Evolution of Transmembrane and Gel-forming Mucins Studied with Bioinformatic Methods

Tiange Lang

2007

Department of Medical Biochemistry and Cell Biology
Institute of Biomedicine
Sahlgrenska Academy
Göteborg University

A doctoral thesis at a university in Sweden is produced either as a monograph or as a collection of papers. In the latter case, the introductory part constitutes the formal thesis, which summaries the accompanying papers. These papers have already been published or are in manuscripts at various stages (in press, submitted or accepted).

Department of Medical Biochemistry and Cell Biology
Institute of Biomedicine, Sahlgrenska Academy,
Göteborg University, Box 440, SE-405 30 Göteborg

Printed by Intellecta Docusys, Göteborg 2007
ISBN 978-91-628-7320-2

Abstract

All mucosal membranes of the body are covered by mucus, largely made up of the family of large glycoproteins called mucins. These are instrumental for the protection of the underlying epithelia and involved in the pathogenesis of many diseases in the lungs and the intestine. Several mucins are also involved in the progression of cancer and can often be linked to bad prognosis. The mucins are classified as membrane-bound or secreted. In human there are eight membrane-bound (MUC1, MUC3, MUC4, MUC12, MUC13, MUC15, MUC16 and MUC17) and five secreted and gel-forming mucins (MUC2, MUC5B, MUC5AC, MUC6, and MUC19). Mucins are characterized by domains rich in proline, serine and threonine that are heavily glycosylated (PTS domains) and typically have either von Willebrand D (VWD) or SEA domains.

To aid in understanding this family of proteins we have taken a bioinformatics approach to mine protein and genomic sequence databases for mucins. We have combined different methods to predict mucin proteins. We developed PTSpred, a method to predict PTS domains characteristic of mucins. We also made use of prediction of signal sequences, transmembrane regions, profile based searches and methods to predict genomic regions encoding specific protein domains. We have examined several animals with respect to mucins and other proteins that contain the VWD and SEA domains and have identified numerous novel mucin homologues and mucin-related proteins.

We first made an comprehsnsive inventory of human, mouse and rat mucins including the human chromosome 7q22 region which encodes MUC3, MUC12, and MUC17. During the analysis of the chicken genome we found that the homologues of human chromosome 11p15 gel-forming mucin group (MUC6, MUC2, MUC5AC and MUC5B) are found with the same order as in human, and Muc13 is encoded by a gene where the PTS domain is divided among several exons, where each exon encodes a repeated unit in the protein.

The mucins in *Xenopus tropicalis* are unusual in many respects. The number of gel-forming mucins has been markedly expanded, and the Muc2 homologues contain an unusual PTS domain rich in cysteines. In addition, *Xenopus tropicalis* has a novel family of mucin-like proteins with alternating PTS and SEA domains, a type of protein also identified in the fishes.

The evolution of the MUC4 mucin seems to have occurred by recruitment of a PTS domain to NIDO, AMOP and VWD domains from a sushi-domain containing family of proteins present in lower animals. *Xenopus tropicalis* is the most deeply branching animal where a protein similar to the mammalian MUC4 was identified.

In the gel-forming mucins, a VWD domain typically occurs together with a TIL domain and a domain we have named VWE. We also demonstrated that the gel-forming mucins, von Willebrand factor (VWF), otogelin and insect hemolectin are evolutionary related. Proteins related to these are found in a range of animals, including a mucin in the deeply branching metazoan *Nematostella vectensis* (sea anemone). This demonstrates an early origin of this group of mucins. In contrast, all the transmembrane mucins do not seem to have evolved until the appearance of the vertebrate lineage.

List of Publications

I

Lang T., Alexandersson M., Hansson G. C. and Samuelsson T. (2004)
Bioinformatic identification of polymerizing and transmembrane mucins in the puffer fish Fugu rubripes.
Glycobiology 14:521-527.

II

Lang T., Hansson G. C. and Samuelsson T. (2006)
An inventory of mucin genes in the chicken genome shows that the mucin domain of Muc13 is encoded by multiple exons and that ovomucin is part of a locus of related gel-forming mucins.
BMC Genomics. 7:197-206.

III

Lang, T., Hansson, G. C. and Samuelsson T. (2007)
Gel-forming mucins appeared early in metazoan evolution.
Proc Natl Acad Sci U S A. 104(41):16209-16214.

Table of contents

Introduction

Mucin proteins

Mucins are large, abundant, filamentous glycoproteins that coat the surfaces of cells lining the respiratory, digestive and urogenital tracts, and in some amphibia, the skin. Several of these mucins are known to form mucus layers, whereas others form the glycocalyx for example on the intestinal enterocytes [1]. They serve as a diffusion barrier against contact with noxious substances and as a lubricant to protect the epithelial cells from infection, dehydration, and physical or chemical injury [2, 3]. The mucins are deemed to mediate many interactions between the cells and microorganism and other components found at the outer mucosal surfaces. Mucins can be considered as powerful two-edged swords, keeping unwanted substances and organisms at an arm's length while, at the same time, allowing specific interactions to be mediated through their highly elaborate structures. Their strategic position places the mucins at center stage in many disease processes in which the interaction of epithelial cells and their surrounding have gone astray, as in inflammatory and infectious diseases, cancer and metastasis [4-8] . For example, inflammation of the large intestine causes a disease similar to the inflammatory bowel disease ulcerative colitis in mice that are deficient in the mucin MUC2 [9].

Each mucin has a characteristic part called a mucin domain, which is rich in the amino acids serine, threonine and proline, often organized in tandem repeats. The function of a mucin domain is probably to serve as scaffolds for O-linked glycans and to bind water and interact with microbial or host lectins (carbohydrate binding proteins). The dense oligosaccharide clusters make this domain proteolytically resistant and give an extended and stiff conformation, which is described as a "bottle brush" (Figure 1). The biophysical properties of mucin domains are related to their extensive O-linked glycosylation rather than directly to their polypeptide sequences [10, 11]. In insects, a number of chitin-binding proteins also contain mucin domains, and chitin composes a matrix whose role is similar to the mucus of animals [12].

The highly O-glycosylated mucin domains are also found in other proteins in addition to the molecules where they make up a major part. Some researches, for example, Yang's studies on the Ebola virus glycoprotein [13], show that a better functional understanding of mucin domains is very important. Generally, their composition consists of the amino acids serine and threonine with up to above 50% of the amino acids, and the content of proline are often larger than 5% in the tandem repeat domain of mucin. We

have used PTS domain as another more specific term to describe mucin domains. Some mucin tandem repeats contain up to 75% serine and threonine. In human it is only MUC1 and MUC7 that have mucin domains with less than 40% serine and threonine. However, the total amount of serine, threonine and proline in MUC7 tandem repeat domain is also larger than 50%. From the number of repeats and length of tandem repeat domain, the sequence length of mucin domains that contain the tandem repeat domain is large, generally from about 100 amino acids (MUC 7) to 5,000 amino acids (MUC12).

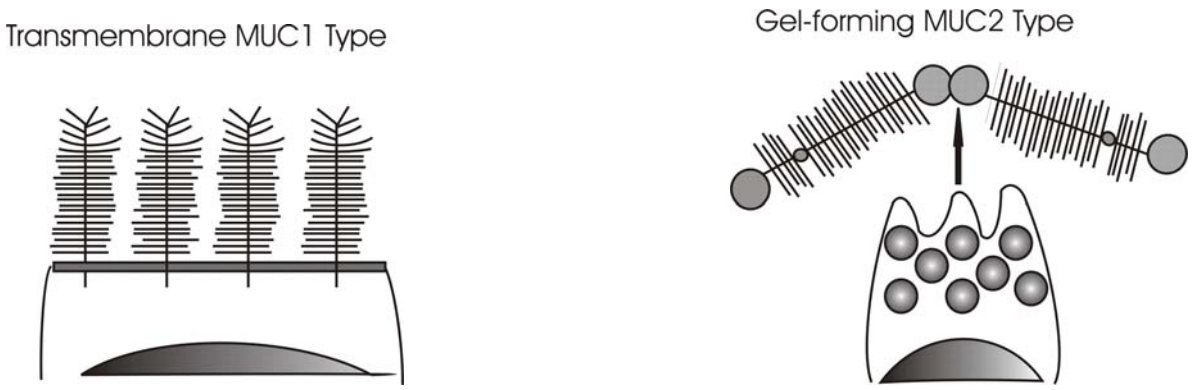

Figure 1. *Mucin types.*

Members of the mucin family vary considerably in size. Some are small and contain a few hundred amino acid residues, whereas others have several thousands of residues and are among the largest known proteins [14]. There are two types of mucins, membrane-bound and secreted [4, 15]. Of the human mucins, eight are membrane-bound (MUC1, MUC3, MUC4, MUC12, MUC13, MUC15, MUC16, MUC17), five are secretory gel-forming (MUC2, MUC5AC, MUC5B, MUC6, MUC19), two are secreted monomeric (MUC7, MUC20).

Transmembrane type mucin MUC1 is believed to play an important role in the normal function of general ductal epithelia and in the pathogenesis of epithelial diseases including adenocarcinoma and cystic fibrosis [4, 16, 17]. MUC1 is known by several names, the most common being PEM (Polymorphic Epithelial Mucin), episialin, DUPAN-2, DF3, HMFG (human milk fat globule), EMA (epithelial membrane antigen), CD227, and MUC1. It has been isolated from various tissue samples including human mammary epithelial cells, ovarian cells, and pancreatic cells. Although in all these tissues, MUC1 apomucin appears to be

identical, each tissue expresses distinct glycoforms with a molecular mass varying from 250 to 500 kDa in the mammary glands or up to 1,000 kDa in the pancreas and intestine [2, 4]. MUC1 interacts directly with the beta-catenin via the SXXXXXSSL motif in its cytoplasmic tail [4, 18]. The beta-catenin is a protein that has important functions in the formation of the cell-cell junction by interaction with the E-cadherin. Several other alternative spliced variants of the MUC1 gene have been discovered. MUC1/SEC is a splice form where the transmembrane and cytoplasmic tail are lacking and thus encoding a secreted mucin [19]. MUC1/Y is characterized by the deletion of the central domain (encoded completely by exon 2), corresponding to the mucin domain. Other splice variants with a deleted tandem repeat array have also been identified, such as MUC1/X or MUC1/Z, but less is known on their surface for these variants [20]. The MUC1-CT80 is still another splice variant where the last intron is not spliced out and the reading into the intron generates an alternative cytoplasmic tail lacking the beta-catenin binding sequence [21].

The gel-forming family of mucins is composed of MUC2, MUC5AC, MUC5B, MUC6, and MUC19. All of these gel-forming mucins are very large; they also share a similar structure and substantial sequence homology in the conserved regions [11, 22-24] [25] . The cDNA sequences of these mucins encode multiple “cysteine-rich” VWC and VWD domains in the flanking region of the mucin-like threonine/serine-rich repeats and Cysteine knot (CK) domains in their C-terminal regions. Both the cysteine number and their positions are extremely conserved in those domains, which play an essential role in forming disulfide-linked dimers and trimers. Their large size and the capability to form multimers support the notion that these mucins have played a pivotal role in forming the mucus gel. Indeed, those gel-forming mucins have been proven to be major components of the mucus secretion of various organs. The genes of MUC2, MUC5AC, MUC5B and MUC6 are clustered in a complex of 400 kb very rich in CpG islands on chromosome 11 in region p15.5 [26]. The cluster is localized between the HRAS and IGF2 genes. The five genes are organized in a complex in which the symmetric distribution of the restriction sites and repetition seems to demonstrate the existence of many duplication events. Computational and phylogenetic analyses suggested an evolutionary history of the five human mucin genes from an ancestor gene common to the human von Willebrand factor gene [20].

MUC2 was first identified and described by Gum et al in 1989 and its full sequence in 1994. The central domain of MUC2 is composed of two mucin or PTS domains [27]. These two sequences are rich in the amino acids serine, threonine, and proline. The second from the N-terminus and largest, is characterized by the perfect repetition of one motif of 23 amino acids. The first, located N-terminal, is composed of an

irregular sequence repeated in tandem with a unit of 347 amino acids. MUC2 also possesses four von Willebrand D domains (VWD or D domains), so called because of their homology with the VWD domains of the von Willebrand factor. The D1, D2 and D3 domains are localized in its N-terminal part, whereas the D4 domain is localized in the C-terminal position in relation to the central mucin domains. D3 is responsible for trimerization in the late secretory pathway [11, 28]. At the end of the C-terminal there is one cysteine knot (CK) domain whose function is to form a dimer [29, 30]. The CK domain is also found in other secreted proteins such as the NDP (Norries Disease Protein). Sequence pattern searches and three-dimensional modeling suggest that the CK domain of the mucins and NDP CK have a tertiary structures similar to that of the transforming growth factor beta and several growth factors [11]. CysD domains can be found adjacent to or within the PTS regions and is also present in other proteins like the vertebrate cartilage intermediate layer protein (CILP) and the oikoplastic epithelium of tunicate [31].

The central domain of MUC5B encodes a large exon containing all mucin domains [32]. Composed of 3570 amino acid residues, the MUC5B central domain is organized by six PTS domains and seven CysD domains which are similar with those of MUC2. The amino acid sequence codes for four VWD domains and CK domain similar to those of MUC2 and von Willebrand factor.

The N-terminal and C-terminal Cys-rich domains of MUC5AC are similar to those of MUC2 and MUC5B [24, 33, 34]. The structural organization of the central part is similar to that of MUC5B, which has several CysD domains with high sequence similarity inserted between PTS domains. The exact number and sequence of central mucin and CysD domains is still not revealed. Comparative molecular analysis of the entire sequence of the 3'-region from MUC5AC and MUC5B points to a remarkable similarity in the size and the distribution of exons, and in the type of splice sites. The N-terminal region, 1,858 amino acid residues long, is composed of the D1, D2 and D3 domain in a way comparable to MUC2 and MUC5B.

