Volume 13 Number 19 1985 Nucleic Acids Research

The nucleotide sequence of a nematode vitellogenin gene

John Spieth, Karen Denison, Erin Zucker and Thomas Blumenthal

Program in Molecular, Cellular and Developmental Biology, and Department of Biology, Indiana University, Bloomington, IN 47405, USA

Received 12 August 1985; Accepted 4 September 1985

ABSTRACT

The nematode, Caenorhabditis elegans, contains a family of six genes that code for vitellogenins. Here we report the complete nucleotide sequence of one of these genes, vit-5. The gene specifies a mRNA of 4869 nucleotides, including untranslated regions of 9 bases at the 5' end and 51 bases at the 3' end. Vit-5 contains four short introns totalling 218 bp. The predicted vitellogenin, yp170A, has a molecular weight of 186,430. At its N terminus it is clearly related to the vitellogenins of vertebrates. However, the vit-5-encoded protein does not contain a serine-rich sequence related to the vertebrate vitellin, phosvitin. In fact, the amino acid composition of the nematode protein is very similar to that of the vertebrate protein without phosvitin. Vit-5 has a highly asymmetric codon choice dictionary. The favored codons are different from those favored in other organisms, but are characteristic of highly expressed C. elegans genes. The strong selection against rare codons is not as great near the 5' end of the gene; rare codons are 15 times more frequent within the first 54 bp than in the next 4.8 kb.

INTRODUCTION

Four distinct yolk proteins have been identified in the nematode, Caenorhabditis elegans (1). Two of these, yp170A and yp170B, are translated as polypeptides estimated to be 170,000 daltons (2). These vitellogenins undergo little if any processing before they are packaged into yolk platelets. In contrast, the two smaller yolk proteins, yp115 and yp88, are cut from a single precursor polypeptide, VIT180 (3). yp170A and B and VIT 180 are synthesized only in the intestine of the adult hermaphrodite worm, from which they are cotranslationally secreted into the body cavity (3, 4). There, VIT180 is cleaved and the yolk proteins are subsequently taken up into developing oocytes.

The vitellogenins are encoded by a small family of genes (5, 6). yp170A is specified by vit-3, vit-4, and vit-5. These three genes are at least 95% homologous to one another at the nucleotide sequence level. Vit-5 mRNA is much more abundant than that from vit-3 and vit-4 (5; Cane, unpublished observations). yp170B is specified by vit-2. Vit-1 is a pseudogene, about

95% identical to vit-2 (7). The vit-1,2 sub-family is about 70% homologous to the vit-3,4,5 sub-family. VIT180 is encoded by a very distantly related family member, vit-6 (6).

We have recently reported the nucleotide sequences of the regions surrounding the 5' ends of five of the C. elegans vitellogenin genes (7). Our results indicated that these genes were surprisingly closely related to the vitellogenin genes of vertebrates. Like the nematode genes, the vertebrate genes are expressed in a tissue of endodermal origin, the liver. Like yp170A and yp170B, the vertebrate vitellogenins are transported intact to the oocytes. However, unlike the nematode vitellogenins, the vertebrate proteins are cleaved in the oocyte to form vitellins: phosvitin and the lipovitellins (8). In this paper we report the complete nucleotide sequence of vit-5, the first member of the vertebrate/nematode vitellogenin gene family to be sequenced. Our data indicate that although the nematode gene is closely related to the vertebrate gene near the 5' end, the region encoding the serine-rich vertebrate protein, phosvitin, is not present in vit-5.

RESULTS AND DISCUSSION

Characteristics of vit-5

The complete nucleotide sequence of vit-5, along with 261 bp of 5' flanking and 27 bp of 3' flanking DNA, is shown in Fig. 1. The location of the 5' end of the mRNA was determined by primer extension using a synthesized 15 bp oligonucleotide primer, accompanied by dideoxy-sequencing. A single, unequivocal extension product was found (data not shown). Similar experiments were performed with vit-2, vit-4, and vit-6. In each case a single extension product, the same distance 3' from a canonical TATA box sequence was found. The 3' end of the gene was located by sequencing a 3' terminal cDNA clone containing a 50 base poly(A) stretch. The gene contains four short introns, totaling 218 bp. The presence of the two introns near the 3' end of the gene was directly demonstrated by sequencing cDNA clones which overlap this region. The presence of the other two introns is inferred: they contain canonical splice sites at their 5' and 3' boundaries (Table 1); they contain stop codons in all three reading frames; and they are at least 70% A + T. Excision of these proposed introns at the positions indicated allows translation utilizing primarily the highly favored codons typical of the C. elegans vitellogenin genes (see below).

