FEBS Letters 527 (2002) 114–118 FEBS 26438

# A novel repeat in the melanoma-associated chondroitin sulfate proteoglycan defines a new protein family

Eike Staub*, Bernd Hinzmann, André Rosenthal

*metaGen Pharmaceuticals GmbH, Oudenarder Str. 16, D-13347 Berlin, Germany*

Received 18 June 2002; revised 26 July 2002; accepted 1 August 2002

First published online 14 August 2002

Edited by Gunnar von Heijne

**Abstract The human melanoma-associated chondroitin sulfate proteoglycan (MCSP) and its rat ortholog NG2 are thought to play important roles in angiogenesis-dependent processes like wound healing and tumor growth. Based on electron microscopy studies, the highly glycosylated ectodomain of NG2 has been subdivided into the globular N-terminus, a flexible rod-like central region and a C-terminal portion in globular conformation. We identified a novel repeat named CSPG in the central ectodomain of NG2, MCSP and other proteins from fly, worm, human, sea urchin and a cyanobacterium which shows similarity to cadherin repeats. As earlier electron microscopy studies indicate, the folding of the tandem repeats compresses the length of the proposed repeat region by a factor of ~10 compared to the fully extended peptide chain. We identified two conserved negatively charged residues which might govern the binding properties of CSPG repeats. The phyletic distribution of CSPG repeats suggests that horizontal gene transfer contributed to their evolutionary history.** 

*Key words:* Chondroitin sulfate proteoglycan; Kringle domain; Protein repeat; Cadherin repeat; β-Sandwich; CSPG repeat

## 1. Introduction

In multicellular organisms, interactions of cells with the extracellular matrix (ECM) are of fundamental importance, ensuring the remodelling and maintenance of tissue architecture of multicellular organisms. Different families of membrane proteins mediate these interactions with varying degrees of specificity to their binding partners in the ECM [1]. The most prominent members are the heterodimeric receptors of the integrin family [2], but in recent years several membrane glycoproteins have been identified as ECM-binding components, such as the syndecans which are important for tissue homeostasis and cancer development [3–5]. The functionality of the syndecans is governed by the attached heparan sulfate chains which can interact with a wide range of ligands, albeit with low ligand specificity [6]. Another family of membrane-bound heparan sulfate proteoglycans, the glypicans, has also been implicated in tumor formation as mutant glypican-3 causes the Simpson–Golabi–Behmel overgrowth syndrome [7,8]. The same molecule was recently shown to be a negative regulator of breast cancer [7,8].

Here we focus on the melanoma-associated chondroitin sulfate proteoglycan (MCSP) and its putative rat ortholog NG2. Human MCSP was first identified by its function as a high molecular weight melanoma-associated antigen [9]. Even before the gene was known, a monoclonal antibody directed against an anti-MCSP antibody and thus mimicking the unavailable natural MCSP protein proved to be an effective suppressor of anchorage-independent tumor growth [10]. NG2 was identified as a developmentally regulated membrane protein in various developing tissues [11]. The rat NG2 protein comprises 2326 amino acids and has a signal peptide, followed by a large extracellular domain, a transmembrane domain, and a 76 residue cytoplasmic tail. The ectodomain was subdivided into the D1, D2 and D3 domains based on sequence features. Four internal repeats of ~200 amino acids and two of ~30 amino acids length were described for the ectodomain. Apart from a 12 residue segment which was noted to resemble a $Ca^{2+}$-binding fragment in the second chicken *N*-cadherin repeat, no similarities to other proteins were noted. In electron microscopy images of NG2, the ectodomain appeared to be subdivided into three parts: the globular N-terminus, the globular C-terminus and the rod-like central region [12].

