In silico identification and microarray analysis of human mucin-like genes

Chikako Nishi-Takaoka [1,2], Tatsunari Nishi [1,2], Takahiro Shimamura[2], Shogo Yamamoto[2], Yoshitaka Hippo[2], Hiroyuki Aburatani[2]

[1]Genaris, Inc., [2]Genome Science Division, Research Center for Advanced Science and Technology, The University of Tokyo

(Received April 10, 2007)

Abstract

Mucins, which are highly *O*-glycosylated proteins, are involved in a variety of diseases. Most of the known mucins show little sequence conservation. To identify novel candidate mucins that might be helpful for diagnosis and therapy of diseases, bioinformatics approaches other than homology searches are required. We have developed two bioinformatics approaches that focus on either the amino acid compositions or the repeated sequences in regions carrying *O*-linked glycans. Both approaches identify candidate mucins efficiently from the human reference sequence database. The former and latter approaches identified three and two novel mucin-like genes, respectively. We used Affymetrix GeneChip oligonucleotide microarrays to analyze the mucins identified from the database and found that three human mucin genes differed significantly in expression between normal tissues/cells and cancer cell lines. A combination of our bioinformatics approaches and the oligonucleotide microarray method will be useful for the efficient identification of candidate human mucins. Further study of the mucins in cancer tissues will facilitate the direct evaluation of their potential for cancer diagnosis and therapy.

Introduction

Mucins were originally characterized as highly glycosylated proteins that are major constituents of the mucus covering epithelial surfaces, such as those in the trachea and colon [1]. To date, 18 human epithelial mucin proteins (MUC1, MUC2, MUC3A, MUC3B, MUC4, MUC5AC, MUC5B, MUC6, MUC7, MUC8, MUC11, MUC12, MUC13, MUC15, MUC16, MUC17, MUC19 and MUC20) have been identified and assigned to the MUC family [2–5]. MUC3A and MUC3B proteins have been reported to derive from the MUC3 gene, and MUC5AC and MUC5B have been shown to be the MUC5 gene products. These MUC-type mucins play important roles in protecting epithelial cells and in cell-cell interactions. Alteration of the glycan structures on cell surface molecules, including mucins, is recognized as a general feature of carcinogenesis; the anti-MUC16 antibody CA125 has been used as a marker for cancer diagnosis [6]. Analyzes of *Muc1*-deficient mice have shown that this gene is important in the progression of mammary carcinoma [7]. Recently, the observation that deletion of the Muc2 gene in mice frequently leads to colon cancer has provided more direct evidence for the relationship between MUC-type mucins and carcinogenesis [8].

The MUC-type mucins are characterized by domain(s) rich in Pro, Thr and Ser, which are thought to be scaffolds for many *O*-linked carbohydrate side chains. Several of these mucin domains, also referred to as PTS regions, contain PTS-rich repeats [2]. In addition, non-MUC-type mucins such as P-selectin ligand [9] and mucosal vascular addressin cell adhesion molecule 1 [10] have been characterized. Several of these non-MUC-type mucins also contain PTS-rich repeats, but their PTS regions are generally shorter than those of the MUC-type mucins. Some non-MUC-type mucins are involved in inflammation and infection, and antibodies against them may eventually be of clinical value. Identification of novel human mucins might therefore be useful for diagnosis and therapy of various diseases.

The publicly available draft sequence database of the human genome has made it easier to identify novel mucins *in silico*. Homology search programs such as BLAST [11] are not appropriate for identifying novel mucins because most of the known mucins show little sequence conservation and they have no common protein motifs. In this study, therefore, we developed two new bioinformatics approaches, which we used to identify several novel human mucins from human reference sequence (RefSeq) database. In addition, we used Affymetrix GeneChip oligonucleotide microarrays to identify human mucin genes that differ significantly in expression between normal tissues/cells and cancer cell lines.

Materials and methods

Data

Authentic amino acid sequences of 18 human MUC-type mucins derived from 16 genes (*MUC1*, *MUC2*, *MUC3*, *MUC4*, *MUC5*, *MUC6*, *MUC7*, *MUC8*, *MUC11*, *MUC12*, *MUC13*, *MUC15*, *MUC16*, *MUC17*, *MUC19* and *MUC20*) were obtained from UniProtKB/Swiss-Prot database (http://www.expasy.uniprot.org/), because UniProtKB/Swiss-Prot database containing manually well-annotated records with information extracted from literature provided a high level of annotation. These mucin sequences were used for protein motif searches and for analysis of amino acid contents in their PTS regions. The human RefSeq database (release 19) that contains of 34,128 protein records was downloaded ftp://ftp.ncbi.nih.gov/refseq/ and used for screening for human novel mucins, because this database contains protein sequences from several sources including computational analysis at National Center for Biological Information and is supposed to contain novel mucin sequences. The RefSeq database includes protein sequences derived from 15 human MUC-type mucin genes except the MUC8 gene.

Human normal cells and carcinoma cell lines

Twenty-three human tumor cell lines derived from different tumor types (gastric, liver, colon, lung, bladder, cervical, esophagea and breast) were obtained from the American Type Culture Collection (Manassas, VA), Riken Cell Bank (Tsurumi, Japan), the Cell Resource Center for Biomedical Research at Tohoku University (Sendai, Japan) and the Japanese Collection of Research Bioresources (Tokyo, Japan). Human skin fibroblasts (KMS-6) were purchased from Dainippon Pharmaceutical Co. Ltd (Osaka, Japan). Information about the other human cell lines is available at http://157.82.78.238/refexa/tissue.jsp.

