PROTEINS: Structure, Function, and Genetics 4:19–30 (1988)

Criteria That Discriminate Between Native Proteins and Incorrectly Folded Models

J. Novotný,[1] A.A. Rashin,[2] and R.E. Bruccoleri[1]
[1]*Molecular & Cellular Research Laboratory, Massachusetts General Hospital & Harvard Medical School, Boston, Massachusetts 02114;* [2]*Department of Physiology & Biophysics, Mount Sinai School of Medicine, New York, New York 10029*

ABSTRACT **Various theoretical concepts, such as free energy potentials, electrostatic interaction potentials, atomic packing, solvent-exposed surface, and surface charge distribution, were tested for their ability to discriminate between native proteins and misfolded protein models. Misfolded models were constructed by introducing incorrect side chains onto polypeptide backbones: side chains of the α-helical hemerythrin were modeled on the β-sheeted backbone of immunoglobulin VL domain, whereas those of the VL domain were similarly modeled on the hemerythrin backbone. CONGEN, a conformational space sampling program, was used to construct the side chains, in contrast to the previous work,[1] where incorrect side chains were modeled in *all trans* conformations. Capability of the conformational search procedure to reproduce native conformations was gauged first by rebuilding (the correct) side chains in hemerythrin and the VL domain: constructs with r.m.s. differences from the x-ray side chains 2.2–2.4 Å were produced, and many calculated conformations matched the native ones quite well. Incorrectly folded models were then constructed by the same conformational protocol applied to incorrect amino acid sequences. All CONGEN constructs, both correctly and incorrectly folded, were characterized by exceptionally small molecular surfaces and low potential energies. Surface charge density, atomic packing, and Coulomb formula-based electrostatic interactions of the misfolded structures and the correctly folded proteins were similar, and therefore of little interest for diagnosing incorrect folds. The following criteria clearly favored the native structures over the misfolded ones: 1) solvent-exposed side-chain nonpolar surface, 2) number of buried ionizable groups, and 3) empirical free energy functions that incorporate solvent effects.**

Key words: Conformation search, CONGEN, misfolded structures, solvent-modified potentials, protein folding

INTRODUCTION

Protein folding is one of the major unsolved problems of present-day biochemistry. The laws of folding embody a set of rules that describe how individual side chains of an amino acid sequence determine backbone conformation in the native, folded proteins. Knowledge of these rules, among other things, would provide guidelines to homologous modeling, that is, to the derivation of three-dimensional structures from known amino acid sequences, based on atomic coordinates of homologous proteins (see, e.g., the recent review of Blundell et al.[2]).

Side chains in proteins show strong preferences for either *trans* (180°) or *gauche* (±60°)[3–8] conformations. Some insight into the ways side chains determine native structures has been gained by studying computer-generated structures in which the side chains were purposefully misplaced. Novotný et al.[1] constructed two incorrectly folded protein models: seaworm hemerythrin and the variable domain of mouse immunoglobulin K chain. These proteins have no sequence homology. The former is composed of a bundle of four α-helices, and the latter consists of two 4-stranded β-sheets (Fig. 1). With the program CHARMM, hemerythrin side chains were substituted into the immunoglobulin domain and vice versa. The structures were energy-minimized and compared with the correctly folded forms. It was found that the incorrect side chains can be incorporated readily into both types of structures with only small structural adjustments. Empirical potential energy values and atomic packing of the misfolded models were comparable to those of the correct structures, although Zehfus and Rose[9] found the misfolded structures to be less compact than average native proteins, as indicated by their coefficient of compactness. The incorrectly folded structures possessed distinctly larger solvent-accessible surfaces, with a greater fraction of nonpolar side-chain atoms exposed to solvent. Bryant and Amzel[10] pointed out that as a consequence of an excessive exposure of nonpolar atoms in misfolded structures, these models make only about half the number of hydrophobic contacts among their atoms, compared with the contacts of correctly folded proteins. Electrostatic and van der Waals energy terms calculated by modified equations that included an approximate representation of solvent effects[1] showed significant dif-

Received March 14, 1988; revision accepted May 2, 1988.

Address reprint requests to Dr. Jiří Novotný, Molecular and Cellular Research Laboratory, Massachusetts General Hospital, Fruit Street, Boston, MA 02114.

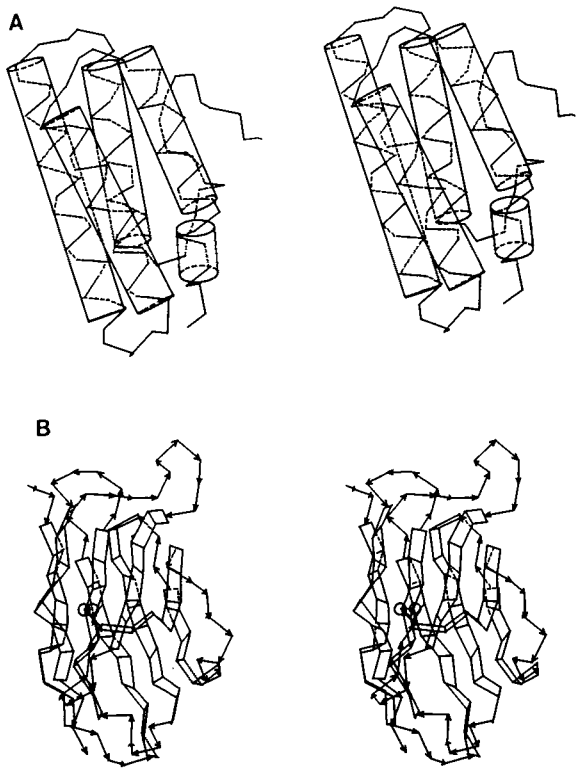

Fig. 1. Schematic stereodiagram of three-dimensional structures of hemerythrin (**A**) and the immunoglobulin VL domain (**B**). Cα backbone atoms are connected by virtual bonds; ribbons indicate β-strands and cylinders indicate α-helices. Cystine is schematically represented as balls-and-sticks. The diagram was prepared with the computer program of Lesk and Hardman.[31]

ferences between the correctly and incorrectly folded structures: the total solvent-modified, noncovalent potential energies of the two misfolded models were 70 kcal/mol and 273 kcal/mol, respectively, whereas those of the correctly folded ones were −117 kcal/mol and −77 kcal/mol, respectively. Eisenberg & McLachlan[11] used empirically derived atomic solvation parameters to estimate stabilities of the incorrectly folded models, and reported large free-energy differences between the models and their correctly folded counterparts; the net stabilization free energies of correct structures, $\Delta G_{s,}$ amounted to −34 kcal/mol and −17 kcal/mol, respectively.

