

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

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

Received: 9 December 2009 / Accepted: 3 May 2010
Springer Science+Business Media, LLC 2010

Abstract The ROK (repressor, open reading frame, mutational events within the carbohydrate-binding sites of kinase) protein family (Pfam 00480) is a large collection of ROK proteins that facilitated the expansion of substrate bacterial polypeptides that includes sugar kinases, carbohydrate specificity within this family. This study provides new hydrate responsive transcriptional repressors, and many insight into the evolutionary relationship of ROK glucose-functionally uncharacterized gene products. ROK family kinases and non-ROK glucokinases (Pfam 02685), reveal-sugar kinases phosphorylate a range of structurally distinct hexoses including the key carbon source glucose, various two protein families, which diverged from a common glucose epimers, and several acetylated hexosamines. The ancestor in ancient times.

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 analysis 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

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

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

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 specificities of individual ROK proteins is more difficult, 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-glucosyltransferase 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 family as a model system for understanding the determinants of substrate specificity. This class of enzymes was valuable in formulating several theories of substrate discrimination by protein catalysts, including the original lock-and-key model of enzyme-substrate 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 1958 (Koshland 1958). Koshland's model became known as the induced fit theory of enzyme specificity and

successfully predicted the conformational flexibility 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 specificity 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 specificities (Miller and Raines 2004). In particular, we found that alkaline phosphatase (AlsK) and *N*-acetylmannosamine kinase (NanK) possess weak phosphoryl transfer activity toward the alternate β -glucose (Miller and Raines 2004; Miller and Raines 2005). The identification 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 specificities. Using laboratory-based directed evolution, we identified a pair of single amino acid substitutions that increased the specificity of AlsK and NanK toward glucose by 78-fold and 24-fold, respectively (Larion et al. 2007). In the case of AlsK, a stimulatory substitution of glycine for alanine at position 73 restored a conserved glycine residue found within many ROK family members.

In this study, we expand upon our previous experimental 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 identification of discrete amino acids that dictate the carbohydrate specificities 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 specificities within ROK sugar kinases from *Cyanobacteria* and *Chlorobi*.

sequences. A first pass with MAFFT's EINSI mode (Katoh et al. 2005) reduced the data set to 1134 sequences and vice versa. The only way by which we were able to eliminating exact duplicates. T-Coffee's (Notredame et al. 2000) seq_reformat trim operations two data sets was by using HMMerAlign (Eddy 1998) 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 the ROK data set's HMM profile and its associated alignment to exclude sequences with greater than 75% identity, which further reduced the data set to 609 sequences. The inclusion through a similar iterative refinement procedure as the parameter was used with T-Coffee's seq_reformat trim ROK data set using FastTree (Price et al. 2009), subjectively including and excluding sequences based on their with solved structures, all Archaea and Eukaryotic entries behavior. This sequentially reduced the data set from 467 and all entries from the previous study. At that point, to its final size of 57 members. The original alignment was several iterative rounds of subjective inclusion and exclusion decisions were made based on a series of FastTree neighbor-joining tree analyses with data supplemental material.

sets that had been masked so that regions more than 95% divergent were not considered. Sequences with extra-long phylogenetic Analysis 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 preliminary phylogenetic analyses to identify near representative sequences, unless they were of those entries that could be excluded from further analysis. RAXML desired based on the criteria in the trim operation. This version 7.0.4 (Stamatakis 2006) was used for all subsequent phylogenetic inferences. ProtTest (Abascal et al. 2005) was first used on the data sets from the previous study to estimate the preferable amino acid substitution length (740 amino acids [including gaps]), and the 5% model (WAG; which was specified to RAXML) and gamma masked version (339 amino acids [including gaps]) used alpha parameter. RAXML was used in its new rapid bootstrap algorithm (the Δf option) in combination with a ML (ML) search. This procedure will discover the optimal ML tree and superimpose the bootstrap support values upon it.

