Structure-based Phylogenetic Analysis of Short-chain Alcohol Dehydrogenases and Reclassification of the 17beta-Hydroxysteroid Dehydrogenase Family

Rainer Breitling, Daniela Laubner, and Jerzy Adamski

Institute of Experimental Genetics, Genome Analysis Center, GSF-National Research Center for Environment and Health, Neuherberg, Germany

Short-chain alcohol dehydrogenases (SCAD) constitute a large and diverse family of ancient origin. Several of its members play an important role in human physiology and disease, especially in the metabolism of steroid substrates (e.g., prostaglandins, estrogens, androgens, and corticosteroids). Their involvement in common human disorders such as endocrine-related cancer, osteoporosis, and Alzheimer disease makes them an important candidate for drug targets. Recent phylogenetic analysis of SCAD is incomplete and does not allow any conclusions on very ancient divergences or on a functional characterization of novel proteins within this complex family. We have developed a 3D structure-based approach to establish the deep-branching pattern within the SCAD family. In this approach, pairwise superpositions of X-ray structures were used to calculate similarity scores as an input for a tree-building algorithm. The resulting phylogeny was validated by comparison with the results of sequence-based algorithms and biochemical data. It was possible to use the 3D data as a template for the reliable determination of the phylogenetic position of novel proteins as a first step toward functional predictions. We were able to discern new patterns in the phylogenetic relationships of the SCAD family, including a basal dichotomy of the 17beta-hydroxysteroid dehydrogenases (17beta-HSDs). These data provide an important contribution toward the development of type-specific inhibitors for 17beta-HSDs for the treatment and prevention of disease. Our structure-based phylogenetic approach can also be applied to increase the reliability of evolutionary reconstructions in other large protein families.

Introduction

The short-chain alcohol dehydrogenase (SCAD) family (SCOP family; Tyrosine-dependent oxidoreductases; CATH superfamily 3.40.50.720, nucleotide binding Rossmann-like domains) is one of the largest and most diverse of protein families (Jörnvall 1999). It contains mostly NAD(P)- or NAD(P)H-dependent oxidoreductases, catalyzing the conversion of a wide variety of chemical compounds, including important vertebrate signaling molecules, such as steroid hormones, prostaglandins, and retinoids. Typical family members have a monomer size of about 250 amino acid residues, and are characterized by an N-terminal Thr-Gly-(X)3-Gly-X-Gly cofactor binding site and a Tyr-(X)2-Lys active center motif. The overall similarity among family members is low, usually in the range of 15%–30% identity for non-orthologous proteins (Persson et al. 1999). More distantly related forms (extended SCAD family) contain about 350 amino acids per subunit and usually exhibit dehydratase, epimerase, or isomerase activity toward nucleotide sugar substrates (Jörnvall et al. 1995). Hidden Markov models (Bailey, Baker, and Elkan 1997; Grundy et al. 1997) identified several further relatives of the extended SCAD family, including an mRNA-binding ribonuclease from spinach (CSP41, Baker, Grundy, and Elkan 1998), a human transcriptional coactivator (TIP30, Baker, 1999), and a subunit of the mitochondrial proton-translocating NADH:ubiquinone oxidoreductase (complex I, Baker, Grundy, and Elkan 1999). Structural comparisons of the nucleotide binding fold (Rossmann fold) have revealed even more distant relationships with other NAD(P)-dependent dehydrogenases (Rossmann, Moras, and Olsen 1974) and with S-adenosyl-L-methionine–dependent methyltransferases (Lauster 1989; Buñicki 1999).

The low degree of amino acid conservation is, however, a major obstacle for reconstructing evolutionary relationships within the SCAD family. Attempts to identify a robust deep-branching pattern by analyzing multiple-sequence alignments were unsuccessful (Persson et al. 1999). To escape from the twilight zone of protein sequence alignments (Rost 1999) we decided to exploit the large numbers of atomic coordinates available for SCAD family members, assuming that 3D structures are more conserved than primary sequence and should also reflect phylogenetic patterns. This approach has been pioneered by Rossmann et al. (Eventoff and Rossmann 1975; Matthews and Rossmann 1985), as well as by Johnson, Sutcliffe, and Blundell (1990), and more recently by Buñicki (Buñicki 1999, 2000), who used 3D comparisons to determine the phylogeny of protein families. Examination of a limited number of Rossmann-fold dehydrogenases (but no SCAD protein) demonstrated the general feasibility and reliability of the method, but because of the small data set, only very general conclusions were possible at that time. Since then the number of available 3D structures has expanded rapidly, including a large number of SCAD proteins, beginning with rat dihydropteridine reductase in 1992 (Varughese et al. 1992). We used these data to determine a reliable evolutionary tree of the SCAD proteins. In this way it was possible for the first time to get a com-
comprehensive structure-based insight into structural relationships within this important family.

