Structural genomics of enzymes involved in sterol/isoprenoid biosynthesis

Jeffrey B. Bonanno*†‡, Carme Edo*‡, Narayanan Eswar*, Ursula Pieper*, Michael J. Romanowski*, Valentin Ilyin*, Sue Ellen Gerchman§, Helen Kycia§, F. William Studier§, Andrej Sali*, and Stephen K. Burley*†¶

*Laboratories of Molecular Biophysics, †Howard Hughes Medical Institute, The Rockefeller University, 1230 York Avenue, New York, NY 10021; and §Biology Department, Brookhaven National Laboratory, Upton, NY 11973

Contributed by F. William Studier, September 4, 2001

X-ray structures of two enzymes in the sterol/isoprenoid biosynthesis pathway have been determined in a structural genomics pilot study. Mevalonate-5-diphosphate decarboxylase (MDD) is a single-domain α/β protein that catalyzes the last of three sequential ATP-dependent reactions which convert mevalonate to isopentenyl diphosphate. Isopentenyl disphosphate isomerase (IDI) is an α/β metalloenzyme that catalyzes interconversion of isopentenyl diphosphate and dimethylallyl diphosphate, which condense in the next step toward synthesis of sterols and a host of natural products. Homology modeling of related proteins and comparisons of the MDD and IDI structures with two other experimentally determined structures have shown that MDD is a member of the GHMP superfamily of small-molecule kinases and IDI is similar to the nudix hydrolases, which act on nucleotide diphosphate-containing substrates. Structural models were produced for 379 proteins, encompassing a substantial fraction of both protein superfamilies. All three enzymes responsible for synthesis of isopentenyl diphosphate from mevalonate (mevalonate kinase, phosphomevalonate kinase, and MDD) share the same fold, catalyze phosphorylation of chemically similar substrates (MDD decarboxylation involves phosphorylation of mevalonate diphosphate), and seem to have evolved from a common ancestor. These structures and the structural models derived from them provide a framework for interpreting biochemical function and evolutionary relationships.

High-throughput genome sequencing has dramatically expanded our knowledge of the proteins of the natural world. Next steps in understanding these macromolecules will involve studies of biochemical and biological function, depending critically on three-dimensional structural information. Nascent structural genomics efforts are aimed at developing experimental and computational pipelines for studying protein structure and integrating their results into the mainstream of biomedical research (1).

The New York Structural Genomics Research Consortium (http://www.nysgrc.org/, including Albert Einstein College of Medicine, Brookhaven National Laboratory, Mount Sinai School of Medicine, The Rockefeller University, and Weill Medical College of Cornell University) has been conducting a structural genomics pilot study under the auspices of the National Institutes of Health Protein Structure Initiative (http://www.nih.gov/nigms/funding/psi.html). Initial effort focused on proteins from *Saccharomyces cerevisiae*, an intensively studied eukaryotic organism with a fully sequenced genome that contains numerous human gene homologues. Target selection was aided by the results of automated comparative protein structure modeling by using the *S. cerevisiae* genome (2). We focused on proteins for which no structural information was available, with emphasis on members of large protein families that would permit homology modeling of as many related proteins as possible. Among the first 111 targets selected for x-ray crystallographic structure determination were two enzymes [Target 100, mevalonate-5-diphosphate decarboxylase (MDD), and Target 109, isopentenyl disphosphate isomerase (IDI)] from the sterol/isoprenoid biosynthesis pathway (ref. 3; Fig. 1).

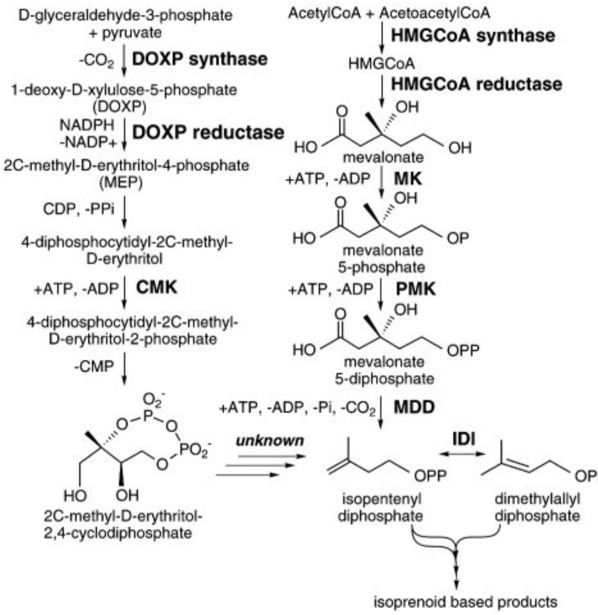

Fig. 1. Pathways for biosynthesis of isopentenyl diphosphate. Isopentenyl diphosphate, the central intermediate in sterol/isoprenoid biosynthesis, is produced by two independent pathways, which have different evolutionary distributions (39, 40).

Materials and Methods

Protein Expression/Purification. DNA fragments encoding MDD (*S. cerevisiae* gene *mvd1*, ORF YNR043W, 396 residues) and IDI (*S. cerevisiae* gene *idi1*, ORF YPL117C, 288 residues) were amplified from the genome of *S. cerevisiae* strain S288C by PCR and cloned into a T7 RNA polymerase-dependent *Escherichia coli* expression vector (4) that furnished a protease-cleavable N-terminal His$_6$ tag. Gel electrophoresis of cell extracts demonstrated that expressed MDD and IDI were both in the soluble fraction. Nickel-ion affinity chromatography and gel filtration

Abbreviations: GK, galactokinase (EC 2.7.1.6); HSK, homoserine kinase (EC 2.7.1.39); MK, mevalonate kinase (EC 2.7.1.36); PMK, phosphomevalonate kinase (EC 2.7.4.2); GHMP, kinase superfamily containing GK, HSK, MK, and PMK; IDI, isopentenyl diphosphate isomerase (EC 5.3.3.2); MDD, mevalonate-5-disphosphate decarboxylase (EC 4.1.1.33); rmsd, rms deviation.

