PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Generating stereochemically acceptable protein pathways

Daniel W. Farrell, Kirill Speranskiy, and M. F. Thorpe*

Department of Physics and Center for Biological Physics, Arizona State University, Tempe, Arizona

ABSTRACT

We describe a new method for rapidly generating stereochemically acceptable pathways in proteins. The method, called geometric targeting, is publicly available at the webserver http://pathways.asu.edu, and includes tools for visualization of the pathway and creating movie files for use in presentations. The user submits an initial structure and a target structure, and a pathway between the two input states is generated automatically. Besides visualization, the structural quality of the pathways makes them useful as input pathways into pathway refinement techniques and further computations. The approach in geometric targeting is to gradually change the system's RMSD relative to the target structure while enforcing a set of geometric constraints. The generated pathways are not minimum free energy pathways, but they are geometrically plausible pathways that maintain good covalent bond distances and angles, keep backbone dihedral angles in allowed Ramachandran regions, avoid eclipsed side-chain torsion angles, avoid nonbonded overlap, and maintain a set of hydrogen bonds and hydrophobic contacts. Resulting pathways for over 20 proteins featuring a wide variety of conformational changes are reported here, including the very large GroEL complex.

Proteins 2010; 78:2908–2921.

Key words: conformational change; conformational sampling; geometric simulation; target; geometric targeting; transition; movies.

INTRODUCTION

The ability to determine pathways between different conformational states in proteins is key to understanding how structure influences function. Computational techniques of varying levels of sophistication have been introduced to find pathways in proteins, many of which are computationally intensive. In this work, we present a rapid and computationally inexpensive method to produce pathways between two states of a protein called geometric targeting, publicly available on a webserver at http://pathways.asu.edu. The webserver provides a simple interface to the targeting method and includes features for visualization of the pathway and generating movie files.

The approach can be summarized as a gradual changing of the system's RMSD (root-mean-square distance) relative to the target structure while enforcing a set of geometric constraints. Underlying geometric targeting is the philosophy that the essence of conformational changes in proteins can be modeled purely from geometric considerations. Geometric targeting generates complicated, highly non-linear, all-atom pathways, and is broadly applicable to many classes of conformational changes and works even for very large systems. The generated pathways are not optimal pathways, but they are stereochemically acceptable pathways in the sense that they prevent atom overlap, preserve bond distances and angles, preserve hydrogen bonds and hydrophobic contacts, and keep backbone and side chain dihedral angles away from unfavorable configurations.

Two primary uses of the method and webserver are the visualization of conformational changes and the generation of input pathways for further computation or refinement. For visualization, there are obvious advantages to looking at a movie compared to looking at two superimposed static protein structures. It can be difficult to visually compare static structures, identify the regions that differ, and mentally figure out what motions could be involved in the transition. A movie showing a pathway between two states is a more natural way to learn what has changed and how the change takes place. The other primary use of geometric targeting is to create input pathways for use in more sophisticated techniques. Several techniques that explore the energy landscape to search for optimal pathways such as transition path sampling,[1] string method,[2–4] and nudged elastic band,[5,6] require an initial pathway to get started that is typically produced by simple interpolation between end states. In systems where interpolation produces a poor initial guess, pathways produced from geometric targeting may make better candidate input pathways. In an article that is currently in preparation[7] with collaborators Tatyana Mamonova and Maria Kurnikova, we

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: National Science Foundation; Grant number: DMS-0714953
*Correspondence to: Michael F. Thorpe, Physics Department, Arizona State University, Tempe, AZ 85287-1504. E-mail: mft@asu.edu.
Received 10 December 2009; Revised 20 May 2010; Accepted 15 June 2010
Published online 29 June 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.22810

will show that a pathway generated from geometric targeting can be used as input into an umbrella sampling[8] free energy calculation.

Besides the geometric targeting method introduced in this article, various other approaches exist for finding pathways in proteins. Sophisticated techniques that perform rigorous searching of energy landscapes to determine optimal pathways include the aforementioned transition path sampling,[1] string method,[2–4] and nudged elastic band,[5,6] as well as the probabilistic roadmap method[9] and the finite temperature non-local exploration method of Branduardi *et al.*,[10] and others.[11,12] See also these review articles.[13,14] Steered molecular dynamics,[15] targeted molecular dynamics,[16,17] and restricted perturbation-targeted molecular dynamics[18,19] are intensive approaches that use biased dynamics to create pathways, recently extended to determine minimum free energy pathways.[19] In contrast to these approaches, geometric targeting lacks a molecular mechanical force field and does not sample according to a Boltzmann distribution. Although less rigorous than these approaches, geometric targeting can be thought of as a "back of the envelope" pathway calculation that considers only geometry and rapidly produces a plausible result at atomic-level detail. For very large systems, the sophisticated techniques may be intractable, making geometric targeting an alternative that can produce a stereochemically correct all-atom pathway.

Other methods for creating pathways include elastic network models, which invoke a small-amplitude approximation on a system of interconnected springs to produce a simplified, smooth harmonic energy landscape. Examples include Elastic Network Interpolation models[20–23] and the Plastic Network Model,[24] which all use coarse-graining at the level of C_α atoms and in some cases rigid clusters of C_α atoms. Iterative cluster-normal mode analysis[25] (ic-NMA) includes all atoms, grouped into rigid clusters no smaller than a residue, with springs that attach pairs of atoms in distinct clusters. Although successful at generating approximate transition pathways, some of the limitations of these models are the lack of atomic-level detail (excepting ic-NMA), the neglect of atomic-overlaps in the elastic network energy function, and an overly flexible protein backbone because of the neglect of covalent bond geometry. Compared to elastic network models, the advantages of geometric targeting are the all-atom geometric detail of the snapshots produced, and the dynamic prevention of atomic overlaps which allows more complicated motions in which atoms bump and move around each other. However, in some cases, elastic network based models do better at capturing the relative timing of events along the pathway (see Discussion).

Linear interpolation with energy minimization is a rapid technique for making pathways used by the Yale Morph Server,[26–28] however, these pathways are often not physically plausible because atoms and chains can pass through each other. In the Results section, we will show examples for which the Yale Morph Server's technique results in unphysical pathways, but for which geometric targeting produces plausible, non-linear, complex motions without chains passing through each other.

We wish to point out the relationship of the present geometric targeting method to prior work in constraint-based exploration of protein structures. Wells *et al.* described a method for exploring freedom in protein structures based on geometric constraints, called FRODA.[29] One of the applications of FRODA described by the authors is targeting.[27,29] FRODA was the original idea that sparked the ideas for the geometric targeting method presented here, and the two methods share a similar philosophy but use different underlying mathematical techniques. Some differences will be pinpointed in the Discussion section. Another technique, called tCONCOORD, also uses geometric constraints for sampling of protein structures.[30,31] Targeting and pathway generation are not possible uses of tCONCOORD, because each generated structure is completely uncorrelated with the previously generated structure.

In this article, we present results for over 20 proteins of various sizes and exhibiting a wide variety of conformational changes, including hinge motions, shear motions, loop rearrangements, side chain rearrangements, domain swapping, and other complex changes that are not easily classified.

WEBSERVER USAGE

The webserver is located at http://pathways.asu.edu. The webserver prompts the user to submit the initial and target protein structures in PDB format. The two proteins need not have identical atoms. Mutational differences and incomplete target structures are acceptable. The files also do not need to contain hydrogen atoms, as these will be added automatically. Missing residues or atoms in the initial structure will not be modeled, however. If multiple chains exist, the webserver will prompt the user to decide how chains from one structure map to the chains of the other structure. After submitting the two structures, some automatic preprocessing takes place, and then the targeting begins. The targeting often completes in a few minutes. Depending on the size of the protein and the amount of structural difference between the two states, some runs can require an hour or longer. During the targeting, the web page is continuously updated, showing the current RMSD-to-target and current number of generated snapshots.

When the pathway is complete, the user views an animation of the pathway in an interactive Jmol window.[32] The user can adjust the zoom level, rotate, and translate the protein while watching the pathway. A movie file can optionally be generated and downloaded. The atomic trajectory can also be saved to disk in PDB format for further analysis.

