PROTEINS: Structure, Function, and Genetics 22:81–99 (1995)

RESEARCH ARTICLES

LINUS: A Hierarchic Procedure to Predict the Fold of a Protein

Rajgopal Srinivasan and George D. Rose
Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205

ABSTRACT **We describe LINUS, a hierarchic procedure to predict the fold of a protein from its amino acid sequence alone. The algorithm, which has been implemented in a computer program, was applied to large, overlapping fragments from a diverse test set of 7 X-ray-elucidated proteins, with encouraging results. For all proteins but one, the overall fragment topology is well predicted, including both secondary and supersecondary structure. The algorithm was also applied to a molecule of unknown conformation, groES, in which X-ray structure determination is presently ongoing. LINUS is an acronym for *L*ocal *I*ndependently *N*ucleated *U*nits of *S*tructure. The procedure ascends the folding hierarchy in discrete stages, with concomitant accretion of structure at each step. The chain is represented by simplified geometry and folds under the influence of a primitive energy function. The only accurately described energetic quantity in this work is hard sphere repulsion—the principal force involved in organizing protein conformation [Richards, F. M. Ann. Rev. Biophys. Bioeng. 6:151–176, 1977]. Among other applications, the method is a natural tool for use in the human genome initiative.**

Key words: protein folding

INTRODUCTION

The protein folding problem was first formulated more than half a century ago by Mirsky and Pauling[1] and Wu.[2] Their early observations culminated in later experiments by Anfinsen and co-workers showing that ribonuclease can be denatured reversibly,[3] and leading to the contemporary view that protein conformation is determined solely by the amino acid sequence.

The central verity of protein folding is the existence of a unique native conformation, in which residues distant in sequence but proximate in space engender a close-packed core[4] enriched in hydrophobic residues.[5] How does the protein screen conformations effectively and select the native one reliably on a biological time-scale?[6] This search process is spontaneous for the protein but vexing for the protein folder.[7]

Current approaches to solving the folding problem can be classified into direct methods and template-based methods. Direct methods seek the native fold as a low(est) energy point in some suitably defined hyperspace of conformational possibilities. Template-based methods compare a sequence of unknown conformation against a library of solved structures and, using a suitable metric, score good matches as likely folds. Many papers on such approaches to three-dimensional structure have appeared in recent years.[8–26]

In this paper, we present a direct prediction method which capitalizes upon the observation that globular proteins are organized as a structural hierarchy.[27,28] The existence of hierarchic organization prompted an early model, termed *folding by hierarchic condensation,*[28] wherein neighboring chain sites interact to form primitive folding modules, which, in turn, further interact in iterative fashion, resulting in larger modules—supersecondary structure, domains, and ultimately whole protein monomers. In such a pathway, all folding events are local at an appropriate step in the self-assembly process. Similar ideas have been proposed more recently by Oas and Kim[29] and Dill et al.[30]

Consistent with folding by hierarchic condensation, protein secondary structure can be codified into only four categories—helix,[31] sheet,[32] turns,[33–35] and loops.[36,37] A complex fold can be decomposed into the elements contained within these categories,

Received March 19, 1995; accepted March 31, 1995.
Address reprint requests to George D. Rose, Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, 725 N. Wolfe Street, Baltimore, MD 21205.

together with their superstructures.[38,39] Such a model is supported by recent studies on the persistence of structure in isolated peptides[40–43] and the existence of equilibrium intermediates in the unfolding of cytochrome *c*.[44]

The algorithm described here, LINUS, is an implementation of our hierarchic folding model.[28,45] LINUS operates on a polypeptide chain of defined sequence, starting in a fully extended conformation. No other input information is required. Using idealized geometry and a highly simplified energy function, the chain is folded in hierarchic stages, with concomitant accretion of structure at each step.

The algorithm has been developed for use with monomeric, globular, water-soluble proteins. In this initial implementation, disulfide bonds and prosthetic groups have been ignored.

LINUS was applied to a diverse test set of 7 X-ray elucidated proteins and one additional molecule, groES, for which a structure has yet to be determined experimentally. The test set contains 2 all-α, 2 all-β, and 3-mixed proteins that have been investigated extensively in folding studies. In this paper, we have attempted to assess the algorithm's effectiveness in predicting structure from short and medium range interactions. Accordingly, entire molecules have not been attempted here. Instead, proteins were subdivided into overlapping 50-residue fragments, each of which was treated independently.

The algorithm predicts both secondary and supersecondary structure effectively, with two exceptions: (1) apocytochrome *c*, a protein that is known *not* to fold under conditions which favor the folding of the holo-protein,[46] and (2) dihydrofolate reductase, which folds incorrectly. In other cases the predicted conformation of the fragments matches the overall topology of their X-ray elucidated counterparts. Full details of the algorithm are given in the next section, followed by vignettes that compare the predicted and X-ray structures. Implications of the algorithm and future directions are then discussed.

METHODS

In brief, the algorithm accumulates favorable structure within a fixed, sequential interval of allowed interaction, then repeats this process in stages as the interval size increases. At each stage, the chain is allowed to move at random under the influence of an energy function. Hierarchy is established by recognizing favorable conformations in early stages and constraining them to persist during later stages. The move set was chosen to represent populated regions in ϕ,ψ-space: helix, sheet, and turns. Only three types of interaction are taken into account—one repulsive, the other two attractive. Repulsive interactions measure steric overlap; two nonbonded atoms cannot occupy the same space at the same time. Attractive interactions measure hydrogen bonds and the tendency of apolar residues to cluster.

In greater detail, the starting point of a simulation run is an amino acid sequence, S, where $S = s_1\text{–}s_2\text{–}s_3\text{–}\cdots\text{–}s_N$. Here, residues are represented by an indexed, lower case 's', with N residues in the sequence. The algorithm keeps track of two particular chain conformations: (1) the current conformation, C; and (2) a trial conformation, C*. At the onset, the interaction interval, Δ, is 6 residues in length, and the starting conformation is set to an extended backbone conformation ($\phi = -150°$, $\psi = 150°$, $\omega = 180°$), with the side chains pointed away from the main chain ($\chi = 180°$).

A simulation run proceeds as follows. Progressing sequentially (one residue at a time) from the N- to C-terminus, three consecutive residues are perturbed simultaneously to generate C*, a new trial conformation. C* is rejected if any two atoms in the entire sequence overlap. Otherwise, the energy of C* is computed by summing interactions between all residues separated in sequence by no more than the current interaction interval, Δ. Again, C* is rejected if a random number between 0 and 1 is less than the Boltzmann-weighted energy, $e^{-\beta\Delta E/T}$, where $\beta = 2$ and $T = 1$. If not rejected, then the trial conformation is accepted; i.e., the current conformation, C, is set to C*. A complete progression from N to C is termed a *cycle*. For each interaction interval, 6000 cycles are performed, consisting of 1000 equilibration steps followed by 5000 trial structures. Equilibration steps are discarded; trial structures are retained. Chain segments in the trial ensemble that adopt a persisting conformation throughout an interval are constrained to remain in that conformation during subsequent intervals.

In the work described next, protein monomers have been subdivided into overlapping 50-residue fragments, although the ultimate fragment at the C-terminus is allowed to exceed 50 residues. Each fragment is autonomous. The interval size is allowed to grow in 6-residue increments until $\Delta = 48$ is attained.

Fragmentation into autonomous 50-residue segments is arbitrary and can result in spurious end-effects for larger values of Δ because fragment termini can be recruited into non-native intrafragment interactions when deprived of native interactions with those residues that happen to partition into an adjacent fragment. Of course, this problem would not have occurred within the context of the whole protein. Surprisingly, the problem appears to be infrequent in our test set, except in the case of dihydrofolate reductase, where it is pronounced, resulting in a grossly incorrect structure.

A complete simulation can be represented diagramatically as:

$$C^{start} \Rightarrow \overset{N}{\underset{6}{\Delta}} \left(\overset{j}{\underset{1}{\text{cycle}}} \left(\overset{N-1}{\underset{2}{\text{step}}} \left\{ \begin{array}{l} \text{Generate Trial Conformation, C*} \\ \text{i. Randomly choose a} \\ \quad \text{backbone conformation} \\ \quad \downarrow \\ \text{ii. Bump check C*} \\ \quad \downarrow \\ \text{iii. Calculate energy of C*, } U(\text{C*}) \\ \quad \downarrow \\ \text{iv. If } U(\text{C*}) < U(\text{C}) \text{ or } x < e^{-\beta \Delta E}, \text{ where } x \text{ is} \\ \quad \text{random and } 0 \leq x \leq 1 \text{ then C} = \text{C*} \end{array} \right\} \right) \right)$$

where the innermost brackets surround the inner loop. The energy of the current conformation will decrease over the course of each interval and from each interval to the next. Upon completion, the minimum energy structure among the ensemble of conformations from the final interval is accepted as the predicted conformation.

The energy units used in these simulations, though not entirely arbitrary (see below), are crude, and a Monte Carlo-motivated mechanism[47] was employed to permit escape from local energy traps (see above). It should be emphasized that the method is authentic Monte Carlo within intervals but not between intervals, because the "freezing" of persisting structure abolishes microscopic reversibility.

Chain conformations are generated by randomly varying backbone and sidechain torsion angles ϕ, ψ, ω, and χ. Two important issues have been incorporated into the sampling strategy. Since the smallest local conformations—a turn of helix or a β-turn—involve several consecutive residues, the move set perturbs the chain three residues at a time, with moves chosen at random from the possibilities listed below.

Chain Geometry

In this initial implementation, LINUS uses backbone atoms—N, C_α, C, O, and C_β—and a highly simplified sidechain (Fig. 1). Ideal values of backbone bond lengths and scalar angles are maintained throughout.[48] Conformations are generated in internal coordinates and then transformed to Cartesian geometry for evaluation.

The Move Set

During each cycle, a pointer is advanced sequentially from N to C. With the pointer at i, a new backbone conformation is chosen at random from one of four move types:

1. *Helix*—three consecutive residues, $(i-1, i, i+1)$, are set to helical conformation, viz. $\phi = -64 \pm 15°$, $\psi = -43 \pm 15°$, $\omega = 180 \pm 10°$.
2. *Sheet*—three consecutive residues, $(i-1, i, i+1)$, are set to strand conformation, viz. $\phi = -120 \pm 30°$, $\psi = 120 \pm 30°$, $\omega = 180 \pm 10°$.
3. *Turn*—three consecutive residues, $(i-1, i, i+1)$, are set at random to one of 9 turn types for glycine or proline, or 8 turn types for all other residues, viz.

