PROTEINS: Structure, Function, and Genetics 28:72–82 (1997)

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 $(\beta\alpha)_8$ barrel with a conserved metal binding site at the C-terminal end of strands β_1, β_5, β_6, and β_8 (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

*Correspondence to: Dr. Liisa Holm, EMBL-EBI, Wellcome Trust Genome Campus, Cambridge CB 10 1SD, U.K.
Received 20 May 1996; Accepted 4 November 1996

TABLE I. Superfamily Members Identified in Representative Organisms

	Archaea	Eubacteria		Eukaryota	
Function	*M. jannaschii*	*H. influenzae*	*E. coli*[#]	*S. cerevisiae*	*C. elegans*[#]
adenosine deaminase (E.C.3.5.4.4)			P22333	P53909	CEC06G3_3 CEC44B7_8
AMP deaminase (E.C.3.5.4.6)				P15274 P40361* P38150*	CEC34F11_5
adenine deaminase (E.C.3.5.4.2)	MJU67586_9		P31441		
cytosine deaminase (E.C.3.5.4.1)		P44058	P25524		
urease (E.C.3.5.1.5)		HI00074_60	Q03284		CEUREA_1
hydantoinase			EC28375_23		CER06C7_5
developmental proteins					CEUNC33G_3* CEC47E12_5
dihydroorotase (E.C.3.5.2.3)	MJU67590_2		P05020	P20051 P07259*	CED2085_1
allantoinase (E.C.3.5.2.5)				P32375	
aminoacylase (E.C.3.5.1.-)					
imidazolonepropionase (E.C.3.5.2.7)					CET12A2_8
phosphotriesterase			P45548		
arylphosphatase					
chlorohydrolase	MJU67516_13 MJU67595_3		EC28375_29 EC28375_33	SCYDL238C_1	CEF38E11_3
formylmethanofuran dehydrogenase	MJU67558_13				
total	5	2	9	8	10

Sequences are labelled by Swissprot accession number or Trembl identifier, and classified into functional categories by homology (asterisks denote catalytically defective proteins).
[#]Genome sequencing has not yet been completed at the time of analysis.

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[3]), phosphotriesterase (1pta[4]), and adenosine deaminase (1fkx[5]) were aligned structurally (without reference to amino acid sequences) using the Dali program[6] (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 around 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.[2] 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

A

B

catalytic domain small domain

Fig. 1.

space starts from a seed sequence, uses standard search tools (here: Fasta3[7] 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 Fasta3 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 superfamily 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 align fairly 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

