Structural comparisons lead to the definition of a new superfamily of NAD(P)(H)-accepting oxidoreductases: the single-domain reductases/epimerases/dehydrogenases (the 'RED' family)

G. LABESSE,*† A. VIDAL-CROS, J. CHOMILIER,* M. GAUDRY,† and J.-P. MORNON*

*Laboratoire de Minéralogie-Cristallographie, Universités PVI et PVII, C.N.R.S. U.R.A. 09, case 115, and †Laboratoire de Chimie Organique Biologique, Université PVI, C.N.R.S. U.R.A. 493, 75252 Paris Cédex 05, France

Using both primary- and tertiary-structure comparisons, we have established new structural similarities shared by reductases, epimerases and dehydrogenases not previously known to be related. Despite the low sequence identity (down to 10%), short consensus segments are identified. We show that the sequence, the active site and the supersecondary structure are well conserved in these proteins. New homologues (the protochlorophyllide

reductases) are detected, and we define a new superfamily composed of single-domain dinucleotide-binding enzymes. Rules for the cofactor-binding specificity are deduced from our sequence alignment. The involvement of some amino acids in catalysis is discussed. Comparison with two-domain dehydrogenases allows us to distinguish two general mechanisms of divergent evolution.

INTRODUCTION

Reductases and dehydrogenases are among the most widespread enzymes in the living organisms, playing a key role in many metabolic and biosynthetic pathways. NDP-sugar epimerases have also been shown to be essential for sugar metabolism. Many sequences have been published, and similarities have been detected, allowing one to delineate the short-chain dehydrogenase family (Persson et al., 1991), a set of proteins, about 250 amino acids (aa) long, homologous with 20 β -hydroxysteroid dehydrogenase (HSD) whose crystal structure has been reported (Ghosh et al., 1991). Both reductases and dehydrogenases from either prokaryotic or eukaryotic origin belong to the HSD family (Persson et al., 1991; Nakajima et al., 1993; Vidal-Cros et al., 1994).

Independently, the three-dimensional (3D) structure of UDP-galactose 4-epimerase (G4E) complexed with a substrate analogue has been published (Bauer et al., 1992). This protein belongs to another protein family that includes various similar NDP-sugar epimerases and some medium-chain dehydrogenases/reductases of about 330–370 aa in length (Baker and Blasco, 1992).

The rat dihydropteridine reductase (DHR), whose structure was solved by Varughese et al. (1992), is another NAD⁺-binding enzyme, 244 aa in length, with no similar sequences having yet been reported.

Despite their common NAD(P)(H) affinities, their low sequence similarities, as well as their distinct functions and substrate specificities, did not suggest that any important structural similarities would be shared by the proteins of these three families. However, our extensive sequence database screening and 3D structure analysis led us to propose that all these proteins belong to the same superfamily, which we term the 'RED' (Reductases/Epimerases/Dehydrogenases) family.

The three crystal-structure determinations described above

have shown that these dinucleotide diphosphate-binding enzymes (DDBP) possess only one single-domain comprising both the NAD⁺-binding and the active site, whereas other DDBP, such as lactate dehydrogenases, are composed of two distinct domains, one involved in cofactor binding and the other in substrate binding, the active site lying at their interface (Rossmann et al., 1974). The 3D structure analysis revealed that HSD, G4E and DHR possess the common regular (β/α)₆ fold usually observed in DDBP, but some specific structural and functional features differ from those of the two-domain DDBP. In the present paper we discuss the structural similarities shared by the proteins of this new enzyme superfamily. A precise comparison of the crystal structures and sequences led us to identify residues likely to be implied in the catalysis and to propose rules for the cofactor specificities. Short consensus sequences were defined that allowed a specific recognition of the proteins of the superfamily. Some residues were already identified as potentially involved in catalysis, but our alignment proposes additional aa as catalytically important residues. The conservation of a common active-site structure implies similarities in the enzymic mechanisms of the RED proteins.

EXPERIMENTAL

Protein databank searches were performed using the BLAST network server at the National Center for Biotechnology Information (N.C.B.I.) with default parametrization (Altschul et al., 1990). Hydrophobic cluster analysis (HCA) was used for pairwise and multiple alignments (Gaboriaud et al., 1987). Editing of multiple alignment and profile database screenings were respectively performed with the programs LINEUP and PROFILESEARCH from the software package GCG 7.0 (Genetics Computer Group Inc., Madison, WI, U.S.A.). The SWISSPROT Database (Release 27, October 1993) was used for the profile screenings. The statistical validation of the similarities

Abbreviations used: 3D, three-dimensional; aa, amino acid; RMSD, root-mean-square deviation; the one-letter code for amino acids is used; PChIR, protochlorophyllide reductase; RED, Reductases/Epimerases/Dehydrogenases; HSD, 20 β -hydroxysteroid dehydrogenase; G4E, UDP-galactose 4-epimerase; DHR, dihydropteridine reductase; DDBP, dinucleotide diphosphate-binding protein; N.C.B.I., National Center for Biotechnology Information; HCA, hydrophobic cluster analysis.

