Nucleic Acids Research, 1996, Vol. 24, No. 5 **803–807**

RFX proteins, a novel family of DNA binding proteins conserved in the eukaryotic kingdom

Patrick Emery, Bénédicte Durand, Bernard Mach and Walter Reith*

Department of Genetics and Microbiology, University of Geneva Medical School, Centre Médical Universtaire, 1211 Geneva 4, Switzerland

Received November 27, 1995; Revised and Accepted January 19, 1996

ABSTRACT

Until recently, the RFX family of DNA binding proteins consisted exclusively of four mammalian members (RFX1–RFX4) characterized by a novel highly conserved DNA binding domain. Strong conservation of this DNA binding domain precluded a precise definition of the motif required for DNA binding. In addition, the biological systems in which these RFX proteins are implicated remained obscure. The recent identification of four new RFX genes has now shed light on the evolutionary conservation of the RFX family, contributed greatly to a detailed characterization of the RFX DNA binding motif, and provided clear evidence for the function of some of the RFX proteins. RFX proteins have been conserved throughout evolution in a wide variety of species, including *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, mouse and man. The characteristic RFX DNA binding motif has been recruited into otherwise very divergent regulatory factors functioning in a diverse spectrum of unrelated systems, including regulation of the mitotic cell cycle in fission yeast, the control of the immune response in mammals, and infection by human hepatitis B virus.

The RFX family was first defined by the identification, in man and mouse, of three highly homologous site-specific DNA-binding proteins called RFX1, RFX2 and RFX3 (1,2) (Fig. 1). These proteins were found to share a novel 76 amino acid DNA binding domain that was called the RFX DNA binding motif (1,2). The sequence of a fourth member of the RFX family (RFX4) was subsequently found fused to the oestrogen receptor in two aberrant cDNA clones derived from a human breast tumour (2,3) (Fig. 1). Until recently, these four mammalian proteins were the only ones in which the RFX DNA binding motif had been identified. Moreover, the strong homology existing between their DNA binding domains (Fig. 2, 76–96% in pairwise comparisons) revealed little about the nature of this novel DNA-binding motif. Finally, although RFX1 was clearly shown to be a cellular transactivator that is used by the highly pathogenic hepatitis B virus (4,5), little was known about the cellular functions of RFX proteins; candidate genes that may be controlled by RFX1 are the c-myc gene (6) and the ribosomal protein L30 gene (7), while for RFX2, RFX3 and RFX4 no potential target genes have yet been reported. In the past year, however, four additional members of the RFX family have been identified. These are RFX5 in man and mouse (8), sak1 in *Schizosaccharomyces pombe* (9), a gene (ScRFX) in *Saccharomyces cerevisiae* and a gene (CeRFX) in *Caenorhabditis elegans* (Fig. 1). The identification of these RFX genes has permitted a precise definition of the consensus motif for this novel DNA binding domain, and has shed new light on the evolutionary conservation and functional importance of RFX proteins in a surprisingly diverse range of biological systems.

RFX5 is the 75 kDa subunit of a nuclear complex called RFX (8,10–12). This RFX complex is a transcription factor that is essential and highly specific for the expression of MHC class II genes, the Ii chain gene and DM genes (8,10–12). These genes play a key role in the immune system because they control the presentation of foreign antigenic peptides to CD4+ helper T lymphocytes. RFX5 is thus a crucial regulator of the immune response. Mutations disrupting the human RFX5 gene (8) result in MHC class II deficiency (also referred to as the bare lymphocyte syndrome), a debilitating primary immunodeficiency that is due to a complete absence of MHC class II gene transcription in all cell types and tissues (reviewed in refs 12–14).

Sak1 is the first member of the RFX family discovered in a non-mammalian organism. It was isolated on the basis of its ability to suppress mutants of the cAMP dependent protein kinase (cAPK) pathway in *S.pombe* (9). Sak1 is an essential regulatory gene in the life cycle of *S.pombe*. It appears to function downstream of cAPK to allow cells to exit the mitotic cycle and enter either stationary phase or the pathway leading to sexual differentiation. The target genes of sak1 are not known.

The genes, ScRFX and CeRFX, are of unknown function, present in the genomes of *S.cerevisiae* and *C.elegans*, respectively. We identified these genes in the EMBL data library by means of a search for homology with the RFX DNA binding motif. ScRFX corresponds to an 811 amino acid open reading frame containing a centrally placed RFX DNA binding motif (Fig. 1). Three exons of the CeRFX gene were identified (Fig. 1); the RFX DNA binding motif is contained within two exons separated by a small 66 bp intron, and a third exon was localized by virtue of the fact that it contains additional regions homologous to RFX1–3 (see below).