Data deposition: The atomic coordinates have been deposited in the Protein Data Bank, www.rcsb.org [PDB ID codes 1I9A (IDI) and 1FI4 (MDD)].

‡J.B.B. and C.E. contributed equally to this work.

¶To whom reprint requests should be addressed. E-mail: burley@rockvax.rockefeller.edu.

www.pnas.org/cgi/doi/10.1073/pnas.181466998

Table 1. Statistics of the crystallographic analysis

Data set	Resolution, Å	Reflections measured/unique	Completeness (%) overall/outer shell	R_{sym} (%) overall/outer shell
MDD structure determination (9 Se sites) MAD figure of merit (18–2.28 Å) 0.58				
λ1 = (0.97923 Å)	18.0–2.25	299538/40942	99.1/90.3	3.0/22.7
λ2 = (0.97912 Å)	18.0–2.25	348900/40675	99.1/87.3	2.7/30.9
λ3 = (0.96863 Å)	18.0–2.25	349634/40447	99.1/82.9	3.1/39.6
IDI structure determination (6 Se sites) MAD figure of merit (20–2.50 Å) 0.30				
λ1 = (0.97940 Å)	20.0–2.5	134131/34202	97.9/91.4	3.8/30.8
λ2 = (0.97900 Å)	20.0–2.5	134573/34188	98.0/91.2	4.0/34.1
MDD refinement		*R* factor	*R* free	
All λ1 data	18.0–2.3	0.239	0.268	
rmsd	Bond lengths, 0.007 Å	Bond angles, 1.30°		
IDI refinement				
All λ1 data	20.0–2.5	0.229	0.279	
rmsd	Bond lengths, 0.006 Å	Bond angles, 1.39°		

$R_{sym} = \Sigma|I - \langle I\rangle|/\Sigma I$, where I, observed intensity; ⟨I⟩, average intensity obtained from multiple observations of symmetry related reflections. rms bond lengths and rms bond angles are the respective rmsds from ideal values. *R* free was calculated with 10% of data omitted from the refinement.

yielded highly purified proteins, as judged by mass spectrometry. After difficulties with *S. cerevisiae* IDI crystallizability, *E. coli* IDI (gene *idi*, 182 residues) was cloned from UT5600 genomic DNA, expressed as a glutathione *S*-transferase fusion, and purified to homogeneity by using published procedures (5).

Crystallizability Screening and Crystallization. Conformationally stable, monodisperse macromolecular preparations represent good crystallization candidates (6, 7). Conformational heterogeneity can be detected by using fluorescence and circular dichroism spectroscopy or proteolysis combined with MS (8). Polydispersity (nonspecific aggregation) can be detected by using dynamic light scattering (9). *S. cerevisiae* MDD behaved as a monodisperse dimer with an apparent molecular mass of 110 kDa (predicted mass for the dimer is 94 kDa) in aqueous solution. *S. cerevisiae* IDI was aggregated under similar conditions and was abandoned in favor of the more suitable homologue from *E. coli*, which behaves as a monodisperse monomer.

As expected from the results of crystallizability screening, hanging-drop/vapor-diffusion trials with *S. cerevisiae* MDD and *E. coli* IDI succeeded with minimal effort. MDD produced trapezoidal crystals in 100 mM Tris·HCl, pH 8.5/15% (wt/vol) polyethylene glycol (PEG) 4K/1 M NaCl/5% (vol/vol) ethylene glycol (space group, $P2_12_12$; unit cell: a = 78.9 Å, b = 126.5 Å, c = 47.3 Å; 1 molecule per asymmetric unit). IDI yielded square pyramidal crystals in 100 mM MES, pH 6.5/15% (wt/vol) PEG 8K/8% (wt/vol) PEG 1K/100 mM NaCl (space group, $P4_12_12$; unit cell: a = 72.4 Å, c = 204.4 Å; 2 molecules per asymmetric unit). Selenium-methionine proteins were produced and crystallized by using essentially identical procedures.

Data Collection and Structure Determination. X-ray diffraction data for MDD and IDI were recorded under standard cryogenic conditions at Beamline X25 at the National Synchrotron Light Source and Beamline F2 at the Cornell High Energy Synchrotron Source, respectively, and processed by using DENZO/SCALEPACK (10). Structures of MDD and IDI were obtained with the multiwavelength anomalous dispersion method (11) from diffraction data recorded at 3 (MDD) or 2 (IDI) x-ray wavelengths (Table 1). Selenium atomic positions (9 and 6 selenium sites, respectively) were located with SNB (12). MLPHARE phasing followed by density modification (13) yielded high-quality experimental phases for both systems. Electron density map interpretation (14) and refinement (15) produced structures of MDD (391 residues, 164 waters, *R* factor = 23.9%, *R* free = 26.8%, 2.3 Å resolution) and IDI (354 residues, 261 waters, *R* factor = 22.9%, *R* free = 27.9% 2.5 Å resolution).

Multiple Sequence Alignments. Proteins similar to *S. cerevisiae* MDD and *E. coli* IDI (*E* value $<10^{-4}$) were identified by using ψ-BLAST (16). Multiple sequence alignments were prepared by using CLUSTAL (17), and conservation was calculated with BLOSUM62 (18). Alignments of selected MDDs and IDIs are given in Figs. 2 and 3, respectively.

