Peptide-*N*-glycanases and DNA repair proteins, Xp-C/Rad4, are, respectively, active and inactivated enzymes sharing a common transglutaminase fold

Vivek Anantharaman, Eugene V. Koonin and L. Aravind*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received March 5, 2001; Revised and Accepted June 12, 2001

Yeast RAD4, its human ortholog Xp-C and their orthologs in other eukaryotes are DNA repair proteins which participate in nucleotide excision repair through a ubiquitin-dependent process. However, no conserved globular domains that might have shed light on their origin or functions have been reported for these proteins. By using sequence profile analysis, we show that RAD4/Xp-C proteins contain the ancient transglutaminase fold and are specifically related to the recently characterized peptide-*N*-glycanases (PNGases) which remove glycans from glycoproteins during their degradation. The PNGases retain the catalytic triad that is typical of this fold and are predicted to have a reaction mechanism similar to that involved in transglutamination. In contrast, the RAD4/Xp-C proteins are predicted to be inactive and are likely to only possess the protein interaction function in DNA repair. These proteins also contain a long, low-complexity insert in the globular transglutaminase domain. The RAD4/Xp-C proteins, along with other inactive transglutaminase-fold proteins, represent a case of functional re-assignment of an ancient domain following the loss of the ancestral enzymatic activity.

INTRODUCTION

Eukaryotes have a complex system of nucleotide excision repair (NER) (1,2) that has been the subject of extensive studies, particularly in connection with the inactivation of various components of this system in human diseases such as Xeroderma pigmentosum [XP (3)]. Studies with the yeast model system and with XP complementation groups have led to the identification of a variety of repair enzymes, including DNA helicases and ATPases, such as the ERCC2/3, and nucleases, such as ERCC1/XP-F and XP-G. Additional components of the excision repair system include Xp-C/Rad4, RAD23 and RAD7 that mediate DNA–protein and protein–protein interactions. Most of these proteins contain recognizable, conserved globular domains which are consistent with the corresponding biochemical activities (4,5). Yeast Rad4 and its human ortholog XP-C play an important role in the recognition of DNA damage and recruitment of the TFIIH complex for excision repair (6–9), but contain no previously identified domains. Here we show that Rad4/XP-C is an inactive homolog of the recently identified peptide-*N*-glycanases (PNGases) that are involved in glycoprotein degradation (10) and that these proteins share a core transglutaminase fold.

RESULTS AND DISCUSSION

Iterative searches of the non-redundant protein sequence database (National Center for Biotechnology Information, NIH, Bethesda) using the PSI-BLAST program (11) revealed statistically significant sequence similarity between the Xp-C/Rad4 proteins and the PNGases. These searches also showed a statistically significant similarity between the Xp-C/Rad4, PNGases and the proteins of the transglutaminase superfamily (12) (Fig. 1). PFAM search tools that utilize Hidden Markov Models based on alignments from the PFAM database (13) also identified the transglutaminase domain in the yeast Rad4 and PNGases (E-values: 10^{-4}–10^{-3}), but not in XP-C or other RAD4 orthologs. The presence of the transglutaminase fold in these proteins was further confirmed by carrying out sequence–structure threading using the hybrid fold recognition method (14), with the yeast Rad4 as a query. This resulted in the detection of PDB:1FIE as the best hit. The transglutaminase superfamily includes, in addition to the well-characterized transglutaminases such as the vertebrate clotting-factor XIIIA′, several proteases and many uncharacterized proteins that are found in a broad range of prokaryotes and eukaryotes (12). The majority of the proteins of this superfamily are known or predicted to be active enzymes that utilize a catalytic triad comprised of a histidine, a cysteine and an aspartate (Fig. 1) and resembling the active site of papain-like proteases (12,15). The reactions catalyzed by these enzymes involve either the formation of amides by linking alkylamines to the glutamate side chains of proteins or hydrolysis of peptide bonds in the case of proteases (12,15). Thus, the finding of the transglutaminase fold in the PNGases is consistent with the reaction catalyzed by these enzymes that involves breakage of the amide bond between *N*-acetylglucosamine and an asparagine side chain (10). This is confirmed by the inactivation of the

