UCSF
UC San Francisco Electronic Theses and Dissertations

Title
Modeling loop flexibility via conformational ensembles

Permalink
https://escholarship.org/uc/item/2vs3816j

Author
Wong, Sergio E

Publication Date
2006

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

To my parents and sister (Mom, Dad and chubby girl)

To my parents and sister (Mom, Dad and chubby girl)

This thesis would not have been possible without the support of my family, friends and colleagues and my research advisor, Professor Matt Jacobson

Acknowledgements

This thesis would not have been possible without the support of my family, friends and colleagues and my research advisor, Professor Matt Jacobson

Modeling Loop Flexibility via Conformational Ensembles

by

Sergio E. Wong

Abstract

The field of loop modeling almost exclusively focuses on finding a single loop conformation given a loop sequence and its protein environment. Here, I propose to move beyond this limitation and model loops as a conformational ensemble. This is a better description of loop structures given their flexibility and the biological role this flexibility plays. First this approach was used, in combination with the conformational selection hypothesis, to predict loop-latching motions that occur upon ligand binding. That effort yielded structures remarkably close to the known holo coordinates and docking into these structures produced results superior to the corresponding apo calculations. Second, I examined flexibility loss in antibody CDR loops during affinity maturation which is hypothesized to help pre-organize CDR loops and lower the entropic cost of binding. Molecular dynamics simulations of 4 antibodies captured the flexibility loss and analysis of their trajectories shed light into how individual mutations, particularly serine to asparagines, restrict CDR loops. A conformational selection binding mechanism, for germline antibody 7G12g, was evident from equilibrium between the bound and unbound H3 loop conformations in the absence of ligand.

Phosphorylation was studied at length because it shifts the conformational

ensemble population distribution. Since simulations of phosphorylated residues remain largely unexplored, I explored the suitability of current solvation models to properly react to a phosphate moiety. Hydrogen bonding of phosphorylated residues explained phosphorylation induced shifts in $^{3}J_{NH\text{-}C_{\alpha}H}$ coupling constants for short tetrapeptides; this underscored the need for the solvation model to modulate hydrogen bond strength. A distance dependent dielectric performed consistently poorly in our simulations while a Generalized Born treatment and the explicit water models TIP3P, TIP4P and SPC/E provided satisfactory results. In addition, the role of polarizability in interactions of phosphorylated residues was explored. *Ab initio* calculations made it clear polarizability contributes to phosphorylated residues' preference for argenine over lysine as salt-bridge partners. It was also found the protonated phosphate moiety (-1) and the carboxylate ion have similar interaction energies with argenine and lysine, which may explain why carboxylic acids can sometimes substitute for phosphorylation.

Prof. Matthew P. Jacobson

Thesis Advisor

Table of Contents

Copyright ii

Dedication iii

Acknowledgements iv

Abstract v

Table of Contents vii

Introduction 1

Chapter 1: Conformational selection *in silico*: Loop latching-motions and ligand binding enzymes..... 17

Chapter 2: Limitations of the replica exchange method..... 62

Chapter 3: How do somatic mutations rigidify CDR loops during affinity maturation? 87

Chapter 4: Competition between intramolecular hydrogen bonds and solvation in phosphorylated peptides: Simulations with explicit and implicit solvent ... 116

Chapter 5: Polarizability contributes to salt-bridge preferences of phosphorylated residues 166

Conclusion 197

Introduction

The topic of this thesis is modeling loop flexibility using conformational ensembles. In this section I explain the motivation for this project and the topics of the chapters that follow. First, however, I discuss the motivation to perform loop modeling; it starts with a discussion of the availability of genome-wide sequence data and is followed by an introduction to comparative modeling.

The post-genomic era

Genomic sequencing efforts yielded thousands of protein sequences-nearly 25,000 in the human genome alone(Venter *et al.*, 2001). There is great promise in understanding the biological function of these proteins and identifying drug targets to treat human disease. The next challenge is to annotate their function(s). Sequence similarity to other proteins provides some information, however, their structures are necessary to further elucidate their functional role(s). Structural genomics(Stevens *et al.*, 2001) efforts aim to ameliorate the situation, however, it is not their goal to yield a set of coordinates for every protein sequence in existence. Instead the objective is to determine representative structures of each protein family so these can be used as templates in comparative modeling.

Comparative modeling and the motivation for loop modeling

Comparative modeling is based on the observation that proteins with similar sequences have similar structures. This is used to build models given similar protein sequences whose structures are known. A sequence identity above 40% is usually sufficient to yield a reasonable model (Marti-Renom *et al.*, 2000). Briefly, the sequence of the unknown (target) protein is aligned against a similar sequence whose structure is known (template). The alignment is then used to build the unknown sequence's structure. In places where the alignment agrees, the coordinates are typically copied, while insertions and deletions need to be modeled in. These insertions often involve loops, the segments connecting secondary structure, and predicting their conformation is termed loop modeling. Since in a given protein family the overall fold is shared among its members, functional differences must arise from mutations in the protein scaffold and insertions or deletions between them. Thus, it is critical to accurately model these differences for the final structures to be useful. In the case of immunoglobulins, these loop differences are crucial. Here, 6 complementary determining region (CDR) loops form the binding site for antigens. The rest of the protein is largely conserved, both in sequence and structure. However, by changing the length and composition of these loops, immunoglobulins can bind essentially any chemical species with high affinity and specificity. Most of the CDR loops can be described using the canonical conformations of Lesk and Chothia (Chothia & Lesk, 1987). However, the H3 loop, which has the largest variation in length and sequence, remains elusive to this type of approach. Antibodies can be a model

UCSF LIBRARY

system for understanding loop conformational preferences in general, since the protein framework is largely constant. Once CDR conformations can be reliably predicted, loop modeling can be key in engineering antibodies that bind a given target. The latter would also improve our understanding of ligand binding in general. A similar example, that is less well understood, are G-coupled protein receptors (GCPR)(Gether, 2000). They are transmembrane proteins that span the cell membrane. Their loops on the extracellular side bind signaling molecules which trigger a response, coupled to a G-protein, in the GCPR's cytosolic domain. The transmembrane helical bundle is common among members of this family, so it is the extracellular loops' composition and conformation that control which molecule will elicit their response. The loops on the cytosolic side control the nature of the response. Once these insertions can be correctly modeled, one can imagine being able to: 1) start searching for possible ligands that bind these loops and 2) once such predictive power is achieved, mutate these loops to induce a given response for a new molecule, for example.

To further complicate matters loops are more flexible than other secondary structure and can easily adapt multiple conformations. Surveys of the PDB(Berman *et al.*, 2000) verify that a given a particular loop sequence can adapt several conformations. The latter include loops involved in ligand binding; proteins with this type of loops include triosephosphate isomerase(Williams & Mcdermott, 1995) and protein tyrosine phosphatases (Juszczak *et al.*, 1997; Williams & Mcdermott, 1995), members of the enolase superfamily(Babbitt *et al.*,

UCSF LIBRARY

1995), and kinases(Johnson & Lewis, 2001). Thus, in order to fully uncover the function of proteins, as initially discussed in this introduction, loop flexibility must be included in the description of loops. However, loop modeling techniques typically seek a single conformation for a given loop.

How is loop modeling currently performed?

Loop modeling methods fall into 2 main categories: *ab initio* and database-based. *Ab initio* methods search for possible loop conformations and then select, using an energy function, the most stable conformation. Ideally, an exhaustive conformational search coupled to an accurate energy function would identify the correct loop conformation. However, the systematic search for a 12 residue loop with 2 backbone and 2 viable side chain conformations, per residue, would involve 2^{24} (~16.8 million) structures, which makes this approach impractical. In addition to sampling issues, the energy function appears to be the leading cause of error in some approaches (Fiser *et al.*, 2000). Techniques for loop modeling include simulated annealing (Zhang *et al.*, 1997), systematic backbone dihedral search guided by empirical distribution(Moulton & James, 1986), random backbone dihedral perturbations with loop closure restraints(Shenkin *et al.*, 1987), uniform conformational sampling(Bruccoleri & Karplus, 1986), high temperature molecular dynamics (Bruccoleri & Karplus, 1986), a global energy optimization approach (Dudeck *et al.*, 1998), biased monte carlo sampling(Abagyan & Totrov, 1994) and modeling of segments within the loop to make the problem more

UCSF LIBRARY

manageable(Tosatto *et al.*, 2002). The energy function depends on the representation chosen, such as an atomistic description versus a united atom description, for example. It was found, however, that a full atomic resolution force field with an implicit solvent model outperforms a statistical potential (de Bakker *et al.*, 2003) .

Database based methods search the PDB for putative loop conformations; the selections are typically based on sequence similarity. These segments are then inserted into the target structure and refined using an energy function which may be statistical in nature (Deane & Blundell, 2000; Debnath, 1997; Levefelt & Lundh, 2006; Mandal *et al.*, 1996; Martin & Thorton, 1996; Oliva *et al.*, 1997; Pellequer & Chen, 1997; Rohl *et al.*, 2004; Rufino *et al.*, 1997; Shepherd *et al.*, 1999; Wintjens & Rooman, 1996).

These approaches to loop modeling, however, seek a single, unambiguous structure. Loops, however, can exist in more than a single conformation. A more realistic representation should include conformational sub-states of the loop. Here the methodology is pushed beyond the search of a single structure and into modeling loop conformational ensembles.

Modeling loop flexibility using conformational ensembles

UCSF LIBRARY

One can imagine two ways of describing flexibility. In the kinetic sense one may consider a loop flexible if has low energy barriers between conformational sub-states; that is, if it is floppy. In the thermodynamic sense flexibility may be assessed by the number of accessible conformational sub-states i.e. entropy in statistical mechanics; it is intuitive to think a flexible loop has a high conformational entropy. Clearly a kinetically flexible loop that accesses many conformational states will also be thermodynamically flexible, however, the converse is not true. One can imagine numerous but deep and narrow free energy basins that would not permit frequent interchange among them. Here, the focus will be on the thermodynamic definition of flexibility and, thus, on identification of conformational sub-states as well as factors that modulate their relative populations.

The conformational sub-states in question make up the canonical ensemble at 300 K. This is the most natural definition since it is the simplest ensemble that approximates experimental conditions in the laboratory (constant temperature and nearly constant volume with a fixed number of molecules).

In chapter 1, an ensemble representation is applied to a set of proteins where a loop occludes the active site upon ligand binding. In some of these cases, it is experimentally known that, in the absence of ligand, the loop exists in equilibrium between the apo and holo conformations. Thus, by identifying their conformational sub-states it was possible to find holo-like conformations.

UCSF LIBRARY

Docking calculations, which are particularly sensitive to receptor geometry, against the structures found were able to select the holo-like loop conformations. Thus, the conformational selection process was reproduced *in silico*.

During the calculations of chapter 1, methodological issues surfaced and involved the efficacy of the replica exchange method (REM)(Sugita & Okamoto, 1999) given largely entropic barriers. In chapter 2, model systems were used to evaluate the performance of REM given these types of barriers. REM takes advantage of higher kinetic energy at higher temperatures to overcome barriers in molecular dynamics simulations (or a higher acceptance probability at higher temperature in monte carlo simulations). Limitations of this approach become evident when the free energy barriers in question are largely entropic rather than enthalpic.

In chapter 3 the flexibility of antibody CDR loops is examined via molecular dynamics simulations. CDR loop regions undergo mutations during an immune response to improve their affinity and specificity for a particular antigen. There is evidence suggesting these mutations rigidify the binding site and pre-arrange it for binding(Foote, 2003; Manivel *et al.*, 2000; Sagawa *et al.*, 2003; Yin *et al.*, 2003). In chapter 3 the results of an exploratory study on well-characterized systems are discussed. Interestingly, the predicted differences were captured in the simulations and the reasons behind the changes in flexibility became evident, in some cases.

UCSF LIBRARY

Among factors that modulate loop sub-state populations, interactions involving phosphorylated residues were chosen for study. Phosphorylation is a prevalent cellular signaling mechanism(Johnson & Lewis, 2001); among its many functions, it plays a role in extra-cellular signal transduction via modulation of mitogen-associated protein (MAP) kinases(Chen *et al.*, 2001), cell cycle regulation via cyclin dependent kinases(Johnson & Lewis, 2001; Vermeulen *et al.*, 2003), insulin receptor response(Johnson & Lewis, 2001), growth signaling pathways via Src family kinases(Johnson & Lewis, 2001) and regulation of ligand-gated ion channels(Swope *et al.*, 1999) as well as voltage-gated sodium and potassium channels(Catterall, 2000; Martens *et al.*, 1999). Phosphorylation often occurs on loop segments and drastically shifts its conformational equilibrium. Despite the importance of this problem, simulation of phosphorylated peptides remains largely unexplored. This motivated the focus on phosphorylation for two chapters of the present work. The calculation of adequate atomic charges for atomistic simulation of phosphorylated residues and the suitability of solvation models for this application are considered in chapter 4. The latter are explored in the context of local hydrogen bond interactions.

Chapter 5 focuses on interactions that occur largely with residues distant in sequence. Arginine and lysine are the two most frequent salt-bridge partners for phosphorylated residues with arginine occurring ~3 times as often as lysine (see PDB survey in chapter 5). The difference in charge distribution between the

UCSF LIBRARY

amino and guanido group prompted an inquiry about the role of polarizability in the bias towards arginine.

Abagyan, R., & Totrov, M. (1994). Biased probability monte carlo conformational searches and electorstatic calculations for peptides and proteins. *Journal of Molecular Biology, 235*, 983-1002.

Babbitt, P. C., Mrachko, G. T., Hasson, M. S., Huisman, G. W., Kolter, R., Ringe, D., et al. (1995). Functionally diverse enzyme superfamily that abstracts the alpha-protons of carboxylic-acids. *Science, 267*(5201), 1159-1161.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucl. Acids Res., 28*(1), 235-242.

Bruccoleri, R., & Karplus, M. (1986). Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers, 26*(1), 137-168.

Catterall, W. A. (2000). From ionic currents to molecular mechanisms: The structure and function of voltage-gated sodium channels. *Neuron, 26*(1), 13-25.

Chen, Z., Gibson, T. B., Robinson, F., Silvestro, L., Pearson, G., Xu, B., et al. (2001). Map kinases. *Chem Rev, 101*(8), 2449-2476.

UCSF LIBRARY

References

Abagyan, R., & Totrov, M. (1994). Biased probability monte carlo conformational searches and electorstatic calculations for peptides and proteins. *Journal of Molecular Biology, 235*, 983-1002.

Babbitt, P. C., Mrachko, G. T., Hasson, M. S., Huisman, G. W., Kolter, R., Ringe, D., et al. (1995). Functionally diverse enzyme superfamily that abstracts the alpha-protons of carboxylic-acids. *Science, 267*(5201), 1159-1161.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucl. Acids. Res., 28*(1), 235-242.

Bruccoleri, R., & Karplus, M. (1986). Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers, 26*(1), 137-168.

Catterall, W. A. (2000). From ionic currents to molecular mechanisms: The structure and function of voltage-gated sodium channels. *Neuron, 26*(1), 13-25.

Chen, Z., Gibson, T. B., Robinson, F., Silvestro, L., Pearson, G., Xu, B., et al. (2001). Map kinases. *Chem Rev, 101*(8), 2449-2476.

UCSF LIBRARY

Chothia, C., & Lesk, a. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *Journal of Molecular Biology, 196*(4), 901-917.

de Bakker, P., De Pristo, M., Burke, D., & Blundell, T. (2003). Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by all-atom statistical potential and the amber force field with the generalized born solvation model. *Proteins-Structure Function and Bioinformatics, 51*, 21-40.

Deane, C., & Blundell, T. (2000). A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. *Proteins-Structure Function and Bioinformatics, 40*, 135-144.

Debnath, A. (1997). A "fragment fitting approach" to model disulfide loops by utilizing homologous peptide fragments from unrelated proteins of known structures: Application to the v3 loop of the hiv-1 envelope glycoprotein gp120. *Journal of Molecular Modeling, 3*, 31-47.

Dudeck, M., kamnarayan, K., & Ponder, J. (1998). Protein structure prediction using a combination of sequence homology and global energy minimization: Ii energy functions. *Journal of Computational Chemistry, 19*(5), 548-573.

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

Foote, J. (2003). Isomeric antibodies. *Science, 299*(5611), 1327-1328.

Gether, U. (2000). Uncovering molecular mechanisms involved in activation of g protein-coupled receptors. *Endocrine Reviews, 21*(1), 90-113.

Johnson, L. N., & Lewis, R. J. (2001). Structural basis for control by phosphorylation. *Chem Rev, 101*(8), 2209-2242.

Juszczak, L. J., Zhang, Z. Y., Wu, L., Gottfried, D. S., & Eads, D. D. (1997). Rapid loop dynamics of yersinia protein tyrosine phosphatases. *Biochemistry, 36*(8), 2227-2236.

Levefelt, C., & Lundh, D. (2006). A fold-recognition approach to loop modeling. *Journal of Molecular Modeling, 12*, 125-139.

Mandal, C., Kingery, B., Anchin, J., Subramaniam, S., & Linthicum, D. (1996). Abgen: A knowledge-bsed automated approach to antibody structure modeling. *Nature, 14*, 323-328.

Manivel, V., Sahoo, N. C., Salunke, D. M., & Rao, K. V. S. (2000). Maturation of an antibody response is governed by modulations in flexibility of the antigen-combining site. *Immunity, 13*(5), 611-620.

Martens, J. R., Kwak, Y. G., & Tamkun, M. M. (1999). Modulation of kv channel alpha/beta subunit interactions. *Trends Cardiovasc Med, 9*(8), 253-258.

Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure, 29*, 291-325.

Martin, A., & Thorton, J. (1996). Structural families in loops of homologous proteins: Automatic classification, modelling and application to antibodies. *Journal of Molecular Biology, 263*, 800-815.

Moulton, J., & James, M. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins-Structure Function and Bioinformatics, 1*, 146-163.

Oliva, B., Bates, P., Querol, E., Aviles, F., & Sternberg, M. (1997). An automated classification of the structure of protein loops. *Journal of Molecular Biology, 266*, 814-830.

Pellequer, J., & Chen, S. (1997). Does conformational free energy distinguish loop conformations in proteins? *Biophysical Journal, 73*, 2359-2375.

Rohl, C., Strauss, C., Chivian, D., & Baker, D. (2004). Modeling structurally variable regions in homologous proteins with rosetta. *Proteins-Structure Function and Bioinformatics, 55*, 656-677.

Rufino, S., Donate, L., Canard, L., & Blundell, T. (1997). Predicting the conformational class of short and medium size loops connecting regular secondary structures: Applications to comoparative modeling. *Journal of Molecular Biology, 267*, 352-367.

Sagawa, T., Oda, M., Ishimura, M., Furukawa, K., & Azuma, T. (2003). Thermodynamic and kinetic aspects of antibody evolution during the immune response to hapten. *Molecular Immunology, 39*(13), 801-808.

Shenkin, P., Yarmush, D., Fine, R., Wang, H., & Levinthal, C. (1987). Predicting antibody hypervariable loop conformations: Ensembles of random conformations for ring-like structures. *Biopolymers, 26*, 2053-2085.

Shepherd, A., Gorse, D., & Thorton, J. (1999). Prediction of the location and type of beta-turns in proteins using neural networks. *Protein Science, 8*(1045-1055).

Stevens, R. C., Yokoyama, S., & Wilson, I. A. (2001). Global efforts in structural genomics. *Science, 294*(5540), 89-92.

Sugita, Y., & Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters, 314*(1-2), 141-151.

Swope, S. L., Moss, S. J., Raymond, L. A., & Huganir, R. L. (1999). Regulation of ligand-gated ion channels by protein phosphorylation. *Ion Channel Regulation, 33*, 49-78.

Tosatto, S. C. E., Bindewald, E., Hesser, J., & Manner, R. (2002). A divide and conquer approach to fast loop modeling. *Protein Engineering, 15*(4), 279-286.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-+.

Vermeulen, K., Van Bockstaele, D. R., & Berneman, Z. N. (2003). The cell cycle: A review of regulation, deregulation and therapeutic targets in cancer. *Cell Prolif, 36*(3), 131-149.

Williams, J. C., & Mcdermott, a. E. (1995). Dynamics of the flexible loop of triosephosphate isomerase - the loop motion is not ligand-gated. *Biochemistry, 34*(26), 8309-8319.

Wintjens, R., & Rooman, M. (1996). Automatic classification and analysis of alpha-alpha turn motifs in proteins. *Journal of Molecular Biology, 255*, 235-253.

Yin, J., Beuscher, A. E., Andryski, S. E., Stevens, R. C., & Schultz, P. G. (2003). Structural plasticity and the evolution of antibody affinity and specificity. *Journal of Molecular Biology, 330*(4), 651-656.

Zhang, H. Y., Lai, L. H., Wang, L. Y., Han, Y. Z., & Tang, Y. Q. (1997). A fast and efficient program for modeling protein loops. *Biopolymers, 41*(1), 61-72.

Conformational selection *in silico*: Loop latching motions and ligand binding in enzymes

Abstract

Ligand binding frequently induces significant conformational changes in a protein receptor. Understanding and predicting such conformational changes represents an important challenge for computational biology, including applications to structure-based drug design. Previous methods usually search, to some extent, the receptor conformational space for many putative ligand conformations. We describe an alternative approach based on the assumption that the holo state is at least transiently populated in the absence of a ligand; this hypothesis has been referred to as "conformational selection". Here we apply a method that tests this hypothesis on a challenging class of ligand-induced conformational changes, which we refer to as loop-latching: the closing of a loop around an active site that sequesters the ligand from solvent. The method uses a combination of replica exchange molecular dynamics and a loop prediction algorithm to generate low energy loop structures, and docking to select the conformation appropriate for binding a particular ligand. On a test set of 6 proteins, it yields loop structures including holo-like conformations, generally below 2 Å RMSD from the liganded structure, for loops that span up to 15 residues. Docking served as a stringent test of the predictions and verified their ability to improve docking results and yield correct poses, in most cases.

Introduction

Conformational changes that occur upon ligand binding have been described using two different models generally referred to as "induced fit" and "conformational selection". Induced fit, as originally formulated by Koshland(Koshland, 1958), postulated that conformational changes in the macromolecular receptor would follow initial binding. It should be noted however that the term is often used colloquially to refer to any conformational change accompanying ligand binding. Conformational selection(Freire, 1998; Ma *et al.*, 1999; Tsai *et al.*, 1999; Tsai *et al.*, 2001) refers to an alternate hypothesis that postulates: 1) the free enzyme exists in equilibrium among several conformational sub-states and 2) the ligand preferentially binds to one or more sub-states, shifting the equilibrium in that direction. The two models can be distinguished in some cases by kinetic measurements(Berger *et al.*, 1999; Foote & Milstein, 1994; James & Tawfik, 2005) where the binding constants behave differently with varying ligand concentrations. From a thermodynamic perspective, however, induced fit and conformational selection can be viewed as two extremes of a continuum, with most systems displaying some combination of both. The key thermodynamic factor is the free energy difference between the free and bound forms of the macromolecule, which determines whether a state similar to the bound form will be populated significantly at thermal equilibrium and whether it will play a major role in binding.

Predicting ligand-induced conformational changes is relevant both for understanding sequence-structure-function relationships in enzymes (e.g., the well-studied triosephosphate isomerase system(Alvarez *et al.*, 1998; Aparicio *et al.*, 2003; Mande

et al., 1994; Williams & Mcdermott, 1995)) and for structure-based drug design. A number of computational methods have been developed to address this problem(Carlson, 2002a; Teague, 2003; Wong & McCammon, 2003), which usually involve sampling possible conformations of both the receptor and the ligand simultaneously (as in some docking algorithms, such as FlexE(Claussen *et al.*, 2001) and others(R. Taylor *et al.*, 2003; Totrov & Abagyan, 1997)) or iteratively(Sherman *et al.*, 2006). Such strategies have been successful in a number of cases, particularly when the conformational changes in the receptor are relatively small, e.g., side chain rotamer flips.

The conformational selection hypothesis is attractive from a computational standpoint because it implies that ligand-induced conformational changes can be studied, in part, by simulations of the free macromolecule; for example, MD simulations of Pitici et al(Pitici *et al.*, 2002) suggest the holo conformation of protein U1A is stable in the absence of ligand and flexibility measurements by Jimenez et al(Jimenez *et al.*, 2003) are inconsistent with a lock-and-key binding mechanism in antibodies. Another approach correlated dynamic fluctuations to ligand induced changes(Ikeguchi *et al.*, 2005) and motivated the use of "relevant" normal modes to predict loop re-arrangements in kinase ligand binding(Cavasotto *et al.*, 2005). The latter study showed examination of the free enzyme dynamics can be useful in predicting receptor rearrangements in ligand binding. This is particularly attractive for studying large conformational changes, since it would greatly restrict the number of conformations relevant to binding. In general, the number of "states" in experimentally characterized systems is relatively few i.e. usually less than five(Cao *et al.*, 1998; Juszczak *et al.*, 1997; Williams & Mcdermott, 1995). A general approach proposed

here is to 1) sample the free enzyme thermodynamic ensemble, 2) identify conformational sub-states and 3) dock into the structures found. From the computational standpoint, it is not necessarily critical for the bound form to be well populated at ambient temperatures, only that the bound form represents a quasi-stable state that can be populated in a simulation, which could in principle be performed at higher temperatures or, as in this work, performed subject to restraints that prevent sampling of irrelevant conformations. That is, as long as the bound conformation of the macromolecule is not too much higher in energy than the ground state, and represents a local minimum or basin on the energy surface, it should be possible to locate it, along with other conformations which may or may not be relevant to ligand binding. Docking simulations can then identify the conformations relevant to binding a given ligand, if one is known.

In this manuscript we propose and evaluate a method of this type to predict protein "loop-latching" movements upon ligand binding. "Loop-latching" refers to the case where a loop covers the ligand in the holo structure and adopts an "open" conformation in the apo form. Predicting a large structural change of this type is challenging but important because it is observed in many enzymes, including many members of the enolase superfamily(Babbitt *et al.*, 1995) (Gulick *et al.*, 2001; Klenchin *et al.*, 2004; Wedekind *et al.*, 1994), triosephosphate isomerase(Williams & Mcdermott, 1995), *Yersinia* protein tyrosine phosphatase(Juszczak et al., 1997), Inosine 5'-monophosphate dehydrogenase(McMillan *et al.*, 2000) (IMD), lactate dehydrogenase(Gerstein & Chothia, 1991), beta-lactoglobulin(Wu & Deem, 1999), S-adenosylmethionine synthatase(J. C. Taylor & Markham, 2003), pepsin(Okoniewska *et al.*, 2000), 1,4 galactosesyl transferase,

phosphoribosylglycinamide formyltransferase(Dahms *et al.*, 2005; Gunasekaran *et al.*, 2003) (GART), tryptophan synthase(Rhee *et al.*, 1997) and Ribulose 1,5-biphosphate carbosylase/oxygenase(T. C. Taylor & Andersson, 1997). Protein tyrosine phosphatase family members, such as PTP 1B(Tonks, 2003; Zhang, 2001), IMD(Franklin *et al.*, 1999; Pankiewicz *et al.*, 2002), and spermidine synthase(Haider *et al.*, 2005; Thomas & Thomas, 2003) , all of which undergo this loop-latching motion, are targets for drug design.

Our own interest in loop latching is motivated in part by our ongoing efforts to investigate and predict enzyme function by metabolite docking(Kalyanaraman *et al.*, 2005; Tricot *et al.*, 1989). In previous work, Kalyanaraman et al(Kalyanaraman et al., 2005) demonstrated that the known substrates of alpha-beta barrel proteins were ranked highly when docking a large library of possible metabolites against the enzyme active sites. However, the method worked best when applied to holo enzyme structures; apo structures gave more variable results, due to large conformational changes in loops covering the active site, i.e., loop latching. Structural genomics efforts are now generating structures for many enzymes with uncertain function or specificity, and virtually all of these are apo structures. Tools for predicting how these enzymes may interact with potential substrates, including associated conformational changes such as loop latching, may aid in leveraging the huge increase in structural data.

This case of protein loop latching motions is a particularly challenging subset of the general induced fit problem. The lengths of the loops, 6-15 residues in this study, and their inherent flexibility, permit a vast number of possible rearrangements. Energy

filters or other heuristics could filter putative structures, but a large fraction would remain. To trim this number, some approaches only explore pre-selected degrees of freedom(Cavasotto et al., 2005; Kovacs *et al.*, 2005; Sandak *et al.*, 1998). But given the flexibility of these loops, it would not be *a priori* clear that only a few degrees of freedom would capture most of the possible rearrangements. Therefore, it is necessary to restrict the conformational space search differently.

In this manuscript, we investigate whether we can predict the holo-like closed form of flexible "latch" loops starting from the apo structures. We assess success in several ways, including very challenging applications where we evaluate the suitability of the predicted holo-like structures for docking calculations, which are notoriously sensitive to fine details of the binding site geometry. The resulting loop structures were remarkably similar, mostly below 2 Å RMSD, to the known holo coordinates. Known ligands 'select' the best loop conformation, as assessed by RMSD to the holo structure, by yielding the lowest (best) docking score. Thus, in this sense, the computational method mimics the conformational selection process *in silico*.

We do not claim, however, that these results constitute a rigorous test of the induced fit hypothesis, or that this method is completely general. The method contains a number of heuristics, the most important of which are restraints during REMD that prevent the loop from sampling irrelevant "open" conformations. This itself does not greatly reduce generality, because the restraints rely only in knowledge of the binding site location. The loops sample several different basins in these simulations i.e. the restraints are not so restrictive that only holo-like conformations are sampled. Altogether we consider these results to represent an encouraging approach to a rather

challenging type of ligand-induced conformational change.

Methodology

Overview

The approach proposed in the introduction consists of three parts: 1) sampling of the unbound state, 2) identification of the conformational sub-states and 3) docking against structural representatives of each sub-state.

In the first stage it is imperative to sample efficiently to include most of the loop conformational sub-states in the analysis. Replica exchange molecular dynamics (REMD) accelerates sampling by overcoming barriers in its high temperature replicas. However, in early tests, our unrestricted high temperature loop simulations drifted away from the protein and, explored conformations in bulk solvent and did not sample "holo"-like conformations within several nanoseconds of simulation time. This is because the favorable entropy of "open" loop conformations dominates at higher temperatures. As described in more detail below, a harmonic restraint on the loop was applied to focus the sampling on the relevant space, in which the loop forms contacts with the protein. The use of the restraint only requires knowledge of where the active site is located, and thus does not significantly reduce the applicability of the method.

Clustering identified conformational sub-states. Since centroids are average structures, bond lengths and angles can be distorted and there can be significant steric clashes. A loop prediction calculation, with Cα atoms restrained to the Cartesian

vicinity of the centroid, yielded low energy structures that were used for docking. This ensured that the final structure is chemically sound and not highly strained. The accuracy of the prediction was measured via backbone RMSD to the known holo structure and by the success of docking applications using the models.

Test cases

The cases considered here are: L-Ala-D-Glu epimerase(Klenchin et al., 2004) (AEE), *Yersinia* protein tyrosine phosphatase(Juszczak et al., 1997; Stuckey *et al.*, 1994), enolase, triosephosphate isomerase(Aparicio et al., 2003; Mande et al., 1994), phosphoribosylglycinamide formyltransferase (GART)(Almassy *et al.*, 1992; Dahms et al., 2005) and spermidine synthase(Haider et al., 2005; Kobayashi *et al.*, 2005; Korolev *et al.*, 2002; Thomas & Thomas, 2003) (Table 1). They were selected based on the following criteria: 1) one and only one loop latches over active site upon ligand binding, 2) crystal structures of the apo and holo forms are available from the PDB, 3) the biologically relevant ligand is known and 4) the ligand-gated backbone conformational change involves only loop residues. In addition, preference was given to cases where experimental evidence suggested the "closed" structure is partially populated in the absence of ligand *e.g. Yersinia* protein tyrosine phosphatase and TIM. The number was restricted to six due to the high computational expense of the replica exchange simulations.

In the following sections, the preparation of the systems for the simulations is explained and details of the replica exchange simulation, clustering and docking calculations follow.

System Preparation

This section describes the steps taken to prepare the protein structures for the REMD simulations. X-ray crystal structures for the apo and holo forms of all the systems examined were taken from the Brookhaven Protein Database (PDB)(Berman *et al.*, 2000). The proteins used in this study and their PDB structure codes are listed in Table 1. To minimize the number of atoms simulated, only portions of the protein near the mobile loop were included. Pymol(Delano, 2002) was used to visualize the apo structures, select the loop residues, expand the selection to residues surrounding the loop in question and write out the coordinates. All residues with any atom inside the distance cutoff were included in their entirety. The selection radius used varied because of variations in loop length and structure and is specified in Table 1.

In addition, a point in the active site and "tip" residues were identified. This information was used to impose a harmonic distance restraint between Cα atoms of "tip" residues and the active site point. The tip residues were visually chosen to be at the turn of the loop, that is, approximately where the loop reverses direction. Table 1 details, for each protein, PDB codes for the structures used, the loop residues, selection radius, tip residues, reference active site point and the restraint radius. The apo structures were structurally aligned to the holo structure via their Cα atoms using the "align" command. In some occasions, points on ligands were used as the active site reference point. In those cases, an artificial atom took the place of the ligand atom in the REMD simulations. This artificial atom is a neutral ion with a mass of 12 au.

The tleap module in AMBER 8(D.A Case, 2004) processed the partial structures to

produce a topology and structure file. The parm99.dat force field with modified backbone dihedral parameters(Simmerling *et al.*, 2002) was used. Atomic radii were modified, per instructions in the AMBER 8 manual, using the command "set default PBradii mbondi2".

Distance restraint files, between the "tip" residues and a point on the active site, were produced using the makeDIST program (part of the AMBER8 program suite). The 7 column file input was prepared; that is, only a maximum distance was specified. The restraint potential is a flat potential until it reaches the specified maximum distance. At that point, it becomes a harmonic potential with a harmonic constant of 20 kcal/mol/A^2 that drives the particle back towards the flat portion. The restraints permit the loop only to sample relevant conformations where the loop is relatively near the active site and not pointing out into solution.

Replica exchange-molecular dynamics

The AMBER 8 simulation package was used to perform all of the molecular dynamics simulations in this study. All of them were carried out using a Generalized Born model with a surface area correction term (GB/SA)(Onufriev *et al.*, 2004) (igb=5 and gbsa=1). Berendsen's thermostat(Berendsen *et al.*, 1984) (ntt=1) with a 1.0 ps relaxation time was used for the temperature equilibration stage (discussed below). Langevin dynamics(Pastor *et al.*, 1988) mimicked solvent collisions and regulated the temperature during the replica exchange simulations. A gamma value of 2 ps^{-1} was used for the Langevin dynamics because a low value improves barrier crossing(Loncharich *et al.*, 1992). For further efficiency, forces involving interactions beyond 8 Å were only evaluated every 4 time steps(Tuckerman *et al.*,

1991) (nrespa=4). A time step of 1 fs and distance cutoff of 12 Å was used for the equilibration and production runs.

Cα atoms were harmonically restrained to their crystallographic coordinates unless they were part of the loop. The restraint force constant was 0.3 kcal/mol/Å^2. This is a relatively mild restraint in the sense that it allows for some departure from the equilibrium position. Each system underwent a 2000 step steepest descent minimization (with Cα restraints) before any equilibration.

