

Peptide Backbone Reconstruction Using Dead-End Elimination and a Knowledge-Based Forcefield

STEWART A. ADCOCK

Department of Chemistry and Biochemistry, University of California–San Diego,
4234 Urey Hall, 9500 Gilman Drive, La Jolla, California 92093-0365

Received 26 February 2003; Accepted 24 April 2003

Abstract: A novel, yet simple and automated, protocol for reconstruction of complete peptide backbones from C_α coordinates only is described, validated, and benchmarked. The described method collates a set of possible backbone conformations for each set of residue triads from a structural library derived from the PDB. The optimal permutation of these three residue segments of backbone conformations is determined using the dead-end elimination (DEE) algorithm. Putative conformations are evaluated using a pairwise-additive knowledge-based forcefield term and a fragment overlap term. The protocol described in this report is able to restore the full backbone coordinates to within 0.2–0.6 Å of the actual crystal structure from C_α coordinates only. In addition, it is insensitive to errors in the input C_α coordinates with RMSDs of 3.0 Å, and this is illustrated through application to deliberately distorted C_α traces. The entire process, as described, is rapid, requiring of the order of a few minutes for a typical protein on a typical desktop PC. Approximations enable this to be reduced to a few seconds, although this is at the expense of prediction accuracy. This compares very favorably to previously published methods, being sufficiently fast for general use and being one of the most accurate methods. Because the method is not restricted to the reconstruction from only C_α coordinates, reconstruction based on C_β coordinates is also demonstrated.

© 2003 Wiley Periodicals, Inc. J Comput Chem 25: 16–27, 2004

Key words: dead-end elimination; knowledge-based forcefield; protein backbone

Introduction and Aims

The process of reconstructing an all-atom peptide, or protein, model from a subset of the atomic coordinates is the subject of a considerable body of literature. Numerous applications for this process exist, including enhancement of low-resolution models from crystallography into usable models or conversion of the coarse structures typical of *ab initio* folding computations or comparative protein modeling techniques into all-atom models. The author's requirement for such a procedure derives from the use of a low-resolution protein threading algorithm, which yields predicted C_β coordinates.

It is common to split the process of reconstructing all-atom structures into two separate tasks. The first task is to predict the full backbone, and the second is to build side chains onto this. This partitioning of the problem is considered reasonable, because it is observed that the backbone geometry of well-defined secondary structure elements is mostly invariant to the identity of its amino acid residues. Side-chain modeling is covered in a large number of reports, and thus is not discussed further here.

A variety of approaches are described in the literature for the determination of complete backbone models from C_α coordinates.

Many of these approaches utilize fragment libraries derived from known structures to locate possible structures that do not violate a specified C_α trace. The most favorable fragments to construct the entire backbone are selected using energy-based, homology-based, or geometric criteria.^{1–5} An alternative class of methods perform *de novo* construction of the backbone using either geometric or energy-based criteria. Payne described a dynamic programming method that located optimal rotations of peptide units with respect to a potential of mean force between adjacent residues.⁶ Rey and Skolnick applied frequency tables derived from the PDB to locate C_β coordinates and then generated suitable coordinates for the remaining atoms to fit with those.⁷ Bassolino–Klimas et al. minimized an empirical potential by applying a directed conformational search.⁸ The very rapid method of Milik et al. used information extracted from known structures to generate statistical positions for reconstructed atoms.⁹ Feldman and Hogue built the backbone sequentially, selecting conformations according to database derived trajectory distributions.¹⁰ Liwo et al. maximized the

Correspondence to: S. A. Adcock; e-mail: adcock@mccammon.ucsd.edu

Contract/grant sponsors: NSF, NIH, HHMI, SDSC, the W. M. Keck Foundation, and NBCR.

peptide dipole alignments to predict the backbone conformation.¹¹ Iwata et al. devised an analytical approach to selecting coordinates compatible with favored regions on the Ramachandran map.¹² Several groups have applied molecular dynamics^{12–15} or Monte Carlo procedures^{16–18} to construct, or refine, backbone conformations, usually through the evaluation of standard molecular mechanics forcefields.

The author's need for a reliable backbone reconstruction procedure stems from the development of a protein threading method that outputs predicted structures in the form of C_β coordinates. No reliable and freely distributable software was known to be available, which could, in addition to building structures from C_α traces, also build structures from C_β traces. The devised scheme uses dead-end elimination (DEE) to find the conformation at the global energy minimum, as evaluated by a database-derived empirical forcefield.

Materials and Methods

Knowledge-Based Forcefield

In the present context, a forcefield (FF) is a set of functions that describe a potential energy surface. A FF is usually a considerable approximation with respect to the actual physical characteristics forming the energy surface. For molecular mechanics (MM), a set of simple functions that are only dependent upon atom types, connectivities, and coordinates are used. By avoiding any consideration of electrons (by invoking the Born Oppenheimer approximation) evaluations of points on the energy surface are rapid, and thus molecular dynamics (MD) simulations become tractable. Nonetheless, MM FFs typically endeavor to capture the physics involved by deriving their approximations from physical principles. Effective potentials do not attempt to describe the underlying physics. They do not represent the actual interaction energy of particles, but are parameterized such that they reproduce the sum of all terms that lead to the potential surface. Many-body effects, solvent effects, entropic effects, polarization effects, and so on, are all included in an average sense. In the case of usual MM FFs, it is often suggested that some of the many-body effects or physical terms (e.g., polarization) that are not explicitly included are absorbed into the parameters. However, there is no consideration of entropic terms, such as the hydrophobic effect, which are known to be very important in protein folding. This manifests itself in the inability of MM FFs to correctly characterize native vs. misfolded structures. This deficiency is often corrected with, computationally expensive, procedures such as MM/PBS.¹⁹

Effective residue–residue potentials^{20–26} (also known as inter-residue database potentials) have been successfully used to distinguish between native and misfolded globular proteins.^{24,26} The computational efficiency of effective potentials make them very attractive, particularly in protein threading and protein folding studies. An abundance of experimental structures is now available, enabling knowledge-based potentials to be extracted.

The fundamental assumption adopted in deriving effective potentials for this application is the Boltzmann principle. The Boltzmann principle states that the probability of the occurrence of a given conformational state of energy E scales with the Boltzmann

factor $\exp\{-E/RT\}$, where R is the gas constant and T is the absolute temperature. The probability, or frequency, of an occurrence of a given state, or interaction, can therefore be calculated from the energy of that state, or interaction. Thus, the inverse Boltzmann relationship provides a method to calculate energies from the probabilities, or in other words, from the natural logarithm of the observed frequencies. To be rigorously correct, this requires the complete ensemble of available conformational states. Evidence suggests that the observed ensemble of database structures does not conform to this requirement.²⁷

