PROTEINS: Structure, Function, and Genetics 47:393–402 (2002)

Increasing the Precision of Comparative Models with YASARA NOVA—a Self-Parameterizing Force Field

Elmar Krieger,[1] Günther Koraimann,[2] and Gert Vriend[1]
[1]*Center for Molecular and Biomolecular Informatics (CMBI), Nijmegen, The Netherlands*
[2]*Institute of Molecular Biology, Biochemistry, and Microbiology, Graz, Austria*

ABSTRACT **One of the conclusions drawn at the CASP4 meeting in Asilomar was that applying various force fields during refinement of template-based models tends to move predictions in the wrong direction, away from the experimentally determined coordinates. We have derived an all-atom force field aimed at protein and nucleotide optimization in vacuo (NOVA), which has been specifically designed to avoid this problem. NOVA resembles common molecular dynamics force fields but has been automatically parameterized with two major goals: (i) not to make high resolution X-ray structures worse and (ii) to improve homology models built by WHAT IF. Force-field parameters were not required to be physically correct; instead, they were optimized with random Monte Carlo moves in force-field parameter space, each one evaluated by simulated annealing runs of a 50-protein optimization set. Errors inherent to the approximate force-field equation could thus be canceled by errors in force-field parameters. Compared with the optimization set, the force field did equally well on an independent validation set and is shown to move in silico models closer to reality. It can be applied to modeling applications as well as X-ray and NMR structure refinement. A new method to assign force-field parameters based on molecular trees is also presented. A NOVA server is freely accessible at http://www.yasara.com/servers Proteins 2002;47:393–402.**

Key words: protein modeling; structure refinement; force-field parameter optimization; WHAT IF

INTRODUCTION

The search for Nature's folding function has been a tempting scientific adventure ever since Linus Pauling predicted the α-helix in 1951.[1] Because a precise quantum chemical calculation of the true energy function is still hardly feasible for macromolecules, one works with approximations, such as the AMBER,[2] CHARMM,[3] or GROMOS[4] molecular dynamics force fields.

When developing a new force field, the first step is to set up a general equation that matches the various forces present in the studied system. Then one defines rules to derive force-field parameters from quantum chemical calculations or experimental measurements on (usually) small molecules. Here we took a different approach and just defined three goals:

1. For every global (or lowest accessible local) minimum of the true conformational free energy, a minimum of NOVA should lie close by.
2. The regions around the minima of NOVA need to be as smooth as possible and thereby facilitate energy minimization algorithms.
3. NOVA should be a function of solute atom coordinates only; the solvent must thus be implicitly included. (See Roux and Simonson[5] for a recent review of implicit solvent models).

The force field was allowed to parameterize itself while trying to optimally fulfill these goals. This was achieved by randomly changing force-field parameters and evaluating the "fitness" of the resulting force field with a protocol step by step matching the three goals defined above:

1. Energy minimization of high resolution X-ray structures. The smaller the root-mean-square deviation (RMSD) from the initial structure, the closer are the NOVA minima to reality.
2. Energy minimization of homology models built for high-resolution X-ray structures. The smaller the RMSD to the experimental structure, the better is the energy landscape suited for getting there. (Other methods for smoothing and reducing the height of energy barriers include umbrella sampling and soft-core potentials[6–8]).
3. All energy minimizations are done as in vacuo.

To make such a global search in force-field parameter space computationally feasible, the number of optimized parameters had to be kept small. Precisely known parameters (i.e., equilibrium bond lengths and angles) were not optimized. Well-known parameters (i.e., bond stretching and angle-bending force constants) came from the AMBER force field and were rescaled together by using two scaling

This article contains Supplementary Material that can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/2002/47_3/v47.393.html.

*Correspondence to: Elmar Krieger, CMBI, Toernooiveld 1, NL-6525 Nijmegen, The Netherlands. E-mail: elmar.krieger@cmbi.kun.nl

Received 7 September 2001; Accepted 8 January 2002

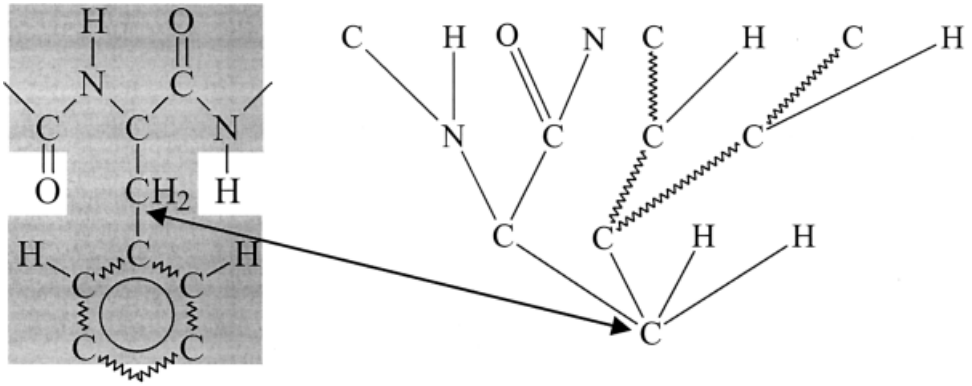

Fig. 1. A phenylalanine residue (**left**). The Cβ carbon atom marked with an arrow serves as the root for creating the molecular tree shown on the right. All atoms in the gray area are within recursion depth 3 of the root and thus part of the tree. Beside the topology, the tree stores data about the bond types as well: single, double, and resonance bonds are shown (the weaker resonance effects of the peptide bond are not considered). From top to bottom, the parts of the tree are called "depth 3," "depth 2," "depth 1," and "root." A comparable tree is generated for every atom in the simulated system, and force-field parameters are then assigned on the basis of the most similar tree in the force-field definition file.

factors: one for the bonds and one for the angles. All other parameters (e.g., Van der Waals interactions and off-center point charges) were optimized independently. To further reduce computational requirements, the energy minimization algorithm searched for NOVA's closest minimum. Therefore, it can only be guaranteed that NOVA has minima close to real protein structures, but not that these minima are also global ones.

