PREDICTION OF THE SECONDARY STRUCTURE OF MOUSE NERVE GROWTH FACTOR AND ITS COMPARISON WITH INSULIN

Patrick Argos
Department of Physics
Southern Illinois University
Edwardsville, Illinois 62026

Received March 23,1976

Summary Five secondary structure prediction methods based on amino acid sequence have been used to predict the secondary structure of mouse nerve growth factor (NGF). The regions predicted helical do not correlate well with the proposal, based on the alignment of primary sequences, that the NGF peptide chain is structurally and evolutionarily related to proinsulin.

INTRODUCTION

Frazier et. al. (1,2) have proposed that mouse nerve growth factor (NGF) consists of peptide units joined together in a single chain of 118 amino acids that are structurally and evolutionarily related in a sequential manner to the insulin B chain, proinsulin C peptide, and insulin A and B chains. This scheme is based on the alignment of the primary sequence of these proteins and on the reactivity of selected residues in NGF. The parts of the NGF molecule supposedly corresponding to the A and B peptides of insulin should then be in a conformation similar to insulin (3). Predictions of the tertiary structure of the C chain in proinsulin indicate the presence of two regions of α-helix (4). Since the crystallographic studies of NGF are not yet complete (5), methods to predict secondary structure were applied to the NGF primary sequence to determine if helices are appropriately predicted in regions homologous with insulin.

Several schemes have been developed to provide rules which enable probable locations of secondary structure to be determined from a knowledge of the primary sequence. For adenylate kinase (6) and T4 phage lysozyme (7), the structure predicted by several of the methods has been compared to that determined by x-ray crystallographic techniques. While no individual prediction scheme was clearly superior, the "joint prediction histogram," obtained by summing the individual predictions for each amino acid, generally agreed well

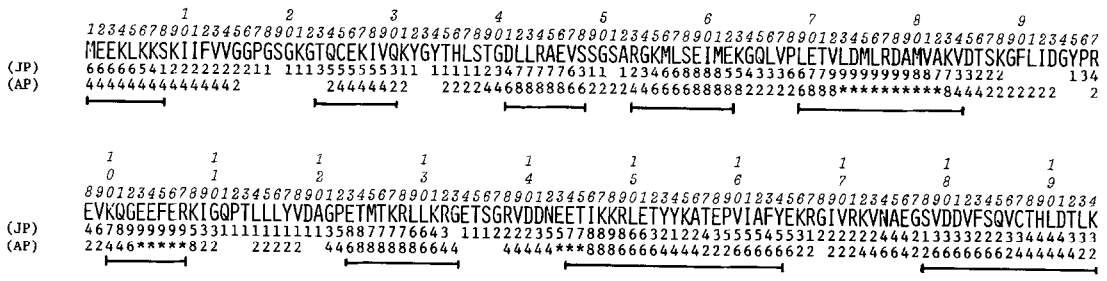

FIGURE 1. Helical regions in adenylate kinase.
(JP) Results of predictions published previously (6). The numbers correspond to the number of methods (out of a total of 10) predicting that a given residue is in an α-helix.
(AP) Results of the automated predictions described here. The numbers were obtained by doubling the prediction score (maximum 5) to put it on the same scale as (JP). Asterisk denotes 10.
(Heavy Line) Regions of α-helix as experimentally determined.

with the observed helical structures. Since sheet and bend predictions agreed less well with those experimentally determined, this study was limited to a comparison of the observed helical regions in insulin with those predicted for proinsulin and for mouse NGF.

METHODS

Five of the various predictive methods have been fully automated and computer programmed to obtain a joint prediction histogram. The techniques include those of Nagano and Hasegawa (8,9), Burgess et al. (10), Barry and Friedman (6), Kabat and Wu (11), and Chou and Fasman (12,13). Several of the other methods (7) were not included due to unavailability of computer programs or difficulty in programming the complex rules of the technique.

Computer programs were obtained from Nagano, Burgess, and Barry. The Nagano program tends to overpredict while that of Burgess underpredicts. Barry's program results require considerable judgment in assessing beginning and ending probabilities for helices. In the former two cases, the program predictions were accepted without modification. In the latter case a routine was written to automate judgment on starting and terminating helices; it predicted to within 90% Barry's final helical prediction for T4 phage lysozyme (7). The methods of Chou and Fasman was programmed by adhering to their prediction rules as nearly as possible (12). The technique of Kabat and Wu predicts peptide regions as permissable, but not necessarily helical; in the program written to apply their method, a chain section of four or more residues in sequence considered possibly helical was assigned helical. In no case was an attempt made to modify automated predictions as is typically done by the predictors themselves. Despite elimination of this evaluation process, Figure 1 clearly demonstrates for adenylate kinase that the joint automated prediction of helical structures based on only five techniques used here is virtually equivalent to the joint helical prediction made by ten methods and submitted by the predictors themselves (6).

