ACKNOWLEDGMENTS

We thank R. Everett for amino acid composition analysis, E. Toth for in vitro bioassays, Dr. N. Stebbing for critically reading the manuscript, and J. Bennett for typing.

REFERENCES

Aggarwal, B. B., Kohr, W. J., Hass, P. E., Moffat, B., Spencer, S. A., Henzel, W. J., Bringman, T. S., Nedwin, G. E., Goeddel, D. V., & Harkins, R. N. (1985) *J. Biol. Chem. 260*, 2345–2354.

Carswell, E. A., Old, L. J., Kassel, R. L., Green, S., Fiore, N., & Williamson, B. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 3666–3670.

Chang, C. T., Wu, C.-S. C., & Yang, J. T. (1978) *Anal. Biochem. 91*, 13–31.

Chou, P. Y., & Fasman, G. D. (1978) *Annu. Rev. Biochem. 47*, 251–276.

Freedman, R. B., & Millson, D. A. (1980) in *Enzymology of Post-Translational Modification of Proteins* (Freedman, R. B., & Hawkins, H. C., Eds.) Vol. 1, pp 157–212, Academic, London.

Green, S., Dobrjansky, A., Carswell, E. A., Kassel, R. L., Old, L. J., Fiore, N., & Schwartz, M. K. (1976) *Proc. Natl. Acad. Sci. U.S.A. 73*, 381–385.

Greenfield, N., & Fasman, G. (1969) *Biochemistry 8*, 4108–4116.

Habeeb, A. F. S. A. (1972) *Methods Enzymol. 25*, 457–465.

Hammerstrom, J. (1982) *Scand. J. Immunol. 15*, 311–318.

Helson, L., Green, S., Carswell, E., & Old, L. J. (1975) *Nature (London) 258*, 731–732.

Illangasekare, M. P., & Woody, R. W. (1986) *Biophys. J. 49*, 296a.

Jirgensons, B. (1976) *Biochim. Biophys. Acta 434*, 58–68.

Kuntz, I. D. (1972) *J. Am. Chem. Soc. 94*, 4009–4012.

Laemmli, U. K. (1970) *Nature (London) 227*, 680–685.

Marmenout, A., Fransen, L., Tavernier, J., Van Der Heyden, J., Tizard, R., Kawashima, E., Shaw, A., Johnson, M.-J., Semon, D., Muller, R., Ruysshaert, M.-R., Van Vliet, A., & Fiers, W. (1985) *Eur. J. Biochem. 152*, 515–522.

Matthews, N. (1981) *Immunology 44*, 135–142.

Matthews, N. (1982) *Br. J. Cancer 45*, 615–617.

Palladino, M. A., Kohr, W. J., Aggarwal, B. B., & Goeddel, D. V. (1984) *Nature (London) 312*, 724–729.

Pennica, D., Nedwin, G. E., Hayflick, J. S., Seeburg, P. H., Derynck, R., Palladino, M. A., Kohr, W. J., Aggarwal, B. B., & Goeddel, D. V. (1985) *Nature (London) 312*, 724–729.

Reeke, G. N., Becker, J. W., & Edelman, G. M. (1975) *J. Biol. Chem. 250*, 1525–1547.

Ruff, M. R., & Gifford, G. E. (1981) *Infect. Immun. 31*, 380–385.

Schoemaker, J. M., Brasnett, A. H., & Marston, F. A. O. (1985) *EMBO J. 4*, 755–780.

Shirai, T., Yamaguchu, H., Ito, H., Todd, C. W., & Wallace, R. B. (1985) *Nature (London) 313*, 803–806.

Spofford, B., Dayness, R. A., & Granger, G. A. (1974) *J. Immunol. 112*, 2111–2115.

Strickland, E. H. (1974) *CRC Crit. Rev. Biochem. 2*, 113–175.

Sugarman, B. J., Aggarwal, B. B., Hass, P. E., Figari, I. S., Palladino, M. A., Jr., & Shepard, H. M. (1985) *Science (Washington, D.C.) 230*, 943–945.

Wang, A. M., Creasey, A. A., Ladner, M. B., Lin, L. S., Strickler, J., Van Arsdell, J. N., Yamamoto, R., & Mark, D. F. (1985) *Science (Washington, D.C.) 228*, 149–154.

Sequence Comparisons of Complementary DNAs Encoding Aequorin Isotypes

Douglas C. Prasher,[‡] Richard O. McCann,[‡] Mathew Longiaru,[§] and Milton J. Cormier*[,‡]

Department of Biochemistry, University of Georgia, Athens, Georgia 30602, and Hoffmann-La Roche, Nutley, New Jersey 07110

Received September 5, 1986; Revised Manuscript Received November 14, 1986

ABSTRACT: Aequorin is the Ca^{2+}-activated photoprotein which participates in the bioluminescence from the circumoral ring of the hydromedusa *Aequorea victoria*. The nucleotide sequences of five aequorin cDNAs have been compared and shown to code for three aequorin isoforms. The cDNA AEQ1 contains the entire protein coding region of 196 amino acids. The other four cDNAs contain only 70–90% of the coding region and apparently code for at least two other isoforms whose amino acid sequences differ significantly from that encoded by AEQ1. The nucleotide sequences coding for the three isotypes differ at a minimum of 54 positions out of a total of 588 nucleotides necessary to code for apoaequorin. Of these nucleotide differences, 24 account for 23 amino acid replacements, substantiating the microheterogeneity observed during sequencing of purified native aequorin [Charbonneau, H., Walsh, K. A., McCann, R. O., Prendergast, F. G., Cormier, M. J., & Vanaman, T. C. (1985) *Biochemistry 24*, 6762–6771]. Comparison of the deduced cDNA translations with the native protein sequences suggests the loss of seven residues from the amino terminus during purification of aequorin from *Aequorea*. Aequorin rapidly extracted from the jellyfish using conditions to minimize proteolysis is shown to have a larger molecular weight than that of purified native aequorin. *Escherichia coli* expressed aequorin encoded by AEQ1 is shown to have the same molecular weight and isoelectric point as those of one of the isotypes rapidly extracted from *Aequorea*.

