Solution Structure of the SEA Domain from the Murine Homologue of Ovarian Cancer Antigen CA125 (MUC16)*

Human CA125, encoded by the MUC16 gene, is an ovarian cancer antigen widely used for a serum assay. Its extracellular region consists of tandem repeats of SEA domains. In this study we determined the three-dimensional structure of the SEA domain from the murine MUC16 homologue using multidimensional NMR spectroscopy. The domain forms a unique α/β sandwich fold composed of two α helices and four antiparallel β strands and has a characteristic turn named the TY-turn between α1 and α2. The internal mobility of the main chain is low throughout the domain. The residues that form the hydrophobic core and the TY-turn are fully conserved in all SEA domain sequences, indicating that the fold is common in the family. Interestingly, no other residues are conserved throughout the family. Thus, the sequence alignment of the SEA domain family was refined on the basis of the three-dimensional structure, which allowed us to classify the SEA domains into several subfamilies. The residues on the surface differ between these subfamilies, suggesting that each subfamily has a different function. In the MUC16 SEA domains, the conserved surface residues, Asn-10, Thr-12, Arg-63, Asp-75, Asp-112, Ser-115, and Phe-117, are clustered on the β sheet surface, which may be functionally important. The putative epitope (residues 58–77) for anti-MUC16 antibodies is located around the β2 and β3 strands. On the other hand the tissue tumor marker MUC1 has a SEA domain believed to another subfamily, and its GSVVV motif for proteolytic cleavage is located in the short loop connecting β2 and β3.

CA125 is a serum marker that is widely used to monitor ovarian cancer because it is overexpressed in ovarian cancer cells and secreted into the blood. An elevated serum CA125 level is a useful indicator of ovarian cancer, but it is also observed in a number of benign conditions (1, 2). CA125 is a mucin-type O-linked glycoprotein (3, 4), but other details about its molecular nature remain unclear. Recently two research groups cloned CA125 (5–8), revealing that CA125 is a membrane protein with some splicing variants. The splicing variants have the same intracellular and transmembrane regions. The extracellular domain consists of the SEA1 domains, which are repeated 7, 12, or 60 times, according to the variant. The gene was named MUC16, after the mucin-like nature of CA125. The elucidation of the amino acid sequence has made it possible to specify the approximate position of the epitope. A previous study showed that the peptide epitope position of CA125 is located between two conserved cysteines in the SEA domain (7).

A cDNA of the murine MUC16 homologue, cloned in the RIKEN FANTOM project (9), has a total of 258 amino acids and a transmembrane domain. It is 66% identical to the C terminus of human MUC16 and has only one SEA domain in its extracellular region. However, our investigation of the mouse and human genomic sequences showed that they share the same characteristic repeat structure of MUC16. Thus, the murine MUC16 appears to have splicing variants, as in the case of the human form. The SEA domain of the clone by the RIKEN FANTOM project (9) corresponds to that most membrane-proximal of the repeat SEA domains.

The SEA domain, or SEA module, was first identified as a module commonly found in sea urchin sperm protein, enterokinase, and agrin (10). Most of the proteins with SEA domains are membrane proteins and usually have only one SEA domain, unlike MUC16. The SEA domains always exist in the extracellular region and are accompanied by an O-linked glycosylation chain at the N-terminal side. SEA domains consist of about 120 residues, of which about 80 residues are highly conserved. Previous studies have speculated about the properties of SEA domains. Perlecan is a type of proteoglycan that has both the SEA domain and glycosaminoglycan (11). The deletion of the SEA domain of perlecan affects heparan sulfate synthesis (11, 12).

* This work was supported by the RIKEN Structural Genomics/Proteomics Initiative, the National Project on Protein Structural and Functional Analysis, Ministry of Education, Culture, Sports, Science, and Technology of Japan. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The atomic coordinates and structure factors (code 1HVZ) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/).

To whom correspondence should be addressed: RIKEN Genomic Sciences Center 1-7-22 Suehiro-cho, Tsurumi, Yokohama 220-0045, Japan, Tel.: 81-45-503-9196; Fax: 81-45-503-9195; E-mail: yokoyama@biochem.s.u-tokyo.ac.jp.