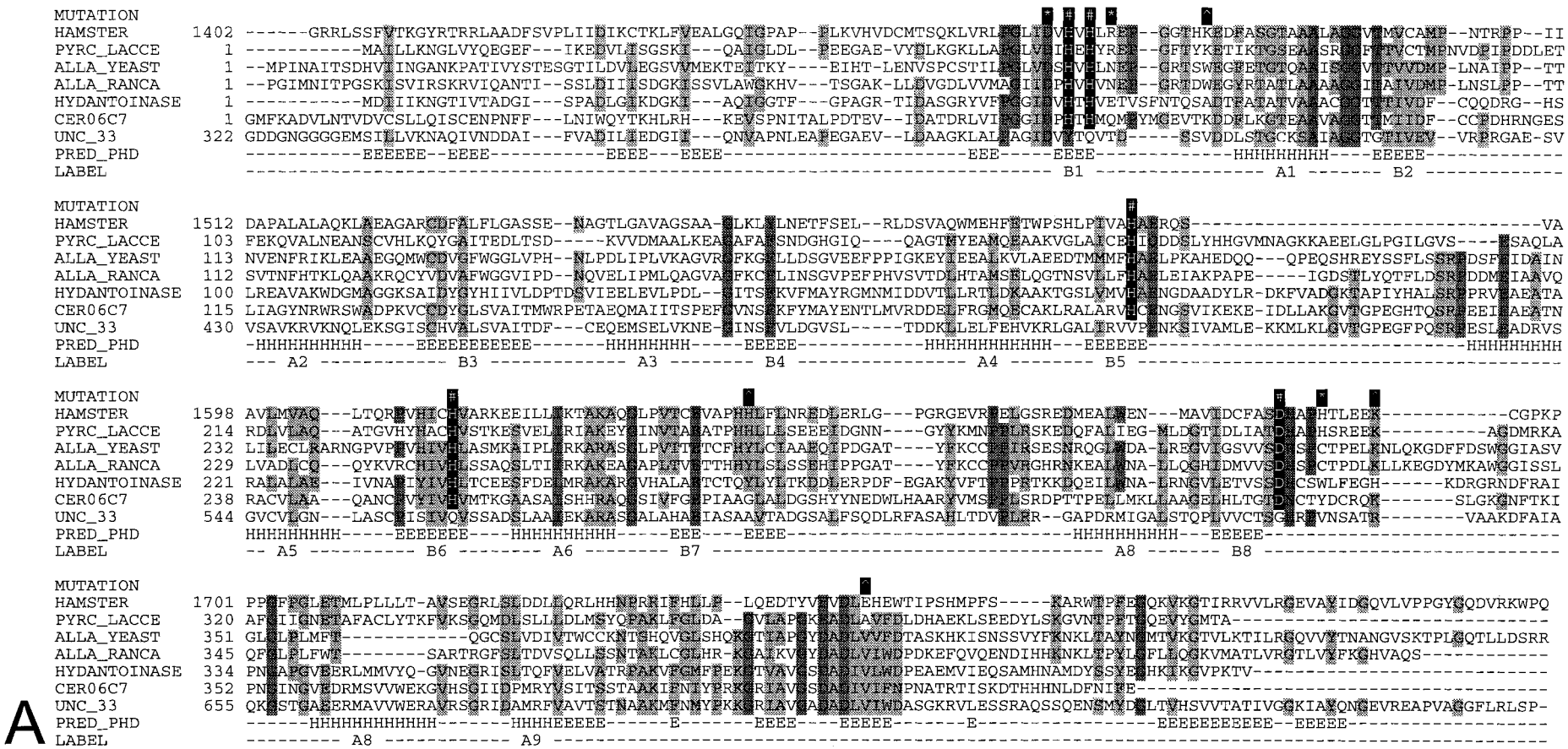
Searching protein sequence databases revealed a number of partial matches to the functionally required His-Asp signature pattern or showed grossly nonuniform sequence similarity over the predicted $(\beta\alpha)_8$ barrel structural unit relative to sequence relatives. Such conflicts were resolved at 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 frameshifts, for example, in the nucleotide sequences of the D-hydantoinase gene from *Pseudomonas putida* (Genpept acc. no. L24157; three frameshifts), 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 dihydroorotases, allantoinases, and hydantoinases.[19] When originally sequenced, aryldialkylphosphatase, 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

Fig. 1. Detection of remote homologues by conservation of three-dimensional structure. **A:** The three-dimensional structures of urease (2kauC[3]), phosphotriesterase (1pta[4]), and adenosine deaminase (1fkx[5]) aligned by using the Dali program.[6] Uppercase letters denote regions that are structurally equivalent with the topmost structure. No sequence information is used for alignment. The set of structurally equivalent residues (in pairwise comparison) maximizes the similarity of intramolecular Cα–Cα distances. Bold conserved residues map to the active site. The binuclear metal centres of urease and phosphotriesterase use a carbamoylated lysine, which is replaced by an aspartic acid in the mononuclear metal centre of adenosine deaminase. Secondary structure is marked as *helix* and strand. Columns with hydrophobic character (A,G,P,I,L,V,M,Y,F,W,T,C) are highlighted. **B:** Ribbon diagram[40] of the crystal structure of urease. Successive βαβα units of the barrel are red, yellow, green, and cyan. A particular structural feature of the common fold is that the bottom of the $(\beta\alpha)_8$ barrel, i.e., the side that has the N termini of β strands, is capped by helix α_9 (blue), which contacts the beginning of strand β_8 via a strongly conserved asparagine residue. Two nickel atoms at the active site of urease are shown as black spheres and side chains are shown for the metal binding histidines (β_1, β_5, β_6) and carbamoylated lysine (β_4) and catalytic Asp (β_8). The small domain is unique to urease among the three known structures. The small domain makes intimate contacts with the catalytic domain and is composed of segments that are both N-terminal (purple/white) and C-terminal (blue/white) to the $(\beta\alpha)_8$ barrel.


