Proc. Natl. Acad. Sci. USA
Vol. 78, No. 1, pp. 243–246, January 1981
Biochemistry

Nucleotide sequence of cloned rat serum albumin messenger RNA

(molecular evolution/amino acid sequence/intragenic duplication)

THOMAS D. SARGENT, MARIA YANG, AND JAMES BONNER

Division of Biology, California Institute of Technology, Pasadena, California 91125

Contributed by James F. Bonner, October 15, 1980

ABSTRACT **The nucleotide sequences of the recombinant DNA inserts of three bacterial plasmid clones containing nearly all of the rat serum albumin mRNA have been determined. A statistical analysis of the nucleotide sequence reveals a pattern of repeated internal homology that confirms the "intragenic triplication" model of albumin evolution.**

The protein serum albumin has several attributes that make it an attractive subject for experimental investigation. It is the predominant and characteristic synthetic product of adult vertebrate liver and is therefore a convenient example of controlled gene expression in terminally differentiated cells. In mammalian embryos, there is a reciprocal relationship between the expression of albumin and its fetal counterpart, α-fetoprotein, which is an interesting problem of developmental biology (1). Perhaps the most striking property of serum albumin is the remnants in its amino acid sequence of the evolutionary history of this protein. Disulfide crosslinks generate a pattern of loops that is repeated threefold, defining the three structural domains of serum albumin. These domains exhibit significant amino acid homology in addition to the cysteine residues, and it has been suggested by Brown (2) that albumin evolved by intragenic triplication of a smaller protein corresponding to one domain, which may have in turn evolved from a much smaller sequence by an earlier series of duplications and partial deletions. However, this evolutionary hypothesis is based upon amino acid sequence homology between domains that is not overwhelming and that could conceivably be due to convergent evolution of originally nonhomologous sequences.

To address this question and as a basis for further research, we have cloned the rat serum albumin messenger RNA as a series of recombinant DNA plasmids and have determined the nucleotide sequences of these clones, which include all of the albumin mRNA from the amino-terminal codon to within approximately 30 nucleotides of the site at which poly(A) is attached. A statistical analysis of these data reveals extensive internal homology in the albumin mRNA that verifies the intragenic triplication hypothesis of albumin evolution.

METHODS

Cloning Procedures. The production of two of the plasmid clones used in the present experiments (pRSA57 and pRSA13) has been described (3). The plasmid clone pRSA510 was produced by "extending" a primer fragment of pRSA57 toward the 5′ end of the albumin mRNA (Fig. 1). Briefly, the *Hin*dIII fragment nearest the 5′ end of pRSA57 was isolated and hybridized to rat liver mRNA, which is approximately 7% albumin-encoding sequences by mass. The specific heteroduplex thus formed was treated with reverse transcriptase from avian myeloblastosis virus and then with sodium hydroxide to generate a cDNA that extended from the second *Hin*dIII site to within a few nucleotides of the cap of the albumin mRNA. This material was converted to double-stranded cDNA and "tailed" with oligo(dG) according to published procedures (4, 5). The tailed albumin DNA was mixed with plasmid pBR322 DNA that had been cleaved with the restriction endonuclease *Cla* I and tailed with oligo(dC). This recombinant DNA was used to transform the *Escherichia coli* strain MC1061 (6) according to the method of Kushner (7). Clones that were sensitive to tetracycline and resistant to ampicillin were further screened by hybridization to a restriction endonuclease fragment of the rat serum albumin gene containing the "leader" exon.

Sequencing. Determination of DNA sequence was done according to the procedures of Maxam and Gilbert, with minor modifications (8).

Statistical Analysis. The validity of a given homology between two sequences was evaluated by calculation of an "accident probability," *Pa*, which is the probability that a homology equal to or greater than that being considered might arise accidentally. The equation is a summation of the Poisson distribution,

$$Pa = \sum_{i=n}^{N} \frac{e^{-Np}(Np)^i}{i!},$$

in which N is the length in nucleotides over which the homology is measured, n is the number of matches in this interval, and p is the probability that any given position will be a match, which will be equal to 0.25 if there is no preference for any of the four nucleotides at any given position. In fact, there is some deviation from ideal randomness. We have empirically determined various values of p and find that they fall between 0.25 and 0.28 for rat serum albumin mRNA. Unless specified otherwise (i.e., Table 1), *Pa* values were calculated by assuming $p = 0.25$. The mRNA sequence was divided into several segments, each of which was compared with all the others. A computer program, written by R. F. Murphy and J. W. Posakony, was used to search for stretches of sequence that met or exceeded arbitrary criteria of homology. No allowance was made for gaps, and thus the "homology blocks" tend to have rather sharp boundaries near sites of deletions. Homology extending over a minimum of 100 nucleotides with a maximum *Pa* of 7×10^{-4} was considered "legitimate" and used to establish the internal alignments of the albumin mRNA sequence.

