Protein Structure Prediction Using a Combination of Sequence Homology and Global Energy Minimization: II. Energy Functions

MICHAEL J. DUDEK,[1] K. RAMNARAYAN,[1] JAY W. PONDER[2]

[1] *Structural Bioinformatics Inc., 10929 Technology Place, San Diego, California 92127*
[2] *Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri 63110*

Received 22 April 1997; revised 22 October 1997

ABSTRACT: A protein energy surface is constructed. Validation is through applications of global energy minimization to surface loops of protein crystal structures. For 9 of 10 predictions, the native backbone conformation is identified correctly. Electrostatic energy is modeled as a pairwise sum of interactions between anisotropic atomic charge densities. Model repulsion energy has a softness similar to that seen in *ab initio* data. Intrinsic torsional energy is modeled as a sum over pairs of adjacent torsion angles of 2-dimensional Fourier series. Hydrophobic energy is that of a hydration shell model. The remainder of hydration free energy is obtained as the energetic effect of a continuous dielectric medium. Parameters are adjusted to reproduce the following data: a complete set of *ab initio* energy surfaces, meaning one for each pair of adjacent torsion angles of each blocked amino acid; experimental crystal structures and sublimation energies for nine model compounds; *ab initio* energies over 1014 conformations of 15 small-molecule dimers; and experimental hydration free energies for 48 model compounds. All *ab initio* data is at the Hartree–Fock/6–31G* level. © 1998 John Wiley & Sons, Inc. J Comput Chem 19: 548–573, 1998

Keywords: surface loops; structure prediction; global energy minimization; energy functions; hydration free energy; atomic multipoles

Contract/grant sponsor: NIH; contract/grant number P01 GM24483

CCC 0192-8651 / 98 / 050548-26

Introduction

This article is the second in a series, the goal of which is to develop a method for predicting protein structure that obtains structural information both from the crystal structure of a homologous protein and from global minimization of a function representing energy. The crystal structure of a homologous protein provides information that is approximately equivalent to distance constraints for some subset of atom pairs. Typically, these distance constraints enable accurate prediction of the core of the native structure but do little to restrict the space of conformations that is available to the surface. In principle, information about the surface of the native structure can be obtained from global energy minimization. In practice, however, the attainment of any significant amount of structural information from a molecular mechanics based method is extremely difficult because of the following two problems: the multiple minima problem and the requirement that the energy surface be accurate.

The primary results of our earlier work[1] were algorithms and code that enable very effective global energy minimization of protein surface loops and some clues concerning the origin of errors in the energy surface that was used. This latter result was made possible by the former through our use of global energy minimization of protein surface loops as a tool for identifying and correcting problems with energy surfaces. In this article we substitute a more accurate energy surface, perhaps splitting the difference between what was available earlier and what is possible given the limitations of a rigid geometry, nonpolarizable model. We describe the construction of the energy surface and validation through structure prediction of protein surface loops.

This work closely follows a strategy that was outlined earlier.[2] The energy surface consists of the following forms. The electrostatic component is represented by a multipole expansion with truncation at the inverse fifth power of distance.[2] The repulsion + dispersion component is represented by a buf14-7 functional form[3] with the addition of a third independent parameter used to adjust softness. The intrinsic torsional component is represented by a 2-dimensional (2-D) Fourier series of order 6.[2] The hydration free energy component is represented as the combined energy of two distinct models. The energetic effects of a continuous dielectric medium are calculated using a boundary element solution to the Poisson equation.[4] The hydrophobic effect is estimated using a Gaussian volume implementation of a hydration shell model.[5] The data on which the parameterization is based consists of: a complete set of *ab initio* energy surfaces,[2] meaning one for each pair of adjacent torsion angles of each blocked amino acid; experimental crystal structures and sublimation energies for 9 model compounds; *ab initio* energy surfaces for a collection of small-molecule dimers; and experimental hydration free energies for 48 model compounds. All of the *ab initio* data is at the Hartree−Fock (HF)/6−31G* level.

It is useful to view this work in an abstract way as an interaction between two complex, imperfect objects: an energy surface and a procedure for global energy minimization of protein surface loops. The adequacy of the energy surface is judged on the basis of success in applications of global energy minimization to surface loops of protein crystal structures. This measure of accuracy forms the primary selective pressure under which the energy surface evolves. In turn, the attainment of an energy surface that is accurate enough to enable reliable structure prediction would provide further incentive for increasing the efficiency of global energy minimization, such that progressively less structural information would be required from sequence homology. From this perspective, the substitution of the current energy surface, which was inspired by earlier feedback from global energy minimization of protein surface loops,[1] constitutes a first iteration of a systematic process. The progress that corresponds to this iteration is substantial.

Methods

GEOMETRY

The energy surface is based on a rigid geometry model. The primary definition of geometry, which is propagated throughout all programs, *ab initio* and molecular mechanics, is a collection of z matrices, one for each blocked amino acid, the blocking groups being acetyl and *N*-methyl. The z matrices allow easy comprehension of the geometry. The use of one primary source of geometry, which

is converted into all others, enforces consistency between programs.

The bond lengths and bond angles of the peptide group are those suggested by Benedetti.[6] Other elements of the geometry were taken from Momany et al.[7]

AB INITIO ENERGY SURFACES

For each pair of adjacent torsion angles of each blocked amino acid, an *ab initio* energy surface was calculated (or assumed equivalent to another previously calculated) over a 24×24 grid (15° spacing between grid points). All torsion angles other than the adjacent pair that characterizes a surface were maintained rigid. We refer to this database as a complete set of *ab initio* energy surfaces. Table I shows the composition. Here, rows correspond to blocked amino acids, columns to pairs of adjacent torsion angles. Entries specify grids over which *ab initio* energy surfaces were calculated (subject to the following screen) using the program SPARTAN.[8] Except for the (ϕ, ψ) energy surfaces of blocked alanine and glycine, wavefunctions were not calculated over grid points that were predicted to contain overlaps based on an ECEPP repulsion + dispersion energy of greater than 5 kcal/mol. Blank entries indicate transfer rather than *ab initio* computation, based on expected similarity to other energy surfaces in the database.

For a given pair of adjacent torsion angles (a given column of Table I), two amino acids can be considered equivalent if the connectivity is identical for each of the three atoms that form the two adjacent bonds. This definition of equivalence forms the basis for a partitioning of each column into classes. For each class, at least one member was obtained by *ab initio* computation. Consider, for example, the (ϕ, ψ) pair. For all of the blocked amino acids with the exceptions of glycine and proline, the connectivities about N, C^{α}, and C' are identical. In the database this class is represented only once by the (ϕ, ψ) energy surface of blocked alanine.

For each *ab initio* energy surface, interpolation was carried out to a larger 72×72 grid (5° spacing between grid points).[2] To partially correct for the lack of electron correlation,[9] each *ab initio* energy surface was scaled by a factor of $(\frac{1}{1.12})$ squared. We

TABLE I.
Complete Set of *Ab Initio* Energy Surfaces.

	Pairs of Adjacent Torsion Angles					
aa	(ϕ, ψ)	(ϕ, χ_1)	(ψ, χ_1)	(χ_1, χ_2)	(χ_2, χ_3)	(χ_3, χ_4)
Ala	24×24					
Asp				12×12		
Cys		12×12	12×12	12×12	12×12	
Glu					12×12	
Phe		24×24	24×24	24×24		
Gly	24×24					
His				24×24		
Ile				24×24		
Lys				24×24	12×12	12×12
Leu				24×24		
Met		12×12	12×12	12×12	12×12	
Asn				24×24		
Pro	24×24					
Gln				24×24	12×12	
Arg						12×12
Ser		24×24	24×24	24×24		
Thr		24×24	24×24	24×24		
Val		24×24	24×24			
Trp				12×12		
Tyr					12×12	

aa, blocked amino acid.

note that this element of the parameterization may not be optimal. Although we expect that the application of this scaling factor should be useful for bringing the electrostatic components of the correlated and uncorrelated *ab initio* energy surfaces into rough alignment, its correctness when similarly applied to repulsion + dispersion or intrinsic torsional components is not yet demonstrated.

ELECTROSTATIC COMPONENT

The electrostatic component is represented by a multipole expansion (interaction sites at nuclei) with truncation at the inverse fifth power of distance.[2] Interaction energies are calculated between all pairs of atoms, including types 1-2 and 1-3. The parameters consist of, for each atom of each amino acid, point multipoles through hexadecapole located at the nucleus.

For each wavefunction of each *ab initio* energy surface, atomic multipoles were calculated out to the seventh moment.[2,10] For each *ab initio* energy surface, a distinct set of averaged atomic multipoles (for that blocked amino acid) was obtained by averaging over the domain of that surface intersected with the set of grid points whose *ab initio* energies lie within 12 kcal/mol of the lowest. The atomic multipoles of equivalent hydrogens were averaged further. Also, planar or threefold symmetry was enforced when appropriate.

For each amino acid (unblocked), a final set of atomic multipoles was assembled as follows. For each atom, atomic multipoles were taken from the set of averaged atomic multipoles that corresponds to the energy surface for which fragments adjacent to the atom are rotated. If the required energy surface was not available, the atomic multipoles were transferred from an equivalent atom of another blocked amino acid. The atomic monopole moments were adjusted slightly, such that the net charge (excluding blocking groups) was either neutral or ± 1 electron charge unit (ecu). To partially correct for neglect of electron correlation, atomic multipoles were scaled uniformly by a factor of $(\frac{1}{1.12})$. As a consequence of this scaling, the final net charge on Asp, Glu, Lys, and Arg is $\pm(\frac{1}{1.12})$ ecu.

REPULSION + DISPERSION COMPONENT

The repulsion + dispersion component is represented by a three-parameter buf14-7 functional form[3]

$$f(r_{(a,b)}) = \varepsilon_{(T_a,T_b)}\left(\frac{1+\delta_{(T_a,T_b)}}{\left(\frac{r_{(a,b)}}{\rho_{(T_a,T_b)}}\right)+\delta_{(T_a,T_b)}}\right)^7 \times\left[\frac{1.12}{\left(\frac{r_{(a,b)}}{\rho_{(T_a,T_b)}}\right)^7+0.12}-2\right], \quad (1)$$

where $r_{(a,b)}$ is the distance of atom pair (a, b); T_a is the atom type of atom a; and $\varepsilon_{(T_a,T_b)}$, $\rho_{(T_a,T_b)}$, and $\delta_{(T_a,T_b)}$ are parameters that control the depth, the position of the minimum, and the softness, respectively, for the interaction between a pair of atoms having types T_a and T_b. In his introduction of this functional form (to fit highly accurate experimental and theoretical data for rare gas dimers), Halgren has suggested that the parameter δ be fixed at 0.07.[3] We allow δ to vary. This was needed to fit our *ab initio* data. The set of atom types is specified in Table II. To reduce the number of independently adjustable parameters, combination rules, which express $(\varepsilon_{(T_a,T_b)}, \rho_{(T_a,T_b)}, \delta_{(T_a,T_b)})$ as a function of $(\varepsilon_{(T_a,T_a)}, \rho_{(T_a,T_a)}, \delta_{(T_a,T_a)})$ and $(\varepsilon_{(T_b,T_b)}, \rho_{(T_b,T_b)}, \delta_{(T_b,T_b)})$, were used for all non-hydrogen-bonding pairs of atom types. For ε and ρ, the combining rules,

$$\rho_{(T_a,T_b)} = \left(\frac{\rho^6_{(T_a,T_a)}+\rho^6_{(T_b,T_b)}}{2}\right)^{1/6} \quad (2)$$

and

$$\varepsilon_{(T_a,T_b)}\rho^6_{(T_a,T_b)} = \left[\left(\varepsilon_{(T_a,T_a)}\rho^6_{(T_a,T_a)}\right)\left(\varepsilon_{(T_b,T_b)}\rho^6_{(T_b,T_b)}\right)\right]^{1/2} \quad (3)$$

are those suggested by Waldman.[11] For the softness parameter δ, the combining rule,

$$\left(\frac{1+\delta_{(T_a,T_b)}}{\delta_{(T_a,T_b)}}\right) = \left[\left(\frac{1+\delta_{(T_a,T_a)}}{\delta_{(T_a,T_a)}}\right)\left(\frac{1+\delta_{(T_b,T_b)}}{\delta_{(T_b,T_b)}}\right)\right]^{1/2}, \quad (4)$$

follows (approximately) from the expectation that the finite energy maxima that occur as the separa-

TABLE II.
Hydration Shell Radii (Å).

Type	Description	Radii	
		Inner	Outer
H_{00}	H bonded to tetrahedral C		
H_{02}	H bonded to N	0.8500	0.8500
H_{03}	H bonded to planar C or S		
H_{05}	H bonded to O	0.8500	0.8500
C_{08}	Tetrahedral C	1.8500	3.1869
C_{09}	Planar C (amide, acid, or carbonyl)	1.6000	2.1418
C_{10}	Planar C (other than C_{09})	1.8000	3.1187
N_{11}	Planar N with three bonds	1.7500	1.9774
N_{12}	Tetrahedral N or N with two bonds	1.7000	1.7000
O_{13}	O (amide, acid, or carbonyl)	1.7000	1.7000
O_{14}	O (other than O_{13})	1.6500	1.6500
S_{15}	S	1.9000	3.1351

tion distances r approach 0 might reasonably combine as the geometric mean. For pairs of atom types (T_a, T_b) that were judged to be capable of hydrogen bond formation, the parameters ($\varepsilon_{(T_a, T_b)}$, $\rho_{(T_a, T_b)}$, and $\delta_{(T_a, T_b)}$) were varied independently of any combining rules.

