Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 6667–6671, November 1984
Biochemistry

Genomic structure and possible retroviral origin of the chicken *CR1* repetitive DNA sequence family

(long terminal repeat/transposable elements/*Alu* sequences/chromatin structure/genomic evolution)

WILLIAM E. STUMPH*, CLAGUE P. HODGSON†, MING-JER TSAI†‡, AND BERT W. O'MALLEY†

*Department of Chemistry, San Diego State University, San Diego, CA 92182; and †Department of Cell Biology, Baylor College of Medicine, Houston, TX 77030

Communicated by James Bonner, July 13, 1984

ABSTRACT We have analyzed the sequence and structure of three *CR1* family repetitive elements found in the region adjoining the 3′ end of a chicken calmodulin gene. Members of this family are ≈300 base pairs long and are dispersed throughout the chicken genome. The present data, when taken together with that from four *CR1*s sequenced previously, reveal that the *CR1* family has an overall structure possessing several features associated with the long terminal repeats of avian retroviruses. This finding implies that a retroviral mechanism may be responsible for the dispersion of *CR1* sequences throughout the chicken genome. The seven different *CR1* repeats that have been analyzed exist at defined locations in the chicken genome relative to nearby structural genes. A directional polarity has been assigned to the *CR1* family based upon limited sequence homology to mammalian *Alu*-type sequences. Interestingly, whether present in 5′- or 3′-flanking DNA, the *CR1* sequences have an inverse orientation such that they all "point toward" the nearby structural genes. This is consistent with the previously proposed concept that chicken *CR1* sequences may be involved in defining the boundaries of active chromosomal domains of gene expression.

A significant portion of the genomes of higher organisms consists of moderately repetitive DNA sequences dispersed among regions of single-copy DNA (1). We have previously identified and characterized a family of short dispersed repetitive DNA sequences in the chicken genome, and this family of repeats has been named the *CR1* family (2, 3). The first *CR1* family sequence that was characterized lies ≈2 kilobase pairs (kb) upstream from a chicken *U1* RNA gene (2). Molecular cloning methods have revealed that members of the *CR1* family also exist at defined locations both upstream and downstream of the chicken *X–Y*–ovalbumin multigene cluster (3). Interestingly, the *CR1* sequences are found at locations in the DNA where there is a change in oviduct chromatin structure from a DNase I-sensitive conformation to a relatively DNase I-resistant conformation (3). It is not yet known whether this genomic arrangement represents a fortuitous occurrence or if it actually reflects a functional role of the *CR1* sequences in determining or regulating chromatin structure.

Because of this potentially important role of the *CR1* family in the chicken genome, we wished to define the nucleotide sequence and overall structure of the *CR1* family in still greater detail. To do this, we have now sequenced three additional closely related family members. These three new family members lie in the vicinity of a chicken structural gene that codes for a calmodulin-like protein. In this report, we present the sequences of these three *CR1* family members as well as an updated consensus sequence and a structural model for the *CR1* family. Interestingly, the overall structure of *CR1*s resembles that of avian retroviral long terminal repeats (LTRs). These structural features include 5′ and 3′ termini ending in "T-G...T-T-C-A" and regions adjoining the termini that are similar to retroviral primer binding sites. In addition, several features suggest that a pair of *CR1*s upstream of *X–Y*–ovalbumin gene cluster may have entered the chicken genome at that location as the LTRs of an intact retrovirus-like element.

MATERIALS AND METHODS

DNA Clones and Hybridization Probes. The clone CL10 contains a gene, described by Stein *et al.* (4), that codes for a calmodulin-like protein. It was isolated from a partial *Alu* I/*Hae* III library of chicken DNA cloned in the λ phage vector Charon 4A. The chicken DNA insert is flanked by artificial *Eco*RI sites. Subclones of the 2.8-, 11.5-, and 2.5-kb *Eco*RI fragments of CL10 in the plasmid vector pBR322 were available from previous work (4).

The *CR1* sequence probe used for Southern hybridizations to localize the *CR1* sequences in CL10 and its subclones has been thoroughly described (2). It contains the *CR1* sequence known as *CR1U1a*, and it consists of a 390-base-pair (bp) *Eco*RI–*Hin*fI fragment that lies ≈2 kb 5′ of a previously cloned chicken *U1* RNA gene (2, 5).

Procedures. Nick-translations and Southern hybridizations were carried out as described (3). DNA sequencing was carried out by using the chemical degradation method of Maxam and Gilbert (6) as described by Catterall *et al.* (7). Nucleotide sequence comparisons were made by using the programs of the Stanford Molgen group (Palo Alto, CA).

