The gene *G13* in the class III region of the human MHC encodes a potential DNA-binding protein

Anu KHANNA* and R. Duncan CAMPBELL†

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

G13 is a single-copy gene lying approx. 75 kb centromeric of the complement gene cluster in the class III region of the human MHC. The gene spans approx. 17 kb of DNA and has been shown to encode mRNA of approx. 2.7 kb that is present in cell lines representing lymphoid and non-lymphoid tissues, indicating that it is ubiquitously expressed. The complete nucleotide sequence of the 2.7 kb mRNA has been derived from cDNA and genomic clones. The longest open reading frame obtained for *G13* codes for a 703 amino acid protein of approx. 77 kDa in molecular mass. Comparison of the putative G13 amino acid sequence with the protein databases revealed significant similarities with DNA-binding proteins of the leucine zipper class, including a human cAMP response element binding protein. G13 contains a bZIP motif, a region rich in basic amino acids adjacent to a coiled-coil leucine zipper domain, common to this class of proteins that is known to be involved in dimerization and DNA binding. Antibodies raised against a fragment encoding the C-terminal half of the putative G13 protein recognized a major polypeptide of approx. 86 kDa and a minor polypeptide of approx. 78 kDa on immunoblotting of U937 cell extracts; this has been confirmed by immunoprecipitation experiments. Even though it contained at least one potential bipartite nuclear localization signal, the G13 protein was present both in the cytoplasm and the nucleus of the fibroblast cells. Thus G13 might be a novel DNA-binding protein that is perhaps translocated to the nucleus in a regulated manner.

INTRODUCTION

The human MHC is located on the short arm of chromosome 6, in the Giemsa light band 6p21.3. The class I and class II loci, which encode cell-surface molecules involved in regulation of the immune response, are located at the telomeric and centromeric ends of the MHC respectively [1]. The approx. 1.1 Mb segment of DNA separating the class I and class II regions is generally termed the class III region. Over the past few years this region has been characterized in detail by a combination of physical mapping techniques using pulsed-field gel electrophoresis and by the construction of a contig of genomic clones comprising both cosmids and yeast artificial chromosomes [2–6]. The pulsed-field gel electrophoresis mapping studies identified a large number of sites for rare-cutting restriction enzymes such as *Bss*HII, *Ksp*I, *Eag*I and *Not*I [2,4,7] clustered at CpG islands. These CpG islands are short stretches of DNA that, unlike the rest of the human genome, are not depleted in unmethylated CpG dinucleotides and they are invariably found at the 5′ ends of genes [8,9]. In particular, characterization of a cloned segment of DNA extending 200 kb centromeric of the complement gene cluster revealed the presence of seven CpG islands, and subsequent screening of a cDNA library with cosmid genomic DNA inserts has led to the isolation of cDNAs corresponding to genes associated with these CpG islands. One of these CpG-island-associated genes, labelled *G13*, was found to encode a ubiquitously expressed 2.7 kb mRNA [4,10].

Here we report the characterization of the gene *G13* at the DNA and protein levels. The derived G13 amino acid sequence was compared with the protein sequence databases, which revealed that it contains motifs similar to those found in the bZIP (basic-leucine zipper) class of DNA-binding proteins. In particular the N-terminal half of the putative G13 protein revealed 23% identity with human cAMP response element binding protein (CREB) over a stretch of approx. 300 amino acids. CREB is known to be involved in the regulation of transcription on specific induction by various signalling molecules during growth and development [11,12]. Chracterization of novel genes in the class III region of the MHC is of particular interest because genetic studies have indicated that some of the genes could be involved in susceptibility to autoimmune disorders [13].

MATERIALS AND METHODS

Southern and Northern blotting and hybridization techniques

Genomic DNA (5 μg) or cosmid/plasmid DNA (1 μg), digested with the appropriate restriction enzymes, was subjected to electrophoresis on 0.8% (w/v) agarose gels by standard procedures [14,15]. Southern blotting was performed by the procedure described in [16].

For Northern blotting, total RNA from tissue culture cells was isolated by either the caesium chloride method [17] or the RNAzol B method [18]. The preparation of RNA samples (20 μg of total RNA from each cell line) for running on agarose/formaldehyde gels and their subsequent transfer on to nitrocellulose membranes was as described in [4].

Abbreviations used: ATF, activation transcription factor; AMPS, alignment of multiple protein sequences; bZIP, basic-leucine zipper; CREB, cAMP response element binding protein; FITC, fluorescein isothiocyanate; GST, glutathione S-transferase; NLS, nuclear localization signal.

* Present address: Sir William Dunn School of Pathology, University of Oxford, Oxford OX1 3RE, U.K.

† To whom correspondence should be addressed.

The nucleotide sequences reported will appear in DDBJ, EMBL and GenBank Nucleotide Sequence Databases under the accession numbers X98053 (*G13*) and X98054 (G13 mRNA).

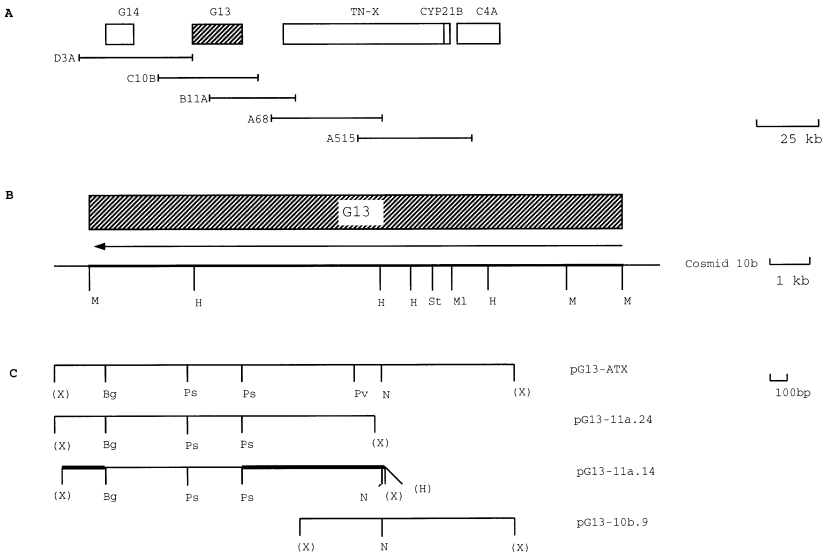


Figure 1 Localization of the gene *G13* in the class III region of the MHC

(**A**) Overlapping cosmids derived from the cell line ICE5 (HLA haplotype A2, B7, C2C, BfS, C4A3, C4BQ0, DR2) [4] illustrating the location of the gene *G13* approx. 75 kb centromeric of the CYP21B gene. (**B**) The detailed localization of gene *G13* to an approx. 17 kb segment of DNA in the cosmid C10B is based on the data in [4] and is indicated by the bold line. The arrow shows the direction of transcription. (**C**) The overlapping cDNAs corresponding to gene *G13* isolated from the U937 cDNA library. The 5′ (*Bgl*II/*Xho*I) and the 3′ (*Hin*dIII/*Pst*I) end probes made from pG13-11a.14 are shown in bold lines. The restriction sites in parentheses are located in the vector. Restriction sites in (**B**) and (**C**) are for *BamH*I (M), *Xho*I (X), *Hin*dIII (H), *Bgl*II (Bg), *Mlu*I (Ml), *Nco*I (N), *Stu*I (St), *Pst*I (Ps) and *Pvu*II (Pv).

