THE MOUSE EPISIALIN (MUC1) GENE AND ITS PROMOTER:
RAPID EVOLUTION OF THE REPETITIVE DOMAIN IN THE PROTEIN

Hans L. Vos,*Yne de Vries and John Hilkens

Division of Tumor Biology, Netherlands Cancer Institute,
Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands

Received September 24, 1991

We have cloned the Muc1 gene of the mouse, encoding the murine equivalent of human episialin (also known as EMA or PEM), a mucin-like glycoprotein that is overexpressed in carcinoma cells. The extracellular domain of the mouse protein, that mainly consists of tandem repeats, contains 16 repeats of variable length and sequence, whereas the human protein usually contains between 30 and 90 nearly identical repeats. The murine repeats contain more potential O-glycan side chains and this may result in a more extended conformation of the murine protein. The transmembrane and cytoplasmic domains of the protein show about 90% conservation. The promoter region shows many conserved regions that could function as transcription factor binding sites.

Human episialin (PEM; EMA; DF3 antigen; NPGP; PAS-O; CA15.3 antigen) is a large sialoglycoprotein, that is encoded by the MUC1 gene (for reviews, see 1,2). It is expressed at the apical side of glandular epithelial cells. The amino acid sequence of human episialin, as determined from cDNA clones, predicts that it is a type I transmembrane protein with a large extracellular domain (3-6). The size of the extracellular domain ranges between approximately 1000 and 2200 amino acids as a result of variations in the number of tandem repeats, that each consist of twenty amino acids. Both the repeats and the adjacent amino acid sequences contain a large number of serine and threonine residues that carry sialylated O-linked glycans, that are typical of mucins. The high density of sialylated carbohydrate side chains on mucin molecules restricts the flexibility of the protein backbone as a result of steric hindrance between adjacent glycans (7). Therefore, mucins are rigid, rod-shaped molecules, that project high above the glycocalyx of the cell. Since the extracellular domain of

*fax (31) 20 6172625

0006-291X/91 $1.50

episialin mainly consists of such a large mucin-like domain, its normal function probably is to prevent access of macromolecules to the apical cell membrane. Expression of episialin at this membrane might also be required for the formation and maintenance of the ducts by preventing stable interactions between the cells. Recently, our group has shown that overexpression of episialin, as occurs on tumor cells, does indeed affect the capacity of cells to aggregate (8), suggesting a role of episialin in invasion and metastasis.

The structural requirements for the functioning of episialin can be assessed by sequencing the MUC1 gene of another species. Not only the sequence of the transcribed region is of interest, but that of the promoter as well, since many human carcinoma cells overproduce mRNA encoding episialin (9), suggesting that the overproduction of episialin is caused by increased transcription. The identification of conserved elements in the promoter can serve as a first step in studying regulation of the gene both in normal and in tumor cells.

Materials and Methods

Cloning and sequencing. The mouse episialin gene was isolated from a genomic library made from Balb/c spleen DNA, that had been cloned in bacteriophage lambda gt10 using partially digested EcoRI fragments (10). This library was screened with a ^{32}P-labeled human cDNA clone that encoded the majority of the non-repetitive part of the episialin message, but lacked the repeats. Hybridization was performed according to standard procedures (11). The blots were washed in 2x SSC, 0.1% SDS at 58°C for 1.5 hours. A single positive clone was identified containing two EcoRI fragments of 2.1 and 11 kb, that were separately subcloned in pEMBL18 (12). Relevant parts of these subclones were sequenced using the dideoxy chain termination method and the Multiwell system (Amersham). The sequence of the entire episialin coding region was determined on two strands. Sequence results were analyzed using the GCG program (13) and the entire sequence was compared with the Genbank database (release 68) using the FastA program (14).

