UNIVERSITE DES SCIENCES ET TECHNOLOGIES DE LILLE I

Année 1997 N° d'ordre : 0000

THESE

de DOCTORAT D'UNIVERSITE

pour l'obtention du grade de

DOCTEUR EN SCIENCES DE LA VIE ET DE LA SANTE

présentée par

Jean-Luc DESSEYN

ORGANISATION DU GENE DE MUCINE HUMAINE *MUC5B*.

BASES MOLECULAIRES D'UNE NOUVELLE CLASSIFICATION DES MUCINES

Présentée le 15 octobre 1997 devant la commission d'examen

JURY

Président :	Professeur André Verbert
Rapporteurs :	Docteur Tony Corfield
	Docteur Michel Raymondjean
Membres :	Professeur Steven Ball
	Professeur Pierre Degand
	Docteur Anne Laine

A mes parents,

à mes frères, à ma soeur,

à toute ma famille,

à mes amis.

Je tiens à exprimer ici ma profonde reconnaissance à tous ceux qui m'ont soutenu, aidé et qui m'ont accordé leur confiance tout au long de ces 4 années de recherches.

J'espère avoir mené mes travaux comme ma meilleure partie d'Echecs.

JE TIENS TOUT PARTICULIEREMENT A REMERCIER :

Monsieur le Professeur Pierre Degand ; vous m'avez accueilli au sein de votre laboratoire pour le DEA puis pour préparer cette thèse de sciences. Vous m'avez soutenu dans des moments très importants. J'espère que mon parcours scientifique à venir vous fera toujours honneur.

Monsieur le Docteur Jean-Pierre Aubert ; tu m'as accueilli pour préparer mon DEA en 1993. Tu m'as toujours fait confiance pour mener à bien mes travaux sur le sujet que tu as bien voulu me confier. J'espère avoir su tirer parti de tes connaissances scientifiques et avoir donné le maximum de moi-même à ton groupe.

Madame le Docteur Nicole Porchet ; tes connaissances et nos discussions scientifiques ainsi que tes remarques constructives sur mon travail m'ont beaucoup apporté.

Madame le Docteur Anne Laine ; tu as accepté d'être mon tuteur de DEA et de m'accompagner dans mes premiers pas en recherche. Nous avons travaillé de concert durant cette thèse avec un réel plaisir personnel. Notre symbiose m'a été très profitable. J'admire ta disponibilité de tout instant, ta bonne humeur de tous les jours et ta rigueur scientifique.

Je remercie mes collègues qui m'ont supporté. Qu'ils sachent que c'est grâce à eux que j'ai toujours apprécié m'adonner à mon travail de recherche dans la bonne humeur.

Je remercie pour l'honneur qu'ils me font :

Monsieur le Professeur André Verbert

en acceptant de présider ce jury ;

Monsieur le Docteur Tony Corfield

Monsieur le Docteur Michel Raymondjean

en acceptant d'être les rapporteurs de cette thèse ;

Monsieur le Professeur Steven Ball

Monsieur le Professeur Pierre Degand

Madame le Docteur Anne Lainé, mon directeur de thèse

qui ont bien voulu juger ce travail.

Je tiens à remercier ici Monsieur le Docteur Philippe Lagant et Messieurs les Professeurs Gérard Vergoten et Roman Efremov du Centre de Recherches et d'Etudes en Simulation et Modélisations Moléculaires (Université des Sciences et Technologies de Lille) pour la modélisation du CK de MUC5B.

Ce travail de recherches a été effectué sous la direction scientifique du Dr. Anne Laine, Chargée de Recherches à l'I.N.S.E.R.M., au sein de l'unité lilloise U-377 de "Biologie et Physiopathologie des Cellules Mucipares" dirigée par le Professeur Pierre Degand.

SOMMAIRE

Ce travail a fait l'objet de publications et communications :

Publications :

Identification of a 42 kDa Nuclear Factor (NF1-MUC5B) from HT-29 MTX Cells That Binds to the 3' Region of Human Mucin Gene *MUC5B*. (1996)
Pigny P., Van Seuningen I., **Desseyn J.L.**, Nollet S., Porchet N., Laine A. and Aubert J.P.
Biochem. Biophys. Res. Commun. 220 : 186-191

Human Mucin Gene *MUC5B*: the 10.7 kb Large Central Exon Encodes Various Alternate Subdomains Resulting in a Super-Repeat. Structural Evidence for a 11p15.5 Gene Family. (1997)
Desseyn J.L., Guyonnet Dupérat V., Porchet N., Aubert J.P. and Laine A.
J. Biol. Chem. **272** : 3168-3178

Genomic Organization of the 3' Region of the Human Mucin Gene *MUC5B*. (1997)
Desseyn J.L., Aubert J.P., Van Seuningen, I., Porchet N. and Laine A.
J. Biol. Chem. **272** : 16873-16883

Evolutionary History of the 11p15 Human Mucin Genes. (1997)
Desseyn J.L., Buisine, M.P., Porchet N., Aubert, J.P., Degand, P. and Laine A.
J. Mol. Evol. **45** (sous presse)

VNTR Polymorphism of the Seventh Intron Downstream of the Central Exon of *MUC5B*. (1997)
Desseyn J.L. and Laine A.
Soumis pour publication.

Le travail présenté concernant la région 5' de *MUC5B* fera l'objet d'un article.

Communications écrites (affiches) :

Organisation structurale des régions centrale et N-terminale du gène de mucine humaine *MUC5B*

Desseyn J.L., Guyonnet Dupérat V., Buisine M.P., Pigny P., Aubert J.P., Porchet N. et Laine A. *Association Française pour l'Etude du Foie* (26 et 27 janvier 1996 - Cabourg)

Organisation génomique du gène codant pour la mucine humaine MUC5B et identification dans la région 3' d'éléments spécifiques de la lignée cellulaire HT-29 MTX

Van Seuningen I., **Desseyn J.L.**, Pigny P., Porchet N., Aubert J.P. et Laine A., *Club d'Etude des Cellules Epithéliales Digestives* (30 et 31 janvier 1996 - Toulouse)

Structural Organization of the Central Exon of the Human Mucin Gene *MUC5B* and Evolutionary History of the 11p15 Mucin Gene Family

Desseyn J.L., Guyonnet Dupérat V., Porchet N., Aubert J.P. and Laine A., *4th International Workshop on Carcinoma-Associated Mucins* (juillet 1996 - Cambridge, UK)

Central and 3' end Genomic Organizations of the Human Mucin Gene *MUC5B*

Desseyn J.L., Aubert J.P., Porchet N., and Laine A., *European Society of Human Genetics* (17-20 mai 1997 - Gênes, Italie)

Résumé publié dans *Med. Genetik* 9(2), 1997 **94** ; P4-190

Communications orales :

Organisation génomique de *MUC5B*, gène de mucine humaine

Desseyn J.L., Porchet N., Aubert J.P. et Laine A., *XXIV Forum des Jeunes Chercheurs de la Société Française de Biochimie et Biologie Moléculaire* (8-11 juillet 1997 - Corte, Corse)

Evolutionary history of the 11p15 human mucin genes

Desseyn J.L., Porchet N., Aubert J.P. et Laine A. *Meeting of the mucin club* (26-28 septembre 1997 - Bristol, Grande Bretagne)

Annexes

ANNEXE I : ABREVIATIONS UTILISEES

AMPc	AMP cyclique
ADN	Acide Déoxyribonucléique
ADNc	ADN complémentaire à l'ARNm
AMV	Avian Myeloblastosis Virus
ARN	Acide Ribonucléique
ARNm	ARN messager
ASGP	Ascites Sialomucin Glycoproteins
BEt	Bromure d'éthidium
BSM	Bovin Submaxillary Mucin
CK	Cystine knot
ddNTP	didéoxynucléotide triphosphate
DEPC	Diéthylpyrocarbonate
DNase	Déoxyribonucléase
dNTP	Déoxynucléotide triphosphate
DO	Densité optique
DTT	Dithiothreitol
EDTA	acide éthylène diamine tétraacétique
EGF	Epidermal growth factor
FIM	Frog tegumentary mucin
G3PDH	Glycéraldéhyde-3 phosphate déshydrogénase
GDNF	Glial derived neurotrophic growth factor
IPTG	Isopropyl-β-D-thiogalactopyranoside
kb	kilobase (10^3 bases)
LB	milieu Luria Bertani
MMLV	Moloney-murine leukemia virus
NDP	Norrie disease protein
NGF	Nerve Growth factor
pb	paire de base
PCR	Polymerase Chain Reaction
PSM	Porcin submaxillary mucin
qsp	quantité suffisante pour
RNase	Ribonucléase
SAB	Sérumalbumine bovine

SDS	Sodium Dodécyl Sulfate
SSC	Standard Sodium Citrate
ssDNA	Salmon Sperm DNA
SSPE	Standard Sodium Phosphate EDTA
STE	Sodium Tris EDTA
TBE	Tris Borate EDTA
TE	Tris EDTA
TEA	Tris EDTA Acetate
TGF	Transforming growth factor
Tm	température de fusion
TR	Tandem Repeat
UV	Ultraviolet
VNTR	Variable Number of Tandem Repeat
vWF	facteur von Willebrand
X-Gal	5-bromo-4-chloro-3-indolyl-β-galactoside

ANNEXE II : SYMBOLES UTILISES POUR LES BASES NUCLEIQUES

Les symboles utilisés pour les bases		Signification
A		**A**dénine
C		**C**ytosine
G		**G**uanine
T		**T**hymidine
U		**U**racile (dans l'ARN)
Y	C ou T (U)	p**Y**rimidine
R	A ou G	pu**R**ine
M	A ou C	a**M**ino
K	G ou T (U)	**K**eto
S	G ou C	**S**trong interaction (3 ponts hydrogènes)
W	A ou T (U)	**W**eak interaction (2 ponts hydrogènes)
H	A ou C ou T (U)	pas G; **H** suit le G dans l'alphabet
B	G ou T (U) ou C	pas A; **B** suit le A dans l'alphabet
V	G ou C ou A	ni T, ni U; **V** suit le U dans l'alphabet
D	G ou T (U) ou A	pas C; **D** suit le C dans l'alphabet
N	G, A, C ou T (U)	a**N**y

ANNEXE III : SYMBOLES UTILISES POUR LES ACIDES AMINES

Les acides aminés et leurs symboles			Codons					
A	Ala	Alanine	GCA	GCC	GCG	GCU		
C	Cys	Cystéine	UGC	UGU				
D	Asp	Acide aspartique	GAC	GAU				
E	Glu	Acide glutamique	GAA	GAG				
F	Phe	Phénylalanine	UUC	UUU				
G	Gly	Glycine	GGA	GGC	GGG	GGU		
H	His	Histidine	CAC	CAU				
I	Ile	Isoleucine	AUA	AUC	AUU			
K	Lys	Lysine	AAA	AAG				
L	Leu	Leucine	UUA	UUG	CUA	CUC	CUG	CUU
M	Met	Méthionine	AUG					
N	Asn	Asparagine	AAC	AAU				
P	Pro	Proline	CCA	CCC	CCG	CCU		
Q	Gln	Glutamine	CAA	CAG				
R	Arg	Arginine	AGA	AGG	CGA	CGC	CGG	CGU
S	Ser	Sérine	AGC	AGU	UCA	UCC	UCG	UCU
T	Thr	Thréonine	ACA	ACC	ACG	ACU		
V	Val	Valine	GUA	GUC	GUG	GUU		
W	Trp	Tryptophane	UGG					
Y	Tyr	Tyrosine	UAC	UAU				
B	Asx	Acide aspartique ou Asparagine						
Z	Glx	Acide glutamique ou Glutamine						

Avant-Propos

AVANT-PROPOS

Les fonctions du mucus ont fait l'objet de nombreuses investigations. Le mucus joue, selon l'espèce, l'organe et le moment où il est produit, des rôles parfois contradictoires ; par exemple, si il est indispensable à la survie de la flore intestinale, il est aussi l'un des éléments majeurs responsables de l'épuration de l'épithélium bronchique et donc de l'élimination des bactéries. Une interrogation se pose alors : à quoi peut être attribuée cette différence de fonction ?

La mucine est le constituant majoritaire du mucus. Cette macromolécule est étudiée depuis plusieurs décennies par les biochimistes. Ceux-ci ont focalisé leur attention plus particulièrement sur la fraction *O*-glycosylée très caractéristique de cette molécule. Depuis les premiers travaux de la fin de la dernière décennie qui ont permis de caractériser le premier gène « MUC », de nombreuses équipes se sont tournées vers les techniques de l'ADN recombinant pour étudier la mucine. Ainsi, depuis ces cinq dernières années, la littérature scientifique abonde en articles illustrant le clonage de nouveaux gènes ou de transcrits de gènes MUCs. C'est pourquoi l'introduction de ce mémoire commence par un descriptif des mucines. Le catalogue ne cesse de s'allonger avec la découverte de nouveaux gènes. Ce travail tente de faire ressortir, à partir des données structurales déduites du séquençage des différents gènes, des caractéristiques communes et donc des fonctions biologiques identiques mais aussi des différences structurales aboutissant probablement à des fonctions plus spécifiques de chacune d'elles.

Des neuf gènes identifiés codant des apomucines humaines, quatre (*MUC6*, *MUC2*, *MUC5AC* et *MUC5B*) ont été localisés sur le même chromosome, sur un fragment d'environ 400 kb dans la région télomérique du bras court du chromosome 11. Un transcrit de *MUC2* a été publié par l'équipe de San Francisco alors que notre équipe à Lille étudie les gènes *MUC5AC* et *MUC5B*. Les résultats des études structurales de ces gènes et les travaux présentés dans cette thèse, mettent en exergue

une construction identique pour *MUC5AC*, *MUC2*, et *MUC5B* et conservée dans l'évolution, définissant ainsi une première sous-famille des gènes de mucines. Cette sous-famille comprend toutes les mucines présentant les caractéristiques suivantes : en plus de motifs répétés en tandem caractéristiques de toutes les mucines mais différents d'une mucine à l'autre, elles possèdent un motif d'une centaine d'acides aminés comprenant 10 résidus de cystéine. Cette sous-famille appartient à la superfamille des protéines ayant des modules protéiques du facteur de von Willebrand, et plus loin dans l'évolution, des molécules de l'hémolymphe comme les vitellogénines. Ces mucines sont donc des protéines mosaïques. Enfin, nous avons pu souligner que les membres de cette superfamille possèdent à leur extrémité C-terminale un module appelé noeud cystine, ou encore CK (pour "cystine knot"). La présence de ce module permet d'élargir la mégafamille des protéines à CK, qui comprend déjà des protéines circulantes, des facteurs de croissance et/ou de différenciation et des facteurs morphogènes.

Toutes ces données récentes ainsi que nos travaux sur MUC5B entrouvrent de nouveaux champs d'études des fonctions des mucines qui pourront être explorés dans les prochaines années grâce aux technologies des protéines recombinantes.

Introduction

INTRODUCTION

I LE MUCUS

I.1 Définition

Le mucus est un terme générique désignant les sécrétions physiologiques visqueuses et viscoélastiques, qui recouvrent les épithéliums des vertébrés (poissons, mammifères) et invertébrés (coelomates, mollusques).

Chez les mammifères, le terme de mucus est restreint à la sécrétion qui recouvre et protège les épithéliums des tractus respiratoire, gastro-intestinal et urogénitaux ainsi que les muqueuses, sous-muqueuses et sécrétions des glandes exocrines. Ce mucus constitue l'interface entre l'environnement extérieur et l'épithélium. Les fonctions dévolues au mucus ne sont pas uniquement des fonctions de protection et/ou lubrification. En effet, le mucus a d'autres fonctions biologiques comme par exemple dans la locomotion des gastéropodes. Il joue aussi le rôle de barrière imperméable protégeant ainsi les organismes des changements soudains de la pression osmotique. Chez le ver par exemple, le mucus le pourvoit d'une paroi semi-perméable permettant le passage d'oxygène et de dioxyde de carbone. L'anguille a une peau recouverte d'un gel de mucus; si ce gel est ôté et si l'anguille est placée dans de l'eau distillée alors l'anguille gagne du poids. Au contraire, si l'anguille est placée dans un milieu salin hypertonique alors elle perd du poids. Le mucus est donc un système ancien de protection de l'organisme et ses rôles biologiques se sont diversifiés au cours de l'évolution. Cette adaptation évolutive doit donc être le fruit des variations des compositions biochimiques des différents mucus.

I.2 Constitution

Le mucus, chez les vertébrés, est une sécrétion physiologique qui forme un tapis continu d'épaisseur variable à la surface des épithéliums constituant ainsi une barrière de protection efficace entre l'environnement et la muqueuse. Ce film protecteur est formé de deux phases: une phase aqueuse au contact des cellules, de faible viscosité, proche de celle de l'eau et une phase superficielle, ou phase gel, se caractérisant par une viscosité et une élasticité élevées.

Le mucus est très hétérogène, il est constitué d'eau (95%), d'électrolytes (Na^+, K^+, Ca^{2+}, Mg^{2+} ...) et de composés organiques, majoritairement des glycoprotéines de haute taille moléculaire. Dans le cas de la muqueuse bronchique par exemple, une proportion significative de lipides (lipides neutres glycolipides, et phospholipides), d'IgA sécrétoires, d'IgG, d'albumine, de lactotransferrine, de lysozyme, de protéases et d'antiprotéases (antileucoprotéases, α1–antitrypsine, élafine) mais aussi d'ADN, dans certains cas pathologiques, a été trouvée.

I.3 Fonctions du mucus

Des fonctions différentes peuvent être attribuées au mucus selon l'organe dans lequel il est produit.

I.3.1 Mucus du tractus respiratoire

Dans le tractus respiratoire, le mucus joue un rôle vital dans l'épuration mucociliaire: l'épithélium respiratoire, qui tapisse l'ensemble des voies aériennes supérieures depuis les fosses nasales jusqu'aux bronchioles terminales, est recouvert d'un tapis muqueux. Le mucus forme un tapis continu (d'une épaisseur de 0,5 µm à 2 µm) à la surface de l'épithélium respiratoire, constituant ainsi une barrière de protection efficace entre l'environnement et la muqueuse des voies aériennes (Figure

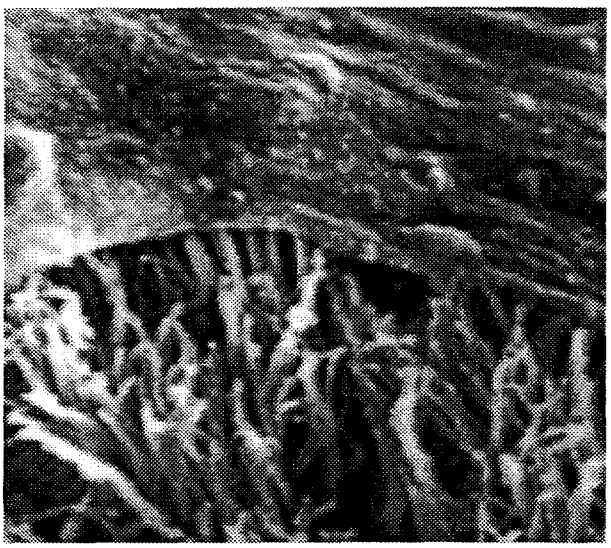
1A). Ce mucus représente donc la première défense de la muqueuse respiratoire, qui est en permanence exposée à de multiples agents agresseurs (gaz et particules toxiques, bactéries, virus et allergènes) inhalés au cours de la respiration. Dans des conditions non-pathologiques, la fine couche de mucus qui recouvre les cellules épithéliales ciliées piège les particules exogènes; la vibration des cils fait remonter ce tapis muqueux jusqu'aux pharynx où il est dégluti.

Des modifications du mucus ont été décelées par exemple au cours de la bronchite où l'on observe une augmentation significative du volume de mucus. Cette augmentation est associée à une multiplication des cellules sécrétant le mucus, les cellules caliciformes, et une diminution des cellules ciliées.

I.3.2. Le mucus du tractus gastro-intestinal

L'épaisseur et la composition de la couche de mucus qui tapisse l'épithélium du tractus gastro-intestinal contribuent à la défense de l'épithélium contre l'adhésion et la pénétration des toxines, et à la protection contre l'invasion bactérienne et les dommages biochimiques. Parallèlement à ce rôle de protection, le gel muqueux régule le mouvement des molécules vers les cellules absorbantes. Le mucus forme un réseau polymérisé dont le support est constitué de glycoprotéines associées à des lipides, par des liaisons faibles non covalentes ou par des liaisons stables et covalentes avec la partie protéique. La composition qualitative en lipides du mucus gastrique est identique chez diverses espèces, que ce soit chez l'Homme, le rat, le chien, et elle est semblable à celle trouvée dans les sécrétions trachéobronchiques. Au sein de ce réseau s'immiscent des débris cellulaires, des nucléotides, des micro-organismes et des macromolécules. Allen et Carroll ont décrit deux formes physiques distinctes: un mucus insoluble formant une couche mince, stable, adhérente à la muqueuse gastroduodénale, et un mucus soluble visqueux, non adhérent à la surface de l'épithélium. Ce mucus "soluble" limiterait l'érosion due aux forces mécaniques de la digestion (Allen *et al.*, 1985).

A



B

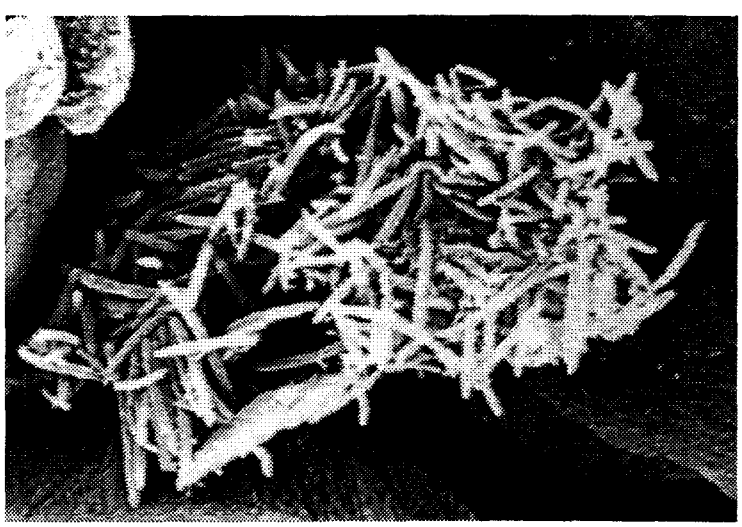


Figure 1 : Vues en microscopie électronique à balayage (A) du tapis muqueux présent à la partie apicale des cils de l'épithélium respiratoire (Gaillard *et al.*, 1992) et (B) de bactéries engluées dans une boule de mucus à la surface d'une crypte intestinale (Boureau *et al.*, 1995)

Bien que le mucus soit produit par l'hôte pour protéger la surface de l'épithélium gastro-intestinal, il peut aussi favoriser la colonisation par des micro-organismes commensaux ou pathogènes en fonctionnant comme source de nutriments et comme matrice dans laquelle ils peuvent proliférer (Figure 1B). La microflore abondante dans l'intestin des mammifères, particulièrement au niveau du côlon, semble être principalement concentrée dans la couche de mucus. Il peut y avoir accumulation dans le gel lorsque la vitesse à laquelle les organismes se multiplient dépasse la vitesse à laquelle ils sont expulsés via le turnover et l'érosion de la couche de mucus. Ce mucus peut favoriser ainsi la prolifération de nombreuses espèces de bactéries, y compris certaines bactéries de souches virulentes comme *Yersinia enterocolitica* et *Shigella flexneri* (Mantle, 1994).

Des changements dans l'organisation du gel sont généralement considérés comme associés à l'ulcère duodénal. De même, il a été suggéré que la diminution de la couche de mucus dans l'antre et le fundus gastrique soit à l'origine des ulcères de ces deux régions. Normalement, l'organisation en réseau du mucus est dégradée par la bile et la pepsine. Un juste équilibre entre sécrétion de mucus et altération de ce mucus est nécessaire pour que se maintienne l'intégrité de l'épithélium sous-jacent.

Les maladies inflammatoires cryptogénétiques de l'intestin sont des affections inflammatoires chroniques de cause inconnue touchant avec prédilection les adolescents et adultes jeunes. Elles comprennent deux entités : la rectocolite hémorragique (RCH) et la maladie de Crohn (MC). Une déficience des mécanismes de protection de la muqueuse semble être associée à ces pathologies. Par exemple, des études *in vitro* ont mis en évidence dans la RCH une érosion permanente de la barrière du mucus.

I.3.3. Le mucus du tractus génital

Dans l'appareil urogénital féminin, le mucus joue un rôle primordial dans la reproduction. Les fonctions du mucus dans le tractus génital dépendent de ses propriétés physico-chimiques, lesquelles reflètent la complexité de la nature biochimique de ce mucus. Il est principalement sécrété au niveau du col utérin (endocol), mais l'endomètre, les trompes de Fallope, et les glandes de Bartholin participent également à la production de ce mucus. Les variations endocriniennes observées au cours du cycle menstruel semblent déterminer les changements des propriétés du mucus cervical. Les études ultrastructurales de ce mucus permettent de corréler ses propriétés physiques aux variations d'orientation de l'hydrogel. Filance, quantité, et viscosité sont autant de variations des propriétés du mucus cervical au cours du cycle menstruel. Durant la phase folliculaire, on observe une augmentation progressive du volume d'un mucus de plus en plus fluide. Ce volume de mucus atteint son maximum lors de l'ovulation. Durant la phase lutéale, la quantité et la fluidité du mucus diminuent. La cristallisation rend parfaitement compte de toutes ces variations temporelles dévolues aux oestrogènes et progestagènes. En effet le phénomène dit de «cristallisation en feuille de fougère» (période péri-ovulatoire) permet de visualiser indiscutablement ces variations de propriétés du mucus. Il est donc plus que probable que des altérations de ce mucus puissent contribuer à certaines formes de stérilité chez la femme.

Les fonctions du mucus cervical sont nombreuses et peuvent être très différentes d'une espèce à une autre. Chez la jument ou la truie, l'éjaculation a lieu dans l'utérus même; les fonctions du mucus sont donc plus restreintes à la protection et à la lubrification (Gibbons, 1978) alors que dans la plupart des espèces animales, le mucus facilite ou au contraire, selon la phase du cycle, stoppe le passage des spermatozoïdes; il protège les spermatozoïdes de l'environnement vaginal hostile et de la phagocytose ; il filtre les spermatozoïdes arrêtant les spermatozoïdes anormaux ou peu motiles ; il peut aussi servir éventuellement de réservoir de sperme chez certaines espèces.

II LES MUCINES

II.1. Etude biochimique des mucines

II.1.1 Définition

Les mucines sont le constituant majoritaire du mucus et lui confèrent ses propriétés rhéologiques. Les mucines, d'une façon générale, peuvent se définir comme de grandes molécules dont la masse atteint plusieurs millions de daltons. Ce sont des *O*-glycoprotéines dont les chaînes glycanniques représentent plus de 50 % du poids sec. Les chaînes oligossacharidiques sont liées aux nombreux résidus de Ser et Thr de l'apomucine via des *N*-acétylgalactosamines (GalNAc). Ces chaînes oligosaccharidiques sont composées de galactose (Gal), de glucose (Glc), de *N*-acétylglucosamine (GlcNAc), de GalNAc et d'acide sialique (NeuAc).

II.1.2 Les différentes mucines

II.1.2.1 Le type mucine

Des *O*-glycoprotéines de type mucine ont été trouvées associées à des cellules humaines non-épithéliales, comme par exemple la glycophorine, glycoprotéine intégrée dans la membrane des hématies ou encore la leucosialine (parfois appelée sialophorine ou encore CD34) qui s'exprime à la surface des cellules B, T, des macrophages, des plaquettes et des granulocytes. Ces mucines, parfois appelées "mucin-like", jouent le rôle de ligand pour les sélectines (Shimizu *et al.*, 1993 ; pour revue voir Van Klinken *et al.*, 1995).

En 1995, Noia *et al.* ont décrit une famille de gènes codant des « mucin-like » exprimées à la surface du protozoaire *Trypanosoma cruzi*. Les peptides déduits des ADNc clonés ont des caractéristiques communes qui ne sont pas sans rappeler les caractéristiques des mucines. Les peptides signaux des différentes apomucines sont très homologues les uns aux autres ainsi que les parties N- et C-terminales; l'apomucine est essentiellement constituée d'un motif, riche en Thr (motif trouvé dans la souche CA1/72 : TTTTTTTTKPP) dont certains sont potentiellement *O*-glycosylables; ce motif, non conservé d'une souche à l'autre, est répété en tandem dans la région centrale de l'apomucine; les motifs répétés en aval de la région centrale sont moins parfaits; et enfin, le nombre de motifs parfaits est variable selon les parasites d'une même souche; ce motif est donc de type VNTR (Noia *et al.*, 1995). Une famille de gènes similaires comprenant au moins 10 membres a été décrite chez *Leishmania major* (Murray *et al.*, 1991) avec la particularité cependant que le peptide déduit d'un des transcrits étudiés (PSA-2) code en son extrémité 3' une région riche en résidus Cys.

Parce que la fonction de ces mucines ne correspond pas à la définition donnée plus haut quant à leur rôle, nous nous focaliserons désormais uniquement sur les mucines exprimées à la surface des épithéliums.

II.1.2.2 Les mucines humaines

Les mucines sécrétées sont synthétisées par des cellules glandulaires de la sous-muqueuse et des cellules épithéliales spécialisées, appelées cellules sécrétrices au niveau des épithéliums bronchiques, des cellules caliciformes dans l'épithélium colique ou cellules à pôle fermé au niveau de l'estomac. Elles sont sécrétées sous forme de granules, qui, libérés dans la lumière, forment au contact de l'eau un gel viscoélastique à la surface de l'épithélium. Ces cellules sont polarisées et présentent des caractéristiques communes en rapport avec la production massive de mucus : réticulum

endoplasmique et appareil de Golgi développés et granules de sécrétion accumulés vers le pôle apical (Figure 2).

Bien que la mucine membranaire MUC1 diffère grandement des mucines sécrétées, dans sa structure et probablement dans son rôle, elle présente cependant des caractéristiques communes à celles des mucines sécrétées. Cette mucine est appelée elle aussi "mucin-like" car elle est membranaire. Elle a de plus été retrouvée sous une forme soluble (Verma, 1994 ; Hilkens *et al.*, 1995).

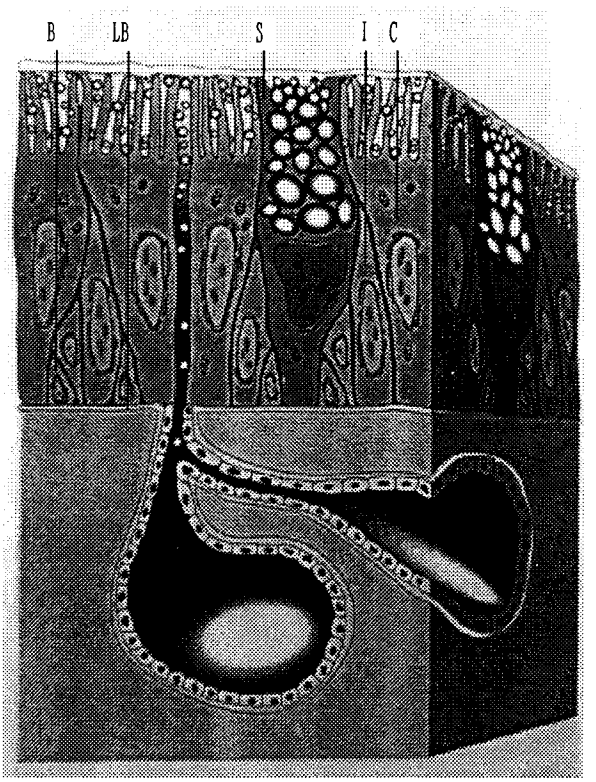
II.1.3 Composition des mucines

II.1.3.1 Purification des mucines

L'isolement des mucines est couramment basé sur leurs grandes tailles et leurs densités. Classiquement, des chromatographies par gel filtration sur Sepharose CL-2B et des ultracentrifugations en chlorure de césium sont employées pour isoler ces grandes molécules très glycosylées et donc très denses (Strous *et al.*, 1992 ; Gum, 1992). La purification peut alors se poursuivre selon 2 stratégies différentes : les glycoprotéines sont traitées soit dans des conditions non-dénaturantes pour préserver les ponts disulfures entre les polypeptides et ainsi déterminer la composition; soit en milieu réducteur avant de purifier la composante monomérique. De plus, comme les parties C- et N-terminales non-glycosylées des mucines sont extrêmement sensibles à la protéolyse, la purification de ces molécules se fait généralement en présence d'inhibiteurs de protéases. Si ces précautions ne sont pas prises, les mucines purifiées sont tronquées des extrémités nues (Dekker *et al.*, 1989). Pour la plupart des mucines, il semble que la purification par centrifugation isopycnique par gradient de densité (chlorure de césium/chlorure de guanidium) semble la méthode la plus appropriée pour isoler une mucine très pure. Le désavantage de cette technique est que la mucine purifiée a perdu sa capacité de former un gel avec les mêmes propriétés rhéologiques que la mucine originelle (Snary *et al.*, 1974).

A

B



B : cellule basale ; LB : lame basale ; S : cellule sécrétrice ; I : cellule intermédiaire ; C : cellule ciliée.

Figure 2 : (A) Vue en microscopie électronique à transmission de l'épithélium respiratoire et (B) schéma de l'architecture respiratoire (Gaillard *et al.*, 1992)

Un autre point important est le choix du matériel biologique de départ. En effet, les mucines peuvent être endommagées par les protéases ou les forces de cisaillement ou encore par les produits sécrétés par les cellules épithéliales dans le mucus. De plus, la purification d'une mucine à partir de mucus peut être contaminée par des glycoprotéines, du mucus, originaires d'autres organes. Par exemple, le mucus bronchique provenant d'expectorations peut être « contaminé » par du mucus salivaire ou encore, le même mucus salivaire « contamine » physiologiquement le mucus gastrique. Des molécules pures et totalement intactes peuvent uniquement être isolées et purifiées à partir des mucines intracellulaires stockées. Les techniques développées pour isoler ainsi les mucines ont abouti à des compositions en aa très différentes de celles des mucines isolées à partir de mucus (Fahim *et al.*, 1987).

II.1.3.2 La composante saccharidique

La composante saccharidique des mucines représente généralement 80% du poids sec de la molécule (Gum, 1992). Les chaînes *O*-glycosidiques des mucines sont liées sur les groupements hydroxyls des résidus Ser et Thr par les liaisons de type α-*O*-glycosidique via des résidus de GalNAc. Ces chaînes oligosaccharidiques maintiennent la mucine sous une forme linéaire (Rose *et al.*, 1984). Outre la GalNAc, du fucose (Fuc), du Gal, du GlcNAc et des acides sialiques ont été trouvés dans les mucines ainsi que parfois des groupements sulfates et de faibles quantités de mannose. A noter que les mucines ne contiennent pas d'acides uroniques.

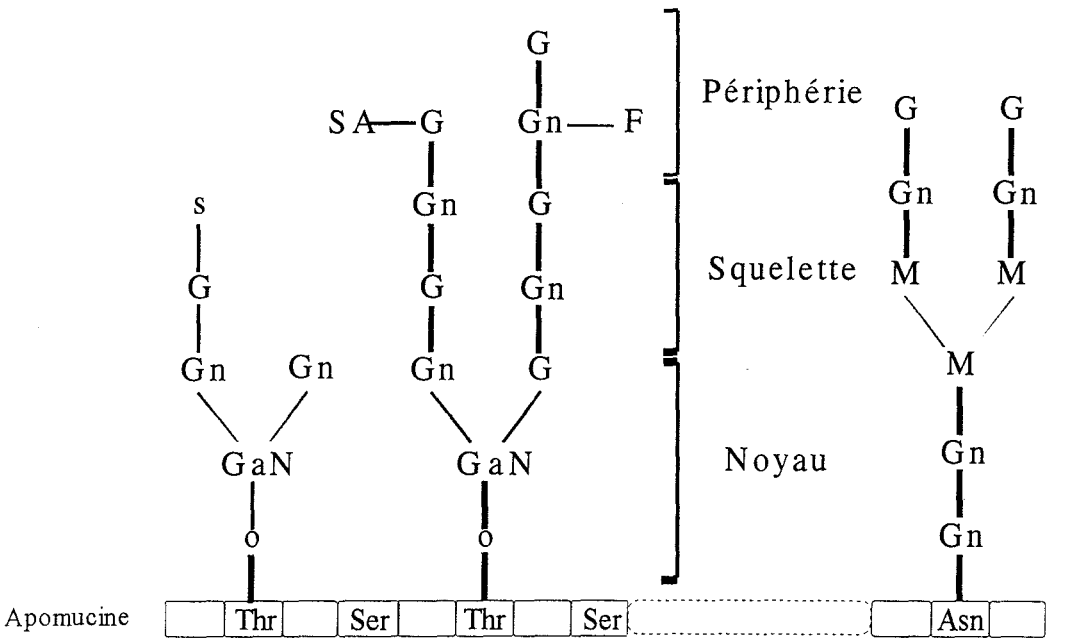
La grande diversité des chaînes *O*-glycosidiques, constituées de ces 5 résidus monosaccharidiques, engendrée par la variation dans leur composition, dans leur longueur, par les différents types de branchement et par leur acidité a constitué un obstacle majeur à leur détermination. Par exemple, 88 chaînes différentes ont été isolées dans les mucines respiratoires d'un même individu. L'étude de MUC1 a montré que l'apomucine est exprimée par différents tissus. La forme mature de MUC1, glycosylée, a une M_r dans les glandes mammaires de 250 000 à 500 000 alors que dans

le pancréas elle a une M_r de plus de 1 000 000 et une M_r de 600 000 à 800 000 dans les carcinomes du côlon (Gendler *et al.*, 1995). Il est donc vraisemblable qu'une même apomucine est glycosylée de façon différente selon l'individu, l'organe et l'état physiologique.

Les chaînes *O*-glycosidiques sont constituées de 3 régions (pour revue, voir Roussel *et al.*, 1996) faisant intervenir des séries de réaction de glycosylation. Le "noyau" est formé par les premières réactions de glycosylation, puis est formé le "squelette" et enfin la "périphérie" (Figure 3). La synthèse des chaînes *O*-glycosidiques est initiée par l'action sur l'apomucine d'enzymes très spécifiques, les UDP-GalNAc-polypeptide-α-*N*-acétylgalactosaminyltransférases. Ces enzymes sont probablement localisées dans le *cis*-Golgi. Ce résidu de GalNAc ainsi que les monosaccharides directement attachés à ce résidu, constituent le "noyau" des chaînes *O*-glycosidiques des mucines. Ce premier résidu de GalNAc peut être substitué sur le groupement hydroxyle du carbone C3 via une liaison β1-3 par un résidu de Gal ou de GalNAc donnant naissance respectivement aux noyaux 1 et 3. L'addition de GlcNAc en β1-6 sur ces 2 noyaux donne naissance aux 2 nouveaux noyaux 2 et 4. D'autres noyaux ont été décrits dans le méconium. Différents types de noyaux, résultant de l'action de nombreuses glycosyltransférases, peuvent être trouvés dans les mucines sécrétées d'un même individu.

Ensuite, la formation du squelette résulte de l'action de différentes glycosyltransférases qui catalysent le transfert en alternance de résidus Gal ou GlcNAc. Ainsi, le squelette est constitué par des disaccharides Gal(β1-3)GlcNAc ou Gal(β1-4)GlcNAc. Ces disaccharides peuvent être liés sur tous les types de noyaux ou encore être branchés sur un résidu interne de Gal du squelette via une liaison β1-3 et/ou β1-6.

La périphérie des chaînes *O*-glycosidiques des mucines est caractérisée par la présence de Fuc, Gal, GalNAc, d'acide sialique, le plus souvent en liaison α. Des résidus de sulfate peuvent aussi être additionnés. Tout ces résidus sont branchés par l'action de différentes glycosyltransférases.



GaN, *N*-acétylgalactosamine; G, galactose; F, fucose; s, sulfate;

Gn, *N*-acétylglucosamine; SA, acide sialique

Figure 3 : Représentation schématique des chaînes glycosylées des mucines (Lamblin *et al.*, 1993)

II.1.3.3 La composante peptidique

L'axe peptidique représente, selon l'espèce animale et selon l'origine du mucus étudié, de 14 à 37 % du poids sec de la molécule (Murty *et al.*, 1978, Hill *et al.*, 1977, Roussel *et al.*, 1975, Snyder *et al.*, 1982). Les résidus Ser, Thr et Pro représentent 20 à 50 % de la composition totale en aa (Van Klinken *et al.*, 1995). La teneur en cystéine a fait l'objet de nombreuses controverses. Aujourd'hui, il semble bien que ces résidus soient présents en quantité significative dans la plupart des mucines et qu'ils jouent un rôle primordial dans l'organisation tridimensionnelle du gel par l'établissement de ponts disulfures intra et intermoléculaires.

II.1.3.3 La composante lipidique

A côté des domaines hydrophiles glycosylables des apomucines sont présentes des régions plus hydrophobes et capables d'établir des liaisons avec des lipides. Différentes classes de ces lipides ont pu être identifiées: lipides neutres (mono-, di- et triglycérides, cholestérol et esters de cholestérol), glycolipides et phospholipides. Les acides gras majoritairement trouvés sont l'acide palmitique, stéarique et oléique (Woodward *et al.*, 1982). Il a été proposé que ces acides gras liés de façon covalente protègent la mucine de la dégradation protéolytique (Slomiany *et al.*, 1983).

II.1.4 Organisation moléculaire

II.1.4.1 Les modèles structuraux

Une caractéristique essentielle des mucines sécrétées est leur structure oligomérique. La réduction des ponts disulfures provoque une diminution de la viscosité du mucus ; l'intégrité de ces ponts disulfures est indispensable à la fonction de ces mucines. Certains auteurs ont suggéré que la réduction pourrait de plus activer des enzymes protéolytiques étroitement associées aux mucines. Elles coupent l'axe peptidique en des régions accessibles et donc peu glycosylées provoquant une

diminution de masse. Deux modèles structuraux ont été proposés pour expliquer ces caractéristiques des mucines (pour revue voir Pasquier *et al.*, 1990).

II.1.4.1.1 Le modèle du «moulin à vent»

Allen en étudiant les mucines gastriques de porc a proposé un premier modèle dit « en moulin à vent » où 4 sous-unités de mucines sont unies à une sous unité centrale de 90 kDa (Allen, 1983).

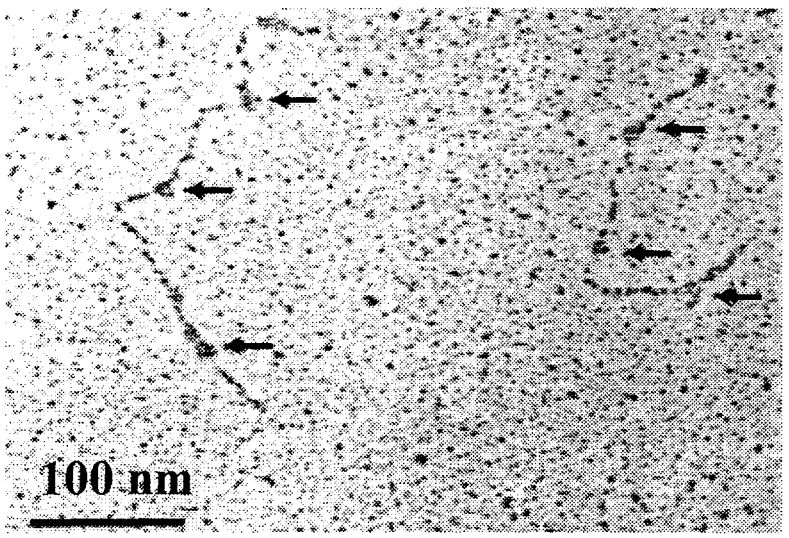
II.1.4.1.2 Le modèle linéaire et les zones nues

Un second modèle, qui s'appuie sur les études en microscopie électronique de mucines bronchiques et cervicales humaines et de mucines gastriques de porc, a été proposé par Carlstedt et Sheehan (Sheehan *et al.*, 1986). Ces mucines apparaissent en microscopie électronique comme de longs filaments linéaires et flexibles et de longueur extrêmement variable (Thornton *et al.*, 1991).

La digestion protéolytique des mucines cervicales et bronchiques met en évidence une alternance de zones hautement glycosylées et résistantes à la protéolyse et de zones peu glycosylées, nues, et sensibles aux protéases. Ces données biochimiques ont été confortées par la microscopie électronique : des anticorps dirigés contre les régions nues se fixent régulièrement le long des filaments de mucines (Figure 4 ; Sheehan *et al.*, 1991). Il a été proposé que les mucines s'oligomérisent par la formation de ponts disulfures interchaînes dans les extrémités N- et C-terminales des mucines (Carlstedt *et al.*, 1984 ; Sheehan *et al.*, 1986 ; Dekker *et al.*, 1991), conformément aux données de la microscopie électronique (Figure 5). Ce modèle qui rend parfaitement compte des données récentes de la biologie moléculaire est schématisé par la figure 6.

Les zones nues, plus hydrophobes, pourraient former des liaisons non-covalentes avec d'autres composants du mucus, dont des cations, des enzymes, et des

A



B

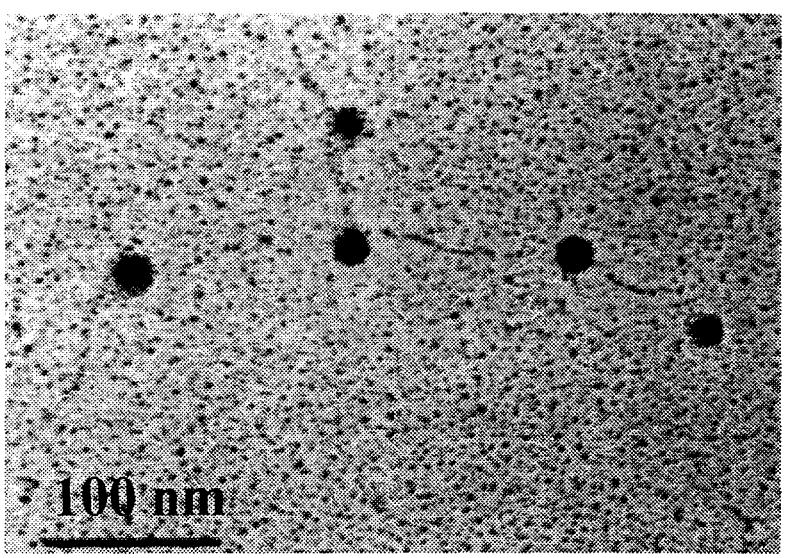


Figure 4 : Vues en microscopie électronique de mucines cervicales humaines après réduction (Sheehan *et al.*, 1991).
(A) Incubées avec un antisérum dirigé contre les zones nues
(B) La même préparation qu'en (A), incubée avec de la protéine A marquée à l'or coloïdal

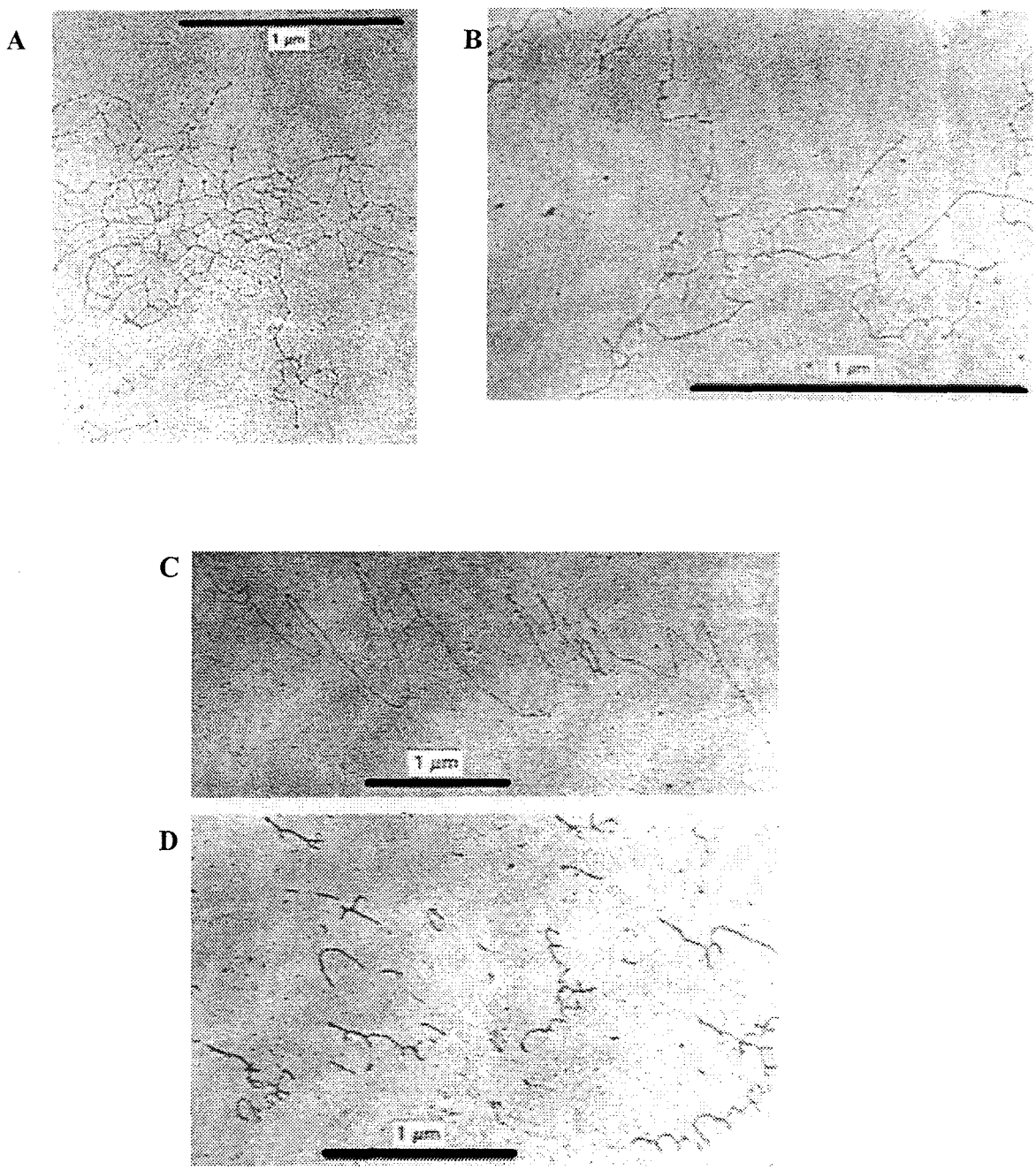


Figure 5 : Mucines bronchiques humaines (A) et (B) et gastriques de porc (C) et (D) (Sheehan *et al.*, 1986)

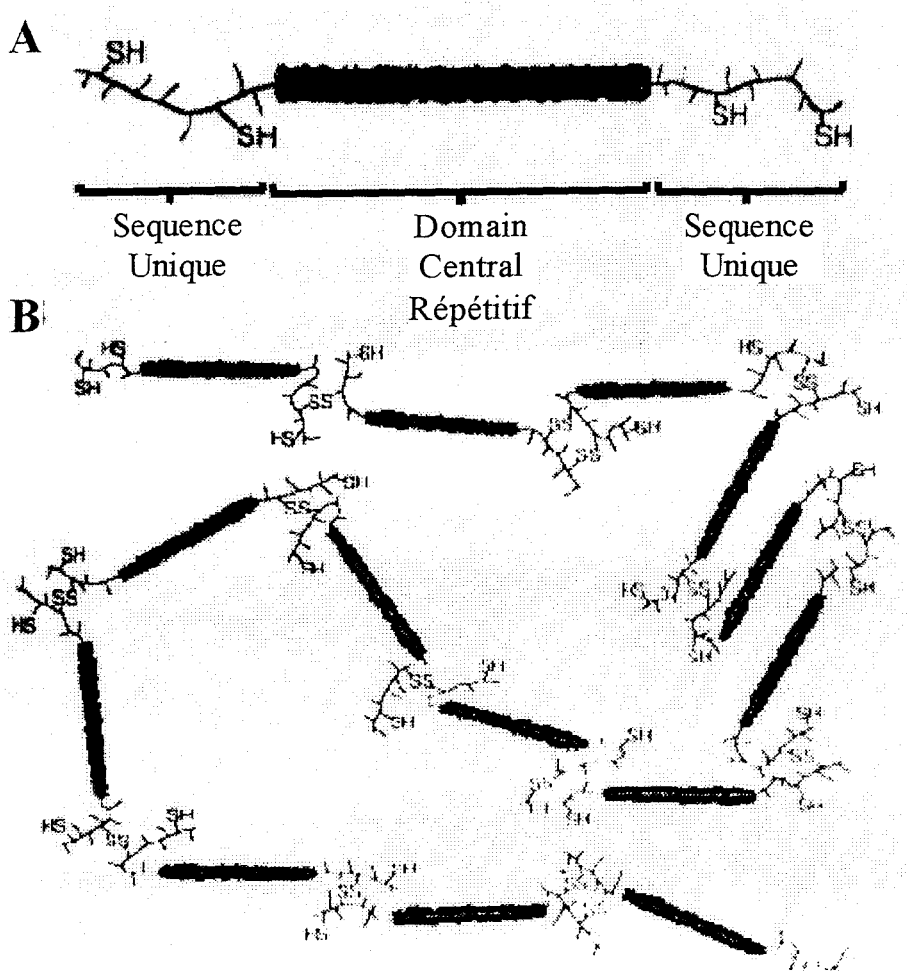


Figure 6 : Modèle de mucine proposé par Carlstedt (pour revue : Gum, 1992)

protéines (Shankar *et al.*, 1990). Dans la vésicule biliaire, il est probable que ces régions sont impliquées dans la liaison avec des lipides biliaires dont le cholestérol (Lee *et al.*, 1981a). Ces calculs se forment en 3 étapes : forte sécrétion de cholestérol et de mucine, puis processus de nucléation et enfin agglomération et croissance des cristaux. Il semblerait que le mécanisme de formation des cristaux biliaires est dû à une hypersécrétion des mucines dans la vésicule biliaire (Lee *et al.*, 1981b) et la formation d'un complexe mucine/bilirubine par des interactions non-covalentes (Smith *et al.*, 1985) et de type hydrophobe entre des zones peu ou pas glycosylées des mucines (Smith *et al.*, 1984).

II.1.4.2 Le peptide de liaison ou "peptide link"

Depuis le modèle de Allen, de nombreux groupes ont tenté de caractériser le peptide de liaison qui relierait quatre grandes sous-unités de mucine. Il a une taille de 70 kDa dans l'estomac de porc, (Pearson *et al.*, 1981, Allen *et al.*, 1984). Des peptides similaires d'un poids moléculaire de 118 kDa ont été décrits dans l'intestin de différentes espèces animales et chez l'Homme. Ce peptide a un poids moléculaire de 150 kDa dans la salive humaine (Roberton *et al.*, 1989 ; Slomiany *et al.*, 1992). Cependant, ce modèle n'a jamais pu être validé par les observations en microscopie électronique. Les travaux de Roberton *et al.* (Roberton *et al.*, 1989) effectués en collaboration avec 4 autres laboratoires ont montré en 1989 que cette sous-unité de 118 kDa, purifiée par différentes techniques biochimiques à partir de nombreux échantillons de différents organes et de différentes espèces, était présente dans les granules de sécrétion des cellules en gobelets de l'intestin, et dérivait d'une réduction d'un pont disulfure d'une plus grande sous-unité d'une mucine.

Dans la salive, une sous unité de 150 kDa a été isolée après réduction des ponts disulfures à partir de la mucine MG1. Cette sous-unité de 150 kDa contient des chaînes *N*-glycosidiques tout comme le peptide de 118 kDa (Fahim *et al.*, 1987 ; Mantle *et al.*, 1984 ; Kawagishi *et al.*, 1990). Compte tenu que l'antisérum dirigé contre la protéine de liaison de 118 kDa isolé à partir d'intestin de rat reconnaît aussi le

peptide salivaire de 150 kDa, il a été suggéré que la différence de masse observée entre ces deux peptides pourrait être due à une glycosylation différente d'un même peptide ou de peptides très similaires (Kawagashi *et al.*, 1990). Ainsi, le peptide de liaison semble être un peptide libéré par protéolyse d'une mucine et lié par un ou des ponts disulfures à cette même mucine.

Le peptide de liaison est plus hydrophobe que la région *O*-glycosylable de l'apomucine. Ce peptide de liaison est impliqué, par l'intermédiaire d'un récepteur localisé dans ses chaînes oligomannosidiques, dans l'interaction de bactéries pathogènes *Escherichia coli* (sérotype O157:H7). Ce récepteur serait en partie caché par des lipides liés à l'apomucine par des liaisons non-covalentes et serait démasqué, dans certaines circonstances physiologiques, par la réduction de ponts disulfures ou la "délipidation" (Sajjan *et al.*, 1990a; Sajjan *et al.*, 1990b).

II.2 Les gènes de mucines

II.2.1 Stratégie d'étude et caractéristiques des mucines

Du fait de la structure même des mucines, la détermination de leurs séquences peptidiques par la dégradation d'Edman s'est révélée inopérante. Les techniques de l'ADN recombinant ont permis de pallier les difficultés de l'étude des apomucines. Ces techniques ont permis d'isoler des clones d'ADNc partiels de mucines. La stratégie générale mise en oeuvre est toujours identique : les mucines ont été purifiées à partir du mucus produit par un tissu sain ou tumoral en utilisant l'ultracentrifugation, le tamisage moléculaire et l'échange ionique. Les mucines sont ensuite déglycosylées presque totalement par solvolyse en milieu acide fort. L'axe peptidique est alors injecté à des lapins ou souris afin de produire un immunsérum dirigé contre l'apomucine. Cet immunsérum, testé par immunohistochimie ou immunotransfert, permet de cribler une banque d'expression construite à partir de la muqueuse ou de la lignée cellulaire étudiée. Les clones positifs sont alors séquencés et la localisation du gène

correspondant peut être entreprise par les techniques d'hybridation de lignées Homme-rongeurs ou par hybridation fluorescente *in situ* (technique FISH).

Les séquences déduites montrent que l'axe peptidique des mucines est constitué de motifs répétés riches en résidus Ser, Thr, Pro et Ala. Ces motifs répétés varient en longueur et dans leur séquence d'un gène à l'autre, d'une espèce à l'autre (voir tableau I). La plupart des gènes de mucines présentent un polymorphisme de type VNTR, c'est-à-dire que le nombre de répétitions varie d'un individu à l'autre. De plus, les gènes de mucines sécrétées, analysés par Northern blot, apparaissaient complexes. Les ARNm préparés à partir de tissus humains avaient la caractéristique quasi unique de produire un continuum d'ARN messagers allant de plus de 20 kb à moins de 1 kb et ceci dans l'ensemble des muqueuses où ils s'expriment. Cependant, des travaux récents au laboratoire ont permis d'affiner la méthode de préparation des ARN et de montrer que le continuum observé en Northern blot est un artefact de préparation aggravé par des difficultés de transfert (Figure 7 ; Debailleul, 1997).

II.2.2 Les mucines humaines : les gènes, les transcrits et protéines déduites

II.2.2.1 MUC1, mucine membranaire

De nombreux travaux ont été réalisés sur la mucine membranaire MUC1. Cette mucine a été isolée dans plusieurs laboratoires sous les noms de HMFG, EMA, PAS-0, DUPAN-2, PUM, Cal, NPGP, NCRC11, épitectine, MAM-6, antigène DF3, SGA, antigène H23, PEM ou encore épisialine (Gendler *et al.*, 1995). Cette grande variété de noms portait évidemment à confusion ; et par convention, les loci de gènes de mucines portent la dénomination de MUC d'après l'« Human Genome Mapping » (HGM) suivi par un numéro indiquant l'ordre de découverte du gène. Ainsi, PEM a pris la dénomination de MUC1, le premier gène identifié codant une mucine.

Gène	AA	pb	Séquence consensus	Référence
MUC1	20	60	PDTRPAPGSTAPPAHGVTSA	Gendler *et al*, 1988
MUC2	23	69	PTTTPITTTTTVTPTPTPTGTQT	Gum *et al.*, 1989
MUC3	17	51	HSTPSFTSSITTTETTS	Gum *et al.*, 1990
MUC4	16	48	TSSASTGHATPLPVTD	Porchet *et al.*, 1991
MUC5AC	8	24	TTSTTSAP	Guyonnet Dupérat *et al.*, 1995
MUC5B	29	87	dégénéré	Aubert *et al.*, 1991; Dufossé *et al.*, 1993
MUC6	169	507	SPFSSTGPMTATSFQTTTTYPTPSHP QTTLPTHVPPFSTSLVTPSTGTYITP THAQMATSASIHSTPTGTIPPPTTLK ATGSTHTAPPMTPTTSGTSQAHSSFS TAKTSTSLHSHTSSTHHPEVTPTSTT TITPNPTSTGTSTPVAHTTSATSSRL PTPFTTHSPPTGS	Toribara *et al.*, 1992
MUC7	23	69	TTAAPPTPSATTPAPPSSSAPPE	Bobek *et al.*, 1993
MUC8	13	39	TSCPRPLQEGTRV	Shankar *et al.*, 1994

Tableau I : Séquences peptidiques consensus des domaines répétitifs des mucines humaines

Gène	Localisation Chromosomique	Référence
MUC1	1q21-24	Swallow *et al.*, 1987
MUC2	11p15.5	Griffiths *et al.*, 1990
MUC3	7q22	Fox *et al.*, 1992
MUC4	3q29	Gross *et al.*, 1992
MUC5AC	11p15.5	Nguyen *et al.*, 1990
MUC5B	11p15.5	Nguyen *et al.*, 1990
MUC6	11p15.5	Toribara *et al.*, 1993
MUC7	4q13-q21	Bobek *et al.*, 1996
MUC8	12q24.3	Shankar *et al.*, 1995

Tableau II : Localisation chromosomique des mucines humaines

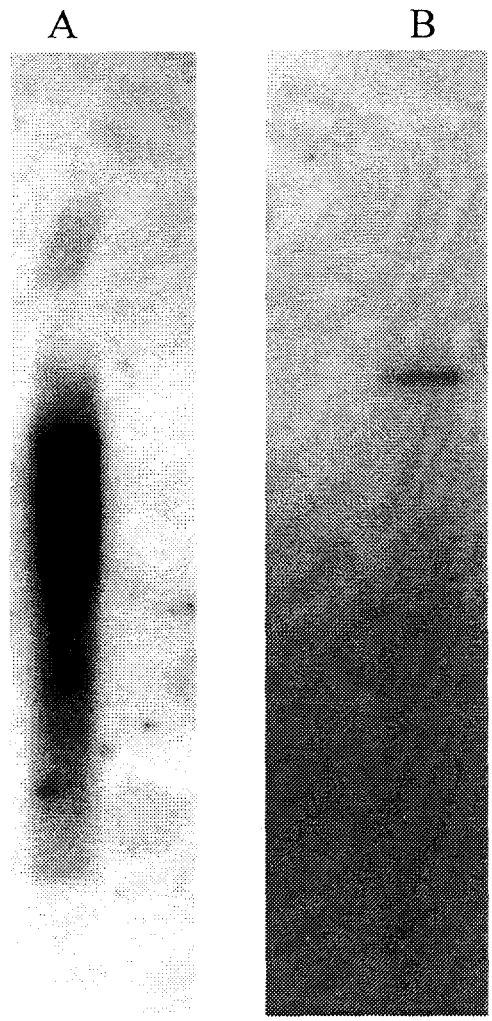
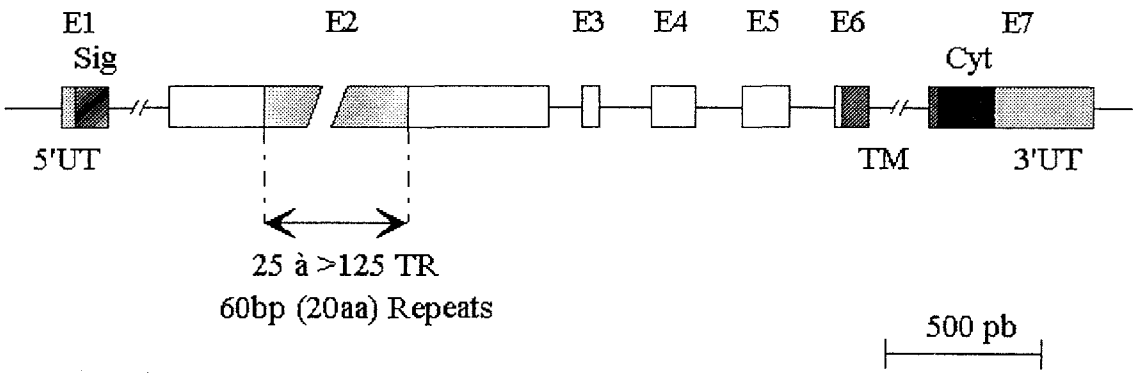


Figure 7 : Northern blot de trachée humaine (sonde non répétitive de *MUC5B*) selon
(A) la technique classique de préparation des ARNm
(B) la technique décrite par Debailleul (Debailleul, 1997)

Le gène *MUC1* est le seul gène de mucine pour lequel les séquences de l'ADNc et génomique ont été complètement déterminées. Le gène, composé de 7 exons et 6 introns (Figure 8), a été localisé sur le bras long du chromosome 1 humain dans une région dense en gènes en q21-24 (Tableau II). Les ADNc ont été isolés à partir de glandes mammaires, de pancréas et d'ovaires. Le gène s'étend sur 4 à 7 kb et code une protéine comportant des domaines distincts : un peptide signal suivi d'un motif dégénéré répété en tandem, la partie centrale de l'apomucine déduite est faite de répétitions d'un motif dont le nombre varie selon les individus de 25 à plus de 125. Ce motif, de 20 aa, est moins parfait dans la région la plus en aval. Ce domaine est suivi d'une séquence unique également de type mucine (riche en Ser, Thr, Pro) de 227 aa, puis d'une séquence hydrophobe transmembranaire de 31 aa et d'une queue cytoplasmique de 69 aa. Il semblerait que 2 motifs, l'un situé à la jonction du domaine transmembranaire et du domaine cytoplasmique et l'autre dans le domaine extracellulaire non répétitif, soient impliqués dans l'adressage à la surface de la cellule (Pemberton *et al.*, 1996). Il a de plus été suggéré que la queue cytoplasmique interagit avec des protéines de structure du cytosquelette (Parry *et al.*, 1990). Deux variants, proposés comme l'un codant la forme membranaire et l'autre la forme soluble de MUC1, seraient engendrés par épissage alternatif du site accepteur de l'exon 2 (Ligtenberg *et al.*, 1990, Wreschner *et al.*, 1990). Un variant serait alors amputé de 27 pb de la partie 5' de l'exon 2. Un autre variant a été décrit : l'ARNm serait colinéaire au gène au delà de la partie répétitive; aucun épissage n'aurait lieu engendrant juste en aval de la région répétitive un peptide déduit de 160 aa, sans séquence peptidique transmembranaire, et une longue partie 3' non traduite (Wreschner *et al.*, 1990). Cependant ces variants n'ont jamais été observés par d'autres équipes. Il est en fait probable que la forme soluble de MUC1 ne soit pas engendrée par épissage alternatif mais uniquement par l'action d'une protéase. En effet, il a été observé que la mucine tout d'abord transmembranaire est recyclée dans le compartiment *trans* du Golgi où MUC1 serait sialylé et clivé donnant ainsi naissance à la forme sécrétée de MUC1 (Litvinov *et al.*, 1993).

Enfin, un ADNc de 1,2 kb de MUC1, sans région répétitive, a été décrit par Zrihan-Licht *et al.* (Zrihan-Licht *et al.*, 1994) dans des cancers primitifs du sein. Ce variant a pris la dénomination de MUC1/Y.



E : Exon
Sig : séquence signal
TM : région codant le domaine transmembranaire
Cyt : région codant le domaine cytoplasmique
UT : région non-traduite (untranslated region)

Figure 8: Organisation génomique de *MUC1* et organisation peptidique déduite

II.2.2.2 Les mucines d'origine intestinale MUC2 et MUC3

II.2.2.2.1 Le gène *MUC2*

Les travaux de J. Gum et Y. Kim à San Francisco ont permis d'identifier trois clones d'ADNc partiels dénommés SMUC 40, 41 et 42 de respectivement 502, 836 et 229 pb. Ces trois clones ont été isolés à partir d'une banque d'ADNc de jéjunum humain grâce à un immunsérum dirigé contre l'axe peptidique des mucines purifiées de la lignée tumorale colique LS174T (Gum *et al.*, 1989). Ces trois clones possèdent le même arrangement répétitif d'un domaine élémentaire de 69 nucléotides, soit 23 aa (Tableau I).

Par criblage, à l'aide d'un oligonucléotide synthétisé d'après les séquences des clones SMUC, d'une banque d'ADNc construite à partir d'ADNc de trachée d'un patient atteint de mucoviscidose, Gérard *et al.* ont isolé un nouvel ADNc nommé AMN-22 (Gerard *et al.*, 1990). Cet ADNc code une région très similaire à la partie répétitive codée par les clones SMUC. En 1991, l'équipe de C. Basbaum isole un autre ADNc, HAM-1, à partir d'une banque de trachée humaine criblée par l'immunsérum utilisé par l'équipe de Y. Gum (Jany *et al.*, 1991). Tous ces ADNc sont des transcrits partiels du gène *MUC2*, localisé sur le bras court du chromosome 11 en p15.5 (Tableau II ; Griffith *et al.*, 1990).

En 1991, l'équipe de J. Gum à San Francisco a isolé un clone génomique de 11 kb appelé GMUC (Toribara *et al.*, 1991). Ce clone établit une relation physique entre la partie répétitive du gène MUC2 (42 à 43 motifs répétés sans interruption) et un nouveau domaine en amont codant une protéine riche en Ser, Thr et Pro. Les auteurs suggèrent que ce domaine est constitué d'un motif répété mais très irrégulier de 16 aa. Ce domaine est par ailleurs encadré de part et d'autre par deux séquences peptidiques très similaires et contenant des résidus Cys. Ce motif d'environ 110 aa sera appelé ensuite "motif Cys". Les études de polymorphisme (Toribara *et al.*, 1991) ont montré que le gène *MUC2* présente un polymorphisme de type VNTR, et ce,

uniquement pour le domaine composé du motif répété en tandem de 23 aa. Les individus étudiés possèdent entre 51 et 115 répétitions de ce motif.

Des expériences de "RACE-PCR" (voir chapitre stratégie) ont permis à la même équipe de déterminer ensuite les séquences des régions transcrites en amont et en aval des parties répétitives de *MUC2* (Gum *et al.*, 1992, Gum *et al.*, 1994). Une séquence de 15720 pb d'un transcrit "virtuel" de ce gène (numéro d'accession L21998 dans la GenBank), codant une apomucine de 5177 aa, a permis pour la première fois de schématiser sur des bases moléculaires une mucine sécrétée humaine (Figure 9). Le gène *MUC2* est constitué de trois parties. La partie centrale est faite d'une énorme région purement répétitive de 2 à 3000 acides aminés précédée d'un domaine répétitif plus petit mais irrégulier. La partie C-terminale est divisée elle aussi en deux parties: un petit domaine de type mucine de 139 acides aminés suivi d'un long domaine de 845 acides aminés riche en résidus Cys et possédant des sites potentiels de *N*-glycosylation. Le domaine N-terminal est lui aussi très complexe, riche en résidus Cys et présente, comme le domaine C-terminal, des homologies avec certains domaines du pro-facteur de von Willebrand.

II.2.2.2.2 Le gène *MUC3*

Une banque d'ADNc intestinal construite en vecteur d'expression λgt11 a été criblée par des anticorps dirigés contre de la mucine intestinale humaine déglycosylée. Des quatre clones isolés, deux (SIB 124 et 139) codent un motif de 17 aa riche en Ser et Thr, motif répété en tandem (Gum *et al.*, 1990). Ces deux clones définissent le gène *MUC3*, localisé sur le chromosome 7 en q22 (Fox *et al.*, 1992) ce qui permet d'affirmer que ces ADNc n'appartiennent pas aux gènes *MUC1* ou *MUC2* (Tableaux I et II).

Un clone génomique codant sur 1000 pb une vingtaine de répétitions en tandem caractéristiques de *MUC3* a été isolé et étudié. La partie répétitive est suivie en aval d'une séquence codant 617 aa, dont les 500 premiers sont de type «mucine» puisque le peptide correspondant est riche en résidus Ser et Thr (50%) et pauvre en résidus Cys (moins de 1%). Cette séquence est suivie d'une séquence

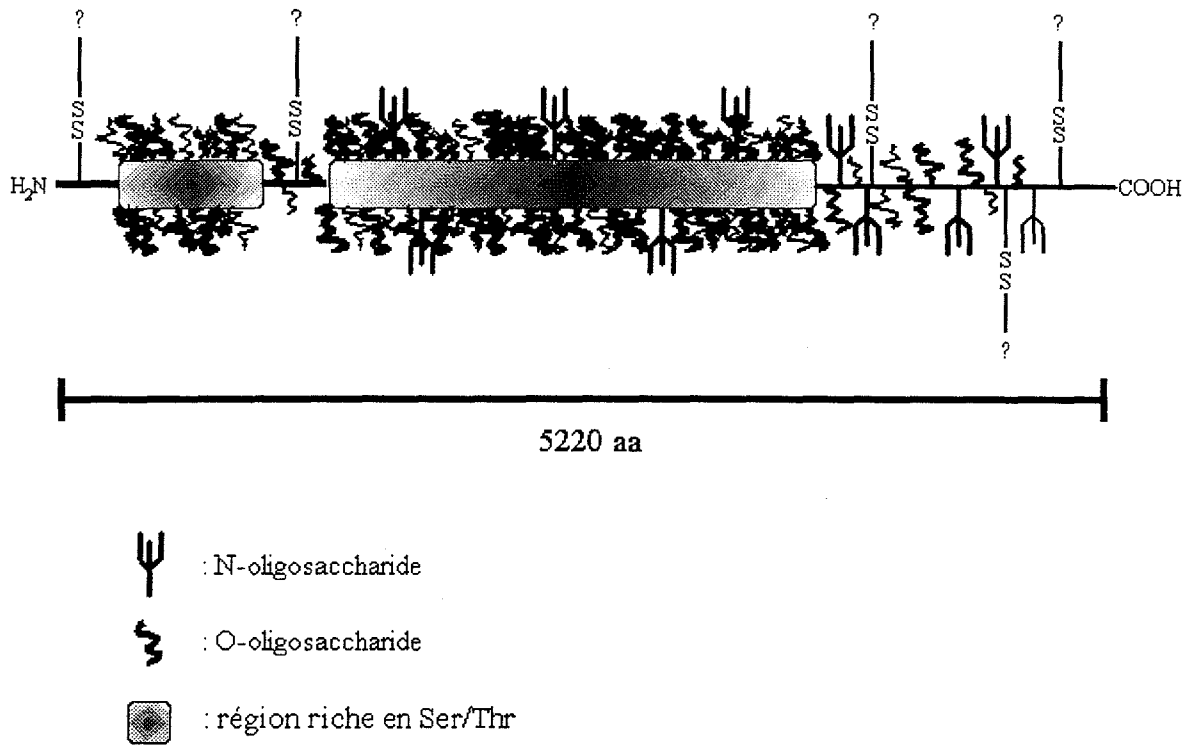


Figure 9 : Schéma de la mucine sécrétée MUC2 (Toribara *et al.*, 1991)

codante un peptide d'une centaine d'aa riche en résidus Cys (10%) ; le nombre de ces résidus ainsi que l'espacement entre eux sont caractéristiques du motif EGF-like retrouvé dans de nombreuses autres protéines (Kim *et al.*, 1994).

Une étude sur 40 familles montre que la partie répétitive de *MUC3* est de type VNTR mais les auteurs n'excluent pas l'existence d'un autre gène (ou pseudogène ?) qui aurait une zone qui serait reconnue par la partie répétitive de *MUC3* (Hill *et al.*, 1996).

II.2.2.3 Les mucines d'origine trachéobronchique MUC4, MUC5AC et MUC5B

En 1990, 20 clones positifs ont été isolés dans notre laboratoire d'une banque d'ADNc de trachée humaine, construite en vecteur d'expression λgt11 et criblée par un immunsérum de lapin dirigé contre de la mucine bronchique déglycosylée chimiquement (Crépin *et al.*, 1990).

II.2.2.3.1 Le gène *MUC4*

Le clone JER64, d'une longueur de 1,8 kb a été entièrement séquencé. Il est constitué uniquement de la répétition (au moins 39 fois) presque parfaite d'un motif de 48 pb codant une apomucine partielle riche en Thr et Ser. La sonde JER64, utilisée sur des Southerns de cellules hybrides Homme-rongeur a tout d'abord permis de localiser le gène correspondant, *MUC4*, sur le chromosome 3 (Porchet *et al.*, 1991). La localisation de *MUC4* a ensuite été précisée en q29 (Gross *et al.*, 1992).

II.2.2.3.2 Le gène *MUC5B*

Des 20 clones obtenus précédemment, JER47, JER57 et JER58 ont été localisés en 11p15 (Tableau II ; Nguyen *et al.*, 1990). Les clones JER28 et JER57 ont permis d'isoler d'une banque de trachée humaine construite en vecteur λgt10 les 2

nouveaux clones JUL7 et JUL10. Ces 4 clones contiennent une répétition de 87 pb dont la dégénérescence, due à des insertions et des délétions nucléotidiques, engendre au niveau peptidique une alternance de zones hydrophobes et hydrophiles (Dufossé *et al.*, 1993).