MUTATION
HAMSTER
PYRC_LACCE
ALLA_YEAST
ALLA_RANCA
HYDANTOINASE
CER06C7
UNC_33
PRED_PHD
LABEL
B1
A1
B2
B3
A3
B4
A4
B5
A2
B6
A6
B7
A8
B8
A5
A9

A

Fig. 2.

Fig. 2. From family alignment to fold recognition: a worked example. One of the largest subfamilies of the superfamily is formed by hydantoinases/dihydropyriminidases, dihydroorotase, allantoinase, and animal developmental proteins. These proteins are related to each other by clear sequence similarity, yet they are sufficiently diverse to bring out distinct conserved blocks. Moreover, functional residues have been characterized by direct experiment. **A:** Representative sequences are chosen from different subfamilies: hamster dihydroorotase domain from the multifunctional CAD enzyme, *Lactobacillus* dihydroorotase, allantoinase from yeast and bullfrog, hydantoinase from *Agrobacterium radiobacter*, and two nematode proteins, a dihydropyriminidase homologue and axonal guidance protein Unc-33. Conserved blocks are shaded according to average similarity by the Blosum62 matrix. Site-directed mutagenesis of conserved residues in hamster dihydroorotase has revealed which are responsible for zinc binding (#), are otherwise essential for activity (*) or are involved in substrate binding (/\).[27–28] Note systematic absence of all zinc binding residues in Unc-33. Secondary structure prediction by neural network (PRED-PHD[15]) indicates an alternating α/β fold (H, helix; E, strand). **B:** The metal binding residues must cluster together in the folded polypeptide. The derived prediction of the three-dimensional structure of dihydroorotase is based on the remarkable correlation between the location of the functional histidines and aspartic acid at the C terminus of predicted strands labeled B1, B5, B6, and B8 in dihydroorotase and the architectural arrangement of the active site of the $(\beta\alpha)_8$ barrel proteins urease, phosphotriesterase, and adenosine deaminase (cylinders, helix; arrows, strand; phosphotriesterase was crystallized with cadmium ions, although the natural protein binds two zinc atoms per molecule[4]).

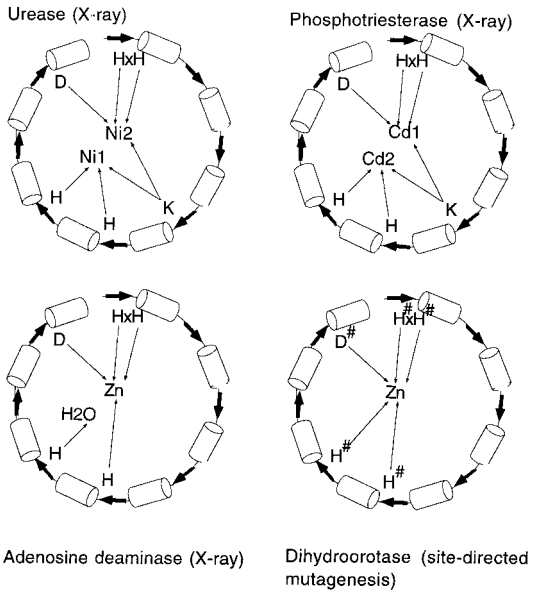


Figure 2. (Continued.)

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 I 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 $(\alpha\beta)_8$ 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 $(\alpha\beta)_8$ 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 segment GKLADLVVWS of urease (first blue β strand in Fig. 1b). Noting that the dihydroorotase from *Methanococcus* belongs to the group that has both the catalytic and small domain suggests that an ancestral form already was a two-domain entity and that along some evolutionary paths the small domain has been lost.

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, aryldialkylphosphatase, *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 atrazine[23] appears surprising (see Table I). 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.