An important further objective was to develop a method to integrate proteins with unknown 3D structure into the structure-based scaffold, using multiple-sequence alignments. This approach was then applied to the 17beta-hydroxysteroid dehydrogenases (17beta-HSDs), one of the best characterized subgroups of the SCAD family that activates and inactivates vertebrate sex hormones by reduction and oxidation at position 17 of the steroid backbone (Peltoketo et al. 1999; Duax, Ghosh, and Pletnev 2000; Labrie et al. 2000a; Adamski and Jakob 2001). Ten different isoforms of 17beta-HSDs have been described so far in vertebrates, and the number is still growing. Isotypes of 17beta-HSDs differ in tissue distribution, substrate specificity, and preferred direction of catalysis. They generally show a low level of sequence conservation (<30% identity). Dysregulation of several members of this subfamily has been implicated in a variety of human diseases, such as pseudohemaphroditism (17beta-HSD3: Geissler et al. 1994), Zellweger-like syndrome (17beta-HSD4 = MFP2: van Grunsven et al. 1998), polycystic kidney disease (17beta-HSD8 = HKE6: Fomiticheva et al. 1998), Alzheimer disease (17beta-HSD10: He et al. 1999), osteoporosis (17beta-HSD4: Jakob et al. 1997; Janssen et al. 1999), and a wide variety of endocrine related cancers (17beta-HSD1, 17beta-HSD2, and others: English et al. 2000; Labrie et al. 2000b; Sasano et al. 2000).

Phylogenetic analysis of the 17beta-HSDs has so far been limited in scope and included only the closest relatives of some isozymes (retinol dehydrogenases, 11beta-hydroxysteroid dehydrogenases: Krozowski 1992, 1994; Baker 1996, 1998, 2001; Lanisnik Rizner et al. 1999) or an arbitrary selection of other SCAD proteins (Grundy et al. 1997). The small number of proteins surveyed and the lack of an obvious outgroup for the 17beta-HSDs seriously confounded the analysis, hampering any reliable evolutionary reconstruction. As a detailed understanding of the relationships of different 17beta-HSDs is a prerequisite for functional assignments of new HSDs (Breitling et al. 2001a, 2001b) and for the rational design of drugs targeting selected isoforms (Tremblay and Poirier 1998; Ngatcha, Luu-The, and Poirier 2000; Penning et al. 2001; Poirier et al. 2001), we decided to use the structural data for a comprehensive reevaluation of the evolutionary history of 17beta-HSDs.

### Materials and Methods

Delimitation of the SCAD Family and Development of a Consensus Tree

Structural coordinates for members of the extended SCAD family were extracted from the Protein Data Bank (PDB, Berman et al. 2000) using pairwise 3D alignments calculated by the DALI program (Holm and Sander 1995), using the X-ray structure of 17beta-HSD type 1 as template. Further, DALI searches with additional members of the SCAD family were used to corroborate the results. As several enzymes were present in multiple data sets, either as cocrystals with ligands or as point mutants, we decided to use only those data from wild-type proteins with the highest resolution and quality according to the R-factor (table 1). The DALI algorithm was also used for the pairwise comparison and alignment of all SCAD protein structures. The DALI-Z-score (indicating the significance of the similarity in standard deviations above database average) as well as the root mean square distance (the square root of the average squared Euclidean distance over all topologically equivalent pairs of alpha-carbon positions) were used as indicators of protein similarity, resulting in two distance matrices for all SCAD structures. Both matrices were used to calculate phylogenetic (similarity) trees using the Fitch-Margoliash algorithm (Fitch and Margol