i. *Type I followed by left-hand conformation*
residue $(i-1)$: $\phi = -60 \pm 15°$, $\psi = -30 \pm 15°$, $\omega = 180 \pm 10°$
residue (i): $\phi = -90 \pm 15°$, $\psi = 0 \pm 15°$, $\omega = 180 \pm 10°$
residue $(i+1)$: $\phi = 90 \pm 15°$, $\psi = 0 \pm 15°$, $\omega = 180 \pm 10°$

ii. *Type I followed by extended conformation*
residue $(i-1)$: $\phi = -60 \pm 15°$, $\psi = -30 \pm 15°$, $\omega = 180 \pm 10°$
residue (i): $\phi = -90 \pm 15°$, $\psi = 0 \pm 15°$, $\omega = 180 \pm 10°$
residue $(i+1)$: $\phi = -90 \pm 15°$, $\psi = 120 \pm 15°$, $\omega = 180 \pm 10°$

iii. *Extended conformation followed by Type I*
residue $(i-1)$: $\phi = -90 \pm 15°$, $\psi = 120 \pm 15°$, $\omega = 180 \pm 10°$
residue (i): $\phi = -60 \pm 15°$, $\psi = -30 \pm 15°$, $\omega = 180 \pm 10°$
residue $(i+1)$: $\phi = -90 \pm 15°$, $\psi = 0 \pm 15°$, $\omega = 180 \pm 10°$

iv. *Type I′ followed by extended conformation*
residue $(i-1)$: $\phi = 60 \pm 15°$, $\psi = 30 \pm 15°$ $\omega = 180 \pm 10°$
residue (i): $\phi = 90 \pm 15°$, $\psi = 0 \pm 15°$, $\omega = 180 \pm 10°$
residue $(i+1)$: $\phi = -90 \pm 15°$, $\psi = 120 \pm 15°$, $\omega = 180 \pm 10°$

v. *Type II followed by extended conformation*
residue $(i-1)$: $\phi = -60 \pm 15°$, $\psi = 120 \pm 15°$, $\omega = 180 \pm 10°$
residue (i): $\phi = 80 \pm 15°$, $\psi = 0 \pm 15°$, $\omega = 180 \pm 10°$
residue $(i+1)$: $\phi = -90 \pm 15°$, $\psi = 120 \pm 15°$, $\omega = 180 \pm 10°$

vi. *Extended conformation followed by Type II*
residue $(i-1)$: $\phi = -90 \pm 15°$, $\psi = 120 \pm 15°$, $\omega = 180 \pm 10°$
residue (i): $\phi = -60 \pm 15°$, $\psi = 120 \pm 15°$ $\omega = 180 \pm 10°$
residue $(i+1)$: $\phi = 80 \pm 15°$, $\psi = 0 \pm 15°$, $\omega = 180 \pm 10°$

vii. *Type II′ followed by extended conformation*
residue (i−1): $\phi = 60 \pm 15°$, $\psi = -120 \pm 15°$, $\omega = 180 \pm 10°$
residue (i): $\phi = -80 \pm 15°$, $\psi = 0 \pm 15°$, $\omega = 180 \pm 10°$
residue $(i+1)$: $\phi = -90 \pm 15°$, $\psi = 120 \pm 15°$, $\omega = 180 \pm 10°$

viii. *Type V followed by extended conformation*
residue $(i-1)$: $\phi = -80 \pm 15°$, $\psi = 80 \pm 15°$, $\omega = 180 \pm 10°$
residue (i): $\phi = 80 \pm 15°$, $\psi = -80 \pm 15°$, $\omega = 180 \pm 10°$
residue $(i+1)$: $\phi = -90 \pm 15°$, $\psi = 120 \pm 15°$, $\omega = 180 \pm 10°$

ix. *Glycine half-turn*[49]
residue $(i-1)$: $\phi = -100 \pm 15°$, $\psi = -150 \pm 15°$, $\omega = 180 \pm 10°$
residue (i): $\phi = 110 \pm 15°$, $\psi = -160 \pm 15°$, $\omega = 180 \pm 10°$
residue $(i+1)$: $\phi = -90 \pm 15°$, $\psi = 140 \pm 15°$, $\omega = 180 \pm 10°$

x. *Type VI (cis-proline) followed by extended conformation*
residue $(i-1)$: $\phi = -120 \pm 15°$, $\psi = 120 \pm 15°$, $\omega = 0 \pm 10°$
residue (i): $\phi = -70 \pm 15°$, $\psi = 150 \pm 15°$, $\omega = 0 \pm 10°$
residue $(i+1)$: $\phi = -90 \pm 15°$, $\psi = 120 \pm 15°$, $\omega = 0 \pm 10°$

4. *Coil*—three consecutive residues, $(i-1, i, i+1)$, are set to have random values of ϕ and ψ between $-180°$ and $180°$, with $\omega = 180 \pm 10°$.

Each side chain has at most one torsion angle which is set to have a random value of χ between $-180°$ and $180°$.

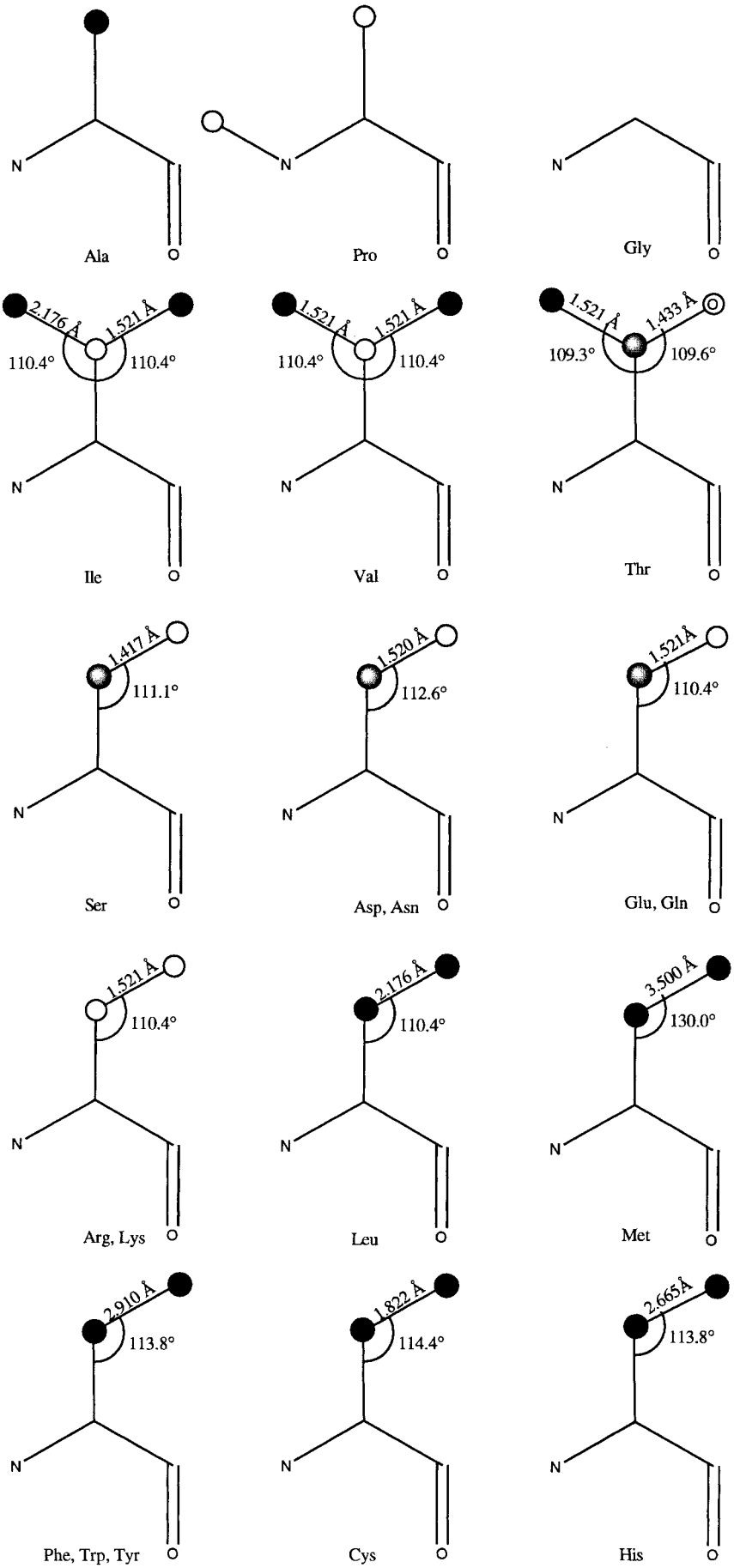
Ala
Pro
Gly
2.176 Å
1.521 Å
110.4°
110.4°
Ile
1.521 Å
1.521 Å
110.4°
110.4°
Val
1.521 Å
1.433 Å
109.3°
109.6°
Thr
1.417 Å
111.1°
Ser
1.520 Å
112.6°
Asp, Asn
1.521Å
110.4°
Glu, Gln
1.521 Å
110.4°
Arg, Lys
2.176 Å
110.4°
Leu
3.500 Å
130.0°
Met
2.910 Å
113.8°
Phe, Trp, Tyr
1.822 Å
114.4°
Cys
2.665Å
113.8°
His
N
O

Fig. 1.

The Energy Function

The energy function is extremely simple and includes just three components: (1) a scaled contact energy with both attractive and repulsive terms, (2) a primitive hydrogen bonding potential, and (3) a main chain torsional potential that helps "chase" the backbone away from values of $\phi > 0$ for residues other than glycine.

The repulsive component of the contact energy (i.e., steric repulsion) is the only accurately represented term in our energy function.[50] It is implemented by rejecting conformations in which the interatomic distance between any two atoms is less than the sum of their respective van der Waals radii, which are taken as 1.7 Å for carbon and sulfur, 1.45 Å for nitrogen, and 1.28 Å for oxygen. All pairwise interactions are evaluated except those involving carbonyl carbons, which are ignored. In addition, atoms connected covalently or by scalar angles (i.e., 1–3 interactions) or torsion angles (i.e., 1–4 interactions) are not evaluated.

The simple attractive term is applied between the side chain pseudo-atoms defined in Figure 1. For residues i and j, the pairwise distance, d_{ij}, between the two pseudo-atom interaction sites, C_i and C_j, is measured and assigned a maximal contact energy of 2 units when both i and j are hydrophobic, 1 unit when either i or j is hydrophobic and the other amphipathic, and 0 when neither i nor j is hydrophobic. As shown in Figure 1, the γ-position of side chains can be represented by either one or two pseudo-atoms. The maximal contact energy associated with a hydrophobic residue is one unit, which is split into two half-unit contributions in hydrophobic residues with two γ-pseudo-atoms. Hydrophobic resides include Cys, Ile, Leu, Met, Phe, Trp, and Val. Amphipathic residues include Ala, His, Thr, and Tyr. The energy units are arbitrary, of course, but are motivated by an analysis of Lesser and Rose[51] showing that, on average, all hydrophobic residues bury approximately the same fraction of their available hydrophobic surface.

This attractive term is scaled by distance, like a potential. The value is assigned to be maximal for C_i–C_j distances less than or equal to σ, the sum of the assigned radii in Figure 1, and it diminishes to 0 beyond σ + 1.4 Å. In the interval [σÅ, σ+1.4Å], the contact energy is calculated as

$$\textit{Contact Energy} = \text{maximum value} \times \left[1.0 - \frac{d_{ij}^2 - \sigma^2}{(\sigma+1.4)^2 - \sigma^2}\right]$$

Both backbone-to-backbone and backbone-to-side chain hydrogen bonds are allowed within the current Δ-interval. To realize a backbone-to-backbone H-bond, an N···O pair must be separated by at least one-residue in sequence and no more than 3.5 Å in space. Also, the out-of-plane dihedral angle between the oxygen and the peptide plane of the nitrogen (C–N–C_α) must not exceed 40°. If these criteria are satisfied, an H-bond energy of 0.5 units is assigned for $\Delta \leq 12$ and 1.0 units for $\Delta > 12$.

Backbone-to-side chain hydrogen bonds are allowed between a backbone nitrogen (at residue i), and a side chain acceptor situated at the β-carbon (at residue j). Only short range H-bonds are allowed. For Asn, Asp, Thr, and Ser, short range is defined as $i-4 \leq j \leq i+4$, $i \neq j$. For Gln and Glu, short range is defined as $i-4 \leq j \leq i-2$. In globular proteins, most backbone-to-side chain hydrogen bonds are found within these intervals.[52] Since our residues lack complete side chains, the C_β is used in lieu of the actual acceptor atom. The allowed out-of-plane dihedral angle may not exceed 40° (as above), and the C_β to nitrogen distance may not exceed 5 Å. An H-bond energy of 1.0 unit is assigned when these criteria are satisfied.