II.2.2.3.3 Le gène *MUC5AC*

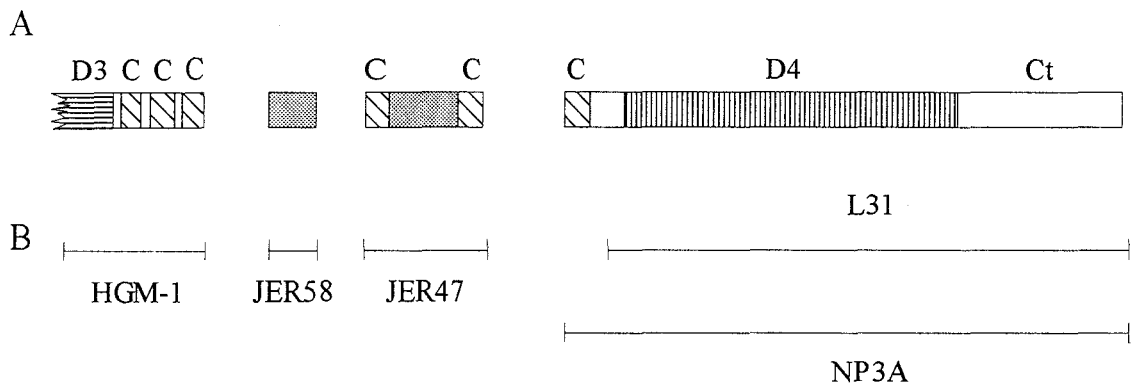
Des 20 clones isolés de la banque d'expression, les 2 clones JER47 et JER58 sont issus du gène *MUC5AC* (Guyonnet Dupérat *et al.*, 1995). Par ailleurs, Lesuffleur *et al.* ont isolé d'une banque préparée à partir de cellules muco-sécrétantes (cellules HT29-MTX, dérivées d'un cancer colique et traitées au méthotrexate) le clone d'ADNc L31 de 3310 pb et correspondant à l'extrémité C-terminale de MUC5AC (Lesufflleur *et al.*, 1995). Un clone très similaire en séquence, NP3a, avait été décrit auparavant par l'équipe de Rose (Meerzaman *et al.*, 1994). Il semble qu'un certain nombre d'erreurs de séquences avaient été commises par cette équipe. L'équipe de Strous a, quant à elle, isolé le clone d'ADNc HGM-1 de 2551 pb (Klomp *et al.*, 1995).

Ces séquences d'ADNc, probablement issues du même gène *MUC5AC* localisé en 11p15.5, ne sont pas encore toutes reliées entre elles physiquement. Le peptide déduit peut se décomposer en 3 domaines (Figure 10) :

- une région centrale constituée de l'alternance de régions répétées en tandem, riches en résidus Ser et Thr, et de motifs Cys d'environ 110 aa et identiques aux 2 motifs Cys retrouvés dans MUC2

- une région C-terminale dont un ADNc a été séquencé complètement ; cet ADNc code un motif Cys (~110 aa) suivi de domaines du vWF très semblables à ceux retrouvés dans l'extrémité C-terminale de MUC2

- une région amino-terminale qui reste incomplète vers le côté 5' du gène. Le peptide déduit de l'ADNc (HGM-1) est similaire à la partie 3' du domaine D3 du vWF, domaine retrouvé aussi en amont de la région répétitive de MUC2. Le peptide contient dans sa région 3' trois motifs Cys identiques au motif retrouvé dans la région C-terminale de MUC5AC, trouvé 2 fois dans MUC2 et au moins 3 fois dans la région centrale de MUC5AC (Guyonnet Dupérat *et al.*, 1995).



: zone répétitive
C : motif Cystéine (110 aa)
D4 : domaines retrouvés dans le vWF

Figure 10 : Organisation peptidique de MUC5AC déduite de l'agencement hypothétique des sous-clones d'ADNc de *MUC5AC*

II.2.2.4 La mucine d'origine gastrique MUC6

A partir d'une banque commerciale d'ADNc gastrique préparée en vecteur d'expression λgt11, un clone a été isolé par immunocriblage (Toribara *et al.*, 1993). Sa séquence révèle un motif de 507 pb répété en tandem codant un motif peptidique répété de 169 aa (Tableau II). Il s'agit du plus grand motif décrit pour une apomucine. La méthode d'hybridation fluorescente *in situ* a montré que le gène codant cette apomucine est, comme *MUC2*, *MUC5AC* et *MUC5B*, localisé en 11p15.4-11p15.5.

Très récemment, Toribara *et al.* ont publié l'organisation génomique de la région 3' de *MUC6* (Toribara *et al.*, 1997). Ils ont isolé un clone génomique contenant dans sa partie 5' des répétitions caractéristiques de *MUC6* et une séquence unique dans sa partie 3'. Par la méthode de «3'RACE-PCR», sur de l'ARN total d'estomac et en utilisant un oligonucléotide synthétisé à partir de la région répétitive de *MUC6*, un ADNc de 1735 pb a été cloné. Le clone génomique est constitué de 3 exons ; le premier étant le plus long et contenant la région répétitive dans sa partie 5' et une région de 805 pb dans sa partie 3' ne présentant pas de motif répété. Le peptide déduit C-terminal (en aval de la région répétitive) de MUC6 peut se décomposer en 2 domaines : un premier domaine de 270 aa entièrement codé par l'exon 1, riche en Ser-Thr-Pro (61,5%) et sans cystéine. Le second domaine (91 aa), codé par les 2 exons suivants, est pauvre en Ser-Thr-Pro (21,7%) mais contient 11 résidus Cys (12%) aux positions conservées par rapport aux C-terminaux de MUC2 et MUC5AC.

II.2.2.5 Les mucines salivaires et le gène *MUC7*

La cavité orale est la porte d'entrée du système digestif et constitue donc le premier segment du tube digestif. Parce que les mucines salivaires peuvent être facilement récoltées, et dans des conditions non forcément pathologiques, la bouche est un modèle privilégié d'étude du mucus et des mucines (pour revues, voir Tabak, 1995; Amerongen *et al.*, 1995; Levine *et al.*, 1987). Les rôles dévolus au mucus salivaire sont nombreux (voir tableau III).

Tableau III (Levine *et al.*, 1987)

Fonctions proposées pour la salive
1 Nettoyage physique et mécanique de la cavité orale
2 Lubrification et barrière perméable
a) recouvre les tissus mous et concentre les molécules protectrices
b) forme un fin film de rétention et amortit les chocs
c) intervient dans la formation de l'émail dentaire
3 Modulation de la flore orale
a) clearance sélective et adhésion de la flore microbienne
b) action anti-microbienne
c) utilisation comme substrat
4 Action anti-acide et neutralisation d'éléments nuisibles
a) pouvoir tampon
b) formation d'un complexe avec le tanin
5 Régulation de l'équilibre calcium/phosphate
a) intracellulaire : maturation des granules de sécrétion
b) extracellulaire : minéralisation
6 Digestion
a) acuité du goût
b) neutralisation du contenu oesophagien
c) dilution du contenu gastrique
d) formation du bol alimentaire
7 Processus extracellulaire de maturation des molécules salivaires

Tableau IV (Loomis *et al.*, 1987)

Composition en aa des mucines salivaires		
	Nombre / 1000 résidus	
Composés	MG1	MG2
Asp	66	48
Thr	163	207
Ser	100	160
Glu	86	69
Pro	92	232
Gly	83	14
Ala	78	142
Cys	23	1
Val	61	31
Met	10	1
Iso	26	15
Leu	71	28
Tyr	23	1
Phe	28	9
Lys	28	14
His	21	9
Arg	34	7
Trp	7	12

Tableau V (Loomis *et al.*, 1987)

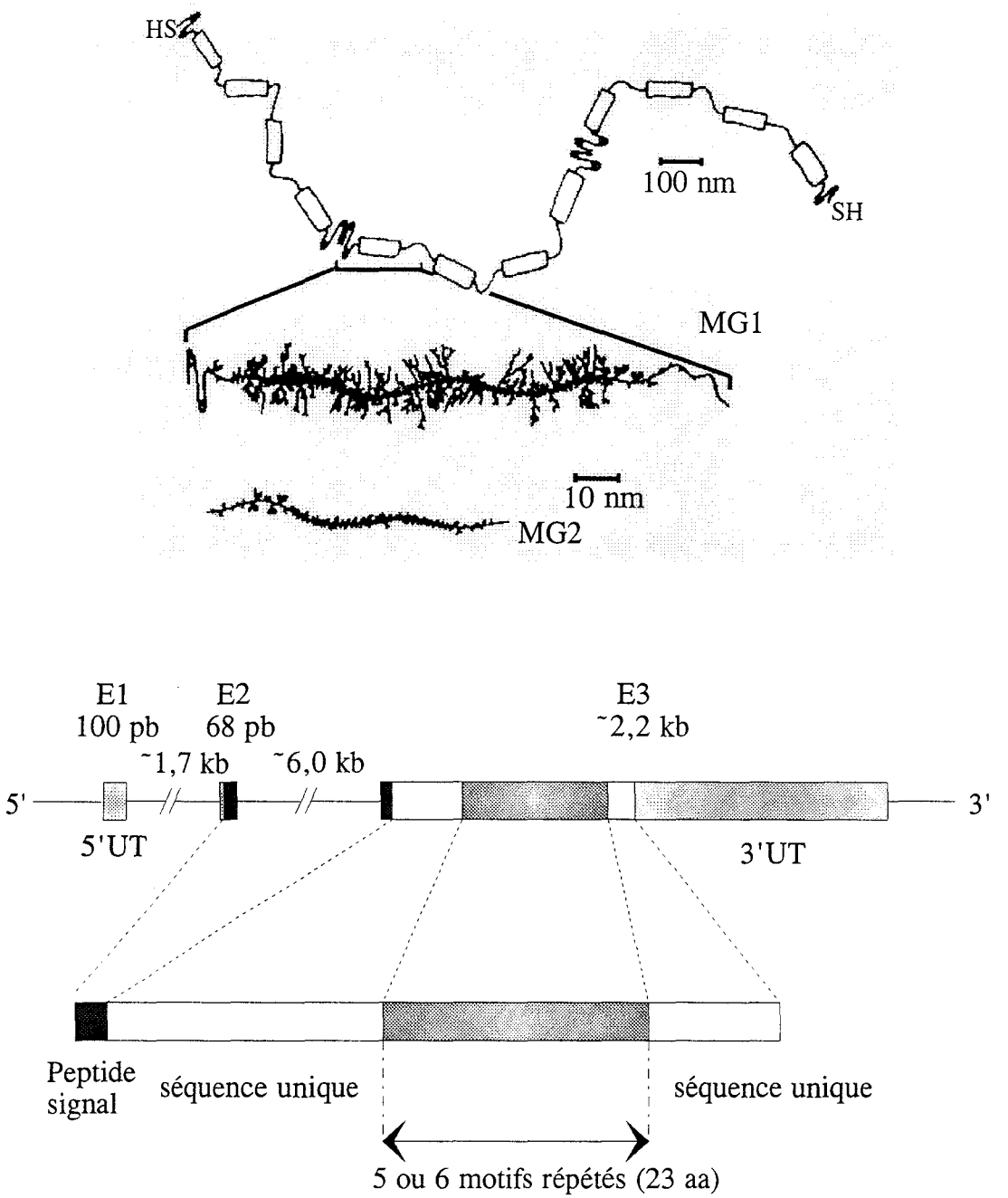
Comparaison des compositions chimiques et des caractéristiques des mucines salivaires		
	MG1	MG2
Protéine	14,9%	30,4%
O-glycanne	78,1%	68,0%
Sulfate	7,0%	1,6%
Acides gras	0,06%	Negligeable
Taille (kDa)	>1000	200-250
Sous-unités	Oui	Non
Résidus/chaîne saccharidique	4-16 résidus	2-7 résidus

Deux catégories de mucines ont été caractérisées: une mucine de haut poids moléculaire, appelée MG1, et une mucine de faible poids moléculaire, appelée MG2 (Tabak *et al.*, 1982; Prakobphol *et al.*, 1982; Loomis *et al.*, 1987; Eversole, 1972). Ces 2 mucines sont structurellement et fonctionnellement distinctes (Al-Hashimi *et al.*, 1989 ; voir tableaux IV et V). En effet, MG2 est trouvée sous forme de monomères alors que MG1 est trouvée sous forme de polymères reliés par des ponts disulfures dans les parties amino et carboxy-terminales (voir Figure 11A).

II.2.2.5.1 La mucine MG2 est codée par le gène *MUC7*

En 1982, Prakobphol *et al.* (Prakobphol *et al.*, 1982) ont purifié et caractérisé MG2, une *O*-glycoprotéine de type mucine, faiblement acide et de faible poids moléculaire. Ces auteurs ont montré que MG2 est une entité monomérique non associée de façon covalente à d'autres molécules et ayant une M_r d'environ 225 000. L'axe peptidique représente 30% du poids sec de la mucine, l'apomucine étant principalement composée de résidus Thr, Ser, Pro et Gly (75%). La composante glycannique est constituée en majorité de *N*-acétylgalactosamine, *N*-acétylglucosamine, galactose, fucose, acide sialique et acide *N*-acétylneuraminique. Des traces de mannose ont également été décelées. Des études préliminaires suggéraient que MG2 pouvait être, dans la cavité orale, une molécule d'adhésion privilégiée pour des souches de streptocoques (Levine *et al.*, 1978). Ces travaux ont été confirmés depuis par d'autres laboratoires (Tabak, 1995) et attribuent à MG2 un rôle essentiellement dans la modulation de la clearance bactérienne dans la cavité orale.

Des clones d'ADNc ont été isolés par immunocriblage d'une banque d'expression de glande sous-maxillaire humaine à l'aide d'un immunsérum de lapin dirigé contre l'axe peptidique de MG2 (Reddy *et al.*, 1992, Bobek *et al.*, 1993). Le transcrit du gène correspondant (*MUC7*), d'une taille de 2,7 kb, a pu être entièrement séquencé grâce à 2 clones chevauchants et un clone obtenu par la technique de



E : Exon
UT : région non-traduite

Figure 11 : Représentations schématiques (A) des mucines salivaires MG1 et MG2 (Amerongen *et al.*, 1995), (B) de l'organisation génomique de *MUC7* et (C) de la protéine déduite MG2.

"RACE-PCR" qui a permis de cloner l'ADNc 5' de *MUC7* (Bobek *et al.*, 1996). Le gène *MUC7*, localisé sur le chromosome 7 en q13-q21, s'étend sur environ 10 kb. Il comprend 2 introns d'environ 1,7 et 6 kb et 3 exons de respectivement 100 pb, 68 pb et environ 2,2 kb. Le premier exon ainsi qu'une partie du second correspondent à la partie 5' non traduite (Figure 11B et C). L'exon 2 et le début de l'exon suivant code un peptide signal de 20 aa. L'apomucine MUC7 (377 aa) est caractérisée par la répétition en tandem (5 ou 6) d'un motif de 23 aa. Cette région répétitive est encadrée par 2 séquences uniques de type mucine (riches en Ser et Thr). La partie 3' non traduite s'étend, sans compter la queue polyA, sur 1,1 kb. Les études de polymorphisme montrent que la majorité des allèles contiennent 6 répétitions en tandem et une minorité d'individus ont un allèle de 6 répétitions et l'autre de 5; ce gène n'est donc pas vraiment de type VNTR. Le territoire d'expression de *MUC7* semble réduit aux glandes salivaires et à la trachée (Bobek *et al.*, 1993 ; Biesbrock *et al.*, 1997 et Buisine, communication personnelle).

II.2.2.5.2 La protéine MG1 et le clone d'ADNc pSM2-1

Des études comparables à celles effectuées sur MG2 ont permis de montrer que la seconde mucine exprimée dans la cavité orale, MG1, a une structure plus complexe (Loomis *et al.*, 1987). MG1 est une molécule de haut poids moléculaire qui se trouve dans la salive sous forme d'homopolymères. MG1 est associée de façon non-covalente à des IgA sécrétoires. Son axe peptidique est composé essentiellement de résidus Ser, Thr et Pro (voir tableau IV), et contient, au regard de MG2, un pourcentage non négligeable de résidus Cys dont certains seraient impliqués dans les ponts disulfures formant le polymère. De plus, MG1 possède des régions nues et accessibles. Ces régions apparaissent hydrophobes, plus riches en aa chargés négativement et en aa aromatiques. La réduction des polymères de MG1 accroît le nombre de ces sites hydrophobes. MG1 possède probablement aussi des régions globulaires. Les chaînes *O*-glycosidiques sont constituées de 4 à 16 résidus comprenant des résidus de *N*-acétylgalactosamine, de *N*-acétylglucosamine, de galactose, et d'acide sialique.

MG1 pourrait jouer le rôle de lubrifiant, pourrait participer activement à la formation de l'émail dentaire (Al-Hashimi *et al.*, 1989, Kawagishi *et al.*, 1990) et à la protection de l'épithélium et de l'émail contre les caries (voir tableau III).

En décembre 1995, Troxler *et al.* publient la séquence d'un ADNc partiel (pSM2-1) obtenu par immunocriblage d'une banque d'ADNc de glande salivaire à l'aide d'un immunsérum dirigé contre la partie peptidique déglycosylée de MG1. Ce clone d'ADNc de 588 pb code une protéine riche en aa acides (Asp et Glu), basiques (Lys et Arg), contenant de nombreux résidus Cys (Troxler *et al.*, 1995). Cette séquence semble très similaire aux séquences des parties carboxy-terminales des mucines codées par les gènes localisés en 11p15. Bien que déjà en 1963 S. Kent avait émis l'hypothèse, à partir de ses études par anticorps fluorescents, qu'une mucine était exprimée à la fois au niveau des glandes salivaires, des glandes trachéales de la sous-muqueuse, de la vésicule biliaire, des conduits pancréatiques, des glandes de Bartholin et des glandes cervicales (Kent, 1963) Troxler *et al.* n'ont pas étudié l'expression de leur ADNc par Northern blot sur les tissus correspondants. Ils ont, par contre, montré que pSM2-1 n'était pas exprimé au niveau de l'intestin, du foie, de l'estomac, du rein et du coeur (Troxler *et al.*, 1995).

II.2.2.6 Le gène *MUC8*

Un immunsérum dirigé contre de la mucine trachéobronchique humaine déglycosylée chimiquement a permis à Shankar *et al.* d'isoler en 1994 un clone d'ADNc appelé pAM1 (Shankar *et al.*, 1994) à partir d'une banque d'expression construite en vecteur Uni-ZAP. Ce clone code un peptide de 313 aa, peptide composé d'une répétition (22 fois) en tandem d'un motif dégénéré de 13 aa suivi d'une séquence de 16 aa (Tableau I). Le motif répété est riche en Ser, Thr et Pro et comprend un résidu Cys par répétition. Par hybridation *in situ*, le gène correspondant et proposé pour *MUC8*, a été localisé sur le chromosome 12 en q24.3 (Tableau II ; Shankar *et al.*, 1995).

Par la méthode de 3'RACE-PCR, la même équipe a isolé un clone d'ADNc correspondant à la partie C-terminale de MUC8 (Shankar *et al.*, 1997). Ce clone contient, en aval du motif répété dégénéré de 41 pb, une région non traduite de 458 pb et une queue poly(A)$^+$.

II.2.3 Expression des gènes de mucines humaines

L'étude de l'expression des mucines humaines en situation normale et pathologique est étudiée par hybridation *in situ*, Northern et Dot blots et en immunohistochimie depuis plusieurs années et confirme l'intérêt que peuvent avoir ces gènes comme marqueurs de diagnostic et/ou de pronostic.

II.2.3.1 En situation normale

Il a été montré une spécificité tissulaire de l'expression des gènes de mucines, mais aussi une spécificité cellulaire. Par exemple, *MUC4* s'exprime indifféremment dans les cellules caliciformes et les glandes mucipares de la plupart des muqueuses alors que, au contraire, *MUC5B* a une expression tissulaire limitée aux glandes sous-maxillaires et aux bronches, à la vésicule biliaire, au pancréas et une expression cellulaire limitée aux glandes. L'ensemble des résultats est résumé par le tableau VI.

Puisque le mucus, et donc les mucines, recouvrent les muqueuses des tractus respiratoire, digestif et urogénitaux, il n'est pas surprenant que la production de mucus et de mucines soit perturbée quantitativement et qualitativement lors de pathologies affectant ces organes. Par exemple, les maladies bronchiques de l'adulte s'accompagnent de modifications du mucus. Au cours de la bronchite, le poids sec des sécrétions augmente ainsi que le volume de la phase gel (Reid *et al.*, 1982). Les sulfomucines puis les mucines neutres augmentent aux dépens des sialomucines (Lamblin *et al.*, 1977a). Au cours de la mucoviscidose, on observe une augmentation des sialo et sulfomucines (Lamblin et al., 1977b).

	SONDES						
MUQUEUSES HUMAINES	**MUC 2**	**MUC 3**	**MUC 4**	**MUC 5B**	**MUC 5AC**	**MUC 6**	**MUC 7**
Glandes salivaires	G : -	G : -	G : -	G : +/++	G : -	G : -	G : +++
Bronches	S : ++	S : -	S : ++	S : -	S : -/+++	S : -	S : -
	G : -	G : -	G : -	G : ++	G : -	G : -	G : +++
Fundus	S : -	S : -/+	S : -/+	S : -	S : ++++	S : - Ct : +	S : -
	G : -	G : -	G : -	G : -	G : -	G : -	G : -
Antre pylorique	S : -	S : ++	S : -/+	S : -	S : ++++	S : -	S : -
	G : -/+	G : +	G : -	G : -	G : -	G : ++/+++	G : -
Duodénum	S : ++++	S : +++	S : -	S : -	S : -	S : -	S : -
	C : ++++	C : -/+	C : -	C : -	C : -	C : -	C : -
	G : +	G : -	G : -	G : -	G : -	G : ++/+++	G : -
Jéjunum	S : ++++	S : +++	S : -	S : -	S : -	S : -	S : -
	C : ++++	C : -/+	C : -	C : -	C : -	C : -	C : -
Iléon	S : ++++	S : +++	S : -/+	S : -	S : -	S : -	S : -
	C : ++++	C : -/+	C : -/+	C : -	C : -	C : -	C : -
Côlon	S : ++++	S : ++	S : ++	S : -	S : -	S : -	S : -
	C : ++++	C : -	C : ++	C : -	C : -	C : -	C : -
Vésicule biliaire	S : +	S : +++	S : -	S : +	S : +	S : -/+	S : -
	I : +	I : +++	I : -	I : ++	I : +	I : +++	I : -
Pancréas	D : -/+	D : +++	D : -	D : ++	D : -	D : +	
Prostate	G : -	G : -	G : ++	G : -	G : -	G : -	G : -
Endocol	S : +	S : -	S : ++	S : +	S : ++	S : +	G : -
	G : +	G : -	G : ++	G : +	G : ++	G : +	S : -

S : épithélium de surface; G : épithélium glandulaire; C : cryptes; I : invaginations de l'épithélium; Ct : collet;
D : cellules des canaux.
Marquage : ++++ : de très forte intensité; +++ : de forte intensité; ++ : d'intensité modérée; + : de faible intensité;
-/+ : quelques cellules; - : absent,

Tableau VI: Profil d'expression des gènes de mucines chez l'adulte par hybridation in situ. (Audié *et al.*, 1993; Audié *et al.*, 1995; Balagué *et al.*, 1995 ; Ho *et al.*, (soumis) ; Vandenhaute *et al.*, 1997)

Les études des pathologies touchant ces organes ont souvent montré des anomalies de glycosylation des mucines mais les nouvelles données concernant les gènes de mucines entrouvrent de nouvelles voies de recherche qui permettront de mieux comprendre l'implication des mucines dans les pathologies des épithéliums.

Il ne s'agit pas ici de donner une liste exhaustive des premiers résultats et nous nous limiterons à la pathologie la plus illustrée dans la littérature jusqu'à présent pour les apomucines : les cancers épithéliaux.

II.2.3.2 En situation pathologique : l'exemple des cancers épithéliaux

Plus de 90% des tumeurs humaines se développent aux dépens des cellules épithéliales. Lors du processus de cancérisation, 5 catégories d'altérations de la fraction glycosylée peuvent être relevées (pour revues voir Ho *et al.*, 1991; Ho *et al.*, 1995a) :

- augmentation du nombre d'antigènes en comparaison avec le tissu sain
- résurgence d'antigènes foetaux
- expression d'antigènes incompatibles avec les déterminants de groupe sanguin du patient
- disparition d'antigènes exprimés dans les tissus sains, dévoilant de nouvelles structures antigéniques qui peuvent alors être la cible d'un traitement anticancéreux. Ces structures peuvent aussi être utilisées comme vaccin.
- synthèse de nouveaux antigènes.

De nombreuses études à des fins thérapeutiques ont été menées ces dernières années sur MUC1. En effet, le gène correspondant est surexprimé dans la majorité des tumeurs mammaires. Le produit d'expression est donc une cible thérapeutique potentielle intéressante. En effet, lorsque la protéine transmembranaire est clivée, la fraction soluble devient antigène circulant et marqueur tumoral (CA15.3) détectable par de nombreux anticorps. Ces anticorps permettent aujourd'hui le dépistage précoce des récidives et le suivi du traitement des formes métastasées. Les nouveaux épitopes

peptidiques et glycosidiques rencontrés lors de la cancérogenèse font de MUC1 un outil précieux comme cible des essais d'immunothérapie (Scholl *et al.*, 1997). Des études comparables sur les autres mucines fourniront de nouveaux outils de diagnostic et/ou de pronostic du cancer ainsi que de nouvelles cibles potentielles d'agents anticancéreux.

Trois grands types d'altération dans la structure des mucines ont été décrites : glycosylation anormale, accumulation de précurseur (glycosylation incomplète) et expression anormale des apomucines. Ce dernier point est de plus en plus étudié grâce aux connaissances acquises concernant les gènes de mucines. Il a été montré, par exemple, que MUC4 et MUC5AC, normalement exprimés dans les canaux pancréatiques le sont plus fortement dans les tumeurs pancréatiques (Balagué *et al.*, 1995). Des variations d'expression des gènes de mucines ont aussi été observées au cours des processus cancéreux sur des tissus et des lignées cellulaires (Yu *et al.*, 1996; Nguyen *et al.*, 1996; Balagué *et al.*, 1994; Ho *et al.*, 1993; Ogata *et al.*, 1992; Niv, 1994). Par exemple, des variations d'expression importantes de *MUC5AC* ont été détectées dans des biopsies de tumeurs villeuses rectosigmoïdiennes. Ce gène est normalement non exprimé dans cette muqueuse. Une expression aberrante a été observée dans les zones lésées et dans certaines zones « saines ». Ces tumeurs sont bénignes mais possèdent un haut potentiel de récidive et de malignité. Ainsi, *MUC5AC* peut être considéré comme un marqueur précoce de transformation car son expression est fonction du grade de dysplasie de la tumeur (Buisine *et al.*, 1996). Pour *MUC5B*, les études faites dans le laboratoire sur des muqueuses foetales nous amènent à envisager que dans les processus cancéreux il puisse y avoir expression et augmentation d'expression de ce gène dans les cancers de l'estomac et de l'oesophage respectivement.

Les mucines augmentent le pouvoir d'adhésion des cellules transformées, stimulent l'activité collagénasique et invasive des cellules tumorales. Il a été suggéré que les mucines protègent les cellules tumorales contre le système immunitaire en masquant des épitopes et en inhibant l'activité des cellules NK (pour "Natural Killer").

II.2.4 Les mucines animales

II.2.4.1 Les mucines de rongeurs

II.2.4.1.1 L'homologue murin *Muc1*

La protéine déduite de l'ADNc de *Muc1* présente des homologies avec la protéine MUC1 humaine particulièrement pour les régions transmembranaires et cytoplasmiques (87%). Ceci suggère un rôle très important de ces domaines. A noter que des homologies similaires ont été trouvées pour les protéines MUC1 d'autres espèces (Gendler *et al.*, 1995). Par contre, la région répétitive est moins homologue à celle de MUC1 (seulement 34%) et est plus courte. De plus, le nombre de répétitions de 20 ou 21 aa ne varie pas d'un animal à l'autre. La structure génomique du gène murin *Muc*1 est-elle aussi très proche de celle du gène humain *MUC1* : les introns ont la même position et sont, en séquence, homologues de 50 à 62% pour les 5 premiers. Les régions promotrices de *MUC1* et *Muc1* sont homologues à 72%, ce qui n'est pas surprenant car ces 2 gènes ont des profils d'expression superposables. Enfin, le gène murin *Muc1* est localisé sur le chromosome 3 dans la région qui correspond à la région du chromosome 1 humain où a été localisé le gène *MUC1* (Gendler *et al.*, 1995).

II.2.4.1.2. L'homologue murin *Muc5ac*

Un immunsérum dirigé contre de la mucine purifiée et déglycosylée chimiquement d'estomac de souris a permis d'isoler un clone d'ADNc d'une banque d'expression (Shekels *et al.*, 1995). La séquence nucléotidique correspondante code un motif répété de 16 aa (Tableau VII) suivi de deux motifs en tandem d'une centaine d'aa et riches en résidus Cys. Ces 2 motifs d'environ 110 aa sont similaires aux motifs Cys retrouvés dans MUC2 et MUC5AC. Le gène correspondant, *Muc5ac*, a été localisé sur le chromosome 7 de la souris qui regroupe des gènes localisés chez l'Homme en 11p15.

Gène ou ADNc	AA	pb	Séquence consensus	Espèce	Référence
Muc-2	11	33	SPTTSPTTSTT	rat	Ohmori *et al.*, 1994
rMUC176	6	18	TTTPDV	rat	Gum *et al.*, 1991
CTM	6	18	TPTPTP et TTTTPV	chien	Verma *et al.*, 1993
PSM	3 et 5	9-15	TEATT et GTT	porc	Eckhardt *et al.*, 1991
pGBM	127	381	ALRLVNGSDRCQGRVEVLYGGSWGTVCDD SWDTNDANVVCRQLGCGWAISAPGDARFG QGSGPIVLDDVGCSGYETYLWSCSHSPWN THNCGHSEDASVICSASQTQSTVVPDLWY PTTDYGTESGL	bovin	Nunes *et al.*, 1995
BSM	5-11	15-33	TTSLG et GTTVAPGSSNT	bovin	Bhargava *et al.*, 1990
FIM-A.1	9	27	VPTTPETTT	xénope	Hoffman, 1988
FIM-B.1	11	33	GESTPAPSETT	xénope	Probst *et al.*, 1992
FIM-C.1	8	24	TTTKATTT	xénope	Hauser *et al.*, 1992
Mucsmg	13	39	PTTDSTTPAPTTK	rat	Albone *et al.*, 1996
Muc5ac	16	48	QTSSPNTGKTSTISTT	souris	Shekels *et al.*, 1995
PGM	16	48	SVQPSSSSSXPTTS(A/T)T	porc	Turner *et al.*, 1995
SM2-1R	13	39	NATTPAPTTKPTT	souris	Denny *et al.*, 1996

Tableau VII : Séquences consensus des domaines répétitifs de quelques mucines animales.

II.2.4.1.3 La sialoglycoprotéine ASGP

Les sialomucines ASGP-1 et -2 ("Ascites Sialomucin Glycoproteins") sont 2 *O*-glycoprotéines qui, isolées à la surface de cellules d'un adénocarcinome mammaire de rate (Hull *et al.*, 1990), sont associées l'une à l'autre. Ces 2 glycoprotéines pourraient jouer un rôle de protection des cellules cancéreuses. De plus, elles sont impliquées dans le processus métastatique et leur expression est associée à la résistance de la cellule cancéreuse aux cellules NK (Moriarty *et al.*, 1990).

L'ASGP-2 a été caractérisée comme étant une molécule transmembranaire. En fait, les 2 glycoprotéines, ASGP-1 et -2 sont issues d'un même précurseur polypeptidique. Sheng *et al.*, en s'appuyant sur leurs études biochimiques, ont proposé un modèle dans lequel le précurseur peptidique appelé pSMC-1 (Figure 12 ; Sheng *et al.*, 1990) est synthétisé dans le réticulum endoplasmique et *N*-glycosylé (Bharathan *et al.*, 1990). Ce précurseur est ensuite *O*-glycosylé et est clivé pour donner les 2 entités ASGP-1 et -2, et ce probablement avant d'arriver dans le compartiment *médian* du Golgi. Ces 2 molécules restent associées par des liaisons non covalentes. La même équipe a ensuite montré que l'ASGP-1, une fois à la surface de la cellule épithéliale sécrétrice, peut être internalisée rapidement dans un compartiment intracellulaire de la cellule où il y a addition et élongation de nouvelles chaînes oligosaccharidiques (Hull *et al.*, 1991).

Par immunocriblage d'une banque d'expression en vecteur λgt11, un premier clone d'ADNc de 1,3 kb a été isolé. L'insert s'hybride en Northern blot à un transcrit d'environ 9 kb (Sheng *et al.*, 1992). La protéine déduite contient 24 sites potentiels de *N*-glysosylation, un peptide hydrophobe de 20 aa prédit comme transmembranaire ainsi que 2 séquences EGF-like ayant des résidus Cys aux positions conservées. Il est connu que de nombreuses molécules possèdent des domaines EGF-like mais qu'elles n'ont pas pour autant une activité de type EGF. Cependant, les similarités entre les séquences consensus des molécules possédant un ou plusieurs motifs EGF-like et ayant une activité de type EGF et les motifs EGF-like de l'ASGP-2, laissent supposer une telle activité autocrine de facteur de croissance pour l'ASGP-2. Les premiers

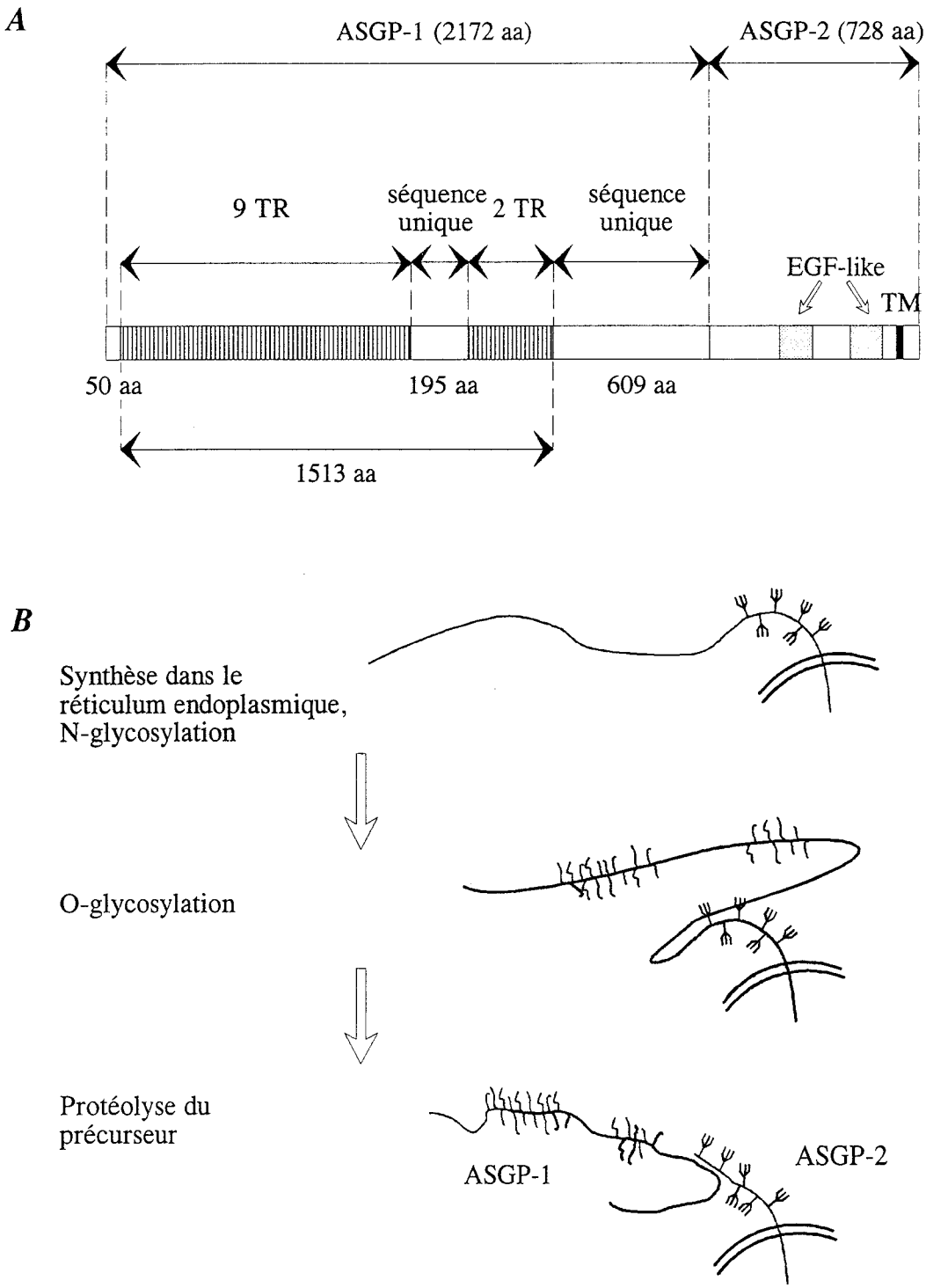


Figure 12 : (A) Schéma d'organisation du précurseur peptidique pSMC-1 et (B) modèle de biosynthèse de l'ASGP-1 et de l'ASGP-2 (TM : transmembranaire ; TR : « tandem repeat »).

éléments de preuve d'une telle activité ont été apportés par l'équipe de Carraway (Juang *et al.*, 1996) puisque l'ASGP-2 transfecté dans des cellules d'insectes se lie au domaine extracellulaire du récepteur tyrosine-kinase $p185^{neu}$.

L'ADNc complet de pSMC-1 a ensuite été séquencé (Wu *et al.*, 1994). Le transcrit a une taille de 9270 pb. L'ASGP-1 déduit contient une séquence signal suivie d'une courte région non répétitive puis d'un motif irrégulier répété 11 fois de 117 à 124 aa. Ce motif, riche en résidus Ser et Thr, a une homologie de 70 à 90% avec la séquence consensus déduite de celui-ci. Ces motifs répétés sont entrecoupés entre la 9ème et la 10ème répétition par une courte séquence unique. L'organisation peptidique du précurseur pSMC-1 et le modèle de biosynthèse de cette sialomucine sont schématisés sur la figure 12.

Fin 1996, l'équipe de Carraway (Rossi *et al.*, 1996) a montré par des études sur la protéine et par la technique de RT-PCR que le complexe ASGP-1/ASGP-2 existe sous une forme sécrétée et membranaire dans les glandes mammaires et le lait mais sous une forme uniquement soluble dans le côlon. Cette dernière forme résulte de l'absence du domaine C-terminal de l'ASGP-2, c'est-à-dire de la région transmembranaire. En conclusion, l'ASGP semble désormais le meilleur candidat pour la mucine non encore caractérisée de haut poids moléculaire décrite dans le lait et proposée sous le nom de MUCX.

II.2.4.1.4 Le gène *Muc2* chez le rat

* La région 5' de *Muc2*

Afin d'étudier la régulation de la transcription des mucines, l'équipe de Basbaum aux Etats-Unis a isolé, en utilisant une sonde humaine de *MUC2* correspondante à la région en amont de la zone répétitive, trois ADNc dont 2 sont chevauchants. Le peptide déduit du cadre de lecture ouvert de 4546 pb est très similaire à l'extrémité 5' de MUC2 (Ohmori *et al.*, 1994). Outre les domaines D1-D2-D'-D3 du pro-vWF, l'ADNc de *Muc2* code en sa région 3' un motif d'environ 110 aa et contenant 10 résidus Cys, c'est-à-dire le motif retrouvé 2 fois dans MUC2 et au

moins 6 fois dans MUC5AC. La région 3' de l'ADNc code une région semi-répétitive riche en Thr, Ser et Pro.

* Le peptide de liaison et le clone MLP

L'équipe de Forster a isolé le peptide de liaison de 118 kDa à partir de mucine intestinale de rat purifiée puis réduite. Des oligonucléotides ont été synthétisés à partir de séquences peptidiques partielles des extrémités N-terminales du peptide natif et de ce peptide hydrolysé par le bromure de cyanogène. Le produit d'amplification de 1,2 kb obtenu par PCR a permis de cribler une banque d'ADNc et d'isoler un clone de 2,6 kb, appelé MLP ("mucin link protein"), contenant un signal de polyadénylation suivi, 14 nucléotides en aval, d'une queue polyA$^+$ (Xu *et al.*, 1992b). L'extrémité 5' code un motif répété de type mucine. Le reste de l'ADNc code un domaine riche en sites potentiels de *N*-glycosylation (13) et en résidus Cys. Ce domaine est très similaire aux régions C-terminales de MUC5AC et MUC2 démontrant que le peptide de liaison est probablement la résultante de l'hydrolyse de la région carboxy-terminale de certaines mucines. De plus, une partie de cette région utilisée comme sonde a montré une hybridation dans la région p15.5 du chromosome humain 11 (Xu *et al.*, 1992a) apportant une nouvelle preuve de l'appartenance du peptide de liaison à une ou à des mucines humaines localisées en 11p15.

* Le clone VR-1A

Un immunsérum, préparé à partir de mucine intestinale de rat déglycosylée, a permis de cribler une banque d'ADNc et d'isoler un clone positif de 705 pb et appelé VR-1A. Il code un peptide de 235 aa constitué dans sa région 3' d'un motif Cys homologue aux motifs Cys de 110 aa retrouvés dans MUC2 et MUC5AC. Il est suivi d'une région de 182 aa riche en résidus Thr et Ser mais qui ne contient pas de résidus Cys.

Les 2 clones VR-1A et MLP sont probablement issus du même gène *Muc2*. En effet, l'insert du clone VR-1A et la sonde MLP s'hybrident tous les deux sur un fragment de 500 kb lorsque l'ADN est hydrolysé avec l'enzyme de restriction *Mlu*I et sur un fragment de 380 kb lorsque l'ADN est hydrolysé avec l'enzyme de restriction *Sal*I. De plus, ces 2 clones ont le même profil d'expression en Northern Blot. Enfin, le

gène correspondant à VR-1A a été localisé sur le chromosome 1 du rat (Hansson *et al.*, 1994). Klings-Levan *et al.* ont ensuite confirmé la co-localisation sur le chromosome 1 de rat de VR-1A et MLP (Klings-Levan *et al.*, 1996).

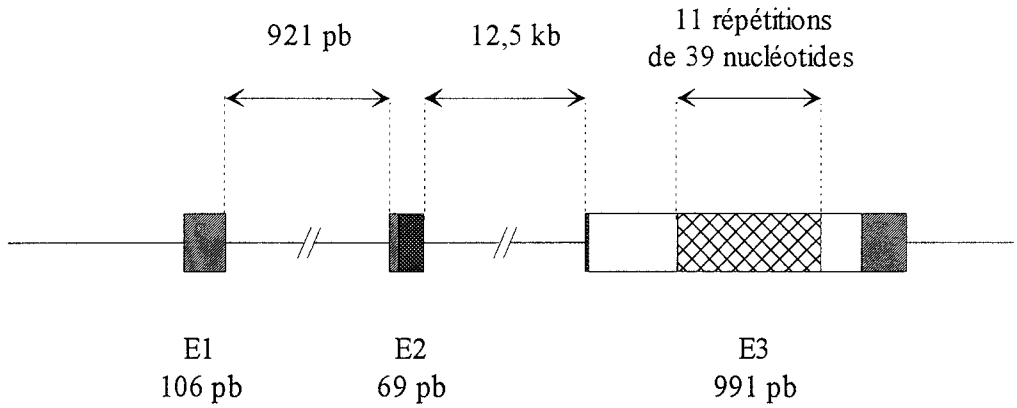
II.2.4.1.5 Le clone Rmuc176

Un immunsérum préparé à partir de mucine de jéjunum de rat déglycosylée chimiquement a permis d'isoler 4 clones d'ADNc par l'équipe de Y. Kim à San Francisco (Gum *et al.*, 1991). Les 4 clones codent un même motif répété de 6 aa. Le plus grand des clones, RMUC176, a une taille 1341 pb et code en aval de la partie répétitive une région unique de 92 aa.

II.2.4.1.6 Le gène de mucine sous-maxillaire de rat *Mucsmg*

L'équipe de Tabak a purifié en 1994 à partir de glande sous-maxillaire de rat une mucine de faible poids moléculaire (114 kDa). A partir de la séquence peptidique partielle de la mucine déglycosylée, un oligonucléotide antisens dégénéré a été synthétisé et a permis de cribler par PCR asymétrique une banque d'ADNc en utilisant comme seconde amorce un oligonucléotide du vecteur. Un produit d'amplification de 201 pb a été obtenu et a permis de cribler à nouveau une banque d'ADNc. Quatre nouveaux clones de plus grande taille ont été obtenus (Albone *et al.*, 1994). Ainsi, l'ADNc complet a été cloné et séquencé. Il a une taille de 1228 pb sans la queue poly(A)$^+$. La région 5' non traduite de 134 pb est suivie par une séquence codant un peptide signal de 22 aa, puis d'une région, polymorphe d'un animal à l'autre, codant un peptide de 104 aa riche en résidus Gln, Pro et Tyr suivi de la région répétitive (motif de 13 aa répété 11 fois) et enfin d'une région C-terminale codant 117 aa.

Le gène correspondant a ensuite été cloné en vecteur phagique par la même équipe (Albone *et al.*, 1996) et la répartition des exons et introns déterminée. Le gène, localisé sur le chromosome 14 du rat, comprend 3 exons; un premier exon de 106 pb et non codant, séparé du second exon de 69 pb par un intron de 921 pb. L'exon 3 de 991



E : Exon
▒ : Région non traduite
■ : Séquence codant le peptide signal
□ : Séquence unique
⊠ : Séquence codant la région répétitive

Figure 13 : Organisation génomique du gène de rat *Mucsmg* et organisation peptidique déduite.

pb comprend toute la partie répétitive du gène et est séparé de l'exon 2 par un intron d'environ 12,5 kb (Figure 13).

II.2.4.1.7 La mucine sous-maxillaire de souris

L'un des clones d'ADNc de rat du gène *Mucsmg* a été utilisé comme sonde pour cribler une banque d'ADNc de glande sous-maxillaire de souris. 27 clones ont été isolés et étudiés. L'analyse des séquences complétée par des expériences de PCR permet de reconstituer l'intégralité du transcrit qui a une taille de 1042 pb sans la queue poly(A)$^+$. Ce transcrit code une petite apomucine de 273 aa. L'ADNc est très homologue au transcrit de *Mucsmg*, particulièrement dans les régions 5' et 3' non traduites. Chacun des dix motifs répétés a la particularité de contenir un site potentiel de *N*-glycosylation (Denny *et al.*, 1996).

II.2.4.2 Les mucines de bovin et de porc

II.2.4.2.1 La mucine sous-maxillaire de bovin BSM

Un transcrit complet de 2 kb a été isolé à partir d'une banque de glande sous-maxillaire de bovin et cloné (Bhargava *et al.*, 1990). La séquence peptidique déduite de 563 aa se compose d'un premier domaine de 310 aa typique des mucines (riche en Ser, Thr et Pro) suivi d'un domaine riche en résidus Cys, domaine homologue à la région C-terminale de FIM-B.1 et aux C-terminaux des mucines humaines MUC2 et MUC5AC localisées en 11p15.5 (voir plus loin). Ce domaine C-terminal est constitué d'un domaine C-like de 140 aa retrouvé dans le vWF suivi d'un domaine dit noeud cystine ou encore CK (pour «Cystine knot») (Gendler *et al.*, 1995 ; Figure 14).

II.2.4.2.2 La mucine de vésicule biliaire de bovin

A l'aide d'un immunsérum dirigé contre la partie peptidique déglycosylée chimiquement de mucine de vésicule biliaire de bovin, 25 clones ont été isolés et étudiés (Nunes *et al.*, 1995). Tous les peptides déduits montrent l'agencement en 2 régions. La première région est constituée d'un motif répété de 127 ou 131 aa. Ce

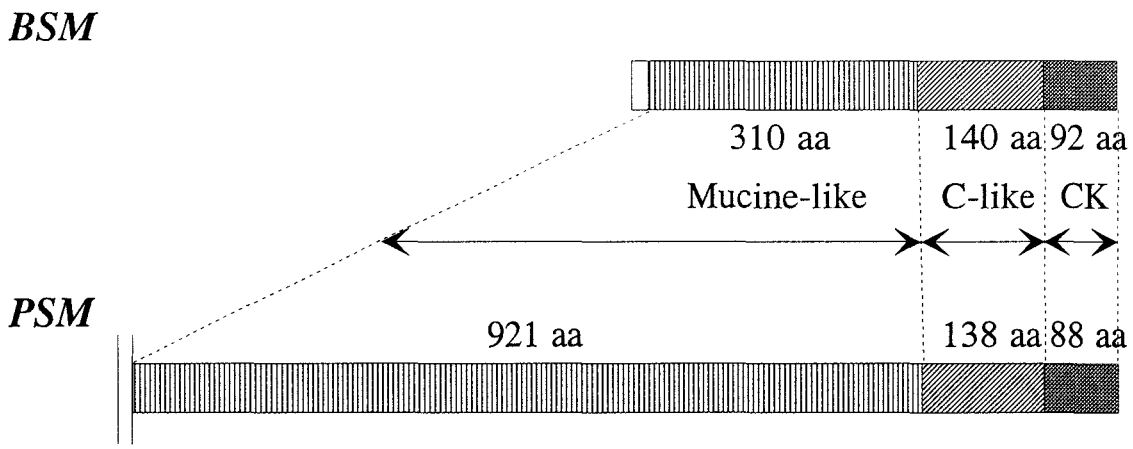


Figure 14 : Organisations peptidiques des mucines sous-maxillaires de boeuf et de porc

motif contient toujours 8 résidus Cys aux positions conservées. La seconde région est composée d'un motif répété de 20 aa ; ce motif est riche en résidus Thr et ne contient aucune cystéine. Les quelques différences en séquence observées dans ces clones suggèrent qu'ils proviennent de différentes régions du même gène. Un fragment d'un des ADNc a été utilisé pour cribler la même banque d'ADNc. Les dix clones alors isolés contiennent 5 fois le motif répété en tandem de 60 pb codant le peptide de 20 aa décrit précédemment. De plus, chaque clone contient le motif de 131 aa avec 8 résidus Cys. Ainsi, l'apomucine correspondante est formée par l'alternance de répétitions en tandem de motifs de 131 aa et de motifs de 20 aa.

II.2.4.2.3 La mucine sous-maxillaire de porc PSM

L'équipe de Hill a cloné un ADNc de 3,65 kb d'une banque d'expression de glande sous-maxillaire de porc (Eckhardt *et al.*, 1991). Le peptide déduit de 1150 résidus contient un domaine de type mucine de 921 aa constitué de répétitions parfaites d'un motif de 81 aa. Ce domaine est suivi de la région C-terminale très homologue à celle décrite chez le bovin (BSM) (Figure 14). La même équipe a ensuite démontré par transfection de la région C-terminale de 240 aa de PSM, dans des cellules COS7, que cette apomucine dimérise dans le réticulum endoplasmique par cette région et que cette dimérisation est indépendante de la *N*-glycosylation de ce domaine (Perez-Villar *et al.*, 1996).

II.2.4.2.4 La mucine gastrique de porc

Dix clones (0,6 à 2,5 kb) d'ADNc d'une banque d'expression d'estomac de porc ont été isolés à l'aide d'un immunsérum dirigé contre de la mucine gastrique de porc purifiée et déglycosylée chimiquement (Turner *et al.*, 1995). Deux de ces clones, PGM-9B et 2A, ont été étudiés. Le peptide déduit de PGM-9B est uniquement constitué de répétitions en tandem (33 fois) d'un motif de 16 aa riche en résidus Ser (47%). Le peptide déduit du clone PGM-2A comprend 16 répétitions en tandem du motif trouvé dans PGM-9B. Cette séquence est précédée de la partie 3' d'un motif cystéine homologue aux motifs cystéines d'une centaine d'aa déjà décrits pour MUC2 et MUC5AC.

II.2.4.3 La mucine trachéobronchique canine

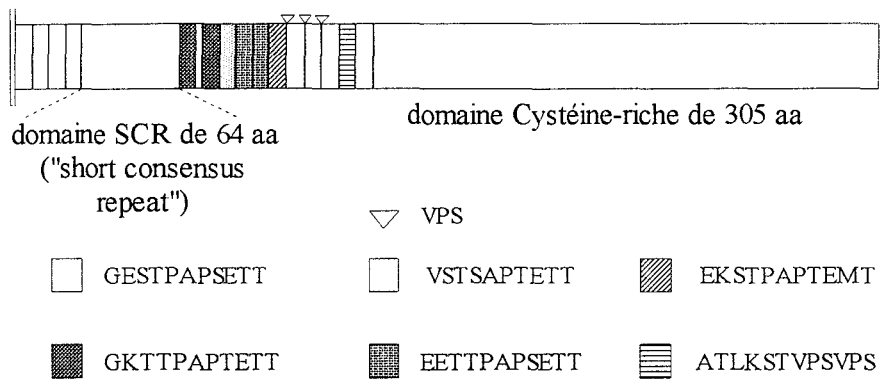
Grâce à un immunsérum dirigé contre de la mucine, de haut poids moléculaire, trachéobronchique de chien et déglycosylée, un ADNc de 1,7 kb codant la région C-terminale de la mucine a été isolé et étudié en 1992 par l'équipe de Sachdev (Shankar *et al.*, 1992). Le peptide déduit de 445 aa ne contient pas de répétitions en tandem et ne présente aucune homologie avec d'autres séquences.

Par la même technique, Verma et Davidson (Verma *et al.*, 1993) ont isolé des clones d'ADNc de mucine trachéobronchique de chien. Par la méthode de RACE-PCR, à partir de la séquence du plus grand des clones d'ADNc (1,8 kb), la séquence complète du transcrit d'un gène de mucine trachéobronchique a été déterminée. Il a une taille de 3766 pb et code une petite apomucine de 1118 aa. Cette apomucine contient quelques motifs TPTPTPTG ou TTTTTT(M/V) non répétés en tandem. Les régions riches en résidus O-glycosylables (Ser/Thr) de cette mucine et de MUC2 sont très homologues. L'apomucine contient 3 sites potentiels de *N*-glycosylation et de nombreux résidus Cys, particulièrement dans la région C-terminale.

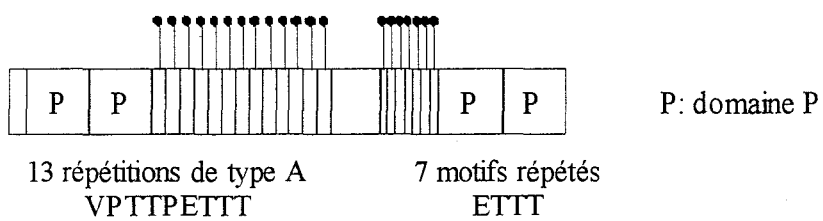
II.2.4.4 Les mucines tégumentaires de *Xénope*

Jusqu'à présent, 3 mucines tégumentaires de xénope (FIM pour "Frog Integumentary Mucin") ayant différents éléments répétitifs ont été caractérisés et nommés FIM-A.1, FIM-B.1 et FIM-C.1 (Figure 15 ; Tableau VII). Chacune de ces FIMs est constituée de sous-domaines riches en résidus Thr et Cys (pour revues voir Hoffman *et al.*, 1993b; Hoffman *et al.*, 1995; Gendler *et al.*, 1995).

FIM-B.1



FIM-A.1



FIM-C.1

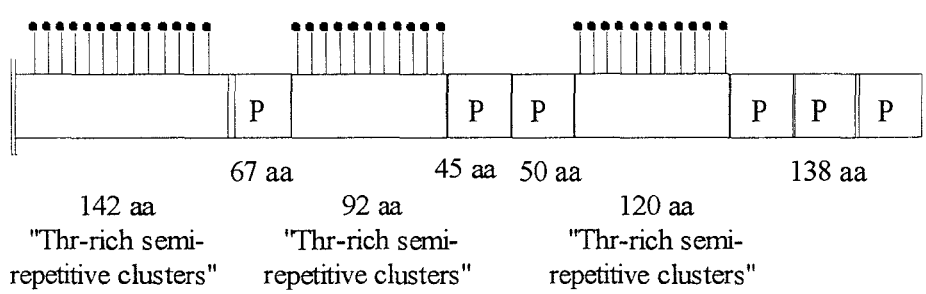


Figure 15 : Représentation schématique des protéines codées par les 3 gènes de mucines tégumentaires de *X. laevis* (Gendler *et al.*, 1995)

II.2.4.4.1 Les domaines P

FIM-A.1 et FIM-C.1 contiennent des domaines «P» d'environ 50 aa et riches en Cys (6 résidus) aux positions très conservées. Ces mêmes domaines ont été trouvés dans le peptide spasmolytique de pancréas de porc. Ces 6 résidus invariants de Cys forment 3 ponts intra-chaînes donnant naissance à une structure dite peptide en trèfle («trefoil peptide») qui confère à la protéine une grande résistance à la dégradation protéolytique. Ce motif est un modulateur potentiel de la croissance cellulaire. Des protéines contenant ce domaine P et appelées pS2 pour peptides spamolytiques ont été trouvées chez l'Homme, le rat, le porc, la souris ainsi que chez le xénope (protéines xP2 et xP4) (Hoffman *et al.*, 1993b). Elles semblent promouvoir la migration des cellules épithéliales lorsque l'épithélium est endommagé (Dignass *et al.*, 1994). Normalement exprimés à la surface des épithéliums par les cellules caliciformes (Suemori *et al.*, 1991; Podolsky *et al.*, 1993), les peptides en trèfle sont aussi exprimés par d'autres tissus dans des situations pathologiques (Hoffman *et al.*, 1993a).

Chez l'Homme, les gènes codant les peptides en trèfle (*hITF*, *pS2* et *hSP*) ont été localisés sur le chromosome 21 en q22.3 (Chinery *et al.*, 1996). Ces 3 gènes sont situés sur un fragment commun de 55 kb (Seib *et al.*, 1997). Les 2 gènes *pS2* et *hITF* comportent 3 exons, le domaine P étant entièrement codé par l'exon 2. Le gène *hSP*, au centre de ce cluster de gènes, comporte quant à lui 4 exons; les exons 2 et 3 codent 2 domaines P (Figure 16).

II.2.4.4.2 FIM-A.1

FIM-A.1 semble, des 3 mucines caractérisées chez le xénope, la mucine prédominante. Appelée à l'origine spasmolysine (Hoffman, 1988), cette protéine mosaïque de 43 kDa est constituée d'une séquence signal suivie de 2 domaines P puis d'une zone répétitive de «type A» de 13 aa riches en résidus Thr puis de 7 répétitions d'une séquence ETTT et enfin de 2 autres domaines P (Figure 15). La protéine mature a une masse d'environ 130 kDa.

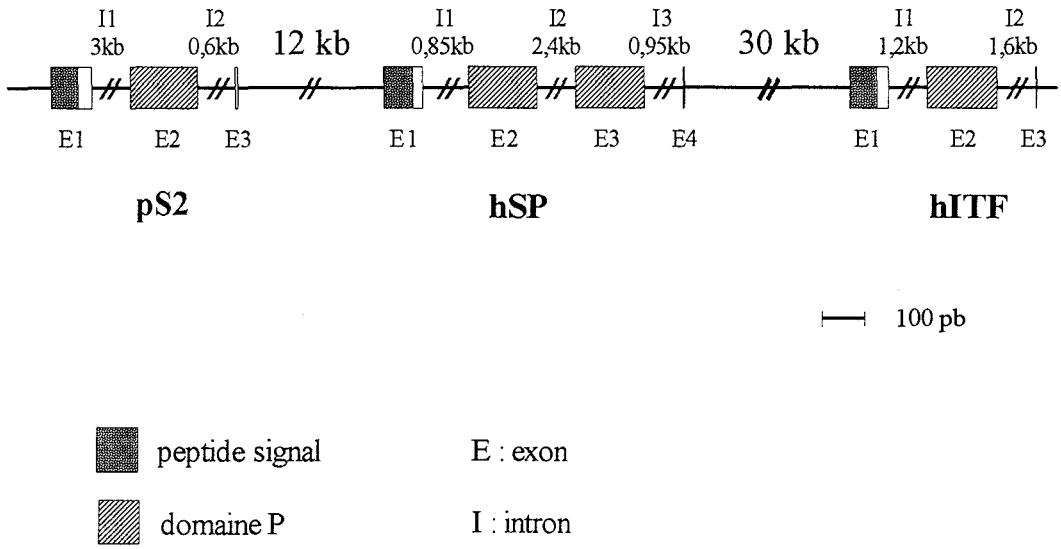


Figure 16 : Organisation génomique des 3 gènes humains codant les peptides en trèfle (Seib *et al.*, 1997)

II.2.4.4.3 FIM-C.1

Une séquence partielle de FIM-C.1 et correspondant à la partie C-terminale de la protéine a été publiée en 1992 par Hauser et Hoffman (Hauser *et al.*, 1992). FIM-C.1 contient au moins 3 régions semi-répétitives riches en résidus Thr (80%) dont le motif prédominant est TTTKATTT (Tableau VII) ainsi que 6 domaines P (Figure 15). De plus, la protéine FIM-C.1 et le gène correspondant sont polymorphes et ceci est probablement dû, selon ces mêmes auteurs, à des variations interindividuelles de longueur des régions répétitives (polymorphisme de type VNTR).

II.2.4.4.4 FIM-B.1

FIM-B.1 n'a pas la même construction que FIM-A.1 et FIM-C.1. Aucun domaine P n'y a été trouvé. C'est la première mucine décrite comme ayant dans sa partie C-terminale des homologies avec le facteur vWF (Probst *et al.*, 1990). La protéine FIM-B.1 est en majeure partie constituée par un motif répété "acide" de 11 aa (Figure 15 ; tableau VII) présentant des similarités avec le motif répété caractérisant FIM-A.1. Ces répétitions, au nombre variable selon l'animal (polymorphisme de type VNTR), sont entrecoupées en un endroit par une région appelée SCR ("Short Consensus Repeat") de 64 aa. Ce motif, contenant 4 résidus Cys, formant probablement des ponts disulfures, ainsi que des résidus Phe, Tyr, Gly et Trp, a été décrit dans de nombreuses protéines dont la plupart appartiennent au système du complément et interagissent avec le C3b ou le C4b. Ce motif a aussi été décrit dans d'autres molécules n'appartenant pas au système du complément comme la β_2-glycoprotéine, l'haptoglobine, le récepteur de l'interleukine-2, le polypeptide sécrété de 35 kDa du virus de la vaccine et la sous-unité b du facteur XIII (Probst *et al.*, 1992).

La région C-terminale de FIM-B.1 est similaire aux régions C-terminales de MUC2 et du pro-vWF. Récemment, la partie codante de la région 5' de FIM-B.1 a été entièrement clonée et séquencée (Joba *et al.*, 1997). FIM-B.1 a unr région amino-terminale semblable aux N-terminaux de MUC2 et du pro-vWF.

II.2.5 Similarités des domaines peptidiques des mucines humaines

II.2.5.1 Les gènes de mucines en 11p15

Des 8 gènes de mucines humaines, 4 ont été localisés sur le même chromosome, en 11p15.5. Une étude en champs pulsés au laboratoire en utilisant comme sonde les régions répétées en tandem des gènes, a permis de déterminer l'ordre de ces 4 gènes sur le chromosome 11 et de les localiser par rapport aux autres gènes de la région (Pigny *et al.*, 1996a) :
Télomère-*HRAS-MUC6-MUC2-MUC5AC-MUC5B-IGF2*-Centromère (Figure 17).
Pour cela, notre équipe a étudié les îlots CpG, encore appelés "HTF" (pour "*Hpa* II tiny fragment"). En effet, ces îlots indiquent la présence de gènes ou pseudogènes car ils leur sont presque toujours associés. Ces îlots sont en outre facilement repérables grâce aux nombreux sites de coupures d'enzymes à site rare, sites riches en G et C. Les îlots CpG signalent très souvent un gène à expression ubiquitaire et sont alors situés dans la région 5' de ces gènes. Ils y couvrent alors tout ou partie des premiers exons et comportent généralement des séquences transcrites (Larsen *et al.*, 1992; Craig *et al.*, 1994; Jordan, 1991).

L'organisation peptidique de MUC2 est connue, celle de MUC5AC vient d'être déterminée pour sa région carboxy-terminale. Quelques éléments concernant sa région amino-terminale ont été publiés (Klomp *et al.*, 1995). Les 2 produits peptidiques des gènes respectifs ont une construction identique: la région centrale, comportant la zone riche en résidus Ser et Thr est encadrée par 2 régions uniques riches en résidus Cys et très similaires aux domaines C- et N-terminaux du pro-vWF, en particulier par la position des résidus Cys.

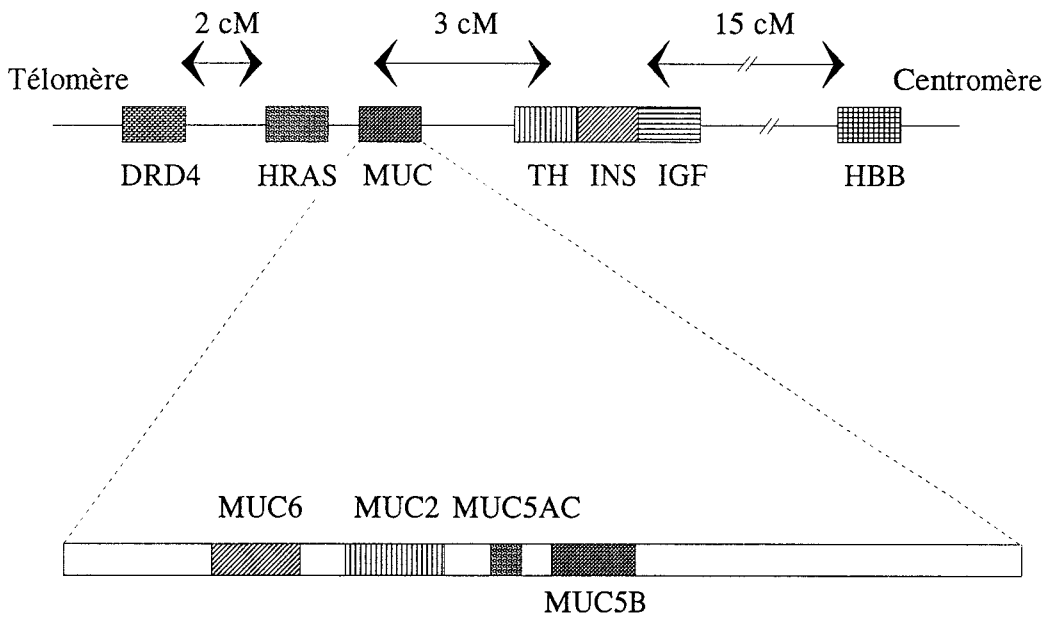


Figure 17 : Organisation du «cluster» de gènes en 11p15.5 (Pigny *et al.*, 1996a)

II.2.5.1 Le prépro-facteur von Willebrand

Le vWF est une grande glycoprotéine d'adhésion qui forme des homopolymères ayant 2 rôles biologiques distincts: le vWF sert d'intermédiaire dans l'adhésion des plaquettes et dans la formation des caillots sanguins lors des lésions vasculaires, et il sert de transporteur du facteur VIII de coagulation dans la circulation sanguine. Il est fabriqué par les mégacaryocytes et les cellules endothéliales où il est stocké dans des organelles spécifiques, les corpuscules de Weibel-Palade. Ce stockage en fait sa particularité puisque le vWF est la seule molécule d'adhésion qui est stockée.

II.2.5.1.1 Organisation générale

Le gène du vWF s'étend sur 178 kb et contient 52 exons (Figure 18). Il a été localisé sur le chromosome 12 en p12-pter. Une séquence homologue et correspondant à un pseudogène du vWF (exons 23 à 34) a été localisée en 22q11-13 (Mancuso *et al.*, 1991). Le transcrit du *vWF* a une taille de 8,7 kb.

Les régions codantes du gène ainsi que toutes les jonctions exon/intron ont été déterminées (Mancuso *et al.*, 1989). Le codon initiateur est dans le second exon; le produit de traduction est un précurseur polypeptidique de 2813 aa (prépro-facteur). Il consiste en un peptide signal de 22 aa, un inhabituel grand pro-peptide de 741 aa et une unité mature de 2050 résidus (pour revues, voir Ruggeri *et al.*, 1993; Mayadas *et al.*, 1991 ; Nichols *et al.*, 1997).

Le pro-peptide avait initialement été identifié comme l'antigène II du vWF. Le rôle de cet antigène, après clivage du peptide mature au niveau de la séquence peptidique KRS (Foster *et al.*, 1987; Verweij *et al.*, 1988), reste inconnu. Certains auteurs ont proposé que ce propeptide intervient dans le stockage du vWF dans des granules spécifiques (Wagner *et al.*, 1991).

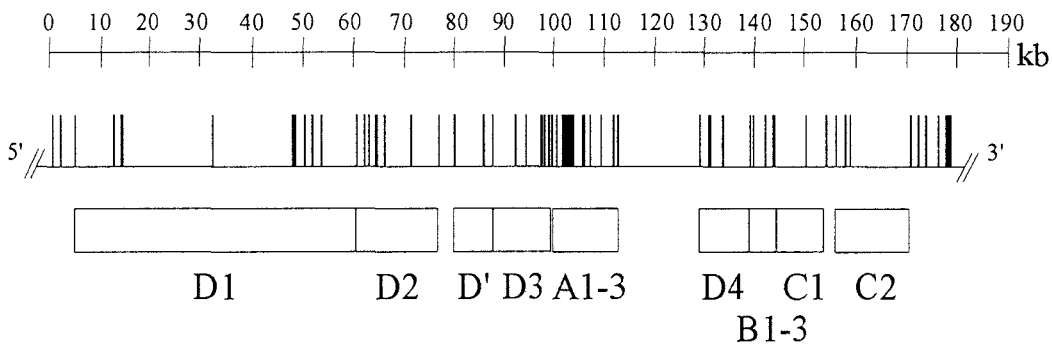


Figure 18 : Représentation schématique du gène vWF et des domaines peptidiques. Les rectangles verticaux représentent les exons (Mancuso *et al.*, 1989)

II.2.5.1.2 Les domaines du pro-vWF et des mucines

Le pro-vWF est constitué presque exclusivement de 4 types de domaines répétés suite à des phénomènes de duplications successives (Shelton-Inloes *et al.*, 1987) ou de brassage d'exons. Ces domaines sont désignés de A à D, et arrangés dans l'ordre suivant: D1-D2-D'-D3-A1-A2-A3-D4-B1-B2-B3-C1-C2 (Figure 18); la région constituée des 2 domaines D1-D2 constitue le pro-peptide (Shelton-Inloes *et al.*, 1986; Verweij *et al.*, 1986; Mancuso *et al.*, 1989). La région constituée des domaines D1-D2-D'-D3 est similaire à la région en amont de la zone répétitive de MUC2 et de FIM-B1 (Joba *et al.*, 1997). Un motif D3 a été décrit par Klomp *et al.* dans la zone en amont de la région répétitive de MUC5AC (Klomp *et al.*, 1995).

Les domaines D4, B et C du vWF ont par ailleurs été retrouvés avec le même agencement dans les parties C-terminales de MUC2, MUC5AC et FIM-B.1. Les domaines C ont été retrouvés dans les régions C-terminales des mucines animales PSM et BSM (pour revue, voir Gendler *et al.*, 1995). Cependant, les mucines contiennent un seul domaine B et un seul domaine C, au lieu des 3 domaines B et des 2 domaines C dans le vWF.

II.2.5.1.3 Le pro-vWF : une molécule mosaïque

* Le domaine A

Le domaine de type "A" du vWF définit une superfamille de protéines possédant un ou plusieurs exemplaires de ce module d'environ 200 aa (Colombatti *et al.*, 1991). Ce module confère une fonction moléculaire à des protéines:

- du système immunitaire (intégrines LFA-1, Mac-1, p150-95 des leucocytes, facteur B)
- du système hémostatique (vWF)
- de la machinerie d'adhésion cellulaire (intégrines VLA-1 et -2, protéine de la matrice cartilagineuse CMP, collagène de type VI).

* Le domaine C

Des homologies de séquence ont été rapportées entre les domaines C (~120 aa) du vWF, des procollagènes de types I et III et la thrombospondine (Hunt *et al.*, 1987), particulièrement en ce qui concerne 9 résidus Cys et, dans une moindre mesure, d'autres aa comme les résidus Trp, Tyr, Gly. L'une des caractéristiques biologiques communes de ces 4 molécules est la polymérisation puisque le vWF se trouve sous forme d'homopolymères. Les 3 autres peptides forment des trimères.

* Le domaine D

Les domaines D du vWF ont une taille d'environ 360 aa. Ils ont été trouvés dans la zonadhésine (Hardy *et al.*, 1995), une protéine membranaire du spermatozoïde du porc impliquée dans la reconnaissance de l'ovule. L'ADNc complet de 7785 pb a été entièrement séquencé. Il code une « mucin-like » de 2476 aa. On peut distinguer 4 régions: un peptide signal potentiel de 29 aa, une région extracellulaire de 2418 aa, un segment transmembranaire puis une région intracellulaire de 36 aa riche en acides aminés basiques. La région extracellulaire de 2418 aa est constituée sur presque 400 aa d'un motif imparfait répété 53 fois de 7 aa dont la séquence consensus est de type mucine : PTE(K/R)(P/T)T(V/I). Cette région est suivie par 5 domaines homologues en tandem appelés D0 à D5, car ces domaines sont similaires aux domaines D du pro-vWF, mis à part que le domaine D0 est beaucoup plus petit.

Le domaine D a été trouvé dans des molécules plus lointaines dans l'évolution. Les vitellogénines constituent une famille de protéines conservées des nématodes aux vertébrés (poulet, xénope). Chez les nématodes et chez l'oursin de mer, elles sont synthétisées et sécrétées dans l'intestin puis transportées jusqu'aux ovocytes. Chez les vertébrés, les vitellogénines sont synthétisées par le foie, transportées dans la circulation sanguine jusqu'aux ovocytes qui possèdent des récepteurs spécifiques de ces molécules. Elles sont les précurseurs des «yolk proteins», principales protéines utilisées par l'embryon comme source de nourriture (pour revue voir Byrne *et al.*, 1989).

Ces «yolk proteins» sont codées par une famille de gènes comprenant : 1 gène chez l'oursin, 2 chez le criquet, 3 chez la drosophile, le xénope et le poulet et 6 (dont un pseudogène) chez le nématode *Caenorhabditis elegans*. Le nombre de gènes dans une espèce est à mettre en relation avec le temps nécessaire pour former les ovocytes. Ces gènes ont, de plus, une expression qui varie en fonction du stade de développement, du sexe et du tissu (pour revues, voir Spieth *et al.*, 1991; Byrne *et al.*, 1989). Ces vitellogénines possèdent un domaine très similaire à la partie 3' du domaine D2 du pro-vWF (Baker, 1988), particulièrement en ce qui concerne la position des résidus Cys. Toutes ces molécules possèdent un motif peptidique CG(L/I)CG.

Toutes les protéines qui comportent des modules communs au pro-vWF sont des gènes mosaïques issus de la duplication de gènes et du brassage d'exons. Les domaines D ont pour origine un module ancestral qui existait déjà il y a plus de 450 millions d'années chez les vertébrés primitifs. Il est remarquable que les résidus Cys aient été conservés au cours de l'évolution, et plus particulièrement les cystéines vicinales CXXC.

II.2.5.1.4 Biosynthèse du vWF

Douze sites du pro-vWF sont *N*-glycosylés dans le réticulum endoplasmique (Titani *et al.*, 1986). Cette *N*-glycosylation semble nécessaire à la formation d'homodimères du pro-peptide par les régions C-terminales en aval du domaine C2 (domaines CK) dans le réticulum (Voorberg *et al.*, 1991). Les dimères sont ensuite *O*-glycosylés et sulfatés dans l'appareil de Golgi.

