Conformation Spaces of Proteins and Implications for Protein Folding

by

David Clifford Sullivan

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Chemistry

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Preface

Having come to UCSF with only vague ideas about how best to spend my time here, I was swayed to theoretical studies on protein structure within months after arrival. For that I thank an excellent physical chemistry course by Professor Dill and the Biophysics department's survey on biophysical chemistry taught by several faculty members. Professor Babbitt provided the first opportunity to dabble in theory and gave me the confidence to use past experience as a source of strength, not limitation. My gratitude to Irwin Kuntz extends on a number of levels. I appreciated the freedom to pursue a course of study perhaps only possible in graduate research, the 'childhood' of my scientific career. However, this freedom is only valuable if properly balanced by advice, direction, and feedback, which Tack likewise provided in kind, amazing me with his ability to critique nearly any idea laid in front of him. The inspiration and provocation to make the initial leap into "measuring conformational space", the dominating topic of my research, was largely generated locally, beginning with seminars on protein dynamics by Peter Kollman. The initial source of encouragement and background knowledge came from the Kuntz group, particularly from discussions with Geoff Skillman and former rotation student Noel Southall. Curiously, I could not suppose at the time that this subject, and related tangents, would consume the majority of my graduate career. I also thank my readers, Professors Ken Dill, Patricia Babbitt, and Irwin Kuntz for reading and critiquing this work.

The dissertation is divided into three chapters with an introduction. Each chapter is in the form of a manuscript submitted for publication. Chapter one has been published (Sullivan and Kuntz, 2001). Chapter two has been accepted for publication. Chapter three has been submitted for publication.

Sullivan DC and Kuntz ID, *Proteins* **42**, 495-511 (2001).

Abstract

The following three chapters characterize structural variance seen across protein conformational ensembles.

In chapter one, two approaches to measuring conformational space are taken. A method called N-cube analysis (NCA) is developed for assigning a gross size to a conformational ensemble. Using NCA, the conformational space size of villin headpiece subdomain's compact unfolded state is calculated to be 10^7 times larger than the native state's. In addition, Euclidean based methods for characterizing the behavior of conformational space surrounding any given structure are developed. In the limit of small displacements, the apparent number of model Euclidean dimensions is found to correspond to the true number of mechanical degrees of freedom for a particular ensemble.

In chapter two we show how protein dynamics of the scale associated with protein folding can be modeled as a diffusive search in a high dimensional space. The size of the space is determined using principles developed in Chapter 1. Our numerical approach permits simple comparison to molecular dynamics. Time dependent displacement in the diffusion model agrees with results from molecular dynamics over many orders of magnitude. We delimit the search space into 'native' and 'unfolded' regions and test different energy biases in protein folding simulations. Our results provide insight into plausible energy landscape models for protein folding.

Chapter three shows how crystal structures of HIV-1 protease, a well-studied drug target with many crystal structures containing different inhibitors and point mutations, cluster by crystal type, with apparent independence of inhibitor type and mutation. The many studies

Table of Contents

List of Tables

List of Figures

Introduction

Internal motion is an essential feature of proteins. Functionally important motions span a large range in terms of time and positional displacement, with particular regions in the range of consequence to different biological problems. Consider, for example, the internal motion of the native state of an enzyme, perhaps one of the smallest functionally important motions (Ansari et al, 1985). The amplitude of low energy loop motions near the active site is of great interest to drug design efforts (Knegtel et al, 1997). These motions elude direct observation at atomic detail. Structural variance across NMR ensembles built from NOE constraints generally reflect lack of data more than observance of motion (MacArthur et al, 1994), though other NMR experiments, such as ^{15}N spin relaxation, show some promise in offering a detailed description of structural fluctuation if combined with other biophysical data (Palmer, 2001). Multiple crystal structures of the same system offer the opportunity to view structural variance at high resolution, however the crystalline environment introduces its own set of issues, which is addressed in this thesis. Contrast native state flexibility with protein folding. The larger refolding motions qualitatively differ from the globular vibrations of the native state since the reference frame for any given chain element is in flux. Structural knowledge of the unfolded state is therefore much less detailed and open to speculation (Ptitsyn, 1995). The next largest structural displacement of biological importance is protein degradation, involving the breaking of covalent bonds, which is outside the scope of this work.

The complexity of protein dynamics motivates separating its modeling into two parts: first, generating an ensemble of conformations that exemplifies some structural variance (i.e. a molecular dynamics trajectory) and, second, analyzing this ensemble. While the generation step seeks to produce a correct answer, which in the ideal case involves comparison to some experiment, the analysis step seeks to produce an informative answer, such as insight into a poorly understood mechanism. This thesis focuses on the latter problem of structural variance analysis.

Implicit to the 'holy grail' of the post-genomic era, prediction of structure from sequence, is consideration of a huge space of alternate, though wrong, set of decoy structures which is analogous to the ensemble of misfolded compact structures visited by a folding protein on its path to the native state. Whether trying to understand protein folding or predict native protein structure from sequence, an understanding of the physical principles that stabilize the native state is crucial to both fields. Methods for characterizing conformational space, a topic this thesis addresses, provide a means for measuring the entropic cost paid by a folding protein and for assessing the challenge decoy conformations pose to structure prediction.

Two approaches to measuring conformational space are explored in Chapter 1. Both assume a Euclidean description of conformational space. First, a global measure of conformational space (N-Cube Analysis) is developed which addresses the problem of assigning a size to the conformational space spanned by a given ensemble of structures. Second, the local behavior of conformational space surrounding any given reference structure is also probed.

Measuring the size of a state's conformational space is one step towards measuring the conformational entropy of that state. Combined with enthalpy calculations and other entropic quantities such as vibrational and solvation terms, a free energy can be calculated. Of course, if all minima can be enumerated, and if anharmonic factors can be safely neglected or corrected for, then the partition function can be solved yielding free energy quantities directly without resorting to adding energetic terms. This is not yet computationally feasible for a protein. A related approach is to estimate the number of conformational minima through a residue-based multiplicative argument as in the classic Levinthal calculation (Levinthal, 1969). Two issues are often lumped together in this respect. First is the assumption that the number of minima scales exponentially with the number of residues. This is a reasonable initial assumption convenient for bookkeeping. The more problematic extension of this idea is that the local shape of the energy landscape (i.e. the inter-minima landscape) is determined by energetic terms local in nature (bond angle, length and most importantly, torsional terms) without consideration to non-bonded terms. This model may be reasonable for some polymer systems where non-bonded interactions are not important, but not for proteins in water . It has been shown that for short polyalanyl chains, local steric effects reduce independence among backbone torsion angles (Pappu et al, 2000). In proteins, correlations between torsion angles should be greater because of constraints arising from compactness (Dill, 1985). The alternative extreme model is that the non-bonded terms exclusively determine the arrangement of minima in the energy landscape, with correlated torsional displacements functioning only to maintain inter-chain contact. In such a model, the molecular surface features of secondary structure elements create and place minima in the energy landscape. Pursuing this model requires a shape-based description of the amino acid chain's surface, which is very complex from the perspective of a modeler. This operational obstacle probably explains why the former model is more frequently developed. The manner in which non-bonded and bonded terms combine to shape the energy landscape is complicated owing to terms being highly correlated and of many dimensions. In summary, attempts at direct monitoring of individual degrees of freedom in order to characterize the global energy landscape in an automated fashion force simplifications that result in an unacceptable loss of information.

Instead of monitoring a protein's many individual degrees of freedom, this work uses global, averaged displacement measures, such as the root-mean-squared displacement (RMSD) of atoms (generally only the C-α atoms). This approach suffers from (1) specific structural features being lost to averaging and (2) since RMSD is a pairwise measure, displacement can only be measured assuming some reference structure. Conformational substate clustering on picosecond time scales observed in two-dimensional (time versus time) RMSD plots of molecular dynamics trajectories (Troyer and Cohen, 1995) provides anecdotal evidence for the usefulness of RMSD in reporting energy landscape features.

One motivation for measuring conformational space of proteins is to connect structure with dynamic events such as protein folding. Chapter 1 culminates with a very simple dynamic model that relates (1) the total number of conformational substates in the compact-unfolded state to (2) the number of conformational substates in the native state with (3) the time scale of substate transitions to arrive at a predicted folding time for a protein based on a random exploration of the conformational substates. The discrepancy between the folding time based on random exploration and the actual experimental folding time (about 2 orders of magnitude difference in the case of villin headpiece subdomain) is offered as a measure of the energetic slope of the "folding funnel". Chapter 2 resumes with the subject of energetic bias in protein folding using the averaged conformational space model developed in Chapter 1. Specifically, the protein is modeled as a point in a high-dimensional hyper-rectangular Euclidean space. Displacement of the point represents changes to the proteins conformation. The relationship between time and displacement is parameterized using all-atom molecular dynamics simulations, resulting in a numerical diffusion model for protein dynamics. The simplicity of the model facilitates connecting the results of short time-scale molecular dynamics simulations with long time-scale protein folding events. Since the molecular dynamics model is parameterized with spectroscopic data from the sub-picosecond time scale (as well as high-level ab initio calculations) and the experimental folding times are on a microsecond-plus time-scale, the diffusion model developed provides the missing theoretical link between two sets of experiments with time-scales differing by

~8 orders of magnitude (10^{-13} sec. versus 10^{-5} sec.). Encouraging to the theorist, we show the connection to be both seamless and reasonable.

Chapter 3 provides a conformational analysis of the native state of liganded HIV-1 protease as revealed by x-ray crystallography. This chapter likewise addresses quantification of conformational variation, however, since the magnitude of the displacements across the native state is much smaller than through the unfolded state, the appropriate analysis differs compared to protein folding studies. The aspiration of this chapter is to relate crystallographically observed structural differences to sequence mutations and the particular ligand bound to the enzyme, both of which vary across the selected set of 63 high-resolution structures (resolution < 2.5 Å). The central finding of the work is that global structure correlates strongly with the crystal type, and thus the space group in which the structure was solved. No clear relationship between sequence, ligand type, and experimental conditions emerges from our analysis. We do find that hypothesized functionally important displacements resulting from mutations are generally smaller in magnitude than space-group associated displacements. If it assumed that cross-crystal type structural variance is small relative to structural variance in solution, then factors not seen in the liganded crystal structure, such as solute assisted loop motions and the dynamics of the less ordered non-liganded state, must be at least equally important to understanding structural perturbations resulting from mutations. Alternatively, crystal packing forces may be inducing structures not highly populated in solution.

Ansari A, Berendzen J, Bowne SF, Frauenfelder H, Iben IET, Sauke TB, Shyamsunder E, Young RD *Proc Natl Acad Sci* **82**, 5000-5004 (1985).

Dill KA, *Biochemistry* **24**, 1501-1509 (1985).

Knegtel RMA, Kuntz ID, Oshiro CM, *J Mol Biol* **266**, 424-440 (1997).

Levinthal C, How to fold graciously. In: Debrunner P, Tsibris JCM, Munck E, editors. Mossbauer Spectroscopy in Biological Systems. Urbana, IL: Univ Illinois Press; 21-24 (1969).

MacArthur MW, Laskowski RA, Thornton JM, *Curr Op Struct Biol* **4**, 731-737 (1994).

Palmer AG, *Annu Rev Biophys Biomol Struct* **30**, 129-155 (2001).

Pappu RV, Srinivasan R, Rose GD, *Proc Natl Acad Sci USA* **97,** 12565-12570 (2000).

Ptitsyn OB, *Curr Opin Struct Biol* **5**, 74-78 (1995).

Troyer JM and Cohen FE, *Proteins* **23**, 97-110 (1995).

Conformation Spaces of Proteins

David C. Sullivan and Irwin D. Kuntz

Department of Pharmaceutical Chemistry
University of California at San Francisco
San Francisco, California   94143-0446

Keywords: protein dynamics, protein folding kinetics, molecular dynamics, lattice model, conformational search

Abbreviations: DRT, discrete-random-torsion; MD, molecular dynamics; NCA, N-Cube analysis; PCA, principal component analysis; RMSD, root-mean-squared deviation;

Abstract

We report a simple method for measuring the accessible conformational space explored by an ensemble of protein structures. The method is useful for diverse ensembles derived from molecular dynamics trajectories, molecular modeling and molecular structure determinations. It can be used to examine a wide range of time scales.

The central tactic we use, which has been previously employed, is to replace the true mechanical degrees of freedom of a molecular system with the conformationally effective degrees of freedom as measured by the root-mean squared Cartesian distances among all pairs of conformations. Each protein conformation is treated as a point in a high dimensional Euclidean space. In this paper, we model this space in a novel way by representing it as an N-dimensional hypercube describable with only two parameters: the number of dimensions and the edge length.

To validate this approach we provide a number of elementary test cases and then use the N-cube method for measuring the size and shape of conformational space covered by molecular dynamics trajectories spanning over 10 orders of magnitude in time. These calculations were performed on a small protein, the villin headpiece subdomain, exploring both the native state and the misfolded/folding regime. Distinct features include single, vibrationally-averaged, substate minima on the 0.1- 1 ps time scale, thermally-averaged conformational states that persist for 1-100 ps and transitions between these local minima on nanosecond time scales. Large-scale refolding modes appear to become uncorrelated on the microsecond time scale. Associated length scales for these events are 0.2 Å for the vibrational minima; 0.5 Å for the conformational minima; and 1-2 Å for the nanosecond events. We find that the conformational space that is dynamically accessible during folding of villin has enough volume for ~10^9 minima of the variety that persist for picoseconds. Molecular dynamics

trajectories of the native protein and experimentally-derived solution ensembles suggest the native state to be composed of ~10^2 of these thermally accessible minima. Thus, based on random exploration of accessible folding space alone, protein folding for a small protein is predicted to be a milliseconds time scale event. This time can be compared with the experimental folding time for villin of 10 - 100 μs. One possible explanation for the 10-100 fold discrepancy is that the slope of the 'folding funnel' increases the rate 1-2 orders of magnitude above random exploration of substates.

Introduction

The size of conformational space has long been of interest to structural biologists, particularly as it relates to the kinetics of protein folding. Levinthal[1,2] is credited with first appreciating the vastness of the search space for a folding protein and for recognizing that a folding mechanism based on random folding requires long and non-biological time scales. Levinthal's paradox becomes less of an issue when energetic biases toward the native state are considered.[3] A useful framework for discussing the energetic landscape is the folding funnel which incorporates such a bias directly .[4]

Estimates of the size of conformational space for a polymer typically begin by assessing the number of states that each residue may assume, and then extrapolating to multi-residue cases. In classical isomeric rotation models, each (non-glycine) peptide has 9 states: three each minima for the ϕ and ψ torsions. The vast majority of these conformations will be energetically disallowed by excluded volume. For proteins with compact globular structures, the number of states of the backbone, per residue, is estimated[5] to be 1.7 still yielding large numbers of states; e.g. 2×10^8 for the 36 residue villin headpiece backbone.

Because of the difficulty of summing over individual states of such complex systems, we approach the problem of measuring conformational space from an alternate perspective. Instead of using a residue-based strategy, we estimate conformational volumes of entire molecules using ensembles of conformations which sample specific protein constructs such as the native state or the molten globule[6]. With conformational 'volumes' for appropriate substates in hand, we can estimate times for randomly finding the native ensemble from among all globular structures. These times, which can be thought of as the "Levinthal" times[1,2] for random exploration of accessible states, can be compared to experimental estimates of the times for folding, to assess the 'slopes' of folding landscapes,[4] that is, the increase in rate over random exploration conferred by the energy landscape. We make extensive use of molecular dynamics (MD) which produces ensembles explicitly referenced in time and temperature. The villin headpiece subdomain, is used as our example for MD calculations. In addition to MD, we look at an NMR ensemble for the villin headpiece subdomain and explore some issues using rotamer-based and lattice-based conformational ensembles. NMR ensembles and lattice-based ensembles differ from ensembles generated by MD in that the conformations are not kinetically connected.

A critical tactic in our overall approach is to relate conformational events to the root-mean-square differences (RMSD) between pairs of molecules in an ensemble.[7] A basic question is how many accessible structures lie within a given RMSD threshold of any arbitrary structure. We explore the pair-wise distribution function p(r), which is defined as the number of structures that lie within a shell at distance r of a reference structure or an integrated version, v(r), which is the number of structures less than a distance r from the reference structure. In this paper, r is the C_α-RMSD after optimal superpositioning[8]. Representing molecular ensembles in this way leads to a high dimensional "conformation space". We assume that structures are distributed in a reasonably uniform way in this space. A simple way to model a bounded conformational space is with an N-dimensional hypercube. The cube edge length corresponds to the average amplitude of the effective modes while the dimensionality of the N-cube (i.e. the value of N) corresponds to the number of high-amplitude conformationally active modes. The edge length is the RMSD between two structures at opposite ends of single modes, averaged over all the effective modes. In this paper we develop this N-cube analysis (NCA) by deriving formulas for the average number of dimensions, N, and the average displacement, A, from the mean and variance of the pair-wise distribution function (see Methods). For the conformational space of a protein ensemble, N only loosely corresponds to the number of mechanical degrees of freedom, because degrees of freedom are not equal in terms of impacting the RMSD. For example, large-scale hinging motions alter the structure more than covalent bond stretching. More properly, this parameter corresponds to the number of degrees of freedom that are effective in producing significant Cartesian coordinate displacements.

It is useful to compare NCA with its more sophisticated cousin, principal component analysis (PCA). PCA has been used to distill out the "essential subspace" of conformational space sampled by MD trajectories and has demonstrated that a small set of anharmonic and multiple-minima modes captures the majority of fluctuations for an MD simulation.[9-12] The remaining modes are small scale, near harmonic fluctuations. However, the significance of any single PCA mode is unclear. Balsera et al.[12] demonstrated that the largest PCA modes derived from two 235 ps MD trajectories are not stable between trajectories. Dimensionally-reduced phase space projections of a 150 ps MD trajectory of myoglobin by Clarage et al.[13] show that the largest PCA modes are poorly sampled. In 500 ps, the trajectory of Clarage et al.[13] still does not close back on itself, but simply explores more phase space, which they point out agrees with Fig. 9a in Amadei et al.[9]. The simple hypercube analysis developed in this paper allows us to focus on the size and shape of the accessible conformational space and ignore the nature of particular modes, which appear to be of debatable significance. More importantly, our simplification of space to two parameters greatly aids in analysis between multiple ensembles. We point out that neither NCA nor PCA results can have a rigorous relationship to equilibrium statistical mechanics since neither method directly counts conformational states, but both give insight into the conformational events that characterize particular macromolecular ensembles.

Methods

Derivation of N-Cube properties from Distributions of RMS Distances

We wish to find a hypercube of dimension N and edge length A such that the distribution of squared distances between two points sampled uniformly from within the hypercube will have the same mean M and variance V as our observed squared RMSDs. Begin by noting that the squared distance between two points is equal to the sum of the squared distances in each coordinate, and that sampling points uniformly means sampling each coordinate independently from a uniform distribution on [0, A]. This in turn is equivalent to sampling from a uniform distribution on [0,1] and then multiplying by A. The mean M and variance V can therefore be derived from known results for the standard uniform distribution.

Let X1 and X2 be uniform random variables over [0,1]. The mean of X1 and X2 is known[41] to have density

$$2 - 4|x-0.5| \text{ for } 0 < x < 1 \qquad (1)$$

so that their sum has density

$$1 - |x-1| \text{ for } 0 < x < 2 \qquad (2)$$

(Vertical bars denote absolute value.) Now X3 = 1-X2 is also uniform on [0,1], and the sum of X1 and X3 follows the above density, implying that X1-X2 also has the same density, but shifted by 1, since X1-X2 = X1+X3-1:

$$1 - |x| \text{ for } -1 < x < 1 \qquad (3)$$

This in turn implies that the distance $D = |X1-X2|$ has density

$$2 - 2x \text{ over } [0,1] \tag{4}$$

and that the squared distance $Y = D^2$ has density

$$\frac{1}{\sqrt{y}} - 1 \text{ for } 0 < x < 1 \tag{5}$$

The mean and variance corresponding to this density are 1/6 and 7/180. Multiplying the original variables X1 and X2 by A will multiply Y by A^2, so that Y will have mean $(A^2)/6$ and variance $7(A^4)/180$. The squared distance between two points sampled uniformly within the hypercube is the sum of N such independent variables, Y1, . . ., YN, and will therefore have mean

$$M = \frac{(N)(A^2)}{6} \tag{6}$$

and variance

$$V = \frac{7(N)(A^4)}{180} \tag{7}$$

Solving for N and A in terms of M and V thus gives:

$$N = \frac{7M^2}{5V} \tag{8}$$

$$A = \sqrt{\frac{30V}{7M}} \tag{9}$$

If a particular dimension is to be imposed on the distribution, then solving for the edge length is an over-determined problem. A simple solution for the edge length is to solve for A for the expressions for both M:

$$M = \frac{NA^2}{6} \tag{10}$$

$$A = \sqrt{\frac{6M}{N}} \tag{11}$$

and V:

$$V = \frac{7NA^4}{180} \tag{12}$$

$$A = \sqrt{\sqrt{\frac{180V}{7N}}} \tag{13}$$

and report the mean value for the edge length:

$$A = \frac{\sqrt{\frac{6M}{N}} + \sqrt{\sqrt{\frac{180V}{7N}}}}{2} \tag{14}$$

Root-Mean-Square Differences

RMSDs were calculated by independently overlaying all pairs of structures using the McLachlan algorithm[8] as implemented in the program ProFit [Martin (1998), University College, London, http://www.biochem.ucl.ac.uk/~martin/programs/] followed by taking the root-mean of the squared atom displacements. Note that RMSDs calculated in this way are not proper mathematical metrics because they include pair-wise rotations that are not transferable across pairs of structures.[14] We show that this issue is not of quantitative significance in the calculations of interest here by comparing to results from a global alignment. For this calculation we use the global alignment algorithm developed by Diamond[44] as implemented in the program polypose which is distributed in the CCP4 suite[45].

Lattice Ensembles

The "lattice_ssfit"[15] lattice ensembles were acquired from the Decoys 'R' Us database at http://dd.stanford.edu/. "Lattice_ssfit" has ensembles for eight proteins which range in size from 55 to 98 residues.

Random Ensembles without Excluded Volume

A discrete-random torsion angle ensemble (DRT) was constructed by randomly enumerating 10,000 conformations of a 36-mer polyalanine chain with ϕ and ψ taking values of -60°, 60° and 180°. Native values for bond lengths, bond angles, and the ω torsion were retained. A compact ensemble (DRT-C) was created from DRT by extracting the 1721 members of DRT with a C_α-R_γ less than 9.5 Å.

Random Ensembles with Excluded Volume

The program YARN [Gregoret (1991), University of California, San Francisco, gregoret@chemistry.ucsc.edu] was used to generate random polyalanine conformations that obey excluded volume constraints. In the default mode, combinations of ϕ and ψ are chosen based on statistics from a reference set of proteins.[16] Alternatively, ϕ and ψ can be selected randomly, while still maintaining excluded volume and user-specified constraints. A user-specified ellipsoid constrains the size of generated conformations.

Error due to under-sampling

We tested the dependence of error on the number of structures in the ensemble by analysis of point distributions in hyperdimensional cubes. Using both a 5 dimensional box and a 15 dimensional box, we found the variance in the dimension for a given number of randomly picked points. For 10 points, the average dimensions are 17.1 for the 15 dimensional box and 5.7 for the 5 dimensional box. Relative standard deviations are 26% for both sets. For 25 points picked randomly from the boxes, the average dimensions reported are 15.6 ± 11% and 5.2 ± 11% for the 15 and 5 dimensional boxes, respectively. Given these results, we use ensembles of at least 25 structures for NCA.

Molecular Dynamics

All molecular dynamics calculations used AMBER[17]. The procedure for the long simulations on villin has been published.[18] We also carried out a number of studies on short trajectories of 300 ps or less. The starting structure coordinates were the minimized average NMR structure of villin headpiece subdomain[19] (PDB[20] access code 1VII) which is the same as used by Duan and Kollman[18]. We performed two sets of simulations for native and misfolded starting structures. For 'native' trajectories, the NMR structure was immersed in 2731 water molecules. The 'misfolded' starting structure was prepared by first performing 1 ns of *in vacuo* MD at 1000 K before solvating with 3102 water molecules. For all systems, water molecules were minimized for 4000 conjugate gradient steps followed by 4000 steps on the whole system. The systems were equilibrated for 10 ps at 10 K, 10 ps at 100 K, 10 ps at 200 K, and 300 ps at 300K. The final equilibrated 'native' structure had a C_α-R_γ of 9.0 Å and C_α-RMSD to 1vii of 3.4 Å. The final equilibrated 'misfolded' structure had a C_α-R_γ of 9.6 Å and C_α-RMSD to 1vii of 7.3 Å.

The Cornell et al.[21] force field was employed with full representation of solvent with the TIP3P water model[22]. An 8.0 Å residue-based cutoff was applied to the non-bonded protein-water and water-water interactions. Intramolecular protein-protein interactions were calculated without truncation. The non-bonded pair list was updated every 50 steps. A time step of 1 fs was employed, except for the very short simulations where a time step equal to the time interval between saved snapshots was used. All bonds involving hydrogen atoms were constrained using the SHAKE[23] algorithm.

A high temperature MD ensemble was generated using the same method as the solvated villin simulations, with the exceptions that: explicit solvent was not included, a distance dependent dielectric function was used for electrostatic calculations and a non-bonded cut-off of 10 Å was imposed. For this calculation, villin was first simulated *in vacuo* at 1000 K for 650 ps. The simulation was continued for another 30 ns, saving snapshots every 100 ps giving an ensemble of 300 high temperature conformations. Each of these conformations was also equilibrated at 300 K *in vacuo* for 300 ps, producing a second ensemble. The high temperature (1000 K) permits large conformational transitions. The 300 ps of room temperature equilibration, in which structures move an average of 2.7 Å (± 0.7 Å) by C_α-RMSD, allows the high temperature structures to move into nearby basins in the energy landscape. For this high temperature ensemble, there is little structural correlation between successive snapshots separated by 100 ps relative to pairs of snapshots further separated in time. The average C_α-RMSD between successive high temperature conformations is 5.3 Å (± 1.1 Å) compared to 7.4 Å (± 1.3 Å) over all pairs of conformations.

Ensembles from the room temperature dynamics were calculated as follows. For short time windows ($t \leq 100$ ps), a trajectory of length 10*t was run and divided into 10 consecutive sections. An ensemble of 25 equally spaced snapshot conformations was taken and NCA parameters were calculated for each ensemble and averaged over the 10 ensembles. Time windows of length 300 ps have only three windows (i.e. a single 900 ps trajectory was divided into three 300 ps sections). Time windows longer than 300 ps use the trajectories calculated by Duan and Kollman[18], which are of total length of 1 µs for folding and 70 ns for native. These longer time points likewise average over ensembles of 25 conformations but differ from the shorter time points in that windows are uniformly distributed over the entire trajectory. For example, the folding time window of 25 ns uses ten 25 ns time windows each spaced by 75 ns. Exceptions include the longest time window, which has only one ensemble and therefore no variance information, and the penultimate window length, which has three windows for averaging.

Principal component analysis

Principal component analysis employed standard techniques of diagonalizing the covariance matrix of C_α atom positions after a least squares fit onto a reference structure. A projection $p_v(t)$ of an MD trajectory $\mathbf{r}^{3N}(t)$ on a given eigendirection $\mathbf{m_v}$ uses the formula[10]: $p_v(t) = \mathbf{r}(t) \cdot \mathbf{m_v}$.

N-cube conformational volumes

An important feature of NCA is that it allows straightforward calculation of conformational volume:

$$\text{conformational volume} = (A)^N \quad (15)$$

where N is the NCA dimensionality and A the NCA edge length. The units are (distance metric unit)^N. Note that direct comparison of these conformational volumes can only be done for systems with the same dimensionality. We utilize two methods to overcome this limitation. First, for some systems, an average dimension can be imposed and an edge length that best matches this dimension can be calculated for a distance distribution (Equation 14). This method is particularly suited for cases where the number of dimensions are very similar. Alternatively, if the dimension difference is meaningful, we examine the possibility of ignoring the 'extra' dimensions of the higher dimensional volume, particularly when the larger NCA edge length space has a lower NCA dimensionality. An example of this concept would be the calculation of how many 3-dimensional boxes one could place upon a 2-dimensional floor. Here, the height of the box is of no concern and this extra dimension can be legitimately ignored.

Results

Introduction

A primary focus of this paper is the use of NCA for analyzing MD trajectories. Our aims are to understand features in conformational space that emerge at various time scales and to assess the energy landscape's influence on the kinetics of protein folding. However, given the novelty of our method, we first explore test systems where the dominant degrees of freedom are known in advance to determine the reasonableness of the results. We then explore ensembles of random villin-like conformations generated by simple geometric manipulation as well as some lattice models of protein folding. We next study the MD trajectories of the villin headpiece subdomain. Knowing the conformational space spanned by these MD trajectories, via NCA, helps us understand protein folding dynamics and energy landscapes. Finally, we return to the question of how the distribution of structures in this representation of conformation space, expressed by p(r), can be used to estimate the probability of finding a structure arbitrarily close to a target structure.