Generation and amplification of episialin cDNAs. An mRNA preparation of lactating mammary glands was reverse transcribed using AMV reverse transcriptase and amplified using the polymerase chain reaction method with two pairs of primers, chosen in such a way that all splice junctions were spanned by the two resulting PCR products, named A and B. The sequences of these primers were: A1 5'gctaggatcctgttcaccaccaccatg 3', A2 5' gccggaattctggcaatagtgctgtggc 3', B1 5'gctaggatccactcagccttcagtgccaag 3', and B2 5' gccggaattcagaccgccaaagctgccccaa 3'. The relative position of these primers is shown in Fig. 1. Following digestion with BamHI and EcoRI, that recognize the underlined sites, the two PCR fragments were subcloned and sequenced.

Northern and Southern blot analysis. The agarose gels (1%) for separation of RNA contained 6.6% formaldehyde. Northern and Southern blots were made, hybridized and washed according to standard procedures (11).

Results

The structure of the murine episialin gene.

Following screening of a genomic library with a human episialin cDNA probe, a single phage was obtained, that contained two EcoRI fragments of 2.1 and 11 kb, which were subcloned separately. Nucleotide sequencing of relevant parts of these clones showed that the entire murine Muc1 gene and long stretches of flanking sequences were present on the 11 kb EcoRI fragment. Subsequently, the nucleotide sequences of the promoter region, the coding regions and the introns (except parts of intron 6) were determined and compared to the mouse genomic sequence (15) and to human sequences (3-6, 16-18). We find few discrepancies between the two murine sequences in the regions where they overlap, with the notable exception of one additional triplet in repeat 5 in our sequence (see below). The entire nucleotide sequence has been deposited in the Genbank database (accession numbers M76179 and M77226) and is not shown here. Most exon-intron junctions could be predicted on the basis of the strong sequence similarity between the mouse and human genes and their location was confirmed by sequencing cDNA clones. The coding region is interrupted by six introns (Fig. 1), as is the case for human gene.

The coding sequence.

Like the human protein, mouse episialin is predicted to be a type I transmembrane protein. The level of similarity of the various domains to their human counterparts is highly variable (Fig. 2). The extracellular domain of the protein is not very conserved, with the exception of some parts between the

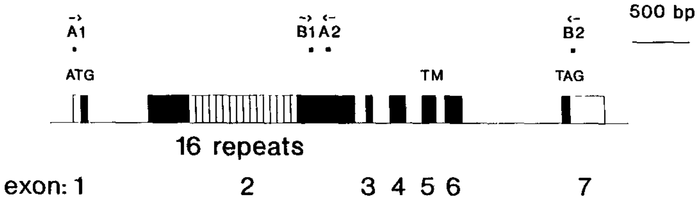

Figure 1.
Structure of the murine mouse episialin gene. The exons are indicated by horizontal bars, the coding regions are shown filled in black, with the exception of the repeats present in exon 2 that are depicted by the open blocks. The locations of the startcodon (ATG), the transmembrane (TM) domain and the stopcodon (TAG) are indicated, as well as the positions and orientations of the PCR primers.

