The G9a gene in the human major histocompatibility complex encodes a novel protein containing ankyrin-like repeats

Caroline M. MILNER and R. Duncan CAMPBELL*

MRC Immunochemistry Unit, Department of Biochemistry, South Parks Road, Oxford OX1 3QU, U.K.

The class III region of the human major histocompatibility complex spans approx. 1.1 Mbp on the short arm of chromosome 6 and is known to contain at least 36 genes. The complete nucleotide sequence of a 3.4 kb mRNA from one of these genes, G9a (or BAT8), has been determined from cDNA and genomic DNA clones. The single-copy G9a gene encodes a protein product of 1001 amino acids with a predicted molecular mass of 111518 Da. The C-terminal region (residues 730–999) of the G9a protein has been expressed in *Escherichia coli* as a fusion protein with the 26 kDa glutathione *S*-transferase of *Schistosoma japonicum* (Sj26). The fusion protein has been used to raise antisera which, in Western-blot analysis, cross-react specifically with an intracellular protein of approx. 98 kDa. The function of the G9a protein is unknown. However, comparison of the derived amino acid sequence of G9a with the protein databases has revealed interesting similarities with a number of other proteins. The C-terminal region of G9a is 35% identical with a 149 amino acid segment of the *Drosophila* trithorax protein. In addition the G9a protein has been shown to contain six contiguous copies of a 33-amino acid repeat. This repeat, originally identified in the Notch protein of *Drosophila* and known as the cdc10/SW16 or ANK repeat, is also found in a number of other human proteins and may be involved in intracellular protein–protein interactions.

INTRODUCTION

The human major histocompatibility complex (MHC) spans approx. 4000 kb of DNA on the short arm of chromosome 6 (Trowsdale et al., 1991). The class I and class II regions lie at the telomeric and centromeric ends of the MHC respectively (Dunham et al., 1987, 1989). These contain genes that encode the highly polymorphic histocompatibility antigens required for the presentation of antigens to T-cells (Strachan, 1987; Trowsdale, 1987; Davis and Bjorkman, 1988). The 1100 kb of DNA between the class I and class II regions is generally termed the class III region (Carroll et al., 1987; Dunham et al., 1987). The MHC is associated with susceptibility to a large number of diseases, many of which are autoimmune in nature (Tiwari and Terasaki, 1985; Todd et al., 1988). The identification and characterization of novel genes in this region is therefore of particular interest.

The whole of the class III region has been cloned in overlapping cosmids and yeast artificial chromosomes (Sargent et al., 1989a; Spies et al., 1989a,b; Kendall et al., 1990; Ragoussis et al., 1991). The characterization of the cloned DNA has shown that this region is densely populated and contains at least 36 genes most of which are unrelated to each other and to the histocompatibility antigens. These include the genes encoding the complement components C2, C4 and factor B (Carroll et al., 1984), the enzyme cytochrome *P*-450 steroid 21-hydroxylase (Carroll et al., 1985; White et al., 1985), the cytokines tumour necrosis factors α and β (Spies et al., 1986; Dunham et al., 1987), two members of the 70 kDa heat-shock protein family (Sargent et al., 1989b; Milner and Campbell, 1990) and the enzyme valyl-tRNA synthetase (Hsieh and Campbell, 1991).

One of the newly identified genes in the class III region, which lies approx. 40 kb telomeric of the C2 gene, has been termed G9a (Dunham et al., 1990) or BAT8 (Spies et al., 1989a). This gene encodes a 3.4 kb mRNA which is expressed in all the cell types analysed to date. Here we describe the cloning and sequencing of the G9a cDNA. Antisera have been raised against the C-terminal region of the predicted protein product of the G9a gene and the expression of this protein in the cell has thus been confirmed by Western-blot analysis. The G9a protein has been shown to contain six copies of a 33-amino acid repeat also found in at least 20 other proteins including the Notch protein of *Drosophila* (Wharton et al., 1985) and a number of human proteins such as NF-κB, ankyrin and the proto-oncogene *bcl*-3 (Kieran et al, 1990; Lux et al., 1990; Ohno et al., 1990; Ellisen et al., 1991; Haskill et al., 1991; LaMarco et al., 1991).

MATERIALS AND METHODS

Isolation of cDNA clones

The 46.5 kb DNA insert from the cosmid J8b (Sargent et al., 1989a) was radiolabelled and used to probe a U937 cDNA library (a gift from D. Simmons, Institute of Molecular Medicine, Oxford, U.K.) constructed in the CDM8 vector (Seed, 1987; Simmons and Seed, 1988). Forty-six positives were detected out of the approx. 5×10^5 colonies screened. Of these positives, 17 were shown to correspond to the G9 gene (Sargent et al., 1989a). The remainder were rescreened and 20 cDNA clones were identified for G9a.

Cloning and nucleotide sequence analysis

For shotgun sequencing, the 3.2 kb insert of the G9a-4C7 cDNA was purified from low-gelling-temperature (LGT) agarose and digested with *Hin*fI, *Dde*I, *Msp*I and *Hin*fI–*Eco*O109. The small DNA fragments generated were ligated into M13mp10 and sequenced. The sequence data were assembled into contiguous sequences using the SAP program of Staden (1986). Specific fragments obtained by digestion of the G9a-4C7 cDNA with the enzymes *Nco*I, *Bgl*II, *Pvu*II, *Bst*EII and *Xho*I were also cloned into M13mp10 and sequenced. Genomic DNA fragments were

Abbreviations used: MHC, major histocompatibility complex; LGT agarose, low-gelling-temperature agarose; PMA, phorbol 12-myristate 13-acetate; IPTG, isopropyl β-thiogalactoside; MTPBS, mouse tonicity phosphate-buffered saline; GABP, guanine adenine binding protein.

* To whom correspondence should be addressed.

cloned into the pGEM-3Zf(+) (Promega, Madison, WI, U.S.A.) or pBluescript (Stratagene, La Jolla, CA, U.S.A.) vectors. Single-stranded DNA was recovered from these clones using the helper phage M13K07 in the presence of kanamycin. All sequencing was carried out by the dideoxy chain termination method (Sanger et al., 1977) using the Sequenase system (US Biochemicals, Cleveland, OH, U.S.A.) with the M13 universal primer (5′-GTAAAACGACGGCCAGT-3′).

The derived amino acid sequence was compared with the National Biomedical Research Foundation (NBRF) and SwissPot protein databases using the FASTA program (Pearson and Lipman, 1988). The significance of protein sequence similarity detected by database searching was determined using the ALIGN program (Dayhoff et al., 1983). ALIGN analysis was carried out using the 250 PAMs mutation data matrix with a bias of 6 added to each term of the matrix and a break penalty of 6. Some 100 random runs were performed to determine the mean random score.

Southern-blot analysis

Cosmid cloned DNA (1 μg) and genomic DNA (5 μg) were digested with restriction enzymes under the conditions recommended by the supplier. The digested DNA was fractionated on 0.8% (w/v) agarose gels, transferred to nitrocellulose membranes (Southern, 1975) and hybridized with ^{32}P-labelled probes. Probes were labelled directly in LGT agarose by random hexanucleotide priming (Feinberg and Vogelstein, 1984). Blots were hybridized and washed as described in Hsieh and Campbell (1991) and autoradiographed between two intensifying screens at −70 °C for 1–5 days.

Isolation of RNA and Northern-blot analysis

The cell lines U937, activated U937, HepG2, Raji, Molt4, HeLa and HL60 were grown in tissue culture to densities of 1×10^6–2×10^6 cells/ml. U937 cells were activated by stimulation with phorbol 12-myristate 13-acetate (PMA) (Sigma) for 3 days before their collection. Total RNA was extracted by the guanidine isothiocyanate lysis method and caesium chloride ultracentrifugation (Chirgwin et al., 1979; Sambrook et al., 1989). Samples of RNA (15 μg) were fractionated in 0.8% (w/v) agarose/formaldehyde denaturing gels and transferred to nitrocellulose (Fourney et al., 1988). Northern blots were hybridized with ^{32}P-labelled probes and processed as described for Southern blots.