Algorithms that search for global minima with big steps in conformational space (e.g., ab initio fold prediction[9]) can be applied safely if the search is restricted to a specific region with additional data (e.g., NMR NOESY restraints). NOVA is useful for applications that require a search for a local minimum near by, such as refinement of experimental low-resolution structures, models built by homology or docked complexes. The force field is shown to significantly reduce the Cα RMSD between experimental structures and theoretical models during an energy minimization.

MATERIALS AND METHODS

The NOVA force field (protei*n* + *n*ucleotide *o*ptimization in *va*cuo) has been implemented as part of the newly developed interactive real-time molecular dynamics program YASARA ("Yet Another Scientific Artificial Reality Application," http://www.yasara.com). It looks like common molecular dynamics force fields, with the total energy being expressed as a sum of individual contributions: bonds, angles, planarity, Van der Waals, and electrostatic terms. Most negative point charges are placed outside the nuclei (off-center charges). Van der Waals interactions are modeled by Born-Mayer Exp6 instead of the familiar Lennard Jones 12-6 potentials. Planarity is treated by least-squares plane fitting instead of improper torsions.

Molecular Trees Define the Chemical Environment

One of the aims during the development of NOVA was the possibility to extend the force field to ligands without the need for manual intervention. This was achieved by using molecular trees (Fig. 1) instead of predefined atom types to assign the force-field parameters. Normally, a topology file lists all atoms by name (e.g., N, H, CA, 1HA, 2HA, C, and O for Glycine) and assigns at least an atom type and often also a point charge. Bond lengths, angles, and so forth are specified for all combinations of atom types.

In contrast, YASARA builds a molecular tree to define the chemical environment of every atom and then chooses force-field parameters based on the closest reference tree found in the NOVA definition file (electronic supplement). Starting from every atom in the molecule (the root), the program follows the various branches (the chemical bonds) of the molecule—up to a certain recursion depth (usually 3).

An example is the molecular tree built from the Cβ atom of phenylalanine (Fig. 1). Bond types (single, double, triple, and resonance) are an integral part of the tree. These types are taken from a connectivity table that contains atom names and bonds for every residue or ligand. These tables can be generated automatically by analyzing heavy atom coordinates, predicting their hybridization state and adding hydrogens where needed. For example, this is done by the Dundee PRODRG Server[10] (http://davapc1.bioch.dundee.ac.uk/programs/prodrg/prodrg.html). Each molecular tree can also be written as a single string, for example, for the Cβ of Phe:

$$\mathrm{CH2(—C(—NH—C)—C(=\!=O)—N)}\\ \mathrm{—C(\div CH\div C)\div CH\div C} \quad (1)$$

Comparable approaches have been suggested before (e.g., by Levitt[11] or Weininger[12]). Here we extend this approach to assign a complete set of force-field parameters.

Building Reference Trees

The NOVA definition file (available from http://www.yasara.com/nova) does not explicitly specify equilibrium

$$E_a = \sum_{j<a} 0.5 * k_j * (R_{j0} - R_j)^2 + 0.5 * l_p * R_p^2 + \sum_{i<a} (A_i * e^{-B_i * R_i} - C_i / R_i^6) + \sum_{k<a} \sum_{m} \sum_{n} \frac{q_m * q_n}{4\pi\varepsilon_0 * R_{mn}}$$

$$\mathbf{F_a} = \sum_{i} k_j * (\mathbf{R_{j0}} - \mathbf{R_j}) - l_p * \mathbf{R_p} + \sum_{i} (\mathbf{R_i} / R_i) * (A_i * B_i * e^{-B_i * R_i} - C_i / R_i^7) + \sum_{k} \sum_{m} \sum_{n} \frac{\mathbf{R_{mn}} * q_m * q_n}{4\pi\varepsilon_0 * R_{mn}^3}$$

Fig. 2. The YASARA NOVA force field. Atom distances are named R, equilibrium values R_0. Vectors are shown in bold print. The energy contribution of atom a (E_a) is the sum over all chemical bonds (to atom j, with bond-stretching force constant k_j, $j<a$), plus a planarity term ($0.5*l_p*R_p^2$), plus the sum over all non-bonded Van der Waals interactions (with atom i, using EXP6-potential parameters A_i, B_i and C_i, $i<a$), plus the electrostatic Coulomb interactions between all m point charges on atom a and n point charges on atom k ($k<a$). More details are given in Materials and Methods and the electronic supplement.

TABLE I. Optimization Parameters of the NOVA Force Field†

Optimization parameters	Parameter description
1	Common scaling factor for all AMBER bond-stretching force constants
2	Common scaling factor for all AMBER angle-bending force constants
3	Planarity force constant of peptide plane
4	Planarity force constant of all planar sidechains (D, E, F, H, N, Q, R, W, Y)
5	Charge c at H (+c) and N (−c) in peptide bond
6	Charge c at C (+c) and O (−c) in C═O groups
7	Distance from O-nucleus of the negative lone "pair" in C═O groups
8	Charge c at H (+c) and N (−c) in aromatic rings (H, W) and NE/HE of R.
9	Charge c at H (+c) and N (−2c or −3c) at NH_2 and NH_3 groups (N, Q, K, R, N-term.)
10	Ionic charge (D, E, K, H-protonated, R, N-terminus, C-terminus)
11	Charge c at C (+2c) and O (−c) in deprotonated carboxyl groups (E, D, C-term.)
12	Distance from O-nucleus of the negative lone "pair" in carboxyl groups
13	Charge c at H (+c), O (−c per lone pair) and C (+c) at hydroxyl groups (S, T, Y, D-protonated, E-protonated, C-terminus protonated)
14–16	Born-Mayer parameters of H—H interaction
17–19	Born-Mayer parameters of H—C interaction
20–22	Born-Mayer parameters of H—N interaction
23–25	Born-Mayer parameters of H—O interaction
26–28	Born-Mayer parameters of C—C interaction
29–31	Born-Mayer parameters of C—N interaction
32–34	Born-Mayer parameters of C—O interaction
35–37	Born-Mayer parameters of N—N interaction
38–40	Born-Mayer parameters of N—O interaction
41–43	Born-Mayer parameters of O—O interaction

