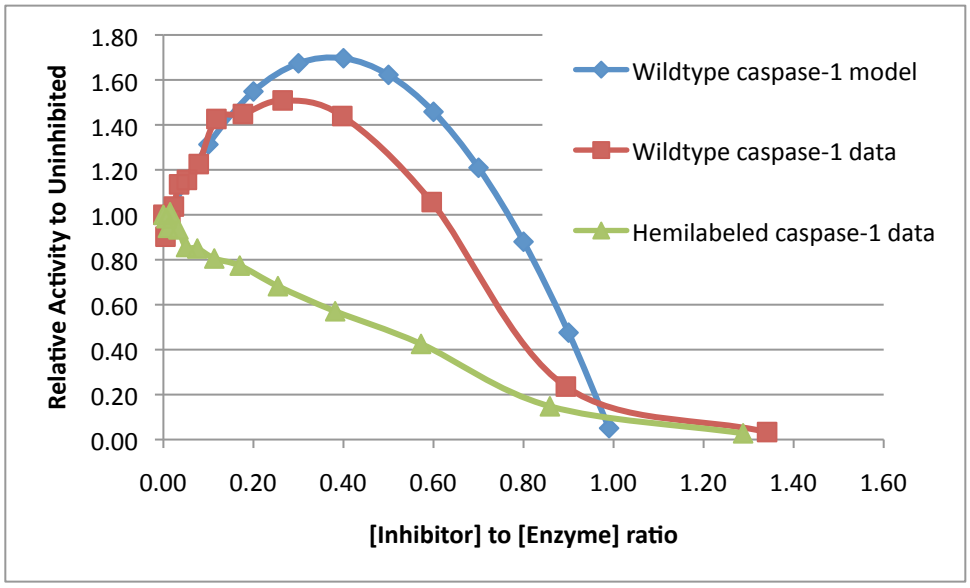


Parameters:	Value		Source
K_D (k_{1off}/k_{1on})	$5x10^{-6}$	M	Fig. 2
K_M ($k_r+k_{cat})/k_f$	$4.36x10^{-8}$	M	Fitted
k_{cat}	1.74	s^{-1}	Table I

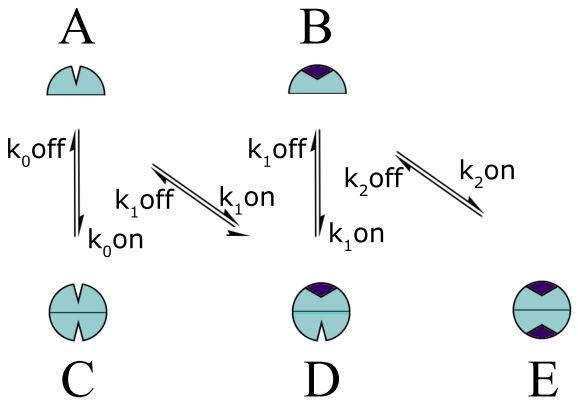
Figure 4. A simplified kinetic model fits activity data for caspase-1 hemi-labeled with z-VAD-fmk. **Panel A.** Scheme showing the preparation of the hemi-labeled caspase-1. Two separate p20 subunits were fused with either a 6X His or Strep affinity tag, and refolded with the p10 to generate the pure homo-dimers. The refolded 6X His tag caspase-1 was fully inhibited with z-VAD-fmk and complete labeling was confirmed by mass spectrometry (see supplemental materials). After gel filtration to remove the active site inhibitor, the enzyme was mixed in excess amounts of unlabeled Strep tagged caspase-1. The proteins were denatured and refolded allowing for scrambling of the subunits. The hemi-labeled species was isolated by affinity purification of the protein that bound to both the nickel and avidin column, and confirmed by mass spectrometry. **Panel B.** Simplified kinetic scheme for the hemi-labeled enzyme assuming all activity comes from the dimer. **Panel C.** Fit of K_M of simplified kinetic scheme in **Panel B** to the activity as a function of [Ac-WEHD-afc] concentration at an enzyme concentration of 50nM, given dimerization affinity and chemical rate k_{chem} from the wild-type model (k_9).

Figure 5. Active site titration of wild-type caspase-1 (blue labels) or hemi-labeled hybrid caspase-1 (red labels). Each enzyme was titrated with increasing amounts of the active site-inhibitor z-VAD-fmk in sub-stoichiometric increments up to 1.4 equivalents per active site. These were allowed to react to completion for 2 hrs at room temperature, and enzyme activity measured as described in the Materials and Methods. **Panel A.** Measured relative activity of wild-type and hemi-labeled caspase-1 and predicted relative activity of wild-type caspase-1 from the steady-state model. **Panel B.** A simplified steady-state model of inhibitor-assisted dimerization (assuming complete labeling with inhibitor), taking parameters from **Table I**. **Panel C.** Predicted amount of hemi-labeled caspsae-1 species formed during the active site titration using a quasi-steady-state solution to the model in **Panel B**.

Panel A.



Panel B.



Panel C.

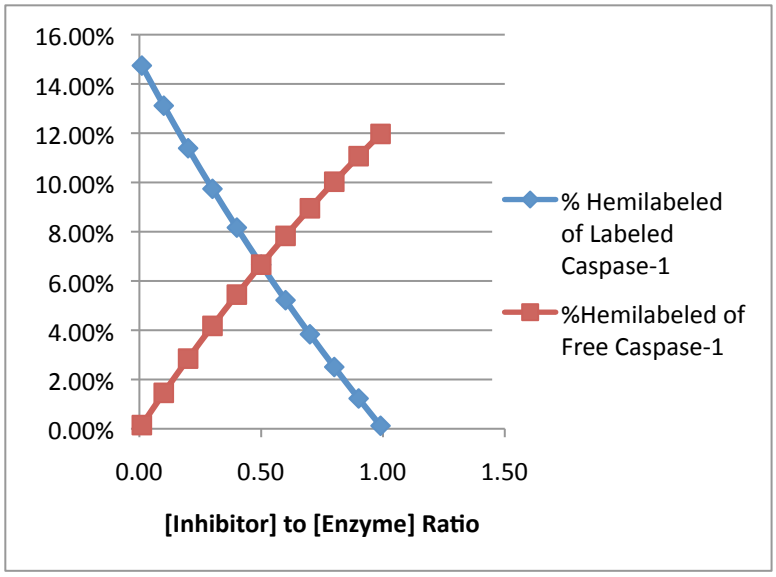

16.00%
14.00%
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
0.00
0.50
1.00
1.50
[Inhibitor] to [Enzyme] Ratio
% Hemilabeled of Labeled Caspase-1
%Hemilabeled of Free Caspase-1

Figure 6. Active site titration of caspase-3 constructs under similar conditions to Figure 5 for caspase-1. Figure courtesy of D. Datta.

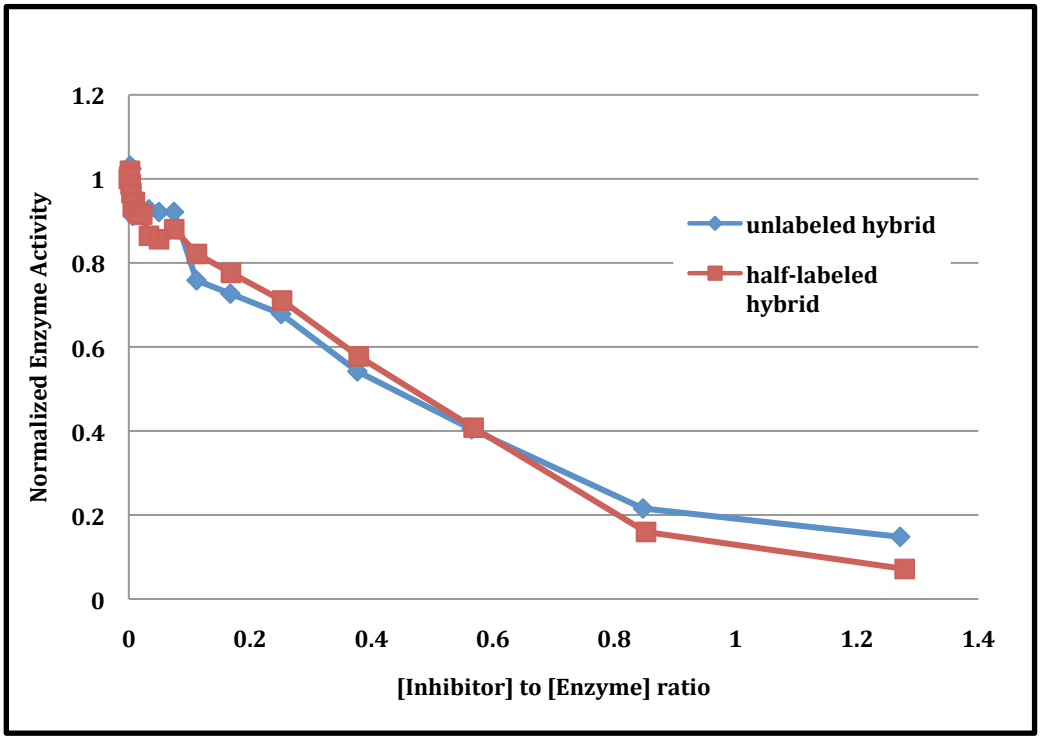


Table I. Parameters for caspase-1 kinetic steady-state model fit to assay data

Parameter	Value	Units
k_{1r}	$2.30x10^{-6}$	s^{-1}
k0off	5.22	s^{-1}
k3	$2.26x10^{-1}$	s^{-1}
k1off	$2.43x10^{1}$	s^{-1}
k1f_prime	$2.60x10^{5}$	$M^{-1}s^{-1}$
k0on	$5.22x10^{4}$	$M^{-1}s^{-1}$
k1on	$4.35x10^{6}$	$M^{-1}s^{-1}$
k2off	$5.34x10^{1}$	s^{-1}
k2on	$1.06x10^{7}$	$M^{-1}s^{-1}$
k4r	1.91	s^{-1}
k7	1.60	s^{-1}
k4f_prime	$1.53x10^{2}$	$M^{-1}s^{-1}$
k8r	$2.44x10^{1}$	s^{-1}
k9	1.74	s^{-1}
k8f_prime	$1.24x10^{8}$	$M^{-1}s^{-1}$

Table II. Hemi-labeling of caspase-1 dimer leads to enzyme activation. Kinetic constants for hybrid tagged caspase-1 either with or without a single z-VAD-fmk on the His-tagged p20 subunit **(see Figure 4).** Table courtesy of D. Datta.

Construct	K_M μM	k_{cat} sec^{-1}	k_{cat}/K_M $M^{-1}\bullet sec^{-1}$	Ratio k_{cat}/K_M
Unlabeled caspase-1 control	1.9	0.11	5.6×10^4	1
Hemi-labeled caspase-1	3.8	1.93	5.1×10^5	9.1

*Standard errors within 10% of reported values based on data collected in triplicate.