Transcription mapping by RNAase protection and primer extension

For transcription mapping (Melton et al., 1984), RNA probes of high specific activity were synthesized *in vitro* using the Riboprobe Gemini System II (Promega). A 700 bp *Sma*I genomic fragment containing the 5′ end of the G9a gene was subcloned into the pGEM-3Zf(+) vector (Promega). The DNA (1 μg) was linearized by digestion with *Eco*RI, and a ^{32}P-labelled antisense RNA probe was transcribed using bacteriophage SP6 polymerase under the conditions specified by the supplier (Promega). Transcription mapping was carried out as described in Hsieh and Campbell (1991) using 1×10^5 c.p.m. of RNA probe and 10 μg of total RNA in each reaction. The ^{32}P-labelled DNA duplexes were analysed by electrophoresis in 6% (w/v) polyacrylamide/7 M urea gels.

Primer extension was carried out as described by Wu et al. (1987). Reactions were performed using 10 μg of total RNA and 5 ng of an oligonucleotide primer (5′-GCCACCTCCTGAGTTCAGCTTCCTCC-3′) end-labelled to a specific radioactivity of approx. 10^8 c.p.m./μg with [γ-^{32}P]ATP. Extended products were analysed by electrophoresis in 6% (w/v) polyacrylamide/7 M urea gels.

Expression of fusion proteins in *Escherichia coli*

The pGEX-2T vector was used for the *in vivo* expression of a fusion protein containing a G9a polypeptide in *E. coli* (Smith and Johnson, 1988). An 810 bp *Hin*cII fragment encoding amino acids 730–999 of the G9a protein was isolated from the G9a-4C7 cDNA and cloned into *Eco*RI-cut pGEX-2T vector (Smith and Johnson, 1988) with the maintenance of the Sj26 reading frame. An overnight culture of the *E. coli* (NM554) transformant of this clone was diluted 1:10 in Luria broth/ampicillin (50 μg/ml) and grown for 5 h, with isopropyl β-thiogalactoside (IPTG) being added (final concentration 0.1 mM) after the first hour. *E. coli* cells were pelleted by centrifugation and resuspended in mouse tonicity phosphate-buffered saline (MTPBS, 150 mM NaCl/16 mM Na_2HPO_4/4 mM NaH_2PO_4, pH 7.3). Cells were lysed by sonication, and Triton X-100 was added [final concentration 1% (v/v)]. Insoluble material was removed by centrifugation (10 min; 1000 *g*; 4 °C) and the supernatant was mixed with 0.1 vol. of preswollen glutathione–agarose beads at room temperature. The beads were collected by centrifugation (5 min; 500 *g*; 4 °C) and washed three times with MTPBS. Fusion protein was eluted from the beads with 5 mM GSH in 50 mM Tris/HCl (pH 8)/0.03% (w/v) SDS. Expression of fusion protein was determined by analysis of samples by SDS/PAGE [10% (w/v) gels].

Preparation of antisera

Rabbits were immunized with the G9a–Sj26 fusion protein by multiple intradermal injections containing 100–200 μg of antigen in 50% (v/v) Fruend's complete adjuvant. Sera were recovered from blood samples taken 5–14 days after the second and subsequent immunizations.

Western-blot analysis

U937 cells in PBS were mixed with an equal volume of 2 × sample loading buffer [6% (w/v) SDS/20% (v/v) glycerol/1.4 mM 2-mercaptoethanol/0.12 M Tris/HCl, pH 6.8/0.025% (v/v) Bromophenol Blue] and cells were lysed by boiling for 10 min. Samples were electrophoresed on SDS/8% (w/v) polyacrylamide gels and proteins were transferred to nitrocellulose by electroblotting. Strips of nitrocellulose containing transferred protein were incubated for 1 h in PBS/0.1% (v/v) Tween 20. Sera were then added at appropriate dilutions and blots were incubated for a further 2 h. After being washed three times in PBS, the strips were incubated for 2 h with alkaline phosphatase-conjugated goat anti-(rabbit IgG) antibody (Sigma) diluted 1:2000 in PBS/0.1% (v/v) Tween 20. The strips were then washed three times in PBS before incubation in freshly prepared alkaline phosphatase substrate solution [0.07 M Tris/acetate, pH 8.6/0.05% (w/v) 4-Nitroblue Tetrazolium chloride/0.0024% (w/v) 5-bromo-4-chloro-3-inoxoyl phosphate *p*-toluidine/4 mM $MgCl_2$]. Development was stopped by transfer of strips to PBS. All incubations were carried out at room temperature with constant agitation.

RESULTS

Isolation and characterization of the G9a cDNA

Twenty positive clones corresponding to G9a, with insert sizes ranging from 0.6 kb to 3.2 kb, were isolated from a U937 cDNA

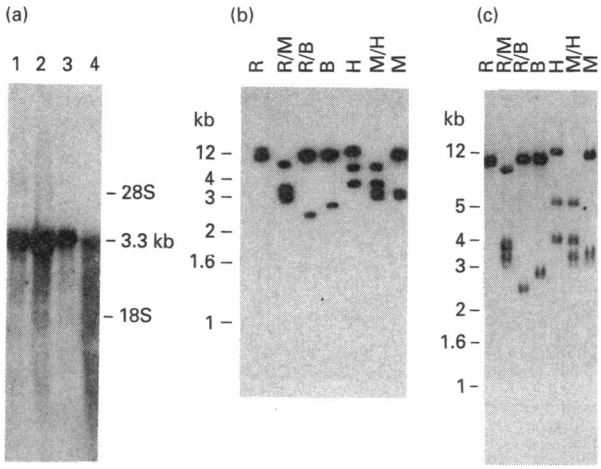

Figure 1 Southern- and Northern-blot analysis

(**a**) Northern-blot analysis. The 4C7 probe was hybridized to a Northern blot containing approx. 15 μg of total RNA from the cell lines U937 (lane 1), U937 stimulated with PMA (lane 2), HepG2 (lane 3) and HL60 (lane 4). The positions of migration of 28S and 18S RNA are indicated. (**b**) Cosmid and (**c**) genomic Southern-blot analysis. The 4C7 probe was hybridized to Southern blots of cosmid J8b and genomic (ICE5) DNA digested with *Eco*RI (R), *Eco*RI–*Bam*HI (R/M), *Eco*RI–*Bgl*II (R/B), *Bgl*II (B), *Hin*dIII (H), *Bam*HI–*Hin*dIII (M/H) and *Bam*HI (M). The positions of DNA markers are indicated.

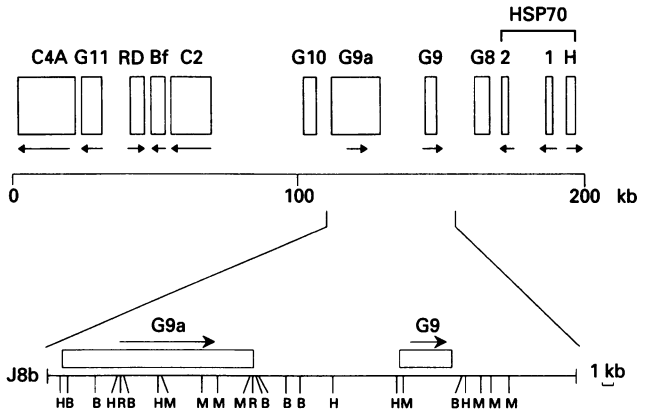

Figure 2 Location of the G9a gene in the MHC class III region

A molecular map of the central part of the class III region. The positions of genes are shown by open boxes with their directions of transcription indicated by arrows. A complete map of this region can be found in Trowsdale et al. (1991). The expanded region in the lower part of the figure shows a restriction map of the cosmid J8b (Sargent et al., 1989a) which contains the G9a and G9 genes. Sites are shown for the restriction enzymes *Bgl*II (B), *Bam*HI (M), *Eco*RI (R) and *Hin*dIII (H).

