J. Mol. Biol. (1994) **235**, 983–1002

Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins

Ruben Abagyan and Maxim Totrov

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

Two major components are required for a successful prediction of the three-dimensional structure of peptides and proteins: an efficient global optimization procedure which is capable of finding an appropriate local minimum for the strongly anisotropic function of hundreds of variables, and a set of free energy components for a protein molecule in solution which are computationally inexpensive enough to be used in the search procedure, yet sufficiently accurate to ensure the uniqueness of the native conformation. We here found an efficient way to make a random step in a Monte Carlo procedure given knowledge of the energy or statistical properties of conformational subspaces (e.g. ϕ–ψ zones or side-chain torsion angles). This biased probability Monte Carlo (BPMC) procedure randomly selects the subspace first, then makes a step to a new random position independent of the previous position, but according to the predefined continuous probability distribution. The random step is followed by a local minimization in torsion angle space. The positions, sizes and preferences for high-probability zones on ϕ–ψ maps and χ-angle maps were calculated for different residue types from the representative set of 191 and 161 protein 3D-structures, respectively. A fast and precise method to evaluate the electrostatic energy of a protein in solution is developed and combined with the BPMC procedure. The method is based on the modified spherical image charge approximation, efficiently projected onto a molecule of arbitrary shape. Comparison with the finite-difference solutions of the Poisson–Boltzmann equation shows high accuracy for our approach. The BPMC procedure is applied successfully to the structure prediction of 12- and 16-residue synthetic peptides and the determination of protein structure from NMR data, with the immunoglobulin binding domain of streptococcal protein G as an example. The BPMC runs display much better convergence properties than the non-biased simulations. The advantage of a true global optimization procedure for NMR structure determination is its ability to cope with local minima originating from data errors and ambiguities in NMR data.

Keywords: Monte Carlo; conformational search; global energy minimization; NMR structure determination; protein folding

1. Introduction

An efficient global optimization procedure and appropriate energy terms are the most important, yet still problematic, components of any procedure aimed at structure prediction of proteins and peptides. Many approaches can be used to sample the conformational space: molecular dynamics in Cartesian space (Bruccoleri & Karplus, 1990; van Gunsteren & Berendsen, 1990; Brünger *et al.*, 1986) or in torsion angle space (Mazur *et al.*, 1991), systematic search (Leach, 1991; Schaumann *et al.*, 1990), build-up procedures (Vásquez & Scheraga, 1985; Braun & Go, 1985; Vajda & Delisi, 1990; Simon *et al.*, 1991) and the Monte Carlo methods including simulated annealing (Kirkpatrick *et al.*, 1983; Kawai *et al.*, 1989; Wilson & Cui, 1990). The Monte Carlo methods can be subdivided into local step and non-local step procedures, the former tending to make a random step in the vicinity of a current local minimum and the latter trying to jump to a different minimum (in general, not even to the neighbouring one) at each step. Rather sophisticated local step methods have been developed (Noguti & Go, 1985; Vanderbilt & Louie, 1984; Shin & Jhon, 1991). They find the appropriate search directions (related to the covariance matrix) in an attempt to make a step along low-energy valleys. However, these methods have limitations in their global sampling capacity because they rely

0022–2836/94/030983–20 $08.00/0

upon local harmonic approximation of the energy surface which is valid only in the near vicinity of the conformation. This feature makes them adequate for sampling of the local environment of a certain conformation, rather than for large-scale searches.

In the alternative approach with non-local random steps, the main problem is how to make the step, so that both the fraction of accepted random moves (so-called acceptance ratio) and the performance are sufficient. High-dimensionality of protein systems clearly calls for much more efficient sampling algorithms than the existing ones. It has been established that a full local minimization after each random step greatly improves the efficiency of the procedure. As far as the random step itself is concerned, it was concluded that changing one angle at a time with 180° amplitude is better than changing several angles or reducing the amplitude (Li & Scheraga, 1987; Abagyan & Argos, 1992).

To increase the samplimg efficiency of the Monte Carlo procedure, we propose another type of random step such that (1) knowledge about statistical and/or energy properties of some small fragments is used in an optimal way, (2) the search still continuously covers the conformational space, so that any conformation may be achieved, and (3) the acceptance ratio is sufficiently high. The method is based on a theorem establishing the optimal probability distribution function used to generate a random step in the conformational subspace. In this paper we apply this general principle to the most obvious conformational zones; namely, ϕ–ψ zones or side-chain torsion angle zones of individual residues, although conformational clusters for more extended fragments can be used in the same way. The shape and size of the statistical probability distributions for both the side-chain conformations (Ponder & Richards, 1987) and the main-chain torsion angles were calculated from a representative set of protein domains (Heringa *et al.*, 1992).

The structure of a 12-residue synthetic peptide (Hill *et al.*, 1990) was predicted by the biased probability Monte Carlo (BPMC†) simulation from the extended state, in contrast to the non-biased runs which failed to find the minimum. Equivalent α-helical conformations were found as the global minimum for two sets of energy terms: the first, the ECEPP energy supplemented with accessibility-based solvation energy (Wesson & Eisenberg, 1992); and the second with the solvation energy represented by the electrostatic reaction field energy as well as hydrophobicity and side-chain entropic terms evaluated *via* accessibilities of reference atoms. The structure of another 16-residue peptide designed and characterized by Scholtz *et al.* (1991) was also successfully predicted by the BPMC procedure.

† Abbreviations used: BPMC, biased probability Monte Carlo procedure; ECEPP, empirical conformational energy program for peptides; MC, Monte Carlo procedure; MD, molecular dynamics; MIMEL, modified image electrostatic approximation.

The BPMC procedure was also applied for determination of the three-dimensional structure of a domain of streptococcal protein G using real NMR data (Gronenborn *et al.*, 1991; Brünger, 1992). In this case, in contrast to the variable target function approach (Braun & Go, 1985), all distance restraints were imposed simultaneously and the experimental variable restraints were used as zones of biased probability. The advantages of the BPMC method over variable target function minimization are its ability to bypass local minima and higher tolerance to ambiguous or erroneous bits of experimental data. The method operates in torsion angle space and allows consideration of multimeric structures (Abagyan *et al.*, 1994).

The electrostatic free energy of a protein in solution is an important energy term that is often missing or inadequately presented in MD or MC calculations. Calculations including water molecules explicitly are too computationally demanding and have problems with a proper treatment of electron polarizability (Gilson & Honig, 1991). There are many implicit methods ranging from simple and markedly inaccurate Coulomb or distance dependent dielectric constant approximations (McCammon *et al.*, 1979) (in existing molecular dynamics/minimization methods), to sophisticated, albeit computationally intensive, algorithms finding the solution of the Poisson equation numerically (for a review, see Davis & McCammon, 1990). The methods of the first group are fast, but all of them, even with an intricate distance dependence of the dielectric constant (Mehler & Eichele, 1984), lack an important contribution; namely, self-energy of the charge.

We have developed a method which may be called modified image electrostatic approximation (MIMEL). It is sufficiently accurate and fast to be incorporated in the MC simulation. We found an analytical correction term to the image approximation of the reaction field energy in a sphere. In the spherical case the energy can be expressed *via* intercharge distances and the charge depths (Imoto, 1984). The same formulae may be applied to arbitrarily shaped proteins. Our algorithm takes advantage of the fast surface calculation algorithm (Abagyan *et al.*, 1994) and weak dependence of the potential for deeply buried charges (Friedman, 1975). Comparison of the MIMEL reaction field energies evaluated for a set of model objects as well as for several proteins, with those calculated by the DelPhi program (Gilson & Honig, 1987; Nicholls & Honig, 1991) solving the Poisson equation numerically, shows high accuracy of the MIMEL method. MIMEL energies are used in the BPMC structure prediction of the synthetic 12- and 16-residue peptides.

2. Materials and Methods

(a) *Biased probability Monte Carlo procedure*

The global optimization method employed in this work consists of the following basic procedures, repeated itera-

tively: (1) random conformational change; (2) local minimization of the ECEPP/2 energy function (Momany *et al.*, 1975; Nemethy *et al.*, 1983) using analytical derivatives; (3) evaluation of additional energy terms weakly dependent on the local conformational adjustments (e.g. solvation energy, electrostatic free energy and entropy); (4) acceptance decision based on the total energy (criterion of Metropolis *et al.*, 1953). The way step (1) is effected, drastically influences the search efficiency (Abagyan & Argos, 1992). We attempted to design a more rational random change.

The idea of the biased probability Monte Carlo procedure is to sample with larger probability those regions of the conformational space which we know *a priori* are, on the average, highly populated and to sample with less probability regions known to be less populated. It is hard to get this statistical or energy information for all combinations of variables. However, local probability distributions of a small number of variables (e.g. ϕ–ψ distributions, χ-distributions, backbone conformations for 2, 3, 4 residue fragments, etc.) are either known or can be evaluated. To be mathematically strict in justifying our way of making a random step, let us consider the following problem.

Imagine we have a function of N_θ variables which are divided into N_v subsets of $n_v = N_\theta/N_v$ variables. For example, if each subset is a pair of the backbone torsion angles (ϕ, ψ), then N_v is a number of residues and n_v is 2. Let us consider a step of the simplified MC prediction procedure consisting of a random selection of the subset v (a residue in the above example) and a random choice of n_v new values of variables in this subset. The choice is effected according to some probability function $f(\mathbf{x})$, where $\mathbf{x}$ is an n_v-dimensional vector $\theta_1^v, \ldots, \theta_n^v$ ($f(\phi, \psi)$ in the example). Suppose that the final average distribution $\rho(\mathbf{x})$ of n_v angles in all N_v subsets is known (e.g. from statistics or energy precalculations of the corresponding short fragments). The question is what is the optimal random step distribution function $f(\mathbf{x})$? In Appendix I it is proven that the $f(\mathbf{x})$ which maximizes the probability of correct prediction is equal to $\rho(\mathbf{x})$. This result can be easily generalized to the situation where all variables are divided into several types of subsets with different numbers of variables and different statistical distributions.

In our example all ϕ–ψ angles are divided into N_{residues} pairs and a random move is made by selection of a residue and a change of both angles by some values. Correctly predicted residues are somehow recognized and kept. The theorem says that, in the fastest way to predict all the correct ϕ–ψ angles, the optimal change for each ϕ–ψ pair should be made according to the expected average distribution of ϕ–ψ angles.

The implication of this theorem is clear, random steps chosen according to the expected probability distribution should be preferred over any other way to make a random move. Another question is how to divide all variables into subsets of associated variables? Firstly, variables in a subset should be correlated. Secondly, using subsets with larger number of variables would increase the efficiency. For example, considering combined distributions of ϕ–ψ–χ variables of different residues would improve the search compared to separate zones for the backbone torsion angles and the side-chain angles. However, for large n_v values it is more difficult to collect enough statistics (if statistical distributions are used) and also to describe the distribution. Two ways for the description could be proposed. One is the grid description of the distribution combined with some interpolation rules. The second is identifying high probability zones and approximating the distribution around the zone by a bell-shaped analytical function with positions, sizes and the heights of the bells as parameters.

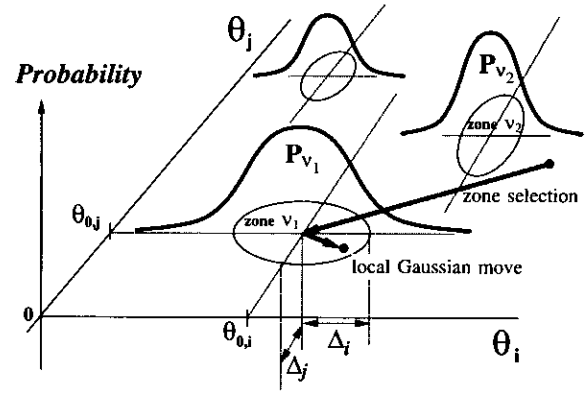


Figure 1. A random move in the biased probability Monte Carlo procedure. After a subspace is chosen (in the example shown, it consists of 2 variables, θ_i and θ_j), a preferred zone is selected and a normally distributed move in the vicinity of the zone is performed. Parameters of the probability distributions are taken from the statistical analysis of the known 3-dimensional structures.

In the current implementation of the BPMC procedure, we describe the probability distribution by a set of Gaussian distributions (Fig. 1). The random move is effected by the following procedures: (1) randomly select an internal variable (normally a torsion angle); (2) identify all high-probability zones $\nu_1, \nu_2, \ldots, \nu_\kappa$, associated with the variable (e.g. all side-chain rotamers for a particular residue type if a χ-angle is picked, or all ϕ–ψ zones if a backbone torsion is picked); (3) select one zone ν_κ according to the probability P_{ν_κ}; (4) make a normally distributed step in the vicinity of ν_κth zone, i.e. the displacement from the centre of the zone by a random vector having components distributed with the probability density ρ (Fig. 1):

$$\rho(\theta_i) = \prod_{\substack{\text{all } \theta_i \text{ of} \\ \text{zone } \nu_\kappa}} \frac{1}{\sqrt{2\pi}\Delta_i} e^{-\frac{(\theta_i - \theta_{0,i})^2}{2\Delta_i^2}}. \qquad (1)$$

(b) *Statistical analysis of local conformational preferences*

The local conformational preferences are represented by multidimensional ellipsoidal zones in subspaces of associated internal variables. To evaluate the positions, sizes and probabilities of preferred zones in ϕ–ψ and χ subspaces, we carried out statistical analysis for a representative set of known protein structures. Both ϕ–ψ maps and χ-maps were divided into regions (Fig. 2). The division of maps into regions was done either from visual inspection or according to a maxima of the torsion potential (e.g. χ_1 of Lys with 3-fold torsion potential had 3 regions ($-120°$, $0°$), ($0°$, $120°$) and ($120°$, $-120°$)). The division is somewhat arbitrary; however, it is not critical since we use the continuous distribution rather than fixed rotamers. Each region corresponds to a preferred zone, which was approximated by an ellipse with the centre

$$\theta_{0,i} = \frac{1}{n}\sum_{p=1}^{n} \theta_i^p, \text{ half-axis } \Delta_i = \sqrt{\frac{1}{n}\sum_{p=1}^{n} (\theta_i^p - \theta_{0,i})^2}$$

