DEOXYNUCLEOTIDE SEQUENCE OF AN INSECT cDNA CODES FOR AN UNREPORTED MEMBER OF THE CHIRONOMUS THUMMI GLOBIN FAMILY

Daad A. Saffarini, Patrick M. Trewitt, *Miguel Castro,
Peter J. Wejksnora and Gerald Bergtrom

University of Wisconsin-Milwaukee,
Dept. of Biological Sciences, Milwaukee, WI 53201
*OCS Laboratories, Denton, TX 76202

Received November 1, 1985

Synthetic oligonucleotides served as probes to isolate insect globin clones from a Chironomus thummi cDNA bank. The cDNA insert of one clone (pC-S9) was completely sequenced by the dideoxy termination procedure. Beginning at the 5' end of the coding region, the 584 base pair sequence encodes most of an N-terminal hydrophobic signal sequence and the complete sequence for a mature secreted globin, and contains a polyadenylation recognition site 3' to an appropriate stop codon. The inferred amino acid sequence is that of an unreported variant of hemoglobin VIIB. Based on the number of differences between Hb VIIB chains, the pC-S9 gene has been evolutionarily independent longer than the other (two) members of the globin VIIB subfamily.

Though found among diverse invertebrate phyla, the occurrence of hemoglobins (Hbs) in invertebrates is nonetheless rare (1,2). The hemolymph of the insect Chironomus thummi contains a large number of hemoglobins which presumably confer a survival advantage to the larvae in their anoxic benthic environment. The Hbs are synthesized as preglobins by cells of the fat body (3-5). Recently we showed that immunoreactive Hb precursors are cotranslationally processed in vitro, yielding proteins of the same molecular weight as mature, secreted globins (6). For globins III and IV, the presence of an N-terminal hydrophobic signal sequence has been demonstrated by the sequencing of their genes (7).

Twelve amino acid sequences have been published for C. thummi Hbs (8). This remarkable polymorphism has arisen by periodic duplications of an original globin locus followed by point mutational divergence. Even greater complexity in the C. thummi globin family was revealed by the report of multiple gene copies for globins III and IV (7). We have used synthetic oligonucleotides, corresponding to a highly conserved amino acid sequence, to isolate a number of different globin cDNAs. Sequencing of a nearly full-

0006-291X/85 $1.50

length cDNA reveals the existence of a previously unreported globin. The structure and evolution of this and related globin sequences are discussed.

MATERIALS AND METHODS

C. thummi poly(A)+ RNA was obtained from predominantly 4th instar larvae by homogenization and phenol-chloroform extraction as reported (6), followed by oligo(dT)-cellulose chromatography (9). In some cases, poly(A)+ RNA was further purified by sucrose density gradient centrifugation. The integrity of the mRNA thus isolated was monitored by *in vitro* translation as described (6).

cDNA clone banks were constructed from reverse transcripts (10,11) of total poly(A)+ RNA and 7S-12S poly(A)+ RNA. cDNAs were C-tailed and ligated into pBR322 which had been opened at the PstI site and G-tailed (12,13). Based on their ampicillin sensitivity, about 90% of the colonies obtained in each cDNA bank were recombinant. Approximately 300 colonies from each (amplified) bank were grown on nitrocellulose filters, replica-transferred to a series of additional filters, and lysed *in situ* to bind the DNA (14).

Screening for Globin Clones. Synthetic 17-mer probes were constructed corresponding to a stretch of six highly conserved amino acids (Gln-Phe-Ala-Gly-Lys-Asp) common to all *C. thummi* globins whose amino acid sequences have been reported (8). Four separate 17-mer deoxyribonucleotide mixtures, which included all of the 128 coding options for these amino acids, were prepared by solid phase phosphotriester chemistry (15,16); control pore glass was used as the solid support (17). The syntheses were performed using 33 mg of the glass beads (1 μmole of loaded nucleoside), 14 μmoles of 5'-O-dimethoxytrityl-3'-(2-chlorophenyl) phosphate nucleoside, and 28 μmoles of mesytelene tetrazolide as the coupling activator (18); the reaction volume was 100 μl of anhydrous pyridine. The oligomer mixtures were:

Probe 1: 5'-TC(C,T)TTNCCNGCAAATTG-3'
Probe 2: 5'-TC(C,T)TTNCCNGCGAATTG-3'
Probe 3: 5'-TC(C,T)TTNCCNGCAAACTG-3'
Probe 4: 5'-TC(C,T)TTNCCNGCGAACTG-3'

(N = any one of the 4 deoxynucleotides). The 17-mers were end-labelled by the T4 polynucleotide kinase exchange reaction (19) using γ[32P]dATP. Successful labelling was confirmed by ascending chromatography on DE81 ion-exchange chromatography paper (Whatman) with 0.35M KCl as the solvent.

