Biotechnol Lett (2015) 37:417–427
DOI 10.1007/s10529-014-1687-y

ORIGINAL RESEARCH PAPER

Fatty acid synthesis enzyme clans

Ngoc N. Phan · Yuen Keong Lee · Peter J. Reilly

Received: 2 August 2014 / Accepted: 12 September 2014 / Published online: 26 September 2014

Abstract Ketoacyl reductases (KRs), hydroxyacyl dehydratases (HDs), and enoyl reductases (ERs) are part of the fatty acid/polyketide synthesis cycle. They are known as acyl dehydrogenases, enoyl hydratases, and hydroxyacyl dehydrogenases, respectively, when catalyzing their reverse reactions. Earlier, we classified these enzymes into four KR, eight HD, and five ER families by statistical criteria. Members of all four KR families and three ER families have Rossmann folds, while five HD family members have HotDog folds. This suggests that those proteins with the same folds in different families may be distantly related, and therefore in clans, even though their amino acid sequences may not be homologous. We have now defined two clans containing three of the four KR families and two of the eight HD families, using manual and statistical tests. One of the ER families is related to the KR clan.

Keywords Enoyl reductases · Hydroxyacyl dehydratases · Ketoacyl reductases · Phylogeny

Abbreviations

RMSD	Root mean square deviation
P	Percentage of amino acid residues in the shorter (reference) molecule to that of the longer molecule used to calculate RMSDs

Introduction

The fatty acid and polyketide synthesis cycles comprise eight main enzyme groups (Fig. 1). We have gathered links to their amino acid sequences (primary structures) and three-dimensional structures (tertiary structures) into the ThYme database (Cantu et al. 2011). Three of these enzyme groups, 3-ketoacyl reductases (KRs), hydroxyacyl dehydratases (HDs), and enoyl reductases (ERs), reduce a 3-keto group in an acyl thioester to a hydroxyl group, remove a water molecule to form an enoyl group, and hydrogenate the resulting double bond to yield a saturated chain, respectively. When they catalyze their reverse reactions as part of the fatty acid β-oxidation pathway, they are known in turn as 3-hydroxyacyl dehydrogenases, enoyl hydratases, and acyl dehydrogenases.

We have divided KRs, HDs, and ERs into four, eight, and five families, respectively, by using the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) with an *E*-value of 0.001 and by employing a number of other tests (Cantu et al. 2010, 2011, 2012). This ensures that members of the same

Electronic supplementary material The online version of this article (doi:10.1007/s10529-014-1687-y) contains supplementary material, which is available to authorized users.

N. N. Phan · Y. K. Lee · P. J. Reilly (✉)
Department of Chemical and Biological Engineering, Iowa State University, Ames, IA 50011-2230, USA
e-mail: reilly@iastate.edu

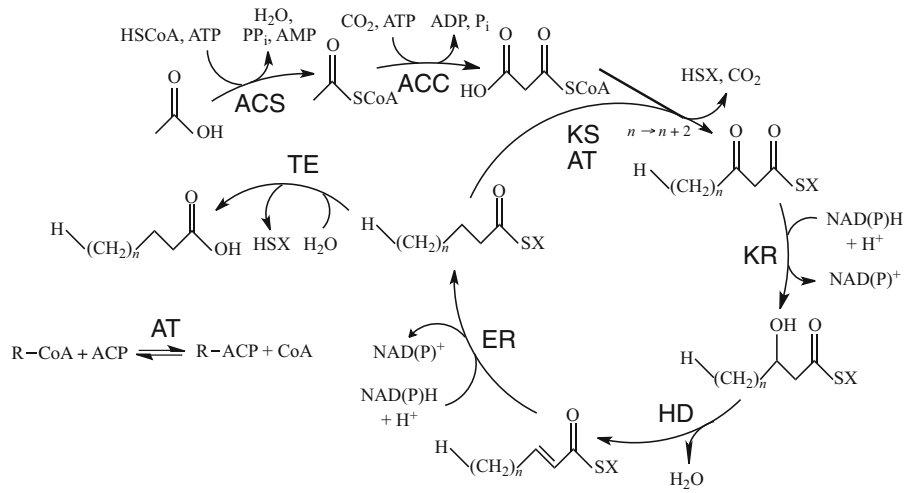

Fig. 1 The fatty acid synthesis cycle and the enzyme groups that are part of it. *ACC* acetyl-CoA carboxylase; *ACS* acyl-CoA synthase; *AT* acyl transferase; *ER* enoyl reductase; *HD* hydroxyacyl dehydratase; *KR* ketoacyl reductase; *KS* ketoacyl synthase; *TE* thioesterase; *SX* coenzyme A or acyl carrier protein. Modified and reprinted from Cantu et al. (2011, 2012) with permission of Oxford University Press

family will have homologous primary structures and superimposible tertiary structures, even though members of different families of the same enzyme group may have almost totally nonhomologous primary structures.