We presume that translation begins at the first AUG in each mRNA. This assumption is supported by: 1) The presence of a highly hydrophobic amino

```
                                                                -261  acttcattcaagcaatcattcagaaaaactta  -230
taacggtcacatatacgtcgaccacgatatgcacgttgaattagaagagtgacagctgtactctgatcgctacctgaacgacatgcgcacttcagtttcaactgataatgcggtaaaag  -111
ttcagattgacattgactttatcgaataaatctgttaagataggatgatttgaaggaatagtgtcaatttcagttgcatctataaaaagggtaacggaggaaccatagtcACTCTCGCA  +9

ATG AAG TCG ATA ATC ATT GCC TCT CTT GTC GCC TTG GCG ATT GCC GCC TCA CCG GCT CTT GAC CGT ACC TTC TCT CCA AAG AGC GAA TAC  +99
 M   K   S   I   I   I   A   S   L   V   A   L   A   I   A   A   S   P   A   L   D   R   T   F   S   P   K   S   E   Y

GTC TAC AAA TTT GAC GGA CTT CTT CTC TCT GGA CTC CCA ACC CGA TCT TCC GAT GCT TCC CAA ACC CTG ATT TCT TGC CGT ACC CGT CTT  +189
 V   Y   K   F   D   G   L   L   L   S   G   L   P   T   R   S   S   D   A   S   Q   T   L   I   S   C   R   T   R   L

CAA GCT GTT GAT GAT CGT TAC ATT CAT CTT CAA TTG ACT GAT ATT CAA TAC TCT GCT TCC CAC ATT CCA CAA TCT GAG CAA TGG CCA AAG  +279
 Q   A   V   D   D   R   Y   I   H   L   Q   L   T   D   I   Q   Y   S   A   S   H   I   P   Q   S   E   Q   W   P   K

ATC GAA TCT TTG GAG CAA CGT GAG CTT TCC GAT GAG TTC AAG GAG CTT CTT GAG CTT CCA TTC CGT GCT CAA ATC AGA AAT GGA CTT ATT  +369
 I   E   S   L   E   Q   R   E   L   S   D   E   F   K   E   L   L   E   L   P   F   R   A   Q   I   R   N   G   L   I

TCT GAG ATC CAA TTC TCT TCC GAA GAT GCC GAG TGG TCC AAG AAC GCC AAA AGA TCG ATT CTC AAT CTC TTC TCT CTC CGC AAG TCA GCT  +459
 S   E   I   Q   F   S   S   E   D   A   E   W   S   K   N   A   K   R   S   I   L   N   L   F   S   L   R   K   S   A

CCA GTT GAT GAG ATG AAC CAA GAT CAG AAA GAT ATG GAA TCC GAC AAG GAT TCT GTT TTC TTC AAT GTT CAT GAA AAG ACC ATG GAA GGA  +549
 P   V   D   E   M   N   Q   D   Q   K   D   M   E   S   D   K   D   S   V   F   F   N   V   H   E   K   T   M   E   G

GAC TGC CGA AGT CGC TTA CAC ATT GTT CAA GAG GGA GAG AAG ACC ATC TAC ACC AAA TCT GTC AAC TTC GAC AAA TGC ATC ACT CGC CCA  +639
 D   C   R   S   R   L   H   I   V   Q   E   G   E   K   T   I   Y   T   K   S   V   N   F   D   K   C   I   T   R   P

GAG ACT GCT TAT GGT CTT CGT TTT GGA TCT GAG TGC AAG GAA TGC GAG AAG GAG GGG CAA TTT GTT AAG CCA CAA ACT GTC TAC ACC TAC  +729
 E   T   A   Y   G   L   R   F   G   S   E   C   K   E   C   E   K   E   G   Q   F   V   K   P   Q   T   V   Y   T   Y

ACC TTC AAG AAC GAG AAA TTG CAA GAA TCT GAG GTT CAT TCC GTT TAC ACT TTG AAT GTC AAT GGA CAA GAG GTC GTC AAG TCT GAG ACT  +819
 T   F   K   N   E   K   L   Q   E   S   E   V   H   S   V   Y   T   L   N   V   N   G   Q   E   V   V   K   S   E   T

CGC GCC AAG GTC ACT TTC GTC GAG GAG AGC AAG ATC AAC AGA GAG ATC AAG AAG gtattctatttcaaaatatattttacgggcttttgatagctaactct  +920
 R   A   K   V   T   F   V   E   E   S   K   I   N   R   E   I   K   K

aacagcagcacacaagttttcag GTT TCT GGA CCA AAG GAG GAG ATC GTC TAC TCC ATG GAA AAC GAG AAG CTT ATT GAG CAA TTC TAC CAA CAA  +1015
                        V   S   G   P   K   E   E   I   V   Y   S   M   E   N   E   K   L   I   E   Q   F   Y   Q   Q

GGA GAC AAG GCT GAG GTC AAC CCA TTC AAG GCT ATC GAG ATG GAG CAG AAG GTT GAG CAA CTT CAA GAA ATC TTC CGC CAA ATT CAA GAG  +1105
 G   D   K   A   E   V   N   P   F   K   A   I   E   M   E   Q   K   V   E   Q   L   Q   E   I   F   R   Q   I   Q   E