We employed replica exchange-molecular dynamics to improve loop sampling. In this method, *N* replicas of the system at *N* temperatures run simultaneously. Periodically, every 0.5 ps in this study, an exchange attempt is made between replicas of adjacent temperatures. The exchange is accepted according to the Metropolis criterion(Metropolis, 1953; Sugita & Okamoto, 1999). If an exchange attempt is successful, the momenta are rescaled such that:

(1) $$p^{new} = \sqrt{\frac{T_{new}}{T_{old}}} p^{old}$$

The temperatures employed in this study were 300, 308, 316, 325, 333, 342, 351, 361, 370 and 380 K. Rather than equilibrating each of these individually, simulations were equilibrated at 300, 325, 350 and 375K. The latter were heated up to the target temperature, from 0 K, over 5 ps and a subsequent equilibration of 45 ps; the timestep was 1 fs. Then the restart files were copied over as initial positions and velocities for the replica exchange runs. Specifically, the 300 K restart file was used for the 300 and 308 K replicas, the 325 K restart file was used for the 316, 325 and the 333 K replicas, the 350 K restart file was used for the 342, 351 and 361 K replicas and the

375 K restart file was used for the 370 and 380 K replicas. A replica exchange simulation for 100 ps, for further equilibration of each individual temperature, followed. During this stage: 1) exchange between replicas were attempted every 0.5 ps, and 2) a harmonic restraint was imposed between Cα atoms of "tip" loop residues and a point in the active site; the specifics are found in Table 1. The latter were in addition to the restraints on the Cα atoms not in the loop. A 5 ns production run, using the same Cα and loop restraints, followed.

Clustering

To identify conformational sub-states, the trajectories at 300 and 325 K were clustered using the MMTSB(Michael Feig, 2001) tools set. The metric for hierarchical clustering was the RMSD between all atoms (-jclust option in cluster.pl). This software tool chooses an optimized number of clusters based on the criterion by Xu and co-workders(Xu *et al.*, 1993). A total of 5000 structures were subjected to clustering from each trajectory (snapshots were taken at 1 ps intervals). The centroids of each cluster were refined as described below.

Loop prediction refinement of cluster centroids

The purpose of this step is to identify low-energy conformations that are as similar as possible to the cluster centroids, which may be quite strained because they are average structures. The loop region of each centroid was grafted onto the corresponding apo structure. A loop prediction calculation followed. These calculations were performed using the protein loop optimization program (PLOP)(Jacobson *et al.*, 2004). PLOP uses a hierarchical algorithm to search loop conformational space. Briefly, the first step is to build a set of putative backbone conformations from each termini to the

center residue. Those pairs that form a closed loop are further screened against steric clashes with the protein body and poor junction geometry. Loops that pass this filter are clustered and a representative for each cluster is refined. In the refinement stage, side chains are added and optimized and the entire loop is minimized. The resulting loops are ranked by energy. The energy function consists of the all-atom OPLS force field(Jorgensen & Tiradorives, 1988) and a Generalized Born(Feig *et al.*, 2004) solvation model with a solvent accessible surface area correction (GB/SA)(Gallicchio *et al.*, 2002).

During the loop refinement stage, Cα carbons were kept within 2 Å of their corresponding centroid coordinates, so that the predicted loops were quite similar to the centroid. In the absence of restraints, the loop prediction algorithm should find the apo structure. Tighter restraints usually failed to produce any solutions, i.e. loops that satisfied the restraints and were free from steric clashes or high internal strain. Side-chains within 3 Å of the loop were optimized for the L-Ala-D-Glu epimerase, triosephosphate isomerase and tyrosine phosphatase cases while a 4 Å cutoff was used for the others. Expanding the side-chain cutoff to 4 Å for AEE, triosephosphate isomerase and the tyrosine phosphatase cases included only a few more residues, distant from the active site, which changed little after the optimization (~10-15 degrees of rotation in methyl groups, for example), but deteriorated the loop prediction. Except for the cases of L-Ala-D-Glu epimerase and *Yersinia* tyrosine protein phosphatase, the loop prediction calculation included an extra residue on each loop termini. This was useful when the superposition was slightly offset and caused an awkward juncture geometry.

In some cases the loop prediction failed to generate low energy loops similar to the cluster centroids (i.e., subject to the Cα constraints). It was possible in these cases to generate low energy loops by relaxing the constraints to 3.5 Å. However, these problematic centroids were not close to the holo conformation, i.e., they are not likely to be relevant to ligand binding.

Docking into representative structures of each cluster

The docking calculations against the resulting low-energy loops were performed using GLIDE 3.5(Schrodinger Inc, 2004). Docking calculations were also performed against the apo and holo structures for comparison. The visualization and structure preparation were performed using Maestro 6.5. For every target, hydrogens were added, if missing, and ions were assigned the correct Macromodel atom type. Active sites were identified by either listing active site residues or identifying a bound ligand. The active site residues of L-Ala-D-Glu epimerase are Lys160, Lys162, Asp191, Glu219, Asp244, Lys268, Asp321 and the magnesium ion. For the epimerase, histidines 223 and 309 were protonated. The enolase active site consists of Lys396, Glu168, Asp246, Asp320, Lys345, Ser372 and both magnesium ions. In the other cases, either a ligand existed in the crystal structure or its coordinates were copied from the holo structure in order to mark the active site. In the case of spermidine synthase, the ligand coordinates were copied, after structural superposition, from the structure in the PDB file 1JQ3.pdb. For triosephosphate isomerase, the ligand coordinates were copied from PDB file 1R2R.pdb. A nitrate ion was bound in the active site of *Yersinia* protein-tyrosine phosphatase (PDB code 1YTN). Note that the correct stereochemical geometry and bond order were assigned to the crystallographic ligand that indicated the active site.

In the cases of AEE, Yersinia tyrosine protein phosphatase, spermidine synthase and triosephosphate isomerase the output of the loop prediction calculation was used in the docking calculation directly. In the other cases complications outside the scope of this work arose. In the case of GART, a cofactor, taken from the holo structure, was included in the docking calculations. This entailed superimposing the Cα atoms of the apo and holo proteins and copying the coordinates of the co-factor. The same procedure was performed in the case of enolase, where a second magnesium ion had to be included.

10,004 compounds from the Kyoto Encyclopedia of Genes and Genomics (KEGG)(Kanehisa & Goto, 2000; Kanehisa M, 2006) were obtained to generate a database for the docking calculations. The database preparation was described before(Kalyanaraman et al., 2005); briefly, polymers and monoatomic ions were removed and different chiral forms were explicitly enumerated using the Daylight software package("Daylight"). A total of 19007 entries made up the final database. In the case of the epimerase, all 400 possible dipeptides were added to the database. The substrates and co-factors for each system are listed in Table 2.

Ligand pose RMSD calculations excluded hydrogens. In cases where the crystallographic ligand and the known metabolic ligand differed, the calculation involved the intersection between the two structures as long as it appeared sensible. For the GART ligand, only the ring system was included in the RMSD calculation because it is the segment that fits in the pocket made by the loop in question. *Yersinia* tyrosine protein phosphatase was not included because the holo crystallographic

ligand is only a nitrate ion.

Hydrogen bonding statistics

Hydrogen bondings statistics were calculated using the ptraj module of Amber 8. The hbond command with the option "distance 3.5" calculated the hydrogen bond probabilities reported here.

Results and Discussion

Replica exchange simulation diagnostics

The exchange probabilities for the REMD simulations ranged from 31 to 41%. There was a significant amount of overlap in the potential energy probability distributions between replicas at adjacent temperatures. Figure 1 shows these distributions for the L-Ala-D-Glu Epimerase and is representative of the other cases. The corresponding plots for the other systems are included in supplementary materials.

Does the method produce holo-like loop conformations?

The proposed method involved sampling loop conformations via replica exchange molecular dynamics, clustering the resulting trajectories and subsequently refining the cluster centroids, via restrained loop prediction, to produce a representative structure per cluster (see Methods section for details). A set of representative structures resulted per case. Within each set the member closest to the holo form differed from it, with one exception, by less than a 2 Å RMSD from the holo coordinates (Table 3). The exception, the spermidine synthase case, appears to stem from insufficient sampling; a control calculation starting from the closed conformation yielded much

lower RMSD's (1.9 vs 2.9 Å) from the bound coordinates; several other representative structures from the control simulation spanned the 1.9-2.5 A RMSD range. A more efficient loop sampling algorithm, such as those based on a robotics approach(Coutsias *et al.*, 2004; Lee *et al.*, 2005), may resolve this issue.

The cluster centroids (Figure 2) provide a qualitative sense of the extent of sampling, but must not be over interpreted since the conformational space spanned by each cluster varies.

Conserved contacts between the apo and holo structures likely contributed to the success of the method by reducing the effective number of degrees of freedom during the simulations. For example, beta-sheet type hydrogen bonds in the AEE loop (2 involving Val26 and Phe10 and another between the carbonyl group of Arg24 and the amide proton of Thr21) were present for 77-87% of the 326K replica trajectory. Similarly backbone hydrogen bonds that form a helical turn in the triosephosphate isomerase loop were present 89-94% of the simulation time (326K replica). The dynamics predict which contacts remain fixed, so they need not be selected *a priori*. Control loop prediction calculations, with the loop midpoint restrained to the vicinity of its holo coordinates, did not yield the closed conformation which is consistent with the the need to restrain a subset of the hydrogen bond contacts during the conformational search.

Do docking scores "select" a holo-like conformer among the representative structures?

In theory, any of the representative structures could be the "correct" one and even that

structure could undergo some rearrangement after binding, so docking, being particularly sensitive to active site geometry, poses a highly stringent test for the predicted structures. Docking known ligands against the representative structures yielded lowest (best) docking scores for loops with lowest RMSD from the holo structure (Table 3), essentially mimicking the conformational selection mechanism *in silico*. Spermidine synthase and GART were omitted because of the poor sampling in the former and poor docking poses for both.

Do the predicted structures improve docking results with respect to the apo structure? We docked a metabolite library, prepared as described in the methods section, to dock against the predicted structures and evaluate their performance in the context of a virtual screening calculation. The aim of this type of calculation, where a metabolite library is used, is to identify substrate candidates for enzymes whose biological substrates are unknown, under the assumption that a high binding affinity correlates with the likelihood of being the actual substrate. Our method differed from other studies that docked into MD(Frembgen-Kesner & Elcock, 2006; K.M.Masukawa, 2000; Lin *et al.*, 2003; Wong *et al.*, 2005) snapshots in that it: 1) focuses on database screening and 2) examines large backbone rearrangements.

Docking into the selected structure resulted in improvement, sometimes significant, over the corresponding apo calculation. Table 3 lists the ranks, expressed as percentage out of the database, of the known substrate in the KEGG database when docked against the apo, holo and "selected" structures. Ranks for the holo structure are largely within 1% of the database, while the corresponding apo structure ranks are largely over the 8% mark. In 5 of the 6 cases the predicted structure results

outperform the corresponding apo calculations and the ranks were mostly below 2% of the database; the exception, spermidine synthase, is due to poor docking poses in both, the predicted and apo calculations (see poses section below). These results are clearly of practical importance for computer aided drug design, where receptor flexibility can pose a significant obstacle(Carlson, 2002b).

Are the docked poses correct?

Evaluating docked ligand poses is an even more stringent test because favorable dock rankings do not necessarily correspond to correct ligand poses. For example, docking against the GART apo structure yielded an excellent ligand rank (0.2 %) but an incorrect pose (13.2 Å RMSD).

We evaluated ligand poses from docking calculations against the predicted structures (Figure 4) and compared them to the corresponding apo and holo results. Docking into apo structures yielded very poor poses for the enolase, GART and spermidine synthase cases (Figure 4) while the corresponding holo results yielded a poor pose only for the spermidine synthase case. The predicted structure poses are comparable to the corresponding apo results except for the enolase case where it performed significantly better than the apo structure (1.4 vs 3.6 Å RMSD).

For GART and spermidine synthase the resulting poses were consistently poor using the predicted and apo structures (6.8 and 3.5 Å RMSD for GART and 12.5 and 13.2 Å RMSD for spermidine synthase). If the holo loop structure was grafted on the apo protein body and used as a docking target, the GART ligand remained outside the binding pocket while the spermidine synthase ligand docked correctly (almost perfect

overlap with the holo structure result within the binding site). So, in the GART case factors other than the loop structure contribute significantly to the poor docking pose and could not be rescued regardless of the quality of the loop prediction. Spermidine synthase is the only case where a poor loop prediction causes a poor pose, but that is not surprising given the high backbone RMSD (Table 3). The spermidine synthase simulation sampling was inadequate, as discussed above.

Interestingly, the resulting pose when docking into the predicted triosephosphate isomerase structure performs better than the corresponding holo result (2.3 vs 1.7 Å RMSD). Poses for the tyrosine protein phosphatase case were not examined because the holo structure had a nitrate ion as a ligand; however, the phosphate group binds in the pocket occupied by the nitrate ion.

Would the perfect loop structure yield holo-like docking results?

The results in Table 3 show differences between the predicted and holo structure docking results that may be partially due to structural variations outside the loop region. In order to isolate the effect of the predicted loop, docking was performed against the following two structures: 1) the predicted loop grafted on the holo protein structure (test case 1) and 2) the holo loop grafted on the apo protein structure (test case 2). In combination with the other docking runs, the first case isolates the effect of the loop geometry while the second one highlights the effect of other differences in the active sites.

Together, the results indicate subtle differences in the active site between the apo and holo form, in addition to the loop conformation, affect the docking calculations (Table

5). This is clear from the AEE case results where using the holo loop coordinates, grafted on the apo protein structure, did not improve the docking results relative to the predicted loop. The holo protein active site, in addition to the correct loop conformation, was required to improve the results further. Docking calculations on triosephosphate isomerase and enolase yielded similar results.

Conclusion

This work evaluates an approach that mimics certain aspects of conformational selection process, to account for ligand-induced loop latching in docking calculations. While there are limitations in our sampling methodology, the data suggests the approach is viable. The methodology proposed here was able to 1) produce holo-like loop structures 2) improve docking results in comparison to apo results and 3) yield accurate docking poses in 4 of the 6 cases. Except for 1 case all predicted loops are below 2 Å backbone RMSD from the crystallographic holo coordinates. In addition, the known ligand docking score was able to "select" the holo-like conformation from among the several conformational sub-states identified.

In the absence of a known ligand to select among cluster representatives, one can imagine: 1) combining the results from each cluster weighting according to the cluster population or 2) perform the virtual screening and empirical testing for all clusters since only a few are produced. A complementary approach would be to filter dubious structures; for example, loops with side-chains contacting catalytic residues could be discarded.

Sampling limited the accuracy of the spermidine synthase loop structural prediction and resulted in a poor docking pose. More efficient sampling methodology may ameliorate that limitation. Structural factors, other than the loop conformation, thwarted the predicted docking pose for GART. Thus, these cases do not undermine the viability of the general scheme proposed, but rather motivate work on these particular areas.

More broadly, this general approach may be applicable to other types of large receptor rearrangements. The important feature here is that it avoids an exhaustive receptor conformational search for every ligand pose; the flexibility is examined *a priori* and without information about putative ligands.

References

Almassy, R., Janson, C., Kan, C., & Hostomska, Z. (1992). Structures of apo and complexed escherichia coli glycinamide ribonucleotide transformylase. *Proceedings of the National Academy of Sciences of the United States of America, 89*(6114-6118).

Alvarez, M., Zeelen, J., Mainfroid, V., Rentier-Delrue, F., Martial, J., Wyns, L., et al. (1998). Triose-phosphate isomerase (tim) of the psychrophilic bacterium vibrio marinus. *Journal of Biological Chemistry, 273*(4), 2199-2206.

Aparicio, R., Ferreira, S. T., & Polikarpov, I. (2003). Closed conformation of the active site loop of rabbit muscle triosephosphate isomerase in the absence of substrate: Evidence of conformational heterogeneity. *Journal Of Molecular Biology, 334*(5), 1023-1041.

Babbitt, P. C., Mrachko, G. T., Hasson, M. S., Huisman, G. W., Kolter, R., Ringe, D., et al. (1995). Functionally diverse enzyme superfamily that abstracts the alpha-protons of carboxylic-acids. *Science, 267*(5201), 1159-1161.

Berendsen, H. J. C., Postma, J. P. M., Vangunsteren, W. F., Dinola, A., & Haak, J. R. (1984). Molecular-dynamics with coupling to an external bath. *Journal of Chemical Physics, 81*(8), 3684-3690.

Berger, C., Weber-Bornhauser, S., Eggenberger, J., Hanes, J., Pluckthun, A., & Bosshard, H. (1999). Antigen recognition by conformational selection. *Febs Letters, 450*, 149-153.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucl. Acids. Res., 28*(1), 235-242.

Cao, Y., Musah, R. A., Wilcox, S. K., Goodin, D. B., & McRee, D. E. (1998). Protein conformer selection by ligand binding observed with crystallography. *Protein Science, 7*(1), 72-78.

Carlson, H. A. (2002a). Protein flexibility and drug design: How to hit a moving target. *Current Opinion in Chemical Biology, 6*, 447-452.

Carlson, H. A. (2002b). Protein flexibility is an important component of structure-based drug design. *Current Pharmaceutical Design, 8*, 1571-1578.

Cavasotto, C. N., Kovacs, J. A., & Abagyan, R. A. (2005). Representing receptor flexibility in ligand docking through relevant normal modes. *Journal Of The American Chemical Society, 127*(26), 9632-9640.

Claussen, H., Buning, C., Rarey, M., & Lengauer, T. (2001). Flexe: Efficient molecular docking considering protein structure variations. *Journal Of Molecular Biology, 308*(2), 377-395.

Coutsias, E., Seok, C., Jacobson, M. P., & Dill, K. (2004). A kinematic view of loop closure. *Journal Of Computational Chemistry, 25*, 510-528.

D.A Case, T. A. D., T.E. Cheatham, III, L.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, B. Wang, D.A. Pearlman, M. Crowley, S Brozell, V. Tsui, H. Gohlke, J.Hornak, G. Cui, P. Beroza, C. Schafmeister, J.W. Caldwell, W.S. Ross, and P.A. Kollman. (2004). Amber (Version 8).

Dahms, T. E. S., Sainz, G., Giroux, E. L., Caperelli, C. A., & Smith, J. L. (2005). The apo and ternary complex structures of a chemotherapeutic target: Human glycinamide ribonucleotide transformylase. *Biochemistry, 44*(29), 9841-9850.

Daylight. Los Altos, CA.

Delano, W. L. (2002). The pymol molecular graphics system. San Carlos, CA, USA.

Feig, M., Onufriev, A., Lee, M. S., Im, W., Case, D. A., & Brooks, C. L. (2004). Performance comparison of generalized born and poisson methods in the calculation

of electrostatic solvation energies for protein structures. *Journal of Computational Chemistry, 25*(2), 265-284.

Foote, J., & Milstein, C. (1994). Conformational isomerism and the diversity of antibodies. *Proceedings of the National Academy of Sciences of the United States of America, 91*, 10370-10374.

Franklin, T. J., Morris, W. P., Jacobs, V. N., Culbert, E. J., Heys, C. A., Ward, W. H. J., et al. (1999). A novel series of non-nucleoside inhibitors of inosine 5 '-monophosphate dehydrogenase with immunosuppressive activity. *Biochemical Pharmacology, 58*(5), 867-876.

Freire, E. (1998). Statistical thermodynamic linkage between conformational and binding equilibria. In *Advances in protein chemistry, vol 51* (Vol. 51, pp. 255-279). San Diego: Academic Press Inc.

Frembgen-Kesner, T., & Elcock, A. (2006). Computational sampling of a cryptic drug binding site in a protein receptor: Explicit solvent molecular dynamics and inhibitor docking to p38 map kinase. *Journal Of Molecular Biology, 359*(1), 202-214.

Gallicchio, E., Zhang, L., & Levy, R. (2002). The sgb/np hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *Journal Of Computational Chemistry, 23*, 517-529.

Gerstein, M., & Chothia, C. (1991). Analysis of protein loop closure - 2 types of hinges produce one motion in lactate-dehydrogenase. *Journal of Molecular Biology, 220*(1), 133-149.

Gulick, A. M., Schmidt, D. M. Z., Gerlt, J. A., & Rayment, I. (2001). Evolution of enzymatic activities in the enolase superfamily: Crystal structures of the l-ala-d/l-glu epimerases from escherichia coli and bacillus subtilis. *Biochemistry, 40*(51), 15716-15724.

Gunasekaran, K., Ma, B. Y., Ramakrishnan, B., Qasba, P. K., & Nussinov, R. (2003). Interdependence of backbone flexibility, residue conservation, and enzyme function: A case study on 1,4-galactosyltransferase-i. *Biochemistry, 42*(13), 3674-3687.

Haider, N., Eschbach, M. L., Dias, S. D., Gilberger, T. W., Walter, R. D., & Luersen, K. (2005). The spermidine synthase of the malaria parasite plasmodium falciparum: Molecular and biochemical characterisation of the polyamine synthesis enzyme. *Molecular and Biochemical Parasitology, 142*(2), 224-236.

Ikeguchi, M., Ueno, J., Sato, M., & Kidera, A. (2005). Protein structural change upon ligand binding: Linear response theory. *Physical Review Letters, 94*, 078102.

Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J. F., Honig, B., Shaw, D. E., et al. (2004). A hierarchical approach to all-atom protein loop prediction. *Proteins-Structure Function And Bioinformatics, 55*(2), 351-367.

James, L. C., & Tawfik, D. S. (2005). Structure and kinetics of a transient antibody binding intermediate reveal a kinetic discrimination mechanism in antigen recognition. *Proceedings of the National Academy of Sciences of the United States of America, 102*(36), 12730-12735.

Jimenez, R., Salazar, G., Baldridge, K., & Romesberg, F. (2003). Flexibility and molecular recognition in the immune system. *Proceedings of the National Academy of Sciences of the United States of America, 100*(1), 92-97.

Jorgensen, W. L., & Tiradorives, J. (1988). The opls potential functions for proteins - energy minimizations for crystals of cyclic-peptides and crambin. *Journal of the American Chemical Society, 110*(6), 1666-1671.

Juszczak, L. J., Zhang, Z. Y., Wu, L., Gottfried, D. S., & Eads, D. D. (1997). Rapid loop dynamics of yersinia protein tyrosine phosphatases. *Biochemistry, 36*(8), 2227-2236.

K.M.Masukawa, H. A. C., J.A.McCammon. (2000). Technique for developing a pharmacophore model that accommodates inherent protein flexibility: An application to hiv-1 integrase. In O. F. Guner (Ed.), *Pharmacophore perception, development and use in drug design* (pp. 537). La Jolla, CA: International University Line.

Kalyanaraman, C., Bernacki, K., & Jacobson, M. P. (2005). Virtual screening against highly charged active sites: Identifying substrates of alpha-beta barrel enzymes. *Biochemistry, 44*(6), 2059-2071.

Kanehisa, M., & Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research, 28*(1), 27-30.

Kanehisa M, G. S., Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. (2006). From genomics to chemical genomics: New developments in kegg. *Nucl. Acids. Res., 34*), D354-357.

Klenchin, V. A., Schmidt, D. M., Gerlt, J. A., & Rayment, I. (2004). Evolution of enzymatic activities in the enolase superfamily: Structure of a substrate-liganded complex of the l-ala-d/l-glu epimerase from bacillus subtilis. *Biochemistry, 43*(32), 10370-10378.

Kobayashi, M., Watanabe, T., Xu, Y. J., Tatemori, M., Goda, H., Niitsu, M., et al. (2005). Control of spermidine and spermine levels in rat tissues by trans-4-methylcyclohexylamine, a spermidine-synthase inhibitor. *Biological & Pharmaceutical Bulletin, 28*(4), 569-573.

Korolev, S., Ikeguchi, Y., Skarina, T., Beasley, S., Arrowsmith, C., Edwards, A., et al. (2002). The crystal structure of spermidine synthase wih a multisubstrate adduct inhibitor. *Nature Structural Biology, 9*, 27-31.

Koshland, D. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America, 44*(98-104).

Kovacs, J. A., Cavasotto, C. N., & Abagyan, R. (2005). Conformational sampling of protein flexibility in generalized coordinates: Application to ligand docking. *Journal Of Computational And Theoretical Nanoscience, 2*(3), 354-361.

Lee, A., Streinu, H., & Brock, O. (2005). A methodology for efficiently sampling the conformational space of molecular structures. *Physical Biology, 2*, s118-s125.

Lin, J. H., Perryman, A. L., Schames, J. R., & McCammon, J. A. (2003). The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme. *Biopolymers, 68*(1), 47-62.

Loncharich, R. J., Brooks, B. R., & Pastor, R. W. (1992). Langevin dynamics of peptides - the frictional dependence of isomerization rates of n-acetylalanyl-n'-methylamide. *Biopolymers, 32*(5), 523-535.

Ma, B. Y., Kumar, S., Tsai, C. J., & Nussinov, R. (1999). Folding funnels and binding mechanisms. *Protein Engineering, 12*(9), 713-720.

Mande, S., Mainfroid, V., Kalk, K., Martial, J., & Hol, W. (1994). Crystal structure of recombinant human triosephosphate isomerase at 2.8 a resolution. Triosephophate isomerase-related human genetic disorders and comparison with the trypanosomal enzyme. *Proteins-Structure Function And Bioinformatics, 3*(5), 810-821.

McMillan, F. M., Cahoon, M., White, A., Hedstrom, L., Petsko, G. A., & Ringe, D. (2000). Crystal structure at 2.4 a resolution of borrelia burgdorferi inosine 5 '-

monophosphate dehydrogenase: Evidence of a substrate-induced hinged-lid motion by loop 6. *Biochemistry, 39*(15), 4533-4542.

Metropolis, N., Rosembluth, A., Rosembluth, M. and Teller, A. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087-1092.

Michael Feig, J. K., Charles L. Brooks, III. (2001). Mmtb tool set. The Scripps Research Institute.

Okoniewska, M., Tanaka, T., & Yada, R. Y. (2000). The pepsin residue glycine-76 contributes to active-site loop flexibility and participates in catalysis. *Biochemical Journal, 349*, 169-177.

Onufriev, A., Bashford, D., & Case, D. A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins-Structure Function and Bioinformatics, 55*(2), 383-394.

Pankiewicz, K. W., Lesiak-Watanabe, K. B., Watanabe, K. A., Patterson, S. E., Jayaram, H. N., Yalowitz, J. A., et al. (2002). Novel mycophenolic adenine bis(phosphonate) analogues as potential differentiation agents against human leukemia. *Journal of Medicinal Chemistry, 45*(3), 703-712.

Pastor, R. W., Brooks, B. R., & Szabo, a. (1988). An analysis of the accuracy of langevin and molecular-dynamics algorithms. *Molecular Physics, 65*(6), 1409-1419.

Pitici, F., Beveridge, D., & Baranger, A. (2002). Molecular dynamics studies of induced fit and conformational capture in u1a-rna binding: Do molecular substates code for specificity? *Biopolymers, 65*, 424-435.

Rhee, S., Parris, K. D., Hyde, C. C., Ahmed, S. A., Miles, E. W., & Davies, D. R. (1997). Crystal structures of a mutant (beta k87t) tryptophan synthase alpha(2)beta(2) complex with ligands bound to the active sites of the alpha- and beta-subunits reveal ligand-induced conformational changes. *Biochemistry, 36*(25), 7664-7680.

Sandak, B., Wolfson, H. J., & Nussinov, R. (1998). Flexible docking allowing induced fit in proteins: Insights from an open to closed conformational isomers. *Proteins-Structure Function And Genetics, 32*(2), 159-174.

Schrodinger Inc. (2004). Glide. New York, New York.

Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A., & Farid, R. (2006). Novel procedure for modeling ligand/receptor induced fit effects. *Journal Of Medicinal Chemistry, 49*(2), 534-553.

Simmerling, C., Strockbine, B., & Roitberg, A. E. (2002). All-atom structure prediction and folding simulations of a stable protein. *Journal of the American Chemical Society, 124*(38), 11258-11259.

Stuckey, J., Schubert, H., Fauman, E., Zhang, Z., Dixon, J., & Saper, M. (1994). Crystal structure of yersinia protein tyrosin phosphatase at 2.5 a and the complex with tungstate. *Nature, 370*, 571-575.

Sugita, Y., & Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters, 314*(1-2), 141-151.

Taylor, J. C., & Markham, G. D. (2003). Conformational dynamics of the active site loop of s-adenosylmethionine synthetase illuminated by site-directed spin labeling. *Archives of Biochemistry and Biophysics, 415*(2), 164-171.

Taylor, R., Jewsbury, P., & Essex, J. (2003). Fds: Flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. *Journal Of Computational Chemistry, 24*(13), 1637-1656.

Taylor, T. C., & Andersson, I. (1997). The structure of the complex between rubisco and its natural substrate ribulose 1,5-bisphosphate. *Journal of Molecular Biology, 265*(4), 432-444.

Teague, S. J. (2003). Implications of protein flexibility for drug discovery. *Nature Reviews Drug Discovery, 2*(7), 527-541.

Thomas, T., & Thomas, T. J. (2003). Polyamine metabolism and cancer. *Journal of Cellular and Molecular Medicine, 7*(2), 113-126.

Tonks, N. K. (2003). Ptp1b: From the sidelines to the front lines! *Febs Letters, 546*(1), 140-148.

Totrov, M., & Abagyan, R. (1997). Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins-Structure Function And Bioinformatics, Suppl. 1*, 215-220.

Tricot, G. J., Jayaram, H. N., Lapis, E., Natsumeda, Y., Nichols, C. R., Kneebone, P., et al. (1989). Biochemically directed therapy of leukemia with tiazofurin, a selective blocker of inosine 5'-phosphate dehydrogenase-activity. *Cancer Research, 49*(13), 3696-3701.

Tsai, C. J., Kumar, S., Ma, B. Y., & Nussinov, R. (1999). Folding funnels, binding funnels, and protein function. *Protein Science, 8*(6), 1181-1190.

Tsai, C. J., Ma, B. Y., Sham, Y. Y., Kumar, S., & Nussinov, R. (2001). Structured disorder and conformational selection. *Proteins-Structure Function And Genetics, 44*(4), 418-427.

Tuckerman, M. E., Berne, B. J., & Rossi, a. (1991). Molecular-dynamics algorithm for multiple time scales - systems with disparate masses. *Journal of Chemical Physics, 94*(2), 1465-1469.

Wedekind, J. E., Poyner, R. R., Reed, G. H., & Rayment, I. (1994). Chelation of serine-39 to mg2+ latches a gate at the active-site of enolase - structure of the bis(mg2+)-complex of yeast enolase and the intermediate analog phosphonoacetohydroxamate at 2.1-angstrom resolution. *Biochemistry, 33*(31), 9333-9342.

Williams, J. C., & Mcdermott, a. E. (1995). Dynamics of the flexible loop of triosephosphate isomerase - the loop motion is not ligand-gated. *Biochemistry, 34*(26), 8309-8319.

Wong, C. F., Kua, J., Zhang, Y., Straatsma, T. P., & McCammon, J. A. (2005). Molecular docking of batanol to dynamics snapshots of protein kinase a. *Proteins-Structure Function And Bioinformatics, 61*(4), 850-858.

Wong, C. F., & McCammon, J. A. (2003). Protein flexibility and computer-aided drug design. *Annual Review Of Pharmacology And Toxicology, 43*, 31-45.

Wu, M. G., & Deem, M. W. (1999). Efficient monte carlo methods for cyclic peptides. *Molecular Physics, 97*(4), 559-580.

Xu, S., Kamath, M. V., & Capson, D. W. (1993). Selection of partitions from a hierarchy. *Pattern Recognition Letters, 14*(1), 7-15.

Zhang, Z. Y. (2001). Protein tyrosine phosphatases: Prospects for therapeutics. *Current Opinion in Chemical Biology, 5*(4), 416-423.

Protein	PDB structure codes	Loop Residues	Selection Radius (A)	Tip residues	Active site reference point	Restraint radius (A)
Yersinia Tyrosine protein phosphatase	1YPT chain A, 1YTN	351-360	12	Gln357	Cα Ala405	13
L-Ala-D-Glu Epimerase	1JPM chain B, 1TKK	14-28	8	Leu23, Arg24, Ala22	Position of Mg ion in active site	12
Triosephosphate isomerase	1R2T chains A and B	166-177	12	Ile170, Gly171	P atom in PGA ligand from pdb structure 2YPI	8
Enolase	1EBH, 1EBG	35-45	12	Ser39, Thr40, Gly41	Mg ion in active site	12
GART	1CDE and 1GCR chain B	141-146	12	Gly142, Leu143, Asp144	Cα of Leu92	12
Spermidine Synthase	1INL chain A and C	171-180	8	Gly176, Gly177, Gly178	C1* atom in AAT ligand in pdb stucture 1JQ3	10

Table 1. Summary of parameters used to prepare protein systems for simulation.

System	Substrate	Co-factors or ions
Yersinia Protein Tyrosine Phosphatase	Capped phosphotyrosine	
L-Ala-D-Glu Epimerase	L-Ala-L-Glu	
Triosephosphate Isomerase	D-Glyceraldehyde 3-phosphate (C00118)	
Enolase	D-Glycerate 2-phosphate (C00631)	Mg^{2+} ions. The ion co-crystallized with ligand was included for all Enolase docking calculations
GART	10-Formyltetrahydrofolate (C00234) (tautomer)	5'-Phosphoribosylglycinamide (GAR) (C03838)
Spermidine Synthase	S-Adenosylmethioninamine (C01137)	

Table 2. List of ligands designated substrate and co-factor for each system. In parenthesis is the KEGG database entry. In some cases, the ligands were not part of the database and had to be newly prepared.

	Holo vs Apo loop RMSD (Å)	Predicted Backbone RMSD (Å)	Docking ranks in percentages from the database		
			Apo	Holo	Predicted
L-ala-D-Glu Epimerase	6.7	1.3	17.7	0.9	3.1
Yersinia Tyrosine Protein Phosphatase	3.3	1.4	14.0	1.1	1.2
Enolase	5.6	1.8	6.0	0.04	2.3
Triosephosphate Isomerase	4.3	1.9	8.9	0.8	4.5
phosphoribosylglycinamide formyltransferase (GART)	4.3	1.6	7.9*	0.06*	4.9*
Spirmidine Synthase	5.6	2.9	0.2*	0.03	2.0*

Table 3. Summary of docking results into apo, holo and predicted structures. The rankings are expressed in percentages from the database (19000 total). An asterisk indicates the docking calculation produced an incorrect pose.

Cluster ID/Description	Backbone RMSD (A)	Docking (Glide) Score
L-Ala-D-Glu Epimerase		
Apo	6.72	-9.88
Pred-1	3.11	(No reasonable poses found)
Pred-2 (SELECTED)	1.3	-10.0
Holo	0.0	-13.3
Enolase		
Apo	5.6	-13.6
Pred-1 (SELECTED)	1.8	-14.1
Holo	0.0	-19.6
TIM		
Apo	4.3	-6.8
Pred-1	3.1	-6.1
Pred-2 (SELECTED)	1.9	-6.6
Holo	0.0	-9.0
***Yersinia* protein tyrosine phosphatase**		
Apo	3.3	-6.3
Pred-1	1.4	-5.1
Pred-2	1.8	-5.2
Pred-3	1.5	-5.8
Pred-4	1.6	-6.6
Pred-5 (SELECTED)	1.4	-7.2
Holo	0.0	-9.4

Table 4. Lowest (best) docking score corresponds identifies lowest RMSD cluster centroid. The GART and spermidine synthase results are skewed because the docking poses are incorrect. The latter introduces other variables, in addition to the loop conformation, that affect the docking score.

	L-Ala-D-Glu Epimerase	Yersinia Protein Tyrosine Phosphatase	Enolase	Triosephosphate Isomerase	GART	Spermidine Synthase
Apo	9.3	14	6.0	8.9	7.9*	0.16*
Holo	0.1	1.1	0.04	0.06	0.06	0.03
Predicted structure	3.1	1.2	2.3	4.5	4.9*	2.0*
Predicted Loop on holo structure	1.3	**	0.5	4.5	15.6*	0.9*
Holo loop on apo structure	3.3	23	0.6	4.9	12.5*	0.1*

Table 5. Docking results against apo, holo, predicted and limiting cases. 1) Grafting the holo loop on to the apo structure and 2) grafting the predicted loop onto the holo protein structure produced the two limiting case structures. The results are expressed in percentages of the 19,000 compound database. An asterisk indicates the docking calculation produced an incorrect pose. A double asterisk indicates no poses were found during the docking calculation.

