ICM—A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation

RUBEN ABAGYAN,* MAXIM TOTROV, and DMITRY KUZNETSOV

European Molecular Biology Laboratory, Postfach 10.2209, Meyerhofstrasse 1, 69012 Heidelberg, Germany

Received 16 June 1993; accepted 15 November 1993

ABSTRACT

An efficient methodology, further referred to as ICM, for versatile modeling operations and global energy optimization on arbitrarily fixed multimolecular systems is described. It is aimed at protein structure prediction, homology modeling, molecular docking, nuclear magnetic resonance (NMR) structure determination, and protein design. The method uses and further develops a previously introduced approach to model biomolecular structures in which bond lengths, bond angles, and torsion angles are considered as independent variables, any subset of them being fixed. Here we simplify and generalize the basic description of the system, introduce the variable dihedral phase angle, and allow arbitrary connections of the molecules and conventional definition of the torsion angles. Algorithms for calculation of energy derivatives with respect to internal variables in the topological tree of the system and for rapid evaluation of accessible surface are presented. Multidimensional variable restraints are proposed to represent the statistical information about the torsion angle distributions in proteins. To incorporate complex energy terms as solvation energy and electrostatics into a structure prediction procedure, a "double-energy" Monte Carlo minimization procedure in which these terms are omitted during the minimization stage of the random step and included for the comparison with the previous conformation in a Markov chain is proposed and justified. The ICM method is applied successfully to a molecular docking problem. The procedure finds the correct parallel arrangement of two rigid helixes from a leucine zipper domain as the lowest-energy conformation (0.5 Å root mean square, rms, deviation from the native structure) starting from completely random

*Author to whom all correspondence should be addressed.

Journal of Computational Chemistry, Vol. 15, No. 5, 488–506 (1994)

CCC 0192-8651/94/050488-19

configuration. Structures with antiparallel helixes or helixes staggered by one helix turn had energies higher by about 7 or 9 kcal/mol, respectively. Soft docking was also attempted. A docking procedure allowing side-chain flexibility also converged to the parallel configuration starting from the helixes optimized individually. To justify an internal coordinate approach to the structure prediction as opposed to a Cartesian one, energy hypersurfaces around the native structure of the squash seeds trypsin inhibitor were studied. Torsion angle minimization from the optimal conformation randomly distorted up to the rms deviation of 2.2 Å or angular rms deviation of 10° restored the native conformation in most cases. In contrast, Cartesian coordinate minimization did not reach the minimum from deviations as small as 0.3 Å or 2°. We conclude that the most promising detailed approach to the protein-folding problem would consist of some coarse global sampling strategy combined with the local energy minimization in the torsion coordinate space.

Introduction

Modeling techniques applicable to complex biomolecular systems so far are largely confined to those based on Cartesian coordinate representation of atoms.[1,2] This approach has an advantage of simplicity and the disadvantages of a large number of variables and mixing of hard and soft degrees of freedom. Using fixed covalent geometry and torsion angles as variables[3] is an alternative approach, which has a principal advantage of significantly reduced number of free variables, allowing efficient conformational searches. A principal contribution to the torsion method was the development of fast analytical algorithms to calculate derivatives of the energy function.[4] However, restricted mechanical models where only torsion angles are variable made applications of the method to complex situations requiring specific constraints of the system and/or consideration of many interacting molecules problematic. Prediction of side-chain conformations, docking of two molecules with either rigid or flexible surface residues, and molecular design where one part of the structure is modeled in the field of the rest of the molecule provide such examples. It has been demonstrated that two or more molecules can be handled by adding three translation vector components and three Eulerian angles to a set of variable dihedral angles.[5,6] A different solution to both multimolecular problems and arbitrary fixation was proposed recently.[7,8] In this approach, arbitrarily fixed molecules were connected by virtual bonds into a unified tree of rigid bodies containing bond lengths, bond angles, and dihedral angles rather than only dihedral angles. Explicit equations of motion were obtained for this system. The proposed methodology and corresponding computer program had some limitations, such as the absence of the variable dihedral phase angle and a restricted algorithm for recurrent summation of components of energy derivatives. This led to restrictions on intermolecular connectivity and torsion angle definitions.

In this article we propose a complete internal coordinate system, where all four types of variables (bond lengths, bond angles, torsion angles, phase dihedral angles) are allowed and the main branch at the branching point of the tree is defined arbitrarily. We will consider the following components of the internal coordinate modeling (ICM) methodology (for the availability of the ICM software, contact R. A.): basic description of the system, summation algorithm for energy derivatives, treatment of bond angle and phase angle deformation energies and their derivatives at nodes containing dependent bond angles, and some new or improved energy/penalty terms which may be used as tools for various molecular modeling tasks. Fast modification of the Shrake and Rupley[9] algorithm for accessible surface evaluation essential for calculations of solvation energy and electrostatics is described. A new type of variable restraint represented by a bell-shaped function and attracting a structure to a multidimensional ellipsoidal zone in the internal coordinate subspace (e.g., a zone around a certain side-chain rotamer) provides a flexible tool for adding statistical or energetical information about torsion angle distributions to modeling.

The ICM methodology has been applied previously to several problems such as structure pre-

diction of peptides,[10] molecular modeling and design,[11,12] and side-chain placement in modeling by homology.[13] Here we address a long-standing problem of protein–protein recognition,[14] which implies prediction of intermolecular association from two individual structures. There are two principal components of a fully automatic docking procedure: the energy (or "fitness") function and the search procedure used to identify the optimal docking configuration. As far as the fitness function is concerned, most of the methods proposed so far use both simplified molecular representation and simplified models of interaction usually based on some measures of shape complementarity and electrostatics.[15–21] Different kinds of grid/systematic searches or Monte Carlo simulated annealing procedures have been used.[22–26] However, simplified models, incomplete set of energy terms, and lack of local energy adjustments (i.e., by minimization) make unambiguous prediction rather difficult especially when some side-chain flexibility should be allowed.[20] Final full-energy refinement for a limited set of configurations found by a rough-energy docking procedure[20,25] improves the situation but does not solve it because the rough energy/fitness function may well miss a configuration potentially converging to the correct answer. Recently, Caflisch et al.[27] refined the "manually" coarsely docked heptapeptide on HIV1 aspartic proteinase by the Monte Carlo procedure combined with minimization (abbreviated as MCM) as proposed by Li and Scheraga.[28] The MCM refinement worked well for the starting conformations close to the correct structure (rmsd of about 2.4 Å). ICM offers a technical possibility of efficient global energy optimization of two rigid or partially flexible molecules with an arbitrary set of free torsion angles (e.g., χ-angles of exposed side chains).

Complex formation of two helixes from GCN4 leucine zipper domain[29] is an example of protein–protein recognition. Nilges and Brünger[30] started from a parallel arrangement of two helixes interacting via leucine side chains and applied a simulated annealing procedure to reproduce the coiled coil geometry without knowledge of the X-ray structure. Comparison carried out after the X-ray structure had been published showed a good prediction accuracy (rmsd for backbone atoms of 1.26 Å).[31] Here we aim at *ab initio* automated docking of the two leucine zipper helixes (both "rigid" and "soft") in full atom representation and ECEPP/2 interaction energy combined with a solvation energy term[32] starting from a completely random relative position of two helixes.

In a general problem of structure prediction by global energy optimization, one of the most important questions is, how large is the vicinity of the native energy minimum from which one can still get to the minimum by a standard energy minimization procedure, or (phrased differently) what does the energy hypersurface around the minimum look like? Using the structure of trypsin inhibitor from squash seeds[33] as an example, we investigated the problem for both torsion angle space and Cartesian space and concluded that a radius of convergence for the minimization is much larger when the covalent geometry is preserved. It emphasizes the potential of the internal coordinate approaches in problems related to the protein structure prediction by global energy minimization.

Methods

MOLECULAR TOPOLOGY

Formal geometrical description imposed on the multimolecular system forms a basis of the method. It is designed to allow efficient manipulations with arbitrarily fixed multimolecular system in the space of internal variables (i.e., dihedral and bond angles and bond lengths). This work extends previously published descriptions[7,8,34] by (1) introducing variable phase angle; (2) using conventional definition of the torsion angles according to the main molecular branch in contrast to the shortest branch, used previously; (3) allowing multiple virtual connections to any node (e.g., to the origin of coordinate frame); (4) using generalized atom node descriptions facilitating molecular editing operations; and (5) abandoning the idea of a tree composed of rigid bodies (so-called BKS-tree[7]), and a tree of variables superimposed on it at nodes. The standard basic regular atomic tree (which does not depend on the variable set) and both variable and rigid body order firmly attached to atomic tree are used instead.

The variable parameters determining the geometry of the multimolecular system are called internal coordinates and are torsion angles, phase angles, bond angles, and bond lengths (Fig. 1a). Some internal coordinates may be replaced by constants, which process will be referred to as fixation. Thus we can consider any system ranging from a

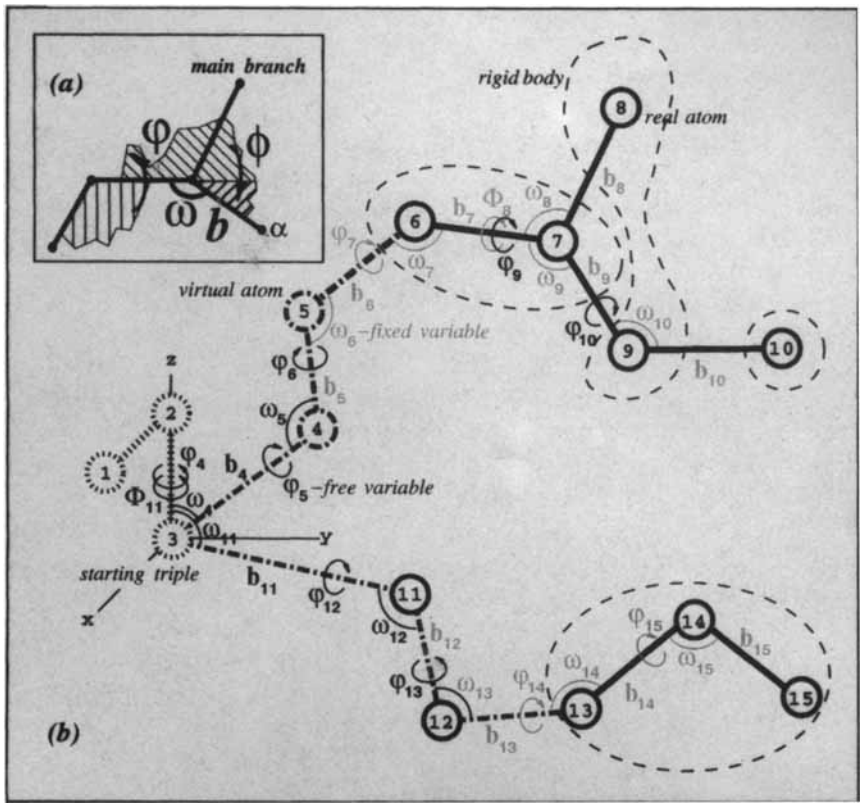

FIGURE 1. (a) Four types of internal variables considered in ICM. (b) The ICM-tree representing the geometry of multimolecular arbitrarily fixed system and containing both real atoms and bonds (continuous lines) and virtual ones (dot-dashed lines). Atoms are numbered so that any atom in the directed graph starts a subtree with a continuous numeration,[7] which is a generalization of the shortest branch rule.[34] An arbitrary subset of free internal variables is shown in bold black characters, all the others being fixed (gray characters). The atomic regular directed graph is the basic one, the order of variables and rigid bodies following it. The numeration does not change as a result of refixation and redefinition of the rigid bodies. The attribution of the main (torsion) branch at the branching point is arbitrary and does not necessarily follow the atomic numeration.