Although the misfolded models were shown to possess aberrant molecular properties, some of the aberrations clearly were due to the arbitrary nature of the modeling process. For the ease of modeling, side chains were placed on inproper backbones in all *trans* (E) conformations, and adjustments of solvent-exposed side chains changed significantly solvent accessible surfaces of the models.[1] In the present study, we used the conformational space search procedure CONGEN[12] to place side chains into their optimal, energetically preferred conformations, and we report here on properties of these new models. We show that the newly generated misfolded structures have smaller solvent-exposed surfaces and better noncovalent energies than the previous models generated "by hand." Nevertheless, the new models still have excessive nonpolar solvent-exposed side-chain surfaces and can be distinguished from the correctly folded proteins by empirical energy functions that incorporate approximate solvent effects.[1,11,14]

METHODS OF COMPUTATION

X-Ray Crystallographic Coordinates

Atomic coordinates of mouse myeloma McPC 603 light chain variable region (N-terminal 113 residues) were obtained from Dr. David Davies; those for *Themiste dyscritum* hemerythrin, from the Brookhaven Data Bank.[13]

Model-Building the Incorrect Structures

Protein structures were built in the computer with aliphatic hydrogens combined with their heavy atoms into "extended atoms," whereas hydrogens serving as potential hydrogen-bond donors were treated ex-

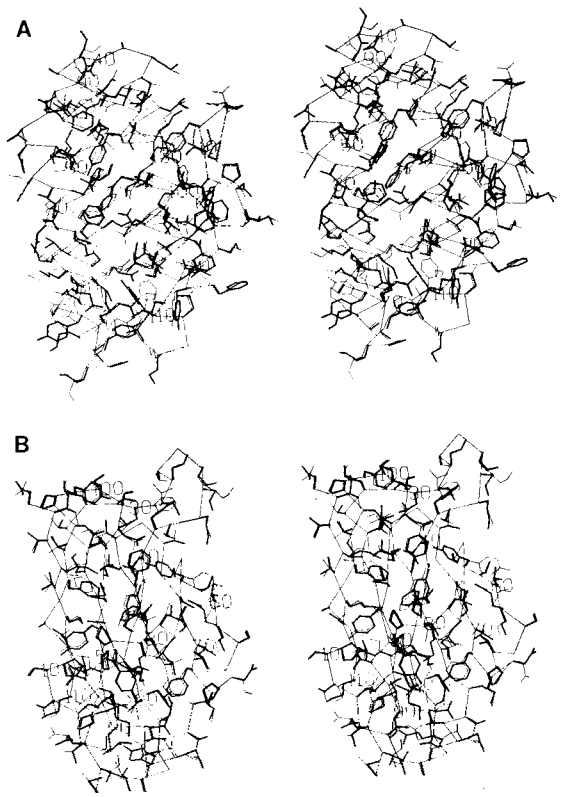

Fig. 2. CONGEN-reconstructed hemerythrin (**A**) and VL domain (**B**). Light lines depict the original crystallographic structures. Cα backbone atoms are connected by virtual bonds, and all the side-chain nonhydrogen atoms are drawn. Heavy lines show side-chain positions after CONGEN reconstruction. The r.m.s. shift between the native side chains and CONGEN-placed side chains was 2.4 Å for hemerythrin, and 2.2 Å for the VL domain.

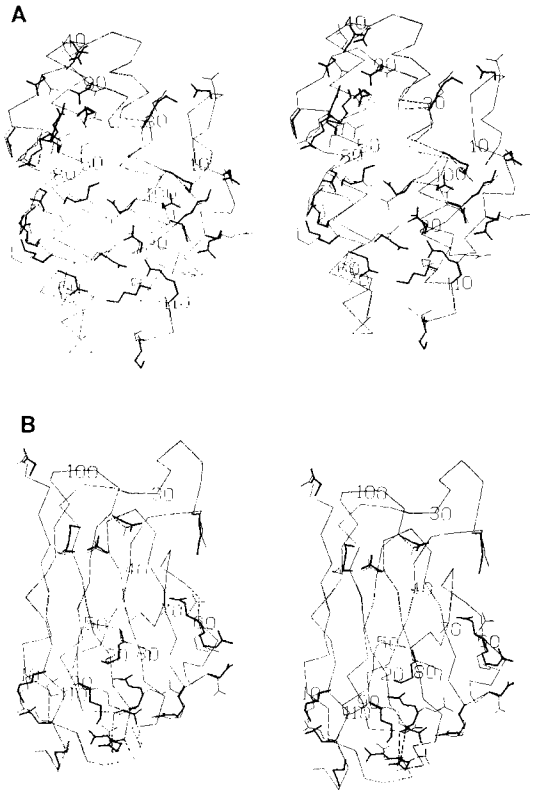

Fig. 3. Comparison of formally charged residues, arginine, lysine, glutamate, and aspartate, in native and CONGEN-reconstructed hemerythrin (**A**) and the VL domain (**B**). Light lines depict the Cα backbone and native side chains; heavy lines denote side chains after CONGEN reconstruction.

plicitly. Coordinates for all the backbone atoms were copied directly from the original set.

Missing side-chain atoms were introduced with the side-chain-construction protocol of the conformational search program CONGEN.[12] The protocol performs an iterative, nested search over all side-chain torsion angles χ_i. The search grid is adjustable, with the coarsest grid option sampling merely three points, ±60° and 180° (i.e., *gauche* and *trans* conformations) of the torsional space. Based on previous experience with the program, a finer search grid was employed, namely 60° for arginine and lysine side chains and 30° for all the other side chains. The side-chain placement protocol proceeded as follows. First, the starting side chain was determined, as described below. Then, a conformation of this side chain was found that avoided van der Waals repulsions greater than 20 kcal/mol. The complete conformational space of each other side chain was thereafter examined, and the lowest energy configurations were retained. The coordinates of the lowest-energy configuration of the whole chain then replaced the previous atomic coordinates. Finally, the search protocol was repeated, beginning with the same starting side chain, until either: 1) the total energy of the structure remained unchanged after a new search cycle, or 2) the maximum number of cycles (typically 100) was exceeded.