* To whom correspondence should be addressed

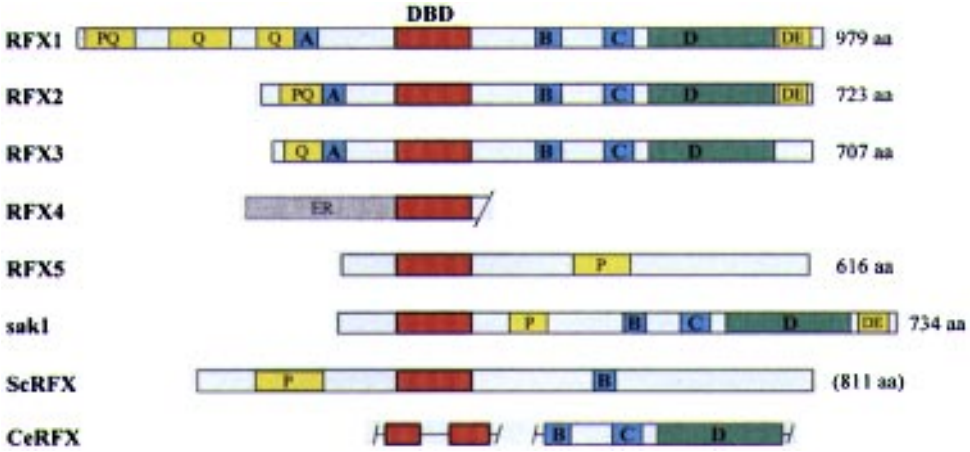
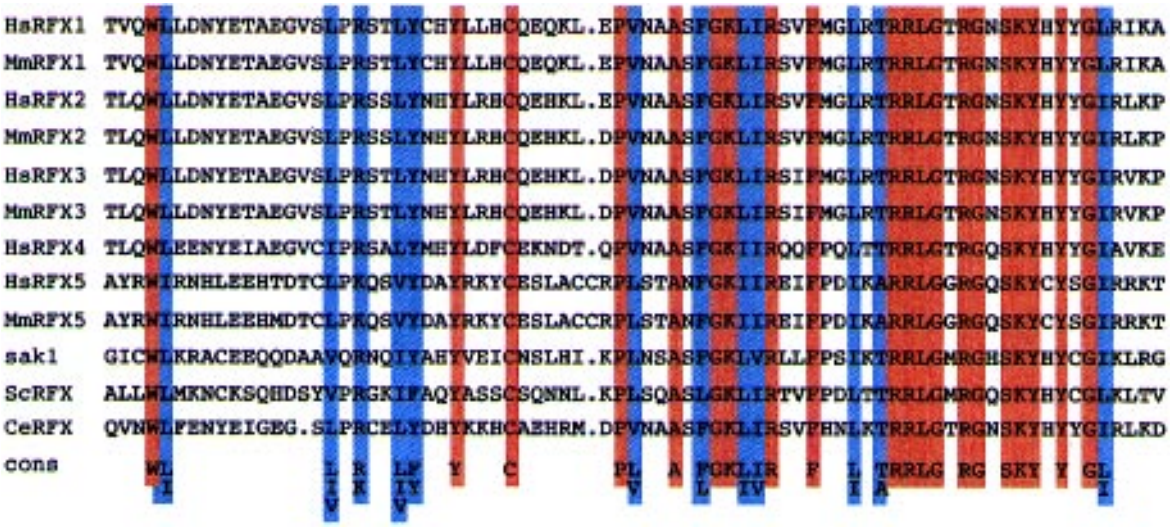


Figure 1. Schematic representation of the RFX1–5 and sak1 proteins, and of the open reading frames corresponding to ScRFX and CeRFX. The RFX DNA binding domain motif (DBD) is indicated in red. The conserved regions A, B and C are indicated in blue. The dimerization domain homology (D) is indicated in green. Regions rich in proline (P), glutamine (Q) or acidic amino acids (DE) are indicated in yellow. The RFX4 DNA binding domain was identified at the C-terminus of two aberrant oestrogen receptor (ER) cDNA clones: the ER portion of these clones is indicated in grey. ScRFX represents a single 811 aa open reading frame. The DNA binding domain of CeRFX is contained in two adjacent exons separated by a 66 bp intron. The B, C and D regions of CeRFX are present on a single exon situated~500 bp downstream of the two DNA binding domain exons. Where known, the size of the proteins is indicated. For ScRFX the size is predicted from the open reading frame present in the genomic sequence. ScRFX (genome sequencing identification name L9470.18) and CeRFX (genome sequencing identification name F33H1.1) were identified by a database search using the BLAST server (16). The MmRFX5 sequence is unpublished. The EMBL accession numbers are HsRFX1 (A20492), MmRFX1 (X76088), HsRFX2 (X76091), MmRFX2 (X76089), HsRFX3 (X76092), MmRFX3 (X76090), HsRFX4 (M69296), HsRFX5 (X85786), sak1 (U19978), CeRFX (Z48783) and ScRFX (U17246).