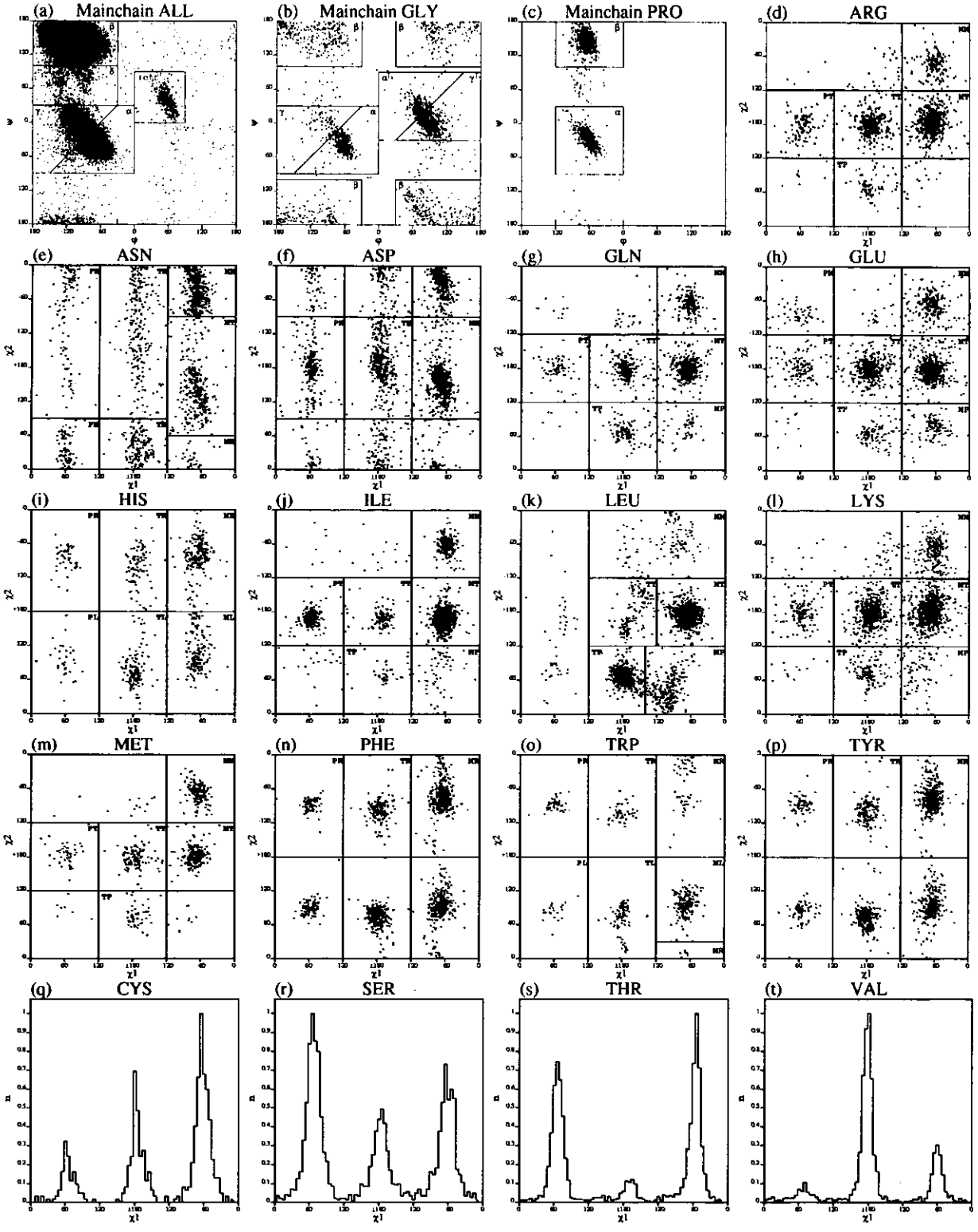


Figure 2. Torsion angle distributions and the boundaries of 163 zones (Table RSTY) used for variable restraints, for biasing random steps in the BPMC procedure and for the evaluation of side-chain entropies. (a) The distribution of the ϕ–ψ torsion angles in all residues but Gly and Pro. Gly and Pro distributions are shown separately in (b) and (c), respectively. The main-chain statistics come from 191 protein domains with sequence identity less than 30% and resolution better than 2·4 Å. (d) to (p) $\chi1$–$\chi2$ distributions for 13 residue types collected from 166 protein domains (identity $<50\%$, resolution $\leq 2{\cdot}0$ Å). Zones containing less than 1% of the points have no label and were ignored. Identically marked pairs of rectangles on the Asn and Trp maps should be connected at $\chi2 = 0°$. $\chi2$ angles of Tyr, Phe, Asp having periodicity of 180° are reduced to (0°, 180°) range ((−90°, 90°) for Asp), so that there are only 3 labelled zones. (q) to (t) $\chi1$ distributions, n is relative number of cases. Three zones are M ($-60\pm60°$), P ($60\pm60°$) and T ($180°\pm60°$).

and probability n/N, where i is a variable contributing to the zone, p is an index of a point, n is a number of points in the region and N is a total number of points.

For the ϕ–ψ statistics, we selected 191 protein chains which is the largest possible subset of protein chains with the pairwise sequence identities less than 35%, resolution less than 2·4 Å and sequence length greater than 30 residues. The list was calculated by an algorithm of Heringa *et al.* (1992). The protein codes (Bernstein *et al.*, 1977) and chain specifiers -X (when necessary) were the following: ACT1-A, ACT1-D, ROP2-B, 2SOD-Y, ZIF1-C, 1AAP-B, 1ACX, 1ALC, 1ALD, 1BBP-D, 1CHO-I, 1CLA, 1COX, 1CRN, 3CRO-A, 1CSC, 1CTF, 1DTX, 1ECD, 1FDX, 1FNR, 1FX1, 1FXI-C, 1GCR, 1GD1-P, 1GOX, 1GP1-B, 1HDS-B, 1HIP, 1HNE-E, 1HOE, 1HRH-A, 1I1B, 1IFB, 1L67, 1LH2, 1MBO, 1MSB-A, 1MVP-A, 1PHH, 1PPT, 1PRC-C, 1PRC-L, 1PRC-M, 1PRC-H, 1R69, 1RBP, 1RDG, 1SGC, 1SGT, 1SN3, 1TAB-I, 1UBQ, 1YCC, 256B-A, 2AZA-B, 2CAB, 2CCY-A, 2CD4, 2CDV, 2CYP, 2ER6-E, 2FB4-L, 2FBJ-H, 2FCR, 2FGF, 2FXB, 2GBP, 2GN5, 2LBP, 2LHB, 2LTN-A, 2MBA, 2MHR, 2PAB-B, 2PAL, 2PAZ, 2PCY, 2PKA-A, 2PKA-B, 2RNT, 2SDH-B, 2SNI-I, 2SOD-B, 2TRX-A, 2TS1, 2UTG-B, 2YHX, 351C, 3ADK, 3B5C, 3BCL, 3BLM, 3C2C, 3CNA, 3DFR, 3FAB-H, 3FXN, 3PEP, 3RN3, 3RP2-A, 3TEC-E, 3TEC-I, 3WRP, 4FD1, 4GR1, 4HHB-C, 4INS-B, 4P2P, 4PFK, 4SGB-I, 5ACN, 5APR-E, 5RUB-B, 5TLN, 5TNC, 6CPA, 6GCH, 6LDH, 6XIA, 7ICD, 8ADH, 8CPP, 8DFR, 9ABP, 1BBH-B, 1BBQ-B, 1BIA, 1BOV-D, 1CGJ-I, 1COL-A, 1CWG-A, 1DFN-B, 1DRB-B, 1END, 1F3B, 1FKB, 1FXD, 1GKY, 1GLY, 1GMF-B, 1GST-B, 1HSA-A, 1HSA-B, 1IPD, 1LFG, 1LMB-A, 1LPE, 1LTS-A, 1LTS-C, 1LTT-F, 1NPX, 1NSB-A, 1PE6, 1PGX, 1PII, 1RNB, 1TGI, 1TPK-C, 1TRB, 2CTX, 2LPR-A, 2MAD-L, 2MCM-A, 2PF1, 2POR, 2REB, 2RN2, 2SAR-B, 2SCP-B, 2SNM, 2TPR-A, 2ZTA-B, 3CHY, 3MT2, 3PRK-E, 4FIS-A, 4ICB, 4PHV-B, 5SIC-I, 6EBX-A, 6FAB-L, 6TIM-B, 7AAT-A, 8EST-E, 8GPB, 1NRD, 1OVA-A, 2LHM, 3ENL, 1AKE-B.

The ϕ–ψ maps were divided into 5 regions for a glycine (Fig. 2(b)), 3 regions for a proline (Fig. 2(c)), and 5 regions for all other residue types (Fig. 2(a)). The low-populated areas (referred to as "others") which are not explicitly represented by any zone, may be still achieved by the BPMC procedure from the neighbouring zones because of the tails of the Gaussian probability distribution of the random step. The average backbone dihedral angles and their root-mean-square deviations are presented in Table 1.

For side-chain statistics another set of 161 protein chains having higher resolution ($\leq$2·0 Å) and higher sequence identity threshold (50%) was constructed with the same algorithm (Heringa *et al.*, 1992). The resolution requirement was enforced to ensure sufficient accuracy of the side-chain torsion angles (backbone torsion angles have comparable accuracy at lower resolution), whereas the identity threshold was raised to increase the number of proteins in the data set, taking advantage of greater diversity of the side-chain angles as compared to the main-chain angles in the distantly related proteins. The protein codes and chain specifiers were the following: ACT1-A, ACT1-D, ROP1-A, SOD0-R, TIM2, ZIF1-C, 1AAP-A, 1ACX, 1AK3-A, 1ALC, 1ALD, 1BBP-B, 1COX, 1CRN, 1CSE-I, 1CTF, 1ECN, 1ER8-E, 1FDX, 1GCR, 1GD1-R, 1GOX, 1GP1-B, 1HIP, 1HMO-D, 1HNE-E, 1HOE, 1IFB, 1L06, 1OMD, 1PAZ, 1PPD, 1PPT, 1R69, 1RBP, 1RNS, 1SAR-B, 1SGC, 1SGT, 1SIC-I, 1SN3, 1TGL, 1TGS-I, 1THB-C, 1TMN-E, 1UBQ, 1UTG, 1YCC, 1YPI-B, 256B-B, 2ACT, 2ALP, 2APR, 2AZA-B, 2BLM-A, 2CCY-B, 2CDV, 2CI2-I, 2CNA, 2CSC, 2CYP, 2FBJ-L, 2FBJ-H, 2FCR, 2GBP, 2HHB-B, 2LH7, 2LHB, 2LTN-B, 2LTN-C, 2MCG-2, 2MHR, 2OVO, 2PAB-A, 2PCY, 2PKA-A, 2PKA-Y, 2PRK, 2RSP-A, 2SOD-Y, 2TEC-E, 2TRX-B, 2TSC-B, 2WRP-R, 31BI, 351C, 3B5C, 3BCL, 3BLM, 3C2C, 3CBH, 3CLA, 3DFR, 3FAB-H, 3FGF, 3GRS, 3RP2-B, 4BP2, 4CPV, 4ENL, 4FD1, 4FXN, 4INS-D, 4LYZ, 4MBA, 4MBN, 4PEP, 4PTI, 4PTP, 5EBX, 5HVP-A, 5P21, 5RUB-B, 5RXN, 5TNC, 6CHA-A, 6CPA, 6CPP, 6LDH, 6RXN, 7XIA, 8DFR, 1AKE-A, 1APT, 1BBH-B, 1C53, 1CWG-A, 1DFN-B, 1DRB-B, 1END, 1FIA-A, 1FKB, 1FXD, 1GKY, 1GPB, 1IMM, 1LMB-B, 1LTE, 1LTS-H, 1LTS-A, 1LTS-C, 1MEE-A, 1OVA-D, 1PGX, 1PII, 1PK4, 1RNB, 1TRB, 2CBC, 2FX2, 2MCM-A, 2POR, 2RN2, 2SCP-A, 2SNM, 2ZTA-A, 3CHY, 3MT2, 4ICB, 5ABP, 5EST-E, 5PAL, 6FAB-H, 6RNT, 7AAT-A, 7ACN.

Division of χ-maps into rotamer regions was done according to the observed statistical distributions and in most cases was in accordance with the rotamers of Ponder & Richards (1987) from a smaller dataset of 19 proteins. Figure 2 shows the zones and Table 2 the appropriate parameters.

(c) *Side-chain entropies*

The statistical distribution of side-chain conformations may also be used to calculate the side-chain entropy changes upon folding. The underlying assumption is that the observed probability distribution is close to that of the unfolded state, whereas in the folded state the conformation of the buried side-chain is confined to 1 zone only and hence its entropy is assumed to be 0. Considering every preferred χ-zone listed in Table 2 as 1 state with probability P_v, one can evaluate the side-chain entropy S_{stat} using the Boltzmann formula:

$$S_{stat} = -R \sum_v P_v \ln (P_v), \quad (2)$$

where the summation is over all preferred conformational zones of a particular residue type, and R is the gas constant. Since some of the χ-angles were ignored in the rotamer list, the contributions from these angles should be added. The additional entropy can be estimated as:

$$S_{add} = R \ln (N_{add}), \quad (3)$$

where N_{add} is the additional number of states for the angles missing, assuming that all states have equal probabilities. We took the following additional N_{add} values: 9 for χ3 and χ4 of Lys and Arg, 2·5 for χ2 of Cys, Thr and Ser, 2 for χ6 of Tyr, 3 for χ3 of Glu and Met, and 6 for χ3 of Gln. In Lys, Arg, Met, Cys, Ser and Thr cases the additional number of states chosen results from the energy barriers, for the χ3 of Gln and Glu the number originates from the assumed 60° fluctuation range in the buried state, *versus* 360° (180° for Glu because of the symmetry) in the unrestricted accessible state. The numbers of χ2-states for Cys, Thr and Ser were taken as 2·5 because, depending on the χ1 and ψ angles of the same residue, the number of accessible states is either 3 or 2. Table 3 shows the entropic contributions to the free energy difference between the exposed and buried states.

