Prediction of structure and functional residues for *O*-GlcNAcase, a divergent homologue of acetyltransferases

Jörg Schultz*, Birgit Pils

Computational Molecular Biology Department, Max-Planck-Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany

Received 15 July 2002; revised 23 August 2002; accepted 23 August 2002

First published online 4 September 2002

Edited by Robert B. Russell

Abstract ***N*-Acetyl-β-D-glucosaminidase (*O*-GlcNAcase) is a key enzyme in the posttranslational modification of intracellular proteins by *O*-linked *N*-acetylglucosamine (*O*-GlcNAc). Here, we show that this protein contains two catalytic domains, one homologous to bacterial hyaluronidases and one belonging to the GCN5-related family of acetyltransferases (GNATs). Using sequence and structural information, we predict that the GNAT homologous region contains the *O*-GlcNAcase activity. Thus, *O*-GlcNAcase is the first member of the GNAT family not involved in transfer of acetyl groups, adding a new mode of evolution to this large protein family. Comparison with solved structures of different GNATs led to a reliable structure prediction and mapping of residues involved in binding of the GlcNAc-modified proteins and catalysis.**

© 2002 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: *O*-Glycosylation; Signalling enzymes; Function prediction

1. Introduction

The function of many proteins is regulated by posttranslational modifications. This is especially apparent in the case of signal transduction, where the activity of proteins has to be fine-tuned to react accordingly to intra- and extracellular signals.

Only recently it became clear, that in addition to phosphorylation of proteins also their glycosylation by *O*-linked β-*N*-acetylglucosamine (*O*-GlcNAc) might be involved in signal transduction [1]. In contrast to the more complex *N*- and *O*-linked glycosylation of extracellular proteins, this modification is added to serine or threonine residues of intracellular proteins. A plethora of proteins from species all over the eukaryotic kingdom has been reported to contain *O*-GlcNAc. The development of novel methods usable in proteomics approaches like antibodies or mass-spectrometry can be expected to further increase this number.

These studies might add to the increasing evidence of an implication of *O*-GlcNAc into different human diseases like cancer [2], diabetes [3] and neurodegenerative diseases. Tau for example, which is a major component of neurofibrillary tangles in Alzheimer's diseased brains, is multiply *O*-GlcNAc modified in normal brains, whereas it is highly phosphorylated in association with Alzheimer's disease [4].

Not only this case links glycosylation and phosphorylation, many different proteins are modified by both systems. As the same types of amino acids are modified, reciprocity was suspected. Indeed it was shown in different cases like the oestrogen receptor [5] and the protooncogene c-Myc [6], that the same residues are either phosphorylated or glycosylated. A fine-tuned regulation is crucial for this 'ying-yang' relationship. The responsible enzymes are the *O*-GlcNAc transferase (OGT) and the *N*-acetyl-β-D-glucosaminidase (*O*-GlcNAcase). The OGT consists of an N-terminal part build of TP repeats and a C-terminal catalytic unit. In depth sequence analysis revealed that the catalytic unit is homologous to the glycogen phosphorylase superfamily [7]. Less is known about the *O*-GlcNAcase, which was originally cloned as a putative hyaluronidase and only later shown to additionally possess *N*-acetylglucosaminidase activity. The region responsible for the *O*-GlcNAcase activity has not been mapped and neither the substrate binding site nor catalytic residues have been described, hindering detailed experimental characterisation of the GlcNAcase and restricting the analysis of *O*-GlcNAc modification.

2. Materials and methods

Members of the *O*-GlcNAcase family were found by BLASTp [8] searching with the human MGEA5 sequence against NCBI's non-redundant database. To identify known domains, sequences were checked against against SMART [9] and Pfam [10]. The C-terminal sequence of MGEA5 (700–916) was further analysed by PSI-BLAST. The obtained putative *O*-GlcNAcase sequences were aligned with ClustalX [11] and the resulting alignment was manually optimised in Seaview [12].

Secondary structure elements were predicted by Jpred [13]. To identify homologues with known structures, intermediate sequence search [14] was used. Within the *N*-acetyltransferase family four additional proteins were selected for a structural alignment: two closely related GNA5 histone acetyltransferase (PDB Id. 1I12 and 1CJW) and two aminoglycoside *N*-acetyltransferases (1bo4 and 1b87) because of their abilities to bind GlcNAc similar substrates. Pairwise structural alignments performed by VAST [15] were obtained from NCBI's MMDB [16]. By comparing each of the four structures to each other a multiple alignment was build based on secondary structure elements.