RESULTS

The strategy used in determining the nucleotide sequence of rat serum albumin mRNA is shown in Fig. 1. Most of the sequencing operations were repeated at least twice, and the resulting data are probably free of errors, although mutations associated with the cloning procedure cannot be ruled out. Approximately 35 nucleotides at the extremities of the albumin mRNA failed to

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.

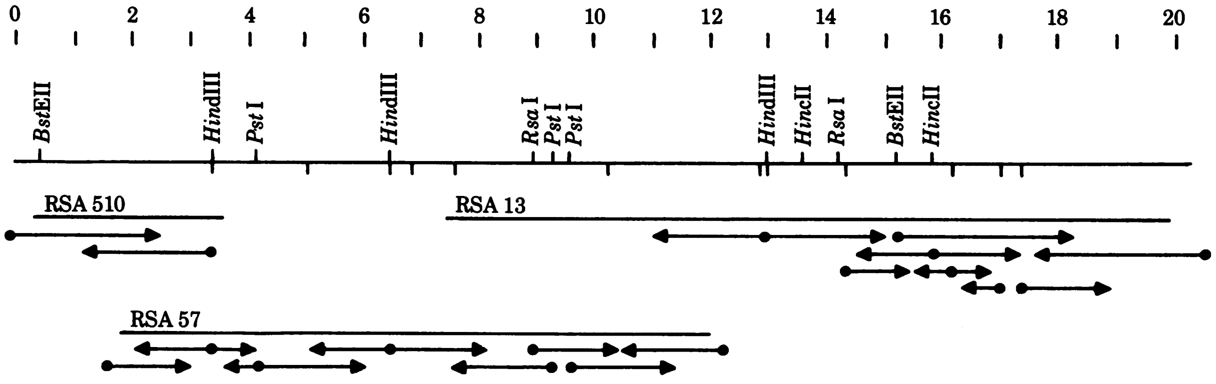

FIG. 1. Sequencing strategy and restriction endonuclease map of albumin cDNA clones. Downward ticks denote *Alu* I sites. The rat serum albumin (RSA) mRNA inserts of the recombinant DNA plasmids pRSA510, pRSA57, and pRSA13 are shown. The nucleotide sequence was determined by labeling restriction endonuclease-digested DNA at the sites indicated by ●s and subjecting this "end-labeled" DNA to the chemical sequencing procedure of Maxam and Gilbert (8). The direction and extent of the sequencing determinations are indicated by the arrows. Most determinations were performed at least twice. The scale is in hundreds of nucleotides.

appear in any of the cDNA clones. The balance extends 1956 nucleotides from the middle of ATG/Met, corresponding to the amino terminus of pre-pro-albumin, to 130 nucleotides into the 3′ untranslated region. The sequence data are shown in Fig. 2. By assuming code universality, the amino acid sequence of rat pre-pro-albumin is readily inferred, and it is listed above the nucleotide sequence. This amino acid sequence concurs with published sequences that have been determined by conventional methods (refs. 9 and 10; T. Ikenaka, personal communication). There are a few discrepancies, however. Amino acid positions 353 (Thr), 357 (Glu), 402 (Gln), 453 (Asn), 454 (Leu), and 456 (Arg) are specified in the literature as Lys, Asp, Ala, Leu, Gly, and Glx, respectively. There are several possible explanations for these differences, but they are not due to erroneous DNA sequence determinations. When the amino acid sequences of rat, human, and bovine serum albumins are compared (11), 61% of the positions are found to be identical in all three proteins. This homology is rather evenly distributed over the length of the protein. However, it is interesting to note that of the 35 cysteine residues, 34 are exactly conserved in all three albumins. The one difference is due to absence in rat albumin of the residues between the 17th and 18th cysteines. This near-perfect conservation implies that changes in the size of the loops generated by cystine crosslinks are highly deleterious, whereas the primary sequence of the protein can diverge rather freely, at a rate typical of eukaryotic proteins (12). There is nothing in the 3′ untranslated portion similar to the A-A-T-A-A-A sequence found near the poly(A) addition sites of most mRNAs (13). This is due to a cloning failure rather than a unique feature of albumin mRNA. The DNA sequence of the 3′ end of the rat serum albumin gene has been determined (unpublished data) and the sequence G-C-A-A-T-T-A-A-T-A-A-A-A-A-A-T-G-G is found 134 nucleotides downstream from the termination codon, TAA.

