Sequence analysis, and chromosomal localization of a gene encoding a cystatin-like protein from *Drosophila melanogaster*

Margaret L. Delbridge and Leonard E. Kelly

Department of Genetics, University of Melbourne, Parkville, Victoria 3052, Australia

Received 5 September 1990

Using polyclonal antibodies raised against a *Drosophila* Ca^{2+}-binding protein (DCABP-23), clones were isolated from a *Drosophila* head cDNA library constructed in the expression vector λgt11. Two non-homologous clones have been isolated and are being subjected to sequence analysis. One of these clones, though not encoding DCABP-23, does encode a *Drosophila* cystatin-like protein. This presumed *Drosophila* cystatin shows homology to mammalian cystatins, chicken egg white cystatin and the rice oryzacystatin. The *Drosophila* cystatin has been mapped, by in situ hybridization, to region 88C on the right arm of the third chromosome.

Protease inhibitor; Cystatin; DNA sequence; *Drosophila*

1. INTRODUCTION

Cystatin was first identified as a component of egg white [1] and shown to be an inhibitor of cysteine proteinases. Equivalent inhibitors have also been described in a number of tissues of both animal and plant origin [2–8], and have been divided into 3 families. These consist of the stefins (family 1), cystatins (family 2) and the blood plasma kininogens (family 3) [9,10]. Members of the cystatin family bind tightly to papain-like cysteine proteinases [11], and cysteine proteinase activity is often found to co-purify with cystatin. Both family 1 and family 2 proteins are non-glycosylated, often heat-stable, small molecular mass proteins (11–13 kDa), whereas the kininogens are glycosylated high molecular mass inhibitors. Structurally, the cystatin of plant origin, oryzacystatin, appears to be intermediate between families 1 and 2 [12].

We have identified and purified a major heat-stable Ca^{2+}-binding protein from *Drosophila* (DCABP-23) [13]. The purified Ca^{2+}-binding protein has a molecular mass of 23 kDa, but has low amounts of a smaller molecular mass protein that co-purifies with the major component [13]. More recently it has been shown that, co-purifying with DCABP-23, is a cryptic cysteine proteinase activity (Kelly and Bell, unpublished results). Antibodies raised against DCABP-23 have been used to screen a *Drosophila* head cDNA expression library. Two non-homologous cDNA clones have been isolated, one of which encodes a cystatin-like protein. It is this clone that is the subject of this report.

Correspondence address: L.E. Kelly, Department of Genetics, University of Melbourne, Parkville, Victoria 3052, Australia

2. MATERIALS AND METHODS

2.1. *cDNA library screening*

Affinity-purified antibodies against DCABP-23 were prepared as previously described [13]. A *Drosophila* head cDNA library in λgt11 [14], was plated at a density of 2×10^5 plaques/150 mm plate, using the bacterial strain RY1090, and probed with affinity-purified antibody using the method described by Young and Davis [15].

For the isolation of the longer cDNA clones, the E.1 insert was excised, nick translated and used to probe Benton lifts of the same library plated at similar densities.

2.2. *Isolation of genomic clones*

The genomic library used was obtained from Dr J. Tamkun (Dept of MCO Biology, University of Colorado), made from a *Drosophila* strain isogenic for all 4 chromosomes. The E.1 insert was used to probe filters from this genomic library.

2.3. *Molecular studies of the cDNA and genomic clones*

Sequence analysis of the clones was carried out using the dideoxy chain termination method described by Sanger et al. [16].

Northern blot analysis was carried out using RNA prepared from whole addult flies run on 1.5% agarose/formaldehyde gels and blotted to nitrocellulose according to Maniatis et al. [17].

DNA probes to be used for in situ hybridization to *Drosophila* polytene chromosomes were labelled using biotinylated dUTP (BRL) in a nick translation reaction, and hybridisation was localized using alkaline phosphatase conjugated to streptavidin [18].

