MOLECULAR GENETICS AND METABOLISM **63,** 230–234 (1998)
ARTICLE NO. GM972652

BRIEF COMMUNICATION

The Fanconi Anemia Complementation Group A Protein Contains a Peroxidase Domain[1]

Computational analysis of the Fanconi anemia (FA) complementation group A protein suggests that it contains a peroxidase domain. FA proteins may be part of a general mechanism that protects cells from oxidative damage.

Key Words: **Fanconi anemia; FAA protein; hidden Markov model; peroxidase; oxidative damage.**

Fanconi anemia (FA) is an autosomal recessive disorder characterized by chromosomal instability, developmental abnormalities, bone marrow failure, and increased risk of malignancy (reviewed in (1)). FA patients exhibit increased sensitivity to ionizing radiation, DNA cross-linking and alkylating agents, and selected chemotherapeutic compounds, suggesting that FA is a disorder of DNA repair. Lymphoblasts from FA patients exhibit increased formation of 8-hydroxydeoxyguanosine, a frequently occuring and highly mutagenic base modification that is a major product of hydroxyl radical damage to DNA (2). Among the FA complementation groups, subtype A is the most prevalent and its protein (FAA) has been cloned recently (3,4). The apparent lack of similarity of FAA to any other protein leads to the suggestion that the FA gene pathway is novel and unrelated to known cellular mechanisms of defense or repair. The results here indicate that FAA possesses a well-studied peroxidase domain and that one of its functions may be as part of a complex that serves as a general defense mechanism protecting cells from DNA or protein cross-linking damage. These observations may yield insights not only into FA but also into organismal development such as hemopoietic cell growth and differentiation.

Peroxidases are heme-containing enzymes that use hydrogen peroxide (H_2O_2) as the electron acceptor to catalyze a number of oxidative reactions and often provide protection under conditions of oxidative stress (reviewed in (5–8)). Two superfamilies have been defined: animal heme (AH) peroxidases (comprising myeloperoxidase, eosinophil peroxidase, lactoperoxidase, thyroid peroxidase, prostaglandin H synthase, peroxidasin) and fungal, plant, and bacterial heme (FPBH) peroxidases (yeast cytochrome *c* peroxidase, ascorbate peroxidase, bacterial catalase peroxidases, lignin peroxidase, manganese peroxidase, secretory plant peroxidases). Crystal structures of several FPBH peroxidases have been determined and reveal a fold in which two antiparallel α-helices from two α-domains form a crevice in which the heme prosthetic group is inserted. The proximal heme iron pocket is formed by a His (the proximal heme iron ligand), an Asp, and a large hydrophobic residue (Trp, Phe, Leu). The distal H_2O_2 binding pocket is formed by a His (the distal heme iron ligand), an Arg, and an aromatic residue (Trp, Phe). Examination of the FPBH structures permits rationalization of some of the known FAA mutations.

MATERIALS AND METHODS

BLAST database searches (9) were performed using default parameters and FAA (1455 residues) as the query sequence. An *Escherichia coli* GUG start protein (235 residues databank code ECU28375) had the highest scoring segment pair(s) but none were considered to be statistically significant ($P = 0.45$). Given the importance of oxidative damage in the

[1] The submitted manuscript has been authored by a contractor of the U.S. Government under Contract No. DE-AC03-76SF00098. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

1096-7192/98 $25.00

etiology of FA, the resemblance of the conserved sequence [SG]GHTLG in the region of greatest similarity (GUG: 109–172, FAA: 281–344) to the proximal heme iron pocket of peroxidases was pursued. A statistical model, a hidden Markov model (HMM) of a peroxidase domain was trained as described elsewhere (10,11) using a training set composed of FAA, *E. coli* GUG protein, AH peroxidases, and FPBH peroxidases. HMMs (12,13) have been used recently to characterize the common features of a family of related sequences, to generate a multiple sequence alignment, and to recognize related, but divergent, family members present in databases (14–22). HMMs used to model sequence families can be viewed as "profiles" recast in a probabilistic framework. A profile is a model for a family consisting of a primary sequence consensus and position-specific residue scores and insertion/deletion penalties (23–27).

