790–803 *Nucleic Acids Research, 2013, Vol. 41, No. 2* *Published online 2 December 2012*
doi:10.1093/nar/gks1184

Functional diversification of ROK-family transcriptional regulators of sugar catabolism in the Thermotogae phylum

Marat D. Kazanov[1,2], Xiaoqing Li[1], Mikhail S. Gelfand[2], Andrei L. Osterman[1] and Dmitry A. Rodionov[1,2,*]

[1]Sanford-Burnham Medical Research Institute, La Jolla, CA 92037, USA and [2]A.A.Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russia

Received September 21, 2012; Revised October 26, 2012; Accepted October 28, 2012

ABSTRACT

Large and functionally heterogeneous families of transcription factors have complex evolutionary histories. What shapes specificities toward effectors and DNA sites in paralogous regulators is a fundamental question in biology. Bacteria from the deep-branching lineage Thermotogae possess multiple paralogs of the repressor, open reading frame, kinase (ROK) family regulators that are characterized by carbohydrate-sensing domains shared with sugar kinases. We applied an integrated genomic approach to study functions and specificities of regulators from this family. A comparative analysis of 11 Thermotogae genomes revealed novel mechanisms of transcriptional regulation of the sugar utilization networks, DNA-binding motifs and specific functions. Reconstructed regulons for seven groups of ROK regulators were validated by DNA-binding assays using purified recombinant proteins from the model bacterium *Thermotoga maritima*. All tested regulators demonstrated specific binding to their predicted cognate DNA sites, and this binding was inhibited by specific effectors, mono- or disaccharides from their respective sugar catabolic pathways. By comparing ligand-binding domains of regulators with structurally characterized kinases from the ROK family, we elucidated signature amino acid residues determining sugar-ligand regulator specificity. Observed correlations between signature residues and the sugar-ligand specificities provide the framework for structure functional classification of the entire ROK family.

INTRODUCTION

DNA-binding transcription factors (TFs) in bacteria are classified in at least 50 protein families based on sequence similarity and domain composition (1–5). Prokaryotic TFs are usually composed of two domains: (i) a DNA-binding domain that provides the basic function in the recognition of specific DNA sequences and (ii) an effector-sensing domain that modulates the TF activity by monitoring cellular signals and binding specific ligands. The distribution of TFs by families varies among bacterial lineages, mainly due to massive lineage-specific expansions of specific TF families and frequent horizontal gene transfer of individual TFs (6). Gene duplication followed by functional diversification of the duplicated genes is a major driver of evolution. Evolution of diverse DNA-binding and ligand-binding activities in a given TF family is a widely observed phenomenon. However, our understanding of the evolutionary mechanisms driving the observed diversity in large and functionally heterogeneous families of TFs is very limited.

Several families of transcriptional regulators in bacteria possess effector-sensing domains that are homologous to ligand-binding domains of non-regulatory proteins. For instance, the sugar-binding domains in regulators from the LacI and DeoR families share the fold with periplasmic binding proteins (PBPs) of sugar uptake ABC transporters (7) and enzymes from the sugar isomerase family (8), respectively. Structural similarity between the effector-sensing domains of these regulators and the ligand-binding domains of enzymes and transporters suggests that the respective protein families are evolutionarily related. Phylogenetic analysis of PBPs and LacI family regulators revealed that the acquisition of the DNA-binding domain occurred in the last common ancestor of bacteria, and that both functional groups have since undergone extensive gene duplication with

*To whom correspondence should be addressed. Tel: +1 858 646 3100; Fax: +1 858 795 5249; Email: rodionov@burnham.org

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

parallel evolution of ligand specificity (7). Whether similar evolutionary scenarios could explain the origin of functional divergence of ligand specificities in other TF families remains an open question.

To understand the mechanisms of diversification of ligand specificity, we selected the ROK (repressor, open reading frame, kinase) protein family that includes two functionally diverse groups of proteins: (i) catalytically active sugar kinases and (ii) sugar-responsive transcriptional repressors that possess an N-terminal DNA-binding fused to a C-terminal sugar-binding domain from the ROK family (9). The ROK protein family is characterized by the PF00480 domain that belongs to the Actin-ATPase clan in the Pfam database (10). The broad distribution of this family in bacterial genomes is illustrated by ~9600 proteins in Pfam (on August 2012). Among these, nearly 1700 proteins (18%) are potential regulators that possess an N-terminal DNA-binding domain. Some taxonomic groups of bacteria, such as the deep-branching phylum Thermotogae analyzed in this work, demonstrate lineage-specific expansion of putative regulators from the ROK family. However, the specificity and functional role of most ROK regulators remain unknown. The only regulators in this family that have been characterized so far are sugar-responsive NagC from *Escherichia coli*, XylR from *Bacillus subtilis* and CsnR from *Streptomyces lividans*. The NagC regulator negatively controls the *N*-acetylglucosamine (GlcNAc) utilization pathway (11) with GlcNAc-6-phosphate being a negative effector (12). XylR functions as a xylose-responsive repressor of the xylose utilization pathway (13,14). The transcriptional repressor CsnR controls the chitosane utilization pathway by responding to D-glucosamine-containing oligosaccharides (15).

