SIMS: A Hybrid Method for Rapid Conformational Analysis

Bryant Gipson, Mark Moll, Lydia E. Kavraki
Department of Computer Science, Rice University, Houston, Texas, United States of America

Abstract

Proteins are at the root of many biological functions, often performing complex tasks as the result of large changes in their structure. Describing the exact details of these conformational changes, however, remains a central challenge for computational biology due the enormous computational requirements of the problem. This has engendered the development of a rich variety of useful methods designed to answer specific questions at different levels of spatial, temporal, and energetic resolution. These methods fall largely into two classes: physically accurate, but computationally demanding methods and fast, approximate methods. We introduce here a new hybrid modeling tool, the Structured Intuitive Move Selector (SIMS), designed to bridge the divide between these two classes, while allowing the benefits of both to be seamlessly integrated into a single framework. This is achieved by applying a modern motion planning algorithm, borrowed from the field of robotics, in tandem with a well-established protein modeling library. SIMS can combine precise energy calculations with approximate or specialized conformational sampling routines to produce rapid, yet accurate, analysis of the large-scale conformational variability of protein systems. Several key advancements are shown, including the abstract use of generically defined moves (conformational sampling methods) and an expansive probabilistic conformational exploration. We present three example problems that SIMS is applied to and demonstrate a rapid solution for each. These include the automatic determination of “active” residues for the hinge-based system Cyanovirin-N, exploring conformational changes involving long-range coordinated motion between non-sequential residues in Ribose-Binding Protein, and the rapid discovery of a transient conformational state of Maltose-Binding Protein, previously only determined by Molecular Dynamics. For all cases we provide energetic validations using well-established energy fields, demonstrating this framework as a fast and accurate tool for the analysis of a wide range of protein flexibility problems.

Introduction

Proteins lie at the root of nearly all biological processes and often accomplish functions through conformational changes in their structure. An understanding of conformational variability therefore would provide valuable insight into protein function, in addition to aiding pharmaceutical drug design – given that drug binding sites often become exposed as the result of conformational changes. The development of computational methods for the analysis of protein flexibility has a long history [1–3], with several broad classes of analytical frameworks having been developed over the years. Rigorously accurate, yet computationally demanding physics-based methods were among the first and best attempts to address such questions by solving equations of motion defined by a particular protein system. While definitive for high-resolution and physically accurate interpretations, such methods have typically been limited by protein size due to computational complexity [4,5]. More recently, a class of methods has been developed that use approximations to quickly provide analytical insight into key biological processes. This class includes a broad range of methods, such as coarse-grained energy calculations [6], multi-scale models [7] and alternative representations of flexibility, such as Normal Mode Analysis [8–11] and Dynamic Elastic Networks [12–14], among others.

Recently, a hybrid class of mechanistic approaches has gained traction for the analysis of molecular structures, inspired by the field of robotic motion planning [15,16]. Such methods attempt to bridge the divide between the above classes and are capable of using highly accurate energetics for representation, while additionally employing long-range moves for conformational exploration. In the motion planning inspired approach, molecules can be regarded as long articulated chains with atoms as links and bonds as joints. Using this representation, the energy of a particular protein conformation (computed using any available method) is used as a selection criterion during conformational exploration.
Exploration can occur by sampling new conformations through the perturbation of known “good” conformations using any available move. The resulting conformation is checked for feasibility by the provided energy function. If it is feasible, it is added to the set of “good” conformations.

The central strength in motion planning-inspired approaches lies in their ability to adaptively guide exploration based on estimates of the density of known conformational samples. Typically, they use some notion of coverage to “push” the exploration away from well-explored conformations (i.e., redundant and highly similar sampled states) and towards unexplored parts of conformational space. This process can rapidly lead to an increasingly accurate approximation of the local conformational flexibility of a protein and typically operates orders of magnitude faster than a random thermodynamic walk [17].

While motion planning-inspired methods are not designed to specifically model physically accurate molecular motions, they are capable of rapidly producing a representative approximation of the local conformational variability of a protein under study. Motion planning has recently been applied to a wide range of biologically important subjects including RNA folding [18], protein loop modeling [19–22], protein folding/bindng [23–25], conformational flexibility [17,26] and conformational transitions [27,28], among others.

This paper introduces a highly general framework, the Structured Intuitive Move Selector (SIMS), used for the automatic or expert-guided discovery and analysis of the conformational variation of arbitrary protein molecules. We demonstrate several key advances that allow us to revisit and significantly improve upon results obtained by earlier robotics based methods. We show that SIMS can identify and use “active” residues (i.e., residues most likely to be involved in conformational transitions) in the exploration of hinge-based systems such as Cyanovirin-N. SIMS is also shown to be capable of identifying significant, long-range, correlated changes in Ribose Binding Protein and is shown to discover a “hidden” (experimentally unobserved) conformation of Maltose-Binding Protein, at a fraction of the computational cost of Molecular Dynamics (MD) simulations.

The contributions of SIMS as a method can be summarized as follows. It adopts a state-of-the-art motion planning algorithm for conformational sampling. It also introduces structured local move selection: a unified approach to intelligently perturbing conformations to obtain new conformations that combines loop sampling, energy minimization, and dihedral angle sampling. Thanks to the level of abstraction the approach provides, other moves can easily be added. The moves are applied to protein “fragments,” groups of possibly non-contiguous residues meant to approximate functional, structural or dynamically correlated regions of the protein. The decomposition of a protein into fragments can be done automatically, but allows an expert user to define fragments as well. Finally, SIMS is designed to run in parallel and requires only minimal communication, allowing it to be run on a large scale.

Generic planning algorithms for conformational sampling

The initial ideas regarding the application of robotic motion planning to proteins were introduced in [29] and used the Probabilistic Roadmap Method (PRM) [30] to build a roadmap for the motion of a small ligand around a protein. The roadmap is a graph representation of conformational transitions, where each node represents a conformation and each edge a transition between two conformations. During the construction of such roadmap, an energy function is used to verify whether a conformation or transition is biophysically plausible. If a conformation or transition is not feasible, it is simply discarded. These initial results followed from significant advances in motion planning around the same time. Rather than developing algorithms for exact, optimal solutions (which is, computationally, prohibitively expensive), motion planning research shifted in the 1990s to the development of sampling-based planning algorithms, which have been very successful in practice and are currently the main way to plan paths for complex robots. Subsequent work on applying sampling-based motion planning to conformational sampling [31] introduced the stochastic roadmap simulation, established the connection with Monte Carlo methods and dealt with problems involving conformations of much larger protein molecules, prms for the computation of folding pathways given the 3D structure of the protein have also been investigated at length in a series of papers that span a decade (see [32] for a detailed discussion). This line of work has provided important insights into the order of formation of secondary structures that agree with experiments [24]. Two recent surveys [33,34] provide an extensive overview of geometric and kinematic modeling of protein structures as well as the application of motion planning techniques for modeling protein motion. Below, we give a brief overview of such algorithms. The algorithm used in this paper will be described in more detail in Methods.

