PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Building alternate protein structures using the elastic network model

Qingyi Yang and Kim A. Sharp*

Johnson Research Foundation and Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, Pennsylvania 19104

ABSTRACT

We describe a method for efficiently generating ensembles of alternate, all-atom protein structures that (a) differ significantly from the starting structure, (b) have good stereochemistry (bonded geometry), and (c) have good steric properties (absence of atomic overlap). The method uses reconstruction from a series of backbone framework structures that are obtained from a modified elastic network model (ENM) by perturbation along low-frequency normal modes. To ensure good quality backbone frameworks, the single force parameter ENM is modified by introducing two more force parameters to characterize the interaction between the consecutive carbon alphas and those within the same secondary structure domain. The relative stiffness of the three parameters is parameterized to reproduce B-factors, while maintaining good bonded geometry. After parameterization, violations of experimental Cα—Cα distances and Cα—Cα—Cα pseudo angles along the backbone are reduced to less than 1%. Simultaneously, the average B-factor correlation coefficient improves to $R = 0.77$. Two applications illustrate the potential of the approach. (1) 102,051 protein backbones spanning a conformational space of 15 Å root mean square deviation were generated from 148 nonredundant proteins in the PDB database, and all-atom models with minimal bonded and nonbonded violations were produced from this ensemble of backbone structures using the SCWRL side chain building program. (2) Improved backbone templates for homology modeling. Fifteen query sequences were each modeled on two targets. For each of the 30 target frameworks, dozens of improved templates could be produced In all cases, improved full atom homology models resulted, of which 50% could be identified blind using the D-Fire statistical potential.

Proteins 2009; 74:682–700.

Key words: elastic network model; normal mode analysis; coarse-grain model; protein conformation generation.

INTRODUCTION

Generation of alternative protein conformations, starting from a known protein structure, has a variety of uses in modeling. It can be used for exploring conformational space, modeling dynamic properties, producing alternative templates for homology modeling, generating putative transition pathways, and can aid sampling in thermodynamic calculations. It is particularly important for protein design and comparative modeling, since both applications rely on reasonable backbone frameworks. Some applications such as homology modeling require only backbone structures, other applications require a full atomic model with backbone and side chain coordinates. A major barrier, however, is the difficulty in obtaining an ensemble of structures that are both significantly different from each other and structurally plausible. By the latter, we mean that they must not violate bond length and bond angle constraints (bonding geometry) imposed by the covalent structure, and steric clashes (overlap of nonbonded atom) must be avoided. These are necessary although not sufficient requirements for generating a low energy structure that might actually be adopted by the protein with reasonable probability. Other requirements such as good packing, optimization of H-bonding and salt bridges, burial of hydrophobic groups, and so forth will be required for a low energy structure, but these cannot sensibly be addressed if the first two requirements are missing. Given a well-packed, highly optimized native structure, generating plausible alternatives is not easy. An attractive solution to maintain good bonding geometry is to randomly or uniformly sample in torsional coordinate space. However, with a large conformational space just from backbone dihedral angles, ϕ and Ψ, alone, it is difficult to find combinations that produce a conformation that is sterically correct. In the flexible backbone design method of Kuhlman *et al.*,[1] the backbone candidates are generated from small random changes in selected torsion angles, followed by structure optimization. This method has been used successfully in designing proteins of modest size. For large proteins, it requires a lot of computation. The randomized backbone ensemble design

*Correspondence to: Kim Sharp, Johnson Research Foundation and Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA 19104.
E-mail: sharpk@mail.med.upenn.edu
Received 14 January 2008; Revised 27 May 2008; Accepted 8 June 2008
Published online 14 August 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22184

method of Desjarlais and Handel[2] is also expensive, and the root mean squared (rms) difference from the starting structure is limited to a small range, for example, 1 Å. An alternative is to do a Monte-Carlo search. This method also requires significant memory and computation. Both flexible and random backbone design methods first generate structures by manipulating torsion angles and then eliminating those energetically unfavorable ones. Of the large torsional phase space, only a tiny fraction has conformations a proteins can adopt, thus most of the computational time is used in filtering out unfavorable structures. Another approach is to perform molecular dynamics, either letting the protein trajectory evolve freely, or by targeting it toward a given structure.[3] Since all atoms are explicitly represented, along with their full bonded and nonbonded interactions, and the total energy of the protein is limited by running simulations at constant temperature, covalent and steric correctness is well-maintained. However, this method is even more time-consuming, and since the time step is quite short (usually 1–2 fs), many nanoseconds of simulations must be run to produce an ensemble of structures that are significantly different from each other. Alternate backbone structures can also be generated using classical normal mode methods,[4] although only a limited distortion can be achieved with a single set of modes without producing bad geometries. An alternative approach, which we describe here, is to use a multiscale approach in which structural ensembles produced using a coarse-grain model are then used to generate plausible fine grained, all-atom models. We take advantage of the remarkable success of the Tirion-type of elastic network, or Gaussian Network type potential coarse-grain models in a variety of protein applications, such as reproduction of X-ray B-factor[5,6] or NMR order parameters,[7] refinement of X-ray structures,[8] generation of conformational transition pathways.[9–25]

Our goal was to develop an efficient method for protein structure generation, which is able to (1) provide a flexible backbone scaffold with reasonable geometry for application that require backbones only, (2) support rapid generation of plausible all-atom models. Several design criteria were considered. The first step is to generate a candidate backbone framework, for which there are two important considerations. First is the correctness and the reasonableness of the backbone. There should be no bad contacts, bond angles and lengths should be correct, and all the dihedral angles should be in the allowed regions. Second, the ensemble of protein backbones should reflect the structural flexibility of that protein. If the ensemble of backbones sample with some fidelity conformations likely to be adopted under realistic conditions, it would greatly reduce the structure searching or optimization time, and make the resulting structures more useful. The third criterion is that the method be rapid and require minimal user input, so that potentially thousands of structures can be produced. Furthermore, we require the method to be robust and applicable to different size proteins. In outline, the method starts with backbone candidates generated at a coarse-grained level using a carefully parameterized nonlinear elastic network model (ENM), reconstructing the structures along the low-frequency normal modes. A full atom backbone model is produced from this by parsimonious modeling of the peptide units, to produce an intermediate level model. Finally a side chain building program, SCWRL[26] is used to generate the final all-atom fine grained model. We first describe how the ENM is parameterized and tested, and then describe the generation of peptide backbone and side chain atom positions. We then describe two applications of the method to illustrate its usefulness: Generation of a large ensemble of conformations from multiple protein structures, and application to improved homology modeling templates.

METHODS

Protein ENM: multiforce parameter potential

Our starting point is a coarse-grain model, which is derived from the single-parameter type elastic network potential.[11] In this model, the Ca atom is the only interacting site and all of the interactions between interacting sites are pairwise spring potentials with a single force constant K

$$E(a,b) = \frac{K}{2}(r_{a,b} - r^0_{a,b})^2 \qquad (1)$$

where, if $\boldsymbol{r}_a$, $\boldsymbol{r}_b$ denote the coordinates of Ca atoms a and b, $r_{a,b} = |\boldsymbol{r}_a - \boldsymbol{r}_b|$ is the distance between them. The zero superscript indicates the coordinates of the original conformation. The mechanical/dynamical properties of an elastic network, consisting of multiple springs of the form of Eq. (1), are easily obtained using the method of normal modes. This simple model has proven remarkably useful, giving a quite realistic description of B-factor profiles along the polypeptide chain,[11] modeling realistic large scale conformational changes in the ribosome,[8] and other examples.[9,27,28] However, it suffers drawbacks as a basis for generating fine-grained protein structures: All Ca–Ca interactions are treated identically, whereas in practice $i - i + 1$ and $i - i + 2$ Ca–Ca distances are tightly bounded by covalent bond interactions. Thus, with a one-spring constant ENM (1K-ENM) one must either keep net displacements very small, or tolerate serious violations of the covalent geometry that prevents from building up of reasonable all-atom models. In addition, a single force constant cannot capture the various types of interactions within a protein. For example, if a structure contains two tightly packed helices, a single

force constant ENM would not distinguish interactions between sites within the same helix and those between different helices. Thus, it would be prone to either underestimate the relative stiffness of interactions within a H-bonded secondary structure element, or overestimate tertiary structure interactions. Our solution is to use three different spring potentials: K_a for $i - i + 1$ interactions, K_b for interactions between Ca's that are in the same secondary structure domain P (as defined by the DSSP program,[29] and K_c for the remaining (Tertiary structure) Ca–Ca interactions within some distance cutoff R. The total potential is now

$$E_p = \sum_{i}^{N} K_a (r^0_{i,i+1} - r_{i,i+1})^2 + \sum_{i,j \in p}^{N,N} K_b (r^0_{i,j} - r_{i,j})^2 + \sum_{i,j,|r_{i,j}|<R_c}^{N,N} K_c (r^0_{i,j} - r_{i,j})^2 \quad (2)$$

which defines a three spring constant ENM (3K-ENM). For R_c we use a value of 13 Å, which was previously shown to be appropriate for normal mode studies of proteins.[27]

Upon generation of the Ca network, $\boldsymbol{r}_{ij} = \boldsymbol{r}^0_{ij}$ by construction, and the structure is at its minimum energy. The normal mode frequencies ω_i and vectors $\mathbf{W}_i$ are then obtained in the usual manner by solving for the Eigen values λ.[30] Each normal mode vector specifies a set of atomic displacements

$$q_k = W_{ki} A_i \cos(\omega_i t + \varepsilon_i) \quad k = 7, 3N \quad (3)$$

where A_i and ε_i are the amplitude and phase of the mode and q_k is an Ca atomic coordinate component. To produce a new structure a small displacement δ is applied to the current coordinates $\mathbf{q}^o$ along a specific mode i, $q_k = W_{ki}\,\delta + q^0_k$, and a new Ca network is generated. From the new Ca network coordinates, we recalculate the normal modes and apply the displacement along a certain newly calculated mode. This procedure is repeated as necessary, with appropriate selection of modes, to produce an ensemble of Ca coordinates with a desired root mean squared (rms) displacement from the starting structure. Although the normal modes are generated by assuming that the fluctuations are Gaussian in distribution, the method is effectively nonlinear since the Ca network is regenerated as the conformational transition evolves. Thus the final structure does not correspond to a linear perturbation along some mode or combination of modes of the starting structure. Since the procedure is very rapid, only a small displacement is applied at each step before regenerating the Ca network. This avoids large undesirable conformational distortions. We have found setting the displacement $\delta = 0.1$ Å to be suitable for most applications. In principle, any one of the current 3N-6 modes or combinations of several modes may be used to generate each step. In practice we select from the first 3 to 6 nonzero lowest frequency modes, because studies show that large-scale conformational transitions are usually dominated by low-frequency modes. The method for choosing a suitable mode depends on the application, as described below.

