Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 4462–4465, May 1992
Biophysics

Environment affects amino acid preference for secondary structure

(proteins/oligopeptides/circular dichroism)

LINGXIU ZHONG AND W. CURTIS JOHNSON, JR.

Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR 97331-6503

Communicated by K. E. van Holde, February 11, 1992

ABSTRACT **Three equivocal amino acid sequences were synthesized that are predicted to be α-helical from amino acid preference but are found to be primarily β-strand from x-ray diffraction of their respective proteins. In some solvent systems we recover the α-helical structure predicted by amino acid preference, whereas in other systems we mimic the interior of the protein and produce a β-strand. These results are experimental proof that the environment is important in determining the secondary structure formed by an amino acid sequence; therefore schemes that predict secondary structure from amino acid sequence alone can never be totally successful.**

It is generally accepted that the amino acid sequence of a protein determines its ultimate three-dimensional structure. In the hierarchic view, the primary structure determines regular repeating secondary structures, which in turn fold up into a tertiary structure. Researchers have noted that certain amino acids have preference for a given secondary structure, and a number of schemes have been developed that use amino acid preference to predict secondary structure from primary structure (1–4). There are also predictions of secondary structure based on homology (5–9), linear optimization of predictors (10), and neural networks (11, 12). The results of methods for predicting secondary structure from amino acid sequence were initially impressive but have failed to improve substantially; generally, about 60% of the residues can be classed correctly as α-helix, β-strand, β-turn, or other.

Here we view the protein folding problem in reverse and ask the question: Why is each amino acid found in every type of secondary structure? We investigate three equivocal sequences of amino acids that are predicted to be in one secondary structure from amino acid preferences but are actually found in another secondary structure. These are the interesting sequences, because they are the demonstrated failures of the prediction methods. For all three equivocal sequences we recover the predicted secondary structure in some solvent systems. We then follow the secondary structure as a function of the solvent, ultimately mimicking the environment inside a protein and producing the observed secondary structure. Our research demonstrates experimentally that the environment is important in determining the secondary structure formed by an amino acid sequence.

MATERIALS AND METHODS

Choosing the Peptide Sequence. Our first choice of an equivocal sequence came from earlier work (13) on *Eco*RI endonuclease (ERE). We predicted residues 103–115 (ERE) to be an α-helix by several primary sequence prediction methods, but it is a β-strand in the protein (14, 15). To obtain more equivocal sequences we applied the Chou and Fasman method (1) to the Kabsch and Sander (16) data base. Some sequences that were predicted to be an α-helix were shown primarily as a β-strand in the data base. We chose the two sequences that were predicted to be the longest α-helices, 77–90 from γ-chymotrypsin (CMT) and 5–19 from liver alcohol dehydrogenase (ADH; liver apoenzyme). The sequences we synthesized (Table 1) are the three predicted to be an α-helix. The portions found to be β-strand from inspection of the x-ray diffraction (14, 15, 17, 18) are underlined in Table 1.

Peptide Synthesis and Purification. Peptide sequences were synthesized by solid-phase methods on automated Applied Biosystems peptide synthesizer model 430A or 431A. To avoid end effects we blocked the N terminus with acetyl and the C terminus with amine. Position 77 in CMT was changed to Trp so that all three peptides would have an aromatic residue for UV detection at 280 nm. This change did not alter the potential of the peptide for α or β structure. The synthesized peptide sequences were purified by HPLC using a VYDAC C_{18} reverse-phase column, with a gradient of water to acetonitrile in the presence of 0.1% and 0.06% trifluoroacetic acid. Peptide identity was verified by amino acid analysis and mass spectroscopy.

Spectroscopy. Absorption was measured on a Cary 15 purged with nitrogen. Absorption data were obtained for the same preparations at 280 nm and 190 nm, and extinction coefficients at both wavelengths were established by the guanidine hydrochloride method (19). We also used amino acid analysis results to confirm the protein concentration.