Table III. Hemi-labeling of the caspase-3 dimer results in minor enzyme activation. Table courtesy of D. Datta.

Construct	K_M μM	k_{cat} sec^{-1}	k_{cat}/K_M $M^{-1}•sec^{-1}$	Ratio k_{cat}/K_M
Unlabeled caspase-3 control	13	1.4	1.1×10^5	1
Hemi-labeled caspase-3	28	3.7	1.3×10^5	1.3

*Standard errors within 10% of reported values based on data collected in triplicate.

References

1. Datta, D.; Scheer, J. M.; Romanowski, M. J.; Wells, J. A., An allosteric circuit in caspase-1. *J Mol Biol* **2008,** *381* (5), 1157-67.

2. Scheer, J. M.; Romanowski, M. J.; Wells, J. A., A common allosteric site and mechanism in caspases. *Proc Natl Acad Sci U S A* **2006,** *103* (20), 7595-600.

3. (a) Romanowski, M. J.; Scheer, J. M.; O'Brien, T.; McDowell, R. S., Crystal structures of a ligand-free and malonate-bound human caspase-1: implications for the mechanism of substrate binding. *Structure* **2004,** *12* (8), 1361-71; (b) Scheer, J. M.; Wells, J. A.; Romanowski, M. J., Malonate-assisted purification of human caspases. *Protein Expr Purif* **2005,** *41* (1), 148-53.

4. Talanian, R. V.; Dang, L. C.; Ferenz, C. R.; Hackett, M. C.; Mankovich, J. A.; Welch, J. P.; Wong, W. W.; Brady, K. D., Stability and oligomeric equilibria of refolded interleukin-1beta converting enzyme. *J Biol Chem* **1996,** *271* (36), 21853-8.

5. Denault, J. B.; Bekes, M.; Scott, F. L.; Sexton, K. M.; Bogyo, M.; Salvesen, G. S., Engineered hybrid dimers: tracking the activation pathway of caspase-7. *Mol Cell* **2006,** *23* (4), 523-33.

6. Rozzelle, J. E.; Dauber, D. S.; Todd, S.; Kelley, R.; Craik, C. S., Macromolecular inhibitors of HIV-1 protease. Characterization of designed heterodimers. *J Biol Chem* **2000,** *275* (10), 7080-6.

7. Friedlander, R. M.; Gagliardini, V.; Hara, H.; Fink, K. B.; Li, W.; MacDonald, G.; Fishman, M. C.; Greenberg, A. H.; Moskowitz, M. A.; Yuan, J., Expression of a dominant negative mutant of interleukin-1 beta converting enzyme in transgenic mice prevents neuronal cell death induced by trophic factor withdrawal and ischemic brain injury. *J Exp Med* **1997,** *185* (5), 933-40.

8. Ona, V. O.; Li, M.; Vonsattel, J. P.; Andrews, L. J.; Khan, S. Q.; Chung, W. M.; Frey, A. S.; Menon, A. S.; Li, X. J.; Stieg, P. E.; Yuan, J.; Penney, J. B.; Young, A. B.; Cha, J. H.; Friedlander, R. M., Inhibition of caspase-1 slows disease progression in a mouse model of Huntington's disease. *Nature* **1999,** *399* (6733), 263-7.

9. Agard, N. J.; Maltby, D.; Wells, J. A., Inflammatory stimuli regulate caspase substrate profiles. *Mol Cell Proteomics* **2010,** *9* (5), 880-93.

10. (a) Ferrari, D.; Pizzirani, C.; Adinolfi, E.; Lemoli, R. M.; Curti, A.; Idzko, M.; Panther, E.; Di Virgilio, F., The P2X7 receptor: a key player in IL-1 processing and release. *J Immunol* **2006,** *176* (7), 3877-83; (b) Andrei, C.; Margiocco, P.; Poggi, A.; Lotti, L. V.; Torrisi, M. R.; Rubartelli, A., Phospholipases C and A2 control lysosome-mediated IL-1 beta secretion: Implications for inflammatory processes. *Proc Natl Acad Sci U S A* **2004,** *101* (26), 9745-50; (c) MacKenzie, A.; Wilson, H. L.; Kiss-Toth, E.;

Dower, S. K.; North, R. A.; Surprenant, A., Rapid secretion of interleukin-1beta by microvesicle shedding. *Immunity* **2001,** *15* (5), 825-35.

11. Mahrus, S.; Trinidad, J. C.; Barkan, D. T.; Sali, A.; Burlingame, A. L.; Wells, J. A., Global Sequencing of Proteolytic Cleavage Sites in Apoptosis by Specific Labeling of Protein N Termini. *Cell* **2008**.

12. Schuck, P., Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys J* **2000,** *78* (3), 1606-19.

13. Vistica, J.; Dam, J.; Balbo, A.; Yikilmaz, E.; Mariuzza, R. A.; Rouault, T. A.; Schuck, P., Sedimentation equilibrium analysis of protein interactions with global implicit mass conservation constraints and systematic noise decomposition. *Anal Biochem* **2004,** *326* (2), 234-56.

14. Stennicke, H. R.; Salvesen, G. S., Caspases: Preparation and characterization. *Methods* **1999,** *17* (4), 313-319.

15. (a) Rano, T. A.; Timkey, T.; Peterson, E. P.; Rotonda, J.; Nicholson, D. W.; Becker, J. W.; Chapman, K. T.; Thornberry, N. A., A combinatorial approach for determining protease specificities: application to interleukin-1beta converting enzyme (ICE). *Chem Biol* **1997,** *4* (2), 149-55; (b) Talanian, R. V.; Quinlan, C.; Trautz, S.; Hackett, M. C.; Mankovich, J. A.; Banach, D.; Ghayur, T.; Brady, K. D.; Wong, W. W., Substrate specificities of caspase family proteases. *J Biol Chem* **1997,** *272* (15), 9677-82; (c) Thornberry, N. A.; Rano, T. A.; Peterson, E. P.; Rasper, D. M.; Timkey, T.; Garcia-Calvo, M.; Houtzager, V. M.; Nordstrom, P. A.; Roy, S.; Vaillancourt, J. P.; Chapman, K. T.; Nicholson, D. W., A combinatorial approach defines specificities of members of

the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis. *J Biol Chem* **1997,** *272* (29), 17907-11.

Appendix: Solutions of Caspase-1 Steady-State Kinetics Models

Caspase-1 wildtype kinetics

First, we present a self-consistent analytical solution of the caspase-1 wildtype steady-state kinetic model shown in Fig. X in the text. This is the simplest possible model that includes dimerization in the presence/absence of one or two substrates and up to two-sites binding and catalysis.

$$
\begin{aligned}
\partial_t A(t) =& k_{1r} + k_3 B + 2k_{0off} C + k_{1off} D - (k_{1f} S + 2k_{0on} A + k_{1on} B) A \\
\partial_t B(t) =& k_{1f} SA + k_{1off} D + 2k_{2off} E - (k_{1r} + k_3 + k_{1on} A + 2k_{2on} B) B \\
\partial_t C(t) =& k_{0on} A^2 + (k_{4r} + k_7) D - (k0off + k_{4f} S) C \\
\partial_t D(t) =& k_{4f} SC + k_{1on} AB + (k_{8r} + k_9) E - (k_{4r} + k_7 + k_{8f} S + k_{1off}) D \\
\partial_t E(t) =& k_{8f} SD + k_{2on} B^2 - 2 * (k_{8r} + k_9) * E \\
\partial_t P(t) =& k_3 B + k_7 D + 2k_9 E
\end{aligned}
$$

We will assume that Substrate S is a constant value in our steady-state limit equal to the starting substrate concentration. We will assume that the time derivatives of each species except B are equal to zero and then invoke mass conservation to yield five equations for our five state variables.

$$
\begin{aligned}
\partial_t E(t) = 0 \rightarrow E =& \frac{k_{8f} SD + k_{2on} B^2}{2(k_{9r} + k_9) + k_{2off}} \\
\partial_t D(t) = 0 \rightarrow D =& \frac{k_{4f} SC + k_{1on} AB + 2(k_{8r} + k_9) E}{k_{4r} + k7 + k_{8f} S + k_{1off}} \\
\partial_t C(t) = 0 \rightarrow C =& \frac{k_{0on} A^2 + k4r + k_7 D}{k_{0off} + k_{4f} S} \\
\partial_t At = 0 \rightarrow A =& \frac{(k_{1r} + k_3) B + 2k_{0off} C + k_{1off} D}{k_{1f} S + 2k_{0on} A + k_{1on} B}
\end{aligned}
$$

Aggregating parameters into composite constants,

$$
\begin{aligned}
\kappa &= 2(k_{8r} + k_9) \\
Y &= k_{4r} + k_7 + k_{8f}S + k_{1off} \\
\gamma &= \kappa Y^{-1} \\
\chi &= k_{4r} + k_7 + k_{8f}S + k_{1off} \\
Z &= k_{2off} + \kappa \\
\zeta &= k_{8f}SZ^{-1} \\
\varepsilon &= k_{2on}Z^{-1} \\
\eta &= k_{1r} + k_3 \\
\rho &= \frac{k_{1on}}{k_{1f}S} \\
\psi &= \frac{k_{4r} + k_7}{k_{0off} + k_{4f}S} \\
\sigma &= \frac{k_{0on}}{k0off + k_{4f}S} \\
\alpha &= k_{4f}SY^{-1} \\
\delta &= k_{1on}Y^{-1} \\
\mu &= \frac{\frac{k_{1r}+k_3}{k_{1f}S}}{1 + \rho B} \\
\nu &= \frac{k_{1off} + 2k_{0off}\psi k_{1f}S}{1 + \rho B}
\end{aligned}
$$

Plugging these in the above equations,

$$
\begin{aligned}
E &= \zeta D + \varepsilon B^2 \\
D &= \alpha C + \delta AB + \gamma E \\
C &= \sigma A^2 + \psi D \\
A &= \frac{\mu B + \nu D}{1 + \rho B + \frac{2k_{0on}A}{k_{1f}S}}
\end{aligned}
$$

Substituting, we obtain

$$
D = \frac{B^2\left(\frac{\alpha\sigma\mu^2}{(1+\rho B)^2} + \frac{\delta\mu}{1+\rho B} + \gamma\varepsilon\right)}{1 - \psi - \gamma\eta - \frac{\alpha\sigma\nu^2 D}{(1+\rho B)^2} - \frac{\delta\nu B}{1+\rho B}}
$$

Our strategy to obtain an analytical form of the solution is analogous to the self-consistent field approximation used in numerically obtaining solutions to the Schrodinger equation in quantum mechanics. We (1) bundle terms into aggregated parameters, (2) obtain an analytical solution in terms of the aggregated parameters, (3) neglect some terms in these parameters to obtain a fully analytical

approximate solution in terms of the input kinetic rate constants, (4) replace the terms that were neglected with state populations from the analytical approximate solution, (5) iterate until the solution converges. Then, we have a solution in the self-consistent limit.