Les dimères se polymérisent dans le Golgi tardif. Ceci requiert la présence du pro-peptide et des cystéines vicinales présentes dans chacun des domaines D. Il est probable que les domaines D1-D2 de 2 dimères établissent des interactions non covalentes qui permettent la formation d'un pont disulfure entre les domaines D3 des 2 dimères (Mayadas *et al.*, 1992; Verweij *et al.*, 1987). Enfin, le pro-peptide de 741 aa est clivé dans le *trans*-Golgi et dans les granules de stockage (Figure 19; Mayadas *et al.*, 1991; Ruggeri *et al.*, 1993). Les polymères, stockés dans les granules de sécrétion,

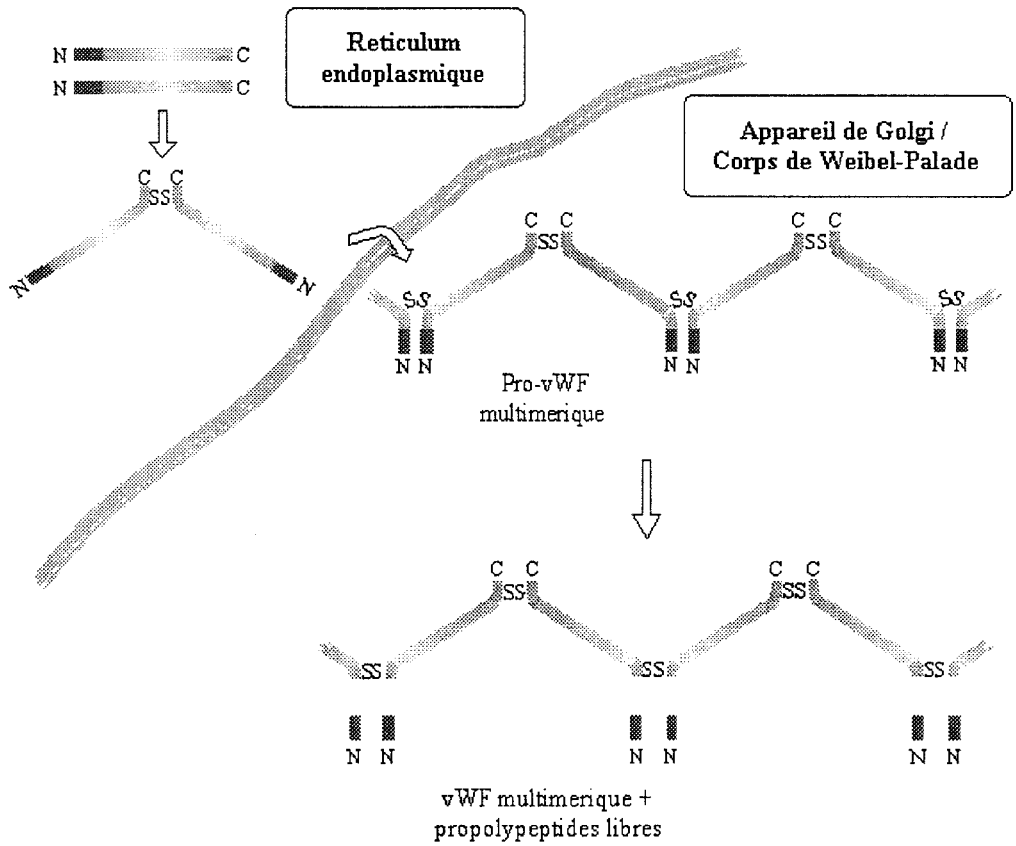


Figure 19 : Représentation schématique de la biosynthèse du vWF (Mayadas *et al.*, 1991)

sont constitués parfois de plus de 100 monomères. Ils sont sécrétés sous forme de longs filaments qui peuvent atteindre la taille de 1300 nm et sont visibles en microscopie électronique (Fretto *et al.*, 1986).

II.2.5.2 La superfamille des CK ou « noeuds cystines »

La superfamille des protéines à CK peut se décomposer en plusieurs familles (tableau VIII) dont la famille des TGF-β qui compte aujourd'hui plus de 40 membres impliqués dans des fonctions biologiques aussi diverses que la fonction immunitaire, le contrôle de la croissance, la différenciation cellulaire, la formation squelettique et la morphogenèse embryonnaire (pour revues voir Sun *et al.*, 1995; Massagué, 1990; Sporn *et al.*, 1992). Un alignement des séquences peptidiques de l'extrémité C-terminale de quelques membres de cette superfamille a été réalisé sur la figure 20.

II.2.5.2.1 Modélisation de la protéine NDP

Le gène *NDP* («Norrie Disease Protein») a été localisé par génétique inverse en Xp11.3-11.4. Il a une taille de 28 kb et un ARNm de 1.8 kb. Le gène contient 3 exons de respectivement 201, 380 et 1257 pb. Le 1er exon est non codant. Le cadre de lecture des 2 autres exons est ouvert sur 399 pb et code un peptide sécrété de 133 aa. Le fragment peptidique, appelé CK pour «Cystine knot» (ou noeud cystine) codé par l'exon 3 est riche en résidus Cys et est homologue aux extrémités C-terminales de nombreuses protéines extracellulaires : gènes précoces, CEF-10 chez le poulet, Cyr61 chez la souris, la protéine Slit chez la drosophile, le vWF et certaines mucines : FIM-B.1, MUC2, MUC5AC, BSM, PSM (Figure 20).

La protéine NDP est très conservée entre l'Homme et la souris, plus particulièrement pour la région CK où l'homologie atteint 99% (Chen *et al.*, 1995). Des mutations touchant dans ce domaine les résidus Cys détruisent la structure

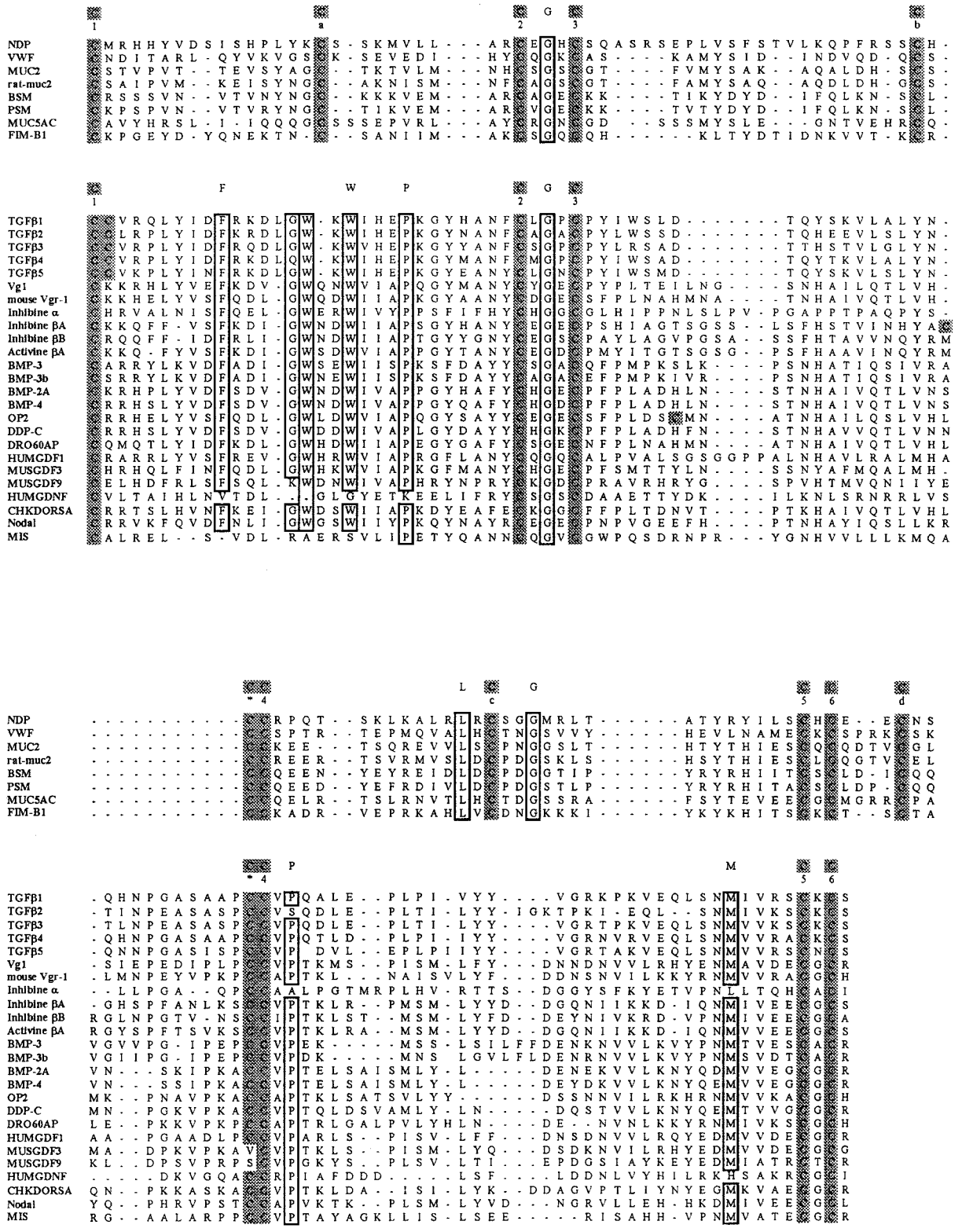


Figure 20 : Alignement des motifs CK

tridimensionnelle et sont à l'origine des phénotypes sévères, aboutissant à la cécité, à la surdité, au retard mental (Strasberg *et al.*, 1995; Meindl *et al.*, 1992).

La structure tridimensionnelle de la protéine NDP (Meitinger *et al.*, 1993) est superposable à la structure tridimensionnelle du TGF-β (Figure 21) déterminée auparavant par cristallographie (Schlunegger *et al.*, 1992). TGF-β et NDP ont 7 résidus Cys communs dont 6 sont engagés dans des ponts disulfures intrachaînes, le septième résidu Cys (noté avec un astérisque sur les figures 20 et 21A) formant, au moins pour le TGF-β, un pont interchaînes avec un autre monomère TGF-β. Ces noeuds maintiennent les feuillets β entre eux et contribuent à la résistance des protéines qui possèdent un module CK aux températures élevées, aux agents dénaturants et aux pH extrêmes.

II.2.5.2.2 La famille des TGF-β

* La sous-famille des facteurs TGF-β

Quatre TGF-β ont été caractérisés dont 3 (TGF-β 1 à 3) chez l'Homme et un chez le xénope. Un 5ème membre a été décrit chez le poulet comme homologue au TGF-β1 humain et appelé TGF-β4. Pro-peptides et peptides matures sont très conservés (82-86% et 95-99%) entre les 4 TGF-β et les homologues trouvés chez les animaux (Burt *et al.*, 1992).

Le précurseur peptidique d'environ 400 aa est clivé au niveau d'une séquence tri- (TGF-β1) ou tétrabasique (TGF-β2 à 5) pour donner naissance au peptide mature de 112 aa (114 pour le TGF-β4) (Kondaiah *et al.*, 1990) qui, selon le TGF, s'homo ou s'hétérodimérise (Tableau VIII) pour former le complexe actif.

Le profil d'expression des TGF-β est assez large. Par exemple, le transcrit du TGF-β2 est détecté dans la sous-muqueuse de la trachée, de l'intestin et de l'estomac, dans la peau, le cartilage et l'os.

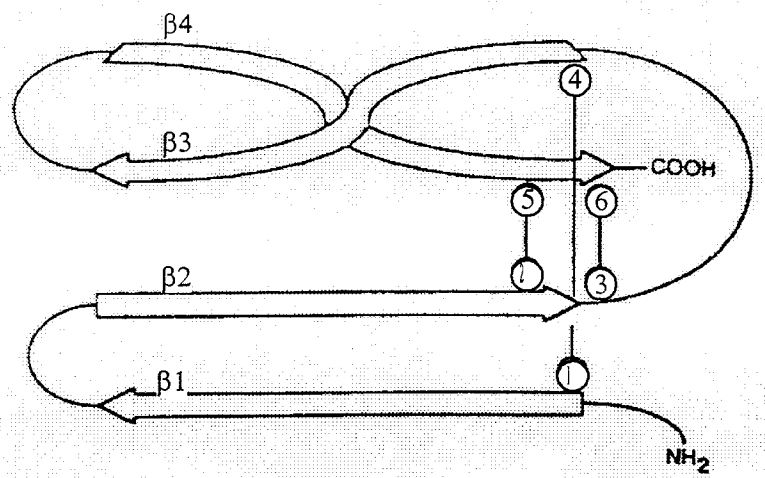
A

1 a 2 3 b *4 c 5 6 d

NDP : $_2$HN-X_{18}-C-X_{15}-C-X_9-C-X_1-G-X_1-C-X_{23}-C-X_1-CC-X_{23}-C-X_{15}-C-X_1-C-X_3-C-X_2-COOH

TGF-β : $_2$HN-X_{15}-C-----X_{26}-----C-X_1-G-X_1-C-----X_{29}-----CC------X_{30}-----C-X_1-C-X_1---------COOH

B



C **D**

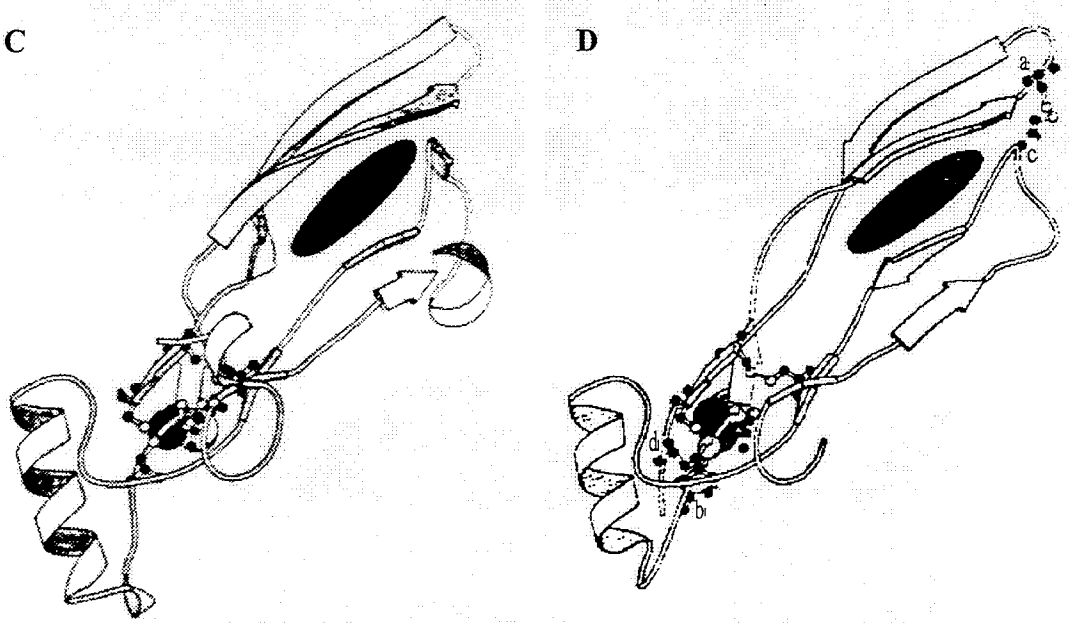


Figure 21 : Modélisation moléculaire de la protéine NDP (Meitinger *et al.*, 1993)
(A) Alignement des séquences du NDP et du TGF-β
(B) Représentation schématique de la structure secondaire du TGF-β
(C) Diagramme en ruban de la structure 3D expérimentale du TGF-β
(D) Diagramme en ruban de la structure 3D modélisée du NDP

sous-familles	Gène	Dimères bioréactifs		Origine	Chromosome	
		Nom	Composition		humain	souris
TGF-β	TGF-β1	TGF-β1	homodimère	humaine	19q13	7
	TGF-β2	TGF-β2	homodimère	humaine	1q41	1
	TGF-β3	TGF-β3	homodimère	humaine	14q24	2
	TGF-β4			poulet		
	TGF-β5	TGF-β5	homodimère	X. laevis		
		TGF-β1-β2	homodimère			
Inhibines	α	Inhibine A	dimère α.βA	humaine		
		Inhibine B	dimère α.βB			
	βA	Activine A	homodimère βA	humaine		
	βB			humaine		
		Activine AB	dimère βA.βB			
DVR	nodal			Souris (embryonnaire)		
	BMP2			Humaine		13
	BMP4			Humaine		5
	dpp-C			Drosophile		
	60A			Drosophile		
	Vg1			X. laevis		
	GDF1			Humaine		
	GDF3/Vgr2			Souris (os)		
	dorsaline			poulet		
	lefty			Souris		
Autres	MIS			Humaine	19	
	GDF9			Souris		
	GDNF			Humaine		

Tableau VIII : La super-famille des «noeuds cystines » de type TGF-β

L'activité de ces facteurs est contrôlée par la régulation de leur transcription, la production du facteur sous forme latente et la séquestration par des protéines de la matrice extracellulaire et des protéines circulantes. Deux récepteurs transmembranaires qui lient le TGF-β avec une forte affinité ont été caractérisés.

* La sous-famille des inhibines.

Activines et inhibines sont des polypeptides dimériques formés d'une sous-unité α associée soit à une sous-unité βA (inhibine A) ou βB (inhibine B). Ces hétérodimères inhibent la production d'hormone FSH, de stéroïdes et d'hormones placentaires. Des dimères βA-βA et βB-βB ont été décrits et auraient un rôle antagoniste aux inhibines et ont été appelés pour cette raison activines.

* La sous-famille des DVR

La sous-famille des DVR comprend des gènes BMP («bone morphogenetic protein»), les gènes Nodal, dpp-C («decapentaplegic»), 60A, Vg1, GDF1, GDF3, lefty et de la dorsaline. Ces facteurs ont une activité inductrice sur la formation du cartilage et de l'os (Rosen *et al.*, 1992; Zhou *et al.*, 1993; McPherron *et al.*, 1993; Kingsley, 1994; Sampath *et al.*, 1993).

* Les autres protéines

A ces groupes de protéines, il faut ajouter le gène MIS, codant l'hormone anti-müllérienne produite par les cellules de Sertoli et responsable de l'atrophie des canaux de Müller, le gène GDF9 dont les transcrits sont synthétisés seulement dans les ovaires des animaux adultes et le GDNF («glial derived neurotrophic growth factor») qui a la propriété de promouvoir la survie et la différenciation des neurones dopaminergiques du tronc cérébral.

II.2.5.2.3 La famille des CK de type NGF

Bien que les séquences peptidiques du NGF («nerve growth factor») et du TGF-β soient très peu similaires, leurs 2 structures tridimensionnelles déterminées par cristallographie (McDonald *et al.*, 1991; Schlunegger *et al.*, 1992) sont superposables (Daopin *et al.*, 1992). Le NGF possède, dans sa région C-terminale, 6 résidus Cys au

lieu de 7 dans le TGF-β. Le résidu manquant correspond à la cystéine impliquée dans la dimérisation du TGF-β. En effet, le NGF possède la propriété de former des dimères par sa région C-terminale et ce, sans former de liaison covalente (McDonald *et al.*, 1993). Ainsi, tous les membres de la famille du NGF font partie d'une sous-famille des protéines à CK.

II.3 Biosynthèse des mucines

II.3.1 Régulation de la transcription

Mis à part pour le gène MUC1, très peu de travaux ont jusqu'à présent pu être menés sur la régulation de la transcription des gènes de mucines. Quatre publications récentes concernent la régulation de MUC2 et celle d'une mucine trachéobronchique canine (TBM). Cependant, plusieurs laboratoires ont cloné ces derniers mois des gènes humains et animaux de mucines. Il est donc probable que d'autres études de la régulation verront le jour très prochainement.

II.3.1.1 Régulation de la transcription de *MUC1*

Quelques laboratoires étudient la régulation de la transcription du gène *MUC1*. Plusieurs régions, en amont du site d'initiation de la transcription, ont été caractérisées comme jouant un rôle dans la régulation. Une séquence, appelée E-MUC1 et localisée entre les nucléotides -84 et -72 pb, semble déterminer la spécificité d'expression tissulaire du gène (Kovarik *et al.*, 1993). Une seconde région, en -505/-485, lie une protéine de 45 kDa (Abe *et al.*, 1993). Cette région est chevauchante avec une séquence similaire retrouvée dans les séquences promotrices de gènes codant des protéines du lait. Une 3ème région, contenant une séquence répétée

inversée et retrouvée dans le promoteur du gène de la mucoviscidose, lie une protéine de 27 kDa (Hollingsworth *et al.*, 1994).

II.3.1.2 Régulation de la transcription de *MUC2*

Des premiers éléments de la régulation du gène *MUC2* ont été publiés dans trois articles récents.

Velcich *et al.* ont isolé des clones génomiques contigus qui couvrent le gène MUC2 en entier ainsi qu'une région d'environ 50 kb en amont de son site d'initiation (Velcich *et al.*, 1997). Des expériences de sensibilité à la DNase I ont montré que cette région possède probablement des éléments de régulation. Le fragment de 12 kb en amont du site d'initiation de la transcription de *MUC2* a alors été étudié par sous clonage en vecteur rapporteur. La région comprise entre les nucléotides +1 et -848 confèrent une activité transcriptionelle maximale au vecteur rapporteur alors que la région plus en amont contient des éléments « silencer ».

L'équipe de Kim a par ailleurs montré qu'un court segment compris entre les bases -91 et -73 est important pour l'activité basale de transcription d'un gène rapporteur transfecté dans plusieurs lignées cellulaires. Ce segment comprend une boîte CACCC qui lie des protéines Sp1 et des protéines non-Sp1 de la famille Sp (Gum *et al.*, 1997). Enfin, la région comprise entre les bases -228 et -171 semble conférer au vecteur rapporteur la spécificité cellulaire d'expression.

L'équipe de Basbaum a quant à elle étudié la régulation de *MUC2* par *Pseudomonas aeruginosa*. Le promoteur de *MUC2* possède un élément de réponse spécifique à *Pseudomonas aeruginosa* compris entre les nucléotides -2864 et -73 (Li *et al.*, 1997). A partir de souches mutantes de cette bactérie impliquée dans la surinfection dans la mucoviscidose, les auteurs ont montré que la surexpression du gène *MUC2* est contrôlée par la copule polysaccharidique (LPS) de l'agent pathogène,

et ce par l'intermédiaire de la phosphorylation d'un facteur intracellulaire encore inconnu.

II.3.1.3 Régulation de la transcription de *TBM*

La région en amont du site d'initiation de la transcription du gène *TBM* (mucine trachéobronchique de chien) a été clonée et séquencée. Cette région qui s'étend sur environ 1 kb ne contient ni de boîte TATA, ni de boîte CAAT. Les expériences de transfection, de fragments obtenus par des délétions et clonés dans des vecteurs rapporteurs, montrent l'importance de la régulation de ce gène par le facteur CREB (Verma *et al.*, 1996). Ce facteur est activé par phosphorylation par la protéine kinase A (PKA), en réponse à un stimulus extracellulaire, et ce, par l'intermédiaire d'une augmentation intracellulaire du taux d'AMPc.

II.3.2 Maturation de l'apomucine

La maturation des mucines sécrétées a fait l'objet de nombreuses études. Le précurseur peptidique, synthétisé dans le réticulum endoplasmique, a une masse de 470 à 700 kDa. Les monomères dimérisent dans ce compartiment et y sont *N*-glycosylés. Cette *N*-glycosylation semble indispensable pour l'oligomérisation des mucines. L'oligomérisation et la *O*-glycosylation pourrait avoir lieu dans les compartiments du Golgi. Selon les auteurs, la mucine, et l'espèce étudiée, chaque oligomère comprend de 3 à plus de 12 monomères (Klomp *et al.*, 1994a; Sheehan *et al.*, 1996; Fogg *et al.*, 1996; Tytgat *et al.*, 1995; Dekker *et al.*, 1990; Klomp *et al.*, 1994b; Asker *et al.*, 1995). Aucun de ces travaux n'apporte la preuve d'une homo- ou au contraire d'une hétéropolymérisation exclusive des mucines.

Les mucines sont ensuite stockées dans des granules de sécrétion. Elles y sont concentrées. Les granules matures contiennent une teneur élevée en ions Ca^{2+}. Ces granules possèdent probablement des canaux calciques actifs (Forstner, 1995).

Dans l'introduction de ce mémoire, nous avons tout d'abord vu les fonctions essentielles du mucus. Nous nous sommes ensuite intéressé aux mucines, qui sont les glycoprotéines majoritaires du mucus et lui confèrent ses propriétés rhéologiques et biologiques.

Après avoir donné quelques caractéristiques des mucines, nous avons dressé le catalogue des apomucines connues maintenant grâce aux résultats acquis par l'utilisation de techniques de l'ADN recombinant. Nous avons indiqué les caractéristiques structurales de leur axe peptidique. Cette revue reflète le nombre très important de séquences de mucines publiées au cours de ces trois dernières années. Ces éléments sont et seront de précieux outils pour étudier les pathologies qui touchent les muqueuses. Certaines mucines présentent de nombreuses similarités entre elles grâce à certains de leurs domaines peptidiques. Il est donc justifié de penser qu'à chaque domaine correspond une/des fonction(s) différente(s).

Dans ce mémoire, nous allons présenter les travaux qui ont permis de déterminer l'organisation génomique complète du gène de mucine humaine *MUC5B*, mucine intervenant dans la formation du gel de mucus, particulièrement au niveau de l'arbre respiratoire, de la vésicule biliaire, des canaux pancréatiques et de l'endocol. Ce premier modèle d'organisation structurale d'un gène de mucine du chromosome 11 nous a permis ensuite (chapitre Discussion et Perspectives) de proposer de nouvelles bases moléculaires pour une classification des mucines.

Stratégie

STRATEGIE

I DONNEES PRELIMINAIRES

I.1 Travaux antérieurs à 1993

I.1.1 Clonage d'ADNc de *MUC5B*

Un immunsérum, préparé à partir de mucines trachéobronchiques humaines déglycosylées, a permis de cribler une banque d'expression en vecteur λgt11 (Crépin *et al.*, 1990). Parmi les clones obtenus, JER28 (0,56 kb) et JER57 (1,83 kb), présentent le même motif répétitif dégénéré de 87 pb soit 29 acides aminés. Ces 2 ADNc ont été utilisés comme sonde pour isoler les clones JUL7 et JUL10 d'une banque d'ADNc construite en vecteur λgt10 (Dufossé *et al.*, 1993). Ces 2 derniers clones, de 1631 et 991 pb respectivement, contiennent aussi le même motif de 87 pb.

Le clone TH71 (380 pb) a été isolé de la même banque d'expression en vecteur λgt11. Des anticorps, purifiés à partir de l'immunsérum par adsorption sur la protéine recombinante JER57, reconnaissent le peptide de fusion TH71 (Crépin *et al.*, 1990). Ceci indique que les 2 clones TH71 et JER57 sont probablement issus du même gène.

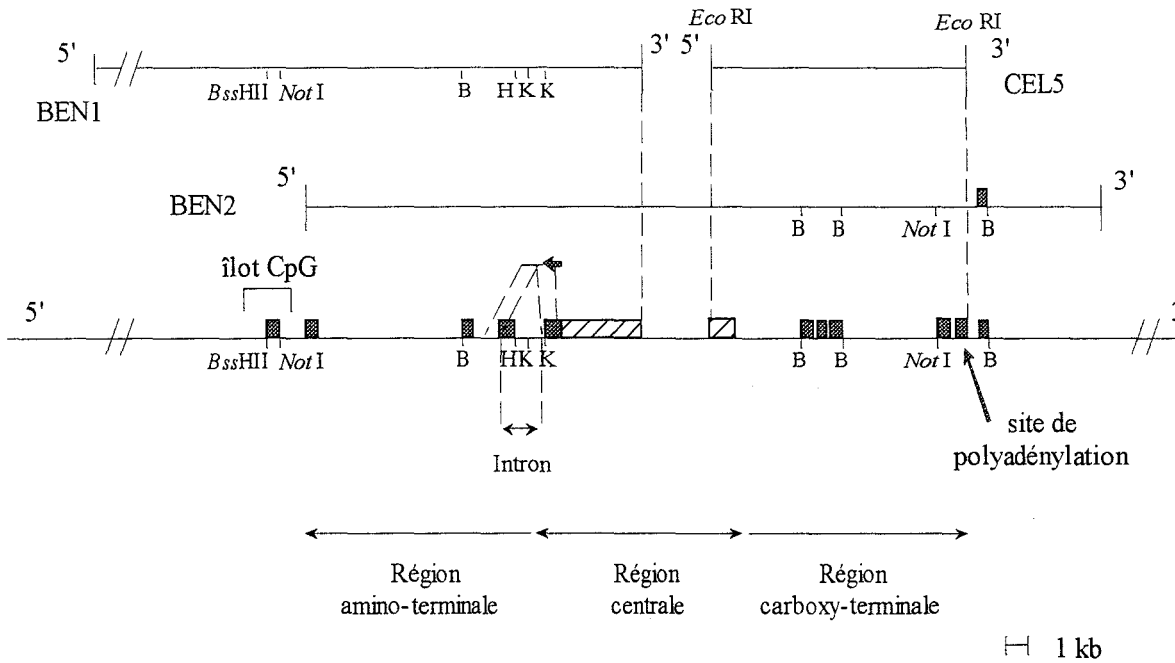
I.1.2 Clonage génomique de *MUC5B*

Le criblage par la sonde JER57 d'une banque génomique humaine en vecteur λEMBL4 a permis d'isoler le clone phagique CEL5 de 12,6 kb. Ce clone contient dans sa partie 5' sur 1445 pb des motifs répétés en tandem de 87 pb caractéristiques de *MUC5B*. Un signal de polyadénylation a été trouvé dans la partie 3' de ce clone indiquant que CEL5 peut correspondre à la partie 3' du gène *MUC5B*. Les autres clones qui ont pu être isolés de cette banque en vecteur λEMBL4 présentaient tous la même carte de restriction. Le criblage avec la sonde JER57 d'une banque génomique

commerciale en vecteur cosmidique pWE15 a alors été entrepris et a livré les deux clones chevauchants BEN1 et BEN2. Les cartes de restriction de ces clones ont été établies et ont permis d'aligner les 3 clones, cosmidiques et phagique, les uns par rapport aux autres (Figure 22). BEN1 contient un îlot CpG et c'est en aval de celui-ci (à environ 12 kb vers le 3') que la sonde JER57 s'hybride. BEN2 possède un signal de polyadénylation en 3' qui correspond à celui trouvé dans le clone CEL5. BEN2 contient tous les fragments obtenus après coupure par des enzymes de restriction et visualisés sur Southern blot de l'ADN humain après hybridation avec la sonde JER57. BEN2 recouvre CEL5. D'après ces éléments, on peut penser que le gène *MUC5B* est entièrement contenu dans ces clones chevauchants et il s'étendrait sur environ une trentaine de kb. Les séquences nucléotidiques des fragments s'hybridant avec la sonde JER57, qui ont été déterminées sur BEN1, BEN2 et CEL5 et qui sont constituées de séquences répétées en tandem de 87 pb, permettent de déduire qu'il doit exister une séquence codante répétitive sur plus de 9 kb (Figure 22).

I.2 Travaux du DEA

Pour progresser dans la recherche des exons dans la région en 5' de cette séquence répétitive, nous n'avons pas choisi la procédure classique qui aurait consisté à cribler une banque d'ADNc à l'aide d'une sonde nucléique. En effet cette méthode aurait nécessité plusieurs mois avec probablement peu de succès puisque l'obtention d'ADNc en 5' est généralement très difficile surtout pour les grands transcrits. Nous avons donc opté pour la méthode RACE-PCR (pour Rapid Amplification cDNA End-PCR) qui a été développée pour amplifier spécifiquement la partie 5' d'un ARNm (Frohman *et al.*, 1988). Cette technique m'a permis, lors de mon DEA, de localiser sur la carte génomique, par la mise en évidence d'un intron, l'extrémité 5' de la région centrale de *MUC5B*. Le principe de la méthode qui a été utilisée est schématisé sur la figure 23 : la partie 5' d'un ARNm est copié à l'aide d'un oligonucléotide P1 s'hybridant avec la séquence connue. Puis après avoir ligué un oligonucléotide «anchor» simple brin à l'extrémité 5' de l'ADNc, il s'agit d'amplifier cet ADNc par PCR



parties séquencées :

séquence non répétitive

séquence répétitive (s'hybridant avec la sonde JER57)

B : *Bgl* II

H : *Hin* dIII

K : *Kpn* I

Clone de RACE-PCR (985 pb) :

Figure 22 : Résultats acquis en début de thèse

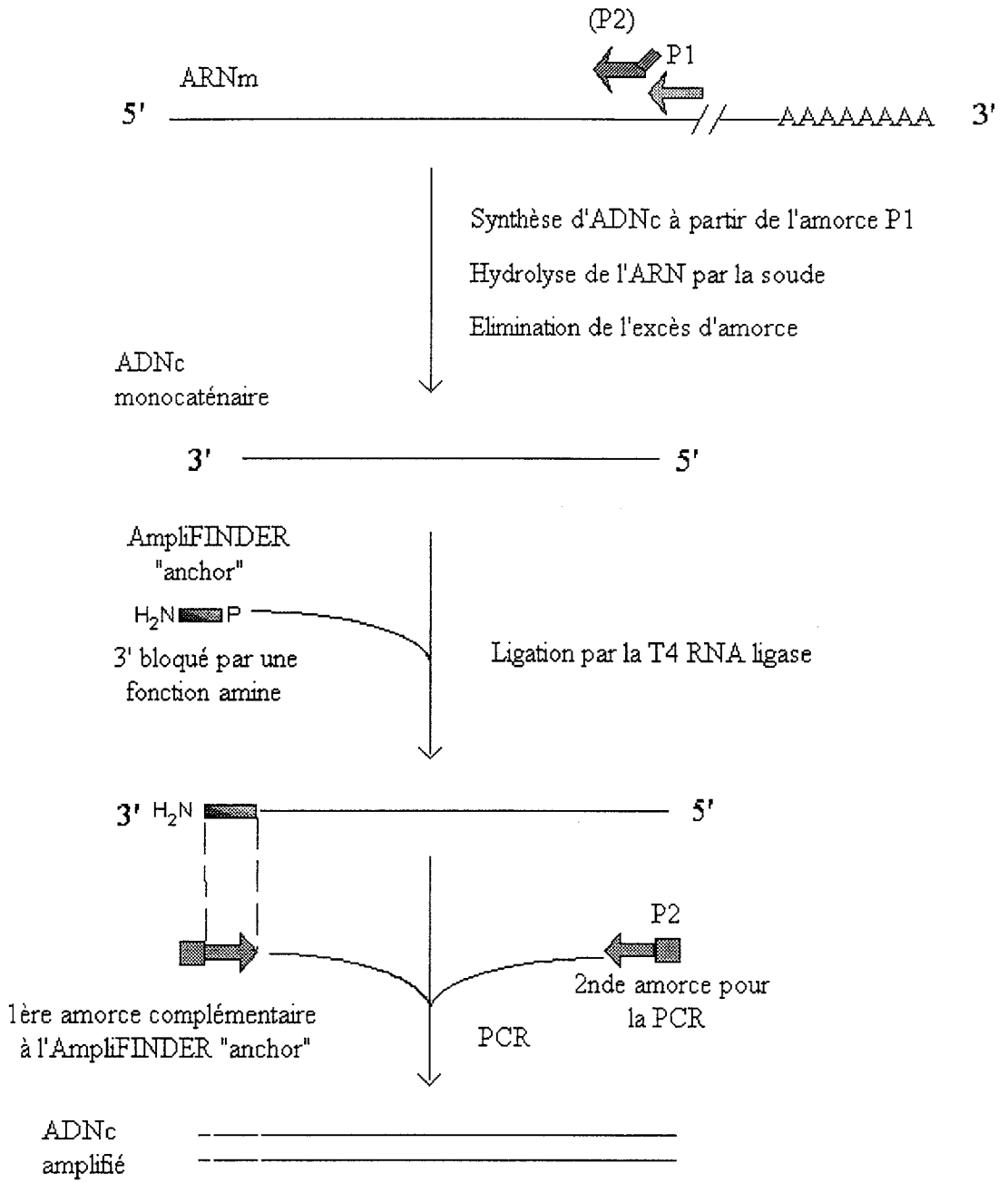


Figure 23 : RACE-PCR selon le protocole préconisé par Clontech

en utilisant une amorce complémentaire à «l'anchor» et une seconde amorce P2 choisie en amont de P1.

Nous avons ainsi pu déterminer la séquence de l'extrémité 5' d'un motif riche en résidus de cystéine qui se trouve juste en amont de la partie répétitive. La région génomique correspondante à cet ADNc a été sous-clonée à partir du cosmide BEN1 et séquencée parallèlement au laboratoire. L'alignement de la séquence du plus long ADNc (985 pb) obtenu par RACE-PCR et de la séquence génomique nous a permis de mettre en évidence une région intronique de 500 pb.

I.3 Expression de *MUC5B*

Les études en dot blots, Northern blots et en hybridation *in situ* menées au laboratoire montrent que les sites préférentiels d'expression de *MUC5B* sont la muqueuse respiratoire, les glandes sous-maxillaires, le pancréas et la vésicule biliaire. Nous avons donc choisi d'utiliser pour nos études du/des messager(s) de *MUC5B* de l'ARN poly(A)$^+$ commercial (Clontech) de trachée humaine, de glande salivaire et de l'ARN total de vésicule biliaire préparé au laboratoire.

II OBJECTIFS DE CES TRAVAUX

Le but de ce travail est :

- de déterminer l'organisation génomique complète du gène de mucine humaine *MUC5B*
- de déterminer s'il existe un ou plusieurs transcrits
- de mettre en évidence un possible épissage alternatif

Notre travail mettra en évidence de fortes similitudes avec d'autres transcrits de mucines ; si ceci est un outil intéressant pour cloner de nouveaux gènes de mucines,

c'est aussi un obstacle, car il nous faudra pour chaque ADNc cloné, vérifier qu'il appartient, ou non, au gène *MUC5B*.

Ce travail doit apporter des réponses ou des éléments de réponse aux questions suivantes :

- les séquences répétées font-elles partie, comme il a été montré pour MUC1 et MUC7, d'un seul et unique exon ?

- nous avons supposé que le gène s'étend sur une trentaine de kb, entre un îlot CpG et un signal de polyadénylation. Les îlots CpG indiquent la présence de gènes car ils leur sont presque toujours associés. Ces îlots recouvrent généralement les premiers exons des gènes qu'ils balisent. Quelle est la taille exacte du gène *MUC5B* ? L'îlot CpG recouvre-t-il les premiers exons du gène ?

- peut-on comprendre par l'étude en parallèle du gène et du/des transcrit(s) de *MUC5B* l'extrême hétérogénéité des transcrits des gènes de mucines *MUC2*, *MUC3*, *MUC4*, *MUC5AC*, *MUC5B*, *MUC6* et *MUC8* observée en Northern blot ?

La séquence peptidique déduite, validera peut-être le modèle linéaire ou le modèle du « moulin à vent » des mucines. L'apomucine MUC5B est-elle capable de dimériser et de polymériser comme cela a été proposé pour MUC2 ou reste-t-elle sous forme monomérique comme MUC7 ? Enfin, MUC5B est-elle une apomucine sécrétée ? transmembranaire ? à la fois sécrétée et transmembranaire ?

Le peptide déduit nous permettra d'appréhender les fonctions biologiques de MUC5B et, plus généralement, des mucines. De plus, la bonne connaissance de la topographie du gène *MUC5B* est une première étape pour l'étude des éléments régulateurs du gène, qu'ils soient en amont, en aval ou à l'intérieur même du gène.

III STRATEGIE D'ETUDE

L'état de nos connaissances sur la topographie du gène *MUC5B* au début de ce travail de thèse est schématisé sur la figure 22.

III.1 La région centrale

III.1.1 Détermination de l'extrémité 3' de la région centrale

Une amorce sens a été synthétisée d'après la séquence nucléotidique trouvée dans JER57 et dans le cosmide BEN2, et une autre amorce, NAU82, antisens et en aval de la précédente, d'après des éléments de séquence de BEN2 situés en aval de la partie répétitive (Figure 24). Ces 2 amorces ont permis de cloner un produit de RT-PCR et un produit de PCR effectuée parallèlement sur BEN2. La comparaison des séquences des clones obtenus par RT-PCR et PCR nous a permis de mettre en évidence un intron en aval de la région répétitive.

III.1.2 Recherche d'autres introns entre les 2 introns précédemment définis

Toute la région qui s'étend entre les 2 introns (intron mis en évidence lors du DEA et intron en aval de la partie répétitive) a entièrement été sous-clonée et séquencée au laboratoire. Nous avons alors choisi des amorces pour amplifier par RT-PCR et parallèlement par PCR sur le cosmide BEN2 les différentes régions de la partie centrale de *MUC5B* (voir Figure 24) afin de trouver éventuellement d'autres introns.

L'analyse des séquences nucléotidiques révèle trois types de domaines répétés dans la région centrale. Un de ces domaines est constitué de la répétition en tandem du motif élémentaire de 87 pb. Ces répétitions à plus ou moins grande échelle se retrouvent aussi pour les sites de restriction. Mais une certaine asymétrie de quelques sites de restriction nous a aidé pour savoir de quelle zone précise de la partie répétitive

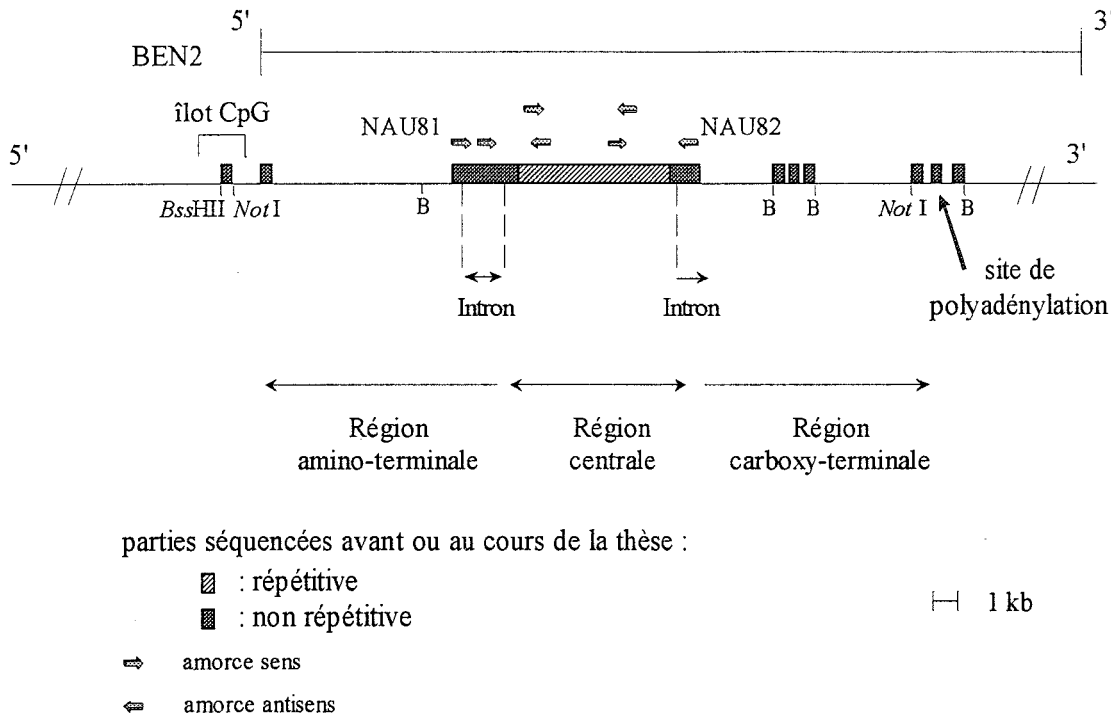


Figure 24 : Stratégie d'étude de la région centrale de *MUC5B*

étaient issus les fragments. De même, il a été intéressant de pouvoir travailler avec trois clones génomiques ne se superposant que sur des longueurs limitées. Ainsi, pour préparer certains fragments ayant des tailles très proches mais situés dans des régions différentes, nous avons pu partir de l'insert cosmidique BEN1 qui ne contient que les 4,3 kb du côté 5' de la partie centrale. Si nous n'avions travaillé que sur BEN2, il aurait été très difficile de placer dans leur agencement correct les fragments qui se ressemblent très fortement.

Nous avons également effectué des expériences de longue PCR sur du cosmide BEN2 en utilisant des amorces (NAU81 et NAU82 ; Figure 24) qui encadrent la région répétitive pour confirmer la taille de cette région.

III.1.3 Analyse phylogénétique

Le domaine répété riche en résidus Cys (voir Résultats, §I.3) est présent dans d'autres mucines humaines et animales. Nous avons alors utilisé le programme PC/Gene pour aligner les séquences nucléotidiques puis le logiciel Excel pour calculer les homologies entre ces séquences. Ce travail nous a permis d'élaborer un modèle cohérent de l'évolution des gènes humains de mucines localisés en 11p15.5.

III.2 La région carboxy-terminale

III.2.1 Le signal de polyadénylation

Un ADNc de 380 pb, TH71, contenant une queue poly(A)$^+$ et un signal de polyadénylation avait été cloné auparavant. Nous avons tout d'abord déterminé sa séquence nucléotidique complète. La séquence nucléotidique cosmidique correspondante a été déterminée. Cette séquence correspond à l'extrémité 3' de l'insert phagique CEL5. En fait, elle contient une séquence identique à celle trouvée dans TH71 sur 67 pb du côté 3' et le même signal de polyadénylation. La région 5' de cet ADNc comporte des motifs répétés de 87 pb identiques à ceux retrouvés dans la

région centrale de *MUC5B*. L'alignement de la séquence du clone TH71 et des séquences partielles génomiques met en évidence une région intronique putative de plusieurs kilobases sans signature de site d'épissage (Mount, 1982). Nous avons alors tenté une première expérience de RACE-PCR afin de remonter sur l'ADNc vers la région centrale. Les ADNc néosynthétisés sont petits (moins de 250 pb), rendant la progression très lente.

III.2.2 La région Carboxy-terminale : région en aval de la région répétitive

Une séquence partielle du cosmide BEN2 contenant un cadre ouvert de lecture sur plus de 200 pb a été obtenue. Cette séquence se situe dans le fragment *Bgl*II-*Bgl*II (Figure 25) qui, utilisé comme sonde, révèle un continuum sur Northern blot de trachée humaine, suggérant ainsi que le fragment d'ADN est tout ou partie exonique.

A partir de cette séquence, nous avons fait synthétiser une amorce sens et une amorce antisens juste en aval de l'amorce sens.

Nous avons effectué deux RT-PCR chevauchantes en utilisant une troisième amorce sens (NAU71) choisie dans la région exonique juste en aval de l'exon central (où se situe NAU82) et un quatrième oligonucléotide, antisens et choisi dans l'extrémité 3' du TH71. L'ADNc complet de la partie C-terminale a ainsi pu être synthétisé et cloné.

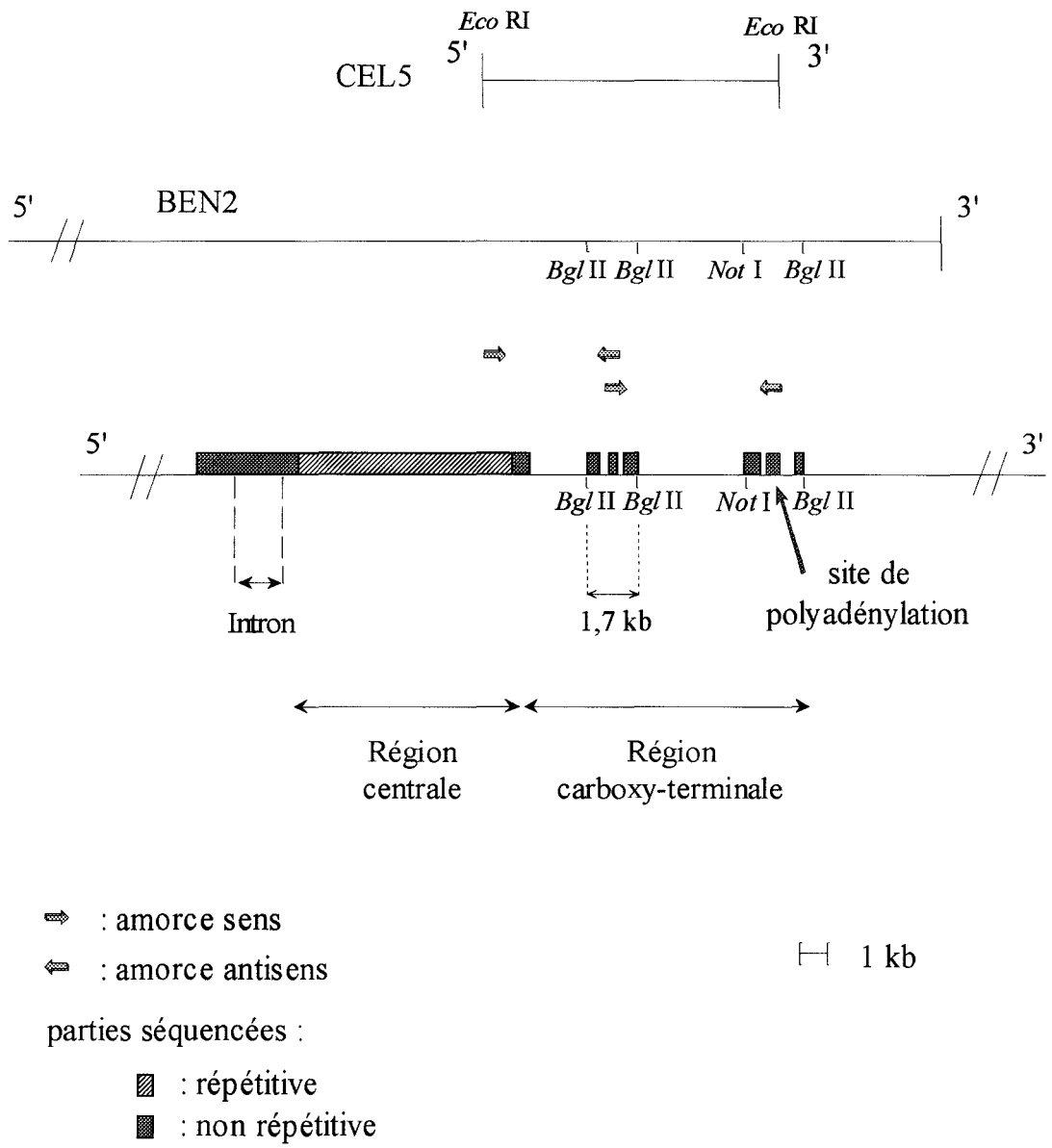


Figure 25 : Stratégie d'étude de la région carboxy-terminale de *MUC5B*

III.3 La région amino-terminale

III.3.1 RACE-PCR selon le protocole préconisé par Clontech

Une première expérience de RACE-PCR nous avait permis de délimiter lors de mon DEA l'extrémité 5' de la région centrale. Nous avons effectué ensuite 2 réactions de RACE-PCR successives afin de remonter vers l'extrémité 5' de l'ARNm (Figure 26). Les 2 ADNc chevauchants clonés (600 et 200 pb) ont été séquencés. Les fragments cosmidiques correspondants ont été sous-clonés et séquencés afin de vérifier que les ADNc amplifiés proviennent bien du gène *MUC5B* et non d'un gène homologue.

III.3.2 RT-PCR utilisant des amorces dégénérées

La séquence peptidique déduite des ADNc précédents est très homologue à la séquence peptidique du domaine D3 du pro-vWF (région N-terminale), domaine déjà retrouvé dans la partie amino-terminale de MUC2, de l'homologue de MUC2 chez le rat et de HGM-1 (ADNc de *MUC5AC*). Afin d'avancer plus rapidement vers l'extrémité N-terminale de MUC5B, nous avons adopté une nouvelle stratégie en supposant que MUC5B devait avoir une construction très similaire à celles de MUC2, de l'homologue de MUC2 chez le rat, de MUC5AC et donc aussi du pro-vWF. Nous avons donc supposé que l'apomucine issue du gène *MUC5B* possédait en amont de sa région répétitive de type mucine, 3 domaines D (D1, D2 et D3) similaires à ceux retrouvés dans le pro-vWF. Une amorce sens, appelée NAU101, ayant une séquence dégénérée déduite de l'alignement des séquences nucléotidiques publiées d'HGM-1 et des régions les plus conservées des parties amino-terminales du pro-vWF, de Rat-Muc2 et de MUC2, a été synthétisée (Tableau IX). Cette amorce est un 19-mers représentant 128 (2^7) combinaisons possibles. Par ailleurs, nous avons choisi, d'après les séquences des ADNc obtenus par RACE-PCR, une deuxième amorce antisens spécifique de *MUC5B* (NAU76).

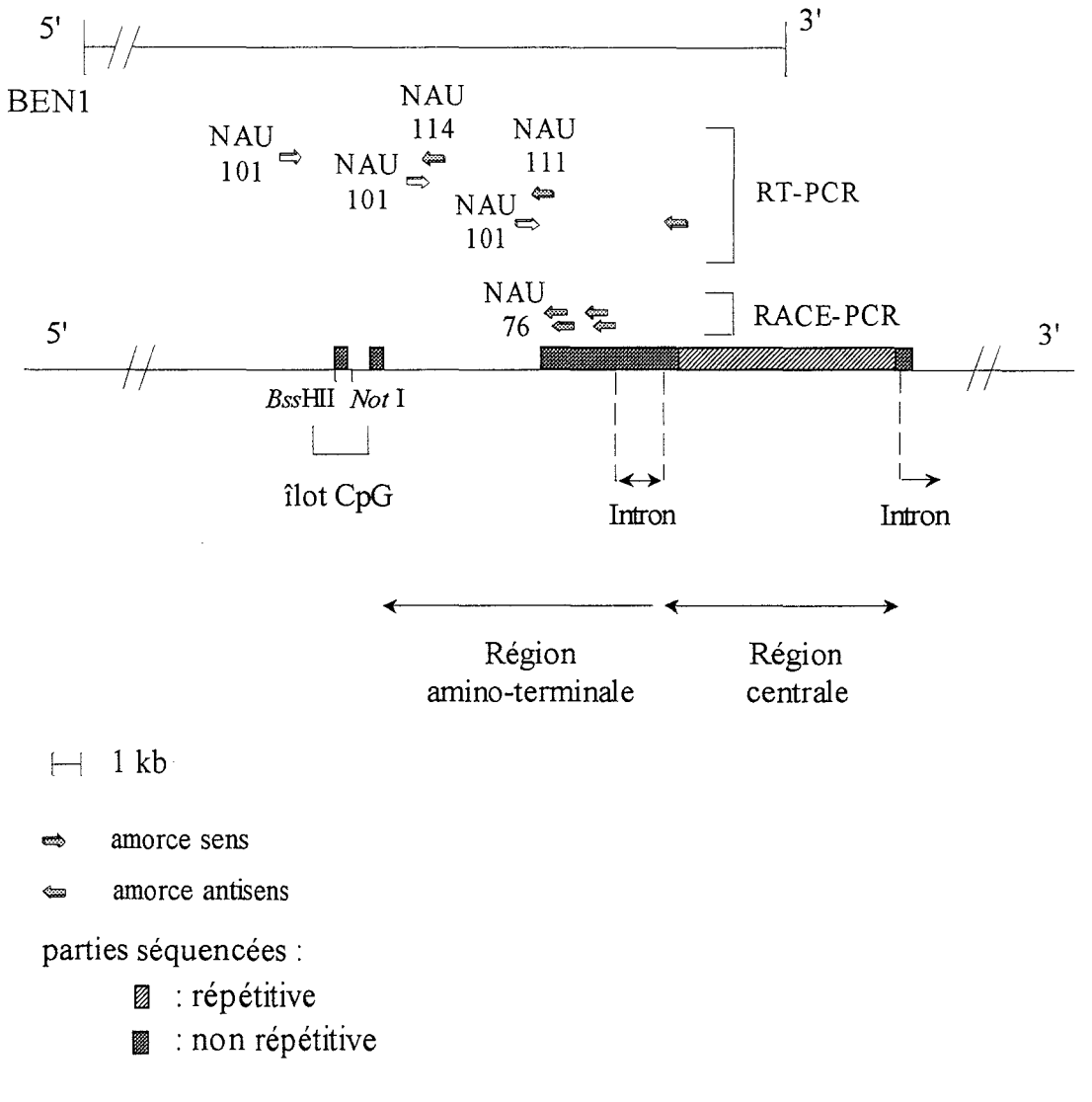


Figure 26 : Stratégie d'étude de la région amino-terminale de *MUC5B*

Tableau IX : Choix de l'oligonucléotide dégénéré NAU101

pro-vWF D1	N	K	T	C	**G**	**L**	**C**	**G**	N	F	N	I
pro-vWF D2	G	K	T	C	**G**	**L**	**C**	**G**	N	Y	N	G
pro-vWF D3	E	K	V	C	**G**	**L**	**C**	**G**	N	F	D	G
MUC2 D1	N	H	T	C	**G**	**L**	**C**	**G**	D	Y	N	G
MUC2 D2	G	Q	V	Q	**G**	**L**	**C**	**G**	N	F	N	G
MUC2 D3	G	T	V	C	**G**	**L**	**C**	**G**	N	F	D	H
rat-*Muc2* D1	N	H	T	C	**G**	**L**	**C**	**G**	D	F	N	G
rat-*Muc2* D2	G	Q	V	Q	**G**	**L**	**C**	**G**	D	F	N	G
rat-*Muc2* D3	G	K	V	C	**G**	**L**	**C**	**G**	N	N	N	G
HGM-1	G	R	V	C	**G**	**L**	**C**	**G**	N	F	D	D
Séquence peptidique consensus			T V	C Q	**G**	**L**	**C**	**G**	N D	F Y N	N D	
Séquence nucléotidique consensus déduite			Y R I	G G I	Y T I	T G Y	G G I		R A Y	W		

Nous avons vérifié que l'amorce NAU101 marquée au $\gamma^{32}P$ s'hybride bien sur des Southern blots de BEN1 à trois régions de la partie amino-terminale de *MUC5B*. Un premier ADNc a été amplifié par RT-PCR (NAU101/NAU76), cloné et séquencé. Une nouvelle amorce antisens, NAU111, spécifique du gène *MUC5B* a été synthétisée d'après la séquence nucléotidique amino-terminale de cet ADNc et ceci, après avoir vérifié par séquençage direct sur le cosmide BEN1, que cet ADNc était bien issu de *MUC5B*. Couplée à l'amorce sens dégénérée NAU101, elle a permis d'amplifier un nouvel ADNc. Nous avons vérifié par la même technique que précédemment son appartenance au gène *MUC5B*. Une nouvelle amorce antisens, NAU114, a été synthétisée d'après la séquence amino-terminale de l'ADNc néosynthétisé. Une dernière expérience de RT-PCR a été effectuée avec cette amorce et l'amorce sens dégénérée (Figure 26). De même, nous avons vérifié que cet ADNc était bien un produit du gène MUC5B par séquençage direct sur le cosmide BEN1.

Nous avons effectué parallèlement 2 longues PCR sur BEN1 afin de mesurer les distances génomiques entre NAU101 et les 2 amorces NAU111 et NAU114.

III.3.3. RACE-PCR selon le protocole préconisé par Boehringer

Puisque les séquences de ces ADNc de *MUC5B* sont très homologues à la séquence amino-terminale de MUC2 et du pro-vWF, il y a tout lieu de penser qu'il reste à peu près 600 pb d'ADN codant à cloner en amont du dernier ADNc obtenu. Des amorces antisens ont été choisies d'après la séquence nucléotidique amino-terminale de cet ADNc afin d'effectuer une réaction de RACE-PCR.

Le système commercialisé par la société Clontech nous semblant peu efficace, nous avons effectué une expérience de RACE-PCR en utilisant le système « 5'/3' RACE » de Boehringer. Le principe reste identique à celui décrit précédemment mis à part qu'on ne ligue pas un oligonucléotide monobrin aminé à l'extrémité 3' de l'ADNc mais on y synthétise une queue poly(A) (Figure 27). Une première amplification utilise

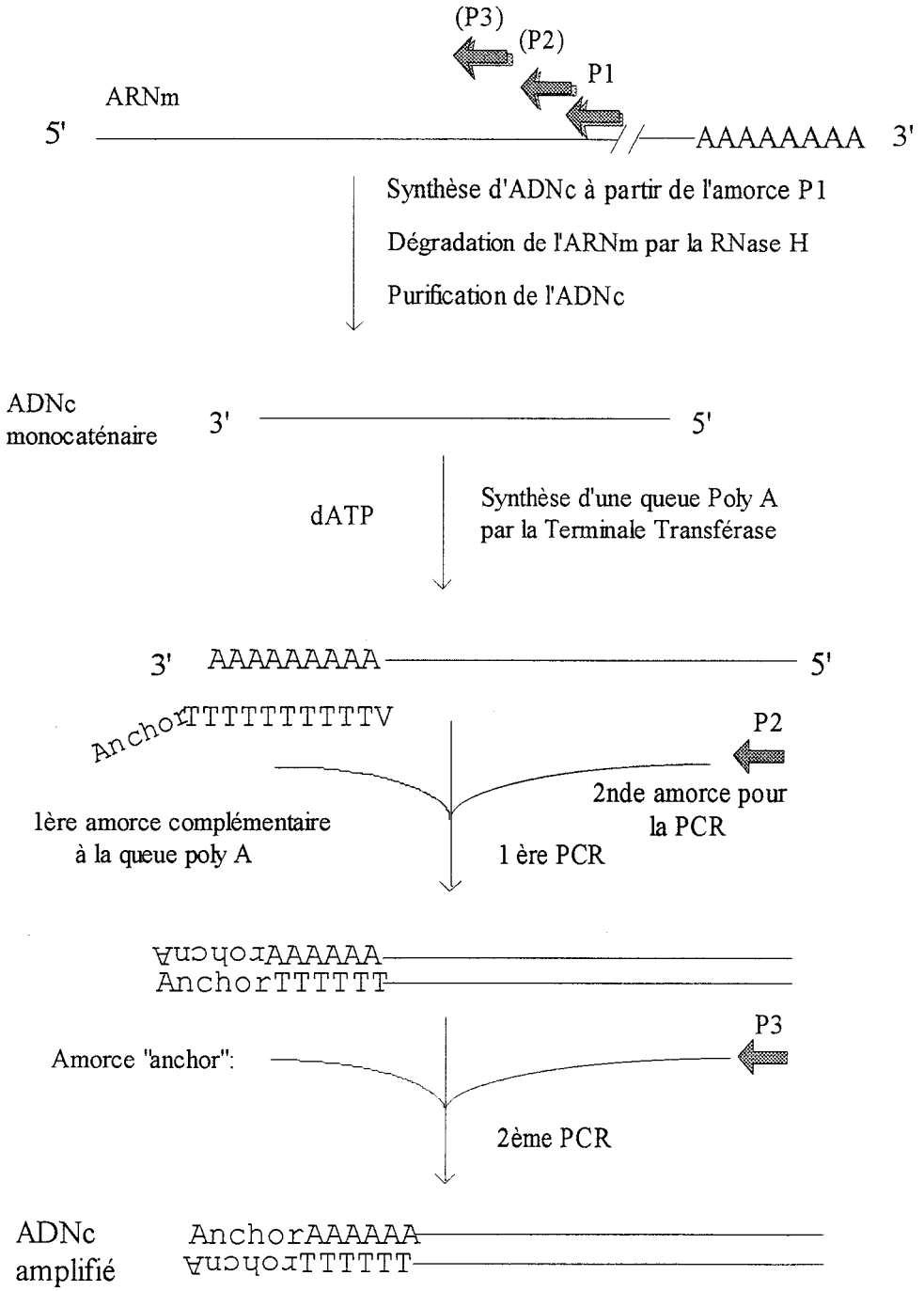


Figure 27 : RACE-PCR selon le protocole préconisé par Boehringer

une amorce poly(T)-anchor. Puis, une seconde amplification sur le produit de la 1ère PCR utilise l'amorce «anchor».

III.3.4 Extension d'amorce

Afin de confirmer que le dernier ADNc cloné par RACE-PCR contient l'extrémité 5' du/des transcrits de *MUC5B*, nous avons effectué des expériences d'extension d'amorce à partir de 2 oligonucléotides monobrins antisens distants entre eux de quelques dizaines de paires de bases et situés à moins de 300 nucléotides de l'extrémité 5' supposée.

Résultats

RESULTATS

I LA REGION CENTRALE

I.1 Les domaines répétés en tandem de MUC5B appartiennent à un unique exon central.

Le gène *MUC5B* est cloné, tout ou partie, dans les 2 cosmides chevauchants BEN1 et BEN2. La partie « tandem repeat » de MUC5B est codée par un unique exon de 10713 pb. La séquence nucléotidique est disponible dans la banque de données GenBank™ sous le numéro d'accession Z72496.

Les résultats sont présentés dans la publication suivante :

« **Human mucin gene *MUC5B*, the 10.7-kb large central exon encodes various alternate subdomains resulting in a super-repeat.** STRUCTURAL EVIDENCE FOR A 11P15.5 GENE FAMILY »

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 272, No. 6, Issue of February 7, pp. 3168–3178, 1997
Printed in U.S.A.

Human Mucin Gene *MUC5B*, the 10.7-kb Large Central Exon Encodes Various Alternate Subdomains Resulting in a Super-repeat

STRUCTURAL EVIDENCE FOR A 11p15.5 GENE FAMILY*

(Received for publication, May 30, 1996, and in revised form, September 30, 1996)

Jean-Luc Desseyn‡§, Véronique Guyonnet-Dupérat‡§, Nicole Porchet‡¶, Jean-Pierre Aubert‡¶, and Anne Laine‡‖

From the ‡INSERM 377 Laboratoire Gérard Biserte, place de Verdun, 59045 Lille Cedex, France and the ¶Laboratoire de Biochimie et de Biologie Moléculaire de l'Hôpital C. Huriez, CHRU de Lille, 59037 Lille Cedex, France

Human mucin gene *MUC5B* is mapped clustered with *MUC6*, *MUC2*, and *MUC5AC* on chromosome 11p15.5. We report here the isolation of three overlapping genomic clones of human *MUC5B* spanning approximately 40 kilobases. We have determined their partial restriction maps and the intron-exon boundaries of the central region encoding a single open reading frame. This coding region has been completely sequenced. Its length is 10,713 base pairs, and it encodes a 3570-amino acid peptide. Nineteen subdomains have been individualized. Some subdomains show similarity to each other, creating larger composite repeat units that we have called super-repeats. Four super-repeats of 528 amino acid residues are thus observed within the central exon. Each comprises (i) a subdomain composed of 11 repeats of the irregular repeat of 29 amino acid residues, (ii) a unique conserved subdomain with no typical repeat, and (iii) a cysteine-rich subdomain. This latter subdomain has high sequence similarity to the cysteine-rich domains described in MUC2 and MUC5AC. Sequence data of these three genes, together with their clustered organization, lead us to suggest that they may be a part of a multigene family. The super-repeat present in *MUC5B* is the largest ever determined in mucin genes and the central exon of this gene is, by far, the largest reported for a vertebrate gene.

Mammalian respiratory, gastrointestinal, and reproductive tracts are protected by mucus secretions, of which the major components are the mucins. The mucins form a heterogeneous group of high molecular mass, polydisperse, highly glycosylated macromolecules. They are synthesized and secreted by specialized cells in the epithelium.

Considerable advances have been made over the past years toward our understanding of the structure and function of mucin glycoproteins. The isolation of mucin cDNA clones introduced a new approach to the structure of the mucins. Until now, at least eight human mucin genes have been identified (see Ref. 1 for review). The chromosomal localization of these genes has been established: four of them, *MUC2*, *MUC5AC*, *MUC5B*, and *MUC6*, are clustered on 11p15.5 between the *HRAS* and *IGF2* genes, *MUC1* is on 1q21–24, *MUC3* on 7q22, *MUC4* on 3q29, and *MUC7* on chromosome 4q13–21. Recently, a cDNA called pAM1 has been cloned from a human tracheal library and localized on chromosome 12 (2). Three novel cDNAs have been reported: NP3a from a human nasal polyp library (3), L31 from a HT29-MTX cell line library (4), and HGM-1 from a human stomach library (5). Their sequences show that they correspond to some parts of the *MUC5AC* gene. More recently, a novel cDNA (pSM2–1) from human sublingual gland has been described (6). *MUC1*, which is developmentally regulated and aberrantly expressed by carcinomas, encodes a membrane-associated mucin-like glycoprotein. In contrast, the other described genes code for secreted mucins (1). Mucins present extended arrays of tandemly repeated sequences, producing a protein core rich in potential *O*-glycosylation sites and having a high content of serine, threonine, proline, glycine, and alanine. The tandem repeat units vary in length from as few as 24 bp[1] in *MUC5AC* (7) to 507 bp in *MUC6* (8). The tandem repeat domain is flanked on either side by nonrepeat regions. In the *MUC2* gene product, these tandem repeats are flanked by cysteine-rich subdomains of approximately 845 residues upstream and 700 residues downstream. Both cysteine-rich subdomains have sequences similar to the D-domains of human pro-von Willebrand factor (9, 10). Some parts of these D-domains are also found in NP3a (3) and in HGM-1 (5). Moreover, the *MUC2* gene has upstream to the tandem repeat a region of imperfectly conserved repeats flanked by another type of cysteine-rich domain (11). This latter domain has also been described in *HGM-1* (5), twice in MUC5AC (12), and also in its related cDNAs (3, 4).

Mucins are essential for the protective properties of the mucus (13, 14). In addition it is becoming apparent that an abnormal expression of the mucin genes occurs in various disease states and in conditions associated with a high risk of adenoma or carcinoma (15–17). Therefore, the study of the factors responsible for the regulation of mucin expression is of great interest. With this goal, it is necessary to acquire a detailed knowledge of the genomic structure of the mucin genes. The complete sequences of MUC1 (18), MUC2 (11, 9, 19), and MUC7 (20) cDNAs have been described. However, only partial cDNA sequences are published for the other mucin genes. The whole genomic structure is only known for *MUC1* (21), although the partial genomic organization of *MUC7* has

* This work was supported by "La Ligue contre le Cancer," "l'Association de Recherche contre le Cancer," and "l'Association Française de Lutte contre la Mucoviscidose." The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) Z72496 (HSMUC5BEX).

§ The first two authors contributed equally to this work and should therefore be considered as equal first authors.

‖ To whom correspondence should be addressed. Tel.: 33-3-20-29-88-59; Fax: 33-3-20-53-85-62; E-mail: laine@biserte.lille.inserm.fr.

[1] The abbreviations used are: bp, base pair; kb, kilobase(s); nt, nucleotide; aa, amino acid; PCR, polymerase chain reaction; RACE, rapid amplification of cDNA ends; RT-PCR, reverse transcription PCR.

TABLE I
Sizes (kb) of the DNA fragments obtained with various restriction enzymes and hybridizing with the JER57 probe

Bam HI	*Bgl* II	*Eco* RI	*Hind* III	*Kpn* I	*Pst* I	*Xba* I	*Xho* I
1.6	18	>30	25	8	1.4	>30	25
1.1				2.2	1.25		3.7
0.4				2.1	0.95		1.6
				1.6	0.65		

also been reported recently (22).

We have previously described four human tracheobronchial cDNA clones with degenerate 87-bp tandem repeats belonging to the human *MUC5B* gene (23). Our laboratory has focused considerable attention upon this gene which is expressed in mucous glands of tracheobronchial tissue, submaxillary glands, gall bladder, and endocervix (24–26).

In this paper, we present the partial restriction map of the three overlapping genomic clones of the *MUC5B* gene spanning more than 40 kb and the complete sequence of its central exon which encompasses 10,713 bp and codes for a 3570-amino acid peptide. Nineteen subdomains are individualized. Some subdomains show similarity to each other, creating larger composite repeat units that we call super-repeats. In addition to the tandem repeat of 29 amino acid residues as described previously (23), which we now define as being irregular or imperfectly conserved, four repeats of 528 amino acid residues are observed within the central exon. Each comprises 11 repeats of the irregular repeat of 29 amino acid residues, a unique conserved domain of 111 amino acid residues with no typical repeat but rich in threonine, serine, and alanine and a cysteine-rich region of 108 amino acid residues. This latter subdomain has sequence similarity with cysteine-rich domains of MUC2 and MUC5AC or its related cDNAs. This super-repeat is the largest ever determined in mucin genes.

EXPERIMENTAL PROCEDURES

Southern Blot Analysis—DNA from 20 healthy unrelated volunteers was prepared from leukocytes. It was digested with the following restriction endonucleases: *Bam*HI, *Bgl*II, *Eco*RI, *Hin*dIII, *Kpn*I, *Pst*I, *Xba*I, and *Xho*I. Fragments were separated by electrophoresis in phosphate buffer through 1% agarose gel, transferred to a nylon membrane, and hybridized as described previously (23).

Library Screening—JER57, the longest cDNA of *MUC5B* isolated previously (23), was first used as a probe to screen a genomic EMBL4 phage library (12). One positive clone CEL5 was isolated and studied. Other screenings only gave the same clone.

To isolate larger genomic clones of the *MUC5B* gene, we screened a human placenta genomic DNA library in pWE15 cosmids provided by Stratagene using the JER57 probe. Two positive clones BEN1 and BEN2 were isolated and studied.

Restriction Mapping of Cosmids—The restriction mapping strategy of Wahl *et al.* (27) was slightly modified as follows. Cosmids were digested to completion with the restriction enzyme *Not*I. For each of the other restriction enzymes used, one part of this *Not*I-digested cosmid was digested to completion and a second part partially digested with the enzyme in order to generate a set of fragments that began at the T7 or T3 promoters and ended at the site of cleavage of the chosen enzyme. These *Not*I-terminated digestion products were fractionated on an agarose gel (0.6%) and blotted to Hybond™-N$^+$ membrane (Amersham Corp.) by capillary blotting overnight. The fragments were then mapped relative to the T7 or T3 promoters by hybridizing the blot with end-labeled oligonucleotide-sequencing primers specific for these promoters.

5′ RACE—The 5′ AmpliFINDER RACE kit (Clontech, Inc., Palo Alto, CA) was used to synthesize first strand cDNA from 2 μg of human tracheal poly(A)$^+$ RNA obtained from Clontech with NAU58 as first primer (5′-TTGTAGCACATCTTGAAGACGCCC-3′, antisense nt 776–799) followed by the ligation of the 5′ anchor adapter. The RACE-PCR was performed in 50-μl reaction volumes containing 5 μl of 10 × buffer, 5 μl of 10 mM deoxynucleoside triphosphates, 2.5 μl of reversed transcribed target cDNA, 10 pmol of each primer (NAU57, 5′-ACGGATCCCCTGCACACCAGGCCGAAGTG-3′, antisense nt 741–762 with underlined nucleotides added in 5′ to generate a *Bam*HI restriction site and 5′ anchor primer), and 1.5 units of *Taq* DNA polymerase (Boehringer Mannheim). After overlaying with 50 μl of mineral oil (Sigma), the mixture was denatured at 94 °C for 3 min followed by 30 cycles at 94 °C for 1 min, 71 °C for 1 min, 72 °C for 2 min. The elongation step was extended for an additional 10-min period. Secondary amplification was performed using 2.5 μl of the primary amplification product. The thermal cycling protocol used was the same as for the primary RACE amplification.

RNA Extraction—Total RNA was extracted from a human gall bladder using the guanidine isothiocyanate/CsCl method (28, 29).

Reverse Transcription and Amplification—A sample (0.5 μg) of human tracheal poly(A)$^+$ RNA (Clontech) and a sample (1 μg) of total RNA extracted from human gall bladder were reverse-transcribed with the 1st-STRAND™ cDNA synthesis kit (Clontech) using random primers according to the manufacturer's instructions.

The first strand cDNA (8 μl) and cosmid DNA (30 ng) were amplified by PCR with various primers: NAU112 (antisense) 5′-ACCAGGCTGGGCCTGGGCACGGCA-3′ (nt 3105–3129), NAU113 (sense) 5′-GACGACTACAGCCACTGCCCCAGTACCCTA-3′ (nt 1671–1697), NAU81 (sense) 5′-CCAACTGGACCCTGGCACAGGTG-3′ (nt 694–716), NAU82 (antisense) 5′-GACTGAGGAGGACACAGTGGACACG-3′ (nt 10601–10625), NAU128 (sense) 5′-CGTGTCCACTGTGTCCTCCTCAGTC-3′ (nt 10601–10625), NAU71 (antisense) 5′-AGTGCTGATTGCACACTGCGT-3′ (in the first exon downstream the central exon), NAU136 (sense) 5′-TTCAACTATGAAATCCGTGTGTTC-3′ (nt 8493–8516)

The thermal cycling protocol used was the same as the one described above except that the annealing temperature was 62 °C. PCR experiments were performed using a Perkin-Elmer apparatus.

Cloning of Amplification Products—RACE-PCR products were separated by electrophoresis. Parts of the gel containing bands of interest were excised and the DNA was purified using Glassmilk (BIO 101, Inc.) and cloned into pGEMT vector (Promega). PCR products were purified using Preps DNA purification resin (Promega) and cloned into pMOS*blue* vector (Amersham).

Cloning in pKS—The fragments of interest from phage or cosmid clones were subcloned into the pBluescript KS(+) vector from Stratagene.

Plasmid DNA Purification—We used the Wizard™ minipreps DNA purification system (Promega).

DNA Sequencing and Sequence Analyses—The clones were sequenced on both strands by the dideoxy chain termination method using α-^{35}S-dATP with Sequenase version 2.0 (U. S. Biochemical Corp.), Sequitherm (TEBU), or the ^{T7}Sequencing™ kit (Pharmacia Biotech Inc.). They were sequenced using synthetic oligonucleotides corresponding to the T7 and T3 primers of the pKS plasmid, to the T7 and −40 primers of the pGEMT or pMOS*blue* vector. Part of the sequence was determined by primer walking using primers specific to the *MUC5B* gene. Analyses of nucleic acid and protein sequence data were performed using PC/GENE Software.

To perform DNA sequencing directly on cosmids (2 μg), we had to anneal at 37 °C for 30 min and to use 5 pmol of primers instead of 0.5 pmol when sequencing DNA in plasmids using the Sequenase version 2.0. The nucleotide sequence reported in this paper has been submitted to the EMBL Data Bank with accession number Z72496.

RESULTS

Southern Analysis—Human genomic DNA from leukocytes of 20 healthy unrelated volunteers was digested with *Bam*HI, *Bgl*II, *Eco*RI, *Hin*dIII, *Kpn*I, *Pst*I, *Xba*I, and *Xho*I. The sizes of the fragments obtained and hybridized with the JER57 probe are indicated in Table I. These results indicated the fragments of interest recognized with the JER57 probe. We isolated all these fragments from a phage or a cosmid genomic library and sequenced them to obtain the complete sequence recognized with the JER57 probe (see below).