set of geometrically free atoms to a system with standard covalent geometry or a completely rigid molecule. A directed treelike graph is imposed on all atoms of the system as well as on some auxiliary virtual atoms (Fig. 1b). Ring closing covalent bonds are ignored in the graph to allow regular connectivity. The closures may be treated by ring/loop closing algorithms[8,35,36] or considered as fixed rigid bodies. Each atom in this basic description has three geometrical parameters determining its position with respect to a preceding part of a tree. The parameters are bond length b^{α}, bond angle ω^{α}, and torsion (φ^{α}) or phase (Φ^{α}) dihedral angles for side or main branches, respectively (Fig. 1a). The main direction at the branching point defining the torsion branch (all the others are phase branches) is described in a residue library according to normal conventions and convenience. Residue subtrees are joined according to the covalent structure of a polymer molecule. Different molecular subtrees are connected by their root virtual bonds to form the final tree. The default connection is to the third virtual atom fixed at the origin of the coordinate frame. The first two virtual atoms are also fixed and reside at (1, 0, 1, Å) and (0, 0, 1 Å, 1 Å = 0.1 nm) points (Fig. 1b). Each molecule has two virtual atoms attached to the molecule origin in a fixed configuration. They allow resolution of possible singularities in three out of the six variables (two virtual bond angles and one virtual bond length) defining molecular rotation and translation. These two atoms can also be rearranged so that the first of them resides at the geometrical center of the molecule; this makes the specification of the absolute molecular position in space more natural.

The precedence of atoms and variables results

from a geometrical construction procedure and does not depend on further redefinition of the set of variables. The order of atoms is such that the numeration is continuous within any branch. Variable order is attributed to the atomic one so that $n_{\phi^\alpha} = 3n_\alpha$ or $n_{\Phi^\alpha} = 3n_\alpha$, $n_{\omega^\alpha} = 3n_\alpha + 1$, $n_{\beta^\alpha} = 3n_\alpha + 2$, where n_α stands for a sequential number of atom in the regular tree, and n_{b^α}, n_{ω^α}, n_{φ^α}, n_{Φ^α} are numbers of the associated bond length, bond angle, torsion or phase angle, respectively. Correspondingly, $n_\alpha = \text{int}(n_{\theta^\alpha}/3)$, where θ^α is any of three variables attributed to atom α.

Rigid bodies are defined as sets of atoms depending on the same free internal variables. They appear upon fixation of some variables. We will need two more definitions based on the regular order of atoms (see above). The order of rigid bodies with monotonically increasing numbers of their first atoms will be referred to as a regular ascending order. Its inverse will be referred to as a regular descending order.

ENERGY TERMS

The following energy and penalty terms are considered in the ICM program:

$$E = (E_{vw} + E_{hb} + E_{torsion} + E_{el}) + E_{solv} + E_{bonds} + E_{angles} + E_{phases} + E_{dist.restr.} + E_{tethers} + E_{var.restr}$$

The basic energy function (the first four terms) currently used in the program is ECEPP/2.[3,37] It contains a van der Waals potential, a hydrogen-bonding potential, a torsion energy, and the Coulomb electrostatic potential. Functional form and derivatives are given in Tables I and II. The solvation energy term is represented by a sum of atomic accessible surfaces multiplied by solvation parameters.[32,38]

A distance-dependent dielectric constant $\varepsilon = d_{\alpha\beta}$ or $\varepsilon = \varepsilon_0 d_{\alpha\beta}$ (refs. 39 and 40) can be used instead of ε = constant. Although not physically justified, it excludes from the energy calculation the only expression containing a computationally expensive square root (Table I) and at least in some way mimics solvent screening by reducing interaction energy between charges usually residing on the surface with average distances greater than 2–4 Å. Another alternative is using a more elaborate and realistic electrostatic energy based on improved image charge approximation (MIMEL) as described by Abagyan and Totrov.[41] This approximation implies also modified parameters of the solvation energy term.[41]

In most cases, chemical (as opposed to virtual) bond angles, phase angles, and bond lengths are fixed and their deformation parameters are not needed. However, for those rare cases when some

TABLE I.
Energy Function $E_{pairwise}$ and Its Derivatives. Terms, Depending on Distances.

Term	Function	$\frac{1}{d_{\alpha\beta}}\frac{\partial E_{\alpha\beta}}{\partial d_{\alpha\beta}}$[a]
	van der Waals	
E_{vw}[b]	$FA/d^{12} - C/d^6$	$-12FA/d^{14} + 6C/d^8$
	Hydrogen bonding	
E_{hb}	$A'/d^{12} - B/d^{10}$	$-12A'/d^{14} + 10B/d^{12}$
	Electrostatics	
E_{el}[c]	$332q_\alpha q_\beta/\epsilon d$	$-332q_\alpha q_\beta/\epsilon d^3$
	Distance-dep. epsilon	
$E_{\epsilon=\epsilon_0}$[d]	$332q_\alpha q_\beta/\epsilon_0 d^2$	$-664q_\alpha q_\beta/\epsilon_0 d^4$
	Distance restraints	
	$0.25W_\rho(d^2 - D_U^2)^2/D_U^2$,	$W_\rho(d^2 - D_U^2)/D_U^2$ if $d > D_U$
E_ρ[d]	0	0 if $D_L \le d \le D_U$
	$0.25W_\rho(d^2 - D_L^2)^2/D_L^2$,	$W_\rho(d^2 - D_L^2)/D_L^2$ if $d < D_L$
	Tethers	
E_τ	$W_\tau d^2$[e]	$2W_\tau$

[a] $d_{\alpha\beta}$ is the distance between atoms α and β. Indexes α and β are later omitted for clarity.
[b] Factor F is 0.5 for atoms separated by three covalent bonds, and 1.0 otherwise.
[c] Terms E_{vw}, E_{hb}, and E_{elec} are part of the ECEPP/2 potential.
[d] Parameters D_U and D_L are upper and lower bounds for a particular restraint type.
[e] For the tethering term $d = |\mathbf{r}_\alpha - \mathbf{r}_\tau^\alpha|$, where $\mathbf{r}_\tau^\alpha$ is a fixed target point of atom α.

TABLE II.
Energy Function E_{var} and Its Derivatives. Terms, Depending Explicitly on Variables.

Term	Function	$\frac{\partial E_{var}}{\partial \theta}$
	Torsion energy	
$E_{\varphi}^{a,d}$	$K_{\varphi}(1 \pm \cos(n\varphi))$	$\mp nK_{\varphi} \sin(n\varphi)$
	Phase deformation	
$E_{\Phi}^{b,d}$	$0.5K_{\Phi}(\Phi - \Phi_0)^2$	$K_{\Phi}(\Phi - \Phi_0)$
	Bond angle bending	
$E_{\omega}^{c,d}$	$0.5K_{\omega}(\omega - \omega_0)^2$	$K_{\omega}(\omega - \omega_0)$
	Bond stretching	
$E_{b}^{c,d}$	$0.5K_{b}(b - b_0)^2$	$K_{b}(b - b_0)$
	Variable restraints	
E_{ν}^{e}	$\frac{U(\delta^2 - 1)^2(2\delta^2 - 3F^2 + 1)}{(1 - F^2)^3}$	$\frac{24U(\delta^2 - 1)(\delta^2 - F^2)(\theta_i - \theta_{0,i})}{(1 - F^2)^3\Delta_i^2}$
	if $F^2 < \delta^2 < 1$	if $F^2 < \delta^2 < 1$
	U if $\delta^2 \le F^2$, and 0 otherwise	0 if $\delta^2 \le F^2$, or $\delta^2 \ge 1$

[a] Torsion energy E_{φ} is a part of ECEPP/2 potential.
[b] Phase angle deformation energy E_{Φ} is equivalent to improper torsion energy. Parameters are adopted from AMBER force field.
[c] Parameters are adopted from AMBER force field.
[d] K_{φ}, K_{Φ}, K_{ω}, K_b, and φ_0, Φ_0, ω_0, b_0 are force constants and equilibrium variable values for the torsion, phase angle deformation, bond angle bending, and bond stretching energy terms, respectively.
[e] Index ν is omitted for brevity. Contributions from one-variable restraint to the function and ith component of the gradient are given. For explanation of parameters, see text.

of them are allowed to be free, parameters from the AMBER force field[2] are currently used for bond angle bending, bond stretching, and phase angle bending (or improper torsion) terms. This may happen if, for example, one wants to allow more flexibility to fit the model to a poorly refined X-ray structure, or to describe flexible rings. Combining nonbonded parameters from one system of potentials and deformation parameters from the other is undesirable in general; however, as a first approximation, coupling between these two sets can be neglected.