The independent ε, ρ, and δ parameters were adjusted to reproduce the experimental crystal structures and sublimation energies of nine small organic model compounds in addition to gas phase HF/6−31G* energies for 1014 conformations of 15 small-molecule dimers. The collection of crystal structures and characteristic properties are listed in Table III.[12−22] Six of these structures were determined at low temperature. The atom types of the collection span the range of atom types found in proteins. For each crystal, a reliable heat of sublimation has been determined. The lattice energies that were used in the parameterization were extrapolated to the temperatures of the structure determinations. The collection of dimers was formed by taking all pairs from the following five model compounds: ethane, pyridine, formamide, ethanol, and ethanethiol.

For each dimer, the following algorithm was used to obtain the conformations and corresponding *ab initio* energies that contribute to the target of the parameterization. A large collection of conformations, $12 \times 7 \times 12 \times 12 \times 7 \times 20$ on a 6-D grid, was generated by stepping through three Euler angles of molecule 1, two Euler angles of molecule 2, and distance for the closest atom pair. Distance for the closest atom pair ranges from the sum of the hard core radii to this minimal distance plus 1.52 Å in increments of 0.08 Å. This collection of conformations was partitioned into $n \times 20$ subcollections, corresponding to the n distinct pairs of atom types and the 20 distinct values of distance for the closest atom pair. From each subcollection, a single conformation, the most favorable based on a molecular mechanics estimate of energy, was retained. For each retained conformation, a single point *ab initio* energy was calculated using GAMESS.[23] Monomer energies were subtracted out. Conformations for which the repulsion energy (defined here as the *ab initio* energy minus the electrostatic component of the molecular mechanics energy) was greater than 16 kcal/mol were excluded.

For each crystal, a molecular geometry was selected as follows. The positions of the heavy atoms were taken from the experimental crystal structure. Hydrogen atoms were placed using standard bond lengths, bond angles, and torsion angles. This geometry was held fixed throughout all remaining calculations.

In both the crystal packing and gas phase dimer calculations, the molecular mechanics energy surface consists of two components: repulsion + dispersion and electrostatic. In both calculations, the electrostatic component was represented by a multipole expansion with truncation at the inverse eighth power of distance. The following calculations were carried out for each model compound. A wavefunction was calculated at the HF/6−31G* level for an isolated molecule using GAMESS.

TABLE III.
Crystal Structures Used in Parameterization.

Molecule	Physical Properties						
	Group[a]	Z[b]	T_m[c]	$\Delta H_{s\rightarrow g}$[d]	T_h[e]	Ref[f]	ΔE_{lat}[g]
Ethane	$P2_1/n$	2	89.9	4.90	90	21	−4.75
Heptane	$P\bar{1}$	2	182.6	13.84	183	21	−13.97
Benzene	Pbca	4	278.7	10.61	298	21	−10.18
Pyrazine	Pmnn	2	326.2	13.46	298	21	−13.54
Formamide	$P2_1/n$	4	275.8	17.30	265	22	−17.46
Oxamide	$P\bar{1}$	1	623.2	27.70	387	22	−27.46
Urea	$P\bar{4}2_1m$	2	405.9	23.20	351	22	−25.03
Trithiane	$Pmn2_1$	2	488.2	17.14	347	22	−16.75
Acetic acid	$Pna2_1$	4	289.8	16.09	223	21	−16.19

	Refinement Properties						
	Year[h]	Ref[f]	T_x[i]	#Refl[j]	B(H / Heavy)[k]	R[l]	Atom types
Ethane	78	12	85	610	an / an	0.052	H_{00} C_{08}
Heptane	77	13	100	1112	iso / an	0.080	H_{00} C_{08}
Benzene	58	14	270	284	iso / an	0.099	H_{03} C_{10}
Pyrazine	76	15	184	605	iso / an	0.047	H_{03} C_{10} N_{12}
Formamide	78	16	90	1125	an / an	0.038	H_{00} H_{02} C_{09} N_{11} O_{13}
Oxamide	77	17	293	1936	iso / an	0.058	H_{02} C_{09} N_{11} O_{13}
Urea	84	18	12	342	an / an	0.030	H_{02} C_{09} N_{11} O_{13}
Trithiane	69	19	298	324	iso / an	0.069	H_{00} C_{08} S_{15}
Acetic acid	71	20	133	316	an / an	0.092	H_{00} H_{05} C_{08} C_{09} O_{13} O_{14}

[a] Space group.
[b] Number of molecules in the unit cell.
[c] Melting temperature (K).
[d] Heat of sublimation (kcal / mol).
[e] Temperature of heat of sublimation determination (K).
[f] Reference.
[g] Estimated lattice energy (kcal / mol) (average intermolecular energy per molecule in the crystal) at the temperature of structure determination obtained using $\Delta E_{lat}(T_x) \approx \Delta H_{s\rightarrow g}(T_h)$-$RT_h$-$3R(T_x - T_h)$, where R is the gas constant.
[h] Year of structure determination.
[i] Temperature of structure determination (K).
[j] Number of reflections used in refinement.
[k] Quality of thermal parameters used in refinement for hydrogen / heavy atoms; an, anisotropic; iso, isotropic.
[l] R factor.

Atomic multipoles were calculated from the wavefunction out to the seventh moment using the distributed multipole analysis (DMA)[10] method. No uniform scaling of atomic multipoles by a factor of $(\frac{1}{1.12})$ was used in these calculations. Given the near identity of this electrostatic component (no scaling, truncation at $1/R^8$) with that previously constructed for the amino acids (uniform scaling, truncation at $1/R^5$), the derived repulsion + dispersion component is expected to be consistent with both.

Parameter local minimization was accomplished by a program developed by the authors. At each step, the program calculates as follows. For each crystal, energy is calculated at the experimental geometry, along with first and second derivatives of energy with respect to the nine Cartesian components of the three unit cell vectors and, for each molecule in the unit cell other than the first molecule, six parameters that specify the position and orientation of that molecule, treated as a rigid body, relative to the first molecule. Interactions are included between a central unit cell and all unit cells within 16 Å of the central cell. The position of the minimum is calculated for the harmonic surface defined by the energy and its derivatives. This

is used as a measure of the projected movement with energy minimization of the crystal away from the experimental structure. For each dimer conformation, energy is calculated. A target function is formed as a weighted sum of harmonic constraints on

1. deviations between calculated and experimental lattice energies,
2. components of the first derivative,
3. projected movements away from experimental crystal structures, and
4. deviations between molecular mechanics and *ab initio* dimer energies.

First and second derivatives of the target function are calculated with respect to the independent ε, ρ, and δ parameters. It is this target function surface that guides the walk through parameter space. All parameters are adjusted simultaneously to fit all target data. Final values of the independent parameters are presented in Table IV.

To validate this new representation, energy minimization was carried out for the nine crystals that were used in the parameterization. For each model compound, the energy of the crystal was minimized, starting from the experimental structure, with respect to the nine Cartesian components of the three unit cell vectors and, for each molecule in the unit cell other than the first molecule, six parameters that specify the position and orientation of that molecule relative to the first molecule. The individual molecules were treated as rigid bodies. In this way, the experimental symmetry of the crystal was not imposed. Interaction energies were calculated between a central unit cell and all unit cells within 16 Å of the central unit cell. Energy minimizations were accomplished with the use of a crystal packing program that was developed by the authors. This program enables the use of atomic multipoles out to the seventh moment.

A comparison of calculated crystal structures and lattice energies to experiment is shown in Table V. Here, the unit cell vectors are specified by the lattice parameters a, b, c and α, β, γ. Differences between experimental and calculated values are shown in parentheses. For example, for formamide, the movement away from the experimental structure is about 0.5 Å in unit cell lengths and about 3° in unit cell angles.

The following are points of reference useful for interpreting these results. Experimental errors in bond lengths are ~ 0.02 Å. Errors in H positions due to placement using standard bond angles and torsion angles are ~ 0.10 Å. This estimate assumes that errors in bond angles and torsion angles can be as large as 4°. For benzene, unit cell lengths a, b, and c decrease by 0.168, 0.189, and 0.288 Å, respectively, as temperature decreases from 270 to 78 K.

Despite the relatively sophisticated functional forms, the calculated structures are not without problems. In particular, formamide and benzene are not well predicted, both unit cells consisting of four small planar molecules. Only in these two cases is the relative movement of the individual molecules within the unit cell (not seen in this presentation of results) larger than the movement of the unit cell vectors. The formamide crystal consists of puckered hydrogen bonded sheets of dimers. These flatten out. It is not clear why. The

TABLE IV.
Parameter Values for Repulsion + Dispersion Component.

Type	ε	ρ	δ
H_{00}	0.0051	3.4815	0.1914
H_{02}	0.0168	2.0087	0.0343
H_{03}	0.0055	3.1780	0.0913
H_{05}	0.0321	1.8379	0.0675
C_{08}	0.1748	3.7101	0.2013
C_{09}	0.1885	3.3416	0.0711
C_{10}	0.0985	3.9443	0.0739
N_{11}	0.1630	3.7915	0.0699
N_{12}	0.1433	3.5676	0.0705
O_{13}	0.0415	3.7358	0.0726
O_{14}	0.0613	3.6577	0.0703
S_{15}	0.2395	4.1637	0.0719
Pair of Types	ε	ρ	δ
(H_{02}, N_{11})	0.0201	3.3253	0.0597
(H_{02}, N_{12})	0.0200	3.0578	0.0664
(H_{02}, O_{13})	0.0200	2.8189	0.0684
(H_{02}, O_{14})	0.0200	2.8219	0.0684
(H_{02}, S_{15})	0.0200	3.9309	0.0640
(H_{05}, N_{11})	0.0200	2.8497	0.0700
(H_{05}, N_{12})	0.0200	2.9609	0.0684
(H_{05}, O_{13})	0.0200	2.6652	0.0758
(H_{05}, O_{14})	0.0200	2.6758	0.0758
(H_{05}, S_{15})	0.0200	3.9465	0.0604

Units of energy and distance are 1 kcal / mol and 1 Å, respectively.

TABLE V.
Comparison of Calculated Crystal Structures and Lattice Energies to Experimental.

	Unit Cell Vectors					
	Formamide		Oxamide		Urea	
	Exptl	Calcd	Exptl	Calcd	Exptl	Calcd
a	3.60	3.67 (0.07)	3.61	3.62 (0.00)	5.56	5.51 (−0.05)
b	9.04	9.37 (0.33)	5.18	5.17 (0.00)	5.56	5.51 (−0.05)
c	6.99	6.49 (−0.49)	5.65	5.60 (−0.04)	4.68	4.76 (0.08)
α	90.00	90.01 (0.01)	83.77	84.09 (0.32)	90.00	90.01 (0.01)
β	100.50	103.33 (2.83)	113.97	114.68 (0.71)	90.00	90.01 (0.01)
γ	90.00	90.02 (0.02)	114.94	115.90 (0.96)	90.00	90.01 (0.01)
E_{lat}	−17.46	−17.61 (−0.15)	−27.46	−27.79 (−0.33)	−25.03	−24.66 (0.37)
	Ethane		Heptane		Benzene	
	Exptl	Calcd	Exptl	Calcd	Exptl	Calcd
a	4.22	4.40 (0.18)	4.15	4.19 (0.04)	7.46	7.07 (−0.38)
b	5.62	5.56 (−0.05)	19.97	20.09 (0.12)	9.66	9.49 (−0.16)
c	5.84	5.71 (−0.13)	4.69	4.54 (−0.14)	7.03	7.13 (0.10)
α	90.00	89.99 (0.00)	91.30	91.16 (−0.13)	90.00	90.00 (0.00)
β	90.41	89.60 (−0.80)	74.30	74.66 (0.36)	90.00	89.99 (0.00)
γ	90.00	89.99 (0.00)	85.10	86.57 (1.47)	90.00	89.99 (0.00)
E_{lat}	−4.75	−4.56 (0.19)	−13.97	−14.62 (−0.65)	−10.18	−10.20 (−0.02)
	Pyrazine		Trithiane		Acetic Acid	
	Exptl	Calcd	Exptl	Calcd	Exptl	Calcd
a	9.32	9.43 (0.10)	7.66	7.52 (−0.14)	13.22	13.50 (0.27)
b	5.85	5.73 (−0.11)	7.00	7.58 (0.58)	3.96	3.86 (−0.09)
c	3.73	3.76 (0.03)	5.28	5.31 (0.02)	5.76	5.70 (−0.06)
α	90.00	90.00 (0.00)	90.00	89.99 (0.00)	90.00	89.99 (0.00)
β	90.00	90.00 (0.00)	90.00	90.00 (0.00)	90.00	90.00 (0.00)
γ	90.00	90.00 (0.00)	90.00	90.03 (0.03)	90.00	89.93 (−0.06)
E_{lat}	−13.54	−13.62 (−0.08)	−16.75	−17.53 (−0.78)	−16.19	−16.24 (−0.05)

Units of cell lengths and angles are angstroms and degrees, respectively. The unit of energy is 1 kcal / mol. All unit cells within 16 Å of the central unit cell are included in the calculation.

benzene crystal is at a moderate temperature and held together by weak forces. It may require some ensemble averaging. Remarkably, these movements away from the experimental structures correspond to only small decreases in energy: 0.21, 0.24, 0.14, 0.05, 0.28, 0.27, 0.08, 0.88, and 0.10 kcal/mol for formamide, oxamide, urea, ethane, heptane, benzene, pyrazine, trithiane, and acetic acid, respectively. Such small energy adjustments with minimization, in some cases corresponding to large structural adjustments, reflect smooth flat energy surfaces not unlike those seen in protein local minimization. They demonstrate the sensitivity of the crystal data to errors of the type that remain in the energy functions. The relatively larger energy adjustment for trithiane may reflect larger errors in the *ab initio* calculations for sulfur than for the first row elements. Given the limitations of the current functional form, most notably neglect of polarization and nonspherical repulsion, accuracy greater than a few tenths of a kilocalorie per mole is not expected. We note also the limitations of potential energy minimization as a method for reproducing finite temperature crystal structures, particularly when the energy surface is flat over large structural rearrangements and the temperature of structure determination is high.