### Table 1

<table>
<thead>
<tr>
<th>SCAD Family X-ray Structures Used for the Phylogenetic Analysis, PDB = Database Identifier for Coordinate File</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PDB</strong></td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>1AE1</td>
</tr>
<tr>
<td>1AEn</td>
</tr>
<tr>
<td>1B16</td>
</tr>
<tr>
<td>1BBD</td>
</tr>
<tr>
<td>1BXX</td>
</tr>
<tr>
<td>1CD1</td>
</tr>
<tr>
<td>1DB3</td>
</tr>
<tr>
<td>1DFG</td>
</tr>
<tr>
<td>1DHR</td>
</tr>
<tr>
<td>1ENP</td>
</tr>
<tr>
<td>1ENY</td>
</tr>
<tr>
<td>1FDS</td>
</tr>
<tr>
<td>1FCM</td>
</tr>
<tr>
<td>1GFS</td>
</tr>
<tr>
<td>1HDC</td>
</tr>
<tr>
<td>1OAA</td>
</tr>
<tr>
<td>1QR1</td>
</tr>
<tr>
<td>1QTL</td>
</tr>
<tr>
<td>1YBV</td>
</tr>
</tbody>
</table>
liash 1967) as implemented in the FITCH program of the PHYLIP package (Felsenstein 1989). The structure-based alignment provided by DALI or multiple-sequence alignments calculated by the Clustal W algorithm (Thompson, Higgins, and Gibson 1994) were used as input for consensus-based secondary structure prediction using the JPred algorithm (Cuff et al. 1998).

Integration of the 17beta-HSD Enzymes into the Phylogenetic Tree

To allow the reliable insertion of new sequences into the structure-based phylogenetic tree, we employed a multi-step procedure. For every sequence to be integrated, we collected a number of homologous sequences that were sufficiently similar to enable a reliable alignment. The sequences were aligned by two iterations of PSI Blast (Altschul et al. 1997) and filtered for redundancy at the 75% identity level at the JPred server. Sequence fragments and ambiguously aligned sequences were removed from the alignment after visual inspection. The secondary structure was predicted for the complete alignment using the JPred consensus method. For illustration, a representative output of the JPred algorithm as employed on the structure-based alignment (i.e., performing under optimal conditions) is included in the study (see Supplementary Material, figs. 2 and 3).

The prediction result was used for manual alignment of the multiple-sequence alignment with that of the crystal structures. The combined sequence alignment was used as an input for the parsimony algorithm. Parsimony reconstruction was the method of choice, as the results of this method agreed best with the structure-based and consensus trees (see later). The most parsimonious tree was inspected visually to determine the most likely position of the sequence of interest. The homologous sequences which were used to avoid the distortion of the results by spurious similarities were not included in the final result. To illustrate the potential of this method, the procedure was applied to all vertebrate 17beta-HSDs. Maximum likelihood trees and maximum likelihood mapping were done using the quartet puzzling algorithm implemented in the TREE-PUZZLE program (Strimmer and von Haeseler 1997).

Results

The Structure-based Trees

The database search identified X-ray structures of 19 members of the extended SCAD family. These structures form a well-delimited cluster with pairwise Z-scores larger than 13, whereas the similarity of other Rossmann-fold dehydrogenases is below Z = 10. Both structure-based trees are moderately resolved with very short internal branches (fig. 1). The tips of all branches are almost equidistant from the tree center in both reconstructions, indicating a uniformly ticking structural clock within the SCAD family.