RNA extraction and high-density oligonucleotide array analysis

Total RNA was isolated from frozen samples of human normal and carcinoma cells with ISOGEN (Nippon Gene, Tokyo, Japan) according to the manufacturer's protocol. The quality of the total RNA was examined by gel electrophoresis to confirm that the ribosomal 28S and 18S RNA bands were intact. The preparation of the RNA samples from human tissues was described previously [12]. Experimental procedures for high-density oligonucleotide microarrays were performed according to the Affymetrix GeneChip Expression Analysis Technical Manual. Briefly, 10 µg of RNA was used to synthesize biotin-labeled cRNA, which was then hybridized to the high-density oligonucleotide array (GeneChip Human U133 Array; Affymetrix, Santa Clara, CA). After washing, the arrays were stained with streptavidin-phycoerythrin, and image data were collected and analyzed with a Hewlett-Packard scanner. The GeneChip Analysis Suite software (version 5.0) was used to calculate the average difference for each gene probe on the array. The average differences were normalized for each array to generate a mean value.

PTS region-scanning approach

The PTS region-scanning approach identifies PTS regions by moving windows of size *w* in 1-aa increments along a peptide sequence and calculating the compositions of Pro, Thr/Ser, Cys, Met, charged amino acids and hydrophobic amino acids. A protein is considered a candidate mucin when the amino acid contents of a window all exceed the thresholds listed in Table 1. To obtain an optimal selection condition for human mucins, we varied *w* from 20 to 70 amino acids in 5-amino acid increments.

PTS repeats-searching approach

The PTS-searching approach identifies proteins carrying at least three repeats of a sequence of 6–9 residues with a Pro + Thr + Ser content exceeding 35%. For example, we can search for repeated occurrence of an 8-residue sequence such as PXXTSXTX (X indicates any amino acid) that contains one or two Pro residue(s) and at least four Pro/Ser/Thr residues. Pro, Thr or Ser should be present at the leftmost position and in at least three of the remaining seven positions. The number of ways to arrange these residues is calculated as: ${}_7C_3 \times {}_4C_1 \times 2^3 \times {}_7C_3 \times {}_4C_2 \times 2^2 = 1960$. When each of these 1960 sequence patterns is scanned along a protein sequence and a pattern is found at least three times in the protein, the protein is identified as a candidate mucin.

Other bioinformatics methods

The BLAST program [11] for protein homology searches and the Pfam database [13] for protein motif searches were used via the Internet at http://www.ncbi.nlm.nih.gov/BLAST/ and at http://www.sanger.ac.uk/Software/Pfam/, respectively. Mucin-type GalNAc *O*-glycosylation sites in mammalian proteins were predicted with the NetOGlyc 3.1 server [14] at http://www.cbs.dtu.dk/services/NetOGlyc/. Signal peptides were predicted with the SignalP 3.0 server [15] at http://www.cbs.dtu.dk/services/SignalP/. Transmembrane domains in proteins were identified with TMHMM Server v. 2.0 [16] at http://www.cbs.dtu.dk/services/TMHMM/.

Results

Development of a bioinformatics approach to identify mucins on the basis of amino acid compositions of PTS regions

It is widely accepted that most MUC-type mucins are characterized by the presence of motifs such as von Willebrand-factor (vWF) domains, Sea urchin sperm protein Enterokinase Agrin (SEA) domains and epidermal growth-factor (EGF)-like domains [2]. The 18 human mucins belonging to the MUC family were selected from UniProtKB/Swiss-Prot database at http://www.expasy.uniprot.org/. These mucins were subjected for protein motifs searches using the Pfam database [13]; no protein motifs were detected in the sequences of MUC7, MUC11, MUC15, MUC19 and MUC20, indicating that the presence of these motifs is not a prerequisite for mucins (data not shown).

To identify candidate mucins, we therefore developed a method similar to that of Lang et al. [17], which focused on the Pro, Ser and Thr contents of the PTS regions of MUC-type mucins. Using this approach, we attempted to select candidate mucins from human RefSeq database but obtained a number of false positives with the selected molecules. Therefore, we asked whether the PTS regions of 16 authentic human MUC-type mucins (excluding MUC3A and MUC5AC) in the UniProtKB/Swiss-Prot database had characteristic features other than the abundance of Pro, Thr and Ser (Table 1). MUC3A and MUC5AC were omitted because neither has a PTS region and NetOGlyc programs predict only a few *O*-linked glycosylation sites [14]. The compositions of various amino acids in the PTS region were compared with the average compositions of protein sequences in the RefSeq database, and this approach identified the specific characteristics of MUC-type mucins shown in Table 1. The Ser and Thr contents were much higher, and the Pro content was relatively higher, in the PTS regions than in the human protein averages. Cys residues were extremely rare in the PTS regions. Moreover, charged amino acids (Lys, Arg, Glu and Asp) showed much less than the average contents, and Met and the hydrophobic amino acids (Leu, Ile, Val and Phe) were also below the human protein averages. On the basis of these results we developed an approach, which we call PTS region-scanning, that identifies PTS regions by moving a window along a protein sequence and calculating the contents of Cys, Met, charged amino acids and hydrophobic amino acids in addition to Pro and Thr/Ser. The threshold values for amino acid contents are based on the values from the PTS regions of the characterized human MUC-type mucins.

We attempted to optimise the setting for identifying candidate human mucins from the RefSeq database by varying these threshold values and the window size. We first examined protein sequences derived from 15 MUC-type mucin genes included in the RefSeq database and optimised the amino acid content threshold values, for window sizes between 40 and 70 amino acids, by varying the thresholds with reference to the amino acid contents of the PTS regions in the known human MUC-type mucins (Table 2). The MUC1 protein in the RefSeq database was not selected with the PTS region-scanning approach, because a MUC1 sequence in the