Interaction lists for van der Waals energy, hydrogen bonding energy, local and remote Coulomb electrostatics, and distance restraints are recalculated every time after abrupt conformational changes such as a random Monte Carlo step or a step in a systematic search, as well as after a sufficiently big change during local minimization. To accelerate this essential part of the calculations, we associate atoms close in space into groups. Thus, redundant distance checks for atom pairs are avoided when their group centers are too far apart. Currently Gly, Ala, Cys, Ser, Val, Leu, Thr, Ile, Asp, Asn residues are individual groups with their centers located at C^{α} atoms, whereas larger residues such as Met, Arg, His, Phe, Glu, Gln, Tyr, Trp are divided into two groups, one containing the backbone and C^{β} H_2 atoms and centered at C^{α} and the second containing the rest of a side chain. Usually both atoms of local dipoles belong to the same group. Inclusion of only one charge from the dipole in the neighbor list may lead to considerable inaccuracies for the long-range electrostatic interaction. Therefore, these lists contain the whole group if at least one representative atom resides within the cutoff sphere. Remote electrostatic calculation considers all charges not included in the lists for local electrostatics. Hydrogen-bonding lists are calculated with their own cutoff distance of 3.0 Å. Recalculating interaction lists is still an expensive part of the calculations. A reasonable criterion for the update is based on the maximum absolute atomic displacement during conformational change.[42]

FAST CALCULATION OF ACCESSIBLE SURFACE

The accessible surface of a molecule in the ICM program is used for evaluation of solvation energy[38] and charge depths in electrostatic free-energy calculations. Both terms might be evaluated at each Monte Carlo or systematic search step, so

it is of principal importance that we are able to calculate them sufficiently fast. We used the Shrake and Rupley algorithm,[9] modified to speed up calculations. The surface of each nonhydrogen atom is represented by a set of points nearly uniformly distributed on a sphere. The exposed surface is calculated by counting dots not occluded by neighboring atoms. Only atoms belonging to groups whose centers are separated by less than $2R_{\text{group}} + R_\alpha + R_\beta$, where R_{group} is a radius of a predefined atom group, and R_α, R_β are van der Waals radii of atoms increased by the radius of water molecule, are taken into consideration. This allows us to avoid redundant tests. Second, surface dots are distributed and sorted in such a way that the Z-coordinate is incremented regularly whereas the angular position in a corresponding Z-slice is defined by Fibonacci numbers (Piotr Zielenkiewicz, personal communication). Regular Z-ordering allows us to check accessibility not of all dots on the surface, but only of those between Z-slices Z_l and Z_u (Fig. 2). They can be calculated by the formula

$$Z_l = \begin{cases} Z_c - \delta, & \text{if } \rho_l = \rho_\beta Z_c/Z_\beta + \delta Z_\beta/\rho_\beta > 0; \\ -R_\alpha, & \text{otherwise} \end{cases}$$

$$Z_u = \begin{cases} R_\alpha, & \text{if } \rho_u = \rho_\beta Z_c/Z_\beta - \delta Z_\beta/\rho_\beta \le 0; \\ Z_c + \delta, & \text{otherwise} \end{cases}$$

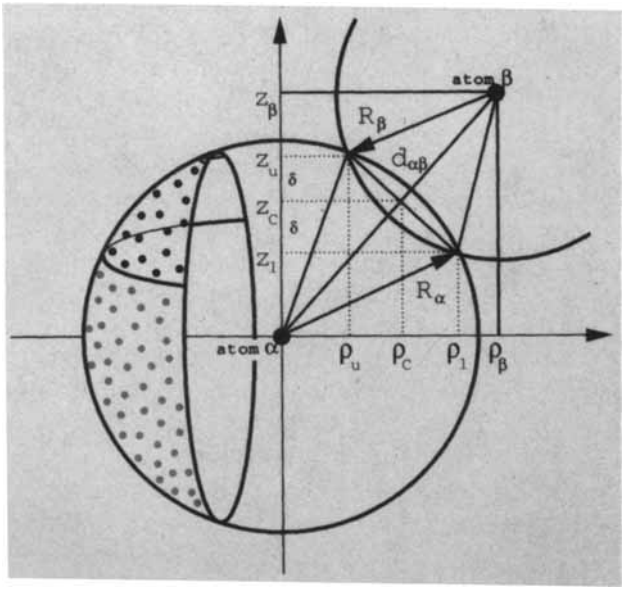

FIGURE 2. Only spherical dots with Z-coordinate between Z_l and Z_u should be checked in fast calculation of the accessible surface of atom α screened by atom β. The cross-section plane goes through atom centres separated by distance $d_{\alpha\beta}$ and the unit vector from atom α in the direction of Z-axis.

where $\rho_\beta = \sqrt{(x_\alpha - x_\beta)^2 + (y_\alpha - y_\beta)^2}$ and $Z_\beta = z_\beta - z_\alpha$ are two cylindrical coordinates of atom β in a coordinate frame at atom α (Fig. 2), and Z_c and δ are expressed as

$$Z_c = \frac{1}{2} Z_\beta \left(\frac{R_\alpha^2 - R_\beta^2}{d_{\alpha\beta}^2} - 1 \right)$$

$$\delta = \sqrt{\rho_\beta^2 \left(\frac{R_\beta^2}{d_{\alpha\beta}^2} - \frac{Z_c^2}{Z_\beta^2} \right)}$$

Note that multiplication of expressions for ρ_l and ρ_u by the positive value ρ_β does not alternate the corresponding conditions but makes possible to use ρ_β^2 instead of ρ_β and, therefore, not calculate the square root. Computer tests showed that substantial reduction of the number of points to be checked in our algorithm results in more than a threefold increase of the speed of surface calculations compared to the standard way.

MULTIDIMENSIONAL VARIABLE RESTRAINTS

Multidimensional ellipsoidal zones in a hyperspace of internal variables described in this section serve two purposes: First, they allow the imposition of a penalty term restraining groups of variables to certain regions (e.g., α or β regions on the ψ-φ map, side-chain rotamers, secondary structure elements); second, the same description is used for making a random step in what we call a biased probability Monte Carlo procedure. A multidimensional variable restraint ν (Fig. 3) imposed on n_ν

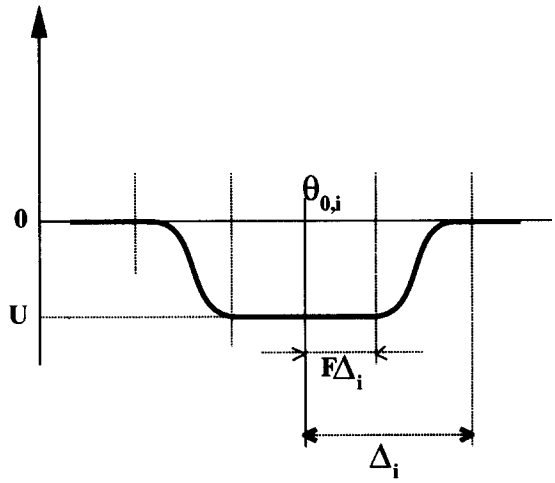

FIGURE 3. Cross-section of the multidimensional ellipsoidal trap in the space of internal variables. It is defined by its center $\theta_{0,i}$, half-axes Δ_i, of the outer ellipsoid, relative size of the flat bottom F, and the trap depth U.

variables θ_i^ν from a subset V_ν is a function of a normalized distance δ:

$$\delta_\nu^2(\theta_1^\nu, \ldots, \theta_{n_\nu}^\nu) = \sum_{i=1}^{n_\nu} \frac{(\theta_i^\nu - \theta_{0,i}^\nu)^2}{(\Delta_i^\nu)^2}$$

where $\theta_{0,i}^\nu$ is the value of ith coordinate of the center of the well and Δ_i^ν is a distance from the center in the θ_i^ν dimension where the penalty function E_ν becomes zero.

The multidimensional variable restraint is represented by a function

$$E_\nu = \begin{cases} U_\nu, & \text{if } \delta_\nu^2 \le F_\nu^2 \\ \dfrac{U_\nu(\delta_\nu^2 - 1)^2(2\delta_\nu^2 - 3F_\nu^2 + 1)}{(1 - F_\nu^2)^3}, & \text{if } F_\nu^2 < \delta_\nu^2 < 1 \\ 0, & \text{if } \delta_\nu^2 \ge 1 \end{cases}$$

where F_ν is a fraction of the well dimension Δ_i^ν occupied by a flat bottom of the function and U_ν is a well depth, which is normally negative to make the well attractive. The derivative of the function with respect to the squared distance δ^2 reads

$$\frac{\partial E_\nu}{\partial \delta_\nu^2} = \frac{6U_\nu(\delta_\nu^2 - 1)(\delta_\nu^2 - F_\nu^2)}{(1 - F_\nu^2)^3}$$

It is seen readily that conditions $\delta_\nu^2 = 1$ and $\delta_\nu^2 = F_\nu^2$ define two ellipsoids in the variable subset V_ν, dividing the subspace into three zones: an internal zone where the penalty function E_ν has the constant function value U_ν, an intermediate zone with a slope toward the center, and an external zone where the penalty is equal to zero. The intermediate part of a function provides a smooth connection (i.e., matches the function and its gradient) between the internal and external parts. Finite size of the well allows combining of many restraints imposed on the same subset of variables (e.g., one can set restraints to the allowed rotameric states of a protein side chain). Restraint zones may also overlap or share the same center. Conformation is attracted to the well only if it gets inside the external ellipsoid.

The multidimensional variable restraints are well suited for representation of the statistical information about the torsion angle distributions in proteins. Parameters for both the main-chain torsion angle distributions and the side-chain rotameric zones for all 20 residues are given elsewhere.[41]

ENERGY DERIVATIVES

We shall consider three functional forms of energy and penalty function components:

$$E = \sum E_{\text{pairwise}}(d_{\alpha\beta}) + \sum E_{\text{var}}(\theta) + \sum E_{\text{complex}}(\theta, \mathbf{r})$$

where the interatomic distance between atoms α and β is $d_{\alpha\beta} = |\mathbf{r}_\alpha - \mathbf{r}_\beta|$. The pairwise term, depending on atom–atom distances, could be composed of a Lennard–Jones energy, a hydrogen-bonding energy, a Coulomb energy, a distance restraint, and a tethering restraint. The second group of energy terms, namely torsion energy, bond stretching, bond angle bending, improper torsion (or phase angle deformation) energy, and simple or multidimensional variable restraints, depends on internal variables explicitly and can be differentiated directly. The third term represents energy components which are not differentiated and not used during local minimization. Energies such as solvation energy and more elaborate electrostatic energy are added only at the end of local minimization to participate in the Monte Carlo selection procedure.

