Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 5916–5920, August 1990
Genetics

Structural and functional comparisons of the *Drosophila virilis* and *Drosophila melanogaster* rough genes

(homeobox/eye development/evolution)

ULRIKE HEBERLEIN AND GERALD M. RUBIN

Howard Hughes Medical Institute and Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720

Contributed by Gerald M. Rubin, May 23, 1990

ABSTRACT **We have isolated the homeobox gene rough (*ro*) from *Drosophila virilis*. Comparison of the predicted amino acid sequences of the *D. melanogaster* and *D. virilis* rough proteins reveals that domains of high conservation, including the homeodomain, are interspersed with highly diverged regions. Stretches of significant sequence conservation are also observed in the 5′ promoter region and in the introns. The *D. virilis* rough gene rescues the rough mutant phenotype and is properly regulated when introduced into the *D. melanogaster* genome. Thus the rough protein as well as the cis-regulatory elements that ensure proper temporal and spatial regulation are functionally conserved between these *Drosophila* species.**

The compound eye of *Drosophila* consists of several hundred units, or ommatidia, each containing a stereotyped arrangement of photoreceptor, pigment, and cone cells. These ommatidia develop during late larval and pupal life in the eye imaginal disc, in a process that involves the recruitment of undifferentiated epithelial cells into gradually growing ommatidial clusters (for review see refs. 1 and 2). The rough (*ro*) mutation disrupts cellular interactions at an early stage of ommatidial assembly, leading to irregularly arranged clusters containing variable numbers of photoreceptor cells (3).

The rough gene encodes a homeodomain protein (3, 35) and is believed to specify the identity of a subset of photoreceptor cells in the developing retina (4, 5). The rough protein is restricted to the eye imaginal disc, where it is expressed in a complex and dynamic pattern (4). Unlike mutations in other *Drosophila* homeobox genes, flies carrying complete loss-of-function alleles of rough are viable (unpublished data). This, together with the relatively small size of the rough gene, provides a unique opportunity to study structure–function relationships of a homeodomain protein in its natural developmental context.

As a first step to identify functionally relevant domains of the rough protein, as well as cis-regulatory DNA sequences required for proper regulation, we have compared the sequences of the rough genes from two distantly related *Drosophila* species, *D. melanogaster* and *D. virilis*.* These two species are separated by an evolutionary period of ≈60 million years (6), which is sufficiently distant for unconstrained DNA sequences to have diverged extensively, allowing putative functional elements to be identified by sequence conservation. To test whether the observed conservation is of functional importance, we introduced the *D. virilis* rough gene into the *D. melanogaster* genome and analyzed its function.

MATERIALS AND METHODS

A genomic *D. virilis* library in bacteriophage λ EMBL3 (ref. 7; a gift of M. Scott, Stanford University) was screened with a full-length *D. melanogaster* rough cDNA (3). Hybridizations were carried out at 42°C in 2× SSC (1× SSC is 0.15 M NaCl/0.015 M sodium citrate, pH 7) containing 35% formamide, 0.1% Ficoll, 0.1% bovine serum albumin, 0.1% polyvinylpyrrolidone, and 100 μg of sonicated salmon sperm DNA per ml. Washing conditions were 1× SSC/0.1% SDS, 50°C. Approximately one clone was obtained per genome screened. DNA blot analysis of the isolated phage DNA identified an 8-kilobase (kb) genomic *Sal* I fragment that hybridized with the *D. melanogaster* rough cDNA. This fragment was subcloned into pBluescript KS(+) (Stratagene), and random clones were generated by sonication and subcloning into phage M13. Sequencing was done by the chain-termination method (8). Both strands were sequenced except for two 100-base-pair (bp) regions in the first intron. Sequences were compiled and analyzed using the IntelliGenetics and University of Wisconsin Genetics Computer Group software packages.

The 8-kb DNA fragment containing the *D. virilis* rough gene was cloned into the *P*-element transformation vector pDM30 (9) and germ-line transformants were obtained by standard techniques (10).

Antibody staining of eye imaginal discs with the rough monoclonal antibody (MAbrol) was carried out exactly as described (4). Fixation and sectioning of adult *Drosophila* heads were performed as described (11).

RESULTS AND DISCUSSION

The *D. virilis* rough gene was isolated from a genomic library by virtue of its cross-hybridization with a *D. melanogaster* rough cDNA (see *Materials and Methods*). The regions of homologous sequence in the two genes were found to be completely contained within an 8-kb *D. virilis Sal* I fragment. The DNA sequence of most of this genomic fragment is shown in Fig. 1, which includes alignments with *D. melanogaster* protein-coding sequences. From the analysis of these alignments we conclude that the *D. virilis* DNA fragment contains all the protein-coding sequences, as well as ≈1 kb each of 5′ and 3′ noncoding DNA. The *D. melanogaster* rough protein is encoded by three exons. The DNA sequences of the splice sites and adjacent regions are conserved in the two species, arguing that the overall genomic organization is the same. A dot-matrix comparison of the *D. virilis* and *D. melanogaster* rough sequences is shown in Fig. 2. Although the homologies are concentrated in the coding regions, several stretches of highly conserved sequence are observed in each intron. The conservation at the DNA level in the three exons, calculated as percent nucleotide identity relative to the total number of nucleotides in the *D. melanogaster* sequence, is 46%, 81%, and 69% in the first, second, and third exons, respectively. It is difficult to calculate the overall conservation in the introns due to significant differences in their length. However, ≈20% of the sequence