Of the four mucins genes clustered on chromosome 11p15, MUC6 is the least studied [35]. The CK domain of MUC6 has approximately 25% amino acid similarity to the CK domain of the human mucins MUC2, 5AC, and 5B and the von Willebrand factor. The MUC6 mucin differs in that it lacks CysD domains and has a short C-terminal region without a VWD 4 domain. MUC19 is a newly found gel-forming mucin. Expression analyses demonstrate that MUC19 is mainly expressed in the mucous cells of some glands, including the major salivary glands (sublingual and submandibular glands), and the submucosal gland of

large airways [36, 37]. As another protein is also encoded from the same gene, the actual expression of MUC19 protein is debated.

MUC4 was initially identified after screening a cDNA library constructed from human tracheo-bronchial mucosa with a polyclonal antiserum raised against deglycosylated glycopeptides from human bronchial mucins [38-40] . It possesses one NIDO domain, one AMOP domain, one VWD domain lacking cysteine, three EGF-like domains, a transmembrane sequence and a cytoplasmic tail. MUC4 is a membrane-bound mucin, in which MUC4alpha is the mucin type-associated subunit and MUC4beta is the growth factor-like subunit. MUC4/X and MUC4/Y forms are expressed by cancer tissues including lung and pancreas [41]. The rat orthologue of the humanMUC4 has been extensively studied by Carraway et al under the name SMC and ASGP1/2 [42].

In human chromosome 7q22, the transmembrane type mucins MUC3, MUC11, MUC12 and MUC17 have been predicted [43]. Of the mucins in the 7q22 cluster, the MUC3 was initially found after the screening of a small intestinal cDNA library using antibodies raised against deglycosylated small intestinal mucins [44]. Two partial cDNA clones encoded 17 amino acid residues repeated in tandem. MUC3 was also the first to be partly cloned, although this was a minor splice variant lacking the transmembrane domain [45]. The MUC3 mucin has later also been suggested to be two genes, MUC3A and MUC3B, with an almost identical sequence [46-48]. The organization of the central repetitive domain of MUC3 remains still unclear. The MUC3 gene, by an alternative splicing mechanism, encodes a family of proteins that can be membrane-bound or secreted [49]. Partial sequences of the MUC11 and MUC12 mucins were first described as down regulated in colon cancer [50]. The first cDNA fragment of MUC12 has been identified using a differential display procedure using colorectal cancer and normal colon samples. The cDNA fragment clone encodes a 28-amino acid degenerated tandem repeat that presents 71% similarity with that of MUC11. The C-terminus of MUC12 was characterized early [50]. It presents the same structure as MUC3, with two EGF-like domains, one transmembrane sequence, and a cytoplasmic tail. Recently, the full sequence of the MUC17 mucin with a PTS domain of about 4073 amino acids was independently described by two groups [51, 52].

The first cDNA clone of MUC7 was isolated by screening a human submandibular gland cDNA library with a rabbit antibody, anti apo-MG2 [53]. MG2 is a low-molecular-mass mucin population (150 to 200 kDa) secreted by the submandibular and sublingual salivary glands. Compared to the other human mucins, MUC7

has a very simple architectural organization. The cDNA sequence of MUC7 encodes a 39 kDa protein of 377 amino acid residues. The distal regions of MUC7 do not exhibit any cysteine rich domain, and only two cysteine residues are present toward its N-terminus. Although MUC7 is a low molecular weight mucin with a simple structure, it might have an important function as an antimicrobial agent in the oral cavity [54]. The two cysteine residues located in the N-terminal region of MUC7 seem to be directly implicated in these activities. Moreover, MUC7 has an anti-candidal activity via its histatin-like domain [55].

Domain structure of mucins

Both secretory gel-forming type and transmembrane type mucin proteins are typically multidomain proteins. For a gel-forming type mucin, a signal sequence is followed by three VWD domains and one PTS domain. At the C-terminal end there is a Cys-Knot (CK) domain. Some gel-forming mucins have a 4th VWD domain between the PTS and CK domain (human MUC2, MUC5AC, MUC5B), while others lack the VWD4 domain (human MUC6, MUC19). Some PTS domains have several CysD domains inserted (human MUC2, MUC5AC, MUC5B) (Figure 2).

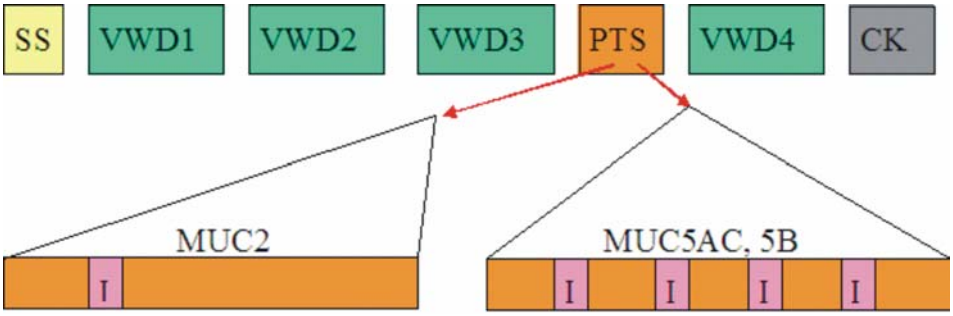

Figure 2. *Principle domain structure of secretory gel-forming type mucins.* I represents CysD domain. The MUC6 and MUC19 mucins have no VWD4 domain.

For a transmembrane type mucin, a signal sequence is followed by a PTS domain and a stalk region. At the C-terminal end there is a transmembrane (TM) domain and a cytoplasmic tail. Commonly, the stalk region part has one SEA domain and some also have one or several EGF domain(s) (human MUC3, MUC12, and MUC17). The stalk of MUC16 is made up of a high number of repeated SEA domains (Figure 3).

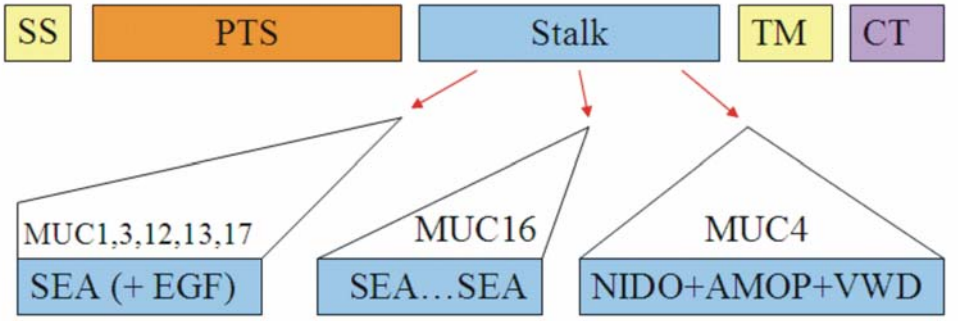

Figure 3. *Domain structure of transmembrane type mucins.*

For the secreted monomeric type mucins, there are no typical domain structures. However, the PTS domain and the signal sequence can always be found in all kinds of mucin proteins. Different human mucin types and domains are listed in Table 1 and selected Pfam domains are in Table 2. Pfam is a comprehensive collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families [56-59].

Table 1. *Mucin type and domains of human* (MUC10 is found only in mouse, MUC11 is part of MUC12, and MUC8, 9, and 18 are other non-mucin proteins).

Mucin	**Subfamily**	**Domains**
MUC1	Membrane-bound	PTS and SEA
MUC2	Secreted gel-forming	PTS, VWC, VWD, CysD and CK
MUC3	Membrane-bound	PTS, SEA and EGF
MUC4	Membrane-bound	PTS, NIDO, AMOP, VWD and EGF
MUC5AC	Secreted gel-forming	PTS, VWD, CysD and CK
MUC5B	Secreted gel-forming	PTS, VWD, CysD and CK
MUC6	Secreted gel-forming	PTS, VWD and CK
MUC7	Secreted monomeric	PTS
MUC12	Membrane-bound	PTS, SEA and EGF
MUC13	Membrane-bound	PTS, SEA and EGF
MUC14	Membrane-bound	PTS and SEA
MUC15	Membrane-bound	PTS
MUC16	Membrane-bound	PTS and SEA
MUC17	Membrane-bound	PTS, SEA and EGF
MUC19	Secreted gel-forming	PTS, VWD, VWC and CK
MUC20	Secreted monomeric	PTS

Table 2. *Selected Pfam domains present in mucin proteins.*

Pfam domain	**Annotation**	**Found-in human**
VWD	von Willebrand factor type D domain.	MUC2, MUC4, MUC5AC, MUC5B, MUC6, MUC19
VWC	von Willebrand factor type C domain.	MUC2, MUC5AC, MUC19
CK	Cys-knot domain. Comprises glycoprotein hormones and the C-terminal domain of various extracellular proteins. It is believed to be involved in disulfide-linked dimerisation.	MUC2, MUC5AC, MUC5B, MUC6, MUC19
SEA	Found in sea urchin sperm protein, Engerokinase, Agrin. Proposed function of regulating or binding carbohydrate side chains.	MUC1, MUC3, MUC12, MUC13, MUC16, MUC17
NIDO	Nidogen-like domain, an extracellular domain found in nidogen and hypothetical proteins of unknown function.	MUC4
AMOP	This domain may have a role in cell adhesion. It is called the AMOP domain after Adhesion associated domain in MUC4 and Other Proteins. This domain is extracellular and contains a number of cysteines that probably form disulphide bridges.	MUC4
EGF	Has 8 conserved cysteines. Includes some cytokine receptors. A large and heterogenous family	MUC3, MUC4, MUC12, MUC17

Materials and methods

Sources of protein and nucleotide sequences

Databases and tools used include NCBI databases (http://www.ncbi.nlm.nih.gov/), Celera Discovery System (http://www.celera.com/), Ensembl databases (http://www.ensembl.org/), UCSC genome browser (http://genome.ucsc.edu/), HapMap project (http://www.hapmap.org/), and Hugo project (http://www.gene.ucl.ac.uk/nomenclature/).

Table 3. *Data downloaded from Ensembl database.*

Species	Version	Protein-regular		Protein-abinitio		Genomic contigs	
		Size	Entries	Size	Entries	Size	Entries
Human	16.33	19.2Mb	32035	23.9Mb	65771	3.21Gb	32035
Human	18.34d	17.5Mb	27941	24.1Mb	65010	3.26Gb	26614
Human	32.35e	21.4Mb	33869	24.1Mb	57898	3.27Gb	26881
Mouse	21.32b	16.5Mb	32281	24.0Mb	76790	3.23Gb	128285
Elephant	42.1b	10.4Mb	15717	31.2Mb	107250	Not downloaded	
Opossum	42.3b	23.3Mb	32690	39.8Mb	109135	Not downloaded	
Chicken	35.1k	18.2Mb	28416	24.6Mb	77600	1.08Gb	111864
X. tropicalis	34.1b	34.2Mb	52786	32.2Mb	70441	1.66Gb	27064
X. tropicalis	37.4	19.0Mb	29843	57.2Mb	123446	1.54Gb	19759
Fugu	11.2	11.2Mb	33609	15.0Mb	32615	338Mb	12403
Fugu	31.2f	22.7Mb	33003	14.5Mb	34080	336Mb	20379
Fugu	39.4b	14.1Mb	22102	18.3Mb	29477	400Mb	7213
Tetraodon	37.1c	15.2Mb	28005	47.5Mb	100735	350Mb	25773
Zebrafish	31.4d	20.5Mb	32062	26.7Mb	56799	1.60Gb	27677
Zebrafish	33.5a	20.5Mb	32143	51.1Mb	87821	1.72Gb	21440
Zebrafish	39.6	15.6Mb	26549	51.4Mb	86391	1.79Gb	16644
Ciona	35.195b	13.4Mb	21574	8.2Mb	11724	119Mb	2242
Ciona	40.2b	11.1Mb	20000	40.8Mb	72979	178Mb	19826
Drosophila	40.43	13.2Mb	19789	18.8Mb	32035	135Mb	2658
C. elegans	40.150a	14.2Mb	26032	20.4Mb	41734	105Mb	3268

We have downloaded mucin sequences from the information available at NCBI for human, chimp, rhesus, mouse, rat, cat, dog, cow, opossum, chicken, *Caenorhabditis elegans* and *Drosophila melanogaster*. Some of the entries are fragments of large mucins, and some of the redundancy was removed by merging sequences. We also make use of proteins listed in the Ensembl database. There are two types of protein sets in these databases. One is the regular set of proteins and the other set are proteins as predicted by abinitio methods. The data we downloaded from Ensembl is listed in table 3. For *Strongylocentrotus purpuratus*, we use the Sea Urchin Genome Project database; for *Apis mellifera*, the Honey Bee genome project database; for *Branchiostoma floridae*, the Sanger database and for *Nematostella vectensis*, we used the database produced by the US Department of Energy Joint Genome Institute.

Bioinformatic tools

PTSpred

The mucin domains pose a specific problem as they are very poorly conserved in sequence. Consequently they cannot be identified using standard methods based on sequence similarities such as BLAST searches. For this project the program PTSpred was developed, based on the unusual amino acid compositional bias in mucin domains. The algorithm is based on: 1. High frequency of amino acids Ser, Thr and/or Pro (Ser+Thr usually >50%, Pro usually > 5%). 2. The domains should be long (usually > 100 amino acids).

Local regions with the compositional bias typical for mucin proteins are identified by moving a window along the sequence and determining the composition in that window. Commonly we set the window length to 100 amino acids and the window is moved in steps of 10 amino acids. For nucleotide sequences, protein products of all six potential reading frames were considered. Figure 4 shows an example of the result of PTSpred used to analyze Ensembl human proteins.