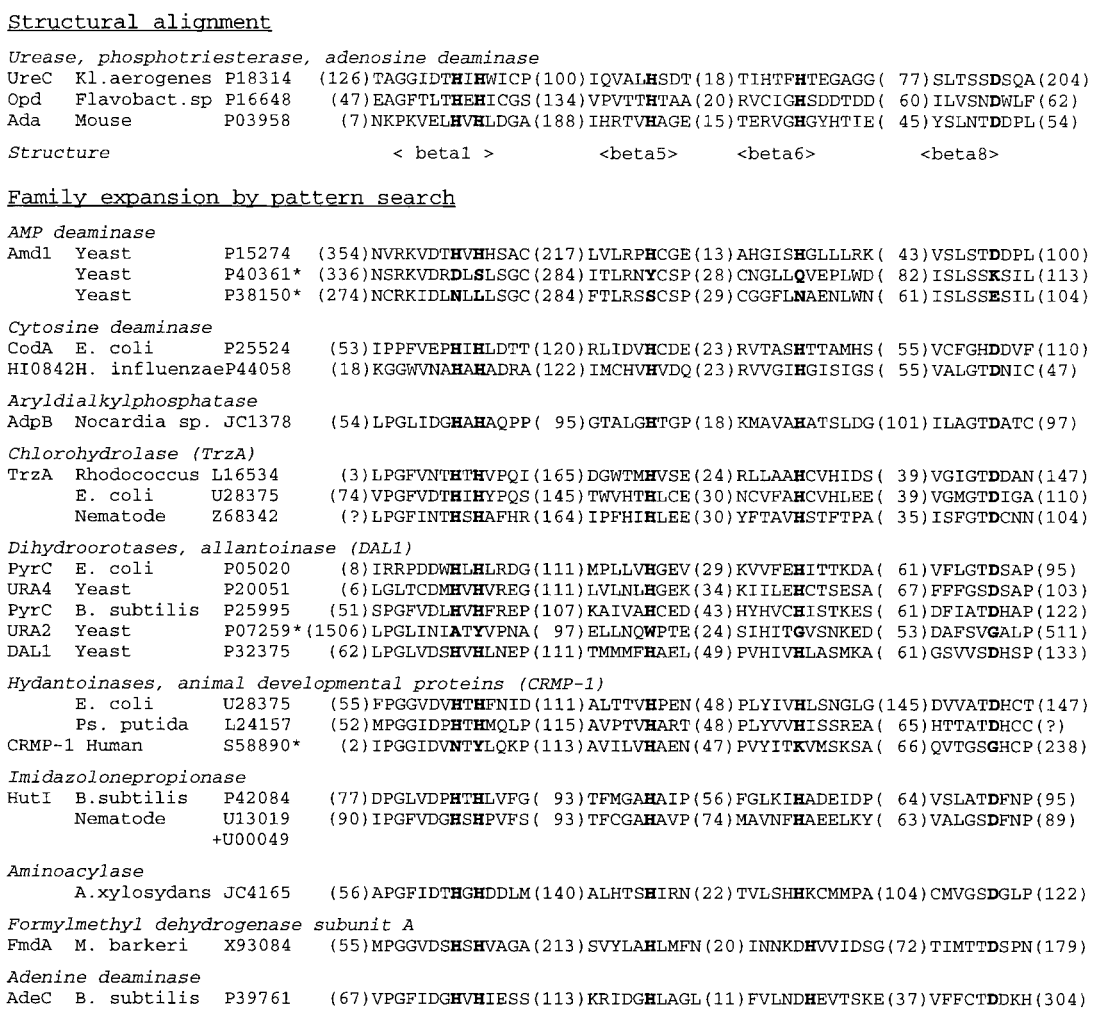


Fig. 3. Sequence conservation across the superfamily. Alignment of a representative subset of sequences (identified by gene name [*UreC*], species [*Klebsiella aerogenes*], and database accession number [P18314]; examples in square brackets) covering all member families. The conserved signature pattern (bold) consists of the four histidines and aspartic acid that, although dispersed in the linear sequence, come together (cf. Fig. 2**B**) in three dimensions in the folded structure (cf. Fig. 1**B**). Apparently nonfunctional proteins are indicated by an asterisk. Numbers in parentheses are the length of the intervening sequence.