Here, we first make the following three approximations for the initial solution:

$$\begin{aligned}
&2k_{0on}A << k_{1f}S + k_{1on}B \\
&1 + \rho B \approx 1 \\
&\alpha\sigma\nu^2 D + \delta\nu B << 1 - \psi - \gamma\zeta
\end{aligned}$$

Then,

$$\begin{aligned}
D =& \phi B^2 \\
E =& (\zeta\phi + \varepsilon)B^2 \\
A =& \mu B + \nu D \\
Etot =& \mu B + \nu\phi B^2 + B + 2\sigma\mu^2 B^2 \\
&+ 2\psi\phi B^2 + 2\sigma\nu^2\phi^2 B^4 + 2\phi B^2 + 2\zeta\phi B^2 + 2\varepsilon B^2
\end{aligned}$$

Assuming the B^4 term is negligible, we can solve this last equation quadratically:

$$\begin{aligned}
\Sigma =& \nu\phi + 2\sigma\mu^2 + 2\phi + 2\zeta\phi + 2\varepsilon + 2\psi\phi \\
B =& \frac{-(\mu+1) + \sqrt{(1+\mu)^2 - 4\Sigma Etot}}{2\Sigma}
\end{aligned}$$

Next, we substitute this value for B above to obtain populations of all the states in terms of total enzyme concentration $Etot$ and the substrate concentration S. Then, we replace the neglected terms that were removed above. Next, for each iteration of the self-consistent procedure, we (1) update the neglected terms with the state populations from the last solution, and then (2) obtain new state populations using the above analytical solution. We found that 10 iterations in practice was more than sufficient to converge the solution in the self-consistent limit.

Caspase-1 "Hemilabeled" model

To construct a simplified model for use in analyzing the "hemilabeled" experiment, we assume a three-state model shown in Fig. X in the text, and that catalysis only comes from the inhibitor-bound dimer.

$$
\begin{aligned}
\partial_t A(t) &= k_{1off} C - k_{1on} AB \\
\partial_t B(t) &= k_{1on} AB + k_{chem} C - k_f SB + k_r C \\
\partial_t Ct &= k_f SB - (k_r + k_{chem}) C \\
\partial_t Dt &= k_c hemC
\end{aligned}
$$

Solving,

$$
\begin{aligned}
\partial_t A(t) &= 0 \rightarrow k_{1on} A^2 = k_{1off} B \\
B &= \frac{k_{1on} A^2}{k_{1off}} \\
\partial_t C(t) &= 0 \rightarrow C = \frac{k_f SB}{k_r + k_{chem}} \\
K_M &= \frac{k_r + k_{chem}}{k_f} \\
K_D &= \frac{k_{1off}}{k_{1on}} \\
\Sigma &= \frac{k_{on}}{k_{off}} \left(1 + \frac{k_f S}{k_r + k_{chem}}\right) \\
\Sigma &= \frac{1}{K_D} \left(1 + \frac{S}{K_M}\right) \\
\frac{1}{2} Etot &= A + B + C \\
\frac{1}{2} Etot &= A + \frac{k_{1on} A^2}{k_{1off}} + \left(\frac{k_f S}{k_r + k_{chem}}\right) \frac{k_{1on} A^2}{k_{1off}} \\
0 &= \Sigma A^2 + A - \frac{1}{2} Etot \\
A &= \frac{-1 + \sqrt{1 + 2\Sigma Etot}}{2\Sigma} \\
B &= \frac{2 - 2\sqrt{1 + 2\Sigma Etot} + 2\Sigma Etot}{4\Sigma^2 K_D} \\
\partial_t Pt &= \frac{k_c hem k_f S}{k_r + k_{chem}} \left(\frac{2 - 2\sqrt{1 + 2\Sigma Etot} + 2\Sigma Etot}{4\Sigma^2 K_D}\right) \\
\partial_t Pt &= \frac{k_c hem S}{K_M} \left(\frac{2 - 2\sqrt{1 + 2\Sigma Etot} + 2\Sigma Etot}{4\Sigma^2 K_D}\right)
\end{aligned}
$$

Chapter 7. Conclusion

The novel information theory-based molecular dynamics simulation analysis methods described in this work, namely Mutual Information and Kullback-Leibler Divergence, are part of a software package called "MutInf", available online through Stanford's simulation toolkit "SimTK" (https://simtk.org/home/mutinf) . Now the Mutual Information (and later Kullback-Leibler Divergence) code is available online through the following Subversion repository: " svn checkout https://simtk.org/svn/mutinf ". MutInf is written in python using NumPy and SciPy, and features inline C code using the "weave" package to optimize some of the high-frequency loops. An online manual is available here: http://www.jacobsonlab.org/mutinf_manual/ .

There are several natural extensions that could be made to MutInf. First, the Kullback-Leibler Divergence needs the capability to handle Cartesian coordinates – this is difficult because of the "frame-fitting", alignment-and-superposition problem, and practically challenging in this histogramming framework due to fixed and different bounds for different datasets. This could be relieved by using a histogram-independent, K-nearest-neighbors approach for the Kullback-Leibler Divergence. Furthermore, the code does not yet handle the second-order version of the Kullback-Leibler Divergence, though initial efforts towards this were made. Another code development effort focused on "weighted" ensembles, such as from replica exchange or from a Markov Model, where each data point has a different weight attached to it. A foundation for this is in a branch of the code that needs to be tested then merged in appropriately.

One of the obvious applications of this work is to identify new allosteric sites for drug discovery, and to help predict how a potential allosteric molecule might alter the protein's conformational ensemble and thereby inhibit or activate the protein. Often, these allosteric sites are protein-protein interaction sites or, if they are on the surface, may share commonalities with protein-protein interaction sites. This work finishes with a review co-authored with Jim Wells on small molecule inhibitors of protein-protein interactions, as some of the same principles may apply to novel allosteric compound binding sites on the surfaces of proteins.

There is probably no other class of macromolecular interactions that rivals the complexity, diversity, and regulatory impact of protein-protein interfaces (PPIs)[1-6]. Although there is tremendous therapeutic interest in PPIs, these interfaces present huge challenges for small molecule drug discovery. Protein-protein interfaces are large (~1500-3000 $Å^2$)[7, 8] compared to typical small molecule contact interfaces(~300-1000 $Å^2$)[9, 10]. Protein-protein interfaces are generally flat and often lack the grooves and pockets seen for proteins known to bind small molecules[11]. Unlike classic small molecule targets such as enzymes or G-protein coupled receptors (GPCR's), these interfaces naturally bind a large protein instead of a small molecule partner. Thus, drug discovery efforts do not have the luxury of starting from a small natural substrate or ligand. Most PPIs involve protein residues that are not continuous in the polymer chain. Short contiguous peptides taken from the interface are rarely good chemical starting points. There are some notable exceptions where the protein partner presents a contiguous tri- or tetra-peptide sequence for which small molecule peptide mimetics have been built (for examples see[12, 13]). High through-put screening (HTS) does not routinely identify

compounds that will disrupt PPIs[14, 15]. Although biopharmaceuticals such as monoclonal antibodies or polypeptide hormones almost always bind to protein-protein interaction surfaces, there are few approved small molecule drugs for this target class.

Despite these challenges, several lines of evidence suggest there may be hope for finding small molecules for PPIs after all. Although these interfaces are large, mutational studies show a small subset of residues contributes the majority of binding free energy[16-19] (Figure 1, for a recent review see[20]). Such "hot-spots" are usually centered in the contact interface and represent less than half of the contact surface. PPIs can be promiscuous, and bind several other targets using the same hot-spot region[21]. Structural studies show that adaptability at these promiscuous interfaces allows one partner to engage structurally diverse partners. Peptides selected in phage display experiments for binding to one of the partners in a protein-protein pair will often compete with the natural partner for binding to the hot-spot[20-24]. Thus, there appear to be many chemical solutions for tight binding and these large interfaces can be engaged by more compact scaffolds.

How small can we go

Research in the area of finding small organic compounds that disrupt PPIs has made significant strides in the past 5 years (for recent reviews see [25-28]). Here, we focus on six recently published examples of discontinuous protein-protein interfaces where small molecules have been discovered that directly compete with one of the protein partners (Table I). This set has publicly available crystal structures for both the protein-protein and protein-small molecule complexes in the Protein Data Bank (PDB). These case studies provide a unique opportunity to structurally compare how a small molecule directly competes with a natural protein partner and provide instructive patterns for how

to advance drug discovery in this important field. We compare these protein-protein complexes to their competing protein-small molecule complexes in Fig. 2. and Supplemental Movies 1-6.

A. IL-2 binders: Il-2 is a key cytokine involved in T-cell activation and a causative agent in graft rejection. A series of small molecules with dissociation constants in the mid-nM range (K_i = 68.5 nM, K_D = 100 nM) were produced at Sunesis Pharmaceuticals that bind to IL-2 and disrupt the interaction between IL-2 and the IL-2α receptor[29-31]. These were assembled in a fragment-based approach guided by X-ray structures and medicinal chemistry, and inspired by previous discovery efforts by Tilley and co-workers at Roche Pharmaceuticals[32]. Although the molecules were built before the structure of the IL-2/ IL-2α receptor complex was reported[33], they bind near the center of the receptor contact interface[34, 35] (Figure 2a) and do not bind to IL-15, the closest homolog of IL-2[31].

A number of interesting features emerge when we compare the structural and functional epitopes for binding the small molecule (SP4206) versus the large receptor[19] (Fig. 2, http://www.jacobsonlab.org/mutinf_manual/wells_mcclendon2007_s1.mov (Supplementary Movie 1)). The contact epitope for the small molecule is about half the size of that for the receptor. Because the small molecule and receptor bind with nearly equivalent affinity, the ligand efficiency (binding free energy per heavy contact atom[36]) for the small molecule is more than twice that for the larger receptor (Table I). When IL-2 binds the α-receptor, the binding epitope on IL-2 is very flat as is typical of protein-protein complexes. In contrast, the small molecule traps a groove in IL-2 and repositions a loop to embrace the furanoic acid moiety at one end of the small molecule. Alanine-

scanning mutational studies show the small molecule and receptor bind the same hot-spot residues on IL-2[19]. Although the structures of the small molecule and α-receptor are very different, the electrostatic potential surfaces presented are quite similar and likely reflect a need to establish electrostatic complementarity with IL-2. Electrostatic and surface shape complementarity[37, 38] plus specific hydrogen-bonding interactions likely account for the high selectivity of these interactions.

These studies show that the binding interface on IL-2 is very adaptive and can bind a small molecule with good affinity using the same primary hot-spot residues. It is notable that the design of these compounds, driven mostly by functional SAR (Structure-Activity Relationships) and structure, did not require the structure of the bound receptor. Indeed, the small molecule is not a faithful atomic mimic of the receptor, and would not have been found assuming one needed to capture the precise structure of the receptor bound form of IL-2.