† To whom correspondence should be sent.

Figure 1 Sequence alignment of DHR, G4E and HSD

The alignment was deduced from the COMPOSER superimposition [cut-off = 0.35 nm (3.5 Å)] of known crystal structures of RED enzymes and used to screen sequence databases. The sequence identity ranges from 13 to 16% over the whole alignment. Strictly conserved and similarly replaced aa are respectively black- and grey-boxed. The PCHr sequence from *Pinus* (Sparo et al., 1992) was aligned, showing with RED 15–20% sequence identity clustering in consensus segments. The aa numbering corresponds to the position in the HSD sequence. Asterisks indicate sequence truncations. A box is drawn surrounding the conserved segments, and secondary structures are underlined and labelled as described in the text.

Figure 2 Superimposed 3D structures of the common core of DHR, G4E and HSD

The corresponding co-ordinate files 1DHR, 1UDP and 1HSD were from the Brookhaven databank (Bernstein et al., 1977). NAD(P)(H)- and substrate-binding sites are labelled respectively by 'NAD(P)(H)' and 'S' symbols. The numbering of the secondary structures follows that defined in the text.

came from the Poisson-Law probabilities calculated by the program BLAST and from the PROFILESEARCH Z score. 3D visualization was done on a PS390/VAX3200 workstation using the program MANOSK (Cherfils et al., 1988). The program COMPOSER was used to perform 3D superimposition and root-mean-square-deviation (RMSD) calculations (Sutcliffe et al., 1987).

RESULTS

Alignment of primary structures

Database screenings were first performed, to show strong sequence similarities using either the HSD, the G4E or the DHR sequence as a probe. Multiple alignments of highly similar proteins were then deduced and used to build a profile for database screenings. Our results agreed with already published alignments of the sequences from the HSD subfamily (Ghosh et al., 1991). New similarities were detected between HSD proteins and the members of the G4E family as well as with DHR sequences (Figure 1). This prompted us to re-analyse these similarities. Additional searches and pairwise comparisons were performed and confirmed our first results. New homologues were found that shared unreported similarities, among which were the protochlorophyllide reductases (PChlRs) from angiosperms (Figure 1; Spano et al., 1992). The region comprising the NAD(P)(H)-binding site and the active site is well conserved in the whole superfamily, whereas larger insertion/deletion mutations occur mainly in the C-terminal part (Figure 1). Comparison at a secondary-structure level, through HCA, showed the conservation of the global fold in the whole RED superfamily containing more than 150 sequences (results not shown).

Global survey of 3D structures

The topology of the N-terminal and central parts (aa 1–200) comprising the cofactor-binding site as well as part of the active

site shows similarities with the dinucleotide-binding domain of the two-domain dehydrogenases. The numbering of the secondary-structure elements was chosen to evidence this fact and referred to the classical $(\beta/\alpha)_6$ fold described by Rossmann et al. (1974).

The automatically refined superimposition of the α -carbon traces of the three crystal structures (a member of each subfamily: HSD, G4E and DHR) allowed the definition of a common core (137 aa long; see Figure 2). The RMSD of the α -carbon positions ranged from 0.145 to 0.185 nm (1.45–1.85 Å), indicating that these structures are closely related, despite the low sequence identity in this core, scoring around 15% (Figures 1 and 2). It showed quite perfect fitting of the central parallel β -sheet (Figure 2). Four helices, namely $\alpha 1$, $\alpha 4$, $\alpha 5$ and $\alpha 6$, superimposed well (Figure 2). The cofactor-binding site (aa 1–80) is well conserved, while the substrate-binding site (aa 80–250) has diverged more. In the RED proteins, the secondary-structure elements surrounding the active site ($\beta 4$, $\alpha 4$, $\beta 5$, $\alpha 5$, $\beta 6$ and $\beta 7$) are twice as long as those usually met with in DDBP. The external loops were highly variable in geometry and length.