Several criteria have been used to indicate the accuracy of the predictions: the precentage of the observed helical residues ($\%_\alpha$) predicted correctly (13); the precentage of non-helical residues ($\%_{n\alpha}$) predicted correctly (13); the mean of the former two precentages (Q_α); and the correlation (C_α) between prediction and observation (7, 14). A perfect prediction would yield Q=100% and $C_\alpha = 1$. If $C_\alpha = 0$, the prediction is considered random; $C_\alpha = -1$ indicates total disagreement between prediction and observation. Using the criterion that a residue is assigned helical if two or more of the five automated schemes or three or more of the ten submitted techniques designate the amino acid helical, the agreement factors for the adenylate kinase prediction are $\%_\alpha = 91.7\%$, $\%_{n\alpha} = 72.1\%$, $Q_\alpha = 81.9\%$, and $C_\alpha = 0.657$ using the five programmed methods and are $\%_\alpha = 89.8\%$, $\%_{n\alpha} = 72.1\%$, $Q_\alpha = 80.9\%$, and $C_\alpha = 0.631$ for the ten non-automated techniques.

RESULTS OF PREDICTIONS FOR INSULIN, PROINSULIN AND NGF

Five automated predictions for the presence of α-helical regions in three proinsulins and four insulins are summarized in Figure 2. The segments underlined with continuous lines correspond to the regions of α-helix observed in the x-ray structure of pig insulin (3). The helix present in the B chain of insulin (B9-B19) is predicted very well in all seven molecules. All residues belonging to this helix score at least two predicting methods on the joint probability histogram, except for the terminal cysteine in the guinea pig insulin which scores one. The scores within the B chain outside the helical region are generally zero or one. The second helix in the A chain (A13-A19) can be assigned with little difficulty, with scores of two or three within the helix and with at most two residues of uncertainty at the N and C terminal points. Amino acids of the helix between residues A2-A8 generally score two in proinsulins and chicken insulin, but are missed in mouse and guinea pig insulins. Probable causes for the weak prediction are the termination effect in a short sequence (as evidenced by the improvement when B and A chains are not separated), difficulty in assigning two consecutive cysteines as helical, and the distorted nature of the observed helix in pig insulin (3).

The best possible helical predictions from the automated histogram for insulin structures is obtained using the following rule: if a sequence of three or more residues each scored two or more predictions as helical, the residue run is assigned helical. This format was used for all helical pre-