DNA probes were labelled to a specific radioactivity of more than 10^8–10^9 c.p.m./μg by the random hexanucleotide priming method [19] using the Multiprime DNA labelling system (Amersham International). Hybridization to DNA/RNA blots was performed by the method described in [4].

Isolation of *G13* cDNA clones

Screening of a cDNA library constructed with mRNA from the monocytic cell line U937 (a gift from D. Simmons, IMM, Oxford, U.K.) with the inserts from the cosmids C10B and B11A (Figure 1A) led to the detection of 40 positive clones [4]. DNA was prepared from each of the positive areas to identify the clone with the largest cDNA insert and this clone (pG13-11a.14) was rescreened to obtain a single colony. The 1.8 kb insert of clone pG13-11a.14 detected the same 2.7 kb mRNA in all the cell lines analysed as described in [4]. To identify the clones that extended the clone pG13-11a.14, an *Xho*I/*Bgl*II fragment and a *Hin*dIII/*Pst*I fragment representing the 5′ and 3′ ends of the clone respectively were used to probe Southern blots of *Xho*I, *Xho*I/*Bgl*II and *Xho*I/*Nco*I digests of the DNA prepared from each of the positive areas. The clones that contained the largest extension at each end of clone pG13-11a.14 were rescreened to obtain single colonies.

DNA sequencing and data analysis

For sequencing, restriction digests of the cDNA or genomic clones were performed. This was followed by end filling if required, and recovery from agarose or polyacrylamide gels. Fragments were ligated into pGEM-3Zf(+) or M13mp10 vectors. Transfections of the ligation mixtures were performed in competent TG1 cells. Single-stranded DNA was prepared from the recombinant clones by the methods described in [20] (for M13mp10 vector) and [21] (for pGEM-3Zf vector).

Sequencing of DNA was performed by extension from the M13 universal primer (5′-GTAAAACGACGGCCAGT-3′) in the presence of dideoxynucleoside triphosphate chain-terminating inhibitors. Sequencing reactions were performed with Sequenase (USB, Amersham) in accordance with the supplier's instructions [22].

The sequence data obtained were analysed with the Oxford University Molecular Biology system on the VAX/VMS cluster. The G13 protein sequence was predicted from the translation of the cDNA open reading frames. The programs of Staden [23] and the GCG (Genetics Computer Group) packages [24] were used for the characterization of the primary structure of the putative protein. Sequence similarities with other proteins were determined by screening the National Biomedical Research Foundation (NBRF) and SwissProt databases by using database search programs FASTA [25] and BLAST [26]. The significance of sequence similarities and comparison of multiple sequences were performed with the AMPS program [27,28].

RNase protection analysis

Plasmid constructs consisting of the 5′ end of the gene [cloned in the 5′ → 3′ (pG13-GEM.S/M2) and 3′ → 5′ (pG13-GEM.S/M7) orientations in pGEM-3Zf] were linearized by digestion to completion with an appropriate restriction enzyme. The linearized templates were used to produce radiolabelled single-stranded anti-sense riboprobes as run-off transcripts from the T7 and SP6 promoters in the pGEM-3Zf vector with the reagents provided in the Riboprobe Gemini System II kit (Promega) in accordance with the supplier's instructions. The full-length probes were purified from incompletely synthesized products on acrylamide gels. U937 RNA (20 μg) and tRNA (2 μg, as a control) were used for RNase protection with the purified probe (5×10^5 c.p.m.) by the method of [29]. An M13mp10 sequence ladder was subjected to electrophoresis along with the protected products to allow size

determination of the protected RNA fragments. After fixation and baking, the gel was autoradiographed at −70 °C with a preflashed film and intensifying screens.

Cell culture

Human cell lines were grown at 37 °C in 5 % CO_2 in RPMI 1640 medium containing 10 % (v/v) fetal calf serum, 100 i.u. of penicillin, 100 mg of streptomycin and 100 mg of kanamycin per 500 ml of medium. U937 cells were stimulated by the addition of phorbol 12-myristate 13-acetate to the tissue culture medium at a final concentration of 100 ng/ml. After gentle mixing, the cells were allowed to grow as above for 5–7 days, after which they were harvested.

Protein expression in *Escherichia coli*

A 1.2 kb *Xho*I fragment spanning nucleotides 1372–2136 was excised from the *G13* cDNA clone pG13-10b.9, end-filled and subcloned into the *Sma*I-cut pGEX-2T expression vector, in frame with the thrombin-cleavable linker and the glutathione S-transferase (GST) gene, and transformed into NM554 competent cells. The expression of the GST–G13 fusion protein was induced by isopropyl β-D-thiogalactoside added to a final concentration of 0.1 mM after 1 h of incubation of the cells at 37 °C. The cells were harvested after 5–7 h of incubation at 37 °C. The purification of the fusion protein was performed in accordance with the procedure described in [30].

Preparation and purification of polyclonal antibodies

Uncleaved fusion protein (GST–G13) was used as an antigen for the immunization of rabbits after 5 ml blood samples (prebleed) had been taken. Rabbits were immunized by multiple intradermal injections containing 100–200 μg of the antigen as an emulsion in 50 % (v/v) Freund's complete adjuvant (Difco). For booster injections, the antigen was emulsified with Freund's incomplete adjuvant. Approx. 5–10 ml blood samples were withdrawn from rabbits 5–14 days after an immunization, and the serum was separated and stored at −20 °C [31]. To remove antibodies against GST, the antiserum was purified by affinity chromatography on a GST–CNBr-activated Sepharose 4B column prepared in accordance with the supplier's instructions (Pharmacia). After absorption of the anti-GST antibodies the beads were washed with PBS and the concentration of the unbound protein was determined by measuring the A_{280} of the unabsorbed fraction and the washes (which should contain antibodies only against the G13 polypeptide).

Western blotting

Approx. 10^8 U937 cells were pelleted by centrifugation (5 min at 500 ***g***, at room temperature) and washed three times with PBS (Dulbecco's; Gibco-BRL). The final cell pellet (about 500 μl in volume) was resuspended in 10 volumes (5 ml) of sample buffer [6 % (w/v) SDS, 20 % (v/v) glycerol, 1.4 mM 2-mercaptoethanol, 0.12 M Tris/HCl, pH 6.8]. The cell suspension was incubated in a boiling water bath for 5 min, centrifuged at 12000 ***g*** for 10 min, at room temperature, and the supernatant recovered. Cell lysates were subjected to SDS/PAGE [10 % (w/v) gel] and the proteins transferred on to membranes by using an LKB 2117 Multiphore II electrophoresis unit.

For antigen detection by the enhanced chemiluminescence method (Amersham), the membranes were incubated at room temperature for 14–16 h in blocking solution [PBS containing 0.1 % (v/v) Tween-20, 0.02 % (w/v) BSA and 10 % (w/v) non-fat dry milk], then washed three times in large volumes of PBS/0.1 % (v/v) Tween-20, each time for 5–10 min. This was followed by incubation for 1 h with the purified antiserum diluted 1:375 (v/v) in 20 % (v/v in PBS) blocking solution or the preimmune serum at an equivalent concentration and subsequent washing of the blots as before. The blots were then incubated for 1 h with horseradish peroxidase-labelled donkey anti-rabbit IgG (Amersham) [1:10000 dilution in 20 % (v/v) blocking solution] and washed again as above. Blots were then treated with the detection reagents (Amersham) in accordance with the supplier's instructions in a darkroom. Finally the blots were covered in Saran wrap and autoradiographed for a period from approx. 30 s to 10 min.