Organisms within each of four kingdoms are bioluminescent; they include bacteria, animals (insects, fish, earthworms), protoctists (dinoflagellates), and fungi (Herring, 1978). A majority of bioluminescent animals are of marine origin and include a large number of coelenterates (cnidarians and ctenophores). These coelenterate luminescent species have closely related bioluminescent systems (Hori et al., 1973, 1977; Ward & Cormier, 1975). Those systems found in the cni-

* Address correspondence to this author.
[‡] University of Georgia.
[§] Hoffmann-La Roche.

0006-2960/87/0426-1326$01.50/0

darians *Aequorea* and *Renilla* are particularly interesting due to the presence of energy-transfer systems and have been especially well characterized biochemically [reviewed in Cormier (1978)]. *Aequorea* bioluminescence is intriguing because the luminescent reaction is not catalyzed by a luciferase but by the photoprotein aequorin which was first described by Shimomura et al. (1962). Two proteins, aequorin and the green fluorescent protein (GFP),[1] are responsible for the green luminescence of *Aequorea* hydromedusae. The light emission from the hydromedusae is presumed to be induced by an increase of free calcium within the photocytes, generally caused by mechanical stimulation. The increase in Ca^{2+} causes oxidation of the aequorin-bound luciferin with the transfer of energy to the chromophore of GFP which ultimately releases green light (λ_{max} = 509 nm) (Johnson et al., 1962; Ward & Cormier, 1979).

Purified native aequorin consists of a single polypeptide (apoaequorin) containing an equimolar amount of bound coelenterate luciferin. In vitro, the addition of Ca^{2+} causes the oxidation of the luciferin with the release of blue light (λ_{max} = 469 nm). The oxyluciferin product remains tightly bound to the polypeptide in the presence of Ca^{2+}, and this complex is called discharged aequorin. The oxyluciferin product can be separated from the apoaequorin upon gel filtration of the discharged aequorin in the presence of EDTA (Shimomura & Johnson, 1975).

Aequorin purified from many thousand jellyfish has been sequenced by Charbonneau et al. (1985). The jellyfish-derived protein was shown to contain 189 amino acids. The N-terminal sequence was demonstrated to consist of Val-Lys-Leu-Thr .... Although the aequorin preparations used in the sequencing studies were electrophoretically homogeneous, the presence of at least 3 isotypes was demonstrated by the location of 17 sites of sequence microheterogeneity. Two amino acid variants were observed at each of 16 positions while 1 position had 3 different amino acid variants.

Two laboratories have recently cloned apoaequorin cDNAs (Prasher et al., 1985; Inouye et al., 1985). This report shows that the nucleotide sequences of several apoaequorin cDNAs differ significantly. This is not unexpected, since considerable microheterogeneity has been observed in the aequorin amino acid sequence (Charbonneau et al., 1985). However, more extensive microheterogeneity is observed in the cDNA translations reported here. In addition to this internal heterogeneity, limited proteolysis causes further heterogeneity. The cDNA sequence of one clone, AEQ1, suggests that the primary translation product contains an extra seven amino acids at the amino terminus compared to purified native aequorin. This makes the protein coding region of AEQ1 196 amino acids. An immunoblot analysis reported here suggests that aequorin is proteolytically cleaved during purification from *Aequorea* tissue, artifactually producing a large number of aequorin forms (Blinks & Harrer, 1975; Shimomura, 1986). Expression of the pAEQ1 cDNA in *Escherichia coli* reveals that it encodes a protein which has the same molecular weight and isoelectric point as that observed for one of the isotypes in *Aequorea* extracts.

Experimental Procedures

Materials. Restriction enzymes and Klenow fragment were purchased from BioLabs, Inc., Boehringer Mannheim, or Pharmacia P-L Biochemicals and used according to the manufacturer's directions. Pharmalytes and protein molecular weight standards were purchased from Pharmacia, Inc.

The *E. coli* strains JM105 (Yanisch-Perron et al., 1985) and SK1592 were obtained from J. Messing and S. Kushner, respectively. *E. coli* AEQ1 is SK1592 transformed with pAEQ1 (Prasher et al., 1985), and *E. coli* AEQ8 is JM105 transformed with pAEQ8. Plasmids pAEQ1 and pAEQ8 are derivatives of pBR322 and pUC9, respectively.

Methods. Both strands of each cDNA were sequenced according to the enzymatic method of Sanger et al. (1979) while the AEQ1 cDNA was also sequenced by using the chemical cleavage method of Maxam and Gilbert (1977).

SDS–polyacrylamide gel electrophoresis was done according to Laemmli and Favre (1973). Two-dimensional gels were run according to O'Farrell (1975). Pharmalytes (pH 4.2–4.9) were used in the IEF gels, and 15% acrylamide was used in the SDS–PAGE gels. Proteins were transferred overnight onto nitrocellulose using a Bio-Rad Trans-Blot at low field strength according to the manufacturer's instructions. The transfer buffer consisted of 192 mM glycine, 25 mM Tris, and 20% (v/v) methanol, pH 8.3. The nitrocellulose blots were stained for protein according to Hancock and Tsang (1983). Apoaequorin immunoblots were performed according to the Bio-Rad Immuno-Blot assay kit instructions. Unbound sites on the nitrocellulose were blocked with 3% gelatin followed by a 3–4-h incubation with the first antibody, rabbit anti-apoaequorin (diluted 500-fold), and a 1-h incubation with the second antibody, goat anti-rabbit IgG/HRP conjugate. Apoaequorin antigens were visualized by using the Bio-Rad color development reagent 4-chloro-1-naphthol.

Purified native apoaequorin was prepared by discharging native aequorin (Blinks et al., 1978) with excess calcium and removing bound oxyluciferin on Sephadex G-25 in the presence of 10 mM EDTA, pH 5.5. The protein was lyophilized and dissolved in SDS-denaturing buffer prior to electrophoresis.