1 The abbreviations used are: SEA, a module first identified in sea urchin sperm protein, enterokinase, and agrin; ER, endoplasmic reticulum; GSV, correlation spectroscopy; NOESY, nuclear Overhauser effect spectroscopy; NOE, nuclear Overhauser effect; TOCSY, total correlation spectroscopy; MUC, mucin; PDB, Protein Data Bank; r.m.s.d., root mean square deviation.
Recently it was reported that the membrane protein mucin 1 (MUC1), a tumor tissue marker, undergoes posttranslational proteolytic cleavage in the ER in a specific site of the SEA domain within its membrane proximal region (13–15). The site-specific cleavage of SEA domains has also been reported for Ig-hepta, seven-transmembrane proteins, and mucin 3 (MUC3) (16, 17). All of these cleavages occur between glycine and serine in the GSVVV motif, located in the middle of the SEA domains (14, 16, 17). This motif is not necessarily conserved in all SEA domains. MUC1 and its relatives have the GSVVV motif, but neither human nor murine MUC16 has the GSVVV motif.

In this study we determined the three-dimensional structure of the SEA domain from the murine MUC16 homologue using NMR solution structure of the SEA domain. A, stereo view showing the best fit superposition of the backbone atoms of the well structured regions (residues 5–65 and 71–119) in the 20 lowest energy structures. B, ribbon representation of the lowest energy structure. Helices are colored orange, and strands are colored green. Secondary structure elements and chain termini are labeled. C, the $^{1}H$-$^{15}N$ NOE and $T_{1}/T_{2}$ values. Secondary structure elements of the SEA domain are also indicated. Smaller values indicate faster motion of the protein backbone.
multi-dimensional NMR spectroscopy. The structure determination revealed a novel fold of the SEA domain. This three-dimensional structure, the first determined for a SEA domain, allowed us to identify the conserved residues clustered on the surface, which are probably important for the function. We have classified the SEA domains into several subfamilies on the basis of a tertiary structure-based sequence analysis. This research provides a basis for future studies not only on CA125 but also on other SEA domains.

EXPERIMENTAL PROCEDURES

Protein Expression and Purification—The DNA encoding the murine MUC16 SEA domain, which corresponds to the most membrane proximal one of the repeated SEA domains in human MUC16, was cloned by PCR from the mouse full-length cDNA clone with the ID RIKEN cDNA 1110008I14 (9). This clone encodes a total of 258 amino acid residues. The SEA domain (residues 67–185; total 119 amino acids) was cloned into the expression vector pCR2.1 (Invitrogen) as a fusion with an N-terminal His6 affinity tag and a tobacco etch virus (TEV) protease cleavage site. The fusion protein was synthesized by a cell-free protein expression system (18). To label the protein 13C/15N-labeled amino acids were used. The protein was first adsorbed to a nickel nitriilotriacetic acid affinity column that was washed by a concentration gradient of buffer A (20 mM sodium phosphate buffer (pH 7.2) containing 0.3 M sodium chloride and 10 mM imidazole) and buffer B (20 mM sodium phosphate buffer (pH 7.2) containing 0.3 M sodium chloride and 500 mM imidazole). The protein was eluted by a concentration gradient of buffer C (20 mM sodium phosphate buffer (pH 7.2) containing 0.3 M sodium chloride and 500 mM imidazole). The His tag was then removed by incubation at 30 °C for 1 h with the tobacco etch virus (TEV) protease. The imidazole and sodium phosphate were removed by an overnight dialysis at 4 °C against 20 mM sodium phosphate buffer (pH 7.0) containing 1 mM dithiothreitol. Finally, the protein was purified on a HiTrap SP cation exchange column by elution with a concentration gradient of 0–1 M sodium chloride. The purified protein was concentrated to ~1.5 mg in 1H2O/3H2O (9:1) or 1H2O 20 mM sodium phosphate buffer (pH 6.0) containing 100 mM NaCl, 1 mM dithiothreitol and 0.02% NaN3.