Mpred

Mpred program is built on a generalized hidden Markov model that is currently composed of two states, mucin or nonmucin, but can be extended to include additional features. The algorithm runs through the

The total number of proteins in database is 32035

No:1
Translation:ENSP00000307150 Database:core Gene:ENSG00000169889 Clone:AC105446
Contig:AC105446.4.1.139517 Chr:7
Basepair:100189651 Status:known
Length: 951 Start position: 0 End position: 951
Sequence:
HTTAFPGSTTMPGVSQESTASHSSPGSTDTTLSPGSTTASSLGPESTTFHSGPGSTETTLLP
DNTTASGLLEASTPVHSSTGSPHTTLSPAGSTTRQGESTTFQSWPNSKDTTPAPPTTTSAF
VELSTTSHGSPSSTPTTHFSASSTTLGRSEESTTVHSSPVATATTPSPARSTTSGLVEESTT
YHSSPGSTQTMHFPESDTTSGRGEESTTSHSSTTHTISSAPSTTSALVEEPTSYHSSPGSTA
TTHFPDSSTTSGRSEESTASHSSQDATGTIVLPARSTTSVLLGESTTSPISSGSMETTALPG
STTTPGLSEKSTTFHSSPRSPATTLSPASTTSSGVSEESTTSHSRPGSTHTTAFPDSTTTPGL
SRHSTTSHSSPGSTDTTLLPASTTTSGSSQESTTSHSSSGSTDTALSPGSTTALSFGQESTT
FHSSPGSTHTTLFPDSTTSSGIVEASTRVHSSTGSPRTTLSPASSTSPGLQGESTAFQTHPA
STHTTPSPPSTATAPVEESTTYHRSPGSTPTTHFPASSTTSGHSEKSTIFHSSPDASGTTPSS
AHSTTSGRGESTTSRISPGSTEITTLPGSTTTPGLSEASTTFYSSPRSPTTTLSPASMTSLGV
GEESTTSRSQPGSTHSTVSPASTTTPGLSEESTTVYSSSPGSTETTVFPRSTTTSVRREEPT
TFHSRPASTHTTLFTEDSTTSGLTEESTAFPGSPASTQTGLPATLTTADLGEESTTFPSSSG
STGTKLSPARSTTSGLVGESTPSRLSPSSTETTTLPGSPTTPSLSEKSTTFYTSPRSPDATLS
PATTTSSGVSEESSTSHSQPGSTHTTAFPDSTTTSGLSQEPTTSHSSQGSTEATLSPGSTTA
SSLGQQSTTFHSSPGDTETTLLPDDTITSGLVEASTPTHSSTGSLHTTLTPASSTSTGLQEE
STTFQSWPSSSDTTPSPP
Pro rate: 9.56887
Ser rate: 24.5005
Thr rate: 24.1851

Figure 4. *Top-ranking hit when PTSpred was applied to analyze Ensembl human proteins.* This particular hit is a part of the Muc12 PTS domain.

protein sequence and determines which amino acids belong to which state, resulting in a set of start and end coordinates for potential mucin domains along with a probability indicating a reliability of the prediction. The probability distributions incorporated in the model, such as state transitions, domain lengths, sequence composition, and so on, are based on empirical data. Currently the performance of the method is limited by the relatively small number of available training sequences. With more training sequences, the specificity of the predictions would be improved [Paper I].

Other bioinformatics software

There are two types of mucin proteins. One is membrane-bound and has one transmembrane domain. The other is secreted and has no transmembrane part. TMHMM (TransMembrane Hidden Markov Model) is a program to predict the trans-membrane helices in proteins using a Hidden Markov Model [60, 61]. Thus it can be used to distinguish between transmembrane and secreted type mucins.

Each mucin, either membrane-bond type or secretory type, has a signal sequence in their N-terminus. The position of the cleavage site is < 30 amino acids from the N-terminal. In order to predict signal sequence we have used SignalP [62]. The present version of SignalP (version 3.0) comprises two signal peptide prediction methods [63], SignalP-NN (based on neural networks) and SignalP-HMM (based on hidden Markov models). For eukaryotic data, SignalP-HMM has a substantially improved discrimination between signal peptides and uncleaved signal anchors, but it has a slightly lower accuracy in predicting the precise location of the cleavage site [64]. The HMM method has been used here.

Sequence alignments provide a powerful way to compare novel sequences with previously characterized genes. For studies of mucin relationships we have used BLAST [65, 66] and SPIDEY [67]. For multiple alignments Clustalw was used [68, 69].

Genewise is designed to align a protein sequence to a genomic sequence, taking splicing into account [70]. It can also use a profile HMM such as a Pfam model for alignment to genomic sequence. The Genewise approach builds on the success of hidden Markov models (HMMs) for modeling both protein family information and gene predictions. Genewise can be thought of as considering every possible gene prediction in a genomic sequence and comparing each one to the protein profile-HMM. The best combined score of both the gene prediction and the protein profile-HMM is used to provide a simultaneous gene prediction and protein alignment. To use Genewise for gene prediction one needs a source of homology information. Examples are profile-HMMs from Pfam.

For phylogenetic analysis, VWD or SEA domain sequences were extracted and aligned using ClustalW using default parameters. The bootstrapping option of ClustalW was used to generate bootstrapped trees. As an alternative, PHYLIP programs NEIGHBOR and PROTPARS were used to generate neighbor-joining and parsimony trees, respectively [71, 72]. 100 replicates were analyzed for bootstrapping analysis and

CONSENSE was used to generate consensus trees. We also used the programs T-coffee [73-75], PAUP [76], NJPLOT [77], TREEVIEW [78] and PHYLODRAW [79-82]. TREEDYN was used for drawing and inspecting the trees [83].

We also used GENSCAN for exon prediction [84], DOTTUP, DOTPLOT and RADAR for examining the repeats [85, 86], PSIPRED for secondary structure prediction [87] and PERL (Practical Extraction and Report Language) for programming. The PERL are used to deal with the result of almost all the bioinformatic softwares we used, to extract and analyze thousands of sequences, and to build our online databases.

Genome sequencing projects

In this project we made use of the results of several genome sequencing projects. Information for some of these projects are listed below and are from web pages of the respective projects. Figure 5 lists the taxonomic relationships of these species.

Homo sapiens (human)**.** The Human Genome Project was one of the great feats of exploration in history [84, 88]. It was a 13-year project coordinated by the U.S. Department of Energy and the National Institutes of Health. The project aimed at identifying all the approximately 20,000-25,000 genes in human DNA, determining the sequences of the 3 billion base pairs that make up human DNA, storing this information in databases, improving tools for data analysis, transfering related technologies to the private sector, and addressing the ethical, legal, and social issues that may arise from the project (http://www.ornl.gov/sci/techresources/Human_Genome/project/).

Mus musculus (mouse) and *Rattus norvegicus* (rat)**.** As the most powerful model organism in biomedical research, the mouse was the second mammal to be sequenced as part of the Human Genome Project. Initial analysis of the differences and similarities between the mouse and human genomes suggested that as much as 5% of the human genome might be highly conserved between mammals based on function (http://www.broad.mit.edu/mouse/).

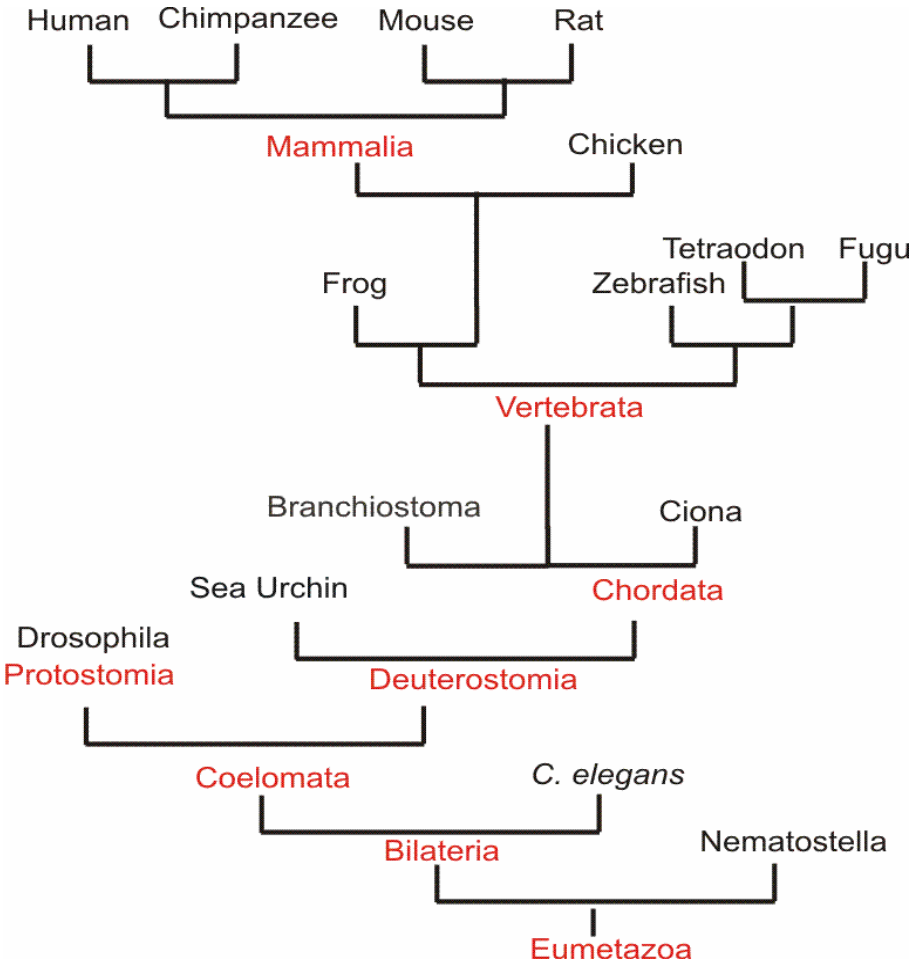

Figure 5. *Schematic phylogenetic tree showing organisms whose genomes have been sequenced.*

Gallus gallus (chicken). The chicken is an advanced model system for studying embryonic development and gene function in higher vertebrates and can offer unique experimental advantage and used for experimental embryology.

Figure 6. *Xenopus tropicalis.*

Xenopus tropicalis (frog) is a unique resource for two critical areas in vertebrate biology: early embryonic development and cell biology. It has become a major vertebrate model for the cellular and developmental biology research. It offers one of the smallest genomes among the amphibians (which number more than 4,500 species) and it is the only *Xenopus tropicalis* species that is diploid, greatly simplifying genetic studies. Also, the integument of

amphibians is covered by a mucus layer which shields against dehydration, physical damage and microbial infections. A *Xenopus laevis* integumentary mucin has been discovered which has typical threonine-rich highly O-glycosylated repetitive domains, trefoil factor like sequences and VWD domains [89].

Figure 7. The pufferfish Fugu.

Takifugu rubripes (Fugu), Japanese pufferfish, has the smallest known genome among vertebrates. The Fugu genome project was initiated in 1989 by Sydney Brenner and his colleagues Greg Elgar, Sam Aparicio and Byrappa Venkatesh [90]. The genome is only 390 Mb which is about eight times smaller than the 3,000 Mb human genome, yet it contains a similar repertoire of genes [91]. The compact nature of the genome is largely due to lack of dispersed repetitive sequences (<10%). Fugu was proposed as a model vertebrate genome to understand the more complex human genome and other vertebrate genomes [92]. Subsequent studies have shown that the intergenic regions and introns in the Fugu are highly compressed and the average gene density is about one gene per 10 kb. Also the gene order over short range is conserved between the Fugu and human. The Fugu and human lineages diverged about 450 million years ago.

Figure 8. *Zebrafish*.

Danio rerio (zebrafish) is a small freshwater fish native to south Asia and is a common aquarium fish worldwide. It is an important model organism for the study of vertebrate development and disease, organ function, behavior, and toxicology. Some of the features that make the zebrafish so experimentally amenable include its short generation time, large numbers of embryos produced per mating, and the development of transparent embryos outside the mother, allowing all stages of development to be observed. Adults are about 4 cm long, the female can lay 200 eggs per week, and the embryos reach maturity in 2-3 months. Precise tools have been developed to generate and analyze alterations in the zebrafish genome. These features mean that a comparison of human and zebrafish projects will accelerate progress (http://www.ncbi.nlm.nih.gov/genome/guide/zebrafish/).

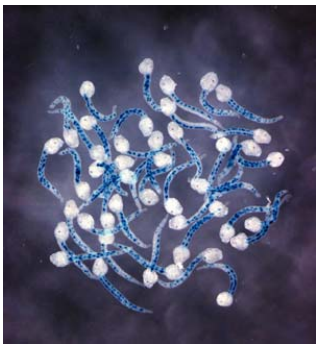

Figure 9. *Ciona intestinalis.*

The *Ciona intestinalis* genome is the smallest of any experimentally manipulable chordate. It provides a good system for exploring the evolutionary origins of the chordate lineage, from which all vertebrates sprouted. The organism has a good genomic infrastructure, easily visualized

cells and morphogenetic processes, existing methods for transient transgene expressions, and is available throughout the world all year long. The complete *Ciona intestinalis* genome sequence will provide a foundation for genome-scale analysis of regulatory networks through development. In addition, there is a deep classical literature on ascidian development, and an active community of researchers worldwide (http://genome.jgi-psf.org/ciona4/ciona4.home.html).

Figure 10. *Fruit fly.*

The common fruit fly, *Drosophila melanogaster,* has been the workhorse of biology and genetics laboratories for the past 90 years. It has been one of the most influential model systems for geneticists (http://www.hhmi.org/news/rubin3.html).