RESULTS

Location of *CR1* Sequences 3′ of a Calmodulin Gene. A restriction map of the clone CL10 is presented in Fig. 1. Coding sequences for a calmodulin-related protein are present in the region designated by the hatched box. This gene has been designated the *cCM1* gene (4). RNA transfer blot experiments indicate that an RNA product of this gene is expressed with relative tissue specificity in muscle cells (4). This gene has no importance to the present study other than as a genomic location marker for a new group of *CR1* sequences.

Restriction endonuclease mapping and Southern blot analysis revealed the presence of three *CR1* related sequences in the 3′-flanking region of clone CL10 (Fig. 1). Nucleotide sequencing was carried out on a total of 2184 bp containing and flanking these three *CR1* sequences.

Nucleotide Sequences of *CR1* Family Members. In Fig. 2, the three *CR1CM* sequences are compared to each other and to four *CR1* family members from other regions of the chicken genome whose sequences had been determined in earlier

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.

Abbreviations: kb, kilobase pair(s); bp, base pair(s); LTR, long terminal repeat; RAV, Rous-associated virus.

‡To whom reprint requests should be addressed.

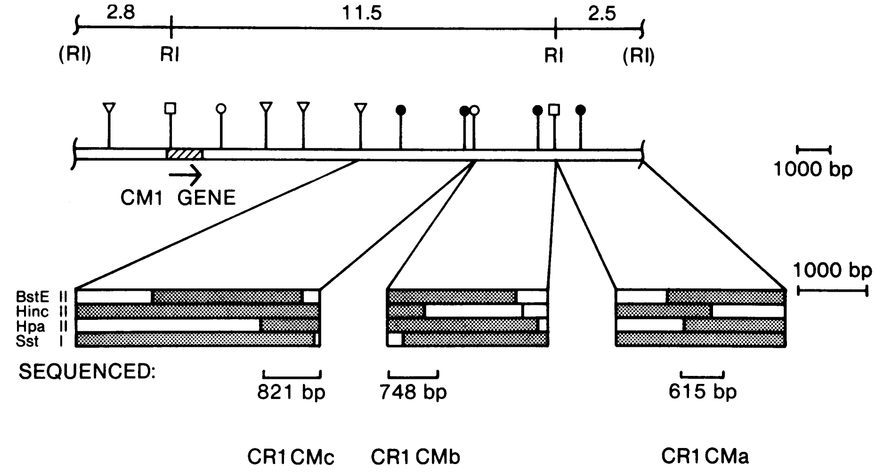

FIG. 1. Location of *CR1* sequences 3′ of a chicken calmodulin gene. A restriction map of the λ phage clone CL10 containing the *cCM1* gene is shown in the center of the figure. The location of the *cCM1* gene is shown by the hatched box (4). The arrow shows the direction of transcription. Three fragments hybridized to a *CR1* sequence probe: the 3.6-kb *Bgl* I–*Cla* I fragment, the 2.4-kb *Cla* I–*Eco*RI fragment, and the 2.5-kb *Eco*RI–(*Eco*RI) fragment. These three fragments were isolated and mapped in greater detail, as shown in the lower part of the figure. Subfragments that hybridized to the *CR1* sequence probe are depicted in each case by the shaded areas. These results defined the locations of three *CR1* sequences (*CR1CMc*, *CR1CMb*, and *CR1CMa*) in the clone. Complete sequence data were obtained from the three regions shown. Restriction enzymes: ▽, *Bgl* I; □, *Eco*RI; ○, *Cla* I; ●, *Bst*EII.

work (2, 3). In the upper line an updated consensus sequence is shown based upon the complete sequence data from all seven of the *CR1* family members. A polarity has already been assigned to the *CR1* family based upon limited sequence homology to the human *Alu* and mouse *B1* repetitive DNA sequence families (2). When this polarity is adhered to, it is found that the three sequences *CR1CMa*, *CR1CMb*, and *CR1CMc* all have the same orientation with respect to each other, but an orientation which is opposite to that of the direction of transcription of the *cCM1* gene. Thus, the sequences of *CR1CMa*, *CR1CMb*, and *CR1CMc* (as shown in Fig. 2) are taken from the DNA strand that runs 5′ to 3′ from right to left in Fig. 1.