*To whom correspondence should be addressed. Tel: +1 301 594 2445; Fax: +1 301 480 9241; Email aravind@ncbi.nlm.nih.gov

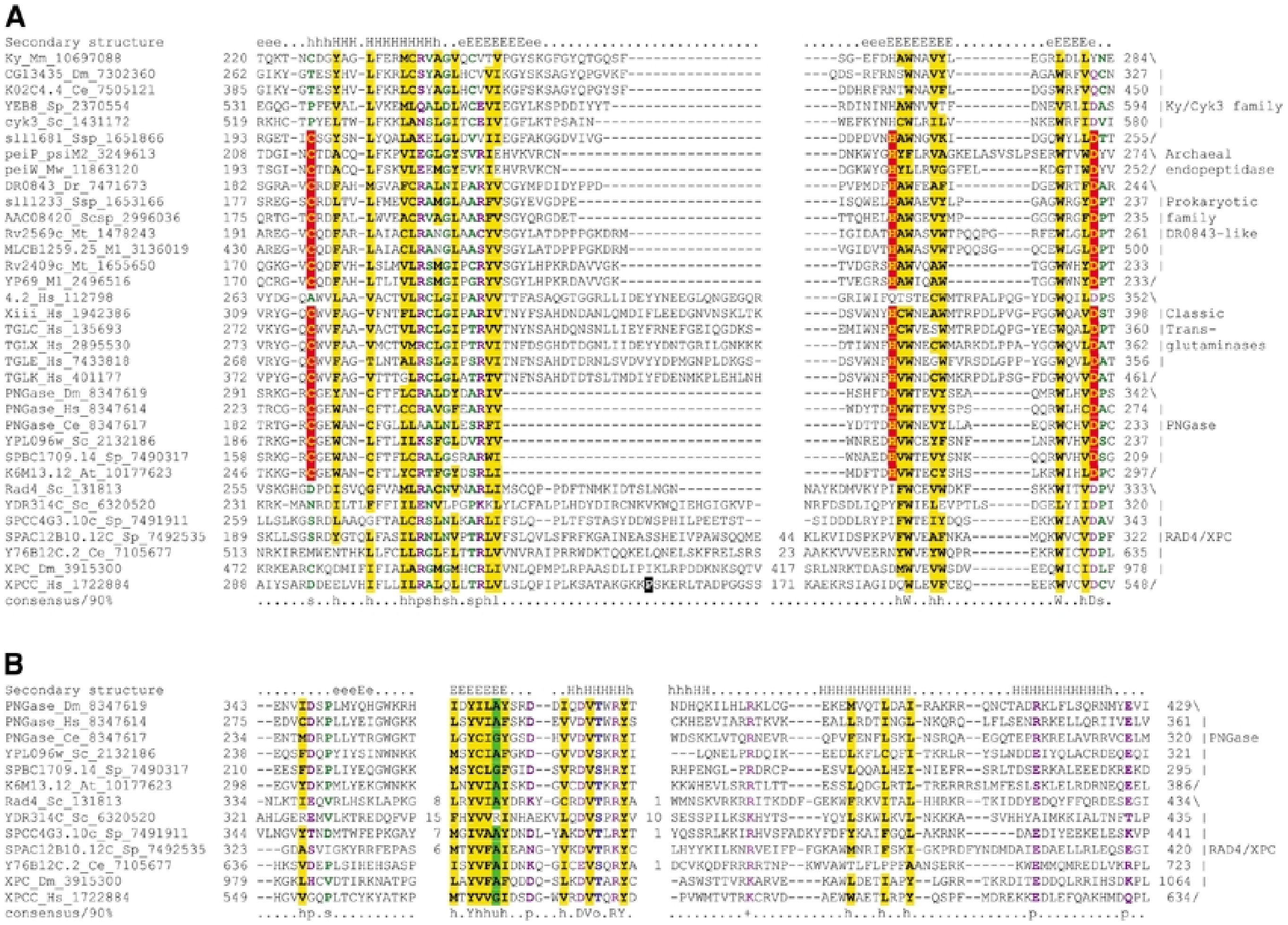

Figure 1. Multiple alignment of the Rad4/XP-C and PNGase sequences with previously identified members of the transglutaminase superfamily. (**A**) Alignment of the transglutaminase core domain. (**B**) Alignment of the C- terminal extension specific to the XP-C/Rad4 and PNGase-like proteins. The different families are denoted on the right. A PSI-BLAST search started with the yeast PNGase (PNG1p) sequence, with a profile inclusion threshold of E = 0.01, revealed a shared conserved region with the Rad4/XP-C proteins (for example, RAD4p was detected in this search with an E-value of 10^{-6} in the second iteration). Further search iterations with this region allowed the detection of several transglutaminase family members including factor XIIIA′ for which a crystal structure is available (E-value 10^{-4} in the seventh iteration). The multiple alignment was constructed using ClustalW, followed by manual adjustment on the basis of PSI-BLAST search results and secondary structure predictions. The secondary structure shown above the alignment is derived from the crystal structure of factor XIIIA′ (PDB: 1FIE) for the transglutaminase core domain and predicted using the PHD program (26) for the C-terminal extension specific to the XP-C/Rad4 and PNGase-like proteins. The 90% consensus shown below the alignment was derived using the following amino acid classes: hydrophobic (h: ALICVMYFW, yellow highlight); the aliphatic subset of these (l: ALIVMC, yellow highlight); alcohol (o: ST, blue), small (s: ACDGNPSTV, green), the 'tiny' subset of these (u: GAS, green highlight), polar (p: CDEHKNQRST, violet), positively charged (+: HKR, pink); charged (c: DEHKR, pink). Black shading indicates the position of mutation in Xeroderma pigme tosum (28,29). The residues of the catalytic triad are shown in reverse shading (shaded in red with yellow letters) in those proteins that retain all three of them. The numbers on each side indicate the limits of the conserved domain in the corresponding protein sequences. The numbers within the alignment are inserts that are not shown. The sequences are denoted by their gene name followed by the species abbreviation and GenBank identifier. The species abbreviations are: Dr, *Deinococcus radiodurans*; Ml, *Mycobacterium leprae*; Mt, *Mycobacterium tuberculosis*; Scsp, *Synechococcus* PCC7002; Ssp, *Synechocystis* PCC6803; Mw, *Methanothermobacter wolfeii prophage psiM100*; PsiM2, *Methanobacterium phage psiM2*; At, *Arabidopsis thaliana*; Ce, *Caenorhabditis elegans*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Sc, *Saccharomyces cerevisiae*; Sp, *Schizosaccharomyces pombe*.