The *Aequorea* extract used in Figures 3 and 4 was prepared as follows: a frozen circumoral ring, cut from a single organism (Prasher et al., 1985), was homogenized in a microfuge tube with a pestle (Kontes Glass) in 4 volumes of 10 mM EDTA and 15 mM Tris, pH 7.5. The protein was immediately precipitated by adding $^1/_5$th volume of 50% trichloroacetic acid. The precipitate was pelleted in the microfuge tube and washed 3 times with cold acetone. The pellet was dried in a Savant Speed-Vac and dissolved in SDS- or urea-denaturing buffer for one- and two-dimensional gels, respectively.

The level of apoaequorin expression in the *E. coli* strain containing pAEQ1 was estimated to be less than 0.0002% of the soluble protein (Prasher et al., 1986). The level of expression was easily increased 600-fold by subcloning the *Pst*I insert of pAEQ1 into pUC9 (Viera & Messing, 1982). The plasmid which resulted from the construction, pAEQ8, contained the cDNA in the correct orientation to utilize the inducible *lac* promoter (Prasher et al., 1986). Apoaequorin partially purified (Prasher et al., 1985) from AEQ8 was used in the immunoblot analyses described above. Lyophilized apoaequorin was dissolved in the appropriate electrophoresis sample buffer.

An extract of JM105 was used as a control. It was prepared by suspending the pellet from 0.5 mL of an overnight culture in 0.5 mL of SDS denaturing buffer and heating the extract at 100 °C for 5 min prior to electrophoresis.

Results

Sequence Comparisons of Aequorin cDNAs. The cDNA AEQ1 was previously shown to encode apoaequorin (Prasher

[1] Abbreviations: GFP, green fluorescent protein; IEF, isoelectric focusing; SDS–PAGE, sodium dodecyl sulfate–polyacrylamide gel electrophoresis; HRP, horseradish peroxidase; EDTA, ethylenediaminetetraacetic acid; M_r, relative molecular weight; Tris, tris(hydroxymethyl)aminomethane; HPLC, high-performance liquid chromatography.

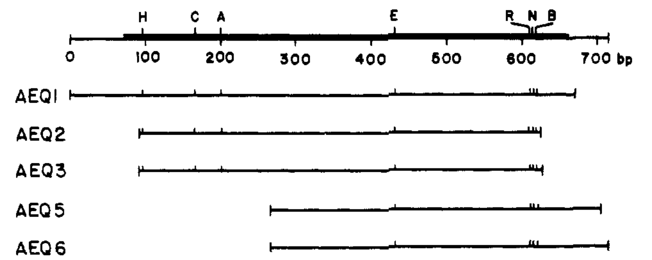

FIGURE 1: Restriction maps of aequorin cDNAs AEQ1, AEQ2, AEQ3, AEQ5, and AEQ6. The aequorin coding region (heavy line) reads from left to right with *Hin*dIII and *Bam*HI sites located at nucleotides 94 and 618, respectively, in Figure 2. The unique restriction enzyme sites indicated are *Hin*dIII (H), *Hin*cII (C), *Acc*I (A), *Eco*RV (E), *Rsa*I (R), *Nco*I (N), and *Bam*HI (B).

et al., 1985). Its nucleotide sequence and those of four other aequorin cDNAs have been determined. AEQ1 is the only one to contain the entire coding region. A restriction map of this clone and other aequorin cDNAs is shown in Figure 1. The DNA sequence of AEQ1 and its deduced amino acid sequence (Figure 2) indicate that it encodes a protein of 196 amino acids, 7 more than observed in the sequence of purified native aequorin (Charbonneau et al., 1985). The N-terminal peptide Met-Thr-Ser-Glu-Gln-Tyr-Ser (residues −7 through −1 in Figure 2), encoded near the 5′ end of AEQ1, was not detected at the N-terminus of purified aequorin. The N-terminal sequence of purified aequorin agrees with the residues beginning with Val-1 in Figure 2. Data presented below suggest the native polypeptide rapidly extracted from *Aequorea* also contains the N-terminal heptapeptide given above.