Overall, these fits are at least as good as any obtained previously.[24] Due to the use of a softer functional form along with the fit to *ab initio* data,

we expect that this representation is much improved in regions of 2–16 kcal/mol overlaps. Although the level of the *ab initio* data is quite low for this kind of work, any misguidance in regions of minima is corrected through the fit to crystal data.

INTRINSIC TORSIONAL COMPONENT

The intrinsic torsional component is represented by a 2-D Fourier series of order 6; for each pair of adjacent torsion angles, the parameters consist of 169 Fourier coefficients.[2] For each *ab initio* energy surface, the following calculations were used to obtain an initial (unjoined) set of coefficients. Corresponding energy surfaces were calculated for the electrostatic component and for the repulsion + dispersion component, each over a 72×72 grid. The 2-D Fourier coefficients were adjusted to fit the difference between the *ab initio* energy surface and the repulsion + dispersion + electrostatic energy surface.[2]

The target function consists of a measure of distance between corrected model and *ab initio* energy surfaces and a harmonic constraint on the size of each coefficient. The measure of distance is the one introduced earlier[2]: a weighted root mean square deviation (RMSD) over the grid points of a 72×72 grid, but with the added exclusion from the domain of the function of grid points such that

1. the repulsion + dispersion component is greater than 8 kcal/mol (relative to the global minimum of the surface);
2. the norm of the gradient of the repulsion + dispersion component with respect to all torsion angles other than the adjacent pair (θ_1, θ_2) whose variation defines the surface is greater than 5 kcal/mol rad; and
3. the distance from a grid point for which an *ab initio* value is available is greater than $16\sqrt{2}\,^\circ$.

In regions of the second type, the dependence of errors on torsion angles is not a form that can be safely corrected using a 2-D Fourier series.[2] Physically, 1-D corrections can become as large as 4 or 5 kcal/mol; 2-D corrections should be smaller. Accordingly, coefficients of 2-D terms were more strongly constrained than coefficients of 1-D terms.

For all of the corrected model energy surfaces, distances to the *ab initio* targets are within a few tenths of a kilocalorie/mole. Figure 1 shows energy surfaces for (ϕ, ψ) of blocked alanine. These include the target *ab initio* energy surface; the repulsion + dispersion surface; and the final model surface, which includes the 2-D Fourier correction. Contour levels range from 1 to 16 kcal/mol in increments of 1 kcal/mol. Anything above 16 kcal/mol has been shaded. Figure 1b shows that the repulsion + dispersion surface has a softness that is closely similar to that of the *ab initio* surface. The bridge region is surprisingly high at about 7 kcal/mol. Causes of this barrier include rigid geometry, no scaling of 1-4 interactions, and the hardness of amide hydrogen. It is easily reduced by the intrinsic torsional component. We note that the energy contour map of Figure 1c is altered only slightly by relaxing, at each grid point of the underlying 72×72 grid, all torsion angles other than ϕ and ψ. Figure 2 shows the corresponding energy surfaces for (χ_1, χ_2) of blocked histidine. Figure 3 shows the corresponding energy surfaces for (ψ, χ_1) of blocked threonine.

We can think about these corrections in the following way. First, we generalize the functional form that is used to represent intrinsic torsional energy from a sum of the 1-D Fourier series to a 2-D Fourier series. Second, we compensate for some of the errors in other energy components by introducing offsetting errors into the intrinsic torsional component. Errors that can be corrected are any that depend on a pair of adjacent torsion angles (θ_1, θ_2) but not on a third adjacent torsion angle θ_3. The net effect is to cause easily computable molecular mechanics model energy surfaces to reproduce more accurate quantum mechanical energy surfaces.

For pairs of adjacent torsion angles for which the required *ab initio* energy surface was not available, 2-D Fourier coefficients were transferred from an equivalent environment of a different blocked amino acid. For example, corrections obtained for (ϕ, ψ) of blocked alanine were transferred to (ϕ, ψ) of blocked tyrosine and the other blocked nonglycine, nonproline amino acids.

At this point, for each pair of adjacent torsion angles of each blocked amino acid, a correction has been obtained in the form of a 2-D Fourier series of order 6 for differences between the model and *ab initio* energy surfaces. These corrections were then merged to create a sum of overlapping 2-D Fourier

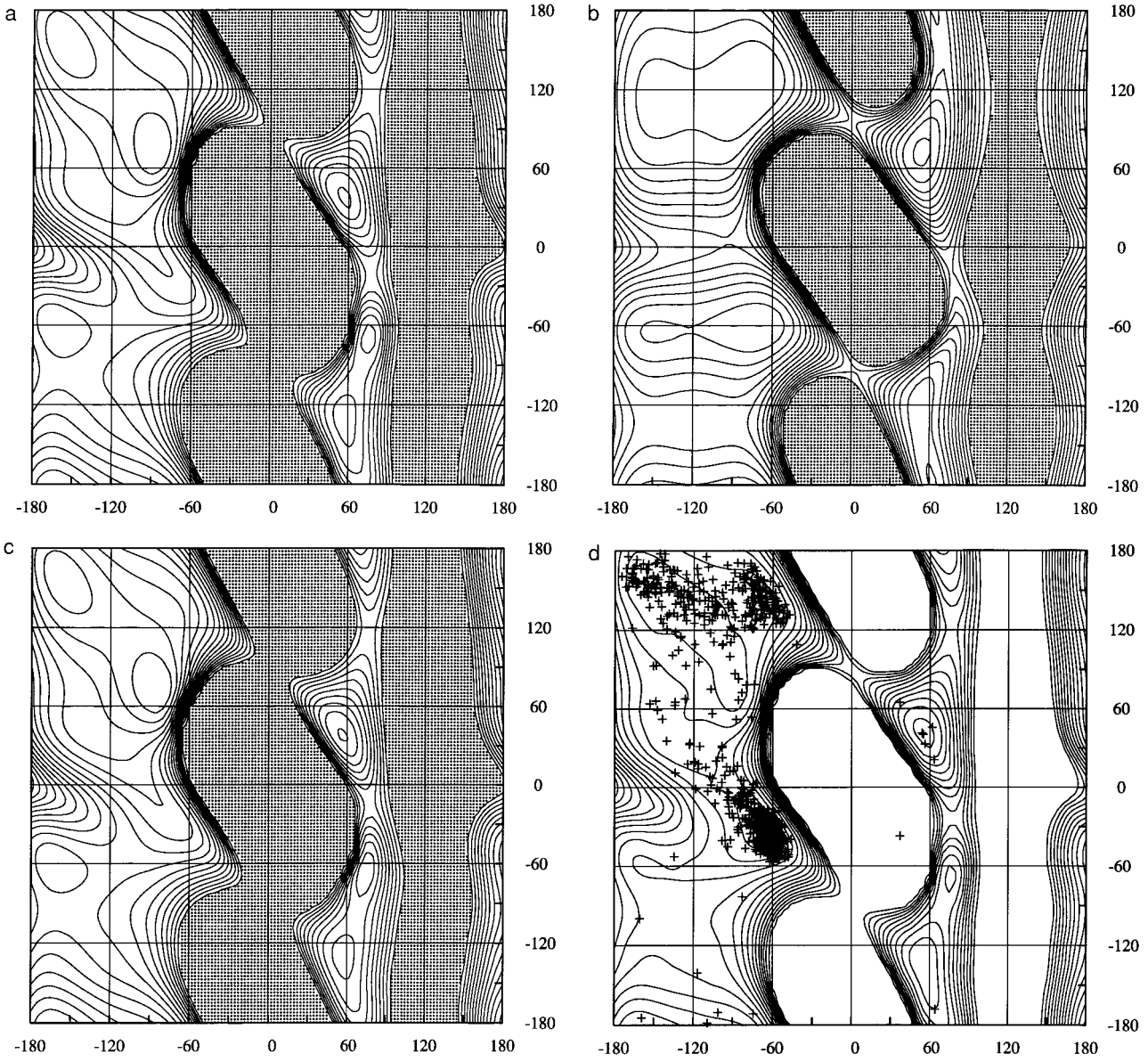

FIGURE 1. Contour plots of energy surfaces for (ϕ, ψ) of blocked alanine: (a) target *ab Initio*, (b) repulsion + dispersion, (c) repulsion + dispersion + electrostatic + intrinsic torsional, and (d) hydrated *ab initio* with superimposed scatter plot of experimental (ϕ, ψ) values for alanine residues. Contour levels range from 1 to 16 kcal / mol in increments of 1 kcal / mol.

series that corrects the model function throughout all of torsion space. For each amino acid, corrections were added to the model in the order of the columns of Table I. As each new correction was added, the corresponding initial set of 2-D Fourier coefficients was adjusted, if necessary, such that contributions from overlapping, previously added corrections were not counted twice. If all sources of corrections were dependent on one or two torsion angles (but not three or more), then joining by considering each adjacent pair in any given order should result in a unique set of 2-D Fourier coefficients.

For methyl groups and large two-fold barriers (for example ω), the intrinsic torsional component is represented by a 1-D Fourier series of order 6.

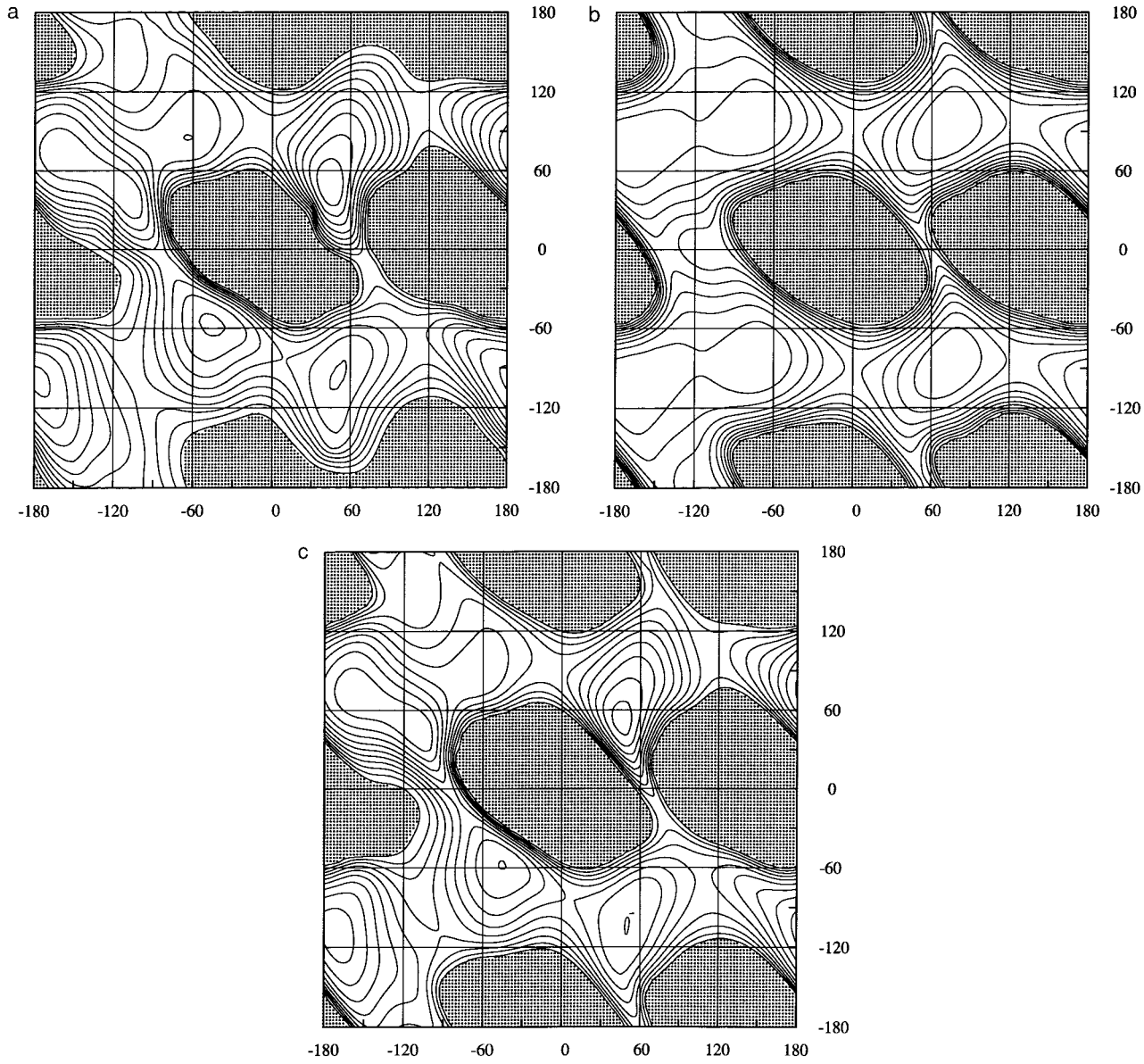

FIGURE 2. Contour plots of energy surfaces for (χ_1, χ_2) of blocked histidine: (a) target *ab initio*, (b) repulsion + dispersion, and (c) repulsion + dispersion + electrostatic + intrinsic torsional. Contour levels range from 1 to 10 kcal / mol in increments of 1 kcal / mol.

These coefficients were adjusted to fit experimental data.