The DNA sequence has been subjected to a statistical analysis designed to identify regions of extensive internal homology. The results of this analysis are summarized in Table 1. Depending upon how one defines "extensive homology," a large variety of alignments can be found. When the specification is made that the homologous regions be at least 100 base pairs long and have a *Pa* of 7 × 10^{-4} or less, only five regions of homology qualify. Four of these, I–IV are in approximate phase with one another. The fifth homology block is distinct from the others, being offset 212 nucleotides. It is also not aligned in phase with the codon reading frame, as are blocks I–IV, and thus seems unlikely to represent "legitimate" homology, although it is certainly statistically significant ($Pa = 6.4 \times 10^{-4}$). Block V has been disregarded in establishing the domain boundaries. Its relationship, if any, to albumin evolution remains to be elucidated. There are no other homologous alignments with *Pa* values within a factor of 10 of those listed in Table 1. The four pairs of homologous sequence unambiguously define an internal alignment, shown in Fig. 3. The protein is divided into three blocks that correspond almost exactly to the three "domains" described in bovine and human albumins by Brown (2). This alignment places most of the cysteine residues in phase. Inspection of the positions of these homology blocks and cysteine residues leads to the inference that there have been several small deletions and insertions of 3, 6, 9, or 12 nucleotides at various positions in the albumin mRNA.

When the amino acid sequences within each homology block are compared, the results are in accord with the DNA homology, except that the percentage of matches is significantly lower for amino acids than for nucleotides, especially for blocks I, III, and IV (Table 1).

Table 1. Homology summary

Homology block	Locations	DNA homology (%)	*Pa*(0.25)	*Pa*(0.28)	Amino acid homology (%)
I	80– 209 vs. 656– 785	57/130 (44)	6.3 × 10^{-5}	9.5 × 10^{-4}	10/43 (23)
II	427– 603 vs. 1003–1179	97/177 (55)	2.0 × 10^{-10}	2.4 × 10^{-8}	24/59 (41)
III	616– 835 vs. 1192–1411	84/220 (38)	1.7 × 10^{-4}	3.8 × 10^{-3}	11/73 (15)
IV	1707–1815 vs. 1113–1221	49/111 (44)	1.7 × 10^{-4}	1.8 × 10^{-3}	7/36 (19)
V	1038–1180 vs. 1402–1544	57/143 (40)	6.4 × 10^{-4}	6.7 × 10^{-3}	—