NMR Spectroscopy—All spectra were recorded on Bruker Avance 600 and 800 MHz spectrometers at 298 K. Spectra were processed using the program NMRPipe (19), and analyses of the processed data were performed with the program NMRView (20). Sequential assignments of the 1H, 13N, and 15N resonances were achieved by means of the through-bond heteronuclear scalar correlations along the protein backbone and side chains (21). Backbone resonances were assigned from three-dimensional HNCA, CBCA(CO)NH, HNCC, and HN(CA)CO spectra (all spectra were measured on a 600-MHz NMR spectrometer, unless otherwise noted), whereas aliphatic side chain resonances were assigned from three-dimensional H(CC)O/NH, C(CC)O/NH, and HC/CH TOCSY (21). Aromatic side chain resonances were assigned from two-dimensional HC/CH/TOCSY, DQF-COSY, TOCSY, and NOESY (22) (the latter three spectra were measured with a non-labeled protein sample in 1H2O buffer on an 800-MHz NMR spectrometer). Interproton distance restraints were obtained from three-dimensional 15N- and 13C-edited NOESY recorded with a mixing time of 100 ms. Hydrogen bond information was obtained from transverse relaxation optimized spectroscopy HNCO (23, 24) and three-dimensional NOESY spectra. We used the 13C-labeled protein sample to measure the spectra for a backbone mobility analysis. The pulse sequences used to measure 15N-(1H) NOE, and T1/T2 were those described by Farw et al. (25), with minor modifications. A total of 10 data sets were collected to measure T1, with delay values of 4.02, 44.22, 124.62, 245.22, 406.02, 607.02, 808.02, 1099.02, 1441.02, and 1813.02 ms. A total of 8 data sets were collected to measure T2, with delay values of 3.34, 23.38, 50.1, 83.5, 126.66, 217.1, and 270.54 ms. The 15N/(1H) T1Y values were determined from the ratio of the intensity of two spectra recorded with and without saturation of protons.

Constraint Generation and Structure Calculation—Interproton distance restraints were assigned from the three-dimensional NOE spectra using the program ARIA1.0 (26). The constraints on the backbone dihedral angles were obtained by using TALOS (27) and analyzing NOE data. The initial constraints and the hydrogen bond information were used to calculate the initial structure on the program CNS1.0 (28), in which the structure calculation was performed with molecular dynamics and energy minimization by simulated annealing. On the basis of the 10 lowest energy structures of the 200 calculated structures, the side chain dihedral angle constraints were obtained by an analysis of the NOE spectra. These steps were repeated several times, whereas the interproton distance constraints were modified.

The final structure was calculated from these sets of constraints on CNS1.0. The 20 lowest energy structures out of the 290 calculated structures were deposited in the Protein Data Bank (code 1IVZ) as a final structure. The program PROCHECK-NMR (29) was used to evaluate the structures. All three-dimensional structures were depicted by the program MOLMOL (30). ClustalX 1.81 (31) was used for the sequence alignment, the dendrogram calculation, and their illustrations. A structure-based sequence alignment was performed by modifying the gap penalties for the secondary structure regions (four times larger penalties within the secondary structures and two times larger penalties at the boundaries). The dendrogram of all SEA family members was obtained from the structure-based sequence alignment.

RESULTS

Structure Determination—The boundary of the SEA domain was determined by the sequence comparison. The well dispersed signals in the 1H, 15N heteronuclear single quantum coherence spectra showed that the SEA domain is a well folded structural domain. The protein sample used for the structure determination has linker sequences GSGSGSG and PSSG, both derived from the vector, at the N and the C termini, respectively. These linker peptides were completely unstructured.

The SEA domain structure was determined from 2270 NOE-derived interproton distance restraints, 178 dihedral angle restraints, and 78 hydrogen bond restraints. Residues 1–4, after the N-terminal linker, and residues 66–69 were not well defined due to a lack of NOEs. Except for these residues the structure is well defined (see Fig. 1A). The root mean square deviations of residues 5–65 and 72–119 are 0.42 Å for the backbone heavy atoms and 0.92 Å for all non-hydrogen atoms. The structural statistics of the SEA domain are shown in Table I.