In contrast to regulators, a number of ROK-family kinases that are active on various sugar hexoses including D-glucose, D-allose, D-mannose and D-fructose, as well as on amino sugars have been biochemically characterized (9,16). The level of substrate selectivity varies among different ROK kinases, e.g. the *B. subtilis* fructokinase is specific toward D-fructose, whereas the glucokinase from *Thermotoga maritima* has a broader specificity showing activity on a range of hexoses including D-mannose and GlcNAc (17). Tertiary structures of multiple kinases from the ROK family have been solved, both in a sugar-free form and in complexes with their natural substrates. Among structurally characterized ROK-family proteins are a fructokinase from *B. subtilis* (18), a glucokinase from *Streptomyces griseus* (19), hexokinases from *Arthrobacter* sp. (20) and *Thermus thermophilus* (21) and human *N*-acetyl-D-mannosamine (ManNAc) kinase (22).

The knowledge of exact functions of multiple diverse family representatives is essential for the analysis of evolution and distribution of substrate specificity in the entire family. We use an integrative genomic approach for the analysis of a reference set of regulons associated with a particular family of TFs (23,24). This approach combining genomics-based reconstruction of regulons and associated pathways allows us to accurately identify regulatory DNA motifs, regulated genes and effectors (25) as previously demonstrated for a number of transcriptional regulons in various groups of bacterial genomes (26,27). Among them were detailed reconstructions of transcriptional regulons controlled by two groups of orthologous TFs from the ROK family, NagC in Proteobacteria (28) and XylR in Firmicutes (29,30). Experimental validation of the predicted TF regulons and effectors provides the basis for propagation of solid annotations in functionally heterogeneous TF families and further analysis of molecular determinants of functional specificities.

In this study, we combined bioinformatics and experimental approaches to identify functions and specificities of multiple ROK-family regulators present in hyperthermophilic bacteria from the Thermotogae phylum. For the majority of these regulators, we predicted and verified their candidate DNA-binding motifs, biological roles and molecular effectors. Using the expanded reference set of characterized ROK regulators, we established the signature amino acid residues in their ligand-binding domains and proposed a model of functional diversification in the ROK family in Thermotogae.

MATERIALS AND METHODS

Bioinformatics techniques

Computational tools and resources

Regulators from the ROK protein family were identified in the Thermotogae genomes based on homology to NagC from *E. coli* (REFSEQ: NP_415202) and XylR from *B. subtilis* (REFSEQ: NP_389641) using BLAST with an *E*-value threshold 1.0E−5 (Supplementary Table S1). Multiple sequence alignment of the ROK family proteins was performed with Clustal Omega (31); the phylogenetic tree with bootstrap values was computed using RaxML (32) and visualized with Dendroscope (33). Orthologous groups of the ROK regulators were identified as bidirectional genome-wide similarity hits using the Genome Explorer package with a 30% protein sequence identity threshold (34) and were validated by the analysis of the phylogenetic tree. The MicrobesOnline database and tools (http://www.microbesonline.org/) (35) were used for comparative analysis of genomic contexts for the ROK-family regulator genes, as well as for the identified regulon members. The Thermotogae species tree was constructed using the concatenated alignment of 78 universal proteins in MicrobesOnline (http://www.microbesonline.org/cgi-bin/speciesTree.cgi). Functional annotation of genes from the predicted regulons and reconstruction of the associated metabolic pathways was performed in the SEED genomic platform (http://pubseed.theseed.org/) (36).

Genomic reconstruction of regulons

For reconstruction of regulons for ROK-family TFs, we used the established approach based on the comparative genomic analysis of candidate TF-binding sites (TFBSs) in closely related genomes [reviewed in (37)]. For each group of TF orthologs, a probabilistic model of a putative TFBS motif was constructed and used to search for additional candidate binding sites in a set of related genomes and to