Test Systems: NCA on Protein-like conformational ensembles

To test the hypothesis that the hypercube dimension reflects the number of dominant modes of variation and that the edge lengths are sensible, we created ensembles of protein-like structures with known numbers of near-orthogonal, activated modes through simple geometric manipulations. From the 148 residue two domain protein, calmodulin (PDB[20] code 1CFD[24]), we generated an ensemble of 50 structures by rotating about a single linker region torsion angle, ϕ_{80}, in fixed increments. NCA, using C_α-RMSD as the distance metric, yields a dimension of 1.05 for the ensemble. An ensemble of 125 structures with three active modes was generated by independent rotations about ϕ_{80}, ψ_{80}, and ψ_{81}. NCA yields a dimension of 2.7. To examine motions smaller than domain-level, we created 100 calmodulin-like structures by randomizing all C_α coordinates in *x, y,* and *z* (438 modes) by up to 1 Å. This ensemble, analyzed by NCA, has a dimension of 443 and edge of 0.082 Å. The effective amplitude in this case corresponds to the C_α RMSD between two structures where a single C_α has been shifted by 1 Å, which is also 0.082 Å. To study the behavior of mixing modes of different amplitude, we varied a linker region torsion across the structures randomized in *x*, *y*, and *z*. (See ensemble overlay in Figure 1). The single "domain motion" dominates the variance and the reported dimension is 1.6.

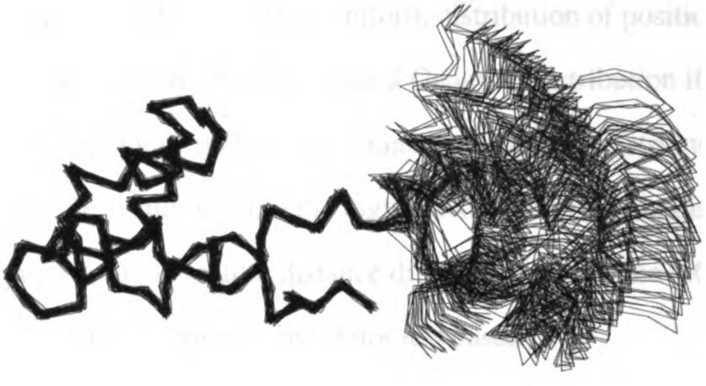


Figure 1: Calmodulin overlay. Shown are C_α traces of calmodulin-like structures generated by random and independent sub-Angstrom shifts in *x, y,* and *z* in addition to a single linker torsion variation. The size of the ensemble's conformational space is $(3.8\ \text{Å})^{1.6}$ (see text). Created by the MIDAS[43] program.

Test Systems: Harmonic Oscillator

A potentially more complex test is provided by an harmonic oscillator. The position distribution of a mass oscillating harmonically will not be uniform. For two points connected by an ideal spring, with one point fixed and the other point oscillating harmonically with unit amplitude, distances between the two points were recorded at random times. Absolute differences between all pairs of distances were generated, from which NCA parameters were calculated. The NCA dimension is 1.12 ± 0.01 (instead of 1) and the NCA amplitude is 1.15 ± 0.01 (instead of 1). These slightly larger amplitudes and dimensions result from the point slowing down near the extremes of its range of motion. This leads to a skew in the distance distribution towards relatively long and short distances, which gives a higher distribution of larger distance differences. Damping the harmonic oscillator would decrease the variation in the point's speed, which would give a more uniform distribution of positions. Conversely, the moving point in this harmonic oscillator has a Gaussian distribution if the probability of a position is Boltzmann energy weighted. For a Gaussian distance distribution, the NCA dimension is 0.71 ± 0.01, resulting from the higher distribution of lengths near equilibrium and thus a higher proportion of smaller distance differences. Thus the NCA results reflect the expected physics in these harmonic oscillator test cases.

Pairwise Alignment Compared to Global Alignment

In this paper, the RMSD between two structures is calculated after optimally superpositioning the two structures. An alternative method for calculating the RMSD distribution is to use a global alignment[44] for calculating all pairwise RMSDs. RMSDs calculated in this fashion are true distance metrics that are embeddable in a space of dimension at most 3 × (number of atoms). NCA analysis of two villin MD-generated ensembles of 50 members equally spaced throughout 1 ns and 1 μs and the villin NMR ensemble are given in Table I. Both pairwise alignment and global alignment results are given. For ensembles with small displacements (1 ns and NMR), the results are nearly identical. The pairwise-based NCA volume for the 1 μs MD ensemble, which has large structure variation, is noticeably smaller than the NCA volume from global alignment . However, for calculations of the nature presented here, this difference is not significant.

Table I: Pairwise versus global alignment* in RMSD calculations.

	Standard Pairwise Based NCA			Global Alignment Based NCA			
	Average RMSD	Dimension	Edge Length	Average RMSD	Dimension	Edge Length	Volume Ratio (Pairwise /Global)
1 μs	6.85	8.70	5.83	6.97	8.33	6.07	0.815
1 ns	0.780	2.708	1.234	0.780	2.709	1.234	0.9999
NMR Ensemble	1.381	3.750	1.819	1.382	3.749	1.820	0.998

* - Standard NCA uses RMSDs calculated after optimally superpositioning pairs of structures. Standard NCA results are compared to results using RMSDs calculated from a single global alignment[44] of all structures. Volumes were calculated using the average dimension between the two NCA methods. The three ensembles analyzed include MD trajectories over 1 μs and 1 ns and the NMR ensemble.

The Conformational Space of Random Protein Folds

We first carry out NCA on ensembles of randomly folded peptide backbones that obey various levels of constraints. We then explore ensembles created from lattice models of proteins and from molecular structural experiments. Lastly, we study MD trajectories.

The first systems we take up are ensembles built from discrete random torsion (DRT) angles. The DRT conformations are distributed uniformly (in torsion space) and no excluded volume or self-avoiding constraints are imposed. We use the DRT ensemble as a reference for other chain models. For more physically reasonable systems, we generate ensembles that include excluded volume, constraints, compactness, and, finally, an ensemble with a limit to the distance between terminal residues. We then compare these random structures with an ensemble generated from a high temperature MD simulation for villin.

Constrained ensembles were built with the program YARN (see Methods). All ensembles contain 2500 members each. No sequence information was encoded - all chains are taken as 36 residue polyalanine. Members of the compact ensembles (Y-C) lie within a constraining ellipsoid while extended conformations (Y-E) do not have this constraint. Very-compact ensembles (Y-V) were built within a constraining ellipsoid with a volume 73% of the standard compact ellipsoid. An ensemble of conformations with terminal C_{α}s separated by less than 6 Å (Y-6) was constructed by selecting from the large set of structures. Generally, there was little difference between the assemblies built using "biased" dihedral angles (Y-B) compared with those using random angles (Y-N). As noted in the Methods section, a high temperature MD ensemble for villin was generated using the same method as the solvated villin simulations, with the exceptions that explicit solvent was not included, a distance-dependent dielectric function was used for electrostatic calculations and a non-bonded cut-off of 10 Å was imposed.

Average C_{α}-R_{γ} and NCA parameters for all ensembles in this section are given in Table II. Distribution functions for some of these ensembles are given in Figure 2 and NCA parameters are plotted in Figure 3.

Table II: NCA parameters and average Rγ for constructed and high temperature MD ensembles.

Ensemble	NCA Dimensionality	NCA Edge Length (Å)	Cα-Rγ (Å)	10-D Volume*
DRT-C	25.3	4.1	8.5 ± 0.7	1.80
DRT	7.6	8.9	12.2 ± 2.8	47.9
Y-CN	23.1	4.0	8.2 ± 0.3	1.14
Y-CB	20.9	4.2	8.2 ± 0.3	1.25
Y-VB	20.8	4.0	7.7 ± 0.2	0.75
Y-VB6	14.9	4.6	7.7 ± 0.2	0.86
Y-EN	8.7	8.3	13.0 ± 2.5	36.1
Y-EB	7.5	9.5	14.1 ± 2.8	81.2
Villin 1000 K (Equil.)	13.1	4.9	8.5 ± 0.4	0.82
Villin 1000 K	11.7	5.3	9.1 ± 0.4	1.42
Villin 300K, 1 μs	10.3	5.4	9.9 ± 1.2	1

* - 10-D conformational volumes are calculated by imposing 10 as the dimensionality in calculation of NCA edge length parameter A. Volume equals A^{10}. All volumes are normalized against the volume for the 1 μs folding simulation[18], equal to 2.44×10^7 Å^{10}.

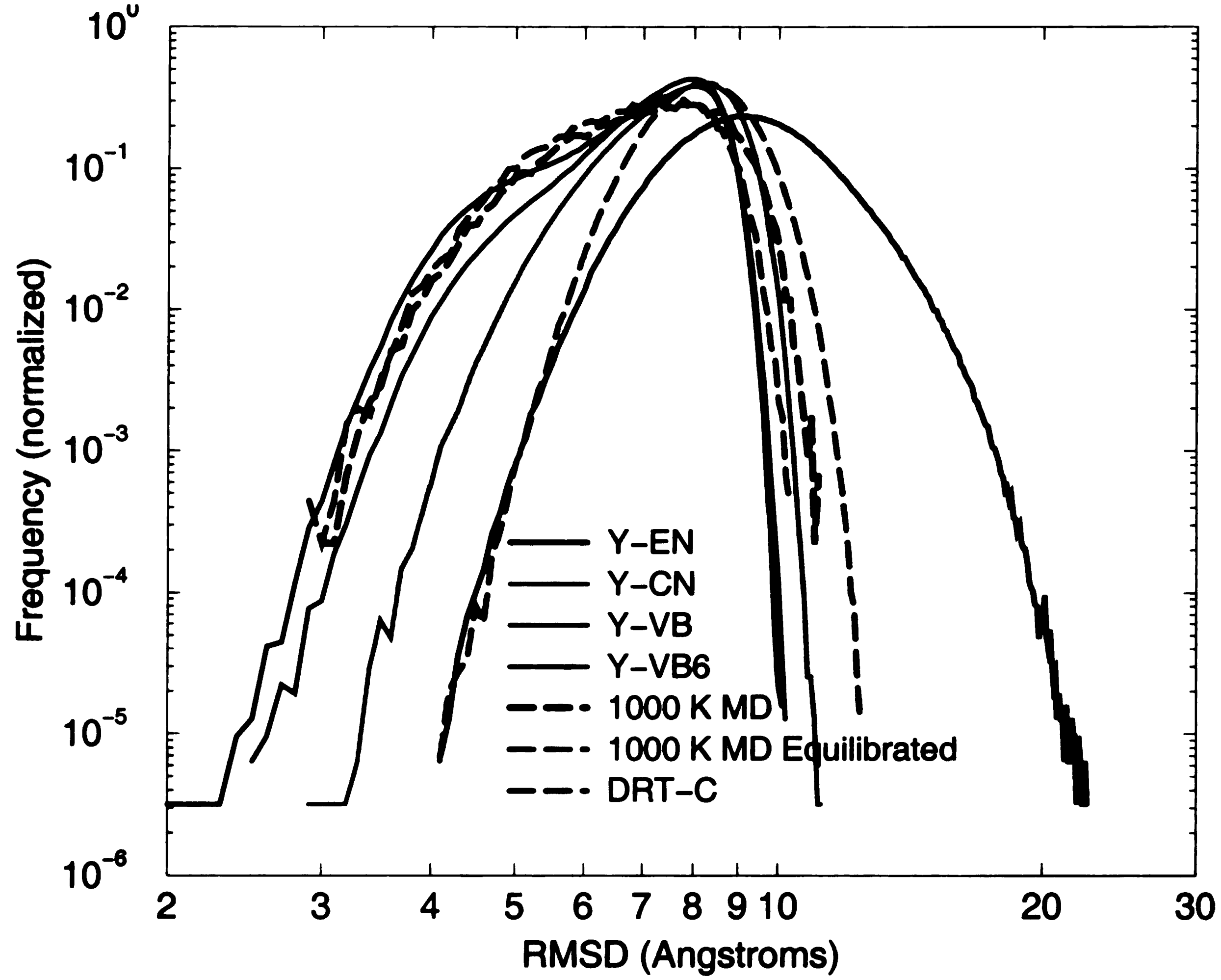
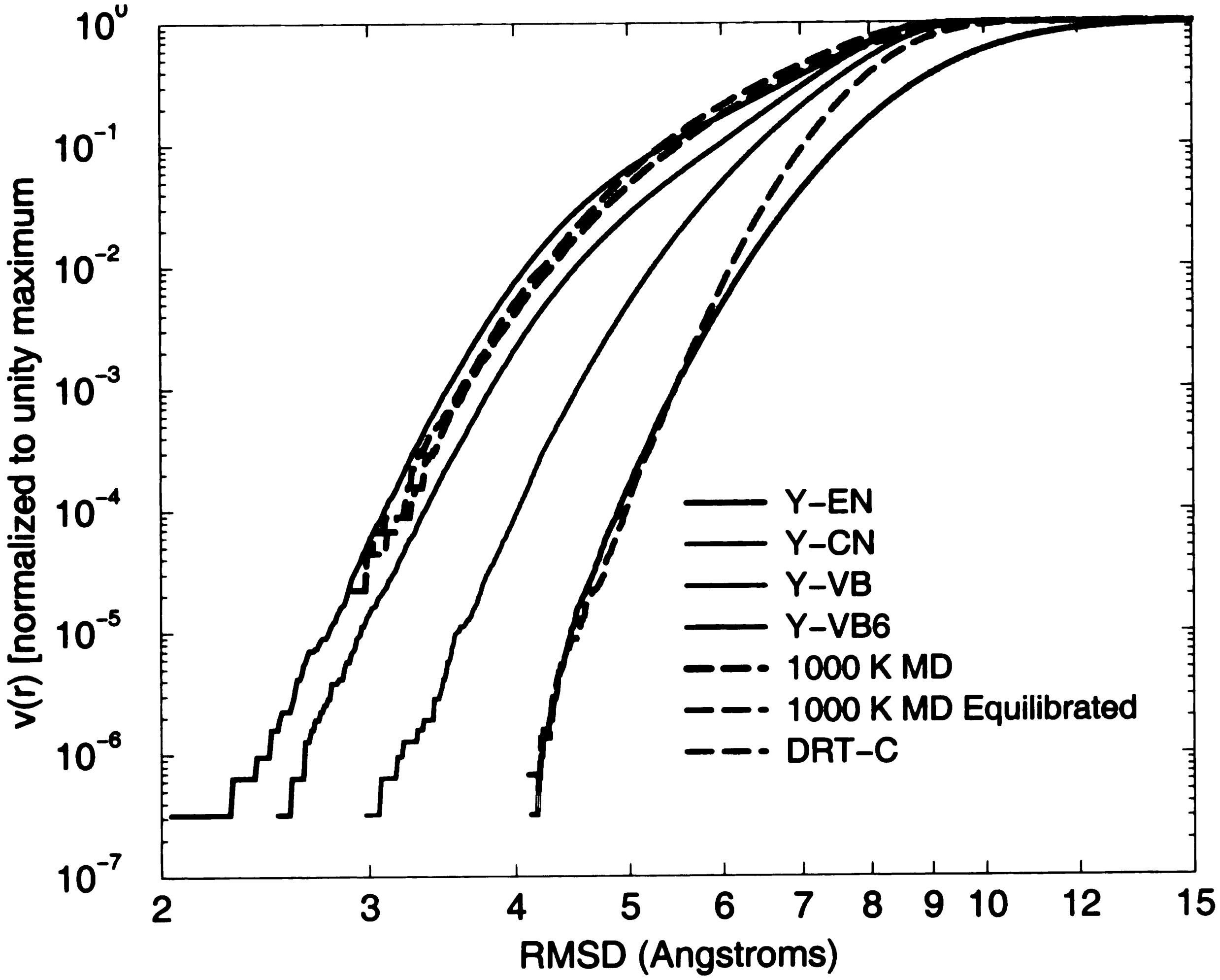


Figure 2a

Figure 2: Comparison of effects of geometric and energetic constraints. Results from seven ensembles [Y-EN (black), Y-CN (green), Y-VB(red), Y-VB6(blue), 300 member 1000 K MD ensemble (black, dashed), 300 member room-temperature equilibrated high temperature MD ensemble (red, dashed), and DRT-C (green, dashed)] are plotted. 2a shows the p(r) distribution normalized to an area of one. For the purposes of displaying the y-axis on a log scale, zero value bins surrounded by non-zero value bins were set to one before normalizing. No more than two bins were artificially non-zeroed in any ensemble. 2b plots v(r), normalized to have a maximum of one, as a function of RMSD.



10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
v(r) [normalized to unity maximum
2
3
4
5
6
7
8
9
10
12
15
RMSD (Angstroms)
Y-EN
Y-CN
Y-VB
Y-VB6
1000 K MD
1000 K MD Equilibrated
DRT-C

Figure 2b

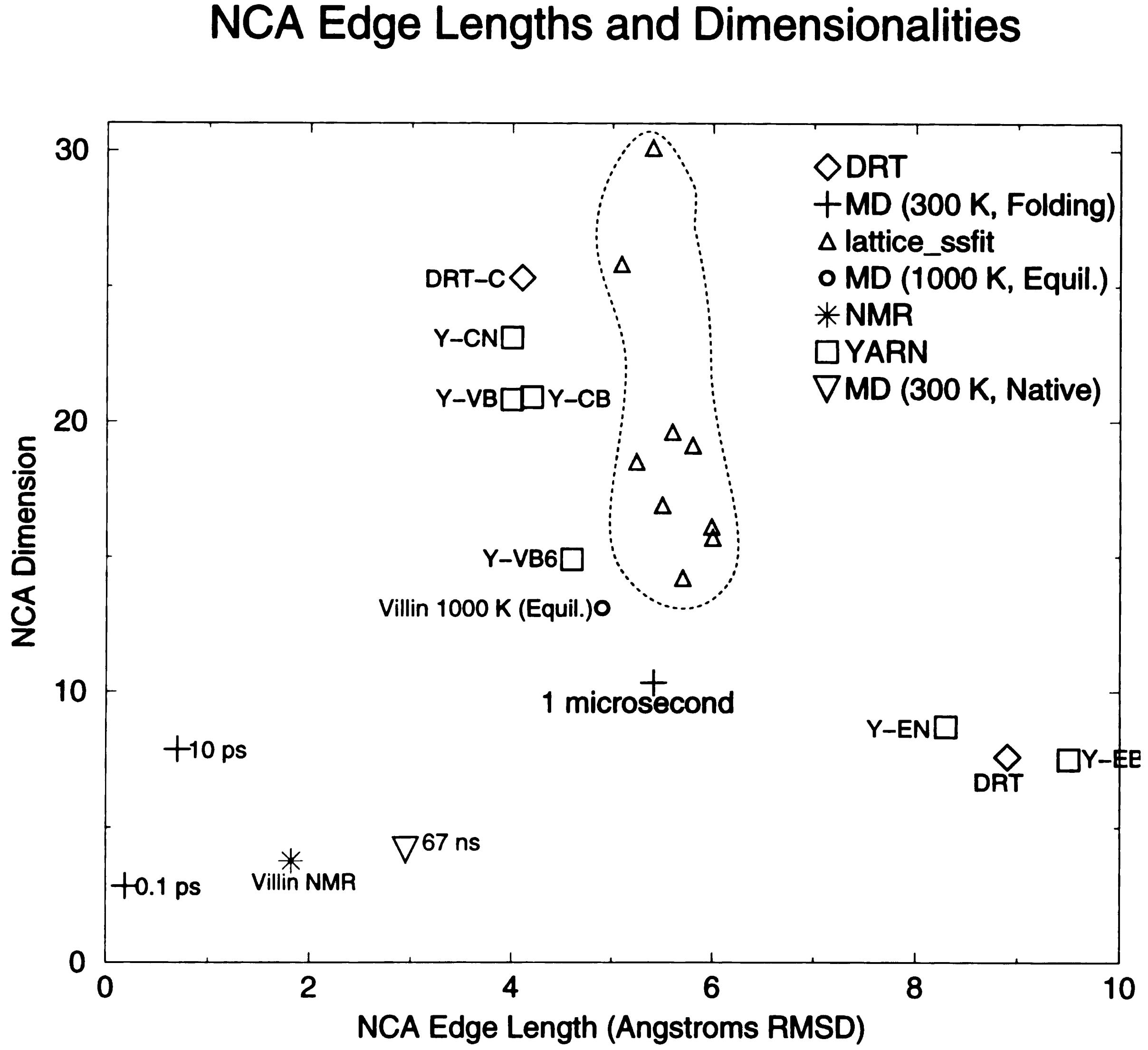


Figure 3: NCA comparison of multiple ensembles. NCA edge lengths and dimensionalities are plotted for DRT and DRT-C (diamond), room temperature MD on misfolded villin (plus), room temperature MD on native villin (triangle-down), lattice_ssfit (triangle-up), high temperature equilibrated MD (circle), villin NMR ensemble (star), and the YARN ensembles (square).

NCA on Lattice Model Ensembles

Our results on random conformations show that constraints as basic as excluded chain volume cause conformations to be distributed non-uniformly. We tested the magnitude of the effects of these deviations by calculating conformational volumes of the "lattice_ssfit" ensembles[15]. Using volumes for the entire ensemble and for sub-ensembles, we can check for self-consistency. Each of their ensembles contains 2000 conformations. From these 2000 structures, sub-ensembles of 100 structures were randomly picked for NCA (set 1). A set of three such sub-ensembles was chosen to provide an estimate of the standard deviation. Another three sub-ensembles of 100 structures each were randomly selected from the 1000 conformations closest via RMSD to the experimental structure (set 2). If the structures are distributed uniformly in conformational space, set 2 should occupy one-half the conformational volume of set 1. NCA parameters were also calculated using the 100 conformations closest to the experimental structure (set 3). For additional comparison, we calculated NCA parameters for the 100 structures with lowest RMSD to the conformation furthest from the experimental structure (set 4). This comparison was inspired by the results of Shortle et al.[25], who found that the density of accepted conformations in ensembles from lattice models is often greater than average in the vicinity of the experimentally determined native structure. While algorithms for generating conformations may search conformational space uniformly, the ensemble of energetically acceptable conformations will reflect properties of the energy landscape.

The NCA results on these lattice model ensembles are given in Table III. We used the average dimension for all sub-ensembles of each model to determine volumes. If the conformations are uniformly distributed throughout conformational space, we would expect the volumes to be of ratio 1 : 0.5 : 0.05 : 0.05 for sets 1 through 4. On average, the ratio expected from uniform distribution between the whole space (set 1) and one-half the space (set 2) is obeyed (0.46 compared to 0.5). However, the near-native sub-ensembles (set 3) on average occupy a smaller space than uniform sampling would predict. This is in agreement with Shortle et al's result[25] discussed above. The sub-ensembles far from the experimental structure but centered about a single particular conformation, set 4, on average occupy a slightly larger space than uniform sampling would predict and is nearly five-fold larger than the near-experimental sub-ensembles' conformational volume. Together, sets 3 and 4 on average occupy 9.4 % of the total conformational volume, which agrees with the uniform prediction of 10 %. In summary, these results suggest that using NCA for conformational volume calculations does give internally consistent results and that the deviations for uniform distribution, at least in these lattice-generated ensembles, are relatively minor.

NCA parameters for these lattice models show a significant dependence of N on the chain length of the model protein (Table III). When the lattice model results are added to Figure 3, we see that they are slightly displaced from the off-lattice YARN models, with a slightly longer edge length and a similar range of dimensionality. While edge lengths are relatively constant for all lattice_ssfit ensembles, dimensionalities correlate with the number of residues with $r^2 = 0.71$. This is consistent with sub-unit size being constant between these ensembles while the number of sub-units increases with the number of residues.

Table III: NCA parameters# and volumes of lattice ensembles[15] divided into sub-ensembles by RMSD from native or outlier.

	NCA Dimension ± Standard Deviation NCA Edge Length (Å) ± Standard Deviation (Å)				
model (number of residues)	Whole space (set 1)	50% with lowest RMSD to Crystal (set 2)	5% with lowest RMSD to Crystal* (set 3)	5% with lowest RMSD to Outlier* (set 4)	volume ratio 1 : 2 : 3 : 4
1fca (55)	18.5 ± 0.3	17.7 ± 1.0	15.9	19.2	1 : 0.64 : 0.0089 : 0.0027
	5.24 ± 0.05	5.2 ± 0.1	4.3	3.7	
1pgb (56)	16.9 ± 0.6	18.1 ± 0.2	17.8	14.7	1 : 0.33 : 0.016 : 0.069
	5.5 ± 0.1	5.0 ± 0.1	4.3	5.0	
1trl-A (62)	14.2 ± 0.5	15.5 ± 0.5	19.5	9.0	1 : 0.31 : 0.0088 : 0.097
	5.7 ± 0.1	5.1 ± 0.1	3.7	5.8	
1ctf (68)	15.7 ± 1.8	16.3 ± 0.6	16.0	18.2	1 : 0.23 : 0.0051 : 0.29
	6.0 ± 0.3	5.38 ± 0.04	4.3	4.5	
1dkt-A (72)	25.8 ± 0.8	21.4 ± 1.6	14.0	14.3	1 : 0.95 : 0.049 : 0.28
	5.09 ± 0.04	5.5 ± 0.2	5.6	6.0	
4icb (76)	16.1 ± 0.1	16.9 ± 0.8	16.6	13.1	1 : 0.21 : 0.0012 : 0.084
	5.99 ± 0.01	5.3 ± 0.2	3.9	5.6	
1nkl (78)	19.1 ± 1.1	17.9 ± 0.7	17.8	14.3	1 : 0.37 : 0.0089 : 0.046
	5.8 ± 0.1	5.6 ± 0.2	4.6	5.4	
1beo (98)	30.1 ± 1.4	28.7 ± 0.5	17.2	15.4	1 : 0.61 : 0.041 : 0.014
	5.4 ± 0.1	5.4 ± 0.1	6.0	6.0	
Averages (71 ± 14)	19.6 ± 5.5	19.1 ± 4.3	16.9 ± 1.6	14.8 ± 3.1	1 : 0.46 : 0.017 : 0.077
	5.6 ± 0.4	5.3 ± 0.2	4.6 ± 0.8	5.3 ± 0.8	(± 0.25) (± 0.018) (±0.087)

# - The NCA dimension is given in the top of each cell, with the NCA edge length (divided by Å) given below.

* -Standard deviations are not reported for the 100 member sub-ensembles for which only one NCA calculation was performed, but can be inferred from the calculations on the larger sub-ensembles performed in triplicate.

Villin

1 microsecond trajectory

We next undertake an assessment of the conformational space of a MD folding trajectory. A study of villin headpiece subdomain folding, derived from a 1 μs MD simulation, has been reported.[18] The protein exhibits large conformational fluctuations over much of the trajectory. A marginally stable conformation is formed in the interval between 240 to 400 ns. This state is characterized by a consistently low radius of gyration (R_γ), RMSD from native structure, the solvation component of the free energy, and structural clustering. The difference between the folding regime and the marginally stable (mis)folded regime is readily detected by NCA. We divided the trajectory into 1 ns windows, each window containing 50 structures, and calculated the number of dominant modes over time (Figure 4). During the folded (low R_γ) time region, the dimension jumps to values greater than 8, compared to an average of less than 5 in the remainder of the trajectory. Principal component analysis, with selection of the dominant modes, leads to a comparable but noisier result (data not shown). Table IV lists correlations between NCA dimensions and the number of PCA dominant modes for several threshold values. The ratio of the average number of NCA dimensions to PCA modes is also given.

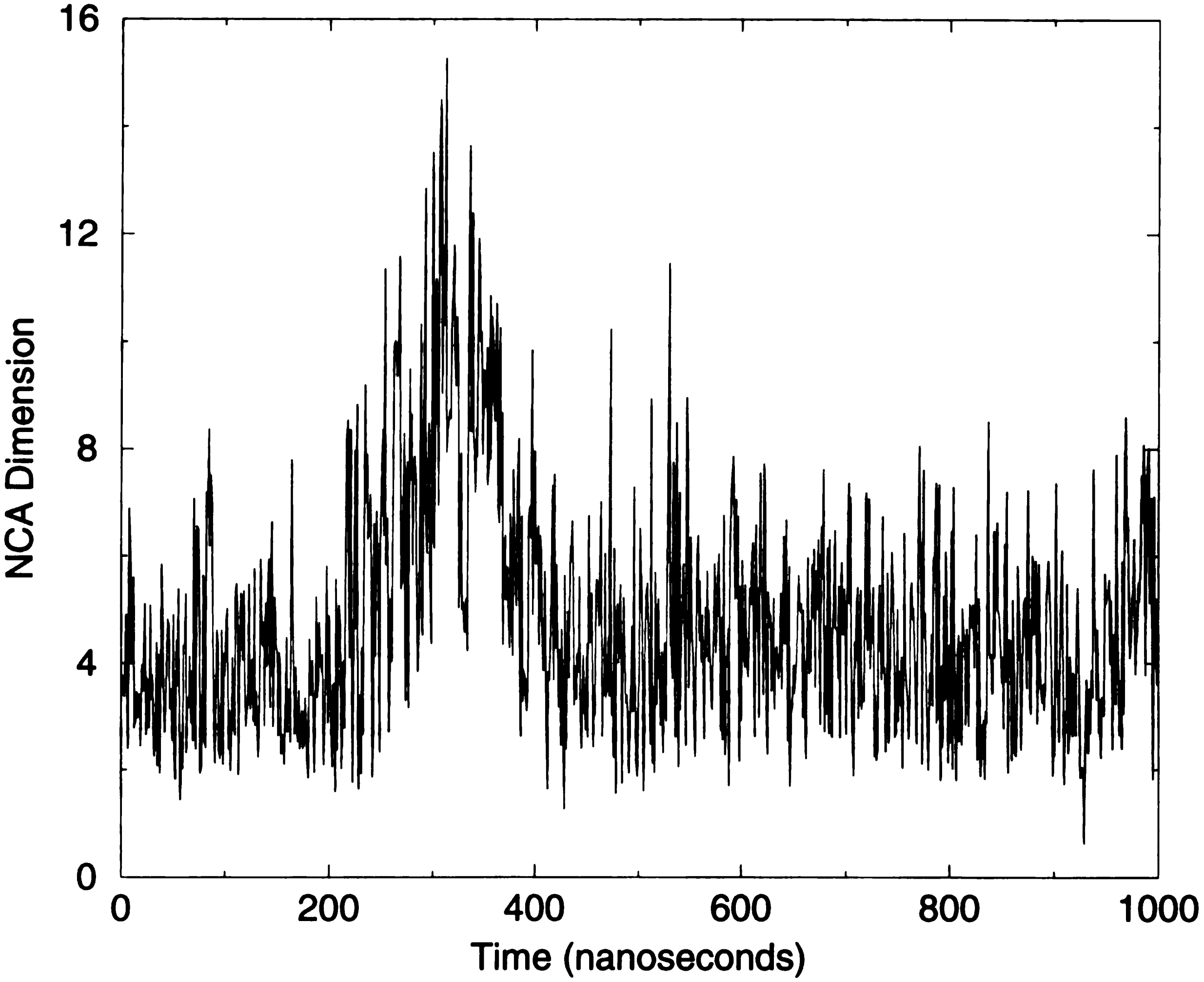


Figure 4: Number of dominant modes for villin 1-μs MD trajectory. The number of dominant modes for 1-ns time windows was calculated using NCA.