Homology Modeling. Automated comparative protein structure modeling (hereafter referred to as "homology modeling") with MODPIPE (19) was carried out by using each experimentally determined structure (hereafter referred to as "structures") as a modeling template. Candidates for modeling were obtained from the National Center for Biotechnology Information (ftp://ftp.ncbi.nlm.nih.gov/blast/db) by using an iterative ψ-BLAST search (*E* value <100 for at least 30 residues) with the sequence of the modeling template. Homology models were computed with the ψ-BLAST alignments by means of satisfaction of spatial restraints, as implemented in MODELLER (20), and assessed by computing a model score that uses a statistical energy function, sequence similarity with the modeling template, and a measure of structural compactness (2). Tests with known structures have shown that models with scores from 0.7 to 1.0 have the correct fold at a 95% confidence level (2). All models discussed below have scores >0.7.

Results

MDD Structural Overview. MDD is a single α/β domain (Fig. 4*A*, dimensions 75 Å × 42 Å × 38 Å) with a deep cleft. The order of secondary structural elements is β1-β2-β3-β4-β5-β6-αA-β7-αB-αC-αD-β8-β9-β10-αE-αF-αG-αH-β11-β12-αI-αJ-β13-β14, and the structure consists of three antiparallel β-sheets (β1-β6-β7-β14, β2-β5-β8-β9, β10-β11-β12-β13) and three sets of α-helices (αA-D, αE-G, αH-J). The crystallographic two-fold symmetry axis parallel to the *c* axis generates a symmetric dimer, which presumably corresponds to the oligomerization state detected by dynamic light scattering.

MDD alignments (Fig. 2) show conserved segments that surround a deep cleft (Fig. 4*A*). Within this surface concavity, Pro-113-Ala-122 (outlined in red in Fig. 2 and colored red in Fig. 4*A*) resembles P loops responsible for ATP binding in other

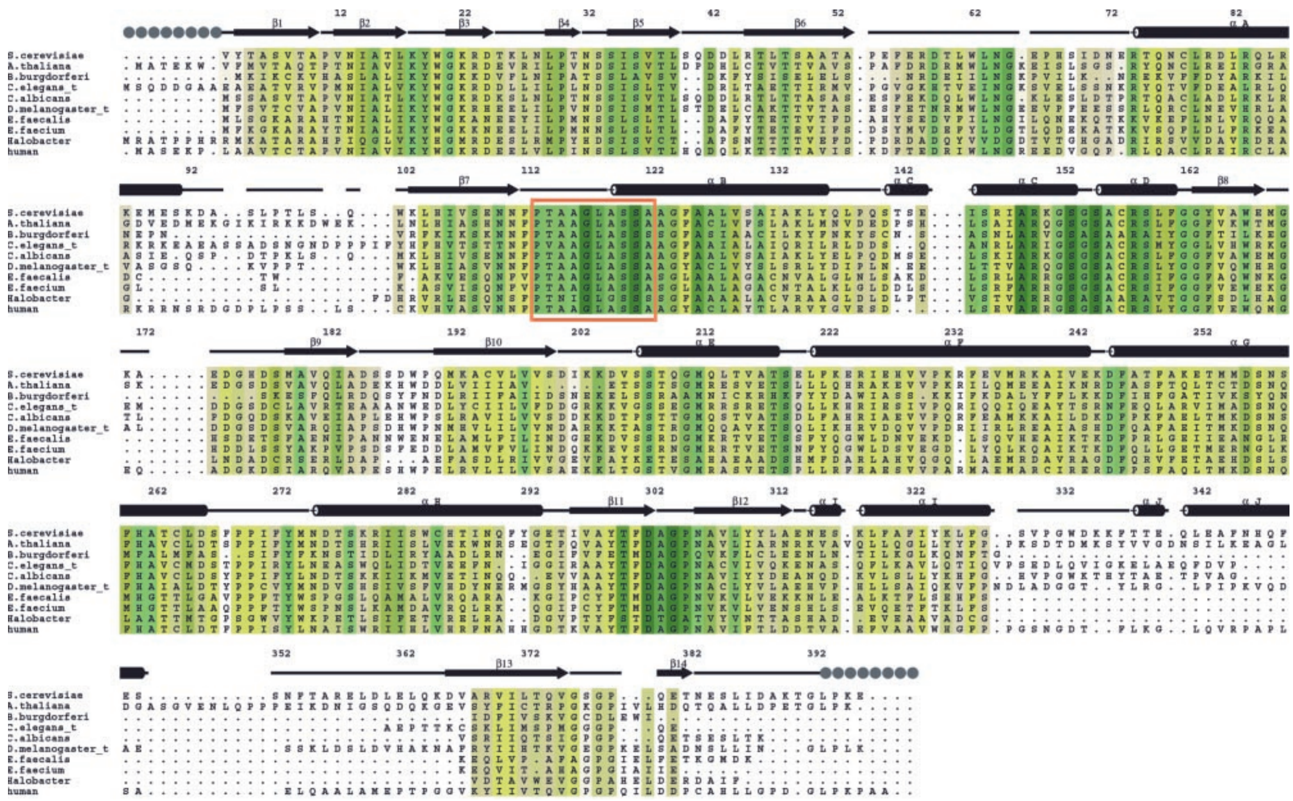

Fig. 2. MDD sequence alignment. Secondary structural elements of *S. cerevisiae* MDD are shown with cylinders (α-helices) and arrows (β-strands). Gray dots denote poorly resolved residues in the final electron density map. Color-coding denotes sequence conservation among MDDs (white $\rightarrow$ green ramp, 30 $\rightarrow$ 100% similarly). Red box denotes the putative ATP-binding P loop.

enzymes (21). GRASP calculations (22) revealed a surface patch with a positive electrostatic potential within the cleft (data not shown), which represents an excellent candidate for binding the anionic substrate, mevalonate-5-diphosphate.