Three closely related aminoacylases (*N*-acyl-D-glutamate amidohydrolase, D-aminoacylase, *N*-acyl-D-aspartate amidohydrolase) from *Alcaligenes xylosudans*[24] 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 truncted or lost during evolution. For example, the pathway of uric acid degradation (see Fig. 5) has been truncated through the successive loss of allantoicase, 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*. Dihydroorotase is absent, and, in fact, this organism lacks all genes encoding the first three steps of pyrimidine biosynthesis pathway.[25] *E. coli* has more functions in common with *Methanococcus* than with *Haemophilus*, although the latter is a closer sister species phylogenetically.

Evolution of Metal Centers

Members of the superfamily employ a fascinating variety of divalent metal ligands for catalysis. Adenosine deaminase binds a single zinc ion in the active site. Phosphotriesterase contains two zinc ions in the binuclear metal center where a carbamoylated ly-

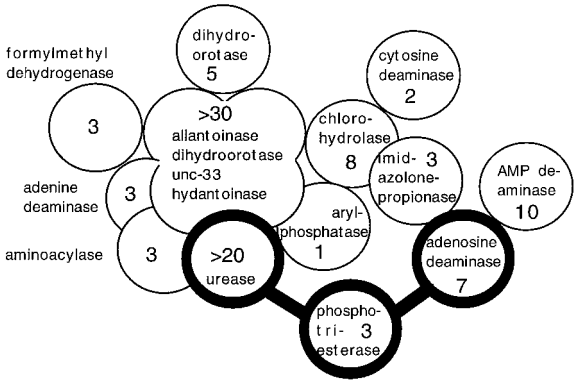


Fig. 4. Detection of remote homologues by walking in sequence space. This conceptual map of sequence space in the region of the 13 related families represents each protein family by a circle labeled with the number of known sequences in the family. Three families of known three-dimensional structure (thick circles, linked by 3D similarity) were used as starting points for walking in sequence space. The walk consists of a series of sequence database searches, starting at the center of a family; a peripheral member is used as the starting point for the next database search. Three regimes of sequence similarity across family boundaries were defined as follows: merged circles, close relatives, Fasta[7] alignment score $>$ 200; intersecting circles, remote relatives, Fasta *E* value (expected number of hits in database) less than 0.01; circles that just touch, twilight relationships, Fasta *E* value less than 1.0 (not all twilight relationships between families are represented in this graph). All families shown share the signature pattern from Figure 1a. Shaded families contain members in which sequence similarities extend from the catalytic domain to the small domain of urease. The sequence relationships suggest that all families derive from a common ancestor, sharing three-dimensional fold and principle of metal-assisted catalysis.

sine (located on β_4) acts as bridging ligand.[4] The same constellation as in phosphotriesterase occurs in the metal center of urease, except that nickel substitutes for zinc ions.[3] This use of a carbamoylated lysine appears unique to urease and phosphotriesterase, since the other member families have no invariant lysines mapping to the same region. Dihydroorotase resembles adenosine deaminase in that it binds one zinc atom per molecule.[26] Importantly, the metal binding residues have been determined experimentally for dihydroorotase,[27–28] and they coincide exactly with the signature pattern derived from the three known structures (Fig. 2). Aminoacylhydrolase contains two zinc atoms per molecule,[29] cytosine deaminase accepts several metal ions, giving the highest turnover with Fe^{2+},[30] and hydantoinase[31] and cytosine deaminase are also experimentally known to be metalloenzymes. For the remaining sequences that match the signature pattern, our results imply testable predictions of metal requirement and active site residues. (Note that the formylmethanofuran dehydrogenases from methanogenic archaebacteria occurs as tungsten- and molybdenum-dependent isoenzymes, which also contain iron-sulfur clusters, however, in other subunits than the A subunit discussed here.[32])

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[33]; serine proteases/haptoglobin, a plasma protein that binds but does not cleave hemoglobin[34]; and repeated recruitments of enzymes as structural proteins in the eye lens.[35] There is some information about the function of the noncatalytic members of the superfamily. A defective dihydroorotaselike 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.[36] In *Pseudomonas putida*, inactive dihydroorotaselike subunits are required for the correct assembly of active aspartate transcarbamoylase subunits into the dodecameric (6 + 6) holoenzyme.[37] 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[13]; 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 dihydrouracil, 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.

