Prediction of Protein Helices with a Derivative of the Strip-of-Helix Hydrophobicity Algorithm*

(Received for publication, March 24, 1989)

Victor E. Reyes§, Lisa Phillips‡‖, Robert E. Humphreys‡‖‖, and Robert A. Lew† ‡‡

From the Departments of §Pharmacology, ‡Medicine, and ‖‖Biostatistics, University of Massachusetts Medical School, Worcester, Massachusetts 01655

The strip-of-helix hydrophobicity algorithm was devised to identify protein sequences which, when coiled as α or 310 helices, had one axial, hydrophobic strip and otherwise variably hydrophilic residues. The strip-of-helix hydrophobicity algorithm also ranked such sequences according to an index, the mean hydrophobicity of amino acids in the axial strip. This algorithm well predicted T cell-presented fragments of antigenic proteins. A derivative of this algorithm (the structural helices algorithm (SHA)) was tested for the prediction of helices in crystallographically defined proteins. For the SHA, eight amino acid sequences, 2 cycles plus one amino acid in an α helix, with strip-of-helix hydrophobicity indices greater than 2.5, were selected with overlapping segments joined. These selections were terminated according to simple "capping rules," which took into account the roles of N-terminal Asn or Pro and C-terminal Gly in the stability of helices. In analyses of 35 crystallographically defined proteins with known α and 310 helices, the predictions with the SHA overlapped (had overlap indices x ≥ 0.5) with 34% of known helices, touched (had overlap indices 0.5 > x > 0) or overlapped with 68% of known helices, or were neighboring (came within 6 residues) or touched or overlapped with 82% of known helices. At each level of judging the quality of prediction, the SHA was usually less sensitive (correct predictions/total number of known helices) and more efficient (correct predictions/total number of predictions) than the Chou-Fasman and Garnier-Robson methods. It was simpler in design and calculation. The chemical mechanisms underlying these algorithms appear to apply both to protein folding and to selection of T cell-presented antigenic sequences.

* This work was supported in part by Diabetes and Endocrinology Research Center Grant DK 32530 and National Institutes of Health Grant CA 37645. The costs of publication of this article were defrayed through usually axial, generically hydrophobic surfaces, was described by Kaiser and Kézdy (1). DeLisi and Berzofsky (2) and a fellow of the Juvenile Diabetes Foundation International Summer Research Program at the University of Massachusetts Medical Center.

† Fellow of the Jeane B. Kempner Fund.
‡ A student from the College of the Holy Cross, Worcester, MA, and a fellow of the Juvenile Diabetes Foundation International Summer Research Program at the University of Massachusetts Medical School.

The American Society for Biochemistry and Molecular Biology, Inc.
Vol. 264, No. 22, Issue of August 5, pp. 12854-12858, 1989

© 1989 by The American Society for Biochemistry and Molecular Biology, Inc.

MATERIALS AND METHODS

Program—The program, originally created by Stille (10), was rewritten to display in tabular form ranked SOH hydrophobicity indices for six helical segments/100 amino acids at 2-6 cycles of α and 310 antigens.

The abbreviations used are: MHC, major histocompatibility complex; SHA, structural helices algorithm (for prediction of helices in native proteins); SOH, strip-of-helix; SOHHA, strip-of-helix hydrophobicity algorithm (for the identification of sequences with the potential to coil as an amphipathic helix for T cell presentation).

1 The abbreviations used are: MHC, major histocompatibility complex; SHA, structural helices algorithm (for prediction of helices in native proteins); SOH, strip-of-helix; SOHHA, strip-of-helix hydrophobicity algorithm (for the identification of sequences with the potential to coil as an amphipathic helix for T cell presentation).
helices. Overlapping 8-amino acid α-helical predictions with SOH hydrophobicity indices greater than 2.5 were joined. Termini of these unions or freestanding segments were "capped" or modified according to the following rules derived from the observations of Richardson and Richardson (15). Within the first four N-terminal amino acids, the most C-terminal Asn became the N terminus. Within the first four amino acids, any position -1 or -2 of the position preceding the most C-terminal Pro became the N terminus. Because a proline usually permits hydrogen bond formation between the amide nitrogen of the amino acid preceding it and the carboxyl of the second amino acid following it (16) if that second amino acid is within a predicted helix, the amino acid preceding the proline could be the first member of the proline-induced turn initiating that helix. The first Gly or Pro following an N terminus became the C terminus. In the absence of any modification of a terminus, it stood as selected in the unions or freestanding segments. A final selection had to be at least five amino acids long. This program was written in Fortran and run on the Harris 1000 mainframe of the computing facility of the University of Massachusetts Medical Center. The graphical display of the sequence and the various indicators of segments predicted by each method were generated as a two-dimensional array, facilitating both extensions to more complex displays and portability. Copies of the program are available upon request.