Finally, a main chain torsion potential is imposed

TABLE IA. Comparison of X-Ray vs. Predicted Secondary Structure for Cytochrome b_{562}

Structure*	X-Ray†	Prediction‡
Helix	3–19	3–18
Helix	23–41	24–41
Helix	56–79	56–81
Helix	84–104	84–104
Turn	80–82	80–82

*Secondary structure possibilities include helix, strand, and hydrogen-bonded turn. Recognition of each category was based solely on dihedral angles: helix is defined as three or more residues with $\phi = -60 \pm 20°$; $\psi = -40 \pm 20°$; strand is defined as three or more residues where $\phi < -80°$; $\psi > 80°$; and turns are defined in Methods.
†Residues that satisfy the given secondary structure definition in the X-ray structure.
‡Residues that satisfy the given secondary structure definition in the predicted structure.

TABLE IB. Frozen Segment List for Cytochrome b_{562}*

Interval = 6	3–7, 7–10, 10–14, 26–33
Interval = 18	34–38, 58–76

*List of segments that were conformationally constrained based on criteria specified in Methods (see hierarchic accretion). Residues that become constrained are listed for each interval of interaction, Δ.

Fig. 1. Atomic structure of the simplified side chains used in this work. The detailed geometry of all 20 residues is shown. Each side chain has at most one torsion angle (χ-angle). All side chain atoms occupy space and contribute to excluded volume. Atoms designated by opaque circles can participate in hydrophobic interactions. Atoms designated by shaded circles can serve as acceptors for backbone-to-side chain hydrogen bonds. Side chain radii have been assigned as follows (in Å): all β-carbons are 2.0. C_γ^1 and C_γ^2 of Ile and Val, C_γ^R of Thr, and S_γ of Cys are also 2.0. Other C_γs, in order of increasing size, are 2.50 for His; 3.00 for Met and Leu; and 3.25 for Phe, Tyr, and Trp.

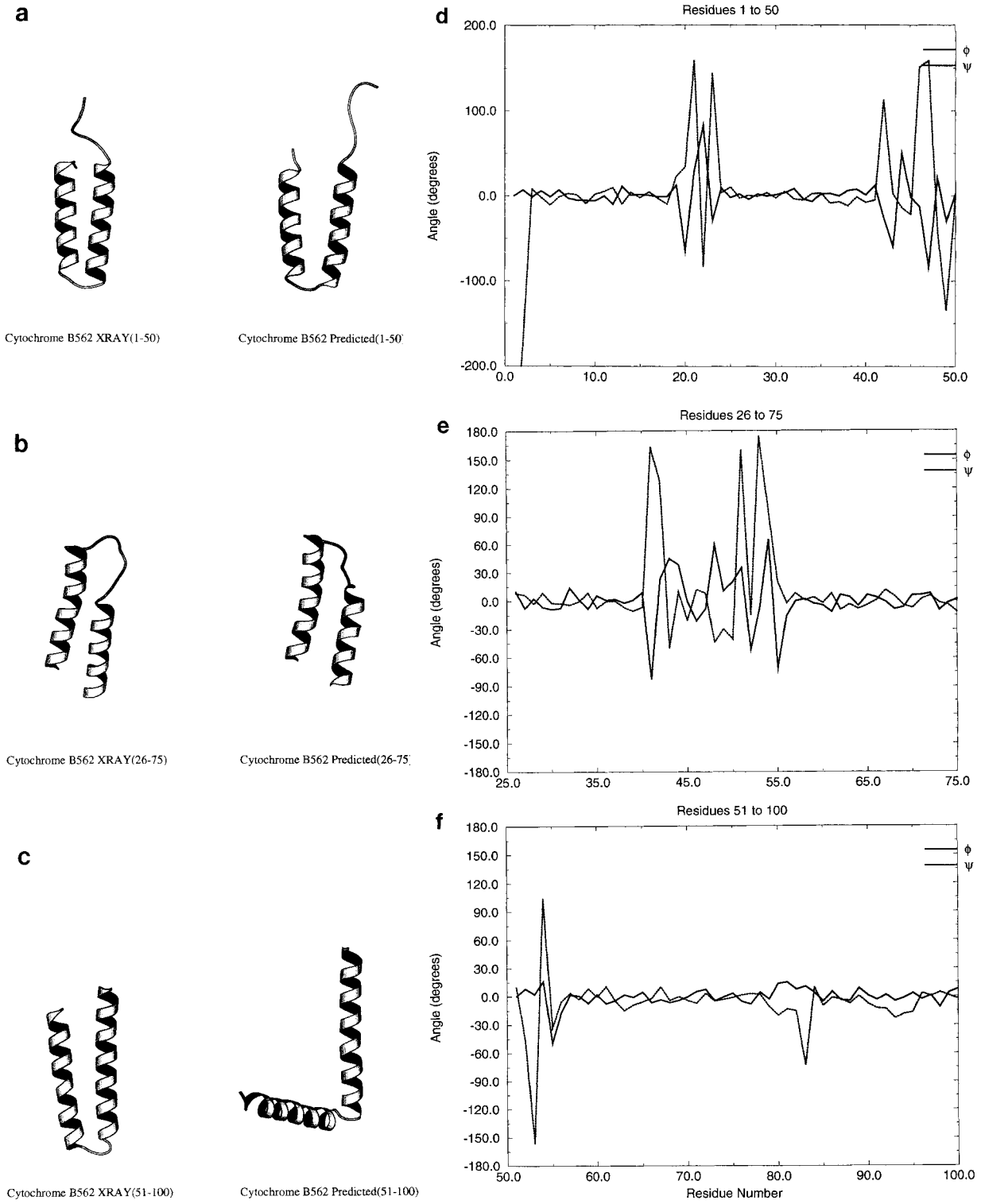

Fig. 2. Cytochrome b_{562}. Molscript drawing of X-ray determined vs. predicted structures for overlapping 50-residue fragments **(a)**–**(c)**, and difference dihedral angle plots for the same fragments **(d)**–**(f)**. A difference dihedral angle plot graphs either $\phi_{x\text{-}ray}$ (i) − $\phi_{predicted}$ (i) (solid line) or $\psi_{x\text{-}ray}$ (i) − $\psi_{predicted}$ (i) (broken line), for every residue i. **(a)** Residues 1–50. RMS difference for residues 3–41 = 2.8 Å. **(b)** Residues 26–75. RMS difference for residues 26–72 = 6.4 Å. **(c)** Residues 51–100. RMS difference for residues 56–82 = 1.2 Å.

TABLE IIA. Comparison of X-Ray vs. Predicted Secondary Structure for Fatty-Acid Binding Protein

Structure*	X-Ray†	Prediction‡
Helix	14–21	14–21
Helix	25–32	25–32
Strand	5–7	6–7
Strand	10–12	10–12
Strand	37–42	36–42
Strand	46–52	46–53
Strand	56–63	56–63
Strand	68–71	66–71
Strand	77–85	77–79, 81–85
Strand	88–92	88–92
Strand	101–109	101–109
Strand	112–119	112–119
Strand	122–131	122–129
Turn	—	3–4
Turn	—	20–21
Turn	44–45	44–45
Turn	64–65	64–65
Turn	73–74	73–74
Turn	86–87	86–87
Turn	110–111	110–111
Turn	120–121	120–121

*,†,‡Legends given in Table IA.

TABLE IIB. Frozen Segment List for Fatty-Acid Binding Protein*

Interval = 6	14–23, 47–49, 58–62, 68–72, 84–89, 102–105
Interval = 18	26–30, 62–66, 118–124, 96–101

*Legend given in Table IB.

TABLE IIIA. Comparison of X-Ray vs. Predicted Secondary Structure for Plastocyanin

Structure*	X-Ray†	Prediction‡
Strand	2–5	2–5
Strand	12–13	13–15
Strand	18–22	18–21
Strand	26–30	27–30
Strand	37–41	37–41
Strand	61–63	61–63
Strand	69–74	69–75
Strand	79–84	79–85
Strand	93–99	93–99
Turn	8–9	8–9
Turn	23–24	23–24
Turn	43–44	43–44
Turn	—	44–45
Turn	48–49	48–49
Turn	—	52–53
Turn	—	54–55
Turn	59–60	59–60
Turn	66–67	66–67
Turn	85–86,86–87, 87–88,88–89, 89–90 (sequential β-turns)	86–87,87–88, 88–89,89–90

*,†,‡Legends given in Table IA.

TABLE IIIB. Frozen Segment List for Plastocyanin*

Interval = 6	12–15, 19–21, 27–29, 80–84
Interval = 18	1–5, 8–10, 23–24, 66–67, 70–74, 86–90

*Legend given in Table IB.

to "chase" residues (except glycine) away from the right side of the ϕ,ψ-map (i.e., $\phi>0°$). Specifically, a residue with $\phi > 0°$ is penalized 1.0 units, unless it is Gly. In this latter case, the residue is rewarded 2.0 units if $0°\leq\phi\leq150°$.

Hierarchic Accretion

Hierarchy is realized by analyzing chain segments at the end of each interval. Structurally persisting segments—those that have adopted a preferred conformation throughout the interval—are then constrained to retain this conformation in subsequent intervals. Specifically, continuous 3-residue segments are constrained whenever two conditions are satisfied: (1) during an interval, the segment is found in the same region of ϕ,ψ-space (helix, extended, or the same turn type) more than 70% of the time, and (2) in this preferred conformation, the segment can realize a hydrophobic contact, i.e., a hydrophobic residue either within or immediately adjacent to the segment maintains a contact with another such residue. When so constrained, the segment, though "frozen," continues to sample other ϕ,ψ values within its preferred region. Longer persisting segments—ranging from helices or β-strands to supersecondary structure—become frozen as the concatenation of consecutive smaller segments.

RESULTS

Vignettes of protein fragments in the test set are given below, with the exception of apocytochrome *c*, which fails to adopt a folded conformation both in our simulation and under solution conditions that favor the folding of the holoprotein.[46] A folded conformation does emerge for all fragments of the other 6 molecules in the test set. Typically, the topology of predicted fragments matches that of X-ray elucidated counterparts, except in the case of dihydrofolate reductase.