Immunoprecipitation

U937 cells were washed in RPMI 1640 incomplete medium (without methionine) (prewarmed at 37 °C). The cell pellet was resuspended in the above medium at a concentration of approx. 10^7 cells/ml and transferred to a well (1.7 cm × 1.6 cm; 3.5 ml well capacity) of a 24-well tissue-culture plate (Linbro, Flow Laboratories). After methionine deprivation for approx. 30 min, [^{35}S]methionine (ICN) was added at a concentration of 500 μCi/ml. The cells were grown for 3, 6 or 12 h to allow incorporation of the radiolabel. For the 6 and 12 h labellings the medium was supplemented with 10 % (v/v) RPMI 1640 complete medium (containing methionine).

At the end of the pulse, the cells were harvested and lysed in IP buffer [10 mM Tris/HCl, pH 7.5, containing 150 mM NaCl, 1 % (w/v) sodium deoxycholate, 1 % (v/v) NP40, 0.1 % (w/v) SDS, 1 mg/ml BSA, 0.02 % sodium azide and 1 mM PMSF] (1 ml/10^7 cells) and incubated at 37 °C for 30 min. After centrifugation for 15 min at 13000 ***g***, at room temperature, the cell lysate was recovered and cleared by incubation with the preimmune serum adsorbed on Protein A–Sepharose beads (500 μl of lysate per 50 μl of beads) for 2 h at 4 °C. Precleared cell lysate (250 μl) was then added to 50 μl of the purified anti-G13 serum or the preimmune sera (at equivalent concentrations) and, after incubation for 1 h at 4 °C, 50 μl of Protein A–Sepharose beads [25 % (v/v)] were added. The beads with bound antigen–antibody complexes were washed two or three times with IP buffer (1 ml per wash). Finally the beads were resuspended in 800 μl of the IP buffer and overlaid on to 300 μl of a 30 % (w/v) sucrose cushion [10 mM Tris/HCl, pH 7.6, containing 150 mM NaCl, 1 % (w/v) sodium deoxycholate, 1 % (v/v) NP40, 0.1 % SDS, 0.02 % sodium azide and 30 % (w/v) sucrose]. The gradient was centrifuged for 1 min at 13000 ***g***, at room temperature. The beads were resuspended in 80 μl of sample buffer and incubated for 2 min at 95 °C. The supernatant was recovered and fractionated by SDS/PAGE [10 % (w/v) gel]. After electrophoresis the gels were stained with Coomassie Blue. After destaining, the gels were soaked in Amplify (Amersham) with agitation for 15–30 min and dried under vacuum. The gels were then autoradiographed for 6–10 days.

Immunolocalization

The human fibroblast cell line GMO2922 was prepared for cell staining by growing the cells in 8.9 mm × 20 mm wells (approx. 1 ml capacity) in plastic tissue-culture dishes (4-Chamber Lab-Tek Chamber slide; Nunc, Gibco-BRL). After growth to approx. 70 % confluence, the medium was removed and the cells were washed once with PBS. The cells were fixed in ice-cold methanol and washed in PBS before incubation with 2 % (v/v) goat serum in PBS. This was followed by incubation with 0.5 ml of the

purified anti-G13 serum (diluted 1:250) or preimmune sera (diluted 1:312.5) in PBS for 1 h at room temperature. These dilutions ensured that equivalent concentrations of the preimmune serum and the purified anti-G13 serum were used. The primary antibody-bound cells were washed with PBS containing 1% (v/v) Triton X-100 and then treated with fluorescein isothiocyanate (FITC)-conjugated goat anti-rabbit IgG (Sigma), diluted 1:40 in PBS containing 1% (v/v) goat serum in accordance with the supplier's instructions. After incubation for 1 h, the cells were washed again in PBS and the slides were mounted with Vectashield mounting medium (Vector Lab.). The cells were observed with an Axioscope microscope (Zeiss) and were photographed on Kodak 1600 film with the microscope camera unit (MC-100).

RESULTS

Characterization of cDNA clones

Of the 40 cDNA clones that corresponded to the gene *G13*, one of the largest clones, pG13-11a.14 (approx. 1.8 kb), detected a transcript of approx. 2.7 kb. To identify clones that could extend the existing clone pG13-11a.14 at the 5′ and 3′ ends, a *Bgl*II/*Xho*I fragment of approx. 0.3 kb was isolated from one end of the cDNA insert and a *Hin*dIII/*Pst*I fragment of approx. 0.9 kb was isolated from the other end of the cDNA insert (Figure 1C). For the 5′ extensions, Southern blots containing *Xho*I/*Bgl*II digests of DNA preparations from the sets of primary cDNA pools were hybridized with the approx. 0.3 kb *Bgl*II/*Xho*I probe. The probe hybridized to a *Bgl*II/*Xho*I fragment of approx. 325–350 bp from the clone pG13-11a.24 (approx. 1.8 kb) compared with a *Bgl*II/*Xho*I fragment of approx. 300 bp in the original clone pG13-11a.14.

Similarly, for 3′ extensions the above procedure was repeated with a *Xho*I/*Nco*I double digest of the DNA preparations, which were probed with the approx. 0.9 kb *Hin*dIII/*Pst*I probe. In the clone pG13-10b.9 the probe hybridized to a *Xho*I/*Nco*I fragment of approx. 0.5 kb. In addition this clone did not reveal any hybridization to the 5′-end probe. Colony rescreening and subsequent restriction enzyme analysis revealed that the insert size was approx. 1.2 kb. Because the clone pG13-10b.9 overlapped with the clone pG13-11a.24 by approx. 500 bp and extended it at the 3′ end by approx. 750 bp, it was chosen for further analysis (Figure 1C).

The region of overlap of the two *G13* cDNA clones contained a unique *Pvu*II site (Figure 1C). The cDNAs were ligated at this site and a full-length cDNA was generated in the *Pvu*II-cut pATX vector. The region around the *Pvu*II site in the re-ligated *G13* cDNA clone was sequenced and this confirmed the authenticity of the clone. The approx. 2.7 kb cDNA insert from pG13-ATX was isolated by using a *Bam*HI/*Hin*dIII double digest and was used in Southern blot and Northern blot analysis. This analysis confirmed the earlier observations [4] that *G13* is a single-copy gene and is transcribed into an mRNA of approx. 2.7 kb in the cell lines tested.