Caenorhabditis elegans is a small, free-living, round worm found in nutrient- and microorganism-rich habitats such as in compost, mushroom beds and garden soil where it feeds on bacteria and probably other microorganisms. It is easy to maintain in the laboratory, grow quickly on a bacterial lawn. The genome is small compared to humans (about 30 times smaller), yet it encodes over 22,000 proteins, only slightly fewer than humans. About 35% of *Caenorhabditis elegans* genes are closely related to human genes (http://genome.wustl.edu/).

Figure 11. *Caenorhabditis elegans.*

The sea urchin (*Strongylocentrotus purpuratus*) has been an important model system for many years in the study of basic biology, particularly in developmental biology. The sea urchin occupies an important evolutionary position with respect to vertebrates and humans. There is a large body of information about gene expression in the sea urchin and there are a number of genomic resources available, making the sea urchin an ideal organism for learning how pathways of genes and proteins regulate growth and development, with potentially profound implications for understanding human biology (http://www.hgsc.bcm.tmc.edu/projects/seaurchin/).

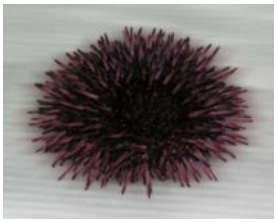

Figure 12. *Sea urchin.*

The Florida lancelet *Branchiostoma floridae* is a representative of the invertebrate subphylum Cephalochordata. Lancelets, along with tunicates, are members of the chordate phylum, and share the same

basic body plan as vertebrates. This includes a stiffening structural rod, or notochord, a dorsal hollow nerve tube, segmented muscle blocks and pharyngeal gill slits. However, they have a simpler anatomy and physiology than vertebrates, with fewer distinct tissue types and organs. Anatomical, embryological and molecular genetic comparisons between the lancelet and the vertebrates continue to shed light on the evolutionary origin of the vertebrates. It is a model organism for understanding evolution of the chordate body plan (http://genome.jgi-psf.org/Brafl1/Brafl1.home.html).

Nematostella vectensis (scarlet sea anemone) belongs to the class Anthozoa. Molecular and morphological studies suggest that Anthozoa is the basal group among the Cnidaria. *Nematostella vectensis* is easy to culture in the lab and gametogenesis can be induced. Thus, embryonic development is easily accessible (http://www.sars.no/research/TechnauGrp.php).

Figure 13. *Branchiostoma floridae*.

Figure 14. *Nematostella vectensis*

Results

Mucins are proteins with highly O-glycosylated PTS domains (also called mucin domains) that are rich in the amino acids serine, threonine and proline. Our understanding of mucins has been obstructed by experimental difficulties that are due to their large size, oligomerization, and extensive glycosylation. To facilitate the studies of mucins, we deduced that a better knowledge and realization of the evolution and distribution of them should be of great assistance. However, the PTS domains in mucins are poorly conserved and cannot be used for studying evolutionary relationships between mucin proteins of different categories and from different species. Instead we have made use of the VWD, SEA and AMOP domains characteristic of mucins. With the detailed analysis of a range of metazoa including mammals, chicken (*Gallus gallus)*, frog (*Xenopus tropicalis*), fishes (Fugu, zebrafish and *Tetraodon nigroviridis*), the sea squirt *Ciona intestinalis*, the lancelet *Branchiostoma floridae*, the sea urchin *Strongylocentrotus purpuratus*, the fruit fly *Drosophila melanogaster*, the worm *Caenorhabditis elegans*, and the starlet sea anemone *Nematostella vectensis*, we were able to examine evolutionary relationships among mucins and mucin related proteins and to analyze relationships of orthology and paralogy.

When no particular reference is given, paper III is referred to. The papers I and II describe our analysis of the mucins of the puffer fish Fugu and chicken, respectively.

Mucins and related proteins with multiple VWD domains

In human, four gel-forming mucin genes are localized in one locus on chromosome 11; MUC2, MUC5AC, MUC5B and MUC6. Computational and phylogenetic analyses made it possible to propose an evolutionary history of these human mucin genes from an ancestor gene common to the human von Willebrand factor gene [20, 26]. In order to better understand the evolution of the VWD domain containing mucins as well as the VWD domain containing proteins in general we have studied these proteins systematically and identified interesting new features. Many VWD domain containing proteins have been considered such as the von Willebrand factor, otogelin, sco-spondin, zonadhesin, tectorin, IgG-binding protein, hemolectin and vitellogenin [93, 94]. The species listed above were taken into account.

For all the proteins with multiple VWD domains, our analysis allowed us to place them into two categories. One contains the gel-forming mucins, otogelin, von Willebrand factor, and insect hemolectin. The other category contains the remaining VWD-containing proteins; IgG-binding protein, tectorin, sco-spondin and zonadhesin.

Characteristic of the gel-forming mucins and mucin related proteins is that these proteins have at least 3 VWD domains. In the N-terminal of the protein, three VWD domains (called D1, D2 and D3 domains) are found followed by a domain characteristic of the respective protein (PTS for mucin, PTS and AbfB for otogelin, three VWA domains for von Willebrand factor and F5/8 type C for insect hemolectin). Table 4 lists the Pfam annotation of these characteristic domains. At the C-terminal side of this domain some of the proteins have another VWD domain (called D4 domain), thus achieving a "3+1" structure. In the absence of such a VWD the protein is referred to as having a "3+0" structure. Figure 15 shows the domain structure of selected proteins with multiple VWD domains.

Table 4. *The characteristic Pfam domains appearing in mucin related proteins.*

Pfam domain	**Annotation**	**Protein with domain**
VWA	von Willebrand factor type A domain.	von Willebrand factor
AbfB	Alpha-L-arabinofuranosidase B domain. This family consists of several fungal alpha-L-arabinofuranosidase B proteins. L-Arabinose is a constituent of plant cell wall poly saccharides. It is found in a polymeric form in L-arabinan, in which the backbone is formed by 1,5-a-linked l-arabinose residues that can be branched via 1,2-a- and 1,3-a-linked l-arabinofuranose side chains. AbfB hydrolyses 1,5-a, 1,3-a and 1,2-a linkages in both oligosaccharides and polysaccharides, which contain terminal non-reducing l-arabinofuranoses in side chains	Otogelin
F5/8 type C	The F5/8 type C domain is also known as the discoidin (DS) domain family. It is a C-terminal and twice repeated domain of about 150 amino acids which is found in blood coagulation factors V and VIII.	Insect hemolectin

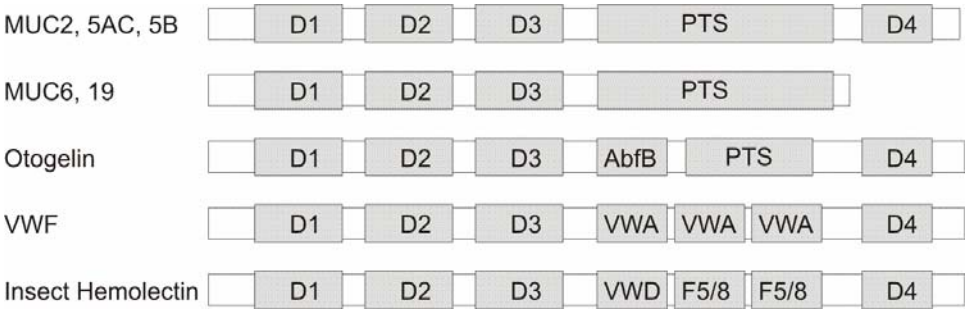

Figure 15. *Domain structures of gel-forming type mucins and related proteins.*

The phylogenetic analysis of VWD domains in gel-forming type mucins and mucin related proteins shows that these VWD domains cluster such that all D1 domains are in one branch, all D2 in one branch, etc. This would suggest that these domains and their order have been conserved during evolution.

MUC2

MUC2 mucins were identified in human, mouse, chicken, *Xenopus tropicalis*, Fugu, and zebrafish. In all these species, MUC2 proteins have a D1-D2-D3-PTS region-D4-CK organization. The PTS region contains PTS domains and CysD domains. In human and mouse, there are two CysD domains. In chicken, at least four CysD domains were detected, whereas the Fugu Muc2 had no CysD domains.

In *Xenopus tropicalis*, the repertoire of Muc2 proteins was markedly expanded and altogether there are 16 Muc2 proteins (Figure 16). Ten of the Muc2 proteins in *Xenopus tropicalis* have PTS domains that are unusually rich in cysteines and we have referred to these domains as ‘CPTS’. These domains are organized in tandem repeats as has been typically found for other mucin PTS domains. All PTS domains in human and mouse mucins are encoded by single exons, but each tandem repeat unit in the CPTS domain is encoded by one or two exons. This is similar to the PTS domains in chicken Muc13 and zebrafish Muc13 [Paper II]. Two of the Muc2 mucins have CysD domains and these have standard PTS domains. It can thus be noted the CPTS domains never appear together with CysD domains.

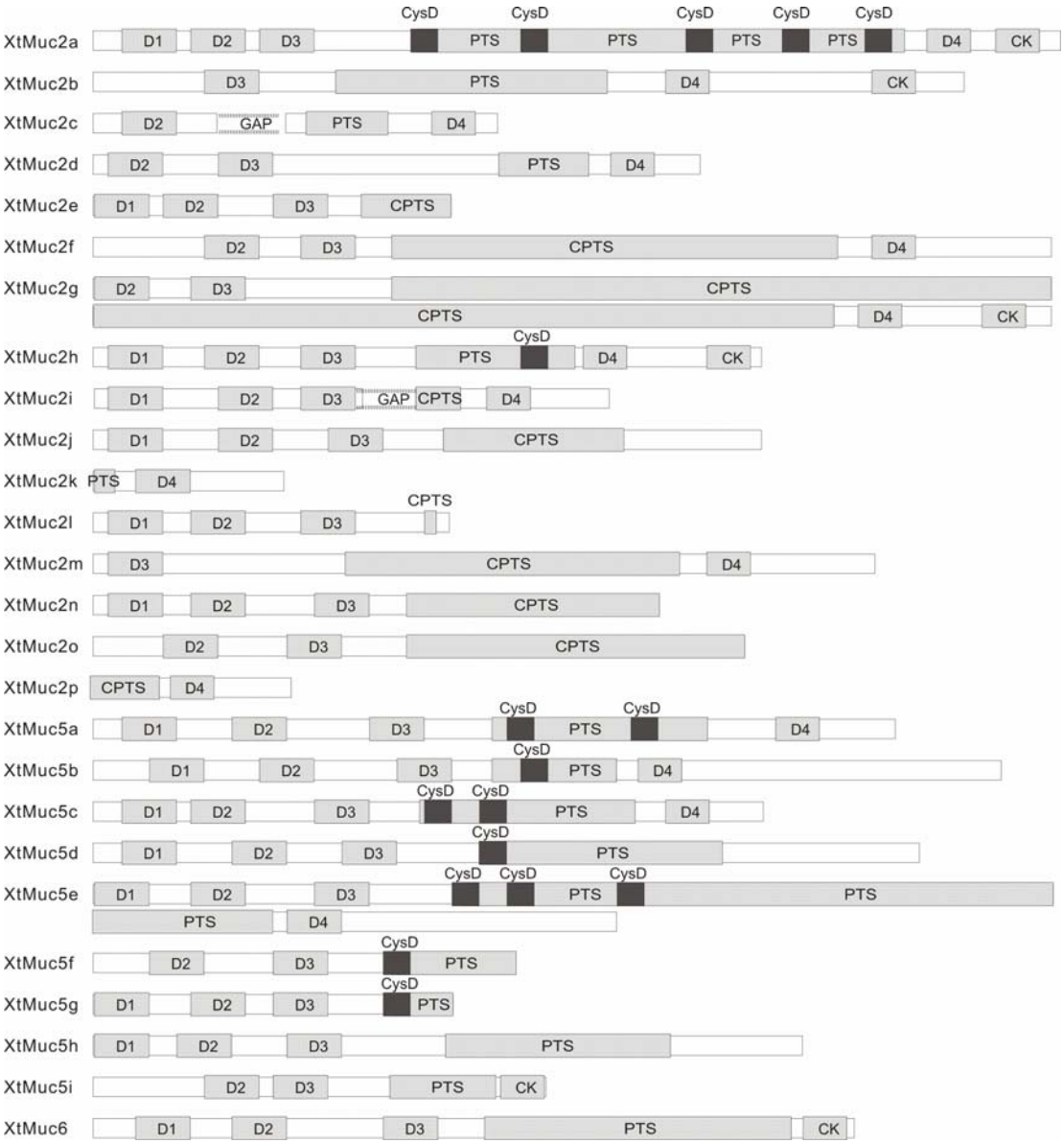

Figure 16. *Gel-forming mucin proteins in Xenopus tropicalis.*

MUC5

The MUC5 mucins are paralogues to the MUC2 mucins. In human and mouse, the MUC5 proteins have a D1-D2-D3-PTS region-D4-CK domain structure. However, in the Muc5b of chicken and some of the Muc5 in *Xenopus tropicalis*, Fugu, and zebrafish we did not find a D4 domain. This is interesting, but might also be the result of errors in genome assembly or gene prediction. CysD domains were identified in human and mouse MUC5AC, MUC5B, in chicken Muc5ac, in seven *Xenopus tropicalis* Muc5 mucins (Figure 16), and

in two zebrafish Muc5. We identified nine and six Muc5 proteins in *Xenopus tropicalis* and zebrafish, respectively.

In the case of human MUC5AC, there are still gaps in the genomic sequence corresponding to the PTS and CysD regions. In our prediction of MUC5AC, we have combined four NCBI mRNA entries (AJ298317, AJ298318, AJ298319 and AJ001402) and were thus able to predict the major part of the PTS region with six large PTS domain blocks and 10 CysD domains of highly repetitive sequences. Based on the size of the gap in genomic region we expect at least two more CysD domains in this region.