In this search procedure, potential energies are computed and compared at each grid position, and for every torsion angle. Thus, all atoms of the system in turn, including those introduced by the previous search operations, influence the search. Although the protocol cannot guarantee finding the lowest possible energy of the polypeptide chain, it nevertheless discards energetically unfavorable options. For example, if all side-chain conformations turn out to have large van der Waals repulsions, the search fails. The order in which the side chains are searched influences the final results. We therefore compared two different construction protocols as described in the section Test of CONGEN procedure, reserving other possible search schemes for future tests.

Energy Minimization and Structure Manipulations

Energy evaluations and manipulations were performed with the program CONGEN, which incorporates many features of the program CHARMM.[15] The empirical potential-energy function used in this program is a sum of atomic covalent energy terms for bond lengths E_b, bond angles E_θ, dihedral angles E_ϕ, and improper torsion angles E_ω (designed to maintain chirality or planarity around certain atoms): $E_b = \Sigma k_b (r - r_0)^2$; $E_\theta = \Sigma k_\theta (\theta - \theta_0)^2$; $E_\phi = \Sigma k_\phi - k_\phi \cos(n\phi)$; $E_\omega = \Sigma k_\omega (\omega - \omega_0)^2$ (in all the terms, the *k*s denote energy constants, and the zero-subscripted values denote geometric constants of optimal bond length r, bond angle θ, torsion angle ϕ, and improper dihedrals, ω) and noncovalent energy terms for van der Waals E_{VDW}, hydrogen bond E_{HB}, and electrostatic E_{EL} interactions:

$$E_{VDW} = \Sigma\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right);$$

$$E_{HB} = \Sigma\left(\frac{A'}{r_{AD}^{10}} - \frac{B'}{r_{AD}^{12}}\right)\cos^4(\theta_{A\text{-}H\text{-}D});$$

$$E_{EL} = \Sigma\frac{Q_iQ_j}{4\pi\epsilon r_{ij}}$$

(A, B, A′, B′, π constants; r_{ij} is distance between the *i* th and *j* th atoms, respectively, and the subscripts AD and A-H-D relate to hydrogen acceptor-donor distance and hydrogen acceptor-hydrogen–hydrogen-donor angle). The constants and parameters (e.g., the partial atomic charges) used in evaluations of these formulas were those of CHARMM version 16, as described.[15]

The structures were minimized with the Adopted Basis set Newton-Raphson (ABNR) minimization, a pseudosecond derivative method.[15] Two hundred cycles of minimization were applied, with hydrogen bond and nonbonded lists updated every 25 steps. The hydrogen bond lists were generated with distance and angle cutoffs of 4.5 Å and 90°, respectively. For the sake of computational efficiency, the noncovalent interactions were limited by a cutoff distance of 8 Å; a distance-dependent dielectric constant was used.[16]

Solvent Accessibility

Accessibility values[17] were computed with a probe radius of 1.4 Å and a fractional error of z coordinate sectioning of 0.05 Å. Atomic radii used were those specified in the parameter input file of the CONGEN program and were identical with those used in the CHARMM program.[15]

Approximate Free Energy Evaluations

We used three different free-energy approximations to estimate the stability of incorrectly folded models.

1. Our previous solvent-modified potential[1] introduced solvent screening into the Coulomb potential, and represented the hydrophobic effect by modifying the van der Waals term of solvent-exposed nonpolar atoms. The solvent-screened electrostatics, as originally used by Northrup et al.,[18] is computed with charges on the atoms multiplied by a factor depending on the ratio of the distance of a given charge from the center of the molecule to the distance from the center of the closest surface atom. The resulting dielectric screening factor decreases the effective charges linearly from a value of unity in the protein center to 0.3 on the protein surface. In our hydrophobic approximation, we considered the repulsive (i.e., steric hindrance) and attractive van der Waals terms separately, classifying nonpolar atoms into interior and surface types. Energetic cost of the exposed nonpolar atoms arising from hydrophobic solvation effects was introduced as an attractive-term modification; that is, the attractive van der Waals interaction for the ij th pair of atoms, $-(\frac{B}{r_{ij}^{-6}})$, was omitted for all the carbon atoms other than carbonyl and for sulphur atoms, whenever either one of them had a solvent-accessible surface in excess of 0.1 $Å^2$. Thus, all pairs with one atom even slightly exposed to solvent have their interaction modified, guaranteeing that the whole surface layer of the molecule is included.

2. In the approximate free-energy treatment of Rashin,[14] protein stability is proportional to the empirical formula $\Delta G_s - T\Delta S_c + T\Delta S_d$, where G_s stands for the buried surface contribution and is equal to $0.022 \times$ *buried surface* [kcal/mol] and T is temperature (degrees Kelvin). S_c, the conformational entropy, is estimated from the molecular weight, Mw, as follows: $S_c = (-60.64 + 0.01612 \times Mw)/T$ for molecular masses $>10{,}375$; $S_c = (-16.96 + 0.02341 \times Mw)/T$ for molecular masses in the range 2,326–10,375; $S_c = 0.01612 \times Mw/T$ for molecular masses $<2{,}326$. Finally, S_d is entropy change resulting from the presence of disulfide bonds. Numerical constants of these formulas were obtained from analysis of crystallographic structures of proteins with known stabilities.[14]

3. Protein solvation energy of Eisenberg and McLachlan[11] considers the free energy of interaction of a solute with water as a sum of energies of atomic groups; solvent accessible surface areas[17] are used as a measure of atomic interactions with solvent. The sign and strength of the water-solute interaction are specified by the atomic solvation parameter as described,[11] and the atomic contributions are summed to obtain the total solvation energy of a protein.

Internal Cavities

Cavities were evaluated as described previously.[19] Briefly, the Shrake and Rupley[20] algorithm was used to generate points representing protein surface. A probe of radius 1.2 Å was used, together with the following atomic radii: tetrahedral C, N or S with hydrogen atoms attached, 2.0 Å; trigonal C and NH, 1.7 Å; trigonal CH, CH^2 and S, 1.85 Å; OH and trigonal N, 1.5 Å; trigonal NH_2, 1.8 Å; O and water, 1.4 Å.[19] Examination of connectivity of the computed points, based on a distance criterion, allows the points to be segregated into separate sets. Any such set is connected only to itself. The largest set obtained represents the outer surface of the protein, whereas the other sets represent cavities.

Results of atomic-volume calculations depend to some degree on the algorithm used and the probe radius employed to define the surface. For example, Connolly[28] reported essentially no packing defects in 20 protein structures of high resolution, including the immunoglobulin VL domain RHE, whereas Rashin et al.[19] found cavities of variable sizes, including most of those that contain crystallographically detected water molecules.