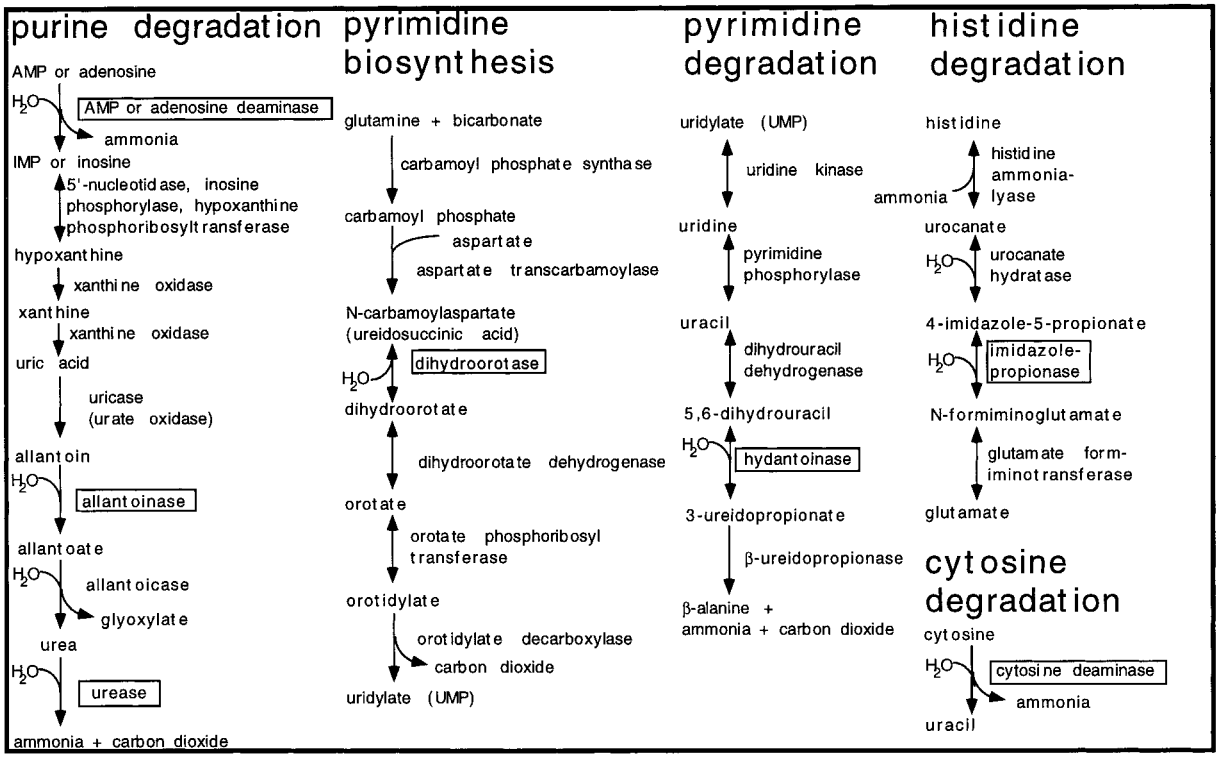


Fig. 5. Members of the superfamily in the context of metabolic pathways. The superfamily illustrates the evolutionary principle of reusing the same chemical principle in different steps of a set of related pathways. Boxes indicate superfamily members. Dihydroorotase, which catalyzes the third step in 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.

The sequence signature (mapping to β_1, β_5, β_6, and β_8) binds together both ends of the $(\beta\alpha)_8$ structural motif; recognition of homology was based on identifying invariantly conserved functional residues in a structural context of alternating α helices and β 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 leads 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 families. As a result, we find two examples for which family analysis refines functional assignments made in recent large-scale automated sequence analyses.[38] 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[39] 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, imidazolonepropionase, aryldialkylphosphatase, *s*-triazine hydrolase, aminoacylase, subunit A of formylmethanofuran dehydrogenase, and proteins involved in guiding animal neuronal development. As experimental struc-