NG2 can be proteolytically processed, resulting in the release of almost the entire ectodomain [13]. Some biochemical studies on NG2/MCSP concentrated on their ligand-binding properties. In electron microscope images, Tillet et al. observed that collagen fibers aligned with the central flexible rod-like D2 domain and collagen V and VI were shown to bind specifically to the D2 domain in ligand-binding assays [12]. In addition, NG2 binds plasminogen and its fragments like angiostatin, as long as they harbor positively charged kringle domains. It was proposed that multiple kringle binding sites in NG2 exist, that the interaction does not depend on chondroitin sulfate (CS) chains and that positively charged residues on kringle domains bind acidic clusters in NG2, which leads to sequestering of angiostatin in gliomas [14,15]. The same mode of binding was also suggested to explain the interaction of NG2 with the PDGF-α receptor in the developing rat brain [16]. In adherent cell lines, NG2 was shown to be organized in arrays and to co-localize with actin and myosin-containing stress fibers [17]. Although the exact function of NG2 and MCSP is still unknown, they may be implicated in angiogenesis, tissue invasion and cell spreading [18–21].

*Corresponding author. Fax: (49)-30-45082 101.
*E-mail address:* eike.staub@metagen.de (E. Staub).

0014-5793/02/$22.00 
PII: S0014-5793(02)03195-2

## 2. Materials and methods

The non-redundant protein database from the NCBI was used as the basic pool of protein sequences in this study. Furthermore, we searched for sequence similarities in smaller databases of different proteomes from yeast, fly and worm (yeast.aa, drosoph.aa, wormpep) available from the FTP servers of the NCBI or the Sanger Institute. We used the BLASTP program [22] of the BLAST package with standard parameters to detect pairwise sequence similarities in these databases. The iterative PSIBLAST method was used to construct sequence profiles starting from single sequence fragments. During the iterations the inclusion of sequences, which are in the twilight zone of sequence similarity, into the profile was carefully checked. Inclusion thresholds were adjusted to include only true homologs, but were never raised above expectation ($E$) values of 0.008. Recursive searches using identified sequences were applied to confirm the detected similarities. As an independent and more sensitive method Hidden Markov Models (HMMs) of sequence alignments were applied to search protein sequence databases for additional homologous sequence fragments. To build, calibrate and apply profile HMMs we used the programs of the HMMER package [23]. Intramolecular repeats were visualized by a dot plot analysis using the program DOTTER [24]. The significance of the similarity between putative intramolecular repeats was confirmed by cross-comparisons using two additional methods, the PRSS program [25] from the FASTA package and the PROSPERO program [26]. Multiple alignments were created using CLUSTALX [27] and edited using JALVIEW (Clamp, M., unpublished). For the coloring of the alignment according to consensus rules we used the CHROMA program [28]. Signal peptides were predicted using the SIGNALP program in version 2.0 [29]. We predicted transmembrane helices using TMHMM version 2.0 [30]. The secondary structure of proteins was predicted on the basis of protein sequence alignments using the PHD prediction server [31].

## 3. Results and discussion

### 3.1. Identification of the CSPG repeat

As the knowledge about protein domains increased dramatically during the last years, we rescanned the NG2 sequence using the Smart and Pfam domain databases [32,33]. We discovered two laminin-G domains in the N-terminus of NG2. These domains occupy a large part of the formerly defined D1 region in the ectodomain of NG2. Though laminin-G domains are widespread among many extracellular proteins, their general function is not known. In laminin-G, this domain is implicated in heparin binding. It shows sequence similarity to pentraxins and thrombospondin-like molecules and a common fold was predicted for members of this superfamily [34].