3. RESULTS AND DISCUSSION

3.1. *DNA cloning and sequencing*

The antibodies raised against DCABP-23 were used to screen the *Drosophila* λgt11 head cDNA library. Two non cross-hybridizing antigen-positive clones were isolated and analysed by restriction mapping. One of these λE.1 was found to have a 450 bp insert. This insert was subcloned into pUC19 and subjected to further restriction analysis. The complete restriction analysis is

Published by Elsevier Science Publishers B. V. (Biomedical Division)

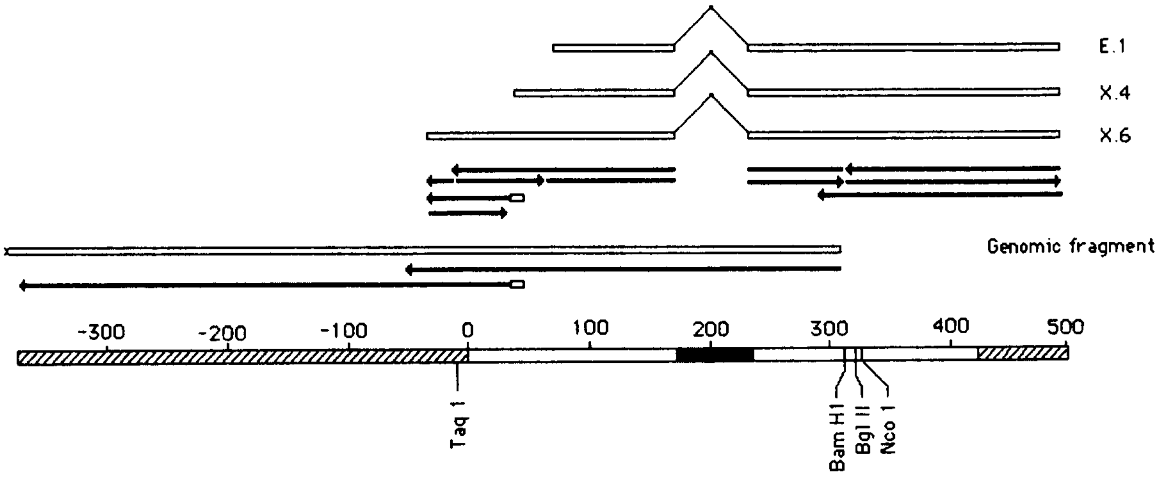

Fig. 1. Restriction map and strategy used to sequence the cDNA and genomic clones encoding the cystatin-like protein. E.1, X.4 and X.6 represent cDNA clones. The arrows show the direction and length of sequence obtained using either restriction fragments and M13 primers, or a synthetic oligonucleotide primer (□).

shown in Fig. 1. As clones isolated by screening expression vectors with antibodies are, by definition, incomplete, the 450 bp insert was used to probe a Northern blot of whole adult mRNA. The analysis indicated that the E.1 clone hybridizes to a single mRNA species of ~800 bases (Fig. 2). The E.1 clone was, therefore, used to reprobe the λgt11 library, and two further clones of ~750 bp were isolated. These clones, λX.4 and λX.6, both possessed the *Bam*HI, *Bgl*II and *Nco*I restriction sites found in E.1. All 3 clones were sequenced as indicated in Fig. 1. It became clear during the course of the sequence analysis that clones X.4 and X.6, while overlapping the entire E.1 clone, differed in their 5′ sequence, and further Northern blot analysis using the 5′ regions of X.4 and X.6 indicated that both of these clones were chimeric cDNAs (data not shown). In order to confirm that we had the complete coding sequence of this gene, a 1.95 kb *Sal*/*Bam*HI genomic fragment that is homologous to the 5′ region of the E.1 clone was also partially sequenced. The sequence shown in Fig. 3, therefore, is derived from the complete E.1 sequence, the X.4 and X.6 sequences up to the point where the latter two diverge, and part of the genomic clone. The reading frame of the cDNA can be defined by the E.1 clone, as it must have been in frame with the β-galactosidase gene of λgt11 in order for the protein to be translated and act as an antigen in the selection. Interestingly there is evidence of polymorphism between the E.1 and the X.4/X.6 and genomic sequences. Although 5 of the 6 polymorphisms are in the third base of the codons and hence do not change the amino acid sequence, one of the base-pair changes results in an amino acid substitution of *asparagine* for *aspartic acid* at position 36.