Retaining correct covalent geometry in the ENM

Introduction of different force constants for neighboring, secondary structure and tertiary Ca–Ca interactions provides a more realistic description of the mechanics of a protein, but it does not by itself ensure that conformations have correct covalent geometry. The distance D between neighboring Ca's in sequence, i and $i + 1$, is tightly constrained by the bond lengths and angles involving the Ca, peptide C, and peptide N atoms. The allowable range for D was taken from the standard structure verification program PROCHECK[31] as 3.6–4.0 Å (excluding *cis*-Pro).

Similarly, because of the bonding geometry of the Ca and peptide units comprising the backbone (in particular, the C—Ca—N bond angle) the virtual angle θ between three consecutive Ca's $i - 1$, i, and $i + 1$ is limited in range. The alanine dipeptide, with PROCHECK ideal bond lengths and angles, was used to determine the allowable range. The dipeptide dihedral angles, Φ and Ψ, were sampled from -180° to 180° in 15° increments, and it was found that the virtual angle θ must be in the range of 95°–135°.

When a new Ca network is generated by the normal mode stepping procedure described above, all the Ca i, $i + 1$ distances and Ca $i - 1$, i, and $i + 1$ angles are checked against the allowable ranges on D and θ. If there are any violations, the structure is discarded. At this point, there are several options. The program chooses the next low frequency mode vectors for the trial step and tries again. If all six low frequency modes give violations, the program backtracks to the beginning and chooses a different mode for the first deformation step. Failure to find a suitable mode or series of modes causes the program to terminate, and output the maximum rms deviation structure it has found as well as the number of generated structures. The current implementation of the program only examines the six lowest modes to find violation-free structures. In principle, all modes could be considered at each step, but we have found in practice that higher frequency modes are unproductive as they usually involve displacements of local groups of atoms, and do not lead to meaningfully different structures. Also, given a choice of six modes at each step, it would be possible (although expensive) to do a full tree search over the hexa-branched tree of possible mode choices to find some "optimal" sequence of mode choices. Again,

we have found in practice that the "local" tree search described above is sufficient to produce good variant structures for the majority of proteins examined here. More "difficult" protein folds that yield only small rms deviations may need a deeper tree search, a point which will be examined in future refinements.

Building full backbone geometry

Given a candidate set of m Ca coordinates, the peptide units are then built in as follows. The N-terminus, C-terminus, and peptide units between Ca's 1 and 2 and between Ca's m − 1 and m are directly taken from the original coordinates and translated into the correct position without any rotation in Cartesian space.

For the other internal peptide units, ideal bond lengths (C—NH, C═O) and bond angles (Ca—C—O, Ca—C—N, C—N—Ca and O—C—N) are taken from Procheck. For a peptide unit between say Ca i and Ca $i + 1$, the orientation of the peptide plane is calculated by superimposing the original coordinates of four consecutive Ca^0 $i - 1$, i, $i + 1$, and $i + 2$ onto the new Ca $i - 1$, i, $i + 1$, and $i + 2$. Then the translation and rotation matrix from the superposition is applied to the peptide plane in the original coordinates, which results in the new orientation of the desired plane.

As the description of this backbone building strategy indicates, the newly built backbone geometry, while satisfying most covalent geometry constraints, is not necessarily optimized. This may be addressed by a short energy minimization applied to each structure after the mode stepping is finished to refine backbone geometry. However, the basic method is very fast and efficient in providing a large set of protein backbone structures, which have significant conformational difference. For many applications unminimized output is sufficient.

Three spring constant potential parameterization strategy

For the training set of proteins, we picked six increasingly large data sets, ranging from 19 to 157 proteins from the PDB database using Dunbrack's protein culling program PISCES[32] program. All the chosen proteins are high-quality X-ray structures with resolution better than 2.0 Å and R_{free} less than 20%. The sequence identity is less than 15% and lengths range from 60 to 398 amino acids. In normal mode analysis, the mean square fluctuation of atom i can be given as follows:

$$\langle q_i^2 \rangle = k_B T \sum_{k=1}^{3N-6} \left(\frac{W_{ik}}{\omega_k} \right)^2$$

where $\langle q_i^2 \rangle$ is the mass-weighted coordinate of atom i, k_B is the Boltzmann constant, T is the absolute temperature, and ω_k is frequency of k^{th} mode, which is related to the Eigenvalue as $\lambda_k^{1/2}$. Thus, the atomic temperature factor B_i can be calculated as

$$B_i = \left(\frac{8\pi^2}{3} \right) \langle q_i^2 \rangle$$

In all the comparisons, we used the original temperature factor values deposited in the PDB database as is. The B-factor of each Ca in the entire training set of proteins was calculated, and the Pearson's correlation coefficient between experimental and theoretical values of the temperature factor, R^0, obtained for each protein in the training set as a reference. In the original ENM, there is only one force parameter, K_c, which describes all the interactions, and its value acts only as a uniform scaling constant of the calculated B_i's, thus requiring no parameterization. In the three parameter cases, clearly from the form of the elastic network potential [Eq. (2)] the modes are invariant to uniform scaling of the spring constants. Only their relative ratios are relevant in calculating dynamical behavior of the system. Our parameterization goal was thus to find the optimal ratio between these three force parameters, K_a, K_b, and K_c. For a given set of three force parameters, K_a, K_b, and K_c, the correlation coefficient R, is recalculated for each protein. Then to optimize K_a and K_b, we used a grid search to maximize the sum of the percentage improvements in R

$$\Delta\% = \sum_{l}^{M} \frac{(R_l - R_l^0)}{R_l^0}$$

over all the training set, where R_l is the correlation coefficient of protein l, and M is the number of proteins in training set. Keeping the tertiary interaction force constant K_a fixed at 40, the other two force parameters, K_b and K_c were varied from 0.2 to 0.7, 1 to 40 with a step size of 0.1 and 1.0. This optimization was done on all six data sets, to determine the effect of set size on parameterization.

After optimization, the modified three force-parameter 3K-ENM was tested against a large test set of 1374 proteins. The proteins in the test set were chosen with the same criteria as those in the training set except that the length was allowed to be longer than 400 amino acids. The 3K-ENM is freely available as a server at: crystal.med.upenn.edu/software.

Homology modeling

Data set

To examine the use of the 3K-ENM for generating alternate backbone templates for homology modeling, we selected 15 query sequences, for each of which two template structures were used, for a total of 30 homology models. We selected the same set of query/template

structures as Misura *et al.*,[33] so that we would be able to compare to the all-atom models produced by Rosetta.[34] Misura *et al.* chose the test set to represent several different folds for which corresponding experimentally determined structures were available and for which homologous template structures could be confidently identified with PSI-BLAST.[35] According to Misura *et al.*, the range of sequence identity between the template and query sequences was 22–44% over the aligned regions. Alignments always covered >50% of the query sequence residues, with most test cases having >80% coverage. The size of the query sequences ranged from 58 to 255 residues, and all consisted of a single domain. All these criteria are well suited to the test of alternate backbone templates made here.

Structure-based alignments

Again following the strategy of Misura *et al.*,[33] sequence alignment of query and target structures was performed by first using a structure-based alignments, here using the combinatorial expansion (CE) method[36] with a medium similarity level, which allows a less-rigorous match. Structure-based alignment is used in order to minimize sequence-based alignment errors, so that template structure generation errors and sequence alignment based errors are not conflated. The goal of this work was to see whether generation of alternate structures using the coarse-grained model could in principle generate better template structures than the initial PDB structure. The problems of correct sequence alignment and blind identification of a superior template are different issues. In particular, the latter can only be sensibly addressed if it is known that better template structures exist in the pool of candidates in the first place.

Model generation

After structure alignment, an all-atom model was produced from a backbone template in three different ways.

1. The initial PDB-derived backbone templates were used to generate all-atom models by using Modeller[37] without any further backbone adjustment or structure manipulation. This is the reference model.
2. Generation of alternative template structures, followed by all-atom modeling. From each template structure, an ensemble of protein backbones is generated by using the 3K-ENM. The backbones were generated by following either (a) the lowest normal mode, or (b) the three lowest normal modes, or (c) the six lowest normal modes. The best template backbone is selected based on the rms distance to the target structure. Side chains are repacked in by SCWRL. Then Modeller builds the full atom models based on the rebuilt chosen template.
3. Full-atom structures are generated by using Modeller. The lowest energy structure was selected, and from it an ensemble of protein backbones is generated by using the 3K-ENM. The side chains are then repacked by SCWRL using the Dunbrack backbone-dependent side chain rotamer library.

For each of the Modeller runs in the protocols above, five output all-atom structures were requested, and the one with the lowest Modeller spatial restrain energy was selected for further steps. Protocols 2 and 3 differ in that in Protocol 2, the alternate structures of the aligned backbone sections are produced first, and then Modeller builds any required loops (for insertions), or anneals deletion regions for each alternate backbone structure. In Protocol 3, insertions/deletions are dealt with by Modeller first using the original PDB template, and then alternate backbone conformations are generated from the entire backbone (aligned segments plus built-in loops).

All-atom model structures produced by Protocols 1–3 were either evaluated as is, or subject to a short minimization to relieve steric clashes and locally optimize torsion angles. The minimization was carried out using the CHARMm (c27b2) force field with 400 ABNR steps with all of Ca atoms fixed in order to keep the generating framework unchanged.

Model evaluation

Structures were compared by computing the root mean square difference (rmsd) of two protein backbones. All-atom structures were evaluated with the DFIRE statistical protein potential[38] before and after minimization.

