```
sim.set_simulation_parameters(numsave=500, trials=50)
NUMRES = sim.protein.num_residues
sim.set_hbond_parameters(hbond_winmax=NUMRES)
sim.set_contact_parameters(contact_winmax=NUMRES)
sim.run()
```

Acknowledgments

We thank Nicholas Fitzkee, Haipeng Gong, Nicholas Panasik, and Timothy Street for critical reading of the manuscript. Support from the NIH and Mathers Foundation is gratefully acknowledged.

[4] Protein Structure Prediction Using Rosetta

By Carol A. Rohl, Charlie E. M. Strauss, Kira M. S. Misura, and David Baker

Introduction

Double-blind assessments of protein structure prediction methods, held biannually in the community-wide critical assessment of structure prediction (CASP) experiments, have documented significant progress in the field of protein structure prediction and have indicated that the Rosetta algorithm is perhaps the most successful current method for *de novo* protein structure prediction.[1–3] In the Rosetta method, short fragments of known proteins are assembled by a Monte Carlo strategy to yield native-like protein conformations. Using only sequence information, successful Rosetta predictions yield models with typical accuracies of 3–6 Å Cα root mean square deviation (RMSD) from the experimentally determined structures for contiguous segments of 60 or more residues. In such low- to moderate-accuracy models of protein structure, the global topology is correctly predicted, the architecture of secondary structure elements is generally correct, and functional residues are frequently clustered to an active site region. Models obtained by *de novo* prediction methods have been demonstrated to have utility for obtaining biological insight, either through

---

[1] R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C. E. M. Strauss, and D. Baker, *Proteins Struct. Funct. Genet.* **45**(Suppl. 5), 119 (2001).

[2] P. Bradley, D. Chivian, J. Meiler, K. M. S. Misura, C. A. Rohl, W. R. Schief, W. J. Wedemeyer, O. Schueler-Furman, P. Murphy, J. Schonbrun, C. E. M. Strauss, and D. Baker, *Proteins Struct. Funct. Genet.* **53**(Suppl. 6), 457 (2003).

[3] A. M. Lesk, L. Lo Conte, and T. J. Hubbard, *Proteins Struct. Funct. Genet.* **S5,** 98 (2001).

Copyright 2004, Elsevier Inc.
All rights reserved.

0076-6879/04 $35.00

functional site recognition or functional annotation by fold identification.[4,5] The Rosetta method is sufficiently fast to make genome-scale analysis possible: a recent study predicted structures for ≈500 PfamA families with no link to known structure.[6] On the basis of previous performance, one of the five models reported for each Pfam family is expected to be a reasonable match to the true structure for about 50–60% of the families, and many of these predictions suggest a homology unapparent in their sequences

Because of its success in *de novo* structure prediction, the Rosetta method has also been successfully extended to other protein-modeling problems including structure determination using limited experimental constraints,[7,8] *de novo* protein design,[9,10] protein–protein docking,[11] and loop modeling.[12] Structure determination by using Rosetta in combination with limited experimental constraints generally yields structures of higher overall accuracy, often with an RMSD of 2–3 Å over the entire protein. Loop modeling is carried out in the context of a homology-based template that is also frequently only ≈2 Å from the true structure. For design of novel protein structures, sequence selection algorithms require backbone structures of accuracy equivalent to experimentally determined X-ray crystal structures. To address these problems, as well as to refine *de novo* models, improvements to the Rosetta method have focused on increased detail in the potential functions and finer control of chain motion in the search algorithm.

Although *de novo* structure prediction with the Rosetta algorithm has been previously described, here we summarize the current method in its entirety. The benefits and limitations of the fragment assembly strategy utilized by Rosetta are discussed, and we describe adaptations of the Rosetta method for structural modeling with finer resolution. Enhancements to the fragment assembly strategy that allow more local modifications of protein conformation are described, and the effectiveness of

---

[4] R. Bonneau, J. Tsai, I. Ruczinski, and D. Baker, *J. Struct. Biol.* **13,** 186 (2001).

[5] J. A. Di Gennaro, N. Siew, B. T. Hoffman, L. Zhang, J. Skolnick, L. I. Neilson, and J. S. Fetrow, *J. Struct. Biol.* **134,** 232 (2001).

[6] R. Bonneau, C. E. M. Strauss, C. A. Rohl, D. Chivian, P. Bradley, L. Malmström, T. Robertson, and D. Baker, *J. Mol. Biol.* **322,** 65 (2002).

[7] P. M. Bowers, C. E. M. Strauss, and D. Baker, *J. Biomol. NMR* **18,** 311 (2000).

[8] C. A. Rohl and D. Baker, *J. Am. Chem. Soc.* **124,** 2723 (2002).

[9] B. Kuhlman, J. W. O'Neill, D. E. Kim, K. Y. Zhang, and D. Baker, *J. Mol. Biol.* **315,** 471 (2002).

[10] B. Kuhlman, G. Dantas, G. Ireton, G. Varani, B. Stoddard, and D. Baker, *Science* **302,** 1364 (2003).

[11] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker, *J. Mol. Biol.* **331,** 281 (2003).

[12] C. A. Rohl, C. E. M. Strauss, D. Chivian, and D. Baker, *Proteins Struct. Funct. Genet.* in press (2004).

these operators for energy function minimization is illustrated. In addition, in Appendix I we derive a new, efficient approach to screening local moves; that is, finding short sets of torsional angle changes that permit local changes in a protein chain while collectively minimizing global changes. Our formulation is computationally fast while offering better correlation to global distance changes appropriate to the atomic interaction potentials than previous popular methods (e.g., Gunn[12a]). The method is applicable to both the problem of screening discrete moves as well as allowing gradient descent of continuous multiangle moves.

Rosetta Strategy

A guiding principle of the Rosetta algorithm is to attempt to mimic the interplay of local and global interactions in determining protein structure. The method is based on the experimental observation that local sequence preferences bias but do not uniquely define the local structure of a protein. The final native conformation is obtained when these fluctuating local structures come together to yield a compact conformation with favorable nonlocal interactions, such as buried hydrophobic residues, paired β strands, and specific side-chain interactions. In the Rosetta algorithm, the structures sampled by local sequences are approximated by the distribution of structures seen for those short sequences and related sequences in known protein structures: a library of fragments that represent the range of accessible local structures for all short segments of the protein chain are selected from a database of known protein structures. Compact structures are then assembled by randomly combining these fragments, using a Monte Carlo simulated annealing search. The fitness of individual conformations with respect to nonlocal interactions is evaluated on the basis of a scoring function derived from conformational statistics of known protein structures.

Rosetta utilizes a torsion space representation in which the protein backbone conformation is specified as a list of backbone ϕ, ψ, and ω torsion angles. Conformation modification occurs in torsion space, although for purposes of evaluating the energy of the conformation the corresponding Cartesian space protein representation is generated with atomic coordinates for all heavy atoms in the protein backbone, assuming ideal bond lengths and angles for individual residues.[13] Two alternate representations of side chains are utilized depending on the requirements of the energy function in use (see below). For residue-based potential terms, a reduced description is used in which each side chain is represented by a centroid located at the side-chain center of mass ($C\beta$ and beyond). For glycine, the centroid is coincident with the $C\alpha$ atom. Centroid positions for each residue type are determined by averaging over observed side-chain conformations in known protein structures. For increased detail, atomic coordinates for all side-chain atoms, including hydrogens, are utilized. Side chains are restricted to discrete conformations as described by a backbone-dependent rotamer library.[14] Side-chain conformations are added to the backbone structure by means of a Monte Carlo simulated annealing search.[15]

[12a] J. R. Gunn, *J. Chem. Phys.* **106,** 4270 (1997).

[13] R. Engh and R. Huber, *Acta Crystallogr. A* **47,** 392 (1991).

Derivation of the Rosetta scoring function or potential energy surface (PES) is based on a Bayesian separation of the total energy into components that describe the likelihood of a particular structure, independent of sequence, and those that describe the fitness of the sequence given a particular structure.[16,17] The terms in this scoring function in their current form are summarized in Table I. The original Rosetta scoring function uses a fairly coarse-grained or low-resolution description of structure: terms corresponding to solvation and electrostatic effects are based on observed residue distributions in protein structures. Hydrogen bonding is not described explicitly, but probabilistic descriptions of β-strand pairing geometry and β-sheet patterns are included. Steric overlap of backbone atoms and side-chain centroids is penalized, but favorable van der Waals interactions are modeled only by rewarding globally compact structures. The scoring function does not explicitly evaluate local interactions because these interactions are implicitly included in the fragment library (see below).

For applications requiring finer resolution, more detailed descriptions of the determinants of protein structure are needed and have motivated the development of a more physically realistic, atomic-level potential function that attempts to model the primary contributions to stability and structural specificity (Table II). van der Waals interactions are modeled with a 6–12 Lennard–Jones potential, attenuated to a linear function in the repulsive regime to compensate for the discrete rotamer representation of side chains. Solvation effects are included, using the model of Lazaridis and Karplus,[18] and hydrogen bonding is explicitly included, using a secondary structure- and orientation-dependent potential derived from analysis of hydrogen bond geometries in high-resolution protein structures.[19–21]

---

[14] R. L. Dunbrack and F. E. Cohen, *Protein Sci.* **6,** 1661 (1997).

[15] B. Kuhlman and D. Baker, *Proc. Natl. Acad. Sci. USA* **97,** 10383 (2000).

[16] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, *J. Mol. Biol.* **268,** 209 (1997).