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*The sequence reported in this paper has been deposited in the GenBank data base (accession no. M35372).

```
vir   GTACCCGGGGATCCTCTAGAGTCGACCTTAACGAATTCGTGCACTTCCCGGAACTCGACATCGTTGAGCATCAGGGTCCAGACGTTGTCGCAGAAGCGATAGGTGTTCAGTTTTCCCGCTTTGAATGTAACG  -1050
vir   CGAGCTTTAACCCTCTGATTAAGGGCGTTGTTAATGCTCTTGTCAAACTGCAGCAGCACCTTAAATGCCAATGCGGGCGTTATTTGTCCATACTGCAAATATGTTTAAAGTTAATCTGACGTCAAAATCGAC   -918
vir   ATTTTACATAAATAGCAAATGAAGTAACAAAAGGGGCCACAGCAAATGCAAGCCAATTTAACAATTGTTGACGATGTGAAAATGCAAAAATTTGACAAGACTTCGCCTGAGTCTGCGACGCAGCTGCTGCTG   -786
vir   CGGCCTGTTGGGAATTCGTTTTGTATCGGGCACGTAGCTTACCTGAATTAGTTCGTCAAGGCTCTCCTGTAGTGTATTGCCAAGCGTAGTGTTGCGATATAATTGATAAGACATGTTTGAGTCGCTAAATTC   -654
vir   TGGAAATTAAAAACTTATTTTATTGAGCAGCCAAAATGTATTGTAGCGATGTCAGTAAATGGCTATACTAAAATATTGTATCGATAGCTTAGATATTACGATATGCGATATATGAAACTAGCTCGATATATC   -522
vir   GATATATTTTATGCGATGTATTGTCAAATGCTGGCATTAACAATTTCTAACAATTTTAATACGCTTATCAAATTGTATTTTATTTATGTCTATTTTCGAATATATATAAATGTACCAGTAGATAGGTAATTA   -390
vir   AAACGATAAATTGAGAACATACAAATTAAAACTCGATAAGAGTAAGCCAAAAATGGAGTTACGTATCAATTAGCTTTCAAGCGCGCTGTGCTGCTGGCAACAATAGCTGCCAGCTATAGCTCACACGTACGA   -258
vir   AGTGGGCGGCGCTTATCCTTTTTGCTACCCTTGGAGGATACTTTGAAGCGGAAGTTGACTGAGGGTTGCACTGCAGCTGCAACCAATCAGCGCAACGCGTTGCTTGCTTCAGGTGTGGCGCCAATATTTTTT   -126
                                                                                                                                  1→
vir   TATTAACCACAATTTTAAGCACCTTCTTGTCGATTTTAATGCGGCACATAATTTGTGACGCGCAAGATAAGCGCAAACGCTTGACCAAGACAAGCGACCGGTCAGTTTCGAAGTCGGAAGTAACGCCAGTTT      6
                                                                  1                                         ↓G
mel                                                               M  Q  R  H  K  V  E  I  G  S  P  D  G  S  P  -  -  R  -  D  -  L  D  P     25
mel                                                               ----AA-GT--C--AGTTGAGATTGGCTCA-CC-AC-GCTC-CCG------AGGAGTG-CAG-TTAGATC-C
vir   CAGTCACTCGTCGATTATCCAAGTGTTAAGTTTAGCCGGTGCTCGATCCGTGCGCCCAAAATGCTGCAGCAAAACCCATCATCGCCGCTCCATGCGGAAAAGTTTATCAAATCTTCGAATTCCACGCCGACA    138
vir                                                               M  L  Q  Q  N  P  S  S  P  L  H  A  E  K  F  I  K  S  S  N  S  T  P  T     24

mel   I  A  .  .  .  .  .  .  .  .  .  .  .  N  T  -  I  -  S  V  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  S  .  .     56
mel   ATAG-G.................................-ACACT---A-T--AT-CGT------A-----C--G--G-----A--G--T-----A-----G---------T-G--A-----CT-C......
vir   GCGACACCCGCACCCGCATCCGCAACCGCAACCGCACCCATGCACACGGTGCTTGCGACGCCGCAGCGTCCGTCATCTCCGCGCCAATTCTTCGAGCGTCTTTACGGGCATCTCGAGACGCGGACGCACAAC    270
vir   A  T  P  A  P  A  S  A  T  A  T  A  P  M  H  T  V  L  A  T  P  Q  R  P  S  S  P  R  Q  F  F  E  R  L  Y  G  H  L  E  T  R  T  H  N     68
                                    ↓HKPPPC
mel   -  -  N  -  E  -  -  -  -  -  -  -  D  -  -  -  H  -  -  -  -  -  V  -  -  -  -  -  -  -  -  -  -  -  T  -  -  A  A  Y  -  A  -  D    106
mel   TCG--A-AT--G--A--C--C--C--A-----------C--------T--C-----T-----A----TC---------A---------C--C--C----------AC---GC---C-GCA-A---CG-C--C--T
vir   AGTGAGAGCGGCGACATTGATGTGGGCACACACGCCGAGACGCCCTACCAGAGCGACGGAGGCAGCGCATCCTCGCCGGACATTTCGATAAGTGATGAGCGCGTCTCATTAGTCAGTTTTCCGTCATATGAA    402
vir   S  E  S  G  D  I  D  V  G  T  H  A  E  T  P  Y  Q  S  D  G  G  S  A  S  S  P  D  I  S  I  S  D  E  R  V  S  L  V  S  F  P  S  Y  E    112
                        ↓PQHPSQQHQQHHHHHHHPPQLVHQKLSYVSPPPA                                ↓G
mel   F  Y  G  H  A  K  -  -  I  -  A  -  G  A  -  -  P  V  -  -  H  .  .  .  .  .  .  -  P  S  -  -  -  -  F  S  -  -  -  -  -  -           181
mel   --CT---GT-ACG-CAA---C--T-TT--CGC---T-GAGC------TC-TGTCC-G--A-AC--A......--C--C-----CC-AAG---C-----CT-C-G------C-----C------C
vir   TTGCATGCGCCTACGCCGGATTACAGCGCGTACGGCGCCATTGCGAACGCAACTTTACCGCCGGCTCTATCATTTCCGGCCTTTTCGGCTGATCCGCATATTAACGCCGGATTTTCAGCATTTTGTGAGTAC    534
vir   L  H  A  P  T  P  D  Y  S  A  Y  G  A  I  A  N  A  T  L  P  P  A  L  S  F  P  A  F  S  A  D  P  H  I  N  A  G  F  S  A  F  L          154

vir   ATCCGAGCAAGATTAAACACAAAAATCGACCTACAACTCATTTTACAAATATTTAATACGTATCCAACTATTTTAGTTCCAGTAAAGATTGAGAAAGAGTTAGTGAAAAGTTTTTTATAGAAAATTTAGAGC    666
vir   GGAATAGAGATATTTGCAGATCTGGAGTGCACATATATCTATATAGCATATATATTAATTATTGTGGGATTCTTTATAATCGCATATAAAAACACTCTATGGTAATAATATGAAAGCTTTTGTGTTTTGATC    798
vir   AAAACGTTTTAGGCAAAAGTAAATCTAATGTAAAGAGTTTGGTTGGAATACTTGGAAAGAGTCCTTGTTCCTAGAATCTCATCTATATAGCAGTTATATAATATGAAATTCTGTATAATGAAAAACCTGATT    930
