J Mol Evol
DOI 10.1007/s00239-010-9351-1

Evolutionary Bases of Carbohydrate Recognition and Substrate Discrimination in the ROK Protein Family

Maria S. Conejo · Steven M. Thompson · Brian G. Miller

Received: 9 December 2009 / Accepted: 3 May 2010

Abstract The ROK (repressor, open reading frame, kinase) protein family (Pfam 00480) is a large collection of bacterial polypeptides that includes sugar kinases, carbohydrate responsive transcriptional repressors, and many functionally uncharacterized gene products. ROK family sugar kinases phosphorylate a range of structurally distinct hexoses including the key carbon source glucose, various glucose epimers, and several acetylated hexosamines. The primary sequence elements responsible for carbohydrate recognition within different functional categories of ROK polypeptides are largely unknown due to a limited structural characterization of this protein family. In order to identify the structural bases for substrate discrimination in individual ROK proteins, and to better understand the evolutionary processes that led to the divergent evolution of function in this family, we constructed an inclusive alignment of 227 representative ROK polypeptides. Phylogenetic analyses and ancestral sequence reconstructions of the resulting tree reveal a discrete collection of active site residues that dictate substrate specificity. The results also suggest a series of mutational events within the carbohydrate-binding sites of ROK proteins that facilitated the expansion of substrate specificity within this family. This study provides new insight into the evolutionary relationship of ROK glucokinases and non-ROK glucokinases (Pfam 02685), revealing the primary sequence elements shared between these two protein families, which diverged from a common ancestor in ancient times.

Keywords ROK · Sugar kinase · Repressor · Substrate specificity · Enzyme evolution

Introduction

The ROK (repressor, open reading frame, kinase) protein family (Pfam 00480) represents a functionally diverse collection of polypeptides that includes carbohydrate-dependent transcriptional repressors, sugar kinases, and as yet uncharacterized open reading frames (Titgemeyer et al. 1994). The Pfam database currently contains more than 4700 ROK family members, many of which were discovered as a result of microbial genome-sequencing projects conducted within the last decade (Finn et al. 2008). Although a vast majority of ROK members are prokaryotic in origin, proteins that possess one or more ROK domains have been identified in organisms from all branches of life. The ROK family is a member of the Actin-ATPase clan (Pfam CL 0108) that includes 21 distinct members, many of which couple phosphoryl transfer/hydrolysis to a functionally significant conformational change. In general, a distinction between repressors and kinases can be made based upon the N-terminal sequence of ROK polypeptides. ROK kinases contain a conserved N-terminal ATP binding motif of sequence DxGxT, while ROK repressors possess a

Electronic supplementary material The online version of this article (doi:10.1007/s00239-010-9351-1) contains supplementary material, which is available to authorized users.

M. S. Conejo · S. M. Thompson · B. G. Miller (✉)
Department of Chemistry and Biochemistry, Florida State University, 217 Dittmer Laboratory of Chemistry, Tallahassee, FL 32306-4390, USA
e-mail: miller@chem.fsu.edu

M. S. Conejo
e-mail: mconejo@chem.fsu.edu

S. M. Thompson (✉)
Department of Biology, Valdosta State University, 1500 N Patterson, Valdosta, GA 31698, USA
e-mail: stthompson@valdosta.edu

Published online: 30 May 2010

N-terminal extension that contains a canonical helix-turn-helix DNA binding motif. Although identifying ROK members as putative repressors or kinases is simple, assigning the carbohydrate speciÞcities of individual ROK proteins is more difÞcult, in part due to a lack of structural characterization of this protein family. To date, crystal structures of six ROK polypeptides have been reported; however, the only family member whose structure has been determined in the presence of a bound carbohydrate ligand is the inorganic polyphosphate/ATP-glucomannokinase from *Arthrobacter* sp. strain KM (Mukai et al. 2004; Schiefner et al. 2005).

In addition to the ROK family, microbial glucokinases are also found in two other protein families, the ADP-dependent glucokinases (Pfam 04587) and the non-ROK glucokinases (Pfam 02685). Until recently, the evolutionary relationship between these three families was uncertain. Standard sequence-based search algorithms such as FastA and BLAST do not identify a meaningful similarity between members of each family. Within the last 5 years, however, representative structures from all the three families have been determined (Ito et al. 2001; Lunin et al. 2004; Mukai et al. 2004). These data reveal that the ROK and non-ROK kinases share a common fold that resembles the overall structure of eukaryotic hexokinases. In contrast, the ADP-dependent glucokinase family is a member of the structurally distinct Ribokinase clan (Pfam CL 0118). The structural similarity between the ROK and non-ROK glucokinases provides strong evidence indicating that these protein families are evolutionarily related and likely share a common ancestor.

Our laboratory is interested in utilizing the sugar kinase members of the ROK family as a model system for understanding the determinants of substrate speciÞcity within a group of protein catalysts related by divergent evolution. The evolutionary bases of enzyme speciÞcity are important to understand, as this feature is a deÞning characteristic of protein catalysts. Indeed, most enzymes possess the ability to recognize and transform minor differences in substrate structure into large differences in catalytic efÞciencies, often separated by many orders of magnitude. The phosphoryl transfer reactions performed by ROK kinases are an attractive model system for understanding the chemical and structural bases for enzyme speciÞcity. This class of enzymes was valuable in formulating several theories of substrate discrimination by protein catalysts, including the original lock-and-key model of enzymeÐsubstrate interactions developed by Emil Fischer in 1894 (Fischer 1894). A revised version of the lock-and-key theory was developed by Daniel Koshland Jr. in an attempt to explain the ability of sugar kinases to avoid the wasteful transfer of the γ-phosphoryl group of ATP to solvent water (Koshland 1958). KoshlandÕs model became known as the induced Þt theory of enzyme speciÞcity and it successfully predicted the conformational ßexibility inherent to proteins. Based on these historical precedents and the wealth of sequence information currently available for the ROK family, this group of polypeptides represents an ideal choice for exploring the determinants of enzyme speciÞcity from an evolutionary perspective.

In our previous study, we discovered that several ROK family sugar kinases encoded within the *Escherichia coli* genome have overlapping substrate speciÞcities (Miller and Raines 2004). In particular, we found that allokinase (AlsK) and *N*-acetylmannosamine kinase (NanK) possess weak phosphoryl transfer activity toward the alternate substrate D-glucose (Miller and Raines 2004; Miller and Raines 2005). The identiÞcation of such substrate ambiguity led us to investigate the possibility that these latent catalysts could serve as useful starting points for the evolution of enzymes with altered speciÞcities. Using laboratory-based directed evolution, we identiÞed a pair of single amino acid substitutions that increased the speciÞcity of AlsK and NanK toward glucose by 78-fold and 24-fold, respectively (Larion et al. 2007). In the case of AlsK, a multiple sequence alignment revealed that the glucokinase stimulatory substitution of glycine for alanine at position 73 restored a conserved glycine residue found within many glucose-speciÞc ROK family members.