We wanted to trace the evolution of the Muc5 family of proteins. Based on the analysis of VWD domains in these proteins, Muc5ac and Muc5b can be distinguished in human, mouse and chicken. However, in *Xenopus tropicalis*, Fugu, and zebrafish, Muc5ac and Muc5b cannot be distinguished although Muc5 can be distinguished from Muc2.

MUC6 and MUC19

MUC6 and MUC19 proteins are paralogues to MUC2 and MUC5. They have a D1-D2-D3-PTSregion-CK structure in common, and there are no CysD domains in the PTS region. Phylogenetic analysis of the VWD domains also indicated a close relationship of Muc6 and Muc19. Human MUC19 was reported by Chen Y. et al [36] but only partial sequences were presented. We have been able to present a more complete human sequence by analyzing the genomic sequence of MUC19 and with PTSpred we identified a large PTS block with at least 69 tandem repeats (http://www.medkem.gu.se/mucinbiology/databases/). The mouse Muc19 was presented by Culp *et al* with cDNA [37].

Ovomucin

Ovomucin is a chicken protein identified by Watanabe *et al* [95], made up of two subunits named α and β. We identified the α subunit (often referred to as 'ovomucin') as being encoded by a gene within a gel-forming mucin cluster (see below) [Paper II]. The chicken ovomucin has a D1-D2-D3-X-D4 structure where the PTS region found in mucins is replaced by non-PTS sequence. We also found that the β-subunit of the ovomucin [95] is an orthologue to the human MUC6. The phylogenetic tree analysis of the VWD

domains of chicken ovomucin shows that they are most closely related to the VWD domains of MUC5 mucins.

MUC6-MUC2-MUC5AC-MUC5B gene cluster

In human and mouse the *MUC2, MUC5AC, MUC5B* and *MUC6* genes are found in the same locus. MUC6 has a direction opposite to that of the other mucin genes. In chicken there is an equivalent locus where the ovomucin gene is located between Muc2 and Muc5ac and with the same orientation as these two genes. [Paper II]. In *Xenopus tropicalis*, Muc6, five Muc2 proteins and five Muc5 proteins are clustered together in the same scaffold with the gene order Muc6(-), Muc2a(+), Muc2b(-), Muc5i(+), Muc5a(+), Muc5b(+), Muc2c(+), Muc2d(+), Muc5c(-), Muc5d(-) and Muc2e(+).

The chromosomal localisation of *Xenopus tropicalis* Muc5i gene is similar to the chicken ovomucin gene. Also the VWD domains of Muc5i, with structure D1-D2-D3-PTS-CK, are similar to the VWD domains of chicken ovomucin. Therefore, the Muc5i protein may be a *Xenopus tropicalis* orthologue of ovomucin.

MucMV (Multiple VWDs)

In *Ciona intestinalis, Branchiostoma floridae* and *Strongylocentrotus purpuratus*, Muc2, Muc5 and Muc6 cannot be distinguished on the basis of our phylogenetic analysis. We have referred to these mucins as 'MucMV' (mucin with multiple VWD domains). We found six of these proteins in *Ciona intestinalis*, four in *Branchiostoma floridae* and one in *Strongylocentrotus purpuratus*. These mucins have a 3+0 or 3+1 VWD structure as well as a PTS domain. In higher species they evolved into Muc2/Muc5/Muc6.

Otogelin and VWF

Otogelin is a protein present in the acellular membranes covering the tectorial membrane, the otoconial and accessory membranes of the utricule and saccule, and the cupula of the semicircular canals in the ear [96]. VWF is an important component in the coagulation system and functions as adhesion sites for the platelets and the connective issue [93]. Both these proteins have a 3+1 VWD domain structure and are evolutionary related to gel-forming mucins as evident from phylogenetic analysis of their VWD domains. The vertebrate

gel-forming mucins, otogelin, and VWF are therefore distinct from the other proteins with multiple VWD domains, i.e. Tectorin, Zonadhesin, Sco-spondin and IgG-binding proteins. In our analysis otogelin and VWF first appear in fish, and can be found throughout the vertebrate animals.

Insect hemolectin and gel-forming mucin ancestor

The *Drosophila melanogaster* protein hemolectin has a function in coagulation and immunity [97, 98]. This protein has the D1-D2-D3-X-D4 motif in common with mucins and between the D3 and D4 domains there are two F5/8 type C domains and a fifth VWD domain. We also identified proteins related to hemolectin in *Branchiostoma floridae* and *Nematostella vectensis*. The protein in *Branchiostoma floridae* has the same domain structure and PTS domains inserted into the F5/8 type C region. The protein in *Nematostella vectensis* has a D1-D2-D3 structure followed by two F5/8 type C and two PTS domains. The phylogenetic analysis of the D1, D2, D3 and D4 domains of these proteins shows that they belong to the mucin/otogelin/VWF group. The presence of the mucin type protein in *Nematostella vectensis* suggests that the gel-forming mucin family can be traced to lower metazoa. Our interpretation of the evolution of gel-forming mucins and related proteins is summarized in Figure 17.

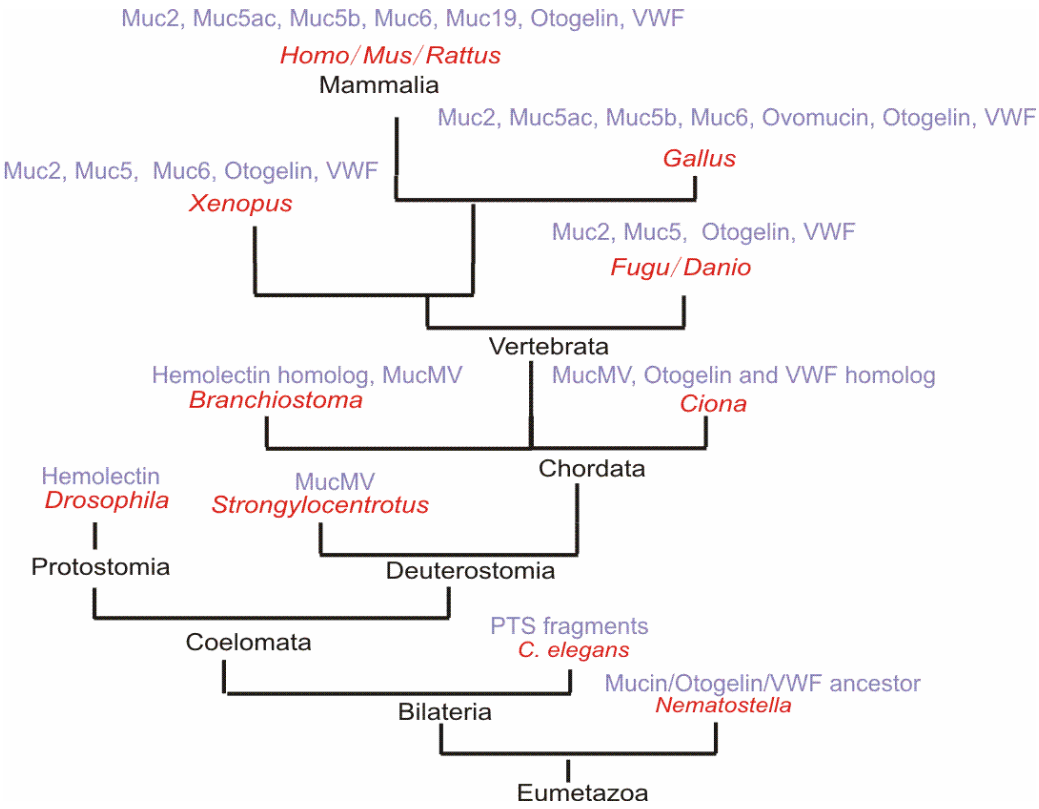

Figure 17. *Evolution of gel-forming mucins and related proteins*

MUC4 and related proteins

The MUC4 mucin is characterized by NIDO-AMOP-VWD domains as well as a TM domain. It forms a group distinct from the other VWD mucins or TM mucins. The only other protein that has a NIDO-AMOP-VWD-TM structure is the sushi domain-containing protein Susd2 that apparently has homologues in *Caenorhabditis elegans* and *Drosophila melanogaster*. The VWD domains of Muc4 and Susd2 are clearly related also in the respect that they do not contain any cysteines. However, the Susd2 protein is different from Muc4 as it lacks PTS and EGF. Because of the relationship between Susd2 and Muc4 and because Susd2 was identified in more deeply branching animals we propose that MUC4 developed from the Susd2 family of proteins by acquiring a PTS domain. An overview of the evolution of Muc4 and related proteins is shown in Figure 18.

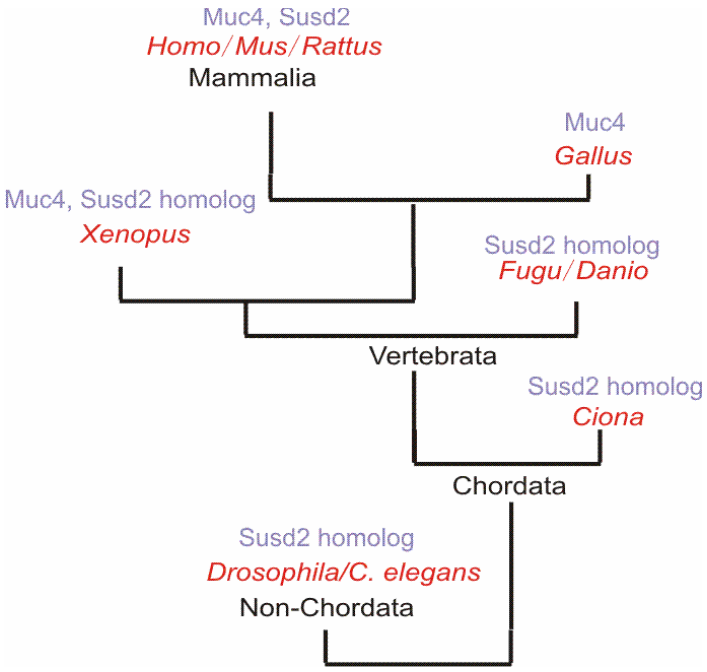

Figure 18. *Evolution of Muc4 and related proteins.*

The MUC4 sequence provided by Choudhurry *et al* which was based a pancreatic carcinoma cDNA and the total protein length was 2169 amino acids [99]. Based on careful analysis of the genomic sequence, we have presented a more complete sequence of the protein (http://www.medkem.gu.se/mucinbiology/databases/). The PTS domain is much longer and is highly repetitive with 197 copies of a unit of 16 amino acids, and the total length of the protein is 4158 amino acids.

Mucins and related proteins with SEA domains

SEA domains have been identified in mucins, agrin, enterokinase (PRSS7), 63 kDa sea urchin sperm protein, HSPG2 (perlecan, heparan sulphate proteoglycan 2) core, the cell surface antigen 114/A10, HAT, a trypsin-like serine protease, in IMPG1 and 2 (interphotoreceptor matrix proteoglycans 1 and 2) and in a

number of uncharacterized proteins from *Drosophila melanogaster* and *Caenorhabditis elegans*. However, only the mucins contain PTS domains. These mucins also typically contain a transmembrane domain (TM), but lack other characteristic domains found in the proteins listed above. The SEA domain is apparently an early metazoan invention as it occurs also in the deeply branching *Nematostella vectensis*.

Five different clusters of mucin-type SEA domains are identified from the phylogenetic analysis and correspond to MUC1, MUC13, MUC16, MUC3 family (MUC3, 12 and 17), and a novel type of mucin. MUC1 appeared at a late stage in the development of mammals. MUC13 were found in chicken [Paper II] and zebrafish, but not in more deeply branching organisms like *Ciona intestinalis*. It would therefore seem that MUC13 is restricted to vertebrates. MUC16 is characterized by multiple SEA domains that are preceded by PTS domains. It was found in human, mouse, chicken and *Xenopus tropicalis* but not in fish. The Muc3 group was found in human, chimpanzee, mouse and *Xenopus tropicalis*. In human, chimpanzee and mouse, this MUC3 group contains MUC3, MUC12 and MUC17 whose genes are tandemly arranged in the same locus. In *Xenopus tropicalis*, seven MUC3 homologues were found and three of these are in the same locus. In addition to previously known SEA-domain-containing mucins, a novel type of mucin was discovered that forms a rather large family in *Xenopus tropicalis* and fish. In this family, the PTS and SEA domains are alternating in a manner not seen in the mucins of mammals.

An overview of the evolution of SEA domain related mucins is shown in Figure 19.

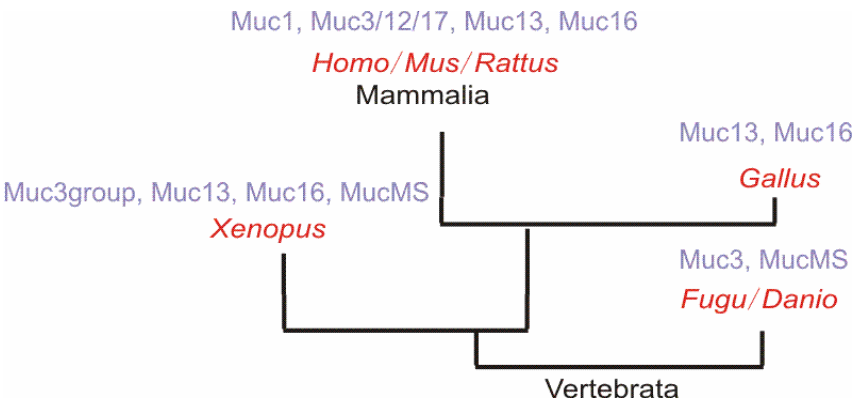

Figure 19. *Evolution of SEA domain related mucins.*

MUC3 group

In the human chromosome 7q22 region, the MUC3, MUC12 and MUC17 genes are located. The region was further analyzed and compared to the corresponding regions in the mouse (chromosome 5) and chimpanzee (chromosome 6).