Each protein is identified by name and by a parenthesized PDB identifier.[53] Agreement between the predicted and X-ray structure is assessed from the partitioning into secondary structure (helix, strand, and hydrogen-bonded turns), Molscript representations,[54] root-mean-square (RMS) differences, and difference-dihedral angle plots. A list of frozen segments and the hierarchic level at which freezing occurred is also presented for each molecule. In the

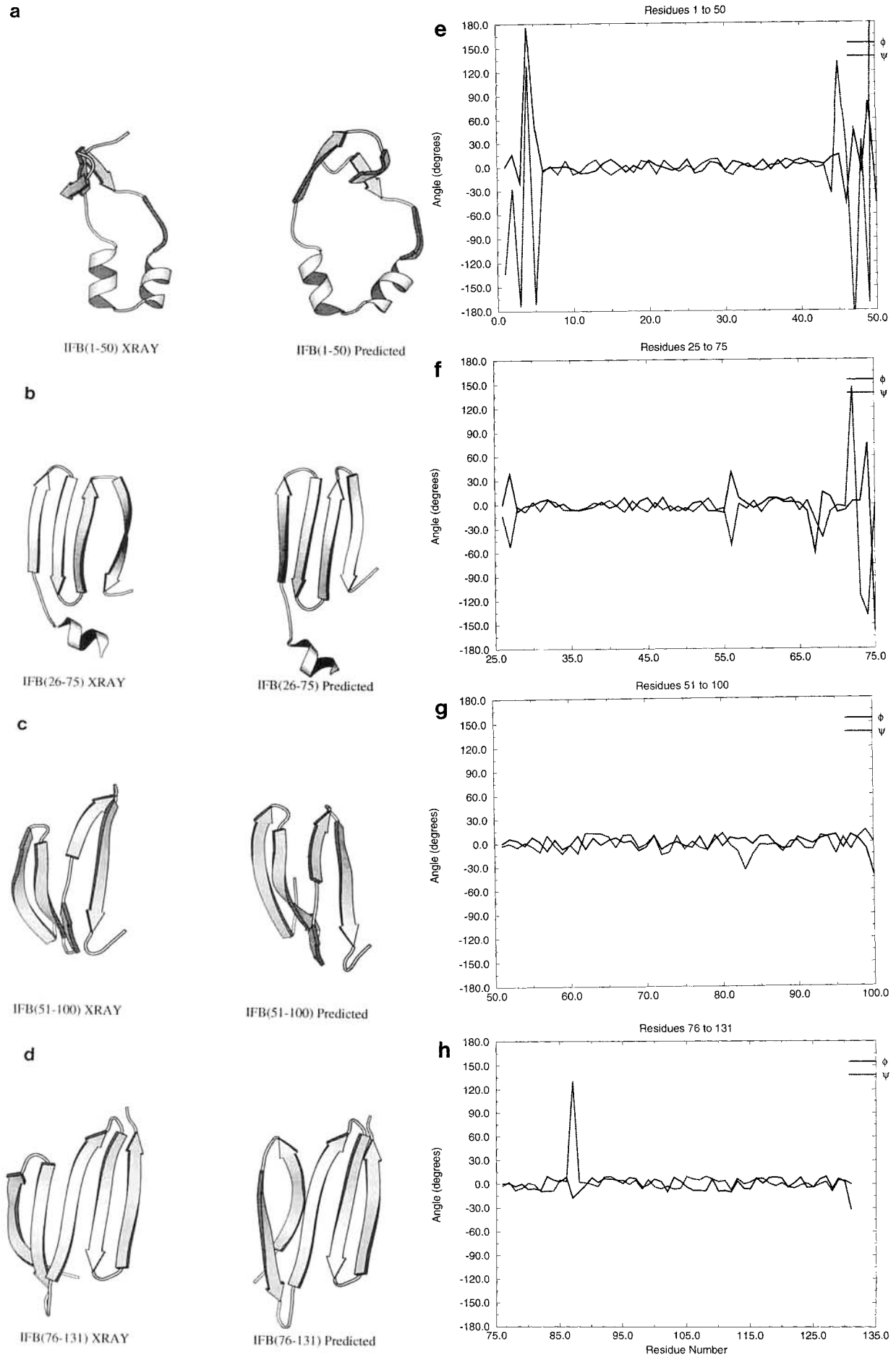
a
IFB(1-50) XRAY
IFB(1-50) Predicted
b
IFB(26-75) XRAY
IFB(26-75) Predicted
c
IFB(51-100) XRAY
IFB(51-100) Predicted
d
IFB(76-131) XRAY
IFB(76-131) Predicted
e
Residues 1 to 50
f
Residues 25 to 75
g
Residues 51 to 100
h
Residues 76 to 131
Angle (degrees)
Residue Number
φ
ψ

Fig. 3.

case of groES, the X-ray structure is not yet available, and only a Molscript drawing together with secondary structure and frozen segment lists are given. Coordinates of all predicted structures are being deposited in the Protein Data Bank.[53]

RMS differences between the predicted and X-ray structures are given in the figure caption for each fragment. To correct for end effects, the sequence interval over which the RMS has been calculated does not include terminal residues in a fragment (fragment i) that adopt a different conformation in the middle of the adjacent overlapping fragment (fragment $i+1$).

Cytochrome b_{562} (256B)

The structure of cytochrome b_{562}, a 4-helix bundle protein, has been solved to 1.4 Å resolution by Mathews and co-workers.[55] The structure of the apoprotein, used in our simulation, has been determined by NMR, where it was found that the C-terminal helix is largely unwound.[56] A sole hydrophobic residue anchors this fourth helix to the remainder of the molecule [viz., Leu (94)], with an energy of association that is correspondingly small.

Secondary structure and conformationally constrained segments are listed in Table I. Molscript representations and difference dihedral angle plots are displayed in Figure 2.

Helix 4 does not associate with the other three helices in our predicted structure for a reason that appears to be related to its behavior in the NMR-determined apoprotein[56]: insufficient hydrophobic surface is available to promote association. As a result, helices 1–3 twist relative to each other to optimize their mutual interaction, as shown in Figure 2.

Fatty-Acid Binding Protein (1IFB)

Intestinal fatty-acid binding protein (IFABP) is a member of the β-clamshell family, so called because the molecule is comprised primarily of two opposing layers of antiparallel β-sheet that surround an inner cavity. The protein has been the object of extensive folding studies by Frieden and co-workers[57] and its structure has been obtained to 1.96 Å resolution by Sacchettini et al.[58]

Secondary structure and conformationally constrained segments are listed in Table II. Molscript representations and difference dihedral angle plots are displayed in Figure 3.

Plastocyanin (1PCY)

Plastocyanin is a β-sheet protein involved in photosynthesis. The molecule, an 8-stranded barrel, has been solved to 1.6 Å resolution by Guss and Freeman.[59] Using NMR, Dyson et al. found that strand peptides tend to populate the extended region of ϕ,ψ space preferentially.[60] Plastocyanin provides a useful test because the strand order is not sequential, i.e., when strands in the β-barrel are indexed consecutively from 1 to 8, they correspond to a sequence order of 5, 4, 7, 8, 2, 1, 3, 6. In addition, the molecule contains 2 *cis*-prolines, at residues 16 and 36.

Secondary structure and conformationally constrained segments are listed in Table III. Molscript representations and difference dihedral angle plots are displayed in Figure 4.

Myoglobin (1MBO)

Apomyoglobin has been studied by NMR[61,62] and the structure of the holoprotein has been solved to 1.6 Å by X-ray crystallography.[63] As emphasized by Cocco and Lecomte,[61] "although it has been assumed for simplicity that native apomyoglobin is structurally the same as holomyoglobin, this view is not supported by circular dichroism data, which report a loss of approximately 20% of helicity when the heme is removed from the protein matrix."[64] The molecule contains 8 helices, conventionally labeled A through H. In comparison with other notable all-helical proteins, adjacent myoglobin helices do not coalesce into bundles, although an A–G–H complex is known to form in both the holoprotein and apoprotein.[61–64]

Secondary structure and conformationally constrained segments are listed in Table IV. Molscript representations and difference dihedral angle plots are displayed in Figure 5.

Eglin C (1CSE)

Eglin C is a tightly bound inhibitor of elastase that has been solved (in complex with subtilisin) to 1.2 Å resolution by Bode et al.[65] Upon binding, residues 1–7 are cleaved and/or disordered and cannot be located in the X-ray structure. The inhibitor consists of a helix followed by a sheet, with a large binding loop situated between two parallel β-strands. Wagner and co-workers have also solved the structure by NMR (including residues 1–7).[66] This molecule, only 70 residues in length, was simulated in its entirety and not partitioned into 50-residue fragments.

Secondary structure and conformationally constrained segments are listed in Table V. Molscript representations and difference dihedral angle plots are displayed in Figure 6.

Fig. 3. Fatty-acid binding protein. Molscript drawing of X-ray determined vs. predicted structures for overlapping 50-residue fragments **(a)**–**(d)**, and difference dihedral angle plots for the same fragments **(e)**–**(h)**. **(a)** Residues 1–50. RMS difference for residues 10–40 = 3.1 Å. **(b)** Residues 26–75. RMS difference for residues 30–72 = 3.8 Å. **(c)** Residues 51–100. RMS difference for residues 56–82 = 3.6 Å. **(d)** Residues 76–131. RMS difference for residues 76–131 = 3.7 Å.

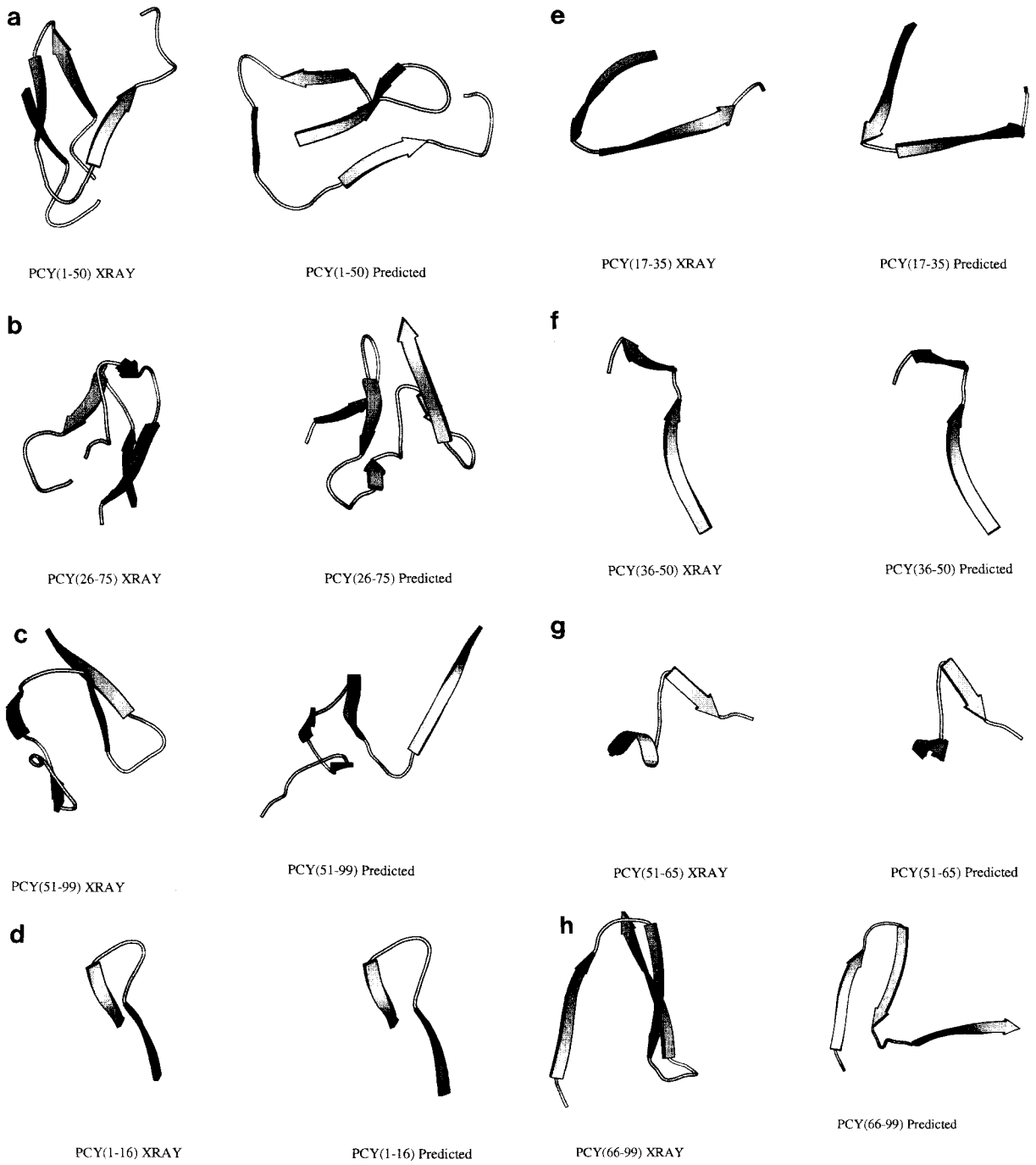

Fig. 4. Plastocyanin. Molscript drawing of X-ray determined vs. predicted structures for overlapping 50-residue fragments **(a)–(c)**, supersecondary structures **(d)–(h)**, and difference dihedral angle plots for the 50-residue fragments **(i)–(k)**. The supersecondary structures shown in **(d)–(h)**, though displayed separately, were not calculated separately. **(a)** Residues 1–50. RMS difference for residues 1–50 = 10.7 Å. **(b)** Residues 26–75. RMS difference for residues 36–65 = 4.1 Å. **(c)** Residues 51–99. RMS difference for residues 51–99 = 9.1 Å. **(d)** Residues 1–16; RMS difference = 1.7 Å. **(e)** Residues 17–35; RMS difference = 3.3 Å. **(f)** Residues 36–50; RMS difference = 1.5 Å. **(g)** Residues 51–65; RMS difference = 3.1 Å. **(h)** Residues 66–84; RMS difference = 3.2 Å.