In this study, we expand upon our previous experimental and phylogenetic investigation of ROK family members to include proteins from seven functionally representative categories. A collection of more than 220 ROK sequences was used to construct a global alignment of this protein family. A phylogenetic analysis of this alignment, which includes ancestral reconstructions of functionally distinct clades, afforded the identiÞcation of discrete amino acids that dictate the carbohydrate speciÞcities of both sugar kinases and carbohydrate-dependent transcriptional repressors. These data can be used as a foundation to tentatively assign function to previously uncharacterized and yet-to-be discovered ROK family members. We also use the results of our phylogenetic analyses to postulate the existence of new carbohydrate speciÞcities within ROK sugar kinases from Cyanobacteria and Chlorobi.

Materials and Methods

Data Set Assembly and ReÞnement

The initial ROK data set was assembled from an HMMerSearch (Eddy 1998) of UniProt (version 12.8, Feb. 2008) with an HMM proÞle from the Þnal alignment of our previous study (Larion et al. 2007) within the GCG SeqLab interface (Genetics Computer Group 1982Ð2008; Smith et al. 1994). An E-value cutoff of 1e-03 yielded 1280

sequences. A Þrst pass with MAFFTÕs EINSI mode (Katoh et al. 2005) reduced the data set to 1134 sequences by eliminating exact duplicates. T-CoffeeÕs (Notredame et al. 2000; Wallace et al. 2006) seq_reformat trim operations were then used to eliminate redundancy by excluding sequences that were 90% or more identical. This procedure reduced the data set to 776 sequences, and it was repeated to exclude sequences with greater than 75% identity, which further reduced the data set to 609 sequences. The in_Þle parameter was used with T-CoffeeÕs seq_reformat trim option each time to retain all SwissProt entries, all entries with solved structures, all Archaea and Eukaryotic entries, and all entries from the previous study. At that point, several iterative rounds of subjective inclusion and exclusion decisions were made based on a series of FastTree (Price et al. 2009) neighbor-joining tree analyses with data sets that had been masked so that regions more than 95% divergent were not considered. Sequences with extra-long branches and those with sequence regions that appeared to be erroneous were eliminated. Furthermore, clades with many similar members were reduced down to a few representative sequences, unless they were of those entries desired based on the criteria in the trim operation. This series of analyses reduced the data set from 609 to the Þnal 227 members. The Þnal data set is available in supplemental material, including both the complete sequence length (740 amino acids [including gaps]), and the 5% masked version (339 amino acids [including gaps]) used for phylogenetic inference.

Merging ROK and Non-ROK Glucokinases

We attempted to force an alignment between a representative set of bacterial glucokinases (Pfam 02685) that were not included in our alignment, after the observation that our entire ROK data set speciÞcally excluded many well-known bacterial glucokinases. This representative set was assembled by searching UniProt (version 12.8, Feb. 2008) with GLK_ECOLI using FastA, (Pearson 1998) while restricting the output to sequences with an E-value less than 5e-05. T-CoffeeÕs seq_reformat trim operation was employed (Notredame et al. 2000; Wallace et al. 2006) to exclude sequences more than 75% identical. This produced a non-ROK GLK data set with 221 members. MAFFTÕs EINSI mode (Katoh et al. 2005) was then used to prepare a multiple sequence alignment from this data set. However, no multiple sequence alignment program, including T-Coffee and MAFFTÕs EINSI mode, nor the proÞle modes of either, could recover a successful alignment between the ROK and non-ROK data sets. Furthermore, no standard sequence-based similarity search tool [FastA (Pearson 1998) nor BLAST (Altschul et al. 1997)] could Þnd any detectable (E-value less than 1.0) sequences from the ROK data set in UniProt when searched with a non-ROK GLK, and vice versa. The only way by which we were able to successfully recover a meaningful alignment between the two data sets was by using HMMerAlign (Eddy 1998) within the GCG package (Genetics Computer Group 1982–2008) to merge an unaligned non-ROK GLK data set with the ROK data setÕs HMM proÞle and its associated alignment. The resulting alignment (467 sequences) was put through a similar iterative reÞnement procedure as the ROK data set using FastTree (Price et al. 2009), subjectively including and excluding sequences based on their behavior. This sequentially reduced the data set from 467 to its Þnal size of 57 members. The original alignment was 456 characters long, but this was reduced to 393 characters by a 5% mask. This data set is made available in the supplemental material.

Phylogenetic Analysis

FastTree version 1.0.0 (Price et al. 2009) was used for preliminary phylogenetic analyses to identify near neighbors that could be excluded from further analysis. RAxML version 7.0.4 (Stamatakis 2006) was used for all subsequent phylogenetic inferences. ProtTest (Abascal et al. 2005) was Þrst used on the data sets from the previous study to estimate the preferable amino acid substitution model (WAG; which was speciÞed to RAxML) and gamma alpha parameter. RAxML was used in its new rapid bootstrap algorithm (the Ðf a option) in combination with a ML (ML) search. This procedure will discover the optimal ML tree and superimpose the bootstrap support values upon it. We performed this procedure with 1000 bootstraps in both analyses (all ROK, and ROK plus non-ROK GLKs).

FastML version 2.02 (Pupko et al. 2002) was run on the data set specifying the optimal RAxML ML fully resolved tree. The FastML program was run assuming a Gamma-distributed rate heterogeneity model with its optimal shape parameter estimated by ML, and the empirical WAG (Whelan and Goldman 2001) model of amino acid evolution. This program maximizes the joint probability of the ancestral node sequences based on the Gamma-distributed model with ML estimated optimal branch lengths on the user speciÞed tree.

Phylogenetic trees were drawn, visualized, and manipulated with FigTree version 1.2.3 (Rambaut 2007). The outgroup was designated as the clade that contained the only two Eukaryotic sequences in the data set: *Paramecium* and *Trichomonas* (both of unknown function, but inferred to be fructokinases in this study). All nodes with less than 50% bootstrap support from 1000 bootstraps were reduced down to the next lower node and drawn as polytomy. Furthermore, bootstrap values are not drawn on any nodes with a value less than 50%.

Results

Large Scale ROK Alignment

A representative alignment of eight functionally distinct ROK family members, selected from the inclusive ROK alignment containing 227 total sequences, is shown in Fig. 1. The location of the helix-turn-helix DNA binding motif found within the xylose and *N*-acetylglucosamine repressors is readily identifiable as an N-terminal extension of approximately 90 residues in length (Kreuzer et al. 1989; Lokman et al. 1991; Sizemore et al. 1991; Angell et al. 1992). The sequences of three functionally characterized ROK kinases are also depicted in Fig. 1. The ATP binding motif, composed of the conserved DxGxT motif (Holmes et al. 1993), and the essential active site Asp residue that functions as a general base to promote phosphoryl transfer are explicitly labeled.