	Xtal vs. Holo pose RMSD (Å)	Holo vs. Apo pose RMSD (Å)	Holo vs. Predicted pose RMSD (Å)
L-Ala-D-Glu Epimerase	1.7	1.8	1.6
Enolase	1.9	3.6	1.4
GART	1.8	3.5	6.8
TIM	2.5	2.4	2.4 (1.7)
Spermidine synthase	1.0	13.2	12.5

Table 5. RMSD between the crystallographic pose and resulting docked pose into the holo structure, between the resulting poses docking between the holo and apo structures and between the resulting poses of docking into the predicted and holo structure. In parenthesis is the comparison against the crystallographic pose rather than pose predicted by docking into the holo structure.

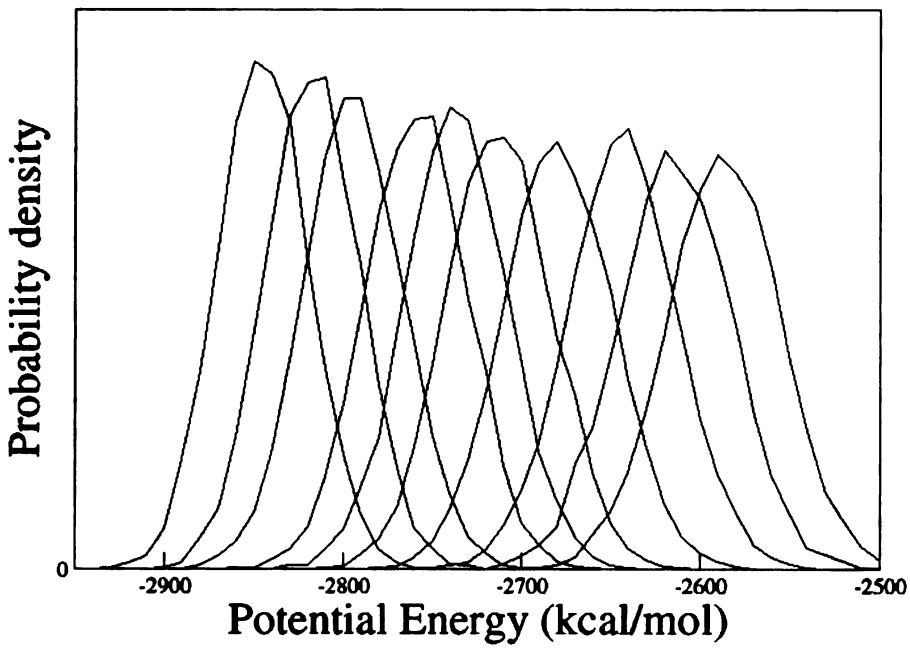


Figure 1. Probability density distributions for replicas during the replica exchange molecular dynamics simulation of L-Ala-D-Glu epimerase.

	Cluster centroids	Refined cluster centroid
L-Ala-D-Glu Epimerase		
Enolase		
GART		
TIM		
***Yersinia* protein tyrosine phosphatase**		
Spermidine Synthase		

Figure 2. Cluster centroids. All centroid structures are colored yellow, the holo structure is colored wheat and the refined cluster centroid is colored green. The second column includes the refined cluster centroid closest (per backbone RMSD) to the holo coordinates and its corresponding raw centroid structure. The holo structure is shown in both columns for reference. Only non-redundant figures clusters are shown; if two clusters overlapped in backbone conformations, only one of them was shown. The clusters of spermidine synthase were separated into three groups for clarity.

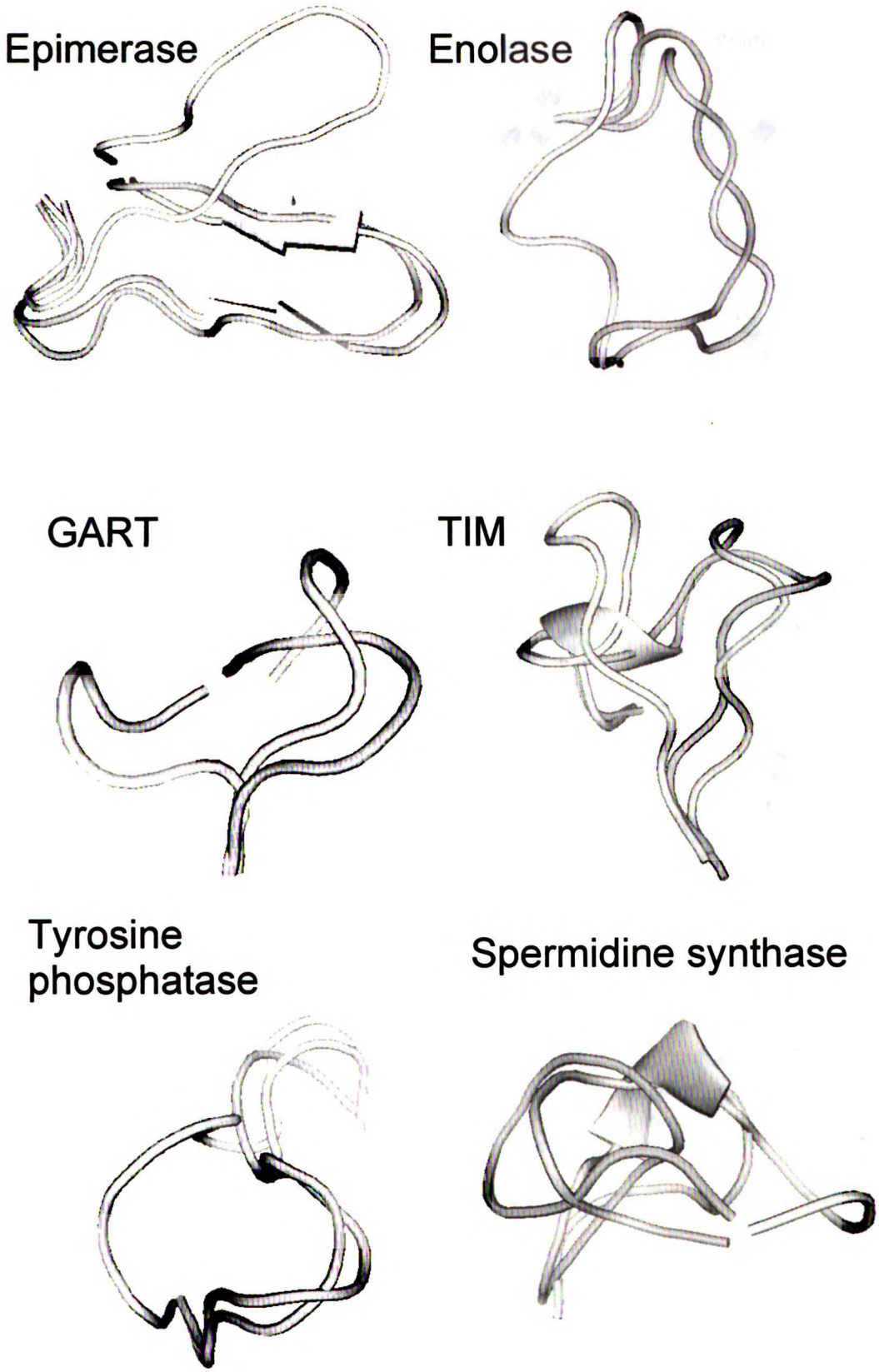


Figure 3. Predicted structure loops for all of the systems examined. In white is the apo structure, in green the predicted structure and in "wheat" the holo structure.

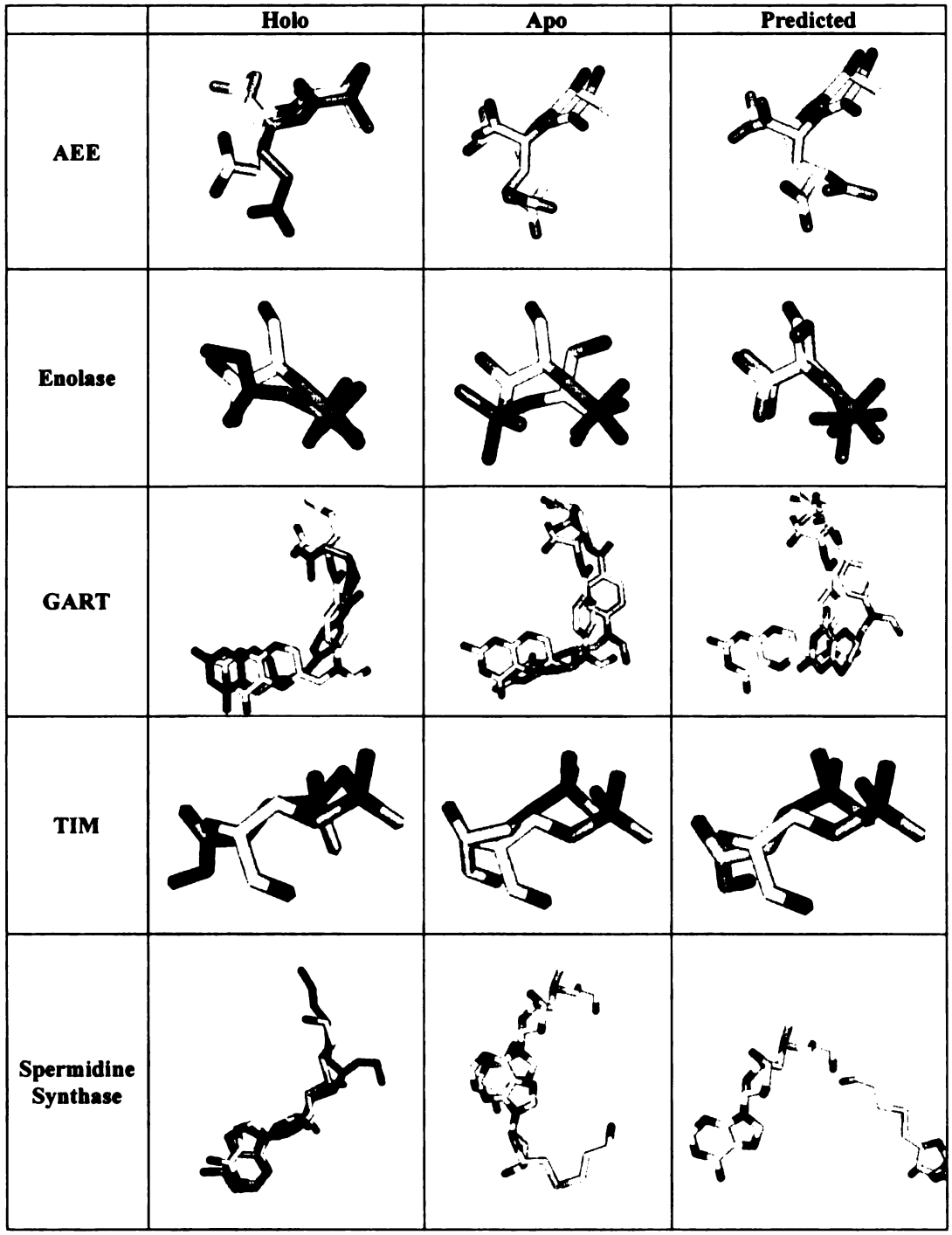


Figure 4. Poses of docked ligands. In each set the ligand with magenta carbon atoms is the crystallographic ligand pose, figures with yellow carbon atoms correspond to the docked ligand into the holo crystal structure and figures with the green carbon atoms correspond to either the docking pose into the apo or predicted structures.

Chapter 2: Limitations of the replica exchange method

Introduction

The simulation of flexible loops requires efficient sampling because the motions of interest can involve backbone rearrangements that are difficult to sample exhaustively. The replica exchange method (Mitsutake *et al.*, 2003a, 2003b) (REM) was developed to ameliorate some of these difficulties by overcoming energy barriers more easily. However, loop simulations using REM failed to explore known conformational sub-states. In this chapter I explore how a strong entropy temperature dependence may in fact diminish efficiency in REM simulations when higher temperature replicas are added.

In the rest of the introduction I discuss the basics of statistical mechanics, Monte Carlo integration, the force field description of bio-molecules and REM. The functional form of the force field description for bio-molecules motivates the need for efficient sampling methodology such as REM.

Formalism of statistical mechanics

Statistical mechanics provides a procedure to calculate bulk properties of materials from their molecular properties. It successfully predicts the equation of state for ideal gases and other ideal cases(Dill & Bromberg, 2002). Given a set of

external conditions (pressure, temperature, ionic strength, etc.) and the resulting quantum states (eigenfunctions and eigenvalues of the Hamiltonian operator) of the molecule in question, thermodynamic properties of an ensemble of these molecules are attainable. This requires the calculation of a so-called partition function. If the external temperature and volume are constant, the partition function, q, can be calculated as follows:

(1) $$q = \sum_i e^{\frac{-E_i}{kT}}$$

Where the index (i) delineates each individual quantum energy level of the molecule, E_i represents the energy of the ith state, T corresponds to the absolute temperature and k is Boltzmann constant (1.38×10^{-23} J/K). The sum is carried out over all of the quantum states of the molecule. For other conditions, e.g. constant pressure and temperature, similar expressions are used(Hill, 1986). For a property m, the bulk value can be calculated by:

(2) $$\langle m \rangle = \frac{\sum_i m(i) e^{\frac{-E_i}{kT}}}{q}$$

Where m(i) indicates the value of the ith state of the molecule. This is a weighted average over all of the quantum states.

If the system in question is described classically, i.e. the energy is not quantized, the partition function can be calculated by the following integral:

$$q = \int_{-\infty}^{\infty} \Omega(E) e^{\frac{-E}{kT}} dE \tag{3}$$

In equation (3) Ω(E) represents the density of states, that is the number of states with an energy between E and E+dE. This is essentially the same as equation 1, except that the "sum" is carried out over the energy not the quantum states.

Evaluating equations 1 and 2 can be challenging unless the quantum states and, subsequently, the sums can be calculated analytically. Otherwise, this problem poses the challenge of enumerating every quantum state. A partial sum may approximate q by ignoring states with high energy, such that they contribute little to the sum. However, this requires the identification of such states *a priori*. The same is true of equation (3).

Monte Carlo integration

Monte Carlo integration is an efficient method to calculate averages of functions of random variables(Kalos & Whitlock, 1986). It is particularly efficient when only a relatively small portion of the range in question contributes significantly to the integral in question. An example illustrates this clearly. Suppose the aim is to evaluate the integral:

(4) $$\int_{-\infty}^{\infty} e^{-ax} f(x)dx$$

Where "a" is a large number and f(x) is bound so that the integrand quickly converges to zero. Naïve standard quadrature methods would include contributions at very large values of x. The latter would be inefficient since only a small interval near the origin actually contributes significantly to the integral. Clearly, inspection of the integrand would suggest the numerical sum should be truncated near the origin, however, more complicated cases, especially those involving many variables, would not be as simple. Monte Carlo integration is a way to focus on the relevant regions of space. Using Monte Carlo integration, the integral in equation (4) would be calculated by computing the average of f(x) given that x is a random variable and has a probability density function:

(5) $$p(x) = e^{-ax}$$

So, the procedure would be to gather a sample of x-coordinates from the distribution (5) and compute their average f(x) value.

Sampling involves generating a set of x-coordinates that obey the target distribution i.e. equation (5) in the example. In a typical calculation, a set of x-coordinates are generated sequentially, making transitions from a previous position to the next one. Each transition is achieved by: 1) postulating a new x-

coordinate and 2) rejecting or accepting it based on a criterion that will reproduce the correct probability distribution. The initial position should not affect, in principle, the final outcome, so it could be selected at random. If the new position is rejected then the previous x-coordinate is repeated in the sequence of coordinates, otherwise the new position is the next coordinate in the sequence. The overall probability of making the transition equals the product of: 1) the probability of proposing that particular transition and 2) the probability of accepting it. The process of sequentially taking these Monte Carlo steps is sometimes called a Monte Carlo simulation.

Coordinate transitions are postulated via a Monte-Carlo "move-set", that is, an algorithm that explores the coordinate space without bias. A bias must be avoided because it would skew the resulting distribution obtained from the calculation. One approach is to make "moves" to nearby points. An example is moving +1 or -1 units. If the x-interval for integration was finite and small (rather than the infinite limits of equation (4)) one could choose random numbers from that interval. If those conditions are not met the simulation will likely be inefficient because, since the probability density is nearly zero far from the origin, most of the proposed moves will be rejected.

The acceptance criterion serves as a restraint to yield the target distribution for x. In order to do so, it must maintain detailed balance. That is, given two x positions q and r:

(6) $$P(q \rightarrow r) \bullet P(q) = P(r \rightarrow q) \bullet P(r)$$

This means the probability of transition from q to r times the probability of being in q must equal the probability of transition from r to q times the probability of being in r. Essentially (6) ensures a microscopic equilibrium between pairs of coordinates. One can insert the probabilities of p(r) and p(q) from the desired probability distribution (from equation (5), for example) and calculate the ratio of the transition probability from q to r versus the transition probability from r to q. So, by maintaining this ratio of transition probabilities the correct distribution is guaranteed. The Metropolis criterion(Metropolis *et al.*, 1953) satisfies detailed balance and maximizes the possible transition probabilities. It can be stated as follows:

(7) $$p(r \rightarrow q) = \begin{cases} 1 & B \geq 1 \\ B & B < 1 \end{cases}$$

Where:

(8) $$B = \frac{p(q)}{p(r)}$$

That is, if B is greater than or equal to one, accept the move, otherwise accept it with the probability equal to B. For the present work, the canonical distribution, that is the distribution for a thermodynamic ensemble at constant temperature, volume and number of particles (NVT) is the most relevant. In that case, B can be expressed as:

(9) $$B = e^{-\frac{(E(q)-E(r))}{RT}}$$

Where R equals the universal gas constant (8.314 J/K), T represents the absolute temperature and E the energy of the given coordinate (q or r).

Thus, to calculate the mean of a function f(x), where x is a random variable that obeys a given probability distribution, one can use the Metropolis criterion to generate a series of x coordinates that obey the distribution and subsequently average them. This can be applied in calculating molecular averages described by equation (2).

The advantage of this approach is that it focuses on the most important regions without user intervention. This is critical as the probability functions become more complex and involve many variables.

Applying statistical mechanics to biomolecules

Bio-molecules, such as proteins or nucleic acids, are unbranched heteropolymers(Alberts *et al.*, 2002). The three-dimensional arrangement, or conformation, of these molecules is critical for their function. In principle one could study them by calculating their quantum states, calculating their partition function and computing their bulk properties as described above. However bio-

molecules do not lend themselves to quantum calculations, in practice, because: 1) they often involve thousands of atoms and 2) can adopt many 3-D orientations or conformations. In the case of a protein of one hundred residues, assuming two backbone states per residue, the backbone conformational space is of the order of 2^{100}. Performing quantum calculations on all of these conformations is not practical, so they are often described using approximate models.

One approach is to use a so-called force field of atomistic resolution i.e., each atom is explicitly included. Each atom is assigned a charge and van der Waals parameters. Atomic bonds and angles are harmonically restrained to their equilibrium positions while a trigonometric function describes a dihedral potential. A potential energy can be calculated based on these interactions and its negative potential yields the force on each atom. This representation tries to reproduce the potential energy surface that govern the nuclei dynamics(Frenkel & Smit, 2001), as per the Born-Oppenheimer approximation. One common implementation has the following form(Schlick, 2002):

(10a) $$E = E_{bond} + E_{Angle} + E_{Tor} + E_{LJ} + E_{coul}$$

(10b) $$E_{bond} = \sum_{i,j} S_{ij}(r_{ij} - \langle r_{ij} \rangle)^2$$

(10c) $$E_{Angle} = \sum_{i,j,k} K_{ijk}(\cos(\theta_{ijk}) - \cos(\langle \theta_{ijk} \rangle))^2$$

(10d) $$E_{tor} = \sum_{ijkl} \sum_{n} \frac{V_{nikl}}{2}\left(1 \pm \cos\left(n\tau_{ijkl}\right)\right)$$

(10e) $$E_{LJ} = \sum_{ij}\left(\frac{-A_{ij}}{r^6_{ij}} - \frac{B_{ij}}{r^{12}_{ij}}\right)$$

(10f) $$E_{coul} = \sum_{ij}\frac{q_i q_j}{\varepsilon r_{ij}^2}$$

Where, the subscripts bond, Angle, tor, LJ and coul represent bond, angle, torsion, Lennard Jones and coulombic respectively. In the expressions for bond, angle and torsion potentials the indices run over the adjacent atoms that make the bond, angle or torsion, respectively. In the case of the Lennard-Jones and Coulombic portions the indices run over the non-bonded atoms.

Given the description from equation (6a), a probability for a given structure, or set of Cartesian coordinates, is given by:

(11) $$P(\vec{x}) = \frac{e^{\frac{-E(\vec{x})}{RT}}}{q}$$

Where the vector $\vec{x}$ denotes the set of Cartesian coordinates of the molecule, T represents the absolute temperature and q the partition function. Monte-Carlo integration could be particularly useful here given that: 1) there are thousands of degrees of freedom in these systems and 2) only a miniscule fraction of the coordinate space significantly contributes to the partition function.

Problems applying Monte Carlo methods to biomolecular systems

The application of Monte-Carlo integration to biomolecules proved difficult due to the rugged nature of the energy function (Equation 10a). The simulations rarely overcome the high energy barriers between the energy minima and the resulting calculation is incorrect. The key problem is devising a move-set that can efficiently interchange among the many energy minima.

Replica Exchange Method

In order to circumvent the short-comings of regular Monte Carlo calculations, Okamoto et al(Mitsutake et al., 2003a) proposed to simulate instead a generalized ensemble that spans several temperatures. Several copies of the same system are concurrently simulated at different temperatures. Periodically, exchanges between them are proposed i.e., they are part of the move-set. High temperature simulations can more easily overcome energy barriers and, by virtue of these exchanges, they allow lower temperature simulations to do so as well. Typically 1,…,n replicas with temperatures corresponding T1,…, Tn are simulated. Exchanges are accepted based on this criterion:

(12a) $$P(x_i^{T1} \leftrightarrow x_j^{T2}) = \begin{cases} 1 & if \quad B \leq 0 \\ \exp(-B) & if \quad B > 0 \end{cases}$$

Where,

(12b) $$B = \left(\frac{1}{T_1} - \frac{1}{T_2}\right) \bullet \left(E(x_j^{T2}) - E(x_i^{T1})\right)$$

Where the function P represents the probability of exchange, the indices (i) and (j) label the particular copy of the system and their temperatures have the subscripts 1 and 2. E(x) denotes the potential energy of the system.

The number of replicas and temperature spacing can be optimized to achieve the highest exchange probabilities. However, it is not clear that increasing the number of replicas, and thus the highest temperature, necessarily improves convergence of the simulation. Furthermore, given limited computing resources, the possible benefits of including higher temperatures must be weighed against having shorter individual trajectories. Here, these issues are explored using a model system and the role of entropy in barrier crossing is explicitly considered.

Methods

Description of the 2-D model

In order to assess the efficacy of the replica exchange method for overcoming entropic barriers, a two-state model system was used (Figure 1). I am particularly interested in cases where, given two macroscopic states, the number of accessible microscopic states within one of them grows significantly more rapidly than in the other. Replica exchange, using high temperature replicas, assumes the higher temperature replicas will improve sampling by avoiding kinetic trapping. However, higher temperatures favor more entropic states, which can trap the

simulations in highly entropic basins and not transition to the states of interest, those with fewer accessible sub-states.

In the model system, there are two states, A and B. Each state consists of a square 2-D grid and they are connected by a line of 10 grid points. The grid spacing is 1 unit and the dimensions of the 2-D regions are temperature dependent. All the grid points in state A have a potential energy of -3 kcal/mol, while all those in state B have -2 kcal/mol. The potential energy of the grid points connecting the two states is 0 kcal/mol. The entropy at a given temperature T_f can be calculated by:

$$S(T_f) - S(T_i) = \int_{T_i}^{T_f} \frac{C_p}{T} dT = k_b \ln\left(\frac{\Omega(T_f)}{\Omega(T_i)}\right) \tag{16}$$

Where S(T) denotes the entropy as a function of temperature, Ω(T) denotes the number of accessible states at temperature T, and C_p represents the heat capacity of the system. Using equation 16 one can calculate the number of accessible states for a given temperature if Cp is given. For this purpose I approximated Cp using the following truncated Taylor expansion:

$$C_p = A + BT + CT^2 \tag{17}$$

Where A=2.0, B=0.08 and C=0.001. The dimensions of states A and B at 300 K were 5x5 units in both cases. Using equations 16 and 17, the state at 300 K was used to calculate the dimensions of each state for each different temperature. Monte Carlo replica exchange simulations using this system were performed 10 times. They involved 64 replicas, spanning the temperatures 300-2367 K. The move set simply involved a single unit displacement in one of the axes. That is, +/-1 in the x or y direction. 100 exchange attempts were performed, where every pair of adjacent temperatures attempted a switch of coordinates. 10^6 Monte Carlo steps separated the exchange attempts. Barrier crossing was defined as, during these 10^6 steps, starting in one state and sampling at least once a point in the other state.

Description of the 1-D system.

A 1-D potential was devised to study convergence in replica exchange simulations given a fixed amount of CPU time. Here we fix the total number of Monte Carlo steps (number of replicas * total number of steps at each temperature). The 1-D potential is represented by a parabola on the negative x-axis and a "V" shaped potential on the positive x-axis. The "V" shape potential is described by a line, with slope m, from the point (0,0) to some minimum (x_min, y_min) on the fourth quadrant of the coordinate system on the interval [0, x_min]. On the interval (x_min, infinity), a line with slope –m describes the potential energy. So, the coordinates of the minimum on the positive x-axis, x_min and

y_min, specify the potential energy function on that side of the y-axis. A similar situation occurs on the negative x-axis, where the minimum of the parabola, and the fact that it must cross the origin, fix its analytical form. Here the Monte Carlo steps were +/- 0.5 units and randomly chosen.

For the current calculation the potential minimum on the positive x-axis was (90.0, -9.2), while on the negative x-axis it was (-2.0, -13.1). These bounds yield, going from the positive to the negative axis basin, an enthalpy difference of -11.0 kcal/mol and an entropic contribution of 9.38 kcal/mol. The overall free energy difference is -1.62 kcal/mol.

The total number of Monte Carlo steps, summed over all replicas, is 10^9. Each replica underwent 10,000 steps between exchange attempts. 1, 10, 20, 50, 80, 100 and 200 replicas were used. The number of exchanges was adjusted so to keep constant the total number of Monte Carlo steps.

The population on the positive x-axis basin, for the 300 K replica, was monitored as a metric of convergence. The calculations were performed 10 times and standard deviations were calculated; error bars spanned 2 times the standard deviation.

Results and discussion

Monte Carlo simulations of biomolecules, explore a particularly rugged energy landscape, as described by equations 10a-10f. It is crucial to overcome free energy barriers between nearby minima in order to obtain the adequate distribution. There will be an average number of Monte Carlo moves between crossings and this quantity determines how long the simulation needs to be in order to converge. Since replica exchange relies on high temperature replicas to overcome barriers, this motivates the question of how the transition frequency, the inverse of the average number of steps between crossings, changes with temperature.

Conceptually, it is useful to draw parallels to transition-state theory and kinetics. In that framework one can imagine two states, A and B, with a free energy barrier between them. Crossing from A to B occurs in two steps. Step 1 involves reaching a "transition-state" or excited state that is apt for barrier crossing. Step 2 involves going from this transition state to state B. Thus the overall barrier crossing probability is the product of the probability of reaching the transition state and the transition probability of going to B.

(13) $$P(A \rightarrow B) = P(A \rightarrow ABts) * P(ABts \rightarrow B)$$

Where $P(A \rightarrow B)$ symbolizes the transition probability from A to B, $P(A \rightarrow ABts)$ corresponds to the probability of reaching the transition state, from A, and $P(ABts \rightarrow B)$ represents the probability of going from the transition state to B. The

transition state is in thermodynamic equilibrium with A, and therefore entropy and enthalpy affect the transition state population. The kinetic rate is expressed by(Dill & Bromberg, 2002):

$$k(T) = \left(\frac{k_b}{h}\right) T e^{\frac{\Delta S^{ts}}{R} - \frac{\Delta H^{ts}}{RT}} \tag{14}$$

Where k(T) denotes the kinetic rate constant, k_b, h, R, ΔS^{ts}, ΔH^{ts}, T and *e* represent the Boltzmann constant (1.380 E-23 J/K), Planck's constant (6.64 E-34 J-s), the universal gas constant, entropy change of reaching the transition state, enthalpy cost of reaching the transition state, temperature and the natural logarithm base (2.718), respectively. Equation 14 describes the dependence of the rate constant as a function of temperature. However, the entropy and enthalpy changes also vary with temperature. The entropy change dependence on temperature can be expressed by a truncated Taylor expansion:

$$\Delta S(T) = \Delta S(300K) + a(T - 300) \tag{15}$$

If the first order coefficient, a, is negative and large, it can dominate the rate constant and diminish it at high temperature. Physically, that corresponds to the case where the number of accessible states in "B" increases with temperature significantly faster than those in the transition state. The latter is not an unreasonable situation given the high dimensionality of the space in question. To illustrate this, consider the volume of a hypercube with n dimensions and sides of

length L, L^n. The effect of a change in L is accentuated by the exponent n and, since biomolecules involve thousands of atoms, the effective n is also of that order of magnitude in these simulations. Figure 1 clearly depicts a maximum in the rate vs temperature plots for values of "a" that are large enough. This analysis does not include the enthalpic temperature dependence and the Taylor expansion may not be valid over this wide temperature range. For this reason I employ model systems to further explore this issue.

While Monte Carlo simulations do not reproduce kinetics and clearly do not have to obey transition-state theory, some of the concepts discussed are certainly useful in examining the temperature dependence of transition probabilities in Monte Carlo simulations. The accessible conformational space will increase at higher temperatures and the probability of barrier crossing will change accordingly. The rate reduction with temperature will be most pronounced if the free energy barrier is mostly entropic.

In high temperature simulations involving long loops (those involving at least 10 residues) it appears such barriers are prevalent. They explore disordered conformational space while extending into solvent and lack intraloop contacts present in the native structure. Performing simulations at higher temperatures did not stimulate the formation of the intraloop contacts, so it appears this strategy of increasing the temperature may be futile in these cases. Since a number of specific contacts must be made to sample native-like structures, it appears this

barrier could possibly be largely entropic. This motivated this inquiry on the limitations of the replica exchange method for cases involving a significant entropic barrier. The latter was assessed in terms of the number of barrier crossings as a function of temperature in a replica exchange simulation, using a 2-D model, and the convergence of the simulation given a fixed amount of computational power (CPU time), was assessed using a 1-D model. Both of these calculations are described in the Methods section.

At this point it may be worth noting that the replica exchange method is not limited to exchanges between replicas at different temperatures. Hamiltonian exchange, of which temperature exchange is one type, involves replica exchanges between systems with slightly different potential energy functions. The advantage of Hamiltonian exchange is that it permits focused modulation of the potential energy function, so that only the relevant barriers are lowered. Using kinetic energy, via temperature regulation, essentially lowers all of the enthalpic barriers simultaneously which opens up a huge conformational space that needs exploration. Here, however, I will continue to only consider replica exchange between systems at different temperatures.

Replica exchange simulations of the 1-D model, with a varying number of replicas, were performed. The simulations using 80 or 100 replicas clearly yield results, within margin of error, that coincide with the theoretical value, as indicated in Figure 2. However, as the number of replicas is increased to 200, the

calculated population of state A drifts away from the correct result by slightly more than 2 standard deviations. This deterioration could stem from: 1) the shortening of the simulations at each temperature or 2) a decrease in barrier crossing at higher temperatures. The latter was explored using the 2-D model.

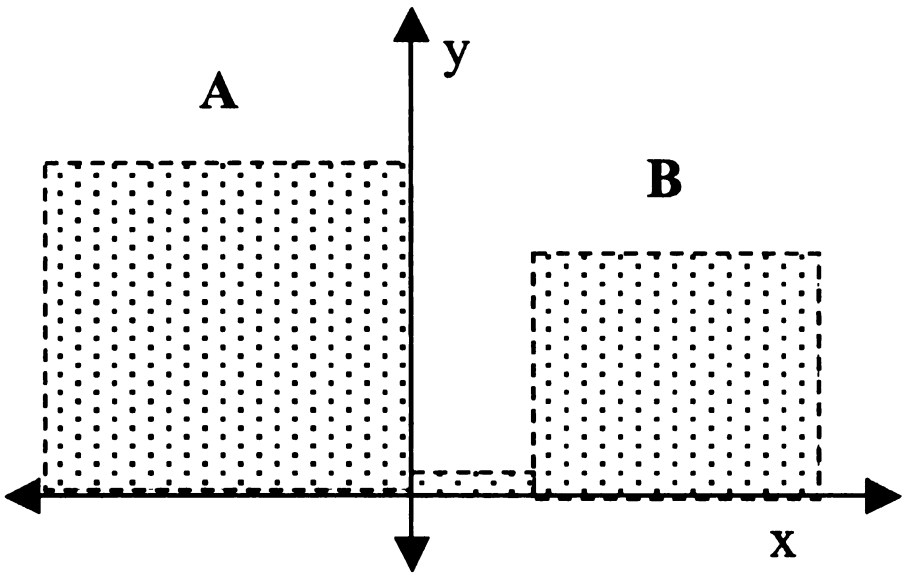
Barrier crossings counted during the 2-D model simulation, as a function of temperature, are shown in figure 3. The simulations were repeated 10 times and their results nearly overlap; the relative error is very small. Clearly the probability of barrier crossing diminishes with temperature. As the temperature increases, more states become available and that reduces the probability of being at the entrance of the "corridor" connecting the two states.

While this model may appear somewhat artificial, kinetic temperature dependence of this sort occurs in real systems. Arrhenius plots of protein folding rates often have a downward concavity(Chan & Dill, 1998) while endothermic chemical reactions also slow down with increased temperature. While the reasons behind these other cases are different, the same problem, of optimizing the number of replicas to maximize efficiency, applies.

Conclusion

The results in this chapter highlight a particular limitation of the replica-exchange method when used to overcome significant entropic barriers. At higher

temperatures more states become available and enthalpic barriers are less hindering. However, the number of states growth, with temperature, can, in extreme cases, have a negative effect on the barrier crossing rate. This can render using higher temperature replicas detrimental. In some cases, such as protein folding, there is a maximum in the rate versus temperature plot, and care must be taken to avoid using temperatures above that extremum.



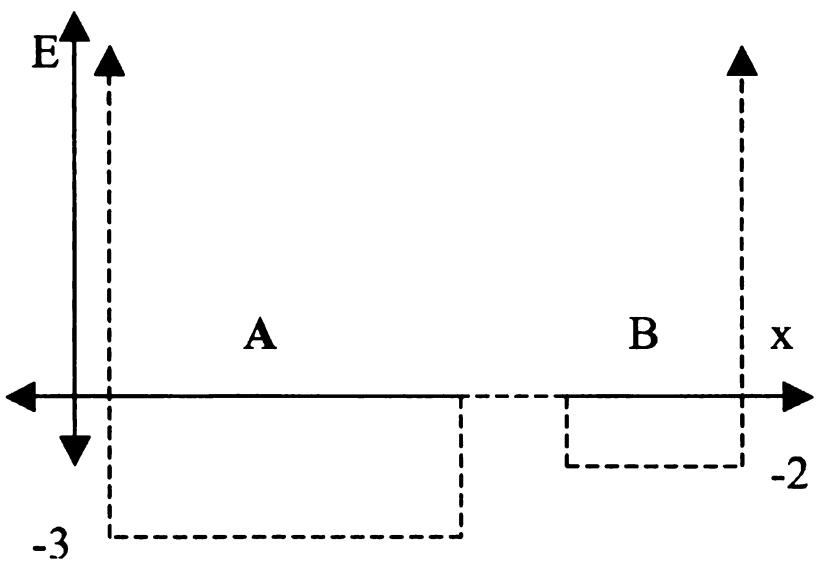


Figure 1. Schematic representation of the 2-D model.