Dihydrofolate Reductase (3DFR)

Dihydrofolate reductase (DFR) is an α/β protein that has been solved to 1.7 Å by Kraut and co-workers.[67] The molecule appears to fold in sequential steps that involve formation of a molten globule intermediate, followed by the emergence of subdomains en route to the native state.[68]

As mentioned previously, fragmentation into au-

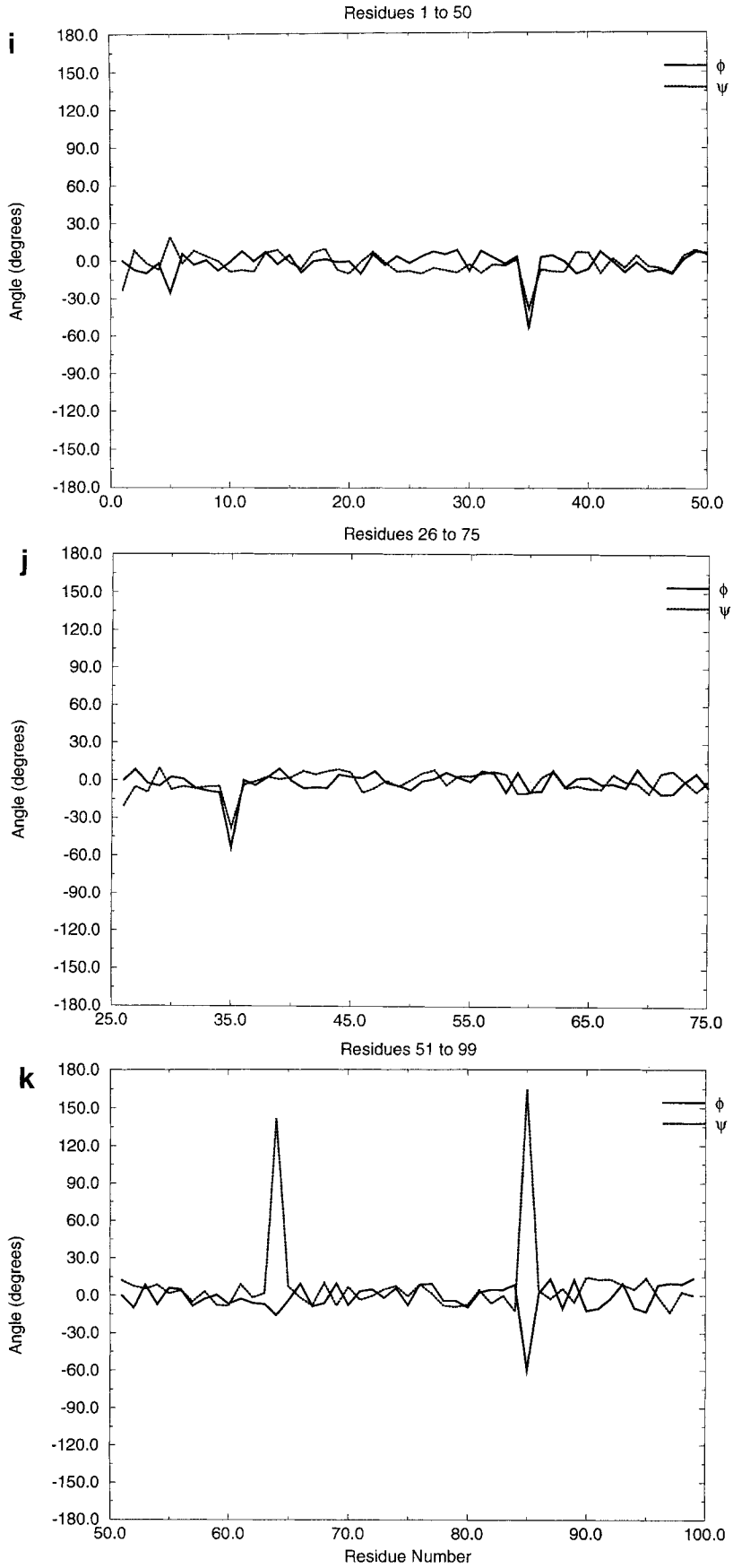
i
Residues 1 to 50
Angle (degrees)
180.0
150.0
120.0
90.0
60.0
30.0
0.0
-30.0
-60.0
-90.0
-120.0
-150.0
-180.0
0.0
10.0
20.0
30.0
40.0
50.0
φ
ψ
j
Residues 26 to 75
Angle (degrees)
25.0
35.0
45.0
55.0
65.0
75.0
k
Residues 51 to 99
Angle (degrees)
50.0
60.0
70.0
80.0
90.0
100.0
Residue Number

Fig. 4i-k.

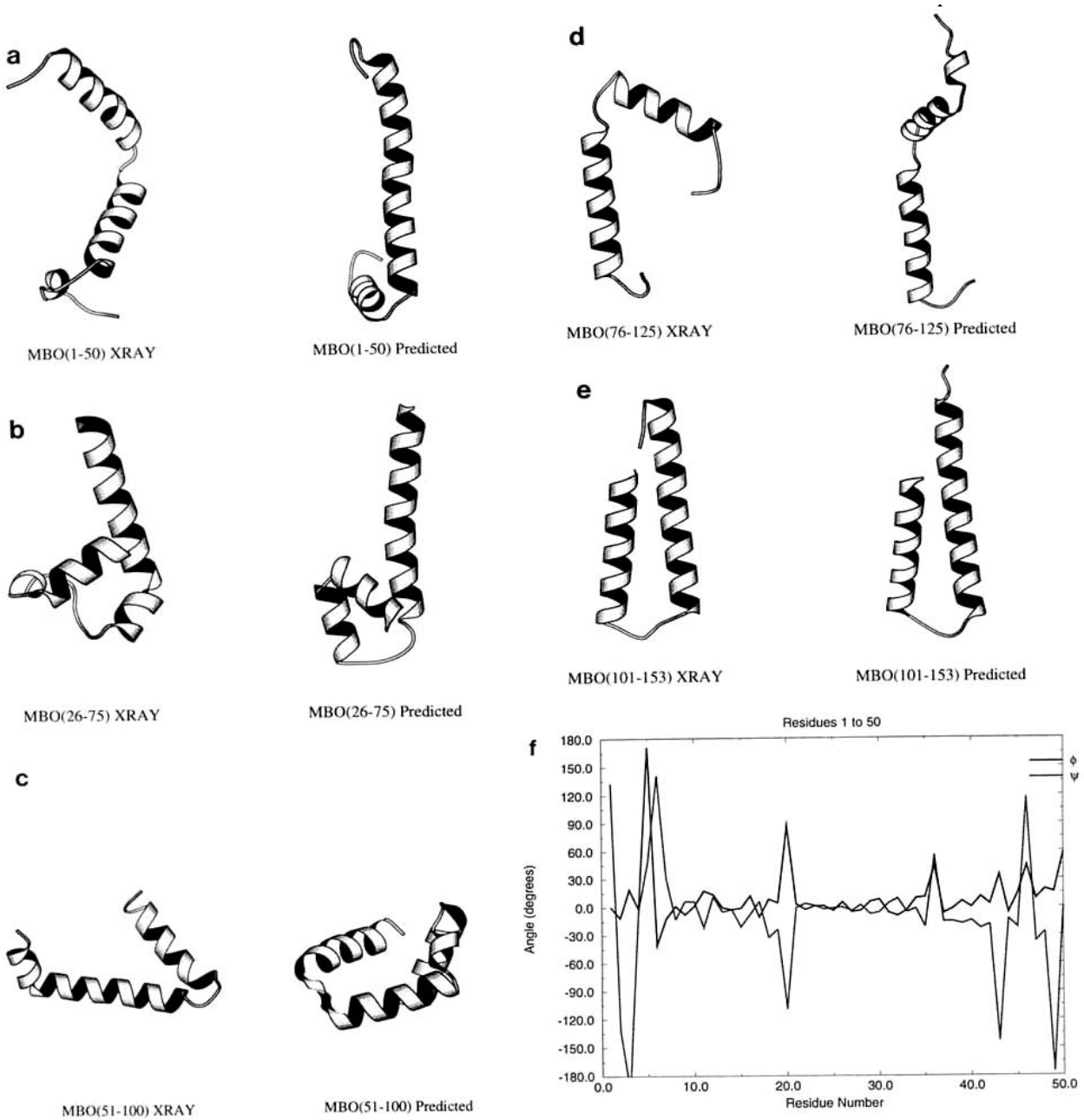

Fig. 5. Apomyoglobin. Molscript drawing of X-ray determined vs. predicted structures for overlapping 50-residue fragments **(a)**–**(e)**, and difference dihedral angle plots for the same fragments **(f)**–**(j)**. **(a)** Residues 1–50. RMS difference for residues 7–35 = 5.2 Å. **(b)** Residues 26–75. RMS difference for residues 26–75 = 8.3 Å. **(c)** Residues 51–100. RMS difference for residues 61–90 = 6.0 Å. **(d)** Residues 76–125. RMS difference for residues 86–115 = 5.5 Å. **(e)** Residues 101–153. RMS difference for residues 101–149 = 2.39 Å.

tonomous 50-residue segments is arbitrary and can result in spurious end-effects for larger values of Δ because fragment termini can be recruited into nonnative intrafragment interactions when deprived of native interactions with residues from adjacent fragments. This problem is pronounced in dihydrofolate reductase and leads to an incorrectly folded molecule, although conformationally constrained segments were in fact correct (Table VI). We view these results for DFR as a failure of the protocol.

groES (Prediction)

The *Escherichia coli* chaperone protein, groES, is the object of much topical interest.[69] The protein is an oligomer of a 97-residue subunit with the following sequence:[69]