A



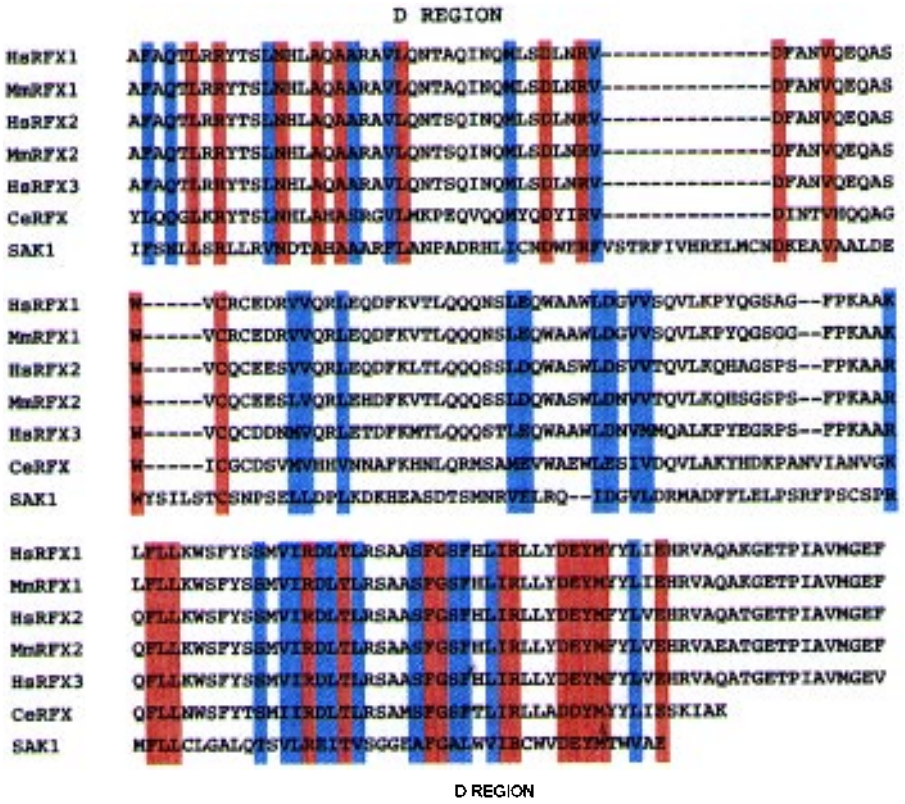
DNA BINDING DOMAIN

B

	HsRFX1	MmRFX1	HsRFX2	MmRFX2	HsRFX3	MmRFX3	HsRFX4	HsRFX5	MmRFX5	SAK1	ScRFX	CeRFX
HsRFX1		100	96	96	96	96	78	57	57	63	62	80
MmRFX1			96	96	96	96	78	57	57	63	62	80
HsRFX2				100	100	100	76	58	58	63	61	83
MmRFX2					100	100	76	58	58	63	61	83
HsRFX3						100	76	58	58	63	61	83
MmRFX3							76	58	58	63	61	83
HsRFX4								57	57	59	57	71
HsRFX5									100	61	57	57
MmRFX5										61	57	57
SAK1											64	61
ScRFX												59
CeRFX												

Figure 2. Conservation of the DNA binding domains of human (Hs) and mouse (Mm) RFX1–5, and homologous sequences found in the *S.pombe* gene sak1, the *S.cerevisiae* gene ScRFX and the *C.elegans* gene CeRFX. (**A**) Amino acid sequence alignment of the DNA binding domains. The consensus sequence is given below. Invariant residues are highlighted in red and similar residues conserved in all sequences are highlighted in blue. (**B**) The percent homology between the DNA binding domains is given for all pairwise alignments. Similar residues were considered to be W/F/Y, K/R, D/E, M/F/I/L/V, S/T/A and N/Q.