3. Results

Searches with MGEA5 (gi: 10835356), the first cloned *O*-GlcNAcase [17], against different domain databases (Pfam [10], SMART [9]) did not reveal any conserved motif, although Comtesse et al. [18] report a C-terminal acetyltrans-

*Corresponding author. Fax: (49)-30-8413 1152.
E-mail address: joerg.schultz@molgen.mpg.de (J. Schultz).

0014-5793/02/$22.00 © 2002 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.
PII: S0014-5793(02)03322-7

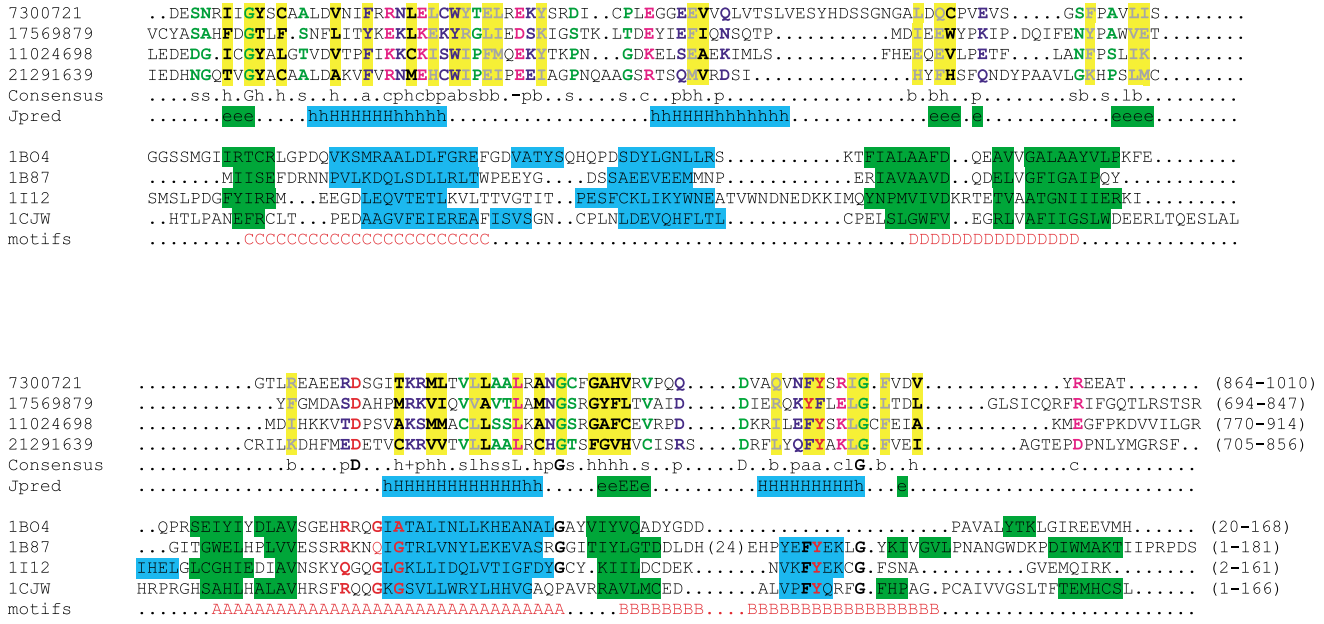

Fig. 1. Multiple alignment of *O*-GlcNAcase sequences and structurally characterised members of the acetyltransferase superfamily. Sequences are labelled with GenBank or PDB identifier on the left side, the region of the sequence used in the alignment is given in parentheses on the lower right side. Numbers in parentheses within the alignment indicate the length of an insert. The upper four sequences (7300721: *Drosophila melanogaster*, 17569879: *Caenorhabditis elegans*, 11024698: *Homo sapiens*, 21291639: *Anopheles gambiae*) are representatives of the *O*-GlcNAcase family. Sequences from *Mus musculus* (gi: 14329484) and *Rattus norvegicus* (gi: 16943639) are not shown, as they are in the aligned region nearly identical to the human sequence. Conserved residues are coloured and the 80% consensus determined by Chroma [29] is given below (capital letters represent amino acids, lower case letters: a, aromatic; c, charged; s, small; p, polar; b, big; h, hydrophobic; −, negative; +, positive). A consensus secondary structure prediction for these sequences was obtained from Jpred and is shown in the corresponding line. α-Helices are denoted by h/H, β-sheets by e/E. Capital letters indicate a reliability score higher than 6. The lower four sequences (1BO4: aminoglycoside 3-*N*-acetyltransferase, 1B87: aminoglycoside 6-*N*-acetyltransferase, 1CJW: serotonin *N*-acetyltransferase, 1I12: glucosamine-phosphate *N*-acetyltransferase) belong to the family of Gcn5-related *N*-acetyltransferases. Structural elements are hallmarked by blue background for helices or green background for β-sheets. Conserved functional residues are highlighted by bold red characters, conserved residues by bold characters. Beneath the alignment positions of the four conserved regions are shown that are described as motifs A, B, C and D by Neuwald et al. [21]

ferase domain found by SMART. Still, a sequence search using BLASTp [8] against the non-redundant subset of GenBank found different, significant homologous proteins. These include bacterial hyaluronidases as reported in Heckel et al. [19] (*E*-value from 10^{-41} to 3×10^{-8}) and a putative acetyltransferase ($E=3\times10^{-8}$). Inspection of the pairwise alignments revealed two regions of homology, separated by a linking sequence. The N-terminal region was homologous to the hyaluronidases, whereas the similarity to the acetyltransferases was restricted to the C-terminus.