HMM analyses were performed with the SAM suite (14,28) running on a MASPAR MP-2201 with a DEC Alpha 3000/300X frontend at the University of California Santa Cruz (UCSC). An initial peroxidase HMM was created such that only the alignment of the pentapeptide containing the proximal heme iron ligand was fixed. Subsequent training revealed the emergence of additional conserved positions, including the H_2O_2 pocket containing the distal His heme iron ligand. Examination of HMM-generated alignments suggested that the GUG protein lacked the amino-terminal region and that the putative FAA peroxidase domain showed little similarity to the AH peroxidases. Thus, further HMM training only utilized FAA and FPBH peroxidases. HMM database searches using a nonredundant protein database obtained from the NCI (29) and updated weekly at UCSC did not reveal any sequences other than FPBH peroxidases with significant log-odds (30,31) scores at the 0.01 level.

RESULTS

Figure 1 shows the HMM-generated alignment for the peroxidase domain in FAA and representative FPBH peroxidases (primarily those for which three-dimensional structures are known). Of the 261 positions in the alignment, only 11 (4.2%) are conserved between FAA and FPBH peroxidases (bold) together with an additional 32 (12.3%) positions that maintain their hydrophobic nature (boxed). Despite this low level of sequence identity, a number of lines of evidence support the proposition that the putative peroxidase domain in FAA is structurally and functionally related to the FPBH peroxidase domain. Inspection of the HMM-generated alignment indicates that three of the six active site residues of FPBH peroxidases are present in FAA: the His heme iron ligands (C, D) and the Trp of the distal H_2O_2 binding pocket (B). Although the Arg at A of the distal pocket is changed to Glu, this seemingly drastic charge reversal is tolerable because change of this Arg to Glu in *S. cerevisiae* cytochrome *c* peroxidase (Sc_CCPR) only has a minor effect on the steady-state kinetics of the enzyme (32). The Thr at F would be capable of forming a hydrogen bond to the proximal His (D) in the same manner as the buried Asp in FPBH peroxidases. The His at E is a position that is variable in the FPBH peroxidases. Of the nonactive site residues that are conserved (green labeled a–i), Fig. 2 shows that they are located in the core of the domain. The Asn at b that forms a hydrogen bond to the distal His (C) is conserved (although absent in At_PX and Gh_PX). Comparison of the structures and alignments suggests that Ps_APX1 represents the minimal peroxidase domain.

Analysis of FAA mutations (33) and secondary structure predictions provides additional support for the presence of a peroxidase domain in FAA. RNA splice site mutations IVS7 + 5G → A and IVS7 + 5G → T result in the change Asp237 → Gly and insertion of AFMTR-CGFLD. Asp 237 corresponds to a position that is conserved in all the peroxidase domains (c in Figs. 1 and 2). Although FPBH peroxidases have been the subject of extensive studies, the site of this FAA mutation highlights the potential importance of the conserved Asp at c that is distant from the heme pocket. Mutation 894–1006del causes a premature stop codon two residues downstream of Cys 297. Cys 297 corresponds to the residue in column 163 so that such a deletion would destroy the proximal heme iron pocket. The known FPBH peroxidase structures are predominantly helical. Apart from a central region around the proximal heme iron pocket, secondary structure predictions for FAA suggest a similar structure. Phylogenetic analysis (data not shown) indicate this FAA peroxidase domain is most similar to the cytosolic metazoan peroxidases and bacterial catalase peroxidases and more distantly related to the fungal ligninases and manganese peroxidase.

CONCLUSIONS

The results here suggest that FAA possesses a domain that is structurally and functionally related to that present in intracellular FPBH peroxidases. These observations are consistent with recent experiments