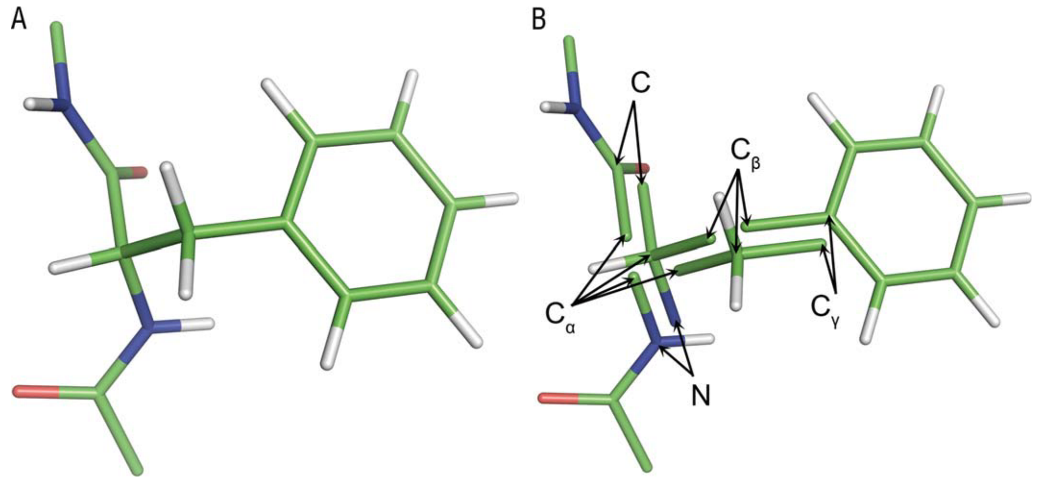

Figure 1

Decomposition of phenylalanine into rigid units. **A**: Stick model of phenylalanine with main-chain atoms of neighboring residues. **B**: Atoms are embedded within rigid units, which lock in place covalent bond distances and angles. Note that a single atom may have multiple copies, each belonging to a different rigid unit, as pointed out with arrows. Graphics were produced with PyMol.[45] [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Various "Advanced Options" are also available when the user submits structures for targeting. As described later in the text, the user can opt to include random motion, enable backtracking, and choose whether to make certain hydrogen bond and hydrophobic contact constraints fixed or breakable. In addition, the user may modify the hydrogen bond cutoff energy. Geometric constraints will be placed in the protein for hydrogen bonds that are stronger than the cutoff energy. Therefore, lowering this cutoff energy will result in fewer hydrogen bonds and hence a more flexible protein.

All targeting results are stored in the "File Cabinet," allowing a user to return to a previous targeting run, visualize the pathway, and download movies or trajectories as needed. Targeting runs continue even if a user navigates away from the page or disconnects from the internet, and the file cabinet can be used to access these runs.

METHODS

Preprocessing

When the user submits the initial and target PDB structures to the webserver, the webserver automatically carries out some preparatory steps behind the scenes before running the targeting. First, waters, hydrogens and ligands are removed. The program Reduce from the Richardson Lab is run on the initial and target structures to add hydrogens and optimally position them.[33] Next, to determine which atoms from the initial structure correspond to which atoms in the target structure, a sequence alignment between the initial and target structures is performed by running ClustalW.[34] This means that mutational differences or incomplete target structures are acceptable. Some atoms in the initial structure may have no matching counterpart in the target structure, and vice versa.

Geometric model

The targeting method begins by constructing a geometric model of the protein, using the initial structure as reference. The geometric model is an all-atom model, including hydrogens. To capture the geometric characteristics of covalent bonds, we subdivide each amino acid of the protein into rigid subcomponents based on the assumption of rigid bond distances, rigid 3-body angles. Dihedral angles for single covalent bonds are not constrained, but dihedral angles for higher bond orders (including peptide bonds) are treated as rigid. The grouping of atoms into rigid units is performed by the software package FIRST,[35] run in a modified fashion so as to only include covalent bonding geometry in making rigid unit assignments. Figure 1 shows an example of how the amino acid phenylalanine is subdivided into rigid units. In phenylalanine, the C_α plus its four covalent neighbors constitute one rigid unit, the C_β plus its four neighbors are a second rigid unit, the phenyl ring's six carbon atoms plus their covalent neighbors are a third, and the peptide planes on both sides of the C_α are rigid. Within the 20 standard amino acids, the largest rigid unit is the planar indole group found in tryptophan. As there are no free atoms in the amino acids, the smallest rigid units have three atoms, such as the C—OH in the side chain of tyrosine or serine.

The rigid units are the mobile units of the system, each having six degrees of freedom (three translational, three rotational). The rigid units contain "embedded atoms," whose positions depend entirely on the degrees of freedom of their corresponding rigid unit. Observe that several "embedded atoms" may correspond to the same (physical) atom, for example, the (split) C_α and C_β atoms in Figure 1(B). We define the position of the atom as being located at the mean of its embedded atoms. Therefore, the positions of the "atoms" depend on the positions of the "embedded atoms," which in turn depend on the rigid unit degrees of freedom.

Having divided the amino acids into rigid units, we then establish various geometric constraints that define allowed and disallowed configurations of the rigid units. The constraints are meant to maintain various aspects of stereochemical quality.

Covalent bond geometry between adjacent rigid units

To make the multiple copies of an atom shared by adjacent rigid units coincide, we constrain the distance between the multiple copies of a shared atom to be zero.

Non-bonded overlap

Between all non-bonded atoms (fourth neighbors and higher), we place a minimum distance constraint that keeps their separation greater than some cutoff value, the cutoffs depending on the types of atoms involved. The cutoff distances have been calibrated from MD simulations. This will be described in more detail in an article currently in preparation with collaborators Tatyana Mamonova and Maria Kurnikova.[7] The cutoff distances are listed in Supporting Information Tables S1 and S2.

Backbone dihedral angles

To keep backbone dihedral angles out of the disfavored and disallowed regions of the Ramachandran plot,[36] we establish additional minimum distance constraints between certain pairs of main-chain atoms. Note that we do not explicitly constrain the dihedral angles, but instead use distance constraints between atoms to block off the disallowed and disfavored backbone dihedral angles. Here, we make use of the work of Ho *et al.*,[37] in which they show that the disallowed and disfavored regions of the Ramachandran map can be understood simply from non-overlap constraints between certain pairs of main-chain atoms. We apply minimum distance constraints to the same main-chain pairs identified in their article (see Table I). These constraints apply to all 20 standard amino acids, including proline and glycine.

Table I
Distance Constraints for Main-Chain Pairs

Main-chain pair	Minimum distance cutoff (Å)
$C_\beta...O$	2.80
$C_\beta...N_{i+1}$	3.00
$O_{i-1}...O$	3.10
$O_{i-1}...N_{i+1}$	3.00
$O_{i-1}...\ C_\beta$	3.05
$H...\ H_{i+1}$	1.85

These minimum distance constraints are imposed on all of the standard 20 amino acids to keep main-chain dihedral angles out of the disallowed and disfavored regions of the Ramachandran plot. Ho *et al.*[37] found that the disallowed and disfavored regions arise because of steric clashes between these pairs of atoms. The distance cutoffs used here are slightly modified from the distances published by Ho *et al.*[37]

Side-chain torsion angles

To keep side-chain torsion angles away from unfavorable eclipsed configurations, we define constraints between 1-4 bonded atom pairs only in cases where atoms 2 and 3 are single-bonded and are each tetrahedrally coordinated. To ensure that these dihedral angles remain staggered, it is not necessary to directly constrain the dihedral angles. Instead, we place a minimum distance constraint between each 1–4 pair, setting the minimum distance such that the dihedral angle between them comes no closer than 55°. These constraints partially account for rotamer configurations, but not completely, because bonds between non-tetrahedrally coordinated atoms are left freely rotatable.

Hydrogen bonds and hydrophobic contacts

Hydrogen bonds and hydrophobic contacts are preserved by placing maximum-distance constraints between pairs of interacting atoms. For hydrogen bonds, we only place a constraint for those that have an energy score better than some cutoff value, typically -1.0 kcal/mol, as measured by an energy function.[35,38] The maximum distance constraint is placed between the hydrogen and the acceptor atoms and is set to the distance that is in the initial structure, but not less than 2.0 Å. For hydrophobic contacts, we place maximum distance constraints between pairs of hydrophobic side-chain carbon or sulfur atoms that are closer than 3.9 Å in the initial structure. We consider only the hydrophobic residues Leu, Ile, Val, Phe, Trp, Met, Ala, Tyr. The maximum distance constraint for hydrophobic pairs is set to the distance in the initial structure plus an extra 0.5 Å.

Geometry in input structures takes precedence

In establishing the minimum distance constraints described above for non-overlap, Ramachandran, and side chain torsion, we make sure that the constraints do not conflict with the geometry in the input structures. If a pair of atoms in either the initial or target structures is found closer than would be normally allowed by a minimum-distance constraint, the minimum distance cutoff for the pair is altered and set to the actual distance in the input structure.

Targeting procedure

The strategy we use to bring the system from the initial to the target state is to impose a constraint on the

RMSD-to-target, gradually decreasing this constraint towards 0 Å RMSD. While bringing the RMSD to zero, we also enforce the structural constraints to keep the snapshots stereochemically acceptable. By enforcing the structural constraints, atoms will be forced to move along curved trajectories, as they must maintain distance and angle relationships while moving towards the target.

We will first describe the targeting procedure in its most basic form, and then describe some optional modifications to the procedure. In the most basic form, the targeting procedure involves no random motion, producing a smooth pathway. Furthermore, only hydrogen bond and hydrophobic constraints that are common to both the initial and target structures are included. Otherwise, incompatible hydrogen bonds or hydrophobic contact constraints could prevent reaching the target state.