HYDROPHOBIC COMPONENT

The hydrophobic effect is estimated using a Gaussian volume implementation of a hydration shell model.[5, 25–27] Historically, the basic parameters of a hydration shell model consist of, for each atom, two radii (inner and outer) that specify the spatial extent of the hydration shell and a factor that relates the volume of the unoccupied portion of this shell to hydration free energy. Because we seek here to represent only the energy of the hydrophobic effect, we need only a subset of the full flexibility of the traditional model. A single parameter is used to relate the volume of any unoccupied portion of any shell to hydrophobic energy. The thickness of the hydration shell changes as a function of atom type. As a consequence, only a single volume needs to be evaluated: the volume of the union of the unoccupied portions of all

shells, or equivalently, the volume of the union of the outer hydration spheres minus the volume of the molecule (the union of the inner hydration spheres). Because atomic overlaps tend to be small or constant, the volume of the molecule is assumed constant. Thus, differences between conformations in hydrophobic energy are taken to be proportional to differences in volume defined by the union of the outer hydration spheres. As a further simplification, separate hydration shells are not included about nonpolar hydrogens.

A starting point for a description of the functional form is the following expression for the volume of a collection of intersecting spheres as an integral over Cartesian space,

$$\int d\mathbf{r}\Big(\sum_a S_a(\mathbf{r}) - \sum_{a<b} S_a(\mathbf{r})S_b(\mathbf{r}) + \sum_{a<b<c} S_a(\mathbf{r})S_b(\mathbf{r})S_c(\mathbf{r}) - \sum_{a<b<c<d} S_a(\mathbf{r})S_b(\mathbf{r})S_c(\mathbf{r})S_d(\mathbf{r}) + \cdots \Big), \quad (5)$$

where S_a is the step function volume density of sphere a. Based on a remarkable discovery by Grant,[5] the step functions, $S_a(\mathbf{r})$, are replaced by Gaussian functions, $G_a(\mathbf{r}) = \sigma_a e^{-\eta_a(\mathbf{r}-\mathbf{r}_a)^2}$, where $\mathbf{r}_a$ is the position of the center of sphere a. Given the imperfect physical basis for the relation between shell volume and hydrophobic energy, this approximation,

$$\int d\mathbf{r}\Big(\sum_a G_a(\mathbf{r}) - \sum_{a<b} G_a(\mathbf{r})G_b(\mathbf{r}) + \sum_{a<b<c} G_a(\mathbf{r})G_b(\mathbf{r})G_c(\mathbf{r}) - \sum_{a<b<c<d} G_a(\mathbf{r})G_b(\mathbf{r})G_c(\mathbf{r})G_d(\mathbf{r}) + \cdots \Big), \quad (6)$$

which greatly simplifies evaluation, introduces no significant loss of accuracy. Further substitution is enabled by the Gaussian product theorem,

$$\sigma_1 \exp\left[-\eta_1(\mathbf{r}-\mathbf{r}_1)^2\right] \sigma_2 \exp\left[-\eta_2(\mathbf{r}-\mathbf{r}_2)^2\right] \cdots \sigma_n \exp\left[-\eta_n(\mathbf{r}-\mathbf{r}_n)^2\right]$$
$$= \sigma_1\sigma_2\cdots\sigma_n \times \exp\left[-\begin{Vmatrix} \eta_1\eta_2 r_{12}^2 & \eta_1\eta_3 r_{13}^2 & \cdots & \eta_1\eta_n r_{1n}^2 \\ & \eta_2\eta_3 r_{23}^2 & \cdots & \eta_2\eta_n r_{2n}^2 \\ & & \ddots & \vdots \\ & & & \eta_{n-1}\eta_n r_{n-1\,n}^2 \end{Vmatrix} \Big/ (\eta_1+\eta_2+\cdots+\eta_n) \right]$$
$$\times \exp\left\{ -(\eta_1+\eta_2+\cdots+\eta_n)\left[\mathbf{r} - \frac{\eta_1\mathbf{r}_1+\eta_2\mathbf{r}_2+\cdots+\eta_n\mathbf{r}_n}{(\eta_1+\eta_2+\cdots+\eta_n)}\right]^2 \right\}, \quad (7)$$

which specifies the relation between the product of two or more Gaussians (itself a Gaussian) and its factors, r_{ij} being used to denote $|\mathbf{r}_i - \mathbf{r}_j|$. Integration, now straightforward, gives

$$\sum_a \sigma_a \left(\frac{\pi}{\eta_a}\right)^{3/2} - \sum_{a<b} \sigma_a\sigma_b \times \exp\left[-\frac{\eta_a\eta_b r_{ab}^2}{(\eta_a+\eta_b)}\right]\left(\frac{\pi}{\eta_a+\eta_b}\right)^{3/2} + \sum_{a<b<c} \sigma_a\sigma_b\sigma_c \times \exp\left[-\frac{\eta_a\eta_b r_{ab}^2 + \eta_a\eta_c r_{ac}^2 + \eta_b\eta_c r_{bc}^2}{(\eta_a+\eta_b+\eta_c)}\right]\left(\frac{\pi}{\eta_a+\eta_b+\eta_c}\right)^{3/2} + \cdots, \quad (8)$$

an easily evaluated expression for Grant's Gaussian volume. For organic molecules, given typical hydration shell thicknesses (about 1 Å for carbon) and a subset of space defined as the union of outer hydration spheres, convergence of this series to within a few percent of its limiting value requires evaluation out beyond seven-body intersections.

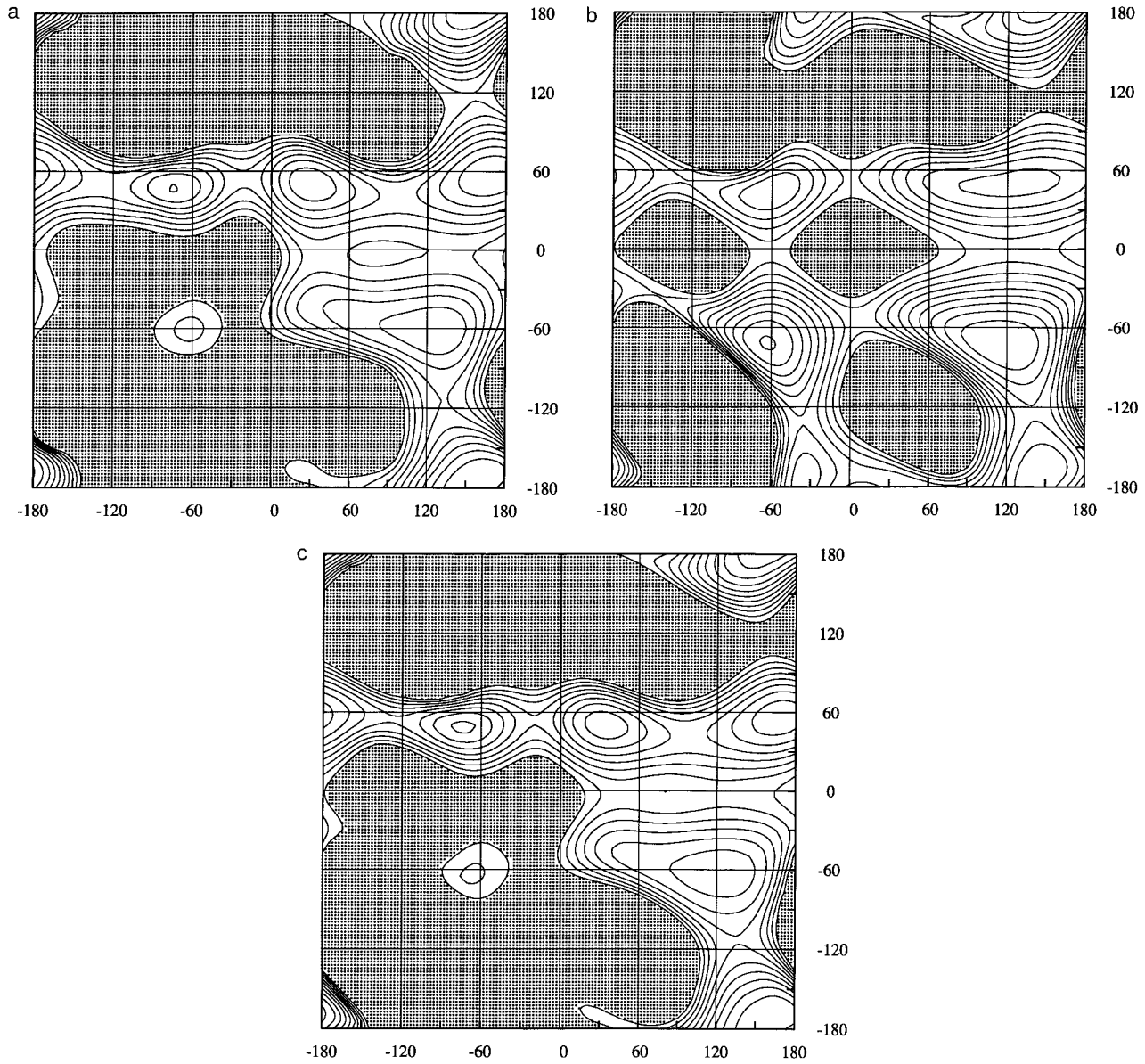

FIGURE 3. Contour plots of energy surfaces for (ψ, χ_1) of blocked threonine: (a) target *ab initio*, (b) repulsion + dispersion, and (c) repulsion + dispersion + electrostatic + intrinsic torsional. Contour levels range from 1 to 10 kcal / mol in increments of 1 kcal / mol.

Let β_a be the radius of step function $S_a(\mathbf{r})$. The general Gaussian function, $G_a(\mathbf{r})$, has two independent parameters, σ_a and η_a. These are specified, as a function of β_a, by the following two conditions. The integrals over all space of $G_a(\mathbf{r})$ and $S_a(\mathbf{r})$ are taken to be equal,

$$\int d\mathbf{r}\, \sigma_a e^{-\eta_a(\mathbf{r}-\mathbf{r}_a)^2} = \sigma_a \left(\frac{\pi}{\eta_a}\right)^{3/2} = \int d\mathbf{r}\, S_a(\mathbf{r}) = \frac{4\pi\beta_a^3}{3}. \quad (9)$$

Alternatively,

$$\sigma_a = \frac{4(\eta_a^{1/2}\beta_a)^3}{3\pi^{1/2}} \quad (10)$$

specifies σ_a as a function of $(\eta_a^{1/2}\beta_a)$. For two identical spheres, the RMSD between the exact volume as a function of separation distance r,

$$\frac{V(r)}{\beta_a^3} = \frac{4\pi}{3} + \pi\left(\frac{r}{\beta_a}\right) - \frac{\pi}{12}\left(\frac{r}{\beta_a}\right)^3 \quad \left(\frac{r}{\beta_a}\right) < 2$$
$$= \frac{8\pi}{3} \quad \left(\frac{r}{\beta_a}\right) > 2 \tag{11}$$

and a Gaussian estimate of volume,

$$\frac{V(r)}{\beta_a^3} = \frac{4\pi}{3}\left\{2 - \left(\frac{2}{\pi}\right)^{1/2} \frac{(\eta_a^{1/2}\beta_a)^3}{3} \times \exp\left[-\frac{(\eta_a^{1/2}\beta_a)^2}{2}\left(\frac{r}{\beta_a}\right)^2\right]\right\}, \tag{12}$$

is taken to be optimal. This condition specifies a universal value (independent of β_a) for $(\eta_a^{1/2}\beta_a)$, 1.49609, obtained by numerical fit in the range $(r/\beta_a) \in [0, 3]$. Substitution in eq. (10) gives σ_a = 2.51905 Å^3, also independent of β_a. Fixing these values, Gaussian volume reproduces exact volume for two identical spheres over the complete range of separation distances (RMSD = 0.10 β_a^3). It is assumed that this level of reproduction also holds for three or more spheres and differing radii.

The inner radii of hydration shells are taken to be atomic van der Waals radii. These are listed in Table II. The coefficient of the linear relationship between unoccupied shell volume and hydrophobic energy is set at 0.032 kcal/mol Å^3. Smaller values lead to thicker (probably more meaningful) shells and greater computational complexity as contributions from six, seven, and higher body intersections increase. The outer radii of hydration shells, or equivalently the η parameters of the substituted Gaussians, were adjusted to fit experimental hydration free energies for a collection of 48 small molecules.

The hydrophobic contribution to hydration free energy, denoted $\Delta F_{\text{hydrophobic}}$, is expected to be proportional to area of exposed nonpolar surface. An estimate of this contribution was obtained using

$$\Delta F_{\text{hydrophobic}} = \Delta F_{\text{hydration}} - \Delta U_{\text{dispersion}} - \Delta U_{\text{electrostatic}}, \tag{13}$$

where $\Delta F_{\text{hydration}}$ is the experimental hydration free energy, $\Delta U_{\text{dispersion}}$ is the dispersion contribution to hydration enthalpy, and $\Delta U_{\text{electrostatic}}$ is the continuous dielectric medium contribution to hydration enthalpy. Values of $\Delta F_{\text{hydration}}$ and its components are listed in Table VI for the 48 model compounds used in parameterization.[28,29] For hydrocarbons, $\Delta U_{\text{dispersion}}$ was estimated using experimental heats of vaporization (measured at T_b but corrected to 298 K) minus RT_b, R being the gas constant and T_b the boiling temperature. For polar compounds, $\Delta U_{\text{dispersion}}$ was taken to be that of the hydrocarbon most similar in size. The subtraction of $\Delta U_{\text{dispersion}}$ from $\Delta F_{\text{hydration}}$, which differs from what has been done before, removes the nonlinearity of the relationship between surface area and hydration free energy for the series methane, ethane, . . . , octane.

For each molecule, the following sequence of calculations was used to obtain an estimate of $\Delta U_{\text{electrostatic}}$. A wavefunction was calculated at standard geometry. Atomic multipoles were calculated from the wavefunction. Interaction energy was calculated between the molecule electric field and a continuous medium of dielectric 80. A description of this calculation is presented in the following section. The boundary between medium and molecule is specified by atomic cavity radii equal to half of the ρ values of Table IV (with some further averaging over similar atom types) plus a displacement of 0.4 Å in combination with a probe radius of 1.6 Å. These are listed in Table VII. The molecule contribution to the electric field at the boundary is that of the atomic multipoles through hexadecapole. These being small molecules, the density of boundary surface elements was much higher than that described in the following section.