CAC GAG CAG AAC ACC CCA GAA ACT GTC CAC CTT ATT GCC CGC GCC GTC CGC ATG TTC CGC ATG TGC ACC ATC GAA GAA CTC AAG AAG GTT  +1195
 H   E   Q   N   T   P   E   T   V   H   L   I   A   R   A   V   R   M   F   R   M   C   T   I   E   E   L   K   K   V

CAC ACC ACC ATC TAC ACT AAG GCC GAG AAG AAG gtaaccaatacacatttcataatcactaatatcatgctctctcttag GTT CAA CTT GTC ATT GAA ACC  +1296
 H   T   T   I   Y   T   K   A   E   K   K                                                V   Q   L   V   I   E   T

AGC ATT GCT GTC GCT GGA ACC AAG AAC ACC ATT CAA CAC TTG ATT CAT CAC TTC GAG AAG AAG AGC ATC ACA CCA TTG AGA GCT GCT GAG  +1386
 S   I   A   V   A   G   T   K   N   T   I   Q   H   L   I   H   H   F   E   K   K   S   I   T   P   L   R   A   A   E

CTT CTT AAA TCT GTT CAA GAA ACT CTT TAC CCA AGT GAG CAC ATT GCT GAT CTT CTT ATC CAA CTC GCC CAA TCG CCA CTC TCT GAG AAG  +1476
 L   L   K   S   V   Q   E   T   L   Y   P   S   E   H   I   A   D   L   L   I   Q   L   A   Q   S   P   L   S   E   K

TAC GAA CCA CTC CGC CAA TCT GCC TGG CTC GCC GCA GGA TCT GTT GTT CGT GGA TTT GCT TCA AAG ACC CAA GAC CTT CCA TTG ATC CGC  +1566
 Y   E   P   L   R   Q   S   A   W   L   A   A   G   S   V   V   R   G   F   A   S   K   T   Q   D   L   P   L   I   R

CCA GCA TCT AGA CAA ACC AAG GAA AAG TAC GTT CGC GTC TTC ATG CAA CAC TTC CGT AAC GCT GAC TCC ACA TAC GAG AAG GTT CTT GCT  +1656
 P   A   S   R   Q   T   K   E   K   Y   V   R   V   F   M   Q   H   F   R   N   A   D   S   T   Y   E   K   V   L   A

CTT AAG ACC CTT GGA AAC GCC GGA ATC GAC CTC TCC GTC TAT GAG CTT GTC CAG ATT ATC CAA GAT CCA CGT CAA CCA CTT TCC ATC CGC  +1746
 L   K   T   L   G   N   A   G   I   D   L   S   V   Y   E   L   V   Q   I   I   Q   D   P   R   Q   P   L   S   I   R

ACT GAG GCT GTA GAT GCT CTT CGT CTT CTC AAG GAC GTT ATG CCA CGC AAG ATC CAG AAG GTT CTC CTT CCA GTC TAC AAG AAC CGC CAA  +1836
 T   E   A   V   D   A   L   R   L   L   K   D   V   M   P   R   K   I   Q   K   V   L   L   P   V   Y   K   N   R   Q
```

```
AAC AAG CCA GAG CTC CGT ATG GCT GCT CTT TGG AGA ATG ATG CAC ACC ATT CCA GAA GAA CCA GTT CTT GCT CAC ATC GTT TCC CAA ATG +1926
 N   K   P   E   L   R   M   A   A   L   W   R   M   M   H   T   I   P   E   E   P   V   L   A   H   I   V   S   Q   M

GAA AAC GAA TCC AAC CAA CAC GTT GCT GCC TTC ACC TAC CAC GTC CTC CGC CAA TTC TAC AAA TCC ACC AAC CCA TGC TAC CAA CAA TTG +2016
 E   N   E   S   N   Q   H   V   A   A   F   T   Y   H   V   L   R   Q   F   Y   K   S   T   N   P   C   Y   Q   Q   L

GCT GTT CGT TGC TCT AAG ATC CTT CTC TTC ACC CGT TAT CAA CCA CAA GAA CAG ATG CTC TCC ACC TAC TCC CAA CTT CCA CTT TTC AAC +2106
 A   V   R   C   S   K   I   L   L   F   T   R   Y   Q   P   Q   E   Q   M   L   S   T   Y   S   Q   L   P   L   F   N

TCT GAG TGG CTC TCC GGA GTT CAA TTC GAC TTC GCC ACC ATT TTC GAG AAG AAC GCT TTC TTG CCA AAG GAA GTT CAA GCA TCA TTG GAA +2196
 S   E   W   L   S   G   V   Q   F   D   F   A   T   I   F   E   K   N   A   F   L   P   K   E   V   Q   A   S   L   E

ACC GTC TTC GGA GGA AAC TGG AAC AAA TAC TTC GCT CAA GTT GGA TTC TCT CAA CAG AAC TTT GAG CAA GTC ATC CTC AAG ACC CTC GAA +2286
 T   V   F   G   G   N   W   N   K   Y   F   A   Q   V   G   F   S   Q   Q   N   F   E   Q   V   I   L   K   T   L   E

AAA CTT TCT CTT TAC GGA AAG CAA TCT GAT GAA CTC CGT TCC CGT CGT GTC CAA TCT GGA ATC CAA ATG CTT CAA GAG ATT GTC AAG AAG +2376
 K   L   S   L   Y   G   K   Q   S   D   E   L   R   S   R   R   V   Q   S   G   I   Q   M   L   Q   E   I   V   K   K

