Sequence of the Structural Gene for Xanthine Dehydrogenase (*rosy* Locus) in *Drosophila melanogaster*

Tim P. Keith,*[,1] Margaret A. Riley,* Martin Kreitman,*[,2] R. C. Lewontin,* Daniel Curtis[†] and Geoffrey Chambers*[,3]

**Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138 and*
[†]*Department of Biological Chemistry, Harvard Medical School, Boston, Massachusetts 02115*

Manuscript received December 1, 1986
Revised copy accepted January 26, 1987

ABSTRACT

We determined the nucleotide sequence of a 4.6-kb *Eco*RI fragment containing 70% of the *rosy* locus. In combination with information on the 5′ sequence, the gene has been sequenced in entirety. *rosy* cDNAs have been isolated and intron/exon boundaries have been determined. We find an open reading frame which spans four exons and would encode a protein of 1335 amino acids. The molecular weight of the encoded protein (xanthine dehydrogenase), based on the amino acid translation, is 146,898 daltons which agrees well with earlier biophysical estimates. Characteristics of the protein are discussed.

THE *rosy* locus (*ry*: 3–52.0) in *Drosophila melanogaster* is of particular interest from two viewpoints. First, it has been the subject of intensive fine structure genetic analysis by those interested in gene structure and regulation (CHOVNICK *et al.* 1977). Electrophoretic variants and null *rosy* mutants have served to delimit the structural boundaries of the gene (MCCARRON, GELBART and CHOVNICK 1974; GELBART, MCCARRON and CHOVNICK 1976). Two putative control variant sites have been genetically mapped to the 5′ region of the gene (CHOVNICK *et al.* 1976). Recently this genetic analysis has been extended to the molecular level using molecular mapping of insertion/deletion mutants (COTE *et al.* 1986) and DNA sequence analysis of putative control mutants (LEE *et al.* 1987).

Quite independently of studies of gene structure and regulation, the *rosy* locus became a focus of interest for population and evolutionary genetics. The encoded protein, xanthine dehydrogenase, is highly polymorphic in natural populations of all species of *Drosophila* where it has been studied. BUCHANON and JOHNSON (1983) found 15 electromorphs in 62 genomes sampled from a single population of *D. melanogaster*. KEITH *et al.* (1985), in a survey of 184 genomes from two California populations of *Drosophila pseudoobscura*, revealed 20 electromorphs, and COYNE (1976) revealed 23 electromorphs in 60 genomes sampled from a single population of *D. persimilis*.

Considering the extent of protein polymorphism, it is of interest to know what amino acid changes correspond to these electromorphs, whether this variation is confined to certain domains of the protein, and how important recombination may be in generating the variation. In addition, the ratio of silent site polymorphism to amino acid substitutions for such a polymorphic gene can be compared to that found at the *Adh* locus (KREITMAN 1983), which has a much lower level of protein polymorphism. Finally, sequence comparisons between species of *Drosophila* can show the rate of evolution of silent sites and intron positions for this highly polymorphic gene as compared to the rate obtained from *Adh* (S. SCHAEFFER and C. AQUADRO, unpublished data) It will be of interest to determine whether there is a correlation between the level of amino acid substitution and the level of overall DNA polymorphism observed.

Because of the interest of both molecular and population geneticists in the expression and evolution of *Xdh*, it is desirable to provide the complete DNA sequence of this locus. In this paper we present an overview of the structure of the *rosy* locus, its DNA sequence and predicted amino acid sequence. In addition, some of the characteristics of the protein, *XDH*, are discussed.

MATERIALS AND METHODS

DNA plasmids and fragments: The *rosy* locus was cloned by BENDER, SPEIRER and HOGNESS (1983) from a Canton-S stock of *D. melanogaster*. We subcloned into pBR322 a 4.6-kb *Eco*RI fragment (Figure 1) from the original 8.1-kb *Sal*I fragment kindly provided by C. S. LEE and W. BENDER. The sequence of the contiguous 5′ region, from the *Pst*I site at −2920 kb to the *Eco*RI site at 0 kb (Figure 1), has been sequenced by LEE *et al.* (1987) from a ry^{+5} laboratory stock and is presented in this issue of GENETICS.

[1] Present address: Collaborative Research Inc., 2 Oak Park, Bedford, Massachusetts 01730.
[2] Present address: Department of Biology, Princeton University, Princeton, New Jersey 08544.
[3] Present address: Department of Biochemistry, Wellington University, Victoria, New Zealand.

The sequence data presented in this article have been submitted to the EMBL/GenBank Data Libraries under the accession number Y00308.