yeast PNGase by a mutation that eliminates the cysteine of the transglutaminase fold catalytic triad (Fig. 1) (10).

The RAD4/XP-C proteins are the closest homologs of the PNGases within the transglutaminase superfamily; these two groups of proteins share a unique feature, a conserved C-terminal extension, to the exclusion of all other members of this superfamily (Fig. 1). All animal Rad4/XP-C proteins and one of the paralogs from *Schizosaccharomyces pombe* contain a large, compositionally biased insert between strands 2 and 3 of the transglutaminase fold (Fig. 1). Examination of the multiple alignment shows that RAD4/XP-C proteins lack the (predicted) catalytic residues, suggesting that these proteins emerged early in the evolution of eukaryotes through a duplication of the PNGase, followed by elimination of the enzymatic activity due to the disruption of the active site triad (Fig. 1). At least two other cases of similar, independent, secondary inactivation of the transglutaminase superfamily proteins following their divergence from active ancestors were noticed, namely the erythrocyte protein-band 4.2 and a highly conserved family of potential cytoskeletal proteins, typified by mouse Ky protein (16) and yeast Cyk3 protein (17) and represented in eukaryotes and cyanobacteria (Fig. 1). Due to the general ability of the ancestral forms of these enzymes to interact with other proteins, some of their inactive descendants probably have been recruited for a non-catalytic interaction function. Notably, the PNGases are required for the early stage of proteasomal degradation of glycoproteins (10), whereas Rad4 has also been shown to interact with the proteasome via the atypical ubiquitin homolog, Rad23 (18,19). Thus, it appears that Rad4/XP-C has evolved from an ancestral *N*-deglycosylase or protease that could have been involved in proteasome-dependent degradation of chromosomal proteins. With the recruitment of the ubiquitin-dependent machinery for a non-proteolytic function in NER, Rad4 could have been recycled to function as an adaptor in protein–protein interactions that are required for this form of repair.