ATG AAC ATC CGT CCA CGT GTC CAA CAA ACC GAT TCT CAA AAT GCT CAC GCT GTT TTC TAC CTT CGC TAC AAG GAG ATG GAC TAC ATC GTT +2466
 M   N   I   R   P   R   V   Q   Q   T   D   S   Q   N   A   H   A   V   F   Y   L   R   Y   K   E   M   D   Y   I   V

CTT CCA ATT GAC ATG GAA ACT ATT GAC ACT CTT GTT GAG AAG TAT GTC AGA AAC GGA GAG TTT GAC ATC AAA TCC CTC CTC ACT TTC TTG +2556
 L   P   I   D   M   E   T   I   D   T   L   V   E   K   Y   V   R   N   G   E   F   D   I   K   S   L   L   T   F   L

ACC AAC GAC TCC AAG TTC GAG CTT CAC CGT GCT CTC TTC TTC TAC GAG GCT GAA CGC AGA ATT CCA ACA ACC ATT GGA ATG CCA CTC ACC +2646
 T   N   D   S   K   F   E   L   H   R   A   L   F   F   Y   E   A   E   R   R   I   P   T   T   I   G   M   P   L   T

ATT TCT GGA AAG ATG CCA ACT ATC CTC TCT ATC AAC GGA AAG GTT TCA ATT GAG CTC GAG AAG CTT GGA GCT CGT CTT GTT CTT GAT ATC +2736
 I   S   G   K   M   P   T   I   L   S   I   N   G   K   V   S   I   E   L   E   K   L   G   A   R   L   V   L   D   I

GTT CCA ACT GTT GCC ACC ACC CAC GTC ACT GAG ATG CCG CTT CTG TAT CCA GTC ATT GAA CAA GGA GTC AAG TCA CTT CAA TCT GCT CGT +2826
 V   P   T   V   A   T   T   H   V   T   E   M   P   L   L   Y   P   V   I   E   Q   G   V   K   S   L   Q   S   A   R

CTC CAC ACT CCA TTG AGA TTC GAA TCA ACT GTT GAA TTG AAG AAG AAC ACT CTC GAA ATC ACT CAC AAG TTT GTT GTC CCA GAG AAC AAG +2916
 L   H   T   P   L   R   F   E   S   T   V   E   L   K   K   N   T   L   E   I   T   H   K   F   V   V   P   E   N   K

AAG ACC ACT GTT TCC GTT CAT ACC CGC CCA GTT GCT TTC ATC CGT GTT CCA AAG AAC CAA GAC TCT GAA TAT GTT GAG GCT GAA GAG AAG +3006
 K   T   T   V   S   V   H   T   R   P   V   A   F   I   R   V   P   K   N   Q   D   S   E   Y   V   E   A   E   E   K

ACT ATT TCC CAC TCA CAA TAC CAA ATG TCT ACT GAA GAG ATT GAT CGT CAA TAT GAG ACC TTT GGA CTC AGA ATC AAT GCC CAA GGA AAT +3096
 T   I   S   H   S   Q   Y   Q   M   S   T   E   E   I   D   R   Q   Y   E   T   F   G   L   R   I   N   A   Q   G   N

GTT CTT TCC CAA TGG ACT CTT CCA ATG GTT TTG ATG ACT GAA CAA GAT TTC GAG TAC ACT CTT GAA AAC AAA AAC CGT CCA GTT GAG TTC +3186
 V   L   S   Q   W   T   L   P   M   V   L   M   T   E   Q   D   F   E   Y   T   L   E   N   K   N   R   P   V   E   F

ACA GCT CGC GTC ACT ATT GGA AAC CTC GAG AAG ACT GAT CTT TCC GAG ATC AAG TTC GAC AAG ATC TTC GAA AAA GAA TTC GAC CTT GAG +3276
 T   A   R   V   T   I   G   N   L   E   K   T   D   L   S   E   I   K   F   D   K   I   F   E   K   E   F   D   L   E

AAC AAC GAA TCT GAG AAC CGC CGC CAA TAC TTC CAC AAG ATG ATC CGT GAG ATT CAA TCT GAG CAA GGA TTC AAG AAC CTC ATC ACC CTC +3366
 N   N   E   S   E   N   R   R   Q   Y   F   H   K   M   I   R   E   I   Q   S   E   Q   G   F   K   N   L   I   T   L

AAG CTT GAA GCC CCA CAA CAA ATG TAC TGG AAC ACT GAA CTT CGT ACC GTC TGT GAC AAA TGG ATC CGT ATG TGC AAG GTT GAG ATG GAT +3456
 K   L   E   A   P   Q   Q   M   Y   W   N   T   E   L   R   T   V   C   D   K   W   I   R   M   C   K   V   E   M   D

GCT CGC CGC TCT CCA ATG GAG CAC GAG AAC AAA GAA TGG ACT CTT CGT ACT GAG CTT CTT GCT GCC CGC CCA CAA ATG CCA TCA TCC CTC +3546
 A   R   R   S   P   M   E   H   E   N   K   E   W   T   L   R   T   E   L   L   A   A   R   P   Q   M   P   S   S   L

CGT CAA CTT CGT GAG CAA CCA CAC CGT GAG GTT CAA CTC GCA TTC AAT GCC AAG TGG GGA TCA TCA AAG AAG AGC GAG ATC ACA GTC AAT +3636
 R   Q   L   R   E   Q   P   H   R   E   V   Q   L   A   F   N   A   K   W   G   S   S   K   K   S   E   I   T   V   N