cDNA sequence analysis

Four contigs were formed upon assimilation of the sequence from shotgun sequencing of the overlapping *G13* cDNA clones pG13-11a.24 and pG13-10b.9. Specific fragments were then isolated from the two cDNA clones to complete the sequencing and to obtain a single contig. The entire sequence was obtained for both strands of DNA with a degeneracy of 5.3. The sequencing confirmed the overlapping of the two cDNA clones over a region of approx. 500 bp, as predicted earlier from the restriction enzyme analysis of the two clones. The nucleotide sequence analysis revealed that the complete *G13* cDNA was 2613 bp long (Figure 2), which is close to the expected length of the *G13* mRNA of approx. 2.7 kb as predicted from the Northern blot analysis. Translation of the cDNA sequence in three phases revealed that the longest open reading frame starts with the first methionine residue at position 26, and terminates at a stop codon UGA at position 2135. Thus the longest open reading frame is 2110 bp long and would encode a product of 703 amino acids. The 3′ end of the *G13* cDNA encompasses 477 bp of untranslated sequence, which terminates in a polyadenylation signal (AAUAAA) at position 2582 separated by 17 nucleotides from

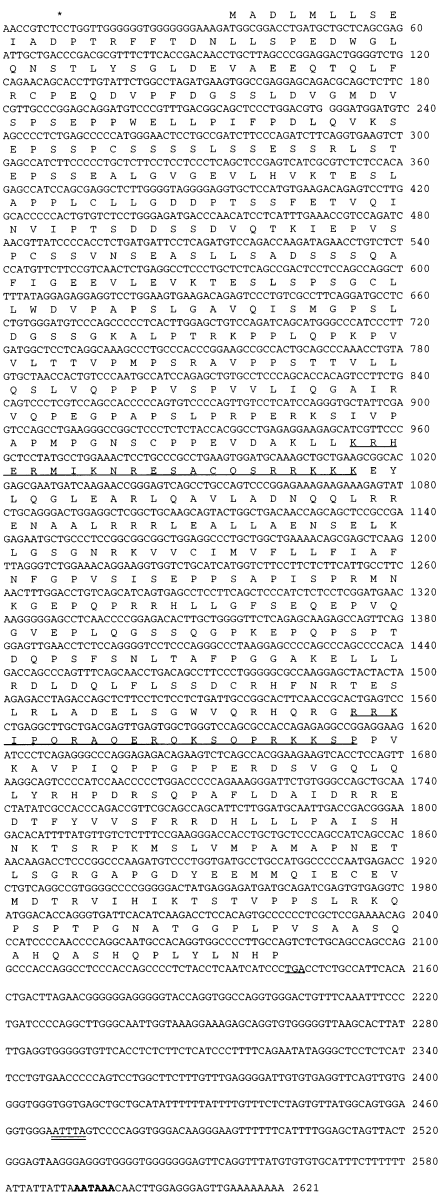
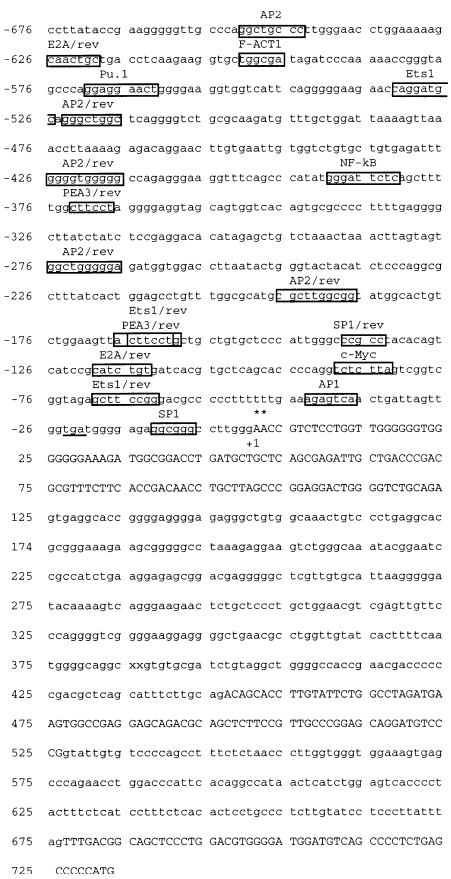
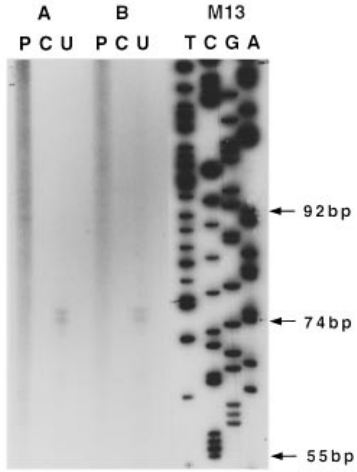
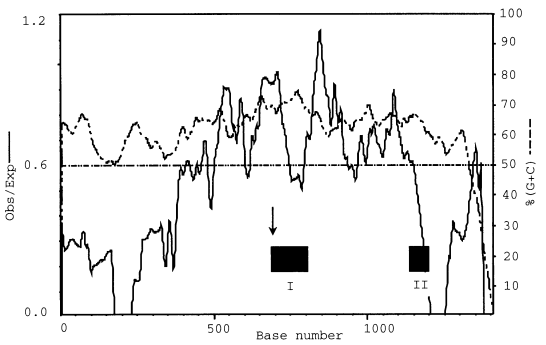


Figure 2 The complete sequence of the cDNA corresponding to the gene *G13*

The transcription initiation site indicated as the first nucleotide was determined by RNase protection analysis. The start of the sequence obtained from the pG13-11a.24 cDNA is indicated by (*). The translation termination codon TGA is underlined. The polyadenylation signal is indicated in bold letters and the signal for mRNA instability is doubly underlined. The potential nuclear localization signals in the G13 protein sequence are underlined.







an 8 bp poly(A) tail. The untranslated sequence also contained U-rich clusters and a single AUUUA motif (Figure 2).

About 1.4 kb of sequence from the 5′ end of the *G13* gene was obtained from genomic clones (Figure 3, top panel). The genomic clones (pG13-GEM S/M2 and pG13-GEM.S/M7), which were approx. 750 bp long and contained 69 bp of the known sequence of exon I (compared with the cDNA sequence), were used to generate anti-sense riboprobes for RNase protection analysis of the U937 RNA. After electrophoresis of the protected fragment, products of 75 and 76 nucleotides were observed that were not present in the control tRNA lane (Figure 3, middle panel). These results suggested that the transcription start site was 8 bp upstream of the start of the cDNA sequence; this nucleotide was therefore designated as +1 (Figure 3, top panel). Thus the *G13* mRNA contains a short 5′ untranslated region of 34 nucleotides. An in-frame stop codon was observed 55 bp upstream of the initiation methionine codon.

A 677 bp sequence upstream of the mRNA start site was searched for potential binding sites for various sequence-specific transcription factors by using the TFD SITES dataset file of the GCG package with default parameters. Several potential binding sites for *trans*-acting factors such as AP2, PEA-3, NFkB, Pu.1, GATA-3, E2A, Ets-1, Pu-F and F-Act 1 were found in the upstream region of the *G13* gene (Figure 3, top panel). The sequence at the 5′ end of the gene was also screened for the presence of a CpG island by using the CpGPLOT program. This indicated the presence of a distinct CpG-rich island that spanned approx. 920 bp of the genomic sequence, encompassing approx. 450 bp of the region upstream of the transcription start site and approx. 475 bp of DNA corresponding to the first exon and the following intron (Figure 3, bottom panel).

