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

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

ABSTRACT

We determined the nucleotide sequence of a 4.6-kb EcoRI 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.

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positions for this highly polymorphic gene as compared to the rate obtained from Adh (S. Schaeffer and C. Aquadro, unpublished data) It will be of interest to determine whether there is a correlation between the level of amino acid substitution and the level of overall DNA polymorphism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 Hoggness (1983) from a Canton-S stock of D. melanogaster. We subcloned into pBR322 a 4.6-kb EcoRI fragment (Figure 1) from the original 8.1-kb SalI fragment kindly provided by C. S. Lee and W. Bender. The sequence of the contiguous 5' region, from the PstI site at -2920 kb to the EcoRI 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.
Rosy Transcriptional Unit

DNA sequencing: The 4.6-kb EcoRI 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 EcoRI 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 (POOL et al. 1985). cDNA phage were plated onto bacterial strain KH802, and 700,000 plaques were screened using nick-translated rosy 4.6-kb EcoRI fragment, or a nick-translated fragment extending from the BclI site at -2.9 kb to the EcoRI site at 0 kb (Figure 1). After plaque purification of phage containing rosy cDNAs, the cDNA inserts were cloned into pEMBL vectors (DENTI, 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 EcoRI fragment (Figure 1). Alignment of the genetic and molecular maps placed rosy cis-acting control sites to the left of this EcoRI fragment and suggested that the entire gene was contained within an 8.1-kb SalI fragment (COTE et al. 1986). Transformation experiments have shown that a 7.3-kb HindIII fragment (Figure 1) contains all sequences necessary to rescue the rosy mutant phenotype (RUBIN and SPRADLING 1982). Insertions into the PstI 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 PstI to HindIII fragment (-2.9 to +4.2) contained all of the rosy sequence. Our laboratory sequenced the 4.6-kb EcoRI 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 PstI-EcoRI fragment (Figure 1) and that sequence is presented in the accompanying paper (1987). The PstI-EcoRI fragment was obtained from a ry+ laboratory stock, whereas the 4.6-kb EcoRI fragment came from a Canton-S stock. The ry+ sequence was extended 200 bp beyond the EcoRI site at 0 kb to ensure that no small EcoRI fragments were lost at the junction. In that overlap, there is one silent polymorphism, a G in Canton-S vs. a T in ry+ 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 ATG at -1407 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, 14-14 and 14-18, 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 EcoRI 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 EcoRI sites, or if restriction enzyme contaminated the methylation 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 14-31 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 14-8, 14-14 and 14-18 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 genomic sequence at the same base (+3859) and this site 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.
FIGURE 2.—The DNA sequence of the 4.6-kb EcoRI fragment containing 70% of the \textit{rosy} locus: the sequence is numbered as in 
\textit{Lee and \textit{et al.}} (1987). \textit{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 \textit{Kyte and Doolittle} (1982). Each hydrophobic stretch of 19 amino acids was assigned an average hydrophy 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 DOLLITTLE 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 Hsp82 (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**


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.


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


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