TABLE I

| Structural statistics for the 10 lowest energy structures | Experimental restraint | NOE restraints (total) | 2277 |
|--------------------------------------------------------|------------------------|----------------------|------|
| Intraresidue (i–j = 0)                                  |                        | 933                  |
| Sequential (i–j = 1)                                    |                        | 476                  |
| Medium range (i–j ≤ 5)                                  |                        | 324                  |
| Long range (i–j > 5)                                    |                        | 544                  |
| Torsion angle restraints                                |                        |                      |
| Backbone                                               | 0.450                  |
| Non-H                                                  | 0.920                  |
| r.m.s.d. from experimental constraints                 |                        |                      |
| Distance restraints                                     | 0.015 ± 0.05           |
| Torsion angle restraints                                | 0.0265 ± 0.0155        |
| r.m.s.d. from covalent geometry                         |                        |                      |
| Bonds                                                  | 0.0006 ± 0.0008        |
| Angles                                                  | 0.230 ± 0.0002         |
| Improper torsions                                       | 0.093 ± 0.0006         |

Structural Features—The three-dimensional structure of the SEA domain has an α/β sandwich fold, with the N and C termini on the same side of the molecule (see Fig. 1B). The α/β sandwich fold can be divided into two layers (Fig. 1A). One layer consists of four-stranded antiparallel β sheets (β1, 6–12; β2, 54–63; β3, 73–80; β6, 117–119) and a short α helix (α1, 19–24), whereas the other layer consists of two long α helices (α2, 27–48; α3, 88–100) and two-stranded short β sheets (β4, 105–106; β5, 109–110). There is a bulge in the β2 strand. The sequence patterns of the hydrophobic residues around the β2 strand can be aligned without gaps in all SEA domains, and thus, the bulge is probably conserved throughout the family, as seen in SH3 domains (32). Two cysteine residues (Cys-57 and Cys-78) are located next to each other on the β2 and β3 strands,
Homologous structures found by DALI search. A, proteins with tertiary structures homologous to the SEA domain are shown in ribbon diagrams. Phe-tRNA synthase (PDB code 1pys-B; Z score, 6.3; r.m.s.d., 2.9; residues 682–785), hydroxymethylglutaryl-CoA reductase fragment (PDB code 1qax-A; Z score, 6.0; r.m.s.d., 4.1; residues 111–220), pseudouridine synthase (PDB code 1dj0-A; Z score, 5.7; r.m.s.d., 3.1; residues 1–100), and elongation factor 1β (PDB code 1b64; Z score, 5.7; r.m.s.d., 3.3; full-length). All structures are viewed in the same orientation as the SEA domain in Fig. 1B, electric representation of the SEA domain and the RNA-binding proteins, with blue and red surfaces representing positive and negative potentials, respectively. The structures on the right are viewed from the same direction as in A, and those on the left are viewed from the 180° opposite direction. i, the SEA domain of murine MUC16; ii, phenylalanyl-tRNA synthetase; iii, elongation factor 1β; iv, pseudouridine synthase.

respectively, and are close enough to form a disulfide bond. However, the mass spectroscopy analysis of the trypsin-digested peptide showed that no disulfide bond is formed (data not shown). The two long α helices, α2 and α3, fit into the four-stranded β sheets to form a hydrophobic core. The α2 helix is so long that its N terminus protrudes from the β sheets. Between α1 and α2 there is a characteristic tight turn, which we named the TY-turn, because it contains well conserved threonine and tyrosine residues. We will discuss this region in more detail below. The overall structure is a βαβαβ fold, except for the insertions of small secondary structures.

The residues that participate in core formation are Phe-7, Leu-9, Phe-11, Ile-13, Leu-16, Tyr-18, Ile-22, Pro-25, Tyr-30, Thr-33, Ile-37, Leu-41, Leu-44, Phe-45, Ile-50, Phe-54, Cys-57, Val-59, Phe-62, Thr-72, Val-74, Ser-76, Cys-78, Phe-80, Val-87, Ile-92, Tyr-93, Phe-96, Leu-98, Phe-100, Leu-106, Leu-107, Phe-109, Leu-111, Val-116, and Val-118. Many of these are hydrophobic residues conserved throughout the family. In particular, Tyr-18, Tyr-30, Phe-45, and Phe-54 are the most conserved. The conservation of hydrophobic residues will be discussed in more detail below (see “Discussion”).