The derivatives $\partial E_{\text{pairwise}}/\partial\theta_i$ can be expressed via partial sums over atom pairs $\mathbf{F}_i$ and $\mathbf{G}_i$ (refs. 7 and 34):

$$\frac{\partial E_{\text{pairwise}}}{\partial \theta_i} = \begin{cases} -\mathbf{e}_i\mathbf{F}_i - (\mathbf{e}_i \times \mathbf{r}_i)\mathbf{G}_i, & \text{if } \theta_i \text{ is torsion, phase or bond angle} \\ \mathbf{e}_i\mathbf{G}_i, & \text{if } \theta_i \text{ is a bond length} \end{cases} \quad (1)$$

where $\mathbf{e}_i$ is a unit vector along a rotation axis of the variable θ_i, and $\mathbf{r}_i$ is a radius–vector from the origin to any point on the rotation axis. To find $\mathbf{F}_i$ and $\mathbf{G}_i$ for the particular θ_i sums, $\mathbf{f}_i$ and $\mathbf{g}_i$ are calculated first for every rigid body ρ:

$$\mathbf{f}_i = \sum_{\alpha\in\rho,\beta\notin\rho} \frac{\partial E_{\alpha\beta}}{\partial d_{\alpha\beta}} \frac{(\mathbf{r}_\alpha \times \mathbf{r}_\beta)}{d_{\alpha\beta}}$$

$$\mathbf{g}_i = \sum_{\alpha\in\rho,\beta\notin\rho} \frac{\partial E_{\alpha\beta}}{\partial d_{\alpha\beta}} \frac{(\mathbf{r}_\alpha - \mathbf{r}_\beta)}{d_{\alpha\beta}}$$

In both cases elementary contributions are antisymmetric upon permutation of atoms α and β. Therefore, each atomic pair can be calculated only once if every time the contribution is added to $\mathbf{f}_i$ or $\mathbf{g}_i$ the same amount is subtracted from the corresponding sums for a rigid body containing atom β.

To compute $\mathbf{F}_i$ and $\mathbf{G}_i$, a summation of $\mathbf{f}_i$ and $\mathbf{g}_i$ on the tree of rigid bodies should be carried out.

It can be done by two procedures. The first one is as follows: Initialize all $\mathbf{F}_i$ and $\mathbf{G}_i$ by $\mathbf{f}_i$ and $\mathbf{g}_i$, then go through the first atoms of every rigid body from the top to the bottom in regular descending order and apply the following recurrent operation:

$$\mathbf{F}_\rho = \mathbf{F}_\rho + \mathbf{F}_i$$
$$\mathbf{G}_\rho = \mathbf{G}_\rho + \mathbf{G}_i \qquad (2)$$

where ρ is a rigid body containing the atom preceding the first atom of a rigid body i in the basic topological tree.

After summation by eq. (2) values, $\mathbf{F}_\rho$ and $\mathbf{G}_\rho$ are used to evaluate derivatives for phase angles, bond angles, and bond lengths, but not for torsion angles. There may be a situation when a torsion angle is free and rotates two or more rigid bodies. It happens if either phase angle, bond angle, or bond length of one of the branches is free. During the second procedure, at all those nodes $\mathbf{F}_\rho$ and $\mathbf{G}_\rho$ should be summed over all branches before using eq. (1) for torsion angle derivatives. Differentiation of E_{var} terms is straightforward, and formulas for the terms mentioned are shown in Table II. However, the deformation energy of dependent bond angles should be treated separately.

Figure 4 shows a fragment with two branches coming out of the central atom β. Bond angle ω_{dep} does not participate in the construction and depends on ω_1, ω_2, and Φ. The cosine theorem for the spherical triangle P_1, P_2, P_3 reads

$$\cos(\omega_{dep}) = \cos(\omega_1)\cos(\omega_2) + \sin(\omega_1)\sin(\omega_2)\cos(\Phi)$$

By differentiating this formula with respect to ω_1, ω_2, and Φ one gets

$$\frac{\partial\omega_{dep}}{\partial\omega_1} = \frac{\sin(\omega_1)\cos(\omega_2) - \cos(\omega_1)\sin(\omega_2)\sin(\Phi)}{\sin(\omega_{dep})}$$

$$\frac{\partial\omega_{dep}}{\partial\omega_2} = \frac{\sin(\omega_2)\cos(\omega_1) - \cos(\omega_2)\sin(\omega_1)\sin(\Phi)}{\sin(\omega_{dep})}$$

$$\frac{\partial\omega_{dep}}{\partial\Phi} = \frac{\sin(\omega_1)\sin(\omega_2)\sin(\Phi)}{\sin(\omega_{dep})}$$

If the bond angle bending energy term E_ω is to be included for ω_{dep}, components of the gradient corresponding to independent variables $\theta_{ind} = \omega_1$, ω_2, and Φ should be changed by

$$\frac{\partial E_{dep}}{\partial\omega_{dep}}\frac{\partial\omega_{dep}}{\partial\theta_{ind}}$$

where the first partial derivative is $K_\omega(\omega_{dep} - \omega_0)$ (Table II) and the second one is one of the three partial derivatives presented above.

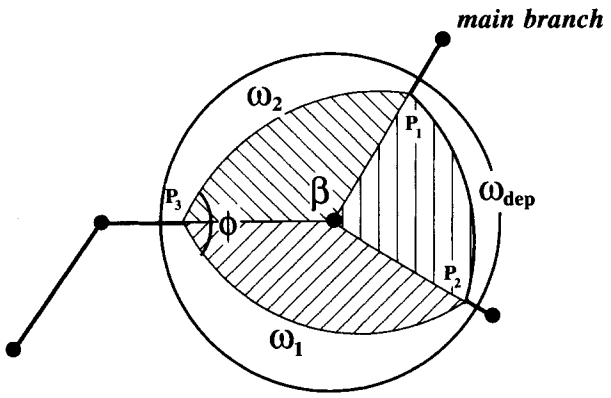

FIGURE 4. Construction needed to calculate derivatives of the bond angle and phase angle deformation energy at the branch point. The relation between two independent bond angles ω_1 and ω_2, dependent bond angle ω_{dep} and phase angle Φ may be obtained using spherical geometry, ω_1, ω_2, and ω_{dep} are sides of the spherical triangle P_1, P_2, P_3, and Φ is one of its angles.

REGULARIZATION, ANNEALING, AND HYDROGEN PLACEMENT

Suppose we have a set of Cartesian coordinates of protein atoms. These coordinates may have small errors and violations of the covalent geometry, and some of the atoms may be missing from the set (e.g., hydrogen atoms). As a first step, the ICM modeling often involves finding a conformation with the ideal covalent geometry and satisfying three requirements: small coordinate rms deviation from the initial data set, low energy, and globally optimized positions of polar hydrogens, which are addressed by regularization, annealing, and hydrogen placement procedures, respectively.

The simplest approach to regularization would be to impose all position restraints as tethers simultaneously and minimize the tethering function with respect to the free internal variables. The set of free internal variables may comprise the dihedral angles such as φ, ψ, and χ angles or any other selection of the internal variables. The fixed variables may have the standard idealized values. However, the straightforward minimization may be caught by local minima and converges extremely slowly because of the complexity of the tethering function surface in the space of internal variables.

The chain-growth procedure described in ref. 13, solves the problem.

The resulting conformation is usually highly strained and should be annealed by iterative minimization of the van der Waals, hydrogen bonding, and torsion energy in conjunction with the tethering penalty term. To prevent the disruption of a structure with strong clashes and, on the other hand, to allow some relaxation of relatively weak distortions, minimization is carried out in several steps, the weight of tethers being reduced successively. In the beginning of each minimization run, van der Waals (E_{vw}) and tethering penalty (E_τ) terms are evaluated and the weighting factor for tethers is recalculated to satisfy the relation $E_\tau = 5E_{vw}$ until the van der Waals energy becomes negative. Some further minimization is performed with a weighting factor of 1 so that the number of function evaluations in the refinement stage adds to 10,000.

The procedure described provides automatic placement of the majority of hydrogens because of preservation of covalent geometry on each step and the final minimization of full energy. Only the polar asymmetric hydrogens with free torsion angle (those in cysteine, serine, threonine, and tyrosine) cannot be placed by simple minimization because they have several rotameric states separated by barriers. In these cases, a systematic search procedure was applied. Each hydrogen was placed into two or three possible rotameric states with subsequent energy minimization and the best energy conformation was chosen. The energy terms used were the van der Waals potential, a hydrogen bonding term, an electrostatic term, and a torsion potential. This procedure is efficient because ICM automatically retains only the interactions between the given hydrogen and its static environment.

DOUBLE-ENERGY MONTE CARLO MINIMIZATION PROCEDURE

The ICM global optimization procedure consists of the following steps:

1. Random conformational change, such as
   - one angle at a time[28]
   - Brownian-like step for docking (see below)
   - biased probability random step[41] (side-chain torsion angles defining positions of nonpolar hydrogens and omegas are excluded);
2. Energy minimization of analytical differentiable energy terms (all free variables are considered);
3. Evaluation of nondifferentiable energy terms such as electrostatics and/or solvation energy;
4. Application of the Metropolis et al.[43] selection procedure to the total energy and return to step 1.

Introduction of two energies (namely, the first one calculated during local minimization and the second one evaluated after minimization) to be used in the Metropolis selection criteria is a way to incorporate complex energy terms as surface-based solvation energy and electrostatics into global optimization process.

DOCKING BY BROWNIAN MONTE CARLO MINIMIZATION

The docking problem can be considered as a particular case of the global energy optimization of a system of several interacting molecules either rigid or with some flexibility on the surface. The double-energy technique described above may be used to find the optimal conformation. A position of each molecule **M** starting with virtual atom $k = k(\mathbf{M})$ in internal coordinate representation is defined by six variables $\nu(\mathbf{M})$: φ_κ (or phase angle Φ_κ), α_κ, b_k, $\varphi_{\kappa+1}$, $\alpha_{\kappa+1}$, $\varphi_{\kappa+2}$ (Fig. 1b). In the course of the MC procedure, molecules should be subjected to certain random rotational and translational movements. Changing variables φ_κ (Φ_κ), α_κ, b_k, $\varphi_{\kappa+1}$, $\alpha_{\kappa+1}$, $\varphi_{\kappa+2}$ independently with a certain amplitude (a natural default in ICM) is not efficient. Depending on b_k as well as molecular size and shape, small changes in the virtual dihedral and planar angles may produce substantial relative displacements of the atoms, causing either atomic clashes or exceedingly large separation of the molecules. We propose random translational and rotational movements which result in atomic displacements of amplitude D_{BMC}. The algorithm is the following:

1. Define the radius of gyration R of the molecule and the coordinates of its center $\mathbf{r}_c$;
2. Find the rotation angle $\vartheta = D_{BMC} \cdot \eta_r \cdot R^{-1}$ and rotate the first three atoms of the molecule by this angle around a randomly oriented axis $\mathbf{e}_r$ passing through the center $\mathbf{r}_c$ (η_r is a random number between 0 and 1);

3. Find the translation vector $\mathbf{t} = D_{BMC} \cdot \eta_t \cdot \mathbf{e}_t$, where $\mathbf{e}_t$ is a randomly oriented unit vector and η_t is a random number between 0 and 1, and translate the first three atoms of the molecule;
4. Calculate new values of six variables φ_κ (Φ_κ), α_κ, b_k, $\varphi_{\kappa+1}$, $\alpha_{\kappa+1}$, $\varphi_{\kappa+2}$ from the Cartesian coordinates of the first three atoms and rebuild the molecule according to the new variables.

