Human mucin genes: genomic organization and expression of MUC4, MUC5AC and MUC5B

N. Porchet†§, P. Pigny†, M.-P. Buisine†, V. Debaillieu†, P. Degand†, A. Laine† and J.-P. Aubert†

*Unité 377 INSERM, Place de Verdun, 59045 Lille cedex, France, †Laboratoire de Biochimie et de Biologie Moléculaire de l’Hôpital C.Huriez, CHRU de Lille, 59037 Lille cedex, France, and §Laboratoire de Biochimie, USN A, CHRU de Lille, 59037 Lille cedex, France

Introduction

Mucins are glycoproteins characterized by a high carbohydrate content in which numerous, heterogeneous oligosaccharide chains are linked to the core peptide (called apomucin) through O-glycoside linkages. Included in this definition are the large mucus glycoproteins and smaller membrane ‘mucin-like’ glycoproteins.

Research in the field of apomucin genes was initiated around 1987, when several partial human or animal mucin cDNAs began to be isolated.

In humans, 11 families of tandemly repeated nucleotide sequences that encode mucin or ‘mucin-like’ peptides have so far been cloned from various tissues. MUC1 was isolated from breast or pancreatic tumour cells [1–3], MUC2 from jejunal mucosa as well as from tracheobronchial mucosa [4–6], MUC3 from jejunal mucosa [7], MUC4, SA, SB and SC from tracheobronchial mucosa [8,9], MUC6 from stomach [10] and MUC7 from the submandibular gland [11]. More recently, three additional cDNAs have been cloned from the trachea (pAM1 clone) [12], from nasal polyps (NP3a clone) [13] and from HT-29 MTX (L31 clone) [14]. They have not yet received a gene symbol in the Human Gene Mapping Committee designation of the 'MUC' nomenclature: MUC4 (JER64) corresponds to the mucin gene located on chromosome 3 (q29) and a common number, MUC5, was recommended to designate three cDNA families that are distinct according to their nucleotide sequence characteristics but exhibit a common chromosome localization in 11p15; MUC5A (main clone JER57), MUC5B (main clone JER64) and MUC5C (main clone JER58). The JER47 clone is also able to hybridize to chromosome 13 [19]. The corresponding locus was designated MUC5L.

In order to determine whether MUC5A, 5B and 5C are products of the same gene locus or distinct loci, we studied physical mapping of these cDNAs using the strategy of CpG island analysis by pulse-field gel electrophoresis. The data indicated that MUC5A and MUC5C are part of the same gene (which we have now called MUC5AC), which was distinct from MUC5B. Other convincing arguments that MUC5AC and MUC5B are distinct genes were provided by tissue/cell expression studies on their mRNA.

We have proceeded to study these genes in the order of their discovery. Thus, the genomic organization of MUC5B is better understood than that of MUC5AC or MUC4.

MUC5B

The MUC5B gene was first characterized by four cDNA clones (JER28 and 57, JUL7 and 10)
mapped to chromosome 11p15. These clones were all found to contain degenerate 87-bp tandem repeats that encode non-repetitive peptides [22]. Numerous deletions or insertions in an otherwise virtually perfect 87-bp tandem repeat create many shifts in reading frame which completely destroy the repetitive peptide structure. Thus, this peptide appeared to be composed of alternate hydrophilic (threonine-rich) and hydrophobic (proline- and histidine-rich) domains that probably differ in the extent to which they are glycosylated.

This exceptional structure curiously does not appear to be unique to MUC5B and has been reported for another airway mucin [12]. In this case, a cDNA was found to contain degenerate 41-bp tandem repeats encoding two types of peptide sequences.

In addition to studying cDNAs from MUC5B, we began chromosome walking using EMBL4 phage and pWE15 cosmids libraries to determine the genomic organization of MUC5B. Two overlapping cosmid clones were obtained: the CpG island, a 0.55-kb CG-rich domain situated at the 5' end, the whole tandem repeat domain and, at the 3' end, several potential polyadenylation sites. The size of this gene is about 35 kb. In the middle there is an enormous unique exon (about 10.5 kb) containing the degenerate 87-bp tandem repeat motif about 120 times. This encodes a non-repetitive domain of about 3500 amino acids. Moreover, this domain contains some cysteine-rich regions also found in MUC2 and MUC5AC. In the region upstream of this exon are found at least four very small exons (from about 50 bp to 500 bp) that encode unique sequences. Some potential regulatory sequences for Sp1 protein have also been observed, and studies on the mechanisms of regulation of MUC5B gene expression are now in progress.