CD spectra of freshly prepared samples were measured from 260 to 178 nm on a McPherson vacuum UV spectrophotometer modified for CD as described elsewhere (20). Measurements were made using quartz cells of various pathlengths. The instrument was calibrated using (+)-10-camphorsulfonic acid, $\Delta\varepsilon = +2.37\ \mathrm{M^{-1}\cdot cm^{-1}}$ at 290.5 nm and $-4.95\ \mathrm{M^{-1}\cdot cm^{-1}}$ at 192.5 nm. The results were digitized at 0.5-nm intervals using an IBM-type computer system, which collected the data at a rate of 1 nm/min. Spectra were smoothed using a cubic spline algorithm. To monitor solutions for aggregation, CD spectra from 260 to 210 nm as a function of concentration were measured on a Jasco J-40 spectrometer. Spectra are presented on a per amide basis.

Data Analysis. The vacuum UV CD spectra from 260 to 178 nm were analyzed by using singular value decomposition combined with variable selection as described earlier (21, 22). We used a 26-protein basis set that contains more spectra of all-β proteins. Since our CD secondary structure prediction program is based on different combinations of secondary structures in proteins, it is not particularly well suited for the single secondary structure induced in our peptides. However, Δε(222 nm) × (−10) is a good estimate of the percentage of α-helix, as Fig. 1 shows. Furthermore, although the position of the bands is variable in the CD of a β-strand (23), the overall CD magnitude from minimum to maximum is fairly constant (15–16.5 Δε units). We compare the overall

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: ERE, *Eco*RI endonuclease; CMT, γ-chymotrypsin; ADH, liver alcohol dehydrogenase; TFE, 2,2,2-trifluoroethanol.

Table 1. Equivocal peptide sequences that are predicted to be α-helical but are observed with the underlined portion as β-strand

Sequence	Residues	Notation
Acetyl EWAVVLVAEAKHQ amide	103–115	ERE
Acetyl WGKIQKLKIAKVFK amide	77–90	CMT
Acetyl KVIKCKAAVLWEEKK amide	5–19	ADH

magnitude of our β-strand CD to that of poly(L-lysine) (24) as an alternate estimate of the percent β-structure.

RESULTS AND DISCUSSION

The isolated peptides are randomly structured in aqueous solution, as based on their CD spectra (Fig. 2). However, we have been able to find a variety of solvents that will recover the predicted α-helix secondary structure. Methanol, ethanol, acetonitrile, 1,1,1,3,3,3-hexafluoroisopropanol at low pH, a mixture of octanol and other alcoholic solvents, 2,2,2-trifluoroethanol (TFE), and high concentrations (about 25 mM) of sodium dodecyl sulfate (SDS) all show a typical α-helical CD. We analyzed our CD spectra for α-helix, antiparallel and parallel β-strand, β-turn, and other structure by using the variable selection algorithm and a 26-protein basis set (22). The highest percentage of α-helix was found in 100% TFE, 0°C (Fig. 2 and Table 2), and these values are confirmed by the depth of the 222-nm band. TFE has a reputation for promoting α-helix (25–27). Nevertheless, there are many reports of stable β-strands in TFE (28–31). TFE is a hydrophilic and hydrogen-bonding solvent that stabilizes peptides in the structure expected from the amino acid preferences used to predict secondary structure. It appears to stabilize the secondary structure for which a sequence has propensity.

To check whether the helical structure in each peptide is unimolecular or results in aggregation, CD spectra were measured versus concentration. The dependence of $\Delta\varepsilon$ at 222 nm on concentration is shown (Fig. 3*A*) for the three peptides at 5°C. No dependence on concentration is observed over a 200-fold range for ERE or CMT. However, for ADH the CD changed with concentration, demonstrating that aggregation is part of the solvent system stabilizing the α-helix. As the helix in ERE and CMT unfolds with increasing temperature, the presence of an initial tight association at low temperature should be revealed as a dependence of the unfolding on peptide concentration (32). Neither ERE (Fig. 3*B*) nor CMT (not shown) showed CD that depended on the peptide con-