Family KR1, which has the most entries of any family in the ThYme database, is composed of 3-ketoacyl-acyl carrier protein (ACP) reductases, 3-hydroxyacyl-coenzyme A (CoA) dehydrogenases, and many other reductases and dehydrogenases (Cantu et al. 2011, 2012). It also incorporates the former ER1, as all members of that family have primary structures homologous with those of KR1. Its catalytic residues are serine, tyrosine, and lysine, with tyrosine thirteen residues to the *C*-terminal end of the protein chain from serine, and lysine a further four residues away. Family KR2 comprises 3-hydroxyacyl-CoA dehydrogenases. Its catalytic residues are histidine and glutamate, twelve residues apart. Members of KR3 and KR4 are acyl-ACP reductases that are part of fatty acid and polyketide synthases. KR3 catalytic residues are serine, tyrosine, and lysine, twelve and four residues apart, almost identical in spacing to those of KR1. The catalytic residues of KR4 are lysine, serine, and tyrosine, as in KR1 and KR3, but in a different order, with serine 24 residues toward the *C*-terminus from lysine and tyrosine thirteen residues further. All four KR families have members with NAD(P)-binding Rossmann folds, KR2 in addition having a *C*-terminal 6-phosphogluconate dehydrogenase fold.

Family HD1 is composed of enoyl-CoA hydratases and hydroxyacyl-ACP dehydratases (Cantu et al. 2011, 2012). Its catalytic residues are aspartate and histidine, five residues apart. HD2 has enoyl-CoA hydratases, with two catalytic glutamate residues 20 residues apart. HD3 contains hydroxyacyl-ACP dehydratases and enoyl-CoA hydratases, like HD1, but they are part of larger proteins. Its catalytic residues are aspartate and histidine, five residues apart, as in HD1. Families HD4 through HD6 are all comprised of hydroxyacyl-ACP dehydratases, but with catalytic residues that vary in identity and placement. In HD4, histidine and aspartate are more than 150 residues apart. In HD5, histidine and aspartate are on different subunits, with the former being located fourteen residues closer to the *N*-terminus of its chain. The same arrangement is found in HD6, but its catalytic residues are histidine and glutamate. Members of HD1 and HD3 through HD6 have HotDog folds. The fold in HD2 is a ClpP/crotonase structure. HD7 and HD8, the first containing acyl-ACP dehydratases and the second containing acyl-CoA dehydratases, do not yet have known tertiary structures.

Families ER2 through ER5 contain mainly enoyl-ACP reductases, with the last also having enoyl-CoA reductases (Cantu et al. 2011, 2012). ER6 is composed of dienoyl-CoA reductases and other reductases. ER3 and ER4 are part of much larger fatty acid synthase complexes. The catalytic residues in ER2 are tyrosine and lysine, seven residues apart. ER3 yet has no clearly identified catalytic residues. Those in ER4 through ER6 are lysine and aspartate, 26 residues apart; tyrosine and tryptophan, separated by over 200 residues; and tyrosine and histidine, over 80 residues apart, respectively. Members of ER2 and ER4 have

NAD(P) Rossmann folds, while ER3 and ER6 families have (α,β)-barrels. ER5 members are composed of a combination of GroES-like and Rossmann folds.

Obvious similarities in enzyme functions, catalytic residue identities and locations, and tertiary structures among different families are evident within these three enzyme groups. Furthermore, KR and ER families both catalyze reductions while ER1 has already been amalgamated into KR1. This suggests that different families may be gathered into clans, where their primary structures may be totally dissimilar but where their tertiary structures may be so similar that they may be superimposed with small root mean square deviations (RMSDs) between adjacent residues. Furthermore, members of the same clan should have similar sequences of secondary structure elements (SSEs) and similar catalytic mechanisms, the latter suggesting that catalytic residues are usually similar and are in the same locations when chains are superimposed upon each other. These considerations imply that members of the same clan either have been the subjects of extensive convergent evolution or have common distant protein ancestors, more distant than the ancestors of proteins within the same family.

A good example of clan determination is found in the CAZy database (Lombard et al. 2014), where many of the approximately 130 glycoside hydrolase families at the time of writing are part of 14 clans containing from two to 19 families.

We have also defined clans. Of the present 25 thioesterase (TE) families (Cantu et al. 2010, 2011), two clans comprise three and four families of acyl-CoA hydrolases whose members have HotDog folds, while two other clans of acyl-ACP hydrolases contain two and three families with α,β-hydrolase folds. Four of five ketoacyl synthase families form a clan of enzymes with α-β-α-β-α sandwiches (Chen et al. 2011). Of 67 carbohydrate binding module families (Lombard et al. 2014), nine clans comprise 27 families (Carvalho et al. 2014). Eight of these clans have tertiary structures consisting of β-sandwiches, while the ninth has all β-trefoils.

This article reports an effort to classify KR, HD, and ER families into clans by manual and statistical means. It also reports an inquiry into whether members of families of fatty acid synthetic enzymes of different substrate specificities, such as those KRs and ERs with Rossmann folds, as well as those ERs and TEs with HotDog folds, can be related to each other.

Computational methods

The methods used to define clans closely follow those that we employed to determine clans of carbohydrate-binding modules (Carvalho et al. 2014). In summary, tertiary structures of one member of each KR, HD, and ER family were obtained from the Protein Data Bank (Berman et al. 2000). They were aligned using MultiProt (Shatsky et al. 2004) with members of other families in the same enzyme group, and sometimes with those in other enzyme groups. Pairs of tertiary structures were visualized with PyMOL (Schrödinger 2014), and the RMSDs between all α-carbon atoms located within the cutoff distance (the average distance between α-carbon atoms in the chain) in two aligned chains were calculated using MATLAB (The MathWorks 2014). The ratio of the number of residues in the shorter molecule to that of the longer molecule used to calculate RMSDs gives a P value (Cantu et al. 2010).