RESULTS AND DISCUSSION

Force constant parameterization

The 3K-ENM was parameterized against X-ray B-factors. The reason we chose B-factors was that in high quality X-ray structures, they are linearly related to the fluctuation in atomic position. The fluctuation arises from internal fluctuations, with additional contributions arising from a combination of lattice disorder and rigid body motion in the crystal.[39] In high resolution crystal structures, contributions from lattice disorder to Ca B-factors is usually small. The internal fluctuation contributions to B-factors in the ENM can be easily calculated from the normal modes using Eq. (5). The correlation coefficient between the experimental B-factors and internal fluctuation contributions from the ENM is independent of the absolute force constant values, which acts as a uniform scaling factor. The assumption behind our parameterization was that a model that fit the B-factors better would provide a more realistic model of protein

Table I
B-factor Correlation for the Training Set of Proteins Using Single and Three Force Parameter Elastic Network Models

		B-factor correlation coefficient			Final B-factor correlation coefficient
PDB ID	No. of residues	1K-ENM	3K-ENM	Improvement (%)	3K-ENM + rigid motion
1dk0a	173	0.49	0.55	13	0.81
1f8ea	388	0.64	0.71	10	0.88
1k2xa	155	0.53	0.57	9	0.85
1kqpa	271	0.30	0.29	−5	0.78
1lko0	190	0.57	0.55	−3	0.94
1mml0	251	0.76	0.85	12	0.83
1mnna	290	0.48	0.57	18	0.81
1muwa	388	0.31	0.35	14	0.69
1mym0	154	0.48	0.68	42	0.82
1o3u0	119	0.38	0.39	1	0.68
1o9ra	162	0.62	0.61	−1	0.80
1od30	131	0.58	0.58	1	0.80
1ouwa	148	0.38	0.38	0	0.79
1pgs0	311	0.66	0.67	1	0.79
1px5a	327	0.63	0.61	−4	0.82
1qwg0	251	0.28	0.25	−12	0.89
1tn6a	315	0.18	0.24	35	0.75
1w0n0	120	0.48	0.48	0	0.88
3nul0	130	0.55	0.50	−9	0.83
Average		0.49	0.52	6	0.81

fluctuations overall, and this expectation is borne out by the subsequent results.

Table I shows the results of the original single parameter 1K-ENM and modified 3K-ENM on the training set of 19 proteins. For this protein set, the correlation coefficients (R) are in the range 0.2–0.9. We find that three proteins have very low R, <0.3, and upon further examination of their structure, we find that they all have one or more very flexible domains, usually at the N or C termini. Those flexible domains typically contact or interact with ligands or other protein chains in the crystal, and these contacts are not included in the normal mode analysis. These domain contact and interactions may help stabilize the structure in the crystal, but in ENM, which models the isolated protein, the fluctuation of those extended terminus or loops are overestimated. For the three low-correlated proteins in the training set, the average improvement is 0.005, from 0.253 to 0.258. Since the correlation coefficient is dominated by the residues that are located in those flexible regions and the fluctuations of those residues are not adequately characterized in the ENM, the improvement of B-factor correlation coefficient is not significant. For the other training set proteins, correlation coefficients of 12 proteins were significantly improved. The average improvement of these is 0.045, from 0.492 to 0.537. The overall improvement on the training set from optimization of just two force constants K_a and K_b is 6.5%, with the average correlation coefficient increasing from 0.485 to 0.517. We note that even after optimization the overall fit to B-factors is rather low, at 0.517. Similar low values were found in previous parameterizations of the ENM.[40] These low values are expected since the model only accounts for internal fluctuation contributions to B-factors. It is well known that rigid body motions can contribute significantly.[39] Thus after we found the optimal pair of force constants for the entire training data set, we subtracted the internal fluctuation contribution calculated using the optimized 3K-ENM from the experimental B-factors, to yield the "residual" B-factors for each protein. We then fit the residuals B-factors for each protein in the training set using the 10-parameter rigid body Translation/Rotation/Screw (TLS) model of Kuriyan and Weis.[39] The correlation coefficient between experimental B-factors and total calculated B-factors (sum of 3K-ENM and TLS contributions) was then determined (Last column of Table I). This shows excellent agreement with experiment, with an average correlation coefficient of $R = 0.81$, indicating a satisfactory modeling of both internal fluctuation and rigid body motions. After paramterization, the values of the three force constants K_a, K_b, and K_c are 40.0, 20.0, and 0.25 kcal/mol/Å^2, respectively. We note that since we included all the available B-factors in the structures, rather than excluding outliers after normalization,[41] our results may not be directly comparable with this previous B-factor fitting calculation.

After parameterization, we compared the experimental B-factors and the calculated B-factors on 1374 proteins in the test set. The proteins in the test set were chosen with the same criteria as those in the training set. Figure 1(a) shows the distribution of B-factor correlation coefficients from both the original 1K-ENM and the 3K-ENM.

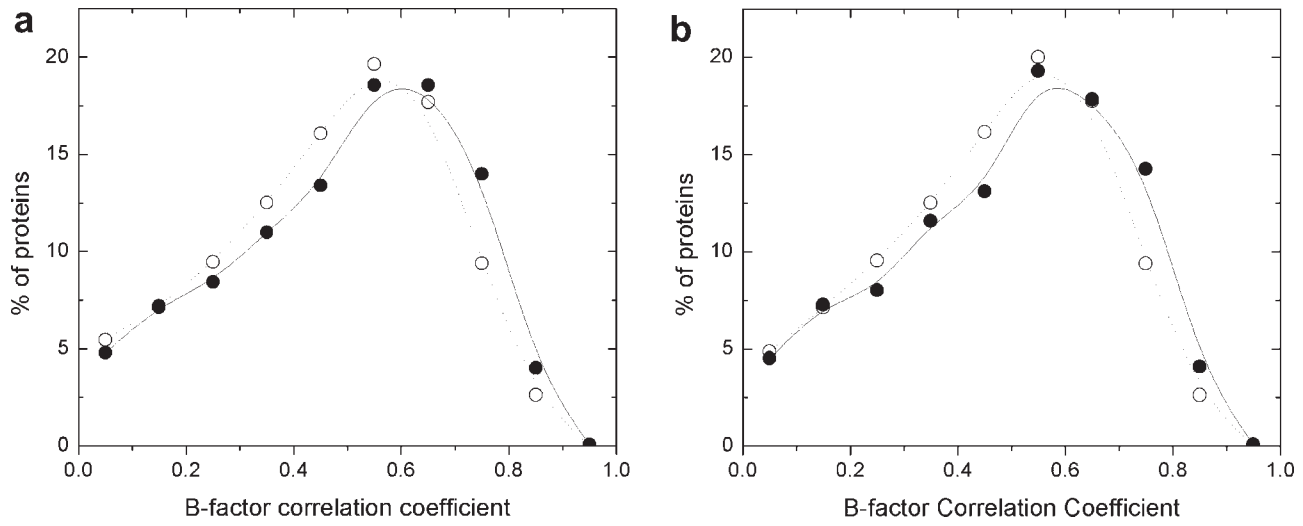

Figure 1

Comparison of distributions of B-factor correlation coefficients computed from the single force parameter ENM (○) and three-force parameter ENM model (●) for the test set of 1347 proteins. The bin size is 0.1. Using a training set of 19 proteins (**a**) and 27 proteins (**b**).

In the test set, very low correlation coefficients ($R < 0.30$) are observed for about 282 proteins. As indicated in the figure, this low correlation set is not much improved because they contain flexible regions whose fluctuations are greatly reduced in the crystal structure. However, for proteins with $R > 0.30$ the figure shows a significant shift to better correlations in the 3K-ENM, indicating general improvement in B-factor prediction. More than 67% of the 1374 proteins show improved B-factor correlations. In the 1K-ENM model, there are 684 proteins with $R > 0.5$, with the average of $R = 0.61$, while in the 3K-ENm 771 proteins show $R > 0.5$ with an average value of $R = 0.65$. Over the entire test set of 1374 proteins, the correlation coefficient is improved from 0.467 to 0.493, which is about a 6% increase. To illustrate the kind of agreement these R-values represent, Figure 2(a) shows a comparison of the B-factor values from the single force parameter and multiparameter

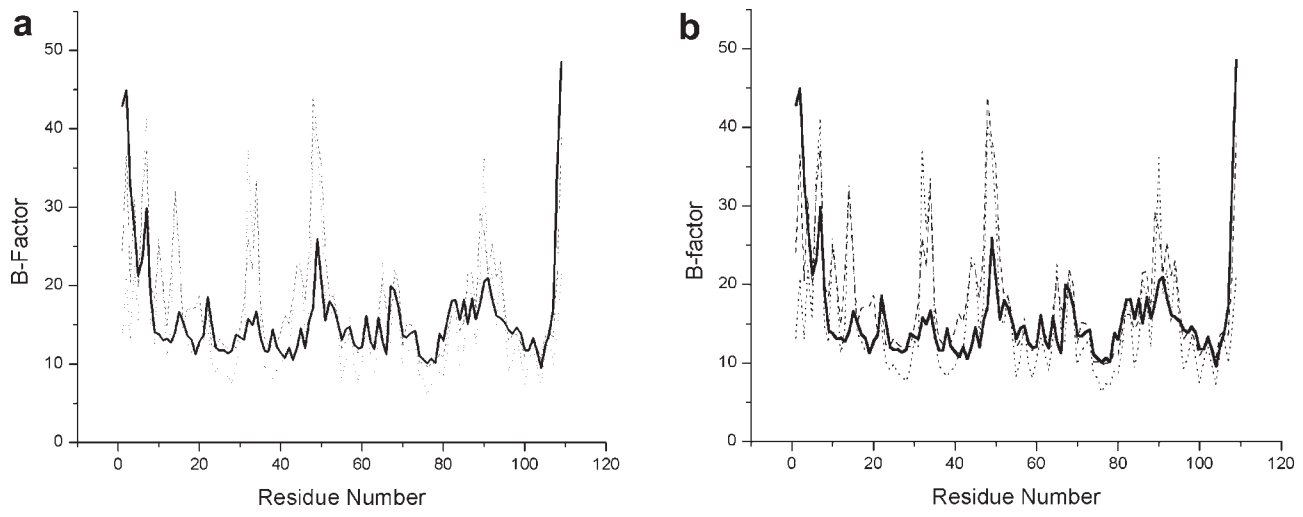

Figure 2

Comparison of B-factors from X-ray diffraction (−) and calculated from single force-parameter (. . .) and 3-force parameter ENM (---) models for RNase protein, PDB ID 1NZ0. Model derived from a training set of 19 proteins (**a**) and 27 proteins (**b**).