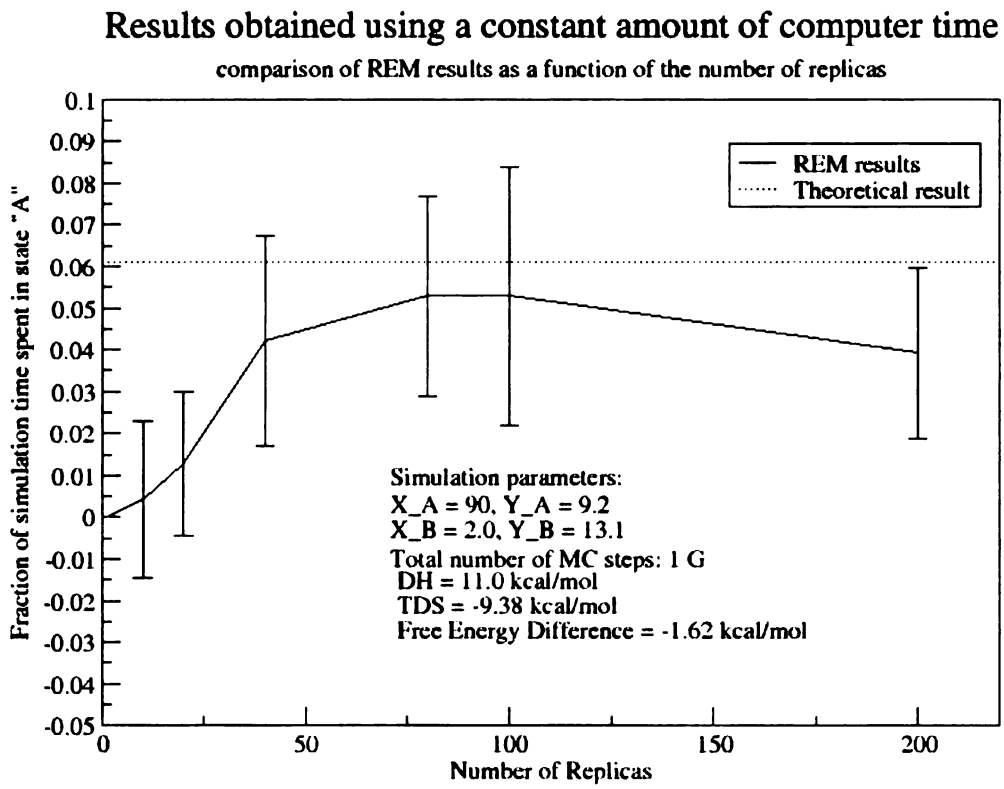


Figure 2. Accuracy of REM calculations deteriorates with increasing replica number.

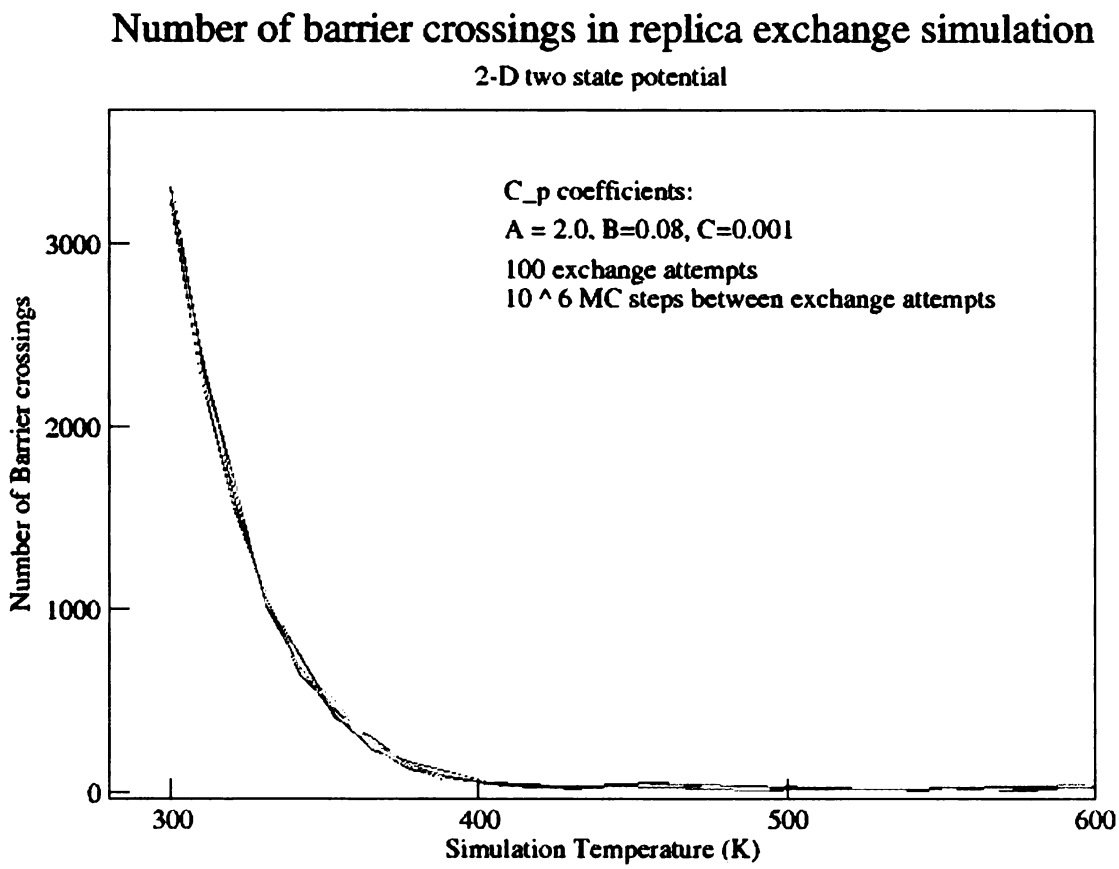


Figure 3. The number of barrier crossings dwindles at higher temperatures for the 2-D model.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular biology of the cell* (4th ed.). New York, NY: Garland Science.

Chan, H. S., & Dill, K. A. (1998). Protein folding in the landscape perspective: Chevron plots and non-arrhenius kinetics. *Proteins-Structure Function and Genetics, 30*(1), 2-33.

Dill, K., & Bromberg, S. (2002). *Molecular driving forces: Statistical thermodynamics in chemistry & biology* (1st ed.). New York, NY: Garland.

Frenkel, D., & Smit, B. (2001). *Understanding molecular simulation* (2 ed.). San Diego, CA: Elsevier.

Hill, T. (1986). *An introduction to statistical thermodynamics*. Mineola, NY: Dover Publications.

Kalos, M., & Whitlock, P. (1986). *Basics, volume 1, monte carlo methods* (Vol. 1): Wiley-Interscience.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087-1091.

Mitsutake, A., Sugita, Y., & Okamoto, Y. (2003a). Replica-exchange multicanonical and multicanonical replica-exchange monte carlo simulations of peptides. I. Formulation and benchmark test. *Journal of Chemical Physics, 118*(14), 6664-6675.

Mitsutake, A., Sugita, Y., & Okamoto, Y. (2003b). Replica-exchange multicanonical and multicanonical replica-exchange monte carlo simulations of peptides. Ii. Application to a more complex system. *Journal of Chemical Physics, 118*(14), 6676-6688.

Schlick, T. (2002). *Molecular modeling and simulation*. New York, NY: Springer.

Chapter 3: How Do Somatic Mutations Rigidify CDR Loops During Affinity Maturation?

Abstract

While some measurements suggest CDR flexibility is lost during affinity maturation, the physical basis for it remains elusive. Here, molecular dynamics simulations captured CDR flexibility differences between 4 mature antibodies (7G12, AZ28, 28B4, 48G7) and their germline predecessors. Analysis of their trajectories: 1) rationalized how mutations during affinity maturation restrict CDR motility, 2) captured the unliganded equilibrium between bound and unbound conformations for the H3 loop of 7G12g, and 3) predicted a set of new mutations that, according to our simulations, should affect binding.

Introduction

The immune system is able to produce highly specific and strongly binding antibodies for essentially any target molecule, or antigen. Germline antibodies, the set initially available at birth, weakly bind suspect molecules and evolve through cycles of mutation and selection, for the strongest binders, to yield "affinity-mature" antibodies. The mutations during affinity maturation are

HOOT LIBRARY
HOOT LIBRARY

focused on the 6 complementarity determining region (CDR) loops that make up the binding site(Alberts *et al.*, 2002).

Because antibodies can bind essentially any antigen, they are an excellent model system to study molecular recognition and binding. Understanding how and why somatic mutations modulate binding affinity may permit the design or disruption of protein-protein interactions, protein-ligand interactions and enable the design of antibodies de novo. These would be useful tools in combating disease as well as developing new applications in biotechnology i.e. biosensors, etc.

One hypothesis suggests germline antibodies are inherently flexible, easily rearrange to accommodate binding and that mutations during maturation restrict CDR loops to their bound conformation. Binding site pre-organization eliminates the free energy cost of rearrangement upon binding and contributes to affinity maturation. Experimental data that supports this position includes: 1) crystallographic and kinetic studies of mature immunoglobulins and their germline predecessors(Berger *et al.*, 1999; Foote, 2003; James & Tawfik, 2005)[8], 2) the entropic cost of binding is lower after maturation(Foote & Milstein, 1994; Manivel *et al.*, 2000), 3) polyreactivity(Notkins, 2004), the ability to bind distinctly different antigens, diminishes as antibodies mature. If this hypothesis is correct, how do mutations modulate flexibility?

Studying the factors that drive flexibility can be experimentally challenging. B-factors from crystallographic studies and NMR order parameters can provide indications of flexibility; however, deciphering the mechanism by which a given mutation modulates flexibility can be difficult. This motivates a theoretical approach to examine these effects to identify atomic-resolution interactions that limit flexibility.

This is an exploratory study where we report the identification of mutations that, in our molecular dynamics simulations, modulate CDR flexibility. Molecular dynamics and loop prediction calculations were performed on 4 mature antibodies and their germline precursors (7G12, AZ28, 28B4 and 48G7). In all four cases, the calculations capture greater flexibility, assessed by calculated B-factors, in germline H3 CDR loop than their mature counterparts. Via analysis of the simulations it was possible to: 1) identify single mutations that, even far from the paratope, can significantly restrict CDR mobility, 2) identify putative new mutations that could affect binding, 3) observe 7G12 visit its bound H3 conformation even in absence of antigen i.e. binds through a conformational selection mechanism, and 4) observe that 48G7g CDR flexibility loss does not drive its affinity maturation. While anecdotal, the finding that single mutations can independently modulate flexibility, and affect the binding affinity, could be particularly useful in the design and optimization of binding sites.

Methodology

Molecular dynamics

Molecular dynamics of 7G12, 28B4, AZ28 and 748G7 were performed using the Amber 8 simulation package(D.A Case, 2004). Simulations were carried out at 300 and 400 K. In order to avoid distortions at 400 K and make RMSD comparisons easier, Cα atoms not on CDR loops were restrained harmonically (force constant =0.3 kcal/mol/A^2) to their crystallographic coordinates. Solvent effects were incorporated by: 1) using langevin dynamics(Pastor *et al.*, 1988) and 2) a Generalized Born(Onufriev *et al.*, 2004) (igb=5) model with a surface area correction for non-polar effects. The collision frequency parameter was set equal to 2 to maximize barrier crossing(Loncharich *et al.*, 1992). The parm99 force field with Simmerling's backbone corrections(Simmerling *et al.*, 2002) were used. For computational efficiency, long range forces were only updated every 4 time-steps (nrespa=4). The inner timestep was 1fs. The production run spanned 15 ns of simulation time.

The PDB structure files, specific chain segments used, and the residues not restrained during the simulations are listed in Table 1. Note that no ligands were included in the simulations. The starting structures underwent 2000 steps of steepest descent minimization; an equilibration run followed. The temperature was gradually increased from 0 K to 300 K over 10,000 time-steps. The temperature was kept at 300 K for the rest of the 100 ps equilibration run. If the

system target temperature was 400 K, it was heated from 300 to 400 K over the 10 ps period following the initial heating (over picoseconds 10-20 of the equilibration run). During this time, the harmonic restraints mentioned above were enforced.

The trajectories were subsequently visualized using VMD(Humphrey *et al.*, 1996). The mutated residues were observed for differences in interactions between the germline and mature species.

Generation of local minima using Protein Local Optimization Program

In order to view differences in the energy landscape of possible H3 loop conformations in the 7G12 antibodies, sampled local minima were generated using the loop prediction algorithm in our Protein Local Optimization Program. We used a hierarchical approach to iteratively optimize loop conformations as described previously(Jacobson *et al.*, 2004) with increased sampling. In order to avoid decreased sampling due to nearby residues blocking alternative loop conformations, side chains with a heavy atom 7.5 Angstroms from the loop were removed during the backbone buildup stage of the loop prediction. This protocol was applied to an additional third initialization stage as well as the two refinement stages. In addition, to better match the loop definition used in the MD work in this paper, two extra residues (one on each side) were included in the prediction beyond the standard H3 loop definition. Residues B:97 B:103

were predicted in both the mature (PDB code 1NGY) and germline (PDB code 1NGZ) structures.

Hydrogen bonds analysis

The ptraj module in Amber was used to hydrogen bond partner distances. The distance cut-off heavy-atom to hydrogen distance cut-off was 2.5 A. In cases where equivalent atoms existed i.e a carboxylic acid, the lowest distance to either atom was taken.

Results and discussion

Molecular dynamics simulations of AZ28 (REFs), 7G12(Romesberg *et al.*, 1998; Yin *et al.*, 2003a; Yin *et al.*, 2003b), 28B4(HsiehWilson *et al.*, 1996; Yin *et al.*, 2001) and 48G7(Patten *et al.*, 1996; Priscilla & Schultz, 1999; Wedemayer & Stevens, 1997) and their corresponding germline predecessors were performed to identify mutations that modulate CDR flexibility and took place during affinity maturation. H3 is the most diverse CDR loop in sequence and length(Davis *et al.*, 1998), is believed to play a central role in specificity during maturation(Arden, 1998; Davis et al., 1998; Xu & Davis, 2000), and, its conformation is far more variable than the CDRs(Chothia *et al.*, 1989; Morea *et al.*, 1998). So, as a first order measure of flexibility we focused on CDR Cα B-factors of H3 and found larger deviations for the germline H3 loop than its

mature counterpart (figure 1). Visualization of the trajectories made it possible to identify contacts made in the mature species, that appear to restrict CDR mobility, but are absent in their germline predecessors.

7G12

The two key observations from simulations of 7G12 and its germline predecessor are: 1) the Ser76^{H}Asn mutation anchors the H1 loop and 2) the H3 loop visits its bound conformation in absence of ligand.

While, Yin et al (Yin et al., 2003a) ignore the Ser76^{H}Asn mutation in their analysis, it substantially hinders the motion of the H1 loop, in our calculations, as measured by calculated B-factors and inter-residue distances to H1. Asn76^{H} in the mature species simultaneously forms hydrogen bonds with the alcohol group of Thr28^{H} and the backbone carbonyl of Tyr27^{H} (figure 3), which does not happen in the germline case where a serine occupies that position. Germline H1 Cα B-factors (up to ~440) were roughly 3 times the magnitude of its mature counterpart (~130); figure 4 shows regular interval snapshots of the H1 loop which clearly show the level of restriction. In addition, the probability density of the distance between the residue centroids at positions 76^{H} and 28^{H} clearly depicts a peak around 6.5 A for the mature simulation which is absent in the germline result; the probability density in the 6-8 A interval is roughly twice for the germline trajectory than the mature (figure 5). If this model is correct: 1) a

significant difference in H1 flexibility should be observed between the 7G12 antibody and its germline predecessor and 2) a Thr28^{H}Ala mutation would increase flexibility the mature species and make no difference in the germline antibody.

While effect of the Ser76^{H}Asn mutation on binding is unknown, it may play a role since H1, in the germline simulation, swings to the binding pocket vicinity and interacts with nearby residues. This will have to be experimentally verified.

The simulation of the mature antibody are consistent with the interpretation, of Yin et al(Yin et al., 2003a), that the Ser97^{H}Met mutation anchors the H3 loop to its holo conformation, since the methionine contacts are firmly kept throughout the simulation. So, it appears this mutation in fact reduces the flexibility of H3, as predicted. The noteworthy and unexpected observation was the transition from the apo to holo conformation of the H3 loop in the germline simulation, in the absence of ligand. Figure 6 shows the Cα RMSD to the bound and unbound conformations. The corresponding simulation of the mature species fails to display this behavior. This observation is striking because it suggests 7G12g binds through a conformational selection mechanism and the Ser97^{H}Met shifts the equilibrium towards the conformation most auspicious for binding. Conformational selection in antibodies has been proposed(Berger et al., 1999; Bosshard, 2001; Foote & Milstein, 1994; James & Tawfik, 2003, 2005) and, in some cases, experimentally verified(Foote, 2003; James & Tawfik, 2005), but, to

our knowledge, it had not been observed in this type of simulation. Simulating these changes would be useful in predicting putative induced fit effects in antibodies.

The absence of flexibility in the mature H3 could potentially be due to frustration in the MD simulation. To eliminate this possibility, loop prediction calculations, which sample space without overcoming energy barriers, were performed on the H3 loop of 7G12. The results show a large diversity of conformations for the germline species in contrast with a small number of solutions found in the mature species (figure 7). Thus it appears the available H3 CDR conformations are more limited in the mature case.

28B4

The maturation of 28B4 involves the Asp95^{H}Trp mutation which, when reversed, lowers binding affinity by 3.7 kcal/mol(Yin et al., 2001) (more than half of the total binding free energy gained during maturation). Trp95^{H} is located at the base of the H3 loop and makes pi-pi interactions with the bound ligand, however, our simulations suggest it my also restrict H3. H3, in the germline simulation, drifts into the space occupied the Trp95H side-chain in the mature structure and onto the binding pocket (figure 7). The corresponding mature simulation does not display this behavior.

It is tempting to suggest the bulk of the Trp side-chain restricts the H3 loop mobility, but perhaps the mature antibody H3 loop could adopt that conformation given enough simulation time. In order to eliminate this possibility the loop conformation from the germline simulation was grafted onto the mature antibody crystal structure, the residue at H95 mutated manually in the pdb file, and a loop prediction calculations, where loop Cα atoms were kept within 2 A of these coordinates was carried out. The latter failed to find a set of loop coordinates that satisfied these restraints because the Trp sidechain could not be accommodated (even if nearby side-chains were removed). Therefore, it is unlikely the mature antibody H3 loop adopts this conformation or, at the very least, it clearly has fewer accessible states than its germline counterpart, which shifts the equilibrium toward the "bound" structure in the mature case.

The energy contribution to binding of this prearrangement effect versus the pi-pi stacking interaction of the Trp side-chain and ligand are difficult to deconvolute. But it is clear that this bulky non-polar group is able to very specifically shift the conformational equilibrium of H3.

AZ28

During AZ28 maturation, only one mutation occurs at a binding site residue (Ser34^{L}Asn)(Mundorff *et al.*, 2000), however, no structural changes are apparent from the crystallographic structures. In both cases the residue at position 34 L

UCSF LIBRARY

hydrogen bonds with the hydroxyl group of Tyr100^{H}a(Mundorff et al., 2000), which sits at the base of the H3 loop. Yet, it does cause a binding affinity difference of 0.9 kcal/mol(Mundorff et al., 2000; Ulrich *et al.*, 1997). One notable difference is that a crystallographic water molecule mediates this interaction in the germline antibody, which may indicate a weaker interaction in this case.

The differences in dynamics are far more evident. The hydrogen bond to Tyr100^{H}a is quickly lost in the simulation germline simulation (3% occupancy through the simulation), while, in the simulation of the mature species, Asn34^{L} interacts with Asp101^{H} (90% occupancy) and maintains its hydrogen bond with Tyr100^{H} (74% occupancy) (figure 8). The hydrogen bond between Asp101^{H} and Tyr100^{H} is absent in the germline simulation (0.4% occupancy), while, in the mature simulation, it creates an electrostatically auspicious environment for the tyrosine hydroxyl group where it can make two hydrogen bonds simultaneously.

This interaction is important because Tyr100^{H}a sits at the base of the H3 loop and it is the mutation most likely to affect the H3 conformation(Ulrich et al., 1997). In the germline simulation the H3 loop folds over onto the space that would be occupied by the ligand upon binding as Tyr100^{H}a makes contacts with other residues which permits this shift (figure 8). Thus, this mutation likely contributes significantly to the flexibility differences in figure 1. It also suggests

a mutation of Asp101H to alanine would have an effect on the binding affinity of the mature antibody and have no effect on binding for its germline predecessor.

748G7

48G7 is an outlier in this analysis. It appears to mature differently than the other cases because: 1) it undergoes 9 mutations (Ser30^{L}Asn, Ser34^{L}Gly, Asp55^{L}His, Glu42^{H}Lys, Gly55^{H}Val, Asn56^{H}Asp, Gly65^{H}asp, Asn76^{H}Lys and Ala78^{H}Thr)(Wedemayer & Stevens, 1997) rather than 6 or less in the other cases and 2) the ligand bound conformation differs between the mature and germline species. Also interesting is that none of the mutations directly interact with the ligand, most of them only weakly affect the binding affinity(Patten et al., 1996) and there is cooperativity between pairs of mutations(Priscilla & Schultz, 1999), that is, combinations of mutations have a larger effect on binding than the sum of their individual contributions. Priscilla et al(Patten et al., 1996; Priscilla & Schultz, 1999) suggest that initial mutations may change the ligand binding-pose and subsequently introduce new contacts. Thus, mutations that optimize those new contacts are only relevant if the previous mutations are present. Because of the latter observation, the analysis of the simulations performed here focused on those mutations that, alone, most affected the binding affinity.

H3 Cα B-factors from the mature simulation are only slightly lower values than the corresponding germline data (Figure 1). Except for H2, the other CDR loops,

L1, L2, L3 and H1, show the reverse trend; the simulation trajectory of the mature species yields higher B-factors than its germline counterpart (see supplementary data). Most of the mutations induced no new contacts; many of their side-chains simply diffused through solvent during the simulation. The higher rigidity of the mature H3 appears to be due to the Asp55^{L}His mutation which causes the loss of a salt bridge between 55^{H} and Arg46^{L}; the latter is then free to interact with the backbone of H3. The Asn76HLys mutation actually increases the flexibility of H1. This flexibility, as discussed by Wedemeyer et al(Wedemayer & Stevens, 1997) , permits the rearrangements necessary to optimize contacts with the ligand. Interestingly, this is consistent with the effect of Ser76^{H}Asn in the maturation of 7G12 and Ser34^{L}Asn in AZ28. That is, asparagine appears to be able to restrict motion by making two electrostatic interactions simultaneously. In the case of 48G7, its removal causes precisely the expected effect and increases flexibility.

A previous molecular dynamics study on the germline and mature species of this antibody concluded the germline complex is more flexible than its mature counterpart(Chong *et al.*, 1999). In that study the simulation included the ligand and was much shorter (less than a nanosecond). At shorter simulation times (< 2 ns), we find a larger gap between the germline and mature H3 flexibility, so it is unclear if their conclusion would hold at longer simulation times and in the absence of ligand.

Conclusion

While the set of cases is too small to make any generalizations, it was possible to rationalize how mutations, sometimes not at the active site (Ser76^{H}Asn in 7G12), restrict CDR mobility. An unexpected and interesting result was the equilibrium between bound and unbound H3 conformations in the absence of ligand (7G12 germline), which is consistent with a conformational selection mechanism for binding. Together these observations are consistent with structurally diverse and polyreactive gernline antibodies that become specific during maturation by restricting their CDR conformations to the most auspicious binding geometry. The fewer unbound accessible states lower the entropic cost of binding and improves the overall binding affinity.

Mutations to asparagine appear to have a restricting effect, especially since reverse mutations caused an increase in flexibility (48G7). Restrictions due to bulky side chains (Asp95^{H}Trp, 28B4) can also play a role.

Here we focused on rigidification of CDR loops during maturation, however, other mechanisms for affinity maturations exist; 48G7 appears to mature via an alternative route and experiments identified others. Sethi et al(Sethi *et al.*, 2006) found a germline antibody that 1) bound several peptides, but in different conformations and 2) its maturation involved disruption of binding contacts except for those involving the target antigen.

Antibody	Germline Structure PDB code	Mature Structure PDB code	CDR region residues
7G12	1NGZ (L1-110, H1-110)	1NGY (L1-110, H1-110)	L1: L23-34 L2: L50-56 L3: L88-97 H1: H26-33 H2: H50-58 H3: H99-103
28B4	1FL5 (L1-112, H1-113)	1KEM (L1-112, H1-117)	L1: L24-40 L2: L54-61 L3: L94-102 H1: H26-35 H2: H52-56 H3: H101-108
AZ28	1D5I (L1-110, H1-110)	1D5B (L1-110, H1-110)	L1: L24-34 L2: L50-56 L3: L89-97 H1: H26-35 H2: H52-58 H3: H95-102
48G7	2RCS (L1-110, H1-110)	1HKL (L1-110, H1-110)	L1: L24-33 L2: L48-55 L3: L89-97 H1: H26-35 H2: H51-56 H3: 95-103

Table 1. Structural information for simulations setup. PDB structure codes and specific chain segments used in the simulation for germline and mature antibodies are listed. In addition, the assignment of CDR residues is explicitly shown.

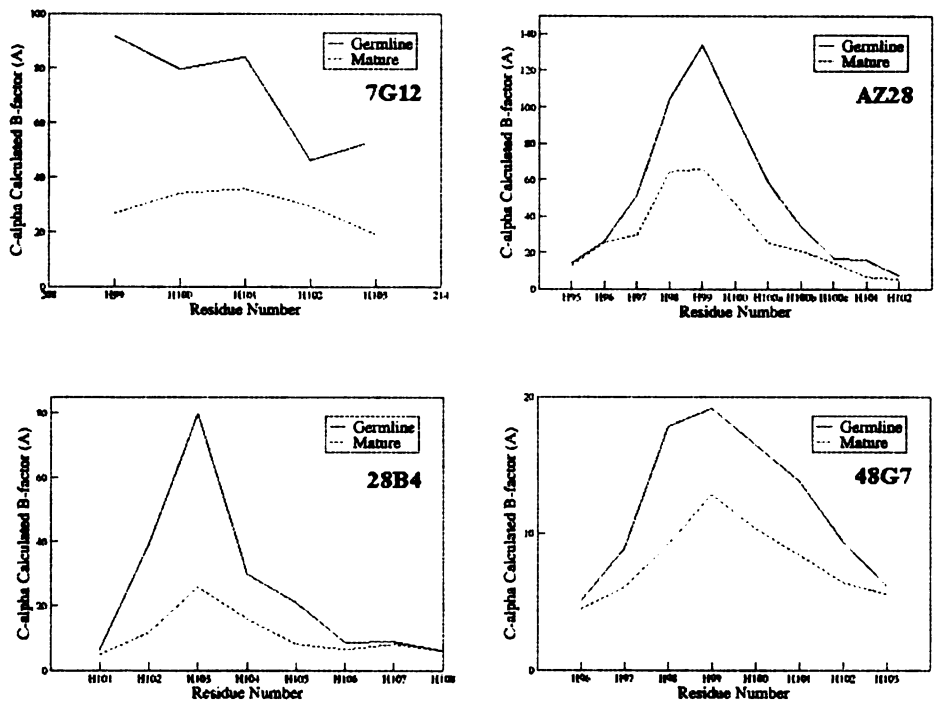


Figure 1. H3 loop Ca B-factors calculated from molecular dynamics trajectories. The germline species consistently yields higher, sometimes by a large margin, B-factors than the mature simulation.

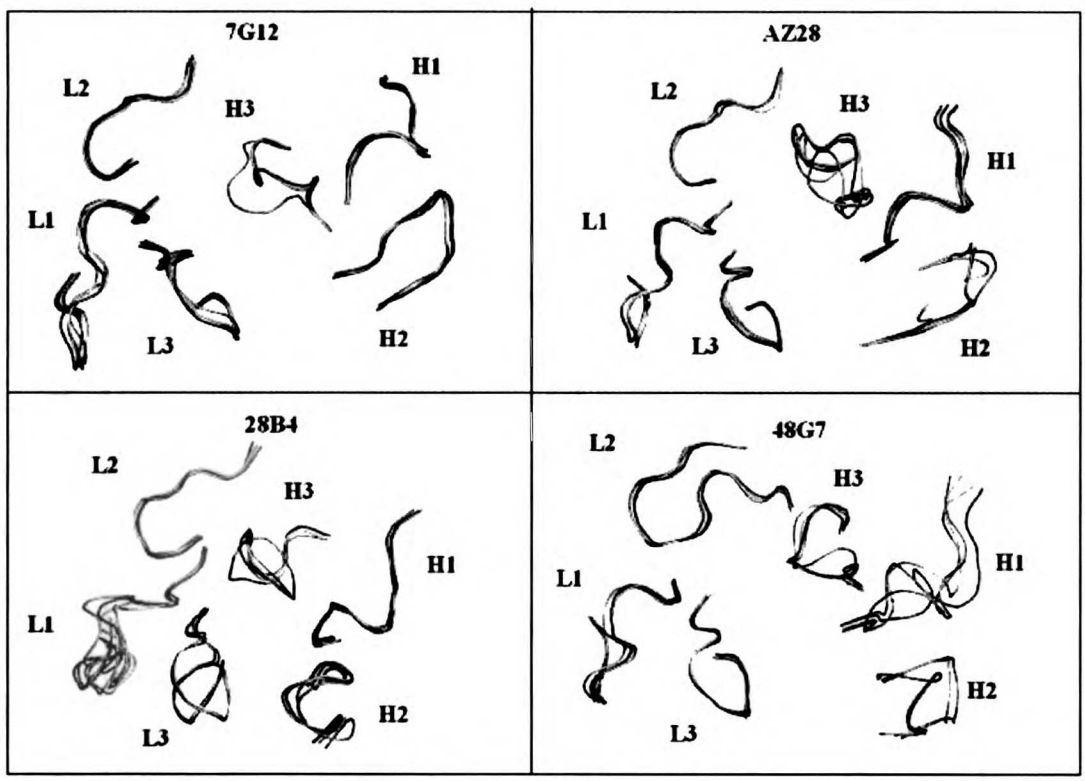


Figure 2. CDR loops for bound and unbound germline and mature antibodies. The bound germline and mature structures are colored yellow and magenta while the free germline and mature structures are colored green and cyan, respectively. It is most clear in the H3 loops that the unbound germline structure (green) is significantly different than the others.

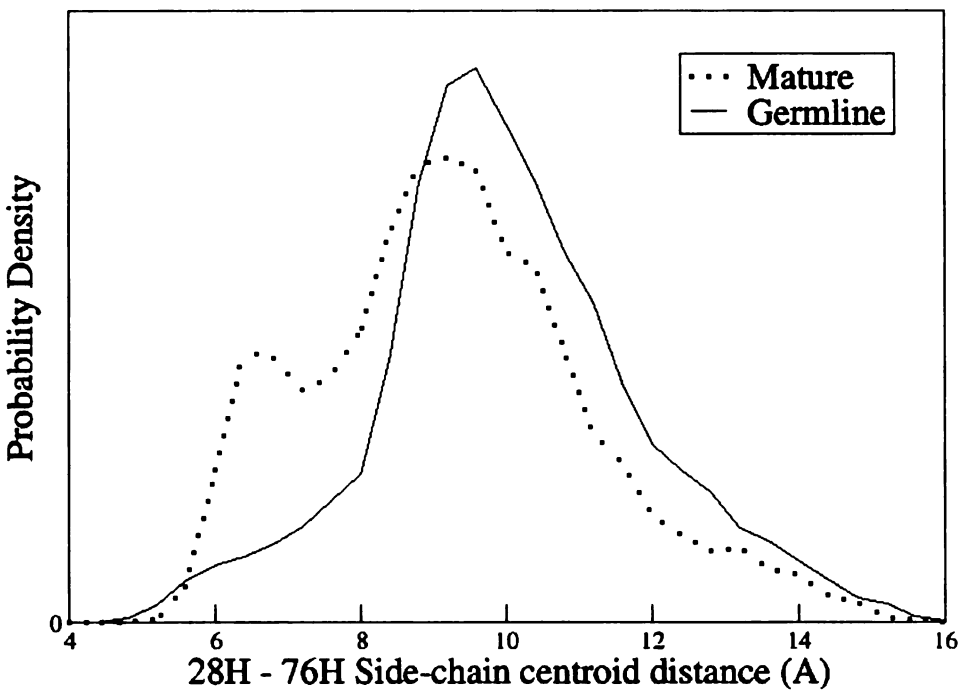


Figure 3. Probability density function of distances between residue centroid at positions 28H and 76H of the germline and mature 7G12 antibody.

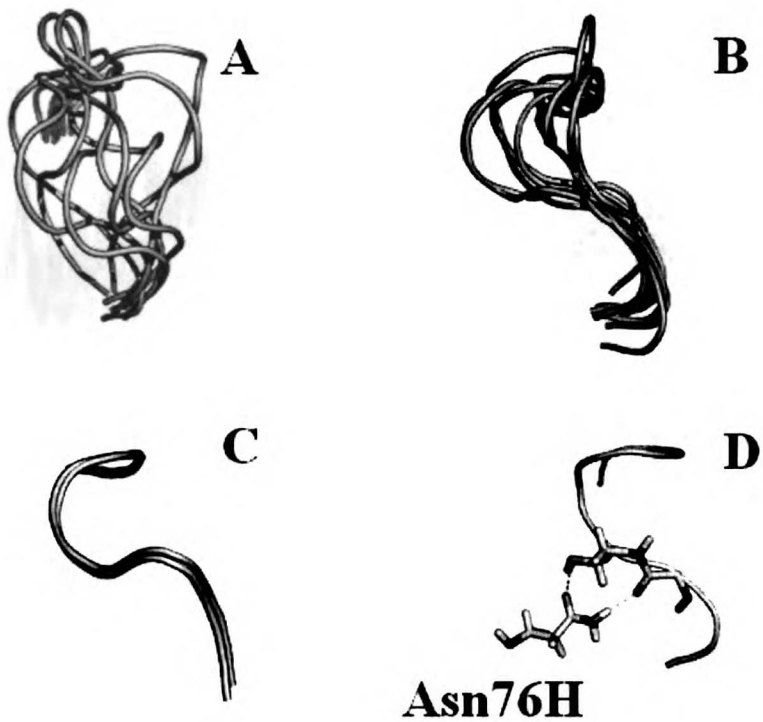


Figure 4. 7G12 H1 loop flexibility is diminished by Ser76^{H}Asn mutation in ways not obvious from crystallographic structures. A. H1 loop snapshots from 7G12 germline predecessor antibody simulation. B. H1 loop snapshots from 7G12 mature antibody simulation. C. H1 crystal structures of the 7G12 mature antibody and its germline predecessor. The crystallographic structures are nearly identical. D. Asn76H hydrogen bonds H1 and restricts its motion in the mature species.