perform further cross-genomic comparison of regulons. The RegPredict web server tool (regpredict.lbl.gov) (38) was used for identification and further comparative analysis of candidate TFBSs of ROK-family regulators. The regulon inference started from construction of training sets of potentially co-regulated genes from conserved genomic neighborhoods of the analyzed TF genes. In each of the seven studied groups of orthologous ROK regulators, the training sets of upstream non-coding regions of potentially regulated operons were extracted and used in the iterative motif detection procedure 'Discovery Profile' implemented in RegPredict. Candidate motifs of palindromic symmetry were analyzed, as it is known that ROK-family TFs bind to palindromic motifs (13,39). For each predicted regulatory motif, a nucleotide frequency positional weight matrix (PWM) was calculated as previously described (40). At the next stage, the constructed PWMs were used by the 'Run Profile' procedure to search for candidate TFBSs in the genomes that had orthologs of a particular ROK regulator and to find additional genes with the same DNA motif in upstream regions. PWM scores of candidate sites were calculated as the sum of positional nucleotide weights. TFBS motifs were visualized as sequence logos using WebLogo (41). The details of the reconstructed ROK-family regulons are captured in the RegPrecise database (http://regprecise.lbl.gov/) (42) as part of the Thermotogae collection.

Prediction of signature residues

For identification of specificity-determining residues (SDRs) in effector-binding domains of ROK-family regulators, we used a computational method that relies on the assumption that binding specificity is conserved among orthologous proteins and is different in paralogous proteins (43). Analysis of SDRs in a protein family starts from the selection of groups of orthologous proteins (the specificity subgroups) and proceeds with identification of residues that can better discriminate between these specificity groups. Candidate SDRs for the analyzed regulators from the ROK family in Thermotogae were identified using the SDPfox tool (44), which requires as input a multiple sequence alignment and a grouping of proteins based on their specificity. As an output, the SDR prediction tool provides a list of rankings for each individual alignment position, and the ranking indicates the significance of the position in distinguishing different isofunctional groups. The ranked list of potential SDRs identified in the analyzed ROK regulators was truncated using the recommended threshold (Supplementary Table S2). To exclude false-positive predictions of SDRs, we filtered out the SDRs located within the hydrophobic core of a protein and SDRs weakly conserved within the specificity groups. The solvent accessibility of residues was calculated using the DSSP tool (45) and the available tertiary structure of the ROK-family regulator TM1224 (PDB code 2HOE). The conservation score for SDRs was calculated using the previously described method (46). The applied thresholds for the solvent accessibility and the conservation score were 20 and 0.9, respectively. The selected seven SDRs were mapped onto the available tertiary structures of the TM1224 regulator and two ROK kinases including the *S. griseus* glucokinase in complex with D-glucose (3VGL) and human ManNAc kinase in complex with ManNAc (2YHW) (Supplementary Figure S1). The signature residues were visualized on the protein structures using the Chimera tool (47).

Experimental procedures

Bacterial strains, reagents and gene cloning

The *E. coli* DL41 strains that carry a pMH2T7-derived plasmid harboring *T. maritima* genes TM0393, TM0808 or TM1224 under control of the arabinose-inducible T7 promoter were a kind gift from Dr S. Lesley at the Joint Center for Structural Genomics (48). The *T. maritima* genes TM0032, TM0110, TM0411 and TM1847 were amplified by polymerase chain reaction (PCR) using specific primer pairs from *T. maritima* MSB8 genomic DNA and cloned into the expression vectors pET28a (Novagen, Madison, WI, USA), pODC29 (49) or pSMT3 (50) (Supplementary Table S3). The pSMT3 expression vector (50) used for the expression of TM1847 was a kind gift from Dr Lima from Cornell University. The resulting plasmids were confirmed by DNA sequencing and transformed into *E. coli* BL21 (DE3) (Gibco-BRL, Rockville, MD, USA). Enzymes for PCR and DNA manipulations were from New England Biolabs Inc. (Beverly, MA, USA). Plasmid purification kits were from Promega (Madison, WI, USA). PCR purification kits and nickel–nitrilotriacetic acid (Ni–NTA) resin were from QIAGEN Inc. (Valencia, CA, USA). Oligonucleotides for PCR and sequencing were synthesized by Sigma-Genosys (Woodlands, TX, USA).

Protein purification

Recombinant proteins TM0393, TM0808, TM1224, TM0032, TM0110 and TM0411 containing an N-terminal His_6 tag were overexpressed in *E. coli* in 50-ml volume culture and purified using Ni^{2+}-chelating chromatography, as previously described (51). TM1847 was produced as the Smt3-His_6-tagged recombinant protein by the same procedure. After purification, N-terminal poly-histidine Smt3 fusion tag was removed from the recombinant TM1847 protein by incubation with the Ulp-4 protease. Protein size, expression level, distribution between soluble and insoluble forms and extent of purification were monitored by SDS–PAGE. Protein concentration was determined by the Quick Start Bradford Protein Assay kit from Bio-Rad. All proteins were obtained with high yield (>1 mg) and purity (80–90%).