Five other aequorin cDNAs were identified in our *Aequorea* library (Prasher et al., 1985), but only AEQ2, AEQ3, AEQ5, and AEQ6 contain a significant portion of the coding sequence. The DNA sequences of two other aequorin cDNAs, AEQ2 and AEQ3, are compared for amino acids 2 through 177 with those of AEQ1 in Figure 2. The nucleotide sequences of AEQ2 and AEQ3 differ from the AEQ1 sequence by 54 and 50 bases, respectively (Figure 2), and their amino acid translations differ from the AEQ1 translation sequence at 19 and 16 residues, respectively. In contrast, the nucleotide sequence of AEQ2 differs from that of AEQ3 at only four nucleotides (Figure 2) although three of these substitutions

```
                                                                                                 -1
A:                                                                     met thr ser GLU gln tyr ser
B:              ctttgcaccaaaacaccacatcaaatctccagttgataaactaaatcgtcccaacggcaacaggccaac atg acc agc gaa caa tac tca
C:
D:
                                         30                         60                            90

                         10                                      20                                30
val lys leu thr PRO asp phe ASP asn pro LYS trp ile gly arg his lys his met phe asn phe leu asp val asn his asn gly ARG
gtc aag ctt aca cca gac ttc gac aac cca aaa tgg att gga cga cac aag cac atg ttt aat ttt ctt gat gtc aac cac aat gga agg
    ... ... ... t.. ... ... ... ... ... .g. ... ... ... ... ... ... ..t ... ..c ... ..c ... ... ... ... ... ... ... .aa
    ... ... ... t.. ... ... ... ... ... .g. ... ... ... ... ... ... ..t ... ..c ... ..c ... ... ... ... ... ... ... .aa
                         120                                     150                               180

                         40                                      50                                60
ile ser leu asp glu met val tyr lys ala ser asp ile val ile asn asn leu gly ala thr pro glu gln ala lys arg his lys asp
atc tct ctt gac gag atg gtc tac aag gcg tcc gat att gtt ata aac aat ctt gga gca aca cct gaa caa gcc aaa cgt cac aaa gat
... ... ... ... ... ... ... ... ... ..a ..t ... ... ..c ..c ..t ..c ... ... ... ... ... ..g ... ... ... ..a ... ... ...
... ... ... ... ... ... ... ... ... ..a ..t ... ... ..c ..c ..t ..c ... ... ... ... ... ..g ... ... ... ..a ... ... ...
                         210                                     240                               270

                         70                                      80                                90
ala val GLU ALA phe phe gly gly ala gly met lys tyr gly val glu thr GLU trp pro GLU tyr ile glu gly trp lys ARG leu ala
gct gta gaa gcc ttc ttc gga gga gct gga atg aaa tat ggt gta gaa act gaa tgg cct gaa tac atc gaa gga tgg aaa aga ctg gct
... ... ... ... ... ... ... ... ... ... ... ... ... ... ..g ... ... ..t ... ... .c. ..t ..t ... ... ... ... .a. t.. ...
... ... .g. .a. ... ... ... ... ... ... ... ... ... ... ..g ... ... ..t ... ... .c. ... ..t ... ... ... ... .a. t.. ...
                         300                                     330                               360

                         100                                     110                               120
SER GLU glu leu LYS ARG tyr SER lys asn GLN ILE thr leu ile arg LEU trp gly asp ala leu phe asp ile ILE asp lys asp gln
tcc gag gaa ttg aaa agg tat tca aaa aac caa atc aca ctt att cgt tta tgg ggt gat gca ttg ttc gat atc att gac aaa gac caa
a.t ..t ... ... g.g .aa ..c g.c ... ... g.. cca ..g ..c ..c ... a.. ... ... ... ..t ... ..t ... ... g.. ... ... ..t ...
a.t ..t ... ... g.g .aa ..c g.c ... ... g.. cca ..g ..c ..c ... a.. ... ... ... ..t ... ..t ... ... g.. ... ... ..t ...
                         390                                     420                               450

                         130                                     140                               150
asn gly ala ile SER leu asp glu trp lys ala tyr thr lys SER ala gly ile ile gln SER ser glu asp cys glu glu thr phe ARG
aat gga gct att tca ctg gat gaa tgg aaa gca tac acc aaa tct gat ggc atc atc caa tcg tca gaa gat tgc gag gaa aca ttc aga
... ... ..c ... a.. ... ... ... ... ... ... ... ... ... g.. .c. ..t ... ... ... ..a ... ... ... ... ... ... ... ... ...
... ... ..c ... a.. ... ... ... ... ... ... ... ... ... g.. .c. ..t ... ... ... ..a ... ... ... ... ... ... ... ... ...
                         480                                     510                               540

                         160                                     170                               180
val cys asp ile asp glu SER gly gln leu asp val asp glu met thr arg gln his leu gly phe trp tyr thr met asp pro ala cys
gtg tgc gat att gat gaa agt gga cag ctc gat gtt gat gag atg aca aga caa cat tta gga ttt tgg tac acc atg gat cct gct tgc
... ... ... ... ... ... ... ... ..a ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
... ... ... ... ... ... .a. ... ..a ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
                         570                                     600                               630

glu lys leu tyr gly gly ala val pro ...
gaa aag ctc tac ggt gga gct gtc ccc taa gaaactctgcgg
                         660
```

FIGURE 2: DNA sequences of AEQ1, AEQ2, and AEQ3 and the deduced amino acid sequence of AEQ1. The upper line (A) shows the amino acid sequence deduced from the AEQ1 DNA sequence (line B). Lines C and D show only the nucleotide differences of AEQ3 and AEQ2, respectively, with AEQ1. The numbering of the amino acid sequence is that of Charbonneau et al. (1985); the Val at residue 1 is the amino terminal in purified native aequorin. The underlined sequences are putative *E. coli* ribosome binding sites. The boxes indicate the putative Ca^{2+} binding domains. The amino acids in boldfaced type are those residues shown to be heterogeneous by either nucleotide or protein sequence comparisons.

Table I: Heterogeneous Amino Acid Residues in Aequorin

residue	AEQ1	AEQ2	AEQ3	AQ440	protein sequence
−4[a]	Glu			Lys	
5	Pro	Ser	Ser	Ser	Ser (Pro)
8	Asp	Asp	Asp	Asp	Asp (Asn)
11	Lys	Lys	Lys	Lys	Arg (Lys)
30	Arg	Lys	Lys	Lys	Lys (Arg)
63[a]	Glu	Gly	Glu	Glu	Glu
64[a]	Ala	Asp	Ala	Ala	Ala
78	Glu	Asp	Asp	Asp	Asp (Glu)
81	Glu	Ala	Ala	Ala	Ala (Glu)
88	Arg	Lys	Lys	Lys	Lys (Arg)
91[a]	Ser	Thr	Thr	Thr	Thr
92	Glu	Asp	Asp	Asp	Asp (Glu, Cys)
95[a]	Lys	Glu	Glu	Glu	Glu
96	Arg	Lys	Lys	Lys	Lys (Arg)
98	Ser	Ala	Ala	Ala	Ala (Ser)
101	Gln	Glu	Glu	Glu	Glu (Gln)
102	Ile	Pro	Pro	Pro	Pro (Ile)
107	Leu	Ile	Ile	Ile	Ile (Leu)
116	Ile	Val	Val	Val	Val (Ile)
125[a]	Ser	Thr	Thr	Thr	Thr
135	Ser	Ala	Ala	Ala	Ala (Ser)
141	Ser	Ser	Ser	Ser	Ser (Thr)
150	Arg	Arg	Arg	Arg	Arg (Lys)
157[a]	Ser	Asn	Ser	Ser	Ser

[a] These residues were not observed to be heterogeneous during sequencing of purified native aequorin.

create amino acid replacements in codons 63, 64, and 157. The fourth nucleotide difference occurs in codon 82 (Figure 2).

The cDNAs AEQ5 and AEQ6 encode only the C-terminal portion of aequorin (residues 59–196) in addition to 3′-untranslated sequences (data not shown). The coding sequences of AEQ5 and AEQ6 are identical with that of AEQ3 except for a single nucleotide difference in AEQ5. Nucleotide 459 (Figure 2) is a T in AEQ5 instead of a C observed in AEQ3 and AEQ6. AEQ1 also has a T at this position.