The targeting begins with the atoms in their initial positions from the submitted initial structure. The RMSD of the initial structure relative to the target structure, calculated over all targeted atoms, is some number C_0. An RMSD step size δ is chosen, typically 0.1 Å or less. Each targeting step consists of the following actions:

1. Advance the RMSD constraint: Set the RMSD constraint to RMSD $< C_i$, where $C_i = C_{i-1} - \delta$, where subscript i denotes the step number.
2. Enforce constraints: The RMSD constraint and structural constraints are enforced simultaneously, causing the rigid units of the system to move and rotate, often taking atoms in curved paths. The process is described in the section Enforcement of Constraints.
3. Global fitting: Finish the step by globally rotating and translating the entire system to optimize the RMSD to the target.
4. If structure is acceptable, move on to next step: The criteria for judging whether the structure is acceptable are that the non-overlap constraints not be violated by more than 0.2 Å, and that the shared atoms between adjoining rigid units not be more than 0.2 Å apart. In the most basic form of targeting, the targeting steps are terminated here if the structure is not acceptable. This can happen when the targeting has run up against a particularly difficult obstacle that it cannot find a way to get around without violating structural constraints.

Random motion

The basic targeting procedure described above contains no random motion. The resulting pathway is deterministic, and atoms appear to move smoothly. To produce a random pathway, random motion can be optionally added to each targeting step as follows. At the beginning of each step, each rigid unit is randomly displaced and rotated, without regard for any constraints. The rest of the targeting step continues as usual. The constraint violations created by the random perturbation are restored during the "Enforce Constraints" portion of each step. The size of the perturbation is rather large, on the scale of 1 Å for translational displacement and 120° for rotational motion, so that rigid units can hop over disallowed dihedral angle regions. This can cause some rigid units to get stuck during the enforcement of constraints, in which case the problem rigid units are restored to their original positions and orientations.

Options for handling of hydrogen bond and hydrophobic contact constraints

"Common" hydrogen bond and hydrophobic constraints are those that are found in the initial and target structures. In the basic targeting procedure, the common constraints are kept fixed throughout the targeting under the assumption that the interactions are present during the entire pathway. As an option, the common constraints can be made breakable, or can be not included, instead of kept fixed. When a breakable constraint becomes stretched beyond a certain amount, it "breaks" and is removed. This can be helpful if some hydrogen bond or hydrophobic contact that is found in both structures needs to transiently break during the pathway. "Non-common" constraints are those that are in the initial structure but not in the final structure. The basic setting is to simply not include the non-common constraints since they are incompatible with the final structure. Optionally, the user can choose to include the non-common constraints as breakable constraints. Having the non-common constraints included may improve the quality of the pathways, since they preserve favorable interactions until the moment they break.

Recovery methods

In the basic targeting procedure, if the shared-atom constraints and non-overlap constraints cannot be satisfied to within tolerance, the structure is deemed unacceptable and the targeting is terminated. Usually this does not happen until the very end, when the RMSD to target is quite low (<0.5 Å), and all the atoms are very close to their targets. It can sometimes happen earlier, when a particularly difficult obstacle in the pathway can cause the targeting to fail to produce an acceptable structure. A few recovery methods are available to try to help the protein move around the obstacle. The first is called "random retry," which is to retry the last step using a random perturbation of the rigid units as described above in hopes that the random motion will help move past the obstacle. Typically up to five consecutive random retries are attempted.

Another available recovery method is "Backtracking." In backtracking, the targeting steps switch into reverse, taking the RMSD away from the target instead of closer to the target. The sign of the RMSD step δ is reversed so that the RMSD constraint C_i increases instead of decreases at each step. The inequality in the RMSD constraint is also switched to a greater-than sign, RMSD > C_i, to carry the system away from the target. The idea is to go back in RMSD, find a new starting point at the higher RMSD level, then return to forward steps, in hopes that this enables the system to get around an obstacle. The method used to find a new starting point at the elevated RMSD before returning to forward steps is called "momentum steps," described later. The first time that a targeting step fails to produce an acceptable structure, the system is backtracked by 1 Å, a new starting point is found, and then the procedure returns to regular forward steps. If the targeting again gets stuck, the backtracking method tries going back by 2 Å, then 4 Å, then 8 Å, etc., doubling the amount of backward motion each time. The backtracking can even take the protein back in RMSD farther than the initial state. All non-common constraints are removed during backtracking so they do not hinder the system from going back in RMSD.

Momentum steps

During backtracking, when the RMSD has been taken back to some higher value, we use "momentum steps" to find a new starting point before recommencing steps toward the target. Momentum steps are so named because the motion tends to persist in the same direction over many steps. Note that momentum is not actually conserved, since we are not integrating equations of motion, and there are no time steps or velocities. Here, the net translational and rotational change of each rigid unit is recorded for each step and used as a perturbation in the next step. Throughout the momentum steps, the upper-bound RMSD constraint is kept active, ensuring that the RMSD does not go back further. A momentum step involves the following actions.

1. Store current configuration: The six degrees of freedom of each rigid unit are stored in a $6M$-dimensional vector $\mathbf{q}_1$, where M is the number of rigid units.
2. Perturb rigid units by the last $\Delta\mathbf{q}$: The rigid units are translated and rotated by the $\Delta\mathbf{q}$ of the previous momentum step, or 0 if this is the first momentum step. The system is now at a new configuration $\mathbf{q}_2$.
3. Small Random Perturbation: Randomly perturb the rigid units (translationally and rotationally), but do so with a very small amplitude (Atoms move by only about 0.05 Å). The system is now at $\mathbf{q}_3$.
4. Enforce constraints: Both the RMSD constraint and the structural constraints are enforced, bringing the system to state $\mathbf{q}_4$.
5. Global Fit to Target Structure: Remove any global translations and rotations by globally fitting to the target structure, bringing the system to state $\mathbf{q}_5$.
6. Calculate net change: Determine the net change of the degrees of freedom in this momentum step, $\Delta\mathbf{q} = \mathbf{q}_5 - \mathbf{q}_1$, for use in the next step. Then move on to the next step.

Because of the small random component being added in each step, components of the motion along soft directions gradually grow in size. Because constraints are enforced in each step, components of the motion that encounter constraints cannot persist more than one step and cannot grow. After several steps, large-amplitude motions develop, which enables fast movement to a new position.

Note that the RMSD constraint, which is kept active during the momentum steps, is only an upper-bound, so the RMSD of the system is free to decrease. Entropy, however, usually keeps the RMSD as high as the constraint allows, since there tend to exist more states at high RMSD than low RMSD.

Enforcement of constraints

To explain how constraints are enforced, we must clarify the mathematical relationship between the rigid unit degrees of freedom and the positions of the atoms. Recall that the position of an atom is located at the mean position of its corresponding "embedded atoms," which in turn depend on the rigid unit degrees of freedom of their respective rigid units. Let $\mathbf{r}$ be a $3N$-dimensional vector containing the (mean) positions of the N atoms, $\mathbf{u}$ be the $3N'$-dimensional vector containing the positions of the N' embedded atoms, and $\mathbf{q}$ be the $6M$-dimensional vector containing the rigid unit degrees of freedom of the M rigid units. For the translational degrees of freedom of a rigid unit, we use the Cartesian coordinates of the centroid of the rigid unit. For the rotational degrees of freedom of a rigid unit, we use three independent rotor variables from geometric algebra, B_x, B_y, B_z, as described in Wells.[39] These three rotor variables can be interpreted as a three-dimensional vector $\mathbf{B}$ that points along the axis of rotation and has a magnitude $|\mathbf{B}| = 2\sin\frac{\phi}{2}$, where ϕ is the angle of rotation. See Wells,[39] for how these variables can be used to describe rigid body rotations.

To help the rigid units find their way to an acceptable state, we define an "energy function" that measures the total amount of constraint violation in the system. We then perform conjugate gradient minimization to find the configuration of rigid units that minimizes the constraint energy.[40] In the constraint energy function, each constraint is represented by an energy term that is zero if the constraint is met and increases quadratically with the amount of constraint violation. It is important to recognize that the snapshots produced by geometric targeting lie within the flat portion of the energy landscape at energy zero (or near zero when some constraint violations cannot be fully resolved). The non-zero region of the energy landscape only serves to guide the system back to the flat, zero energy region.

$$V = V_{\text{shared atoms}} + V_{\text{min dist}} + V_{\text{max dist}} + V_{\text{RMSD}}$$

$$V_{\text{shared atoms}} = \sum_{i<j}{}' \frac{1}{2} k u_{ij}^2$$

$$V_{\text{min dist}} = \sum_{i<j}{}' \begin{cases} \frac{1}{2} k\left(r_{ij} - d_{ij}^{\min}\right)^2, & r_{ij} < d_{ij}^{\min} \\ 0, & r_{ij} \geq d_{ij}^{\min} \end{cases}$$

$$V_{\text{max dist}} = \sum_{i<j}{}' \begin{cases} \frac{1}{2} k\left(r_{ij} - d_{ij}^{\max}\right)^2, & r_{ij} < d_{ij}^{\max} \\ 0, & r_{ij} \geq d_{ij}^{\max} \end{cases}$$

$$V_{\text{RMSD}} = \begin{cases} \frac{1}{2} kN(\text{RMSD} - C)^2, \text{RMSD} > C \\ 0, \text{RMSD} \leq C \end{cases}$$

In the above equations, u_{ij} is the distance between "embedded atoms" i and j, $r_{i,j}$ is the distance between (mean) atom positions i and j, $d_{ij}^{\min}$ and $d_{ij}^{\max}$ are the minimum and maximum distance constraints for atoms i and j. The prime symbols in the summations denote that sums are only over pairs i, j for which a constraint exists.