Table IV: Correlation between NCA dimension and number of PCA modes which capture X% of variance for 1000 time windows of 1 ns each.

X %	Correlation between NCA Effective Dimension (N) and # of PCA Modes which capture X% of variance	<NCA Dimension>* <PCA # of modes>*
50	0.86	2.1
60	0.87	1.5
70	0.85	1.1
80	0.83	0.71
85	0.82	0.55
90	0.79	0.40
95	0.72	0.25
99	0.49	0.11

* - Closed angle brackets indicate mean value.

Time domain analysis

To understand the peak in dimensionality in Figure 4, we explore the time dependence of the NCA parameters using windows of various lengths for the two villin simulation sets: the folding trajectories from 1 fs up to 1 μs and "native" villin simulations from 1 fs to 67 ns (see Methods).

There are several interesting features seen in the plot of dimensionality versus time interval (Figures 5a and 5c). Starting at the shortest times, both trajectories give rise to an increase in dimensionality from 1 at the shortest time scales, to a weak shoulder at 0.1 ps (amplitude ca. 0.2Å) and a peak at ~8 dimensions in the 1-100 ps interval (amplitude ca. 0.5Å). This behavior is consistent with the dephasing of the elastic vibrations of globular proteins in a single substate within 0.1 - 1 ps,[26,27] followed by exploration of thermally-averaged anharmonic conformational minima on the 1-100 ps time scale. The convergence in dimensionality on 1 at very short time scales for all states reflects the fact that the positions of all C_αs are correlated at the shortest times. This convergence supports the concern noted by Janezic et al.[28] that modes of similar frequency will not dephase in short time simulations and will result in artificial mixing of modes.

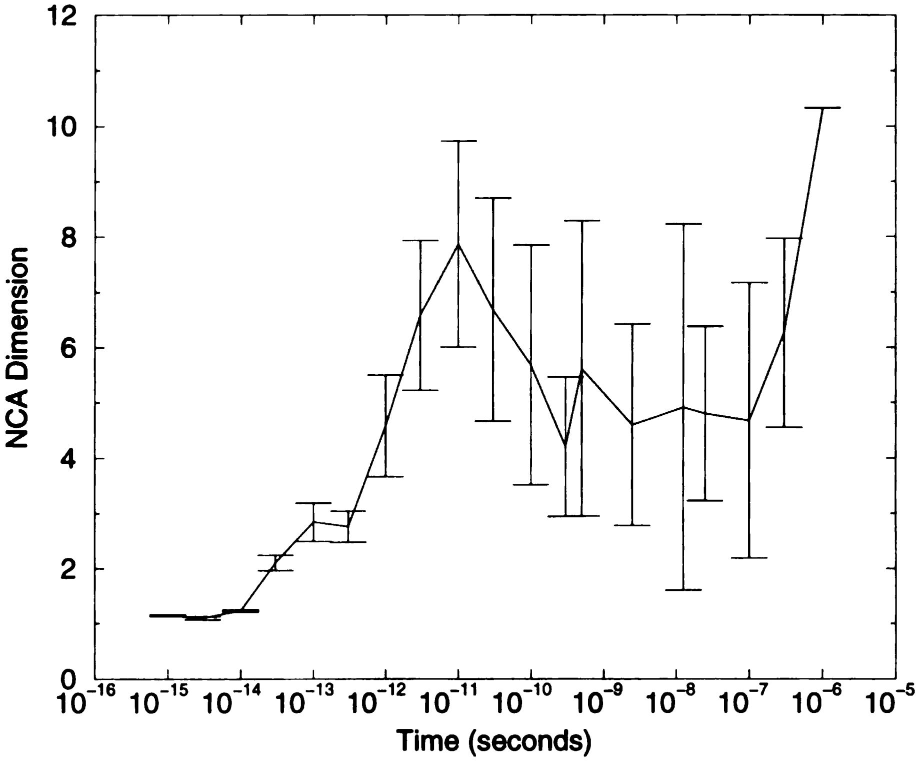


Figure 5a

Figure 5: Conformational volume dependence on time window size. NCA parameters were calculated for MD trajectories of various lengths for a 'folding' regime (dimension: 5a, edge length: 5b) and a 'native' regime (dimension: 5c, edge length: 5d). With some exception (see text), the average NCA parameters of ten trajectories with 25 conformations uniformly spaced in time are graphed.

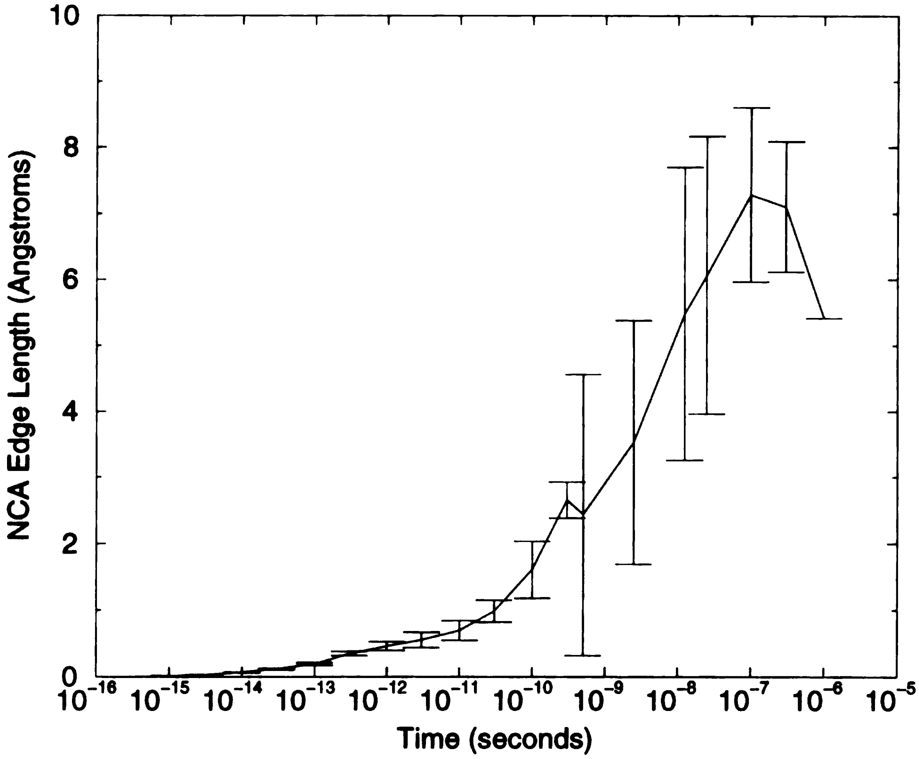


Figure 5b

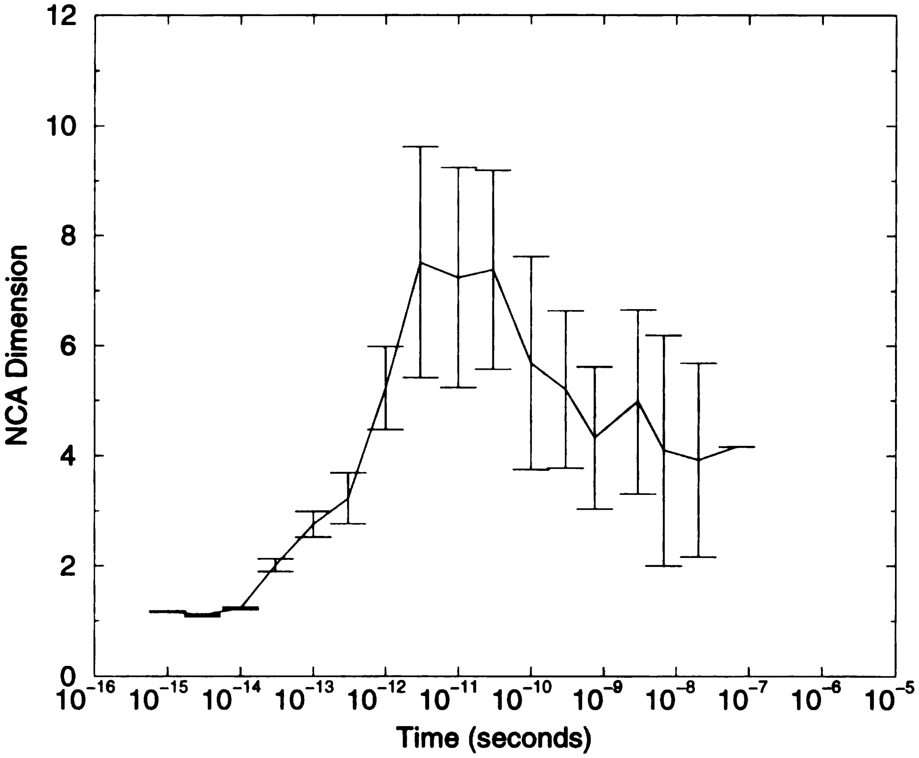


Figure 5c

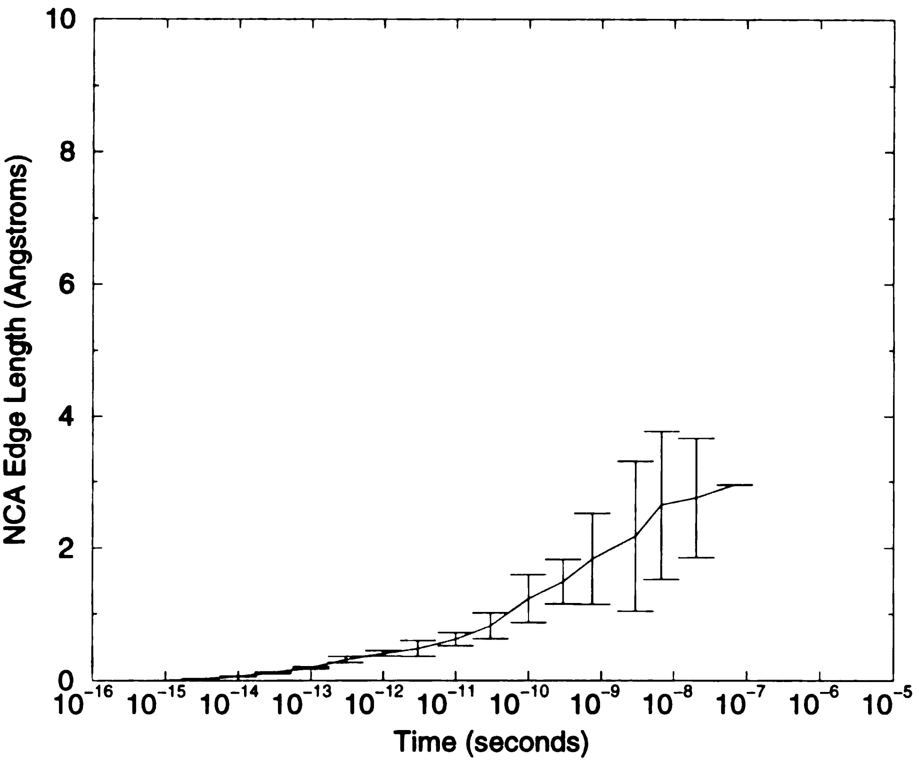


Figure 5d

Very short time intervals

We can look more closely at very short time behavior with a slight modification of our procedures. Instead of comparing all pair-wise conformations in a time window, we take RMSD measurements only between successive snapshot conformations. We refer to this technique as Successive N-Cube Analysis (S-NCA). We used multiple ensembles of 250 structures from the misfolded/folding simulations where ensembles differ from each other by the time interval between successive conformations. We examined time intervals from 0.04 fs to 4 ns. The premise for S-NCA is that a large abundance of vibrational modes decreases the variance in RMSD between consecutive structures, yielding a higher number of dimensions. We would expect the dominant modes at sub-picosecond time scales to be the normal mode vibrations. The C_α RMSD S-NCA dimensionalities peak on the order of 100 in the 1-100 femtosecond time scales (data not shown). This is sensible given that there are 36 C_α atoms with 102 internal degrees of freedom. S-NCA dimensionalities calculated from all-atom or all-heavy-atom (non-hydrogen) RMSD measurements peak significantly higher, on the order of ~1000, since many more atoms (and degrees of freedom) are captured. Villin has 596 total atoms and 295 heavy atoms. The use of the SHAKE[23] algorithm in the simulations likely reduces the S-NCA dimensionality for the all-atom RMSD calculation from what it would otherwise be. Coordinates of conformations were only saved to 0.001 Å limiting the significance of results below 0.1 femtoseconds.

Nanosecond time scales

Returning to our analysis of the MD trajectories, transitions between thermally averaged conformational substates comprise the dominant modes during nanosecond time scales, leading to a lower number of effective dimensions and much increased edge length. The NCA parameters for the native state of villin continue to track those for the folding trajectory although the edge lengths start to deviate at longer times. In the parlance of Clarage et al.[13], this time scale transition is analogous to going from the level of "beads" (<100ps, many intra-substate modes dominate) to "beads on a string" (>1ns, fewer inter-substate modes dominate). The high standard deviation in the dimension values between 100 ps and 1 ns partly arises from the large variation in lifetime of the substates. This variance is seen in the 1 μs trajectory, where the majority of nanosecond time blocks reveal a low dimensional regime, which implies that substates persist for less than 1 ns. However, from 240 ns to 400 ns, substates persist at least 1 ns as indicated by high effective dimensionality using 1 ns time windows (Figure 4).

Microsecond time-scale

For the folding simulations, there is a rise in the dimension as the 1 μs time-window is approached. We interpret this as a dephasing of the "folding modes". This is similar to the dephasing of the substate modes at shorter times, but on a larger scale. These folding modes have NCA amplitudes of ~5-6 Å compared to sub-Angstrom amplitudes in the conformational substates. As expected, the native trajectories do not show this same increase in dimension at longer time points; however, the longest native trajectory is more than an order of magnitude shorter than the longest folding trajectory. We can learn more about longer time-scales of the native structure of villin if we make the naive, yet useful, interpretation that the NMR ensemble[19] is equivalent to a series of conformations separated by long time intervals. The NMR ensemble has an effective dimensionality of 3.8, which is similar to the effective dimensionalities of the longer native villin MD trajectories and an edge length of 1.8 Å. This relatively low dimensionality contrasts with longer-time dimensionalities for the protein folding trajectory. Both MD and NMR results agree with the model that native conformational space consists of relatively few medium amplitude modes that connect many conformational substates which are themselves higher dimensional entities.

Principal Component Analysis

We performed PCA on the entire 1 μs folding simulation and projected the MD trajectory along with the native conformation into the largest PCA modes (Figure 6). These results show that the native structure is contained within the conformational subspace explored by the simulation, even though the native state is not explicitly sampled during the simulation. Finding the native conformation is thus a matter of searching the space at a finer resolution (i.e. more time) and not moving into new spaces. Projection of the MD simulation's coordinates onto the PCA coordinate system does not reveal any obvious progression toward the native state. Instead, fluctuations along the largest eigenmodes appear to have a time period much shorter than the length of the simulation. This lack of progression in the largest eigenmodes is one indicator that the trajectory has 'converged' on the accessible folding space.[12,13] Additionally, the NCA conformational volume of the 1 μs folding simulation is approximately equal to the conformational volumes of the random compact polypeptide ensembles (Table II). We did not explore the physical significance of the PCA eigenmodes, with the exception of the largest eigenmode which has the nature of an expansion/contraction mode, i.e., the correlation coefficient (r^2) of the trajectory along PC 1 with the radius of gyration is -0.86.

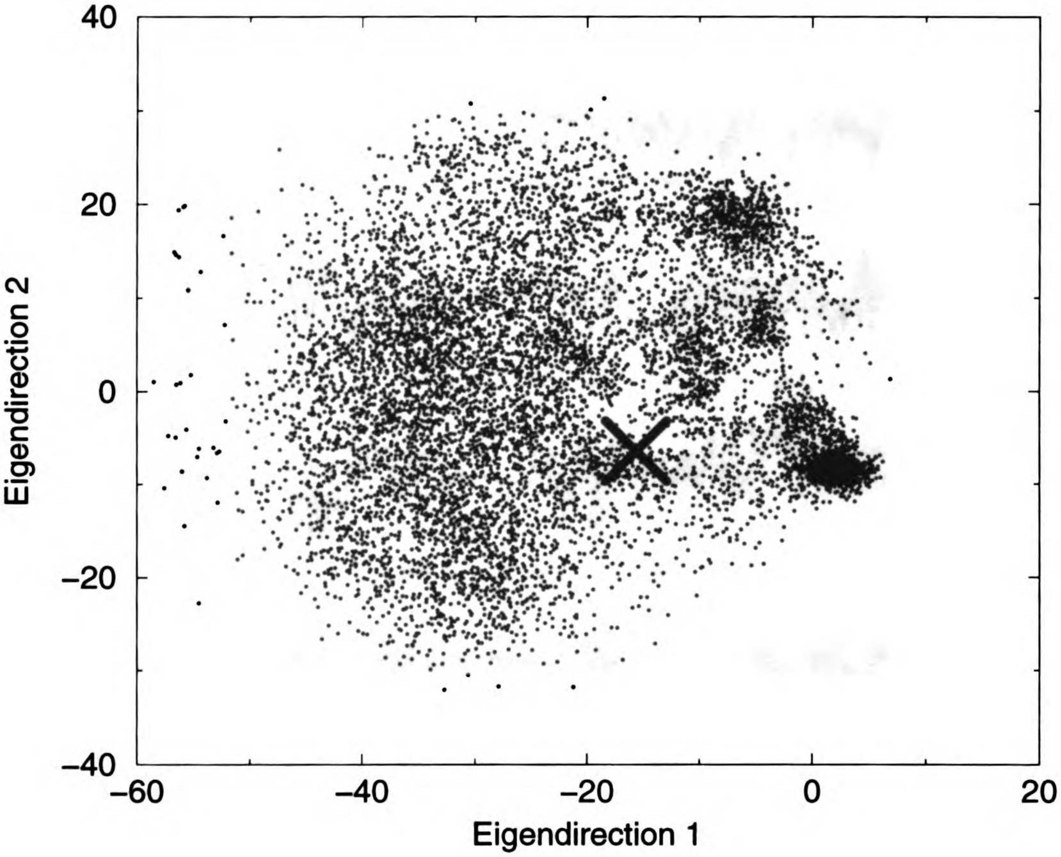


Figure 6a

Figure 6: PCA projection of villin 1-μs MD trajectory compared to native. PCA coordinates were calculated for the villin 1-μs MD trajectory with snapshots spaced by 100 ps. 6a is a two dimensional projection (see Methods for projection procedure) into the two largest principal components. 'X' marks the position of the native conformation in the PCA space. 6b shows the movement in time of the trajectory in the eight largest PCA modes. The thin horizontal line marks the native position.

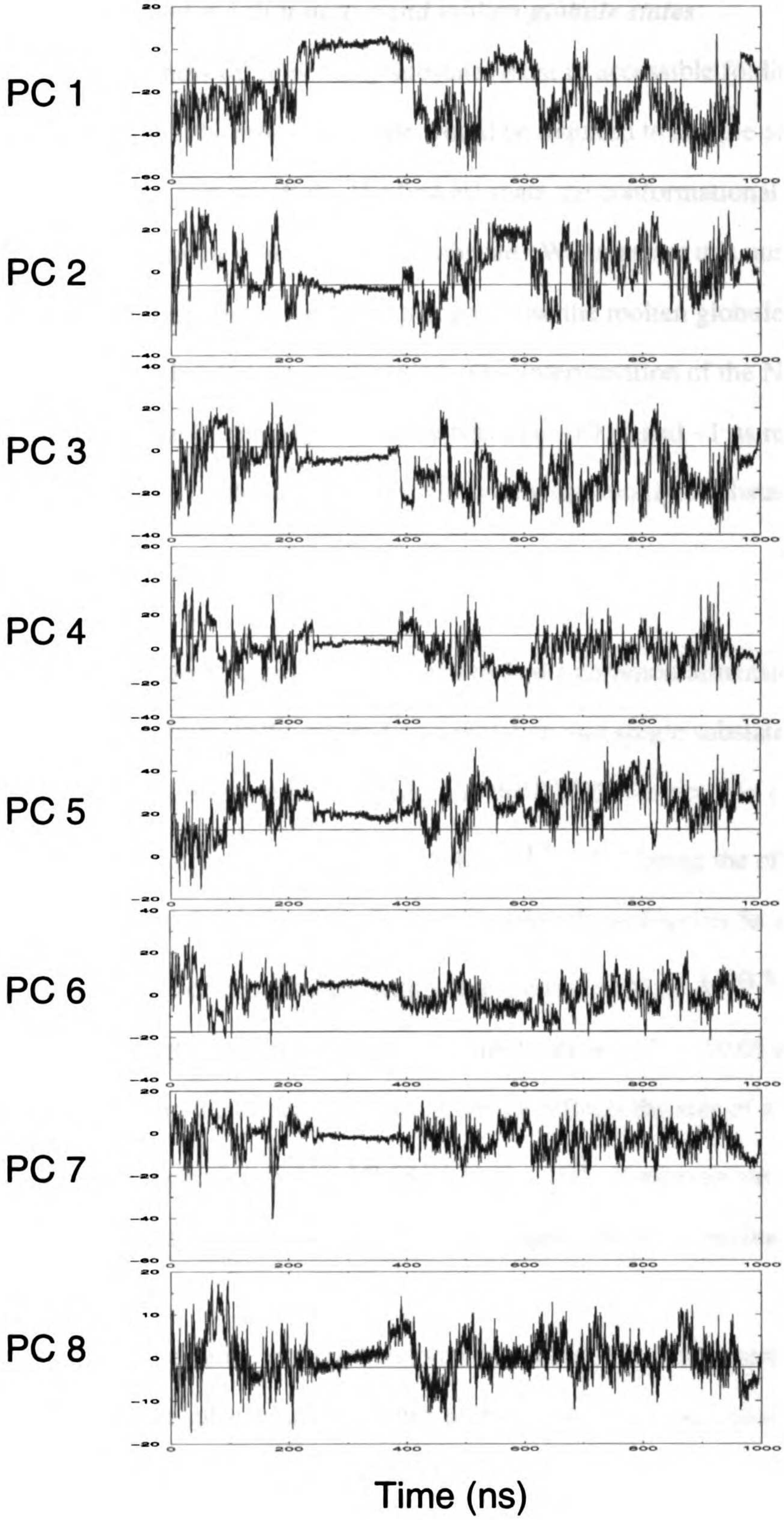

PC 1
20
0
-20
-40
-60
PC 2
40
20
0
-20
-40
PC 3
40
20
0
-20
-40
PC 4
60
40
20
0
-20
-40
PC 5
60
40
20
0
-20
PC 6
40
20
0
-20
-40
PC 7
20
0
-20
-40
-60
PC 8
20
10
0
-10
-20
0
200
400
600
800
1000
Time (ns)

Figure 6b

Size of conformation space for Villin native and molten globule states

It would be interesting to know (a) how many substates exist in accessible folding space (i.e. the molten globule state) and (b) how much time would be required to sample all substates. NCA can be used to answer this question. We first estimate the conformational volume ratio of the molten globule state to the conformational substate. We interpret this number as the maximum number of conformational substates that compose the molten globule. We then need to find the rate at which substates are sampled. Our interpretation of the NCA dimensionality plots, that the fall in dimensionality between ~10 ps and ~1 ns results from a transition from intra-substate exploration to excursions between multiple substates, yields an estimate on this rate.

In order to compare two hypervolumes, we first need to find a common dimensionality. Since the NCA dimensionalities of the entire 1 μs simulation and single substates are similar (~10), we use the method of imposing a dimension on the RMSD distribution (Equation 14). From Figure 5, if we assume folding space to be $(5.4\ \text{Å})^{10.3}$ ('10.3' being the effective number of dimensions and 5.4 Å being the edge length, taken from Figures 5a and 5b at 1 μs), then the maximum number of substates, with volume equal to $(0.68\ \text{Å})^{10.3}$ [NCA parameters of the 30 ps length ensemble with maximum dimension (N = 10.0) with a new edge length found by forcing the dimension to 10.3, which reflects the size of a single substate's conformational space], is $5.4^{10.3}/0.68^{10.3} \approx 2 \times 10^9$. If each minimum persists for 100 ps on average, exploration of all minima would require about 0.2 seconds.

The time required to find the native state by random exploration is a more interesting number than the time to find a single substate in the molten globule, which we just calculated. The time required for protein folding by random exploration of the molten globule can be found by first recognizing that the native state is not a single conformational substate, but a collection of substates.[29] Therefore, finding the native state is an easier task than finding a single substate. NCA can be used to give a crude estimate of the number of substates that compose the native state. We have two independent data sets for estimating this number: a 67 ns MD simulation of native villin[18] and the NMR determined structure ensemble[19]. The conformational volume explored in 67 ns of MD simulation is $(4.1 \text{ Å})^{2.9}$. The NMR conformational volume is $(1.8 \text{ Å})^{3.8}$. Both MD and NMR suggest native conformational space to be of lower dimension than substate space. We assume intra-substate space to be composed of modes coincident with inter-substate modes as well as extra, orthogonal modes. These coincident intra-substate modes are simply less activated forms of inter-substate modes. Therefore, we ignore the extra dimensions unique to intra-substate space for volume comparison. By MD, the number of native substates is 183 $[4.1^{2.9}/0.68^{2.9}]$. By NMR, the number of substates is 40 $[1.8^{3.8}/0.68^{3.8}]$. These results suggest that the native state's conformational space is about two orders of magnitude larger than that of a single substate. Therefore, about two orders of magnitude less time is needed to find the native state of villin than the ~0.2 seconds needed for finding a single substate within the native state. Protein folding by random exploration of the molten globule is therefore calculated to require milliseconds.

The NCA parameters for the native and molten globule states of villin, as well as those for the NMR ensemble, are compared to the other structural representations in Figure 3. We see that the microsecond villin folding ensemble parameters, and especially those for the simulation at 1000 K, fall near the most compact of the random structure arrays (Y-VB6). For time windows of a nanosecond or less, all the folding trajectories look rather similar to each other in the lower left corner of the diagram. The NMR ensemble and the native villin simulation have almost the same NCA dimensionality (4) and edge length (2 Å). At very short times, as noted above, the NCA parameters tend towards a dimensionality of 1, reflecting the artificial coupling of the vibrational degrees of freedom, and an edge length of a fraction of an Angstrom.

Detailed Distribution of RMSDs and Implications for Finding "Native-like" Structures

Up to this point we have used NCA to analyze the distributions of RMSDs for a given ensemble under the assumption that the distribution could be fit to two parameters. In fact, the distributions are more complex and additional information can be extracted. Returning to Figure 2, notice that the p(r) distributions for the compact ensembles all have approximately the same mean values. The variances differ significantly and are the source of the dimensionality changes in this set (see Table II and Methods). It is interesting that several of the most constrained distributions show secondary shoulders at smaller r, indicating some weak clustering of conformations.

We compare DRT's distributions with simple models in Figure 7 (see also Appendix). The normal distribution and a NCA model based on DRT's dimensionality rounded to the nearest integer (N = 8) both fail to describe the asymmetry of the DRT distribution (Figure 7a), under-predicting the frequency of finding longer distances and over-predicting the frequency of shorter distances (Figures 7a and 7b). The NCA distribution, compared to the normal curve, does have the redeeming quality that it does not intersect the y-axis at a non-zero value. Rather than using these simple models, a more accurate assessment of the probability of finding structures that fall "close" to an arbitrary reference structure can be generated by extrapolating (where necessary) the limiting slopes on the v(r) distribution plots (Figures 7c and 7d). For the ensembles listed in Table V, these limiting slopes (not extrapolated) all have approximately the same value (20) on log-log plots (data not shown). The high dimensionality of the conformation space utilized by these ensembles implies that it is very difficult to find structures close to an arbitrary structure. On a numerical basis, the sample populations must be expanded at least by 10^6 to start to see reliable representations within 3 Å. and beyond 10^9 to produce structures within 2 Å for all the distributions. An extrapolation of DRT extended by assuming an increasing slope of v(r) at small r (see Appendix) shows the probability of randomly generating a structure within 2 Å as less than 10^{-22}.