The SEA domain has two short loops (L1, residues 64–72; L2, residues 81–87) but no unstructured region except for residues 66–69. In the hydrophobic core the side chains are well packed, suggesting the rigidity of the SEA domain. The measured 1H–15N NOE and T1/T2 values revealed the low mobility of the main chain, as presented in Fig. 1C. The main chain of the SEA domain exhibits low mobility throughout the entire domain, except for the N- and C-terminal peptides, which are mostly derived from the expression vector. The mobile region is very short; even some loop regions show low mobility. The aromatic residues that contribute to core packing, such as Tyr-18, Tyr-30, Phe-45, and Phe-54, are well conserved among all SEA domains. Phe-54, located on the N terminus of β2, and Phe-45 stack upon each other to bridge distinct elements. In addition the charged residues at the N and C termini of the α helices probably stabilize the helix electric dipole. These residues appear to contribute to the rigidity of the structure.

Structure Comparison and Binding Assays—We searched for structural homologues in the Protein Data Bank with the program DALI (33). The six most similar structures are the anticodon binding domain of phenylalanyl-tRNA synthase (PDB code 1pys-B; Z score, 6.3; r.m.s.d., 2.9), hydroxymethylglutaryl-CoA reductase fragment (PDB code 1qax-A; Z score, 6.0; r.m.s.d., 4.1), pseudouridine synthase (PDB code 1dj0-A; Z score, 5.7; r.m.s.d., 3.1), elongation factor 1β (PDB code 1b64; Z score, 5.7; r.m.s.d., 3.3), hydroxymethylglutaryl-CoA reductase fragment (PDB code, 1qax-S; Z score, 5.5; r.m.s.d., 3.1), and ribosomal protein S6 (PDB code 1qax-A; Z score, 5.5; r.m.s.d., 3.1). Interestingly, many RNA binding domains shared similarities with the SEA domain. The four most similar structures are illustrated in ribbon diagrams in Fig. 2A. We also compared the surface electrostatic potential of the SEA domain with those of the RNA binding domains (see Fig. 2B). The SEA domain is a highly positively charged protein, which suggests that it can bind negatively charged molecules such as nucleic acids or acidic sugars. However, their surface charge patterns differ, and the functional residues of the RNA-binding proteins are not conserved in the SEA domain. The SEA domain is an extracellular domain, and thus, it probably does not share the same function with the RNA-binding proteins.
Fig. 3. Secondary structure-based sequence alignment and dendrogram of the SEA domains. First column, protein name/species abbreviations (SWISS-PROT data base codes are used if available): MUC, mucin; AGRI, agrin; ENTK, entero kinase; PGBM, perlecan; PG10.2, PG10.2 protein; MUA3, transmembrane cell adhesion receptor mua-3; MAT, mouse airway trypsin-like protease; DES1, serine protease DESC1; SP63, 63-kDa sperm protein. The SEA domains are numbered according to the positions from the N terminus. The small numbers represent the N termini of the SEA domain of MUC16 from the long transcript variant with 12 SEA domains. Above the sequences, the secondary structure elements are indicated, and residues that participate in core formation are designated with C. The boxed region indicates the GSVVV motif that is conserved among MUC1 and its family. The dendrogram, generated by ClustalX, is displayed on the right side of the sequence. Subfamilies are indicated by roman numerals.
It was proposed that the SEA domain might have sugar chain binding activity (10). Interestingly, there are chemical and structural similarities between acidic sugar chains and nucleic acids. Thus, we tested the SEA domain for the ability to bind sugar chains by using model sugars. We measured the chemical shift perturbations of the $^{15}$N-labeled SEA domain mixed with candidate compounds by heteronuclear single quantum coherence. The compounds we tested were ATP, sucrose, porcine stomach mucin ($O$-glycosylated glycoprotein; Sigma), and chondroitin sulfate (a type of glycosaminoglycan; EXTRASYNTHESIS S.A.). However, no significant chemical shift perturbation was observed with any of the compounds. We also tried to observe the binding between the SEA domain and sucrose, mucin, and chondroitin sulfate using surface plasmon resonance (34), but no interaction was detected (data not shown). Our SEA domain does not associate with the glyco-chains presented by mucin and chondroitin sulfate under our conditions. Of course, the possibility still exists that it may associate with any of the other compounds under different conditions.

