HUMAN α-CHAIN GLOBIN MESSENGER: PREDICTION OF A NUCLEOTIDE SEQUENCE

Brian Laux, Don Dennis, and Harold B. White, III*

Department of Chemistry, University of Delaware, Newark 19711

Received August 3, 1973

Summary: All nucleotide sequences consistent with the amino acid sequence of the α-chain of human hemoglobin were tested for their complementarity with a known 26-nucleotide sequence from the α-chain messenger. The region with the highest pairing potential is immediately adjacent to the known nucleotide sequence. The existence of this potential hairpin loop requires the specification of nucleotides in at least 6 degenerate positions.

Secondary structure is a characteristic feature of tRNA's[1] and rRNA[2]. The degeneracy of the genetic code allows the possibility for extensive secondary structure in mRNA's as well[3-6]. Certainly for the MS2 coat protein messenger, secondary structure is very evident[7]. Whether such base-paired messengers are favored generally by natural selection is yet to be determined.

Double-standed RNA inhibits protein synthesis in rabbit reticulocytes by binding to initiation factor 3(IF-3)[8]. This suggests that IF-3 normally binds to a double-standed region in mRNA. The hyperchromicity exhibited by rabbit α-chain globin messenger indicates that secondary structure is present in this molecule[9]. We have predicted two possible hairpin loops for the portion of the human α-chain messenger coding for the C-terminal region of the protein[5]. These predictions were made possible by the amino acid sequence of the

*To whom inquiries should be addressed.

chain-termination variant, Constant Spring[10]. The recent discovery of the frame-shift deletion variant, Wayne[11], has permitted the unambiguous assignment of 26 consecutive nucleotides in this region (W. P. Winter, personal communication). This nucleotide sequence is incompatible with both of the predicted structures[5], but is in striking agreement with a third potential hairpin loop (Fig. 1).

This loop represents the structure which provided maximal base-pairing when the deduced 26-nucleotide sequence was checked for complementarity with all regions of the mRNA coding for protein. All possible nucleotide sequences compatible with the amino acid sequence of human α-chain globin were considered. This selection was performed by computer using a pattern recognition algorithm which retains only base-pairing possibilities which would contribute to stable secondary structure[12].

Several features of this proposed loop should be noted:

a. The region that provides maximum complementarity with the 26-nucleotide sequence is a segment of the mRNA immediately adjacent to that known sequence. The hairpin loop which this creates is intrinsically more stable than a structure with complementary segments remote from each other. (The region with the second-best pairing potential codes for Leu 80 to Leu 86 and is located 156 nucleotides away from the known 26-nucleotide sequence.)

b. The size and stability of this hairpin loop is comparable to those in the MS2 RNA[7]. We have

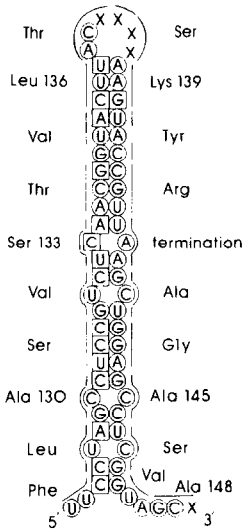

Fig. 1. Predicted hairpin loop for the C-terminal region of human α-chain globin. Circles indicate known nucleotides. Boxes indicate nucleotides in degenerate positions which have been selected to maximize base-pairing. Unspecified nucleotides are indicated by X. The amino acid sequence beyond the termination codon is that of the chain-termination variant, Constant Spring[10].