Evolutionary Relationship of ROK and Non-ROK Glucokinases

Following the initial assembly of ROK family members, we noted that many bacterial glucokinases were excluded from the data set. The excluded sequences included members of the non-ROK glucokinase protein family (Pfam 02685), which encompasses a variety of functionally characterized and putative glucokinases from microbes and invertebrates. Notably, the non-ROK glucokinase protein family is a member of the Actin-ATPase clan (Pfam CL0108) to which ROK family members also belong. In order to gain insight into the evolutionary relationship between the ROK and glucokinase families, we used HMMerAlign to merge non-ROK glucokinases with our large-scale ROK alignment. Functionally significant sequence similarities, both in the ATP binding motif and near the critical active site Asp residue, were observed in

Fig. 1 Representative multiple sequence alignment of eight functionally distinct ROK family members (Glk is glucokinase). Important regions of the ROK scaffold are annotated and delineated with a *bold line*. Residues conserved in all but one sequence are shaded in *light gray*, while those conserved in all sequences are shown on a darker background. Residues in lower case did not conform to the original HMM search query

this comparison. Sufficient differences at the primary amino acid level preclude the identification of non-ROK glucokinases when a ROK family member is used to probe the database using traditional sequence-based similarity search tools such as FastA or BLAST. Nevertheless, our alignment with its subsequent ML phylogeny, establishes a clear evolutionary relationship between the ROK and non-ROK glucokinase families. These results are consistent with structural studies demonstrating that both protein families share a similar fold.

ROK Family Phylogenetic Tree

A phylogenetic analysis was carried out with 227 manually selected sequences from the ROK protein family data set, which yielded a ML tree with many well-supported clades distinguishable either by function, by phylum, or both. A collapsed version of the ML tree is shown in Fig. 2. Two major glucokinase clades (GK) were formed, each representing an individual phylum. One clade was formed entirely of Actinobacteria and is well supported, with a bootstrap value of 99. The second clade, containing proteins from Firmicutes, is subdivided into two clades with bootstrap values of 74 and 95. It is noteworthy that other ROK family member proteins are encoded within the genomes of Actinobacteria and Firmicutes, indicating that these organisms harbor a number of kinases/repressors whose functions are yet to be determined.

Individual clades were also observed for *N*-acetylglucosamine (NagK) and *N*-acetylmannosamine (NanK) kinases. Both clades include functionally and structurally characterized gene products and possess primary protein structure elements specific to each sub-family. The NanK clade is strongly supported with a bootstrap value of 95. Similarly, the grouping of members of the NagK clade is well supported, as indicated by the bootstrap value of 94. Interestingly, coding sequences for NanK and NagK appear only within the genomes of Proteobacteria, suggesting a unique environmental and/or metabolic basis for utilizing *N*-acetylated hexosamines in these organisms.

Our analysis revealed two additional groups that appear to cluster exclusively on the basis of function: the polyphosphate glucomannokinases (PPGMK) and the putative fructokinases (FK, ydhR). Function was assigned to polypeptides within the PPGMK and FK clusters by comparison to proteins with previously described functions. The sequence Q7WT42_ARTSK (Mukai et al. 2004), found within the PPGMK clade, has been identified as an inorganic polyphosphate/ATP-dependent kinase on the basis of structural and functional characterization. The clade tentatively identified as fructokinases includes the putative FK from *Bacillus subtilis*, whose structure has been determined, but for which no kinetic characterization appears to have been performed. The PPGMK and FK clades are absolutely supported at their base with bootstrap values of 100 in both instances. In contrast to the NanK and NagK enzymes, PPMGKs and FKs are not limited to a single phylum, as both clades contain sequences belonging to a range of organisms.

A large, single clade supported with the weak bootstrap value of 67 separates the repressors from the remainder of the ROK family. The clade itself is further subdivided into

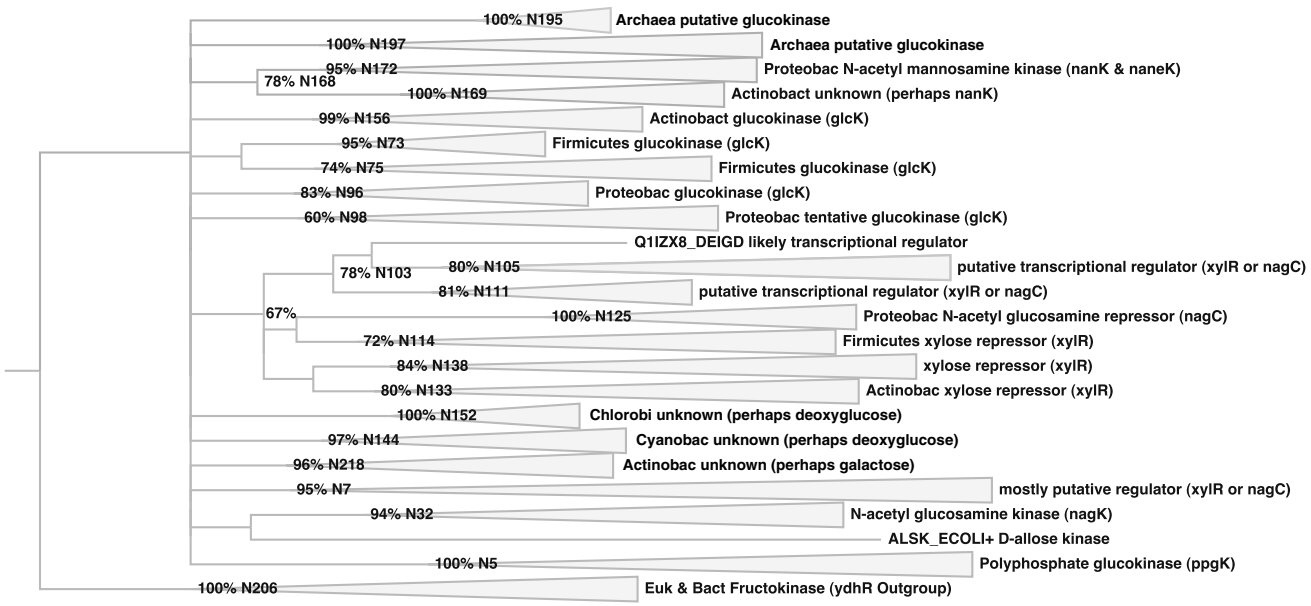

Fig. 2 Collapsed version of the ML tree resulting from the phylogenetic analysis of 227 manually selected sequences from the ROK protein family data set. The percentages represent bootstrap values obtained from 1000 replicates. Bootstrap values less than 50 are not shown

smaller clusters predicted to be xylose or *N*-acetylglucosamine dependent repressors, or that have putative repressor function, as determined by primary structure analysis. It is not surprising that the functionally segregated clade containing the *N*-acetylglucosamine repressor (BS value = 100) is composed entirely of Proteobacteria, since its partnering kinase appears exclusively in Proteobacteria. The experimentally characterized xylose repressor from *Bacillus subtilis* is found within a clade of sequences characterized by a bootstrap value of 72, which exclusively includes members of the Firmicutes phylum. The two remaining clades are of unknown function and have been putatively categorized as repressors given their primary protein structure and their position in the ML tree. Although Actinobacteria is the predominant phylum, species of both the Proteobacteria and Deinococcus-Thermus phyla are present within these clades as well.