FIG. 1. An HMM-generated multiple sequence alignment of the peroxidase domain in (representative) FPBH peroxidases and FAA. Amino acids conserved in at least five sequences including FAA are in bold and columns that are predominantly hydrophobic are boxed. Columns containing "." correspond to insert states and numbers indicate the lengths of insertions in sequences at that position (if present). The sequences are Hs_FAA, *Homo sapiens* Fanconi anemia complementation group A protein (databank code HSCH16FAA); Ps_APX1, *Pisum sativum* cytosolic L-ascorbate peroxidase (APX1_PEA); So_APX, *Spinacia oleracea* stromal ascorbate peroxidase (SPICPSAP); Sc_CCPR, *Saccharomyces cerevisiae* mitochondrial electron transport chain cytochrome *c* peroxidase precursor (CCPR_YEAST); Ss_HPI, *Synechocystis* PCC6803 catalase hpi (D90910); At_PX, *Arabidopsis thaliana* peroxidase ATP17a (ATPO1); Gh_PX, *Gossypium hirsutum* peroxidase (COTPROXDS); Pc_LIG2, *Phanerochaete chrysosporium* ligninase LG2 precursor (LIG2_PHACH); Pc_PERM, *P. chrysosporium* peroxidase manganese-dependent I precursor (PERM_PHACH); Ar_PER, *Arthromyces ramosus* peroxidase precursor (PER_ARTRA). Cylinders and arrows mark the positions of the α-helices and β-strand taken from the CATH database (36): Ps_APX1 (PDB entry 1APX, chain A), Sc_CCPR (1CPG), Pc_LIG2 (1LGA, chain A), Pc_PERM (1MNP), and Ar_PER (1ARU). Residues forming the distal H_2O_2 binding pocket (A–C) and proximal heme iron pocket (D–F) are labeled. a–i indicate residues conserved between FAA and FPBH peroxidases. The predicted secondary structures for the FAA peroxidase domain using two different methods, PHD (37) and DSC (38), are shown.

(34) which indicate that FAA is a cytosolic and not a nuclear protein that possesses nonfunctional nuclear localization and leucine-zipper domains. In conjunction with the cytoplasmic location of FAA and the complementation group C protein (FAC), it seems likely therefore that FA proteins are part of a general

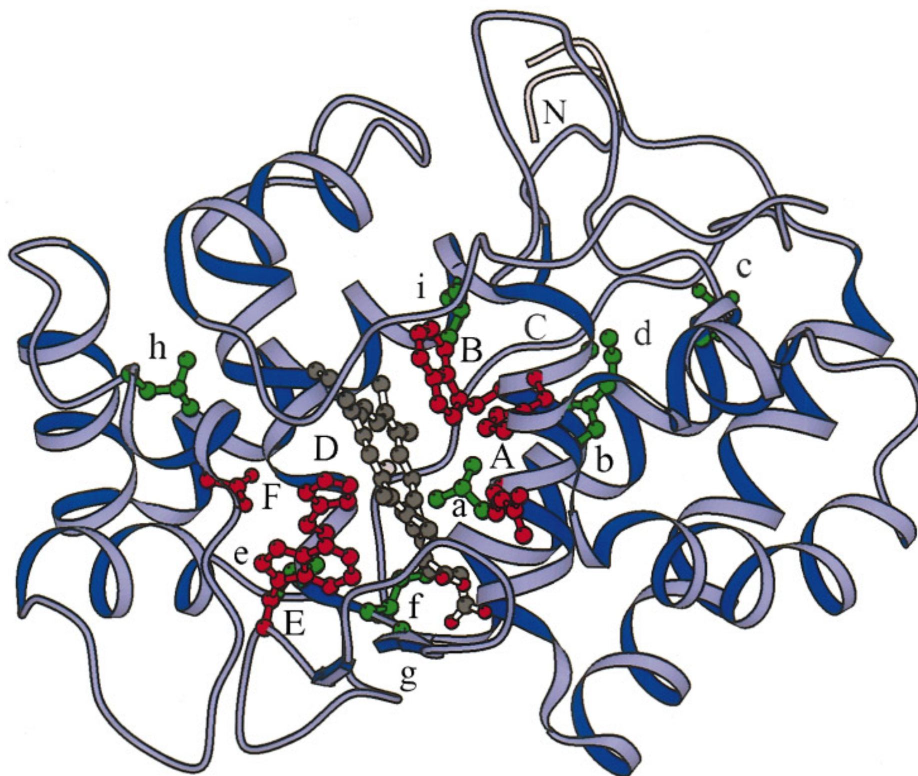

FIG. 2. Ribbon diagram of the peroxidase Ps_APX1 shown in the HMM-generated alignment (Fig. 1). Regions in grey are outside the peroxidase domain modeled here. The heme moiety is shown in ball-and-stick form. Residues in red (labeled A–F) comprise the distal H_2O_2 binding pocket and proximal heme iron pocket. Residues in green are additional conserved residues. N designates the amino-terminus of the protein.

mechanism that prevents oxidative damage to cells. Further support for this notion comes from preliminary results which suggest that FAC may possess a catalase domain (I.S.M., unpublished). Thus, the myriad of symptoms presented by FA patients may be a consequence of a breakdown in this protective mechanism rather than originating from a more direct defect in some function of the individual FA proteins.