Table 1 Amino acid compositions in PTS regions of human MUC-type mucins

Gene name	UniProt: accession no.	Position of PTS region[a]	Length (amino acids)	Pro + Thr + Ser (%)	Thr + Ser (%)	Pro (%)	Met (%)	Cys (%)	Charged amino acids[b] (%)	Hydrophobic amino acids[c] (%)
MUC1	P15941	99-942	844	50.0	25.71	24.9	0.0	0.0	10.0	5.3
MUC2	Q02817	1869-4338	2470	81.2	59.7	21.5	0.1	0.2	0.4	8.9
MUC3	Q9UKW9	1-857	857	64.2	55.1	9.1	1.2	0.5	7.1	16.8
MUC4	Q9NY09	868-1071	204	52.5	40.7	11.8	2.5	0.0	9.3	13.7
MUC5	Q9HC84	2932-3550	619	61.4	48.9	12.4	0.8	0.8	5.2	11.1
MUC6	Q6W4X9	1185-1569	385	60.0	42.3	17.7	1.3	0.5	6.5	9.1
MUC7	Q8TAX7	75-359	285	60.0	35.4	24.5	0.0	0.0	7.0	9.5
MUC8	Q12964	260-297	38	50.0	31.6	18.4	0.0	2.6	13.2	5.3
MUC11	Q9UKN0	1-957	957	58.0	48.2	9.8	0.4	0.0	10.8	10.9
MUC12	Q9UKN	1-240	240	52.9	43.8	9.2	0.8	0.4	9.2	16.3
MUC13	Q9H3R2	27-126	100	57.0	43.0	14.0	0.0	0.0	7.0	15.0
MUC15	Q8N387	149-205	57	47.3	36.8	10.5	0.0	0.0	10.5	26.3
MUC16	Q96RK2	2952-3790	839	52.6	41.2	11.3	1.9	0.0	10.4	18.5
MUC17	Q685J3	71-4146	4076	60.2	48.7	11.6	2.6	0.1	9.3	15.3
MUC19	Q7Z5P9	96-191	96	47.9	43.8	4.2	1.0	0.0	7.3	15.6
MUC20	Q8N307	161-257	97	50.5	37.1	13.4	0.0	0.0	14.1	13.4
Average compositions found in human proteins				20.3	13.7	6.6	2.1	2.3	23.3	23.4

[a] PTS regions were identified as regions within a 30 amino-acid window where the content of Pro, Thr and Ser is at least 50%. Numbers indicate amino acid positions from the first Met. For each MUC-type mucin, a representative PTS region was selected, and the amino acid composition was calculated.

[b] Charged amino acids include Lys, Arg, Glu and Asp.

[c] Hydrophobic amino acids include Leu, Ile, Val and Phe.

Table 2 Optimal conditions for selection of MUC-type mucins and selection of candidate mucins from human RefSeq database

Window size (residues)	Optimal amino acid residues for selection of human MUC-type mucins							Number of mucin candidates identified from human RefSeq database
	Pro + Thr+ Ser	Thr+ Ser	Pro	Met	Cys	Charged amino acids[b]	Hydrophobic amino acids[a]	
40	≥ 20	≥ 13	≥ 3	≤ 1	≤ 2	≤ 4	≤ 7	1684
45	≥ 23	≥ 15	≥ 3	≤ 1	≤ 2	≤ 6	≤ 10	1454
50	≥ 25	≥ 17	≥ 3	≤ 1	≤ 2	≤ 6	≤ 10	1126
55	≥ 28	≥ 18	≥ 3	≤ 1	≤ 2	≤ 7	≤ 12	924
55	≥ 28	≥ 18	≥ 3	(–)[c]	(–)[c]	(–)[c]	(–)[c]	1358
60	≥ 29	≥ 19	≥ 3	≤ 1	≤ 2	≤ 8	≤ 14	1069
65	≥ 30	≥ 20	≥ 3	≤ 1	≤ 2	≤ 9	≤ 15	1168
70	≥30	≥ 20	≥ 3	≤ 1	≤ 2	≤ 9	≤ 15	1307

[a] Hydrophobic amino acids include Leu, Ile, Val and Phe.
[b] Charged amino acids include Lys, Arg, Glu and Asp.
[c] (–) means that no threshold was applied.

RefSeq database contains only an N-terminal portion of the complete MUC1 protein and lacks any PTS regions. We then identified mucin candidates from the RefSeq database using the optimized thresholds for each window size. We found that the optimal window size was 55 amino acids (Table 2). This setting had greater specificity, without sacrificing sensitivity, than that of a selection based only on the Pro and Thr/Ser compositions (Table 2). Using this approach, we screened for human mucin proteins in the RefSeq database. This screen yielded 924 candidates. To evaluate these we designed an index, which we called the "PTS index", that is the ratio of the combined length of the identified PTS regions to the entire protein length. When the mucin candidates were listed in decreasing order with the limits of not less than 0.1 by this index value, known MUC-type mucins except MUC1 in the RefSeq database were found among the mucin candidates. This result confirms that this approach is appropriate for identifying mucins.

Identification of novel human mucins with the PTS region-scanning approach

Among the 924 human mucin candidates identified from the RefSeq database, we used the following approach to identify novel human mucins, and also known human proteins the annotation of which has not been suggested to carry mucin domains. Proteins with the limits of not less than 0.4 by this index value were selected from among the candidates and analyzed for annotations. A number of hypothetical proteins were found and omitted from the selected proteins, which resulted in the identification of 17 proteins including five known MUC-type mucins (MUC2, MUC5, MUC6, MUC7 and MUC17). Because mucins have been shown to carry signal peptides in principle, the 12 proteins other than the MUC-type mucins were analyzed for predicted signal peptide and transmembrane domains using the SignalP and TMHMM programs, respectively [15,16]. Signal peptides were found in seven proteins that were designated DKFZP564O0823 (DKFZP564O0823 protein), PODXL (podocalyxin-like precursor), CDSN (corneodesmosin), C11ORF24 (chromosome 11 open reading frame 24), CD68 (CD68 antigen), SPN (sialophorin) and CD164 (CD164 antigen) (Table 3 and Fig. 1). Five of these proteins were found to have one transmembrane domain (Fig. 1), suggesting that these are plasma-membrane bound proteins. Further analysis using the NetOGlyc program [14] predicted that all these proteins contained a number of *O*-linked glycosylation sites (Fig. 1). Careful investigation of published reports regarding characterization of these seven proteins revealed that PODXL [18], CD68 [19], SPN [20] and CD164 [21] were demonstrated to carry a large number of *O*-linked glycans, supporting that this approach is appropriate for

