Folding kinetics of designer proteins

Application of the diffusion-collision model to a de novo designed four-helix bundle

Kanthi K. Yapa and David L. Weaver
Department of Physics, Tufts University, Medford, Massachusetts 02155 USA

ABSTRACT A folding algorithm is described, based on the diffusion-collision model, combining static and dynamic calculational methods. The algorithm is applied to predict the basic structure and schematic folding pathways of an artificial four-helix bundle.

INTRODUCTION

New techniques in molecular biology have established the possibility of creating peptides of arbitrary length and sequence with novel properties not yet realized in nature. The techniques have great potential because of the enormous number of different artificial sequences that can, in principle, be made (e.g., $\approx 10^{130}$ for a chain of 100 amino acids) (1), and the possibility that a significant number of these sequences have a thermodynamically dominant fold (2, 3). It clearly will not be possible to determine all of their three-dimensional structures by physical methods, so a rational approach to protein engineering has as one of its elements the design of a computational model to predict the structure of a protein from its amino acid sequence. We describe a folding algorithm based on the diffusion-collision model (1), combining static and dynamic methods, and apply it to predict the basic structure and schematic folding pathways of one (4) of the recently designed and synthesized four-helix bundles (4, 5), whose three-dimensional structure has not yet been determined.

The algorithm is illustrated by the flow diagram:

homology → model building → minimization →
simplified dynamics → all atom MD

The diffusion-collision (1) model suggests that the principal, early dynamical events of the folding process are concerned with the properties of microdomains (α-helices, β-sheets, turns) and their interactions, rather than with the individual amino acids of the polypeptide chain, as in a random search. Hence, the folding algorithm emphasizes microdomains. It begins with the static, structural modeling by homology approach (step 1), first used by Browne et al. (6) to model α-lactalbumin on the basis of the homologous lysozyme structure, and since then used for modeling relaxins and insulin-like growth factors on the basis of insulin (7, 8), renins on the basis of aspartic proteinases (9), and so on, and the model building program developed by Sutcliffe et al. (10) (step 2), to produce a starting structure for the initial minimization (step 3), which is carried out to eliminate bad contacts in the model-built structure. The simplified dynamics (step 4), introduced to overcome the local minima problem (see below), can use one of the kinetics algorithms of the diffusion-collision model (1, 11, 12) which are: (*a*) the spherical-microdomain chemical kinetics approximation, which permits the calculation of the time course of the concentration of folding species, the delineation of pathways, and the determination of folding rates; and (*b*) the simplified-residue Langevin kinetics approximation, which circumvents the local minimum problem found in all atom molecular dynamics (MD) simulations. The spherical microdomain chemical kinetics approximation does not allow the introduction of a detailed residue representation to implement step 5. However, because it provides kinetic information with relatively small computational cost, it was used in the present preliminary calculation. All atom MD (step 5) allows the inner-side-chain interdigitation characteristic of a native, folded protein structure.

The α-helices in four-helix bundles are robust secondary structures, which lend themselves to physical characterization. They are used with steps 1–4 of the folding algorithm to predict time scales, schematic folding pathways, and concentrations of kinetic intermediates of the designed four-helix bundle. Previous work with the diffusion-collision model (1) has described folding kinetics and pathways for the all α-helical proteins myoglobin (11) and the operator-binding domain of the λ-repressor (12), both having known x-ray structures.

METHODS

A model of the designed bundle (4) was built, starting from the known x-ray structures of hemerythrin (13) (2MHR) and myohemerythrin (14) (1HMQ A chain), using the homologous model-building program COMPOSER (Sutcliffe et al. [10]). Schematically, the sequence is H_A-T_1-H_B-T_2-H_C-T_3-H_D, where H_i is [Gly-Glu-Leu-Glu-Glu-Leu-Leu-Lys-Lys-Leu-Lys-Glu-Leu-Leu-Lys-Gly] and T_i is [Pro-Arg-Arg]. The helices are labeled A, B, C, and D in order in the sequence. They are considered to be the microdomains that diffuse, collide, and coalesce into higher aggregates in the diffusion-collision model. To construct the model, the hydrophobic packings of 2MHR and 1HMQA were examined to identify the important residues in their helix–helix contacts, by determining the solvent accessible contact area (15, 16) loss upon packing for each helix-pair in the x-ray structures. The model sequence was fitted to the average structure for the C^α atoms of the x-ray structures, placing the centers of the model helices at the centers of the identified hydrophobic packing interactions. The loop sections were selected from structural fragments of correct length and with roughly correct end-point geometry. The preliminary structure was