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 specifically 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-Coffee's seq_reformat trim operation was employed (Notredame et al. 2000; Wallace et al. 2006) to exclude sequences more than 75% identical. This procedure

a non-ROK GLK data set with 221 members. MAFFT's 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 MAFFT's EINSI mode, nor the profile mode 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 find any detectable (E-value less than 1.0) sequences from the ROK

Results

Evolutionary Relationship of ROK and Non-ROK Glucokinases

Large Scale ROK Alignment

Following the initial assembly of ROK family members, A representative alignment of eight functionally distinct ROK family members, selected from the inclusive ROK from the data set. The excluded sequences included alignment containing 227 total sequences, is shown in members of the non-ROK glucokinase protein family Fig. 1. The location of the helix-turn-helix DNA binding motif found within the xylose and N-acetylglucosamine characterized and putative glucokinases from microbes and repressors is readily identifiable as an N-terminal extension in vertebrates. Notably, the non-ROK glucokinase protein of approximately 90 residues in length (Kreuzer et al. 1991) family is a member of the Actin-ATPase clan (Pfam 1989 Lokman et al. 1991; Sizemore et al. 1991; Angell CL0108) to which ROK family members also belong. In et al. 1992). The sequences of three functionally characterized ROK kinases are also depicted in Fig. The ATP binding motif between the ROK and glucokinase families, we used binding motif, composed of the conserved DxGxT motif HMMerAlign to merge non-ROK glucokinases with our (Holmes et al. 1993), and the essential active site residue that functions as a general base to promote phosphoryl transfer are explicitly labeled.



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 original HMM search query in bold line. Residues conserved in all but one sequence are shaded in gray.

this comparison. Sufficient differences at the primary characterized gene products and possess primary protein amino acid level preclude the identification of non-ROK structure elements specific to each sub-family. The NanK 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 to proteins with previously described functions. The 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 100 in both instances. In contrast to the NanK and NagK 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 diverse members. 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

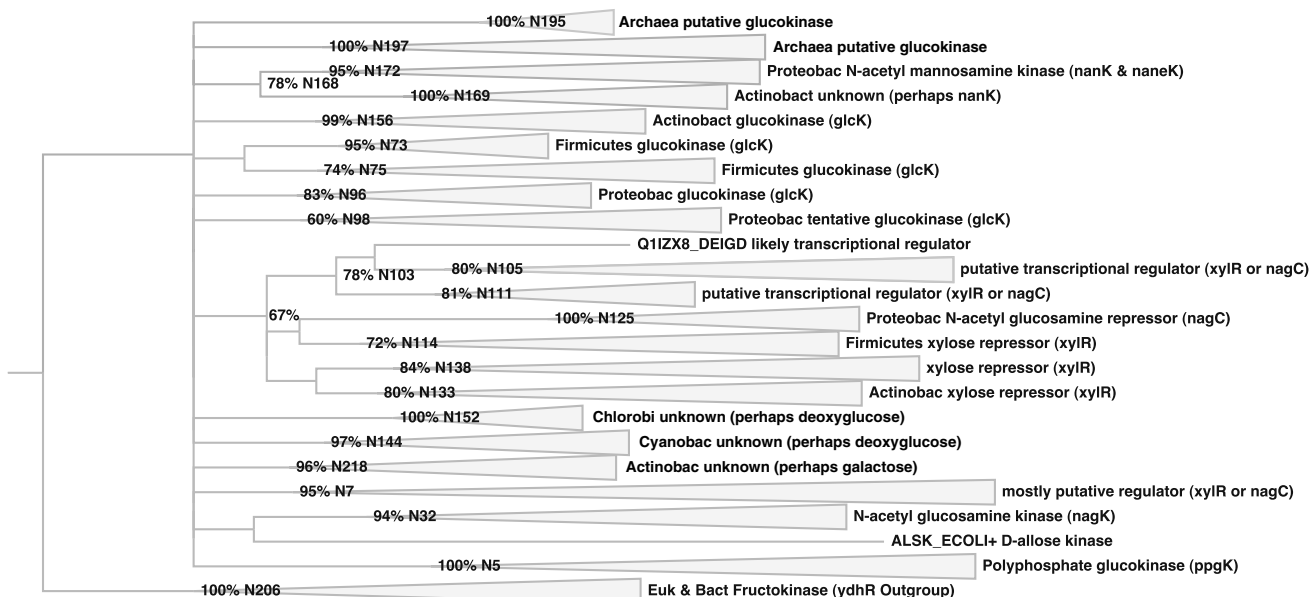


Fig. 2 Collapsed version of the ML tree resulting from the phylogenetic analysis of 227 manually selected sequences from the ROK protein family data set. The percentages represent bootstrap values obtained from 1000 replicates. Bootstrap values less than 50 are not shown

smaller clusters predicted to be xylose 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 group) 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 classified 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 (and the second alpha helix) of the ROK scaffold, make this classification 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.