The nature of the substrate for FAA is unknown but given the similarity of the peroxidase domain to higher plant ascorbate peroxidases, one possibility is ascorbate. Ascorbate is present at high concentrations at sites where oxidative stress is likely to be the most intense, modulates cell growth and differentiation, and can reduce or stimulate the growth of tumor cells (reviewed in (35)). Finally, heme is a constituent of various proteins essential for the function of all living cells. FAA joins this group of enzymes which, while performing completely different chemistry, share features of the iron heme active site: a protophorphyrin IX prosthetic group linked to the protein by a proximal His.

ACKNOWLEDGMENTS

We thank Huber Warner for bringing FAA to our attention. This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Division of the US Department of Energy under Contract No. DE-AC03-76SF00098. The data and multiple alignments are available in electronic form upon request.

REFERENCES

1. D'Andrea A. Fanconi anaemia forges a novel pathway. *Nature Genet* **14:**240–242, 1996.
2. Takeuchi T, Morimoto K. Increased formation of 8-hydroxydeoxyguanosine, an oxidative DNA damage, in lymphoblasts from Fanconi's anemia patients due to possible catalase deficiency. *Carcinogenesis* **14:**1115–1120, 1993.
3. Lo Ten Foe J, Rooimans M, Bosnoyan-Collins L, Alon N, Wijker M, Parker L, Lightfoot J, Carreau M, Callen D, Savoia A, Cheng N, van Berkel C, Strunk M, Gille J, Pals G, Kruyt F, Pronk J, Arwert F, Buchwald M, Joenje H. Expression cloning of a cDNA for the major Fanconi anaemia gene, FAA. *Nature Genet* **3:**320–323, 1996.
4. Apostolou S, Whitmore S, Crawford J, Lennon G, Sutherland G, Callen D, Ianzano L, Savino M, Dapolito M, Notarangelo