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

1. Holm, L., Sander, C. Searching protein structure databases has come of age. Proteins 19:165–173, 1994.
2. Holm, L., Sander, C. DNA polymerase beta belongs to an ancient nucleotidyltransferase superfamily. Trends Biochem. Sci. 20:345–347, 1995.
3. Jabri, E., Carr, M.B., Hausinger, R.P., Karplus, P.A. The crystal structure of urease from *Klebsiella aerogenes*. Science 268:998–1004, 1995.
4. Benning, M.M., Kuo, J.M., Rauschel, F.M., Holden, H.M. Three-dimensional structure of the binuclear center of phosphotriesterase. Biochemistry 34:7973–7978, 1995.
5. Wilson, D.K., Quiocho, F.A. A pre-transition-state mimiv of an enzyme: X-ray structure of adenosine deaminase with bound 1-deazaadenosine and zinc-activated water. Biochemistry 32:1689–1694, 1993.
6. Holm, L., Sander, C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233:123–138, 1993.
7. Pearson, W.R. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. 183:63–98, 1990.
8. Tatusov, R., Altschul, S.F., Koonin, E.V. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. Proc. Natl. Acad. Sci. USA 91:12091–12095, 1994.
9. Benner, S.E., Hubbard, T., Murzin, A., Chothia, C. Gene duplications in *H. influenzae*. Nature 378:140, 1995.
10. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. Basic local alignment search tool. J. Mol. Biol. 215:403–410, 1990.
11. Feng, D.F., Boolittle, R.F. Progressive alignment and phylogenetic tree construction of protein sequences. Methods Enzymol. 183:357–387, 1990.
12. Thompson, J.D., Higgins, D.G., Gibson, T.J. CLUSTALW: Improving the sensitivity of the progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680, 1994.
13. Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
14. King, R.D, Sternberg, M.J.E. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Prot. Sci. 5:2298–2310, 1996.
15. Rost, B., Sander, C., Schneider, R. PHD: An automatic mail server for protein secondary structure prediction. CABIOS 10:53–60, 1994.
16. Holm, L., Sander, C. Evaluation of protein models by atomic solvation preference. J. Mol. Biol. 225:93–105, 1992.
17. Sayle, R., Milner-White, J. RASMOL: Biomolecular graphics for all. Trends Biochem. Sci. 20:374–376, 1995.
18. Chang, Z., Nygård, P., Chinault, A.C., Kellems, R.E. Deduced amino acid sequence of *Escherichia coli* adenosine deaminase reveals evolutionarily conserved amino acid residues: Implications for catalytic function. Biochemistry 30:2273–2280, 1991.
19. LaPointe, G., Viau, S., Leblanc, D., Roberts, N., Morin, A. Cloning, sequencing, and expression in *Escherichia coli* of the D-hydantoinase gene from *Pseudomonas putida* and distribution of homologous genes in other microorganisms. Appl. Environ. Microbiol. 60:888–895, 1993.
20. Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., Kerlavage, A.R., Dougherty, B.A., Tomb, J.F., Adams, M.D., Reich, C.I., Overbeek, R., Kirkness, E.F., Weinstock, K.G., Merrick, J.M., Glodek, A., Scott, J.L., Geoghagen, N.S.M., Weidman, J.F., Fuhrmann, J.L., Nguyen, D., Utterback, T.R., Kelley, J.M., Peterson, J.D., Sadow, P.W., Hanna, M.C., Cotton, M.D., Roberts, K.M., Hurst, M.A., Kaine, B.P., Borodovsky, M., Kenk, H.P., Fraser, C.M., Smith, H.O., Woese, C.R., Venter, J.C. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058–1073, 1996.
21. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F, Dougherty, B.A., Merrick, J.M. et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512, 1995.
22. Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, J.L., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Sandek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.F., Dougherty, B.A., Bott, K.F., Hu, P.C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchinson III, C.A., Venter, J.C. The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403, 1995.
23. Shao, Z.Q., Seffens, W., Mulbry, W., Behki, R.M. Cloning and expression of the *s*-triazine hydrolase gene (trzA) from *Rhodococcus corallinus* and development of *Rhodococcus* recombinant strains capable of dealkylating and dechlorinating the herbicide atrazine. J. Bacteriol. 177:5748–5755, 1995.
24. Wakayama, M., Ashika, T., Miyamoto, Y., Yoshikawa, T., Sonoda, Y., Sakai, K., Moriguchi, M. Primary structure of *N*-acyl-D-glutamate amidohydrolase from *Alcaligense xylosudans* subsp. *xylosudans* A-6. J. Biochem. 118:204–209, 1995.
25. Karp, P.D., Ouzounis, C., Paley, S. HinCyc: A knowledge base of the complete genome and metabolic pathways of *H. influenzae*. In "Intelligent Systems for Molecular Biology." St. Louis, MO: AAAI Press, 1996.
26. Kelly, R.E., Mally, M.I., Evans, D.R. The dihydroorotase domain of the multifunctional protein CAD: Subunit structure, zinc content, and kinetics. J. Biol. Chem. 261:6073–6083, 1986.
27. Williams, N.K., Manthey, M.K., Hambley, T.W., O'Donoghue, S.I., Keegan, M., Chapman, B.E., Christopherson, R.I. Catalysis by hamster dihydroorotase: Zinc binding, site-directed mutagenesis, and interaction with inhibitors. Biochemistry 34:11344–11352, 1995.
28. Zimmermann, B.H., Kemling, N.M., Evans, D.R. Function of conserved histidine residues in mammalian dihydroorotase. Biochemistry 34:7038–7046, 1995.
29. Yang, Y.B., Hsiao, K.M., Li, H., Yano, H., Tsugita, A., Tsai, Y.C. Characterization of D-aminoacylase from *Alcaligenes denitrificans* SA181. Biosci. Biotechnol. Biochem. 56:1392–1395, 1992.
30. Porter, D.J., Austin, E.A. Cytosine deaminase: The role of divalent metal ions in catalysis. J. Biol. Chem. 268:24005–24011, 1993.
31. Mukohara, Y., Ishikawa, T., Watabe, K., Nakamura, H. Biosci. Biotechnol. Biochem. 58:1621–1626, 1994.
32. Vorholt, J.A., Vaupel, M., Thauer, R.K. A polyferredoxin with eight [4Fe-4S] clusters as a subunit of molybdenum formylmethanofuran dehydrogenase from *Methanosarcina barkeri*. Eur. J. Biochem. 236:309–317, 1996.
33. Lewis, P.N., Scheraga, H.A. Prediction of structureal homol-