In calculating the volumes of the unions of inner and outer hydration spheres, the series was truncated at seven-body intersections. For all of the molecules, volume contributions from intersections of eight or more atoms are negligible for most radii (3.2 Å and smaller). Shell outer radii were adjusted to reproduce hydrophobic energies in five stages. At each stage, the subset of molecules fit includes the subset fit at the previous stage. Increments to this subset consist of {aliphatic hydrocarbons}, {aromatic hydrocarbons}, {thiols}, {aldehydes, ketones, water, alcohols, ethers, acids, esters}, and {ammonia, amines, amides}. Respectively, the subset of locked parameters (those held fixed in all subsequent stages) expands as follows: {C_{08}}, {C_{10}}, {S_{15}}, {H_{05}, O_{13}, O_{14}}, {H_{02}, C_{09}, N_{11}, N_{12}}. Outer hydration spheres were constrained to not become smaller than corresponding inner hydration spheres. All of the H bonding atom types (H_{02}, H_{05}, N_{12}, O_{13}, O_{14}) had to be constrained in

this way. Table VI shows the optimized reproduction of the data. The optimal outer radii are presented in Table II.

Function (8) expresses volume as a sum over elements, elements being intersections of collections of spheres. This is the starting point for a second, less elegant approximation that allows for evaluation as a sum of pairwise functions. A reduced model is constructed as follows. The molecules of a system are viewed as a collection of (sometimes overlapping) rigid segments; rigid segments being largest groupings of atoms such that no pair can change distance given torsion angle rotation within molecules and relative translation–rotation between molecules. Elements can be partitioned into classes based on the (minimal) number of rigid segments that is spanned by the intersecting spheres. Depending on whether an element spans one or more rigid segments, its volume will either remain constant or vary as a function of conformation. Considering first only those elements that span a single segment, elements are retained only for intersections of one, two, or three spheres such that any two spheres are separated by no more than two bonds. Given this reduced set of constant elements, variable elements (those that span two or more rigid segments) are retained only for intersections of two (but not three or more) constant elements. These elements span two rigid segments but not three or more. The reduced model is greatly simplified relative to the exact model. The loss of accuracy is not significant.

In Table VIII shell volume eliminated by the bringing together of two identical monomers is compared for two approximate evaluations of function (8): truncation at seven-body intersections and reduced model. Data is presented for heptane and benzene. Only elements that span both molecules contribute. These include all volume elements of monomer 1 interacting with all volume elements of monomer 2, out to intersections of seven spheres. Rows correspond to separation distances between monomers, displacement being perpendicular to the plane of the heavy atoms. Columns correspond to volume and its decomposition with respect to the number of spheres whose intersection defines an element. Shell radii are those of Table II. Table VIII shows that, for total volume eliminated, the reduced model estimate tends to be slightly high. In addition, for small separation distances, function (8) has not fully converged at seven-body intersections. The neglect of eight, nine, and higher body intersections is a source of significant error for dimer shell volumes. In contrast, reduced model estimates remain reasonable even when the order-7 model is wildly unconverged.

The reduced model is used as a component of our protein energy surface. Each element of volume density,

$$(-1)^{\tau-1}\sigma_{a_1a_2\cdots a_\tau}e^{-\eta_{a_1a_2\cdots a_\tau}(\mathbf{r}-\mathbf{r}_{a_1a_2\cdots a_\tau})^2}, \quad (14)$$

is characterized by the general Gaussian parameters $\mathbf{r}_{a_1a_2\ldots a_\tau}$, $\sigma_{a_1a_2\ldots a_\tau}$, and $\eta_{a_1a_2\ldots a_\tau}$, along with an additional parameter τ that specifies the number of hydration spheres whose intersection defines the element. For elements of volume density that span a single rigid segment, the constant values of these characteristic parameters are precalculated. The positions $\mathbf{r}_{a_1a_2\ldots a_\tau}$ of these constant elements can be thought of as interaction sites. Because the integral of volume density over constant elements never changes, it can be neglected. The set of elements of volume density that span two rigid segments is equivalent to all pairwise intersegment intersections of constant elements, excluding pairs that share a common parent atom. The integral of volume density over the variable element formed from the intersection of constant elements i and j is

$$V_{ij} = \sigma_i\sigma_j\left(\frac{\pi}{(\eta_i+\eta_j)}\right)^{3/2}(-1)^{\tau_i+\tau_j-1} \times\exp\left[\frac{-\eta_i\eta_j r_{ij}^2}{(\eta_i+\eta_j)}\right]. \quad (15)$$

Thus, the form of the hydrophobic component is a sum of easily evaluated pairwise interactions.

For crambin, a 46 residue protein, values of the hydrophobic component are -45.5 kcal/mol for the regularized crystal structure and -17.5 kcal/mol for the fully extended conformation. This hydrophobic stabilization of the folded state is enough to overcome entropy of about 2.8 conformations/residue, assuming that the fully extended conformation is representative of the unfolded state.

DIELECTRIC MEDIUM COMPONENT

The energetic effects of a dielectric continuous medium are calculated with the use of a boundary element solution to the Poisson equation.[4, 30–34] Our implementation of this method differs only slightly from previous ones. The impact of this

energy component on predictions of protein structure has been demonstrated using a variety of computational algorithms.[35–38]

The dielectric boundary, taken to be the molecular surface and calculated using Connolly's analytical algorithm,[30] is specified by the atomic cavity radii of Table VII and a probe radius of 1.6 Å. These cavity radii were obtained as half of the ρ values of Table IV plus a displacement of 0.4 Å followed by averaging over $\{H_{00}, H_{03}\}$, $\{H_{02}, H_{05}\}$, $\{C_{08}, C_{10}\}$, $\{N_{11}, N_{12}\}$, and $\{O_{13}, O_{14}\}$. The boundary surface is partitioned into a collection $\{B_j : j \in \mathscr{B}\}$ of boundary surface elements. The following notation is used for quantities associated with boundary surface element B_j. Let A_j be the surface area, $\mathbf{r}_j$ a single representative point on the surface, $\mathbf{d}_j$ a point dipole of magnitude -1 ecu bohr located at $\mathbf{r}_j$ and directed along the normal to the surface at $\mathbf{r}_j$, and σ_j the surface charge density assumed uniform over B_j. For protein molecules, the density of boundary elements is kept low at ~ 0.6 elements/Å^2. For ribonuclease, a 124 residue protein, this density translates into $\#\mathscr{B} \sim 3500$, $\#\mathscr{B}$ being the number of elements in $\mathscr{B}$. The set of boundary elements is partitioned into groups of neighboring elements $\mathscr{B} = \{\mathscr{B}_J : J \in \mathscr{G}\}$. In general, $\#\mathscr{G}$ is reduced relative to $\#\mathscr{B}$ by a factor of about 3.2. The following notation is used for quantities associated with the group of boundary elements $\mathscr{B}_J$. Let $\mathbf{r}_J$ be a single representative point on the surface $\cup\{B_j : j \in \mathscr{B}_J\}$ and $\mathbf{e}_J$ a point monopole of magnitude 1 ecu located at $\mathbf{r}_J$.

To reduce the complexity of computing the surface charge densities $\{\sigma_j : j \in \mathscr{B}\}$, the following simplification is introduced. For each group of neighboring boundary elements $\mathscr{B}_J$, it is assumed that all induced surface charge, equal to

$$\sum_{j \in \mathscr{B}_J} (\sigma_j A_j), \tag{16}$$

acts from the single representative point $\mathbf{r}_J$.

The contribution to the surface charge of group $\mathscr{B}_J$ that is induced by the atomic multipoles of the protein is

$$Q_J = \frac{(1 - D)}{2\pi(1 + D)} \sum_{j \in \mathscr{B}_J} (\mathbf{q}\,\mathbf{d}_j A_j), \tag{17}$$

where $\mathbf{q}\mathbf{d}_j$, the electrostatic interaction energy in atomic units between the protein charge density $\mathbf{q}$ and the point dipole $\mathbf{d}_j$, is equivalent to the normal component of the electric field at the representative surface point $\mathbf{r}_j$; and D is the dielectric constant, here taken to be 80. Included in $\mathbf{q}$ are atomic multipoles through hexadecapole for each atom. The contribution to the surface charge of group $\mathscr{B}_J$ that is induced by the surface charge of group $\mathscr{B}_I$ is

$$\mathbf{K}_{JI}\left[\sum_{i \in \mathscr{B}_I} (\sigma_i A_i)\right] = \frac{(1 - D)}{2\pi(1 + D)} \times \sum_{j \in \mathscr{B}_J} \left(\mathbf{e}_I \left[\sum_{i \in \mathscr{B}_I} (\sigma_i A_i)\right] \mathbf{d}_j A_j\right), \tag{18}$$

where $\mathbf{e}_I \mathbf{d}_j$ is the electrostatic interaction energy between point monopole $\mathbf{e}_I$ and point dipole $\mathbf{d}_j$. Within each grouping of boundary elements $\mathscr{B}_I$, surface–surface interactions between individual elements $(B_{i_1}, B_{i_2}) \in \mathscr{B}_I \times \mathscr{B}_I$ are neglected. Equating for each $J \in \mathscr{G}$ the surface charge of group $\mathscr{B}_J$ to its contributions (17) and (18) gives

$$\sum_{I \in \mathscr{G}} (\mathbf{I} - \mathbf{K})_{JI} \left[\sum_{i \in \mathscr{B}_I} (\sigma_i A_i)\right] = Q_J, \tag{19}$$

$\mathbf{I}$ being the identity matrix, a system of $\#\mathscr{G}$ linear equations that relates the $\#\mathscr{G}$ surface charges. This system of equations is solved using LU decomposition. Required computation time is proportional to $\#\mathscr{G}^3$.

The continuous dielectric medium contribution to hydration free energy is represented as a function of induced surface charges by

$$F_{\mathrm{m}} = \frac{1}{2}\left(\sum_{J \in \mathscr{G}} \mathbf{q}\mathbf{e}_J \left[\sum_{j \in \mathscr{B}_J} (\sigma_j A_j)\right] + \sum_{\substack{(I,J) \in \mathscr{G} \times \mathscr{G} \\ I < J}} \mathbf{e}_I \left[\sum_{i \in \mathscr{B}_I} (\sigma_i A_i)\right] \times \mathbf{e}_J \left[\sum_{j \in \mathscr{B}_J} (\sigma_j A_j)\right]\right), \tag{20}$$

where $\mathbf{q}\mathbf{e}_J$ is the electrostatic interaction energy between the protein charge density $\mathbf{q}$ and the point monopole $\mathbf{e}_J$. The factor $\frac{1}{2}$ roughly corresponds to free energy required to polarize the medium. Alternatively, because the polarization component is neglected from protein–protein interactions, for consistency, it should also be neglected from protein–solvent interactions. Otherwise, the full charges of ionized functional groups can polarize

TABLE VI.

Decomposition of Experimental Hydration Free Energies[a] and Optimal Hydration Shell Model Reproduction of Hydrophobic Contributions.

Molecule	$\Delta F_{hydration}$[b]	$\Delta U_{dispersion}$[c]	$\Delta U_{electrostatic}$[d]	$\Delta F_{hydrophobic}$[e]	Error[f]
Methane	2.00	−0.63	−0.06	2.69	0.80
Ethane	1.83	−2.48	−0.07	4.38	−0.04
Propane	1.96	−3.63	−0.09	5.68	−0.05
Butane	2.08	−4.66	−0.10	6.84	−0.07
Pentane	2.33	−5.61	−0.11	8.05	−0.16
Hexane	2.49	−6.48	−0.12	9.09	−0.07
Heptane	2.62	−7.28	−0.13	10.03	0.12
Octane	2.89	−8.04	−0.14	11.07	0.21
Methylpropane	2.32	−4.35	−0.13	6.80	−0.20
Methylbutane	2.38	−5.32	−0.14	7.84	−0.27
Dimethylpropane	2.50	−4.79	−0.18	7.47	−0.07
Benzene	−0.87	−6.98	−1.30	7.41	0.00
Methanethiol	−1.24	−3.63	−2.25	4.64	−0.26
Ethanethiol	−1.30	−4.66	−2.12	5.48	0.10
Dimethylsulfide	−1.54	−4.66	−2.35	5.47	0.25
Ethanal	−3.50	−3.63	−3.10	3.23	−0.23
Propanal	−3.44	−4.66	−2.79	4.01	0.12
Butanal	−3.18	−5.61	−2.83	5.26	0.11
Pentanal	−3.03	−6.48	−2.83	6.28	0.22
Propanone	−3.85	−4.66	−3.18	3.99	0.85
Butanone	−3.64	−5.61	−2.87	4.84	1.10
Pentanone	−3.53	−6.48	−2.91	5.86	1.30
Water	−6.27	−0.63	−5.51	−0.13	0.13
Methanol	−5.11	−2.48	−3.64	1.01	2.03
Ethanol	−5.01	−3.63	−3.29	1.91	2.19
Propanol	−4.83	−4.66	−3.33	3.16	2.22
Butanol	−4.72	−5.61	−3.33	4.22	2.28
Pentanol	−4.47	−6.48	−3.35	5.36	2.27
Hexanol	−4.36	−7.28	−3.37	6.29	2.47
Heptanol	−4.24	−8.04	−3.38	7.18	2.71
Phenol	−6.62	−8.27	−4.24	5.89	1.25
Dimethylether	−1.90	−3.63	−1.80	3.53	1.42
Ethanoic	−6.70	−4.66	−5.87	3.83	−1.00
Propanoic	−6.47	−5.61	−5.43	4.57	−0.64
Butanoic	−6.36	−6.48	−5.45	5.57	−0.40
Methylethanoate	−3.32	−5.61	−3.61	5.90	−0.62
Ammonia	−4.25	−0.63	−4.28	0.66	−0.66
Methanamine	−4.56	−2.48	−3.38	1.30	1.72
Ethanamine	−4.50	−3.63	−3.01	2.14	1.93
Propanamine	−4.39	−4.66	−3.03	3.30	2.05
Butanamine	−4.29	−5.61	−3.04	4.36	2.11
Pentanamine	−4.10	−6.48	−3.05	5.43	2.17
Dimethylammonia	−4.28	−3.63	−2.24	1.59	3.34
Trimethylammonia	−3.23	−4.35	−1.29	2.41	3.70
Pyridine	−4.70	−6.98	−3.40	5.68	1.03
Acetamide	−9.71	−4.66	−7.31	2.26	0.66
Nmethylacetamide	−10.07	−5.61	−5.80	1.34	3.90
Dimethylacetamide	−8.55	−6.48	−4.81	2.74	3.63

[a] The unit of energy is 1 kcal / mol.
[b] Experimental hydration free energy.[28,29]
[c] Dispersion contribution to hydration enthalpy.
[d] Continuous dielectric medium contribution to hydration enthalpy.
[e] Hydrophobic contribution to hydration free energy. Estimated as $\Delta F_{hydrophobic} = \Delta F_{hydration} - \Delta U_{dispersion} - \Delta U_{electrostatic}$.
[f] Difference between values for the hydrophobic contribution estimated from experiment and calculated using the hydration shell model.

