

# Sequence and Structure Classification of Kinases

Sara Cheek<sup>1</sup>, Hong Zhang<sup>1</sup> and Nick V. Grishin<sup>1,2\*</sup>

<sup>1</sup>Department of Biochemistry  
University of Texas  
Southwestern Medical Center  
5323 Harry Hines Blvd., Dallas  
TX 75390, USA

<sup>2</sup>Howard Hughes Medical  
Institute, University of Texas  
Southwestern Medical Center  
5323 Harry Hines Blvd., Dallas  
TX 75390, USA

Kinases are a ubiquitous group of enzymes that catalyze the phosphoryl transfer reaction from a phosphate donor (usually ATP) to a receptor substrate. Although all kinases catalyze essentially the same phosphoryl transfer reaction, they display remarkable diversity in their substrate specificity, structure, and the pathways in which they participate. In order to learn the relationship between structural fold and functional specificities in kinases, we have done a comprehensive survey of all available kinase sequences (>17,000) and classified them into 30 distinct families based on sequence similarities. Of these families, 19, covering nearly 98% of all sequences, fall into seven general structural folds for which three-dimensional structures are known. These fold groups include some of the most widespread protein folds, such as Rossmann fold, ferredoxin fold, ribonuclease H fold, and TIM  $\beta/\alpha$ -barrel. On the basis of this classification system, we examined the shared substrate binding and catalytic mechanisms as well as variations of these mechanisms in the same fold groups. Cases of convergent evolution of identical kinase activities occurring in different folds are discussed.

© 2002 Elsevier Science Ltd. All rights reserved

\*Corresponding author

**Keywords:** protein classification; fold; homology; enzymes; genomes

## Introduction

Kinases are a ubiquitous group of enzymes that participate in a variety of cellular pathways. By definition, the common name kinase is applied to enzymes that catalyze the transfer of the terminal phosphate group from ATP to an acceptor, which can be a small molecule, lipid, or protein substrate. The cellular and physiological roles of kinases are diverse. Many kinases participate in signal transduction pathways, in which these enzymes are essential components. Other kinases are involved centrally in the metabolism of carbohydrates, lipids, nucleotides, amino acid residues, vitamins, and cofactors. Some kinases have roles in various other processes, such as gene regulation, muscle contraction, and antibiotic resistance. Because of their universal roles in cellular processes, kinases are among the best studied enzymes at the structural, biochemical, and cellular level. Despite the fact that all kinases catalyze essentially the same phosphoryl transfer reaction, they display remark-

able diversity in their structures, substrate specificity, and in the number of pathways in which they participate.

In order to investigate the relationship between structural fold and functional specificities in kinases, we carried out a comprehensive analysis of available kinase structures and sequences. We have classified all available kinase sequences into structure/sequence families with evolutionary implications. We predicted a number of hypothetical proteins in the database that may possess kinase activity. Furthermore, we performed a large-scale structural prediction for kinases with unknown structures. Such classification of protein families can be quite helpful in understanding relationships between protein sequence, structure, and function. Functional information about proteins is often inferred following the identification of homologs based on the observation that evolutionary relatives commonly have similar or related functions. Possible biochemical roles for an uncharacterized protein are suggested by the function of its homologs. This information can then be utilized to aid the experimental determination of a protein's role at the biochemical and cellular level. Evolutionary relatives are often identified by sequence analysis. However, even sequence similarity searches with powerful profile-based tools such as PSI-BLAST<sup>1</sup> and HMMER<sup>2</sup> tend to

Abbreviations used: TRP, transient receptor potential; PEPCK, phosphoenolpyruvate carboxykinase; P-loop, phosphate binding loop; HPPK, 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase.

E-mail address of the corresponding author: [grishin@chop.swmed.edu](mailto:grishin@chop.swmed.edu)

**Table 1.** Classification of kinase activities by family and group

Group	Family and PFAM/COG members	EC no. and kinase activities	Representative PDB or gi
Group 1: protein S/T-Y kinase/lipid kinase/atypical protein kinase/SAICAR synthase/ATP-grasp, 9799 sequences	Protein S/T-Y kinase/lipid kinase/atypical protein kinase: COG0478, COG2112, PF00069, PF00454, PF01163, PF01633, 9600 sequences	2.7.1.32 Choline kinase	PDB: 1cdk
		<b>2.7.1.37 Protein kinase</b>	
		<b>2.7.1.38 Phosphorylase kinase</b>	
		2.7.1.39 Homoserine kinase	
		2.7.1.67 1-Phosphatidylinositol 4-kinase	
		2.7.1.70 Protamine kinase	
		2.7.1.72 Streptomycin 6-kinase	
		2.7.1.82 Ethanolamine kinase	
		2.7.1.87 Streptomycin 3'-kinase	
		<b>2.7.1.95 Kanamycin kinase</b>	
		2.7.1.100 5-Methylthioribose kinase	PDB: 1bo1
		2.7.1.103 Viomycin kinase	
		<b>2.7.1.112 Protein-tyrosine kinase</b>	
		2.7.1.116 [Isocitrate dehydrogenase (NADP <sup>+</sup> )] kinase	
		<b>2.7.1.117 [Myosin light-chain] kinase</b>	
		2.7.1.119 Hygromycin-B kinase	
		<b>2.7.1.123 Calcium/calmodulin-dependent protein kinase</b>	
		2.7.1.125 Rhodopsin kinase	
		2.7.1.126 [Beta-adrenergic-receptor] kinase	
		2.7.1.129 [Myosin heavy chain] kinase	
		2.7.1.135 [Tau protein] kinase	PDB: 1dik
		<b>2.7.1.137 1-Phosphatidylinositol 3-kinase</b>	
		2.7.1.141 [RNA-polymerase]-sub-unit kinase	
	SAICAR synthase: PF01504, 82 sequences	<b>2.7.1.68 1-Phosphatidylinositol-4-phosphate kinase</b>	
	ATP-grasp: PF01326, 117 sequences	2.7.1.133 1D-Myo-inositol-trisphosphate 6-kinase	
		2.7.1.139 1D-Myo-inositol-trisphosphate 5-kinase	
		<b>2.7.9.1 Pyruvate, phosphate dikinase</b>	
		2.7.9.2 Pyruvate, water dikinase	
Group 2: Rossmann-like, 3777 sequences	P-loop kinases: COG0645, COG1618, COG1663, COG1936, COG2019, PF00265, PF00406, PF00485, PF00625, PF00693, PF01121, PF01202, PF01583, PF01591, PF01712, PF02223, PF02224, PF02283, 1756 sequences	2.7.1.12 Gluconokinase	PDB: 1qf9
		<b>2.7.1.19 Phosphoribulokinase</b>	
		<b>2.7.1.21 Thymidine kinase</b>	
		<b>2.7.1.22 Ribosylnicotinamide kinase</b>	
		<b>2.7.1.24 Dephospho-CoA kinase</b>	
		<b>2.7.1.25 Adenylylsulfate kinase</b>	
		<b>2.7.1.33 Pantothenate kinase</b>	
		2.7.1.37 Protein kinase (bacterial)	
		2.7.1.48 Uridine kinase	
		<b>2.7.1.71 Shikimate kinase</b>	
		2.7.1.74 Deoxycytidine kinase	
		2.7.1.76 Deoxyadenosine kinase	
		2.7.1.78 Polynucleotide 5'-hydroxyl-kinase	
		<b>2.7.1.105 6-Phosphofructo-2-kinase</b>	
		<b>2.7.1.113 Deoxyguanosine kinase</b>	
		2.7.1.130 Tetraacyldisaccharide 4'-kinase	
		2.7.4.2 Phosphomevalonate kinase	
		<b>2.7.4.3 Adenylate kinase</b>	
		2.7.4.4 Nucleoside-phosphate kinase	
		<b>2.7.4.8 Guanylate kinase</b>	

(continued)

Table 1 Continued

Group	Family and PFAM/COG members	EC no. and kinase activities	Representative PDB or gi
	Phosphoenolpyruvate carboxykinase: COG1493, PF01293, PF00821, 212 sequences	<b>2.7.4.9 Thymidylate kinase</b> <b>2.7.4.10 Nucleoside-tri-phosphate-adenylate kinase</b> <b>2.7.4.13 (Deoxy)nucleoside-phosphate kinase</b> <b>2.7.4.14 Cytidylate kinase</b> <b>2.7.4.- Uridylate kinase</b>	
		<b>2.7.1.37 Protein kinase (HPr kinase/phosphatase)</b> 4.1.1.32 Phosphoenolpyruvate carboxykinase (GTP) <b>4.1.1.49 Phosphoenolpyruvate carboxykinase (ATP)</b>	PDB: 1aq2
	Phosphoglycerate kinase: PF00162, 271 sequences	<b>2.7.2.3 Phosphoglycerate kinase</b> 2.7.2.10 Phosphoglycerate kinase (GTP)	PDB: 13pk
	Aspartokinase: PF00696, 420 sequences	<b>2.7.2.2 Carbamate kinase</b> 2.7.2.4 Aspartate kinase 2.7.2.8 Acetylglutamate kinase 2.7.2.11 Glutamate 5-kinase 2.7.4.- Uridylate kinase	PDB: 1b7b
	Phosphofructokinase-like: PF00365, PF00781, PF01219, PF01513, 451 sequences	<b>2.7.1.11 6-phosphofructokinase</b> 2.7.1.23 NAD(+) kinase 2.7.1.56 1-Phosphofructokinase 2.7.1.90 Diphosphate-fructose-6-phosphate 1-phosphotransferase 2.7.1.107 Diacylglycerol kinase	PDB: 4pfk
	Ribokinase-like: PF00294, PF01256, PF02110, 517 sequences	<b>2.7.1.2 Glucokinase</b> 2.7.1.3 Ketohexokinase 2.7.1.4 Fructokinase 2.7.1.11 6-Phosphofructokinase <b>2.7.1.15 Ribokinase</b> <b>2.7.1.20 Adenosine kinase</b> 2.7.1.35 Pyridoxal kinase 2.7.1.45 2-Dehydro-3-deoxy-gluconokinase 2.7.1.49 Hydroxymethylpyrimidine kinase <b>2.7.1.50 Hydroxyethylthiazole kinase</b> 2.7.1.56 1-Phosphofructokinase 2.7.1.73 Inosine kinase 2.7.1.144 Tagatose-6-phosphate kinase 2.7.1.146 ADP-dependent phosphofructokinase 2.7.1.147 ADP-dependent glucokinase 2.7.4.7 Phosphomethylpyrimidine kinase	PDB: 1rkd
	L-2-Haloacid dehalogenase, 2 sequences	2.7.1.39 Homoserine kinase	PDB: 1j97
	Thiamin pyrophosphokinase, 148 sequences	<b>2.7.6.2 Thiamin pyrophosphokinase</b>	PDB: 1ig0
	Nucleoside-diphosphate kinase: PF00334, 200 sequences	<b>2.7.4.6 Nucleoside-diphosphate kinase</b>	PDB: 2bef
	HPPK: PF01288, 70 sequences	<b>2.7.6.3 2-Amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase</b>	PDB: 1eqo
Group 3: ferredoxin-like fold kinases, 1798 sequences	Guanido kinases: PF00217, 151 sequences	2.7.3.1 Guanidoacetate kinase <b>2.7.3.2 Creatine kinase</b> <b>2.7.3.3 Arginine kinase</b> 2.7.3.5 Lombricine kinase	PDB: 1bg0
	Histidine kinase: PF00512, COG2172, 1377 sequences	<b>2.7.1.37 Protein kinase (histidine kinase)</b>	PDB: 1i59

(continued)

Table 1 Continued

Group	Family and PFAM/COG members	EC no. and kinase activities	Representative PDB or gi
		<b>2.7.1.99 [Pyruvate dehydrogenase(lipoamide)] kinase</b> <b>2.7.1.115 [3-Methyl-2-oxobutanoate dehydrogenase (lipoamide)] kinase</b>	
Group 4: ribonuclease H-like, 723 sequences	COG0837, PF00349, PF00370, PF00871	<b>2.7.1.1 Hexokinase</b> 2.7.1.2 Glucokinase 2.7.1.4 Fructokinase 2.7.1.5 Rhamnulokinase 2.7.1.12 Gluconokinase 2.7.1.16 L-Ribulokinase 2.7.1.17 Xylulokinase 2.7.1.27 Erythritol kinase <b>2.7.1.30 Glycerol kinase</b> 2.7.1.47 D-Ribulokinase 2.7.1.51 L-Fuculokinase 2.7.1.53 L-Xylulokinase 2.7.1.55 Allose kinase 2.7.1.58 2-Dehydro-3-deoxy-galactonokinase 2.7.1.59 N-Acetylglucosamine kinase 2.7.1.60 N-Acylmannosamine kinase 2.7.1.63 Polyphosphate-glucose phosphotransferase 2.7.1.85 Beta-glucoside kinase <b>2.7.2.1 Acetate kinase</b> 2.7.2.7 Butyrate kinase 2.7.2.14 Branched-chain-fatty-acid kinase	PDB: 1dgg
Group 5: TIM beta/alpha-barrel kinase, 231 sequences	PF00224	<b>2.7.1.40 Pyruvate kinase</b>	PDB: 1a49
Group 6: GHMP kinase, 382 sequences	COG1685, COG1907, PF00288, PF01971	2.7.1.6 Galactokinase  <b>2.7.1.36 Mevalonate kinase</b> <b>2.7.1.39 Homoserine kinase</b> 2.7.1.71 Shikimate kinase <b>2.7.4.2 Phosphomevalonate kinase</b> 2.7.1.148 4-(Cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase	PDB: 1h72
Group 7: AIR synthetase-like, 251 sequences	PF00586	2.7.4.16 Thiamine-phosphate kinase 2.7.9.3 Selenide, water dikinase	PDB: 1cli
Group 8: integral membrane kinases, 63 sequences	Dolichol kinase: PF01879, 24 sequences Undecaprenol kinase, 39 sequences	2.7.1.108 Dolichol kinase 2.7.1.66 Undecaprenol kinase	gi 6323655 [349..513] gi 1705428
Group 9: polyphosphate kinase, 63 sequences	PF02503	2.7.4.1 Polyphosphate kinase	gi 7465499 [48..730]
Group 10: riboflavin kinase, 69 sequences	PF01687	2.7.1.26 Riboflavin kinase	gi 6320442
Group 11: inositol 1,4,5-trisphosphate 3-kinase, 57 sequences		2.7.1.127 1D-Myo-inositol-trisphosphate 3-kinase	gi 10176869
Group 12: inositol 1,3,4,5,6-pentakisphosphate 2-kinase, 2 sequences			gi 6320521

(continued)

Table 1 Continued

Group	Family and PFAM/COG members	EC no. and kinase activities	Representative PDB or gi
Group 13: tagatose 6-phosphate kinase, 8 sequences		2.7.1.101 Tagatose kinase	gi 1168382
Group 14: pantothenate kinase, 6 sequences		2.7.1.33 Pantothenate kinase	gi 4191500
Group 15: glycerate kinase, 20 sequences		2.7.1.31 Glycerate kinase	gi 2495546
Group 16: putative glycerate kinase, 18 sequences	COG2379	2.7.1.31 Glycerate kinase	gi 1907334
Group 17: dihydroxyacetone kinase, 43 sequences		2.7.1.29 Glycerone kinase	gi 7387627

The EC (Enzyme Commission) activities shown in bold have solved structures.

miss more distant homologs. In some cases, comparative analysis of protein structural folds allows for the inference of biochemical and biological functional properties. Structure analysis methods are able to detect evolutionary relationships that sequence similarity searches miss, because protein structure conservation persists after sequence similarity disappears. However, similarity of fold alone does not necessarily indicate a common ancestor. Furthermore, structural information is much less readily available than sequence information. Thus, the most effective route to the identification of homologs and the prediction of protein function is provided by the integration of sequence and structure data.

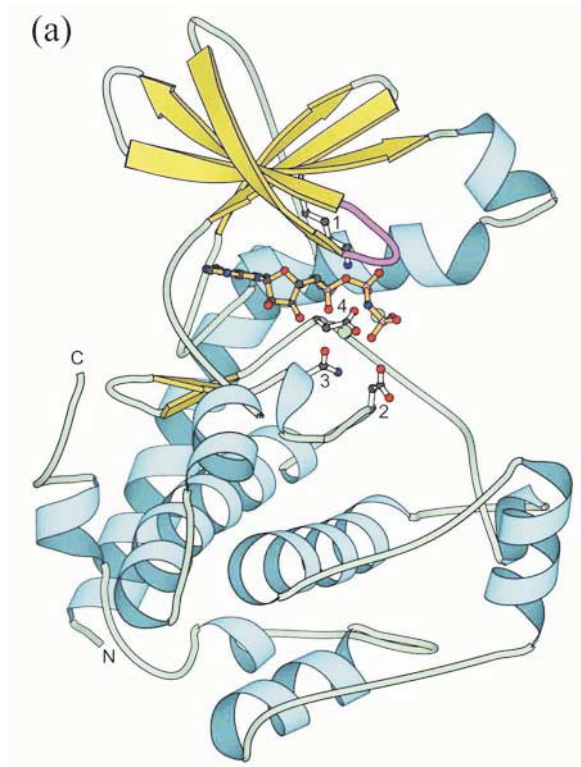
Currently, several protein classification schemes such as SCOP,<sup>3</sup> CATH,<sup>4</sup> and Pfam<sup>5</sup> have been developed for the purpose of cataloging all protein sequences and structures. Here, we present the classification of a single group of proteins that catalyze a phosphoryl transfer reaction, and we subsequently examine the relationships between the fold and biochemistry within this group. The availability of thousands of kinase sequences and hundreds of kinase structures coupled with a wealth of biochemical data make kinases an ideal group of enzymes for such structural/functional classification and analysis. We have analyzed over 17,000 kinase sequences and found that ~98% of them fall into seven general fold groups with known structures. We describe the common structural features of each group and the families therein, emphasizing the shared catalysis and substrate-binding mechanisms as well as variations within the same fold groups. In particular, we try to address the questions of how different kinase structural folds accomplish the same required steps in the common phosphoryl transfer reaction and, in some cases, even recognize exactly the same substrate, and conversely, how kinases of the

same fold recognize substrates of very different structures.