The G13 protein sequence

The longest open reading frame, beginning with the first methionine residue at nucleotide 34 (from the start of transcription initiation) in the *G13* mRNA predicts a protein product of 703 amino acids with a molecular mass of 76706 Da. The G13 protein is rich in serine and proline, which together account for 25% of the amino acids in the protein. Sequences rich in proline (P), glutamic acid (E) or aspartic acid, serine (S) and threonine (T), and flanked by basic residues, are signals for proteolytic degradation and are termed PEST sequences [32]. Several possible PEST sequences might also be present in the G13 protein. These include regions encompassing residues 50–88, 88–106, 106–123, 125–164 and 164–199. A charge plot analysis reveals an overall

Figure 3 RNase mapping and CpGPLOT analysis of the gene *G13*

Top panel: sequence of the 5′ region of the gene *G13*. The exonic sequences (exons I, II and part of III) are shown in capital letters. The intronic sequence and the sequence of the region upstream of the transcription start site are shown in lower-case letters. The position of the initiation site of the *G13* mRNA, denoted by +1 below the corresponding nucleotide, was determined by RNase protection analysis as a doublet, which is indicated by (*) above the start site. The upstream stop signal (TGA) in-frame with the translation initiation codon (nucleotides 34–36) is underlined. The putative binding sites for transcription factors are shown by open boxes. Middle panel: identification of the transcription start site by RNase protection analysis. Anti-sense RNA probes were prepared from the pG13-GEM.S/M2 and pG13-GEM.S/M7 genomic clones by using SP6 (A) and T7 (B) RNA polymerases respectively. The riboprobes were hybridized to U937 RNA (approx. 20 μg; lane U) or tRNA (20 μg; lane C). After digestion of the unhybridized RNA, the samples were subjected to electrophoresis on a 6% acrylamide/7 M urea gel. Aliquots of the RNA probes generated with T7 and SP6 polymerases were run in lane P. Products of 76 and 75 nucleotides were observed in lane U that were not observed in lanes C or P. An M13mp10 sequence ladder was run as a size marker. Bottom panel: CpGPLOT analysis of the gene *G13*. The screening parameters were based on the method described in [9]. The positions and extents of exons I and II are indicated by black boxes and the arrow denotes the transcription start site.

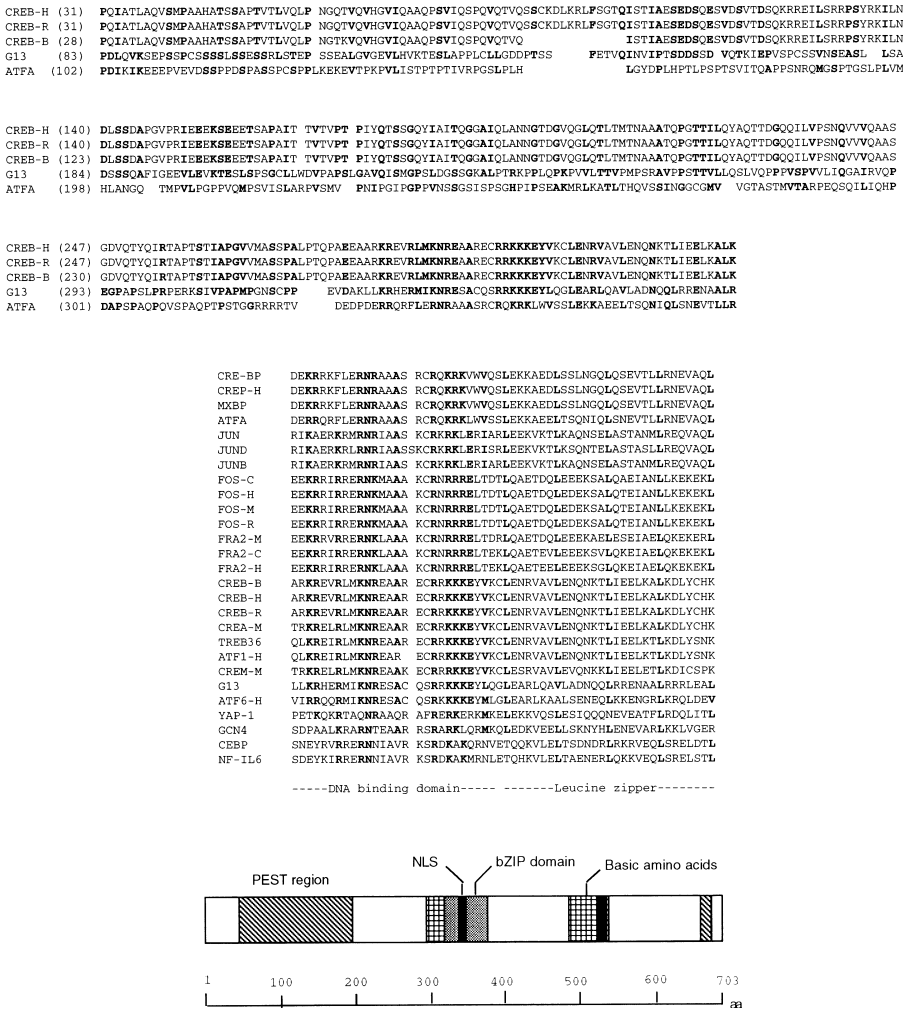


Figure 4 The sequence of the putative G13 protein and its similarity to the bZIP proteins

Top panel: alignment of the G13 protein with the CREB and ATFA proteins. The G13 protein sequence (residues 83–375) is aligned with CREB-H (human; residues 31–333), CREB-R (rat; residues 31–333), CREB-B (bovine; residues 28–317) and ATFA (residues 102–382). The residues that are identical or represent conserved substitutions between the G13 protein and the rest of the proteins are shown in bold letters. Middle panel: residues 325–381 of the G13 protein are aligned with the bZIP domains of a variety of proteins. The residues involved in DNA binding and dimerization via the leucine zipper are indicated with dashes. Identical amino acids and conserved substitutions between G13 and 70% of the other sequences are indicated in bold letters. Bottom panel: schematic representation of the primary structure of the G13 protein. The putative bZIP domain, bipartite NLSs, the region rich in basic amino acid residues and the region rich in proline, serine and threonine residues (PEST region) are indicated by different patterns.

negatively charged region spanning amino acids 1–200 (Asp + Glu = 18%; Arg + Lys = 4%) and two largely positively charged areas between residues 300 and 400 (Arg + Lys = 24%; Asp + Glu = 12%) and between residues 500 and 550 (Arg + Lys = 34%; Asp + Glu = 8%) (Figures 2 and 4).

The N-terminal sequence does not have any potential signal sequence, according to the criteria described in [33] and [34]. The hydrophobicity analysis of the G13 protein with the algorithm PEPWINDOW of the STADEN package at a window size of 19 revealed several short hydrophobic stretches (six to eight residues) in the region of residues 1–300 and a hydrophobic peak spanning residues 388–413 (results not shown). The MOTIFS search program of the GCG package displayed possible signals for N-glycosylation, lipid modification and phosphorylation of serine, threonine and tyrosine in the primary sequence of the G13 protein. There is a plausible bipartite nuclear localization signal (NLS) between residues 327 and 347 comprising two basic amino acids (Lys and Arg), a spacer of any 14 residues and a basic cluster at residues 343–347 (Arg-Arg-Lys-Lys-Lys) [35]. The sequence at residues 343–350 (Arg-Arg-Lys-Lys-Lys-Glu-Tyr-Leu) might also contain a nuclear targeting capability without the assistance of the dibasic region 14 amino acids upstream. A potential bipartite NLS might also be present between residues 527 and 547 (Figures 2 and 4).