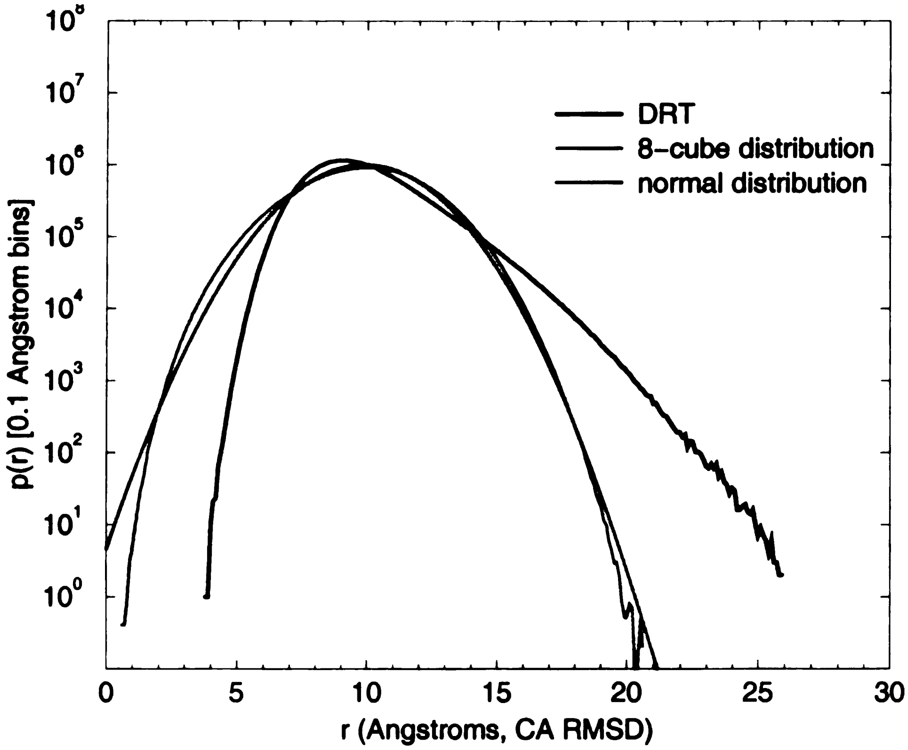


Figure 7a

Figure 7: Comparison of multiple distributions. In 7a, DRT's p(r) distribution (black) is compared to a normal distribution (red) and an n-cube distribution (green). 7b plots v(r) for the three distributions and also plots the unbounded 8-space derived v(r), Equation A-5 (blue). 7c gives the slope of log(r) versus log(v(r)) for v(r) calculated for the four sets of v(r). The slope of log(r) versus log(v(r)) was calculated in a 0.2 Å sliding window (20 data points) using the method of determinants[40]. 7d uses Equation A-12 (cyan) to extrapolate DRT's v(r) to r = 0.2 Å. v(r) distributions in 7d are normalized to unity maximum.

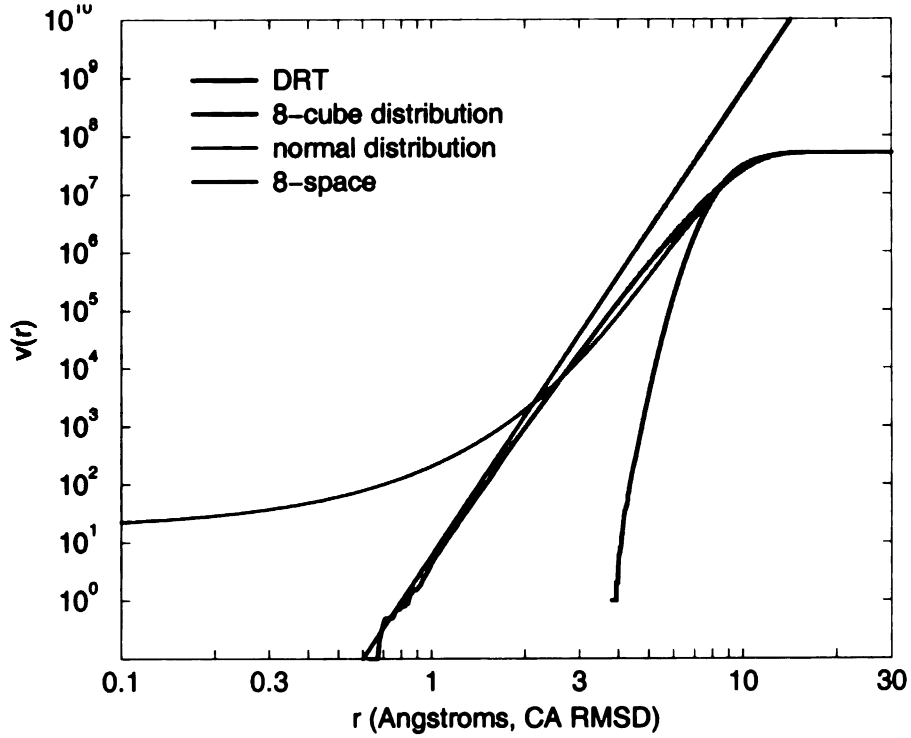


Figure 7b

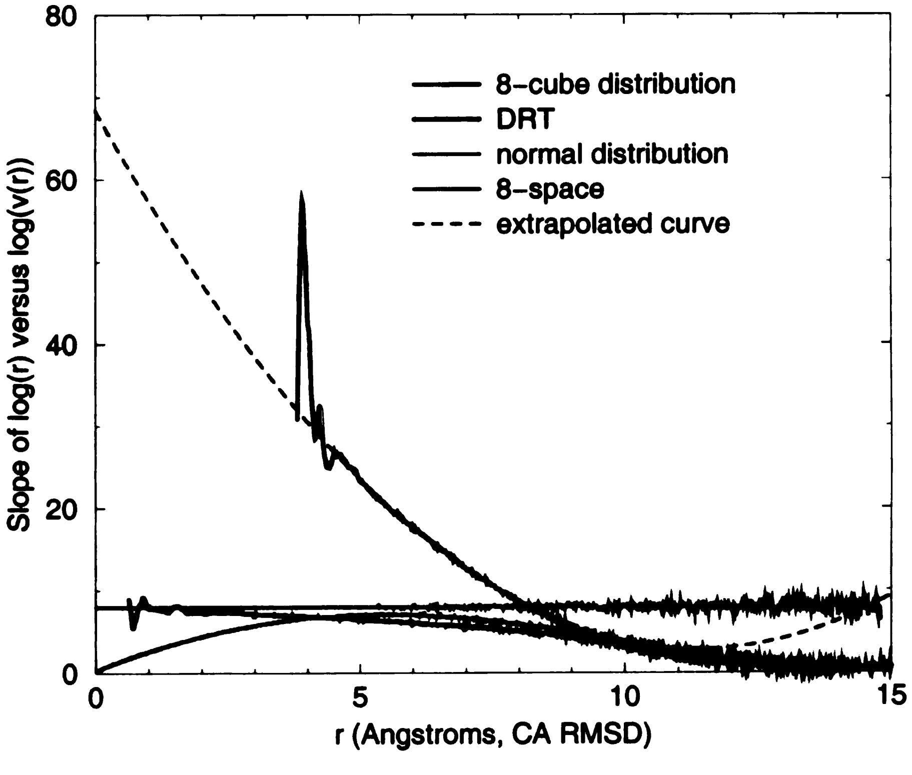


Figure 7c

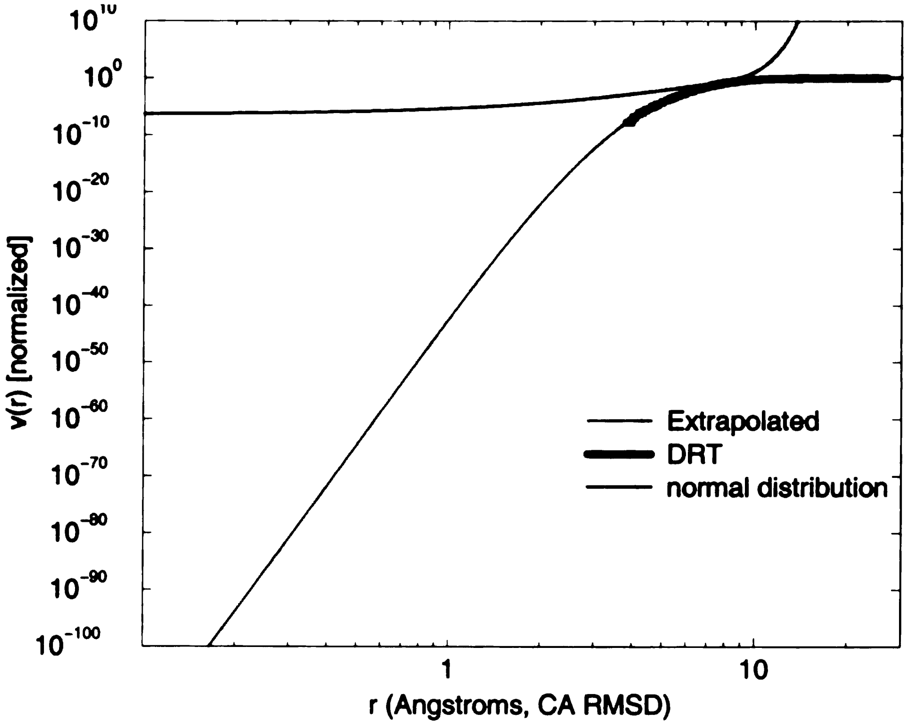


Figure 7d

Table V: Probability* of finding a structure within a given RMSD.

Ensemble	RMSD		
	2 Å	3 Å	4 Å
Y-VB6	<1.9 × 10^{-8}	5.6 × 10^{-5}	7.5 × 10^{-3}
1000 K Equilibrated	<1.2 × 10^{-8}	4.5 × 10^{-5}	5.2 × 10^{-3}
Y-VB	<4.9 × 10^{-9}	1.4 × 10^{-5}	2.1 × 10^{-3}
DRT-C	<9.2 × 10^{-13}	<3.0 × 10^{-9}	<9.6 × 10^{-7}
Y-EN	<6.3 × 10^{-13}	<2.1 × 10^{-9}	<6.6 × 10^{-7}

*-Probabilities are the normalized value of v(r) at a given RMSD. Probabilities that are prefixed by a '<' were calculated by performing linear extrapolations of log(r) versus log(v(r)). The point where v(r) = 10 (not normalized) is taken as the point of extrapolation in order to decrease the effect of noise at low v(r) values. A slope of 20 for log(r) versus log(v(r)) is used for all extrapolations. Since the slope appears to be rising at decreasing RMSD for all ensembles, extrapolations should be treated as overestimates.

The Effects of Geometric and Energetic Constraints on Conformational space

As shown in Figure 3, we find that the NCA parameters and average C_α-R_γ separate the ensembles into three major regimes - the extended sets have a dimensionality ca. 8 and edge lengths of 8-9 Å similar to those found for the random flight ensemble (DRT). The compact random structures have dimensionalities of 15-25 and edge lengths of 4-5 Å, with C_α-R_γ relatively close to that of structures from the equilibrated version of the 1000 K MD simulation of villin. Even though the degree of compactness as measured by R_γ is relatively constant, the number of NCA dimensions varies inversely with the level of constraints with the compact random flight ensemble DRT-C having the largest dimensionality (25) and the "very compact, close ends" Y-VB6 set most closely approaching the molten globule simulation (see below). This result is consistent with there being fewer effective modes of achieving a fixed displacement in the more highly constrained ensembles. There are small effects of ϕ/ψ bias in determining the spread of YARN conformations. The compact, biased Y-CB ensemble is in a tighter conformational space than the unbiased sample Y-CN (Figure 3). Similarly, the equilibration procedure tightens the conformational space of the high-temperature MD ensembles (data not shown). The native structures (NMR, MD) occupy the third regime of small dimensionality and small edge length.

Non-Uniform Distributions from Conformational Model Systems

To explore further the effects of extreme non-uniformity of the RMSD distribution, we considered a system composed of M conformational minima. Each conformational minimum is equidistant from all the others and is equally populated (same well depth). For M minima, an M-1 dimensional space is required. We generated RMSD distributions by randomly placing points into minima and recording the point-to-point distance for all pairs, yielding a distribution function composed of two delta functions at 0 and L, the inter-minima distance.

NCA yields the data in Table VI, showing that reasonable results are obtained even with extremely non-uniform distributions.

Table VI: NCA on minima equidistantly spaced by one unit.

Number of Minima - 1*	N	A
1	1.40	1.46
2	2.81	1.19
3	4.23	1.03
5	7.06	0.84
10	14.0	0.62
20	28.1	0.45
100	140	0.21
1000	1400	0.065

*-Since a set of M equidistant points embeds into a space with M-1 dimensions, M-1 is listed for convenience in comparison with NCA results.

Discussion

We have presented new tools for protein ensemble analysis that are based on a Euclidean model for conformational space. While this model requires simplifying assumptions and does not provide a rigorous statistical mechanical treatment of conformational degrees of freedom, we find we can get interesting answers to three basic questions for models of protein conformations. First, the simplicity of the approach allows comparisons among static and dynamic ensembles generated in very different ways, revealing general features about effective mechanical motions. Second, the time dependence of dynamics trajectories suggests limits on random search folding times. Third, we see that the probability of finding structures within a particular ensemble arbitrarily close to another structure is very small in high dimensional conformation spaces. We consider the implications of each of these results below.

If we look at the NCA parameters for all the static and dynamic ensembles considered in this paper, the most striking results are for compact systems which differ little in their radius of gyration or NCA edge length but differ dramatically in the NCA dimensionality. As one might expect intuitively, the effective degrees of freedom decrease smoothly as the constraints increase. The difference between the most constrained random structure ensemble (Y-VB6) and the high temperature or microsecond ensembles of villin are rather small, suggesting that relatively few additional constraints are needed to generate molten globule-like states from random chain programs. The lattice models span approximately the same number of degrees of freedom as the random chain ensembles. Native-like states of proteins (e.g. the NMR ensemble of structures or a native MD simulation) are distinguished by much smaller over-all geometric distortions - edge lengths of 2 Å versus ~5 Å and lower dimensionality. The effective degrees of freedom to produce RMSD changes in such highly constrained systems are likely to be associated with displacements in a few flexible features such as mobile loops.[30]

To summarize the dynamics results from the villin MD trajectories, there are five interesting time intervals (Table VII). From 0.1 to 10 fs, thermal excitation of all the harmonic or near harmonic vibrational motion dominates as shown by the S-NCA results. This regime would be classically explored with normal mode analysis without solvent coupling. A weak maximum in dimensionality at 100 fs suggests local minima in the near-harmonic regime spanning 0.1-0.2 Å and most likely corresponds to thermally-averaged vibrational motions. Such states have been described by previous workers[30,31]. A major peak in dimensionality at 1-100 ps with an edge length of 0.5 Å is associated with thermally-averaged conformational substates. Troyer and Cohen[30] specifically note that snapshots minimized over these time scales do not fall into the same conformational minima. In the nanosecond regime, transitions occur among these substates that are separated by 2-5 Å. Finally, in the longest villin folding trajectory, there is some evidence for the dephasing of the major folding modes. By exploring the major modes and comparison of MD and NMR results, we estimate that the conformational volume of the native state of villin is about 100 times larger than a single substate and that the unfolded molten globule state contains ca. 10^9 substates, leading to an estimated time to fold villin randomly of ca. 1 millisecond.

Table VII: NCA analysis of important time intervals in protein dynamics.

Time Scale	Assignment	NCA parameters	
		N	A
1 fs	full normal mode excitation	70*	0.003 Å*
100 fs	averaged vibrations	2.8	0.2 Å
1-100 ps	averaged conformations	6.5	0.5 Å
1-100 ns	conformational transitions	4.8	2 Å
1 μs	compact state transitions	10	>5 Å

*-S-NCA parameters listed for 1 fs data.

The probability distributions can be used either directly or with extrapolation to ask about the chance of finding "near native" structures in any particular ensemble. We find that the simple chain-generating models and, by implication, the lattice model ensembles, have a reasonable chance of finding structures in the 4-5 Å regime, but a very small chance of finding structures much closer to an arbitrary structure. Normal distribution curves greatly overestimate this probability, as does a distribution function generated by the simple NCA parameters for a distribution. This latter problem presumably arises from the likelihood that the dimensionality of the space grows inversely to the size of the distortion.

Implications for Protein Folding

Our calculated folding time assuming random exploration of 1 millisecond is 1-2 orders of magnitude larger than the number found by experiment, which is 10-100 microseconds (Duan and Kollman[18] and references within). We offer two alternative explanations. First, one could interpret these two numbers as being essentially the same, particularly given the level of error in our calculation. Assessing the random error is possible by repeating the experiment (i.e. calculating another microsecond of simulation) and by simply carrying through uncertainties in the calculation. There might well also be systematic errors associated with modeling the asymmetric pair distributions with a simple hypercube. We currently have no method for determining systematic errors. If the experimental folding time and predicted random search folding time are the same, within error, the energy landscape seen by the folding protein is not biased toward the native state once collapse has occurred. Alternatively, the discrepancy between our time prediction for random protein folding and experiment may be meaningful and result from the energy landscape of the molten globule directing the protein toward the native state. The logarithmic difference (1-2 in this case) can thus be taken as a folding funnel 'slope'. Thirdly, if we again assume the time difference is real, the speed-up in actual folding could also result from a large transition region surrounding the native state in accordance with the results of Sali et al.[32]. Conformations in this region, while not themselves native, rapidly transition to the native structure. If the search is random, then our results suggest this transition region to be 1-2 orders of magnitude larger than the native state.

The apparent lack of progression to native in the PCA projection of the 1 microsecond simulation reveals the complexity of the energy surface in a manner consistent with Crippen and Ohkubo's[14] model for a protein folding trajectory where winding streams in a mountainous landscape lead to the lake which represents the native state. The streams, however, do not always point toward the lake in an arbitrary coordinate system. These streams (representing folding trajectories) can approach very close to the native state in a Cartesian coordinate based system, as revealed by the PCA projections (Figure 6), while still being quite far from native in the time coordinate. This mountain stream model also explains the lack of correlation between RMSD-from-native and energy score observed for 'decoy' ensembles.[33]

The conformational substate

We treat the conformational substate as the common currency of conformational space and protein dynamics. It is modeled here using only two parameters: dimension (number of dominant degrees of freedom) and edge length (average amplitude of dominant degrees of freedom). The simplicity in description is the main advantage over a more detailed description such as that from normal mode analysis. However, the conformational volume has no intrinsic physical meaning and requires additional assumptions to convert total accessible conformational volume measurements to an estimate of the number of substates.

The more traditional method of conformer accounting, where the size of conformational space is presented as z^n, where z is the number of states per residue and n is the number of residues, also requires corrections for excluded chain volume. Conformer counting based on extrapolation of the number of conformations available to a single chain segment has an additional problem because it does not consider the extent to which non-bonded energy terms determine the shape of the energy surface. For example, the helix-interface shear mechanism of domain closure, which is based on conformational differences observed in experimentally determined structures, illustrates how non-bonded attractions and steric effects can provide the dominant barriers.[34] In this mechanism, helices bump over each other as rigid bodies. Sidechain torsions are important but repacking appears to present the dominant barrier to unhindered shearing. Likewise, the hinging mechanism of domain closure is marked by a lack of packing constraints at the hinge, suggesting sterics to be the dominant modulator of this motion.[35] From MD calculations on myoglobin aimed at sampling multiple conformational states, Elber and Karplus[31] found that substate transitions are perhaps best described by rigid body translations and rotations of substructure subunits. The NCA approach avoids the complexities of defining a structure-based framework for quantifying conformational freedom.

The molten globule and unfolded states

Our current model for the molten globule state is an ensemble of compact regimes. NCA parameters for a single such regime have high dimensionality (ca. 10) and relatively small edge length (ca. <1 Å). Transitions between such regimes presumably lead to the decrease in dimensionality and increase in edge length in the 1-100 ns time intervals. The ultimate uncoupling of such transitions gives the rise in dimensionality in the longest time windows. These observations are consistent with the view of molten globules as partially ordered compact structures with local secondary structure but large-scale tertiary disorder .[6,36-38]

We note that the analysis of conformational volumes produced here yields a change in entropy from native to "unfolded" state of $R\ln(V_u/V_n)$ or 0.9 eu/residue for villin backbone structures. This is close to the estimate from Dill[5] for the transition between folded and <u>compact</u> unfolded states of 1.05 eu/residue. These numbers are much less than the conformational entropy associated with the transition from native to fully unfolded states of

4-6 eu.[39] It is not clear, at this point, that NCA models can treat fully unfolded states in a quantitative manner since none of the structural ensembles studied here produce the very much larger "volumes" associated with the experimental entropy changes.

Conclusions

We have developed a method, NCA, for quantifying the conformational volume given an ensemble of representative conformations. The premise for our method is that conformational space expands exponentially with distance from a reference conformation. The exponent by which space expands implies a dimensionality to the space, which is physically likened to the number of large amplitude, normal degrees of freedom. Our studies employ C_α-RMSD as the distance metric, but the method is general.

Analysis of the folding of villin using NCA on long time MD simulations reveals the mechanism for folding from the molten globule to the native state to occur by near-random exploration of conformational substates. To our knowledge, this is the first study aimed at resolving "Levinthal's paradox" which uses an all-atom model for a protein with a simulation method that explicitly references time.

NCA results with MD trajectories reveal local conformational clustering consistent with the substate model. Using constructed conformational ensembles and high temperature MD we have shown that the excluded chain volume constraint in compact structures is a primary cause of a secondary clustering of conformations at the level of different folds.

NCA was shown to give self-consistent conformational volume measurements for lattice ensembles, however NCA is generally too simple to give meaningful statistics on the RMSD distribution in the low RMSD range. In order to determine the probability of randomly folding a protein less than a given RMSD from native, calculations based directly on the distribution functions must be used.

Acknowledgments

We are extremely grateful to Peter Bacchetti for developing the derivation of the N-cube parameters. We would like to thank Peter Kollman and Yong Duan for providing the coordinates of the villin headpiece subdomain trajectory analyzed in this paper. We thank Lydia Gregoret for making her YARN software available to us. Ken Dill, Charles Brooks, Kenneth Foreman, and Gordon Crippen provided helpful comments. Support for this work was provided by the Pharmaceutical Chemistry Training Grant to DCS and GM-31497 from the National Institute of General Medical Sciences, NIH to IDK.

References

1. Levinthal C. Are there pathways for protein folding? J Chim Phys 1968;65:44-45.

2. Levinthal C. How to fold graciously. In: Debrunner P, Tsibris JCM, Munck E, editors. Mossbauer Spectroscopy in Biological Systems. Urbana, IL: Univ Illinois Press; 1969. p 21-24.

3. Zwanzig R, Szabo A, Bagchi B. Levinthal's paradox. Proc Natl Acad Sci USA 1992;89:20-22.

4. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nature Struct Biol 1997;4:10-19.

5. Dill KA. Theory for the folding and stability of globular proteins. Biochemistry 1985;24:1501-1509.

6. Kuwajima K. The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. Proteins 1989;6:87-103.

7. Cohen FE, Sternberg MJE. On the prediction of protein structure: The significance of the root-mean-square deviation. J Mol Biol 1980;138:321-333.

8. McLachlan AD. Rapid comparison of protein structures. Acta Crystallog sect A 1982;38:871-873.

9. Amadei A, Linssen ABM, Berendsen JC. Essential dynamics of proteins. Proteins 1993;17:412-425.

10. Garcia AE. Large-amplitude nonlinear motions in proteins. Phys Rev Lett 1992;68:2696-2699.

11. Hayward S, Kitao A, Go N. Harmonic and anharmonic aspects in the dynamics of BPTI: A normal mode analysis and principal component analysis. Protein Sci 1994;3:936-943.

12. Balsera MA, Wriggers W, Oono Y, Schulten K. Principal component analysis and long time protein dynamics. J Phys Chem 1996;100:2567-2572.

13. Clarage JB, Romo T, Andrews BK, Pettitt BM, Phillips GN Jr. A sampling problem in molecular dynamics simulations of macromolecules. Proc Natl Acad Sci USA 1995;92:3288-3292.

14. Crippen GM, Ohkubo YZ. Statistical mechanics of protein folding by exhaustive enumeration. Proteins 1998:32:425-437.

15. Samudrala R, Xia Y, Levitt M, Huang ES. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. Proceedings of the Pacific Symposium on Biocomputing 1999;4:505-516.

16. Gregoret L, Cohen FE. Novel method for the rapid evaluation of packing in protein structures. J Mol Biol 1990;211:959-974.

17. Case DA, Pearlman DA, Caldwell JW, Cheatham III TE, Ross WS, Simmerling CL, Darden TA, Merz KM, Stanton RV, Cheng AL, Vincent JJ, Crowley M, Ferguson DM, Radmer RJ, Seibel GL, Singh UC, Weiner PK, Kollman PA. AMBER 5. University of California, San Francisco. 1997.

18. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740-744.

19. McKnight CJ, Matsudaira PT, Kim PS. NMR structure of the 35-residue villin headpiece subdomain. Nature Struct Biol 1997;4:180-184.

20. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rogers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535-542.

21. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins and nucleic acids. J Am Chem Soc 1995;117:5179 -5197.

22. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys 1983;79:926-935.

23. Ryckaert J, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 1977;23:327-341.

24. Kuboniwa H, Tjandra N, Grzesiek S, Ren H, Klee CB, Bax A. Solution structure of calcium-free calmodulin. Nature Struct Biol 1995;2:768-776.

25. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 1998;95:11158-11162.

26. Suezaki Y, Go N. Breathing mode of conformational fluctuations in globular proteins. Int J Pept Prot Res 1975;7:333-334.

27. McCammon JA, Harvey S. Dynamics of proteins and nucleic acids. Cambridge: Cambridge University Press; 1987. p 28-29.

28. Janezic D, Venable RM, Brooks BR. Harmonic analysis of large systems III Comparison with molecular dynamics. J Comput Chem 1995;16:1554-1566.

29. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscape and motions of proteins. Science 1991;254:1598-1603.

30. Troyer JM, Cohen FE. Protein conformational landscapes: Energy minimization and clustering of a long molecular dynamics trajectory. Proteins 1995;23:97-110.

31. Elber R, Karplus M. Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin. Science 1987;235:318-321.

32. Sali A, Shakhnovich E, Karplus M. How does a protein fold? Nature 1994;369:248-251.

33. Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J Mol Biol 1996;258:367-392.

34. Lesk AM, Chothia C. Mechanisms of Domain Closure in Proteins. J Mol Biol 1984;174:175-191.

35. Gerstein M, Schulz G, Chothia C. Domain closure in adenylate kinase: Joints on either side of two helices close like neighboring fingers. J Mol Biol 1993;229:494-501.

36. Balbach J, Forge V, Lau WS, Jones JA, van Nuland NAJ, Dobson CM. Detection of residue contacts in a protein folding intermediate. Proc Natl Acad Sci USA 1997;94:7182-7185.

37. Ptitsyn OB. Structures of folding intermediates. Curr Opin Struct Biol 1995;5:74-78.

38. Pollack L, Tate MW, Darnton NC, Knight JB, Gruner SM, Eaton WA, Austin RH. Compactness of the denatured state of a fast-folding protein measured by submillisecond small-angle x-ray scattering. Proc Natl Acad Sci USA 1999;96:10115-10117.

39. Cooper A. Thermodynamics of Protein Folding and Stability. In: Allen G, editor. Protein. vol 2. Stamford, CT: JAI Press Inc; 1999. p 244.

40. Bevington PR, Robinson DK. Data reduction and error analysis for the physical sciences, 2nd Edition. New York: McGraw-Hill, Inc; 1992. 104 p.

41. Mood AM, Graybill FA, Boes DC. Introduction to the theory of statistics, 3rd Edition. New York: McGraw-Hill; 1974. 238-9 p.

42. Conway JH, Sloane NJA. Sphere Packings, Lattices and Groups. New York: Springer-Verlag; 1988. 9 p.

43. Ferrin TE, Huang CC, Jarvis LE, Langridge R. The MIDAS display system. J Mol Graphics 1988;6:13-27.

44. Diamond R. On the multiple simultaneous superposition of molecular structures by rigid body transformations. Protein Sci 1992;1:1279-1287.

45. Bailey S. The CCP4 suite - programs for protein crystallography. Acta Crystallog sect D 1994;50:760-763.

Appendix. Derivations of integrated probability, v(r), for different distribution functions of conformational space.