Table 3 Mucins identified from human RefSeq database by either the PTS region-scanning approach or the PTS repeat-searching approach

Gene name	GenBank: accession no.	Gene product	Length (amino acids)	PTS index[a]	REPEAT index[a]
DKFZP564O0823	NP_056208.2	DKFZP564O0823 protein	310	0.574	(–)[b]
PODXL	NP_005388.2	podocalyxin-like precursor	558	0.562	(–)[b]
CDSN	NP_001255.3	corneodesmosin precursor	529	0.495	0.006
C11ORF24	NP_065693.2	chromosome 11 open reading frame 24	449	0.450	(–)[b]
CD68	NP_001242.2	CD68 antigen isoform A	354	0.438	0.017
SPN	NP_001025459.1	sialophorin	400	0.430	(–)[b]
CD164	NP_006007.2	CD164 antigen, sialomucin	197	0.411	(–)[b]
PRG4	NP_005798.2	proteoglycan 4	1404	(–)[b]	0.232
HAVCR1	NP_036338.1	hepatitis A virus cellular receptor 1	359	0.198	0.206
TNFRSF10C	NP_003832.2	tumor necrosis factor receptor superfamily, member 10c precursor; decoy receptor 1	259	(–)[b]	0.174
C6ORF205	NP_001010909.1	hypothetical protein LOC394263 (chromosome 6 open reading frame 205)	566	(–)[b]	0.070
GP1BA	NP_000164.3	platelet glycoprotein Ib alpha polypeptide precursor	626	0.145	0.067
DMBT1	NP_015568.1	deleted in malignant brain tumors 1 isoform b precursor	2413	(–)[b]	0.053
LPA	NP_005568	lipoprotein, Lp(a)	2040	(–)[b]	0.041

[a] Refer to the text for descriptions of the PTS index and REPEAT index.

[b] (–) The index could not be calculated because the gene was not selected from the human RefSeq database.

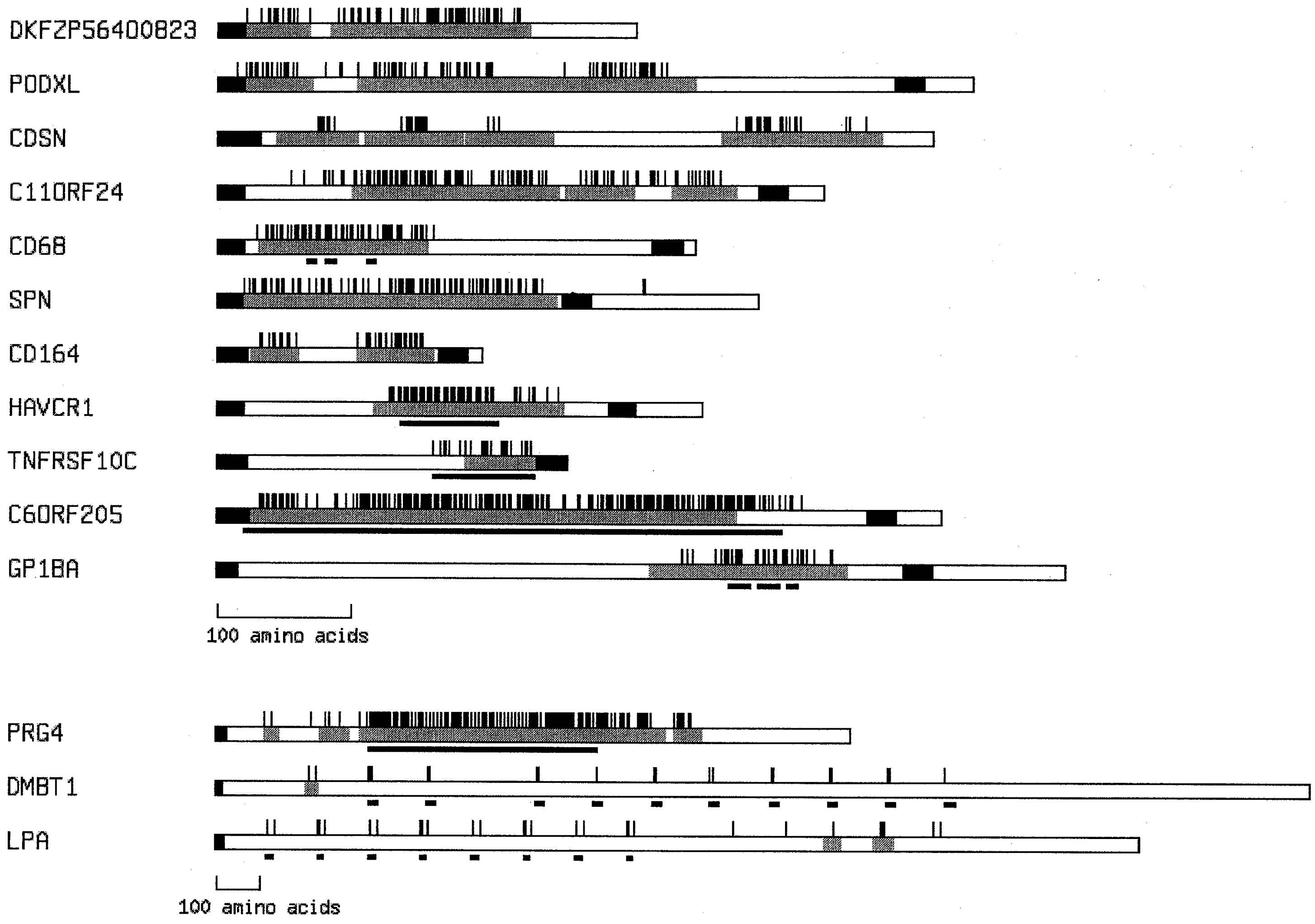

Fig. 1 Structures of mucin proteins selected from human RefSeq database using two different bioinformatics approaches.
Fourteen candidate mucins identified from the human RefSeq database are shown. A filled area indicates a signal peptide or transmembrane domain(s). Shaded areas represent PTS regions predicted by our approach as described in Table 1. Vertical lines represent *O*-glycosylation sites predicted by the NetOGlyc program. Thick bars beneath the protein indicate tandem repeat sequences.