In recent years, specific motion planning algorithms have seen significantly increased use with regard to the protein flexibility problem. In particular, the application of the Rapidly-exploring Random Tree (RRT) algorithm [35] to molecular simulations has expanded dramatically. This algorithm attempts to explore protein conformational variabilty by growing a tree of conformations, starting from a known structure. The algorithm iteratively samples a uniformly random conformation, finds the most similar conformation in the tree, and extends the tree from this conformation towards the random conformation. The transitions between conformations are typically obtained by simple interpolation of the Degrees of Freedom (DOFs). Protein loops have been successfully analyzed using this method [19] (though generating the “random” loop conformations required special attention). More recently, long-range protein conformational analysis has been performed [27]. To reduce the computational cost, the authors used a priori information in the form of “predicates” to solve certain highly constrained planning problems (see Results). This work highlighted that RRT-based approaches are difficult to scale up to proteins with hundreds and hundreds of thousands of DOFs. Perhaps this is to be expected as protein conformations with uniformly random backbone angles almost always represent an unfolded protein, often with many steric clashes. Moving toward random conformations may therefore not represent an ideal method for efficiently exploring conformational changes. In our implementation we use a recently proposed alternative to RRT called Kinodynamic Planning by Interior-Exterior Cell Exploration (KPECE) [36] which is a member of a class of expansive planners [37]. This specific algorithm will be described in more detail in Methods. Like RRT, expansive planners grow a tree of conformations. Unlike RRT, these planners use estimates of local state density to push tree growth towards unexplored regions of the conformational space (i.e., regions with low density). While RRT and expansive planners may seem somewhat similar, they exhibit markedly different behavior in practice, especially as the number of DOFs increases.

The mechanism that expansive planners use to create a new conformation in a neighborhood of a previously generated conformation can incorporate techniques that increase the probability of sampling energetically feasible conformations. In this work, we define a library of moves that each individually has
been used in prior work for conformational sampling, but not in an integrated way as is done here. This library includes: energy minimization, loop sampling [22], random dihedral angle perturbation, and "natural moves" similar to [38]. An expansive planning algorithm thus grows a tree of conformations that preferentially expands away from a set of starting states towards less-explored regions of the energetic landscape.

Proteins and energy functions
Typically, important biological functions are performed by folded, compact proteins existing in one of a few stable conformations available at cellular conditions. Stable conformational ensembles represent groups of protein states at low free energy and are typically associated with basins about the minima of the potential energy field [39,40]. An understanding of protein stability therefore requires an accurate notion of potential energy. Many potential energy functions have been proposed (see [41,42] for detailed discussions), typically for md simulations. Energy calculation typically represents the largest computational cost when modeling changes in proteins, and the use of the above models can prove prohibitively expensive. While sim is not restricted to any particular energy function, in this paper we rely on the Rosetta library, which contains efficient implementations of many full-atom energy models, striking a good balance between accuracy and speed of computation.

As in earlier work, “active” DOFs are limited to the $\phi$ and $\psi$ backbone angles [17,27,28,44] and side-chain positions are automatically determined by Rosetta’s side-chain minimization protocol [27]. We used the Rosetta “score12_full” energy function for the experiments, which provides an atomic representation of all atoms, implicitly modeling solvation and related energetic terms. At the end of the paper we show energy validations against a variety of other methods, demonstrating that Rosetta energy calculations were sufficiently accurate for the studies in this paper.

Motivating problems
To demonstrate the range and generality of analysis that sim can provide, we present three important problems often encountered in computational biology. Below, we introduce protein systems that are shown to characterize these problems and in later sections present results for each. The first two problems have been previously studied by a related robotic motion planning-inspired method [27], and were specifically chosen to enable a direct comparison. These problems involve the use and determination of active DOFs, especially in the context of previously defined (and possibly incomplete or inaccurate) expert knowledge. By active DOFs we mean a set of dihedral angles from a range of residues that represent the minimal set of angles that must change in order to allow a particular type of conformational transition to occur. The final problem has been investigated primarily by md and, though sim is not designed as an alternative to such methods, is presented as a case where sim can be used to replicate valuable conformational insights quickly and automatically.

Cyanovirin-N (CVN) is a two-domain bacterial anti-viral protein, capable of binding to the surface sugars of a range of viruses including HIV. CVN is known to occur in monomeric [47] and domain-swapped [48] forms, with the domain-swapped conformation found to possess higher anti-viral affinity than the monomer [49]. It is known that these two conformations co-exist in solution [49] and transitions between them were previously computed [27], but depended on expert knowledge.

Ribose-Binding Protein (RBP) is part of a ribose transport system in bacteria and is additionally involved in chemotaxis. It is composed of two domains connected by a hinge formed by three well-separated loops. Both closed [50] and open [51] forms of RBP are known for this system. While the active DOFs in this system are known to occur almost exclusively in the hinge region, domain movement can only occur as a result of coordinated motion among the three loop regions. The two forms of RBP are separated by just over 4Å, and the required domain transition seems deceptively simple; a visually convincing transition between the forms can be quickly computed with, e.g., UCSF Chimera [52]. However, solving this problem in an energetically feasible manner that preserves the kinematic bond structure of the protein is quite challenging. As a result, prior work [27] relied on artificial distance restraints to maintain “reasonable” structures during sampling.