N-space model

In the limit of small distances compared to the extent of the smallest dimensions, v(r) in a bounded space approaches the hypersphere volume formula $C(r)^n$. C is a scaling factor equal to the volume of a sphere of radius 1 and dimension n. We can numerically calculate the analogous scaling constant for conformational space, C', by considering a hyperdimensional grid, with edge length and dimensionality (forced to the nearest integer) given by NCA. This gridded n-cube is created to contain as many grid points as DRT ensemble members. Our equation for v(r)

$$v(r) = C'r^n \qquad \text{(A-1)}$$

is first expanded by setting:

$$C' = Cb^n \qquad \text{(A-2)}$$

to give:

$$v(r) = (Cb^n)r^n \qquad \text{(A-3)}$$

where b is the number of points per Angstrom along any axis. C is equal to 4.06 for an 8-dimensional sphere.[42] b is found by solving:

$$(bA)^N = \text{ensemble size} \qquad \text{(A-4)}$$

where A and N are the NCA parameters, edge length and dimensionality, respectively. In our case, N=8, A=8.767 Å, and ensemble size is 10,000, to give b=0.361. C' therefore equals 1.16×10^{-3}. C' was multiplied by 5000 to scale to v(r) calculated from DRT. (While there are 10,000 ensemble members in DRT, only unique pairs are considered.) In the limit of small RMSD, NCA predicts DRT's v(r) to approach:

$$v(r) = 5.8 \times r^{8} \qquad \text{(A-5)}$$

Normal distribution

Cohen and Sternberg[7] suggest that the distribution of RMSD values for a conformational ensemble against a reference structure follows a normal distribution. Our all-pairs RMSD distribution differs only slightly from the type of distribution they consider in that effectively we average over all conformations serving as reference. To calculate v(r) derived from a normal curve, the function:

$$P(r) = 4.9995 \times 10^{5} \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}\right] \qquad \text{(A-6)}$$

[4.9995×10^{5} is a normalizing constant, μ (=9.861) and σ (=1.174) are calculated from DRT's RMSD distribution] is calculated at 0.01 intervals and numerically integrated (starting at -10).

Extrapolation of DRT's v(r)

Assuming the hypervolume formula (Equation A-1) for conformational space, the slope of log(v(r)) versus log(r) is the dimensionality n in the limit of small r where boundary effects become negligible (i.e. where C' is constant).

$$\log(v(r)) = \log(C') + n(\log(r)) \qquad \text{(A-7)}$$

Because C' generally depends on r, the slope of log(r) versus log(v(r)) (the "log-log slope") does not generally yield a dimensionality of direct physical significance. In this paper, our primary use of the log-log slope is for informing the extrapolation of v(r).

Figure 7c plots the log-log slope for DRT, the normal distribution and the n-cube distribution. The divergence at low r between the log-log slopes of DRT and the 8-cube likely reflects the non-cubic nature of DRT's conformational space. At a finer resolution of conformational space (smaller r), more modes become significant. DRT's log-log slope is arguably converging on a number near 68, the total number of degrees of freedom (torsion angles) that impact DRT's C_α RMSD. We extrapolate DRT's log-log slope back to zero RMSD by fitting the curve over a well-behaved portion (4.64 Å to 8 Å) to a quadratic equation. If we ignore the contribution of C' to this slope we can express this fit as the function n(r):

$$n(r) = 68.36 - 11.48r + 0.5031r^2 \qquad \text{(A-8)}$$

The y-intercept is very close to 68, thus supporting our interpretation of the log-log slope at small r. We can now determine the dependence of log(C') on r by rearranging Equation A-7 to give:

$$\log(C') = \log(v(r)) - n(\log(r)) \qquad \text{(A-9)}$$

Log(C') was plotted using Equation A-8 for n and the observed function for v(r). We fitted log(C') to a quadratic function over the interval from r = 3.82 Å to r = 8 Å yielding the equation:

$$\log(C') = -41.35 + 6.754r - 0.2123r^2 \tag{A-10}$$

If a Euclidean model, with uniformly distributed points, perfectly described DRT's conformational space, we would expect log(C') at r=0 (i.e. -41.4) to correspond to that of a 68 dimensional hypersphere. We test this by first taking the log of both sides of Equation A-2:

$$\log(C') = \log(C) + n(\log(b)) \tag{A-11}$$

Log(C) equals -21.6 for a 68 dimensional hypersphere.[42] Estimating the grid spacing for the ensemble in a 68 dimensional space can be very crudely approximated by forcing a dimension of 68 on DRT's RMSD distribution and calculating A. This gives A equal to 4.08 Å, and, by Equation A-4, b equal to 0.28 Å with n equal to 68. The right side of Equation A-11 thus equals -59.2 compared to -41.4. We currently have no method for assessing the significance of this discrepancy.

Combining equations A-7, A-8 and A-10 we can extrapolate v(r)'s approach to zero RMSD:

$$v(r) = 10^{y} \tag{A-12}$$

where y equals $\left[\left(-41.35 + 6.754r - 0.2123r^2\right) + \left(68.36 - 11.48r + 0.5031r^2\right)(\log r)\right]$.

Protein Folding as Biased Conformational Diffusion

David C. Sullivan and Irwin D. Kuntz

Department of Pharmaceutical Chemistry
University of California at San Francisco
San Francisco, California 94143-0446

Summary

Analysis of molecular dynamics trajectories for a small protein, the villin headpiece subdomain, shows that its internal dynamics can be modeled as a random walk on a bounded high dimensional lattice where each lattice point corresponds to a protein conformation. The time evolution of the dynamics trajectory can be described as a three stage process. The first stage, over a very short time scale of 10 fs, is consistent with an all-atoms ballistic flight with an average path length of 0.068 Å. The apparent conformational displacement "velocity" of 680 meters/sec is of similar magnitude to free flight monatomic carbon gas at room temperature. The second stage is unobstructed random walk diffusion in a high dimensional space for ~300 fs, average displacement of 0.4 Å. In the third stage, at longer times, "harmonic-like" barriers confine the conformational diffusion to a few large amplitude modes. We use this description to develop a lattice model that can reproduce the time-dependence of the root-mean square displacement (RMSD) along the dynamics trajectory with surprisingly few parameters.

The lattice model allows exploration of specific energy landscapes on the folding of the villin headpiece subdomain. For example, we test a simple funnel landscape with a Monte Carlo search in which displacement is biased toward a corner of the lattice-space representing the native state. An energy difference of ~15 kcal/mol between native and unfolded states is sufficient to give a reasonable folding time of 10 μs with two-state kinetics. To our knowledge, these are the first protein folding simulations that successfully predict the protein folding dynamics using a model with explicit time representation, a stochastic search and without a native-state specific potential function.

Keywords: Protein folding, energy landscape, lattice models, molecular dynamics, folding kinetics.

Introduction

Lattice models have been critical for understanding protein folding. Examples include development of energy landscape theories such as the "folding funnel",[1] in which a global energy bias directs the folding protein toward the native state, and the importance of the hydrophobic effect in explaining collapse in the initial stages of folding.[2] Simplified atom representation permits simulation of events outside the range of computationally demanding methods such as molecular dynamics (MD) treatments of all-atom models with explicit solvent representation. Lost in most minimalist models is explicit time representation. Its absence ultimately limits the level of kinetic interpretation derivable from these models. In this paper we develop a lattice model that is parameterized from MD simulations over femtosecond (fs) to microsecond (μs) time scales. With this lattice, we can exchange explicit atom representation in favor of explicit time dependence. Each point in the high dimensional lattice represents a single conformation although there is no direct mapping between points and actual structures as explained below.

Our lattice corresponds to the ensemble of compact unfolded conformations for the 36-mer villin headpiece subdomain protein. The lattice is parameterized using the one microsecond MD simulation by Duan and Kollman[3] on villin. While their simulation does not successfully fold the protein to its experimentally determined native state, it does search a conformational space that contains the native state,[4] arguably providing a reasonable model for the conformational readjustments associated with the folding process. We use additional simulations on misfolded villin to examine time scales shorter than possible with the 1 μs simulation which saved conformations only every 20 picoseconds (ps).

An important question in protein folding is whether there is a bias of the global energy landscape that favors the native state. Such a bias would be difficult to assess by MD. The Duan and Kollman[3] 1 μs simulation, the longest to date on a solvated protein, does not reveal any obvious downhill trend in energy after initial collapse. This lack of folding is not thought to reflect a force-field limitation. Free energy calculations using the AMBER molecular mechanics force-field[5] with an additional solvation free energy term (Poisson Boltzmann/surface area[6]) show the native state to be 15-35 kcal/mol lower in energy than snapshots taken from the 1 μs simulation.[7] In this paper we show that an energetic difference of this magnitude, expressed as a constant slope between native and unfolded regions of our conformation space lattice, yields reasonable folding times while unbiased models fold much too slowly.

The primary tool we introduce for studying protein dynamics is exploration of a lattice representing conformation space that can be explored by diffusional processes (i.e. random walks). Distances traveled on this lattice conform to the time-dependent conformational displacement of a protein as measured by the root mean square deviation (RMSD) of Cα positions after optimal pairwise superpositioning.[8] Our conformation-space lattice has relatively few parameters. They are: 1) the inter-point spacing, 2) the time associated with a single move on the lattice, 3) the number of dimensions of the lattice, 4) the length of dimensions, and thus the gross size of the lattice, and 5) the roughness of the energy landscape. To derive these parameters we analyze specific features of the trajectory recast as the time auto-difference (TAD) function, which is roughly equivalent to the time autocorrelation function but uses pairwise RMSDs instead of a similarity measure. Specifically, point spacing and time equivalence come from analyzing the transition between the free-flight and diffusive time scales. The displacement variance at short time separation determines the total number of dimensions of the lattice (see below). RMSDs between pairs of random conformations taken from the state of interest (e.g. the compact unfolded state of villin headpiece) determines the size of the lattice. This is done by requiring that the distribution of RMSDs between random protein conformations and the distribution of distances between random points on the lattice be similar. Finally, estimates of energetic roughness are taken from the increase in the RMSD variance at longer time scales. We will model landscape roughness with a set of minima (traps). Our trap parameters were found by iteratively adjusting trap size, depth, and total number to give good fit between random walk and MD TAD functions.

Methods

Molecular Dynamics.

All molecular dynamics calculations used AMBER.[9] Details of the solvated MD simulations have been described elsewhere.[3,4] For in vacuo simulations, a distance-dependent dielectric function was used for electrostatic calculations and a non-bonded cut-off of 10 Å was imposed. The Cornell et al[5] force field was used. The non-bonded pair list was updated every 100 steps. A time step of 1 fs was employed, except for the very short simulations where a time step equal to the time interval between saved snapshots was used. All bonds involving hydrogen atoms were constrained using the SHAKE algorithm.[10] The starting structure (PDB code 1VII[11]) was minimized and equilibrated first at 300 K for 1 ns and then for another 100 ps at the temperatures reported (10 K, 80 K, 150 K, 200 K, 250 K, 300 K, and 500 K).

The primary entity we extract from MD is a time auto-difference (TAD) function, which is the spread of RMSD values for conformations separated by a particular time (Figure 1a). The TAD function is similar to a time autocorrelation function with multiple time origins, with the main exception that we are measuring a difference (RMSD) instead of a similarity (dot product). Since we are interested in examining behavior over several orders of time, we wish to spread the calculated RMSDs uniformly over a logarithmic time scale while keeping the total number of RMSD measurements at a reasonable level. To meet these needs, the TAD function for the solvated folding MD studies was calculated as follows. A trajectory of length $10 \times t$ was divided into 10 consecutive sections of length t. An ensemble of 25 equally spaced snapshots from each section was extracted and all pairwise RMSDs calculated. The time separation between pairs of structures was noted. For example, a single 1 ns trajectory with 250 snapshots would yield 3000 RMSDs [$300 = (24 \times 25)/2$ from each 100 ps section, multiplied by 10 sections] with time separations ranging from 4 ps to 96 ps. MD trajectories range in length from 10 fs (0.04 fs separation between successive snapshots) to 1 μs in approximately one-half log time units (i.e. 10 fs, 30 fs, 100 fs . . . 1 μs). The shortest time scale RMSDs (< 0.1 fs) suffer from rounding errors since our coordinate files are saved only to the thousandth of an Angstrom. In sum, 48,300 RMSD measurements were calculated from these studies. RMSDs for the in vacuo simulations were calculated by the same approach. For each temperature, five trajectories were calculated (100 fs, 1 ps, 10 ps, 100 ps and 1 ns) which were each divided into 5 sections of 50 equally spaced snapshots. All pairwise RMSDs in each section were calculated yielding 1225 RMSDs per section, 6125 RMSDs per trajectory, and 30,625 for all five trajectories. For further details see Sullivan and Kuntz.[4]

RMSDs were calculated using only alpha-carbons by independently overlaying all pairs of structures using the McLachlan algorithm[8] as implemented in the program ProFit[12] followed by taking the root-mean of the squared atom displacements.

Random Walk Lattice Model.

The first issue in developing our lattice model is defining the set of allowable steps. We use an N-dimensional hypercubic lattice with steps confined to neighboring lattice points. Our basic move is between points along the N-dimensional diagonals to simulate the MD model where all atoms move simultaneously in a protein. Steps in our simulations are attempted along all dimensions at every time step. As expected, we find tighter correspondence to the behavior of MD RMSDs with this diagonal step compared to steps along a single randomly chosen dimension. Since there are 2n possible steps along a single randomly chosen dimension where n is the total number of dimensions, while there are 2^n possible diagonal steps, for spaces where $n \geq 3$ there are far more possible diagonal steps than possible orthogonal steps. The chance of back-stepping is much lower if diagonal steps are taken. Lattice points adjacent along an all-dimension diagonal are spaced 0.07 Å. The spacing of points along a single dimension will thus be less than 0.07 Å (for lattices with more than one dimension) and depend on the number of dimensions. In the 65-dimension lattices used here, points have a spacing of 0.00868 Å along any axis. Each step has a time equivalence of 10 fs. For time scales shorter than 10 fs, we linearly interpolate between trajectory points.

Number of dimensions.

Having established, for a particular system, the diagonal step size and the fundamental time per step, we turn to an analysis of the dimensionality needed to simulate the RMSD-time data. The number of dimensions of the lattice determines the displacement variance after a given time. For example, in the limit of an infinite number of dimensions, displacement variance goes to zero as each step will be in a direction normal to all previous steps.[13] That is, for a walk with unit-length steps in a space with an infinite number of dimensions, displacement simply equals the square-root of the number of steps taken with zero variance. In spaces with a finite number of dimensions, there will be a distribution of possible displacement magnitudes for a given number of steps. We use the variance observed in MD RMSDs for a given time separation to determine the number of dimensions of the lattice.

Barriers associated with conformational substates should also affect this variance, so we limit this analysis to a short time scale where barrier effects should be minimal.[14] Specifically, we performed 12 step random walks, equal to 120 fs, with diagonal steps of length 0.07 Å. The 610 MD RMSD measurements at 120 fs have an average of 0.26 Å and standard deviation of 0.021 Å. The 610 random walks in 65 dimensions have an average displacement of 0.24 Å and standard deviation of 0.021 Å. We use 65 as the number of dimensions for our lattices. Based on measured variance in the standard deviation for random walk displacements, the number of dimensions can be changed by ±5 without changing our results.

Conformational space boundaries.

The next concern is whether to treat all these dimensions equally. In earlier work[4] we found that the villin trajectory could be best represented as sampling a few large amplitude modes and many small amplitude modes. Using a hypercube model related to principal component analysis that we termed N-Cube analysis (NCA),[4] we found the number of large displacement modes for the unfolded state[3] to be ~10, while the small scale motions had a dimensionality of ~70. To model these large scale and small scale changes, we allowed the "large" lattice dimensions to contain hundreds of points while the "small" lattice dimensions contain tens of points (see below). At the lattice edge, exiting steps are disallowed so that no move occurs in that dimension on that step. Moves in different dimensions are treated independently. So, for a given step, there may be displacement in some dimensions with no displacement in other dimensions yielding a displacement magnitude less than 0.07 Å. Starting points are randomly chosen from within the allowed space.

Trap parameters.

We add roughness to the energy landscape with local "traps". A trap is defined as an energetic barrier that hinders exiting the defined region of the lattice. Traps have the effect of slowing lattice exploration for part of the time. Analogous to a pit in the ground with vertical walls, there is no impediment nor bias to "falling" into the pit. Also, displacement within the trap is not slowed compared to outside the traps. In contrast to a two dimensional trap in the ground, our traps take the form of 12-dimensional hypercubes (i.e. they are located in the "large" dimensions, see below). There is no bias to enter a trap, but attempts to exit the trap are permitted only a fraction of the time in accordance to the energetic depth of the trap, expressed as a Boltzmann probability and calculated once per time step. If "escape" is rejected on a particular time step, all moves except exiting components of the move are accepted. A variety of trap parameters were explored. They were evaluated by comparing random walk simulations to the time dependence and variance of the RMSD from MD studies. The following parameters were chosen: Traps are 6.86 kcal/mol deep. A single trap occupies 30 points in all 12 "large" dimensions. Traps are spaced by 27 points. 30% of each dimension at one end is free of traps. In energy-biased random walks, the trap-free region is at the high energy end of the dimension (see below). In sum, 0.0015% of all lattice points are part of a trap.

To explore the kinetic implications of our "conformational" lattice, we defined three sets of parameters:

Parameter set A: Bounded lattice with minima traps - "Molten Globule State".

There are 65 dimensions of which 12 allow large displacements. The 12 large dimensions are 600 points long (5.21 Å). The remaining 53 dimensions permit local displacement with a maximum displacement of 30 lattice points (0.260 Å). The average distance between two randomly selected points is 7.3 ± 1.3 Å. Trap parameters are taken from above.

Parameter set B: Trap-less Random Walk - "Molten Globule State".

Set B random walks use the same parameters as set A, however the minima traps have been removed.

Parameter set C: Reduced Size Trap-less Random Walk - "Native State".

In this lattice there are only 5 large displacement modes that are 176 points long (1.57 Å). The remaining 60 dimensions are 30 points long. The average distance between any two points is 1.6 ± 0.3 Å. For comparison, a structurally diverse set of 11 NMR conformations of villin[11] has an N-Cube Analysis (NCA)[4] dimensionality of 4.8, NCA dimension length of 1.9 Å, and average pairwise RMSD of 1.7 ± 0.4 Å.

Monte Carlo sampling of lattices.

Protein folding was simulated by designating a corner of the 12 long dimensions in the molten globule lattice as 'native'. The size of the native region was set equal to the hyperrectangle formed by the parameter set C lattice, with five medium length edges and seven short length edges. Energy bias was only added to the large displacement modes since small displacement modes of the folding lattice are already congruent with the native state small displacement modes. Along all biased axes, an energy difference, *dE,* was imposed between successive points. Thus, the total energy difference between the termini of an axis simply equals the number of points on the axis, minus one, multiplied by *dE.* This quantity, multiplied by twelve, is the maximum energy difference between any two points. The time at point of first passage into native is taken as the folding time. The actual energy drop on any folding run on average will be slightly less than half the largest energy difference, since the initial point will be at the energy midpoint, on average, and the first passage into native will be slightly above the lowest energy point. For simplicity, we designate the difference between the lattice midpoint and the lowest energy point as ΔE. Actual energy differences traversed in our folding simulations are ~0.9 ΔE reported here. Reported ΔE values do not consider local minima traps. For any given lattice, ΔE is 3600 times larger than *dE.* Monte Carlo searches use the Metropolis acceptance criterion[15] where an uphill move within the lattice bounds is accepted only if a random number on [0,1] is less than exp(-*dE*/kT). Moves in different dimensions are treated independently.

Results

Our first task is to establish the diffusive character of protein internal motion. We first analyze the very short time behavior in terms of a mean free path length and collision interval using MD. From these basic diffusion parameters, we create a random walk model for conformational diffusion on a high dimensional lattice that represents conformational space. The measured mean free path length defines the grid spacing on the lattice. The collision interval provides the time equivalence for individual steps in the random walk. We vary additional lattice parameters, as described in the Methods, so that random walk displacement as a function of time reproduces the MD TAD function. Finally, we extend our model to protein folding by adding a funnel energy landscape to the lattice. We replace the random walk mechanism with a Monte Carlo search for a "native" corner in the unfolded conformational space. We calculate folding times as a function of the folding funnel slope.

Diffusive behavior of protein dynamics.

Figure 1a plots the MD TAD function. This raw form of the TAD function suggests transitional times exist. These characteristic time intervals are more clearly shown by simple data transformations, below.

Dividing displacement by time separation gives the magnitude of a conformational "velocity". Constant conformational speed is observed up to ~10 fs (data not shown). Villin undergoes a 0.068 Å displacement in 10 fs (Figure 1a), giving a speed of 680 m/s. For comparison, ideal gas at 300 K with the mass of monatomic carbon would have a free-flight RMS speed of $\sqrt{3RT/M} = 789\ m/s$, where R is the gas constant, T is the absolute temperature and M

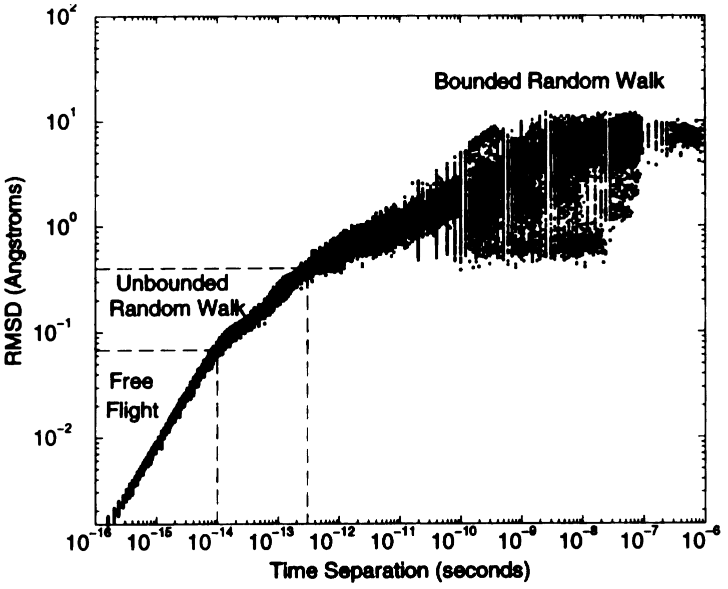


Figure 1a

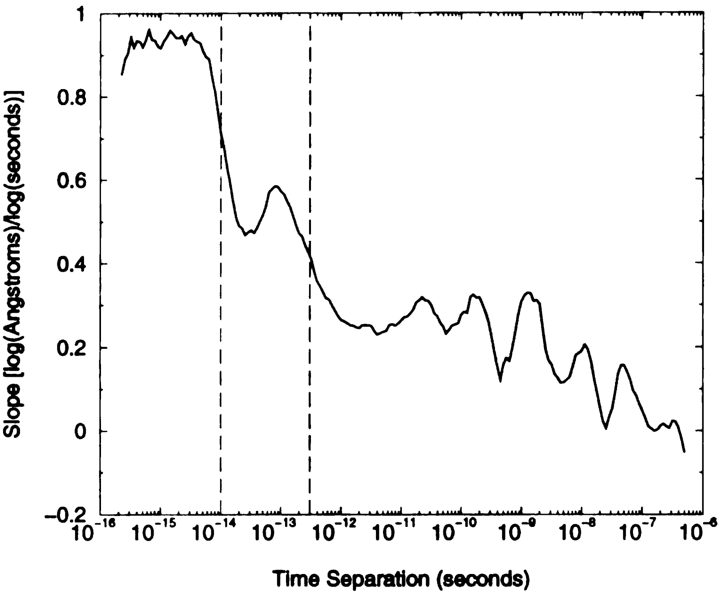


Figure 1b

Figure 1 (A) RMSD time auto-difference (TAD) function from solvated MD simulations on misfolded villin and (C) on *in vacuo* MD simulations on the native structure. The logarithmic slope of the RMS average RMSDs for (B) solvated misfolded and (D) *in vacuo* native villin. For the native simulations, only RMS average values are shown for each temperature. For all slope calculations, RMS average displacements were first calculated with a sliding time window of 0.1 logarithmic units (base 10), followed by slope calculation over 0.5 logarithmic units by the method of determinants.[36] Reference lines are given at 10 fs and 300 fs for (A) and (B). *In vacuo* simulations were performed at 10 K (●), 80 K (○), 150 K (■), 200 K (□), 250 K (▲), 300 K (Δ), and 500 K (◆).

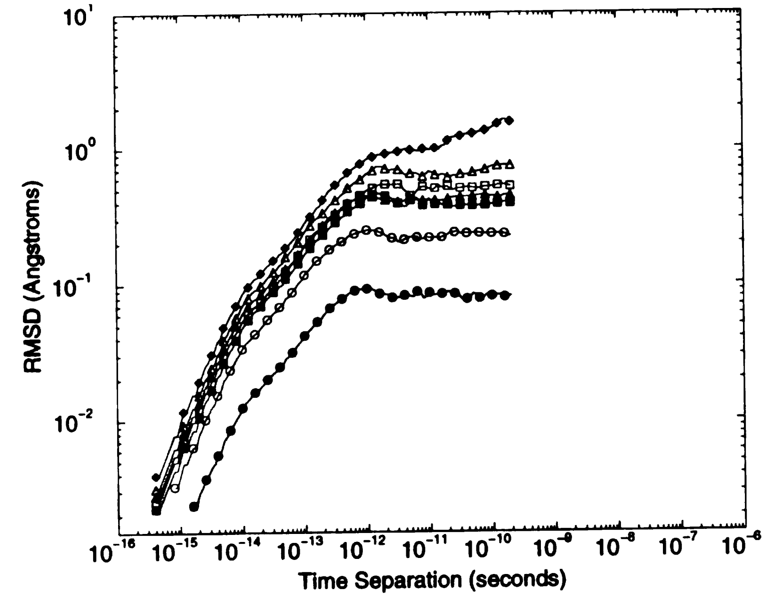


Figure 1c

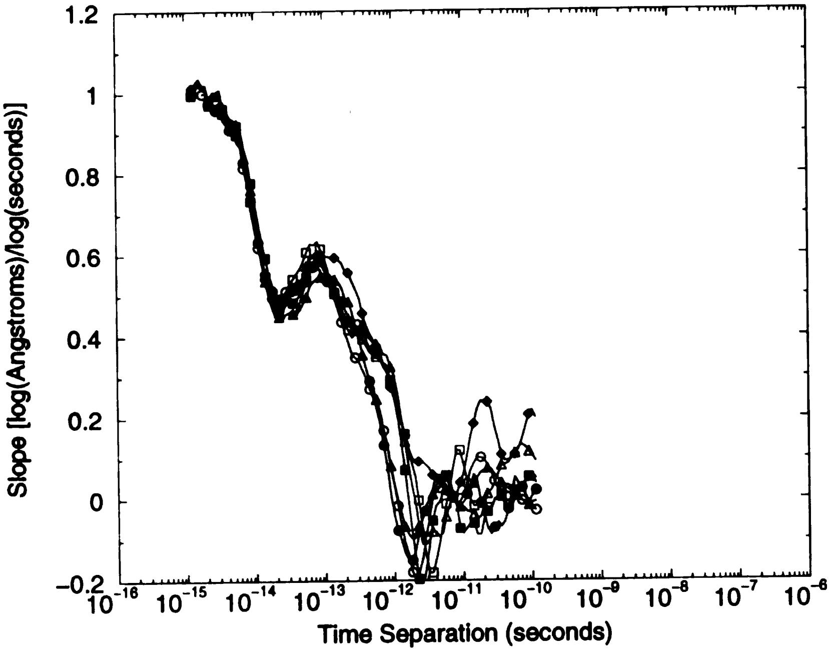


Figure 1d

is the molar mass. The 10 fs "free flight" has a time length comparable to a single C-C bond vibration (~33 fs). The time observed by RMSD analysis is less since conformational displacement originates from averaging effects of out-of-phase vibrations.

The transition from the very short time free-flight regime to simple diffusion is indicated by the change in slope of RMSD vs. time, in double logarithmic form (Figure 1b). The initial slope of log(displacement) vs. log(time) is unity as expected for free-flight. The drop in slope to ~0.5 starting at ~10 fs (Figure 1b) suggests a transition to an "unbounded" random walk regime. Thus villin's dynamics up to ~300 fs can be approximated as a simple diffusion process composed of very short free-flight steps. From this description, the diffusion constant, D, is calculated from the Einstein-Smoluchowski equation, $D=\lambda^2/2\tau$, where λ is the mean free path length (0.068 Å) and τ is the time step (10 fs), giving a value of 2.3×10^{-9} m^2/s.

The decrease in slope beyond ~300 fs (Figure 1b) presumably arises from barriers slowing the diffusion process. What is the nature of these barriers? Multiple TAD functions (Figure 1c) calculated from *in vacuo* MD simulations on the native structure of villin,[11] each for a different temperature, have slope functions (Figure 1d) that are essentially independent of temperature over the observed time range, 1 fs to 100 ps. Further, the mean squared displacements (i.e. $RMSD^2$) correlate strongly with temperature for a given time separation, with r^2 of 0.998 at 100 fs and dropping only to 0.91 at 100 ps. This correlation is consistent with harmonically bound motions dominating the displacement behavior at short time scales, as expected. This time range would include even the lowest frequency normal modes ($\nu \sim 2$ cm^{-1}) associated with coupled vibrations in proteins.[16]

Lattice-based Random Walk Simulations.

Random walks on lattices using optimized parameters (set A) give displacement magnitudes with good fit to the MD RMSD time distribution function (Figure 2). In this set of simulations, we added an equilibration period of 100 ns (10^7 steps) before tracking displacement and time. Figure 2b plots the means for the MD RMSD time distribution function and for displacement in the parameter set A random walks. Figure 2c plots the standard deviations of the logarithms of the same two data sets. There is clearly a tight correspondence between the random walk and MD results. Even the variance (Figure 2c) is reproduced in a semi-quantitative manner.