The inactive transglutaminase is the only globular domain (20) in the RAD4/Xp-C proteins, which makes this domain a candidate for the role of the determinant of the specific protein–protein interaction of these proteins. Consistent with this, the portion of RAD4 that encompasses the tranglutaminase domain is required for the interaction with the leucine-rich-repeat containing protein RAD7 (21).

Recruitment of inactivated proteases has been described as one of the evolutionary sources of eukaryotic transcription factors (22). The present observations indicate that evolutionary exaptation (recycling of proteins that have lost their original activity) of the transglutaminase fold for protein–protein interaction has occurred on several independent occasions in contexts as different as NER complex and cytoskeletal organization.

MATERIALS AND METHODS

The non-redundant database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda, MD) was searched using the BLASTP program (11). Profile searches were conducted using the PSI-BLAST program with either a single sequence or an alignment used as the query, with a profile inclusion expectation (E) value threshold of 0.01, and were iterated until convergence (11,23). Previously known, conserved protein domains were detected using the corresponding position-specific scoring matrices, which were constructed using PSI-BLAST (24). Multiple alignments of protein sequences were constructed using the ClustalW program (25) and protein secondary structure was predicted using a multiple alignment as the input for the PHD program (26). Structure visualization was done with Swiss PDB Viewer (27). Sequence–structure threading was performed using the hybrid fold prediction method that combines multiple alignment information with secondary structure prediction (14).