B. Bcl-x_L binders: Bcl-2 family members are important regulators of apoptosis[39, 40]. These can form homo- and hetero-dimers with pro- or anti-apoptotic relatives. Bcl-2 and Bcl-x_L inhibit apoptosis by binding a 16-residue α-helical portion of pro-apoptotic Bak[40] or a 26-residue α-helical portion of pro-apoptotic Bad[41] (Figure 2b). The importance of these targets in cancer has generated considerable interest in synthetic inhibitors. Several groups have produced smaller helical peptide mimics with high affinities (K_i ~5-100 nM in the best cases)[42-45]. Recently, a team at Abbott Laboratories generated high-affinity organic compounds (K_i = 0.6 nM; molecular weight MW ~880) that bind to the hydrophobic helical binding domain of Bcl-x_L, Bcl-2, and Bcl-w but not to other homologs such as Mcl-1 or Bcl-B[46] (Table I). The affinity of the small molecule

is comparable to that of the peptide, and given its smaller contact interface, the ligand efficiency is more than two-fold higher. The compounds were discovered using a fragment-based NMR method (SAR-by-NMR) and advanced using NMR structure-guided medicinal chemistry[47-49]. The compounds showed significant cell-based activity, good activity in mice with tumor xenografts, and synergy with radiation and a number of other chemotherapeutics. A derivative of ABT-737 (ABT-263) is currently in Phase I/II cancer clinical trials (S. Fesik, personal communication).

NMR structures of the bound fragments (K_D's ~ 0.3-4 mM) show reasonable correspondence to the elaborated high affinity compound ABT-737 (Fig. 2) and analogs (http://www.jacobsonlab.org/mutinf_manual/wells_mcclendon2007_s2.mov (Supplementary Movie 2)). However, compared to the α-helical peptide there are some striking differences (Figure 2b). Alanine-scanning of the Bak peptide identified several critical residues: Leu-78, Asp-83, Val-74, and Ile-81, and Ile-85[40, 45]. Although the small molecule binds where these peptide residues fit, it does not closely mimic the atomic details of peptide partner. Instead, the small molecule traps a slightly different conformation of Bcl-$x_{L,}$ binding in deeper cavities with more puckered groves. Thus, the small molecule and peptide bind to the same hot-spot region but the small molecule binds deeper and in a slightly different way.

C. Hdm-2 binders: Hdm-2 emerged as an excellent drug target in cancer when it was found that the mouse homolog (Mdm-2) binds to the tumor suppressor p53 and enhances its degradation, thus blocking p53 tumor suppressing transcriptional activity (for review see [50]). Mdm-2 can bind a 15-residue α-helical region of p53 (K_D ~ 600nM) and the structure of the complex shows a largely hydrophobic interface[51]. Alanine-

scanning mutational analysis of the peptide identified three dominant determinants in the center of the interface: Leu22, Phe19 and Trp23[52]. A high through-put screening and medicinal chemistry effort at Roche in Nutley, NJ identified a series of tetra-substituted imidazoles, termed Nutlins (Table I). After considerable chemical optimization, the most potent of these disrupted the interaction with a IC_{50} of 90 nM, and showed potent cellular and xenograft activity[53]. At Johnson and Johnson, a parallel screen of 338,000 compounds that monitored binding by changes in thermostability (ThermoFluor) identified a series of benzodiazepinediones[54]. After chemical optimization one of these was found with a K_D of 67 nM and an IC_{50} of 420 nM[55]. Although these compounds were initially selected for binding to Hdm-2 and not for functional disruption of the complex, they promoted rapid dissociation of p53 from Hdm-2 in cells overexpressing Hdm-2[56]. Furthermore, a benzodiazepinedione with an added solubilizing moiety inhibited cell proliferation with an IC_{50} of ~10μM and showed synergistic activity with doxorubicin in mice[56].

The structures of both small molecules have been solved in complex with Hdm-2: Nutlin-2 by X-ray crystallography[53], Nutlin-3 by NMR[57], and the benzodiazapinedione by X-ray crystallography[54] (Figure 2c, http://www.jacobsonlab.org/mutinf_manual/wells_mcclendon2007_s3.mov (Supplementary Movie 3)). Both compounds bind precisely over the p53 peptide site and insert aromatic or aliphatic moieties into hot-spot pockets on Hdm-2 that bind key p53 peptide residues Phe 19, Trp 23 and Leu 26. The contact epitopes for the small molecules are again about half the size of the minimal peptide binding epitope. The conformation of Hdm-2 is more open at the ends for binding the peptide whereas it closes more tightly

over the small molecules, presenting a more concave binding site as in IL-2 and Bcl-x_L. It is remarkable that these dissimilar small molecule scaffolds, discovered from very different starting points, have converged on a very similar binding mode.

D. E2 binders: Papallomavirus (HPV) is of significant medical interest, as it is the causative agent in warts and is a malignant agent in cervical cancers. There is currently no small molecule therapeutic. The interaction between the viral E2 transcription factor and the E1 helicase is critical for the viral life cycle and thus is an important PPI target. A group at Boehringer-Ingelheim identified a class of indandiones by HTS that weakly disrupted this interface (K_D ~20μM)[58]. Medicinal chemistry efforts further optimized the compounds' affinity, achieving IC_{50}'s as low as 6 nM (Table I)[59-61]. Direct binding of [^{3}H]-labeled compounds and titration calorimetry showed the compounds bound to the transactivation domain of the E2 protein with one-to-one stoichiometry. Interestingly, an X-ray structure of a compound bound to the transactivation domain showed two copies of the small molecule; one penetrates into a cavity in the three helix domain and the other sits on top[60].

Soon after the release of this X-ray structure, the structure of the HPV 18 E2-activation domain bound to the E1 helicase was reported[62]. The protein-protein interface completely covers the small molecule-E2 contact interface in HPV 11 (Figure 2d). Of the 20 residues that E2 uses to contact E1, the small molecule contacts only 7. Importantly, the small molecule accesses a cavity not seen in the protein-protein interface (http://www.jacobsonlab.org/mutinf_manual/wells_mcclendon2007_s4.mov (Supplemental Movie 4)). The small molecule achieves higher ligand efficiency than E1

(Table I), presumably by deeply burying its hydrophobic surface area rather than spreading it across the interface.

E. ZipA binders: Separation of bacterial cells during division depends on a septal ring composed of at least two proteins in some gram-negative strains: FtsZ, a homolog of eukaryotic tubulin, and ZipA, a membrane-anchored protein. These form a protein-protein complex using their C-terminal domains. A high resolution X-ray structure (1.5Å) with a 17-residue peptide from the C-terminus of *E. coli* FstZ shows that the peptide binds to a cavity in ZipA as an extended β-strand followed by an α-helix[63] (http://www.jacobsonlab.org/mutinf_manual/wells_mcclendon2007_s5.mov (Supplemental Movie 5)). The FtsZ peptide binds about 100 times weaker than the full-length FtsZ (~7μM), but it serves as a useful surrogate. Although ten of the side-chains of the peptide make direct interactions with ZipA, alanine-scanning mutagensis shows that only four of these (three hydrophobic and one acidic) dominate the binding affinity and constitute a hot-spot. The structure of the unbound form of ZipA was also reported and found to have a similar structure as the bound ZipA except for notable differences in some side-chains that allow the critical hot-spot residues from the peptide to penetrate more deeply into the surface of ZipA. Thus, ZipA appears to be an adaptive surface that accommodates the binding of hot-spot residues.

A NMR fragment screening effort of a diverse set of 825 compounds at Wyeth Pharmaceuticals yielded seven hits that bound to the FtsZ site[64]. Even though this represents a high hit rate of 0.8%, which may suggest this target to be druggable[65], extensive medicinal chemistry and SAR from selected hits was unsuccessful in achieving significantly high potency[66]. In the search for other small molecule possibilities, an HTS

of 250,000 compounds identified a pyridyl-pyrimidine which had a K_i of 12μM (Table I) and whose X-ray structure was solved[67]. The structure shows it binds entirely in the hot spot-region and only contacts 740Å^2 compared with 1350Å^2 for the 17-mer peptide (Supplemental Movie 5). Though the surfaces of the small molecules are more complementary to ZipA's surface than the surface of the peptide these molecules were not able to penetrate deep into the ZipA surface.

F. TNFα disruptors: TNF represents a major drug target as it is a key cytokine that drives inflammation. Important biological therapeutics are approved for this target for treating arthritis. Not surprisingly, there has been considerable interest in developing small molecules or peptides that can disrupt the interaction between TNF and its TNF–receptors (TNFR-1 and TNFR-2). For example, small peptides (13-mers) taken from TNFR-1 were found that could bind modestly to TNF (K_D ~ 5μM)[68] and small molecule photo-active inhibitors have been discovered that label a site near the receptor binding site[69].

More recently[70] another class of small molecule inhibitors was discovered that disrupt the trimeric cytokine (K_D ~13 μM, Table I) by binding and displacing one of the monomers from the timer. These compounds, discovered by fragment screening, bind to an adaptive cluster of tyrosine residues at the core of the timer interface (Figure 3a). Two aromatic groups from the compound occupy the position of the tyrosine residues from the displaced monomer (http://www.jacobsonlab.org/mutinf_manual/wells_mcclendon2007_s6.mov (Supplemental Movie 6)). Although these compounds are too weak to be serious drug candidates, they illustrate that even constitutive oligomeric interfaces can bind small

molecules. In another example, small molecule inhibitors of survivin have recently been discovered that bind within survivin's homodimer interface[71].

It is known that TNF monomers can exchange with other momoners in the trimer, albeit slowly. Remarkably, these small molecules *enhance* the dissociation kinetics of the displaced monomer by over 600-fold. Thus, the compound need not wait for a monomer to completely dissociate (Figure 3b, Model 1); it can actually intercalate into the dynamic trimer complex and displace the monomer (Figure 3b, Model 2). Presumably, breathing motions allow the small molecule to intercalate the interface and prevent the displaced monomer from reforming a high-affinity complex with the remaining dimer.

Myths about disrupting protein-protein interfaces

"These are large flat interfaces without cavities for small molecules to dock".

All of the interfaces above show some adaptability that opened up cavities not seen in the free protein. Most of this flexibility involves side-chain motions and small loop perturbations. In each case in Fig. 2, the small molecule or fragment accesses smaller pockets or grooves that the larger and more constrained protein or peptide does not. Thus, one should not assume that the best small molecule site is seen from static structures of either the free protein target or even the protein-protein complex. For example, Bcl-x_L appears to have a rather flat surface in the static apo structure, but during molecular dynamics simulations of less than one nanosecond, transient pockets open up[72, 73]. Similar transient binding pocket openings were found in simulations with IL-2 and Hdm-2[73].

"Screening does not work for PPIs".