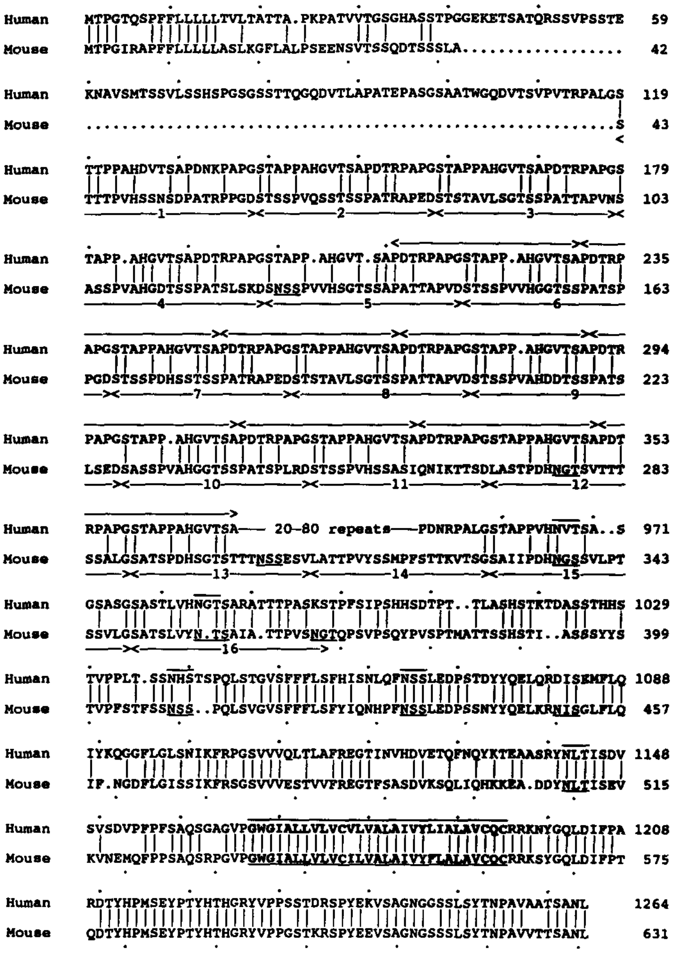
Human MTPGTQSPFFLLLLLTVLTATTA.PKPATVVTGSGHASSTPGGEKETSATQRSSVPSSTE 59
Mouse MTPGIRAPFFLLLLLASLKGFLALPSEENSVTSSQDTSSSLA.................. 42
Human KNAVSMTSSVLSSHSPGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTSVPVTRPALGS 119
Mouse ....S 43
Human TTPPAHDVTSAPDNKPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGS 179
Mouse TTTPVHSSNSDPATRPPGDSTSSPVQSSTSSPATRAPEDSTSTAVLSGTSSPATTAPVNS 103
1
2
3
Human TAPP.AHGVTSAPDTRPAPGSTAPP.AHGVT.SAPDTRPAPGSTAPP.AHGVTSAPDTRP 235
Mouse ASSPVAHGDTSSPATSLSKDSNSSPVVHSGTSSAPATTAPVDSTSSPVVHGGTSSPATSP 163
4
5
6
Human APGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPP.AHGVTSAPDTR 294
Mouse PGDSTSSPDHSSTSSPATRAPEDSTSTAVLSGTSSPATTAPVDSTSSPVAHDDTSSPATS 223
7
8
9
Human PAPGSTAPP.AHGVTSAPDTRPAPGSTAPPAHGVTSAPDTRPAPGSTAPPAHGVTSAPDT 353
Mouse LSEDSASSPVAHGGTSSPATSPLRDSTSSPVHSSASIQNIKTTSDLASTPDHNGTSVTTT 283
10
11
12
Human RPAPGSTAPPAHGVTSA 20–80 repeats PDNRPALGSTAPPVHNVTSA..S 971
Mouse SSALGSATSPDHSGTSTTTNSSESVLATTPVYSSMPFSTTKVTSGSAIIPDHNGSSVLPT 343
13
14
15
Human GSASGSASTLVHNGTSARATTTPASKSTPFSIPSHHSDTPT..TLASHSTKTDASSTHHS 1029
Mouse SSVLGSATSLVYN.TSAIA.TTPVSNGTQPSVPSQYPVSPTMATTSSHSTI..ASSSYYS 399
16
Human TVPPLT.SSNHSTSPQLSTGVSFFFLSFHISNLQFNSSLEDPSTDYYQELQRDISEMFLQ 1088
Mouse TVPFSTFSSNSS..PQLSVGVSFFFLSFYIQNHPFNSSLEDPSSNYYQELKRNISGLFLQ 457
Human IYKQGGFLGLSNIKFRPGSVVVQLTLAFREGTINVHDVETQFNQYKTEAASRYNLTISDV 1148
Mouse IF.NGDFLGISSIKFRSGSVVVESTVVFREGTFSASDVKSQLIQHKKEA.DDYNLTISEV 515
Human SVSDVPFPFSAQSGAGVPGWGIALLVLVCVLVALAIVYLIALAVCQCRRKNYGQLDIFPA 1208
Mouse KVNEMQFPPSAQSRPGVPGWGIALLVLVCILVALAIVYFLALAVCQCRRKSYGQLDIFPT 575
Human RDTYHPMSEYPTYHTHGRYVPPSSTDRSPYEKVSAGNGGSSLSYTNPAVAATSANL 1264
Mouse QDTYHPMSEYPTYHTHGRYVPPGSTKRSPYEEVSAGNGSSSLSYTNPAVVTTSANL 631

Figure 2.