TABLE VII.
Atomic Cavity Radii (Å).

Type	Radius
H_{00}	2.06
H_{02}	1.36
H_{03}	2.06
H_{05}	1.36
C_{08}	2.31
C_{09}	2.07
C_{10}	2.31
N_{11}	2.22
N_{12}	2.22
O_{13}	2.25
O_{14}	2.25
S_{15}	2.48

the medium but not, for example, a protein methyl group, thus creating a strong energetic preference for contact with medium. In this interpretation, the factor $\frac{1}{2}$ roughly corresponds to a removal of polarization from F_m.

The above representation of the F_m energy component contains no parameters that have been adjusted to reproduce experimental data. It is entirely theoretically based. Potentially adjustable parameters include the atomic cavity radii, the probe radius, the dielectric constants inside and outside the boundary surface, and the factor $\frac{1}{2}$ used in the evaluation of F_m. For this reason, we conclude that the dielectric medium contribution to hydration free energy is a large term that is represented here only crudely.

Code for evaluating F_m was developed by the authors. Of the components of our energy surface, F_m is the only one for which code is not yet completed for calculating first and second derivatives. As a consequence, the F_m component has not yet been used in local energy minimization. Its inclusion in global energy minimization is as a single point function evaluation at end points of local minimization on an energy surface that neglects this component.

As an initial validation of the model, the dielectric medium contribution to hydration free energy was calculated for blocked alanine over the *ab initio* (ϕ, ψ) energy surface. It is useful to view this validation in the context of several previous studies in which the conformational energetics of blocked alanine has been gradually uncovered using *ab initio* methods in combination with models of hydration free energy.[39–45] Because blocked alanine is a small molecule, the density of boundary elements was higher than that used for proteins. For each conformation of a 24×24 grid, the charge density **q** of the molecule was taken to be the atomic multipoles through hexadecapole of the corresponding vacuum wavefunction. The hydrated *ab initio* energy surface was then interpolated to a larger 72×72 grid. Figure 1d shows a scatter plot of 865 experimental (ϕ, ψ) values for alanine residues superimposed over the hydrated *ab initio* rigid geometry (ϕ, ψ) energy contour map of blocked alanine. The experimental (ϕ, ψ) values were taken from a set of 57 good protein structures (resolution 1.8 Å or better, R factor 18% or better). Terminal residues were excluded, as were residues that precede or follow a cis peptide bond. As a first approximation, we can relate the densities in the various regions to energies by assuming Boltzmann weights; but, of course, this is not completely justified because the experimental data includes long-range interactions and the energy contour map does not. For example, the long-range interactions of helices and sheets account for the relatively high concentrations in the α and extended regions of the map. Adjusting roughly for long-range interactions, the agreement that is seen in Figure 1d is encouraging. Relative to the vacuum *ab initio* energy surface (Fig. 1a), the changes in the stabilities of the various regions tend to be in the direction of reproducing the experimental distribution.

CRYSTAL STRUCTURE REGULARIZATION

Code for implementing the substituted energy surface in crystal structure regularization, surface loop global minimization, and ligand binding was developed by the authors. Regularization of a protein crystal structure consists of rigid geometry local minimization with respect to all torsion angles of energy plus a sum of ~ 4000 harmonic distance constraints; target distances were taken from the experimental structure.[1] The F_m component of the substituted energy surface was neglected. The full charges of ionized functional groups were scaled by a factor $\frac{1}{8}$. A cutoff distance of 12 Å was used in regularization.

The quality of the fits to experimental coordinates is monitored in Table IX for a collection of protein crystal structures as a function of the weight assigned to the distance constraints. Table rows correspond to protein crystal structures, columns to a gradual reduction of the force constant of the harmonic constraints. Entries indicate RMSDs over backbone heavy atoms. The regular-

TABLE VIII.
Shell Volume Eliminated[a] by Bringing Together Two Identical Monomers, Comparison of Two Approximate Evaluations of Function (8): Truncation at Seven Body Intersections and Reduced Model.

				Heptane				
				Truncation at Seven-Body Intersections				
Dist.[b]	v1[c]	v2[d]	v3	v4	v5	v6	v7	v[e]
3.9	0.0	−523.5	916.6	−848.1	465.9	−160.1	36.6	−112.6
4.5	0.0	−300.4	437.1	−339.6	153.3	−40.8	6.7	−83.7
5.1	0.0	−159.2	187.5	−121.5	45.7	−9.6	1.1	−56.1
5.7	0.0	−78.0	72.4	−39.0	12.5	−2.2	0.2	−34.1
6.3	0.0	−35.3	25.1	−11.3	3.2	−0.5	0.0	−18.7
				Reduced Model				
3.9	0.0	−523.5	800.6	−511.1	117.0	−12.6	0.0	−129.6
4.5	0.0	−300.4	381.8	−200.3	30.9	−2.4	0.0	−90.5
5.1	0.0	−159.2	163.8	−70.1	6.7	−0.4	0.0	−59.2
5.7	0.0	−78.0	63.2	−22.1	1.2	0.0	0.0	−35.6
6.3	0.0	−35.3	21.9	−6.3	0.2	0.0	0.0	−19.4
				Benzene				
				Truncation at Seven-Body Intersections				
	v1	v2	v3	v4	v5	v6	v7	v
3.9	0.0	−476.5	1108.7	−1497.7	1325.4	−783.2	309.2	−14.2
4.5	0.0	−266.8	511.7	−581.2	431.8	−206.7	60.0	−51.2
5.1	0.0	−137.5	211.4	−200.7	127.2	−51.2	11.2	−39.4
5.7	0.0	−65.2	78.2	−61.9	34.3	−12.2	2.2	−24.6
6.3	0.0	−28.5	25.9	−17.0	8.4	−2.7	0.4	−13.5
				Reduced Model				
3.9	0.0	−476.5	943.2	−803.5	238.2	−34.2	0.0	−132.9
4.5	0.0	−266.8	435.3	−303.9	59.2	−6.0	0.0	−82.2
5.1	0.0	−137.5	179.9	−102.3	12.1	−0.8	0.0	−48.7
5.7	0.0	−65.2	66.6	−30.8	2.0	−0.1	0.0	−27.5
6.3	0.0	−28.5	22.1	−8.3	0.3	0.0	0.0	−14.5

[a] Shell radii are those of Table II.
[b] Separation distance (Å) between monomers. Displacement is perpendicular to heavy atom planes.
[c] Volume contribution (Å^3) from one-body intersections.
[d] Volume contribution from two-body intersections.
[e] Total volume eliminated.

ized structures that were selected for use as starting points for surface loop global minimization are those obtained using force constants equal to 1 kcal/mol Å^2. These have RMSDs of less than 0.2 Å. As the constraints are relaxed, the regularized structures deviate further. Because of our current inability to include the F_m component, movements with unconstrained local minimization away from the experimental crystal structures were judged to be a poor measure of the accuracy of the complete energy surface. Also, we note that the packing of proteins in crystals tends to neutralize full surface charges, either through intermolecular salt contacts or binding of counterions. For example the 1PPT, 3EBX, and 5RSA crystal structures contain bound zinc, sulfate, and phosphate ions, respectively. As a consequence, a side chain conformation that is optimal in a crystal environment may

TABLE IX.
RMSD[a] of Regularized Protein Crystal Structures as Function of Weighting Factor for Harmonic Distance Constraints.

Protein Crystal Structure[b]	Weighting Factor[c] 10.00	1.00	0.10	0.01	0.00
1PPT (avian pancreatic polypeptide)	0.12	0.12	0.20	0.54	2.20
1CRN (crambin)	0.10	0.11	0.18	0.35	0.55
4PTI (bpti)	0.16	0.17	0.26	0.41	1.02
3EBX (erabutoxin B)	0.13	0.14	0.26	0.64	0.93
2RHE (immunoglobulin domain)	0.13	0.14	0.23	0.51	0.96
5RSA (ribonuclease A)	0.15	0.16	0.26	0.71	1.25
1LZ1 (lysozyme)	0.14	0.16	0.27	0.54	0.75

[a] Root mean square deviation (Å) from crystal structure over backbone heavy atoms.
[b] Specified as Brookhaven protein data bank entries.
[c] Weighting factor (kcal / mol Å^2) for a sum of harmonic distance constraints. Target distances are taken from the experimental structure.

be highly strained by electrostatic forces when transferred to aqueous solution. Local minimization, unable to explore alternative combinations of side chain rotamers, reduces this strain with motion distributed throughout the structure. To avoid such physically unnecessary motion, energy function validation through protein local minimization would first require a careful repositioning of surface side chains.

In Table X a decomposition of energy is monitored for crambin as the structure descends the energy surface. The movement is seemingly driven by the 39 kcal/mol decrease in the intrinsic torsional component, although the electrostatic and repulsion components also participate. The disulfide bond component is represented by a sum of harmonic distance constraints, with parameters taken from ECEPP.

SURFACE LOOP GLOBAL MINIMIZATION

Validation of the substituted energy surface is through surface loop global minimization for proteins of known structure. The global search algorithm, used here as a tool for generating predicted structures, remains essentially unchanged from that described previously.[1] A summary of modifications and current timings is presented in Table XI.

As described in Table XI, local minimizations are carried out on the F_{vac} energy surface. The dielectric medium component F_m is added as a single point function evaluation at end points of

TABLE X.
Energy Components[a] of Energy Minimized 1CRN (Crambin) as Function of Weighting Factor for Harmonic Distance Constraints.

Wt.[b]	F_{vac}[c]	F_r	F_e	F_s	F_t	F_c	F_h	RMSD[d]
100.00	5152.6	331.4	−1721.6	11.2	−717.4	7294.6	−45.6	0.10
10.00	−1414.8	328.8	−1722.0	10.5	−718.3	731.8	−45.6	0.10
1.00	−2083.7	321.9	−1725.5	8.2	−724.8	81.9	−45.4	0.11
0.10	−2175.4	316.5	−1734.4	6.9	−738.7	19.4	−45.2	0.18
0.01	−2207.1	314.0	−1741.8	7.1	−750.5	9.2	−45.1	.35
0.00	−2219.8	315.4	−1739.9	6.1	−756.7	0.0	−44.8	0.55

[a] The unit of energy is 1 kcal / mol.
[b] Weighting factor (kcal / mol Å^2) for a sum of harmonic distance constraints.
[c] Total energy. $F_{vac} = (F_r + F_e + F_s + F_t + F_c + F_h)$. Components include repulsion + dispersion F_r, electrostatic F_e, disulfide bond F_s, intrinsic torsional F_t, harmonic distance constraint F_c, hydrophobic F_h, and dielectric medium F_m.
[d] Root mean square deviation (Å) from crystal structure over backbone heavy atoms.