†Equilibrium bond lengths and angles are taken from high-resolution X-ray structures without optimization. Bond stretching and angle-bending force constants come from the AMBER force field and are rescaled (parameters 1 and 2 above).

bond lengths and angles. These are extracted from the 25 highest resolution X-ray structures in the PDB, with <30% sequence identity, obtained from the PDB-SELECT algorithm[13] (a list with PDBID codes of all proteins used in this work is available from http://www.yasara.com/nova). Missing hydrogen atoms were added with the WHAT IF hydrogen-bonding network optimizer[14] and then relaxed to the closest energy minimum with the AMBER force field[1] (parm94), whereas all heavy atoms were kept fixed. For each atom in these proteins, a molecular tree was built. The distances between the root and depth 1 atoms (the bond lengths in Fig. 1) and between any two depth 1 atoms (the bond "angles") are stored within the tree. If two atoms share the same local covalent structure, their molecular trees are identical. Therefore, bond lengths (and also angles) were averaged over identical trees. In this way, the program obtained a set of partly residue-dependent bond lengths and angles, which can capture features that are missed when using residue-independent parameters.

Currently, there are 426 different trees available from http://www.yasara.com/novatree.

Bonds

Chemical bond stretching is described by a harmonic potential. Bond lengths (R_{j0} in Fig. 2) are taken from the reference trees. The initial bond stretching force constants came from the AMBER Parm94 set. During force-field parameter optimization, one common scaling factor (parameter 1 in Table I) was assigned to the force constants. The final optimized values (k_j in Fig. 2) are listed in the NOVA definition file. To associate a force constant to a type of bond, slightly modified molecular trees are used (as in the case of VdW parameters). Instead of choosing any of the

two bonded atoms, the bond itself becomes the root of the tree. Thus, there are always two atoms with depth 1, and the bond type is the same for both of them.

Angles

Bond angles are treated like true bonds between 1 and 3 bonded atoms (Urey-Bradley method). Equilibrium distances (R_{j0} in Fig. 2) are taken from the reference trees; the initial angle-bending force constants were converted to the required distance-dependent form from the AMBER Parm94 set. Again, one common scaling factor (parameter 2 in Table I) was assigned to all the angle-bending force constants. The final optimized values (k_j in Fig. 2) can be found in the NOVA definition file.

We chose the Urey-Bradley approach for two reasons: (i) Numerical stability: In many typical NOVA applications, very high temperatures (5000 K) temporarily cause bond angles to approach 180°. This creates problems with angle-dependent formulations, which contain a singularity at 180° and assign very large forces close to this angle that can trap part of the molecule in an unrealistic local minimum (mainly if it belongs to a planar group). (ii) At least qualitatively, the Urey-Bradley method implicitly contains a bond/angle cross-term that is normally missing. (A change in bond lengths influences the bond angles and vice versa.)

Dihedrals

Accurate potentials are more difficult to obtain for torsions. Therefore, we decided to optimize all parameters. To achieve this goal without making the set of optimization parameters too large, we reduced the torsion forces to the repulsion between 1–4 bonded atoms. Thus, 1–4 interactions are treated exactly like non-bonded interactions. However, because Lennard-Jones 12-6 potentials do not model the torsion energy properly, we chose the more flexible Born-Mayer Exp6 potentials,[15] combined with "distant geometry links."

A difficulty with this approach is the hydrogen-bonding problem: The electrostatic attraction between the point charges on polar protons and H-bond acceptors is normally not large enough to compensate for the Van der Waals repulsion. This requires a large reduction of the Van der Waals radii of polar protons, which in turn leads to unrealistic torsional energy profiles that must be corrected with additional terms. By using negative off-center point charges, we increased the electrostatic attraction between proton and acceptor sufficiently to reproduce the experimental H-bond lengths of about 1.9 Å while still keeping realistic VdW parameters.

Distant Geometry Links

Many planar groups contain charged atoms in close proximity. For example, the terminal hydrogen atoms HH12 + HH22 and HH21 + HE in arginine are separated by four bonds, but they lie very close to each other. The distance is about 2.3 Å, leading to repulsive Van der Waals forces. The atoms are all positively charged, adding further repulsive forces with a non-negligible influence on the molecular geometry. But these forces are already implicit in the equilibrium bond lengths and angles taken from the high-resolution structures. Distant geometry links (DGLs) define such critical atom pairs and link them with a pseudo-bond to exclude them from the calculation of non-bonded interactions.

Planarity

All force-field terms considered so far were a function of the distance between two atoms. Planarity of atom groups is one of the features that cannot be based on atomic distances only, because out-of-plane bending is accompanied by only small changes in distances. Here we present a method that differs from the normally used "improper dihedrals." It is fast and has some advantageous features when used with the NOVA force field. Our approach calculates the optimal plane through all members of a planar group. Knowing the normal vector of the plane, a force toward the plane is applied to every atom. For a given atom A_i, this force is simply the distance from the plane ($\mathbf{R_p}$ in Fig. 2) times the "plane stretching force constant" l_p (parameters 3 + 4 in Table I): $\mathbf{F_p} = -\mathbf{R_p} * l_p$. A compensating force $-\mathbf{F_p}/n$ is applied to all the n atoms bound to A_i. Although not entirely conserving energy, this approach offers the advantage that the least-squares plane fitting has to be performed only once per planar group every timestep and that the resulting normal vector can be used to apply non-spherical Van der Waals potentials in DNA base pair stacking.

Van der Waals Interactions

The NOVA force field uses Born-Mayer potentials[15] to describe interatomic forces. This function consists of an attractive R^{-6} term and a short-range exponential repulsion term:

$$E = A * e^{-B*R} - C/R^6 \qquad (2)$$

Our reasons for choosing this function were as follows. First, the Born-Mayer Exp6 function contains three adjustable parameters and thus allows us to shift the root independent of the minimum, whereas the more common Lennard Jones 12-6 potential always has its root at 0.89 times the distance of its minimum. Second, because there is no need to minimize numerical noise as in long-time MD simulations, a simple look-up table can be used to avoid the very expensive evaluation of the exponential term. This approach is particularly handy because the Exp6 function requires special care at short distances: When R approaches zero, the repulsive term reaches A (see Eq. 3), whereas the attractive part tends toward infinity. Thus, the potential becomes attractive again at close separations, somehow resembling "nuclear fusion." Therefore, we replaced the interval from 0 to the first root of the third derivative with an R^{-12} damping function.