[17] K. T. Simons, I. Ruczinski, C. Kooperberg, B. Fox, C. Bystroff, and D. Baker, *Proteins Struct. Funct. Genet.* **35,** 82 (1999).

[18] T. Lazaridis and M. Karplus, *Proteins Struct. Funct. Genet.* **35,** 133 (1999).

[19] T. Kortemme, A. V. Morozov, and D. Baker, *J. Mol. Biol.* **326,** 1239 (2003).

TABLE I
COMPONENTS OF ROSETTA ENERGY FUNCTION[a]

Name	Description (putative physical origin)	Functional form	Parameters (values)
env[b]	Residue environment (solvation)	$\sum_i - \ln [P(\mathrm{aa}_i \| \mathrm{nb}_i)]$	i = residue index aa = amino acid type nb = number of neighboring residues[c] (0, 1, 2… 30, >30)
pair[b]	Residue pair interactions (electrostatics, disulfides)	$\sum_i \sum_{j>i} - \ln \left[\frac{P(\mathrm{aa}_i, \mathrm{aa}_j \| s_{ij} d_{ij})}{P(\mathrm{aa}_i \| s_{ij} d_{ij}) P(\mathrm{aa}_i \| s_{ij} d_{ij})} \right]$	i, j = residue indices aa = amino acid type d = centroid–centroid distance (10–12, 7.5–10, 5–7.5, <5 Å) s = sequence separation (>8 residues)
SS[d]	Strand pairing (hydrogen bonding)	SchemeA : $\mathrm{SS}_{\phi,\theta} + \mathrm{SS}_{hb} + \mathrm{SS}_d$ SchemeB : $\mathrm{SS}_{\phi,\theta} + \mathrm{SS}_{hb} + \mathrm{SS}_{d\sigma}$ where $\mathrm{SS}_{\phi,\theta} = \sum_m \sum_{n>m} - \ln [P(\phi_{mn}, \theta_{mn} \| d_{mn}, \mathrm{sp}_{mn}, s_{mn})]$ $\mathrm{SS}_{hb} = \sum_m \sum_{n>m} - \ln [P(\mathrm{hb}_{mn} \| d_{mn}, s_{mn})]$ $\mathrm{SS}_d = \sum_m \sum_{n>m} - \ln [P(d_{mn} \| s_{mn})]$ $\mathrm{SS}_{d\sigma} = \sum_m \sum_{n>m} - \ln [\mathrm{P}(d_{mn} \sigma_{mn} \| \rho_m, \rho_n)]$	m, n = strand dimer indices; dimer is two consecutive strand residues V = vector between first N atom and last C atom of dimer m = unit vector between $\hat{V}_m$ and $\hat{V}_n$ midpoints x = unit vector along carbon–oxygen bond of first dimer residue y = unit vector along oxygen–carbon bond of second dimer residue ϕ, θ = polar angles between $\hat{V}_m$ and $\hat{V}_n$ (36° bins) hb = dimer twist, $\sum_{k=m,n} 0.5(\|\hat{m} \cdot \hat{x}_k\| + \|\hat{m} \cdot \hat{y}_k\|)$ (< 0.33, 0.33–0.66, 0.66–1.0, 1.0–1.33, 1.33–1.6, 1.6–1.8, 1.8–2.0) d = distance between $\hat{V}_m$ and $\hat{V}_n$ midpoints (< 6.5 Å) σ = angle between $\hat{V}_m$ and $\hat{M}$ (18° bins) sp = sequence separation between dimer-containing strands (< 2, 2–10, > 10 residues) s = sequence separation between dimers (>5 or >10) ρ = mean angle between vectors $\hat{m}$, $\hat{x}$ and $\hat{m}$, $\hat{y}$ (180° bins)

sheet[e]	Strand arrangement into sheets	$-\ln[P(n_{\text{sheets}} n_{\text{lonestrands}} \mid n_{\text{strands}})]$	n_{sheets} = number of sheets $n_{\text{lone strands}}$ = number of unpaired strands n_{strands} = total number of strands
HS	Helix–strand packing	$\sum_m \sum_n -\ln[P(\phi_{mn}, \psi_{mn} \mid sp_{mn} d_{mn})]$	m = strand dimer index; dimer is two consecutive strand residues n = helix dimer index; dimer is central two residues of four consecutive helical residues $\hat{V}$ = vector between first N atom and last C atom of dimer ϕ, θ = polar angles between $\hat{V}_m$ and $\hat{V}_n$ (36° bins) sp = sequence separation between dimer-containing helix and strand (binned < 2, 2–10, >10 residues) d = distance between $\hat{V}_m$ and $\hat{V}_n$ midpoints (< 12 Å)
rg	Radius of gyration (vdw attraction; solvation)	$\sqrt{\langle d_{ij}^2 \rangle}$	i, j = residue indices d = distance between residue centroids
cbeta	Cβ density (solvation; correction for excluded volume effect introduced by simulation)	$\sum_i \sum_{sh} -\ln\left[\frac{P_{\text{compact}}(\text{nb}_{i,sh})}{P_{\text{random}}(\text{nb}_{i,sh})}\right]$	i = residue index sh = shell radius (6, 12 Å) nb = number of neighboring residues within shell[f] P_{compact} = probability in compact structures assembled from fragments P_{random} = probability in structures assembled randomly from fragments
vdw[g]	Steric repulsion	$\sum_i \sum_{j>i} \frac{\left(r_{ij}^2 - d_{ij}^2\right)^2}{r_{ij}}; \; d_{ij} < r_{ij}$	i, j = residue (or centroid) indices d = interatomic distance r = summed van der Waals radii[h]

[a] All terms originally described in Refs. 16 and 17.

[b] Binned function values are linearly interpolated, yielding analytic derivatives.

(continued)

TABLE I (*continued*)

[c] Neighbors within a 10-Å radius. Residue position defined by Cβ coordinates (Cα for glycine).

[d] Interactions between dimers within the same strand are neglected. Favorable interactions are limited to preserve pairwise strand interactions, that is, dimer m can interact favorably with dimers from at most one strand on each side, with the most favorable dimer interaction ($\mathrm{SS}_{\phi,\theta} + \mathrm{SS}_{hb} + \mathrm{SS}_d$) determining the identity of the interacting strand. $\mathrm{SS}_{d\sigma}$ is exempt from the requirement of pairwise strand interactions. SS_{hb} is evaluated only for m, n pairs for which $\mathrm{SS}_{\phi,\theta}$ is favorable. $\mathrm{SS}_{d\sigma}$ is evaluated only for m, n pairs for which $\mathrm{SS}_{\phi,\theta}$ and $\mathrm{S\tilde{S}}_{hb}$ are favorable. A bonus is awarded for each favorable dimer interaction for which $|m - n| > 11$ and strand separation is more than eight residues.

[e] A sheet is composed of all strands with dimer pairs <5.5 Å apart, allowing each strand having at most one neighboring strand on each side. Discrimination between alternate strand pairings is determined according the most favorable dimer interaction. Probability distributions fitted to $c(n_{\text{strands}}) - 0.9n_{\text{sheets}} - 2.7n_{\text{lone strands}}$ where $c(n_{\text{strands}}) = (0.07, 0.41, 0.43, 0.60, 0.61, 0.85, 0.86, 1.12)$.

[f] Residue position defined by Cβ coordinates (Cα for glycine).

[g] Not evaluated for atom (centroid) pairs whose interatomic distance depends on the torsion angles of a single residue.

[h] Radii determined from (1) 25th closest distance seen for atom pair in pdbselect25 structures, (2) the fifth closest distance observed in X-ray structures with better than 1.3-Å resolution and <40% sequence identity, or (3) X-ray structures of <2-Å resolution, excluding $i, i + 1$ contacts (centroid radii only).

TABLE II
COMPONENTS OF ROSETTA ALL-ATOM ENERGY FUNCTION[a]

Name	Description (physical origin)	Functional form	Parameters	Ref.
rama	Ramachandran torsion preferences	$\sum_i -\ln[P(\phi_i, \psi_i \| \mathrm{aa}_i, \mathrm{ss}_i)]$	i = residue index ϕ,ψ = backbone torsion angles (36° bins) aa = amino acid type ss = secondary structure type[b]	7, 12
LJ[c]	Lennard–Jones interactions	$\sum_i \sum_{j>i} \begin{cases} \left[\left(\frac{r_{ij}}{d_{ij}}\right)^{12} - 2\left(\frac{r_{ij}}{d_{ij}}\right)^{6}\right] e_{ij}, & \text{if } \frac{d_{ij}}{r_{ij}} > 0.6 \\ \left[-8759.2\left(\frac{d_{ij}}{r_{ij}}\right) + 5672.0\right] e_{ij}, & \text{else} \end{cases}$	i, j = residue indices d = interatomic distance e = geometric mean of atom well depths[d] r = summed van der Waals radii[e]	15
hb[f]	Hydrogen bonding	$\sum_i \sum_j (-\ln[P(d_{ij} \| h_j \mathrm{ss}_{ij})]$ $-\ln[P(\cos\theta_{ij} \| d_{ij} h_j \mathrm{ss}_{ij})]$ $-\ln[P(\cos\psi_{ij} \| d_{ij} h_j \mathrm{ss}_{ij})])$	i = donor residue index j = acceptor residue index d = acceptor–proton interatomic distance h = hybridization (sp^2, sp^3) ss = secondary structure type[g] θ = proton–acceptor–acceptor base bond angle ψ = donor–proton–acceptor bond angle	19–21
solv	Solvation	$\sum_i \left[\Delta G_i^{\mathrm{ref}} - \sum_j \left(\frac{2\Delta G_i^{\mathrm{free}}}{4\pi^{3/2}\lambda_i r_{ij}^2} e^{-d_{ij}^2} V_j + \frac{2\Delta G_i^{\mathrm{free}}}{4\pi^{3/2}\lambda_j r_{ij}^2} e^{-d_{ij}^2} V_i\right)\right]$	i,j = atom indices d = distance between atoms r = summed van der Waal radii[e] λ = correlation length[h] V = atomic volume[h] $\Delta G^{\mathrm{ref}}, \Delta G^{\mathrm{free}}$ = energy of a fully solvated atom[h]	15, 18

