An Evolutionary Treasure: Unification of a Broad Set of Amidohydrolases Related to Urease

Liisa Holm* and Chris Sander
European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Cambridge, United Kingdom

ABSTRACT  The recent determination of the three-dimensional structure of urease revealed striking similarities of enzyme architecture to adenosine deaminase and phosphotriesterase, evidence of a distant evolutionary relationship that had gone undetected by one-dimensional sequence comparisons. Here, based on an analysis of conservation patterns in three dimensions, we report the discovery of the same active-site architecture in an even larger set of enzymes involved primarily in nucleotide metabolism. As a consequence, we predict the three-dimensional fold and details of the active site architecture for dihydroorotases, allantoinases, hydantoinases, AMP-, adenine and cytosine deaminases, imidazolonepropionase, aryldialkylphosphatase, chlorohydrolases, formylmethanofuran dehydrogenases, and proteins involved in animal neuronal development. Two member families are common to archaea, eubacteria, and eukaryota. Thirteen other functions supported by the same structural motif and conserved chemical mechanism apparently represent later adaptations for different substrate specificities in different cellular contexts. Proteins 28:72–82, 1997 © 1997 Wiley-Liss, Inc.

Key words: protein family analysis; genome analysis; homology modeling; molecular evolution; protein structure comparison

INTRODUCTION

Inference by homology is currently the most powerful computational method of function assignment and structure prediction for the protein products of new genes. The success of the method is based on two remarkable evolutionary phenomena. Protein structure can be conserved over long evolutionary distances, in spite of strong sequence divergence; and, embedded in conserved 3D structure, proteins can adapt their biochemically active site to catalyze reactions on a variety of substrates. In recent years, these phenomena have been amply illustrated by the discovery of many unexpected distant evolutionary relationships as a result of systematic comparison of three-dimensional protein structures.1 For example, the catalytic domains of kanamycin nucleotidyltransferase and DNA polymerase-β are structurally similar, although sequence identity is limited to a few key functional residues; translating this similarity into a structure-based position-specific sequence profile to search sequence databases led to the identification of five additional terminal nucleotidyltransferase families as descending from the same ancestor as the initial pair and conserving structure and catalytic principle.2

Here, we discover an evolutionary treasure based on the striking similarity of the enzyme architectures of urease, phosphotriesterase and adenosine deaminase.3 Residue-by-residue optimal alignment and superimposition of three-dimensional structures reveals a common structural core consisting of an ellipsoidal (βα)8 barrel with a conserved metal binding site at the C-terminal end of strands β1, β5, β6, and β8 (Fig. 1). In the common reaction mechanism the metal ion (or ions) deprotonate a water molecule for nucleophilic attack on the substrate. The metal ligands, four histidines and one aspartic acid residue, are strictly conserved in the three enzyme families, and define a subtle but sharp sequence signature of this emerging superfamily.

We set out to detect additional members of the superfamily by computational sequence analysis. The analysis has five steps:

1. Structural alignment and identification of functional residues common to the three seed families
2. Sequence space walk, i.e., searching for neighbors in sequence databases
3. Analysis of conserved patterns/functions within each neighbor family
4. Correlation of these patterns/functions with those extracted from the known structures
5. Multiple alignment of neighbor families with the known structures ("threading") and 3D model building

In the process, we link 10 additional families to urease and adenosine deaminase as mechanistically
and structurally conserved homologues. We summarize their diverse metabolic roles and discuss methodological implications for family analysis in genome research.

**METHODS**

**Structure Alignment**

The three-dimensional structures of urease (2kauC), phosphotriesterase (1pta), and adenosine deaminase (1fkx) were aligned structurally (without reference to amino acid sequences) using the Dali program (Fig. 1). The evolutionary constraints common to the three enzyme families are (1) a precisely defined histidine-aspartic acid signature required for metal binding and catalysis, and (2) a structural context of alternating α-helix and β-strand secondary structure elements in which the functional residues map to the C-terminal end of strands 1, 5, 6, and 8 (Fig. 1).