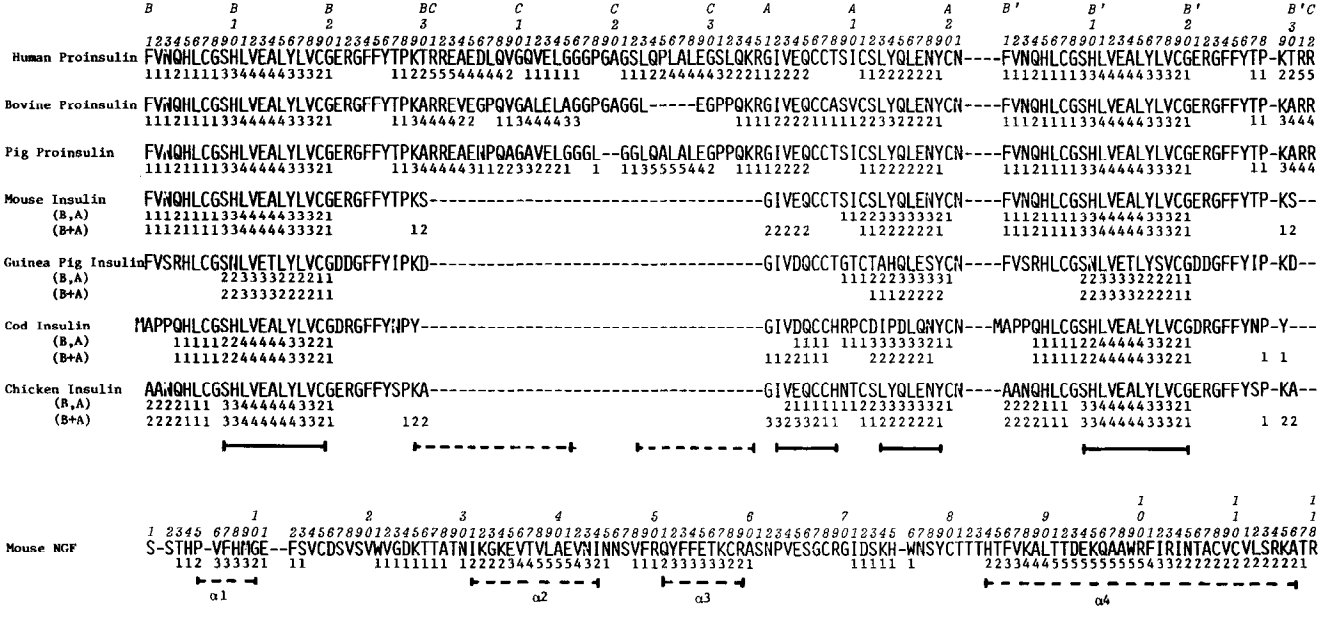

FIGURE 2. Prediction of the helical regions in three proinsulins, four insulins and nerve growth factor. The chains were aligned as in (1). The number under each residue represents the number of methods predicting it to be helical (out of a maximum of 5). For insulins, (B,A) means the prediction for separate chains, (B+A) for a single chain containing both peptides. The regions of α-helix in insulin observed in the x-ray structure (3) are underlined with continuous lines, those predicted for NGF and proinsulin by two or more methods by broken lines.

Table 1. The agreement factors between the observed and predicted helical regions in various insulins. The observed helical structure for all insulins and the insulin A & B segments of proinsulins was assumed to be that observed for pig insulin (3). The factors given for NGF indicate the correlation between the helical structures predicted from the joint histogram and those assumed from the NGF-insulin primary sequence homology (NGF residue segments 1-26 and 62-118). For all cases, the C peptide was not included in the agreement factor calculations.

Polypeptide	%α	%nα	Qα(%)	Cα
Human Proinsulin	84	100	92	0.85
Bovine Proinsulin	84	92	88	0.77
Pig Proinsulin	84	100	92	0.85
Mouse Insulin (B+A)	84	96	90	0.81
Guinea Pig Insulin (B+A)	64	100	82	0.69
Cod Insulin (B+A)	72	96	84	0.70
Chicken Insulin (B+A)	88	81	84	0.69
Nerve Growth Factor (NGF)	41	59	50	0.00

dictions given in the present paper. Table 1 lists the agreement factors for insulins and the insulin A and B segments of proinsulins shown in Figure 2. In all cases the pig insulin helical structure was assumed for the homologous regions in proinsulins and other insulins. This seems reasonable given that hagfish insulin (15) has been shown to have a structure similar to pig insulin. The average correlation C_α is 0.77, indicating excellent helical predictions (16).

Two or three helices are predicted in the C segments of various proinsulins, for which there is no x-ray structure. These helical regions generally agree with those predicted for the isolated C chain (4). However, the first region involves the two final amino acids of the B chain and extends between residues B29-C14 in pig proinsulin, B29-C16 for bovine, and B29-C9 in human. Since the B and C chains of proinsulin are connected, the extension of this helix seems significant. The second helical region in the C chain can be located in the sequence C24-C30, in good agreement with (4).

NGF was aligned with human proinsulin, as previously postulated (1), with appropriate deletions and the repeat of the B chain. Four helical regions indicated α1 to α4 in Figure 2 can be assigned to the NGF sequence. The predictions would place them as side chains 5-10, 31-44, 51-59, and 84-116.

DISCUSSION

The excellent agreement (Table 1) between the helical regions predicted for various insulins and the experimental data suggest that the use of the prediction schemes for proinsulin and NGF is warranted. The predictions for NGF indicate strongly four helical regions while six are predicted for the appropriately aligned proinsulin molecules with the exception of the bovine case in which one helix of the C chain is deleted.