In fact, all of the examples presented involved empirical screening, both fragment screening and HTS. In several examples, the starting compounds were identified by HTS using large numbers of compounds (>250,000) to identify very weak hits (K_i's in the mid-μM range). Extensive biophysics was applied to validate that these hits were "real" and stoichiometric prior to investment of medicinal chemistry. In four of the cases cases presented here, medicinal chemistry advanced these hits to compounds with K_D's of mid to low-nM, and in two cases they did not. The ability to progress a hit was not well predicted by the initial compound behavior or inspection of the site, but may be suggested by hit rates from fragment libraries[65] or by druggability indices[9] applied to conformational ensemble samples from computer simulations[72].

One reason that HTS may not be more successful is that the compounds used for screening are derived mostly from historical medicinal chemistry efforts in pharmaceutical companies. These chemotypes have been dominated by past drug discovery on GPCR's, enzyme targets, and traditional "druggable" targets. Every new target class is seeded invariably by new chemotypes that are somewhat distinct from the targets that preceded it. Is it possible that PPI inhibitors require different chemotypes? As a small-scale analysis, we took high-affinity protein-protein inhibitors from the IL-2, Bcl-x_L, Hdm-2, and E2 class and compared these to sets of compounds directed against targets in the MDDR and WOMBAT chemical databases using a compound Similarity Ensemble Approach (http://sea.docking.org)[74]. The PPI inhibitors did not show high similarity to any set of compounds against other known targets. Thus, if traditional libraries are used, large compound collections may be required to find bonafide[75] hits with K_D's in the 10-100 μM range. Moreover, one should not assume there are a few

privileged scaffolds that will unlock this entire target class as there has been for protein kinases and GPCR's. Except for close homologs, each PPI is different and thus chemotypes are likely to be more isolated in chemical space.

It is possible that fragment screening will be more successful than HTS when applied to PPI targets. Although anecdotal, several successes did come from fragment screening even though there probably have been far fewer fragment screens than HTS screens. In theory, fragments (MW 150-250) have higher ligand efficiencies than typical HTS compounds (MW 400-500) and enable a greater search of compound diversity space per atom[36, 76].

"The native protein-protein complex is tighter and cannot be competed away once formed".

In most of the cases, the optimized small molecule bound with an affinity comparable to that of the partner protein or peptide. In several examples (IL-2, Hdm-2, E2), compound K_i or IC_{50} values measured by competitive inhibition of the compounds were in the mid to low-nM range and comparable to the compound binding affinity (K_D) measured by direct binding methods. This would indicate that under equilibrium conditions the small molecule is not disadvantaged to displace the protein partner.

From a kinetic perspective the small molecule may actually have an advantage over a large protein competitor like an antibody. For example, in the TNF case the compounds actually accelerated the dissociation of the monomer from the complex by >600-fold. Thus, inhibition was not rate-limited by the off-rate of a TNF monomer. It would be very interesting to see if the other small molecule inhibitors can accelerate dissociation of their protein-protein interaction partner. Recent paramagnetic NMR studies on protein

complexes suggest that protein-protein interfaces may be inherently "wobbly"[77, 78]. If this were generally the case, a small molecule may be expected to penetrate these dynamic "encounter" complexes and have a kinetic advantage over a large antibody therapeutic, whose association depends on complete (not partial) dissociation of the competing protein partner.

"Small molecules to protein-protein interfaces will be too large to be drugs".

Most orally active drugs have molecular weights below 500, and neurological drugs usually require even lower molecular weights to cross the blood brain barrier [79, 80]. Such rules, derived from the limited set of known drugs, have notable exceptions such as cyclosporine (MW ~ 1000). In fact, ABT-737 (MW ~880; Table I) has a respectable 70% bioavailability in rodents[46], and a derivative (ABT-263) of comparable size has begun clinical trials. Moreover, many useful drugs including many antibiotics and cancer drugs are given as injections where molecular weight considerations are not driven by oral bioavailability.

There is always a trade-off between compound binding affinity and other properties such as pharmacokinetics, solubility, toxicity, and synthetic tractability that together determine the likelihood that a compound will ultimately make it as a drug. For these latter properties, lower molecular weight is clearly better. All of the PPI inhibitors described here with K_i values < 1μM have molecular weights in the range of 500-900. We therefore wondered if there is a limiting relationship between compound potency and size for small molecule PPI inhibitors. For this analysis, we selected only compounds with extensive medicinal chemistry data and those where structures were solved showing the compounds or close analogs bound to their targets. We plotted the free energy of

binding (kcal/mol) vs. the number of heavy atoms for highest affinity fragments and optimized compounds against these target proteins (Figure 4). These data form a reasonably linear plot with a correlation coefficient of 0.88. It is remarkable that all of these very different targets with different chemotypes share similar ligand efficiencies. The slope of the line gives a ligand efficiency (LE) value of 0.23 kcal/mol per non-hydrogen atom. This LE value is considerably below that for the tightest binding small molecules (~1.5)[36] but not far from many other kinase inhibitors (LE = 0.3-0.4 kcal/mol per non-hydrogen atom) and comparable to many protease inhibitors (LE ~ 0.25-0.35 kcal/mol per non-hydrogen atom)[36, 81]. A survey of a number of less-optimized small molecule protein interface inhibitors shows, with some exceptions, ligand efficiencies similar to those here (Table II). Assuming a value of 0.23 kcal/mol per non-hydrogen atom, a compound with a K_D of 10nM, typical of many drugs, would require roughly 47 non-hydrogen atoms (MW ~ 660). We suggest that medicinal chemistry efforts that exceed this curve are doing exceptionally well and those that are significantly below this curve have significant optimization to do if nanomolar affinity and oral bioavailability are desired.

Prospects and challenges for drug discovery at PPIs

There is always a trade-off between compound binding affinity and other properties such as pharmacokinetics, solubility, toxicity, and synthetic tractability that together determine the likelihood that a compound will ultimately make it as a drug. For these latter properties, lower molecular weight is clearly better. All of the PPI inhibitors described here with K_i values < 1μM have molecular weights in the range of 500-900. We therefore wondered if there is a limiting relationship between compound potency and

size for small molecule PPI inhibitors. For this analysis, we selected only compounds with extensive medicinal chemistry data and those where structures were solved showing the compounds or close analogs bound to their targets. We plotted the free energy of binding (kcal/mol) vs. the number of heavy atoms for highest affinity fragments and optimized compounds against these target proteins (Figure 4). These data form a reasonably linear plot with a correlation coefficient of 0.88. It is remarkable that all of these very different targets with different chemotypes share similar ligand efficiencies. The slope of the line gives a ligand efficiency (LE) value of 0.23 kcal/mol per non-hydrogen atom. This LE value is considerably below that for the tightest binding small molecules (~1.5)[36] but not far from many other kinase inhibitors (LE = 0.3-0.4 kcal/mol per non-hydrogen atom) and comparable to many protease inhibitors (LE ~ 0.25-0.35 kcal/mol per non-hydrogen atom)[36, 81]. A survey of a number of less-optimized small molecule protein interface inhibitors shows, with some exceptions, ligand efficiencies similar to those here (Table II). Assuming a value of 0.23 kcal/mol per non-hydrogen atom, a compound with a K_D of 10nM, typical of many drugs, would require roughly 47 non-hydrogen atoms (MW ~ 660). We suggest that medicinal chemistry efforts that exceed this curve are doing exceptionally well and those that are significantly below this curve have significant optimization to do if nanomolar affinity and oral bioavailability are desired.

In the past 5 years there has been remarkable progress in identifying, characterizing, and developing small molecules that bind to protein interface sites. In addition to the interfacial inhibitors presented here, it is also possible to inhibit PPIs through allosteric sites[82, 83], and by promoting abherent protein-protein interactions (e.g..

by cyclosporins)[84]. However we still have a long way to go. It is not clear that we are screening the optimal region of compound space or that compounds we find can be easily optimized for these diverse interfaces. Fragment screening methods offer the greatest opportunities to cover a wider swath of synthetically-feasible chemical space per atom. Hot-spots enable ligand-efficient 'footholds' to be established by initial fragments. However, except for Hdm-2, hot-spot binding alone did not yield high-affinity inhibitors; in IL-2, Bcl-x_L, E2, and ZipA, additional sources of small-molecule affinity were needed and subsequently found, except in ZipA. The highest-affinity small molecules engaged residues that the natural protein partner did not interact with, often exploiting 'cryptic' pockets. If fragments are to be used, we need new sensitive, low-cost, high-throughput fragment screening technologies for such screens to become commonplace, and possibly the biggest challenge is growing fragments that bind into higher affinity small molecules. Improved computational methods to design ligand-efficient elaborated compounds at flexible protein sites would be quite helpful in focusing medicinal chemistry efforts against these adaptive targets. To develop such methods, one would want to know whether these compounds "induce" conformational changes or whether they "select" a conformation of the target protein or endogenous complex from an ensemble of states sampled during the normal dynamic excursions these proteins and complexes make at physiological temperatures. A recent study that docked high-affinity inhibitors of IL-2, Bcl-x_L, and Hdm2 to protein conformational snapshots from 10ns Molecular Dynamics simulations found ligand poses docked to some snapshots that were close to those observed experimentally, with corresponding protein conformations that were roughly similar to those observed in inhibitor-bound structures. These results suggest that most

but not all of the conformational differences seen when comparing apo to inhibitor-bound structures are due to conformational selection by the ligand. Moreover, the structural changes seen at these protein-protein interfaces are smaller than those that appear in biologically evolved examples of induced-fit such as in hexokinase[85-87].

If one accepts that this class of proteins generally has a lower ceiling for ligand efficiency than more traditional targets, the drug discovery community we will have to get better at managing the ADME properties of larger compounds. Though these compounds are larger than typical drugs, the compounds were quite specific for their targets, as was the case for IL-2, Bcl-x_L, E2, Hdm-2. The compelling biology surrounding PPIs and the fact that more small molecules are winding their way through clinical trials gives us hope that we may have more of these drugs on the shelf in the future. Clearly, the new efforts have moved us a rung higher toward reaching this class of high hanging fruit.

Figure 1. Examples of "hot-spots" based upon alanine-scanning mutational analysis of four protein-protein interfaces. The effect of the alanine mutation on the free energy of binding relative to wild type ($\Delta\Delta G$) is color coded from red (most disruptive--hot spots) to dark blue (little or no effect). Figure courtesy of W. DeLano[98].