For the derivation of early effective potentials, it was often assumed that specific residue–residue interactions dominate the stability of a given sequence in a given fold.^{20–22,24,25} The potential of mean force between two residues, A and B , is expressed relative to the average potential between all pairs of residues:²¹

$$\Delta E_{AB}(r) \equiv E_{AB}(r) - E_{XX}(r) \equiv -kT \ln \left[\frac{\bar{g}_{AB}(r \pm \Delta r)}{\bar{g}_{XX}(r \pm \Delta r)} \right] \quad (1)$$

where $\bar{g}_{AB}(r \pm \Delta r)$ is the normalized radial pair distribution function with respect to a distance range of $r \pm \Delta r$. $\bar{g}_{XX}(r \pm \Delta r)$ is the average radial pair distribution function over all amino acid types:

$$\bar{g}_{XX}(r) \equiv \frac{\sum_{A=1}^N \sum_{B=1}^N \bar{g}_{AB}(r)}{N^2} \quad (2)$$

$E_{XX}(r)$ has been regarded as a homogeneous interaction energy between an average pair of residues in the native state of globular proteins, upon which particular preferences, $\Delta E_{AB}(r)$, are superimposed to give the specific interaction potential $E_{AB}(r)$.²⁵

Effective distance-dependent residue–residue energy of interaction between a pair of residues, A and B , at distance r , is:

$$e_{AB}(r) = E_{AB}(r) - \frac{E_{AA}(r) + E_{BB}(r)}{2} \quad (3)$$

This is the energy associated with an occurrence of a residue pair AB at distance r at the expense of pairs AA and BB at that distance.

This scheme incorporates a large number of approximations and assumptions.²⁷ Equation (1) makes an arbitrary choice of reference state, the assumption that Boltzmann statistics are applicable (i.e., the radial distribution is assumed representative of an equilibrium ensemble, and the native state is at the thermodynamic equilibrium), and divides the conformational space into finite spheres of interval $2\Delta r$. The potential energy is also expressed as a sum of pairwise interactions, which act at a single, arbitrary point. Each specific residue pair is assumed to behave independently of other residues and of the environment. Finally, the experimental data set is assumed large enough to sample the full range of residue–residue interactions in soluble protein structures. The PDB is biased toward small globular proteins that are easy to crystallize, and therefore, does not fully represent the set of structures exhibited in nature. All those concerns aside, the success of similar approaches in native fold detection studies lends significant

support to this approach. An excellent review of these assumptions is available.²⁸

For the purposes of this study, these equations are extended to atomic interactions, with each atom type within each residue being considered as an interaction centre.

A triangular sampling function was employed. This sampling function required minor modification to the standard equations detailed above, and Δr is redefined to be the width of each spherical shell.

$$E_{AB}(r) = -kT \ln g_{AB}(r) \quad (4)$$

$g_{AB}(r)$ is the radial pair distribution function, which is proportional to the number of atoms of type B within a differential spherical shell at a distance r from a central atom of type A , $N_{B|A}(r)$. The volume of this shell is $dr = 4\pi r^2 dr$. The number of $A - B$ pairs at separation r , $N_{AB}(r)$, is given by:

$$N_{AB}(r) = N_{B|A}(r) N_A = N_A \rho_B g_{AB}(r) 4\pi r^2 dr \quad (5)$$

where ρ_B is the density of atom B , $\rho_B \equiv x_B \rho$, the mole fraction of B multiplied by the mean number density of the system. In this analysis, $N_{AB}(r)$ is evaluated by counting the atom pairs at separation r occurring in the database of structures:

$$N_{AB}(r) = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \delta(|r_{Ai} - r_{Bj}| - r) \quad (6)$$

$$\delta = \begin{cases} 0 & -\frac{3\Delta r}{2} < R - r \\ \frac{R-r}{\Delta r} & -\frac{3\Delta r}{2} \leq R - r \leq 0 \\ 1 & R - r = 0 \\ \frac{r-R}{\Delta r} & 0 \leq R - r \leq \frac{3\Delta r}{2} \\ 0 & \frac{3\Delta r}{2} > R - r \end{cases} \quad (7)$$

This summation is performed over all pairs of atom types A and B , except for those that are from the same residue or from residues that are adjacent in the sequence. The coordinate of occurrence i of type A is indicated by r_{Ai} . From eq. (5), the normalized radial pair distribution may be written as:

$$\bar{g}_{AB}(r) \equiv \frac{g_{AB}(r)}{\sum_r g_{AB}(r)} = \frac{\frac{N_{AB}(r)}{4\pi r^2}}{\sum_r \frac{N_{AB}(r)}{4\pi r^2}} \quad (8)$$

$\bar{g}_{AB}(r)$ reflects the normalized interaction probability for a given pair of atoms. $\bar{g}_{AB}(r)$ is not biased by the particular frequencies of those atoms occurring in the database. This provides the corresponding potential of mean force:

$$E_{AB}(r) = -kT \ln \left(\frac{(4\pi r^2)^{-1} N_{AB}(r)}{\sum_r (4\pi r^2)^{-1} N_{AB}(r)} \right) \quad (9)$$

The size of each interval, ΔR , is typically of the order of 0.2 Å. A cutoff distance of 15.2 Å was used during sampling. During evaluation using the derived potentials, any cutoff up to 15.2- ΔR may be selected. In all work presented here, the cutoff is 15.0 Å. Beyond 15.0 Å the potential function is close to negligible, if parameterized to such distances.

An analogous potential was created to deal with atoms in the same residue or in residues adjacent in the sequence. In this case, no cutoff was required.

Preparation of Structures

Seven thousand, one hundred, thirty cleaned, high-resolution, structures were used to derive each potential. These structures selected from the PDB²⁹ using the following criteria: (a) resolution of 2.0 Å or better; (b) solved using X-ray crystallography or neutron diffraction; and (c) deposited prior to October 30th 2002.

The identities of the chains in the data set of 7130 structures were used to characterize these chains to remove the bias caused by degenerate structures. Each chain was compared with every other chain using the BLAST³⁰ algorithm as implemented in the NCBI toolkit.³¹ Using the BLAST p -value as the distance measure, they were then clustered using a single-linkage clustering procedure. With a p -value cutoff of 1.0e-6, 1362 clusters were determined from the 12,022 chains in the 7130 structures. These clusters are considered to be sets of chains with significant sequence similarity. Rather than using the more common approach of selecting a single representative chain from each cluster, the contribution to the derived potentials are weighted according to the size of the cluster by modifying eq. (6). The scaling factor S is the inverse of the total number of elements in A_i 's chain's cluster.

$$N_{AB}(r) = \sum_{i=1}^{N_A} S_{Ai} \sum_{j=1}^{N_B} \delta(|r_{Ai} - r_{Bj}| - r) \quad (10)$$

The cleaning was performed using the CLEAN program distributed with the HBPLUS software package,³² which corrects atom labeling, followed by HBPLUS itself to add or correct polar hydrogen atom coordinates.

Internally, the forcefield engine, named SPLIFF-AA, has a hash-based implementation. There are several thousand possible atom-atom interaction types, and the potential for each of these is stored as a lookup array. These lookup arrays are indexed using a fast hashing function, resulting in rapid access to molecular energies.

Only a subset of the SPLIFF-AA forcefield is required in this application because side-chain atoms are not considered.