Comparisons of tertiary structures of members of the same family of carbohydrate-binding modules generally yielded RMSD and P values < 1.6 Å and > 80 %, respectively (Carvalho et al. 2014). Tertiary structures of members of different families in the same clan generally gave RMSDs of <2.2 Å and P values > 60 %. RMSD and P values of tertiary structures not in the same clan can have similar RMSD and P values if they have similar tertiary structures, but usually their RMSDs are > 2.2 Å and their P values are $\ll 60$ %, the latter value often caused by significant differences in chain lengths between the proteins being compared. RMSD and P values of different families in the four TE clans were similar to those of the carbohydrate-binding domains (Cantu et al. 2010).

Secondary structure elements (SSEs) were tabulated and their lengths were measured from illustrations based on DSSP (Kabsch and Sander 1983) in the Protein Data Bank, which often display more SSEs than PyMol illustrations.

Results and discussion

We separated KR, HD, and ER families into clans by success in superimposing representative tertiary structures of different families, as gauged by RMSD and P values between structures, success in superimposing

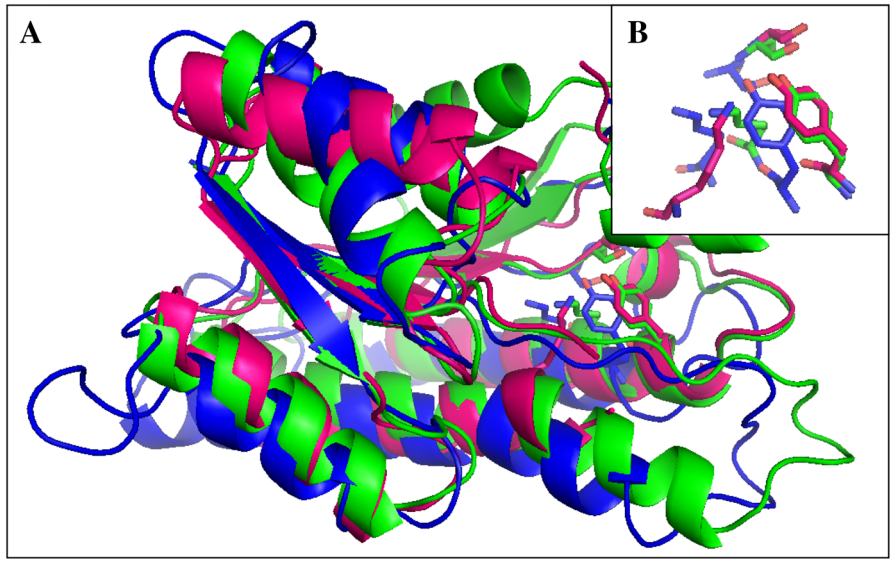

Fig. 2 **a** Superimposed representative tertiary structures of clan KR-A: KR1 (PDB 1EDO β-ketoacyl-ACP reductase from *Brassica napus*, *green*); KR3 (PDB 2PFF ketoacyl-ACP reductase in fatty acid synthase from *Saccharomyces cerevisiae*, *blue*); KR4 (PDB 2VZ8 NADPH-dependent β-keto-ATP reductase in fatty acid synthase from *Sus scrofa*, *red*). **b** Superimposed active-site side chains of the same KR family representatives, with colors as in **a**

catalytic residues, SSE location and order, substrate specificities, and chain lengths, in decreasing order of importance.

Clan KR-A

All four KR families have Rossmann folds. Three of the four, KR1, KR3, and KR4, are part of clan KR-A. This is confirmed by the very close superimposition of representatives of each family (Fig. 2a). In addition, they have identical catalytic residues (serine, tyrosine, and lysine, although they are in a different order in KR4, lysine being to the *N*-terminal side of serine and tyrosine in the protein chain rather than to the *C*-terminal end, as in KR1 and KR3 (Cantu et al. 2012). The catalytic residues are in the same general locations in the active site, with those of KR3 being uniformally displaced from those of KR1 and KR4 (Fig. 2b). Serine and tyrosine catalytic residues in KR1 and KR4 quite closely overlap, but the side chains of their lysine residues are aligned differently.

Further confirmation of the membership of this clan comes from relative RMSD and *P* values (Table 1). RMSDs linking KR1, KR3, and KR4 are 1.87 Å (KR1 and KR3), 1.57 Å (KR1 and KR4), and 1.83 Å (KR3 and KR4), much lower than those between KR2 and KR1, KR3, and KR4, which are 2.13 Å, 2.25 Å, and 2.23 Å, respectively. Furthermore, *P* values within clan KR-A, 75.8 % (KR1 and KR3), 64.8 % (KR1 and KR4), and 77.5 % (KR3 and KR4), are in general higher than those between KR2 and KR1, KR3, and KR4, which are 48, 55, and 69.3 %, respectively.