Environment of Buried Ionizable Groups

Buried, formally charged side chains (lysine, arginine, histidine, aspartate, glutamate) were analyzed as described by Rashin and Honig.[27] Interatomic distance criteria were used to identify putative salt bridges and hydrogen bonds. The "hard" distance criterion required the distance between donor and acceptor atoms to be 3.5 Å or less; the "soft" criterion extended this distance to 4 Å for hydrogen bonds and 5.5 Å for salt bridges. Three different classes of buried ionizable groups were defined, based on solvent accessibility of their polar atoms (see Table VIII).

TEST OF CONGEN PROCEDURE: SIDE-CHAIN RECONSTRUCTION ON NATIVE PROTEINS

Before the attempt was made to model incorrectly folded structures, the side-chain-placement procedure was tested on native structures. Side chains of hemerythrin and the VL domain were reconstructed onto their backbones with the CONGEN program and compared with side-chain orientations in the native proteins. Two different protocols of side-chain placement were tested. In the first, side chains were introduced on the backbone, starting from the N-terminus and continuing sequentially along the chain. In the

TABLE I. Root-Mean-Square Shifts of CONGEN-Constructed Side Chains From Their Original Positions in Hemerythrin*

Residue	Sequence no.	R.M.S.
PHE	2	4.0
ILE	4	3.1
ASP	6	3.7
TYR	8	6.5
CYS	9	0.1
TRP	10	0.3
ASP	11	2.5
ILE	12	1.1
SER	13	0.3
PHE	14	0.7
ARG	15	3.8
THR	16	2.4
PHE	17	0.4
TYR	18	0.6
THR	19	2.4
ILE	20	1.2
VAL	21	2.4
ASP	22	0.3
ASP	23	1.0
GLU	24	3.0
HIS	25	1.6
LYS	26	1.1
THR	27	0.5
LEU	28	0.2
PHE	29	0.6
ASN	30	0.3
ILE	32	0.9
LEU	33	1.5
LEU	34	0.2
LEU	35	2.3
SER	36	2.4
GLN	37	0.6
ASP	39	0.8
ASN	40	0.3
ASP	42	1.9
HIS	43	1.6
LEU	44	2.0
ASN	45	3.0
GLU	46	1.2
LEU	47	0.4
ARG	48	1.4
ARG	49	3.5
CYS	50	0.1
THR	51	2.3
LYS	53	2.1
HIS	54	0.4
PHE	55	0.8
LEU	56	2.9
ASN	57	1.8
GLU	58	2.4
GLN	59	3.2
GLN	60	1.3
LEU	61	2.1
MET	62	3.0
GLN	63	3.7
SER	65	2.2
GLN	66	3.0
TYR	67	2.2
TYR	70	4.6
GLU	72	2.6
HIS	73	1.4
LYS	74	2.4
LYS	75	4.6
HIS	77	1.7
ASP	78	1.2
ASP	79	3.1
PHE	80	4.6
ILE	81	1.7
HIS	82	4.6
LYS	83	4.6
LEU	84	2.0
ASP	85	0.1
THR	86	0.1
TRP	87	0.6
ASP	88	1.7
ASP	90	2.4
VAL	91	2.6
THR	92	2.5
TYR	93	4.8
LYS	95	2.3
ASN	96	1.4
TRP	97	5.4
LEU	98	0.2
VAL	99	0.1
ASN	100	2.8
HIS	101	1.7
ILE	102	0.2
LYS	103	3.6
THR	104	0.1
ILE	105	3.3
ASP	106	1.3
PHE	107	2.6
LYS	108	1.5
TYR	109	5.4
ARG	110	5.8
LYS	112	5.4
ILE	113	1.2

*The r.m.s. values were computed on all atoms other than the fixed backbone, i.e., other than N, Cα, C, O and Cβ. Glycine, alanine, and proline residues were not reconstructed and are not included in the r.m.s. computation.

second, the protein core was rebuilt before the protein surface. Amino acid residues were rank ordered according to their proximity to the center of mass, and side chains were sequentially introduced in that order.

Figure 2 shows results of the sequential side-chain-placement protocol, which gave better results for both hemerythrin and the VL domain, as judged by the r.m.s. shifts between side-chain positions in the native and the reconstructed proteins (2.4 Å and 2.2 Å in the case of hemerythrin and the VL domain, respectively, compared with 2.8 Å and 2.5 Å, respectively, obtained with the center-of-mass rebuilding protocol). Many of the side chains in the reconstructed proteins match the native conformations quite well (Tables I, II). This correspondence holds true particularly for the buried hydrophobic side chains, which make the core of the proteins. The 20 side chains of the VL domain that are buried (that is, have solvent

TABLE II. Root-Mean-Square Shifts of CONGEN-Constructed Side Chains From Their Original Positions in the VL Domain*

Residue	Sequence no.	R.M.S.
ASP	1	1.2
ILE	2	3.5
VAL	3	0.1
MET	4	1.5
THR	5	0.1
GLN	6	0.2
SER	7	2.3
SER	9	0.0
SER	10	2.4
LEU	11	2.9
SER	12	2.1
VAL	13	0.3
SER	14	2.3
GLU	17	2.9
ARG	18	6.1
VAL	19	0.2
THR	20	1.9
MET	21	2.1
SER	22	2.5
CYS	23	0.0
LYS	24	1.0
SER	25	2.4
SER	26	1.7
GLN	27	0.8
SER	28	0.1
LEU	29	1.8
LEU	30	3.3
ASN	31	1.8
SER	32	2.5
ASN	34	0.3
GLN	35	5.0
LYS	36	1.0
ASN	37	0.7
PHE	38	2.0
LEU	39	0.2
TRP	41	0.6
TYR	42	0.2
GLN	43	1.6
GLN	44	3.9
LYS	45	3.0
GLN	48	1.0
LYS	51	4.8
LEU	52	3.5
LEU	53	3.1
ILE	54	0.2
TYR	55	6.7
SER	58	0.0
THR	59	0.7
ARG	60	2.0
GLU	61	3.3
SER	62	0.1
VAL	64	2.5
ASP	66	3.4
ARG	67	1.7
PHE	68	0.4
THR	69	0.2
SER	71	0.1
SER	73	0.2
THR	75	2.5
ASP	76	1.7
PHE	77	0.4
THR	78	0.2
LEU	79	2.1
THR	80	0.1
ILE	81	0.9
SER	82	0.3
SER	83	2.3
VAL	84	0.1
GLN	85	1.6
GLU	87	5.2
ASP	88	1.6
LEU	89	0.5
VAL	91	0.6
TYR	92	1.7
TYR	93	2.0
CYS	94	0.0
GLN	95	1.6
ASN	96	3.8
ASP	97	1.8
HIS	98	4.5
SER	99	2.3
TYR	100	1.8
LEU	102	2.1
THR	103	2.2
PHE	104	0.5
THR	108	2.3
LYS	109	1.4
LEU	110	2.7
GLU	111	2.4
ILE	112	2.1
LYS	113	4.9