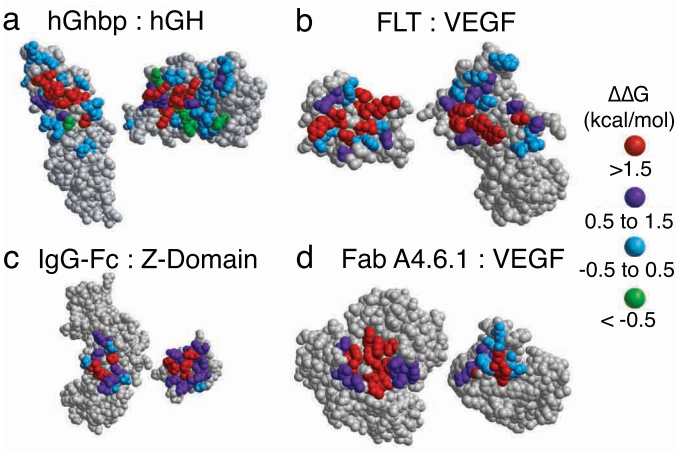


Figure 2. Four examples (Panels **a-d**) comparing how a protein binds its natural protein or peptide partner relative to an unnatural small molecule. The left column shows the structure of the protein-protein or protein-peptide complexes where the target protein is rendered in grey filled surface and the binding protein or peptide is shown in yellow ribbons with selected side chains in sticks. The contact surface (within 4.5Å of the binding partner) is shown in green. The right column shows the structure of the small molecule in yellow sticks bound to the protein rendered in gray surface and the contact interface shown in orange. The middle column shows the small molecule (yellow) superimposed onto the surface of the protein in the conformation used to bind its natural protein or peptide partner, whose contact surface is shown in green. Note how much larger and flatter the protein-protein contact surface (green) is compared to the small molecule-protein contact surface (orange). Panel **a** compares Il-2/IL-2 a receptor vs. IL-2/small molecule (SP4206). Panel **b** compares Bcl-x_L /Bad peptide vs. Bcl-x_L /small molecule (ABT737). Panel **c** compares Hdm-2/p53 peptide vs. Hdm-2/Nutlin-2 (top) or Hdm-2/benzodiazepinedione (bottom). Panel **d** compares HPV-18 E2/E1 vs. HPV-11 E2/B.I. cmpd23. The middle column is not shown for Panel **d** since HPV-18 and HPV-11 are not identical but are homologs.

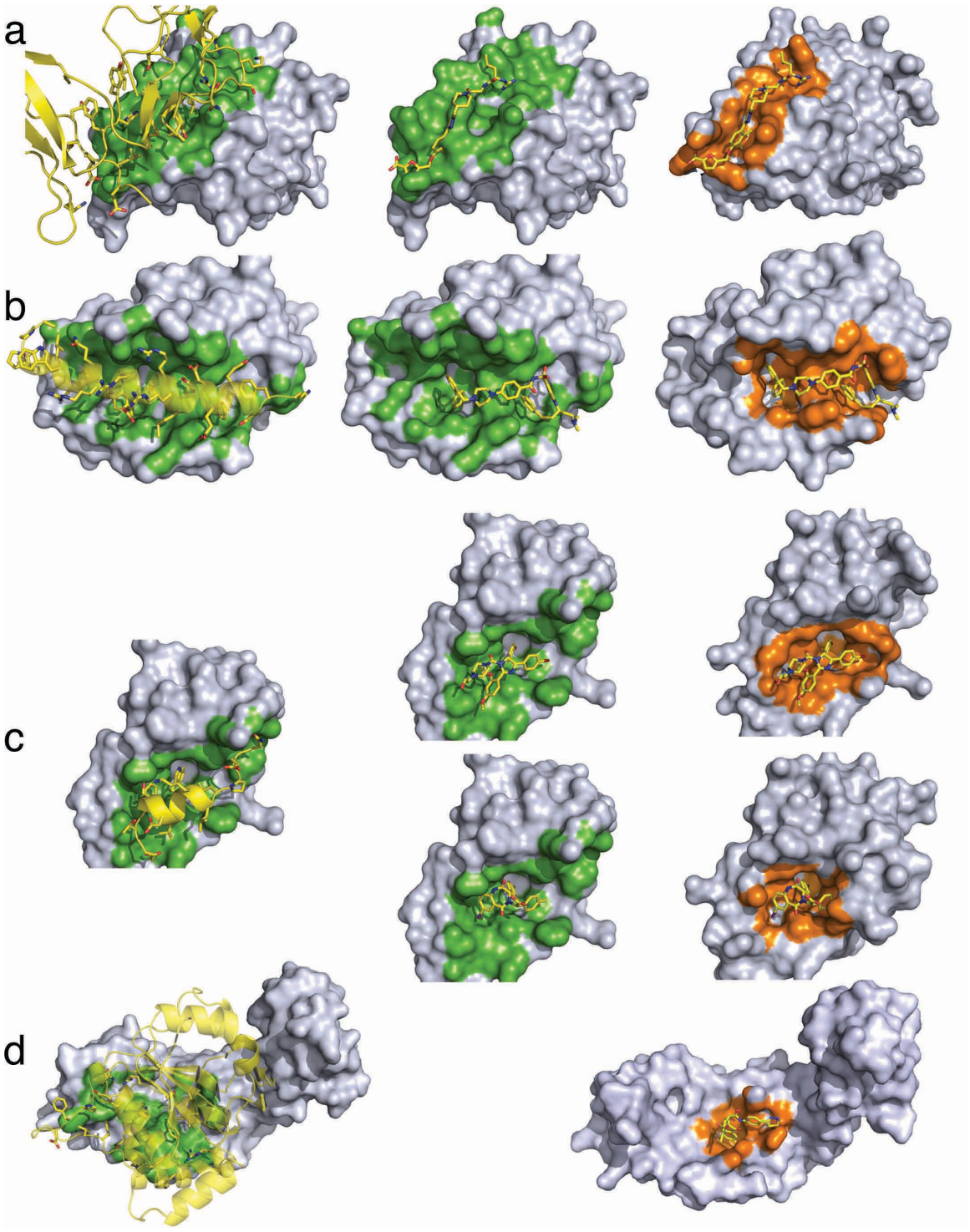

a
b
c
d

Figure 3. Panel **a**. Structure of the TNF timer versus the TNF dimer/SP403 small molecule. Panel **b**. Two models for how small molecules could block formation of TNF trimers. Model 1 requires complete dissociation of one of the monomers before the small molecule can bind. Model 2 allows the small molecule to associate with the trimer and facilitate dissociation. The fact that the small molecule accelerates the rate of dissociation of the monomer (by >600-fold) supports Model 2.

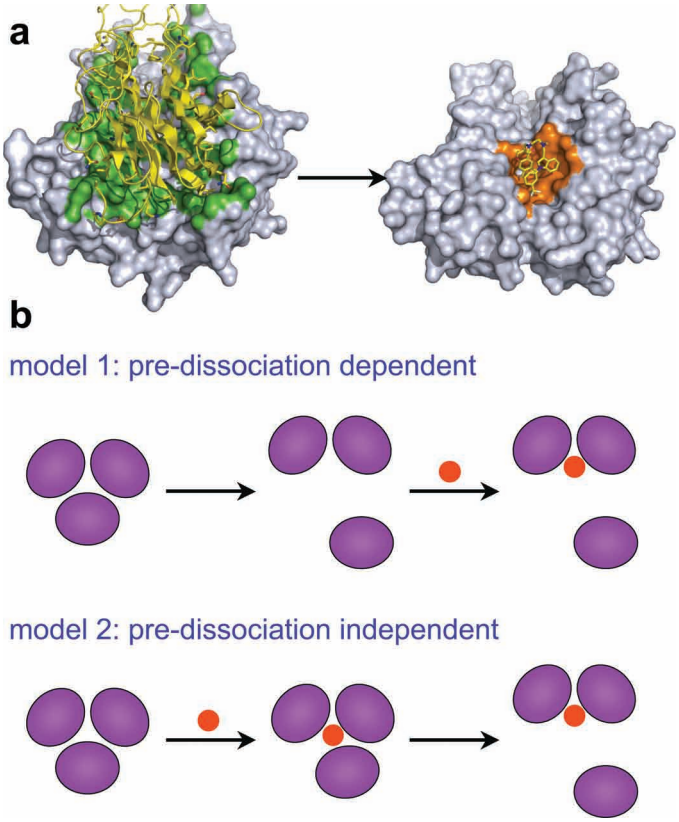


Figure 4. Plot of binding free energy versus number of heavy (non-hydrogen) atoms in highest affinity fragments and small molecules for the protein-protein interfaces. K_D values were converted to free energy (kcal/mol) using standard-state conditions of 1 M concentration at a temperature of 300K. Where direct binding affinity was not available, K_i or IC_{50} was used as an estimate. The slope can be described by $y = 0.24x$, and the correlation coefficient is 0.77. The linear relationship implies that there is a uniform ligand efficiency for these targets. Inhibitors: **IL-2:** Ro26-4550(X)—Roche, SP4206(♦)—Sunesis. **Bcl-x_L:** Biphenyl fragment(*), Napthalenyl fragment (+), ABT-737(■)—Abbott. **Hdm-2:** Nutlin-3(●)—Roche; Benzodiazepine dione(■)—Johnson&Johnson. **E2:** Compound 23(♦)—Boehringer Ingleheim. **ZipA:** Hexahydroquinolizinone fragment (○), Compound 1(◊)—Wyeth. **TNFα**: SP403(Δ) —Sunesis-Biogen-Idec. **Survivin**: Compound 1(□), Compound 23b(▲)—Abbott.

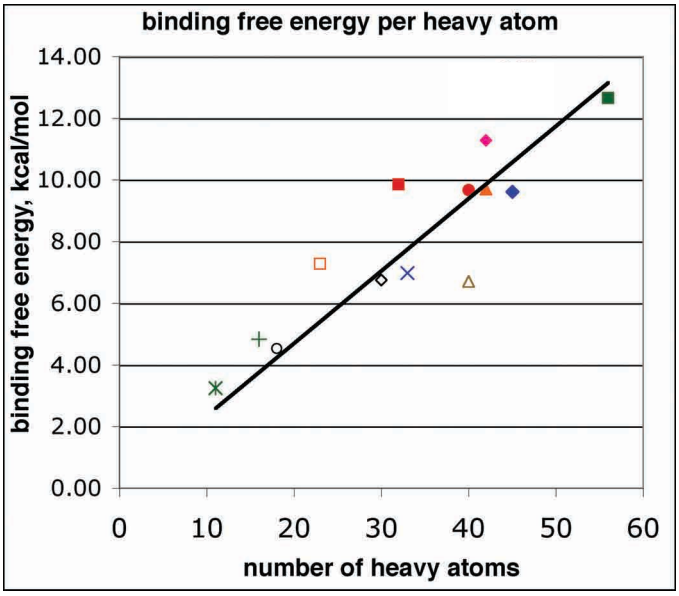


Table I. Comparison of Protein vs. Small Molecule Binding Partners. Examples of six proteins whose complex structures have been solved both with their natural protein partner and with the small molecule that binds them, and for which affinities have been measured. Molecular weights for each are given in units of Daltons. Ligand efficiency (LE) values for the protein-protein pair are given as binding free energy ($-\Delta G$) per non-hydrogen contact atom because so little of the protein is actually in contact. LE values for the small molecule are given as $-\Delta G$/non-hydrogen atom[36]. *Ligand in X-ray structure is very similar to the compound shown.