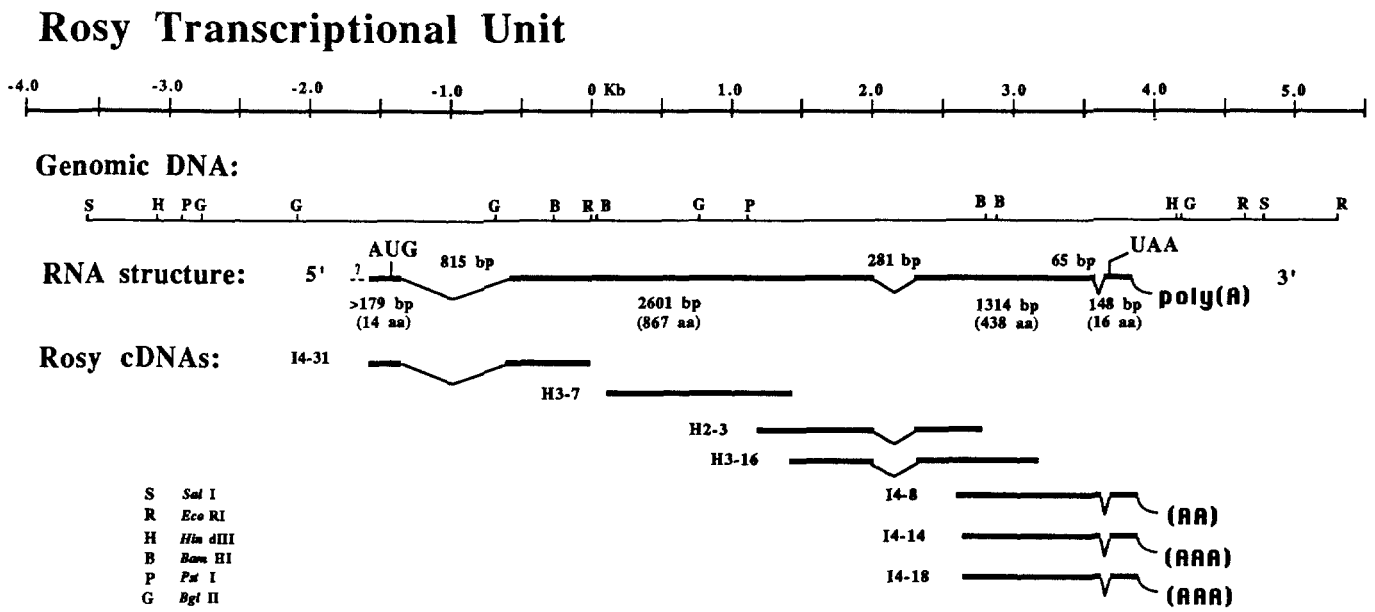

FIGURE 1.—The *rosy* transcriptional unit: a genomic restriction map for the region of the *rosy* gene is pictured in the second line, with coordinates in kilobases shown above. The coordinate of 0 kb is placed at the *Eco*RI site near the center of the gene. The RNA structure shown on the third line is a composite deduced from partial cDNAs. The locations of the seven *rosy* cDNAs we have isolated are given below the composite picture. The DNA sequence reported in this paper extends from the 4.6-kb *Eco*RI site at 0 kb to the *Eco*RI site at +4.6 kb.

DNA sequencing: The 4.6-kb *Eco*RI fragment was self-ligated and sonicated, and random fragments of approximately 600 bp were subcloned into M13 strain MP8 (NORANDER, KEMPE and MESSING 1983) according to the BANKIER and BARREL (1983) protocol. The clones obtained were sequenced according to the methods of SANGER, NICKLEN and COULSON (1977, 1980) on TBE buffer gradient gels (BIGGIN, GIBSON and HONG 1983). Sequenced fragments were overlapped into a single sequence using the programs of STADEN (1982, 1984). Confidence in the sequence was obtained by repeatedly sequencing the same region on both strands wherever possible. On average, a specific nucleotide was covered six times by independent clones. One percent of the *Eco*RI fragment was sequenced only once. In addition, 696 bp were sequenced on only one strand but these regions were repeatedly covered by four to six independent clones.

cDNA isolation: cDNA libraries made from *D. melanogaster* early and late third instar larval RNA in the vector lambda gt10 were kindly provided by L. KAUVAR, B. DRESS, S. POOLE and T. KORNBERG (POOLE *et al.* 1985). cDNA phage were plated onto bacterial strain KH802, and 700,000 plaques were screened using nick-translated *rosy* 4.6-kb *Eco*RI fragment, or a nick-translated fragment extending from the *Bcl*I site at −1837 kb to the *Eco*RI site at 0 kb (Figure 1). After plaque purification of phage containing *rosy* cDNAs, the cDNA inserts were cloned into pEMBL vectors (DENTE, CESARINI and CORTESE 1983). Convenient restriction sites were used to subclone smaller fragments of the cDNAs into pEMBL, and sequence was determined by the Sanger dideoxy method (SANGER, NICKLEN and COULSON 1977).

Protein analysis: The translated sequence of *XDH* was analyzed for amino acid composition and hydrophobicity using the programs from International Biotechnologies Incorporated (IBI) written by JAMES PUSTELL. Secondary structure predictions of the protein were determined using the method of CHOU and FASMAN (1978).