GCT CAA CTC GAA CAA TCC ACG GAA CAA AAG AAG TTC ATC CGC AAC ATC GAG CGT GAG TAC AAG GGA ATT CCA GAG TAC GAA CTT TTG ATC +3726
 A   Q   L   E   Q   S   T   E   Q   K   K   F   I   R   N   I   E   R   E   Y   K   G   I   P   E   Y   E   L   L   I
```

```
AAG GCT GCT CGT CTT AAC CAA GTC AAT GTT GTC TCT GAG TAC AAG CTC ACC CCA CAG TCT GAA TAC ACT TTC TCC CGC ATT TTC GAC CTT +3816
 K  A  A  R  L  N  Q  V  N  V  V  S  E  Y  K  L  T  P  Q  S  E  Y  T  F  S  R  I  F  D  L

ATC AAG GCA TAC AAC TTC TGG ACT GTT TCT GAG AAG CGT GTC CAA AAC GAG AAT CGC CGC GTT GTT CTT CAA CTT TCT GTT GAG CCA CTT +3906
 I  K  A  Y  N  F  W  T  V  S  E  K  R  V  Q  N  E  N  R  R  V  V  L  Q  L  S  V  E  P  L

TCC CGC CAA TCA CAT GAA CAT GAC CAT CAG ACT CCA GAA CAA GAA GTT GAG TTG AAG AAT GCT CGT ATT CCA CGA GTC GTT CTC CCA ACT +3996
 S  R  Q  S  H  E  H  D  H  Q  T  P  E  Q  E  V  E  L  K  N  A  R  I  P  R  V  V  L  P  T

ATT GCT CGT AGT GCC ATG TTC CAA CAA ACC TGG GAA AAG ACC GGA GCC ACC TGC AAG GTT GAC CAA TCT GAG GTT TCT ACC TTT CAC AAC +4086
 I  A  R  S  A  M  F  Q  Q  T  W  E  K  T  G  A  T  C  K  V  D  Q  S  E  V  S  T  T  H  N

GTG ATC TAC CGC GCT CCA CTC ACC ACC TGC TAC TCT CTT GTT GCC AAG GAT TGC TCT GAA CAG CCA AGA TTC GCT GTT CTT GCC AAG AAG +4176
 V  I  Y  R  A  P  L  T  T  C  Y  S  L  V  A  K  D  C  S  E  Q  P  R  F  A  V  L  A  K  K

ATC AAC AAG AAC TCT GAG GAG CTT CTC GTT AAG GTT GTC CGC CGT GAG GAA GAA ATT GTT GTG AAG AAG TCT GAC GAT AAG TTC CTT GTC +4266
 I  N  K  N  S  E  E  L  L  V  K  V  V  R  R  E  E  E  I  V  V  K  K  S  D  D  K  F  L  V

AAG GTT GAC GGA AAG AAG GTT AAC CCA ACT GAA CTT GAA CAA TAC AA  gtaagctttaactctaaatcacatgaatttttctaatctcgtttttcag T ATC +4366
 K  V  D  G  K  K  V  N  P  T  E  L  E  Q  Y  N                                                                  I

GAA ATT CTT GGA GAT AAC CTT ATT GTT ATT CGT CTT CCA CAA GGA GAG GTT CGT TTC GAT GGA TAC ACT GTC AAG ACC AAC ATG CCA TCC +4456
 E  I  L  G  D  N  L  I  V  I  R  L  P  Q  G  E  V  R  F  D  G  Y  T  V  K  T  N  M  P  S

GTT GCT TCA CAA AAC CAA CTT TGC GGA CTT TGC GGA AAC AAT GAC GGT GAG AGA GAC AAT GAG TTC ATG ACC GCT GAC AAC TAC GAA ACT +4546
 V  A  S  Q  N  Q  L  C  G  L  C  G  N  N  D  G  E  R  D  N  E  F  M  T  A  D  N  Y  E  T

GAG GAT GTT GAG GAA TTC CAC CGG TCT TAC CTT CTC AAG AAT GAG GAA TGC GAG TTT GAG TTC GAC CGC ATC TCC GAG AAG AAG AAC TAC +4636
 E  D  V  E  E  F  H  R  S  Y  L  L  K  N  E  E  C  E  F  E  F  D  R  I  S  E  K  K  N  Y

AGA AAC AAA TGG AAC AGA GAA GAG AAG AAG TCC GAC TAC GAG AGC AGC TCC GAC TAC GAG AGC AAC TAC GAT GAG AAG GAA ACT GAA GAG +4726
 R  N  K  W  N  R  E  E  K  K  S  D  Y  E  S  S  S  D  Y  E  S  N  Y  D  E  K  E  T  E  E

G  gttagtctaaggctgaatgatgtgtagttttaacaaaatacatatttttcag  AA CTC GTC AAG AAG ACC CTC ATC AAG GAG TTC TCC AAC CGC GTC TGC +4826
                                                          E  L  V  K  K  T  L  I  K  E  F  S  N  R  V  C

TTC TCC ATC GAG CCA GTC TCT GAG TGC CGC CGT GGA CTC GAA TCC GAG AAG ACT TCC AAC AAG AAG ATC CGT TTC ACT TGC ATG CCA CGT +4916
 F  S  I  E  P  V  S  E  C  R  R  G  L  E  S  E  K  T  S  N  K  K  I  R  F  T  C  M  P  R

CAC AGG CAA GAA CGT AGT CGT TTT CTT CAA GGA AGC TCT GAG CAA ACT GTT GCC GAG TTG GTC GAT TTC CCA GTC TCC TTC GTT GAG TCT +5006
 H  R  Q  E  R  S  R  F  L  Q  G  S  S  E  Q  T  V  A  E  L  V  D  F  P  V  S  F  V  E  S

GTC AAG ATC CCA ACC GCC TGC GTT GCC TAT TAG ATTTCATATGTTTTATTAATTTTCTATTAAATAAAGCATTTTCACTAGaatgatttttttttctcacacttctaga +5114
 V  K  I  P  T  A  C  V  A  Y                    ======
```