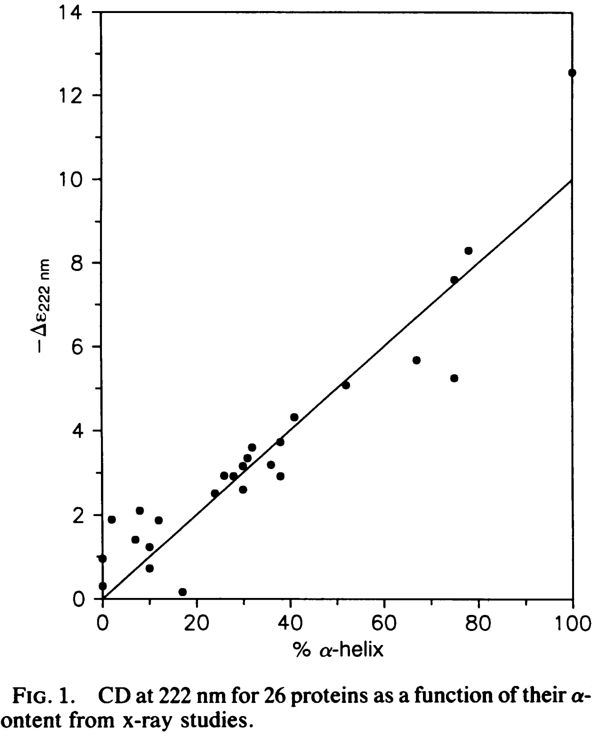

FIG. 1. CD at 222 nm for 26 proteins as a function of their α-helix content from x-ray studies.

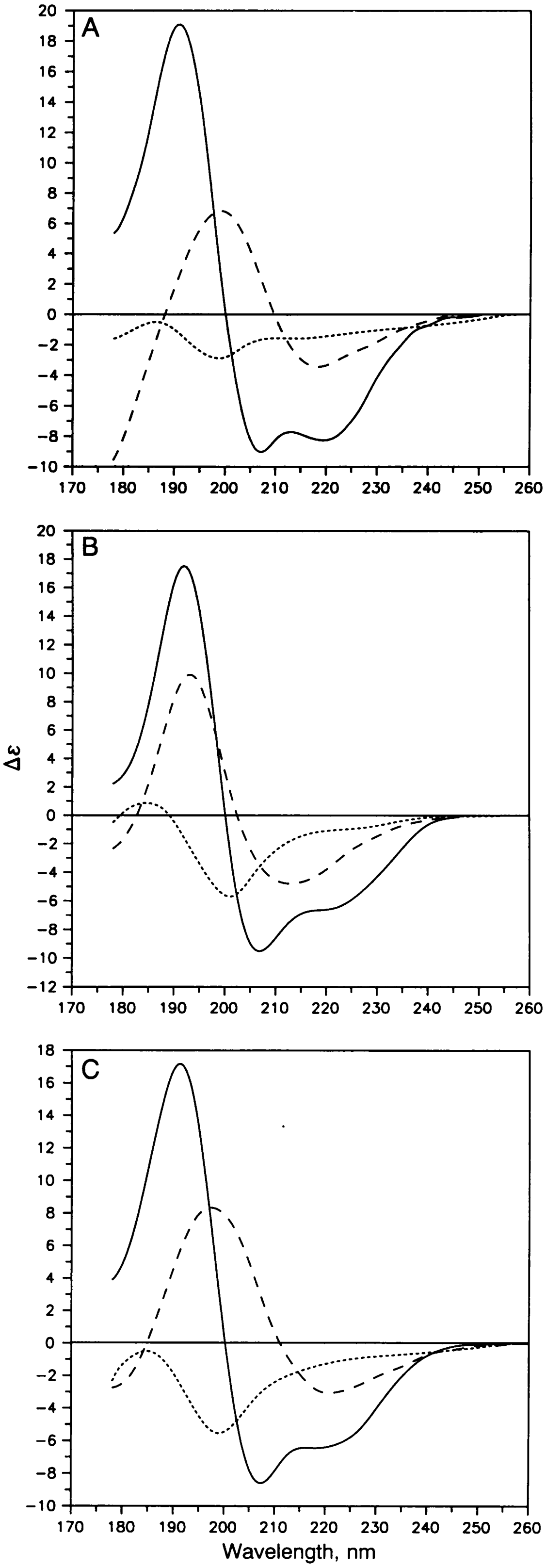

FIG. 2. CD of the equivocal amino acid sequences as a random coil in 10 mM sodium phosphate buffer at pH 7.0 (· · ·), as an α-helix in 100% TFE (——), and as a β-strand in 2–4 mM SDS (– – –). (*A*) ERE. (*B*) CMT. (*C*) ADH.