GENERAL STRUCTURE OF THE GENE

Extensive genetic (CHOVNICK, BALLANTYNE and HOLM 1971; GELBART, MCCARRON and CHOVNICK 1979; CLARK *et al.* 1984) and molecular (COTE *et al.* 1986) mapping of *rosy* mutants indicated that most or all of the *XDH* protein coding sequences were contained within a single 4.6-kb *Eco*RI fragment (Figure 1). Alignment of the genetic and molecular maps placed *rosy cis*-acting control sites to the left of this *Eco*RI fragment and suggested that the entire gene was contained within an 8.1-kb *Sal*I fragment (COTE *et al.* 1986). Transformation experiments have shown that a 7.3-kb *Hin*dIII fragment (Figure 1) contains all sequences necessary to rescue the *rosy* mutant phenotype (RUBIN and SPRADLING 1982). Insertions into the *Pst*I site at −2.9 kb have no effect on *rosy* expression (CLARK and CHOVNICK 1986), which further limits the extent of the putative control region. We therefore were confident that the *Pst*I to *Hin*dIII fragment (−2.9 to +4.2) contained all of the *rosy* sequence. Our laboratory sequenced the 4.6-kb *Eco*RI fragment containing the majority of the structural gene. That sequence and the accompanying protein translation is presented in Figure 2. LEE *et al.* simultaneously sequenced the contiguous 2.9-kb *Pst*I-*Eco*RI fragment (Figure 1) and that sequence is presented in the accompanying paper (1987). The *Pst*I-*Eco*RI fragment was obtained from a ry^{+5} laboratory stock, whereas the 4.6-kb *Eco*RI fragment came from a Canton-S stock. The ry^{+5} sequence was extended 200 bp beyond the *Eco*RI site at 0 kb to ensure that no small *Eco*RI fragments were lost at the junction. In that overlap, there is one silent polymorphism, a G in Canton-S *vs.* a T in ry^{+5} at position +74 in the DNA sequence.

The *rosy* gene is transcribed from left to right, as determined by hybridization of single stranded probes to the *rosy* message (COTE *et al.* 1986). Examination

of the complete sequence reveals a long open reading frame in the correct orientation which begins at the ATG at −1407 in the first exon, splices across three introns and terminates at the TAA codon at +3760 in the fourth exon. Analysis of *rosy* point mutations supports our belief that the −1407 ATG is the translational start site (LEE *et al.* 1987).

In order to determine the precise limits of the transcribed regions of the *rosy* gene, we searched for *rosy* cDNA clones. From the Oregon-R early and late third instar cDNA libraries of POOLE *et al.* (1985) we isolated seven partial *rosy* cDNAs. Two of these, I4-14 and I4-18, appear identical, although they were isolated in separate screenings. The cDNA clones overlap as diagramed in Figure 1, and when combined they cover nearly the entire gene. There is a gap in the coverage of 158 bases, at the *Eco*RI site. Two cDNAs end near or at this site, which is probably an artifact of the construction of the libraries (if the double stranded cDNA were not completely methylated at internal *Eco*RI sites, or if restriction enzyme contaminated the methylase preparation, these sites would be cut). Since the *rosy* open reading frame continues uninterrupted through this region, it is unlikely that any additional introns are located within this 158-bp gap.

The cDNAs reveal the positions of the four exons and three introns in the *rosy* gene (Figure 1). The intron/exon splice junction sequences agree well with the consensus sequences derived by MOUNT (1982) and KELLER and NOON (1985). Although the cDNAs were not sequenced in entirety, mapping with 4-base-recognition enzymes shows that there are no additional introns in the regions covered by the cDNA clones. The 5′ cDNA clone I4-31 extends 132 bp 5′ of the AUG codon at −1407 which initiates the *rosy* long open reading frame, but we have not determined if the cDNA is complete at its 5′ end.

A total of 2676 bp of the cDNAs was sequenced. There were seven nucleotide substitutions between the Oregon-R cDNA sequence and the Canton-S genomic sequence, all conservative third position changes (data not shown). *rosy* mRNA is polyadenylated (COVINGTON, FLEENOR and DEVLIN 1984). The cDNA I4-8, I4-14 and I4-18 all depart from the *rosy* genomic sequence at the same base (+3859) and this base is followed in the cDNAs by poly-A tracts of 19–20 residues. At 19 bp preceding the site of poly-A addition is the sequence AATTAAA, a variation of the conserved polyadenylation signal AATAAA (reviewed by BIRNSTIEL, BUSSLINGER and STRUB 1985). Base pair +3859 is apparently the 3′ boundary of the *rosy* mature mRNA.

There is an additional open reading frame of 115 codons within the sequence, beyond the 3′ end of *rosy*. This frame reads in the opposite orientation from *rosy*. It begins at the right boundary of our sequence, so the full size of the reading frame is unknown. The next characterized gene 3′ to *rosy* is *snake*, but this open reading frame does not correspond to the *snake* gene. A *rosy* null mutation, ry^{506}, is a 3.4-kb deletion beginning at about +1.1 kb and extending into the next distal *Eco*RI fragment, (+4.6 to +5.3 kb, Figure 1) (COTE *et al.* 1986), and this deletion has no *snake* phenotype. In addition, *snake* cDNAs have been isolated and do not extend into the 4.6-kb *Eco*RI fragment (DELOTTO and SPIERER 1986).