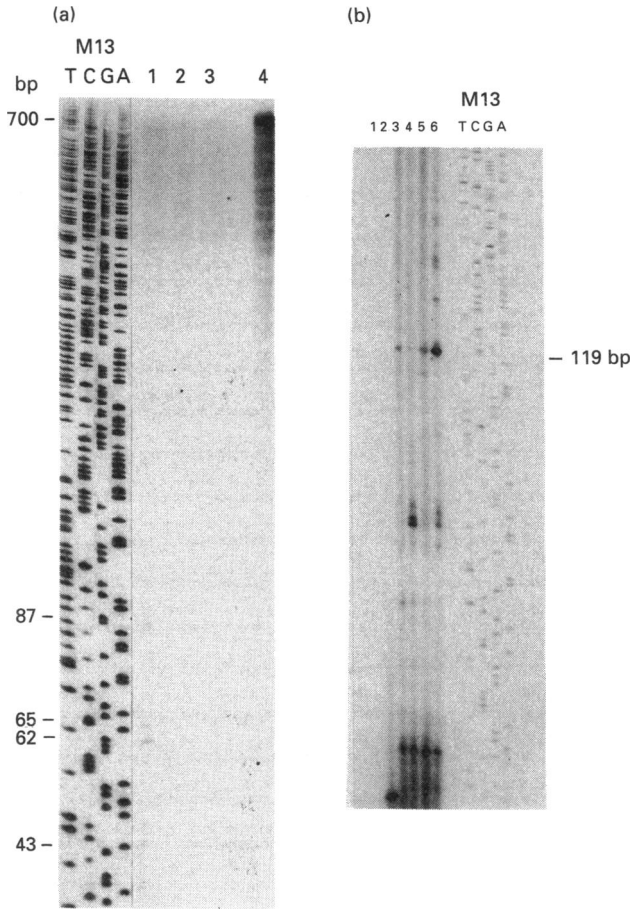

Figure 3 Mapping of the G9a transcription start site

(**a**) RNAase protection analysis using a 700 bp probe (lane 4) complementary to exon 1, exon 2 and part of exon 3 of the G9a gene. The probe was annealed to total RNA from U937 cells (lane 1), calf thymus tRNA (lane 2) and no RNA (lane 3). After digestion with RNAase A and T1, the resultant protected fragments were fractionated in a 6% (w/v) polyacrylamide/7M urea gel. The sizes of fragments in lane 1 are shown on the left. Sizes are taken from the M13mp10 sequencing ladder. (**b**) Primer extension using an oligonucleotide primer complementary to nucleotides 100–125 of the G9a gene. The labelled primer was hybridized to no RNA (lane 1), calf thymus tRNA (lane 2) and total RNA from U937 (lane 3), HepG2 (lane 4) and HL60 (lane 5) cells and a HLA homozygous B lymphoblastoid cell line (lane 6). The products of primer extension were fractionated in a 6% polyacrylamide/7M urea gel. An M13mp10 sequencing ladder was used as a size standard.

library using the genomic insert of cosmid J8b as a probe. The 3.2 kb insert of the largest cDNA (G9a-4C7) was separated from the CDM8 vector by digestion with *Xho*I and used as a probe (probe 4C7) in Northern- and Southern-blot analysis. In Northern-blot analysis of total RNA prepared from cell lines representing monocytes (U937), macrophages (U937 stimulated with PMA), hepatocytes (HepG2), T-lymphocytes (Molt4), B-lymphocytes (Raji and Daudi), epithelial cells (HeLa) and neutrophilic promyelocytes (HL60), a single mRNA of approx. 3.4 kb was detected (examples are shown in Figure 1a).

The 4C7 probe was hybridized to Southern blots of cloned DNA from cosmid J8b and uncloned genomic DNA from the HLA homozygous cell line ICE5, which was originally used to construct the cosmid library from which J8b was isolated (Sargent et al., 1989a). The pattern of hybridization observed on the blot of cosmid J8b DNA (Figure 1b) localized the G9a gene approx. 40 kb telomeric of the C2 gene, between G9 and G10 (Figure 2) and indicated that it extended over at least 14 kb. On a genomic Southern blot the 4C7 probe hybridized uniquely to fragments corresponding to those detected on the cosmid blot (Figure 1c). This result suggests that the G9a gene is a single-copy gene in the human genome.