**Walking in Sequence Space**

There are two ways to explore sequence space between and among structurally identified members of an emerging superfamily. Profiles combining sequence information from remote relatives are a powerful search method if an active site signature is contained in a contiguous stretch of conserved positions, for example, the ATP-binding helix-turn-strand motif in terminal nucleotidyltransferases. Here, we have adopted a second, complementary, strategy of neighbor searching. This approach is based on pairwise comparisons of proteins to identify candidate sequences and profile-profile comparisons to verify consistency of family membership. This strategy exploits “neutral” variation within protein families, which may result in statistically significant overlap (in terms of sequence similarity) between functionally distinct subfamilies. A walk in sequence space...
catalytic domain  small domain

Fig. 1.
space starts from a seed sequence, uses standard search tools (here: Fasta\(^3\) with optimized scores and ktup = 1, searching a nonredundant database of protein sequences) to collect first neighbors, and then branches out collecting second neighbors, that is, those of peripheral members, third neighbors, and so on. Whether a walk explodes (collects spurious members) or results in a closed set containing nontrivial discoveries depends on the bounding constraints, that is, cutoffs used to decide between true and false similarity links.\(^8\) We set the Fasta\(^3\) cutoff for statistically significant links to 0.01 expected hits in a nonredundant protein database of 207,645 sequences. This value has been used as a conservative cutoff by other researchers, based on empirical observations.\(^7,9\) Only the matching domain was used for subsequent search cycles if a match was found to a multidomain protein. Hits were listed as twilight matches if the expectation value was less than 1. Families identified through twilight hits were included in the superfamilies if the signature pattern (bold histidines and aspartic acid in Fig. 1a) was present. A similar set of sequences results from Blast\(^10\) neighboring and is available over the NCBI server (http://www3.ncbi.nlm.nih.gov/Entrez/).

**Verification Steps**

Evolutionary constraints describing the new candidate families were analyzed from family alignments generated by progressive alignment.\(^11\) Automatic programs\(^7,12,13\) were used with default parameters to allow closely related sets of sequences, but for multiple alignments involving different families we took a shortcut through alignment parameter space (gap penalties, substitution matrices, sequence- and position-specific weights\(^13\)) by hand editing. The multiple alignments were used as input for secondary structure predictions for each family by linear discrimination function\(^14\) and neural network\(^15\) methods. The signature patterns were identified by inspection of conservation patterns within families (e.g., see Figs. 2 for example and 3 for summary). Threading the sequences onto the known 3D structures phased on the active site pattern preserved the hydrophobicity of the structural core.\(^16\) Within each family, conserved regions map to the expected structural core, for example, there are conserved blocks in predicted β strands preceding the metal ligands. The full multiple alignment and 3D model coordinates are available over the Internet from http://www.sander.embl-ebi.ac.uk/urease/ and can be viewed graphically using Belvu (E. Sonnhammer, unpublished) and Rasmol.\(^17\)

**Error Detection by Homology Arguments**

Searching protein sequence databases revealed a number of partial matches to the functionally required His-Asp signature pattern or showed grossly nonuniform sequence similarity over the predicted (β\(\alpha_8\)) barrel structural unit relative to sequence relatives. Such conflicts were resolved at the level of gene prediction from DNA sequence. Optimal alignment, using the PairWise (E. Birney, unpublished) program to compare alternative translation frames of the DNA against protein sequences detected frame-shifts, for example, in the nucleotide sequences of the D-hydantoinase gene from Pseudomonas putida (GenPept acc. no. L24157; three frame-shifts), of s-triazine hydrolase (GenPept acc. no. L16534; replacement of a composition biased N terminus), of Haemophilus influenzae gene HI0482 (Swissprot acc. no. P44058; extension of the C terminus), and allantoinase (Swissprot acc. no. P40757, change in C-terminal region). Similarly, a missing exon that contains the N-terminal H×H motif of the signature pattern was identified for the Caenorhabditis elegans gene CEF38E11-3 (EMBL acc. no. Z68342).

**RESULTS**

**A Novel Amidohydrolase Superfamily**

The stepwise search leads to the unification of a large number of remotely related enzyme families with conserved substructures and catalytic principle (Fig. 4). Local regions of sequence similarity have been reported earlier for the pair AMP deaminase and adenosine deaminase,\(^18\) and between dhydroorotases, allantoinases, and hydantoinases.\(^19\) When originally sequenced, arylalkylphosphatase, cytosine deaminase, formylmethanofuran dehydrogenase subunit A, s-triazine hydrolase (a chlorohydrolase), and imidazolonepropionase appeared to be "pioneer" sequences, that is, they had no relatives in the sequence database. The identification of the common signature pattern establishes previously
unknown evolutionary links (Fig. 3). Members of the superfamily catalyze the hydrolysis of amide or, in a few families, amine bonds, in more than a dozen different substrates, and are responsible for 7 of some 20-odd steps along four important metabolic pathways (Fig. 5). Table 1 compares the presence or absence of functional homologues (assigned by sequence similarity to experimentally characterized proteins) in the completely sequenced genomes of Methanococcus jannaschii,20 Haemophilus influenzae,21 Mycoplasma genitalium,22 and yeast (http://speedy.mips.biochem.mpg.de/mips/YEAST/), and the partially sequenced genomes of Escherichia coli and nematode.