When the various *CR1* sequences are examined and compared, it is discovered that the homologies between individual members of the family do not begin at exactly the same position. As a consequence, the beginning of homology to the consensus sequence varies from one individual family member to another. For example, if the sequence homology between *CR1CMc* and *CR1CMb* is used to define the first nucleotide of the consensus sequence (as shown in Fig. 2), the *CR1CMa* homology does not begin until about position 129 of the consensus sequence. The *CR1U1a* homology starts at position 149. The three *CR1* sequences of the ovalbumin domain (*CR1OVb*, *CR1OVc*, *CR1OVa*) exhibit a degree of homology beginning at positions 182 and 193 of the consensus sequence.

Similarity of *CR1* Sequences to Avian Retroviral LTRs. Comparison of these seven *CR1* sequences reveals the composite structure as shown in Figs. 2 and 3. *CR1CMc* and *CR1CMb* have an extended homology to the left but converge with the other sequences in the region of a short polypurine tract that ends at position 192. This is followed by the main body of the *CR1* family consisting of about 291 nucleotides. This main body begins with a dinucleotide T-G (position 193–194) and ends with the consensus sequence T-T-C-A (positions 480–483). The characteristic structural features of *CR1*s are summarized in Fig. 3*A* and closely resemble those of avian retroviral LTRs and some eukaryotic transposable elements. Avian retroviral LTR sequences are also ≈300 bp long, are terminated by the sequences T-G.....T-T-C-A, and, in the case of LTRs located at the 3′ end of the retrovirus, are preceded by a polypurine tract. The polypurine tract serves as the (+)-strand primer site during retroviral replication (for a review, see ref. 8). For LTRs that possess the polypurine tracts, the homology may be expected often to extend further in the upstream direction, since the body of the virus lies in that direction. Thus, in this respect, five of the *CR1* family members sequenced resemble 3′ LTRs. However, two of the *CR1*s (*CR1OVa* and *CR1OVc*) lack the polypurine tract preceding position 192. It is therefore possible that these two *CR1*s represent 5′-LTR sequences of retroviral DNAs.

The similarity of *CR1*s and LTRs is further supported by an analysis of the combined structure of the two *CR1*s that exist at the 5′ end of the *X–Y*–ovalbumin gene domain. It is possible that *CR1OVc* and *CR1OVb* represent the 5′ and 3′ LTRs of an intact proviral element. This is shown in Fig. 3*B*. (For the genomic location of these two *CR1* family members, see Fig. 4.) These two *CR1*s are separated by a distance of 8 kb, which is a typical overall length observed for avian retroviruses (9). Each *CR1* is flanked by the short inverted repeat T-G...C-A, which is also a characteristic of viral LTRs. The main body of *CR1OVc*, the 5′ element, is closely followed by the sequence T-T-T-G, which forms part of the binding site of the tRNA primer involved in the initiation of (−)-strand DNA synthesis of avian retroviruses (8). Conversely, *CR1OVb*, the 3′ element, is immediately preceded by a polypurine tract, which can serve as the (+)-strand primer binding site (8). Moreover, the entire "proviral" element is flanked at each end by the short direct repeat C-C-T-(A)-T. It is a characteristic of retroviruses that they duplicate 4–6 bp of host cell DNA sequences at the site of integration, thereby leading to such flanking short direct repeats (8).

As further support for the relationship between *CR1*s and viral LTRs, Fig. 3*C* shows a comparison of sequences from *CR1OVc* and *CR1OVb* with sequences from the 5′ and 3′ LTRs of RAV-O, an endogenous virus spontaneously released by cells from certain lines of chickens (9). The similarity of overall structure and of nucleotide sequence homology between the *CR1*s and the LTRs of RAV-O is very striking.