A comparison of the G13 protein sequence with those in the National Biomedical Research Foundation (NBRF), Protein Identification Resource (PIR) and SWISSPROT protein databases revealed that the N-terminal half of the protein was 23% identical with human CREB over a stretch of approx. 300 residues, between residues 83 and 375. When conservative substitutions were included, the similarity increased to approx. 33%. CREB is a sequence-specific DNA-binding protein belonging to the family of transcription factors that use dimerization to control function [36]. The putative G13 protein also revealed 20% identity over this stretch of amino acids with a related protein, termed the activating transcription factor, from

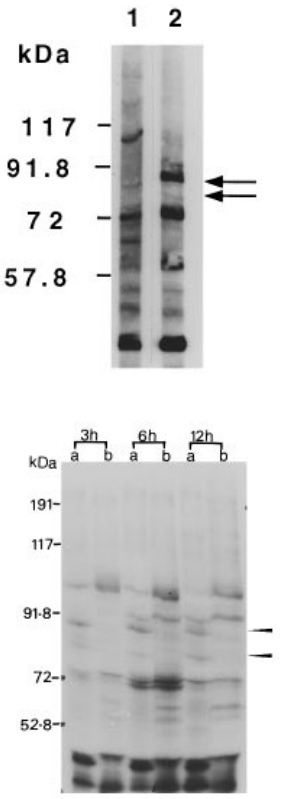


Figure 5 Expression of the G13 protein in mammalian cells

Upper panel: immunoblot analysis of the G13 protein with the enhanced chemiluminescence system. Western blots of U937 cells were prepared and probed with the purified antiserum (lane 2). The preimmune serum was employed as a negative control (lane 1). The anti-G13 antiserum recognized a major polypeptide of approx. 86 kDa and a minor polypeptide of approx. 78 kDa (indicated by arrows). These bands were not seen in lane 1. The positions of size markers are shown at the left-hand side. Lower panel: immunoprecipitation analysis of the G13 protein with radiolabelled antigen. The cell lysates prepared after a radioactive pulse of 3, 6 or 12 h were incubated with the purified anti-G13 serum (lane a) or the preimmune serum (lane b). The immune complexes formed were purified on Protein A–Sepharose beads and subjected to SDS/PAGE. The specific products with molecular masses of 79 and 88 kDa in lanes (a) are indicated by arrows. The positions of size markers are shown at the left-hand side.

human (ATFA) (Figure 4, top panel). The sequence between residues 83 and 375 of the G13 protein and the corresponding sequence from CREBs (human, bovine and rat) and ATFA were aligned by using the multiple-sequence-alignment package AMPS [27,28]. G13 revealed a score of 9.03 with human and rat CREB, 7.51 with bovine CREB and 7.14 with ATFA. An alignment score of more than 5 S. D. units has been suggested as indicating significant similarity [27]. Thus G13 revealed significant similarity to the CREB/ATF family of proteins.

The DNA-binding domain of the CREB/ATF family is composed of a region rich in basic amino acids followed by a bZIP domain [37,38]. The bZIP domain is shared by the G13 protein between amino acid residues 327 and 381. The amino acids corresponding to the putative G13 bZIP domain were searched separately through the various protein databases. The sequence corresponding to residues 327–381 of the G13 protein and the bZIP domains of the leucine zipper class of proteins were also analysed by using the AMPS programme to align these domains and to determine the significance of this similarity. This domain came up with significant identities (20–61 % identity; 3.64–17.79 SD alignment scores) with a large number of proteins of the bZIP family (Figure 4, middle panel).

The C-terminal half of the G13 protein did not show any significant similarities to known proteins, even when it was searched through the database separately from the N-terminal half. Similarly, the N-terminal 80 amino acids of the G13 protein did not disclose any identities with known proteins.

Expression of the G13 polypeptide

The region spanning nucleotides 1372–2136 in the *G13* cDNA and encoding amino acid residues 451–703 constitutes the most hydrophilic region of the putative G13 protein. This region was cloned into the pGEX-2T expression vector. A fusion protein of approx. 54 kDa (26 kDa from GST and approx. 28 kDa from G13 peptide) was expressed on induction of the correctly oriented recombinants with isopropyl β-D-thiogalactoside. Thrombin cleavage of this protein could not be accomplished successfully irrespective of whether the protein was attached to the glutathione–agarose matrix or whether it was in solution after elution. This was probably due to the masking of the cleavage site by unusual folding of the fusion protein.

The antiserum to the GST–G13 fusion protein obtained after the fourth booster dose was checked for its specificity by Western blot analysis. After purification on the GST-linked affinity resin, the polyclonal antibodies specifically recognized only the G13 polypeptide and not GST (results not shown). Western blots of U937 cell lyates (with the purified antiserum) revealed a strong specific signal at a position corresponding to an approx. 86 kDa protein and a minor signal corresponding to a protein of approx. 78 kDa molecular mass (Figure 5, upper panel, lane 2). The molecular mass of the putative G13 protein is 76706 Da as calculated from its amino acid composition. The strong specific product of approx. 86 kDa might represent the G13 protein that has been post-translationally modified. The minor product of approx. 78 kDa is most likely to correspond to the unmodified G13 protein.

The immunoprecipitation experiments (with the purified antiserum) revealed a prominent signal of molecular mass approx. 88 kDa, which was obtained in each of the samples pulsed with [^{35}S]methionine for 3, 6 and 12 h (Figure 5, lower panel). A minor product of approx. 79 kDa was also noted. The product of approx. 88 kDa recognized in immunoprecipitation is very close to the approx. 86 kDa product obtained by Western analysis.

The immunofluorescence results with the purified antiserum indicated that the G13 protein is not localized to the nucleus as expected from its similarity to the bZIP class of transcription factors and the presence of the potential nuclear localization signals. Rather, it is present all over the cell, and with hints of membrane staining in some parts. The preimmune serum used as a negative control produced little specific signal in the cell. The secondary antibodies (FITC-conjugated anti-rabbit antibodies raised in goat) were also used to check the level of background fluorescence produced in the absence of specific primary antibodies (Figure 6). They produced very small amounts of background fluorescence.