Dead-End Elimination

Stochastic algorithms, although often highly tunable for a given problem, are never guaranteed to locate the global energy minima. Systematic searching, on the other hand, would always find the global minima but, unfortunately, this is often not a tractable approach due to the computational expense of enumerating and evaluating every possible configuration. An approach that is com-

monly applied to resolve the shortcomings of systematic searching is known as dead-end elimination.^{33,34}

DEE procedures are generally regarded as some of the best approaches for both protein design and side-chain placement.³⁵

The theoretical basis of DEE requires that the energy function may be decomposed into pairwise contributions. By virtue of the natural pairwise decomposition of the potential functions described earlier, they are ideal for application in DEE search procedures.

The DEE algorithm attempts to systematically eliminate states for each variable that are unable to contribute to the global energy minimum. If DEE converges to a single solution, it may be proven that the global energy minimum is found.³³ Often, DEE does not converge, but in such cases one would hope that a sufficient number of possible rotamers are eliminated to ensure that complete enumeration of the remaining configurations is tractable.

The energy function that this particular DEE implementation seeks to solve is:

$$E_{\text{total}} = \sum_{r=1}^N E_{\text{single}}(r_i) + \sum_{r=1}^N \sum_{s=1}^N E_{\text{pair}}(r_i, s_j) \quad (11)$$

In this expression, N is the total number of residues. r_i represents residue r with its observed local backbone conformation i . Likewise, s_j is residue s with backbone conformation j . $E_{\text{single}}(r_i)$ represents the energy of r_i , which may be considered as including both internal energy terms, and any interactions with the environment. In the presented work, E_{single} is a harmonic penalty function ensuring that the backbone fragment overlays the C_{α} trace sufficiently well. It is assumed that the internal energies of the backbone fragments are very similar, at least relative to the potentially large residue-residue interactions. The applied harmonic constant was $2.0 \text{ RT } \text{\AA}^2$. $E_{\text{pair}}(r_i, s_j)$ represents the interaction energy between r_i and s_j . E_{pair} is the value derived from the SPLIFF-AA force field for the central residue of each fragment, with an additional superpositioning term for cases when atoms in the pair of fragments overlap. Again, this is a simple harmonic penalty function, with a harmonic constant of $1.5 \text{ RT } \text{\AA}^2$. In the case of protein design, i and j would indicate residue identity in addition to rotameric state. Locating the global minimum configuration of rotamers via a systematic search would obviously require significant computation.

State i for variable r , that is r_i , will not contribute to the global energy configuration if the net energy contribution from its best case pairwise interactions with all states for all other variables is higher than that of the worst case pairwise interactions with an alternative state, r_k . This statement is known as the original dead-end elimination criterion.³³ Formally, this may be stated as:

$$E_{\text{single}}(r_i) - E_{\text{single}}(r_j) + \sum_{s=1, s \neq r}^N \left\{ \min_k E_{\text{pair}}(r_i, s_k) \right\} - \sum_{s=1, s \neq r}^N \left\{ \max_k E_{\text{pair}}(r_j, s_k) \right\} > 0 \quad (12)$$

State r_i may also be eliminated if the contribution to the total energy is always higher than that of an alternative state, r_k . This is known as the simple dead-end elimination criterion,³⁴ and may be formally stated as:

$$E_{\text{single}}(r_i) - E_{\text{single}}(r_j) + \sum_{s=1, s \neq r}^N \left\{ \min_k [E_{\text{pair}}(r_i, s_k) - E_{\text{pair}}(r_j, s_k)] \right\} > 0 \quad (13)$$

When applied in practice, both of the above elimination criteria converge on sets of more than one possible configuration. Further eliminations may be made by determining pairs of states that cannot coexist in the global energy minimum configuration. Adapting the original DEE criterion³⁶ gives:

$$\begin{aligned} \varepsilon(r_i, s_k) - \varepsilon(r_j, s_l) + \sum_{t=1, t \neq r \neq s}^N \left\{ \min_m E_{\text{pair}}(r_i, t_m) + E_{\text{pair}}(s_j, t_m) \right\} \\ - \sum_{t=1, t \neq r \neq s}^N \left\{ \max_m E_{\text{pair}}(r_i, t_m) + E_{\text{pair}}(s_k, t_m) \right\} > 0 \quad (14) \end{aligned}$$

where

$$\varepsilon(a_x, b_y) = E_{\text{single}}(a_x) + E_{\text{single}}(b_y) + E_{\text{pair}}(a_x, b_y) \quad (15)$$

Similarly, adapting the simple DEE criterion³⁴ gives:

$$\begin{aligned} \varepsilon(r_i, s_k) - \varepsilon(r_j, s_l) + \sum_{t=1, t \neq r \neq s}^N \left\{ \min_m [E_{\text{pair}}(r_i, t_m) + E_{\text{pair}}(s_j, t_m) \right. \\ \left. - E_{\text{pair}}(r_i, t_m) + E_{\text{pair}}(s_k, t_m)] \right\} > 0 \quad (16) \end{aligned}$$

Eliminating a pair of states using the above two criteria does not directly allow elimination of the individual states because either could exist individually in the global energy minimum configuration. However, in subsequent single-state elimination calculations, eliminated pairs will no longer contribute, and this could result in additional single residue eliminations.

A number of further enhancements have been described in the literature, for example, residue unification³⁷ where a so-called "super-residue" is formed from all of the possible rotamer pairs at two positions. This super-residue is then treated as a single residue for the remainder of the calculation. Search efficiency may also be improved by estimating which of the possible rotamer pairs are unlikely to be eliminated and therefore skipping their evaluation.³⁸ A trivial trick for enhancing the efficiency of any search algorithm, including DEE, is to ignore any rotamer or rotamer pair that contributes a very high energy to any configuration in which it is present, for example, from a clash with the backbone. This last approach, "threshold elimination" is the only accelerating enhancement applied in the presented procedure, at this time.

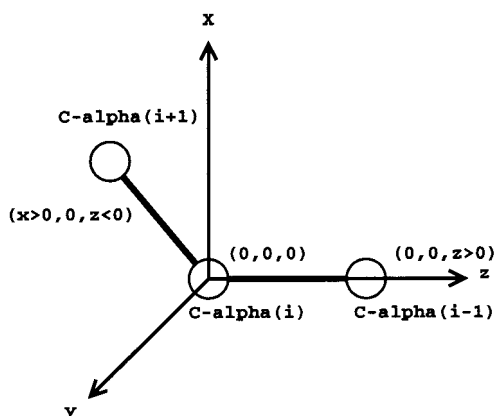


Figure 1. The standard orientation enforced on the three-residue fragment, centered on residue i . In the case that C_α coordinates are not provided (e.g., for the reconstruction of backbones from C_β coordinates) then the alternative atoms are used instead.

Backbone Fragment Library

A library of three-residue backbone fragments was extracted from 1336 randomly selected nonredundant PDB structures. Each structure was considered redundant if its longest chain had greater than 95% sequence identity with a previously selected structure. Any fragments with missing atoms, or bonds outside of generally accepted lengths, are also discarded. This fragment library was sorted according to the presence of glycine, proline, and other residues in the first instance and then by the C_α - C_α - C_α angle, discretized to 6° increments. For reconstruction of backbones from atoms other than the C_α traces, alternative libraries are built with appropriate angular criteria. Duplicate fragments, as determined by clustering to an adjustable backbone atom RMSD tolerance with the fragments aligned in the standard frame, were discarded. For all of the presented results, the RMSD clustering tolerance was 1.0 Angstroms.