Iterative BLAST searches with the C-terminal part (aa 700–916) found different acetyltransferases. All found proteins belong to the GCN5-related family of acetyltransferases (GNAT), for which a couple of structures have been solved (for review see [20]). Yet, direct searches with the C-terminus of MGEA5 against the sequences of proteins with known structure did not reveal significant similarities. This link could only be established using intermediate sequence searches [14], that is, searching with sequences retrieved in previous searches. For example, the putative acetyltransferase sequence from *Streptomyces coelicolor* (gi: 7160091) found in a search with MGEA5 ($E=7\times10^{-12}$) did detect the PDB sequence 1CJW, a serotonin acetyltransferase ($E=0.091$).

To further strengthen the link between the *O*-GlcNAcases and acetyltransferases selected members of the GNAT family were aligned according to their structures. A secondary structure prediction based on an alignment of the C-terminal region of the *O*-GlcNAcase family revealed a significant congruence between the secondary structure elements (Fig. 1). Taken together, these results show that the *O*-GlcNAcase is a divergent member of the GNAT family.

The degree of conservation between the GlcNAcase family and the acetyltransferases varies in different regions of the alignment (Fig. 1). The C-terminal region, motif B, is highly conserved between the GlcNAcases and a subfamily of acetyltransferases including glucosamine-6-phosphate *N*-acetyltransferases (PDB Id. 1I12) and serotonin *N*-acetyltransferases (PDB Id. 1CJW). This region with the consensus FYxxxG is not present in other GNATs with differing substrates like the aminoglycoside 3′*N*-acetyltransferases. Only in the *Caenorhabditis elegans* sequence, two of these residues have been 'swapped'. Less conservation can be found in motif A containing the acetyl-CoA binding site of the GNATs which can be described by the consensus sequence R/QxxGxG/A [21]. This motif is not present in the corresponding region of the GlcNAcases, which in contrast contain a conserved D not found in the GNATs.

4. Discussion

4.1. Structure and functional residues of O-GlcNAcase

The human gene encoding a *O*-GlcNAcase, MGEA5, was cloned by two independent studies, one purifying a putative hyaluronidase [19], the other a *N*-acetylglucosaminidase [17].