Evolutionary Origins

It is plausible that the extended family of urease-related amidohydrolases began to diverge from a common ancestor that had similar structure and biochemical function at a very early evolutionary stage, that is, before the divergence of archaea, prokaryota, and eukaryota. Subsequently, the sequence signature required for the basic catalytic mechanism has remained invariant in spite of considerable functional specialization. Dihydroorotase is a possible ancestral activity of the superfamily, as it is present in diverse species, has many neighbor families in sequence space, and appears more important being a biosynthetic enzyme than the other members, which are catabolic. Dihydroorotases are a diverse enzyme family with three subgroups. In eukaryotes, dihydroorotase is part of a fused trifunctional enzyme (CAD) that catalyzes the first three steps of pyrimidine biosynthesis (Fig. 5). In some dihydroorotases, sequence similarity extends to the small domain of urease in both N- and C-terminal regions relative to the (αβ)8 barrel. By contrast, the N termini of, for example, E. coli dihydroorotase and the yeast URA4 gene product coincide with the start of the catalytic (αβ)8 barrel domain (see Fig. 3). Sequence similarities to the small domain of urease is also seen in a number of other member families. For example, the conserved block GADADLVIWD (Unc-33 sequence in Fig. 2) has 60% identity with the yeast URA4 gene product coincide with the start of the catalytic (αβ)8 barrel domain (see Fig. 3).

Evolution of Metabolic Pathways

The powerful evolutionary potential of the metal-assisted catalytic framework is illustrated by the apparently rapid emergence and perfection, in modern times, of three detoxifying enzyme activities in bacteria (phosphotriesterase, arylalkylphosphatase, s-triazine hydrolase). The wide phylogenetic distribution of genomic homologues of the s-triazine hydrolase gene (trzA) from Rhodococcus corallinus, which is capable of dechlorinating dealkylated metabolites of the herbicide atrazine23 appears surprising (see Table 1). We conjecture that the s-triazine hydrolase activity might actually be a recent adaptation and that the original substrate is an as yet unidentified “natural” compound.
Three closely related aminoacylases (N-acyl-D-glutamate amidohydrolase, D-aminoacylase, N-acyl-D-aspartate amidohydrolase) from Alcaligenes xylosoxidans share significant sequence similarities with three other member families of the superfamily (Fig. 4) but are unrelated to Bacillus and animal aminoacylase sequences. Apparently, the Alcaligenes aminoacylase group represents convergent evolution of similar enzymatic activity on different structural frameworks.

In addition to the invention of new catalytic activities or reinvention of catalytic activities existing in other organisms, metabolic pathways may be truncated or lost during evolution. For example, the pathway of uric acid degradation (see Fig. 5) has been truncated through the successive loss of allantoinase, allantoinase and urate oxidase during phylogenetic evolution of vertebrates. No member of the superfamily has been retained in the parasitic Mycoplasma genitalium. Surprisingly, only two members were identified in Haemophilus influenzae. Dehydroorotase is absent, and, in fact, this organism lacks all genes encoding the first three steps of pyrimidine biosynthesis pathway. E. coli has more functions in common with Methanococcus than with Haemophilus, although the latter is a closer sister species phylogenetically.
Evolution of New Cellular Functions

The definition of the superfamily was initially guided by the signature pattern for the metal center. Surprisingly, three member families contain branches (marked by asterisks in Fig. 2 and Table I) in which the catalytic residues are not conserved, yet family membership is clear at 30–40% sequence identity with the closest relatives. In general, the mutations are correlated in that all four histidines and the aspartic acid have disappeared, although sequence conservation remains very clear in surrounding structural positions. The implication is that these subfamilies no longer function as enzymes but rather reuse the fold for another purpose, presumably another type of biological function. Such evolutionary behavior has numerous precedents, for example, lysozyme/α-lactalbumin, the regulatory subunit of lactose synthetase; serine proteases/haptoglobin, a plasma protein that binds but does not cleave hemoglobin, and repeated recruitments of enzymes as structural proteins in the eye lens. There is some information about the function of the noncatalytic members of the superfamily. A defective dihydroorotase-like domain forms the middle part of the yeast URA2 gene, which is homologous with CADs but has only carbamoyl phosphate synthase and aspartate transcarbamoylase activities. In Pseudomonas putida, inactive dihydroorotase-like subunits are required for the correct assembly of aspartate transcarbamoylase subunits into the dodecameric (6 + 6) holoenzyme. An interesting puzzle is presented by a set of animal proteins involved in neuronal development, typified by the nematode axonal outgrowth and guidance protein unc-33 (Swissprot acc. no. Q01630). These appear to have recently diverged from the hydantoinases, with sequence identity as high as 40%, which implies conservation of fold; but they also have lost the residues characteristic for the metal binding site. The precise role of these proteins in the development of neurons is not yet known. Based on analogy to hydantoinase, it is plausible to propose that their function involves binding (but not catalysis) of a molecule chemically related to dihydrooracil, the substrate of hydantoinase.