Initial parameters were taken from a table published by Mirsky,[16] manually adjusted by up to 5% so that secondary structure elements did not fall apart and a stable starting guess was available; then parameters were optimized for all atom pairs, and no combining rules were

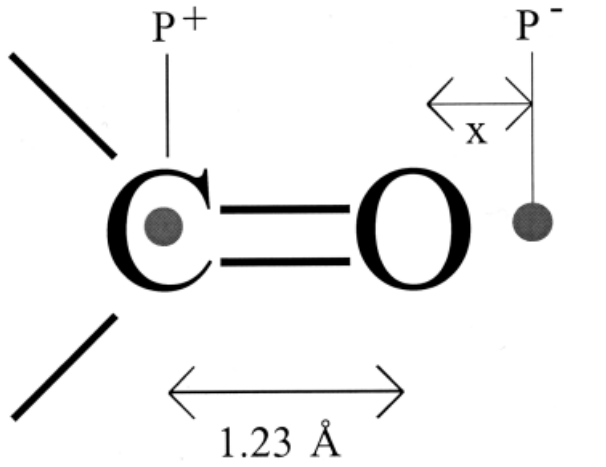

Fig. 3. Placing off-center point charges as a linear combination of atom coordinates. If the distance x of the partial charge P^- from the O nucleus was 0.8 Å, the weighting factors i, j for coordinates **O**, **C** to obtain $\mathbf{P}^-$ would be: $\mathbf{P}^- = \mathbf{O} + (0.8/1.23) * (\mathbf{O} - \mathbf{C})$, $\mathbf{P}^- = (0.8/1.23 + 1) * \mathbf{O} - (0.8/1.23) * \mathbf{C}$ and thus i = 1.65, j = −0.65. In this case, the position of the point charge depends on the length of the C=O bond.

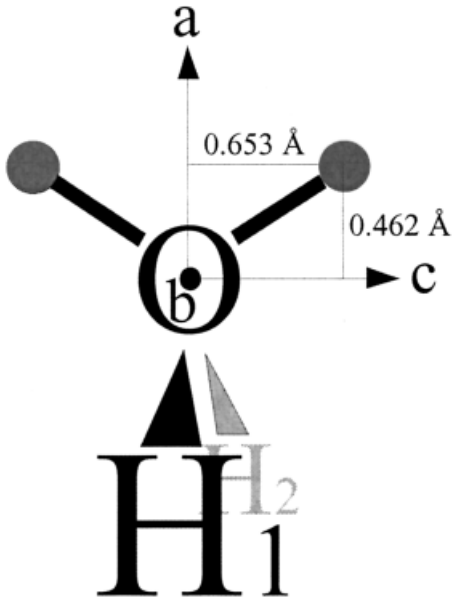

Fig. 4. Placing off -center point charges at sp3 hybridized atoms. This example shows the application of the method to the ST2 water model. The required parameters are the number of atoms and weighting factors needed to calculate the vectors **a** and **b**, the number of charges placed in the plane defined by **a** and **c** (**c** = **a** × **b**), and finally for each of these charges the plane coordinates and size.

used. To keep the number of parameters below a reasonable limit, all interactions involving sulfur were not optimized but taken from AMBER Parm94 and transformed to Exp6 format.

The Exp6 parameters make the largest contribution to the optimization set (parameters 14–43, Table I).

Electrostatics

The electrostatic forces are added by assigning point charges to the atoms. To maximize the level of consistency with the remaining force field, all electrostatic parameters were optimized. This required a description of the essential features with a minimal set of nine parameters (Table I). These included the charges and also their positions, if placed outside the nucleus. Using off-center point charges was required mainly for hydrogen bonds. Because numerical long-time stability is not an issue, charges can be assumed massless. Thus, forces are calculated between the charges, but they act on the associated nuclei. After each change in atom positions, the coordinates of the point charges are recalculated on the basis of the following rules:

Central charges: These are simply placed at the coordinates of the nucleus.

Charges with positions that can be obtained as a linear combination of atom coordinates: A typical example is the lone pair of the carbonyl group (Fig. 3).

Charges with positions that require the calculation of a vector product: The lone pair coordinates of sp3 hybridized atoms [like the oxygen in hydroxyl groups or water (Fig. 4)] cannot be determined with a simple linear combination of atom coordinates. The point charges are placed in a plane N defined by vectors **a** and **c** (Fig. 4), where **c** is **a** × **b**. **a** and **b** are themselves linear combinations of atom coordinates.

Initial guess values for the various charges were obtained from the experimentally measured dipole moments of small molecules; the ST2 parameters[17] were used for hydroxyl groups. Ionic charges were set to 0.3e. Then the parameters were optimized (Table I).

Optimization and Validation Sets

The aim was to create an optimized force field that does not make highest resolution X-ray structures worse while at the same time improves approximate models during an energy minimization procedure. The generation of the required optimization and validation sets of proteins (each one consisting of 25 high-resolution structures and 25 models of high-resolution structures) was done automatically on the basis of a list of highest quality PDB chains with <30% sequence identity, no chain breaks, resolutions better than 1.9 Å and R-factors below 0.19, generated by the PDB-SELECT program.[13] For each of these chains, the script searched for a modeling template in the corresponding FSSP file.[18] If a template existed that allowed modeling the chain without insertions or deletions and had <93% sequence identity, the model was built with WHAT IF[19] and added to the group of models (M); otherwise, the chain became part of the structures group (S). The lowest value that reached the required number of 50 structure-model pairs was 93%. The first 25 odd entries of M and S were taken as optimization set; the even entries as validation set. The 25 entries in group S of the optimization set were also used to generate the reference trees.