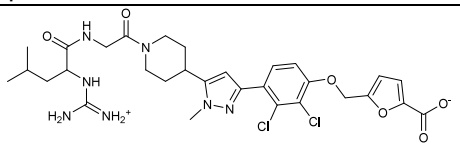
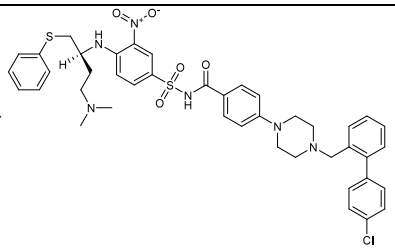
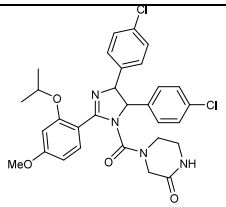
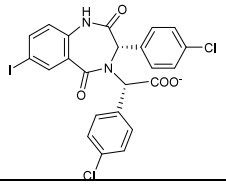
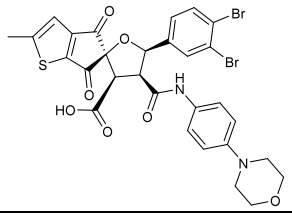
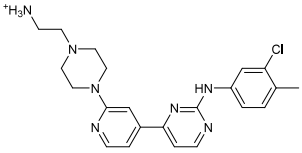
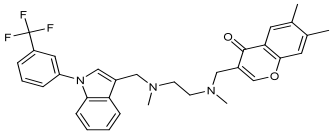
Table I		Comparison of Protein vs. Small Molecule Binding Partners			
Receptor	PDB	Ligand	Molec. Weight	Affinity (μM)	LE(kcal/mol per non-H atom)
IL-2	1Z92	IL-2a receptor[33]	24790	0.0105	0.11
IL-2	1PY2	SP4206[31] 	663	0.1	0.21
Bcl-x_L	2BZW	Bad peptide[88]	3110	0.0006	0.16
Bcl-x_L	2YXJ	ABT-737[47] 	813	0.0006	0.23
HDM-2	1YCR	p53 (peptide 15-29)[51]	1808	0.6	0.12
HDM-2	1RV1*	Nutlin-3[53] 	581	0.09	0.24
HDM-2	1T4E	Benzodiazepine dione[55] 	566	0.067	0.31
HPV E2	1TUE	E1[62]	24630	0.06	0.14
HPV E2	1R6N*	compound 23[61] 	684	0.006	0.28
ZipA	1F47	FtsZ (peptide 367-383)[63]	2024	21.6	0.13
ZipA	1Y2F	compound 1[67] 	425	12	0.23
TNFα	1TNF	Subunit protein	17381	ND	ND
TNFα	2AZ5	SP403[70] 	548	13	0.17

Table II. Ligand Efficiencies of Additional Small Molecule Protein Interface Inhibitors. The Ligand Efficiencies (LE) for a number of additional sub-optimal or un-optimized small molecule protein interface inhibitors from the recent literature generally fall near the 0.23 kcal/mol per non-hydrogen atom trend for highest affinity fragments and small molecules shown in Fig. 4.

Table 2	**Ligand Efficiencies of Additional Small Molecule Protein Interface Inhibitors**			
Target	Compound	Affinity (μM)	LE (kcal/mol per non-H atom)	PDB
Bcl-x_L	Compound 31 (Abbot)[49]	0.036	0.27	1YSI
HPV E2	Compound 18 (Boehringer)[60, 61]	0.04	0.25	1R6N
ZipA	Compound 3 (Wyeth)[67]	83.1	0.22	1Y2G
BoNT/B	Doxorubicin (Brookhaven)[89]	9.4	0.18	1I1E
β-catenin	PNU-74654 (Nerviano)[90]	0.45	0.36	
Arf1/ARNO	LM11(Montpellier, France)[91]	49.7	0.22	
Disheveled	FJ9 (UCSF, St. Jude's)[92]	29	0.23	
Rac	NSC23766 (St. Jude's)[93]	50	0.19	
CD4 D1	J2 (Inst. Basic Med. Sci, Beijing)[94]	100	0.22	
HIV gp120	NBD-556 (Johns Hopkins)[95]	47	0.26	
eIF4E	4EGI-1 (Harvard Med. Sch.)[96]	25	0.22	
CD80	Compound 9 (Active Biotech)[97]	0.28	0.37	

References

1. Krogan, N. J. et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440, 637-43 (2006).
2. LaCount, D. J. et al. A protein interaction network of the malaria parasite Plasmodium falciparum. Nature 438, 103-7 (2005).
3. Komurov, K. & White, M. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. Mol Syst Biol 3 (2007).
4. Strong, M. & Eisenberg, D. The protein network as a tool for finding novel drug targets. Prog Drug Res 64, 191, 193-215 (2007).

5. Pu, S., Vlasblom, J., Emili, A., Greenblatt, J. & Wodak, S. J. Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. Proteomics 7, 944-60 (2007).

6. Collins, S. R. et al. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature 446, 806-810 (2007).

7. Jones, S. & Thornton, J. M. Principles of protein-protein interactions. Proc Natl Acad Sci U S A 93, 13-20 (1996).

8. Conte, L. L., Chothia, C. & Janin, J. The atomic structure of protein-protein recognition sites. Journal of Molecular Biology 285, 2177-2198 (1999).

9. Cheng, A. C. et al. Structure-based maximal affinity model predicts small-molecule druggability. Nat Biotechnol 25, 71-5 (2007).

10. Smith, R. D. et al. Exploring protein-ligand recognition with Binding MOAD. Journal of Molecular Graphics and Modelling 24, 414-425 (2006).

11. Hopkins, A. L. & Groom, C. R. The druggable genome. Nat Rev Drug Discov 1, 727-30 (2002).

12. Marsters, J. C., Jr. et al. Benzodiazepine peptidomimetic inhibitors of farnesyltransferase. Bioorg Med Chem 2, 949-57 (1994).

13. Zobel, K. et al. Design, synthesis, and biological activity of a potent Smac mimetic that sensitizes cancer cells to apoptosis by antagonizing IAPs. ACS Chem Biol 1, 525-33 (2006).

14. Robin, W. S. High-throughput screening of historic collections: Observations on file size, biological targets, and file diversity. Biotechnology and Bioengineering 61, 61-67 (1998).

15. Cochran, A. G. Antagonists of protein-protein interactions. Chem Biol 7, R85-94 (2000).

16. Clackson, T. & Wells, J. A. A hot spot of binding energy in a hormone-receptor interface. Science 267, 383-6 (1995).

17. Clackson, T., Ultsch, M. H., Wells, J. A. & de Vos, A. M. Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. J Mol Biol 277, 1111-28 (1998).

18. Muller, Y. A. et al. Vascular endothelial growth factor: crystal structure and functional mapping of the kinase domain receptor binding site. Proc Natl Acad Sci U S A 94, 7192-7 (1997).

19. Thanos, C. D., DeLano, W. L. & Wells, J. A. Hot-spot mimicry of a cytokine receptor by a small molecule. Proc Natl Acad Sci U S A 103, 15422-7 (2006).

20. Moreira, I. S., Fernandes, P. A. & Ramos, M. J. Hot spots - A review of the protein-protein interface determinant amino-acid residues. Proteins: Structure, Function, and Bioinformatics 9999, NA (2007).

21. DeLano, W. L., Ultsch, M. H., de Vos, A. M. & Wells, J. A. Convergent solutions to binding at a protein-protein interface. Science 287, 1279-83 (2000).

22. Sidhu, S. S., Lowman, H. B., Cunningham, B. C. & Wells, J. A. Phage display for selection of novel binding peptides. Methods Enzymol 328, 333-63 (2000).

23. Wrighton, N. C. et al. Small peptides as potent mimetics of the protein hormone erythropoietin. Science 273, 458-64 (1996).

24. Livnah, O. et al. Functional mimicry of a protein hormone by a peptide agonist: the EPO receptor complex at 2.8 A. Science 273, 464-71 (1996).

25. Arkin, M. R. & Wells, J. A. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. Nat Rev Drug Discov 3, 301-17 (2004).

26. Yin, H. & Hamilton, A. D. Strategies for targeting protein-protein interactions with synthetic agents. Angew Chem Int Ed Engl 44, 4130-63 (2005).

27. Fry, D. C. Protein-protein interactions as targets for small molecule drug discovery. Biopolymers 84, 535-52 (2006).

28. Arkin, M. Protein-protein interactions and cancer: small molecules going in for the kill. Current Opinion in Chemical Biology 9, 317-324 (2005).

29. Arkin, M. R. et al. Binding of small molecules to an adaptive protein-protein interface. Proc Natl Acad Sci U S A 100, 1603-8 (2003).

30. Braisted, A. C. et al. Discovery of a potent small molecule IL-2 inhibitor through fragment assembly. J Am Chem Soc 125, 3714-5 (2003).

31. Raimundo, B. C. et al. Integrating fragment assembly and biophysical methods in the chemical advancement of small-molecule antagonists of IL-2: an approach for inhibiting protein-protein interactions. J Med Chem 47, 3111-30 (2004).

32. Tilley, J. W. et al. Identification of a Small Molecule Inhibitor of the IL-2/IL-2Rα Receptor Interaction Which Binds to IL-2. J. Am. Chem. Soc. 119, 7589-7590 (1997).

33. Rickert, M., Wang, X., Boulanger, M. J., Goriatcheva, N. & Garcia, K. C. The structure of interleukin-2 complexed with its alpha receptor. Science 308, 1477-80 (2005).

34. Thanos, C. D., Randal, M. & Wells, J. A. Potent small-molecule binding to a dynamic hot spot on IL-2. J Am Chem Soc 125, 15280-1 (2003).

35. Emerson, S. D. et al. NMR characterization of interleukin-2 in complexes with the IL-2Rα receptor component, and with low molecular weight compounds that inhibit the IL-2/IL-Rα interaction. Protein Sci 12, 811-822 (2003).

36. Kuntz, I. D., Chen, K., Sharp, K. A. & Kollman, P. A. The maximal affinity of ligands. Proc Natl Acad Sci U S A 96, 9997-10002 (1999).

37. Lee, L. P. & Tidor, B. Optimization of binding electrostatics: charge complementarity in the barnase-barstar protein complex. Protein Sci 10, 362-77 (2001).

38. Midelfort, K. S. et al. Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody. J Mol Biol 343, 685-701 (2004).

39. Adams, J. M. & Cory, S. The Bcl-2 protein family: arbiters of cell survival. Science 281, 1322-6 (1998).

40. Sattler, M. et al. Structure of Bcl-xL-Bak Peptide Complex: Recognition Between Regulators of Apoptosis. Science 275, 983-986 (1997).