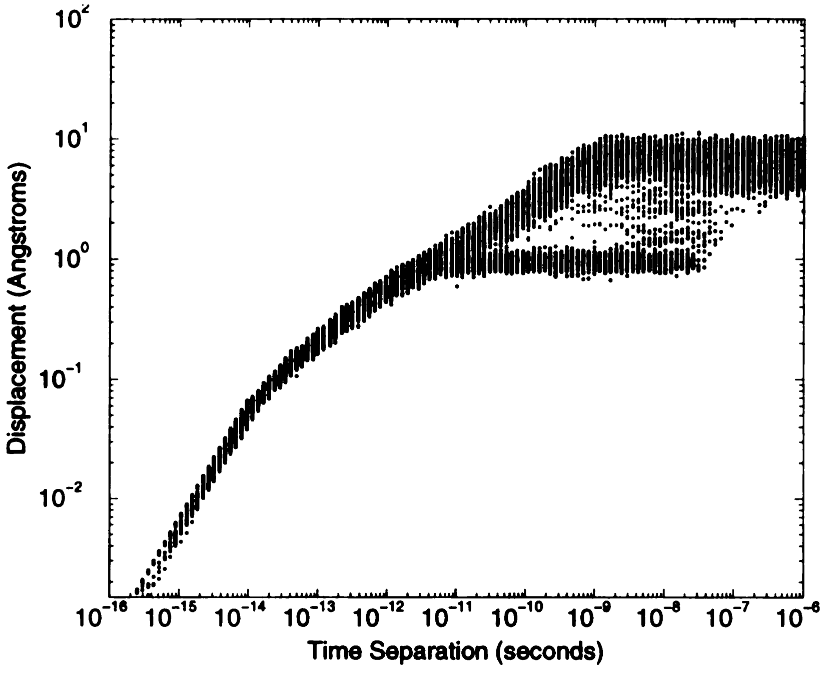


Figure 2a

Figure 2 (**A**) Random walk TAD function on a 65 dimensional lattice with minima traps (set A) and the (**B**) mean and (**C**) standard deviation for random walk displacements (solid line) compared to values for solvated misfolded MD on villin (dashed line). Mean and standard deviations were calculated using a 0.1 logarithm unit sliding window for MD and at discrete time points for random walk data. Standard deviations are calculated on the logarithms of the displacements.

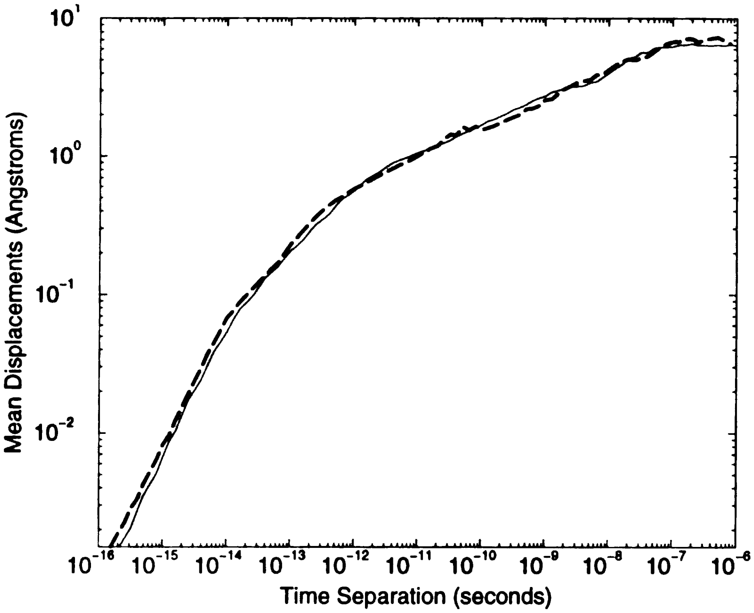


Figure 2b

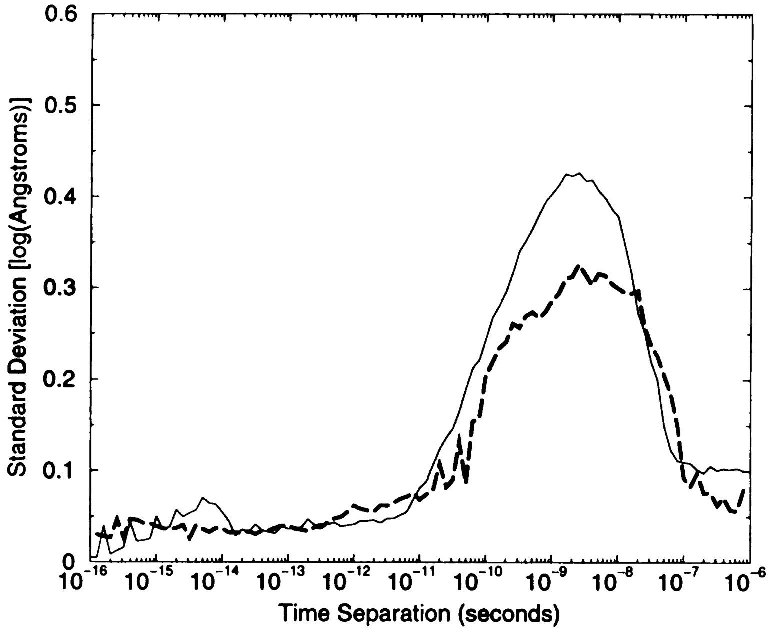


Figure 2c

High dimensional random walks do not explore each dimension at an equal rate even when the dimensions are geometrically equivalent.[13] A simple way to monitor the extent of exploration is to calculate the principal radii of gyration (the variance along the principal axes[13]) as a function of time on the path drawn by the trajectory through the conformational space. On bounded lattices, given enough time, the radii of gyration should converge to values reflecting the extent of each lattice dimension. The convergence time is a property of the lattice and the random walk parameters and is of interest to us because it establishes a time frame for effective sampling of the lattice. If we ignore traps and barriers, we can establish a lower bound on the time required for sampling the "molten globule" and "native" lattices by following the radii of gyration in each principal axis using parameter sets B and C (see Methods). For the molten globule lattice, 1-10 μs are required for the bounds of the lattice to impose convergence (Figure 3a). The same level of convergence occurs at ~100 ns in the smaller, native sized lattice (Figure 3b). At shorter times, variance among principal axes extents simply reflects the stochastic nature of the random walk trajectory rather that an enduring quality of the "energy landscape". As these calculations are performed on energetically flat landscapes free of traps and conformational substate barriers, the convergence times should be taken as underestimates of what would be seen in MD trajectories. These results are consistent with earlier observations that show that short MD trajectories (235 ps) do not yield stable principal moments.[17]

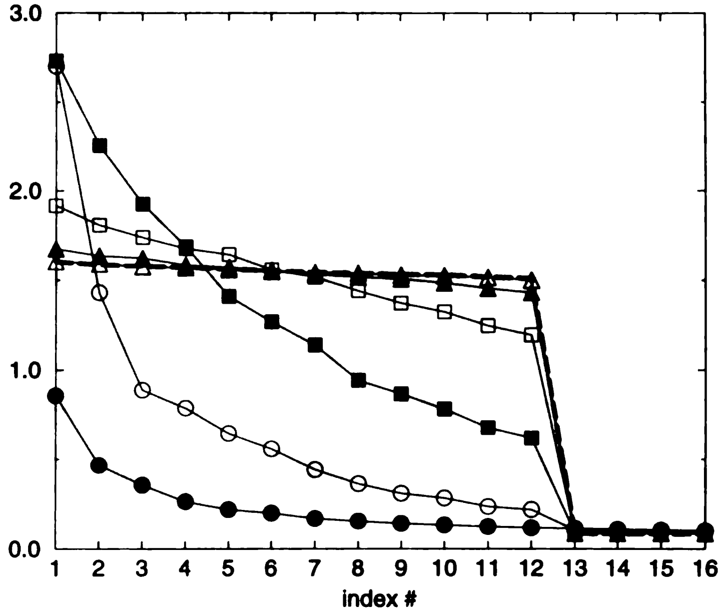


Figure 3a

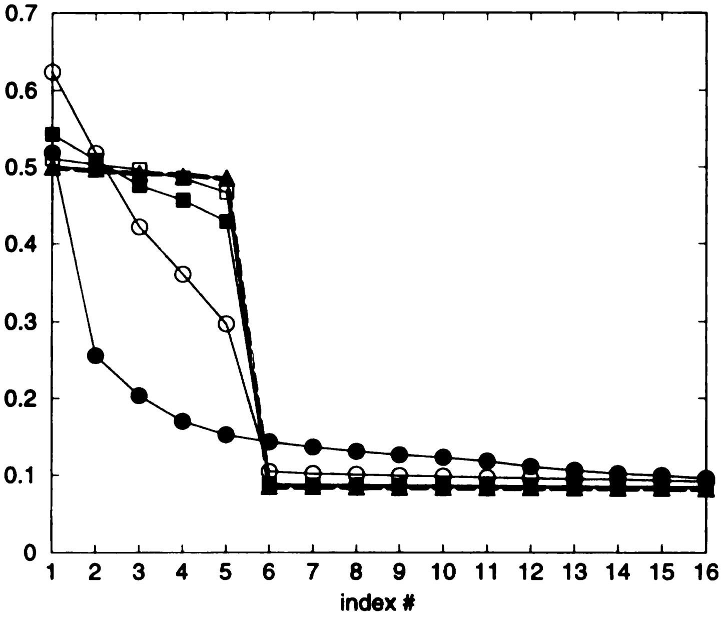


Figure 3b

Figure 3 Principal radii of gyration spectra for (**A**) "folding" (set B) and (**B**) "native" (set C) random walk trajectories. Random walks were performed in triplicate and the mean radii of gyrations are plotted. Simulation lengths are 100 ps (●), 1 ns (○), 10 ns (■), 100 ns (□), 1 μs (▲), and 10 μs for folding only (Δ). The dashed line gives the spectra for 10,000 points randomly placed on the lattice to serve as reference for limiting spherical symmetry. Random walk radii spectra likewise use 10,000 points spaced uniformly along the trajectory.

Energy-biased random walks: Protein Folding.

Our lattices are excellent systems for testing which energy landscapes are consistent with experimentally determined folding times. Here we focus on simplified funnel landscapes, though other landscape models such as a Go potential[18] and pathway models[19] could be explored. First, the dependence of folding time on funnel slope was determined (Figure 4). Since the minima traps are arguably the most *ad hoc* parameter in our simulations, we present results on lattices without minima traps (set B lattice parameters) to gauge their effects. In general, the traps slow folding about two orders of magnitude for a given energy difference (Figure 4).

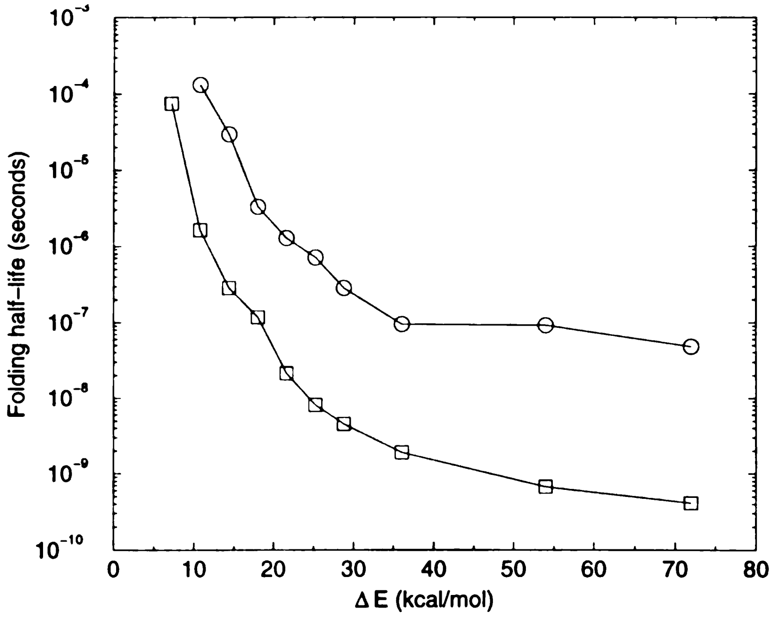


Figure 4 Folding half-times as a function of ΔE, the energy difference between the mean lattice energy (i.e. the energy of the lattice's center point) and the global minimum for optimized (set A) parameters (○) and for trap-free (set B) parameters (□). For each energy value, 20 simulations from random starting points were performed, with the point of first-passage into the native state being the folding time for that run. The time for one-half of the 20 runs to complete is plotted.

Two-state folding behavior is a hallmark of most fast folding single domain proteins.[20,21] We tested for two-state folding behavior on our lattice by performing many (>100) folding simulations for a given set of conditions and checking for a linear drop in the logarithm of the unfolded population versus time (data not shown). We observe two-state folding behavior for $\Delta E \leq \sim 1000$ kcal/mol with set A lattice parameters (with traps) and $\Delta E \leq \sim 25$ kcal/mol with set B lattice parameters (no traps). For both sets of parameters, two-state folding behavior begins when the folding half life is greater than ~10 ns. For larger energetic biases, folding times appear to become proportional to the distance between the starting conformation and the native state.

Preliminary dynamic NMR measurements suggest a folding time for villin headpiece subdomain on the order of 10 μs (MingHui Wang, Liliya Vugmeyster & Daniel P. Raleigh. Personal communication.). Figure 4 shows that an energy gap of ~15 kcal/mol gives a folding time of ~10 μs. Figure 5a gives the distance to the global minimum for a conformation in a single folding simulation with $\Delta E = 15$ kcal/mol ($dE = 0.00417$ kcal/mol). Since we assume the native state to be a hyperrectangle, there are some lattice points closer to the lowest energy point (and lower in energy) than some native points, which explains why the point of first passage into native is not necessarily the lowest point for a simulation. The interesting result is that the downhill trend (in energy and displacement) is lost in the noise even though the bias significantly shortens the time to folding. Figure 5b plots the same information as in Figure 5a except that the trap-free parameters are used (set B). In both cases, an entropic barrier slows progression to native at ~5 Å from the global minimum. The traps in parameter set A slow the folding process but do not qualitatively change the progression pattern.

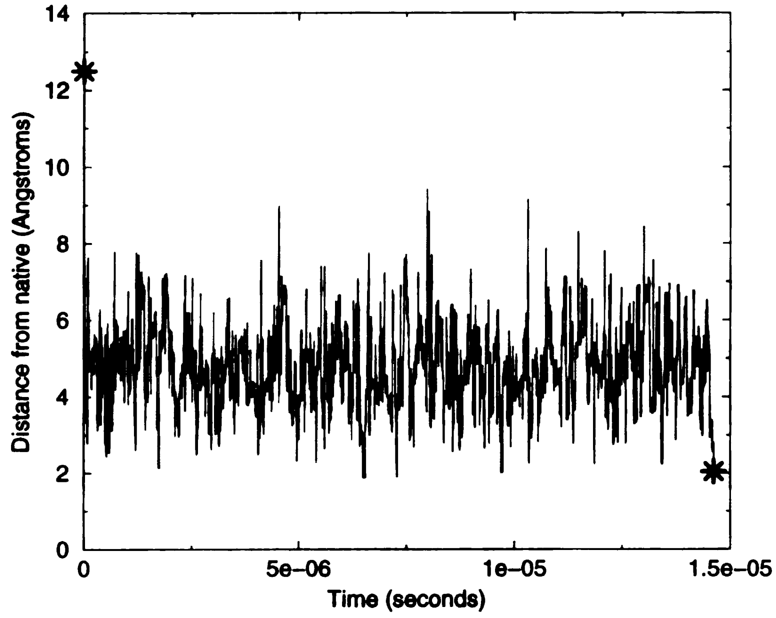


Figure 5a

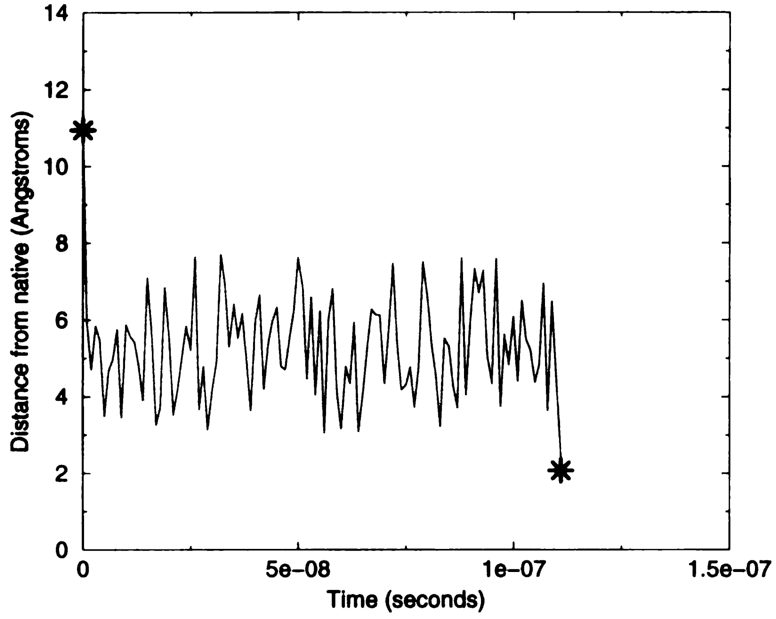


Figure 5b

Figure 5 The displacement from the global minimum is given as a function of time for folding simulations on a lattice with ΔE = 15 kcal/mol using **(A)** optimized parameters (set A) and **(B)** trap-free parameters (set B). The final and initial conformational distances are marked with a star.

Experimental free energies of unfolding further constrain theoretical models for a folding system. Denaturation studies on villin headpiece show the native state to be more stable than the unfolded state by about 3 kcal/mol.[22] Free energies of unfolding on the lattice $\Delta E = 15$ kcal/mol (set A parameters) were calculated by measuring the equilibrium constant for 200 μs simulations. Six simulations were run, three starting from the native state and three starting in the unfolded state. These calculations converged on free energy differences of 10.0 (± 0.3) kcal/mol in favor of the unfolded state. This ~13 kcal/mol discrepancy (10 kcal/mol in favor of the unfolded state versus 3 kcal/mol in favor of the native state) can be explained as the extra stability that would be gained upon formation of the native state from the transition state. Here we are defining the transition state as simply the set of 11-dimensional hyperplanes that divides the native and unfolded states in our model. Including a 13 kcal/mol drop at the native/unfolded transition zone could correct for this thermodynamic discrepancy without affecting our first-passage folding times. This precipitous drop in energy at the transition zone would be in agreement with existing energy landscape models (see Discussion). Figure 6 diagrams a quantitative energy landscape of villin headpiece in the typical reaction coordinate form.

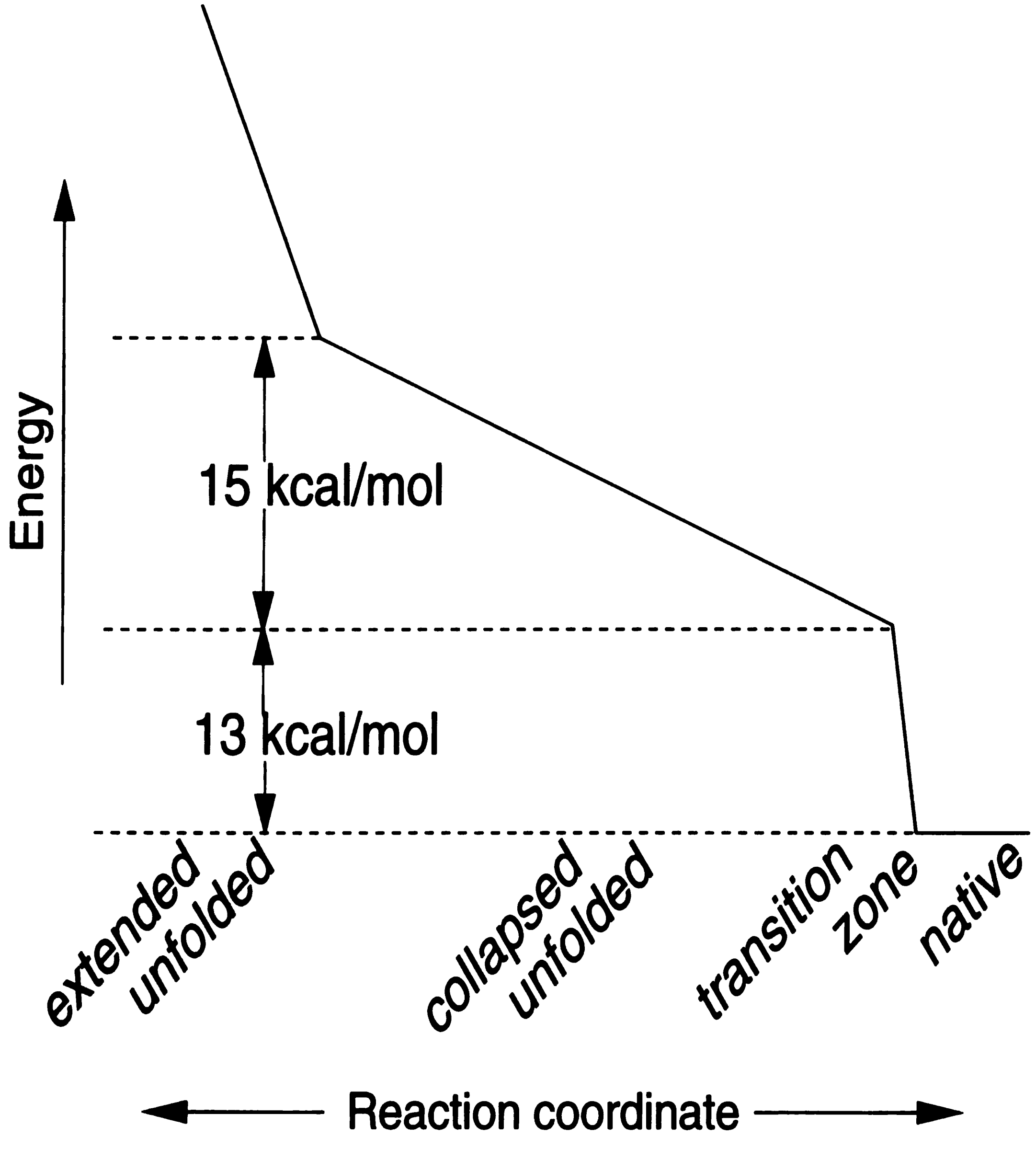


Figure 6 Reaction coordinate for villin headpiece subdomain that conforms to experimental kinetic and thermodynamic constraints assuming a funnel landscape .

Discussion

Short-time dynamics.

The basic time step of our random walk models (~10 fs) and the number of dimensions required (~65) are consistent with a physical mechanism based on bond vibrations. While the total number of internal degrees of freedom for the C_αs of villin (102) is similar in magnitude to the number of lattice dimensions for sub-picosecond time scales, bonding restraints between successive C_αs may explain why the best-fit number of dimensions is closer to 2N "active" modes rather than 3N.

The diffusion constant (2.3×10^{-9} m^2/s) calculated early in the time evolution agrees with values for very small molecules, such as methane in carbon tetrachloride (2.9×10^{-9} m^2/s) and water in water (2.3×10^{-9} m^2/s).[23] This is not surprising given that the Cαs are of similar mass as methane or water and are undergoing thermal motion in a medium (the protein) of similar density to common solvents. This contrasts with the much lower apparent diffusion constants of ~5×10^{-11} m^2/s between residues separated by tens of residues in denatured proteins for microsecond time scale motions.[24,25] In these larger displacement motions, constraints arising from solvent friction, internal friction and excluded volume become significant.

Energy bias.

The energy difference necessary for 10 μs folding (~15 kcal/mol) is on the low end for the energy difference between native and compact unfolded ensemble calculated by Lee et al[7] of 15-35 kcal/mol. Our thermodynamics calculations likewise show the free energy of unfolding to be ~13 kcal/mol too low, compared to experiment. Together, these results suggest that the energy gradient steepens significantly at the unfolded/native transition zone.

This would give the native state additional stability compared to our simple uniform slope model and would affect only thermodynamics but not first-passage folding times. Experimental thermodynamics studies suggest many small fast folding proteins to have a transition state that is an ensemble of slightly expanded and distorted native-like structures.[26] MD studies of unfolding pathways for small proteins also support the model of a transition state ensemble that is near-native, yet heterogeneous in that particular ensemble members from different MD trajectories are equally (dis)similar to each other as to the native state.[27,28] Free energy (intraprotein energy plus solvation free energy) calculations on 24 unfolding MD simulations of chymotrypsin inhibitor 2 by Lazaridis and Karplus[29] show the final ordering of structure to the native state, measured by the fraction of native contacts, is accompanied by a sharp drop of ~10-15 kcal/mol (their[29] Fig 4a). This contrasts with the gentle energy gradient they see in earlier stages of folding.[29] It is quite reasonable that a locking-in of native structure by proper side chain ordering in the final stage of folding would give a pronounced drop in internal energy relative to near-native structures.[30] In our lattice, an additional drop of ~13 kcal/mol at the native/unfolded transition zone would give a model that conforms to experimental kinetics and thermodynamics results as well as detailed free energy calculations.

Comparison to other dynamics models.

A primary issue with dynamics studies using lattice models[31,32] or very simplified atom representation[33] is placing the observed events on a time axis. This is an issue with Monte Carlo studies in general, but reduced atom representation further obscures the connection. Our random walk model with explicit time representation shows excellent agreement with MD time behavior at the cost of sacrificing direct atom representation. In this sense, our model is complementary to existing tools for studying protein dynamics and testing energy landscape models. For example, our real-time behavior can extend analytic dynamic models, such as the model of Zwanzig et al[34], who derive the equation for folding mean first-passage time as approximately $(1/Nk_0)(1 + k_0/k_1)^N$, where N is the number of bonds (dimensions in our case) and k_0 (k_1) is the rate of entering (leaving) the native state for any dimension. Zwanzig et al[34], who were mostly interested in resolving Levinthal's paradox, simply assert "reasonable" microscopic rate constants for their calculation. Our model can be transformed into their model by refitting our native space such that all twelve large dimensions are equal (i.e. made hypercubic instead of hyperrectangular). Microscopic rate constants can be arrived at numerically by one-dimensional random walk simulations as a function of overall landscape bias, shape, and roughness. Such models could then be quantitatively compared to MD to test for reasonable displacement behavior.

Generalization to other systems.

While these results derive entirely from one protein, we expect generalization of many model features to other small single domain proteins. The shortest time dynamic features (collision interval and path length) clearly reflect basic material properties of protein, not fold or sequence, and should be constant for all proteins at room temperature. Preliminary MD calculations on other systems uphold this assertion. We explain the total number of dimensions of the lattice in terms of amino acid structure and so expect the ~2N number of dimensions to be transferable to other proteins. The number of large displacement modes should scale with the number of secondary structural elements. The RMSD distribution in the MD TAD function at times longer than 100 ns is similar to that of randomly folded compact 36-mer polypeptides. A scheme for arriving at the total lattice size is iteratively comparing distance distributions for randomly chosen lattice point-pairs with the RMSD distribution of randomly folded conformations for the protein of interest. The least understood aspect of our model is the appropriate roughness of conformational space. The parameters for our local minima traps are largely influenced by a single semi-stable "folding intermediate" that persists for 150 ns in the Duan and Kollman[3] trajectory. Despite this dearth of long time data, we are comforted that on our trap-free lattice (set B parameters), which is unrealistically smooth, folding times are only ~2 orders of magnitude faster than our optimal lattice.

Finally, we try to put this work in context with the large-scale issues in protein folding. There are two quite distinct challenges: modeling protein folding kinetics and predicting the most stable protein geometry. This paper focuses on the first problem. In fact, we explicitly remove knowledge of specific geometries by averaging coordinate displacements generated during MD trajectories. In exchange for the loss of coordinate details, our approach appears to yield a plausible time scale for folding events. It also establishes a protocol for using molecular dynamics information to test diverse folding models.

The formulation of folding as a bounded, biased lattice walk suggests a set of basic variables that differ in significant ways from other formulations. These variables include the number and extent of the "conformational" dimensions; the fundamental time/displacement step on the lattice; the trap depth and trap distribution; and the energy surface configuration and bias. The ranges of these variables - the feasible "solution space" - are strongly bounded by experimental data: experimental folding information, exponential vs. non-exponential kinetics, equilibrium protein stability, kinetic rates as a function of temperature or denaturants.

In the future, approaches such as this one can be used to explore pathway and funnel models of protein folding, the nature and effects of entropic transition states, and unfolding phenomena. On the other hand, this method will most likely not be able to explore specific geometric issues such as the effect of contact order models[35], or the spatial compactness or arrangement of native proteins. While it is too early to know whether a unique description of the folding of the villin headpiece can be developed, it is encouraging that the experimental results currently available fit a simple lattice diffusion model.

Conclusions

We show that a random walk, lattice-diffusion model explains many features of conformational evolution in MD. While individual motions may be oscillatory in nature, they sum into a diffusive motion through conformation space. It is possible to interpret RMSD dependence on time separation using this model up to microsecond time scales for a folding protein by introducing a hierarchical energy landscape. A simple funnel landscape with an energy difference between native and unfolded of ~28 kcal/mol, with 15 kcal/mol spread over the compact unfolded state and 13 kcal/mol at the unfolded/native transition, gives reasonable folding times and thermodynamics.