Table 2. Analysis of secondary structure of CD of equivocal peptide sequences in 100% TFE or 2–4 mM SDS

Peptide	Solvent	H	A + P	T	O	Total
ERE	TFE	0.77 ± 0.01	0.08 ± 0.04	0.15 ± 0.01	0.03 ± 0.01	1.03
CMT	TFE	0.65 ± 0.04	0.12 ± 0.10	0.21 ± 0.07	0.02 ± 0.02	1.03
ADH	TFE	0.66 ± 0.03	0.07 ± 0.07	0.26 ± 0.03	0.00 ± 0.03	0.99
ERE	SDS	0.00 ± 0.00	0.60 ± 0.07	0.01 ± 0.01	0.39 ± 0.04	1.00
CMT	SDS	0.38 ± 0.03	0.51 ± 0.06	0.07 ± 0.03	0.07 ± 0.04	1.03
ADH	SDS	0.16 ± 0.02	0.54 ± 0.07	0.01 ± 0.01	0.31 ± 0.02	1.03

H, α-helix; A, antiparallel β-sheet; P, parallel β-sheet; T, β-turn; O, other secondary structure.

centration at any point in the unfolding, demonstrating that helix formation by these two peptides is not the result of aggregation.

Many of the solvents that one might expect to mimic the environment inside a protein caused precipitation of these equivocal sequences, but low percentages of TFE at pH 11, 0.08% digitonin/0.016% cholate with 10 mM phosphate buffer at pH 7.4, 50 mM octyl glucoside, and 1 M sucrose in 25 mM Mops at pH 7.4 gave CD spectra typical of a β-strand for some sequences. The hydrophobic environment created by SDS at 2–4 mM when the ratio of SDS to peptide is 2:1 to 4:1 consistently gave a large percentage of β-strand (Fig. 2). Here we cannot test for aggregation because the structure depends on SDS concentration and the ratio of SDS to peptide. Analysis of the CD (Table 2) for ERE (25°C), CMT (45°C), and ADH (45°C) gives 60%, 51%, and 54% β-strand, respectively, which is lower than that found in the proteins (86%, 73%, and 56% β-strand). If we compare the overall magnitude of the β-strand CD for our three equivocal sequences to the overall magnitude of poly(L-lysine) as a β-strand, we obtain 62% for ERE, 97% for CMT, and 76% for ADH.

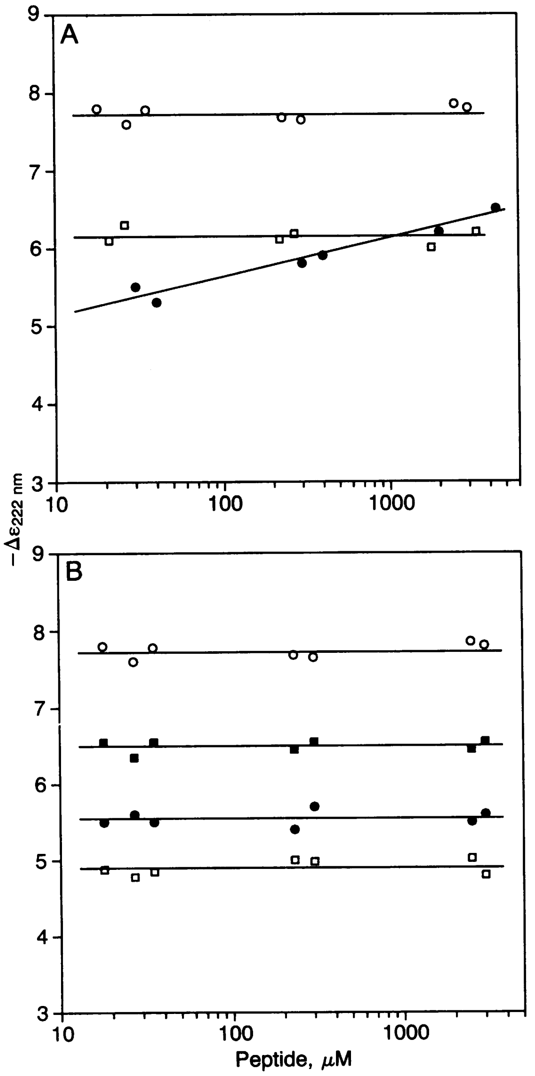

FIG. 3. CD at 222 nm in TFE as a function of concentration. (*A*) Peptides ERE (○), CMT (□), and ADH (●) at 5°C. (*B*) Peptide ERE at 5°C (○), 25°C (■), 45°C (●), and 60°C (□).