All fragments in the library were aligned in a standard reference frame, as shown schematically in Figure 1. This allowed several evaluations to be optimized out of the algorithm used to build structures from these fragments.

Prior to the DEE process, the least frequently observed fragment conformations are discarded. In the results presented, the threshold is 0.5%.

Backbone Construction

Fragment Overlap

Individual fragments are overlaid on the known backbone coordinates by determining the orthogonal transformations that minimize the sum of squared distances between the corresponding pairs of atoms in the fragments with those in the known backbone. One should note that this method is also applicable to the case that atomic coordinates other than the C_α are known. Although such results are not presented, this protocol is not specific to only the reconstruction of backbones from C_α traces, and consequently, simple three-point superpositioning is not suitable.

The orthogonal transformations are determined using a slightly modified form of Kearsley's method,³⁹ which is an analytical solution of a constrained least-squares problem by stating it as an eigenvalue problem with quaternion algebra. Practically, this method is elegant because: (a) it is not an iterative procedure; instead, it only requires construction and diagonalization of a symmetric 4×4 matrix; (b) improper rotations are never produced; (c) it is general, with no special cases to consider separately; and (d) the RMSD between the superimposed vector sets falls out of the calculation with no significant computational overhead.

In the case of molecular structures, it was assumed to be advantageous to weight the individual contributions to the solution by the atomic masses. Thus, an extension to Kearsley's noniterative superpositioning algorithm has been implemented for molecular structures, in which the vector-pair contributions are weighted by their associated atomic masses.

We relate a pair of corresponding vector sets of dimension N , a reference set a , and a rotated set b , which are both mass centered, by a general quaternion q . The contribution of each pair of vectors, a_i and b_i , is weighted using vector W , which in our application is the vector of atomic masses. We wish to determine q by minimizing the weighted sum of squared magnitudes of the residual quaternions, as given by eq. (18)

$$r_i = [0, a_i] - q^{-1}[0, b_i]q \quad (17)$$

where r_i is the residual quaternion.

By multiplying the above equation by q , we may construct our weighted least-squares residual function, ε .

$$\begin{aligned} \varepsilon &= \sum_{i=0}^{N-1} |W_i(qr_i)|^2 = |q|^2 \sum_{i=0}^{N-1} |W_i(r_i)|^2 \\ &= \sum_{i=0}^{N-1} |W_i[-(\xi, \eta, \zeta) \cdot (a_i - b_i), \\ &\quad \chi(a_i - b_i) + (\xi, \eta, \zeta) \times (a_i - b_i)]|^2 \quad (18) \end{aligned}$$

Because q is a general quaternion, rather than unit quaternion, the transformation will include a component of isotropic dilation. To ensure that the transformation is a pure rotation, we constrain the norm of q to unity using a Lagrangian multiplier λ , as in eq. (19). The Lagrangian term removes the nonlinear contribution to the weighted least-squares procedure. With the norm of q constrained to unity, solutions for r will reduce to our original weighted least-squares residual function, eq. (18)

$$\begin{aligned} \varepsilon &= \sum_{i=0}^{N-1} W_i \{ [\xi(a_{ix} - b_{ix}) + \eta(a_{iy} - b_{iy}) + \zeta(a_{iz} - b_{iz})]^2 \\ &\quad + [\chi(a_{ix} - b_{ix}) + \eta(a_{iz} - b_{iz}) + \zeta(a_{iy} - b_{iy})]^2 \\ &\quad + [\chi(a_{iy} - b_{iy}) + \zeta(a_{ix} - b_{ix}) + \xi(a_{iz} - b_{iz})]^2 \end{aligned}$$

$$+ [\chi(a_{iz} - b_{iz}) + \xi(a_{iy} - b_{iy}) + \eta(a_{ix} - b_{ix})]^2 + \lambda(1 - \chi^2 - \xi^2 - \eta^2 - \zeta^2) \quad (19)$$

with $a_i = (a_{ix}, a_{iy}, a_{iz})$, and likewise, $b_i = (b_{ix}, b_{iy}, b_{iz})$.

$$\lambda \begin{pmatrix} \chi \\ \xi \\ \eta \\ \zeta \end{pmatrix} = \begin{pmatrix} \sum W_i \{(x_i^+)^2 + (y_i^-)^2 + (z_i^-)^2\} & \sum W_i \{y_i^+ z_i^- - y_i^- z_i^+\} & \sum W_i \{x_i^- z_i^+ - x_i^+ z_i^-\} & \sum W_i \{x_i^+ y_i^- - x_i^- y_i^+\} \\ \sum W_i \{y_i^+ z_i^- - y_i^- z_i^+\} & \sum W_i \{(x_i^-)^2 + (y_i^+)^2 + (z_i^+)^2\} & \sum W_i \{x_i^- y_i^- - x_i^+ y_i^+\} & \sum W_i \{x_i^- z_i^- - x_i^+ z_i^+\} \\ \sum W_i \{x_i^- z_i^+ - x_i^+ z_i^-\} & \sum W_i \{x_i^- y_i^- - x_i^+ y_i^+\} & \sum W_i \{(x_i^+)^2 + (y_i^-)^2 + (z_i^+)^2\} & \sum W_i \{y_i^- z_i^- - y_i^+ z_i^+\} \\ \sum W_i \{x_i^+ y_i^- - x_i^- y_i^+\} & \sum W_i \{x_i^- z_i^- - x_i^+ z_i^+\} & \sum W_i \{y_i^- z_i^- - y_i^+ z_i^+\} & \sum W_i \{(x_i^+)^2 + (y_i^+)^2 + (z_i^-)^2\} \end{pmatrix} \begin{pmatrix} \chi \\ \xi \\ \eta \\ \zeta \end{pmatrix} \quad (20)$$

where \sum indicates the sum over all values for

$$i, \sum_{i=0}^{N-1}.$$

In the case that all atomic weights are unity, eq. (20) reduces to Kearsley's equation. The eigenvector with the smallest eigenvalue indicates the rotation required to minimize the weighted sum of squared distances, with the eigenvalue equal to the corresponding residual error (i.e., the weighted sum of the residuals squared for the superimposed vector sets). The RMSD is conveniently given by:

$$\text{RMSD} = \sqrt{\frac{\lambda}{m}} \quad (21)$$

where

$$m = \sum_{i=0}^{N-1} W_i,$$

the sum of all individual atomic masses.

For the special case that all weights are unity, we get Kearsley's original result that:

$$\text{RMSD} = \sqrt{\frac{\lambda}{N}} \quad (22)$$

The required matrix diagonalization was performed using a Jacobian-based eigenvector solving procedure.

Search Procedure

DEE, as described above, is used to search for an optimal sequence of fragments. On occasion, the DEE procedure is unable to converge. In these situations, the least likely fragments, based on the frequency of occurrence in the set of PDB structures used to generate the fragment library, are successively removed from the

By differentiating ε with respect to each component of q and defining as zero, the resulting equations may be organized as an eigenvalue problem. Using the notation, $\alpha_{i\alpha}^+ = (a_{i\alpha} + b_{i\alpha})$ and $\alpha_{i\alpha}^- = (a_{i\alpha} - b_{i\alpha})$, we have:

allowable list, and further DEE elimination steps are taken until convergence occurs.