*The r.m.s. values were computed on all atoms other than the fixed backbone, i.e., other than N, Cα, C, O and Cβ. Glycine, alanine, and proline residues were not reconstructed and are not included in the r.m.s. computation; cysteine residues 23 and 94, involved in a disulfide bond, were not reconstructed.

accessible surfaces less than 5 Å^2) show an r.m.s from the crystal structure of 1.6 Å. The 29 side chains buried in hemerythrin show an r.m.s 2.3 Å^2.

Some of the surface side chains with greater conformational freedom show large deviations from the positions occupied in the crystal. These often involve formally charged residues of arginine, lysine, aspartate, and glutamate, and their rearrangements resulted in remarkable improvement of electrostatic interactions (cf. Table III, Fig. 3). For example, in the case of the VL domain, there is a 100% increase in stabilizing electrostatic energy, as computed in vacuo with the distance-dependent dielectric constant. In native structures, electrostatic interactions on the surface of the protein would be efficiently screened by solvent, and the large electrostatic stabilization observed on reconstruction with an in vacuo potential is likely to be ficticious. The unsatisfactory nature of methods using approximate in vacuo models was further indicated by pilot CONGEN computations employing constant dielectrics with $\epsilon = 50$,[22] which

TABLE III. Empirical Potential Energies (kcal/mol) and Solvent-Accessible Surfaces [Å^2] of CONGEN-Reconstructed Hemerythrin and VL Domain

		CONGEN-hemerythrin			CONGEN-VL domain	
Energy term	Hemerythrin	Constructed	Energy-minimized	VL domain	Constructed	Energy-minimized
Bond lengths	21	18	18	21	19	19
Bond angles	151	126	156	155	130	148
Torsion angles	111	162	128	178	201	151
Improper torsions	25	20	26	22	19	24
van der Waals forces	−1,098	−874	−1,112	−836	−706	−841
Electrostatics	−432	−497	−755	−246	−305	−491
Hydrogen bonds	−361	−325	−411	−266	−249	−364
Total energy	−1,578	−1,369	−1,950	−965	−891	−1,354
r.m.s. shift [Å]			0.5			0.6
Surface	6,165		5,670	5,880		5,473

TABLE IV. Empirical Potential Energies (kcal/mol) and Solvent-Accessible Surfaces (Å^2) of the *all-trans*[1] and CONGEN-Constructed, Incorrectly Folded Hemerythrin and VL Domain

	VL-like hemerythrin		Hemerythrin-like VL domain	
Energy term	*all-trans*	CONGEN	*all-trans*	CONGEN
Bond lengths	23	22	20	18
Bond angles	185	189	128	129
Torsion angles	166	167	91	103
Improper torsions	30	30	24	25
van der Waals forces	−916	−981	−761	−776
Electrostatics	−199	−588	−258	−442
Hydrogen bonds	−235	−302	−335	−393
Total energy	−942	−1,461	−1,088	−1,335
Surface	7,349	6,234	7,120	6,421

produced models with larger r.m.s. deviations from the original structures than those obtained with the ϵ = R protocol.

The quality of the sequentially constructed models was further assessed by comparing their empirical potential energy and molecular surfaces to those of native proteins. Both models compared favorably with native structures, and their empirical potential energy further improved with energy minimization (Table III). A large portion of this additional stabilization, however, comes from attractive electrostatic interactions among CONGEN-selected side-chain conformations (Fig. 3), as based on the ϵ = R electrostatic model computed for pairwise interactions within a cutoff distance of 8 Å. The solvent-accessible surface of the energy-minimized models was somewhat smaller than that of the native, energy-minimized structures. Although the significance of this phenomenon is unclear, it once again may relate to the absence of solvent and to a concomitant neglect of attractive solvent-solute interactions during model construction and energy minimization.

In summary, the results presented in Figs. 2, 3 and Tables I–III make it clear that uniform conformational sampling, as implemented in the program CONGEN, is able to produce protein models with side chains placed at positions of major (perhaps global) energetic minima, as defined by the empirical potential used to evaluate potential energies of sampled conformations. The agreement of many CONGEN-constructed conformations with the native ones is encouraging. At the same time, the use of more realistic potential functions that incorporate solvent-screened noncovalent interactions (see, e.g., ref. 21) is needed to improve on the accuracy of placement of polar and charged side chains.

CONGEN CONSTRUCTION OF INCORRECTLY FOLDED MODELS

Figure 4 compares the CONGEN-constructed incorrectly folded models with the original, *all trans* misfolded structures.[1] Empirical potential energies of energy-minimized models show a substantial improvement in noncovalent interactions as a result of CONGEN-modeling, most significantly in electrostatic energies (Table IV). Visually, the CONGEN-generated models appear to have smoother surfaces,

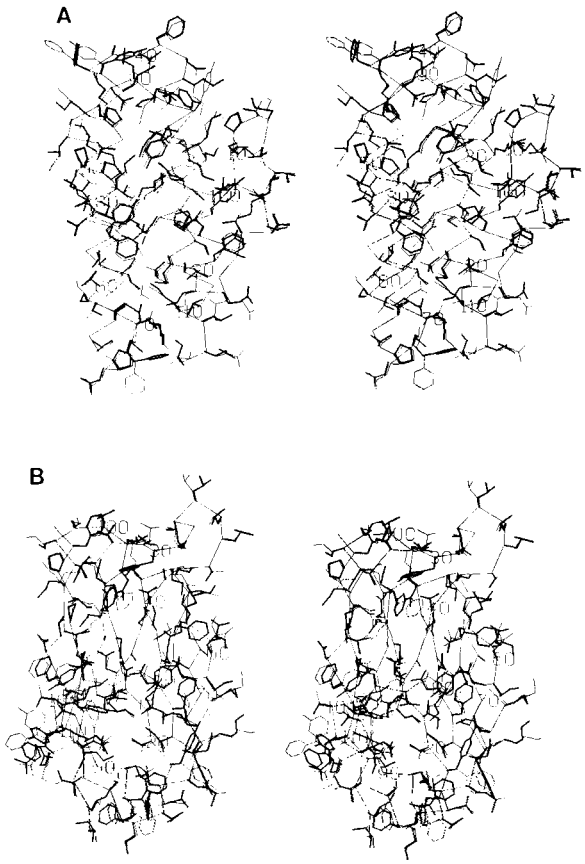

Fig. 4. Incorrectly folded hemerythrin (**A**) and VL domain (**B**), as constructed by the CONGEN program. Cα backbone atoms are connected by virtual bonds, and all the side-chain nonhydrogen atoms are drawn. Heavy lines show side-chain positions after CONGEN reconstruction.