The entropic effects can be incorporated into energy calculations by relating the entropy with accessible surface. The solvation energy surface densities can be modified to make them represent the entropic term, using the observation that the accessibility of some reference atoms at the tip of a side-chain may reflect the number of reachable states for the side-chain. Division of the

Table 1

Probabilities, average positions and sizes of the preferred zones for main-chain torsion angles

Residue (total number) Zone	P†	ϕ‡ (°)	Δ_ϕ§ (°)	ψ‡ (°)	Δ_ψ§ (°)
Alanine (2232)					
α	0·54	−63·2	(9·6)	−38·5	(10·2)
β	0·31	−107·8	(35·6)	144·4	(16·6)
γ	0·08	−92·6	(18·1)	−5·1	(14·0)
δ	0·03	−108·7	(31·7)	72·1	(19·7)
Left	0·02	54·1	(15·7)	43·9	(19·7)
Others	0·03				
Arginine (974)					
α	0·50	−64·0	(9·1)	−40·4	(10·6)
β	0·33	−111·4	(29·1)	142·2	(17·9)
γ	0·10	−98·7	(17·8)	−5·8	(16·1)
δ	0·04	−119·0	(25·5)	72·0	(21·0)
Left	0·02	61·3	(8·7)	35·3	(16·7)
Others	0·01				
Asparagine (1027)					
β	0·29	−108·6	(30·1)	140·0	(24·4)
α	0·29	−65·0	(11·1)	−38·6	(13·0)
γ	0·19	−100·4	(19·3)	3·2	(16·4)
δ	0·11	−112·8	(25·9)	71·2	(19·6)
Left	0·11	55·7	(11·0)	40·1	(15·5)
Others	0·03				
Aspartic acid (1448)					
α	0·38	−65·6	(11·6)	−38·5	(12·2)
β	0·32	−98·4	(31·4)	138·1	(24·4)
γ	0·16	−98·6	(18·0)	−1·1	(16·7)
δ	0·07	−105·2	(27·4)	74·2	(21·6)
Left	0·05	56·0	(11·7)	42·6	(17·8)
Others	0·03				
Cysteine (319)					
β	0·49	−108·4	(32·3)	138·7	(19·0)
α	0·31	−63·2	(11·3)	−38·3	(10·5)
γ	0·11	−99·5	(20·5)	−8·2	(20·5)
δ	0·05	−122·1	(23·8)	80·8	(17·1)
Left	0·02	60·2	(10·3)	37·7	(19·7)
Others	0·01				
Glutamine (806)					
α	0·48	−64·2	(9·4)	−38·7	(9·6)
β	0·35	−105.5	(29·9)	140·2	(17·9)
γ	0·10	−98·6	(17·6)	−5·8	(17·1)
Left	0·03	57·6	(12·0)	38·7	(20·1)
δ	0·02	−113.2	(28·6)	75·2	(13·9)
Others	0·02				
Glutamic acid (1425)					
α	0·55	−64·7	(9·5)	−38·7	(10·4)
β	0·29	−105·5	(29·0)	137·7	(17·1)
γ	0·11	−96·5	(18·6)	−7·6	(15·2)
δ	0·02	−106·7	(25·4)	71·7	(18·9)
Left	0·02	60·1	(10·6)	37·3	(21·8)
Others	0·01				
Glycine (1944)					
β	0·41	−184·1	(77·9)	178·1	(30·3)
γ^L	0·22	92·5	(14·7)	0·2	(13·9)
α	0·17	−62·8	(10·6)	−39·8	(12·8)
α^L	0·11	68·6	(12·0)	31·4	(13·9)
γ	0·05	−99·8	(20·7)	−3·4	(19·5)
Others	0·04				
Histidine (450)					
β	0·38	−112·3	(32·4)	144·1	(20·0)
α	0·31	−65·0	(9·7)	−39·7	(11·3)
γ	0·18	−99·3	(18·2)	−2·0	(15·6)
δ	0·07	−122·2	(19·6)	64·9	(18·5)
Left	0·05	58·3	(9·3)	43·1	(15·2)
Others	0·02				

Table 1 *(continued)*

Residue (total number) Zone	P†	ϕ‡ (°)	Δ_ϕ§ (°)	ψ‡ (°)	Δ_ψ§ (°)
Isoleucine (1273)					
β	0·52	−109·2	(22·1)	132·2	(15·8)
α	0·39	−65·8	(10·8)	−42·4	(10·0)
γ	0·07	−101·8	(15·5)	−9·6	(19·2)
δ	0·02	−117.4	(16·2)	78·8	(21·5)
Left	0·01	41·7	(20·1)	46·0	(9·4)
Leucine (1881)					
α	0·48	−64·6	(9·0)	−40·0	(10·1)
β	0·38	−101·5	(26·4)	136·6	(16·2)
γ	0·09	−95·2	(15·4)	−7·4	(16·5)
δ	0·03	−107·7	(23·3)	76·3	(19·9)
Left	0·01	58·0	(9·8)	37·7	(20·5)
Lysine (1497)					
α	0·46	−63·7	(10·1)	−39·1	(10·6)
β	0·35	−105·2	(30·4)	140·0	(17·6)
γ	0·12	−98·3	(17·6)	−8·9	(16·5)
Left	0·03	55·2	(9·1)	41·7	(13·5)
δ	0·02	−108·1	(23·9)	72·4	(19·8)
Others	0·01				
Methionine (428)					
α	0·52	−65·5	(8·6)	−39·4	(9·9)
β	0·35	−113·3	(28·8)	141·2	(17·3)
γ	0·07	−93·1	(13·1)	−3·0	(15·4)
δ	0·04	−94·6	(21·5)	76·6	(16·3)
Left	0·02	54·2	(12·6)	40·2	(22·9)
Phenylalanine (896)					
β	0·46	−110·4	(29·8)	141·4	(17·7)
α	0·35	−62·8	(9·5)	−42·5	(10·8)
γ	0·12	−102·3	(16·4)	−4·4	(17·1)
δ	0·05	−116·2	(21·9)	75·5	(19·0)
Left	0·01	65·1	(9·7)	29·6	(10·2)
Proline (966)					
β	0·51	−66·5	(10·4)	146·4	(14·8)
α	0·44	−62·6	(12·2)	−27·4	(15·5)
Others	0·04				
Serine (1487)					
β	0·42	−107·8	(34·1)	148·8	(17·8)
α	0·35	−64·9	(11·8)	−36·8	(13·2)
γ	0·15	−96·9	(19·1)	−4·7	(16·2)
δ	0·03	−123·7	(28·2)	71·7	(22·1)
Left	0·01	58·2	(12·6)	37·4	(20·8)
Others	0·03				
Threonine (1293)					
β	0·49	−111·7	(25·9)	145·2	(19·8)
α	0·31	−66·2	(12·6)	−40·2	(11·9)
γ	0·15	−103·9	(17·7)	−6·1	(16·4)
δ	0·02	−121·7	(17·8)	56·4	(20·6)
Left	0·01	48·3	(14·3)	35·9	(32·2)
Others	0·02				
Tryptophan (320)					
β	0·43	−105·8	(29·8)	139·6	(18·9)
α	0·42	−64·0	(11·1)	−40·8	(10·7)
γ	0·11	−100·1	(18·7)	−3·4	(20·2)
δ	0·03	−96·0	(17·2)	70·7	(17·6)
Left	0·02	63·5	(9·3)	28·8	(12·8)
Tyrosine (790)					
β	0·48	−114·0	(29·1)	142·5	(18·0)
α	0·33	−63·5	(9·6)	−42·3	(10·4)
γ	0·12	−103.0	(16·8)	−2·8	(16·2)
δ	0·03	−114·9	(21·9)	78·6	(14·8)
Left	0·03	61·8	(11·6)	32·9	(16·7)
Others	0·01				

Table 1 *(continued)*

Residue (total number) Zone	P†	ϕ‡ (°)	Δ_ϕ§ (°)	ψ‡ (°)	Δ_ψ§ (°)
Valine (1603)					
β	0·55	−112·7	(23·4)	135·5	(16·4)
α	0·36	−65·0	(9·4)	−41·9	(9·9)
γ	0·06	−106·5	(18·5)	−11·0	(19·9)
δ	0·02	−108·1	(24·1)	81·5	(19·9)
Left	0·01	36·0	(21·5)	42·0	(17·3)

Rotamer abbreviations are shown in Fig. 2.
† Fraction of points within a given zone.
‡ Average backbone angle.
§ Standard deviation from the average backbone angle.

Table 2

Probabilities, average positions and sizes of preferred zones for the side-chain torsion angles

Residue (total number) Rotamer	P	χ_1 (°)	(Δ_1) (°)	χ_2 (°)	(Δ_2) (°)
Arginine (1140)					
MT	0·43	−67·5	(15·4)	−176·6	(20·1)
TT	0·24	−174·3	(17·1)	179·9	(18·3)
MM	0·12	−63·7	(18·1)	−73·5	(20·7)
PT	0·09	63·7	(17·2)	174·8	(21·1)
TP	0·07	−173·9	(20·5)	71·9	(18·7)
Others	0·05				
Asparagine (1302)					
MN	0·33	−70·2	(14·4)	−40·2	(32·4)
TN	0·29	−169·7	(16·9)	−39·0	(88·7)
MT	0·21	−69·3	(16·6)	136·1	(38·8)
PN	0·16	62·9	(13·0)	−30·1	(81·5)
Aspartic acid (1754)					
MN	0·51	−70·1	(14·2)	159·1	(33·5)
TN	0·31	−170·9	(16·0)	−173·4	(42·7)
PN	0·18	62·5	(13·7)	174·9	(40·0)
Glutamine (1029)					
MT	0·36	−67·7	(15·3)	179·9	(17·1)
TT	0·21	−173·0	(18·1)	178·6	(19·0)
MM	0·15	−64·8	(15·8)	−66·8	(19·5)
TP	0·11	−172·8	(21·2)	68·4	(16·8)
PT	0·07	64·3	(20·6)	−179·2	(20·6)
MP	0·05	−73·6	(22·8)	74·1	(19·9)
Others	0·04				
Glutamic acid (1620)					
MT	0·35	−67·5	(16·3)	179·3	(18·4)
TT	0·24	−174·1	(19·7)	−179·9	(18·9)
MM	0·15	−66·6	(20·6)	−66·4	(20·2)
PT	0·07	59·8	(23·3)	−178·1	(21·5)
MP	0·07	−66·8	(21·7)	76·7	(18·1)
TP	0·06	−166·5	(20·4)	65·9	(18·2)
PM	0·03	55·1	(22·6)	−82·6	(15·9)
Others	0·02				
Histidine (629)					
ML	0·30	−65·1	(14·3)	−82·0	(31·0)
MR	0·24	−65·8	(12·1)	109·8	(34·8)
TR	0·19	−176·5	(11·9)	78·3	(31·3)
TL	0·15	−170·1	(13·9)	−101·5	(33·6)
PL	0·07	62·7	(12·0)	−86·9	(20·2)
PR	0·05	58·7	(15·0)	96·1	(31·9)

Table 2 *(continued)*

Residue (total number) Rotamer	P	χ_1 (°)	(Δ_1) (°)	χ_2 (°)	(Δ_2) (°)
Isoleucine (3014)					
MT	0·58	−64·2	(10·1)	168·4	(14·4)
PT	0·13	62·2	(12·6)	169·0	(14·4)
MM	0·13	−57·6	(10·6)	−62·4	(15·4)
TT	0·07	−175·3	(20·6)	167·7	(16·0)
MP	0·04	−70·4	(20·3)	72·8	(31·3)
TP	0·03	−165·3	(26·5)	71·4	(16·3)
Others	0·03				
Leucine (2277)					
MT	0·51	−66·3	(12·2)	175·5	(12·9)
TP	0·26	−177·9	(11·9)	65·5	(12·9)
MP	0·10	−99·6	(19·3)	44·3	(25·3)
TT	0·06	−159·8	(19·0)	−179·3	(30·6)
MM	0·05	−104·4	(40·3)	−56·5	(30·0)
Others	0·02				
Lysine (1762)					
MT	0·39	−68·9	(16·8)	−178·3	(21·4)
TT	0·26	−173·3	(17·4)	178·6	(21·9)
MM	0·12	−64·1	(17·0)	−71·1	(23·3)
PT	0·07	62·4	(18·8)	−179·9	(21·9)
TP	0·07	−174·3	(20·4)	76·6	(19·2)
MP	0·04	−87·6	(19·7)	75·6	(27·8)
Others	0·04				
Methionine (545)					
MT	0·34	−69·9	(13·6)	−178·0	(15·9)
MM	0·24	−64·0	(11·9)	−66·5	(16·4)
TT	0·19	−173·6	(17·1)	178·0	(17·8)
TP	0·09	−170·7	(13·4)	76·1	(18·7)
PT	0·08	62·7	(17·0)	−175·2	(17·7)
Others	0·06				
Phenylalanine (1142)					
MR	0·52	−66·8	(11·9)	98·7	(30·0)
TR	0·34	−177·4	(12·4)	76·9	(19·0)
PR	0·13	63·1	(11·6)	91·1	(13·2)
Tryptophan (426)					
MR	0·37	−67·0	(11·4)	98·4	(15·9)
TR	0·20	−179·4	(11·4)	71·4	(24·8)
ML	0·16	−69·0	(13·2)	−38·5	(47·5)
TL	0·12	179·6	(14·4)	−101·1	(14·0)
PL	0·10	62·0	(12·6)	−87·9	(9·2)
PR	0·05	61·0	(14·5)	84·3	(13·9)
Tyrosine (1086)					
MR	0·53	−66·1	(11·9)	99·8	(26·6)
TR	0·35	179·7	(11·8)	76·3	(20·6)
PR	0·12	64·3	(12·3)	87·4	(15·7)
Cysteine (554)					
P	0·16	63·1	(18·5)		
T	0·30	−177·1	(12·9)		
M	0·54	−64·8	(13·7)		
Serine (2058)					
P	0·43	63·6	(16·1)		
T	0·24	−179·0	(19·9)		
M	0·33	−64·8	(18·7)		
Threonine (1807)					
P	0·42	62·7	(13·4)		
T	0·10	−179·2	(25·3)		
M	0·48	−60·5	(14·2)		
Valine (2108)					
P	0·10	61·7	(26·5)		
T	0·67	174·7	(11·7)		
M	0·23	−60·9	(16·6)		

Table 3

Conformational entropy differences between free and buried states of amino acid side-chains

Residue	S_{stat} angles	TS_{stat} (kcal/mole)	S_{add} angles	N_{add}	TS_{add} (kcal/mole)	TS_{total} (kcal/mole)
Alanine						0·00
Glycine						0·00
Proline						0·00
Cysteine	$\chi1$	0·59	$\chi2$	2·5	0·55	1·14
Serine	$\chi1$	0·64	$\chi2$	2·5	0·55	1·19
Threonine	$\chi1$	0·57	$\chi2$, 1	2·5	0·55	1·12
Valine	$\chi1$	0·50		1	0	0·50
Asparagine	$\chi1$, $\chi2$	0·81		1	0	0·81
Aspartic acid	$\chi1$, $\chi2$	0·61		1	0	0·61
Histidine	$\chi1$, $\chi2$	1·00		1	0	0·99
Isoleucine	$\chi1$, $\chi2$	0·75		1	0	0·75
Leucine	$\chi1$, $\chi2$	0·75		1	0	0·75
Phenylalanine	$\chi1$, $\chi2$	0·58		1	0	0·58
Tryptophan	$\chi1$, $\chi2$	0·97		1	0	0·97
Tyrosine	$\chi1$, $\chi2$	0·58	$\chi6$	2	0·42	0·99
Glutamine	$\chi1$, $\chi2$	0·94	$\chi3$	6	1·08	2·02
Glutamic acid	$\chi1$, $\chi2$	1·00	$\chi3$	3	0·66	1·65
Methionine	$\chi1$, $\chi2$	0·87	$\chi3$	3	0·66	1·53
Arginine	$\chi1$, $\chi2$	0·82	$\chi3$, $\chi4$	9	1·32	2·13
Lysine	$\chi1$, $\chi2$	0·89	$\chi3$, $\chi4$	9	1·32	2·21

$S = S_{stat} + S_{add}$, the first term is calculated by the Boltzmann formula from the statistical distributions observed in 161 protein domains. $T = 300$ K.

entropic part of the free energy (TS) at $T = 300$ K over the solvent-accessible surface of the reference atom in the extended residue surrounded by glycines gives the following corrected surface energy densities to be used with MIMEL electrostatics: −16·0 for N^{ζ} of Lys; 0·0 for $N^{\eta1}$ and $N^{\eta2}$ of Arg; 7·0 for $N^{\delta2}$ and $O^{\delta1}$ of Asn; 6·0 for $N^{\varepsilon2}$ and $O^{\varepsilon1}$ of Gln; −18·0 for $N^{\varepsilon1}$ of Trp; 1·0 for $O^{\varepsilon1}$ and $O^{\varepsilon2}$ of Glu; −20·0 for S^{δ} of Met and 20·0 calories per mole/Å^2 for all other heavy atoms. These parameters were used in the BPMC simulations of the 12- and 16-residue peptides when electrostatic energy was calculated with the MIMEL approximation.