The SSEs of KR1, KR3, and KR4 can be aligned fairly well (Fig. 3), as should be expected of representatives of different families in the same enzyme group whose tertiary structures can be closely superimposed. Interestingly, the SSEs in KR2 can be aligned with those of the other three families even though the KR2 tertiary structure is less well superimposed with those of the other three families than they are with each other. This is probably caused by all four KR families being composed of Rossmann folds.

A further observation separating KR2 from KR1, KR3, and KR4 is that KR2 members are 3-hydroxyacyl-CoA dehydrogenases, and therefore they are part of the fatty acid β-oxidation pathway. On the other hand, KR1 members that attack acyl chains are mainly 3-ketoacyl-ACP reductases, while members of KR3 and KR4 are almost all 3-ketoacyl-ACP reductases (Cantu et al. 2012), and therefore they are all members of the fatty acid synthesis cycle.

It may be noted with KRs (Table 1), as well as with HDs and ERs later (see Tables 2 and 3 below), that intrafamily values of RMSD and *P*, in general determined from more than two tertiary structures, are usually much lower and much higher, respectively, than those between single representatives of two families. This occurs because the homologous primary structures within a family dictate very similar secondary and tertiary structures, while nonhomologous

Table 1 Lengths, RMSDs, and *P* values between different KR families

Families	Number of residues		KR1	KR2	KR3	KR4
KR1 (PDB 1EDO)	244	RMSD (Å)	2[a]			
		P (%)	84.1[a] (>50)			
KR2 (PDB 1WDK)	179	RMSD (Å)	2.13	1.2[a]		
		P (%)	48	91.3[a] (5)		
KR3 (PDB 2PFF)	200	RMSD (Å)	1.87	2.25	0.93[a]	
		P (%)	75.8	55	98.5[a] (2)	
KR4 ((PDB 2VZ8)	180	RMSD (Å)	1.57	2.23	1.83	1.28[a]
					77.5	98.1[a] (4)

Numbers in parentheses: Number of tertiary structures used for calculation of RMSD and *P* values within a family

[a] From Cantu et al. (2012). PDB notations: Protein Data Bank identification codes for protein tertiary structures

Families and PDB designations	Number of residues	Progression of secondary structure elements
KR1 (PDB 1EDO)	244	β T α T β α β T α β 3 α β α T α T β b α T α T b α 3 β T
KR2 (PDB 1WDK)	179	β α T β α T α β T 3 β α T T β α 3 3 β T β α T β
KR3 (PDB 2PFF)	200	β α β T α T T β T α T β b T α T α T β T b T α T T β
KR4 (PDB 2VZ8)	180	β T α T β α T β α β T α T β α T T α T β b
ER2 (PDB 1C14)	262	T β T α T β T α T β T α T β 3 α α 3 β 3 T α 3 β T 3 T α T α 3 T β T 3
ER4 (PDB 2VZ8)	138	α T β α T T β α T β α T β α T β 3 β 3 T α
ER5 (PDB 2VCY)	186	α T α T β T α T β T α T β α T β α T β T β α β α α T

α – α-helix; β – β-helix; b – β-bridge; T – turn; 3 – 3/10 helix.

Fig. 3 Order of secondary structure elements of KRs and ERs with Rossmann folds

primary structures of different families in a clan dictate secondary and tertiary structures that vary more.

Clan HD-A

Families HD1 and HD3–HD6 have HotDog folds, while HD2 members have ClpP/crotonase folds and HD7 and HD8 have no known tertiary structures. HD5 and HD6 make up clan HD-A. They can be closely superimposed (Fig. 4a), as can HD1, while HD3 and HD4 can neither be closely superimposed on HD1, HD5, and HD6, nor on each other. However, the catalytic aspartate and histidine residues of HD1 are in different locations than the aspartate and histidine residues of HD5 or the glutamate and histidine residues of HD6, which overlap each other very closely (Fig. 4b). This suggests that HD1 is not in the same clan as HD5 and HD6, despite being closely superimposed on them. Cantu et al. (2012) had already successfully superimposed HD5 and HD6 tertiary structures, but without employing statistical measures, superimposing their catalytic residues, or aligning their SSEs.

RMSDs between HD2 and the five families with HotDog folds range from 2.39 to 2.52 Å, while those among the latter are from 1.47 to 2.20 Å, with only four of the ten RMSDs being >2.00 Å (Table 2). It is of interest that RMSDs between HD3 and HD4, on one hand, and HD1, HD5, and HD6, on the other hand, are low, even though they cannot be superimposed very well. Values of *P* range from 35.6 to 87.1 %, the lowest values being associated with comparisons of HD3 and HD4 with HD1, HD5, and HD6 because of the great disparity in their lengths, and the highest values stemming from comparisons among HD1, HD5, and HD6, which are roughly the same length.