In support of the orthologous relationship we found that the mucins in human, chimpanzee as well as in mouse are flanked by the *ACHE* (5´) and *TRIM56* (3´) genes. The three SEA domains that we identified in the human genome were identical to the previously described MUC3, MUC12 and MUC17 mucins. In the public protein sequence databases we identified two protein sequences that correspond to the q22 region and they are assigned as MUC11 and MUC12. Our analysis of the human genomic sequence reveals that the sequence between the known MUC11 and MUC12 sequences is very short (5400 bp) and one long open reading frame spans this entire region. It would therefore seem that the proteins described as MUC11 and MUC12 are parts of the same protein where the region corresponding to MUC11 and the N-terminal part of MUC12 are parts of the same exon.

In human, the sequences C-terminal to the PTS domain of MUC3, MUC12 and MUC17 are quite similar (Figure 20). Furthermore, downstream of the PTS domain all three mucins are encoded by 10 exons and the exon-intron borders have almost the same positions. In all the three mucins, the first EGF domain is encoded by exon 1 after the PTS exon, the SEA domain by the second, third, and fourth exons and the second EGF domain by exons 6 and 7. The transmembrane domains are all in the seventh exon and the cytoplasmic tail in exons 7 to 10. A comparison of the C-terminal regions of MUC3, MUC12 and MUC17 is shown in Figure 20.

For the human MUC3 protein there are four non-overlapping PTS-containing protein sequences in the sequence databases. In these sequences there is one repeat sequence covering 377 amino acids and downstream of that there are 24 repeats, each 13 amino acids in length [45]. The PTS domain of MUC12 is made up of repetitive elements of different lengths, mostly around 55 to 56 amino acids. In the PTS domain of MUC17, 60 tandem repeats are found of which 57 repeats have 59 amino acids each and 3 repeats have 58 amino acids each [51, 52].

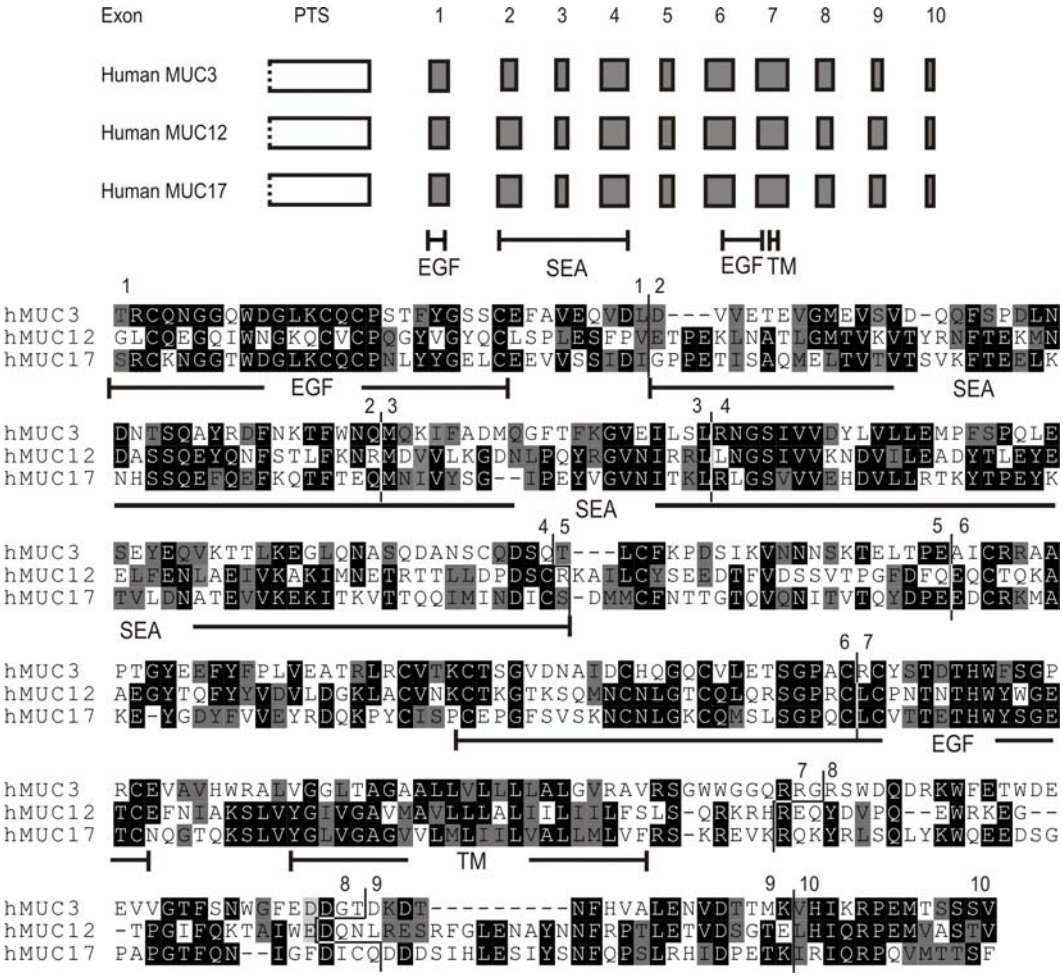

Figure 20. *Comparison of C-terminal regions of human MUC3, MUC12 and MUC17.* In the upper part, each box represents one exon. The length of the each box is proportional to the length of each exon. In the lower part, a multiple sequence alignment is shown where the exons borders are indicated.

The analysis of the human MUC3 mucin causes special problems as the protein seems to be extremely variable in sequence. It has been suggested [46, 47] that MUC3 in fact is made up of two genes named MUC3A and MUC3B that are about 95% identical. However, if we consider the NCBI and Ensembl genomic dataset, only one transmembrane domain, one SEA domain, and one signal sequence is found between the ACHE and MUC12 genes. This suggests that the variable sequences could be due to alleles of one MUC3 gene. However, PCR studies of individuals have revealed that within 8 different genomes, both the MUC3A and MUC3B sequences were found [47]. This makes it less likely that the MUC3A and 3B sequences are allelic and suggests that the human genome is incorrectly assembled here. If this is the case,

the genes must be very close together as there is apparently not much room for another gene between ACHE and MUC12.

We also examined the homologues of the human 7q22 mucins in chimpanzee. Three neighbouring transmembrane type mucins were found in chromosome 6 with the same domain structure as human MUC3, MUC12 and MUC17, and with the same gene order as in human and chimpanzee.

The corresponding genomic region for mouse is on chromosome 5. The mouse genome contains only one SEA domain in this region and has 43% identity to MUC17, but only 28% and 32% identity with the human MUC3 and MUC12 SEA domains, respectively. Furthermore, the SEA domain is located close to the TRIM56 gene. The mouse mucin gene in this region has been annotated as Muc3, but according to our analysis it is likely to be the mouse Muc17 homologue. In the mouse genome assembly there is a large gap in this region upstream of Muc17, and the length of this region is comparable to that of man and chimpanzee which encodes MUC3 and MUC12. In an analysis of predicted mouse proteins we identified sequences corresponding to Muc3 and Muc12 C-terminals that had not been possible to place in available genomic assemblies. Therefore, we conclude that the gap region in mouse is likely to encode Muc3 and Muc12.

Seven Muc3 homologues were identified in *Xenopus tropicalis*. Three of them are on the same scaffold where each gene is separated by approximately 30,000 nt. It is possible that this region is homologous to the MUC3 cluster referred to above in higher animals.

MUC1, MUC13 and MUC16

The MUC1 mucin was only found in mammals, while MUC13 homologues were identified in human, mouse, chicken, Fugu and zebrafish. They all have a PTS-SEA-TM structure. The chicken Muc13 protein is unusual in that the PTS domain is encoded by many different exons. Each of these exons has 60 nucleotides and encodes a 20 amino acid long sequence. It is interesting to note that the sequence encoded by each exon is a repeated element in the PTS domain [Paper II]. Such a characteristic of Muc13 was also observed in one of the two zebrafish Muc13 homologues. The organization of repeats in multiple exons allows alternative splicing to generate transcripts encoding mucins with different repeat lengths.

MUC16 proteins have multiple SEA domains C-terminal of the PTS domain (more than 33 for human, 10 for mouse, at least 4 for chicken, and at least 3 for *Xenopus tropicalis*), and in each species the repeated SEA domains are similar.

MucMS (Multiple SEAs)

A new type of mucins, designated MucMS, was discovered. This is characterized by alternating PTS and SEA domains and most of them have a transmembrane domain. They form a large family in *Xenopus tropicalis*, Fugu, zebrafish and *Ciona intestinalis* (Figure 21), but are absent in higher animals.

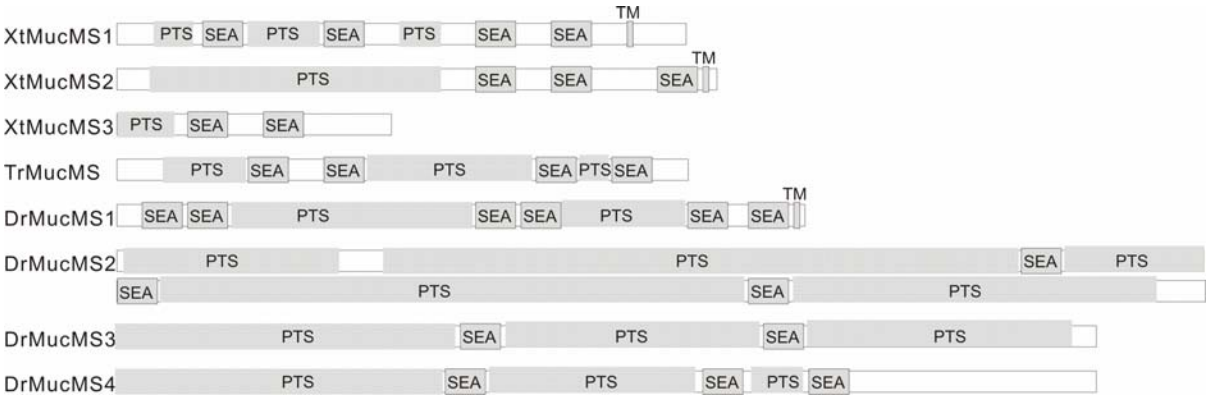

Figure 21. *New type of mucin with multiple PTS and SEA domains.* Organisms shown are *Xenopus tropicalis* (Xt), T*akifugu rubripes* (Tr, Fugu), and *Danio rerio* (Dr, zebrafish).

Cleavage site of mucins

A cleavage in the GDPH (Gly-Asp-Pro-His) sequence located in the VWD D4 region of the human MUC2 and MUC5AC mucin were reported by Martin E. Lidell *et al* [100, 101]. The protein can be cleaved at this site by a slow, non-enzymatic process. We have examined gel-forming mucins for GDPH sites suggesting that these mucins also should be cleaved. For the mouse and chicken MUC2 mucins, the D4 domains have a GDPH cleavage site. In *Xenopus tropicalis*, five of the MUC2 have GDPH cleavage sites and five have not. The D4 domain of zebrafish MUC2 has no GDPH cleavage site. For MUC5, all D4 domains in MUC5AC in

mouse and chicken have GDPH cleavage sites but MUC5B has not. In *Xenopus tropicalis*, Fugu, and zebrafish, some of the D4 domains in Muc5 have GDPH cleavage sites and some do not.

Transmembrane mucins typically contain 110-residue SEA domains located next to the membrane. These domains undergo post-translational cleavage between glycine and serine in a characteristic GSV sequence [102]. We have analyzed the cleavage site of all the mucin proteins with SEA domains. A GSV/GST/GSI was found in MUC1, MUC3 group, and the new type of mucin MucMS. In MUC16 we could not find any potential cleavage site. This could be used to distinguish MucMS and MUC16 since they both have multiple SEA domains.

Information presented at mucin web site

Much of the information that we have collected on studies of mucins and mucin related proteins is presented at http://www.medkem.gu.se/mucinbiology/databases/. Compilations of protein, mRNA, and DNA sequences, domain structures, and exon-intron structures for mucins in human, mouse, chicken, *Xenopus tropicalis*, Fugu, zebrafish, *Ciona intestinalis*, *Branchiostoma floridae*, *Strongylocentrotus purpuratus, Drosophila melanogaster*, and *Caenorhabditis elegans* are deposited there. In addition, there is information on mucin classification, PTS repeats, EST support of gene predictions, phylogenetic trees with bootstrapping values, GDPH cleavage sites and alignment of CysD domains.

Discussion

Prediction of mucin proteins

In order to identify mucin proteins we have used a combination of methods, including PTSpred that aims at predicting the unusual amino acid composition which is typical of mucin domains. We also made use of VWD and SEA domains and domain structures of mucin proteins. Our initial analysis of the human genome showed that the method was effective in predicting mucins. The approach taken cannot find mucins that only have a small PTS fragment like the human MUC7 and cannot find mucins that have frequencies of proline, threonine or serine that are different from our amino acid frequency model of the mucin domains.

Prediction of the PTS domain

We used the program PTSpred to predict large PTS domains [Paper I]. This program has been used on different genomes and predicted proteins. Our test on human using a threshold of S+T > 40% identified all the mucins including MUC7, MUC19 and MUC20. However, we did not identify the PTS domains of MUC1 as the frequency of T and S is only 25%.

As a rule we found that all organisms have proteins that contain both a PTS domain as well as either a VWD or a SEA domain. An exception is *Drosophila melanogaster* and *Caenorhabditis elegans*. In *Drosophila* PTSpred identified at least 12 large PTS tandem repeats and in *Caenorhabditis elegans* at least 8. Some of the corresponding proteins are previously annotated as ‘mucin’ or belonging to ‘mucin family’.