Additional Structural Features of the Chicken *CR1* Family. Within the body of the *CR1* sequence, the three ovalbumin domain *CR1*s (*CR1OVa/b/c*) exhibit fairly poor homology in the region preceding position 261 (only about 60% homology to the consensus sequence compared to an average 91% homology for the other four *CR1* family members). Following position 261, however, all seven of the *CR1* sequences are

```
                                1        10        20        30        40        50        60        70        80        90       100
CONSENSUS                       GGCCACGARAGNNGATNAGRGGGCTGNAGYACCTYCCCTRYGAGGACAGGCTGAGGGAGCTGGGCTTRTTCAGCCTGAGAAGAGAAGGCTGYGGNGTRAC
CR1CMc    GAGCCTTGAATGGCACCTGT--••----G--CG---G--G------G--T----T----AC--------------------------A---------------------•-C--G--A--
CR1CMb    TTGGATTGGGTCCAGAGGAG------•-A--AT---C--A------C--C----C----GT--------------------------G-----------------------T--T--G--

          101      110       120       130       140       150       160       170       180       190             200
CONSENSUS CTCANTGCAGCCTTTCAGTACCTRAAGGGAGCCTATAAACAGGANGGGAGTCAACCCTCTGAAAGGGTAGATAAYAGCAGGACAAGGGGAATGGT      TTGA
CR1CMc    ----T------------------G--------------------A------A------T----------A----C-----------------C----------T--
CR1CMb    ----G------------------A-----------------A--G--------------•ACG-----------T--------G----••----------------
CR1CMa    TGCACGTGACAGGAGGTTGGAGCTTGAT--•T--T---•••---C••--------T--A---------------T--------------A-----------C----
CR1U1a    TGGGGTTGTGTGATCCGACTGGGGCTTCCCCAAAGGCACTAATGCTTC-T-------------------C----CG------T------A---C--------G---
CR1OVb    CCAAGGGAAACTCATTCTCCTTATGAGGTTCTGAGTGACTCTTGCTTAGTTTAGGAAAGCAATGGAGATCGAGTACTCTCA--------------------G-CT-
CR1OVc    CAGACTGGAGAAGATAAGGCTGAAGGGGATCATATCAGTGATTACAAATACTTAAAGGGCAGAAGCCAAGTGAATAGGGCCAGGCTCCTTT----ATCCTG---C
CR1OVa    GGTTAAATTTTGCAATTCTACCTACAAATGGAAGCATGTCTTAAGTATTGCATAAGGTGAAGGTATCACAGGATGCTGTACTAATCAGTTG--•-ACCC--•-•C

          201      210       220            230       240          250       260                 270       280       290       300
CONSENSUS AGTTGAAGGAGGGAAGATTNAGGT     TGGATATCAGGGGAAG   TTCTTTACTATGAGAGTGG        TGAGGT GCTGGAACAGGCTGCCCAGAGAGGTTGTGGATGC
CR1CMc    ----A--A-----------G--------------G-T-------------------C-A-------------------------------G----•------------------A-
CR1CMb    ----------AA--T----T--------------G-----AA-G---------------------------------------T--------•••••••••-----C---------
CR1CMa    -------------------T-------------------------------------C----------------------------------•-------------C---------
CR1U1a    -----G-------------G-------------C----------------------------------------------------------•-----------------------
CR1OVb    -AC-A--A---C-G-A---T---------AA---G-T---CAT-TA----------•-C•----•--C---A--A------CACCA-C-------T--------A-C-----•---
CR1OVc    --GAA-T--•-C---A---C-•••-----ACACC-A--A-A----TAC----C----T-••---•---T--AACA•----•-A----------------G-------------•-T
CR1OVa    -T---•----A--C-A-CAG--AGAGAAAATA-A-G--A•--------G-G--•-GC----CT--A-CTTGATCA-----•CA----------T------A---------------

          301        310       320       330       340       350       360         370       380       390        400
CONSENSUS C  GTCCATCCCTGGAGGTGTTCAARGCCAGGYTGGATGGGGCCCTGGGCAGCCTGGTCTAGTA YTG  TGTGGAGGYTGATGGCCCTGCCYATAGCA GGGGGGG
CR1CMc    -CT----------A-----------A---GA-T-------T------•••••••••••••••••••••••••••••••••••••••••••••--••------------•
CR1CMb    G--T---T-T-A-A-T-AGCACCTTAAGACTGATGATTCTCTTATAGGAGTCTCTCTGAATTGCTCTCAGGTTGCAGCAAAGCATGCTGTTCTGGCTGCAGCTGATC
CR1CMa    -CC----------------------G-T----C-------------------A------------T----••------T--GC---T------TG-GA-•--------
CR1U1a    -CC----------------------G---G--T----C-T-•---------------G--G----C--AA-------TC--------------T•••-••---------
CR1OVb    G--••----------------CAC----A------T-------------------A----AG--G--•-CC•••••-••-•-C••---••---------C-C-----T-----TT-
CR1OVc    ---•---T--T-----AA-A-----AA--T-CC-------CTTT----T---A---AC-----•••••••••••----CT-•••••••----••••----••--A-T-A-
CR1OVa    ---••---------------CA-----G------C-------T---T---AA----------------GGT--G---•••-••••••••••------AC-------------A•

          401      410       420        430       440       450       460          470         480
                                                                                                      GAANNNGCAC
CONSENSUS GGGTTGAAGCTTGGTGATC CTTGAGGTCCCTTCCAACCCAGGCCATTCTATGATTCTATGATC  TAATGCACCTT G   GGGA   TTCACTYANCCA
CR1CMc    •••---G--A--C•••--••••••---A•------••---•--•T--••--G--•---AG--T••••••G-----•••A-AT--•-••••--CTCAGCCATCTGCTGAGCTGCAGCTT
CR1CMa    ---•--G------AG-------CA------------------A-----------------G---G-T----------C-••••----GT-A--GAAACAGCACCAGCCCGGTTGCTGGG
CR1U1a    ------C--A--T----C-----------------------------G-------•--T-C---••---T----T-CA----GT----TCCCTCACCCTACCAGCTCTGAAGAA
CR1OVb    ---C---•-••---•-••--T----------------------AA--T----------AG-•-A--•A-•---CT--•••••••••••-•••---CTTATCCAGGACAGCCCAGTAAATCT
CR1OVc    •••-----•--CA--A---T-CA-----------------TCTA-A------------C--••------C-•G----A-AA---•••---GAATTTGCACATACGGTATAATGTTC
CR1OVa    •••-----A--A-A-------A---T-----T--T------------------------------AT--G•••-•---••••--•••••---•TATACCTTCTGGAAACAGGCACGGTG
```