vir   GATAAGTCGTTGCCATATAGGAACCTACACCGATCTCTCCTATAGCAGCTAGCTATATGATAAGCTTGTCCCATGTTCATAGAAATTTCCCTATGAGGATCATAATGGAACTACTAGAACTATCAAGTTACC   1062
vir   CTAATATAAAAATAGGTTTCCAACTTTTCGATCAAACTTTCCTATAGCAGCTATATGATATGGCTAGCCGATGTTGATAAGATTTTGCATACATATAGGGCACGTTCTTAAACTTATAAATGCCAAGTTTAG   1194
vir   GTAGGATATCTTCTGCTTTTGGTCCGATCGTTCCTATGGCAACTGTATAATATAGTGGTCCGATCCGCCCAGTTCCGACTGAGAGGCAAGTTCGCATAGAAACAGACGGACCGACTTGGCTGTTGGCACTGA   1326
vir   TCAGGAATATATATACTTTATGTGGTCGAAGTTATAATACCCTCTATAATACCCTACATAACAAGGAATAAAAAAACTTTCTGAATACGATTCGCTCATCCCAATGAAATGAAAGATGTTAAAAAGATTTAC   1458
vir   TAAAATCTTAAATGATAGAAAAACAACGGTCGTAGCTTTAAGATTTGGTACATTACCCAAAAAGCAGTCAAAAATACTTCTAAGGGTGACCCAGAAAATAAGTTTAATTAAAGAAAAAGGGAGATACTCCAG   1590
vir   TTTAACACTTAACCATTATCTAATCATATTTTATAATTGAAATTTCGGCAATTCTTAGTAGAATCAAGAATTTTCATTCGGATATTATAGGGTATTTAATTATTAAAAAACGGCTTTAGATAGATTTGGGAA   1722
vir   ATTAATTGCCTTAATTGAACCTGGAATATTTCTATATACAGCCGATTTTCAACTCACCAACTGTGTTGTATAGTGTATTTATATGGTCGAGCGAAACGTTAAGCATATCAATAGACAAGACACCTTCGATGA   1854
vir   CATGTCATTAAGCTTGTCTTGCAAGCTGACGCCAGGGCATAGATCGGATTGCAGCCTATTCCCTTAACATGTCCGATTGATTGTTGTTGATATAGATGTTGTTCCTTTCAGATCAGTTAAAATTAATTTGGT   1986
vir   TATGGCATAATCGAATTTTGCTAATGTATTTGAATACAATTTGTTGCTACAGTTTGGATTATTTGGGTTTCTTTTTCGCCTCCTCTGCGCTGTCTCGCAGCTGTTTCAAGTTGCGCAGCTCCATGCTCAAAA   2118
vir   ACATACGCAAATTGGGATCTTTGATCTTGGCCAGAACAATCTCACGGGCACGCTCATCGTTGTCCCCCAGCAGCTTGGCCACAGTCTGCGCGTACTCCACCCAGACGCAATTGGGGCAACCGGACATGCAGC   2250
vir   AGGTGGTGGGTTCCGGGGGCAACTGTCGGTTTGCCAGACATTAGACGAGCTATGTTTCTCCATTTGCAACACTTACCACAATATCCGAGGCGTCTTTGTTCGCCAGATGTCTGGCCTGGAAGGCTGCGGCTA   2382
vir   ATTGGATTGTCGCCCGGCCAAGCAGGCAGCGCTGGCAGCCGAAGCACAGAGGAGCACGCATTCTGATTGGATCTGTAATATGGTCAGAAATGGGCAAATGTTAAATGCAACGGGATTTTCACACGCTGGCGC   2514
vir   ATGCAGGTGCAAGAGTTTAACGAAAGTAGGAGCAAAAAAAAAATTAGTTCCGCAGGAACTGCGTAAAAAGTAGCCAGGCTATTAAATATCCTGGCACTTGTCGGCAAATCCTTGATGAAAAAAAAAAAAAGT   2646
vir   ATGCAAAAGCAAAGAAAAGAAATGCTAAAGAAAACCGAAAAGTCGAGAGAGAAAGGGCTTTAAGAAGAAGAAAGCAAAAAAAAAACAAGAAATTAACAAAATATTAAAACGCATTCAGCCGTAAAAATTTTC   2778
vir   ATTTACTTGCGGTAATCTCGCTTAAAGGTCGCGCGGGAGAGAAGGTACGCAAATTGAGAATGGACAATGCCAGAAGAAGATTGATAAGGGAGAGATGATATTACAGGCCACATCGCGATTAATTGGCACCCC   2910
vir   CAATAAATATGCCACTTTGAGAGTTAGCTAAAAAAAAAAAGGAAGAAGCAGAAGAGATGATGAAAAAACGCATTGCTAAAAAAATTAACATTGTAATTGAGCTTAACAGGAAACCTAAAAAACACACACACAC   3042
vir   ACACACACACAAACACAAAGGACACTTGTCGTATCGAATTTCACACGTGATTTAAAAAAAATATATATATGTATGTATATATGTGCATATAAGCCGAAAGCAGAGCTTAAGGCTCTTGTGAGTTAGTCAGGC   3174
vir   GGCGGGTAATCAGGCGAGAACTAGAGATTAGCAACCCCTACCCCCGCACTACCCCCGCTCTACCACCACACTCACACAAAAAGTCGCGCACTAGCGGAAACTTTACAGCTTCGCAAGAGGCAAGGCAGCAGA   3306
vir   GAGTGAGAGTATCAGAGCGAGAGAGACAGAGAGAGAGAGAGAGAGCGTGGAGAGTGGAGAGACGGACTGGAACAATTAAGCGGTAATCAGCGTACGCGAACAGACAATGGGAGTTGAGTACGCATCGTGTGA   3438
vir   GCGGGCGTGGCTGAATATTGTTCGCTCGAATGACGTCTACGTCACAGCGCAATGCAAACAAATATTCATTAGGATTGCGCTACAACGAAGTGCAGATGCCCTTACAAACGAATTGATCAAACTGTTGTCGAT   3570
vir   ACTCTCTTCGCTTAAAGCGCTGCGATACTTATTTTCGAAATTTACAAAAAATATATACAAAATATTTTCAAAATTCCATTAAAATGAATCATTTCGAATACATTTTAATATTTGTACATTATTGTCTTTTGC   3702
vir   TGCGGCTCTCATTTTCGAAGTCTACAAAACACTAAAATCTTTTTTTGTATTCAATTAAACGAATTATATATGTTTAATTATGTTCTTTGCGATCCTCATTTTCGAAAGATACAAAATACTAAAACATTTGTT   3834
vir   GTATTAAATCAAAATTCATTGTTTTTTGTTTTTGTAATATGTATTAAATTATCTATTGTTGTGAAAAATTGAAGTACGACTCGACGTTTCGATAGATATTAAAAAATGTTTAGGACAAAATAAAATTAAAAA   3966
vir   TGTTAATATGAGCATATTGAAAATTCGCTCTAAAGAAACCTTCGACCTAAACAAATGCTTCATTGAAACCAAATATGTTGTAGATTTTTTTTTCTAAGTCAACGTAATATATATATATTTTTTTAATATGTTTTT   4098
vir   ATATATTTTGACGAAAGTTTTCAAGCTGAAATTGTCCCGAAATTGAATACAACCAATGGCCTCAAAAAAACATTTATTTTAAAATAGTTTCTATTATTTTAATATATAAAATGTATTCTCCATTTATTTCTT   4230
vir   AAGCAAAAAGTGAGAAATGAATTATTAGCAAACTTGTATGCGAAATTAAAAATGGCGAACCGTATGAAAAAACTAATAATACCCTTTCAAGAAGTGTATAGTTATGATGACTGTTCGCAGCATAACAAAAAT   4362
vir   AACAATGGGAGGATGGGAGACAGCTGTTGGTATTGTTGAAAGTTTCCCGGCTCAGATTGCCCAAAGTATGAATTGCTGGGACGCAGGATGGGAGTGGAGAGGGGGGTAAAGGGGGCACGGCACTGATGGACA   4494
vir   GTCTAGGAACACGTGTAAAATGCATTTTAATGTGCGCCATAAACGCGATCACAAACATCGTTTAGCAGTAGAAAAGGGGCGGGGTGGGCAACAGCCTGGTTAATTATTTGTTTAACTTGTTCGCACGCAAGG   4626
vir   ACACTAAGCCGTACGCTGATATATGTGCACGTGCAGGCGGGTCTGAGGGTAATTTTATGGTTTATTAATGTGGCATTTGGTACGCTCTAACATCTCCATTGTGTGTGCGTGATATGCCCAATGATATGCCTA   4758

mel                                                                  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -     204
mel                                                             -------CA-G--A-----------A-A--CC-C--C---A-G-----T--------A-----G--G--