Comparison with Sequence-based Trees

The alignment of structurally equivalent residues in all 19 X-ray structures defined a conserved 160 amino acid structural core of SCAD proteins. A maximum likelihood mapping of this alignment indicated that a sequence-based reconstruction of phylogeny would yield only a moderately resolved tree: 18.6% of the likelihood vectors are located in the unresolved central area of the likelihood map (fig. 2). This observation agrees well with the low degree of sequence conservation close to mutational saturation at most positions, with the exception of the catalytically indispensable sites. The multi-
parsimony tree. Eight clades are also supported by the chotomies, all of which are also present in the maximum solved tree is presented in figure 4. It contains 10 di-
are fairly independent. The resulting moderately re-
and assumed that the results of all the five calculations
a priori which reconstruction method is the most reliable
results sharing 10 and 5 dichotomies with the parsimony method agrees best with the structure-based
same-score structure-based tree, whereas the other two
Dots in the corners indicate branches for which one topology has a
monophyly of bacterial enoyl-acyl±carrier protein re-
ment is between Fitch-Margoliash and maximum par-
sugar converting enzymes). The highest degree of agree-
from a super®cial analysis of the sequences (dichotomy
zymatic data (monophyly of tropinone reductases and
the constraints of the algorithms, contradicting the re-
results sharing 10 clades, whereas the maximum
was previously unexamined, as this group was consid-
ered paraphyletic (Baker 2001). Our data indicate that the
oxidative isoforms 17beta-HSD4, 17beta-HSD8, and
17beta-HSD10 are members of a monophylum with re-
sence of this approach is demonstrated by the superior
consensus tree of ®ve independent attempts to
classes are common only to the three sequence-based reconstructions.
Integration of the 17beta-HSD Enzymes
Figure 5 shows nine sequences coding for 17beta-
HSD enzymes integrated in the obtained scaffold tree. The main new feature is a clustering of the 17beta-HSDs in two large, independent groups, which was not evident
by means of conventional enzymatic data or previous phylogenetic analysis (Jörnvall 1999; Baker 2001 and references therein). Group 1 consists of 17beta-HSD type 7, 2, 6, and 9 plus the only crystallized isoform, 17beta-HSD type 1. The second group includes 17beta-
HSD type 3, 4, 8, 10. Also the topology within each group differs in several points from the previously
hypothesized pattern (Baker 2001). In group 1, 17beta-
HSD7 turns out to be the closest relative of 17beta-
HSD1, whereas 17beta-HSD type 2, 6, and 9 form their sister groups. The internal relationships within group 2 were previously unexamined, as this group was consid-
ered paraphyletic (Baker 2001). Our data indicate that the
oxido
tive isoforms 17beta-HSD4, 17beta-HSD8, and
17beta-HSD10 are members of a monophylum with re-
spect to 17beta-HSD3. The nearest neighbor of group 2 among the crystallized proteins is the 7alpha-hydroxy-
steroid dehydrogenase from E. coli.
Discussion
Phylogenetic Relationships Within the SCAD Family
In this paper, we present the ®rst comprehensive, structure-assisted genealogy of SCADs. The ®nal tree is
based on a consensus of ®ve independent attempts to
determine the phylogenetic relationships within the fam-
ily. Therefore, it is considerably less resolved than the individual trees. If one of the methods used to calculate
these trees reconstructed the correct phylogeny, combin-
ing it with the other scenarios would not only decrease the resolution but might also introduce errors. However,
as we cannot decide a priori which method is the most
reliable, the consensus tree should at least provide a way
to visualize areas of agreement and facilitate the com-
parison of the different results. The consensus tree of
the SCAD structures evinces much more information
than is available by a superficial inspection of their se-
quencies. Although all of this information is shared with the
maximum parsimony tree, it should be kept in mind
that the high quality of the sequence-based trees is de-
pendent on an unambiguous alignment which was avail-
able only from the structural comparisons. One further
reason for the good performance of the parsimony meth-
od may be the action of a uniform molecular clock, as
indicated by the results of the structure-based analysis
(Steel and Penny 1993). We used the Z-score (Holm and
Sander 1997), instead of root mean square deviation, to
evaluate similarities for tree reconstructions. The suc-
cess of this approach is demonstrated by the superior
congruence with the sequence-based results and the bet-
ter agreement with enzymological and phylogenetic data
as compared with the root mean square-based tree (e.g.,
monophyly of bacterial enoyl-acyl±carrier protein re-

Fig. 2.—Maximum likelihood mapping of the structure-based se-
quen
ple-sequence alignment was used to calculate phyloge-
netic trees using three different tree-reconstruction
methods (®g. 3). The comparison of these trees with the
two structure-based reconstructions is shown in table 2.
Both Fitch-Margoliash and maximum parsimony anal-
ysis yield almost completely resolved trees, because of
the constraints of the algorithms, contradicting the re-
results of the maximum likelihood mapping and indicating
that several of the reconstructed nodes are probably
artificial. On the other hand, maximum likelihood analysis
yields only a poorly resolved tree, with only 7 dichot-
omies, most of which are already obvious from the en-
zymatic data (monophyly of tropinone reductases and
enoyl-acyl±carrier protein reductases, respectively) or
from a superficial analysis of the sequences (dichotomy
between SCADs in the strict sense and the nucleotide
sugar converting enzymes). The highest degree of agree-
ment is between Fitch-Margoliash and maximum par-
simony trees sharing 10 clades, whereas the maximum parsimony method agrees best with the structure-based results sharing 10 and 5 dichotomies with the Z-score
tree and root mean square tree, respectively (table 2).
Consensus Scenario
The five phylogenetic trees were condensed into a
single majority-rule consensus tree. We did not decide
a priori which reconstruction method is the most reliable
and assumed that the results of all the five calculations
are fairly independent. The resulting moderately re-
solved tree is presented in figure 4. It contains 10 di-
}
ductases). It should also be noted that the tree based on Z-score similarity shows a more uniform rate of evolution, indicating a better reflection of the true evolutionary distances. This might be because of the fact that the Z-score takes into account the length of the equivalent segments of protein structure, whereas the root mean square deviation is calculated only for equivalent alpha-carbons defined by an arbitrary and floating cut-off of 3 Å and ignoring the length of the compared sequence (Holm and Sander 1997).