Dot plot analysis [24] of the NG2 sequence (GenPep accession CAA39884.2) revealed extensive repeat structures between residues 420 and 2135 in the NG2 ectodomain. Most diagonals were separated by approximately 100 amino acids, which is an indicator for the repeat size. We chose the subsequence from 1124–1226 as a prototype of the putative repeat because it appeared to be similar to the highest number of other subsequences. When we compared this fragment with the whole NG2 ectodomain sequence using PROSPERO [26], four copies of the repeat were detected with expectation ($E$) values below 1e−5. When we used this subsequence as a seed in a PSIBLAST [22] query of the non-redundant protein database at the NCBI (nr) with an inclusion threshold $E$ value of 0.008, the search converged after four rounds having identified 63 repeat copies in 10 different proteins from human, rat worm, fly, sea urchin and a cyanobacterium. We detected 10 copies of the repeat in the NG2 sequence. We confirmed this finding by extensive reciprocal PSIBLAST searches using different repeat copies as queries. The PSIBLAST searches usually converged after three to five rounds, having identified largely overlapping sets of subsequences. In addition, we were able to proof the significance of the similarity between repeats by extensive pairwise comparisons using the PROSPERO algorithm and the PRSS algorithm [25]. Hereafter, we refer to the repeat as CSPG repeat.

To achieve maximum sensitivity in database searches we aligned the CSPG repeat sequences and constructed a HMM using the HMMer program package [23]. A search in the nr database using the HMM revealed eight non-redundant protein sequences with a total of 74 repeat copies ($E < 1e{-}3$) (see Fig. 1). For each of the rat NG2 and human MCSP proteins 15 repeat copies were now found covering the whole region that was expected to contain repeats after dot plot analysis. We predicted the secondary structure on the basis of the manually edited alignment using the PHD prediction server (Fig. 1). The CSPG repeat is likely to obtain an all-β fold, possibly comprising eight β-strands. The sixth β-strand starts with a conserved aromatic residue, which is followed by a conserved serine or threonine. Conserved acidic residues are present in the subsequent loop regions between strands 6 and 7 as well as between 7 and 8. For most β-strands one can observe a typical alternating pattern of hydrophobic and non-hydrophobic residues. Hydrophobic side chains that point to the same side of the β-sheet are probably buried in the interior of the CSPG domain. We applied several fold recognition methods using single sequences or the whole alignment as queries, but no significant predictions were obtained.

We also determined the domain architectures of the CSPG repeat proteins (Fig. 2). The CSPG repeat occurs in 1–15 tandem copies per protein. In some proteins the CSPG repeat is combined with laminin-G domains and EGF-like domains. These domains are common in extracellular proteins and are known to mediate interactions between cell surface molecules and extracellular ligands or matrix components. We found 12 copies of the CSPG repeat in the embryonic blastocoelar matrix protein ECM3 (GenPep AAG00570.1) from the sea urchin *Lytechinus variegatus* next to a five-fold tandem repeat of Calx-β motifs. These motifs were originally found to occur in cytoplasmic regulatory regions of $Na^{+}$–$Ca^{2+}$-exchange proteins and integrin-β4. In ECM3 they reside in a putative extracellular region between a predicted signal peptide and a single transmembrane domain. The function of these, presumably extracellular, Calx-β motifs is not clear, although it is assumed that they bind calcium [35]. Support for the hypothesis that extracellular Calx-β motifs bind calcium comes from the sponge MAFp4 ECM protein which requires calcium for self-association.

One CSPG repeat prediction in ECM3 (residues 1145–1240) overlapped with a weak prediction of a cadherin repeat (residues 1169–1260; $E = 0.54$) detected by Smart [33]. This was an indicator of similarity between cadherin-like repeats and CSPG repeats. PSIBLAST searches starting with this sea urchin sequence fragment converged on a set of CSPG repeats without identification of cadherin repeats, although marginal similarity to cadherins was detected in weak hits with $E$ values below the profile inclusion threshold. To gain sensitivity we built a profile HMM from an alignment of CSPG repeats with less than 70% pairwise identity. When we applied the HMM to the sequences of cadherins we obtained nine matches with $E$ values in the range between 0.1 and 0.0095 which can be