Conjugate gradient minimization of V takes the system to a local minimum, using the gradient of the energy function to guide the search for the minimum.[40,41] The gradient must be taken with respect to the system's degrees of freedom **q** (the rigid unit degrees of freedom). Since the various terms of V are explicitly expressed as functions of the positions of the atoms **r** or the positions of the embedded atom copies **u**, rather than as functions of **q**, chain rules must be used to obtain the derivatives $\partial V/\partial q_i$ for each degree of freedom q_i. To make the system better-conditioned for conjugate gradient, the unitless rotor degrees of freedom are each scaled by a characteristic length-scale so that they are comparable in magnitude with the translational degrees of freedom. In addition, diagonal elements of the second derivative matrix, $\partial^2 V/\partial q_i^2$, are calculated and used as a preconditioner.[40]

To determine when to stop the conjugate gradient minimization, we make an estimate of how close each degree of freedom is from the local minimum. In the approximation that each degree of freedom lies in an independent parabolic well $\frac{1}{2}k(q_i - q_{i0})^2$ for some unknown minimum-energy position q_{i0}, taking the ratio of $\partial V/\partial q_i$ to $\partial^2 V/\partial q_i^2$ gives $q_i - q_{i0}$, which is the estimate of the error. We stop the conjugate gradient minimization when this error estimate is below some tolerance value, typically 0.005 Å.

Ideally, conjugate gradient minimization would bring the energy to zero, meaning that all constraints are satisfied. In practice, local minima in the energy function can arise from mutually incompatible constraints (an RMSD constraint pulling a side chain through an eclipsed configuration, for example), preventing certain constraints from being fully satisfied.

RESULTS

We applied the geometric targeting method to over twenty proteins of various sizes and exhibiting a wide variety of conformational changes, including hinge motions, shear motions, loop rearrangements, side chain rearrangements, domain swapping, and other complex changes that are not easily classified. Some of these examples were selected from the Database of Macromolecular Movements.[28] Results are summarized in Table II. For each system, an initial targeting attempt was made using the following settings: no random motion, no backtracking, RMSD step size of 0.1, common hydrogen bonds and hydrophobic contacts treated as fixed constraints, non-common hydrogen bonds and hydrophobic contacts left unconstrained. Random retry steps were used as a recovery method. With these settings, the targeting was successful for most systems. Proteins that could not reach their targets under these targeting settings were re-run with backtracking enabled in order to get around significant obstacles in the pathway. One system was unsuccessful even with backtracking enabled, but was successful when the common set of hydrogen bonds and hydrophobic contacts were allowed to break, instead of keeping them fixed (Table II). All examples successfully reached their targets within very low all-atom RMSD (<0.5 Å). Runs typically completed within minutes, with the largest case GroEL requiring almost 2 h. Run times scaled roughly in proportion to the number of atoms and the RMSD difference between the two states. Rather than describing in detail the results for each protein, we highlight below a few examples that suffice to demonstrate the versatility and robustness of the method (movies for these examples can be found in the Supporting Information). Protein Data Bank[42] (PDB) IDs and chain information for all examples are listed in Supporting Information Table S3.

A few of the successful examples discussed below are known to yield unphysical pathways under the linear interpolation method of the Yale Morph Server,[26] namely diphtheria toxin and GroEL, with groups of atoms passing through each other as discussed in their article and available for viewing at their website.[28] A third unphysical

Table II
Webserver Pathway Results

Protein name	No. of sub-units	No. of atoms	Initial RMSD (Å)	Final RMSD (Å)	CPU Time (min)	No. of steps	CPU time per step per No. of atoms (ms)	Figure	Movie
Basic settings[a]									
Toy Model 1	1	129	7.6	**0.0**	**0.0**	76		2A	1
Collagenase	1	1770	8.1	**0.2**	**0.4**	85	0.16		
Calmodulin	1	2262	5.5	**0.0**	**0.2**	55	0.08		
Dihydrofolate Reductase	1	2489	1.9	**0.1**	**0.3**	26	0.31		
Pyrophosphokinase	1	2535	3.0	**0.1**	**0.3**	36	0.22		
Spindle Assembly Checkpoint Protein	1	3033	10.2	**0.2**	**1.9**	106	0.36	2B	2
CD2	2	3096	23.2	**0.2**	**3.4**	237	0.28	2D	3
Adenylate Kinase	1	3341	7.2	**0.1**	**0.6**	77	0.13		
Alcohol Dehydrogenase	1	5639	2.1	**0.2**	**1.2**	25	0.49		
Heparin Cofactor II	1	6931	6.1	**0.0**	**1.9**	61	0.27		
Diphtheria Toxin	1	7972	16.0	**0.2**	**4.7**	164	0.21	2C	4
5′-Nucleotidase	1	8120	10.1	**0.2**	**2.8**	105	0.20		
Citrate Synthase	2	13,182	3.3	**0.5**	**6.4**	34	0.85		
Pyruvate Phosphate Dikinase	1	13,420	11.7	**0.1**	**4.2**	118	0.16		
DNA Polymerase	1	14,563	6.6	**0.2**	**4.9**	70	0.29		
HIV-1 Reverse Transcriptase	2	15,299	4.1	**0.3**	**8.6**	45	0.75		
Phosphofructokinase	4	19,140	2.1	**0.2**	**8.2**	28	0.92		
Replication Factor C	6	29,966	14.0	**0.0**	**16.9**	145	0.23		
Rho Transcription Termination Factor	6	37,136	2.1	**0.2**	**15.9**	25	1.03		
GroEL	14	109,718	11.2	**0.2**	**115.0**	115	0.55	2E	5
Basic settings[a] + backtracking									
Toy Model 2	1	1611	10.8	**0.3**	**9.3**	—	—	3A	6
HIV Protease	2	3144	2.0	**0.2**	**1.6**	—	—	3B	7
Dengue 2 Virus Envelope Glycoprotein	1	6129	12.1	**0.2**	**5.7**	—	—	3C	8
Basic settings[a] + breakable hydrogen bond and hydrophobic contact constraints									
Immunoglobulin E SPE7	2	3467	2.7	**0.5**	**1.3**	29	0.77		

[a]Basic settings: no backtracking, no random motion (except during retry steps), five random retry steps enabled, 0.1 Å RMSD step size, hydrogen bonds and hydrophobic contacts common to both structures treated as fixed constraints, non-common hydrogen bonds and hydrophobic contacts not included as constraints.
Pathways were generated for the listed examples. As indicated, many completed successfully even without random motion or backtracking. Other pathways were only successful when backtracking was enabled or when common constraints were made breakable, as indicated. The number of atoms in the initial state includes hydrogens. Initial and final RMSD are computed with respect to the target structure, using all targeted atoms including hydrogens. The reported CPU times correspond to a single processor. The number of steps reported includes random retry steps. The number of steps and time per step is not reported for backtracking runs, because these runs include a mixture of forward steps, backward steps, and momentum steps, each of which have different characteristic times per step. The "Figure" column lists figure numbers for those pathways represented in the figures, and the "Movie" column lists the Supporting Information Movie number.

example not discussed in their article but available on their website[28] is spindle assembly checkpoint protein.

Figure 2(A) shows results for a toy model system designed to illustrate how the targeting procedure can produce highly non-linear pathways, without the use of backtracking or random motion. Toy model 1 is a polyalanine fragment of length 12 residues starting in an extended beta-like configuration that was targeted to an alpha-helix configuration. The initial structure had no hydrogen bonds or hydrophobic contacts, so the only constraints active in the system were the minimum distance constraints between atoms (preventing overlap, disfavored Ramachandran regions, and eclipsed side-chain configurations), and the shared atom constraints between connected rigid units. Random motion was not used, so the motion is driven solely by the gradually changing RMSD constraint which pulls the system closer and closer to the target. As atoms are pulled towards their targets, they follow curved trajectories due to the enforcement of structural constraints. The Ramachandran constraints and non-eclipsing side chain constraints pose particularly difficult obstacles, creating disallowed regions in dihedral angle space that must somehow be crossed in order to reach the target. Sometimes, the rigid units are observed to make sudden jumps as the RMSD constraint pulls them from one allowed region to another, moving over a disallowed region. The RMSD constraint energy term in the constraint energy function lifts the system over a barrier created by a minimum-distance energy term, and minimization carries the system downhill to the other side.

The conformational change in spindle assembly checkpoint protein is complicated [Fig. 2(B)], involving a strand of beta sheet (red) that passes under a loop and joins with an alpha helix, while a second strand of beta sheet (green) moves from the top to bottom side of the beta sheet. On the Yale Morph Server[26] which uses linear interpolation with energy minimization, the polypeptide chains can be seen to pass through each other in an unphysical manner. With geometric targeting, the chains are observed to bump into each other and move around each other to avoid atomic overlap in reaching the target.