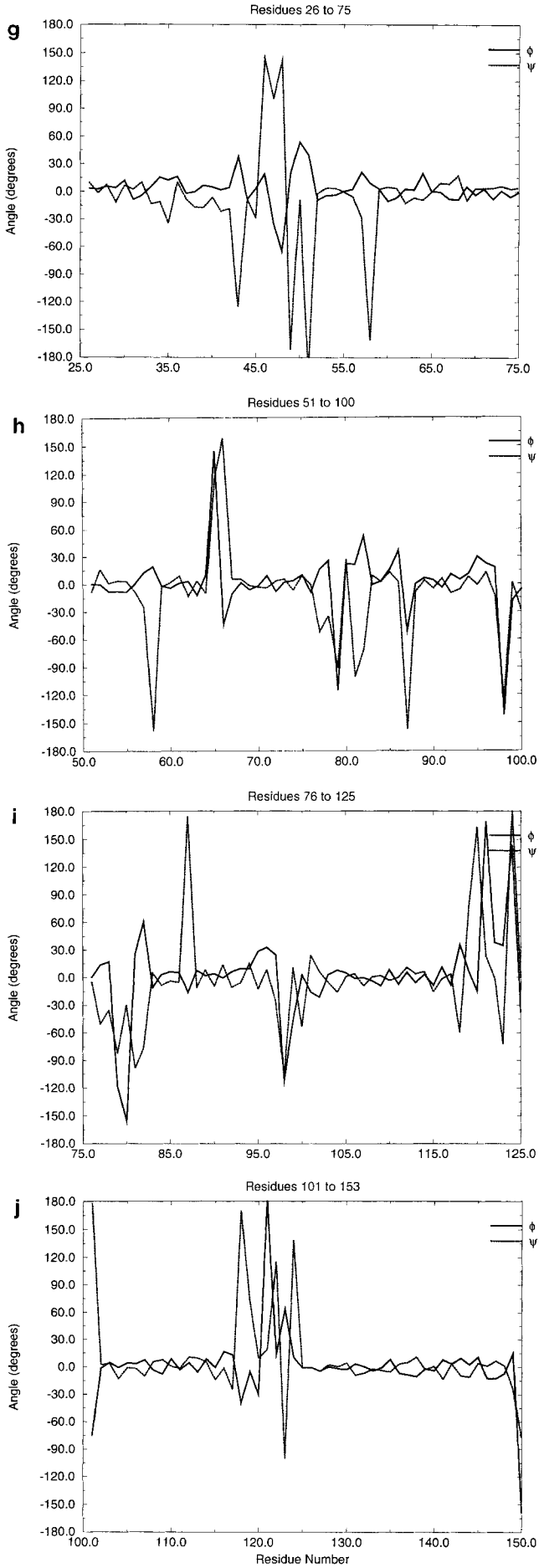
g
Residues 26 to 75
Angle (degrees)
180.0
150.0
120.0
90.0
60.0
30.0
0.0
-30.0
-60.0
-90.0
-120.0
-150.0
-180.0
25.0
35.0
45.0
55.0
65.0
75.0
ϕ
ψ
h
Residues 51 to 100
50.0
60.0
70.0
80.0
90.0
100.0
i
Residues 76 to 125
75.0
85.0
95.0
105.0
115.0
125.0
j
Residues 101 to 153
100.0
110.0
120.0
130.0
140.0
150.0
Residue Number

Fig. 5g-j.

TABLE IVA. Comparison of X-Ray vs. Predicted Secondary Structure for Apomyoglobin

Structure*	X-Ray†	Prediction‡
Helix	4–18	7–35
Helix	21–35	(see 7–35)
Helix	38–41	37–45
Helix	52–76	52–79
Helix	83–95	81–85,88–98
Helix	101–118	101–118
Helix	125–149	123–149
Turn	44–45,45–46, 46–47,47–48§	47–48,48–49
Turn	77–78,78–79,79–80	(see helix 52–79)
Turn	120–121,121–122	120–121,121–122

*,†,‡Legends given in Table IA.
§Our definitions classify the D helix as a series of turns.

TABLE IVB. Frozen Segment List for Apomyoglobin*

Interval = 6	7–21, 28–32, 52–66, 108–117, 126–138
Interval = 24	108–138

*Legend given in Table IB.

Met-Asn-Ile-Arg-Pro-Leu-His-Asp-
Arg-Val-Ile-Val-Lys-Arg-Lys-Glu-
Val-Glu-Thr-Lys-Ser-Ala-Gly-Gly-
Ile-Val-Leu-Thr-Gly-Ser-Ala-Ala-
Ala-Lys-Ser-Thr-Arg-Gly-Glu-Val-
Leu-Ala-Val-Gly-Asn-Gly-Arg-Ile-
Leu-Glu-Asn-Gly-Glu-Val-Lys-Pro-
Leu-Asp-Val-Lys-Val-Gly-Asp-Ile-
Val-Ile-Phe-Asn-Asp-Gly-Tyr-Gly-
Val-Lys-Ser-Glu-Lys-Ile-Asp-Asn-
Glu-Glu-Val-Leu-Ile-Met-Ser-Glu-
Ser-Asp-Ile-Leu-Ala-Ile-Val-
Glu-Ala

Unlike other molecules, groES has been simulated in its entirety, i.e. the simulation was run from $\Delta = 6$ to $\Delta = 97$. LINUS predicts an all β-protein, with 11 strands and 9 hydrogen-bonded turns (Table VII).

Summary of Predicted Secondary Structure

LINUS predicts secondary structure effectively, particularly helix and strand. The algorithm does fail for one molecule in the test set, viz., dihydrofolate reductase. However, the failure is a consequence of the protocol adopted in this study that calls for subdividing molecules into autonomous 50-residue fragments. When applied to the complete DFR sequence, LINUS's accuracy as a secondary structure predictor matched its performance on other molecules in the test set (data not shown).

Secondary structure recognition in both X-ray and predicted molecules was based solely on dihedral angles. Three or more residues are classified as (1) helix when $\phi = -60 \pm 20°$; $\psi = -40 \pm 20°$; (2) strand when $\phi < -80°$; $\psi > 80°$; and (3) hydrogen-bonded turn, based on conventional definitions,[34] as described in Methods.

TABLE VA. Comparison of X-Ray vs. Predicted Secondary Structure for Eglin C

Structure*	X-Ray†	Prediction‡
Helix	18–28	18–28
Strand	32–37	32–37
Strand	46–48	45–47
Strand	51–57	51–57
Strand	62–67	62–66
Turn	11–12, 12–13, 13–14, 14–15	—
Turn	30–31	30–31
Turn	39–40	39–40
Turn	49–50	49–50
Turn	58–59	58–59

*,†,‡Legends given in Table IA.

TABLE VB. Frozen Segment List for Eglin C*

Interval = 6	34–36, 43–49, 52–56
Interval = 12	18–24

*Legend given in Table IB.

Using these definitions, all predicted helices and strands have boundaries within two residues of their X-ray counterparts in cytochrome b_{562}, fatty-acid binding protein, plastocyanin, and eglin C, without underprediction or overprediction (tables IA, IIA, IIIA, VA). Hydrogen bonded turns are predicted almost as well. There are two overpredicted turns in fatty-acid binding protein, three overpredicted turns in plastocyanin, and an underpredicted sequence of four consecutive turns in eglin C; all other predicted turns match their X-ray counterparts precisely. Apomyoglobin is more difficult to assess because its structure may differ significantly from that of myoglobin.[64] In the predicted structure (Table IVA), the F helix and most of the C and D helices do not become conformationally constrained by our criteria, although these regions are found to populate helical conformations much of the time in latter stages of the simulation. Also, the algorithm does not partition the structure into A and B helices but predicts one long helix instead.

DISCUSSION

LINUS effectively determines the secondary and supersecondary structure of 5 proteins in our test set and the overall fold of most of their fragments. Of course, extensive atomic detail is beyond the scope of an algorithm that uses highly simplified side chains, lumps hydrophobics into a single class, ignores charges altogether, and forces geometry to maintain idealized values. Yet, many local features including the location of helices and strands of β-sheet are identified to within a residue or two; and turn types,

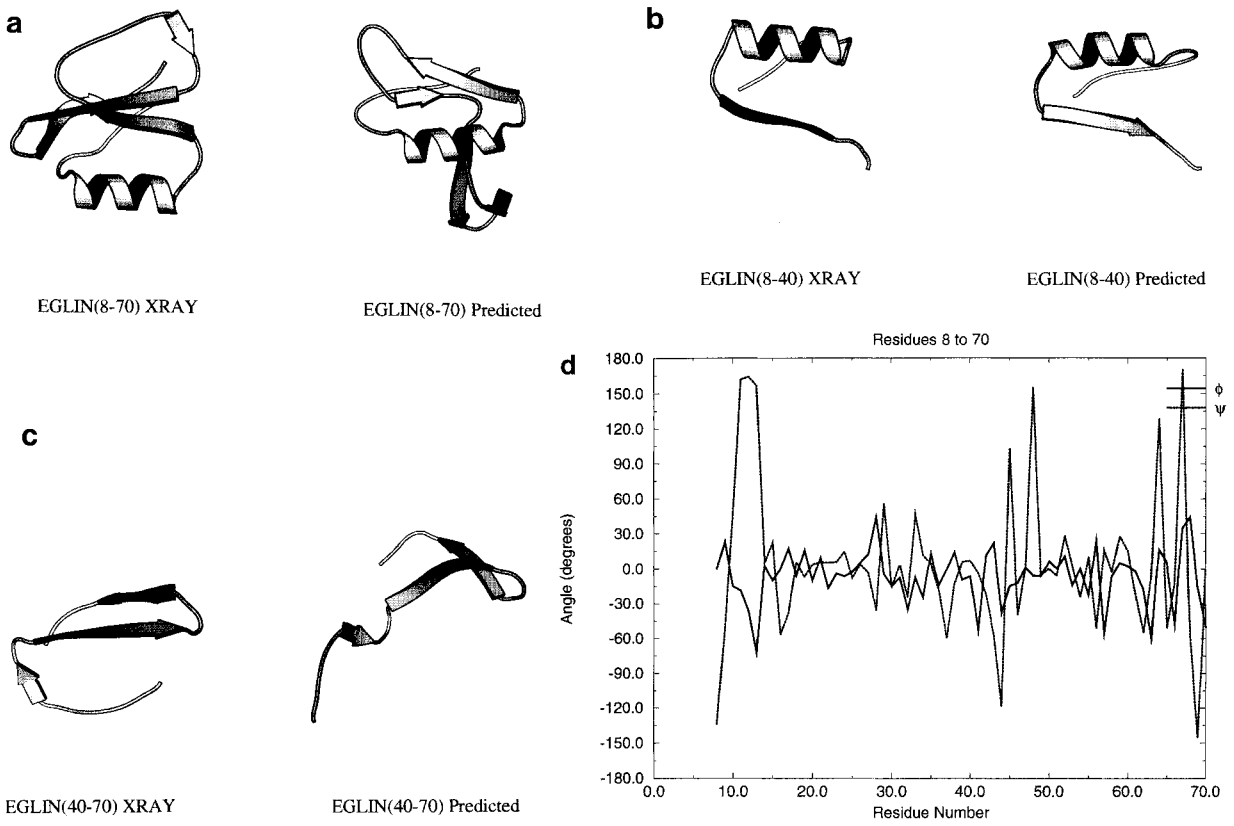

Fig. 6. Eglin C. Molscript drawing of X-ray determined vs. predicted structures for residues 8–70 **(a)–(c)** and difference dihedral angle plot for the same fragment **(d)**. **(a)** Residues 8–70. RMS difference for residues 8–70 = 12.1 Å. **(b)** Residues 8–40; RMS difference = 3.31 Å. **(c)** Residues 41–70; RMS difference = 7.42 Å.

TABLE VI. Frozen Segment List for Dihydrofolate Reductase*

Interval = 6	21–23, 48–50, 60–63, 71–73, 88–90, 113–115
Interval = 12	38–40, 80–91

*Legend given in Table IB.

cis-prolines, specific helix capping motifs,[70,71] and hydrogen bonding patterns are captured successfully, as seen in the Molscript diagrams and in more detailed representations (data not shown).

We felt obliged to assess the results using a standard figure of merit, the root-mean-square difference (RMS) between the predicted and X-ray structures, where

$$\mathrm{RMS} = \sqrt{\sum_i \frac{\Delta x^2 + \Delta y^2 + \Delta z^2}{N}}$$

and $\Delta x^2 + \Delta y^2 + \Delta z^2$ is the squared distance between corresponding points, the sum being performed over all N atoms. The RMS values are encouraging for large fragments but poor for entire molecules. As a single figure of merit, RMS differences are misleading in this case because they are dominated by a few large errors in dihedral angles, which are then propagated by the internal coordinate system. The use of internal coordinates is convenient, but it magnifies such errors because chain segments behave like lever arms, wherein a small change in an angle can result in a disproportionately large displacement. The worst errors occur when the predicted conformation of a residue or two is grossly in error, leading to global deformation of the structure.