PROTEIN PROPERTIES

The translated polypeptide (xanthine dehydrogenase EC 2.1.37) is predicted to be 1335 amino acids long. The xanthine dehydrogenase amino acid sequence shows no discernible homologies with any of the proteins in the Protein Identification Resource Database (March 1986; National Biomedical Research Foundation). The program searches for 40% homology over 40 amino acids or seven consecutive amino acids between two proteins. An additional protein homology search was performed using the LIPMAN and PEARSON fast protein homology search programs contained in the MBCRR distributed Molecular Biology Analysis Programs. With a possible score of 6661 for 100% homology for this protein, the highest score produced in this search was 59 in a comparison with baker's yeast histone H3. As xanthine dehydrogenase is both a dehydrogenase and a molybdenum binding enzyme it was of particular interest to compare the amino acid sequence in more detail with other dehydrogenases and molybdenum binding enzymes, in order to search for limited regions of homology that may be related to the proteins structural requirements. No dehydrogenases or molybdenum binding proteins included in the database were shown to have even short regions of homology with the *XDH* sequence.

The amino terminus of the protein has been examined for indications of a signal sequence. HEIJNE (1985) describes three well-defined functional domains that are highly conserved in all eukaryotic signal sequences examined to date. These include a short, positively charged *n*-terminal region, a strongly hydrophobic region, and a short polar stretch terminating in a cleavage site. *XDH* begins with three polar residues, followed by six hydrophobic amino acids, followed by a polar region including a potential signal sequence cleavage site between amino acids 12 and 13. The *n*-terminal region and the cleavage site sequence fall within the limits for eukaryotic signal sequences. However, the hydrophobic region is one residue shorter than the shortest example in a large survey of signal sequences (HEIJNE 1985). Thus we are uncertain if the amino terminus of *XDH* can function as a secretion signal.