Most of the smaller clades found within our ML tree contain sequences not previously characterized, but which could be classiÞed on the basis of phylum and/or domain. Although functions cannot be assigned to sequences contained within these clades without experimental characterization, their primary structures indicate that they are putative sugar kinases. The ROK proteins from Archaea form two individual highly supported clades, both of which have bootstrap values of 100. Polypeptides within both clusters have been tentatively assigned as glucokinases; however, certain amino acid substitutions, particularly within the loop located between the fourth beta sheet (β4) and the second alpha helix (α2) of the ROK scaffold, make this classiÞcation equivocal. Interestingly, the allokinase from *Escherichia coli* K-12 occupies an independent branch within our tree and does not cluster with other ROK sugar kinases, suggesting that the ability to metabolize allose may be a unique attribute of this particular proteobacterium.

SpeciÞcity Determinants in ROK Family Members

The identiÞcation of amino acids responsible for carbohydrate recognition and substrate speciÞcity within functionally distinct ROK family members was facilitated via the multiple sequence alignment (Fig. 1) and the clade-speciÞc ancestral reconstruction sequences (Fig. 3). Guided by these data and the crystal structure of the inorganic polyphosphate/ATP-glucomannokinase from *Arthrobacter* sp. strain KM in complex with glucose (Mukai et al. 2004), we were able to postulate speciÞc interactions within the ligand binding sites of seven functionally distinct ROK clades. These putative interactions are graphically presented in Fig. 4, and are summarized later using residue numbering of the structurally characterized *Arthrobacter* kinase.

Gluco- and Glucomannokinases

The anomeric OH group of bound glucose interacts with Asn-96 and Glu-180 in PPGMK. Similarly, the 6^0-OH group forms interactions with Glu-168 and His-171. The 3^0-OH group of glucose forms two hydrogen bonds, one with the side chain of Glu-168 and the other with Asn-122. The 4^0-OH group of glucose interacts with the side chain of the putative catalytic base, Asp-123, as does the 6^0-OH group, which is the site of phosphorylation. Finally, Van der Waals contacts with glucose are formed via two consecutive residues, Pro-83 and Gly-84, located in a loop region between the fourth β-sheet and the second α-helix of the ROK scaffold.

N-acetylglucosamine Kinases

The active site residues of NagKs are identical to those found in PPGMK and Firmicute glucokinases. Based on our multiple sequence alignment, it is unclear how NagK active sites accommodate the larger size of the acetyl group using the same residues that form interactions with glucose. A repositioning of the loop that harbors the ExGH motif that provides interactions with the 2^0 and 3^0 moieties of the bound carbohydrate is one possibility.

N-acetylmannosamine Kinases

The active site architectures of *N*-acetylmannosamine kinases appear to be highly similar to PPGMK except for a single substitution of a His residue for Glu-168 in the conserved ExGH motif that interacts with the 2^0- and 3^0-OH groups. Based on a comparison of the glucose bound structure of PPGMK, we postulate that the NanK-speciÞc His side chain interacts with the acetylated 2^0-NH group. The substitution of His for Glu-168 may also provide steric bulk to the active site, thereby providing discrimination against carbohydrates with an inappropriate stereochemistry at the 2^0 position. This postulate is supported by the observation that *N*-acetylglucosamine kinases retain the glucose-speciÞc Glu residue, despite possessing a modiÞed 2^0-amino group (vide supra). The conserved Pro-Gly loop sequence located between β4 and α2 of glucokinases is altered to Thr-Gly in NanKs. The lack of a Pro within this loop, a residue with a restricted conformational space, likely affords additional Þexibility in the loop and may provide more space for the acetyl moiety.

Allokinase

The active site architecture of the single allokinase identiÞed to date is predicted to be identical to PPGMK except