(continued)

TABLE II *(continued)*

pair	Residue pair interactions (electrostatics, disulfides)	$\sum_i \sum_{j>i} -\ln\left[\frac{P(\mathrm{aa}_i,\ \mathrm{aa}_j \mid d_{ij})}{P(\mathrm{aa}_i \mid d_{ij})P(\mathrm{aa}_i \mid d_{ij})}\right]$	i, j = residue indices aa = amino acid type d = distance between residues[i]	15
dun	Rotamer self-energy	$\sum_i -\ln\left[\frac{P(\mathrm{rot}_i \mid \phi_i, \psi_i)P(\mathrm{aa}_i \mid \phi_i, \psi_i)}{P(\mathrm{aa}_i)}\right]$	i, j = residue indices rot = Dunbrack backbone-dependent rotamer aa = amino acid type ϕ, ψ = backbone torsion angles	14, 15
ref	Unfolded state reference energy	$\sum_{\mathrm{aa}} n_{\mathrm{aa}}$	aa = amino acid type n = number of residues	15

---

[a] All binned function values are linearly interpolated, yielding analytic derivatives, except as noted.

[b] Three-state secondary structure type as assigned by DSSP.[22]

[c] Not evaluated for atom pairs whose interatomic distance depends on the torsion angles of a single residue.

[d] Well depths taken from CHARMm19 parameter set.[23]

[e] Radii determined from fitting atom distances in protein X-ray structures to the 6–12 Lennard–Jones potential using CHARMm19 well depths.

[f] Evaluated only for donor acceptor pairs for which $1.4 \le d \le 3.0$ and $90^\circ \le \psi, \theta \le 180^\circ$. Side-chain hydrogen bonds in involving atoms forming main-chain hydrogen bonds are not evaluated. Individual probability distributions are fitted to eighth-order polynomials and analytically differentiated.

[g] Secondary structure types for hydrogen bonds are assigned as helical ($j - i = 4$, main chain); strand ($|j - i| > 4$, main chain), or other.

[h] Values taken from Lazaridis and Karplus.[18]

[i] Residue position defined by $C\beta$ coordinates ($C\alpha$ of glycine).

Electrostatics are modeled using a residue-based pair potential, similar to that utilized at low resolution. In addition, because finer modifications of conformation permit introduction of angles not part of the original discrete fragment set (see below), energetic description of local interactions is included in the total energy, including an amino acid- and secondary structure-dependent torsion potential for the backbone dihedral angles.

De Novo Structure Prediction with Rosetta

Fragment Selection

The basic conformation modification operation employed by Rosetta is termed a "fragment insertion." For each fragment insertion, a consecutive window of three or nine residues is selected, and the torsion angles of these residues are replaced with the torsion angles obtained from a fragment of a protein of known structure. For each query sequence to be predicted, a customized library of fragments defining the conformational space to be searched is selected by comparison of short windows of the query sequence with known protein structures. All three- and nine-residue windows in the query are scored against all windows in a nonredundant database of proteins of known structure composed of X-ray structures of 2.5 Å resolution or better and <50% sequence identity. All bond lengths and bond angles in these structures have been set to ideal values and the backbone torsion angles fine tuned using small perturbations to maintain agreement with the X-ray-determined atomic coordinates, minimize steric overlap (Table I; vdw), and maintain favorable values of backbone torsion angles as evaluated by the torsion potential (Table II; rama).

Sequence profiles for the query sequence and each sequence in the structure database are constructed by two rounds of PSIBLAST[24] with a cutoff of 9×10^{-4}. Over each sequence window, a profile–profile similarity score is calculated as the sum of the absolute value of the differences of the probabilities of each amino acid at each position (L1 norm, widely known as the city block or taxi cab distance). In addition, the predicted secondary structure of the query sequence is compared with the DSSP[22]-assigned secondary structure of the known structure in each sequence window. Currently, three secondary structure predictions are utilized: Psipred,[25]

---

[20] W. Wedemeyer and D. Baker, *Proteins Struct. Funct. Genet.* **53,** 262 (2003).

[21] J. Schonbrun, C. A. Rohl, and D. Baker, unpublished results (2003).

[22] W. Kabsch and C. Sander, *Biopolymers* **22,** 2577 (1983).

[23] E. Neria, S. Fischer, and M. Karplus, *J. Chem. Phys.* **105,** 1902 (1996).

[24] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, *Nucleic Acids Res.* **25,** 3389 (1997).

SAM-T99,[26] and JUFO.[27] The similarity score for each secondary structure prediction is calculated as the negative sum of the three-state confidence for the correct secondary structure type at each position in the sequence window. For each secondary structure prediction, the overall similarity score is the sum of the sequence similarity score and half the secondary structure similarity score. Fragments containing backbone torsion angles inconsistent with the torsional preferences of the residue types specified by the query sequence are also discarded (e.g., *cis* peptide bonds are allowed only at proline residues in the query sequence).

A ranked list of the top fragments in each sequence window is assembled iteratively, adding the top scoring fragment according to each secondary structure prediction to the combined ranked list and eliminating redundancies. Fragments selected according to Psipred secondary structure prediction are incorporated into the list with a threefold greater frequency than fragments selected on the basis of other secondary structure predictions. As this round robin assembly of the fragment list proceeds, the proportion of helix, strand, and other secondary structure types at each residue is balanced to be consistent with the average three-state prediction of all secondary structure predictions utilized, supplementing the final list as needed with fragments with the desired secondary structure type at a particular position, ranked according to their agreement with the average secondary structure prediction and sequence profile. The final fragment list for a query sequence is composed of 200 nine-residue and 200 three-residue fragments for every overlapping insertion window in the query.

Fragment Assembly

The assembly of fragments into protein-like structures occurs by a Monte Carlo search. The search is arbitrarily started with the protein in a fully extended conformation. A 9-residue fragment insertion window is randomly selected and a fragment for this window is randomly selected from the top 25 fragments in the ranked list for this position. After replacing the torsion angles in the protein chain with the torsion angles from the selected fragment, the energy of the resulting conformation is evaluated. Moves that decrease the energy are retained; those that increase the energy are retained according to the Metropolis criterion. If no moves are accepted in 150 attempted insertions, the probability of accepting a move of increased energy is incrementally increased. After an accepted

---

[25] D. T. Jones, *J. Mol. Biol.* **292,** 195 (1999).

[26] K. Karplus, R. Karchin, C. Barrett, S. Tu, M. Cline, M. Diekhans, L. Grate, J. Casper, and R. Hughey, *Proteins Struct. Funct. Genet.* **S5,** 86 (2001).

[27] J. Meiler, M. Müller, A. Zeidler, and F. Schmäschke, *J. Mol. Model* **7,** 360 (2001).

move, the acceptance probability is returned to its initial value. Each simulation begins from a different random seed and attempts 28,000 nine-residue fragment insertions.

The complete scoring function used for *de novo* prediction is given in Table I. During the course of the simulated annealing protocol, terms are progressively added to the total potential. Initially, only the steric overlap term (Table I; vdw) is evaluated, and this stage continues until all initial torsion angles have been replaced. Over the next 2000 fragment insertion attempts, secondary structure is accumulated in the chain, and all terms except those rewarding compactness (Table I; cbeta, rg) are evaluated. The strand pairing score is evaluated according to Scheme A (Table I; SS) at 0.3 of its final weight. For the next 20,000 attempted fragment insertions, the SS score is increased to its full weight, and the cbeta term is added at half its final weight. During this stage of the simulation, the SS score alternates every 2000 cycles between encouraging local strand pairing and relaxing it. This relaxation in strand pairing requirements is accomplished by evaluating interactions only between residues separated by more than 10 residues in sequence. For the last 6000 attempted moves, strand pairing is encouraged, using the standard sequence separation cutoff of 5 residues (see Table I). For the final 4000 attempted moves, the complete scoring function as described in Table I is utilized with all terms at their full weight. The SS score is evaluated according to Scheme B, using the 5-residue sequence separation cutoff. After the assembly of decoy structures from 9-residue fragments, each decoy is subjected to a short refinement of 8000 attempted 3-residue fragment insertions of the "gunn" type (see below), using the complete scoring function.

For each structure prediction, many short simulations starting from different random seeds are carried out to generate an ensemble of "decoy" structures that have both favorable local interactions and protein-like global properties. This set is then clustered by structural similarity to identify the broadest free energy minima; the structure predictions for a sequence are generally the centers of the largest clusters.[6] Examples of successful predictions made by means of this strategy at CASP 5 are shown in Fig. 1.