vir   ATCTTAAGCCATTGTTGTTGTGGTCTATGCAACAATTTGATGATGTGCTCTCGATTTGGCAGTGGCTCGACGTCGCCGCAAGGAGGGACGCCAGAGACGGCAGCGCACCACGTTCAGCACGGAGCAAACACT   4890
vir                                                                  A  R  R  R  R  K  E  G  R  Q  R  R  Q  R  T  T  F  S  T  E  Q  T  L     177

mel   -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  T  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -    248
mel   T--G----------------------------C--C--C--G-----------C-----------A--------TC--A-C--A--G--------G--------C--G-----CA-G--------C-----
vir   GCGACTGGAGGTGGAGTTCCATCGGAACGAGTATATATCGAGAAGTCGTCGCTTTGAGCTGGCCGAGACGCTGCGCTTGTCGGAGACACAAATCAAAATCTGGTTTCAAAATCGTCGCGCCAAGGATAAACG   5022
vir   R  L  E  V  E  F  H  R  N  E  Y  I  S  R  S  R  R  F  E  L  A  E  T  L  R  L  S  E  T  Q  I  K  I  W  F  Q  N  R  R  A  K  D  K  R    221

mel    -  -  -  -  -  -  -  -  -  -  -                                                                                                       259
mel   ---T--A-----T--G-----------C-----
vir   CATCGAGAAGGCGCAAATCGATCAGCATTACAGGTAAGCCAACAGAGTGGAGCAGGAGTAGCAATTCAGGGAAATGTCACAATATAATTGATGTACTCCCATTAAATTAACACGAAATTAATTTAACGACAT   5154
vir    I  E  K  A  Q  I  D  Q  H  Y  R                                                                                                       232
mel                                                                                                                            -  -  -  -    263
mel                                                                                                                            A----------
vir   TTACGGCCCAATTAGACAAATGTCTGCGCGTCCGGCTTTGGGTTTCGGGCTCGGGCTCGAGCTCGGGCTCGGTTAATGTTCGACATTTTTTATGTAATGCACGATACTATCAATTTGCAGGAACTTTGTCGT   5286
vir                                                                                                                            N  F  V  V    236
                                                                                                                     ↓AA
mel    -  -  -  -  -  -  S  -  -  -  -  A  A  T  T  M  P  P  G  -  V  T  G  G  V  -  V  G  V  G  L  N  .  .  .  -  -  -  -  -  -  -  -  P    306
mel   ------------------T-T--A-----T---G-TGCCACGA-C-T-C-CC-CGGA--CGT--CTGGG-GC-T----GT--GC-T--G-C-AAAC........--------C-------------GCC
vir   GGCCAATGGTTTTATGAGCACCATTATGGGCCAGACGAGTTACGCGACGTCTGCGCCGGGGACCAGCTCTGCGGCGGCGACGGCGGCGGCCATGCTAGGCAGCAGCTACTATGCAGCGACGCCGGCGCTCGT   5418
vir    A  N  G  F  M  S  T  I  M  G  Q  T  S  Y  A  T  S  A  P  G  T  S  S  A  A  A  T  A  A  A  M  L  G  S  S  Y  Y  A  A  T  P  A  L  V    280

mel    K  D  N  T  Q  .  .  .  .  .  .  -  A  -  F  I  -  -  D  D  -  F  -  R  -  -  -  .  .  .  .  .  .  .  -  K  -  -  -  -  -  -  -  -    337
mel   CAA-GAT-ACACGCAG....................----CC----TCA-C-----C--T-AC--GTT---G-GC--G------.......................---A-A--A--A--A-------------
vir   GCCGTCAACGCTCGTAAATAATGCGCATCGTGCGGATGTTAATTATGTTGACATTGAGGCACAACACCAACAACAACAACAACACCACCAAGTGCATCCTGACCAACAGCAGCAGCAGCAGCAGCAGCA   5550
vir    P  S  T  L  V  N  N  A  H  R  A  D  V  N  Y  V  D  I  E  A  Q  H  Q  Q  Q  Q  Q  H  H  Q  V  H  P  D  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  324

mel    -  -  R  .  .  .  E  T  T  -  P  -  -  I  -  *                                                                                      349
mel   ------TC-A.........GAAA-AA-------C-----CATT--C
vir   GCGACGCAGCCGTTTTGTGCCTCCGCCAACGCTTATTAATTCGTGTTAAGTGAATTCTTGTTTTAAGCTGATTTAAATGCGTCAAATAAACTAAGTTAAATGAAAAAAAACGTTTTTAGCTTTTAAATAGTC   5682
vir    R  R  S  R  F  V  P  P  P  T  L  I  N  S  C  *                                                                                      339

vir   AAAACTTGCATCACTGACGGCTGCTTATAACTATCTATGACAACAGTTTAATTGGCGGTCTAACGCGCACTGTGACACTTACCTTAGTTCACTGCATTTGTTGTTGCAGCAACTGCTGGGTTTTTTTTTTT   5814
vir   TTTGGTTTTTGTTTTTAATCCAACTGAGCTCTTAATTTAGTTAAATATTATACAATTTGACTAAACAACGTAATAATAACAATAACAAACCAGGGCTGCCAAAGGTCTGTGCGTCAAAATAATATATCGAAG   5946
vir   GCAAGCAACTTATCGATAGTCAGGTAGTACGATGTATATTGCACATCCCTGAAATTTTCTTCGCGTGCGTGTCTGACGAGCTACAGCTCTTTGCATTTTTTTTTTTGTAAATTTAATTAAACGTGCGCCAAA   6078
vir   TATGGCAAACACACCGCAAGTAGTCAAATTAGAGGATAACGCCGAAGAGCCCGAGATATTTGAAATCAAAGACGAAGATGTAAGTATCAATAACAATTGCGGCGTCCGCACTACTCATTCTACTGCGTGCAC   6210
vir   CTTGTGGTACATAAAAAAAAAAAAAACCTGCGCTGTTTCTGACAAACGAACCCATGCATATGCACATGCACACATGCGAATATATATTTGTGTTTCATTTGGCGTGCAGGACATTTGCGAATTGGAGCACTT   6342
vir   GCGCAAGCTGTTTATTGGCGGCCTGGCGCCCTACACCACAGAGGAGGGCCTCAAAGTGTTCTATT                                                                      6400
```