FIG. 2. Sequence comparison of seven members of the chicken *CR1* repetitive DNA family. A consensus sequence, based upon the most common nucleotide at any one position, is shown in the upper line. The numbering system refers to positions along the consensus sequence. Nucleotide 1 is the position at which *CR1CMc* and *CR1CMb* begin to show a significant degree of homology. Other members of the *CR1* family exhibit homology beginning only at later positions. Dashes indicate identity of the individual sequence to the consensus sequence at any given position. Dots, on the other hand, indicate the absence of a nucleotide at any particular position. Gaps have been introduced at certain positions in the consensus sequence when nucleotides are present in <50% of the individual sequences. Nucleotides that are boxed do not show any reasonable degree of homology to the consensus sequence and were not included when determining the consensus sequence. R stands for purine, Y for pyrimidine, and N for any nucleotide.

very highly homologous. With the exception of *CR1CMc* and *CR1CMb*, this particular stretch of very high homology (at least 80%) continues up to position 361. Following position 361, however, there is a stretch of about 60 nucleotides that are less well conserved. In this region (position 362 to about position 421), the individual *CR1* sequences are characterized by an abundance of deletions relative to the consensus sequence. In turn, this relatively divergent region is followed by a second very strongly conserved region extending over a length of about 30–40 nucleotides. In this second highly conserved region (positions 423–456), five of the *CR1* sequences (*CR1CMc* and *CR1CMb* excepted) are 85–97% homologous to the consensus sequence.

Thus, five of the *CR1* family members shown in Fig. 2 share the following characteristic internal features: two regions of highly conserved sequence are separated by an intervening region of relatively lower sequence homology. However, the two family members *CR1CMb* and *CR1CMc* do not fit precisely into this same general pattern. As shown in Fig. 2, the family member *CR1CMb* appears to be prematurely truncated at its 3′ end with respect to the other *CR1* family members. No significant homology in *CR1CMb* has been detected beyond position 315 of the consensus sequence.

In the case of *CR1CMc*, a deletion of a single block of about 39 nucleotides between positions 346 and 384 appears to have occurred. A possible mechanism for this deletion is revealed by the nucleotide sequences in this region. In the consensus sequence, there is a nearly perfect direct repetition of 15 nucleotides immediately preceding the beginning and the end of this block of 39 deleted nucleotides. To be more specific, the sequence between positions 328–345 (*A-G-G-Y-T*-G-G-A-*T*-G-G-G-G-C-C-C-T-G) is repeated at positions 371–384 (*A-G-G-Y-T-G-A-T-G-G-C-C-C-T-G*). If such sequences existed at both positions in an earlier evolutionary version of *CR1CMc*, it is quite possible that homologous re-