Homology Modeling with MDD. Comparison of *S. cerevisiae* MDD with the Protein Data Bank (23) in November 2000 revealed no similar structures, as judged by DALI (24). Homology modeling with *S. cerevisiae* MDD gave models for the other MDDs plus various GHMP kinases (21), including the canonical galactokinases (GK), homoserine kinases (HSK), mevalonate kinases (MK; Fig. 1), and phosphomevalonate kinases (PMK; Fig. 1), plus diphosphocytidyl-2-*C*-methyl-D-erythritol kinases [CMK (EC 2.7.1.-), an enzyme in the 1-deoxy-D-xylulose-5-phosphate pathway to isopentenyl diphosphate; Fig. 1], and some hypothetical proteins. A kinase mechanism of action for MDD is not unreasonable. MDD phosphorylates the substrate C-3 hydroxyl group, followed by elimination of both the added phosphate and carboxylate groups to give isopentenyl diphosphate, ADP, and P_i. Our modeling results suggest that all three enzymes respon-

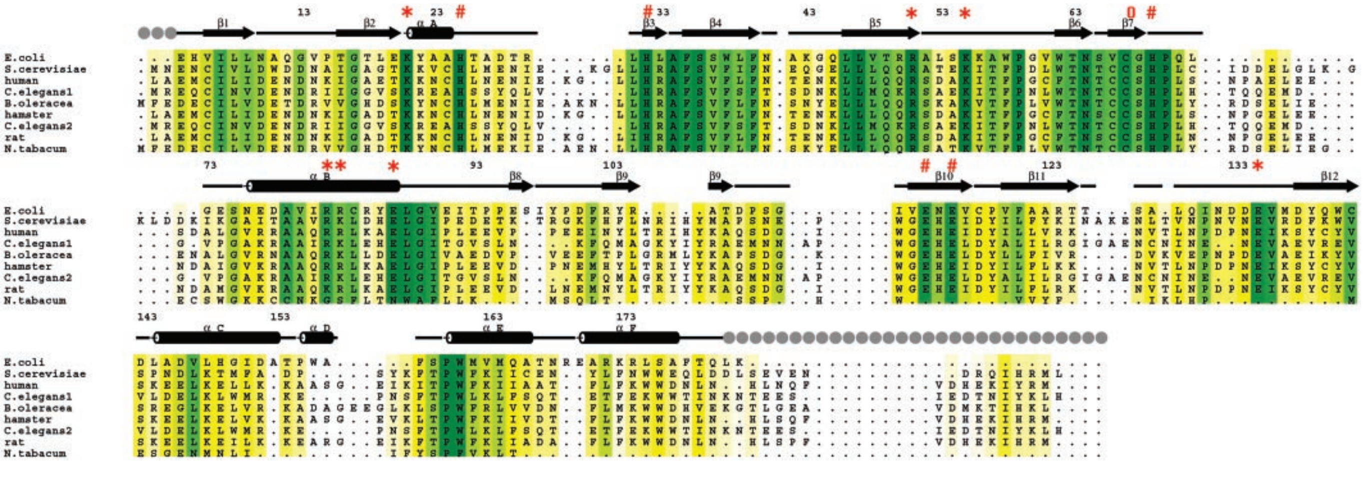

Fig. 3. IDI sequence alignment. Secondary structural elements, poorly resolved residues, and sequence conservation are denoted as in Fig. 2. #, metal-binding residues; *, conserved residues in the cleft; o = the active-site Cys. Some of the N- and C-terminal residues of IDI sequences other than *E. coli* have been excluded for clarity.