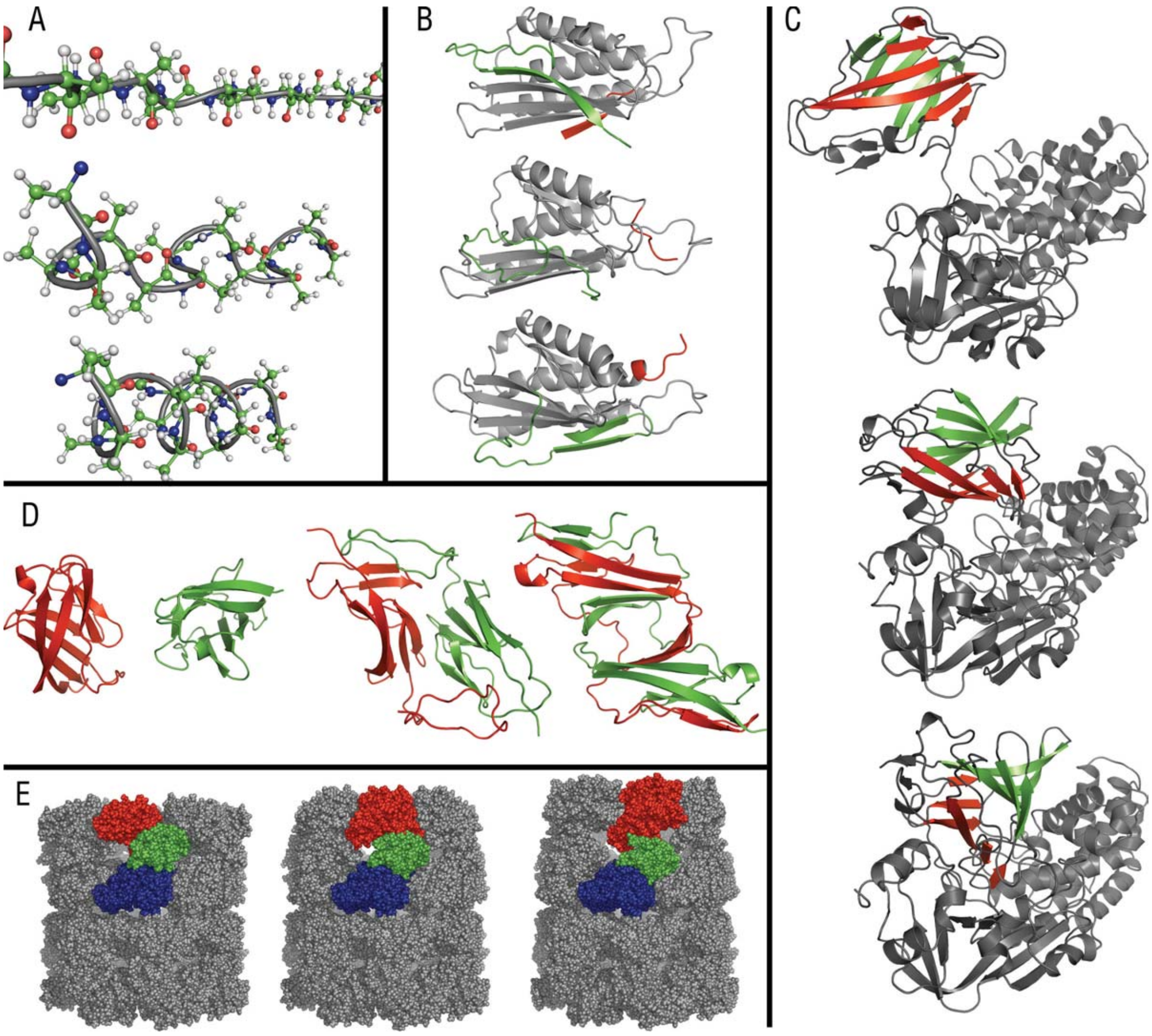

Figure 2

Example pathways that completed successfully without the use of backtracking or random motion. Each panel shows three pathway snapshots: the initial structure, an intermediate snapshot, and the final snapshot. Colored regions in green, red, and blue (panels **B**–**E**) highlight particular motions, described in the text. **A**: Toy Model 1. A polyalanine strand of beta sheet gradually folds to alpha helix while maintaining geometric constraints. **B**: Spindle assembly checkpoint protein. A strand of beta sheet (red) passes under a loop and joins with an alpha helix, while a second strand of beta sheet (green) moves from the left to right side of the beta sheet **C**: Diphtheria Toxin. A very large domain rotation of nearly 180° is shown. Observe the relative position of the red and green colored beta sheets over the course of the rotation, in which the green beta sheet begins behind the red beta sheet, then rotates to be on top, then rotates further to end up on the right of the red. **D**: CD2. Two monomers (red and green) dimerize, forming a domain-swapped dimer. **E**: GroEL. Each subunit in the upper ring of the large 14-subunit complex undergoes a transition involving a large clockwise rotation and upward tilt of the apical domain (red) and a downward tilt of the intermediate domain (green), while the equatorial domain (blue) remains relatively unchanged. For clarity, only one subunit is colored.

The pathway generated for diphtheria toxin shows a very large domain rotation of nearly 180° [Fig. 2(C)], created without any backtracking or random motion. Observe the relative position of the red and green colored beta sheets over the course of the rotation, in which the green beta sheet begins behind the red beta sheet, then rotates to be on top, then rotates further to end up on the right of the red. The initial state was taken from a domain-swapped state (only one monomer shown), and the final state was the native, non-swapped state. The pathway generated here by geometric targeting is in contrast to the result obtained from the linear-interpolation-based method at the Yale Morph Server,[26] which produces an unphysical pathway with atoms passing through each other. Krebs *et al.*[26] declared the conformational change in diphtheria toxin to be "impossible" to compute, speculating that only a complete unfolding and refolding of the domain could explain the conformational change. The insight gained from the geometric targeting approach is that a plausible pathway does exist that does not involve unfolding/refolding, although the method makes no prediction as to the actual or optimal pathway. This is in harmony with the results from elastic network interpolation on this protein.[22]

A misfolding pathway is shown in Figure 2(D), as two monomers of the protein CD2 dimerize to form a domain-swapped dimer. The motion is complicated and non-linear, involving the domains opening up and inter-digitating, which is notable considering that the pathway was generated without random motion and without any backtracking. In the figure, the two monomers were given separate colors to highlight how their beta-strands intermingle.

The conformational change in the large, 14-subunit GroEL complex is shown in Figure 2(E). Two 7-subunit rings are stacked on top of each other, viewed from the side as the top ring undergoes an opening and twisting transition. Although the initial and final states have seven-fold rotational symmetry, symmetry was not enforced in the pathway and all atoms were explicitly simulated. In the figure, the apical, intermediate, and equatorial domains of one subunit are given distinct colors to show how they change in the pathway. GroEL is another case that results in an unphysical pathway when run on the Yale Morph Server.[26] It is important to note that a targeted molecular dynamics[43] study of a single GroEL monomer indicates that the intermediate domain motion occurs first, followed by the apical domain motion. In the pathway generated from geometric targeting, all motions occur simultaneously because it is stereochemically plausible to do so, and because the energetics of the system are not considered. Still, the types of movements involved in the transition can be seen in the pathway, even if the relative timing of events is not accurate.

Next we present some results for pathways that required backtracking in order to successfully reach the target. Toy model 2 was created to help illustrate how difficult obstacles can cause the targeting to get stuck, and how backtracking can sometimes help in these cases. The model is a 4-helix bundle in the initial state. The target state was created by moving one of the helices to the opposite side of the bundle, shown in Figure 3(A) (fourth snapshot). The other three helices form a wall that the mobile helix must somehow pass to reach its target state on the other side. The helices making up the wall cannot separate because there are hydrogen bond and hydrophobic constraints between helices that are common to both the initial and target structures, which are kept fixed. Without backtracking, the gradually decreasing RMSD simply pulls the mobile helix into the wall formed by the other three (not shown). With backtracking enabled, instead of quitting when stuck, the RMSD is backed, momentum steps are performed to find a new starting state, and the forward steps commence again. However, as shown in the second snapshot of Figure 3(A), the system is still stuck, with the mobile helix trying to go around the side of the wall, but unable to because of its loop attachment to the wall. After more backtracking attempts, the helix is observed to flip over the top of the wall to reach the other side [Fig. 3(A), third snapshot].

In HIV protease, initial and target structures were chosen that would require the two flexible arm regions to move past each other [Fig. 3(C)]. The green-colored region is behind the red region in the initial state, but is in front of the red region in the target state. Without backtracking and without the use of random motion, the RMSD constraint pulls these two arm regions into each other, causing them to collide and get stuck [Fig. 3(C), second snapshot]. With backtracking enabled, the arms back away after colliding. During the momentum steps that follow the backward steps, the red region moves over the top of the green region [Fig. 3(C), third snapshot]. As forward steps begin again, the system has moved around the obstacle, enabling the RMSD constraint to pull the system to the target state.

The main part of the motion in dengue 2 virus envelope glycoprotein, shown in Figure 3(C), is the closing of a hinged domain (green) against the stable portion of the protein (gray). This motion is accomplished easily, however a small loop region (red) that needs to move gets pinned by the closing domain [Fig. 3(C), second snapshot]. With backtracking enabled, the system finds a new starting configuration that is not obviously very different [Fig. 3(C), third snapshot], however, when forward steps recommence the red loop region is able to slip out and move to its target position [Fig. 3(C), fourth snapshot].