Sequence comparison of the human and murine episialin amino acid sequences. Identical amino acid residues in both sequences are connected by rods. Eight human repeats are shown and are overlined, and the position of the remaining ones is indicated. The mouse repeats are indicated in a similar manner. The start of the human repeats is arbitrary, that of the mouse protein have been chosen in such a way that there is a maximum number of repeats (sixteen). The homology between human and murine sequences in the repetitive regions is very low and several other, slightly different alignments might be possible. The putative N-linked glycosylation sites and the transmembrane region have been indicated as well. Each tenth amino acid in every sequence is indicated by a dot, except in the repetitive region. The numbering of the human episialin sequence is based on an episialin molecule containing 40 repeats.

repeats and the transmembrane domain. There are ten consensus-sites for N-linked glycosylation present in the extracellular domain of murine episialin, some of which are located within the repeat region. There are only five sites in the human protein, all located C-terminal from this region and their position has been conserved in the mouse (Fig. 2).

Ninety percent of the residues in the transmembrane domain of episialin are conserved. The last two cysteines in this domain are ideally positioned at the inside of the plasma membrane for the covalent linkage of a fatty acid residue (19), as has been previously proposed for human episialin (3).

The 69 amino acid cytoplasmic domain of episialin is very well conserved between the two species (87%). It is probably involved in the routing of the molecule during internalization and subsequent recycling (Litvinov et al., submitted), an event that has also been described for ASGP1, a similar mucin in the rat (20). Tyrosine residues in the cytoplasmic domain of some transmembrane receptors are involved in the process of internalization (21-23) and all seven tyrosine residues present in the cytoplasmic tail of episialin are conserved.

The promoter region and the expression pattern.

The nucleotide sequence of the promoter region of the mouse episialin gene is in general conserved with respect to its human counterpart (Fig. 3). Detectable mRNA levels were present in lactating mammary glands, uterus, lung, and, surprisingly, muscle (Fig. 4). Except for this latter tissue, the tissue distribution of murine episialin mRNA is comparable to that of its human counterpart. The size of the episialin message is about 2.5 kb.

Discussion

Our current model of the structure of the episialin molecule predicts that above a certain level of expression, episialin prevents the interaction between molecules on the episialin-producing cell and macromolecules in the immediate surroundings of the cell. We would therefore expect that a large extracellular domain and an extended structure, caused by the presence of a large number of O-linked glycosylation sites, are essential for the function of episialin. However, the lower number of repeats in murine episialin indicates that the extracellular domain of this protein is much shorter than its human counterpart and this might suggest that the length of this domain is not crucial. However, it should be realized that the murine extracellular domain has a predicted length of 70 nm (7) and still extends far above the glycocalyx, which has a thickness of only 10 nm (7). In addition, high-field NMR studies have established that a single non-glycosylated repeat unit of human episialin contains at least one beta turn (24). This would predict a zig-zag structure for human episialin and the overall length of the molecule would critically depend on the bend angle of the turn. Murine

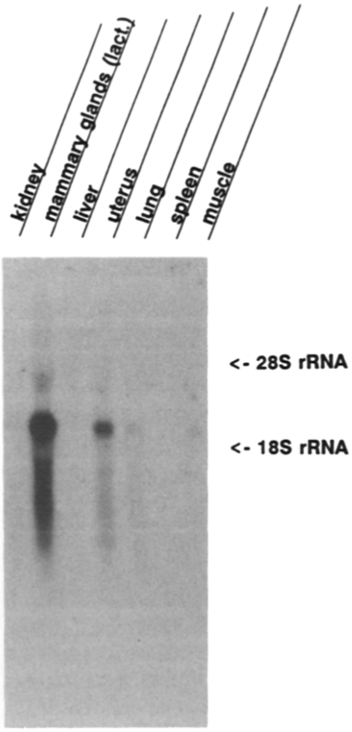

Figure 3.
Episialin mRNA expression in various murine tissues. 20 μg of RNA was used in each lane. The gel was a 1% agarose-formaldehyde gel, and the Northern blot was probed with a genomic DNA fragment containing exons 5 and 6 of the murine episialin gene.

episialin may therefore adopt a more extended structure as a result of the higher density of O-linked glycans or the lower proline content, that both may decrease the bend angle.