Isolation and Restriction Mapping of MUC5B Genomic DNA

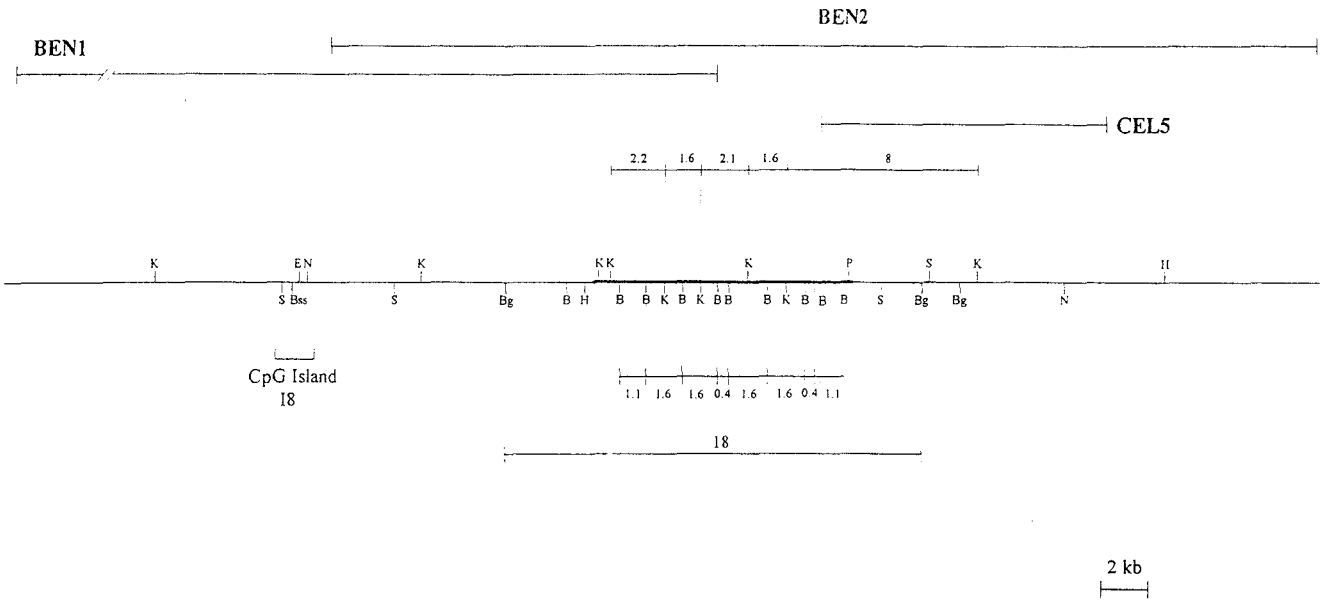


FIG. 1. **Partial restriction map of the three overlapping genomic clones.** CEL5 is a phage clone. BEN1 and BEN2 are cosmid clones. The *thicker line* represents the central exon. *Numbers* stand for the lengths (kb) of the fragments obtained with various restriction enzymes: *B*, *Bam*HI; *Bg*, *Bgl*II; *E*, *Eco*RI; *H*, *Hin*dIII; *K*, *Kpn*I; *P*, *Pst*I; *N*, *Not*I; *Bss*, *Bss*HII; and *S*, *Sac*II.

Clones—The EMBL4 human genomic library has been screened with the JER57 probe, and one positive clone with an insert of approximately 12 kb called CEL5 was obtained. The JER57 probe hybridized with one *Bam*HI-*Bam*HI fragment of 1.1 kb (indicated in Figs. 1 and 2*A*) situated in the 5′ part of this clone. This fragment was completely sequenced. Other screenings of the EMBL4 genomic library only gave the same clone. This led us to screen a human placenta genomic DNA library in cosmid vector pWE15 using JER57 as probe. Two cosmid clones containing inserts of approximately 40 kb were obtained and called BEN1 and BEN2. The partial restriction map of the three clones (CEL5, BEN1, and BEN2) is indicated in Fig. 1. BEN1 contains a CpG island, since restriction sites such as *Bss*HII and *Not*I are close together (30, 31). This island is located at 24 kb from the 5′-end of the insert and corresponds to the I 8 CpG island on the macrocartography performed in the 11p15.5 region (32). BEN2 overlaps the 3′ region of BEN1 on 16 kb and overlaps completely the CEL5 clone. We have found on these clones all the restriction fragments corresponding to the fragments recognized with the JER57 probe and observed on Southern blots of genomic DNA (Fig. 1 and Table I).

Strategy for the Sequencing of the Cosmid Genomic Fragments Hybridizing with the JER57 Probe—We prepared, subcloned into pBluescript KS(+) and sequenced all the fragments indicated in Fig. 2*A* at the *top*. Various restriction fragments derived from these clones were also subcloned to determine the entire sequence. We distinguished the fragments obtained from BEN2 only (with an *asterisk*). One fragment *Bam*HI-*Bam*HI of 1.1 kb was obtained from BEN2 and from CEL5 (noted with an *asterisk* and ++++). Since some fragments had the same lengths but were at various positions, we had to carefully isolate them starting from larger fragments unambiguously positioned, obtained for example only from BEN2, or for others only from BEN1. Due to an extremely high proportion of GC residues, many sequence problems were encountered. Subclones were thus sequenced several times before a reliable sequence was obtained.

Determination of Intron-Exon Boundaries on Each Side of the Central Exon—To obtain cDNAs upstream of the central tandem repeat region we used 5′ RACE-PCR with the primers NAU58 and NAU57 chosen in a nonrepeat region in 5′ of the central region. The amplification products were analyzed by agarose gel electrophoresis. Four ethidium bromide-stained bands ranging from 0.4 to 1 kb were obtained (data not shown) and cloned into pGEMT vector. Several transformants were isolated and sequenced. The nucleotide sequence showed that the 3′-ends of all the clones sequenced overlap with the genomic sequence. The comparison of the sequence of the longest cDNA (985 bp) with the sequence of the fragment *Bam*HI-*Kpn*I of 1.35 kb (noted with a *vertical arrow* on the *left* in Fig. 2*A*) from the cosmid BEN1 indicates an intron of 468 bp (nucleotide sequence not shown), thus identifying the 5′-end of the central exon. Splice acceptor and donor sequences agree with the "GT-AG" rule (33).

To determine the 3′-end of the central exon we performed RT-PCR using two primers (NAU128 and NAU71, their positions are shown on Fig. 2*A*) on human RNAs from tracheobronchial tissue and from gall bladder. Then we compared the sequences obtained after subcloning into pMOS*blue* vector with those determined on fragments of the cosmid located 3′ of the central exon: fragment *Bam*HI-*Sac*II ~ 1.7 kb and *Pst*I-*Sac*II ~ 1.3 kb (noted with *vertical arrows* on the *right* in Fig. 2*A*). The presence of an intron at 27 bp downstream from the site *Pst*I evidenced by comparing the sequences of the two cDNAs obtained with the genomic sequence. Splice acceptor and donor sequences agree with the GT-AG rule (33).

In order to confirm that the sequence was in the correct order, we performed restriction maps of overlapping fragments such as the 8-kb *Kpn*I-*Kpn*I fragment from BEN2 or the 5.5-kb *Hin*dIII-*Not*I 3′-end fragment from BEN1 (Fig. 1).

To confirm that this sequence forms only one exon, additional PCR products were produced using various pairs of primers (see Fig. 2*A*). We compared the lengths and the sequences of the fragments obtained first by RT-PCR on gall bladder or tracheobronchial RNA and by PCR on genomic DNA (cosmids BEN1 or BEN2). Using NAU81 and NAU112 we obtained fragments of ~2500 bp, using NAU113 and NAU112 fragments of ~1400 bp, using NAU82 and NAU136 fragments of ~2100 bp. In each case, we obtained the same lengths and the same sequences with either cDNAs or genomic DNA as matrices. No intron was detected in this way. Thus the sequenced region consists of a single long open reading frame which extends over

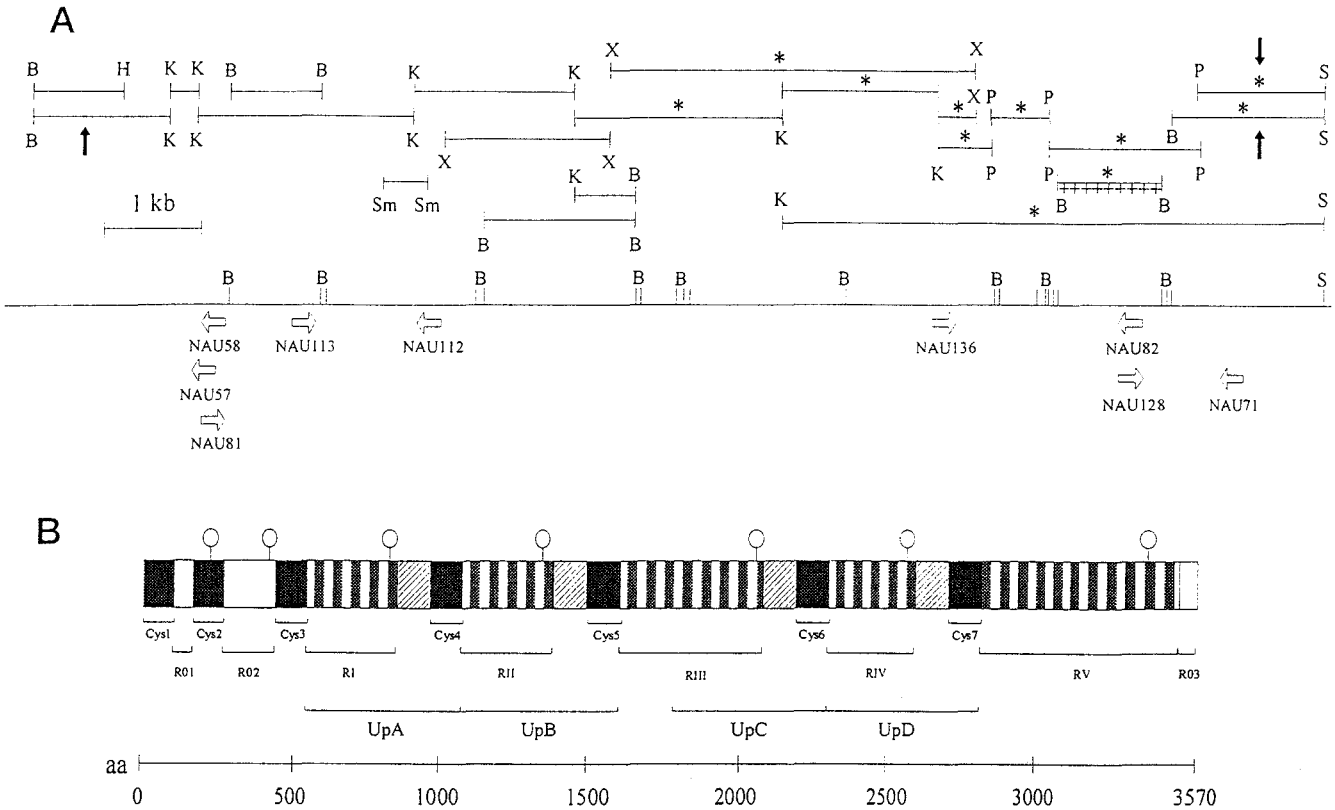


FIG. 2. **Sequencing strategy, *Bam*HI restriction map, and schematic protein subdomains diagram of the open reading frame.** *A* at the *top* shows the subcloned fragments of the central exon. The following restriction sites are displayed: *B*, *Bam*HI; *H*, *Hin*dIII; *K*, *Kpn*I; *P*, *Pst*I; *X*, *Xho*I; *Sm*, *Sma*I and *S*, *Sac*II. The fragments with an *asterisk* are only from BEN2. The *Bam*HI-*Bam*HI fragment with ++++ and an *asterisk* is from BEN2 and CEL5. The explanation of *vertical arrows* is given in the text (see "Results"). *A* also shows the distribution of the *Bam*HI sites within the central exon (underlying diagram). The primers and their directions (not to scale) are indicated by *horizontal open arrows* and their NAU numbers (their sequences are given under "Experimental Procedures"). *B*, representation of the deduced amino acid sequence of the open reading frame. The 19 subdomains are depicted: ▩, cysteine-rich subdomains; ▩□▩□, R subdomains containing the imperfectly conserved 87-bp repeat; ▨, R-end subdomains; *lollipop symbol*, potential *N*-glycosylation site. The four super-repeats UpA, UpB, UpC, and UpD are indicated. The amino acid scale is depicted beneath.

10,713 bp (submitted to the EMBL Data Bank with accession number Z72496) coding for 3570 amino acid residues and leading to a polypeptide core with a calculated M_r of 370,000.

Comparison with the cDNAs Described Previously—The nucleotide sequences of the previously isolated cDNAs JUL10, JER28, and JUL7 (23) have been positioned within the central exon. JUL10 is at position 5272–6258. JUL7 overlaps JER28. They are respectively positioned at 8510–10143 and at 9172–9732. Some small differences in sequence were observed between the genomic sequence determined here and these cDNAs as determined previously by us (23). As we did not succeed to position the JER57 sequence as reported by us (23), we redetermined its nucleotide sequence (accession number X74955 with correction submitted to the EMBL Data Bank). The reason for this discrepancy between this newly determined sequence of JER57, which is at position 8361–10222, and our previous work is explained under "Discussion."

These cDNAs were obtained by screening a human tracheobronchial tissue cDNA expression library using antibodies (23). Their deduced amino acid sequences allowed us to choose the appropriate open reading frame of the central exon continuous with that of their sequences.

Analysis of the Nucleotide Sequence and of the Deduced Amino Acid Sequence—Nineteen subdomains are individualized and are indicated in Fig. 2*B*. Seven code for cysteine-rich subdomains called Cys1 (aa 6–112), Cys2 (aa 176–283), Cys3 (aa 457–565), Cys4 (aa 986-1093), Cys5 (aa 1515–1622), Cys6 (aa 2213–2320), and Cys7 (aa 2742–2849). Their amino acid sequences are displayed in Fig. 3. Their average amino acid composition is given in Table II (column Cys). These subdomains are rich in cysteine residues (9.3%). In Fig. 3 is also shown the similarity between these cysteine-rich subdomains and homologous domains found in other mucins such as human MUC5AC or related cDNAs (3, 5), mouse Muc5ac (34), pig gastric mucin (35), human MUC2 (9), and rat MUC2 homologue (36). In addition to the remarkable conservation of the cysteine residues, the conservation of numerous amino acid residues (*bold boxes* or *boxes* in Fig. 3) such as tryptophan, proline, arginine, and glycine is to be noted, especially between MUC5B and MUC5AC and its related cDNAs. There is one conserved potential *O*-glycosylation site in each of these cysteine-rich subdomains (*boxed* together with a tryptophan residue in the *upper part* of Fig. 3). No potential *N*-glycosylation site exists. The Cys4 to Cys7 subdomains of MUC5B mucin present a perfect sequence similarity except in three positions. The average amino acid composition of the seven cysteine-rich subdomains (see column Cys in Table II) is very similar to that of domain III in HGM-1 (5) (Fig. 3). It is noticeable that all but 2 of the cysteine residues of the central exon are in the cysteine-rich subdomains. 31 out of the 32 tyrosine residues present in the central exon are in the cysteine-rich subdomains.

The first two subdomains between Cys1-Cys2 and Cys2-Cys3, subdomains R01 (64 amino acid residues) and R02 (174 amino acid residues), respectively, are enriched with threonine, serine, proline, and alanine (Table II), but no typical repeats or even imperfect repeats can be discerned. Searching of the GenBank™ data base indicated that these sequences were not identical with any registered sequence. A certain similarity

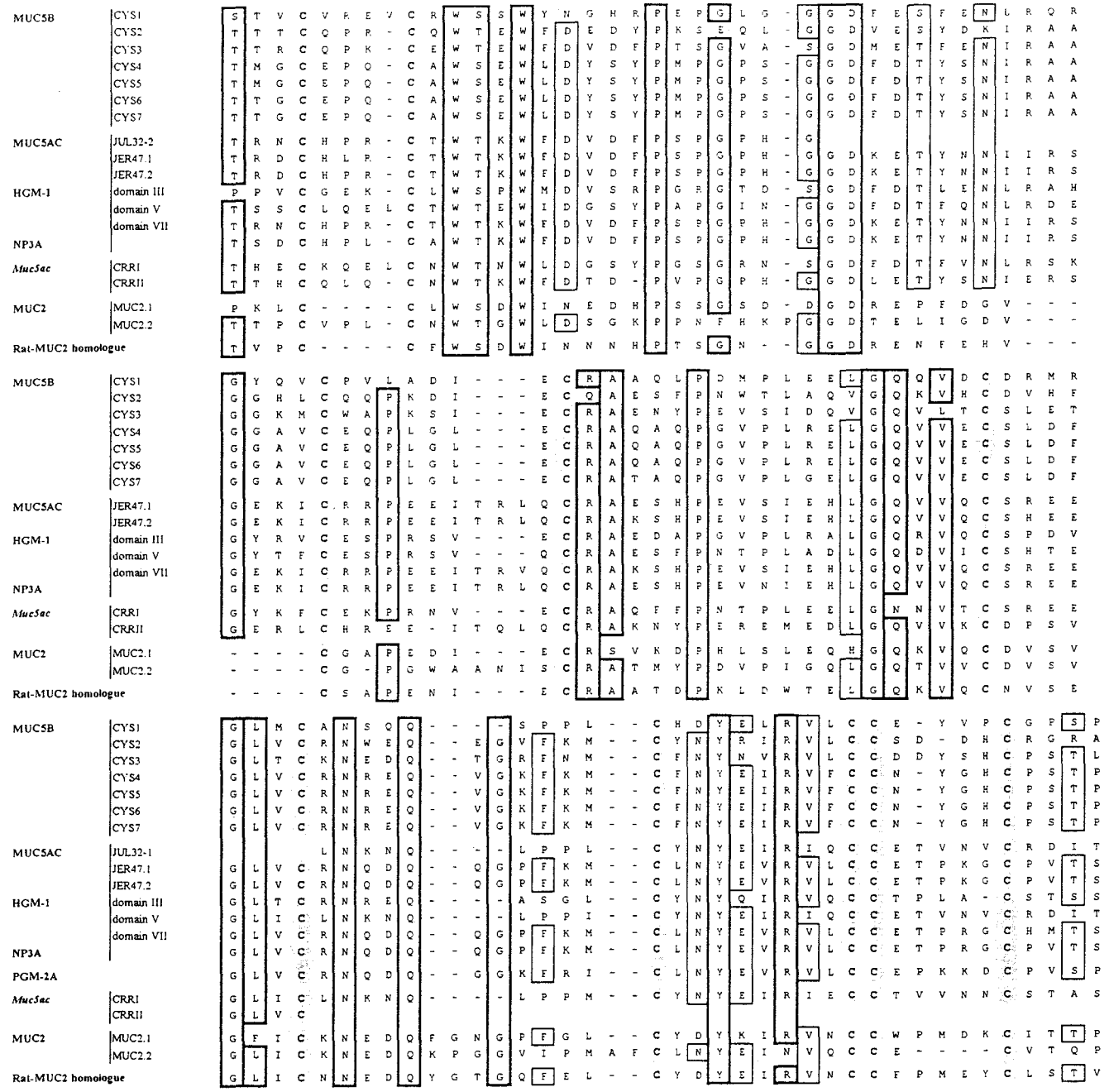


FIG. 3. **Amino acid sequence similarity between the cysteine-rich subdomains of human MUC5B mucin and other human or animal mucins.** Sequences were deduced from nucleotide sequences and were aligned to give maximum identity with a minimal number of gaps. JUL32, JER47-1, and JER47-2 are parts of MUC5AC (12); domains III, V, and VII are from HGM-1 (5); NP3a (3), CRRI, and CRRII are from mouse Muc5ac (34); MUC2-1 and MUC2-2 (9); rat MUC2 homologue (36); PGM-2A is from pig (35). Conserved cysteine residues are shadowed. *Bold boxed-in* positions have 90–100% conservation, and *boxed* positions have from 70 to 90% conservation. Serine and threonine residues were *boxed* together.

with other mucins, especially MUC1 or MUC2, is due to the typical amino acid composition of mucins.

Five subdomains (RI to RV) composed of various numbers of 87-bp imperfect tandem repeats are encountered downstream of the regions coding for the cysteine-rich subdomains Cys3 to Cys7. The alignment of the corresponding nucleotide sequences is shown in Fig. 4*A*. Domains RI, RII, and RIV are very similar. Each contains 11 imperfectly conserved repeats of 87 bp. The comparison is more obvious when looking at the deduced amino acid sequences (Fig. 4*B*). The subdomain RIII contains the same 11 irregular repeats of 29 amino acid residues from RIII-6 to RIII-17 (except RIII-7). Moreover, RIII-1 to RIII-5 are very similar to the first five repeats of subdomains RI, RII, and RIV. The subdomain RV is composed of 22 irregular repeats. RV-1 to RV-5 are also very similar to the first five repeats of the subdomains RI to RIV. RV-6 to RV-17 (except RV-7 and RV-9) are similar to the 11 imperfectly conserved repeats of the other R-subdomains except for the insertion of the pentapeptide PTTTT in RV-14. Beneath the alignment is indicated the consensus sequence of this imperfectly conserved repeat and the percentages of the indicated amino acid residue(s). One serine residue has 100% conservation and 6 amino acid residues have more than 80% conservation. As far as all the five R-subdomains are concerned, 44 units have the 29-amino acid residue repeat out of the aligned 72 units. The others are composed of 28, 27, or 26 amino acid residues, although one contains 32

TABLE II

Amino acid compositions deduced from the nucleotide sequences of the various types of subdomain in the central exon

The number of residues (Nb) and their percentage (%) are indicated for each subdomain type. The first five amino acid residues of the central exon are not included. Results for amino acid residues of interest (Cys, Ser, Thr) are in bold type.

Amino acid	Subdomains											
	R01		R02		Cys		R		R-end		R03	
	Nb	%	Nb	%	Nb	%	Nb	%	Nb	%	Nb	%
Ala	7	11.1	16	9.2	38	5.0	212	10.3	35	7.9	6	8.0
Arg	1	1.6	8	4.6	44	5.8	30	1.5	19	4.3	3	4.0
Asn	0	0.0	1	0.6	28	3.7	11	0.5	0	0.0	0	0.0
Asp	0	0.0	0	0.0	37	4.9	0	0.0	0	0.0	0	0.0
Cys	**0**	**0.0**	**0**	**0.0**	**70**	**9.3**	**0**	**0.0**	**1**	**0.2**	**1**	**1.3**
Gln	3	4.8	6	3.5	43	5.7	18	0.9	5	1.1	1	1.3
Glu	2	3.2	8	4.6	58	7.7	22	1.1	8	1.8	0	0.0
Gly	4	6.3	12	6.9	73	9.7	109	5.3	25	5.6	3	4.0
His	0	0.0	0	0.0	11	1.5	35	1.7	22	5.0	1	1.3
Ile	1	1.6	0	0.0	15	2.0	47	2.3	10	2.3	0	0.0
Leu	3	4.8	15	8.7	48	6.3	92	4.5	20	4.5	5	6.7
Lys	1	1.6	1	0.6	17	2.2	24	1.2	4	0.9	0	0.0
Met	0	0.0	2	1.2	17	2.2	22	1.1	0	0.0	2	2.7
Phe	0	0.0	1	0.6	31	4.1	12	0.6	0	0.0	7	9.3
Pro	10	15.9	24	13.9	55	7.3	237	11.5	63	14.2	9	12.0
Ser	**11**	**17.5**	**23**	**13.3**	**41**	**5.4**	**314**	**15.3**	**81**	**18.2**	**15**	**20.0**
Thr	**15**	**23.8**	**51**	**29.5**	**32**	**4.2**	**765**	**37.2**	**135**	**30.4**	**15**	**20.0**
Trp	1	1.6	1	0.6	17	2.2	23	1.1	4	0.9	0	0.0
Tyr	0	0.0	1	0.6	31	4.1	0	0.0	0	0.0	0	0.0
Val	4	6.3	3	1.7	50	6.6	81	3.9	12	2.7	7	9.3
Total	63	100	173	100	756	100	2054	100	444	100	75	100

residues. The amino acid composition (Table II) shows an extremely high threonine content (37%) and high serine, proline, and alanine contents. There are numerous potential *O*-glycosylation sites. The sequence T*X*XP, which has been implicated as a major site for GalNAc addition, is found in most of the 72 repeats (5, 37). These regions are likely to be heavily glycosylated. Moreover, one potential *N*-glycosylation site exists in each R-subdomain, and it is noticeable that the surrounding sequences are exactly the same for the potential *N*-glycosylation sites situated in RI, RII, RIII, and RIV (TPPVP N TTATT).

Concerning the sequences called RI-end, RII-end, RIII-end, and RIV-end, the alignment of their nucleotide sequences (Fig. 5*A*) shows a striking similarity. A consensus sequence is indicated. The amino acid sequences are aligned in Fig. 5*B*. The 111 amino acid residues present a perfect sequence similarity to each other except in eight positions. Searching of the GenBank™ data base indicated that these sequences are not identical with any registered sequence. A certain sequence similarity with other mucin genes is only due to the high percentage of threonine, serine, proline, and alanine (Table II) encountered in these sequences. There are numerous potential *O*-glycosylation sites and five T*X*XP sequences are present in each R-end subdomain.

The central exon ends with a 75-amino acid peptide that we have called R03, since it consists of a sequence different from those of the other subdomains of the *MUC5B* central exon. It is enriched in threonine, serine, proline, phenylalanine, and valine (Table II), but no typical repeat exists. It is different from the sequences found at the 3′-ends of the tandem repeats of *MUC5AC* (3, 4) and of *MUC2* (9).

Close examination brought to the foreground the existence of a super-repeat found four times within the MUC5B central exon that we have called UpA, UpB, UpC, and UpD (Fig. 2*B*). Restriction mapping and sequence determination have shown that the *Bam*HI digestion sites approximately flank the super-repeats (Fig. 2*A*). The *Bam*HI digestion sites are quite regularly distributed within the central exon. This super-repeat consists of a R-subdomain with 11 imperfect repeats of 29 amino acid residues followed by a R-end subdomain and a cysteine-rich subdomain. It contains 528 amino acid residues. The alignment of these four super-repeats has been performed (not shown since the alignments of the various sequences existing in these super-repeats have already been shown in Figs. 3, 4*B*, and 5*B*). They present a striking similarity except in the 11 positions noticed before (in the R-end and the cysteine-rich subdomains) and in 54 positions in the first part of the super-repeat composed of the 11 imperfectly conserved repeats of 29 amino acid residues.

DISCUSSION

We report here the isolation of three overlapping genomic clones of the human *MUC5B* mucin, one from a phage library (CEL5) and two from a cosmid library (BEN1 and BEN2) using the longest cDNA of *MUC5B* that we had previously isolated in the laboratory and called JER57 (23). *MUC5B* has been mapped clustered with *MUC6*, *MUC2*, and *MUC5AC* to chromosome 11p15.5. The partial restriction map of the three overlapping clones has been determined. We have also determined the intron-exon boundaries of the central region encoding a single open reading frame. This coding region has been completely sequenced. Its length is 10,713 bp, and it codes for 3570 amino acid residues. *MUC5B* is expressed in mucous glands of tracheobronchial tissue, submaxillary glands, gall bladder, and endocervix (24–26). RT-PCR experiments carried out on total RNA from human gall bladder or on human tracheal poly(A)$^+$ RNA unambiguously demonstrated that this open reading frame consists of a single exon and does not contain any intron. It was not possible to amplify the complete length of this open reading frame by RT-PCR using the two primers, NAU81 and NAU82, that are on both ends of the central exon. However, by performing different amplifications using more closely situated primers, in particular in the *MUC5B* specific part of the cysteine-rich domains, we were able to produce fragments exhibiting the same lengths and the same sequences as those in the cosmid clones. No difference was found in the common sequences determined on the genomic clones from the two libraries, even though they came from two unrelated individuals. This gene differs from *MUC1* and *MUC2* for which variable numbers of randem repeats polymorphisms have been demonstrated (38, 11). Since the repeat unit of 87 bp is imperfectly

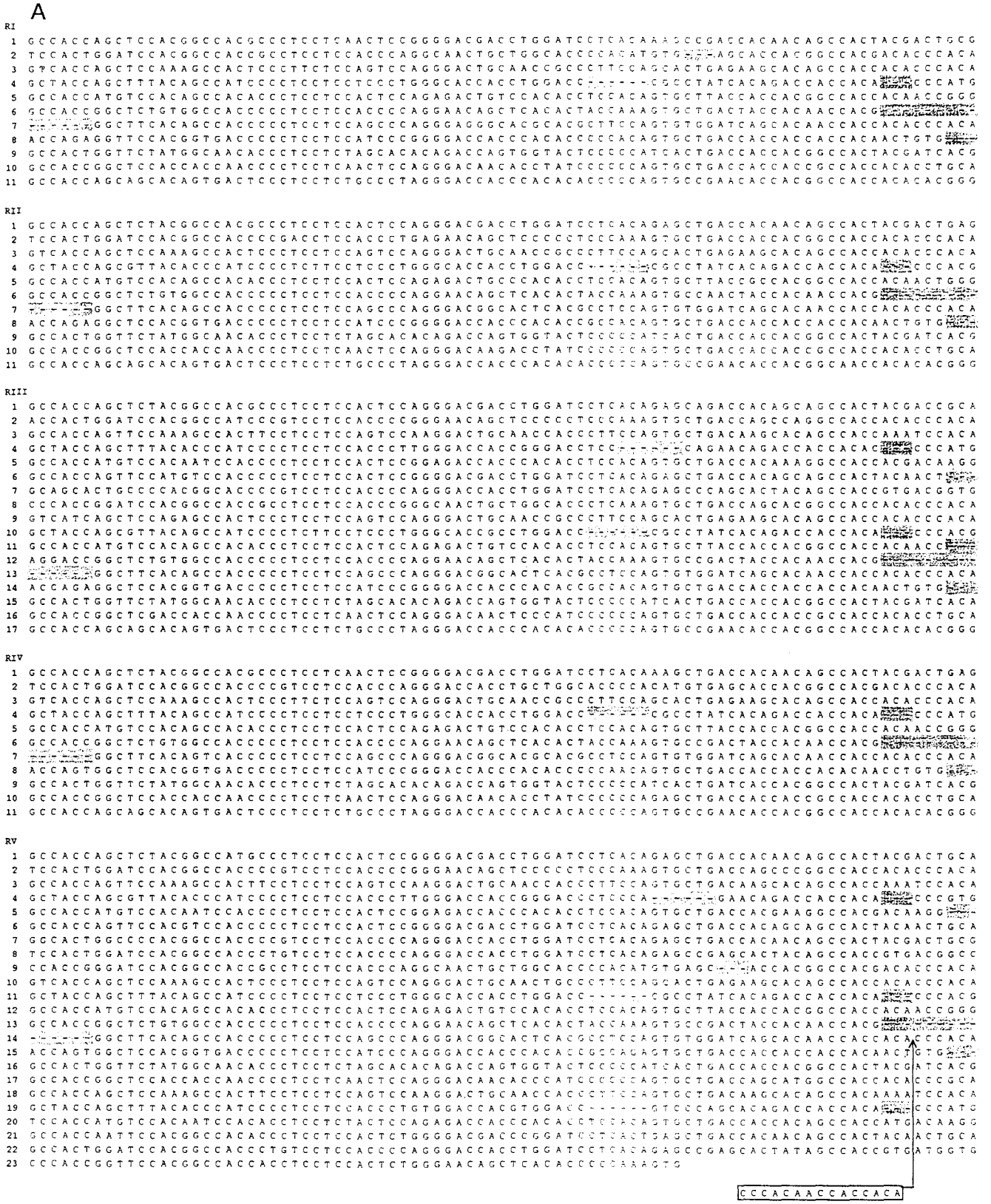


FIG. 4. **Nucleotide sequences (*A*) and deduced amino acid sequences (*B*) of the R-subdomains.** The residues of serine and threonine are *shadowed* in *B*.

conserved, we had to sequence the whole central exon, while in *MUC2* the authors were able to extrapolate the sequence repeated up to 100 times (9, 11).

The length of the central exon is much larger than that of *MUC1*, which varies from 3.5 to 6.2 kb, depending on the number of repeats (38). *MUC7*, a nonforming-gel mucin, possesses a central exon of 2.2 kb (22). As far as *MUC2* is concerned, the intron/exon distribution is not as yet known. Toribara *et al.* (11) have indicated that the 5′ region of the GMUC clone isolated in their laboratory forms, together with

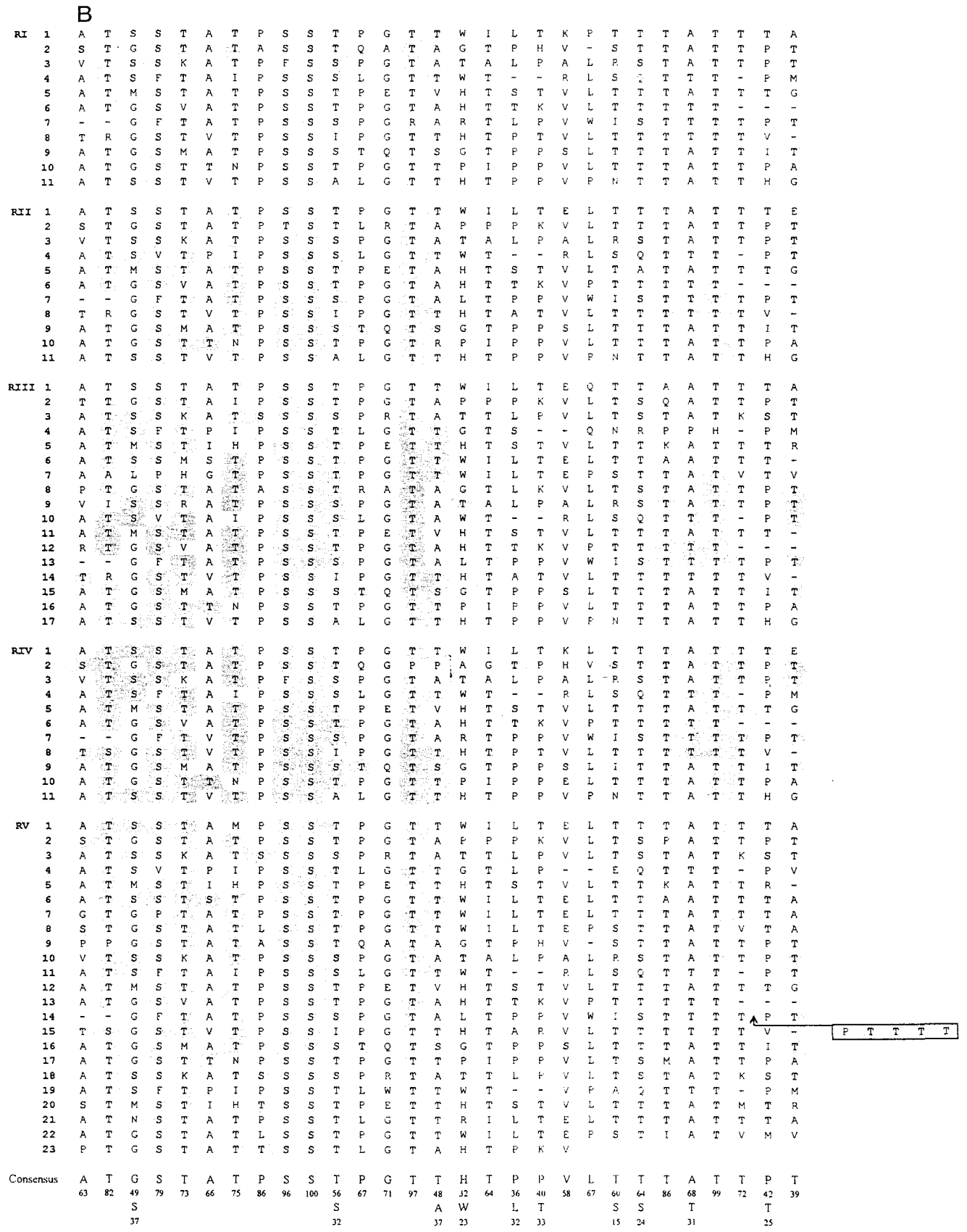


FIG. 4—*continued*

the tandem repeat array, a single large exon. In the most common genotype for *MUC2*, this region extends over 8,700 bp. No information is available about the size of the exonic repeat region of the other mucin genes. A survey on intron and exon lengths has indicated that in vertebrates few exons are over 800 bp (39). Most vertebrate genes are composed of long introns

A

```
Consensus  CGATCCCTGTCCCCCAGCAGTCCCCACACGGTGCGCACACCCTGGACTTCGGCCACCTCRGGCAYCTTG

RI End     CGATCCCTGTCCCCCAGCAGTCCCCACACGGTGTGCACAGCCTGGACTTCGGCCACCTCGGGCATCTTG
RII End    CGATCCCTGTCCCCCAGCAGTCCCCACACGGTGCGCACAGCCTGGACTTCGGCCACCTCAGGCACCTTG
RIII End   CGGTCCCTGCCCCCCAGCAGTCCCCACACGGTGCCCACAGCCTGGACTTCGGCCACCTCGGGCATCTTG
RIV End    CGATCCCTGTCCCCCAGCAGTCCCCACACGGTGCGCACAGCCTGGACTTCGGCCACCTCAGGCACCTTG

Consensus  GGCACCACCCACATCACAGAGCCTTCCACGGGGACTTCCCACACCCCAGCAGCAACCACCGGTACC

RI End     GGCACCACCCACATCACAGAGCCTTCCACGGGGACTTCCCACACCCCAGCAGCAACCACCGGTACC
RII End    GGCACCACCCACATCACAGAGCCTTCCACGGGGACTTCCCACACCCCAGCAGCAACCACCGGTACC
RIII End   GGCACCACCCACATCACAGAGCCTTCCACGGGGACTTCCCACACCCCAGCAGCAACCACCGGTACC
RIV End    GGCACCACCCACATCACAGAGCCTTCCACGGGGACTTCCCACACCCCAGCAGCAACCACCGGTACC

Consensus  ACCCAGCMCTCGACTCCAGCCCTTTCCAGCCCTCACCCTAGCAGCAGRACCACCGAGTCACCCCCT

RI End     ACCCAGCACTCGACTCCAGCCCTTTCCAGCCCTCACCCTAGCAGCAGAACCACCGAGTCACCCCCT
RII End    ACCCAGCACTCGACTCCAGCCCTTTCCAGCCCTCACCCTAGCAGCAGAACCACCGAGTCACCCCCT
RIII End   ACCCAGCCCTCGACTCCAGCCCTTTCCAGCCCTCACCCTAGCAGCAGGACCACCGAGTCACCCCCT
RIV End    ACAACGACCTCGACTCCAGCCCTGTCCAGCCCTCACCCTAGCAGCAGGACCACCGAGTCACCCCCT

Consensus  TCTCCAGGGACGACCACCCCGGGCCACACCACGGCCACCTCCAGGACCACAGCCACGGCCACACCC

RI End     TCTCCAGGGACGACCACCCCGGGCCACACCACGGCCACCTCCAGGACCACAGCCACGGCCACACCC
RII End    TCTCCAGGGACGACCACCCCGGGCCACACCACGGCCACCTCCAGGACCACAGCCACGGCCACACCC
RIII End   TCTCCAGGGACGACCACCCCGGGCCACACCAGGGGCACCTCCAGGACCACAGCCACAGCCACACCC
RIV End    TCCCCAGGGACGACCACCCCGGGCCACACCACGGCCACCTCCAGGACCACGGCCACGGCCACACCC

Consensus  AGCAAGACCCGCACCTCGACCCTGCTGCCCAGCMRSCCCACATCGGCCCCCCATAACCACGGTGGTG

RI End     AGCAAGACCCGCACCTCGACCCTGCTGCCCAGCCAGCCCACATCGGCCCCCCATAACCACGGTGGTG
RII End    AGCAAGACCCGCACCTCGACCCTGCTGCCCAGCAGCCCCACATCGGCCCCCCATAACCACGGTGGTG
RIII End   AGCAAGACCCGCACCTCGACCCTGCTGCCCAGCAGCCCCACATCGGCCCCCCATAACCACGGTGGTG
RIV End    AGCAAGACCCGCACCTCGACCCTGCTGCCCAGCCAGCCCACATCGGCCCCCCATAACCACGGTGGTG
```

B

```
RI End     RSLSPSSPHTVCTAWTSATSGILGTTHITEPSTGTSH

RII End    RSLSPSSPHTVRTAWTSATSGTLGTTHITEPSTGTSH

RIII End   RSLPPSSPHTVPTAWTSATSGILGTTHITEPSTGTSH

RIV End    RSLSPSSPHTVRTAWTSATSGTLGTTHITEPSTGTSH


RI End     TPAATTGTTQHSTPALSSPHPSSRTTESPPSPGTTTP

RII End    TPAATTGTTQHSTPALSSPHPSSRTTESPPSPGTTTP

RIII End   TPAATTGTTQPSTPALSSPHPSSRTTESPPSPGTTTP

RIV End    TPAATTGTTTTSTPALSSPHPSSRTTESPPSPGTTTP


RI End     GHTTATSRTTATATPSKTRTSTLLPSQPTSAPITTVV

RII End    GHTTATSRTTATATPSKTRTSTLLPSSPTSAPITTVV

RIII End   GHTRGTSRTTATATPSKTRTSTLLPSSPTSAPITTVV

RIV End    GHTTATSRTTATATPSKTRTSTLLPSQPTSAPITTVV
```

FIG. 5. **Nucleotide sequences (*A*) and deduced amino acid sequences (*B*) of the R-end subdomains.** In *A*, identical nucleotide sequences are *shaded*, and a consensus sequence is given: *R* is A or G, *Y* is T or C, *M* is A or C, and *S* is C or G. The residues of serine and threonine are *shadowed* in *B*.

and short exons. Composed of 10,713 bp, the central exon of *MUC5B* is even much larger than the 7572-bp exon of the gene for lipoprotein ApoB considered to be the largest one in vertebrates (40). The corresponding polypeptide coded by the central exon of *MUC5B* has a calculated M_r of 370,000.

Nineteen subdomains have been individualized. Seven subdomains of 108 amino acid residues, called Cys1 to Cys7, contain 10 cysteine residues and show a very similar organization to that observed twice in human (11) and rat (36) MUC2, at least twice in human (12), and mouse (34) MUC5AC as well as in other mucin cDNAs, which can now be identified as representing parts of the *MUC5AC* gene, *i.e.* NP3a (3), HGM-1 (5),

and L31 (4). The structure of the cysteine-rich subdomains has been conserved over a long evolutionary time scale, and the evolutionary constraint has likely been maintained because of crucial disulfide bonds. Thus, this typical cysteine-rich domain seems to be a feature of at least three of the four human mucin genes located on 11p15.5, little information is available as to whether there is a similar domain in *MUC6* (8). It is more and more tempting to speculate, as have Toribara *et al.* (8), that at least *MUC2*, *MUC5AC*, and *MUC5B* genes on chromosome 11p15.5 are part of a multigene family whose members code for secreted mucins. The conservation of numerous amino acid residues in addition to the cysteine residues suggests an important role for these domains. It is noticeable that the *MUC5B* and *MUC5AC* genes show greater similarity with each other than with the *MUC2* genes (from human and rat). Deletions of amino acid residues occur at the same positions in *MUC5B* and *MUC5AC*, while some observed deletions are characteristics of the *MUC2* type. Moreover, *MUC5B* and *MUC5AC* are very close together in the cluster (32). Overall, these observations suggest that *MUC5AC* and *MUC5B* may have the same ancestral gene.

It is noteworthy that the mouse mucin gene homologous to human *MUC5AC* is mapped to a site on mouse chromosome 7 homologous to the location of the human secretory mucin gene cluster on human chromosome 11p15.5 (34). We do not as yet know if mouse possesses both *Muc5ac* and *Muc5b*. Perhaps only one gene exists.

The presence of multiple cysteine residues strongly supports the idea that the *MUC5B* gene codes for a secreted mucin as disulfide bonding is necessary for the formation of a mucus gel. Moreover, we can postulate that interactions may occur via cysteine-rich domains between several mucin molecules producing a complicated network. This can explain the observations made using electron microscopy of bronchial mucins which have been shown to form complex entangled structures (41) or "bush-like" aggregates (42). The cysteine-rich domains may play a highly conserved role such as the packaging of the mucins or in mediating interactions with numerous other proteins such as, for example, the association with bilirubin in the gall bladder matrix (43). Further studies will be required to more adequately clarify the role of these domains.

An examination of the newly determined sequence of JER57 cDNA led us to conclude that the changes in the reading frame emphasized in our previous work (23) were in fact due to errors in sequence determination because of difficulties encountered in determining the sequence of such imperfectly conserved repeat domains. This is a problem that we have now overcome. We will now call the 87-bp repeat "imperfectly conserved" or "irregular" rather than "degenerate" repeat as designated previously. The cDNAs previously isolated (23) essentially consist of parts of the R-subdomain. In fact, the repeat of 29 amino acid residues of the R-subdomains appears to be only one of the components of the super-repeat. This super-repeat is the largest ever determined in mucin genes. The largest described until now was the 507-bp repeat unit of *MUC6* (8). The super-repeat in *MUC5B* is more than three times as long as the *MUC6* repeat. Its particularity is that it is made of three various subdomains. A variable number of this super-repeat has not been as yet observed. An interesting aspect of the deduced polypeptide sequence is the alternating arrangement of the three types of subdomains. The four R-end subdomains show a striking similarity to each other. Their role is likely to be crucial.

The conservation of the potential *N*-glycosylation sites within the repeat may be important for the proper maturation of MUC5B providing the positions at which *N*-glycosylation occurs. Recently, Denny *et al.* (44) reported that molecular cloning of mucin apoproteins showed that the consensus sequence(s) for *N*-glycosylation is usually found outside of the repeat domain. For MUC5B, five out of the seven potential *N*-glycosylation sites are within the repeat region. The subdomains R01, R02, and R03 display amino acid compositions typical of mucins but do not contain any repeat.

The *MUC5B* gene is clustered in 11p15.5 with *MUC2*, *MUC6*, and *MUC5AC* (8, 32). Although mucins are characterized by having tandem repeat regions containing significant amounts of threonine, serine, alanine, and proline, the individual repeats units for each of these genes are very different in length and amino acid sequence. Moreover, the number of the cysteine-rich subdomains emphasized here, which are different from the D-domains of human pro-von Willebrand factor (10), considerably differs between MUC2 and MUC5B. The role of these two types of secreted mucins may be highly specialized as suggested by *in situ* hybridization experiments which show a different expression pattern for each gene (24). In fact, we will now be able to design new oligonucleotides and produce fusion proteins and antibodies in order to reinvestigate the expression of the *MUC5B* gene using *in situ* hybridization and immunohistochemistry.

Experiments performed to elucidate the entire genomic organization of *MUC5B* are now progressing as is the study of the regulatory regions. Whether the regulatory elements of *MUC5B* function coordinately with the regulatory elements of adjacent mucin genes in the cluster is an important question to answer in the future.

Acknowledgments—We are indebted to Dr. A. T. Nurden for help in improving the style of the paper. We thank Pascal Mathon and Danièle Petitprez for technical assistance.

REFERENCES

1. Gendler, S. J., and Spicer, A. P. (1995) *Annu. Rev. Physiol.* **57,** 607–634
2. Shankar, V., Gilmore, M. S., Elkins, R. C., and Sachdev, G. P. (1994) *Biochem. J.* **300,** 295–298
3. Meerzaman, D., Charles, P., Daskal, E., Polymeropoulos, M. H., Martin, B. M., and Rose, M. C. (1994) *J. Biol. Chem.* **269,** 12932–12939
4. Lesuffleur, T., Roche, F., Hill, A. S., Lacasa, M., Fox, M., Swallow, D. M., Zweibaum, A., and Real, F. X. (1995) *J. Biol. Chem.* **270,** 13665–13673
5. Klomp, L. W. J., Van Rens, L., and Strous, G. J. (1995) *Biochem. J.* **308,** 831–838
6. Troxler, F. R., Offner, G. D., Zhang, F., Iontcheva, I., and Oppenheim, F. G. (1995) *Biochem. Biophys. Res. Commun.* **217,** 1112–1119
7. Aubert, J. P., Porchet, N., Crépin, M., Duterque-Coquillaud, M., Vergnes, G., Mazzuca, M., Debuire, B., Petitprez, D., and Degand, P. (1991) *Am. J. Respir. Cell. Mol. Biol.* **5,** 178–185
8. Toribara, N. W., Roberton, A. M., Ho, S. B., Kuo, W.-L., Gum, E., Hicks, J. W., Gum, J. R., Jr., Byrd, J. C., Siddiki, B., and Kim, Y. S. (1993) *J. Biol. Chem.* **268,** 5879–5885
9. Gum, J. R., Jr., Hicks, J. W., Toribara, N. W., Rothe, E.-M., Lagace, R. E.,and Kim, Y. S. (1992) *J. Biol. Chem.* **267,** 21375–21383
10. Mancuso, D. J., Tuley, E. A., Westfield, L. A., Worrall, N. K., Shelton-Inloes, B. B., Sorace, J. M., Alevy, Y. G., and Sadler, J. E. (1989) *J. Biol. Chem* **264,** 19514–19527
11. Toribara, N. W., Gum, J. R., Culhane, P. J., Lagace, R. E., Hicks, J. W., Petersen, G. M., and Kim, Y. S. (1991) *J. Clin. Invest.* **88,** 1005–1013
12. Guyonnet-Dupérat, V., Audié, J. P., Debailleul, V., Laine, A., Buisine, M. P., Zouitina-Galiègue, S., Pigny, P., Degand, P., Aubert, J. P., and Porchet, N. (1995) *Biochem. J.* **305,** 211–219
13. Neutra, M. R., and Forstner, J. F. (1987) in *Physiology of the Gastrointestinal Tract* (Johnson, L. R., ed) pp. 975-1009, Raven Press, New York
14. Van Klinken, B. J. W., Dekker, J., Büller, H. A., and Einerhand, A. W. C. (1995) *Am. J. Physiol.* **269,** G613–G627
15. Ho, J. J. L., and Kim, Y. S. (1995) *Int. J. Oncol.* **7,** 913–926
16. Lesuffleur, T., Zweibaum, A., and Real, F. X. (1994) *Crit. Rev. Oncol./Hematol.* **17,** 153–180
17. Buisine, M. P., Janin, A., Maunoury, V., Audié, J. P., Delescaut, M. P., Copin, M. C., Colombel, J. F., Degand, P., Aubert, J. P., and Porchet, N. (1996) *Gastroenterology* **110,** 84–91
18. Gendler, S., Taylor-Papadimitriou, J., Duhig, T., Rothbard, J., and Burchell, J. (1988) *J. Biol. Chem.* **263,** 12820–12823
19. Gum, J. R., Jr., Hicks, J. W., Toribara, N. W., Siddiki, B., and Kim, Y. S. (1994) *J. Biol. Chem.* **269,** 2440–2446
20. Bobek, L. A., Tsai, H., Biesbrock, A. R., and Levine, M. J. (1993) *J. Biol. Chem.* **268,** 20563–20569
21. Lancaster, C. A., Peat, N., Duhig, T., Wilson, D., Taylor-Papadimitriou, J., and Gendler, S. J. (1990) *Biochem. Biophys. Res. Commun.* **173,** 1019–1029
22. Bobek, L. A., Liu, J., Sait, S. N. J., Shows, T. B., Bobek, Y. A., and Levine, M.

J. (1996) *Genomics* **31,** 277–282
23. Dufossé, J., Porchet, N., Audié, J. P., Guyonnet-Dupérat, V., Laine, A., Van-Seuningen, I., Marrakchi, S., Degand, P., and Aubert, J. P. (1993) *Biochem. J.* **293,** 329–337
24. Audié, J. P., Janin, A., Porchet, N., Copin, M. C., Gosselin, B., and Aubert, J. P. (1993) *J. Histochem. Cytochem.* **41,** 1479–1485
25. Audié, J. P., Tétaert, D., Pigny, P., Buisine, M. P., Janin, A., Aubert, J. P., Porchet, N., and Boersma, A. (1995) *Hum. Reprod.* **10,** 98–102
26. Campion, J. P., Porchet, N., Aubert, J. P., L'Helgoualc'h, A., and Clément, B. (1995) *Hepatology* **21,** 223–231
27. Wahl, G. M., Lewis, K. A., Ruiz, J. C., Rothenberg, B., Zhao, J., and Evans, G. A. (1987) *Proc. Natl. Acad. Sci. U. S. A.* **84,** 2160–2164
28. Glisin, V., Orkvenjakov, R., and Byus, C. (1974) *Biochemistry* **13,** 2633–2637
29. Crépin, M., Porchet, N., Aubert, J. P., and Degand, P. (1990) *Biorheology* **27,** 471–484
30. Brown, W. R. A., and Bird, A. P. (1986) *Nature* **322,** 477–481
31. Bickmore, W. A., and Bird, A. P. (1992) *Methods Enzymol.* **216,** 224–244
32. Guyonnet-Dupérat, V. (1993) *Study of the Human Mucin Genes Located in 11p15.* Ph.D. thesis, University of Lille, France
33. Jacob, M., and Gallinaro, H. (1989) *Nucleic Acids Res.* **17,** 2159–2180
34. Shekels, L. L., Lyftogt, C., Kieliszewski, M., Filie, J. D., Kozak, C. A., and Ho, S. B. (1995) *Biochem. J.* **311,** 775–785
35. Turner, B. S., Bhaskar, K. R., Hadzopoulou-Cladaras, M., Specian, R. D., and Lamont, J. T. (1995) *Biochem. J.* **308,** 89–96
36. Ohmori, H., Dohrman, A. F., Gallup, M., Tsuda, T., Kai, H., Gum, J. R., Jr., Kim, Y. S., and Basbaum, C. B. (1994) *J. Biol. Chem.* **269,** 17833–17840
37. Elhammer, A. P., Poorman, R. A., Brown, E., Maggiora, L. L., Hoogerheide, J. G., and Kezdy, F. J. (1993) *J. Biol. Chem.* **268,** 10029–10038
38. 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
39. Hawkins, J. D. (1988) *Nucleic Acids Res.* **16,** 9893–9908
40. Blackhart, B. D., Ludwig, E. M., Pierotti, V. R., Caiati, L., Onasch, M. A., Wallis, S. M., Powell, L., Pease, R., Knott, T. J., Chu, M.-L., Mahley, R. W., Scott, J., McCarthy, B. J., and Levy-Wilson, B. (1986) *J. Biol. Chem.* **261,** 15364–15367
41. Sheehan, J. K., Oates, K., and Carlstedt, I. (1986) *Biochem. J.* **239,** 147–153
42. Slayter, H. S., Lamblin, G., Le Treut, A., Galabert, C., Houdret, N., Degand, P., and Roussel, P. (1984) *Eur. J. Biochem.* **142,** 209–218
43. Smith, B. J., and LaMont, J. T. (1985) *J. Clin. Invest.* **76,** 439–445
44. Denny, P. C., Mirels, L., and Denny, P. A. (1996) *Glycobiology* **6,** 43–50

L'exon central de *MUC5B* est équidistant d'une dizaine de kilobases d'un îlot CpG en 5' et d'un signal de polyadénylation en 3'. Cet exon est le plus grand jamais caractérisé chez les mammifères. Nous avons par ailleurs vérifié par PCR longue sur le cosmide BEN2 que la région comprise entre les oligonucléotides NAU81 et NAU82 (voir la localisation des oligonucléotides dans l'article) a bien une taille de plus de 10 kb (Figure 28). Cet énorme exon code un peptide déduit de 3570 aa composé majoritairement (3254 aa / 3570 aa) de l'agencement en alternance de 3 sous-domaines :

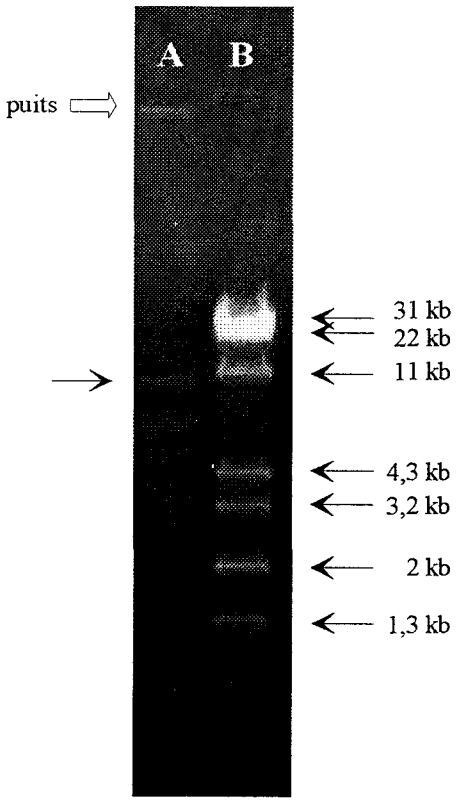
- un domaine de 108 aa, appelé sous-domaine Cys (« Cys subdomain »), riche en résidus Cys (10 Cys soit 9,3%) et trouvé 7 fois dans MUC5B, 2 fois dans MUC2, au moins 6 fois dans MUC5AC et retrouvé dans des mucines animales *Muc5ac*, rat-*Muc2* et PGM-2A. Ce domaine est pauvre en résidus Thr et Ser (Thr+Ser=9,6%).

- un domaine de 111 aa, appelé R-end, retrouvé 4 fois dans MUC5B et très conservé. Ce domaine est de type mucine, c'est-à-dire riche en résidus Thr et Ser (Thr+Ser=48,6%) et en résidus Pro (14,2%). Il a, de plus, la particularité de contenir un taux appréciable d'histidine (5%).

- un domaine trouvé 5 fois et appelé R. Ce domaine est constitué de la répétition en tandem (11, 17 ou 23 fois) du motif élémentaire et irrégulier de 29 aa (87 pb). Ces sous-domaines R sont riches en résidus Ser et Thr (Ser+Thr=52,5%), ainsi qu'en résidus Pro (11,5%). Chaque sous-domaine R possède un site potentiel de *N*-glycosylation dont l'environnement peptidique, pour les 4 premiers, est très conservé. De plus, ces 4 sites sont d'excellents sites potentiels de *N*-glycosylation puisqu'ils ont une séquence Asn-Xaa-Thr où Xaa est un aa hydroxylé (Thr) (Kasturi *et al.*, 1997).

Les deux sous-domaines de type mucine (R et R-end) ont un rapport Ser/Thr d'environ 0,5.

L'alternance des sous-domaines Cys / R / R-end donne naissance au motif élémentaire répété le plus grand (528 aa) jamais décrit jusqu'à présent pour une mucine.



A : PCR NAU81-NAU82 sur le cosmide BEN2

B : marqueur de tailles (13i)

Figure 28 : Electrophorèse en agarose (0.5%) du produit d'amplification obtenu par PCR sur BEN2 en utilisant les amorces NAU81 et NAU82

I.2 Travaux publiés simultanément concernant la région répétitive

Il avait été montré par Thornton *et al.* (Thornton *et al.*, 1996) que le mucus respiratoire contient 3 mucines oligomériques majoritaires. De plus, l'équipe de Carlstedt avait suggéré, en s'appuyant sur des résultats obtenus par des travaux de purification des mucines trachéobronchiques et par immunohistochimie, que MUC5AC, comparativement à MUC2, est une mucine abondamment produite dans les sécrétions respiratoires (Hovenberg *et al.*, 1996).

Peu après la publication de nos travaux de clonage et séquençage de la région centrale de *MUC5B*, l'équipe de Sheehan publie dans le même journal une étude biochimique du mucus trachéobronchique (Thornton *et al.*, 1997). Ces auteurs ont purifié une fraction moins acide du mucus respiratoire alors que la fraction la plus acide correspondrait à MUC5AC. Cette fraction contient 2 glycoformes, moins chargées que MUC5AC, de la même mucine identifiée comme MUC5B. Ces glycoformes ont été séparées, purifiées, réduites puis fragmentées par un traitement à la trypsine. Les 4 fractions peptidiques majoritaires obtenues, séparées par chromatographie, ont été séquencées par la dégradation de Edman. Les 4 séquences peptidiques de 10, 14, 17 et 8 aa sont homologues au peptide déduit (à l'exception d'un aa) de la séquence nucléotidique corrigée de JER57.

En fait, le 1er peptide pourrait être issu du motif Cys4, Cys5 ou Cys6 de MUC5B selon notre nomenclature (voir article précédent). Les 3 autres peptides sont 100% homologues à 3 autres régions présentes dans les motifs Cys4, Cys5, Cys6 et Cys7 de MUC5B.

Les études biochimiques de Thornton *et al.* confortent nos résultats concernant la région centrale de MUC5B. Ces mêmes auteurs suggèrent que le peptide MUC5B est en partie fait de l'agencement en alternance d'un motif riche en résidus Cys et de régions glycosylées plus ou moins longues.

I.3 Les gènes de mucines en 11p15: phylogenèse d'une famille de gènes

Les travaux ci-dessus montrent que certaines mucines, dont les 3 mucines humaines MUC2, MUC5AC et MUC5B issues du cluster de gènes localisé en 11p15.5, possèdent le motif Cys de 108 aa. Ce sous-domaine a la particularité de contenir 10 résidus Cys (ainsi que d'autres résidus comme Trp, Gln et Pro) aux positions très conservées. Ce sous-domaine définit ainsi une famille de gènes regroupant, jusqu'à présent, 3 mucines humaines et 3 mucines animales.

L'alignement des séquences nucléotidiques des sous-domaines Cys des différents membres de cette famille et le calcul des homologies entre ces séquences nous ont permis de bâtir un modèle cohérent de l'évolution des gènes humains *MUC2*, *MUC5AC* et *MUC5B* à partir d'un gène ancestral.

Ce modèle est présenté dans l'article ci-dessous :

« **Evolutionary history of the 11p15 human mucin gene family** »

J Mol Evol (1997) 45:000–000

JOURNAL OF MOLECULAR EVOLUTION

Evolutionary History of the 11p15 Human Mucin Gene Family

Jean-Luc Desseyn,[1] Marie-Pierre Buisine,[1,2] Nicole Porchet,[1,2] Jean-Pierre Aubert,[1,2] Pierre Degand,[1,2] and Anne Laine[1]

[1] INSERM U377 Laboratoire Gérard Biserte, place de Verdun, 59045 Lille Cedex, France
[2] Laboratoire de Biochimie et de Biologie Moléculaire de l'Hôpital C. Huriez, CHRU de Lille, 59037 Lille Cedex, France

Received: 30 January 1997 / Accepted: 17 April 1997

Abstract. The four human mucin genes *MUC6*, *MUC2*, *MUC5AC*, and *MUC5B* are located at chromosome 11p15.5. It has been demonstrated that the three mucins MUC2, MUC5AC, and MUC5B contain several Cys-subdomains of 108 amino acid residues. In contrast, little information is available concerning MUC6. These Cys-subdomains contain 10 cysteine residues that have a highly conserved position. We present here a coherent probable evolutionary history of this human gene family after comparison of the nucleotide sequences of these Cys-subdomains. The three MUC loci *MUC2*, *MUC5AC*, and *MUC5B* may have evolved from a common ancestral gene by two successive duplications. Moreover, we can postulate that *MUC5AC* and *MUC5B* have evolved in a concerted manner, while *MUC2* has evolved separately.

Key words: Human mucin — Cluster — 11p15.5 — Evolution — Cys-rich domain

Introduction

Mucins are the major component of mucus and constitute a group of high-molecular-mass glycoproteins. They have been implicated in a large number of pathological situations. Abnormal expression of the mucin genes occurs in various disease states and in conditions associated with a high risk of colonic adenoma or carcinoma (Buisine et al. 1996; Ho and Kim 1995). Over the last few years there has been a considerable effort in research on mucin genes. Until now, at least eight human mucin genes have been identified (for a more complete review, see Gendler and Spicer 1995). Four of them have been localized, in a cluster between *HRAS* and *IGF2* genes, at the short arm of chromosome 11 at band 15.5 The order of these mucin genes in the cluster has been determined to be *MUC6/MUC2/MUC5AC/MUC5B* (Pigny et al. 1996).

Until recently, the characteristic feature of the mucin core was to contain variable numbers of repeats with very different sizes and amino acid sequences. Another typical feature is the presence of the unique sequences of *MUC5AC*, *MUC2*, and *MUC5B* that flank the tandem repeat arrays and that encode the D domains that were found in the pro-von Willebrand factor (pro-vWF) (Gum et al. 1992; Klomp et al. 1995; Lesuffleur et al. 1995). In contrast, little information is available concerning *MUC6*. Furthermore, MUC2, MUC5AC, and MUC5B contain cysteine-rich subdomains of 108 amino acid residues. These Cys-subdomains which contain 10 well-conserved Cys residues have been found twice in MUC2, seven times in MUC5B, at least four times in MUC5AC and in its related cDNAs, twice in *Muc5ac* (mouse mucin gene), once in Pig Gastric Mucin (PGM-2A) and twice in rat-*Muc2* (Gum et al. 1992; Klomp et al. 1995; Desseyn et al. 1997; Meerzaman et al. 1994; Guyonnet Dupérat et al. 1995; Shekels et al. 1995; Turner et al. 1995; Ohmori et al. 1994). The role of this Cys-subdomain type is not yet known, although the fact that the cysteine residues and numerous other amino acid residues are well con-

Correspondence to: A. Laine

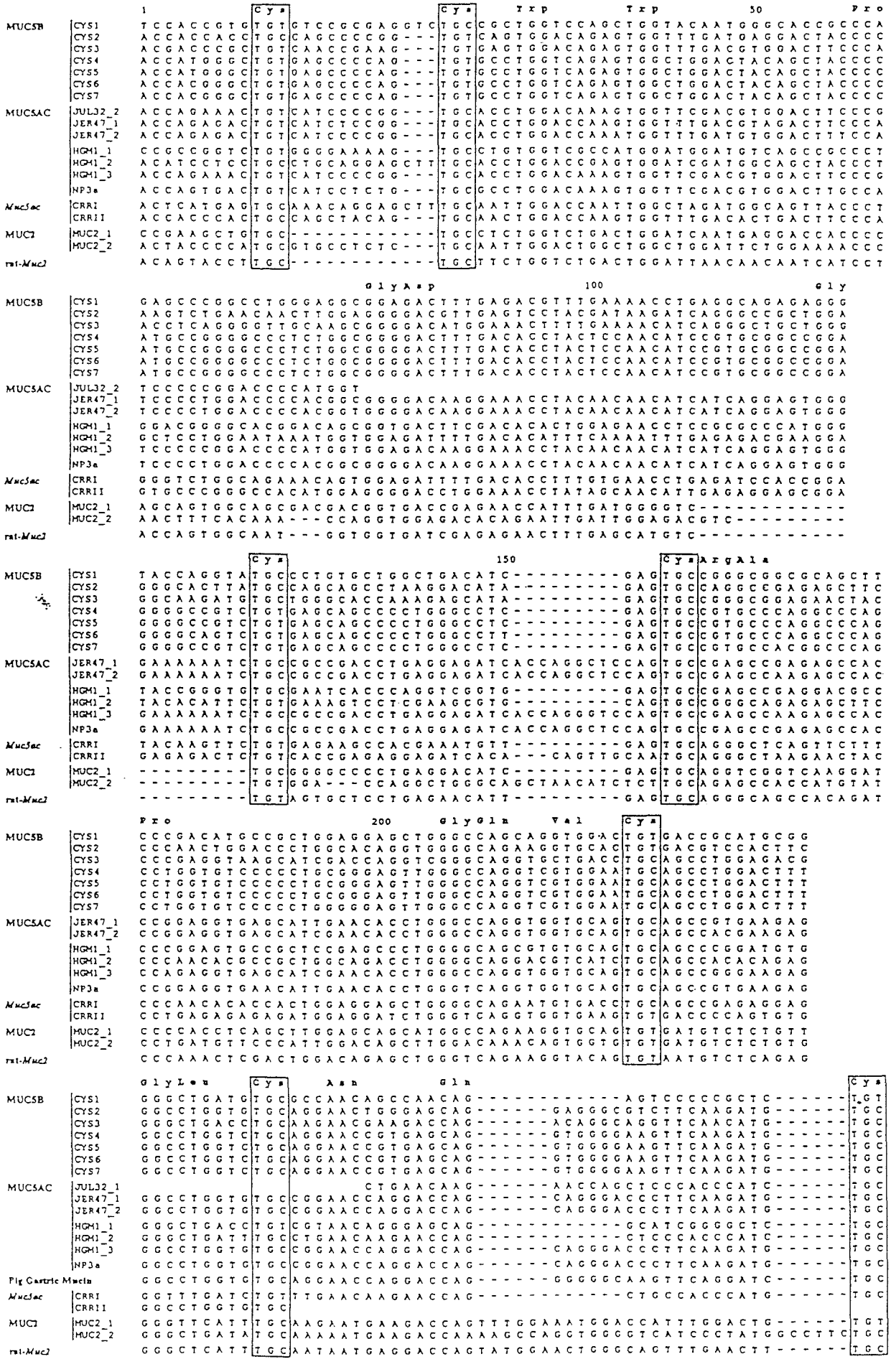


Fig. 1. Nucleotide sequences of the Cys-subdomains. The sequences are aligned to allow comparison (see methods). Gaps are shown by *hyphens*, one hyphen per nucleotide. The well-conserved amino acid residues are shown. Cysteine residues and nucleotide sequences coding for cysteine residues are *boxed*.

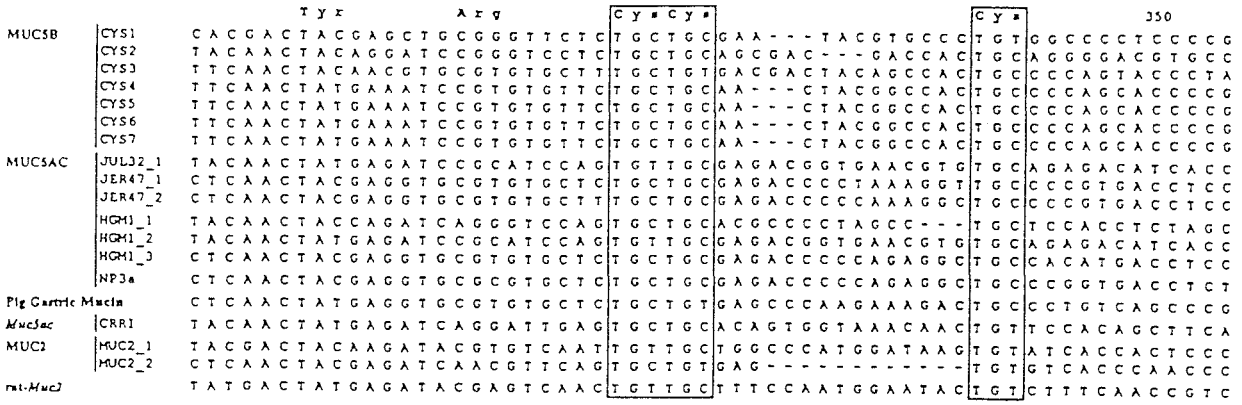


Figure 1. Continued.

served suggests that this peptide may be very important. It may, for instance, play a role in the packaging of mucins or in interactions with other proteins.

Methods

Nucleotide sequences

Nucleotide sequences are available from the EMBL data base with the following accession numbers, and the cysteine subdomains are defined as follows: *MUC5B:* Z72496; Cys1, nt 18–338; Cys2, nt 528–851; Cys3, nt 1371–1697; Cys4, nt 2958–3281; Cys5, nt 4545–4868, Cys6, 6639–6962; Cys7, 8226–8549. *MUC5AC:* Z34277; JER47_1, nt 55–390; JER47_2, nt 949–1284; Z34280, JUL32_1, nt 3–92; JUL32_2, nt 267–341; NP3a: U06711; nt 32–367; HGM1: X81649; HGM1_1, nt 1127–1444; HGM1_2, nt 1709–2032; HGM1_3, nt 2207–2542. *Muc5ac:* L42292; CRRI, nt 340–743; CRRII, nt 826–1062. *MUC2:* L21998; MUC2_1, nt 3922–4224; MUC2_2, nt 5362–5673. Rat-*Muc2:* U07615; nt 3981–4179. Pig gastric mucin: U12768 (PGM-2A); nt 1–111.

Analyses of Nucleic Acid Sequences

Computer analyses were performed with the NALIGN and CLUSTAL programs from PC/GENE Software. Gaps were introduced to optimize alignment.

Table 1. Sequence relationships between nucleic sequences coding mucin Cys-subdomains (see methods). Percent nucleotide identities are shown for all pairwise combinations

		MUC5B							*MUC5AC*									*Muc5ac*		*MUC2*		
		CYS1	CYS2	CYS3	CYS4	CYS5	CYS6	CYS7	JUL32_2	JUL32_1	JER47_1	JER47_2	HGM1_1	HGM1_2	HGM1_3	NP3a	PGM2A	CRRI	CRRII	MUC2_1	MUC2_2	Rat-*Muc2*
MUC5B	CYS1	100																				
	CYS2	56	100																			
	CYS3	56	66	100																		
	CYS4	55	66	66	100																	
	CYS5	55	66	66	100	100																
	CYS6	55	66	66	99	99	100															
	CYS7	56	67	66	99	99	99	100														
MUC5AC	JUL32_2	52	69	65	65	65	65	65	100													
	JUL32_1	49	54	49	51	51	51	51		100												
	JER47_1	53	62	65	62	62	62	62	89	59	100											
	JER47_2	53	62	66	61	61	61	61	89	59	96	100										
	HGM1_1	57	54	56	59	59	59	58	46	55	53	53	100									
	HGM1_2	59	58	55	55	55	55	55	59	91	56	56	60	100								
	HGM1_3	53	62	66	63	63	63	63	100	57	94	95	53	55	100							
	NP3a	53	61	66	62	62	62	62	86	59	95	94	53	55	93	100						
Pig	PGM2A	58	68	73	75	75	75	75		58	82	82	54	61	81	81	100					
Muc5ac	CRRI	59	56	56	56	56	56	55	48	64	50	50	61	72	49	49	52	100				
	CRRII	54	60	58	57	57	58	58	74		68	67	46	55	67	67	100	52	100			
MUC2	MUC2_1	50	53	52	46	46	46	47	48	49	48	49	50	46	48	47	53	46	43	100		
	MUC2_2	38	44	41	41	41	41	42	41	40	42	43	40	40	43	43	48	40	44	50	100	
Rat-*Muc2*		45	49	48	47	47	47	47	44	49	46	46	46	51	45	46	57	47	44	75	47	100

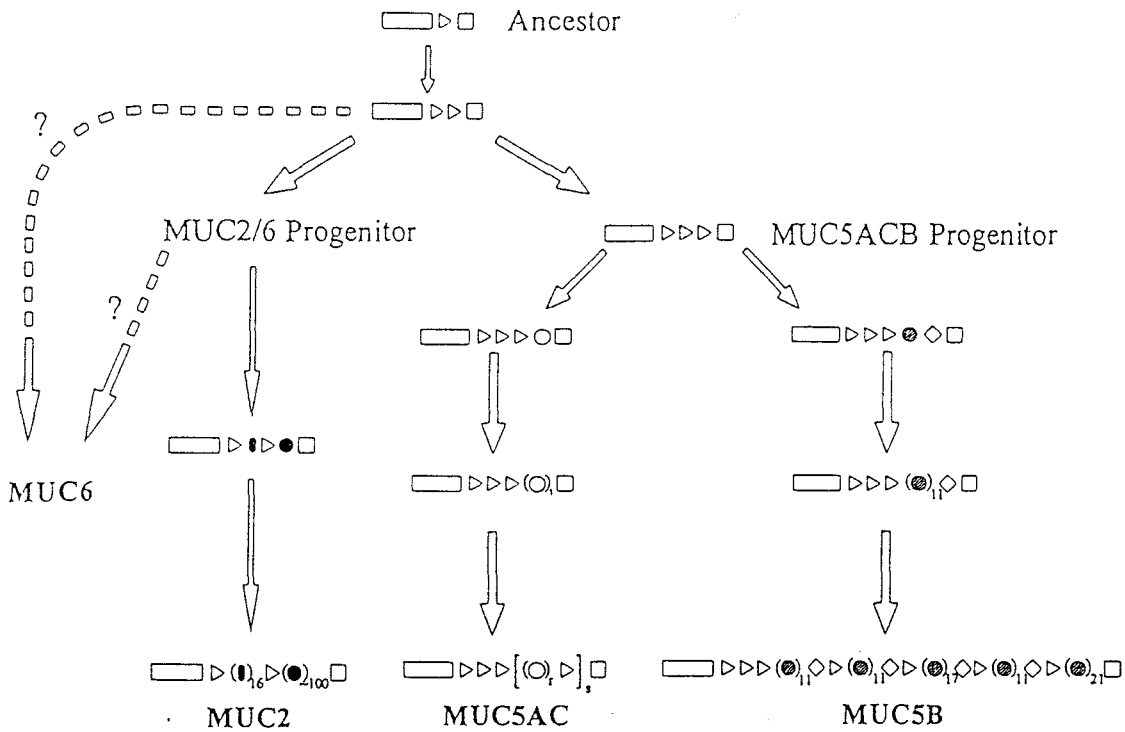


Fig. 2. Hypothetical diagram showing evolution of the 11p15.5 mucin genes *MUC6*, *MUC2*, *MUC5AC*, and *MUC5B* from a single ancestral gene. Unique domains flanking the tandem repeats arrays, conserved in each protein, are represented with *rectangular* (D1-D2-D'-D3 pro-von Willebrand domains) and *square boxes* (D4 von Willebrand domain). *Triangle* represents the Cys-subdomain type. The repetitive central domains are represented by the following symbols: circles (*stippled*, *empty*, and *hatched*) for the 23 amino acid tandem repeat polypeptides of MUC2, for the eight amino acid tandem repeat polypeptides of MUC5AC, for the 29 amino acid tandem repeat polypeptides of MUC5B (Gendler and Spicer 1995; Desseyn et al. 1997), respectively. The *oval* represents the irregular 16 amino acid tandem repeat polypeptide of MUC2; *t*, *r*, and *s* represent the unknown number of repeats. *Diamonds* represent the MUC5B R-end subdomain type (Desseyn et al. 1997). The *stippled arrows* indicate the two hypothetical routes that would link *MUC6* more closely to *MUC2*.