A, Memeo E, Piemontese M, Zelante L, Savoia A, Gibson R, Tipping A, Morgan N, Hassock S, Jansen S, Deravel T, Vanberkel C, Pronk J, Easton D, Mathew C, Levran O, Verlander P, Batish S, Erlich T, Auerbach A, Cletonjansen A, Moerland E, Cornelisse C, Doggett N, Deaven L, Moyzis R. Positional cloning of the Fanconi anaemia group A gene. *Nature Genet* **3**:324–328, 1996.
5. Degtyarenko K, North A, Findlay J. PROMISE: A new database of information on prosthetic centres and metal ions in protein active sites. *Protein Eng* **10:**183–186, 1997. [The database is available at URL http://bmbsgill.leeds.ac.uk/bmbknd/promise/MAIN.html].
6. Welinder K. Bacterial catalase-peroxidases are gene duplicated members of the plant peroxidase superfamily. *Biochim Biophys Acta* **1080:**215–220, 1991.
7. Li H, Poulos T. Structural variation in heme enzymes: A comparative analysis of peroxidase and P450 crystal structures. *Structure* **2:**461–464, 1994.
8. Everse J, Everse KE, Grisham MB. (Eds.) Peroxidases in Chemistry and Biology. Boca Raton, Fl: CRC Press, 1990.
9. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. *J Mol Biol* **215:**403–410, 1990.
10. Mian I. Sequence analysis of ribonucleases HII, III, II, PH and D. *Nucleic Acids Res* **25:**3187–3195, 1997.
11. Dalgaard J, Moser M, Hughey R, Mian I. Statistical modeling, phylogenetic analysis and structure prediction of a protein splicing domain common to inteins and hedgehog proteins. *J Comp Bio* **4:**193–214, 1997.
12. Rabiner L, Juang B. An introduction to hidden Markov models. *IEEE ASSP Mag* **3:**4–16, 1986.
13. Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* **77:**257–286, 1989.
14. Krogh A, Brown M, Mian I, Sjölander K, Haussler D. Hidden Markov models in computational biology: Applications to protein modelling. *J Mol Biol* **235:**1501–1531, 1994.
15. Eddy S. Hidden Markov models. *Current Opin Struct Biol* **6:**361–365, 1996. [The hidden Markov model software can be accessed at URL http://genome.wustl.edu/eddy/hmm.html]
16. Fujiwara Y, Asogawa M, Konagaya A. Stochastic motif extraction using hidden Markov model. *ISMB* **2:**121–129, 1994.
17. Bateman A, Chothia C. Fibronectin type III domains in yeast detected by a hidden Markov model. *Curr Biol* **6:**1544–1547, 1996.
18. Bateman A, Eddy S, Chothia C. Members of the immunoglobulin superfamily in bacteria. *Protein Sci* **5:**1939–1941, 1996.
19. Hazes B. The (QxW)3 domain: A flexible lectin scaffold. *Protein Sci* **5:**1490–1501, 1996.
20. Shub D, Goodrich-Blair H, Eddy S. Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. *TIBS* **19:**402–404, 1994.
21. Baldi P, Chauvin Y, Hunkapiller T, McClure M. Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci USA* **91:**1059–1063, 1994.
22. Grundy W, Bailey T, Elkan C, Baker M. Hidden Markov model analysis of motifs in steroid dehydrogenases and their homologs. *Biochem Biophys Res Commun* **231:**760–766, 1997.
23. Waterman M, Perlwitz M. Line geometries for sequence comparisons. *Bull Math Biol* **46:**567–577, 1986.
24. Barton G, Sternberg M. Flexible protein sequence patterns: A sensitive method to detect weak structural similarities. *J Mol Biol* **212**(2):389–402, 1990.
25. Gribskov M, McLachlan A, Eisenberg D. Profile analysis: detection of distangly related proteins. *Proc Natl Acad Sci USA* **84:**4355–4358, 1987.
26. Taylor W. Identification of protein sequence homology by consensus template alignment. *J Mol Biol* **188:**233–258, 1986.
27. Bowie J, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253:**164–170, 1991.
28. Hughey R, Krogh A. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *CABIOS* **12:**95–107, 1996. [The hidden Markov model software can be accessed at URL http://www.cse.ucsc.edu/research/compbio/sam.html]
29. NCI (1997). NRP (Non-Redundant Protein) and NRN (Non-Redundant Nucleic Acid) Database. Distributed on the Internet via anonymous FTP from ftp.ncifcrf.gov, under the auspices of the National Cancer Institute's Frederick Biomedical Supercomputing Center.
30. Altschul S. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol* **219:**555–565, 1991.
31. Barrett C, Hughey R, Karplus K. Scoring hidden Markov models. *CABIOS* **13:**191–199, 1997.
32. Bujons J, Dikiy A, Ferrer J, Banci L, Mauk A. Charge reversal of a critical active-site residue of cytochrome-c peroxidase: Characterization of the arg48 → glu variant. *Eur J Biochem* **243:**72–84, 1997.
33. Auerbach A, Mathew C, Verlander P. Fanconi Anemia Mutation Database. The database is available at URL http://www.rockefeller.edu/fanconi/mutate, 1997.
34. Kruyt F. Characteristics of the FA-A Polypeptide, The Science Newsletter is available at URL http://www.rio.com/fafund/FAHTML/SL21, 1997.
35. Alcain F, Buron M. Ascorbate on cell growth and differentiation. *J Bioenerget Biomembr* **26:**393–398, 1994.
36. Orengo C, Michie A, Jones S, Swindells M, Jones D, Thornton J. Cath: Protein structure classification, The CATH database is available at URL http://www.biochem.ucl.ac.uk/bsm/cath, 1997.
37. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19:**55–72, 1994.
38. King R, Sternberg M. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci* **5:**2298–2310, 1996.

I. Saira Mian[2]
Michael J. Moser

Life Sciences Division (Mail Stop 29-100),
Lawrence Berkeley National Laboratory,
1 Cyclotron Road, Berkeley,
California 94720

Received July 28, 1997, and in revised form October 21, 1997

[2] To whom correspondence should be addressed. Fax: (510)486-6949. E-mail: SMian@lbl.gov.