The heterogeneity among the amino acid translations of AEQ1, AEQ2, and AEQ3 is consistent with the microheterogeneity observed at 17 residues during sequencing of the native protein (Charbonneau et al., 1985). In the native protein, 1 amino acid predominated in abundance at each of these 17 positions. Those heterogeneous residues in the sequenced native aequorin are compared in Table I to the respective residues encoded by AEQ1, AEQ2, and AEQ3 as well as to the aequorin cDNA AQ440 isolated by Inouye et al. (1985). At 14 of these residues, AEQ1 codes for the less abundant amino acid detected during protein sequencing. Ignoring the seven extra amino acids at the 5′ end of the coding region of AEQ1, the cDNA differs from the protein sequence at three additional residues. Two are conservative replacements (Ser-91, Ser-125) whereas the third results in a change in charge (Lys-95). On the other hand, the amino acids encoded by AEQ2, AEQ3, and AQ440 at 16 of the 17 heterogeneous residues are in higher abundance in the protein sequence (Table I). The translation of AEQ2 also contains three additional amino acid replacements in the protein sequence: Glu-63 → Gly, Ala-64 → Asp, and Ser-157 → Asn.

Bacterial Expression of Apoaequorin. Expression of apoaequorin is detected in two *E. coli* strains we have constructed. One strain contains pAEQ1, the original cDNA construction isolated from the *Aequorea* library (Prasher et al., 1985). The other strain contains pAEQ8, a derivative of pUC9 which uses the *lac* promoter to increase expression (Prasher et al., 1986). We do not believe the expressed polypeptides to be fusion products because no qualitative differences in molecular weight or charging ability have been observed on the apoaequorins isolated from either strain. The

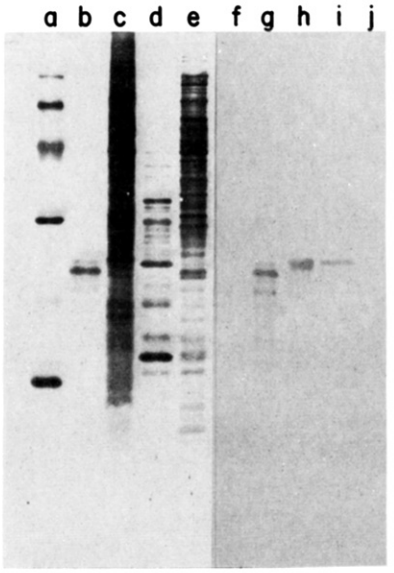

FIGURE 3: Immunoblot analysis of purified native aequorin, *Aequorea* circumoral ring extracts, and *E. coli* expressed apoaequorin. Protein extracts were prepared, separated on a SDS–PAGE gel, and transferred to nitrocellulose as described under Methods. Half of the filter was stained with India ink (lanes a–e) and the other half treated with anti-aequorin (lanes f–j). The following were applied to the SDS–PAGE gel: protein standards (lanes a and f); purified native aequorin (1 μg in lane b and 0.5 μg in lane g); extract of a single *Aequorea* circumoral ring (lanes c and h); partially pufied extract of AEQ8 (1 μg in lane d and 5 μg in lane i); and an extract of the control stain JM105 (lanes e and j). See Materials for a description of AEQ8. The gel is calibrated with phosphorylase *b* (M_r 94000), bovine serum albumin (M_r 67000), ovalbumin (M_r 43000), carbonic anhydrase (M_r 30000), soybean trypsin inhibitor (M_r 20100), and α-lactalbumin (M_r 14400). The soybean trypsin inhibitor protein standard (lane a) does not stain well with India ink (A. Harmon, unpublished observations).

expression can be explained by the fortuitous occurence of two procaryotic ribosome binding sites within the cDNA insert occurring at 4 and 11 bases upstream from the initiation codon (Figure 2). Transcription of the apoaequorin cDNA probably initiates at the *bla* promoter in pAEQ1 and the *lac* promoter in pAEQ8 (see Methods). Other supporting evidence for translation initiation within the cDNA includes comigration on two-dimensional gels of both the *E. coli* derived and the *Aequorea*-derived apoaequorins (see below). In addition, the two apoaequorins have similar charging kinetics (Prasher et al., 1985).

The apoaequorin expressed in *E. coli* has a different molecular weight from aequorin highly purified from *Aequorea*. By gel filtration, we previously determined the molecular weight of apoaequorin isolated from *E. coli* AEQ1 to be 20600; this is approximately 1000 daltons larger than purified native aequorin (Prasher et al., 1985). A similar phenomenon is observed if apoaequorin purified from *E. coli* AEQ8 is compared to purified native aequorin by SDS–polyacrylamide gel electrophoresis. The migration of these proteins was determined by immunoblot analysis as described under Experimental Procedures. Lanes g and i of Figure 3 show that *E. coli* expressed apoaequorin migrates as a higher molecular weight species when compared to purified native aequorin. Their apparent molecular weight difference is 1900. No significant difference is observed in the migration of native aequorin and discharged aequorin on an SDS gel (data not shown).