The nucleotide sequence from Tyr 140 to Val 147 was deduced by a comparison of the amino acid sequences of the Constant Spring variant and the frame-shift deletion variant, Wayne[11]. The Wayne variant can be explained by a deletion of any one of the three nucleotides coding for Lys 139 or either of the last two nucleotides of Ser 138. In order to maximize base-pairing, we have assumed that the Wayne variant results from a deletion of G from the Lys 139 codon. (The other possible deletions for the Wayne variant require an A in this position.) The first nucleotide of the Leu 136 codon is likely to be C since the hemoglobin variant Bibba (Leu 136α→Pro)[16] would then be possible with a single nucleotide substitution.

calculated the stabilization energy of this loop to be -17.6 kcal/mole[13] or -10.8 kcal/mole[14] depending upon the method of calculation used. The

loop as shown in Fig. 1 is 25-nucleotide pairs long; there are four interior loops but no bulge loops[14]. The codons are in 2:2 registry[5].

It is conceivable that the hairpin could start at the third nucleotide of Ala 130 and end at the first nucleotide of Ala 145. In this case the loop would be 19 nucleotides long. Likewise, it is possible that the hairpin could start at the third nucleotide of Ser 133 and end at the first nucleotide of the termination codon. While the longest hairpin is slightly more stable theoretically, there may be other factors which would favor these shorter hairpin loops.

c. The loop in Fig. 1 permits the prediction of 15 nucleotides in degenerate positions. The probability of the random occurrence of these nucleotides in the positions indicated is less than 10^{-6}.

d. The proposed loop codes for the C-terminal region of the α-chain globin, a region whose amino acid sequence is rather constant among vertebrates[1]. It may be that a stable segment of a messenger contributes to the retention of a particular amino acid sequence.

The sequencing of human globin messengers now in progress[15] will confirm or refute our predicted sequence. If our prediction is confirmed, it is highly likely that a hairpin loop is present. Furthermore, it would suggest that secondary structure is also important for nonviral mRNA. The methods used in this paper combined with additional selection

rules may then make it possible to predict messenger sequences from the many protein sequences available[1].

We thank Drs. Jean S. White and William P. Winter for their comments and criticisms. This work was supported by the University of Delaware.

REFERENCES

1. Dayhoff, M. O., Atlas of Protein Sequence and Structure, Vol. 4. National Biomedical Research Foundation, Silver Spring, Md. 543 pp (1972).
2. Fellner, P., Ehresmann, C., Stiegler, P., and Ebel, J.-P., Nature New Biology, 239, 1 (1972).
3. Adams, J. M., Jeppeson, P. G. N., Sanger, F., Barrell, B. C., Cold Spr. Harb. Symp. Quant. Biol., 34, 611 (1969).
4. Fitch, W. M. and Margoliash, E., Evol. Biol., 4, 67 (1970).
5. White, III, H. B., Laux, B. E., and Dennis, D., Science, 175, 1264 (1972).
6. Mark, A. J. and Petruska, J. A., J. Mol. Biol., 72, 609 (1972).
7. Min Jou, W., Haegeman, G., Ysebaert, M., and Fiers, W., Nature, 237, 82 (1972).
8. Kaempfer, R. and Kaufman, J., Proc. Natl. Acad. Sci. U.S., 70, 1222 (1973).
9. Gianni, A. M., Giglioni, B., Ottolenghi, S., and Comi, P., Nature New Biology, 240, 183 (1972).
10. Clegg, J. B., Weatherall, D. J., and Milnor, P. F., Nature, 234, 337 (1971).
11. Seid-Akhavan, M., Winter, W. P., Abramson, R. K., and Rucknagel, D. L., Blood, 40, 927 (1972).
12. Fitch, W. M., J. Molec. Evolution, 1, 185 (1972).
13. Delisi, C. and Crothers, D. M., Proc. Natl. Acad. Sci. U.S., 68, 2682 (1971).
14. Tinoco Jun, I., Uhlenbeck, O. C., and Levine, M. D., Nature, 230, 362 (1972).
15. Marotta, C. A., Verma, I. M., McCaffery, R. P., and Forget, B. G., Fed. Proc. 32, 455 (1973).
16. Kleinauer, E. F., Reynolds, C. A., Duzy, A. M., Wilson, J. B., Moores, R. R., Berenson, M. P., Wright, C. S., Hnisnan, T. H. J., Biochem. Biophys. Acta, 154, 220 (1968).