The amino acid sequence of the repeats is only poorly conserved, but both the murine and human repeats are rich in serine and threonine residues, that may serve as O-linked glycosylation sites. There are twice as many of these residues in the murine repeats, that number only 16, whereas most human alleles contain between 30 and 90 repeat units. The murine repeats are much more variable in length and in sequence than the human repeats. The high rate of change in the murine repeats, that vary in length between 20 and 22 amino acids, is illustrated by the presence of an extra alanine residue in repeat 5. It is absent in the sequence of Spicer et al. (15), that shows several additional amino acid changes in this region. This suggests that the repetitive domain is evolving very rapidly and that the protein backbone merely serves as a scaffold for the attachment of O-linked glycans.

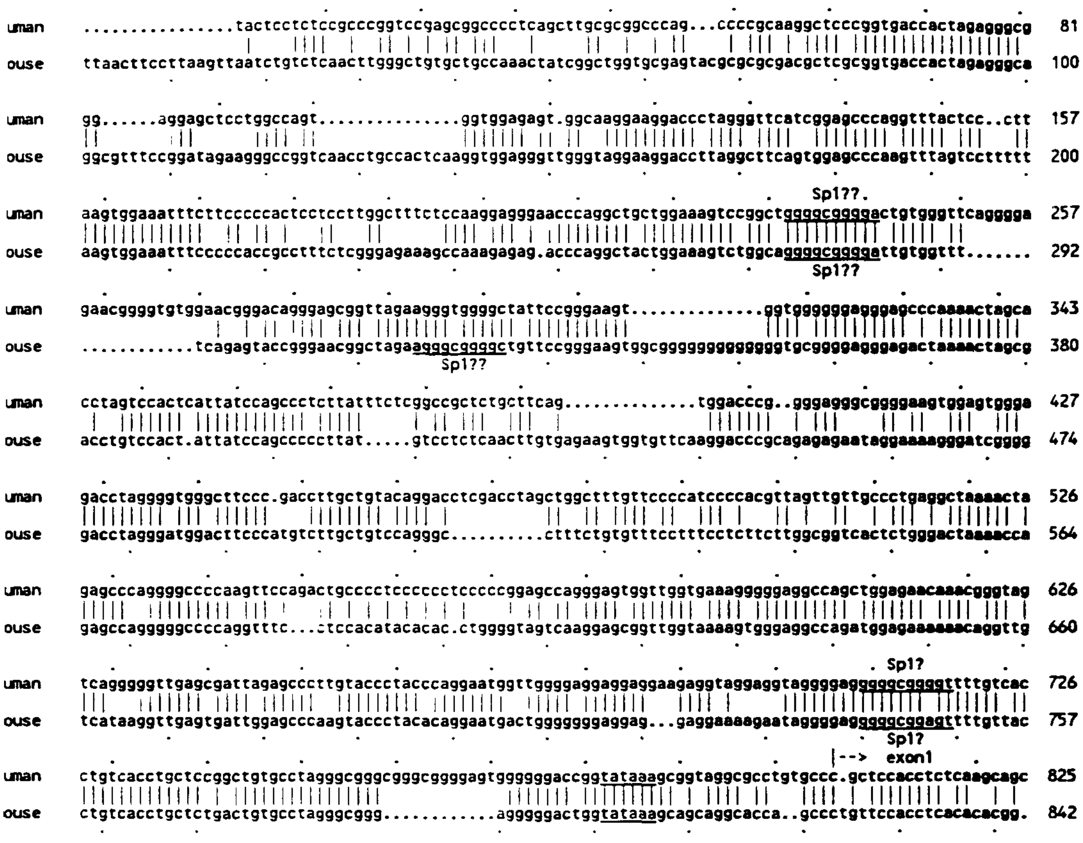

Figure 4.