Mature Form of Aequorin in Aequorea. The immunoreactive proteins in an *Aequorea* extract, prepared such that proteolysis was minimized (Figure 3, lane h), also have higher

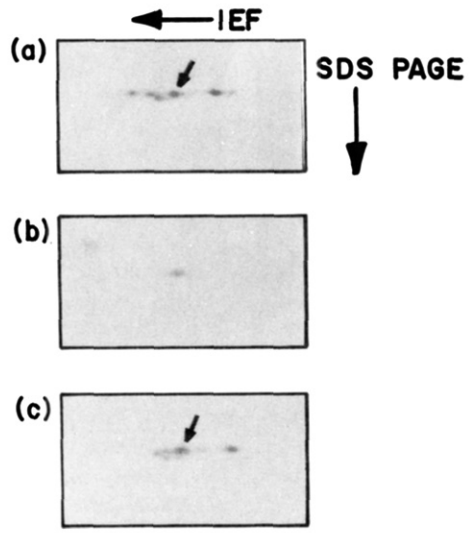

FIGURE 4: Immunoblot analysis of aequorin isotypes. Extracts of a single *Aequorea* ring (a), AEQ8 (b), and both together (c) were separated on two-dimensional gels, transferred to nitrocellulose, and treated with anti-aequorin described under Methods. The arrow indicates the aequorin isotype that comigrates with the *E. coli* expressed apoaequorin.

molecular weights than purified "native" aequorin (lane g). The predominant and higher molecular weight immunoreactive species (M_r 27 100) of the *Aequorea* extract migrates with the same mobility as the apoaequorin expressed in *E. coli* AEQ8 (lane i). The faster migrating immunoreactive protein detected in the *Aequorea* extract migrated with a molecular weight of 25 900. Purified native aequorin (lane g) migrated corresponding to a molecular weight of 25 200. These molecular weights are slightly larger than the molecular weight of 22 470 calculated from the amino acid sequence encoded by the AEQ1 cDNA. These results suggest limited proteolysis occurs during the isolation of native aequorin from *Aequorea*.

Evidence for Multiple Isotypes in a Single Aequorea. We are able to separate and detect on immunoblots of two-dimensional gels five aequorin species from the *Aequorea* extract, four of which migrate with the same molecular weight (Figure 4a). To demonstrate that the apoaequorin expressed in *E. coli* AEQ8 represents only one isotype, its migration on two-dimensional gels is compared in Figure 4 to that of the aequorin isotypes in an extract of a single *Aequorea* circumoral ring. The results indicate that pAEQ8-encoded apoaequorin comigrates with an isotype in high abundance in the *Aequorea* extract (Figure 4b,c). Note that the aequorin isotypes from the circumoral ring separate into two groups having slightly different molecular weights (Figure 4a). This is also observed by one-dimensional SDS–PAGE (Figure 3, lane h).

Discussion

Comparison of the coding regions of three cDNAs encoding aequorin shows there are a large number of amino acid differences among the isotypes represented by these cDNAs. These differences explain most of the microheterogeneity (Charbonneau et al., 1985) observed at 17 positions during the sequencing of native aequorin. There are also 17 amino acid differences in the translations of AEQ1 and AQ440 (Inouye et al., 1985), both of which contain the entire coding region. This represents 8.6% variation in their amino acid sequences. There are even more replacements in AEQ2 which does not contain the entire coding sequence. Our cDNA comparisons demonstrate that there are seven more heterogeneous residues in aequorin not detected in the protein sequencing (Table I). One of those occurs at residue −4 present only in cDNAs AEQ1 and AQ440. These additional replacements were not noted during protein sequencing because they probably existed below the detection limit.

In general, the amino acids observed at the corresponding heterogeneous sites in the AEQ1 translation were those detected in low abundance during protein sequencing (Table I). In contrast, the amino acids at these same sites in the AQ440 translation were the major variants. For example, Arg is found at residue 30 in the AEQ1 cDNA translation where Lys is the deduced amino acid in the AQ440 translation. When native aequorin was sequenced by Charbonneau et al. (1985), Arg was in lower abundance than Lys at residue 30. Exceptions to this major/minor rule occur in the AEQ1 amino acid sequence at residues 8, 141, and 150. The one exception in the AQ440 sequence occurs at residue 11. Lys is encoded in both cDNAs at position 11, but Arg was reported to be the more dominant amino acid in purified aequorin. Protein sequencing of the intact protein indicated only Lys at residue 11 and apparently gave no indication for the presence of Arg at this position. However, upon sequencing of smaller variant peptides, Arg predominated over Lys at this position [see Figure 2 of Charbonneau et al. (1985)]. The codons of residue 11 in all the cDNAs reported here support the existence of only Lys as the predominant species. The generalization above suggests that AQ440 represents one of the more abundant aequorin isotypes and the AEQ1 cDNA represents a less abundant one. Despite this, the *E. coli* expressed apoaequorin comigrates on two-dimensional gels with the isotype in highest abundance in the *Aequorea* extract. This conflicting data can be explained if several isotypes comigrate on the two-dimensional gels. In fact, the isotypes encoded by AEQ1 and AQ440 can be expected to comigrate because the net charge of their amino acid side chains is identical (calculated to equal zero). Additional heterogeneity in aequorin cDNAs not represented by AEQ1, AEQ2, or AEQ3 can be expected because certain amino acids observed in the protein sequence are not found in the cDNA translations. Those residues include Asn-8, Arg-11, Cys-92, Thr-141, and Lys-150.

Assuming that all of the isotypes present in *Aequorea* are luminescent, the multiple forms represented by cDNAs AEQ1, AEQ2, and AEQ3 may have different physical or chemical properties due to their sequence variations. This is not supported by the work of Shimomura (1986), who compared physical properties of various molecular forms of aequorin. Only slight differences were observed between the isoforms which had been isolated by HPLC. Their similarity may be due to an inherent problem with amino-terminal proteolysis when aequorin is isolated from large numbers of *Aequorea* (see below). The N-terminal residues may affect light emission because preliminary evidence shows that the rate of light emission is significantly greater from "charged" recombinant apoaequorin than from purified native aequorin (D. Prasher, unpublished observations). There is a precedent for the existence of isotypes of photoproteins having different physical properties. Two isoforms of a related photoprotein, mnemiopsin, were shown to have different kinetics (Ward & Seliger, 1974). There may be environmental or behavioral reasons for *Aequorea* possessing numerous aequorin isotypes. However, the benefit that *Aequorea* gains from its bioluminescence remains to be determined.