identification of mucins. Thus, DKFZP564O0823, CDSN and C11ORF24 also are likely novel mucin-like proteins, although there is no biochemical evidence for the presence of a large number of *O*-linked glycans on these proteins. Because the functions of DKFZP564O082 and C11ORF24 are not known, we performed BLAST searches [11] and Pfam searches [13] for protein homology and motifs, respectively. However, there were no significant homologies or protein motifs including the motifs frequently detected on MUC-type mucins i.e., vWF, SEA, and EGF-like domains. Moreover, Pfam searches showed that CDSN carried none of vWF, SEA, and EGF-like domains.

Development of a bioinformatics approach based on PTS-rich repeats to identify mucins

Some non-MUC-type mucins, such as P-selectin ligand and mucosal vascular addressin cell adhesion molecule 1, have been shown to carry *O*-linked glycans on tandem repeats of short amino acid stretches containing Pro and Thr/Ser [9, 10]. The PTS region-scanning approach did not identify either of these non-MUC-type mucins, and this led us to develop another bioinformatics approach, which we call the PTS repeats-searching approach, to identify mucins on the basis of these repeated sequences. In this approach, the presence of at least three repeats of a short sequence containing Pro and Thr/Ser was considered a criterion for mucins. As described in the Materials and methods section, we tested the selection conditions both by varying the length of the repeat unit between six and nine amino acids and by altering the Pro and Ser/Thr contents for 15 human MUC-type mucins in the RefSeq database. The results of this test allowed us to select optimal repeat unit length (6 residues) and Pro and Ser/Thr contents ($P \geq 1, T + S \geq 3$ or $P \geq 2, T + S \geq 2$) for identifying mucin candidates (Table 4). Searches for human mucins in the RefSeq database using these optimal conditions led to the identification of 955 candidate mucins. To evaluate these, we created a REPEAT index, which is the frequency with which PTS-rich repeats are detected in the entire protein length. When we used this index to list the selected mucin candidates in decreasing order with the limits of not less than 0.006 by this index value, 13 out of 15 known MUC-type mucins in the RefSeq database were found among the mucin candidates. This result confirmed the usefulness of this approach for identifying mucins *in silico.*

Identification of novel human mucins using the PTS repeats-searching approach

To identify human proteins that have not been described as mucins, proteins with a REPEAT index of at least

Table 4 Bioinformatics approach to identifying mucins based on PTS-rich repeats

Characteristics of a repeat unit		Unselected MUC-type mucins	Number of mucin candidates identified from the human RefSeq database
Length (residues)	Number of Pro,Thr and Ser residues[a]		
6	$P \geq 1, T+S \geq 2$	none	4839
7	$P \geq 1, T+S \geq 2$	none	5574
8	$P \geq 1, T+S \geq 2$	none	6104
6	$P \geq 1, T+S \geq 3$ or $P \geq 2, T+S \geq 2$	MUC12, MUC13 and MUC15	714
7	$P \geq 1, T+S \geq 3$ or $P \geq 2, T+S \geq 2$	MUC13 and MUC15	971
8	$P \geq 1, T+S \geq 3$ or $P \geq 2, T+S \geq 2$	MUC15	1220
9	$P \geq 1, T+S \geq 3$ or $P \geq 2, T+S \geq 2$	MUC15	1519
9	$P \geq 1, T+S \geq 3$ or $P \geq 2, T+S \geq 2$	MUC12, MUC13 and MUC15	248

[a] P, T and S refer to the number of Pro, Ser and Thr residues, respectively.

0.05 were selected from among the 955 proteins and analyzed for annotations. A number of hypothetical proteins were found and omitted from the selected proteins, which resulted in the identification of 28 proteins including seven known mucins (MUC2, MUC5, MUC7, MUC11, MUC12, MUC17, and mucosal vascular addressin cell adhesion molecule 1). The 21 proteins except known mucins were analyzed for predicted signal peptide and transmembrane domains using the SignalP and TMHMM programs, respectively [15,16]. Signal peptides were predicted in seven proteins that were designated PRG4 (proteoglycan 4), HAVCR1 (hepatitis A virus cellular receptor 1), TNFRSF10C (tumor necrosis factor receptor superfamily, member 10c), C6ORF205 (chromosome 6 open reading frame 205), GP1BA (platelet glycoprotein Ib alpha polypeptide), DMBT1 (deleted in malignant brain tumors 1) and LPA (lipoprotein, Lp(a)) (Table 3 and Fig. 1). Four of these proteins were predicted to have one transmembrane domain (Fig. 1), suggesting that these are plasma-membrane bound proteins. The NetOGlyc program [14] predicted that all these proteins contained potential *O*-linked glycosylation sites (Fig. 1). Investigation of published reports regarding characterization of these seven proteins revealed that PRG4 [22-23], HAVCR1 [24], GP1BA [25], DMBT1 [26] and LPA [27] were shown to carry a large number of *O*-linked glycans, demonstrating that this approach is appropriate for identification of mucins. Thus, TNFRSF10C and C6ORF205 also are likely novel mucin-like proteins, although there is no biochemical evidence for the presence of a large number of *O*-linked glycans on these proteins. BLAST [11] and Pfam [13] analyzes of C6ORF205 detected no significant homologies and protein motifs. Furthermore, Pfam search of TNFRSF10C showed that both carried none of vWF, SEA, and EGF-like domains.