The Practical Limit of Fold Prediction

Comparisons of experimental crystal structures indicate limits on effective accuracy that need to be specified in optimizing the parameters. They also reveal practical limits of fold prediction.

Using the PDBFINDER database,[20] we identified chains with identical sequences that have been solved by different authors and refinement programs at resolutions better than 1.9 Å. For these structure pairs, Cα, backbone and heavy-atom RMSDs were calculated, defining the experimental uncertainty of coordinates obtained at high resolution (top and bottom outliers were removed). We used these values (0.48 Å for Cα, 0.95 Å for all heavy atoms) to define how much the force field may modify a structure during an energy minimization before we know that it got worse.

Force-Field Optimization Methods

The NOVA force field described in this study has been optimized by Monte Carlo moves in parameter space. After every step, the quality of the force field was evaluated by running "simulated annealing" molecular dynamics simulations for the 50 structures in the optimization set with YASARA and the following protocol. The non-bonded force cutoff was set to 10.5 Å. One hundred steps of steepest descent minimization with a maximum step size of 0.05 Å removed any sources of conformational stress that might lead to a collapse of the following simulation. Velocity vectors were initialized to average values found at 298 K[4] followed by 3800 integration steps of the equations of motion with the leapfrog algorithm, using a timestep of 2 fs for electrostatic plus Van der Waals interactions and 1 fs for all harmonic forces including planarity.

The random initial velocities were required to avoid optimization toward force-field vectors with zero length, which are guaranteed not to introduce errors. The distortion at the beginning made sure that a realistic parameter set with the ability to "pull the structure back to where it belongs" was obtained. Every 20 fs, all velocity vectors were scaled by 0.9, thus the protein was slowly frozen. After 3.8ps, the timesteps were reduced by 50% to 1 fs and 0.5 fs, respectively, for another 200 cycles. Finally the Cα, backbone, and heavy-atom RMSDs were calculated [with respect to the starting structure (group S) or the modeling target (group M)]. The heavy atom RMSDs of all 50 proteins were averaged and used as a progress indicator. A move in parameter space was accepted with a probability of

$$p = \exp(-(\mathrm{RMSD}_{\mathrm{now}} - \mathrm{RMSD}_{\mathrm{best}})/0.00045) \qquad (3)$$

The value of 0.00045 for kT was empirically chosen so that progress was steady but local minima could still be escaped. A total of 43 force-field parameters were subjected to this minimization procedure (Table I). Each Monte Carlo move was performed by picking one parameter randomly and then either increasing or decreasing it 1–10 times the minimal step size. The minimal step size was predefined for every parameter and equivalent to the final precision required (0.002e for charges, 2% for scaling factors, planarity force constants and VdW contact energies, 0.01 Å for VdW radii and VdW potential roots).

Force-Field Evaluation Methods

After the optimization converged, the force field was evaluated with an extensive minimization: 250 steps of steepest descent and long 40 ps simulated annealing runs without initial velocities. Timesteps were 1 and 0.5 fs for the first 4 ps and 2 and 1 fs for the remaining 36 ps. Force-field energies were calculated without a cutoff distance every 200 fs. If the energy did not drop during five of these measurements (1000 integration steps), the energy minimum was reached and the procedure was stopped (corresponding to horizontal lines in Fig. 8). This avoided the problem that simulated annealing does not stop at the true energy minimum if a cutoff distance is used (it proceeds further to the minimum of the truncated energy function, leading to an increase in true energy).

Computational Requirements

The force-field optimization required about 20,000 h of CPU time. The calculations were performed on 26 PCs at the CMBI, using Models@Home, a freely available screensaver that turns a network of normal, non-dedicated PCs into a distributed computing cluster (http://www.cmbi.nl/models). Two months of computer time could be saved by implementing the NOVA force field in Assembly language.

RESULTS

Force-Field Optimization

Our goal was to parameterize a force field for energy minimization of proteins that does not "mess up" high-resolution X-ray structures and that moves predicted models closer to reality. Most parameters were not chosen on the basis of measured or calculated physicochemical properties, but instead freely optimized with Monte Carlo moves in force-field parameter space. Each move was evaluated with energy minimizations (by simulated annealing) of a 50-protein optimization set: 25 high-resolution X-ray structures and 25 homology models built by WHAT IF.[19]

The parameter optimization progress is shown in Figure 5 for the 25 high-resolution structures. Three distinct regions are visible. During the first 100 optimization steps, the force field improved rapidly. The damage done to the Cα coordinates during the 4-ps minimization decreased by 0.15 Å. At step 100, the "low hanging fruits" were gone, and the first local minimum was reached. From here on, a Monte Carlo algorithm that can escape local minima was obligatory. Nevertheless, there was slow but steady progress until step 450. From step 450 on, improvement was minimal but still measurable. Around step 600, the optimizer finally passed the experimental uncertainty barrier of 0.48 Å (Fig. 5) for Cα atoms: Below 0.48 Å RMSD, it is impossible to decide whether the structure got worse during the minimization. However, the heavy atom RMSD barrier lies much higher, at about 0.95 Å. This can be attributed to the influence of crystal contacts on surface rotamers, which can not be exactly determined experimentally and thus increase the RMSD. After 1000 steps (and thus 50,000 simulated annealing runs) the procedure converged.

Force-Field Evaluation

The above results apply to the optimization set only and were obtained for short 4-ps simulating annealing runs.