REFERENCES

1. Prakash, S. and Prakash, L. (2000) Nucleotide excision repair in yeast. *Mutat. Res.*, **451**, 13–24.
2. Batty, D.P. and Wood, R.D. (2000) Damage recognition in nucleotide excision repair of DNA. *Gene*, **241**, 193–204.
3. Cleaver, J.E., Thompson, L.H., Richardson, A.S. and States, J.C. (1999) A summary of mutations in the UV-sensitive disorders: xeroderma pigmentosum, Cockayne syndrome, and trichothiodystrophy. *Hum. Mutat.*, **14**, 9–22.
4. Aravind, L., Walker, D.R. and Koonin, E.V. (1999) Conserved domains in DNA repair proteins and evolution of repair systems. *Nucleic Acids Res.*, **27**, 1223–1242.
5. Eisen, J.A. and Hanawalt, P.C. (1999) A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat. Res.*, **435**, 171–213.
6. Yokoi, M., Masutani, C., Maekawa, T., Sugasawa, K., Ohkuma, Y. and Hanaoka, F. (2000) The xeroderma pigmentosum group C protein complex XPC-HR23B plays an important role in the recruitment of transcription factor IIH to damaged DNA. *J. Biol. Chem.*, **275**, 9870–9875.
7. Jansen, L.E., Verhage, R.A. and Brouwer, J. (1998) Preferential binding of yeast Rad4.Rad23 complex to damaged DNA. *J. Biol. Chem.*, **273**, 33111–33114.
8. Guzder, S.N., Sung, P., Prakash, L. and Prakash, S. (1998) Affinity of yeast nucleotide excision repair factor 2, consisting of the Rad4 and Rad23 proteins, for ultraviolet damaged DNA. *J. Biol. Chem.*, **273**, 31541–31546.
9. Wang, Z., Wei, S., Reed, S.H., Wu, X., Svejstrup, J.Q., Feaver, W.J., Kornberg, R.D. and Friedberg, E.C. (1997) The RAD7, RAD16, and RAD23 genes of *Saccharomyces cerevisiae*: requirement for transcription-independent nucleotide excision repair *in vitro* and interactions between the gene products. *Mol. Cell Biol.*, **17**, 635–643.
10. Suzuki, T., Park, H., Hollingsworth, N.M., Sternglanz, R. and Lennarz, W.J. (2000) PNG1, a yeast gene encoding a highly conserved peptide:*N*-glycanase. *J. Cell Biol.*, **149**, 1039–1052.
11. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
12. Makarova, K.S., Aravind, L. and Koonin, E.V. (1999) A superfamily of archaeal, bacterial, and eukaryotic proteins homologous to animal transglutaminases. *Protein Sci.*, **8**, 1714–1719.
13. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
14. Fischer, D. (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pac. Symp. Biocomput.*, 119–130.
15. Pedersen, L.C., Yee, V.C., Bishop, P.D., Le Trong, I., Teller, D.C. and Stenkamp, R.E. (1994) Transglutaminase factor XIII uses proteinase-like catalytic triad to crosslink macromolecules. *Protein Sci.*, **3**, 1131–1135.
16. Blanco, G., Coulton, G.R., Biggin, A., Grainge, C., Moss, J., Barrett, M., Berquin, A., Marechal, G., Skynner, M., van Mier, P. *et al.* (2001) The kyphoscoliosis (ky) mouse is deficient in hypertrophic responses and is caused by a mutation in a novel muscle-specific protein. *Hum. Mol. Genet.*, **10**, 9–16.
17. Korinek, W.S., Bi, E., Epp, J.A., Wang, L., Ho, J. and Chant, J. (2000) Cyk3, a novel SH3-domain protein, affects cytokinesis in yeast. *Curr. Biol.*, **10**, 947–950.
18. Masutani, C., Sugasawa, K., Yanagisawa, J., Sonoyama, T., Ui, M., Enomoto, T., Takio, K., Tanaka, K., van der Spek, P.J., Bootsma, D. *et al.* (1994) Purification and cloning of a nucleotide excision repair complex involving the xeroderma pigmentosum group C protein and a human homologue of yeast RAD23. *EMBO J.*, **13**, 1831–1843.
19. Schauber, C., Chen, L., Tongaonkar, P., Vega, I., Lambertson, D., Potts, W. and Madura, K. (1998) Rad23 links DNA repair to the ubiquitin/proteasome pathway. *Nature*, **391**, 715–718.
20. Wootton, J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **18**, 269–285.
21. Wei, S. and Friedberg, E.C. (1998) A fragment of the yeast DNA repair protein Rad4 confers toxicity to *E. coli* and is required for its interaction with Rad7 protein. *Mutat. Res.*, **400**, 127–133.
22. Aravind, L. and Koonin, E.V. (1998) Eukaryotic transcription regulators derive from ancient enzymatic domains. *Curr. Biol.*, **8**, R111–R113.

23. Aravind, L. and Koonin, E.V. (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.*, **287**, 1023–1040.
24. Schaffer, A.A., Wolf, Y.I., Ponting, C.P., Koonin, E.V., Aravind, L. and Altschul, S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
25. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
26. Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.
27. Guex, N. and Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
28. Chavanne, F., Broughton, B.C., Pietra, D., Nardo, T., Browitt, A., Lehmann, A.R. and Stefanini, M. (2000) Mutations in the *XPC* gene in families with xeroderma pigmentosum and consequences at the cell, protein, and transcript levels. *Cancer Res.*, **60**, 1974–1982.
29. Li, L., Bales, E.S., Peterson, C.A. and Legerski, R.J. (1993) Characterization of molecular defects in xeroderma pigmentosum group C. *Nat. Genet.*, **5**, 413–417.