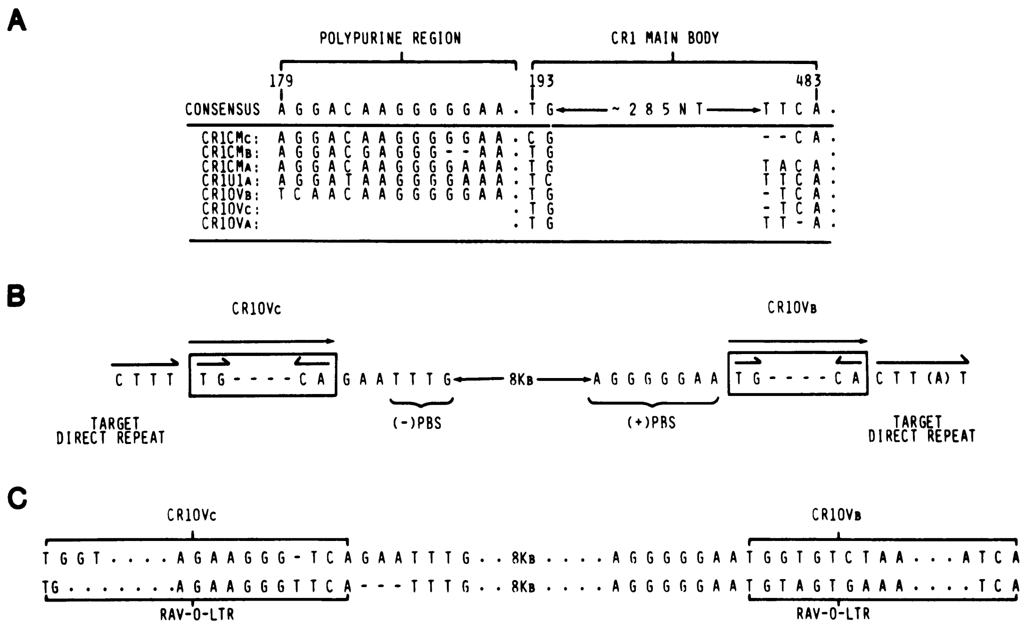
A
POLYPURINE REGION
CR1 MAIN BODY
179
193
483
CONSENSUS A G G A C A A G G G G G A A . T G ~ 2 8 5 N T T T C A .
CR1CMc: A G G A C A A G G G G G A A . C G - - C A .
CR1CMb: A G G A C G A G G G - - A A . T G .
CR1CMa: A G G A C A A G G G G A A A . T G T A C A .
CR1U1a: A G G A T A A G G G G A A A . T C T T C A .
CR1OVb: T C A A C A A G G G G G A A . T G - T C A .
CR1OVc: . T G - T C A .
CR1OVa: . T G T T - A .
B
CR1OVc
CR1OVb
C T T T
T G - - - - C A
G A A T T T G
8Kb
A G G G G G A A
T G - - - - C A
C T T (A) T
TARGET DIRECT REPEAT
(-)PBS
(+)PBS
TARGET DIRECT REPEAT
C
CR1OVc
CR1OVb
T G G T . . . . A G A A G G G - T C A G A A T T T G . . 8Kb . . . . A G G G G G A A T G G T G T C T A A . . . A T C A
TG . . . . A G A A G G G T T C A - - - T T T G . . 8Kb . . . . A G G G G G A A T G T A G T G A A A . . . . T C A
RAV-O-LTR
RAV-O-LTR

FIG. 3. Similarity between sequences flanking *CR1* repeats and avian retroviral LTRs. (*A*) Nucleotide sequences at the beginnings and ends of seven *CR1* family members are shown. The main body of the *CR1* sequence begins with T-G and ends with C-A. In five of the sequences, the T-G is preceded by a polypurine tract. (*B*) Two *CR1* family members (*CR1OVc* and *CR1OVb*), separated by about 8 kb, may form the ends of an intact retroviral-like element. The following binding characteristics are shown: (*i*) short inverted repeats (T-G...C-A) at the ends of each *CR1*; (*ii*) sequences corresponding to the primer binding sites for (+)- and (−)-strand synthesis during viral replication; (*iii*) a 4-nucleotide direct repeat corresponding to host cell DNA duplicated at the insertion site. (*C*) A comparison is shown of *CR1* sequences to LTR sequences from an avian retrovirus, Rous-associated virus O (RAV-O). The sequences from the *CR1*s are shown in the upper line, and the LTR sequences are shown in the lower line. PBS stands for primer binding site.

combination could be responsible for the apparent 39-nucleotide deletion observed in *CR1CMc*.

DISCUSSION

To clearly define the genomic structure of the chicken *CR1* family, we have now sequenced a total of seven members of this family. We were somewhat surprised to find that the length of the tract of sequence homology extending in the 5′ direction was quite variable when different *CR1* sequences were compared. This is different from human *Alu* sequences, in which case an unique 5′ end can usually be assigned precisely (10). The retroviral LTR model presented here would predict that homology should extend in the 5′ direction only in those elements that have polypurine tracts and therefore would be expected to represent 3′ LTRs. This holds true for all of the *CR1* sequences examined and provides an explanation for the variable lengths of homology.