Structure Prediction by Fragment Assembly

The fragment assembly approach has multiple benefits for *de novo* protein structure prediction. First, and foremost, the fragment library approximates Gibbs sampling of the populated regions of the local potential energy surface of the backbone. The Rosetta philosophy is that during the folding process of real proteins, the local structure fluctuates between

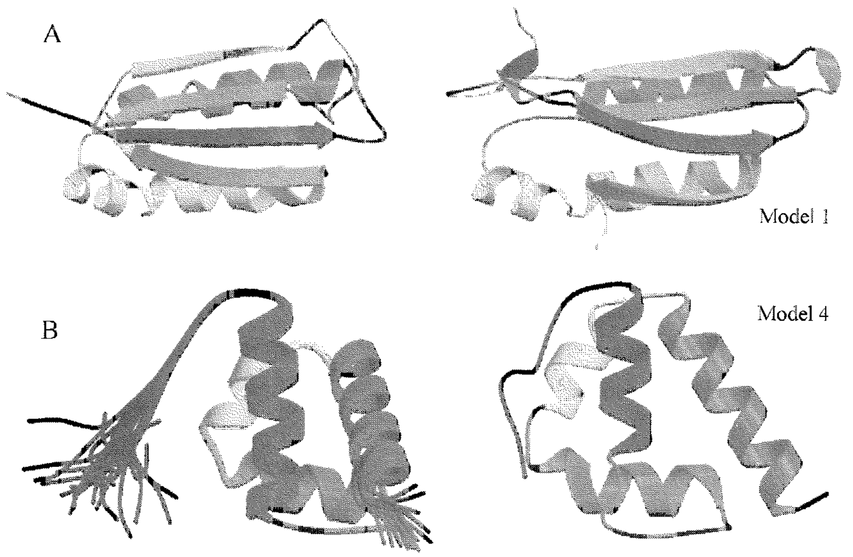


FIG. 1. Rosetta-predicted protein structures for CASP 5 targets. *Right:* Models predicted using the *de novo* prediction protocol. *Left:* Experimental structure of each protein. Protein chains are colored in a blue-to-red gradient along the length of the chain to highlight correctly predicted secondary structure elements. (A) T0135. The predicted model has 54 residues (of 106 total) predicted at a Cα RMSD of 4 Å to the experimental structure. (B) T0171. The predicted model has 60 residues (of 69 total) predicted at a Cα RMSD of 4 Å to the experimental structure. The global Cα RMSD between the prediction and the experimental structure is 4.2 Å. (See color insert.)

alternative local conformations and each fragment is a likely conformation of the local sequence. The use of a preset library of low-energy local structures means the local interaction energy need not be explicitly calculated with each move. This simplification is both efficient and crucial; computing the interaction energy assumes that an accurate potential energy surface is known, which may not be possible. Fragments, on the other hand, allow an accurate, but implicit, representation of the potential energy surface for local interactions. In the Rosetta fragment move set, a single-fragment substitution moving the protein from one topological isomer to another is like instantly transporting from one local energy minimum on the local PES to another; something a more continuous molecular dynamics or gradient search algorithm would be hard pressed to mimic.

Having dispensed with the need for an accurate local PES, the remaining global PES in Rosetta can be coarse grained in distance, and discrete in the combinatorics of strand pairing (Table I). Such a potential is well suited to the large search space inherent in the folding problem, and a second major advantage of the fragment insertion strategy is that it trades precise atomic positioning in favor of rapidly and coarsely sampling the large conformational space. Because fragment insertion modifies a consecutive set of backbone torsion angles, the effect of a move is not localized as it would be with a Cartesian move: the orientation and displacement of atoms on either side of an insertion position can change dramatically as a result of a single backbone torsion angle rotation. Angular changes made in the insertion window are not continuous, but rather are selected from a discrete library. Multiple angles are changed simultaneously, and the original values of the torsion angles being changed are not considered when selecting the angles that will be used for the new move. Consequently, the angular changes from move to move can be quite dramatic, allowing the conformation to evolve rapidly and escape local minima.

The single-fragment insertion approach makes many global conformers dynamically inaccessible on the search trajectory. In effect, the space of accessible conformations is cut off and thereby dramatically reduced. Whether this effect is beneficial depends only on whether the set of included conformations contains a close neighbor of the native protein conformer. The Rosetta philosophy has been shaped by empirical observations that the folding dynamics of protein domains (see e.g., Plaxco *et al.*[28]) are consistent with a process dominated by a quick quench: proteins may repeatedly quickly collapse from an extended chain to a compact structure and either unfold if the structure is distant from the native conformation or, in the rare case that the chain collapses to the native free energy basin, stay folded rather than sampling many energy basins while in a compact form. The "single move at a time" philosophy is not consistent with the physical fact that all the torsion angles are free to move simultaneously, but empirically and intuitively it does bias the final set of accessible structures toward the set achievable by a rapid, unorchestrated collapse.

Enhancements of Fragment Insertion Strategy

For *de novo* fold prediction, the benefits of fragment insertion allow rapid convergence on collapsed structures of plausible topology. Once this initial collapse has occurred, however, the fragment insertion strategy hinders efficient model refinement. Within a compact structure, any randomly selected, rigid body transformation of part of the chain is likely to create a clash with neighboring atoms or break favorable contacts. In addition, once the structure is coarsely established, the scale of conformation modification must be appropriate for optimizing more fine-grained potentials. It is thus desirable to define conformation modification operators for which the scale of global perturbation can be adjusted. Single-fragment insertion also lacks the notion that two or more consecutive moves taken together might offset the harmful effect of each other, yielding a net improvement. Each single insertion is rejected or accepted on the basis of the new energy of the entire conformation, and two moves can be coupled only indirectly via the Metropolis acceptance criteria: a move that increases the energy of the structure is on occasion accepted, setting the stage for a subsequent compensating move. Finally, the Monte Carlo fragment insertion strategy lacks the concept of using the gradient of the potential function to bias conformation modifications toward those that are more likely to be accepted.

---

[28] K. W. Plaxco, I. S. Millet, D. J. Segel, S. Doniach, and D. Baker, *Nat. Struct. Biol.* **6,** 54 (1999).

For modeling scenarios requiring finer sampling about compact structures such as loop modeling, model refinement, and protein design, we have supplemented the original fragment insertion move set of Rosetta with additional conformation modification operators. Five basic concepts are combined to generate these novel operators: (1) random torsion angle perturbation, (2) selection of globally nonperturbing fragments, (3) rapid torsion angle optimization to offset global backbone perturbations, (4) optimization of the scoring function by gradient descent after a backbone modification (Monte Carlo plus minimization), and (5) rapid optimization of side-chain rotamers. In some of these operations, detailed below, deviations from backbone torsion angles of the fragment library are permitted, but such deviations are small so that the assumption that the fragment structures approximate low-energy local interactions is not violated.

Random Angle Perturbation

The simplest approach to local sampling about a compact structure is to perturb the torsion angles from their current values. We employ either small perturbations of randomly selected (ϕ, ψ) pairs ("small") or perturbation of a randomly selected ϕ angle coupled with a compensating rotation of equal magnitude but opposite direction of the preceding ψ angle. The latter case effects a "shear" motion in which the intervening peptide plane is rotated with minimal perturbation to the rest of the chain. Modification of residues in α helices is not allowed, and random perturbations have an upper limit of 2° for residues in β strands and 3° for all other residues. In addition, perturbations that increase the Ramachandran score (Table II; rama) are discriminated against, using a Metropolis criterion.

Selection of Globally Nonperturbing Fragments: A Local Move

The second approach retains reliance on fragment insertion, but biases fragment selection toward those that are most similar to the existing fragment in the model. Two different methods are used to estimate the similarity of two fragments. In the "chuck" strategy, the relative displacement of all the atoms in the protein on fragment replacement is measured, whereas in the "gunn" method, the net rotation and translation effected by different fragments are directly compared.

In the chuck method of fragment selection, the rigid body displacement of the downstream chain resulting from a fragment replacement is computed. The smaller the mean square deviation (MSD) of the atoms in the rigid body, the more the move is regarded as local. The MSD computation can be done extremely efficiently for a library of candidate fragments by relying on the fact that the differential rotation and translation need not be applied to every atom in order to compute the MSD, but only to the inertial ellipsoid of the rigid body (Appendix I). Further advantage is obtained by pretabulating the rotation and translation implied by every fragment in the discrete library. The library is winnowed to those fragments with a total downstream MSD change below a specified threshold, and a fragment in this set is then chosen at random for insertion.

In the gunn strategy, the rotation and translation effected by a fragment on the downstream portion of the chain are summarized by six degrees of freedom that are defined such that they are independent of the absolute origin and orientation of the coordinate system. Consequently, two fragments with similar net rotations and translations will have six-parameter descriptions that are almost numerically equal. The parameterization originally described by Gunn[12a] is used, but the arbitrary cost function is chosen such that large parameter differences are attenuated more than smaller ones: in other words, a fragment with five closely matching parameters and one poor match is preferred over a fragment with six mediocre matches (Appendix II). The gunn cost is computed for each library fragment at the selected insertion window, and a random selection is made among fragments with costs lower than the specified threshold.

Torsion Angle Variation to Offset Global Perturbation

In the "wobble" operation, the global perturbation of a fragment insertion (or any initial conformation modification) is offset by continuous variation of backbone (ϕ, ψ) angles within or adjacent to the insertion window. This operation is similar to the chuck strategy in that the MSD of downstream atoms is the measure of global perturbation. The wobble gradient descent is accelerated by analytic derivatives and, conveniently, our choice of the downstream MSD is differentiable and efficiently computed (Appendix I). In addition to the MSD, the cost function for optimization includes the Ramachandran potential (Table II; rama) that is derived from a smoothed, highly flattened version of the residue- and secondary structure-specific frequency with which a given (ϕ, ψ) pair occurs. The flattening deliberately weakens discrimination in allowed (ϕ, ψ) regions so that the MSD term dominates the result. The smoothing erases local minima and creates gradients leading from low-frequency areas to allowed regions. The look-up table is linearly interpolated between adjacent bins, making the derivative analytic.