**MUCSAC**

The MUCSAC gene also mapped to 11p15.3–15.5, where a mucin gene cluster exists near the gene encoding the c oncogene h-ras, extended by about 400 kb, and which also contains MUC2, MUC5B and MUC6. Sequencing of cDNAs clones from MUCSAC (JER47, 58 and 62, JUL32, MAR2, 10 and 11) showed that this gene was characterized by a 24-bp tandem repeat that encodes the consensus peptide motif TTSTTTSAP. In this mucin gene, the tandem repeat domains are interrupted several times with a subdomain encoding a 130 amino acid cysteine-rich peptide [23].

Thus subdomain, comprising 10 cysteine residues, involves invariant positions for most of the cysteines and also for some typical amino acids surrounding these cysteine: tryptophan, tyrosine, proline and glycine. The consensus peptide sequence of this subdomain was also found encoded several times in MUC2 and MUC5B genes. In the MUC5B gene, some of the cysteine-rich-encoding domains are missing.

The tandem repeat domain is shorter than that of MUC5B gene and is seen in an 8-kb PstI fragment in Southern blot. The size of the MUCSAC gene is about 20 kb according to the CpG island restriction map. Chromosome walking is under way but is difficult owing to the great instability of the genomic clones. Genetic polymorphism studies have been reported [24].

**MUC4**

JER64 was the first MUC4 cDNA that we isolated. It contains 39 repeats of 48 bp and encodes a serine/threonine-rich domain that is potentially a glycosylated domain. MUC4 was mapped to chromosome 3 in the region q29 and genetic polymorphism in the number of tandem repeats has been demonstrated [8,16].

Recently, we isolated two overlapping cosmid clones for this gene. Their restriction maps indicate that MUC4 sequences extend about 60 kb. Thus, this gene seems to be larger than MUCSAC (20 kb) and MUC5B (35 kb). We have encountered the same problem as seen with MUCSAC, i.e. of isolating and maintaining genomic clones in phages or in cosmids.

However, during the chromosome walking, we have recently isolated a novel coding sequence within MUC4. This sequence is composed of 15-bp tandem repeats that are perfectly conserved extending for at least 4 kb with a high degree of polymorphism between individuals. We are presently investigating whether these two domains are coded by separate exons.

The expression patterns of this novel MUC4 sequence seem to be identical to those seen with MUC4 probe on Northern blots prepared from a few specimens of bronchus, colon, small intestine and gastric mucosa.

**Expression of MUC5B, MUCSAC and MUC4**

Several techniques are available to study the expression of genes, including immunohistochemistry, Northern blot analysis and in situ hybridization.
hybridization. Each offers advantages and disadvantages when used for studying mucin gene expression, which may explain some of the discrepancies observed. Immunohistochemical analysis of apomucin expression necessitates the isolation of antibodies that are very specific and, at best, whose epitopes are not obscured by glycosylation. Owing to the presence of tandem repeats rich in hydroxylated amino acids, the different mucin genes may share structural homology at the conformational level and so specificity of the antibodies must be carefully evaluated.

With the exception of MUC1, all the other MUC genes exhibit complex patterns on Northern blots. Depending on the tissue or cell line studied, a broad smear or a few distinct RNA transcripts associated with a ‘background or smear’ are generally seen. The significance of this feature of mucin mRNA remains unclear. We are trying in our laboratory to perform experiments (preparation of mucin RNA poly-somes, in vitro transcription of MUCSB genomic sequences, pulse-field gel electrophoresis applied to RNAs) with the aim of throwing some light on this confused subject.

The in situ hybridization technique appears to be a sensitive method of studying the expression of the mucin genes. Indeed, the tandem repeat organization and the choice of oligonucleotide probes specific for the random repeat of each gene is a convenient way of obtaining a large signal. We have used this technique, and the choice of ‘consensus’ oligonucleotides allowed the amplification of the signal by hybridizing a maximum number of small probes all along the same mRNA molecule. Table 1 shows the results we have previously obtained and an update of our data [21].

**MUC5B**

Submaxillary glands and respiratory tract MUC5B mRNAs are strongly expressed in submaxillary glands and in bronchus glands (in situ hybridization [21]). Using Northern blot analysis, we were able to study MUC5B expression on specimens of human trachea and bronchus and