Another way in which *CR1* sequences differ from their *Alu* counterparts is that no poly(A) tracts are observed in the 3′-flanking DNA. That property is a characteristic feature of most *Alu* sequences (10). To explain the dispersion of *Alu* sequences throughout the human genome, models have been proposed utilizing the idea of transcription–reverse transcription–reinsertion (11, 12). These models, which are sufficient to explain the constant 5′ ends, include as essential elements the presence of the 3′ poly(A) tract and the ability of *Alu* sequences to act as promoters for RNA polymerase III. *CR1* sequences, in contrast, have variable 5′ ends, lack 3′ poly(A) tracts, and are not known to function as promoters for RNA polymerase III (unpublished results).

Moreover, individual *Alu* sequences are usually flanked by short direct repeats (10). In our first report describing the *CR1* family (2), short direct repeats flanking two members of the *CR1* family were identified. However, it is now evident that those repeats were a fortuitous occurrence. The more complete data now available clearly show that individual members of the *CR1* family generally are not flanked by short direct repeats. Therefore, it seems likely from the cumulative data that many of the details involved in the dispersal of *CR1* repeats through the chicken germ line must be different from those involved in the dispersal of modern mammalian *Alu*-like repeats.

The present data show that chicken *CR1* sequences possess certain essential structural features of integrated retroviral LTRs. These include (*i*) terminal inverted repeats (T-G...C-A); (*ii*) (+)- and (−)-strand primer binding sites; (*iii*) terminal sequence homology of the *CR1OVb/c* pair with the LTRs of the avian retrovirus RAV-O; and (*iv*) structural homology of the *CR1OVb/c* to the overall structure of an integrated retrovirus. On the other hand, certain transcriptional regulatory signals ("TATA" boxes and polyadenylylation signals), which are associated with viral LTRs and required for viral gene expression, are not readily apparent in most of the *CR1*s studied. Although we can find a TATA box (T-A-T-A-T at position 238) and poly(A) addition signal (A-A-T-A-A-A at position 460) in *CR1OVb*, these signals are generally not conserved in the other *CR1*s. It is likely that the absence of these regulatory signals reflects the long residence of the *CR1*s in the chicken genome.

Indeed, if *CR1* sequences enter into the chicken genome as the result of integrations by an intact retrovirus, it must be the case that most of the *CR1*s described in this report have had a very long residence in the chicken genome, because it appears from the data that most of them are not components of intact proviral elements. For example, all three of the *CR1CM* sequences most closely resemble 3′ LTRs in that all of them possess the polypurine tract preceding the T-G dinucleotide. Two of the other *CR1*s (*U1* and *OVa*) are not paired in the genome with other LTR-like sequences, thus indicating that they are not part of intact proviruses. Examples have been described in the literature of chicken endogenous proviral elements that consist only of sequences corresponding to a single LTR and no additional retroviral sequences (9). It has been postulated in these cases that homologous

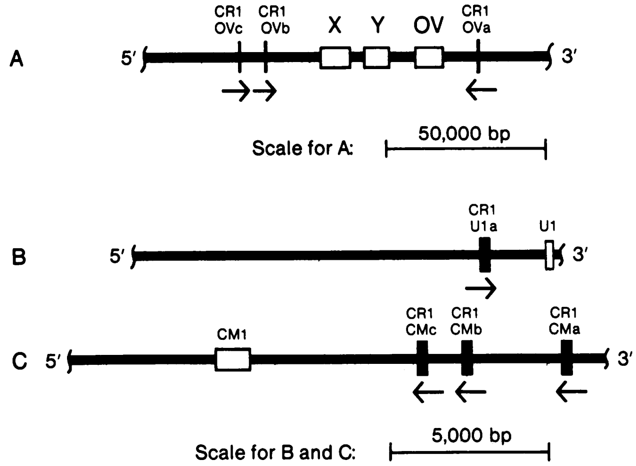

FIG. 4. Genomic locations of *CR1* sequences. *CR1* sequences have been characterized near three different structural gene loci in the chicken genome. The extents of the cloned regions around these three loci are shown. The structural genes are represented as unshaded boxes, and in all cases the direction of transcription is from left to right. The locations of *CR1* sequences are shown by solid vertical bars. In all seven of the cases examined, the orientations of the *CR1* sequences are such that they "point" toward the nearby structural genes. Note that the scale for *B* and *C* is 10-fold expanded relative to that of the ovalbumin region shown in *A*.

recombination between the 5′ and 3′ LTRs has resulted in the excision of sequences corresponding to the virus and to one of the LTR equivalents. This is a plausible mechanism for the generation of individual *CR1* sequences.