Fig. 1. Alignment of selected CSPG repeats. The identifier of each sequence in the non-redundant database (NCBI) is followed by the position of each repeat in the sequence. The protein names and species can be taken from the legend of Fig. 2. The alignment was colored according to a 70% consensus using the following amino acid classification: negatively charged: white on red, DE (-); hydroxylic: black on yellow, ST (*); aliphatic: white on dark blue, ILV (l); positively charged: black on green, HKR (+); tiny: white on brown, AGS (t); aromatic: white on purple, FHWY (a); polar: white on green, CDEHKNQRST (p); hydrophobic: black on light blue, ACFGHILMTVWY (h). Below the alignment of CSPG repeats the predicted secondary structure as determined using the PHD prediction server and the corresponding reliability values are printed (0–9; 9 is most reliable) [31]. The predicted secondary structure can easily be compared to the secondary structure of a cadherin repeat in the solved 3D structure of an *N*-cadherin fragment (PDB code 1NCJ) which is given in the last two lines together with its protein sequence. Secondary structure code: *E* stands for β-strand, *H* for α-helix.

considered significant. Therefore, we hypothesize that CSPG and cadherin repeats are distantly related. Apart from this similarity, the CSPG repeat predictions did not overlap with any predictions of known domains from Pfam and Smart.

### 3.2. Structural and functional implications

The identification of laminin-G domains and the novel CSPG repeats permitted a finer partitioning of the ectodomains of the NG2 and MCSP oncoproteins compared to the originally proposed D1/D2/D3 division. The presence of CSPG repeats in proteins from worm, fly and sea urchin shed light on the phyletic distribution of the CSPG repeats. The presence of a single CSPG repeat in a cyanobacterium may suggest that the CSPG repeat is an ancient protein module that was preserved during evolution. Alternatively, it may be an example of domain accretion by horizontal gene transfer

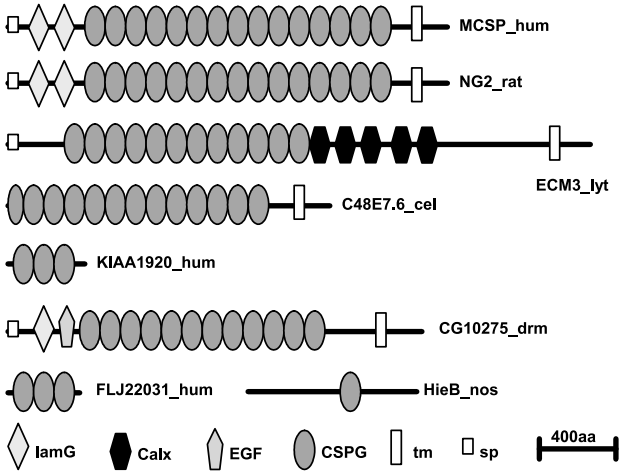


Fig. 2. Domain architecture of CSPG repeat proteins. The following domain abbreviations were used: lamG, laminin-G domain; EGF, EGF-like domain; Calx, Calx-β domain, CSPG, CS proteoglycan repeat; sp, signal peptide; tm, transmembrane helix. The protein identifiers consist of a protein and a species abbreviation. The protein names and their GenBank identifiers are: MCSP, gi:4503099; NG2, gi:13591932; embryonic blastocoelar matrix protein (ECM3), gi:9837427; KIAA1920, gi:15620899; hypothetical protein FLJ-22031, gi:13640532; CG10275, gi:7298451; C48E7.6, gi:7497613; HieB, gi:9858192. The species abbreviations are hum for *Homo sapiens*, rat for *Rattus norvegicus*, lyt for *L. variegatus*, drm for *Drosophila melanogaster*, cel for *Caenorhabditis elegans*, nos for *Nostoc* species PCC9229. The domain positions on the sequences are drawn approximately to scale.

from an unknown multicellular eukaryote to a cyanobacterium. We noticed the similarity of the phyletic distributions of CSPG repeats and the Calx-β motif which was also found in higher eukaryotes and cyanobacterial proteins. Both motifs are not present in other eukaryotic organisms like yeast, fly and worm. We think that a scenario in which these motifs are deleted from yeast, fly and worm genomes is less likely than the horizontal transfer of genes or gene fragments with CSPG repeats and Calx-β motifs between a marine eukaryote and a cyanobacterium.