XDH has been characterized as a soluble protein.

```
+1
(EcoRI)   11  (BamHI) 23          35          47  (Pvu2) 59          71          83          95         107         119
GAATTCCAGCCCTTGGATCCCAGCCAGGAACCCATCTTCCCACCGGAACTTCAGCTGAGTGACGCCTTCGATTCGCAGAGTTTGATCTTTAGTTCGGATAGGGTGACCTGGTATCGTCCC
GluPheGlnProLeuAspProSerGlnGluProIlePheProProGluLeuGlnLeuSerAspAlaPheAspSerGlnSerLeuIlePheSerSerAspArgValThrTrpTyrArgPro

          131         143         155  (5'end of H3-7 cDNA clone)     191         203         215         227         239
ACCAATCTGGAGGAGCTGCTTCAGCTGAAGGCCAAACATCCGGCTGCCAAGCTGGTCGTGGGCAATACGGAAGTGGGCGTTGAGGTTAAGTTCAAGCACTTCCTCTACCCGCACCTCATC
ThrAsnLeuGluGluLeuLeuGlnLeuLysAlaLysHisProAlaAlaLysLeuValValGlyAsnThrGluValGlyValGluValLysPheLysHisPheLeuTyrProHisLeuIle

          251         263         275         287         299         311         323 (ClaI)  335         347         359
AATCCCACCCAGGTGAAGGAGCTGCTGGAGATCAAAGAGAACCAGGATGGCATTTACTTCGGTGCGGCTGTCAGTTTGATGGAGATCGATGCGCTTCTGCGGCAGAGAATCGAGCTGCTG
AsnProThrGlnValLysGluLeuLeuGluIleLysGluAsnGlnAspGlyIleTyrPheGlyAlaAlaValSerLeuMetGluIleAspAlaLeuLeuArgGlnArgIleGluLeuLeu

          371         383         395         407         419         431         443         455         467         479
CCGGAATCGGAGACCAGATTGTTCCAGTGCACCGTGGATATGCTTCACTACTTTGCCGGCAAGCAGATCCGCAACGTCGCCTGTTTGGGTGGAAACATCATGACCGGCAGTCCCATTTCC
ProGluSerGluThrArgLeuPheGlnCysThrValAspMetLeuHisTyrPheAlaGlyLysGlnIleArgAsnValAlaCysLeuGlyGlyAsnIleMetThrGlySerProIleSer

          491         503  (SstI)         527         539         551         563         575         587         599
GATATGAATCCTGTGCTCTCGGCAGCAGGAGCTCAACTGGAGGTGGCCAGTTTTGTGGATGGAAAGCTCCAAAAGAGATCAGTTCACATGGGAACTGGGTTCTTCACTGGCTATCGCAGG
AspMetAsnProValLeuSerAlaAlaGlyAlaGlnLeuGluValAlaSerPheValAspGlyLysLeuGlnLysArgSerValHisMetGlyThrGlyPhePheThrGlyTyrArgArg

          611         623         635         647         659         671         683         695         707         719
AATGTTATCGAAGCCCACGAGGTGCTGCTGGGCATCCACTTTCGGAAGACCACTCCGGACCAGTATATCGTTGCTTTTAAGCAGGCCAGAAGAAGGGATGATGACATAGCCATCGTAAAT
AsnValIleGluAlaHisGluValLeuLeuGlyIleHisPheArgLysThrThrProAspGlnTyrIleValAlaPheLysGlnAlaArgArgArgAspAspIleAlaIleValAsn

          731         743         755         (Bgl2)      779         791         803         815         827         839
GCCGCAATAAACGTTCGCTTTGAGGAAAAATCCAACATTGTGGCGGAGATCTCGATGGCTTTTGGTGGAATGGCACCAACCACAGTGCTCGCTCCTCGAACTTCCCAACTGATGGTTGGG
AlaAlaIleAsnValArgPheGluGluLysSerAsnIleValAlaGluIleSerMetAlaPheGlyGlyMetAlaProThrThrValLeuAlaProArgThrSerGlnLeuMetValGly

          851         863         875         887         899         911         923         935  (SstI)         959
CAGGAGTGGAGCCACCAGCTCGTGGAGCGCGTGGCGGAGAGCTTGTGCACGGAGCTGCCTTTGGCTGCCTCCGCTCCGGGTGGCATGATCGCCTATCGTCGAGCTCTGGTGGTGAGCCTG
GlnGluTrpSerHisGlnLeuValGluArgValAlaGluSerLeuCysThrGluLeuProLeuAlaAlaSerAlaProGlyGlyMetIleAlaTyrArgArgAlaLeuValValSerLeu

          971         983         995        1007        1019        1031        1043        1055        1067        1079
TTCTTCAAGGCCTATCTGGCCATTTCCCTGAAGCTGAGCAAGTCAGGGATCACATCTTCCGATGCTCTGCCACCGGAGGAGCGAAGTGGTGCCGAGACATTCCACACACCAGTACTCAAA
PhePheLysAlaTyrLeuAlaIleSerLeuLysLeuSerLysSerGlyIleThrSerSerAspAlaLeuProProGluGluArgSerGlyAlaGluThrPheHisThrProValLeuLys

         1091        1103  (PstI)        1127        1139        1151        1163        1175        (5'end of H2-3cD
AGTGCCCAGCTCTTCGAGCGCGTCTGCAGCGATCAACCCATCTGTGATCCCATTGGGAGACCAAAAGTTCATGCGGCTGCTTTGAAACAGGCCACTGGTGAAGCTATCTACACAGATGAC