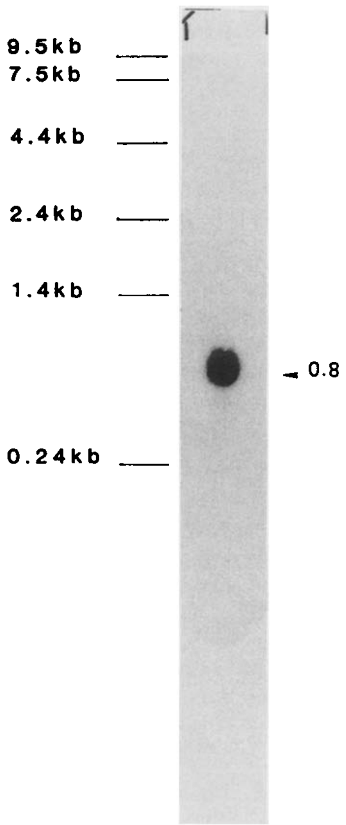

Fig. 2. Northern blot of whole RNA from adult flies probed with the E.1 clone. A single band at ~800 bases is observed.

The initiation codon was decided upon by virtue of the CAAC sequence immediately preceeding the AUG. This sequence conforms to the consensus sequence of translation initiating signals for *Drosophila* genes [19]. However there is another potential, in phase, initiation codon starting at position −69, and the use of this codon to initiate translation in vivo cannot be ruled out. If this sequence is translated, it produces an *arg arg ala ser* sequence which is a potential cAMP-dependent pro-

```
   C GCA CCT TGA ATA TTG ACT ATA AAT ATT GAA TAT TGA CTA TAA ATA TTT GGA TAG TAG
   -358                 1                        1

 CTG CGA AAA TAT CAG GAC ACG TAG ATA CCG TAG TTT ATG GTT CTA AAA AAA AAT ATT TCG
-300

 AAT CTT AAA ACT ATT TTT TCA CTT ATT TTG CCT TTC TAA GCA ATT TTC TTA AAT ATT GTA
-240

 AAC CCC GAG TGA GGT TCT GGT TAT TGA TAG TTC TGA TGA TGG GCC AGT CAA GTC TTG GAT
-180
                                                                    (met ala gly
 CGA TGG CAT AGA GAG CCT ATT GTA TTT TTC CAT TGT TGA TCT CCC ACT GAG ATG GCA GGT
-120
 ile lys trp ala thr ser pro gly val his gln leu ser gln ser arg arg ala ser asn)
 ATA AAA TGG GCC ACA TCG CCG GGT GTT CAT CAG TTC TCT CAG AGT CGT CGA GCG AGC AAC
-60                                                                           2


 met asn val val lys ser leu cys ile leu gly leu val leu val ser leu ile ala thr
 ATG AAC GTA GTA AAA TCT TTG TGT ATT TTG GGT CTG GTT CTC GTC AGC TTG ATT GCC ACC
 1             G
                                                                 asp
 gln ala ala asp glu gln val val gly gly val ser gln leu glu gly asn ser arg lys
 CAA GCA GCC GAT GAG CAG GTG GTA GGC GGT GTC AGC CAG TTG GAG GGA AAC AGC AGG AAG
 61                            T                                 G

 glu ala leu glu leu leu asp ala thr leu ala gln léu ala thr gly asp gly pro ser
 GAG GCT CTG GAA CTT CTG GAT GCC ACT CTC GCA CAG TTG GCC ACC GGA GAT GGT CCC AGC
 121                                                       T

 val lys
 TAC AA  GTG AGT TAG AAC TGA GTT CCG AAT TTA AAA TTC ATA CTA ATT GAT CTG CTG ATC
 181

         ala ile asn val thr ser val thr gly gln val val ala gly ser leu asn thr
 TTA G G GCA ATC AAT GTG ACC TCT GTG ACG GGT CAG GTC GTA GCT GGA AGT CTT AAC ACC
 240                                                               C

 tyr glu val glu leu asp asn gly ser asp lys lys gln cys thr val lys ile trp thr
 TAC GAG GTG GAA CTT GAC AAT GGA TCC GAC AAA AAG CAG TGC ACC GTG AAG ATC TGG ACT
 299               G

 gln pro tro leu lys glu asn gly thr asn ile lys ile lys cys ser gly asp asp gly
 CAG CCA TGG CTC AAG GAG AAC GGC ACC AAC ATC AAG ATC AAG TGC TCT GGT GAC GAT GGC
 359

 glu leu asp arg thr trp stop
 GAA CTG GAC CGA ACC TGG TAG AAG ATT CTT CGT GAG AAT TGC CCT GAA AGA AAT AAT AAA
 419                                                                        3

 AAA AAA AAC AAG ATA CAC GTG ATT T (AAAAAAAAAAA.....AAAAAA)
 479                             503
```