A related problem concerns the subjective assessment of error in Molscript-rendered molecules. The visual perception of apparent similarity between complex objects is a complicated issue. Identical objects appear similar, but the eye is easily confounded by small changes in orientation, and perceived differences tend to accumulate in a nonlinear way.

A better comparison between X-ray and predicted structures can be achieved using the difference-dihedral angle plots that are presented for each molecule. These plots provide two figures-of-merit per residue, from which it is apparent that, residue-by-residue, the X-ray and predicted structures are in close agreement in most cases.

The approach employed here is remarkably sim-

TABLE VIIA. Predicted Secondary Structure for groES

Structure*	Prediction‡
Strand	1–3
Strand	9–14
Strand	17–19
Strand	25–28
Strand	31–33
Strand	40–44
Strand	57–61
Strand	64–67
Strand	71–74
Strand	83–87
Strand	90–96
Turn	15–16
Turn	29–30
Turn	45–46
Turn	51–52
Turn	62–63
Turn	68–69
Turn	76–77
Turn	80–81
Turn	88–89

*,‡Legend given in Table IA.

Table VIIB. Frozen Segment List for groES*

Interval = 12	1–3, 10–12, 17–19, 25–28, 40–44, 57–70, 78–96

*Legend given in Table IB.

ple, and its success makes a strong case for the proposition that protein conformation is determined largely by just four factors: (1) excluded volume, (2) a tendency for the polypeptide chain to populate preferred regions on the ϕ,ψ map, (3) a drive to make hydrophobic contacts and hydrogen bonds between sequentially proximate groups, and (4) a folding mechanism that progresses hierarchically toward the native conformation. Respectively, these four factors are implemented by using (1) a "bump check" procedure that will not allow two atoms to occupy the same space at the same time, (2) a move set that favors helix, sheet, and known turns, (3) an energy function that includes hydrophobic contacts and hydrogen bonds, and (4) a search strategy that allows only those interactions within a fixed linear window to "feel" one another, as the window size progresses from 6 residues (i.e. highly local) to the full length of the molecule (i.e., maximally global).

LINUS is somewhat insensitive to the individual character for residues. For example, a "computer mutation" of one hydrophobe to another would not be expected to have a large effect on the predicted conformation. Detailed side chain packing cannot play a dominant role in a model that uses such simplified side chains. Yet, minute changes in hydrophobic burial and packing are clearly measurable in actual proteins. Here it is important to emphasize the difference between stability and specificity.[7,72] *Stability* involves the question of why a protein molecule does not denature under physiological conditions, while *specificity* is concerned with the question of why that molecule adopts its particular fold and not another (e.g., why the lysozyme sequence adopts the lysozyme fold and not the ribonuclease fold). LINUS was designed to predict the protein fold (i.e., specificity), not to assess conformational free energy (i.e., stability).

The fact that most atoms are omitted from our simplified side chains is a welcome simplification, but it comes at the expense of the single most potent factor influencing protein conformation—excluded volume.[4] How does the algorithm achieve its current degree of success without the side chain's complete contribution to excluded volume? Apart from their polarity, side chains appear to play two distinct roles: one involves conformational organization, exerted through excluded volume and entropic effects[73,74]; the other is efficient internal packing.[75] Lacking detailed side chains, our predicted structures fail to achieve the conspicuous compactness that is a hallmark of proteins.[4] Yet, despite this deficiency, the predictions reproduce the overall topology quite faithfully, with only occasional catastrophes.

In essense, proteins can be regarded as unidirectional segments—helices and β-strands—joined by turns and loops that establish the chain trajectory.[76] Thus, if local structures—helices, strands, and turns—are identified accurately, then the overall fold will be largely correct, albeit lacking in atomic detail. It is these very features of the structure—helices, strands, and turns—that LINUS appears to capture successfully, and they are determined principally by local and medium range interactions ($\Delta \leq 18$), demonstrating the importance of such determinants in the global fold.

The issue of hierarchy deserves special mention. It has been known for some time that globular proteins are organized as a hierarchy.[27,28] This architecture immediately suggests a model,[28] called *folding by hierarchic condensation*, wherein nearby chain sites interact to form primitive folding modules which, in turn, interact iteratively to form larger modules of increasing complexity. An implicit benefit of this model is that it requires only a crude energy function because, at each step in the folding hierarchy, there are, at most, only a few competing interactions and, consequently, few energy-dependent branch points.

The free energy difference between the native and denatured states in proteins is quite small (~−5 to −15 kcal/mol), but it results from large opposing changes in entropy and enthalpy.[77] For this reason, we have come to think of the native fold as representing a precarious balance of forces (though Latt-

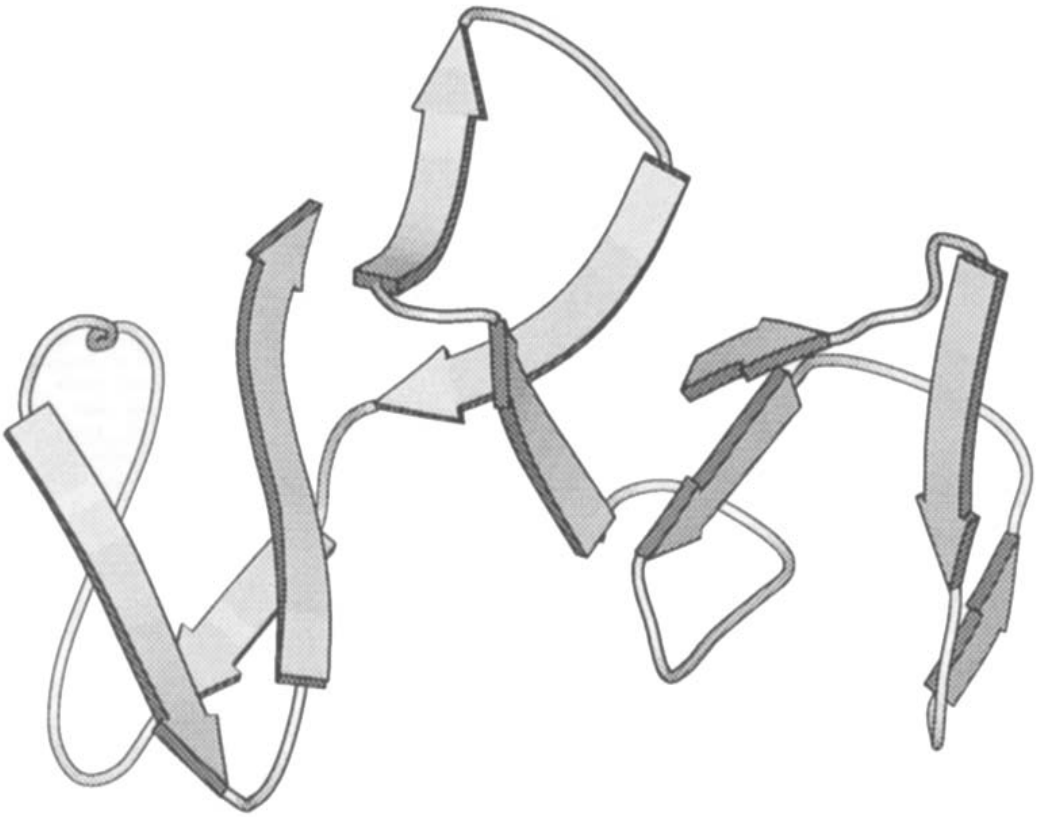

Fig. 7. GroES. Unlike the molecules in our test set, groES has been simulated in its entirety. The figure is a molscript drawing of the groES monomer, residues 1–97. The predicted supersecondary structure consists of 4 antiparallel strands connected by tight turns. In sequence, the strand-turn-strand complexes are (1) residues 9–19, (2) 25–33, (3) 64–74, and (4) 83–96. Complexes (1) and (2) interact as do (3) and (4); in each case the first antiparallel pair is roughly orthogonal to the second antiparallel pair.

man and Rose[7] suggest a different view). Therefore, it is worth commenting that the method described above is little more than a hodge-podge of crude approximations, introduced with little attempt at optimization, and the success of the results attests to an especially robust phenomenon, not a precarious one.

The absence of complete side chains blurs the identity of our residues, though not their classification into polar, apolar, and amphipathic categories. The argument that a given fold is specified primarily by the spacing of polar or apolar residues along the sequence, without regard for their exact identity, has been made persuasively by Dill and colleagues[78] and Hecht and colleagues.[79]

The emphasis in this initial version of LINUS has been on predicting the secondary and supersecondary structure and the overall fold of 50-residue fragments. The method is now being extended to prediction of entire molecules. We anticipate that the approach will find ready application in many topical problems, including genome initiatives[80] and protein threading.

ACKNOWLEDGMENTS

We owe Rajeev Aurora and Trevor Creamer an inestimable debt of gratitude for their creative ideas and critical assessment. We thank Jeffrey Seale for suggesting groES as an attractive candidate for prediction. G.D.R. also thanks Eaton Lattman for many

years of collaboration, inspiration, and encouragement. Supported by NIH GM29458.

NOTE ADDED IN PROOF

After submission of this manuscript, we communicated with Johann Deisenhofer and John Hunt, who are solving the X-ray structure of groES, and learned that repeating units in the heptamer include chain from two separate monomers. In retrospect, the molecule is a poor choice for prdection because LINUS is currently limited to protein monomers.