FIG. 1. DNA sequence of the *D. virilis* rough gene and comparison with the *D. melanogaster* protein-coding sequence. The nucleotide sequence and predicted amino acid sequence from the *D. virilis* (vir) genomic fragment and the *D. melanogaster* (mel) cDNA are shown. Positions of nucleotide or amino acid identity are indicated with a dash. Positions that are absent in the *D. melanogaster* sequence are indicated with a period. Insertions in the *D. melanogaster* sequence are shown above the aligned amino acid sequence, with the exact position indicated by a vertical arrow. The presumed start site of transcription for the *D. virilis* gene (by homology to *D. melanogaster*) is indicated as nucleotide 1→. The homeodomain, as well as highly conserved DNA sequences located in the introns, is underlined.

located in the first intron (2773 bp) of the *D. melanogaster* rough gene shows significant homology (>80%) with the *D. virilis* sequence (underlined in Fig. 1).

Comparison of Protein-Coding Regions. The coding regions of the *D. virilis* and *D. melanogaster* rough genes have been aligned such that amino acid identities are optimized (Fig. 1).

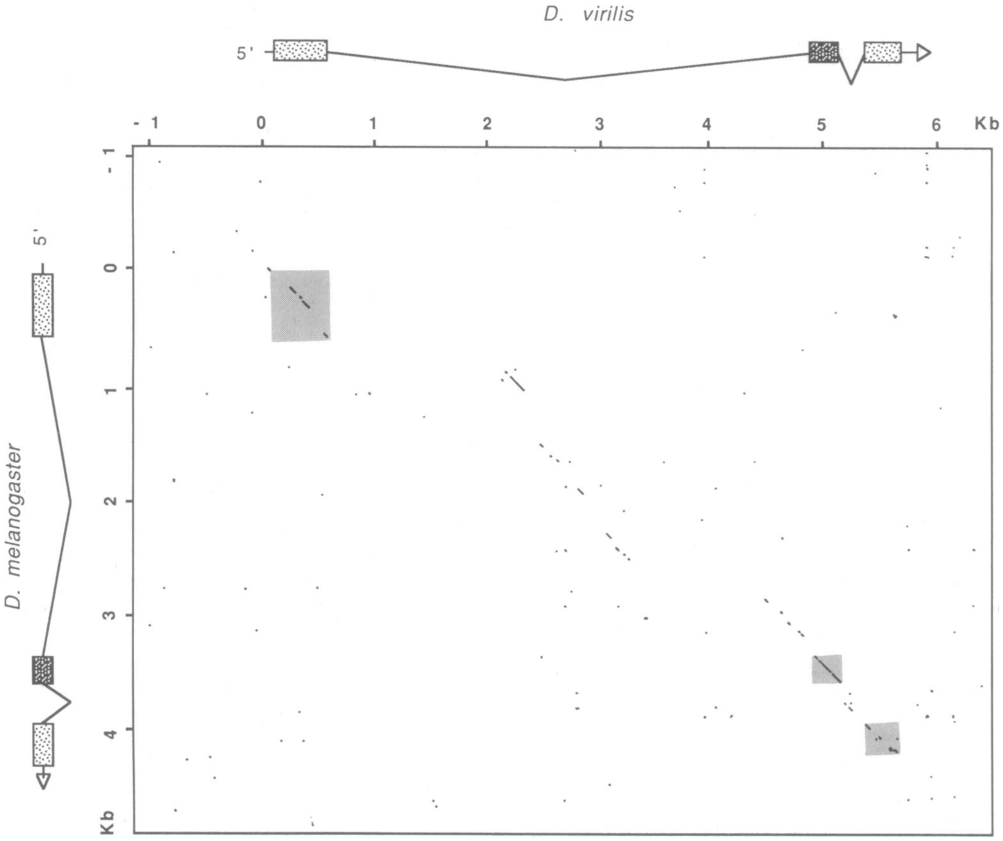

FIG. 2. Dot-matrix comparison of the *D. melanogaster* and *D. virilis* genomic rough sequences. The structure of the transcription units (presumed for the *D. virilis* gene) is shown; exons are represented as shaded boxes; introns and 5′ and 3′ untranslated regions are shown as lines. Position 0 corresponds to the transcription initiation site (3). Protein-coding regions are shaded in the dot-matrix graph. The genomic fragment of the *D. melanogaster* rough gene used in the analysis contains all the sequences necessary for function. The criteria were such that one dot corresponds to a match of 16 out of 21 nucleotides.