Acknowledgments

We are grateful to Ken Dill, Kim Sharp and Peter Kollman for useful discussion. This work was supported by grants from General Medical Sciences Institute of the National Institutes of Health.

References

(1) Leopold, P.E.; Montal, M.; Onuchic, J.N. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 8721-8725.

(2) Dill, K.A.; Bromberg, S.; Yue, K.Z.; Fiebig, K.M.; Yee, D.P.; Thomas, P.D.; Chan, H.S. *Protein Sci.* **1995**, *4*, 561-602.

(3) Duan, Y.;Kollman, P.A. *Science* **1998**, *282*, 740-744.

(4) Sullivan, D.C.; Kuntz, I.D. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 495-511.

(5) Cornell, W.D.; Cieplak, P.; Bayly, C.I.; Gould, I.R.; Merz, K.M.; Jr., Ferguson, D.M.; Spellmeyer, D.C.; Fox, T.; Caldwell, J.W.; Kollman, P.A. *J. Am. Chem. Soc.* **1995**, *117*, 5179 -5197.

(6) Srinivasan, J.; Cheatham, T.E.; Cieplak, P.; Kollman, P.A.; Case, D.A. *J. Am. Chem. Soc.* **1998**, *120*, 9401-9409.

(7) Lee, M.R.; Duan, Y.; Kollman, P.A. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 309-316.

(8) McLachlan, A.D. *Acta. Crystallogr., Sect A: Found. Crystallogr.* **1982**, *38*, 871-873.

(9) Case, D.A.; Pearlman, D.A.; Caldwell, J.W.; Cheatham, T.E., III; Ross, W.S.; Simmerling, C.L.; Darden, T.A.; Merz, K.M.; Stanton, R.V.; Cheng, A.L.; Vincent, J.J.; Crowley, M.; Ferguson, D.M.; Radmer, R.J.; Seibel, G.L.; Singh, U.C.; Weiner, P.K.; Kollman, P.A. *AMBER 5 (UCSF)*; University of California, San Francisco, 1997.

(10) Ryckaert, J.; Ciccotti, G.; Berendsen, H.J.C. *J. Comp.Phys.* **1977**, *23*, 327-341.

(11) McKnight, C.J.; Matsudaira, P.T.; Kim, P.S. *Nat. Struct. Biol.* **1997**, *4*, 180-184.

(12) Martin, A.C.R. *ProFit V1.8*; University College, London, 1998.

(13) Rudnick, J.; Gaspari, G. *Science* **1987**, *237*, 384-389.

(14) Troyer, J.M.; Cohen, F.E. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 97-110.

(15) Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. *J. Chem. Phys.* **1953**, *61*, 813-826.

(16) Levitt, M.; Sander, C.; Stern, P.S. *J. Mol. Biol.* **1985**, *181*, 423-447.

(17) Balsera, M.A.; Wriggers, W.; Oono, Y.; Schulten, K. *J. Phys. Chem.* **1996**, *100*, 2567-2572.

(18) Ueda, Y.; Taketomi, H.; Go, N. *Biopolymers* **1978**, *17*, 1531-1548.

(19) Englander, S.W. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 213-238.

(20) Jackson, S.E.; Fersht, A.R. *Biochemistry* **1991**, *30*, 10428-10435.

(21) Plaxco, K.W.; Baker, D. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 13591-13596.

(22) McKnight, C.J.; Doering, D.S.; Matsudaira, P.T.; Kim, P.S. *J. Mol. Biol.* **1996**, *260*, 126-134.

(23) Gray, D.E. (editor). *American Institute of Physics handbook.* McGraw-Hill: New York, 1972.

(24) Buckler, D.R.; Haas, E.; Scheraga, H.A. *Biochemistry* **1995**, *34*, 15965-15978.

(25) Hagen, S.J.; Hofrichter, J.; Szabo, A.; Eaton, W.A. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 11615-11617.

(26) Fersht, A.R. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 1525-1529.

(27) Fulton, K.F.; Main, E.R.G.; Daggett, V.; Jackson, S.E. *J. Mol. Biol.* **1999**, *291*, 445-461.

(28) Li, A.J.; Daggett, V. *J. Mol. Biol.* **1996**, *257*, 412-429.

(29) Lazaridis, T.; Karplus, M. *Science* **1997**, *278*, 1928-1931.

(30) Levitt, M.; Gerstein, M.; Huang, E.; Subbiah, S.; Tsai, J. *Annu. Rev. Biochem.* **1997**, *66*, 549-579.

(31) Levitt, M.; Warshel, A. *Nature* **1975**, *253*, 694-698.

(32) Sali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248-251.

(33) Zhou, Y.Q.; Karplus, M. *Nature* **1999**, *401*, 400-403.

(34) Zwanzig, R.; Szabo, A.; Bagchi, B. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 20-22.

(35) Plaxco, K.W.; Simons, K.T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985-994.

(36) Bevington, P.R.; Robinson, D.K. *Data reduction and error analysis for the physical sciences,* 2nd ed.; McGraw-Hill: New York, 1992; Chapter 6.

HIV-1 Protease Structures Cluster by Space Group

David C. Sullivan and Irwin D. Kuntz

Department of Pharmaceutical Chemistry
University of California at San Francisco
San Francisco, California 94143-0446

Abstract

We show that much of the structural variations among the liganded human immunodeficiency virus type 1 protease crystal structures is simply correlated with space group. Structures solved in the same space group are reproducibly similar to each other, while structures of different space groups are different. The pattern of main-chain disorder, as measured by B-factors, is also strongly crystal form dependent. Most of the differences across space groups are in the surface loops (residues 15-19, 36-41, and 67-71) as well as the active site ψ-loop (residues 78-84). Structural variation in the ψ-loop is largely independent of mutations to this substructure. The hinges of the flap (residues 45-56) covering the active site show space-group-associated displacement while the ligand-contacting tips of the flap do not. Across the ensemble of structures, variations between the A and B chains of the homodimer are of the same order as variations between the same chain in different space groups. Regardless of whether these observations arise from crystal packing forces, refinement artifacts or some combination of such factors, it is clear that structural inferences made from comparisons from differing space groups or between crystallographic and NMR results must be done with caution.

Keywords: HIV, proteinase, lattice packing effects, comparison of X-ray structures, B-factors

Introduction

Human immunodeficiency virus type 1 protease (HIV-1 PR) protease proved a valuable target of structure-based drug design in the early 1990s[1]. Crystal structures of the inhibited protease were pursued both for hypothesis generation and testing of inhibitor binding modes. From these efforts, many tens of crystal structures solved in multiple space groups, and an NMR solution structure, of the liganded protease were deposited in the Protein Data Bank (PDB)[2]. This ensemble provides a test set for measuring the relative roles played by variations in ligand, space group, mutations, and experimental conditions in perturbing structure.

Differences in structure within this ensemble are presumed to result primarily from differences in the ligand[3,4]. The many liganded wildtype crystal structures of HIV-1 PR are structurally very similar to each other, with Cα RMSDs between structures generally less than 1 Å. By some analyses, similarities between structures break down near the level of error[5] with any real differences presumed to result from plastic deformations unique to each inhibitor[3,4], variations in experimental conditions, and experimental error and different model building biases. On the assumption that variation among structural models has physical meaning, Van Aalten et al[4] proceeded to extract concerted motions implied by this ensemble which resemble those seen in a molecular dynamics trajectory. In summary, the published literature suggests that while the magnitude of structural divergence is not large, the changes may not be random.

Lange-Savage et al[6] solved the crystal structure of the S37N mutant of HIV-1 PR bound to HOE/BAY 793 in two crystal forms. Significant structural differences are seen between the two crystal forms. Here we have further investigated lattice packing effects by analyzing a large number of HIV-1 PR crystal structures. The PDB and National Cancer Institute's HIV Protease Databank were searched for liganded structures of HIV-1 PR. A data set of 63 high-resolution (resolution < 2.5 Å) structures was created. These structures display a number of mutations (See Table 1) and a variety of ligands. A distinguishing feature of the HIV-1 PR ensemble is the presence of many structures from each of several packing environments. This characteristic allows us to study the effects of crystal packing in more detail than previous studies. There is a one-to-one correspondence between space group and crystal type among the structures in this ensemble, which can be seen by rebuilding the unit cells using the crystal parameters given in the PDB files and checking for proper overlay. Herein we use the space group to distinguish between the crystal types.

Table 1: Mutations in Test Set

PDB i.d.	Space Group	Mutations
pdb1hih	$P2_12_12_1$	
pdb1hpx	$P2_12_12_1$	
pdb1hvi	$P2_12_12_1$	
pdb1hvj	$P2_12_12_1$	
pdb1hvk	$P2_12_12_1$	
pdb1hvl	$P2_12_12_1$	
pdb1dif	$P2_12_12_1$	
pdb1ohr	$P2_12_12_1$	
pdb1hiv	$P2_12_12_1$	I3V
pdb1odw	$P2_12_12_1$	I3V
pdb1hvs	$P2_12_12_1$	V82A
pdb1ytg	$P2_12_12_1$	Q7K K14R S37N R41K L63P I64V
pdb1mtr	$P2_12_12_1$	Q7K K14R L33I S37N R41K L63P I64V C67B C95B *
pdb1yth	$P2_12_12_1$	Q7K K14R S37N R41K L63P I64V
pdb2aid	$P2_12_12_1$	Q7K K14R S37N R41K L63P I64V
pdb4hvp	$P2_12_12_1$	K14R S37N R41K L63P I64V C67B C95B *
pdb7hvp	$P2_12_12_1$	K14R S37N R41K L63P I64V C67B C95B *
pdb1cpi	$P2_12_12_1$	K14R S37N R41K L63P I64V C67B C95B *
pdb1hos	$P6_1$	
pdb1htf	$P6_1$	
pdb1hps	$P6_1$	
pdb1hpv	$P6_1$	
pdb1hbv	$P6_1$	
pdb1gno	$P6_1$	
pdb1sbg	$P6_1$	S37N
pdb1pro	$P6_1$	S37N
pdb1vij	$P6_1$	S37N
pdb1axa	$P6_1$	A28S
pdb1gnm	$P6_1$	V82D
pdb1gnn	$P6_1$	V82N
pdb1hxb	$P6_1$	I3V
pdb1mer	$P6_1$	I3V I84V
pdb1mes	$P6_1$	I3V I84V
pdb1bv9	$P6_1$	I3V I84V
pdb1bv7	$P6_1$	I3V V82F
pdb1met	$P6_1$	I3V V82F
pdb1meu	$P6_1$	I3V V82F I84V
pdb1bwa	$P6_1$	I3V V82F I84V
pdb1bwb	$P6_1$	I3V V82F I84V
pdb1qbr	$P6_1$	I3V C95A
pdb1qbt	$P6_1$	I3V C95A
pdb1qbu	$P6_1$	I3V C95A
pdb1hvh	$P6_1$	I3V C95A

pdb1hvr	$P6_1$	I3V C95A
pdb1odx	$P6_1$	A71T V82A
hiv10msd	$P2_12_12$	
hiv11msd	$P2_12_12$	
hiv9msd	$P2_12_12$	
pdb1htg	$P2_12_12$	
pdb1hsg	$P2_12_12$	
pdb7upj	$P2_12_12$	
pdb1ajv	$P2_12_12$	
pdb1ajx	$P2_12_12$	
pdb1tcx	$P2_12_12$	V32I I47V V82I
pdb1hxw	$P2_12_12$	S37N
pdb5hvp	$P2_12_12$	S37N
pdb4phv	$P2_12_12$	S37N
pdb1vik	$P2_12_12$	S37N
pdb1a30	$P2_12_12$	Q7K L33I L63I
hiv22ulk	$P2_12_12$	Q7K L33I L63I
hiv21ulk	$P2_12_12$	Q7K L33I R41G L63I
pdb1aid	$P4_1$	K14R S37N R41K L63P I64V
pdb1a9m	I222	G48H

* - "B" symbolizes alpha-aminobutyric acid.

Results

Structure Comparison

Figure 1 shows a Cα overlay of all structures in the ensemble. PDB structure 1HIH is used as the reference molecule. Using different structures as the reference molecule gives a qualitatively similar picture. The greatest variation is seen in the surface loops, namely in the vicinity of residues 18, 40 and 68. Less variation is seen in the position of the flap residues which clamp down over the ligand. 1AID, which has a haloperidol-based ligand in a binding mode not commonly observed[7], is the obvious exception. The flaps of the non-liganded structure (3HVP -- not shown) are similarly open. Flap opening appears necessary for protease activity and has been studied extensively[8,9].

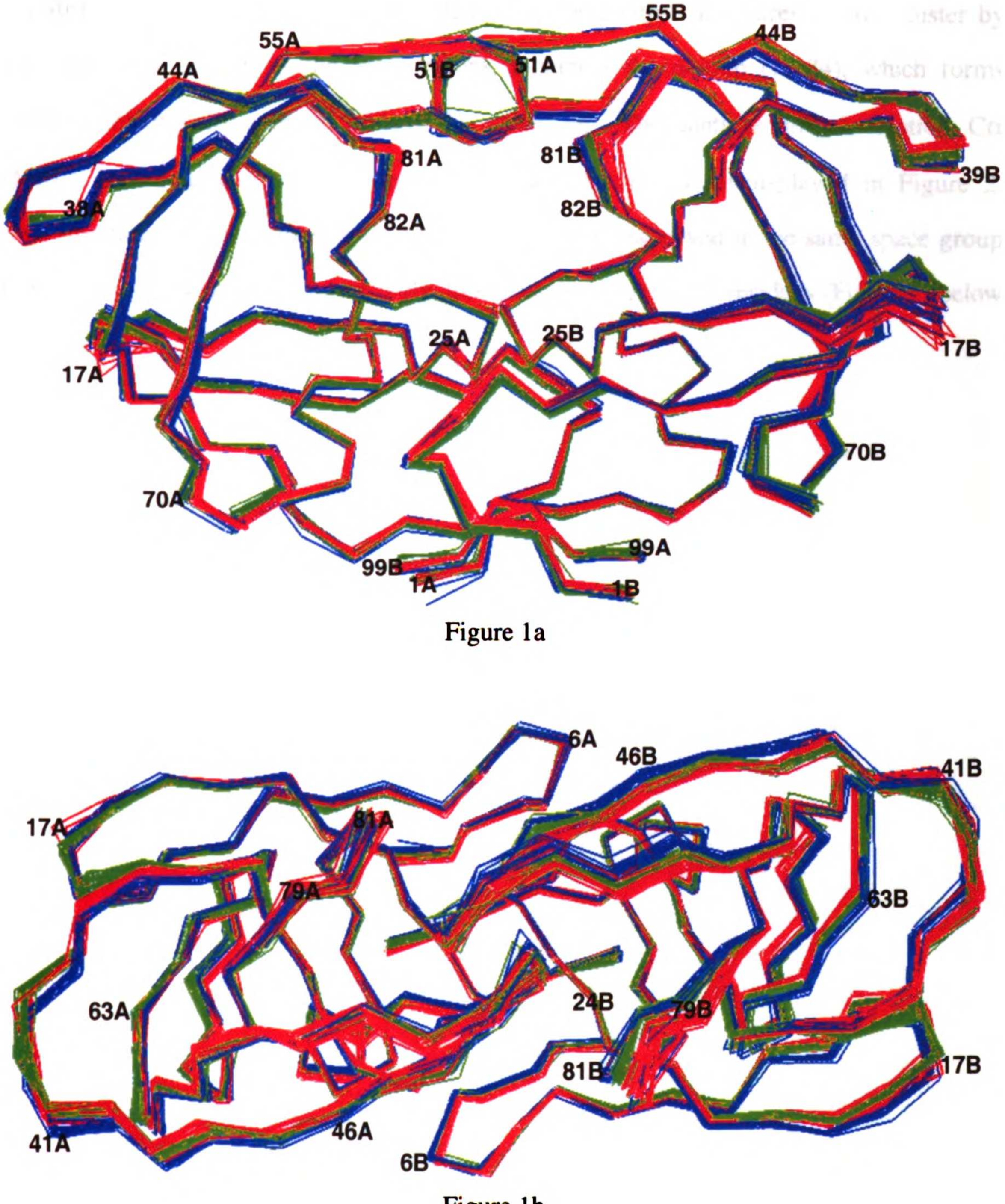


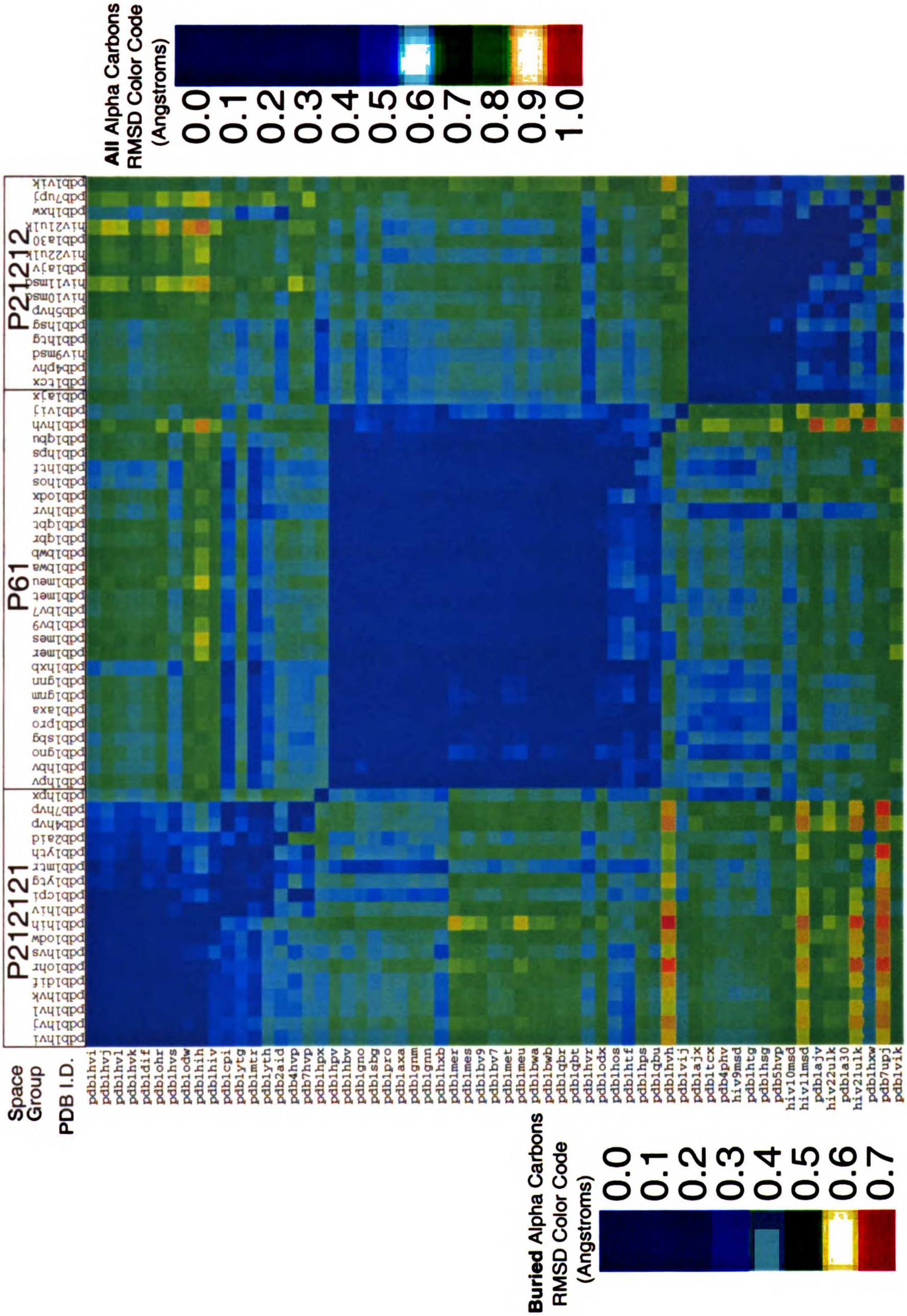
Figure 1a

Figure 1b

Figure 1 Structural variation within ensemble. The Cαs of all structures in our test set are superimposed onto 1HIH. Structures of space group $P2_12_12_1$ are colored red, $P6_1$ are green, and $P2_12_12$ are blue. Perspective A looks through the active site. Perspective B is looking down onto the active site flaps.

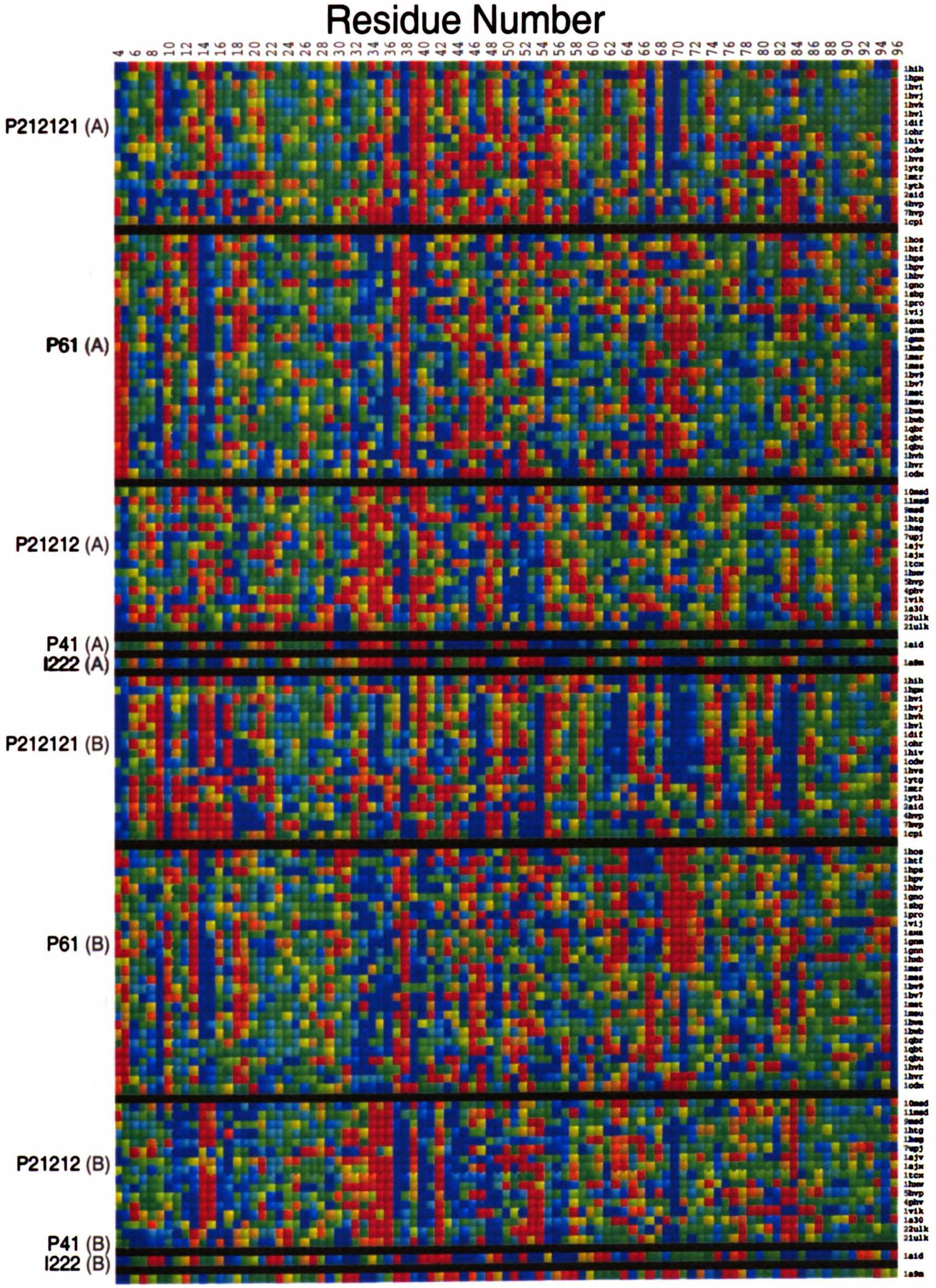
The structures in Figure 1 are color coded by space group. Structures clearly cluster by space group in several of the loop regions, including the ψ-loop (78-84), which forms pseudo-symmetrically opposed walls of the active site. To quantitate this observation, Cα RMSDs for all structure pairs were calculated and the results are displayed in Figure 2. Most structures show the most similarity to other structures solved in the same space group (Figure 2, above diagonal). Surprisingly, this is also true for buried residues (Figure 2, below diagonal).

Figure 2 (next page) Pairwise Cα RMSD. Pair-wise Cα RMSDs are color-coded by the scale indicated. For values in the above-diagonal region, all Cα receptor atoms are included in the calculation. In contrast, Cαs of surface residues were excluded from both the superpositioning and RMSD calculation of values displayed below the diagonal. Excluded residues are: 1-4, 6-8, 10, 12, 14, 16-21, 29-30, 34-46, 51-55, 57-58, 60-61, 63, 65, 67-70, 72-74, 79-82, 87-88, 92-94, and 98-99.


All Alpha Carbons
RMSD Color Code
(Angstroms)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Buried Alpha Carbons
RMSD Color Code
(Angstroms)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Space Group
PDB I.D.
P212121
P61
P21212
pdb1hvi
pdb1hvj
pdb1hvl
pdb1hvk
pdb1dif
pdb1ohr
pdb1hvs
pdb1odw
pdb1hih
pdb1hiv
pdb1cpi
pdb1ytg
pdb1mtr
pdb1yth
pdb2aid
pdb4hvp
pdb7hvp
pdb1hpx
pdb1hpv
pdb1hbv
pdb1gno
pdb1sbg
pdb1pro
pdb1axa
pdb1gnm
pdb1gnn
pdb1hxb
pdb1mer
pdb1mes
pdb1bv9
pdb1bv7
pdb1met
pdb1meu
pdb1bwa
pdb1bwb
pdb1qbr
pdb1qbt
pdb1hvr
pdb1odx
pdb1hos
pdb1htf
pdb1hps
pdb1qbu
pdb1hvh
pdb1vij
pdb1ajx
pdb1tcx
pdb4phv
hiv9msd
pdb1htg
pdb1hsg
pdb5hvp
hiv10msd
hiv11msd
pdb1ajv
hiv22ulk
pdb1a30
hiv21ulk
pdb1hxw
pdb7upj
pdb1vik

We probed for the location of differences along sequence by examining internal coordinates. The space group associated positional variance seen in the Cα overlay was not readily apparent from an analysis of the torsion angles. The structural features that did emerge from torsion angle analysis include: space group independent variation in the main chain torsion angles of Ile 50 and Gly 51, which NMR experiments show to undergo microsecond time scale conformational changes in solution[10], as well as mainchain torsion variation in some turns (Gly 16-Gly 17 and Cys 67-Gly 68). The side chains of Val 11, Leu 24, and Leu 97, which are well-defined, buried, and spatially contiguous, populate multiple rotomer types in correlation with space group. Crystal packing associated differences seen in the Cα overlay are spread over several torsions. To capture these differences, we measured angles defined by vectors $C\alpha_{(i-1)}$ to $C\alpha_{(i-3)}$ and $C\alpha_{(i+1)}$ to $C\alpha_{(i+3)}$ (Figure 3). A plot of the deviation from the average angle value shows regions of substructure-level hinging which is associated with crystal type. Centers of space group associated hinging are seen at residues 14-15, 36-40, and 69-72. In the ψ-loop region, hinging centers around residues 83 and 84.