(d) *Modified image electrostatics*

Here we shall find the approximate solution of the electrostatic free energy of an arbitrarily shaped body (e.g. a protein molecule) with low dielectric constant ε_p containing fixed charges q_i ($i = 1, n$), which is surrounded by high dielectric media (ε_w). The approximation consists of two components: (1) a compact analytical solution of the problem when the body has an exact spherical shape; and (2) a fast method projecting the spherical solution onto the non-spherical body without building explicit image charges. An important feature of the projection method is its stable behaviour and ability to achieve a reasonable accuracy for wide ranges of shapes. The goal of this development is to incorporate the electrostatic free energy term into the global optimization procedure, a development which might be of principal importance for successful structure prediction.

(i) *Compact solution for a sphere*

Consider a sphere of radius R and dielectric constant ε_p embedded in a homogeneous, isotropic medium of dielectric constant ε_w (Fig. 3(a)). A potential $\Phi(r, \theta)$ created by a charge q residing at distance x from the centre of the sphere at any point inside the sphere consists of a standard Coulomb part and a reaction field potential (Friedman, 1975):

$$\Phi(r, \theta) = \frac{Cq}{\varepsilon_p r^q} - \frac{Cq(\varepsilon_w - \varepsilon_p)}{\varepsilon_p} \sum_{n=0}^{\infty} \left(\frac{n+1}{\varepsilon_p n + \varepsilon_w (n+1)} \right) \frac{x^n r^n}{R^{2n+1}} P_n(\cos(\theta)), \quad (4)$$

where C is a constant (C = 332, if charges are in electron units, distances in Å and energy in kcal/mole); $P_n(\cos(\theta))$ is the Legendre polynomial. The term $(n+1)/(\varepsilon_p n + \varepsilon_w(n+1))$ may be expanded in a Taylor series. There may be two ways to do the expansion based on the following rearrangements:

$$\left(\frac{n+1}{\varepsilon_p n + \varepsilon_w (n+1)} \right) = \frac{1}{\varepsilon_w} \left(1 + \frac{\varepsilon_p n}{\varepsilon_w (n+1)} \right)^{-1} \quad (5)$$

or

$$\left(\frac{n+1}{\varepsilon_p n + \varepsilon_w (n+1)} \right) = \frac{1}{(\varepsilon_w + \varepsilon_p)} \left(1 - \frac{\varepsilon_p}{(\varepsilon_w + \varepsilon_p)(n+1)} \right)^{-1}. \quad (6)$$

These expressions, when expanded into an infinite series, lead to 2 ways of representing the reaction potential: $\mathfrak{R} = B(0) + B(1) + B(2) + \ldots$ or $\mathfrak{R} = \mathfrak{R}(0) + \mathfrak{R}(1) + \mathfrak{R}(2) + \ldots$, respectively. The first has been used by Kirkwood (1934) and the second by Friedman (1975). The zero and first order terms of these series can be analytically summed. B(0) and $\mathfrak{R}(0)$ give the image approximation with the image charges of

$$-\frac{(\varepsilon_w - \varepsilon_p)}{\varepsilon_w} \frac{R}{x} q \quad \text{or} \quad -\frac{(\varepsilon_w - \varepsilon_p)}{(\varepsilon_w + \varepsilon_p)} \frac{R}{x} q,$$

respectively, located at the point $r = R^2/x$, $\theta = 0$.

The $\mathfrak{R}(0) + \mathfrak{R}(1) + \mathfrak{R}(2) + \ldots$ series converges more rapidly and provides a better approximation of the reaction potential than the Kirkwood expansion $B(0) + B(1) + B(2) + \ldots$. However, using only the $\mathfrak{R}(0)$ and $\mathfrak{R}(1)$ terms, for which a compact analytical form was found

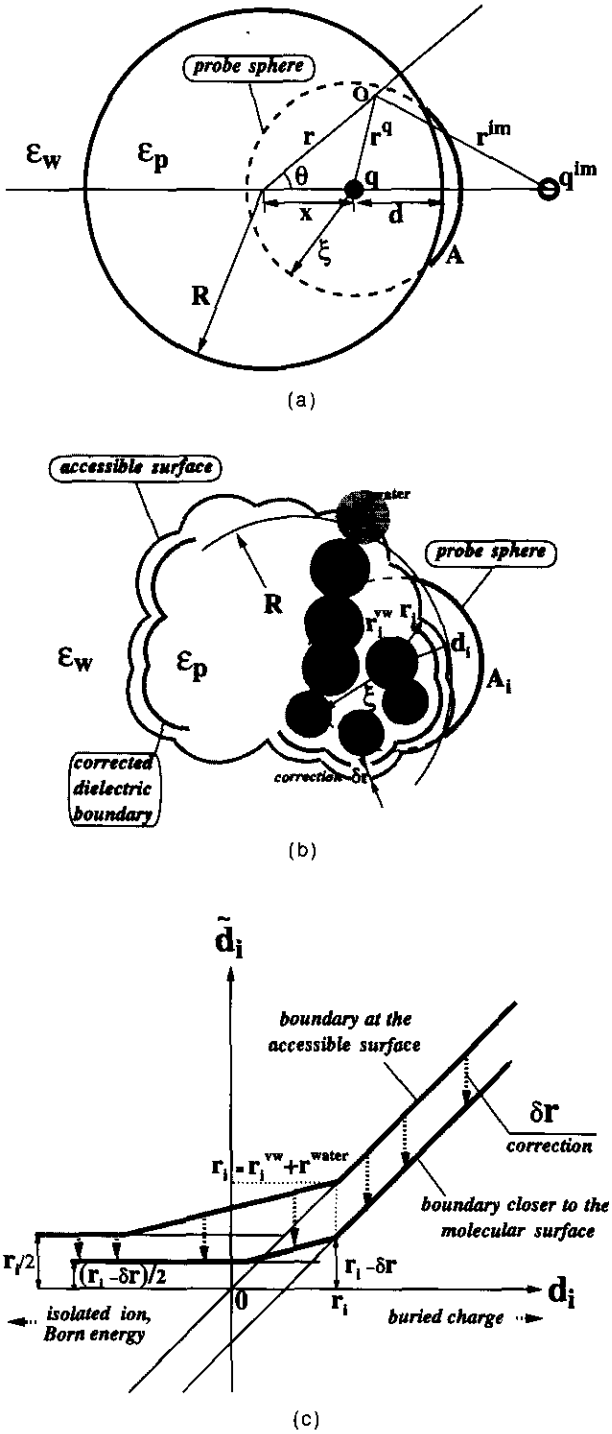


Figure 3. (a) Spherical body of radius R in a dielectric medium. The charge and its image as well as the point of observation O are shown. Accessible surface A of a probe sphere can be used to define depth d of the charge. (b) Protein in water. Atom i has van der Waals radius r_i^{vw}. To calculate the protein solvent-accessible surface, increased radii $r_i = r_i^{vw} + r^{water}$ are used. Accessible surface A_i of the probe sphere is used to assess the depth d_i by eqn (9). Distances d_i are later corrected to move the effective dielectric boundary from the protein solvent-accessible surface closer to the molecular surface. (c) Derivation of the final set of distances $\tilde{d}_i$ between the charge and the effective dielectric boundary in order to satisfy 2 conditions: for large positive d_i the distance should be decreased by δr, whereas for negative distances the asymptotic value of $\tilde{d}_i$ should be such that interaction with the image charge reproduces the Born energy of the charge of r_i–δr radius. Simple linear functions satisfying the above conditions for the initial dielectric boundary (bold dotted line) and the corrected one (bold solid line) are shown. $\delta r = r^{water}$ brings the effective dielectric boundary close to the molecular surface.

(Friedman, 1975), provides insufficient accuracy. For example, when x tends to zero so that charge is located in the centre of the sphere, the reaction field energy based on the $\Re(0)$ potential does not reproduce the Born formula.

We have found an approximation which is more accurate and less computationally intensive than $\Re = \Re(0) + \Re(1)$. It exploits the progressively weaker dependence of $\Re(m)$ on the charge position x inside the sphere and uses the x-independent part of the whole series $\Re(1) + \Re(2) + \Re(3) + \ldots$ rather than the exact expression for $\Re(1)$ as a correction term which is added to the image charge approximation $\Re(0)$. The electrostatic free energy of charges q_i in a sphere in the modified image approximation (MIMEL) is given by a formula consisting of a Coulomb image term and a correction term (see Appendix I), the last two terms approximating the reaction field energy:

$$\sum_{q_i, q_j, i<j} \frac{Cq_iq_j}{\varepsilon_p d_{ij}} + \frac{1}{2}\sum_{q_i, q_k^{im}} \frac{Cq_iq_k^{im}}{\varepsilon_p d_{ik}} - \frac{1}{2}\frac{C(q^{total})^2(\varepsilon_w - \varepsilon_p)}{R\varepsilon_w(\varepsilon_w + \varepsilon_p)}, \quad (7)$$

where r_{ij} is the distance between charges i and j, the image charge

$$q_i^{im} = -\frac{(\varepsilon_w - \varepsilon_p)}{(\varepsilon_w + \varepsilon_p)}\frac{R}{x_i} q_i \quad (8)$$

is located at the inverse point, and q^{total} is the net charge of all real charges in the system.

To calculate the electrostatic free energy for a real protein, the interaction energy between 2 charges i and j and 2 corresponding image charges should be expressed in terms of 2 depths d_i and d_j of the charges from the protein surface and their interatomic distance r_{ij} (Imoto, 1983). Different methods to estimate the distance d_j between charge and dielectric boundary have been proposed. Schaefer & Froemmel (1990) define the planar dielectric boundary from homogeneous redistribution of solute atoms inside the probe sphere. Tanford & Roxby (1972) and Imoto (1983) related d_j with accessible areas of charged atoms, which makes their method position-sensitive only for the surface atoms, whereas many partial charges like peptide group N and O atoms are not accessible, even though they play an important role in the electrostatic effect (Yang *et al.*, 1992).

Let us notice that the distance between the charge and the spherical dielectric boundary can be exactly expressed as a function of the accessible area of an artificial probe sphere of sufficiently large radius, centred at the charge. For an ideal sphere of radius R and probe radius ξ the distance d_i (Fig. 3(a)) may be exactly calculated as:

$$d_i = R\left(1 + \kappa(1 - 2a_i) - \sqrt{1 - \kappa^2(1 - (1 - 2a_i)^2)}\right), \quad (9)$$

where a_i is the exposed surface A_i of the probe sphere divided by its total surface and $\kappa = \xi/R$. The same formula can be applied to proteins. We evaluate an effective distance d_i between a charge and a dielectric boundary from the accessible surface of a probe sphere of increased radius $\xi = \kappa R$, which is set to be half of the effective molecular radius R (i.e. $\kappa = 0{\cdot}5$; Fig. 3(b)). The

molecular radius is estimated as

$$R = \sqrt[3]{\frac{3}{4\pi}(17{\cdot}5N_{\text{heavy}})},$$

where 17·5 Å^3 is the average volume of non-hydrogen atoms in proteins. Values of κ greater than 0·5 result in problems for proteins of remarkably non-spherical shape, because for atoms near the centre of the molecule the probe sphere may have too many exposed patches leading to unrealistic estimates of the depth. The chosen value of κ provides a reasonable estimate of charge depths, in spite of some inaccuracies for centrally located charges, since (1) d_i may be evaluated for about 7/8th fraction of molecular volume (in the worst case of spherical protein, the centrally located sphere where d_i is uncertain occupies $(1-\kappa)^{-3} = 1/8$th fraction of the total volume); (2) the remaining part contains on the average fewer charges; and (3) the electrostatic energy depends weakly on d_i for deeply buried charges.

The distances calculated from eqn (9), however, should be further corrected (1) to modify the position of the dielectric boundary (Fig. 3(b), (c)), and (2) to ensure correct asymptotic behaviour for the exposed charges having very small or negative d_i values. To calculate the accessible surface A_i of the probe sphere (Fig. 3(b)), all van der Waals radii are increased by the radius of a water molecule to exclude intramolecular cavities smaller than one water molecule sphere. Therefore distances d_i initially correspond to the effective dielectric boundary placed at the protein solvent-accessible surface. Although there is no consensus on the optimal atom radii and hence position of the dielectric boundary (Davis & McCammon, 1990), it might be necessary to move the boundary by decreasing distances d_i. The correction procedure, moving the dielectric boundary by δr towards the molecular surface, is shown in Fig. 3(b). To place the effective boundary at the van der Waals distance from the centres of surface atoms, the correction displacement δr should be set to r^{water}. The second requirement mentioned above (2) stems from the overexposed probe spheres (which could be the case for atoms at protruding protein regions) resulting in values d_i calculated by eqn (9) which are smaller than $r_i - \delta r$ or even negative. These values should be transformed so that the corrected distances $\tilde{d}_i$ tend to the half of the corrected atom radii $r_i - \delta r$ to ensure that the self-energy of the charge (the second term in eqn (7)) tends to the Born energy of isolated ion of the same radius. The actual d_i to $\tilde{d}_i$ transformation is shown in Fig. 3(c). We used the following van der Waals radii: H, 1 Å; C, 1·6 Å; O, 1·35 Å; N, 1·45 Å; the water molecule radius was set to 1·6 Å.