Structure Generation

The middle portion of the three-residue fragment selected at each position in the sequence is used to build that part of the backbone. This middle portion includes the carbonyl group of residue one and all of residue two, except for its carbonyl group. At the termini and sequence breaks the relevant portions of the adjacent fragments are used instead.

Results

Forcefield Validation

Although the results are not shown, extensive studies were performed to validate the SPLIFF-AA forcefield. As an illustrative example, it is applied to the evaluation of standard decoy sets in Table 1.⁴⁰⁻⁴²

An Example: Immunoglobulin-Binding Protein (IIGD)

Kazmierkiewicz et al.¹⁸ illustrated the application of their method by reconstructing the backbone coordinates of an immunoglobulin-binding protein. They reported an RMSD over all backbone atoms of 0.50 Å, with maximum deviations of 0.7 Å for MET-1, THR-21, THR-58, and LYS-59. It is not clear to the author why LYS-59 is reported as such, because the crystal structure contains a valine at position 59. They also reported a correlation coefficient of 0.86 for ϕ dihedrals and 0.97 for ψ dihedrals. With the present method, an overall backbone RMSD of 0.345 Å is observed, with maximum deviations of 1.39, 0.99, 0.63, and 0.72 Å, for MET-1, ASP-45, TRP-46, and GLU-60, respectively. Only seven residues had a backbone RMSD of 0.5 Å. The correlation coefficients for the backbone dihedrals, ϕ and ψ are 0.96 and 0.97, respectively.

Figure 2 provides a view of the resulting structure.

This calculation took approximately 10 s on a 1.4 GHz PC, which compares to a reported time of 22 min on a 650 MHz PIII PC for the method of Kazmierkiewicz et al.

Table 1. Results of Decoy Evaluation.

Decoy set	Protein PDB ID	Number of decoys	SPLIFF-AA (all-atoms)		SPLIFF-AA (backbone-atoms)		Samudrala and Moul, 1998		Fain et al., 2002	
			Z	Rank	Z	Rank	Z	Rank		
Levitt ⁴⁰	1CTF	630	-9.97	1	-9.08	1	-2.9	1	-2.9	1
	1R69	675	-10.77	1	-10.32	1	-2.4	1	-2.0	11
	1SN3	660	-13.73	1	-13.82	1	-3.3	1	-2.7	5
	2CRO	674	-9.99	1	-9.36	1	-2.5	1	-1.7	22
	3ICB	653	-8.16	1	-7.38	1	-1.7	22	-1.4	55
	4PTI	687	-12.86	1	-12.97	1	-3.3	1	-2.0	13
	4RXN	677	-14.17	1	-13.91	1	-2.6	1	-3.4	2
	Lattice ⁴⁶	1BEO	2000	-8.23	1	-8.80	1	-9.4	1	-10.1
	1CTF	2000	-1.31	125	-1.06	262	-6.7	1	-6.6	1
	1DKT/A	2000	-7.34	1	-7.24	1	-5.6	1	-5.4	1
	1FCA	2000	-8.22	1	-0.18	933	-5.6	1	-5.1	1
	1NKL	2000	-1.43	53	-1.55	34	-7.3	1	-7.8	1
	1PGB	2000	-7.84	1	-7.96	1	-8.9	1	-9.9	1
	1TRL/A	2000	-2.62	1	-2.04	1	-3.9	1	-4.9	1
	4ICB	2000	-5.68	1	-5.60	1	-4.3	1	-4.9	1
Fisa ⁴²	1FC2	500	-3.63	1	-4.47	1				
	1HDD/C	500	-4.87	1	-5.59	1				
	2CRO	500	-5.28	1	-5.32	1				
	4ICB	500	-8.30	1	-8.56	1				

Only the native 1CTF and 1NKL structures are incorrectly ranked (against the Lattice Secondary Structure Fit Decoys) using the SPLIFF-AA potential energy terms for all-atoms. Using the backbone atoms' contributions to the effective energy produces an incorrect rank for 1FCA, in addition. The columns labeled Z contain the Z-scores for the native structure with respect to the decoy set. Z-scores express the score of the native structure in terms of the standard deviation of scores in the decoy set: $Z = S_o - \langle S \rangle / \sigma$. The more negative the Z-score, the better the predictive ability of the scoring function. The Samudrala et al. columns were calculated using the freely available code and potentials described by Samudrala and Moul.⁴⁶ The Fain et al. columns were directly taken from ref. 47.

Initially, the total number of fragment permutations that could describe the global energy minimum is astronomical. In the case demonstrated here, there are 1×10^{102} possible permutations with the default fragment library. Evaluating these through systematic searching would clearly be ridiculous. The initial threshold elimination step does not remove any possibilities, in this case. One iteration of the original singles elimination reduces the number of permutations to 1×10^{71} and one iteration of the simple singles elimination further reduces this to 1×10^{12} , at which point 20 of the 61 variables are fully solved. One iteration of the original pairs elimination reduces the number of possible permutations to 1×10^7 . This is followed by one iteration of the simple pairs elimination that yields a single solution.

Reconstruction of a Series of Structures

A set of structures, chosen based on their use in prior backbone reconstruction literature, were built using the described protocol. The input C_α coordinates were extracted directly from the crystal structures. The results are shown in Table 2. For the cases in which alternative atomic coordinates are provided in the crystal structure, the conformation with greatest occupancy is selected for the comparison.

Comparison to MaxSprout

MaxSprout, an interactive server provided at the European Bioinformatics Institute,⁴³ is a widely used modeling tool. MaxSprout uses a fragment-based protocol.⁴ The accuracy of structures built using the MaxSprout server were compared to those built using the present method. As is evident from Table 3, MaxSprout is roughly 20% less accurate than the present method, and this value ignores the cases for which MaxSprout failed to generate a complete set of satisfactory backbone coordinates.

Reconstruction of a Series of Distorted Structures

Often, modeled C_α coordinates will contain errors. Distortions were deliberately made to input C_α traces to gauge the effect of such errors. For each of the three Cartesian axes, a random translation from a Gaussian distribution was applied to each C_α coordinate. The resulting models were compared to the models built from the original C_α coordinates, and the results are plotted in Figure 3. It is seen that errors in the input coordinates have an almost linear effect on the reconstructed backbone coordinates. The excess error in the output coordinates is approximately 80% of the error in the input coordinates, implying that the erroneous coordinates are improved by the backbone reconstruction process.