Protein Sequences and Helices—Amino acid sequences were obtained from the Swiss Protein Sequence and the Protein Identification Resource of the National Biomedical Research Foundation data banks. Structural helices were defined to be those used by Richardson and Richardson (15) and for additional proteins as identified by Presta and Rose (17). The Chou-Fasman (18) and Garnier-Robson (19) predictions were found with The Protein Program (Comprehensive Microcomputer Systems for Molecular Biology, DNASTAR Inc., Madison, WI).

Statistical Analyses—A correct prediction was defined in terms of the overlap of the predicted segment with the true segment, with arbitrary definitions of the required amount of overlap. If merely touching the true segment counted as a correct prediction, then a perfect score would result from predicting the entire protein to be one helical segment. Therefore, we defined a match in three ways, one of which (true segment) made that one segment strategy a poor method of prediction. The definitions are nested. In other words, any prediction which overlaps a true structure will also touch and be near.

1) Overlap occurred when the intersection of the true and predicted segments was 50% or more of the union of the two segments. For example, if the true segment occupied sequence positions 1-10 and the prediction occupied positions 4-13, then the union of the two covered positions 1-13, while the intersection covered positions 4-10 (seven positions), and 7 out of 13 exceeded 50%. With this definition, only one prediction could match a true segment. If two predicted segments overlapped with one true helix, then the two predictions must have touched and would have been merged. The other two definitions of a correct prediction did not have this property.

2) Touching of the prediction with the true segment was defined as a nonempty intersection of the two segments less than overlap. Two predictions can touch one true segment.

3) Nearness of the prediction to the true segment was defined as the two segments being at least six amino acids of each other. Again, two predictions can be near a true segment.

We compared the rates of correct predictions for the SHA, Chou-Fasman, and Garnier-Robson methods. The rates were compared by means of the t-test and the χ-square test. For the second and third definitions of a match if several predictions matched a true segment, then only one correct prediction resulted. The extra predictions were not counted as errors. Similarly, when one prediction matched two (or more) true segments, then two (or more) correct predictions were recorded.

RESULTS

Development of the Algorithm—Rather than progressively refining our algorithm to fit a data base, we started with a biological model and then tested how well the predictions with the algorithm derived from the model actually fit a set of established data. Simply put, the model stated that, during protein folding, nucleations of helices could be catalyzed by surfaces which recognized and stabilized the axial, hydrophobic surfaces of those helices. The axial, hydrophobic strips of nucleated helices could either associate against each other through interdigititation of hydrophobic amino acids or bind to hydrophobic regions of a growing protein structure. In addition, the initial nucleation could catalyze the extension of helical structures into adjacent segments of the protein to be preserved when an axial, hydrophobic surface also exists in such evolving helical regions or when other favorable interactions exist. Although the initial nucleating region might be the one with the strongest axial, hydrophobic strip, touching or neighboring helices with weaker axial, hydrophobic strips could also be present. Because analyses of protein structure had reported a high frequency of 2-cycle, α and 3₁₀ helices (20), our basic window for evaluation was eight amino acids (2 cycles + 1 amino acid, or 3 turns) of an α helix. Unions of these short helices could form longer helices. Finally, local structural interactions with Asn or Pro at N termini or C termini and/or interactions with the peptide backbone of the helix could promote stability of the termini, as described by Richardson and Richardson (15) and Presta and Rose (17). These principles were reflected in the algorithm formally presented under "Materials and Methods."
Prediction of Protein Helices

Comparisons of helices which are known, predicted by SHA, Chou-Fasman, and Gamier-Robson methods, or are T cell-presented

| Protein                  | Known† | Predicted                  | T cell-presented† |
|--------------------------|--------|----------------------------|-------------------|
| Hen egg lysozyme         |        |                            |                   |
| 4-16                     | 1-27   | 9-16                       | 1-4               |
| 24-37                    | 45-50  | 23-31                      | 12-19             |
| 86-102                   | 69-76  | 46-52                      | 23-33             |
| 108-116                  | 92-101 | 111-116                    | 49-53             |
| 106-113                  | 123-128| 107-113                    | 108-119           |
| 138-145                  |        | 126-130                    |                   |
| Sperm whale myoglobin    | 4-17   | 21-28                      | 1-32              |
| 21-35                    | 61-75  | 35-64                      | 38-115            |
| 37-42                    | 107-114| 69-116                     | 124-149           |
| 52-56                    | 131-138| 122-149                    |                   |
| 89-95                    |        |                            |                   |
| 101-118                  |        |                            |                   |
| 125-148                  |        |                            |                   |
| Porcine insulin          | A2-A8  | A3-A13                     | A10-A17           |
| 19-26                    | A13-A18| A18-B2                     | B9-B14            |
| 31-40                    | B9-B13 | B7-B12                     |                   |
| Staphylococcal nuclease  | 54-67  | 19-26                      | 5-17              |
| 99-106                   | 31-40  | 21-26                      | 19-23             |
| 122-134                  | 69-73  | 48-53                      | 39-47             |
| 93-100                   | 56-76  | 56-81                      |                   |
| 105-115                  | 94-110 | 94-109                     | 80-97             |
|                            | 120-139| 120-142                    | 91-110            |

