Volume 283, number 2, 203–206 FEBS 09754 June 1991

ADONIS 0014579391004631F

The yeast DNA repair proteins RAD1 and RAD7 share similar putative functional domains

Rainer Schneider and Manfred Schweiger

Institut für Biochemie, Universität Innsbruck, Peter Mayr Str. 1a, A-6020 Innsbruck, Austria

Received 28 March 1991

Sequence information on eukaryotic DNA repair proteins provided so far only few clues concerning possible functional domains. Since the DNA repair process involves a strict sequential complex formation of several proteins [(1988) FASEB J. 2, 2696–2701], we searched for special protein-protein interacting domains, which consist of tandemly repeated leucine rich motifs (LRM). Search algorithms, capable of detecting even largely divergent repeats by assessing their significance due to the tandem repetitivity, revealed that the yeast DNA repair proteins RAD1 and RAD7 contain 9 and 12 tandem LRM repeats, respectively. These results represent the first clues concerning specific domains in these proteins and assign them to the LRM superfamily, which includes such members as yeast adenylate cyclase, cell surface protein receptors and ribonuclease/angiogenin inhibitor, all exerting their function by specific protein-protein interactions involving LRM domains [(1988) EMBO J. 7, 4151–4156; (1990) Proc. Natl. Acad. Sci. USA 87, 8711–8715; (1989) Science 245, 494–499; (1990) Mol. Cell. Biol. 10, 6436–6444; (1989) Proc. Natl. Acad. Sci. USA 86, 6773–6777].

DNA repair; RAD1; RAD7; Leucine-rich motif; Protein-protein interaction; Superfamily

DNA repair is a complex biochemical process requiring the precise coordination of several protein-protein and protein-DNA interactions [1]. The excision of bulky adducts from DNA, referred to as nucleotide excision repair, has been extensively studied in *E. coli*, where the products of five genes are involved in this repair process (uvrA, uvrB, uvrC, uvrD and polA). These proteins have to interact in a strict sequential manner to recognize the damage (uvrA + uvrB) and incise the afflicted DNA strand (uvrAB + uvrC). Additionally, the uvrD and polA proteins are required for postincision events that include turnover of the uvrABC protein complex, release of the damaged stretch of nucleotides and repair synthesis. Eukaryotic systems are more complex and less well characterized. In the yeast *Saccharomyces cerevisiae* several genes have been identified as having a role in nucleotide excision repair [7], 6 of which (RAD1, RAD2, RAD3, RAD4, RAD10 and MMS19) are absolutely required for damage-specific excision of DNA. Mutants in the RAD7, RAD14, RAD16 and RAD23 genes show partial incision defectiveness. Although several of the yeast DNA repair genes have been cloned and sequenced, only few clues concerning the nature and function of the corresponding proteins could be deciphered from their predicted amino acid sequences. Only in two cases, namely RAD6 and RAD3, was it possible to identify obvious homologies and sequence motives suggesting certain functions: RAD6 [8] and its human homologue [9] were shown to have an ubiquitin conjugating activity and RAD3 [10] and its putative human homologue [11] were shown to have similarities to DNA helicases.

The fact that no biochemical function has been identified for most of the cloned eukaryotic DNA repair proteins, is not totally unexpected, since purified *E. coli* uvrB and uvrC proteins have no catalytic activity and uvrB protein alone does not bind to damaged DNA. However, this latter protein has a high affinity for purified uvrA protein, and it has recently been suggested that when bound to DNA an uvrBC protein complex constitutes a catalytically active endonuclease [12]. Thus the interaction of specific proteins seems to be essential to assess their catalytic function. Therefore we searched for special protein-protein interacting domains in the sequences of DNA repair proteins. Since no obvious structural motifs like EGF-, kringle- or complement repeats could be identified, we analyzed the sequences for less conserved structures, namely the leucine-rich motifs (LRM) [2]. These motifs are composed of a leucine-rich amino acid consensus sequence: $x\alpha\alpha xxLxxLxxLx\alpha xx^{N}/_{C}xLxx\alpha xx$ (α means hydrophobic amino acid) and have been found multiply repeated in a vast variety of proteins, constituting the LRM superfamily. The LRM domains of these proteins seem to be involved in tight and specific protein-protein interactions. The identification of the proteins of this superfamily is often hampered by a very weak conservation of each motif, which can lead to an overall conser-