Pa(0.25) and *Pa*(0.28) are accident probabilities calculated for random match probabilities, *p*, of 0.25 and 0.28, respectively. Block V is not aligned in phase with the codon reading frame, so an amino acid comparison is meaningless.

```
met lys trp val thr phe leu leu leu leu phe ilu ser gly ser ala phe ser arg gly val phe arg arg glu ala his lys ser glu ilu ala his
ATG AAG TGG GTA ACC TTT CTC CTC CTC CTC TTC ATC TCC GGT TCT GCC TTT TCT AGG GGT GTG TTT CGC CGA GAA GCA CAC AAG AGT GAG ATC GCC CAT     99

arg phe lys asp leu gly glu gln his phe lys gly leu val leu ilu ala phe ser gln tyr leu gln lys cys pro tyr glu glu his ilu lys leu
CGG TTT AAG GAC TTA GGA GAA CAG CAT TTC AAA GGC CTA GTC CTG ATT GCC TTT TCC CAG TAT CTC CAG AAA TGC CCA TAT GAA GAG CAT ATC AAA TTG    198

val gln glu val thr asp phe ala lys thr cys val ala asp glu asn ala glu asn cys asp lys ser ilu his thr leu phe gly asp lys leu cys
GTG CAG GAA GTA ACA GAC TTT GCA AAA ACA TGT GTC GCT GAT GAG AAT GCC GAA AAC TGT GAC AAG TCC ATT CAC ACT CTC TTC GGA GAC AAG TTA TGC    297

ala ilu pro lys leu arg asp asn tyr gly glu leu ala asp cys cys ala lys gln glu pro glu arg asn glu cys phe leu gln his lys asp asp
GCC ATT CCA AAG CTT CGT GAC AAC TAC GGT GAA CTG GCT GAC TGC TGT GCA AAA CAA GAG CCC GAA AGA AAC GAG TGT TTC CTG CAG CAC AAG GAT GAC    396

asn pro asn leu pro pro phe gln arg pro glu ala glu ala met cys thr ser phe gln glu asn pro thr ser phe leu gly his tyr leu his glu
AAC CCC AAC CTG CCA CCC TTC CAG AGG CCG GAG GCT GAG GCC ATG TGC ACC TCC TTC CAG GAG AAC CCT ACC AGC TTT CTG GGA CAC TAT TTG CAT GAA    495

val ala arg arg his pro tyr phe tyr ala pro glu leu leu tyr tyr ala glu lys tyr asn glu val leu thr gln cys cys thr glu ser asp lys
GTT GCC AGG AGA CAT CCT TAT TTC TAT GCC CCA GAA CTC CTT TAC TAT GCT GAG AAA TAC AAT GAG GTT CTG ACC CAG TGC TGC ACA GAG TCT GAC AAA    594

ala ala cys leu thr pro lys leu asp ala val lys glu lys ala leu val ala ala val arg gln arg met lys cys ser ser met gln arg phe gly
GCA GCC TGC CTG ACA CCG AAG CTT GAT GCC GTG AAA GAG AAA GCA CTG GTC GCA GCT GTC CGT CAG AGG ATG AAG TGC TCC AGT ATG CAG AGA TTT GGA    693

glu arg ala phe lys ala trp ala val ala arg met ser gln arg phe pro asn ala glu phe ala glu ilu thr lys leu ala thr asp val thr lys
GAG AGA GCC TTC AAA GCC TGG GCA GTA GCT CGT ATG AGC CAG AGA TTC CCC AAT GCT GAG TTC GCA GAA ATC ACC AAA TTG GCA ACA GAC GTT ACC AAA    792

ilu asn lys glu cys cys his gly asp leu leu glu cys ala asp asp arg ala glu leu ala lys tyr met cys glu asn gln ala thr ilu ser ser
ATC AAC AAG GAG TGC TGT CAC GGC GAC CTG TTG GAA TGC GCG GAT GAC AGG GCA GAA CTT GCC AAG TAC ATG TGT GAG AAC CAG GCC ACT ATC TCC AGC    891

lys leu gln ala cys cys asp lys pro val leu gln lys ser gln cys leu ala glu thr glu his asp asn ilu pro ala asp leu pro ser ilu ala
AAA CTG CAG GCT TGC TGT GAT AAG CCA GTG CTG CAG AAA TCC CAG TGT CTC GCT GAG ACA GAA CAT GAC AAC ATT CCT GCC GAT CTG CCC TCA ATA GCT    990

ala asp phe val glu asp lys glu val cys lys asn tyr ala glu ala lys asp val phe leu gly thr phe leu tyr glu tyr ser arg arg his pro
GCT GAC TTT GTT GAG GAT AAG GAA GTG TGT AAG AAC TAT GCT GAG GCC AAG GAT GTC TTC CTG GGC ACG TTT TTG TAT GAA TAT TCA AGA AGG CAC CCC   1089

asp tyr ser val ser leu leu leu arg leu ala lys lys tyr glu ala thr leu glu lys cys cys ala glu gly asp pro pro ala cys tyr gly thr
GAT TAC TCC GTG TCC CTG CTG CTG AGA CTT GCT AAG AAA TAT GAA GCC ACA CTG GAG AAG TGC TGT GCT GAA GGC GAT CCT CCT GCC TGC TAC GGC ACA   1188

val leu ala glu phe gln pro leu val glu glu pro lys asn leu val lys thr asn cys glu leu tyr glu lys leu gly glu tyr gly phe gln asn
GTG CTT GCA GAA TTT CAG CCT CTT GTA GAA GAA CCT AAG AAC TTG GTC AAA ACT AAC TGT GAG CTT TAC GAG AAG CTT GGA GAG TAT GGA TTC CAA AAC   1287

ala val leu val arg tyr thr gln lys ala pro gln val ser thr pro thr leu val glu ala ala arg asn leu gly arg val gly thr lys cys cys
GCC GTT CTG GTT CGA TAC ACC CAG AAA GCA CCT CAG GTG TCG ACC CCA ACT CTC GTG GAG GCA GCA AGA AAC CTG GGA AGA GTG GGC ACC AAG TGT TGT   1386

thr leu pro glu ala gln arg leu pro cys val glu asp tyr leu ser ala ilu leu asn arg leu cys val leu his glu lys thr pro val ser glu
ACC CTT CCT GAA GCT CAG AGA CTG CCC TGT GTG GAA GAC TAT CTG TCT GCC ATC CTG AAC CGT CTG TGT GTG CTG CAT GAG AAG ACC CCA GTG AGC GAG   1485

lys val thr lys cys cys ser gly ser leu val glu arg arg pro cys phe ser ala leu thr val asp glu thr tyr val pro lys glu phe lys ala
AAG GTC ACC AAG TGC TGT AGT GGG TCC TTG GTG GAA AGA CGG CCA TGT TTC TCT GCT CTG ACA GTT GAC GAG ACA TAT GTC CCC AAA GAG TTT AAA GCT   1584

glu thr phe thr phe his ser asp ilu cys thr leu pro asp lys glu lys gln ilu lys lys gln thr ala leu ala glu leu val lys his lys pro
GAG ACC TTC ACC TTC CAC TCT GAT ATC TGC ACA CTC CCA GAC AAG GAG AAG CAG ATA AAG AAG CAA ACG GCT CTC GCT GAG CTG GTG AAA CAC AAG CCC   1683

lys ala thr glu asp gln leu lys thr val met gly asp phe ala gln phe val asp lys cys cys lys ala ala asp lys asp asn cys phe ala thr
AAG GCC ACA GAA GAT CAG CTG AAG ACG GTG ATG GGT GAC TTC GCA CAA TTC GTG GAC AAG TGT TGC AAG GCT GCC GAC AAG GAT AAC TGC TTC GCC ACT   1782

glu gly pro asn leu val ala arg ser lys glu ala leu ala ter
GAG GGG CCA AAC CTT GTT GCT AGA AGC AAA GAA GCC TTA GCC TAA ACACATCACAACCATCTCAGGCTACCCTGAGAAAAAAAGACATGAAGACTCAGGACTCATCTCTTCTGTTG   1881

GTGTAAAACCAACACCCTAAGGAACACAAATTTCTTTGAACATTTGACTT CTTTTCTC
```