Figure 1. The nucleotide sequence of vit-5 and predicted amino acid sequence of yp170a. The sequence was determined by the method of Sanger (20). Restriction fragments of genomic clone 2017 containing all of vit-5 (5) were subcloned into pUC8 from which random Sau3a fragments were cloned into mp8 and mp9 for sequencing. Sequences were aligned by comparison of restriction sites found by sequencing, with restriction maps of the sub-clones. Alignments were confirmed by sequencing predicted fragments which overlapped the boundaries of adjacent clones. Both strands were sequenced over most of the length of the gene. Portions of cDNA clone 1728 (5) were sequenced to confirm the predicted intron 3 and 4 boundaries. Numbering is from the first base in the mRNA, determined by premer extersion accompained by dideoxy sequencing. Introns and 5' and 3' untranslated region is double-underlined. Rare codons (as defined in the test) are underlined.

Table 1
Intron/Exon Border Sequences

Intron 1	AAG|GTATTC.....TTTTCAG|GTT
2	AAG|GTAACC.....CTCTTAG|GTT
3	CAA|GTAAGC.....TTTTCAG|TAT
4	AGG|GTTAGT.....TTTTCAG|AAC
C. elegans Consensus*	AG|GTAAG.....TTTTCAG|A/G

*The C. elegans consensus is taken from a compilation of 29 sequenced C. elegans intron borders (Blumenthal, unpublished).

acid sequence at the N-terminus of the proposed protein, which is probably the signal sequence responsible for cotranslational secretion. 2) The sequences of this region of the other vitellogenin genes show that all of the proteins would start with a very similar signal sequence if the first AUG in each mRNA is used for translation initiation (7). 3) This signal sequence would be homologous to signal sequences at the N-termini of vertebrate vitellogenins (9). 4) The AUG at the proposed site is preceded by G at position -3 and followed by A at position +4, and so should be favored for initiation (10). 5) The proposed site for initiation precedes the only long, open reading frame, one with a very asymmetric codon usage typical of abundantly-expressed C. elegans genes (11-14).