While DHR and G4E are dimeric, HSD is tetrameric (Persson et al., 1991; Bauer et al., 1992; Varughese et al., 1992). However, the tetramer could be structurally described as a dimer of dimers. The interfaces, between the A and C, or between the B and D, subunits are composed of two long parallel helices ($\alpha 4$ and $\alpha 5$) from each subunit (Persson et al., 1991). The dimer interface is composed of the same helices in DHR and G4E. The tetramer is stabilized by tight interactions involving the C-terminal part of HSD (aa 220–250) (Persson et al., 1991). This region differs from those of the dimeric DHR or G4E.

Two long loops ($\beta 5$ – $\alpha 5$, $\beta 7$ –C-terminal end) in G4E face the substrate-binding site closing a deep buried cleft (Bauer et al., 1992). The UDP moiety of the G4E substrate lies in this groove, as shown in the analogue–NAD⁺–enzyme ternary complex structure (Bauer et al., 1992). The residues conserved in RED proteins lie in this cleft. These structural features suggest that the

substrates point toward the active site in a similar way in the RED enzymes, while the substrate specificities would be due to interactions with the variable loops.

DISCUSSION

Structure-function relationship

The common active site of the RED enzymes was delineated from residue conservation and 3D structure survey. The two loops ($\beta 1$ - $\alpha 1$, $\beta 2$ - $\alpha 2$) involved in phosphate groups and ribose binding are well conserved, and the positions of the complexed cofactor perfectly superimposed in the three holoproteins HSD, G4E and DHR. The nicotinamide ring points towards a conserved long bulge in the fourth strand. Five residues conserved in the proteins of the RED family are spatially clustered around this bulge. In the G4E structure the substrate analogue is in close contact with Y152 and K156 ($\alpha 5$ helix), which have been already identified by sequence alignment and directed mutagenesis as being in the active site of the HSD proteins (Persson et al., 1991; Chen et al., 1993). In the RED proteins, the $\alpha 4$ helix and the $\beta 5$ strand contribute two residues to the active site (Figure 1), namely N111 and S139 (in HSD), which are well conserved in the RED proteins, but in DHR they are replaced by a serine and an alanine residue respectively. However, the crystal structure of DHR reveals the presence of internal water molecules that could counterbalance the aa mutations and the resulting lack of two polar groups. This observation correlates well with the proposed involvement of solvent molecule for proton delivery to 'quinonoid' dihydropteridine (Varughese et al., 1992). Similarly, the T187 is well conserved in the HSD family and lies in close contact with the nicotinamide end in the HSD structure. In DHR, a threonine residue is similarly positioned.

This suggests that the three residues, N111, S139 and T187, facing the side chains of the catalytic Y152 and K156 residues, belong to the active site. Directed mutagenesis studies remain to be done to test their role.

The similarity in the active-site structure implies some similarities in the catalysis mechanism. HSD catalyses a reversible oxidation of hydroxysteroid (Persson et al., 1991). Some dehydrogenases also possess a specific reductase activity (Krook et al., 1993). It has been observed that most of the RED proteins are reversible oxidoreductases (Persson et al., 1991; Bauer et al., 1992; Varughese et al., 1992; Vidal-Cros et al., 1994). The epimerase reaction has been viewed as a coupled oxidation-reduction that would explain the close relationship between epimerases and oxidoreductases/dehydrogenases (Bauer et al., 1992).

Signature of RED proteins

Seven consensus segments were deduced from the alignment of all the RED sequences (~150). Along with the consensus segments of the NAD⁺-cofactor-binding site, the sequence stretches including the active-site residues N111, S139, Y152, K156 and T187 constitute a signature for all the RED enzymes. In this superfamily, consensus sequence comprising the $\beta 1$ strand and the phosphate binding loop is ***tgxxxx*g where * stands for any hydrophobic aa, x for any aa, lower-case letters for the most frequent aa and upper-case letters (see below) for strictly conserved aa. The loop is one aa longer than in other DDBP (except malate dehydrogenases) and preceded by the $\beta 1$ strand containing a β -bulge in RED enzymes (Figure 1). The segment ($\beta 2$ strand) involved in the ribose binding has distinct consensus sequences ****dx and ***(nyx/*sr), for NAD(H)- and

NADP(H)-specific enzymes respectively. The $\beta 4$ strand sequence usually encountered over the whole family is *d***nxag. This strand contains a large bulge stabilized by a conserved glycine residue. The consensus for $\alpha 4$ helix and $\beta 5$ strand, *dxn* and ****ss comprise, in HSD, the N111 and S139 residues respectively. The $\alpha 5$ helix comprises the consensus segment YxxsKxx* (Persson et al., 1991). The pgxxx sequence contributes aa (T187) to the active site in HSD and DHR. The use of this signature allowed us to confirm the similarity shared between PChlR and the proteins of the HSD subfamily (Figure 1).