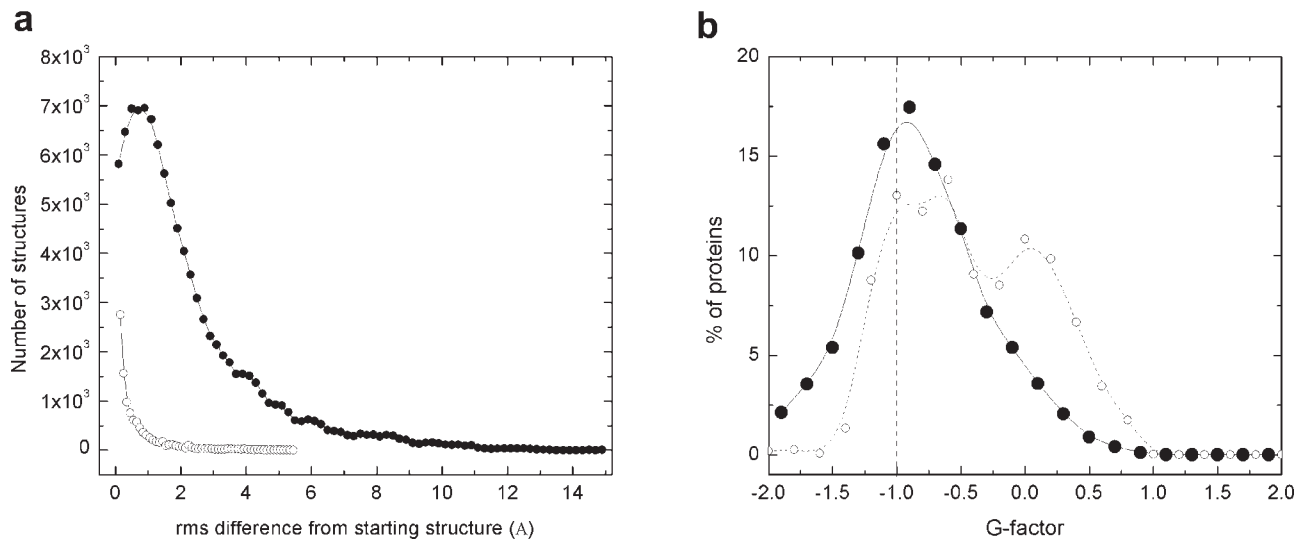

Figure 3

(**a**) Distribution of root mean square (rms) differences from the original structure for alternate protein backbones. The rms is calculated for Ca atoms only. (**b**) Distribution of ProCheck torsion angle G-factors for alternate protein backbones. Dotted line at −1 indicates cutoff for good quality structures. Single force parameter ENM (○), three-force parameter ENM model (●). Structures were generated from 148 proteins. The one and three parameter models produced a total of 10,571 and 102,051 alternate backbones respectively. The bin size is 0.2 Å.

ENM to the experimental data for the protein 1NZ0. This 109 amino acid long protein has three α-helices and one β-sheet. The B-factors range from 4.9 to 48.5. The 1K-ENM shows a correlation coefficient of $R = 0.389$ while the 3K-ENM has a much better fit with $R = 0.631$, with a lot of improvement coming from better modeling of the flexibility in the more rigid regions. This figure illustrates the key result of our parameterization, an R value of 0.6 or better implies the pattern of high and low flexibility within a protein is captured in significant detail.

Backbone stereochemical quality

To examine the ability of the 3K-ENM to generate good stereochemical quality backbones, from the subset of the 1374 proteins that had $R > 0.5$ we randomly chose 148 proteins. We used a cutoff of $R = 0.5$ as we expect the ENM model to work reliably for proteins where it predicts B-factor profiles well. Following the stepwise backbone distortion/building strategy, a total of 102,051 structures were generated from these 148 starting structures, by choosing distortions along the first six nonzero mode directions. Generation of an ensemble of structures from a starting PDB structure typically took 1–10 s of Cpu time on a 2-GHz workstation. Figure 3(a) shows the distribution of backbones frames according to the rms deviation of the Ca's from the starting PDB structures. 93.6% of the 102,051 frames lie in the conformational space within 6 Å. 6 Å rms is very reasonable space for protein design and comparative modeling, since in both applications it is rare that one encounters a final structure that is more than 6 Å rms away from the starting template or model. However, the method is also able to generate hundreds of backbone structures up to ∼13 Å rms deviation away from the starting conformation. Although most of the proteins can be pushed to 6 Å rms, the ability to generate larger deviations (>10 Å) without violating stereo-chemical constraints depends on the "tolerance" of the starting protein. Figure 4 shows some generated backbones for one such tolerant case, flagellar transcriptional activator FlhD protein, PDB ID 1g8e. The backbones are 5 Å (B), 10 Å (C), and 13 Å (D) rms away from the starting structure (A).

To show the improvement given by the 3K-ENM, we also generated structures using the single force 1K-ENM. In this model, there are only 10,571 structures generated, which is about 10% of the number of structures generated by the 3K-ENM. Also 93% of these 10,571 structures are within 2.0 Å of the starting structures. The single force parameter model thus cannot push the structure to deform as much as the three force parameter can within the limits of good geometry.

In our implementation of the ENM, the full backbone is generated from the Ca positions using standard peptide bond lengths and angles from the PROCHECK program. Any Ca frame that has an improper Ca i, $i + 1$ distance or an improper Ca $i - 1$, i, $i + 1$ angle is discarded. The dihedral angles, phi and psi, are based on the orientation of the peptide plane in the starting struc-

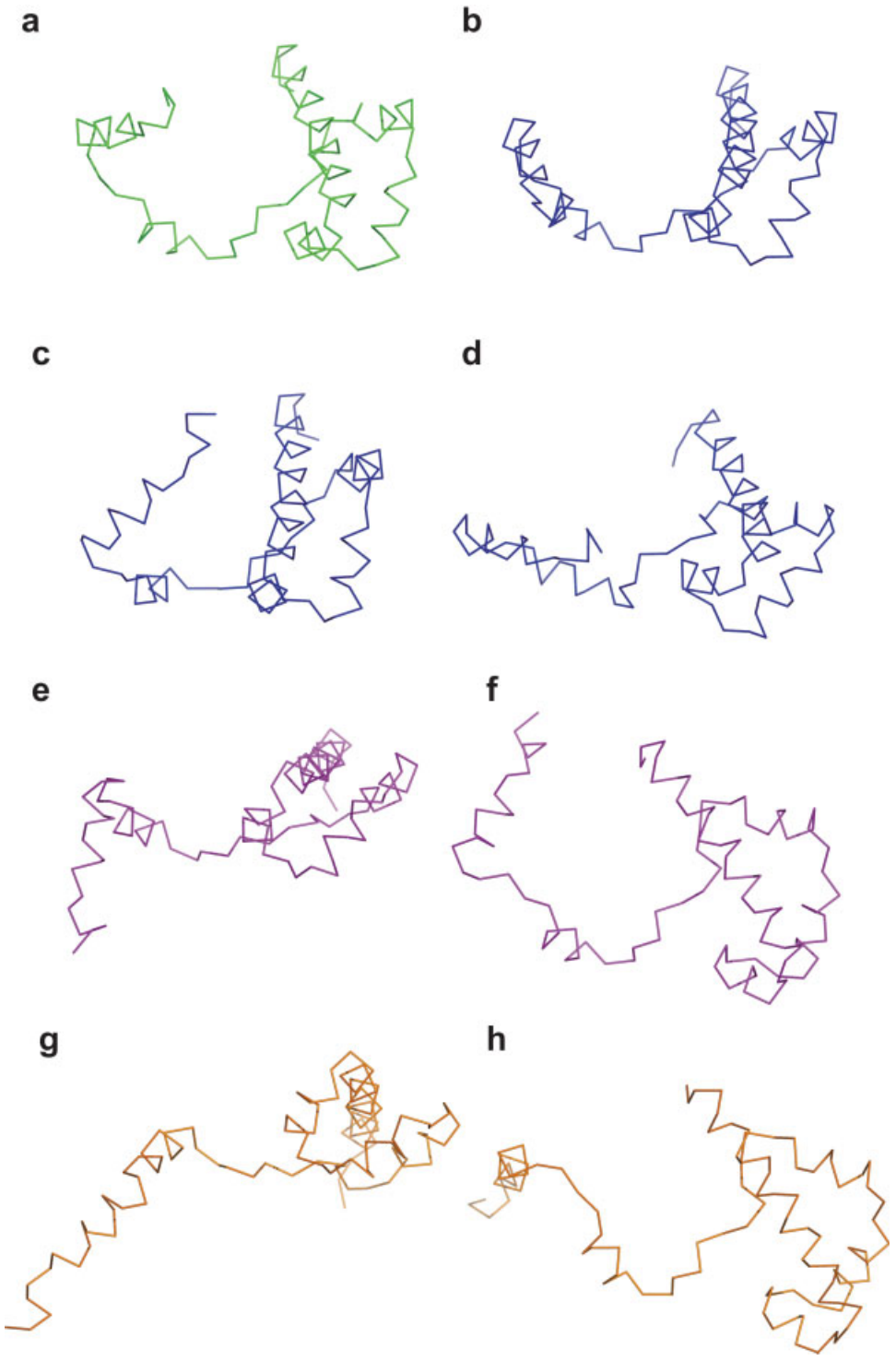

Figure 4

Backbone traces of flagellar transcriptional activator FlhD, PDB ID 1G8E. X-ray structure: (**a**). Alternate backbones with rms deviations from X-ray of 5 Å (**b**–**d**), 10 Å (**e**,**f**), and 13 Å (**g**,**h**) respectively.