Specificity Determinants in ROK Family Members

The identification of amino acids responsible for carbohydrate recognition and substrate specificity within functionally distinct ROK family members was facilitated via the multiple sequence alignment (Fig. 1) and the clade-specific ancestral reconstruction sequences (Fig. 2) guided by these data and the crystal structure of the inorganic polyphosphate/ATP-glucomannokinase from *Throbacter* sp. strain KM in complex with glucose (Mukai et al. 2004), we were able to postulate specific 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 *Throbacter* kinase.

The anomeric OH group of bound glucose interacts with Asn-96 and Glu-180 in PPGMK. Similarly, the 2'-OH group of glucose forms two hydrogen bonds, one with the side chain of Glu-168 and the other with Asn-122. The 4'-OH group of glucose interacts with the side chain of the putative catalytic base, Asp-123, as does the 6'-OH group, which is the site of phosphorylation. Finally, Van Waals contacts with glucose are formed via two consecutive residues, Pro-83 and Gly-84, located in a loop between the fourth beta-sheet and the second alpha-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' and 3' 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' and 3'-OH groups. Based on a comparison of the glucose bound structure of PPGMK, we postulate that the NanK-specific His side chain interacts with the acetylated 2'-OH 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' position. This postulate is supported by the observation that N-acetylglucosamine kinases retain the glucose-specific Glu residue, despite possessing a modified 2'-amino group (vide supra). The conserved Pro-Gly loop sequence located between $\beta 4$ and $\alpha 2$ 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 flexibility in the loop and may provide more space for the acetyl moiety.