As should be expected, the SSE sequences of members of HD1, HD5, and HD6 are similar (Fig. 5). Those of HD3 and HD4 are like them only near their

Table 2 Lengths, RMSDs, and *P* values between different HD families

Families	Number of Residues		HD1	HD2	HD3	HD4	HD5	HD6
HD1 (PDB 2C2I)	151	RMSD (Å)	1.47[a]					
		P (%)	78.1[a] (4)					
HD2 (PDB 1DUB)	260	RMSD (Å)	2.52	1.37[a]				
		P (%)	10.7	85.5[a] (14)				
HD3 (PDB 3KH8)	297	RMSD (Å)	1.85	2.51	1.2[a]			
		P (%)	41.7	20.4	79.7[a] (9)			
HD4 (PDB 2VZ8)	232	RMSD (Å)	2.09	2.47	2.25	–		
		P (%)	39.7	48.1	54.9	–		
HD5 (PDB 1MKA)	171	RMSD (Å)	1.89	2.5	1.99	2.13	0.62[a]	
		P (%)	64.3	13.4	35.6	49.6	91.5[a] (2)	
HD6 (PDB 2GLL)	171	RMSD (Å)	1.87	2.39	2.01	1.63	1.47	1.04[a]
		P (%)	71.1	12.2	36.9	36.6	87.1	89.9[a] (4)

Numbers in parentheses: Number of tertiary structures used for calculation of RMSD and *P* values within a family

– Only one known tertiary structure

[a] From Cantu et al. (2012). PDB notations: Protein Data Bank identification codes for protein tertiary structures

Table 3 Lengths, RMSDs, and *P* values between different ER families

Families	Number of residues		ER2	ER3	ER4	ER5	ER6
ER2 (PDB 1C14)	262	RMSD (Å)	1.45[a]				
		P (%)	77.9[a] (15)				
ER3 (PDB 2PFF)	287	RMSD (Å)	2.59	0.76[a]			
		P (%)	59.9	99.7[a] (2)			
ER4 (PDB 2VZ8)	138	RMSD (Å)	2.16	2.26	–		
		P (%)	45.7	29.6	–		
ER5 (PDB 2VCY)	186	RMSD (Å)	2.09	2.33	1.92	1.43[a]	
		P (%)	54.7	40.8	63.4	90.9[a] (2)	
ER6 (PDB 1PS9)	369	RMSD (Å)	2.44	2.34	2.47	2.46	1.31[a]
		P (%)	50.9	54.5	29	37.4	84.5[a] (4)

[a] From Cantu et al. (2012). PDB notations: Protein Data Bank identification codes for protein tertiary structures

Numbers in parentheses: Number of tertiary structures used for calculation of RMSD and *P* values within a family

–: Only one known tertiary structure

C-terminal ends. In addition, they differ greatly from each other. HD2, with a completely different fold, has a totally different SSE sequence than the others.

A further measure that differentiates HD2 from HD1 and HD3–HD6 is that HD2 members have two catalytic glutamate residues, while the other families have either catalytic aspartate and histidine residues (HD1 and HD3–HD5) or catalytic glutamate and histidine residues (HD6) (Cantu et al. 2012).

Enoyl reductases

Families ER2, ER4, and ER5, all with Rossmann folds, can be closely superimposed. However, their catalytic amino acid residues are not only different, tyrosine and lysine in ER2, lysine and aspartate in ER4, and tyrosine and tryptophan in ER5, but they are in different locations in the active site. Families ER3 and ER6, with (α,β)-barrels, are neither well

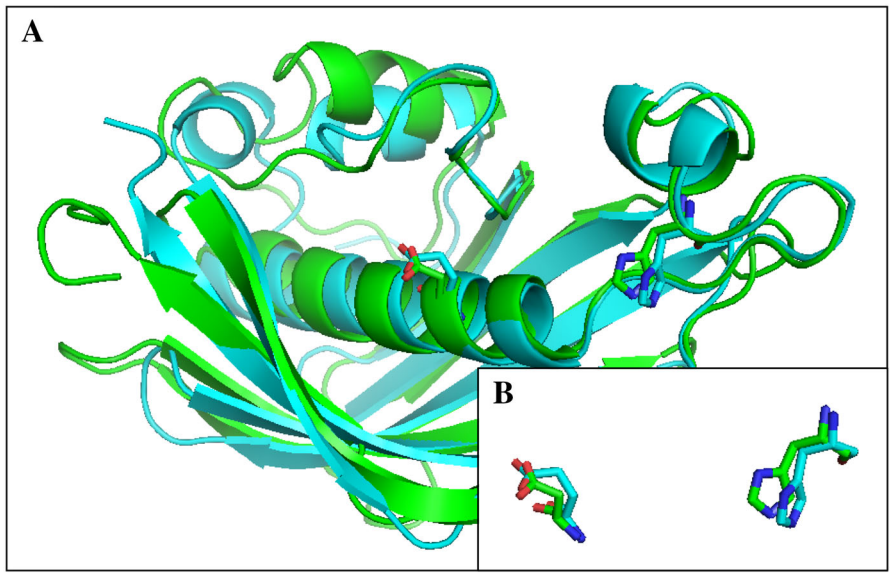

Fig. 4 **a** Superimposed representative tertiary structures of clan HD-A: HD5 (PDB 1MKA β-hydroxydecanoyl thiol ester-ACP dehydrase from *Escherichia coli*, *green*); HD6 ((3R)-hydroxyacyl-ACP dehydrat-ase (FabZ) from *Helicobacter pylori*, *cyan*). **b** Superimposed active-site side chains of the same HD family representatives, with colors as in **a**