Each probe was hybridized to one of five replica filters of colonies from each cDNA bank. The fifth filter was hybridized with an equimolar combination of the four probes. The hybridization solution was 6X SSC (0.9M NaCl, 0.09M sodium citrate) containing 0.2% bovine serum albumin, 0.2% Ficoll, 0.2% polyvinylpyrolidone, 20 μg/ml wheat germ tRNA, and 0.05% sodium pyrophosphate. Hybridization was carried out at 37°C overnight in sealed plastic bags. Filters were washed with 6X SSC/0.05% sodium pyrophosphate, first at 37°C for 1 hour and then at 47°C for 10 minutes. After air drying, the filters were autoradiographed at -70°C overnight. Under these conditions, each oligomer probe mixture yielded essentially the same signals on the replicate filters. Colonies that showed strong signals on the initial screening were streaked, and single colony isolates were re-screened with the oligomer probes.

Analysis of Globin cDNAs. Each globin-positive colony was grown in liquid medium, and its plasmid DNA was isolated (20). To analyze for the presence of restriction site(s) within their cDNA insert, plasmid DNAs were digested with PstI and EcoRI (separately). The restriction fragments were separated by agarose gel electrophoresis and sized according to their migration relative to marker DNAs of known length.

Double digestion of DNA from one clone (pC-S9) with PstI and EcoRI generated cDNA fragments compatible with M13 subcloning. These fragments were separated by electrophoresis through low-melting-point agarose. After recovery from the gel and purification on NACS columns (BRL), pC-S9 cDNA fragments were ligated into the polycloning region of M13 strains mp18 and mp19 (21). Sequencing was accomplished by the dideoxy termination method (22) as modified by Messing (21). α[32P]ATP was incorporated in each reaction mix and the samples were double- or triple-loaded (as necessary to obtain complete sequence data) onto 8% polyacrylamide/urea gels. After electrophoresis, gels were exposed directly to AR film (Kodak), or were fixed with 10% methanol/10% glacial acetic acid for 20 minutes and dried before autoradiography.

RESULTS AND DISCUSSION

Screening and Analysis of Globin cDNAs. cDNA clone banks were generated from total poly(A)+ RNA (P bank) and from 7S-12S poly(A)+ RNA (S bank). Initially, 385 and 270 transformant colonies from amplified P and S banks, respectively, were probed with end-labelled globin-specific 17-mer mixtures. Strong hybridization signals were obtained and confirmed by a second round of screening for 16 P colonies and 16 S colonies. The relatively high number of globin-positive colonies is not surprising since 10% of the *in vitro* translation products of total 4th instar poly(A)+ RNA are immunoprecipitable with globin-specific antiserum (6).

Plasmid DNA preparations from each globin-positive colony were cut separately with PstI and EcoRI, and run on calibrated agarose gels to estimate cDNA insert sizes. The length of the inserts ranged up to 600 base pairs (data not shown). Full-length *C. thummi* globin cDNAs were expected to be 530-850 base pairs (bp), based on the sedimentation of preglobin mRNAs through sucrose density gradients (M. Cigan, unpublished data). This prediction had been corroborated by the nucleotide sequence of the genes for *C. thummi* Hbs III and IV (7). Furthermore, the coding regions of the Hb III and Hb IV genes lack EcoRI sites (7). Therefore, plasmid clones likely to contain full-length cDNA inserts (>500 bp), and which contained an internal EcoRI site were candidates for further analysis. Among the clones that met these criteria, pC-S9 (∿590 bp cDNA insert) was chosen for sequencing.