Table II
Effect of Data Set Size on Parameterization

No. of pdb structures used	19	27	54	80	115	157
% improvement in B-factor fits (average)	6.50	6.4	5.1	5.6	5.6	5.1
Optimal second, third force constants[a]	20, 0.25	6.0, 0.4	6.0, 0.6	6.0, 0.6	6.0, 0.6	8.0, 0.4
Tolerance in second, third force constants[b]	8, 0.2	2, 0.2	7, 0.2	8, 0.2	8, 0.2	8, 0.2

[a]With a fixed first constant of 40.
[b]Range of values about optimum that produce a fit within 15% of optimal.

ture followed by the minimal degree of rotation to meet the N—Ca—C bond angle constraint. Although this produces stereochemically correct backbones, it does not necessarily produce optimal dihedral angles. We thus ran all of the generated 102,051 protein backbones through the PROCHECK program with a nominal X-ray resolution of 2.0 Å to check the quality of the torsion angles. The resulting torsion G-factor computed in PROCHECK thus provides a quality measure of just of phi, psi backbone torsion angles. Figure 3(b) shows the resulting distribution of G-factors. The G-factor is essentially a log-odds score based on the observed distribution of the stereochemical parameter. The average G-factor is −0.79 and 68% of the protein structures we generate have G-factor larger than −1.0 (i.e., sterochemistry typical of >90% of protein structures). It is important to emphasize that this means that the majority of backbone variants are of sufficient quality to build side chains. This is a key requirement in being able to go from a coarse-grained model to a fine grained, all-atom model of an alternate structure.

We also observed that when the generated backbones are close to the starting conformation, for example, rms deviations <3 Å, the torsion angles computed in this methodology are mostly in allowable range. When the protein backbone shows larger rms deviations, it becomes more difficult to find the appropriate torsion angles for the structure without applying additional optimization methods. The distribution of the G-factor of the 10,571 protein backbones from the 1K-ENM is slightly better than that from 3K-ENM model: The average G-factor is −0.41 and 74% of backbones have a G-factor greater than −1.0. This merely reflects the fact that the single force parameter model cannot produce very large rms deviations.

Effect of parameter set size and force constants

The original parameterization was performed against a rather small training set of proteins, 19. Nevertheless, the model was able to better fit the B-factor data on the much larger test set of 1374 proteins, and to produce an order of magnitude larger number of good conformations than the original single force constant ENM (see Fig. 3). To judge whether this training set was adequate, and whether further improvement could be achieved, we reparameterized the 3K-ENM on successively larger training sets of 27, 54, 80, 115, and 157 proteins, again selected using the PISCES database for high quality and low mutual sequence homology. The results are summarized in Table II. Increasing to 27 proteins, the secondary (K_b) and tertiary (K_c) force constants change somewhat, the former decreasing from 20 to 6, the latter increasing from 0.25 to 0.4 [Expressed here relative to a fixed primary (K_a) force constant of 40]. The tolerance of the fit was estimated by finding the range allowed for the two fitted force constants that produced fits within 15% of the optimal fit. For the larger sets, the relative force constant magnitudes hardly change within the tolerance of the fit, indicating insensitivity to training set size. However, the ability to fit B-factors is not changed significantly upon moving to 27 proteins [Comparing Figs. 1(b) and 2(b)], and in fact it declines somewhat upon further increases in training set size. The ability of the 27-protein set model to fit individual protein B-factor profiles was almost indistinguishable from that of the 19 protein set [Figs. 2(a,b)]. Similarly, its ability of generating quality alternate structures was almost indistinguishable from the 19-protein set shown in Figure 3 (data not shown). Thus, for further homology modeling applications we chose the set with force constants of 40, 6, 0.4 from the 27 protein data, as large enough to have "converged" in training set size, but close to best in B-factor fitting.

Regardless of training set size, the relative strengths of the three force constants are similar: The primary structure constant is about one order of magnitude larger than the secondary constant, and two orders of magnitude larger than the tertiary constant. This reflects the relative strengths of these interactions. Typical bond length and angle force constants in all atom force fields are 10–100's kcal/mol/Å, while nonbond interactions typically involve energies about 1 kcal/mol. The low value of tertiary constants does not mean that this term is unimportant, however. First, there are many more tertiary interactions within the cutoff distance. Second, precisely because they are weaker, displacements are dominated by motions that distort the third class of spring, while the much stiffer primary and secondary constants allow little displacement along those coordinates, so acting as effective restraints for good geometry and stereochemistry, as intended. Kondrashov *et al.*[6] previously examined a two-parameter spring constant model, and found an optimal ratio of 10:1 between the force constants for bonded and

Table III
Structures Generated From Starting PDB Structure Using 3K-ENM by Selecting From One, Three, or Six Modes

				One mode[b]		Three modes[c]		Six modes[d]	
No.	pdb	NumCa[a]	B-factor correlation coefficient	rmsd[e]	NumFrame	rmsd	NumFrame	rmsd	NumFrame
1	1dt7	92	n/a	3.2	40	5.2	182	7.0	459
2	1jwd	90	n/a	4.1	88	4.1	248	4.4	588
3	1awi	138	0.69	0.7	11	1.2	34	1.5	149
4	1cqa	123	0.17	1.2	21	1.7	78	1.7	198
5	1c0b	124	0.73	1.8	34	2.1	131	2.1	275
6	1km9	110	0.36	1.6	50	2.0	169	2.3	345
7	1b1c	166	0.71	0.8	13	1.5	98	1.7	236
8	2fcr	173	n/a	1.0	16	1.3	86	1.5	226
9	1gyb	122	0.50	0.6	10	1.8	100	1.6	183
10	1jkg	139	0.74	1.2	14	2.6	76	2.8	248
11	1qwe	56	0.54	0.8	12	1.4	82	1.0	129
12	2sem	58	0.71	0.6	10	0.9	51	1.2	162
13	1i92	91	0.29	0.3	2	1.0	73	1.5	190
14	1qlc	95	n/a	0.7	6	1.0	20	1.3	140
15	1a81[f]	100	0.40	1.0	17	1.5	116	1.5	197
16	1csy	112	n/a	0.4	4	0.5	22	1.3	95
17	1awe	130	n/a	0.6	10	1.5	108	1.6	239
18	1dro	122	n/a	0.5	13	0.6	38	1.1	101
19	1ak6	174	0.84	0.7	11	1.4	64	1.7	122
20	1f7s	124	0.69	0.9	14	0.9	64	1.5	228
21	1alb	131	0.36	0.9	15	1.0	72	1.1	168
22	1o1u	127	n/a	1.1	19	1.1	83	1.3	184
23	1eqt	67	0.57	1.7	17	2.6	99	2.7	203
24	1je4	69	0.91	3.0	104	3.1	61	3.1	304
25	1t7p	105	0.27	1.0	17	0.9	57	1.2	205
26	2trx	107	0.72	0.6	6	0.6	6	0.6	6
27	1gqm	87	0.32	2.8	41	4.5	234	5.0	403
28	1mr8	90	0.76	2.3	27	2.7	123	3.5	381
29	1ahr[f]	72	0.39	3.6	79	3.7	242	3.8	505
30	1m8q[f]	70	n/a	4.1	116	4.1	271	4.1	662
31	1ihj	94	0.62	1.0	17	0.9	59	1.1	146
32	1kef	93	0.54	0.4	3	1.7	22	1.1	118

n/a, the B-factor is not available in pdb.
[a]The number of carbon-α atoms whose coordinates are available and used.
[b]The structures are generated by following the lowest mode.
[c]The structures are generated by following the first three lowest modes.
[d]The structures are generated by following the first six lowest modes.
[e]The largest rmsd between the generated and starting structures.
[f]Only the aligned region is used as the starting structure.

nonbonded interactions. This smaller ratio relative to our primary/tertiary ratio probably results from their use of different cutoff criteria and a single force constant for both secondary and tertiary interactions. Their model effectively picks a compromise value for their second force constant somewhere between our secondary and tertiary values. Also that model was not optimized to preserve good backbone geometry and stereochemistry as was the 3K-ENM here.

Homology modeling

To assess the ability of the 3K-ENM to provide good templates for homology modeling, we build homology models for 16 query sequences. For each query structure, two template structures were chosen based on the sequence homology, and aligned using structural alignment.[36] Two different modeling protocols were used with the 3K-ENM. In the first, (Protocol 2 of methods) the 3K-ENM was used to generate ensembles of alternate backbones of each template. Ensembles of backbones with different conformation are generated by following either the lowest one, three or six modes. Table III shows that as expected, more stereochemically correct backbones can be generated using six modes than three modes. It also shows that by following the six lowest modes, larger conformational variations can be achieved. From Table IV and Figure 5, we see that in all 32 template cases, the 3K-ENM model generates many alternative templates that are closer to the target than the original PDB template structures, in terms of the rmsd of the aligned carbon α atoms.

Table IV shows that one can generate improved template backbones using either the lowest three or six

Table IV
Alternative Template Models Generated Using 3K-ENM by Selecting From Three or Six Modes

						Three modes			Six modes	
Target	Template	Sequence identity (%)[a]	Aligned size[a]	Original rms[b]	Total no. of frames	No. of frames with better rmsd	Best rms[b]	Total no. of frames	No. of frames with better rmsd	Best rms[b]
1a4p	1dt7	39	80	2.4	182	35	2.2	459	147	2.2
	1jwd	31	80	2.2	248	22	2.2	588	56	2.2
1acf	1awi	21	121	1.8	34	15	1.7	149	36	1.7
	1cqa	37	105	1.6	78	15	1.6	198	58	1.5
1agi	1c0b	37	110	1.5	131	36	1.5	275	48	1.5
	1km9	32	99	2.0	169	8	1.9	345	17	1.9
1ahn	1b1c	21	145	2.5	98	27	2.3	236	88	2.3
	2fcr	37	165	1.5	86	3	1.5	226	13	1.5
1ar0	1gyb	42	117	1.1	100	30	0.9	183	29	0.9
	1jkg	25	116	1.2	76	53	1.2	248	120	1.1
1b07	1qwe	40	54	1.2	82	24	1.0	129	22	1.0
	2sem	37	56	1.0	51	6	0.9	162	42	0.8
1be9	1i92	26	83	2.0	73	11	2.0	190	14	2.0
	1qlc	38	84	2.1	20	1	2.1	140	57	2.1
1bmb	1a81	31	93	2.0	116	16	1.9	197	46	1.9
	1csy	28	91	2.4	22	6	2.4	95	46	2.3
1btn	1awe	27	87	2.8	108	89	2.6	239	104	2.6
	1dro	38	102	3.2	38	8	3.2	101	16	3.2
1cfy	1ak6	35	130	2.3	64	13	2.3	122	39	2.2
	1f7s	39	116	1.3	64	27	1.2	228	70	1.2
1crb	1alb	37	128	1.7	72	10	1.6	168	24	1.6
	1o1u	26	124	1.8	83	28	1.8	184	44	1.7
1dol	1eqt	25	63	1.1	99	33	1.0	203	44	1.0
	1je4	34	57	2.2	61	1	2.2	304	61	2.2
1erv	1t7p	25	102	1.2	57	0	1.2	205	11	1.2
	2trx	25	102	1.2	6	1	1.2	6	1	1.2
1ig5	1gqm	39	70	2.2	234	23	2.1	403	53	2.1
	1mr8	36	69	2.2	123	59	2.1	381	176	2.1
1pva	1ahr	35	68	2.3	242	17	2.3	505	40	2.2
	1m8q	26	63	2.0	271	23	2.0	662	11	2.0
1qav	1ihj	30	80	1.3	59	10	1.3	146	55	1.3
	1kef	43	82	2.1	22	1	2.1	118	59	1.9

[a]The number is based on structural alignment.
[b]The rmsd is calculated for all the aligned atoms.

modes but, as Figure 5 shows, the best template using six normal modes is significantly better than using the three modes, because a choice of six modes allows more choice for the structure to deform, therefore covering a larger conformational space. We have found (results not shown) that using more than six modes does not improve template generation significantly.