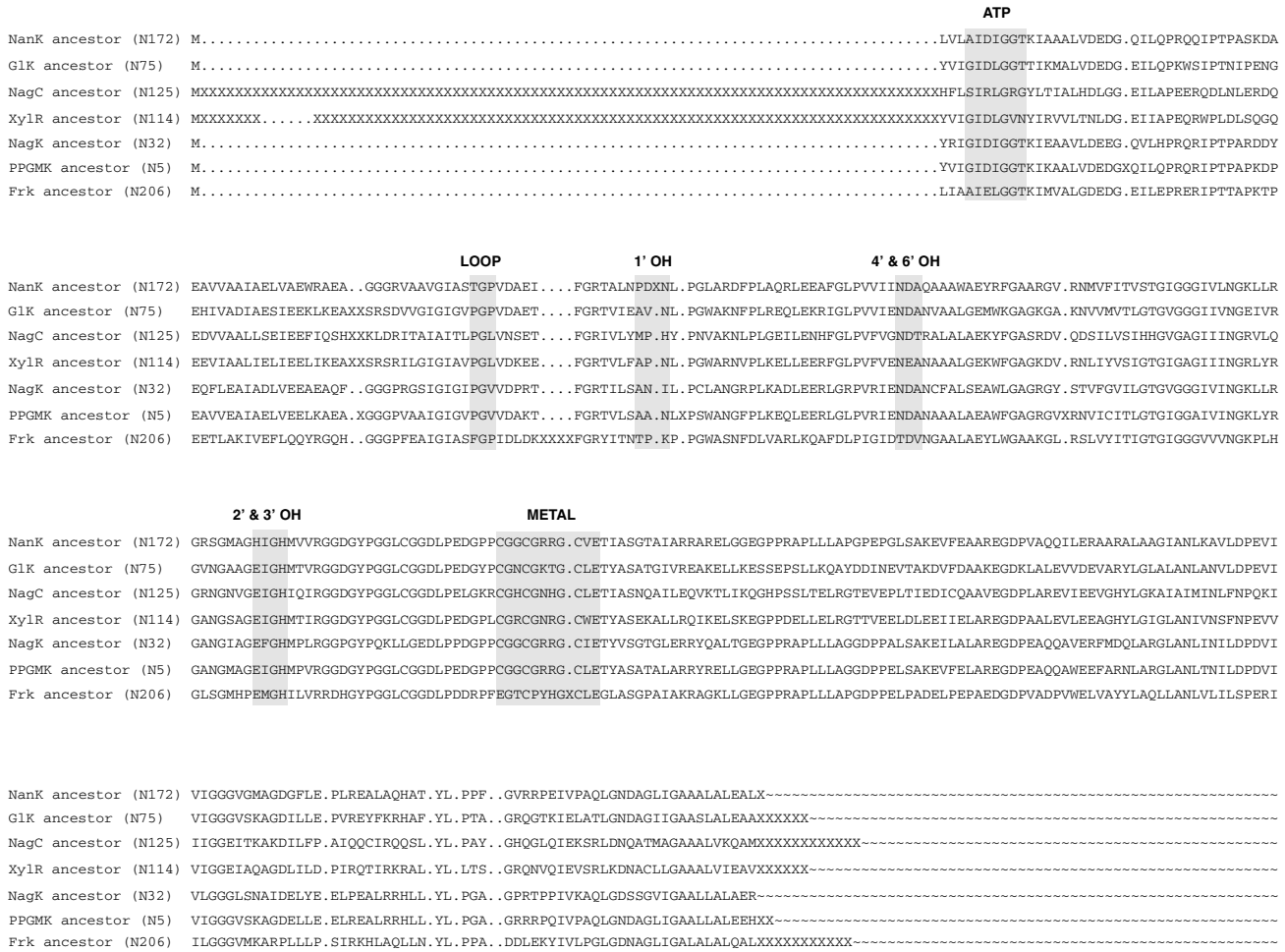

Fig. 3 Multiple sequence alignment of ancestral reconstruction sequences for seven functionally distinct ROK clades. The metal-binding site and location of the active site loop are indicated. Also depicted are the postulated sites of interaction with ATP and specific hydroxyl groups of bound carbohydrates. X represents ancestral sites that could not be reconstructed because the region was excluded from the phylogenetic analysis due to lack of homologous sequence data

for two residues. First, the conserved Asn-96 that interacts with O3 of glucose is replaced with Arg. The longer side chain of Arg may be necessary to interact with the more distant O3 atom of allose, whose stereochemistry is opposite to that of glucose at the 3 position. Second, the Pro-Gly loop sequence located between $\beta 4$ and $\alpha 2$ in glucokinases is altered to Pro-Ala in AlsK. In the structure of PPGMK in complex with glucose, the O3 atom of glucose comes in close proximity to the α proton of Gly-84. We postulate that the addition of a methyl group at this position enforces the preference of the active site for allose compared to its epimer, glucose. This hypothesis is consistent with our previous experimental study demonstrating that replacement of this Ala residue in allokinase with Gly increased the glucokinase activity of the enzyme by 60-fold (Larion et al. 2007).

Fructokinases

Fructokinases lack an Asn residue (Asn-96 in PPGMK), present in all other ROK members, which interacts with the O1 group. Instead, fructokinases harbor a conserved Thr-Pro-Lys triad found at a similar position within the primary amino acid sequence. The Glu side chain that forms an additional hydrogen bond with the 1-OH group in glucokinases is retained in fructokinases. The ExGH motif that interacts with the 2- and 3-OH groups of glucose is also retained in fructokinases. The side chain of Asn-122, which interacts with the 3-OH of glucose in PPGMKs is replaced with Thr, a substitution that appears to be a defining characteristic of fructose-specific ROK kinases. The loop that separates the fourth beta sheet from the second alpha helix contains a Phe-Gly duet that replaces the Pro-Gly

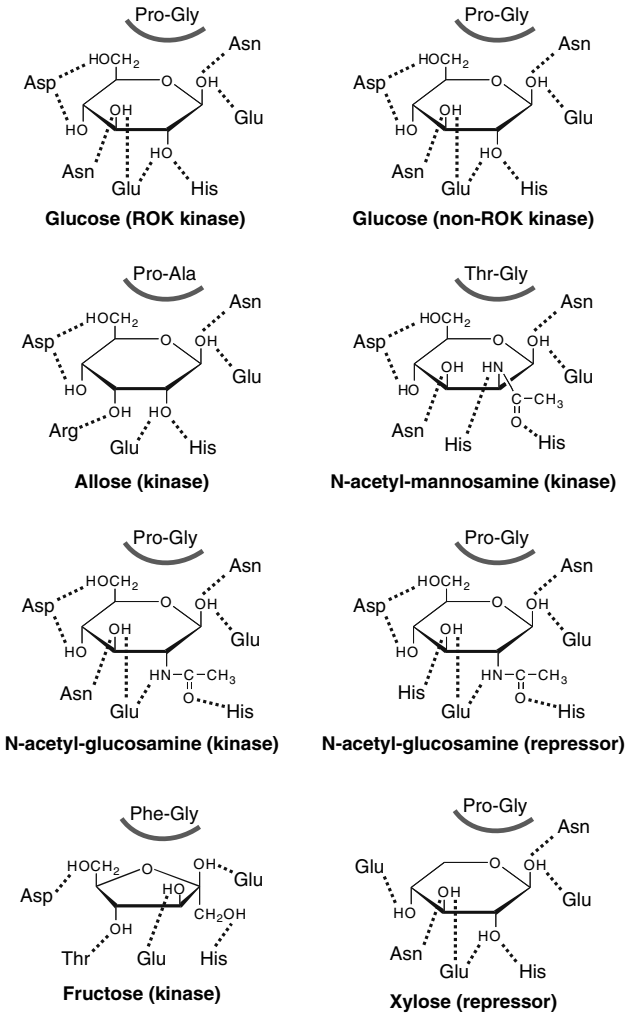

Fig. 4 Active site interactions within seven functionally distinct groups of ROK polypeptides postulated on the basis of the results of phylogenetic analyses. Glucose interactions within the active sites of both ROK and non-ROK kinases were obtained from the crystal structures of *Arthrobacter* inorganic polyphosphate glucomannokinase (PDB entry 1WOQ) and *E. coli* glucokinase (PDB entry 1SZ2), respectively

sequence found within glucokinases and *N*-acetylglucosamine kinases. Finally, there is a conspicuous replacement of Glu for Asp in the conserved N-terminal ATP-binding motif, a region that adopts the consensus DxGxT sequence in all other ROK kinases.