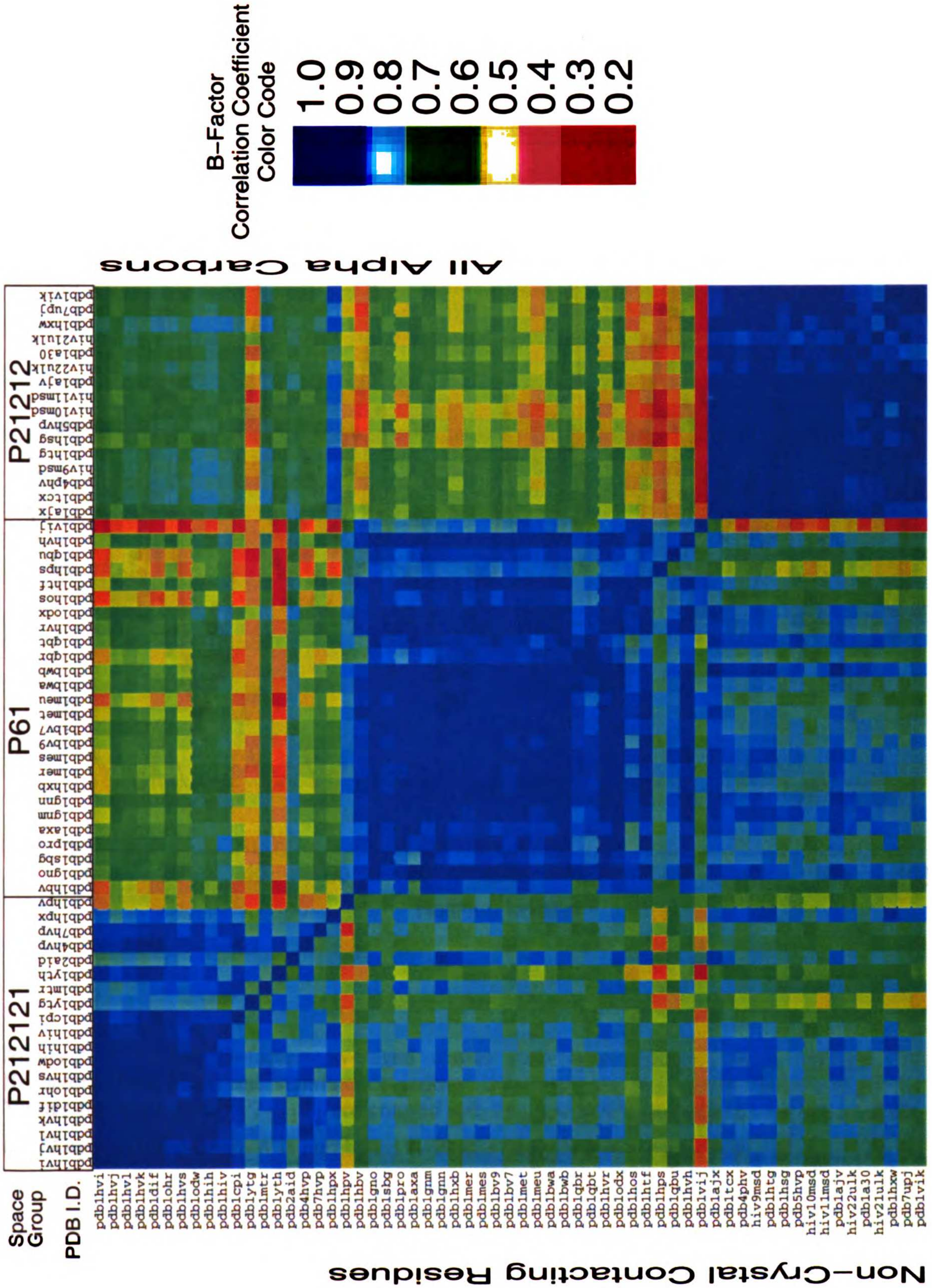
Figure 3 (next page) Cα vector angle deviations. For residue i, an angle defined by vectors $C\alpha_{(i-1)}$ to $C\alpha_{(i-3)}$ and $C\alpha_{(i+1)}$ to $C\alpha_{(i+3)}$ was calculated. Deviation from the average angle for residue i across all structures for both chains is color coded in the figure. Positive deviations 3° or greater are colored blue, negative deviations 3° or greater are colored red. Green is neutral. The structures are listed in the same order as in Table 1. '(A)' and '(B)' denote chain letter.


Residue Number
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
56
58
60
62
64
66
68
70
72
74
76
78
80
82
84
86
88
90
92
94
96
P212121 (A)
P61 (A)
P21212 (A)
P41 (A)
I222 (A)
P212121 (B)
P61 (B)
P21212 (B)
P41 (B)
I222 (B)
1hih
1hpx
1hvi
1hvj
1hvk
1hvl
1dif
1ohr
1hiv
1odw
1hvs
1ytg
1mtr
1yth
2aid
4hvp
7hvp
1cpi
1hos
1htf
1hps
1hpv
1hbv
1gno
1sbg
1pro
1vij
1axa
1gnm
1gnn
1hxb
1mer
1mes
1bv9
1bv7
1met
1meu
1bwa
1bwb
1qbr
1qbt
1qbu
1hvh
1hvr
1odx
10msd
11msd
9msd
1htg
1hsg
7upj
1ajv
1ajx
1tcx
1hxw
5hvp
4phv
1vik
1a30
22ulk
21ulk
1aid
1a9m
1hih
1hpx
1hvi
1hvj
1hvk
1hvl
1dif
1ohr
1hiv
1odw
1hvs
1ytg
1mtr
1yth
2aid
4hvp
7hvp
1cpi
1hos
1htf
1hps
1hpv
1hbv
1gno
1sbg
1pro
1vij
1axa
1gnm
1gnn
1hxb
1mer
1mes
1bv9
1bv7
1met
1meu
1bwa
1bwb
1qbr
1qbt
1qbu
1hvh
1hvr
1odx
10msd
11msd
9msd
1htg
1hsg
7upj
1ajv
1ajx
1tcx
1hxw
5hvp
4phv
1vik
1a30
22ulk
21ulk
1aid
1a9m

B-factor correlations associated with space group

The Cα disorder along sequence in HIV-1 PR is strongly space group dependent. We calculated correlation coefficients between Cα factors plotted as a function of residue number for pairs of structures (See Figure 4, above diagonal). The calculation was repeated, this time excluding residues which make crystal contacts in any of the three packing environments. The maintenance of class segregation when crystal contacting residues are excluded (Figure 4, below diagonal) points to the extensive impact of crystal packing on the protein's dynamic environment. The spread in B-factors averaged across the space groups (see Figure 5) exceeds 10 Å^2 at residues 14-21A, 43-46A, 63-71A, 14-20B, 40-46B and 52-53B and exceeds 20 Å^2 at residues 17-19A, 17-18B and 41-42B.

Figure 4 (next page) Cα B-factor correlation coefficients. Correlation coefficients between each pair of structures' set of Cα B-factors over the whole dimer are displayed by color code. In the below-diagonal data, residues making crystal contacts in either of the three main crystal types were not considered in the correlation calculation. For the crystal contact calculation, three representative structures were used: 1HVI ($P2_12_12_1$), 1HPV ($P6_1$), and 1HTG ($P2_12_12$). A contact distance of 4.5 Å was employed. Ninety-five of 198 residues were thus excluded in the correlation calculation.


B-Factor
Correlation Coefficient
Color Code
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
All Alpha Carbons
Space Group
PDB I.D.
P212121
P61
P21212
pdb1hvi
pdb1hvj
pdb1hvl
pdb1hvk
pdb1dif
pdb1ohr
pdb1hvs
pdb1odw
pdb1hih
pdb1hiv
pdb1cpi
pdb1ytg
pdb1mtr
pdb1yth
pdb2aid
pdb4hvp
pdb7hvp
pdb1hpx
pdb1hpv
pdb1gno
pdb1hbv
pdb1sbg
pdb1pro
pdb1axa
pdb1gnm
pdb1gnn
pdb1hxb
pdb1mer
pdb1mes
pdb1hv9
pdb1hv7
pdb1met
pdb1meu
pdb1bwa
pdb1bwb
pdb1qbr
pdb1qbt
pdb1hvr
pdb1odx
pdb1hos
pdb1htf
pdb1hps
pdb1qbu
pdb1hvh
pdb1vij
pdb1ajx
pdb1tcx
pdb4phv
hiv9msd
pdb1htg
pdb1hsg
pdb5hvp
hiv10msd
hiv11msd
pdb1ajv
hiv22ulk
pdb1a30
hiv21ulk
pdb1hxw
pdb7upj
pdb1vik
Non-Crystal Contacting Residues

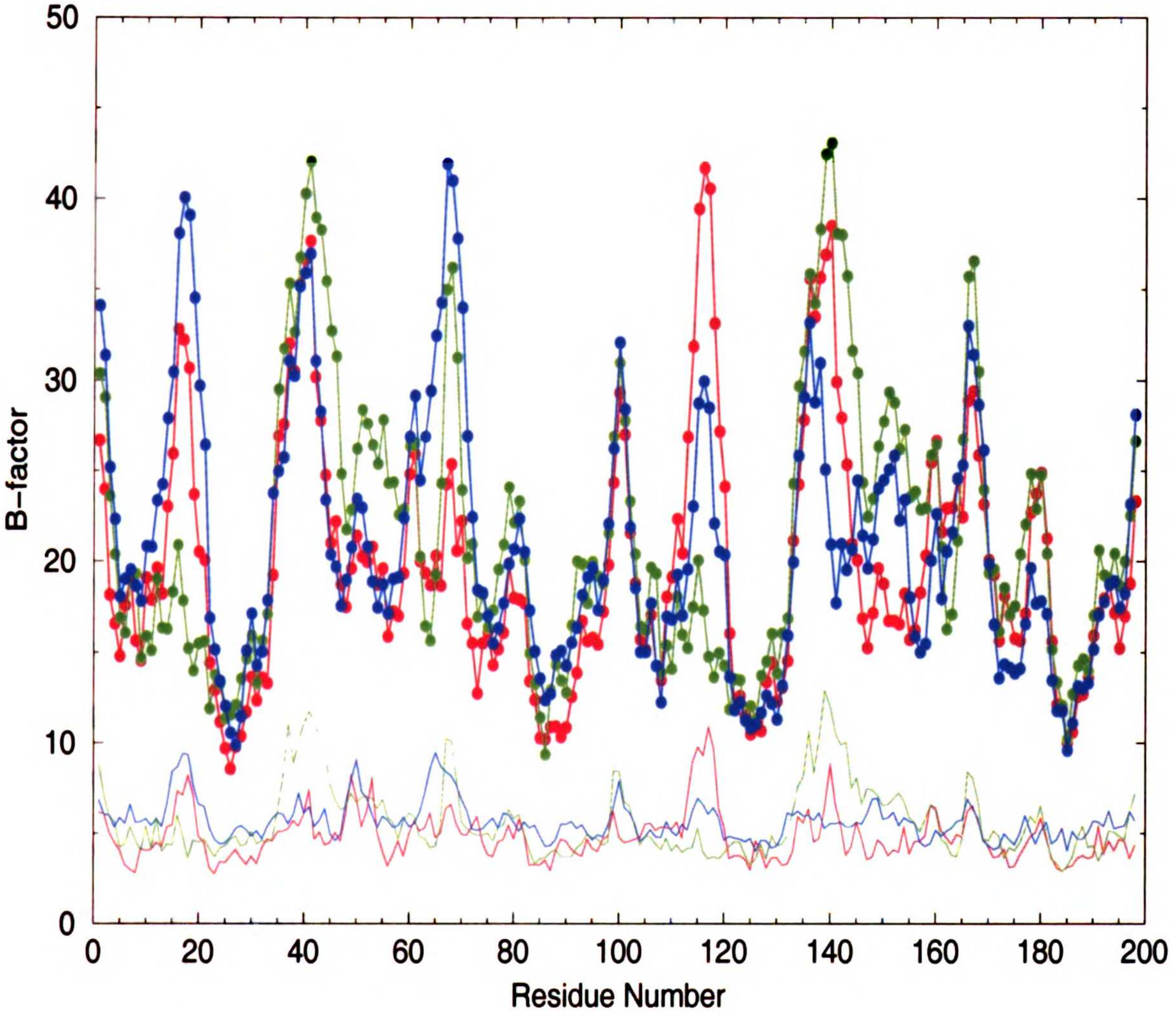


Figure 5 Cα B–factors averaged across each space group. Both average Cα B–factors (bold lines) and standard deviations (thin lines) are color coded by space group. Values for space group $P2_12_12_1$ are colored red, $P6_1$ are green, and $P2_12_12$ are blue. Units are (Angstroms)^2. Chain A is numbered 1–99 and chain B is numbered 100–198.

Discussion

Clearly, HIV-1 PR crystal structures segregate into classes that correlate with space group both with respect to geometry and B-factors. The first critical question is to what extent this clustering arises from physical properties of the crystal packing arrangements or if it is, at least in part, an artifact of the refinement procedures. The most decisive way to answer this question is to solve structures of closely related or identical systems independently using isomorphous replacement, anomalous phases or ab initio methods. This experiment has not been carried out for HIV-1 PR to the best of our knowledge. Figure 6 shows the connections among most of the structures discussed in this paper. There are no MIR solutions. All liganded structures were developed using molecular replacement techniques.

If we assume, for purposes of discussion, that the HIV-1 protease structures reflect true physical variability, we are struck by the smallness of the geometric perturbations associated with changing space group, with mutated sequences or by changing ligation. In fact, neither sequence nor inhibitor decisively fixes the space group since liganded wild-type HIV-1 PR has been repeatedly crystallized in three well-populated crystal types. Lange-Savage et al's[6] results and results of others anecdotally rule out the importance of ligand. Among structures of this data set, HIV-1 PR inhibited by SB203386 has been solved in both $P2_12_12$ (1TCX[11]) and $P6_1$ space groups(1SBG[12]). HIV-1 PR inhibited by U89360E has been solved in both $P6_1$ (1GNM[13], 1GNN[13], 1GNO[13] and 1AXA[14]) and I222 (1A9M[15]). HIV-1 PR inhibited by BMS-182193 has been solved in both $P2_12_12_1$ (1ODW[16]) and $P6_1$ (1ODX[16]). Rather, the literature suggests that electrostatics is the primary determinant of crystal packing among nearly identical systems. Jelsch et al[17] show that crystal type can be modulated by changing the charge distribution on the surface of cutinase via mutations and ligand changes. Gallagher and Croker[18] varied pH to elucidate a "trigger" mechanism in bovine pancreatic trypsin inhibitor where the protonation state of a particular residue determines crystal type. There are not such clear results for HIV-1 PR. The determinants of crystal type have not been heavily explored. A pH dependent crystal phase transition was observed between pH 4.8 and 5.0. Small tetragonal rods (unusable for data collection) grow below this range and orthorhombic plates (space group $P2_12_12$) grow above this range for a particular set of conditions[19]. Jordan Tang's group[13,14] has reported on searches for optimal crystal conditions in terms of pH and salt concentrations (ammonium sulfate), but do not report conditions associated with crystal type transitions.

An intriguing question is whether the different crystal packing arrangements generate forces which perturb the protein into conformations not significantly populated in solution or do the conformations preferred in each space group correspond to true conformational minima? Kossiakoff et al's[20] comparisons of BPTI molecular dynamics trajectories to crystal structures of different crystal types suggest that crystal packing stabilizes conformations found in solution. The average pairwise $C\alpha$ RMSD between the HIV-1 PR NMR ensemble members (PDB code: 1BVE) is 1.14 ± 0.35 Å. This is much more than the average inter-space group $C\alpha$ RMSD for structures in the ensemble here (0.66 Å ± 0.08 Å; excluding 1AID and 1A9M), and much more than the average intra-space group $C\alpha$ RMSD of 0.38 ± 0.11Å. If each NMR ensemble member is taken to correspond to a region of the solutions structure's accessible conformational space, we can speculate that any crystal environment will restrict the protein to a particular region of conformational space and particular crystal environments will further differentiate possible conformations.

To put these RMSD values in perspective it is worth comparing with other retro-virus proteases. The average $C\alpha$ RMSD between HIV-1 PR of space groups $P2_12_12_1$, $P6_1$, and $P2_12_12$ and a representative high resolution (1.7 Å) HIV-2 PR liganded structure, 1IDA, is

0.98 ± 0.06 Å. 1YTI, a simian immunodeficiency virus structure of 2.2 Å resolution has an average Cα RMSD with HIV-1 PR of 1.12 ± 0.06 Å. These numbers agree well with the Chothia and Lesk[21] expression for relating sequence identity to crystal structure backbone RMSD: 1IDA, with 49% sequence identity to wild-type HIV-1 PR is predicted to differ by 1.05 Å backbone RMSD. 1YTI, of 51% sequence identity, is predicted to have 1.01 Å RMSD with HIV-1 PR.

From comparing crystal packing associated differences with those associated with mutation, we conclude that mutations do not greatly influence the global structure of the protease as revealed by crystallography. However, the structural consequences of mutations are undoubtedly of great importance in solution. Cross space group analysis adds proper perspective for interpreting crystallographic structures, particularly mutant structures, many of which come from studies aimed at understanding drug resistance. Hypotheses for the mechanism of resistance generally consider perturbations to the shape of the active site and/or to dynamics as it relates to enzyme function. The degree of each argument invoked depends on the particular mutation at hand. To explain mutations distal from the active site, Rose et al[3] found five rigid domains in the dimer by comparing an open, unliganded SIV protease crystal structure with liganded and unliganded HIV-1 PR crystal structures. Many resistance-inducing mutations cluster at the inter-domain interfaces allowing them to alter the dynamics of domain motions. They propose that these domain interface mutations change substrate and inhibitor binding kinetics in favor of the substrate. This theory does not necessarily predict changes to the static structure by domain interface mutations.

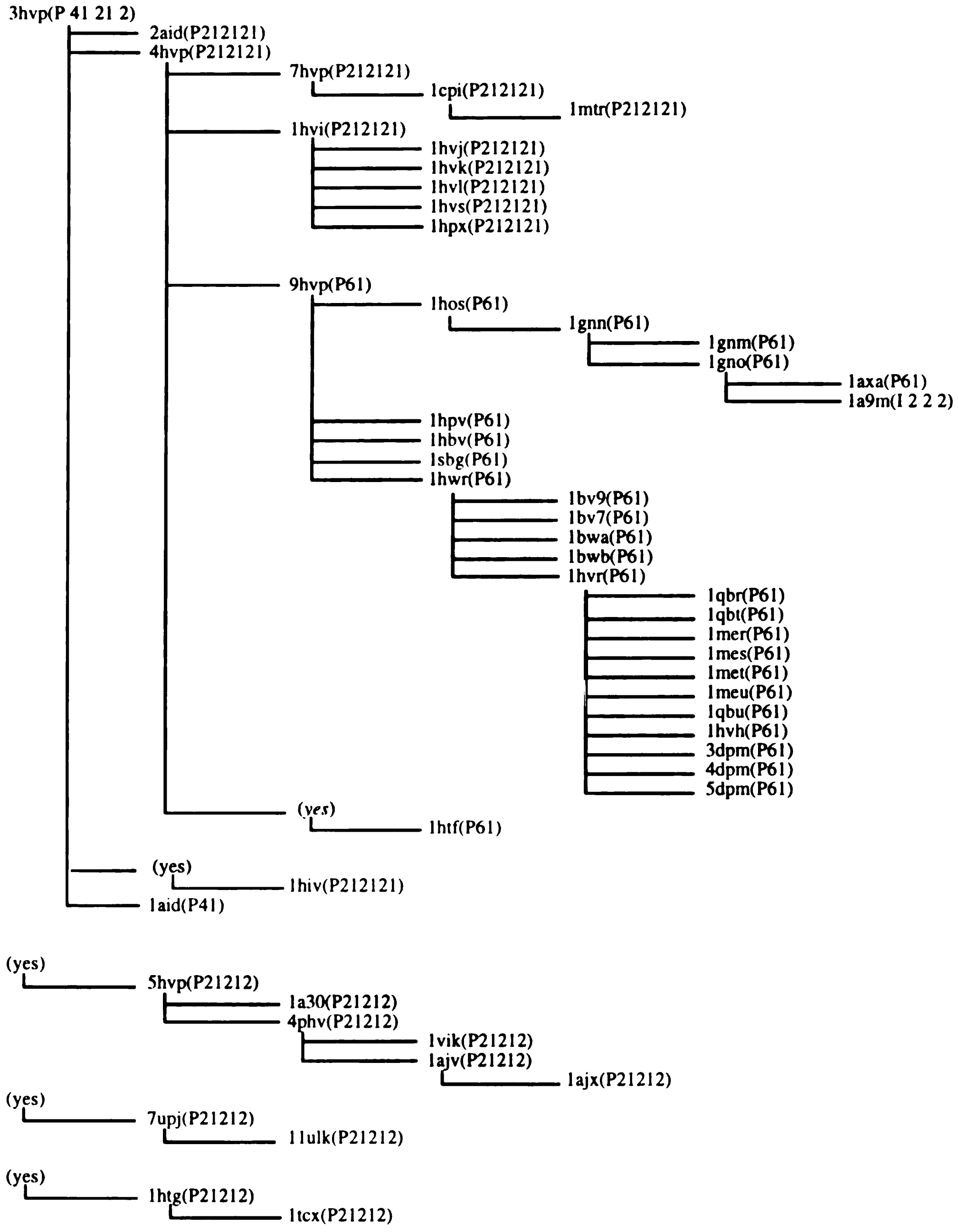


Figure 6 Molecular replacement evolution tree. The history of molecular replacement, where accessible from the literature, is here listed for the ensemble. Structures higher in the hierarchy were used to aid in phase determination of those below. We use the term 'molecular replacement' in the liberal sense of using a previously solved model to aid in phasing another set of data. The top structure (3HVP[28]) was itself solved using a Rous sarcoma virus protease-based homology model of HIV-1 PR[29]. 'Yes' indicates that the PDB accession number for that parent structure is not available, either because it was not reported or because the coordinates were not deposited.

Ala et al[22] studied differences in binding to DMP323 by the active site mutants: V82F, I84V and the double mutant, V82F/I84V. Structure based energy calculations are consistent with the hypothesis that the inhibitor is less potent against resistance mutants due to loss of van der Waals's interactions. In the V82A mutant[23], contacts predicted to be lost are partially compensated for by local backbone plasticity. This type of plasticity should not be confused with the variation associated with crystal packing. Compared to other structures of space group $P2_12_12_1$, the Cα positions of Pro81 and Asn 83 in the V82A mutant are within the spread. The mainchain of Ala 82 is at the edge of the spread. The average pairwise Cα displacement (i.e. the distance between corresponding Cαs after optimal pairwise fitting over all Cαs) at 82A across all pairs of $P2_12_12_1$ structures is 0.43 Å and for 82B is 0.61 Å. For 1HVI (WT) fitted to 1HVS (V82A), 82B is 0.56 Å shifted and for 82B is 0.58 Å shifted. The average inter space group shift for the Cα of 82A is 0.48 Å and 82B is 0.65 Å. For 1HVI fitted to 1HVS, the Cα of neighboring residue 81 is shifted by 0.35 Å (chain A) and 0.68 (chain B). The inter space group spread of residue 81's Cα is generally larger: the average Cα shifts between $P2_12_12_1$ and $P6_1$ at residue 81 are 0.63 Å (chain A) and 1.27 Å (chain B); between structures of $P2_12_12_1$ and $P2_12_12$ are 0.80 Å (chain A) and 1.71 Å (chain B). Compared to crystal packing associated conformation changes, point mutation induced atomic shifts are small and localized.

Energy calculations which involve inhibitors that contact the ψ-loop could be very sensitive to crystal type. Both Ala et al[22] and Baldwin et al[23] compared structures solved in the same space group (Ala et al[22] compared structures of $P6_1$ and Baldwin et al[23] compared structures of $P2_12_12_1$). Clearly, the removal of sidechain atoms via mutation will reduce inhibitor-protein contact. Just as clear, overall backbone position is associated with space group and will modulate the importance of a reduction in sidechain surface area. The fact that crystal environment plays such a large role in determining the ψ-loop conformation speaks to the relative instability of the loop, even in the bound form of the protease. The hinging centered around residues 83 and 84 suggests that mutations in this area could act to modulate loop dynamics or stability.

Conclusions

The basic observation in this paper is the strong clustering of the HIV-1 protease structures and B-factors with space group. This clustering demonstrates that the molecular replacement refinement procedures are responsive to crystal packing. However, it is difficult to assess whether the current practice of obtaining a large number of similar structures using only one set of independent phases has led to a significantly over constrained structural ensemble. This unexpectedly strong clustering by space group raises important caveats that should be resolved before these structures are routinely used for general biophysical investigations.

Methods

RMSDs were calculated by fitting the structures using the McLachlan algorithm[24] as implemented in the program ProFit[25] followed by taking the root-mean of the squared atom displacements. Crystal contacts were calculated using the program Crispack[26], which rebuilds neighboring structures in the crystal by using symmetry operations based on space group and crystal dimensions given in the PDB file. Contact is threshold based, and a cut-off distance of 4.5 Å between non-hydrogen atoms was employed herein. Pairwise RMSD plots, angle plots and molecular images were produced using MidasPlus[27].

Acknowledgments

The authors are grateful for support for DCS through NIH Training Grant GM 07175-23 (TL James, PI) and GM 31497 to IDK.

References

1. Wlodawer, A.; Erickson, J.W. Annu Rev Biochem 1993, 62, 543-585.

2. Bernstein, F.C.; Koetzle, T.F.; Williams, G.J.B.; Meyer, E.F.; Brice, M.D.; Rodgers, J.R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. J Mol Biol 1977, 112, 535-542.

3. Rose, R.B.; Craik, C.S.; Stroud, R.M. Biochemistry 1998, 37, 2607-2621.

4. Van Aalten, D.M.F.; Conn, D.A.; De Groot, B.L.; Berendsen, H.J.C.; Findlay, J.B.C.; Amadei, A. Biophys J 1997, 73, 2891-2896.

5. Stroud, R.M.; Fauman, E.B. Protein Sci 1995, 4, 2392-2404.

6. Lange-Savage, G.; Berchtold, H.; Liesum, A.; Budt, K.H.; Peyman, A.; Knolle, J.; Sedlacek, J.; Fabry, M.; Hilgenfeld, R. Eur J Biochem 1997, 248, 313-322.

7. Rutenber, E.; Fauman, E.B.; Keenan, R.J.; Fong, S.; Furth, P.S.; Ortiz de Montellano, P.R.; Meng, E.; Kuntz, I.D.; DeCamp, D.L.; Salto, R.; Rose, J.R.; Craik, C.S.; Stroud, R.M. J Biol Chem 1993, 268, 15343-15346.

8. Harte Jr., W.E.; Swaminathan, S.; Mansuri, M.M.; Martin, J.C.; Rosenberg, I.E.; Beveridge, D.L. Proc Natl Acad Sci USA 1990, 87, 8864-8868.

9. Collins, J.R.; Burt, S.K.; Erickson, J.W. Nat Struct Biol 1995, 2, 334-338.

10. Nicholson, L.K.; Yamazaki, T.; Torchia, D.A.; Grzesiek, S.; Bax, A.; Stahl, S.J.; Kaufman, J.D.; Wingfield, P.T.; Lam, P.Y.S.; Jadhav, P.K.; Hodge, C.N.; Domaille, P.J.; Chang, C.H. Nat Struct Biol 1995, 2, 274-280.

11. Hoog, S.S.; Towler, E.M.; Zhao, B.G.; Doyle, M.L.; Debouck, C.; Abdel-Meguid, S.S. Biochemistry 1996, 35, 10279-10286.

12. Abdel-Meguid, S.S.; Metcalf, B.; Carr, T.J.; Demarsh, P.; DesJarlais, R.L.; Fisher, S.; Green, D.W.; Ivanoff, L.; Lambert, D.M.; Murthy, K.H.M.; Petteway, S.R.; Pitts, W.J.; Tomaszek, T.A.; Winborne, E.; Zhao, B.G.; Dreyer, G.B.; Meek, T.D. Biochemistry 1994, 33, 11671-11677.

13. Hong, L.; Treharne, A.; Hartsuck, J.A.; Foundling, S.; Tang, J. Biochemistry, 1996, 35, 10627-10633.

14. Hong, L.; Hartsuck, J.A.; Foundling, S.; Ermolieff, J.; Tang, J. Protein Sci 1998, 7, 300-305.

15. Hong, L.; Zhang, X.J.; Foundling, S.; Hartsuck, J.A.; Tang, J. FEBS Lett 1997, 420, 11-16.

16. Kervinen, J.; Thanki, N.; Zdanov, A.; Tino, J.; Barrish, J.; Lin, P.F.; Colonno, R.; Riccardi, K.; Samantha, H.; Wlodawer, A. Protein Pept Lett 1996, 3, 399.

17. Jelsch, C.; Longhi, S.; Cambillau, C. Proteins Struct Funct Genet 1998, 31, 320-333.

18. Gallagher, W.H.; Croker, K.M. Protein Sci 1994, 3, 1602-1604.

19. Fitzgerald, P.M.D.; McKeever, B.M.; VanMiddlesworth, J.F.; Springer, J.P.; Heimbach, J.C.; Leu, C.T.; Herber, W.K.; Dixon, R.A.F.; Darke, P.L. J Biol Chem 1990, 265, 14209-14219.

20. Kossiakoff, A.A.; Randal, M.; Guenot, J.; Eigenbrot, C. Proteins Struct Funct Genet 1992, 14, 65-74.

21. Chothia, C.; Lesk, A.M. EMBO J 1986, 5, 823-826.

22. Ala, P.J.; Huston, E.E.; Klabe, R.M.; McCabe, D.D.; Duke, J.L.; Rizzo, C.J.; Korant, B.D.; DeLoskey, R.J.; Lam, P.Y.S.; Hodge, C.N.; Chang, C.H. Biochemistry 1997, 36, 1573-1580.

23. Baldwin, E.T.; Bhat, T.N.; Liu, B.S.; Pattabiraman, N.; Erickson, J.W. Nat Struct Biol 1995, 2, 244-249.

24. McLachlan, A.D. Acta Crystallogr, Sect A: Found Crystallogr 1982, 38, 871-873.

25. Martin, A.C.R. ProFit V1.8. University College London, 1998; http://www.biochem.ucl.ac.uk/~martin/programs/.

26. Rodier, F.; Crosio, M.P.; Chiadmi, M. Acta Crystallogr, Sect A: Found Crystallogr 1990, 46, 37 (suppl).

27. Huang, C.C.; Pettersen, E.F.; Klein, T.E.; Ferrin, T.E.; Langridge, R. J Mol Graphics Modell 1991, 9, 230-236.

28. Wlodawer, A.; Miller, M.; Jaskolski, M.; Sathyanarayana, B.K.; Baldwin, E.; Weber, I.T.; Selk, L.M.; Clawson, L.; Schneider, J.; Kent, S.B.H. Science 1989, 245, 616-621.

29. Weber, I.T.; Miller, M.; Jaskolski, M.; Leis, J.; Skalka, A.M.; Wlodawer, A. Science 1989, 243, 928-931.