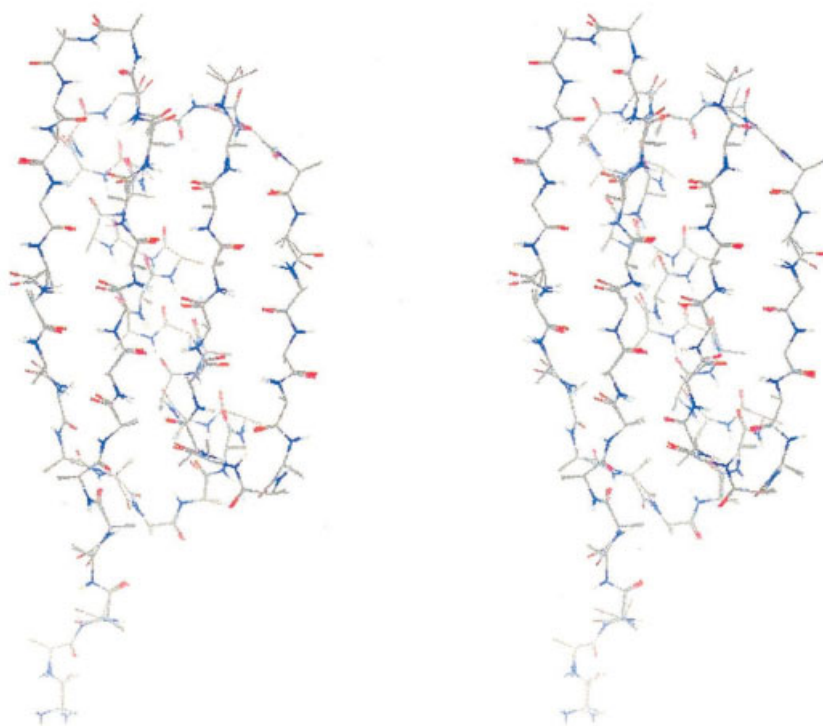


Figure 2. A stereoview of the superposition of the crystal structure backbone atoms (with hydrogens added by HBPLUS³²) and the predicted backbone. The measured RMSD is 0.345 Å.

Analysis of Predicted ϕ and ψ Dihedrals

Side-chain placement algorithms are often dependent upon the accuracy of the backbone ϕ and ψ dihedrals, particularly those algorithms that use backbone-dependent rotamer libraries or χ angle probabilities. Therefore, the correlation between predicted and measured dihedrals was evaluated and presented in Table 4.

Reconstructions from C_{β} Coordinates

As an illustration of this methods applicability to rebuilding backbones from atom subsets other than the C_{α} trace, Table 5 shows the results of reconstruction from C_{β} traces.

Discussion

Once the backbone of a protein has been constructed, only side-chain placement is needed to produce a full all-atom three-dimensional structure. There is a wide range of reliable and readily available methods for this task, and the overall accuracy of resulting structures are primarily dependant upon the choice of these sidechain construction methods.

When applied in actual protein modeling tasks, structures resulting from this procedure are subsequently minimized. Once again, a knowledge-based forcefield is usually applied. This minimization process relaxes stressed portions of the structure, preparing it for further studies or simulation. If the C_{α} coordinates are

fixed at their crystal positions, the minimization procedure further improves the agreement between backbone atoms' predicted and experimentally measured coordinates. Most backbone reconstruction algorithms described in the literature do not apply significant minimization, so using this to generate results for this article was not considered a fair test. The reader should bear in mind that the presented RMSDs are, therefore, at the lower end of the quality scale of those for protein models derived using this procedure followed by refinement.

The method is highly tunable, with certain parameters giving backbone structures for a typical 300 residue protein in approximately 2 s. However, this is at the expense of reliable results, with around two residues per 100 being unassigned and an overall RMSD of the remaining backbone atoms of the order of 1.0–1.5 Å. Such tuning options include aggressive distance cutoffs and automatic elimination of a large proportion of the least-probable fragments, based on the frequency of existence in the original structural library. In addition, the relative weights of the superpositioning score and the energetic score may be modified so that when the backbone is known to be inaccurate, the method can compensate, to some extent. No systematic study of this aspect of the method has yet been undertaken; however, preliminary results show that with C_{α} RMSD distortions of around 1 Å, the predicted backbone can be seen to consistently improve the C_{α} trace.

Comparison of Kazmierkiewicz et al.'s IIGD calculation shows that this present method is a superior approach in terms of

Table 2. A Sample of the Structures, and Their Corresponding PDB Index Codes, Used to Compare with Results of Previous Work.

Protein	Number of residues	PDB index	Main-chain RMSD (Å)		Reference
			Present work	Prior method	
Myoglobin	154	111M	0.31	0.49 ^d	Feldman and Hogue, 2000
Crambin	46	1CRN	0.44	0.56 0.20 ^c 0.41 ^d 0.64 ^h	Levitt, 1992 Payne, 1993 Milik et al., 1997 Gan et al., 1997
Ribosomal protein	74	1CTF	0.41	0.29 0.19 ^c 0.46 ^d	Levitt, 1992 Payne, 1993 Milik et al., 1997
Immunoglobulin binding protein	61	1IGD	0.34	0.50	Kazmierkiewicz et al., 2002
Oncomodulin	107	1OMD	0.39	0.41 ^d	Feldman and Hogue, 2000
Signal transduction protein	129	1SEM	0.50	0.64 ^{d,e}	Feldman and Hogue, 2000
Triose phosphate isomerase	494	1TIM	0.56	0.55 0.50 ^c 0.53/0.54 ^{a,b} 0.60 ^d 0.58 ^h 0.63 0.70 ^{d,f}	Claessens et al., 1989 Payne, 1993 Iwata et al., 2002 Milik et al., 1997 Gan et al., 1997 Rey and Skolnick, 1992 Feldman and Hogue, 2000
Ubiquitin	76	1UBQ	0.37	0.42 0.25 ^c 0.32 ^d	Holm and Sander, 1991 Payne, 1993 Milik et al., 1997
Serine protease	198	2ALP	0.47	0.19 0.30 ^c 0.45 ^d	Correa, 1990 Payne, 1993 Milik et al., 1997
Citrate synthase	437	2CTS	0.37	0.54 0.33 ^c 0.41/0.42 ^{a,b} 0.37 ^d	Claessens et al., 1989 Payne, 1993 Iwata et al., 2002 Milik et al., 1997
Lysozyme	129	2LYM	0.32	0.69 ^d 0.76	Feldman and Hogue, 2000 Rey and Skolnick, 1992
Myohemeythrin	118	2MHR	0.33	0.71 0.22 ^c 0.46 ^d 0.76 ^d	Rey and Skolnick, 1992 Payne, 1993 Milik et al., 1997 Feldman and Hogue, 2000
Apo-plastocyanin	99	2PCY	0.48	0.86 ^d 0.72	Feldman and Hogue, 2000 Rey and Skolnick, 1992
Serine proteinase	279	2PRK	0.37	0.49 0.26 ^c 0.36 ^d	Holm and Sander, 1991 Payne, 1993 Milik et al., 1997
DNA binding protein	107	2WRP	0.24	0.46 0.18 ^c 0.21 ^d	Holm and Sander, 1991 Payne, 1993 Milik et al., 1997
Hydrolase	323	3APP	0.40	0.42 ^d	Milik et al., 1997
Tyrosyl-transfer RNA synthase	317	3TS1	0.40	0.58 ^{d,g}	Feldman and Hogue, 2000
Trypsin inhibitor	58	4PTI	0.51	0.39 ^a 0.56 ^d 0.61 ^h 0.63	Iwata et al., 2002 Feldman and Hogue, 2000 Gan et al., 1997 Rey and Skolnick, 1992
Hydrolase	307	5CPA	0.48	0.61 0.33 ^c 0.42 ^a 0.48 ^d	Claessens et al., 1989 Payne, 1993 Iwata et al., 2002 Milik et al., 1997

(continued)

Table 2. (Continued)

Protein	Number of residues	PDB index	Main-chain RMSD (Å)		Reference
			Present work	Prior method	
Flavodoxin	138	5NLL	0.42	0.47 ^a	Iwata et al., 2002
Trypsin inhibitor	58	6PTI	0.46	0.51	Levitt, 1992
				0.32 ^c	Payne, 1993
				0.38 ^d	Milik et al., 1997
Agglutinin	342	9WGA	0.52	0.45 ^d	Milik et al., 1997

These particular protein structures were deleted from the sets used to generate the forcefield and the fragment library, although homologous proteins or alternative solutions of the same proteins remained. The RMSD is calculated between the unmodified X-ray structure vs. the generated backbone model over the N, H, C_α, C_β, C, and O atoms.