DISCUSSION

The single-copy gene *G13*, which is located in the MHC class III region approx. 75 kb centromeric of the C4A gene (Figure 1A), encodes a 703-residue polypeptide (Figure 2). The similarity of the putative G13 protein to the CREB/ATF class of transcription factors, which use their bZIP domains for DNA binding, suggests that G13 is probably a novel member of this class of proteins. The degree of identity of G13 with different CREB/ATF family

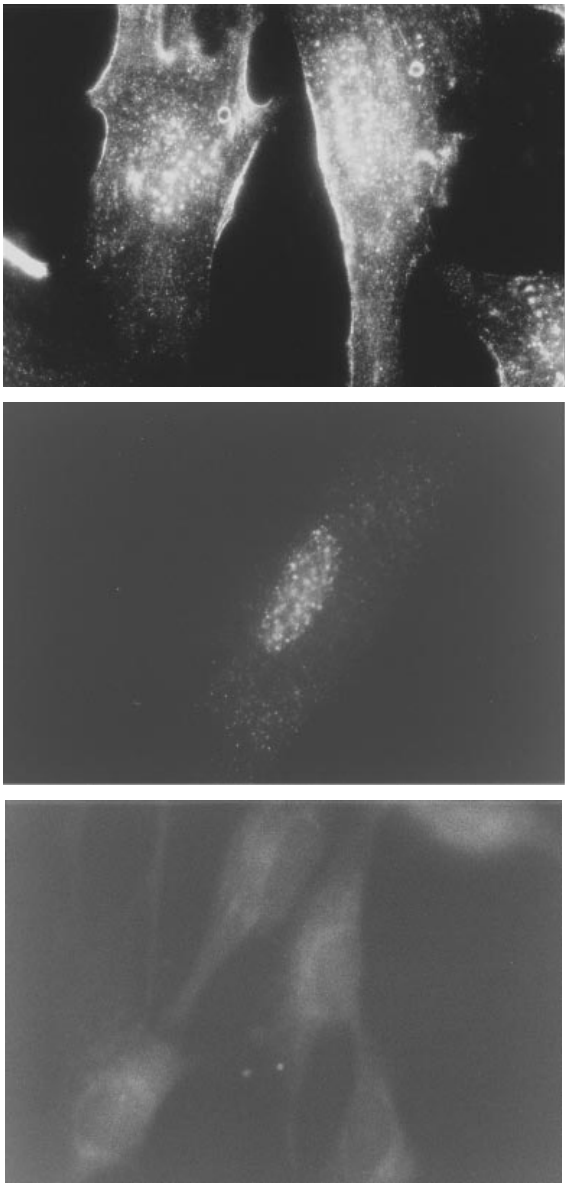


Figure 6 Immunolocalization of the G13 protein in the human fibroblast cell line GM02922

Fluorescent antigen detection was facilitated by treatment of the cells with FITC-labelled secondary antibody after incubation of the cells with the purified anti-G13 serum (top panel) or the preimmune serum (middle panel). The labelled secondary antibody did not show any significant fluorescence (bottom panel) when used on its own.

members varies from 23% identity over a stretch of more than 300 amino acids with CREB, to 27.8% identity over 115 amino acids with cAMP response element modulator. The residues 325–381 display the highest degree of identity (varying between 20% and 61%) with the bZIP domain of transcription factors of the CREB/ATF, AP-1 and related families, suggesting that this domain is probably the bZIP domain of G13 (Figure 4). The degree of similarity is less outside the bZIP domain. This is also true of other known members of the bZIP class, which are dissimilar to each other outside their bZIP domains. It is interesting to note that residues 291–320, preceding the bZIP similarity domain, contain a large proportion of proline residues (27.5%) interspersed with several serine residues. It has been demonstrated that such proline-rich domains are indispensable for transcriptional activation of several families of transcription factors [39].

The bZIP family can also be subdivided into subgroups depending upon the degree of similarity between amino acids in the leucine zipper and basic region. From the comparison of the bZIP domains it is clear that the G13 protein is more closely related to the CREB/ATF family of proteins than are Fos, Jun and others (Figure 4, middle panel). However, throughout the bZIP domain there are several conserved features in the leucine zipper and the basic domain structure that seem important for its proper function, as described below.

The term bZIP was coined for those site-specific DNA-binding proteins that contained a basic region and leucine zipper as their DNA-binding motif [40,41]. The leucine zipper mediates dimerization, putting the two basic regions together in parallel, to form a DNA-binding site. The basic amino acid region is the most conserved region in the bZIP domain. As shown in Figure 4 (middle panel) this region in the G13 protein conforms to the consensus sequence Asn-Xaa-Xaa-(Ala)-(Ala)-Xaa-Xaa-(Cys/Ser)-Arg. Notably, many of the basic residues contacting DNA phosphates in the GCN4–DNA complex [42] are also conserved in the G13 protein. Mutational and physico-chemical studies of several proteins containing leucine repeats have demonstrated that the region is folded into α-helices and mediates dimerization. The bZIP family bind to DNA as homodimers or heterodimers. The sequence in coiled coils follow the heptad pattern $(abcdefg)_n$. Leucine zipper motifs conform to this pattern with the leucine (in 40% of cases) or hydrophobic residues at position 'd'. In G13, which contains four such repeats, leucines are conserved in position 'd' at residues 350, 357 and 378. Another feature of the leucine zipper is that it is rich in charged and polar residues at positions 'e' and 'g', which is also true of the potential bZIP domain of G13.

Indirect immunofluorescence studies on a fibroblast cell line, with antisera raised against the C-terminal half of the G13 polypeptide fused to the GST protein, indicated that the protein was present all over the cell. The G13 protein harbours two potential bipartite nuclear targeting sequences. The NLS at residues 337–347 is part of the putative bZIP domain of the G13 protein. The NLS of the CREB protein is also a part of the basic region of its bZIP domain and has been shown to be functional [43]. Not all nuclear proteins, including transcriptional activators, synthesized in the cytoplasm are immediately transported into the nucleus. Their translocation can be regulated by cytoplasmic factors that mask the NLS and prevent its recognition by the nuclear import machinery [44,45]. The translocation of the bZIP proteins, Jun, Fos and C/EBPb, into the nucleus is known to be regulated [46–48] and this might also be so for the G13 protein.

In the absence of any functional results it is difficult to assign a function to the putative G13 protein. However, the presence of a highly conserved bZIP domain in the G13 protein suggests that it might be a novel member of this class of proteins and is probably involved in functions mediated by interactions with DNA. Unlike the class I and class II regions, the class III region of the MHC does not encode a 'classical molecular structure'. Although the functions of most of the novel genes in the class III region have not been established yet, the similarities of their putative products to other proteins suggest that they encode a variety of different functions [49].

It is noteworthy that most of the novel gene products are not directly linked to immune functions and are also unrelated to each other. This feature is also exhibited in some novel genes identified in the class I and class II regions [50]. However, *G13*

and some novel genes in the MHC seem to have one feature in common: they contain motifs for binding to DNA in their primary amino acid sequence. These include the *Oct-3* (class I region), *RING1* (class II region) and *G17* (class III region) genes. This motif in *RING1* is involved in protein–DNA interactions and perhaps protein–protein interactions [51]. *Oct-3* and *G17* contain a homeodomain in their primary amino acid sequences [52,53]. The homeodomain motif was first recognized in proteins that regulate *Drosophila* development, although it is now clear that it has a broader role in eukaryotic gene regulation [54]. The proteins with bZIP domains are known to be regulators of transcription. They can function constitutively or in a regulatable manner through post-translational modifications in response to external stimuli. Many bZIP proteins show a cell-specific or developmentally regulated pattern of expression and contribute to tissue differentiation [55]. A detailed functional analysis of the G13 protein will have to be performed to establish its role in the cell.

Note added in proof (received 2 August 1996)

After our manuscript was completed we became aware of the paper by Min et al. [56] which describes a novel CREB family gene telomeric of HLA-DR in the HLA complex. This gene corresponds to *G13*. However, the predicted size of the protein described by Min et al. is only 700 amino acids, while that reported here is 703 amino acids. Other differences in the nucleotide and derived amino acid sequences reported by Min et al. from those reported here can be found in the DDBJ, EMBL and GenBank databases under the accession number X98054.