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Expression patterns of mucin genes in human adult mucosae ascertained by in situ hybridization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probes</td>
<td>E, epithelium; GI, glands. Labelling: ++++, very high intensity; ++, high intensity; +, moderate intensity; +, weak intensity; -, absent.</td>
</tr>
<tr>
<td>Human mucosae</td>
<td>MUC4</td>
</tr>
<tr>
<td>---------</td>
<td>-----</td>
</tr>
<tr>
<td>Submaxillary glands</td>
<td>Gl: -</td>
</tr>
<tr>
<td>Bronchus</td>
<td>E: ++</td>
</tr>
<tr>
<td>Fundus</td>
<td>E: +</td>
</tr>
<tr>
<td>Antrum</td>
<td>Gl: -</td>
</tr>
<tr>
<td>Jejunum</td>
<td>E: +</td>
</tr>
<tr>
<td>Ileum</td>
<td>Gl: +</td>
</tr>
<tr>
<td>Colon</td>
<td>Gl: +</td>
</tr>
<tr>
<td>Gall bladder</td>
<td>E: --</td>
</tr>
<tr>
<td>Prostate</td>
<td>Gl: ++</td>
</tr>
<tr>
<td>Endocervix</td>
<td>E: ++</td>
</tr>
<tr>
<td>Gl: ++</td>
<td>Gl: +</td>
</tr>
</tbody>
</table>
the signal seen with the MUC5B probe was intense.

Gall bladder
A strong expression of MUC5B is specifically observed in gall bladder by Northern blot and in situ hybridization. This observation is in good agreement with data obtained by Northern blot analysis of ex vivo gall bladder cells [25]. In these cells, the MUC5B gene appears to be the most highly expressed of the presently known mucin genes. The role of MUC5B gene products could be the binding of hydrophobic ligands by the 'naked' domains, and thus this mucin may contribute to the pathogenesis of cholesterol choledolithiasis [26].

Digestive tract
This gene is also weakly expressed in the colon (by Northern blot analysis), but is not expressed in other digestive tissues (by in situ hybridization studies). However, MUC5B is significantly expressed in the colon cancer cell line HT-29 MTX.

Pancreas
MUC5B expression was negative in normal pancreatic tissue using Northern blot analysis [27]; using in situ hybridization a weak and homogeneous labelling was detected, predominantly in interlobular ductal cells. No significant change in the expression level of MUC5B was observed in pancreatic cancer or in chronic pancreatitis adjacent to pancreas cancer [28].

MUC5AC
Bronchus and stomach
The MUC5AC gene is predominantly expressed in airways and stomach (by Northern blot and in situ hybridization). In bronchial mucosa, labelling is seen in the goblet cells of the epithelium and generally a weaker signal is seen in mucus glands. In gastric mucosa, labelling is present in the goblet cells of the fundus and in the antral epithelia.

Intestinal tract
In normal adult intestinal tract, the MUC5AC gene is not expressed (by Northern blot and in situ hybridization). Interestingly, however, studying mucin expression in fetuses ranging in age from 12 to 27 weeks' gestation, we found a significant expression of this gene at 12 weeks' gestation. This signal disappears at 13 weeks and is not observed again until 27 weeks' gestation. This suggests a possible role for this gene in the development and differentiation of epithelial cells in embryos (L. Devisme, M. P. Buisine, A. Janin, C. Gespach, F. Puech, J. P. Aubert, N. Porchet and B. Gosselin, unpublished work).

MUC5AC is strongly expressed in the colon cancer cell line HT-29 MTX [29] and in recto-sigmoid villous adenomas (RVAs), while normal adult mucosa is negative for this gene expression.

This aberrant expression of MUC5AC in RVAs is more intense in RVAs with low-grade dysplasia than in RVAs with high-grade dysplasia [29a]. This indicates that dysregulation of MUC5AC expression might be an early event in some phenomena occurring during tumorigenesis.

Pancreas
Studies on MUC5AC gene dysregulation in malignant pancreas tissues also support this idea. MUC5AC transcripts are undetectable in normal exocrine pancreas by Northern blot and by in situ hybridization. However, a strong expression of MUC5AC is associated with mucinous hyperplasias and pancreatic cancers [28]. The diagnostic implications of these observations need to be carefully studied on large series with the aim of defining markers useful to pathologists.

Uterine endocervix
The MUC5AC gene was also found to be expressed in uterine endocervix. The expression level analysed by in situ hybridization is not significantly different in biopsies obtained from women at different stages of the ovulatory cycle [30].

Gall bladder
The MUC5AC gene was also studied in gall bladder biliary epithelial cells ex vivo or cultured under various conditions — medium, temperature, presence or absence of stagnant bile — which mimic changes occurring in the management of liver graft during orthotopic liver transplantation. Addition of bile and rewarming of the cells resulted in a marked increase in MUC5AC mRNAs [24].