DNA-binding assays

The interaction of the purified recombinant tagged proteins with their cognate DNA-binding sites in *T. maritima* was assessed using the electrophoretic mobility shift assay (EMSA). Two sets of 3′-labeled DNA oligonucleotides containing the predicted 23-bp DNA-binding sites from *T. maritima* were synthesized by Integrated DNA Technologies. DNA fragment sequences, sizes and labels used for testing of candidate

regulator binding sites are given in Supplementary Table S4. Biotin-labeled DNA oligonucleotides were obtained for 11 candidate binding sites of six ROK regulators and included the surrounding genomic regions of these sites. Oligonucleotides with a fluorescence label (6-carboxyfluorescin) were synthesized for one TreR-binding site and two GluR-binding sites. The double-stranded labeled DNA fragments were obtained by annealing the labeled oligonucleotides with unlabeled complementary oligonucleotides at a 1:10 ratio. The biotin-labeled DNA fragments (0.1 nM) were incubated with increasing concentrations of the purified regulator protein (0.25–100 nM) in a total volume of 20 µl of the binding buffer containing 20 mM Tris–HCl (pH 8.0), 150 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 1 mM EDTA, 0.05% NP-40 and 2.5% glycerol. Poly(dI-dC) was added to the reaction mixture as a non-specific competitor DNA at 1 µg to suppress non-specific binding. After 30 min of incubation at 60°C, the reaction mixtures were separated by electrophoresis on a 5% native polyacrylamide gel. The DNA was transferred by electrophoresis onto a Hybond-N^+ membrane and fixed by UV cross-linking. Biotin-labeled DNA was detected with the LightShift chemiluminiscent EMSA kit (Thermo Fisher Scientific Inc, Rockford, IL, USA). The fluorescence-labeled DNA fragments (2 nM) were tested using the same assay and detected with the FLA-5100 fluorescent image analyzer. To assess DNA motif specificity, each regulator was tested for binding with DNA fragments containing regulatory sites of another *T. maritima* regulator from the ROK family. To identify effectors of the characterized ROK-family transcriptional regulators, additional EMSA experiments were performed to test the effect of various mono- and disaccharides on the DNA-binding affinity of regulators. For each regulator, one target DNA fragment containing a regulator binding site was selected, and the effect of carbohydrates was tested by their addition to the incubation mixture.

RESULTS

Genomic reconstruction of regulons for ROK-family regulators in Thermotogae

To investigate the functional diversification of ROK-family regulators in the Thermotogae phylum, we start with the analysis of genomic context and reconstruction of associated regulons and pathways that allowed us to infer biological functions of these regulators and predict their specificities at two levels: DNA-binding sites and possible effectors.

Genomic repertoire and features of ROK regulators in Thermotogae

Putative transcriptional regulators from the ROK protein family were identified in the genomes by homology-based scanning with two previously characterized ROK-family regulators, NagC from *E. coli* and XylR from *B. subtilis*. As the ROK protein family contains both DNA-binding regulators and kinases (9), we checked the identified proteins for the presence of putative N-terminal DNA binding domains. Among 66 protein candidates identified by similarity searches, 13 proteins were filtered out as lacking a DNA-binding domain and containing solely the ROK kinase domain. Hence, we identified 53 putative ROK-family regulators that are unevenly distributed among 11 analyzed genomes (Supplementary Table S1). The phylogenetic tree of the ROK-family regulators (Figure 1) identified eight distinct groups of orthologous proteins in Thermotogae, with seven groups present in the model bacterium *T. maritima* (Figure 2). The largest group containing the TM0393 regulator has representatives in all but one of the Thermotogae genomes, whereas the representatives of the TM0808 group are present in all *Thermotoga* and *Thermosipho* species but not in other Thermotogae. Five groups of ROK regulators—including TM0032, TM0110, TM0411, TM1224 and TM1847—were found only in the *Thermotoga* spp. The last group contains only two orthologs found in the genomes of *Thermotoga lettingae* (Tlet_1225) and *Kosmotoga olearia* (Kole_1973). Further analysis was aimed at reconstruction of the ROK-family regulons and associated metabolic pathways that allowed us to predict DNA motifs and possible effectors for ROK regulators in Thermotogae.

Reconstruction of regulons and prediction of DNA motifs

Analysis of genomic context revealed that all studied ROK-family regulators are located within gene loci encoding components of carbohydrate utilization pathways such as sugar ABC transporters and carbohydrate-active enzymes. To find sets of genes regulated by ROK regulators, we used the standard comparative genomics procedure for *de novo* regulon inference implemented in the RegPredict tool (38) that was previously used for the genomic reconstruction of multiple novel TF regulons (23,24,52–54). For each group of orthologous ROK regulators, we collected training sets of DNA upstream regions of prospective target operons identified by the genomic context analysis, and applied the DNA motif recognition program to derive a conserved 23-bp palindromic motif. Then, we used the identified regulatory motifs to search for additional sites in genomes encoding the orthologous regulators. Multiple alignment of non-coding regulatory regions of orthologous genes from closely related *Thermotoga* genomes (phylogenetic footprinting) confirms a high degree of conservation of the predicted regulatory sites (Supplementary Figure S2).