Families and PDB designations	Number of residues	Progression of secondary structure elements
HD1 (PDB 2C2I)	151	β α T β β α α T b α T β β b T β T β T β
HD5 (PDB 1MKA)	171	b α T T β T β T 3 T α T β β T β β T β T
HD6 (PDB 2GLL)	171	b α T β T β 3 T b α α T β b β T β T β
HD3 (PDB 3KH8)	297	α β α T α T T T 3 α T α T T β T β β T β β β T α 3 3 T α T 3 β T β T β T β 3
HD4 (PDB 2VZ8)	232	3 β T β α T 3 β β β T β T β 3 β α 3 β T β α T β b β α β β T β T β β

α – α-helix; β – β-helix; b – β-bridge; T – turn; 3 – 3/10 helix.

Fig. 5 Order of secondary structure elements of HDs with HotDog folds

superimposed with each other nor with ER2, ER4, and ER5. These conclusions suggest that the present five ER families do not form a clan.

These observations are supported by Table 3 where RMSDs among ER2, ER4, and ER5 are 2.16 Å (ER2 and ER4), 2.09 Å (ER2 and ER5), and 1.92 Å (ER4 and ER5), while RMSDs between ER3 and ER6, on one hand, and ER2, ER4, and ER5, on the other hand, range from 2.26 to 2.59 Å. Furthermore, the RMSD between ER3 and ER6 is 2.34 Å. Corresponding values of *P* are less clearly predictive, those within among ER2, ER4, and ER5 being 45.7 % (ER2 and ER4), 54.7 % (ER2 and ER5), and 63.4 % (ER4 and ER5), caused by the significant differences in chain lengths among them. The same or lower *P* values are found when ER3 and ER6 are compared with each other and with ER2, ER4, and ER5.

The SSE patterns of ER2, ER4, and ER5 are similar (Fig. 3) while those of ER3 and ER6 are dissimilar to each other and are only superficially similar to those of ER2, ER4, and ER5.

Comparisons between KR and ER families

As previously mentioned, ER1 was added to KR1 in the ThYme database because all of its primary structures isolated by BLAST were also found in the latter family (Cantu et al. 2012). In addition, all four

KR families (KR1, KR3, and KR4 in clan KR-A plus KR2) and three of the five remaining ER families (ER2, ER4, and ER5) have Rossmann folds, and KRs and ERs all use NADH or NADPH to reduce acyl-ACP and acyl-CoA moieties. However, the catalytic residues in the various families vary, with KR1, KR3, and KR4 having serine, tyrosine, and lysine, KR2 using histidine and glutamate, ER2 having tyrosine and lysine, ER4 using lysine and aspartate, and ER5 using tyrosine and tryptophan (Cantu et al. 2012). These facts in general indicate that superimposing the structures of these KR and ER families and comparing the positions of their SSEs are needed to show if they are related.

Superimpositions of tertiary structures of KR1 and ER2 are extremely close (Fig. 6a). Their catalytic lysine residues are close while their catalytic tyrosine residues are aligned at different angles, with their hydroxyl groups being closest to each other (Fig. 6b). KR1 has a catalytic serine residue not identified in ER2. Superimpositions of tertiary structures of KR3 and KR4, the other two members of clan KR-A, with those of ER2, ER4, and ER5 are in general close, with those between KR4 and ER4 and ER5 being more distant. The catalytic residues of ER2 can be superimposed on those of KR3 and KR4, although not as closely as with KR1. ER4 and ER5 have catalytic residues in different locations than those of KR1, KR3, and KR4. The KR2 tertiary structure, also with a Rossmann fold but not part of clan KR-A, cannot be well superimposed with the Rossmann folds of ER2, ER4, and ER5.

All RMSDs between KR1, KR3, KR4, ER2, ER4, and ER5 are $\leq$2.25 Å (Table 4). The average of RMSD values of KR2 is higher than that of the three other KR families with the three ER structures. Values of *P* are very variable, again often because of the different lengths of the family members in the two clans.

SSEs of the four KR and three ER families with Rossmann folds are fairly well aligned (Fig. 3).

These findings suggest that ER2 has a distant common ancestor with KR1, KR3, and KR4, the three members of clan KR-A, or has convergently evolved with them, reinforcing the similarities between members of the ER and KR enzyme groups already noted when the former ER1 was incorporated into KR1. This is reasonable, since both enzyme groups catalyze reducing reactions, and they are performed on similar substrates.

Comparisons between HD and TE families

The fact that five of the six HD families with known tertiary structures (all but HD2) have HotDog folds brings up the possibility of a link with those thioesterases (TEs), also part of the fatty acid and polyketide synthetic pathways, that have HotDog folds. These include families TE4–TE15, TE24, and TE25 (Cantu et al. 2010, 2011). This would be analogous to the link between ER2 and clan KR-A, but it is less likely, as HDs and TEs catalyze quite different reactions.

At present, two TE clans of families with HotDog folds have been defined, one of TE5, TE9, TE10, and TE12, and the other of TE8, TE11, and TE13 (Cantu et al. 2010). The first clan has one main α-helix and four main β-strands, while the second has one main α-helix, a peripheral α-helix parallel to it, and six main β-strands. Members of both clans attack acyl-CoA substrates, as do most TEs with HotDog folds, but TE14 and TE15 members are specific for acyl-ACP substrates and TE24 and TE25 members are almost all uncharacterized (Cantu et al. 2011). TEs with HotDog folds have a variety of catalytic residues.