Correspondence address: R. Schneider, Institut für Biochemie, Universität Innsbruck, Peter Mayr Str. 1a, A-6020 Innsbrück, Austria. Fax: (43) (512) 507 2419.

vation of the LRM domain of less than 25%. Thus without the consideration of the tandem repetitivity, which helps to recognize the significance of the motifs, relationships to this family might easily be overlooked. Therefore we devised search algorithms that specifically recognize the repeated consensus sequences, even if they have been largely diverged by conservative amino acid substitutions or gaps and insertions. For our analysis we used the program FIND of the GCG program package [13] and applied two specific search patterns, which take into account that there are two groups of LRM repeats; one has a length of 22–25 amino acids and has a highly conserved asparagine at position 17 of its consensus sequence, the other, occurring in the module B type of ribonuclease/angiogenin inhibitor (RAI) is about 5 amino acids longer and shows substitution of the conserved asparagine by cysteine [2] (Fig. 1b). Additionally, the algorithms allow for gaps/insertions at positions which show variations in some family members (Fig. 1b). The algorithms practically compare a search pattern, representing the core conserved residues of two tandemly arranged repeats, against all known protein sequences. The used search patterns were A: $LX_{(2-3)}LXXXXXNX_{(12-15)}LX_{(2-3)}$-LXXXXN and B:$LX_{(2-3)}LXL^*XXCX_{(17-21)}$-$LX_{(2-3)}LXL^*$ where L* means either the amino acids I, V, L or F. The tandem repeats are, due to gap, length and amino acid variations, highly degenerate, therefore these algorithms might select false proteins by chance. Thus we further strengthened the criteria by only accepting proteins as family members whose sequences can be matched at least twice with the search patterns and the matches must lie about 24n amino acids apart. This means that we only consider proteins, that contain at least 3 tandem repeats ($n = 1$) or 2 times 2 tandem repeats ($n > 2$) separated by a distance that corresponds to the length of one or several repeats which are too largely diverged to be recognized by the chosen search patterns.

Applying these algorithms to the PIR protein database (release 26), which contains about 26,000 entries, results in the selection of just 14 proteins meeting all criteria. Algorithm A selected 10 accepted members of the family and two novel possible candidates, one of which indeed is a DNA repair protein, namely RAD1 from *S. cerevisiae*. An alignment of the tandem repeats found in RAD1 (Fig. 1a) shows several additionally conserved positions and 6 less conserved tandem repeats, which were not detected by the chosen algorithm. Furthermore, the N-terminal region of the proposed LRM domain shows similarities to the corresponding N-terminally conserved regions of several family members [6]. The C-terminal region is followed by a cysteine containing segment, which is also typical for this family. This assigns RAD1 to the proteins containing an LRM domain. Additionally, the region contains 15% leucines, 80% of which occur at positions corresponding with the periodic pattern typical for leucine-rich repeats. A prediction of the secondary structure and hydrophobicity profile of RAD1 also suggest that the LRM structures form a distinct domain in this protein (not shown).

Algorithm B selected 2 proteins: RAI, as of course expected, and also a DNA repair protein, RAD7 of *S. cerevisiae*. For the latter protein the algorithm finds 3 matches 26 and 136 amino acids apart, which would account for at least 5 LRM repeats. Further analysis of the RAD7 amino acid sequence actually revealed 12 tandem repeats of about 27 amino acids each, showing many additional conserved positions (Fig. 1c). Thus, RAD7 can clearly be assigned to the LRM protein family, too. Interestingly, both DNA repair proteins show a closer similarity to the consensus sequences of the RAI modules, than to most of the other family members. RAI has been discovered 25 years ago [16] and has since then been widely used in molecular biology. This protein consists almost entirely of LRM repeats and the protein inhibits ribonucleases by the formation of a 1:1 complex of extraordinary stability. Thus we propose that the LRM domains now located in RAD1 and RAD7 also have some role in the interaction with other proteins, namely other components of the DNA repair machinery. In this respect it should be noted that variations of the cellular levels of RAI have been reported after whole body γ-irradiation of rats [17] and that an involvement of RAI in cellular repair has been discussed before [18].