SDS is a surfactant that can provide a hydrophobic environment for polypeptides in proteins. At high concentrations it forms micelles, and it is well documented that these conditions usually stabilize α-helical structure (33–38). Yang and coworkers (35, 36) successfully used low concentrations (2–4 mM) of SDS to induce β-structure. We followed their methods and dissolved our peptides in aqueous solution without salt, the solutions being self buffering because of the high concentration of peptide. In the absence of salt the low SDS concentration is far below the critical micelle concentration (39). Our equivocal sequences assume the expected α-helical structure in the hydrophilic solvent TFE. The interior of a protein is usually hydrophobic, so we would expect that a hydrophobic environment would create the β-strand structure for our equivocal sequences, as is found from inspection of the x-ray diffraction for the parent proteins. The nature of this solvent system is unknown, but we do have two to four SDS molecules for each amino acid in the sequence. We presume that the hydrophobic tail of the SDS interacts with the equivocal sequence to mimic the hydrophobic environment found in the interior of the parent protein, whereas the hydrophilic end of the SDS molecule keeps the β-strand in solution.

Our results are consistent with recent work (40), which showed that the inverse protein folding problem can be effectively attacked by finding sequences that are most compatible with the environments of the residues in the three-dimensional structure. Here we show that equivocal sequences can be made to assume the α-helix and the β-strand conformation by proper choice of solvent system. The results indicate that there is a common thread in the behavior of these equivocal sequences, since they all can form a stable α-helix in 100% TFE at 0°C and a β-strand in low concentrations of SDS.

Our results demonstrate that the solvent is in control; therefore schemes that predict secondary structure from primary structure alone can never be totally successful. Tertiary structure must be taken into account so that we know what solvent the peptide sequence effectively sees when protein folding is complete. Since the environment can indeed change the secondary structure, the hierarchic model will have to be modified to take folding feedback into account: (*i*) primary structure determines secondary structure, (*ii*) secondary structures fold into a tertiary structure, (*iii*) some secondary structures change as a result of their new environment, and (*iv*) minor rearrangements occur in the tertiary structure. On the other hand, these results fit in with Dill's nonhierarchic model (41) that involves random condensation and then segment rearrangement and would contribute to folding simulations (42, 43). If we are going to understand how proteins fold, we must understand the environmental effects as well as the sequence effects.

It is a pleasure to acknowledge Ping-Jung Chou for his smoothing program that uses the cubic spline algorithm and Dean Malencik for his help with the HPLC. This work was supported by Public Health Service Grant GM-21479 from the National Institute of General Medical Sciences.