Maltose-Binding Protein (MBP) is a well-studied bacterial protein involved in chemotaxis, biosensing, the maltose/maltodextrin system of E. coli and is also used as an affinity tag in protein purification and expression. MBP is important for biological and experimental reasons. Though many structures have been determined for MBP by X-Ray crystallography and other methods, most of these fall into the classes of “open” and “closed” states, as determined by the degree of bending between the C and N terminal domains. A third “hidden” semi-closed intermediate was recently identified by accelerated MD [53], though it had been previously identified by NMR [54] and earlier computational studies [55]. While ligand binding is known to drive conformational change, NMR [54] studies have shown that MBP exists in solution in a mixture of these states. An analysis of the available set of MBP proteins [56-84] shows a high degree of spatial and torsional variation for essentially all residues, excepting several short stretches in core helical regions. Further, the difference between open and closed forms is one of tightly constrained long-range “bending” occurring across the entire molecule, as opposed to simple rigid body changes in sub-domains, producing extensive side-chain interactions. This system represents a difficult challenge in that no clear set of active DOFs exists and the motion is extremely coordinated.

Methods
The central problem we address here is of how to vary the DOFs of a protein in such a way that the energy never exceeds biologically feasible bounds when attempting to find low-energy conformational transformations between known states. As was done in prior work [17,27,28,44,65], we represent a conformation of a protein by just the backbone angles. The positions of side-chain atoms for any given conformation are determined by side-chain optimization and bond angles and lengths are always idealized. This representation significantly reduces the computational difficulty of the problem.

Below we first describe the primitive “moves” that will be used to perturb conformations. These moves typically do not affect the entire structure, but instead correspond to local changes. We propose a way to automatically define a collection of residue subsets called a schema on which the moves operate. Finally, the high-level planner maintains state density estimates which it uses to apply moves to conformations in relatively sparsely sampled parts of the conformational space.

Structured move selection
Computationally generating new conformations based on known states involves applying some type of perturbation of the DOFs of the system. We call such a perturbation a move. Many
different types of moves have been proposed, including dihedral perturbations [86] and Normal Modes [8–11], as well as moves based on Dynamic Elastic Networks [12–14,87], to name a few. In our method, moves can be applied to both small protein fragments (such as loop regions) and the whole structure. We use a schema to define subsets of DOFs on which moves operate. Such a schema can automatically be constructed based on the structure of a protein. For example, one can define a subset for each domain, each secondary structure element or even each residue. Each residue can be part of multiple subsets. Often, an expert may wish to define additional subsets. For example, the three loop regions in RBP that connect two domains can form an additional subset, since motions of the DOFs within that subset are highly coordinated. Note that a subset of residues does not need to correspond to a contiguous sequence of residues. Associated with each subset is a probability for selecting that subset for a move. These probabilities can be defined heuristically based on what is known from the literature about the relative flexibility of, e.g., secondary structure elements: flexible loop fragments will be sampled with a higher probability than more rigid alpha helices. Associated with each subset is a probability distribution over the “allowed” moves. In our experiments described below we used the following moves:

**Dihedral angle sampling.** This is simply a uniformly random perturbation (up to 6°) of each dihedral angle within a subset.

**Loop sampling.** Here, a random conformation of a loop region (or collection of loop regions) is generated, subject to the constraint that the endpoints of each loop are kept in the same position.

**Rigid body movements.** This type of move corresponds to a small displacement of one loop endpoint relative to another while maintaining the kinematic constraints of the loop. This move enables fast sampling of whole domain rearrangements.

**Energy minimization.** This move is applied with low probability to the entire protein since it is computationally expensive.

A schematic overview of a schema and move selection is shown in Figure 1. Although the figure shows a hierarchical decomposition, this does not have to be the case (unlike [88]). As mentioned, a default schema can be automatically computed from the primary structure, but expert knowledge can easily be incorporated as well. Not only can extra subsets be defined, also the types of moves and the probabilities of selecting a move can be changed, if there exists prior knowledge about a suspected mechanism underlying some conformational change.

**Rosetta**

The Rosetta Library [43] has been applied to a considerable number of protein systems and problems in recent years [89–93], due to its powerful algorithmic flexibility and extensive library of protocols for protein modeling. While not strictly dependent on the library, SIMS is able to take advantage of Rosetta for structure representation and modification as well as minimization and energy analysis. This allows any experiment performed with SIMS to be run in centroid mode or with the full atom representation mode, along with user-specified weightings to energy terms as needed (though the “score12_full” scoring function was used for all simulations presented here). Moreover, conformational sampling can be performed by taking advantage of the extensive library of moves available in Rosetta’s sampling protocols, including minimization, CCD loop closure [94], and loop-sampling [95]. The SIMS moves described above have been implemented using Rosetta’s moves. It is important to note, however, that any alternative representation or energy calculation library could have been used in its place.

**Efficient conformational sampling using a motion planning algorithm**

The moves described above can be used by an expansive motion planning algorithm to grow a tree of conformations, where each conformation is derived from its parent through a move. Many robot motion planning algorithms have been proposed over the years, and many of them are implemented in a very abstract way in the Open Motion Planning Library (ompl) [96]. This level of abstraction makes it possible to adapt them for conformational exploration. While in robotics, a collision checker is often used to decide whether a robot configuration is valid, here we use an energy threshold as a criterion for accepting sampled conformations. We used ompl’s default high-dimensional planner, called kPIECE [36], for all experiments presented in this paper. kPIECE has previously been shown to be very effective in high-dimensional spaces, including kinematic chains of rigid bodies – systems similar to proteins. We will give a brief description of kPIECE algorithm below; for details see [36].

kPIECE approximates the density of sampling of the conformational space through a projection of all the DOFs. High-dimensional systems are often constrained to move on a low-dimensional manifold embedded in a high-dimensional space. Proteins are no exception: once proteins are folded the DOFs are often very constrained. Using a low-dimensional projection allows for efficient estimation of sampling density. The default projection we have defined is a random, linear 2D projection of the cosines and sines of the dihedral angles. This projection is computed as follows. For a conformation with \( n \) dihedral angles, a vector of size \( 2n \) is computed with the cosines and sines of all angles. This vector is projected to a 2D point with a matrix \( P \) of size \( 2 \times (2n) \). The matrix \( P \) is constructed by first drawing its entries from a normal distribution with mean 0 and variance 1. Next, the first row is normalized to be of length 1. Finally, the second row is made orthogonal to row 1 and then also normalized. This process can be generalized to any \( m \times (2n) \) projection matrix. The projection is chosen randomly because (a) there is no natural choice of projection in general and (b) prior work has shown that a random projection often captures sample density quite well compared to an optimal or expert-chosen projection [97]. Given a 2D projection, all conformations can (for the purpose of density estimates) be represented by 2D points. kpiCE defines a 2D grid and maintains a count of the number of conformations per grid cell. It then (1) samples a grid cell with probability inversely proportional to its density, (2) samples a conformation uniformly at random from that cell, (3) applies a random move selected in the manner described in the previous section, and (4) checks if the conformation’s energy is below a user-specified threshold. If the new conformation is accepted, it is connected to its parent conformation and inserted into the grid. This process continues until a desired conformation is reached or a time limit is reached.