^aPrior to minimization by molecular dynamics.

^bReported for each chain.

^cAlthough the method of Payne gives the most favorable results, it is the author's understanding that all proteins in the test set, except 1TIM, were used in the small training set to parameterize the applied functions, and therefore, the results are meretriciously enhanced.

^dOnly N, C, O, and C_β atoms considered in calculation of the RMSD, which is likely to lead to lower values than if the H atom was also considered.

^eOnly the first 58 residues were constructed by Feldman and Hogue.

^fOnly one subunit, 247 residues were constructed by Feldman and Hogue.

^gOnly the first 211 residues were constructed by Feldman and Hogue.

^hPrior to energy minimization.

both computational time and accuracy of the solution. This conclusion is likely to be true for any given protein.

This method may be easily adapted for reconstruction from other atom subsets, as demonstrated for C_β coordinates. It is feasible that it could be trivially extended to instead deal with virtual atoms. This would be useful for cases where, for example, side-chain geometric centers are used in coarse protein models. It is interesting to note that the performance of this algorithm is better for reconstruction from C_α coordinates than from C_β coordinates. The backbone conformational freedom given particular C_α coordinates is less than that given C_β coordinates. This results in a larger C_α fragment library than C_β fragment library for the same clustering threshold. Because a constant probability threshold of 0.5% is used, the backbone conformational space for specific C_β coordinates is not as well sampled. A lower likelihood threshold improves the accuracy of the results, but at a considerable computational expense (results not shown).

The implementation of this procedure allows the scoring functions to be readily changed. For example, one could envision calculating the single-residue contributions by comparison to electron density maps. One advantage that the current scoring function possesses is that it considers long-range interactions, something that the majority of the older methods do not. Long-range interactions are expected to determine, to at least some extent, the local backbone conformation.

By virtue of using the DEE algorithm, the solution is guaranteed to be the global minimum for the selected scoring function, assuming successful convergence. A trivial extension to the DEE algorithm enables location of all structures within a specified energy of the global energy minimum. Coupled to the A* search algorithm, this has been shown to be an effective method for sampling the protein sidechain conformational energy surface,⁴⁴

Table 3. A Comparison of Structures Generated by This Method vs. Those Generated by MaxSprout.

PDB index	Main-chain RMSD (Å)	
	Present work	MaxSprout
111M	0.31	0.43
1CTF	0.41	0.76
1IGD	0.34	0.49
1OMD	0.39	0.44
1SEM	0.50	13.71
1TIM	0.56	17.31
1UBQ	0.37	0.38
2ALP	0.47	Incomplete structure
2CTS	0.37	0.46
2LYM	0.32	0.46
2MHR	0.33	0.56
2PRK	0.37	0.45
2WRP	0.24	Incomplete structure
3TS1	0.40	0.43
4PTI	0.51	0.44
5NLL	0.42	0.46
6PTI	0.46	Incomplete structure
9WGA	0.52	Incomplete structure
Mean ^a	0.39/0.38 ^b	0.48
Std. Dev. ^a	0.08/0.05 ^b	0.10

The remaining structures in the test set (1CRN, 3APP, and 5CPA) are present in the fragment database utilized by MaxSprout, and were therefore ignored.

^aIgnoring 1TIM and 1SEM results.

^bExcluding results for structures that MaxSprout was unable to satisfactorily construct.

Effect of input coordinate distortions on output accuracy

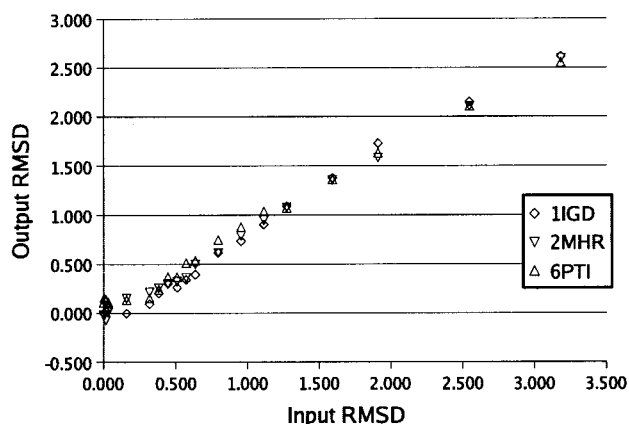


Figure 3. The effect of errors in the input C_{α} trace. The output RMSD is the excess RMSD when compared to the model created with the initial C_{α} coordinates.

and may be anticipated as being equally useful for sampling the backbone conformational energy surface.

An interesting observation is that the backbone predictions are remarkably accurate, even in the absence of side-chain interactions. This almost certainly results from use of the effective forcefield. To

Table 4. The Correlation between Predicted and Measured Dihedrals.

PDB index	Correlation	
	ψ	ϕ
111M	0.94	0.82
1CRN	0.99	0.89
1CTF	0.99	0.89
1IGD	0.97	0.96
1OMD	0.96	0.93
1SEM	0.91	0.86
1TIM	0.91	0.76
1UBQ	0.97	0.96
2ALP	0.96	0.94
2CTS	0.94	0.82
2LYM	0.95	0.96
2MHR	0.93	0.71
2PCY	0.94	0.89
2PRK	0.96	0.92
2WRP	0.99	0.89
3APP	0.94	0.93
3TS1	0.94	0.80
4PTI	0.96	0.93
5CPA	0.94	0.88
5NLL	0.95	0.87
6PTI	0.97	0.83
9WGA	0.96	0.91
Mean	0.95	0.88

Terminal residues are ignored.

Table 5. C_{β} Reconstruction Results.

PDB index	Main-chain RMSD (Å)	
	C_{α}	C_{β}
111M	0.31	0.53
1CRN	0.44	0.57
1CTF	0.41	0.68
1IGD	0.34	0.54
1OMD	0.39	0.64
1SEM	0.50	0.71
1TIM	0.56	0.79
1UBQ	0.37	0.62
2ALP	0.47	0.74
2CTS	0.37	0.67
2LYM	0.32	0.51
2MHR	0.33	0.69
2PCY	0.48	0.65
2PRK	0.37	0.76
2WRP	0.24	0.54
3APP	0.40	0.70
3TS1	0.40	0.71
4PTI	0.51	0.81
5CPA	0.48	0.72
5NLL	0.42	0.65
6PTI	0.46	0.66
9WGA	0.52	0.91
Mean	0.41	0.67

some degree, backbone-side-chain interactions are included in an average sense. The evaluation of native vs. decoy structures was not significantly disrupted by the use of backbone atoms only. In fact, the Z-scores for several protein sets were improved.