Tertiary structures representing the five HD families with known HotDog folds were superimposed individually on the tertiary structures of the thirteen TE families with HotDog folds (TE7 members are known to have HotDog folds, but none has a fully characterized tertiary structure). The HD4 structure could not be superimposed on any TE structure, nor could the TE4 structure on any HD structure. Other pairwise superimpositions are often quite close. However, in no case could catalytic residues of HDs be satisfactorily superimposed on those of TEs. The closest were the catalytic histidine residues of HD1, HD5, and HD6 on the catalytic aspartate residues of clan TE-A (TE5, TE9, TE10, and TE12). This suggests that the evolutionary relationships between HDs and TEs are fairly distant.

RMSD values of individual pairings between HD and TE tertiary structures are usually fairly low (Supplementary Table 1). Values of *P* are relatively high when chain lengths of the two structures being superimposed are roughly the same [short (117–179 residues), as they are with HD1, HD5, and HD6 and with all the TEs except TE4 and TE14, and otherwise long (225–297 residues)].

The SSEs of all HD and TE families with HotDog folds were aligned (Supplementary Fig. S1). The SSEs

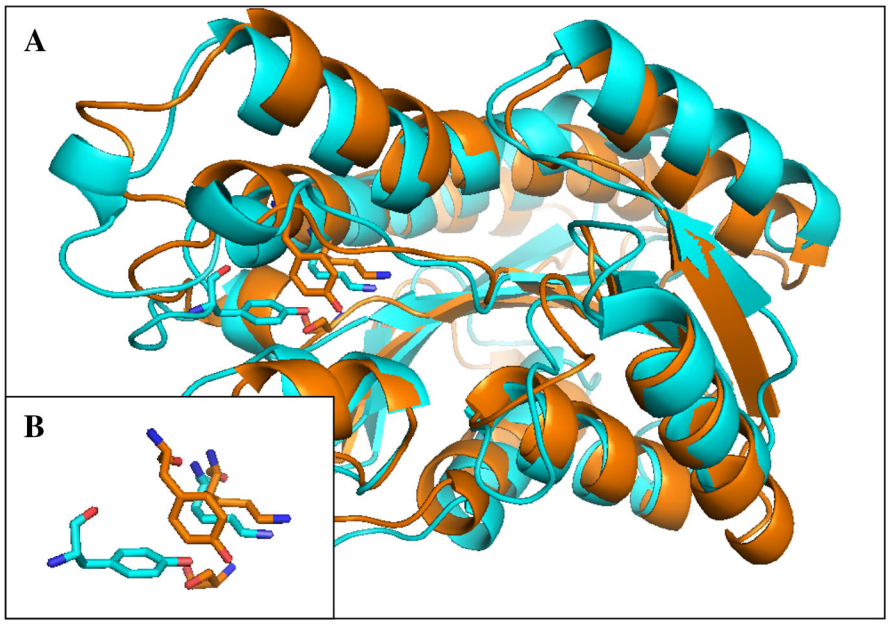

Fig. 6 **a** Superimposed representative tertiary structures of KR1 (PDB 1EDO β-ketoacyl-ACP reductase from *Brassica napus*, *orange*) and ER2 (PDB 1C14 enoyl-(ACP) reductase (NADH) from *Escherichia coli*, *cyan*). **b** Superimposed active-site side chains of the same KR1 and ER2 family representatives, with colors as in A

Table 4 RMSD and *P* values between different KR and ER families with Rossman folds

Families		ER2 (PDB 1C14)	ER4 (PDB 2VZ8)	ER5 (PDB 2VCY)
KR1 (PDB 1EDO)	RMSD (Å)	1.68	2.01	2.14
	P (%)	91.4	45.1	52.5
KR2 (PDB 1WDK)	RMSD (Å)	2.35	2.2	2.06
	P (%)	55.5	53.1	58.6
KR3 (PDB 2PFF)	RMSD (Å)	2.05	1.97	2.05
	P (%)	66.8	55	55
KR4 (PDB 2VZ8)	RMSD (Å)	1.87	2.24	2.25
	P (%)	59.8	56.1	57.5

PDB notations: Protein data bank identification codes for protein tertiary structures

of HD1, HD5, and HD6 match those of most of the TE families, except TE4, TE24, and less so TE25, much better than do those of HD3 and HD4.

Evolutionary implications

Clearly enzymes in the same family, with homologous primary structures leading to very similar secondary and tertiary structures, have the same common ancestor and have been formed by divergent evolution. If the primary structures of a sufficient number of family members are available, multisequence alignment of them may lead toward the primary structure of the ancestral protein. Family members may be produced by one or more domains of life, as shown in the ThYme database (Cantu et al. 2011) and in other databases organized in the same fashion. Genes coding for the same protein cannot only mutate but can migrate to very different organisms.