The sampling of grid cells is actually slightly more complicated than described above. For each grid cell the algorithm also keeps track of the number of neighboring grid cells that are empty (i.e., ones that contain no conformations). Non-empty grid cells with at least one empty neighbor grid cell are called exterior cells while the other non-empty grid cells are called interior cells. The sampling of grid cells is heavily biased towards exterior cells to improve the expansiveness of the conformational search. Note that the sampling bias towards low-density and exterior cells does not
preclude exploration of higher-density and interior cells, albeit with a lower probability.

The overall behavior of the algorithm can be summarized as follows. The conformational sampling algorithm requires as input one or more known structures and a schema that defines the subsets of residues and associated moves. It then performs an expansive conformational search by iteratively applying a random move to a previously generated conformation. Conformations are selected inversely proportional to the local conformation density. This process can be considered an undirected search: the algorithm attempts to expand the tree of conformations equally in all directions. It is also possible to provide a goal conformation and have the search bias sampling with a small probability towards this goal conformation. This is called a directed search. (In robot motion planning, this is in fact the more common use case.) These modes of operation, directed and undirected search, can also be combined: a directed search can be performed first to find a transition between two conformations, and a transition envelope can be subsequently (or simultaneously) explored using an undirected search. Such generality allows for rapid exploration of conformational variability, both between and near known structures, as well as into unknown regions where experimentally unobserved (yet energetically stable) conformations may be hidden.

To enable analysis of extremely large systems, SIMS has been written to take advantage of all available computational resources (clusters, desktops, laptops) simultaneously and without special configuration. This is achieved by having each computational core perform a small run of SIMS and write the generated conformations back to a central database. The density estimates are then updated and a core can pick a random starting conformation from the database in a sparsely sampled part of the conformational space.

Storing all generated data in a database also permits real-time analysis during a run.

Results

In our computational experiments we explore how well SIMS performs with different schema s. The automatically-generated schema is defined as follows. There is a subset for each secondary structure element and one set containing all residues. Each loop, sheet, and helix has a sampling weight of 1.0, 0.2, and 0.1, respectively. With 9% probability the set with all residues is selected, while the remaining probability mass is distributed over the secondary structure elements proportional to their weight. The set of moves and their relative probabilities for each subset are the same: dihedral angle sampling, loop sampling, and rigid body movements are all sampled with equal probability. Note that loop sampling and rigid body movements are also applied to sheets and helices to allow these secondary structure elements to dissolve. However, since the sampling weight of loops is much larger, most of the conformational sampling is focused on loop changes. The energy (as a function of all backbone angles) is minimized 1% of the time. The expert-informed schema s described below either limit the degrees of freedom by only allowing moves for a small number of subsets or define additional subsets for residues whose motion need to be coordinated. In the first case, we can potentially explore the conformational space faster, but we risk eliminating a motion that is necessary for some conformational transition. In the second case, we simply encourage sampling particular degrees of freedom but do not sacrifice completeness of the algorithm.

In this work, all experiments were run on a multi-core cluster, typically using 200 cores. Though the times described in this section are measured in hours (assuming 200 cores) it should not be assumed that the problems necessarily represent 200×
flexibility that help overcome steric constraints in transitions and two secondary flex regions required for "breathing" primary flexibility arises from a central hinge.

Difficulty of this problem. The second CVN experiment therefore the search space, planning with flex.

Transition between the monomeric (PDB:2EZM) and domain-three experiments the goal is to find a low-energy conformational (2) the hinge and flex regions, or (3) all residues, respectively. In all experiments the goal is to find a low-energy conformational transition between the monomeric (PDB:2EZM) and domain-swapped (PDB:1L5E) forms of CVN. We are interested in how fast swapped (PDB:1L5E) forms of CVN. We are interested in how fast the hinge region (0.16), each individual flex region (0.16 each), a subset containing the hinge region and both flex regions (0.50), and the set of all residues (0.01). Each subset has the default moves, except for the set of all residues, which only has a minimization move. This experiment took approximately 1.3 hours and showed almost identical torsional activity in the hinge region to the previous experiment, including 26–35 (see Figures 2 and 3). The flex regions, however, were very active in this run, though the expanded conformational freedom in these regions produced a 3-fold computational increase relative to the first experiment. The increase in computation time, combined with the findings from the first experiment – that the flex regions were not strictly required for solving this problem – appears to imply that the flex regions in fact do not play a significant role in the transition between monomeric and domain-swapped forms of CVN. This conclusion was reinforced by the results of the final experiment for CVN.

The final experiment for CVN involved a case where no expert knowledge was assumed. In this case, the automatic schema of DOFs was applied to the system, alongside a second, overlapping subset of DOFs composed of all residues – moves for this subset were sampled at 10% the rate of the automatically partitioned subset. This resulted in searching the full 198 DOFs for CVN. Again a transition was determined, this time in around 30 minutes, a similar time to the first experiment, despite searching with a number of DOFs nearly an order of magnitude greater than before. That is, though the hinge region performed a search using 20 dihedral angles and this experiment used 198, computation times were nearly identical. In this case, the transition determined employed nearly all torsional dofs, with the exception of those found within rigid sub-regions and, surprisingly, the flex regions (see Figures 2C and 3C).

All three runs were qualitatively similar at their start and end points, with the beginning of the paths defined by slow progression away from the highly-constrained starting state, and the end of the path characterized by alignment with the final position and a slow counter-rotation of the first and second half of the central hinge. The middle of the paths were relatively unconstrained with rotation mostly about the hinge. Qualitative transition smoothness was clearly the best for the auto-schema experiment, likely due to the availability of full conformational freedom.
Analysis of residue level torsional changes (Figure 2) for the three experiments revealed a number of common and unique features. Here, cumulative, residue-wise torsional change for residue $i$ was calculated according to $\sum_{j=1}^{n} |\phi_{i,j} - \phi_{i,j-1}| + |\psi_{i,j} - \psi_{i,j-1}|$, where $j$ is the index of the conformation along the path. Unsurprisingly, the hinge region was active for all runs, though significantly less motion occurred in this region in the automatic run. As shown in the experiments, 26–35 represented secondarily important residues in all runs. More generally, the most active regions outside of the hinge for the auto-schema experiment were residue ranges 26–35 and 75–87, representing anti-complementary halves of sub-domains A and B respectively (i.e., one half of the $\beta$-sheets defining these domains). It is clear from this analysis that the hinge region of CVN plays a dominant role in driving conformational transitions though, based on the results (and combined with the relative smoothness of the final experiment), there is also a large-scale sub-domain flexing that appears to aid this process.