The predicted ROK regulator-binding motifs and the content of reconstructed regulons in *T. maritima* are illustrated in Figure 3. Detailed information about binding sites and downstream regulated genes is provided within the Thermotogae collection in the RegPrecise database (http://regprecise.lbl.gov/RegPrecise/collection_tax.jsp?collection_id=4). The reconstructed TF regulons vary both by number of cognate binding sites and by number of genes within candidate-regulated operons. Each of the three local regulators, TM0032, TM0411 and TM0808, has a single target DNA site per genome and controls one operon containing five to nine genes. The other three regulators, TM1224,

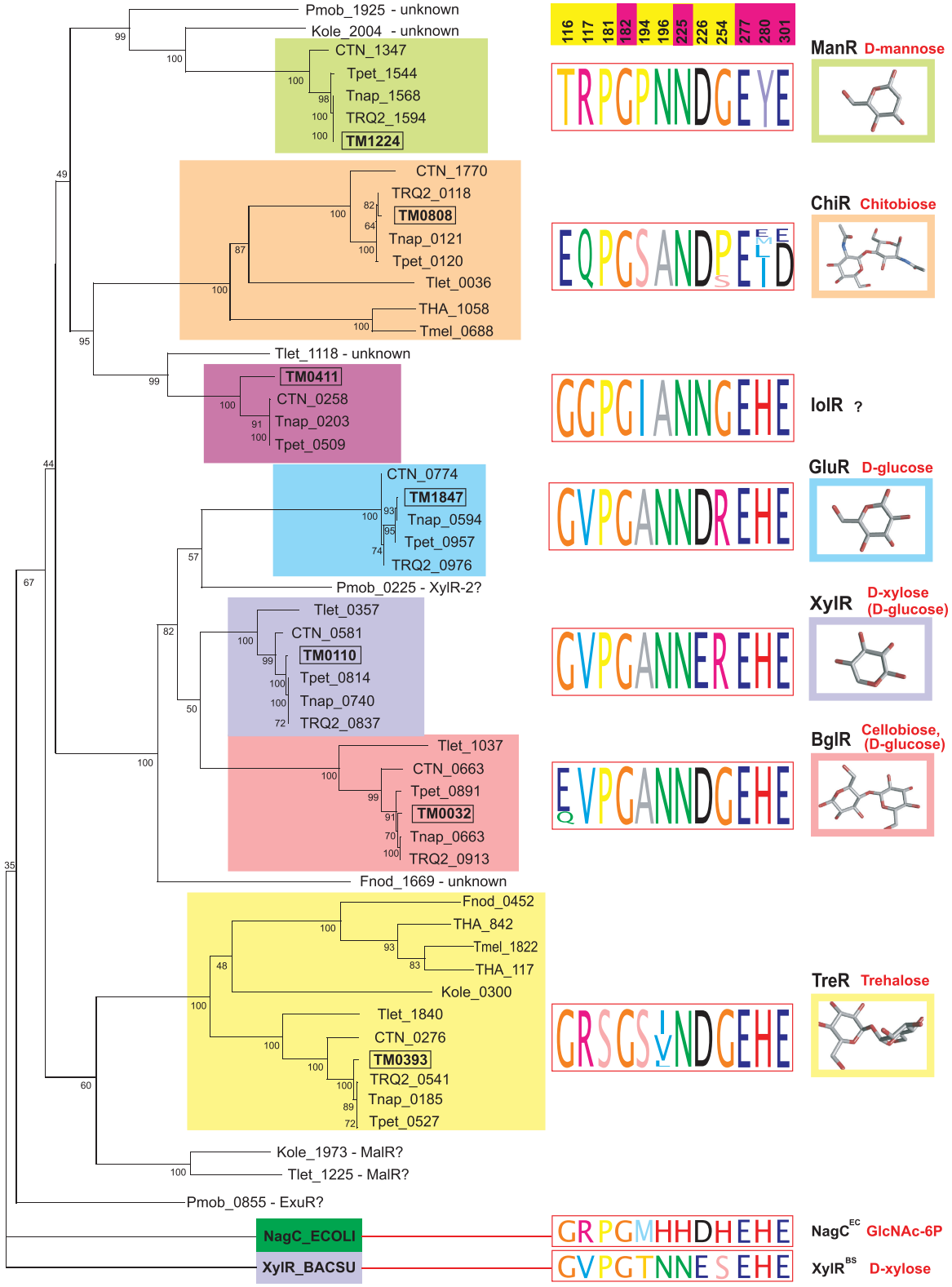

Figure 1. Phylogeny and specificities of ROK-family regulators in the Thermotogae phylum. The maximum likelihood phylogenetic tree of all ROK proteins identified in 11 Thermotogae genomes was reconstructed using the RaxML program with 1000 bootstrap replicates. The numbers in nodes represent bootstrap values in percentages. The *E. coli* NagC and *B. subtilis* XylR proteins were used as outgroups. Attribution of gene locus tags to the genomes is given in Supplementary Table S1. The branches are colored based on their predicted functional specificities. Experimentally determined sugar ligands are shown in red font; predicted ligands are in black. D-glucose is a secondary effector for XylR and BglR regulators. The amino acid distributions of seven SDRs and five conserved residues are shown as a logo. Positions of the SDRs and conserved residues are according to the alignment in Supplementary Figure S4 and are highlighted in yellow and magenta, respectively.

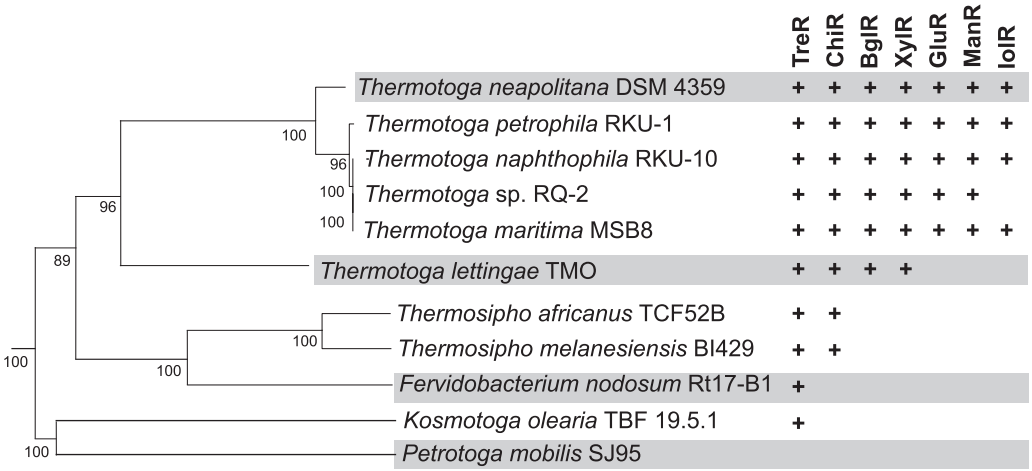

Figure 2. Distribution of eight ROK-family regulators encoded in 11 Thermotogae genomes.