Gene expression analysis of novel human mucin genes using oligonucleotide microarrays

To characterize the selected novel mucins further, we performed gene expression analysis using Affymetrix GeneChip oligonucleotide microarrays and cRNA probes from 30 normal human tissues shown in Fig. 2. In addition to the normal tissues, the following 16 normal culture cells and 23 carcinoma cell lines were analyzed: human umbilical vein endothelial cell (HUVEC); small airway epithelial cell (SAEC); bronchial smooth muscle cell (BSMC); coronary artery smooth muscle cell (CASMC); human mesangial cell (NHMC); umbilical artery smooth muscle cell (UASMC); uterine smooth muscle cell (UtSMC); skeletal muscle cell (SkMC); human coronary artery endothelial cell (HCAEC); human dermal microvascular endothelial cell (HMVEC); human bronchial/tracheal epithelial cell (NHBE); human adult epidermal keratinocyte (NHEK-Ad); human neonate

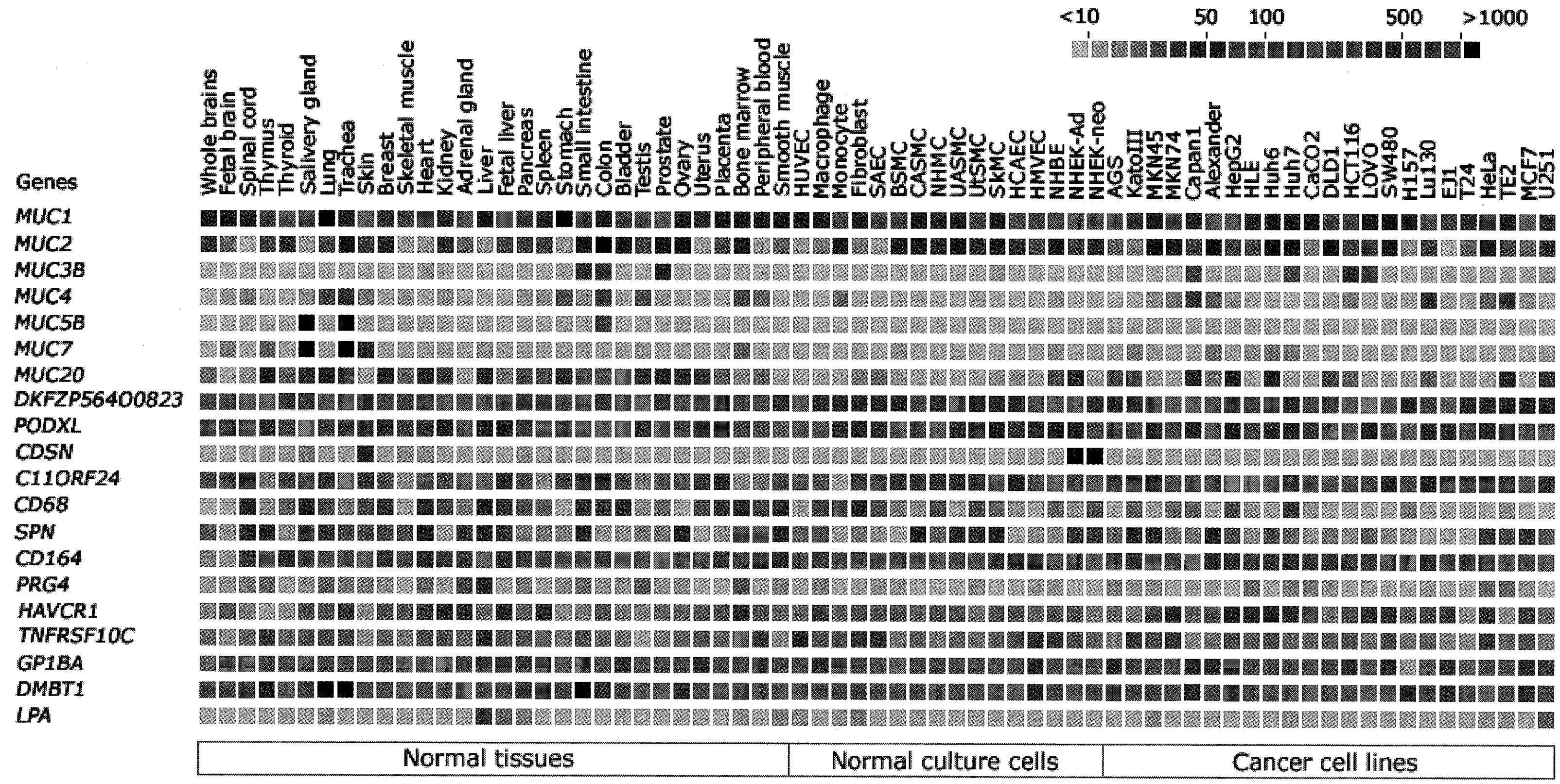

Fig. 2 Oligonucleotide microarray analysis of mucin genes selected from human RefSeq database using two different bioinformatics approaches.
The expression level is indicated by box color and is based on normalized average differences. The color changes from red to light blue in proportion to normalized average differences in log scale. Abbreviations of normal culture cells and cancer cell lines refer to the text.

epidermal keratinocyte (NHEK-neo); gastric carcinoma cell lines AGS, KatoIII, MKN45 and MKN74; a pancreatic carcinoma cell line Capan1; liver carcinoma cell lines Alexander, HepG2, HLE, Huh6 and Huh7; colon carcinoma cell lines CaCO2, DLD1, HCT116, LOVO and SW480; lung carcinoma cell lines H157 and Lu130; bladder carcinoma cell lines EJ1 and T24; a cervical carcinoma cell line HeLa; an esophageal carcinoma cell line TE2; a breast carcinoma cell line MCF7; and a brain carcinoma cell line U251 (Fig. 2). Oligonucleotides for four (*DKFZP564O0823*, *CDSN*, *C11ORF24*, and *TNFRSF10C*) of the five novel mucin genes were found on the Affymetrix U133 GeneChip, and their hybridization signals were analyzed along with the signals of seven MUC-type mucin genes (*MUC1*, *MUC2*, *MUC3B*, *MUC4*, *MUC5B*, *MUC7*and *MUC20*), and nine other known mucin-like genes (*PODXL*, *CD68*, *SPN*, *CD164*, *PRG4*, *HAVCR1*, *GP1BA*, *DMBT1* and *LPA*). *DKFZP564O0823* was expressed in many tissues but at high levels in thyroid, salivary gland, prostate and colon (Fig. 2). *CDSN* was highly expressed in skin and epidermal keratinocyte cells (NHEK-Ad and NHEK-neo). *C11ORF24* was expressed in most normal culture cells and several cancer cell lines, and the *TNFRSF10C* was expressed in bone marrow, peripheral blood and human coronary artery endothelial cells (HCAEC) (Fig. 2).