In the antibody Immunoglobulin E SPE7, we performed targeting between two structures that exhibited some loop and side chain conformational differences in the heavy chain. We found that we had to make the common hydrogen bond and hydrophobic contact constraints breakable to successfully reach the target. Typically, common constraints are kept fixed, under the assumption that if the interactions are present in the initial state and in the final state, they are also present at intermediate states. In this protein, all targeting attempts with common constraints kept fixed were unsuccessful, even with backtracking and random motion activated. When common constraints are made breakable, however, targeting is successful without backtracking and without random motion, indicating that the pathway requires some hydrogen bond or hydrophobic contact to transiently break and reform.

All examples presented so far did not use random motion. To demonstrate the use of random motion, we performed additional targeting runs on the HIV protease system with random motion activated. Recall that without random motion, the arms of the protein could only move past each other if backtracking was used. With random motion added to each RMSD step, and using an RMSD step size of 0.01 Å, we find that targeting is successful without backtracking. The random motion enables the two colliding arms to find a way to slip past each other without getting stuck. Figure 4 shows the variability in the pathway introduced by the random motion. The superimposed snapshots shown were taken

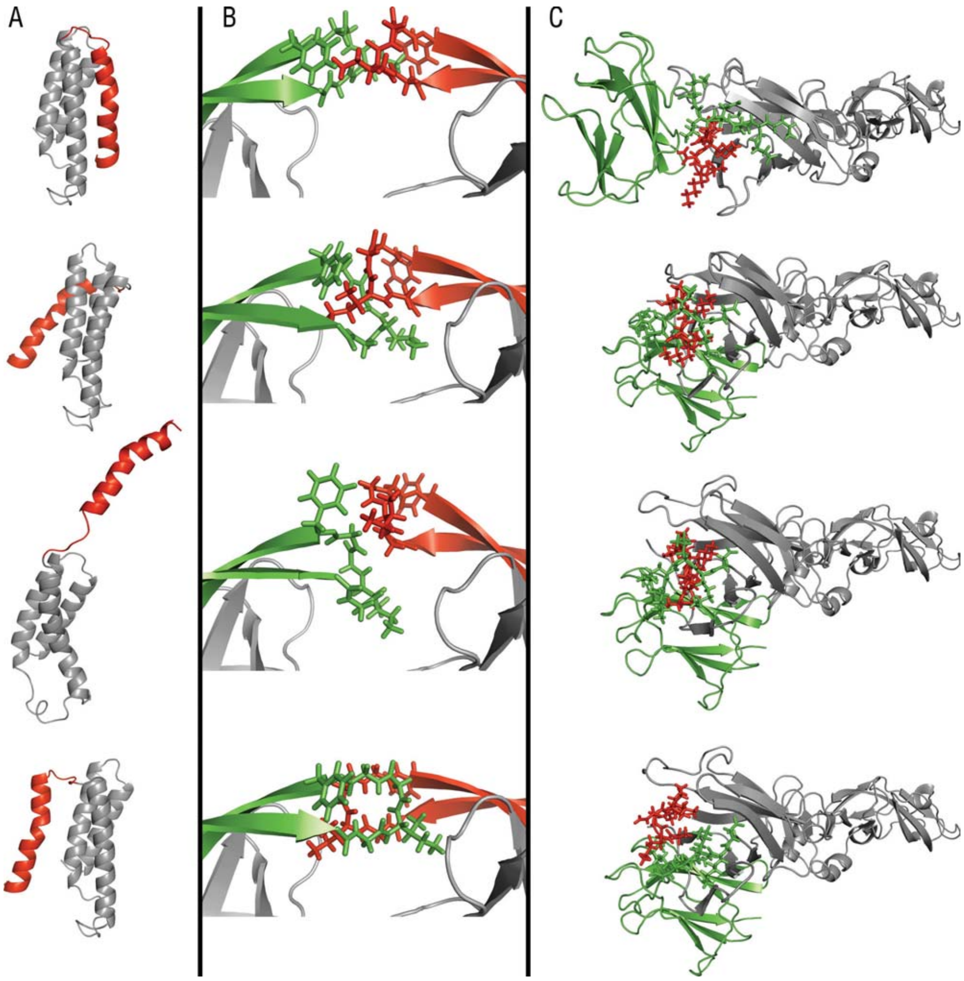

Figure 3

Example pathways that required backtracking. Each panel shows four pathway snapshots: the initial structure, a snapshot when the system encounters an obstacle, a snapshot when the protein moves around the obstacle, and the final snapshot. **A**: Toy Model 2. One helix (red) of a 4-helix bundle must move from the right side to the left, the other three helices forming a wall. The helix tries to go around the side of the wall (second snapshot), but is unable to because of its loop attachment to the wall. After more backtracking attempts, the helix is observed to flip over the top of the wall (third snapshot) to reach the other side. **B**: HIV Protease. Two arm regions (green and red) must somehow switch places, with the red-colored arm moving from front to back. Targeting initially gets stuck (second snapshot), but after backtracking finds a way around the obstacle (third snapshot). **C**: Dengue 2 Virus Envelope Glycoprotein. A hinged domain (green) closes up against the stable portion of the protein (gray), but a small loop region (red) that needs to move gets pinned by the closing domain (second snapshot). With backtracking enabled, the system finds a new starting configuration that is not obviously very different (third snapshot), however, when forward steps recommence the red loop region is able to slip out and move to its target position (fourth snapshot).

from three independent targeting runs, at 1.24 Å RMSD from the target. Interestingly, the random motion was not successful when used in conjunction with a larger RMSD step size of 0.1 Å. With a large RMSD step, the RMSD changes so rapidly that the random motion does not have enough opportunity to prevent the arms from getting stuck.

DISCUSSION

Several successful examples of the application of geometric targeting have been presented, illustrating that the technique is generally applicable to a wide variety of conformational changes. Especially promising is the application to very large systems, demonstrated in the 14-subunit GroEL complex, for which an all-atom pathway was produced in under 2 h on a single CPU. A significant improvement compared to linear-interpolation techniques[26] has been demonstrated in that the pathways produced by geometric targeting do not have chains passing through each other. The geometric constraints between atoms serve to keep the system in plausible geometric configurations, redirecting atoms along curved paths as the target is approached. Furthermore, backtracking,

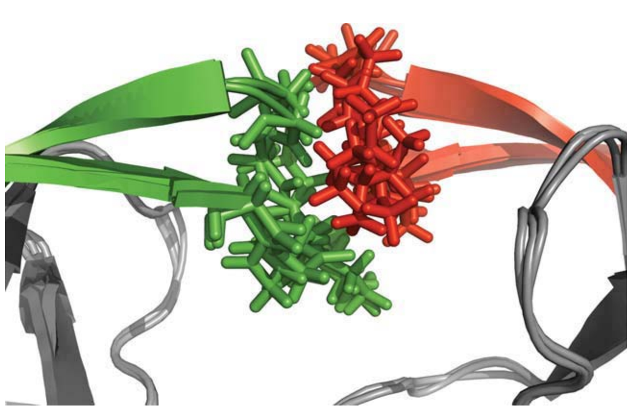

Figure 4

Pathways with random motion. Three snapshots are superimposed, each taken from the same intermediate point of three different random runs in HIV Protease. The random motion enables the arm regions to find a way to pass each other. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

random motion, and breakable constraints have been shown to enable the system to get past particularly difficult obstacles.

Beyond the results presented, we have also found that creating a geometric pathway between an unfolded state (linear polypeptide) of a protein and its folded state is successful in some cases (data not presented). It is not successful for some folds, for example in knotted proteins. Though broadly applicable to diverse kinds of conformational changes, there are sure to be other examples of pathological transitions where geometric targeting is not successful.

By including random motion in the pathways, it is possible to create random pathways. Because the random perturbations are performed at the level of individual rigid units, however, the diffusive motion of the rigid units has little effect on domains. The domain motions are more heavily influenced by the changing RMSD constraint than they are by the random motion. In principle, using a very small RMSD step size would give the random motion more opportunity to have effect at the level of domains.

A limitation to be aware of in geometric targeting is the tendency for events to happen simultaneously along the pathway, due to the flat zero-energy landscape within which all configurations reside that meet the geometric constraints. In the GroEL case, for example, the intermediate domain and apical domain motions occur simultaneously in the geometric pathway, whereas targeted molecular dynamics[43] shows the domain movements occurring sequentially. Similarly, in adenylate kinase, the geometric pathway has the ATP-binding domain and the lid domain closing together simultaneously, rather than sequentially, as was found using elastic network models[23,24] and in the work of Arora *et al.*[44] which used nudged elastic band[5,6] and umbrella sampling.[8] As an attempt to induce sequential domain movement in GroEL and adenylate kinase, we tried rerunning them with non-common constraints included as breakable constraints (instead of the usual setting, which is to not include the non-common constraints), but the correct timing was still not obtained. What is learned from geometric targeting in these two cases is that there is no steric reason why the timing of events must be a certain way, and the change can be performed while preserving all the geometric constraints. The timing must be due to energetics.

In many cases, the motions determined by geometric targeting may accurately capture the essential geometric features of a transition. On the other hand, it is important to recognize the fundamental limitations of geometric targeting that arise from its neglect of thermodynamics and lack of Boltzmann weighting. Some geometric features may be successfully captured, but features that depend on energetic considerations will be missed, such as transient hydrogen bonds and the timing of events, which could have significant effects on the pathway.