0006-3495/92/07/296/04 $2.00

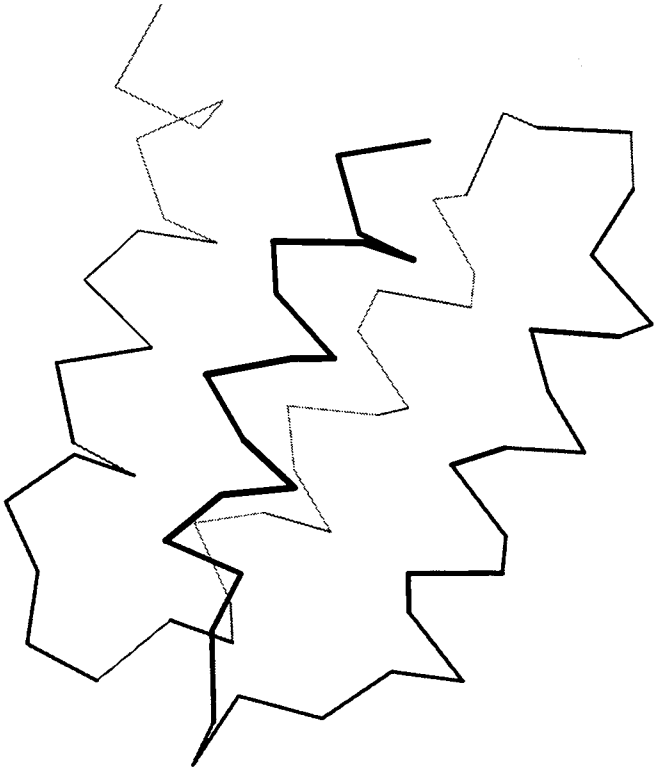


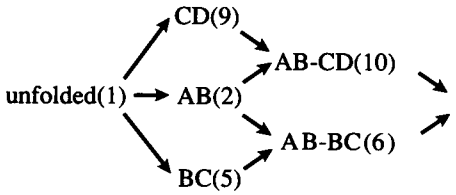
FIGURE 1 Backbone C^{α} trace of the model-built four-helix bundle after step 3 of the folding algorithm.

refined using the energy minimization program CHARMM (17) to remove bad contacts. The backbone C^{α} trace of the resulting model structure is shown in Fig. 1. Using this model, the helix–helix hydrophobic interactions, as measured by contact area loss (16) were found (see Table 1). These contacts stabilize the secondary structure of the microdomains in their helical conformations. The equivalent helix–helix contact area losses in 1HMQA chain and 2MHR are also listed in Table 1. Because the interacting residues in the interfaces of the helix-pairs are different in the three cases, the contact area losses and, hence, the hydrophobic interactions are not identical. However, in all cases, the BD pair contact area loss is much smaller than that of the other pairs, and the pair (AC) not mentioned has negligible packing area loss.

One of the problems of modeling starting with homologous structures is that there may be relative movements of up to 7 Å and rotations of up to 30° of secondary structural elements (18) in different members of the homologous family. Because of limitations in potential energy functions and computing power, this means that starting structures may be caught in local minima (during the feasible simulation time) if all atom MD is used immediately after step 3. A kinetics step (step 4) based on a simplified representation of the structure from step 3 is introduced to overcome the local minimum problem. The diffusion-collision model (1, 11, 12) was used for the simplified kinetics (step 4). The model suggests that the principal, early dynamical events of the folding process are concerned with the properties of microdomains and their interactions rather than with the individual amino acids of the polypeptide chain, as in a random search. In this application, the chemical kinetics approximation for calculating diffusion-collision dynamics was used in the simplified dynamics of the folding algorithm. The approximation (11, 19) is essentially analytic and simulates the dynamics of folding by a set of diffusion equations that describe the motion of microdomains in aqueous solution, and by coupled boundary conditions that provide for their collision and possible coalescence. Solving the diffusion equations provides the rate constants to be used in a set of first-order rate equations. The geometric parameters in the rate constants (see reference 11 for a complete description) were evaluated in a spherical approximation of the type often used to simplify diffusion calculations. Only collisions leading to the principal helix–helix native contacts (AB, BC, CD, AD) were followed.

RESULTS

The four-helix bundle microdomains and their physical properties are listed in Table 2 and the possible states of the folding protein (neglecting the BD pairing) are given in Table 3. Using the parameters for the microdomains and microdomain pairs in Tables 1–2, and the diffusion-collision rate equations (see reference 11), the probabilities of states 1–16 (Table 3) were determined numerically as a function of time for several choices of the stabilities of the individual helical microdomains and microdomain clusters (see Fig. 2 for the results of a typical simulation run). From the simulations, the schematic folding pathways shown below (the numbers refer to Table 3) have been deduced.