**DISCUSSION**

The rapid increase in the number of known three-dimensional protein structures will increase the scope and importance of structure-based pattern identification. The subtle but sharp signature pattern used in this work was based on the structural alignment of urease, phosphotriesterase, and adenosine deaminase, which allowed us to identify a large set of additional relatives. Sequence conservation and structural considerations provide evidence for a common fold shared by the different families of enzymes that supports an active site constrained to perform metal-assisted hydrolysis of amide bonds.
The sequence signature (mapping to $\beta_1$, $\beta_5$, $\beta_6$, and $\beta_8$) binds together both ends of the ($\beta_1\alpha_2$$\beta_6\alpha_1$) structural motif; recognition of homology was based on identifying invariantly conserved functional residues in a structural context of alternating $\alpha$ helices and $\beta$ strands; the fact that predicted member families either are metalloenzymes or simultaneously lack the metal ligands and are known to be catalytically defective is congruent with the identification of active sites and fold prediction. The multiple alignment of the ten new member families implies that three-dimensional models can be built for the more than 70 member sequences by using any of the three known structures as template. The detailed understanding of the active site in the known structures lead to precise predictions concerning mechanism of function. The new functional and structural insights are expected to provide strong impetus to experimental studies of these enzymes.

Evolutionary discontinuity of enzyme function was observed in three groups of the superfamily. Simplistic function assignment based merely on a threshold in sequence similarity can both under- and overpredict function. In the present work, patterns of sequence conservation were examined within each family and scrutinized for consensus between fami-

lies. As a result, we find two examples for which family analysis refines functional assignments made in recent large-scale automated sequence analyses. In the analysis of more than 5000 yeast genes (http://www.sander.embl-heidelberg.de/genequiz/), we find a probable false-positive assignment of AMP deaminase function to two ORFs from yeast (acc. nos. P40361, P38510 in Fig. 3). They are closely related to AMP deaminases by overall sequence similarity, but the subtle effect of losing the metal ligands suggests they are probably catalytically defective. In the other example, the presence of the His-Asp signature pattern confirms the tentatively assigned cytosine deaminase function in H. influenzae (HI0842 in Fig. 3).

Prediction of three-dimensional protein folds from amino acid sequence, using physical principles, remains basically unsolved. This work has exploited analysis of evolutionary constraints by structure and sequence comparisons to arrive at a new fold prediction for dihydroorotase, allantoinase, hydantoinase, cytosine and adenine deaminase, imidazolonepropionate, aryldialkylphosphatase, s-triazine hydrolase, aminoacylase, subunit A of formylmethanofuran dehydrogenase, and proteins involved in guiding animal neuronal development. As experimental struc-

---

**Fig. 5.** Members of the superfamily in the context of metabolic pathways. The superfamily illustrates the evolutionary principle of reusing the same chemical principle in different steps of a set of related pathways. Boxes indicate superfamily members. Dihydroorotase, which catalyzes the third step in pyrimidine biosynthesis, is common to all forms of life (archaea, eubacteria, eukaryota) and it might be closest in function to the most ancient ancestor. Catabolic member enzyme families (of which six map to pathways shown) have a more patchy phylogenetic distribution (cf. Table I), apparently as a result of evolutionary changes in these pathways in some organisms.
tural biology slowly but surely will approach complete coverage of all basic types of three-dimensional protein structures, we believe this family analysis approach combined with model building by homology will eventually be able to provide a plausible structural model for almost any new protein sequence.

ACKNOWLEDGMENTS

We thank Antoine de Daruvar for database updates, Andy Karplus for comments on an early version of the manuscript, and Alexey Murzin for discussions and pointing out the importance of the sequence similarities to the small domain of urease.

REFERENCES

33. Lewis, P.N., Scheraga, H.A. Prediction of structural homol-