REFERENCES

1. Mirsky, A. E., Pauling, L. On the structure of native, denatured, and coagulated proteins. Proc. Natl. Acad. Sci. U.S.A. 22:439–447, 1936.
2. Wu, H. Studies on denaturation of proteins. XIII. A theory of denaturation. Chinese J. Physiol. V:321–344, 1931.
3. Anfinsen, C. B. Principles that govern the folding of protein chains. Science 181:223–230, 1973.
4. Richards, F. M. Areas, volumes, packing, and protein structure. Annu. Rev. Biophys. Bioeng. 6:151–176, 1977.
5. Kauzmann, W. Some factors in the interpretation of protein denaturation. Adv. Prot. Chem. 14:1–64, 1959.
6. Levinthal, C. Are there pathways for protein folding? J. Chim. Phys. 65:44–45, 1968.
7. Lattman, E. E., Rose, G. D. Protein folding—what's the question? Proc. Natl. Acad. Sci. U.S.A. 90:439–441, 1993.
8. Levitt, M., Warshel, A. Computer simulation of protein folding. Nature (London) 253:694–698, 1975.
9. Richmond, T. J., Richards, F. M. Packing of alpha-helices: geometrical constraints and contact areas. J. Mol. Biol. 119:537–555, 1978.
10. Lim, V. A mechanism of protein folding. A stereochemical theory of the tertiary structure of globular proteins. In: "Protein Folding." Jaenicke, R., ed. Amsterdam: Elsevier/North Holland, 1980.
11. Ptitsyn, O. B., Finkelstein, A. V. Self-organization of proteins and the problem of their three-dimensional structure prediction. In: "Protein Folding." Jaenicke, R., ed. Amsterdam: Elsevier/North-Holland, 1980.
12. Sternberg, M., Cohen, F. E., Taylor, W. R. A combinatorial approach to the prediction of the tertiary fold of globular proteins. Biochem. Soc. Trans. 10:299–301, 1982.
13. Kolinski, A., Skolnick, J., Yaris, R. Monte Carlo simulations on an equilibrium globular protein folding model. Proc. Natl. Acad. Sci. U.S.A. 83:7267–7271, 1986.
14. Crawford, I. P., Niermann, T., Kirschner, K. Prediction of secondary structure by evolutionary comparison: application to the α subunit of tryptophan synthase. Proteins Struct. Funct. Genet. 2:118–129, 1987.
15. Benner, S. A., Gerloff, D. Patterns of divergence in homologous proteins as indicators of secondary and teritary structure: a prediction of the structure of the catalytic domain of protein kinases. Adv. Enzyme Regul. 31:121–181, 1990.
16. Sippl, M. J. Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol. 213:859–883, 1990.
17. Bryant, S. H., Lawrence, C. E. An empirical energy function for threading protein sequence through the folding motif. Proteins Struct. Funct. Genet. 16:92–112, 1993.
18. Crippen, G. M. Prediction of protein folding from amino acid sequence over discrete conformation spaces. Biochemistry 30:4232–4237, 1991.
19. Friedrichs, M. S., Goldstein, R. A., Wolynes, P. G. Generalized protein tertiary structure recognition using associative memory hamiltonians. J. Mol. Biol. 222:1013–1034, 1991.
20. Hinds, D. A., Levitt, M. A lattice model for protein structure prediction at low resolution. Proc. Natl. Acad. Sci. U.S.A. 89:2536–2540, 1992.
21. Luthy, R., Bowie, J. U., Eisenberg, D. Assessment of protein models with three-dimensional profiles. Nature (London) 356:83–85, 1992.
22. Jones, D. T., Taylor, W. R., Thornton, J. M. A new approach to protein fold recognition. Nature (London) 358: 86–89, 1992.
23. Liwo, A., Pincus, M. R., Wawak, R. J., Rackowsky, S., Scheraga, H. A. Prediction of protein conformation on the basis of a search for compact structures: test on avian pancreatic polypeptide. Protein Sci. 2:1715–1731, 1993.
24. Monge, A., Friesner, R. A., Honig, B. An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. Proc. Natl. Acad. Sci. U.S.A. 91:5027–5029, 1994.
25. Sali, A., Shakhnovich, E. I., Karplus, M. How does a protein fold? Nature (London) 369:248–251, 1994.
26. Berg, J. M. Zinc finger domains: From predictions to design. Acc. Chem. Res. 28:14–19, 1995.
27. Crippen, G. M. The tree structural organization of proteins. J. Mol. Biol. 126:315–332, 1978.
28. Rose, G. D. Hierarchic organization of domains in globular proteins. J. Mol. Biol. 134:447–470, 1979.
29. Oas, T. G., Kim, P. S. A peptide model of a protein folding intermediate. Nature (London) 336:42–48, 1988.
30. Dill, K. A., Fiebig, K. M., Chan, H. S. Cooperativity in protein-folding kinetics. Proc. Natl. Acad. Sci. U.S.A. 90: 1942–1946, 1993.
31. Pauling, L., Corey, R. B., Branson, H. R. The structures of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. U.S.A. 37: 205–210, 1951.
32. Pauling, L., Corey, R. B. The pleated sheet, a new layer configuration of polypeptide chains. Proc. Natl. Acad. Sci. U.S.A. 37:251–256, 1951.
33. Venkatachalam, C. M. Stereochemical criteria for polypeptides and proteins. V Conformation of a system of three linked peptide units. Biopolymers 6:1425–1436, 1968.
34. Rose, G. D., Gierasch, L. M., Smith, J. A. Turns in peptides and proteins. Adv. Prot. Chem. 37:1–109, 1985.
35. Efimov, A. V. Common structural motifs in small proteins and domains. FEBS Lett. 355:213–219, 1994.
36. Leszczynski, J. F., Rose, G. D. Loops in globular proteins: a novel category of secondary structure. Science 234:849–855, 1986.
37. Ring, C. S., Kneller, D. G., Langridge, R., Cohen, F. E. A taxonomy and conformational analysis of loops in proteins. J. Mol. Biol. 224:685–699, 1992.
38. Fetrow, J. S., Zehfus, M. H., Rose, G. D. Protein folding: New twists. Biotechnology 6:167–171, 1988.
39. Richards, F. M., Kundrot, C. E. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins 3:71–84, 1988.
40. Padmanabhan, S., Marqusee, S., Ridgeway, T., Laue, T. M., Baldwin, R. L. Relative helix-forming tendencies of nonpolar amino acids. Nature (London) 344:268–270, 1990.
41. Lyu, P. C., Liff, M. I., Marky, L. A., Kallenbach, N. R. Side chain contributions to the stability of alpha-helical structure in peptides. Science 250:669–673, 1990.
42. O'Neil, K. T., DeGrado, W. F. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. Science 250:646–651, 1990.
43. Dyson, H. J., Cross, K. J., Houghten, R. A., Wilson, I. A., Wright, P. E., Lerner, R. A. The immunodominant site of a synthetic immunogen has a conformational preference in water for a type-II reverse turn. Nature (London) 318:480–483, 1985.
44. Bai, Y., Sosnick, T. R., Mayne, L., Englander, S. W. Protein folding intermediates by native-state hydrogen exchange. Science, in press.
45. Rose, G. D., Wolfenden, R. Hydrogen bonding, hydrophobicity, packing, and protein folding. Annu. Rev. Biophys. Biomol. Struct. 22:381–415, 1993.
46. Dumont, M. E., Corin, A. F., Campbell, G. A. Noncovalent binding of heme induces a compact apocytochrome c structure. Biochemistry 33:7368–7378, 1994.
47. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092, 1953.

48. Pauling, L., Corey, R. B. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. Proc. Natl. Acad. Sci. U.S.A. 37:235–240, 1951.
49. Efimov, A. V. Standard structures in proteins. Prog. Biophys. Mol. Biol. 60:201–239, 1993.
50. Bondi, A. Van der Waals volumes and radii. J. Phys. Chem. 68:441–451, 1964.
51. Lesser, G. J., Rose, G. D. Hydrophobicity of amino acid subgroups in proteins. Proteins Struct. Funct. Genet. 8:6–13, 1990.
52. Stickle, D. F., Presta, L. G., Dill, K. A., Rose, G. D. Hydrogen bonding in globular proteins. J. Mol. Biol. 226: 1143–1159, 1992.
53. Bernstein, F. C., Koetzle, T. G., Williams, G., Meyer, E., Jr., Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
54. Kraulis, P. J. Molscript: a program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst. 24:946–950, 1991.
55. Lederer, F., Glatigny, A., Bethge, P. H., Bellamy, H. D., Mathews, F. S. Improvement of the 2.5 Angstroms resolution model of cytochrome B562 by redetermining the primary structure and using molecular graphics. J. Mol. Biol. 148:427–448, 1981.
56. Feng, Y., Sligar, S. G., Wand, A. J. Solution structure of apocytochrome b_{562}. Nature Struct. Biol. 1:30–35, 1994.
57. Ropson, I. J., Gordon, J. I., Frieden, C. Folding of a predominantly beta-structure protein: Rat intestinal fatty acid binding protein. Biochemistry 29:9591–9599, 1990.
58. Sacchettini, J. C., Gordon, J. I., Banaszak, L. J. Refined apoprotein structure of rat intestinal fatty acid binding protein produced in escherichia coli. Proc. Natl. Acad. Sci. U.S.A. 86:7736–7740, 1989.
59. Guss, J. M., Freeman, H. C. Structure of oxidized poplar plastocyanin at 1.6 Angstroms resolution. J. Mol. Biol. 169:521–563, 1983.
60. Dyson, H. J., Sayre, J. R., Merutka, G., Shin, H. C., Lerner, R. A., Wright, P. E. Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding: II. Plastocyanin. J. Mol. Biol. 226:819–835, 1992.
61. Cocco, M. J., Lecomte, J. T. J. The native state of apomyoglobin described by proton NMR spectroscopy: interaction with the paramagnetic probe by HyTEMPO and the fluorescent dye ANS. Protein Sci. 3:267–281, 1994.
62. Hughson, F. M., Wright, P. E., Baldwin, R. L. Structural characterization of a partly folded apomyoglobin intermediate. Science 249:1544–1548, 1990.
63. Phillips, S. E. V. Structure and refinement of oxymyoglobin at 1.6 angstroms resolution. J. Mol. Biol. 142:531–554, 1980.
64. Breslow, E., Koehler, R. Properties of protoporphyrin-apomyoglobin complexes and related compounds. J. Biol. Chem. 240:PC2266–PC2268, 1965.
65. Bode, W., Papamokos, E., Musil, D., Seemueller, U., Fritz, H. Refined 1.2 A crystal structure of the complex formed between subtilisin Carlsberg and the inhibitor eglin C. Molecular structure of eglin and its detailed interaction with subtilisin. Embo J. 5:813–818, 1986.
66. Hyberts, S. G., Goldberg, M. S., Havel, T. F., Wagner, G. The solution structure of eglin C based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. Protein Sci. 1:736–751, 1992.
67. Bolin, J. T., Filman, D. J., Matthews, D. A., Hamlin, R. C., Kraut, J. Crystals structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 A resolution. I. General features and binding of methotrexate. J. Biol. Chem. 257:13650–13662, 1982.
68. Jones, B. E., Jennings, P. A., Pierre, R. A., Matthews, C. R. Development of nonpolar surfaces in the folding of Escherichia coli dihydrofolate reductase detected by 1-anilinonapthalene-8-sulfonate binding. Biochemistry 33:15250–15258, 1994.
69. Hemmingsen, S.M., Wooford, C., van der Views, S.M., Tilly, K., Dennis, D.T., Georgopoulos, C.P., Hendrix, R.W., Ellis, R.J. Homologous plant and bacterial proteins chaperone oligomeric protein assembly. Nature (London) 333: 330–334, 1988.
70. Aurora, R., Srinivasan, R., Rose, G.D. Rules for alpha-helix termination by glycine. Science 264:1126–1130, 1994.
71. Seale, J. W., Srinivasan, R., Rose, G. D. Sequence determinants of the capping box, a stabilizing motif at the N-termini of alpha-helices. Protein Sci. 3:1741–1745, 1994.
72. Rose, G. D., Creamer, T. P. Protein folding: Predicting predicting. Proteins Struct. Funct. Genet. 19:1–3, 1994.
73. Creamer, T. P., Rose, G. D. Side-chain entropy opposes alpha-helix formation but rationalizes experimentally determined helix-forming propensities. Proc. Natl. Acad. Sci. U.S.A. 89:5937–5941, 1992.
74. Creamer, T. P., Rose, G. D. alpha-Helix-forming propensities in peptides and proteins. Proteins Struct. Funct. Genet. 19:85–97, 1994.
75. 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.
76. Rose, G. D., Seltzer, J. A new algorithm for finding the peptide chain turns in a globular protein. J. Mol. Biol. 113:153–164, 1977.
77. Brandts, J. F .The thermodynamics of protein denaturation. II. A model of reversible denaturation and interpretations regarding the stability of chymotrypsinogen. J. Am. Chem. Soc. 86:4302–4314, 1964.
78. Yue, K., Dill, K. A. Forces of tertiary structural organization in globular proteins. Proc. Natl. Acad. Sci. U.S.A. 92: 146–150, 1995.
79. Brunet, A. P., Huang, E. S., Huffine, M. E., Loeb, J. E., Weitman, R. J., Hecht, M. H. The role of turns in the structure of an α-helical protein. Nature (London) 364: 355–358, 1993.
80. Potter, B. B., ed. "Mapping Our Genes—The Genome Projects: How Big, How Fast? Washington, DC: U.S. Government Printing Office, 1988.