Allokinase

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

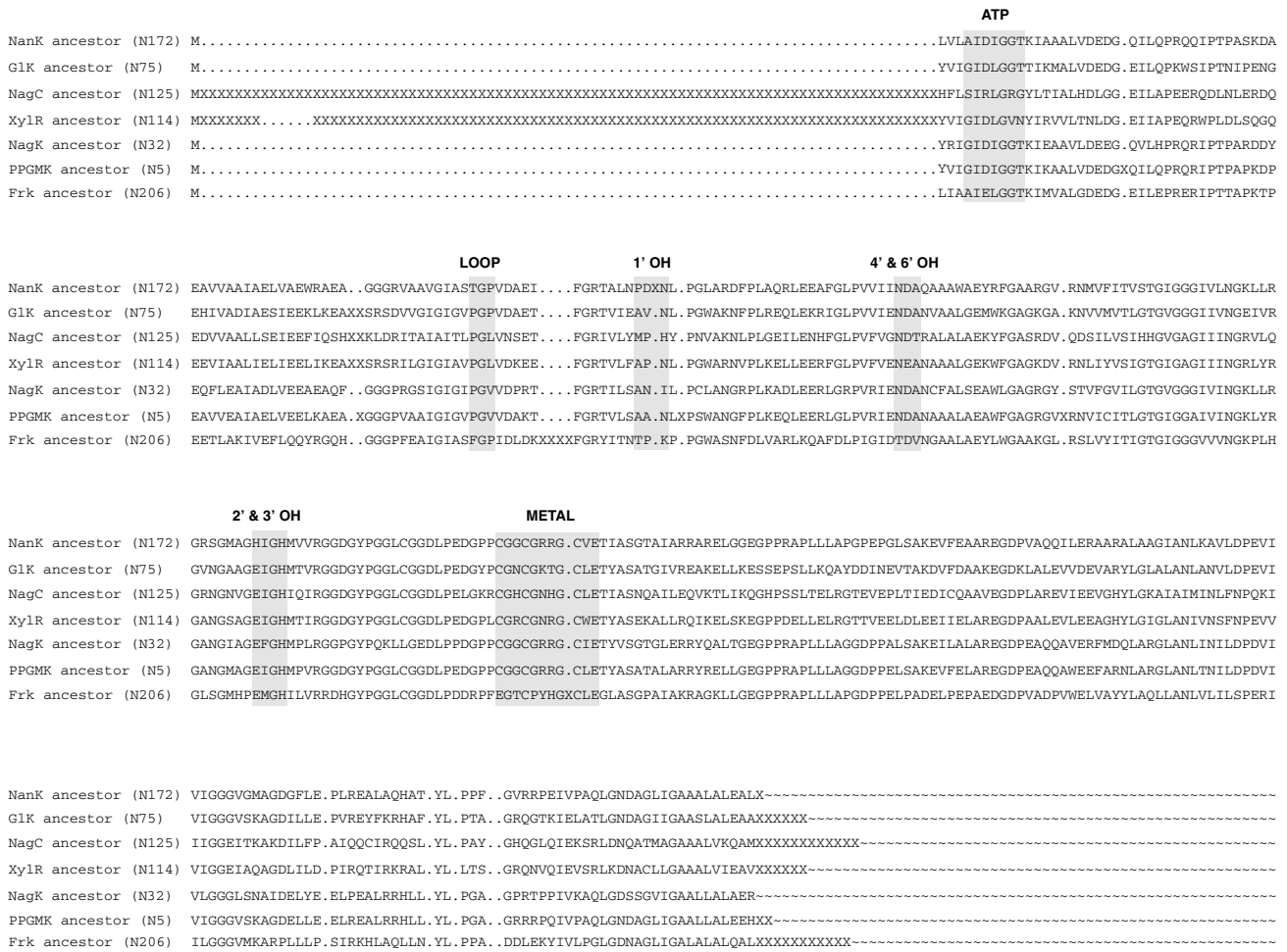


Fig. 3 Multiple sequence alignment of ancestral reconstruction hydroxyl groups of bound carbohydrates. X represents ancestral sites sequences for seven functionally distinct ROK clades. The metal that could not be reconstructed because the region was excluded from binding site and location of the active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

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 position. Second, the O1 group. Instead, fructokinases harbor a conserved Thr-Pro-Gly loop sequence located between $\alpha 1$ and $\alpha 2$ in Pro-Lys triad found at a similar position within the primary 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 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 allokinaise with Gly increased the glucokinase activity of the enzyme by 60-fold (Larion et al.2007). The loop that separates the fourth beta sheet from the second alpha helix contains a Phe-Gly duet that replaces the Pro-Gly