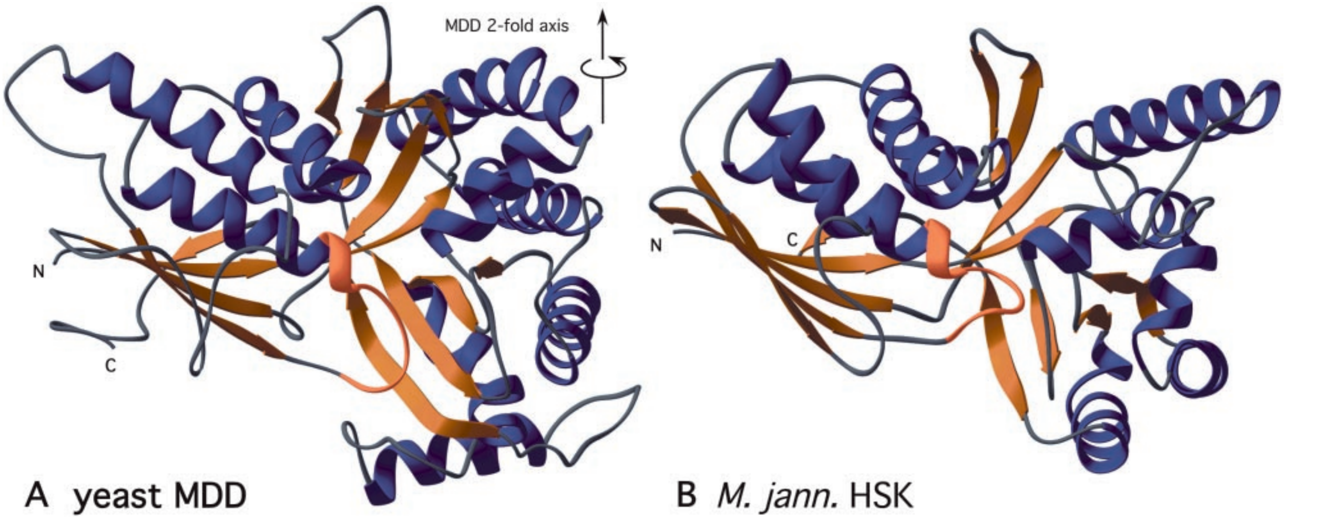

Fig. 4. *S. cerevisiae* MDD and *M. jannaschii* HSK. Ribbon drawings of MDD (*A*) and HSK (*B*) in the same orientation. The two-fold rotational symmetry axis that generates the MDD homodimer is indicated. Putative P loops are colored red.

sible for sequential conversion of mevalonate to isopentenyl diphosphate (MK, PMK, and MDD; Fig. 1) have the same fold, indicating that they arose from a common precursor and may represent an example of retrograde evolution (25). It is remarkable that diphosphocytidyl-2-*C*-methyl-D-erythritol kinase (CMK), which is part of the mevalonate-independent pathway for sterol/isoprenoid biosynthesis in plastids and bacteria, also seems to have evolved from the same ancestral small-molecule kinase.

Following our initial analyses, a *bona fide* GHMP kinase structure was reported (ref. 26; *Methanococcus jannaschii* HSK, Protein Data Bank ID code 1FWL). The only significant structural differences between *S. cerevisiae* MDD and *M. jannaschii* HSK are two insertions within MDD (corresponding to β3 plus β4 and αJ), providing direct experimental confirmation that MDD is indeed a GHMP kinase superfamily member [Fig. 4; 276 α-carbon pairs, rms deviation (rmsd) = 3.0 Å, 13% identity, Z score = 21.9; all rmsds and Z scores reported herein were obtained with DALI]. The P loop within HSK (colored red in Fig. 4*B*) was shown to bind ADP, albeit in a distinct orientation from that seen in other P loop-containing enzymes (26).

By using both structures (MDD and HSK), MODPIPE produced models (length > 200 residues) for 181 proteins (as of March 2001)—113 with both templates, 60 with HSK only, and 8 with MDD only. MDD and HSK each yielded models for the other. The HSK model derived from the MDD template vs. the structure of HSK gave an rmsd of 3.8 Å (241 α-carbon pairs, 13% identity, Z score = 15.0). The MDD model derived from the HSK template vs. the structure of MDD gave an rmsd of 3.4 Å (272 α-carbon pairs, 16% identity, Z score = 17.7). The correspondence between our computed models and the structures of MDD and HSK provides further evidence of the reliability of homology modeling with MODPIPE. In cases where models for the same sequence were obtained from both structural templates, the average rmsd between alternative models was 3.4 Å with <Z score> = 16.2 ± 3.0.

Models were produced for members of all subgroups of the GHMP kinase superfamily, including 22 MDDs, 31 GKs, 33 HSKs, 25 MKs, 9 PMKs, 25 diphosphocytidyl-2-*C*-methyl-D-erythritol kinases (CMK), 7 archael shikimate kinases (27), and 8 D-glycero-D-manno-heptose 7-phosphate kinases (28). The remaining 21 structural models fall into 4 sequence-similarity groups (one of which contains 11 members), suggesting that additional enzyme activities are encompassed within the GHMP kinase superfamily. Models covered at least 80% of the full length for 172 of 181 (95%) proteins modeled, suggesting that the GHMP kinase superfamily encompasses a rather narrow range of variations on the basic fold. Modeling summary statistics are given in Tables 2–5, which are published as supporting information on the PNAS web site, www.pnas.org, and atomic coordinate files for each of the models can be downloaded from MODBASE (http://guitar.rockefeller.edu/modbase/; select 1FI4 and 1FWL datasets on the advanced search page).

IDI Structural Overview. IDI is a single α/β domain (Fig. 5*A*, dimensions 28 Å × 40 Å × 27 Å) with a divalent metal ion bound within a deep cleft. The order of secondary structural elements is β1-β2-αA-β3-β4-β5-β6-β7-αB-β8-β9-β10-β11-β12-αC-αD-αE-αF, and the structure consists of two mixed polarity (β2-β1-β3-β10-β9, β7-β4-β11-β8) and one anti-parallel (β12-β5-β6) β-sheets. The second and third β-sheets are backed by α-helices.

Conserved metal-binding residues (His-25, His-32, His-69, Glu-114, and Glu-116; Fig. 3) form a distorted square pyramidal coordination with a water molecule near the open coordination site ($Mn^{2+} - H_2O$ = 5.5 Å; Fig. 5*A*). The metal ion was treated as Mn^{2+} for the purpose of crystallographic refinement, but it may be Mg^{2+} *in vivo* (29). Conserved polar residues (Lys-21, Arg-51, Lys-55, Arg-82, Arg-83, Glu-87, and Glu-135; Fig. 3) line the remainder of the cleft (Fig. 5*A*), and we presume that some of the basic side chains contribute to binding of the anionic substrates isomerized by IDI. Experimental support for our assignment of the IDI active site is provided by the results of inhibition studies, which showed that nearby Cys-67 (Fig. 5*A*) is essential for catalysis (29).