Evidence that CSPG repeats fold into structural modules comes from the reconsideration of earlier electron microscopy studies in the light of the repeat discovery. Assuming a maximum extension of the polypeptide chain and a per-residue distance of 0.36 nm, the ~1700 residues of the CSPG repeat region would result in an extended polypeptide chain of 612 nm. The length of the rod-like central domain of the NG2 ectodomain was estimated to be in the range of 30–110 nm in electron microscopy images [12]. This implies that the folding of the repeat region results in a significant (~10-fold) compression of the length of the polypeptide chain. It is likely that this protein shrinking is conferred by the folding of CSPG repeats into structural units.

Further evidence for the relatedness between cadherin and CSPG repeats came from a comparison of the secondary structures of cadherin repeats with known 3D structures and the predicted secondary structures of CSPG repeats. The cadherin repeat in the second domain of an *N*-cadherin fragment folds into a β-sandwich (PDB code 1NCJ) [36]. We aligned the 1NCJ sequence to the CSPG repeat alignment (see Fig. 1). The six β-strands of the second cadherin domain aligned to a large extent with the predicted β-strands in CSPG repeats. Furthermore, cadherin-like and CSPG repeats are both thought to obtain a rod-like structure and the size of the repeat units from both families is ~100 residues. Therefore, we hypothesize that CSPG repeats and cadherin repeats share a common ancestor and structural fold. Do they also have similar biochemical properties? Compared to the many negatively charged residues involved in calcium binding of cadherin repeats, the CSPG repeats contain only two negatively charged positions in their C-terminal half (see Fig. 1). A calcium-binding capacity has not been reported for CSPG repeat proteins yet. It cannot be inferred from sequence analysis alone whether CSPG repeats bind calcium by their two acidic residues.

Insights into the biochemical function of CSPG repeats can be gained by reviewing the literature on the MCSP/NG2 proteins. One function of CSPG repeats may be the binding and presentation of the CS chains which then determine the functional properties of the molecule. However, the binding of the D2 and D3 regions of NG2 to positively charged kringle domains of the plasmin(ogen)/angiostatin system seemed to be independent of the presence of CS chains. As these regions comprise most of the CSPG repeats and multiple binding sites seem to exist, the binding of kringle domains is possibly facilitated by negatively charged conserved residues in the CSPG repeats (see Fig. 1). Another function of CSPG repeats is the binding of collagen. The D2 region of NG2 was shown to bind collagen and almost completely consists of CSPG repeats.

We conclude that the CSPG repeat is a novel cadherin-like and tumor-relevant protein module which we expect to mediate interactions between cells and the ECM in species as divergent as cyanobacteria, fly, worm, sea urchin and human. Furthermore, we propose that horizontal gene transfer contributed to the evolutionary history of genes which encode CSPG repeats.