Fig. 3. Sequence derived from the cDNA and genomic clones as outlined in Fig. 1. The sequence from the genomic clone indicated the presence of a 58 bp intron from bases 186 to 243 with the consensus GT & AT termini. Sequence blocks labelled '1' show the position of the 20 bp direct repeat each of which contains a TATA transcription initiation sequence. The block labelled '2' is the consensus sequence found to precede initiation codons in *Drosophila* genes [19], and the sequence labelled '3' is a polyadenylation signal. The bases typed in bold represent polymorphisms observed between the E.1 clone and the X.4, X.6 and genomic clones. The amino acid sequence in brackets is that produced where the ATG at position −69 used to initiate translation, and the 3′ terminal adenines in brackets represent the polyadenylated tail found in all three cDNA clones.

tein kinase phosphorylation site. In the 5′ untranslated region of the genomic sequence, there is a 20 bp direct repeat. Each of these repeats contains a TATA box consensus. The significance, if any, of this repeated TATA sequence is unknown. From the sequence of the genomic clones, a 58 bp intron was identified extending from bases +187 to +244 inclusive (Fig. 3). In the 3′ untranslated sequence there is a polyadenylation signal consensus sequence, and the E.1 clone possessed a polyA sequence extending 44 base pairs at the 3′ end.

3.2. *Amino acid sequence*

The cDNA derived amino acid sequence gives a protein of 126 amino acids with a molecular mass of 13 429 Da. This amino acid sequence was compared with those in a number of data bases, resulting in the detection of homology between this protein and members of the cystatin superfamily of proteins (Fig. 4). The maximum identity between the *Drosophila* protein and individual members of the family is 22%, exhibited between the *Drosophila* protein and oryzacystatin [12]. In the *Drosophila* protein there is strong conservation of the presumed active-site amino acids, including G(32), the highly conserved Q V V A G sequence beginning at position 77, and the P(127), W(128) amino acids. Each of these 3 regions appear, from X-ray crystallographic analysis of the chicken cystatin, to form a 'wedge' that can interact with the active site of papain [20]. The

```
                 *      * **           *  *****        *  * **
a)         MNVVKSL  CILGLVLVSL  IATQAADEQV  VGGVSQLEGD  SRKEALELLD
                                                     N
b)                                MSSDGGPV  LGGVEPVGNE  NDLHLVDLAR

c)                                      MI  PGGLSEAK-P  ATPEIQEIVD
d)                                      MM  CGAPSATQ-P  ATAETQHIAD
e)                                      MM  CGAPSATM-P  ATTETQEIAD

f)                              SSPGKPPRL  VGGPMDAS-V  EEEGVRRALD
g)                                      RL  LGGLMEAD-V  NEEGVQEALS
h)      MAGARGCVVL  LAAALMLVGA  VLGSEDRSRL  LGAPVPVD-E  NDEGLQRALQ
                10          20          30          40          50


           ***          *   *     **  ****  * *    ***    *  **
a)      ATLAQLATGD  GPSYK--AIN  VTSVTGQVVA  GSLNTYEVEL  DNGSDKKQ--

b)      FAVTEHNKKA  NSLLE--FEK  LVSVKQQVVA  GTLYYFTIEV  KEGDAKKL--

c)      KVKPQLEEKT  NETYG--KLE  AVQYKTQVVA  GTNYYIKVRA  GDNKYMHLK-
d)      QVRSQLEEKY  NKKFP--VFK  AVSFKSQVVA  GTNYFIKVHV  GDEDFVHLR-
e)      KVKSQLEEKA  NQKFD--VFK  AIDFRRQVVA  GTNFFIKVDV  GEEKCVHLR-

f)      FAVGEYNKAS  NDMYHSRALQ  VVRARKQIVA  GVNYFLDVEL  GRTTCTKTQP
g)      FAVSEFNKRS  NDAYQSRVVR  VVRARKQVVS  GMNYFLDVEL  GRTTCTKSQA
h)      FAMAEYNRAS  NDKYSSRVVR  VISAKRQLVS  GIKYILQVEI  GRTTCPKSSG
                60          70          80          90         100


                                    **   ***  ****        * **
a)      ----------  --------CT  VKIWTQPWLK  ENGTNIKIKC  SGDDGELDRT W

b)      ----------  --------YE  AKVWEKPWMD  FKELQ-EFKP  VDASANA

c)      -----VFKSL  PGQNE--DLV  LTGYQVDKNK  DDELT-GF
d)      -----VFQSL  PHENK--PLT  LSNYQTNKAK  HDELT-YF
e)      -----VFEPL  PHENK--PLT  LSSYQTDKEK  HDELT-YF

f)      NLDNCPFHDQ  PHLKRKAFCS  FQIYAVPSQG  TMTLS-KSTC  QDA
g)      NLDSCPFHNQ  PHLKREKLCS  FQVYVVPWMN  TINLV-SFSC  QD
h)      DLQSCEFHDE  PEMAKYTTCT  FVVYSIPWLN  QIKLL-ESKC  Q
               110         120         130         140         150
```