TABLE XI.
Current Timings for Surface Loop Global Minimization.[a]

Action	#Confs	Time
Analytical backbone deformations[b]	559	Instantaneous
Cluster[c]	293	Instantaneous
Remove overlaps[d]	254	Instantaneous
Single point F_h calculation[e]	254	Instantaneous
Cluster	220	Instantaneous
Exclude deformations having $F_h > 8$[f](kcal / mol)	212	Instantaneous
Local energy minimization[g] of $F_{vac} = (F_r + F_e + F_s + F_t + F_c + F_h)$, 15 Å cutoff, F_e truncated at r^{-2}	212	6:53 h
Single point F_m calculation[h]	212	Part of above
Cluster	46	Instantaneous
Exclude conformations having $F_{tot} > 24$[i](kcal / mol)	41	Instantaneous
Side chain global energy minimization[j] of F_{tot}, 15 Å cutoff, F_e truncated at r^{-2}	41	5:16 h
Local energy minimization[k] with respect to all torsions of $F_{vac} = (F_r + F_e + F_s + F_t + F_c + F_h)$,15 Å cutoff, F_e truncated at r^{-5}	42	4:39 h
Single point F_m calculation	42	Part of above
Cluster	41	Instantaneous
Local energy minimization[l] of $F_{vac} = (F_r + F_e + F_s + F_t + F_c + F_h)$, no cutoff, F_e truncated at r^{-5}	10	:30 h
Single point F_m calculation	10	Part of above
Cluster	10	Instantaneous

[a] For loop (33–39) of crambin using an MIPS R10000 processor.
[b] Deformations span six (ψ, ϕ) pairs. Deformations having (ϕ, ψ) values that do not occur in nature are immediately screened out. In practice, at most $\sim$ 2000 deformations are generated.
[c] Avoids redundancy from multiple deformations within a single potential well.
[d] Extremely fast procedure generates overlap free deformations for use as starting points for local energy minimization. Reduces movement needed to accomplish local energy minimization. Deformations having unremovable overlaps are excluded.
[e] Orders overlap free deformations.
[f] Relative to the lowest value obtained. At most 512 overlap free deformations are passed through to the next step.
[g] Energy components include repulsion + dispersion F_r, electrostatic F_e, disulfide bond F_s, intrinsic torsional F_t, harmonic distance constraint F_c, hydrophobic F_h, and dielectric medium F_m.
[h] At the endpoint of local minimization.
[i] $F_{tot} = (F_{vac} + F_m + F_h)$. In addition, a penalty of 0.5 kcal / mol is assessed against each occurrence of backbone conformational regions C, C*, X, and X*. Relative to the lowest value obtained. At most 128 low energy backbone conformations are passed through to the next step.
[j] Carried out separately for each backbone conformation. For each pair of side chains that can contact, a fast buildup type procedure is used to generate starting points for local energy minimization. Consider, for example, the Arg-Glu pair. For Arg, the number of rotamers is $3 \times 3 \times 3 \times 3$; for Glu, $3 \times 3 \times 6$; $\sim$ 4000 total for the pair. The buildup starts with the first two torsions of Arg. These are placed in 3×3 ideal rotameric conformations. The rest of the Arg side chain and all of the Glu side chain are neglected. Overlaps are removed if possible. The resulting overlap free conformations are ordered based on a single point evaluation of F_h. A limited number is passed to the next stage of the buildup in which the first two torsions of Glu are added in 3×3 ideal rotameric conformations. The procedure continues in this manner, generating in very little time a collection of overlap free conformations for the Arg-Glu pair. At most 16 are selected as starting points for local minimization on the F_{vac} energy surface. F_m is added as a single point function evaluation. The lowest energy conformation is retained, and the procedure progresses to the next pair of side chains. In a worst case scenario, where several side chains are present and many backbone conformations have been carried over from the backbone search, there can be as many as 300 side chain pair searches.
[k] Allows the backbone and the side chains to relax together. At this point of the search, the starting structure, in this case the native structure, is inserted into the collection of low energy conformations. This allows the loop search algorithm to be used for generating steps of a trajectory. It guarantees that each conformation of a trajectory, regardless of the completeness of the search, will have lower energy than the previous conformation.
[l] At most 10 low energy conformations are passed through to this step.

TABLE XII.
Energy,[a] Sequence of (ϕ, ψ) Regions, and RMSD from Regularized Crystal Structure[b] for Low Energy Structures of Avian Polypeptide Surface Loops[c].

			Loop (6–12): Thr-Tyr-Pro-Gly-Asp-Asp-Ala							
No.	F_{tot}	(ϕ, ψ) Regions	F_{vac}	F_m	F_r	F_e	F_s	F_t	F_h	RMSD
1	−500.5	F E A F* B B F	−237.7	−256.6	38.2	−220.3	0.0	−49.4	−6.1	0.52 (1.22)
N 2	−493.5	F F A E*A B F	−230.5	−256.8	33.8	−215.7	0.0	−42.6	−6.1	0.53 (0.62)
3	−493.1	X E A E*A B F	−242.4	−246.4	50.6	−228.0	0.0	−60.2	−4.7	1.63 (2.41)
4	−481.1	C D A G* X E E	−220.4	−256.3	43.1	−227.6	0.0	−30.5	−5.3	3.74 (4.07)
5	−474.8	D D B C B B F	−199.9	−269.8	57.0	−193.7	0.0	−57.5	−5.6	2.36 (4.62)
6	−473.9	D D C E A B F	−211.3	−257.7	55.4	−211.7	0.0	−49.8	−5.3	2.13 (4.58)
7	−470.2	F E A A*A*B E	−195.6	−269.1	56.7	−198.8	0.0	−48.0	−5.5	1.57 (3.36)
8	−469.0	D D F C*A* X E	−200.5	−264.4	53.6	−204.6	0.0	−44.5	−5.0	2.78 (4.76)

[a] Calculated without a cutoff distance. The unit of energy is 1 kcal / mol.
[b] Calculated over backbone heavy atoms (and alternatively over all heavy atoms) of the surface loop. The unit of distance is 1 Å.
[c] The structure that corresponds to the regularized crystal structure is marked by an N in column 1.

local minimization. The quantity F_{tot}, on which ordering of local minima is based, includes a doubling of the hydrophobic component F_h, with the justification that long-range electrostatic contributions to energy differences between conformations would be damped by roughly half had the polarization component been included. Also included in F_{tot} is a penalty of 0.5 kcal/mol on any occurrences of backbone conformations C, C*, X, or X*; conformational regions are those defined previously.[1]

Ten surface loops were selected. These are specified in Tables XII–XV. Global energy minimization was applied. The adequacy of the proposed energy functions is judged based on their ability to distinguish the crystal structures from the resulting collections of low energy local minima.

TABLE XIII.
Energy, Sequence of (ϕ, ψ) Regions, and RMSD from Regularized Crystal Structure for Low Energy Structures of Crambin Surface Loops[a].

			Loop (17–25): Arg-Leu-Pro-Gly-Thr-Pro-Glu-Ala-Ile							
No.	F_{tot}	(ϕ, ψ) Regions	F_{vac}	F_m	F_r	F_e	F_s	F_t	F_h	RMSD
N 1	−526.4	A A B A*F F A A A	−410.3	−105.0	63.9	−410.9	0.0	−52.5	−10.9	0.52 (0.50)
2	−513.8	B A F C F F A A A	−404.2	−100.6	60.5	−403.9	0.0	−51.5	−9.4	0.86 (1.19)
3	−505.5	B A B B*F F A A A	−396.8	−98.3	65.5	−396.8	0.0	−55.3	−10.3	0.56 (0.99)
4	−498.3	B A X C G X A A A	−387.5	−103.5	74.3	−402.7	0.0	−50.5	−8.7	2.19 (2.49)
5	−496.9	C A F X D X X A A	−378.8	−111.2	74.9	−385.5	0.0	−59.2	−8.9	2.11 (3.01)
6	−494.7	B A B A*X F X A A	−392.7	−94.1	68.9	−396.3	0.0	−56.5	−8.8	2.17 (2.56)
7	−494.4	C A A C*A X A A A	−395.4	−91.6	83.0	−408.3	0.0	−61.3	−8.8	2.00 (2.49)
8	−488.3	B A B B*D C E A A	−371.0	−108.7	87.5	−379.4	0.0	−70.2	−9.0	2.44 (3.85)
			Loop (33–39): Ile-Ile-Ile-Pro-Gly-Ala-Thr							
N 1	−396.1	E E E A F B E	−281.7	−104.0	42.4	−279.7	0.0	−34.2	−10.3	0.38 (0.41)
2	−392.1	E E E A E C D	−274.6	−108.3	46.4	−279.9	0.0	−31.6	−9.6	0.63 (0.87)
3	−390.0	E E E A F D E	−275.8	−104.3	46.7	−275.7	0.0	−37.1	−9.8	0.50 (0.97)
4	−387.1	E E E A F C D	−268.9	−109.3	50.5	−280.3	0.0	−29.8	−9.3	0.71 (1.26)
5	−385.5	E D D A C*X F	−263.3	−112.8	51.2	−260.5	0.0	−43.8	−10.2	3.35 (3.57)
6	−384.7	E D D F A X F	−266.2	−109.0	45.8	−261.8	0.0	−40.3	−9.9	3.49 (3.64)
7	−384.4	E E E A E*F*E	−270.3	−104.5	52.4	−273.2	0.0	−40.0	−9.5	1.86 (1.86)
8	−382.1	E E E C E A E	−265.4	−107.4	46.2	−274.5	0.0	−27.4	−9.7	1.95 (1.82)

[a] See footnotes of Table XII.

Results and Conclusions

With each application, the global search algorithm generates a collection of 10 well-separated low energy conformations.[1] In Tables XII–XV, these conformations are characterized by the sequence of (ϕ, ψ) conformational regions, energy decomposition, and RMSD from the regularized crystal structure. For each surface loop, the crystal structure is indicated by an N in column 1.

For 4 of the 10 surface loops [loops (17–25) and (33–39) of crambin and loops (14–20) and (44–50) of bpti], the search was unable to find any confor-

TABLE XIV.
Energy, Sequence of (ϕ, ψ) Regions, and RMSD from Regularized Crystal Structure for Low Energy Structures of BPTI Surface Loops[a].

No.	F_{tot}	(ϕ, ψ) Regions	F_{vac}	F_m	F_r	F_e	F_s	F_t	F_h	RMSD
			Loop (14–20): Cys-Lys-Ala-Arg-Ile-Ile-Arg							
N 1	−395.8	F D F D F F E	98.7	−485.1	54.0	80.3	2.1	−28.3	−9.3	0.36 (0.33)
2	−392.5	F D F D E F E	95.9	−479.8	60.4	81.1	2.0	−39.1	−8.5	0.36 (1.27)
3	−386.7	F E E D E F E	109.8	−488.6	53.1	97.6	1.6	−34.7	−7.9	0.78 (1.98)
4	−383.3	F D F D E F E	103.0	−477.7	55.8	84.0	2.0	−30.2	−8.5	0.28 (1.60)
5	−381.3	F E E D E E E	108.3	−481.2	63.0	87.6	1.8	−35.6	−8.5	1.21 (2.22)
6	−380.6	F D F C E E E	104.7	−477.8	62.4	84.7	2.0	−36.4	−8.0	0.72 (1.78)
7	−373.7	A A* X D E F E	123.0	−488.4	64.6	94.6	5.1	−32.6	−8.7	1.14 (1.74)
8	−367.3	F D F D A X* E	126.7	−485.2	65.2	102.4	1.9	−33.5	−9.3	0.81 (1.51)
			Loop (23–29): Tyr-Asn-Ala-Lys-Ala-Gly-Leu							
1	−725.4	F E A B BA* E	−222.8	−489.5	34.0	−205.3	0.0	−38.5	−13.0	0.19 (1.52)
N 2	−723.5	F E A B B B* E	−229.6	−481.2	36.0	−217.4	0.0	−35.7	−12.6	0.26 (0.22)
3	−713.0	F E A B X B* E	−222.3	−478.3	35.3	−209.4	0.0	−35.3	−12.9	0.69 (1.97)
4	−706.9	F A F* C* D F X	−218.2	−476.9	39.1	−206.0	0.0	−38.7	−12.7	2.45 (2.40)
5	−705.4	F C F* B A D* X	−206.0	−487.6	42.9	−197.6	0.0	−38.7	−12.7	1.71 (2.25)
6	−705.2	X B* A* A D X* X	−208.2	−483.6	49.9	−199.3	0.0	−44.0	−14.8	2.49 (3.30)
7	−704.6	E X* C A* A* D E	−211.3	−481.4	33.1	−206.2	0.0	−25.4	−12.7	2.94 (3.11)
8	−702.4	F E A B D C* X	−223.4	−467.1	36.5	−213.8	0.0	−33.3	−12.8	0.76 (2.00)
			Loop (36–42): Gly-Gly-Cys-Arg-Ala-Lys-Arg							
1	−388.2	B D* E A* F F A	86.3	−464.2	43.1	80.9	3.2	−30.6	−10.3	0.30 (2.30)
2	−373.7	B D* E X* B F* G	103.0	−466.8	52.0	86.0	3.4	−28.0	−10.4	0.97 (2.19)
3	−367.6	F C E A* F F A	112.1	−470.4	36.6	104.4	3.5	−22.6	−9.8	0.73 (2.35)
4	−366.5	B D* A E F F B	109.4	−466.2	42.4	94.7	6.2	−24.3	−9.7	0.92 (2.21)
5	−366.1	E C X E E F B	108.5	−466.0	31.2	97.5	6.8	−17.4	−9.6	1.13 (2.57)
6	−365.6	B D* E C* X* E A	114.1	−470.1	52.3	94.7	3.2	−25.6	−10.6	0.73 (2.84)
7	−365.1	C X* E A* F F A	111.0	−467.1	43.8	84.7	4.2	−11.6	−10.0	0.52 (2.36)
N 8	−365.1	B D* E A* F F A	115.5	−470.0	43.8	110.6	3.1	−31.5	−10.5	0.30 (0.33)
			Loop (44–50): Asn-Phe-Lys-Ser-Ala-Glu-Asp							
N 1	−946.4	D E B E A A A	−461.0	−475.3	24.8	−433.7	0.0	−41.9	−10.1	0.20 (0.24)
2	−929.2	D E E X* A A A	−447.6	−472.0	21.5	−422.7	0.0	−36.4	−10.0	0.69 (1.25)
3	−927.4	D E E A* C* A A	−448.5	−470.2	43.2	−438.9	0.0	−43.7	−9.1	1.49 (1.68)
4	−916.9	D E B E A A A	−466.4	−440.5	29.9	−442.3	0.0	−44.0	−10.0	0.22 (0.95)
5	−916.5	D E B D C* B A	−458.6	−448.8	44.1	−445.2	0.0	−48.1	−9.5	1.68 (1.79)
6	−907.8	X* B C C X A A	−436.0	−465.6	53.2	−436.7	0.0	−44.2	−8.2	3.46 (4.67)
7	−901.5	X* X X X A A A	−463.4	−432.1	64.8	−456.2	0.0	−64.1	−7.8	2.55 (4.08)
8	−891.6	X* F A* B C A* B	−395.3	−489.5	59.9	−384.8	0.0	−62.5	−7.7	4.14 (5.10)

[a] See footnotes of Table XII.

mations with energy below that of the crystal structure. We note that, if not for the reinsertion of the initial undeformed conformation (footnote k of Table XI), the crystal conformation would not, in these four cases, have been recovered by the global search. In each case, a similar backbone conformation was recovered, but these were combined with less stable side chain conformations. Because current values of clustering and side chain buildup parameters have been selected to balance completeness with efficiency, the searches are not always complete. Also, the attainment of true completeness is complicated by our current inability to carry out local minimization on the full energy surface F_{tot}. For another four of the surface loops [loop (6–12) of avian polypeptide, loops (23–29) and (36–42) of bpti, and loop (18–24) of lysozyme], the search finds what is essentially the native backbone conformation in combination with a more stable positioning of the side chains. Due largely to salt contacts in the crystal environment, the most stable side chain conformation is not expected to remain unchanged with transfer to aqueous solution. In addition, for loop (118–124) of lysozyme, the predicted backbone is deformed only slightly relative to the native conformation.