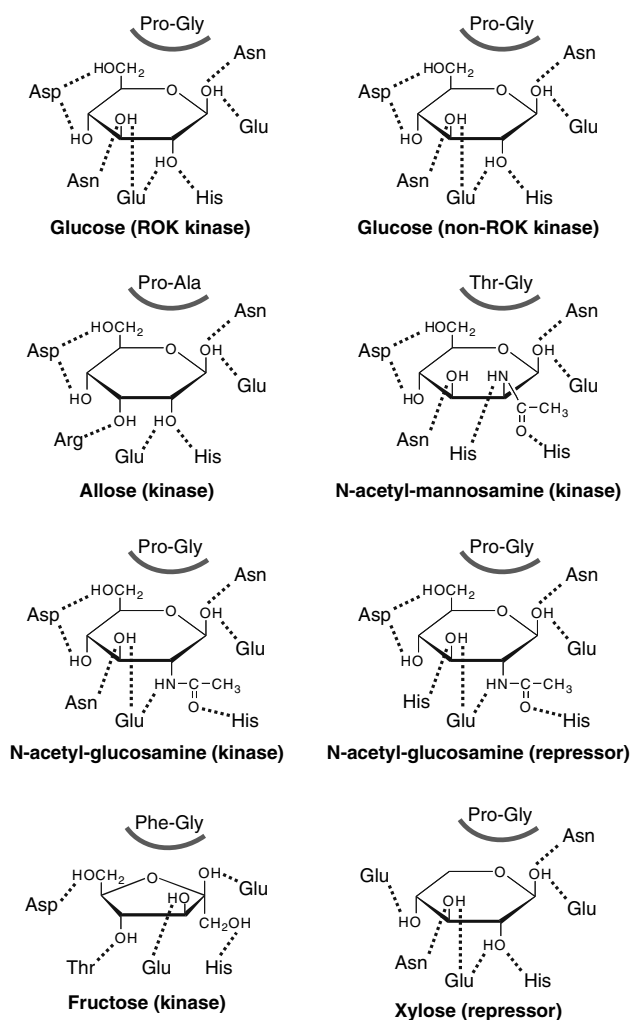


Fig. 4 Active site interactions within seven functionally distinct groups of ROK polypeptides postulated on the basis of the results of phylogenetic analyses. Glucose interactions within the active sites of both ROK and non-ROK kinases were obtained from the crystal structures of *Arthrobacter* inorganic polyphosphate glucomannokinase (PDB entry 1WOQ) and *E. coli* glucokinase (PDB entry 1S22), 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 D_xG_xT 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 (G₁)ID_xG_xT N-terminal motif found within NanKs, glucokinases, allokinases, and NagKs has been altered to GID_xLGVN in xylose repressors, emphasizing the importance of the Pnal Thr in phosphate recognition. Significantly, 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 a hydroxymethyl group. By analogy to the interactions in the PPGMK active site, this Glu side chain may also form hydrogen bonding contacts with the 4-OH of xylose. The remaining carbohydrate interacting residues, including those that interact with the 1,2^o and 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 identifiable 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 C_xCG_xGC_x(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 significance 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 C_xCG_xGC_x(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 C_xH_xC_x(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 site 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. The inorganic polyphosphate-dependent kinases characterized to date possess broad substrate specificities, both with respect to the identity of the phosphoryl donor and the 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 and non-ROK phylogenetic analyses, the polyphosphate-dependent kinases represent an arid group located far from the base of each tree. The two major clades of Firmicute sugar kinases lack this conserved metal-binding 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. Ancestral ROK prototype.

Our ancestral sequence reconstruction (Fig. 5) indicates that the CxCGxxGCx(E/D) motif is present within the carbohydrate acceptor, is consistent with their ancient 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 of the Firmicute clade 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. Ancestral ROK prototype. 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 pathways 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 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

ized to date possess broad substrate specificities, both with respect to the identity of the phosphoryl donor and the 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 and non-ROK phylogenetic analyses, the polyphosphate-dependent kinases represent an arid group located far from the base of each tree. The two major clades of Firmicute sugar kinases lack this conserved metal-binding 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. Ancestral ROK prototype. 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. 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'-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 4'-OH and 6'-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