(e) *Energy calculations*

The basic description of the molecular system and fast algorithms to calculate energy terms and their derivatives with respect to the internal variables as implemented in the Internal Coordinate Mechanics (ICM) program are given elsewhere (Abagyan *et al.*, 1994). In the BPMC simulations, the following energy terms were calculated: (1) in the course of local minimization following every random step, the objective function consisted of ECEPP/2 energy potentials (Momany *et al.*, 1975; Nemethy *et al.*, 1983) with the distance-dependent dielectric constant $\varepsilon = 4r$ (McCammon *et al.*, 1979; Pickersgill, 1988), and these energy terms were calculated along with their analytical derivatives; (2) for the evaluation of the trial conformation in the Metropolis selection criterion (Metropolis *et al.*, 1953) 2 sets of energy terms have been tried, the 1st one consisted of the ECEPP/2 energy without Coulomb electrostatics, the MIMEL energy and the hydrophobic energy with the entropic correction (see section (c)), while the 2nd one consisted of the terms used in the local minimization supplemented with the solvation energy of Wesson & Eisenberg (1992). The hydrophobic energy density before entropic correction was assumed to be 20 cal/mole per Å^2. This number is a subject of some controversy (16 to 31 cal/mole per Å^2 for the microscopic surface energy density according to Sharp *et al.* (1991)) and was chosen somewhat arbitrarily, bearing in mind the future optimization of all the solvation and electrostatic parameters involved. Using two different energies for the minimization and the final evaluation at every MCM step (Abagyan *et al.*, 1994) was justified by the relatively weak dependence of the solvation and electrostatic terms on small conformational changes in the local minimization compared to the van der Waals term, and allowed an efficient local minimization using analytical energy derivatives.

3. Results

(a) *Accuracy of the modified image electrostatic calculations*

To evaluate the accuracy of MIMEL and $\Re(0)+\Re(1)$ approximations, we calculated the electrostatic reaction field energy with the DelPhi program which implements the finite-difference algorithm (Gilson & Honig, 1987; Nicholls & Honig, 1991). The DelPhi reaction field energy was calculated with the grid size 1 Å, ionic strength set to 0, $\varepsilon_w = 80$, and $\varepsilon_p = 4$. The default DelPhi van der Waals radii, increased by the water molecule radius, were used. The first comparison was made for the artificial quasi-spherical randomly charged "protein" (Fig. 4(a)). Atoms of van der Waals radius 1·6 Å were placed in the nodes of a three-dimensional cubic grid within the 10 Å sphere. In each comparison, 20 of them were randomly selected and charged by +1 or −1 electron charge unit (e.u.), to provide certain net charge of the globule. Forty different charge distributions were generated for the net charge of 0, 2, 4, 6, 8 and 10 e.u.

Three approximations were tested. The first one is the pure spherical image charge approximation (term $\Re(0)$). Comparison of this approximation with the DelPhi energies gives the regression line $E^{\text{MIMEL}} = 0{\cdot}952E^{\text{DelPhi}} - 1{\cdot}775$ with the standard deviation of points from the regression line of 2·26. The 5% error in the slope is caused by the stepwise jumps between the sets corresponding to the different net charges. Addition of $\Re(1)$ logarithmic terms to the pure image approximation almost eliminates these jumps and improves the gradient up to the value of 0·995. However, when the net charge and corresponding electrostatic energies are small, the error is quite large since the intercept becomes +7·37 kcal/mole. The MIMEL approximation which corresponds to $\Re(0)+\Re^{\text{corr}}$ reaction potential solves both problems, as it brings the gradient to 1·0 and the intercept to less than 1 kcal/mole. Figure 4(b)

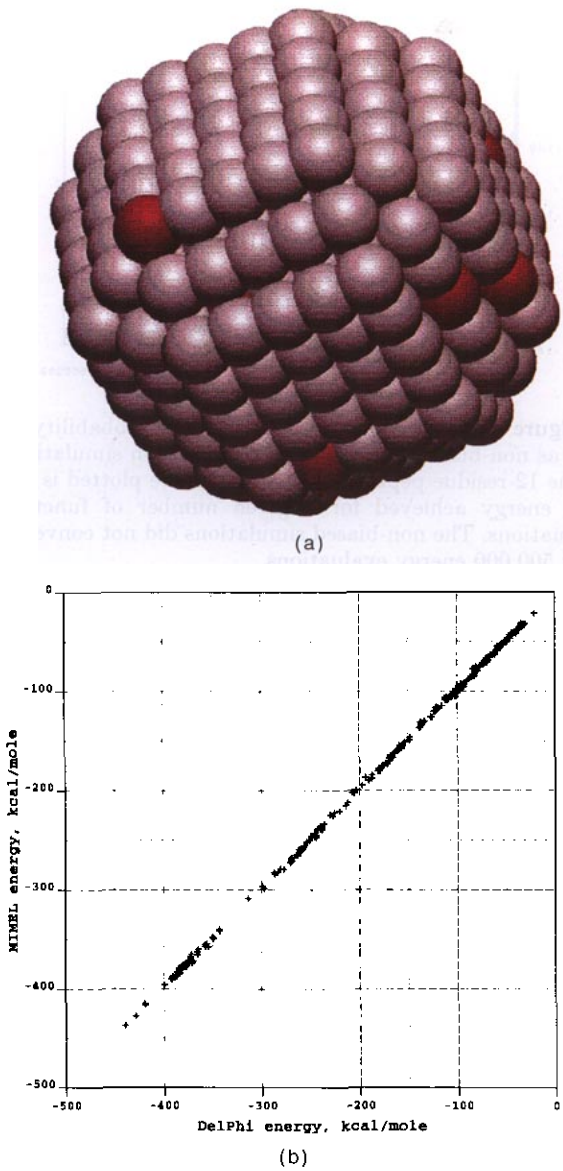


Figure 4. (a) Model quasi spherical protein containing 485 atoms. 240 randomly charged sets having net charge between 0 and 10 e.u. were generated. In each set, 20 randomly selected atoms were charged with either 1 or −1 e.u. (b) Comparison of the electrostatic reaction field energies calculated by the DelPhi program and the modified image approximation.

demonstrates the high precision of the MIMEL approximation. The exact linear regression line is $E^{\text{MIMEL}} = 0{\cdot}997E^{\text{DelPhi}} + 0{\cdot}614$ (the standard deviation of points from the regression line is 1·99).

The MIMEL approximation turned out to be also rather accurate for proteins of rather different shapes and sizes. Four proteins were taken from the PDB data base and regularized in full-atom representation by a standard ICM procedure (Eisenmenger *et al.*, 1993). These proteins are: catabolite gene activator (3GAP, Weber & Steitz, 1987), glycoprotein G (Gronenborn *et al.*, 1991), avian pancreatic polypeptide (1PPT, Blundell *et al.*, 1981) and a trypsin inhibitor from squash seeds (2CTI, Holak *et al.*, 1989).

To test the accuracy of the electrostatic calculation for the intermediate protein conformations we also generated a set of low-energy conformations for the trypsin inhibitor. We started from a random conformation and six different low-energy conformations were accumulated during 5000 steps of the BPMC procedure. One of them is shown in Figure 5(c). The correspondence between reaction field energies, calculated by our method and by DelPhi, for the proteins and intermediate conformations of the trypsin inhibitor is less accurate than for the quasi-spherical protein model, but is still rather high (Fig. 5(a)). The linear regression line for energies spanning about 200 kcal/mole is $E^{\text{MIMEL}} = 0{\cdot}93E^{\text{DelPhi}} - 1{\cdot}53$ with the standard deviation of points from the regression line of 2·47.

In previous calculations we used van der Waals radii increased by the water molecule radius to define the dielectric boundary. So far, no consensus on what is the optimal set of radii has been reached. To test the influence of the radii on the agreement, we recalculated the energies with effective atomic radii equal to the van der Waals radii (see Materials and Methods, end of section (d)). The regression line in this case reads: $E^{\text{MIMEL}} = 0{\cdot}91E^{\text{DelPhi}} - 8{\cdot}4$ with a standard deviation of 7·4 (Fig. 5(b)). The correspondence is still good, although worse than with the increased radii, when a protein looks more "spherical".

(b) *Structure prediction of 12-residue synthetic peptide*

The protein folding problem remains the main challenge of structural biology. In this section we demonstrate the efficiency of the BPMC procedure as a global search method. We also analyse what energy terms are necessary for a successful prediction. The 12-residue synthetic peptide Acetyl-Glu-Leu-Leu-Lys-Lys-Leu-Leu-Glu-Glu-Leu-Lys-Gly-COOH was crystallized and solved by Hill *et al.* (1990). The monomer forms an α-helix which is associated into higher order structures such as tetramers and hexamers. We tried to predict the conformation of the monomer in solution, believing that the α-helical conformation is a quasi-stable intermediate on the way to the ultimate higher order structure formation.

The BPMC simulation started from a random conformation. The preferred angular zones used to modify the probability distribution for a random step, were those listed in Tables 1 and 2. The parameters for van der Waals, hydrogen bonding, torsion energy terms and electric charges were taken from the ECEPP/2 potential (Momany *et al.*, 1975; Nemethy *et al.*, 1983). The electrostatic energy was calculated with a distance dependent dielectric constant (McCammon *et al.*, 1979). Additionally, solvation energy was calculated from the accessible surface using the atom parameters of Wesson & Eisenberg (1992). To speed up the calculations, the

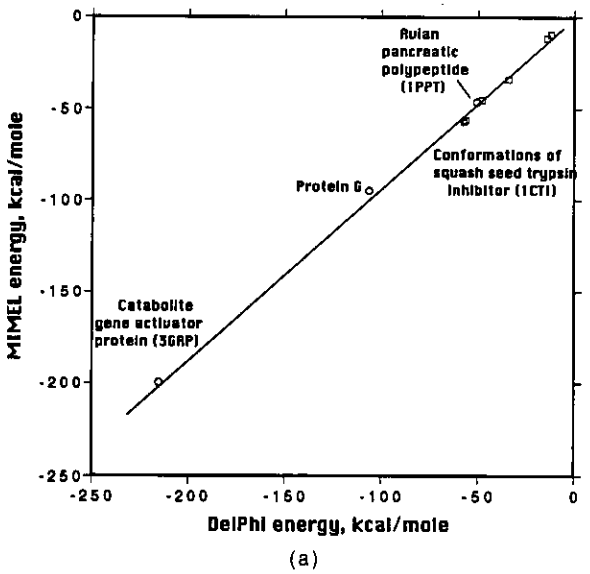


Figure 5. (a) Comparison of the MIMEL approximation with the DelPhi reaction field energies for 3 proteins (open circles) and generated intermediate conformations of trypsin inhibitor from squash seeds (open squares); (b) the same graph calculated with the dielectric boundary at the molecular surface. (c) Example of the trypsin inhibitor conformation.

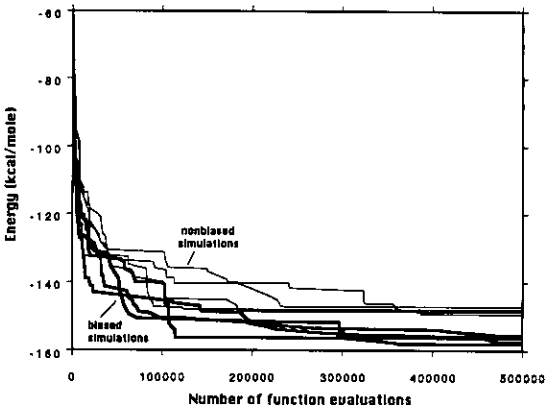


Figure 6. Energy profiles for the biased probability as well as non-biased Monte Carlo minimization simulations of the 12-residue peptide. The energy value plotted is the best energy achieved for a given number of function evaluations. The non-biased simulations did not converge over 500,000 energy evaluations.

solvation energy was not included in the local minimization, instead, it was added to other terms to decide whether a new conformation should be accepted (this procedure is described in detail by Abagyan *et al.* (1994)). Up to 35 low-energy conformations with a pairwise ϕ–ψ r.m.s. deviation greater than 25° were accumulated in a so-called conformational stack (a special data structure designed to keep track of different low-energy conformational families met during simulation, Abagyan & Argos (1992)). The maximum number of energy evaluations in every local minimization was set to 200. The simulation temperature was set to 600 K. All dihedral angles, including ω, were used as minimization variables. A non-biased (evenly distributed) random step was applied to angles not included in the preferred variable zones. The backbone ω angles were modified only during the local minimization.

To compare the efficiency of the BPMC procedure with the non-biased Monte Carlo minimization procedure (Li & Scheraga, 1987), we performed four simulations of each type starting from different random conformations. Typically, 500,000 energy evaluations are sufficient for the BPMC procedure to converge to the α-helical conformation, which corresponds to the global energy minimum for the energy terms described above. This is not the case for some other sets of energy terms (see below). Figure 6 shows the progression of the best energy achieved with the time of simulation for both evenly distributed random steps (Li & Scheraga, 1987) and the biased ones. None of the non-biased simulations converged over 500,000 energy evaluations. The best energy conformations achieved in these four runs had an energy ranging from -147 to -149 kcal/mole with conformations ranging from β-hairpins to assorted coils, often with α-helical fragments. In contrast, four BPMC runs achieved conformations with much lower energies between -155 and -158 kcal/mole, all of them being

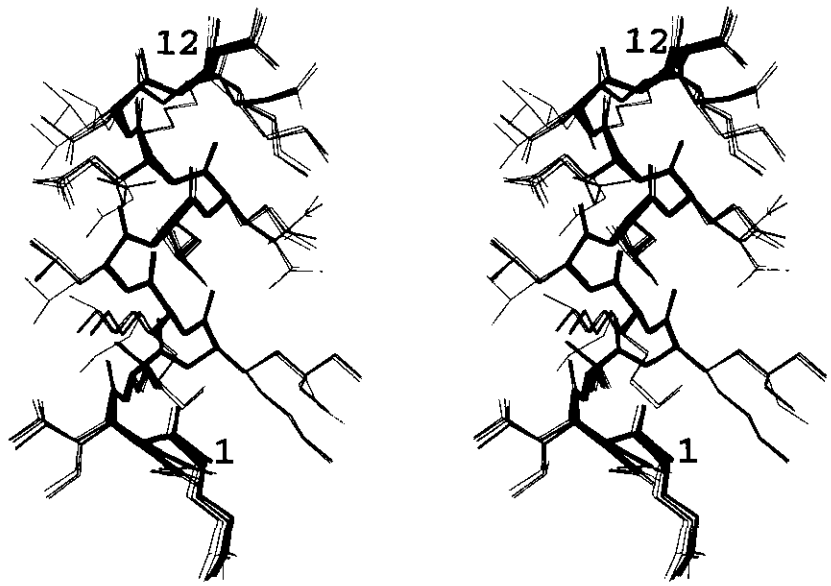


Figure 7. Stereo diagram of the superimposed low-energy conformations of the 12-residue peptide achieved in four BPMC runs. The C^{α}, N and C atoms of residues 3–10 were used for the superposition.