The repetitive nature and large size of the PTS domains causes difficulties in sequencing projects and in genome assembly. This is reflected in our work as several genes encoding mucins are incomplete. For this reason it is very important to examine different databases and extract relevant information from each of these.

In human and mouse, an important characteristic of the PTS domain in mucin proteins is the ‘one-exon rule’, i. e. each PTS domain is encoded by only one exon. At that time all mucin domains discovered were organized in this way and it was thus assumed that this was true for all mucins. However, this rule does clearly not apply to some lower animals as chicken, *Xenopus tropicalis* and zebrafish. Still most of the

mucins in these organisms follow the single exon rule. In chicken Muc13, zebrafish Muc13 and Muc2 proteins in *Xenopus tropicalis*, the PTS tandem repeats are encoded by multiple exons and each tandem repeats unit is encoded by one exon (two neighbour exons in some of the Muc2 proteins in *Xenopus tropicalis* with CPTS domain). During evolution from lower to higher animals the number of exons is increasing, but some mucins domains considered here do not seem to conform to this principle.

The use of domain structure in the prediction of mucin proteins

There are two major difficulties with PTS domains in the context of mucin analysis. One is that there are many false positives using PTSpred. For instance, the genome contains repetitive DNA sequences with no relationship to mucins. There are also intracellular proteins that contain sequences that could be classified as mucin domains. However, the hallmark of the mucins is that these PTS domains become heavily glycosylated when they pass the Golgi appratus and thus it is only secreted proteins that should be classified as mucins.

Another difficulty is that PTS domains are not well conserved between mucin homologues. Therefore, the mucin domain itself is of limited value for the identification and analysis of mucins. Instead, we have taken into consideration other domains of the mucins. These domains, such as VWD and SEA are well conserved as compared to PTS domains and also the domain structure of the mucin proteins tends to be conserved. Thus, most of the transmembrane type mucins have a SEA domain (MUC16 has multiple SEA domains) and all of the gel-forming type mucins have multiple VWD domains. MUC4 proteins always have an AMOP domain (Figure 2 and 3). In *Xenopus tropicalis*, zebrafish and *Ciona intestinalis*, we have also found a new type of mucin with alternating PTS and SEA domains.

Among all the Pfam domains found in mucins, VWD, SEA and AMOP domains have been particularly useful, but the CK, CysD, TIL and VWE domains have also been very informative in our analyses. We have also explored the EGF domain, but this large family is more difficult to predict as it seems to be quite poorly conserved in sequence except for the localization of the cysteines. It would perhaps be possible to identify a more mucin-specific EGF domain.

Some non-mucin proteins such as otogelin and IMGP2 have a domain structure which is similar to mucin proteins. Thus, otogelin has the same D1-D2-D3-PTS-D4-CK structure as MUC2 and MUC5, and IMPG2 has the same PTS-SEA structure as MUC1, MUC13 and the MUC3 family of proteins. This means that we cannot rely simply on domain architecture to assign homology relationship, but we need also a phylogenetic analysis based on sequence of individual domains or of full-length protein sequences.

For correct classification it is also important to consider as many sequences as possible. This is illustrated by our results regarding the mucins of Fugu. We originally identified three Muc2 homologues which we called fMUC2A, fMUC2B and fMUC2C, one Muc1 homolog which we called fMUC1, and one protein similar to the Muc2 family which we called fMUC2D [Paper I]. After a more rigorous analysis of VWD and SEA domains from a larger number of species (including chicken, *Xenopus tropicalis*, fish, and *Ciona intestinalis*) we found that fMUC2A is actually a Muc5 (Muc5b), fMUC2B is the otogelin homologue, fMUC2D is the von Willebrand factor homologue, and fMUC1 is a IMPG2 homologue. The fMuc2C is the only Muc2 mucin that was correctly described in Paper I.

Synteny and the prediction of mucin genes

The order of genes tends to be preserved between closely related organisms. This could be exploited in gene prediction, for instance to verify orthology as predicted by sequence analysis. For instance, in human the gene ACHE is located upstream of the MUC3 group of genes (MUC3, MUC12 and MUC17) and the gene TRIM56 is located downstream. We made use of this information to identify the C-terminal portion of Muc17 in mouse which was originally misclassified as Muc3. The ACHE and TRIM56 genes were also helpful to identify the Muc3, Muc12 and Muc17 genes in chimpanzee.

When we used PTSpred to analyze the human genome, we found a protein which has a large PTS domain, two EGF domains and a transmembrane domain whose domain structure is similar to that of a transmembrane type of mucin. Eventually we found that this protein is HEG (heart of glass) which was first identified in zebrafish where it regulates the growth of the heart [103]. It turned out through analysis of synteny that MUC13 is adjacent to HEG in human, mouse and chicken. Together with the fact that these proteins are similar, this might indicate that there is an evolutionary relationship between the two proteins [Paper II].

Recently R. Kawahara *et al* have analyzed the relationship of mucins to spiggins [105], a fish protein used in nest building. It has two VWD domains as well as TIL and VWD domains that are always adjacent to the VWD domains. The phylogenetic analysis was based on concatenated domains of spiggins and mucins and demonstrated that the closest relative of spiggin is MUC19 [104, 105]. This is a case where the orthology may be further explored by analysis of synteny.

Study of mucin evolution with a phylogenetic analysis of protein domains

We have made a phylogenetic analysis of SEA, VWD, AMOP, TIL and VWE domains. SEA, VWD, TIL and VWE domains were present at an early stage of metazoan evolution. Other domains characteristic of mucins appear at a later stage. The CysD domain is observed in Coelomata and the cysteine-knot in vertebrates. In the case of MUC4, PTS and TM domains are added to pre-existing AMOP-NIDO-VWD modules already present in lower animals such as *Xenopus tropicalis*. The most deeply branching animals where we observe a typical mucin protein are *Branchiostoma floridae* and *Nematostella vectensis* where we identified gel-forming mucin-like proteins.

The evolution of the PTS domains is difficult to study but still of great interest. As the PTS domains are very poorly conserved in sequence it is not possible to follow their evolution and to use these domains for tree reconstruction. However, the analysis of tandem repeats, exon-intron structure, the possibility of alternative splicing for PTS domains as well as the inserted CysD domains is very useful in the analysis of mucin evolution.

In chicken we have identified a mucin with one PTS domain and only one SEA domain that mapped to the MucMS group of mucins in the phylogenetic analysis. Therefore, this protein seems to be some type of intermediate between the MucMS type of mucins and the mammalian transmembrane type mucins.

Through the analysis of mucin proteins we have obtained an overview of mucin evolution across a wide range of species. From the present results it seems that during evolution, the number of types for secretory gel-forming mucin stays the same, while the number of types for transmembrane mucins increases from lower to higher vertebrates. However, the absolute number of mucin proteins increases dramatically in *Xenopus tropicalis* and fish. The gel-forming mucins can be traced to early metazoan evolution, while all the

membrane anchored mucins, both the mucins with SEA domains and the VWD-containing MUC4 seem to have appeared in the vertebrate lineage. This might be related to their possible function in signaling processes in vertebrates.

Acknowledgements

My supervisors **Gunnar C. Hansson** and **Tore Samuelsson**. My most sincere gratitude to you for giving me the opportunity to begin and complete this PhD project; continuous supporting in the PhD programme; providing me with the direction, ideas, methods and guidance; and helping me organizing, writing and proofreading all the papers and thesis. This work will never be finished without your infinite patience and indispensable assistance. My special thanks for **Gunnar** for bring me to the world of biology and biochemistry, accepting me as a member of the mucin biology group (ever since the master project), and teaching me the knowledge and technology. My special thanks for **Tore** for bring me to the world of bioinformatics and computer science, accepting me as a member of the bioinformatics group, and solving all kinds of tough problems whenever I am in need.

All the people in Gunnar's Mucin biology group during this time. **Fredrik J. Olson, Julia Fernandez-Rodriguez, Christian X. Andersson, Malin Bäckström, Malin Johansson, Martin E. Lidell, Tina Thomsson, Jessica M. Holmén Larsson, Emily K. Malmberg, Sirle Laos, Åsa Petersson, Ana Maria Rodriguez**, and all other people who have been in the group. Thanks for helping me with the biology knowledge as well as the Swedish language. All the people in Tore's bioinformatics group. **Magnus Alm Rosenblad, Paul Piccinelli, Liqun He,** and **Marcela Dávila López.** Thanks for discussing the bioinformatics problems and issues.

The teachers in the Bioinformatics of International Master Programme of Chalmers University of Technology. **Peter Jagers, Olle Nerman, Marita Olsson, Serik Sagitov, Anders Blomberg,** and **Graham Kemp**. Thanks for teaching me the knowledge and enlarging my insight in the bioinformatic field.

I am indebted for **Nils Lycke**, Clinical Immunology, Sahlgrenska University Hospital for acting as a co-supervisor.

This work was supported by the Swedish Research Council (#7461), Wilhelm and Martina Lundgren's fund, the Sahlgrenska University Hospital and by the Swedish Knowledge Foundation through the Industrial PhD program in Medical Bioinformatics at the Strategy and Development Office (SDO) at the Karolinska Institute that provided most of the Ph.D. fellowship.

References

1. Toribara, N.W., et al., *Human gastric mucin. Identification of a unique species by expression cloning.* J Biol Chem, 1993. **268**(8): p. 5879-85.
2. Gendler, S.J. and A.P. Spicer, *Epithelial mucin genes.* Annu Rev Physiol, 1995. **57**: p. 607-34.
3. Taupin, D. and D.K. Podolsky, *Trefoil factors: initiators of mucosal healing.* Nat Rev Mol Cell Biol, 2003. **4**(9): p. 721-32.
4. Hollingsworth, M.A. and B.J. Swanson, *Mucins in cancer: protection and control of the cell surface.* Nat Rev Cancer, 2004. **4**(1): p. 45-60.
5. Gendler, S.J., et al., *Molecular cloning and expression of human tumor-associated polymorphic epithelial mucin.* J Biol Chem, 1990. **265**(25): p. 15286-93.
6. Swallow, D.M., et al., *The human tumour-associated epithelial mucins are coded by an expressed hypervariable gene locus PUM.* Nature, 1987. **328**(6125): p. 82-4.
7. Olson, F.J., et al., *Blood group A glycosyltransferase occurring as alleles with high sequence difference is transiently induced during a Nippostrongylus brasiliensis parasite infection.* J Biol Chem, 2002. **277**(17): p. 15044-52.
8. Swallow, D.M., et al., *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, 1987. **51**(Pt 4): p. 289-94.
9. Van der Sluis, M., et al., *Muc2-deficient mice spontaneously develop colitis, indicating that MUC2 is critical for colonic protection.* Gastroenterology, 2006. **131**(1): p. 117-29.
10. Davis, C.G., et al., *Deletion of clustered O-linked carbohydrates does not impair function of low density lipoprotein receptor in transfected fibroblasts.* J Biol Chem, 1986. **261**(6): p. 2828-38.
11. Perez-Vilar, J. and R.L. Hill, *The structure and assembly of secreted mucins.* J Biol Chem, 1999. **274**(45): p. 31751-4.
12. Shi, X., et al., *Modeling the structure of the type I peritrophic matrix: characterization of a Mamestra configurata intestinal mucin and a novel peritrophin containing 19 chitin binding domains.* Insect Biochem Mol Biol, 2004. **34**(10): p. 1101-15.
13. Yang, Z.Y., et al., *Identification of the Ebola virus glycoprotein as the main viral determinant of vascular cell cytotoxicity and injury.* Nat Med, 2000. **6**(8): p. 886-9.
14. Debailleul, V., et al., *Human mucin genes MUC2, MUC3, MUC4, MUC5AC, MUC5B, and MUC6 express stable and extremely large mRNAs and exhibit a variable length polymorphism. An improved method to analyze large mRNAs.* J Biol Chem, 1998. **273**(2): p. 881-90.
15. Wreschner, D.H., et al., *Generation of ligand-receptor alliances by "SEA" module-mediated cleavage of membrane-associated mucin proteins.* Protein Sci, 2002. **11**(3): p. 698-706.
16. Gendler, S.J., *MUC1, the renaissance molecule.* J Mammary Gland Biol Neoplasia, 2001. **6**(3): p. 339-53.
17. Vinall, L.E., et al., *Altered expression and allelic association of the hypervariable membrane mucin MUC1 in Helicobacter pylori gastritis.* Gastroenterology, 2002. **123**(1): p. 41-9.
18. Yamamoto, M., et al., *Interaction of the DF3/MUC1 breast carcinoma-associated antigen and beta-catenin in cell adhesion.* J Biol Chem, 1997. **272**(19): p. 12492-4.
19. Smorodinsky, N., et al., *Detection of a secreted MUC1/SEC protein by MUC1 isoform specific monoclonal antibodies.* Biochem Biophys Res Commun, 1996. **228**(1): p. 115-21.
20. Moniaux, N., et al., *Structural organization and classification of the human mucin genes.* Front Biosci, 2001. **6**: p. D1192-206.