Our main motivation in studying *CR1* sequences arises from their possible role in the regulation of chromatin structure surrounding active genes (3). In the oviduct where the *X*, *Y*, and ovalbumin genes are expressed, these three genes exist in a 100,000-bp domain of preferential DNase I sensitivity (13). At each end of this domain, the chromatin structure reverts to the relatively DNase I-resistant configuration characteristic of the bulk of the chromatin (13). Previous work has shown that two *CR1* sequences (*CR1OVb* and *CR1OVc*) are located at the 5′ end of this domain, and one *CR1* sequence (*CR1OVa*) is located at the 3′ end of the DNase I-sensitive domain (3). *CR1OVa* exists in an inverse orientation with respect to *CR1OVb* and *CR1OVc*—that is, *CR1OVb* and *CR1OVc*, located 5′ of the gene cluster, have the same orientation as the *X*, *Y*, and ovalbumin genes, whereas *CR1OVa*, located 3′ of the structural genes, has the reverse orientation. This is diagramed in Fig. 4*A*. This same pattern holds true for the other known *CR1* sequences *CR1U1a* is located 5′ of a *U1* gene and has the same orientation as the *U1* gene (Fig. 4*B*). *CR1CMa*, *CR1CMb*, and *CR1CMc*, on the other hand, lie 3′ of the *cCM1* gene and have orientations that are the opposite to that of the *cCM1* gene (Fig. 4*C*). Thus, in all seven of the cases examined, the *CR1* sequences are oriented such that they all point toward the nearby structural genes. The repeated finding of this structural arrangement would not be expected on a random basis and further suggests that the *CR1* sequences may play a functional role in the regulation of gene expression, possibly via an effect on chromatin structure.

If such an effect exists, several factors could be involved. One possibility is that *CR1* sequences may serve as recognition sites for proteins involved in the regulation of chromatin structure. Recently, we have been able to isolate a protein factor that binds specifically to restriction fragments that contain *CR1* sequences (unpublished results). The existence of a specific *CR1* binding protein raises the possibility that this protein may be involved in the observed alteration of chromatin structure in regions of transcriptional activity.

In summary, the data presented here suggest that retroviral mechanisms have been responsible for the duplication and dispersal of *CR1* repetitive elements in the evolving chicken genome. In addition, the interesting orientation and location of *CR1*s with respect to expressed cellular genes suggest a possible role for *CR1*s in chromatin structure.

We acknowledge Ms. Debbie Spector for excellent technical assistance and Dr. Christina Chang for helpful discussions and sequence analysis. This work was supported in part by a Summer Faculty Fellowship awarded to W.E.S. by the San Diego State University Foundation and by National Institutes of Health grants (HD-08188 and HD-07495) to B.W.O.

1. Britten, R. J. & Davidson, E. H. (1969) *Science* **165**, 349–357.
2. Stumph, W. E., Kristo, P., Tsai, M.-J. & O'Malley, B. W. (1981) *Nucleic Acids Res.* **8**, 5383–5397.
3. Stumph, W. E., Baez, M., Beattie, W. G., Tsai, M.-J. & O'Malley, B. W. (1983) *Biochemistry* **22**, 306–315.
4. Stein, J. P., Munjaal, R. P., Lagace, L., Lai, E. C., O'Malley, B. W. & Means, A. R. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 6485–6489.
5. Roop, D. R., Kristo, P., Stumph, W. E., Tsai, M.-J. & O'Malley, B. W. (1981) *Cell* **23**, 671–680.
6. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
7. Catterall, J. F., Stein, J. P., Kristo, P., Means, A. R. & O'Malley, B. W. (1980) *J. Cell Biol.* **87**, 480–487.
8. Varmus, H. E. (1982) *Science* **216**, 812–820.
9. Hughes, S. H., Toyoshima, K., Bishop, J. M. & Varmus, H. E. (1981) *Virology* **108**, 189–207.
10. Schmid, C. W. & Jelinek, W. R. (1982) *Science* **216**, 1065–1070.
11. Van Arsdell, S. W., Denison, R. A., Bernstein, L. B., Weiner, A. M., Manser, T. & Gesteland, R. F. (1981) *Cell* **26**, 11–17.
12. Jagadeeswaran, P., Forget, B. G. & Weissman, S. M. (1981) *Cell* **26**, 141–142.
13. Lawson, G. M., Knoll, B. J., March, C. J., Woo, S. L. C., Tsai, M.-J. & O'Malley, B. W. (1982) *J. Biol. Chem.* **257**, 1501–1507.