## Results

Our kinase classification is summarized in [Table 1](#). For each group and family therein, the Pfam/COG members are listed as well as the kinase activities and a corresponding representative PDB or gi accession number. The total number of sequences in each family and group is specified as well. The EC (Enzyme Commission) activities shown in bold have solved structures. It should be noted that the activity lists are not exhaustive, as they include only kinase activities with designated EC numbers. Of the 184 enzymes listed in the EC system over our chosen range (EC2.7.1.- through EC2.7.4.-), 112 activities were placed in our kinase families. Sequences for 70 of the remaining kinase activities from our chosen EC range have not been identified and thus could not be included in our analysis. The two remaining kinase activities were excluded intentionally (see Methods).

Overall, 17,310 sequences were analyzed and classified into 30 families. Sequences in each family are supported by statistically significant sequence similarities, indicating that they are homologs. Some of these families unify several Pfam/COG members. In the case of the P-loop kinase family, 18 Pfam/COG members are found to contain statistically significant links and therefore belong to the same protein family. There are nine families, each containing between two and 148 sequences, which are not present in Pfam or COG ([Table 1](#)). These kinase families are assembled into fold groups on the basis of similarity of structural fold. Within a fold group, the core of the nucleotide-binding domain of each family has the same architecture, and the topology is either identical or



(b)

			.....-EEEEEE-....HHHHHHHH.....-EEE-EEE-.....	
cdk2	gi 116051	[ 26]	GEV <b>VALKKI</b> R[10] <b>TAIREISLI</b> [66]-RVLHRD <b>LKPQ</b> <b>LLINTEGA</b> <b>IKLAD</b> FGL[150]	298
capk	gi 7110693	[ 66]	GNH <b>YAMKILD</b> [11] <b>HTLNEKRIL</b> [65]-DLIYRD <b>LKPEN</b> <b>LLIDQQGY</b> <b>IQVTD</b> FGL[163]	351
csk	gi 417209	[215]	GNK <b>VAVKCI</b> K-[ 6] <b>AFLAEASVM</b> [68]-NFVHRD <b>LAARN</b> <b>VLVSEDNV</b> <b>AKVSD</b> FGL[115]	450
I	gi 125397	[192]	-----HLVHAD <b>FGSN</b> <b>NVLT</b> DNGRI <b>TAVID</b> WSE[122]	341
II	gi 135001	[148]	-----CPLHGD <b>LHHEN</b> <b>VLDFGDRG</b> <b>WLAI</b> DPHG[91]	266
III	gi 11545906	[ 34]	AGP <b>VFVKVNR</b> -[ 6] <b>MFEGEVASL</b> [152]ALLHGD <b>LWSGN</b> <b>VAE</b> -DDVG <b>PIIYD</b> PAS[72]	309
IV	gi 146444	[329]	GFDR <b>VEKVI</b> K-----[112]-----NIFPGD <b>MLFKN</b> <b>FVTRHGR</b> <b>VVFYD</b> YDE[100]	578
V	gi 66882	[ 36]	NEN <b>LYLKMTD</b> -[ 8] <b>DVEREKDMM</b> [120]VFSHGD <b>LGDSN</b> <b>IFVKDGKV</b> <b>SGFID</b> LGR[53]	263
VI	gi 2144279	[ 34]	ARDR <b>VERFP</b> K-----[136]-----GLVHGD <b>LGGEN</b> <b>VLW</b> [7]RL <b>TGIVD</b> WDE[70]	281
VII	gi 5542182	[ 90]	SGV <b>FIVKR</b> ST---N <b>IESETFCS</b> [88]-IVNNSD[13] <b>NIML</b> [4]AT <b>VVPID</b> SKI[108]	342
VIII	gi 418468	[ 48]	RRR <b>FVVKFYR</b> -[ 7] <b>QILEHQFA</b> [121]LRLHGD <b>CHAGN</b> <b>ILW</b> --RDG <b>PMFVD</b> LDD[108]	328
IX	gi 14488515	[ 91]	GHL <b>YIIKSFL</b> [18] <b>LCLREIQQQ</b> [83]-ELLVLD <b>T</b> ---Q <b>GVG</b> ---EN <b>LTD</b> PSV[50]	280
X	gi 6681275	[162]	ASN <b>YVAKRYI</b> -----[95]-----QLIVVD <b>T</b> ---Q <b>GVG</b> ---D <b>LYTD</b> PQI[438]	724
XI	gi 3420749	[ 79]	DQG <b>LVGKFST</b> -----[96]-----ELLIVD <b>T</b> ---Q <b>GVN</b> ---D <b>FYTD</b> PQI[547]	751
			(1) (2) (3) (4)	

(c)

I:	.....HH--HHH....-EEEE-.....-EEEE-EEE-E....-EEEEEE-EEEEEE-.....	
gi 1709955	[510]LTED <b>RRRA</b> [31]GY <b>PVLV</b> RAAF <b>AVGGL</b> [23]-Q <b>VLVD</b> KSLKGWKE[63]GECN <b>VQYALNP</b> -ESE <b>QYYII</b> EVNARLS[1537]	1999
gi 115631	[145]TSEDRDL[31]KY <b>PVIV</b> RSAYAL <b>GGL</b> [23]-Q <b>ILVEK</b> SLKGWKE[63]GECN <b>VQYALQP</b> -DGLD <b>YRVIE</b> VNARLS[825]	1118
gi 3873545	[203]TSEDRDL[31]SY <b>PVVI</b> RSAYSL <b>GGL</b> [23]-Q <b>ILVEK</b> SLKGWKE[63]GECN <b>VQYALSP</b> -NSLEY <b>RVIE</b> VNARLS[779]	1160
gi 115627	[124]KAEDRRR[31]GF <b>PCI</b> IRPSFTMGGS[24]KE <b>LLID</b> ESLIKGWKE[64]GGSN <b>VQFAVNP</b> -KNGR <b>LIVIE</b> MNPVRS[768]	1073
II:	.....HHHHHHH....EEEE-.....-EEEE-EEE-E....-EEEE-EEEE-EEEEEE-.....	
gi 11282581	[ 90]RLHNRVS[41]K <b>FPVIAK</b> PLDAD <b>GSA</b> [18]AP <b>IVLQ</b> EFVN <b>HGGV</b> [82]NLFN <b>EDVI</b> RDAKDAN <b>RYLI</b> DI <b>IN</b> YFPG[25]	319
gi 7487924	[ 90]RLHNRVS[41]K <b>FPVIAK</b> PLDAD <b>GSA</b> [18]AP <b>IVLQ</b> EFVN <b>HGGV</b> [82]NLFN <b>EDVI</b> RDAKDAN <b>RYLI</b> DI <b>IN</b> YFPG[25]	319
gi 7487925	[122]HLRNRQS[41]RL <b>PLVAK</b> PLVAD <b>GSA</b> [18]PP <b>LVLQ</b> EFVN <b>HGGV</b> [77]RLFN <b>IDII</b> REHGTRDR <b>FYVI</b> DI <b>IN</b> YFPG[17]	338
gi 7657244	[101]TLLDRSK[42]TF <b>FI</b> CKTRVAH <b>GT</b> -[18]PP <b>CVVQ</b> NFINNAV[79]SLFG <b>IDII</b> INN <b>Q</b> -TGQ <b>HAVI</b> DI <b>IN</b> AFPG[113]	414

Figure 1 (Legend opposite)

related by circular permutation. Homology between families in a fold group is not implied. The structural features of the nucleotide-binding domain of each group are described in the fold group descriptions below.

Most kinase sequences (~98%) belong to families with known structures and could be placed in one of the seven known fold groups. The seven fold groups that describe kinase structures are all either of the  $\alpha + \beta$  or the  $\alpha/\beta$  class in SCOP,<sup>3</sup> with approximately half of the families in these seven groups belonging to each of these two classes. Because the complete genomes of many model organisms have been sequenced, we believe that the sequences for most, if not all genes encoding the remaining 70 kinase activities are already known, but remain to be annotated or experimentally characterized. It is likely that most of these kinases would fall into one of the existing families or fold groups of the current classification scheme.

## Fold group descriptions

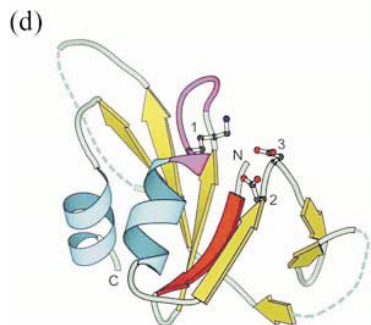
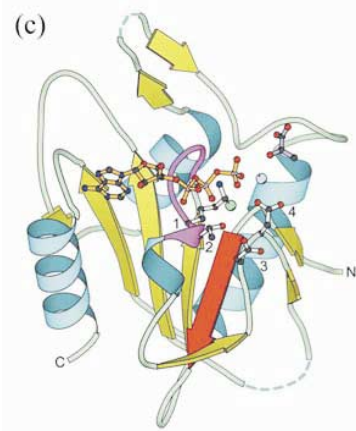
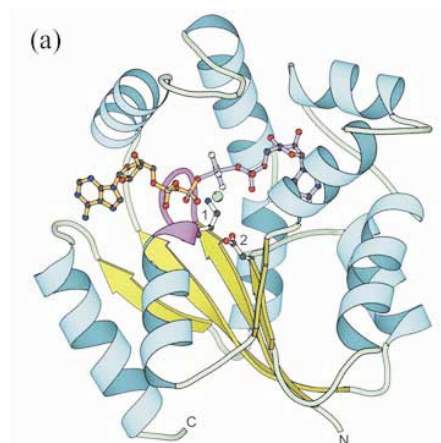
### Group 1: S/T-Y protein kinase-like/SAICAR synthase/ATP-grasp/atypical protein kinase

Group 1 is composed of three families: the protein serine/threonine-tyrosine kinase-like family (referred to here as the protein kinase family), the SAICAR synthase family, and the ATP-grasp family. An evolutionary link between these three families based on structural similarity has been described.<sup>6</sup> In each of the group 1 families, the active site of the enzyme is located in the cleft between two  $\alpha + \beta$  domains (Figure 1(a)). The protein kinase-like family members and SAICAR synthase family members, such as PIPK, have very similar N-terminal domains, but their

C-terminal domains are different, except for a region containing a three-stranded anti-parallel  $\beta$ -sheet and associated  $\alpha$ -helices, which includes the  $\alpha\beta\beta$  unit that is essential for nucleotide binding.<sup>6</sup> The SAICAR synthase and ATP-grasp proteins, on the other hand, share very similar C-terminal domains but have different N-terminal domains. All three families bind ATP along the  $\beta$ -sheet of the  $\alpha\beta\beta$  unit core of the C-terminal domain.<sup>6</sup>

**Family 1a: protein serine/threonine-tyrosine kinase-like.** The protein kinase family has six Pfam/COG members. The majority of the 9600 sequences in this family are eukaryotic protein S/T-Y kinases. However, protein kinase homologs in bacterial and archaeal species have been identified.<sup>7</sup> Links between the six Pfam/COGs are trivial. Two families have a trivial link if at least one sequence in one family produces a significant hit to at least one sequence in the other family within three iterations of PSI-BLAST<sup>1</sup> (*E*-value cutoff 0.001). In this family, several sequences in each of the Pfam/COGs give significant PSI-BLAST<sup>1</sup> hits to multiple sequences in PF00069 (protein kinase domain), which is the largest of the six Pfam/COG members in this family. In addition, there are several kinases with different specificities that can be linked to the protein kinase family but are not assigned by automatic methods. These sequences can be identified by the presence of conserved active-site residues, which are well described for the protein kinase family,<sup>8,9</sup> and by the predicted secondary structure patterns (Figure 1(b)). These proteins include hygromycin-B kinase, streptomycin 3''-kinase, fructosamine 3-kinase, isocitrate dehydrogenase kinase/phosphatase,<sup>10</sup> kanamycin kinase, viomycin kinase, actin-fragmin kinase, homoserine kinase from several proteobacterial species, the kinase

**Figure 1.** The PK family and the ATP-grasp family. (a) The protein kinase fold (cAMP-dependent protein kinase, PDB: 1cdk).<sup>89</sup> Residue 1 (Lys72) interacts with the  $\alpha$  and  $\beta$ -phosphate groups of ATP. Residue 2 (Asp166) is believed to have a role in catalysis. Residues 3 and 4 (Asn171 and Asp184, respectively) each coordinate a magnesium cation. The glycine-rich loop is shown in magenta. In this and all other Figures, the ATP analog is orange, the phosphate-accepting substrate is purple, magnesium cations are green balls, manganese cations are light-blue balls, and potassium cations are orange balls. (b) Addition of distant members to the protein kinase family. The first three sequences are members of the PK family with known structures: cell division protein kinase 2 (cdk2), cAMP-dependent protein kinase (capk), tyrosine-protein kinase CSK (csk). I–IX are representative sequences of kinase groupings that were not assigned by automatic methods to any of the kinase profiles according to the criteria used for analysis, but are in fact evolutionarily related to the PK family: hygromycin-B kinase (I), streptomycin 3''-kinase (II), fructosamine 3-kinase (III), isocitrate dehydrogenase kinase/phosphatase (IV), kanamycin kinase (V), viomycin kinase (VI), actin-fragmin kinase (VII), homoserine kinase (VIII), the kinase domain of ChaK (IX), eukaryotic elongation factor-2 kinase (X), and myosin heavy chain kinase (XI). (c) Addition of distant members to the ATP-grasp family. I, representative sequences of the ATP-grasp fold (large subunit of carbamoyl-phosphate synthase); II, representative sequences of inositol 1,3,4-trisphosphate 5/6 kinase. In these and later alignments, conserved residues known to be involved in catalysis or substrate binding are highlighted in black and are shown in bold white letters. Other highly conserved residues are highlighted in gray. Uncharged residues in mainly hydrophobic sites are highlighted in yellow. The numbers to the far right after each sequence indicate the total length (in amino acid residues) of the sequence. Numbers in brackets specify the number of residues in an insert that are not shown. Secondary structure predictions by Jpred<sup>70</sup> are above each alignment, with E signifying  $\beta$ -strand residues and H signifying  $\alpha$ -helix residues.



(b)

COG1663  
gi|14423746 [41]APVPVIVVGNLSVGGNGKTPVVIWLV[51]VCISANRQHAIELLQHQPDCLGISDDGL[177] 325  
gi|14423747 [48]APLPVLVVGNIIVGGTGKTPMILWMI[51]LMIDPDRPRALQALLAEE-QLDLVLCDLGL[178] 332  
gi|7225899 [53]LPVPVVVVGNIHAGGTGKTPPIVAALV[50]TAVGSSRAEAGRALLAAHPDIGLIVADDGL[185] 344  
gi|6968257 [54]---PVISVGNLSFGGNGKTPPLCKAIA[51]VIVSEDRVKGIEKAF--ELGAKIVVLDDAF[152] 308

PF02283  
gi|1168992 [0]-----MILVTGGARSGKSRHAEALI[33]WRTVERWQHLDLHADINPNEVVLEECVT[97] 181  
gi|13471360 [0]MPDRNKLTFTIIGGARSGKSAHAEILV[34]WTTIDAPLDLAGAL-EALPDNQPVLVDCLT[82] 171  
gi|9655721 [0]----MTVHFVLGGARSGKSSYAELLA[38]WNEHEVPLALAEAL-TRFQRQDLVLVDCLT[90] 179  
gi|10174207 [0]-----MFVFISGAVRSGKSSYAEARA[42]WNVIEAQVDIGSLA-ANFDERDIVLIDCLT[96] 188

PMK  
gi|5729980 [4]GGAPRLVLLFSGKRKSGKDFVTEALQ[55]ADPGFFCR-KIVEGISQ---PIWLVSDTRR[81] 192  
gi|8928269 [0]---MMKIVLISGKRKCGKDIYISERLQ[54]QDYG YFCRVAMEEALS RQQT PYLVSDVRR[82] 189

(e)

PF01293  
gi|1000080 [229]LTLHSSANQGIQNGDVTLFFGLSGTGKTTLSADPHRLLLIGDDEHCWS[54]ITENTRCAY[210] 549  
gi|1084989 [201]LCMHASANVG-KSGDVTVFFGLSGTGKTTLSADPRRNLLIGDDEHVWT[54]ICKNTRVAY[215] 525  
gi|5672673 [344]LSLHSGCNMG-KDGDVALFFGLSGTGKTTLSTDHNGLLIGDDEHCWS[54]VTENTRAAY[213] 666  
gi|7449802 [349]LSLHSGCNMG-KDGDVALFFGLSGTGKTTLSTDHNRLLIGDDEHCWT[54]VTENTRAAY[213] 671  
gi|1172572 [228]ASMHCASANVG-EKGDVAVFFGLSGTGKTTLSTDPKRRLLIGDDEHGWD[53]KTENTRVSY[204] 540

PF00821  
gi|130757 [260]LAEHMLILG[6]KKKYLAFAAFPSACGKTINLA[9]KVECVGDDIAWMK[83]AHPNSRFCT[214] 622  
gi|212540 [278]LAEHMLILG[6]EKRYMAFAAFPSACGKTINLA[9]RIHCVGDDIAWMK[83]AHPNSRFCA[214] 640  
gi|464437 [278]MAEHMLIMG[6]EERFVAAAFPSACGKTINLA[9]KVRVIGDDIAWMK[84]AHPNSRFTA[216] 643  
gi|6225864 [245]LAEHMLILK[6]KAYYFAFAAFPSACGKTINLA[9]RAETLIGDDIAWMR[82]AHPNSRYCT[214] 606  
gi|7436548 [242]LSEHMFIMR[6]RKYFTGAYPSMCGKTISTA[4]E-NIVGDDLVIK[86]SHKNARFTV[235] 621

COG1493  
gi|10176214 [134]TAVHGVLDI--YGTGVLITGSSGVGKSETA[6]GHRLVADDSEVIRR-----[127] 310  
gi|6647570 [134]TSVHGVLDI--YGMGVLIQGDGIGKSETG[6]GHRLVADDRVDVYA-----[126] 309  
gi|7225965 [137]SVKHGVFLDV--FETGVLITGHSGLGKSELA[6]GHSLIADDAVELFR-----[134] 320  
gi|9106416 [139]ATLHGVMFEI--YSTGVLITGEAGSGKSELA[6]GHRLVADDAPEFTQ-----[128] 316  
gi|13357632 [137]SLVHGTLLSV--FTGVILMGESGVGKSELA[6]GHI FVGDDAILVNR-----[122] 308

Figure 2 (Legend opposite)

domain of ChaK (a transient receptor potential (TRP) channel), eukaryotic elongation factor-2 kinase, and myosin heavy chain kinase. Solved structures for kanamycin kinase,<sup>11</sup> actin-fragmin kinase,<sup>12</sup> and the kinase domain of ChaK<sup>13</sup> have confirmed that these three kinases are indeed structurally similar to protein kinases. Multiple alignment of inositol 1,4,5-trisphosphate 3-kinase (group 11) sequences indicates that these sequences may belong to the protein kinase-like family, although such a prediction cannot be made with confidence.