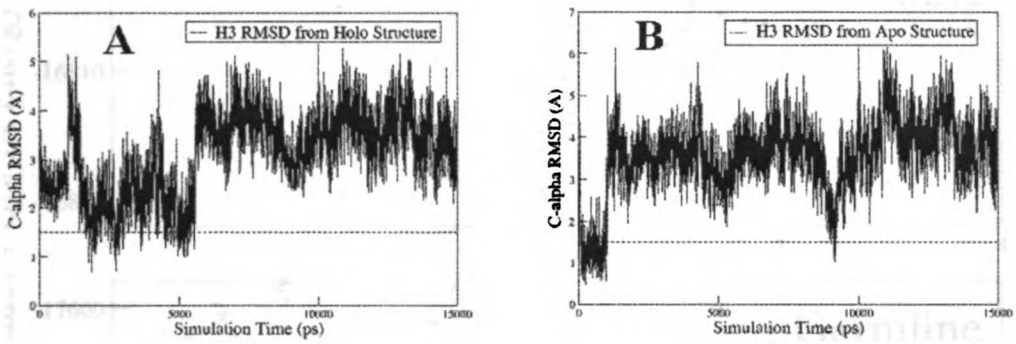


Figure 5. 7G12 germline H3 loop visits bound and unbound conformations during simulation . A. H3 loop RMSD from the germline holo structure as a function of simulation time. B. H3 loop RMSD from the germline apo structure as a function of simulation time. The dotted line marks a 1.5 RMSD.

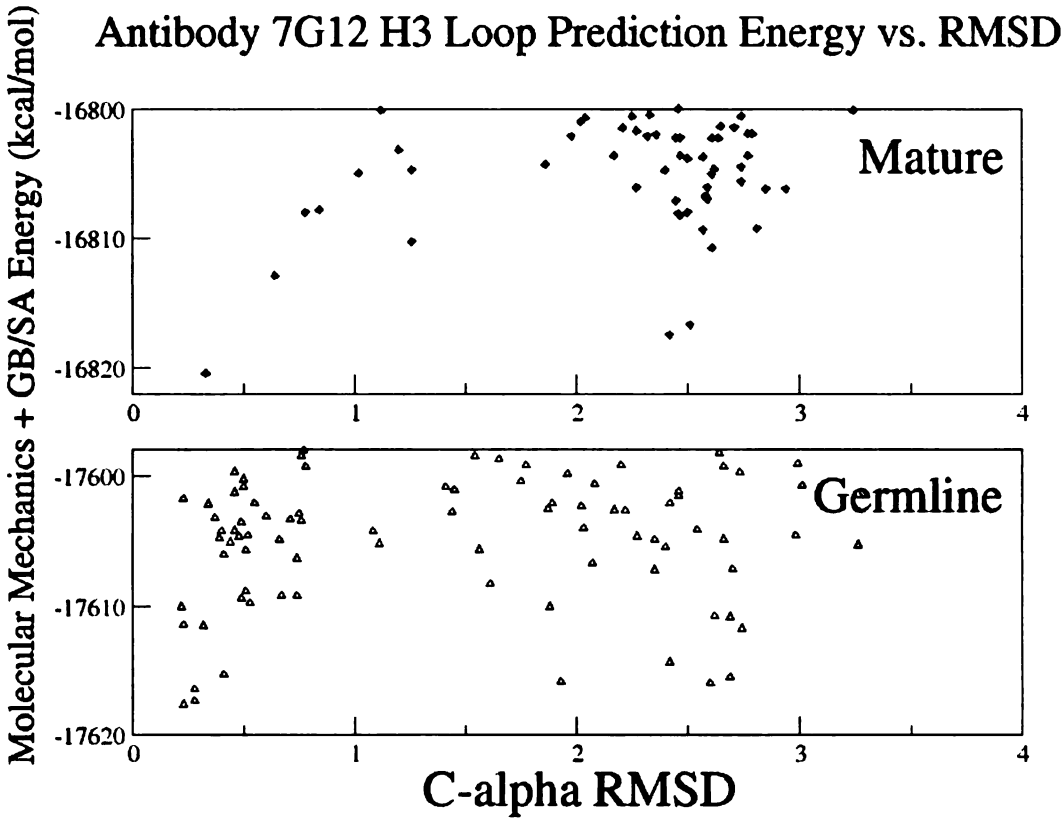


Figure 6. Plots of energy (kcal/mol) versus backbone RMSD for local minima identified during optimization of the H3 loop in both the apo mature 7G12 antibody (top) and the apo germline 7G12 antibody (bottom.) All samples from each prediction stage (see Methods) were combined in these plots. RMSD is measured against all heavy atoms in the loop backbone between the predicted and native structures with all atoms in the unperturbed chains aligned (excluding the loop.)

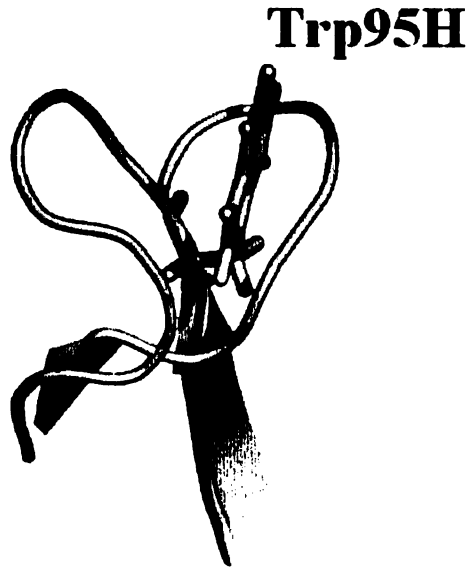


Figure 7. Simulation of the H3 loop in the germline antibody (green) explores conformational spaces occupied by a bulky Trp95H sidechain in the mature species (green). The latter is the result of a somatic mutation during maturation. The bulky Trp sidechain restricts the available conformational space of H3.

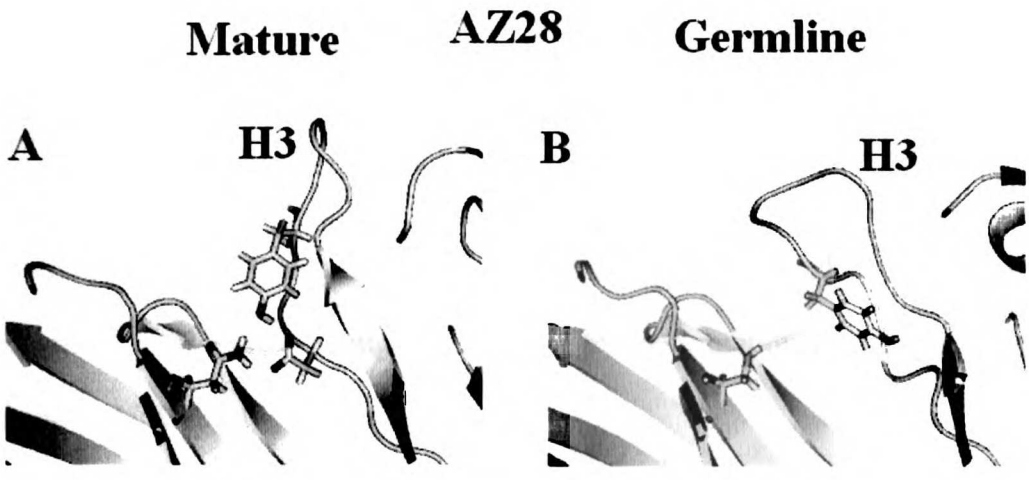


Figure 8. Simulation dynamics differences between AZ28 and its germline predecessor. A. Simulation snapshot showing the hydrogen bond contacts of Asn34L of the mature AZ28 antibody. Asn34L makes contacts with Asp101H and His100aH. B. Simulation snapshot showing the lack of a hydrogen bond between Ser34L and His100aH in the germline species simulation. These snapshots are representative of the entire simulation.

References:

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular biology of the cell* (4th ed.). New York, NY: Garland Science.

Arden, B. (1998). Conserved motifs in t-cell receptor cdr1 and cdr2: Implications for ligand and cd8 co-receptor binding. *Current Opinion in Immunology, 10*, 74-81.

Berger, C., Weber-Bornhauser, S., Eggenberger, J., Hanes, J., Pluckthun, A., & Bosshard, H. R. (1999). Antigen recognition by conformational selection. *Febs Letters, 450*(1-2), 149-153.

Bosshard, H. R. (2001). Molecular recognition by induced fit: How fit is the concept? *News in Physiological Sciences, 16*, 171-173.

Chong, L., Duan, Y., Wang, L., Massova, I., & Kollman, P. (1999). Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48g7. *Proceedings of the National Academy of Sciences of the United States of America, 96*(25), 14330-14335.

Chothia, C., Lesk, a. M., Tramontano, a., Levitt, M., Smithgill, S. J., Air, G., et al. (1989). Conformations of immunoglobulin hypervariable regions. *Nature, 342*(6252), 877-883.

D.A Case, T. A. D., T.E. Cheatham, III, L.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, B. Wang, D.A. Pearlman, M. Crowley, S Brozell, V. Tsui, H. Gohlke, J.Hornak, G. Cui, P. Beroza, C. Schafmeister, J.W. Caldwell, W.S. Ross, and P.A. Kollman. (2004). Amber (Version 8).

Davis, M. M., Boniface, J. J., Reich, Z., Lyons, D., Hampl, J., Arden, B., et al. (1998). Ligand recognition by alpha beta t cell receptors. *Annual Review of Immunology, 16*, 523-+.

Foote, J. (2003). Isomeric antibodies. *Science, 299*(5611), 1327-1328.

Foote, J., & Milstein, C. (1994). Conformational isomerism and the diversity of antibodies. *Proceedings of the National Academy of Sciences of the United States of America, 91*(22), 10370-10374.

HsiehWilson, L. C., Schultz, P. G., & Stevens, R. C. (1996). Insights into antibody catalysis: Structure of an oxygenation catalyst at 1.9-angstrom resolution. *Proceedings of the National Academy of Sciences of the United States of America, 93*(11), 5363-5367.

Humphrey, W., Dalke, A., & Schulten, K. (1996). Vmd: Visual molecular dynamics. *Journal of Molecular Graphics, 14*(1), 33-&.

Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J. F., Honig, B., Shaw, D. E., et al. (2004). A hierarchical approach to all-atom protein loop prediction. *Proteins-Structure Function and Bioinformatics, 55*(2), 351-367.

James, L. C., & Tawfik, D. S. (2003). Conformational diversity and protein evolution - a 60-year-old hypothesis revisited. *Trends in Biochemical Sciences, 28*(7), 361-368.

James, L. C., & Tawfik, D. S. (2005). Structure and kinetics of a transient antibody binding intermediate reveal a kinetic discrimination mechanism in antigen recognition. *Proceedings of the National Academy of Sciences of the United States of America, 102*(36), 12730-12735.

Loncharich, R. J., Brooks, B. R., & Pastor, R. W. (1992). Langevin dynamics of peptides - the frictional dependence of isomerization rates of n-acetylalanyl-n'-methylamide. *Biopolymers, 32*(5), 523-535.

Manivel, V., Sahoo, N. C., Salunke, D. M., & Rao, K. V. S. (2000). Maturation of an antibody response is governed by modulations in flexibility of the antigen-combining site. *Immunity, 13*(5), 611-620.

Morea, V., Tramontano, A., Rustici, M., Chothia, C., & Lesk, A. M. (1998). Conformations of the third hypervariable region in the vh domain of immunoglobulins. *Journal of Molecular Biology, 275*(2), 269-294.

Mundorff, E. C., Hanson, M. A., Varvak, A., Ulrich, E., Schultz, P. G., & Steens, R. C. (2000). Conformational effects in biological catalysis: An antibody-catalyzed oxy-cope rearrangment. *Biochemistry, 39*(4), 627-632.

Notkins, a. L. (2004). Polyreactivity of antibody molecules. *Trends in Immunology, 25*(4), 174-179.

Onufriev, A., Bashford, D., & Case, D. A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins-Structure Function And Bioinformatics, 55*(2), 383-394.

Pastor, R. W., Brooks, B. R., & Szabo, a. (1988). An analysis of the accuracy of langevin and molecular-dynamics algorithms. *Molecular Physics, 65*(6), 1409-1419.

Patten, P. A., Gray, N. S., Yang, P. L., Marks, C. B., Wedemayer, G. J., Boniface, J. J., et al. (1996). The immunological evolution of catalysis. *Science, 271*(5252), 1086-1091.

Priscilla, Y., & Schultz, P. G. (1999). Mutational analysis of the affinity maturation of antibody 48g7. *Journal of Molecular Biology, 294*, 1191-1201.

Romesberg, F. E., Santarsiero, B. D., Spiller, B., Yin, J., Barnes, D., Schultz, P. G., et al. (1998). Structural and kinetic evidence for strain in biological catalysis. *Biochemistry, 37*(41), 14404-14409.

Sethi, D. K., Agarwal, A., Manivel, V., Rao, K. V. S., & Salunke, D. M. (2006). Differential epitope positioning within the germline antibody paratope enhances promiscuity in the primary immune response. *Immunity, 24*(4), 429-438.

Simmerling, C., Strockbine, B., & Roitberg, A. E. (2002). All-atom structure prediction and folding simulations of a stable protein. *Journal Of The American Chemical Society, 124*(38), 11258-11259.

Ulrich, H. D., Mundroff, E., Santarsiero, B. D., Driggers, E. M., Stevens, R. C., & Schultz, P. G. (1997). The interplay between binding energy and catalysis in the evolution of a catalytic antibody. *Nature, 389*(6648), 271-275.

Wedemayer, G. J., & Stevens, R. C. (1997). Structural insights into the evolution of an antibody combining site (vol 276, pg 1665, 1997). *Science, 277*(5331), 1423-1423.

Xu, J., & Davis, M. (2000). Diversity in the cdr3 region of vh is sufficient for most antibody specificities. *Immunity, 13*, 37-45.

Yin, J., Andryski, S. E., Beuscher, A. E., Stevens, R. C., & Schultz, P. G. (2003a). Structural evidence for substrate strain in antibody catalysis. *Proceedings of the National Academy of Sciences of the United States of America, 100*(3), 856-861.

Yin, J., Beuscher, A. E., Andryski, S. E., Stevens, R. C., & Schultz, P. G. (2003b). Structural plasticity and the evolution of antibody affinity and specificity. *Journal of Molecular Biology, 330*(4), 651-656.

Yin, J., Mundorff, E. C., Yang, P. L., Wendt, K. U., Hanway, D., Stevens, R. C., et al. (2001). A comparative analysis of the immunological evolution of antibody 28b4. *Biochemistry, 40*(36), 10764-10773.

Chapter 4: Competition Between Intramolecular Hydrogen Bonds and Solvation in Phosphorylated Peptides: Simulations with Explicit and Implicit Solvent

Abstract

In this work we simulate a capped phosphorylated peptide, Ace-Gly-Ser-pSer-Ser-Nme, using a variety of solvent models, and compare the results to each other and to experimental observables from an NMR experiment. The overall goal is to compare the performance of water models (explicit and implicit) in treating the phosphate moiety. Three explicit water models (TIP3P, TIP4P and SPC/E), the Generalized Born-surface area (GB/SA) treatment, and a distance dependent dielectric (DDD) (with dielectric constants of r and 4r) were employed in molecular dynamics simulations. The phosphate group can have a net charge of -2 (unprotonated) or -1 (singly protonated) near physiological pH, and we investigated both protonation states. Atomic charges for the phosphorylated residue were calculated using several methods including high-level quantum mechanics with approximate treatment of correlation and solvent, and several charge sets were tested in this analysis. The key conclusion of this study is that the explicit solvent models and GB/SA implicit solvent model both give good agreement with experiment given appropriate partial charges, while the DDD simulations do not. The dynamics are dominated by transitions between a state in which the phosphate group forms no intramolecular hydrogen bonds, and one in which it forms multiple hydrogen bonds with backbone NH groups N-terminal to the phosphate group. Thus, the strength of the P-O/H-N hydrogen bond, which depends on both the partial charges and solvent model used, is a key parameter controlling the accuracy of the simulations. The singly

and doubly charged phosphopeptide simulations show similar ensemble averages, with the overall hydrogen bonding propensity being slightly higher for the -1 species. The similarity of the strength of hydrogen bonds, despite the change in net charge, is due to near cancellation between changes in the favorable Coulomb interactions and the free energy cost of desolvation incurred upon forming a hydrogen bond.

Introduction

Solvation of a phosphorylated species can potentially pose a significant challenge for current water/solvation models. Since simulation of phospho-proteins is a relatively unexplored area (compared to biomolecular simulation in general), the adequacy of current solvation models, for this purpose, is not yet well-established. In this work we compare the ability of some explicit (TIP3P, TIP4P and SPC/E) and implicit (Generalized Born-Surface Area and a distance dependent dielectric) models to solvate the phosphate moiety. In addition, we go beyond simply gauging the quality of results and seek insight into the reasons behind poor performance. The latter effort led us to consider hydrogen bonds between the phosphate moiety and backbone amides. As explained later, we choose to study a relatively small system because 1) NMR measurements were made on this system and 2) it is computationally tractable and 3) it isolates the local effects of phosphorylation.

Post-translational phosphorylation serves as a control mechanism in a myriad of cellular processes including metabolic pathway regulation(Audette *et al.*, 2000), extracellular

signal transduction(Chen *et al.*, 2001), ion channel regulation(Martens *et al.*, 1999; Patel & Honore, 2001) and cell cycle progression(Vermeulen *et al.*, 2003)(Feng *et al.*, 1996; Johnson & Lewis, 2001). The hydroxyl groups of tyrosine, serine and threonine are targeted for phosphorylation in eukaryotic cells while histidine and aspartic acid are the most frequent targets in prokaryotes(Johnson & Lewis, 2001; Johnson & O'Reilly, 1996). The introduction of the phosphate ion, which predominantly carries a negative two charge at neutral pH, perturbs the electrostatic potential and quite often the conformation of the modified protein(Johnson & Lewis, 2001). Even in absence of rearrangement, the change in electric field and the steric hindrance from a phosphate group can have biologically significant consequences, e.g., promoting or opposing protein-protein interactions(Audette et al., 2000).

Theoretical studies can in principle be used to 1) gain insight into the physical phenomena that drive the effects of phosphorylation, and 2) predict the effects of phosphorylation on proteins where the phosphorylated structure remains unsolved. To date, there have been relatively few studies of this type, and force field parameters for phosphorylated residues are not well established. Prior studies using phosphorylated residues have obtained partial charges and other parameters from nucleotide parameter sets(Smart & McCammon, 1999; Stultz *et al.*, 2002), quantum mechanical calculations(Feng et al., 1996) or alternative methods(Shen *et al.*, 2001).

Performing molecular mechanics simulations of phosphorylated species requires careful consideration of solvation effects. The standard 20 amino acids have net charges

between +1 and –1. The effect of the solvent on the behavior of the -2 phosphate group is likely to be different than for the standard amino acid side chains, and is likely to pose a stringent test for both explicit and implicit solvent models. The role of the solvent in stabilizing or destabilizing polar interactions, especially hydrogen bonding and salt bridges, will be critical in establishing the correct balance between potential interactions. There is no guarantee that the appropriate balance will be achieved with currently employed water models, either explicit or implicit.

Some recent simulations of nucleic acids provide reason for optimism about the ability of widely used solvent models to treat highly charged systems. Molecular dynamics simulations of the dodecamer d(CGCGAATTCGCG) have explored several issues in nucleic acid dynamics(Beveridge & McConnell, 2000; Norberg & Nilsson, 2002). Success in reproducing the 2D NMR specturum(Arthanari *et al.*, 2003) of this oligomer suggests that fixed charge water models are able to adequately reproduce the behavior of the solvent. Hydration around nucleic acids is more ordered than around proteins(Makarov *et al.*, 2002) and, hence, can potentially pose a challenge for implicit solvent treatments. However, using the Generalized Born model, Tsui *et al*(Tsui & Case, 2000a) were able to mimic the results obtained in explicit water simulations. With a slight modification of atomic radii of hydrogens bonded to oxygen or nitrogen, their implicit solvent results compares well with an explicit solvent simulation of d(CCAACGTTGG).

Previously, Brownian dynamics simulations were performed on the phosphorylated peptide, Ace-Gly-Ser-pSer-Ser-Nme, but only a distance dependent dielectric (DDD) was used to incorporate solvent effects(Shen et al., 2001). Both the phosphorylated and unphosphorylated peptides were simulated and $^3J_{NH-CH}$ constants for all residues in both molecules were calculated. The calculated coupling constants agreed with experiment, within 0.1 Hz, for the unphosphorylated peptide simulations, but the agreement was poorer for the phosphorylated peptide. The use of a distance dependent dielectric was cited as a possible reason for disagreement with experiment.

In this work, we follow the same approach to compare several explicit (TIP3P(Jorgensen *et al.*, 1983), SPC/E(Berendsen *et al.*, 1987) and TIP4P(Jorgensen et al., 1983)) and implicit water models (Generalized Born(Dominy & Brooks, 1999; Tsui & Case, 2000b) and DDD). The simulations focus mainly on the unprotonated state (-2) but the protonated species (-1) is also considered. As in the previous study, we employ $^3J_{NH-CH}$ coupling constants for validation. Simulation results of Ace-Gly-Ser-pSer-Ser-Nme, using the models mentioned are compared to each other and to experimental data(Tholey *et al.*, 1999). While NMR measurements exist for phosphotyrosine and phosphothreonine in the same short peptide model, we focused on the phosphoserine peptide because ensemble averages converge more rapidly. The $^3J_{NH-CH}$ ensemble average as a function of simulation time was monitored to make this determination.

The DDD treatment yielded poor results before, yet we test it here because 1) comparison with the previous work(Shen et al., 2001) is desirable, 2) it is commonly used in docking

HOPE LIBRARY

studies (particularly relevant to docking into active sites with a phosphate moiety or nucleic acids) and 3) we go beyond noting a poor performance and gain insight into the failings of the model. Because force field partial charges for phosphorylated residues are not well established and can be expected to strongly impact the accuracy of the simulations, we also invest considerable effort in determining appropriate phosphate moiety partial charges by a combination of quantum chemistry and empirical testing. Note that we expect the phosphate group partial charges to drive the phosphorylation effects; this is discussed further in the results section. The key conclusion is that, given appropriate partial charges, the explicit solvent models and, perhaps more surprisingly, GB/SA implicit solvent model both produce good agreement with experiment. The DDD simulations do not give good agreement, over the range of partial charges we tested.

Finally, observations made during the analysis of the MD trajectories motivated an analysis of the phosphate moiety hydrogen bonding propensities. We sought to understand the balance of hydrogen bonding of the phosphate group to backbone amide and hydroxyl hydrogens. Backbone amides and hydroxyl hydrogens are two of the three most common hydrogen bond partners of phosphorylated amino acids in the protein data bank (PDB)(Berman *et al.*, 2000). Both of these groups are present in the phosphopeptide under study. In our simulations, the dynamics are dominated by two states: one in which the phosphate forms no intramolecular hydrogen bonds, and one in which it forms multiple hydrogen bonds with backbone NH groups N-terminal to the phosphate group. The strength of the P-O/H-N hydrogen bond, which depends on both the partial charges and solvent model, is the critical determinant of agreement with

experiment. The singly and doubly charged phosphopeptide simulations show similar ensemble averages, with the overall hydrogen bonding propensity being slightly higher for the -1 species.

Methods

Molecular Dynamics Simulations

All molecular dynamics simulations in this study were performed using the Amber 7 package(D.A. Case, 2002) with the parm99 parameter file. The peptides were capped with an acetyl group (Ace) at the N-terminus and N-methyl amine (Nme) at the C-terminus, i.e., the sequence was Ace-Gly-Ser-pSer-Ser-Nme. Atom types for the phosphate moiety are already present in the parameter file (for the nucleic acids), therefore no new atom types were needed. The ff99 force field was the best all-around biomolecular force field in the Amber 7 simulation package, however, it overstabilizes helical conformations(Okur *et al.*, 2003). This may affect the coupling constant calculations by imposing a 1-4 contact, however, this was not a prevalent conformation in the simulation trajectories. In all simulations, a weak-coupling algorithm(Berendsen *et al.*, 1984) kept the temperature and/or pressure constant. The temperature coupling constant, tcoup, for all production phase runs was 2 ps. During equilibration of the implicit solvent simulations, the temperature coupling constant was 1 ps. It was temperature coupling constant 2 ps during equilibration runs that included explicit

NCSF LIBRARY

solvent. For all constant pressure simulations, the pressure coupling constant was 2 ps as well. Trajectory snapshots were saved every picosecond.

Explicit solvent simulations were prepared by immersing the solute in a rectangular water box with a 10 Å clearance from the box walls in every axis. In the case of TIP3P, a pre-equilibrated water box (WATBOX216) was employed as solvent. For TIP4P and SPC/E, the solvent molecules were placed in the box using the "solvatebox" command in tleap (Amber module for system preparation). The initial box volume was 49968 A^3 (37.40 A x 42.62 A x 31.35 A) for all explicit solvent simulations. Sodium counterions were used to maintain electroneutrality. Periodic boundary conditions were imposed, and electrostatics were treated using the particle mesh Ewald method(Darden *et al.*, 1993; Essmann *et al.*, 1995; Toukmaji *et al.*, 2000). Default parameters, in Amber 7, for the particle mesh Ewald method were used. The charge grid dimensions were 36, 40 and 27 A and the nonbonded interation cut-off in real space was 8 A. A fourth order B-spline interpolation was employed. Before equilibration, 2000 steps of steepest descent minimization helped relax the system. The equilibration protocol was as follows: the system was heated to 300 K in 25 ps; a constant pressure and temperature equilibration phase of 100 ps followed; subsequently, the system underwent a further NVT equilibration period of 50 ps. Following these equilibration phases, a production run of at least 15 ns was performed. The production run was performed at constant temperature and volume; the integration step was 1 fs and bonds involving hydrogen were constrained using the SHAKE algorithm(J.-P. Ryckart, 1977).

Implicit solvent simulations were performed using the Generalized Born-Surface Area (GB/SA) model using the radii by Tsui et al(Tsui & Case, 2000b). GB/SA simulations were equilibrated for 100 ps. The system was heated to 300K over 5000, 1 fs, time-steps and maintained at that temperature for the remainder of the equilibration. A production run of 40 ns followed. The integration time step was 1 fs. The simulations were run at a constant temperature of 300K.

The simulations using distance-dependent dielectric (DDD) were performed using a replica exchange method(Mitsutake *et al.*, 2003a, 2003b; Sugita & Okamoto, 1999). Due to the poor shielding of electrostatics in this type of simulation and the large charge on the phosphate ion, strong salt bridges form, and thermal energy (at 300K) is insufficient to break them on the timescale of the simulation. As a result, the trajectories can become trapped in local minima at room temperature. The replica exchange method was used to sample more efficiently the conformational space of the peptide. In this method, N replicas of the system at N temperatures are run simultaneously. Periodically, every 50 ps in this study, an attempt to exchange the Cartesian coordinates among the replicas occurs, based on the Metropolis criterion(Sugita & Okamoto, 1999). If an exchange attempt is successful, the momenta are rescaled such that:

(3) $$p^{new} = \sqrt{\frac{T_{new}}{T_{old}}} p^{old}$$

The temperatures employed in this study were 300, 321, 348, 380, 418, 465, 524, 600 and 680 K. These were chosen to maintain a ~50% exchange probability between replicas.

An equilibration run of 200 ps in which no exchanges were attempted was followed by 1 ns of exchange equilibration, and finally a 20 ns production run. Only the replica at 300 K was employed for the trajectory analysis.

Statistical analysis of the trajectories was performed following the procedure outlined by Janke(Janke, 2002). The analysis consists of calculating an autocorrelation function for the observable, determining a correlation time from the correlation function, and then using the correlation time to estimate the error of the mean. The correlation function calculated is given by

$$C(t) = \frac{\langle O_i O_{i+t} \rangle - \langle O_i \rangle^2}{\langle O_i^2 \rangle - \langle O_i \rangle^2} \tag{4}$$

where O_i represents the value of the observable O at the ith trajectory time point. The integrated autocorrelation time is

$$\tau = \tfrac{1}{2} + \sum_{j=1}^{N} C(j)\left(1 - \frac{j}{N}\right) \tag{5}$$

where N is the total number of points in the trajectory. As t gets larger, the statistical sampling for the calculation of $C(t)$ deteriorates and the sum diverges. In order to avoid this problem, the sum in equation 5 is truncated once $N>6\tau$. Finally, error of the estimator of the expectation (mean) of the observable O is

(6) $$\varepsilon = \sqrt{2\tau \frac{\sigma^2}{N}}$$

where σ^2 represents the variance of the observable O and N equals 6τ.

Quantum Chemistry/Partial Charge Calculations

All quantum mechanical calculations were performed on ethyl phosphate with the Jaguar package(Jaguar 5.0, 1991-2003) (Schrödinger, Inc.). The initial geometry employed had a P-O bond distance of 1.55 Å for the ester bond and 1.46 Å for the acidic oxygens. Geometry optimization of the phosphate ion was carried out at the HF/6-31G** level, incorporating a condensed phase environment via a self-consistent reaction field (SCRF) algorithm(Marten *et al.*, 1996; Tannor *et al.*, 1994). Single point calculations were also performed at the LMP2/cc-pvtz(-f) level, *in vacuo* and with a SCRF.

Calculation of Experimental Observables

NMR vicinal coupling constants were calculated using the Karplus equation(Karplus, 1959)

(7) $$J(\theta) = A\cos^2(\theta) + B\cos(\theta) + C$$

with A = 6.51, B = -1.3 and C = 1.5(Vuister & Bax, 1993). The angle θ is the dihedral angle between the amide hydrogen and the alpha carbon hydrogen in the backbone of

UCSF LIBRARY

each residue. The average coupling constant over the trajectory was calculated for comparison with experiment. We note that coupling constant data for glycine is usually not included in the fitting of the Karplus equation parameters.

Hydrogen bonds were identified using a distance cutoff of 2.5 Å between a polar hydrogen and the hydrogen bond acceptor, and no angular cutoff. The hydrogen bond probabilities were calculated by counting the number of trajectory frames where a given hydrogen bond was formed and dividing by the total number of frames.

Strictly speaking, a dipole moment is defined for a neutral species by the following expression,

(8) $$\vec{\mu} = \sum_i \vec{R}_i \delta_i ,$$

where $\vec{\mu}$ is the dipole moment, $\vec{R}_i$ is the position vector for the ith atomic charge and δ_i is the charge at that position. In order to capture the changes in both, the phosphorous and acidic oxygen atoms, we choose to define the bond dipole moment of the phosphate P-O bond by equation 8 and set the origin at the bond midpoint. The total bond length used was 1.46 Å, and hence the value of R in equation 8 was 0.73 Å.

Clustering of Trajectories

To gain insight into the structural states visited during the simulations, the trajectories were clustered using the MMTSB tools set(Michael Feig, 2001). The metric for hierarchical clustering was the RMSD between heavy atoms (-jclust option in cluster.pl). This software tool chooses an optimized number of clusters based on the criterion by Xu and co-workders(Xu *et al.*, 1993). In this approach a function, E, of the minimum of between-cluster distances (MBCD) and sum-of-square error (SSE) is evaluated at every hierarchical level; the optimal level corresponds to a maximum in E. E is defined by:

(9) $$E(i) = \frac{M(i)-M(i+1)}{\sqrt{J(i)}-\sqrt{J(i+1)}}$$

where 'i' is the partition level, M(i) and J(i) corresponds to the MBCD and SSE at the partition level 'i', correspondingly. The MBCD, denoted M(i), is defined by:

(10) $$M(i) = \min[d_{ij}] \quad i<j \qquad i,j = 1,2,...,n$$

Where d_{ij} is defined by:

(11) $$d_{jk} = \sqrt{\frac{n_j n_k}{n_j + n_k}}\left|m_j - m_k\right| \quad j,k= 1,2,....,i$$

where n_j and m_j denote the number of members in cluster j and its mean, respectively. SSE (J), denoted J(i), is defined by:

(12) $$J(i) = \sum_{k=1}^{i} J_k$$

where

(13) $$J_k = \sum_{x \in X_k} \|x - m_k\|^2$$

In equation 13 m_k corresponds to the mean of cluster k, as previously defined.

A total of 2000-4000 structures were subjected to clustering from each trajectory (snapshots were taken at 10 ps intervals). A representative structure for each cluster was identified as the member closest to the cluster centroid.

GB/SA single point energy calculations

To study the energetics of hydrogen bonding between a phosphate group and an amide hydrogen, single point energy calculations, using GB/SA as a solvent model, were performed. The calculations were carried out along a collinear P-O – H-N coordinate, and used ethylphosphate and N-methyl ethanamide ($CH_3C(=O)NHCH_3$) as a model system for the hydrogen bonding. In the case where the protonated phosphate moiety was examined, ethyl biphosphate was employed. The energy calculations were performed at 0.125 Å intervals for O—H distances between 1.5 and 11 Å.

Results and Discussion

In this section, results from simulations of the capped Gly-Ser-pSer-Ser peptide using several solvent models are compared with each other and with experimental backbone $^3J_{NH-CH}$ coupling constants. We first consider the choice of atomic charges of the phosphate ion and their effect on the computed observable. This analysis is performed using the TIP3P water model, the Generalized Born–Surface Area (GB/SA) implicit solvent model, and a distance dependent dielectric treatment. We focus primarily on the -2 pSer residue, but present some results for the protonated, -1 pSer peptide for comparison.

Calculation of atomic charges for phosphate ion

We have invested considerable effort in determining atomic partial charges for the phosphate ion, because interactions with the surrounding solvent and intramolecular hydrogen bonding partners depend sensitively on the particular charges assigned. Here we carry out quantum mechanical calculations at several levels of theory, ranging from semi-empirical to *ab initio* calculations with approximate treatment of configuration interaction and solvent effects. We provide recommendations for partial charges that we believe to be most accurate. We also aim to deconvolute, insofar as possible, the effects of different solvent models from the choice of partial charges. Therefore, in much of the rest of the paper, we employ several different sets of partial charges for molecular dynamics simulations.

It is important to note that the electrostatic potential fitting and subsequent test of several charge sets is a means to evaluate the different solvation models. This, however, is not a full parameterization effort of the phosphoserine residue. Modifying a single parameter at a time, allows us to correlate the parameter to changes in the calculated observable.

Many methodologies have been described for computing atomic charges for use in a force field description of a chemical group. The ff99AMBER force field atomic charges were originally computed using a restricted electrostatic potential (RESP)(Cieplak *et al.*, 1995) fitting protocol, with HF/6-31G* quantum calculations. This level of theory overestimates the dipole moment relative to experimental gas phase results, but frequently provides a good estimate of dipole moments in water. Quantum calculations that use larger basis sets and implicitly account for a condensed phase environment are now computationally tractable, and we have chosen to explore the charges produced by some of these more sophisticated *ab initio* calculations as well.

Atomic charges were obtained by electrostatic potential fitting(Chirlian & Francl, 1987; Woods *et al.*, 1990). The structure of the ethyl phosphate ion was taken from the work of Feng et al(Feng et al., 1996). We performed calculations using this geometry as well as a minimized structure, using the 6-31G** basis set and a self consistent reaction field (SCRF)(Marten et al., 1996; Tannor et al., 1994) representation of aqueous solvation. The highest level of theory used for obtaining partial charges was cc-pVTZ(-f)/LMP2/SCRF (i.e., correlation treated at the level of "local" Moller-Plesset second-order perturbation theory(Saebo & Pulay, 1993), and an SCRF treatment of solvent

effects). For comparison, we calculated a set of charges using the semi-empirical AM1-BCC method as implemented in the Antechamber module in AMBER 7. Furthermore, we also compared our results with the charges from the OPLS force field(Jorgensen & Tiradorives, 1988). The results are shown in Table 1.