21. Hinojosa-Kurtzberg, A.M., et al., *Novel MUC1 splice variants contribute to mucin overexpression in CFTR-deficient mice.* Am J Physiol Gastrointest Liver Physiol, 2003. **284**(5): p. G853-62.
22. Gum, J.R., J.W. Hicks, and Y.S. Kim, *Identification and characterization of the MUC2 (human intestinal mucin) gene 5'-flanking region: promoter activity in cultured cells.* Biochem J, 1997. **325 (Pt 1)**: p. 259-67.
23. Desseyn, J.L., et al., *Evolution of the large secreted gel-forming mucins.* Mol Biol Evol, 2000. **17**(8): p. 1175-84.
24. Baeckstrom, D. and G.C. Hansson, *The transcripts of the apomucin genes MUC2, MUC4, and MUC5AC are large and appear as distinct bands.* Glycoconj J, 1996. **13**(5): p. 833-7.
25. Rousseau, K., et al., *The complete genomic organization of the human MUC6 and MUC2 mucin genes.* Genomics, 2004. **83**(5): p. 936-9.
26. Desseyn, J.L., et al., *Evolutionary history of the 11p15 human mucin gene family.* J Mol Evol, 1998. **46**(1): p. 102-6.
27. Gum, J.R., Jr., et al., *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, 1994. **269**(4): p. 2440-6.
28. Godl, K., et al., *The N terminus of the MUC2 mucin forms trimers that are held together within a trypsin-resistant core fragment.* J Biol Chem, 2002. **277**(49): p. 47248-56.
29. Asker, N., et al., *Dimerization of the human MUC2 mucin in the endoplasmic reticulum is followed by a N-glycosylation-dependent transfer of the mono- and dimers to the Golgi apparatus.* J Biol Chem, 1998. **273**(30): p. 18857-63.
30. Axelsson, M.A., N. Asker, and G.C. Hansson, *O-glycosylated MUC2 monomer and dimer from LS 174T cells are water-soluble, whereas larger MUC2 species formed early during biosynthesis are insoluble and contain nonreducible intermolecular bonds.* J Biol Chem, 1998. **273**(30): p. 18864-70.
31. Spada, F., et al., *Molecular patterning of the oikoplastic epithelium of the larvacean tunicate Oikopleura dioica.* J Biol Chem, 2001. **276**(23): p. 20624-32.
32. Desseyn, J.L., et al., *Genomic organization of the human mucin gene MUC5B. cDNA and genomic sequences upstream of the large central exon.* J Biol Chem, 1998. **273**(46): p. 30157-64.
33. Li, D., et al., *Cloning of the amino-terminal and 5'-flanking region of the human MUC5AC mucin gene and transcriptional up-regulation by bacterial exoproducts.* J Biol Chem, 1998. **273**(12): p. 6812-20.
34. Chen, Y., et al., *Characterization of human mucin 5B gene expression in airway epithelium and the genomic clone of the amino-terminal and 5'-flanking region.* Am J Respir Cell Mol Biol, 2001. **25**(5): p. 542-53.
35. Toribara, N.W., et al., *The carboxyl-terminal sequence of the human secretory mucin, MUC6. Analysis Of the primary amino acid sequence.* J Biol Chem, 1997. **272**(26): p. 16398-403.
36. Chen, Y., et al., *Genome-wide search and identification of a novel gel-forming mucin MUC19/Muc19 in glandular tissues.* Am J Respir Cell Mol Biol, 2004. **30**(2): p. 155-65.
37. Culp, D.J., et al., *The gene encoding mouse Muc19: cDNA, genomic organization and relationship to Smgc.* Physiol Genomics, 2004. **19**(3): p. 303-18.
38. Desseyn, J.L., I. Clavereau, and A. Laine, *Cloning, chromosomal localization and characterization of the murine mucin gene orthologous to human MUC4.* Eur J Biochem, 2002. **269**(13): p. 3150-9.
39. Moniaux, N., et al., *Complete sequence of the human mucin MUC4: a putative cell membrane-associated mucin.* Biochem J, 1999. **338 (Pt 2)**: p. 325-33.
40. Escande, F., et al., *Genomic organization of MUC4 mucin gene. Towards the characterization of splice variants.* Eur J Biochem, 2002. **269**(15): p. 3637-44.
41. Moniaux, N., et al., *Alternative splicing generates a family of putative secreted and membrane-associated MUC4 mucins.* Eur J Biochem, 2000. **267**(14): p. 4536-44.

42. Wu, K., N. Fregien, and K.L. Carraway, *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, 1994. **269**(16): p. 11950-5.
43. Crawley, S.C., et al., *Genomic organization and structure of the 3' region of human MUC3: alternative splicing predicts membrane-bound and soluble forms of the mucin.* Biochem Biophys Res Commun, 1999. **263**(3): p. 728-36.
44. Gum, J.R., et al., *Molecular cloning of cDNAs derived from a novel human intestinal mucin gene.* Biochem Biophys Res Commun, 1990. **171**(1): p. 407-15.
45. Gum, J.R., Jr., et al., *MUC3 human intestinal mucin. Analysis of gene structure, the carboxyl terminus, and a novel upstream repetitive region.* J Biol Chem, 1997. **272**(42): p. 26678-86.
46. Pratt, W.S., et al., *Multiple transcripts of MUC3: evidence for two genes, MUC3A and MUC3B.* Biochem Biophys Res Commun, 2000. **275**(3): p. 916-23.
47. Gum, J.R., Jr., et al., *Initiation of transcription of the MUC3A human intestinal mucin from a TATA-less promoter and comparison with the MUC3B amino terminus.* J Biol Chem, 2003. **278**(49): p. 49600-9.
48. Kyo, K., et al., *Associations of distinct variants of the intestinal mucin gene MUC3A with ulcerative colitis and Crohn's disease.* J Hum Genet, 2001. **46**(1): p. 5-20.
49. Williams, S.J., et al., *The MUC3 gene encodes a transmembrane mucin and is alternatively spliced.* Biochem Biophys Res Commun, 1999. **261**(1): p. 83-9.
50. Williams, S.J., et al., *Two novel mucin genes down-regulated in colorectal cancer identified by differential display.* Cancer Res, 1999. **59**(16): p. 4083-9.
51. Gum, J.R., Jr., et al., *MUC17, a novel membrane-tethered mucin.* Biochem Biophys Res Commun, 2002. **291**(3): p. 466-75.
52. Van Klinken, B.J., et al., *Molecular cloning of human MUC3 cDNA reveals a novel 59 amino acid tandem repeat region.* Biochem Biophys Res Commun, 1997. **238**(1): p. 143-8.
53. Liu, B., et al., *The recombinant N-terminal region of human salivary mucin MG2 (MUC7) contains a binding domain for oral Streptococci and exhibits candidacidal activity.* Biochem J, 2000. **345 Pt 3**: p. 557-64.
54. Lehmann, J., et al., *[Expression of antimicrobial peptide MUC7 in kidneys with pyelonephritis].* Urologe A, 2006. **45**(10): p. 1300, 1302-7.
55. Bobek, L.A., et al., *Molecular cloning, sequence, and specificity of expression of the gene encoding the low molecular weight human salivary mucin (MUC7).* J Biol Chem, 1993. **268**(27): p. 20563-9.
56. Bateman, A., et al., *Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins.* Nucleic Acids Res, 1999. **27**(1): p. 260-2.
57. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2000. **28**(1): p. 263-6.
58. Sonnhammer, E.L., et al., *Pfam: multiple sequence alignments and HMM-profiles of protein domains.* Nucleic Acids Res, 1998. **26**(1): p. 320-2.
59. Studholme, D.J., et al., *A comparison of Pfam and MEROPS: two databases, one comprehensive, and one specialised.* BMC Bioinformatics, 2003. **4**: p. 17.
60. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.* J Mol Biol, 2001. **305**(3): p. 567-80.
61. Sonnhammer, E.L., G. von Heijne, and A. Krogh, *A hidden Markov model for predicting transmembrane helices in protein sequences.* Proc Int Conf Intell Syst Mol Biol, 1998. **6**: p. 175-82.
62. Nielsen, H., et al., *Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.* Protein Eng, 1997. **10**(1): p. 1-6.
63. Bendtsen, J.D., et al., *Improved prediction of signal peptides: SignalP 3.0.* J Mol Biol, 2004. **340**(4): p. 783-95.

64. Nielsen, H. and A. Krogh, *Prediction of signal peptides and signal anchors by a hidden Markov model.* Proc Int Conf Intell Syst Mol Biol, 1998. **6**: p. 122-30.
65. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
66. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
67. Wheelan, S.J., D.M. Church, and J.M. Ostell, *Spidey: a tool for mRNA-to-genomic alignments.* Genome Res, 2001. **11**(11): p. 1952-7.
68. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
69. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**(9): p. 755-63.
70. Birney, E., M. Clamp, and R. Durbin, *GeneWise and Genomewise.* Genome Res, 2004. **14**(5): p. 988-95.
71. Felsenstein, J., *Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods.* Methods Enzymol, 1996. **266**: p. 418-27.
72. Retief, J.D., *Phylogenetic analysis using PHYLIP.* Methods Mol Biol, 2000. **132**: p. 243-58.
73. O'Sullivan, O., et al., *3DCoffee: combining protein sequences and structures within multiple sequence alignments.* J Mol Biol, 2004. **340**(2): p. 385-95.
74. Notredame, C., D.G. Higgins, and J. Heringa, *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* J Mol Biol, 2000. **302**(1): p. 205-17.
75. Notredame, C., L. Holm, and D.G. Higgins, *COFFEE: an objective function for multiple sequence alignments.* Bioinformatics, 1998. **14**(5): p. 407-22.
76. Maddison, D.R., D.L. Swofford, and W.P. Maddison, *NEXUS: an extensible file format for systematic information.* Syst Biol, 1997. **46**(4): p. 590-621.
77. Perriere, G. and M. Gouy, *WWW-query: an on-line retrieval system for biological sequence banks.* Biochimie, 1996. **78**(5): p. 364-9.
78. Page, R.D., *TreeView: an application to display phylogenetic trees on personal computers.* Comput Appl Biosci, 1996. **12**(4): p. 357-8.
79. Lehtonen, J.V., et al., *Finding local structural similarities among families of unrelated protein structures: a generic non-linear alignment algorithm.* Proteins, 1999. **34**(3): p. 341-55.
80. Page, R.D. and M.A. Charleston, *From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem.* Mol Phylogenet Evol, 1997. **7**(2): p. 231-40.
81. Day, W.H., *Computational complexity of inferring phylogenies from dissimilarity matrices.* Bull Math Biol, 1987. **49**(4): p. 461-7.
82. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol, 1987. **4**(4): p. 406-25.
83. Chevenet, F., et al., *TreeDyn: towards dynamic graphics and annotations for analyses of trees.* BMC Bioinformatics, 2006. **7**: p. 439.
84. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA.* J Mol Biol, 1997. **268**(1): p. 78-94.
85. Heger, A. and L. Holm, *Rapid automatic detection and alignment of repeats in protein sequences.* Proteins, 2000. **41**(2): p. 224-37.
86. Durand, P., et al., *Browsing repeats in genomes: Pygram and an application to non-coding region analysis.* BMC Bioinformatics, 2006. **7**: p. 477.
87. Jones, D.T., *Protein secondary structure prediction based on position-specific scoring matrices.* J Mol Biol, 1999. **292**(2): p. 195-202.
88. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

89. Probst, J.C., E.M. Gertzen, and W. Hoffmann, *An integumentary mucin (FIM-B.1) from Xenopus laevis homologous with von Willebrand factor.* Biochemistry, 1990. **29**(26): p. 6240-4.
90. Aparicio, S., et al., *Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes.* Science, 2002. **297**(5585): p. 1301-10.
91. Brenner, S., et al., *Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome.* Nature, 1993. **366**(6452): p. 265-8.
92. Elgar, G., et al., *Generation and analysis of 25 Mb of genomic DNA from the pufferfish Fugu rubripes by sequence scanning.* Genome Res, 1999. **9**(10): p. 960-71.
93. Sadler, J.E., *von Willebrand factor.* J Biol Chem, 1991. **266**(34): p. 22777-80.
94. Meiniel, O. and A. Meiniel, *The complex multidomain organization of SCO-spondin protein is highly conserved in mammals.* Brain Res Rev, 2007. **53**(2): p. 321-7.
95. Watanabe, K., et al., *Amino acid sequence of alpha-subunit in hen egg white ovomucin deduced from cloned cDNA.* DNA Seq, 2004. **15**(4): p. 251-61.
96. Cohen-Salmon, M., et al., *Otogelin: a glycoprotein specific to the acellular membranes of the inner ear.* Proc Natl Acad Sci U S A, 1997. **94**(26): p. 14450-5.
97. Goto, A., et al., *A Drosophila haemocyte-specific protein, hemolectin, similar to human von Willebrand factor.* Biochem J, 2001. **359**(Pt 1): p. 99-108.
98. Lesch, C., et al., *A role for Hemolectin in coagulation and immunity in Drosophila melanogaster.* Dev Comp Immunol, 2007.
99. Choudhury, A., et al., *Human MUC4 mucin cDNA and its variants in pancreatic carcinoma.* J Biochem (Tokyo), 2000. **128**(2): p. 233-43.
100. Lidell, M.E. and G.C. Hansson, *Cleavage in the GDPH sequence of the C-terminal cysteine-rich part of the human MUC5AC mucin.* Biochem J, 2006. **399**(1): p. 121-9.
101. Lidell, M.E., M.E. Johansson, and G.C. Hansson, *An autocatalytic cleavage in the C terminus of the human MUC2 mucin occurs at the low pH of the late secretory pathway.* J Biol Chem, 2003. **278**(16): p. 13944-51.
102. Macao, B., et al., *Autoproteolysis coupled to protein folding in the SEA domain of the membrane-bound MUC1 mucin.* Nat Struct Mol Biol, 2006. **13**(1): p. 71-6.
103. Mably, J.D., et al., *heart of glass regulates the concentric growth of the heart in zebrafish.* Curr Biol, 2003. **13**(24): p. 2138-47.
104. Kawahara, R. and M. Nishida, *Multiple occurrences of spiggin genes in sticklebacks.* Gene, 2006. **373**: p. 58-66.
105. Jones, I., et al., *Molecular cloning and characterization of spiggin. An androgen-regulated extraorganismal adhesive with structural similarities to von Willebrand Factor-related proteins.* J Biol Chem, 2001. **276**(21): p. 17857-63.