Sequence comparison of the promoter regions of the human and mouse genes. The human sequence is a consensus sequence, that is based on all available promoter sequences (16; 17; 18; and our own results). However, this consensus sequence only starts at position 382 in the human sequence. The sequence upstream of this position is mainly based on Lancaster et al. (18). However, there are several differences between the sequences of refs. 17 and 18, the only two remaining sequences covering this region. Possible Sp1 sites that show at least a 9 out of 10 match with the consensus sequence $(G/_T)(G/_A)GGCG(G/_T)(G/_A)(G/_A)(C/_T)$; less favored nucleotides shown in small case (27)) have been underlined. The TATAAA-box and the ATG startcodon are similarly indicated. A comparison of the complete genes is available upon request.

The first ten repeats are clearly the result of a relatively recent duplication event involving five repeat units. As a result, repeat unit 3 is nearly identical to repeat unit 8, whereas the repeats 2 and 4 also show a strong similarity, both in length and in sequence, to repeats five units downstream. In view of these results it seems likely that the human repeats, that are almost identical to one another, are the result of a recent expansion of a single repeat type. Similar evolutionary changes in the repetitive domain of the involucrin gene in primates have recently been documented (25).

A proteolytic cleavage event, that occurs in the endoplasmic reticulum, has been described for human episialin (8). The two cleavage products remain associated, probably via non-covalent interactions (8). The cleavage site has not yet been determined exactly, but it should be located between positions 1101 and 1119 (IKFR...FREG) in human episialin (Fig. 2; 8). This region is relatively well conserved in the mouse protein, since most non-identical positions show conservative substitutions. Another region of possible functional significance is a conserved, relatively hydrophobic region that is bordered by two N-linked glycosylation sites and that is characterized by a high number of phenylalanine residues (residues 1033-1053 in the human episialin sequence). The hydrophobic nature of this part suggests that it might be located on the inside of a globular subdomain of this otherwise highly elongated protein or that it is shielded in some other way from the hydrophilic environment, e.g. by interacting with the cleaved carboxyl-terminal domain or with another protein.

The promoter region of the episialin gene shows a relatively high level of conservation between the two species. The TATAAA-box, that serves as the binding site for transcription factor TFIID (26), is completely conserved (Fig 3). One possible Sp1 binding site, that consists of a perfect consensus sequence (27), is present in both promoters about 80 bp upstream of the TATAAA-box. In both promoters, there are several additional motifs present, that contain the GGGCGG core binding site of the transcription factor Sp1, but that deviate from the full-length consensus. There are many additional conserved stretches of nucleotides in the promoter region and these might represent other transcription factor binding sites. Few of these elements correspond to potential transcription factor binding sites identified before (17,18). This underscores the need for functional assays to establish the significance of these promoter elements. From preliminary experiments using CAT assays we have established that sequences that are located more than 250 bp upstream of the transcription start site are still of importance for promoter function and this is reflected in the continuing high level of homology upstream of this region. However, we have evidence that there is another gene not far upstream of the episialin gene, that shows some similarities to the gene for human thrombospondin 1 (manuscript in preparation in collaboration with Dr. P. Bornstein). It is therefore unclear at the moment if conserved elements present far upstream of the episialin gene are important for its transcriptional regulation. This should be resolved using the CAT assays.

Acknowledgments

We would like to thank Dr. Elly Holthuizen for the gift of the genomic DNA library, Marja van Meijer for a gift of mouse RNA preparations, Dr. Sandra Gendler for communicating her results prior to publication and Dr. Marjolijn Ligtenberg for critical reading of the manuscript. H.L.V. is supported by a grant from Centocor BV, Leiden, the Netherlands.