Xylose Repressors

Although not active in phosphoryl transfer, a remnant of the ATP binding signature sequence can be found within the consensus sequence for xylose repressors. The (G/A)IDxGxT N-terminal motif found within NanKs, glucokinases, allokinases, and NagKs has been altered to GIDLGVN in xylose repressors, emphasizing the importance of the Þnal Thr in phosphate recognition. SigniÞcantly, the key catalytic base, Asp-123 in PPGMK, is altered to a Glu in the xylose repressor. We postulate that this substitution prevents binding of ligands that possess an α-1,6-hydroxymethyl group. By analogy to the interactions in the PPGMK active site, this Glu side chain may also form hydrogen bonding contacts with the α-2-OH of xylose. The remaining carbohydrate interacting residues, including those that interact with the α-1, β-2 and β-3-OH groups of glucose in PPGMK, are retained in the ligand binding site of the xylose repressor.

N-acetylglucosamine Repressors

Unlike the xylose repressor, the ATP signature binding motif is not clearly identiÞable in the NagR consensus sequence. The remaining residues in the carbohydrate binding pocket, however, are identical to those found in ROK kinases that act upon glucose and *N*-acetylglucosamine. Unlike the xylose repressor, the active site Asp residue that functions as a general base in ROK sugar kinases is retained in the NagR collection of proteins.

Discussion

A Conserved Metal Binding Site in the ROK Family

Previous investigators (Mesak et al. 2004) reported the presence of a conserved Cys rich motif of sequence CxCGxxGCx(E/D) within microbial glucokinases that belong to the ROK family. Removal of any of the three Cys residues within this motif via site-directed mutagenesis produced an inactive enzyme, demonstrating the functional signiÞcance of these amino acids (Mesak et al. 2004). The crystal structures of three ROK proteins, MLC from *Escherichia coli* (Schiefner et al. 2005; PDB entry 1ZR6), *N*-acetyl mannosamine kinase from *Escherichia coli* (PDB entry 2AA4), and the putative *N*-acetyl glucosamine kinase from *Salmonella typhimurium* (PDB entry 2AP1), demonstrate that the CxCGxxGCx(E/D) motif constitutes a metal-binding site. In each of these structures, a single Zn atom is coordinated to the thiolate side chains of the three Cys residues. Interestingly, in the fructokinase functional clade, this motif has been altered to the sequence CxxHxxCx(E/D). The crystal structure of the putative fructokinase from *Bacillus subtilis* (1XC3) reveals that the imidazole nitrogen from the His side chain provides the third coordination interaction for the Zn atom. The Archaeal clade of functionally characterized ROK sugar kinases contains sequences that also contain an altered metal-binding motif. These polypeptides contain a His residue in place of the third Cys residue. By analogy to the fructokinases, we

speculate that this substitution is unlikely to disrupt metal binding.

Our ancestral sequence reconstruction (Fig. 3) indicates that the CxCGxxGCx(E/D) motif is present within the ancestors of most functionally characterized ROK clades. This finding suggests that the capability to bind metals may be an important functional feature of the ROK scaffold. Nevertheless, our multiple sequence alignment indicates that not all ROK polypeptides contain a metal-binding site. For example, the polyphosphate glucomannokinases and two major clades of Firmicute sugar kinases lack this motif. Cyanobacterial ROK kinases have a partially conserved metal-binding motif in which the first two Cys residues are substituted by Ser in all but one sequence. It seems unlikely that these polypeptides retain metal-binding capabilities, as oxygen provides a much weaker coordination interaction with Zn that does Cys, in part due to its lower basicity.

Past investigators have used the presence of the CxCGxxGCx(E/D) metal-binding motif to indicate membership in the ROK family. Based on our results, this qualification appears to be valid. The converse is not true, however, as the absence of the metal-binding motif does not preclude ROK family membership. Whether the metal-binding site found within many ROK polypeptides plays a structural role, or whether it is more intimately linked to function remains to be experimentally investigated. It is noteworthy that the CxCGxxGCx(E/D) motif is absolutely conserved in all ROK proteins with putative repressor activity, and the appearance of metal-dependent transcriptional repressors is well established in many microorganisms (Hantke 2001). Thus, ROK family repressors may constitute another example of polypeptides that link transcriptional control with the physiological concentrations of metal ions. The validity of this hypothesis and the implications of metal binding to ROK repressors are worthy of future exploration.

Divergent Evolution of Function in ROK Polypeptides

A goal of this research project was to understand the evolutionary history of the ROK family. In particular, we wanted to understand the potential evolutionary pathway(s) that led to the functional divergence of carbohydrate specificity in ROK family sugar kinases. Similar phylogenetic analyses have been conducted on other protein families, including the serine proteases and glutathione transferases, to reveal evolutionary relationships between distantly related polypeptide sequences (Krem and Di Cera 2001; McGoldrick et al. 2005). Past investigators have postulated that the origins of the ROK family may be traced back to polyphosphate-dependent gluco/mannokinases (Mukai et al. 2004; Kawai et al. 2005). The fact that the inorganic polyphosphate-dependent kinases characterized to date possess broad substrate specificities, both with respect to the identity of the phosphoryl donor and the carbohydrate acceptor, is consistent with their ancient lineage. Despite these considerations, the results of our phylogenetic analyses do not support PPGMK as the ancestral prototype of the ROK family. In both the ROK exclusive (Fig. 2) and merged (Fig. 5) phylogenetic analyses, the polyphosphate-dependent kinases represent an outgroup located far from the base of each tree. The combined analysis of ROK and non-ROK family members suggests that the clade containing the glucokinase from *Streptomyces coelicolor* is a more likely candidate for an ancestral ROK prototype.

Putatively, one of the earliest evolutionary events that occurred during the history of the ROK family was the divergence of the non-ROK glucokinases. No evolutionary relationship between the ROK protein family and the non-ROK glucokinases is apparent using simple sequence-based search algorithms. However, when the crystal structures of representative members of each family were determined, a clear structural similarity between both groups was revealed. Our multiple sequence alignment and ancestral reconstruction data are consistent with a model in which the ROK and non-ROK glucokinases diverged from a common ancestor long ago (Kawai et al. 2005). Interestingly, the active site residues that interact with glucose have been completely conserved in both ROK and non-ROK glucokinases (Fig. 4), despite significant sequence divergence in other regions of the polypeptides.