The 5' untranslated region of vit-5 is exceptionally short, only nine bases long. The other members of the vit-1-vit-5 group also have unusually short 5' untranslated regions, 9-11 bases long (7). Translation terminates at a UAG codon which is followed by an AAUAAA sequence 30 bases further on and the 3' end of the mRNA 18 bases later. Hence, of the 4869 bases in the vit-5 mRNA, only 60 are untranslated.

The vit-5 introns range in length from 47 to 70 base pairs. Only 218 base pairs of the total gene length of 5087 base pairs are in introns. Thus vit-5 is a very compact gene. Very little RNA is spliced out and almost all that remains is translated. This gene structure is typical of the C. elegans genes sequenced thus far (11-14). In contrast, the members of the vitellogenin gene family of vertebrates are about 25 kb in length and contain 33 introns (15, 16).

Codon Usage

The codon usage is very asymmetric (Table 2). In the 1603 codons of

Table 2
Codon Usage in vit-5

Codon	Amino Acid	Residues	Codon	Amino Acid	Residues	Codon	Amino Acid	Residues	Codon	Amino Acid	Residues
UUU	F	11	UCU	S	51	UAU	Y	8	UGU	C	1
UUC	F	59	UCC	S	38	UAC	Y	45	UGC	C	19
			UCA	S	13						
UUA	L	1	UCG	S	3	CAU	H	8	UGG	W	14
UUG	L	18	AGU	S	4	CAC	H	25			
CUU	L	76	AGC	S	9				CGU	R	45
CUC	L	44				CAA	Q	90	CGC	R	34
CUA	L	0	CCU	P	0	CAG	Q	10	CGA	R	3
CUG	L	2	CCC	P	0				CGG	R	1
			CCA	P	61	AAU	N	17	AGA	R	14
AUU	I	40	CCG	P	2	AAC	N	56	AGG	R	1
AUC	I	51									
AUA	I	1	ACU	T	43	AAA	K	16	GGU	G	2
			ACC	T	49	AAG	K	108	GGC	G	0
AUG	M	35	ACA	T	5				GGA	G	41
			ACG	T	1	GAU	D	27	GGG	G	1
GUU	V	68				GAC	D	31			
GUC	V	46	GCU	A	46						
GUA	V	1	GCC	A	28	GAA	E	66			
GUG	V	2	GCA	A	5	GAG	E	106			
			GCG	A	1						

vit-5, 19 codons are used three times or less, including four that are not used at all (CUA, CCU, CCC, GGC) and nine others that are used only once (UUA, AUA, GUA, ACG, GCG, UGU, CGG, AGG, GGG). In general, third position purines are rare. For instance, 114 out of 117 valine codons have third position pyrimidines. However, there are exceptions to this rule: 61 out of 63 proline codons are CCA; and 41 out of 44 glycine codons are GGA. This degree of asymmetry has been observed previously only in single-celled organisms (17, 18). An examination of published sequences of other abundantly expressed C. elegans genes, including actin (14), myosin (12), collagen (13), and the major sperm protein (11), reveals the same highly asymmetric codon usage.

Interestingly, the 19 rare codons occur at a much higher frequency near the beginning of the gene. While only 23 of 1603 total codons, or 1.4%, are rare codons, four of the first 18 codons, or 22 percent are rare codons, and two of the four occur nowhere else in the protein. We have also sequenced portions of the other vitellogenin genes (unpublished observations) and observed the same phenomenon: the same codons are rare, and each gene has several rare codons within the first 18 (a total of 16 rare codons in this region of the 5 vitellogenin genes for which sequence data is available). We do not know why rare codons are more acceptable near the 5' end of the gene. Assuming that rare codons are decoded by rare tRNAs and that selection is for rapid translation of abundantly expressed genes as has been shown to be the case in other systems (17, 18), we hypothesize there is no strong selection

Table 3
Amino Acid Composition of Vitellogenins

	yp170A	Xenopus	Chick
Asp + Asn	131	148	154
Thr	98	87	79
Ser	118	171	236
Glu + Gln	272	229	179
Pro	63	84	79
Gly	44	84	84
Ala	80	135	126
Val	117	102	102
Met	35	42	39
Ile	92	81	89
Leu	141	141	138
Tyr	53	50	48
Phe	70	62	43
His	33	60	55
Lys	124	113	138
Arg	98	87	118
Cys	20	10	ND
Trp	14	ND	ND

The amino acid composition of the nematode vitellogenin yp170a is inferred from the DNA sequence of vit-5 (Fig. 1). The amino acid compositions of Xenopus and chicken vitellogenins were determined by amino acid analysis and are taken from Wiley and Wallace (21) and Wang et al. (22), respectively.

for rapid translation of the region of the mRNA encoding the signal sequence. Indeed the ubiquity of rare codons in this region of the five genes sequenced suggests there may be selection for slow translation of the region. Perhaps slow translation allows interaction between the nascent polypeptide and some component required for secretion.