Results and Discussion

Sequencing data of the three genes *MUC2*, *MUC5AC*, and *MUC5B*, coding for large mucins, together with their clustered chromosomal localization strongly suggest that these genes have been created by a process of gene duplication. This observation provides an insight into the mechanisms that have shaped the evolution of this multigene family. An alignment of the nucleotide sequences of the Cys-subdomains is displayed in Figure 1.

For comparison, each base mismatch and each gap is counted as one difference on the basis of the alignment in Figure 1. Sequence relationships between the nucleic sequences coding for the Cys-subdomains are given in Table 1. The similarity between the Cys-subdomains of rat-*Muc2* and *Muc5ac* is less than 50%. As is made clear by Table 1, *MUC2* Cys-subdomains versus rat-*Muc2* Cys-subdomains are more similar (75% for rat-*Muc2* with MUC2-1) than *MUC2* Cys-subdomains versus *MUC5B* Cys-subdomains (38–53%) or *MUC5AC* Cys-subdomains (40–50%). Moreover, the 20-base pair deletion at nucleotide (nt) 109, for example, shows that rat-*Muc2* and *MUC2* are very close (Fig. 1). Furthermore, MUC5AC Cys-subdomains versus *Muc5ac* Cys-subdomains are more similar (72% for HGM1-2 with CRRI, moreover, see the 3-base pair insertion CTT at nt 22 in Fig. 1) than *MUC5AC* Cys-subdomains versus *MUC5B* Cys-subdomains (54–59% between HGM1_1 and the Cys-subdomains of *MUC5B*). In contrast, the Cys-subdomains of the rat-*Muc2* and *Muc5ac* (CRRI and CRRII) differ significantly (47% and 44%). It clearly follows that the greatest similarity between two human mucin genes is for *MUC5B* versus *MUC5AC*, suggesting that these two genes have evolved more recently. Nucleotide deletions at nt 262 (Fig. 1) for these two genes confirm our hypothesis. It must be emphasized that *MUC5B* and *MUC5AC* are close together in the cluster of 11p15.5 genes (Pigny et al. 1996). This is in good agreement with the idea that *MUC5B* and *MUC5AC* genes have the same more recent progenitor. *MUC6* and *MUC2* are in close proximity on chromosome 11p15.5 (Pigny et al. 1996), suggesting a possible duplication into *MUC2* and into *MUC6* from the same progenitor gene *MUC2/6*. Further studies on *MUC6* gene will be required to clarify this assumption, in particular the presence of D domains and Cys-subdomains. Hence, we inferred that a common ancestral gene has duplicated and diverged into at least two progenitor genes, the progenitor *MUC2/6* and the progenitor *MUC5ACB* (Fig. 2). Further duplication of the progenitor *MUC5ACB* gene gave rise to at

least two genes *MUC5B* and *MUC5AC*. Moreover, it might be postulated, therefore, that the duplication of the ancestral gene into the *MUC2/6* progenitor gene and into the *MUC5AC/B* progenitor gene had occurred before the rodent–primate divergence about 100 million years ago (Novacek 1992).

Presumably, the ancestral gene contained the pro-vWF domains conserved in MUC2, MUC5AC (Lesuffleur et al. 1995) and MUC5B (Desseyn et al. submitted). It contained also one Cys-subdomain which duplicated. This common ancestor gene diverged into two progenitor genes. Then, the internal repeat domains appeared. The last four Cys-subdomains of *MUC5B* share a high degree of similarity to each other (99% and 100%, Table 1). In contrast, the first three Cys-subdomains yield fewer similarities with each other (55% to 66%, Table 1). This suggests that the crossing-over events are more ancient for the first three Cys-subdomains. Likewise, the same observations could be made for the HGM-subdomains (i.e., the first three Cys-subdomains of the *MUC5AC* gene). HGM1-1, -2, and -3 nucleotide sequences yield fewer similarities with each other (53% to 60%) than JER47_1 with JER47_2 (96%) or JER47_1 with JUL32_2 (89%), for example. Moreover, the last four MUC5B Cys-subdomains are preceded by R-end subdomains that are very similar. This R-end subdomain type has not yet been found in MUC5AC. Its sequence has not been shown to be similar to that of any other protein (Desseyn et al. 1997). These structural features are suggestive of an involvement in *MUC5B* of unequal crossing over after the *MUC5B/MUC5AC* duplication, a process by which the last four Cys-subdomains, R-end subdomains, and tandem repeats have been generated. Our model is in good agreement with the scheme proposed by Gum (Gum 1992) A simple representation of the hypothetical evolutionary history of the four 11p15.5 human mucin genes is given in Figure 2.

At present, it is not possible to assess the precise molecular mechanisms that have shaped the structure of these four human mucin genes, however, gene duplication and unequal crossing over seem to be good candidates.

Acknowledgments. This work was supported by La Ligue contre le Cancer, l'Association de Recherche contre le Cancer and L'Association François Aupetit.

References

Buisine MP, Janin A, Maunoury V, Audié JP, Delescaut MP, Copin MC, Colombel JF, Degand P, Aubert JP, Porchet N (1996) Aberrant expression of a human mucin gene (*MUC5AC*) in rectosigmoid villous adenoma. Gastroenterology 110:84–91

Desseyn JL, Guyonnet Dupérat V, Porchet N, Aubert JP, Laine A (1997) Human mucin gene *MUC5B*: the 10.7 kb large central exon encodes various subdomains resulting in a super-repeat. Structural evidence for a 11p15.5 gene family. J Biol Chem 272:3168–3178

Gendler SJ, Spicer AP (1995) Epithelial mucin genes. Annu Rev Physiol 57:607–634

Gum JR (1992) Mucin genes and proteins they encode: structure, diversity, and regulation. Am J Respir Cell Mol Biol 7:554–564

Gum JR, Hicks JW, Toribara NW, Rothe EM, Lagace RE, Kim YS (1992) The human MUC2 intestinal mucin has cysteine-rich subdomains located both upstream and downstream of its central repetitive region. J Biol Chem 267:21375–21383

Guyonnet Dupérat V, Audié JP, Debailleul V, Laine A, Buisine MP, Galiègue-Zouitina S, Pigny P, Degand P, Aubert JP, Porchet N (1995) Characterization of the human mucin gene *MUC5AC*: a consensus cysteine-rich domain for 11p15 mucin genes? Biochem J 305:211–219

Ho JJL, Kim YS (1995) Do mucins promote tumor cell metastasis? Int J Oncol 7:913–926

Klomp LWJ, Van Rens L, Strous GJ (1995) Cloning and analysis of human gastric mucin cDNA reveals two types of conserved cysteine-rich domains. Biochem J 308:831–838

Lesuffleur T, Roche F, Hill AS, Lacasa M, Fox M, Swallow DM, Zweibaum A, Real FX (1995) Characterization of a mucin cDNA clone isolated from HT-29 mucus-secreting cells. The 3'end of MUC5AC? J Biol Chem 270:13665–13673

Meerzaman D, Charles P, Daskal E, Polymeropoulos MH, Martin BM, Rose MC (1994) Cloning and analysis of cDNA encoding a major airway glycoprotein, human tracheobronchial mucin (MUC5). J Biol Chem 269:12932–12939

Novacek MJ (1992) Mammalian phylogeny: shaking the tree. Nature 356:121–125

Ohmori H, Dohrman AF, Gallup M, Tsuda T, Kai H, Gum JR, Kim YS, Basbaum CB (1994) Molecular cloning of the amino-terminal region of a rat MUC2 mucin gene homologue. J Biol Chem 269: 17833–17840

Pigny P, Guyonnet Dupérat V, Hill A, Pratt WS, Galiègue-Zouitina S, Collyn d'Hooge M, Laine A, Van Seuningen I, Gum JR, Kim YS, Swallow DM, Aubert JP, Porchet N (1996) Human mucin genes assigned to 11p15.5: identification and organization of a cluster of genes. Genomics 38:340–352

Shekels LL, Lyftogt C, Kieliszewski M, Filie JD, Kozak CA, Ho SB (1995) Mouse gastric mucin: cloning and chromosomal localization. Biochem J 311:775–785

Turner BS, Bhaskar R, Hadzopoulou-Cladaras M, Specian RD, LaMont JT (1995) Isolation and characterisation of cDNA clones encoding pig gastric mucin. Biochem J 308:89–96

Les 3 apomucines MUC2, MUC5AC et MUC5B appartiennent à la même famille de gènes et ont leurs gènes localisés sur la même région télomérique p15 du chromosome 11 humain. Les 3 apomucines possèdent, comme nous le verrons plus loin pour MUC5B, des domaines D du pro-vWF de part et d'autre de la région riche en résidus Ser et Thr (région centrale). Il est donc plus que probable que ces gènes sont issus d'un unique gène ancestral à la suite de phénomènes de duplication de gènes.

Ces 3 mucines ont plusieurs motifs "Cys" retrouvés jusqu'à présent uniquement dans des mucines. Nos analyses phylogénétiques des séquences de ces motifs, l'ordonnancement de ces motifs et des gènes sur le chromosome 11 montrent que *MUC5AC* et *MUC5B* sont plus proches l'un de l'autre dans l'évolution que de *MUC2*. Ceci nous amène à penser que *MUC5AC* et *MUC5B* ont un gène progéniteur commun que nous avons baptisé *MUC5ACB*. Ce gène provient d'un gène ancestral commun également à *MUC2*. Aucun élément de séquence concernant *MUC6* ne permettait alors de placer ce gène dans notre arbre phylogénétique.

Les travaux sur l'organisation des régions carboxyl- et amino-terminales des gènes de mucines localisés en 11p15 permettront d'infirmer ou au contraire de confirmer notre modèle.

II LA REGION CARBOXY-TERMINALE

II.1 Organisation génomique de la région Carboxy-terminale de *MUC5B*

Par comparaison des séquences d'ADNc et des séquences génomiques déterminées à partir du sous-clonage de fragments de BEN2, nous avons déduit l'organisation génomique complète de la région en aval de l'exon central de *MUC5B*.

Les résultats sont exposés dans l'article suivant :

« **Genomic organization of the 3' region of the human mucin gene *MUC5B*** »

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 272, No. 27, Issue of July 4, pp. 16873–16883, 1997
Printed in U.S.A.

Genomic Organization of the 3′ Region of the Human Mucin Gene *MUC5B**

(Received for publication, January 28, 1997, and in revised form, April 28, 1997)

Jean-Luc Desseyn‡, Jean-Pierre Aubert‡§, Isabelle Van Seuningen‡, Nicole Porchet‡§, and Anne Laine‡¶

From the ‡Unité 377 INSERM, Place de Verdun, 59045 Lille Cedex, and §Laboratoire de Biochimie et de Biologie Moléculaire de l'Hôpital C. Huriez, Centre Hospitalier Régional et Universitaire, 59037 Lille Cedex, France

***MUC5B*, mapped clustered with *MUC6*, *MUC2*, and *MUC5AC* to chromosome 11p15.5, is a human mucin gene of which the genomic organization is being elucidated. We have recently published the sequence and the peptide organization of its huge central exon, 10,713 base pairs (bp) in length. We present here the genomic organization of its 3′ region, which encompasses 10,690 bp. The genomic sequence has been completely determined. The 3′ region of *MUC5B* is composed of 18 exons ranging in size from 32 to 781 bp, contrasting thus with the very large central exon. The sizes of the 18 introns range from 114 to 1118 bp. Some repetitive sequences were identified in four introns. The peptide deduced from the sequence of the 18 exons consists of an 808-amino acid peptide. This carboxyl-terminal region exhibits extensive sequence similarity to MUC2, MUC5AC, and von Willebrand factor, particularly the number and the positions of the cysteine residues, suggesting that this domain may be derived from a common ancestral gene. The presence in these components of a cystine knot also found in growth factors such as transforming growth factor-β is of particular interest. Moreover, one part of this peptide is identical to the 196-amino acid sequence deduced from the cDNA clone pSM2-1, which codes for a part of the high molecular weight mucin MG1 isolated from human sublingual gland. Considering the expression pattern of *MUC5B* and the origin of MG1, we can thus conclude that *MUC5B* encodes MG1.**

Mucus is the layer that covers, protects, and lubricates the luminal surfaces of epithelial respiratory, gastrointestinal, and reproductive tracts. These basic properties are due to the viscous and viscoelastic properties of mucins, the major glycoprotein components of mucus. Mucins constitute a family of high molecular mass glycoproteins synthesized by the goblet cells of the epithelia and in some cases by submucosal glands (for more complete reviews, see Refs. 1–3).

Alterations of the biosynthesis of mucins affecting the protein core and/or the carbohydrate content linked to the peptide have been observed in numerous pathological situations such as various adenomas and carcinomas, inflammatory diseases such as cystic fibrosis, asthma, chronic bronchitis, or inflammatory bowel diseases (4–7). Moreover, the hypersecretion of mucins and the presence of alternating hydrophobic and hydrophilic domains in mucins have been shown to play a central role in the pathogenesis of cholesterol gallstones (8, 9).

All apomucins contain tandemly repeated sequences rich in threonine and/or serine. Due to the high carbohydrate content, the peptide moiety of mucins has been difficult to characterize. cDNA cloning has enabled researchers to approach the study of the mucins over the past decade. Today, the membrane-associated mucin MUC1 and the secreted MUC7 are the only mucins for which the full-length cDNA and the genomic organization have been reported (10–13). Both were revealed to be, in fact, small mucins. A complete cDNA of the large secreted mucin MUC2 (14–17) has been described. Partial cDNAs have been identified for the other human mucin genes that code for secreted mucins: MUC3 (18), MUC4 (19), MUC5AC (20–24), MUC5B (25), and MUC6 (26).

Four mucin genes are mapped to 11p15.5: *MUC5AC*, *MUC5B*, *MUC2*, and *MUC6*. (26–28). Recently, we have determined that the order of the four clustered 11p15.5 human mucin genes is tel-*MUC6*/*MUC2*/*MUC5AC*/*MUC5B*-cen (29). We have also established that MUC2, MUC5AC, and MUC5B have a consensus cysteine-rich domain found twice in MUC2 (16), at least four times in MUC5AC (21, 22, 24), and seven times in MUC5B (30).

MUC5B is expressed mainly in bronchus glands and also in submaxillary glands, endocervix, gall bladder, and pancreas (31–35). The structural organization of the peptide deduced from the nucleotide sequence of the central region of *MUC5B* has been published recently (30). The single large exon of 10,713 bp,[1] containing all the tandem repeat domain, is, to our knowledge, the biggest described for a vertebrate gene. It codes for a 3570-amino acid peptide. Nineteen subdomains have been individualized. Most of the MUC5B subdomains show similarity to each other, creating four larger composite super-repeat units of 528 amino acids. Each super-repeat is made up of repeats consisting of an irregular repeat of 29 amino acids, one cysteine-rich subdomain (10 cysteine residues, 108 aa), and one unique sequence of 111 amino acid residues also rich in serine and threonine. The complete organization of the region downstream of the central region of the human *MUC5B* gene, *i.e.* its complete 3′ region, is reported in this paper; we present here the complete genomic nucleotide sequence, the exon-intron organization, and the full cDNA sequence coding for the carboxyl-

* This work was supported by La Ligue contre le Cancer, L'Association de Recherche contre le Cancer, and L'Association François Aupetit. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) Y10080 and Y09788.

¶ To whom correspondence should be addressed: INSERM U-377, 59045 Lille Cedex, France. Tel.: 33-3-20-29-88-59; Fax: 33-3-20-53-85-62; E-mail: laine@lille.inserm.fr.

[1] The abbreviations used are: bp, base pair(s); aa, amino acid(s); BSM, bovine submaxillary gland mucin-like; CK, cystine knot; FIM, frog integumentary mucin; PCR, polymerase chain reaction; PSM, porcine submaxillary mucin; TGF, transforming growth factor; RACE, rapid amplification of cDNA ends; RT, reverse transcription; vWF, von Willebrand factor; ORF, open reading frame.

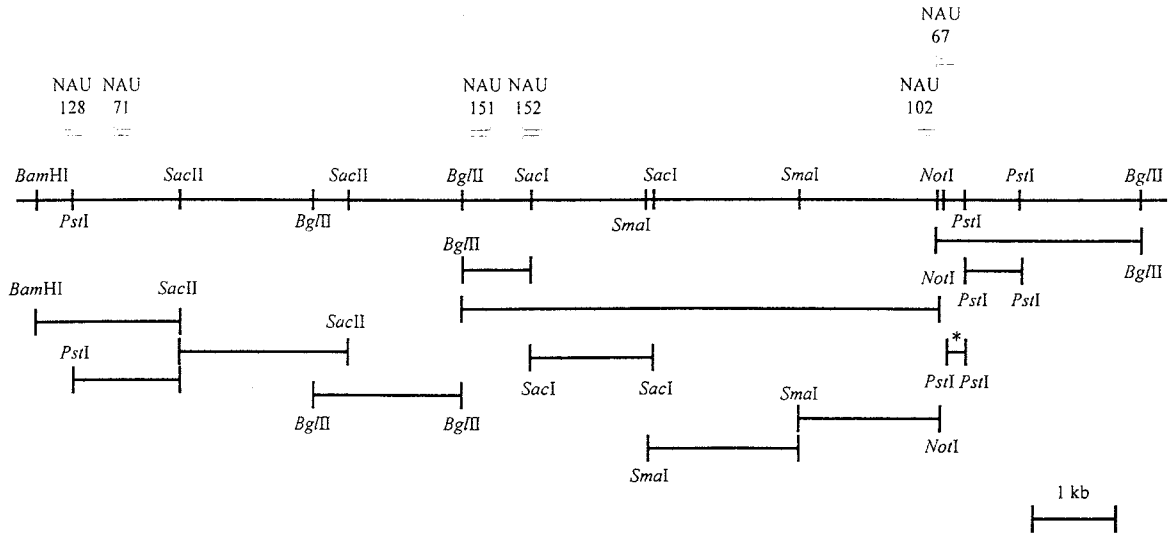


FIG. 1. **Restriction map of the 3′ region of *MUC5B* and sequencing strategy.** The fragments indicated were subcloned into pBluescript KS(+) vector. Some primers and their directions are indicated (not to scale) by *horizontal arrows* and their NAU *numbers* (their sequences are given in Table I).

terminal domain of the human MUC5B apomucin. This domain stretches 808 amino acid residues and can be divided into six subdomains. The last five cysteine-rich subdomains exhibit extensive sequence similarity to MUC2, MUC5AC, and vWF (17, 22, 23, 36), particularly the number and the positions of the cysteine residues, suggesting that this domain may be derived from a common ancestral gene. Moreover, with the exception of one substitution, which does not change the coded amino acid, one part of the cDNA sequence we determined is identical to the nucleotide sequence of pSM2-1. This cDNA codes for 196 amino acids in the carboxyl-terminal region of the high molecular weight mucin MG1 isolated from human sublingual gland (37). Considering the expression pattern of MUC5B (31) and the origin of MG1, we can thus conclude that *MUC5B* encodes MG1.

EXPERIMENTAL PROCEDURES

Screening of cDNA and Genomic Libraries—A λgt11 cDNA library constructed from human tracheal mucosa was screened with rabbit antibodies raised to deglycosylated Pronase glycopeptides from bronchial mucins (38). Among the various positive clones obtained, the one designated TH71 and containing a poly(A) tail was of particular interest in the present study.

A human genomic λEMBL4 phage library was screened using hybridization with the JER57 probe (25). One positive clone, CEL5, was isolated and studied.

A human placental genomic DNA library in pWE15 cosmid provided by Stratagene was screened using the JER57 probe. Two positive clones, BEN1 and BEN2, were obtained (30). BEN2 was the useful clone in the present study.

Oligonucleotide Primers—Oligonucleotide primers used in PCR, RACE-PCR, RT-PCR, and sequencing experiments were synthesized by Eurogentec (Liège, Belgium). Their sequences and locations are indicated in Table I.

3′-End Amplification of cDNAs—The 5′-AmpliFINDER RACE kit (CLONTECH) was used to synthesize first-strand cDNA from human trachea poly(A)$^+$ RNA (1 μg) obtained from CLONTECH using NAU61 as a primer (Table I), followed by ligation of the 5′-ANCHOR adapter. The PCR was then performed using the nested primer NAU67 (Table I and Fig. 1) and the 5′-ANCHOR primer. Nested PCRs involving a second or third round amplification were carried out with 1 μl of the reaction mixture obtained from each previous round of PCR as template.

RT-PCR Amplification—Total RNA of human gall bladder was extracted as described previously (39). Single-stranded cDNA was performed using the 1st STRAND Synthesis kit (CLONTECH), random hexamers and human trachea poly(A)$^+$ mRNA (0.5 μg) (CLONTECH) or total gall bladder RNA (1 μg). PCR amplification reaction mixtures (50 μl) contain 0.3 mM dNTPs, 2.5 units of *Taq* DNA polymerase (Boehringer Mannheim), 15 pmol of the appropriate primers, the buffer system purchased with *Taq* DNA polymerase, and an aliquot of cDNA. The PCR was performed using a Perkin-Elmer Thermal Cycler 480. PCR parameters were 94 °C for 2 min, followed by 30 cycles at 94 °C for 30 s, 60 °C for 1 min, and 72 °C for 2 min, followed by a final extension at 72 °C for 15 min. The amplified products were electrophoresed on a 1% Seaplaque gel (FMC, Rockland, ME) and stained with ethidium bromide. The band was cut out, purified using Preps DNA purification resin (Promega), and subcloned into the T/A cloning vector, pMOS*Blue* T-vector (Amersham). Thereafter, cDNA clones were subcloned into pBluescript KS(+) vector (Stratagene) using the restriction enzymes (Boehringer Mannheim) *Pst*I, *Sac*I, and/or *Sma*I. Subclones were sequenced as described below using either universal primers or a series of oligonucleotides specific for both strands of the inserts (Table I).

Isolation and Sequencing of MUC5B Genomic Clones and Sequence Analyses—Fragments of the genomic clones CEL5 and BEN2 corresponding to the region downstream of the central exon were subcloned into pBluescript KS(+) vector as described previously (30). The double-stranded plasmid inserts were sequenced manually using the dideoxynucleotide chain termination method (40) using [α-^{35}S]dATP (Amersham) and Sequenase 2.0 (U. S. Biochemical Corp.) according to the protocol indicated by the manufacturer. Universal primers or a series of specific oligonucleotides were used. Sequencing reaction mixtures were electrophoresed on 6% polyacrylamide gel (Sequagel-6™, National Diagnostics). The clones were sequenced on both strands several times. Direct DNA sequencing on cosmid was performed as described previously (30). Computer analyses were performed using PC/GENE Software. The whole genomic sequence reported in this paper has been submitted to the EMBL Data Bank with accession number Y09788. The sequence of TH71 has been submitted to the EMBL Data Bank with accession number Y10080.

Study of the Intron G—To determine the exact number of repeats in the intron G, first we cut the genomic subcloned *Bgl*II-*Bgl*II fragment using *Sac*I and *Rsa*I that flank the region containing these direct 59-bp repeats. The complete digestion with *Sma*I was obtained using 10 units/μg DNA for 3 h. The partial digestions were performed using 1 unit/μg DNA and 0.25 unit/μg DNA for 1 h. After electrophoretic separation on 1.5% agarose gel, the blot analysis was conducted using the antisense oligonucleotide NAU199 (5′- AGAGCCGAGGGGTCTGGG-3′), which had been previously radiolabeled using T4 polynucleotide kinase (Boehringer Mannheim) and [γ-^{32}P]ATP from Amersham.

RESULTS AND DISCUSSION

Characterization of the Genomic Fragments of the 3′ Region of MUC5B—The partial restriction maps of the genomic clones CEL5 and BEN2 were determined. Their overlapping parts present the same restriction map. The partial restriction map of the 3′ region of *MUC5B* is shown in Fig. 1 together with the

TABLE I
Primers used for cDNA synthesis and DNA and cDNA sequencing

Primer designation	Primer sequence (5′ to 3′)	Position	Orientation[a]
NAU61	ACTCAATGCTCAGGGTTTATTTGC	10582–10605	AS
NAU67	GGGTTTATTTGCAAAACTG	10575–10593	AS
NAU71	AGTGCTGATTGCACACTGCGT	838–859	AS
NAU102	CCTGTCGCAGCTTCCTGGCAG	10446–10466	AS
NAU106	CAGTGAGCATAGGGGAAGCCT	3387–3407	S
NAU127	AGGCTTCCCCTATGCTCACTG	3387–3407	AS
NAU128	CGTGTCCACTGTGTCCTCCTCAGTC	1–25	S
NAU140	GATGGCGGAGGGCTGCTTCTG	5139–5159	S
NAU141	CAGACCGTGTGCACGCAGCAC	1001–1021	AS
NAU142	CCAGGGTAGGACTCCTGAGTG	10246–10266	AS
NAU151	TGAGCAGCGGTTTCAGCAAGA	3168–3188	S
NAU152	CAAGGTTGTGGCACTCAGCAA	3837–3857	AS
NAU196	CGAGGGTTCAGTGTCGGTG	6013–6031	S
NAU200	CAGTGTCCTTACCGGGAGA	2221–2239	AS
NAU203	ATTTAGGAAACCCATCGGGT	5689–5708	AS
NAU207	CGCGGGGTGCCACACACAGGCC	10142–10163	AS
NAU208	GGGTGTAGGTGTGCAGGATGG	9927–9947	AS
NAU219	GCAGGGAAGGGCGCCTGGGAA	7394–7414	AS
NAU226	AGCGGAAGGTGGGACAGCAGT	6620–6640	AS
NAU227	ACTGCTGTCCCACCTTCCGCT	6620–6640	S
NAU232	CTTCCCAGGCGCCCTTCCCTGC	7393–7414	S
NAU233	CTGCGAGACCGAGGTCAACATC	9113–9134	S
NAU234	GATGTTGACCTCGGTCTCGCAG	9113–9134	AS
NAU249	CTCCTCACAGGAGTAGCAGC	8814–8832	AS
NAU277	CAGTGACTGGCGAGGTGCAACTG	3973–3995	S
NAU278	GTATGGGGCCGCATGCGTTGTACACT	4624–4649	AS
NAU280	TGGACAGATGCCCAGGGTTGA	5901–5921	S
NAU281	TGCCATTGTACGAACACAGCT	6776–6796	AS
NAU282	CTGCAGGCCCCATTGGGTCAT	7297–7317	S
NAU293	ATGAGCCGTGGATGGGGTCCC	1195–1215	S
NAU297	TCATGGTCCTGGGCGGCTCCT	5277–5297	AS

[a] Strand orientation: sense (S), antisense (AS).

overlapping fragments, which were separated and subcloned into pBluescript KS(+) vector. The fragments *Bam*HI-*Sac*II and *Pst*I-*Sac*II (in the *left part* of Fig. 1) contain the 3′ end of the central exon. All these clones were entirely sequenced after restriction digestion and subcloning. Primer walking using specific oligonucleotides (Table I) was also performed.

3′ Region of MUC5B —Several cDNA-positive clones were obtained by screening the λgt11 cDNA library using antibodies as described previously (38). The clone designated TH71 is 380 bp in length. Its sequence (Fig. 2), submitted to the EMBL data bank with accession number Y10080, revealed a poly(A) tail with 73 A, 16 bp downstream from a polyadenylation signal (AAUAAA). By sequencing the *Pst*I-*Pst*I subclone (noted with an *asterisk* in Fig. 1) obtained from the fragment *Not*I-*Bgl*II of the BEN2 clone, an identical 67-bp sequence was observed (Fig. 2), up to the A where the poly(A) addition occurs, indicating that the clone BEN2 contains the 3′ end of the *MUC5B* gene. Using the two synthesized oligonucleotides NAU61 and NAU67 chosen in this sequence, a 5′-RACE-PCR experiment was performed. After cloning of the fragment obtained, the insert of 88 bp designated RACE67 was sequenced. This sequence is identical to the 88-bp sequence determined in the *Pst*I-*Pst*I clone (Fig. 2). In contrast, the first 34 nucleotides differ from the sequence of TH71. The TH71 clone, which has been found using the antibodies directed against the repeat part of the MUC5B apomucin (38), begins with a 132-bp sequence we found in the central exon. Between this sequence and the 3′ end identical to the RACE67, TH71 seems to have been rearranged; moreover, the following results show that an important part of the cDNA has been lost. We will discuss these data below.

The *Not*I-*Bgl*II fragment from BEN2 contains two other clustered canonical polyadenylation signals, AATAAA. The first was located about 2 kilobase pairs downstream from the first polyadenylation signal and the second 298 base pairs downstream from this latter AATAAA. The significance of these two additional polyadenylation signals is not known. It will be interesting to determine if several forms of *MUC5B* mRNA can be transcribed by selection of alternative polyadenylation signals.

The dinucleotides TG and GT were found with oligo(T) stretches in the region downstream from the first AATAAA motif within the *Pst*I-*Pst*I subclone. This region, referred to as "GT cluster," is important for 3′ processing of polyadenylated mRNAs (41). Moreover, the pentanucleotide CATTG was found between the AATAAA sequence and the poly(A) site addition (Fig. 3). This CAYTG recognition element has been described to be related to cleavage site selection by Berget (42). The author suggested that pre-polyadenylated RNA hybridized with the AAUAAA recognition element as related to primary site selection, and with CAYUG recognition element within the U4 small nuclear ribonucleoproteins as related to cleavage site selection. Hence, *MUC5B* combines some common features of the 3′ mRNA processing. From this nucleotide sequence, the new oligonucleotide NAU102 was synthesized to perform RT-PCR.

RT-PCR —Two specific overlapping cDNAs were synthesized by RT-PCR experiments. The locations of the oligonucleotides used in these experiments are indicated in Fig. 1. The oligonucleotide primer NAU151 was designed on the basis of the sequence determined for the *Bgl*II-*Bgl*II cosmid fragment (Fig. 1). This fragment hybridized with human tracheal RNA on Northern blot and probably contains coding sequences. An amplification product was obtained when the RT-PCR was performed with the two primers NAU151 and NAU102 using human tracheal first-strand cDNA as template. It was designated RT151-102 and is 2209 nucleotides in length. An other RT-PCR was then performed with the following oligonucleotide primers: NAU152, designed with the sequence of RT151-102, and NAU128, chosen in the 3′ end sequence of the *MUC5B* central exon (30). The resultant 1166-bp amplification product, called RT128-152, and the RT151-102 were cloned into pMOS-

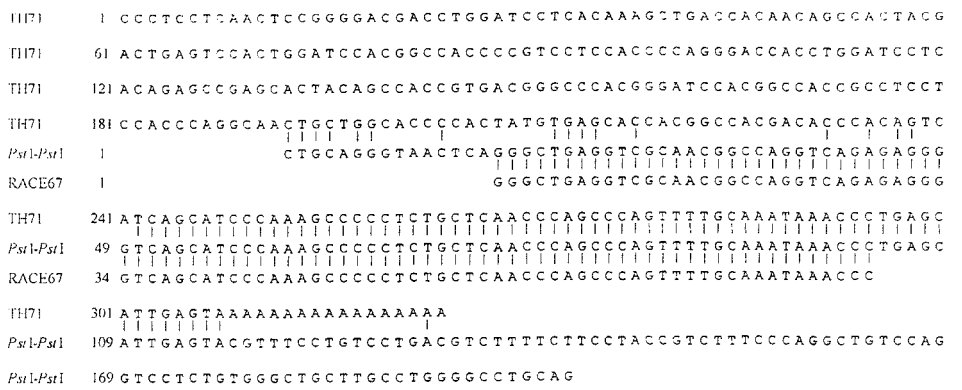


FIG. 2. **Alignments of the sequences, genomic or cDNA, containing the polyadenylation signal of *MUC5B*.** The cDNA clone called TH71 was aligned with the sequence of the genomic *Pst*I-*Pst*I fragment (with an *asterisk* in the *right part* of Fig. 1), and the clone obtained by performing RACE-PCR called RACE67.

Blue T-vector. They were subsequently subcloned into pBluescript KS(+) vector after cutting with the restriction enzymes *Pst*I, *Sac*I, and/or *Sma*I. The subclones were entirely sequenced on both strands several times using T3 and T7 primers and specific oligonucleotides (see in Table I). The two amplification products RT128-152 and RT151-102 have overlapping sequences of 416 nucleotides.

Sequencing Data and Genomic Organization—The 3′ region of the human *MUC5B* gene shown in Fig. 3 encompasses 10,690 bp, of which the first 113 nucleotides correspond to the 3′ end of the central exon we recently published (30). The full-length sequence has been submitted to the EMBL data bank with accession number Y09788.

The 3′ region of *MUC5B* gene is composed of 18 exons ranging in size from 32 to 781 bp (Table II) in good agreement with the mean length of exons (43), in contrast to the extraordinary large central exon of *MUC5B* (30). The last exon is the largest one. It codes for the 72-amino acid COOH terminus of the core protein and comprises the 3′-untranslated region, 564 bp in length, of the *MUC5B* gene. The sizes of the 18 introns range from 114 bp to 1118 bp. Each intron begins with a GT and ends with an AG (Table II), obeying strictly the GT/AG rule of splice-junction sequences proposed by Mount (44).

Sequencing Data and Amino Acid Analyses—The 2423 nucleotides open reading frame (Fig. 3) encodes a 808-amino acid peptide rich in cysteine (10.1%) and proline (9.5%). This region is relatively poor in threonine and serine (8.3 and 7.3%, respectively). It is thus different from a mucin-like domain. The comparison of a part of the deduced peptide sequence of RT151-102 (aa 634–829 in Fig. 3) with the deduced amino acid sequence of the cDNA clone pSM2-1 (37) shows 100% identity. In nucleotide sequence only one codon differs, since the proline in position 769 in our sequence (Fig. 3) is coded by CCC instead of CCG in the sequence of Troxler *et al.* (37). Consequently, this suggests that pSM2-1 is a part of the *MUC5B* gene. The pSM2-1 clone was isolated from a human sublingual gland cDNA library, screened with a polyclonal antiserum against deglycosylated MG1, the high molecular weight mucin from human sublingual gland. MG1 is a candidate, among other roles, for participation in enamel pellicle formation (45). MG1 is made up of multiple disulfide-linked subunits and contains numerous hydrophobic binding sites in naked regions with negatively charged amino acid residues (46). These characteristics are in very good agreement with our data, since such regions do exist in the 3570-amino acid peptide encoded by the central exon of *MUC5B* (30). Seven nonadjacent domains, termed Cys subdomains, have been individualized among the 19 subdomains encoded by the central exon. These Cys subdomains, found in several other apomucins, are richer in Cys (9.3%), Asp (4.9%), and Glu (7.7%) than the 12 other subdomains. The Cys subdomains are poor in Ser and Thr (Ser+Thr: 9.6%) *versus* tandem repeat domains, termed R domains (Ser+Thr: 52.5%). Moreover, Loomis *et al.* (46) suggested that aromatic residues of MG1 are buried within the hydrophobic domains. In fact, the Tyr and Trp amino acid residues are strikingly clustered in the Cys subdomains for the MUC5B apomucin. What is more, the MG1 glycoprotein and *MUC5B* mRNAs are both expressed in salivary glands among other mucosa for *MUC5B* (31, 37). We can thus conclude that *MUC5B* encodes the MG1 apomucin.

The deduced amino acid sequence of the carboxyl-terminal region of MUC5B contains 15 consensus sequences for attachment of *N*-linked oligosaccharides (*italic* in Fig. 4). Studies were performed using the computer PC Gene software (47). The secondary structure of the carboxyl-terminal region of MUC5B was predicted to contain 62% β turn conformation and 13% helix structure located between aa 219 and 250, 407 and 421, 776 and 801. The rest of the structure consists of extended and coil structures. The rigid conformation could be essential for the oligomerization process. In fact 88% of the cysteine residues are located in or near a β turn as well as 11 out of the 15 potential *N*-glycosylation sites. Moreover, a serumalbumin family signature with the consensus sequence YX_6CCX_7C has also been found between residues 658 and 682. As mucins are well known to bind various hydrophobic substances such as cholesterol, fatty acids, or bilirubin, this small region could be important in the formation of gallstones for example in which mucins have been described to be involved (8, 9, 48, 49).

As far as TH71 is concerned, we can now evaluate that more than 2000 bp have been lost in this last cDNA. We were unable to reproduce this cDNA using RT-PCR. It may be concluded that there has been a problem when producing this clone, which has been otherwise of great interest in determining the location of the polyadenylation signal in the genomic DNA.

Deduced Amino Acid Sequence of MUC5B Carboxyl-terminal Region: Comparison with Other Proteins—Some partial alignments with the sequence of vWF have been made by other authors, for example for MUC2 (16) and for MUC5AC-related cDNA clones, like NP3a (22) and L31 (23). Fig. 4 shows that an alignment on longer sequences can be accomplished. The conservation of nearly all of the cysteine residues and of several other amino acids of vWF and of the three 11p15.5 human mucins MUC2, MUC5AC, and MUC5B is readily apparent, suggesting a very similar tertiary structure. The comparison of the deduced carboxyl-terminal MUC5B peptide with vWF especially allows us to dissect this region into six domains: one domain called MUC11p15-type, which follows the central exon, one 56-amino acid domain with similarities to what we called

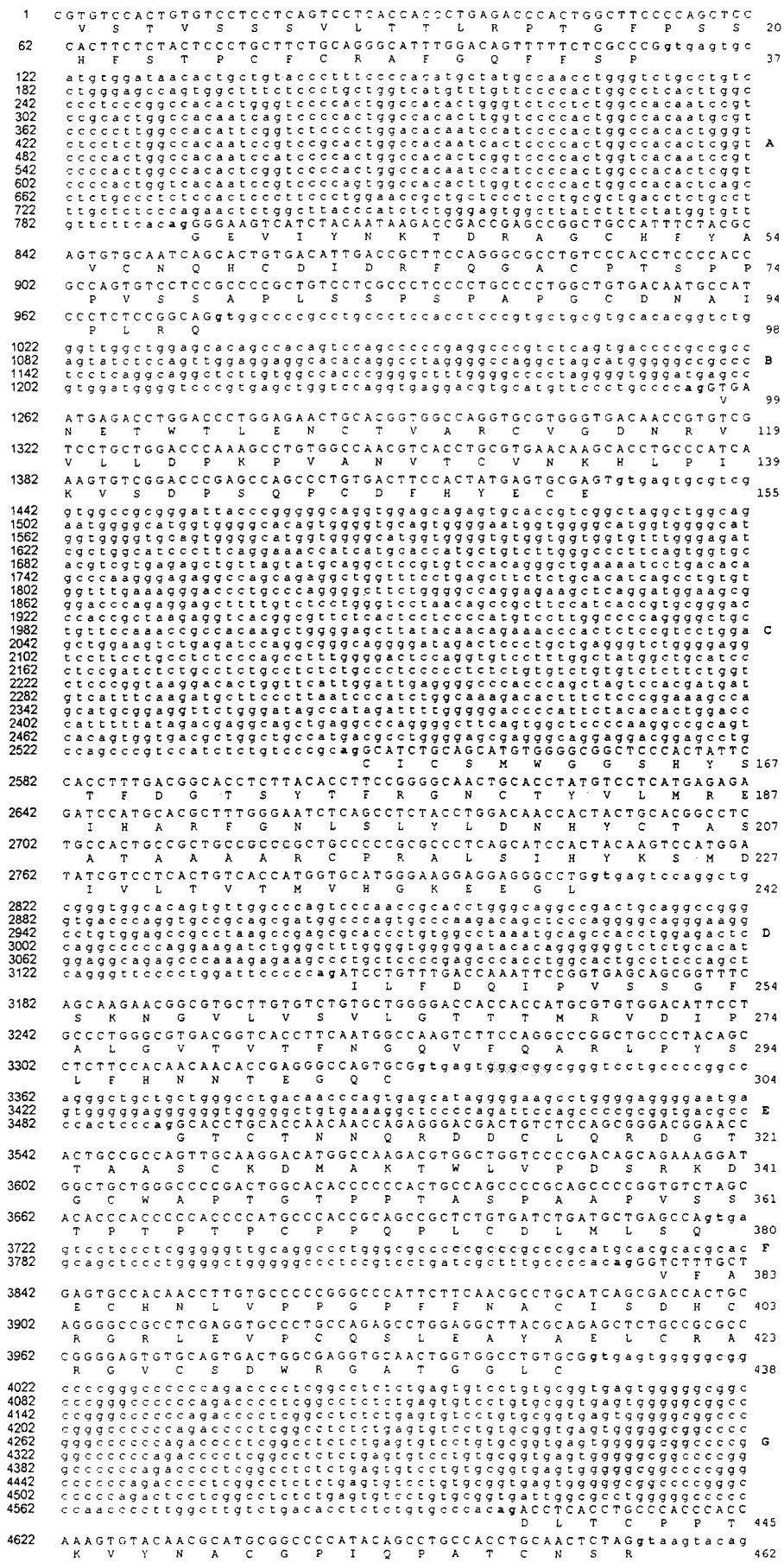


FIG. 3. **Sequence of the 3′ region of the *MUC5B* gene.** The sequence shows the entire region beginning at the sequence of the oligonucleotide NAU128 used for the RT-PCR. The sequences of exons are indicated in *uppercase letters*, and the sequences of introns are indicated in *lowercase letters*. Encoded amino acids are shown in *single-letter code* and *numbered 1–845* in the *right margin*. *Bold letters* on the *right* indicate the names of the introns. GC boxes are *shaded*. The polyadenylation signal is *bold boxed*. GT stretches and the pentanucleotide CATTG that may be involved in processing or polyadenylation are *underlined*. The position of the signal of poly(A) addition is marked by an *arrow*. The splice donor and acceptor sites are in *bold*.

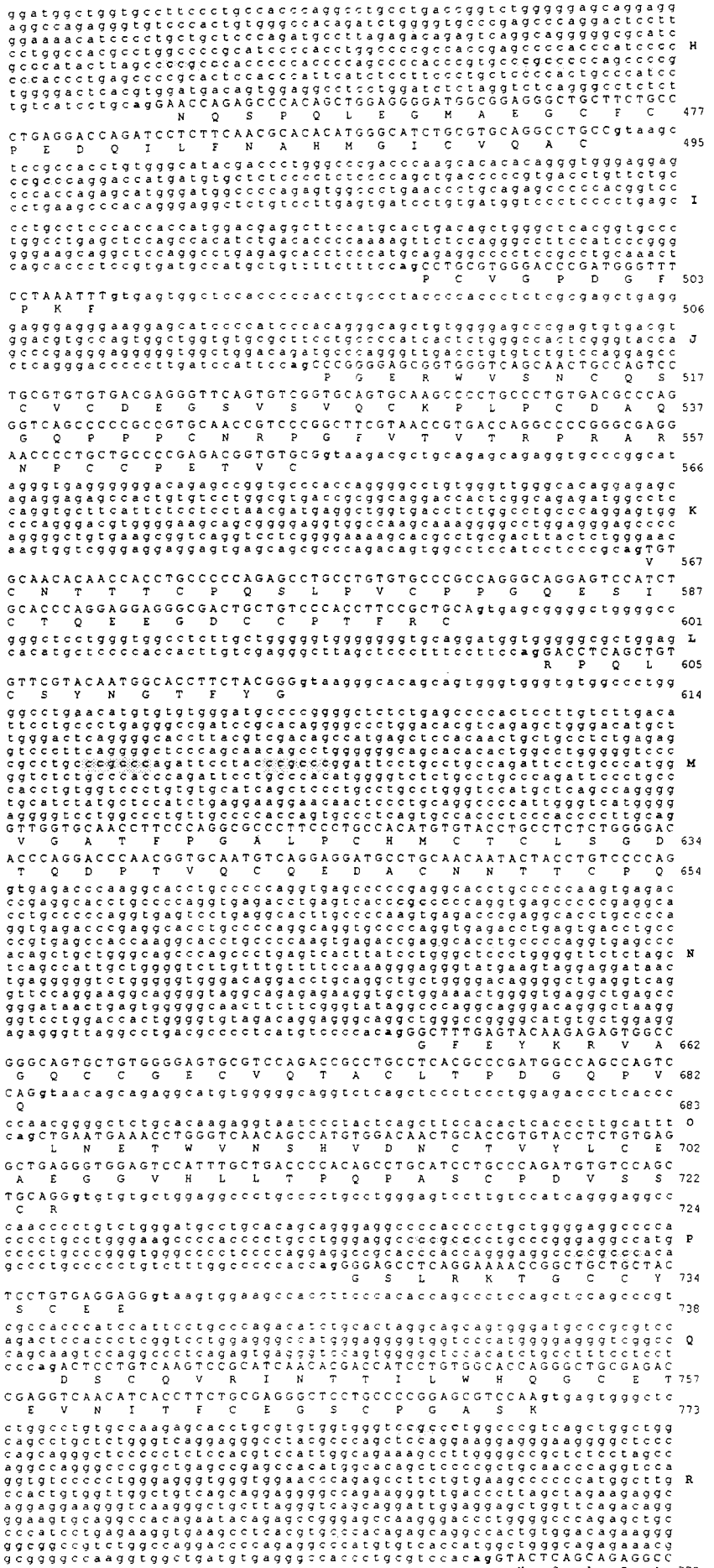


FIG. 3—*continued*

FIG. 3—*continued*

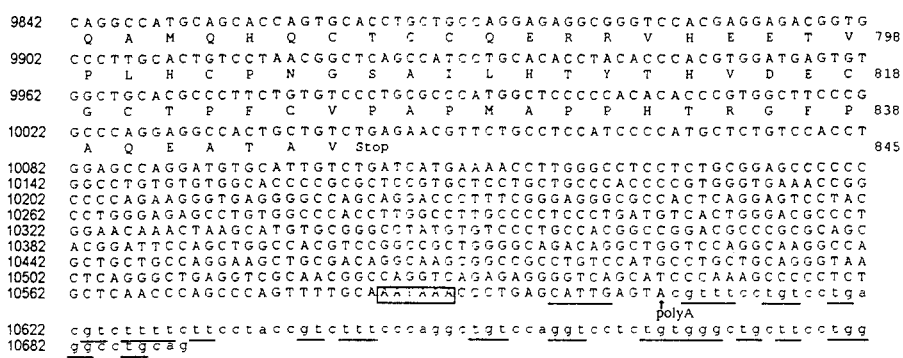


TABLE II

Characteristics of the exon-intron junctions of the 3′ region of the MUC5B gene

Protein domain	Exon		5′ Splice donor	Intron			3′ Splice acceptor
	No.	Size		Name	Size	Class	
		bp			*bp*		
Central	0	10,713	CGCCCG**gt**gagtgcatgtgga	A	679	1	tgttgttcttcac**ag**GGGAAG
MUC11p15	1	182	CGGCAG**gt**ggccccgcctgcc	B	283	0	tgttccctgcccc**ag**GTGAAT
A3uD4	2	172	GCGAGT**gt**gagtgcgtcggtg	C	1118	1	atctctgtcccgc**ag**GCATCT
D4-like	3	260	GGCCTG**gt**gagtccaggctgc	D	338	0	cctggattcccccc**ag**ATCCTG
D4-like	4	187	AGTGCG**gt**gagtgggcggcgg	E	159	1	acgccccactccc**ag**GCACCT
D4-like	5	226	GAGCCA**gt**gagtcctccctcg	F	114	2	cgctttgccccac**ag**GGTCTT
D4-like	6	176	TGTGCG**gt**gagtgggggcggc	G	594	1	tctctgtgcccac**ag**ACCTCA
D4-like	7	70	CTCTAG**gt**aagtacagggatg	H	443	2	cttgtcatcctgc**ag**GAACCA
D4-like	8	101	CCTGCC**gt**aagctccgccacc	I	463	1	ctgttttctttcc**ag**CCTGCG
D4-like	9	32	AAATTT**gt**gagtggctccacc	J	258	0	cttgatccattcc**ag**CCCGGG
D4+B-like	10	181	TGTGCG**gt**aagacgctgcaga	K	389	1	tccatcctcccgc**ag**TGTGCA
B+C-like	11	105	GCTGCA**gt**gagcggggctggg	L	126	1	tccctttccttcc**ag**GACCTC
C-like	12	38	TACGGG**gt**aagggcacagcag	M	574	0	tcccacccctt**gc**agGTTGGT
C-like	13	120	CCCCAG**gt**gagacccaaggca	N	696	0	ctcatgtccccac**ag**GGCTTT
C-like	14	87	GTCCAG**gt**gtaacagcaggca	O	120	0	ccttgcatttcag**ag**CTGAAT
C-like	15	123	TGCAGG**gt**gtgtgctggaggc	P	264	0	tttggcccccacc**ag**GGGAGC
C-like	16	43	AGGAGG**gt**aagtggaagccac	Q	232	1	ctttcctcctccc**ag**ACTCCT
C-like	17	103	GTCCAA**gt**gagtgggctcctg	R	656	2	accctgcgtccac**ag**GTACTC
CK+3′-UTR	18	781					
Consensus (44)			C a AG**gt** agt A g				ccccccccccc c n **ag** ttttttttttt t

the A3uD4 domain (located between the A3 and the D4 domains in vWF), one D4-like domain, one B-like domain, one C-like domain, and one CK domain of 86 amino acid residues (see Fig. 5*A*).

The domain called MUC11p15-type (Fig. 5*A*) from aa 38 to 84 in MUC5B follows the central domain described previously (30). It shows similarities to MUC2 and MUC5AC, particularly with regard to the cysteine residues. This domain is somewhat different from the A3 domain of vWF.

The second domain, called A3uD4, also present in vWF, is 69 aa in length in MUC5B and spans aa 85–154. This domain is also found in MUC2 and L31, *i.e.* MUC5AC.

The vWF-D4 domain was found in MUC5AC, MUC2, and MUC5B (aa 155–533). D4 was found in zonadhesin, a sperm membrane protein that binds in a specific manner to the egg extracellular matrix from pig (50). Moreover, the D4 domain shows similarity to a part of vitellogenin found in nematode *Caenorhabditis elegans*, chicken, and frog (51). Among the well conserved peptide sequences, of which the positions are indicated for MUC5B, NC(S/T)YVL (aa 180–185), TXGXCGXC (aa 300–307), YAXLC (aa 417–421), CXDWR (aa 427–431), EGCFCP (aa 473–478) found in MUC2, MUC5AC, vWF and MUC5B, the TXGXCGXC octapeptide contains the vicinal cysteine residue motif CGXC, which also exists in vitellogenin and zonadhesin. Mayadas *et al.* (52) showed that these motifs are similar to the amino acid sequences at the active site of disulfide isomerases that catalyze thiol protein disulfide interchange. These vicinal cysteines may have the capacity to catalyze disulfide interchain formation, but Voorberg *et al.* (53) indicated that the dimerization resides in the last 151 residues for the vWF. Recently Perez-Vilar *et al.* (54) validated this hypothesis for PSM, showing that this apomucin can very likely form dimers between its carboxyl-terminal domains. Hence, our sequencing data suggest that the MUC5B apomucin is also able to form dimers between its carboxyl-terminal domains. The presence of disulfide-linked subunits was already predicted by Loomis *et al.* (46) in MG1. Moreover, Kawagishi *et al.* reported that MG1 contains at least two subunits (55), one of which is the salivary link component that weakly cross-reacts with antiserum to the human small intestinal link component, which contains *N*-linked carbohydrate (56).

Following the D4-like domain, one B-like domain of 40 aa residues was found in MUC5B (aa 534–574), MUC2, and MUC5AC instead of the three B domains defined for vWF (36). This B-like domain has not yet been found in other protein sequences.

Instead of the two C domains (C1 and C2) present in vWF, only one C-like domain was found in the three 11p15.5 mucins MUC5B, MUC5AC, and MUC2 (aa 575–740 in MUC5B). One C-like domain was previously reported in the frog integumentary mucin FIM-B.1 (57). This C-like domain has been described to be more related to the C1 than to the C2 domain. In contrast to vWF, in which the homologous C1 and C2 domains arose by duplication, this duplication did not occur in FIM-B.1

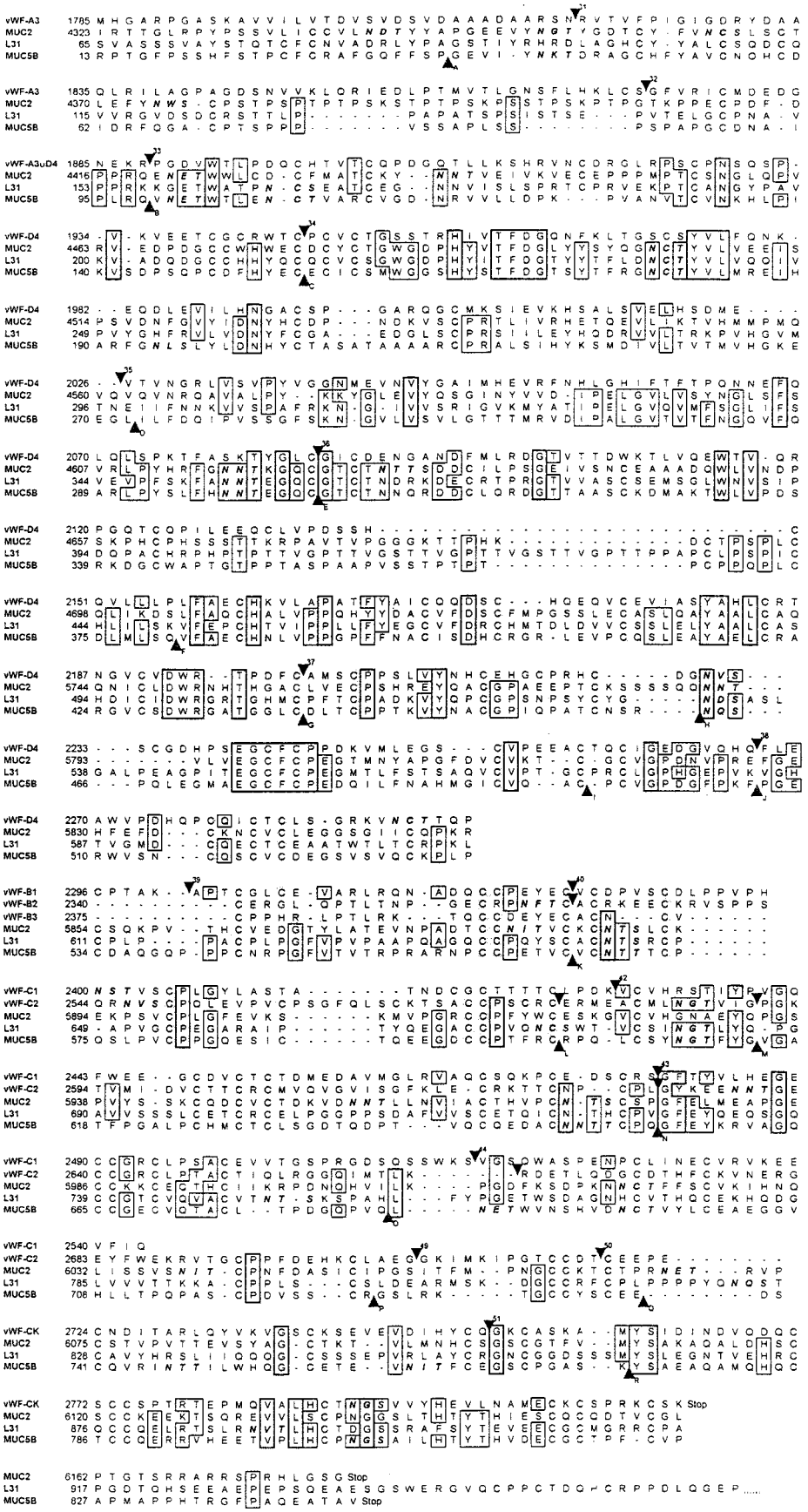


FIG. 4. **Comparison of MUC5B carboxyl-terminal protein with other proteins: vWF (36), MUC2 (16), and L31 (23).** *Dashes* indicate gaps introduced in the sequence for alignment purposes. Cysteine residues are *shaded*. Potential *N*-glycosylation sites are indicated by *bold italic letters*. *Black down-pointing arrowheads* indicate the positions of the introns of human vWF and are named according to Mancuso *et al.* (36). *Black up-pointing arrowheads* indicate the positions of *MUC5B* introns and are marked with the intron letter according to the Table II. Identical amino acids in vWF, MUC2, L31 (MUC5AC-3′ end), and MUC5B are *bold boxed*. *Thin boxes* indicate identical amino acids in at least three proteins.

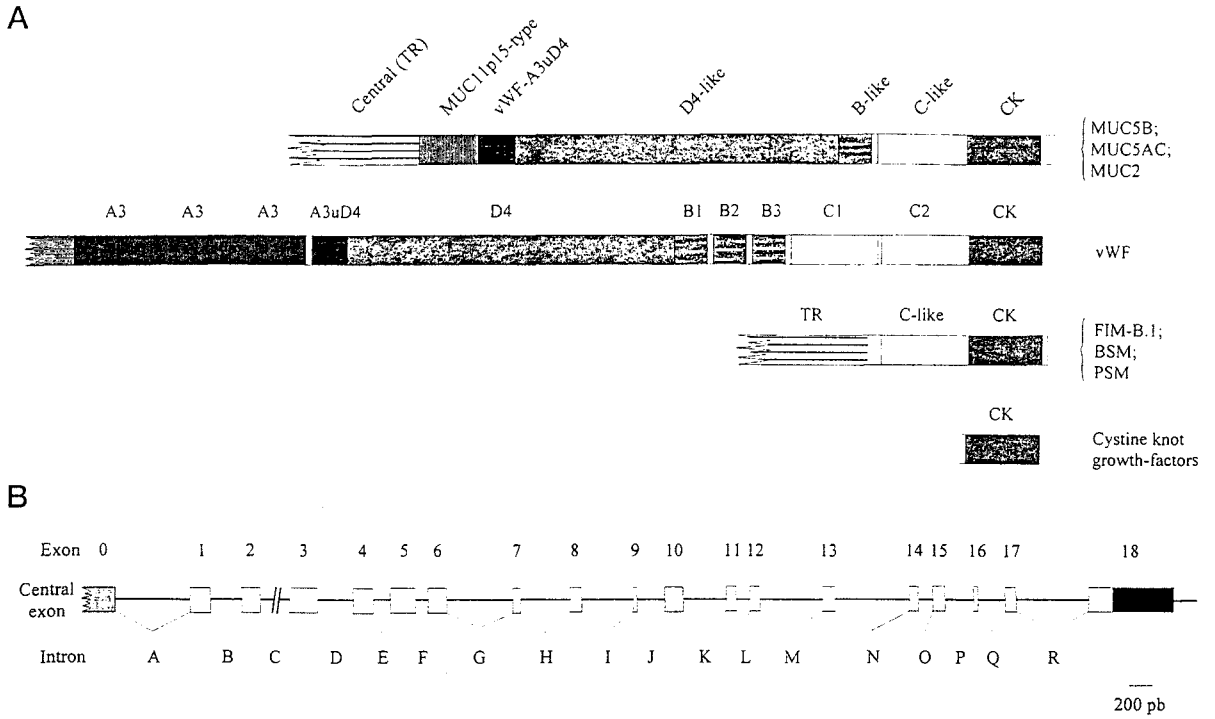


FIG. 5. **Organization of the 3′ region of MUC5B.** *A*, schematic diagram of the carboxyl-terminal MUC5B core protein and similar proteins. *B*, organization of exons and introns in the 3′ region of *MUC5B*. Exons are indicated by *open boxes* and *numbered* consecutively with 0 for the central exon (*light gray box*). *Black box* indicates the 3′-untranslated region. Introns as well as 3′-flanking sequence are indicated by *lines*. Each intron is named with a *letter* according to Table II.

(57). In fact, our genomic study shows that the C-like domain of MUC5B is more related to the C2 domain. Extremely conserved intron positions can be shown for introns 45 and L, and for introns 46 and M. Thus C1 and C2 in vWF, and C-like domains in 11p15.5 mucin genes, probably have a common ancestor domain, which has duplicated into C1 and C2 in vWF.

The last domain found in MUC5B is the CK domain (for cystine knot) from aa 741 to 826. The CK domain was also found in the 3′ end of the secreted proteins MUC2, MUC5AC, FIM-B.1, and rat-Muc2 (58, 59). The CK domain exists in other secreted proteins (60–62). Eleven cysteine residues and some other amino acid residues within this CK domain are nearly invariant. Molecular modeling of the Norrie disease protein (61) predicts that this domain has a tertiary structure similar to that of transforming growth factor β (TGF-β). In TGF-β, seven cysteine residues, corresponding to cysteines 741, 764, 768, 787, 788, 818, and 820 in MUC5B, are nearly invariant (61). Crystallography studies of TGF-β2 have shown that six of these cysteine residues are closely grouped to make a rigid structure called the cystine knot (for reviews see 60, 63). Moreover, the determination of the crystal structure of dimeric nerve growth factor and platelet-derived growth factor revealed a structure similar to the one of TGF-β2. Tertiary structure similarities probably account for a strong resistance, as suggested by these authors, to heat, denaturants, and extremes of pH. The remaining cysteine residue in each monomer, corresponding in MUC5B to Cys-787, forms an additional disulfide bond that was found to link two TGF-β2 monomers into a dimer. Hence, the human mucins MUC5B, MUC5AC, and MUC2, the animal mucins PSM, BSM, rat-Muc2, and FIM-B.1 and Norrie disease protein may be members, with their 11 cysteine residues, of a new CK subfamily.

Out of the 15 consensus sequences for attachment of *N*-linked oligosaccharides (*italic* in Fig. 4), 10 sites are close to those observed in the 3′ ends of MUC2, and/or MUC5AC (L31) and/or vWF. Four of them (positions aa 179, 298, 299, and 569 in MUC5B) have the same positions in the deduced peptides of the three human mucin genes mapped on chromosome 11p15.5. One site has the same position in the three mucins and in vWF (aa 2223 in vWF and aa 463 in MUC5B). In addition to the typical and expected *O*-glycosylation that occurs in MUC5B, it is very tempting to speculate that this apomucin, synthesized in the endoplasmic reticulum, is rapidly *N*-glycosylated. The polypeptide might fold to form intramolecular interactions and then dimers through intermolecular disulfide bridges within the carboxyl-terminal region. Although previous studies on bovine vWF suggested that *N*-glycosylation is not necessary for dimerization (64), Wagner *et al.* (65) more recently reported that *N*-linked carbohydrate addition onto human vWF is important for dimerization. In contrast, Perez-Vilar *et al.* (54) demonstrated that PSM dimerization is not dependent on the *N*-linked oligosaccharides within its carboxyl-terminal domain. Further studies on MUC5B apomucin using recombinant proteins synthesis to obtain antibodies and culture of mucus-secreting cells such as HT29-MTX in presence of tunicamycin will be required to clarify the role of *N*-glycosylation.

Sequence Analyses of the Introns—The schematic organization of the carboxyl-terminal *MUC5B* gene is given in Fig. 5*B*. No alternative splicing was found using total RNA from gall bladder or poly(A)$^+$ mRNA from trachea and the following pairs of primers: NAU128/NAU152, NAU151/NAU203, NAU140/NAU219, NAU227/NAU208, and NAU233/NAU67 (Table I).

Introns A, C, E, G, I, K, L, and Q are class 1, where each intron interrupts the coding sequence between the first and second bases of the codon (66). Introns F, H, and R are class 2 (the intron interrupts the second and the third bases of the codon), and introns B, D, J, M, N, O, and P are class 0 (the intron occurs between codons). Introns B, C, E, G, J, K, L, M, and N have the same positions as the introns 33, 34, 36, 37, 38, 40/41, 45, 46, or 43/47, respectively, in vWF (Fig. 4). It must be emphasized that introns C and 34, E and 36, G and 37, K and

40 or 41, L and 45 are class 1, while introns J and 38, M and 46, and N and 43 or 47 are class 0. We can then observe that the ORFs between the symmetrical introns C and E, between E and G, between G and K, between K and L, between J and M and between M and N have flanking introns of the same phase class at both their ends. Consequently these ORFs are good candidates for exon shuffling especially when both ORF flanking introns are class 1 (67). It would be interesting to determine if *MUC2*, *MUC5AC*, and *MUC5B* have a common 3′ end gene organization. Then it would be proposed that exon shuffling mechanisms may have played an important role in the formation of genes coding for proteins with D/B/C/CK domains, while a single ancestral gene may have evolved by successive duplications to give rise to the 11p15.5 human mucin gene family. Clearly, much work remains to be done and new data have to be collected concerning the three other 11p15.5 mucin genes *MUC2*, *MUC5AC*, and *MUC6* to confirm our hypothesis; in particular, the exon-intron repartitions have to be elucidated.

In some introns, unique tandemly repeated sequences that are more or less perfect are found: 23 copies of an imperfect 20-bp repeat in intron A, 11 copies of an imperfect 10-bp repeat in intron C, 9 copies of a perfect 59-bp repeat in intron G, and 12 copies of an imperfect 20-bp repeat in intron P. Searching of the GenBank data base indicated that the consensus sequences of these four distinct repeats were not identical with any registered sequence. It is striking that intron G is 75% (G+C)-rich and it is almost entirely built up of copies of a perfect 59-bp repeat, CCTGTGCGGTGAGTGGGGGCGGCCCCGGGCCCCCCAGACCCCTCGGCCTCTCTGAGTGT. Each repeat contains one GC box binding site and one *Sma*I enzyme recognition site. The first copy of this repeat begins in the 3′ end of the exon 6. To determine the exact number of repeats in this intron, first we cut the genomic subcloned *Bgl*II-*Bgl*II fragment using *Sac*I and *Rsa*I enzymes that flank the region containing these perfect 59-bp repeats. Then we performed a complete and two partial restriction digestions with *Sma*I (for details see "Experimental Procedures"). NAU199 is an oligonucleotide that recognizes a part of the 59-bp repeat. The results shown in Fig. 6 led us to conclude that there are nine 59-bp repeats. In a previous study where single-stranded oligonucleotides were used, we found that a nuclear factor called NF1-MUC5B (68), extracted from the colonic mucus-secreting subclone HT29-MTX, binds this GC site. This factor, with a M_r of 42,000, has been demonstrated not to be Sp1. Biochemical studies are currently in progress in our laboratory to characterize this nuclear factor.

In summary, we have cloned and sequenced the whole genomic 3′ region of *MUC5B* and defined the exon-intron repartition. We have proved that this gene codes for the high molecular weight salivary mucin MG1. *MUC5B* is expressed essentially at high levels in acini mucous cells of salivary and respiratory submucosal glands and in epithelial cells of gall bladder, endocervix, and pancreas. This study provides the first genomic organization of the 3′ region for a large size secreted gel-forming mucin gene. Our recent work showed that the central domain of MUC5B is encoded by a single large exon (10,713 bp), the largest one described to date in vertebrates. The deduced protein contains 19 subdomains. Some of them show similarity to each other, creating repeat units called super-repeats of 528 amino acid residues, which are the biggest ever determined in mucin genes. Each super-repeat comprises a 108-amino acid cysteine-rich subdomain. This last subdomain, found seven times in MUC5B, has thus been found several times in at least three of the four human mucin genes mapped to 11p15.5. (30). It seems that 11p15.5 human mucin genes are characterized by (i) a large exon encoding the repetitive domain as demonstrated for MUC5B and as suggested by Toribara *et al.* for *MUC2* (15), (ii) the presence of Cys subdomains with 10% Cys residues (30), and (iii) a unique sequence just downstream from the repetitive domain typical of the 11p15.5 mucin genes and a cysteine-rich region, which is divided in several subdomains similar to vWF-D4, vWF-B, vWF-C, and CK domains (Fig. 4). It will be interesting to determine if the three other mucin genes *MUC2*, *MUC5AC*, and *MUC6* have the same 3′ end genomic organization as *MUC5B*. Moreover, it is clear with our previously published data (21, 29, 30) and with our present results, that *MUC5AC* and *MUC5B* are two distinct genes; therefore, it would be preferable for all authors to be precise in specifying which gene is concerned when they write *MUC5*.

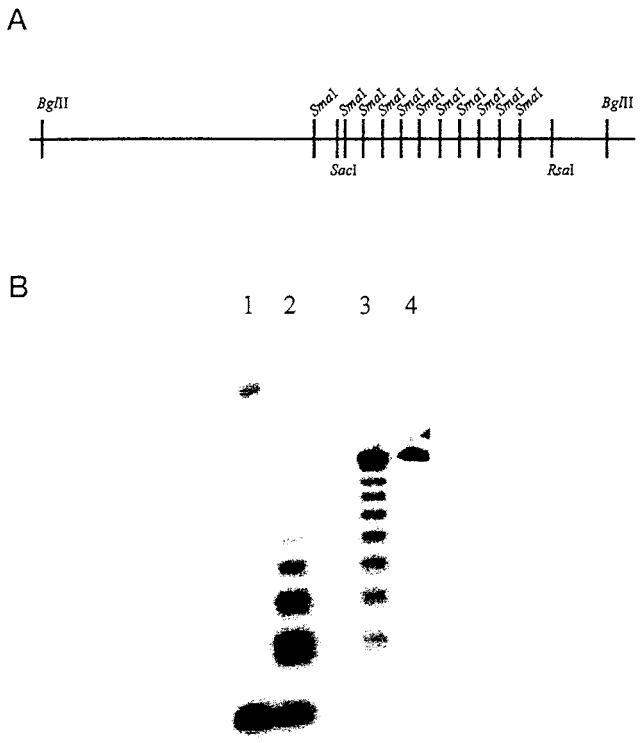


FIG. 6. **Characterization of the 59-bp repeat in the intron G.** *A*, partial restriction map of the *Bgl*II-*Bgl*II fragment from cosmid genomic clone BEN2. *B*, electrophoretic separation in agarose gel. The *Bgl*II-*Bgl*II fragment containing the intron G was cut with *Sac*I and *Rsa*I that flank the intron G and thereafter partially digested using *Sma*I. Blot analysis was conducted using the radiolabeled oligonucleotide NAU199. *Lane 1*, *Bgl*II-*Bgl*II fragment totally digested with *Sma*I; *lane 2*, *Sac*I-*Rsa*I fragment partially digested for 1 h with *Sma*I (1 unit/μg DNA); *lane 3*, *Sac*I-*Rsa*I fragment partially digested for 1 h with *Sma*I (0.25 unit/μg DNA); *lane 4*, *Sac*I-*Rsa*I fragment.

Acknowledgments—We are grateful for the technical assistance of Michel Crépin, Evelyne Destailleur, Viviane Mortelec, and Danièle Petitprez.

Note Added in Proof—While this manuscript was in review, Nielsen, P. A., Bennett, E. P., Wandall, H. H., Therkildsen, M. H., Hannibal, J., and Clausen, H. ((1997) *Glycobiology* **7**, 413–419) identified MG1 as tracheobronchial mucin MUC5B. On the other hand, Keates, A. C., Nunes, D. P., Afdhal, N. H., Troxler, R. F., and Offner, G. D. ((1997) *Biochem. J.* **324**, 295–303) published a partial genomic organization of the 3′ end of *MUC5B* with some differences from our data.

REFERENCES

1. Gendler, S. J., and Spicer, A. P. (1995) *Annu. Rev. Physiol.* **57**, 607–634
2. Bansil, R., Stanley, E., and LaMont, J. T. (1995) *Annu. Rev. Physiol.* **57**, 635–657
3. Forstner, G. (1995) *Annu. Rev. Physiol.* **57**, 585–605
4. Verma, M. (1994) *Cancer Biochem. Biophys.* **14**, 151–162

5. Ho, S. B., Roberton, A. M., Shekels, L. L., Lyftogt, C. T., Niehans, G. A., and Toribara, N. W. (1995) *Gastroenterology* **109,** 735–747
6. Buisine, M. P., Janin, A., Maunoury, V., Audié, J. P., Delescaut, M. P., Copin, M. C., Colombel, J. F., Degand, P., Aubert, J. P., and Porchet, N. (1996) *Gastroenterology* **110,** 84–91
7. Kaliner, M., Shelhamer, J. H., Borson, B., Nadel, J., Patow, C., and Marow, Z. (1986) *Am. Rev. Respir. Dis.* **134,** 612–621
8. Lee, S. P., LaMont, J. T., and Carey, M. C. (1981) *J. Clin. Invest.* **67,** 1712–1723
9. Klinkskpoor, J. H., Tytgat, G. N. J., and Groen, A. K. (1993) *Eur. J. Gastroenterol. Hepatol.* **5,** 226–234
10. Lan, M. S., Batra, S. K., Qi, W. N., Metzgar, R. S., and Hollingsworth, M. A. (1990) *J. Biol. Chem.* **265,** 15294–15299
11. Lancaster, C. A., Peat, N., Duhig, T., Wilson, D., Taylor-Papadimitriou, J., and Gendler, S. J. (1990) *Biochem. Biophys. Res. Commun.* **173,** 1019–1029
12. Bobek, L. A., Tsai, H., Biesbrock, A. R., and Levine, M. J. (1993) *J. Biol. Chem.* **268,** 20563–20569
13. Bobek, L. A., Liu, J., Sait, S. N. J., Shows, T. B., Bobek, Y. A., and Levine, M. J. (1996) *Genomics* **31,** 277–282
14. Gum, J. R., Byrd, J. C., Hicks, J. W., Toribara, N. W., Lamport, D. T. A., and Kim, Y. S. (1989) *J. Biol. Chem.* **264,** 6480–6487
15. Toribara, N. W., Gum, J. R., Culhane, P. J., Lagace, R. E., Hicks, J. W., Petersen, G. M., and Kim, Y. S. (1991) *J. Clin. Invest.* **88,** 1005–1013
16. Gum, J. R., Jr., Hicks, J. W., Toribara, N. W., Rothe, E. M., Lagace, R. E., and Kim, Y. S. (1992) *J. Biol. Chem.* **267,** 21375–21383
17. Gum, J. R., Jr., Hicks, J. W., Toribara, N. W., Siddiki, B., and Kim, Y. S. (1994) *J. Biol. Chem.* **269,** 2440–2446
18. Gum, J. R., Hicks, J. W., Swallow, D. M., Lagace, R. E., Byrd, J. C., Lamport, D. T. A., Siddiki, B., and Kim, Y. S. (1990) *Biochem. Biophys. Res. Commun.* **171,** 407–415
19. Porchet, N., Nguyen, V. C., Dufossé, J., Audié, J. P., Guyonnet Dupérat, Gross, M. S., Denis, C., Degand, P., Bernheim, A., and Aubert, J. P. (1991) *Biochem. Biophys. Res. Commun.* **175,** 414–422
20. Aubert, J. P., Porchet, N., Crépin, M., Duterque-Coquillaud, M., Vergnes, G., Mazzuca, M., Debuire, B., Petitprez, D., and Degand, P. (1991) *Am. J. Respir. Cell. Mol. Biol.* **5,** 178–185
21. Guyonnet Dupérat, V., Audié, J. P., Debailleul, V., Laine, A., Buisine, M. P., Galiegue-Zouitina, S., Pigny, P., Degand, P., Aubert, J. P., and Porchet, N. (1995) *Biochem. J.* **305,** 211–219
22. Meerzaman, D., Charles, P., Daskal, E., Polymeropoulos, M. H., Martin, B. M., and Rose, M. C. (1994) *J. Biol. Chem.* **269,** 12932–12939
23. Lesuffleur, T., Roches, F., Hill, A. S., Lacasa, M., Fox, M., Swallow, D. M., Zweibaum, A., and Real, F. X. (1995) *J. Biol. Chem.* **270,** 13665–13673
24. Klomp, L. W. J., Van Rens, L., and Strous, G. J. (1995) *Biochem. J.* **308,** 831–838
25. Dufossé, J., Porchet, N., Audié, J. P., Guyonnet Dupérat, V., Laine, A., Van Seuningen, I., Marrakchi, S., Degand, P., and Aubert, J. P. (1993) *Biochem. J.* **293,** 329–337
26. Toribara, N. W., Roberton, A. M., Ho, S. B., Kuo, W. L., Gum, E., Hicks, J. W., Gum, J. R., Jr., Byrd, J. C., Siddiki, B., and Kim, Y. S. (1993) *J. Biol. Chem.* **268,** 5879–5885
27. Nguyen, V. C., Aubert, J. P., Gross, M. S., Porchet, N., Degand, P., and Frézal, J. (1990) *Hum. Genet.* **86,** 167–172
28. Griffiths, B., Matthews, D. J., West, L., Attwood, J., Povey, S. M., Swallow, D. M., Gum, J. R., and Kim, Y. S. (1990) *Ann. Hum. Genet.* **54,** 277–285
29. Pigny, P., Guyonnet Dupérat, V., Hill, A. S., Pratt, W. S., Galiegue-Zouitina, S., Collyn D'Hooge, M., Laine, A., Van Seuningen, I., Gum, J. R., Kim Y. S., Swallow, D. M., Aubert, J. P., and Porchet, N. (1996) *Genomics* **38,** 340–352
30. Desseyn, J. L., Guyonnet Dupérat, V., Porchet, N., Aubert, J. P., and Laine, A. (1997) *J. Biol. Chem.* **272,** 3168–3178
31. Audié, J. P., Janin, A., Porchet, N., Copin, M. C., Gosselin, B., and Aubert, J. P. (1993) *J. Histochem. Cytochem.* **41,** 1479–1485
32. Audié, J. P., Tetaert, D., Pigny, P., Buisine, M. P., Janin, A., Aubert, J. P., Porchet, N., and Boersma, A. (1995) *Hum. Reprod.* **10,** 98–102
33. Balagué, C, Gambus, G., Carrato, C., Porchet, N., Aubert, J. P., Kim, Y. S., and Real, F. X. (1994) *Gastroenterology* **106,** 1056–1061
34. Campion, J. P., Porchet, N., Aubert, J. P., L'Helgoualc'h, A., and Clément, B. (1995) *Hepatol.* **21,** 223–231
35. Balagué, C., Audié, J. P., Porchet, N., and Real, F. X. (1995) *Gastroenterology* **109,** 953–964
36. Mancuso, D. J., Tuley, E. A., Westfield, L. A., Worrall, N. K., Shelton-Inloes, B. B., Sorace, J. M., Alevy, Y. G., and Sadler, J. E. (1989) *J. Biol. Chem.* **264,** 19514–19527
37. Troxler, R. F., Offner, G. D., Zhang, F., Iontcheva, I., and Oppenheim, G. O. (1995) *Biochem. Biophys. Res. Commun.* **217,** 1112–1119
38. Crépin, M., Porchet, N., Aubert, J. P., and Degand, P. (1991) *Biorheology* **27,** 471–484
39. Glisin, V., Orkvenjakov, R., and Byus, C. (1974) *Biochemistry* **13,** 2633–2637
40. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467
41. Birnstiel, M. L., Busslinger, M., and Strub, K. (1985) *Cell* **41,** 349–359
42. Berget, S. M. (1984) *Nature* **309,** 179–181
43. Hawkins, J. D. (1988) *Nucleic Acids Res.* **16,** 9893–9905
44. Mount, S. M. (1982) *Nucleic Acids Res.* **10,** 459–472
45. Tabak, L. A., Levine, M. J., Mandel, I. D., and Ellison, S. A. (1982) *J. Oral Pathol.* **11,** 1–17
46. Loomis, R. E., Prakobphol, A., Levine, M. J., Reddy, M. S., and Jones, P. C. (1987) *Arch. Biochem. Biophys.* **258,** 452–464
47. Chou, P. Y., and Fasman, G. D. (1974) *Biochemistry* **13,** 211–249
48. Smith, B. F., and LaMont, J. T. (1984) *J. Biol. Chem.* **259,** 12170–12177
49. Smith, B. F., and LaMont, J. T. (1985) *J. Clin. Invest.* **76,** 439–445
50. Hardy, D. M., and Garbers, D. L. (1995) *J. Biol. Chem.* **270,** 26025–26028
51. Baker, M. E. (1988) *Biochem. J.* **256,** 1059–1063
52. Mayadas, T. N., and Wagner, D. D. (1992) *Proc. Natl. Acad. Sci. U. S. A.* **89,** 3531–3535
53. Voorberg, J., Fontijn, R., Calafat, J., Janssen, H., van Mourik, J. A., and Pannekoek, H. (1991) *J. Cell Biol.* **113,** 195–205
54. Perez-Vilar, J., Eckhardt, A. E., and Hill, R. L. (1996) *J. Biol. Chem.* **271,** 9845–9850
55. Kawagishi, S., Fahim, R. E. F., Wong, K. H., and Bennick, A. (1990) *Arch. Oral Biol.* **35,** 265–272
56. Roberton, A. M., Mantle, M., Fahim, R. E. F., Specian, R. D., Bennick, A., Kawagishi, S., Sherman, P., and Forstner, J. F. (1989) *Biochem. J.* **261,** 637–647
57. Probst, J. C., Gertzen, E. M., and Hoffmann, W. (1990) *Biochemistry* **29,** 6240–6244
58. Hoffmann, W., and Joba, W. (1995) *Biochem. Soc. Trans.* **23,** 200–205
59. Van Klinken, B. J. W., Dekker, J., Büller, H. A., and Einerhand, A. W. C. (1995) *Am. J. Physiol.* **269,** G613–G627
60. Sun, P. D., and Davies, D. R. (1995) *Annu. Rev. Biophys. Biomol. Struct.* **24,** 269–91
61. Meitinger, T., Meindl, A., Bork, P., Rost, B., Sander, C., Haasemann, M., and Murken, J. (1993) *Nat. Genet.* **5,** 376–380
62. Hoffmann, W., and Hauser, F. (1993) *Comp. Biochem. Physiol.* **105B,** 465–472
63. Kingsley, D. M. (1994) *Genes Dev.* **8,** 133–146
64. Lynch, D. C., Williams, R., Zimmerman, T. S., Kirby, E. P., and Livingston, D. M. (1983) *Proc. Natl. Acad. Sci. U. S. A.* **80,** 2738–2742
65. Wagner, D. D. T, Mayadas, T., and Marder, V. J. (1986) *J. Cell Biol.* **102,** 1320–1324
66. Patthy, L. (1987) *FEBS Lett.* **214,** 1–7
67. Bork, P. (1992) *Curr. Opin. Struct. Biol.* **4,** 383–392
68. Pigny, P., Van Seuningen, I., Desseyn, J. L., Nollet, S., Porchet, N, Laine, A., and Aubert, J. P. (1996) *Biochem. Biophys. Res. Commun.* **220,** 186–191

La région 3' du gène *MUC5B* s'étend sur près de 11 kb (la séquence complète est accessible sous le numéro Y09788 dans la banque de données GenBank™) et comprend 18 introns qui ont une taille comprise entre 114 et 1118 pb. L'intron G ($7^{ième}$ intron en aval de l'exon central) a une taille de 597 pb. Il est majoritairement constitué d'un motif parfait de 59 pb répété 9 fois en tandem. Le premier motif est en partie sur l'exon précédant l'intron G. Ces motifs, très riches en G+C, contiennent tous un site de restriction *Sma*I ainsi qu'une boîte GC. Une protéine nucléaire de 42 kDa (Pigny *et al.*, 1996b) reconnaît cette boîte et pourrait être impliquée dans la régulation de *MUC5B* ou d'un gène proche de *MUC5B*. La purification de cette protéine et sa caractérisation biochimique sont en cours au laboratoire.