After the divergence of the non-ROK glucokinases from a common ancestor, further functional specialization occurred with the ROK family. The acquisition of an N-terminal DNA binding domain, perhaps through domain swapping, led to the realization of carbohydrate responsive transcriptional repressors that belong to the ROK family. Based on our ancestral reconstruction phylogenetic tree, this event appears to have occurred following a significant level of functional divergence within the ROK sugar kinases. Further specialization of this repressor lineage then yielded polypeptides specific for a range of sugars including xylose and *N*-acetylglucosamine. In particular, the substitution of a His residue in place of an Asn residue that interacts with the 3^{0}-OH group of the carbohydrate ligand promoted repressor specificity toward *N*-acetylglucosamine. Similarly, the substitution of a Glu residue for an Asp residue that interacts with the 2^{0}-OH and 3^{0}-OH group of the carbohydrate in glucose-specific ROK family members appeared to trigger repressor specificity for xylose. Analogous single amino acid substitutions within the active sites of ROK sugar kinases appeared to drive the expansion of substrate specificity. For example, the

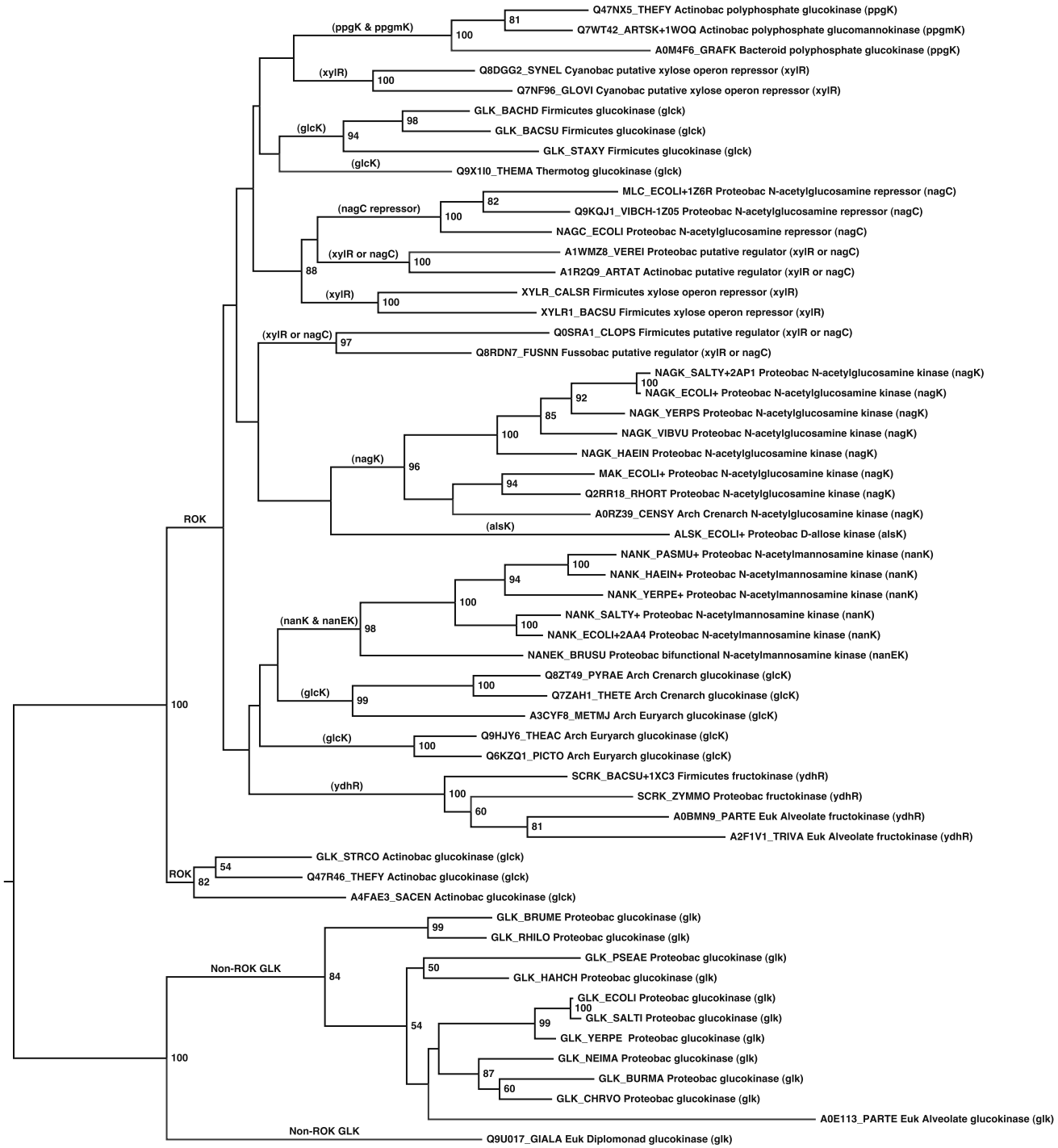

Fig. 5 Maximum likelihood tree resulting from the phylogenetic analysis of a merged multiple sequence alignment data set including both ROK (Pfam 0480) and non-ROK (Pfam 02685) protein family members. The numbers located at each node represent bootstrap values obtained from 1000 replicates. Bootstrap values below 50 are not shown

combination of a Gly to Ala and an Asn to Arg substitution near the 3′-OH group of bound ligand promoted activity toward allose. Similarly, a Glu to His replacement in the vicinity of the 2′-OH group afforded kinase activity toward *N*-acetylmannosamine, and an Asn to Thr substitution enabled the divergence of fructose specific kinases. Significantly, the degeneracy of the genetic code enables several of these amino acid replacements to occur via a single base pair change. Thus, minimal redecoration of the ROK active site architecture, often via a single mutational event, appears to have led to a wealth of functional divergence.

Predicting New Carbohydrate Specificities Within the ROK Family

A vast majority of ROK polypeptides described to date have not been experimentally characterized. As a result, the physiological function of these proteins remains unknown. Using the minimal active site architecture developed on the basis of our multiple sequence alignment and phylogenetic analysis, the carbohydrate specificities for some of these uncharacterized polypeptides can now be predicted. More importantly, our results hint at the possible existence of as-yet-uncharacterized carbohydrate specificities contained with the ROK family. One such example can be observed in the primary sequence of several Actinobacterial ROK sugar kinases that contain an Asn to His substitution in the amino acid that precedes the critical active site Asp catalytic base. In the structure of PPGKM, this Asn residue forms hydrogen bonding interactions with the O3H group of glucose and serves to orient the neighboring Asp residue with respect to the position of the reactive O3H moiety. Replacement of Asn with the bulkier His side chain has the potential to impact the identity of the carbohydrate substrate. Galactose is the 4 epimer of glucose, and although no galactose-specific ROK kinases have been described to date, this hexose is a likely candidate for ROK family members possessing the Asn to His substitution. The prevalence of this hexose in nature makes the emergence of a ROK family member with specificity for this sugar likely.