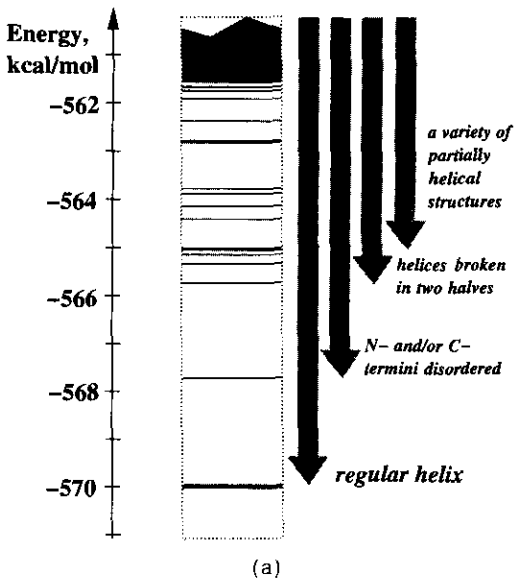
α-helices and differing mainly in their side-chain orientation. The best-fit superposition of 11 low-energy conformations accumulated in the BPMC simulations is shown on Figure 7. Two conformations with the lowest energies (−157·5 and −158·0 kcal/mole) are basically identical with a backbone r.m.s. deviation of only 0·05 Å. The backbone and all heavy atom r.m.s. deviations from the X-ray structure are 0·46 Å and 1·25 Å, respectively (disordered in the crystallographic structure Lys side-chains, and two C-terminal residues are omitted). The average pairwise coordinate r.m.s. deviation of the 11 conformations is 0·43 Å for the C^{α}, N and C atoms and 1·36 Å for all non-hydrogen atoms.

The correct conformation also appears to be the global energy minimum, if the electrostatic free energy in the MIMEL approximation and the surface tension with entropic correction (see section (c) of Materials and Methods) substitute for the Coulomb electrostatics with the distance-dependent dielectric constant and for the Wesson & Eisenberg (1992) solvation energy. However, the globally optimal conformation is different, when the solvation energy is left out and the calculation is carried out *in vacuo*, e.g. with ECEPP/2 potentials including the Coulomb electrostatics with $\varepsilon = 4$. The BPMC procedure easily finds non-helical conformations with lower energies. Typical features of these conformations are (1) clustering of positively and negatively charged atoms and (2) reduced compactness.

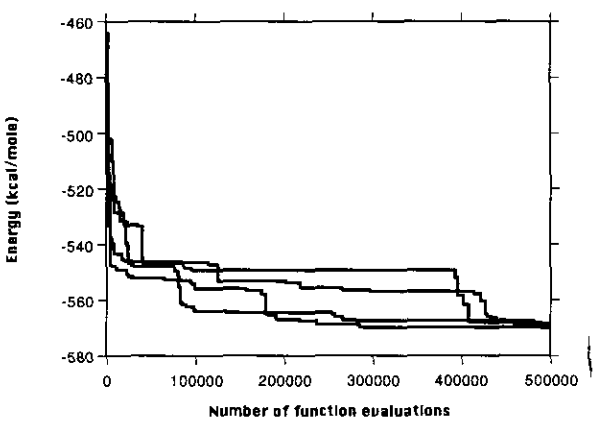
A possible criticism of the prediction described above would be that either the potentials or the global optimization method are biased towards α-helices. To check this, we carried out another series of simulations with the modified sequence Acetyl-Asp-Leu-Val-Lys-Lys-Val-Val-Asp-Asp-Val-Lys-Gly-COOH, where aspartic acid residues replace all glutamic acid residues and valines replace leucines at positions 3, 6 and 10. Such a modification preserves both the hydrophobicity pattern and the charge distribution, yet could possibly modify the conformational behaviour of the peptide, i.e. potentially disrupt the helix. Seven simulations of 1 million energy evaluations each resulted in 15 low-energy conformations having energies between −134 kcal/mole and −137 kcal/mole. In contrast to the original peptide where all low-energy conformations within 3 kcal/mole from the lowest one contained similar α-helices, low-energy conformations of the modified peptide were quite different and none of them contained more than one turn of α-helix. The lowest energy conformation had a β-turn at Asp8 and Asp9. The second best one forms a β-hairpin conformation (fragment Leu2-Val6 interacting with Asp8-Gly12 in an antiparallel manner). Inaccuracy in the potential functions (we observe many hydrogen bonds O_i–H_{i+2}, which might be an artefact of the non-directional hydrogen bonding potential in ECEPP/2) and incomplete treatment of the free energy (the backbone entropic term is missing) do not allow us to predict whether the conformation found will be sufficiently stable in solution. However, this conformation is certainly preferred over the α-helical one.

(c) *Structure prediction of the neutral water-soluble 16-residue peptide*

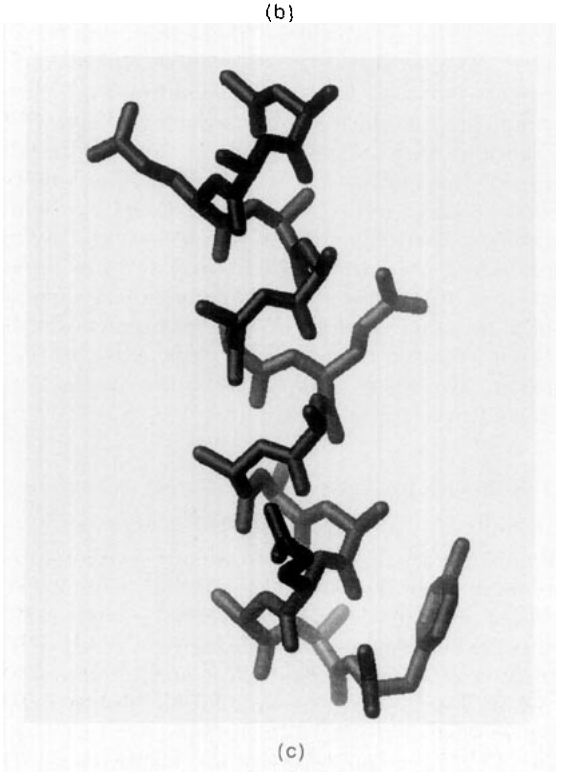
Following a suggestion of one of the referees, we also made a prediction for the neutral water-soluble α-helical peptide, described by Scholtz *et al.* (1991). The peptide sequence is Ac-$(AAQAA)_3Y(NH_2)$. The peptide was shown to be helical and monomeric in solution. We applied the same BPMC protocol as in the previous case. The following energy terms were used: ECEPP/2 potentials plus electrostatic free



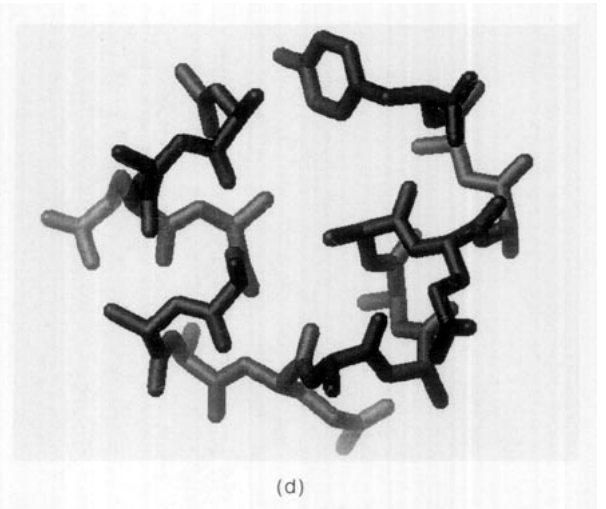
(a)



(b)



(c)



(d)

Figure 8. Structure prediction of 16-residue peptide (Scholtz *et al.*, 1991) by the BPMC procedure with MIMEL electrostatics and surface energy with entropic correction. (a) Filtered energy spectrum of low-energy conformations accumulated in 4 simulations. The filtering procedure was the following. All the low-energy conformations in the conformational stacks were sorted by energy. Then, for each conformation, starting from the lowest energy, all similar conformations (backbone r.m.s. deviation <1 Å) with higher energies were removed. (b) Energy profiles of 4 simulations. The best energy achieved over the specified number of energy evaluations is shown. (c) The lowest energy conformation. (d) A broken helix having energy 4 kcal/mole higher than the lowest one.

energy in the MIMEL approximation with the dielectric boundary at the van der Waals distance from the surface atom centres ($\delta r = r^{\text{water}}$), the surface tension with entropic correction (see sections (c), (d) and (e) of Materials and Methods).

Figure 8(a) shows a filtered spectrum of conformations retained in the conformational stacks by four BPMC simulations (Fig. 8(b)). The simulations show rather good convergence, all of them achieved the completely α-helical conformation, having energies of about -570 kcal/mole. The spectrum shows relative stability of the α-helical conformation (Fig. 8(a)). The lowest energy structure is shown in Figure 8(c). The lowest energy conformation in the spectrum is followed by a group of conformations with partial N- or C-terminal fragment rearrangements, and only after that does a topologically different conformation (Figure 8(d)) appear, containing two interacting α-helical fragments. The use of ECEPP/2 energies with the solvation energy as parameterized by Wesson & Eisenberg (1992) gives similar results, since most of the polar atoms were exposed so that solvation parameters account for the electrostatic polarization effects.

(d) *NMR structure determination*

Two families of methods are currently used for structure determination from NMR data: restrained molecular dynamics in Cartesian coordinate space

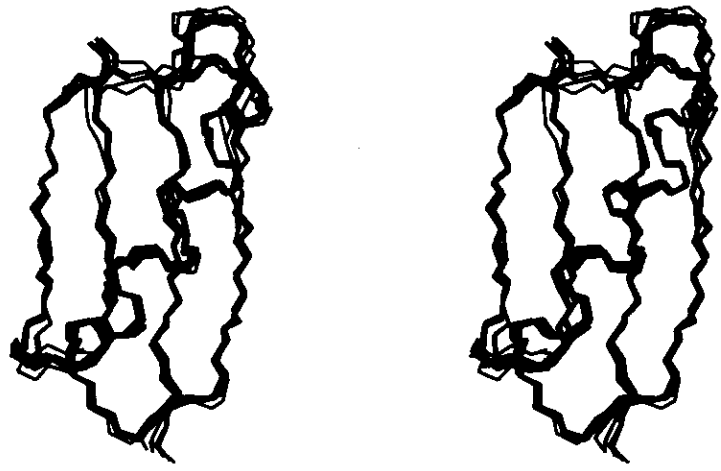


Figure 9. Stereo views of the 7 superimposed protein G conformations determined by the straightforward BPMC procedure with all restraints imposed simultaneously.

(Kaptein *et al.*, 1985; Clore *et al.*, 1985; Brünger *et al.*, 1986) and variable target function minimization in torsion angle space (Braun & Go, 1985). The advantage of methods operating in torsion angle space is a considerable reduction in the number of free variables, which is achieved by preserving the covalent geometry. The size of the search hyperspace depends exponentially on the number of free variables, which becomes important for medium and large proteins. The simplest approach would be just a minimization with all restraints superimposed, but this does not work because of local minima. The idea of the variable target algorithm, therefore, is to start minimization in small fragments and then gradually increase the window size. However, as pointed out by Güntert & Wüthrich (1991), the method has convergence problems, particularly for β-proteins. That is why those authors proposed a more sophisticated algorithm, accumulating and applying the dihedral angle restraints derived from preliminary variable target function calculations.

We tried to return to the simplest approach of imposing all restraints simultaneously while solving the local minima problem radically, by using the biased probability Monte Carlo minimization procedure instead of only minimization. A true global optimization procedure could be important not only for β-proteins and/or complex topologies. It is also much more stable with respect to data inaccuracies and ambiguities often inherent in NMR data. Ambiguous assignment of distance restraints can create multiple minima such that the total number of combinations may be much larger than any feasible number of random start conformations in the variable target function minimization procedure, and it is only the global search method which may approach the conformation with the lowest energy/penalty function.

To test the ability of the BPMC method to find the solution in a single simulation with all restraints imposed simultaneously, we took the immunoglobulin binding domain of streptococcal protein G solved by Gronenborn *et al.* (1991). The original experimental data, namely, 922 distance restraints and 105 dihedral angle restraints, are available as a test part of the X-PLOR program (Brünger, 1992). The preceding heavy atoms were used instead of pseudo-atoms where stereospecific assignments were not available. The functional form of distance restraints was biquadratic with upper and lower boundaries (Braun & Go, 1985). The variable restraints were imposed in the form described above. The same variable ranges were used in the BPMC procedure to make a random step if a restrained variable is selected. Apart from the variable and distance restraints, the penalty function also contained a soft van der Waals term and a torsion energy term (Momany *et al.*, 1975). A maximum of 250 function calls were allowed for the minimization part of the random step.

In seven of nine BPMC simulations starting from random conformations, the procedure converged to the solution over less than 100,000 function calls. The two unsuccessful simulations resulted in basically correct conformations with one or two local defects. The average pairwise r.m.s. deviation for the backbone N, C^{α} and C atoms is 0·7 Å, and for all heavy atoms 1·6 Å. The average r.m.s. deviations from the average structure determined originally by Gronenborn *et al.* (1991) are 1·0 Å and 1·8 Å, respectively. The best-fit superposition is shown in Figure 9.

4. Discussion

Available techniques of global energy optimization have not yet solved the protein folding problem. The reason is twofold: insufficient accuracy of the energy function, which should ideally represent the true free energy of a protein in solution, as well as insufficient efficiency of the optimization procedure. This paper makes a step towards solution of both problems.