The nucleotide sequence of the G9a-4C7 cDNA was determined by a combination of shotgun cloning and specific fragment cloning as described in the Materials and methods section. The complete sequence of the G9a-4C7 insert (3199 bp) was determined on both strands with a degeneracy of approx. 5. A polyadenylation signal (AATAAA) and a 17-base poly(A) tail were identified at one end of the G9a-4C7 cDNA, indicating that this corresponded to the 3′ end of G9a. However, the longest open reading frame from the cDNA (ORF1), which comprised

```
                                                  M  S  D  D  V  H  S  L  G  K  V  T  S  D    14
AGCCCCCGGTCCCTGAGAAGCGGCCCCCTGAAATACAGCATTTCCGCATGAGTGATGATGTCCACTCACTGGGAAAGGTGACCTCAGATC    90

L  A  K  R  R  K  L  N  S  G  G  G  L  S  E  E  L  G  S  A  R  R  S  G  E  V  T  L  T  K    44
TGGCCAAAAGGAGGAAGCTGAACTCAGGAGGTGGCCTGTCGGAGGAGTTAGGTTCTGCCCGGCGTTCAGGAGAAGTGACCCTGACGAAG   180

G  D  P  G  S  L  E  E  W  E  T  V  V  G  D  D  F  S  L  Y  Y  D  S  Y  S  V  D  E  R  V    74
GGGACCCCGGGTCCCTGGAGGAGTGGGAGACGGTGGTGGGTGATGACTTCAGTCTCTACTATGATTCCTACTCTGTGGATGAGCGCGTGG   270

D  S  D  S  K  S  E  V  E  A  L  T  E  Q  L  S  E  E  E  E  E  E  E  E  E  E  E  E  E  E    104
ACTCCGACAGCAAGTCTGAAGTTGAAGCTCTAACTGAACAACTAAGTGAAGAGGAGGAGGAGGAAGAGGAAGAAGAAGAGGAAG      360

E  E  E  E  E  E  E  E  E  E  D  E  E  S  G  N  Q  S  D  R  S  G  S  S  G  R  R  K  A  K    134
AGGAGGAGGAAGAGGAAGAAGAAGAGGAAGATGAGGAGTCAGGGAATCAGTCAGATAGGAGTGGTTCCAGTGGCCGGCGCAAGGCCAAGA   450

K  K  W  R  K  D  S  P  W  V  K  P  S  R  K  R  R  K  R  E  P  P  R  A  K  E  P  R  G  V    164
AGAAATGGCGAAAAGACAGCCCATGGGTGAAGCCGTCTCGGAAACGGCGCAAGCGGGAGCCTCCGCGGGCCAAGGAGCCACGAGGAGTGA   540

N  G  V  G  S  S  G  P  S  E  Y  M  E  V  P  L  G  S  L  E  L  P  S  E  G  T  L  S  P  N    194
ATGGTGTGGGCTCCTCAGGCCCCAGTGAGTACATGGAGGTCCCTCTGGGGTCCCTGGAGCTGCCCAGCGAGGGACCCTCTCCCCCAACC   630

H  A  G  V  S  N  D  T  S  S  L  E  T  E  R  G  F  E  E  L  P  L  C  S  C  R  M  E  A  P    224
ACGCTGGGGTGTCCAATGACACATCTTCGCTGGAGACAGAGCGAGGGTTTGAGGAGTTGCCCCTGTGCAGCTGCCGCATGGAGGCACCCA   720

K  I  D  R  I  S  E  R  A  G  H  K  C  M  A  T  E  S  V  D  G  E  L  S  G  C  N  A  A  I    254
AGATTGACCGCATCAGCGAGAGGGCGGGGCACAAGTGCATGGCCACTGAGAGTGTGGACGGAGAGCTGTCAGGCTGCAATGCCGCCATCC   810

L  K  R  E  T  M  R  P  S  S  R  V  A  L  M  V  L  C  E  T  H  R  A  R  M  V  K  H  H  C    284
TCAAGCGGGAGACCATGAGGCCATCCAGCCGTGTGGCCCTGATGGTGCTCTGTGAGACCCACCGCGCCCGCATGGTCAAACACCACTGCT   900

C  P  G  C  G  Y  F  C  T  A  G  T  F  L  E  C  H  P  D  F  R  V  A  H  R  F  H  K  A  C    314
GCCCGGGCTGCGGCTACTTCTGCACGGCGGGCACCTTCCTGGAGTGCCACCCTGACTTCCGTGTGGCCCACCGCTTCCACAAGGCCTGTG   990

V  S  Q  L  N  G  M  V  F  C  P  H  C  G  E  D  A  S  E  A  Q  E  V  T  I  P  R  G  D  G    344
TGTCTCAGCTGAATGGGATGGTCTTCTGTCCCCACTGTGGGGAGGATGCTTCTGAAGCTCAAGAGGTGACCATCCCCCGGGGTGACGGGG  1080

V  T  P  P  A  G  T  A  A  P  A  P  P  P  L  S  Q  D  V  P  G  R  A  D  T  S  Q  P  S  A    374
TGACCCCACCGGCCGGCACTGCAGCTCCTGCACCCCCACCCCTGTCCCAGGATGTCCCCGGGAGACGCAGACACTTCTCAGCCCAGTGCCC  1170

R  M  R  G  H  G  E  P  R  R  P  P  C  D  P  L  A  D  T  I  D  S  S  G  P  S  L  T  L  P    404
GGATGCGAGGGCATGGGGAACCCCGGCGCCCGCCCTGCGATCCCCTGGCTGACACCATTGACAGCTCAGGGCCCTCCCTGACCCTGCCCA  1260

N  G  G  C  L  S  A  V  G  L  P  L  G  P  G  R  E  A  L  E  K  A  L  V  I  Q  E  S  E  R    434
ATGGGGGCTGCCTTTCAGCCGTGGGGCTGCCACTGGGGCCAGGCCGGGAGGCCCTGGAAAAGGCCCTGGTCATCCAGGAGTCAGAGAGGC  1350

R  K  K  L  R  F  H  P  R  Q  L  Y  L  S  V  K  Q  G  E  L  Q  K  V  I  L  M  L  L  D  N    464
GGAAGAAGCTCCGTTTCCACCCTCGGCAGTTGTACCTGTCCGTGAAGCAGGGCGAGCTGCAGAAGGTGATCCTGATGCTGTTGGACAACC  1440

L  D  P  N  F  Q  S  D  Q  Q  S  K  R  T  P  L  H  A  A  A  Q  K  G  S  V  E  I  C  H  V    494
TGGACCCCAACTTCCAGAGCGACCAGCAGAGCAAGCGCACGCCCCTGCATGCAGCCGCCCAGAAGGGCTCCGTGGAGATCTGCCATGTGC  1530

L  L  Q  A  G  A  N  I  N  A  V  D  K  Q  Q  R  T  P  L  M  E  A  V  V  N  N  H  L  E  V    524
TGCTGCAGGCTGGAGCCAACATAAATGCAGTGGACAAACAGCAGCGGACGCCACTGATGGAGGCCGTGGTGAACAACCACCTGGAGGTAG  1620

A  R  Y  M  V  Q  R  G  G  C  V  Y  S  K  E  E  D  G  S  T  C  L  H  H  A  A  K  I  G  N    554
CCCGTTACATGGTGCAGCGTGGTGGCTGTGTCTATAGCAAGGAGGAGGACGGTTCCACCTGCCTCCACCACGCAGCCAAAATCGGGAACT  1710

L  E  M  V  S  L  L  L  S  T  G  Q  V  D  V  N  A  Q  D  S  G  G  W  T  P  I  I  W  A  A    584
TGGAGATGGTCAGCCTGCTGCTGAGCACAGGACAGGTGGACGTCAACGCCCAGGACAGTGGGGGGTGGACGCCCATCATCTGGGCTGCAG  1800

E  H  K  H  I  E  V  I  R  M  L  L  T  R  G  A  D  V  T  L  T  D  N  E  E  N  I  C  L  H    614
AGCACAAGCACATCGAGGTGATCCGCATGCTACTGACGCGGGGCGCCGACGTCACCCTCACTGATAACGAGGAGAACATCTGCCTGCACT  1890

W  A  S  F  T  G  S  A  A  I  A  E  V  L  L  N  A  R  C  D  L  H  A  V  N  Y  H  G  D  T    644
GGGCCTCCTTCACGGGCAGCGCCGCCATCGCCGAAGTCCTTCTGAATGCGCGCTGTGACCTCCATGCTGTCAACTACCATGGGGACACCC  1980

P  L  H  I  A  A  R  E  S  Y  H  D  C  V  L  L  F  L  S  R  G  A  N  P  E  L  R  N  K  E    674
CCCTGCACATCGCAGCTCGGGAGAGCTACCATGACTGCGTGCTGTTATTCCTGTCACGTGGGGCCAACCCTGAGCTGCGGAACAAAGAGG  2070

G  D  T  A  W  D  L  T  P  E  R  S  D  V  W  F  A  L  Q  L  N  R  K  L  R  L  G  V  G  N    704
GGGACACAGCATGGGACCTGACTCCCGAGCGCTCCGACGTGTGGTTTGCGCTTCAACTCAACCGCAAGCTCCGACTTGGGGTGGGAAATC  2160

R  A  I  R  T  E  K  I  I  C  R  D  V  A  R  G  Y  E  N  V  P  I  P  C  V  N  G  V  D  G    734
GGGCCATCCGCACAGAGAAGATCATCTGCCGGGACGTGGCTCGGGGCTATGAGAACGTGCCCATTCCCTGTGTCAACGGTGTGGATGGGG  2250

E  P  C  P  E  D  Y  K  Y  I  S  E  N  C  E  T  S  T  M  N  I  D  R  N  I  T  H  L  Q  H    764
AGCCCTGCCCTGAGGATTACAAGTACATCTCAGAGAACTGCGAGACGTCCACCATGAACATCGATCGCAACATCACCCACCTGCAGCACT  2340

C  T  C  V  D  D  C  S  S  S  N  C  L  C  G  Q  L  S  I  R  R  W  Y  D  K  D  G  R  L  L    794
GCACGTGTGTGGACGACTGCTCTAGCTCCAACTGCCTGTGCGGCCAGCTCAGCATCCGGCGCTGGTATGACAAGGATGGGCGATTGCTCC  2430

Q  E  F  N  K  I  E  P  P  L  I  F  E  C  N  Q  A  C  S  C  W  R  N  C  K  N  R  V  V  Q    824
AGGAATTTAACAAGATTGAGCCTCCGCTGATTTTCGAGTGTAACCAGGCGTGCTCATGCTGGAGAAACTGCAAGAACCGGGTCGTACAGA  2520

S  G  I  K  V  R  L  Q  L  Y  R  T  A  K  M  G  W  G  V  R  A  L  Q  T  I  P  Q  G  T  F    854
GTGGCATCAAGGTGCGGCTACAGCTCTACCGAACAGCCAAGATGGGCTGGGGGGTCCGGCGCCCTGCAGACCATCCCACAGGGGACCTTCA  2610

I  C  E  Y  V  G  E  L  I  S  D  A  E  A  D  V  R  E  D  D  S  Y  L  F  D  L  D  N  K  D    884
TCTGCGAGTATGTCGGGGAGCTGATCTCTGATGCTGAGGCTGATGTGAGAGAGGATGATTCTTACCTCTTCGACTTAGACAACAAGGATG  2700

G  E  V  Y  C  I  D  A  R  Y  Y  G  N  I  S  R  F  I  N  H  L  C  D  P  N  I  I  P  V  R    914
GAGAGGTGTACTGCATAGATGCCCGTTACTATGGCAACATCAGCCGCTTCATCAACCACCTGTGTGACCCCAACATCATTCCCGTCCGGG  2790

V  F  M  L  H  Q  D  L  R  F  P  R  I  A  F  F  S  S  R  D  I  R  T  G  E  E  L  G  F  D    944
TCTTCATGCTGCACCAAGACCTGCGATTTCCACGCATCGCCTTCTTCAGTTCCCGAGACATCCGGACTGGGGAGGAGCTAGGGTTTGACT  2880

Y  G  D  R  F  W  D  I  K  S  K  Y  F  T  C  Q  C  G  S  E  K  C  K  H  S  A  E  A  I  A    974
ATGGCGACCGCTTCTGGGACATCAAAGCAAATATTTCACCTGCCAATGTGGCTCTGAGAAGTGCAAGCACTCAGCCGAAGCCATTGCCC  2970

L  E  Q  S  R  L  A  R  L  D  P  H  P  E  L  L  P  E  L  G  S  L  P  P  V  N  T  *         1001
TGGAGCAGAGCCGTCTGGCCCGCCTGGACCCACACCCTGAGCTGCTGCCCGAGCTCGGCTCCCTGCCCCCCTGTCAACACATGAGAACGGA  3060

CCACACCCTCTCTCCCCAGCATGGATGGCCACAGCTCAGCCGCCTCCTCTGCCACCAGCTGCTCGCACCCATGCCTGGGGGTGCTGCCAT  3150

CTTCTCTCCCCACCACCCTTTCACACATTCCTGACCAGAGATCCCAGCCAGGCCCTGGAGGTCTGACAGCCCCTCCCTCCCAGAGCTGGT  3240

TCCTCCCTGGGAGGGCAACTTCAGGGCTGGCCACCCCCCGTGTTCCCCATCCTCAGTTGAAGTTTGATGAATTGAAGTCGGGCCTCTATG  3330

CCAACTGGTTCCTTTTGTTCTCAATAAATGTTGGGTTTGGTAATAAAAAAAAAAAAAAA                                  3390
```

Figure 4 cDNA sequence and derived amino acid sequence of G9a

The complete nucleotide sequence of the G9a mRNA from the transcription start site (defined by RNAase protection) to the poly(A) tail is shown. The derived amino acid sequence of G9a is shown above the nucleotide sequence. The stop codon (TGA) is marked by an asterisk (*) and the polyadenylation signal is underlined.