Another potential example of new carbohydrate specificity within the ROK family is provided by a clade of Cyanobacterial sugar kinases that possess a conspicuous substitution of Leu in place of His within the ExGH motif that interacts with the 2OH and 3-OH groups of bound carbohydrates. This His residue is highly conserved in functionally characterized ROK family members, and the imidazole chain appears to be tolerant of alterations in stereochemistry and acetylation at the ligand O2 group (Fig.4). The substitution of a hydrophobic Leu residue at this position indicates that the putative substrate of the Cyanobacterial ROK kinases may be less hydrophilic than glucose. Similarly, four putative sugar kinases from *Chlorobi*, which cluster together into a single clade with a bootstrap value of 100, possess a Phe in place of this His residue. The large size and hydrophobicity resulting from this substitution suggests that sugars lacking a 2 group may be transformed with reasonable efficiency by these proteins. Although often considered an antimetabolite, 2-deoxyglucose is transformed by certain fungi (Greene1969) and could be a logical substrate for ROK kinases bearing hydrophobic residues at a position within the active site near the expected location of the 2 position of bound ligands. Our multiple sequence alignment also revealed a subset of functionally uncharacterized ROK sugar kinases, largely from Firmicutes, which contain a Tyr residue in place of the His side chain. We speculate that this class of enzymes might be specialized for mannose as a substrate since the His to Tyr substitution retains hydrogen bonding capability, but adds steric bulk near the vicinity of the 2 position, which could enforce preference for a 2 stereoisomer. It is noteworthy that no mannose-specific ROK sugar kinases have been described to date. Instead, mannose has only been characterized as a substrate for the broadly specific kinases that also transform glucose.

In conclusion, our studies provide a foundation on which to classify and tentatively assign function to ROK family members discovered in the future. Moreover, these results provide the opportunity to experimentally redesign the substrate specificity of individual ROK family members using the carbohydrate recognition motifs developed herein. Such work promises to provide new insight into the molecular features that dictate the evolution of substrate selectivity in this highly divergent protein family.

Acknowledgement The authors wish to thank Adi Doron-Faigenboim and Tal Pupko for their technical assistance with the FastML program execution.

References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 9:2104–2105

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 17:3389–3402

Angell S, Schwarz E, Bibb MJ (1992) The glucose kinase gene of *Streptomyces coelicolor* A3(2): its nucleotide sequence, transcriptional analysis and role in glucose repression. Mol Microbiol 19:2833–2844

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 9:755–763

Finn RD, Tate J, Mistry J, Coggill PC, Samut JS, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A (2008) The Pfam protein families database. Nucleic Acids Res 36:D281–D288

Fischer E (1894) Einfluss der configuration auf die Wirkung der enzyme. Ber Dtsch Chem Ges 27:2984–2993

Greene GL (1969) Enzymes of glucose catabolism pathways in *Colletotrichum* and *Gloeosporium*. Mycologia 61:902–914

Hantke K (2001) Iron and metal regulation in bacteria. Curr Opin Microbiol 4:172–177

Holmes KC, Sander C, Valencia A (1993) A new ATP-binding fold in actin, hexokinase and Hsc70. Trends Cell Biol 2:53–59

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 archae*. Structure 9:205–214

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518

Kawai S, Mukai T, Mori S, Mikami B, Murata K (2005) Hypothesis: structure, evolution, and ancestor of glucose kinases in the hexokinase family. J Biosci Bioeng 4:320–330

Koshland DE Jr (1958) Application of a theory of enzyme speciÞcity to protein synthesis. Proc Natl Acad Sci USA 44:98Ð104

Krem MM, Di Cera E (2001) Molecular markers of serine protease evolution. EMBO J 20:3036Ð3045

Kreuzer P, Gärtner D, Allmansberger R, Hillen W (1989) IdentiÞcation and sequence analysis of the *Bacillus subtilis* W23 xylR gene and xyl operator. J Bacteriol 7:3840Ð3845

Larion M, Moore LB, Thompson SM, Miller BG (2007) Divergent evolution of function in the ROK sugar kinase superfamily: role of enzyme loops in substrate speciÞcity. Biochemistry 46: 13564Ð13572

Lokman BC, van Santen P, Verdoes JC, Kerr J, Leer RJ, Posno M, Pouwels PH (1991) Organization and characterization of three genes involved in D-xylose catabolism in *Lactobacillus pentosus*. Mol Gen Genet 1Ð2:161Ð169

Lunin VV, Li Y, Schrag JD, Iannuzzi P, Cygler M, Matte A (2004) Crystal structure of *Escherichia coli* ATP-Dependent glucokinase and its complex with glucose. J Bacteriol 186:6915Ð6927

McGoldrick S, OÕSullivan SM, Sheehan D (2005) Glutathione transferase-like proteins encoded in genomes of yeasts and fungi: insights into evolution of a multifunctional protein superfamily. FEMS Microbiol Lett 242:1Ð12

Mesak LR, Mesak FM, Dahl MK (2004) *Bacillus subtilis* GlcK activity requires cysteines within a motif that discriminates microbial glucokinase into two lineages. BMC Microbiol 4:6

Miller BG, Raines RT (2004) Identifying latent enzyme activities: substrate ambiguity within modern bacterial sugar kinases. Biochemistry 43:6387Ð6392

Miller BG, Raines RT (2005) Reconstitution of a defunct glycolytic pathway via recruitment of ambiguous sugar kinases. Biochemistry 44:10776Ð10783

Mukai T, Kuwai S, Mori S, Mikami B, Murata K (2004) Crystal structure of bacterial inorganic polyphosphate/ATP-glucomannokinase. J Biol Chem 48:50591Ð50600

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 1:205Ð217

Pearson WR (1998) Empirical statistical estimates for sequence similarities searches. J Mol Biol 1:40Ð47

Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with proÞles instead of a distance matrix. Mol Biol Evol 7:1641Ð1650

Pupko T, PeÕer I, Graur D, Hasegawa M, Friedman N (2002) A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: application to the evolution of Þve gene families. Bioinformatics 18:1116Ð1123

Rambaut A (2007) http://tree.bio.ed.ac.uk/software/Þgtree/

Schiefner A, Gerber K, Seitz S, Welte W, Diederichs K, Boos W (2005) The crystal structure of Mlc, a global regulator of sugar metabolism in *Escherichia coli*. J Biol Chem 32:29073Ð29079

Sizemore C, Buchner E, Rygus T, Witke C, Götz F, Hillen W (1991) Organization, promoter analysis and transcriptional regulation of the *Staphylococcus xylosus* xylose utilization operon. Mol Gen Genet 3:277Ð284

Smith SW, Overbeek R, Woese CR, Gilbert W, Gillevet PM (1994) The genetic data environment an expandable GUI for multiple sequence alignment. Comput Appl Biosci 10:671Ð675

Stamatakis A (2006) RAxML-VL-HPC: ML-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 21:2688Ð2690

Titgemeyer F, Reizer J, Reizer A, Saier MH Jr (1994) Evolutionary relationships between sugar kinases and transcriptional repressors in bacteria. Microbiol 140:2349Ð2354

Wallace IM, OÕSullivan O, Higgins DG, Notredame C (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res 34:1692Ð1699

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 5:691Ð699