We thank Dr. D. Simmons for providing the U937 cDNA library, Dr. E. Kendall for the cDNA clones, Dr. G. Barton and Dr. T. Day for their help with protein alignments, and Ken Johnson for photographic work. This work was supported by the Commowealth Scholarship Commission, the Radhakrishnan Memorial Trust and the University of Oxford.

REFERENCES

1 Campbell, R. D. and Trowsdale, J. (1993) Immunol. Today **14**, 349–352
2 Sargent, C. A., Dunham, I., Trowsdale, J. and Campbell, R. D. (1989) Proc. Natl. Acad. Sci. U.S.A. **86**, 1968–1972
3 Spies, T., Blanck, G., Bresnahan, M., Sands, J. and Strominger, J. L. (1989) Science **243**, 214–217
4 Kendall, E., Sargent, C. A. and Campbell, R. D. (1990) Nucleic Acids Res. **18**, 7251–7257
5 Fukugawa, T., Sugaya, K., Matsumoto, K., Okumura, K., Ando, A., Inoko, H. and Ikemura, T. (1995) Genomics **25**, 184–191
6 Abderrahim, H., Sambucy, J.-L., Iris, F., Ougen, P., Billault, A., Chumakov, I. M., Dausset, J., Cohen, D. and Le Paslier, D. (1994) Genomics **23**, 520–527
7 Dunham, I., Sargent, C. A., Dawkins, R. L. and Campbell, R. D. (1989) Genomics **5**, 787–796
8 Bird, A. P. (1987) Trends Genet. **3**, 342–347
9 Gardiner-Garden, M. and Frommer, M. (1987) J. Mol. Biol. **196**, 261–282
10 Khanna, A. and Campbell, R. D. (1993) in Proceedings of the 11th International Histocompatibility Workshop (Tsuji, K., Aizawa, M. and Sasazuki, T., eds.), pp. 198–201, Oxford University Press
11 Lee, K. A. W. (1991) Curr. Opin. Cell Biol. **3**, 953–959
12 Karin, M. and Smeal, T. (1992) Trends Biochem. Sci. **17**, 418–422
13 Campbell, R. D. and Milner, C. M. (1994) Curr. Opin. Immunol. **5**, 887–893
14 Bell, G. I., Karam, J. H. and Ruther, W. J. (1981) Proc. Natl. Acad. Sci. U.S.A. **78**, 5759–5763
15 Birnboim, H. C. and Doly J. (1979) Nucleic Acids Res. **7**, 1513–1523
16 Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd edn., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
17 Chirgwin, J. M., Przybyla, A.E, MacDonald, R. J. and Ruther, W. J. (1979) Biochemistry **18**, 5294–5299
18 Chomczynski, P. and Sacchi, N. (1987) Anal. Biochem. **162**, 156–159
19 Feinberg, A. P. and Vogelstein, B. (1984) Anal. Biochem. **137**, 26–267
20 Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. S. H. and Roe, B. A. (1980) J. Mol. Biol. **143**, 161–178
21 Viera, J. and Messing, J. (1987) Methods Enzymol. **153**, 3–11
22 Tabor, S. and Richardson, C. C. (1987) Proc. Natl. Acad. Sci. U.S.A. **84**, 4767–4771
23 Staden, R. (1982) Nucleic Acids Res. **10**, 2951–2961
24 Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. **12**, 387–395
25 Pearson, W. R. and Lipman, D. J. (1988) Proc. Natl. Acad. Sci. U.S.A. **85**, 2444–2448
26 Pearson, W. R. and Miller, W. (1992) Methods Enzymol. **210**, 575–601
27 Barton, G. J. (1990) Methods Enzymol. **183**, 403–428
28 Barton, G. J. and Sternberg, M. J. E. (1990) J. Mol. Biol. **212**, 389–402
29 Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. and Struhl, K. (eds.) (1989) in Short Protocols in Molecular Biology, pp. 154–157, John Wiley and Sons, New York
30 Smith, D. B. and Johnson, K. S. (1988) Gene **67**, 31–44
31 Harlow, E. and Lane, D. (1988) Antibiotics: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
32 Rogers, S., Wells, R. and Rechsteiner, M.(1986) Science **234**, 364–368
33 von Heijne, G. (1988) Biochim. Biophys. Acta **947**, 307–333
34 Gierash, L. M. (1989) Biochemistry **28**, 923–930
35 Dingwall, C. and Laskey, R. A. (1991) Trends Biochem. Sci. **16**, 478–481
36 Ziff, E. B. (1990) Trends Genet. **6**, 69–72
37 Landshultz, W. H., Johnson, P. E. and McKnight, S. L. (1988) Science **240**, 1759–1764
38 Brindle, B. K. and Montminy, M. R. (1992) Curr. Opin. Genet. Dev. **2**, 199–204
39 Mermod, N., O'Neill, E. A., Kelly, J. J. and Tijan, R. (1989) Cell **58**, 741–753
40 Vinson, C. R. and Garcia, K. C. (1992) New Biol. **4**, 396–403
41 McKnight, S. L. (1991) Sci. Am. **264**(4), 54–64
42 Ellenberger, T. E., Brandle, C. J., Struhl, K. and Harrison, S. C. (1992) Cell **71**, 1223–1237
43 Waeber, G. and Haberner, J. H. (1991) Mol. Endocrinol. **5**, 1431–1438
44 Silver, P. A. (1991) Cell **64**, 489–497
45 Davis, L. I. (1992) Curr. Opin. Cell Biol. **4**, 424–429
46 Chida, K. and Vogt, P. K. (1992) Proc. Natl. Acad. Sci. U.S.A. **89**, 4290–4294
47 Roux, P., Blanchard, J. M., Fernandez, A., Lamb, N., Jeanteur, P. and Piechaczyk, M. (1990) Cell **63**, 341–351
48 Metz, R, and Ziff, E. (1991) Genes Dev. **5**, 1754–1766
49 Aguado, B., Milner, C. M. and Campbell, R. D. (1996) in HLA/MHC: Genes, Molecules and Functions (Browning, M. J. and McMichael, A. J., eds.), BIOS Scientific Publishers, Oxford, in the press
50 Trowsdale, J. and Powis, S. H. (1992) Curr. Opin. Cell Genet. Dev. **2**, 492–497
51 Lovering, R., Hanson, I. M., Borden, K. L. B., Martin, S., O'Reilly, N. J., Evan, G. I., Rahman, D., Pappin, D. J. C., Trowsdale, J. and Freemont, P. S. (1993) Proc. Natl. Acad. Sci. U.S.A. **90**, 2112–2116
52 Takeda, J., Seino, S. and Bell, G. I.(1992) Nucleic Acids Res. **20**, 4613–4620
53 Aguado, B. and Campbell, R. D. (1995) Genomics **25**, 650–659
54 Scott, M. P., Tamkum, J. W. and Hartzell, G. W. (1989) Biochim. Biophys. Acta **989**, 25–48
55 Hirst, H. C. (1994) Protein Profile **1**, 123–153
56 Min, J., Shukla, H., Kozono, H., Bronson, S. K., Weissman, S. M. and Chaplin, D. D. (1995) Genomics **30**, 149–156

Received 17 January 1996/22 May 1996; accepted 7 June 1996