Identification of human mucin genes differentially expressed in cancer cell lines by gene expression analysis

To determine whether any of the candidate mucins showed altered expression in cancer cells, we analyzed the gene expression profiles of the above-selected mucins by the Affymetrix oligonucleotide microarray method. Novel and known mucin genes of which the oligonucleotides were present on the Affymetrix GeneChip were selected from among the genes identified by either the PTS region-scanning approach or the PTS repeats-searching approach. Comparison of expression between normal tissues/normal culture cells and cancer cell lines revealed that the expression levels of MUC20, HAVCR1 and PODXL genes increased in some cancer cell lines (Table 3, Figs 1 and 2). Mucin 20 was recently reported as a novel MUC-type mucin that is upregulated in renal tissues in nephropathy and other renal injuries [5]. *MUC20* was expressed at relatively high levels in the AGS (gastric carcinoma), Capan1 (pancreatic carcinoma), DLD1 (colon carcinoma) and TE2 (breast carcinoma) cell lines. *HAVCR1* was very recently reported to be over-expressed in renal carcinoma tissues [28], was expressed prominently in the Huh6 and Huh7 (liver carcinoma), CaCO2 (colon carcinoma) and

EJ1 (bladder carcinoma) cell lines. *PODXL*, a member of the sialomucin family [18] the expression of which is associated with breast cancer progression [29, 30], was expressed abundantly in the GT3 (gastric carcinoma), Huh6 (liver carcinoma), CaCO2 and LOVO (colon carcinoma) and HeLa (cervical carcinoma) cell lines.

Discussion

In the present study, we developed two different approaches for identifying mucins based on the characteristics of mucin domains. We found that both approaches can identify mucins efficiently from the human RefSeq protein database. The PTS region-scanning approach is similar to the method described by Lang et al. [17] but differs in two respects. As shown in Table 1 and Fig. 1, we most often observed relatively short mucin domains (40–50 amino acids long); therefore, the window size of 100 amino acids used by Lang et al. [17] seemed too large for selecting many mucin molecules. We found that the specificity, i.e. [true positives/(true positives + false positives)], was highest when a window size of 55 amino acids was used for identifying human MUC-type mucins (Table 2). We also analyzed the contents of various amino acids in addition to those of Pro, Ser and Thr in the PTS regions and found that relative paucities of Cys and Met residues, charged amino acids and hydrophobic amino acids were characteristic of these regions. Lang et al. [17] focused only on the abundance of Pro, Ser and Thr, whereas we included these additional characteristics in the selection settings; this improved the specificity of mucin selection. Hydrophobic amino acids are often found in transmembrane regions and inside protein folds, whereas PTS regions should be located on the outer surfaces of proteins. This may explain the low numbers of hydrophobic amino acids in PTS regions. Most Cys residues are thought to be involved in the formation of S-S bridges, which might hamper *O*-glycosylation. This may explain why Cys residues are relatively rare in PTS regions. An abundance of charged amino acids may also interfere with efficient *O*-glycosylation; hence there are fewer charged amino acids in PTS regions.

We also designed a PTS repeats-searching approach. Most of the known MUC-type mucins were identified by at least one of the two approaches, but each approach also identified different novel mucins in the human RefSeq database (Table 3). Searches for repeated sequences revealed that all the novel mucins identified by the PTS repeats-searching approach contained PTS-rich repeats with a number of potential *O*-glycosylation sites (Fig. 1). However, among the novel mucins selected by the PTS region-scanning approach, only CDSN was identified with either PTS region-scanning or PTS repeats-searching approaches. Furthermore, the amino acid contents of the putative mucin domains for the novel mucin-like proteins identified by PTS repeats searching were quite different from those of the PTS regions of MUC-type mucins (data not shown). It appears that two different classes of mucins are selected preferentially by these approaches.

Both approaches select human MUC-type mucins effectively. The 924 mucin candidates identified by PTS region scanning were subjected to PTS repeats searching, resulting in the selection of 346 mucin candidates, including all the known MUC-type mucins in the RefSeq database. Thus, simultaneous use of these approaches increased the specificity for selection of human MUC-type mucins, suggesting that combined use of these screens would be suitable for identifying MUC-type mucins.

Two bioinformatics approaches developed in this study were shown to identify mucins efficiently. Furthermore, the oligonucleotide microarray method allowed us to identify mucins that were differentially expressed in normal tissues/cells and cancer cell lines. Further study of the selected mucins in cancer tissues will facilitate direct evaluation of their potential for cancer diagnosis and therapy.

Acknowledgements

We thank Hiroko Meguro for technical assistance.

Abbrebiations: Refseq, reference sequence; vWF, von Willebrand-factor; SEA, sea urchin sperm protein enterokinase agrin domains; EGF, epidermal growth-factor.