Thus, this methodology is able to generate better templates for all the test proteins. To see whether this would result in a better complete homology model, the best alternative template was chosen based on the smallest rmsd of the aligned C α atoms between target and query. Side chains of aligned regions of the chosen template were built by SCWRL and then the full-atom model was built in Modeller as described in methods, Protocol 2. Here, we excluded 1PVA:1AHR and 1PVA:1M8Q queries, because the percentage of the aligned Cα's according to CE is below 60%. Modeller produces five alternate final models, and the lowest energy all-atom model was chosen for analysis. In addition, the origin PDB template was input to Modeller, and the best all-atom model output was used as a reference (Protocol 1 of methods). Table V and Figure 6(a) show the comparison of the models built from the original template and the alternative template. In all the test cases, we observed the following: (a) For those aligned carbon α atoms, the rmsd between the Modeller model and the target structure is very close to the initial difference between the template and target structures no matter whether the template is chosen from the original pdb structure or from the generated alternative candidates. This means that, in building the full atom model, the program Modeller retains the starting template structure as much as possible. Therefore, if the template structure is closer to the target structure, in the output model, those aligned carbon α atoms, at least, are closer to the target. (b) In terms of the rmsd comparison of all carbon α atoms, the models built from generated alternative templates do not always show significant improvement, as shown in Figure 6(a). By examining the cases that do not show improvement, we found that typi-

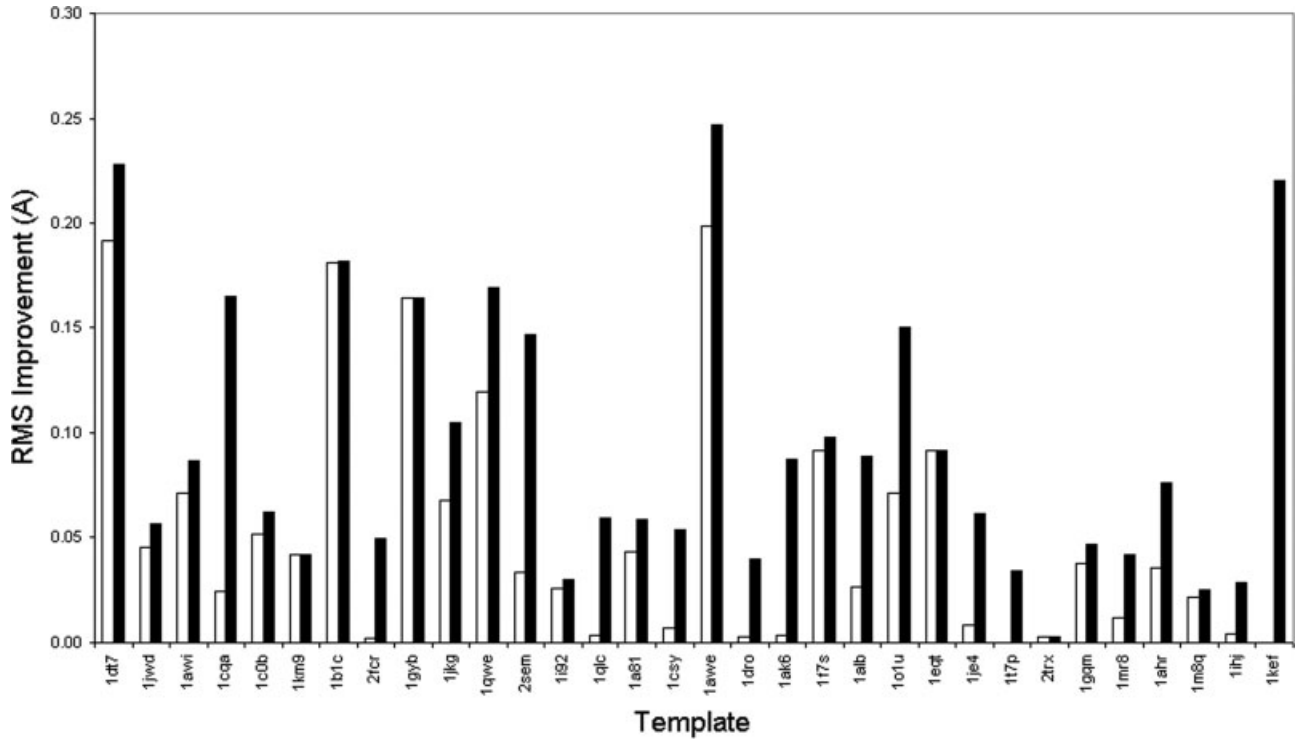

Figure 5

Template improvement by three parameter ENM, choosing from three (□) or six (■) lowest modes.

cally the loops which are inserted between the alternate template fragments are not in an appropriate position. This leads to a large rmsd for the entire structure. From this we concluded that in order to improve the entire all-atom model, it is not enough to generate alternative templates or optimized templates for the aligned region alone. Correctly, constructing the fragments (insertions of loops) that are not in the template structure or closing gaps from deletion is also crucial.

To address this, we used another modeling strategy (Protocol 3 of Methods), in which a Modeller all-atom model is first produced from the original unperturbed PDBD template. From this the side chains are discarded, and complete backbone (with insertions and deletions) is used to generate the Ca network for the 3K-ENM, and then alternate backbones are generated, followed by side chain rebuilding with SCWRL. In this case, as Table VI and Figure 6(b) show, in all of the 30 test cases a better final all-atom model is built from one of the alternative templates. This is true even for difficult cases like 1I92, and 1QLC, which were not originally well-built due to the lack of appropriate template. Figure 6(b) also shows an interesting point when comparing the improvement for aligned fragments versus all Ca atoms. For those well-predicted models, (rmsd to the target $<$3.0 Å), the conformational change produced by the 3K-ENM could provide moderate improvement, and most of the improvement comes from the aligned fragments. In contrast, for those less well-predicted models, (rmsd to target structure $>$3.0 Å), the rmsd improvement resulting from 3K-ENM is larger and mostly due to the conformational changes in the inserted loops or extended terminals rather than the changes in aligned fragments. A summary of model improvement versus initial template error is shown in Figure 7. Query sequences are ordered by increasing error of the initial PDB template. Improvement in the model is seen in all cases, in some cases quite modest, usually for queries with small initial template error. Cases with larger initial template error also show larger improvement using the 3K-ENM, in some cases up to 1 Å or more, a dramatic improvement. Of course initial templates with larger errors permit potentially larger improvements, but there is no *a priori* guarantee that one can realize them with a given method. In fact, the results show that the 3K-ENM template generation method is capable of providing better models regardless of the quality of the starting template, especially improving the poorer ones.

Following Misura *et al.*,[33] we tested the template improvement method on known structures, using opti-

Table V
Comparison of Models Built From the Original Fixed (PDB) Template and Templates Generated From Original Template Using 3K-ENM[a]

				rms Error			
Target	NumRes[a]	Template	NumRes[b]	All Ca (PDB)	All Ca 3K-ENM	Aligned Ca (PDB)	Aligned Ca 3K-ENM
1a4p	92	1dt7	92	3.4	3.6	2.2	2.1
		1jwd	90	3.0	3.3	2.3	2.5
1acf	125	1awi	138	2.0	1.9	1.7	1.7
		1cqa	123	2.4	2.2	1.6	1.4
1agi	125	1c0b	124	2.8	2.8	1.5	1.4
		1km9	110	4.8	4.8	1.9	1.9
1ahn	169	1b1c	166	4.9	5.4	2.4	2.4
		2fcr	173	1.6	1.5	1.4	1.4
1ar0	125	1gyb	122	2.0	2.0	1.0	0.9
		1jkg	139	2.2	2.2	1.2	1.1
1b07	58	1qwe	56	1.7	1.4	1.2	1.0
		2sem	58	1.0	1.0	1.0	1.0
1be9	115	1i92	91	9.9	10.0	2.1	2.1
		1qlc	95	12.7	12.7	2.1	2.1
1bmb	98	1a81	100	2.9	2.8	2.6	2.5
		1csy	112	2.8	2.8	2.3	2.3
1btn	106	1awe	130	4.7	4.1	2.8	2.9
		1dro	122	3.2	3.3	3.2	3.1
1cfy	133	1ak6	174	2.2	2.1	2.0	2.0
		1f7s	124	8.6	7.4	1.3	1.3
1crb	134	1alb	131	1.8	1.8	1.7	1.6
		1o1u	127	5.2	5.3	5.1	5.1
1dol	71	1eqt	67	3.8	2.6	1.5	1.1
		1je4	69	6.1	6.5	2.1	2.4
1erv	105	1t7p	105	1.2	1.1	1.2	1.1
		2trx	107	1.2	1.2	1.2	1.2
1ig5	75	1gqm	87	2.4	2.3	2.1	2.1
		1mr8	90	2.5	2.6	2.1	2.1
1qav	90	1ihj	72	3.0	2.6	1.3	1.3
		1kef	70	3.3	2.4	2.0	1.7

[a]Protocol 2 of Methods.
[b]The number of residues whose coordinates are available and used for query, template.

mal (structure-based) sequence alignment. The reason is that we need to establish that the method can in principle give better models, independent of errors in sequence alignment. If not, there is no point in pursuing the method. The results in Tables IV–VI establish that (a) it is relatively easy to produce better templates, even with the less than exhaustive conformational space search used by the 3K-ENM, (b) these can reliably be used to produce, using Protocol 3, a better final all-atom model. A crucial question is whether one could identify the better model in the absence of the query structure, that is, in a real modeling application. To examine this, we used the DFIRE statistical protein scoring potential[38] to evaluate each of the full-atom models generated from alternative 3K-ENM templates. If, in each query set, the structure with the lowest DFIRE potential is closer to the target structure than that produced by the fixed PDB template (judged by the rmsd of all carbon-α atoms), we call that a detectable or realizable improvement. As shown in Table VII, of the 30 tested cases the DFIRE potential can detect 36% of the better models. In other words, for those models with lowest DFIRE potentials, 36% of them are better than the initial model. After applying a short energy minimization to the all-atom models with all the carbon-α atoms fixed, DFIRE's detection rate improves to 50%. That means that half of structures with the lowest DFIRE potentials are better than the initial models. 50%, though not perfect, is still a promising number. It shows that these better models are potentially detectable. There is still a need to develop better scoring functions, or at least investigate different scoring functions, but it seems probable that with improvements in this step almost all better models will be detectable.