† Crystallographically defined helices data are compiled from Refs. 15, 17, and 28.
† T cell-presented peptides data are compiled from Refs. 21-27.
† T cell-presented peptides which were predicted by the SHA but were not crystallographically defined helices.

Analysis of comparisons of predicted to known helices

| Helices                  | Known | SHA | Chou-Fasman | Garnier-Robson |
|--------------------------|-------|-----|-------------|----------------|
| Total                    | 146   | 180 | 243         | 194            |
| Median number/protein    | 4     | 5   | 6           | 4              |
| Mean number/protein (± S. D.) | 4.1 ± 2.5 | 5.1 ± 2.4 | 6.9 ± 3.9 | 5.5 ± 5.1 |
| Mean length (± S. D.)    | 12.0 ± 9.0 | 10.2 ± 4.6 | 10.4 ± 6.6 | 10.8 ± 11.6 |
| Predictions              |       |     |             |                |
| Overlapping              | 49    | 60  | 39          |                |
| Touching + overlapping   | 97    | 120 | 100         |                |
| Neighboring + touching + overlapping | 119 | 128 | 110 |                |
| Wrong                    | 61    | 115 | 84          |                |
| Missed                   | 27    | 18  | 36          |                |
| Sensitivity†             |       |     |             |                |
| Overlapping              | 0.34  | 0.41| 0.27        |                |
| Touching + overlapping   | 0.66  | 0.82| 0.68        |                |
| Neighboring + touching + overlapping | 0.82 | 0.88 | 0.75 |                |
| Efficiency†              |       |     |             |                |
| Overlapping              | 0.27  | 0.25| 0.21        |                |
| Touching + overlapping   | 0.53  | 0.49| 0.52        |                |
| Neighboring + touching + overlapping | 0.66 | 0.53 | 0.57 |                |

† Correct predictions/number of known helices.
† Correct predictions/number of predictions.

touching or neighboring became overlapping or touching, respectively. Also, previously missed relationships became predicted. However, such a lower threshold also resulted in more wrong predictions. Overall, the lower threshold did not improve the efficiency or sensitivity of this approach.

Prediction of T Cell-presented Sequences—For four proteins in this study, some T cell-presented peptides had been experimentally determined. Such identifications did not come from exhaustive surveys of all peptides to be presented by multiple MHC alleles but, rather, from studies of principal determinants recognized in one or a few mouse strains (21–27). Nevertheless, several peptides which were T cell-presented
were found to originate in helices and were predicted by the SHA (Table 1). T cell-presented staphylococcal nucleic peptides 11–30, 66–78, and 89–97 were predicted to be structural helices with high SOH hydrophobicity indexes but were not helices in the native protein. This observation supported the view that binding to class I or class II MHC desetopes depends on the potential of a peptide to coil as a helix with an axial, hydrophobic strip after proteolytic excision from a protein. Alternatively, such “potential” helical peptides acquire their helical conformation while in the presence of MHC molecules.

**DISCUSSION**

The SOH hydrophobicity index of the SOHHA predicted in a ranked fashion, T cell-presented sequences in antigenic proteins (10, 14). A derivative of this procedure (with the addition of simple rules to cap termini), the SHA, also predicted many helices found in proteins of known crystallographic structure. In comparison to the helix-predicting methods of Chou-Fasman (18) and Garnier-Robson (19), the SHA predicted fewer helices at an overlapping or touching level of quality than did the Chou-Fasman algorithm. The SHA is also computationally simpler than the methods of Chou-Fasman, Garnier-Robson, Eisenberg-Terwilliger (29), Cornette et al. (30), and Finer-Moore and Stroud (31).