The kinase domain of ChaK, eukaryotic elongation factor-2 kinase, and myosin heavy chain kinase are also known as  $\alpha$ -kinases or atypical kinases. These kinases do not have significant sequence similarity to classical protein kinases;<sup>14</sup> thus, the structural similarity between the  $\alpha$ -kinases and classical protein kinases was unexpected.

Protein kinases are among the most thoroughly studied protein families. This family has several highly conserved active-site residues. A conserved lysine residue from the N-terminal domain interacts with the  $\alpha$  and  $\beta$ -phosphate groups of ATP. The aspartate residue and the asparagine residue in a highly conserved DXXXXN motif play a role in catalysis and in coordinating a secondary magnesium divalent cation, respectively. The primary  $Mg^{2+}$  is coordinated by the conserved aspartate residue of the DFG motif (Figure 1(b)). Protein kinases also have a glycine-rich loop (GXGXXGXV) that interacts with the phosphate groups of the ATP.

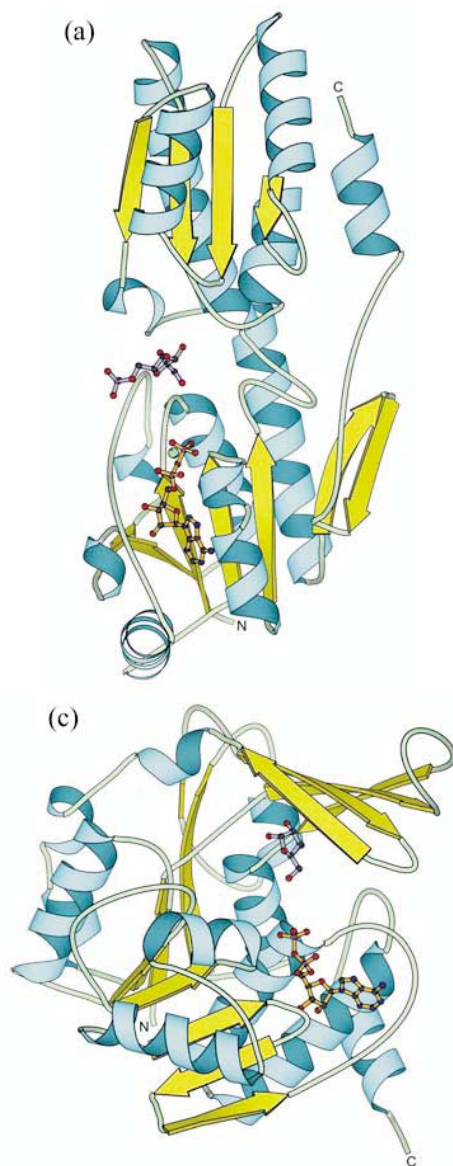
The mechanism of protein kinases was historically thought to be acid–base catalysis. However, recent studies have questioned this hypothesis and led to the proposal that phosphoryl transfer is accomplished by the simultaneous transfer of a proton from the substrate hydroxyl group to an oxygen atom of the  $\gamma$ -phosphate group.<sup>15</sup> The conserved aspartate residue (of the DXXXXN motif)

that was once thought to act as the base catalyst is now suggested to have a role in stabilization of the protonated form of the transferred phosphate.

**Family 1b: SAICAR synthase (as defined in SCOP<sup>3</sup>).** The only kinase member of this family is type II $\beta$  phosphatidylinositol phosphate kinase (PIPK). Since the N-terminal domains of PIPK and protein kinases are very similar, the glycine-rich phosphate-binding loop and the conserved lysine residue located at the N-terminal domain are preserved in both structures. Additionally, PIPK has two conserved aspartate residues, Asp278 and Asp369, which can be aligned with the catalytic aspartate residue of the DXXXXN motif, and the  $Mg^{2+}$ -coordinating aspartate residue of the DFG motif in protein kinases, respectively.<sup>6,16</sup> The roles of these residues are likely to be the same as those of their protein kinase counterparts.

**Family 1c: ATP-grasp.** The ATP-grasp fold describes several different ATP-hydrolyzing enzymes.<sup>17</sup> However, there is only one kinase Pfam member of this family (PF01326: pyruvate phosphate dikinase, PEP/pyruvate-binding domain). In this family, ATP is held between two anti-parallel  $\beta$ -sheets. Hence, the term ATP-grasp is used to describe this fold in SCOP.<sup>3</sup> The mechanism of the phosphotransfer reaction in pyruvate phosphate dikinase involves the reversible phosphorylation of a histidine residue.<sup>18,19</sup> In this enzyme, the binding sites of the nucleotide and the pyruvate are in distant locations on the protein, and the shuttling of the phosphorylated histidine residue between the two active sites is accomplished by a dramatic swivel of the phospho-histidine domain.<sup>20</sup> Inspection of the multiple alignment of inositol 1,3,4-trisphosphate 5/6-kinase sequences has indicated that this kinase also belongs to the ATP-grasp family. Figure 1(c) shows a sequence alignment of the

**Figure 2.** The P-loop kinase family and the PEPCK family. (a) UMP/CMP kinase (PDB: 1qf9)<sup>90</sup> of the P-loop kinase family. Residues 1 and 2 are the conserved Lys19 of the Walker A motif and the conserved Asp89 of the Walker B motif, respectively. Lys19 coordinates the  $\beta$  and  $\gamma$  phosphate groups of ATP, while Asp89 coordinates the magnesium cation. The P-loop is shown in magenta. (b) The alignment of representative sequences of each of the two Pfam/COG members with non-trivial links to the P-loop kinase family. Phosphomevalonate kinase sequences (PMK) containing the Walker A and Walker B motifs are shown. (c) Phosphoenolpyruvate carboxykinase (PDB: 1aq2)<sup>30</sup> of the PEPCK family. Residues 1 and 2 are Lys254 and Thr255 of the Walker A motif, respectively. Residues 3 and 4 are the pair of conserved Asp residues (Asp268 and Asp269) in the PEPCK Walker B motif. Lys254 coordinates the  $\beta$  and  $\gamma$  phosphate groups of ATP, while Thr255, Asp268, and Asp269 coordinate the magnesium cation. The anti-parallel  $\beta$ -strand is shown in red (see the text), and the P-loop is shown in magenta. Most of the N-terminal domain and some elements of the C-terminal domain were removed for clarity. (d) HPr kinase/phosphatase (PDB: 1jb1)<sup>29</sup> of the PEPCK family. Residue 1 is Lys161 of the Walker A motif and is involved in coordination of the  $\beta$  and  $\gamma$  phosphate groups of ATP. Residues 2 and 3 are the pair of conserved Asp residues (Asp178 and Asp179) in the Walker B motif. Asp178 and Asp179 coordinate the magnesium cation. The anti-parallel  $\beta$ -strand is shown in red (see the text), and the P-loop is shown in magenta. Only the core of the C-terminal domain is shown. (e) The alignment of three Pfam members of the PEPCK family: PF01293 (phosphoenolpyruvate carboxykinase-ATP), PF00821 (phosphoenolpyruvate carboxykinase-GTP), and COG1493 (HPr kinase/phosphatase). The conserved histidine and arginine residues are believed to be involved in substrate binding and are catalytically important in PEPCK.<sup>28</sup> In these alignments, blue and green letters signify the Walker A and the Walker B motifs, respectively.



(b)

```

PF00781      ....EEEE-----HHHHHHHHHH....HHHHHHH-----HHHHHHHHHH-----
gi|9506535   [437]VFVNPKSGGKQG-ERIYRKFQ[43]LDCTEKANVY[12]GNDLA[11]ENLMKILKDIESS[250] 801
gi|4503315   [436]VLVNPKSGGKQG-ERILRKFH[43]LDCTIDKANFA[12]GNDLA[11]GSLTKILKDIEQS[241] 791
gi|6978761   [433]VLVNPKSGGKQG-ERILQKFH[43]LDCTIDKANFT[12]GNDLA[11]GSLTKILKEIEQS[241] 788
gi|3551830   [378]VFVNPKSGGKQG-QRVIWKFQ[43]LETIDKANLP[12]GNDLA[11]QNLAKILKDLEMS[75] 567

PF01219      ...-----HHHHHHHHHH....HHHHHHH-----HHHHH-----HHHHHHHHHH---HH....
gi|11260583   [4]SYAAEKKG-KGGIRRVINAFG[50]VI-VELFNTA-VEAAVDHTSTEKHELAKRAKDAGSA[19] 127
gi|7226805   [4]SYAAEKKG-KGGIRRVINAFG[50]VI-VELFNTA-VEAAVDHTSTEKHELAKLAKDAGSA[19] 127
gi|7380388   [4]SYAAEKKG-KGGIRRVINAFG[50]VI-VELFNTA-VEAAVDHTSTEKHELAKRAKDAGSA[19] 127
gi|1170635   [0]----MYK-TTGLTHLINSTK[49]VMALELLNSA-VETTVDRIGTERHELSGRAKDQGSA[19] 118

```

(d)

```

I:      ....---EEE---EEEEEE---...-EE-----HHHHHHHHHHHH---.....-----HHHHHHH....-EEEE-----HHHHHHHHHHHH....
rk gi|132128 [ 0]MGITVIGSINYDLDTF[16]ETHAGGKGLNQAAAIGKL-K[132]WELVDLLVVNEIEGLQI[71]NVSVVDTTGAGDTFLGGLVTQL[40] 333
fk gi|7434221 [ 4]PLIVSFGEMLIDFVPD[12]LKAPGGAPANVCAITKL-G[137]WNEADIIKVSDDE---[49]AVKAVDTTGAGDSFVGAFLVSL[52] 324
rk gi|3915836 [ 1]RNICVIGSCMDLVVT[16]QTVPGGKANQAVAAARL-G[121]IDHATYLTPNEHEASIL[41]-VEPVDTTGAGDTFNAAFAVAL[41] 293
ak gi|6322565 [26]YSLKENDAILVDAKSG[16]KLVAGGAQNTARGAAYVLG[130]LPYATVIIANESEAEAF[57]SSKIVDTNGAGDAFAGGFMAGL[37] 340
gk gi|14484944 [19]-----[11]DVIYGGGESNVAVSLANY-G[141]TSRCDIVLGNEEDAEMH[71]ITDIVDRVGGSFMGGLIYGL[50] 350
              sub              sub              nb              nb

II:      ....EEEEEE---HHHHHH....HHH-----HHHHHHHHHH-----.....-----HHHHHHH....-EE-----HHHHHH....
pfk gi|11496153 [11]LSYTAYNTIVDAIVY[69]VERMGQAGIIANLLSNLDF[169]FPLADSVGMDEAEIAYV[128]QNPVSTVGLGDTISTGAFASY[10] 461
pfk gi|5726296 [ 7]LGYTAYNANVDAIVN[69]TERIGQAGIIANILVGLKV[168]FPMVDSVGMDEAEIAYI[128]PNPVSTVGLGDTISTGTFLSY[8] 454
hyp gi|7518298 [ 6]VSIFTAYNANVDAITK[69]LERIGQAGIIANVLAGLGI[167]LPIADSVGIDEAEIAQL[128]PNPVLTVGLGDTISAGAFITY[7] 451
gk gi|5726298 [22]KGVLLGYNTINDAIKY[68]ELRMGQAGIMANLLGGVYG[156]LGMFYSVGLNEVELASI[130]AKPKSTVGIGDTISSAFIGE[5] 455
hyp gi|7518729 [25]RGVLLAYNTINDAIKY[68]ELRMGQVGIMANLLGGVYG[156]LKHFYSVGLNEVELASV[130]KKPKSTVGIGDTISSAFYSE[4] 457
              sub              sub              nb              nb

```

Figure 3 (Legend opposite)

large subunit of *Escherichia coli*, human, and yeast carbomoyl-phosphate synthases (a representative of the ATP-grasp fold) with inositol 1,3,4-trisphosphate 5/6-kinases.

### Group 2: Rossmann-like

The second kinase fold group contains eight Rossmann-like families. The common structural feature of these families is that the architecture of their nucleotide-binding domain core is three layers ( $\alpha/\beta/\alpha$ ) composed of  $\beta\alpha$  repeats, with the central  $\beta$ -sheet mostly parallel. There is always a change in direction of strand order in the middle of the sheet,<sup>21</sup> resulting in the most common strand order of 321456, although modifications of this topology are frequent. The total number of strands and the strand order of the  $\beta$ -sheet may differ between the families in this group. There is a wide range of insertions or additional domains that are associated mostly with phosphoryl-acceptor substrate binding, accounting for the extremely diverse substrate specificities in this group of kinases. In addition to sharing a common fold of the nucleotide-binding domain, the families in the Rossmann-like group have similar nucleotide-binding patterns. In each family, the nucleotide binds at the C-terminal end of the  $\beta$ -sheet of the core, with the phosphate tail always located at the N-terminal end of one or more  $\alpha$ -helices (Figures 2 and 3). More thorough descriptions of nucleotide-binding specifics for the families within this group are provided below.

**Family 2a: P-loop kinases.** The largest family in group 2 is the P-loop-containing kinases, which unifies 18 Pfam/COG members. Of these Pfam/COG members, 16 have trivial PSI-BLAST<sup>1</sup> links. Alignments identifying the Walker A and Walker B motifs for the remaining two members are shown in Figure 2(b). Additionally, a small number of phosphomevalonate kinase sequences from animals can be assigned to this family.<sup>22</sup> Alignments identifying the conserved diagnostic motifs for these phosphomevalonate kinase sequences are shown in Figure 2(b).

The P-loop kinases contain one three-layered ( $\alpha/\beta/\alpha$ ) domain. For the majority of the members of this family, the central parallel  $\beta$ -sheet is five-stranded with strand order 23145. Nucleotide binding in this family is distinguished by the presence of the conserved Walker A (GXXXXGKT/S)

and Walker B (ZZZZD, where Z is any hydrophobic residue) motifs.<sup>23</sup> The Walker A motif forms a phosphate-binding loop (P-loop) and is found in a variety of different proteins that bind nucleotides.<sup>24</sup> In this family of kinases, the P-loop is located at the end of the first  $\beta$ -strand and includes the first half-turn of the following  $\alpha$ -helix. The conserved lysine residue of the Walker A motif binds to and orients oxygen atoms of the  $\beta$  and  $\gamma$ -phosphate groups of ATP. The essential magnesium cation is coordinated directly by the hydroxyl group of the conserved threonine/serine of the Walker A motif and indirectly by the conserved aspartate residue of the Walker B motif. Figure 2(a) illustrates the Walker A and Walker B motifs, and metal-coordinating residues in UMP/CMP kinase.

The Walker A and Walker B motifs are common in nucleotide-binding proteins. In SCOP,<sup>3</sup> the P-loop-containing nucleotide triphosphate hydrolases fold currently contains 82 proteins in 15 families. Of these proteins, 13 are kinases; thus, ~84% of P-loop-containing proteins with known structures are not kinases. The non-kinase, P-loop-containing proteins are mostly NTPases, which are likely to have come before the kinases, since they catalyze simpler reactions and are involved in more fundamental biological processes.

A variety of catalytic mechanisms are utilized by the kinases in this family. For example, an iso-random Bi-Bi mechanism has been suggested for adenylate kinase.<sup>25</sup> A mechanism involving the synchronous shift of a proton to the transferred phosphate group (similar to that proposed for protein kinases) has been suggested for the UMP/CMP-kinase reaction.<sup>26</sup> If correct, such a mechanism could apply to the other nucleotide-phosphorylating kinases in this family. However, the phosphorylation of some metabolites appears to require a base catalyst.<sup>27</sup> This would apply to certain members of the P-loop kinase family including, for example, phosphoribulokinase and shikimate kinase.

**Family 2b: phosphoenolpyruvate carboxykinase.** The phosphoenolpyruvate carboxykinase (PEPCK) family consists of three Pfam/COG members: PF01293 (phosphoenolpyruvate carboxykinase-ATP), PF00821 (phosphoenolpyruvate carboxykinase-GTP), and COG1493 (HPr kinase/phosphatase). Members of this family are distinguished by their shared nucleotide-binding region,

**Figure 3.** (a) Phosphofructokinase (PDB: 4pfk).<sup>91</sup> (b) The two Pfam members containing diacylglycerol kinase in the phosphofructokinase-like family. Representative diacylglycerol kinase sequences from PF00781 (diacylglycerol kinase catalytic domain, presumed) and PF01219 (prokaryotic diacylglycerol kinase) are shown. (c) Ribokinase (PDB: 1rkd).<sup>37</sup> (d) Addition of distant members of the ribokinase-like family. I, representative sequences of ribokinase-like family: ribokinase (rk), fructokinase (fk), adenosine kinase (ak), gluconate kinase (gk); II, representative sequences of archaeal phosphofructokinase/glucokinase: phosphofructokinase (pfk), glucokinase (glk), related hypothetical proteins (hyp). The underlined residues are conserved motifs in the nucleotide-binding pocket (nb) and the substrate-binding pocket (sub).<sup>92</sup>

characterized by the topology of the  $\beta$ -sheet and the placement of the Walker A motif and an atypical Walker B motif within the nucleotide-binding fold. Solved structures of representatives from PF01293 and COG1493 illustrate this shared topology (Figure 2(c) and (d)). Although no structure has yet been solved for a representative of PF00821, members of this family contain the characteristic Walker A and Walker B motifs. Similar predicted secondary structure distributions and conserved potential catalytic residues indicate that proteins from PF00821 and PF01293 share similar fold and active site architecture (Figure 2(e)).