AEQ2 and AEQ3 may represent alleles because their amino acid translations differ by only three residues. Generally, alleles at a single locus are assumed to differ by only one to three residues. For example, greater than 200 hemoglobin

variants were shown to differ by no more than three residues (Hung et al., 1972). Likewise, allelic forms of bovine carboxypeptidase (Petra et al., 1969), human immunoglobulin κ chains (Terry et al., 1969), and human haptoglobulins (Black & Dixon, 1968) differ by less than four residues. On the other hand, two allelic forms of human placental alkaline phosphatase (Henthorn et al., 1986) differ by seven amino acids. However, the substitutions in the latter protein represent only 1.3% variation in the primary sequence due to the polypeptide length (513 residues). In contrast, the deduced amino acid sequence of AEQ1 differs from that of AEQ2 and AEQ3 by 19 (10.6% variation) and 16 (8.9% variation) residues, respectively. We suspect AEQ1 is not an allele of AEQ2 or AEQ3 due to the high level of substitutions.

The heterogeneity of aequorin isotypes in a single organism can be explained either by a multigene family or by alternative splicing of a primary mRNA transcript from multiple exons comprising the aequorin gene. A definitive explanation of the microheterogeneity must wait until genomic clones become available.

There is now considerable experimental evidence to support the view that native aequorin undergoes limited proteolysis at the N-terminus. Most of the proteolysis occurs between amino acid residues −1 (Ser) and 1 (Val) while the mature protein has seven additional N-terminal amino acid residues beginning with N-terminal methionine (Figure 2). This conclusion contrasts with that of Inouye et al. (1985), who proposed that the N-terminus of the mature protein begins with valine and lacks these seven additional N-terminal residues. The evidence which suggests that the mature form of aequorin exists as a 196 amino acid polypeptide, the same number as observed in the AEQ1 and AQ440 cDNA translations, follows. (1) The molecular weight determined by SDS–PAGE of *E. coli* expressed aequorin corresponds to that of aequorin contained in extracts of a single circumoral ring, and both are slightly greater than the molecular weight of purified aequorin (Figure 3). (2) This molecular weight difference can be accounted for by an additional seven amino acids. (3) The N-terminus of native aequorin begins with valine (Charbonneau et al., 1985) whereas the N-terminus of the *E. coli* expressed aequorin contains seven additional amino acids beginning with methionine (data not shown). (4) Aequorin translated in vitro migrated during SDS–PAGE corresponding to a slightly higher molecular weight than that of purified native aequorin (Prasher et al., 1985). The putative Ser_{-1}–Val_1 cleavage site is most unusual because we are unaware of any protease reported to have such substrate specificity. Perhaps an amino peptidase in *Aequorea* is able to degrade the N-terminal sequentially only as far as Val-1.

Proteolysis of aequorin during isolation may explain the presence of at least 13 molecular species of the protein in a purified preparation of native aequorin (data not shown), whereas only 5 isotypes are observed in a crude extract of a single *Aequorea* (Figure 4). Furthermore, four of the latter aequorin isoforms have similar mobilities in SDS gels, suggesting that the lower molecular weight form might also be derived from proteolysis (Figure 4). Shimomura (1986) has observed eight molecular forms of aequorin from a large population of jellyfish. The results presented here suggest that some of these isotypes may also have been derived via proteolysis.

Charbonneau et al. (1985) have shown that aequorin is structurally related to calmodulin which has four high-affinity calcium binding domains, sometimes referred to as EF hands (Kretsinger, 1980). Aequorin binds 3 mol of calcium per mole of protein (Shimomura & Johnson, 1970; Allen et al., 1977; Blinks et al., 1982). This agrees with the three putative calcium binding domains observed in the protein sequence. Aequorin appears to have evolved from a four-domain precursor by a small insertion within the second calcium binding domain. This may be the location of luciferin binding since there is a larger number of nonpolar amino acids in residues 40–75, a sequence which is highly conserved in the aequorin isotypes (Figure 2). The only heterogeneous residues (63 and 64) in this region were observed in AEQ2. Nevertheless, a nonpolar amino acid and an acidic amino acid are retained in these replacements (Table I). The carboxy-terminal region of aequorin may also be involved in luciferin binding because it too contains many nonpolar amino acids and appears to be conserved in the aequorin cDNA translations. Also of interest is a single heterogeneous residue within each putative Ca^{2+} binding domain (30, 125, and 157). We do not expect any of these replacements to affect Ca^{2+} sensitivity because all are compatible as Ca^{2+} ligands in EF-hand domains (Kretsinger, 1980).

ACKNOWLEDGMENTS

We thank Alice Harmon, Claiborne Glover, Virginia Eckenrode, and Jon Shuman for critical comments on the manuscript. We are indebted to Debbie Vaughn for her assistance in preparation of the manuscript.

REFERENCES

Allen, D. G., Blinks, J. R., & Prendergast, F. G. (1977) *Science (Washington, D.C.) 195*, 996–998.

Black, J. A., & Dixon, G. H. (1968) *Nature (London) 218*, 736–741.

Blinks, J. R., & Harrer, G. C. (1975) *Fed. Proc., Fed. Am. Soc. Exp. Biol. 34*, 474.

Blinks, J. R., Mattingly, P. H., Jewell, B. R., van Leeuwen, M., Harrer, G. C., & Allen, D. G. (1978) *Methods Enzymol. 57*, 292–328.

Blinks, J. R., Wier, W. G., Hess, P., & Prendergast, F. G. (1982) *Prog. Biophys. Mol. Biol. 40*, 1–114.

Charbonneau, H., Walsh, K. A., McCann, R. O., Prendergast, F. G., Cormier, M. J., & Vanaman, T. C. (1985) *Biochemistry 24*, 6762–6771.

Cormier, M. J. (1978) in *Bioluminescence in Action* (Herring, P., Ed.) pp 78–108, Academic Press, London.

Feinberg, A. P., & Vogelstein, B. (1983) *Anal. Biochem. 132*, 6–13.

Hancock, K., & Tsang, V. C. W. (1983) *Anal. Biochem. 133*, 157–162.

Henthorn, P. S., Knoll, B. J., Raducha, M., Rothblum, K. N., Slaughter, C., Weiss, M., Lafferty, M. A., Fischer, T., & Harris, H. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 5597–5601.

Herring, P. (1978) *Bioluminescence in Action*, Academic Press, New York.