The fact that many of our predictions overlap or touch known helices supports the hypothesis that a generic feature of the axial, hydrophobic strip is a propensity to catalyze helix formation. In studies of the folding of cytochrome c by Roder et al. (32), nucleation of helices at the N and C termini was indicated by early protection of protein backbone amide protons from deuterium exchange, presumably by hydrogen bonding in a helix. Other than for retention of some residual structure about the heme group in the denatured form of cytochrome c, those helices were the earliest structural elements to demonstrate deuterium exchange protection. The N- and C-terminal helices made contact with each other about Glyx and Tyrx, thereby stabilizing the intermediate and leading to a general condensation of the remainder of the molecule. A similar study of the folding of ribonuclease S by Udgaonkar and Baldwin (33) demonstrated early protection of β-sheet amide protons. The core of ribonuclease is composed of two such sheets without helices. In another study of bovine pancreatic trypsin inhibitor, Os and Kim (34) demonstrated (through analysis of both synthetic peptides and the protein) that a C-terminal α helix, a central antiparallel β-sheet, and the hydrophobic core between them were stabilized by the formation of a disulfide bond between residues 30 and 51 in the β-sheet and α helix, respectively. Our finding that putative axial, hydrophobic strips predicted structural helix formation supports the view that distinct intermediates with local helical structure stabilized by hydrophobic interactions and, in some instances, by disulfide bridge formation can form in the early time course of protein folding.

Not all predicted helices with strong axial, hydrophobic strips were found to be actual helices in the native proteins. This observation was consistent with the potential extension or migration of newly formed helices to adjacent regions with axial, hydrophobic strips. The finding of known helices flanked by touching and/or neighboring predictions supported this view. In attempting to design a migration indicator based on analysis of axial, hydrophobic strips with SOH indexes 2.5 > x > 2.0, we found that although some helices improved their quality of prediction, e.g., going from the touching to overlapping category, other helices did not improve and some wrong selections were made. In general, we would not decrease the stringency of initial selections to include putative helices with SOH hydrophobicity indices less than 2.5. Nevertheless, because the SOH index is a continuum of values, an investigator focusing on a single protein, for example, in the context of determining a crystallographic structure, might wish to explore such additional calculations for clues which could be evaluated in the light of an evolving structural model.

The finding that a method to predict helices in proteins also identified T cell-presented epitopes relates to the mechanism of antigen processing and presentation. DeLisi, Bartzosky, and colleagues (2, 35–37) initially found that amphipathicity correlated to potential for T cell presentation in sequences of antigenic proteins. Kaiser and Kézdy (1) observed that amphipathicity was a generic feature leading to the binding of many polypeptide hormones to their receptors or lipid bilayers. The β-plated sheet floors of the MHC class I desetope and of the deduced MHC class II desetope are composed of hydrophobic or uncharged amino acids with allelic variations conserving the hydrophobic character of the floor with few exceptions (only in positions 114 and 116 of HLA A2 and 29 of Eβ (11, 12)). The principal allelic-nonspecific force in binding of T cell-presented peptides could be hydrophobic interaction of the axial, hydrophobic strip with the hydrophobic floor of the desetope. It is not known whether, in addition, the hydrophobic strip catalyzes folding of the T cell-presented peptide so that the resulting helical dipole orients and attracts the peptide to the helices binding the desetope or permits scavingning of digested peptides by molecules, which could protect antigenic epitopes from further proteolysis and/or catalyze their transfer to desetopes (38). The view that helical coiling of digested peptides is a step in processing of T cell antigenic sequences is supported by the finding that some sequences predicted both with this structural helices derivative of the SHA and with the parental SOHHA algorithm are T cell-presented, whereas, in fact, they are actually not helices in the native protein.

Acknowledgments—We thank Dr. Larry Hardy for discussions of this work and Erik Skinner for assistance with the programming. We thank Marion Bonin for preparation of this manuscript.