Comparison with other NAD(P)(H)-binding proteins

The common features of the two-domain dehydrogenases have been already described (Rossmann et al., 1974). We compared these proteins with those of the newly defined superfamily.

The characteristic connectivities ($-1x, -1x, +3x, +1x, +1x$; according to the nomenclature of Richardson, 1976) linking the first six β -strands are conserved in both single- and two-domain DDBP (Persson et al., 1991; Bauer et al., 1992; Varughese et al., 1992). This common topology does not extend beyond the sixth β -strand. The dyad symmetry of the Rossmann fold is lost in RED proteins, as the secondary-structure elements are longer in the C-terminal part than in the N-terminal one, and an additional strand ($\beta 7$) is present (Figure 2).

The DDBP fold signature comprises few conserved regions. Among them are those involved in tight interaction with phosphate groups and ribose binding. An unexpected common feature of DDBP is the conserved pattern, *d****, of the fourth strand core, hydrophobic as are most of the internal parallel strands, but usually preceded in DDBP by an aspartate residue. The role of this aspartate residue in the stability or the folding of the DDBP structure remains an open question.

The divergence of DDBP from a common ancestor possessing a single-domain is suggested by the sequence and structure conservations. The single-domain enzymes have longer secondary-structure elements, forming a substrate-binding pocket. Mutational events, mainly occurring in loops of the DDBP ancestor, would lead to the RED family, while we believe that the two-domain enzymes could result from the addition of a new domain at a protein end, by gene or intron rearrangement.

We thank Dr. M. J. Sutcliffe and Dr. N. Srinivasan, who kindly provided us with the program COMPOSER. We thank Dr. D. Ghosh and Dr. W. L. Duax for providing us with the newly refined structure of HSD. We thank the Service de Bioinformatique (INSERM-CNRS at Villejuif, France) for access to its computational facilities, and N. Colloc'h for critical reading of the manuscript before its submission. This work was supported by the Universités Paris VI and Paris VII, the C.N.R.S., I.N.S.E.R.M. (contract 910912). G.L. is supported by a fellowship from the Ministère de l'Enseignement Supérieur et de la Recherche.

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403-410
- Baker, M. E. and Blasco, R. (1992) *FEBS Lett.* **301**, 89-93
- Bauer, A. J., Rayment, I., Frey, P. A. and Holden, H. M. (1992) *Proteins: Struct. Funct. Genet.* **12**, 372-381
- Bernstein, F. C., Koetzel, T. F., Williams, G. J. B., Meyer, E. F., Brice, J. M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M., (1977) *J. Mol. Biol.* **112**, 535-542
- Chen, Z., Jiang, J. C., Lin, Z.-G., Lee, W. R., Baker, M. E. and Chang, S. H. (1993) *Biochemistry* **32**, 3342-3346
- Cherfils, J., Vaney, M.-C., Morize, I., Surcouf, E., Colloc'h, N. and Mornon, J.-P. (1988) *J. Mol. Graph.* **6**, 155-160
- Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J.-P. (1987) *FEBS Lett.* **224**, 149-155
- Ghosh, D., Weeks, C. M., Grochulski, P., Duax, W. L., Erman, M., Rimsay, R. L. and Orr, J. C. (1991) *Proc. Natl. Acad. Sci. U.S.A.* **88**, 10064-10068

- Krook, M., Ghosh, D., Strömberg, R., Carlquist, M. and Jörnvall, H. (1993) *Proc. Natl. Acad. Sci. U.S.A.* **90**, 502–506
- Nakajima, K., Hashimoto T. and Yamada, Y. (1993) *Proc. Natl. Acad. Sci. U.S.A.* **90**, 9591–9595
- Persson, B., Krook, M. and Jörnvall, H. (1991) *Eur. J. Biochem.* **200**, 537–543
- Richardson, J. S. (1976) *Proc. Natl. Acad. Sci. U.S.A.* **73**, 2619–2623
- Rossmann, M. G., Moras, D. and Olsen, K. W. (1974) *Nature (London)* **250**, 194–199
- Spano, A. J., He, Z. and Timko, M. P. (1992) *Mol. Gen. Genet.* **236**, 86–95
- Sutcliffe, M. J., Haneef, I., Carney, D. and Blundell, T. (1987) *Protein Eng.* **1**, 377–384
- Varughese, K. I., Skinner, M. M., Whiteley, J. M., Matthews, D. A. and Xuong, N. H. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 6080–6084
- Vidal-Cros, A., Viviani, F., Labesse, G., Boccara, M. and Gaudry, M. (1994) *Eur. J. Biochem.* **219**, 985–992

Received 3 March 1994/26 May 1994; accepted 9 June 1994