SerAlaGlnLeuPheGluArgValCysSerAspGlnProIleCysAspProIleGlyArgProLysValHisAlaAlaAlaLeuLysGlnAlaThrGlyGluAlaIleTyrThrAspAsp

NA clone)         1223        1235        1247        1259        1271        1283        1295        1307        1319
ATTCCCCGCATGGATGGTGAAGTTTATCTGGCCTTTGTCCTTAGTACCAAGCCACGTGCCAAGATCACCAAGCTGGATGCCAGTGAAGCTCTGGCCCTGGACGGAGTGCATCAGTTCTTT
IleProArgMetAspGlyGluValTyrLeuAlaPheValLeuSerThrLysProArgAlaLysIleThrLysLeuAspAlaSerGluAlaLeuAlaLeuAspGlyValHisGlnPhePhe

         1331        1343        1355        1367        1379        1391        1403        (5'end of H3-16 cDNA clone)
TGCTACAAGGACTTAACGGAGCACGAGAACGAAGTGGGACCCGTCTTTCATGATGAGCACGTCTTTGCCGCTGGAGAAGTGCATTGCTATGGTCAGATAGTGGGCGCCATAGCTGCCGAT
CysTyrLysAspLeuThrGluHisGluAsnGluValGlyProValPheHisAspGluHisValPheAlaAlaGlyGluValHisCysTyrGlyGlnIleValGlyAlaIleAlaAlaAsp

(3'end of H3-7)   1463        1475        1487        1499        1511        1523        (SstI)        1547        1559
AATAAGGCGTTGGCCCAAAGAGCCGCTCGGCTAGTGAAAGTGGAGTACGAGGAGCTGAGCCCGGTTATCGTGACCATAGAGCAGGCCATCGAGCTCAAGTCCTATTTCCCGGACTACCCC
AsnLysAlaLeuAlaGlnArgAlaAlaArgLeuValLysValGluTyrGluGluLeuSerProValIleValThrIleGluGlnAlaIleGluLeuLysSerTyrPheProAspTyrPro

         1571        1583        1595        1607        1619        1631        1643        1655        1667        1679
CGATTCGTGACCAAGGGCAATGTGGAGGAGGCTTTATCCCAGGCGGATCACACTTTCGAGGGCACCTGTCGAATGGGCGGACAGGAGCACTTCTATCTGGAGACCCATGCTGCATTGGCC
ArgPheValThrLysGlyAsnValGluGluAlaLeuSerGlnAlaAspHisThrPheGluGlyThrCysArgMetGlyGlyGlnGluHisPheTyrLeuGluThrHisAlaAlaLeuAla

         1691        1703        1715        1727        1739        1751        1763        1775        1787        1799
GTACCTCGTGACAGCGATGAGCTGGAACTCTTTTGCTCCACGCAGCATCCCTCGGAGGTGCAGAAGCTAGTGGCCCATGTAACCGCACTTCCTGCCCACCGTGTCGTCTGTCGTGCCAAG
ValProArgAspSerAspGluLeuGluLeuPheCysSerThrGlnHisProSerGluValGlnLysLeuValAlaHisValThrAlaLeuProAlaHisArgValValCysArgAlaLys

         1811        1823        1835        1847        1859        1871        1883        1895        1907        1919
CGTTTGGGAGGCGGTTTCGGCGGCAAGGAGTCCAGAGGCATCTCCGTGGCCCTACCCGTTGCCCTGGCCGCCTATCGAATGGGTCGTCCTGTGCGCTGTATGTTGGATCGCGATGAGGAC
ArgLeuGlyGlyGlyPheGlyGlyLysGluSerArgGlyIleSerValAlaLeuProValAlaLeuAlaAlaTyrArgMetGlyArgProValArgCysMetLeuAspArgAspGluAsp

         1931        1943        1955        1967        1979        (Bcl1)      2003        2015        2027        2039
ATGCTTATCACCGGCACCAGGCATCCCTTCCTCTTCAAATACAAAGTGGGCTTCACCAAGGAGGGTCTGATCACTGCCTGCGACATTGAGTGCTACAACAATGCCGGTTGGTCCATGGAT
MetLeuIleThrGlyThrArgHisProPheLeuPheLysTyrLysValGlyPheThrLysGluGlyLeuIleThrAlaCysAspIleGluCysTyrAsnAsnAlaGlyTrpSerMetAsp

         2051        2063        2075        2087        2099        2111        2123        2135        2147        2159
CTGTCATTTTCGGTAAGAGTGGTGGGTTTATGGAAAATCCATATGTAGGTTTCTTTATGCATACCATGTTTGCTCAAATATCCATGTGTTGTATTTTCCATTGAACATCTGGTTAGTGTG
LeuSerPheSer(-- intron 2

         2171        2183        2195        2207        2219        2231        2243        2255        2267        2279
GGAATTTCTAGGACCCCTATATGTTATCATCCTGAGTGCAAATAGTTTGAAACTTTTAAGAATGTTTAGGTCTAATCAGGAAAGCAAGCACCAGGATTATATTGACATAAACAATTATAA

         2291        2303        2315        2327        2339        2351        2363        2375        2387        2399
ATAATAATCTAAAAATTCTTAAAAATGATCTAATATATAAATCCTATGTTTAGGTTCTTGAGCGCGCCATGTTCCACTTTGAGAATTGCTACAGGATTCCCAACGTTCGCGTGGGTGGAT
                                                                intron 2 --)ValLeuGluArgAlaMetPheHisPheGluAsnCysTyrArgIleProAsnValArgValGlyGlyT
```