As the substrates of these enzymes share some similarity, it was speculated that both reactions are performed by one, unspecific catalytic centre [18]. Based on sequence similarity to bacterial proteins, Hanover [22] tentatively assigns this catalytic centre to the N-terminal 300 aa of MGEA5. Here we showed that MGEA5 contains two regions similar to distinct enzymes. The N-terminal region is homologous to bacterial hyaluronidases, whereas the C-terminus shows homology to acetyltransferases. Evidence that the C-terminal part contains the *N*-acetylglucosaminidase activity comes from the observation that one of the acetyltransferases most similar to MGEA5 is GNA1. This protein, whose structure has been solved [23], is involved in the biosynthesis of UDP–GlcNAc by catalysing the formation of GlcNAc6P. This led to the question, whether the ability to bind GlcNAc might be conserved in MGEA5. The region of GNA1 responsible for interaction with GlcNAc6P, commonly denominated Motif B [21], is highly conserved within the *O*-GlcNAcases but not in other acetyltransferases with strongly differing substrates like aminoglycosides (Fig. 1, sequence 1bo4). The most prominent feature is the conservation of a tyrosine residue, which is involved in the binding of GlcNAc and the catalytic mechanism of GNA1 [23]. This tyrosine as well as two other positions is conserved within all members of the *O*-GlcNAcases. Taken together, this led to the prediction, that the C-terminal, acetyltransferase homologous region of MGEA5 is responsible for binding *O*-GlcNAc-modified proteins. Furthermore, the conserved tyrosine might be involved in the catalytic function of MGEA5.

Contrasting the well conserved proposed GlcNAc binding site, regions which are in the classical acetyltransferases involved in acetyl-CoA binding have diverged in the GlcNAcases. Most of the interactions between the acetyltransferases and acetyl-CoA are performed by a helix termed motif A (Fig. 1). Within this motif, residues which can be described by a Q/RxxGxG pattern are involved in direct interaction [24]. This pattern is not conserved in the GlcNAcases. Instead, they contain a conserved D within the region of this pattern. As, according to the structures of the GNATs, this residue is solvent accessible and close to the proposed catalytic tyrosine, it might either be involved in binding of the *O*-GlcNAcase-modified protein or directly in the catalytic process.

Additional experimental evidence for a C-terminal location of the *N*-acetylglucosaminidase comes from a splice variant of MGEA5, which misses the C-terminal region (MGEA5s). No in vitro enzymatic activity was found for this variant [25]. Furthermore, the full-length variant is mainly localised in the cytoplasm, whereas the short form has a nuclear localisation [18]. Together with the finding that major parts of the *O*-GlcNAcase catalytic activity are present in the cytoplasm [25], these experimental results indicate that the catalytic activity is localised in the C-terminal region of MGEA5, which is homologous to acetyltransferases.

4.2. *O-GlcNAcases as linkers of different regulatory processes?*

Two catalytic functions have been ascribed to MGEA5, a hyaluronidase [19] and a *N*-acetylglucosaminidase [25] activity. The mapping of two independent catalytic domains raises the question, why these are co-localised in the same protein. The presence of intracellular hyaluronan, the substrate of hyaluronidases, was described only recently and an involvement in regulatory processes is assumed [26]. The GlcNAc modification also takes part in regulatory mechanisms and is highly interwoven with phosphorylation [1]. This might indicate that MGEA5 links two regulatory mechanisms. With the mapping of the responsible catalytic domains, both processes can now be studied independent of each other.

The prediction of functional residues for the *N*-acetylglucosaminidase will allow creating non-functional mutants. As these should not be able to remove GlcNAc groups from modified proteins, they will influence the cellular GlcNAcase modification state. Complementary to described specific inhibitors, these mutants might therefore be usefull in the determination of proteins, which undergo glycosylation in response to various stimuli.

4.3. *A novel mechanism in the evolution of acetyltransferases*

In the evolution of novel enzymes one can imagine two different scenarios. Either a protein which has a defined catalytic activity evolves a new substrate affinity or an enzyme with a given substrate affinity changes the underlying catalytic mechanism. All so far described members of the GCN5-related acetyltransferase family transfer an acetyl group from acetyl-CoA to a wide range of substrates like small molecules and proteins using a conserved catalytic mechanism [20]. On the structural level, GNATs are related to *N*-myristoyltransferases [27]. Although the transferred molecule is changed, the major catalytic mechanism and the mode of cofactor binding are conserved [28]. This indicates that the prevalent mode of evolution in this superfamily is the change of substrate affinity by keeping catalytic mechanism and cofactor binding constant. The membership of *O*-GlcNAcase to the family of acetyltransferases adds a new mode of evolution to this family. Not only a novel catalytic mechanism has developed but also cofactor-binding capabilities were lost.