1. Chou, P. Y. & Fasman, G. D. (1978) *Adv. Enzymol.* **47**, 45–148.
2. Burgess, A. W., Ponnuswamy, P. K. & Scheraga, H. A. (1974) *Israel J. Chem.* **12**, 239–286.
3. Lim, V. I. (1974) *J. Mol. Biol.* **88**, 857–894.
4. Garnier, J., Osguthorpe, D. J. & Robson, B. (1978) *J. Mol. Biol.* **120**, 97–120.
5. Pongor, S. & Szaley, A. A. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 366–370.
6. Sweet, R. M. (1986) *Biopolymers* **25**, 1565–1577.
7. Nishikawa, K. & Ooi, T. (1986) *Biochim. Biophys. Acta* **871**, 45–54.
8. Levin, J. M., Robson, B. & Garnier, J. (1986) *FEBS Lett.* **205**, 303–308.
9. Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987) *J. Mol. Biol.* **195**, 957–961.
10. Edelman, J. & White, S. H. (1989) *J. Mol. Biol.* **210**, 195–209.
11. Qian, N. & Sejnowski, T. J. (1988) *J. Mol. Biol.* **160**, 865–884.
12. Holley, L. H. & Karplus, M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 152–156.
13. Manavalan, P., Johnson, W. C., Jr., & Modrich, P. (1984) *J. Biol. Chem.* **259**, 11666–11667.
14. McClarin, J. A., Frederick, C. A., Wang, B.-C., Green, P., Boyer, H. W., Grable, J. & Rosenberg, J. M. (1986) *Science* **234**, 1526–1541.
15. Kim, Y., Grable, J. C., Love, R., Greene, P. J. & Rosenberg, J. M. (1990) *Science* **249**, 1307–1309.
16. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22**, 2577–2637.
17. Cohen, G. H., Silverton, E. W. & Davies, D. R. (1981) *J. Mol. Biol.* **148**, 449–479.
18. Colonna-Cesari, F., Perahia, D., Karplus, M., Eklund, H., Branden, C. I. & Tapia, O. (1986) *J. Biol. Chem.* **261**, 15273–15280.
19. Elwell, M. L. & Schellman, J. A. (1977) *Biochim. Biophys. Acta* **494**, 367–383.
20. Johnson, W. C., Jr. (1988) *Annu. Rev. Biophys. Biophys. Chem.* **17**, 145–166.
21. Hennessey, J. P., Jr., & Johnson, W. C., Jr. (1981) *Biochemistry* **20**, 1085–1094.
22. Manavalan, P. & Johnson, W. C., Jr. (1987) *Anal. Biochem.* **167**, 76–85.
23. Woody, R. W. (1985) *The Peptides* **7**, 15–114.
24. Greenfield, N. & Fasman, G. (1969) *Biochemistry* **8**, 4108–4116.
25. Nelson, J. W. & Kallenbach, N. R. (1986) *Proteins Struct. Funct. Genet.* **1**, 211–217.
26. Nelson, J. W. & Kallenbach, N. R. (1989) *Biochemistry* **28**, 5256–5261.
27. Merutka, G. & Stellwagen, E. (1989) *Biochemistry* **28**, 352–357.
28. Goodman, M., Verdini, A. S., Choi, N. S. & Masuda, Y. (1970) in *Topics in Stereochemistry*, ed. Eliel, E. & Allinger, N. L. (Wiley, New York), pp. 69–166.
29. Balcerski, J. S., Pysh, E. S., Bonora, G. M. & Toniolo, C. (1976) *J. Am. Chem. Soc.* **98**, 3470–3473.
30. Kelly, M. M., Pysh, E. S., Bonora, G. M. & Toniolo, C. (1977) *J. Am. Chem. Soc.* **99**, 3264–3266.
31. Narayanan, U., Keiderling, T. A., Bonora, G. M. & Toniolo, C. (1986) *J. Am. Chem. Soc.* **108**, 2431–2437.
32. Ho, S. P. & Degrado, W. F. (1987) *J. Am. Chem. Soc.* **109**, 6751–6758.
33. Jirgensons, B. (1977) *Biochim. Biophys.* **473**, 352–358.
34. Jirgensons, B. (1981) *Makromol. Chem. Rapid Commun.* **2**, 213–217.
35. Wu, C. C. & Yang, J. T. (1981) *Mol. Cell. Biochem.* **40**, 109–122.
36. Wu, C. C., Ikeda, K. & Yang, J. T. (1981) *Biochemistry* **20**, 566–570.
37. Wu, C. C. & Yang, J. T. (1988) *Biopolymers* **27**, 423–430.
38. Gierasch, L. M. (1989) *Biochemistry* **28**, 923–930.
39. Tanford, C. (1980) in *The Hydrophobic Effect: Formation of Micelles and Biological Membranes* (Wiley, New York), pp. 66–68.
40. Bowie, J., Luthy, R. & Eisenberg, D. (1991) *Science* **253**, 164–170.
41. Dill, K. A. (1990) *Biochemistry* **29**, 7133–7155.
42. Cohen, F. E., Abarbanel, R. A., Kuntz, I. D. & Fletterick, R. J. (1986) *Biochemistry* **25**, 266–275.
43. Skolnick, J. & Kolinski, A. (1990) *Science* **250**, 1121–1125.