References

1. Hilkens, J. (1988). Cancer Rev. 11-12, 25-54.
2. Taylor-Papadimitriou, J., Lalani, E.-N., Burchell, J., and Gendler, S.J. (1990).J. Nucl. Med. Allied Sci. 34, 144-150.
3. Ligtenberg, M.J.L., Vos, H.L., Gennissen, A.M.C., Hilkens, J. (1990). J. Biol. Chem. 265, 5573-5578.
4. Gendler, S.J., Lancaster, C.A., Taylor-Papadimitriou, J., Duhig, T., Peat, N., Burchell, J., Pemberton, L., Lalani, E.-N., and Wilson, D. (1990). J. Biol. Chem. 265, 15286-15293.
5. Lan, M.S., Batra, S.K., Qi, W.-N., Metzgar, R.S., and Hollingsworth, M.A. (1990). J. Biol. Chem. 265, 15294-15299.
6. Wreschner, D.H., Hareuveni, M., Tsarfaty, I., Smorodinsky, N., Horev, J., Zaretsky, J., Kotkes, P., Weiss, M., Lathe, R., Dion, A., Keydar, I. (1990). Eur. J. Bioch. 189, 463-473.
7. Jentoft, N. (1990). Trends Biochem. Sci. 15, 291-294.
8. Ligtenberg, M.J.L. (1991) Ph. D. thesis, Univ. of Amsterdam.
9. Zaretsky, J.Z., Weiss, M., Tsarfaty, I., Hareuveni, M., Wreschner, D.H., and Keydar, I. (1990). FEBS Lett. 265, 46-50.
10. Davis, M.M., Calame, K., Early, P.W., Livant, D.L., Joho, R., Weissman, I., and Hood, L. (1980). Nature 283, 733-739.
11. Maniatis, T., Fritsch, E., and Sambrook, J. (1982) Molecular cloning (A laboratory manual). Cold Spring Harbor, NY.
12. Dente, L., Cesareni, G., and Cortese, R. (1983). Nucl. Acids Res. 11, 1645-1655.
13. Devereux, J., Haeberli, P., and Smithies, O. (1984). Nucl. Acids Res. 12, 387-395.
14. Pearson, W.R., and Lipman, D.J. (1988). Proc. Natl. Acad. Sci. USA 85, 2444-2448.
15. Spicer, A.P., Parry, G., Patton, S.P., and Gendler, S.J. (1991). J. Biol. Chem. 266, 15099-15109.
16. Abe, M., Siddiqui, J., and Kufe, D. (1989). Biochem. Biophys. Res. Comm. 165, 644-649.
17. Tsarfaty, I., Hareuveni, M., Horev, J., Zaretsky, J., Weiss, M., Jeltsch, J.M., Garnier, J.M., Lathe, R., Keydar, I., and Wreschner, D.H. (1990). Gene 93, 313-318.
18. Lancaster, C.A., Peat, N., Duhig, T., Wilson, D., Taylor-Papadimitriou, J., and Gendler, S.J. (1990). Biochem. Biophys. Res. Comm. 173, 1019-1029.
19. Veit, M., Kretzschmar, E., Kuroda, K., Garten, W., Schmidt, M.F.G., Klenk, H.-D., and Rott, R. (1991). J. Virol. 65, 2491-2500.
20. Hull, S.R., Sugarman, E.D., Spielman, J., and Carraway, K.L. (1991). J. Biol. Chem. 266, 13580-13586.

21. Chen, W.-J., Goldstein, J., and Brown, M.S. (1990). J. Biol. Chem. 265, 3116-3123.
22. Peters, C., Braun, M., Weber, B., Wendland, M., Schmidt, B., Pohlmann, R., Waheed, A., and von Figura, K. (1990). EMBO J. 9, 3497-3506.
23. Ktistakis, N., Thomas, D., and Roth, M.G. (1990). J. Cell Biol. 111, 1393-1407.
24. Tendler, S.J.B. (1990). Biochem. J. 267, 733-737.
25. Djian P., and Green, H. (1991). Proc. Natl. Acad. Sci. USA 88, 5321-5325.
26. Davison, B.L., Egly, J.-M., Mulvihill, E.R., Chambon, P. (1983). Nature 301, 680-686.
27. Briggs, M., Kadonaga, J., Bell, S., and Tjian, R. (1986). Science 234, 47-52.