A



B

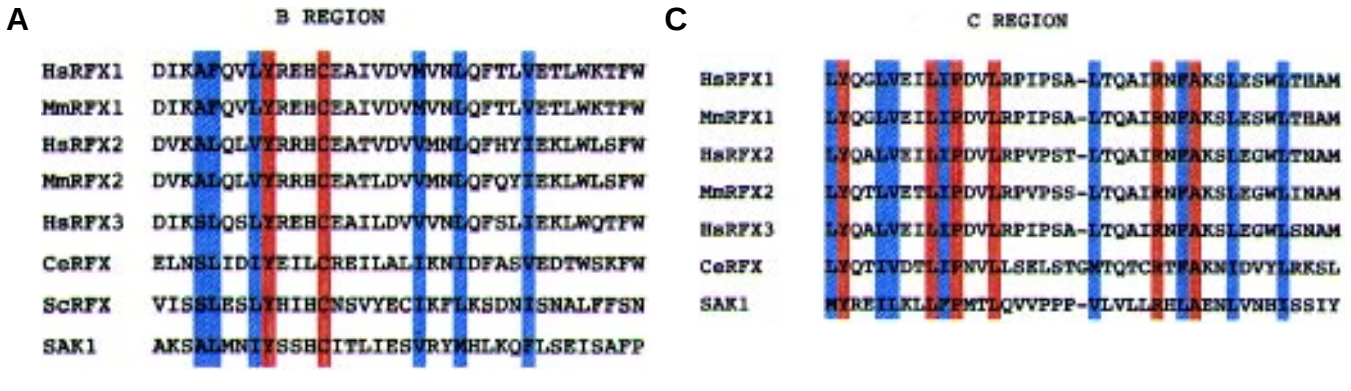
	HsRFX1	MmRFX1	HsRFX2	MmRFX2	HsRFX3	CeRFX	SAK1
HsRFX1		100	93	93	92	67	44
MmRFX1			93	93	92	67	44
HsRFX2				98	94	69	43
MmRFX2					95	70	43
HsRFX3						69	43
CeRFX							47
SAK1							

Figure 3. Conservation of the dimerization (D) domains. The amino acid sequence alignement (**A**) and the percent homology in all pairwise comparisons (**B**) are shown as in Figure 2.

The products, RFX5, Sak1, ScRFX and CeRFX, all contain a 75–77 amino acid segment showing homology with the DNA binding domains of RFX1–4 (Fig. 2). These four new motifs are sufficiently divergent among themselves and with the previously known RFX1–4 sequences to permit a detailed characterization of the RFX DNA binding domain (Fig. 2). The consensus sequence for this domain (Fig. 2A) shows no significant homology to any other known DNA binding motif. It is characterized by 20 invariant and 12 similar amino acids. Among these, aromatic residues (W, F, Y), basic residues (K, R), hydrophobic residues (I, L, V) and four glycine residues are particularly prominent. The majority of the conserved amino acids, particularly the invariant residues, are clustered in the C-terminal half of the domain, which consequently exhibits an overall homology that is considerably greater than that observed in the N-terminal half.

Despite the strong conservation of the DNA binding motif, RFX proteins are likely to have different target site specificities. This is already clear for the RFX proteins for which target site specificity has been analysed in detail. Optimal target sites for the three most closely related proteins, RFX1–3, are inverted repeats, referred to as either EF-C or MDBP motifs, that are present in several cellular genes as well as in the enhancers of polyomavirus, cytomegalovirus and hepatitis B virus (see refs 2,4 and references therein). On the other hand, RFX1–3 bind with lower affinity the X box motifs of MHC class II promoters, sequences that do not show a perfect match to the consensus EF-C/MDBP site (2,15). The opposite is observed for the complex of which RFX5 is a subunit. The optimal known target sites for this complex are the MHC class II X box motifs rather than EF-C/MDBP sites (11). In addition, RFX1–3 complexes exhibit the peculiar and characteristic feature of being able to bind to certain EF-C/MDBP sites only when they contain methylated CpG dinucleotides (see refs 2,4 and references therein). This dependence on methylation of certain target sites is not observed for the complex containing RFX5 (11).

RFX1–3 bind DNA as homo- or heterodimeric complexes. The domain responsible for dimerization has been mapped to a conserved C-terminal region in RFX1–3 (Fig. 1, region D) (1,2). Two of the new RFX genes, sak1 (9) and CeRFX, contain a region exhibiting significant homology to this dimerization domain (Fig. 3), suggesting that they also bind as homo- or heterodimers.