## References

[1] Pluschke, G., Vanek, M., Evans, A., Dittmar, T., Schmid, P., Itin, P., Filardo, E.J. and Reisfeld, R.A. (1996) Proc. Natl. Acad. Sci. USA 93, 9710–9715.
[2] Hynes, R.O. (1999) Trends Cell Biol. 9, M33–M37.
[3] Woods, A. and Couchman, J.R. (1998) Trends Cell Biol. 8, 189–192.
[4] Liu, W., Litwack, E.D., Stanley, M.J., Langford, J.K., Lander, A.D. and Sanderson, R.D. (1998) J. Biol. Chem. 273, 22825–22832.
[5] Alexander, C.M., Reichsman, F., Hinkes, M.T., Lincecum, J., Becker, K.A., Cumberledge, S. and Bernfield, M. (2000) Nat. Genet. 25, 329–332.
[6] Carey, D.J. (1997) Biochem. J. 327, 1–16.
[7] Pilia, G. et al. (1996) Nat. Genet. 12, 241–247.
[8] Xiang, Y.Y., Ladeda, V. and Filmus, J. (2001) Oncogene 20, 7408–7412.
[9] Bumol, T.F., Wang, Q.C., Reisfeld, R.A. and Kaplan, N.O. (1983) Proc. Natl. Acad. Sci. USA 80, 529–533.
[10] Harper, J.R. and Reisfeld, R.A. (1983) J. Natl. Cancer Inst. 71, 259–263.
[11] Nishiyama, A., Dahlin, K.J., Prince, J.T., Johnstone, S.R. and Stallcup, W.B. (1991) J. Cell Biol. 114, 359–371.
[12] Tillet, E., Ruggiero, F., Nishiyama, A. and Stallcup, W.B. (1997) J. Biol. Chem. 272, 10769–10776.
[13] Nishiyama, A., Lin, X.H. and Stallcup, W.B. (1995) Mol. Biol. Cell 6, 1819–1832.
[14] Goretzki, L., Lombardo, C.R. and Stallcup, W.B. (2000) J. Biol. Chem. 275, 28625–28633.
[15] Chekenya, M. et al. (2002) FASEB J. 12, 12.

[16] Nishiyama, A., Lin, X.H., Giese, N., Heldin, C.H. and Stallcup, W.B. (1996) J. Neurosci. Res. 43, 315–330.
[17] Lin, X.H., Dahlin-Huppe, K. and Stallcup, W.B. (1996) J. Cell Biochem. 63, 463–477.
[18] Burg, M.A., Pasqualini, R., Arap, W., Ruoslahti, E. and Stallcup, W.B. (1999) Cancer Res. 59, 2869–2874.
[19] Iida, J., Pei, D., Kang, T., Simpson, M.A., Herlyn, M., Furcht, L.T. and McCarthy, J.B. (2001) J. Biol. Chem. 276, 18786–18794.
[20] Eisenmann, K.M. et al. (1999) Nat. Cell Biol. 1, 507–513.
[21] Iida, J., Meijne, A.M., Spiro, R.C., Roos, E., Furcht, L.T. and McCarthy, J.B. (1995) Cancer Res. 55, 2177–2185.
[22] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[23] Eddy, S.R. (1998) Bioinformatics 14, 755–763.
[24] Sonnhammer, E.L. and Durbin, R. (1995) Gene 167, GC1–GC10.
[25] Pearson, W.R. (2000) Methods Mol. Biol. 132, 185–219.
[26] Mott, R. (2000) J. Mol. Biol. 300, 649–659.
[27] Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G. and Gibson, T.J. (1998) Trends Biochem. Sci. 23, 403–405.
[28] Goodstadt, L. and Ponting, C.P. (2001) Bioinformatics 17, 845–846.
[29] Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Int. J. Neural Syst. 8, 581–599.
[30] Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) J. Mol. Biol. 305, 567–580.
[31] Rost, B. and Sander, C. (1994) Proteins 19, 55–72.
[32] Bateman, A. et al. (2002) Nucleic Acids Res. 30, 276–280.
[33] Letunic, I. et al. (2002) Nucleic Acids Res. 30, 242–244.
[34] Beckmann, G., Hanke, J., Bork, P. and Reich, J.G. (1998) J. Mol. Biol. 275, 725–730.
[35] Hodor, P.G., Illies, M.R., Broadley, S. and Ettensohn, C.A. (2000) Dev. Biol. 222, 181–194.
[36] Tamura, K., Shan, W.S., Hendrickson, W.A., Colman, D.R. and Shapiro, L. (1998) Neuron 20, 1153–1163.