Finally, all conformational transitions were analyzed using the Amber99 force field to calculate energies for the entire transition (Figure 4). Energies calculated for the raw output of SIMS were occasionally quite high, likely indicating some level of steric overlap between neighboring atoms. Using 100 steps of energy minimization always yielded an extremely low-energy structure, occasionally quite high, likely indicating some level of steric overlap between neighboring atoms. Using 100 steps of energy minimization always yielded an extremely low-energy structure, reflecting the relative distance of two domains. The transition between the two forms of RBP seems relatively simple, given that the two conformations are only 4Å apart, but computationally producing energetically feasible transitions presents a formidable challenge (described more fully in the section Motivating Problems). Similar to the previous example, the goal is to compare an expert-determined schema with the default one. The expert-determined schema consists of a number of DOFs used and tightness of the energy constraint, it is very likely that both transitions represent slight variations of the hinge region. Somewhat surprisingly, the transition determined by the expert guided run was essentially identical to the automatically guided run, with slightly more domain level variation occurring in the expert run (see Figure 5). Given the large differences in the number of DOFs used and tightness of the energy constraint, it is very likely that both transitions represent slight variations of the minimum energy transition between these two states.

This experiment demonstrated that in spite of the significant kinematic challenge of making coordinated changes to non-sequential hinge residues using torsional DOFs, SIMS is able to rapidly determine solutions using only unbiased energetic constraints, requiring no a priori knowledge.

Ribose-binding protein

RBP is known to exist in bound (PDB:1URP) and unbound (PDB:2DR1) states, reflected by the relative distance of two domains and the volume of the ligand binding space between them. Movement between the domains occurs via coordinated changes in three non-sequential loop regions connecting the two domains. The transition between the two forms of RBP seems relatively simple, given that the two conformations are only 4Å apart, but computationally producing energetically feasible transitions presents a formidable challenge (described more fully in the section Motivating Problems). Similar to the previous example, the goal is to compare an expert-determined schema with the default one. The expert-determined schema consists of a number of DOFs used and tightness of the energy constraint, it is very likely that both transitions represent slight variations of the hinge region. Somewhat surprisingly, the transition determined by the expert guided run was essentially identical to the automatically guided run, with slightly more domain level variation occurring in the expert run (see Figure 5). Given the large differences in the number of DOFs used and tightness of the energy constraint, it is very likely that both transitions represent slight variations of the minimum energy transition between these two states.

This experiment demonstrated that in spite of the significant kinematic challenge of making coordinated changes to non-sequential hinge residues using torsional DOFs, SIMS is able to rapidly determine solutions using only unbiased energetic constraints, requiring no a priori knowledge.

Maltose-binding protein

In [33] it was shown that a “hidden” energetically semi-stable conformation of MBP likely exists as an intermediate between known open and closed forms that has been only indirectly observed experimentally. Described as “semi-closed”, this distinct state is characterized by changes in the so-called balancing interface, a loop region that acts as a “spring” between the C-terminal and

Figure 3. Plots of angular change in DOFs for hinge (A), hinge-flex (B) and auto (C) experiments. Residues are colored by absolute total angular change, with blue indicating a small change and red a large change. Hinge and Flex (A, B) experiments show relatively low activity outside of the planning regions. The automatically guided experiment (C) shows high activity in the hinge and $\beta$-sheet regions of both subdomains.

doi:10.1371/journal.pone.0068826.g003
N-terminal domains. The goal of the experiments described here was to see if this hidden state could be determined using SIMS, when searching for a direct transition between open and bound forms of MBP. Changes between known open and bound forms of MBP represent a dominant bending deformation across the entire protein that involves changes in nearly all residues. As a result, no expert-determined set of active DOFs was available for this system and an automatic schema was used. However, we will demonstrate that an initial run of SIMS can be used to determine active DOFs. This is in itself may provide useful insight into the mechanism of MBP's function, but we will show that this can also be used to create a new schema that enables for a more rapid exploration of conformational space.

The first experiment used the default schema to find a transition from the unbound form (PDB:1OMP [73]) to the bound form (PDB:3MBP [84]). The search took approximately 15 hours to complete, coming to a state less than 1Å away from the goal – after which progress became significantly slower. The differences between the final state and the goal state were observed to occur...
almost exclusively around the end of the balancing interface loop region (Figure 6).

To visualize how SIMS has explored the conformational space, we computed a low-dimensional embedding of all conformations using Principal Component Analysis (PCA) [98]. Specifically, pca was applied to the Cartesian coordinates of all conformations generated during the search. By plotting each conformation as a point with coordinates given by the first two principal components we obtain a low-dimensional embedding of the conformations (see Figure 7). Similar to the results of [53], the open and bound forms of MBP were observed to cluster into two relatively tight groups, with the conformational transition (the red path in Figure 7) tracing a nearly direct transition between the two groups. Almost identical to the md results of [53], a large, relatively stable basin of intermediate conformations was observed almost directly between the bound and unbound groups. Calculating the centroid state of a representative set of low energy conformations from this basin yielded a structure that matched a known NMR structure [80] (PDB:2H25) for the “semi-closed” state of MBP to within the resolution of the experiment (Figure 8). Moreover, the low-energy conformational transition determined in this experiment was also found to pass extremely close to this state (to within less than 1Å full atom RMSD), lending likelihood to the proposition that the semi-closed state of MBP represents a necessary transition intermediate between open and bound forms. These results were quickly determined using only an automatically generated schema, producing both a low-energy conformational transition between known states of MBP as well as a model for the semi-closed transition intermediate.