The first possible translational start signal in the appropriate reading frame of the *D. virilis* gene is indicated as amino acid 1. The sequence preceding this methionine codon (GCCCAAA) is a 6/7 match to the *Drosophila* consensus translational initiation signal (12). The predicted *D. virilis* rough protein is 339 amino acids long, which is 10 amino acids shorter than the *D. melanogaster* protein. Although the overall identity at the amino acid level is 60%, domains of high conservation are found interspersed among regions displaying little or no similarity. A schematic representation of the comparison between the *D. melanogaster* and *D. virilis* rough proteins is shown in Fig. 3. The most striking conservation is observed in the homeodomain and the regions adjacent to it. Among the 60 amino acids that define the homeodomain, only one conservative, serine-for-threonine substitution is observed. Moreover, 14 amino acids located immediately N-terminal and 17 amino acids immediately C-terminal of the homeodomain are identical in both species. Comparison of the *D. melanogaster* and *D. virilis* engrailed (*en*) genes also shows remarkable conservation in the homeodomain and the surrounding regions (13).

Another salient structural feature of the *D. melanogaster* rough protein is the presence of two regions rich in glutamine and histidine. These regions, which are encoded by CAX (where X is A, C, G, or T) repeats in the DNA, are often found in genes involved in important developmental processes (14, 15). The CAX repeat (17 glutamines/histidines in a 23-amino acid stretch) found in the first exon of the *D. melanogaster* rough gene is absent in *D. virilis*. The CAX repeat located in the third exon is retained; it is, however, 7 amino acids longer in *D. virilis* (21 glutamines/histidines among 24 amino acids). Regions of simple repeated sequences, such as CAX repeats, show a much higher local rate of sequence divergence than adjacent unique sequences (13, 16, 17); it has been suggested that this high local divergence could provide a mechanism for the evolution of regulatory patterns (17). Although the function of these repeats is unclear, glutamine-rich regions have been shown to be important for the transcriptional activity of the mammalian Sp1 factor (18). Outside of the conserved domains described above, several shorter regions of amino acid sequence conservation are observed, particularly in the first exon (Figs. 1 and 3). The overall divergence observed for the *D. melanogaster* and *D. virilis* rough proteins (≈40%) is significantly higher than that described for other genes, such as *en* (13), hunchback (17), period (19), and parts of Ultrabithorax (*Ubx*; ref. 20), where only about 20% divergence is observed.

Comparison of Promoter and Intron Sequences. Interspecies comparisons of non-protein-coding sequences have aided in the identification of important cis-regulatory DNA sequences in several *Drosophila* genes (13, 20–24). In comparing the sequences of ≈1 kb of DNA located upstream of

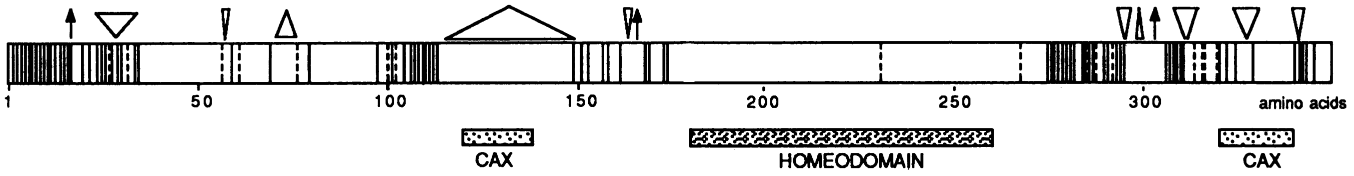

FIG. 3. Schematic representation of the divergence of the predicted *D. melanogaster* and *D. virilis* rough proteins. The structure of the *D. melanogaster* protein is diagramed. Nonconservative amino acid substitutions are marked with vertical lines. Conservative changes (E/D, S/T, Q/N, Y/F, V/L/A/I) are represented with broken vertical lines. Insertions in the *D. virilis* protein are represented by ▽ (width of triangle corresponds to number of amino acids). Deletions in the *D. virilis* protein are shown as ↑ (single amino acid) or △ (several amino acids).

the transcription start sites of the *D. melanogaster* and *D. virilis* rough genes, we have found areas of sequence conservation in the 350 bp immediately 5′ of the start site, where overall conservation was 43%. Several short stretches of sequence identity (10–16 bp) are found amid nonconserved DNA. In the 600 nucleotides located upstream of this conserved region, only one homology (13/14 bp) of more than 4 consecutive nucleotides was detected. These homologous regions are candidate recognition elements for trans-acting factors involved in transcriptional regulation of the rough gene.

Although the overall sequence conservation in the introns is low, comparative analyses yielded a surprising number of highly similar regions, which are interspersed with completely divergent DNA. Nine regions, 40–140 bp in length, displaying between 75% and 89% sequence identity were identified. Eight of these conserved regions are located in the first intron (underlined in Fig. 1), and, although the spacing between them is variable, their order is the same in both species. One of these regions, located between nucleotides 2122 and 2268 in the *D. virilis* sequence, is particularly striking; a stretch of 146 consecutive nucleotides is 81% conserved. Although we have not completely ruled out the possibility that some of these conserved sequences are part of another transcription unit, we have not found any long open reading frames or good consensus splicing signals flanking them. However, we have found that sequences located in the first intron are critical for proper rough expression in *D. melanogaster* (unpublished data), which suggests that some of the conserved regions between *D. melanogaster* and *D. virilis* found in this intron may be important cis-regulatory elements. The enhancers of several genes have been found to be located in introns (for example, see refs. 25–28). They are, however, usually composed of a series of short (7–20 bp) and often irregularly spaced DNA sequence elements that are recognized by regulatory proteins (for recent reviews see refs. 29 and 30). The homologies found in the first intron of the rough gene are significantly longer than predicted for protein–DNA binding sites and, perhaps more surprisingly, are almost completely devoid of gaps. Similar observations have been reported previously for other *Drosophila* genes, such as the *Gart* locus (31), *en* (13), and *Ubx* (20).