PEPCK contains two  $\alpha/\beta$  domains. The nucleotide-binding fold is located in the C-terminal domain and is composed of the six-stranded mixed  $\beta$ -sheet and the surrounding  $\alpha$ -helices.<sup>28</sup> The PEPCK family proteins contain the typical Walker A motif and a deviant Walker B motif (Figure 2(e)). Figure 2(a) and (c) illustrate the phosphate-binding loops of a P-loop kinase and PEPCK, respectively. Note the similar structures of the Walker A motifs (in magenta) and the different spatial locations of the Walker B aspartate residues between the two proteins. The topology of the nucleotide-binding fold of PEPCK differs from that in P-loop kinases. The central  $\beta$ -sheet of the PEPCK nucleotide-binding C-terminal domain is a mixed six-stranded sheet with strand order 312456, and strands 1 and 5 are anti-parallel to the rest of the sheet,<sup>28</sup> whereas the central  $\beta$ -sheet of P-loop kinases consists typically of five parallel  $\beta$ -strands of strand order 21345. Furthermore, while the  $\beta$ -strand preceding the Walker A motif and the  $\beta$ -strand preceding the Walker B motif are neighboring structural elements in space in P-loop kinases, PEPCK has an anti-parallel  $\beta$ -strand that lies between the two  $\beta$ -strands associated with the Walker A and Walker B motifs. This strand is colored red in Figure 2(c) and (d). Thus, whether PEPCK is evolutionarily related to P-loop kinases is still a subject of debate.

The structure of the C-terminal segment of the other kinase in PEPCK family, HPr kinase/phosphatase (HPrK/P), has been solved.<sup>29</sup> The fold of this segment of HPrK/P is similar to the C-terminal domain of PEPCK. In HPrK/P, the  $\beta$ -sheet associated with nucleotide binding has at least five strands (strand order 31245, with strand 1 anti-parallel to the rest of the sheet). A sixth strand in this sheet seems likely, although its presence is uncertain, due to poor electron density in the crystal structure in this region.<sup>29</sup> This sixth strand would make the topology of this  $\beta$ -sheet in HPrK/P identical with that of the corresponding  $\beta$ -sheet in PEPCK. Additionally, the placement of the Walker A and Walker B motifs is similar in both HPrK/P and PEPCK (Figure 2(e)). In both kinases, the Walker A motif is located on a  $\beta\alpha$  loop following  $\beta$ -strand 2. The Walker B motif is found at the C-terminal end of  $\beta$ -strand 3.

Two divalent metal cations are present in the active site of PEPCK (Figure 2(c)). The magnesium cation is coordinated by the threonine hydroxyl group of the Walker A motif and indirectly by the conserved pair of aspartate residues in the Walker B motif. The manganese cation is coordinated by the side-chain nitrogen atoms of a histidine residue and a lysine residue in addition to the second aspartate residue of the Walker B motif. The magnesium cation interacts with oxygen atoms of the  $\beta$  and  $\gamma$ -phosphoryl groups of ATP while the manganese cation associates with an oxygen atom of the  $\gamma$ -phosphoryl group and hypothetically with the enolate oxygen atom of pyruvate during catalysis.<sup>30</sup> PEPCK is different from other kinases, in that this enzyme catalyzes both the decarboxylation and phosphorylation of its substrate, oxaloacetate, to form phosphoenolpyruvate. However, the details of this mechanism are unknown.

**Family 2c: phosphoglycerate kinase.** Phosphoglycerate kinase is composed of two  $\alpha/\beta/\alpha$  domains. Both domains adopt a Rossmann-like fold and each contains a six-stranded parallel  $\beta$ -sheet. The N-terminal domain  $\beta$ -sheet has strand order 342156, while the C-terminal domain  $\beta$ -sheet has strand order 321456. The active site of this enzyme is located in the cleft between these two domains.<sup>31</sup> In this enzyme, the C-terminal domain binds ATP, while the N-terminal domain binds the 3-phosphoglycerate substrate. The nucleotide binds at the edge of the  $\beta$ -sheet in the C-terminal domain and is roughly perpendicular to the  $\beta$ -strands. There are no sequence or structural motifs resembling the Walker A or Walker B motifs in phosphoglycerate kinase.

The primary factor contributing to catalysis in phosphoglycerate kinase is transition state stabilization. In this enzyme, all three peripheral oxygen atoms of the transferred phosphate group are stabilized by interactions with protein residues or the required divalent cation in the transition state. However, only two of these oxygen atoms are stabilized when the phosphate group is fully bonded to either the phosphate donor (ATP) or acceptor (phosphoglycerate substrate).<sup>32</sup>

**Family 2d: aspartokinase.** The only member of this family with a solved structure is carbamate kinase from *Enterococcus faecalis*<sup>33</sup> and *Pyrococcus furiosus*.<sup>34</sup> Carbamate kinase and aspartokinase (along with acetylglutamate kinase, glutamate 5-kinase, and uridylylate kinase) belong to the same Pfam profile (PF00696: amino acid kinase family). Therefore, the structure of carbamate kinase can serve as a prototype for other members of the family. The nucleotide-binding domain in carbamate kinase is composed of three layers ( $\alpha/\beta/\alpha$ ), including a  $\beta$ -sheet with Rossmann-like topology.<sup>33</sup> The bound nucleotide is located at the edge of the  $\beta$ -sheet strands and is approximately perpendicular to the direction of the  $\beta$ -strands, as

is shown in the complex structure of the carbamate kinase-like carbamoyl phosphate synthetase from *P. furiosus*.<sup>34</sup>

**Family 2e: phosphofructokinase-like.** The phosphofructokinase-like family contains four Pfam members. Links between three of these members are trivial *via* PSI-BLAST.<sup>1</sup> The link between another Pfam member, PF01219 (prokaryotic diacylglycerol kinase), to the phosphofructokinase-like family was established through multiple sequence alignment analysis and secondary structure predictions (Figure 3(b)).

The fold of phosphofructokinase consists of two  $\alpha/\beta/\alpha$  domains. The nucleotide-binding N-terminal domain has a seven-stranded mixed  $\beta$ -sheet of strand order 3214567, where strands 3 and 7 are anti-parallel to the rest of the sheet (Figure 3(a)). The active site is located in the cleft between the two domains.<sup>35</sup> The ATP moiety, which is positioned above the  $\beta$ -sheet of the N-terminal domain and is approximately parallel with the  $\beta$ -strands, sits between an  $\alpha$ -helix and a long loop segment (Figure 3(a)).

**Family 2f: ribokinase-like.** The ribokinase-like family contains several carbohydrate kinases. All three Pfam members in this family have trivial PSI-BLAST<sup>1</sup> links. An additional grouping of archaeal phosphofructokinase/glucokinase sequences can be placed in the ribokinase-like family. An alignment for this assignment is shown in Figure 3(d). The solved structure of glucokinase from *Thermococcus litoralis* shows that the structure of this archaeal ADP-dependent kinase is indeed similar to that of the other members of the ribokinase-like family.<sup>36</sup>

The core of the ribokinase-like fold is a three-layered domain ( $\alpha/\beta/\alpha$ ) with a central eight-stranded  $\beta$ -sheet. The strand order of this sheet is 21345678, with strand 7 anti-parallel to the rest of the  $\beta$ -sheet. Ribokinase has an additional subdomain composed of a four-stranded  $\beta$ -sheet. In addition to acting as a “substrate lid”, this  $\beta$ -sheet forms the dimerization surface.<sup>37</sup> In ribokinase, the nucleotide-binding site is found along a shallow groove in the core of the sole  $\alpha/\beta/\alpha$  domain.<sup>37</sup> The ATP moiety lies at the edge of the  $\beta$ -sheet and is roughly perpendicular to the adjacent  $\beta$ -strands (Figure 3(c)).

**Family 2g: L-2-haloacid dehalogenase (HAD-like).** The HAD-like family contains only two kinase sequence members, which are the bifunctional homoserine kinase/phosphoserine phosphatase (ThrH) from *Pseudomonas aeruginosa*.<sup>38</sup> PSI-BLAST<sup>1</sup> establishes a link between these sequences and phosphoserine phosphatase, which has the HAD-like fold.<sup>39,40</sup> The core of the HAD-like fold is a six-stranded parallel  $\beta$ -sheet of strand order 321456 with an additional four-helix bundle subdomain.<sup>41</sup> The nucleotide-binding site in a HAD-like enzyme is suggested to be located in

the cleft between the two domains of the fold,<sup>41</sup> however, the details of the orientation of the ATP moiety are unknown.

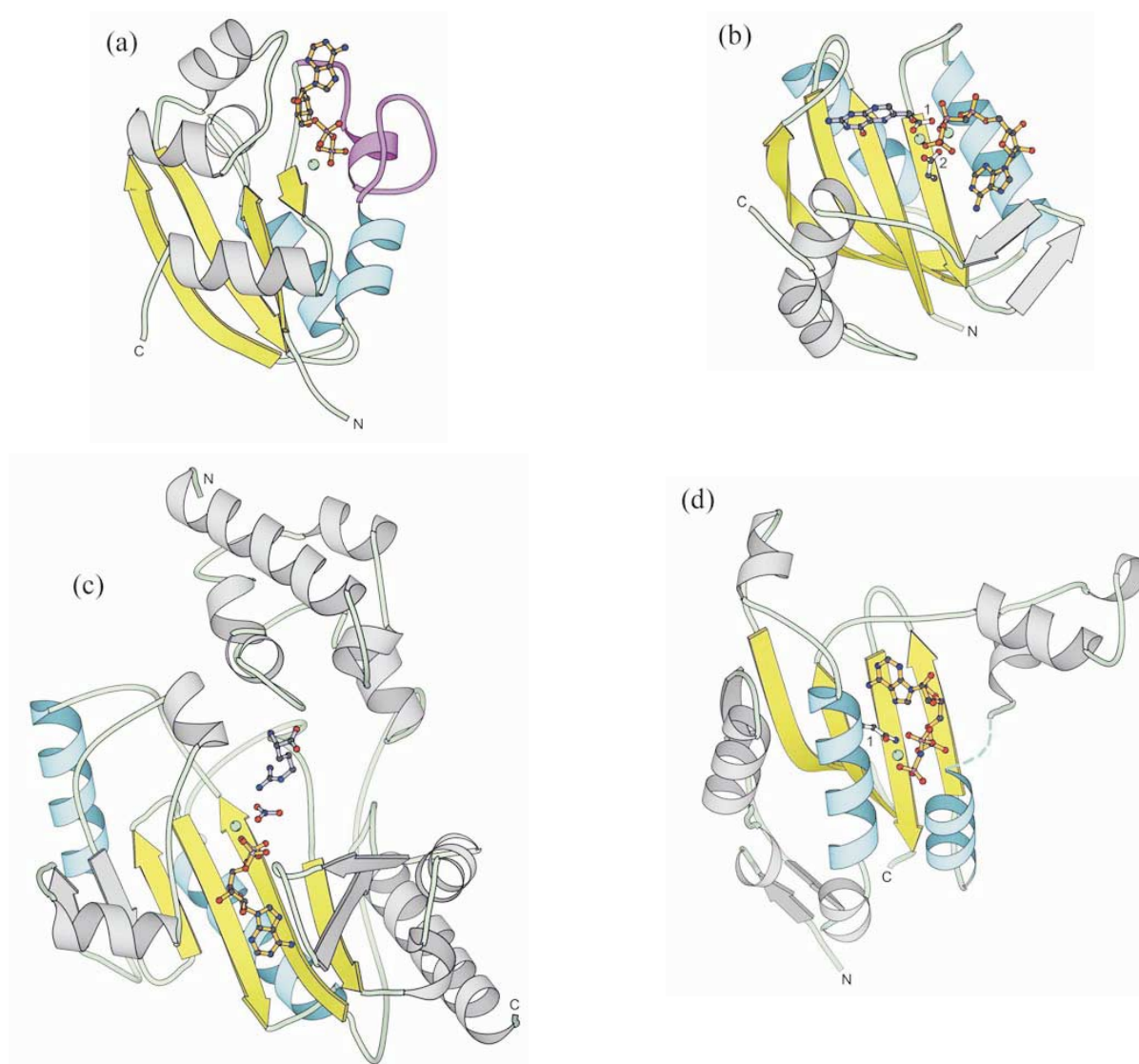
HAD-like family enzymes typically utilize a conserved aspartate residue for nucleophilic catalysis in a reaction that includes a covalent intermediate involving the phosphate group and the aspartate residue.<sup>42,43</sup> This mechanism could be used for the kinase phosphotransfer reaction. Alignment of the homoserine kinase isozyme with phosphoserine phosphatase suggests that Asp7 (gi|4138297) is the residue that is phosphorylated for the formation of the covalent intermediate in the homoserine kinase isozyme.

**Family 2h: thiamin pyrophosphokinase.** Thiamin pyrophosphokinase (TPK) is a two-domain protein. The ATP-binding C-terminal domain has Rossmann-like topology and is composed of three layers ( $\alpha/\beta/\alpha$ ) with a central six-stranded parallel  $\beta$ -sheet of strand order 432156, while the N-terminal domain consists of a four-stranded  $\beta$ -sheet and a six-stranded  $\beta$ -sheet that form a flattened sandwich (“jelly-roll topology”).<sup>44</sup> TPK is a homodimer with the active site located in a cleft at the interface of the component monomers. Thus, active-site residues are contributed by the N-terminal domain of one monomer and the C-terminal domain of the opposing monomer.<sup>44</sup> The precise orientation of the nucleotide in the active site has not been determined.

### Group 3: ferredoxin-like fold kinases

Group 3 is composed of kinases whose nucleotide-binding domain core resembles the ferredoxin-like fold. The four families in this group are as follows: nucleoside-diphosphate (NDP) kinase, 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (HPPK), guanido kinases, and histidine kinase. The ferredoxin fold (also known as the  $\alpha - \beta$  plait fold or the  $\alpha + \beta$  sandwich) is characterized by its  $\beta\alpha\beta\beta\alpha\beta$  unit with the four-stranded anti-parallel  $\beta$ -sheet having strand order 2314 and two helices on one side of the sheet. One exception is the histidine kinase family, in which the core of the nucleotide-binding domain is related to the ferredoxin-like topology by a circular permutation (see below). An interesting feature of this group is that the mode of nucleotide binding differs significantly between each of the families, in terms of structural elements utilized in nucleotide-protein interactions, and in the orientation and position of the nucleotide relative to the ferredoxin-like core. Figure 4 illustrates the orientation of the bound nucleotide in NDP kinase, HPPK, arginine kinase, and histidine kinase.

**Family 3a: NDP kinase.** In addition to the ferredoxin-like core of NDP kinase, this enzyme has several other secondary structural elements, including the Kpn loop, an  $\alpha$ -helix hairpin, and a C-terminal extension (Figure 4(a)). The Kpn loop



**Figure 4.** Nucleotide binding in the ferredoxin-like group. (a) Nucleoside-diphosphate kinase (PDB: 2bef)<sup>93</sup> of the NDP kinase family. The Kpn loop is shown in magenta. (b) 6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase (PDB: 1eqo)<sup>47</sup> of the HPPK family. Residues 1 (Asp95) and 2 (Asp97) are involved in  $Mg^{2+}$  coordination. (c) Arginine kinase (PDB: 1bg0)<sup>50</sup> of the guanido kinase family. (d) Histidine kinase (PDB: 1i59)<sup>51</sup> of the histidine kinase family. Residue 1 (Asn409) is involved in the coordination of the magnesium cation. In these structures, the ferredoxin-like core of the proteins is shown in yellow  $\beta$ -strands and blue  $\alpha$ -helices. Any additional secondary structural elements are shown in gray.

is a small, compact structural element containing an interesting combination of helical structures: a turn of  $3_{10}$  helix, a turn of polyproline II left-handed helix, and a turn of standard  $\alpha$ -helix.<sup>45</sup> The Kpn loop and the  $\alpha$ -helix hairpin constitute a nucleotide-binding site that is unique to NDP kinase (Figure 4(a)). In terms of the orientation and position of this substrate, the nucleotide lies at the edge of the  $\beta$ -sheet that defines the ferredoxin-like core, with the adenine base more distant from the sheet and the phosphate tail extending towards the sheet.<sup>46</sup>

Catalysis in NDP kinase occurs by a ping-pong mechanism in which the phosphoryl group is transferred first from the nucleoside triphosphate to the enzyme, involving the formation of a

covalent intermediate of the phosphate with a histidine residue in the active site, followed by the transfer of the phosphate group to the nucleoside diphosphate substrate.

**Family 3b: HPPK.** In addition to the ferredoxin-like core of HPPK, this enzyme has a pair of  $\beta$ -strands and a pair of  $\alpha$ -helices at the C-terminal end of the protein.<sup>47,48</sup> In HPPK, ATP is bound in the region between the  $\alpha 2$ - $\beta 4$  connecting loop and a short C-terminal  $\beta$ -strand that is not part of the ferredoxin-like core (Figure 4(b)). The ATP moiety lies at the edge of the  $\beta$ -sheet of the ferredoxin-like core and is curled such that the adenine base and ribose sugar angle away from the  $\beta$ -sheet, while the triphosphate tail points towards the  $\beta$ -sheet. The

adenine base and the ribose sugar residue lie in the same plane as the  $\beta$ -sheet. However, the triphosphate tail reaches over the  $\beta$ -sheet on the opposite side as the  $\alpha$ -helices associated with the ferredoxin-like core.<sup>47,48</sup>

HPPK utilizes two magnesium cations in its active site, each of which is coordinated by two aspartate residues. The mechanism of HPPK has not been elucidated, but it is presumed to be a direct in-line transfer of the pyrophosphate group. Suggestions for the mechanism include roles for the two magnesium cations and acid–base catalysis with a water molecule acting as the general base.<sup>47</sup>

**Family 3c: guanido kinases.** The fold of the guanido kinase family consists of two domains. The smaller N-terminal domain is composed entirely of  $\alpha$ -helices, and the nucleotide-binding C-terminal domain is composed of an eight-stranded  $\beta$ -sheet of strand order 23451687, which is flanked by seven  $\alpha$ -helices.<sup>49,50</sup> The middle four strands of the  $\beta$ -sheet and associated  $\alpha$ -helices have ferredoxin-like fold topology. Figure 4(c) illustrates the ferredoxin-like topology within this fold.