The high level of accuracy of the predictions illustrates that the SPLIFF-AA forcefield captures the essential physical interactions that dominate backbone structures. As far as the SPLIFF-AA forcefield is concerned, one major deficit that is currently exhibited is the lack of directional components in the interactions. For example, hydrogen bonds are directional, but only the distances are significant in the atom-based terms. On a larger scale, the direction of interaction between two residues can be important. This may be captured to some extent by the multiple minima in the sum of potentials between multiple atoms, but a study into direction-dependent SPLIFF-AA terms may be worthwhile. Additionally, rotamer specific potentials have been suggested, and may warrant investigations.⁴⁵ Another limitation may be the lack of any distinction between buried and exposed atoms. However, adding this to the FF introduces considerable computational overhead without any appreciable gain in the quality of predictions.

The terminal residues have significantly worse RMSD deviations, on average, than nonterminal residues (results not shown). The author relies on subsequent minimization to resolve this minor issue.

Summary

The protocol described in this report is able to restore the full backbone coordinates to within 0.2–0.6 Å of the actual crystal

structure from C_{α} coordinates only. The entire process is rapid, taking of the order of a few minutes for a typical protein on a typical desktop PC. This compares very favorably to previously published methods, being both one of the fastest and one of the most accurate methods.

When compared to the widely used MaxSprout server, the present approach yields backbones with RMSD deviations from the native structure of approximately 20% less.

The software implemented for this work is available from <http://mccammon.ucsd.edu/~adcock/bb.html>, along with the potential function data files.

Acknowledgments

The author gratefully acknowledges Professor J. A. McCammon for useful and enlightened advice. The author is also grateful to Dr. J. E. Nielsen for thoughtful comments and suggestions regarding the manuscript.

References

1. Jones, T.; Thirup, S. *EMBO J* 1986, 5, 819.
2. Reid, L.; Thornton, J. *Proteins Struct Funct Genet* 1989, 5, 170.
3. Claessens, M.; van Cutsem, E.; Lasters, I.; Wodak, S. *Protein Eng* 1989, 2, 335.
4. Holm, L.; Sander, C. *J Mol Biol* 1991, 218, 183.
5. Levitt, M. *J Mol Biol* 1992, 226, 507.
6. Payne, P. *Protein Sci* 1993, 2, 315.
7. Rey, A.; Skolnick, J. *J Comput Chem* 1992, 13, 443.
8. Bassolino-Klimas, D.; Buccoleri, R. *Proteins Struct Funct Genet* 1992, 14, 465.
9. Milik, M.; Kolinski, A.; Skolnick, J. *J Comput Chem* 1997, 18, 80.
10. Feldman, H.; Hogue, C. *Proteins Struct Funct Genet* 2000, 39, 112.
11. Liwo, A.; Pincus, M.; Wawak, R.; Rackovsky, S.; Scheraga, H. *Protein Sci* 1993, 2, 1697.
12. Iwata, Y.; Kasuya, A.; Miyamoto, S. *J Mol Graph Model* 2002, 21, 119.
13. Correa, P. *Protein Struct Funct Genet* 1990, 7, 366.
14. van Gelder, C.; Leusen, F.; Leunissen, J.; Noordik, J. *Proteins Struct Funct Genet* 1994, 18, 174.
15. van Hoof, P.; Holtje, H.-D. *J Computer-Aided Mol Des* 2000, 14, 719.
16. Mathiowetz, A.; Goddard, W., III. *Protein Sci* 1995, 4, 1217.
17. Gan, K.; Coxon, J.; McKinnon, A.; Worth, G. *Biopolymers* 1997, 41, 391.
18. Kazmierkiewicz, R.; Liwo, A.; Scheraga, H. *J Comput Chem* 2002, 23, 715.
19. Lee, M.; Duan, Y.; Kollman, P. *Proteins Struct Funct Genet* 2000, 39, 309.
20. Miyazawa, S.; Jernigan, R. L. *Macromolecules* 1985, 18, 534.
21. Sippl, M. *J Mol Biol* 1990, 213, 859.
22. Rooman, M. J.; Kocher, J. P.; Wodak, S. J. *J Mol Biol* 1991, 221, 961.
23. Jones, D. T.; Taylor, W. R.; Thornton, J. *Nature* 1992, 358, 86.
24. Kocher, J. A.; Rooman, M. J.; Wodak, S. J. *J Mol Biol* 1994, 235, 158.
25. Bahar, I.; Jernigan, R. L. *J Mol Biol* 1997, 266, 195.
26. Wodak, S. J. In *Encyclopedia of Computational Chemistry*; Schleyer, P. R., Ed.; John Wiley & Sons: New York, 1998.
27. Thomas, P.; Dill, K. *J Mol Biol* 1996, 257, 457.
28. Jernigan, R. L.; Bahar, I. *Curr Opin Struct Biol* 1996, 6, 195.
29. Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. *Nucleic Acids Res* 2000, 28, 235.
30. Altenbach, C.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res* 1997, 25, 3389.
31. <http://www.ncbi.nlm.nih.gov/IEB/ToolBox/index.cgi>, "NCBI Tool-Box," 2003.
32. McDonald, I.; Thornton, J. *J Mol Biol* 1994, 238, 777.
33. Desmet, J.; De Maeyer, M.; Hazes, B.; Lasters, I. *Nature* 1992, 356, 539.
34. Goldstein, R. *Biophys J* 1994, 66, 1335.
35. Looger, L.; Hellinga, H. *J Mol Biol* 2001, 307, 429.
36. Lasters, I.; Desmet, J. *Protein Eng* 1993, 6, 717.
37. Desmet, J.; De Maeyer, M.; Lasters, I. In *The Protein Folding Problem and Tertiary Structure Prediction*; Merz, K. M., Jr.; Le Grand, S. M., Eds.; Birkhauser: Boston, 1994.
38. Gordon, D.; Mayo, S. *J Comput Chem* 1998, 19, 1505.
39. Kearsley, S. K. *Acta Crystallogr Sect A* 1989, 45, 208.
40. Park, B.; Levitt, M. *J Mol Biol* 1996, 258, 367.
41. Samudrala, R.; Xia, Y.; Huang, E.; Levitt, M. *Proteins Struct Funct Genet* 1999, Suppl. 3, 194.
42. Simons, K.; Kooperberg, C.; Huang, E.; Baker, D. *J Mol Biol* 1997, 268, 209.
43. <http://www.ebi.ac.uk/maxsprout/>, "MaxSprout: Reconstruction of 3D coordinates from C(alpha) trace," 2002.
44. Leach, A. R.; Lemon, A. *Proteins* 1998, 33, 227.
45. Lemak, A.; Gunn, J. *J Phys Chem B* 2000, 104, 1097.
46. Samudrala, R.; Moult, J. *J Mol Biol* 1998, 275, 893.
47. Fain, B.; Xia, Y.; Levitt, M. *Protein Sci* 2002, 11, 2010.