For loop (68–74) of lysozyme, six conformations were found with energies below that of the crystal conformation. The predicted conformation looks physically reasonable. It is compact with star conformational regions only at glycine positions. Analysis of this error suggests that its origin may lie in the inability of the dielectric continuum model to account for two water molecules that are

TABLE XV.
Energy, Sequence of (ϕ, ψ) Regions, and RMSD from Regularized Crystal Structure for Low Energy Structures of Lysozyme Surface Loops[a].

No.	F_{tot}	(ϕ, ψ) Regions	F_{vac}	F_m	F_r	F_e	F_s	F_t	F_h	RMSD
			Loop (18–24): Asp-Gly-Tyr-Arg-Gly-Ile-Ser							
1	−1076.	C B* F A* B* E F	−375.6	−690.7	30.2	−351.8	0.0	−43.5	−10.4	0.27 (0.99)
2	−1071.	A D F A* B* F F	−385.2	−675.7	30.9	−358.1	0.0	−47.5	−10.5	0.65 (1.01)
N 3	−1062.	F B* F A* B* E F	−371.2	−680.7	26.1	−355.7	0.0	−31.1	−10.5	0.26 (0.28)
4	−1054.	A X A D X* C F	−372.5	−673.4	34.2	−351.1	0.0	−46.2	−9.4	1.15 (1.32)
5	−1048.	A D X B B C* E	−363.9	−677.2	50.8	−356.1	0.0	−50.0	−8.6	2.14 (2.43)
6	−1048.	A D C A* D* F F	−366.1	−673.6	32.6	−351.9	0.0	−37.9	−9.0	1.30 (1.64)
7	−1047.	X F* B X D* E F	−368.1	−674.2	43.2	−356.3	0.0	−48.7	−6.4	3.13 (5.84)
8	−1047.	E F F A A D F	−337.7	−704.6	41.9	−325.1	0.0	−50.0	−4.5	4.25 (7.26)
			Loop (68–74): Gly-Lys-Thr-Pro-Gly-Ala-Val							
1	−896.7	A* X D A A* F A	−199.4	−692.0	43.6	−201.9	0.0	−35.3	−5.7	2.09 (2.73)
2	−891.7	A* D E A E D A	−193.3	−693.2	42.0	−200.1	0.0	−30.1	−5.0	1.28 (1.73)
3	−890.6	A* X D A F* E A	−196.0	−689.8	42.4	−205.3	0.0	−28.0	−5.2	2.20 (2.85)
4	−889.3	A* X D B C* A* A	−192.0	−692.6	51.1	−207.2	0.0	−30.2	−5.6	2.19 (2.79)
5	−889.2	A* D E A D A* A	−186.7	−696.5	42.3	−188.4	0.0	−34.7	−5.8	1.06 (1.51)
6	−881.6	C* C E A E D A	−184.8	−692.3	36.4	−195.9	0.0	−19.8	−5.5	1.08 (1.40)
N 7	−880.8	C* C E F A* C A	−177.1	−699.9	35.1	−188.2	0.0	−18.7	−5.2	0.69 (0.68)
8	−879.4	G* A* A X F* D A	−158.3	−715.5	34.5	−162.7	0.0	−24.2	−5.9	2.62 (4.22)
			Loop(118–124): Asn-Arg-Asp-Val-Arg-Gln-Tyr							
1	−965.0	A* X C A A B B	−304.5	−652.1	71.8	−299.9	0.0	−67.0	−9.3	1.07 (2.56)
2	−956.5	A* D C B A* A B	−311.8	−637.1	70.2	−307.6	0.0	−66.4	−8.0	2.14 (2.84)
N 3	−952.7	A* E C B A A B	−274.4	−669.5	60.3	−269.1	0.0	−56.5	−9.1	0.49 (0.69)
4	−950.4	E* X A* A A A B	−276.7	−664.8	87.3	−281.4	0.0	−73.4	−9.2	1.43 (4.04)
5	−948.6	A* X C B X A X	−282.5	−659.3	60.8	−287.8	0.0	−46.8	−8.7	2.18 (2.58)
6	−943.1	A* B A* C B A B	−269.7	−664.3	70.2	−262.2	0.0	−68.3	−9.5	1.13 (1.59)
7	−941.8	A* E B A X B X	−249.5	−683.8	73.3	−248.6	0.0	−64.8	−9.4	2.76 (3.37)
8	−941.4	A* E B A X C X	−253.1	−680.0	72.8	−248.6	0.0	−67.7	−9.7	2.81 (3.33)

[a] See footnotes of Table XII.

partially buried, along with an aspartic acid side chain, in the crystal structure by the surface loop. In 9 of 10 searches, the native backbone conformation was identified correctly, or very nearly correctly, by the substituted energy surface. Although the one failed prediction is most likely a reflection of errors that remain in the energy surface, other explanations are possible. The rotameric states of neighboring charged side chains, here fixed at crystal values, are complimentary to the crystal environment and could conceivably force the movement of the surface loop when modeled in the aqueous environment. Also possible is that the global energy minimum structure has a nativelike backbone but it was not found by the search.

Tables XII–XV show that the full energy surface F_{tot} is a complex balance between offsetting components with no single component dominating. The ranges of the energy components tend to decrease in the order $F_e > F_m > F_r > F_t > F_h$. Because of our neglect of the polarization component, F_{tot} energy differences between conformations are large, perhaps as much as 4 times what would be expected from experiment. One notable success of the current energy surface is the low frequency of occurrence, especially among the lowest energy structures, of (ϕ, ψ) conformational regions G, D, C, X, and C*. In nature these regions occur only rarely in comparison to A, B, E, and F. Another partial success is the predictive capability of the F_h component that, with only a few exceptions [loop (23–29) of bpti and loops (68–74) and (118–124) of lysozyme], tends to be lowest for the native conformation.

Discussion

Returning to the broader perspective that was put forward in the Introduction, substitution of the current energy surface constitutes a first step of an iterative procedure moving toward the attainment of a reliable molecular mechanics based method for predicting protein structure. This first iteration corresponds to a measurable increase in the accuracy with which structure can be predicted. The progress that can be made per iteration with a given amount of effort is relevant to a proper appreciation of the difficulty of the problems that remain and perhaps the time frame for solutions.

Errors in the current energy surface include the following. The polarization component is neglected as is anisotropic repulsion + dispersion. The model assumes rigid geometry. No additional component is included to represent energy of hydrogen bond formation. Although much improved, the hydration free energy component continues to be represented only crudely in comparison to the repulsion + dispersion, electrostatic, and intrinsic torsional components. Parameterization could be improved through increases in the quality and quantity of the *ab initio* data. In future work, a detailed analysis of errors in predicted structures will be used to identify those components of the current energy surface that need to be treated more carefully. We note, however, that before such an analysis can be carried out, it will be necessary to find further instances in which the current energy surface breaks down. This will probably require predictions over segments larger than seven residues, the number of competing conformations increasing exponentially with segment length.

Acknowledgments

A part of this work was supported by NIH Program Project Grant P01 GM24483. The following material is available on the internet via anonymous ftp from dasher.wustl.edu: the contour plot of Figure 1d without superposition of experimental (ϕ, ψ) values, z matrices for 20 blocked amino acids, the *ab initio* energy surfaces of Table I, and the protein structures of Tables XII–XV.

References

1. M. J. Dudek and H. A. Scheraga, *J. Comput. Chem.*, **11**, 121 (1990).
2. M. J. Dudek and J. W. Ponder, *J. Comput. Chem.*, **16**, 791 (1995).
3. T. A. Halgren, *J. Am. Chem. Soc.*, **114**, 7827 (1992).
4. A. A. Rashin and K. Namboodiri, *J. Phys. Chem.*, **91**, 6003 (1987).
5. J. A. Grant and B. T. Pickup, *J. Phys. Chem.*, **99**, 3503 (1995).
6. E. Benedetti, In *Peptides*, M. Goodman and J. Meinhofer, Eds., Wiley, New York, 1977, p. 257.
7. F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, *J. Phys. Chem.*, **79**, 2361 (1975).
8. *SPARTAN Version 2.0*, Wavefunction, Inc., Irvine, CA.
9. S. L. Price, J. S. Andrews, C. W. Murray, and R. D. Amos, *J. Am. Chem. Soc.*, **114**, 8268 (1992).
10. A. J. Stone, *Chem. Phys. Lett.*, **83**, 233 (1981).
11. M. Waldman and A. T. Hagler, *J. Comput. Chem.*, **14**, 1077 (1993).

12. G. J. H. van Nes and A. Vos, *Acta Crystallogr.*, **B34**, 1947 (1978).
13. A. M. Merle, M. Lamotte, S. Risemberg, C. Hauw, J. Gaultier, and J. P. Grivet, *Chem. Phys.*, **22**, 207 (1977).
14. E. G. Cox, D. W. J. Cruickshank, and J. A. S. Smith, *Proc. R. Soc. A*, **247** (1958).
15. G. de With, S. Harkema, and D. Feil, *Acta Crystallogr.*, **B32**, 3178 (1976).
16. E. D. Stevens, *Acta Crystallogr.*, **B34**, 544 (1978).
17. G. de With and S. Harkema, *Acta Crystallogr.*, **B33**, 2367 (1977).
18. S. Swaminathan, B. M. Craven, and R. K. McMullan, *Acta Crystallogr.*, **B40**, 300 (1984).
19. G. Valle, V. Busetti, M. Mammi, and G. Carazzolo, *Acta Crystallogr.*, **B25**, 1432 (1969).
20. P. G. Jonsson, *Acta Crystallogr.*, **B27**, 893 (1971).
21. J. S. Chickos, In *Molecular Structure and Energetics*, Vol. 2, J. F. Liebman and A. Greenberg, Eds., VCH Publishers Inc., New York, 1987, p. 67.
22. H. G. M. de Wit, J. C. van Miltenburg, and C. G. de Kruif, *J. Chem. Thermodyn.*, **15**, 651 (1983).
23. M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery, *J. Comput. Chem.*, **14**, 1347 (1993).
24. K. T. No, O. Y. Kwon, S. Y. Kim, K. H. Cho, C. N. Yoon, Y. K. Kang, K. D. Gibson, M. S. Jhon, and H. A. Scheraga, *J. Phys. Chem.*, **99**, 13019 (1995).
25. Y. K. Kang, G. Némethy, and H. A. Scheraga, *J. Phys. Chem.*, **91**, 4105 (1987).
26. K. D. Gibson and H. A. Scheraga, *Mol. Phys.*, **62**, 1247 (1987).
27. C. E. Kundrot, J. W. Ponder, and F. M. Richards, *J. Comput. Chem.*, **12**, 402 (1991).
28. S. Cabani, P. Gianni, V. Mollica, and L. Lepori, *J. Solution Chem.*, **10**, 563 (1981).
29. R. Wolfenden, L. Andersson, P. M. Cullis, and C. C. B. Southgate, *Biochemistry*, **20**, 849 (1981).
30. M. L. Connolly, *J. Appl. Crystallogr.*, **16**, 548 (1983).
31. R. J. Zauhar and R. S. Morgan, *J. Mol. Biol.*, **186**, 815 (1985).
32. R. J. Zauhar and R. S. Morgan, *J. Comput. Chem.*, **9**, 171 (1988).
33. A. A. Rashin, *J. Phys. Chem.*, **94**, 1725 (1990).
34. R. Bharadwaj, A. Windemuth, S. Sridharan, B. Honig, and A. Nicholls, *J. Comput. Chem.*, **16**, 898 (1995).
35. W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, *J. Am. Chem. Soc.*, **112**, 6127 (1990).
36. A. A. Rashin, *Proteins*, **13**, 120 (1992).
37. K. C. Smith and B. Honig, *Proteins*, **18**, 119 (1994).
38. R. Abagyan and M. Totrov, *J. Mol. Biol.*, **235**, 983 (1994).
39. J. N. Scarsdale, C. Van Alsenoy, V. J. Klimkowski, L. Schafer, and F. A. Momany, *J. Am. Chem. Soc.*, **105**, 3438 (1983).
40. S. J. Weiner, U. C. Singh, T. J. O'Donnell, and P. A. Kollman, *J. Am. Chem. Soc.*, **106**, 6243 (1984).
41. A. G. Anderson and J. Hermans, *Proteins*, **3**, 262 (1988).
42. J. A. Grant, R. L. Williams, and H. A. Scheraga, *Biopolymers*, **30**, 929 (1990).
43. T. Head–Gordon, M. Head–Gordon, M. J. Frisch, C. L. Brooks III, and J. A. Pople, *J. Am. Chem. Soc.*, **113**, 5989 (1991).
44. H. S. Shang and T. Head–Gordon, *J. Am. Chem. Soc.*, **116**, 1528 (1994).
45. T. J. Marrone, M. K. Gilson, and J. A. McCammon, *J. Phys. Chem.*, **100**, 1439 (1996).