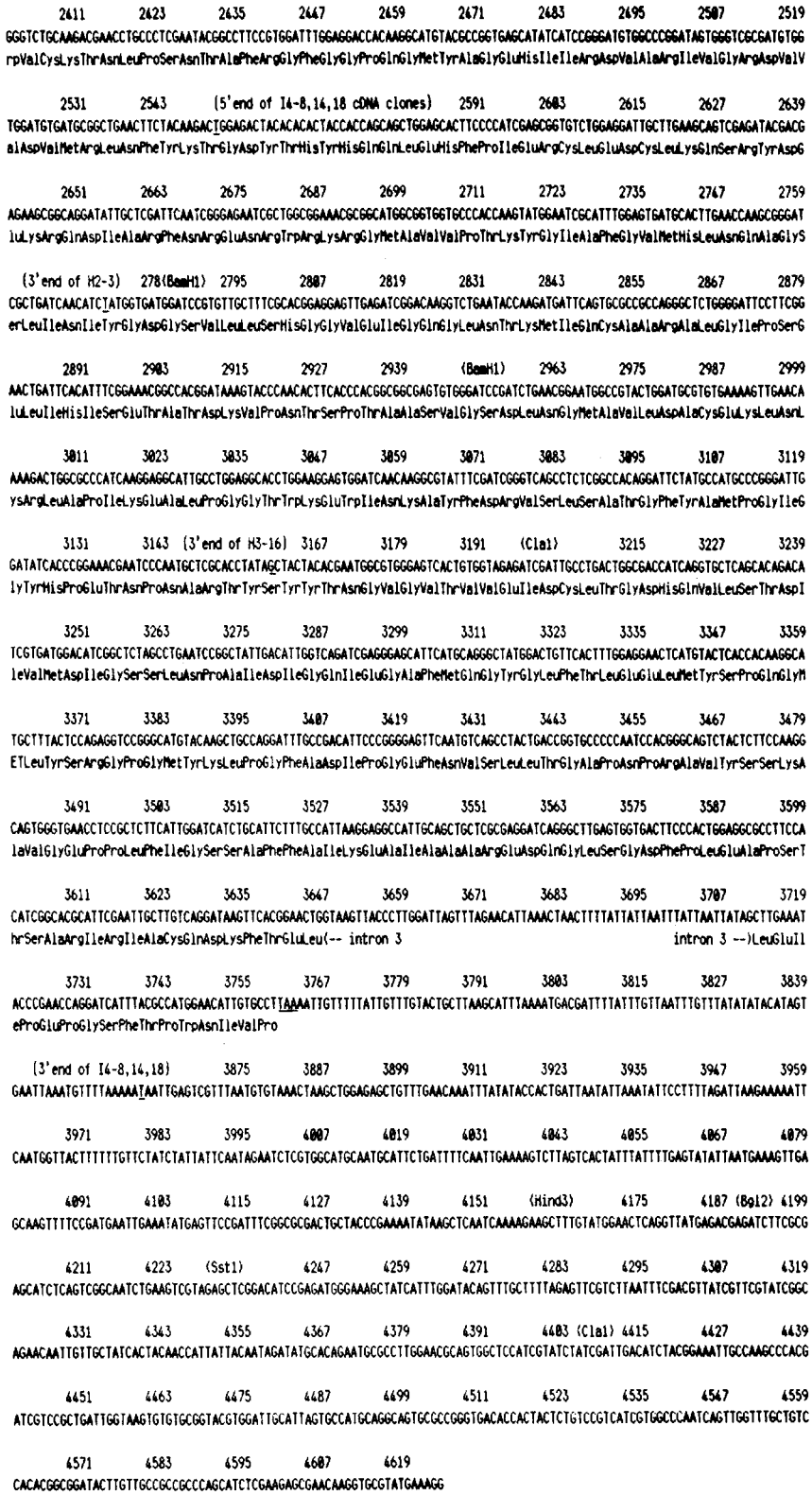

FIGURE 2.—The DNA sequence of the 4.6-kb *Eco*RI fragment containing 70% of the *rosy* locus: the sequence is numbered as in LEE *et al.* (1987) (1+ is underlined). Selected restriction sites of six-base recognition enzymes are indicated above the DNA sequence. The predicted *XDH* protein sequence is shown below the DNA sequence. cDNA boundaries are underlined and the numbers noted above the DNA sequence. Intron boundaries are noted below the DNA sequence. The TAA stop codon is underlined.

Consistent with that, we do not find any transmembrane-like stretches of amino acids. This feature was examined according to the procedure outlined by KYTE and DOOLITTLE (1982). Each hydrophobic stretch of 19 amino acids was assigned an average hydropathy value. None of these averages were equal

to or above the value of 1.6, which is the lower limit value associated with transmembrane amino acid sequences (KYTE and DOOLITTLE 1982).

The functional protein is a homodimer with a subunit molecular weight of 146,898 daltons as determined from the translated sequence. This is in good agreement with the subunit weight of 150,000 previously estimated by SDS gel electrophoresis (EDWARDS and CANDIDO 1977).

DISCUSSION

We have determined the sequence of the *rosy* locus to serve as the basis for subsequent sequence comparisons. Both the genomic and cDNA sequences, reported in this paper, have helped define the limits of the mature mRNA. This information will help direct a search for *cis*-acting control regions of the gene (LEE *et al.* 1987).

In addition, structural features of *Xdh* described in this paper make it an interesting locus for population and genetic studies. Since the gene is composed of both introns and exons, it allows the comparison of different functional regions. The large size of the locus permits more powerful statistical analyses of these comparisons than were possible in the smaller loci analyzed to date, including *Hsp*82 (BLACKMAN and MESELSON 1986) and *Adh* (SCHAEFFER and AQUADRO 1987). We plan to focus future work on a sequence comparison of different *XDH* alleles isolated from natural populations (KEITH *et al.* 1985). These data will provide information on the distribution and type of amino acid substitutions permitted in the molecule and on the origin and maintenance of genetic variation at this locus.

We thank DOREEN LEWIS for assistance with fly food. This work was supported by grants from the National Institutes of Health to R.C.L.

LITERATURE CITED

BANKIER, A. T. and B. G. BARREL, 1983 Shotgun DNA Sequencing. Laboratory of Molecular Biology: MRC Centre, Cambridge.

BENDER, W., P. SPEIRER and D. S. HOGNESS, 1983 Chromosomal walking and jumping to isolate DNA from the *Ace* and *rosy* loci and the bithorax complex in *Drosophila melanogaster*. J. Mol. Biol. **168:** 17–33.

BIGGEN, M.D., T. J. GIBSON and G. F. HONG, 1983 Buffer gradient gels and 35S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80:** 3963–3965.

BIRNSTIEL, M. L., M. BUSSLINGER and K. STRUB, 1985 Transcription, termination and 3′ processing: the end is in site! Cell **41:** 349–359.

BLACKMAN, R. K. and M. MESELSON, 1986 Interspecific nucleotide sequence comparisons used to identify regulatory and structural features of the *Drosophila hsp82* gene. J. Mol. Biol. **188:** 499–515.

BUCHANON, B. A. and D. L. E. JOHNSON, 1983 Hidden electrophoretic variation at the xanthine dehydrogenase locus in a natural population of *Drosophila melanogaster*. Genetics **104:** 301–315.

CHOU, P. Y. and G. D. FASMAN, 1978 Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol **47:** 45–148.

CHOVNICK, A., G. H. BALLANTYNE and D. G. HOLM, 1971 Studies on gene conversion and its relationship to linked exchange in *Drosophila melanogaster*. Genetics **69:** 179–209.

CHOVNICK, A. W., W. GELBART, M. MCCARRON, B. OSMOND, E. P. M. CANDIDO and D. L. BAILLIE, 1976 Organization of the *rosy* locus in *Drosophila melanogaster*. Evidence for a control element adjacent to the xanthine dehydrogenase structural element. Genetics **84:** 233–255.