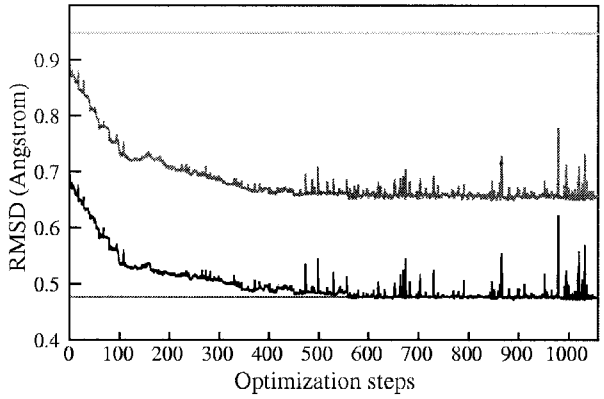

Fig. 5. Force-field optimization progress shown for the 25 high-resolution structures in the optimization set. The average heavy atom (upper curve, gray) and Cα RMSDs (lower curve, black) after ≈4-ps simulated annealing runs are drawn as a function of the optimization step. The two horizontal lines mark the border of experimental uncertainty [i.e., 0.95 Å for the heavy atom (top) and 0.48 Å for the Cα RMSDs (bottom) observed if the same structure is solved at high resolution by different authors and refinement programs]. Anything above these lines surely got worse during the minimization. To make the optimization procedure computationally feasible, every simulated annealing run lasted only 4 ps (and did not always reach the energy minium). Exhaustive simulated annealing runs that proceed until the energy minium is reached are shown in Figure 6.

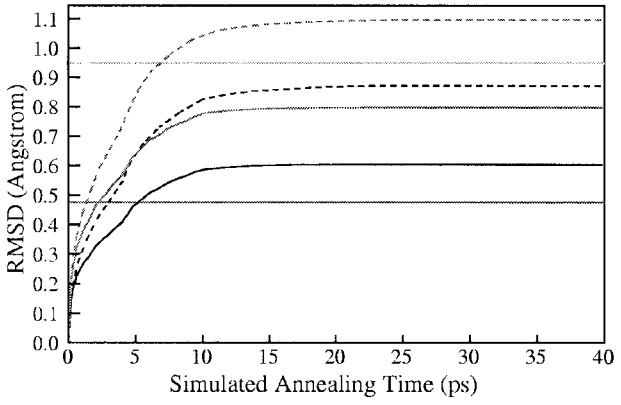

Fig. 6. Extensive minimization of the structure validation set (25 proteins). The two horizontal lines mark the experimental boundaries described in Figure 5. The average heavy atom (gray curves) and Cα RMSDs (black curves) are shown as a function of simulated annealing time. Dashed curves were obtained with initial force-field parameters, and solid curves with final optimized parameters.

(By making the minimization time short enough, one can stay arbitrarily close to the initial structures.) The truly important question is: How far are the NOVA minima away from reality? This can only be answered with an extensive simulated annealing run that proceeds till the energy converges. The result is shown in Figure 6 for an independent validation set of another 25 structures.

Before parameter optimization, the force field undoubtedly made the 25 high-resolution structures worse. The energy minima lay 0.15 Å (heavy atom RMSD) and 0.39 Å (Cα RMSD) above their experimental boundaries (Fig. 6). During optimization, the minima moved closer to reality by 0.30 Å (heavy atoms) and 0.27 Å (Cα). Thus, the Cα RMSD came close to the boundary, whereas the heavy

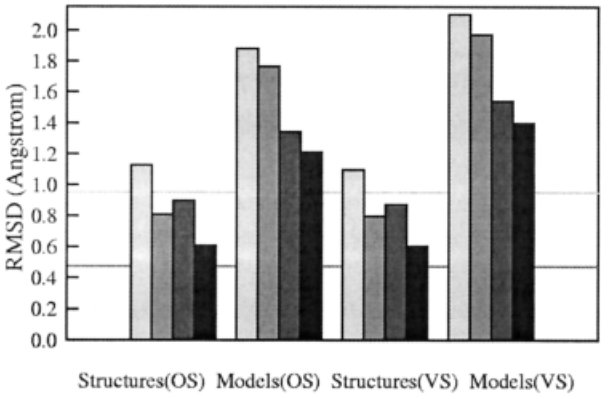

Fig. 7. Force-field parameter optimization results for optimization sets (OS) and validation sets (VS) of 25 structures and 25 models. For each of the four groups, the following values are shown: Average heavy atom RMSD before and after parameter optimization (bright gray and gray bars) and Cα RMSD before and after optimization (dark gray and black bars). RMSDs are measured after a 40-ps simulated annealing run. The two horizontal lines mark the experimental boundaries described in Figure 5. Time-dependent results for the structure validation set (group 3) are shown in Figure 8.

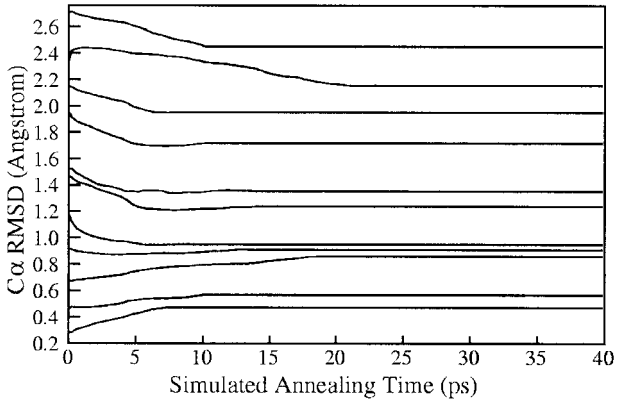

Fig. 8. Extensive minimization of the model validation set (25 proteins) with the optimized force-field parameters. Eleven non-overlapping of 25 trajectories are shown.

atom RMSD crossed it and converged 0.15 Å below. Figure 7 shows the results for all 100 proteins involved.

Model Improvements

Not to mess up high-resolution structures can be regarded as a basic requirement. But to be of practical use, the force field must also be able to move models toward reality. We concentrate on the evaluation of Cα RMSDs to indicate that a true improvement in backbone geometry and not just rotamer prediction accuracy was obtained. It is important to note that WHAT IF was *only* used to mutate the side-chains; the backbones of the templates were simply copied to the models.