TM1847 and TM0393, have two target DNA sites per genome and control two target operons each. In the case of TM1847, the target operons are adjacent to each other. In contrast, the TM0110 regulator exhibited a more global pattern of gene regulation, with each genome bearing five to six binding sites controlling seven or eight target operons containing up to 27 target genes. The predicted DNA motifs were experimentally tested as described in the 'Materials and Methods' section.

Functional content of reconstructed regulons and prediction of possible effectors

We assessed the functional content of the reconstructed regulons to tentatively predict possible biological functions and effectors of ROK-family regulators in *Thermotoga* spp. Metabolic reconstruction of the respective sugar catabolic pathways, and prediction and refinement of functions of co-regulated genes were performed using the subsystem-based approach implemented in the SEED genomic platform (36). The results of this analysis are captured in the SEED subsystem 'Sugar utilization in Thermotogae', which is available online (http://pubseed.theseed.org/). Overall, the analyzed regulons are associated with specific sugar catabolic pathways that include components of three functional types: (i) secreted enzymes for breakdown of poly- and oligosaccharides; (ii) sugar uptake transporters and (iii) intracellular enzymes constituting a sugar catabolic pathway. Thus, the predicted regulons in *T. maritima* comprise 34 components of ABC-type transporters for mono- and oligosaccharides, 19 intracellular sugar catabolic enzymes, 6 secreted sugar hydrolases, 7 ROK-family regulators (that are subject to autoregulation) and 3 proteins of unknown function (Figure 3). Each of these novel regulators was functionally annotated and named using an abbreviation of the target sugar catabolic genes as described below.