La partie 3' de *MUC5B* code un peptide de 808 aa très similaire à la région 3' du vWF, c'est-à-dire en aval des domaines A. MUC2, MUC5AC et MUC5B possèdent des régions carboxy-terminales très homologues entre elles, tout particulièrement en ce qui concerne les positions des résidus Cys. Notre analyse des séquences peptidiques nous a permis de subdiviser en 6 domaines la région carboxy-terminale de MUC5B comparativement aux régions carboxy-terminales de MUC5AC et MUC2 :

① un premier domaine que nous avons appelé «MUC11p15-type» car il est retrouvé uniquement dans les 3 apomucines

② un petit domaine commun aux 3 mucines et situé dans le vWF entre le domaine A3 et le domaine D4. Nous l'avons nommé A3uD4.

③ un domaine D4-like semblable à celui du vWF

④ un domaine B-like unique au lieu des 3 domaines B du vWF

⑤ un domaine C-like unique au lieu des 2 domaines C1 et C2 du vWF

⑥ un domaine CK qui fait de MUC5B, MUC5AC et MUC2 des membres de la mégafamille des protéines à CK. Ce domaine est probablement impliqué dans la dimérisation des protéines possédant ce domaine.

Enfin, la séquence peptidique de MUC5B comporte une région 100% homologue à la séquence peptidique déduite du clone pSM2-1. Cet ADNc code 196 aa de la région carboxy-terminale de la mucine salivaire de haut poids moléculaire MG1 (Troxler *et al.*, 1995). Comme, de plus, *MUC5B* s'exprime dans les glandes salivaires, on peut conclure que *MUC5B* code la mucine salivaire MG1.

Alors que nos travaux présentés ci-dessus étaient sous presse, 2 articles ont été publiés concernant la région 3' de *MUC5B* et un article présente la région carboxy-terminale de MUC6.

II.2 « Identification of a major human high molecular weight salivary mucin (MG1) as tracheobronchial mucin MUC5B »

Nielsen *et al.* avaient auparavant produit et caractérisé un anticorps monoclonal PANH2 contre la fraction déglycosylée de MG1 (Nielsen *et al.*, 1996). Cet anticorps a permis de cribler une banque d'expression (Nielsen *et al.*, 1997). Un clone, SAL1, de 849 pb, a été isolé, séquencé et étudié. La séquence nucléotidique est très homologue aux séquences d'ADNc JER57 et JUL7. Le clone d'ADNc PSM2-1 (Troxler *et al.*, 1995) a été utilisé comme sonde sur une banque génomique construite en bactériophage P1. Un clone a été isolé. Les sondes PSM2-1 et SAL1 s'hybrident à un fragment *Eco*RI-*Eco*RI de 25 kb de ce clone génomique. Ces auteurs, en utilisant un oligonucléotide sens dans la région 3' de JER57 et un oligonucléotide antisens dans la région 5' de PSM2-1 ont amplifié, par RT-PCR sur de l'ARN salivaire, une bande de 2,3 kb. Ce produit de PCR a été cloné et les séquences nucléotidiques des extrémités correspondent aux séquences 3' et 5' des clones JER57 et PSM2-1. Ces auteurs ont ensuite vérifié par Northern blot d'ARN de glande salivaire que SAL1 donne un continuum de bandes typique des mucines. Enfin, les études d'hybridation *in situ* menées par ce groupe montrent que *MUC5B* est exprimé par les cellules des glandes salivaires sublinguales, sous-maxillaires, palatine et labiales confortant les études d'immunohistochimie menées avec l'anticorps monoclonal (Nielsen *et al.*, 1996).

II.3 « Molecular cloning of a major gall bladder mucin : complete C-terminal sequence and genomic organization of MUC5B »

Ce second article qui a été publié en mai 1997 dans *Biochem. J.* présente en fait l'organisation génomique partielle de la région 3' de *MUC5B*. Ces travaux du groupe d'Offner

nous ont semblé, au regard de nos résultats, très critiquables. Au delà de la polémique qui oppose notre groupe au groupe américain concernant la région 3' de *MUC5B*, ces résultats mettent surtout l'accent sur la difficulté d'étudier des gènes comportant des séquences répétées en tandem.

Nous avons donc analysé cet article de très près et nous avons soumis pour publication nos remarques et critiques au même journal sous forme d'une lettre :

« Complete C-terminal sequence and genomic organization of *MUC5B* ? »

Complete C-terminal Sequence and Genomic Organization of MUC5B ?

Anne Laine

Unité 377 INSERM, Place de Verdun, 59045 Lille Cedex, France.

BJ LETTER

* Correspondence :

Dr. Anne LAINE

INSERM U-377

59045 LILLE Cedex-France

Tel : (33) 3 20 29 88 59

Fax : (33) 3 20 53 85 62

e-mail : laine@lille.inserm.fr

Keates *et al.* [1] recently published in this journal a paper entitled "Molecular cloning of a major human gall bladder mucin : complete C-terminal sequence and genomic organization of MUC5B". Concomitantly, we determined the complete genomic organization of the 3' region of *MUC5B* [2]. Since discrepancies on several points exist between our data and the findings published by Keates *et al.*, I wish to argue in this letter and present evidence that numerous shortcomings in [1] come closest to explaining these differences.

1- The authors state that five nucleotide sequences have been deposited in the GenBank, but no information is provided to show what sequence each number corresponds to. We have analyzed these sequences. They are pieces of a puzzle in which some parts are missing. The readers will appreciate the explanations I give here : the first one concerns the hGBM4-1 : U78550. The second one, U78551, concerns hGBM2-3, but its length is not 1565 bp as is written at the beginning of the fourth paragraph p : 299 , but 1439 bp- as can be deduced when the authors specified : nucleotides 1779 (not 1780) -3217 in Fig.2. Then, U78552 contains 700 bp with one exon from nucleotide 1 to 526, but there is no indication about the exact link with U78553, which, in fact, corresponds to the sequence of one part of the fragment of 5.5 kb obtained after *Sac*I digestion of hGBM G1-4/7.5 kb and which contains only 3616 bp. U78554 very probably concerns the 1.4 kb *Sac*I fragment of hGBM G1-4/7.5 kb.

2- In Fig. 1 the consensus sequence for the MUC5B repeat is not from ref. 28 but from ref. 4.

3- There is a lack of any physical link between hGBM4-1 and hGBM G1-4, and the size of exon 1 is given without any supporting evidence. Moreover, the authors should have mentioned in an added note, when they corrected the proofs, that the complete sequence of the central exon of *MUC5B* had recently been published [3]. It is not clear how the authors have been able to ascertain that the tandem repeat domain is quite large .

4- The hGBM G1-4/7.5 kb fragment was split with *Sac*I into 3 fragments of 5.5 ,1.4 and 0.4, i.e. 5.5 + 1.4 + 0.4 = 7.3 kb ... How can it be that the hGBM2-3 probe hybridizes the 1.4 kb fragment and yet does not hybridize the 0.4 kb fragment ?

5- Further detail would have been helpful to explain how the authors managed to perform all the RT-PCR experiments which were needed to assert that the exon-intron organization was properly determined (specific primers and their positions should have been specified).

6-We aligned the two cDNA nucleotide sequences of the two groups. After resequencing or rereading the sequences, we corrected 1 substitution (nt 3688) and one inversion (nt 4633-4634 in [2]) with two amino acid changes (see the corrected version of Y09788). Then, the comparison revealed six substitutions and four inversions leading to six amino acid

substitutions between the two deduced peptides. One inversion which exists in the sequence with accession number U78551 (nt 778-780 GAC) does not exist anymore in [1] (GCA). Moreover, the genomic sequence alignment allowed us to notice some bp insertions and/or deletions, specially in introns. Therefore, we compared the length of the introns published by the two groups (Table 1). We highlighted some differences and then we made our calculation of the size of the introns according to the sequences published by the group of Offner (Table 1). We demonstrate that they made mistakes concerning the length of the introns corresponding to the introns C, D, G, H and I according to our designation. But the differences in the intron sizes cannot be explained by mistakes in calculations only and we cannot exclude a polymorphism between the two genomic clones.

We determined that the size of the intron A is 679 bp in length while Keates *et al.* inferred that the first intron is about 900 bp in length without any experimental evidence, in fact, the boundaries of this intron overlap the two genomic sequences submitted to the GenBank™ with the accession numbers U78552 and U78553 [1].

Our laboratory has focused a great deal of attention on another intron designated as G [2]. After automatic resequencing, intron G is 597 bp in length (3 nucleotides were missing in [2]) and contains nine perfect direct repeats of 59 bp with one potential binding site SP1 that leads to a specific interaction with a nuclear factor, designated NF1-MUC5B, from mucus-secreting cells [4]. No similar repeat exists in the EMBL data bank.

Keates *et al.* determined the size of the intron J at precisely 195 bp. It is unclear how they obtained this length since they cloned the 5' part of this intron but not its 3' end. Hence, they cannot deduce its length exactly and the authors should have written '>209 bp' according to their data (Table 1).

Concerning the sizes of exons, we would like to point out some inconsistent results: exon 1, that we have designated as the central exon of 10,713 bp [3], is noted ~ 1.8 kb . How did the authors determine this size ? They should have added the size of hGBM4-1 (984 bp) to the size of the repeated part of hGBM G1-4/7.5 (525 bp) and noted > 1.509 kb. This would have been more reliable since they had no further indications. Exon 2 is not 181 but 182 bp in length and exon 8 is 70 instead of 71 bp in length, this is important with regard to the classes of the introns the authors specified. It would have been helpful to the reader if the authors had explained why the sizes of exons 1-10 in Figure 4 are not represented to scale. Moreover, the complete C-terminal sequence of MUC5B codes for 808 amino acids - and not 807 as written

by the authors -, this is just a calculation problem. The authors did not mention a polyadenylation signal nor a poly-A tail.

7- For the alignment performed in Fig. 5, the authors had the opportunity to choose between two published sequences, one by Meerzaman *et al.* (ref. 24) and the other by Lesuffleur *et al.* (ref. 26). They chose the one presented in ref. 26 but they always mentioned both references. This is confusing.

8. As far as the sequence of hGBM 2-3 is concerned, the authors could have found in the GenBank as well as in their own results, the pSM2-1 sequence (ref. 31) which is 100 % identical to the amino acid sequence 772-967 in Fig 2. The accession number of pSM2-1 is S80993 and has been given by the GenBank staff that created this entry from the original journal article ! One conclusion is thus missing from the article : *MUC5B* encodes MG1 as has also been demonstrated by Desseyn *et al.* [2], and at the same time by Nielsen *et al.* [5].

9- The authors mentioned a genomic clone of approx. 18 kb but no data is presented for the remaining part of this clone, apart from the 7.5 kb *Hinc*II fragment. If they had determined the complete sequence of this fragment, the authors should have been able to obtain the physical link with the hGBM4-1 and/ or, why not, the rest of the genomic organization of the 3' end of *MUC5B* with the 8 introns they did not find, but which are presented in our article [2]. In fact, the second part of the title previously provided as ref. 25 in the paper of Gipson *et al.* [6] would have been more accurate. It was : "Complete carboxyl-terminal sequence and genomic organization of the D4 domain of MUC5B" and at that time noted as being in press.

10- In the introduction, the authors specified : 'MUC5 (now referred to as MUC5AC)'. The fact is that some authors have persisted in naming MUC5 the gene which had been previously identified as MUC5AC [7]. One additional reference should have been included, namely Klomp *et al.* [8], as these researchers also have identified several cysteine-rich domains in the *N*-terminal region of MUC5AC as in MUC2.

Thus, there are assertions made by the authors that lack supportive evidence. This Letter will allow the readers to understand why discrepancies exist between the two papers.

I am deeply indebted to Jean-Luc Desseyn for his help in sequence analyses and in preparing this Letter. I also thank Michel Crépin for performance of automated DNA sequencing and Marcus Hurt for help in improving the style of this Letter.

References

1. Keates, A.C., Nunes, D.P., Afdhal, N.H., Troxler,R.F. and Offner, G.D. (1997) Biochem. J. **324**, 295-303.

2. Desseyn, J.L., Aubert, J.P., Van Seuningen, I., Porchet, N. and Laine, A. (1997) J. Biol. Chem. **272,** 16873-16883.

3. Desseyn, J.L., Guyonnet Dupérat, V., Porchet, N., Aubert, J.P. and Laine A. (1997) J. Biol. Chem., **272**, 3168-3178.

4. Pigny, P., Van Seuningen, I., Desseyn, J.L., Nollet, S., Porchet, N., Laine, A., and Aubert, J.P. (1996) Biochem. Biophys. Res. Commun., **220**, 186-191.

5. Nielsen, P.A., Bennett, E.P., Wandall, H.H., Therkildsen, M.H., Hannibal, J. and Clausen, H. (1997) Glycobiology, **7**, 413-419.

6. Gipson,I.K., Ho, S.B., Spurr-Michaud, S.J., Tisdale, A.S., Zhan, Q., Torlakovic, E., Pudney, J., Anderson, D.J., Toribara, N.W. and Hill, J.A. (1997) Biol. Reprod.,**56**, 999-1011.

7. Guyonnet Dupérat, V., Audié, J. P., Debailleul, V., Laine, A., Buisine, M. P., Zouitina-Galiègue, S., Pigny, P., Degand, P., Aubert, J. P. and Porchet, N. (1995) Biochem. J. **305**, 211-219

8. Klomp, L. W. J., Van Rens, L., and Strous, G. J. (1995) Biochem. J. **308**, 831-838

Table 1 Comparison of the intron sizes of *MUC5B* published by Keates *et al.* [1] with the corresponding introns published by Desseyn *et al.* [2]

Ref [1]		Our calculation*	Our work (Ref [2])	
Between exons	Size (bp)	bp	Name	Size (bp)
1-2	~ 900	>439	A	679
2-3	285	285	B	283
3-4	1,118	1,116	C	1,118
4-5	344	342	D	338
5-6	165	165	E	159
6-7	114	114	F	114
7-8	210	226	G	594
8-9	444	445	H	443
9-10	469	469	I	463
10-11	195	>209	J	258

* : according to the three accession numbers U78552, U78553 and U78554 [1]

Il faut noter que cette équipe ne mentionne pas l'existence de répétitions en tandem dans certains introns. Cependant, la séquence publiée par Keates *et al.* nous suggère qu'au moins l'intron G possède un polymorphisme de type VNTR. Nous avons été amenés à donner des résultats préliminaires concernant ce polymorphisme dans un article soumis pour publication :

« VNTR polymorphism of the seventh intron downstream of the central exon of *MUC5B.* »

VNTR Polymorphism of the Seventh Intron Downstream of the Central Exon of MUC5B.

Jean-Luc DESSEYN and Anne LAINE

Unité 377 INSERM, Place de Verdun, 59045 Lille Cedex - France.

Short title : VNTR polymorphic repeat in MUC5B

* Correspondence :

Dr. Anne LAINE

INSERM U-377

59045 LILLE Cedex - France

Tel : (33) 3 20 29 88 59

Fax : (33) 3 20 53 85 62

e-mail : laine@lille.inserm.fr

Abstract.

After sequencing the huge central exon of *MUC5B* (10,713 bp), we have recently determined the complete sequence and the exon-intron organization of the 3' region of this human mucin gene. Our initial genomic sequencing of *MUC5B* showed a 9 x 59 bp tandem repeat segment in the seventh intron downstream of the central exon. Using sequencing and Southern blot analyses, we demonstrate in this paper a polymorphism within this region probably due to a variable number of the 59 bp tandem repeat, each containing a potential regulatory sequence.

Introduction

Mucus is the layer that covers, protects and lubricates the luminal surfaces of epithelial respiratory, gastrointestinal and reproductive tracts. Mucins are the major glycoprotein components of mucus and constitute a family of high-molecular-mass glycoproteins. Mucins are synthesized by the goblet cells of the epithelia and in some cases by the submucosal glands (for review, see Gendler and Spicer 1995; Bansil et al. 1995; Forstner 1995). They have been implicated in numerous pathological situations such as in malignancy (Verma 1994). In the last few years there has been major efforts made in the research on mucin genes. Nine human mucins genes have been identified. *MUC5B*, mapped clustered with *MUC6, MUC2* and *MUC5AC* to chromosome 11p15.5 (Pigny et al. 1996a), is a human mucin gene expressed in submaxillary and bronchus glands, gall bladder and endocervix (Audié et al. 1993 ; Audié et al. 1995).

MUC5B has been cloned in our laboratory with three overlapping clones, two in cosmid (BEN1 and BEN2) and one in phage (CEL5). We published the genomic organization of the central region of this gel-forming mucin gene (Desseyn et al 1997a). Our research showed that the central domain is encoded by a single huge exon of 10,713 bp in length. Recently, we published the sequence of the 3' region of *MUC5B* and defined the complete exon-intron organization of the region downstream of the central domain (Desseyn et al. 1997b). On the other hand, Keates *et al.* reported a partial genomic organization of the 3' region of *MUC5B* (Keates et al. 1997). Several differences exist between our results and those of Keates *et al.* concerning the length of introns, particularly for the intron G. Our initial genomic sequencing showed a 9 x 59 bp tandem repeat segment in this intron. The sequence published by Keates et al. (1997) contains 2 perfect 59 bp repeats and one incomplete. Using Southern blot

analyses, we demonstrate in this article that at least three alleles exist with 9 x 59 bp, 8 x 59 bp and 6 x 59 bp respectively. The one described by Keates et al. may be a fourth one. Then, we provide evidence that a length polymorphism exists for intron G that might be an extremely valuable tool in linkage analysis.

Materials and methods

Nucleotide Sequences Alignment

Nucleotide sequence data appear in the EMBL Nucleotide Sequence Database under the accession numbers Y09788 (the complete 3'-end genomic sequence of *MUC5B* published in (Desseyn et al. 1997b)), U78552, U78553 and U78554 (the genomic sequences published in (Keates et al. 1997)). Alignment of nucleotide sequences was performed using the CLUSTAL program of the PC/GENE Package Software (IntelliGenetics, Inc., CA) and the alignment was manually optimized.

Southern Blotting

Genomic DNA, isolated from lymphocytes, cosmid DNA (BEN2) or phage DNA (CEL5) was digested with restriction endonucleases *Rsa*I and *Sac*I and separated on 1% agarose gel. The DNA in the gel was denatured, neutralized and transferred to Hybond™-N$^+$ membrane (Amersham Corp.). The membrane was baked for 2h at 80°C in a vacuum oven, pre-hybridized at 65°C in 0.1 X SSC (1X SSC : 0.15 M sodium chloride, 0.015 M sodium citrate, pH 7) and 0.1% SDS, then hybridized overnight at 65°C with the corresponding *Rsa*I-*Sac*I fragment from the cosmid clone BEN2, labeled with [α-^{32}P]dCTP (by random priming using standard procedures). Membrane was washed 2 times for 15 min with 0.1 X SSC, 0.1% SDS at 65°C and exposed for 72 hours to Kodak X-OMAT films

DNA-sequencing

Sequences were performed manually using the dideoxynucleotide chain termination method. Sequencing reaction mixtures were electrophoresed on 6% polyacrylamide gels (Sequagel-6™ , National Diagnostics). Sequences were also performed on an ABI Prism™ 310 automated sequencer (Perkin Elmer). All the sequences were performed using universal and specific primers.

Results and discussion

Four introns of the 3' end of *MUC5B* contain tandemly repeated sequences that are more or less perfect (Desseyn et al. 1997b). Our laboratory has focused a great deal of attention on the intron designated as G. After automatic resequencing, intron G is 597 bp in length (3 nucleotides were missing in (Desseyn et al. 1997b)) and contains nine perfect direct repeats of 59 bp (Fig 1) with one potential binding site SP1 that leads to a specific interaction with a nuclear factor, designated NF1-MUC5B, from mucus-secreting cells (Pigny et al. 1996b).

The shorter length reported by Keates *et al.* (210 bp) for this intron led us to try to explain the difference observed with the length of intron G we had determined. Then we hypothesized that intron G presents some length polymorphism due to a variable number of tandem repeats. Consequently, we performed an experiment using Southern blotting (Fig 2) of seven random unrelated individuals and of the genomic clones (BEN2 and CEL5), after digestion by the endonucleases *Rsa*I and *Sac*I which flank this intron (Fig 1). Analysis with the *Rsa*I-*Sac*I probe of BEN2 clearly shows the presence of three bands. The highest present in lanes 5, 8 and 9 migrates as 630 bp instead of 683 bp according to the sequence determined in BEN2 and CEL5, this is due to the very high proportion of G and C. One band migrates as 570 bp (in lanes 1 to 7) and the third band as 450 bp (in lane 1), thus probably lacking one and three 59 bp repeats respectively when compared with BEN2 or CEL5 (Fig 2). Hence, intron G seems to have a length polymorphism. But these fragments contrast in length with the length of the *Sac*I-*Rsa*I fragment of 309-bp calculated from the sequence reported by Keates *et al.* (Fig 1). The allele lacking 6 repeats, which would correspond to the published genomic sequence from Keates *et al.*, was not found in our experiment. The study has to be performed with more individuals. Nevertheless, these observations suggest that

the polymorphism in length of intron G is probably due to a variable number of repeats of 59 bp sequence since the sequences which flank these repeats are very similar in both studies. This VNTR polymorphism has probably been generated by unequal crossing over during mitosis or meiosis.

Intron G of *MUC5B* shares several features with intron 6 of human interleukin-1α (Bailly et al., 1996). This intron contains several tandem repeats of 46 bp. Bailly *et al.* have shown a polymorphism due to a variable number of repeats of the 46 bp sequence. Six alleles ranging from 5 to 18 repeats have been found. This polymorphism is potentially relevant to gene function, since each repeat was described as containing three potential binding sites for transcriptional factors. The functional consequences of the variation in length has been explored by the same authors. They have suggested that this polymorphism could play a role in the complex regulation of IL-1α gene transcription. The hypothesis of whether this polymorphism could play a role in *MUC5B* gene regulation is currently under investigation. Thus, the length polymorphism of this intron might be an extremely valuable tool in linkage analysis. Moreover, *MUC2*, *MUC5AC* and *MUC5B* may have evolved by gene duplications starting from a common ancestor (Desseyn et al. 1997c) and share probably the same exon-intron structure. It will be interesting in the future to sequence the two introns of *MUC5AC* and *MUC2* corresponding to intron G in *MUC5B*.

Acknowledgments

This work was supported by Le Comité du Nord de La Ligue contre le Cancer, and L'Association de Recherche sur le Cancer. The authors wish to thank Annette Leclercq and Christine Mouton for their technical assistance. We also thank Michel Crépin for performance of automated DNA sequencing and Marcus Hurt for help in improving the style of the paper.

References

Audié, J.P., Janin, A., Porchet, N., Copin, M.C., Gosselin, B., and Aubert, J.P. (1993) J. Histochem. Cytochem. 41 : 1479-1485

Audié, J.P., Tetaert, D., Pigny, P., Buisine, M.P., Janin, A., Aubert, J.P., Porchet, N., and Boersma, A. (1995) Hum. Reprod. 10 : 98-102

Bailly, S., Israel, N., Fay, M., Gougerot-Pocidalo, M.A., and Duff, G.W. (1996) Mol. Immunol. 33 : 999-1006

Bansil, R., Stanley, E., and LaMont, J.T. (1995) Annu. Rev. Physiol. 57 , 635-657

Desseyn, J.L., Guyonnet Dupérat, V., Porchet, N., Aubert, J.P., and Laine, A. (1997a) J. Biol. Chem. 272 : 3168-3178

Desseyn, J.L., Aubert, J.P., Van Seuningen, I., Porchet, N., and Laine, A. (1997b) J. Biol. Chem. 272 : 16873-16883

Desseyn, J.L., Buisine, M.P., Porchet, N., Aubert, J.P., Degand, P., and Laine, A. (1997c) J. Mol. Evol (in press)

Forstner, G. (1995) Annu. Rev. Physiol. 57 , 585-605

Gendler, S.J., and Spicer (1995) A.P., Annu. Rev. Physiol. 57 , 607-634

Keates, A.C., Nunes, D.P., Afdhal, N.H., Troxler, R.F., and Offner, G.D. (1997) Biochem. J. 324 : 295-303

Pigny, P., Guyonnet Dupérat, V., Hill, A.S., Pratt, W.S., Galiegue-Zouitina, S., Collyn D'Hooge, M., Laine, A., Van Seuningen, I., Gum, J.R., Kim Y.S., Swallow, D.M., Aubert, J.P., and Porchet, N. (1996a) Genomics 38 : 340-352

Pigny, P., Van Seuningen, I., Desseyn, J.L., Nollet, S., Porchet, N, Laine, A., and Aubert, J.P. (1996b) Biochem. Biophys. Res. Commun. 220 : 186-191

Verma, M. (1994) Cancer Biochem. Biophys. 14 : 151-162

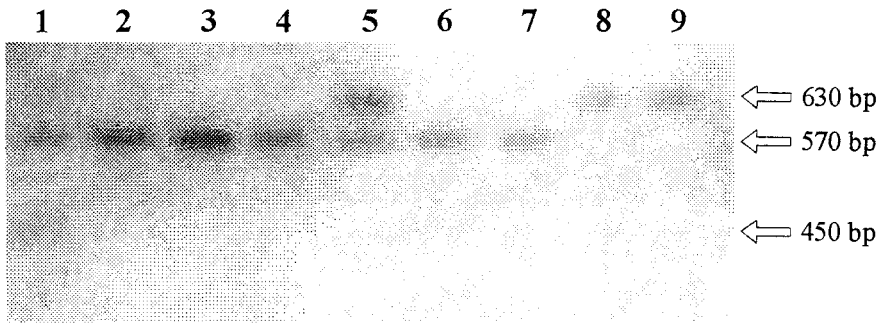
Legends

Figure 1 - Alignment of the intron G and its bordering regions reported by Desseyn *et al.* (1997b) (under accession number Y09788 corrected in the GenBank) and Keates *et al.* (1997) (under accession numbers U78552, U78553 and U78554 in the GenBank). Alignment was done using PC/GENE Software and optimized manually. The *numbers* on the *right* are the nucleotide positions. *Dashes* indicate gaps introduced in the sequence for alignment purposes. Non identical nucleotides are *thin boxed.* The individual tandem repeat units have been *numbered* above them. The *Rsa*I and the *Sac*I sites are *underlined.*

Figure 2 - Southern blot analysis for the determination of the length of the intron G. DNA of 7 random unrelated individuals was isolated using standard procedures as described previously (Desseyn et al. 1997a). An aliquot of 10 µg DNA was digested with *Sac*I and *Rsa*I and subjected to Southern analysis (lanes 1 to 7). Cosmid BEN2, lane 8, and phage CEL5, lane 9, were also digested with *Sac*I and *Rsa*I and analysed. Then the *Sac*I-*Rsa*I fragment of the genomic clone BEN2 was used as probe. The upper arrow correlates with the *Rsa*I-*Sac*I fragment described previously (Desseyn et al. 1997b), given in Fig 1 and having 9 x 59 bp repeats; the second arrow is correlated with a fragment with 8 x 59 bp repeats. The lowest arrow correlates with a fragment lacking 3 x 59 bp repeats when compared with the highest band.

Figure 1

Figure 2



II.4 La région carboxy-terminale de MUC6

L'organisation génomique de la région carboxy-terminale de MUC6 a été publiée en juin 1997 (Toribara *et al.*, 1997). Ce gène est différent des 3 autres gènes de mucines localisés en 11p15.5 puisque *MUC6* ne possède pas derrière sa région répétitive de domaines D4, B et C mais uniquement un domaine CK. La région 3' du gène comprend uniquement 2 introns équivalents aux introns Q et R dans *MUC5B* (mêmes positions). Les classes de ces 2 introns sont par ailleurs identiques pour les 2 gènes (1 et 2 respectivement).

Ces résultats confortent l'arbre phylogénétique que nous avons proposé précédemment (Desseyn *et al.*, 1997b) : MUC6 est, des 4 gènes de mucines localisés sur le chromosome 11, le plus éloigné des trois autres dans la phylogenèse. Ce gène a probablement perdu un fragment par délétion d'une région comprise entre 2 introns de même classe (Patthy, 1996). Cette hypothèse semble vraisemblable car les introns A et Q de *MUC5B* et l'intron Q de *MUC6* sont de même classe (classe 1 ; Figure 29). Ainsi, la délétion génomique de la région bordée par les anciens introns putatifs A et Q de *MUC6* ne modifie pas le cadre de lecture, et le motif peptidique CK reste conservé.

Une première duplication a donc probablement eu lieu donnant naissance tout d'abord à l'ancêtre de *MUC6* et à l'ancêtre commun de *MUC2*, *MUC5AC* et *MUC5B*. Ce dernier s'est ensuite dupliqué comme proposé précédemment dans notre article traitant de l'évolution de la famille des mucines du chromosome 11. L'ancêtre de *MUC6* a ensuite perdu une partie 3' pour donner naissance au gène que nous connaissons actuellement. Ce schéma hypothétique est représenté sur la figure 29.

Afin d'apporter de nouvelles preuves confortant notre schéma, nous avons aligné les motifs peptidiques CK des mucines avec ceux du NDP et du vWF (Figure 30). Nous avons calculé les pourcentages d'homologie entre les différents peptides (Tableau X). On observe que PSM est très homologue pour le domaine CK à BSM (81%) et que MUC2 a une homologie de 70% avec son homologue chez le rat. De plus, des 4 mucines humaines du chromosome 11, MUC5B est plus proche de

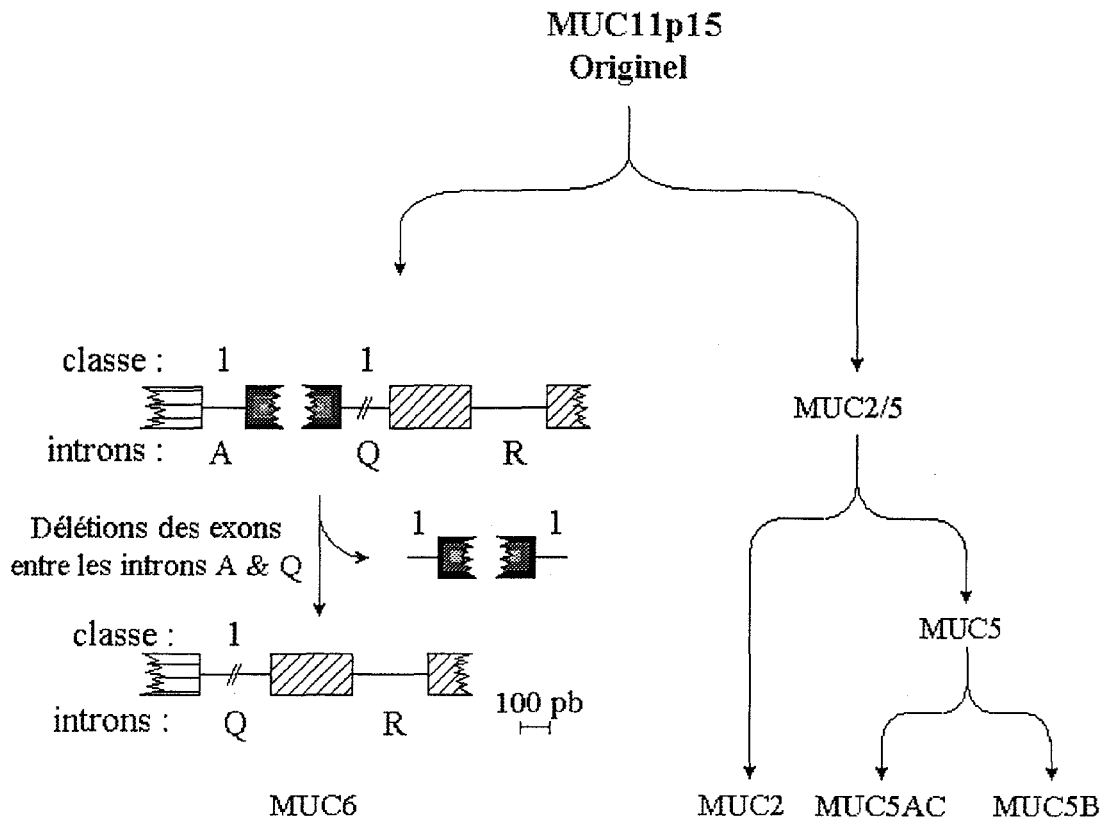


Figure 29 : Schéma hypothétique montrant l'évolution des gènes de mucines localisés en 11p15 à partir d'un ancêtre commun (MUC11p15 originel)

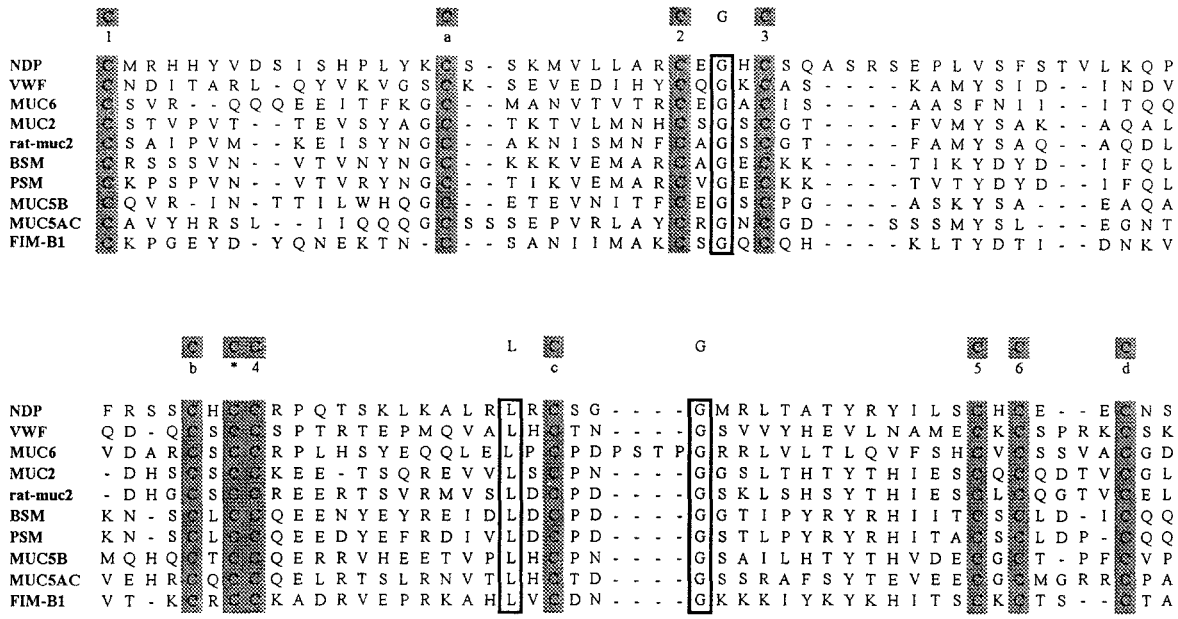


Figure 30 : Comparaison des motifs peptidiques CK du vWF, du NDP, des mucines humaines localisées en 11p15 et de certaines mucines animales

Tableau X. Pourcentage d'homologie entre les différents motifs CK d'après l'alignement présenté ci-dessus.

	NDP	VWF	MUC6	MUC2	Rat-muc2	BSM	PSM	MUC5B	MUC5AC	FIM-B1
NDP	100									
VWF	21	100								
MUC6	25	29	100							
MUC2	29	35	31	100						
rat-Muc2	25	38	33	70	100					
BSM	28	31	28	47	45	100				
PSM	27	32	29	49	46	81	100			
MUC5B	25	35	33	43	42	38	38	100		
MUC5AC	27	38	26	39	43	32	33	44	100	
FIM-B1	25	37	26	36	38	39	43	33	30	100

MUC5AC (44%) ; puis de MUC2 (43%) alors que MUC6 a un taux d'homologie moins élevé avec MUC2 (31%), MUC5B (33%) et MUC5AC (26%). MUC6 est donc plus éloigné des 3 autres gènes de mucines localisés en 11p15. Les résultats sont similaires si on effectue les calculs sur les séquences nucléotidiques. Ces résultats confortent donc notre schéma hypothétique d'évolution des 4 gènes de mucines humaines du chromosome 11 (Figure 29).

III LA REGION AMINO-TERMINALE

Nous avons obtenu très récemment les dernières séquences permettant de présenter l'organisation génomique de la région 5' de *MUC5B*. Certains résultats devront être confirmés ou précisés pour une future publication.

III.1 Résultats

III.1.1 Clonage et séquençage de l'ADNc 5' de *MUC5B*

III.1.1.1 Expériences de RACE-PCR

Nous avons poursuivi les travaux de DEA par 2 expériences de RACE-PCR successives (Figure 31). Les 2 clones chevauchants RACE66 et RACE76 de respectivement 600 pb et 200 pb ont été clonés et séquencés. La progression vers le 5' transcrit de *MUC5B* étant assez lente, nous avons ensuite changé de stratégie.

III.1.1.2 RT-PCR utilisant des oligonucléotides dégénérés

Grâce aux similarités de séquences entre MUC2, le pro-vWF, HGM-1 et les peptides déduits des expériences de RACE-PCR sur *MUC5B*, nous avons déduit la séquence d'un oligonucléotide dégénéré NAU101 qui doit se retrouver 3 fois dans l'extrémité 5' de *MUC5B* (voir Stratégie, §III.3.2). Cet oligonucléotide a été utilisé pour effectuer des RT-PCR en utilisant un second oligonucléotide spécifique de *MUC5B*. Cette technique nous a permis d'isoler successivement les 3 clones chevauchants 101/76 (530 pb), 101/111 (1416 pb) et 101/114 (1120 pb). Les séquences des extrémités de ces clones sont identiques à des séquences obtenues en

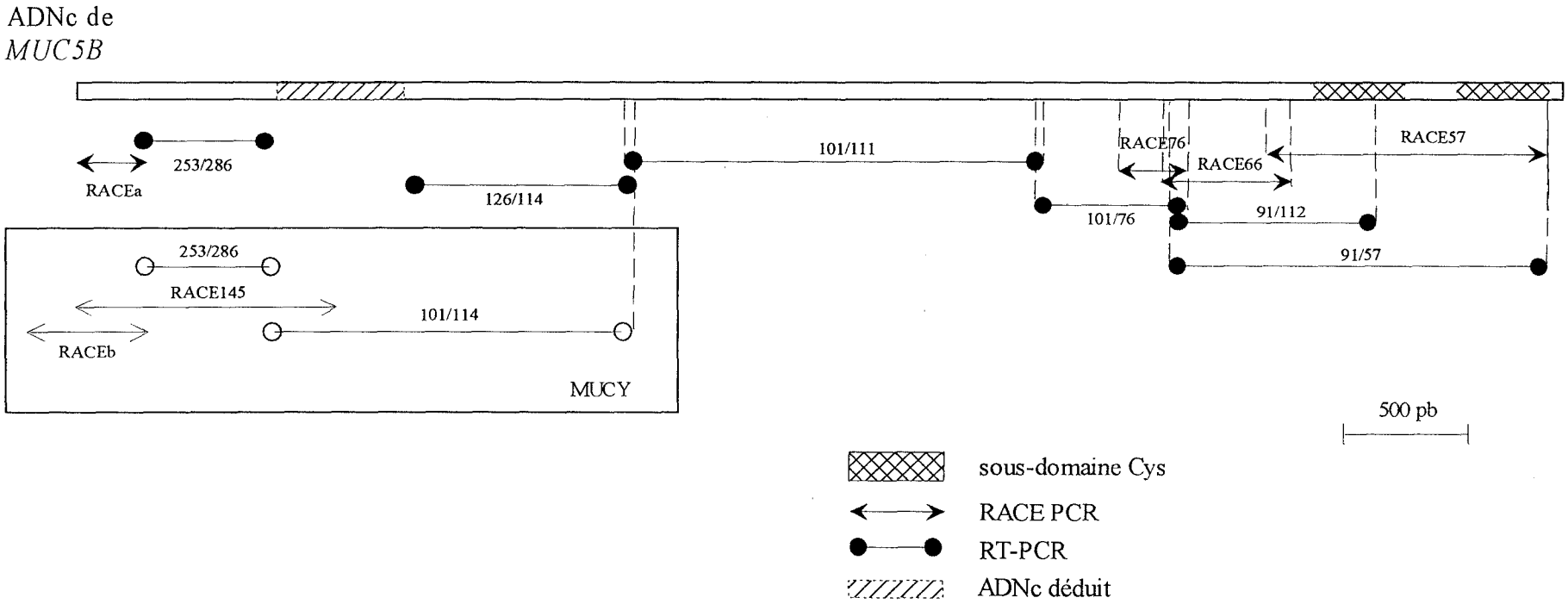


Figure 31 : Stratégie de clonage de l'ADNc 5' de *MUC5B*

utilisant les oligonucléotides sur de l'ADN cosmidique BEN1, suggérant que ces clones sont bien issus du gène *MUC5B*. Les différents clones ont été découpés et sous-clonés en plasmide pKS puis séquencés.

III.1.1.3 Clonage de l'extrémité 5' du transcrit de MUC5B

Nous avons alors effectué une expérience de RACE-PCR (RACE145) qui nous a permis de cloner un ADNc d'environ 600 pb correspondant à la taille attendue d'après les séquences de *MUC2* et du *vWF*. Les clones de RACE-PCR ont été sous-clonés en vecteur pKS comme décrit précédemment (Desseyn *et al.*, 1997a ; Desseyn *et al.*, 1997c) et la séquence a été entièrement déterminée. Afin de vérifier que nous avions bien cloné l'extrémité tout à fait 5' du transcrit de *MUC5B*, nous avons effectué une dernière expérience de RACE-PCR en utilisant un oligonucléotide situé environ 300 pb en aval de l'extrémité 5' du clone RACE145. Nous avons obtenu 2 familles de clones appelées RACEa et RACEb de 170 pb et 300 pb respectivement. Ces 2 familles ont pu être amplifiées en utilisant de l'ARN total de vésicule biliaire préparée au laboratoire ou de l'ARNm commercial de glande salivaire ou de trachée. La séquence nucléotidique du clone RACEb est identique à la séquence 5' du clone RACE145. Les 2 clones RACEa et RACEb ont des séquences presque 100% identiques à leurs extrémités 3' sur 70 nucléotides. Les 2 inserts des clones RACEa et RACEb utilisés comme sondes sur des Southern blots du cosmide BEN1 marquent les mêmes fragments suggérant que ces 2 ADNc ne sont pas issus d'un épissage alternatif de *MUC5B*. Les séquences déterminées parallèlement indiquent que les 3 clones chevauchants (recouvrant en tout environ 1,8 kb) 101/114, RACE145 et RACEb proviennent d'un gène autre que *MUC5B* et que nous avons baptisé *MUCY*. Ce gène est très homologue en séquence à *MUC5B*. Par exemple, les 2 clones 253/286 issus des gènes *MUC5B* et *MUCY* ont une homologie de 72% mais cette homologie est supérieure à 95% pour les extrémités de ces clones.

III.1.1.4 RT-PCR complémentaires utilisant 3 couples d'amorces

Nous avons alors entrepris, d'après les nouvelles séquences génomiques de *MUC5B*, 3 RT-PCR chevauchantes entre les clones RACEa et 101/111. Nous avons pu ainsi amplifier et cloner 2 fragments différents, de taille très proche, avec le couple d'oligonucléotides NAU253 et NAU286 (269 et 272 pb). L'un de ces clones appartient à *MUCY*, l'autre à *MUC5B*. Nous avons aussi cloné et séquencé un fragment de 669 pb (clone 126/114 sur la figure 31).

L'ADNc de *MUC5B* correspondant au fragment compris entre les clones 253/286 et 126/114 n'a pas pu être obtenu jusqu'à présent. Nous avons déduit l'organisation génomique de cette région par comparaison entre les séquences génomiques de *MUC5B* et celles des ADNc de *MUC2* et *MUCY* en repérant les zones les plus conservées entre les régions codantes et en tenant compte des séquences caractéristiques des introns (Mount, 1982).

III.1.1.5 Extension d'amorce

Nous avons fait synthétiser l'oligonucléotide antisens NAU346 d'après la séquence du clone RACEa et la séquence génomique correspondante. L'expérience d'extension d'amorce révèle une bande de 97pb par rapport à une séquence témoin (Figure 32). Nous attendions une bande d'une centaine de nucléotides d'après la longueur du clone RACEa.

III.1.2 Organisation génomique

III.1.2.1 Clonage génomique

La carte partielle de restriction du cosmide BEN1 avait été déterminée. Les différents fragments génomiques reconnus par les clones d'ADNc utilisés comme

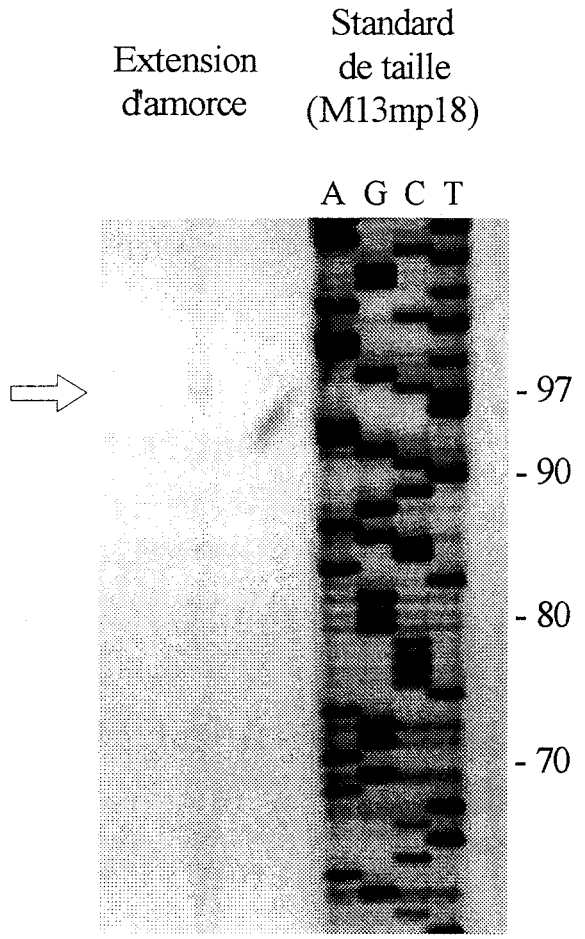


Figure 32 : Extension d'amorce avec l'oligonucléotide NAU346

sondes, ont été parallèlement sous-clonés en vecteur pKS au laboratoire. Ces sous-clones (Figure 33A) ont été séquencés en utilisant des oligonucléotides spécifiques du site de clonage multiple du pKS et des oligonucléotides spécifiques des inserts sous-clonés, que nous avons fait synthétiser.

L'organisation génomique complète de la région 5' de *MUC5B* est schématisée sur la figure 33B.

Trois PCR longues ont été menées avec les 3 oligonucléotides antisens NAU76, NAU111, NAU114 couplés à l'oligonucléotide sens dégénéré NAU101. Les tailles obtenues (respectivement 3,2kb, 4,9kb et 4,5 kb) suggèrent que la région 5' de *MUC5B* s'étend en amont de l'îlot CpG distant de 11 kb seulement de l'exon central. Ceci a été confirmé sur des Southern blots de BEN1 en utilisant les ADNc et les oligonucléotides comme sondes. Ces résultats font apparaître une difficulté supplémentaire qui a retardé notre progression puisque nous n'avions, jusqu'alors, pas entrepris de sous-cloner la région en amont de l'îlot CpG. De plus, la carte de restriction était moins bien déterminée car cette région se situe juste en amont du site *Not*I interne dans le cosmide BEN1.

III.1.2.2 Jonctions exon-intron

Nous avons déduit l'organisation génomique complète de la région 5' de *MUC5B* par comparaison des séquences génomiques et des séquences d'ADNc. Comme nous l'avons vu auparavant, il reste à obtenir l'ADNc d'un petit fragment de 500 pb afin de vérifier l'organisation génomique que nous avons déduite dans la région correspondante. La région 5' du gène s'étend sur 15,2 kb.

Toutes les jonctions exon-intron à l'exception de celle de l'intron 8 sont conformes à la séquence consensus de Mount (Mount, 1982 ; Tableau XI). La répétition parfaite directe CAGgca (encadrée sur le tableau XI) de part et d'autre de

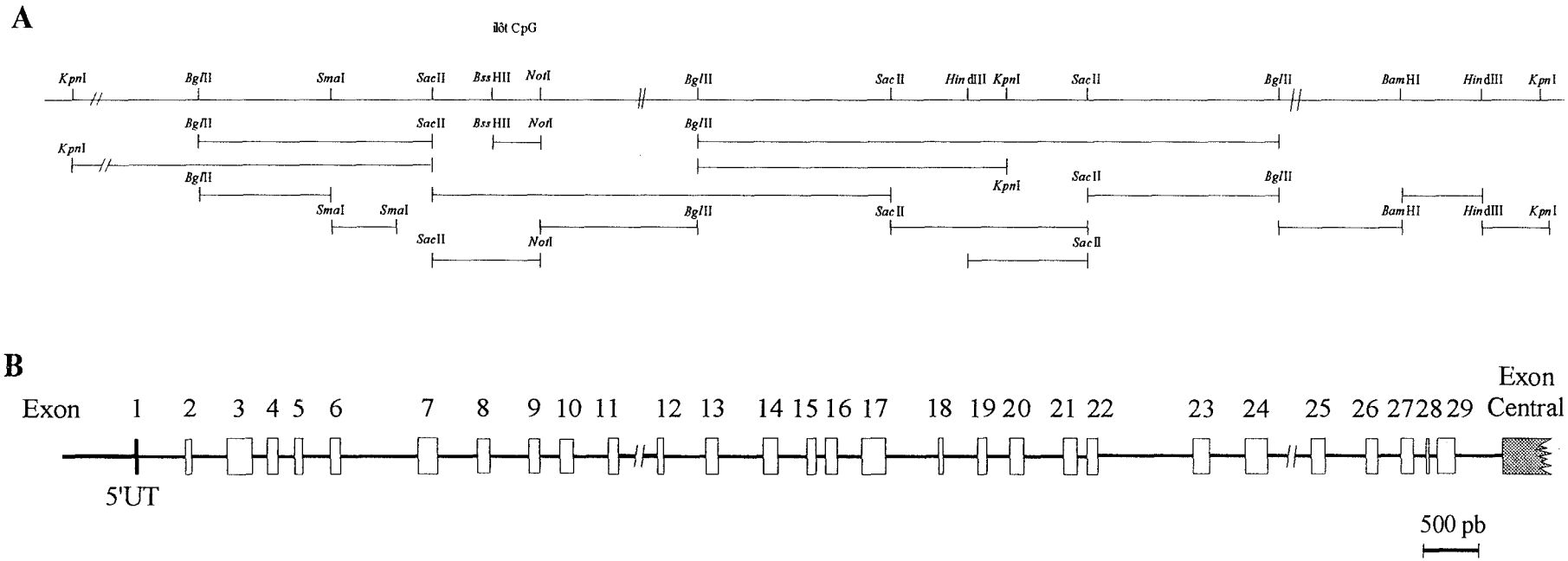


Figure 33 : Stratégie de sous-clonage génomique (A) et organisation génomique déduite (B) de la région 5' de *MUC5B*

Tableau XI. Jonctions exon-intron de la région 5' de *MUC5B*.

Domaine protéique	EXON N°	EXON Taille (pb)	Site donneur	INTRON N°	INTRON Taille (pb)	INTRON Classe	Site accepteur
5'NT+5B	1	46	A T G G C G **g t** a t g t g g c c a g g t t c	1	464	1	c a g c c t t c a c c c a c **a g** G T G C C C
5B+D1-like	2	72	T G A G C C **g t** a a g c a g a t g c t g c c	2	338	1	c c c g c g c t c c c c a c **a g** C C C T G A
D1-like	3	262	G C A G C G **g t** g a g c c g g c c a c c c t	3	154	2	c c c g c t g t c t g c g c **a g** G G A G G A
D1-like	4	115	G C C C T G **g t** g a g g a a g c c c c c t c	4	160	0	c c c t c c c c a c t c t c **a g** C T G G A G
D1-like	5	91	C A G T G A **g t** g c c a c c t g g g t g a g	5	270	1	g c t t t c t t c c c g g c **a g** A C G C C A
D1-like	6	111	A G G T G A **g t** c c c c g c c a c c c c c a	6	782	1	c c a c t g t g c t c c c c **a g** G A G G C A
D1-like	7	201	T C T G C C **g t** g a g t g c t c c c a g g g	7	403	1	g c c t c c t t t t g c g c **a g** C C C G G A
D1-like	8	126	C C C C A G **g c a** g g t c t t g t g t g c c	8	394	1	a c g c c g c g c c c c a **c a g** G C A C G G
D1-like	9	118	C T C C T G **g t** a c t t a t g a g c c c a c	9	196	2	c g g c c t c c c t c c c c **a g** C A C C T G
D1+D2-like	10	142	T C C A A G **g t** c t g g g g c t t g g g g c	10	346	0	g g t c c t t c t c c c c c **a g** A A A T G T
D2-like	11	112	G A C A C G **g t** g g a g g a c c t g g c c t	11	~415 *	1	c c g c c t c c t t c c g c **a g** G C C A T C
D2-like	12	69	C G G C A G **g t** a t g t g g c t c t c c c a	12	420	1	c c c a c a c t c c c g g c **a g** C C A A C A
D2-like	13	138	T G T G C G **g t** g a g g c t g g g c a g g g	13	438	1	c c a a c c c t g g c c c c **a g** G C C T G T
D2-like	14	162	A G A A T G **g t** a c t c c t c g c c c c c a	14	290	1	c a g g t c t t c c c c a c **a g** A G A A C T
D2-like	15	95	C A C T C G **g t** g a g a g g c t g a g g c a	15	98	0	c c t c c c a c c t c c g c **a g** A A C T G C
D2-like	16	127	T C T G C A **g t** g a g t g c c c a c g c t g	16	242	1	g c c c c a c c g a c c a c **a g** C C A A G T
D2-like	17	256	C G T G T G **g t** a a g g g t c t g g g g g g	17	523	2	c g t c t g t c t c t t g t **a g** T T C A T G
D'-like	18	56	G C A C A G **g t** a a g t g c c a c c c c t g	18	347	1	c a a g c t c t g t c c c c **a g** G G T G T G
D'-like	19	101	G G C T G T **g t** g a g t t c c a t g c t t c	19	223	0	t g c c t g g g c c c c c a c **a g** T T C A G C
D'-like	20	152	C A C C T G **g t** g g g t c g t g a g t c t c	20	393	2	t t g t g g c c c c t t g c **a g** C A C C T G
D'+D3-like	21	139	G C C C A G **g t** a c g c c g c c c c c t c g	21	98	0	t c t c t t t c t g g c c c **a g** G A C T A C
D3-like	22	111	G T G G A G **g t** g a g a a c g g c c c c a g	22	975	0	t c t g c c t c c c c t g c **a g** A G C T A C
D3-like	23	177	T A C A A G **g t** g a g c t c g g g c c g t g	23	357	0	c t c g c t g c c t c t g c **a g** G G C A G G
D3-like	24	240	T C C C A G **g t** g g g g c t c t g g t c t t	24	~1700 *	0	t t g g c c t c a c c c g c **a g** G T T G A C
D3-like	25	157	C C T G C C **g t** g a g t c g g g c t c t g t	25	406	1	g c c c c c t g t g c c c c **a g** C C T T G T
D3-like	26	129	T G G A A G **g t** g a g g g g c a g c c t t t	26	231	1	a c c c t g c c g g c c c t **a g** G C T G C T
D3-like	27	145	G A G C T G **g t** g a g g g g g t g g g a a g	27	100	2	t c t c c c c c a c c c c t **a g** T A A C T G
D3-like	28	44	T T G A G G **g t** a a g g a a g g g c c g g g	28	87	1	c c g g c t c c a t c c c c **a g** C C T G C A
M2/5AC/5B	29	198	C G A C A A **g t** a a g c c c t g c c t g g c	29	468	1	c c c a t g c c c c t t g c **a g** G C C C G G
Central	30	10.713	C G C C C G **g t** g a g t g c a t g t g g a t	30			
Consensus (Mount, 1982)			C a A G **g t** a g t A g				c c c c c c c c c c c c n **a g** t t t t t t t t t t t t

5' NT : 5' non traduit

* : les tailles des introns 11 et 24 ont été évaluées grâce à des PCR effectuées sur le cosmide BEN1. Ceci nous a permis de préciser qu'une séquence d'environ 250 pb reste à déterminer pour ces 2 introns.

cet intron, donne 7 possibilités différentes de le placer. La séquence génomique a été vérifiée par séquençage automatique (séquenceur Perkin Elmer ABI Prism™ 310). Quelle que soit la place de cet intron, le site donneur n'est pas consensuel et nous avons placé cet intron entre les triplets codants CAG et GCA pour les 2 raisons suivantes :

- la jonction que nous suggérons, avec la séquence CAGgc dans le site donneur et agG dans le site accepteur, est la plus proche de la séquence consensus (séquences MAGgt dans le site donneur et nag dans le site accepteur). Ainsi, seul le dinucléotide gt du site donneur est substitué par le dinucléotide gc. La jonction exon-intron déduite est, parmi les séquences consensus répertoriées, la plus souvent retrouvée (Jackson, 1991).

- Nous disposons des séquences des extrémités d'un fragment génomique de MUC2, dont l'ADNc correspondant recouvre les aa 264 à 348 (Figure 35). La comparaison des séquences génomiques et de l'ADNc publié de *MUC2* permet de placer sans ambiguïté les 2 introns correspondants aux introns 7 et 8 de *MUC5B*. Ces 2 introns ont, dans les gènes *MUC2* et *MUC5B*, les mêmes positions.

III.1.2.3 Tailles des exons et des introns

La région 5' du gène *MUC5B* comprend 29 exons dont les tailles sont comprises entre 44 et 262 pb. Les introns ont des tailles comprises entre 87 pb et environ 1,7 kb. Les introns ont été entièrement séquencés à l'exception des introns 11 et 24 dont on connaît les séquences sur respectivement 154 et 1435 pb. Leurs tailles (environ 415 pb et 1700 pb) ont été estimées par PCR en utilisant des oligonucléotides introniques (Figure 34). Nous avons effectué des PCR sur le cosmide BEN1 en utilisant 2 couples d'amorces qui nous ont permis de préciser que dans les 2 cas, une séquence d'environ 250 pb reste à déterminer : dans le couloir A (Figure 34), on observe un produit d'amplification de 1,5kb or nous connaissons la séquence entre les 2 amorces sur 1240 pb. Dans le couloir B, on observe une bande à 1,3 kb or nous connaissons la séquence entre les 2 amorces sur 1060 pb.

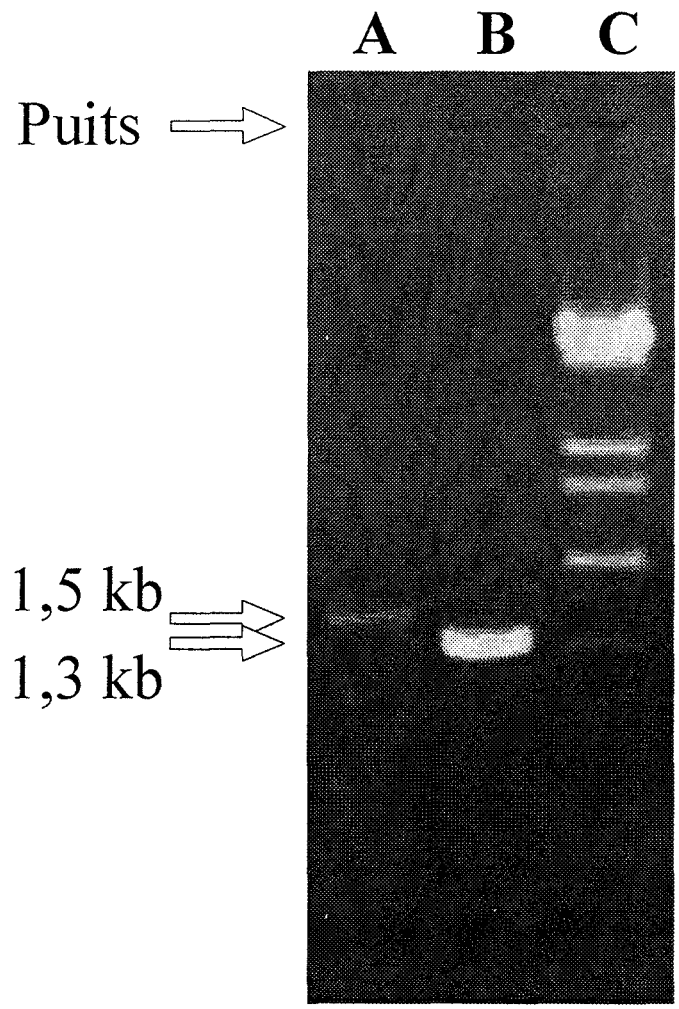


Figure 34 : Détermination des tailles des introns 11 (couloir A) et 24 (couloir B) par PCR sur le cosmide BEN1 en utilisant des oligonucléotides introniques (couloir C : marqueur de tailles 13i)

III.1.2.4 Place des introns et classes

Les gènes *MUC5B* et *vWF* ont 13 introns aux mêmes positions (intron 3, 5-7, 9, 14, 16, 17, 21, 23-25, 27 dans *MUC5B*). A l'exception de l'intron 6, les introns conservés sont de classes identiques (Figure 35).

III.1.2.5 Séquences introniques particulières

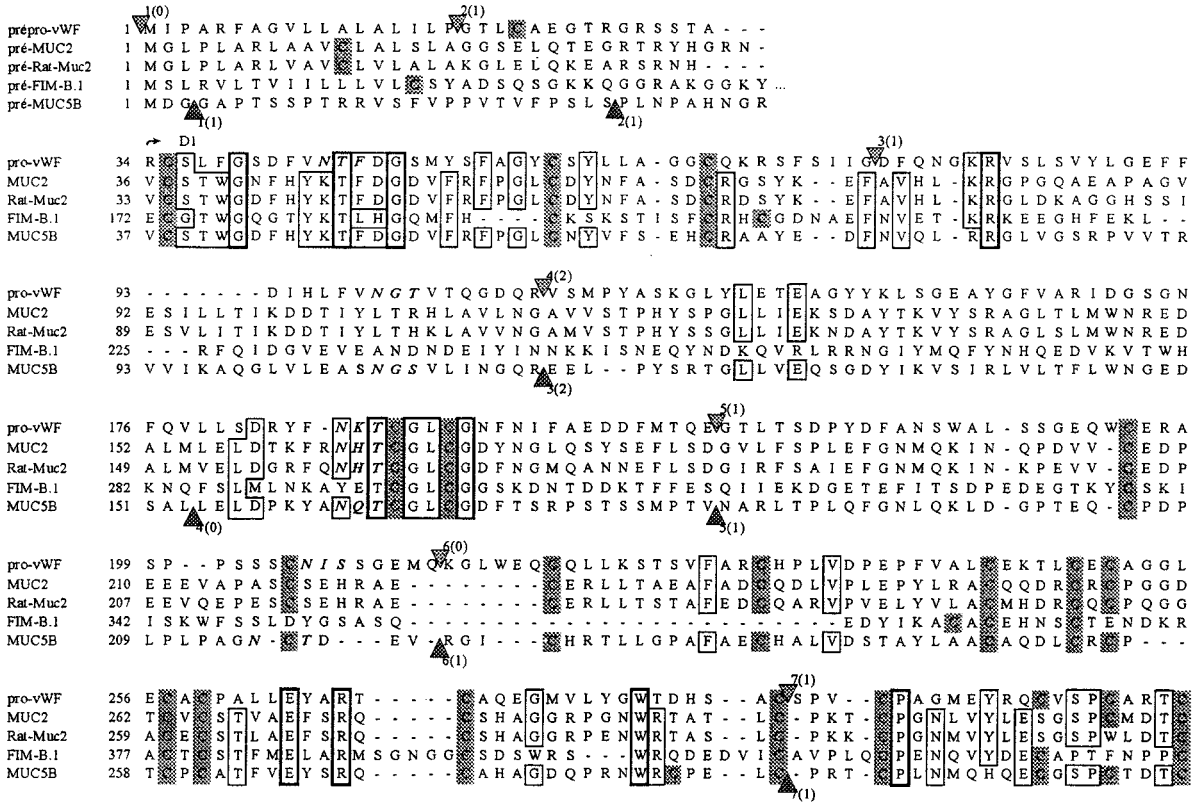
Nous avons noté dans l'intron 9 un motif de 45 pb répété 2 fois et presque parfait (91,1% d'homologie). L'intron 24 a quant à lui presque 6 répétitions imparfaites en tandem d'un motif de 15 pb YAGGTGGGCAGRWBR. Cette séquence n'est pas homologue aux séquences enregistrées dans les banques de données EMBL et GenBank. Il sera intéressant de chercher un éventuel polymorphisme VNTR de cet intron.

III.1.3 Peptide déduit

III.1.3.1 Séquence peptidique

La composition peptidique de la région amino-terminale de MUC5B (Tableau XII) est très riche en résidus Cys (9,6%). Le taux de Ser + Thr + Pro est quant à lui peu élevé (S+T+P=21,5%)

L'alignement des séquences peptidiques déduites du pro-vWF, MUC2, rat-Muc2, FIM-B.1, MUC5AC (clone HGM-1) et MUC5B est montré sur la figure 35. Les régions amino-terminales de ces peptides sont homologues entre elles. Presque tous les résidus Cys (120 résidus) de la région amino-terminale de MUC5B (1286 aa),



.../...

Figure 35 : Peptide déduit de la région amino-terminale de MUC5B et comparaison avec les régions amino-terminales du pré-proWF, de MUC2, rat-Muc2, FIM-B.1 et MUC5AC (clone HGM-1)

Les triangles représentent les introns dans le vWF et MUC5B. Chaque intron est numéroté selon Mancuso *et al.* (Mancuso *et al.*, 1989) et selon le tableau XI. Entre parenthèses figurent les classes des introns. Les résidus Cys sont grisés. Les sites potentiels de *N-glycosylation* sont en italique. Les aa les plus conservés sont encadrés en gras ; les aa un peu moins conservés sont encadrés.

Les domaines D1, D2, D' et D3 du pro-vWF sont délimités par les flèches. Le site de clivage KRS du pro-vWF est souligné en gras.

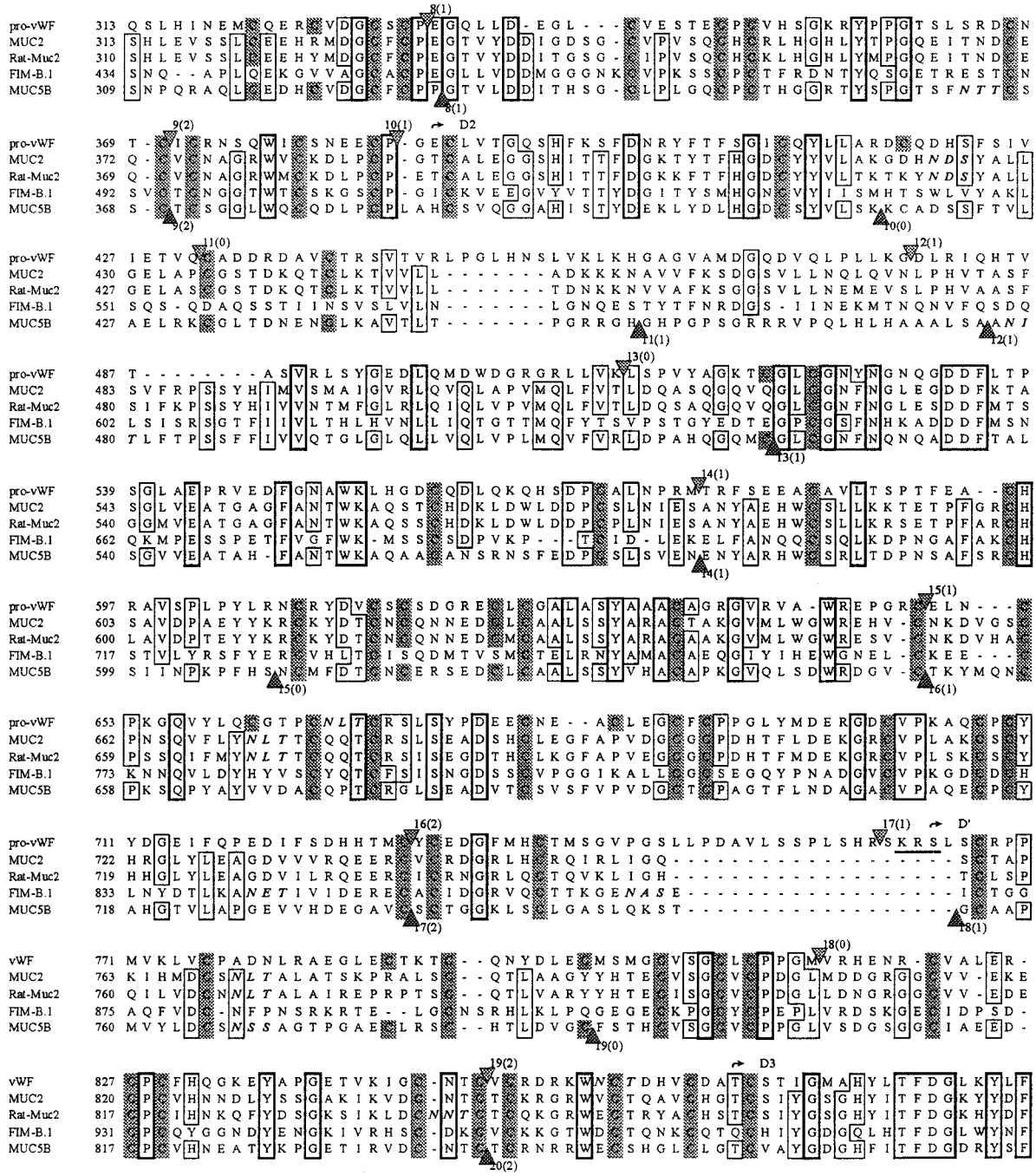


Figure 35 (suite)

.../...