2859 nucleotides (encoding 953 amino acids), did not start with a Met codon, indicating that the G9a-4C7 cDNA was incomplete at the 5′ end.

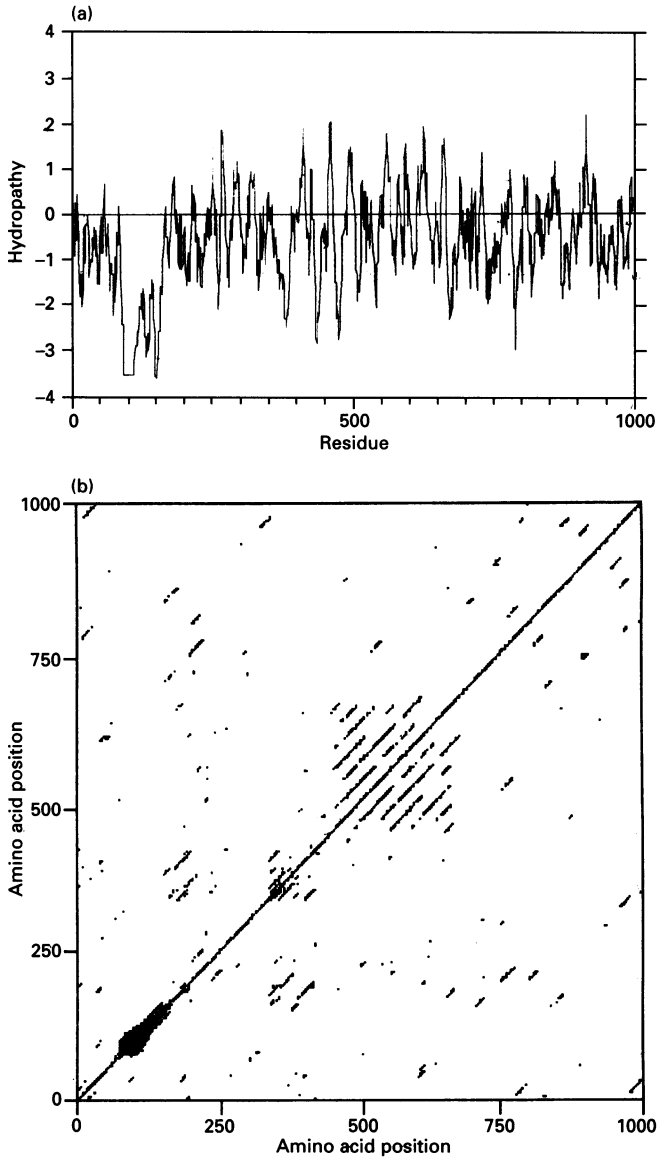

Figure 5 Analysis of the G9a amino acid sequence

(**a**) A hydropathy plot (Kyte and Doolittle, 1982) for the G9a protein obtained using the PepWindow program (Smithies et al., 1981). Amino acid positions are shown on the horizontal axis and hydropathy measures on the vertical axis. Positive values correspond to hydrophobic regions of the protein and negative values to hydrophilic regions. (**b**) Localization of repetitive sequences in the G9a protein using the DIAGON program (Devereux et al., 1984). Amino acid positions for G9a are shown on the horizontal and vertical axes.

5′ End mapping and completion of the cDNA sequence

An approx. 700 bp *Sma*I fragment, containing the 5′ end of the G9a gene, was subcloned in both orientations into *Hin*cII-cut pGEM-3Zf(+) vector (designated pGEM-0.7S-9a-1 and -2) and sequenced. The results of this analysis, in addition to sequence data from anchored PCR studies (results not shown), extended the open reading frame, ORF1, to include a potential initiation Met codon (ATG); 302 bp of DNA was sequenced upstream of this Met codon. An in-frame stop codon (TAA) was identified 105 bp from the putative initiation Met and there were no other in-frame Met codons in the intervening sequence. The 700 bp *Sma*I genomic fragment was shown to contain exon 1 (encoding

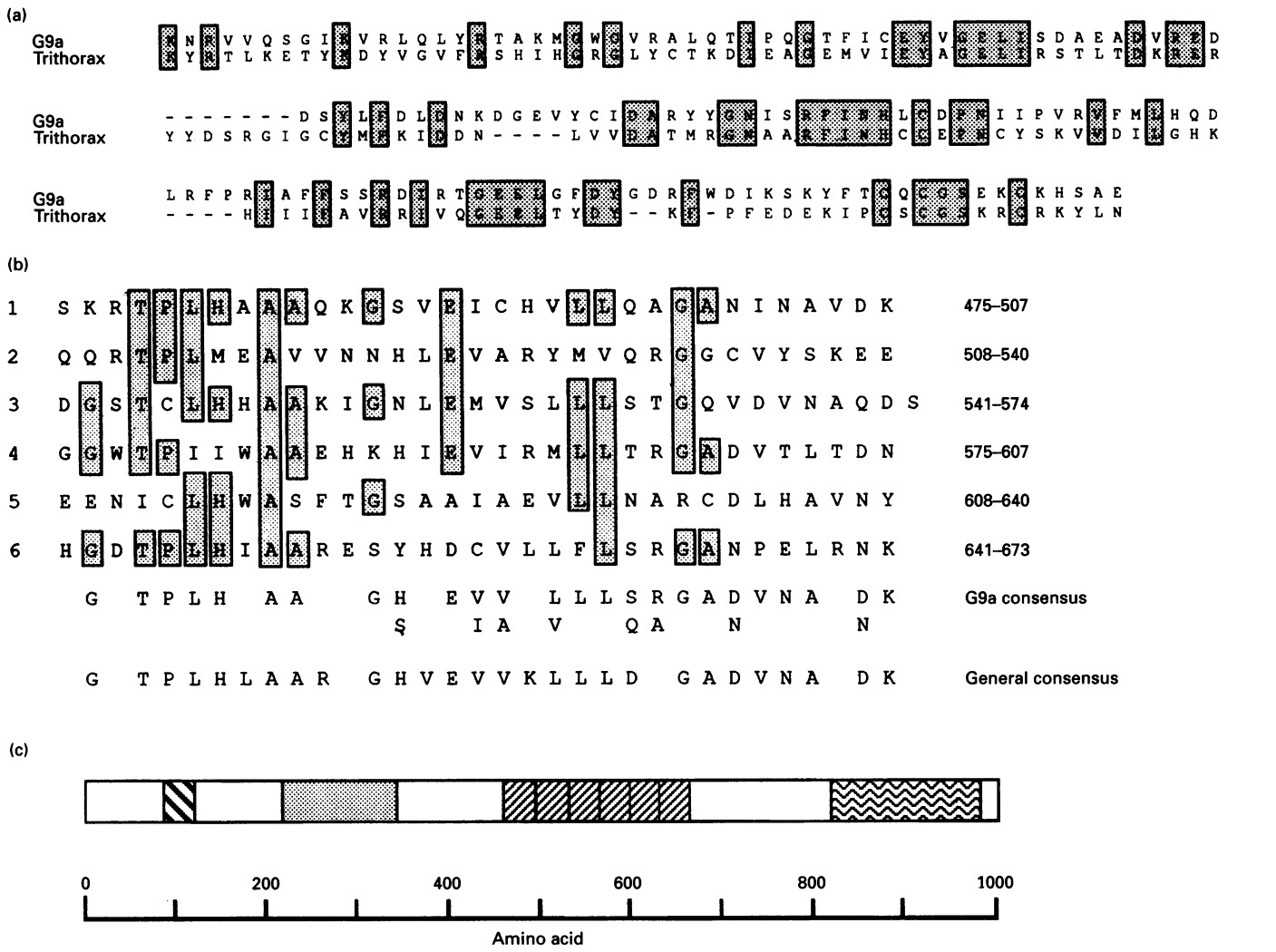

Figure 6 Structure and homology of the G9a protein

(**a**) An alignment of residues 819–971 of G9a and residues 3611–3759 of the *Drosophila* trithorax protein using the ALIGN program of Dayhoff et al. (1983). Identical residues in the two sequences are marked by shaded boxes. Gaps (—) have been inserted to maximize the alignment. An ALIGN score of 11.02 standard deviation units was obtained. (**b**) An alignment of the six ANK repeats in the G9a protein. Residues that are conserved in at least three of the six repeats are marked by shaded boxes. A consensus sequence for the repeats in G9a is shown below together with a general consensus for the ANK repeat. Amino acid positions for G9a are shown on the right. (**c**) A schematic representation of the primary structure of the G9a protein. The positions of interesting sequence elements are marked: ▨, polyglutamic acid; ▩, cysteine-rich region; ■, ANK repeat; ▨, region identical with trithorax.

amino acids 1–13), exon 2 (encoding amino acids 14–27) and part of exon 3 (encoding amino acids 27 onwards).