a feature that is confirmed by solvent-accessibility computations (Table IV). It has been shown[23–25] that the solvent-accessible surface, S, of small, single-domain proteins is a simple function of their molecular weight, Mw: $S = 11.1 \times Mw^{2/3}$. The computed surface of the CONGEN-misfolded VL domain, 6,421 Å^2, is about 9% larger than the value 5,897 Å^2 expected for the VL domain based on its molecular weight. However, the molecular surface of the CONGEN-misfolded hemerythrin, 6,234 Å^2, is virtually identical to its theoretical value of 6,232 Å^2. A similar overall trend is observed in molecular surfaces of CONGEN-reconstructed crystallographic structures (Table III), which are 7–8% smaller than the native molecular surfaces.

The fact that it is possible to construct grossly incorrect models with apparently correct total solvent-accessible surfaces is striking and was not anticipated by our previous study.[1] The important question of molecular surfaces, as well as other structural features potentially diagnostic of incorrect folds, is discussed further in the following paragraphs.

FRACTIONAL MOLECULAR SURFACES

Data assembled in Table V indicate that solvent-exposed *side chain* nonpolar surfaces of the CONGEN-constructed incorrectly folded models are still significantly larger (i.e., 15% larger for VL-like hemerythrin and 18% larger for hemerythrinlike VL domain) than those of the native structures. In fact, the ratios of nonpolar/polar solvent-exposed side-chain surfaces, already exceptionally high in the original misfolded models, are even higher in the new models, namely, 3.37 for CONGEN-misfolded hemerythrin, compared with 1.69 for correctly folded hemerythrin; 2.73 for CONGEN-misfolded VL domain, compared with 1.74 for the correctly folded domain. As much as the CONGEN procedure samples the whole conformational space of the side chains, we can conclude that the smaller total surfaces of incorrectly folded models can be obtained only at the expense of exposing more nonpolar side-chain atoms. This conclusion in turn suggests that in properly folded structures, the side-chain nonpolar surface that remains exposed to solvent is the smallest possible one.

A direct numerical measure of burial of the nonpolar atoms is the nonpolar/polar side-chain surface ratio, with values of about 2.0–2.2 and greater indicating incorrect folding in the case of hemerythrin and the VL domain. Limited pilot calculations carried out on other proteins seem to suggest that the "critical threshold" value of the side-chain surface ratio varies somewhat from case to case. For example, in oligomeric hemoglobin α and β chains, the limiting value appears to be approximately 2.5.

APPROXIMATE FREE ENERGY COMPUTATIONS

Hydrophobic energetic penalties incorporated into solvent-modified empirical free-energy functions represent another way of gauging the ratio of solvent-exposed nonpolar atoms and consequently the correctness of protein models. Table VI lists stabilities of correctly and incorrectly folded structures, respectively, computed according to Novotný et al.,[1] Rashin,[14] and Eisenberg and McLachlan.[11] In all these treatments, the misfolded structures, compared to their correctly folded counterparts, are destabilized by large energy differences. The difference in computed energy values, *correctly folded structure − misfolded structure* can be taken as a measure of the free energy associated with the virtual reaction *incorrect fold → correct fold*. Thus, the Eisenberg-McLachlan approach gives net correct structure stabilization of −30 kcal/mol for the VL domain and −29 for hemerythrin, whereas Rashin's approach gives net stabilization free energies −14.4 kcal/mol for the VL domain and −5.7 kcal/mol for hemerythrin. The net electrostatic and hydrophobic stabilization terms, computed according to Novotný et al., is −366 kcal/mol for the VL domain and −393 kcal/mol for hemerythrin. Incidentally, the solvent-modified van der Waals inter-

TABLE V. Solvent-Accessible Surfaces (Å^2) of Various Structures

	Hemerythrin				VL domain			
	Native	CONGEN rebuilt	Misfolded *all trans*	Misfolded CONGEN	Native	CONGEN rebuilt	Misfolded *all trans*	Misfolded CONGEN
Backbone	617	992	716	1,037	790	1,211	666	1,230
Side chains, polar atoms	2,084	1,391	2,072	1,188	1,878	1,517	1,822	1,391
Side chains, nonpolar atoms	3,484	3,287	4,561	4,009	3,212	2,745	4,632	3,800
Total surface	6,165	5,670	7,349	6,234	5,880	5,473	7,120	6,421

TABLE VI. Empirical Free Energies (kcal/mol) of Correctly and Incorrectly Folded Hemerythrin and VL Domain

	Solvent-modified noncovalent[1] interactions		Empirical free energy	
Structure	Electrostatic	van der Waals	Rashin	Eisenberg and McLachlan
Hemerythrin	−89	−28	−15	−113
CONGEN-misfolded hemerythrin	−99	12	−9	−84
VL domain	−68	−9	−17	−106
CONGEN-misfolded VL domain	−77	314	−3	−76

TABLE VII. Solvent-Accessible Electrically Charged Surfaces*

Protein	Total surface (Å^2)	Polar atoms Arg + Lys (Å^2)	Polar atoms Asp + Glu (Å^2)	Percent charged surface
Insulin monomer	1,150	45	78	10.7
Cytochrome†	1,958	298	139	22.3
Lysozyme	2,042	617	208	40.4
Myoglobin	2,480	374	304	27.4
Papain	2,929	414	130	18.5
Subtilisin	3,188	153	165	10.0
Thermolysin	3,860	319	238	14.5
Hemoglobin	7,439	801	710	20.3
Crambin‡	3,000	99	110	6.9
Anthaxanthin‡	5,464	125	292	7.6
VL domain	5,473	253	380	11.5
CONGEN-misfolded VL domain	6,421	232	568	12.4
Hemerythrin	5,670	267	574	14.8
CONGEN-misfolded hemerythrin	6,234	394	316	11.2

*Values computed with the Lee and Richards algorithm (1971), as described in Methods.
†Computed accessibilities include heme.
‡These proteins are not soluble in water.

action distinguishes the correctly and incorrectly folded structures by sign; that is, the correctly folded structures show negative, i.e., net attractive (stabilizing), terms, and the misfolded structures have positive, i.e., net repulsive (destabilizing) solvent-modified van der Waals interactions.