Characteristics of yp170A

Vit-5 encodes a polypeptide of 186,430 daltons. The predicted amino acid composition is shown in Table 3. The protein is very high in glutamic acid and lysine, as expected for a vitellogenin. Table 3 also presents the amino acid compositions of vitellogenins from Xenopus and from chicken (vitellogenin II) for comparison. On the whole the predicted amino acid composition of yp170A is quite similar to the vertebrate compositions, but there are some notable differences. yp170A has more glutamic acid + glutamine and less glycine, alanine and histidine. Most interestingly, yp170A has much less serine than do the vertebrate vitellogenins. In chicken vitellogenin, 123 of 236 serine residues are contained in the phosvitin region of the precursor (19). We have searched the yp170A sequence for a phosvitin-like region and failed to find it. (In the "core" region of chickin phosvitin, 80 of 99 residues are serine). Thus it appears that phosvitin is missing from yp170A, even though at the N-terminus the protein is clearly related to the chicken vitellogenin (7), and overall, the amino acid compositions of the two proteins are quite similar (Table 3). Since almost all of the phosvitin sequence of chicken is contained in a single exon (19), it seems likely that this exon is missing from the C. elegans gene.

ACKNOWLEDGEMENTS

This work was supported by grant GM 30870 from the National Institute of General Medical Sciences. We are grateful to S. Strome for critical reading of the manuscript.

REFERENCES

1. Klass, M.R., Wolf, N., and Hirsh, D. (1979) Develop. Biol. 69, 329-335.
2. Sharrock, W.J. (1983) Develop. Biol. 96, 182-188.
3. Sharrock, W.J. (1984) J. Mol. Biol. 174, 419-431.
4. Kimble, J. and Sharrock, W.F. (1983) Develop. Biol. 96, 189-196.
5. Blumenthal, T., Squire, M., Kirtland, S., Cane, J., Donegan, J., Spieth, J., and Sharrock, W.J. (1984) J. Mol. Biol. 174, 1-18.
6. Spieth, J. and Blumenthal, T. (1985) Mol. Cell. Biol. In press.
7. Spieth, J., Denison, K., Kirtland, S., Cane, J. and Blumenthal, T. (1985) Nucleic Acids Res., 13, 5283-5295.
8. Tata, J. R. (1976) Cell 9, 1-14.
9. Walker, P., Brown-Luedi, M., Germond, J-E., Wahli, W., Meijlink, F., Vanhet, P.W., Schip, A.D., Roelink, H., Gruber, M., and Ab, G. (1983) EMBO J. 2, 2271-2279.
10. Kozak, M. (1984) Nucleic Acids Res. 12, 857-873.
11. Klass, M.R., Kinsley, S., and Lopez, L.C. (1984) Mol. Cell. Biol. 4, 529-537.
12. Karn, J., Brenner, S., and Barnett, L. (1983) Proc. Natl. Acad. Sci. USA 80, 4253-4257.
13. Kramer, J.M., Cox, G.N., and Hirsh, D. (1982) Cell 30, 599-606.

14. Files, J.G., Carr, S., and Hirsh, D. (1983) J. Mol. Biol. 164, 355-375.
15. Wahli, W., Dawid, I.B., Wyler, T., Weber, R., and Ryffel, G.U. (1980) Cell 20, 107-117.
16. Arnberg, A.C., Meijlink, F.C.P.W., Mulder, J., Van Bruggen, E.F.J., Gruber, M., and AB, G. (1981) Nucleic Acids Res. 9, 3271-3286.
17. Grosjean, H. and Fiers, W. (1982) Gene 18, 199-209.
18. Bennetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. 257, 3026-3031.
19. Byrne, B.M., Van het Schip, A.D., Van de Klundert, J.A.M., Arnberg, A.C., Gruber, M., and AB, G. (1984) Biochemistry 23, 4275-4279.
20. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) Proc Natl. Acad. Sci. USA 74, 5463-5467.
21. Wiley, H.S. and Wallace, R.A. (1981) J. Biol. Chem. 256, 8626-8634.
22. Wang, S-Y., Smith, D.E., and Williams, D.L. (1983) Biochemistry 22, 6206-6212.