Since it was realized that the empirical energy

functions *in vacuo* are not capable of distinguishing between correct and incorrect folds (Novotny *et al.*, 1984), several methods incorporating the solvation effects into structure prediction simulations have been proposed (Vila *et al.*, 1991; Wesson & Eisenberg, 1992; Williams *et al.*, 1992; Perrot *et al.*, 1992). In these models the free energy of solvation is related to the atomic accessible surface. However, the solvation energy consists of hydrophobic energy which can be reasonably related to the atomic accessibilities, and the electrostatic polarization energy which can not, because of the physical nature of electrostatic effect (Davis & McCammon, 1990). For small molecules, where most of the charges are exposed, the problem is not that acute; however, in globular proteins, contribution from the buried charges is significant (Yang *et al.*, 1992). The Coulomb formula with the distance-dependent dielectric constant is another frequently used alternative accounting for the electrostatic polarization (McCammon *et al.*, 1979; Mehler & Eichele, 1984; Pickersgill, 1988). This method may give reasonable results if the charge–charge interaction energy change upon a small conformational rearrangement or a point mutation is to be evaluated. We used this kind of energy term (Pickersgill, 1988) for the local energy minimization which is a part of one random step of the BPMC procedure. However, the lack of physical justification in general and of the self-energy in particular would lead to unacceptable inaccuracies in large-scale structure prediction simulations where the total free energies of rather different conformations are compared.

Our goal was to decouple the hydrophobic and the electrostatic parts of the solvation energy and to develop a computationally efficient algorithm for the electrostatic free energy calculations which can be incorporated in the extensive MC simulation. The approach to the energy function can be summarized as follows: (1) all-atom representation, including hydrogen atoms, is used to provide sufficient accuracy; (2) solvation free energy is represented by the surface-based hydrophobic energy; (3) electrostatic free energy is evaluated by the MIMEL approximation; and (4) the side-chain entropy is taken into account. Empirical parameters such as the surface free energy densities, effective atomic radii used to define the dielectric boundary, relative permittivity ε_p etc., were not optimized in this work. They can be substantially improved by a comparison of all the free energy terms with the experimental free energies.

The MIMEL approximation belongs to the family of methods based on analytical expressions for the electrostatic free energy of a system of charges in a spherical cavity surrounded by a dielectric medium (Kirkwood, 1934; Friedman, 1975). To apply this theory to a real protein, two steps must be taken; the infinite series representing the analytical solution has to be approximated by a simpler analytical expression (e.g. a classical image approximation, Friedman, 1975), and the effective distances d_i between the charge and the spherical boundary have to be evaluated (Imoto, 1983; Schaeffer & Froemmel, 1990). Neither task is trivial and they influence the accuracy greatly. As far as the first task is concerned, the expansion used in this work (Friedman, 1975) has better convergence properties than the frequently used Kirkwood expansion (1934) (Imoto, 1975) and the analytical correction term derived here (eqn (7)) improves the accuracy of the image approximation. The second task, as it is implemented here (Fig. 3(b), (c)), has the advantages of being stable with respect to the surface irregularities and, more importantly, the definition of the effective distances for the charges close to the protein surface (Fig. 3(c)) ensures the correct asymptotic behaviour. The comparison with the finite-difference method (Nicholls & Honig, 1991) illustrates the accuracy of the MIMEL method. The method is sufficiently fast and simple to be used in the MC simulations.

Estimates of the side-chain entropies (Table 3) underline the importance of the inclusion of the entropic free energy term in the structure prediction simulations. Indeed, the exposure of one CH_2 group of a long side-chain to the solvent, accompanied by liberation of three additional rotameric states results in an estimated entropic gain of about $-R \ln 3 \approx -0{\cdot}66$ kcal/mole, which is comparable with a hydrophobic energy loss of about 0·88 kcal/mole (Yang *et al.*, 1992). Relating the entropic term to the accessible surface of certain reference atoms is a simple, albeit not very accurate, way to incorporate this term in the energy calculations. However, in the examples considered (12-residue and 16-residue peptides), the accuracy appeared to be sufficient to uniquely identify the native three-dimensional structure.

The efficiency of the global optimization procedure is perhaps an even more important component of the structure prediction problem. Mentioning an astronomical time, which a systematic search procedure would require to find the global minimum, has become a commonplace. The Monte Carlo method has a large potential as a global optimization method; however, in its straightforward implementation, it is clearly unable to solve the problem. The difference between the BPMC and other methods from the MC family lies in the way the random conformational change (MC-step) is made. Two principal ideas aimed at an improved MC-step for protein modelling have been proposed and successfully implemented so far. The first one relies on the harmonic approximation of the energy surface near its minimum. The idea of Noguti & Go (1985) was to bypass a step size limitation, imposed by a strong anisotropy of the energy surface, by using collective variables corresponding to eigenvectors of the second derivative matrix of the energy function. Explicit calculation of the matrix of second derivatives as well as eigenvectors and eigenvalues can also be avoided in an approximation proposed by Vanderbilt & Louie (1984). However, both methods improve the quality of only the local MC-steps, which become redundant when a

local minimization follows the random change (Li & Scheraga, 1987). The second idea was to direct the movements of permanent protein dipoles by the local electrostatic field created by the whole molecule (Ripoll & Scheraga, 1988). Such a driving force may increase the performance of the global search; however, it is not clear whether the electrostatic forces in particular, calculated *in vacuo*, always steer the search trajectory in the correct direction rather than mislead it by neglecting other components of the energy function. The electrostatically driven conformational changes, consisting of the local rotation of the peptide plane, do not change the overall topology, thus restricting the sampling power of the procedure.

The BPMC procedure introduces a different basis for the random change, which now uses knowledge about local conformational preferences. These preferences, in the form of continuous probability distributions, can be either deduced from energy calculations or taken from known three-dimensional structures. Currently, the probability distributions of the main-chain and side-chain torsion angles in representative protein structures are taken directly as a distribution function of a random conformational change. The optimality of such a choice can be proven mathematically under some simplifying conditions. The comparison between the biased and non-biased simulations of the 12-residue peptide demonstrates the greatly increased sampling efficiency of the BPMC procedure.

Two important factors of any non-local-step-MC procedure are the number of angles changed simultaneously and the fraction of accepted trial conformations (acceptance ratio). Increasing the first would improve the coverage of conformational space, whereas increasing the second would reduce the fraction of wasted trials. These two parameters are conflicting, i.e. increased number of angles leads to a decreased acceptance ratio. Changing only one angle at a time was proven to be optimum for the "non-biased" Monte Carlo-minimization procedure (Li & Scheraga, 1987). The BPMC-step allows changing of several torsion angles simultaneously while still keeping the acceptance ratio rather high (about 50%).

In this work, a relatively rough approximation to the observed probability distribution, a set of Gaussian distributions, was used. We assume that local energy minimization of trial conformations, as well as restricted accuracy of the data-base derived preferences as applied to a particular peptide, make more complex and detailed description of the distribution unnecessary. In the current implementation, variables within the zone (e.g. χ-angle values in conformations corresponding to a certain side-chain rotamer) are considered as uncorrelated. This simplifies the zone description and generation of the random step. For most of the χ-zones, as well as for the β-zone of the main-chain, the shape of the distribution (Fig. 2) suggests the absence of strong correlation between angles. The α-zone, where the correlation is strong, was divided into two roughly round subzones to cover the slanting ellipsoid of the α-zone. However, as noted above, the procedure is rather stable with respect to inaccuracies in the representation of the actual probability distribution.

The concept of the BPMC is general, and not restricted to a particular division of the conformational space into subspaces. The obvious extension would be use of a joint probability distribution of the backbone and the side-chain torsion angles for each residue type. Considering several residue fragments could become another step towards greater efficiency of the BPMC procedure.

Appendix I: Optimal Probability Distribution for a Random Step in the Monte Carlo Procedure

Theorem

Suppose we want to find a global minimum of an objective function dependent on variables, which can be divided into n m-dimensional subspaces of similar type. The target variable set is: $(\mathbf{x}_1^0, \mathbf{x}_2^0, \ldots, \mathbf{x}_i^0, \ldots, \mathbf{x}_n^0)$, where $\mathbf{x}$ is a point in m-dimensional space of one zone. For example, $\mathbf{x}$ may be a pair of (ϕ, ψ), so that $\mathbf{x}_1^0, \ldots, \mathbf{x}_n^0$ describe the backbone conformation of an n-residue peptide. Let us consider a simplified prediction procedure consisting of a series of random selections of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i$, and $\mathbf{x}_n$ generated with a probability function $f(\mathbf{x})$, which is the same for all subspaces. We consider the prediction for subspace i as being successful if $\mathbf{x}$ gets sufficiently close to the target point $\mathbf{x}_i^0$ so that it can be further refined. We assume that a subsequent refinement procedure brings the $\mathbf{x}$ to $\mathbf{x}_i^0$ with the probability $h(\mathbf{x}, \mathbf{x}_i^0)$, which is a certain bell-shaped function centred at $\mathbf{x}_i^0$. Let $S(\mathbf{x})$ be the known distribution function of points $(\mathbf{x}_1^0, \mathbf{x}_2^0, \ldots, \mathbf{x}_i^0, \ldots, \mathbf{x}_n^0)$.

We want to prove the following statement: for $f(\mathbf{x})$ and $S(\mathbf{x})$ belonging to the class $\mathfrak{I}$ of smooth functions with respect to the bell function $h(\mathbf{x}, \mathbf{x}^0)$, so that:

$$\int_x h(\mathbf{x}, \mathbf{x}^0) g(\mathbf{x})\, d\mathbf{x} \approx \text{const}\, g(\mathbf{x}^0) \qquad \text{(A1.1)}$$

for any $g(\mathbf{x}) \subset \mathfrak{I}$, the optimal function $f(\mathbf{x})$ maximizing the probability of successful prediction is equal to $S(\mathbf{x})$.

Proof

The probability of successful prediction is given by a product of n integrals:

$$P = \prod_{i=1}^{n} \int h(\mathbf{x}, \mathbf{x}_i^0) f(\mathbf{x})\, d\mathbf{x}. \qquad \text{(A1.2)}$$

To find a maximum (stationary) value of this product we shall consider variations of the function $f(\mathbf{x})$. That is, we vary $f(\mathbf{x})$ to $f(\mathbf{x}) + \lambda f_1(\mathbf{x})$, where

$f_1(\mathbf{x})$ is an arbitrary function obeying the condition:

$$\int f_1(\mathbf{x})\,d\mathbf{x} = 0 \qquad \text{(A1.3)}$$

which results from two normalization conditions for the original and the varied probability functions:

$$\int f(\mathbf{x})\,d\mathbf{x} = 1 \quad \text{and} \quad \int (f(\mathbf{x}) + \lambda f_1(\mathbf{x}))\,d\mathbf{x} = 1. \qquad \text{(A1.4)}$$

We search for the stationary value of P as a function of the parameter λ. The condition for this is the vanishing of the derivative with respect to λ. Inserting the varied function into the expression for P and differentiating with respect to λ, we obtain:

$$\frac{\partial P}{\partial \lambda} = \sum_{i=1}^{n} \left\{ \int h(\mathbf{x}, \mathbf{x}_1^0) f(\mathbf{x})\,d\mathbf{x} \int \ldots \int h(\mathbf{x}, \mathbf{x}_i^0) f_1(\mathbf{x})\,d\mathbf{x} \ldots \int h(\mathbf{x}, \mathbf{x}_n^0) f(\mathbf{x})\,d\mathbf{x} \right\} = 0. \qquad \text{(A1.5)}$$

Rearranging the sum after multiplication and division of each summand by $\int h(\mathbf{x}, \mathbf{x}_i^0) f(\mathbf{x})\,d\mathbf{x}$:

$$\left(\prod_{i=1}^{n} \int h(\mathbf{x}, \mathbf{x}_i^0) f(\mathbf{x})\,d\mathbf{x} \right) \left(\sum_{i=1}^{n} \frac{\int h(\mathbf{x}, \mathbf{x}_i^0) f_1(\mathbf{x})\,d\mathbf{x}}{\int h(\mathbf{x}, \mathbf{x}_i^0) f(\mathbf{x})\,d\mathbf{x}} \right) = 0. \qquad \text{(A1.6)}$$

Using the above-formulated condition (A1.1) for smoothness of functions $f(\mathbf{x})$ and $f_1(\mathbf{x})$ with respect to the bell-shaped functions $h(\mathbf{x}, \mathbf{x}^0)$, one can get rid of the integrals in numerator and denominator of the last multiplicand and simplify the optimality condition:

$$\sum_{i=1}^{n} \frac{f_1(\mathbf{x}_i^0)}{f(\mathbf{x}_i^0)} = 0. \qquad \text{(A1.7)}$$

This sum may be approximated by an integral using the observed probability distribution $S(\mathbf{x})$:

$$\int S(\mathbf{x}) \frac{f_1(\mathbf{x})}{f(\mathbf{x})}\,d\mathbf{x} = \int \left(\frac{S(\mathbf{x})}{f(\mathbf{x})} \right) f_1(\mathbf{x})\,d\mathbf{x} = 0. \qquad \text{(A1.8)}$$

Condition (A1.3) for $f_1(\mathbf{x})$ leads to: $\dfrac{S(\mathbf{x})}{f(\mathbf{x})} = \text{const}$

or

$$f(\mathbf{x}) = S(\mathbf{x}) \qquad \text{(A1.9)}$$

because of the normalization condition for the probability distributions.