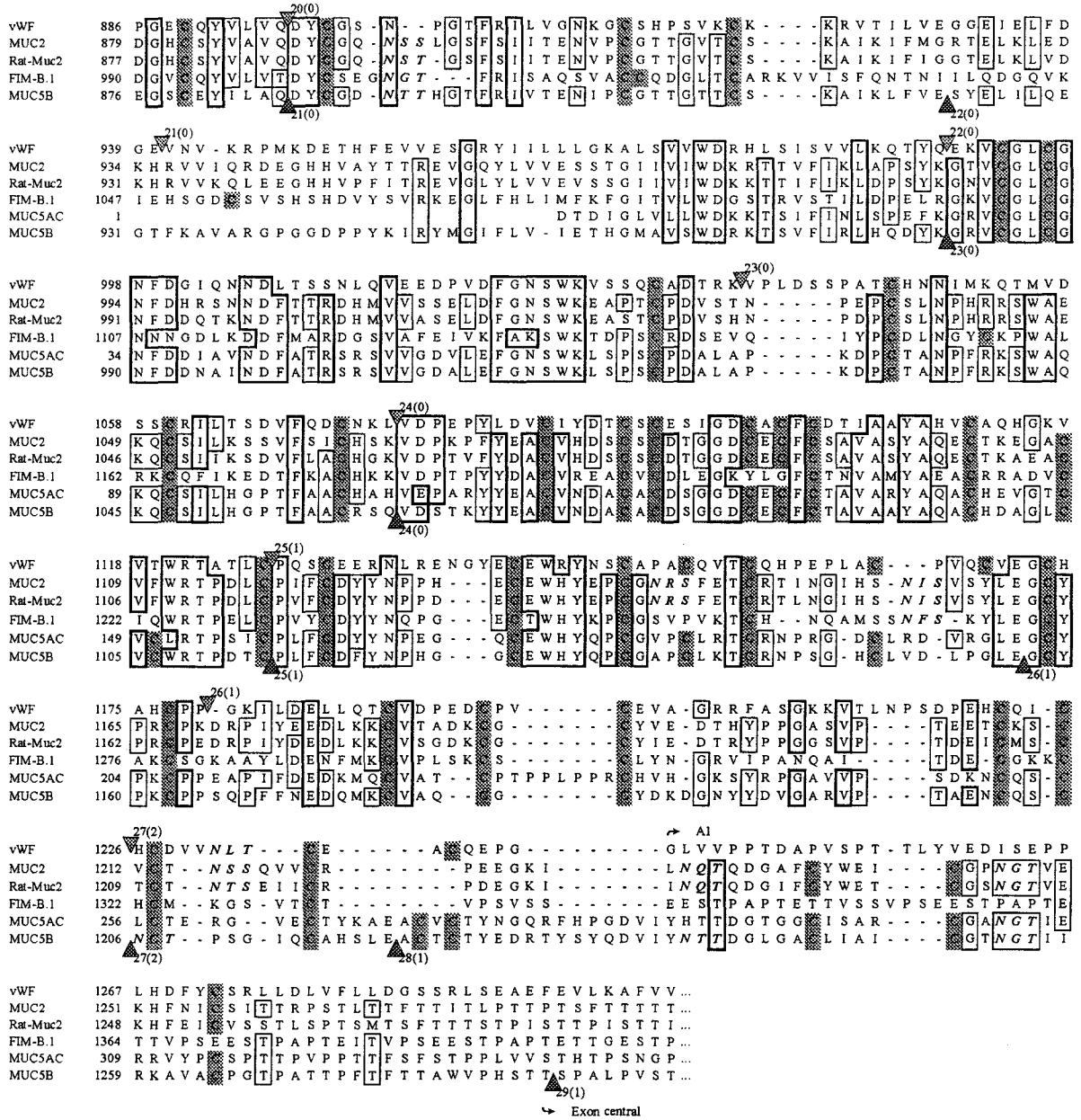


Figure 35 (suite)

Tableau XII. Composition peptidique des différentes régions de MUC5B.
Le nombre de résidus (Nb) et les pourcentages sont indiqués pour chaque région.

aa	Régions							
	Amino-terminale		Centrale		Carboxy-terminale		**Total**	
	Nb	%	Nb	%	Nb	%	Nb	%
Ala	100	7,8	315	8,8	55	6,8	470	8,3
Arg	58	4,5	105	2,9	36	4,5	199	3,5
Asn	52	4,0	40	1,1	35	4,3	127	2,2
Asp	74	5,8	37	1,0	37	4,6	148	2,6
Cys	120	**9,3**	72	**2,0**	82	**10,1**	274	**4,8**
Gln	49	3,8	76	2,1	42	5,2	167	2,9
Glu	50	3,9	98	2,7	38	4,7	186	3,3
Gly	109	8,5	226	6,3	59	7,3	394	7,0
His	40	3,1	69	1,9	23	2,8	132	2,3
Ile	31	2,4	73	2,0	20	2,5	124	2,2
Leu	96	7,5	184	5,2	47	5,8	327	5,8
Lys	34	2,6	47	1,3	16	2,0	97	1,7
Met	11	0,9	43	1,2	12	1,5	66	1,2
Phe	50	3,9	51	1,4	25	3,1	126	2,2
Pro	88	**6,8**	400	**11,2**	77	**9,5**	565	**10,0**
Ser	93	**7,2**	485	**13,6**	50	**6,2**	628	**11,1**
Thr	96	**7,5**	1013	**28,4**	68	**8,4**	1177	**20,8**
Trp	14	1,1	46	1,3	8	1,0	68	1,2
Tyr	41	3,2	32	0,9	19	2,4	92	1,6
Val	80	6,2	158	4,4	59	7,3	297	5,2
Tot	1286	100	3570	100	808	100	5664	100

ont une position généralement très conservée avec le pro-vWF. Comme pour la région carboxy-terminale, des séquences peptidiques, comme l'hexapeptide FGNSWK (aa 1021 à 1026 dans le vWF) sont retrouvées dans les régions amino-terminales des mucines citées ci-dessus. On retrouve le motif CGLCG 3 fois dans MUC5B et dans le pro-vWF.

La séquence peptidique prédite de MUC5B ne contient pas de peptide signal. Nous avons pris soin de vérifier plusieurs fois les séquences génomiques et d'ADNc afin de contrôler qu'il n'y avait pas d'erreur qui pourrait modifier le cadre de lecture. Une dernière vérification de séquence a été réalisée par séquençage automatique. Nous avons appelé domaine 5B le motif de MUC5B en amont du domaine D1.

III.1.3.2 Le domaine MUC11p15

La séquence peptidique comprise entre les introns 28 et 29 de MUC5B (exon 29) correspondant au domaine compris entre les domaines D3-like et central de MUC5B, n'a pas de similarité avec celle de vWF. Une région analogue existe également dans MUC2 et MUC5AC. Les similarités entre ces peptides dans cette région sont plus faibles que pour les régions D du pro-vWF. Nous avons appelé ce domaine MUC2/5AC/5B car il a été retrouvé dans MUC2, MUC5AC et MUC5B. Par ailleurs, un domaine est intercalé aussi entre les domaines D-like et la région répétitive de FIM-B.1. Cette région, plus grande que le domaine MUC2/5AC/5B, n'a pas de similarité de séquence avec ce domaine.

III.1.3.3 Les sites potentiels de *N*-glycosylation

Les régions amino-terminales de MUC5B et MUC2 comportent 10 sites potentiels de *N*-glycosylation (notés en italique sur la figure 35). Cinq de ces sites ont les mêmes positions entre ces 2 peptides (aa 162, 767, 891, 1238 et 1254 dans

MUC5B). Le site 162 est conservé également dans le pro-vWF et le site 891 est conservé dans FIM-B.1. Bien que les domaines MUC2/5AC/5B de MUC2, MUC5AC et MUC5B sont peu similaires entre eux, on remarquera que les 2 sites potentiels de *N*-glycosylation 1238 et 1254 (dans MUC5B) de cette région sont retrouvés dans MUC2 et MUC5AC.

III.2 Discussion

III.2.1 Clonage de l'extrémité 5' de *MUC5B*

Les expériences de RACE-PCR et d'extension d'amorce suggèrent que nous avons cloné la totalité de l'ADNc 5' de *MUC5B*. De plus, l'extrémité 5' du transcrit de MUC5B que nous avons présentée a une taille presque identique à celles de MUC2 et du pro-vWF. Enfin, la séquence nucléotidique gcacaccATGG qui borde le putatif codon d'initiation de MUC5B, est une séquence consensus (gcc(a/g)ccATGG ; Kozak, 1987). Nous n'avons cependant pas trouvé de peptide signal N-terminal hydrophobe, alors qu'il en existe un dans MUC2 et le pro-vWF. Plusieurs protéines membranaires et sécrétées ne possédant pas de séquence signal reconnaissable ont été publiées. Ces peptides gagnent probablement quand même le réticulum endoplasmique (Wen *et al.*, 1992). Il nous faudra vérifier que nous avons bien l'extrémité 5' non traduite par de nouvelles expériences de RACE-PCR et par une expérience de protection à la RNase.

Les similitudes entre les séquences de la région 5' de *MUC5B* avec *MUCY* sont autant de difficultés supplémentaires à l'étude des mucines localisées en 11p15. Nous avons vu que par RACE-PCR et RT-PCR utilisant des oligonucléotides, a priori, spécifiques de MUC5B et utilisant de l'ARN de différentes muqueuses, il est possible de cloner des ADNc d'un gène différent de *MUC5B*. De même, le clone 253/286 issu du gène *MUCY* a été utilisé au laboratoire pour cribler une banque cosmidique. Un clone a été isolé et partiellement caractérisé. Ce clone a été identifié comme appartenant à *MUC5B*.

III.2.2 Jonction atypique de l'intron 8

De tous les introns de MUC5B, seul l'intron 8 a une séquence atypique concernant son site donneur puisque le gt de l'extrémité 5' de l'intron est substitué par un gc (Tableau XI). Cette jonction a été trouvée dans de nombreux gènes (Jackson, 1991). Des études *in vitro* ont montré que cette substitution est la seule qui permet, lors de l'épissage, le clivage correct de l'intron. La seule conséquence de cette substitution semble la vitesse d'épissage qui est alors plus lente (Aebi *et al.*, 1987).

III.2.3 Organisation en domaines de la région amino-terminale de MUC5B

La partie amino-terminale de MUC5B peut être divisée en plusieurs domaines par analogie au pro-vWF (Figure 36). Les résidus Cys ont des positions particulièrement bien conservées.

Cette région comprend 3 domaines D homologues (D1, D2 et D3) et le domaine D tronqué (domaine D') situé entre les domaines D2 et D3. Ces domaines sont nécessaires à la multimérisation du pro-vWF (Wise *et al.*, 1988). Cette étape de multimérisation est indépendante de la dimérisation. De plus, la multimérisation par les domaines D1 et D2 joue un rôle important dans le stockage de la molécule (Wagner *et al.*, 1991).

Presque tous les résidus Cys sont conservés entre le pro-vWF, MUC2, FIM-B.1 et MUC5B mais les similarités des séquences sont peu prononcées. Cependant, les 2 séquences peptidiques TCGLCG (domaine D1) et VCGLCGN (domaine D3) sont parfaitement retrouvées dans les 4 molécules. Les cystéines viscinales de ces 2 motifs ont été retrouvées dans le site actif de PDI (« protein disulfide isomerase »). Il a été suggéré que cette séquence joue un rôle primordial dans la multimérisation du

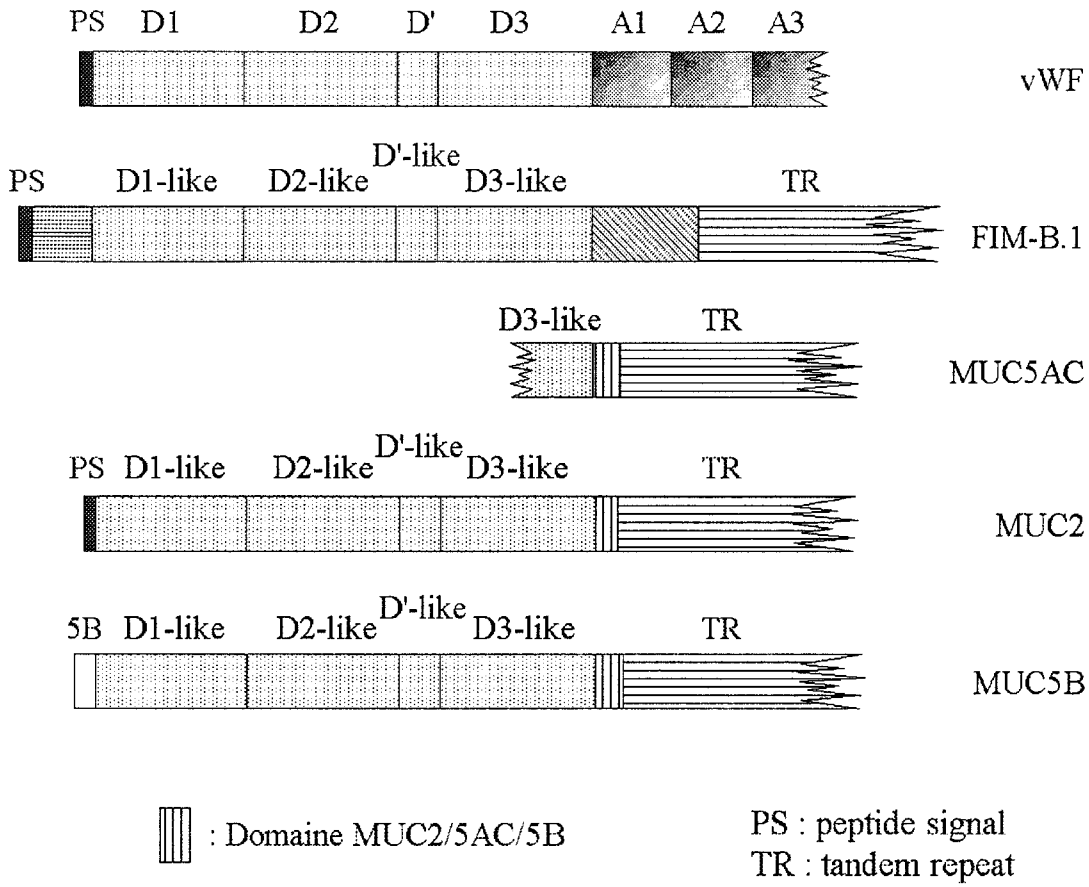


Figure 36 : Organisation peptidique comparée des régions amino-terminales du pro-vWF, de FIM-B.1, de MUC5AC, de MUC2 et de MUC5B

pro-vWF (Mayadas *et al.*, 1992). Il est probable que les mucines qui possèdent ces domaines D multimérisent de la même manière que le pro-vWF. Il a été montré que le clivage du pro-vWF (séquence RSKR↓S entre les domaines D2 et D') n'est pas essentiel à la formation des multimères (Verweij *et al.*, 1988). Un site potentiel de clivage a été relevé dans FIM-B.1 (séquence SRKR↓T ; aa 825 à 829 ; Joba *et al.*, 1997) mais aucune séquence similaire n'est retrouvée dans MUC5B ou MUC2.

Nous avons individualisé un domaine MUC2/5AC/5B. Nous avions de la même façon individualisé un petit domaine intercalé entre le domaine central de MUC5B et la région carboxy-terminale homologue à la région caboxy-terminale du vWF. Ce domaine que nous avons appelé MUC11p15 (Desseyn *et al.*, 1997c) est retrouvé uniquement dans MUC2 et MUC5AC juste en aval des régions centrales. Comme pour le domaine MUC2/AC/5B, les domaines MUC11p15 sont très peu homologues entre eux. Ceci suggère que le gène ancestral des mucines localisées sur le chromosome 11 possédait ces 2 domaines flanquant le sous-domaine Cys (108 aa, 10 résidus Cys) retrouvé ensuite 2 fois dans MUC2, au moins 6 fois dans MUC5AC et 7 fois dans MUC5B.

IV CARACTERISTIQUES STRUCTURALES DE MUC5B

Le gène *MUC5B* est constitué de 48 exons répartis sur 36,5 kb. Les exons ont des tailles qui varient de 32 pb à 10.713 pb. Les 47 introns ont des tailles comprises entre 87 pb et environ 1700 pb.

Le transcrit de *MUC5B* a une taille de 17.592 pb (plus la queue polyA$^+$) et code une apomucine de 5664 aa. Nous avons subdivisé cette molécule en 3 parties :

- la région amino-terminale : 1286 aa ; 10 sites potentiels de *N*-glycosylation ; 29 exons répartis sur environ 15,2 kb.
- la région centrale : 7 sites potentiels de *N*-glycosylation ; 3570 aa codés par un exon unique de 10713 pb.
- la région carboxy-terminale : 808 aa ; 15 sites potentiels de *N*-glycosylation ; 18 exons répartis sur 10,6 kb.

La région centrale de MUC5B, domaine mucine, est composée de l'alternance de plusieurs sous-domaines :

- un domaine de 108 aa riche en résidus Cys (10 Cys) et appelé sous-domaine Cys. Ce sous-domaine est présent 7 fois dans MUC5B, au moins 6 fois dans MUC5AC, 2 fois dans MUC2, au moins 2 fois dans Muc5ac (souris), 2 fois dans Rat-Muc2 et au moins une fois dans PGM-2A (porc)
- un domaine riche en résidus Ser, Thr et Pro de 111 aa, retrouvé 4 fois dans MUC5B et appelé R-end
- 5 sous-domaines de type mucine, constitués de répétitions en tandem (11, 17 ou 23 fois) du motif imparfait de 87 pb.

Les régions amino- et carboxy-terminales de MUC5B sont riches en résidus Cys et sont similaires, comme pour MUC2 et probablement pour MUC5AC, aux régions qui flanquent les domaines A (A1-A2-A3) du pro-vWF (Figure 37).

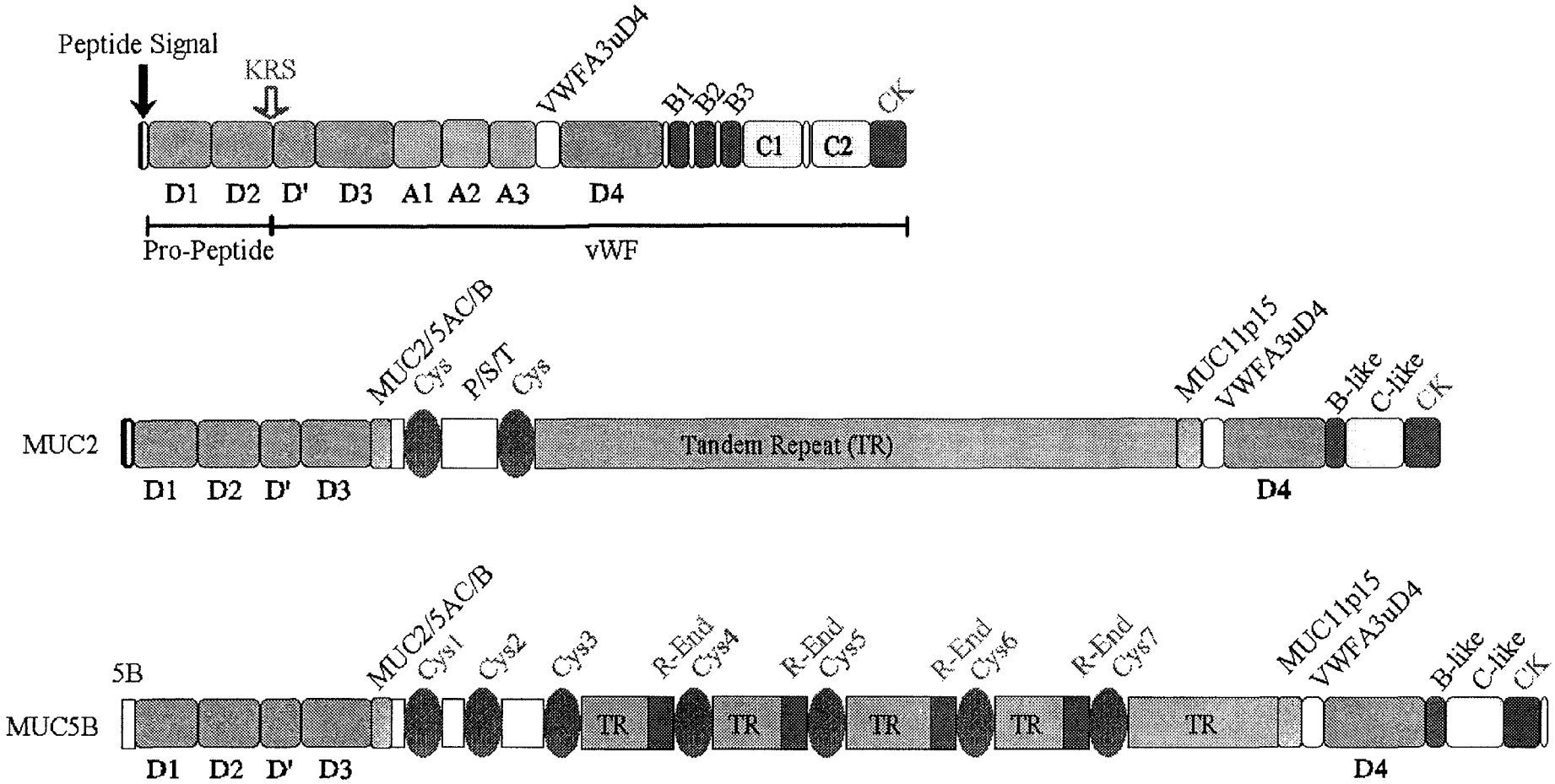


Figure 37 : Organisation peptidique comparée du pro-vWF, de MUC2 et de MUC5B

Discussion & Perspectives

DISCUSSION ET PERSPECTIVES

I MUC5B : MODELE D'ETUDE DES GENES DE MUCINES DU CHROMOSOME 11

I.1 Organisation génomique des mucines du chromosome 11

C'est la première fois qu'une organisation génomique d'un gène codant une « mucine sécrétée formant le gel de mucus » est résolue.

Les 3 gènes de mucines localisés en 11p15.5 *MUC2*, *MUC5AC* et *MUC5B* sont très homologues entre eux et il est probable que les organisations génomiques de *MUC2* et *MUC5AC* soient très similaires à celle de *MUC5B*.

Les travaux récents réalisés sur *MUC6*, localisé lui aussi en 11p15.5, montrent que ce gène est moins similaire aux 3 autres gènes (absence dans MUC6 des domaines D4, B et C) bien qu'il possède des caractéristiques communes (en particulier, présence du domaine CK).

Ainsi, notre étude de l'organisation génomique de *MUC5B* permettra peut-être de localiser plus rapidement les introns des autres gènes de mucines localisés en 11p15 et de leurs homologues animaux. Il sera alors intéressant de séquencer entièrement les introns de ces gènes équivalents aux introns A et G (1er et 7ème intron en aval de la partie répétitive) de *MUC5B* afin de voir s'ils possèdent eux aussi des motifs répétés en tandem. Si c'est le cas, il faudra alors étudier un éventuel polymorphisme VNTR et la relation VNTR / régulation de la transcription.

I.2 Régulation

La séquence quasi complète du gène *MUC5B* a été déterminée. C'est, après MUC1, la seule séquence connue pour un gène de mucine, et donc aussi le premier modèle d'étude de mucine sécrétée.

Il reste à étudier la région promotrice de *MUC5B*, contenue dans le cosmide BEN1. Nous allons en même temps aborder les premières étapes de la recherche des zones régulatrices du gène *MUC5B*. Pour cela, nous nous proposons de chercher les sites hypersensibles à la DNase I. Nous disposons au laboratoire d'une lignée cellulaire n'exprimant pas les gènes de mucines (les cellules HT 29 dérivées d'un cancer colique) et des cellules exprimant certains de ces gènes (cellules HT 29 traitées au méthotrexate), dont *MUC5B*. Nous pourrons ainsi approcher très rapidement les zones régulatrices qu'elles soient en 5', 3' ou même à l'intérieur du gène.

Parallèlement à la caractérisation du facteur nucléaire spécifique des cellules muco-sécrétantes qui interagit avec le motif répété de 59 pb de l'intron G (Pigny *et al.*, 1996b ; Desseyn *et al.*, 1997a), il faut étudier la conséquence du polymorphisme VNTR de cet intron sur la régulation de la transcription. Pour cela, après avoir sous-cloné le promoteur minimal de *MUC5B* en amont d'un gène rapporteur, nous sous-clonerons en aval de ce gène rapporteur des fragments de l'intron G comportant 5 à 9 répétitions du motif de 59 pb. On mesurera alors l'éventuelle augmentation ou diminution de transcription que l'on pourra relier à ce polymorphisme de type VNTR.

L'étude de la régulation de *MUC5B* répondra en partie à la question d'importance : les gènes de mucines, et en particulier ceux du cluster en 11p15, sont-ils soumis à une régulation concertée ? Ces travaux devraient nous permettre de comprendre les fonctions exactes des mucines dans la cancérogenèse et les mécanismes permettant d'augmenter ou de diminuer très nettement l'expression de ces glycoprotéines dans les tumeurs. Les connaissances apportées par l'étude de l'expression et de la régulation de *MUC5B* permettront une meilleure compréhension

des anomalies du mucus, comme la surexpression des mucines observée lors de la genèse des cristaux biliaires, les hypersécrétions etc...

I.3 Polymorphisme VNTR de MUC5B : intronique ou dans la région codante ?

Il est généralement admis que les mucines présentent dans leur région répétitive, codant la région riche en résidus Ser et Thr, un polymorphisme de type VNTR. Qu'en est-il aujourd'hui concernant *MUC5B* ?

Les RT-PCR effectuées en utilisant des oligonucléotides de l'exon central ou des oligonucléotides qui bordent cet exon central l'ont été sur de l'ARN provenant de différents individus. Par cette méthode, aucun polymorphisme de type VNTR n'a pu être visualisé. Le nombre de répétitions en tandem du motif de 87 pb est donc très peu ou pas variable. Ceci a par ailleurs été confirmé dans notre laboratoire par la méthode de Southern blot en utilisant la sonde répétitive de *MUC5B* JER57 sur une cinquantaine d'échantillons d'ADN hydrolysés par l'enzyme de restriction *Bgl*II (Pigny, 1997 ; Figure 38 (seuls 8 individus sont présentés)). Sur la cinquantaine d'individus testés, tous ont un fragment *Bgl*II/*Bgl*II d'environ 17 kb révélé par la sonde JER57. Cette bande correspond à l'allèle du cosmide BEN2. Deux individus ont une bande supplémentaire : l'un à environ 20 kb (Figure 38, couloir A), l'autre à environ 15 kb (Figure 38, couloir B).

Ce polymorphisme n'est probablement pas un polymorphisme de restriction. En effet, des expériences identiques menées sur les mêmes échantillons d'ADN mais en utilisant cette fois l'enzyme de restriction *Hin*dIII donnent un profil semblable à celui observé sur la figure 38 mais décalé vers des tailles plus élevées (le fragment *Hin*dIII/*Hin*dIII du cosmide BEN2 a une taille de 25kb).

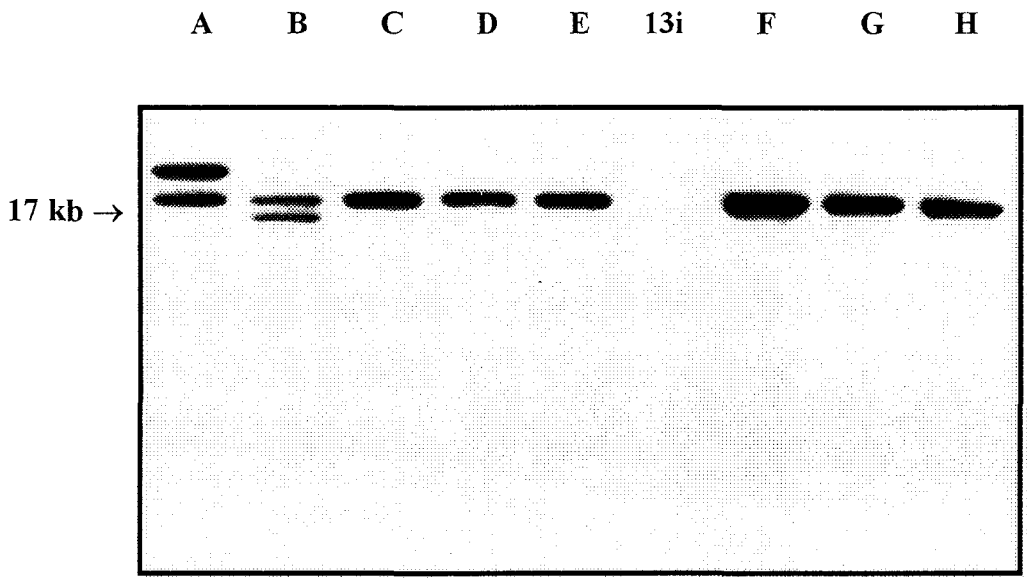


Figure 38 : Polymorphisme *Bgl*II révélé par la sonde JER57 (Pigny, 1997) (13i : marqueur de tailles)

Ces sites de restriction, *Bgl*II et *Hin*dIII, n'encadrent pas exactement l'exon central de *MUC5B*. Les sites *Bgl*II sur le cosmide BEN2 sont situés respectivement à 3,5 et 2,9 kb en amont et en aval de l'exon central. Les sites *Hin*dIII sur le cosmide BEN2 sont situés à 0,5 et 14,5 kb respectivement en amont et en aval de l'exon central. Il faut donc maintenant étudier si les introns compris dans les fragments *Bgl*II/*Bgl*II de 17 kb ou *Hin*dIII/*Hin*dIII de 25 kb ont un polymorphisme de type VNTR qui pourrait expliquer en partie le polymorphisme observé par Southern blot avec la coupure *Bgl*II ou *Hin*dIII. L'intron A est un bon candidat, car il est le seul intron du fragment *Bgl*II/*Bgl*II ayant des répétitions en tandem (23 x 20 pb dans le cosmide BEN2). Grâce à des oligonucléotides choisis de part et d'autre de la région répétitive de cet intron, nous allons mettre au point un protocole d'amplification génomique par PCR afin de calculer le nombre de répétitions en tandem chez plusieurs individus. Nous mènerons parallèlement une étude identique avec l'intron G (compris dans le fragment *Hin*dIII/*Hin*dIII) afin de compléter les résultats précédemment exposés et afin de vérifier que le faible polymorphisme observé pour le fragment *Rsa*I/*Sac*II (fragment contenant l'intron G dans son entier et étudié dans l'article soumis) est dû à un polymorphisme VNTR, comme nous l'avons suggéré, et non à un polymorphisme de restriction.

Quels que soient les résultats que nous obtiendrons, l'éventuel polymorphisme VNTR de MUC5B étudié par la sonde JER57 est très faible comparé par exemple à celui de MUC2 (51 à 115 répétitions du motif élémentaire). Ceci ne fait pas cependant de *MUC5B* une exception dans le monde des mucines puisque le polymorphisme du petit gène MUC7 (~10 kb) est lui aussi très réduit (5 ou 6 répétitions ; Biesbrock *et al.*, 1997).

Il est souvent admis que le nombre de répétitions nucléotidiques en tandem de la région codante des mucines varie d'un individu à l'autre. Certains auteurs font de ce polymorphisme VNTR un critère de définition des mucines (voir par exemple les revues de Devine *et al.*, 1992 ; Van Klinken *et al.*, 1995). Ceci amenait à penser que la taille de l'axe peptidique n'est pas cruciale à la fonction de la molécule (Gendler *et al.*, 1995), or il semble bien que le critère de polymorphisme ne convienne ni à MUC7, ni à

MUC5B. Nous pouvons conclure qu'une mucine ne présente pas systématiquement de polymorphisme de type VNTR dans ses régions codantes.

L'absence de polymorphisme VNTR se conçoit plus aisément pour un petit gène comme MUC7 (c'est-à-dire petite région centrale et donc petit nombre de répétitions codantes). Il est, par contre, beaucoup plus difficile d'admettre l'absence d'un tel polymorphisme sur de très grandes régions construites de motifs répétés en tandem. Il doit donc probablement y avoir un "phénomène" qui préserve MUC5B d'une augmentation ou au contraire d'une diminution du nombre de répétitions. La taille de la région répétitive de MUC5B est certainement un paramètre important dans son rôle biologique. Il n'en est pas exactement de même pour les 3 autres gènes du cluster MUC en 11p15.5. En effet, MUC6 est, de ces 4 gènes, le plus polymorphique (VNTR), puis vient ensuite MUC2, MUC5AC et enfin MUC5B. En comparant les 4 gènes de mucines localisés en 11p15 on remarque que cette différence entre les 4 gènes peut être reliée à la présence ou non des régions codant les motifs que nous avons dénommés sous-domaines Cys (« Cys-subdomains »). Ces domaines ont été retrouvés 2 fois dans MUC2, au moins 6 fois dans MUC5AC et exactement 7 fois dans MUC5B. Il n'en n'existe probablement pas dans MUC6. Ainsi, moins il y a de ces sous-domaines, plus le gène correspondant montre un polymorphisme important de type VNTR dans sa région centrale (Vinall *et al.*, soumis). On peut aussi envisager que les 4 régions appelées R-end, très homologues entre elles, et trouvées uniquement jusqu'à présent dans MUC5B, accentuent ce "phénomène" de stabilité génétique, ce qui expliquerait le polymorphisme VNTR plus marqué pour MUC5AC que pour MUC5B. Ceci pourrait aussi expliquer que les mucines, dont les gènes ne sont pas localisés sur le chromosome 11, présentent toutes un polymorphisme VNTR, car elles ne possèdent pas, dans l'état actuel de nos connaissances, les 2 sous-domaines Cys et R-end.

II LES GRANDS ARNM ET LES GRANDS EXONS

Le transcrit de *MUC5B* a une taille de 17,6 kb. Ceci est en bonne corrélation avec les études de Northern blot menées au laboratoire qui ont montré un unique transcrit pour MUC5B d'une taille supérieure à 15 kb. Un transcrit aussi grand, comme le sont ceux des autres mucines (Debailleul, 1997), est atypique.

II.1 Les grands ARNm

Les grands ARNm, de plus d'une dizaine de kb, sont extrêmement rares. D'après Noara *et al.* (Noara *et al.*, 1987), des acides nucléiques monobrins de plus d'une vingtaine de kb sont trop instables pour exister dans la cellule. Quelques grands ARNm codent des protéines de structure comme par exemple la twitchine (Benian *et al.*, 1989) qui a un ARNm de 23 kb, ou encore la filaggrine (Rothnagel *et al.*, 1990). Les gènes correspondants codent des peptides qui possèdent, comme pour les mucines, des motifs répétés en tandem. Par exemple, la filaggrine a un motif élémentaire de 250 aa (750 pb) répété au moins 20 fois.

II.2 Les grands exons

Lors de nos études sur l'exon central de *MUC5B*, nous avons été amenés à nous demander s'il existait d'autres grands exons dans les génomes d'eucaryotes.

Dans la littérature, bon nombre d'auteurs qui ont étudié de grands transcrits pensent souvent avoir caractérisé en même temps un énorme exon. Cependant, pour la plupart de ces gènes, aucun élément de preuve indiscutable n'a pu être rapporté, car ces gènes codent presque toujours des motifs répétés en tandem. La taille des fragments génomiques portant ces motifs est alors estimée par Southern blot avec des coupures

de restriction partielles et totales. Cette méthode a l'inconvénient de ne pas pouvoir détecter d'éventuels petits introns ni d'exclure qu'une partie du cadre ouvert de lecture (extrapolé le plus souvent), même contenant des motifs répétés en tandem, ne soit en fait intronique.

Grâce aux technologies actuelles, il semble possible de montrer l'existence d'un énorme exon si et seulement si ce putatif exon possède des motifs répétés mais non parfaitement identiques entre eux. En effet, ceci permet alors d'effectuer des RT-PCR chevauchantes et parallèlement des PCR sur de l'ADN génomique. On peut alors avoir la quasi certitude qu'une grande région est totalement exonique et qu'elle ne possède ni site d'épissage alternatif 5' donneur interne, ni site d'épissage alternatif 3' accepteur interne.

Le fait que MUC5B possède une alternance de plusieurs domaines distincts et qu'un motif donné répété en tandem est rarement 100% homologue à un autre motif (motifs imparfaits) a permis de montrer que la région centrale de MUC5B est codée par une énorme région exonique (pas d'intron) et que cette région exonique de 10,7 kb n'a pas de site d'épissage interne. Ceci est par ailleurs en parfait accord avec les résultats de Northern blot préparé selon la technique décrite par V. Debailleul (Debailleul, 1997) qui montrent pour MUC5B un seul transcrit apparent de très grande taille (>15 kb) que la sonde soit répétitive ou non. On ne perçoit pas, par cette technique, d'épissage alternatif. Enfin, aucune des RT-PCR effectuées avec des oligonucléotides spécifiques de *MUC5B* n'a révélé d'épissage alternatif, ce qui tend aussi à montrer que la région centrale de MUC5B est bien codée par un unique exon de 10.713 pb.

Ceci marque une différence vis-à-vis de nombreuses protéines de structure. L'organisation du gène *Unc-22* de *Caenorhabditis elegans* (Benian *et al.*, 1989) codant la protéine musculaire twitchine, était connue pour avoir le record de taille d'exon (Long *et al.*, 1995). Le gène code un transcrit de 19.092 nucléotides avec une séquence codante de 18.147 nucléotides (6049 aa). Le peptide déduit est essentiellement formé de 2 motifs, l'un retrouvé 31 fois et le second 26 fois selon l'agencement schématisé sur la figure 39. A la différence de MUC5B, mais aussi de

MUC7 et MUC1, les motifs répétés sont codés par 12 exons (sur les 14 que comporte le gène), le plus grand ayant une taille calculée de 9613 pb. Cette taille reste cependant putative, car l'organisation génomique de ce gène a été établie grâce à la méthode du "shot gun". Cette technique consiste à séquencer systématiquement tous les sous-clones issus d'un fragment génomique et à compiler par informatique la séquence grâce à des algorithmes, ce qui, bien évidemment, est une méthode peu sûre dans le cas de séquences répétées.

II.3 Les gènes BR

A propos des grands ARNm et des grands exons, il nous faut ici ouvrir une large parenthèse sur la famille des gènes BR (pour « Balbiani ring »), car les auteurs qui les étudient pensent avoir caractérisé les plus grands exons du règne eucaryote. De plus, cette famille nous semble très intéressante, car gènes et protéines possèdent de nombreux points communs avec les mucines.

Chez la larve du diptère *Chironomus tentans*, les glandes salivaires synthétisent et sécrètent au moins une quinzaine de protéines. Ces molécules, une fois libérées dans l'eau, participent au réseau fibreux (Wellman *et al.*, 1989) qui protège la larve du milieu aquatique hostile et qui lui permet de filtrer la nourriture. Plus d'une dizaine de gènes codant ces protéines ont été partiellement ou totalement caractérisés à ce jour (Galli *et al.*, 1993). Quatre d'entre eux appartiennent à une même famille de gènes, la famille BR. Chacun de ces 4 gènes (BR1, BR2.1, BR2.2 et BR6) code un peptide constitué d'une région centrale étirée d'environ 10.000 aa flanquée en amont par une région amino-terminale d'environ 200 aa et en aval par une région globulaire d'environ 110 aa. Les organisations génomiques de ces 4 gènes sont identiques. La région amino-terminale est codée par les 3 premiers exons et le début du quatrième. Le dernier exon, l'exon 5, contient la région 3' non traduite et code la région carboxy-terminale. L'exon 4 de chaque gène BR a une taille comprise entre 26 et 28 kb ! Cet exon est constitué de la répétition (120 à 150 fois) du motif élémentaire, différent d'un

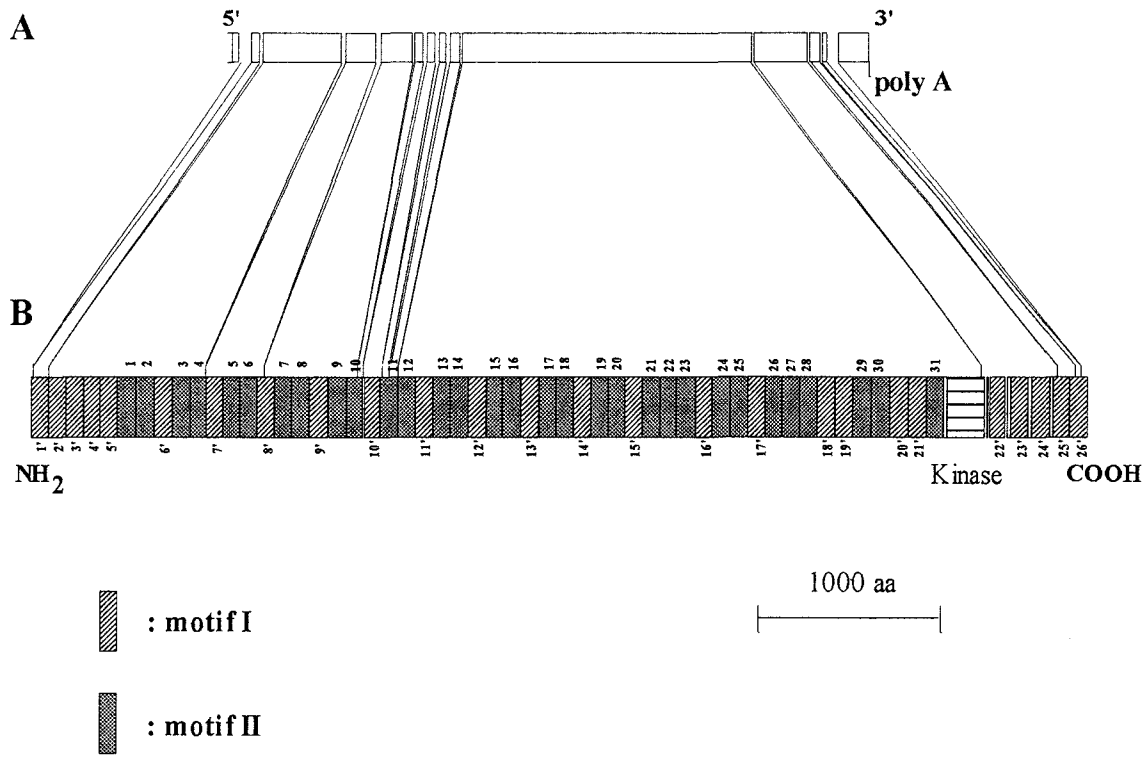


Figure 39 : Représentation schématique du gène *unc-22* (A) et du peptide déduit (B) (Benian *et al.*, 1989)

gène à l'autre. Notons, que comme pour les mucines localisées en 11p15.5, les produits peptidiques des 4 gènes BR sont très homologues entre eux dans leurs régions N- et C-terminales (Paulsson *et al.*, 1992).

Cependant, les tailles des transcrits des gènes BR ont probablement été surévaluées (Noara *et al.*, 1987). Il semble, en fait, que les monomères peptidiques ayant une M_r supérieure à 700 kDa soient extrêmement rares, ce qui correspond à des ARNm ayant des tailles supérieures à 20 kb. Des ARNm d'aussi grande taille sont très certainement instables et doivent être dégradés dans la cellule (Noara *et al.*, 1987).

III PROPOSITION D'UNE NOUVELLE CLASSIFICATION DES MUCINES

Les nouvelles données structurales des apomucines nous laissent entrevoir une nouvelle classification des mucines. Cette classification permettra peut-être de donner une nouvelle définition d'une mucine.

Nos travaux montrent clairement une organisation en modules de MUC5B. Chacun de ces domaines ou modules présente de fortes similitudes avec les domaines retrouvés dans les autres "mucines 11p15" et certaines mucines animales (FIM-B.1, PSM, BSM, Muc5ac, rat-Muc2; Tableau XIII). Parmi tous les domaines de ces mucines, certains sont retrouvés dans le pro-vWF et donc aussi dans les molécules apparentées.

Gènes	Localisation chromosomique humaine	Motif à 10 résidus Cys (108 aa)	domaines du pro-von Willebrand en 5' du TR D1-D2-D'-D3	D4	en 3' du TR B	C	CK	Motif EGF-like ou P
vWF	12p12-pter	0	oui	1	3	2	1	0
MUC5B	11p15.5	7	oui	1	1	1	1	0
MUC5AC	11p15.5	>5	oui[1,2]	1	1	1	1	0
MUC2	11p15.5	2	oui	1	1	1	1	0
MUC6	11p15.5	0 ?	?	0	0	0	1	0
MUC3	7q22	?	?	?	?	?	?	2
MUC4	3q29	?	non[3]	0[3]	0[3]	0[3]	0[3]	?
MUC7	4q13-q21	0	non	0	0	0	0	0
MUC8	12q24.3	?	?	?	?	?	?	?
PSM		? (0)[4]	?	0	0	1	1	?
BSM		0	oui	0	0	1	1	0
FIM-B.1		0	oui	0	0	1	1	0
FIM-A.1		0	non	0	0	0	0	4
FIM-C.1		? (0)[4]	?	0	0	0	0	>5
ASGP		0	non	0	0	0	0	2
Mucsmg		0	non	0	0	0	0	0

? : non déterminé
1: Klomp *et al* ., 1995
2: Buisine, communication personnelle
3: Nollet, communication personnelle
4: Probable d'après les éléments publiés

Tableau XIII. Organisation peptidique comparée de quelques mucines et du pro-vWF (TR : tandem repeat)

III.1 « La famille CK » : les mucines qui forment le gel

D'après le tableau XIII, une nouvelle classification des mucines est possible. La sous-famille 11p15 élargie aux mucines animales citées ci-dessus, comprend uniquement des mucines sécrétées ayant une région carboxy-terminale de type CK. Selon les études menées sur le vWF (Voorberg *et al.*, 1991) et PSM (Perez-Vilar *et al.*, 1996) les mucines et le vWF dimériseraient par cette région riche en résidus Cys. Sur le modèle du pro-vWF, ces mêmes mucines polymériseraient par leur région amino-terminale également très riche en résidus Cys. En effet, les régions amino-terminales des mucines 11p15 déjà connues, sont homologues à la région amino-terminale du pro-vWF.

Ce motif CK est, comme pour la protéine NDP et le vWF, un motif à 11 résidus Cys aux positions très conservées. Cette sous-famille fait elle-même partie de la mégafamille des protéines CK contenant:

- la sous-famille CK à 7 résidus Cys dont un résidu Cys est impliqué dans une liaison intermoléculaire et conservé dans cette sous-famille CK. C'est la famille des TGF-β.

- la sous-famille CK à 6 résidus Cys où la dimérisation est non covalente. Le prototype est le NGF.

III.2 Les mucines solubles

L'organisation génomique de *MUC7* a été entièrement élucidée. La protéine MG2, issue du gène *MUC7*, n'est pas impliquée, directement, dans la formation du gel de mucus, car elle n'établit pas de ponts interchaînes. Ceci n'exclut cependant pas des interactions non covalentes de cette mucine sécrétée avec des mucines qui forment le gel. De plus, cette molécule est fortement *O*-glycosylée et donc très hydrophile. Nous pouvons la qualifier de « mucine soluble ». Il est probable que cette mucine joue essentiellement un rôle dans l'interaction avec d'autres molécules dont les IgAs et avec les bactéries (Tabak, 1995).

III.3 Statut des autres mucines

La plupart des auteurs ont pour habitude de différencier d'un côté les mucines sécrétées, et de l'autre, les mucines membranaires (ou « mucin-like ») dont le prototype est MUC1. Mais que penser maintenant de la forme soluble décrite de MUC1 ? Où classer l'ASGP qui semble être à la fois mucine transmembranaire, soluble et sécrétée ?

III.3.1 « Les sous-familles EGF-like et P-like »

A la lumière des nouvelles données structurales des mucines, il existe peut-être des caractéristiques qui permettent de classer les autres mucines en sous-familles. L'organisation peptidique déduite des séquences d'ADNc a permis de distinguer pour les « mucines 11p15 » (et les homologues animaux) un motif de type facteur de croissance-like (like, car aucune activité de facteur de croissance n'est démontrée) c'est-à-dire un motif CK. Les mucines MUC3, ASGP FIM-A.1 et FIM-C.1 possèdent quant à elles soit 2 motifs au moins de type EGF-like (MUC3 et ASGP), soit de nombreux motifs P (FIM-A.1 et FIM-C.1).

III.3.2 Les autres mucines

Nous avons pu ainsi classer la plupart des mucines humaines et animales en 3 catégories : les mucines CK, les mucines EGF et les mucines P. Nous avons dû classer à part MUC7 qui, contrairement aux autres mucines, est une petite molécule.

Les informations concernant les 2 dernières mucines humaines, MUC4 et MUC8, et nombre de mucines animales, sont insuffisantes actuellement pour les classer parmi les sous-familles que nous avons été amenés à individualiser. Il est cependant probable d'après les travaux menés sur ce gène dans notre laboratoire que MUC4 n'appartienne pas au groupe CK.

Bien que l'organisation peptidique de MUC1 soit parfaitement connue, cette mucine reste ici, comme MG2, à part.

IV Relation Structure / Fonctions des Mucines

Les fonctions biologiques des mucines ont surtout été abordées par des études portant sur la copule glycannique. La molécule MUC5B étirée, linéaire, est constituée d'une alternance de régions fortement *O*-glycosylables (les domaines R et R-end) et de domaines nus, c'est-à-dire peu ou pas glycosylables (sous-domaines Cys, régions amino- et carboxy-terminales). Ces régions dites nues représentent en fait pour MUC5B 50% de l'apomucine. Nous allons discuter plus particulièrement dans ce chapitre des régions nues.

De toutes les mucines humaines, il apparaît jusqu'à présent que seules les 11p15 établissent des ponts disulfures formant le gel de mucus. On peut alors se demander, puisqu'il est toujours difficile de définir ce qu'est une mucine, si de toutes les molécules désignées comme mucines, les mucines 11p15 (et les homologues animaux) ne sont pas « les mucines ». En d'autres termes, les mucines qui ne sont pas en 11p15 sont-elles des mucines ? Participent-elles à la formation du gel de mucus ?

Tout ceci revient finalement à s'interroger sur les fonctions biologiques de ces molécules. Les organisations génomiques de *MUC1* et *MUC7* ainsi que les fonctions des produits issus de ces gènes, respectivement MUC1 et MG2, sont très étudiées. Puisque MG1, étudié depuis 10 ans, est issu du gène *MUC5B*, nous possédons maintenant un excellent modèle d'étude des fonctions des mucines 11p15. Si on compare les domaines des mucines 11p15, on notera que, puisque MUC6 ne possède pas les domaines D4, B et C (Tableau XIII), ces mucines possèdent, par contre, toutes sans doute les modules D1, D2, D', D3, et CK. En effet, il est probable que MUC6 possède une région amino-terminale identique aux 3 autres mucines 11p15 si *MUC6* correspond bien au *MUCX* décrit par Velcich *et al.* (Velcich *et al.*, 1997).

Un certain nombre de questions se posent alors :

- pourquoi certaines de ces mucines (MUC6, PSM, BSM) ne possèdent-elles pas de domaine D4 et B ?

- quel est le rôle du domaine C manquant dans MUC6 ?

- les domaines D1, D2, D' et D3 avec leurs résidus Cys aux positions très conservées sont probablement essentiels aux fonctions des mucines. Le très large domaine amino-terrminal des mucines 11p15 joue-t-il uniquement un rôle dans la polymérisation ?

- quels sont dans ce contexte les rôles biologiques des sous-domaines Cys ?

IV.1 les zones nues

Des travaux font référence aux zones nues des mucines qui correspondent en fait aux régions peu ou pas glycosylables. Ces régions correspondent aux régions amino- et carboxy-terminales ainsi qu'aux sous-domaines Cys (108 aa, 10 résidus Cys), retrouvés chez l'Homme 2 fois dans MUC2, 7 fois dans MUC5B et au moins 6 fois dans MUC5AC.

Plusieurs rôles peuvent être dévolus à ces régions sans que l'on puisse pour l'instant attribuer telle ou telle fonction à une région nue particulière :

- interaction spécifique avec divers composés du mucus comme l'amylase et les PRPs (« Prolin Rich Proteins ») dans la salive (Iontcheva *et al.*, 1997)

- une interaction spécifique entre MG1 et *Haemophilus (para)influenzae* a été décrite par l'équipe de Amerongen (Veerman *et al.*, 1995). Si il est généralement admis que les interactions bactéries/mucines font intervenir la copule glycannique des mucines, ces auteurs ont montré que *H. (para)influenzae* adhère spécifiquement aux zones nues de MG1.

La plupart des interactions décrites entre des bactéries et le mucus salivaire mettent en jeu MUC7. En effet, il a été rapporté que MG2 interagit avec des souches de streptocoques (Murray *et al.*, 1992). Si MUC7 a un rôle essentiellement dans la clearance bactérienne, il ne faut pas réduire cependant le rôle de MUC5B uniquement à la formation du gel de mucus.

IV.2 Les sous-domaines Cys

Plusieurs équipes ont observé par microscopie électronique (voir par exemple Rose *et al.*, 1984; Sheehan *et al.*, 1986 ; Figure 5) que les mucines s'organisent en réseau. Il est possible que les mucines possédant les sous-domaines Cys forment des liaisons entre elles par ces sous-domaines. Ainsi, les mucines MUC2, MUC5AC et MUC5B auraient la capacité de former des mailles où la taille de chaque maille est fonction à la fois de la distance entre les sous-domaines Cys et de la quantité de mucines sécrétées. Les propriétés rhéologiques du mucus seraient donc influencées par deux paramètres, l'un régulé (la sécrétion), et l'autre non régulé (la distance entre les domaines cystéines). Ce dernier paramètre est invariable d'un individu à l'autre, tout au moins pour MUC2 et MUC5B. En effet, à l'inverse de la région purement répétitive, la région T/S/P entre les deux sous-domaines Cys de MUC2 ne présente pas de polymorphisme VNTR (Toribara *et al.*, 1991). Cette région s'étend sur environ 390 aa. Nous avons par ailleurs vu que MUC5B n'a pas de polymorphisme VNTR dans sa région codante mais que la distance entre les sous-domaines Cys est de 430 ou 585 aa. Les propriétés rhéologiques du mucus sont probablement essentiellement fonction des quantités relatives et absolues de ces mucines.

Sheehan *et al.* avaient probablement visualisé ces mêmes sous-domaines en 1991 (Sheehan *et al.*, 1991) grâce à leur immunsérum dirigé contre de la mucine humaine cervicale réduite. Ils avaient observé une alternance de zones nues (3 ou 4) reconnues par les anticorps polyclonaux (voir Figure 4) et de régions glycosylées. La distance entre ces régions nues est d'un peu plus de 100 nm, ce qui correspond, si on considère qu'un aa d'une région étirée a une taille d'environ 2,5 Å (Van Klinken *et al.*, 1995), à environ 400 aa. Ce nombre est en très bonne corrélation avec les tailles des fragments séparant les sous-domaines Cys de MUC5B qui sont sans doute les régions reconnues par les anticorps et observées par microscopie électronique.

IV.3 La région carboxy-terminale

Nous avons vu que MUC5B multimérise probablement par sa région amino-terminale et dimérise par son domaine CK en C-terminal. Est-ce là le seul rôle de ce domaine ?

IV.3.1 Modélisation du CK de MUC5B

Les motifs CK des mucines ont 11 résidus Cys aux positions invariantes. Ce même motif à 11 résidus Cys a été trouvé dans la protéine NDP. La fonction de cette protéine n'est pas connue mais elle pourrait avoir un rôle de facteur de croissance (Meitinger *et al.*, 1993). La prédiction informatique prévoit que le NDP a une conformation 3-D superposable à la structure tertiaire du TGF-β obtenue par cristallographie.

Nous avons aligné le motif CK de MUC5B avec le TGF-β en tenant compte des résidus Cys et de l'espacement entre ces résidus. Le programme informatique utilisé par Meitinger *et al.* prévoit une succession de 5 feuillets β pour le NDP et le TGF-β concluant que les 2 molécules ont très certainement des structures superposables. En utilisant 2 autres programmes de prédiction de structure, nous avons pu montrer que le TGF-β, comme le CK de MUC5B, ont eux aussi 2 structures superposables. Les 2 molécules, selon ces 2 programmes, sont constituées de la succession de 2 feuillets β, d'une hélice α puis de 2 feuillets β, conformément aux données de cristallographie du TGF-β.

Nous avons ensuite réalisé une première modélisation moléculaire en 3 dimensions du CK de MUC5B en remplaçant, dans le modèle 3-D du TGF-β accessible dans la banque de structures protéiques, les acides aminés du TGF-β par ceux du CK de MUC5B. La conformation tridimensionnelle a alors été optimisée par un programme de minimisation d'énergie afin de trouver la conformation la plus stable

pour le CK de MUC5B. La structure expérimentale du TGF-β et la prédiction de structure du CK de MUC5B sont superposables (Figure 40).

IV.3.2 Spéculations sur la fonction biologique du domaine CK

Puisque le motif CK de MUC5B, et donc probablement les motifs CK des autres mucines, peuvent adopter une conformation semblable à celle du TGF-β, il est justifié de penser que les motifs CK peuvent avoir un rôle de type facteur de croissance par liaison avec un récepteur spécifique qui reste à trouver, ou avec un récepteur des facteurs de croissance. Il peut sembler surprenant qu'une énorme molécule comme MUC5B joue un rôle direct de type facteur de croissance. Il a cependant été montré que MG1 contient 2 sous-unités distinctes reliées entre elles par des ponts disulfures. La plus petite des 2 sous-unités contient des *N*-glycannes, a une M_r de 150.000 dans la salive (Kawagishi *et al.*, 1990) et est identifiée au peptide de liaison. Nous avons vu précédemment que le domaine carboxy-terminal de MUC2 correspond au peptide de liaison intestinal (Gendler *et al.*, 1995). Il est donc probable que la région carboxy-terminale de MUC5B, similaire à la région carboxy-terminale de MUC2, riche en sites potentiels de *N*-glycosylation (15 sur les 32 que possède la molécule) est clivée. Ceci reste à démontrer.

Enfin, ces domaines CK peuvent jouer plus simplement un rôle de modulateur de facteur de croissance comme cela a déjà été démontré pour des protéines de la matrice extracellulaire (Ruoslahti *et al.*, 1991 ; Taipale *et al.*, 1997) telle que la décorine (Hildebrand *et al.*,1994).

L'ensemble des résultats présentés nous permettent d'entreprendre l'étude de la régulation de ce gène. De plus, nous possédons maintenant toutes les informations nucléotidiques qui vont nous permettre de préparer des protéines recombinantes correspondant aux différents domaines de MUC5B et d'étudier ensuite leurs fonctions biologiques.

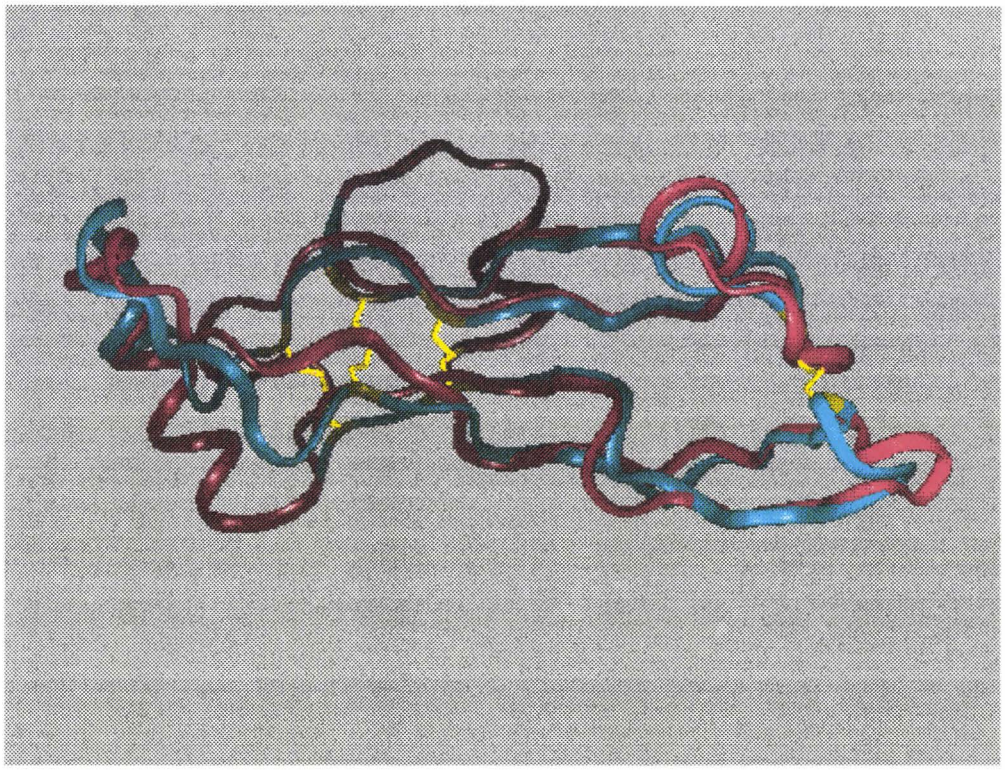


Figure 40 : Superposition de la structure 3-D expérimentale du TGF-β (en rouge) et de la structure prédite du CK de MUC5B (en bleu). Les ponts disulfures sont représentés en jaune.

Appendice Technique

APPENDICE TECHNIQUE

I Préparation des ARN totaux

Le matériel biologique est immergé dans l'azote liquide et y est broyé. On ajoute ensuite 40 ml de tampon GT. Le mélange est homogénéisé à l'aide du «Sorvall» puis filtré sur gaze stérile afin d'éliminer le tissu conjonctif. L'ADN est cassé par des passages successifs à la seringue pour diminuer la viscosité. On ajoute au mélange 16 g de chlorure de césium. Après dissolution, on répartit le mélange dans 6 tubes à la surface d'une solution (3,2 ml) de chlorure de césium 5,7M; EDTA 0,1M pH7,5; DEPC 0,1%. Les tubes sont équilibrés deux à deux par une solution de tampon GT/CsCl (5ml/2g) et centrifugé à 29500 tours/min pendant 15 heures à 18°C. Les culots sont repris par 1 ml de TE 1X-SDS 0,1% puis transférés dans 2 tubes Corex (30 ml). Les ARN sont alors purifiés par 2 volumes de chloroforme - n butanol (4/1, v/v). Les ARN sont ensuite précipités par 0,1 volume d'acétate de sodium 3M pH 5,5 et 2,2 volumes d'éthanol absolu pendant au moins 2 heures à -80°C. Après centrifugation pendant 30 minutes à 10000 tours/min, le culot est séché à l'azote et resolubilisé dans 200 à 500 µl d'eau stérile additionné de DEPC. L'ARN est conservé à -80°C. La concentration de l'ARN est apprécié par la mesure de sa DO à 260 nm : une DO_{260nm} de 0,025 correspond à 1 µg d'ARN par ml. La pureté des ARN est appréciée par le rapport des DO à 260 sur 230 nm qui doit être supérieur à 2 et le rapport des DO à 260 sur 280 nm qui doit être supérieur à 1,75.

II Synthèse d'ADNc

II.1 Méthode d'amorçage aléatoire

De l'ARN poly(A)$^+$ trachéobronchique ou salivaire (0,5 µg) commercial (Clontech) ou de l'ARN total préparé au laboratoire à partir de vésicule biliaire (2 µg) dans un volume de 12,5 µl (H_2O-DEPC) et 1 µl d'un cocktail d'hexanucléotides sont dénaturés par chauffage pendant 2 min à 70°C. Le mélange est ensuite mis dans la glace 5 min puis centrifugé quelques secondes. On y ajoute des dNTPs (1 µl à 10mM chaque), 20 U d'inhibiteur de RNases (0,5 µl), 4 µl de tampon de transcriptase inverse (5X) et 200 U (1 µl) de transcriptase inverse MMLV du kit (1st-STRAND™ cDNA Synthesis Kit; Clontech). Le tout est incubé pendant une heure à 42°C. L'ARN, la polymérase et les DNases sont ensuite détruits par chauffage pendant 5 min à 90°C. Après centrifugation, les 20 µl de solution d'ADNc complétés par 80 µl d'eau-DEPC sont conservés à -80°C. Une PCR contrôle sera effectuée sur 5 µl d'ADNc en utilisant des amorces spécifiques de la G3PDH.

II.2 Méthode RACE-PCR

La RACE-PCR a initialement été décrite par Frohman *et al.* (Frohman *et al.*, 1988). Nous avons essayé différents kits; notre préférence va au kit "5'/3' RACE" (Boehringer) dont le principe a été rappelé dans le chapitre Stratégie.

II.2.1 Synthèse de l'ADNc

Dans un microtube de 0,5 ml sont mélangés puis centrifugés :

tampon de transcription inverse (5X)	4 µl
dNTPs (10 mM chaque)	2 µl
oligonucléotide spécifique (5pmol/µl)	2, 5 µl
ARN poly(A)$^+$ (0,5 µg) ou total (2 µg)	
eau-DEPC	qsp 19 µl
transcriptase inverse AMV (20 U/µl)	1 µl

Le mélange est incubé pendant 60 min à 55°C puis pendant 10 min à 65°C. Il est centrifugé et purifié.

II.2.2 Purification de l'ADNc

Aux 20 µl d'ADNc sont ajoutés 100 µl d'une solution tampon (thiocyanate de guanidine 3M; Tris-HCl 10mM; éthanol 5% (v/v); pH6,6), pour l'adsorption des acides nucléiques sur microcolonne. Le mélange est passé sur un filtre (kit "High Pure Product Purification"; Boehringer) et l'ADNc retenu est lavé plusieurs fois par 0,5 ml d'une solution de lavage (NaCl 20mM; Tris-HCl 2mM; pH7,5; 4 volumes d'éthanol). Après essorage par centrifugation pendant 30 secondes à 13000 x g, l'ADNc est élué par 50 µl d'une solution de Tris-HCl, 10mM pH8-8,3.

II.2.3 Synthèse de la queue poly(A)

Dans un tube de 0,5 ml sont mélangés, sur de la glace, 19 µl d'ADNc purifié; 2,5 µl d'une solution tampon 10X (Tris-HCl, 100mM; $MgCl_2$, 15mM; KCl, 500mM, pH8,3) et 2,5 µl de dATP (2mM). L'ADNc est dénaturé à 94°C pendant 3 min puis remis sur la glace et centrifugé. On ajoute 1 µl de terminale-transférase (10 U) qui va catalyser l'addition de dATP, sans matrice, à l'extrémité 3'OH de l'ADNc. L'incubation se fait à 37°C pendant 10 minutes. La terminale-tranférase est ensuite inactivée par chauffage à 70°C pendant 10 minutes. Le mélange, après centrifugation, est conservé à -80°C. 5µl de ce mélange seront utilisés pour chacune des réactions de PCR.

III Amplification élective

III.1 Principe

La réaction de polymérisation en chaîne (PCR pour "Polymerase Chain Reaction") a été mise au point au milieu des années 80 par K. Mullis (prix Nobel de chimie en 1993). Elle utilise un couple d'amorces qui encadre la région que l'on veut amplifier sélectivement, des dNTPs, de l'ADN monobrin (essentiellement de l'ADNc)

ou double brin (ADN génomique, un produit d'amplification ou un vecteur et son insert) et une polymérase thermostable de type Taq Polymérase et son tampon.

III.2 Les réactifs, le matériel utilisé et généralités

Nous avons utilisé cette méthode pour amplifier de l'ADN à partir de plasmide (moins d'un microlitre de solution de préparation de plasmide, voir § VI), de cosmide (1 à 20 ng par réaction) et de l'ADNc. Les oligonucléotides ont une taille comprise entre 18 à 30 nucléotides. Ils sont synthétisés par Eurogentec (Liège, Belgique) et sont utilisés à des concentrations de 5 pmol/µl. Les amorces sont choisies de façon à ce qu'elles ne puissent pas s'hybrider entre elles et ne puissent pas former d'épingles à cheveux, qu'elles ne contiennent pas de séquences répétées, qu'elles ne puissent pas s'hybrider à plusieurs endroits du matériel à amplifier. Il est préférable que la composition en bases soit équilibrée (éviter les longues répétitions de GC) et que les températures de fusion (Tm) des 2 amorces ne soient pas trop différentes l'une de l'autre. Nous avons utilisé, pour choisir les amorces et calculer les Tm, le programme informatique PCR Plan de PC/Gene. Les réactions de PCR ont été faites sur un volume final de 50 µl avec une solution de dNTPs (Boehringer) à 2,5 mM chaque (5 à 10 µl selon les réactions) dans un thermocycleur Perkin 480. Un corollaire indésirable et inattendu de la puissance d'amplification de la PCR est qu'une contamination même mineure du matériel de départ peut être amplifiée. Nous avons donc préparé les réactions sous une hôte à flux laminaire en utilisant des embouts cotonnés pour les pipettes automatiques. Nous avons aussi fait des témoins négatifs en remplaçant le matériel nucléique par de l'eau. L'enzyme thermostable employée pour les PCR "classiques" est la Taq polymérase de chez Boehringer à 1U/µl ; pour les longues PCR nous avons utilisé un mélange de plusieurs Taq polymérases vendu par Boehringer: kit Expand™ Long Template PCR System.

Pour chaque réaction nous avons effectué un "hot start", c'est à dire que le tube est placé dans l'appareil PCR alors que sa température est supérieure à 80°C.

Les produits de PCR ont soit été purifiés par électrophorèse (voir § IV.1), soit directement sur une matrice de silice (résine Wizard™ PCR Preps DNA Purification System, Promega).

III.3 Réactions

III.3.1 Amplification d'ADN cosmidique

III.3.1.1 Méthode usuelle

Dans un microtube (0,5 ml) on mélange 20 ng de cosmide, 15 pmoles de chaque amorce, 6 à 8 µl de dNTPs, 5 µl de tampon 10 X fourni (Tris-HCl, 20mM; DTT, 1mM; EDTA, 0,1mM; KCl, 0,1M; Nonidet® P40, 0,5% (v/v); Tween® 20, 0,5% (v/v); glycérol, 50% (v/v), pH 8,0) et de l'eau stérile (qsp 48 µl). On recouvre le mélange avec 50 µl d'huile minérale (Sigma) afin d'éviter l'évaporation. Après un "hot start", on injecte sous l'huile 2 unités de Taq polymérase (Boehringer). Après une dénaturation de 2 min à 94°C, on effectue 30 cycles :

95°C: 10 s (dénaturation des brins)

Tm+2°C: 30 s (hybridation)

72°C: 30 s à 2 min (élongation) selon la taille à amplifier

On termine par une élongation finale plus longue de 20 min à 72°C.

III.3.1.2 Amplification de grands fragments: longue PCR

L'Expand™ Long Template PCR System est un mélange d'une Taq polymérase "classique" et de la polymérase Pwo. Ce système nous a permis d'amplifier des fragments d'ADN de plus de 4 kb, et jusqu'à 11 kb. Le mélange est effectué sur la glace dans des microtubes à paroi fine et est composé de 5 µl de tampon 10X ($MgCl_2$, 22,5 mM; Tris-HCl, 500mM; $(NH_4)_2SO_4$, 140 mM; DMSO, 20% (v/v); Tween 20, 1% (v/v)), 10 µl de dNTPs (10 mM chaque), 15 pmoles de chaque amorce, 20 ng de vecteur (cosmide) contenant le fragment à amplifier, 0,75 µl du mélange d'enzyme (2,5 U) et de l'eau stérile (qsp 50 µl).

Après une dénaturation de 2 min à 94°C, on effectue tout d'abord 10 cycles :

94°C: 10 s

Tm+5°C: 30 s

68°C: 45 s à 30 min selon la taille à amplifier

puis les 20 cycles suivants :

94°C: 10 s

Tm+5°C: 30 s

68°C: 45 s à 30 min + 20 s par cycle

On termine par une élongation finale de 20 min à 68°C.

III.3.2 Amplification d'ADNc

III.3.2.1 Minibanque d'ADNc

Les réactions d'amplification d'ADNc synthétisé par la méthode aléatoire utilisent 5 µl d'ADNc, 15 pmoles de chaque amorce, 6 à 10 µl de dNTPs, 5 µl de tampon 10X, 2 U de Taq polymérase et de l'eau stérile (qsp 50 µl).

III.3.2.2 RACE-PCR

Une première réaction d'amplification est réalisée sur 5 µl d'ADNc ayant une queue poly(A) avec 1 µl (12,5 pmoles) d'une amorce dT-anchor, 12,5 pmoles de la seconde amorce antisens et spécifique du transcrit à amplifier (choisie en amont de l'amorce utilisée pour la synthèse de l'ADNc), 1 µl de dNTPs (10mM chaque), 2,5 U de Taq polymérase, 5 µl de tampon d'amplification 10X et de l'eau stérile (qsp 50 µl). Une seconde réaction d'amplification est réalisée ensuite sur 1 µl de la première réaction de PCR diluée au 1/20 dans du TE 1X. Les amorces utilisées sont une amorce antisens spécifique du transcrit choisie en amont de l'amorce précédente et l'amorce "anchor" du kit de RACE-PCR.

IV Clonage des fragments amplifiés

IV.1 Purification des fragments

Deux méthodes de purification ont été utilisées :

IV.1.1 Méthode « GENECLEAN »

Les ADNc sont séparés en gel d'agarose 0,8% contenant du BEt (0,5 µg/ml), et en tampon TEA pour éviter l'action des DNases. Après repérage sous UV grâce aux marqueurs de tailles, les bandes d'agarose d'intérêt sont découpées au scalpel. L'ADN est extrait et purifié à l'aide du kit GENECLEAN II® (Bio 101): l'agarose est liquéfié en ajoutant trois volumes de NaI et en incubant au moins 5 minutes à 50°C. L'ADN est ensuite adsorbé sur la matrice de silice ("Glassmilk") pendant 10 minutes: 5 µl de "Glassmilk" sont ajoutés pour 5 µg d'ADN (ou moins) et 0,5 µl de "Glassmilk" supplémentaire par 0,5 µg d'ADN au delà des 5µg. Après centrifugation, on récupère le culot d'ADN/matrice et on le lave trois fois par des solutions de 300 µl de New Wash dilué (selon les conditions du fournisseur). Après centrifugation, le culot est resuspendu dans 5 à 10 µl de TE 1X et placé 2 à 3 minutes à 50°C. Après 30 secondes de centrifugation à 12000 x g, le surnageant contenant l'ADN est transféré dans un nouveau tube. Une deuxième élution de l'ADN peut être effectuée et permet d'obtenir 20% supplémentaires du matériel élué la première fois. Une partie aliquote est quantifiée après électrophorèse en gel d'agarose contenant du BEt (0,5 µg/ml) en comparaison avec deux marqueurs de tailles moléculaires : le 13i qui est un clone phagique qui, après hydrolyse par l'endonucléase *Eco*RI, libère 8 fragments de 31, 22, 11, 4.3, 3.2, 1.3 et 0.5 kb ; et le marqueur de taille commercial « 100-bp ladder » contenant des fragments d'ADN multiples de 100 pb (Pharmacia Biotech).

IV.1.2 Méthode « Seaplaque »

Les fragments d'ADN à purifier sont déposés dans un gel d'agarose SeaPlaque®GTG® (FMC Bioproducts) 0,8% en tampon borate et contenant du BEt (0,5 µg/ml). Cet agarose a la particularité de se liquéfier à basse température. Après repérage sous UV, les bandes d'agarose d'intérêt sont excisées au scalpel et placées

dans un tube de 1,5 ml. On ajoute 50 µl de TE 1X et 100 µl de tampon de purification de PCR. Les tubes sont mis au bain-marie à 65°C pendant 7 min. On ajoute ensuite au gel liquéfié 0,9 ml de résine (Wizard™ PCR Preps DNA Purification System, Promega). On retourne plusieurs fois les tubes ; l'ADN s'adsorbe sur la matrice. On filtre ensuite les mélanges sur des colonnes (Wizard™ Minicolumns, Promega). Les filtres sont lavés 2 fois par 1 ml d'isopropanol 80% puis essorés par centrifugation pendant 2 min à 13000 x g. On dépose sur les colonnes 50 µl de TE 1X. Après au moins une minute de contact, les produits d'amplification purifiés sont élués par centrifugation pendant 30 s à 13000 x g.

IV.2 Clonage en vecteur plasmidique T/A

IV.2.1 Principe

La Taq DNA Polymérase, comme la plupart des polymérases thermostables, ajoute un nucléotide en 3' des fragments amplifiés, le plus souvent un A. Cette activité d'addition d'un nucléotide semble indépendante de l'information contenu dans le brin copié. Une telle activité catalytique de type terminale transférase a été trouvée pour d'autres polymérases dont l'AMV (Clark, 1988). Les T-vecteurs sont pourvus d'un dT non apparié, ce qui permet une ligation facile des fragments de PCR et du vecteur. En pratique, ils peuvent être préparés de la façon suivante: le vecteur de clonage est hydrolysé par une enzyme de restriction à coupure franche (on emploie généralement les enzymes *Sma*I ou *Eco*RV) et on l'incube à 72°C avec une Taq polymérase et du dTTP. Nous avons choisi comme vecteur le plasmide pMOS*Blue* (Amersham) qui a une taille de 2887 pb.

IV.2.2 Ligation

Un mélange contenant 6,8 µl de produit de PCR purifié, 0,5 µl de DTT 100mM, 0,5 µl d'ATP 10 mM, 1 µl de tampon de ligation 10X, 0,7 µl (35 ng) de vecteur T/A (pMOS*Blue* T-vector, Amersham) et 0,5 µl de T4 ADN ligase (2-3 U) est incubé pendant 2 heures à 15°C.

IV.2.3 Transformation

1,5 µl du produit de ligation sont transférés dans un tube de 0,5 ml. 18 µl de bactéries compétentes sont ajoutées. Le contact se fait pendant 20 minutes dans la glace. On effectue ensuite un choc thermique pendant 45 secondes à 42°C précisément puis on ajoute 80 µl de SOC (milieu de culture riche). Les bactéries sont placées dans l'incubateur à 37°C et agitées à 280 rpm. Après une heure d'agitation, les bactéries sont étalées sur boîte de LB-agar ampicilline (100 µg/ml) et incubées à 37°C pendant une nuit. Seules les bactéries possédant le plasmide seront résistantes à l'antibiotique et pourront pousser sur la boîte.

IV.2.4 Détection des clones recombinants

Pour sélectionner les plasmides recombinants, nous avons utilisé la méthode X-Gal / IPTG. 30 minutes avant d'étaler les bactéries, on imprègne les boîtes avec 100 µl d'IPTG 0,1M et 20 µl d'une solution de X-Gal à 50 mg/ml. Les souches d'*E. coli* utilisées au laboratoire (JM109, XL1-Blue® ou celles fournies avec le vecteur pMOS*Blue*) possèdent un gène codant une β-galactosidase défective, dépourvue des aa 11 à 41 (gène LacZΔM15), utile pour le test de l'α-complémentation. Les vecteurs plasmidiques utilisés au laboratoire possèdent quant à eux, en plus d'un gène de résistance à l'ampicilline, un site de clonage multiple (encore appelé « polylinker ») introduit au début du gène LacZ. Pour mettre en évidence la présence (vecteur religué sans insert) ou non (clonage d'un insert dans le gène LacZ) d'une activité β-galactosidase, 2 réactifs sont indispensables : IPTG et X-Gal. est un inducteur par dérépression de la synthèse de β-galactosidase et le X-Gal est un β-galactoside qui peut être hydrolysé par la β-galactosidase et libérant ainsi, en plus du galactose, une substance colorée en bleu. Finalement, les colonies blanches sont témoins de la présence de recombinants et les bleues d'absence de recombinants.

V Sous clonage en vecteur plasmidique pKS

Le vecteur utilisé pour le sous-clonage des fragments cosmidiques ou plasmidiques est le phagémide *pBlueScript* II KS (Stratagene). Il a une taille de 2961 pb.

V.1 Hydrolyse du vecteur

Le vecteur est tout d'abord clivé par une ou deux enzymes de restriction appartenant au site de clonage multiple. Le vecteur est ensuite purifié par extraction au phénol/chloroforme ou purifié comme un produit de PCR (voir §IV.1.3). Il est ensuite précipité, repris par du TE 1X et dosé en gel d'agarose par comparaison au témoin 13i.