The results in Table 1 vary significantly and are dependent on the geometry and methodology. However, several conclusions can be drawn. First the OPLS force field parameters are clearly outliers and should not be used. Second, the ester oxygen charge is consistent over the set, –0.66±0.03. Third, contrary to expectations, the polarization of the P-O bond is not larger for the 6-31G* results than for results with larger basis sets. Finally, among the different calculations, the relative variation in bond dipole is greater than the relative variation of charges on the individual atoms, especially the phosphorus. Specifically, the larger partial charges for the P on the non-optimized structure seem to partially compensate for the longer bond length. We use the bond-dipole as the *x*-axis in Figure 1, which summarizes many of the key results discussed below.

Since the different methods in Table 1 yield disparate results for the phosphate atomic charges, we tested the effect of varying the atomic charges on our simulation results. In an effort to explore systematically the range of possible atomic partial charges, we chose not to test every charge set in Table 1, but to test a representative group of charge sets that spans these results, as shown in Table 2. While not all of these charges were obtained from the quantum mechanical calculations, they are quite reasonable. One of the charge sets was directly taken from Table 1 (AM1-BCC charges) while another one is

nearly equal to the results from the HF/cc-pvtz(-f)/LMP2/SCRF calculation (phosphorous charge of 1.495 versus 1.50). The charge set with a phosphorous charge of 1.8 has the same dipole moment that the OPLS charges while the charge set with a phosphorous charge of 1.65 serves as an intermediate between the latter two. Once the phosphorous and acidic oxygen charges were chosen, the ester oxygen was adjusted slightly to maintain an integral charge value for the entire molecule. We note that the ester oxygen only rarely forms any hydrogen bonds in any of our simulations; we do not consider it further.

Dependence of results on atomic charges and solvent model

Molecular dynamics simulations were performed utilizing all charge sets listed in Table 2 with the TIP3P explicit water model, the Generalized Born-Surface Area (GB/SA) implicit solvent model, and distance dependent dielectric (with $\varepsilon=r$ and $\varepsilon=4r$). Figure 1a and 1d show the calculated coupling constants as a function of P-O bond dipole moment. The Gly1[1] and Ser4 coupling constants are insensitive to the charge set and are not shown. For the central two residues, Ser2 and pSer3, the results are dramatically sensitive to the charges employed; however, the changes are systematic. The results are

[1] The coupling constant for Gly1 is consistently miscalculated for all the water models. This may reflect an inherent problem in the force field, problems in the interpretation of the experimental data, or the use of Karplus equation constants not explicitly parameterized for Gly.

reassuring in the sense that agreement with experiment, within statistical error, is attainable in the range of charges tested for both explicit solvent and GB/SA.

Comparing the solvation models, it is clear that the distance dependent dielectric (DDD) simulation results are consistently inferior to those of the explicit solvent and GB/SA simulations (Figures 1a and 1d). Since this is true for all charge sets, it is not due to a poor electrostatic representation of the phosphate ion and indicates a fundamental deficiency of DDD for treating highly charged species. We note that our results, using an $\varepsilon=r$ distance dependent dielectric, differ from previous results obtained using Brownian dynamics(Shen et al., 2001). The differences probably arises from a combination of the following three factors: 1) we did not include a non-polar surface area term in the DDD simulations, 2) the Gasteiger charges employed in the previous study are outside the range studied here (much smaller), and 3) the force fields used are different (AMBER vs. CHARMM). While the specific results are not comparable, both studies agree the DDD treatment yields poor results for the phosphopeptide with the -2 charge (much more so in our results than in the previous study). The inadequate performance of the DDD models is not surprising, given the very crude treatment of solvation afforded by DDD models.

More surprising is the similarity of the results obtained from simulations using TIP3P water and using the GB/SA model. While the absolute coupling constants do not agree precisely, the trend as a function of P-O bond dipole is strikingly similar. This result suggests that a version of GB/SA may be appropriate for the simulation of phosphorylated systems, which would permit greatly improved computational efficiency

in the case of large proteins. The best overall agreement with experiment, for TIP3P and GB/SA, occurs with P-O bond dipole of roughly 8.2–8.8. The partial charges obtained with HF/6-31G* lie in this range, and the higher level HF/cc-pvtz(-f)/LMP2/SCRF partial charges lie at the upper end of this range. On this basis, we suggest that the following partial charges for the -2 pSer/pThr group should be appropriate for most molecular mechanics simulations: +1.45 to +1.50 for the P, -0.99 to -1.05 for the acidic oxygens, and approximately -0.66 for the ester oxygen.

While the focus of this work is a phosphorylated peptide, we performed simulations of the corresponding unphosphorylated peptide. The results were deposited in the supplementary materials. Also in the supplementary materials, are typical graphs of the backbone dihedral angle θ (the relevant angle for the coupling constant calculation) as a function of time. The latter are included to show that sampling was adequate.

In the remainder of this section, we explore the ensembles obtained by the various simulations in more detail, emphasizing the strengths of internal hydrogen bonds, which provide a compact explanation for all of the key trends we observe.

Intramolecular hydrogen bonding interactions

Upon visualizing the simulations(Humphrey *et al.*, 1996), we observed that the phosphate group hydrogen bonds most commonly to the backbone amide NH groups. This type of interaction has biological relevance. In a survey of the Protein Data Bank (PDB), we

have found that backbone amides are the second most frequent hydrogen bond partner, after Arg, for the phosphate groups in phosphorylated proteins (results not shown). In one common motif, the phosphate group forms multiple hydrogen bonds with amide groups at the N-terminus of a helix. To explore the role of hydrogen bonding in the dynamics, we calculated the probability of observing phosphate group hydrogen bonds in the MD trajectory. Table 4 shows the probability, for all models tested, of hydrogen bond formation between the phosphate ion and each of the backbone amide groups.

In general, explicit solvent simulations generate significantly lower (below 10%) hydrogen bond probabilities than the implicit treatments (mostly above 50%). The distance dependent dielectric model, $\varepsilon=r$, grossly overestimates the hydrogen bonding probability (essentially 100%). The other DDD treatment, $\varepsilon=4r$, suffers from the same problem, however, it is slightly less pronounced.

Since intramolecular hydrogen bonding appears to play an important role in these simulations, we studied the hydrogen bonding probability as a function of the P-O dipole (Figures 1b and 1e). The results in Figures 1b and 1e may appear counterintuitive since they show that a larger dipole moment reduces the hydrogen bond probability to an amide NH group. However, this simply indicates that a larger dipole is more stable when solvated in bulk water than when forming a hydrogen bond to a backbone amide. A well organized hydration shell and bulk dielectric response stabilize the ion in water. Since hydrogen bonding disrupts both of these, it is not unreasonable that increasing the P-O dipole may destabilize hydrogen bonds. The experimental observable seems to be

dominated by the equilibrium between hydrogen bonding of the phosphate ion to water versus hydrogen bonding to amide hydrogens. As the equilibrium shifts, so does the corresponding coupling constant (Figures 1c and 1f). Hydrogen bonding imposes a geometrical constraint on the backbone angular distribution, which is reflected in the backbone $^{3}J_{NH\text{-}CH}$ coupling constants. Therefore, there is a clear physical connection between the hydrogen bond probability and the angular distribution for the coupling constant calculation.

We also investigated the correlation of hydrogen bond formation between the phosphate group and the various backbone amide groups, and found evidence of a cooperative transition involving the formation of multiple hydrogen bonds. Specifically, we calculated the shortest distance between the acidic phosphate oxygens and backbone amides of residues 1 and 2. These distances are plotted against each other in Figure 2. For clarity, only every tenth trajectory frame was used for this analysis. The red dots represent the subset of the structures in which the phosphate group formed a hydrogen bond to the third backbone amide. Similar graphs were obtained for both GB/SA and explicit solvent and we show only the GB/SA results. It is clear that phosphate hydrogen bonding to the amide group on pSer3 strongly increases the probability of hydrogen bonding to both the first and second residue. The hydrogen bonding probability for amides 1 and 2 once the hydrogen bond to amide 3 is formed ranges from 50-70% (much above the product of individual probabilities). The probability of forming a hydrogen bond increases by up to10-fold when a hydrogen bond to an adjacent NH group is already present (data not shown). Thus, we find that the dynamics can be understood, to a first

approximation, in terms of a cooperative transition between a state in which the phosphate group is well solvated, and one in which it forms multiple (generally 3) hydrogen bonds with amide groups N-terminal to it. This is not surprising since the entropic cost of hydrogen bonding to an adjacent amide hydrogen is effectively lowered.

Clustering of the GB/SA simulation trajectories yielded two clusters. The representative structures (Figure 3a) are consistent with a cooperative transition; the representative structure of cluster 1 shows 3 hydrogen bonds between the phosphate moiety and backbone amides, while the representative structure for cluster 2 shows none. None of the clustering results for the GB/SA and the TIP3P simulation trajectories produced representative structures involving only 1 or 2 hydrogen bonds to backbone amides; Such intermediate states would be necessary if no cooperativity existed. The relative population between cluster 1 and cluster 2 was approximately 3:1.

The other possible hydrogen bond donors on the phosphopeptide, beyond the backbone NH groups, are the hydroxyl groups on the Ser2 and Ser4 side chains. However, for both the explicit solvent and GB/SA simulations, the hydrogen bonding probability to the OH groups was less than one percent. This result is somewhat surprising, given that hydroxyl groups are almost as common hydrogen bonding partners to phosphate ions as amide hydrogens, in our survey of phosphorylated proteins in the PDB. The minor role for the hydroxyl hydrogens in our simulations may be due to the structure of this peptide. Six covalent bonds separate any of the phosphate acidic oxygens and the amide hydrogen on pSer3; the closest hydroxyl hydrogen is ten bonds away. Thus, there is likely a higher

entropic loss incurred upon forming a hydrogen bond between the phosphate ion and a hydroxyl hydrogen than the corresponding hydrogen bond with a backbone amide hydrogen. This argument also suggests that hydrogen bonds involving the phosphate and hydroxyl groups would not show cooperativity as discussed above for the amide NH groups; in fact, our data indicates that the hydrogen bonds involving the hydroxyl groups form independently. These entropy arguments, combined with the 5:2 amide to hydroxyl hydrogen ratio, provide a qualitative rationalization for the low hydrogen bonding probability to the hydroxyl groups in the GB/SA simulations.

In contrast to the explicit solvent and GB/SA results, hydrogen bonding between the phosphate ion and hydroxyl groups is observed in the simulations using a distance dependent dielectric. The percentage is almost invariant as a function of charge and is nearly 100% for $\varepsilon=r$, and 25-30% for $\varepsilon=4r$. The simulations using a $\varepsilon=r$ distance dependent dielectric were typified by the conformation displayed in Figure 4. The peptide is arranged so that every possible hydrogen bond donor interacts with the phosphate group.

Comparison of explicit solvent models

We have performed simulations of the -2 phosphopeptide with 3 explicit water models—TIP3P, TIP4P and SPC/E—for a single set of partial charges (those with the phosphorous charge of 1.5). The calculated coupling constants are displayed in Table 3. The TIP4P

and SPC/E water models generate results that are quantitatively but not qualitatively different than the TIP3P results. In particular, both TIP4P and SPC/E show less hydrogen bonding to the amide NH groups than TIP3P. These results seem reasonable based on known properties of these solvent models. In particular, TIP4P water is known to be more structured than TIP3P(Guillot, 2002; Jorgensen et al., 1983). A recent study(Mark & Nilsson, 2001) has also shown SPC/E water to be more structured than TIP3P. We hypothesize that a larger degree of structure around the phosphate ion increases the desolvation cost for the formation of a hydrogen bond.

Despite these minor differences among the results, all of the explicit solvent models evaluated here appear to be largely successful in modeling the condensed phase environment for a phosphorylated species. This is encouraging since these models lack polarizability.

Simulations of the protonated phosphopeptide

In addition to the doubly negative phosphopeptide, we also simulate the singly charged (-1), protonated species. Except for the extra proton and the partial charges on the phosphate group, the phosphopeptide parameters were unchanged. For computational efficiency, this system was only simulated using the GB/SA model. To obtain atomic charges, methyl biphosphate was subjected to: 1) a gas phase HF/6-31G** geometry optimization followed by 2) a single point electronic structure calculation

HF/LMP2/SCRF/cc-pvtz(-f) and finally 3) electrostatic potential fitting. The partial charges for the phosphorous atom and the acidic oxygens were 1.5 and –0.98 respectively, while the hydroxyl oxygen and hydrogen were –0.7 and 0.5 respectively. These results seem reasonable since the phosphorous and acidic oxygen partial charges are similar to the set with a bond dipole moment of 8.75 D in Table 2. The coupling constant results are reported in Table 5, where they are compared with the experimental results at low pH. The results are in agreement with experiment to within 0.4 Hz (except for Gly1, as discussed above). Moreover, the relatively large shifts in the coupling constants for pSer3 and Ser4 between the singly and doubly charged species are quantitatively reproduced by the simulations, to within 0.2 Hz.

The shifts in the coupling constants with protonation state of the phosphate appear to be caused by quantitative but not qualitative differences in hydrogen bonding propensity. Table 6 shows that the overall P-O/N-H hydrogen bonding probability is slightly larger (up to 12% higher) for the -1 species, consistent with our expectation that a lower overall charge reduces the free energy cost of hydrogen bonding to backbone amides. Unlike the unprotonated phosphate, where all 3 acidic oxygens are equivalent, the hydrogen bonding propensities of the 3 oxygens of the protonated phosphate group are dramatically different. The protonated oxygen, which can acts as either a donor or acceptor, shows little hydrogen bonding; the asymmetry in hydrogen bonding of the other two oxygens is likely due to their stereochemical difference. The unprotonated oxygens compensate for the limited hydrogen bonding of the protonated oxygen and produce a small net increase in the overall hydrogen bonding probability.

The change in overall hydrogen bonding propensity and other ensemble properties is remarkably small between the two protonation states, given the large change in the partial charges on the phosphate group. In Figure 5, we explore the strengths of the P-O/N-H hydrogen bonds in more detail by plotting the energy of interaction (force field + GB/SA) for a model system consisting of ethylphosphate and N-methyl ethanamide. Specifically, we performed single point calculations for a series of conformations in which the molecules were constrained such that the P-O and H-N vectors were collinear, for both unprotonated (-2) and singly protonated (-1) phosphate. (For the -1 species, the P-O group involved in forming the hydrogen bond is unprotonated.) For both protonation states, the hydrogen bond is weakly favorable, but the singly protonated phosphate has a somewhat (~2 kT) deeper well. The individual components of the energy are plotted in Figure 6. The stronger electrostatic interaction of the -2 species is slightly exceeded by the larger desolvation penalty yielding the small overall energy difference.

The ensemble of structures for the singly charged phosphopeptide, like the doubly charged phosphopeptide, shows an equilibrium between two states (no hydrogen bonds and multiple backbone hydrogen bonds), as shown in Figure 3. The representative structures from the two clusters are very similar for the two charge states. As in the case of the doubly charged phosphopeptide, only 2 clusters were found.

Conclusion

In this work we have simulated a capped phosphorylated peptide, Ace-Gly-Ser-pSer-Ser-Nme, using a variety of solvent models, and compared the results to each other and to experimental observables from an NMR experiment. The major conclusions are

1. The quality of agreement to experiment depends sensitively on both the choice of partial charges for the phosphorylated side chain and the solvent model.
2. The explicit solvent models (TIP3P, TIP4P, SPC/E) and a Generalized Born implicit solvent model (GB/SA(Tsui & Case, 2000a)) both gave good agreement with experiment given appropriate partial charges. The "optimal" partial charges are slightly different for each solvent model. The apparently good performance of the GB/SA solvent model is encouraging, because it provides a much more efficient means of simulating phosphorylated proteins than the explicit solvent models. Nonetheless, we do not wish to overstate the significance of this result; more work is clearly needed. First, in other work, one of us (MPJ) has elucidated specific deficiencies of implicit solvent models related to the inability to represent first-shell solvation effects(Yu *et al.*, 2004), and these same deficiencies likely apply to phosphorylated residues as well. Second, in simulations on the unphosphorylated Gly-Ser-Ser-Ser peptide, we have found that explicit solvent simulations give significantly better agreement with experiment than GB/SA (results in supplementary materials).
3. Distance dependent dielectric solvent models were unable to properly simulate the condensed phase environment for the phosphorylated peptide, over the range of

partial charges tested here, due to gross overstabilization of hydrogen bonds involving the phosphate group.

4. We observed a strong correlation between the calculated NMR coupling constants and the propensity to form hydrogen bonds between the phosphate group and backbone amide NH groups. This is reasonable, because hydrogen bonding imposes a restraint on the molecule that it would otherwise not have. To a reasonable approximation, the simulations with explicit solvent and GB/SA show transitions between two states of the phosphopeptide: one in which the phosphate forms no hydrogen bonds to the peptide and is thus fully solvated, and one in which it forms multiple hydrogen bonds to NH groups, especially the NH group on the phosphorylated residue itself and the residues N-terminal to it. Hydrogen bonds to the OH groups on the Ser side chains appear to be unimportant in these simulations.
5. The strengths of hydrogen bonds involving the phosphate group involve a subtle trade-off between the direct Coulomb's law electrostatic attraction and the free energy cost of partially desolvating the phosphate group. Although this conclusion is not new or unique to this study, this trade-off is particularly striking in the case of the phosphate group with the -2 charge. The inability of the distance dependent dielectric simulations to reproduce the experimental results is simply due to the lack of any treatment of desolvation. The strengths of the P-O/N-H hydrogen bonds decrease with increasing dipole on the P-O bond, because the desolvation penalty increases more rapidly than the favorable dipole/dipole interaction.
6. The ensemble averages obtained for the phosphopeptide in the -1 (singly protonated phosphate) and -2 (fully deprotonated phosphate) charge states are surprisingly

similar. The results with the singly protonated phosphate show slightly higher propensities for hydrogen bonding to the backbone amide NH groups. Upon increasing the charge on the phosphate from -1 to -2, the strength of the electrostatic attraction of course increases significantly, but this effect is very nearly cancelled by the increase in the free energy cost of desolvation upon forming a hydrogen bond. This interpretation is slightly complicated by the fact that the protonated oxygen on the -1 phosphate can act as either a donor or acceptor.

In this work, we have examined only a subset of intriguing questions regarding solvation of the 2- phosphorylated residues. In ongoing work, we are performing additional explicit solvent simulations to study in more detail the solvent structures around the phosphate group in solution and when it forms hydrogen bonds, the possible role of dielectric saturation(Jayaram *et al.*, 1989), and the role of solvent polarizability; these results will be reported elsewhere.

Geometry Optimized	Method	Phosphorous Charge	Acidic Oxygen charge	Ester Oxygen Charge	P-O bond dipole (D)
No	AM1-BCC	1.371	-0.969	-0.655	8.193
Yes	HF/6-31G*	1.449	-0.992	-0.665	8.546
Yes	HF/cc-pvtz(-f)	1.568	-1.025	-0.669	9.079
Yes	HF/cc-pvtz(-f)/LMP2/SCRF	1.495	-1.050	-0.657	8.911
No	HF/6-31G*	1.767	-1.077	-0.686	9.957
No	HF/cc-pvtz(-f)	1.778	-1.082	-0.687	10.013
No	HF/cc-pvtz(-f)/LMP2/SCRF	1.719	-1.114	-0.687	9.919
N/A	OPLS	1.920	-1.120	-0.700	10.644

Table 1. Several charge sets proposed for a phosphate group. Abbreviations are defined in the text.

Phosphorous Charge	Acidic Oxygen Charge	P-O bond dipole (D)
1.371	-0.969	8.193
1.400	-0.900	8.053
1.500	-1.000	8.753
1.650	-1.100	9.628
1.800	-1.200	10.504

Table 2. Charge sets used in this study. All charges are in a.u.

	TIP3P	TIP4P	SPC/E	GB/SA	DDD (ε=r)	DDD	Experiment
Gly1	4.47 (0.03)	4.71 (0.04)	4.55 (0.03)	4.12 (0.03)	2.95 (0.01)	3.92 (0.02)	5.63 (0.12)
Ser2	6.28 (0.23)	5.99 (0.18)	6.35 (0.22)	6.62 (0.11)	9.04 (0.10)	8.33 (0.01)	6.65 (0.12)
pSer3	6.24 (0.23)	6.51 (0.28)	6.02 (0.20)	5.73 (0.10)	9.58 (0.40)	8.68 (0.02)	5.48 (0.12)
Ser4	7.72 (0.10)	7.72 (0.06)	7.84 (0.07)	7.37 (0.03)	9.26 (0.04)	8.05 (0.01)	6.93 (0.12)

Table 3. Comparison of calculated J coupling constants with the experimental result. The error of the mean for each result is in parenthesis. For the experimental measurements, the reported error is also in parenthesis.

	TIP3P	TIP4P	SPC/E	GB/SA	DDD (ε=r)	DDD
Gly1	6.4% (3%)	0.0% (0%)	0.7% (0.7%)	56.3% (4%)	100.0% (0.01%)	80.2%
Ser2	7.6% (3%)	0.2% (0.2%)	1.5% (0.7%)	50.8% (4%)	99.9% (0.02%)	78.5%
pSer3	8.4% (3%)	0.5% (0.3%)	2.6% (0.9%)	60.4% (4%)	100.0% (0.0005%)	95.6%
Ser4	0.1% (0.05%)	0.0% (0%)	0.1% (0.05%)	7.8% (2%)	100.0% (0.01%)	81.0%

Table 4. Phosphate group hydrogen bonding probability to peptide backbone amides. Each backbone amide is labeled by its residue name and number. The standard error of the mean is in parenthesis.

	Doubly charged phosphate		Singly charged phosphate	
	GB/SA	Experiment	GB/SA	Experiment
Gly1	4.12 (0.03)	5.63 (0.12)	3.89 (0.04)	5.66 (0.11)
Ser2	6.62 (0.11)	6.65 (0.12)	6.89 (0.34)	6.52 (0.11)
pSer3	5.73 (0.10)	5.48 (0.12)	6.69 (0.14)	6.52 (0.11)
Ser4	7.37 (0.03)	6.93 (0.12)	7.70 (0.03)	7.47 (0.11)

Table 5. Effect of phosphate ionization state on calculated coupling constants. Experimental results are listed for comparison. In parenthesis is standard deviation of calculated results or reported experimental error, in the case of the experimental result.

	Doubly charged phosphate				Singly charged phosphate			
	Overall	O1	O2	O3	Overall	OH	O2	O3
Gly1	56.3% (4%)	19.0% (3%)	18.5% (3%)	18.9% (4%)	57.8% (7%)	0.0% (0%)	44.0% (8%)	13.0% (5%)
Ser2	50.8% (4%)	17.0% (3%)	17.3% (3%)	16.9% (4%)	51.1% (6%)	0.0% (0%)	40.0% (7%)	11.1% (4%)
pSer3	60.4% (4%)	20.3% (3%)	19.9% (3%)	20.7% (4%)	72.2% (8%)	18.2% (1%)	47.2% (8%)	14.2% (5%)
Ser4	7.8% (2%)	2.3% (0.6%)	1.6% (0.5%)	3.8% (1%)	17.1% (4%)	3.1% (0.6%)	15.8% (4%)	4.0% (0.2%)

Table 6. Effect of phosphate ionization state on hydrogen bonding probability. The hydrogen bonding probabilities were computed for each phosphor-oxygen. In the case of the singly charged phosphate, the hydroxyl oxygen is labeled OH. All calculations performed using GB/SA solvent. The standard error of the mean is in parenthesis.

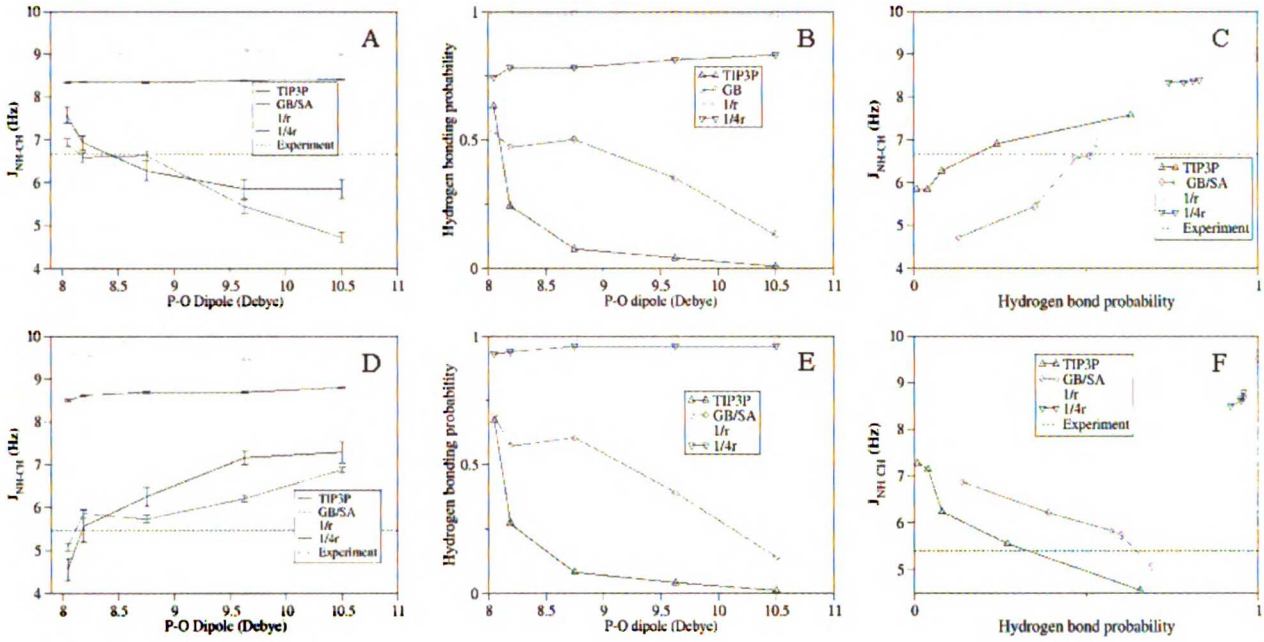


Figure 1. Phosphopeptide backbone coupling constants depend on degree of hydrogen bonding between the phosphate group and backbone amide NH groups. The P-O dipole modulates hydrogen bonding allowing a clear isolation of its effects. (A) Dependence of calculated J coupling constant on P-O dipole moment for Ser2. (B) Dependence of phosphate hydrogen bonding to Ser2 backbone amide on the P-O dipole moment. The larger dipole is better stabilized by solvent rather than by hydrogen bonding. (C) Calculated Ser2 J coupling constant as a function of hydrogen bonding probability to its backbone amide. The dependence observed in this graph links hydrogen bonding to changes in backbone conformational population differences. (D)-(F) are the corresponding graphs for pSer3.

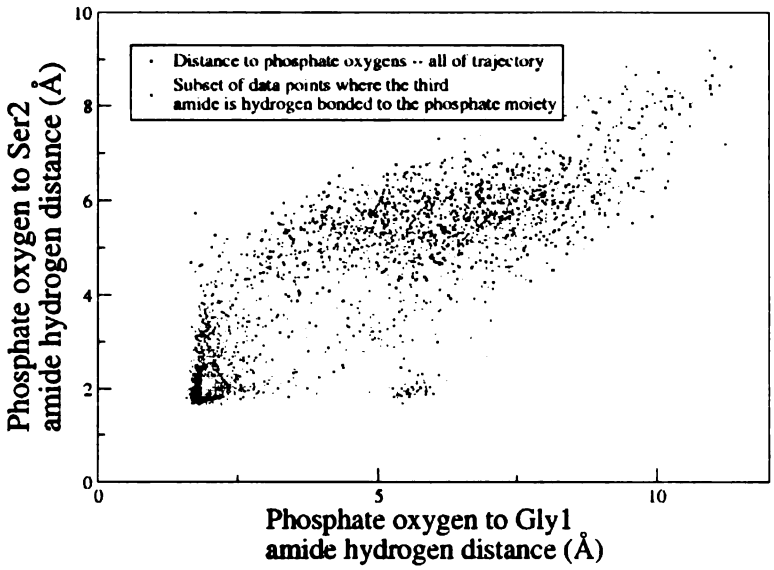


Figure 2. Cooperativity of formation of hydrogen bonds to the backbone NH groups. The axes represent the shortest distance between the phosphate acidic oxygens and the NH groups on Gly1 (x) and Ser2 (y). The points in black were calculated for all structures clustered (every 10 ps). The points in red correspond to a subset of the structures in which the phosphate group forms a hydrogen bond to the pSer3 amide group. The data shown are from the GB/SA simulation. The results for explicit solvent (TIP3P) are similar, but the density of points in the lower left corner is smaller.

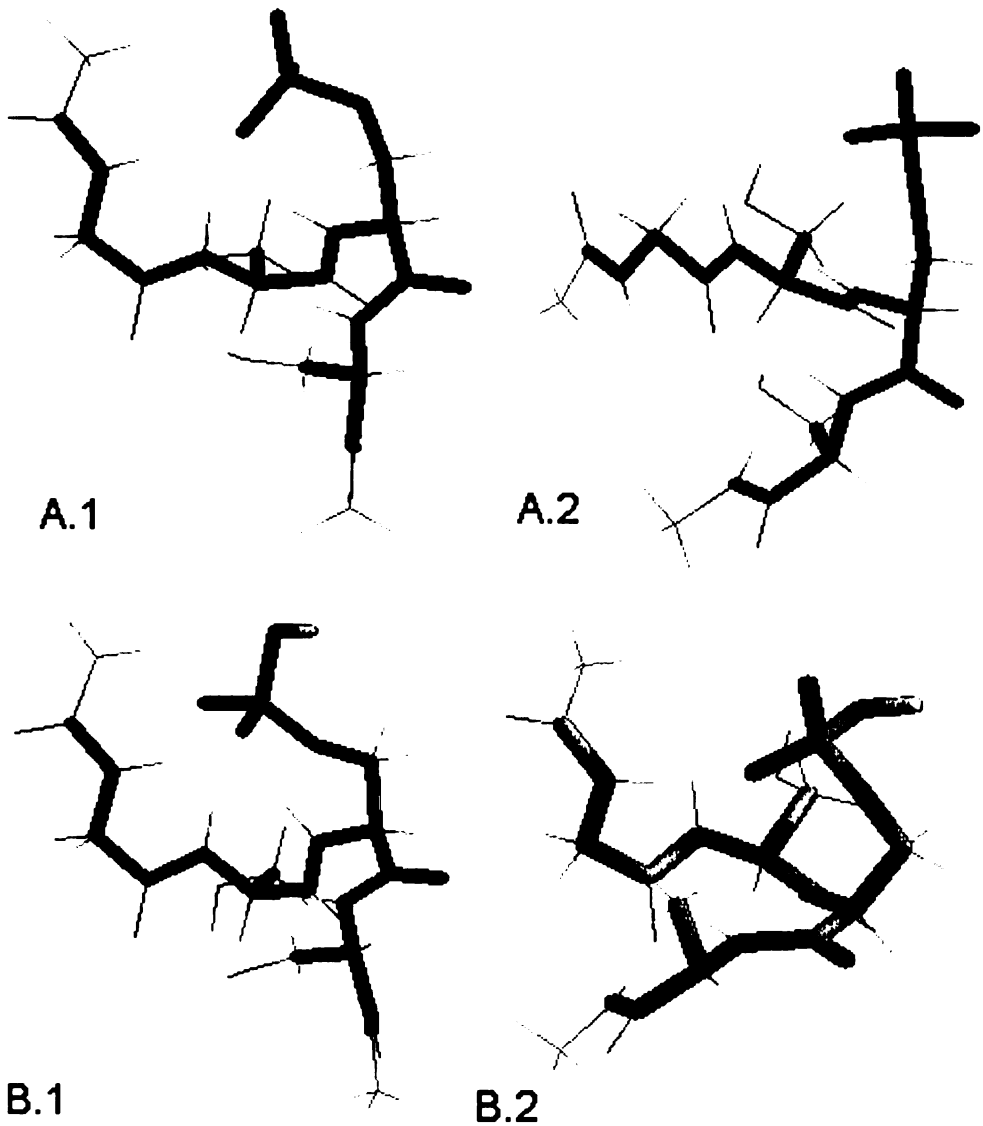


Figure 3. Representative structures from GB/SA simulation of doubly charged (A) and singly charged (B) phosphopeptides. The backbone and phosphoserine side chain is shown in sticks. A.1 and B.1 are similar conformations in which the phosphate forms multiple hydrogen bonds with backbone amide NH groups N-terminal to it. A.2 and B.2 are conformations in which the phosphate does not form internal hydrogen bonds and is well solvated.

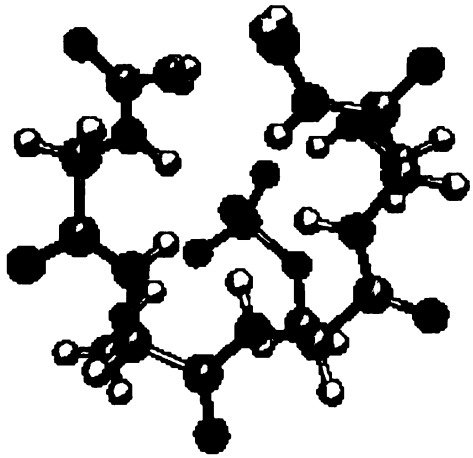


Figure 4. Snapshot from simulation of the phosphopeptide using a distance dependent dielectric ($\varepsilon = r$). The phosphate groups forms hydrogen bonds with all possible hydrogen bond donors.

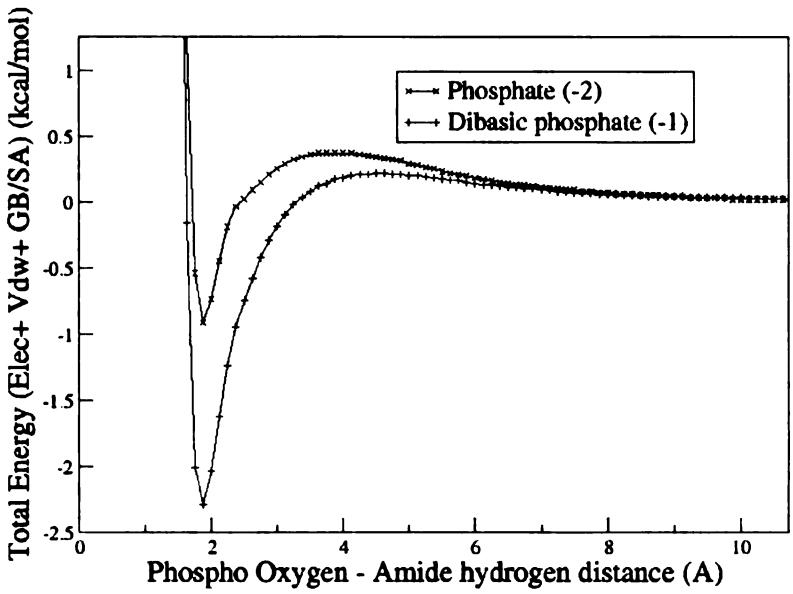


Figure 5. Energy for hydrogen bonding between the phosphate group (singly and doubly charged) and an amide NH group. Solvent effects were incorporated using GB/SA. The reaction coordinate was a direct linear approach between the P-O and N-H bond vectors. For the singly charged, protonated phosphate, the hydrogen bond was made to an unprotonated oxygen.