As is clear from energetic analysis of the input conformations (and basic biological intuition), the bound forms of MBP represent relatively high-energy conformations if the ligand is removed. The first piece of expert knowledge for the final experiments therefore involved reversing the start and goal states, starting instead at the bound form of MBP with a goal of reaching the open state – essentially removing the ligand and observing the energetic consequences. Using the same schema as before, the reversed search took approximately 5 hours to get within 1Å of the goal state (whereas the first took 15 hours). In the previous case, the primary difference between the final state and the goal lay in the tip of the balancing interface (Figure 6), possibly due to energetic stabilizing factors in this region in the two forms, or an insufficiently resolved loop sampling schema in this region. Further, as before, the transition also passed within 1Å of the

Figure 6. Comparison of the final state of the reverse transition (blue) and the direct transition (red). Structures are nearly identical save for a 10 residue relaxation of a loop region in the balancing interface. Both transitions come to within 1Å of their goal.
doi:10.1371/journal.pone.0068826.g006

Figure 7. PCA landscape of all conformations generated in the MBP experiments. Each point represents a unique conformation. The color indicates energy with darker colors representing more energetically stable states. The red path shows the path found with the default schema, starting from the open state with the bound state as the goal. The green path represents the reverse case, with the default schema, starting at the bound state and moving toward the open state. The purple path was produced using the expert-schema, moving from the bound state towards the open state. The yellow star indicates the centroid conformation of the energetic valley between the open and bound states and falls extremely close to all paths, as well as the NMR structure. The circular pattern in the green path was automatically generated and seems to arise from a slight bending reversal that occurs near the semi-closed state.
doi:10.1371/journal.pone.0068826.g007
Figure 8. Comparison of a computationally identified intermediate state and NMR structure PDB:2H25. The identified intermediate state (dark blue) corresponds to the centroid of the low energy region shown in Figure 7. The NMR structure is known to be close to the semi-closed state and traced an essentially direct transition between known open and bound forms (aqua structure) of MBP to PDB:2H25 shows almost perfect agreement with the centroid structure. doi:10.1371/journal.pone.0068826.g008

semi-closed state and traced an essentially direct transition between the bound form of MBP and the open group (green path in Figure 7).

For the final experiment we determined the active residues as measured by total torsional change per residue along the path found in the first experiment. The active residue ranges identified were: 101–104, 234–236, and 261–262. The schema we used, based on this information, included two subsets of residues: one with all the active residues (with the default set of moves) and the set of all residues (with only energy minimization, selected 1% of the time). With this schema it took approximately 2.5 hours to find a path from the bound to a state within 1 Å of the unbound form. This path showed identical features to the previous two transitions (see purple path in Figure 7).

This collection of experiments demonstrated that, even absent expert knowledge about a protein system, SIMS can rapidly generate detailed information about low-energy conformational transitions. The conformational information generated during the search was also shown to be useful for conformational analysis, producing results typically requiring experiment or long-running MD simulation. Finally, expert knowledge was generated from the initial investigation and subsequently used to generate information for new experiments. Besides dramatically improving experimental run times, this expert knowledge serves as a result in itself that could be applied as a constraint in future computational investigations by alternative methods.

Discussion

In this work we have introduced a hybrid method for rapidly analyzing the conformational variability of proteins that combines all-atom energy calculations with abstractly defined long-range moves for conformational sampling. SIMS allows for rapid conformational exploration of input protein systems, producing an increasingly accurate sampling of the energetic landscape. While this method is not a replacement for MD or approximate methods such as Normal Mode analysis, SIMS represents a powerful intermediate tool that benefits from aspects of both.

Moreover, output from SIMS can easily be used as a launching point for more rigorous investigation using physics-based methods, reducing the substantial computational cost such investigation of long-range conformational variability would typically require.

We applied SIMS to three common classes of problems in computational biology: a hinge system, a non-sequential long-range correlated motion problem, and the discovery of a “hidden” conformational state of a protein. The demonstrated solutions to these problems were found rapidly and with minimal information as the result of a number of key features of the presented framework. The inclusion of a powerful schema based on collections of subsets of dofs, to aid successful move selection, simultaneously allowed the incorporation of expert knowledge while allowing likely active dofs to be rapidly explored.

We have shown that while this framework can benefit from expert knowledge when available, it is also capable of investigating systems about which little is known. In such cases automatic generation of a schema, as described earlier, can be used to perform initial explorations and, subsequently, determine active dofs from initial results. In the case of CNN, we showed a key example of how incomplete expert knowledge could negatively influence results and how automatic partitioning was used to refine this information.

Finally, we showed through the MBP experiments that, while not a replacement for MD, SIMS can provide insight into a number of problems that have been traditionally studied by such methods. The ability to rapidly discover transient conformational intermediates (or at least to characterize a range of nearby neighbors), with minimal user input, presents a powerful extension to the range of analytical tools available to researchers.

Future directions

Relative to the available computational power provided by the Rice University clusters (and eventually larger national computing clusters), the systems investigated here are likely far smaller than the limit of computationally tractability for this framework. Future studies will likely focus on significantly larger systems, or more complex problems (such as docking and protein-protein interaction).

While Rosetta proved both powerful and efficient for energy calculation and move generation, the move protocols used here were not necessarily tailored to the protein systems presented here. As the community continues to generate increasingly powerful move types, we hope to continuously extend the exploratory power of SIMS by including such developments into the framework.

Finally, the analysis performed on the datasets presented in this work, while extensive, only hints at the full range of options that could be used. Analyzing the graph-structure of the conformational exploration and casting the network as a Markov Process has previously demonstrated useful theoretical results [25,31,99–101]. Though PCA was used for analysis of the conformational states generated, non-linear analysis of the energetic landscape [102] is another obvious direction for investigation. Most importantly for usability, however, will be a range of visualization output options, possibly benefiting from real-time (i.e., during data generation) interaction with intermediate results. It is expected that this improvement will likely provide the most important development for the computational biology community at large.

Conclusions

The work has demonstrated the power and flexibility of a hybrid method for the investigation of protein conformational variability. Naturally integrating expert knowledge with automatic exploration allows both ease of use and the ability to account for
References

1. Adock S, McCammon J (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev 106: 1589–1615.

2. Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. Nature 450: 964–972.

3. Marsh JA, Teichmann SA, Forman-Kay JD (2012) Probing the diverse landscape of protein exibility and binding. Curr Opin Struc Biol 22: 643–650.

4. Johnston JM, Filizola M (2011) Showcasing modern molecular dynamics simulations of membrane proteins through G protein-coupled receptors. Curr Opin Struc Biol 21: 552–8.