The pGEM-0.7S-9a-2 clone was used in the determination of the G9a transcriptional start site by RNAase mapping, as described in the Materials and methods section. Specific products (not present in control samples) of 43 bp and 62 bp were observed after electrophoresis of the reaction products under denaturing conditions (Figure 3a). These are close to the expected sizes of products corresponding to exon 2 (41 bp) and the 5′ end of exon 3 (60 bp). In addition, a product of 87 bp was observed which was assumed to correspond to exon 1. These results indicate that the position of the transcriptional start site is 47 bp upstream of the initiating Met codon (Figure 4).

To complement the RNAase mapping result, the position of the 5′ end of the G9a gene was also determined by primer extension using an oligonucleotide complementary to nucleotides 100–125 (Figure 4). The primer extension was carried out as described in the Materials and methods section and the extended products were analysed by denaturing gel electrophoresis (see Figure 3b). For each of the four RNA samples used [U937, HepG2 and HL60 cell lines and a HLA homozygous (B44 DR4) B lymphoblastoid cell line], a major product of 119 bp was obtained which was not detectable in the control samples. This result places the G9a transcriptional start site 6 bp downstream of the position predicted by RNAase mapping. For the purposes of numbering, the start site (defined as nucleotide position +1) has been assumed to be that defined by RNAase mapping (Figure 4).

The complete G9a cDNA sequence is shown in Figure 4. The sequence analysis of the G9a-4C7 cDNA yielded 3199 bp of sequence. This was extended at the 5′ end to give 3391 bp of sequence in total (taking the start site defined by RNAase mapping as position +1). The sequence upstream of the putative initiation Met codon at position 48 (TCCGC*ATG*) does not

correspond well to the consensus for eukaryotic initiation sites [CC(A/G)CC*ATG*] (Kozak, 1984). However, this is also true of the next three in-frame Met codons in the open reading frame ORF1. Therefore, it seems clear that the Met codon at position 48 is the initiation codon. The cDNA includes 3003 bp of coding sequence, from the initiation Met codon (ATG) at position 48 to the stop codon (TGA) at position 3051. The 3′ end of the cDNA comprises 338 bp of 3′-untranslated sequence, with a polyadenylation signal (AATAAA) at position 3353, followed by a 17-base poly(A) tail at position 3373. The 5′-untranslated region of G9a is very short, comprising just 47 bp of sequence. The 5′-flanking sequence of the G9a gene up to position −256 does not contain TATA or CAAT box consensus sequences (results not shown).

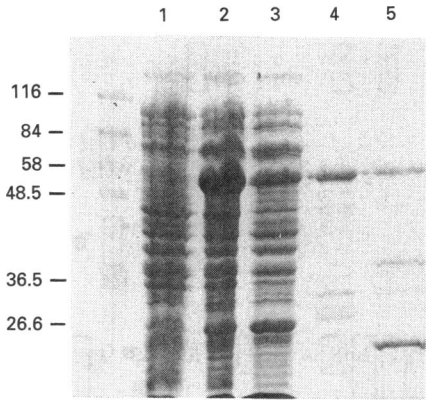

Figure 7 Expression of a G9a–Sj26 fusion protein

Overnight cultures of *E. coli* (NM554) cells transformed with the pGEX-9a-13 construct were diluted 1:10 in Luria broth/ampicillin and grown for 5 h (lane 1). IPTG was added to 0.1 mM after the first hour (lane 2). Cells were lysed by sonication and, after the removal of insoluble debris, the lysate was adsorbed to glutathione–agarose beads. Lane 3 contains cell lysate after adsorption and lane 4 shows the beads to which the G9a–Sj26 fusion protein had been adsorbed. The fusion protein was digested with thrombin while bound to the beads for 2 h at 25 °C in 0.15 M NaCl/2.5 mM $CaCl_2$ (lane 5). Samples were mixed with an equal volume of 2 × sample buffer (Laemmli 1970), boiled for 10 min and electrophoresed on an SDS/10% (w/v) polyacrylamide gel. The gel was stained with Coomassie Blue. The sizes (kDa) of protein markers are shown on the left.

The derived amino acid sequence of G9a

The longest open reading frame identified in the G9a cDNA sequence encodes a protein of 1001 amino acids (Figure 4). The N-terminal sequence of the predicted G9a protein does not fulfil the criteria for a signal sequence (von Heijne, 1985), and a hydropathy plot (Figure 5a) shows that there are no strongly hydrophobic transmembrane segments. On this basis the G9a protein product is expected to be intracellular.

The DIAGON program (Devereux et al., 1984) was used to analyse the G9a amino acid sequence for the presence of repetitive elements. The results, which are shown in Figure 5(b), indicate that the region between residues 450 and 750 contains six (or possibly seven) copies of a repeated sequence.

The G9a amino acid sequence was screened against the National Biomedical Research Foundation (NBRF) and SwissProt protein databases to look for sequence similarity with other known proteins, which might give some indication as to the function of the G9a protein. The N-terminal region of the G9a protein includes 24 contiguous glutamate residues (amino acids 91–114) (see Figures 4 and 6c). Similar long runs of negatively charged acidic residues have been identified in a number of other proteins including nucleolin (Bourbon et al., 1988), the human major centromere autoantigen (CENP-B) (Earnshaw et al., 1987) and the non-histone chromosomal protein HMG-1- (Wen et al., 1989).

Amino acids 819–971 of G9a show 35% identity with amino acids 3611–3759 (C-terminal 149 residues) of the *Drosophila* trithorax protein which is involved in the regulation of the expression of homeotic genes during the development of the *Drosophila* thorax (Mazo et al., 1990). If conservative replacements (Dayhoff et al., 1983) are taken into account, the sequence similarity between the trithorax protein and G9a increases to 48%. These two protein regions were compared, using the ALIGN program of Dayhoff et al. (1983). A score of 11.02 standard deviation units was obtained which is indicative of sequence identity and a high level of structural similarity between these two protein regions. The alignment is shown in Figure 6(a).

The most striking sequence identity associated with the G9a protein was revealed when amino acids 451–720 were screened against the protein databases. Amino acids 469–720 of G9a were shown to have 32% sequence identity with and 44% sequence similarity to residues 1895–2109 of the Notch protein of *Drosophila*. Amino acids 1895–2109 of the Notch protein constitute six contiguous copies of a 33-amino acid repeat which was first observed as an octapeptide (*TXLXLAAR*) (Wharton et al., 1985; Kidd et al., 1986). This repeat has since been shown to extend over 33 amino acids and is known as the cdc10/SW16 (Breeden and Nasmyth 1987) or ANK (Lux et al., 1990) repeat. The sequence similarity between G9a and Notch proteins indicated that amino acids 469–667 of G9a constitute six copies of the ANK repeat (Figure 6b). This is consistent with results of the DIAGON analysis.