Integration of Multiple-sequence Alignments and the Structure-based Tree

The comparison of atomic coordinates of members of the SCAD family yielded not only a phylogenetic tree but also a protein sequence alignment that reliably identified equivalent amino acid residues. Our results show that a maximum parsimony analysis of this alignment determined a phylogenetic pattern that was largely congruent with the structure-based tree. Because of this, it was possible to use sequence-based algorithms to integrate proteins with unknown 3D structure into the evolutionary scenario. The placement of the 17beta-HSDs within the SCAD tree (which in every case reflects the placement of a large number of homologous sequences examined in parallel) indicates that the results are most likely not distorted by long-branch attraction, the most common problem of parsimony tree reconstructions (Kuhner and Felsenstein 1994). This might be because of the uniform rate of evolution on all branches as indicated by the structure-based trees, as well as by the use of a large number of homologous sequences that are

Fig. 3.—Sequence-based phylogenetic trees derived from the structure-based sequence alignment using different algorithms. 3: Neighbor-Joining (100 bootstrap pseudoreplicates); 4: maximum parsimony (100 bootstrap pseudoreplicates); 5: maximum likelihood. Red bars and numbers indicate which branches are shared between different methods of tree reconstruction (compare fig. 1). The bootstrap support is represented by the branch lengths.
Table 2
Number of Clades Shared by the Different Tree Reconstructions

<table>
<thead>
<tr>
<th></th>
<th>Z-score</th>
<th>rms</th>
<th>NJ</th>
<th>MP</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z-score</td>
<td>15/8</td>
<td>16/5</td>
<td>16/5</td>
<td>15/10</td>
<td>7/6</td>
</tr>
<tr>
<td>rms</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>NJ</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>MP</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>7/6</td>
</tr>
<tr>
<td>ML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.—Numbers in the central diagonal: total number of dichotomies/contribution to the consensus tree. rms: root mean square; NJ: Neighbor-Joining; MP: maximum parsimony; ML: maximum likelihood. The structure-based trees are in bold face.

expected to disrupt very long branches. The obtained topology is therefore likely to be very reliable.

Reclassification of 17beta-HSDs

From previous sequence-based analysis it was obvious that 17beta-HSD activity arose convergently in all 17beta-HSD isoforms (Baker 1996). Thus, the subsumption of 17beta-HSDs in two subgroups arising independently in different regions of the SCAD tree was unexpected. It indicates that the aptitude to develop an activity toward position 17 of steroid substrates arose only twice within the SCAD family. Both groups of 17beta-HSDs include oxidative and reductive enzymes. In both the cases the oxidative (steroid inactivating) enzymes form a separate subgroup. This agrees well with the hypothesis of a shared ancestral substrate for these proteins: retinols in the case of 17beta-HSD type 2, 6, and 9 (Baker 1998) and fatty-acids in the case of 17beta-HSD type 4, 8, and 10 (Baker 2001). The functional and structural connection between the oxidative subgroups and their respective reductive paralogues has yet to be determined.

Conclusions

The approach presented here provides a reliable scaffold for comparative analyses of SCAD proteins. By integrating protein sequences in this scaffold it is possible to make basic functional predictions on novel proteins, and the choice of optimal templates for homology

![Diagram](image-url)
Fig. 5.—Phylogenetic tree of the extended SCAD family including vertebrate 17beta-HSDs. 17beta-HSD enzymes with predominantly oxidative, steroid-inactivating activity are marked with an asterisk. The placement of every 17beta-HSD is based on the data from a large number of homologous proteins, most of which do not have 17beta-HSD activity. These homologous sequences are excluded from the tree for simplicity. Colors as in figure 4.

modeling is facilitated. The approach can easily be expanded to include further members of the SCAD family and can also be applied to the analysis of other large protein families. Especially, it provides a basis for the analysis of the structural determinants responsible for the evolution of 17beta-HSD activities in two independent subgroups of the SCAD family. This will serve as an important contribution toward the development of type-specific inhibitors of selected 17beta-HSDs for the treatment and prevention of endocrine-related cancers, osteoporosis, and Alzheimer disease.

Supplementary Material

The structure-based alignment and the distance matrices derived from the comparison of short-chain alcohol dehydrogenase X-ray structures are available in electronic form from the MBE website at http://www.molbiolevol.org/.

Acknowledgment

This work was supported in part by a DFG grant to J.A.

LITERATURE CITED


William Taylor, reviewing editor

Accepted July 20, 2001