**Structure-based Sequence Analysis and Clustering of SEA Domain Family Members**—We performed a multiple alignment of 40 SEA domain sequences from various proteins using ClustalX (see Fig. 3) with higher gap penalties in the secondary structure regions to make a dendrogram reflecting the three-dimensional structural information. The dendrogram based on the alignment shows that the SEA domain is clustered into several subfamilies, each with common features. Subfamily I consists of the SEA domains of the murine MUC16 homologue, human MUC16, and agrin. Subfamilies II, III, and IV consist of proteins with the GSVVV motif, which undergo proteolytic cleavage. Subfamily IV contains Caenorhabditis elegans proteins that have no GSVVV motif. Subfamily V consists of only two proteins, mouse airway trypsin-like protease and DESC1, which are both serine proteases. Subfamily VI has no significant common features.

**DISCUSSION**

**The SEA Domain Family**—The residues conserved in all SEA domains are illustrated in Fig. 3. The SEA domain of the murine MUC16 homologue has 36 residues that participate in core formation. They are well conserved in all SEA domains, indicating that all of the proteins in the family share the same topology. In general, the first half (1–75) of the SEA domain sequence is well conserved, but the second half (76–119) is less conserved, although the tertiary structure is indivisible at the border. Thus, the amino acid sequences of the second half vary between subfamilies, and each subfamily is characterized by the sequence of the second half. Besides the core residues, no residues are conserved throughout the SEA domains. The conserved residues on the surface differ between the subfamilies, as discussed below. These facts suggest that each subfamily functions in a different manner; that is, all of these domains may share a common fundamental function such as sugar binding or protein recognition, but the target molecules or binding mechanisms should be different depending on the subfamily.

**TY-turn**—As described above a characteristic tight turn exists between $\alpha_1$ and $\alpha_2$ (see Fig. 4). Because the two conserved residues Thr-27 and Tyr-30 play important roles in maintaining this structure, we named it the TY-turn. The linker region of this turn contains four amino acids in a $\beta_\alpha a\beta$ conformation. Thr-27 is located in the TY-turn (Fig. 4) and is conserved as serine or threonine residues throughout the family. The OH group of Thr-27 protrudes toward the N terminus of $\alpha_2$, which contains the backbone HN group of Tyr-30. These two groups probably form a hydrogen bond for "capping" the helix (35). A hydrogen bond network seems to stabilize the characteristic structure (Fig. 4). Tyr-30 on $\alpha_2$ and Pro-25 on the turn are stacked upon each other, and they seem to support the structure. Tyr-30 is conserved as aromatic residues throughout the family, and Pro-25 is conserved among subfamily I and some SEA domains. Therefore, this characteristic structure is likely to be conserved in the family.

**Conserved Residues on the Murine MUC16 SEA Domain Surface**—As previously described the extracellular region of human MUC16 consists of SEA domain repeats (6). Thus, the question naturally arises as to whether the repeated SEA domains are equivalent. According to the dendrogram in Fig. 3, the SEA domains in MUC16 are apparently more similar to each other than to other SEA domains. This suggests that the multiplication of the SEA domains occurred after MUC16 appeared. Interestingly, the second SEA domain from the C terminus of human MUC16 is now classified into subfamily II, with the GSVVV motif, and correspondingly has a DSVLV sequence. Therefore, this SEA domain might undergo proteolytic cleavage. In the dendrogram of all SEA domains from human MUC16 (Fig. 3), the fifth and further SEA domains from the C terminus are very similar to each other, whereas the four C-terminal, membrane-proximal SEA domains are more diverse. This suggests that the N-terminal SEA domains may have duplicated after the C-terminal SEA domains duplicated. We suppose that MUC16 used to be a rather short protein, with a small number of SEA domains, which became long recently.

We then performed a detailed analysis of the amino acid sequences of the MUC16 SEA domains (Fig. 5A). The murine MUC16 SEA domains and that of the human long splicing variant were included in the comparison, whereas the second SEA domain from the C terminus was excluded because of its peculiarity. The conserved residues presented on the surface are illustrated in Fig. 5B. Most of the conserved residues are concentrated on the $\beta$ sheet surface. The Asn-10, Thr-12, Arg-63, Asp-75, Asp-112, Ser-115, and Phe-117 residues are completely conserved in the MUC16 SEA domains. There is a pocket on the conserved $\beta$ sheet surface that may be suitable for ligand binding.