The ChiR/TM0808 regulon contains seven genes that are organized in a single operon, *chiR-cbsA-chiEFG-nagBA*. The β-*N*-acetylglucosaminidase CbsA is a cytoplasmic enzyme catalyzing hydrolysis of GlcNAc in chitobiose and other chitin-derived oligosaccharides (55). NagA and NagB are homologous to the GlcNAc-6-P deacetylase and GlcN-6P isomerase from *Shewanella oneidensis*, respectively (28). Based on tentative pathway reconstruction, the ChiR regulon is responsible for utilization of chitin and/or products of its degradation, chitobiose and GlcNAc. The physiological studies suggest that *T. maritima* is unable to grow on GlcNAc or chitin as a single carbon source (V. Portnoy, personal communication). Thus, we propose that the ChiR-controlled catabolic pathway is responsible for utilization of chitobiose. An ABC-type sugar transporter encoded in the ChiR-regulated operon (termed ChiEFG) is likely involved in the chitobiose uptake. Chitobiose can be largely available in natural habitats populated by *Thermotoga* spp. (e.g. the geothermally heated sea floors) due to: (i) chitin-rich sediment rain to the sea floor and (ii) chitinolytic activity of the microbial community in the sea floor (56).

XylR/TM0110 in *T. maritima* co-regulates multiple gene operons encoding enzymes and transporters for the utilization of xylose and xylose-containing xylosides (17,57,58). ManR/TM1224 co-regulates two operons involved in the degradation of β-mannan and the utilization of resulting mannose-containing oligosaccharides (57,58). BglR/TM0032 regulates an operon involved in the utilization of β-linked glucose polymers (e.g. glucans), and the utilization of resulting β-glucosides such as cellobiose (57,58). TreR/TM0393 co-regulates the trehalose transporter operon *treEFG* (59), and the *treTR* operon encoding the trehalose degradation enzyme TreT (60). GluR/TM1847 co-regulates the glucose transporter *gluEFK* (59) and the trehalose transporter *treEFG*. The trehalose and glucose transporters are significantly upregulated in *T. maritima* grown on glucose (61), and the *treEFG* transporter is also upregulated on trehalose (V. Portnoy, personal communication), thus confirming the GluR and TreR regulons. Finally, IolR/TM0411 regulates an operon encoding a novel variant of the *myo*-inositol catabolic pathway (I. Rodionova, personal communication).

Based on the metabolic pathway reconstruction and the knowledge of pathway metabolites, a range of possible molecular effectors was suggested for most ROK-family