As mentioned earlier, enzymes in the same clan, defined as those having the same tertiary structure and much the same secondary structure but usually with little if any remaining similarity in primary structure, can either have evolved convergently from different ancestors, or have evolved divergently from a more distant common ancestor. The question before us is whether we can use information about the different families in the KR and HD clans that we have defined to choose between these two evolutionary pathways.

In clan KR-A, family KR1 has a very large number of members showing many different substrate specificities (Cantu et al. 2011, 2012), which is evidence of divergent evolution to these specificities. Those involved in fatty acid metabolism are primarily 3-ketoacyl-ACP reductases. Different family members are produced by bacteria, eukaryota, and archaea. They are found separate from other protein members of the fatty acid and polyketide synthesis cycles. Families KR3 and KR4 are much smaller and are

composed of 3-ketoacyl-ACP reductases and related enzymes. They are both produced by large numbers of eukaryotic and bacterial species. Family KR3 members are part of fatty acid synthase multifunctional proteins, while KR4 members are found in polyketide synthases and, to a lesser extent, in fatty acid synthases.

In clan HD-A, members of family HD5 are produced almost exclusively by proteobacteria with a few eukaryota (Cantu et al. 2011, 2012). Overwhelmingly they are 3-hydroxyacyl-ACP dehydratases. They preferentially attack medium-length substrates. HD6 members come from bacteria and eukaryota and also are 3-hydroxyacyl-ACP dehydratases. They are most active on short- and long-chain substrates.

Given that these two clans are composed of families whose members, except for the multispecific KR1, have similar substrate specificities, and that their families are produced by at least two domains of life, evidence of convergent evolution from different enzymes is lacking. The great similarity of tertiary structures in different families of the same clan, even in loops very far from the active site (Figs. 2 and 4), gives some support to divergent evolution from a distant common ancestor.

Turning to similarity of tertiary structures, all four KR families have Rossmann folds with three of them in clan KR-A and one not. This suggests that, while all four KR families may be descended from a common very distant Rossmann-fold parent, three have a less distant common ancestor. Five of the six HD families with known tertiary structures have HotDog folds. These five families comprise two in clan HD-A and three others not part of a second clan, therefore with perhaps four different distant common ancestors. The sixth HD family, with a ClpP/crotonase fold, must have joined the others as an HD through convergent evolution. The five ER families are composed of three with Rossmann folds and two with (α,β)-barrel folds, not closely related to each other, suggesting two very distant common ancestors that have adopted the same substrate specificities by convergent evolution.

Two further cases are presented here. The first is a comparison of the four KR families and three ER families with Rossmann folds. Here the three families of clan KR-A but not the fourth KR family are closely related in secondary and tertiary structure with family ER2. Along with the assumption of ER1 into KR1 because of their related primary structures, this suggests a distant common ancestor for the three families in clan KR-A with two ER families, with mutations to differentiate their substrate specificities into the two enzyme groups.

The final case concerns the comparison of the five HD and thirteen TE families with HotDog folds. The very diffuse results, with all but one each of the HD and TE families having similar tertiary structures and all but four or five families having similar secondary structures, suggests that they are all descended from a very distant common HotDog ancestor. No closer relationship seems possible, as in general catalytic residues vary between the two enzyme groups, and they are usually in different locations in their active sites.

Conclusions

We have used manual and statistical means to define two clans, one in each of the KR and HD families. Each clan has members with closely superimposable tertiary structures and catalytic amino acid residues, as well as with SSEs in the same locations on the protein chain. This implies that different families in these clans have common ancestors, but that they are sufficiently distant that members of these different families no longer have significantly homologous primary structures. Furthermore, we have found that one KR clan is related to one ER family, all members having Rossmann folds and catalyzing reducing reactions. The evolutionary distance between HDs and TEs having HotDog folds but catalyzing different reactions is more distant.

Supporting information Supplementary Table 1-RMSD and *P* values between different HD and TE families with HotDog folds.

Supplementary Figure 1-Order of secondary structure elements of HDs and TEs with HotDog folds.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Mol Biol 215:403–410

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acid Res 28:235–242

Cantu DC, Chen Y, Reilly PJ (2010) Thioesterases: a new perspective based on their primary and tertiary structures. Protein Sci 19:1281–1295

Cantu DC, Chen Y, Lemons ML, Reilly PJ (2011) ThYme: a database for thioester-active enzymes. Nucleic Acid Res 39:D342–D346

Cantu DC, Dai T, Beversdorf ZS, Reilly PJ (2012) Structural classification and properties of ketoacyl reductases, hydroxyacyl dehydratases and enoyl reductases. Protein Eng Des Sel 25:803–811

Carvalho CC, Phan NN, Chen Y, Reilly PJ (2014) Carbohydrate binding module clans. Submitted for publication

Chen Y, Kelly EE, Masluk RP, Nelson CL, Cantu DC, Reilly PJ (2011) Structural classification and properties of ketoacyl synthases. Protein Sci 20:1659–1667

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The Carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acid Res 42:D490–D495

Schrödinger LLC (2014) Portland, OR. http://www.pymol.org/

Shatsky M, Nussinov R, Wolfson HJ (2004) A method for simultaneous alignment of multiple protein structures. Protein Struct Funct Bioinf 56:143–156

The MathWorks, Natick, MA (2014) http://www.mathworks.com/products/matlab/