The wobble operation is typically combined with fragment insertion in order to generate a complete conformation modification operator. In the standard combination, a fragment is selected by the chuck method, using an MSD cutoff of 60 Å^2. Subsequently, the torsion angles of a residue at the edge of the insertion window are modified by a wobble operation (wobble move in Fig. 3). Multiple fragment insertions and wobble operations can also be combined to generate more complex modification operators. The "crank" move in Fig. 2 is one such example. The move is initiated by making a chuck insertion at a selected insertion window. Torsion angles of an adjacent residue are then perturbed, using the wobble operation. Finally, at a second site not adjacent to the insertion window, torsion angles of two additional residues are perturbed by a second wobble operation. This type of move also attempts to select insertion windows where perturbations are more likely to be tolerated, biasing selection to residues not part of regular secondary structure elements. We note that the chuck methodology can be easily extended to double-fragment insertions, although in practice we find such operations inefficient and prefer to combine chuck insertions with wobble operations.

Monte Carlo Plus Minimization

Finally, any of the above described operations can be combined with direct optimization of the Rosetta potential energy function, replacing the Monte Carlo search strategy with the Monte Carlo-plus-minimization strategy described by Li and Scheraga.[29] After application of the initial modification, we attempt to rescue conformations with slightly increased energy by gradient descent to a local minimum. Either a single line minimization is carried out along the initial gradient ("lin"; Fig. 3) or an iterative descent to the local minimum is employed, using the variable metric method of Davidon, Fletcher, and Powell ("dfp"; Fig. 3).[30] In either case,

---

[29] Z. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **84,** 6611 (1987).

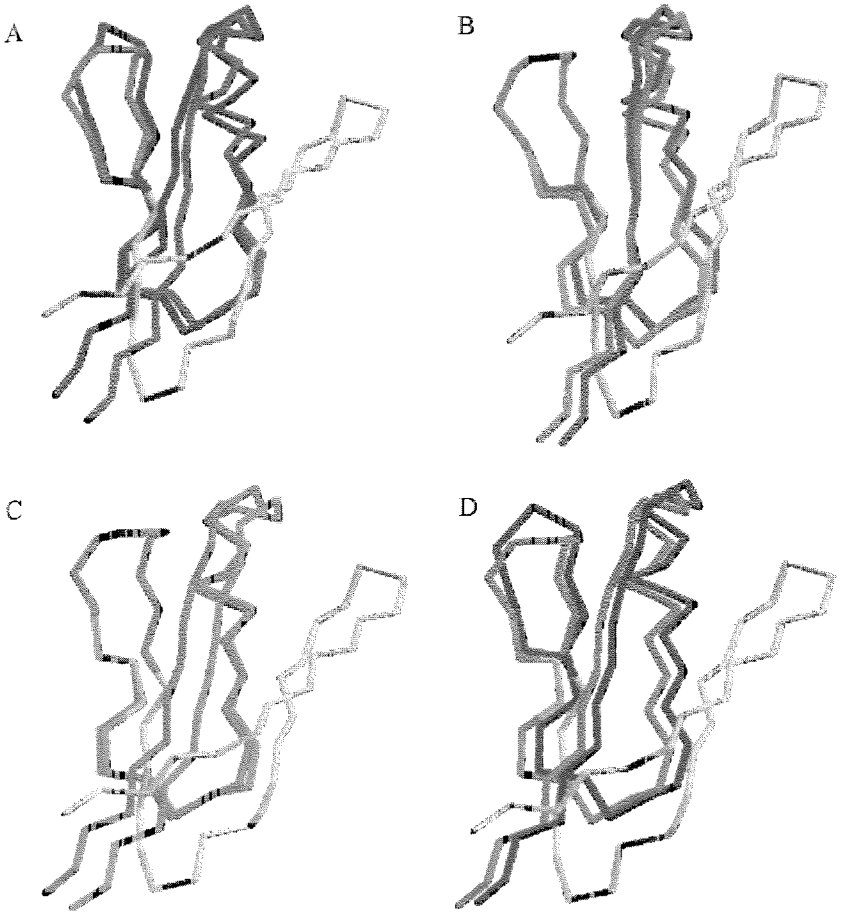
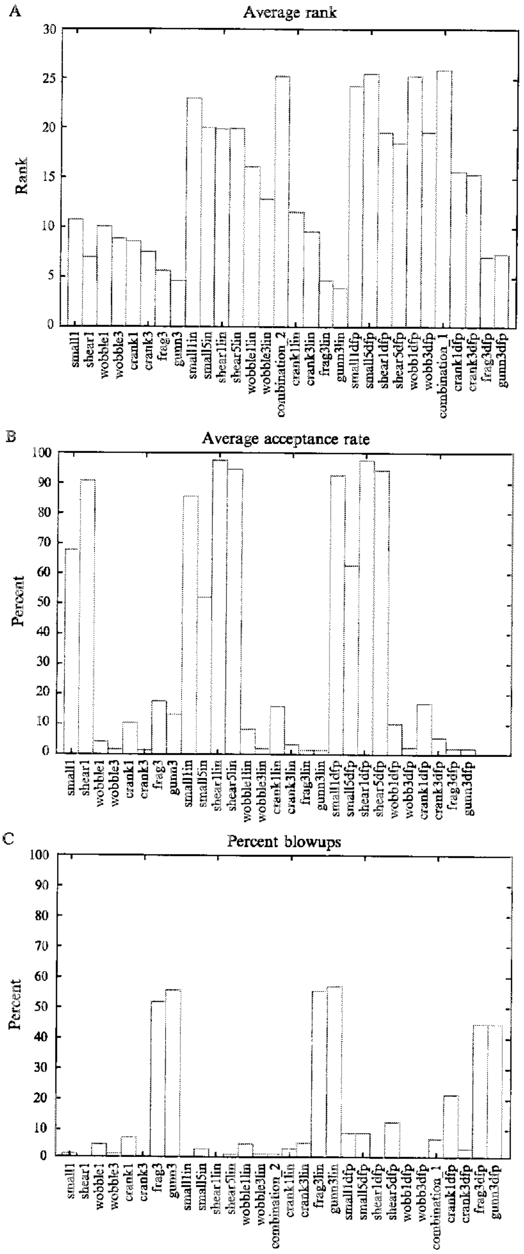


FIG. 2. Modified "crank" fragment insertion into 1 dan. (A) Superposition of the protein conformations preceding (black) and following (blue) insertion of a nine-residue fragment. The fragment insertion window is shown in red. The portion of the chain unperturbed by insertion is shown in gray. (B) Superposition of the protein conformations preceding (blue) and following (green) optimization of angles at a wobble site (cyan) adjacent to the insertion window. (C) Superposition of the protein conformations preceding (green) and following (magenta) optimization of angles at a second wobble site (orange) nonadjacent to the insertion window. (D) Superposition of the original (black) and final (magenta) conformations. (See color insert.)

---

[30] W. H. Press, S. A. Teukolski, W. T. Vetterling, and B. P. Flannery, "Numerical Recipes in Fortran 77: The Art of Scientific Computing," 2nd Ed. Cambridge University Press, New York, 2001.


A
Average rank
Rank
0
5
10
15
20
25
30
small1
shear1
wobble1
wobble3
crank1
crank3
frag3
gumm3
small1lin
small5lin
shear1lin
shear5lin
wobble1lin
wobble3lin
combination_2
crank1lin
crank3lin
frag3lin
gumm3lin
small1dfp
small5dfp
shear1dfp
shear5dfp
wobb1dfp
wobb3dfp
combination_1
crank1dfp
crank3dfp
frag3dfp
gumm3dfp
B
Average acceptance rate
Percent
0
10
20
30
40
50
60
70
80
90
100
small1
shear1
wobble1
wobble3
crank1
crank3
frag3
gumm3
small1lin
small5lin
shear1lin
shear5lin
wobble1lin
wobble3lin
crank1lin
crank3lin
frag3lin
gumm3lin
small1dfp
small5dfp
shear1dfp
shear5dfp
wobb1dfp
wobb3dfp
crank1dfp
crank3dfp
frag3dfp
gumm3dfp
C
Percent blowups
Percent
0
10
20
30
40
50
60
70
80
90
100
small1
shear1
wobble1
wobble3
crank1
crank3
frag3
gumm3
small1lin
small5lin
shear1lin
shear5lin
wobble1lin
wobble3lin
combination_2
crank1lin
crank3lin
frag3lin
gumm3lin
small1dfp
small5dfp
shear1dfp
shear5dfp
wobb1dfp
wobb3dfp
combination_1
crank1dfp
crank3dfp
frag3dfp
gumm3dfp

(ϕ, ψ) torsion angles within five residues of the site(s) of initial perturbation are varied. Torsion angles that are part of helical segments are held fixed. After minimization of the potential function, moves are accepted according to the usual energy criteria.