We further examined the conservation within the subfamily I SEA domains, including agrin. The alignment is illustrated in Fig. 5C, and the surface plot is in Fig. 5D. The surface residue conservation significantly diminished as the comparison range was expanded to the entire subfamily, but several conserved
Fig. 5. Conserved residues in the SEA domain. A, sequence alignment of the murine MUC16 homologue and MUC16 SEA domains. C indicates the residues that participate in core formation, and + indicates the conserved residues. B, solvent accessible surface model of the SEA domain with the residues conserved in MUC16 SEA domains, colored according to their residue types; yellow, hydrophobic; blue, basic; red, acidic; green, Ser or Thr; purple, Asn or Asp. The structure on the right is viewed from the β sheet surface, and that on the left is viewed with a 180° rotation. C, sequence alignment of the SEA domains belonging to subfamily I (see Fig. 3). D, solvent-accessible surface model of the SEA domain, with the residues conserved within the subfamily I SEA domains, colored according to their residue types (the same as in B).
residues still exist on the β sheet surface. This may be a key to discriminate the functionally important residues from the unimportant residues. On the other hand, the surface residues around the TY-turn of the SEA domain are not conserved.

The Epitope Site of CA125—Anti-CA125 antibodies are divided into three groups, OC125-like, M11-like, and Ov197-like, which recognize domains of nonoverlapping epitopes (36). All three types of antibodies can recognize the recombinant CA125 which recognize domains of nonoverlapping epitopes (36). All variants, MUC1/REP, MUC1/SEC, and MUC1/Y (37). Only three types of antibodies can recognize the recombinant CA125, which is a heavily glycosylated protein. The antibodies can recognize either native or denatured CA125. The epitope site of the M11 antibody was further investigated with digested peptide fragments, which revealed that the peptide between two conserved cysteine residues in the SEA domain is sufficient for antibody recognition (17). These cysteines correspond to Cys-57 and Cys-78 of murine MUC16, which are on β2 and β3. This fact suggests that the epitope site is located on the β sheet surface.

The GSVVV Motif as the Proteolytic Cleavage Site—The SEA domains of MUC1, MUC3, and Ig-hepta undergo proteolytic cleavage between the glycine and serine residues of the GSVVV motif, as a posttranslational modification in the ER (14, 16, 17). MUC1, which is a tissue tumor marker, has three splicing variants, MUC1/REP, MUC1/SEC, and MUC1/Y (37). Only MUC1/REP is proteolytically cleaved (13, 14). MUC1/REP is a transmembrane protein with an extracellular domain consisting of one SEA domain and 20 repeated short sequences (13, 38). The two resulting cleavage products of MUC1/REP still form a heterodimer (13). It has been proposed that the SEA domain serves not only as a cleavage site but also as a site for reassociation (15). MUC1/SEC is a secreted form and lacks the transmembrane region (39). MUC1/Y has a short extracellular region that lacks repeated sequences (13, 40). MUC1/SEC and MUC1/Y reportedly function as a ligand and its receptor, respectively (13, 15). Ig-Hepta, which is a seven-transmembrane G protein-coupled receptor (41), undergoes a proteolytic cleavage at the same GSVVV sequence (16). MUC3 is cleaved in the same manner and forms a heterodimer even after cleavage (17).

The GSVVV motif must be exposed to the solvent to be properly cleaved. To examine the position of the motif on the three-dimensional structure, we performed multiple alignments of the SEA domains with the GSVVV motif, as illustrated in Fig. 6. Besides the GSVVV motif, most of the conserved residues appear to be involved in core formation. The GSVVV motif is located in the loop region between β2 and β3, which protrude from the entire domain. The GSVVV motif approximately corresponds to amino acid residues 65–72 (see Fig. 3). The solvent accessibility values of these residues (calculated by MOLMOL (30); the average of 20 structures) are 14, 44, 54, 33, 39, 28, 17, and 0.1%, respectively. This region appears to be sufficiently solvent-exposed for proteolytic cleavage. For MUC1/REP the resulting cleavage products form a heterodimer in vitro, and the resulting secreted and membrane-bound fragments may dissociate and then reassociate through the SEA domain to work as a ligand and receptor (13). The three-dimensional structure information would be useful to explain the molecular mechanism of this phenomenon.