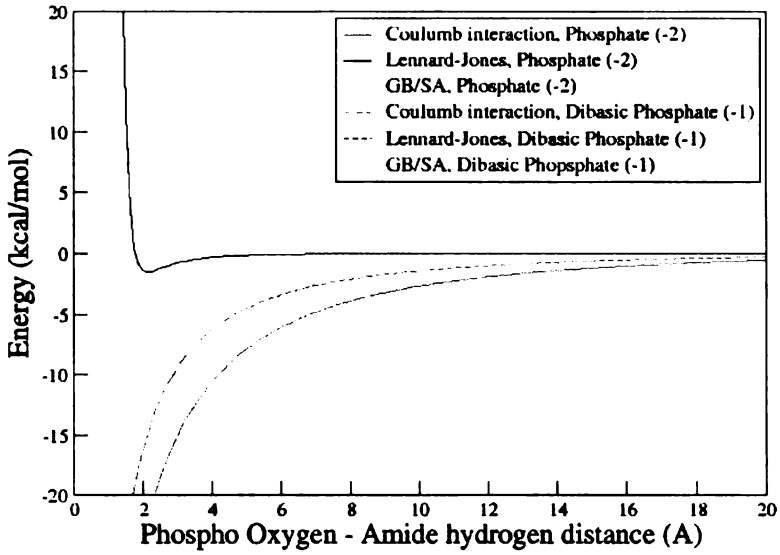


Figure 6. Energy components for the energy curves of Figure 5. Comparison of energetics in hydrogen bonding to the fully –2 phosphate (solid lines) moiety versus the –1, protonated phosphate (dashed lines). The Lennard-Jones nonbonded interactions for the phosphate group are the same in both protonation states, and thus only a single line is seen.

References

Arthanari, H., McConnell, K. J., Beger, R., Young, M. A., Beveridge, D. L., & Bolton, P. H. (2003). Assessment of the molecular dynamics structure of DNA in solution based on calculated and observed nmr noesy volumes and dihedral angles from scalar coupling constants. *Biopolymers, 68*(1), 3-15.

Audette, G. F., Engelmann, R., Hengstenberg, W., Deutscher, J., Hayakawa, K., Quail, J. W., et al. (2000). The 1.9 a resolution structure of phospho-serine 46 hpr from enterococcus faecalis. *J Mol Biol, 303*(4), 545-553.

Berendsen, H. J. C., Grigera, J. R., & Straatsma, T. P. (1987). The missing term in effective pair potentials. *Journal of Physical Chemistry, 91*(24), 6269-6271.

Berendsen, H. J. C., Postma, J. P. M., Vangunsteren, W. F., Dinola, A., & Haak, J. R. (1984). Molecular-dynamics with coupling to an external bath. *Journal of Chemical Physics, 81*(8), 3684-3690.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Research, 28*(1), 235-242.

Beveridge, D. L., & McConnell, K. J. (2000). Nucleic acids: Theory and computer simulation, y2k. *Current Opinion in Structural Biology, 10*(2), 182-196.

Chen, Z., Gibson, T. B., Robinson, F., Silvestro, L., Pearson, G., Xu, B., et al. (2001). Map kinases. *Chem Rev, 101*(8), 2449-2476.

Chirlian, L. E., & Francl, M. M. (1987). Atomic charges derived from electrostatic potentials - a detailed study. *Journal of Computational Chemistry, 8*(6), 894-905.

Cieplak, P., Cornell, W. D., Bayly, C., & Kollman, P. A. (1995). Application of the multimolecule and multiconformational resp methodology to biopolymers - charge derivation for DNA, rna, and proteins. *Journal of Computational Chemistry, 16*(11), 1357-1377.

D.A. Case, D. A. P., J.W. Caldwell, T.E. Cheatham III, J. Wang, W.S. Ross, C.L. Simmerling, T.A. Darden, K.M. Merz, R.V. Stanton, A.L. Cheng, J.J. Vincent, M. Crowley, V. Tsui, H. Gohlke, R.J. Radmer, Y. Duan, J. Pitera, I. Massova, G.L. Seibel, U.C. Singh, P.K. Weiner and P.A. Kollman. (2002). Amber 7. San Francisco, CA: University of California.

Darden, T., York, D., & Pedersen, L. (1993). Particle mesh ewald - an n.Log(n) method for ewald sums in large systems. *Journal of Chemical Physics, 98*(12), 10089-10092.

Dominy, B. N., & Brooks, C. L. (1999). Development of a generalized born model parametrization for proteins and nucleic acids. *Journal of Physical Chemistry B, 103*(18), 3765-3773.

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., & Pedersen, L. G. (1995). A smooth particle mesh ewald method. *Journal of Chemical Physics, 103*(19), 8577-8593.

Feng, M. H., Philippopoulos, M., MacKerell, A. D., & Lim, C. (1996). Structural characterization of the phosphotyrosine binding region of a high-affinity sh2 domain-phosphopeptide complex by molecular dynamics simulation and chemical shift calculations. *Journal of the American Chemical Society, 118*(45), 11265-11277.

Guillot, B. (2002). A reappraisal of what we have learnt during three decades of computer simulations on water. *Journal of Molecular Liquids, 101*(1-3), 219-260.

Humphrey, W., Dalke, A., & Schulten, K. (1996). Vmd: Visual molecular dynamics. *Journal of Molecular Graphics, 14*(1), 33-&.

J.-P. Ryckart, G. C. H. J. C. B. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Computational Physics, 23*, 327-341.

Jaguar 5.0, S., L.L.C. (1991-2003). Jaguar (Version 5.0). Portland, OR.

Janke, W. (2002). Statistical analysis of simulations: Data correlation and error estimation. In M. D. Grotendorst J., Muramatsu A. (Ed.), *Quantum simulations of*

complex many-body systems: From theory to algorithms, lecture notes (Vol. 10, pp. 423-445). Julich: John von Neumann Institute for Computing.

Jayaram, B., Fine, R., Sharp, K., & Honig, B. (1989). Free-energy calculations of ion hydration - an analysis of the born model in terms of microscopic simulations. *Journal of Physical Chemistry, 93*(10), 4320-4327.

Johnson, L. N., & Lewis, R. J. (2001). Structural basis for control by phosphorylation. *Chem Rev, 101*(8), 2209-2242.

Johnson, L. N., & O'Reilly, M. (1996). Control by phosphorylation. *Curr Opin Struct Biol, 6*(6), 762-769.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics, 79*(2), 926-935.

Jorgensen, W. L., & Tiradorives, J. (1988). The opls potential functions for proteins - energy minimizations for crystals of cyclic-peptides and crambin. *Journal of the American Chemical Society, 110*(6), 1666-1671.

Karplus, M. (1959). *Journal of Physical chemistry, 30*, 11-15.

Makarov, V., Pettitt, B. M., & Feig, M. (2002). Solvation and hydration of proteins and nucleic acids: A theoretical view of simulation and experiment. *Acc Chem Res, 35*(6), 376-384.

Mark, P., & Nilsson, L. (2001). Structure and dynamics of the tip3p, spc, and spc/e water models at 298 k. *Journal of Physical Chemistry A, 105*(43), 9954-9960.

Marten, B., Kim, K., Cortis, C., Friesner, R. A., Murphy, R. B., Ringnalda, M. N., et al. (1996). New model for calculation of solvation free energies: Correction of self-consistent reaction field continuum dielectric theory for short-range hydrogen-bonding effects. *Journal of Physical Chemistry, 100*(28), 11775-11788.

Martens, J. R., Kwak, Y. G., & Tamkun, M. M. (1999). Modulation of kv channel alpha/beta subunit interactions. *Trends Cardiovasc Med, 9*(8), 253-258.

Michael Feig, J. K., Charles L. Brooks, III. (2001). Mmtb tool set. The Scripps Research Institute.

Mitsutake, A., Sugita, Y., & Okamoto, Y. (2003a). Replica-exchange multicanonical and multicanonical replica-exchange monte carlo simulations of peptides. I. Formulation and benchmark test. *Journal of Chemical Physics, 118*(14), 6664-6675.

Mitsutake, A., Sugita, Y., & Okamoto, Y. (2003b). Replica-exchange multicanonical and multicanonical replica-exchange monte carlo simulations of peptides. Ii. Application to a more complex system. *Journal of Chemical Physics, 118*(14), 6676-6688.

Norberg, J., & Nilsson, L. (2002). Molecular dynamics applied to nucleic acids. *Accounts of Chemical Research, 35*(6), 465-472.

Okur, A., Strockbine, B., Hornak, V., & Simmerling, C. (2003). Using pc clusters to evaluate the transferability of molecular mechanics force fields for proteins. *Journal of Computational Chemistry, 24*(1), 21-31.

Patel, A. J., & Honore, E. (2001). Properties and modulation of mammalian 2p domain k+ channels. *Trends Neurosci, 24*(6), 339-346.

Saebo, S., & Pulay, P. (1993). Local treatment of electron correlation. *Annual Review of Physical Chemistry, 44*, 213-236.

Shen, T. Y., Wong, C. F., & McCammon, J. A. (2001). Atomistic brownian dynamics simulation of peptide phosphorylation. *Journal of the American Chemical Society, 123*(37), 9107-9111.

Smart, J. L., & McCammon, J. A. (1999). Phosphorylation stabilizes the n-termini of alpha-helices. *Biopolymers, 49*(3), 225-233.

Stultz, C. M., Levin, A. D., & Edelman, E. R. (2002). Phosphorylation-induced conformational changes in a mitogen-activated protein kinase substrate - implications for tyrosine hydroxylase activation. *Journal of Biological Chemistry, 277*(49), 47653-47661.

Sugita, Y., & Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters, 314*(1-2), 141-151.

Tannor, D. J., Marten, B., Murphy, R., Friesner, R. A., Sitkoff, D., Nicholls, A., et al. (1994). Accurate first principles calculation of molecular charge-distributions and solvation energies from ab-initio quantum-mechanics and continuum dielectric theory. *Journal of the American Chemical Society, 116*(26), 11875-11882.

Tholey, A., Lindemann, A., Kinzel, V., & Reed, J. (1999). Direct effects of phosphorylation on the preferred backbone conformation of peptides: A nuclear magnetic resonance study. *Biophysical Journal, 76*(1), 76-87.

Toukmaji, A., Sagui, C., Board, J., & Darden, T. (2000). Efficient particle-mesh ewald based approach to fixed and induced dipolar interactions. *Journal of Chemical Physics, 113*(24), 10913-10927.

Tsui, V., & Case, D. A. (2000a). Molecular dynamics simulations of nucleic acids with a generalized born solvation model. *Journal of the American Chemical Society, 122*(11), 2489-2498.

Tsui, V., & Case, D. A. (2000b). Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers, 56*(4), 275-291.

Vermeulen, K., Van Bockstaele, D. R., & Berneman, Z. N. (2003). The cell cycle: A review of regulation, deregulation and therapeutic targets in cancer. *Cell Prolif, 36*(3), 131-149.

Vuister, G. W., & Bax, A. (1993). Quantitative j correlation - a new approach for measuring homonuclear 3-bond j(h(n)h(alpha) coupling-constants in n-15-enriched proteins. *Journal of the American Chemical Society, 115*(17), 7772-7777.

Woods, R. J., Khalil, M., Pell, W., Moffat, S. H., & Smith, V. H. (1990). Derivation of net atomic charges from molecular electrostatic potentials. *Journal of Computational Chemistry, 11*(3), 297-310.

Xu, S., Kamath, M. V., & Capson, D. W. (1993). Selection of partitions from a hierarchy. *Pattern Recognition Letters, 14*(1), 7-15.

Yu, Z., Jacobson, M. P., Rapp, C. S., & Friesner, R. A. (2004). First-shell solvation of ion pairs: Correction of systematic errors in implicit solvent models. *Journal of Physical Chemistry B, 108*, 6643-6654.

Chapter 5: Polarizability contributes to salt-bridge partner preferences of phosphorylated residues

Introduction

In chapter 4, I considered hydrogen bonding of phosphorylated residues in the context of local contacts. Here, I examine interactions involving residues typically distant, in sequence, from the phosphorylated residues, that form after phosphorylation and which involve a conformational change. These contacts are important since their accompanying conformational changes are important for the biological function of the proteins in question(Johnson & Lewis, 2001).

Arginine and lysine are the most common salt-bridge partners for phosphorylated residues(Johnson & Lewis, 2001) and a preliminary survey of salt-bridge contacts in the Brookhaven protein databank (PDB)(Berman *et al.*, 2000) shows a preference for arginine over lysine. The reason for the preference is unclear, but it would be helpful to determine if the survey reflects a stronger interaction energy between the phosphate moiety and arginine than with lysine. If this is the case, this information would be helpful in protein engineering efforts to mimic structural control by phosphorylation. While arginine and lysine bear the same overall charge, it is distributed markedly differently; thus one could imagine the

guanidino group's delocalized charge may be more malleable and could more easily polarize in response to the phosphate moiety. To test this hypothesis *ab initio* calculations, including a self-consistent reaction field, followed by electrostatic potential charge fitting, were performed; shifts in fitted charges were used as an indication of charge rearrangement. Since the singly charged (-1), biphosphate, species may also be present at physiological pH, it was also included in the analysis.

Occasionally, it is possible to mimic phosphorylation by mutation of the target residue to aspartic or glutamic acid; why this only happens part of the time is unkown. One possibility is that the smaller charge of the carboxylate ion makes less stable salt bridges relative to the corresponding phosphate interactions. This hypothesis is tested here by carrying out the same type of *ab initio* calculations on an aspartic acid analog.

In terms of electron density distributions, the case of phosphor-aspartate is particular because of the possible conjugation with the carbonyl group. This case was also examined to capture any differences in polarizability due to its unique structure.

Methods

Survey of phosphorylated residues salt-bridge partners in the PDB

Salt bridge interactions, involving phospohrylated residues, were surveyed from structures available in the PDB. The PDB codes for the structures employed in this survey are:

14ps 1b4g 1b4i 1bkx 1bx6 14ps 1b4g 1b4i 1bkx 1bx6 1c8l 1f34 1f8a 1fa9 1fu0 1gkk 1h4x ih4z 1hjk 1i7w 1jdy 1k35 1kdx 1khx 1kkm 1lwn 1lwo 1qja 1qjb 1stc 1vkl iygp 2pil 1cm8 1e9h 1f5d 1fot 1fq1 1g6g 1gxc 1i8g 1i8h 1ib1 1j4l 1j4p 1j4q 1j4x 1k2n 1k3n 1k3q 2erk 1a07 1a1b 1a1c 1a1e 1a31 1a81 1ad5 1aot 1aou 1bf5 1bg1 1bm2 1bmb 1crx 1csy 1csz 1d4w 1een 1eeo 1f1w 1fbv 1fhr 1fmk 1fpr 1fu5 1fyr 1g1f 1g1g 1g1h 1gag 1h9o 1irs 1j4k 1jyq 1jyr 1k2m 1ka6 1kc2 1ksw 1m0v 1pic 1gpa 1pty 1qcf 1qg1 1qpc 1qpd 1qpj 1shc 1tce 1tze 1zfp 2cbl 2hck 2nmb 2src 1vr2 1phz 4icd 1qmp 1d5w 1djm 1d4w 1f4v 1dc8 3ezb 1fuo

The structures were found by searching for structures that had the residues TPO, SEP or PTR. Only one protein structure per sequence was allowed. Then, salt bridges to arginine, lysine and histidine side chains were investigated, where the heavy atom distance cut-off was 4.5 A. For reference, the number of these residues within 15 A was also reported. A python script was used to search through the structures, identify the phosphorylated residues, measure their distance to potential salt-bridge partners and report the findings.

The molecules and interactions considered

The following model molecules, which correspond to the side-chain fragments of their target residues, were used: phospho-acetate (singly and doubly charged), methyl phosphate, methyl biphosphate, acetate, 1-aminobutane, and 1-guanidopropane. Their geometries were obtained from the OPLS (Jorgensen & Tiradorives, 1988) force field.

Potential of mean force (PMF) calculations were performed between pairs of these molecules in collinear and planar geometries. A collinear geometry refers to a salt bridge where the chemical bonds involving the electron donor and acceptor align (see figure 1). A planar approach involves two simultaneous hydrogen bonds where the two pairs of atoms and their chemical bonds lie in the same plane. A typical planar approach occurs between two hydrogens from a guanidino group and two acidic oxygens of a phosphate moiety (see figure 1). Similar geometries can be constructed for an amine by using two of its hydrogens to interact with two acidic oxygens.

QM/SCRF calculations

Quantum mechanical calculations coupled to a self consistent reaction field (SCRF)(Marten *et al.*, 1996) were performed using the Jaguar software package (Jaguar 5.0, 1991-2003). For both, planar and collinear approaches, point energy

calculations were carried out at 0.25 A intervals between 3.0 and 12.25 A. The atoms used to measure these distances are listed in Table 2, while the atom names for each molecule studied are evident from figure 2. The Hartree-Fock calculations used a cc-pVTZ(-f) basis set and a subsequent electron correlation correction was included. The electron correlation correction was performed at the level of the local Moller-Plesset second order perturbation theory (LMP2)(Saebo & Pulay, 1993). After an initial gas-phase calculation, atomic charges were derived via ESP fitting(Cieplak *et al.*, 1995). The reaction field response for the latter charges was evaluated using a Poisson-Boltzmann solver and surface charges were derived. The quantum calculation was then repeated with the surface charges from the reaction field and new atomic charges were calculated. This procedure was iterated until the overall energy converged.

For each pair, the interaction energy was defined as the difference between the minimum in the PMF and the energy at 12.25 A. Charges at these two distances were also compared in search of evidence of polarizability.

Results and discussion

PDB survey

Table 1 shows the number of salt-bridge contacts found between phoshorylated residues and arginine, lysine and histidine side-chains. The data shows a clear preference for arginine (60 vs 13) while their relative abundance is far less

disparate (150 vs 100). This observation motivated the following calculations that capture the differences in polarizability between the arginine and lysine side chains.

Arginine versus lysine polarizability

The differences in interaction energy between hydrogen bonding involving lysine versus arginine depend on the geometry of approach (Table 3). In the linear approach, lysine has the strongest interactions by ~1-2 kcal/mol in most cases. That trend is reversed in the co-planar case, where arginine has interactions 4-6 kcal/mol more stable. This behavior is reflected in the charge shift data. In the collinear geometry data, the amino nitrogen charge shifts by ~0.1e or causes polarization of its interaction partner (Table 4). In contrast, that charge shift is absent in the collinear arginine case where only one of the arginine nitrogens is involved in the interaction and little polarization is propagated through the rest of the guanido group.

In the coplanar case, the guanido group is better able to react to the phosphate moiety. This is most clear in the doubly charged species where both hydrogens and nitrogens involved experience a charge shift of roughtly 0.1e and 0.18e respectively. In addition, the central carbon also experiences a shift of ~ 0.07e. Lysine's positive charge appears too densely focused on its single amino group and can only accommodate changes in the amine nitrogen (~0.1e) and the adjacent carbon (~ -0.08e). Given that there are 2 pairs of nitrogens and

hydrogens from the guanido group involved in the co-planar interaction, the shift in charge density is clearly superior to the corresponding shift in the lysine amino group. Thus, polarizability appears to play a role in the preference of phosphorylated residues for arginine over lysine. Determining the exact contribution will require further research.

Effects of the phosphate protonation state

The energy of interaction data show a pronounced effect due to the protonation state of the phosphate moiety. The doubly charged phosphate ion interaction energies appear 2-4 kcal/mol more stable than those of its protonated counterpart. Polarization, assessed by induced charge differences at the energy minimum, only reflects this effect for the phospho-aspartate analog; interactions of the phosphoserine show large charge differences (~0.1e or greater) for both protonation states. Thus, it is unlikely polarization plays a role in this effect.

Oddly, induced charge shifts sometimes seemed inconsistent. For example, the phosphate phosphorous atom in the (serine)biphoshpate-arginine collinear interaction undergoes a 0.17e shift, however, in the phosphoserine-arginine collinear approach the shift is significantly smaller (0.02e). The induced charge differences of the arginine involved seem to complement these changes, that is, they seem to compensate by being smaller in the latter case than in the former (0.10e vs 0.15e for the guanido nitrogen involved). This may reflect limitations

of the fitting algorithm rather than qualitative differences in the nature of the electron densities fitted.

When can a carboxylate ion substitute for phosphorylation??

Some experiments verify a phosphorylation site by introducing a carboxylic acid group at the suspect residue. This approach does not always work; the present calculations show the carboxylic acid group interactions are 2-4 kcal/mol less stable than those of the doubly charged phosphate. However, the preference for the phosphate group is diminished (below 1 kcal/mol) in the protonated species and, in some geometries, even reversed (co-planar approach with arginine). The data suggests a carboxylic acid can mimic the energetics of the biphosphate ion. This could be useful in determining *a priori*, by testing for activity of the phosphorylated protein at lower pH, if mutation to glutamic or aspartic acid will rescue activity. Alternatively, this test may be used *post facto* to better interpret a null result upon mutation to glutamatic or aspartic acid. Clearly more calculations will be needed to establish the validity of such a "test".

Conclusion

The quantum calculations performed here explore the role of polarizability in salt bridge interactions with phosphorylated residues. They suggest that: 1) the guanido side chain is better able to adjust its electron density, in the interactions

examined, than a lysine amino group, 2) the doubly charged phosphate species makes stronger hydrogen bonds than its protonated counterpart and 3) a carboxyl group appears to have interaction stabilities very close to those of biphosphate species. The latter has applications in the interpretation of mutation experiments seeking to identify phosphorylation sites. Polarizability, assessed by shifts in fitted ESP charges, clearly plays an important role in the interactions of phosphorylated residues.

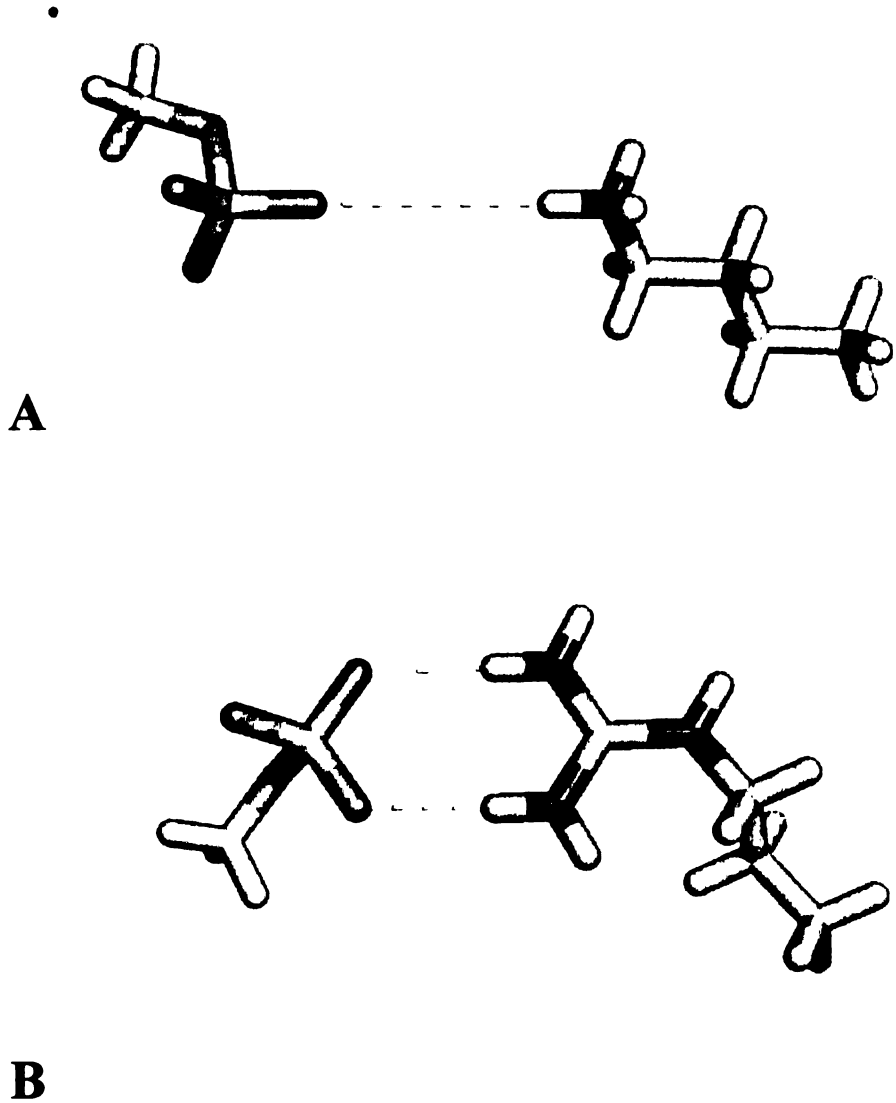


Figure 1. Geometries of interaction for *ab initio* calculations. Figure A displays the "linear" approach, while B displays the "planar" approach.

a) EACT

```
      HA1   HB1  OG2
       |     |    |
HA2—CA - CB - CG - OG1
       |     |
      HA3   HB2
```

b) Methyl Phosphate

```
      HD2          OP1
       |            |
HD1 - CD - OD1 - P - OP2
       |            |
      HD3          OP3
```

c) Methyl biphosphate

```
      HD2          OP1—HO1
       |            |
HD1 - CD - OD1 - P - OP2
       |            |
      HD3          OP3
```

d) Acetophosphate

```
      HB1                O1P
       |                  |
HB1—CB - CG - OD1 - P - O2P
       |     |            |
      HB3   OD2          O3P
```

e) Aceto-biphosphate

```
      HB1                 O1P
       |                   |
HB1—CB – CG – OD1 – P – O2P
       |    |              |
      HB3  OD2            O3P – HO1
```

f) 1-amino butane

```
      HZ3  HE1  HD1   HG1 HB1
       |    |    |     |    |
HZ2 – NZ – CE –  CD –  CG – CB – HB2
       |    |    |     |    |
      HZ1  HE2  HD2   HG2  HB3
```

g) 1-guanido propane

```
      H12
       |
H11—NH1  HE   HD1  HG1   HB1
       |    |    |     |    |
      CZ –  NE – CD – CG – CB – HB2
       |         |     |    |
H22—NH2        HD2  HG2  HB3
       |
      H21
```

Figure 2. Connectivity diagrams for all of the molecules modeled. Atom names, used in the quantum calculation are explicitly listed in this figure. Note that bond order is not included in this figure.

Salt bridges , threshold 4.5 A				Number found within 15 A		
	Arg	Lys	His	Arg	Lys	His
pTyr	27	3	0	64	47	0
pSer	14	4	7	46	35	27
pThr	19	6	0	50	23	0
Total	60	13	7	160	105	27

Table 1. Phosphorylated residue salt bridge count results from survey of PDB structures.

Molecules involved	Geometry	Atoms used to measure distances
Glu – Lys	Collinear	CG (Glu), NZ (Lys)
Glu – Lys	Planar	CG (Glu), NZ (Lys)
Glu – Arg	Collinear	CG (Glu), NH2 (Arg)
Glu – Arg	Planar	CG (Glu), CZ (Arg)
pAsp – Lys	Collinear	P (pAsp), NZ (Lys)
pAsp – Lys	Planar	P (pAsp), NZ (Lys)
pAsp – Arg	Collinear	P (pAsp), NH2 (Arg)
pAsp – Arg	Planar	P (pAsp), CZ (Arg)
pSer – Lys	Collinear	P (pSer), NZ (Lys)
pSer –Lys	Planar	P (pSer), NZ (Lys)
pSer -- Arg	Collinear	P (pSer), NH2 (Arg)
pSer – Arg	Planar	P (pSer), CZ (Arg)

Table 2. List of atoms, for each interaction considered, used to measure the distances in the PMF calculation. Note that the same atoms were used regardless of protonation state.

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

Table 3. Salt-bridge interaction energies in kcal/mol

Phosphoaspartate (-1) --Lysine (Colinear)

	P	OD1	O1P	O2P	O3P	HO1
PMF mimimum	1.6045	-0.5523	-1.0027	-0.9718	-0.8014	0.5114
12.25 A dist	1.492	-0.5396	-0.954	-0.9503	-0.7862	0.509
Difference	0.1125	-0.0127	-0.0487	-0.0215	-0.0152	0.0024

Phosphoaspartate (-1) --Lysine (Planar)

	P	OD1	O1P	O2P	O3P	HO1
PMF mimimum	1.4999	-0.5521	-0.9409	-0.9343	-0.7887	0.5013
12.25 A dist	1.4909	-0.5393	-0.9521	-0.9506	-0.786	0.5085
Difference	0.009	-0.0128	0.0112	0.0163	-0.0027	-0.0072

Phospohaspartate (-1) -- Arginine (Colinear)

	CB	HB1	HB2	HB3	CG	OD1
PMF mimimum	-0.4232	0.1618	0.1641	0.1217	0.8447	-0.5406
12.25 A dist	-0.4133	0.159	0.163	0.1172	0.8564	-0.5388
Difference	-0.0099	0.0028	0.0011	0.0045	-0.0117	-0.0018

Phospohaspartate (-1) -- Arginine (Planar)

	P	OD1	O1P	O2P	O3P	HO1
PMF mimimum	1.4642	-0.5346	-0.9378	-0.9043	-0.7583	0.4913
12.25 A dist	1.4497	-0.5383	-0.9312	-0.927	-0.7765	0.4986
Difference	0.0145	0.0037	-0.0066	0.0227	0.0182	-0.0073

Phosphoaspartate (-2) --Lysine (Colinear)

	P	OD1	O1P	O2P	O3P	CG
PMF mimimum	1.4854	-0.592	-0.9647	-1.0142	-1.0137	0.9496
12.25 A dist	1.4643	-0.5861	-1.0153	-1.0199	-1.0195	0.9366
Difference	0.0211	-0.0059	0.0506	0.0057	0.0058	0.013

Phosphoaspartate (-2) --Lysine (Planar)

	P	OD1	O1P	O2P	O3P	CG
PMF mimimum	1.4208	-0.5666	-0.974	-0.9676	-1.0092	0.9205
12.25 A dist	1.4618	-0.5876	-1.0135	-1.0188	-1.0193	0.9374
Difference	-0.041	0.021	0.0395	0.0512	0.0101	-0.0169

Phosphoaspartate (-2) -- Arginine (Colinear)

	P	OD1	O1P	O2P	O3P	CG
PMF mimimum	1.5024	-0.5938	-1.0025	-1.0232	-1.0232	0.9432
12.25 A dist	1.4686	-0.5863	-1.0167	-1.0214	-1.0211	0.9371
Difference	0.0338	-0.0075	0.0142	-0.0018	-0.0021	0.0061

Phosphoaspartate (-2) -- Arginine (Planar)

	CB	HB1	HB2	HB3	CG	OD1
PMF mimimum	-0.466	0.1687	0.1789	0.0956	0.932	-0.5845
12.25 A dist	-0.4607	0.1714	0.1731	0.0896	0.9386	-0.5883
Difference	-0.0053	-0.0027	0.0058	0.006	-0.0066	0.0038

Phosphoaspartate (-1) --Lysine (Colinear)

	CG	OD2	CB	HB1	HB2	HB3
PMF mimimum	0.8615	-0.6675	-0.4324	0.1668	0.1667	0.1289
12.25 A dist	0.871	-0.6728	-0.4301	0.1657	0.1691	0.1258
Difference	-0.0095	0.0053	-0.0023	0.0011	-0.0024	0.0031

Phosphoaspartate (-1) --Lysine (Planar)

	CG	OD2	CB	HB1	HB2	HB3
PMF mimimum	0.8796	-0.6756	-0.4361	0.1609	0.1693	0.1287
12.25 A dist	0.8701	-0.6721	-0.4327	0.1662	0.1694	0.1271
Difference	0.0095	-0.0035	-0.0034	-0.0053	-1E-04	0.0016

Phospohaspartate (-1) -- Arginine (Colinear)

	OD2	P	O1P	O2P	O3P	HO1
PMF mimimum	-0.6519	1.5258	-0.9614	-0.9408	-0.7822	0.4997
12.25 A dist	-0.6589	1.4522	-0.9321	-0.9275	-0.7767	0.4982
Difference	0.007	0.0736	-0.0293	-0.0133	-0.0055	0.0015

Phospohaspartate (-1) -- Arginine (Planar)

	CG	OD2	CB	HB1	HB2	HB3
PMF mimimum	0.8537	-0.6498	-0.43	0.1606	0.1656	0.1278
12.25 A dist	0.8557	-0.658	-0.4156	0.1596	0.1631	0.1184
Difference	-0.002	0.0082	-0.0144	0.001	0.0025	0.0094

Phosphoaspartate (-2) --Lysine (Colinear)

	OD2	CB	HB1	HB2	HB3	NZ
PMF mimimum	-0.734	-0.4835	0.1776	0.1772	0.0987	-0.3843
12.25 A dist	-0.7342	-0.4608	0.1722	0.1721	0.0908	-0.4787
Difference	0.0002	-0.0227	0.0054	0.0051	0.0079	0.0944

Phosphoaspartate (-2) --Lysine (Planar)

	OD2	CB	HB1	HB2	HB3	NZ
PMF mimimum	-0.7273	-0.447	0.1567	0.1753	0.093	-0.3209
12.25 A dist	-0.7335	-0.4627	0.172	0.174	0.0905	-0.4783
Difference	0.0062	0.0157	-0.0153	0.0013	0.0025	0.1574

Phosphoaspartate (-2) -- Arginine (Colinear)

	OD2	CB	HB1	HB2	HB3	NH2
PMF mimimum	-0.7313	-0.4862	0.1794	0.1792	0.0975	-1.067
12.25 A dist	-0.7352	-0.4578	0.1713	0.1712	0.0895	-1.0379
Difference	0.0039	-0.0284	0.0081	0.008	0.008	-0.0291

Phosphoaspartate (-2) -- Arginine (Planar)

	OD2	P	O1P	O2P	O3P	CB
PMF mimimum	-0.7265	1.4708	-1.0152	-1.0029	-1.004	-0.3278
12.25 A dist	-0.7347	1.4658	-1.0153	-1.0198	-1.0204	-0.3305
Difference	0.0082	0.005	1E-04	0.0169	0.0164	0.0027

Phosphoaspartate (-1) --Lysine (Colinear)

	NZ	HZ1	HZ2	HZ3	CE	HE1
PMF mimimum	-0.4884	0.4313	0.3668	0.3583	0.2488	0.0078
12.25 A dist	-0.4781	0.3921	0.3761	0.376	0.2421	0.0269
Difference	-0.0103	0.0392	-0.0093	-0.0177	0.0067	-0.0191

Phosphoaspartate (-1) --Lysine (Planar)

	NZ	HZ1	HZ2	HZ3	CE	HE1
PMF mimimum	-0.4993	0.3805	0.3693	0.3981	0.2208	0.0291
12.25 A dist	-0.4785	0.3908	0.3755	0.3774	0.2431	0.027
Difference	-0.0208	-0.0103	-0.0062	0.0207	-0.0223	0.0021

Phospohaspartate (-1) -- Arginine (Colinear)

	CB	HB1	HB2	HB3	CG	HG1
PMF mimimum	-0.3436	0.0912	0.069	0.0705	0.2185	-0.035
12.25 A dist	-0.3343	0.09	0.0675	0.069	0.2144	-0.0364
Difference	-0.0093	0.0012	0.0015	0.0015	0.0041	0.0014

Phospohaspartate (-1) -- Arginine (Planar)

	NH2	H21	H22	CZ	NH1	H11
PMF mimimum	-1.1549	0.57	0.5505	1.0505	-1.2879	0.5574
12.25 A dist	-1.0344	0.5529	0.4995	0.9844	-1.0995	0.5383
Difference	-0.1205	0.0171	0.051	0.0661	-0.1884	0.0191

Phosphoaspartate (-2) --Lysine (Colinear)

	HZ1	HZ2	HZ3	CE	HE1	HE2
PMF mimimum	0.3656	0.3186	0.3189	0.2848	0.0035	0.0039
12.25 A dist	0.3909	0.3753	0.3757	0.248	0.0265	0.0263
Difference	-0.0253	-0.0567	-0.0568	0.0368	-0.023	-0.0224