Isotropically distributed random vectors $\mathbf{e}_t$ and $\mathbf{e}_r$ may be obtained by taking three normally distributed random x-, y-, and z-components and normalizing the vector.

An alternative random rotation more suitable for irregularly shaped molecules may be based on the tensor of inertia. A rotation around the randomly isotropically chosen direction $\mathbf{e}_r$ may be found from the following relation:

$$(\vartheta \mathbf{e}_r)^T I (\vartheta \mathbf{e}_r) = D_{BMC}^2 \eta_r^2$$

where the symmetrical second-rank tensor I reads

$$I = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} y_i^2 + z_i^2 & -x_i y_i & -x_i z_i \\ -x_i y_i & x_i^2 + z_i^2 & -y_i z_i \\ -x_i z_i & -y_i z_i & x_i^2 + y_i^2 \end{pmatrix}$$

(summation is over all heavy atoms of the molecule and masses are set to 1). Thus, the rotation angle ϑ is given by

$$\vartheta = \frac{D_{BMC} \eta_r}{\sqrt{(\mathbf{e}_r^T I \mathbf{e}_r)}}$$

Results

DOCKING TWO ALPHA-HELIXES FROM A LEUCINE ZIPPER

Two α-helixes from the leucine zipper form a parallel complex with leucines, creating a dimerization interface. The structure of the dimerization region of the yeast transcriptional activator GCN_4 was determined crystallographically[29] and by NMR.[44] In an attempt to understand the specificity of intermolecular interactions and to develop a fully automated procedure to predict the structure of the complex from the structures of individual molecules, one may formulate three consecutive steps on the way to the ultimate solution of the docking problem: (1) docking rigid molecules in their conformations obtained from the X-ray structure of complex; (2) docking molecules with the backbone rigid and side chains flexible, with their starting conformation being set to the energy optimum for the separated molecules in the solution; (3) docking molecules with both the backbone and the side chains flexible. In this article we address the first two problems using full-atom representation of two molecules and double-energy MCM procedure with Brownian-type random motion to predict the dimerization geometry from a random starting position.

DOCKING RIGID HELIXES

Regularized structures of the two 31-residue long helixes were obtained by fitting a protein model with standard bonding geometry to the atomic positions from the crystallographic structure[29] as described above. All hydrogen atoms were added and set to their optimal energy positions. The root mean square deviation of the regularized structure from the original coordinates was 0.15 Å for the C^α, N, and C backbone atoms and 0.23 Å for all heavy atoms.

Starting conformations were generated by assigning random values to the six variables φ_κ (Φ_κ), α_κ, b_k, $\varphi_{\kappa+1}$, $\alpha_{\kappa+1}$, $\varphi_{\kappa+2}$ determining the position of the second helix. The generated starting conformations had the average backbone rms deviation of 25.8 ± 9.5 Å from the correct structure (as measured for the second helix with the first one exactly superimposed). A weak distance restraint was imposed between the C^α atoms of Lys 15 in both helixes. It was switched on when the C^α—C^α distance occasionally exceeded 15 Å. ECEPP/2 van der Waals energy parameters were used during minimization. Solvation energy[32] was added for the final energy evaluation at every random step. Parameter D_{BMC} defining the average amplitude of random translation and rotation (see above) was set to 2 Å. The maximal number of energy evaluations in every local minimization was set to 250. The simulation was performed at 600 K (see ref. 10 for justification) and no more than 300,000 energy evaluations were allowed in the global optimization procedure. The average simulation took 3.5–4 CPU hours on the Indigo workstation with the R4000 processor.

Figure 5 shows the dependence of the best energy achieved during the time of simulation for nine runs. Eight out of nine runs achieved energies between −330 and −333 kcal/mol. In each simu-

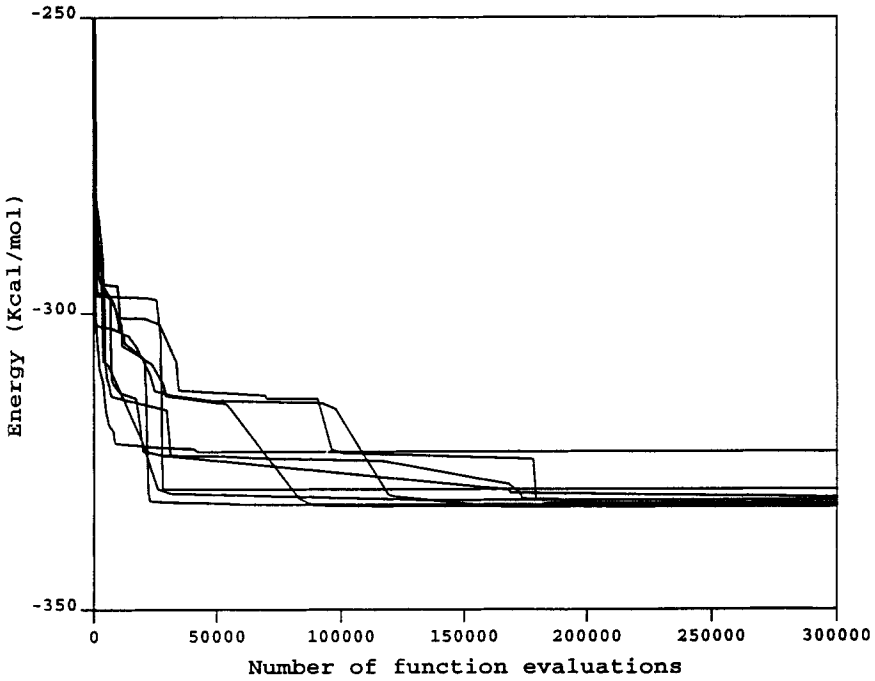

FIGURE 5. Energy profiles for nine Monte Carlo docking simulations of formation of GCN4 leucine zipper domain from two randomly placed rigid helixes. The best energy achieved for a given number of function evaluations is shown.

lation, low-energy conformations visited during the search were accumulated in a "conformational stack."[10] The stack retained conformations within 20RT = 24 kcal/mol energy range from the lowest energy. Three types of complexes were found among 48 stack conformations. Conformations having the best energy (between −333 and −330 kcal/mol depending on the run) were basically identical to the crystallographic structure, with the average all-atom rms deviation being only 0.6 Å (Fig. 6a). The second family turned out to be antiparallel complexes with higher energies from −325 to −323 kcal/mol (Fig. 6b). The third family of conformations had even higher energies ranging from −324 to −322 kcal/mol. In those cases, the helixes were again parallel but staggered by two helix turn (one period of the leucine ladder) and interacting via leucines (Fig. 6c). The rms deviation of these conformations from the crystallographic structure was about 5.0 Å. No other types of possible interaction were found within the specified energy range. Figures 6a, b, and c show how leucines are driving the dimer formation for all three types of complexes. The crystallographic parallel complex appeared to be the most energetically favorable.

DOCKING HELIXES WITH FLEXIBLE SIDE CHAINS

The successful and unambiguous prediction of the docked conformation of two rigid helixes encouraged us to address a much more realistic, though difficult, problem—namely, docking with the side chains set to conformations energetically optimal for the separate helixes, with the conformational flexibility being allowed during association.

All the side chains in the isolated helix were subjected to the 100,000 steps of biased probability Monte Carlo minimization procedure[41] at 600°K. The ECEPP/2 van der Waals, torsion, and hydrogen bonding energies, electrostatics with distance-dependent dielectric constant ($\varepsilon = 4d_{\alpha\beta}$), and the solvation energy[32] were considered. Most of the side chains changed their conformations (Fig. 7). To dock these two helixes from random start positions, we applied a double-energy Monte Carlo minimization procedure with the same energy terms as for the rigid helixes. The difference with the rigid docking procedure was that all the side chains were unfixed during minimization so that they could change their conformation to allow di-