Acknowledgments—We thank Yukiko Fujikura, Natsuko Matsuda, Miyuki Saito, Yukako Miyata, Noriko Hirakawa, Noriko Sakagami, Masoomi Ikari, Yasuko Tomo, Megumi Watanabe, Satoshi Yasuda, Tomoko Nakayama, and Hiroki Ito for technical assistance and Dr. Nobukazu Nameki for discussions and encouragement.

REFERENCES
1. Bast, R. C. Jr., Feeney, M., Lazarus, H., Nadler, L. M., Colvin, R. B., and Knapp, R. C. (1981) J. Clin. Invest. 68, 1331–1337
2. Meden, H., and Fattahi-Meibodi, A. (1998) Int. J. Mol. Biotechnol. 13, 231–237
3. Lloyd, K. O., Yin, B. W., and Lloyd, K. O. (2002) Int. J. Cancer 98, 737–740
4. Parry, S., Silverman, H. S., McDermott, K., Willis, A., Hollingsworth, M. A., and Harris, A. (2001) Biogch. Biochem. Res. Commun. 283, 1128–1130
5. Wreschner, D. H., McGuckin, M. A., Williams, A. J., Baruch, A., Yoeli, M., Ziv, R., Okun, L., Zaretsky, J., Smorodinski, N., Keydar, I., and Wreschner, D. H. (1999) Cancer Res. 59, 1552–1561
6. Abe, J., Fukuzawa, T., and Hirose, S. (2002) J. Biol. Chem. 277, 23981–23988
7. Wang, R., Khatri, I. A., and Forstner, J. F. (2002) Bioch. J. 366, 623–631
8. Igashira, T., Yabuki, T., Yoshihara, T., Matsui, S., Ito, K., and Yokoyama, S. (1999) FEBS Lett. 445, 15–19
9. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) J. Biomol. NMR 6, 277–280
10. Johnson, B., and Blevins, R. (1994) J. Biol. Chem. 269, 408–412
11. Markley, J. L. (1989) Methods Enzymol. 176, 12–64
12. Cordier, F., and Grzesiek, S. (1999) J. Am. Chem. Soc. 121, 1601–1602
34. Cush, R., Cronin, J. M., Stewart, W. J., Maule, C. H., Molloy, J., and Goddard, N. J. (1993) *Biosens. Bioelectronics* 8, 347–354
35. Aurora, R., and Rose, G. D. (1998) *Protein Sci.* 7, 21–38
36. Nap, M., Vitali, A., Nustad, K., Bast, R. C., Jr., O’Brien, T. J., Nilsson, O., Seguin, P., Suresh, M. R., Borner, O. P., Saga, T., de Bruijn, H. W., Nozawa, S., Kreutz, F. T., Jette, D., Sakahara, H., Gadinell, M., Ends, K., Barlow, E. H., Warren, D., Paus, E., Hammarström, S., Kenemans, P., and Hilgers, J. (1996) *Tumour Biol.* 17, 325–331
37. Croce, M. V., Rabassa, M. E., Price, M. R., and Segal-Eiras, A. (2001) *Pathol. Oncol. Res.* 7, 284–291
38. Gendler, S. J., Lancaster, C. A., Taylor, Papadimitriou, J., Dubig, T., Peat, N., Burchell, J., Pemberton, L., Lalani, E. N., and Wilson, D. (1990) *J. Biol. Chem.* 265, 15286–15293
39. Smorodinsky, N., Weiss, M., Hartmann, M. L., Baruch, A., Harness, E., Yaakovovitz, M., Kevdar, I., and Wreschner, D. H. (1996) *Biochem. Biophys. Res. Commun.* 228, 115–121
40. Ligtenberg, M. J., Kruijshaar, L., Buijs, F., van Meijer, M., Litvinov, S. V., and Hilken, J. (1992) *J. Biol. Chem.* 267, 6171–6177
41. Abe, J., Suzuki, H., Notsu, M., Yamamoto, T., and Hirose, S. (1999) *J. Biol. Chem.* 274, 19957–19964