Two helices belonging to the A chain of proinsulin do not have corresponding helices in the NGF sequence, implying that the secondary structure of this region of the two proteins bears little resemblance. Furthermore, the NGF α1 helix does not line up well with the corresponding B9-B19 helix in proinsulin, with only four side chains aligned. The proinsulin B'9-B'19 helix fits within the α4 helix of NGF; however, the latter is much longer. Thus the two segments in NGF corresponding to the B chains of insulin appear very unlike each other, one being strongly helical and the other containing relatively little helix. Table 1 lists the agreement factors between the helical regions predicted from the NGF probability histogram and those assumed from the NGF-insulin primary sequence homology which encompasses NGF residues 1-26 and 62-118. The correlation coefficient C_α was 0.00 which indicates a random correspondence and compares poorly with the average insulin C_α of 0.77. Thus, the predicted secondary structure of NGF does not support the hypothesis (1,2) that NGF and insulin (or proinsulin) have evolved from a common ancestral gene.

The best apparent agreement between helical regions in proinsulin and NGF can be seen in the C-peptide for which no structural data are available. The NGF α3 helix falls completely within the second helix of the proinsulin C chain. Eleven of the fourteen NGF α2 residues line up with the sixteen-residue B29-C14 helix of human proinsulin.

Frazier et al. (1) have observed a number of similarities in primary sequence between human proinsulin, guinea pig insulin, and NGF; however, these are based on residues which are quite variable in other insulins. This lack of exact correspondence between NGF and insulin primary sequences is exemplified at NGF Lys 25, Phe 86, Val 87, and Lys 115, as seen in Figure 2. Furthermore, a reasonable number of residues conserved in most insulins do not have NGF counterparts; for example, NGF Trp 21 and insulin Phe B-25 or Trp 76 and Leu A-16. Once again, the postulated hypothesis of NGF-proinsulin correspondence is not supported.

Given this study, predicted secondary structures may be a generally useful adjunct to the comparison of distant proteins based solely on primary sequence. Application of computerized prediction schemes make such a task quite easy.

ACKNOWLEDGMENTS The author is grateful to Drs. K. Nagano, W. Burgess, and D. Barry for providing computer programs. Mr. John Schwartz and Mr. James Schwartz of Southern Illinois University provided invaluable help in writing programs as well as assembling them into a working unit. The Southern Illinois University computing staff gave considerable technical assistance.

REFERENCES

1. Frazier, W.A., and Hogue-Angeletti, R.A. (1972) Science, 176, 482-488.
2. Frazier, W.A., Hogue-Angelettie, R.A., Sherman, R., and Bradshaw, R.A. (1973) Biochemistry, 12, 3281-3293.
3. Blundell, T., Dodson, G., Hodgkin, D., and Mercola, D. (1972) Advan. Protein Chem., 26, 279-402.
4. Snell, C.R., and Smyth, D.G. (1975) J. Biol. Chem. 250, 6291-6295.
5. Wlodawer, A., Hodgson, K.O., and Shooter, E.M. (1975) Proc. Nat. Acad. Sci. USA, 72, 777-779.
6. Schulz, G.E., Barry, C.D., Friedman, J., Chou, P.Y., Fasman, G.D., Finkelstein, A.V., Lim, V.I., Ptitsyn, O.B., Kabat, E.A., Wu, T.T., Levitt, M., Robson, B., and Nagano, K. (1974) Nature, 250, 140-142.
7. Matthews, B.W. (1975) Biochem. and Biophys. Acta., 405, 442-451.
8. Nagano, K. (1973) J. Mol. Biol., 75, 401-421.
9. Nagano, K. (1974) J. Mol. Biol., 84, 337-372.
10. Burgess, A.W., Ponnuswamy, P.K., and Scheraga, H.A. (1974) Isr. J. Chem., 12, 483-494.
11. Wu, T.T., and Kabat, E.A. (1971) Proc. Nat, Acad. Sci. USA, 68, 1501-1506.
12. Chou, P.Y., and Fasman, G.D. (1974) Biochemistry, 13, 211-222.
13. Chou, P.Y., and Fasman, G.D. (1974) Biochemistry, 13, 222-245.
14. Fisher, R.A. (1958) Statistical Methods for Research Workers, 13th ed., p. 183, Hafner, New York.
15. Cutfield, J.F., Cutfield, S.M., Dodson, E.J., Dodson, G.G., and Sabesan, M.N. (1974) J. Mol. Biol., 87, 23-30.
16. Kabat, E.A., and Wu, T.T. (1974) Proc. Nat. Acad. Sci. USA, 71, 4217-4220.