Phosphoaspartate (-2) --Lysine (Planar)

	HZ1	HZ2	HZ3	CE	HE1	HE2
PMF mimimum	0.3437	0.331	0.3109	0.1765	0.0275	0.0249
12.25 A dist	0.3901	0.3746	0.3761	0.2496	0.0261	0.026
Difference	-0.0464	-0.0436	-0.0652	-0.0731	0.0014	-0.0011

Phosphoaspartate (-2) -- Arginine (Colinear)

	H21	H22	CZ	NH1	H11	H12
PMF mimimum	0.5341	0.5443	0.9788	-1.1063	0.5331	0.5233
12.25 A dist	0.5437	0.5124	0.9848	-1.1027	0.5331	0.5391
Difference	-0.0096	0.0319	-0.006	-0.0036	0	-0.0158

Phosphoaspartate (-2) -- Arginine (Planar)

	HB1	HB2	HB3	CG	HG1	HG2
PMF mimimum	0.0863	0.0638	0.0665	0.2264	-0.0428	-0.0156
12.25 A dist	0.0902	0.0657	0.0681	0.2169	-0.0403	-0.0126
Difference	-0.0039	-0.0019	-0.0016	0.0095	-0.0025	-0.003

Phosphoaspartate (-1) --Lysine (Colinear)

	HE2	CD	HD1	HD2	CG	HG1
PMF mimimum	0.0237	-0.0859	0.0343	0.0337	0.2043	-0.0298
12.25 A dist	0.0271	-0.1127	0.043	0.0426	0.2086	-0.0323
Difference	-0.0034	0.0268	-0.0087	-0.0089	-0.0043	0.0025

Phosphoaspartate (-1) --Lysine (Planar)

	HE2	CD	HD1	HD2	CG	HG1
PMF mimimum	0.0435	-0.1132	0.0442	0.0512	0.1936	-0.0251
12.25 A dist	0.0275	-0.1165	0.0436	0.0443	0.2078	-0.0316
Difference	0.016	0.0033	0.0006	0.0069	-0.0142	0.0065

Phospohaspartate (-1) -- Arginine (Colinear)

	HG2	CD	HD1	HD2	NE	HE
PMF mimimum	-0.0151	0.2418	0.0718	0.0173	-0.8616	0.4904
12.25 A dist	-0.013	0.2273	0.0781	0.0181	-0.8429	0.4907
Difference	-0.0021	0.0145	-0.0063	-0.0008	-0.0187	-0.0003

Phospohaspartate (-1) -- Arginine (Planar)

	H12	NE	HE	CD	HD1	HD2
PMF mimimum	0.6723	-0.8354	0.4782	0.1804	0.0856	0.0251
12.25 A dist	0.5356	-0.836	0.4855	0.2126	0.082	0.023
Difference	0.1367	0.0006	-0.0073	-0.0322	0.0036	0.0021

Phosphoaspartate (-2) --Lysine (Colinear)

	CD	HD1	HD2	CG	HG1	HG2
PMF mimimum	-0.1432	0.0391	0.039	0.2306	-0.0395	-0.0398
12.25 A dist	-0.1166	0.0435	0.0434	0.2071	-0.032	-0.0322
Difference	-0.0266	-0.0044	-0.0044	0.0235	-0.0075	-0.0076

Phosphoaspartate (-2) --Lysine (Planar)

	CD	HD1	HD2	CG	HG1	HG2
PMF mimimum	-0.1232	0.0443	0.0438	0.2229	-0.0339	-0.033
12.25 A dist	-0.1187	0.0433	0.0448	0.2076	-0.0319	-0.0325
Difference	-0.0045	0.001	-0.001	0.0153	-0.002	-0.0005

Phosphoaspartate (-2) -- Arginine (Colinear)

	NE	HE	CD	HD1	HD2	CG
PMF mimimum	-0.8425	0.485	0.2268	0.0741	0.0175	0.2227
12.25 A dist	-0.8433	0.4906	0.2292	0.0769	0.0175	0.2151
Difference	0.0008	-0.0056	-0.0024	-0.0028	0	0.0076

Phosphoaspartate (-2) -- Arginine (Planar)

	CD	HD1	HD2	NE	HE	CZ
PMF mimimum	0.1913	0.0803	0.0237	-0.8574	0.4795	1.0822
12.25 A dist	0.2139	0.0814	0.023	-0.8398	0.4866	0.9871
Difference	-0.0226	-0.0011	0.0007	-0.0176	-0.0071	0.0951

Phosphoaspartate (-1) --Lysine (Colinear)

	HG2	CB	HB1	HB2	HB3
PMF mimimum	-0.0344	-0.2505	0.0604	0.0544	0.0532
12.25 A dist	-0.0325	-0.2419	0.0585	0.0524	0.0524
Difference	-0.0019	-0.0086	0.0019	0.002	0.0008

Phosphoaspartate (-1) --Lysine (Planar)

	HG2	CB	HB1	HB2	HB3
PMF mimimum	-0.0295	-0.2429	0.0605	0.0547	0.0518
12.25 A dist	-0.032	-0.2443	0.0596	0.0533	0.053
Difference	0.0025	0.0014	0.0009	0.0014	-0.0012

Phospohaspartate (-1) -- Arginine (Colinear)

	CZ	NH1	H11	H12	NH2	H21
PMF mimimum	0.9953	-1.1037	0.5322	0.5232	-1.0747	0.5298
12.25 A dist	0.9865	-1.1045	0.5336	0.5396	-1.0398	0.5442
Difference	0.0088	0.0008	-0.0014	-0.0164	-0.0349	-0.0144

Phospohaspartate (-1) -- Arginine (Planar)

	CG	HG1	HG2	CB	HB1	HB2
PMF mimimum	0.2212	-0.0403	-0.0128	-0.3313	0.0879	0.0663
12.25 A dist	0.2144	-0.0387	-0.0116	-0.3343	0.091	0.0672
Difference	0.0068	-0.0016	-0.0012	0.003	-0.0031	-0.0009

Phosphoaspartate (-2) --Lysine (Colinear)

	CB	HB1	HB2	HB3
PMF mimimum	-0.2403	0.0542	0.051	0.051
12.25 A dist	-0.2398	0.0584	0.0519	0.0519
Difference	-0.0005	-0.0042	-0.0009	-0.0009

Phosphoaspartate (-2) --Lysine (Planar)

	CB	HB1	HB2	HB3
PMF mimimum	-0.2648	0.0615	0.0563	0.0577
12.25 A dist	-0.2402	0.0587	0.0521	0.0521
Difference	-0.0246	0.0028	0.0042	0.0056

Phosphoaspartate (-2) -- Arginine (Colinear)

	HG1	HG2	CB	HB1	HB2	HB3
PMF mimimum	-0.0376	-0.0169	-0.3375	0.0897	0.0672	0.0694
12.25 A dist	-0.0369	-0.0133	-0.3358	0.0907	0.0678	0.0694
Difference	-0.0007	-0.0036	-0.0017	-0.001	-0.0006	0

Phosphoaspartate (-2) -- Arginine (Planar)

	NH1	H11	H12	NH2	H21	H22
PMF mimimum	-1.318	0.551	0.695	-1.2075	0.5653	0.6106
12.25 A dist	-1.1008	0.5357	0.5378	-1.032	0.549	0.501
Difference	-0.2172	0.0153	0.1572	-0.1755	0.0163	0.1096

Acetate --Lysine (Colinear)

	NZ	HZ1	HZ2	HZ3	CE	HE1
PMF mimimum	-0.5558	0.4883	0.3699	0.3807	0.2314	0.0268
12.25 A dist	-0.4849	0.3952	0.3779	0.3778	0.2468	0.0235
Difference	-0.0709	0.0931	-0.008	0.0029	-0.0154	0.0033

Acetate--Lysine (Planar)

	CG	OD1	OD2	CB	HB1	HB2
PMF mimimum	0.9	-0.9494	-0.8591	-0.0015	-0.0048	0.0069
12.25 A dist	0.8433	-0.9311	-0.9131	0.1484	-0.0509	-0.0365
Difference	0.0567	-0.0183	0.054	-0.1499	0.0461	0.0434

Acetate-- Arginine (Colinear)

	H21	H22	NH2	CZ	H11	H12
PMF mimimum	0.5327	0.5764	-1.0895	1.0135	0.5325	0.5305
12.25 A dist	0.5577	0.5054	-1.0504	0.9916	0.5369	0.542
Difference	-0.025	0.071	-0.0391	0.0219	-0.0044	-0.0115

Acetate -- Arginine (Planar)

	CG	OD1	OD2	CB	HB1	HB2
PMF mimimum	0.8614	-0.9689	-0.8848	0.1305	-0.0297	-0.0289
12.25 A dist	0.8064	-0.9265	-0.8854	0.179	-0.0595	-0.0459
Difference	0.055	-0.0424	0.0006	-0.0485	0.0298	0.017

Phosphoserine(-1) -- Lysine (Colinear)

	P	OD1	OP1	OP2	OP3	HO1
PMF mimimum	1.6254	-0.5565	-1.0742	-0.9826	-0.785	0.4922
12.25 A dist	1.4531	-0.556	-0.9661	-0.9525	-0.7466	0.479
Difference	0.1723	-0.0005	-0.1081	-0.0301	-0.0384	0.0132

Phosphoserine (-1) --Lysine (Planar)

	P	OD1	OP1	OP2	OP3	HO1
PMF mimimum	1.4437	-0.56	-0.9451	-0.9154	-0.7481	0.4726
12.25 A dist	1.4541	-0.5567	-0.9655	-0.9521	-0.7472	0.4788
Difference	-0.0104	-0.0033	0.0204	0.0367	-0.0009	-0.0062

Phosphoserine (-1) --Arginine (Colinear)

	P	OD1	OP1	OP2	OP3	HO1
PMF mimimum	1.535	-0.5421	-1.0177	-0.9497	-0.7626	0.4743
12.25 A dist	1.4209	-0.5468	-0.9458	-0.9331	-0.7411	0.4673
Difference	0.1141	0.0047	-0.0719	-0.0166	-0.0215	0.007

Phosphoserine (-1) -- Arginine (Planar)

	P	OD1	OP1	OP2	OP3	HO1
PMF mimimum	1.4067	-0.5306	-0.9537	-0.8774	-0.7412	0.4764
12.25 A dist	1.4184	-0.5465	-0.9451	-0.9316	-0.7415	0.468
Difference	-0.0117	0.0159	-0.0086	0.0542	0.0003	0.0084

Acetate --Lysine (Colinear)

	HE2	CD	HD1	HD2	CG	HG1
PMF mimimum	0.0015	-0.0247	0.0204	0.0197	0.1895	-0.0337
12.25 A dist	0.0252	-0.1024	0.0379	0.0393	0.2085	-0.0333
Difference	-0.0237	0.0777	-0.0175	-0.0196	-0.019	-0.0004

Acetate--Lysine (Planar)

	CA	HA1	HA2	HA3	NZ	HZ1
PMF mimimum	-0.0674	0.0202	-0.0044	0.0052	-0.4631	0.4291
12.25 A dist	-0.0657	-0.0002	-0.0018	0.008	-0.4734	0.39
Difference	-0.0017	0.0204	-0.0026	-0.0028	0.0103	0.0391

Acetate-- Arginine (Colinear)

	NH1	HE	NE	CD	HD1	HD2
PMF mimimum	-1.1087	0.4941	-0.8866	0.2688	0.0627	0.011
12.25 A dist	-1.1086	0.489	-0.8441	0.2316	0.0752	0.0155
Difference	-1E-04	0.0051	-0.0425	0.0372	-0.0125	-0.0045

Acetate -- Arginine (Planar)

	CA	HA1	HA2	HA3	H21	H22
PMF mimimum	-0.1058	0.014	0.01	0.0142	0.577	0.636
12.25 A dist	-0.0703	-0.0046	-0.0016	0.0079	0.553	0.5009
Difference	-0.0355	0.0186	0.0116	0.0063	0.024	0.1351

Phosphoserine(-1) -- Lysine (Colinear)

	CD	HD1	HD2	HD3	NZ	HZ1
PMF mimimum	0.0269	0.0921	0.0753	0.0743	-0.5576	0.4935
12.25 A dist	0.1058	0.0685	0.059	0.056	-0.4764	0.392
Difference	-0.0789	0.0236	0.0163	0.0183	-0.0812	0.1015

Phosphoserine (-1) --Lysine (Planar)

	CD	HD1	HD2	HD3	NZ	HZ1
PMF mimimum	0.1088	0.0734	0.0483	0.0579	-0.3589	0.335
12.25 A dist	0.1022	0.0706	0.0587	0.0574	-0.4719	0.3883
Difference	0.0066	0.0028	-0.0104	0.0005	0.113	-0.0533

Phosphoserine (-1) --Arginine (Colinear)

	CD	HD1	HD2	HD3	H21	H22
PMF mimimum	0.0502	0.0785	0.0658	0.0677	0.5353	0.607
12.25 A dist	0.1092	0.0612	0.0547	0.0526	0.5433	0.5133
Difference	-0.059	0.0173	0.0111	0.0151	-0.008	0.0937

Phosphoserine (-1) -- Arginine (Planar)

	CD	HD1	HD2	HD3	H21	H22
PMF mimimum	0.0822	0.0748	0.0517	0.0657	0.5556	0.5352
12.25 A dist	0.1047	0.0633	0.0555	0.0541	0.5439	0.5103
Difference	-0.0225	0.0115	-0.0038	0.0116	0.0117	0.0249

Acetate --Lysine (Colinear)

	HG2	CB	HB1	HB2	HB3	CG
PMF mimimum	-0.0257	-0.2578	0.0618	0.0547	0.0552	0.9938
12.25 A dist	-0.0337	-0.2358	0.0556	0.0522	0.0499	0.8393
Difference	0.008	-0.022	0.0062	0.0025	0.0053	0.1545

Acetate--Lysine (Planar)

	HZ2	HZ3	CE	HE1	HE2	CD
PMF mimimum	0.341	0.3615	0.1226	0.0424	0.0416	-0.0257
12.25 A dist	0.3739	0.3753	0.2444	0.026	0.0261	-0.1132
Difference	-0.0329	-0.0138	-0.1218	0.0164	0.0155	0.0875

Acetate-- Arginine (Colinear)

	CG	HG1	HG2	CB	HB1	HB2
PMF mimimum	0.2286	-0.0357	-0.0221	-0.3357	0.0872	0.0653
12.25 A dist	0.2288	-0.0411	-0.0179	-0.3308	0.0875	0.0659
Difference	-0.0002	0.0054	-0.0042	-0.0049	-0.0003	-0.0006

Acetate -- Arginine (Planar)

	NH2	CZ	H11	H12	NH1	HE
PMF mimimum	-1.2292	1.0895	0.5545	0.6291	-1.2631	0.4871
12.25 A dist	-1.0394	0.9925	0.5348	0.5371	-1.0996	0.4918
Difference	-0.1898	0.097	0.0197	0.092	-0.1635	-0.0047

Phosphoserine(-1) -- Lysine (Colinear)

	HZ2	HZ3	CE	HE1	HE2	CD
PMF mimimum	0.362	0.3746	0.2969	0.0051	0.0043	-0.0866
12.25 A dist	0.3743	0.3754	0.2457	0.0256	0.026	-0.1121
Difference	-0.0123	-0.0008	0.0512	-0.0205	-0.0217	0.0255

Phosphoserine (-1) --Lysine (Planar)

	HZ2	HZ3	CE	HE1	HE2	CD
PMF mimimum	0.3485	0.3491	0.1638	0.0379	0.0475	-0.1331
12.25 A dist	0.3738	0.375	0.2448	0.026	0.0265	-0.1156
Difference	-0.0253	-0.0259	-0.081	0.0119	0.021	-0.0175

Phosphoserine (-1) --Arginine (Colinear)

	NH2	CZ	H11	H12	NH1	HE
PMF mimimum	-1.1111	1.005	0.5314	0.5278	-1.1101	0.4887
12.25 A dist	-1.039	0.9854	0.5326	0.5402	-1.1041	0.4907
Difference	-0.0721	0.0196	-0.0012	-0.0124	-0.006	-0.002

Phosphoserine (-1) -- Arginine (Planar)

	NH2	CZ	H11	H12	NH1	HE
PMF mimimum	-1.1371	1.0557	0.5501	0.6768	-1.2851	0.4882
12.25 A dist	-1.0372	0.9861	0.5327	0.5385	-1.1018	0.4905
Difference	-0.0999	0.0696	0.0174	0.1383	-0.1833	-0.0023

Acetate --Lysine (Colinear)

	OD1	OD2	CB	HB1	HB2	CA
PMF mimimum	-1.0411	-0.9241	0.0507	-0.0004	-0.0121	-0.1549
12.25 A dist	-0.9305	-0.9136	0.1621	-0.0538	-0.0404	-0.0728
Difference	-0.1106	-0.0105	-0.1114	0.0534	0.0283	-0.0821

Acetate--Lysine (Planar)

	HD1	HD2	CG	HG1	HG2	CB
PMF mimimum	0.0297	0.0303	0.1864	-0.0256	-0.0274	-0.2664
12.25 A dist	0.042	0.0428	0.2131	-0.0339	-0.0343	-0.2376
Difference	-0.0123	-0.0125	-0.0267	0.0083	0.0069	-0.0288

Acetate-- Arginine (Colinear)

	HB3	CG	OD1	OD2	CB	HB1
PMF mimimum	0.0666	0.9613	-1.0052	-0.9196	0.0685	-0.0177
12.25 A dist	0.0665	0.8412	-0.9323	-0.9152	0.1663	-0.0554
Difference	0.0001	0.1201	-0.0729	-0.0044	-0.0978	0.0377

Acetate -- Arginine (Planar)

	NE	CD	HD1	HD2	CG	HG1
PMF mimimum	-0.8762	0.2501	0.0691	0.0127	0.2033	-0.0265
12.25 A dist	-0.8583	0.2432	0.0742	0.0144	0.2178	-0.0368
Difference	-0.0179	0.0069	-0.0051	-0.0017	-0.0145	0.0103

Phosphoserine(-1) -- Lysine (Colinear)

	HD1	HD2	CG	HG1	HG2	CB
PMF mimimum	0.032	0.0375	0.1878	-0.0302	-0.0304	-0.2298
12.25 A dist	0.042	0.0422	0.2117	-0.0338	-0.0342	-0.2352
Difference	-0.01	-0.0047	-0.0239	0.0036	0.0038	0.0054

Phosphoserine (-1) --Lysine (Planar)

	HD1	HD2	CG	HG1	HG2	CB
PMF mimimum	0.0518	0.0473	0.2248	-0.0315	-0.0342	-0.2533
12.25 A dist	0.0429	0.0436	0.2106	-0.0329	-0.0333	-0.2387
Difference	0.0089	0.0037	0.0142	0.0014	-0.0009	-0.0146

Phosphoserine (-1) --Arginine (Colinear)

	NE	CD	HD1	HD2	CG	HG1
PMF mimimum	-0.8564	0.2339	0.0725	0.0201	0.2243	-0.0401
12.25 A dist	-0.8422	0.2287	0.0762	0.0168	0.2235	-0.0388
Difference	-0.0142	0.0052	-0.0037	0.0033	0.0008	-0.0013

Phosphoserine (-1) -- Arginine (Planar)

	NE	CD	HD1	HD2	CG	HG1
PMF mimimum	-0.8562	0.2101	0.0793	0.0142	0.2168	-0.0338
12.25 A dist	-0.8423	0.2244	0.0784	0.0177	0.2208	-0.0378
Difference	-0.0139	-0.0143	0.0009	-0.0035	-0.004	0.004

Acetate --Lysine (Colinear)

	HA1	HA2	HA3
PMF mimimum	0.0321	0.027	0.0265
12.25 A dist	0.0007	-0.0005	0.0097
Difference	0.0314	0.0275	0.0168

Acetate--Lysine (Planar)

	HB1	HB2	HB3
PMF mimimum	0.0655	0.0553	0.0568
12.25 A dist	0.0568	0.0508	0.0507
Difference	0.0087	0.0045	0.0061

Acetate-- Arginine (Colinear)

	HB2	CA	HA1	HA2	HA3
PMF mimimum	-0.0197	-0.099	0.017	0.0101	0.0123
12.25 A dist	-0.0421	-0.0663	-0.0013	-0.0019	0.006
Difference	0.0224	-0.0327	0.0183	0.012	0.0063

Acetate -- Arginine (Planar)

	HG2	CB	HB1	HB2	HB3
PMF mimimum	-0.0131	-0.3345	0.0881	0.0674	0.0664
12.25 A dist	-0.0142	-0.3368	0.0902	0.0671	0.0685
Difference	0.0011	0.0023	-0.0021	0.0003	-0.0021

Phosphoserine(-1) -- Lysine (Colinear)

	HB1	HB2	HB3
PMF mimimum	0.0558	0.0489	0.0477
12.25 A dist	0.0561	0.0502	0.0501
Difference	-0.0003	-0.0013	-0.0024

Phosphoserine (-1) --Lysine (Planar)

	HB1	HB2	HB3
PMF mimimum	0.0587	0.0546	0.0555
12.25 A dist	0.0573	0.0512	0.0516
Difference	0.0014	0.0034	0.0039

Phosphoserine (-1) --Arginine (Colinear)

	HG2	CB	HB1	HB2	HB3
PMF mimimum	-0.0181	-0.3224	0.0856	0.0627	0.0638
12.25 A dist	-0.0161	-0.3312	0.0879	0.0664	0.067
Difference	-0.002	0.0088	-0.0023	-0.0037	-0.0032

Phosphoserine (-1) -- Arginine (Planar)

	HG2	CB	HB1	HB2	HB3
PMF mimimum	-0.0156	-0.3261	0.0852	0.0652	0.0666
12.25 A dist	-0.0144	-0.3353	0.0897	0.0673	0.0687
Difference	-0.0012	0.0092	-0.0045	-0.0021	-0.0021

Phosphoserine (-2) -- Arginine (Colinear)

	P	OD1	OP1	OP2	OP3	CD
PMF mimimum	1.5893	-0.5925	-1.0625	-1.0608	-1.0592	0.1363
12.25 A dist	1.5688	-0.6009	-1.0745	-1.0599	-1.0601	0.1663
Difference	0.0205	0.0084	0.012	-0.0009	0.0009	-0.03

Phosphoserine (-2) -- Arginine (Planar)

	P	OD1	OP1	OP2	OP3	CD
PMF mimimum	1.5758	-0.6003	-1.0742	-1.0446	-1.0439	0.1573
12.25 A dist	1.5669	-0.5995	-1.0738	-1.0592	-1.0598	0.1615
Difference	0.0089	-0.0008	-0.0004	0.0146	0.0159	-0.0042

Phosphoserine (-2) -- Arginine (Colinear)

	NH2	CZ	H11	H12	NH1	HE
PMF mimimum	-1.0823	0.9992	0.5358	0.5422	-1.1307	0.4863
12.25 A dist	-1.0365	0.9835	0.5324	0.5397	-1.1026	0.4904
Difference	-0.0458	0.0157	0.0034	0.0025	-0.0281	-0.0041

Phosphoserine (-2) -- Arginine (Planar)

	NH2	CZ	H11	H12	NH1	HE
PMF mimimum	-1.1986	1.0729	0.5447	0.6896	-1.3041	0.4858
12.25 A dist	-1.0383	0.9858	0.5322	0.5387	-1.101	0.49
Difference	-0.1603	0.0871	0.0125	0.1509	-0.2031	-0.0042

Phosphoserine (-2) -- Arginine (Colinear)

	HG1	HG2	CB	HB1	HB2	HB3
PMF mimimum	-0.0406	-0.0201	-0.329	0.0862	0.0641	0.0664
12.25 A dist	-0.0382	-0.0157	-0.3316	0.0886	0.0666	0.0674
Difference	-0.0024	-0.0044	0.0026	-0.0024	-0.0025	-0.001

Phosphoserine (-2) -- Arginine (Planar)

	HG1	HG2	CB	HB1	HB2	HB3
PMF mimimum	-0.0314	-0.0169	-0.3246	0.0843	0.0641	0.065
12.25 A dist	-0.0376	-0.0146	-0.3325	0.0893	0.0667	0.068
Difference	0.0062	-0.0023	0.0079	-0.005	-0.0026	-0.003

Phosphoserine (-2) -- Arginine (Colinear)

	HD1	HD2	HD3	H21	H22
PMF mimimum	0.02	0.0318	0.0323	0.5353	0.5543
12.25 A dist	0.013	0.0232	0.0238	0.5426	0.5126
Difference	0.007	0.0086	0.0085	-0.0073	0.0417

Phosphoserine (-2) -- Arginine (Planar)

	HD1	HD2	HD3	H21	H22
PMF mimimum	0.0185	0.0259	0.0312	0.5513	0.6197
12.25 A dist	0.0143	0.0241	0.0252	0.5438	0.5117
Difference	0.0042	0.0018	0.006	0.0075	0.108

Phosphoserine (-2) -- Arginine (Colinear)

	NE	CD	HD1	HD2	CG
PMF mimimum	-0.85	0.2274	0.0727	0.0166	0.2311
12.25 A dist	-0.8421	0.2291	0.0761	0.0166	0.2212
Difference	-0.0079	-0.0017	-0.0034	0	0.0099

Phosphoserine (-2) -- Arginine (Planar)

	NE	CD	HD1	HD2	CG
PMF mimimum	-0.8728	0.2311	0.0714	0.0101	0.2122
12.25 A dist	-0.8421	0.225	0.0779	0.0175	0.2194
Difference	-0.0307	0.0061	-0.0065	-0.0074	-0.0072

Phosphoserine (-2) -- Lysine (Colinear)

	P	OD1	OP1	OP2	OP3	CD
PMF mimimum	1.6231	-0.5935	-1.0798	-1.0665	-1.0631	0.1105
12.25 A dist	1.5604	-0.5993	-1.0718	-1.0577	-1.0574	0.1617
Difference	0.0627	0.0058	-0.008	-0.0088	-0.0057	-0.0512

Phosphoserine (-2) -- Lysine (Planar)

	P	OD1	OP1	OP2	OP3	CD
PMF mimimum	1.5227	-0.5704	-1.0372	-1.017	-1.0469	0.1188
12.25 A dist	1.5629	-0.5983	-1.0721	-1.0582	-1.0586	0.1567
Difference	-0.0402	0.0279	0.0349	0.0412	0.0117	-0.0379

Phosphoserine (-2) -- Lysine (Colinear)

	HD1	HD2	HD3	NZ	HZ1	HZ2
PMF mimimum	0.0303	0.0396	0.0367	-0.4426	0.3969	0.3448
12.25 A dist	0.0148	0.0247	0.0253	-0.4764	0.3905	0.3744
Difference	0.0155	0.0149	0.0114	0.0338	0.0064	-0.0296

Phosphoserine (-2) -- Lysine (Planar)

	HD1	HD2	HD3	NZ	HZ1	HZ2
PMF mimimum	0.026	0.0322	0.0381	-0.3214	0.3458	0.3424
12.25 A dist	0.0163	0.0255	0.0263	-0.4747	0.389	0.374
Difference	0.0097	0.0067	0.0118	0.1533	-0.0432	-0.0316

Phosphoserine (-2) -- Lysine (Colinear)

	HZ3	CE	HE1	HE2	CD	HD1
PMF mimimum	0.3492	0.2615	0.0118	0.0117	-0.0979	0.0336
12.25 A dist	0.3745	0.2502	0.0255	0.0253	-0.1154	0.0428
Difference	-0.0253	0.0113	-0.0137	-0.0136	0.0175	-0.0092

Phosphoserine (-2) -- Lysine (Planar)

	HZ3	CE	HE1	HE2	CD	HD1
PMF mimimum	0.3083	0.1773	0.0301	0.0263	-0.1461	0.0505
12.25 A dist	0.3746	0.2499	0.0256	0.0253	-0.1168	0.0428
Difference	-0.0663	-0.0726	0.0045	0.001	-0.0293	0.0077

Phosphoserine (-2) -- Lysine (Colinear)

	HD2	CG	HG1	HG2	CB	HB1
PMF mimimum	0.035	0.2107	-0.0339	-0.0343	-0.2448	0.0573
12.25 A dist	0.0425	0.2101	-0.0333	-0.0336	-0.2347	0.0564
Difference	-0.0075	0.0006	-0.0006	-0.0007	-0.0101	0.0009

Phosphoserine (-2) -- Lysine (Planar)

	HD2	CG	HG1	HG2	CB	HB1
PMF mimimum	0.0464	0.2359	-0.036	-0.036	-0.2596	0.058
12.25 A dist	0.0437	0.2092	-0.0327	-0.0333	-0.2358	0.0569
Difference	0.0027	0.0267	-0.0033	-0.0027	-0.0238	0.0011

Phosphoserine (-2) -- Lysine (Colinear)

	HB2	HB3
PMF mimimum	0.0519	0.0514
12.25 A dist	0.0502	0.0502
Difference	0.0017	0.0012

Phosphoserine (-2) -- Lysine (Planar)

	HB2	HB3
PMF mimimum	0.0542	0.0571
12.25 A dist	0.0504	0.0509
Difference	0.0038	0.0062

Table 4. Charge differences induced during salt bridge interactions involving phosphorylated residues and carboxylate ions. The atom names corresponds to those in figure 1. Highlighted in yellow are atoms that undergo large charge differences (> 0.1e) , while cells highlighted in light blue include smaller, but significant changes (~0.07e).

References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucl. Acids. Res., 28*(1), 235-242.

Cieplak, P., Cornell, W. D., Bayly, C., & Kollman, P. A. (1995). Application of the multimolecule and multiconformational resp methodology to biopolymers - charge derivation for DNA, rna, and proteins. *Journal of Computational Chemistry, 16*(11), 1357-1377.

Jaguar 5.0, S., L.L.C. (1991-2003). Jaguar (Version 5.0). Portland, OR.

Johnson, L. N., & Lewis, R. J. (2001). Structural basis for control by phosphorylation. *Chem Rev, 101*(8), 2209-2242.

Jorgensen, W. L., & Tiradorives, J. (1988). The opls potential functions for proteins - energy minimizations for crystals of cyclic-peptides and crambin. *Journal Of The American Chemical Society, 110*(6), 1666-1671.

Marten, B., Kim, K., Cortis, C., Friesner, R. A., Murphy, R. B., Ringnalda, M. N., et al. (1996). New model for calculation of solvation free energies: Correction of self-consistent reaction field continuum dielectric theory for short-range hydrogen-bonding effects. *Journal of Physical Chemistry, 100*(28), 11775-11788.

Conclusion

Loop modeling techniques largely focus on finding a single loop conformation given a segment sequence and its protein environment. Here, methodology was proposed to go beyond this limitation and model loops as an ensemble of loop conformations. This is a more appropriate representation of loops given their high flexibility and this flexibility's role in their biological function; applications of this approach verify its usefulness.

In chapter 1 the need to and advantage of modeling loops as conformational ensembles is clear. There, it was possible to, by making use of the conformational selection hypothesis to identify holo-like loop conformations starting from an apo structure. The predictions were remarkably good, usually below a 2 A RMSD from the known holo coordinates. In addition to a geometric evaluation, the resulting structures were tested by docking calculations, which are notoriously sensitive to imperfections in the receptor conformation. Docking into the predicted structures ranked known ligands better than the corresponding docking against the apo structure, sometimes by a significant amount. In addition, the lowest (best) docking scores for known ligands corresponded to structures with the lowest RMSD from the holo coordinates.

Predicting these loop-latching motions is a subset of the more general induced fit problem in docking calculations. The approach developed here is clearly an

efficient alternative to current methods that account for receptor flexibility explicitly on a per ligand basis. This approach could be used for predicting loop motions in the several examples listed in chapter 1, and also, coupled with a more efficient sampling scheme, be generalized to account for other conformational changes during binding.

Chapter 2 focuses on a methodological challenge encountered in chapter 1, namely how to adequately sample transitions involving a large entropic barrier. It became clear, by using 1 and 2 dimensional models, that the replica-exchange method, based on temperature, has limitations in talking this problem. REM depends on high temperature replicas to quickly overcome energy barriers; however, these replicas actually shift towards highly entropic states. Thus, if the transition barrier is largely entropic exchanging with very high temperature replica may not increase the barrier crossing probability. In extreme cases, even exchanging with moderately higher temperatures can be detrimental because the exchange rates at those temperatures may in fact be lower. The increased computational cost of adding more replicas further compounds the loss in efficiency by increasing the largest temperatures used in the replica exchange simulation, which amounts to increasing the total number of replicas. Considering these types of barriers during the developing of new sampling schemes could certainly improve their overall performance in biomolecular simulations.

Chapter 3 dealt with variations of flexibility in antibody CDR loops during affinity maturation. In short, via analysis of molecular dynamics simulations, it was possible to rationalize how mutations during affinity maturation can restrict CDR loops. In several cases mutations of serine to asparagines yielded the formation of salt-bridge triads, that is, the formation of an electrostatic contact involving three different groups. In addition, it was possible to observe the equilibrium between the bound and unbound H3 loop conformation of 7G12g in the absence of ligand. Such an equilibrium is consistent with a conformational selection mechanism for binding in the germline antibody. Also clear from the simulations was that other mechanisms, in addition to CDR rigidification, exist for affinity maturation; 48G7g maturation involves an alternative mechanism.

The results of chapter 3 are somewhat anecdotal, however, motivate further research to validate these findings. If found general, they would be helpful in protein engineering efforts or de novo antibody design where an antibody could be "matured" via mutations that ridigify its CDRs. Establishing the relation between these mutations and binding affinity would be critical in the design, or disruption, of protein-protein interfaces. In addition, capturing holo-like conformations via sampling of CDR loops could potentially offer new candidate structures for target interactions.

In chapter 4 several methodological issues regarding the simulation of phophorylated peptides were explored. The first issue was the suitability of implicit and explicit solvent models to adequately stabilize the phosphate moiety and modulate its hydrogen bonding interactions. The solvent response is coupled to the atomic charges that represent the phosphate ion in our all-atom force field model. Thus, in order to control for a possible poor set of charges, a wide range of charges were tested. The performance of the models was tested via calculation of $J^3_{NH\text{-}C\alpha}$ coupling constants and comparison with experiment. It is clear from the results the use of a distance dependent dielectric is a poor approximation regardless of the charges used. A Generalized Born model with a surface area non-polar correction performed nearly as well as the explicit solvent simulations using TIP3P, TIP4P and SPC/E. A fortuitous finding in this study was the relationship between hydrogen bonding to backbone amides and its effect on the J-coupling constants. The latter explained why the distance dependent dielectric, which did not properly modulate the hydrogen bonding interaction, yielded poor results.

In chapter 5, the role of polarizability in interactions involving phosphorylated residues was explored. These contacts usually involve residues that were far apart before phosphorylation and that were made due to comformational changes upon the introduction of the phosphate moiety. The *ab initio* calculations, with a self-consistent reaction field, made it possible to reach several conclusions. First, an argenine's guanido group more easily polarizes and forms a stronger interaction

with a phosphate moiety than lysines amino group. In terms of modeling these interactions, it is clear some element of polarizability must be included to yield the correct energetics. Second, the biphosphate species (-1) makes weaker salt-bridge interactions than its doubly charged counterpart. However, there appears to be little induced polarization in these cases. Thirdly, the interactions involving a biphosphate ion had similar interaction energies than those involving a carboxylate ion. This suggests a carboxylic acid may substitute for a phosphorylation if, in the relevant interaction, the phosphate ion is protonated. Clearly, this must be further validated, but, if valid, it could be useful in interpreting mutational studies.

UNIVERSITY OF CALIFORNIA
UC San Francisco LIBRARY

7541017
3 1378 00754 1017
UC San Francisco LIBRARY
UNIVERSITY OF CALIFORNIA