Rapid Side-Chain Optimization

The methods described above implement modifications of the backbone torsion angles and are combined in a variety of ways to generate backbone conformation modifications. In addition to modifying backbone conformation, operators must also allow for changes in side-chain conformation. Using a simulated annealing protocol, side-chain rotamers can be completely reoptimized.[15] In combination with operators that modify the backbone conformation, however, side-chain rotamers are rapidly optimized by cycling through each side-chain position in random order and replacing the current rotamer with the lowest energy rotamer available at that position. Combining both backbone and side-chain modification, complete conformation modification operators follow this general progression of steps: (1) initial backbone modification, either by random angle perturbation or selection of a globally nonperturbing fragment, (2) wobble of selected torsion angles to offset the global perturbation caused by the

Fig. 3. Comparison of move types in optimizing the all-atom energy function. Moves are named according to the type of perturbation made and the number of residues in the original perturbation (see text for details): small, random perturbation of one or more nonconsecutive (ϕ, ψ) pairs; shear, random compensating changes in a ϕ angle and the preceding ψ angle; wobble, insertion of a chuck fragment followed by a wobble of one residue; crank, insertion of a chuck fragment followed by a wobble of one residue adjacent to the insertion window and then by a wobble of two residues nonadjacent to the insertion window (illustrated in Fig. 2); frag, unmodified fragment insertion; gunn, insertion of a fragment selected using the gunn strategy. Addition of lin to the move name indicates the move is followed by a single-line minimization along the gradient of the potential function before evaluation of the Metropolis criterion. Addition of dfp indicates the move is followed by variable metric optimization of the potential function before evaluation of the Metropolis criterion. For combination 1, the attempted moves were cycled between small1dfp, small5dfp, shear5dfp, and wobble3dfp. For combination 2, the attempted moves were cycled between small1lin, shear5lin, wobble1lin, and wobble3lin. (A) Average rank of moves. For each starting decoy in the test set, the energies of the lowest energy decoy obtained from application of each move were sorted from highest energy (1) to lowest (30). The histogram reports the average overall decoys for each move type. (B) Percentage of moves accepted. Acceptance rates are reported for each move type, averaged over all decoys. The percentage was scaled on the basis of the percentage of independent simulations that resulted in an expanded structure, in order to account for the dramatic increase in acceptance rate into expanded models relative to compact models. (C) Frequency of simulations resulting in expanded structures. (See color insert.)

initial modification, (3) rapid optimization of side-chain rotamers, and (4) optimization of the scoring function by gradient descent.

Effectiveness of Conformation Modification Operators for Energy Function Optimization

Modified fragment insertions of the gunn type have been incorporated into the *de novo* prediction protocol, as described above, and permit significant optimization of the scoring function that is often accompanied by improvements in decoy accuracy and/or discrimination of near-native decoys.[31] When the Rosetta strategy is combined with structural constraints, experimentally determined by nuclear magnetic resonance (NMR), the incorporation of the modified moves described here is essential for refining initial decoy conformations generated by fragment assembly.[7,8] Experimental constraints provide an unusually effective scoring function for identifying accurate structures, and optimization of the agreement between the experimental data and the model structure nearly always results in models of increased accuracy.

To more generally evaluate the effectiveness of the modified move sets described here, we applied each move in isolation to a small test set of model structures and evaluated the frequency with which moves were accepted and the overall ability of each move type to optimize the value of the all-atom energy function (Table II). Most of the modified move sets evaluated here involve perturbation of backbone torsion angles to values not described by the original fragment set, necessitating the inclusion of the torsion potential term (Table II; rama) to describe local interactions. In addition, all terms in the all-atom function are differentiable as binned terms are linearly interpolated.

The test set consisted of eight models generated by Rosetta for each of seven small proteins. These models were generated by the standard *de novo* protocol (see above). The backbones were subjected to a short relaxation protocol to remove steric clashes (using the high-res radii set for the vdw term; see Table I), and then side chains were added to the decoys by means of a simulated annealing repacking algorithm.[15] For each decoy, a total of 200 times the number of residues in the sequence moves was attempted for each move type or combination of move types. During the course of the attempted backbone moves, side chains were completely reoptimized, using the simulated annealing protocol. Five independent simulations were performed for each move type for each decoy. Simulations in which the final model differed from the starting structure by more than 4 Å C_α RMSD were discarded, and the simulation resulting in the lowest energy model was selected for analysis.

---

[31] K. T. Simons, C. Strauss, and D. Baker, *J. Mol. Biol.* **306,** 1191 (2001).

Figure 3 illustrates the relative effectiveness of each move type for optimizing the all-atom energy function. For each starting decoy in the test set, the energies of the lowest energy decoy obtained from application of each move type were sorted from highest energy (1) to lowest (30). The average rank over all decoys for each move type is reported in Fig. 3A. The fraction of attempted moves accepted for each move type is reported in Fig. 3B, and Fig. 3C reports the frequency of simulations that resulted in expanded structures. As expected, moves that cause smaller global perturbations (small, shear) are generally accepted with higher frequency than more globally perturbing moves. The critical conclusion is that most of the modified fragment insertion move types, although not accepted with high frequency, are significantly more effective than standard fragment insertions in optimizing the cost function, and are less likely to result in expanded protein conformations. Addition of the Monte Carlo-plus-minimization strategy increases the efficacy of the moves as well. Furthermore, application of the most effective moves in combination further increases both the acceptance rate and the extent of optimization of the cost function (see combination 1 and combination 2 in Fig. 3).

Conclusions

Although any protein-modeling strategy must attempt to find an optimal tradeoff between cost of computation of each move and the effectiveness of modifications in optimizing a cost function, the optimal tradeoff is specific to the particular problem of interest. The random selection of fragment insertions without consideration of gradient information or likelihood of the modification being accepted allows fragment insertion to be an extremely rapid operation and is well suited for *de novo* structure prediction, where coarse topological information is of interest. Conversely, careful selection of a move with a higher probability of acceptance or an operation that modifies many degrees of freedom simultaneously can be expensive to compute, yet may be a significantly more effective modification. Many structural modeling problems, including model refinement and loop modeling, require finer searches of a localized region of conformational space. For such problems, backbone modifications designed to effectively search local conformations are likely worth the extra computational expense.

Our goal in developing Rosetta is to assemble a unified platform for structural modeling that provides scoring functions and move sets applicable to a wide range of resolutions. The scoring function is modular, and individual terms of the energy functions can be combined with weights optimized for specific applications. In addition, the move sets described here provide options for both coarse-grained and local conformational searches. Different combinations of these move sets and scoring functions provide Rosetta modules for *de novo* structure prediction, model refinement, loop modeling, rigid body protein–protein docking, and protein design. Combinations of modules allow complex modeling problems to be approached. For example, combination of the *de novo* prediction, design, and model refinement modules has been used to design a protein with a novel topology.[10] Combining the docking and design modules allows alternative rigid body orientations to be explored for interface design, whereas combining the docking module with model refinement is applicable to docking with backbone flexibility.

Supplemental Materials

Licensing information for Rosetta may be obtained by e-mail (rosettaNMR@rosetta.bakerlab.org, rosettaABINITIO@rosetta.bakerlab.org, rosettaFRAGMENTS@rosetta.bakerlab.org). In addition, automated Rosetta predictions can be obtained from the Rosetta server[32] at http://robetta.bakerlab.org. The Rosetta server uses a combination of *de novo* prediction and homology modeling to produce complete three-dimensional models for proteins. Rosetta fragment libraries can be obtained from the automated server at http://rosetta.bakerlab.org/fragmentssubmit.jsp.

Appendix I

Efficient Computation of MSD Induced by Rigid Body Transformation

After fragment insertion, the relative positions of the protein chains before and after the insertion point undergo a rigid body motion. A convenient and differentiable measure of this global change is to fix one of the chains in space and compute the mean squared deviation (MSD) of the atoms in the other, "downstream" chain. Commonly, we want to compute this quantity quickly for all members of the fragment library at a given insertion position and then screen out the ones with a large MSD. This appendix discusses a general method for accelerating this screen and the computation of the derivative.

---

[32] D. Chivian, D. E. Kim, L. Malmström, P. Bradley, T. Robertson, P. Murphy, C. E. M. Strauss, R. Bonneau, C. A. Rohl, and D. Baker, *Proteins Struct. Funct. Genet.* **53**(Suppl. 6), 524 (2003).

To compute the MSD, we apply the fragment replacement transform to each downstream atom, find its squared deviation, and compute the mean. Specifically, if $\{\hat{V}_k\}$ are the initial coordinates, (V_x, V_y, V_z), of the N_{atoms} atoms downstream from the end of the insertion window; $\hat{\hat{A}}_1$ and $\hat{U}_1$ are, respectively, the rotation and translation caused by the existing torsion angles in the insertion window; and $\hat{\hat{A}}_2$ and $\hat{U}_2$ are, respectively, the rotation and translation induced by the new torsion angles in the insertion window, then after the fragment replacement the new coordinates of the N_{atoms} atoms are $\hat{V}_k{}' \equiv {}^{T}\hat{\hat{A}}_2\,[\hat{\hat{A}}_1\,(\hat{V}_k - \hat{U}_1)] + \hat{U}_2$ and the MSD is given by

$$\text{MSD} = \frac{1}{N_{\text{atoms}}} \sum_{k=1}^{N_{\text{atoms}}} \left\| \hat{V}_k - {}^{\text{T}}\hat{\hat{A}}_2 \left[\hat{\hat{A}}_1 \left(\hat{V}_k - \hat{U}_1 \right) \right] - \hat{U}_2 \right\|^2 \qquad \text{(A1)}$$

As written, this calculation is slow for large values of N_{atoms}, but it can be significantly accelerated by the insight that all relevant properties of the atom positions, (V_x, V_y, V_z), can be summarized by the inertial tensor, $\hat{\hat{E}}$, and the center of mass position, $\hat{V}_{\text{avc}}$.