In arginine kinase, ATP binding is accomplished by interactions with five arginine residues and a magnesium cation. The ATP moiety lies in the plane above the  $\beta$ -sheet, rather than at the edge of the  $\beta$ -sheet, and is oriented approximately parallel with the  $\beta$ -strands (Figure 4(c)). The nucleotide is positioned above the center two strands of the four-stranded section that resembles the ferredoxin-like fold. In this enzyme, the bound nucleotide and the  $\alpha$ -helices that compose the ferredoxin-like topology lie on opposite sides of the  $\beta$ -sheet.<sup>50</sup>

Arginine kinase catalyzes an associative in-line phosphotransfer reaction. The primary factor in catalysis appears to be substrate alignment by positioning reaction components in close proximity and promoting proper alignment of orbitals, although acid–base catalysis, polarization, and transition state stabilization may contribute to the reaction.<sup>50</sup>

**Family 3d: histidine kinases.** There are two Pfam/COG members of this family. The link between the members is trivial *via* PSI-BLAST.<sup>1</sup> Histidine kinases (HKs) catalyze a *trans*-autophosphorylation reaction in the two-component system of signal transduction. The fold of the ATP-domain (the catalytic domain) is an  $\alpha/\beta$  sandwich composed of a five-stranded  $\beta$ -sheet and four  $\alpha$ -helices (Figure 4(d)).<sup>51,52</sup> The core of this fold has a tertiary structure similar to that of the ferredoxin fold. The HK fold has topology  $\alpha\beta\alpha\beta\beta$  with strand order 3421, which can be related to the ferredoxin-like topology by a circular permutation consisting of cutting the loop between the first and second  $\beta$ -strands and connecting the natural termini.

There are two classes of HKs, which can be differentiated by their domain organization.<sup>53</sup> EnvZ

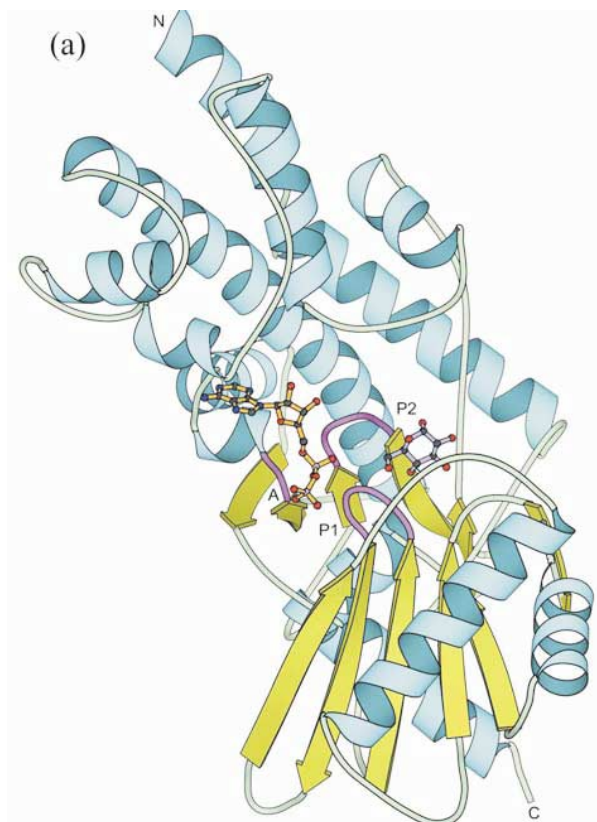
is a member of class I in which the H-domain (the domain that contains the histidine phosphorylation site) directly precedes the ATP domain. CheA is a member of class II in which the H domain and ATP domain are separated by at least one other domain. The ATP domains of HKs are structurally similar to the ATP-binding domains of the GHF family (DNA gyrase/Hsp90/MutL).<sup>51,52</sup> There is currently some debate as to whether both HK classes or only class II bind ATP in the same conformation as the GHF family.<sup>51,52</sup> In both classes, there are four conserved motifs that contribute to the nucleotide-binding pocket; namely, the N, G1, F, and G2 boxes.<sup>54</sup> The required magnesium cation is coordinated by direct interactions with an asparagine residue and indirectly by a histidine residue and an arginine residue.<sup>51</sup>

In HK, the ATP moiety lies above the  $\beta$ -sheet that is part of the ferredoxin-like core (Figure 4(d)). The nucleotide is above two of the  $\beta$ -strands and is found at the edge of the sheet rather than the center of it. In HK, the adenine base is nearer to the  $\beta$ -sheet than the triphosphate tail, which angles away from the  $\beta$ -strands. The nucleotide and the  $\alpha$ -helices associated with the ferredoxin-like core are located on the same side of the  $\beta$ -sheet.<sup>51</sup>

HK, as noted above, operates *via* the two-component system. Here, one HK monomer phosphorylates a histidine residue in the other monomer of the homodimer, which results in a high-energy phosphoryl group. The regulatory domain of the cognate response regulator (RR) then catalyzes the reaction that transfers the phosphoryl group from the histidine residue to an aspartate residue in the RR. The mechanism is somewhat similar to that of NDP kinase, in that both involve the formation of a high-energy phosphohistidine residue. However, NDP kinase phosphorylates a histidine residue in its own active site, while HK phosphorylates a histidine residue in another HK monomer.

#### Group 4: ribonuclease H-like kinases

There are four Pfam/COG members in this group. Three of these members have trivial links *via* PSI-BLAST.<sup>1</sup> The fourth member (PF00871: the acetate kinase family, which contains acetate kinase and butyrate kinase) was predicted to be a member of this family by Buss *et al.*<sup>55</sup> The recently solved structure of acetate kinase shows that this enzyme does in fact adopt the ribonuclease H-like fold.<sup>56</sup> Multiple alignment of 2-dehydro-3-deoxygalactonokinase sequences indicates that this kinase activity belongs to the ribonuclease H-like group (Figure 5(b)). The ribonuclease H-like group contains the ASKHA (acetate and sugar kinase/hsc70/actin) superfamily, the structures of which are characterized by duplicate domains of the ribonuclease H-like fold. The ribonuclease H-like fold is composed of three layers ( $\alpha/\beta/\alpha$ ) (Figure 5(a)). The five-stranded mixed  $\beta$ -sheet has strand order



(b)

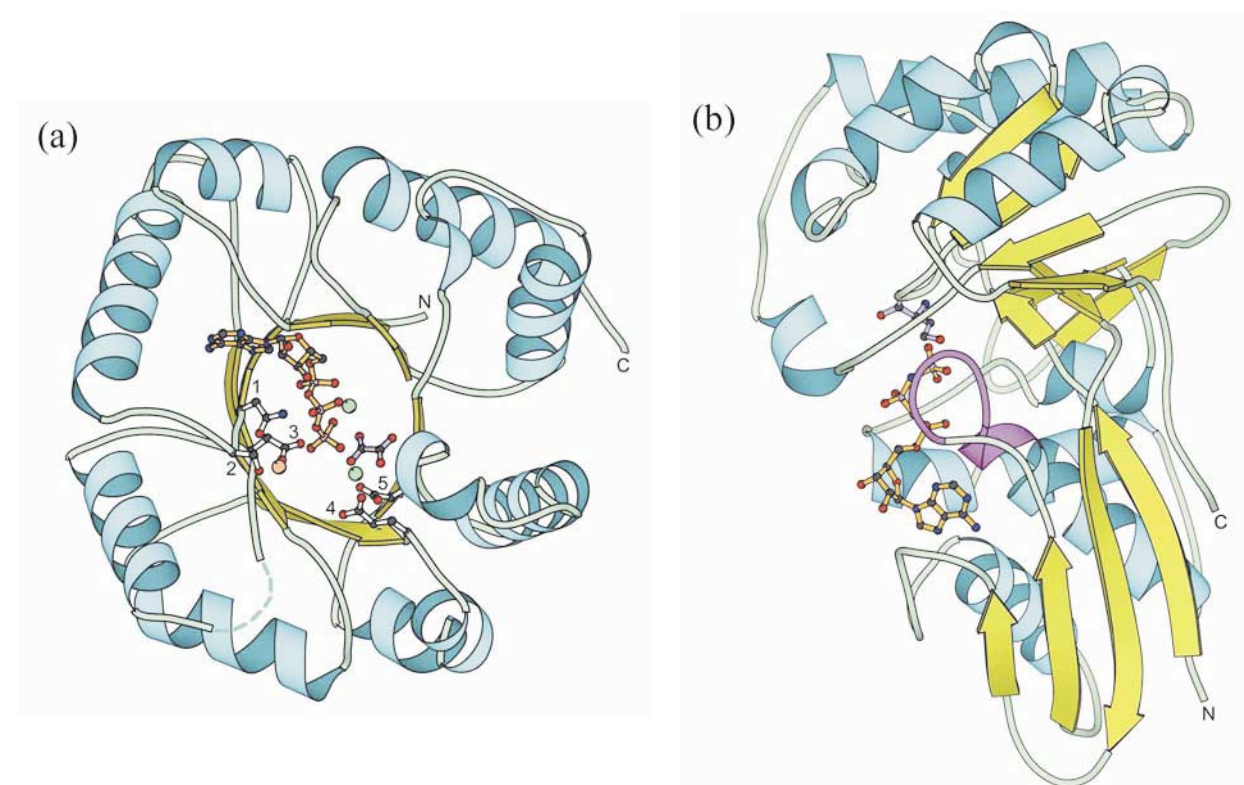
```

GlyK      .....EEEEEE---EEEEEE.....---EEEEEE---EEE.....-EEEE-EEEEE---.....
gi|13431541 [ 5] YIIALDQGTSSRAVL[52] QIAAIGITNQRETTIV[170] GQAKNTYCTGCFMLMHTGD[224] 502
gi|5107527 [ 4] YIVALDQGTSSRAVV[52] QIAAIGITNQRETTIV[170] GMAKNTYCTGCFMLMNTGE[224] 501
gi|2493483 [ 7] YVMALDLGTTGNRAIL[52] QIVAILGLTVQRETCLL[168] GLLKCTYGTGAFLVANTGQ[217] 495
gi|6016137 [ 3] YMLAIDQGTSSRAIL[52] QVAGIGITNQRETTVV[170] GMAKNTYCTGCFMLMNTGE[220] 496
gi|5834428 [11] LVGAVDQGTSSSTRFLV[56] NIKAIGVSNQRETTVV[179] GQAKNTYCTGCFLLCNTGH[262] 559
           phosphate 1                               phosphate 2

DDGK      .....EEEE-EEEEE.....EEEEEEEE-----.....-EEEE-EEEEE---.....
gi|13422026 [ 4] IVIVGDWGGSRRLRLWL[39] YAILCGMVGSRMGWTE[ 65] RHLVLVLPCTHNKWAIVEDG[147] 306
gi|401613 [ 3] RYIATDWGSTNLRAWL[40] PVVMAGMVGSNVGWKV[ 56] SSLYVMPCTHCKWVQADSQ[142] 292
gi|13473974 [337] AVAALDWGTTTLRAWL[42] PIIICGMAGSRQGWIE[ 60] RHLACMPCTHCKWVVVEDG[156] 646
gi|15073756 [ 5] YYAAVDWGTSSFRLWI[42] PIIICGMAGARQGWKE[ 61] SHLVCMPCPTHCKWVRLADD[147] 306
gi|14523461 [19] VTVVLDWGTTGFRAFL[43] DIVAAGMIGSRNGWIE[ 61] -IIVVLPCTHAKWAEIRGG[151] 324
           phosphate 1                               phosphate 2

```

**Figure 5.** The ribonuclease H-like group. (a) Hexokinase (PDB: 1dkg).<sup>58</sup> The loop regions of the conserved nucleotide-binding motifs (PHOSPHATE 1 (P1), PHOSPHATE 2 (P2), ADENOSINE (A)) are shown in magenta. Only the C-terminal domain is shown. (b) Alignment of 2-dehydro-3-deoxygalactonokinase (DDGK) sequences and glycerol kinase (GlyK) sequences. The PHOSPHATE 1 and PHOSPHATE 2 motifs are indicated.



**Figure 6.** (a) Metal cofactor coordination and nucleotide orientation in the TIM  $\beta/\alpha$ -barrel kinase family (pyruvate kinase, PDB: 1a49).<sup>63</sup> Residues 1 (Asn74), 2 (Ser76), and 3 (Asp112) coordinate the potassium cation. Residues 4 (Glu271) and 5 (Asp295) coordinate one of the magnesium cations. The C-terminal subdomain was removed for clarity. (b) Homoserine kinase (PDB: 1h72)<sup>67</sup> of the GHMP kinase group. The novel P-loop is shown in magenta.

32145, with strand 2 anti-parallel to the rest of the sheet. The topology of the core of this fold is  $\beta\beta\beta\alpha\beta\alpha\beta\alpha$ . Nucleotide binding and divalent metal coordination are achieved by interactions of ATP with several motifs conserved within the ASKHA superfamily.<sup>57</sup> These conserved motifs include the ADENOSINE motif that interacts with ribosyl and the  $\alpha$ -phosphoryl group of ATP, the PHOSPHATE 1 motif that interacts with  $Mg^{2+}$  through coordinated water molecules, and the PHOSPHATE 2 motif that interacts with the  $\beta$  and  $\gamma$ -phosphoryl groups of ATP (Figure 5(a)). A modeled active site of hexokinase predicts that the required divalent metal cation is not liganded to the protein directly, but is positioned by coordinated water molecules.<sup>58</sup> The mechanism of kinases in this group is presumed to be acid–base catalysis. In hexokinase, an aspartate residue is the putative catalytic base that deprotonates the 6-hydroxyl group of glucose.<sup>58,59</sup>

#### Group 5: TIM $\beta/\alpha$ -barrel fold kinases

Group 5 is described by the TIM  $\beta/\alpha$ -barrel fold, which consists of an eightfold repeat of  $\beta\alpha$  units that form a closed barrel. The barrel is composed of an inner layer of eight parallel  $\beta$ -strands of strand order 12345678 and an outer layer of eight  $\alpha$ -helices (Figure 6(a)). This fold characterizes many different enzyme families that have

extremely low levels of sequence similarity and catalyze unrelated reactions.<sup>60,61</sup> The active site of all TIM  $\beta/\alpha$ -barrel enzymes is located at the C-terminal end of the parallel  $\beta$ -strands. The only kinase known to adopt this fold is pyruvate kinase, which has an additional domain inserted at the C-terminal end of  $\beta$ -strand 3.<sup>62</sup> The nucleotide-binding pattern of pyruvate kinase is thought to be novel.<sup>63</sup> The triphosphate tail of the ATP moiety is held by hydrogen bond interactions with two arginine residues, an asparagine residue, a lysine residue, and three metal cations (two  $Mg^{2+}$  and one  $K^+$ ). The adenine ring sits in a pocket that is bounded by a histidine residue, a proline residue, and a tyrosine residue. One distinctive feature of the pyruvate kinase active site is the coordination of each of the  $\gamma$ -phosphoryl peripheral oxygen atoms to a different inorganic cofactor.<sup>63</sup> One of the magnesium cations is coordinated by the carboxylate groups of an aspartate residue and a glutamate residue. The other magnesium cation is not liganded to the protein directly. The potassium cation is coordinated by the carbonyl group of a threonine residue, the hydroxyl group of a serine residue, the carboxylate group of an aspartate residue, and the carboxamide group of an asparagine residue. The coordination of the metal cofactors is detailed in Figure 6(a). The reaction catalyzed by pyruvate kinase is presumed to be

direct in-line phosphotransfer *via* acid–base catalysis, although the group(s) responsible for the acid–base catalysis has not been identified. Recent work suggests the possibility that a proton relay through a series of conserved residues may be responsible for acid–base catalysis.<sup>63</sup>

#### Group 6: GHMP kinases

The members of the GHMP kinase superfamily constitute group 6. The GHMP kinase superfamily was named after its original four members: galactokinase, homoserine kinase, mevalonate kinase, and phosphomevalonate kinase.<sup>64</sup> The crystal structure of two members of this family, homoserine kinase and mevalonate kinase, have been solved.<sup>65,66</sup> The fold of this group consists of two  $\alpha + \beta$  domains, with the active site in the cleft between the two domains (Figure 6(b)). The N-terminal domain contains two  $\beta$ -sheets and four  $\alpha$ -helices. The C-terminal domain has a ferredoxin-like core with four additional  $\alpha$ -helices. The nucleotide-binding site resides mostly with the N-terminal domain. Nucleotide binding is accomplished by a novel P-loop with a conserved PXXX-GSSAA motif. The structure of homoserine kinase revealed the presence of an unusual left-handed  $\beta\alpha\beta$  unit in the N-terminal domain. The second  $\beta\alpha$  loop in this unit contains the novel phosphate binding loop (Figure 6(b)).<sup>65</sup> Notably, the orientation of the ATP is different in the GHMP phosphate-binding loop than in the classical Walker A P-loop. A glutamate residue acts to coordinate the essential magnesium cation. In homoserine kinase, it has been suggested that the homoserine hydroxyl group is deprotonated not by a catalytic base, but by interaction with the  $\gamma$ -phosphate group in a mechanism similar to that proposed for protein kinases.<sup>67</sup> There are four Pfam/COG members of this group. The links between the four members are trivial *via* PSI-BLAST.<sup>1</sup>

#### Group 7: AIR synthetase (*PurM*)-like

Two kinases, thiamine-phosphate kinase and selenide, water dikinase, belong to this group. Although the structures of these enzymes have not been solved, the known structure of the homologous aminoimidazole ribonucleotide synthetase<sup>68</sup> (AIR synthetase, *PurM*) can serve as a prototype for this group. AIR synthetase has two  $\alpha/\beta$  domains. The N-terminal domain contains a mixed four-stranded  $\beta$ -sheet with four  $\alpha$ -helices on one side of the sheet. The C-terminal domain has a mixed six-stranded  $\beta$ -sheet flanked by seven  $\alpha$ -helices. Four of the  $\beta$ -strands and two of the  $\alpha$ -helices in the C-terminal domain adopt a tertiary structure and topology that resembles the ferredoxin-like fold. The crystal structure indicates that this enzyme exists as a dimer, with the active site likely to be located in a cleft between the two subunits.<sup>68</sup> A sulfate ion is bound in this cleft and

could indicate a phosphate-binding site, although not necessarily the ATP-binding site, since both substrates of AIR synthetase contain a phosphate group.<sup>68</sup> Mutagenesis and affinity-labeling studies of this enzyme suggest that the ATP-binding site is located close to the N terminus of the enzyme.<sup>69</sup> Thus, it appears that nucleotide-binding is accomplished predominantly by the N-terminal domain of one subunit in the dimer, while the C-terminal domain of the opposing subunit binds the second substrate. A similar situation may be seen in the kinase members of this family. However, as no substrate/product complex structure is available for any member of this group, the exact nucleotide-binding mode is not clear.