Figure 8 shows the energy minimization results for the 25 models in the validation set. There are two different regions to deal wit. If the model is already very close to the true structure (Cα RMSD < 0.9 Å), it gets slightly worse during the minimization; otherwise, it is significantly improved [up to 0.25 Å as in the case of the 1CYO model

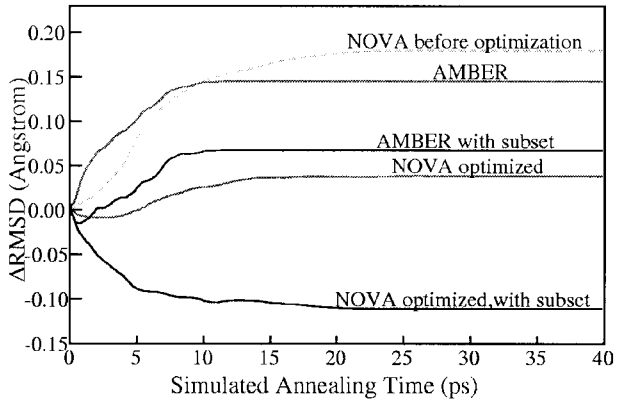

Fig. 9. Average changes of Cα RMSDs during an extensive minimization of the model validation set. Results for the AMBER and NOVA force fields are shown. The minimization protocol was identical in both cases; only the central force-field equation was changed. Gray lines correspond to the complete set (25 models); black lines indicate the subset of those models, where template resolution divided by percentage sequence identity was >0.04 (14 models). Because this subset has a different average RMSD, only changes in RMSD are displayed. The performance of NOVA before optimization is also shown (dashed line).

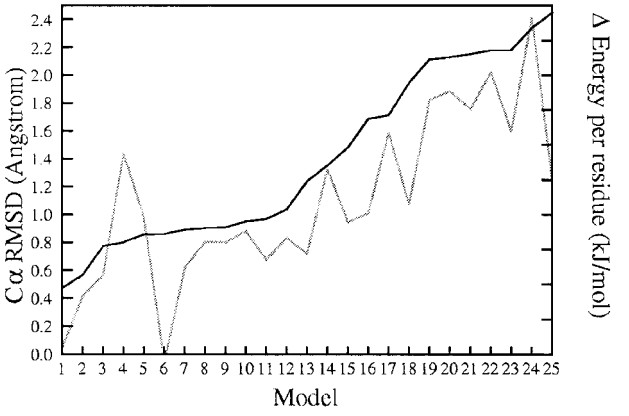

Fig. 10. Comparison of model and structure energies. The Cα RMSDs between the 25 homology models in the validation set and their true structures are shown in black (models have been sorted by RMSD). The gray curve indicates the differences between structure and model energies, whereas the black curve serves as the root: as long as the true structure is lower in energy than the homology model, the gray curve stays below the black one. Every tick on the right axis corresponds to 1kJ/mol NOVA energy difference per residue.

(top curve in Fig. 8)]. This result was to be expected: the closer one gets to the true structure, the more precision in the force field is required to improve the model.

The solution is obvious: not to energy minimize if the model is closer than 0.9 Å to its high-resolution X-ray structure. Because this RMSD is of course not known at the time of modeling, the decision must be based on different grounds. We derived the following empirical rule from the optimization set: Only minimize a model if template resolution (Å) divided by sequence identity (%) is >0.04.

We applied this rule to the validation set and obtained the results shown in Figure 9. Initially, minimizing the models clearly made them worse (Cα RMSD increased from 1.36 to 1.54 Å). During the parameter optimization, the model RMSD dropped from 1.54 to 1.40 Å, 0.04 Å above the initial RMSD without minimization. By energy minimizing only the models that match the selection rule, we found a true improvement: The backbone moved on average 0.111 Å (Cα) closer to reality (bottom curve in Fig. 9).

To investigate the performance of NOVA relative to other force fields, we ran exactly the same energy minimization protocol also with the AMBER force field. Although NOVA with the initial parameters did clearly worse, it is apparent from Figure 9 that the optimization procedure allowed to turn around and go into the right direction.

Force-Field Energy

Ideally, one would like the RMSD to decrease further than the 0.111 Å obtained above, all the way down to the experimental limit of 0.48 Å. It is obvious from Figure 8, that the proteins "get stuck" too early. Two reasons are possible: (i) the NOVA energy function is not precise enough, or (ii) the simulated annealing procedure is not adequate to find the way down.

Low-temperature simulated annealing allows backbone shifts and reorientations of flexible surface rotamers, but certainly not a complete flip of a buried tryptophan. If such a rotamer is initially not predicted correctly, the minimization can easily get stuck in a wrong local minimum. To verify this hypothesis and clarify the possible involvement of point (i), we compared the NOVA energies of the models with those of the true structures. The latter should of course always be lower. Energies were calculated after 40-ps simulated annealing simulations, models, and structures were subjected to the same procedure. The results are shown in Figure 10.

In 3 of 25 cases, the real structure has a higher energy than the model. Structures 4 and 5 are not a big surprise, because Figure 8 has already shown that the force field loses its discriminative power below 0.9 Å Cα RMSD. If it were possible to "home in" from 2.4 to 0.9 Å, this would already be a huge step forward. However, model 24 looks disappointing at first sight: It has a lower energy than the X-ray structure and 2.3 Å Cα RMSD. Closer inspection reveals surprising characteristics. Hirustasin, a serine protease inhibitor, does not have an exactly determined native fold. It is a highly flexible protein,[21] with a very loose residue packing (the WHAT IF packing quality Z-score[22] is −5), and almost no secondary structure (of 51 amino acids in 1BX7, 43 are neither strand nor helix, according to DSSP[23]). The only reason why this protein does not fall apart are five disulfide bonds. By assigning very similar energies to both conformations (the structure 1BX7 and the modeling template bdellastasin 1C9P-B), NOVA predicted this high level of flexibility.

DISCUSSION

We developed an all-atom force field that moves models on average 0.111 Å (Cα RMSD) closer to their true structures, in cases for which template resolution divided by percentage sequence identity is >0.04. This result was achieved with an optimization in force-field parameter space, which allowed to obtain a set of parameters that

optimally fit the approximate force-field equation, disregarding the physical correctness of individual values. This principle has been applied on a small scale since the very beginning of molecular dynamics simulations. The attractive R^{-6} term in the Lennard Jones 12-6 potential is typically a factor 2 larger than suggested by experimental or theoretical data, which is meant to cancel the error caused by neglecting any higher order R^{-8} and R^{-10} terms.[24] With the Monte Carlo search method described here, it was possible to extend this idea to the entire force field.