Sequence of pC-S9. Two EcoRI-PstI fragments (∿200 and ∿400 bp) of pC-S9 were subcloned into M13 and sequenced in both directions. The 584 bp sequence of pC-S9 cDNA is presented in Figure 1. Based on its inferred amino acid sequence, pC-S9 is derived from an mRNA encoding a polypeptide similar to globin VIIB (23). Although the cDNA sequence does not extend into the 5'-untranslated region, it does include most of the 5' signal peptide sequence, coding for predominantly hydrophobic or neutral amino acids. The termination codon (AAT) appears at the end of the coding sequence, and the polyadenylation recognition signal (AATAAA) is located 60 bases further downstream. Twelve bases 3' to the AATAAA sequence, there is a tetranucleotide (TTGT) which is found in the 3' untranslated region of some eukaryotic mRNAs including many vertebrate globin genes (24,25), but whose function is not known. There are no premature termination codons in the coding register.

```
                     K   F   F   A   V   L   A   L   C   I     - 6
      5'-poly(G)-TG AAA TTC TTC GCT GTT CTT GCT CTC TGC ATC

 V   G   A   I   A   S   P   L   T   A   D   E   A   S   L     10
GTT GGA GCT ATT GCT*TCC CCA TTG ACT GCT GAC GAA GCT TCA CTC

 V   Q   S   S   W   K   A   V   S   H   N   E   V   D   I     25
GTC CAA TCA TCA TGG AAG GCT GTT AGC CAC AAT GAA GTT GAC ATC

                                            [m]
 L   A   A   V   F   A   A   Y   P   D   I   Q   A   K   F     40
CTC GCT GCT GTT TTT GCT GCT TAC CCA GAC ATC CAG GCT AAG TTC

[s]
 P   Q   F   A   G   K   D   L   A   S   I   K   D   T   G     55
CCA CAA TTC GCC GGA AAG GAC CTC GCT TCA ATC AAG GAT ACT GGT

 A   F   A   T   H   A   T   R   I   V   S   F   L   S   E     70
GCA TTC GCC ACA CAC GCA ACA AGA ATT GTT TCA TTC TTG TCA GAA

                [m]         [a]              a           q
 V   I   A   L   S   G   N   E   S   N   A   S   A   V   N     85
GTC ATC GCT CTT TCA GGA AAC GAA TCA AAC GCT TCT GCT GTC AAC

 g       l   d
 S   L   V   S   K   L   G   D   D   H   K   A   R   G   V    100
TCA CTC GTC TCA AAG CTT GGA GAT GAC CAC AAA GCT CGT GGA GTT

                                    [l]
 S   A   A   Q   F   G   E   F   R   T   A   L   V   A   Y    115
TCA GCT GCT CAA TTT GGA GAA TTC AGA ACC GCT CTC GTT GCT TAC

     q   a                       n
 L   S   N   H   V   S   W   G   D   N   V   A   A   A   W    130
CTC TCA AAC CAT GTC TCA TGG GGT GAC AAT GTT GCT GCT GCC TGG

 s                          [f]
 N   K   A   L   D   N   T   Y   A   I   V   V   P   R   L    145
AAC AAA GCC CTC GAT AAC ACC TAT GCC ATC GTT GTC CCA CGT CTT

TAA*ATTATTTAGATAAACATCTAGCTGAATAATTTATCATTATAAAATGAAATAGAAG

AATTAATAAAGATCAATAATTCTTTGTGATTTTTAACATCTACCG-poly(C)-3'
```

Figure 1: Nucleotide and inferred amino acid sequences for pC-S9 cDNA. An asterisk denotes the beginning and end of the sequence encoding the mature globin polypeptide. The termination codon is underlined, and the polyadenylation recognition site is double underlined. Centered over triplet codons, upper case letters represent the inferred amino acid sequence, including a 16 amino acid signal peptide (positions -16 to -1). Lower case letters in brackets denote polymorphic positions, based on the published sequence of Hb VIIB variants (21). Unbracketed lower case letters are amino acids in the published sequences which differ from those inferred from the pC-S9 nucleotide sequence.

The single letter code for amino acids is as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; and W, Trp.