Interestingly, a deletion in the RAD7 gene of the first 99 codons, which represent a major portion of the hydrophilic region directly preceding the proposed LRM domain, retains complementation ability in RAD7 deletion strains, which is not observed in strains deleted for both RAD7 and RAD23 [15]. One possible interpretation of these results is that an interaction between the RAD7 and RAD23 proteins could account for the activity or the stability or both of the amino-terminally truncated RAD7 protein. The LRM domain seems to be a good candidate for a site of interaction between RAD7 and RAD23. Mutations of residues highly conserved in the LRM repeats should therefore interfere with the subserving action of RAD23. The repetitive domain of RAD7 also shows a high similarity to the LRM domain in the trypanosome leucine-rich repeat protein [5]. A rather basic region preceding the LRM domain in this parasite protein resembles the metal binding domains of nucleic acid binding proteins, known as zinc fingers. Interestingly, in RAD7 the LRM domain is preceded by a highly basic region, too. Since this basic region is not deleted in the active N-terminally truncated RAD7, it could be involved in DNA-protein interactions necessary for the RAD7 activity.

No possible protein-protein interactions of RAD1 are known thus far [14], but the localization of a putative protein binding domain will help to elucidate possible

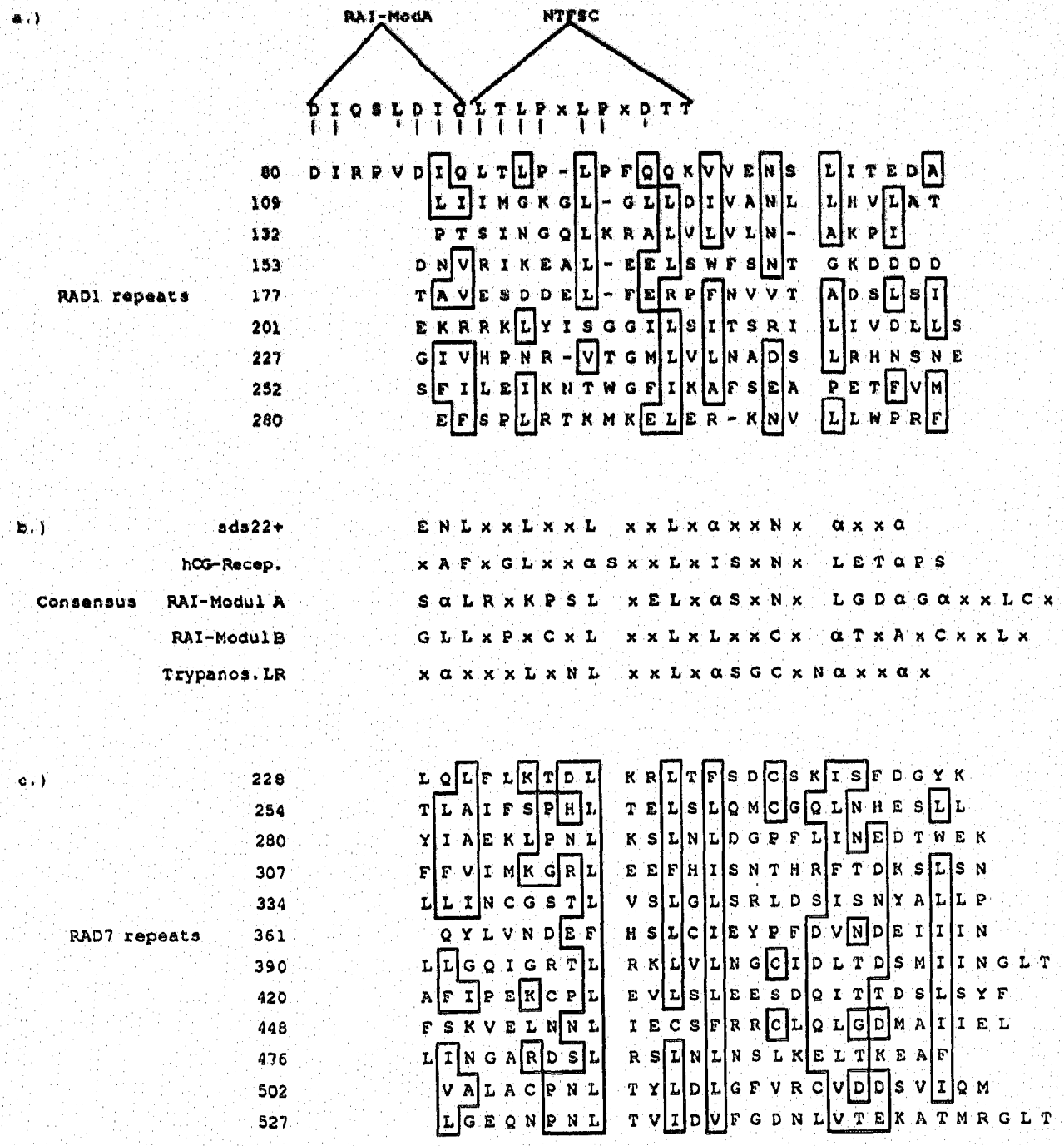

Fig. 1. Alignment of the leucine-rich repeats of RAD1 and RAD7 with representative consensus sequences of the LRM family. (a) Alignment of the 9 leucine-rich repeats in RAD1. Positions showing conservation to the family consensus sequences are boxed (amino acid similarity groups: AFLVIM, HKR, NDEQ, ST, PG, WY, C). Additional similarities to the LRM family are shown in the upper part: RAI-ModA represents a segment found in the N-terminal RAI module A and NTFSC means the 'N-terminal flanking sequence consensus' found in several family members at a similar position as in RAD1 [6] (vertical lines mean identity (long) or similarity (short) between these sequences and RAD1). Several of the repeats show an amino acid insertion similar to the one in the consensus sequence of the LRM domain in the human choriogonadotropin receptor. (b) Five representative consensus sequences of the LRM family, showing the positions of possible insertions and amino acid variations (α stands for hydrophobic amino acids). sds22+: modulator of protein phosphatase 1 in *S. pombe* [26], hCG-recep.: human choriogonadotropin receptor [4], RAI-Modul A,B: the consensus sequences of the two different repeat types of RAI [2], Trypanos. LR: trypanosome leucine-rich repeat gene [5]. (c) Alignment of the 12 leucine-rich repeats found in RAD7. All the repeats show the same insertion of one amino acid as the consensus sequence of the trypanosome leucine-rich repeat gene. In addition to the boxed residues conserved according to the consensus sequences there are in both repeat alignments several positions which also show conservation, but only within the repeat alignment of each protein, thus supporting the repetitive character of these domains. Numbering of the amino acids is according to the published sequences. For topological reasons 2 short insertions in the sequence alignments are not shown: YRS at position 257 of RAD1 and NEE at position 377 of RAD7.

coordinations with other DNA repair components. The fact that the RAD1 and RAD7 proteins do contain a similar structural domain does not necessarily implicate that they have derived from the same progenitor by divergent evolution. Since the primordial ancestor(s) of the domains was (were) rather short and probably showed distinct differences, they might have evolved independently. Nevertheless it is very likely that the found similarity of the present repetitive domains reflects a structural and functional relationship because, despite the apparent high variability between individual repeats, the few conserved positions seem to confer a rigid stereo structure necessary for functional integrity, as has recently been shown by a mutational analysis of the interaction of the LRM domain of adenylate cyclase with RAS proteins [3].

The finding of repetitive motifs not only helps to understand the domain structure and evolution of proteins, but also facilitates the detection of very distant relationships. Thus, our results should be useful to identify homologues of RAD1 and RAD7 in distant organisms as soon as their sequences are available. Major efforts are presently put into the identification and cloning of DNA repair genes that are involved in various inheritable human cancer prone diseases [18-20], either by classical genetic complementation [11,21,22] or reverse genetics [9,23-25]. Informative homologies between human, yeast and *E. coli* DNA repair proteins might be overlooked without consideration of highly sensitive search algorithms, because distant relationships are often lying in the 'twilight zone' of protein sequence homology (<25% amino acid conservation). Therefore this method of searching ancient repetitive structures in proteins seems to be a generally applicable powerful tool; especially in research fields dealing with a rapidly increasing number of protein sequences of unknown function.