41. Petros, A. M. et al. Rationale for Bcl-xL/Bad peptide complex formation from structure, mutagenesis, and biophysical studies. Protein Sci 9, 2528-34 (2000).

42. Sadowsky, J. D., Murray, J. K., Tomita, Y. & Gellman, S. H. Exploration of backbone space in foldamers containing alpha- and beta-amino acid residues:

developing protease-resistant oligomers that bind tightly to the BH3-recognition cleft of Bcl-xL. Chembiochem 8, 903-16 (2007).

43. Yin, H. et al. Terphenyl-Based Bak BH3 alpha-helical proteomimetics as low-molecular-weight antagonists of Bcl-xL. J Am Chem Soc 127, 10191-6 (2005).

44. Walensky, L. D. et al. Activation of apoptosis in vivo by a hydrocarbon-stapled BH3 helix. Science 305, 1466-70 (2004).

45. Sadowsky, J. D. et al. (alpha/beta+alpha)-peptide antagonists of BH3 domain/Bcl-x(L) recognition: toward general strategies for foldamer-based inhibition of protein-protein interactions. J Am Chem Soc 129, 139-54 (2007).

46. Oltersdorf, T. et al. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. Nature 435, 677-81 (2005).

47. Bruncko, M. et al. Studies Leading to Potent, Dual Inhibitors of Bcl-2 and Bcl-xL. J. Med. Chem. 50, 641-662 (2007).

48. Hajduk, P. J. SAR by NMR: putting the pieces together. Mol Interv 6, 266-72 (2006).

49. Petros, A. M. et al. Discovery of a potent inhibitor of the antiapoptotic protein Bcl-xL from NMR and parallel synthesis. J Med Chem 49, 656-63 (2006).

50. Levine, A. J., Hu, W. & Feng, Z. The P53 pathway: what questions remain to be explored? Cell Death Differ 13, 1027-36 (2006).

51. Kussie, P. H. et al. Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. Science 274, 948-53 (1996).

52. Picksley, S. M., Vojtesek, B., Sparks, A. & Lane, D. P. Immunochemical analysis of the interaction of p53 with MDM2;--fine mapping of the MDM2 binding site on p53 using synthetic peptides. Oncogene 9, 2523-9 (1994).

53. Vassilev, L. T. et al. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. Science 303, 844-8 (2004).

54. Grasberger, B. L. et al. Discovery and cocrystal structure of benzodiazepinedione HDM2 antagonists that activate p53 in cells. J Med Chem 48, 909-12 (2005).

55. Parks, D. J. et al. 1,4-Benzodiazepine-2,5-diones as small molecule antagonists of the HDM2-p53 interaction: discovery and SAR. Bioorg Med Chem Lett 15, 765-70 (2005).

56. Koblish, H. K. et al. Benzodiazepinedione inhibitors of the Hdm2:p53 complex suppress human tumor cell proliferation in vitro and sensitize tumors to doxorubicin in vivo. Mol Cancer Ther 5, 160-9 (2006).

57. Fry, D. C. et al. NMR structure of a complex between MDM2 and a small molecule inhibitor. J Biomol NMR 30, 163-73 (2004).

58. Yoakim, C. et al. Discovery of the first series of inhibitors of human papillomavirus type 11: inhibition of the assembly of the E1-E2-Origin DNA complex. Bioorg Med Chem Lett 13, 2539-41 (2003).

59. White, P. W. et al. Inhibition of human papillomavirus DNA replication by small molecule antagonists of the E1-E2 protein interaction. J Biol Chem 278, 26765-72 (2003).

60. Wang, Y. et al. Crystal structure of the E2 transactivation domain of human papillomavirus type 11 bound to a protein interaction inhibitor. J Biol Chem 279, 6976-85 (2004).

61. Goudreau, N. et al. Optimization and determination of the absolute configuration of a series of potent inhibitors of human papillomavirus type-11 E1-E2 protein-protein interaction: a combined medicinal chemistry, NMR and computational chemistry approach. Bioorg Med Chem 15, 2690-700 (2007).

62. Abbate, E. A., Berger, J. M. & Botchan, M. R. The X-ray structure of the papillomavirus helicase in complex with its molecular matchmaker E2. Genes Dev 18, 1981-96 (2004).

63. Mosyak, L. et al. The bacterial cell-division protein ZipA and its interaction with an FtsZ fragment revealed by X-ray crystallography. Embo J 19, 3179-91 (2000).

64. Tsao, D. H. et al. Discovery of novel inhibitors of the ZipA/FtsZ complex by NMR fragment screening coupled with structure-based design. Bioorg Med Chem 14, 7953-61 (2006).

65. Hajduk, P. J., Huth, J. R. & Fesik, S. W. Druggability indices for protein targets derived from NMR-based screening data. J Med Chem 48, 2518-25 (2005).

66. Jennings, L. D. et al. Combinatorial synthesis of substituted 3-(2-indolyl)piperidines and 2-phenyl indoles as inhibitors of ZipA-FtsZ interaction. Bioorg Med Chem 12, 5115-31 (2004).

67. Rush, T. S., 3rd, Grant, J. A., Mosyak, L. & Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. J Med Chem 48, 1489-95 (2005).

68. Takasaki, W., Kajino, Y., Kajino, K., Murali, R. & Greene, M. I. Structure-based design and characterization of exocyclic peptidomimetics that inhibit TNF alpha binding to its receptor. Nat Biotechnol 15, 1266-70 (1997).

69. Carter, P. H. et al. Photochemically enhanced binding of small molecules to the tumor necrosis factor receptor-1 inhibits the binding of TNF-alpha. Proc Natl Acad Sci U S A 98, 11879-84 (2001).

70. He, M. M. et al. Small-molecule inhibition of TNF-alpha. Science 310, 1022-5 (2005).

71. Wendt, M. D. et al. Discovery of a novel small molecule binding site of human survivin. Bioorg Med Chem Lett 17, 3122-9 (2007).

72. Brown, Scott P. & Hajduk, Philip J. Effects of Conformational Dynamics on Predicted Protein Druggability. ChemMedChem 1, 70-72 (2006).

73. Eyrisch, S. & Helms, V. Transient Pockets on Protein Surfaces Involved in Protein-Protein Interaction. J. Med. Chem. 50, 3457-3464 (2007).

74. Keiser, M. J. et al. Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25, 197-206 (2007).

75. Feng, B. Y. et al. A high-throughput screen for aggregation-based inhibition in a large compound library. J Med Chem 50, 2385-90 (2007).

76. Carr, R. A. E., Congreve, M., Murray, C. W. & Rees, D. C. Fragment-based lead discovery: leads by design. Drug Discovery Today 10, 987-992 (2005).

77. Volkov, A. N., Worrall, J. A., Holtzmann, E. & Ubbink, M. Solution structure and dynamics of the complex between cytochrome c and cytochrome c peroxidase

determined by paramagnetic NMR. Proc Natl Acad Sci U S A 103, 18945-50 (2006).

78. Tang, C., Iwahara, J. & Clore, G. M. Visualization of transient encounter complexes in protein-protein association. Nature 444, 383-6 (2006).

79. Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. J Pharmacol Toxicol Methods 44, 235-49 (2000).

80. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46, 3-26 (2001).

81. Erlanson, D. A. Fragment-based lead discovery: a chemical update. Curr Opin Biotechnol 17, 643-52 (2006).

82. Lowe, J., Li, H., Downing, K. H. & Nogales, E. Refined structure of alpha beta-tubulin at 3.5 A resolution. J Mol Biol 313, 1045-57 (2001).

83. McMillan, K. et al. Allosteric inhibitors of inducible nitric oxide synthase dimerization discovered via combinatorial chemistry. PNAS 97, 1506-1511 (2000).

84. Schreiber, S. L. & Crabtree, G. R. The mechanism of action of cyclosporin A and FK506. Immunol Today 13, 136-42 (1992).

85. Fletterick, R. J., Bates, D. J. & Steitz, T. A. The structure of a yeast hexokinase monomer and its complexes with substrates at 2.7-A resolution. Proc Natl Acad Sci U S A 72, 38-42 (1975).

86. Anderson, C. M., Zucker, F. H. & Steitz, T. A. Space-filling models of kinase clefts and conformation changes. Science 204, 375-80 (1979).

87. Yankeelov, J. A., Jr. & Koshland, D. E., Jr. Evidence for Conformation Changes Induced by Substrates of Phosphoglucomutase. J Biol Chem 240, 1593-602 (1965).

88. Kelekar, A., Chang, B. S., Harlan, J. E., Fesik, S. W. & Thompson, C. B. Bad is a BH3 domain-containing protein that forms an inactivating dimer with Bcl-XL. Mol Cell Biol 17, 7040-6 (1997).

89. Eswaramoorthy, S., Kumaran, D. & Swaminathan, S. Crystallographic evidence for doxorubicin binding to the receptor-binding site in Clostridium botulinum neurotoxin B. Acta Crystallographica Section D 57, 1743-1746 (2001).

90. Trosset, J.-Y. et al. Inhibition of protein-protein interactions: The discovery of druglike ?-catenin inhibitors by combining virtual and biophysical screening. Proteins: Structure, Function, and Bioinformatics 64, 60-67 (2006).

91. Viaud, J. et al. Structure-based discovery of an inhibitor of Arf activation by Sec7 domains through targeting of protein-protein complexes. Proc Natl Acad Sci U S A 104, 10370-5 (2007).

92. Fujii, N. et al. An antagonist of dishevelled protein-protein interaction suppresses beta-catenin-dependent tumor cell growth. Cancer Res 67, 573-9 (2007).

93. Gao, Y., Dickerson, J. B., Guo, F., Zheng, J. & Zheng, Y. Rational design and characterization of a Rac GTPase-specific small molecule inhibitor. Proceedings of the National Academy of Sciences 101, 7618-7623 (2004).

94. Xiao, H. et al. Potent inhibition of the CD4-dependent T cell response by J2, a novel nonpeptide organic ligand of CD4 D1. Molecular Immunology 44, 784-795 (2007).

95. Schon, A. et al. Thermodynamics of Binding of a Low-Molecular-Weight CD4 Mimetic to HIV-1 gp120. Biochemistry 45, 10973-10980 (2006).

96. Moerke, N. J. et al. Small-molecule inhibition of the interaction between the translation initiation factors eIF4E and eIF4G. Cell 128, 257-67 (2007).

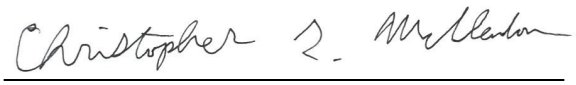
97. Uvebrant, K. et al. Discovery of Selective Small-Molecule CD80 Inhibitors. Journal of Biomolecular Screening 12, 464-472 (2007).

98. DeLano, W. L. Unraveling hot spots in binding interfaces: progress and challenges. Curr Opin Struct Biol 12, 14-20 (2002).

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts 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, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

September 26, 2011
Date