Expanding all of the multiplications implied by the square and rearranging the sum yields

$$\begin{aligned} \text{MSD} = {} & \left\| {}^{\text{T}}\hat{\hat{A}}_2\hat{\hat{A}}_1 \left(\hat{V}_{\text{ave}} - \hat{U}_1 \right) - \left(\hat{V}_{\text{avc}} - \hat{U}_2 \right) \right\|^2 \\ & + \frac{2}{N_{\text{atoms}}} \sum_{k=1}^{N_{\text{atoms}}} {}^{\text{T}}\left(\hat{V}_k - \hat{V}_{\text{ave}} \right) \left(\hat{\hat{1}} - {}^{\text{T}}\hat{\hat{A}}_2\hat{\hat{A}}_1 \right) \left(\hat{V}_k - \hat{V}_{\text{ave}} \right) \qquad \text{(A2)} \end{aligned}$$

where the first term is simply $\|\hat{V}_{\text{avc}} - \hat{V}'_{\text{ave}}\|^2$, the MSD change of the center of mass, and the last term is the contribution to the MSD from the rotation about the center of mass. We can pull the rotational dependence out of the summation as follows. Let $\hat{\hat{R}} \equiv {}^{T}\hat{\hat{A}}_2\hat{\hat{A}}_1$ be the net rotation, and let $\hat{\hat{E}}$ be the second moments of $\{\hat{V}_k\}$. For example,

$$E_{xy} \equiv \sum_{k=1}^{N_{\text{atoms}}} (V_{x_k} - V_{x_{\text{ave}}})(V_{y_k} - V_{y_{\text{ave}}})$$

The summation of the second term in Eq. (A2) may then be written as the dot product of the $\hat{\hat{1}} - \hat{\hat{R}}$ and $\hat{\hat{E}}$ matrices viewed as vectors:

$$\begin{aligned} E_{xx} + E_{yy} + E_{zz} - [& R_{xx}E_{xx} + R_{xy}E_{xy} + R_{xz}E_{xz} + R_{yx}E_{yx} \\ & + R_{yy}E_{yy} + R_{yz}E_{yz} + R_{zx}E_{xx} + R_{zy}E_{zx} + R_{zz}E_{zz}] \end{aligned}$$

After precomputation of $\hat{\hat{E}}$, the cost of computing the MSD for each candidate insertion fragment is independent of N_{atoms}, resulting in more than an order of magnitude speed improvement [over Eq. (A1)] in evaluating the MSD for typical downstream chain lengths.

This MSD is asymmetric because the chain on one side of the fragment insertion is treated as fixed while the chain on the "downstream" side undergoes the rigid body motion. To avoid a directional preference, we fix whichever chain is the larger of the two, on average halving the precomputation of the inertial tensor. The transformation factors, $\hat{A}$, $\hat{U}_1$, $\hat{A}_2$, and $\hat{U}_2$, are dependent on the arbitrary initial protein orientation, but this dependence can be eliminated by moving the fragment insertion position to a standard origin and orientation: this transformation only requires rotating the inertial moments and also allows $\hat{A}_2$ and $\hat{U}_2$ to be precomputed for the entire fragment library. As a final refinement, $\hat{A}_1$ is preapplied to $\hat{E}$, reducing $\hat{R}$ trivially to ${}^{\mathrm{T}}\hat{A}_2$.

The decomposition of the MSD expression into translational and rotational motion is not only illustrative, but also allows fragments to be screened for motions that are both small in MSD and also primarily shearing (or nonshearing) motions. For example, one might seek a motion that twists a helix in place or, conversely, a pure shearing motion that frame shifts a strand pairing without twisting it. Screening for fragments with these effects is possible in this framework by unevenly weighting the relative contribution of individual terms of Eq. (A2).

Efficient Computation of MSD Partial Derivatives

Following a fragment insertion that transforms the downstream chain from $\{\hat{V}_k\}$ to $\{\hat{V}'_k\}$, this transformation can be counteracted by continuously varying a selected set of torsion angles to minimize the MSD by gradient descent (wobble). Here $\{\hat{V}'_k\}$ are the coordinates of the downstream chain before minimization, $\{\hat{V}_k\}$ are the target coordinates of the downstream chain (i.e., the coordinates before the initial fragment insertion), and we must compute the partial derivatives of the MSD with respect to each torsion angle. The differential motion $\partial\hat{V}'$ of a point in space when rotated by $\partial\theta$ about a bond axis $\hat{B}$ passing through an atom $\hat{V}_{\text{atom}}$ is

$$\partial\hat{V}' = \hat{B} \times \left(\hat{V}' - \hat{V}_{\text{atom}}\right)\partial\theta$$

The change in the distance between any two points V' and V caused by this infinitesimal motion is its projection along the line joining them:

$$\partial\|\hat{V} - \hat{V}'\| = \partial\hat{V}' \cdot \left(\hat{V} - \hat{V}'\right)/\|\hat{V} - \hat{V}'\|$$

Thus the partial derivative of the MSD with respect to any torsion angle is

$$\begin{aligned}\frac{\partial \mathrm{MSD}}{\partial \theta} &= \frac{1}{N_{\mathrm{atoms}}}\sum_{k=1}^{N_{\mathrm{atoms}}}\frac{\partial}{\partial\theta}\|\hat{V}_{\mathrm{k}}-\hat{V}'_{\mathrm{k}}\|^2 = \frac{2}{N_{\mathrm{atoms}}}\sum_{k=1}^{N_{\mathrm{atoms}}}\|\hat{V}_{\mathrm{k}}-\hat{V}'_{\mathrm{k}}\|\frac{\partial\|\hat{V}_{\mathrm{k}}-\hat{V}'_{\mathrm{k}}\|}{\partial\theta}\\ &= \frac{2}{N_{\mathrm{atoms}}}\sum_{k=1}^{N_{\mathrm{atoms}}}\hat{B}\times\left(\hat{V}_{\mathrm{k}}-\hat{V}_{\mathrm{B}}\right)\cdot\left(\hat{V}_{\mathrm{k}}-\hat{V}'_{\mathrm{k}}\right) = \frac{-2}{N_{\mathrm{atoms}}}B\cdot\sum_{k=1}^{N_{\mathrm{atoms}}}\hat{V}_{\mathrm{B}}\\ &\quad\times\left(\hat{V}_{\mathrm{k}}-\hat{V}'_{\mathrm{k}}\right)+\hat{V}_{\mathrm{k}}\times\hat{V}'_{\mathrm{k}}\end{aligned} \tag{A3}$$

Applying the cross-product associativity rule,

$$\frac{\partial \mathrm{MSD}}{\partial \theta} = \frac{2}{N_{\mathrm{atoms}}}\sum_{k=1}^{N_{\mathrm{atoms}}} -\hat{B}\cdot\hat{V}_{\mathrm{atom}}\times\left(\hat{V}_{\mathrm{k}}-\hat{V}'_{\mathrm{k}}\right)-\hat{B}\cdot\hat{V}_{\mathrm{k}}\times\hat{V}'_{\mathrm{k}}$$

Finally, the center of mass motion and the rotation about the center of mass can be factored into two separate components:

$$\begin{aligned}\frac{\partial \mathrm{MSD}}{\partial \theta} &= -2\hat{B}\cdot\left(\hat{V}_{\mathrm{ave}}-\hat{V}_{\mathrm{atom}}\right)\times\left(\hat{V}'_{\mathrm{ave}}-\hat{V}_{\mathrm{atom}}\right)\\ &\quad-\frac{2}{N_{\mathrm{atoms}}}\hat{B}\cdot\sum_{k=1}^{N_{\mathrm{atoms}}}\left(\hat{V}'_{\mathrm{k}}-\hat{V}_{\mathrm{ave}}\right)\times\left(\hat{V}'_{\mathrm{k}}-\hat{V}'_{\mathrm{ave}}\right)\end{aligned} \tag{A4}$$

As before, the last sum can be collapsed to a simple vector depending only on R and E. For example, the x coordinate of this sum is

$$R_{zx}E_{xy}+R_{zy}E_{yy}+R_{zz}E_{zy}-R_{yy}E_{xz}-R_{vv}E_{yz}-R_{yz}E_{zz}$$

In computing the gradient, one evaluates the partial derivative for each torsion axis. This calculation can be done efficiently by using the form in Eq. (A4): $\hat{V}_{\mathrm{atom}}$ and $\hat{B}$ differ for each torsion angle, but $\hat{V}_{\mathrm{ave}}$, $\hat{V}'_{\mathrm{ave}}$, R, and E are determined by the configuration and are the same for every torsion angle.

Applications of Efficient MSD Evaluation

Although the MSD is discussed above in the context of single-fragment replacement and subsequent continuous perturbation (wobble) of selected torsion angles, numerous other applications of these methods exist. In loop modeling, for example, the starting and ending points of a loop segment are known and the goal is to locate a fragment that will join these end points. This problem is isomorphic to computing the MSD in fragment replacement; although there is no existing fragment being replaced, the targets $\hat{A}_1$ and $\hat{U}_1$ are defined by the fixed segments. In addition, in the case of incomplete loop closure, the downstream chain position implied by the torsion angles of the terminal loop residue differs from the fixed template coordinates of the downstream chain. Gradient descent minimization of this MSD can be used to perturb torsion angles of loop residues to close this chain break. These methods are in fact incorporated into the loop-modeling strategy used in Rosetta (Rohl *et al.*[12]). In general, the methods described here are applicable to computation and minimization of MSD between any two coordinate sets, $\{\hat{V}_k\}$ and $\{\hat{V}'_k\}$.