ogy between bovine-lactalbumin and hen egg white lysozyme. Arch. Biochem. Biophys. 144:584–588, 1971.
34. Kurosky, A., Barnett, D.R., Rasco, M.A., Lee, T.H., Bowman, B.H. Evidence of homology between the beta-cahin of human haptoglobin and the chymotrypsin family of serine proteases. Biochem. Genet. 11:279–293, 1974.
35. Wistow, G., Piatigorsky, J. Recruitment of enzymes as lens structural proteins. Science 236:1554–1556, 1987.
36. Souciet, J.L., Nagy, M., Le Gouar, M., Lacroute, F., Potier, S. Organization of the yeast URA2 gene: Identification of a defective dihydroorotase-like domain in the multifunctional carbamoylphosphate synthetase-aspartate transcarbamylase complex. Gene 79:59–70, 1989.
37. Schurr, M.J., Vickrey, J.F., Kumar, A.P., Campbell, A.L., Cunin, R., Benjamin, R.C., Shanley, M.S., O'Donovan, G.A. Aspartate transcarbamouylase genes of *Pseudomonas putida*: Requirement for an inactive dihydroorotase for assembly into the dodecameric holoenzyme. J. Bacteriol. 177:1751–1759, 1995.
38. Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C., Sander, C. GeneQuiz. In "Intelligent Systems for Molecular Biology." Menlo Park, CA: AAAI Press, 1994:348–353.
39. Casari, G., Andrade, M.A., Bork, P., Boyle, J., Daruvar, A. de, Ouzounis, C., Schneider, R., Tamames, J., Valencia, A., Sander, C. Challenging times for bioinformatics. Nature 376:647–648, 1995.
40. Kraulis, P. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24:946–950, 1991.