Fig. 4. Comparison of the amino acid sequence of the *Drosophila* cystatin-like protein with other members of the cystatin superfamily. The asterisks above the *Drosophila* sequence indicate identity of amino acids with at least one of the members of the family. The regions underlined indicate those amino acids which are included in a loop formed by intramolecular cysteine bridges in the cystatins and are presumed to be present in the *Drosophila* protein. The sequences shown are: (a) *Drosophila* cystatin (this study), (b) Oryzacystatin (rice) [12], (c) Human cystatin A [2], (d) Human cystatin B [6], (e) Rat cystatin [3], (f) Human cystatin C [4], (g) Bovine cystatin [7], (h) Chicken cystatin β [21].

Drosophila cystatin shows greater resemblance to the family 2 cystatins, especially in the bracketing of amino acids 120–139 by two cysteine residues. These two cysteines form a disulphide bridge in the chicken cystatin, and we presume that such a disulphide bond structure is also present in *Drosophila* cystatin. The amino terminal region of the protein also bears some resemblance to the chicken cystatin in that there is a strongly hydrophobic amino terminal region. In the chicken cystatin it has been proposed that this is a signal peptide necessary for secretion of the protein [21].

3.3. *Chromosomal localization*

Genomic clones which hybridize to the 450 bp insert of clone E.1 have been isolated and hybridized in situ to *Drosophila* polytene chromosomes. The probe hybridized to a single site at position 88C on the right arm of chromosome 3 (Fig. 5).

This study has identified a *Drosophila* gene that encodes a protein that is structurally similar to members of the cystatin superfamily, although it has yet to be proved that this protein does possess cystatin activity. Attempts are presently being made to express this *Drosophila* gene in *E. coli* thus allowing the confirmation that the gene does encode an active cystatin.

The cystatins have been well characterized at the biochemical level, however the in vivo function of these cysteine proteinase inhibitors has yet to be determined. The identification of the gene for this cystatin-like protein in a well defined genetic organism, such as *Drosophila melanogaster*, should facilitate genetic and molecular studies aimed at determining the in vivo function of cystatins.

Acknowledgements: We would like to thank Dr Marie Philips for assisting with the sequence analysis of the genomic clones and for critically reading the manuscript. This work was supported by a grant from the NH & MRC.