Appendix II

Gunn Estimation of Global Perturbation Induced by Fragment Replacement

The net translation and rotation of the chain resulting from a fragment is described by six degrees of freedom that can be chosen such that if two fragments have nearly the same values for these six parameters, they produce nearly the same rotation and translation. Let $\hat{x}_1$ be the unit vector along the N-to-C$_\alpha$ bond at the N terminus of the fragment, let $\hat{x}_2$ be the analogous unit vector along the C-to-C$_\alpha$ bond at the C terminus of the fragment, and let $\hat{R}$ be the vector between the C$_\alpha$ atoms at the fragment N and C termini. Let $\hat{y}_1$ and $\hat{y}_2$ be the normals to the planes defined by N, C$_\alpha$, and C of the N- and C-terminal (respectively) residue of the fragment. The six parameters to describe the net rotation and translation are chosen such that they are independent of the absolute position and orientation. q_1 and q_2, that is,

$$q_1 = \hat{x}_1 \cdot \hat{R}$$
$$q_2 = \hat{x}_2 \cdot \hat{R}$$

are proportional to the cosine of the polar angle of final and initial bond vectors with respect to the R-axis;

$$|q_3| = \arccos\left[\frac{\hat{x}_1 \cdot \hat{x}_2 - \left(\hat{x}_1 \cdot \hat{R}\right)\left(\hat{x}_2 \cdot \hat{R}\right)}{\sqrt{\left(1 - q_1^2\right)\left(1 - q_2^2\right)}}\right]$$

is the dihedral angle between $\hat{x}_1$ and $\hat{x}_2$ along the $\hat{R}$ axis;

$$|q_4| = \arccos\left(\frac{\hat{y}_1 \cdot \hat{R}}{\sqrt{\left(1 - q_1^2\right)}}\right)$$

and

$$|q_5| = \arccos\left(\frac{\hat{y}_2 \cdot \hat{R}}{\sqrt{(1-q_2^2)}}\right)$$

are the angles between $\hat{y}_1$, $\hat{y}_2$ and the $\hat{R}$ axis; and

$$q_6 = \|\hat{R}\|$$

is the fragment length. These definitions match those of Gunn,[12a] other than the specific atoms used to define the coordinate system.

These parameters are independent of the orientation and origin of the coordinate system, allowing two fragments to be compared term by term:

$$\begin{aligned}\text{cost} \equiv\ & c_1 \ln(1 + |\Delta q_1| + |\Delta q_2|) + c_2 \ln(1 + |\Delta q_3|) \\ & + c_3 \ln(1 + |\Delta q_4| + |\Delta q_5|) + c_4 \ln(1 + |\Delta q_6|)\end{aligned}$$

where Δq is the difference between the respective q values of the original and replacement fragment and the absolute value also implies modulo-π for the angular Δq values. The $\{c\}$ coefficients were determined by regression on a test set of single-domain proteins to make the cost function a good discriminator between small and large RMS deviations in the downstream chain induced the fragment swap. In Rosetta, $\{c\} = \{5.72, 2.035, 3.84, 0.346\}$, and typical lower and upper cost thresholds are 0.03 and 4.08, respectively. In ballpark terms, these limits correspond to pure angular changes between ~0.5 and ~90°, or pure displacements of less than ~2 Å. When all the terms contribute equally to the cost, the upper limit on the deviation of any one angular degree of freedom falls to about 10°.

In the above description, the effect of modifying ψ (and ω) of the C-terminal residue of the insertion window is not included in the six q parameters. These angles are inserted into the chain, however, to avoid potential violations of allowed Ramachandran space that may occur if the C-terminal residue (ϕ, ψ) angles are effectively drawn from different library fragments. Although this inaccuracy limits the discriminatory power of the cost function and could be corrected by modification of the q parameter definitions, the cost function works well in practice, presumably because when the rotation and translation of two fragments are similar, the terminal torsion angles are likely to be similar as well.

[5] Poisson–Boltzmann Methods for Biomolecular Electrostatics

By Nathan A. Baker

Introduction

The understanding of electrostatic properties is a basic aspect of the investigation of biomolecular processes. Structures of proteins and other biopolymers are being determined at an increasing rate through structural genomics and other efforts; furthermore, the specific roles of these biopolymers in cellular pathways and supramolecular assemblages are being detected by genetic and other experimental efforts. The integration of this information into physical models for drug discovery or other applications requires the ability to evaluate intra- and interbiomolecular energetics. Among the various components of molecular interactions, electrostatic energetics and forces are of special importance because of their long range and the substantial charges of amino and nucleic acids.

Because of the ubiquitous nature of electrostatic interactions in biomolecular systems, a variety of computational methods have been developed for elucidating these interactions (see Refs. 1–5 and references therein). Popular computational electrostatics methods for biomolecular systems can be loosely grouped into two categories: "explicit solvent" methods, which treat the solvent in full molecular detail, and "implicit solvent" methods, which include solvent influences in averaged or continuum fashion. Although explicit solvent approaches offer more detailed insight into solvent-mediated biomolecular interactions, the necessity to integrate over the numerous solvent degrees of freedom often limits the ability of these methods to calculate thermodynamic quantities for large biomolecular systems.

---

[1] N. A. Baker and J. A. McCammon, *in* "Structural Bioinformatics" (P. Bourne and H. Weissig, eds.), John Wiley & Sons, New York, 2002.

[2] M. Gilson, *in* "Biophysics Textbook Online" (D. A. Beard, ed.), Biophysical Society Bethesda; T. A. Darden, *in* "Computational Biochemistry and Biophysics" (O. M. Becker, A. D. J. MacKerell, B. Roux, and M. Watanabe, eds.), pp. 91–114. Marcel Dekker, New York, 2001.

[3] B. Roux, *in* "Computational Biochemistry and Biophysics" (O. M. Becker, A. D. J. MacKerell, B. Roux, and M. Watanabe, eds.), pp. 133–152. Marcel Dekker, New York, 2001.

[4] M. E. Davis and J. A. McCammon, *Chem. Rev.* **94,** 7684 (1990).

[5] B. Honig and A. Nicholls, *Science* **268,** 1144 (1995).

METHODS IN ENZYMOLOGY, VOL. 383

Copyright 2004, Elsevier Inc.
All rights reserved.

0076-6879/04 $35.00

Brief Overview of Implicit Solvent Methods

Because of the sampling issues associated with explicit solvent treatments, implicit solvent methods have gained increasing popularity for elucidating the electrostatic properties of biomolecules in solution (see Refs. 1–5). As their name implies, these methods implicitly average over the configuration space of solvent and counterion species surrounding the biomolecule. The result is a polarizable continuum representation for the solvent and a mean field charge "cloud" for the counterion distribution. Despite some artifiacts arising from this continuum treatment (see Roux[3] and references therein), implicit solvent methods offer a significant advantage over traditional explicit solvent approaches and have become standard techniques for investigating the energetics and dynamics of biomolecular systems.

The importance of electrostatic interactions in protein behavior was recognized decades ago in work by Linderstrom-Lang[6] in developing protein titration models; Tanford and Kirkwood[7] in investigating the effects of pH and ionic strength on enzyme activity; and Flanagan *et al.*[8] in studying the energetics of dimer–tetramer assembly in hemoglobin. However, electrostatic models for protein systems were improved dramatically in 1982 by Warwicker and Watson.[9] Drawing on the increased knowledge of the three-dimensional structure of proteins increased computer power; they introduced a grid-based, finite difference approach for calculating the electrostatic potential of a nonspherical protein by solving the Poisson–Boltzmann (PB) equation, a nonlinear partial differential equation that incorporates detailed information about the biomolecular shape and charge distribution. Since this pioneering work, the PB equation has become a standard method for the detailed investigation of biomolecular electrostatics.[1,4,5,10]

In addition to PB methods, simpler approximate models have also been constructed for continuum electrostatics, including distance-dependent dielectric functions,[11] analytic continuum methods,[12] and generalized Born

---

[6] K. Linderstrom-Lang, *C. R. Travl. Lab. Carlsberg* **15** (1924).

[7] C. Tanford and J. G. Kirkwood, *J. Am. Chem. Soc.* **79,** 5333 (1957).

[8] M. A. Flanagan, G. K. Ackers, J. B. Matthew, G. I. H. Hanania, and F. R. N. Gurd, *Biochemistry* **20,** 7439 (1981).

[9] J. Warwicker and H. C. Watson, *J. Mol. Biol.* **157,** 671 (1982).

[10] C. Holm, P. Kekicheff, and R. Podgornik, eds. *in* "Electrostatic Effects in Soft Matter and Biophysics," Vol. 46. NATO Science Series. Kluwer Academic, Boston, 2001.

[11] A. D. J. MacKerell and L. Nilsson, *in* "Computational Biochemistry and Biophysics" (O. M. Becker, A. D. J. MacKerell, B. Roux, and M. Watanabe, eds.), Marcel Dekker, New York, 2001; A. R. Leach, *in* "Molecular Modelling: Principles and Applications," 2nd Ed., pp. xxiv and 744, and plate 16. Prentice Hall, New York, 2001.