IDI Resembles MutT. Comparison of our IDI structure with the contents of the Protein Data Bank (PDB) in January 2001 revealed one relative, *E. coli* MutT (PDB ID code 1TUM; 108 α-carbon pairs, rmsd = 3.0 Å, 13% identity, Z score = 6.7). The solution NMR structure of MutT (Fig. 5*B*) resembles the portion of IDI containing the β3-β10-β9 and β7-β4-β11-β8 sheets and two of the flanking α-helices (αB and αF).

MutT is thought to guard against DNA incorporation of a mutagenic product of oxidative damage by hydrolyzing 7,8-dihydro-8-oxoguanine-triphosphate to the monophosphate (30). It is a member of the nudix superfamily of enzymes that act on various substrates, most of which contain a nucleotide diphosphate (31). A conserved motif characteristic of the nudix family

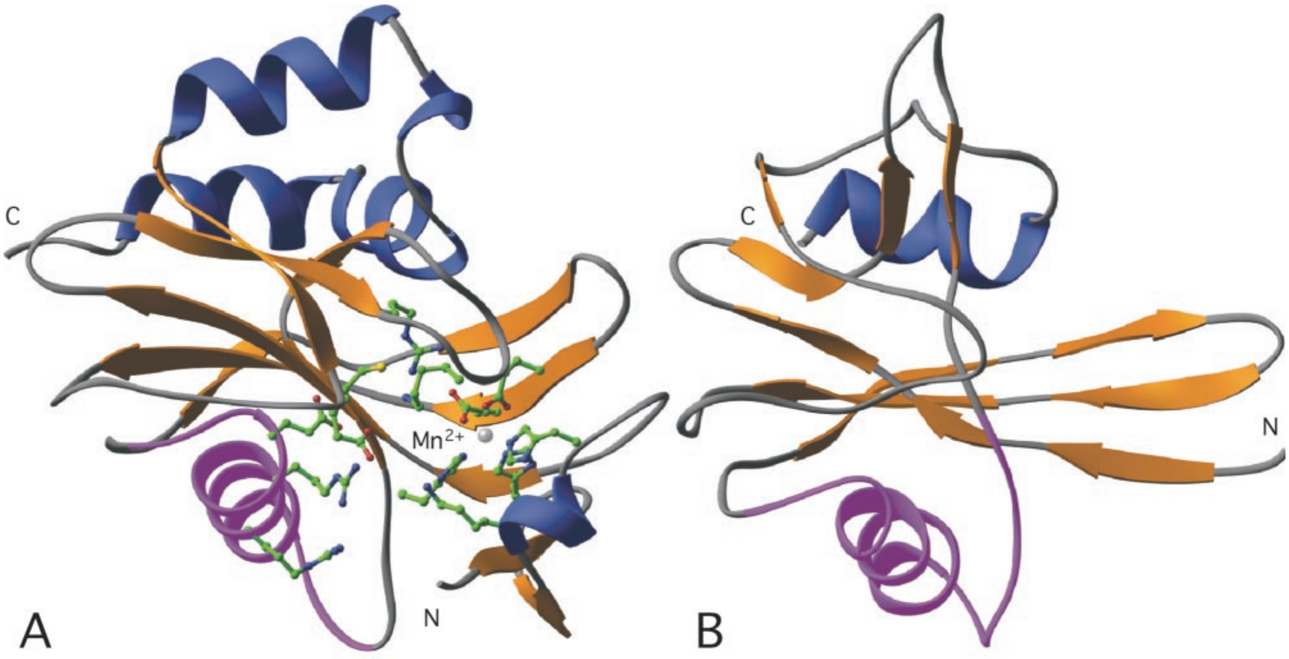

Fig. 5. *E. coli* IDI and MutT. Ribbon drawings IDI (*A*) and MutT (*B*) in the same orientation. Ball-and-stick representations of the divalent metal ion and putative active-site residues are given for IDI (atom type code: C, green; S, yellow; N, blue; O, red; Mn^{2+}, gray). The positions of the equivalent conserved motifs of the two proteins are colored magenta.

is Gly-X_5-Glu-X_7-Arg-Glu-ϕ-X-Glu-Glu-X_2-ϕ, where X is any residue and ϕ denotes a large hydrophobic side chain (31). The corresponding segment within the IDIs contains a similar but different conserved motif Gly-X_3-Ala-X_2-Arg-Arg/Lys-ϕ-X_2-Glu-Leu-Gly-ϕ (residues 75–90; Fig. 3), where Arg-Arg/Lys in IDI corresponds to Arg-Glu in MutT, and X-Glu in IDI corresponds to Glu-Glu in MutT. The positions of these conserved motifs in the IDI and MutT structures are shown in Fig. 5 (colored magenta).

Homology Modeling with IDI and MutT. By using the structures of IDI and MutT as modeling templates, MODPIPE produced models (as of April 2001) for 202 nonredundant protein sequences (length > 100 residues)—62 with both templates, 103 with IDI only, and 37 with MutT only. The MutT model derived from IDI vs. the solution NMR structure of MutT gave an rmsd of 3.0 Å (69 α-carbon pairs, 18% identity, Z score = 2.7). No model for IDI (score > 0.7) was produced with the structure of MutT as a template. In cases where models for the same sequence were obtained from both structural templates (IDI and MutT), the average rmsd between alternative models was 3.6 Å with <Z score> = 5.6 ± 1.0. This <Z score> is considerably lower than the corresponding value obtained by comparing MDD- and HSK-derived alternative models, and presumably reflects the greater structural difference between IDI and MutT. Nonetheless, our structural comparison and modeling results do confirm the recently proposed similarity of the IDIs and the nudix proteins (32).