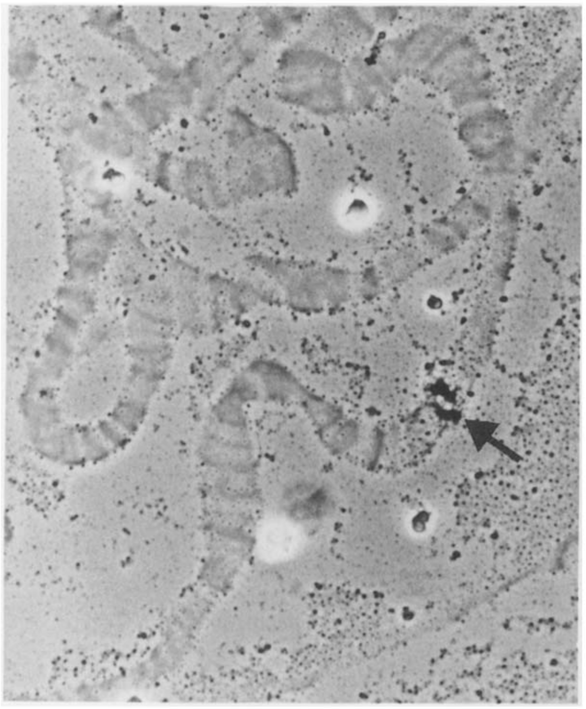

Fig. 5. Localization of the *Drosophila* cystatin-like gene to polytene chromosomes. The arrow indicates the hybridization at 88C on the right arm of chromosome 3.

REFERENCES

[1] Sen, L.C. and Whitaker, J.R. (1973). Arch. Biochem. Biophys. 158, 623-632.

[2] Machleidt, W., Borchart, U., Fritz, H., Brzin, J., Ritonja, A. and Turk, V. (1983) Hoppe-Seyler's Z. Physiol. Chem. 364, 1481-1486.

[3] Takio, K., Kominami, E., Wakamatsu, N., Katunuma, N. and Titani, K. (1983) Biochem. Biophys. Res. Commun. 115, 902-908.

[4] Grubb, A. and Lofberg, H. (1982) Proc. Natl. Acad. Sci. USA 79, 3024-3027.

[5] Isemura, S., Saitoh, E., Sanda, K., Isemura, M. and Ito, S. (1984) J. Biochem. 96, 1311-1314.

[6] Ritonja, A., Machleidt, W. and Barrett, A.J. (1985) Biochem. Biophys. Res. Commun. 131, 1187-1192.

[7] Hirado, M., Tsunasawa, S., Sakiyama, T., Niinobe, M. and Fujii, S. (1985) FEBS Lett. 186, 41-45.

[8] Abe, K., Kondo, H. and Arai, S. (1987) Agric. Biol. Chem. 51, 2763-2768.

[9] Müller-Estrel, W., Fritz, H., Kellermann, J., Lottspeich, F., Machleidt, W. and Turk, V. (1983) FEBS Lett. 191, 221-226.

[10] Barrett, A.J., Rawlings, N.D., Davies, M.E., Machleidt, W., Salvesen, G. and Turk, V. (1986) in: Proteinase Inhibitors, (Barrett, A.J. and Salvesen, G. eds) Vol. 12, Elsevier, Amsterdam, pp. 515-569.

[11] Schwabe, C., Anastasi, A., Crow, H., McDonald, J.K. and Barrett, A.J. (1984) Biochem. J. 217, 813-817.

[12] Abe, K., Emori, Y., Kondo, H., Suzuki, R. and Arai, S. (1987) J. Biol. Chem. 262, 16793-16797.

[13] Kelly, L.E. (1990) Biochem. J. (in press.).

[14] Young, R.A. and Davis, R.W. (1983) Science 222, 778-782.

[15] Itoh, N., Salvaterra, P. and Itakura, K. (1985) D.I.S. 61, 89.

[16] Sanger, F., Coulson, A.R., Barell, B.G., Smith, A.J.H. and Roe, B. (1980) J. Mol. Biol. 143, 161-178.

[17] Maniatis, T., Fritsch, E.F. and Sambrook, J. (1983) Molecular Cloning, Cold Spring Harbor Laboratories, Cold Spring Harbor, New York.

[18] Engels, W.R., Preston, C.R., Thompson, P. and Eggleston, W.B. (1985) Focus 8, 6-8.

[19] Cavener, D.R. (1987) Nucleic Acids Res. 15, 1353-1361.

[20] Bode, W., Engh, R., Musil, D., Thiele, U., Huber, R., Karshikov, A., Brzin, J., Kos, J. and Turk, V. (1988) EMBO J. 7, 2593-2599.

[21] Colella, R., Sakaguchi, Y., Nagase, H. and Bird, J.W.C. (1989) J. Biol. Chem. 264, 17164-17169.