CONCLUSIONS

To build stereochemically correct alternate protein conformations rapidly is not a trivial problem. One fast method is to use a coarse-grained model. The challenge then is to generate a fine-grained atomistic model from the coarse-grained conformations. The approach we develop here uses reconstruction from a series of backbone framework structures, which are obtained from a modified ENM by perturbation along low-frequency normal

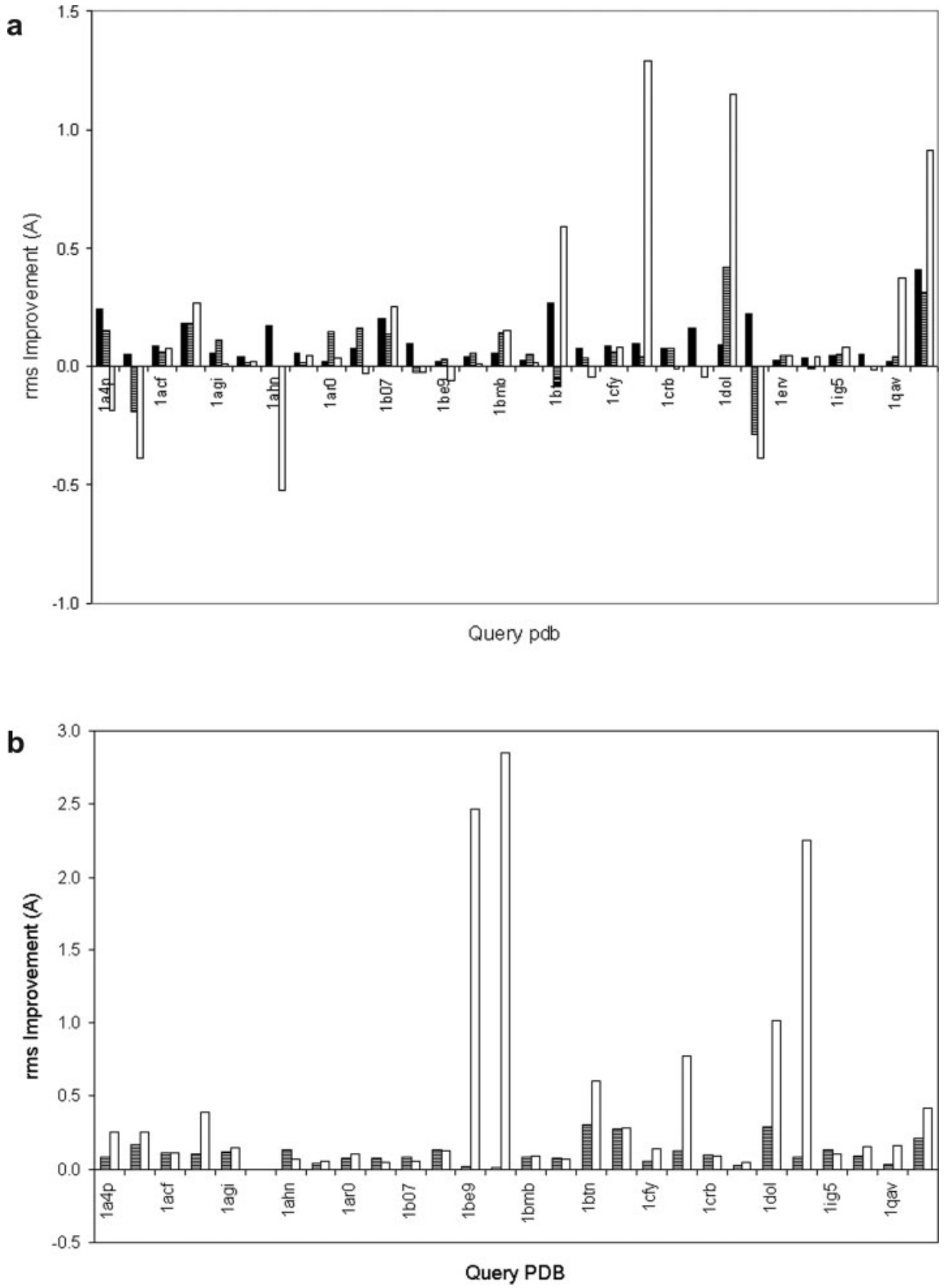

Figure 6

Model improvement using (**a**) Protocol 2, generation of alternate templates followed by building of full model by Modeller. (**b**) Protocol 3, Lowest energy Modeller model used to initiate generation of alternate backbones followed by side chain rebuilding. Template improvement for Protocol 2 (■). Improvement over aligned Ca's (▤). Improvement over all Ca's (□). Alternate structures were generating using 3K-ENM by choosing from six lowest modes.

Table VI
Comparison of Models Built From the Original Fixed (PDB) Template and Templates Generated Using 3K-ENM on a Full Model Built From the Original Template[a]

		rms Error			
Target	Template	All Ca init.	All Ca best	Aligned Ca init.	Aligned Ca best
1a4p	1dt7	3.4	3.1	2.2	2.1
	1jwd	3.0	2.7	2.3	2.2
1acf	1awi	2.0	1.9	1.7	1.6
	1cqa	2.4	2.1	1.6	1.5
1agi	1c0b	2.8	2.7	1.5	1.4
	1km9	4.8	4.8	1.9	1.9
1ahn	1b1c	4.9	4.8	2.4	2.3
	2fcr	1.6	1.5	1.4	1.4
1ar0	1gyb	2.0	1.9	1.0	0.9
	1jkg	2.2	2.2	1.2	1.1
1b07	1qwe	1.7	1.6	1.2	1.1
	2sem	1.0	0.9	1.0	0.8
1be9	1i92	9.9	7.5	2.1	2.1
	1qlc	12.7	9.9	2.1	2.1
1bmb	1a80	2.9	2.8	2.6	2.5
	1csy	2.8	2.7	2.3	2.2
1btn	1awe	4.7	4.1	2.8	2.5
	1dro	3.2	2.9	3.2	2.9
1cfy	1ak6	2.2	2.0	2.0	2.0
	1f7s	8.6	7.9	1.3	1.2
1crb	1alb	1.8	1.7	1.7	1.6
	1o1u	5.2	5.2	5.1	5.1
1dol	1eqt	3.8	2.7	1.5	1.2
	1je4	6.1	3.8	2.1	2.1
1ig5	1gqm	2.4	2.3	2.1	2.0
	1mr8	2.5	2.4	2.1	2.0
1qav	1ihj	3.0	2.8	1.3	1.3
	1kef	3.3	2.9	2.0	1.8

[a]Protocol 3 of Methods.

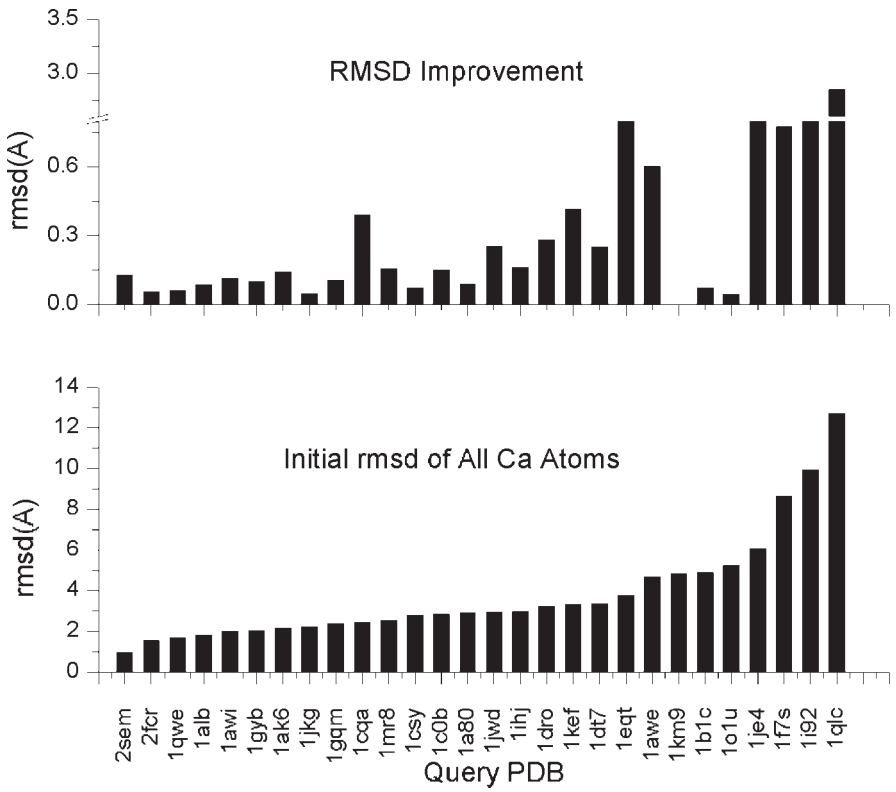

Figure 7
Improvement of model versus initial accuracy of template. Query sequences are ordered by increasing error of initial PDB template (Lower panel). Upper panel shows improvement of final model, assessed over all Ca's.