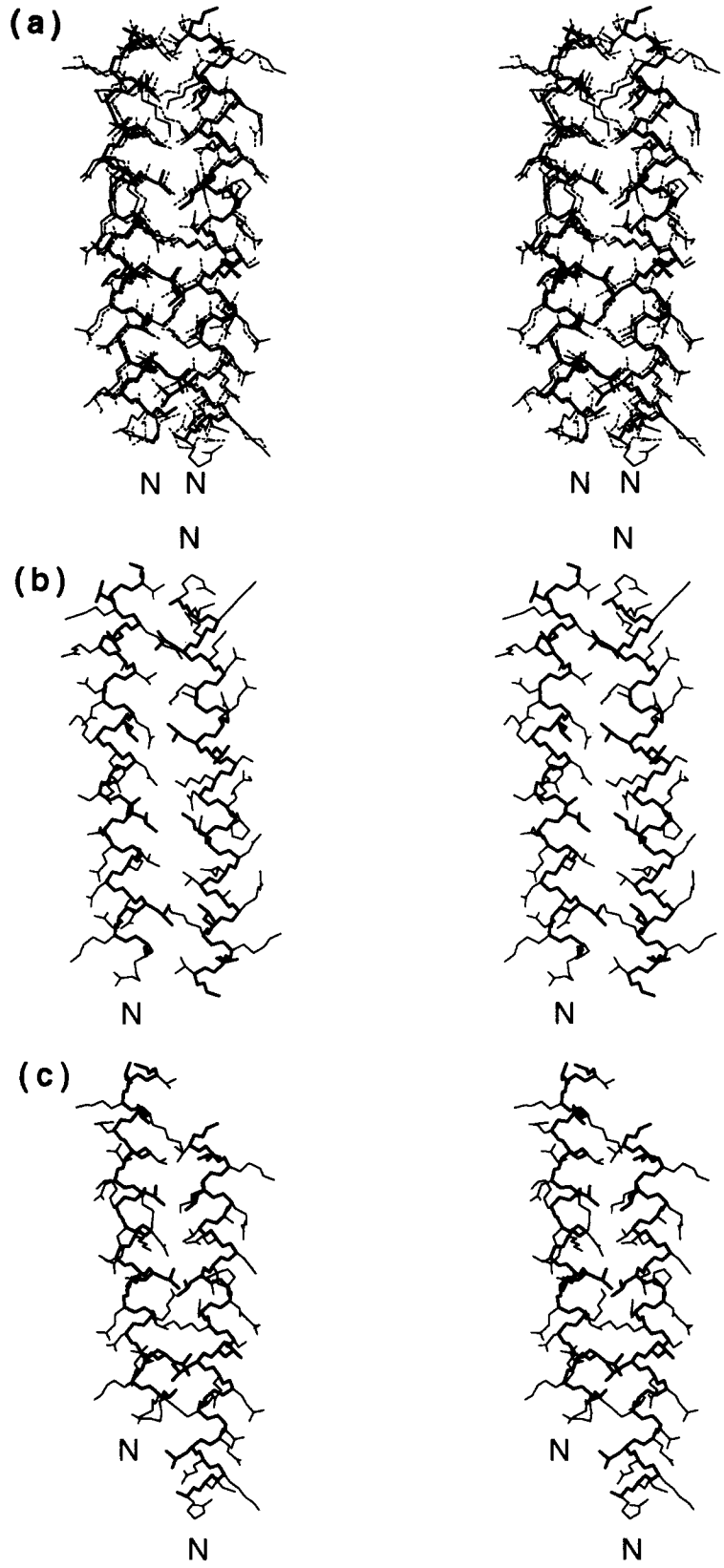

FIGURE 6. Stereoplots of three types of low-energy complexes found by the procedure, namely (a) parallel complex; (b) antiparallel helixes, and (c) helixes staggered by one helix turn. Superposition with the crystallographic structure (dashed line) is shown for the lowest-energy conformation (a). The backbone atoms and the leucines are shown with bold lines. Hydrogens are not shown.

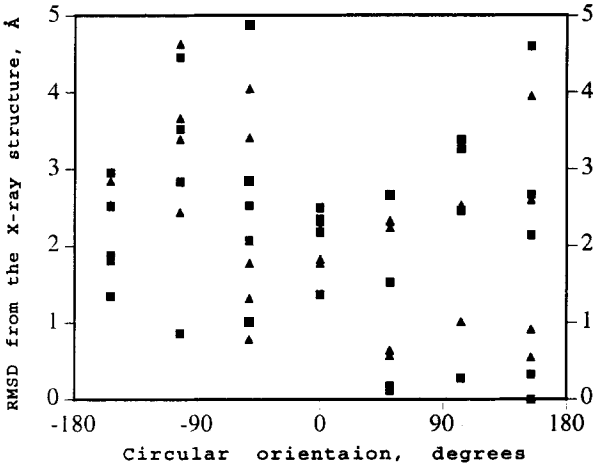

FIGURE 7. rms Deviations of the globally optimized leucine zipper side chains from their conformations in the X-ray structure plotted against the polar angle calculated as $\mathrm{mod}_{360}((360° \times 2/7)i_{res})$. The optimization was carried out for uncomplexed helixes. Triangles and squares depict deviations for the first and second helixes, respectively.

merization. The results of 10 simulations (each took on the average 12–14 CPU hours) were accumulated in a conformational stack, where only the lowest energy representatives of different conformational families were retained. Pairwise coordinate rms deviation within each family was less than 3.0 Å. The results are shown in Table III. The lowest energy conformation found by the procedure was rather close to the X-ray structure with an rms deviation of 1.18 Å for the backbone atoms and 2.34 Å for all heavy atoms (Fig. 8). The best conformation is separated by an energy gap of about 5 kcal/mol from the group of five suboptimals spanning a 12-kcal/mol range (Table III). In the second best conformation, Leu12 of the first helix interacts with Leu13 in the second, with two helixes forming a crossing angle of about 60°. The third conformation is antiparallel as in the rigid case.

POTENTIAL ENERGY HYPERSURFACE IN CARTESIAN COORDINATE SPACE AND TORSION ANGLE SPACE

Minimization and molecular dynamics in Cartesian coordinate space are popular elements of global energy optimization procedures aimed at structure prediction. To justify our usage of internal coordinates (torsion angles in particular), we carried out an investigation of the nature of potential energy hypersurface in both systems of variables. A specific aspect we focused on was the ability of the minimization procedure to restore the lowest-energy conformation after structural distortion.

A small protein, the 29-residue trypsin inhibitor from squash seeds, was used in the study. Its three-dimensional structure was solved by NMR spectroscopy.[33,45] To allow large-scale conformational deformations, we replaced cysteines involved in the S—S bridges by alanines. The model with regular covalent geometry was fitted to the coordinates of the first conformation in the 3cti.brk file[46] with a 0.2-Å rms deviation for all nonhydrogen atoms. Then the model was optimized globally in the vicinity of the crystallographic structure by 5000 steps of the biased probability Monte Carlo minimization procedure at 600°K.[41] The ECEPP/2 van der Waals, hydrogen bonding, torsion energies, and electrostatics with distance-dependent dielectric constant $\varepsilon = 4d_{\alpha\beta}$ (ref. 40) were combined with the solvation energy.[32] In spite of the extensive sampling of the conformational space around the experimental structure and disruption of the di-

TABLE III.
The Low-Energy Conformations of Two Helixes from GCN4 Leucine Zipper Found by the ICM Soft Docking Procedure.

Energy, kcal/mol	rmsd(Å)[a] from X-ray	Configuration, crossing angle, deg.	Interface residues[b] (distances ≤ 2.5 Å)
−597.3	1.18	Parallel	Leucines
−592.7	5.59	≈60°	12–13
−587.6	18.1	Antiparallel	9–27, 19–12, 29–5, 30–6
−585.3	7.05	≈56°	6–12, 9–15
−580.9	9.16	≈47°	23–22, 26–22, 30–30
−580.7	4.76	≈52°	13–16, 20–19

[a] rms Deviation was calculated for the backbone (C^α, C, N) atoms.
[b] Two numbers separated by dash are residue numbers in the first helix and second helix, respectively.

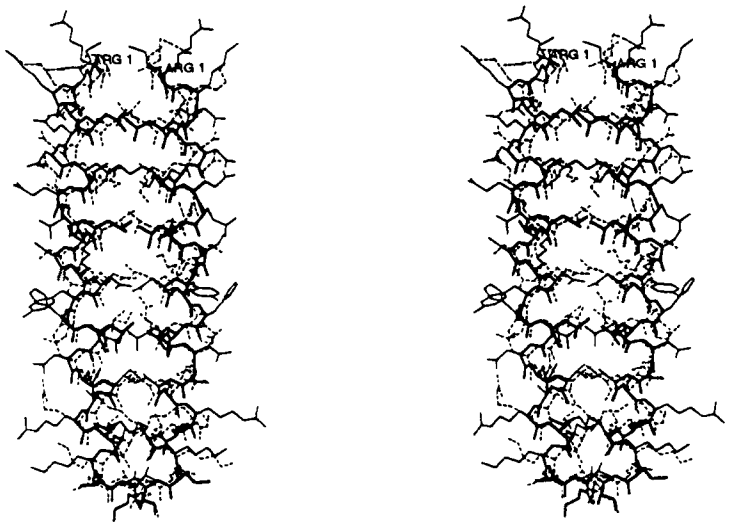

FIGURE 8. Stereoplot of the lowest-energy conformation found by the ICM docking procedure. The leucine side chains are shown by bold lines. The superimposed crystallographic structure (RMSD 1.18 Å for the backbone C^{α}, N, C atoms) is shown by the dashed line.

sulfide bonds, the best energy conformation was rather close to the starting one apart from the *N*-terminal fragment Arg1—Leu7, which changed its location. In the native structure, this fragment seems to be stabilized by the Cys3—Cys20 disulfide bridge. The rms deviation of the backbone atoms of residues 8–29 was only 0.9 Å.

ICM MINIMIZATION FROM A DISTORTED CONFORMATION

The optimized trypsin inhibitor was distorted by adding random rotation angles ξ_i from the range $(-\Xi, +\Xi)$ to all free torsion angles, namely, φ, ψ, and χ angles. At each amplitude Ξ (which was progressively increased from 1° to 43° in 2° steps), 30 randomly distorted conformations were generated. The distorted conformations were energy minimized by the Powell conjugate–gradient method.[47] Up to 1000 energy evaluations were allowed during minimization. Each minimization took about 0.1.5–2 minutes of CPU time depending on starting conformation, which influences the number of interaction lists updates. Figure 9a plots the Cartesian coordinate rms deviation of a distorted conformation before and after minimization. The starting position is always on the diagonal of the plot. Conformational change is represented by vertical movement of the dot downward or upward from the diagonal as the conformation approaches the optimal one or moves away from it, respectively. For amplitudes (Ξ) less than about 10–15° (which corresponds to about 2.0 Å rms deviation), the vast majority of the distorted conformations are brought exactly to the lowest energy conformation (rms deviation less than 0.25 Å) (Figs. 9a and 10). Larger distortions make the task increasingly more difficult, although even at a Ξ of 33° and 35° we still had cases when the optimal conformation was restored successfully.

RESTORING THE OPTIMAL CONFORMATION IN CARTESIAN COORDINATE SPACE

The XPLOR program[42] was used for minimization and dynamics runs in Cartesian coordinate space (further referred to as XYZ). The optimized conformation of the trypsin inhibitor was reoptimized according to the CHARMM potential energy functions by 1000 steps of conjugate gradient minimization followed by 2000 steps of molecular dynamics at 300°K and a final 1000 steps of minimization. We used the united atom representation with polar hydrogens (PARAM19) and distance-dependent dielectric constant $\varepsilon = 4d_{\alpha\beta}$. The XYZ-optimized conformation was distorted in two ways: (1) by randomizing its φ, ψ, and χ torsion angles as in the previous case, and (2) by randomizing Cartesian coordinates of all the atoms with rms deviation increasing from 0 to 3 Å in 0.2 Å steps. Twenty random conformations were generated at each step and then minimized by 1000 steps of the Powell conjugate–gradient method. The average minimization took about 1.1 minutes of CPU time.