5. Piana S, Lindorff-Larsen K, Shaw DE (2012) Protein folding kinetics and thermodynamics from atomistic simulation. P Natl Acad Sci USA 109: 17845–50.

6. Takada S (2012) Coarse-grained molecular simulations of large biomolecules. Curr Opin Struc Biol 22: 130–7.

7. Knight C, Lindberg GE, Voit GA (2012) Multiscale reactive molecular dynamics. J Chem Phys 137: 22A525.

8. Caze D (1994) Normal mode analysis of protein dynamics. Curr Opin Struc Biol 4: 285–290.

9. Skjæraaen L, Hollop SM, Reuter N (2009) Normal mode analysis for proteins. J Mol Struc-theochem 898: 42–48.

10. Venkatraman V, Ritchie DW (2012) Flexible protein docking refinement using pose-dependent normal mode analysis. Proteins 80: 2262–74.

11. Krüger DM, Ahmed A, Göhlke H (2012) NMSim web server: integrated approach for normal mode-based geometric simulations of biologically relevant conformational transitions in proteins. Nucleic Acids Res 40: W510–6.

12. Haliloglu T, Bahar I, Erman B (1997) Gaussian dynamics of folded proteins. Phys Rev Lett 79: 3090–3093.

13. Schroder GF, Brumberg AT, Levit M (2007) Combining eficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. Structure 15: 1630–1641.

14. Zimmermann MT, Kloczkowski A, Jermyn RJ (2011) MAVE-Ns: motion analysis and visualization of elastic networks and structural ensembles. BMC Bioinformatics 12: 264.

15. Latombe JC (1990) Robot Motion Planning. Boston, MA: Kluwer Academic Publishers.

16. Choset H, Lynch KM, Hutchinson S, Kantor G, Burgard W, et al. (2005) Principles of Robot Motion: Theory, Algorithms, and Implementations. MIT Press.

17. Cortés J, Siméon T, Ruiz de Angulo V, Guieysse D, Remaud-Siméon M, et al. (2005) A path planning approach for computing large-amplitude motions of exible molecules. Bioinformatics 21 Suppl 1: i116–25.

18. Tang X, Thomas S, Tapia L, Giedroc DP, Amato NM (2008) Simulating RNA folding kinetics on approximated energy landscapes. J Mol Biol 381: 1055–1067.

19. Cortés J, Siméon T, Remaud-Siméon M, Tran V (2004) Geometric algorithms for the conformational analysis of long protein loops. J Comput Chem 25: 956–967.

20. Camatescu AA, Dunbrack RL (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci 12: 963–972.

21. Yao P, Dhanik A, Marz N, Propper R, Kou C, et al. (2008) Efficient algorithms for the conformational analysis of long protein loops. J Comput Chem 29: 1067.

22. Thomas S, Song G, Amato NM (2005) Protein folding by motion planning. Phys Biol 2: S48–55.

23. Chiang SH, Apaydin MS, Bratlag DL, Hsu D, Latombe JC (2007) Using stochastic roadmap simulation to predict experimental quantities in protein folding kinetics: Folding rates and pvalues. J Comput Biol 14: 578–599.

24. Kirillova S, Cortés J, Stefaniu A, Siméon T (2008) An NMA-guided path planning approach for computing large-amplitude conformational changes in proteins. Proteins 70: 131–43.

25. Shehu A, Kavraki LE (2012) Modeling structures and motions of loops in protein molecules. Entropy 14: 252–290.

26. Thomas S, Tang X, Tapia L, Amato NM (2007) Simulating protein motions with rigidity analysis. J Comput Biol 14: 839–855.

27. Thomas S, Song G, Amato NM (2005) Protein folding by motion planning. Phys Biol 2: S48–55.

28. Bjo¨rkman AJ, Mowbray SL (1998) Multiple open forms of ribose-binding protein kinematics and dynamics: Beyond simulation. Annual Review of Analytical Chemistry 5: 273–291.

29. Singh AP, Latombe JC, Bratlag DL (1999) A motion planning approach to eblig ligand binding. Proc Int Conf Intelligent Syst for Molecular Biology (ISMB): 252–261.

30. Kavraki LE, Svestka P, Latombe JC, Overmars MH (1996) Probabilistic roadmaps for path planning in high-dimensional conuration spaces. IEEE Trans on Robotics and Automation 12: 566–580.

31. Singh AP, Kavraki LE, Guerri C, Hsu D, Latombe JC, et al. (2003) Stochastic roadmap simulation: An eficient representation and algorithm for analyzing molecular motion. J Comput Biol 10: 257–281.

32. Vaidya K, Datta N, Kavraki LE (2008) Roadmap Methods for Protein Folding. Methods in Molecular Biology 413: 219–239.

33. Gipson B, Hsu D, Kavraki LE, Latombe JC (2012) Computational models of protein kinematics and dynamics: Beyond simulation. Annual Review of Analytical Chemistry 5: 273–291.

34. Al-Belawi I, Siméon T, Cortés J (2012) Motion planning algorithms for molecular simulations: A survey. Computer Science Review 6: 125–143.

35. LaValle SM, Kuffner JJ (2001) Randomized kinodynamic planning. Intl J of Robotics Research 20: 378–400.

36. Ugur O, MacKrell AD (2008) Comparison of protein force fields for molecular dynamics simulations. Methods in molecular biology (Clifton, NJ) 443: 63–88.

37. Shabaninejad A, Kavraki LE (2012) A sampling-based tree planner for complex interactions. IEEE Trans on Robotics 28: 17–35.

38. Chen Y, Song G, Amato NM (2004) Hybrid of system dynamics and agent technologies for simulation of large-scale molecular systems. PLoS Comp Biol 7: e1002034.

39. Singh AP, Latombe JC, Bratlag DL (1999) A motion planning approach to eblig ligand binding. Proc Int Conf Intelligent Syst for Molecular Biology (ISMB): 252–261.

40. Kavraki LE, Svestka P, Latombe JC, Overmars MH (1996) Probabilistic roadmaps for path planning in high-dimensional configuration spaces. IEEE Trans on Robotics and Automation 12: 566–580.

41. Arvind MS, Bratlag DL, Guerri C, Hsu D, Latombe JC, et al. (2003) Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion. J Comput Biol 10: 257–281.