ELECTRICALLY CHARGED SURFACES AND BURIED CHARGED GROUPS

Barlow and Thornton[26] reported the average charge density of native proteins to be 1.4 charged groups per 100 Å^2 of protein surface. An alternative way of

TABLE VIII. Environment and Bonding Patterns of Buried Ionizable Groups in Native and Misfolded Structures

		Structure							
		VL domain				Hemerythrin			
Sequence	Group class*	No.	paired	S-bridge	H-bond	No.	paired	S-bridge	H-bond
VL domain	Glu total	4				4			
	I	—				1	1	1	1
	Ia	—				—			
	II	—				—			
	Asp total	5				5			
	I	—				—			
	Ia	1	1	1	—	—			
	II	—				—			
	Lys total	6				6			
	I	—				2	2	2	2
	Ia	—				—			
	II	—				—			
	Arg total	3				3			
	I	—				—			
	Ia	1	1	1	1	—			
	II	—				—			
	His total	1				1			
	I	—				1	1	0	1
	Ia	—				—			
	II	1	1	—	1	—			
Hemerythrin	Glu total	4				4			
	I	—				1	1	1	1
	Ia	2	2	2	2	1	1	1	1
	II	1	1	(1)	1	—			
	Asp total	12				12			
	I	4	4	4	2(4)	1	1	1	1
	Ia	3	3	3	1(2)	1	1	(1)	1
	II	1	1	1	1	—			
	Lys total	9				9			
	I	3	3	3	2(3)	—			
	Ia	—				—			
	II	2	2	1(2)	2	—			
	Arg total	4				—			
	I	—				—			
	Ia	1	1	1	1	—			
	II	—				—			
	His total	7				7			
	I	2	2	1	2	5	5	5	2(3)
	Ia	2	2	(1)	2	—			
	II	—				—			

*Class I: polar atoms of the side chain unaccessible to water; Ia: one polar atom unaccessible, another accessible; II: polar atoms are accessible but less than 5 Å^2. Numbers in parentheses correspond to the "soft" criteria for bonding (see Methods).

expressing surface charge density is to list percentage of solvent-exposed, formally charged nitrogen and oxygen atoms, as we do in Table VII. Paradoxically, CONGEN-misfolded models show a surface-charge density (~12%) comparable to that of its native counterparts (~13%) and exceeding the charge densities of some native soluble proteins, e.g., subtilisin (10%) or insulin (10.7%). Nevertheless, it appears that the incorrectly folded structures have more buried charged residues than is normal. Rashin and Honig[27] found that charged groups are only rarely buried (no more than two buried groups per structure on the average), and that they always become involved in hydrogen bonds or salt bridges. The misfolded VL domain, having four buried side chains, compared with the three marginally accessible groups in the native fold (Table VIII), is only slightly anomalous. Misfolded hemerythrin is clearly abnormal, however: nine ionizable groups become completely buried and an additional 12 have only marginal solvent exposure, whereas the native fold has seven buried groups, five of which are histidines, with only two marginally accessible groups. When 36 native protein structures were examined,[27] only five Nϵ atoms of lysine resi-

TABLE IX. Cavities in Native and Misfolded Structures

	Structure			
	Fab VL		Hemerythrin	
	Cavity volume Å^3		Cavity volume Å^3	
Sequence	Total	Empty*	Total	Empty*
Fab VL	94	70	82	60
Hemerythrin	203	100	134	100

*The cavity volume not filled with water and destabilizing the structure is calculated as described in Methods.

dues were found buried; this finding contrasts with that showing three lysines completely buried in misfolded hemerythrin.

It is reasonable to expect large destabilizing electrostatic effects to arise from burial of charged groups.[29] Such effects, however, are difficult to quantitate by empirical potential-energy calculations and require more detailed approaches.[21,29-30] As pointed out in the accompanying paper by Gilson and Honig,[30] the calculated ion-solvent interactions are good indicators of correctness of the protein structure. Unfortunately, direct comparisons between the values obtained from the macroscopic screening models[1,29] and those reported by Gilson and Honig[30] are difficult to make because of differences in such technical details as different positions of reconstructed polar hydrogen atoms, different partial atomic charge sets, etc.

ATOMIC PACKING

The quality of atomic packing in the CONGEN-constructed incorrectly folded models was analyzed indirectly, by computing volumes of internal cavities[19] in these models and then comparing the cavity volumes to those found in the correctly folded structures (Table IX). It has been noted[1] that different molecular weights of hemerythrin and the VL domain are reflected in their different molecular volumes, 15,679 Å^3 and 14,336 Å^3, respectively. On misfolding, the amino acid sequence of the VL domain accommodates easily into the larger volume of the hemerythrin fold, whereas the amino acid sequence of hemerythrin has to be packed into the volume of the VL domain that was originally 1,343 Å^3 smaller. Accordingly, summed internal cavity volume of the misfolded VL domain, 82 Å^3, is smaller than that found in its correct fold (94 Å^3), whereas the summed cavity volume of hemerythrin increases on misfolding from 134 Å^3 to 203 Å^3. In fact, the total size of internal cavities found in incorrectly folded structures does not deviate in any significant way from the sizes of cavities found in typical native proteins of comparable molecular weight.[19]

It also should be considered that polar cavities, when filled with hydrogen-bonded water molecules, add to the stability of protein structures. In order to account for this possibility, Rashin et al.[19] introduced the notion of empty-cavity volume, V_e, as $V_e = V - V_wN$ (V, cavity volume; V_w 30Å^3 , partial specific volume of water in the bulk phase). If empty-cavity volumes are considered, practically no differences are found between native and misfolded structures (Table IX). Therefore, it would seem that misfolded structures can be reasonably tightly packed and, contrary to the conclusions of Ponder and Richards,[8] packing constraints in proteins are not sufficient to eliminate the possibility that arbitrary sequences will adopt identical folds. A more detailed examination of this problem would be necessary, however, before general conclusions could be made with any certainty.