It is interesting to consider what would happen if an all-atom molecular mechanical force field was used instead of the constraint energy function, and if the rigid units were reduced to single atoms. The procedure would then be to perform energy minimizations at each RMSD level in an attempt to produce a low-energy pathway. We have not tried this, and we do not know whether such a technique would be an improvement or if new problems would arise, such as getting trapped in local minima.

It is also interesting to examine the set of all pathways to the target, under different conditions – keeping common constraints only, adding randomness and adding the option of backtracking. It is also interesting to reverse the direction of the targeting between the two protein conformations. This leads to a plethora of data that will be the subject of a future study in which “bottlenecks” along the pathway are identified—these are narrow regions of phase space through which the structure passes.

The geometric targeting method introduced in this article has some similarities and differences with the FRODA-targeting method published earlier.[29] The similarities are in the overall idea, which in both cases is to move the system towards the target state while enforcing a set of geometric constraints to keep the structure stereochemically acceptable. Though similar in overall idea, the underlying geometric model and manner in which constraints are enforced are quite different, leading to improvements in speed, ability to successfully enforce constraints, and ability to more closely reach the target. Differences in the model and method that facilitate these improvements include the following: the use of a constraint energy function with conjugate gradient minimization to enforce constraints; the use of an explicit RMSD constraint in the targeting; small rigid units instead of large rigid clusters; maximum-distance constraints for hydrogen bonds instead of rigid distance and angle constraints; new minimum-distance constraints calibrated from MD simulations; minimum-distance constraints for maintaining good Ramachandran quality; and minimum-distance constraints for favorable side-chain torsional configurations.

CONCLUSION

We have created a new method for pathway generation in proteins, with an easy-to-use webserver, that is quick and produces stereochemically correct pathways. When compared to more sophisticated and computationally expensive methods like targeted molecular dynamics, this method can be thought of as a "back-of-the-envelope" calculation. It is a quick and easy method to gain preliminary insights into a pathway. The geometric constraints used here model the physical reality that motion in proteins is highly constrained. Although the neglect of energetic considerations certainly affects the details of the outcome, in many cases, geometric considerations alone may be sufficient to capture the essential translational and rotational motions that make up the actual pathway. At a minimum, these pathways are useful for visualization purposes, to easily see what is changing and what motions might be involved in the change. But beyond visualization, the stereochemical quality of these pathways makes them candidates for input to more intensive quantitative approaches.

Future planned developments on the pathways website include extensions of the technique to handle RNA, DNA and ligands. This site is a companion to http://flexweb.asu.edu which uses similar techniques to explore undirected motion.

ACKNOWLEDGMENTS

The authors thank Stephen A. Wells for many interesting discussions, Maria Kurnikova and Tatyana Mamonova for their insights and for providing MD trajectories. M.F.T. thanks the Leverhulme Trust and Sophia Yaliraki for support at Imperial College. D.W.F. thanks Boehringer Ingelheim Fonds for travel funding. They also take this opportunity to thank the anonymous referees of this article for providing additional challenging initial and target structures, all of which have been incorporated into this article.

REFERENCES

1. Bolhuis PG, Chandler D, Dellago C, Geissler PL. Transition path sampling: throwing ropes over rough mountain passes, in the dark. Annu Rev Phys Chem 2002;53:291.
2. Weinan E, Ren WQ, Vanden-Eijnden E. Finite temperature string method for the study of rare events. J Phys Chem B 2005;109:6688–6693.
3. Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G. String method in collective variables: minimum free energy paths and isocommittor surfaces. J Chem Phys 2006;125:24106.
4. Pan AC, Sezer D, Roux B. Finding transition pathways using the string method with swarms of trajectories. J Phys Chem B 2008;112:3432–3440.
5. Jónsson H, Mills G, Jacobsen KW. Nudged elastic band method for finding minimum energy paths of transitions. In: Berne BJ, Ciccoti G, Coker DF, editors. Classical and quantum dynamics in condensed phase simulations. Singapore: World Scientific; 1998. pp 385–404.
6. Chu JW, Trout BL, Brooks BR. A super-linear minimization scheme for the nudged elastic band method. J Chem Phys 2003;119:12708–12717.
7. Farrell D, Mamonova T, Kurnikova M, Thorpe MF. To be published.
8. Torrie GM, Valleau JP. Non-physical sampling distributions in Monte-Carlo free-energy estimation—umbrella sampling. J Comput Phys 1977;23:187–199.
9. Yang HJ, Wu H, Li DW, Han L, Huo SH. Temperature-dependent probabilistic roadmap algorithm for calculating variationally optimized conformational transition pathways. J Chem Theory Comput 2007;3:17–25.
10. Branduardi D, Gervasio FL, Parrinello M. From A to B in free energy space. J Chem Phys 2007;126.
11. Elber R, Karplus M. A method for determining reaction paths in large molecules—application to myoglobin. Chem Phys Lett 1987;139:375–380.
12. Fischer S, Karplus M. Conjugate peak refinement—an algorithm for finding reaction paths and accurate transition-states in systems with many degrees of freedom. Chem Phys Lett 1992;194:252–261.
13. Elber R. Long-timescale simulation methods. Curr Opin Struct Biol 2005;15:151–156.
14. Christen M, Van Gunsteren WF. On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: a review. J Comput Chem 2008;29:157–166.
15. Isralewitz B, Gao M, Schulten K. Steered molecular dynamics and mechanical functions of proteins. Curr Opin Struct Biol 2001;11:224–230.
16. Schlitter J, Engels M, Kruger P. Targeted molecular-dynamics—a new approach for searching pathways of conformational transitions. J Mol Graph 1994;12:84–89.
17. Schlitter J, Engels M, Kruger P, Jacoby E, Wollmer A. Targeted molecular-dynamics simulation of conformational change—application to the T[--]R transition in insulin. Mol Simul 1993;10:291.
18. van der Vaart A, Karplus M. Simulation of conformational transitions by the restricted perturbation-targeted molecular dynamics method. J Chem Phys 2005;122.
19. van der Vaart A, Karplus M. Minimum free energy pathways and free energy profiles for conformational transitions based on atomistic molecular dynamics simulations. J Chem Phys 2007;126.
20. Kim MK, Chirikjian GS, Jernigan RL. Elastic models of conformational transitions in macromolecules. J Mol Graph Model 2002;21:151–160.
21. Kim MK, Jernigan RL, Chirikjian GS. Rigid-cluster models of conformational transitions in macromolecular machines and assemblies. Biophys J 2005;89:43–55.
22. Song G, Jernigan RL. An enhanced elastic network model to represent the motions of domain-swapped proteins. Proteins 2006;63:197–209.
23. Feng Y, Yang L, Kloczkowski A, Jernigan RL. The energy profiles of atomic conformational transition intermediates of adenylate kinase. Proteins 2009;77:551–558.
24. Maragakis P, Karplus M. Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. J Mol Biol 2005;352:807–822.
25. Schuyler AD, Jernigan RL, Qasba PK, Ramakrishnan B, Chirikjian GS. Iterative cluster-NMA: a tool for generating conformational transitions in proteins. Proteins 2009;74:760–776.

26. Krebs WG, Gerstein M. The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. Nucleic Acids Res 2000;28:1665–1675.
27. Flores S, Echols N, Milburn D, Hespenheide B, Keating K, Lu J, Wells S, Yu EZ, Thorpe M, Gerstein M. The database of macromolecular motions: new features added at the decade mark. Nucleic Acids Res 2006;34:D296–D301.
28. Echols N, Milburn D, Gerstein M. MolMovDB: analysis and visualization of conformational change and structural flexibility. Nucleic Acids Res 2003;31:478–482.
29. Wells S, Menor S, Hespenheide B, Thorpe MF. Constrained geometric simulation of diffusive motion in proteins. Phys Biol 2005;2:S127–S136.
30. Seeliger D, Haas J, de Groot BL. Geometry-based sampling of conformational transitions in proteins. Structure 2007;15:1482–1492.
31. Seeliger D, De Groot BL. tCONCOORD-GUI: visually supported conformational sampling of bioactive molecules. J Comput Chem 2009;30:1160–1166.
32. Angel H. Biomolecules in the computer: Jmol to the rescue. Biochem Mol Biol Ed 2006;34:255–261.
33. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285:1735–1747.
34. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and clustal X version 2.0. Bioinformatics 2007;23:2947–2948.
35. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. Proteins Struct Funct Genet 2001;44:150–165.
36. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol 1963;7:95.
37. Ho BK, Thomas A, Brasseur R. Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. Protein Sci 2003;12:2508–2522.
38. Dahiyat BI, Gordon DB, Mayo SL. Automated design of the surface positions of protein helices. Protein Sci 1997;6:1333–1337.
39. Wells SA, Dove MT, Tucker MG. Finding best-fit polyhedral rotations with geometric algebra. J Phys Condens Matter 2002;14:4567–4584.
40. Shewchuk JR. An introduction to the conjugate gradient method without the agonizing pain. Available at: http://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf; 1994.
41. Fletcher R, Reeves CM. Function minimization by conjugate gradients. Comput J 1964;7:149.
42. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol 2003;10:980–980.
43. Ma JP, Sigler PB, Xu ZH, Karplus M. A dynamic model for the allosteric mechanism of GroEL. J Mol Biol 2000;302:303–313.
44. Arora K, Brooks CL, III. Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. Proc Natl Acad Sci USA 2007;104:18496–18501.
45. The PyMOL Molecular Graphics System, version 1.2r1: Schrödinger, LLC. http://www.pymol.org/.