Globin Signal Peptides. Aside from the data presented in this report, the only documented amino acid sequence for the signal peptide of a secreted globin is that inferred for *C. thummi* globin IV (7). The most parsimonious alignment of the published amino acid sequences of Hb VIIB and Hb IV (8) yields a 43.5% homology between these two mature (i.e. processed and secreted) proteins. Their signal peptides exhibit a similar level of homology:

```
              -15                 -10                 -5
VIIB:        -Lys-Phe-Phe-Ala-Val-Leu-Ala-Leu-Cys-Ile-Val-Gly-Ala-Ile-Ala-
  IV:   (Met)-Lys-Leu-Leu-   -Ile-Leu-Ala-Leu-Cys-Phe-Ala-Ala-Ala-Ser-Ala-
```

Each of the signal peptides has a charged N-terminus followed by an extended "core" region of non-polar, largely hydrophobic residues, which are characteristic features of prokaryotic and eukaryotic signal peptides (26). The total hydrophobic free energy of transition from an aqueous to a non-polar phase for the the core region of globin VIIB and globin IV is -109.5 and -86.1 kJ/mol, respectively. These energetically favorable values are in the range calculated for a diverse sample of other signal peptides (27).

The signal peptide sequence of Hb VIIB is at least one amino acid longer than that of Hb IV. Alignment for maximum homology results in either a deletion of one residue in the signal sequence of Hb IV, or an addition of one amino acid to the sequence of Hb VIIB. Based on the length of the globin IV signal sequence and our sequence data, we suspect that pC-S9 includes the last two bases of the methionine initiation codon.

The Globin VIIB Subfamily. The amino acid sequence inferred from pC-S9 differs from the published Hb VIIB sequence (23) in 9 positions. Based on X-ray crystallographic data for *C. thummi* Hb III (28), six of these nine positions (82, 86, 117, 118, 124 and 131) have no assignable function with respect to intramolecular interactions, while three of the substituted positions (85, 88 and 89) do participate in interhelical contact at the surface of the molecule (29). Nevertheless, in *C. thummi* Hbs as well as the α-like and β-like globin sequences of vertebrates, both classes of positions show a relatively high degree of amino acid variability (30). Therefore, the amino acid substitutions specified by pC-S9 would be expected to have minimal, if any effect on the essential function (i.e. reversible oxygen binding) of Hb VIIB. Moreover, the 9 residues in question are clustered in two regions of the globin molecule. Their non-random distribution and their confinement to positions of high variability strongly suggest that pC-S9 differences did not arise by stochastic processes, such as unfaithful transcription *in vivo* or cloning artifacts *in vitro*. pC-S9 differs from other sequences encoding Hb VIIB variants at a minimum of 15 nucleotides. Since it is difficult to imagine how so many differences might be maintained at the same locus, our data implies the existence of another gene copy in the globin VIIB subfamily.

As indicated in Figure 1, the amino acid sequence published for Hb VIIB contains paired alternative residues at each of six positions (23). Although the genetic basis for this polymorphism remains uncertain, the most straightforward explanation of the amino acid data is that two related polypeptide chains were co-purified and sequenced. Based on our sequence data, we suggest that one of these Hb VIIB polymorphs contains Gln, Pro, Ser, Glu, Thr and Tyr at positions 37, 41, 75, 78, 110 and 138, respectively (which are the residues inferred for these positions from the sequence of pC-S9), while the second Hb VIIB variant contains the reported alternatives: Met, Ser, Met, Ala, Leu and Phe.

The globin VIIB subfamily arose by a duplication in the globin VI locus (8). The evolutionary relationship between the amino acid sequence inferred from pC-S9 and the Hb VIIB variants reported by Sladic-Simic *et al.* (23) is more distant than that between the Hb VIIB variants themselves. Therefore, we propose that in a subsequent event the globin VIIB locus was itself duplicated, one branch leading to pC-S9 and the second branch to other members of the subfamily. More recently, the latter branch underwent a second round of duplication, creating the Hb VIIB variants which were identified at the amino acid level (23).

ACKNOWLEDGMENTS: This work was funded in part by the National Science Foundation (DCB 8410667), by the Graduate School of the University of Wisconsin-Milwaukee, by Shaw Discretionary Funds and the Shaw Scholars Program of the James D. Shaw and Dorothy Shaw Fund. Thanks go to Lynn DeOgny for her expertise in the preparation of oligonucleotides.