The 202 nudix/IDI superfamily members for which MODPIPE produced models were grouped on the basis of their amino acid sequences: 156 proteins fall into one of 15 groups in which at least 1 member was annotated with a discrete enzymatic activity (including 37 IDIs), 42 proteins were singletons or fall into small groups with no attributed enzymatic activity, and 4 were larger proteins thought to possess more than 1 biochemical function. In contrast to the GHMP kinase superfamily, the nudix/IDI superfamily fold seems to be rather frequently combined with other structural features or domains. Models based on either IDI or MutT covered at least 80% of the full length for only 109 of 202 (54%) proteins modeled, and less than 50% of the full length for 35 proteins (17%). Modeling summary statistics are given in Tables 2–5, and atomic coordinate files for each of the models can be downloaded from MODBASE (http://guitar.rockefeller.edu/modbase/; select 1I9A and 1MUT datasets on the advanced search page).

Discussion

One of the goals of the New York Structural Genomics Research Consortium is to generate experimentally determined structures with sufficient diversity to support homology modeling of large numbers of protein sequences. Recent analyses by Vitkup *et al.* (33) have suggested that determination of as few as 16,000 selected structures could enable production of "accurate" homology models for 90% of all proteins found in nature (see below).

Homology modeling can be distinguished from all other methods for analyzing relationships among protein sequences because it yields atomic coordinates suitable for direct comparisons with x-ray and solution NMR structures. Acceptable models can be divided into three accuracy classes, characterized with blind tests by using known structures (34). Models based on >50% sequence identity with the template are comparable in accuracy to 3 Å resolution x-ray structures or medium-resolution solution NMR structures (10 long-range restraints per residue). Models obtained with less similar templates (30–50% identity) typically have >85% of their α-carbons within 3.5 Å of the correct position. When sequence identities fall below 30%, models with acceptable model scores (>0.7) may contain significant errors arising from ambiguities in the sequence alignment of the modeling candidate with the template, but can nevertheless be used for protein fold identification.

Targets 100 (MDD) and 109 (IDI) were selected for experimental structure determination to provide modeling templates for two enzyme families in the sterol/isoprenoid biosynthesis pathway and, indeed, we obtained models for 21 MDDs and 36 IDIs. Greater modeling coverage was, however, possible because both MDD and IDI proved to be members of distinct protein superfamilies encompassing a range of different enzyme activities and proteins of unknown biochemical function. Even wider coverage was afforded by the availability of a second modeling template from each of these protein sequence/structure families. Of the 379 sequences modeled with one or more of the MDD, HSK, IDI, and MutT templates, 3 (<1%) fall into the highest accuracy range (>50% identity), 53 (14%) in the medium accuracy range (30–50% identity), and 323 (85%) in the lowest accuracy range (<30% identity). Virtually all of the 56 proteins with medium or highest accuracy models (≥30% identity) are thought to have the same enzymatic function as the template (with two possible exceptions in the nudix/IDI superfamily). In clustering the members of these two superfamilies into groups of the most closely related proteins (by means of sequence comparisons), proteins of unknown function were often grouped with proteins of known enzymatic activity, allowing tentative assignment of biochemical function (Tables 2–5). Isolated cases of patently incorrect functional annotation were also found (Tables 2–5).

The modeling coverage of the GHMP kinase superfamily provided some insights into the problem of target selection for structural genomics. The structure of *S. cerevisiae* MDD produced medium or high accuracy models for 80% of the modeled sequences thought to have MDD activity, but coverage of the HSK family provided by the *M. jannaschii* HSK template was not as broad (only 41% of HSK models are medium or high accuracy). Current GHMP kinase superfamily members can be clustered into 19 discrete subfamilies with at least 4 members, encompassing the MDDs, 3 clusters of GKs, 4 clusters of HSKs, 3 clusters of MKs, the PMKs, the D-glycero-D-manno-heptose 7-phosphate kinases, 3 clusters of diphosphocytidyl-2-*C*-methyl-D-erythritol kinases (CMKs), the archael shikimate kinases, and 2 distinct clusters of hypothetical proteins. We estimate that 17 additional experimentally determined structures will be required to produce medium to high accuracy models for most members (≥80%) of each detectable GHMP kinase subfamily. The nudix/IDI superfamily, on the other hand, seems to have much greater structural diversity, and achieving equivalent coverage by medium to high accuracy models may present a greater logistical challenge.

Structures of GHMP kinase superfamily members and resulting homology models may guide experimentation aimed at defining enzymatic function, cofactor requirements, and mechanism(s) of action, which should allow us to understand better how this family of structurally similar yet functionally diverse enzymes evolved from a common ancestor. The GHMP kinase models may be of some medical relevance in understanding the structural bases of a number of diseases caused by single-nucleotide polymorphisms in coding regions. Impairment of human GK function by missense mutations within the enzyme leads to galactosemia and cataract formation in newborns (35), which can be reversed by restricting dietary galactose. Mutations in the human gene encoding MK lead to either hyperIgD and periodic fever syndrome or mevalonic aciduria, a severe and frequently fatal malady (36, 37). Finally, mutations in both MK and PMK have been implicated in the development of Zellweger syndrome, a genetic disorder characterized by defects in peroxisome biogenesis (38).