CONCLUSIONS

We have found that the uniform conformational space sampling procedure, CONGEN, is capable of placing side chains on improper backbones, thereby generating incorrectly folded structures with remarkable properties. Because the side chains are placed into positions corresponding to major energetic minima, the incorrectly folded models are characterized by large stabilizing empirical potential energies, comparable to those found in correctly folded structures. Similarly, total solvent-accessible surfaces, surface charge density, and atomic packing of the misfolded models compare favorably with those expected for native proteins. The only structural parameter found to be consistently anomalous in the incorrectly folded models was the solvent-exposed side chain nonpolar surface, with its concomitant ratio of side-chain nonpolar/polar solvent-accessible surfaces. This ratio, 1.69 and 1.74, respectively, for hemerythrin and the VL domain, is 2.20 and 2.50, respectively, for the incorrectly folded models previously constructed with side chains in *all trans* conformations,[1] and it becomes 2.73 and 3.37, respectively, in CONGEN-misfolded structures. Our study shows that the smaller total surfaces of incorrectly folded structures can be obtained only by exposing more nonpolar side-chain atoms to solvent, suggesting that the smallest solvent-accessible nonpolar surface of side chains is an important attribute of properly folded structures.

The approximate free-energy treatments that incorporate hydrophobic penalties for solvent-exposed nonpolar atoms[1,11,14] all indicate that incorrectly folded structures are destabilized by large energy differences, compared with the correctly folded structures.

The solvent-modified van der Waals force computations[1] yield net repulsive interactions in incorrectly folded structures and net attractive interactions in native proteins; they may be well suited for empirical gauging of correctness of protein models whose native, correctly folded counterparts are not known. On a relative scale, the number of buried ionizable groups may serve as a useful indicator, as dehydration of charged side chain groups is energetically costly.

ACKNOWLEDGMENTS

We are much indebted to David Eisenberg and Morgan Wesson for providing us with their values of free energies and to David Eisenberg, Michael Gilson, and Barry Honig for stimulating discussions and many helpful comments on the manuscript. We gratefully acknowledge the funding provided by the Office of Naval Research to Jiri Novotný and by the National Science Foundation to Alexander Rashin.

REFERENCES

1. Novotný, J. Bruccoleri R.E., Karplus, M. An analysis of incorrectly folded protein models. Implications for structure predictions. J. Mol. Biol. 177:787–818, 1984.
2. Blundell, T., Sibanda, B.L., Sternberg, M.J.E., Thornton, J.M. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326:347–352, 1987.
3. Janin, J., Wodak, S., Levitt, M., Maigret, B. Conformation of amino acid side chains in proteins. J. Mol. Biol. 125:357–386, 1978.
4. Bhat, T.N., Sasisekharan, V., Vijayan, M. Analysis of side chain conformation in proteins. Int. J. Peptide Protein Res. 13:170–184, 1979.
5. James, M.N.G., Sielecki, A.R. Structure and refinement of penicillopepsin at 1.89 Å resolution. J. Mol. Biol. 163:299–361, 1983.
6. Summers, N., Carlson, W., Karplus, M. Analysis of side chain orientations in homologous proteins. J. Mol. Biol. 196:175–198, 1987.
7. McGregor, M.J., Islam, S.A., Sternberg, M.J.E. Analysis of the relationship between side chain conformation and secondary structure in globular proteins. J. Mol. Biol. 198:295–310, 1987.
8. Ponder, J.W., Richards, F.M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.
9. Zehfus, M.H., Rose, G.D. Compact units in proteins. Biochemistry 25:5759–5765, 1986.
10. Bryant, S.H., Amzel, L.M. Correctly folded proteins make twice as many hydrophobic contacts. Int. J. Peptide Protein Res. 29:46–52, 1986.
11. Eisenberg, D., McLachlan, A.D. Solvation energy in protein folding and binding. Nature 319:199–203, 1986.
12. Bruccoleri, R.E., Karplus, M. Prediction of folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26:137–168, 1987.
13. 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.J. The protein databank: A computer-based archival file for macromolecular structure. J. Mol. Biol. 112:535–542, 1977.
14. Rashin, A.A. Buried surface area, conformational entropy and protein stability. Biopolymers 23:1605–1620, 1984.
15. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comp. Chem. 4:187–217, 1983.
16. Gelin, B., Karplus, M. Side chain torsional potentials and motion of amino acids in proteins: Bovine pancreatic trypsin inhibitor. Proc. Nat. Acad. Sci. USA 72:2002–2006, 1975.
17. Lee, B.K., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55:379–400, 1971.
18. Northrup, S.H., Pear, M.R., Morgan, J.D., McCammon, J.A., Karplus, M. Molecular dynamics of ferrocytochrome c. J. Mol. Biol. 153:1087–1109, 1981.
19. Rashin, A.A., Iofin, M., Honig, B. Internal cavities and buried waters in globular proteins. Biochemistry 25:3619–3625, 1986.
20. Shrake, A., Rupley, J.A. Environment and exposure to solvent of protein atoms: Lysozyme and insulin. J. Mol. Biol. 79:351–371, 1973.
21. Gilson, M.K., Honig, B.H. Energetics of charge-charge interactions in proteins. Proteins 3:32–52, 1988.
22. Warshel, A., Russell, T.S., Churg, A.K. Macroscopic models for studies of electrostatic interactions in proteins: Limitations and applicability. Proc. Nat. Acad. Sci. USA 81:4785–4789, 1984.
23. Chothia, C. Structural invariants of protein folding. Nature 254:304–308, 1975.
24. Janin, J. J. Surface area of globular proteins. Mol. Biol. 105:13–14, 1976.
25. Teller, D.C. Accessible area, packing volumes, and interaction surfaces of globular proteins. Nature 230:729–731, 1976.
26. Barlow, D.J., Thornton, J.M. The distribution of charged groups in proteins. Biopolymers 25:1717–1733, 1986.
27. Rashin, A.A., Honig, B.H. On the environment of ionizable groups in globular proteins. J. Mol. Biol. 173:515–521, 1984.
28. Connolly, M. Atomic size packing defects in proteins. Int. J. Peptide Protein Res. 28:360–363, 1986.
29. Rashin, A.A., Namboodiri, K. A simple method for the calculations of hydration enthalpies of polar molecules with arbitrary shapes. J. Phys. Chem. 91:6003–6012, 1987.
30. Gilson, M.K., Honig, B.H. The energetics of chargesolvent interactions in proteins: Hydration of small molecules, ion-binding, and conformational analysis. Proteins (submitted), 1988.
31. Lesk, A.M., Hardman, K.D. Computer-generated schematic diagrams of protein structures. Science 216:539–540, 1982.