FIG. 2. Nucleotide sequence. Except for approximately 35 nucleotides from either end of the mRNA that were not cloned, this is the complete sequence of the albumin mRNA. The inferred amino acid sequence of rat pre-pro-albumin is also indicated. The "pre" piece is amino acid residues 1–18 and the "pro" piece is residues 19–24.

DISCUSSION

Did albumin evolve by intragenic triplication? There is no doubt that the internal homology we have found in albumin mRNA is statistically significant. The protein is in fact composed of three homologous "domains." There are only two ways to explain partial sequence homology; (*i*) initial identity followed by mutational divergence and (*ii*) convergent evolution of two initially distinct sequences. There are several arguments against the latter alternative, the most convincing of which is based on the fact that the internal homology of albumin is much greater at the level of the DNA than at the level of the amino acid sequence (Table 1). This cannot be explained by convergent evolution driven by selective pressure on albumin protein structures. Intragenic triplication followed by partial divergence is the only reasonable explanation for the observed structure of rat serum albumin. It is possible that other duplication events preceded and followed this triplication. We have not been able to conclusively identify vestiges of intradomain homology that would indicate an earlier series of intragenic duplications, but there is a high background of relatively weak internal homology that is not in phase with the four main blocks that define the domains. A more sophisticated analysis of the rat serum albumin mRNA sequence might reveal periodicity in this background and thereby identify the "proto-albumin" sequence, if it exists. This question is more effectively addressed by analysis of the albumin gene rather than its mRNA. Previous measurements of exon boundaries in this gene, recently augmented by DNA sequence analysis (ref. 3; unpublished data), suggest that duplication