#### Groups 8–17: kinases with unknown structures

The remaining groups, which account for only 2% of the sequences in our analysis, contain kinases with unsolved structures. For each group, Jpred<sup>70</sup> was used to generate secondary structural predictions. On the basis of these predictions, the majority of these groups are expected to be of the  $\alpha/\beta$  or  $\alpha + \beta$  protein classes in SCOP,<sup>3</sup> like their known-structure kinase counterparts, which are all  $\alpha/\beta$  or  $\alpha + \beta$  proteins. The exceptions are the two families in group 8 (dolichol kinase and undecaprenol kinase), which are both predicted to be composed almost entirely of  $\alpha$ -helices. This is not unexpected, because these are both integral membrane proteins.

For some of these groups, it is tempting to predict their fold on the basis of similar substrate specificities. For example, one might conjecture that inositol 1,4,5-trisphosphate 3-kinase (group 11) and inositol 1,3,4,5,6-pentakisphosphate 2-kinase (group 12) are likely to be members of group 1, which includes many of the other kinase participants in the inositol phosphate metabolism pathway. Similarly, dihydroxyacetone kinase (group 17) may belong to the ribonuclease H-like group, which includes glycerol kinase. However, such links are difficult to establish without the expected presence of specific conserved motifs and confirmed active-site residues. Furthermore, functional links such as these can be misleading. One such example is pantothenate kinase. Group 14 contains eukaryotic pantothenate kinase sequences. The solved structure of prokaryotic pantothenate kinase identifies the enzyme as a member of the P-loop kinase family.<sup>71</sup> However, due to the lack of sequence identity between the prokaryotic and eukaryotic versions of this protein,<sup>72</sup> in conjunction with dissimilar secondary structural predictions, eukaryotic pantothenate kinase is expected to adopt a fold distinct from that of its prokaryotic counterpart. Additionally, two distinct families of glycerate kinase sequences are found in groups 15 and 16. Although these proteins have been predicted to have the same biochemical activity, sequence similarity between the members of the two groups is not detected.

Group 15 glycerate kinases are from bacterial species, primarily of the firmicutes group and of the gamma subdivision of the proteobacterial group. Group 16 contains predicted glycerate kinase sequences from eukaryotes and archaeal species in addition to several bacterial species. Most group 16 bacteria are from the alpha subdivision of the proteobacterial group.

Potential active-site components for some of these groups can be predicted by studying multiple alignments of the sequences. Conserved patterns containing several glycine residues and other small residues are typical of nucleotide-binding motifs. For example, the highly conserved VPVPGGA motif in tagatose 6-phosphate kinase (group 13) might be involved in ATP binding (e.g. gi|1168382, residues 176–182). Furthermore, this motif is found in the loop between a predicted  $\beta$ -strand and  $\alpha$ -helix. Nucleotide-binding residues located on  $\beta\alpha$  loops are a theme in kinases. Conserved, positively charged amino acid residues, especially lysine and arginine, are common in binding and orienting of the ATP triphosphate tail.  $Mg^{2+}$  coordination is often accomplished with a conserved aspartate residue or with another conserved hydroxyl-containing group in the active site. Other conserved, charged amino acid residues could be potential catalytic or substrate-binding residues. Multiple alignments identifying potential active-site components from two of the kinase groupings with unknown structures (glycerate kinase and dihydroxyacetone kinase) are shown in Figure 7.

## Discussion

### Common structural mechanisms shared among kinases

Although all kinases catalyze a similar phosphoryl transfer reaction, they adopt a wide variety of structural folds. Our current survey cataloged 30 distinct sequence/structure families of kinases. Of these families, 19, covering 98% of over 17,000 sequences, were further grouped into seven general fold groups. This classification system combined with a wealth of kinase structural and biochemical data enables us to address the question of how these different structural folds accomplish the same chemistry. The common structural features that influence phosphoryl transfer by kinases have been reviewed.<sup>73</sup> Briefly, all phosphotransfer reactions contain the following three principal components: binding and orienting the phosphate donor (ATP); binding and orienting the phosphate acceptor (substrate); and catalysis of the chemical reaction.

### Nucleotide binding

Several distinct modes of nucleotide binding have emerged. One recurring theme is that the

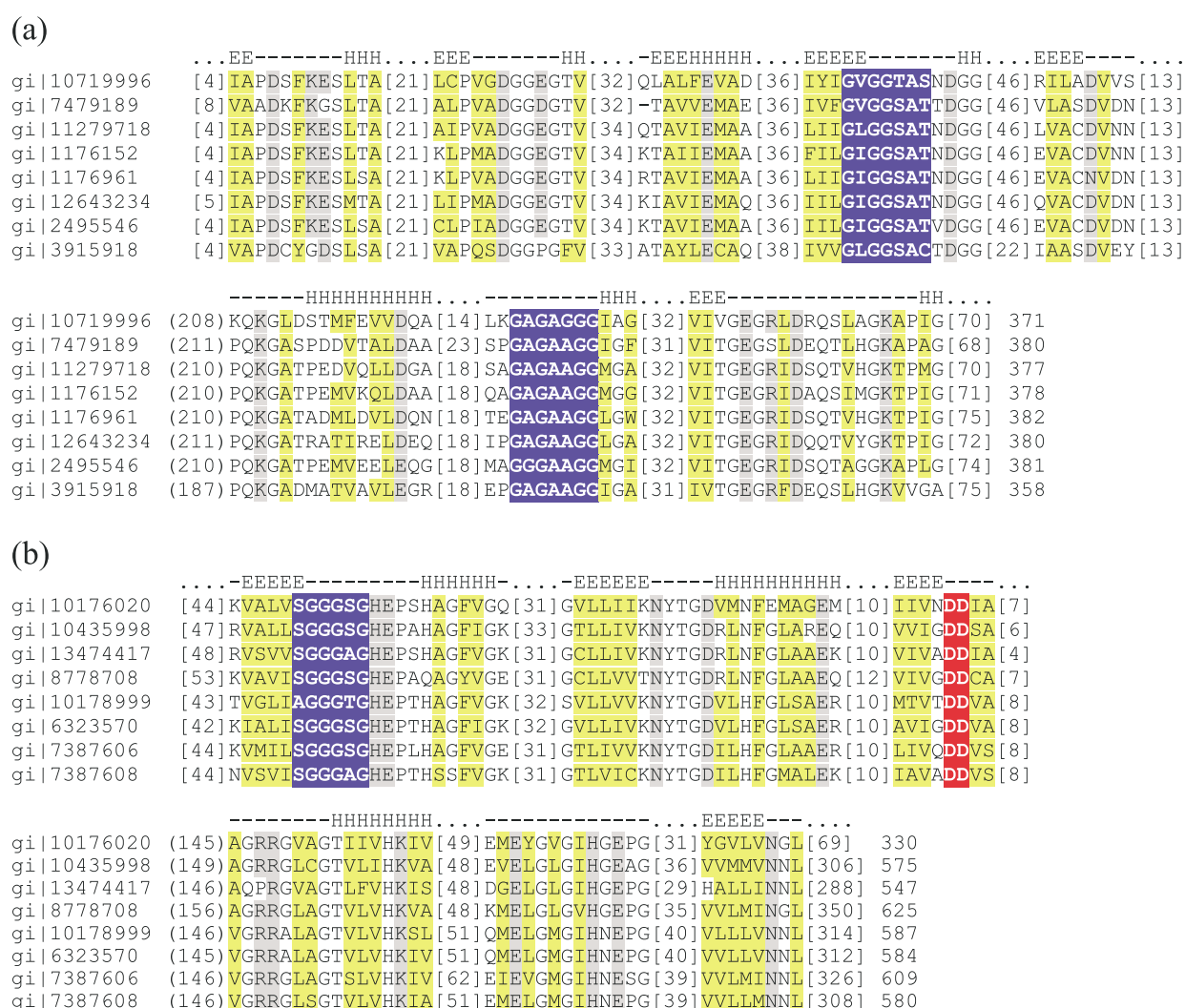
nucleotide binds at the C-terminal end of  $\beta$ -strands and N terminus of  $\alpha$ -helices. This is observed in all Rossmann-like fold families and in the GHMP family. In three of these families, P-loop kinase, PEPCK, and GHMP, the connecting loop between the  $\beta$ -strand and  $\alpha$ -helix are extended, forming the so-called phosphate-binding loop (P-loop) that wraps around the triphosphate tail of the bound nucleotide. The glycine-rich nature of these loops enables them to adopt conformations such that several main-chain amide groups are all pointed towards the bound nucleotide. Together with the positive dipole of the  $\alpha$ -helix and some positively charged lysine or arginine side-chains, a strong anion hole is created for the binding of the nucleotide.

The glycine-rich phosphate-binding loops are observed in families of the protein kinase-like fold and ribonuclease H-like fold groups. However, in these cases, the loops are mostly  $\beta$ -hairpin loops that connect two anti-parallel  $\beta$ -strands. No major contribution from helix dipoles is involved in nucleotide binding in these kinase families. However, several  $\beta$ -hairpin loops, such as in the case of hexokinase (ribonuclease H-like), may congregate at the active site with their main-chain amide groups interacting with the phosphate group. In the above two distinct nucleotide-binding modes, the protein main-chain interacting with the nucleotide is prominent. This is probably one of the reasons why, for example, in all Rossmann-fold kinase families, the ATP moiety binds at roughly the same place.

Nucleotide binding in the ferredoxin-like group and TIM-barrel group are completely different. No commonly shared local structural motif is observed across families. In general, kinases in these two fold groups use mainly positively charged protein side-chains to interact with the nucleotide phosphate groups. As discussed before, nucleotide binding differs greatly between the families of the ferredoxin-like group, probably as a result of a lack of interactions with protein backbones.

### Binding of phosphoryl acceptor substrates

Binding and orientation of the phosphate-accepting substrate in kinases depends on interactions between the substrate and strategically placed active-site residues. The details of such interactions are, of course, contingent upon the specific activity of the kinase in question. Extra structural motifs or domains in addition to the nucleotide-binding core are usually necessary for the recognition of the substrate. Since the size and structure of kinase substrates varies drastically from a small molecule of a few atoms to a whole protein, the substrate-binding motifs also vary significantly. One common phenomenon associated with the substrate binding is induced conformational changes. In some cases, such as pyruvate phosphate dikinase and phosphoglycerate kinase, drastic domain movements occur in order to bring the substrate to the active site.<sup>20,74</sup> While in other cases, such as



**Figure 7.** Amino acid sequence alignment of two kinase groupings with unknown structures. (a) Group 15: glycerate kinase. (b) Group 17: dihydroxyacetone kinase. The conserved pair of aspartate residues (in bold white letters and highlighted in red) following the  $\beta$ -strand is reminiscent of the Walker B motif and are potential  $Mg^{2+}$ -coordinating residues. The C-terminal end of the group 17 proteins is predicted to be composed almost entirely of  $\alpha$ -helices and is therefore unlikely to be the catalytic kinase domain. In these alignments, highly conserved charged residues (highlighted in gray) are potential components in substrate binding or catalysis. Conserved motifs of small residues (potential nucleotide-binding regions) are highlighted in blue and are in bold white letters. Numbers in square brackets specify the number of residues in an insert that are not shown, and numbers in parentheses indicate the position of the first residue on that line.

protein kinase and GHMP kinase family members, local conformational change such as closing the loop conformation is sufficient to sequester the substrate and ensure the transfer of the phosphoryl group from ATP to the substrate.

### Metal binding and catalysis

Almost all kinases require a divalent metal cation in order to function.  $Mg^{2+}$  usually activates ATP for catalysis by weakening the bond between the  $\beta$  and  $\gamma$ -phosphate groups and assists in properly orientating the  $\gamma$ -phosphate group for phosphotransfer. Metal cations are usually coordinated by a conserved glutamate, aspartate, or other hydroxyl-containing residue in the active

site. In some cases, however, the magnesium cations are positioned by coordinated water molecules and have no direct liganding to the enzyme. Several kinases utilize additional metal cofactors such as a secondary magnesium cation, a manganese cation, or a potassium cation.<sup>30,63,75</sup>

There are three main catalytic mechanisms that are widely used by kinases: transition state stabilization; acid-base catalysis; and ping-pong (or double displacement) mechanisms. In general, there is little correlation between the structural fold group and the mechanism of catalysis utilized. One example of different mechanisms in the same fold group includes the ATP-grasp family and the protein kinase family of group 1. The mechanism of pyruvate phosphate dikinase of the ATP-grasp

family involves the reversible phosphorylation of a histidine residue.<sup>18,19</sup> Although the mechanism utilized by protein kinases is currently a matter of debate, the phosphotransfer reaction in this family is thought to proceed either *via* a proposed simultaneous transfer mechanism,<sup>15</sup> which falls into the category of transition-state stabilization, or *via* acid–base catalysis. A second example includes the families of the ferredoxin-like fold group. Nucleoside diphosphate kinase (NDP) and histidine kinase both utilize ping-pong mechanisms. Although the possibility of acid–base catalysis cannot be eliminated in arginine kinase of the guanido kinase family, the primary factor in the mechanism of this enzyme appears to be transition-state stabilization and precise substrate orientation in the active site.<sup>50</sup>

Within most kinase families, all enzymes usually utilize the same type of catalytic mechanism. All protein kinases, for example, are thought to have similarly catalyzed reactions. However, there are exceptions to this tendency. In the GHMP family, for example, the homoserine kinase mechanism is proposed to proceed *via* a simultaneous transfer reaction (transition-state stabilization),<sup>67</sup> while the archaeal shikimate kinase reaction may follow a ping-pong mechanism.<sup>76</sup> In a second example, reactions catalyzed by P-loop family kinases may proceed *via* different mechanisms. Acid–base catalysis is the probable mechanism of P-loop enzymes such as phosphoribulokinase and shikimate kinase for the production of phosphorylated metabolites,<sup>27</sup> while UMP/CMP kinase of the same family has been suggested to utilize a simultaneous transfer mechanism similar to that proposed for protein kinases (transition-state stabilization).<sup>26</sup>

### Same activity, different fold

There are several cases of the same kinase activity that exists in unrelated fold families, reflecting convergent evolution of the same function from different ancestors. For example, Galperin *et al.* have identified analogous enzymes in fructokinase, 6-phosphofructokinase, and gluconokinase.<sup>77</sup> Comparisons of homoserine kinase, phosphomevalonate kinase, shikimate kinase, glucokinase, 1-phosphofructokinase, uridylate kinase, and pantothenate kinase from different folds are summarized below.

Homoserine kinase is thus far the only activity to be found in three distinct fold groups (the protein kinase family, the HAD-like family, and the GHMP family). The protein kinase family contains homoserine kinases from several proteobacterial species (eubacteria). The GHMP kinase family includes homoserine kinases from the majority of eubacteria, from archaeobacteria, and from eukaryotes. The HAD-like family currently contains only one homoserine kinase isozyme, which is the bifunctional ThrH from *Pseudomonas aeruginosa*.<sup>38</sup> It is interesting to note that the same mechanism of cat-

alysis has been proposed for both the protein kinase family and the GHMP family (the “synchronous shift” mechanism). The existence of a HAD-like homoserine kinase/phosphoserine phosphatase, however, demonstrates that nature has accomplished this same activity *via* a completely different mechanism. Catalysis in the HAD-like homoserine kinase isozyme is expected to proceed *via* a phosphoaspartate intermediate similar to the mechanism employed by the other members of the HAD-like family. Another surprising feature is homoserine kinase’s use of the protein kinase fold, which is generally utilized for the accommodation of very large substrates.

Phosphomevalonate kinase (PMK) and shikimate kinase (SK) are each found in both the P-loop kinase family and the GHMP kinase family.<sup>22,76</sup> PMK from higher eukaryotes (including human, pig, and fruit fly) belong to the P-loop kinase family. All other identified PMK sequences (such as those of *Saccharomyces cerevisiae*, *Staphylococcus aureus*, and *Streptococcus pneumoniae*) are found in the GHMP kinase family. A structure has not been solved for PMK from either fold family. SK also contributes members to each of these two families. The P-loop kinase family includes primarily eubacterial SK, while the GHMP kinase family contains archaeobacterial SK.<sup>76</sup> Currently, the only SK structure that has been solved is that of *Erwinia chryanthemi* (eubacteria), which belongs to the P-loop kinase family.<sup>78</sup>

Glucokinase is found in both the ribokinase-like family of the Rossmann-like fold group and the ribonuclease H-like fold group. Glucokinase from archaeal species such as *Thermococcus litoralis* and *Pyrococcus furiosus* belong to the ribokinase-like family. The ribonuclease H-like family contains glucokinases from eukaryote, eubacterial, and a few archaeobacterial species. Unlike the ATP-dependent ribonuclease H-like glucokinases, the archaeal glucokinases of the ribokinase-like family are ADP-dependent. The modes of nucleotide binding differ between these two families.