42. Moll S, Schwartz D, Kavraki L (2008) Roadmap Methods for Protein Folding. Methods in Molecular Biology 413: 219–239.

43. Gipson B, Hsu D, Kavraki LE, Latombe JC (2012) Computational models of protein kinematics and dynamics: Beyond simulation. Annual Review of Analytical Chemistry 5: 273–291.

44. Al-Belawi I, Simeon T, Cortes J (2012) Motion planning algorithms for molecular simulations: A survey. Computer Science Review 6: 125–143.

45. LaValle SM, Kuffner JJ (2001) Randomized kinodynamic planning. Int J of Robotics Research 20: 378–400.

46. Ugur O, MacKrell AD (2008) Comparison of protein force fields for molecular dynamics simulations. Methods in molecular biology (Clifton, NJ) 443: 63–88.

47. Vaidya K, Datta N, Kavraki LE (2008) Roadmap Methods for Protein Folding. Methods in Molecular Biology 413: 219–239.

48. Gipson B, Hsu D, Kavraki LE, Latombe JC (2012) Computational models of protein kinematics and dynamics: Beyond simulation. Annual Review of Analytical Chemistry 5: 273–291.

49. Al-Belawi I, Simeon T, Cortes J (2012) Motion planning algorithms for molecular simulations: A survey. Computer Science Review 6: 125–143.

50. Haspel N, Moll M, Baker ML, Chiu W, Kavraki LE (2010) Tracing conformational changes in proteins. BMC Structural Biology 10: 81.

Author Contributions

Conceived and designed the experiments: BG MM LEK. Performed the experiments: BG. Analyzed the data: BG MM LEK. Contributed reagents/materials/analysis tools: BG. Wrote the paper: BG MM LEK. Designed the software used in analysis: BG.
56. Saul FA, Vallée-Lé Normand B, Lema F, Bentley GA (1998) Crystal structure of a dominant B-cell epitope from the preS2 region of hepatitis B virus in the form of an inserted peptide segment in maltodextrin-binding protein. J Mol Biol 280: 185–92.

57. Sharff AJ, Rodseth LE, Quiocio FA (1993) Refined 1.8 Å structure reveals the mode of binding of P-cyclodextrin to the maltodextrin binding protein. Biochemistry 32: 10553–9.

58. Evdokimov AG, Anderson DE, Routzahn KM, Waugh DS (2001) Structural basis for oligosaccharide recognition by Pyrococcus furiosus maltodextrin-binding protein. J Mol Biol 305: 891–904.

59. Dier J, Diederichs K, Grellier G, Horlacher R, Boos W, et al. (2001) The crystal structure of a liganded trehalose/maltose-binding protein from the hyperthermophilic archaeon Thermococcus litoralis at 1.85 Å. J Mol Biol 305: 905–15.

60. Duan X, Quiocio FA (2002) Structural evidence for a dominant role of nonpolar interactions in the binding of a transport/chemoreceptor to its highly polar ligands. Biochemistry 41: 706–12.

61. Muller GA, Choy WY, Yang D, Forman-Kay JD, Venter RS, et al. (2000) Global folds of proteins with low densities of NOE's using residual dipolar couplings: application to the 370-residue maltodextrin-binding protein. J Mol Biol 300: 197–212.

62. Duan X, Hall JA, Nikaido H, Quiocio FA (2001) Crystal structures of the maltodextrin/maltose-binding protein complexed with reduced oligosaccharides: exibility of tertiary structure and ligand binding. J Mol Biol 306: 1115–26.

63. Liu Y, Manna A, Li R, Marrin WE, Murphy RC, et al. (2001) Crystal structure of the Sac7 protein from Staphylococcus aureus. P Natl Acad Sci USA 98: 6177–82.

64. Saul FA, Vallée-Lé Normand B, Lema F, Bentley GA (1997) Crystal structure of a recombinant form of the maltodextrin-binding protein carrying an inserted sequence of a B-cell epitope from the preS2 region of hepatitis B virus. Protein 27: 1–8.

65. Srinivasan U, Iyer GH, Przybycień TA, Samsonoff WA, Bell JA (2002) Crystine: fibrous biomolecular material from protein crystals cross-linked in a specific geometry. Method Enzymol 13: 895–962.

66. Saul FA, Mourez M, Vallée-Lé Normand B, Sassoon N, Bentley GA, et al. (2003) Crystal structure of a defective folding protein. Protein Sci 12: 577–85.

67. Rubin SM, Lee SY, Ruiz EJ, Pines A, Wenmeter DE (2002) Detection and characterization of xenon-binding sites in proteins by 129Xe NMR spectroscopy. J Mol Biol 322: 425–40.

68. Sharff AJ, Rodseth LE, Szeleckman S, Hofmung M, Quiocio FA (1995) Refined structures of two insertion/deletion mutants probe function of the maltodextrin binding protein. J Mol Biol 246: 8–13.

69. Koler B, Center RJ, Kemp EE, Poulouchours (1999) Crystal structure of human T-cell leukemia virus type 1 gp21 ectodomain crystalized as a maltose-binding protein chimera reveals structural evolution of retroviral transmembrane proteins. P Natl Acad Sci USA 96: 4319–24.

70. Ke A, Wolberger C (2003) Insights into binding cooperativity of MATa1/MATA21 from the crystal structure of a MATa1 homodomain-maltose binding protein chimera. Protein Science: a publication of the Protein Society: 12: 306–12.

71. Shilton BH, Shuman HA, Mosbray SL (1996) Crystal structures and solution conformations of a dominant-negative mutant of Echerichia coli maltose-binding protein. J Mol Biol 264: 364–76.

72. Chao JA, Prasad GS, Stout CD, Williamson JR (2003) Inherent protein structural exibility in the RNA-binding interface of L30e. J Mol Biol 326: 999–1004.

73. Sharff AJ, Rodseth LE, Spurinlo JC, Quiocio FA (1992) Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the maltodextrin binding protein involved in active transport and chemotaxis. Biochemistry 31: 10657–63.

74. Telman PG, Shalton BH (2003) Insights into the conformational equilibria of maltose-binding protein by analysis of high affinity mutants. J Biol Chem 278: 34555–67.

75. Song JJ, Liu J, Tolli NH, Schweidtman J, Smith SK, et al. (2003) The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNA effector complexes. Nat Struct Biol 10: 1026–32.

76. Chao JA, Williamson JR (2004) Joint X-ray and NMR reneforment of the yeast L30e-mRNA complex. Structure 12: 1165–76.