AB-BC-CD(14) → AB-BC-CD-AD(16),

which correspond to four competing, parallel pathways: 1 → 2 → 6 → 14 → 16, 1 → 2 → 10 → 14 → 16, 1 →

TABLE 1 Contact area loss of helix pairs in four-helix bundles

	Residue range				Contact area loss (Å²)		
Helices	Designed	1HMQA	2MHR	Pairs	Designed	1HMQA	2MHR
A	1–16 (16)	21–37 (17)	18–38 (21)	AB	211	310	325
B	20–35 (16)	41–64 (14)	40–65 (16)	AD	130	200	241
C	39–54 (16)	41–64 (14)	40–65 (16)	BC	156	301	296
D	58–73 (16)	90–103 (14)	92–115 (24)	BD	77	50	97
				CD	206	143	250

Residue ranges of the A, B, C, and D helices and contact area loss in Å² of helix pairs in the designed four-helix bundle and in the x-ray structures of 1HMQA chain and 2MHR.

TABLE 2 **Properties of four-helix bundle microdomains**

Label	Radius* (Å)	Mass‡ (amu)	D§ (Å²/ns)
A	9.07	1702	36.19
B	9.06	1700	36.22
C	9.06	1700	36.22
D	9.08	1716	36.15
AB	11.42	3402	28.73
AD	11.43	3418	28.71
ABD	13.08	5118	25.09
BC	11.42	3400	28.75
ABC	13.07	5102	25.10
ABCD	14.39	6818	22.80
BCD	13.08	5116	25.10
CD	11.43	3416	28.72
ACD	13.08	5118	25.09

* Radius in Å of the spherical approximation to the microdomain; ‡mass (in amu) of the microdomain; §relative diffusion coefficient in Å²/ns of the microdomain or microdomain cluster.

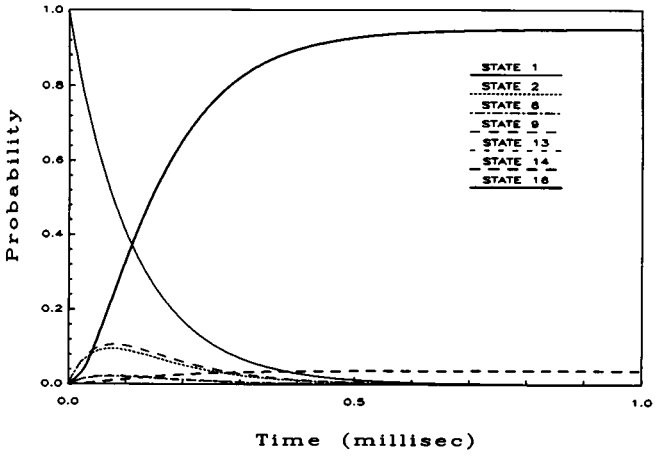


FIGURE 2 An example of a chemical kinetics approximation, diffusion-collision folding run. In this run, the individual helices were assumed to be marginally stable (equilibrium constant 10^{-2}), helix pairs to be more stable (equilibrium constant 10^{-1}), and clusters of three or more helices to be as stable as helix pairs. The time scale is determined by the microdomain stabilities.

5 → 6 → 14 → 16, and 1 → 9 → 10 → 14 → 16. The multiple pathways are a result of the similar physical properties of the four helical microdomains, the assumption of similar stability (helix-coil equilibrium) for each of them and the central role played by the microdomains and their interactions in diffusion-collision folding dynamics. In each pathway, the steps involve a pair of adjacent microdomains and microdomain clusters, as expected by consideration of the size of the available diffusion space, when other factors are equal. Multiple, equivalent pathways suggest the "jigsaw puzzle" model of folding (20), a possible realization of the diffusion-collision model (1), which proposes that proteins fold by a large number of relatively equal, parallel pathways rather than by a single defined sequence of events. It will often be true, however, that "other factors" are not equal. In particular, the stabilities of the individual helical microdomains and microdomain clusters will generally be different. This will affect the relative probabilities of the folding pathways, and perhaps cause a single pathway to dominate the folding kinetics. The relative importance of individual pathways could change in evolution by mutations that affect the stabilities of microdomains and the stability of their pairings. Determining physical properties of microdomains, in particular, their intrinsic stabilities, is important in differentiating folding pathways in the model. The spherical, analytical approximation to the rate constants used here does not allow detailed molecular dynamics (step 5) to compute the final positions of residues in the native structure of the artificial four-helix bundle. It does provide the time course of the kinetic intermediates (microdomain clusters listed in Table 3) formed during the folding reaction (for example, see Fig. 2) and the associated parallel pathways.

TABLE 3 **Folding states**

State	Helix pairs
1	none
2	AB
3	AD
4	AB AD
5	BC
6	AB BC
7	AD BC
8	AB AD BC
9	CD
10	AB CD
11	AD CD
12	AB AD CD
13	BC CD
14	AB BC CD
15	AD BC CD
16	AB AD BC CD

Connection between state number and type of helix pairs.