B

B REGION

	HsRFX1	MmRFX1	HsRFX2	MmRFX2	HsRFX3	CeRFX	ScRFX	SAK1
HsRFX1		100	83	83	92	56	36	44
MmRFX1			83	83	92	56	36	44
HsRFX2				94	83	50	36	42
MmRFX2					83	50	36	42
HsRFX3						56	39	39
CeRFX							33	42
ScRFX								50
SAK1								

D

C REGION

	HsRFX1	MmRFX1	HsRFX2	MmRFX2	HsRFX3	CeRFX	SAK1
HsRFX1		100	95	88	93	65	48
MmRFX1			95	88	93	65	48
HsRFX2				95	93	68	48
MmRFX2					88	70	43
HsRFX3						70	48
CeRFX							38
SAK1							

Figure 4. Conservation of region B (**A** and **B**) and C (**C** and **D**). The amino acid sequence alignment (A and C) and the percent homology in all pairwise comparisons (B and D) are shown as in Figure 2.

The dimerization domain of RFX proteins consists of two strongly conserved subregions separated by a region that is more divergent and variable in length (Fig. 3A). The two conserved subregions exhibit no significant homology to any other known proteins, and do not contain any motifs known thus far to be implicated in dimerization or protein–protein interactions. In particular, no potential coiled coil structure is evident. Thus, like the DNA binding domain of RFX proteins, the dimerization domain appears to represent a novel motif. Surprisingly, although RFX5 is known to be a subunit of a multimeric complex, it does not contain a sequence showing clear homology to the dimerization domain of RFX1–3 (8).

Outside of the DNA binding and dimerization domain, two other features characterize RFX proteins. (i) Of three additional homologous regions (A, B and C) first identified in RFX1–3 (2), two (B and C) have also been maintained in other members of the family (Fig. 1). Region B is present in sak1, ScRFX and CeRFX (Fig. 4), and region C is present in sak1 and CeRFX (Fig. 4). Regions B and C have been conserved both in their sequence and in their positions relative to the DNA binding and dimerization domains (Fig. 1). (ii) RFX1–3, RFX5, sak1 and ScRFX all contain regions that share no sequence homology but are rich in proline, glutamine or acidic amino acids (Fig. 1). These features are characteristic of transcription activation domains, suggesting that RFX proteins are transcription factors. This has in fact been demonstrated to be the case for RFX1 (4) and RFX5 (8,11).

To evaluate the evolutionary relationship between the different RFX genes, a phylogenetic tree was derived from their DNA binding domain sequences (Fig. 5). Together with the presence and extent of homology of the other conserved regions, the analysis of the phylogenetic tree raises a number of interesting issues. First, it shows that the mammalian RFX1–3 genes form a subfamily of recently diverged genes, which is consistent with the high conservation observed in the A, B, C and D regions. Secondly, among the novel RFX genes, CeRFX appears to be the closest relative of the RFX1–3 subfamily. This is evident both from the sequence of its DNA binding domain and from the strong conservation in the B, C and D regions. Surprisingly, CeRFX seems to have diverged more recently than the two mammalian RFX4 and RFX5 genes. This implies that duplication events leading to a multigene family must have preceded the emergence of vertebrates, and indicates that invertebrates such as *C.elegans* should have additional RFX genes, possibly homologues of RFX4 and RFX5. Thirdly, the DNA binding domains of the two yeast

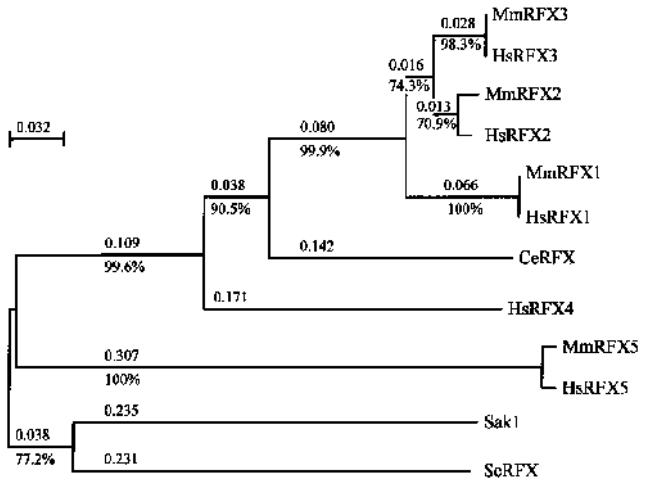


Figure 5. The phylogenetic tree of the RFX DNA binding domain sequences was generated with the clustalW program (17) which uses the neighbour joining method (18). The horizontal branch lengths (indicated above the branches) are proportional to the percent divergence. The reproducibility of the tree was determined using 1000 bootstrap replicas. The percent of replicas containing a given branch is indicated below the branches.