REFERENCES

1. Laufer, H., Bergtrom, G. and Rogers, R. (1976) In *Invertebrate Tissue Culture: Applications in Medicine, Biology and Agriculture*, Ch. 19. (Kurstak, E. and Maramorosch, K., eds.) Academic Press, New York.
2. Vinogradov, S.N. (1985) Comp. Biochem. Physiol. 82B: 1-15.
3. Bergtrom, G., Laufer, H. and Rogers, R. (1976). J. Cell Biol. 69: 264-274.
4. Bergtrom, G. and Robinson, J.M. (1977). J. Ultrastructural Res. 60: 395-405.
5. Baldwin, S., Haralson, M.A. and Bergtrom, G. (1981). Arch. Biochem. Biophys. 211: 390-395.
6. Cigan, A.M., Trewitt, P.M. and Bergtrom, G. (1985). Insect Biochem. (in press).
7. Antoine, M. and Niessing, J. (1984). Nature, London 310: 795-798.
8. Goodman, M., Braunitzer, G., Kleinschmidt, T. and Aschauer, H. (1983). Hoppe-Seyler's Z. Physiol. Chem. 354: 205-217.
9. Aviv, G. and Leder, P. (1972). Proc. Nat. Acad. Sci. U.S.A. 69: 1408-1412.
10. Efstratiadis, A., Kafatos, F.C., Maxam, A.M. and Maniatis, T. (1976) Cell 7: 279-288.
11. Retzel, E.F., Collet, M.S. and Faras, A.J. (1980) Biochemistry 19: 513-518.
12. Villa-Komaroff, L., Efstratiadis, A., Broome, S.,Lomedico, P., Tizard, R., Naker, S.P., Chick, W.L. and Gilbert, W. (1978) Proc. Nat. Acad. Sci. U.S.A. 75: 3727-3731.
13. Rowekamp, W. and Firtel, R.A. (1980) Dev. Biol. 79: 409-418.
14. Grunstein, M. and Hogness, D. (1975). Proc. Natl. Acad. Sci. U.S.A. 72: 3961-3965.
15. Kohli, V., Balland, A., Wintzerith, M., Saueruald, R., Staub, A. and Lecocq, J.P. (1982). Nucl. Acid Res. 10: 7439-7448.

16. Narang, S.A., Brousseau, R., Hsiung, H. M. and Michniewicz, J.J. (1980). In Methods in Enzymol., Vol. 65, pp 610-615. (Wu, R., Grossman, L. and Moldave, K., eds.) Academic Press, N.Y.
17. Gait, M.J. (1984). In Oligonucleicotide Synthesis: A Practical Approach, Ch. 4. (Gait, M.J., ed.) IRL Press, Oxford, U.K.
18. Seth, A.K. and Jay, E. (1980). Nucl. Acid Res. 8: 5445-5459.
19. Berkner, K.L. and Folk, W.R. (1977) J. Biol. Chem. 252: 3176-3184.
20. Holmes, D.S. and Quigley, M. (1981) Anal. Biochem. 114: 193-197.
21. Sanger, F., Nicklen, S. and Coulson, A.R. (1977). Proc. Natl. Acad. Sci. U.S.A. 74: 5463-5467.
22. Messing, J. (1983). In Methods in Enzymol., Vol. 101, pp. 20-89. (Wu, R., Grossman, L. and Moldave, K., eds.) Academic Press, N.Y.
23. Sladic-Simic, D., Kleinschmidt, T. and Braunitzer, G. (1977). Hoppe-Seyler's Z. Physiol. Chem. 385: 591-594.
24. Efstratiadis, A., Posakony, J., Maniatis, T., Lawn, R., O'Connell, C., Spritz,R., DeRiel, J.K., Forget, B., Weissman, S.M., Slightom, J.L., Blechl,A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) Cell 21: 653-668.
25. Knochel, W., Meyerhof, W., Hummel, S. and Grundmann, U. (1983). Nucl. Acid. Res. 11: 1543-1553.
26. von Heijne, G. (1982) J. Mol. Biol. 159: 537-541.
27. von Heijne, G. (1981) Eur. J. Biochem. 116: 419-422.
28. Huber, R., Epp, O., Steigemann, W. and Formanek, H. (1971). Eur. J. Biocem. 19: 42-50.
29. Lesk, A. and Chothia, C. (1980). J. Mol. Biol. 136: 225-270.
30. Trewitt, P.M. and Bergtrom. G. (1984) Genetics 107: s107 (Abstract).