In future work, the Langevin kinetics method (21–24) will be used for the dynamics, on a simplified representation of the modeled peptide chain derived from the first three steps in the algorithm. The helices, sheets, and loops in the model structure will be allowed to explore conformational space to approximate the large movements observed in the comparative structures of homologous family members (18). The resulting structure will then be used in an all atom molecular dynamics simulation (step 5) to allow the detailed inter-side-chain interactions characteristic of a native, folded protein structure.

The authors would like to thank Suhail Islam, Martin Karplus, and Michael Sutcliffe for useful discussions, Tom Blundell for the hospitality of his laboratory, and the National Institute of General Medical Sciences (grant 29553) for support.

REFERENCES

1. Karplus, M., and D. L. Weaver. 1976. Protein folding dynamics. *Nature (Lond.)*. 260:404–406.
2. Shakhnovich, E. I., and A. M. Gutin. 1990. Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature (Lond.)*. 346:773–775.
3. Chan, H. S., and K. A. Dill. 1990. Origins of structure in globular proteins. *Proc. Natl. Acad. Sci. USA*. 87:6388–6392.
4. Regan, L., and W. F. DeGrado. 1988. Characterization of a helical protein designed from first principles. *Science (Wash. DC)*. 241:976–978.
5. Hecht, M. H., J. S. Richardson, D. C. Richardson, and R. C. Ogden. 1990. DeNovo design, expression and characterization of Felix: a four-helix bundle protein of native-like sequence. *Science (Wash. DC)*. 249:884–891.
6. Browne, W. J., A. C. T. North, D. C. Phillips, K. Brew, T. Vanaman, and R. L. Hill. 1969. A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol.* 42:65–86.
7. Bedarkar, S., W. G. Turnell, and T. L. Blundell. 1976. Relaxin has conformational homology with insulin. *Nature* 270:449–451.
8. Blundell, T. L., S. Bedarkar, E. Rinderknecht, and R. E. Humbel. 1978. Insulin-like growth factor: a model for tertiary structure accounting for immunoreactivity and receptor binding. *Proc. Natl. Acad. Sci. USA*. 75:180–184.
9. Blundell, T. L., B. L. Sibanda, and L. Pearl. 1983. Three-dimensional structure, specificity and catalytic mechanism of renin. *Nature (Lond.)*. 304:273–275.
10. Sutcliffe, M. J., I. Haneef, D. Carney, and T. L. Blundell. 1987. Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* 1:377–384.
11. Bashford, D., F. E. Cohen, M. Karplus, I. D. Kuntz, and D. L. Weaver. 1988. Diffusion-collision model for the folding kinetics of myoglobin. *Proteins*. 4:211–227.
12. Bashford, D., D. L. Weaver, and M. Karplus. 1984. Diffusion-collision model for the folding kinetics of the λ-repressor operator-binding domain. *J. Biomol. Struct. Dynam.* 1:1243–1250.
13. Sheriff, S., W. A. Hendrickson, and J. L. Smith. 1987. Structure of myohemerythrin in the azidomet state at 1.7/1.3 Å, *J. Mol. Biol.* 197:273–296.
14. Stenkamp, R. E., L. C. Sieker, and L. H. Jansen. 1983. Adjustment of restraints in the refinement of methhemerythrin and azidomethhemerythrin at 2.0 Angstroms resolution. *Acta Cryst.* B39:697–715.
15. Lee, B. K., and F. M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55:379–400.
16. Richmond, T. J., and F. M. Richards. 1978. Packing of α-helices: geometrical constraints and contact areas. *J. Mol. Biol.* 119:537.
17. Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. CHARMM: a program for macromolecular energy, minimization and dynamics calculations. *J. Comp. Chem.* 4:187–217.
18. Lesk, A. M., and C. Chothia. 1980. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* 136:225–270.
19. Weaver, D. L. 1984. Alternative pathways in diffusion-collision controlled protein folding. *Biopolymers* 23:675–694.
20. Harrison, S. C., and R. Durbin. 1985. Is there a single pathway for the folding of a polypeptide chain? *Proc. Natl. Acad. Sci. USA* 82:4028–4030.
21. McCammon, J. A., S. H. Northrup, M. Karplus, and R. M. Levy. 1990. Helix-coil transitions in a simple polypeptide model. *Biopolymers* 19:2033–2045.
22. Yapa, K., D. L. Weaver, and M. Karplus. 1992. Beta-sheet-coil transitions in a simple polypeptide model. *Proteins* 12:237–265.
23. Lee, S., M. Karplus, D. Bashford, and D. L. Weaver. 1987. Brownian dynamics simulation of protein folding: a study of the diffusion-collision model. *Biopolymers* 26:481–506.