REFERENCES

[1] Grossman, L., Caron, B., Mazur, S.J. and Oh, E.Y. (1988) FASEB J. 2, 2696-2701.
[2] Schneider, R., Schneider-Scherzer, E., Thurnher, M., Auer, B. and Schweiger, M. (1988) EMBO J. 7, 4151-4156.
[3] Suzuki, N., Choe, H., Nishida, Y., Yamawaki-Kataoka, Y., Ohnishi, S., Tamaoki, T. and Kataoka, T. (1990) Proc. Natl. Acad. Sci. USA 87, 8711-8715.
[4] McFarland, K.C., Sprengel, R., Phillips, H.S., Köhler, M., Rosemblit, N., Nikolics, K., Segaloff, D.L. and Seeburg, P.H. (1989) Science 245, 494-499.
[5] Smiley, B.L., Stadnyk, A.W., Myler, P.J. and Stuart, K. (1990) Mol. Cell. Biol. 10, 6436-6444.
[6] Hickey, M.J., Stuart, A.W. and Roth, G.J. (1989) Proc. Natl. Acad. Sci. USA 86, 6773-6777.
[7] Friedberg, E.C. (1988) Microbiol. Rev. 52, 70-102.
[8] Jentsch, S., McGrath, J.P. and Varshavsky, A. (1987) Nature 329, 131-134.
[9] Schneider, R., Eckerskorn, C., Lottspeich, F. and Schweiger, M. (1990) EMBO J. 9, 1431-1435.
[10] Sung, P., Prakash, L., Matson, S.W. and Prakash, S. (1987) Proc. Natl. Acad. Sci. USA 84, 8951-8955.
[11] Weber, C.A., Salazar, E.P., Stewart, S.A. and Thompson, L.H. (1990) EMBO J. 9, 1437-1447.
[12] Sancar, A. and Sancar, G.B. (1988) Annu. Rev. Biochem. 57, 29-67.
[13] Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. 12, 387-395.
[14] Reynolds, P., Prakash, L. and Prakash, S. (1987) Mol. Cell. Biol. 7, 1012-1020.
[15] Perozzi, G. and Prakash, S. (1986) Mol. Cell. Biol. 6, 1497-1507.
[16] Traub, P., Zillig, W., Millette, R.L. and Schweiger, M. (1966) Biol. Chem. Hoppe-Seyler 343, 261-275.
[17] Ferencz, A., Hidvegi, E.J., Szabo, L.D. and Varteresz, V. (1973) Radiat. Res. 55, 304-317.
[18] Schweiger, M., Schneider, R., Hirsch-Kauffmann, M., Auer, B., Klocker, H., Scherzer, E., Thurnher, M., Herzog, H., Vosberg, J.P. and Wagner, E.F. (1988) in The Roots of Modern Biochemistry (Kleinkauf, H., vanDöhren, H. and Jaenicke, L., eds.) pp. 379-387, De Gruyter, Berlin.
[19] Schweiger, M., Auer, B., Burtscher, H.J., Hirsch-Kauffmann, M., Klocker, H. and Schneider, R. (1987) Eur. J. Biochem. 165, 235-242.
[20] Friedberg, E.C., Backendorf, C., Burke, J., Collins, A., Grossman, L., Hoeijmakers, J.H.J., Lehmann, A.R., Seeberg, E., Van der Schans, G.P. and Van Zeeland, A.A. (1987) Mutation Res. 184, 67-86.
[21] Weeda, G., Van Ham, R.C.A., Vermeulen, W., Bootsma, D., van der Eb, A.J. and Hoeijmakers, J.H.J. (1990) Cell 62, 777-791.
[22] Tanaka, K., Miura, N., Satokata, I., Miyamoto, I., Yoshida, M.C., Satoh, Y., Kondo, S., Yasui, A., Okayama, H. and Okada, Y. (1990) Nature 348, 73-76.
[23] Schneider, R., Auer, B., Kühne, C., Herzog, H., Klocker, H., Burtscher, H.J., Hirsch-Kauffmann, M., Wintersberger, U. and Schweiger, M. (1987) Eur. J. Cell. Biol. 44, 302-307.
[24] Herzog, H., Zabel, B.U., Schneider, R., Auer, B., Hirsch-Kauffmann, M. and Schweiger, M. (1989) Proc. Natl. Acad. Sci. USA 86, 3514-3518.
[25] Auer, B. Nagl, U., Herzog, H., Schneider, R. and Schweiger, M. (1989) DNA 8, 575-580.
[26] Ohkura, H. and Yanagida, M. (1991) Cell 64, 149-157.