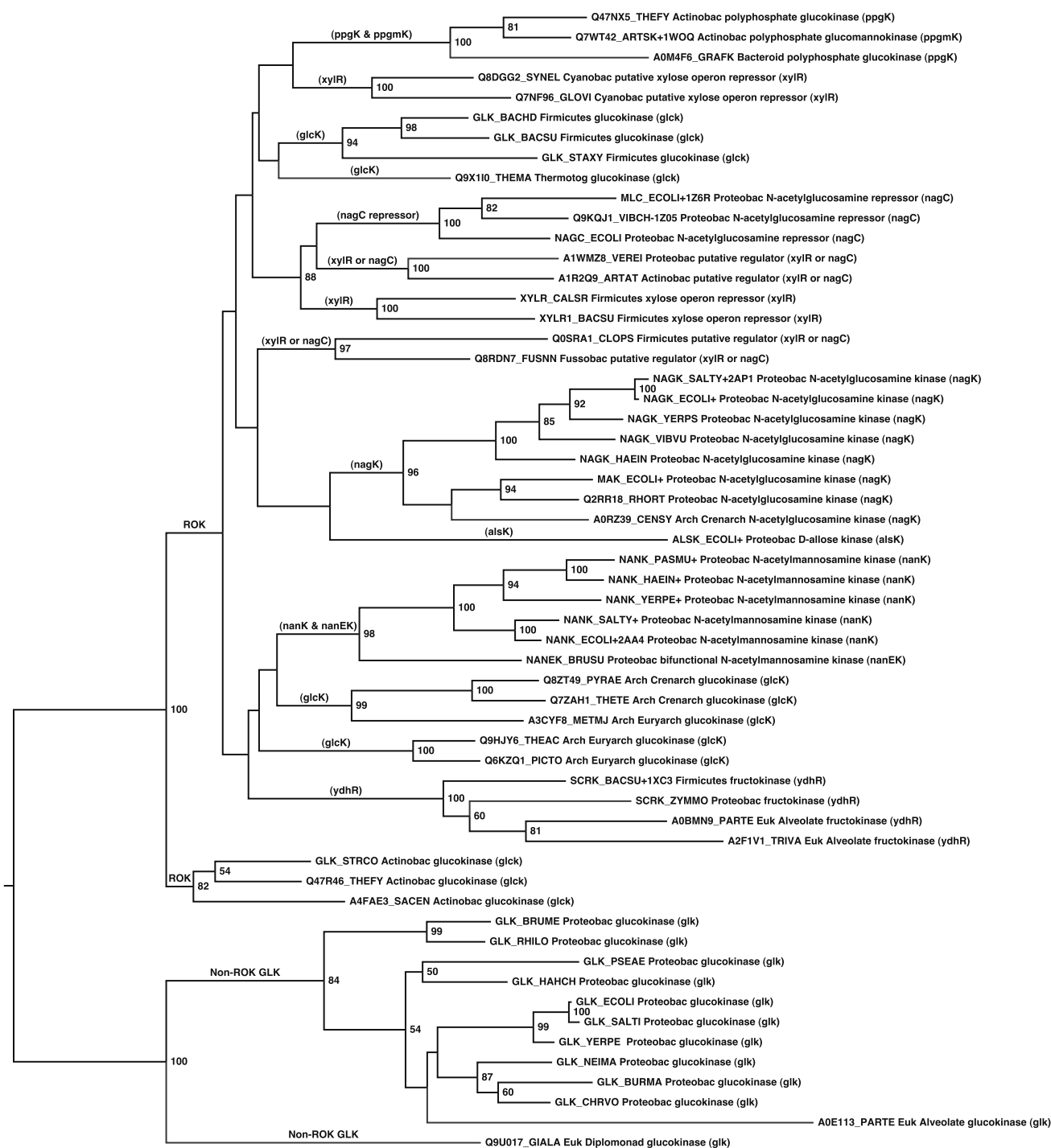


Fig. 5 Maximum likelihood tree resulting from the phylogenetic members. The numbers located at each node represent bootstrap analysis of a merged multiple sequence alignment data set including values obtained from 1000 replicates. Bootstrap values below 50 are both ROK (Pfam 0480) and non-ROK (Pfam 02685)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

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

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 within 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 OH group of glucose and serves to orient the neighboring Asp residue with respect to the position of the reactive OH moiety. Replacement of Asn with the bulkier His side chain has the potential to impact the identity of the carbohydrate substrate. Galactose is the dimer 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 2-OH 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 OH2 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 the OH2 group may be transformed with reasonable efficiency by these proteins. Although often considered an antimetabolite, 2-deoxyglucose is transformed by certain fungi (Greene 1969) 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 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 causing 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, Schaffer 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) ProPle hidden Markov models. *Bioinformatics* 9:755-763
- Finn RD, Tate J, Mistry J, Coghill 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) Einßuss 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 specificity to protein synthesis. *Proc Natl Acad Sci USA* 44:98-104
- Krem MM, Di Cera E (2001) Molecular markers of serine protease evolution. *EMBO J* 20:3036-3045
- Kreuzer P, Gerner D, Allmansberger R, Hillen W (1989) Identification and sequence analysis of the *Bacillus subtilis* W23 xylR gene and xyl operator. *J Bacteriol* 7:3840-3845
- 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 specificity. *Biochemistry* 46:13564-13572
- Lokman BC, van Santen P, Verdoes JC, Kruw, Leer RJ, Posno M, Pouwels PH (1991) Organization and characterization of three genes involved in xylose catabolism in *Lactobacillus pentosus*. *Mol Gen Genet* 162:161-169
- 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* 186:6915-6927
- McGoldrick S, O'Sullivan 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* 242:1-12
- 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* 43:6387-6392
- Miller BG, Raines RT (2005) Reconstitution of a defunct glycolytic pathway via recruitment of ambiguous sugar kinases. *Biochemistry* 44:10776-10783
- Mukai T, Kuwai S, Mori S, Mikami B, Murata K (2004) Crystal structure of bacterial inorganic polyphosphate/ATP-glucomanokinase. *J Biol Chem* 279:50591-50600
- Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 1:205-217
- Pearson WR (1998) Empirical statistical estimates for sequence similarities searches. *J Mol Biol* 1:40-47
- Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with proBles instead of a distance matrix. *Mol Biol Evol* 7:1641-1650
- Pupko T, PeOre 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 pve gene families. *Bioinformatics* 18:1116-1123
- Rambaut A (2007) <http://tree.bio.ed.ac.uk/software/bgtree/>
- 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* 32:29073-29079
- Sizemore C, Buchner E, Rygus T, Witke C, Hillen W (1991) Organization, promoter analysis and transcriptional regulation of the *Staphylococcus xylosus* xylose utilization operon. *Mol Gen Genet* 3:277-284
- 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* 10:671-675
- Stamatakis A (2006) RAxML-VL-HPC: ML-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 21:2688-2690
- Titgemeyer F, Reizer J, Reizer A, Saier MH Jr (1994) Evolutionary relationships between sugar kinases and transcriptional repressors in bacteria. *Microbiol* 140:2349-2354
- Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* 34:1692-1699
- 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* 5:691-699