Table VII
Identification of Improved Models With DFIRE Statistical Potential

			No minimization		After minimization	
Target	Template	Total no. of frames/models	DFIRE score[a]	Detectable?[b]	DFIRE score[a]	Detectable?[b]
1a4p	1dt7	396	3.4	n	3.2	y
	1jwd	200	2.9	y	2.8	y
1acf	1awi	202	2.0	y	2.0	n
	1cqa	196	2.2	y	2.1	y
1agi	1c0b	298	3.0	n	2.8	y
	1km9	17	5.0	n	5.4	n
1ahn	1b1c	184	4.9	n	4.9	n
	2fcr	177	1.7	n	1.8	n
1ar0	1gyb	60	2.0	y	2.0	y
	1jkg	201	2.3	n	2.3	n
1b07	1qwe	158	1.6	y	1.7	y
	2sem	145	1.0	y	1.0	n
1be9	1i92	1258	9.6	y	10.5	n
	1qlc	1165	12.8	n	12.7	y
1bmb	1a80	199	3.3	n	3.0	n
	1csy	156	2.8	n	2.9	n
1btn	1awe	245	4.7	n	4.4	y
	1dro	252	3.2	y	3.1	y
1cfy	1ak6	130	2.2	n	2.1	y
	1f7s	219	9.0	n	8.4	n
1crb	1alb	130	1.9	n	1.8	n
	1o1u	167	5.2	n	5.3	n
1dol	1eqt	236	3.7	y	3.7	y
	1je4	515	6.0	y	5.9	y
1ig5	1gqm	207	2.4	n	2.4	n
	1mr8	191	2.5	y	2.5	y
1qav	1ihj	81	3.0	n	3.0	n
	1kef	126	3.5	n	3.3	y
	Success rate			39%		50%

[a]DFIRE value of model with lowest score.
[b]An improved model is detectable if the one with the lowest DFIRE score is closer to the query than the model built from the original fixed PDB template.

modes. To ensure good quality backbone frameworks, we introduce a carefully parameterized three force constant Ca ENM. Based on physical arguments, we use one force constant to characterize the interaction between the consecutive carbon alphas, which are the stiffest, a second for those within the same secondary structure domain, which is less stiff but still substantial due to a large number of H-bonds, and a third for all remaining (tertiary) interactions within the cutoff distance. After generating alternative Ca coordinates via the ENM normal modes, a parsimonious modeling in of the peptide planes, preserving the relative orientation of the original PDB structure as much as possible, then produces a complete backbone upon which the side chains may be built. This step is done here using the backbone-dependent rotamer library method implemented in SCWRL, since it is fast, flexible, and accurate, but other side chain building methods may be used equally easily. We show that the 3K-ENM model can do a good job of reproducing B-factors, and more importantly, rapidly generate a large number of alternate backbone frameworks with good geometry and stereochemistry, upon which all-atom models can be built. Depending on the protein, the method typically provides conformations up to 6 Å rmsd or more from the starting structure. In a few cases, only small rms deviation models with good geometry are generated. This may reflect something intrinsically "difficult" about that protein fold, or a need to sample more extensively the different modes produced by the 3K-ENM. Aside from filtering out bad geometries, the method does not weigh or rank the conformations, since the criteria for this will vary from application to application. Our aim here was to produce a general method of conformation generation as input for other types of modeling. To demonstrate this we showed an application to improvement of templates for homology modeling.

Homology modeling is a good test of alternative conformation generation, since the putative alternative structures are compared to the experimental X-ray structure of the query sequence. Homology models usually use a fixed template from the aligned regions of a known PDB structure, so that region is inherently limited in final accuracy by the initial error in that PDB template. Using the 3K-ENM model, we showed that in all 30 cases one can easily generate several to many templates which are better than the PDB structure. This ease is perhaps sur-

prising, since the method by no means does a full conformational exploration. For example, we only sample from the six lowest modes, and even here only a depth one tree search is done. A full tree search is not done because of the prohibitive size of the tree, but at this stage a deeper one does not seem warranted for homology modeling. The second significant result from the homology modeling test is that from the better template, a better full atom model could be built in all 30 cases. The key is to first build the complete backbone, that is, insert loops and anneal deletions. The 3K-ENM is then used to generate variants of the complete backbone, followed by side chain building. Generating conformational variants of just the aligned regions first using the 3K-ENM, followed by modeling in of loops was less successful. This is probably because the partial Ca ENM from the aligned fragments allows distortions toward an ensemble of less realistic conformations. Demonstrating that the 3K-ENM can in principle be used to build a better homology model is of course only a start. One also needs to be able to recognize such a model in a real modeling application (i.e., from criteria other than the rms error). We showed here that using the DFIRE statistical scoring potential, this can be done in 50% of the cases. Since better models are present in 100% of the cases, this indicates that the limiting factor is current scoring and energy functions, not the structure generation itself.

We used the homology program Modeller to generate the full backbone from the aligned fragments, and the Dfire scoring potential since both are widely used and among the best programs in their respective areas for accuracy. Again, other programs could be used, since the 3K-ENM approach is quite general. Indeed, as developments in loop generation and protein scoring occur, future directions will be to combine these with the 3K-ENM methodology developed here.

REFERENCES

1. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic level accuracy. Science 2003;302:1364–1368.
2. Desjarlais JR, Handel TM. Side-chain and backbone flexibility in protein core design. J Mol Biol 1999;290:305–318.
3. Ferrara P, Apostolakis J, Caflisch A. Computer simulations of protein folding by targeted molecular dynamics. Proteins 2000;39:252–260.
4. Fu X, Apgar JR, Keating AE. Modeling backbone flexibility to achieve sequence diversity: the design of novel a-helical ligands for Bcl-xL. J Mol Biol 2007;371:1099–1117.
5. Zheng Y, Bailey TL, Teasdale RD. Prediction of protein B-factor profiles. Proteins 2005;58:905–912.
6. Kondrashov D, Cui Q, Phillips G. Optimization and evaluation of a coarse-grained model of protein motion using X-ray crystal data. Biophys J 2006;91:2760–2767.
7. Temiz NA, Meirovitch E, Bahar I. *Escherichia coli* adenylate kinase dynamics: comparison of elastic network model modes with mode-coupling (15)N-NMR relaxation data. Proteins 2004;57:468–480.
8. Tama F, Valle M, Frank J, Brooks CL, III. Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. Proc Natl Acad Sci USA 2003;100:9319–9323.
9. Tama F, Brooks CL, III. The mechanism and pathway of pH induced swelling in cowpea chlorotic mottle virus. J Mol Biol 2002;318:733–747.
10. Tama F, Sanejouand YH. Conformational change of proteins arising from normal mode calculations. Protein Eng 2001;14:1–6.
11. Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 1996;77:1905–1908.
12. Sanders CT, Baker D. Recapitulation of protein family divergence using flexible backbone protein design. J Mol Biol 2005;346:631–644.
13. 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.
14. Delarue M, Sanejouand YH. Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. J Mol Biol 2002;320:1011–1024.
15. Kim MK, Jernigan RL, Chirikjian GS. Efficient generation of feasible pathways for protein conformational transitions. Biophys J 2002;83:1620–1630.
16. Echols N, Milurn D, Gerstein M. MolMovDB: analysis and visualization of conformational change and structural flexibility. Nucleic Acids Res 2003;31:478–482.
17. Goh CS, Milburn D, Gerstein M. Conformational changes associated with protein-protein interactions. Curr Opin Struct Biol 2004;14:104–109.
18. Alexandrov V, Lehnert U, Echols N, Milburn D, Engelman D, Gerstein M. Normal modes for predicting protein motions: a comprehensive database assessment and associated Web tool. Protein Sci 2005;14:633–643.
19. Kim MK, Jernigan RL, Chirikjian GS. Rigid-cluster models of conformational transitions in macromolecular machines and assemblies. Biophys J 2005;89:43–55.
20. 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.
21. Song G, Jernigan RL. An enhanced elastic network model to represent the motions of domain-swapped proteins. Proteins 2006;63:197–209.
22. Yang L, Song G, Jernigan RL. How well can we understand large-scale protein motions using normal modes of elastic network models? Biophys J 2007;93:920–929.
23. Zheng W, Brooks BR, Hummer G. Protein conformational transitions explored by mixed elastic network models. Proteins 2007;69:43–57.
24. Chu JW, Voth GA. Coarse-grained free energy functions for studying protein conformational changes:a double-well network model. Biophys J 2007;93:3860–3871.
25. Kirillova S, Cortés J, Stefaniu A, Siméon T. An NMA-guided path planning approach for computing large-amplitude conformational changes in proteins. Proteins 2008;70:131–143.
26. Canutescu AA, Shelenkov AA, Dunbrack RL, Jr. A graph theory algorithm for protein side-chain prediction. Protein Sci 2003;12:2001–2014.
27. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 2001;80:505–515.
28. Doruker P, Jernigan RL, Bahar I. Dynamics of large proteins through hierarchical levels of coarse-grained structures. J Comput Chem 2002;23:119–127.
29. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.

30. Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical recipes in Fortran 77, 2nd ed. New York, NY: Cambridge University Press; 1992.
31. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291.
32. Wang G, Dunbrack RL, Jr. PISCES. a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
33. Misura KMS, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci USA 2006;103:5361–5366.
34. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D. Automated prediction of CASP-5 structures using the Robetta server. Proteins 2003;53 (Suppl 6):524–533.
35. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
36. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) pf the optimal path. Protein Eng 1998;11:739–747.
37. Fisher A, Sali A. Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 2003;374:461–491.
38. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improve structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726.
39. Kuriyan J, Weis WI. Rigid protein motion as a model for crystallographic temperature factors. Proc Natl Acad Sci USA 1991;88:2773–2777.
40. Kundu S, Melton JS, Sorensen DC, Philips GN. Dynamics of proteins in crystals: comparison of experiment with simple models. Biophys J 2002;83:723–732.
41. Yuan Z, Bailey TL, Teasdale RD. Prediction of protein B-factor profiles. Proteins 2005;58:905–912.