genes, which diverge most from those of RFX1–3, are nevertheless associated with regions retaining homology to the B, C or D regions, suggesting that these regions were present in an ancestral gene. Finally, perhaps the most intriguing finding is that RFX5 is predicted, on the basis of its strong divergence, to be a very ancient member of the family (Fig. 5), yet has acquired a highly specialized role in a system that has evolved only relatively recently, namely the mammalian immune system. It should be mentioned however that the phylogenetic tree was made under the assumption that all the genes in the family diverged at identical rates. Consequently, if RFX5 has diverged more rapidly than the other RFX genes, it may well have emerged more recently than predicted by the phylogenetic tree.

In conclusion, RFX proteins constitute a family of DNA binding proteins that has been conserved in *S.cerevisiae*, *S.pombe*, *C.elegans*, mouse and man, and must thus have appeared over a billion years ago. RFX proteins are characterized by several highly conserved features, among which the most prominent are novel DNA binding and dimerization motifs that are clearly distinct from those found in other known DNA binding proteins. Finally, RFX proteins function as regulatory factors in a wide variety of unrelated systems, including regulation of the mitotic cell cycle in fission yeast, the control of the immune response in mammals, and infection by human hepatitis B virus. It thus appears likely that the RFX family will turn out to be as widespread and functionally important as the well known zinc finger, homeodomain, basic-leucine-zipper and basic-helix–loop–helix families of DNA binding proteins.

REFERENCES

1 Reith,W., Herrero Sanchez,C., Kobr,M., Silacci,P., Berte,C., Barras,E., Fey,S. and Mach,B. (1990) *Genes Dev.*, **4**, 1528–1540.
2 Reith,W., Ucla,C., Barras,E., Gaud,A., Durand,B., Herrero Sanchez,C., Kobr,M. and Mach,B. (1994) *Mol. Cell. Biol.*, **14**, 1230–1244.
3 Dotzlaw,H., Alkhalaf,M. and Murphy,L.C. (1992) *Mol. Endocrinol.*, **6**, 773–785.
4 Siegrist,C.A., Durand,B., Emery,P., David,E., Hearing,P., Mach,B. and Reith,W. (1993) *Mol. Cell. Biol.*, **13**, 6375–6384.
5 Garcia,A.D., Ostapchuk,P. and Hearing,P. (1993) *J. Virol.*, **67**, 3940–3950.
6 Reinhold,W., Emens,L., Itkes,A., Blake,M., Ichinose,I. and Zajac-Kaye,M. (1995) *Mol. Cell. Biol.*, **15**, 3041–3048.
7 Safrany,G. and Perry,R.P. (1995) *Eur. J. Biochem.*, **230**, 1066–1072.
8 Steimle,V., Durand,B., Barras,E., Zufferey,M., Hadam,M.R., Mach,B. and Reith,W. (1995) *Genes Dev.*, **9**, 1021–1032.
9 Wu,S.-Y. and McLeod,M. (1995) *Mol. Cell. Biol.*, **15**, 1479–1488.
10 Reith,W., Satola,S., Herrero Sanchez,C., Amaldi,I., Lisowska-Grospierre,B., Griscelli,C., Hadam,M.R. and Mach,B. (1988) *Cell*, **53**, 897–906.
11 Durand,B., Kobr,M., Reith,W. and Mach,B. (1994) *Mol. Cell. Biol.*, **14**, 6839–6847.
12 Reith,W., Steimle,V. and Mach,B. (1995) *Immunol. Today*, **16**, 539–546.
13 Griscelli,C., Lisowska-Grospierre,B. and Mach,B. (1993) In Rosen,F.S. and Seligman,M. *Immunodeficiencies.* Harwood Academic Publishers, Chur, pp. 141–154.
14 Mach,B., Steimle,V. and Reith,W. (1994) *Immunol. Rev.*, **138**, 207–221.
15 David,E., Garcia,A.D. and Hearing, P. (1995) *J. Biol. Chem.*, **270**, 8353–8360.
16 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
17 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
18 Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.*, **4**, 406–425.