Functional Analysis of the *D. virilis* Rough Gene. *D. melanogaster* bearing mutations in the rough gene that result in complete loss of function are viable, and the eye phenotype can be rescued by introducing a wild-type copy of the gene by germ-line transformation. This allows us to test whether the conserved sequences observed in the two rough genes reflect an underlying conservation of gene function. The *D. virilis* rough gene was introduced into the genome of *D. melanogaster* carrying the *ro*^{1} mutation (32) by *P*-element-mediated transformation. Four independent transformant fly stocks were established and the ability to rescue the *ro*^{1} phenotype was assessed by analyzing ommatidial structure in tangential sections of adult eyes (Fig. 4). One of the transformant lines, P[virro]4, showed complete rescue, defined as the absence of mutant ommatidia, when one copy of the *D. virilis* gene was present in the genome (Fig. 4*F*). In another line, P[virro]1, complete rescue was obtained only when two copies of the transposon were present (Fig. 4*D*). In the remaining two transgenic lines the insertion caused a recessive lethal mutation. However, all transformant lines displayed at least partial rescue of the *ro*^{1} phenotype, manifested as a mixture of wild-type and mutant ommatidia, when one copy of the *D. virilis* rough gene was present in the genome (Fig. 4 *C* and *E*). The *D. melanogaster* rough gene (containing ≈1 kb of both 5′ and 3′ untranslated sequences) can rescue the *ro*^{1} phenotype more efficiently than the *D. virilis* gene; in four of the five transformant lines established for the *D. melanogaster* gene, complete rescue of the *ro*^{1} phenotype was observed when only one copy of the transposon was present (data not shown). This suggests that the *D. virilis* protein is less functional, or perhaps expressed at lower levels, than the *D. melanogaster* homologue. Nevertheless, our results indicate that the *D. virilis* rough protein is expressed when introduced into *D. melanogaster* and that it can, to a large degree, functionally replace the *D. melanogaster* rough protein.

The expression pattern of rough protein in the eye imaginal disc of third-instar *D. melanogaster* larvae is complex (4). Initially, rough is broadly expressed in the morphogenetic furrow, and later, expression is restricted to a subset of developing photoreceptor cells (Fig. 5*A*). Monoclonal and polyclonal antibodies generated against the *D. melanogaster* rough protein were found to cross-react with the *D. virilis* protein. Fig. 5*B* shows a *D. virilis* eye-antennal disc complex stained with a monoclonal antibody (MAbrol; ref. 4). We found that the rough expression pattern in *D. virilis* is very similar, if not identical, to that observed in *D. melanogaster* (compare Fig. 5 *A* and *B*). That the *D. virilis* gene was able to rescue the *ro*^{1} phenotype already suggested that at least some of its regulation was properly maintained in *D. melanogaster*. To test this directly, we used MAbrol to stain eye discs from *ro*^{1} third-instar *D. melanogaster* larvae bearing P[virro]4. The expression pattern was found to be indistinguishable from that of the wild-type *D. melanogaster* rough protein (compare Fig. 5 *A* and *D*). The intensity of the staining, however, varied among different transformant lines; weaker disc staining was observed in lines showing poor

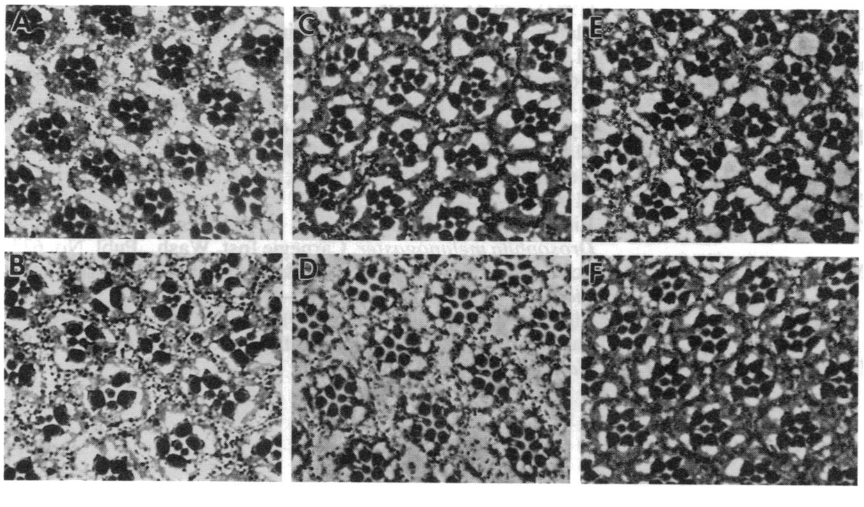

FIG. 4. Comparison of eye morphologies of *D. melanogaster* wild-type flies, *ro*^{1} flies, and *ro*^{1} flies transformed with the *D. virilis* rough gene. (*A*) Tangential 1-μm plastic section through a wild-type eye. A repeating array of precisely structured ommatidia is observed. (*B*) Section through a *ro*^{1} eye. The ommatidial array is disrupted and the number of photoreceptor cells per ommatidium is variable. (*C* and *D*) Sections through the eyes of transformant flies P[virro]1 containing either one or two copies of the *D. virilis* rough gene, respectively, in a *ro*^{1} background. In the presence of one copy of the transposon only 10% of the ommatidia are rescued (*C*). In the presence of two copies, the eye is completely wild-type (*D*). (*E* and *F*) Sections through eyes of transformants P[virro]2 and P[virro]4; 90% and 100% wild-type ommatidia are observed, respectively, when one copy of the transposon is present in a *ro*^{1} background. (×250.)