An alignment of the six ANK repeats of G9a protein is shown in Figure 6(b). Ten out of 33 amino acids are conserved in at least four of the six G9a repeats. The close similarity between the consensus sequence for the ANK repeats of G9a protein and the general consensus for this repeat (Figure 6b) indicates that the repeats in G9a do belong to the ANK repeat family. The phasing of the repeats in Figure 6(b) corresponds to that adopted by Lux et al. (1990) on the basis of limited knowledge of the exonic structure of the human ankyrin gene. However, in other genes, including G9a (C. M. Milner and R. D. Campbell, unpublished work), that encode ANK repeats there is no evidence of single repeats being encoded by discrete exons. This and the common occurrence of these repeats in tandem arrays makes it difficult to define the boundaries of the ANK repeat.

Detection of the G9a protein by Western-blot analysis

To confirm that G9a is expressed at the protein level, it was necessary to raise antibodies against a peptide from the predicted protein sequence. A 810 bp *Hin*cII fragment encoding amino acids 730–999 was isolated from the G9a-4C7 cDNA and ligated into *Eco*RI-cut pGEX-2T and designated pGEX-9a-13.

The expression of the G9a polypeptide, as a fusion protein with Sj26, was induced as described in the Materials and methods section. The pGEX-9a-13 clone expressed a 56 kDa fusion protein, which comprised the Sj26 protein linked to the approx. 30 kDa G9a C-terminal polypeptide (270 amino acids). This fusion protein was isolated from the *E. coli* cell lysate by adsorption to glutathione–agarose beads (Figure 7). The thrombin cleavage of the pGEX-9a-13 fusion protein yielded products of 26 kDa (due to Sj26) and approx. 34 kDa. However, the free G9a peptide remained bound to the glutathione–agarose beads (Figure 7), suggesting that it may be insoluble. It did prove possible to elute the whole pGEX-9a-13 fusion protein from the beads very efficiently as described in the Materials and methods

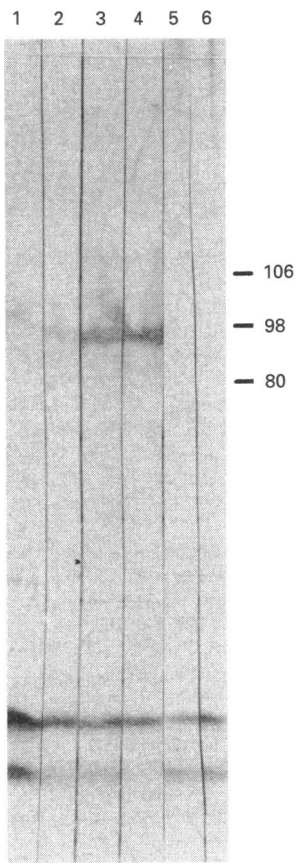

Figure 8 Western-blot analysis

U937 cells were lysed by boiling in sample loading buffer (Laemmli, 1970) and cellular proteins were fractionated by SDS/PAGE [8% (v/v) gel]. Proteins were transferred to nitrocellulose membrane by electroblotting. Strips of nitrocellulose containing bound protein were incubated with antisera from a rabbit immunized with the G9a–Sj26 fusion protein at dilutions of 1:1000 (lane 1), 1:600 (lane 2), 1:300 (lane 3) and 1:150 (lane 4), with preimmune sera from the same rabbit (lane 5) or with no primary antibody (lane 6). Antibody bound to G9a was detected by subsequent incubation with alkaline phosphatase-conjugated goat anti-rabbit IgG followed by alkaline phosphatase substrate solution. The positions of protein standards (kDa) are shown on the right.

section and this was used as antigen for injection into rabbits. The fusion protein was purified with yields of approx. 100 μg/litre of culture which is significantly lower than the reported optimal yield for this expression system (approx. 15 mg/l) (Smith and Johnson, 1988).

Before immunization, a 10 ml sample of blood was taken from each rabbit (the prebleed) and the sera were recovered for use as controls in subsequent experiments. Western blots were prepared from U937 cell lysates. On these blots antisera recovered after four immunizations detected a unique band of approx. 98 kDa at dilutions of 1:150 down to 1:1000 (Figure 8, lanes 1–4). This band was not present on blots incubated with preimmune sera from either of the rabbits, or on a blot to which no primary antibody has been added (Figure 8, lanes 5 and 6). The predicted molecular mass of the G9a protein, on the basis of the amino acid sequence, is 111500 Da and the band detected by the antisera is within 10% of this. This discrepancy may be due to the difficulty in accurately determining the molecular masses of large proteins by SDS/PAGE. Alternatively, it could be due to post-translational proteinase cleavage of the G9a protein. There are four potential N-linked glycosylation sites in G9a, but these are likely to be unglycosylated in a cytoplasmic or nuclear protein.

DISCUSSION

The results described here show that G9a is an approx. 17.5 kb gene which lies in the MHC class III region approx. 40 kb telomeric of the C2 gene. G9a appears to be ubiquitously expressed as an approx. 3.4 kb mRNA and encodes a novel protein product of 1001 amino acids, with a predicted molecular mass of 111500 Da. Western-blot analysis and preliminary cell staining studies (C. M. Milner and R. D. Campbell, unpublished work) have confirmed that the G9a protein is expressed intracellularly.

The function of the G9a protein remains to be determined. However, the sequence of this protein comprises a number of distinct domains. The N-terminal region of G9a contains a run of 24 contiguous glutamate residues. Anionic regions of this type are commonly found in nuclear proteins and it has been suggested that they may be associated with autoantigenicity (Brendel et al., 1991). This is interesting in view of the localization of the G9a gene in the MHC, a region of the genone known to be associated with susceptibility to a large number of autoimmune diseases (Tiwari and Terasaki, 1985; Todd et al., 1988).

The G9a protein contains two cysteine-rich regions, spanning amino acids 210–410 and 720–1001. The first of these regions shows no significant sequence similarity to any known protein. However, the C-terminal region of G9a (amino acids 819–917) is identical with the C-terminus of the *Drosophila* trithorax protein. The trithorax protein contains nine cysteine-rich zinc finger-like domains. In addition, the C-terminal region (amino acids 3611–3759) of this protein has been shown to bind zinc *in vitro* (Mazo et al., 1990), although it does not conform to a zinc finger or zinc twist consensus sequence (Vallee et al., 1991). Preliminary evidence suggests that the C-terminal region of G9a also has zinc-binding properties (C. M. Milner, R. D. Campbell and A. J. Day, unpublished work).

The G9a protein contains six contiguous copies of the 33-amino acid ANK repeat [reviewed by Michaely and Bennett (1992)]. This repeat has been identified in at least 20 other proteins and has been highly conserved throughout evolution, occurring in species as diverse as yeast (Breeden and Nasmyth, 1987) and man (Kieran et al., 1990; Lux et al., 1990; Ohno et al., 1990; Ellisen et al., 1991; Haskill et al., 1991; LaMarco et al., 1991), which suggests that it has an important function.

The proteins that contain ANK repeats occur in a variety of cellular environments, including the cytoplasm [e.g. the human protein NF-κB (Kieran et al., 1990)], the inner face of the plasma membrane [e.g. the Notch protein of *Drosophila* and the glp-1 and lin-12 proteins of *Caenorhabditis elegans* (Austin and Kimble, 1989; Yochem and Greenwald, 1989)], the nucleus [e.g. the human guanine adenine-binding protein (GABP) (Thompson et al., 1991) and SW14 of *Saccharomyces cerevisiae* (Breeden and Nasmyth, 1987)], mitochondria [e.g. human glutaminase (Banner et al., 1988)] and the extracellular space [e.g. α-latrotoxin from the black widow spider (Kiyatkin et al., 1990)].

The proteins containing ANK repeats have diverse functions such as the determination of cell fate (e.g. Notch, lin-12 and glp-1), the cell cycle (e.g. SW14) and the regulation of transcription (e.g. NF-κB and GABP). However, a common feature of many of these proteins is their ability to interact specifically with other proteins. For example, human erythrocyte ankyrin is known to interact with the erythroid anion exchanger via two of its 22 ANK repeats (Davis et al., 1991). ANK repeats may also be involved in the recognition of other macromolecules. For

example, a number of the proteins that contain these repeats are transcription factors which bind to DNA, and the ANK repeats of the GABP β-subunit have been implicated in DNA binding (Thompson et al., 1991). It seems likely that the G9a protein has protein- or DNA-binding properties associated with its ANK repeats but this requires further investigation. The precise localization of the G9a protein in the cell may also be helpful in elucidating its function.

We are very grateful to Dave Simmons for the U937 cDNA library. Thanks also go to Tony Day for oligonucleotide synthesis and helpful discussion and to Ken Johnson for photographic work. C.M.M. held a Medical Research Council studentship.