References

1. S. J. Gendler, A. P. Spicer, *Annu. Rev. Physiol.* **1995**, 57, 607–634
2. J. Dekker, J. W. Rossen, H. A. Buller, A. W. Einerhand, *Trends Biochem. Sci.* **2002**, 27, 126–131
3. J. R. Jr. Gum, S. C. Crawley, J. W. Hicks, D. E. Szymkowski, Y. S. Kim, *Biochem. Biophys. Res. Commun.* **2002**, 291, 466–475
4. Y. Chen, Y. H. Zhao, T. B. Kalaslavadi, E. Hamati, K. Nehrke, A. D. Le, D.K. Ann, R. Wu, *Am. J. Respir. Cell Mol. Biol.* **2004**, 30, 155–165

5. T. Higuchi, T. Orita, S. Nakanishi, K. Katsuya, H. Watanabe, Y. Yamasaki, I. Waga, T. Nanayama, Y. Yamamoto, W. Munger, H. W. Sun, R. J. Falk, J. C. Jennette, D. A. Alcorta, H. Li, T. Yamamoto, Y. Saito, M. Nakamura, *J. Biol. Chem.* **2004**, 279, 1968–1979

6. B. W. Yin, K. O. Lloyd, *J. Biol. Chem.* **2001**, 276, 27371–27375

7. A. P. Spicer, G. J. Rowse, T. K. Lidner, S. J. Gendler, *J. Biol. Chem.* **1995**, 270, 30093–30101

8. A. Velcich, W. Yang, J. Heyer, A. Fragale, C. Nicholas, S. Viani, R. Kucherlapati, M. Lipkin, K. Yang, L. Augenlicht, *Science* **2002**, 295, 1726–1729

9. D. Sako, X. J. Chang, K. M. Barone, G. Vachino, H. M. White, G. Shaw, G. M. Veldman, K. M. Bean, T. J. Ahern, B. Furie, D. A. Cumming, G. R. Larsen, *Cell* **1993**, 75, 1179–1186

10. M. J. Briskin, L. M. McEvoy, E. C. Butcher, *Nature* **1993**, 363, 461–464

11. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucleic Acids Res.* **1997**, 25, 3389–3402

12. S. Jiang, T. Tanaka, H. Iwanari, H. Hotta, H. Yamashita, J. Kumakura, Y. Watanabe, Y. Uchiyama, H. Aburatani, T. Hamakubo, T. Kodama, M. Naito, *Nucl. Recept.* **2003**, 1, 5

13. A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, S. R. Eddy, *Nucleic Acids Res.* **2004**, 32, D138–D141

14. J. E. Hansen, O. Lund, N. Tolstrup, A. A. Gooley, K. L. Williams, S. Brunak, *Glycoconj. J.* **1998**, 15, 115–130

15. J. D. Bendtsen, H. Nielsen, G. von Heijne, S. Brunak, *J. Mol. Biol.* **2004**, 340, 783–795

16. E. L. Sonnhammer, G. von Heijne, A. Krogh, *Proc. Int. Conf. Intell. Sys. Mol. Biol.* **1998**, 6, 175–182

17. T. Lang, M. Alexandersson, G. C. Hansson, T. Samuelsson, *Glycobiology* **2004**, 14, 521–527

18. D. B. Kershaw, S. G. Beck, B. L. Wharram, J. E. Wiggins, M. Goyal, P. E. Thomas, R.C. Wiggins, *J. Biol. Chem.* **1997**, 272, 15708–15714

19. C.L. Holness, D.L. Simmons, *Blood* **1993**, 81, 1607–1613

20. E. Remold-O'Donnell, D.M. Kenney, R. Parkman, L. Cairns, B. Savage, F. S. Rosen, *J. Exp. Med.* **1984**, 159, 1705–1723

21. Y. Masuzawa, T. Miyauchi, M. Hamanoue, S. Ando, J. Yoshida, S. Takao, H. Shimazu, M. Adachi, T.

Muramatsu, *J. Biochem.* **1992**, 112, 609–615

22. B. L Schumacher, J. A. Block, T. M.Schmid, M. B. Aydelotte, K. E. Kuettner, *Arch. Biochem. Biophys.* **1994**, 311, 144-152
23. C. R. Flannery, C. E. Hughes, B.L. Schumacher, D. Tudor, M. B. Aydelotte, K. E. Kuettner, B. Caterson, *Biochem. Biophys. Res. Commun.* **1999**, 254, 535–541
24. D. Feigelstock, P. Thompson, P. Mattoo, Y. Zhang, G.G. Kaplan, *J. Virol.* **1998**, 72, 6621–6628
25. J. A. Lopez, D. W. Chung, K. Fujikawa, F.S. Hagen, T. Papayannopoulou, G. J. Roth, *Proc Natl. Acad. Sci. U S A.* **1987**, 84, 5615–5619
26. R. C. De Lisle, M. Petitt, K. S. Isom, D. Ziemer, *Am. J. Physiol.* **1998**, 275, G219–227
27. B. Garner, A. H. Merry, L. Royle, D. J. Harvey, P. M. Rudd, J. Thillet, *J. Biol. Chem.* **2001**, 276, 22200–22208
28. M. R. Vila, G. G. Kaplan, D. Feigelstock, M. Nadal, J. Morote, R. Porta, J. Bellmunt, A. Meseguer, *Kidney Int.* **2004**, 65, 1761–1773
29. W. M. Schopperle, D. B. Kershaw, W. C. DeWolf, *Biochem. Biophys. Res. Commun.* **2003**, 300, 285–290
30. A. Somasiri, J. S. Nielsen, N. Makretsov, M. L. McCoy, L. Prentice, C. B. Gilks, S. K. Chia, K. A. Gelmon, D. B. Kershaw, D. G. Huntsman, K. M. McNagny, C. D. Roskelley, *Cancer Res.* **2004**, 64, 5068–5073

Chikako Nishi-Takaoka [1,2], Tatsunari Nishi [1,2], Takahiro Shimamura[2], Shogo Yamamoto[2], Yoshitaka Hippo[2], Hiroyuki Aburatani[2]

[1]Genaris, Inc., JRC-106, 1-1-40 Suehirocho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

[2]Genome Science Division, Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo 153-8904, Japan

E-mail address: takaoka@genaris.co.jp