This gives rise to questions about the possible starting point for this novel invention. One step in the biochemical pathway leading to UDP–GlcNAc, the substrate that is transferred by the OGT to a protein, is the addition of an acetyl group to Glc6P, giving GlcNAc6P. The protein catalysing this step, GNA1, belongs to the GNAT family and has an affinity for GlcNAc [23]. Starting from this protein, the *O*-GlcNAcase activity might have evolved by changing the underlying catalytic mechanism. Thus, one could imagine a recruitment of this protein from the metabolic pathway to the regulatory *O*-GlcNAc pathway.

References

[1] Wells, L., Vosseller, K. and Hart, G.W. (2001) Science 291, 2376–2378.
[2] Shaw, P., Freeman, J., Bovey, R. and Iggo, R. (1996) Oncogene 12, 921–930.
[3] Vosseller, K., Wells, L., Lane, M.D. and Hart, G.W. (2002) Proc. Natl. Acad. Sci. USA 99, 5313–5318.
[4] Arnold, C.S., Johnson, G.V., Cole, R.N., Dong, D.L., Lee, M. and Hart, G.W. (1996) J. Biol. Chem. 271, 28741–28744.
[5] Cheng, X., Cole, R.N., Zaia, J. and Hart, G.W. (2000) Biochemistry 39, 11609–11620.
[6] Chou, T.Y., Hart, G.W. and Dang, C.V. (1995) J. Biol. Chem. 270, 18961–18965.
[7] Wrabl, J.O. and Grishin, N.V. (2001) J. Mol. Biol. 314, 365–374.
[8] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol. 215, 403–410.
[9] Letunic, I. et al. (2002) Nucleic Acids Res. 30, 242–244.
[10] Bateman, A. et al. (2002) Nucleic Acids Res. 30, 276–280.

[11] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) Nucleic Acids Res. 25, 4876–4882.
[12] Galtier, N., Gouy, M. and Gautier, C. (1996) Comput. Appl. Biosci. 12, 543–548.
[13] Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M. and Barton, G.J. (1998) Bioinformatics 14, 892–893.
[14] Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. and Chothia, C. (1998) J. Mol. Biol. 284, 1201–1210.
[15] Madej, T., Gibrat, J.F. and Bryant, S.H. (1995) Proteins 23, 356–369.
[16] Wang, Y. et al. (2002) Nucleic Acids Res. 30, 249–252.
[17] Gao, Y., Wells, L., Comer, F.I., Parker, G.J. and Hart, G.W. (2001) J. Biol. Chem. 276, 9838–9845.
[18] Comtesse, N., Maldener, E. and Meese, E. (2001) Biochem. Biophys. Res. Commun. 283, 634–640.
[19] Heckel, D., Comtesse, N., Brass, N., Blin, N., Zang, K.D. and Meese, E. (1998) Hum. Mol. Genet. 7, 1859–1872.
[20] Dyda, F., Klein, D.C. and Hickman, A.B. (2000) Annu. Rev. Biophys. Biomol. Struct. 29, 81–103.
[21] Neuwald, A.F. and Landsman, D. (1997) Trends Biochem. Sci. 22, 154–155.
[22] Hanover, J.A. (2001) FASEB J. 15, 1865–1876.
[23] Peneff, C., Mengin-Lecreulx, D. and Bourne, Y. (2001) J. Biol. Chem. 276, 16328–16334.
[24] Wolf, E., Vassilev, A., Makino, Y., Sali, A., Nakatani, Y. and Burley, S.K. (1998) Cell 94, 439–449.
[25] Wells, L., Gao, Y., Mahoney, J.A., Vosseller, K., Chen, C., Rosen, A. and Hart, G.W. (2002) J. Biol. Chem. 277, 1755–1761.
[26] Lee, J.Y. and Spicer, A.P. (2000) Curr. Opin. Cell. Biol. 12, 581–586.
[27] Bhatnagar, R.S., Futterer, K., Waksman, G. and Gordon, J.I. (1999) Biochim. Biophys. Acta 1441, 162–172.
[28] Bhatnagar, R.S. et al. (1998) Nat. Struct. Biol. 5, 1091–1097.
[29] Goodstadt, L. and Ponting, C.P. (2001) Bioinformatics 17, 845–846.