The improvement in force-field precision was equally large in the optimization and validation sets (Fig. 7). This result is most likely due to the large optimization set with 50 proteins and no restrictions on the number of residues (the largest protein being 1C0P-A with 363 amino acids). Thus, the force-field parameters did not "memorize" any features specific to the optimization set.

The composition of optimization and validation sets has been influenced by two arbitrary choices: first, we split them into 25 structures and 25 homology models, and second, we decided to use only models without insertions or deletions. The latter choice was made to ensure that a signal of progress due to backbone shifts of secondary structure elements (which are the hardest to predict) was not masked by an improvement in loop modeling. To make sure that these choices did not reduce the range of application (e.g., to models without insertions or deletions), we continued the search procedure with models alone and structures alone (25 proteins each) for another 500 steps. Our remarkable finding was that the all-atom RMSDs decreased only by an insignificant amount of 0.003 Å (model optimization set) and 0.008 Å (structure optimization set). This means that keeping a structure in its minimum and improving a model requires the same force-field parameters. It also implies that the force-field parameters *do not depend on the structural characteristics of the models* (the number of insertions, deletions, etc.) and the algorithms used to build the models. The parameters simply provide a precise description of protein structure.

It also became clear that the ability not to make a high-resolution structure worse is a crucial feature: One of the main force-field applications is the calculation and comparison of energies (Fig. 10). For an all-atom force field, this only makes sense after an extensive, unrestrained energy minimization (otherwise, bumps and bond lengths or angles that are slightly off, add a huge, random factor to the force-field energy that makes structure comparisons impossible). If a structure is significantly distorted during the minimization, the whole procedure becomes questionable.

Comparing the initial and optimized values, we observed the smallest changes in experimentally and theoretically well-determined parameters (bond-stretching force constants changed by only 1%), whereas large shifts occurred in the less precisely known parameters (e.g., Van der Waals interactions).

We found that the current model improvement of 0.111 Å is limited by the simulated annealing search rather than the force-field precision. Therefore, future work will include the development of a more flexible minimization algorithm. Getting 0.111 Å closer to reality is a valuable achievement, because the best CASP predictions (http://PredictionCenter.llnl.gov) in homology modeling are often just a few hundredths of an Angstrom ahead of the competitors.

ACKNOWLEDGMENTS

We thank Arthur Lesk and Alexei Finkelstein for carefully reading the manuscript and providing valuable suggestions. We thank all researchers at the CMBI for participating in the Models@Home screensaver project (http://www.cmbi.nl/models), which provided the computational resources for this work.

REFERENCES

1. Pauling L, Corey RB. The structure of proteins: two hydrogen-bonded helical conformations of the polypeptide chain. Proc Natl Acad Sci USA 1951;37:205–211.
2. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. J Am Chem Soc 1995;117:5179–5197.
3. MacKerell AD Jr, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
4. Van Gunsteren WF, Billeter SR, Eising AA, Hünenberger PH, Krüger P, Mark AE, Scott WRP, Tironi IG. In: Biomolecular simulation: the GROMOS96 manual and user guide. ETH Zürich: vdf Hochschulverlag; 1996.
5. Roux B, Simonson T. Implicit solvent models. Biophys Chem 1999;78:1–20.
6. Northrup SH, Pear MR, Lee CY, McCammon JA, Karplus M. Dynamical theory of activated processes in globular proteins. Proc Natl Acad Sci USA 1982;79:4035–4039.
7. Czaplewski C, Rodziewicz-Motowidlo S, Liwo A, Ripoll DR, Wawak RJ, Scheraga HA. Molecular simulation study of cooperativity in hydrophobic association. Protein Sci 2000;9:1235–1245.
8. Tappura K. Influence of rotational energy barriers to the conformational search of protein loops in molecular dynamics and ranking the conformations. Proteins 2001;44:167–179.
9. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio structure prediction of CASP III targets using ROSETTA. Proteins Suppl 1999;3:171–176.
10. Van Aalten DMF, Bywater R, Findlay JBC, Hendlich M, Hooft RWW, Vriend G. PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. J Comput Aid Mol Des 1996;10:255–262.
11. Levitt M. Energy refinement of hen egg-white lysozyme. J Mol Biol 1974;82:393–420.
12. Weininger D. SMILES, a chemical language and information system. J Chem Inf Comput Sci 1993;28:31–36.
13. Hooft RWW, Sander C, Vriend G. Verification of protein structures: side-chain planarity. J Appl Crystallogr 1996;29:714–716.
14. Hooft RWW, Sander C, Vriend G. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins 1996;26:363–376.
15. Born M, Mayer JE. Zur Gittertheorie der lonenkristalle. Z Phys 1932;75:1–6.
16. Mirsky K. In: Schenk R, Olthof-Hazenkamp R, van Koningsveld H, Bassi GC, editors. Computing in crystallography. Twente: Delft University Press; 1978.
17. Rahman A, Stillinger FH. Improved simulation of liquid water by molecular dynamics. J Chem Phys 1974;60:1545–1557.

18. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138. doi:10.1006/jmbi.1993.1489
19. Vriend G. WHAT IF—a molecular modeling and drug design program. J Mol Graph 1990;8:52–56.
20. Hooft RWW, Sander C, Scharf M, Vriend G. The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. Comput Appl Biosci 1996;12:525–529.
21. Uson I, Sheldrick GM, de la Fortelle E, Bricogne G, Di Marco S, Priestle JP, Grutter MG, Mittl PR. The 1.2 A crystal structure of hirustasin reveals the intrinsic flexibility of a family of highly disulphide-bridged inhibitors. Structure Fold Des 1999;7:55–63.
22. Vriend G, Sander C. Quality control of protein models: Directional atomic contact analysis. J Appl Crystallogr 1993;26:47–60.
23. Kabsch W, Sander C. Directory of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 1983;22:2577–2637.
24. Stone AJ. In: The theory of intermolecular forces. Oxford: Clarendon Press; 1996.