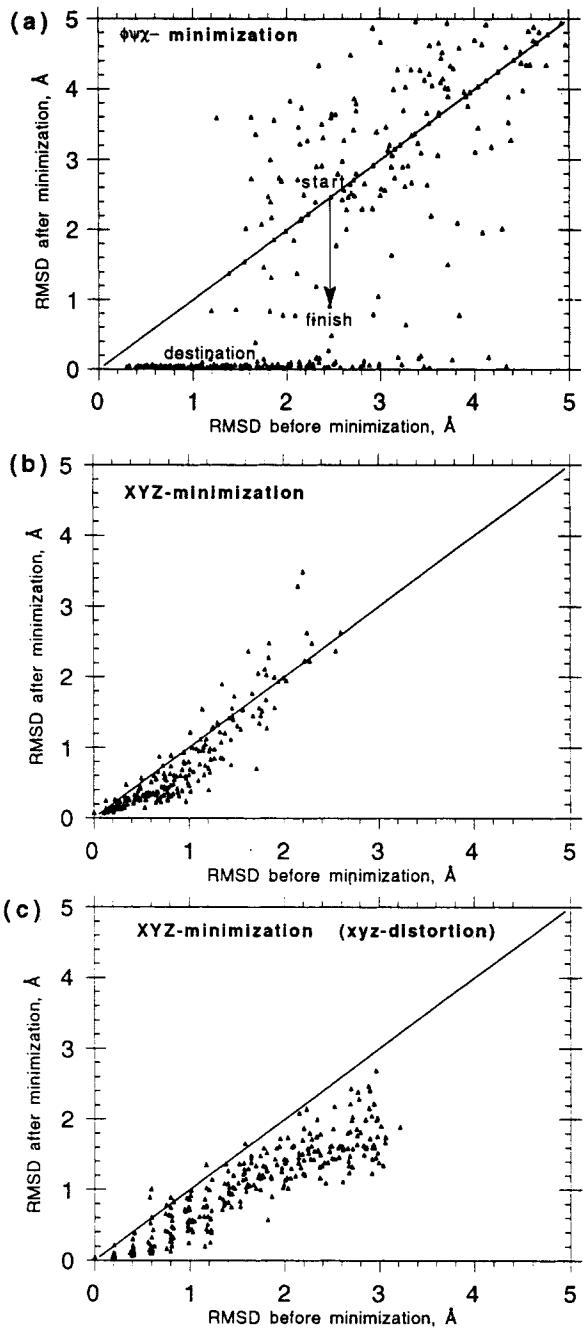

FIGURE 9. Restoration of the lowest-energy structure by energy minimization from distorted conformations. rms Deviations of atom coordinates from the optimal structure are calculated before and after minimization. (a) Torsion angles are randomly distorted and the ECEPP/2 energy is minimized in Z-space; (b) minimization is performed in Cartesian coordinate space using PARAM19 force field; (c) both distortions and minimization were carried out in the Cartesian coordinate space.

Somewhat shorter time of one run as compared to the ICM is due to the reduced number of atoms in the united atom representation used for the XYZ simulations.

The rms deviations from the XYZ-optimized conformation before and after minimization are shown in Figures 9b and c for both kinds of distortion. The most striking difference from the $\phi\psi\chi$ minimization is that even after very small distortions, the minimization does not reach the destination because it gets stuck in some local minima on the way to the initial conformation. Memory of the optimal conformations gradually disappears after 1.1 Å for the $\phi\psi\chi$ distortion (Fig. 9b). For XYZ distortions greater than 1.1 Å, one can also notice a stepwise deterioration of restoration ability, although the majority of conformations still lie under the diagonal (i.e., minimization still results in certain decrease of rms deviation from the initial structure). However, this decrease corresponds mostly to the restoration of regular covalent geometry while preserving the overall distorted backbone topology.

Discussion

The ICM method offers a number of technical improvements to the docking problem. These improvements allowed both "rigid" and "soft" docking of two 31-residue long helixes with their mo-

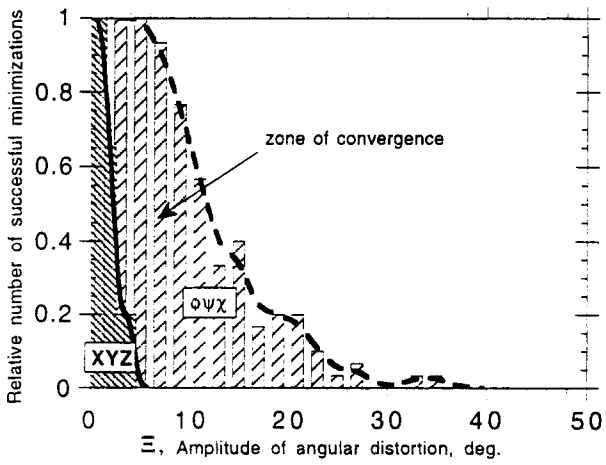

FIGURE 10. Relative number of randomly distorted conformations which were brought back to the initial lowest-energy conformation with rms deviation smaller than 0.25 Å by energy minimization in Cartesian space and in torsion angle space.

lecular structure and energy being represented accurately by a fully automated procedure and in several hours of computer time. Molecular mechanics calculations *in vacuo* without any energy terms accounting for solvation are inadequate for structure prediction of peptides and proteins. Consideration of a large number of explicit water molecules in MC or MD simulation is prohibitively expensive if a considerable portion of peptide configuration space is to be sampled. An alternative way is to use approximations taking solvation into account implicitly by expressing it as a function of the accessible molecular surface.[32] Analytical differentiation of the accessible surface-based solvation energy with respect to atom positions or dihedral angles is possible.[48] However, according to our estimates, calculation of the surface area and its derivatives still takes much more CPU time than all the other energy terms together. The electrostatic free energy which can be calculated by numerical solution of the Poisson–Boltzmann equation on the three-dimensional grid (for review, see ref. 49) or using the boundary element method[50,51] approximates more accurately the solvent polarization part of the solvation free energy. However, it misses the nonelectrostatic part of the solvation free energy, which might be referred to as hydrophobic energy. Therefore, a realistic estimate of the solvation free energy would require both surface-dependent terms and electrostatic polarization term, whatever approximation is used for the latter. In this work we used approximation of the solvation energy[32] combined with distance-dependent dielectric constant electrostatics, which seemed to be sufficient for the docking application considered.

In any case, incorporating solvation energy calculations into extensive MC simulations poses a problem because these terms are computationally heavy and nondifferentiable (at least the electrostatic polarization term). The introduction of two different sets of energy terms for the minimization and evaluation parts of the Monte Carlo minimization random step is a possible solution of the problem. Fortunately, both surface and electrostatic terms do not change much upon small conformational changes during the minimization, which mostly optimizes the sharp Lennard–Jones potential.

Another improvement in the Monte Carlo minimization method is the reduction of the conformational space to be sampled. This is achieved by introduction of two subsets of variables: one (V_{MC}) for the large-scale random changes, and the second (V_{min}) for the subsequent relaxation. V_{MC} is a subset of V_{min} and usually does not contains side-chain torsion angles defining positions of nonpolar hydrogens, the main chain ω, etc., whereas V_{min} contains all free variables. Since minimization is quite sufficient to adjust variables excluded from the V_{MC} variable set, more random steps may be performed in V_{MC} space, which is almost two times smaller than the energy minimization set V_{min}.[10]

The docking problem has physical and technical aspects. Physically, we should decide what model of atomic structure and interactions ensures the unambiguous prediction of the native conformation (i.e., should we use a simplified model with approximations of some energy terms, or alternatively, is the accuracy and availability of all principal physical components of the interaction really crucial?). The results of Shoichet and Kuntz[20] as well as a large number of false positives for simplified models (e.g., ref. 21) indicate that the approximations proposed so far are not sufficiently accurate. We used the most detailed model with all hydrogens and a full ECEPP/2 potential supplemented by the solvation energy. This gave a good separation (5–8 kcal/mol) between the true and the false positives (antiparallel and staggered). One could guess that for a simplified model it would be difficult to distinguish between parallel and antiparallel arrangements.

Surprisingly, making side chains flexible and assigning to them conformations which are different from those in the complex (Fig. 7) and correspond to their energy optimal states in the uncomplexed helixes does not change the main picture. The separation between the correctly predicted parallel arrangement and the false positive is about 5 kcal/mol. However, in contrast to the rigid docking case, adjustment of the interface residues allows wide variety of low-energy conformations. The ability of the procedure to find the correct answer even with flexible side chains is of principal importance in the more general context of the protein-folding problem. Seemingly, both inter- and intraprotein interaction energies are somewhat tolerant of the side-chain rearrangements. This tolerance can be important for a successful recognition/folding event.

Comparison of rigid and flexible geometry force fields, which could be exemplified by ECEPP/2[3,37] and AMBER[2] potentials, respectively, has a long history (e.g., discussion of Roterman et al.[52] and Kolman and Dill[53]). Here we focus rather on com-

parison of Cartesian coordinate and internal coordinate representations in terms of how the potential energy surfaces look (regardless of the specific implementation) and how they influence the global search efficiency. A Cartesian coordinate description of molecular geometry has two main disadvantages for large molecular systems: (1) hard degrees of freedom, such as bond lengths and bond angles, are not excluded so the total number of variables is 7–10 times larger than the number of torsion angles; (2) hard and soft degrees of freedom are mixed in the Cartesian coordinate. The large number of variables leads to an exponential increase of the conformational hyperspace to be searched. The second disadvantage is the reason why it is difficult to get rid of the first one.

We believe that the most promising way to approach the protein-folding problem is a two-level global energy optimization procedure. If the whole of conformational space can be subdivided into zones surrounding local minima, the two-level procedure may be represented as a combination of global jumps between zones according to a certain strategy that avoids impossible sampling of the whole space while following some low-energy pathway, and local searches aimed at finding the representative free energy value in this zone. The most efficient procedure for the latter seems to be a local minimization algorithm because these algorithms (e.g., the conjugate–gradient algorithm used in this article) were designed to achieve the optimum using a minimal number of function evaluations. Under this scheme, apart from the global strategy of jumping between minima, the key determinant of the search efficiency is the size of the zone. Our estimates show (Fig. 10) that radius of convergence is significantly larger if the covalent geometry is fixed.