MUC4
Bronchus and digestive tract
MUC4 was expressed in all the epithelia and in most of the cells, such as the ciliary cells of the
bronchus and enterocytes, except in the gall bladder and the submaxillary glands. In the small intestine of adults, only a few cells were labelled by the MUC4 probe, while this gene is expressed at a higher level and in more cells in the small intestine of fetuses at 12 and 13 weeks’ gestation. MUC4 was not detected from the 18th to the 27th weeks of gestation. This gene is expressed in HT-29-FU [14] but not expressed in HT-29 MTX cells [29], while its expression is not significantly modified in different conditions of gall bladder biliary cell culture [25].

Pancreas

In pancreatic tissues, Northern blot and in situ hybridization studies have confirmed that neoplastic transformation of the exocrine pancreas is accompanied by changes in mucin gene expression [27,28]. MUC4 transcripts were detected in all pancreas cancers examined, whereas they were only detected in a low proportion of cells in normal ducts. As with MUC5AC, MUC4 might be a sensitive marker for cancer precursors.

Uterine endocervix

By studying endocervical biopsies from women at different stages of the ovulatory cycle, we found that the most intense labelling by in situ hybridization is obtained with MUC4 during the luteal phase [30]. In carcinoma of the endocervix, MUC4 alone was detected and was highly expressed (N. Porchet, unpublished work).

A critical issue for the coming years will be to understanding how mucin gene expression is regulated during development, differentiation and in disease states. It will be possible to better elucidate the role of mucins and their associated glycan chains in neoplastic progression and in the metastatic process.

We are very grateful to Dr D. M. Swallow for helpful discussions and reading the manuscript. This work was supported by the Association de Recherche sur le Cancer, the CH et U de Lille (PHRC 1994) and by the Fondation Aupetit.

Biosynthesis and molecular architecture of gel-forming mucins: implications from an amphibian model system

W. Hoffmann and W. Joba
Institut für Molekularbiologie und Medizinische Chemie, Otto-von-Guericke-Universität, D-39120 Magdeburg, Germany

Introduction

Mucins represent a major constituent of mucus gels, which protect delicate epithelial surfaces, e.g. of the alimentary, respiratory and reproductive tracts. Mucus gels serve as a lubricant, as a diffusion barrier and as an anchor for various micro-organisms; furthermore, mucins bind heavy metals, protect the mucosa against proteases and interact with the extracellular matrix, and their secretion is a response to inflammatory-immune processes [1].

However, all these intrinsic properties of mucins are dependent on the formation of a gel-like extracellular matrix with defined viscoelastic properties precisely adjusted to their specific biological role. For example, the rheological properties of cervical mucus vary characteristically during the ovulatory cycle [2]. Furthermore, mucin gels do not form only a rigid matrix; mucin gels rather readily undergo annealing when placed in contact with other gels. This raises the fundamental question on the molecular architecture of mucus gels and the nature of their networks.

Elaborate physicochemical studies of cervical, respiratory and gastric mucus have revealed that these gels consist typically of highly expanded multimeric molecules (Mr about 10^7-10^8) with linear rod-like structures and no major branching points [3,4] which form a three-dimensional network. Generally, two models would describe such a network [5]: (1) a cross-linked network model and (2) an entangled-network model.

These two models predict different kinds of interactions within the polymer gel. The first model implies multiple covalent cross-links among the network. The degree of cross-linking defines the viscoelastic properties of the matrix. The second model assumes aggregation of linear molecules that are held together by physical entanglement and by low-energy bonds but not by covalent cross-links.

Generally, entangled polymer networks would fulfill the criteria that define a dynamic mucus gel. In contrast, a heavily cross-linked static network with restricted diffusion would not explain all the typical features, e.g. annealing. Conversely, covalent cross-links among polymer chains are not favoured in mucins. The question thus arises: How do mucin subunits manage to fulfill all these complex requirements on a molecular level? It is certainly essential to obtain a stiff and highly extended conformation.

Molecular structure and biosynthesis of mucins

Currently, two simple principles are thought to create the unusual length of the molecules found in many mucus gels: (1) formation of rod-like mucin monomer subunits and (2) joining of subunits end to end.

The results of cDNA cloning established only within the last few years suggest that a specific type of domain has been developed for each of these two principles. Thus, mucins show a typical twofold domain structure, i.e. repetitive serine/threonine-rich regions and cysteine-rich modules.

Molecular structure

The serine/threonine-rich regions, which usually also include many proline residues, are responsible for the rigid rod-like structures owing to extensive O-glycosylation; the addition of neutral and acidic oligosaccharide units increases the radius of gyration of mucins by a factor of 2.5-3 when compared with other proteins of similar chain length [6]. Thus, mucin glycoproteins have a characteristic chemical composition that is usually more than 80% carbohydrate by mass. Interestingly, the primary structures of the serine/threonine-rich domains in various mucins do not show pronounced similarities (for compi-