1-Phosphofructokinase can be found in both the phosphofructokinase-like and ribokinase-like families. The sequences found in the ribokinase-like family are from bacterial species, while only human sequences with this activity are found in the phosphofructokinase-like family. Both of these families belong to the Rossmann-like group, and both families are believed to utilize acid–base catalysis in their reactions. Thus, the core of the folds and the mechanisms of phosphotransfer are expected to be similar for 1-phosphofructokinases from each family. However, because the nucleotide-binding patterns are somewhat different between these two families, the precise location and orientation of the bound ATP will most likely differ between human and bacterial 1-phosphofructokinase.

Uridylate kinase is found in two different Rossmann-like families. Uridylate kinase from *Leishmania major* and *S. cerevisiae* belong to the

P-loop kinase family. The structure of yeast uridylate kinase has been solved.<sup>79</sup> Uridylate kinase sequences in the aspartokinase family are predominantly from bacterial and archaeobacterial species. Again, since both families are in the Rossmann-like group, the core of the uridylate kinase structures will be similar, while specificities of nucleotide binding will differ between the uridylate kinase representatives in the two families.

Pantothenate kinase should provide another interesting example. Prokaryotic pantothenate kinase is known to belong to the P-loop kinase family.<sup>71</sup> However, eukaryotic pantothenate kinase, for which there is currently no solved structure, is not similar to its prokaryotic counterpart in either primary sequence<sup>72</sup> or in predictions of secondary structure elements. Thus, the prokaryotic and eukaryotic versions of this enzyme are likely to differ in fold, in mode of nucleotide binding, and in catalytic mechanism.

The examples above describe cases in which nature has developed the same activity in multiple ways. The opposite situation, in which the same structural fold is used for many different substrate specificities, is readily observable as well. Examples of families in which one structural fold accounts for kinase activity on many different substrates include the P-loop kinase family, the ribokinase-like family, and the ribonuclease H-like family.

### Correlation of structural fold with placement in cellular pathway

One of the few generalities that can be made in the correlation between structural fold and cellular pathway is that the protein kinase fold is dedicated predominantly to cellular signaling. The vast majority of the members of the protein kinase family participate in signal transduction. Furthermore, the Rossmann-like fold in kinases is apparently utilized exclusively in metabolic pathways. However, the types of metabolic pathways that the kinases participate in vary between each of the families of the Rossmann-like group. For example, most kinases in the ribokinase-like family are involved in carbohydrate metabolism, although a few do participate in the metabolism of nucleotides or vitamins and cofactors. The aspartokinase family, however, has many members that participate in amino acid metabolism, in addition to a few that are involved in energy metabolism or nucleotide metabolism. P-loop kinases represent the Rossmann-like family whose members participate in the widest variety of metabolic pathway types. While the largest fraction of P-loop kinase family members participate in nucleotide metabolism, a substantial number function in the metabolism of lipids, carbohydrates, amino acid residues, and multiple other types of molecules. As a whole, group 2 (Rossmann-like) kinases are involved in the entire scope of metabolic pathway

types, including carbohydrate, lipid, amino acid, nucleotide, cofactor, vitamin, and energy metabolism.

Kinases of the ferredoxin-like fold group have evident partialities in terms of pathway type. Members of the guanido kinases family function solely in amino acid metabolism, and histidine kinases are signaling enzymes. The other two families in this group each contain only one kinase member: nucleoside-diphosphate kinase functions in nucleotide metabolism while HPPK participates in vitamin metabolism pathways.

Although GHMP kinases participate in several different metabolic pathways, such as carbohydrate, amino acid, and lipid metabolism, their role in the isoprenoid biosynthesis pathways is particularly prominent. Notably, members of the GHMP kinase superfamily (mevalonate kinase, phosphomevalonate kinase, and mevalonate pyrophosphate decarboxylase) catalyze three consecutive steps in the early mevalonate pathway. One other GHMP kinase, 4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase, participates in the recently characterized non-mevalonate isoprenoid biosynthesis pathway.<sup>80</sup> The essentiality of these enzymes has identified them as potential antibacterial drug targets.<sup>81–83</sup>

Trends in other groups are less evident. Members of the ribonuclease H-like group participate mainly in carbohydrate metabolism, although a significant number of kinases are involved in other types of metabolic pathways as well. Notably, ATP-grasp and TIM  $\beta/\alpha$  barrel, two of the most widespread protein folds, are adopted for only four and one kinase activities, respectively. However, each family includes participants in three different types of metabolic pathways (lipid, carbohydrate, and energy metabolism for ATP-grasp and nucleotide, carbohydrate, and energy metabolism for TIM  $\beta/\alpha$ -barrel kinases). This may reflect the overall functional diversity of both folds.

### Distribution of kinases across genomes and fold families

The distribution of kinases from the genomes of several representative species is shown in Table 2. The protein kinase (PK) family of group 1 is by far the largest family in terms of number of sequences from the non-redundant database. The PK family contains over half of all the kinase sequences in our analysis. In the selected representatives of eukaryotic species (*Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *S. cerevisiae*), sequences belonging to the PK family account for between 65% and 85% of all identified kinases in these genomes. In prokaryotes, however, the two-component system (histidine kinase) is the predominant method of signal transduction. This is reflected by the small number of proteins in the PK family from the genomes of *E. coli* and *Methanococcus jannaschii*. The non-signaling kinase groups, which include all families except the

**Table 2.** Distribution of kinases in genomes of various organisms

	Number of kinase sequences		All kinases	% All identified proteins in genome
	Protein kinase family	Non-signaling kinase families		
<i>Homo sapiens</i>	391	100	492	1.5
<i>Drosophila melanogaster</i>	273	91	365	2.5
<i>Caenorhabditis elegans</i>	473	76	549	2.7
<i>Saccharomyces cerevisiae</i>	126	59	187	3.0
<i>Escherichia coli</i>	1	76	107	2.5
<i>Methanococcus jannaschii</i>	2	29	31	1.8

The non-signaling kinase families include all families except the protein kinase family and the histidine kinase family.

protein kinase family and the histidine kinase family, contain predominantly metabolic kinases. For these families, the distribution of kinase sequences is approximately equal for five of the six representative species (Table 2). Although the percentage of metabolic kinases differs between the organisms, the actual numbers of these enzymes in each of the representative genomes are rather similar, ranging from 59 in *S. cerevisiae* to 100 in human. Notably, there are only 31 identified kinase genes in *M. jannaschii*. This may be due to the smaller genome size of *M. jannaschii*, as well as unique aspects of archaeal metabolisms. It needs to be pointed out that the number of kinases listed here is far from complete. There are still many “hypothetical” or non-annotated genes in each of the complete genomes, and genes coding for ~70 known kinase activities have not been identified in any organisms.

In conclusion, we have, for the first time, performed a comprehensive survey of all available kinase sequences, regarding their evolutionary relationships, structural fold diversity, and specific activities. Our sequence and structural classification of all kinase groups should be beneficial to both experimental and computational biologists. With this classification system, the function and mechanism of poorly studied or newly discovered kinases may be deduced from proteins in the same family. The common enzymatic and structural mechanisms of kinases may be generalized for different families or fold groups based on the framework of the current classification system. The insights gained from this analysis will further our understanding of protein structure–function relationships in general and the evolution of various kinases in particular.

## Methods

### Initial groupings of all kinase sequences

A list of all Pfam profiles<sup>5</sup> from version 5.4 and COGs from version 2<sup>84</sup> that describe catalytic kinase domains was constructed. For cases in which a COG's contents were contained completely within a Pfam profile, the COG was removed from the list to avoid redundancy. The reduced list contained 44 Pfam profiles and 12 COG sequence sets. One COG from version 3<sup>85</sup> was later

added, forming a total of 57 kinase profiles (44 from Pfam and 13 from COG). The hmmbuild and hmmcalibrate programs of the HMMER2 package<sup>2</sup> were used to construct profiles for the 13 COG sequence sets. The hmmsearch program of the HMMER2 package<sup>2</sup> was then used to assign sequences from the non-redundant (NR) protein database at the National Center for Biotechnology Information (NIH, Bethesda, MD) to the kinase profiles (*E*-value cutoff 0.1). Additionally, the GREFD program of the SEALS package<sup>86</sup> was used to extract all sequences for which the definition line contained the pattern “kinase”. Three iterations of PSI-BLAST<sup>1</sup> were run for each kinase sequence that was not already assigned to a profile. Any of these sequences that produced hits (*E*-value cutoff 0.001) to sequences already assigned were subsequently placed in those profiles. The remaining unassigned kinase sequences (sequences with the word kinase in the definition line that were not placed in the Pfam/COG kinase profiles) were then filtered manually to remove fragments, non-kinase entries (e.g. kinase inhibitors), and non-catalytic entries (e.g. regulatory subunits). Thus, after manual filtering any remaining entries are considered “true” catalytic kinase sequences that cannot be assigned to existing kinase profiles by automatic methods with the criteria described above. These remaining sequences were clustered by sequence similarity using the GROUPE program (score cutoff 50, single linkage) of the SEALS package.<sup>86</sup> Multiple sequence alignments and secondary structure predictions were performed on these groups. Further sequence similarity searches with PSI-BLAST<sup>1</sup> were carried out with somewhat relaxed thresholds and their results were inspected manually. As a result, some of these initially unassigned groupings could be merged into existing profiles based on the presence of conserved catalytic residues and other distinguishing motifs, as well as matching secondary structure distributions (see examples in Results), while others were placed into novel groupings.

### Homologous sequence family classification

After all the kinase sequences had been assigned to Pfam/COG profiles or to novel groupings, PSI-BLAST<sup>1</sup> was used to detect possible evolutionary links between these Pfam/COG profiles. Sequences from different Pfam/COG profiles with statistically significant similarities were identified and assembled into families. In other words, sequences from each Pfam/COG profile in the same family usually produce significant PSI-BLAST<sup>1</sup> hits to each other. Therefore, homology is inferred to all sequences in the same family. In most cases, finding these links was trivial. A trivial link is one that is

established by three iterations of PSI-BLAST<sup>1</sup> with *E*-value cutoff 0.001. For orphan Pfam/COG profiles or sequence groupings, multiple alignments were constructed in order to reveal conserved active-site motifs. Secondary structure predictions with Jpred<sup>70</sup> and manual inspection of PSI-BLAST<sup>1</sup> search results were performed. In some cases, these orphan groupings can be placed into existing families with confidence (see Results). Others were assigned as novel kinase families.

For the purposes of this study, we chose to examine those enzymes in EC2.7.1.- (phosphotransferases with an alcohol group as acceptor), EC2.7.2.- (phosphotransferases with a carboxyl group as acceptor), EC2.7.3.- (phosphotransferases with a nitrogenous group as acceptor), and EC2.7.4.- (phosphotransferases with a phosphate group as acceptor) of the EC system. Once the groupings had been made on the basis of the profile and sequence similarity searches, a handful of other activities that are not a part of this range fell neatly into the pre-existing groups. These enzymes were added to our analysis. These added activities include 4.1.1.32 (phosphoenolpyruvate carboxykinase-GTP), 4.1.1.49 (phosphoenolpyruvate carboxykinase-ATP), 2.7.9.1 (pyruvate, phosphate dikinase), 2.7.9.2 (pyruvate, water dikinase), 2.7.9.3 (selenide, water dikinase), 2.7.6.2 (thiamin pyrophosphokinase), and 2.7.6.3 (7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase). It should be noted that many of the activities between 2.7.1.- and 2.7.4.- do not have identified sequences and therefore could not be included in the groupings. Kinase activities utilizing different phosphate donors, such as 2.7.2.10 (phosphoglycerate kinase (GTP)), were included in our classification if they were found to belong to a pre-existing kinase group. Kinases that do not hydrolyze ATP and do not belong to existing kinase groups were excluded intentionally. These excluded kinase activities include 2.7.1.69 (protein-N(P)-phosphohistidine-sugar phosphotransferase) and 2.7.3.9 (phosphoenolpyruvate-protein phosphatase).

### Fold group classification

A total of 30 kinase families, containing as few as two sequences and as many as 9600 sequences, were formed. Of these kinase families, 19 contained at least one member with a solved structure. These 19 families could be assembled into seven fold groups on the basis of similarities of structural fold. Families in the same fold group share structurally similar nucleotide-binding domains that are of the same architecture and topology (or related by circular permutation) for at least the core of the domain. The remaining ten groups are composed of the 11 families that contain no members with solved structures.

### Distribution of kinase sequences in the completed genomes

The distribution of kinase sequences in the completed genomes of several representative species was investigated (Table 2). For each representative genome, the number of kinase sequences in each of the families was determined. The hmsearch program (HMMER2<sup>2</sup>) was used to assign sequences from the selected genomes to the 57 kinase profiles (*E*-value cutoff 0.1). In each family, any assigned sequences for which the word kinase was not found in the definition line were identified. BLAST<sup>1</sup> was run for each of these potential non-kinase sequences

in order to identify homologs. Any sequences for which BLAST results did not indicate that the protein was a kinase were removed from the profiles. Additionally, the GREFD program (SEALS<sup>86</sup>) was used to extract all entries with the word kinase in the definition line, and PSI-BLAST<sup>1</sup> was used to place any unassigned kinase sequences into the profiles.

The genomes of *H. sapiens*<sup>†</sup>, *S. cerevisiae* (NC\_001133, NC\_001134, NC\_001135, NC\_001136, NC\_001137, NC\_001138, NC\_001139, NC\_001140, NC\_001141, NC\_001142, NC\_001143, NC\_001144, NC\_001145, NC\_001146, NC\_001147, and NC\_001148), *E. coli* (NC\_000913), and *M. jannaschii* (NC\_000909) were obtained from the NCBI (NIH, Bethesda, MD) site.

The *D. melanogaster* genome (release 2) was downloaded from the Berkeley Drosophila Genome Project site<sup>‡</sup>.

The *C. elegans* genome (wormpep70) was downloaded from the Sanger Institute site<sup>§</sup>.

Multiple alignments were generated using the T-COFFEE program<sup>87</sup> and adjusted manually. Ribbon diagrams of the proteins were drawn using the program MOLSCRIPT<sup>88</sup>.

### Supplementary Material

Supplementary Material is available by anonymous ftp at <ftp://iole.swmed.edu/pub/kinase/>

## Acknowledgments

This work was supported, in part, by National Institutes of Health grant GM63689 (to H.Z.) and the Welch Foundation grant I-1505 (to N.V.G.).

## References

1. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
2. Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
3. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
4. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
5. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucl. Acids Res.* **28**, 263–266.

<sup>†</sup> Downloaded from [ftp://ftp.ncbi.nih.gov/refseq/H\\_sapiens/H\\_sapiens/protein/](ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/H_sapiens/protein/)

<sup>‡</sup> <http://www.fruitfly.org/>

<sup>§</sup> [http://www.sanger.ac.uk/Projects/C\\_elegans/wormpep/](http://www.sanger.ac.uk/Projects/C_elegans/wormpep/)