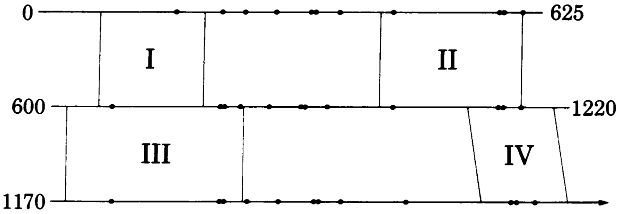

FIG. 3. Homology diagram. The horizontal lines symbolize the mRNA sequence, divided into three overlapping segments, the boundaries of which are indicated at the ends of each line. The vertical lines denote the four homology blocks that define the three domains of albumin. ●, Cysteine residues.

events may have preceded the triplication of domains. As for subsequent duplications, there is evidence that albumin and α-fetoprotein are related sequences (14), which suggests that an intergenic duplication may have taken place. This matter should be resolved when more sequence data become available.

A fundamental problem of biology is to explain the complexity of the eukaryotic genome. Duplication and divergence of genomic DNA may account for much of this diversity. Surveys of the primary sequences of many different proteins reveal a number of clear examples of internal homology (15), most of which are probably due to intragenic duplications. Furthermore, many genes, conceivably most, are members of families that arose by intergenic duplications (16, 17). Coupled with relatively unrestrained accumulation of mutations in redundant sequences, this evolutionary mechanism could convert a simple protein with only one function into a family of complex proteins with many different functions.

We thank R. Bruce Wallace for his helpful suggestions and T. Ikenaka for providing us with unpublished amino acid sequence data. This work was supported in part by Grants 5 T32 GM 07616 and GM 13762 awarded by the National Institute of General Medical Sciences.

1. Sala-Trepat, J. M., Dever, J., Sargent, T. D., Thomas, K., Sell, S. & Bonner, J. (1979) *Biochemistry* **18**, 2167–2178.
2. Brown, J. R. (1976) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **35**, 2141–2144.
3. Sargent, T. D., Wu, J. R., Sala-Trepat, J. M., Wallace, R. B., Reyes, A. A. & Bonner, J. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3256–3260.
4. Higuchi, R., Paddock, G. V., Wall, R. & Salser, W. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 3146–3150.
5. Roychoudhury, R., Jay, E. & Wu, R. (1976) *Nucleic Acids Res.* **3**, 101–116.
6. Casadaban, M. J. & Cohen, S. N. (1980) *J. Mol. Biol.* **138**, 179–207.
7. Kushner, S. R. (1978) in *Proceedings of the International Symposium on Genetic Engineering*, eds. Boyer, H. W. & Nicosia, S. (Elsevier, New York), pp. 17–23.
8. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.*, **65**, 499–560.
9. Strauss, A. W., Bennet, C. D., Donohue, A. M., Rodkey, J. A. & Alberts, A. W. (1977) *J. Biol. Chem.* **252**, 6846–6855.
10. Isemura, S. & Ikenaka, T. (1978) *J. Biochem.* **83**, 35–48.
11. Dayhoff, M. O. (1976) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (National Biomedical Research Foundation, Washington, DC), Vol. 5, Suppl. 2, pp. 266–277.
12. Wilson, A. C., Carlson, S. S. & White, T. J. (1977) *Annu. Rev. Biochem.* **46**, 573–639.
13. Benoist, C., O'Hare, K., Breathnach, R. & Chambon, P. (1980) *Nucleic Acids Res.* **8**, 127–142.
14. Liao, W. & Taylor, J. M. (1980) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **39**, 2016 (abstr.).
15. Barker, W. C., Ketcham, L. K. & Dayhoff, M. O. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (National Biomedical Research Foundation, Washington, DC), Vol. 5, Suppl. 3, pp. 359–362.
16. Hood, L., Campbell, J. H. & Elgin, S. C. R. (1975) *Annu. Rev. Genet.* **9**, 305–353.
17. Long, E. O. & Dawid, I. B. (1980) *Annu. Rev. Biochem.* **49**, 727–764.