**REFERENCES**

1. Kaiser, E. T., and Kézdy, F. J. (1984) Science **233**, 249–255
2. DeLisi, C., and Bartzosky, J. A. (1985) **Proc. Natl. Acad. Sci. U. S. A.** **82**, 7048–7052
3. Novotny, J., Handschumacher, M., and Brucoleri, R. E. (1987) **Immunol. Today** **8**, 26–31
4. Thomas, L. J., Nguyen, Q. V., Elliott, W. L., and Humphreys, R. E. (1988) *J. Immunol.* **140**, 2670–2674
5. Blum, J. S., and Cresswell, P. (1988) **Proc. Natl. Acad. Sci. U. S. A.** **85**, 3975–3979
6. Nguyen, Q. V., Knapp, W., and Humphreys, R. E. (1989) **Hum. Immunol.** **24**, 153–163
7. Nguyen, Q. V., and Humphreys, R. E. (1989) *J. Biol. Chem.* **264**, 1631–1637
8. Stockinger, B., Pessara, U., Lin, R. H., Habicht, J., Crez, M., and Koch, N. (1988) *Cell* **56**, 683–689
9. Elliott, W. L., Stille, C. J., Thomas, L. J., and Humphreys, R. E. (1987) *J. Immunol.* **138**, 2949–2952
10. Stille, C. J., Thomas, L. J., Reyes, V. E., and Humphreys, R. E. (1987) *Mol. Immunol.* **24**, 1021–1027
11. Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L., and Wiley, D. C. (1987) *Nature* **329**, 512–518
12. Brown, J. H., Jardetzky, T., Saper, M. A., Samraoui, B., Bjorkman, P. J., and Wiley, D. C. (1988) *Nature* **332**, 845–850
13. Gorga, J. C., Dong, A., Manning, M. C., Woody, R. W., Caughey, W. S., and Strominger, J. L. (1989) *Proc. Natl. Acad. Sci. U. S. A.* **86**, 2291–2295
14. Reyes, V. E., Chiu, L. T., and Humphreys, R. E. (1988) *Mol. Immunol.* **25**, 867–871
Prediction of Protein Helices

15. Richardson, J. S., and Richardson, D. C. (1988) Science 240, 1648-1652
16. Dwyer, H. J., Cross, K. J., Houghten, R. A., Wilson, I. A., Wright, P. E., and Lerner, R. A. (1985) Nature 318, 480-483
17. Presta, L. G., and Rose, G. D. (1988) Science 240, 1632-1641
18. Chou, P. Y., and Fasman, G. D. (1978) Adv. Enzymol. Relat. Areas Mol. Biol. 47, 45-147
19. Garnier, J., Osguthorpe, D., and Robson, B. (1978) J. Mol. Biol. 120, 97-120
20. Kabsch, W., and Sander, C. (1983) Biopolymers 22, 2577-2637
21. Berkower, I., Buckenmeyer, G. K., and Berzofsky, J. A. (1986) J. Immunol. 136, 2498-2503
22. Cease, K. B., Berkower, I., York-Jolley, J., and Berzofsky, J. A. (1986) J. Exp. Med. 164, 1779-1784
23. Hackett, C. J., Dietzschold, B., Gerhard, W., Ghrist, B., Knorr, R., Gillessen, D., and Melchers, F. (1983) J. Exp. Med. 158, 294-302
24. Hurwitz, J. L., Heber-Katz, E., Hackett, C. J., and Gerhard, W. (1984) J. Immunol. 133, 3371-3377
25. Lamb, J. R., Eckels, D. D., Lake, P., Woody, J. N., and Green, N. (1982) Nature 300, 66-69
26. Sette, A., Buus, S., Colon, S., Miles, C., and Grey, H. M. (1989) J. Immunol. 142, 35-40
27. Lakey, E. K., Casten, L. A., Anderson, M. S., Smolenski, L. A., Smith, J. A., Margoliash, E., and Pierce, S. K. (1987) Eur. J. Immunol. 17, 1605-1609
28. Tucker, P. W., Hazen, E. E., Jr., and Cotton, F. A. (1979) Mol. Cell. Biochem. 23, 67-85
29. Eisenberg, D., Weiss, R. M., and Terwilliger, T. C. (1984) Proc. Natl. Acad. Sci. U.S.A. 81, 140-144
30. Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., and DeLisi, C. (1987) J. Mol. Biol. 110, 859-885
31. Finer-Moore, J., and Stroud, R. M. (1984) Proc. Natl. Acad. Sci. U.S.A. 81, 155-159
32. Roder, H., Elóe, G. A., and Englander, S. W. (1988) Nature 335, 700-704
33. Udgaonkar, J. B., and Baldwin, R. L. (1988) Nature 335, 694-699
34. Oas, T. G., and Kim, P. S. (1988) Nature 336, 42-48
35. Berzofsky, J. A. (1988) Immunol. Lett. 18, 83-92
36. Margalit, H., Spouge, J. L., Cornette, J. L., Cease, K. B., DeLisi, C., and Berzofsky, J. A. (1987) J. Immunol. 138, 2213-2229
37. Spouge, J. L., Guy, H. R., Cornette, J. L., Margalit, H., Cease, K., Berzofsky, J. A., and DeLisi, C. (1987) J. Immunol. 138, 204-212
38. Sorli, C. H., and Humphreys, R. E. (1988) in Immunobiology of HLA: Immunogenetics and Histocompatibility (DuPont, B., ed) Vol. 2, pp. 361-362, Academic Press, New York