6. Grishin, N. V. (1999). Phosphatidylinositol phosphate kinase: a link between protein kinase and glutathione synthase folds. *J. Mol. Biol.* **291**, 239–247.
7. Leonard, C. J., Aravind, L. & Koonin, E. V. (1998). Novel families of putative protein kinases in bacteria and archaea: evolution of the “eukaryotic” protein kinase superfamily. *Genome Res.* **8**, 1038–1047.
8. Goldsmith, E. J. & Cobb, M. H. (1994). Protein kinases. *Curr. Opin. Struct. Biol.* **4**, 833–840.
9. Hanks, S. K. & Hunter, T. (1995). Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J.* **9**, 576–596.
10. Oudot, C., Cortay, J. C., Blanchet, C., Laporte, D. C., Di Pietro, A., Cozzzone, A. J. & Jault, J. M. (2001). The catalytic triad of isocitrate dehydrogenase kinase/phosphatase from *E. coli* and its relationship with that found in eukaryotic protein kinases. *Biochemistry*, **40**, 3047–3055.
11. Burk, D. L., Hon, W. C., Leung, A. K. & Berghuis, A. M. (2001). Structural analyses of nucleotide binding to an aminoglycoside phosphotransferase. *Biochemistry*, **40**, 8756–8764.
12. Steinbacher, S., Hof, P., Eichinger, L., Schleicher, M., Gettemans, J., Vandekerckhove, J. *et al.* (1999). The crystal structure of the *Physarum polycephalum* actin-fragmin kinase: an atypical protein kinase with a specialized substrate-binding domain. *EMBO J.* **18**, 2923–2929.
13. Yamaguchi, H., Matsushita, M., Nairn, A. C. & Kuriyan, J. (2001). Crystal structure of the atypical protein kinase domain of a TRP channel with phosphotransferase activity. *Mol. Cell*, **7**, 1047–1057.
14. Ryazanov, A. G., Ward, M. D., Mendola, C. E., Pavur, K. S., Dorovkov, M. V., Wiedmann, M. *et al.* (1997). Identification of a new class of protein kinases represented by eukaryotic elongation factor-2 kinase. *Proc. Natl Acad. Sci. USA*, **94**, 4884–4889.
15. Hart, J. C., Hillier, I. H., Burton, N. A. & Sheppard, D. W. (1998). An alternative role for the conserved Asp residue in phosphoryl transfer reactions. *J. Am. Chem. Soc.* **120**, 13535–13536.
16. Rao, V. D., Misra, S., Boronenkov, I. V., Anderson, R. A. & Hurley, J. H. (1998). Structure of type II $\beta$  phosphatidylinositol phosphate kinase: a protein kinase fold flattened for interfacial phosphorylation. *Cell*, **94**, 829–839.
17. Murzin, A. G. (1996). Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386–394.
18. Spronk, A. M., Yoshida, H. & Wood, H. G. (1976). Isolation of 3-phosphohistidine from phosphorylated pyruvate, phosphate dikinase. *Proc. Natl Acad. Sci. USA*, **73**, 4415–4419.
19. Goss, N. H., Evans, C. T. & Wood, H. G. (1980). Pyruvate phosphate dikinase: sequence of the histidyl peptide, the pyrophosphoryl and phosphoryl carrier. *Biochemistry*, **19**, 5805–5809.
20. Herzberg, O., Chen, C. C., Kapadia, G., McGuire, M., Carroll, L. J., Noh, S. J. & Dunaway-Mariano, D. (1996). Swiveling-domain mechanism for enzymatic phosphotransfer between remote reaction sites. *Proc. Natl Acad. Sci. USA*, **93**, 2652–2657.
21. Rossmann, M. G., Moras, D. & Olsen, K. W. (1974). Chemical and biological evolution of nucleotide-binding protein. *Nature*, **250**, 194–199.
22. Smit, A. & Mushegian, A. (2000). Biosynthesis of isoprenoids via mevalonate in Archaea: the lost pathway. *Genome Res.* **10**, 1468–1484.
23. Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. (1982). Distantly related sequences in the  $\alpha$ - and  $\beta$ -subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1**, 945–951.
24. Saraste, M., Sibbald, P. R. & Wittinghofer, A. (1990). The P-loop—a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **15**, 430–434.
25. Sheng, X. R., Li, X. & Pan, X. M. (1999). An iso-random Bi Bi mechanism for adenylate kinase. *J. Biol. Chem.* **274**, 22238–22242.
26. Hutter, M. C. & Helms, V. (2000). Phosphoryl transfer by a concerted reaction mechanism in UMP/CMP-kinase. *Protein Sci.* **9**, 2225–2231.
27. Miziorko, H. M. (2000). Phosphoribulokinase: current perspectives on the structure/function basis for regulation and catalysis. *Advan. Enzymol. Relat. Areas Mol. Biol.* **74**, 95–127.
28. Matte, A., Goldie, H., Sweet, R. M. & Delbaere, L. T. (1996). Crystal structure of *Escherichia coli* phosphoenolpyruvate carboxykinase: a new structural family with the P-loop nucleoside triphosphate hydrolase fold. *J. Mol. Biol.* **256**, 126–143.
29. Fieulaine, S., Morera, S., Poncet, S., Monedero, V., Gueguen-Chaignon, V., Galinier, A. *et al.* (2001). X-ray structure of HPr kinase: a bacterial protein kinase with a P-loop nucleotide-binding domain. *EMBO J.* **20**, 3917–3927.
30. Tari, L. W., Matte, A., Goldie, H. & Delbaere, L. T. (1997).  $Mg^{2+}$ – $Mn^{2+}$  clusters in enzyme-catalyzed phosphoryl-transfer reactions. *Nature Struct. Biol.* **4**, 990–994.
31. Watson, H. C., Walker, N. P., Shaw, P. J., Bryant, T. N., Wendell, P. L., Fothergill, L. A. *et al.* (1982). Sequence and structure of yeast phosphoglycerate kinase. *EMBO J.* **1**, 1635–1640.
32. Bernstein, B. E. & Hol, W. G. (1998). Crystal structures of substrates and products bound to the phosphoglycerate kinase active site reveal the catalytic mechanism. *Biochemistry*, **37**, 4429–4436.
33. Marina, A., Alzari, P. M., Bravo, J., Uriarte, M., Barcelona, B., Fita, I. & Rubio, V. (1999). Carbamate kinase: new structural machinery for making carbamoyl phosphate, the common precursor of pyrimidines and arginine. *Protein Sci.* **8**, 934–940.
34. Ramon-Maiques, S., Marina, A., Uriarte, M., Fita, I. & Rubio, V. (2000). The 1.5 Å resolution crystal structure of the carbamate kinase-like carbamoyl phosphate synthetase from the hyperthermophilic Archaeon *Pyrococcus furiosus*, bound to ADP, confirms that this thermostable enzyme is a carbamate kinase, and provides insight into substrate binding and stability in carbamate kinases. *J. Mol. Biol.* **299**, 463–476.
35. Shirakihara, Y. & Evans, P. R. (1988). Crystal structure of the complex of phosphofructokinase from *Escherichia coli* with its reaction products. *J. Mol. Biol.* **204**, 973–994.
36. Ito, S., Fushinobu, S., Yoshioka, I., Koga, S., Matsuzawa, H. & Wakagi, T. (2001). Structural basis for the ADP-specificity of a novel glucokinase from a hyperthermophilic archaeon. *Structure*, **9**, 205–214.
37. Sigrell, J. A., Cameron, A. D., Jones, T. A. & Mowbray, S. L. (1998). Structure of *Escherichia coli* ribokinase in complex with ribose and dinucleotide determined to 1.8 Å resolution: insights into a new family of kinase structures. *Structure*, **6**, 183–193.
38. Patte, J. C., Clepet, C., Bally, M., Borne, F., Mejean, V. & Foglino, M. (1999). ThrH, a homoserine kinase

- isozyme with *in vivo* phosphoserine phosphatase activity in *Pseudomonas aeruginosa*. *Microbiology*, **145**, 845–853.
39. Cho, H., Wang, W., Kim, R., Yokota, H., Damo, S., Kim, S. H. *et al.* (2001).  $\text{BeF}_3^-$  acts as a phosphate analog in proteins phosphorylated on aspartate: structure of a  $\text{BeF}_3^-$  complex with phosphoserine phosphatase. *Proc. Natl Acad. Sci. USA*, **98**, 8525–8530.
  40. Wang, W., Kim, R., Jancarik, J., Yokota, H. & Kim, S. H. (2001). Crystal structure of phosphoserine phosphatase from *Methanococcus jannaschii*, a hyperthermophile, at 1.8 Å resolution. *Structure*, **9**, 65–71.
  41. Hisano, T., Hata, Y., Fujii, T., Liu, J. Q., Kurihara, T., Esaki, N. & Soda, K. (1996). Crystal structure of L-2-haloacid dehalogenase from *Pseudomonas* sp. YL. An  $\alpha/\beta$  hydrolase structure that is different from the  $\alpha/\beta$  hydrolase fold. *J. Biol. Chem.* **271**, 20322–20330.
  42. Collet, J. F., Stroobant, V. & Van Schaftingen, E. (1999). Mechanistic studies of phosphoserine phosphatase, an enzyme related to P-type ATPases. *J. Biol. Chem.* **274**, 33985–33990.
  43. Morais, M. C., Zhang, W., Baker, A. S., Zhang, G., Dunaway-Mariano, D. & Allen, K. N. (2000). The crystal structure of *Bacillus cereus* phosphonoacetaldehyde hydrolase: insight into catalysis of phosphorus bond cleavage and catalytic diversification within the HAD enzyme superfamily. *Biochemistry*, **39**, 10385–10396.
  44. Baker, L. J., Dorocke, J. A., Harris, R. A. & Timm, D. E. (2001). The crystal structure of yeast thiamin pyrophosphokinase. *Structure*, **9**, 539–546.
  45. Janin, J., Dumas, C., Morera, S., Xu, Y., Meyer, P., Chiadmi, M. & Cherfils, J. (2000). Three-dimensional structure of nucleoside diphosphate kinase. *J. Bioenerg. Biomembr.* **32**, 215–225.
  46. Morera, S., Lascu, I., Dumas, C., LeBras, G., Briozzo, P., Veron, M. & Janin, J. (1994). Adenosine 5'-diphosphate binding and the active site of nucleoside diphosphate kinase. *Biochemistry*, **33**, 459–467.
  47. Blaszczyk, J., Shi, G., Yan, H. & Ji, X. (2000). Catalytic center assembly of HPPK as revealed by the crystal structure of a ternary complex at 1.25 Å resolution. *Structure*, **8**, 1049–1058.
  48. Stammers, D. K., Achari, A., Somers, D. O., Bryant, P. K., Rosemond, J., Scott, D. L. & Champness, J. N. (1999). 2.0 Å X-ray structure of the ternary complex of 7,8-dihydro-6-hydroxymethylpterinpyrophosphokinase from *Escherichia coli* with ATP and a substrate analogue. *FEBS Letters*, **456**, 49–53.
  49. Fritz-Wolf, K., Schnyder, T., Wallimann, T. & Kabsch, W. (1996). Structure of mitochondrial creatine kinase. *Nature*, **381**, 341–345.
  50. Zhou, G., Somasundaram, T., Blanc, E., Parthasarathy, G., Ellington, W. R. & Chapman, M. S. (1998). Transition state structure of arginine kinase: implications for catalysis of bimolecular reactions. *Proc. Natl Acad. Sci. USA*, **95**, 8449–8454.
  51. Bilwes, A. M., Quezada, C. M., Croal, L. R., Crane, B. R. & Simon, M. I. (2001). Nucleotide binding by the histidine kinase CheA. *Nature Struct. Biol.* **8**, 353–360.
  52. Tanaka, T., Saha, S. K., Tomomori, C., Ishima, R., Liu, D., Tong, K. I., Park, H. *et al.* (1998). NMR structure of the histidine kinase domain of the *E. coli* osmosensor EnvZ. *Nature*, **396**, 88–92.
  53. Dutta, R., Qin, L. & Inouye, M. (1999). Histidine kinases: diversity of domain organization. *Mol. Microbiol.* **34**, 633–640.
  54. Robinson, V. L., Buckler, D. R. & Stock, A. M. (2000). A tale of two components: a novel kinase and a regulatory switch. *Nature Struct. Biol.* **7**, 626–633.
  55. Buss, K. A., Ingram-Smith, C., Ferry, J. G., Sanders, D. A. & Hasson, M. S. (1997). Crystallization of acetate kinase from *Methanosarcina thermophila* and prediction of its fold. *Protein Sci.* **6**, 2659–2662.
  56. Buss, K. A., Cooper, D. R., Ingram-Smith, C., Ferry, J. G., Sanders, D. A. & Hasson, M. S. (2001). Urkinase: structure of acetate kinase, a member of the ASKHA superfamily of phosphotransferases. *J. Bacteriol.* **183**, 680–686.
  57. Bork, P., Sander, C. & Valencia, A. (1992). An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proc. Natl Acad. Sci. USA*, **89**, 7290–7294.
  58. Aleshin, A. E., Kirby, C., Liu, X., Bourenkov, G. P., Bartunik, H. D., Fromm, H. J. & Honzatko, R. B. (2000). Crystal structures of mutant monomeric hexokinase I reveal multiple ADP binding sites and conformational changes relevant to allosteric regulation. *J. Mol. Biol.* **296**, 1001–1015.
  59. Arora, K. K., Filburn, C. R. & Pedersen, P. L. (1991). Glucose phosphorylation. Site-directed mutations which impair the catalytic function of hexokinase. *J. Biol. Chem.* **266**, 5359–5362.
  60. Wierenga, R. K. (2001). The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Letters*, **492**, 193–198.
  61. Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164.
  62. Larsen, T. M., Laughlin, L. T., Holden, H. M., Rayment, I. & Reed, G. H. (1994). Structure of rabbit muscle pyruvate kinase complexed with  $\text{Mn}^{2+}$ ,  $\text{K}^+$ , and pyruvate. *Biochemistry*, **33**, 6301–6309.
  63. Larsen, T. M., Benning, M. M., Rayment, I. & Reed, G. H. (1998). Structure of the bis( $\text{Mg}^{2+}$ )-ATP-oxalate complex of the rabbit muscle pyruvate kinase at 2.1 Å resolution: ATP binding over a barrel. *Biochemistry*, **37**, 6247–6255.
  64. Bork, P., Sander, C. & Valencia, A. (1993). Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Sci.* **2**, 31–40.
  65. Zhou, T., Daugherty, M., Grishin, N. V., Osterman, A. L. & Zhang, H. (2000). Structure and mechanism of homoserine kinase: prototype for the GHMP kinase superfamily. *Structure*, **8**, 1247–1257.
  66. Yang, D., Shipman, L. W., Roessner, C. A., Scott, A. I. & Sacchettini, J. C. (2002). Structure of the *Methanococcus jannaschii* mevalonate kinase—a member of the GHMP kinase superfamily. *J. Biol. Chem.* **277**, 9462–9467.
  67. Krishna, S. S., Zhou, T., Daugherty, M., Osterman, A. & Zhang, H. (2001). Structural basis for the catalysis and substrate specificity of homoserine kinase. *Biochemistry*, **40**, 10810–10818.
  68. Li, C., Kappock, T. J., Stubbe, J., Weaver, T. M. & Ealick, S. E. (1999). X-ray crystal structure of aminoimidazole ribonucleotide synthetase (PurM), from the *Escherichia coli* purine biosynthetic pathway at 2.5 Å resolution. *Structure*, **7**, 1155–1166.

69. Mueller, E. J., Oh, S., Kavalierchik, E., Kappock, T. J., Meyer, E., Li, C., Ealick, S. E. & Stubbe, J. (1999). Investigation of the ATP binding site of *Escherichia coli* aminoimidazole ribonucleotide synthetase using affinity labeling and site-directed mutagenesis. *Biochemistry*, **38**, 9831–9839.
70. Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. & Barton, G. J. (1998). JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.
71. Yun, M., Park, C. G., Kim, J. Y., Rock, C. O., Jackowski, S. & Park, H. W. (2000). Structural basis for the feedback regulation of *Escherichia coli* pantothenate kinase by coenzyme A. *J. Biol. Chem.* **275**, 28093–28099.
72. Calder, R. B., Williams, R. S., Ramaswamy, G., Rock, C. O., Campbell, E., Unkles, S. E. *et al.* (1999). Cloning and characterization of a eukaryotic pantothenate kinase gene (panK) from *Aspergillus nidulans*. *J. Biol. Chem.* **274**, 2014–2020.
73. Matte, A., Tari, L. W. & Delbaere, L. T. (1998). How do kinases transfer phosphoryl groups? *Structure*, **6**, 413–419.
74. Bernstein, B. E., Michels, P. A. & Hol, W. G. (1997). Synergistic effects of substrate-induced conformational changes in phosphoglycerate kinase activation. *Nature*, **385**, 275–278.
75. Machius, M., Chuang, J. L., Wynn, R. M., Tomchick, D. R. & Chuang, D. T. (2001). Structure of rat BCKD kinase: nucleotide-induced domain communication in a mitochondrial protein kinase. *Proc. Natl Acad. Sci. USA*, **98**, 11218–11223.
76. Daugherty, M., Vonstein, V., Overbeek, R. & Osterman, A. (2001). Archaeal shikimate kinase, a new member of the GHMP-kinase family. *J. Bacteriol.* **183**, 292–300.
77. Galperin, M. Y., Walker, D. R. & Koonin, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Res.* **8**, 779–790.
78. Krell, T., Coggins, J. R. & Laphorn, A. J. (1998). The three-dimensional structure of shikimate kinase. *J. Mol. Biol.* **278**, 983–997.
79. Muller-Dieckmann, H. J. & Schulz, G. E. (1994). The structure of uridylyl transferase with its substrates, showing the transition state geometry. *J. Mol. Biol.* **236**, 361–367.
80. Luttgen, H., Rohdich, F., Herz, S., Wungsintaweeikul, J., Hecht, S., Schuhr, C. A. *et al.* (2000). Biosynthesis of terpenoids: YchB protein of *Escherichia coli* phosphorylates the 2-hydroxy group of 4-diphosphocytidyl-2C-methyl-D-erythritol. *Proc. Natl Acad. Sci. USA*, **97**, 1062–1067.
81. Jomaa, H., Wiesner, J., Sanderbrand, S., Altincicek, B., Weidemeyer, C., Hintz, M. *et al.* (1999). Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science*, **285**, 1573–1576.
82. Lichtenthaler, H. K., Zeidler, J., Schwender, J. & Muller, C. (2000). The non-mevalonate isoprenoid biosynthesis of plants as a test system for new herbicides and drugs against pathogenic bacteria and the malaria parasite. *Z. Naturforsch.* **55**, 305–313.
83. Wilding, E. I., Brown, J. R., Bryant, A. P., Chalker, A. F., Holmes, D. J., Ingraham, K. A. *et al.* (2000). Identification, evolution, and essentiality of the mevalonate pathway for isopentenyl diphosphate biosynthesis in gram-positive cocci. *J. Bacteriol.* **182**, 4319–4327.
84. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucl. Acids Res.* **28**, 33–36.
85. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S. *et al.* (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucl. Acids Res.* **29**, 22–28.
86. Walker, D. R. & Koonin, E. V. (1997). SEALS: a system for easy analysis of lots of sequences. *Proc. Conf. Intell. Syst. Mol. Biol.* **5**, 333–339.
87. Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217.
88. Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.
89. Bossemeyer, D., Engh, R. A., Kinzel, V., Ponstingl, H. & Huber, R. (1993). Phosphotransferase and substrate binding mechanism of the cAMP-dependent protein kinase catalytic subunit from porcine heart as deduced from the 2.0 Å structure of the complex with Mn<sup>2+</sup> adenylyl imidodiphosphate and inhibitor peptide PKI(5–24). *EMBO J.* **12**, 849–859.
90. Schlichting, I. & Reinstein, J. (1999). pH influences fluoride coordination number of the AlFx phosphoryl transfer transition state analog. *Nature Struct. Biol.* **6**, 721–723.
91. Evans, P. R., Farrants, G. W. & Hudson, P. J. (1981). Phosphofructokinase: structure and control. *Philos. Trans. Roy. Soc. ser. B*, **293**, 53–62.
92. Carret, C., Delbecq, S., Labesse, G., Carcy, B., Precigout, E., Moubri, K. *et al.* (1999). Characterization and molecular cloning of an adenosine kinase from *Babesia canis rossi*. *Eur. J. Biochem.* **265**, 1015–1021.
93. Xu, Y. W., Morera, S., Janin, J. & Cherfils, J. (1997). AlF<sub>3</sub> mimics the transition state of protein phosphorylation in the crystal structure of nucleoside diphosphate kinase and MgADP. *Proc. Natl Acad. Sci. USA*, **94**, 3579–3583.

Edited by G. Thornton

(Received 11 February 2002; received in revised form 10 May 2002; accepted 21 May 2002)