V.2 Déphosphorylation

Lorsque le vecteur est hydrolysé par une enzyme unique, on le déphosphoryle afin d'éviter qu'il ne se religue sur lui même. On utilise pour cela une phosphatase (Boehringer). Le vecteur ouvert est incubé une heure au moins à 37°C avec un tampon 10X (Tris-HCl 0,5M; EDTA 1mM pH8,5) et 1 unité de phosphatase alcaline (Boehringer). Le vecteur déphosphorylé est ensuite purifié et repris en TE 1X. Il est dosé après électrophorèse en gel d'agarose.

V.3 Ligation rapide

8 µl d'ADN purifié sont mélangés dans un microtube à 2 µl d'un tampon 5X (kit Rapid DNA Ligation, Boehringer). On ajoute 10 µl de tampon de ligation 2X, 35 ng de vecteur puis 1 µl (5 U/µl) de T4 ADN ligase. Après 5 minutes de ligation à température ambiante, on procède à la transformation en utilisant 1,5 µl de produit de ligation.

V.4 Transformation

La transformation effectuée par la même méthode que celle décrite ci-dessus pour les produits de PCR en utilisant des bactéries lac$^-$ JM109 ou XL1-Blue® ou des HB101 (Promega) vendues compétentes. Les bactéries sont ensuite étalées sur boîtes comme précédemment sur un milieu LB-agar ampicilline. Les boîtes sont placées à 37°C pendant une nuit.

V.5 Recherche des clones d'intérêt

Avec les JM109 ou les XL1-Blue®, il est possible d'utiliser la méthode de détection XGAL/IPTG. Cependant, si on possède la sonde correspondant à un fragment de l'insert à sous-cloner, on procède comme suit.

V.5.1 Empreintes plasmidiques

Une rondelle de nitrocellulose est appliquée sur chaque boîte de Pétri pendant une minute. L'ADN est dénaturé dans la soude 0,5N pendant 5 minutes. La rondelle est ensuite neutralisée pendant 5 minutes dans un bain de Tris-HCl pH7,0. La rondelle est séchée sur papier et le matériel est fixé par la chaleur (une heure à 80°C sous vide).

V.5.2 Hybridation à une sonde nucléique

La méthode est identique à celle décrite ultérieurement pour les Southern mis à part que les rondelles sont préhybridées et hybridées à 42°C dans un tampon d'hybridation 3X pendant au moins 2 heures. L'hybridation est effectuée avec la sonde à raison de 250000 cpm/min par rondelle. Enfin, les rondelles sont lavées par 4 bains de 5 minutes à 55°C dans un tampon de SSC 0,1X-SDS 0,1% et sont mises en autoradiographie pendant 6 heures ou une nuit.

VI Préparation de l'ADN plasmidique recombinant

(kit Wizard™ Minipreps DNA Purification System-Promega)

VI.1 Lyse bactérienne par un détergent

1,5 ml de culture sont centrifugés 5 minutes à 3000 tours/min. Le surnageant est éliminé à la trompe à vide. Les bactéries sont resuspendues dans 200 µl de solution de resuspension (Tris-HCl pH7,5 50mM; EDTA 10mM; RNase A 100µg/ml) puis lysées (200 µl de NaOH 0,2M, SDS 1%). Après 2 minutes d'incubation le mélange est neutralisé (200 µl d'acétate de potassium pH4,8 1,32M) puis centrifugé 10 minutes à 12000 x g ; le surnageant contenant les plasmides est récupéré.

VI.2 Purification de l'ADN plasmidique

1 ml de résine Wizard Preps est ajouté au surnageant. Le tout est filtré sur microcolonne. Après lavage par 2 ml d'une solution de «Wash» (NaCl, 200mM; Tris-HCl 20mM, pH7,5; EDTA 5mM; EtOH 95%, (1,4 v/v), la colonne est essorée par centrifugation pendant 2 minutes.

VI.3 Elution de l'ADN de la matrice

50 µl de TE chaud (65°C) sont appliqués sur la colonne pendant quelques minutes. L'ADN plasmidique est élué de la colonne par 30 secondes de centrifugation.

VI.4 Détermination de la taille des inserts

3 µl de chaque solution de plasmides sont hydrolysés pendant 1 heure à 37°C dans un tube contenant 5 µl de TE 1X, 1 µl de tampon 10X adapté à l'enzyme (Boehringer) et 0,5 µl de chaque enzyme qui coupe de part et d'autre du site de

clonage (*Hin*dIII et *Bam*HI pour le vecteur pMOS*Blue* en tampon B). L'insert et le vecteur sont ensuite séparés par électrophorèse sur gel d'agarose.

Une hydrolyse sur une plus grande quantité (35 µl) permet, après purification comme précédemment (§ IV.1), d'avoir assez de matériel pour utiliser ces inserts comme sonde sur des Southern blots ou/et des Northern blots.

VII Sonde nucléique

VII.1 Sonde d'ADN : technique du multi-amorçage

L'ADN à marquer (25 ng dans 10 µl) est tout d'abord dénaturé 5 minutes dans l'eau bouillante puis gardé dans la glace. La réaction d'élongation, amorcée par un cocktail d'hexanucléotides, est réalisée pendant 30 minutes à 37°C par le fragment Klenow de l'ADN polymérase (kit Random Primed DNA Labeling, Boehringer) en présence de dNTPs dont du dCTP[α^{32}P]. La sonde double brin est ensuite purifiée par gel-filtration sur colonne de Sephadex G50 en tampon d'élution STE-SDS 0,1%, puis dénaturée par 1/10 (v/v) de NaOH 3N pendant 10 minutes à température ambiante et neutralisée par addition d'HCl 3N (5/6ème du volume de soude).

VII.2 Marquage en 5'OH d'un oligonucléotide

L'oligonucléotide (5 pmoles dans un volume maximal de 3 µl) est dénaturé 5 minutes à 95°C, mis 5 minutes dans la glace puis mélangé à 1 µl de tampon de phosphorylation 10X, 5 µl de dATP[γ^{32}P] et 1 µl de T4 polynucléotide kinase (5' end labelling kit, Amersham). Le mélange est incubé 30 minutes à 37°C. La réaction est stoppée par un passage de 5 minutes à 95°C puis 5 minutes dans la glace et addition de 1/10ème du volume de STE 10X. L'oligonucléotide est ensuite purifié par gel-filtration sur "Quick Spin™ Column Sephadex® G25" (Boehringer) puis dénaturé 5 minutes à 95°C et mis 3 minutes dans la glace. Un volume égal de formamide est alors ajouté.

VIII Southern Blots

Après hydrolyse par les enzymes de restriction, les fragments d'ADN sont séparés en gel d'agarose (0,8 à 2%) en tampon borate puis transférés sur membrane et fixés pour repérer le matériel génétique par des sondes radioactives.

VIII.1 Transfert

L'ADN est tout d'abord dénaturé par 2 bains de 15 minutes dans un tampon de dénaturation (NaCl 1,5M; NaOH 0,5M). Le gel est ensuite neutralisé par 2 bains dans un tampon de neutralisation (Tris-HCl 0,5M pH7,5; NaCl 3M). L'ADN est alors transféré sur membrane de Nylon Hybond™-N+ (Amersham) par capillarité en tampon SSC 20X pendant une nuit. L'ADN est ensuite fixé sur la membrane par une exposition de 3 minutes aux UV.

VIII.2 Hybridation

VIII.2.1 Sonde double brin

La membrane humidifiée en tampon SSC 3X est tout d'abord préhybridée au moins 2 heures à 65°C dans un tampon de préhybridation afin de saturer les sites libres de la membrane. Elle est ensuite hybridée dans un tampon d'hybridation pendant une nuit à même température avec la sonde (entre 100000 et 150000 cpm/min par couloir).

VIII.2.2 Sonde oligonucléotidique

Ces étapes sont identiques à celles utilisées pour une sonde double brin mais la température de préhybridation et d'hybridation ainsi que les compositions des tampons sont différentes :

- température de préhybridation et d'hybridation : 42°C
- tampon d'humidification : SSPE 6X
- tampons d'hybridation et préhybridation : cf. lexique des formules.

VIII.3 Lavages et autoradiographie

L'excès de sonde est éliminé de la membrane par rinçage en tampon SSC 3X pour une sonde d'ADN double brin, SSPE 6X pour une sonde oligonucléotidique; la membrane est ensuite lavée deux fois 15 minutes à 65°C en tampon SSC 0,1X-SDS 0,1% pour une sonde d'ADN double brin, à 42°C en tampon SSPE 4X-SDS 0,1% pour une sonde oligonucléotidique. Après un nouveau rinçage en tampon SSC 3X ou SSPE 6X, la membrane est séchée sur papier et mise en autoradiographie pendant des temps pouvant aller de quelques minutes à plusieurs jours. La membrane peut ensuite être déshybridée par des bains bouillants de SDS (0,1 à 0,5%) et le contrôle de la déshybridation est effectué par autoradiographie.

IX Utilisation des sondes sur Northern Blots

IX.1 Hybridation

La membrane est humidifiée en tampon SSPE 6X et préhybridée au moins 2 heures à 42°C en tampon de préhybridation/hybridation préparé selon le protocole préconisé par Clontech. Elle est ensuite hybridée après remplacement du tampon "Clontech" durant une nuit avec la sonde (1.000.000 à 1.500.000 cpm/min par couloir).

IX.2 Lavage et autoradiographie

La membrane est rincée par du tampon SSPE 6X et lavée deux fois pendant 15 minutes à 65°C par du tampon SSPE 0,1X-SDS 0,1%. Elle est alors mise en autoradiographie durant plusieurs jours.

X Séquençage nucléotidique

La technique utilisée est celle de Sanger (Sanger *et al.*, 1977). L'ADN plasmidique ou cosmidique est tout d'abord dénaturé par de la soude en présence d'EDTA puis précipité par l'acétate de sodium et de l'éthanol. Le vecteur monobrin est ensuite hybridé à une amorce oligonucléotidique (T3, T7, M-40 ou oligonucléotide spécifique de l'insert). Puis, 4 réactions de polymérisation en parallèles sont réalisées par l'ADN polymérase I (2 kits ont été utilisés: Sequenase 2.0, U.S. Biochemical corporation, Amersham et ^{T7}Sequencing™ Kit, Pharmacia Biotech) ou par une polymérase thermostable (Kit SequiTherm™ Cycle Sequencing, Epicentre Technologie). La synthèse du second brin est effectuée en présence des 4 désoxyribonucléotides dont l'un, le dCTP, est marqué en α par du ^{35}S, plus stable que le ^{32}P et donnant des images plus nettes. Chacune des 4 réactions est ensuite poursuivie en présence de l'une des 4 bases sous forme di-désoxyribonucléoside triphosphate à très faible concentration qui arrête, au hasard, la polymérisation.

Les fragments de tailles différentes sont ensuite séparés par électrophorèse sur gel de polyacrylamide à 6% en tampon urée (SequaGel-6™, National Diagnostics). La lecture directe de la séquence est ensuite réalisée après autoradiographie pendant une nuit à plusieurs jours. Cette méthode permet de lire des séquences de 200 à 400 pb.

XI Analyse des séquences

Les données sont traitées par informatique à l'aide des programmes de PC/GENE, et plus récemment grâce au serveur NCBI disponibles sur Internet: http://www.ncbi.nlm.nih.gov/.

Afin de calculer le pourcentage d'homologie entre les différents sous-domaines riches en résidus de cystéine et étant donné le grand nombre de séquences à

traiter, nous avons utilisé le programme NALIGN (PC/Gene) puis transféré le résultat dans le logiciel Excel. Nous avons ensuite transformé la matrice des alignements AxB en matrice BxA pour que chaque séquence soit sur une unique colonne. Nous avons ensuite élaboré le programme suivant pour calculer le pourcentage d'homologie entre les différents sous-domaines:

```
'=============================================================================
'          Fonction Calcul du pourcentage d'homologie entre des séquences (EXCEL)
'                    nucléotidiques - tableau vertical
' La première séquence (CYS1) s'étend de la cellule AU3 et AU353 soit de la ligne 3 de la
' colonne 47 à la ligne 353 de la même colonne.
' La dernière séquence à comparer s'étend de la cellule AA3 (ligne 3, colonne 23), à la ligne
' 353 de la colonne 23.
'=============================================================================

Sub phyla()
With Worksheets("Cys")              'Feuille de travail "Cys"
'c=colonne
'cc=les colonnes à comparer
'l=ligne
'lv=nucléotide contenu dans la cellule (l,c)
'lvv=nucléotide contenu dans la cellule à comparer à lv
'nb=nombre de nucléotides comparés
'd=le nombre de nucléotides différents
  For c = 47 To 26 Step -1
    For cc = c To c - 20 Step -1      'décrémentation de 1
      nb = 0                          'initialise nb à 0
      d = 0                           'initialise d à 0
      For l = 3 To 353
        lv = .Cells(l, c).Value
        llv = .Cells(l, cc).Value
        If lv <> "" And llv <> "" Then nb = nb + 1
        If lv = llv Then
          If lv <> "" And llv <> "" Then d = d + 1
          End If
      Next l
      If d > 0 Then .Cells(50 - cc, c - 22) = 100 * d / nb
      'calcul du pourcentage d'homologie et l'inscrit dans une cellule
    Next cc
  Next c
  MsgBox ("Fin de la fonction....")
End With   'fin d'utilisation de la feuille de  travail
End Sub    'fin du programme
'=============================================================================
```

XII Extension d'amorce

XII.1 Marquage de l'oligonucléotide

5 pmoles d'un oligonucléotide antisens spécifique du transcrit sont marqués au $\gamma^{32}P$ par la méthode décrite précédemment. L'arrêt de la réaction de marquage se fait par chauffage à 95°C pendant 5 minutes. L'oligonucléotide, dans un volume réactionnel de 10 µl, est ensuite refroidi dans la glace et le tube est centrifugé. On ajoute alors 10 µl d'eau stérile.

XII.2 Purification et dénaturation

L'oligonucléotide est purifié par tamisage moléculaire sur une colonne G25 (Quick Spin™ columns Sephadex® G-25, Boehringer). L'oligonucléotide est ensuite dénaturé par chauffage pendant 5 minutes à 95°C.

XII.3 Réaction d'extension d'amorce

XII.3.1 Hybridation

Dans un tube de 1,5 ml sont mélangés sur la glace : de l'ARN total (10 µg) ou poly(A)$^+$ (3µg), de l'oligonucléotide marqué (10^6 cpm), 4 µl de tampon de transcriptase inverse 5X et de l'eau traitée au DEPC (qsp 20 µl).

Le mélange est chauffé à 80°C pendant 3 minutes puis on diminue la température jusqu'au Tm de l'oligonucléotide pendant 30 minutes.

XII.3.2 Elongation

A 10 µl de mélange d'ARN/amorce sont ajoutés 2,5 µl de DTT 0,1M, 4 µl de dNTPs 0,25mM chaque, 3 µl de tampon de transcriptase inverse 5X, 1 µl (40 U) d'inhibiteur de RNases (rRNasin®, Promega), 1 µl de transcriptase inverse MMLV

(200 U/µl) et 3,5 µl d'eau-DEPC. Le mélange est incubé pendant 60 minutes à 45°C puis on ajoute 190 µl d'eau-DEPC.

XII.3.3 Purification

L'ADNc est purifié par extraction phénol/chloroforme puis précipité par de l'acétate de sodium, de l'éthanol absolu et de l'ARNt. Le culot est repris par 10 µl de la solution stop du kit de séquençage (Kit Sequenase).

XII.3.4 Electrophorèse

Après dénaturation par chauffage à 80°C pendant 5 minutes, les ADNc et une séquence témoin (séquence d'un plasmide sans insert) sont séparés par électrophorèse sur gel de polyacrylamide à 6% en tampon urée (SequaGel-6™, National Diagnostics).

Composition des Tampons

COMPOSITION DES TAMPONS

TE 10X

Tris-HCl 1M pH 8,0	10 ml
EDTA 0,5M	2 ml
H_2O	qsp 100 ml
pH 7,0	

STE 10X

NaCl 3M	166,6 ml
Tris-HCl 1M pH 8,0	50 ml
EDTA 0,5M	10 ml
H_2O	qsp 500 ml

SSC 20X

NaCl	175,3 g
Citrate trisodique	88,2 g
H_2O	qsp 1 l
pH 7,0	

SSPE 20X

NaCl	174 g
NaH_2PO4, H_2O	27,6 g
EDTA	7,4 g
H_2O	qsp 1 l
pH 7,4	

Tampon TEA 10X

Tris	148,4 g
Acide acétique	11,42 ml
EDTA 0,5M	20 ml
H_2O	qsp 1 l
pH 8,0	

Denhardt's 50X

Ficoll 400	1 g
Polyvinylpyrrolidone	1 g
SAB	1 g
H_2O	qsp 100 ml

Eau DEPC

DEPC	1 ml
H_2O désionisée	qsp 1 l

Solution d'IPTG

IPTG	0,12 g
Eau stérile	qsp 5 ml

Solution de X-Gal

X-Gal	50 mg
N,N' diméthylformamide	qsp 1 ml

LB

Bactotryptone	10 g
Yeast extract	5 g
NaCl	5 g
H_2O	qsp 1 l
pH 7,2	

LB-Agar

LB contenant 15 g d'Agar pour 1l

Tampon Borate 10X

Acide borique	55 g
Tris	108 g
EDTA	9,3 g
H_2O	qsp 1 l
pH 8,3	

Tampon Borate 10X (séquence)

Acide borique	55 g
Tris	108 g
EDTA	6,72 g
H_2O	qsp 1 l
pH 8,3	

Tampon GT

Isothiocyanate de guanidium	23,6 g
Citrate trisodique	73,5 mg
Sarcosyl	250 mg
H_2O stérile	qsp 50 ml
DEPC	50 µl
β mercaptoéthanol (extemporanément)	357 µl

Tampon de purification des produits de PCR

KCl	372,8 mg
Tris-HCl, pH8,8	121,14 mg
$MgCl_2$, 6 H_2O	30,50 mg
Triton X-100	100 mg
Eau stérile	qsp 100 ml

SOC

Bactotryptone (w/v)	2%
Yeast extract (w/v)	0,5%
NaCl	10mM
KCl	25mM
$MgCl_2$	10mM
$MgSO_4$	10mM
Glucose	20mM

Tampon hybridation 3X

formamide désionisée	50 ml
EDTA 0,5M	0,3 ml
Denhardt's 50X	2 ml
Hepes 500mM	10 ml
SSC 20X	15 ml
ssDNA	2,5 ml (25 mg)
H_2O	qsp 100 ml

Tampon hybridation oligonucléotides

SSPE 20X	.....30 ml
Denhardt's 50X	20 ml
ssDNA	1 ml (10 mg)
sulfate de dextran	10 g
SDS 10%	0,5 ml
H_2O	qsp 100 ml

hybridation ARN Clontech

SSPE 20X	.....25 ml
Denhardt's 50X	20 ml
ssDNA	1 ml (10 mg)
formamide	50 ml
SDS	2 g
H_2O	qsp 100 ml

Tampon de préhybridation d'une sonde nucléique (Southern blot)

SSC 20X	30 ml
Denhardt's 50X	10 ml
SDS 10%	5 ml
H_2O	qsp 100 ml

Tampon d'hybridation d'une sonde nucléique (Southern blot)

SSC 20X	30 ml	
Denhardt's 50X	10 ml	
SDS 10%	5 ml	
ssDNA	2,5 ml	(25 mg)
sulfate de dextran 50%	10 g	
H_2O	qsp 100 ml	

Bibliographie

BIBLIOGRAPHIE

Abe, M., Kufe, D. (1993) Characterization of *cis*-acting element regulating transcription of the human DF3 breast carcinoma-associated antigen (*MUC1*) gene.
Proc. Natl. Acad. Sci. USA **90**, 282-286

Aebi, M., Hornig, H., Weissmann, C. (1987) 5' cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5' splice region, not by the conserved 5' GU.
Cell **50**, 237-246

Albone, E.F., Hagen, F.K., VanWuyckhuyse, B.C., Tabak, L.A. (1994) Molecular cloning of rat submandibular gland apomucin.
J. Biol. Chem. **269**, 16845-16852

Albone, E.F., Hagen, F.K., Szpirer, C., Tabak, L.A. (1996) Molecular cloning and characterization of the gene encoding rat submandibular gland apomucin, *Mucsmg*.
Glycoconj. J. **13**, 709-716

Al-Hashimi, I., Levine, M.J. (1989) Characterization of *in vivo* salivary-derived enamel pellicle.
Archs. Oral Biol. **34**, 289-295

Allen, A. (1983) Mucus - a protective secretion of complexity.
TIBS **8**, 169-173

Allen, A., Cunliffe, W.J., Pearson, J.P., Sellers, L.A., Ward, R. (1984) Studies on gastrointestinal mucus.
Scand. J. Gastroenterol. **93**, 101-113

Allen, A., Carroll, N.J.H. (1985) Adherent and soluble mucus in the stomach and duodenum.
Dig. Dis. Sci. **30**, 55-62

Amerongen, A.V.N., Bolscher, J.G.M., Veerman, E.C.I. (1995) Salivary mucins : protective functions in relation to their diversity.
Glycobiology **5**, 733-740

Asker, N., Baeckström, D., Axelsson, M.A.B., Carlstedt, I., Hansson, G.C. (1995) The human mucin apoprotein appears to dimerize before O-glycosylation and shares epitopes with the 'insoluble' mucin of rat small intestine.
Biochem. J. **308**, 873-880

Aubert, J.P., Porchet, N., Crépin, M., Duterque-Coquillaud, M., Vergnes, G., Mazzuca, M., Debuire, B., Petitprez, D., Degand, P. (1991) Evidence for different human tracheobronchial mucin peptides deduced from nucleotide cDNA sequences.
Am. J. Respir. Cell. Mol. Biol. **5**, 178-185

Audié, J.P., Janin, A., Porchet, N., Copin, M.C., Gosselin, B., Aubert, J.P. (1993) Expression of human mucin genes in respiratory, digestive, and reproductive tracts ascertained by in situ hybridization.
J. Histochem. Cytochem. **41**, 1479-1485

Audié, J.P., Tetaert, D., Pigny, P., Buisine, M.P., Janin, A., Aubert, J.P., Porchet, N., Boersma, A. (1995) Mucin gene expression in the human endocervix.
Hum. Reprod. **10**, 98-102

Baker, M.E. (1988) Invertebrate vitellogenin is homologous to human von Willebrand factor.
Biochem. J. (letter) **256**, 1059-1063

Balagué, C., Gambus, G., Carrato, C.,Porchet, N., Aubert, J.P., Kim, Y.S., Real, F.X. (1994) Altered expression of MUC2, MUC4, and MUC5 mucin genes in pancreas tissues and cancer cell lines.
Gastroenterology **106**, 1056-1061

Balagué, C., Audié, J.P., Porchet, N., Real, F.X. (1995) In situ hybridization shows distinct patterns of mucin gene expression in normal, benign, and malignant pancreas tissues.
Gastroenterology **109**, 953-964

Benian, G.M., Kiff, J.E., Neckelmann, N., Moerman, D.G., Waterston, R.H. (1989) Sequence of an unusually large protein implicated in regulation of myosin activity in *C. Elegans*.
Nature (London) **342**, 45-50

Bharathan, S., Moriarty, J., Moody, C.E., Sherblom, A.P. (1990) Effect of tunicamycin on sialomucin and natural killer susceptibility of rat mammary tumor ascites cells.
Cancer Res. **50**, 5250-5256

Bhargava, A.K., Woitach, J.T., Davidson, E.A., Bhavanandan, V.P. (1990) Cloning and cDNA sequence of a bovine submaxillary gland mucin-like protein containing two distinct domains.
Proc. Natl. Acad. Sci. USA **87**, 6798-6802

Biesbrock, A.R., Bobek, L.A., Levine, M.J. (1997) *MUC7* gene expression and genetic polymorphism.
Glycoconj. J. **14**, 415-422

Bobek, L.A., Tsai, H., Biesbrock, A.R., Levine, M.J. (1993) Molecular cloning, sequence, and specificity of expression of the gene encoding the low molecular weight human salivary mucin (*MUC7*).
J. Biol. Chem. **268**, 20563-20569

Bobek, L.A., Liu, J., Sait, S.N.J., Shows, T.B., Bobek, Y.A., Levine, M.J. (1996) Structure and chromosomal localization of the human salivary mucin gene, MUC7.
Genomics **31**, 277-282

Boureau, H., Gomez Trevino, M. (1995) Mucus et écosystème digestif.
Mucus Dialogue **6**, 1-7 *IPSEN*

Buisine, M.P., Janin, A., Maunoury, V., Audié, J.P., Delescaut, M.P., Copin, M.C., Colombel, J.F., Degand, P., Aubert, J.P., Porchet, N. (1996) Aberrant expression of a human mucin gene (*MUC5AC*) in rectosigmoid villous adenoma.
Gastroenterology **110**, 84-91

Burt, D.W., Paton, I.R. (1992) Evolutionary origins of the growth factor-β gene family.
DNA & Cell Biol. **11**, 497-510

Byrne, B.M., Gruber, M., Ab, G. (1989) The evolution of egg yolk proteins.
Prog. Biophys. Molec. Biol. **53**, 33-69

Carlstedt, I., Sheehan, J.K. (1984) Macromolecular properties and polymeric structure of mucus glycoproteins.
Ciba Foun Symp. **109**, 157-172

Chen, Z.Y., Denney, R.M., Breakefield, X.O. (1995) Norrie disease and MAO genes : nearest neighbors.
Hum. Mol. Genet. **4**, 1729-1737

Chinery, R., Williamson, J., Poulsom, R. (1996) The gene encoding human intestinal trefoil factor (TFF3) is located on chromosome 21q22.3 clustered with other members of the trefoil peptide family.
Genomics **32**, 281-284

Clark, J.M. (1988) Novel non-templated nucleotide addition reactions catalyzed by prokaryotic and eucaryotic DNA polymerases.
Nucleic Acids Res. **16**, 9677-9686

Colombatti, A., Bonaldo, P. (1991) The superfamily of proteins with von Willebrand factor type A-like domains: one theme common to components of extracellular matrix, hemostasis, cellular adhesion, and defense mechanisms.
Blood **77**, 2305-2315

Craig, J.M., Bickmore, W.A. (1994) The distribution of CpG islands in mammalian chromosomes.
Nat. Genet. **7**, 376-382

Crépin, M., Porchet, N., Aubert, J.P., Degand, P. (1991) Diversity of the peptide moiety of human airway mucins.
Biorheology **27**, 471-484

Daopin, S., Cohen, G.H., Davies, D. (1992) Structural similarity between transforming growth factor-β2 and nerve growth factor.
Science **258**, 1160-1162

Debailleul, V. Expression des gènes de mucines humaines. Etude de la stabilité, de l'hétérogénéité et du polymorphisme des ARNm.
Thèse d'Université, Lille, 1997

Dekker, J., Van Beurden-Lamers, W.M.O., Oprins, A., Strous, G.J. (1989) Isolation and structural analysis of rat gastric mucus glycoprotein suggests a homogeneous protein backbone.
Biochem. J. **260**, 717-723

Dekker, J., Strous, G.J. (1990) Covalent oligomerization of rat gastric mucin occurs in the rough endoplasmic reticulum, is *N*-glycosylation-dependent, and precedes initial *O*-glycosylation.
J. Biol. Chem. **265**, 18116-18122

Dekker, J., Van Der Ende, A., Aelmans, P.H., Strous, G.J. (1991) Rat gastric mucin is synthesized and secreted exclusively as filamentous oligomers.
Biochem. J. **279**, 251-256

Denny, P.C., Mirels, L., Denny, P.A. (1996) Mouse submandibular gland salivary apomucin contains repeated *N*-glycosylation sites.
Glycobiology **6**, 43-50

Desseyn J.L., Aubert J.P., Van Seuningen, I., Porchet N., Laine A. (1997a) Genomic Organization of the 3' Region of the Human Mucin Gene *MUC5B*.
J. Biol. Chem. **272**, 16873-16883

Desseyn J.L., Buisine, M.P., Porchet N., Aubert, J.P., Degand, P., Laine A. (1997b) Evolutionary History of the 11p15 Human Mucin Genes.
J. Mol. Evol. **45** (sous presse)

Desseyn, J.L., Guyonnet Dupérat, V., Porchet, N., Aubert, J.P., Laine, A. (1997c) Human mucin gene MUC5B: the 10.7 kb large central exon encodes various subdomains resulting in a super-repeat. Structural evidence for a 11p15.5 gene family.
J. Biol. Chem. **272**, 3168-3178

Devine, P.L., McKenzie, F.C. (1992) Mucins : Structure, function, and associations with malignancy.
BioEssays 14, 619-625

Dignass, A., Lynch-Devaney, K., Kindon, H., ,Thim, L., Podolsky, D.K. (1994) Trefoil peptides promote epithelial migration through a transforming growth factor β-independent pathway.
J. Clin. Invest. **94**, 376-383

Dufossé, J., Porchet, N., Audié, J.P., Guyonnet Dupérat, V., Laine, A., Van-Seuningen, I., Marrakchi, S., Degand, P., Aubert, J.P. (1993) Degenerate 87-base-pair tandem repeats create hydrophilic/hydrophobic alternating domains in human mucin peptides mapped to 11p15.
Biochem. J. **293**, 329-337

Eckhardt, A.E., Timpte, C.S., Abernethy, J.L., Zhao, Y., Hill, R.L. (1991) Porcine submaxillary mucin contains a cystin-rich, carboxyl-terminal domain in addition to a highly repetitive, glycosylated domain.
J. Biol. Chem. **266**, 9678-9686

Eversole, L.R. (1972) The mucoprotein histochemistry of human mucous acinar cell containing salivary glands : submandibular and sublingual glands.
Archs. oral Biol. **17**, 43-53

Fahim, R.E.F., Forstner, G.G., Forstner, J.F. (1987) Structural and compositional differences between intracellular and secreted mucin of rat small intestine.
Biochem. J. **248**, 389-396

Fogg, F.J.J., Hutton, D.A., Jumel, K., Pearson, J.P., Harding, S.E., Allen, A. (1996) Characterization of pig colonic mucins.
Biochem. J. **316**, 937-942

Forstner, G. (1995) Signal transduction, packaging and secretion of mucins.
Annu. Rev. Physiol. **57**, 585-605

Foster, P.A., Fulcher, C.A., Marti, T., Titani, K., Zimmerman, T.S. (1987) A major factor VIII binding domain resides within the amino-terminal 272 amino acid residues of von Willebrand factor.
J. Biol. Chem. **262**, 8443-8446

Fox, M.F., Lahbib, F., Pratt, W., Attwood, J., Gum, J., Kim, Y., Swallow, D.M. (1992) Regional localization of the intestinal mucin gene *MUC3* to chromosome 7q22.
Ann. Hum. Genet. **56**, 281-287

Fretto, L.J., Fowler, W.E., McCaslin, D.R., Erickson, H.P., McKee, P.A. (1986) Substructure of human von Willebrand factor. Proteolysis by V8 and characterization of two functional domains.
J. Biol. Chem. **261**, 15679-15689

Frohman, M.A., Dush, M.K., Martin, G.R. (1988) Rapid production of full-length cDNAs from rare transcripts : amplification using a single gene-specific oligonucleotide primer.
Proc. Natl. Acad. Sci. USA. **85**, 8998-9002

Gaillard, D., Plotkowski, C, Puchelle, E. Mucus et protection de la muqueuse respiratoire. Edition et communication médicales, 1992.

Galli, J., Wieslander, L. (1993) A repetitive secretory protein gene of a novel type in *Chironomus tentans* is specifically expressed in the salivary glands and exhibits extensive length polymorphism.
J. Biol. Chem. **268**, 11888-11893

Gendler, S., Taylor-Papadimitriou, J., Duhig, T., Rothbard, J., Burchell, J. (1988) A highly immunogenic region of a human polymorphic epithelial mucin expressed by carcinomas is made up of tandem repeats.
J. Biol. Chem. **263**, 12820-12823

Gendler, S.J., Spicer A.P. (1995) Epithelial mucin genes.
Ann. Rev. Physiol. **57**, 607-634

Gerard, C., Eddy, R.L., Shows, T.B. (1990) The core polypeptide of cystic fibrosis tracheal mucin contains a tandem repeat structure. Evidence for a common mucin in airway and gastrointestinal tissue.
J. Clin. Invest. **86**, 1921-1927

Gibbons, R.A. (1978) Mucus of the mammalian genital tract.
Br. Med. Bull. **34**, 34-38

Griffiths, B., Matthews, D.J., West, L., Attwood, J., Povey, S.M., Swallow, D.M., Gum, J.R., Kim, Y.S. (1990) Assignment of the polymorphic intestinal mucin gene (MUC2) to chromosome 11p15.
Ann. Hum. Genet. **54**, 277-285

Gross, M.S., Guyonnet Dupérat, V., Porchet, N., Bernheim, A., Aubert, J.P., Nguyen, V.C. (1992) Mucin 4 (MUC4) gene : regional assignment (3q29) and RFLP analysis.
Ann. Genet. **35**, 21-26

Gum, J.R., Byrd, J.C., Hicks, J.W., Toribara, N.W., Lamport, D.T.A., Kim, Y.S. (1989) Molecular cloning of human intestinal mucin cDNAs. Sequence analysis and evidence for genetic polymorphism.
J. Biol. Chem. **264**, 6480-6487

Gum, J.R., Hicks, J.W., Swallow, D.M., Lagace, R.E., Byrd, J.C., Lamport, D.T.A., Siddiki, B., Kim, Y.S. (1990) Molecular cloning of cDNAs derived from a novel human intestinal mucin gene.
Biochem. Biophys. Res. Commun. **171**, 407-415

Gum, J.R., Hicks, J.W., Lagace, R.E., Byrd, J.C, Toribara, N.W., Siddiki, B., Fearney, F.J., Lamport, D.T.A., Kim, Y.S. (1991) Molecular cloning of rat intestinal mucin. Lack of conservation between mammalian species.
J. Biol. Chem. **266**, 22733-22738

Gum, J.R. (1992) Mucin genes and proteins they encode: structure, diversity, and regulation.
Am. Respir. Cell. Mol. Biol. **7**, 554-564

Gum, J.R., Hicks, J.W., Toribara, N.W., Rothe, E.M., Lagace, R.E., Kim, Y.S. (1992) The human MUC2 intestinal mucin has cysteine-rich subdomains located both upstream and downstream of its central repetitive region.
J. Biol. Chem. **267**, 21375-21383

Gum, J.R., Hicks, J.W., Toribara, N.W., Siddiki, B., Kim, Y.S. (1994) Molecular cloning of human intestinal mucin (*MUC2*) cDNA. Identification of the amino terminus and overall sequence similarity to prepro-von Willebrand factor.
J. Biol. Chem. **269**, 2440-2446

Gum, J.R., Hicks, J.W., Kim, Y.S. (1997) Identification and characterization of the *MUC2* (human intestinal mucin) gene 5'-flanking region : promoter activity in cultured cells.
Biochem. J. **325**, 259-267

Guyonnet Dupérat, V., Audié, J.P., Debailleul, V., Laine, A., Buisine, M.P., Galiegue-Zouitina, S., Pigny, P., Degand, P., Aubert, J.P., Porchet, N. (1995) Characterization of the human mucin gene MUC5AC: a consensus cysteine-rich domain for 11p15 mucin genes ?
Biochem. J. **305**, 211-219

Hansson, G.C., Baeckström, D., Carlstedt, I., Klinga-Levan, K. (1994) Molecular cloning of a cDNA coding for a region of an apoprotein from the 'insoluble' mucin of rat small intestine.
Biochem. Biophys. Res. Commun. **198**, 181-190

Hardy, D.M,. Garbers, D.L. (1995) A sperm membrane protein that binds in a species-specific manner to the egg extracellular matrix is homologous to von Willebrand factor.
J. Biol. Chem. **270**, 26025-26028

Hauser, F., Hoffmann, W. (1992) P-domains as shuffled cysteine-rich modules in integumentary mucin C.1 (FIM-C.1) from *Xenopus laevis*. Polydispersity and genetic polymorphism.
J. Biol. Chem. **267**, 24620-24624

Hildebrand, A., Romaris, M., Rasmussen, L.M., Heinegard, D., Twardzik, D.R., Border, W.A., Ruoslahti, E. (1994) Interaction of the small interstitial proteoglycans biglycan, decorin and fibromodulin with transforming growth factor beta.
Biochem. J. **302**, 527-534

Hilkens, J., Wesseling, J., Vos, H.L., Storm, J., Boer, B., van der Valk, S.W., Maas, M.C. (1995) Involvement of the cell surface-bound mucin, episialin/MUC1, in progression of human carcinomas.
Biochem. Soc. Trans. **23**, 822-826

Hill, A.S., Pratt, W.S., Fox, M., Vinall, L., Green, E., Ho, J., Kim, Y.S., Swallow, D.M., Gum, J.R. The human *MUC3* gene and its complexities.
4th International Workshop on Carcinoma-Associated Mucins. (Aug 1996) Cambridge

Hill, H.D., Reynolds, J.A., Hill, R.L. (1977) Purification, composition, molecular weight, and subunit structure of ovine submaxillary mucin.
J. Biol. Chem. **252**, 3791-3798

Ho, J.J.L, Kim, Y.S. (1991) Carbohydrate antigens on cancer-associated mucin-like molecules.
Cancer Biol. **2**, 389-400

Ho, S.B., Niehans, G.A., Lyftogt, C., Yan, P.S., Cherwitz, D.L., Gum, E.T., Dahiya, R., Kim, Y.S. (1993) Heterogeneity of mucin gene expression in normal and neoplastic tissues.
Cancer Res. **53**, 641-651

Ho, J.J.L, Kim, Y.S. (1995a) Do mucins promote tumor cell metastasis ? (Review)
Int. J. Oncol. **7**, 913-926

Ho, S.B., Roberton, A.M., Shekels, L.L., Lyftogt, C.T., Niehans, G.A., Toribara, N.W. (1995b) Expression cloning of gastric mucin complementary DNA and localization of mucin gene expression.
Gastroenterology **109**, 735-747

Hoffmann, W. (1988) A new repetitive protein from *Xenopus laevis* skin highly homologous to pancreatic spasmolytic polypeptide.
J. Biol. Chem. **263**, 7686-7690

Hoffmann, W., Hauser, F. (1993a) The P-domains or trefoil motif: a role in renewal and pathology of mucous epithelia ?
Trends Biochem. Sci. **18**, 239-243

Hoffmann, W., Hauser, F. (1993b) Biosynthesis of frog skin mucins : cysteine-rich shuffled modules, polydispersities and genetic polymorphism.
Comp. Biochem. Physiol. **105B**, 465-472

Hoffmann, W., Joba, W. (1995) Biosynthesis and molecular architecture of gel-forming mucins : implications from an amphibian model system.
Biochem. Soc. Trans. **23**, 200-205

Hollingsworth, M.A., Closken, C., Harris, A., McDonnald, C.D., Pahwa, G.S., Maher, L.J. (1994) A nuclear factor that binds purine-rich, single-stranded oligonucleotides derived from S1-sensitive elements upstream of the CFTR gene and the MUC1 gene.
Nucleic Acids Res. **22**, 1138-1146

Hovenberg, H.W., Davies, J.R., Carlstedt, I. (1996) Different mucins are produced by the surface epithelium and the submucosa in human trachea: identification of MUC5AC as a major mucin from the goblet cells.
Biochem. J. **318**, 319-324

Hull, S.R., Sheng, Z., Vanderpuye, O., David, C., Carraway, K.L. (1990) Isolation and partial characterization of ascites sialoglycoprotein-2 of the cell surface sialomucin complex of 13762 rat mammary adenoma cells.
Biochem. J. **265**, 121-129

Hull, S.R., Sugarman, E.D., Spielman, J., Carraway, K.L. (1991) Biosynthetic maturation of an ascites tumor cell surface sialomucin. Evidence for *O*-glycosylation of cell surface glycoprotein by the addition of new olisaccharides during recycling.
J. Biol. Chem. **266**, 13580-13586

Hunt, L.T., Barker, C.W. (1987) Von Willebrand factor shares a distinctive cysteine-rich domain with thrombospondin and procollagen.
Biochem. Biophys. Res. Commun. **144**, 876-882

Iontcheva, I., Oppenheim, F.G., Troxler, R.F. (1997) Human salivary mucin MG1 selectively forms heterotypic complexes with amylase, proline-rich proteins, statherin, and histatins.
J. Dent. Res. **76**, 734-743

Jackson, I.J. (1991) A reappraisal of non-consensus mRNA splice sites.
Nucleic Acids Res. **19**, 3795-3798

Jany, B.H., Gallup, M.W., Yan, P.S., Gum, J.R., Kim, Y.S., Basbaum, C.B. (1991) Human bronchus and intestine express the same mucin gene.
J. Clin. Invest. **87**, 77-82

Joba, W., Hoffmann, W. (1997) Similarities of integumentary mucin B.1 from *Xenopus laevis* and prepro-von Willebrand factor at their amino-terminal regions.
J. Biol. Chem. **272**, 1805-1810

Jordan, B.R. (1991) Ilots HTF: le gène annoncé.
médecine/sciences **2**, 153-160

Juang, S.H., Carvajal, M.E., Whitney, M., Liu, Y., Carraway, C.A. (1996) Tyrosine phosphorylation at the membrane-microfilament interface: a p185neu-associated signal transduction particle containing Src, Abl and phosphorylated p58, a membrane- and microfilament-associated retroviral gag-like protein.
Oncogene **12**, 1033-1042

Kasturi, L., Chen, H., Shakin-Eshleman, S.H. (1997) Regulation of N-linked core glycosylation: use of a site-directed mutagenesis approach to identify Asn-Xaa-Ser/Thr sequons that are poor oligosaccharide acceptors.
Biochem. J. **323**, 415-419

Kawagishi, S., Fahim, R.E.F., Wong, K.H., Bennick, A. (1990) Purification and characterization of subnits of high molecular weight human salivary mucin.
Archs. Oral Biol. **35**, 265-272

Kent, S.P. (1963) Study of tissue mucins using the fluorescent antibody technique. II. The preparation and specificity of human submaxillary gland mucin antibody.
J. Histochem. Cytochem. **11**, 273-282

Kim, Y.S., Hicks, J.W., Ho, J.J.L., Swallow, D, Gum, J.R. (1994) Diverse structures of human intestinal mucins *MUC2* and *MUC3*.
3rd International Workshop on Carcinoma-Associated Mucins. (Aug 1994) Cambridge

Kingsley, D.M. (1994) The TGF-β superfamily : new members, new receptors, and new genetic tests of function in different organisms.
Genes & Dev. **8**, 133-146

Klings-Levan, K., Gum, J.R., Gendler, S.J., Kim, Y.S., Hansson, G.C. (1996) Chromosomal mapping of three mucin genes in the rat.
Mammalian Genome **7**, 248-250

Klomp, L.W.J., Jan De Lely, A., Strous G.J. (1994a) Biosynthesis of a human gall-bladder mucin.
Biochem. J. **304**, 737-744

Klomp, L.W.J., Van Rens L., Strous G.J. (1994b) Identification of a human gastric precursor : N-linked glycosylation and oligomerization.
Biochem. J. **304**, 693-698

Klomp, L.W.J., Van Rens L., Strous G.J. (1995) Cloning and analysis of human gastric mucin cDNA reveals two types of conserved cysteine-rich domains.
Biochem. J. **308**, 831-838

Kondaiah, P., Sands, M.J., Smith, J.M., Fields, A., Roberts, A.B., Sporn, M.B., Melton, D.A. (1990) Identification of a novel transforming growth factor-β (TGF-β5) mRNA in *Xenopus laevis*.
J. Biol. Chem. **265**, 1089-1093

Kovarik, A., Peat, N., Wilson, D., Gendler, S.J., Taylor-Papadimitriou, J. (1993) Analysis of the tissue-specific promoter of the *MUC1* gene.
J. Biol. Chem. **268**, 9917-9926

Kozak, M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs.
Nucleic Acids Res. **15**, 8125-8148

Lamblin, G., Lhermitte, M., Laffite, J.J. (1977a) Etude comparative des mucines bronchiques purifiées à partir de l'expectoration de sujets atteints de mucoviscidose et d'autres affectations bronchiques chroniques.
Bull. Eur. Physiopathol. Respir. **13**, 175-190

Lamblin, G., Laffite, J.J., Lhermitte, M., Degand, P., Roussel, P. (1977b)
Mucins from C.F. sputum.
Mod. Probl. Paediatr. **19**, 53-64

Lamblin, G., Roussel, P. (1993) Airway mucins and their role in defence against micro-organisms.
Respiratory medicine **87**, 421-426

Larsen, F., Gundersen, G., Prydz, H. (1992) Choice of enzymes for mapping based on CpG islands in the human genome.
GATA **9**, 80-85

Lee, S.P., Carey, M.C., LaMont, J.T. (1981a) Aspirin prevention of cholesterol gallstone formation in prairie dogs.
Science **211**, 1429-1431

Lee, S.P., LaMont, J.T., Carey, M.C. (1981b) Role of gallbladder mucus hypersecretion in the evolution of cholesterol gallstones. Studies in the prairie dog.
J. Clin. Invest. **67**, 1712-1723

Lesuffleur, T., Roches, F., Hill, A.S., Lacasa, M., Fox, M., Swallow, D.M., Zweibaum, A., Real, F.X. (1995) Characterization of a mucin cDNA clone isolated from HT-29 mucus-secreting cells.
J. Biol. Chem. **270**, 13665-13673

Levine, M.J., Herzberg, M.C., Levine, M.S., Ellison, S.A., Stinson, M.W., Li, H.C., Van Dyke, T. (1978) Specificity of salivary-bacterial interactions : role of terminal sialic acid residues in the interaction of salivary glycoproteins with *Strepococcus sanguis* and *Streptococcus mutans*.
Infect. Immun. **19**, 107-115

Levine, M.J., Reddy, M.S., Tabak, L.A., Loomis, R.E., Bergey, E.J., Jones, P.C., Cohen, R.E., Stinson, M.W., Al-Hashimi, I. (1987) Structural aspects of salivary glycoproteins.
J. Dent. Res. **66**, 436-441

Li, J.D., Dohrman, A.F., Gallup, M., Miyata, S., Gum, J.R., Kim, Y.S., Nadel, J.A., Prince, A., Basbaum, C.B. (1997) Transcriptional activation of mucin by *Pseudomonas aeruginosa* lipopolysaccharide in the pathogenesis of cystic fibrosis lung disease.
Proc. Natl. Acad. Sci. USA **94**, 967-972

Ligtenberg, M.J.L, Vos, H.L., Gennissen, A.M.C., Hilkens, J. (1990) Episialine, a carcinoma-associated mucin, is generated by polymorphic gene encoding splice variants with alternative amino termini.
J. Biol. Chem. **265**, 5573-5578

Litvinov, S.V., Hilkens, J. (1993) The epithelial sialomucin, episialin, is sialylated during recycling.
J. Biol. Chem. **268**, 21364-21371

Long, M., Rosenberg, C., Gilbert, W. (1995) Intron phase correlations and the evolution of the intron/exon structure of genes.
Proc. Natl. Acad. Sci. USA **92**, 12495-12499

Loomis, R.E., Prakobphol, A., Levine, M.J., Reddy, M.S., Jones, P.C. (1987) Biochemical and biophysical comparison of two mucins from human submandibular-sublingual saliva.
Arch. Biochem. Biophys. **258**, 452-464

Mancuso, D.J., Tuley, L.A., Westfield, L.A., Worrall, N.K., Shelton-Inloes, B.B., Sorace, J.S., Alevy, Y.G., Sadler, J.E. (1989) Structure of the gene for human von Willebrand factor.
J. Biol. Chem. **264**, 19514-19527

Mancuso, D.J., Tuley, L.A., Westfield, L.A., Lester-Mancuso, T.L., Le Beau, M.M., Sorace, J.S., Sadler, J.E. (1991) Human von Willebrand factor gene and pseudogene: structural analysis and differentiation by polymerase chain reaction.
Biochemistry **30**, 253-269

Mantle M., Forstner, G.G., Forstner J.F. (1984) Biochemical characterization of the component parts of intestinal mucin from patients with cystic fibrosis.
Biochem. J **224**, 345-354

Mantle, M. (1994) Rôle antiadhérent du mucus intestinal : mécanismes et physiopathologie.
Mucus Dialogue **2**, 1-4 *IPSEN*

Massagué, J. (1990) The transforming growth factor-β family.
Annu. Rev. Cell Biol. **6**, 597-641

Mayadas, T.N., Wagner, D.D. (1991) von Willebrand factor biosynthesis and processing.
Ann. N. Y. Acad. Sci. **614**, 153-166

Mayadas, T.N., Wagner, D.D. (1992) Vicinal cysteines in the prosequence play a role in von Willebrand factor multimer assembly.
Proc. Natl. Acad. Sci. USA **89**, 3531-3535

McDonald, N.Q., Lapatto, R., Murray-Rust, J., Gunning, J., Wlodawer, A., Blundell, T.L. (1991) New protein fold revealed by 2.3-Å resolution crystal structure of nerve growth factor.
Nature (London) **354**, 411-414

McDonald, N.Q., Hendrickson, W.A. (1993) A structural superfamily of growth factors containing a cystine knot motif.
Cell **73**, 421-424

McPherron, A.C., Lee, S.J. (1993) GDF-3 and GDF-9: two new members of the transforming growth factor-β superfamilly containing a novel pattern of cysteines.
J. Biol. Chem. **268**, 3444-3449

Meerzaman, D., Charles, P., Daskal, E., Polymeropoulos, M.H., Martin, B.M., Rose, M.C. (1994) Cloning and analysis of cDNA encoding a major airway glycoprotein, human tracheobronchial mucin (MUC5).
J. Biol. Chem. **269**, 12932-12939

Meindl, A., Berger, W., Meitinger, T., van de Pol, D., Achatz, H., Dörner, C., Haasemann, M., Hellebrand, H., Gal, A., Cremers, F., Ropers, H.H. (1992) Norrie disease is caused by mutations in an extracellular protein resembling C-terminal globular domain of mucins.
Nat. Genet. **2**, 139-143

Meitinger, T., Meindl, A., Bork, P., Rost, B., Sander, C., Haasemann, M., Murken, J. (1993) Molecular modeling of the Norrie disease protein predicts a cystine knot growth factor tertiary structure.
Nat. Genet. **5**, 376-380

Moriarty, J., Skelly, C.M., Bharathan, S., Moody, C.E., Sherblom, A.P. (1990) Sialomucin and lytic susceptibility of rat mammary tumor ascites cells.
Cancer Res. **50**, 6800-6805

Mount, S.M. (1982) A catalogue of splice junction sequences.
Nucleic Acids Res. **10**, 459-472

Murray, P.J., Spithill, T.W. (1991) Variants of a *Leishmania* surface antigen derived from a multigenic family.
J. Biol. Chem. **266**, 24477-24484

Murray, P.A., Prakobphol, A., Lee, T., Hoover, C.I., Fisher, S.J. (1992) Adherence of oral streptococci to salivary glycoproteins.
Infect. Immun. **60**, 31-38

Murty, V.L., Downs, F.J., Pigman, W. (1978) Rat-colonic, mucus glycoprotein.
Carbohydr. Res. **61**, 139-145

Nguyen P.L., Niehans, G.A., Cherwitz, D.L., Ho, S.B. (1996) Membrane-bound (MUC1) and secretory (MUC2, MUC3 and MUC4) mucin gene expression in human lung cancer.
Tumor Biol. **17**, 176-192

Nguyen, V.C., Aubert, J.P., Gross, M.D., Porchet, N., Degand, P., Frézal, J. (1990) Assignment of tracheobronchial mucin gene(s) to 11p15 and a tracheobronchial mucin-related sequence to chromosome 13.
Hum. Genet. **86**, 167-172

Nichols, W.C., Ginsburg, D. (1997) von Willebrand disease.
Medicine **76**, 1-20

Nielsen, P.A., Mandel, U., Therkildsen, M.H., Clausen, H. (1996) Differential expression of human high-molecular-weight salivary mucin (MG1) and low-molecular-weight salivary mucin (MG2).
J. Dent. Res. **75**, 1820-1826

Nielsen, P.A., Bennett, E.P., Wandall, H.H., Therkildsen, M.H., Hannibal, J., Clausen, H. (1997) Identification of a major human high molecular weight salivary mucin (MG1) as tracheobronchial mucin MUC5B.
Glycobiology **7**, 413-419

Niv, Y.(1994) Mucin and colorectal cancer metastasis.
Am. J. Gastroenterol. **89**, 665-669

Naora, H., Miyahara, K., Curnow, R.N. (1997) Origin of noncoding DNA sequences: molecular fossils of genome evolution.
Proc. Natl. Acad. Sci. USA **84**, 6195-6199

Noia, J.M.D., Sanchez, D.O., Frasch, A.C.C. (1995) The protozoan *Trypanosoma cruzi* has a family of genes resembling the mucin genes of mammalian cells.
J. Biol. Chem. **270**, 24146-24149

Nunes, D.P., Keates, A.C., Afdhal, N.H., Offner, G.D. (1995) Bovine gall-bladder mucin contains two distinct tandem repeating sequences : evidence for scavenger receptor cysteine-rich repeats.
Biochem. J. **310**, 41-48

Ogata, S., Uehara, H., Chen, A., Itzkowitz, S.H. (1992) Mucin gene expression in colonic tissues and cell lines.
Cancer Res. **52**, 5971-5978

Ohmori, H., Dohrman, A.F., Gallup, M., Tsuda, T., Kai, H., Gum, J.R., Kim, Y.S., Basbaum, C.B. (1994) Molecular cloning of the amino-terminal region of a rat MUC2 mucin gene homologue. Evidence for expression in both intestine and airway.
J. Biol. Chem. **269**, 17833-17840

Parry, G., Beck, J.C., Moss, L., Bartley, J., Ojakian, G.K. (1990) Determination of apical membrane polarity in mammary epithelial cell cultures : the role of cell-cell, cell-substratum, and membrane-cytoskeleton interactions.
Exp. Cell Res. **188**, 302-311

Pasquier, M.C., Vatier, J. (1990) Mucus gastro-intestinal: une barrière protectrice complexe.
Gastroenterol. Clin. Biol. **14**, 352-358

Patthy, L. (1996) Exon shuffling and other ways of module exchange.
Matrix Biol. **15**, 301-310

Paulsson, G., Bernholm, K., Wieslander, L. (1992) Conserved and variable repeat structures in the Balbiani ring gene family in *Chironomus tentans*.
J. Mol. Evol. **35**, 205-216

Pearson, J.P., Allen, A., Parry, S. (1981) A 70000-molecular-weight protein isolated from purified pig gastric mucus glycoprotein by reduction of disulphide bridges and its implication in the polymeric structure.
Biochem. J. **197**, 155-162

Pemberton, L.F., Rughetti, A., Taylor-Papadimitriou, J., Gendler, S.J. (1996) The epithelial mucin MUC1 contains at least two discrete signals specifying membrane localization in cells.
J. Biol. Chem. **271**, 2332-2340

Perez-Vilar, J., Eckhardt, A.E., Hill, R.L. (1996) Porcine submaxillary mucin forms disulfide-bonded dimers between its carboxyl-terminal domains.
J. Biol. Chem. **271**, 9845-9850

Pigny, P. Les gènes de mucines humaines localisés en 11p15. Polymorphisme, cartographie physique et approche de régulation.
Thèse d'Université, Lille, 1997

Pigny, P., Guyonnet Dupérat, V., Hill, A.S., Pratt, W.S., Galiegue-Zouitina, S., Collyn d'Hooge, M., Laine, A., Van-Seuningen, I., Degand, P., Gum, J.R., Kim, Y.S., Swallow, D.M., Aubert, J.P., Porchet, N. (1996a) Human mucin genes assigned to 11p15: identification and organization of a cluster of genes.
Genomics **38**, 340-352

Pigny, P., Van Seuningen, I., Desseyn, J.L., Nollet, S., Porchet, N., Laine, A., Aubert, J.P. (1996b) Identification of a 42 kDa nuclear factor (NF1-MUC5B) from HT-29 MTX cells that binds to the 3' region of human mucin gene *MUC5B.*
Biochem. Biophys. Res. Commun. **220**, 186-191

Podolsky, D.K., Lynch-Devaney, K., Stow, J.L., Oates, P., Murgue, B., DeBeaumont, M., Sands, B.E., Mahida, Y.R. (1993) Identification of human intestinal trefoil factor. Goblet cell-specific expression of a peptide targeted for apical secretion.
J. Biol. Chem. **268**, 6694-6702

Porchet, N., Nguyen, V.C., Dufossé, J., Audié, J.P., Guyonnet Dupérat, V., Gross, M.S., Denis, C., Degand, P., Bernheim, A., Aubert, J.P. (1991) Molecular cloning and chromosomal localization of a novel human tracheo-bronchial mucin cDNA containing tandemly repeated sequences of 48 base pairs.
Biochem. Biophys. Res. Commun. **175**, 414-422

Prakobphol, A., Levine, M.J., Tabak, L.A., Reddy, M.S. (1982) Purification of low-molecular-weight, mucin-type glycoprotein from human submandibular-sublingual saliva.
Carbohydr. Res. **108**, 111-122

Probst, J.C., Gertzen, E.M., Hoffmann, W. (1990) An integumentary mucin (FIM-B.1) from *Xenopus laevis* homologous with von Willebrand factor.
Biochemistry **29**,6240-6244

Probst, J.C., Hauser, F., Joba, W., Hoffmann, W. (1992) The polymorphic integumentary mucin B.1 from *Xenopus laevis* contains the short consensus repeat.
J. Biol. Chem. **267**, 6310-6316

Reddy, M.S., Bobek, L.A., Haraszthy, G.G., Biesbrock, A.R., Levine, M.J. (1992) Structural features of the low-molecular-mass human salivary mucin.
Biochem. J. **287**, 639-643

Reid, M.M., Bhaskar, K.R., Coles, S. (1982) Clinical aspects of respiratory mucus.
Adv. Exp. Med. Biol. **14**, 369-391

Roberton, A.M., Mantle, M., Fahim, R.E.F., Specian, R.D., Bennick, A., Kawagishi, S., Sherman, P., Forstner, J.F. (1989) The putative 'link' glycopeptide associated with mucus glycoproteins. Composition and properties of preparations from the gastrointestinal tracts of several mammals.
Biochem. J. **261**, 637-647

Rose, M.C., Voter, W.A., Sage, H., Brown, C.F., Kaufman, B. (1984) Effects of deglycosylation on the architecture of ovine submaxillary mucin glycoprotein.
J. Biol. Chem. **259**, 3167-3172

Rosen, V., Thies, R.S. (1992) The BMP proteins in bone formation and repair.
TIG **8**, 97-102

Rossi, E.A., McNeer, R.R., Price-Schiavi, S.A., Van den Brande J.M.H., Komatsu, M., Thompson, J.F., Carothers Carraway, C.A., Fregien, N.L., Carraway, K.L. (1996) Sialomucin complex, a heterodimeric glycoprotein complex. Expression as a soluble, secretable form in lactating mammary gland and colon.
J. Biol. Chem. **271**, 33476-33485

Rothnagel, J.A., Steinert, P.M. (1990) The structure of the gene for mouse filaggrin and a comparison of the repeating units.
J. Biol. Chem. **265**, 1862-1865

Roussel P., Lamblin, G., Degand, P. (1975) Heterogeneity of the carbohydrate chains of sulfated bronchial glycoproteins isolated from a patient suffering from cystic fibrosis.
J. Biol. Chem. **250**, 2114-2122

Roussel P., Lamblin, G. (1996) Human mucosal mucins in diseases.
Editions J. Montreuil, J.F.G. Vliegenthart et H. Schachter, *Glycoproteins and disease* pp. 351-393

Ruggeri, Z.M., Ware, J. (1993) von Willebrand factor.
FASEB J. **7**, 308-316

Ruoslahti, E., Yamaguchi, Y. (1991) Proteoglycans as modulators of growth factor activities.
Cell **64**, 867-869

Sajjan, S.U., Forstner, J.F. (1990a) Characteristics of binding of *Escherichia coli* serotype O157:H7 strain CL-49 to purified intestinal mucin.
Infect. Immun. **58**, 860-867

Sajjan, S.U., Forstner, J.F. (1990b) Role of the putative "link" glycopeptide of intestinal mucin in binding of piliated *Escherichia coli* serotype O157:H7 strain CL-49.
Infect. Immun. **58**, 868-873

Sampath, T.K., Rashka, K.E., Doctor, J.S., Tucker, R.F., Hoffmann, F.M. (1993) *Drosophila* transforming growth factor β superfamilly proteins induce endochondral bone formation in mammals.
Proc. Natl. Acad. Sci. USA **90**, 6004-6008

Sanger, F., Nicklen, S., Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors.
Proc. Natl. Acad. Sci. U.S.A. **74**, 5463-5467

Schlunegger, M.P., Grütter, M.G. (1992) An unusual feature revealed by the crystal structure at 2.2 Å resolution of human transforming growth factor-β2.
Nature (London) **358**, 430-434

Scholl, S.M., Pouillart, P. (1997) Vaccination contre les cellules tumorales mammaires exprimant la mucine MUC1.
Bull. Cancer **84**, 61-64

Seib, T., Blin, N., Hilgert, K., Seifert, M., Theisinger, B., Engel, M., Dooley, S., Zang, K.D., Welter, C. (1997) The three human trefoil genes TFF1, TFF2 and TFF3 are located within a region of 55 kb on chromosome 21q22.3.
Genomics **40**, 200-202

Shankar, V., Naziruddin, B., Reyes de la Rocha, S., Sachdev, G.P. (1990) Evidence of hydrophobic domains in human respiratory mucins. Effect of sodium chloride on hydrophobic binding properties.
Biochemistry **29**, 5856-5864

Shankar, V., Gilmore, M.S., Sachdev, G.P. (1992) Molecular cloning of the carboxy terminus of canine tracheobronchial mucin.
Biochem. Biophys. Res. Commun. **189**, 958-964

Shankar, V., Gilmore, M.S., Elkins, R.C., Sachdev, G.P. (1994) A novel human airway mucin cDNA encodes a protein with unique tandem-repeat organization.
Biochem. J. **300**, 295-298

Shankar, V., Pichan, P., Eddy, R.L., Sait, S.N.J., Nowak, N., Shows, T.B., Elkins, R.C., Gilmore, M.S., Sachdev, G.P. (1995) A major airway mucin gene, *MUC8* : chromosomal localization and cloning of complementary DNA.
Glycoconj. J. **12**, S19 p497

Shankar, V., Pichan, P., Eddy, R.L., Tonk, V., Nowak, N., Sait, S.N.J., Shows, T.B., Schultz, R.E., Gotway, G., Elkins, R.C., Gilmore, M.S., Sachdev, G.P. (1997) Chromosomal localization of a human mucin gene (*MUC8*) and cloning of the cDNA corresponding to the carboxy terminus.
Am. J. Respir. Cell Mol. Biol. **16**, 232-241

Sheehan, J.K., Oates, K., Carlstedt, I. (1986) Electron microscopy of cervical, gastric and bronchial mucus glycoproteins.
Biochem. J. **239**, 147-153

Sheehan, J.K., Boot-Handford, R.P., Chantler, E., Carlstedt, I., Thornton, D.J. (1991) Evidence for shared epitopes within the 'naked' protein domains of human mucus glycoproteins. A study performed by using polyclonal antibodies and electron microscopy.
Biochem. J. **274**, 293-296

Sheehan, J.K., Thornton, D.J., Howard, M., Carlstedt, I., Corfield, A.P., Paraskeva, C. (1996) Biosynthesis of the MUC2 mucin : evidence for a slow assembly of fully glycosylated units.
Biochem. J. **315**, 1055-1060

Shekels, L.L., Lyftogt, C., Kieliszewski, M., Filie, J.D., Kozak, C.A., Ho, S.B. (1995) Mouse gastric mucin : cloning and chromosomal localization.
Biochem. J. **311**, 775-785

Shelton-Inloes, B.B., Titani, K., Sadler, J.E. (1986) cDNA sequences for human von Willebrand factor reveal five types of repeated domains and five possible protein sequence polymorphisms.
Biochemistry **25**, 3164-3171

Shelton-Inloes, B.B., Broze, G.J., Miletich, J.P., Sadler, J.E. (1987) Evolution of human von Willebrand factor : cDNA sequence polymorphism, repeated domains, and relationship to von Willebrand antigen II.
Biochem. Biophys. Res. Commun. **144**, 657-665

Sheng, Z., Hull, S.R., Carraway, K.L. (1990) Biosynthesis of the cell surface sialomucin complex of ascites 13762 rat mammary *adenocarcinoma* cells from a high molecular weight precursor.
J. Biol. Chem. **265**, 8505-8510

Sheng, Z., Wu, K., Carraway, K.L., Fregien, N.L. (1992) Molecular cloning of the transmembrane component of the 13762 mammary adenocarcinoma sialomucin complex. A new member of the epidermal growth factor superfamily.
J. Biol. Chem. **267**, 16341-16346

Shimizu Y, Shaw S (1993) Cell adhesion. Mucins in the mainstream.
Nature (London) **366**, 630-631

Slomiany, A., Witas, H., Aono, M., Slomiany, B.L. (1983) Covalently linked fatty acids in gastric mucus glycoprotein of cystic fibrosis patients.
J. Biol. Chem. **258**, 8535-8538

Slomiany, A., Tamura, S., Grzelinska, E., Piotrowski, J., Slomiany, B.L. (1992) Mucin complexes: characterization of the "link" component of submandibular mucus glycoprotein.
Int. J. Biochem. **24**, 1003-1015

Smith, B.F., LaMont, J.T. (1984) Hydrophobic binding properties of bovine gallbladder mucin.
J. Biol. Chem. **259**, 12170-12177

Smith, B.F., LaMont, J.T. (1985) Identification of gallbladder mucin-bilirubin complex in human cholesterol gallstone matrix. Effects of reducing agents on in vitro dissolution of matrix and intact gallstones.
J. Clin. Invest. **76**, 439-445

Snary, D., Allen, A., Pain, R.H. (1974) Conformational changes in gastric mucoproteins induced by caesium chloride and guanidium chloride.
Biochem. J. **141**, 641-646

Snyder, C.E., Nadziejko, C.E., Herp, A. (1982) Isolation of bronchial mucins from cystic fibrosis sputum by use of citraconic anhydride.
Carbohyd. Res. **105**, 87-93

Spieth, J., Nettleton, M., Zucker-Aprison, E., Lea, K., Blumenthal, T. (1991) Vitellogenin motifs conserved in nematodes and vertebrates.
J. Mol. Evol. **32**, 429-438

Sporn, M.B., Roberts, A.B. (1992) Transforming growth factor-β : recent progress and new challenges.
J. Cell Biol. **119**, 1017-1021

Strasberg, P., Liede, H.A., Stein, T., Warren, I., Sutherland, J., Ray, P. (1995) A novel mutation in the Norrie disease gene predicted to disrupt the cystine knot growth factor motif.
Hum. Mol. Genet **4**, 2179-2180

Strous, G.J., Dekker, J. (1992) Mucin-type glycoproteins.
Crit. Rev. Biochem. & Mol. Biol. **27**, 57-92

Suemori, S., Lynch-Devaney, K., Podolsky, D.K. (1991) Identification and characterization of rat intestinal trefoil factor: tissue- and cell-specific member of the trefoil protein family.
Proc. Natl. Acad. Sci. USA **88**, 11017-11021

Sun, P.D., Davies, D.R. (1995) The cystine-knot growth-factor superfamily.
Annu. Rev. Biophys. Biomol. Struct. **24**, 269-91

Swallow, D.M., Gendler, S., Griffiths, B., Kearney, A., Povey, S., Sheer, D., Palmer, R.W., Taylor-Papadimitriou, J. (1987) The hypervariable gene locus PUM, which codes for the tumour associated epithelial mucins, is located on chromosome 1, within the region 1q21-24.
Ann. Hum. Genet. **51**, 289-294

Tabak, L.A., Levine, M.J., Mandel, I.D., Ellison, S.A. (1982) Role of salivary mucins in the protection of the oral cavity.
J. Oral Pathol. **11**, 1-17

Tabak, L.A. (1995) In defense of the oral cavity : structure, biosynthesis, and function of salivary mucins.
Annu. Rev. Physiol. **57**, 547-564

Taipale, J., Keski-Oja, J. (1997) Growth factors in the extracellular matrix.
FASEB J. **11**, 51-59

Thornton, D.J., Sheehan, J.K., Lindgren, H., Carlstedt, I. (1991) Mucus glycoproteins from cystic fibrotic sputum. Macromolecular properties and structural 'architecture'.
Biochem. J. **276**, 667-675

Thornton, D.J., Carlstedt, I., Howard, M., Devine, P.L., Price, M.R., Sheehan, J.K. (1996) Respiratory mucins: identification of core proteins and glycoforms.
Biochem. J. **316**, 967-975

Thornton, D.J., Howard, M., Khan, N., Sheehan, J.K. (1997) Identification of two glycoforms of the MUC5B mucin in human respiratory mucus. Evidence for a cysteine-rich sequence repeated within the molecule.
J. Biol. Chem. **272**, 9561-9566

Titani, K., Kumar, S., Takio, K., Ericsson, L.H., Wade, R.D., Ashida, K., Walsh, K.A., Chopek, M.W., Sadler, J.E., Fujikawa, K. (1986) Amino acid sequence of human von Willebrand factor.
Biochemistry **25**, 3171-3184

Toribara, N.W., Gum, J.R., Culhane, P.J., Lagace, R.E., Hicks, J.W., Petersen, G.M., Kim, Y.S. (1991) MUC-2 human small intestinal mucin gene structure (Repeated arrays and polymorphism).
J. Clin. Invest. **88**, 1005-1013

Gum, J.R., Hicks, J.W., Toribara, N.W., Rothe, E.M., Lagace, R.E., Kim, Y.S. (1992) The human MUC2 intestinal mucin has cysteine-rich subdomains located both upstream and downstream of its central repetitive region.
J. Biol. Chem. **267**, 21375-21383

Toribara, N.W., Roberton, A.M., Ho, S.B., Kuo, W.L., Gum, E., Hicks, J.W., Gum, J.R., Byrd, J.C., Siddiki, B., Kim, Y.S. (1993) Human gastric mucin. Identification of a unique sequence by expression cloning.
J. Biol. Chem. **268**, 5879-5885

Toribara, N.W., Ho, S.B., Gum, E., Gum, J.R , Lau, P., Kim, Y.S. (1997) The carboxyl-terminal sequence of the human secretory mucin, MUC6. Analysis of the primary amino acid sequence.
J. Biol. Chem. **272**, 16398-16403

Troxler, R.F., Offner, G.D., Zhang, F., Iontcheva, I., Oppenheim, G.O. (1995) Molecular cloning of a novel high molecular mucin (MG1) from human sublingual gland.
Biochem. Biophys. Res. Commun. **217**, 1112-1119

Turner, B.S., Bhaskar, R., Hadzopoulou-Cladaras, M., Specian, R.D. (1995) Isolation and characterisation of cDNA clones encoding pig gastric mucin.
Biochem. J. **308**, 89-96

Tytgat, K.M.A.J., Bovelander, F.J., Opdam, F.J.M., Einerhand, A.W.C., Büller, H.A., Dekker, J. (1995) Biosynthesis of rat MUC2 in colon and its analogy with human MUC2.
Biochem. J. **309**, 221-229

Vandenhaute, B., Buisine, M.P., Debailleul, V., Clément, B., Moniaux, N., Dieu, M.C., Degand, P., Porchet, N., Aubert, J.P. (1997) Mucin gene expression in biliary epithelial cells.
J. Hepathol. (sous presse)

Van Klinken, B.J.W., Dekker, J., Büller, H.A., Einerhand, A.W.C. (1995) Mucin gene structure and expression : protection vs. adhesion.
Am. J. Physiol. **269**, G613-G627

Velcich, A., Palumbo, L., Selleri, L., Evans, G., Augenlicht, L. (1997) Organization and regulatory aspects of the human intestinal mucin gene (MUC2) locus.
J. Biol. Chem. **272**, 7968-7976

Verma, M., Davidson, E.A. (1993) Molecular cloning and sequencing of a canine tracheobronchial mucin cDNA containing a cysteine-rich domain.
Proc. Natl. Acad. Sci. USA **90**, 7144-7148

Verma, M. (1994) Carcinoma associated mucins : molecular biology and clinical applications.
Cancer Biochem. Biophys. **14**, 151-162

Verma, M., Murthy, V.V.S., Mathew, S., Banerji, D., Kurl, R.N., Olnes, M.J., Yankaskas, J.R., Blass, C., Davidson, E.A. (1996) Promoter of the canine tracheobronchial mucin gene.
Glycoconj. J. **13**, 797-807

Veerman, E.C., Ligtenberg, A.J., Schenkels, L.C., Walgreen-Weterings, E., Nieuw Amerongen, A.V. (1995) Binding of human high-molecular-weight salivary mucins (MG1) to Hemophilus parainfluenzae.
J. Dent. Res. **74**, 351-357

Verweij, C.L., Diergaarde, P.J., Hart, M., Pannekoek, H. (1986) Full-length von Willebrand factor (vWF) cDNA encodes a highly repetitive protein considerably larger than the mature vWF subunit.
EMBO J. **5**, 1839-1847

Verweij, C.L., Hart, M., Pannekoek, H. (1987) Expression of variant von Willebrand factor (vWF) cDNA in heterologous cells : requirement of the pro-polypeptide in vWF multimer formation.
EMBO J. **6**, 2885-2890

Verweij, C.L., Hart, M., Pannekoek, H. (1988) Proteolytic cleavage of the precursor of von Willebrand factor is not essential for multimer formation.
J. Biol. Chem. **263**, 7921-7924

Vinall, L.E., Hill, A.S., Pigny, P., Pratt, W.S., Toribara, N., Gum, J.R., Kim, Y.S., Porchet, N., Aubert, J.P., Swallow, D.M. (1997) Variable number tandem repeat polymorphism of the mucin genes located in the complex on 11p15.
Soumis

Voorberg, J., Fontijn, R., Calafat, J., Janssen, H., van Mourik, J.A., Pannekoek, H. (1991) Assembly and routing of von Willebrand factor variants : the requirements for disulfide-linked dimerization reside within the carboxy-terminal 151 amino acids.
J. Cell. Biol. **113**, 195-205

Wagner, D.D., Saffaripour, S., Bonfanti, R., Sadler, J.E., Cramer, E.M., Chapman, B., Mayadas, T.N. (1991) Induction of specific storage organelles by von Willebrand factor polypeptide.
Cell **64**, 403-413

Wellman, S.E., Case, S.T. (1989) Disassembly and reassembly *in vitro* of complexes of secretory proteins from *Chironomus tentans* salivary glands.
J. Biol. Chem. **264**, 10878-10883

Wen D., Peles, E., Cupples, R., Suggs, S.V., Bacus, S.S., Luo, Y., Trail, G., Hu, S., Silbiger, S.M., Levy, R.B., Koski, R.A., Lu, H.S., Yarden, Y. (1992) Neu differentiation factor: a transmembrane glycoprotein containing an EGF domain and an immunoglobulin homology unit.
Cell **69**, 559-572

Wise, R.J., Pittman, D.D., Handin, R.I., Kaufman, R.J., Orkin, S.H. (1988) The propeptide of von Willebrand factor independently mediates the assembly of von Willebrand multimers.
Cell **52**, 229-236

Woodward, H., Horsey, B., Bhavanandan, V.P., Davidson, E.A. (1982) Isolation, purification, and properties of respiratory mucus glycoproteins.
Biochemistry **21**, 694-701

Wreschner, D.H., Hareuveni, M., Tsarfaty, I., Smorodinsky, N., Horev, J., Zaretsky, J., Kotkes, P., Weiss, M., Lathe, R., Dion, A., Keydar, I. (1990) Human epithelial tumor antigen cDNA sequences. Differential splicing may generate multiple protein forms.
Eur. J. Biochem. **189**, 463-473

Wu, K., Fregien, N.L., Carraway, K.L. (1994) Molecular cloning and sequencing of the mucin subunit of a heterodimeric, bifunctional cell surface glycoprotein complex of ascites rat mammary adenocarcinoma cells.
J. Biol. Chem. **269**, 11950-11955

Xu, G., Huan, L., Khatri, I., Sajjan, U.S., McCool, D., Wang, D., Jones, C., Forstner, G.G., Forstner, J.F. (1992a) Human intestinal mucin-like protein (MLP) is homologous with rat MLP in the C-terminal region, and is encoded by a gene on chromosome 11p15.5.
Biochem. Biophys. Res. Commun. **183**, 821-828

Xu, G., Huan, L., Khatri, I., Wang, D., Bennick, A., Fahim, R.E.F., Forstner, G.G., Forstner, J.F. (1992b) cDNA for the carboxyl-terminal region of a rat intestinl mucin-like peptide.
J. Biol. Chem. **267**, 5401-5407

Yu, C.H., Yang, P.C., Shew, J.Y., Hong, T.M., Yang, S.C., Lee, L.N., Luh, K.T., Wu, C.W. (1996) Mucin mRNA expression in lung adenocarcinoma cell lines and tissues.
Oncology **53**, 118-126

Zhou, X., Sasaki, H., Lowe, L., Hogan, B.L.M., Kuehn, M.R. (1993) *Nodal* is a novel TGF-β-like gene expressed in the mouse node during gastrulation.
Nature (London) **361**, 543-547

Zrihan-Licht, S., Vos, H.L., Baruch, A., Elroy-Stein, O., Sagiv, D., Keydar, I., Hilkens, J., Wreschner, D.H. (1994) Characterization and molecular cloning of a novel MUC1 protein, devoid of tandem repeats, expressed in human breast cancer tissue.
Eur. J. Biochem. **224**, 787-795