CHOVNICK, A. W., M. MCCARRON, A. HILLIKER, J. O'DONNELL, W. GELBART and S. CLARK, 1977 Gene organization in *Drosophila*. Cold Spring Harbor Symp. Quant. Biol. **42:** 1011–1021.

CLARK, S. H. and A. CHOVNICK, 1986 Studies of normal and position-affected expression of *rosy* region genes in *Drosophila melanogaster*. Genetics **114:** 819–840.

CLARK, S. H., S. DANIELS, C. A. RUSHLOW, A. J. HILLIKER and A. CHOVNICK, 1984 Tissue-specific and pretranslational character of variants of the rosy locus control element in *Drosophila melanogaster*. Genetics **108:** 953–968.

CLARK, S. H., M. MCCARRON, C. LOVE and A. CHOVNICK, 1986 On the identification of the *rosy* locus DNA in *Drosophila melanogaster*: intragenic recombination mapping of mutations associated with insertions and deletions. Genetics **112:** 755–767.

COTE, B., W. BENDER, D. CURTIS and A. CHOVNICK, 1986 Molecular mapping of the *rosy* locus in *Drosophila melanogaster*. Genetics **112:** 769–783.

COVINGTON, M., D. FLEENOR and R. B. DEVLIN, 1984 Analysis of xanthine dehydrogenase mRNA levels in mutants affecting the expression of the *rosy* locus. Nucleic Acids Res. **12:** 4559–4573.

COYNE, J., 1976 Lack of genetic similarity between two sibling species of *Drosophila* as revealed by varied techniques. Genetics **84:** 593–607.

DELOTTO, R. and P. SPIERER, 1986 The *Drosophila melanogaster snake* locus: a gene required for specification of dorsal-ventral pattern appears to encode a serine protease. Nature **311:** 223–228.

DENTE, L., G. CESARINI and R. CORTESE, 1983 pEMBL: a new family of single stranded plasmids. Nucleic Acids Res. **11:** 1645–1655.

EDWARDS, T. C. R. and E. P. M. CANDIDO, 1977 Xanthine dehydrogenase from *Drosophila melanogaster*. A comparison of the kinetic parameters of the pure enzyme form two wild-type isoalleles differing at a putative regulatory site. Mol. Gen. Genet. **154:** 1–6.

GELBART, W., M. MCCARRON and A. CHOVNICK, 1976 Extension of the limits of the *XDH* structural element in *Drosophila melanogaster*. Genetics **84:** 211–232.

HEIJNE, G., 1985 Signal sequences: the limits of variation. J. Mol. Biol. **184:** 99–105.

KEITH, T. P., L. D. BROOKS, R. C. LEWONTIN, J. C. MARTINEZ-CRUZADO and D. L. RIGBY, 1985 Nearly identical allelic distributions of xanthine dehydrogenase in two populations of *Drosophila pseudoobscura*. Mol. Biol. Evol. **2:** 206–216.

KELLER, E. B. and W. A. NOON, 1985 Intron splicing: a conserved internal signal in introns of *Drosophila* pre-mRNAs. Nucleic Acids Res. **13:** 4971–4981.

KREITMAN, M. K., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412–417.

KYTE, J. and F. DOOLITTLE, 1982 A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157:** 105–132.

LEE, C. S., D. CURTIS, M. MCCARRON, C. LOVE, M. GRAY, W. BENDER and A. CHOVNICK, 1987 Mutations affecting expression of the *rosy* locus in *Drosophila melanogaster*. Genetics. **116:** 55–66.

McCARRON, M., W. GELBART and A. CHOVNICK, 1974 Intracistronic mapping of electrophoretic sites in *Drosophila melanogaster*: fidelity of information transfer by gene conversion. Genetics **76:** 289–299.

McCARRON, M., J. O'DONNELL, A. CHOVNICK, B. S. BHULLAR, J. HEWITT and E. P. M. CANDIDO, 1979 Organization of the *rosy* locus in *Drosophila melanogaster*: further evidence in support of a *cis*-acting control element adjacent to the xanthine dehydrogenase structural element. Genetics **91:** 275–293.

MOUNT, S. M., 1982 A catalogue of splice junction sequences. Nucleic Acid Res. **10:** 459–472.

NORRANDER, T., T. KEMPE and J. MESSING, 1983 Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. Gene **26:** 101–106.

POOLE, S. J., L. M. KAUVAR, B. DREES and T. KORNBERG, 1985 The *engrailed* locus of *Drosophila*: structural analysis of an embryonic transcript. Cell **40:** 37–43.

RUBIN, G. M. and A. C. SPRADLING, 1982 Genetic transformation of *Drosophila* with transposable element vectors. Science **218:** 348–353.

SANGER, F., S. NICKLEN and A. R. COULSON, 1977 DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci USA **74:** 5463–5467.

SANGER, F., A. R. COULSON, B. G. BARRELL, A. J. H. SMITH and B. A. ROE, 1980 Cloning in single stranded bacteriophage as an aid to rapid DNA sequencing. J. Mol. Biol. **143:** 161–178.

STADEN, R., 1982 Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Res. **10:** 4731–4751.

STADEN, R., 1984 A computer program to enter DNA gel reading data into a computer. Nucleic Acids Res. **12:** 499–504.

Communicating editor: A. SPRADLING