Hori, K., Wampler, J. E., Matthews, J. C., & Cormier, M. J. (1973) *Biochemistry 12*, 4463–4468.

Hori, K., Charbonneau, H., Hart, R. C., & Cormier, M. J. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 4285–4287.

Hunt, L., Sochard, M. R., & Dayhoff, M. O. (1972) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., Ed.) pp 67–88, National Biomedical Research Foundation, Silver Spring, MD.

Inouye, S., Noguchi, M., Sakaki, Y., Takagi, Y., Miyata, T., Iwanaga, S., Miyata, T., & Tsuji, F. I. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 3154–3158.

Johnson, F. H., Shimomura, O., Saiga, Y., Gershman, L. C., Reynolds, G. T., & Waters, J. R. (1962) *J. Cell. Comp. Physiol. 60*, 85–103.
Kretsinger, R. H. (1980) *CRC Crit. Rev. Biochem. 8*, 119–174.
Laemmli, U. K., & Favre, M. (1973) *J. Mol. Biochem. 80*, 575–599.
Maxam, A. M., & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 560–564.
O'Farrell, P. H. (1975) *J. Biol. Chem. 250*, 4007–4021.
Petra, P. H., Bradshaw, R. A., Walsh, K. A., & Neurath, H. (1969) *Biochemistry 8*, 2762–2768.
Prasher, D. C., McCann, R. O., & Cormier, M. J. (1985) *Biochem. Biophys. Res. Commun. 126*, 1259–1268.
Prasher, D. C., McCann, R. O., & Cormier, M. J. (1986) *Methods Enzymol. 133*, 288–298.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.
Shimomura, O. (1986) *Biochem. J. 234*, 271–277.
Shimomura, O., & Johnson, F. H. (1970) *Nature (London) 227*, 1356–1357.
Shimomura, O., & Johnson, F. H. (1975) *Nature (London) 256*, 236–238.
Shimomura, O., Johnson, F. H., & Saiga, Y. (1962) *J. Cell. Physiol. 59*, 223–240.
Southern, E. M. (1975) *J. Mol. Biol. 98*, 503–517.
Sparrow, A. H., Price, H. J., & Underbrink, A. G. (1972) *Brookhaven Symp. Biol. 23*, 451–494.
Terry, W. D., Hood, L. E., & Steinberg, A. G. (1969) *Proc. Natl. Acad. Sci. U.S.A. 63*, 71–77.
Vieira, J., & Messing, J. (1982) *Gene 19*, 259–268.
Wahl, G. M., Stern, M., & Stark, C. R. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 3683–3687.
Ward, W. W., & Seliger, H. H. (1974) *Biochemistry 13*, 1491–1499.
Ward, W. W., & Cormier, M. J. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 2530–2534.
Ward, W. W., & Cormier, M. J. (1979) *J. Biol. Chem. 254*, 781–788.
Yanisch-Perron, C., Vieira, J., & Messing, J. (1985) *Gene 33*, 103–119.

Structure, Polymorphism, and Novel Repeated DNA Elements Revealed by a Complete Sequence of the Human α-Fetoprotein Gene†

Peter E. M. Gibbs, Rita Zielinski, Carol Boyd, and Achilles Dugaiczyk*
Department of Biochemistry, University of California, Riverside, California 92521
Received September 10, 1986; Revised Manuscript Received November 3, 1986

ABSTRACT: The human α-fetoprotein gene spans 19 489 base pairs from the putative "Cap" site to the polyadenylation site. It is composed of 15 exons separated by 14 introns, which are symmetrically placed within the three domains of α-fetoprotein. In the 5′ region, a putative TATAAA box is at position −21, and a variant sequence, CCAAC, of the common CAT box is at −65. Enhancer core sequences $GTGG^{TTT}_{AAA}G$ are found in introns 3 and 4, and several copies of glucocorticoid response sequences $AGA^{T}_{A}CAG^{T}_{A}$ are found on the template strand of the gene. There are six polymorphic sites within 4690 base pairs of contiguous DNA derived from two allelic α-fetoprotein genes. This amounts to a measured polymorphic frequency of 0.13%, or 6.4×10^{-4}/site, which is about 5–10 times lower than values estimated from studies on polymorphic restriction sites in other regions of the human genome. There are four types of repetitive sequence elements in the introns and flanking regions of the human α-fetoprotein gene. At least one of these is apparently a novel structure (designated Xba) and is found as a pair of direct repeats, with one copy in intron 7 and the other in intron 8. It is conceivable that within the last 2 million years the copy in intron 8 gave rise to the repeat in intron 7. Their present location on both sides of exon 8 gives these sequences a potential for disrupting the functional integrity of the gene in the event of an unequal crossover between them. There are three Alu elements, one of which is in intron 4; the others are located in the 3′ flanking region. A solitary Kpn repeat is found in intron 3. The Xba and Kpn repeats were only detected by complete sequencing of the introns. Neither X, Xba, nor Kpn elements are present in the related human albumin gene, whereas Alu's are present in different positions. From phylogenetic evidence, it appears that Alu elements were inserted into the α-fetoprotein gene at some time postdating the mammalian radiation 85 million years ago.

The fundamental role of mutations in the evolution of species, and in the occurrence and persistence of genetic disorders, makes them an important subject for investigation. Phylogeny of lineages may be reconstructed from DNA sequences, and genetic disorders can be traced by following DNA polymorphism in affected families. While the difference between two chromosomes in one individual appears to be 0.1–0.5%, the difference in DNA sequence between two primates can be as low as 1–2% (Sibley & Ahlquist, 1984). Clearly, accurate and quantitative sequence data are required if we wish to extract meaningful information from the sequences of our genomes.

Studies in this laboratory have been directed toward understanding the evolution of a small family of single-copy genes: those for serum albumin, its fetal counterpart α-fetoprotein (AFP)[1] [e.g., Minghetti et al. (1985)], and the vitamin

† This work was supported by Grant 1RO1 HD19281 from the National Institutes of Health.
* Address correspondence to this author.

0006-2960/87/0426-1332$01.50/0