REFERENCES

Austin, J. and Kimble, J. (1989) Cell **58**, 565–571
Banner, C., Hwang, J. J., Shapiro, R. A., Wenthold, R. J., Nakatani, Y., Lampel, K. A., Thomas, J. W., Huie, D. and Curthoys, N. P. (1988) Mol. Brain Res. **3**, 247–254
Bourbon, H. M., Lapeyre, B. and Amalric, F. (1988) J. Mol. Biol. **200**, 627–638
Breeden, L. and Nasmyth, K. (1987) Nature (London) **329**, 651–654
Brendel, V., Dohlam, J., Blaisdell, B. E. and Karlin, S. (1991) Proc. Natl. Acad. Sci. U.S.A. **88**, 1536–1540
Carroll, M. C., Campbell, R. D., Bentley, D. R. and Porter, R. R. (1984) Nature (London) **307**, 237–241
Carroll, M. C., Campbell, R. D. and Porter, R. R. (1985) Proc. Natl. Acad. Sci. U.S.A. **82**, 521–525
Carroll, M. C., Katzman, P., Alicot, E. M., Koller, B. H., Geraghty, D., Orr, H. T., Strominger, J. L. and Spies, T. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 8535–8539
Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. and Rutter, W. J. (1979) Biochemistry **18**, 5294–5299
Davis, L. H., Otto, E. and Bennett, V. (1991) J. Biol. Chem. **266**, 11163–11169
Davis, M. M. and Bjorkman, P. J. (1988) Nature (London) **334**, 395–402
Dayhoff, M. O., Barker, W. C. and Hunt, L. T. (1983) Methods Enzymol. **91**, 524–545
Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. **12**, 387–395
Dunham, I., Sargent, C. A., Trowsdale, J. and Campbell, R. D. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 7237–7241
Dunham, I., Sargent, C. A., Dawkins, R. L. and Campbell, R. D. (1989) Genomics **5**, 787–796
Dunham, I., Sargent, C. A., Kendall, E. and Campbell, R. D. (1990) Immunogenetics **32**, 175–182
Earnshaw, W. C., Sullivan, K. F., Machlin, P. S., Cooke, C. A., Kaiser, D. A., Pollard, T. D., Rothfield, N. F. and Cleveland, D. W. (1987) J. Cell Biol. **104**, 817–829
Ellisen, L. W., Bird, J., West, D. C., Soreng, A. L., Reynolds, T. C., Smith, S. D. and Sklar, J. (1991) Cell **66**, 649–661
Feinberg, A. P. and Vogelstein, B. (1984) Anal. Biochem. **137**, 266–267
Fourney, R. M., Miyakoshi, J., Day, R. S. and Paterson, M. C. (1988) Focus (Bethesda Research Laboratories) **10**, 5–7
Haskill, S., Beg, A. A., Tompkins, S. M., Morris, J. S., Yurochko, A. D., Sampso-Johannes, A., Mondal, K., Ralph, P. and Baldwin, A. S. (1991) Cell **65**, 1281–1289
Hsieh, S. L. and Campbell, R. D. (1991) Biochem. J. **278**, 809–816
Kendall, E., Sargent, C. A. and Campbell, R. D. (1990) Nucleic Acids Res. **18**, 7251–7257
Kidd, S., Kelley, M. R. and Young, M. W. (1986) Mol. Cell. Biol. **6**, 3094–3108
Kieran, M., Blank, V., Logeat, F., Vandekerckhove, J., Lottspeich, F., LeBall, O., Urban, M. B., Kourilsky, P., Baeuerle, P. A. and Israel, A. (1990) Cell **62**, 1007–1018
Kiyatkin, N., Dulubova, I., Chekhosskaya, I. and Grishin, E. (1990) FEBS Lett. **270**, 127–131
Kozak, M. (1984) Nucleic Acids Res. **12**, 857–873
Kyte, J. and Doolittle, R. F. (1982) J. Mol. Biol. **157**, 105–132
Laemmli, U. K. (1970) Nature (London) **227**, 680–685
LaMarco, K., Thompson, C. C., Byers, B. P., Walton, E. M. and McKnight, S. L. (1991) Science **252**, 789–792
Lux, S. E., John, K. M. and Bennett, V. (1990) Nature (London) **344**, 36–42
Mazo, A. M., Huang, D. H., Mozer, B. A. and David, I. B. (1990) Proc. Natl. Acad. Sci. U.S.A. **87**, 2112–2116
Melton, D., Krieg, P. A., Rebagliati, M. R., Maniatis, T., Zinn, K. and Green, M. R. (1984) Nucleic Acids Res. **11**, 857–872
Michaely, P. and Bennett, V. (1992) Trends Cell Biol. **2**, 127–129
Milner, C. M. and Campbell, R. D. (1990) Immunogenetics **32**, 175–182
Ohno, H., Takimoto, G. and McKeithan, T. W. (1990) Cell **60**, 991–997
Pearson, W. R. and Lipman, D. J. (1988) Proc. Natl. Acad. Sci. U.S.A. **85**, 2444–2448
Ragoussis, J., Monaco, A., Mockridge, I., Kendall, E., Campbell, R. D. and Trowsdale, J. (1991) Proc. Natl. Acad. Sci. U.S.A. **88**, 3753–3757
Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
Sanger, F., Nicklen, S. and Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 5463–5467
Sargent, C. A., Dunham, I. and Campbell, R. D. (1989a) EMBO J. **8**, 2305–2312
Sargent, C. A., Dunham, I., Trowsdale, J. and Campbell, R. D. (1989b) Proc. Natl. Acad. Sci. U.S.A. **86**, 1968–1972
Seed, B. (1987) Nature (London) **329**, 840–842
Simmons, D. and Seed, B. (1988) J. Immunol. **141**, 2797–2800
Smith, D. B. and Johnson, K. S. (1988) Gene **67**, 31–44
Smithies, O., Engels, W. R., Devereux, J. R., Slighton, J. L. and Shen, S. (1981) Cell **26**, 345–353
Southern, E. M. (1975) J. Mol. Biol. **98**, 503–517
Spies, T., Morton, C. C., Nedospasov, S. A., Fiers, W., Pious, D. and Strominger, J. L. (1986) Proc. Natl. Acad. Sci. U.S.A. **83**, 8699–8702
Spies, T., Bresnahan, M. and Strominger, J. L. (1989a) Proc. Natl. Acad. Sci. U.S.A. **86**, 8955–8958
Spies, T., Blanck, G., Bresnahan, M., Sands, J. and Strominger, J. L. (1989b) Science **243**, 214–217
Staden, R. (1986) Nucleic Acids Res. **14**, 217–231
Strachen, T. (1987) Br. Med. Bull. **43**, 1–14
Thompson, C. C., Brown, T. A. and McKnight, S. L. (1991) Science **253**, 762–768
Tiwari, J. L. and Terasaki, P. I. (1985) HLA & Disease Associations, Springer Verlag, New York
Todd, J. A., Acha-Orbea, H., Bell, J. I., Chao, N., Fronek, Z., Jacob, C. O., McDermott, M., Sinha, A. A., Timmerman, L., Steinman, L. and McDevitt, H. O. (1988) Science **240**, 1003–1009
Trowsdale, J. (1987) Br. Med. Bull. **43**, 15–36
Trowsdale, J., Raggoussis, J. and Campbell, R. D. (1991) Immunol. Today **12**, 443–446
Vallee, B. L., Coleman, J. E. and Auld, D. S. (1991) Proc. Natl. Acad. Sci. U.S.A. **88**, 999–1003
von Heijne, G. (1985) J. Miol. Biol. **184**, 99–105
Wen, L., Huang, J. K., Johnson, B. H. and Reeck, G. R. (1989) Nucleic Acids Res. **17**, 1197–1214
Wharton, K. A., Jahansen, K. M., Xu, T. and Artavanis-Tsakonas, S. (1985) Cell **43**, 567–581
White, P. C., Grossberger, D., Onufer, B. J., Chaplin, D. D., New, M. I., Dupont, B. and Strominger, J. L. (1985) Proc. Natl. Acad. Sci. U.S.A. **82**, 1089–1093
Wu, L. C., Morley, B. J. and Campbell, R. D. (1987) Cell **48**, 331–342
Yochem, J. and Greenwald, I. (1989) Cell **58**, 553–563

Received 14 August 1992/8 October 1992; accepted 13 October 1992