Appendix II: Formula for the Modified Image Approximation of the Electrostatic Free Energy

Let us consider a sphere with one charge in the dielectric media (Fig. 3(a)). The reaction potential in the sphere reads:

$$\mathfrak{R}(r, \theta) = \frac{Cq(\varepsilon_w - \varepsilon_p)}{\varepsilon_p} \sum_{n=0}^{\infty} \left(\frac{n+1}{\varepsilon_p n + \varepsilon_w (n+1)} \right) \times \frac{x^n r^n}{R^{2n+1}} P_n(\cos(\theta)). \qquad \text{(A2.1)}$$

The factor $(n+1)/(\varepsilon_p n + \varepsilon_p(n+1))$ can be rearranged and expanded in increasing powers of $\varepsilon_p/(\varepsilon_w + \varepsilon_p)(n+1) \ll 1$:

$$\left(\frac{n+1}{\varepsilon_p n + \varepsilon_w (n+1)} \right) = \frac{1}{(\varepsilon_w + \varepsilon_p)} \left(1 + \frac{\varepsilon_p}{(\varepsilon_w + \varepsilon_p)(n+1)} + \left(\frac{\varepsilon_p}{(\varepsilon_w + \varepsilon_p)(n+1)} \right)^2 + \cdots \right). \qquad \text{(A2.2)}$$

Substituting this formula into the reaction potential and noting that the zero-th term contains the expansion of $1/r^{im}$ in spherical harmonics, progressively decreasing contributions to the reaction potential are obtained:

$$\mathfrak{R}(r, \theta) = C \frac{(\varepsilon_w - \varepsilon_p) R}{\varepsilon_p (\varepsilon_w + \varepsilon_p) x} q \frac{1}{r^{im}} + \mathfrak{R}(1) + \mathfrak{R}(2) + \cdots, \qquad \text{(A2.3)}$$

where C is a constant of proportionality, and r^{im} is the distance between the position of image charge and the point of observation inside the sphere. The reaction field energy caused by this potential is:

$$E^{im} = \frac{1}{2} q \mathfrak{R}(0)(x, 0) = \frac{1}{2} \frac{Cq^2 R(\varepsilon_w - \varepsilon_p)}{\varepsilon_p (\varepsilon_w + \varepsilon_p) x r^{im}}. \qquad \text{(A2.4)}$$

Substituting r^{im} by $(R^2/x - x)$ and letting x tend to zero gives the image approximation of the reaction field energy for a charge in the centre of the sphere:

$$E^{im}_{x=0} = -\frac{1}{2} \frac{Cq^2 (\varepsilon_w - \varepsilon_p)}{\varepsilon_p (\varepsilon_w + \varepsilon_p) R}. \qquad \text{(A2.5)}$$

However, for this particular case, the exact electrostatic reaction field energy is known and given by the Born formula:

$$E^{Born} = -\frac{1}{2} \frac{Cq^2 (\varepsilon_w - \varepsilon_p)}{\varepsilon_p \varepsilon_w R}. \qquad \text{(A2.6)}$$

Using the increasingly weaker dependence of $\mathfrak{R}(1)$, $\mathfrak{R}(2)$, $\mathfrak{R}(3)$, ... on x, we propose an approximation where, instead of an exact expression for $\mathfrak{R}(1)$, we use the position independent addition $\mathfrak{R}(\text{corr})$ to the reaction potential, which represents the position independent part of all the terms $\mathfrak{R}(1)$, $\mathfrak{R}(2)$, $\mathfrak{R}(3)$, ..., so that for the particular case of charges in the centre one gets the exact solution. The additional term $\mathfrak{R}^{corr}_{ik}$ can be computed as:

$$\mathfrak{R}^{corr} = \frac{2(E^{Born} - E^{(0)}_{x=0})}{q} = -\frac{Cq(\varepsilon_w - \varepsilon_p)}{R\varepsilon_w(\varepsilon_w + \varepsilon_p)}. \qquad \text{(A2.7)}$$

For a system of charges q_i, $i = 1, \ldots, n$, the reaction field energy in the MIMEL approximation is:

$$E^{\text{MIMEL}} = \frac{1}{2}\sum_{q_i}\sum_{q_k^{\text{im}}}(\mathfrak{R}_{ik}^{\text{im}} + \mathfrak{R}_{ik}^{\text{corr}})q_i, \qquad \text{(A2.8)}$$

where $\mathfrak{R}_{ik}^{\text{im}}$ and $\mathfrak{R}_{ik}^{\text{corr}}$ are the image and the correction reaction potentials produced by atom k at the position of atom i. Substituting the expressions for these terms, noting that

$$\sum_i \sum_k q_i q_k = \left(\sum_i q_i\right)^2 \qquad \text{(A2.9)}$$

and defining q^{total} as $\sum q_i$, we obtain the following expression for the reaction field energy:

$$E^{\text{MIMEL}} = \frac{1}{2}\sum_{q_i, q_k^{\text{im}}} \frac{C q_i q_k^{\text{im}}}{\varepsilon_p r_{ik}} - \frac{1}{2}\frac{C(q^{\text{total}})^2(\varepsilon_w - \varepsilon_p)}{R\varepsilon_w(\varepsilon_w + \varepsilon_p)}. \qquad \text{(A2.10)}$$

We are grateful to Patrick Argos for his support, to Michael Nilges for helpful discussions and NMR data, to Jaap Heringa for the lists of proteins used in the analysis and to Toby Gibson and Patrick Argos for careful reading of the manuscript and valuable suggestions. This work was supported in part by a grant FG5-1075 from the German Bundesministerium für Forschung und Technologie.

References

Abagyan, R. A. & Argos, P. (1992). Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. *J. Mol. Biol.* **225**, 519–532.

Abagyan, R. A., Totrov, M. M. & Kuznetsov, D. A. (1994). ICM: an efficient technique for structure predictions and design. *J. Comp. Chem.*, in the press.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Blundell, T. L., Pitts, J. E., Tickle, I. J., Wood, S. P. & Wu, C.-W. (1981). X-ray analysis (1.4 Å resolution) of avian pancreatic polypeptide: small globular protein hormone. *Proc. Nat. Acad. Sci., U.S.A.* **78**, 4175 4179.

Braun, W. & Go, N. (1985). Calculation of protein conformations by proton–proton distance constraints. *J. Mol. Biol.* **186**, 611–626.

Brünger, A. T. (1992). X-PLOR Software Manual, Version 3.0. Yale University, New Haven, CT.

Brünger, A. T., Clore, G. M., Gronenborn, A. M. & Karplus, M. (1986). Three-dimensional structures of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. *Prot. Nat. Acad. Sci., U.S.A.* **83**, 3801–3805.

Bruccoleri, R. E. & Karplus, M. A. (1990). Conformational sampling using high-temperature molecular dynamics. *Biopolymers*, **29**, 1847–1862.

Clore, G. M., Gronenborn, A. M., Brünger, A. T. & Karplus, M. (1985). Solution conformation of a heptadecapeptide comprising the DNA binding helix F of the cyclic AMP receptor protein of *Escherichia coli*. Combined use of 1H nuclear magnetic resonance and restrained molecular dynamics. *J. Mol. Biol.* **186**, 435–455.

Davis, M. E. & McCammon, J. A. (1990). Electrostatics in biomolecular structure and dynamics. *Chem. Rev.* **90**, 509–521.

Eisenmenger, F., Argos, P. & Abagyan, R. A. (1993). A method to configure protein side-chains from the main-chain trace in homology modelling. *J. Mol. Biol.* **231**, 849–860.

Friedman, H. L. (1975). Image approximation to the reaction field. *Mol. Phys.* **29**, 1533–1543.

Gilson, M. K. & Honig, B. (1987). Calculation of electrostatic potentials in an enzyme active site. *Nature (London)*, **330**, 84–86.

Gilson, M. K. & Honig, B. (1991). The inclusion of electrostatic hydration energies in molecular mechanics calculations. *J. Comp. Mol. Design*, **5**, 5–20.

Gronenborn, A. M., Filpula, D. R., Essig, N. Z., Achari, A., Whitlow, M., Wingfield, P. T. & Clore, G. M. (1991). The immunoglobulin binding domain of streptococcal protein G has a novel and highly stable polypeptide fold. *Science*, **253**, 657–661.

Güntert, P. & Wüthrich, K. (1991). Improved efficiency of protein structure calculations from NMR data using the program DIANA with redundant dihedral angle constaints. *J. Biomol. NMR*, **1**, 447–456.

Heringa, J., Sommerfeldt, H., Higgins, D. & Argos, P. (1992). OBSTRUCT: a program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. *CABIOS*, **8**, 599–600.

Hill, C. P., Anderson, D. H., Wesson, L., DeGrado, W. F. & Eisenberg, D. (1990). Crystal structure of alpha-1: implications for protein design. *Science*, **249**, 543–546.

Holak, T. A., Gondol, D., Otlewski, J. & Wilusz, T. (1989). Determination of the complete three-dimensional structure of the trypsin inhibitor from squash seeds in aqueous solution by nuclear magnetic resonance and a combination of distance geometry and dynamical simulated annealing. *J. Mol. Biol.* **210**, 635–648.

Imoto, T. (1983). Electrostatic free energy of lysozyme. *Biophys. J.* **44**, 293–298.

Kaptein, R., Zuiderweg, E. R. P., Scheek, R. M., Boelens, R. & van Gunsteren, W. F. (1985). A protein structure from nuclear magnetic resonance data. Lac repressor headpiece. *J. Mol. Biol.* **182**, 179–182.

Kawai, H., Kikuchi, T. & Okamoto, Y. (1989). A prediction of tertiary structures of peptide by the Monte Carlo simulated annealing method. *Protein Eng.* **3**, 85–94.

Kirkpatrick, S., Gellatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, **220**, 671–680.

Kirkwood, J. G. (1934). Theory of solutions of molecules containing widely separated charges with special applications to zwitterions. *J. Chem. Phys.* **2**, 351–461.

Leach, A. R. (1991). A survey of methods for searching the conformational space of small and medium-sized molecules. In *Reviews in Computational Chemistry II* (Lipkowitz, K. B. & Boyd, D. B., eds), vol. 1, pp. 1–55, VCH Publishers, Inc., New York.

Li, Z. & Scheraga, H. A. (1987). Monte Carlo minimization approach to the multiple-minima problem in protein folding. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 6611–6615.

Mazur, A. K., Dorofeev, V. E. & Abagyan, R. A. (1991). Derivation and testing of explicit equations of motion for polymers described by internal coordinates. *J. Comput. Phys.* **92**, 261–272.

McCammon, J. A., Wolynes, P. G. & Karplus, M. (1979). Picosecond dynamics of tyrosine side chains in proteins. *Biochemistry*, **18**, 927–942.

Mehler, E. L. & Eichele, G. (1984). Electrostatic effect in water-accessible regions of proteins. *Biochemistry*, **23**, 3887–3891.

Metropolis, N. A., Rosenbluth, A. W., Rosenbluth, N. M., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.

Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (1975). Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions and intrinsic torsional potentials for the naturally occurring amino acids. *J. Phys. Chem.* **79**, 2361–2381.

Nemethy, G., Pottle, M. S. & Scheraga, H. A. (1983). Energy parameters in polypeptides. 9. Updating of geometric parameters, nonbonded interactions and hydrogen bond interactions for the naturally occurring amino acids. *J. Phys. Chem.* **87**, 1883–1887.

Nicholls, A. & Honig, B. (1991). A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. *J. Comput. Chem.* **12**, 435–445.

Noguti, T. & Go, N. (1985). Efficient Monte Carlo method for simulation of fluctuating conformations of native proteins. *Biopolymers*, **24**, 527–546.

Novotny, J., Bruccoleri, R. & Karplus, M. (1984). An analysis of incorrectly folded protein models. *J. Mol. Biol.* **177**, 787–818.

Perrot, G., Cheng, B., Gibson, K. D., Vila, J., Plamer, K. A., Nayeem, A., Maigret, B. & Scheraga, H. A. (1992). MSEED: a program for the rapid analytical determination of accessible surface areas and their derivatives. *J. Comp. Chem.* **13**, 1–11.

Pickersgill, R. W. (1988). A rapid method of calculating charge–charge interaction energies in proteins. *Protein Eng.* **2**, 247–248.

Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775–791.

Ripoll & Scheraga (1988). On the multiple-minima problem in the conformational analysis of polypeptides. II. An electrostatically driven Monte Carlo method-test on poly(L-alanine). *Biopolymers*, **27**, 1283–1303.

Schaefer & Froemmel (1990). A precise analytical method for calculating the electrostatic energy of macromolecules in aqueous solution. *J. Mol. Biol.* **216**, 1045–1066.

Schaumann, T., Braun, W. & Wüthrich, K. (1990). The program FANTOM for energy refinement of polypeptides and proteins using a Newton–Raphson minimizer in torsion angle space. *Biopolymers*, **29**, 679–694.

Scholtz, J. M., York, E. J., Stewart, J. M. & Baldwin, R. L. (1991). A water-soluble, α-helical peptide: the effect of ionic strength on the helix–coil equilibrium. *J. Amer. Chem. Soc.* **113**, 5102–5104.

Sharp, K. A., Nicholls, A., Fine, R. F. & Honig, B. (1991). Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science*, **252**, 106–109.

Shin, J. K. & Jhon, M. S. (1991). High directional Monte Carlo procedure coupled with the temperature heating and annealing as a method to obtain the global energy minimum structure of polypeptides and proteins. *Biopolymers*, **31**, 177–185.

Simon, I., Glasser, L. & Scheraga, H. A. (1991). Calculation of protein conformation as an assembly of stable overlapping segments: application to bovine pancreatic trypsin inhibitor. *Proc. Nat. Acad. Sci., U.S.A.* **88**, 3661–3665.

Tanford, C. & Roxby, R. (1972). Interpretation of protein titration curves. Application to lysozyme. *Biochemistry*, **11**, 2192–2198.

Vajda, S. & Delisi, C. (1990). Determining minimum energy conformation of polypeptides by dynamic programming. *Biopolymers*, **29**, 1755–1772.

Vanderbilt, D. & Louie, S. G. (1984). A Monte Carlo simulated annealing approach to optimization over continuous variables. *J. Comp. Phys.* **56**, 259–271.

van Gunsteren, W. F. & Berendsen, H. J. C. (1990). Computer simulation of molecular dynamics methodology, applications and perspectives in chemistry. *Angew. Chem. Int. Ed. Engl.* **29**, 992–1023.

Vásquez, M. & Scheraga, H. A. (1985). Use of buildup and energy-minimization procedures to compute low-energy structures of the backbone of enkephalin. *Biopolymers*, **24**, 1437–1447.

Vila, J., Williams, R. L., Vásquez, M. & Scheraga, H. A. (1991). Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor. *Proteins: Struct. Funct. Genet.* **10**, 199–218.

Weber, I. T. & Steitz, T. A. (1987). Structure of a complex of catabolite gene activator protein and cyclic AMP refined at 2·5 Å resolution. *J. Mol. Biol.* **198**, 311–326.

Wesson, L. & Eisenberg, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.* **1**, 227–235.

Williams, R. L., Vila, J., Perrot, G. & Scheraga, H. A. (1992). Empirical solvation models in the context of conformational energy searches: application to bovine pancreatic trypsin inhibitor. *Proteins: Struct. Funct. Genet.* **14**, 110–119.

Wilson, C. & Cui, W. (1990). Applications of simulated annealing to peptides. *Biopolymers*, **29**, 149–157.

Yang, A.-S., Sharp, K. A. & Honig, B. (1992). Analysis of the heat capacity dependence of protein folding. *J. Mol. Biol.* **227**, 889–900.

Edited by B. Honig

(Received 5 February 1993; accepted 7 September 1993)