The different behavior of the minimization from distorted conformations (Figs. 9 and 10) can be explained in terms of the appearance of the energy hypersurface in each representation. Let us imagine the energy hypersurface around the minimum energy state (Fig. 11a). The surface for a protein or peptide is complex and contains many extrema because of numerous atom–atom interactions. Our results (Fig. 9a) suggest that in the case of torsion angle space they are mostly energy maxima, which can be bypassed by the minimization procedure. That is the reason of rather large radius of convergence (about 15°) for the torsion coordinate minimization (Fig. 10). The restoration pattern for Cartesian coordinate minimization is quite different (Figs. 9a and b). The optimal conformation can not be restored fully even after small distortions. To interpret this in terms of the energy hypersurface and to compare it with the one shown in Figure 11a, let us project the energy hypersurface in Cartesian coordinate space to the torsion angle basis so that for each point in a projection space the covalent geometry is relaxed fully. As a result of the relaxation, many additional local minima appear around the minimum (Fig. 11b). The radius of con-

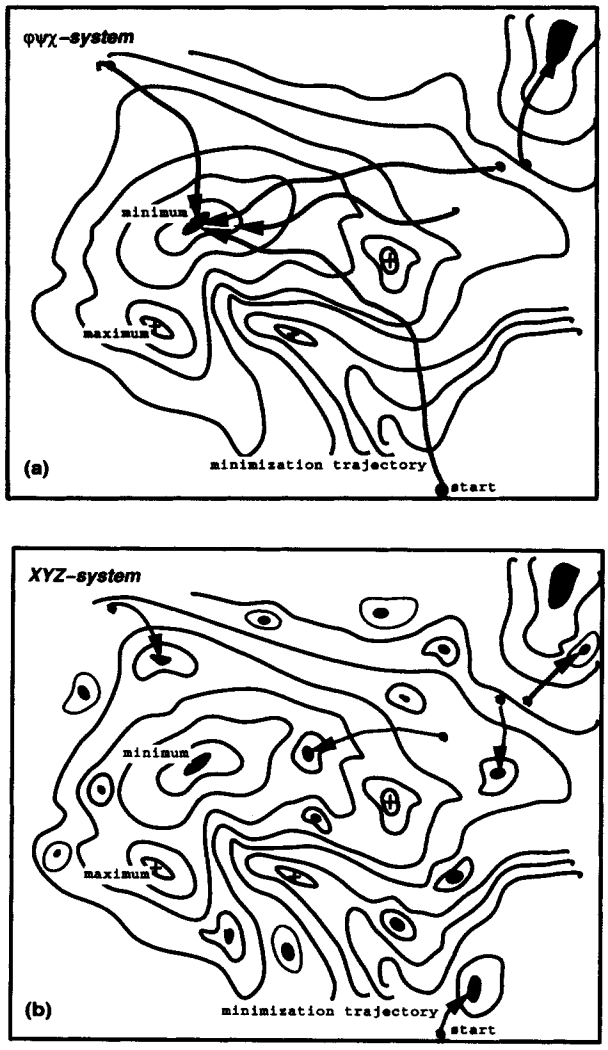

FIGURE 11. Schematic picture of the energy hypersurface for (a) torsion angle representation and (b) Cartesian coordinate approach. The surface is projected onto the same space of torsion angles. In the first case (a) the fixed covalent geometry is imposed, whereas in the second case (b) the covalent geometry is relaxed. Relaxation results in many additional minima trapping the minimization procedure.

vergence becomes smaller and the number of minima to be searched grows drastically.

Acknowledgments

The authors are grateful to Patrick Argos for his help and support. They thank Toby Gibson, Julie Thompson, Frank Eisenmenger, and Dmitry Frishman for their patience, helpful suggestions, and bug reports. They also thank Michael Nilges for challenging them to dock leucine zipper helixes, for interesting discussions, and for his instructions about the XPLOR package. Gerrit Vriend gave them a function filling a sphere with Z-ordered dots. They are grateful to David Thomas for a critical reading of the manuscript.

References

1. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comp. Chem.*, **4,** 187 (1983).
2. S. J. Weiner, P. A. Kollman, D. A. Case, U. Chandra Singth, C. Ghio, G. Alagona, S. Prefeta, Jr., and P. Weiner, *J. Amer. Chem. Soc.*, **106,** 765 (1984).
3. F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, *J. Phys. Chem.*, **79,** 2361 (1975).
4. T. Noguti and N. Go, *Biopolymers*, **24,** 527 (1985).
5. K.-C. Chou, M. Pottle, G. Nemethy, Y. Ueda, and H. A. Scheraga, *J. Mol. Biol.*, **162,** 89 (1982).
6. W. Braun, S. Yoshioki, and N. Go, *J. Phys. Soc. of Japan*, **53,** 3269 (1984).
7. A. K. Mazur and R. A. Abagyan, *J. Biomol. Struct. Dyn.*, **6,** 815 (1989).
8. R. A. Abagyan and A. K. Mazur, *J. Biomol. Struct. Dyn.*, **6,** 833 (1989).
9. A. Shrake and J. A. Rupley, *J. Mol. Biol.*, **79,** 351 (1973).
10. R. A. Abagyan and P. Argos, *J. Mol. Biol.*, **225,** 519 (1992).
11. T. J. Gibson, J. D. Thompson, and R. A. Abagyan, *Protein Engineer.*, **6,** 41 (1993).
12. T. V. Borchert, R. A. Abagyan, K. V. R. Kishan, J. Ph. Zeelen, and R. K. Wierenga, *Structure*, **1,** 205, 1993.
13. F. Eisenmenger, P. Argos, and R. A. Abagyan, *J. Mol. Biol.*, **231,** 839 (1993).
14. C. Levinthal, S. Wodak, P. Kahn, and A. Dadivanian, *Proc. Natl. Acad. Sci. USA*, **11,** 271 (1975).
15. S. J. Wodak and J. Janin, *J. Mol. Biol.*, **124,** 323 (1978).
16. M. L. Connolly, *Biopolymers*, **25,** 1229 (1986).
17. J. Warwicker, *J. Mol. Biol.*, **206,** 381 (1989).
18. D. J. Bacon and J. Moult, *J. Mol. Biol.*, **225,** 849 (1992).
19. F. Jiang and S.-H. Kim, *J. Mol. Biol.*, **219,** 79 (1991).
20. B. K. Shoichet and I. D. Kuntz, *J. Mol. Biol.*, **221,** 327 (1991).
21. P. H. Walls and M. J. E. Sternberg, *J. Mol. Biol.*, **228,** 277 (1992).
22. Yue Shi-Yi, *Protein Engineer.*, **4,** 177 (1990).
23. A. S. Goodsell and A. J. Olson, *Proteins: Struct., Funct. Gen.*, **8,** 195 (1990).
24. K.-C. Chou and L. Carlacci, *Protein Engineer.*, **4,** 661 (1991).
25. J. Cherfils, S. Duquerroy, and J. Janin, *Proteins*, **11,** 271 (1991).
26. T. N. Hart and R. J. Read, *Proteins: Struct., Funct. Gen.*, **13,** 206 (1992).
27. A. Caflisch, P. Niederer, and M. Anliker, *Proteins*, **13,** 223 (1992).
28. Z. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA*, **84,** 6611 (1987).
29. E. K. O'Shea, J. D. Klemm, P. S. Kim, and T. Alber, *Proc. Natl. Acad. Sci. U.S.A.*, **88,** 561 (1991).
30. M. Nilges and A. T. Brünger, *Protein Engineer.*, **4,** 649 (1991).
31. M. Nilges and A. T. Brünger, *Protein Engineer.*, **15,** 133 (1993).
32. L. Wesson and D. Eisenberg, *Protein Sci.*, **1,** 227 (1992).
33. T. A. Holak, D. Gondol, J. Otlewski, and T. Wilusz, *J. Mol. Biol.*, **210,** 635 (1989).
34. H. Abe, W. Braun, T. Noguti, and N. Go, *Comp. Chem.*, **8,** 239 (1984).
35. N. Go and H. A. Scheraga, *Macromolecules*, **3,** 178 (1970).
36. R. A. Abagyan and A. K. Mazur, *Comp. Chem.*, **14,** 169 (1990).
37. G. Nemethy, M. S. Pottle, and H. A. Scheraga, *J. Phys. Chem.*, **87,** 1883 (1983).
38. D. Eisenberg and A. D. McLachlan, *Nature*, **316,** 199 (1986).
39. J. A. McCammon, P. G. Wolynes, and M. Karplus, *Biochemistry*, **18,** 927 (1979).
40. R. W. Pickersgill, *Protein Engineer.*, **2,** 247 (1988).
41. R. A. Abagyan and M. M. Totrov, *J. Mol. Biol.*, **235,** 983, 1994.
42. A. T. Brünger, X-PLOR software manual, version 3.0 New Haven, CT, Yale University, 1992.
43. N. A. Metropolis, A. W. Rosenbluth, N. M. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.*, **21,** 1087 (1953).
44. V. Saudek, A. Pastore, M. A. Castiglione Morelli, R. Frank, H. Gausepohl, T. Gibson, F. Weih, and P. Roesch, *Protein Engineer.*, **4,** 3 (1990).
45. M. Nilges, J. Habazettl, T. A. Holak, and A. T. Brünger, *J. Mol. Biol.*, **219,** 499 (1991).
46. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, **112,** 535 (1977).
47. M. J. D. Powell, *Math. Programming*, **12,** 241 (1977).
48. G. Perrot, B. Cheng, K. D. Gibson, J. Vila, K. A. Palmer, A. Nayeem, B. Maigret, and H. A. Scheraga, *J. Comp. Chem.*, **13,** 1 (1992).
49. M. E. Davis and J. A. McCammon, *Chem. Rev.*, **90,** 509 (1990).
50. R. J. Zauhar and R. S. Morgan, *J. Mol. Biol.*, **186,** 815 (1985).
51. A. H. Juffer, E. F. F. Botta, B. A. M. van Keulen, A. van der Ploeg, and H. J. C. Berendsen, *J. Comp. Phys.*, **97,** 144 (1991).
52. I. K. Roterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga, *J. Biomol. Struct. Dyn.*, **7,** 421 (1989).
53. P. A. Kollman and K. A. Dill, *J. Biomol. Struct. Dyn.*, **8,** 1103 (1991).