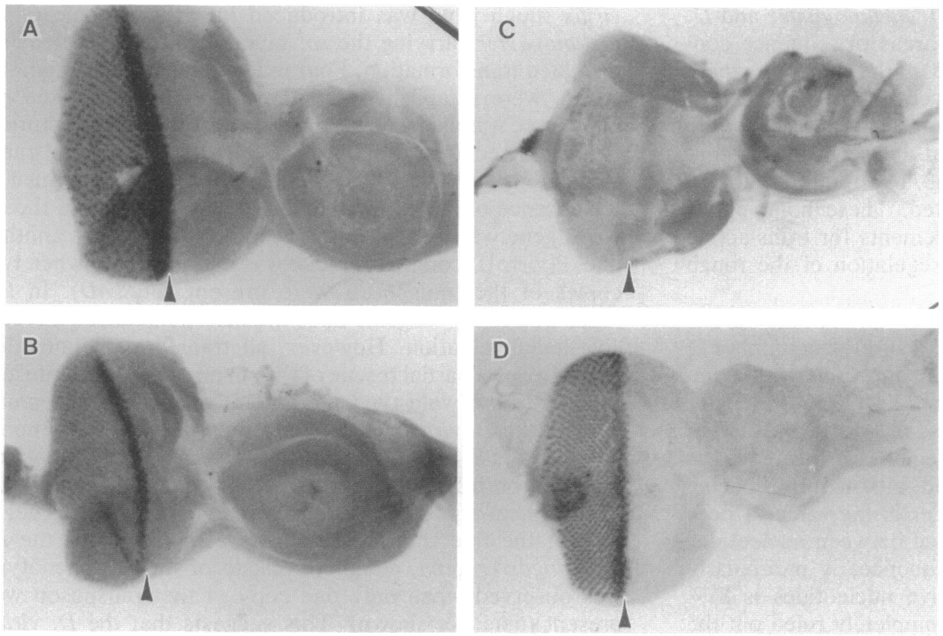

FIG. 5. Pattern of rough protein expression in third-instar larval eye imaginal discs. Eye-antennal imaginal disc complexes were stained with a monoclonal antibody (MAbrol; ref. 4) generated against the *D. melanogaster* rough protein. (*A*) Eye-antennal disc derived from wild-type *D. melanogaster*. Strong staining is observed in the morphogenetic furrow (arrowhead) and in a subset of differentiating photoreceptor cell nuclei posterior to the furrow. (*B*) Eye-antennal disc from a *D. virilis* larva. The staining pattern is essentially identical to what is observed in *A*. (*C*) Eye disc from a ro^1 larva. No specific staining is observed. (*D*) Eye-antennal disc derived from a larva carrying P[virro]4 in a ro^1 background. The expression pattern is the same as in wild type. Anterior is to the left. (×200.)

rescue of the adult phenotype (data not shown). These quantitative effects are likely to be due to the particular genomic environment of each transposon. Taken together, these results show that the cis-regulatory elements that control rough expression in *D. virilis* are sufficiently conserved to be recognized by the *D. melanogaster* transcriptional machinery to impart proper temporal and spatial regulation.

CONCLUDING REMARKS

To help define important functional elements we have compared the rough genes from two distantly related *Drosophila* species. Although the overall conservation of the predicted proteins is only 60%, the homeodomain and its immediately surrounding regions are almost identical. Homeodomain proteins have been shown to possess specific DNA-binding activity (for review see refs. 33 and 34). The very high conservation of the rough homeodomain over ≈60 million years suggests that it is critical for proper function, perhaps conferring the specificity of target site recognition and interaction with other regulatory proteins.

Conserved sequence elements were also identified in both the 5′ promoter region and the introns. These sequences are good candidates for cis-regulatory elements. The ability of the *D. virilis* rough gene to properly function during *D. melanogaster* eye development raises our confidence about the significance of these conserved elements.

We thank Mark Fortini for help with the microinjections and Joe Heilig, Kevin Moses, Mike Ellis, Iswar Hariharan, and Mark Fortini for comments on the manuscript. U.H. was supported by a Jane Coffin Childs Memorial Fund postdoctoral fellowship.

1. Tomlinson, A. (1988) *Development* **104**, 183–193.
2. Ready, D. F. (1989) *Trends Neurosci.* **12**, 102–110.
3. Tomlinson, A., Kimmel, B. E. & Rubin, G. M. (1988) *Cell* **55**, 771–784.
4. Kimmel, B. E., Heberlein, U. & Rubin, G. M. (1990) *Genes Dev.* **4**, 712–727.
5. Basler, K., Yen, D., Tomlinson, A. & Hafen, E. (1990) *Genes Dev.* **4**, 728–739.
6. Beverly, S. M. & Wilson, A. C. (1984) *J. Mol. Evol.* **21**, 1–13.
7. Frischaut, A., Lehrach, H., Poustka, A. & Murray, N. (1984) *J. Mol. Biol.* **170**, 827–842.
8. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
9. Mismer, D. & Rubin, G. M. (1987) *Genetics* **116**, 565–578.
10. Spradling, A. C. & Rubin, G. M. (1982) *Science* **218**, 341–347.
11. Tomlinson, A. & Ready, D. F. (1987) *Dev. Biol.* **123**, 264–275.
12. Cavener, D. R. (1987) *Nucleic Acids Res.* **15**, 1353–1361.
13. Kassis, J. A., Poole, S. J., Wright, D. K. & O'Farrell, P. H. (1986) *EMBO J.* **5**, 3583–3589.
14. Wharton, K. A., Yedvobnick, B., Finnerty, V. G. & Artavanis-Tsakonas, S. (1985) *Cell* **40**, 55–62.
15. Laughon, A., Carrol, S. B., Storfer, F. A., Riley, P. D. & Scott, M. P. (1985) *Cold Spring Harbor Symp. Quant. Biol.* **50**, 253–262.
16. Tautz, D., Trick, M. & Dover, G. A. (1986) *Nature (London)* **322**, 652–656.
17. Treier, M., Pfeifle, C. & Tautz, D. (1989) *EMBO J.* **8**, 1517–1525.
18. Courey, A. & Tjian, R. (1988) *Cell* **55**, 887–898.
19. Colot, H. V., Hall, J. C. & Rosbash, M. (1988) *EMBO J.* **7**, 3929–3937.
20. Wilde, C. D. & Akam, M. (1987) *EMBO J.* **6**, 1393–1401.
21. Blackman, R. K. & Meselson, M. (1986) *J. Mol. Biol.* **188**, 499–515.
22. Bray, S. J. & Hirsh, J. (1986) *EMBO J.* **5**, 2305–2311.
23. Kassis, J. A., Desplan, C., Wright, D. K. & O'Farrell, P. H. (1989) *Mol. Cell. Biol.* **9**, 4304–4311.
24. Fortini, M. & Rubin, G. M. (1990) *Genes Dev.* **4**, 444–463.
25. Banerji, J., Olson, L. & Schaffner, W. (1983) *Cell* **33**, 729–740.
26. Gillies, S. D., Morrison, S. L., Oi, V. T. & Tonegawa, S. (1983) *Cell* **33**, 717–728.
27. Slater, E. P., Rabenau, O., Karin, M., Baxter, J. D. & Beato, M. (1985) *Mol. Cell. Biol.* **5**, 2984–2992.
28. Bowtell, D. D. L., Kimmel, B. E., Simon, M. A. & Rubin, G. M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 6245–6249.
29. Dynan, W. S. (1989) *Cell* **58**, 1–4.
30. Mitchell, P. J. & Tjian, R. (1989) *Science* **245**, 371–378.
31. Henikoff, S. & Eghtedarzadeh, M. K. (1987) *Genetics* **117**, 711–725.
32. Lindsley, D. L. & Grell, E. H. (1986) *Genetic Variations of Drosophila melanogaster*, Carnegie Inst. Wash., Publ. No. 627.
33. Scott, M. P., Tamkun, J. W. & Hartzell, G. W., III (1989) *Biochim. Biophys. Acta* **989**, 25–48.
34. Biggin, M. D. & Tjian, R. (1989) *Trends Genet.* **5**, 377–383.
35. Saint, R., Kalionis, B., Lockett, T. J. & Elizur, A. (1988) *Nature (London)* **334**, 151–154.