ID	Type Name	Type Description
0	CT_A	Amber CT type (sp3 carbon) with only carbon and hydrogen neighbors
1	CT_B	Amber CT type (sp3 carbon) with at least one neighbor that is not carbon or hydrogen
2	C	Amber C type (carbonyl sp2 carbon)
3	CA	Amber CA type (aromatic sp2 carbon in 6-membered rings and CE of Arg)
4	Cother	All other carbon types
5	N	Amber N type (sp2 nitrogen in amides)
6	N3	Amber N3 type (sp3 nitrogen)
7	Nother	All other nitrogen types
8	O	Amber O type (sp2 oxygen in amides)
9	O2	Amber O2 type (sp2 oxygen in anionic acids, COO-)
10	OH	Amber OH type (sp3 oxygen with bonded hydrogen)
11	Oother	All other Amber oxygen types
12	S	All sulfur types
13	HN	Amber H type (hydrogen attached to nitrogen)
14	HS	Amber HS type (hydrogen attached to sulfur)
15	HO	Amber HO and HW type (hydrogen attached to oxygen/water)
16	HA	Amber HA type (hydrogen attached to aromatic carbon)
17	HC	Amber HC type (hydrogen attached to aliphatic carbon with no electron-withdrawing substituents)
18	Hother	All other hydrogen types

Supplementary Table S1. Atom Types for non-overlap distance constraints. Based on Cornell et al., 1995

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. (1995). A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. J. Am. Chem. Soc. 117, 5179-5197.

Minimum Distance Constraints

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
		CT_A	CT_B	C	CA	Cother	N	N3	Nother	O	O2	OH	Oother	S	HN	HS	HO	HA	HC	Hother
0	CT_A	3.45	3.40	3.15	3.40	3.40	3.10	3.10	3.40	3.05	3.35	3.25	3.25	3.30	2.45	2.80	2.80	2.85	2.70	2.65
1	CT_B	3.40	3.40	3.25	3.35	3.50	3.05	3.05	3.20	3.00	3.15	3.10	3.10	3.30	2.55	2.90	2.90	2.70	2.75	2.45
2	C	3.15	3.25	3.10	3.10	3.25	3.00	3.00	3.15	2.80	2.95	3.15	3.15	3.30	2.25	2.35	2.35	2.75	2.60	2.55
3	CA	3.40	3.35	3.10	3.40	3.30	3.05	3.05	3.25	3.15	3.20	3.30	3.30	3.30	2.40	2.95	3.05	2.65	2.65	2.60
4	Cother	3.40	3.50	3.25	3.30	3.30	3.15	3.15	3.30	3.15	3.15	3.30	3.30	3.30	2.40	3.05	3.05	2.65	2.70	2.70
5	N	3.10	3.05	3.00	3.05	3.15	3.10	3.10	3.10	2.75	2.80	2.85	2.85	3.20	2.20	2.60	2.35	2.60	2.50	2.50
6	N3	3.10	3.05	3.00	3.05	3.15	3.10	3.10	3.10	2.75	2.70	2.85	2.85	3.20	2.20	2.60	2.35	2.60	2.50	2.50
7	Nother	3.40	3.20	3.15	3.25	3.30	3.10	3.10	3.10	2.75	2.70	2.85	2.85	3.20	2.10	2.60	2.35	2.60	2.65	2.55
8	O	3.05	3.00	2.80	3.15	3.15	2.75	2.75	2.75	2.95	3.05	2.65	2.65	3.20	1.80	1.95	1.70	2.45	2.45	2.25
9	O2	3.35	3.15	2.95	3.20	3.15	2.80	2.70	2.70	3.05	4.15	2.55	2.55	3.20	1.75	1.95	1.65	2.60	2.50	2.50
10	OH	3.25	3.10	3.15	3.30	3.30	2.85	2.85	2.85	2.65	2.55	2.95	2.95	3.20	1.95	1.95	2.10	2.60	2.55	2.50
11	Oother	3.25	3.10	3.15	3.30	3.30	2.85	2.85	2.85	2.65	2.55	2.95	2.95	3.20	1.95	1.95	2.10	2.60	2.55	2.50
12	S	3.30	3.30	3.30	3.30	3.30	3.20	3.20	3.20	3.20	3.20	3.20	3.20	4.00	2.70	2.70	2.70	2.70	2.70	2.70
13	HN	2.45	2.55	2.25	2.40	2.40	2.20	2.20	2.10	1.80	1.75	1.95	1.95	2.70	1.85	2.25	1.95	2.10	1.85	2.00
14	HS	2.80	2.90	2.35	2.95	3.05	2.60	2.60	2.60	1.95	1.95	1.95	1.95	2.70	2.25	2.45	2.45	2.20	2.25	2.20
15	HO	2.80	2.90	2.35	3.05	3.05	2.35	2.35	2.35	1.70	1.65	2.10	2.10	2.70	1.95	2.45	2.45	2.20	2.25	2.20
16	HA	2.85	2.70	2.75	2.65	2.65	2.60	2.60	2.60	2.45	2.60	2.60	2.60	2.70	2.10	2.20	2.20	2.35	2.25	2.10
17	HC	2.70	2.75	2.60	2.65	2.70	2.50	2.50	2.65	2.45	2.50	2.55	2.55	2.70	1.85	2.25	2.25	2.25	2.20	2.10
18	Hother	2.65	2.45	2.55	2.60	2.70	2.50	2.50	2.55	2.25	2.50	2.50	2.50	2.70	2.00	2.20	2.20	2.10	2.10	2.05

Supplementary Table S2. Non-bonded Minimum-Distance Constraints, defined based on atom types in Table S1.

Protein Name	# Sub-units	Initial PDB	Final PDB	Chain Information
5’-Nucleotidase	1	1HP1	1HPU	Chain A from initial structure targeted to chain C of final.
Adenylate Kinase	1	4AKE	1AKE	Chain A from initial structure targeted to chain A of final.
Alcohol Dehydrogenase	1	8ADH	6ADH	Chain A from initial structure targeted to chain A of final.
Calmodulin	1	1CFD	1CFC	Chain A from initial structure targeted to chain A of final.
CD2	2	1HNG	1CDC	Chains AB from initial structure targeted to chains AB of final.
Citrate Synthase	2	5CSC	6CSC	Chains AB from initial structure targeted to chains AB of final.
Collagenase	1	1NQD	1NQJ	Chain A from initial structure targed to chain B of final structure
Dengue 2 Virus Envelope Glycoprotein	1	1OAN	1OK8	Chain A from initial structure targeted to chain A of final.
Dihydrofolate Reductase	1	1RX2	1RX6	Chain A from initial structure targeted to chain A of final.
Diphtheria Toxin	1	1DDT	1MDT	Chain A from initial structure targeted to chain A of final.
DNA Polymerase	1	1IH7	1IG9	Chain A from initial structure targeted to chain A of final.
GroEL	14	1KP8	1AON	The chains in these two PDB files are not labeled consistently. The correct mapping of chains that we used in targeting was ABCDEFG to FEDCBAG (top ring) and HIJKLMN to HIJKLMN (bottom ring).
Heparin Cofactor II	1	1JMO	1JMJ	Chain A from initial structure targeted to chain A of final.
HIV Protease	2	2HB4	2AZ8	Since both PDB files only contain one subunit (chain A) in the assymmetric unit, the full biological unit (2 subunits) was downloaded from PDB and relabled as chains AB. Chains AB from initial structure were targeted to chains AB of final structure.
HIV-1 Reverse Transcriptase	2	1DLO	2HMI	Chains A+B from initial structure targeted to chains A+B of final.
Immunoglobulin E SPE7	2	1OAQ	1OCW	Chains HL from initial structure targeted to chains HL of final.
Phosphofructokinase	4	4PFK	6PFK	Since initial PDB file 4PFK only contains one subunit (chain A) in the assymetric unit, the full biological unit (4 subunits) was downloaded from PDB and relabeled as chains ABCD. These were targeted to chains ABCD of final structure.
Pyrophosphokinase	1	1HKA	1Q0N	Chain A from initial structure targeted to chain A of final.
Pyruvate Phosphate Dikinase	1	1KBL	2R82	Chain A from initial structure targeted to chain A of final.
Replication Factor C	6	2CHV	2CHQ	Since final PDB file 2CHQ only contains 3 subunits (chains ABC) in the assymetric unit, the full biological unit (6 subunits was downloaded from the PDB and relabeled as chains ABCDEF. These were targeted to chains ABCDEF of the final structure.
Rho Transcription Termination Factor	6	3ICE	3ICE	Same PDB file was used for both initial and final structures. Chains ABCDEF of initial state were targeted to chains BCDEFA, to simulate one step in the cyclic transition of this protein.
Spindle Assembly Checkpoint Protein	1	1DUJ	1KLQ	Chain A from initial structure targeted to chain A of final.
Toy Model 1	1	-	-	
Toy Model 2	1	-	-	

Table S3. PDB IDs and chain information