This article is dedicated to the memory of Konrad Bloch. We are most grateful to Drs. M. Becker, L. Berman, and D. Thiel, and the MacChess staff for synchrotron beamline access. We thank Drs. M. Goger and S. S. Ray for assistance with biophysical measurements; Ms. T. Niven for manuscript preparation; and Drs. M. S. Brown, J. L. Goldstein, D. Jeruzalmi, K. Kamada, G. A. Petsko, and members of the New York Structural Genomics Research Consortium for help and useful discussions. This work was supported by National Institutes of Health Grants P50 GM62529 (New York Structural Genomics Research Consortium) and R29 GM54762 (A.S.), The Rockefeller University (S.K.B. and A.S.), the Mathers Foundation and Merck Genome Research Institute (A.S.), and the Office of Biological and Environmental Research of the Department of Energy (F.W.S.). S.K.B. is an Investigator in the Howard Hughes Medical Institute. A.S. is an Irma T. Hirschl Career Scientist.

1. Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. & Swaminathan, S. (1999) *Nat. Genet.* **23,** 151–157.
2. Sanchez, R. & Sali, A. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 13597–13602.
3. Goldstein, J. L. & Brown, M. S. (1990) *Nature (London)* **343,** 425–430.
4. Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. (1990) *Methods Enzymol.* **185,** 60–89.
5. Roll-Mecak, A., Cao, C., Dever, T. E. & Burley, S. K. (2000) *Cell* **103,** 781–792.
6. D'Arcy, A. (1994) *Acta Crystallogr. D* **50,** 469–471.
7. Ferré-D'Amaré, A. R. & Burley, S. K. (1997) *Methods Enzymol.* **276,** 157–166.
8. Cohen, S. L., Ferre-D'Amare, A. R., Burley, S. K. & Chait, B. T. (1995) *Protein Sci.* **4,** 1088–1099.
9. Schmitz, K. S. (1990) *An Introduction to Dynamic Light Scattering by Macromolecules* (Academic, San Diego).
10. Otwinowski, Z. & Minor, W. (1997) *Methods Enzymol.* **276,** 307–326.
11. Hendrickson, W. (1991) *Science* **254,** 51–58.
12. Weeks, C. M. & Miller, R. (1999) *J. Appl. Crystallogr.* **32,** 120–124.
13. Dodson, E. J., Winn, M. & Ralph, A. (1997) *Methods Enzymol.* **277,** 620–633.
14. Jones, T. A. & Kjeldgaard, M. (1997) *Methods Enzymol.* **277,** 173–208.
15. Brünger, A. T., Adams, P. D., Clore, G. M., Gros, P., Grosse-Kuntsleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S. & Read, R. J. (1998) *Acta Crystallogr. D* **54,** 905–921.
16. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
17. Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1992) *Comput. Appl. Biosci.* **8,** 189–191.
18. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.
19. Sanchez, R., Pieper, U., Mirkovic, N., de Bakker, P. I., Wittenstein, E. & Sali, A. (2000) *Nucleic Acids Res.* **28,** 250–253.
20. Sali, A. & Blundell, T. L. (1993) *J. Mol. Biol.* **234,** 779–815.
21. Bork, P., Sander, C. & Valencia, A. (1993) *Protein Sci.* **2,** 31–40.
22. Gilson, M., Sharp, K. & Honig, B. (1988) *J. Comput. Chem.* **9,** 327–335.
23. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000) *Nucleic Acids Res.* **28,** 235–242.
24. Holm, L. & Sander, C. (1993) *J. Mol. Biol.* **233,** 123–138.
25. Horowitz, N. H. (1945) *Proc. Natl. Acad. Sci. USA* **31,** 153–157.
26. Zhou, T., Daugherty, M., Grishin, N. V., Osterman, A. L. & Zhang, H. (2000) *Structure (London)* **8,** 1247–1257.
27. Daugherty, M., Vonstein, V., Overbeek, R. & Osterman, A. (2001) *J. Bacteriol.* **183,** 292–300.
28. Kneidinger, B., Graninger, M., Puchberger, M., Kosma, P. & Messner, P. (2001) *J. Biol. Chem.* **276,** 20935–20944.
29. Hahn, F. M., Hurlburt, A. P. & Poulter, C. D. (1999) *J. Bacteriol.* **181,** 4499–4504.
30. Maki, H. & Sekiguchi, M. (1992) *Nature (London)* **355,** 273–275.
31. Bessman, M. J., Frick, D. N. & O'Handley, S. F. (1996) *J. Biol. Chem.* **271,** 25059–25062.
32. Smit, A. & Mushegian, A. (2000) *Genome Res.* **10,** 1468–1484.
33. Vitkup, D., Melamud, E., Moult, J. & Sander, C. (2001) *Nat. Struct. Biol.* **8,** 559–566.
34. Marti-Renom, M. A., Stuart, A., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000) *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291–325.
35. Novelli, G. & Reichardt, J. K. (2000) *Mol. Genet. Metab.* **71,** 62–65.
36. Houten, S. M., Koster, J., Romeijn, G.-J., Frenkel, J., Di Rocco, M., Caruso, U., Landrieu, P., Kelly, R. I., Kuis, W., Poll-The, B. T., *et al.* (2001) *Eur. J. Hum. Genet.* **9,** 253–259.
37. Cuisset, L., Drenth, J. P. H., Simon, A., Vincent, M. F., van der Velde Visser, S., van der Meer, J. W. M., Grateau, G. & Delpech, M. (2001) *Eur. J. Hum. Genet.* **9,** 260–266.
38. Wanders, R. J. & Romeijn, G. J. (1998) *Biochem. Biophys. Res. Commun.* **247,** 663–667.
39. Lange, B. M., Rujan, T., Martin, W. & Croteau, R. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 13172–13177. (First Published November 14, 2000; 10.1073/pnas.240454797)
40. Eisenreich, W., Rohdich, F. & Bacher, A. (2001) *Trends Plant Sci.* **6,** 78–84.