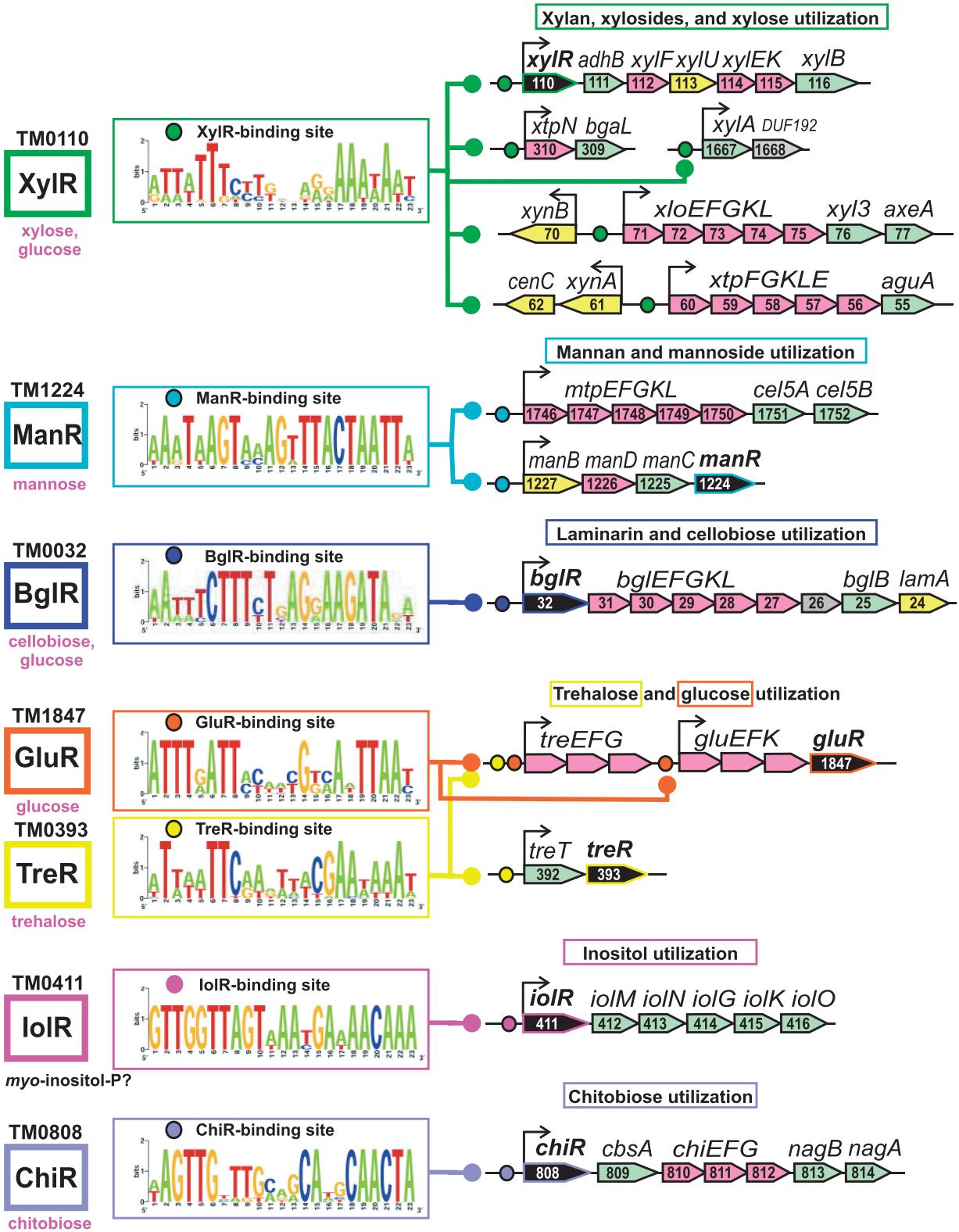

Figure 3. Functional and genomic context of ROK-family regulons in *T. maritima*. Validated and predicted effectors of regulators are listed in red and black, respectively. Regulator binding sites and downstream regulated genes are shown by circles and arrows, respectively. Sequence logos representing the consensus binding site motifs were built using all candidate sites in the Thermotogae genomes. Genes encoding transcriptional regulators and components of sugar uptake transporters are shown in black and pink, whereas the genes encoding the secreted and intracellular sugar catabolic enzymes are in yellow and light green, respectively. Hypothetical genes are in gray. The *T. maritima* gene IDs are given inside the genes.

regulators in Thermotogae (Supplementary Table S1). In contrast with DNA motifs, specific effectors cannot be unambiguously predicted solely from genome analysis. Suggested candidate effectors of seven regulators from *T. maritima* were experimentally tested (as described in the next section), providing the basis for the analysis of structural determinants of effector specificity.

Experimental assessment of predicted DNA motifs and effectors of ROK regulators

Seven ROK-family regulators from *T. maritima*—BglR/TM0032, ManR/TM1224, XylR/TM0110, ChiR/TM0808, TreR/TM0393, GluR/TM1847 and IolR/TM0411—were assessed *in vitro* for their ability to recognize the predicted DNA operator sites and effectors. Each regulator gene was cloned and overexpressed in *E. coli*, and the recombinant protein was purified by Ni^{2+}-chelating chromatography. We used EMSA to test specific DNA binding of the purified ROK proteins to the synthetic DNA fragments containing the predicted 23-bp DNA sites from *T. maritima*. The results of all EMSA experiments are available in Supplementary Figure S3 and are briefly described below.

DNA motif binding specificities

All tested ROK regulators demonstrated a concentration-dependent formation of shifted DNA bands when tested with DNA fragments containing their predicted regulator-binding sites (Figure 4A; the complete results are provided in Supplementary Figure S3). As determined by EMSA experiments, the minimal effective concentration (MEC) of protein that is required to shift at least 80% of target DNA was in the 0.5–25 nM range (Table 1). The ChiR protein demonstrated the complete band shift at minimal concentration of 0.5 nM. For the XylR protein, the band shift for all five target sites was essentially complete at protein concentrations above 2.5–5.0 nM. The BglR and IolR proteins demonstrated similar affinities to their cognate DNA sites, producing a complete shift at 2.5 nM. For the ManR regulator, a high-scoring binding site at the *manB* gene and the low-scoring site at the *mtpE* gene showed the complete band shifts at protein concentrations above 10 and 25 nM, respectively. In summary, the results of *in vitro* binding assays provide experimental validation for all 14 predicted DNA binding sites of seven ROK-family regulators in *T. maritima*.

Effector specificities

To test the effects of various mono- and disaccharides on DNA binding of each ROK regulator, we performed additional EMSA experiments with a selected subset of DNA fragments containing regulator-binding sites (Supplementary Figure S3). The specific DNA-binding ability of all tested ROK regulators except IolR was abolished by one or several carbohydrates in a concentration-dependent manner (Table 2). The binding assays confirm that a disaccharide chitobiose but not monosaccharide GlcNAc is a specific effector of ChiR, a predicted regulator of the chitobiose utilization operon. The XylR regulator of xylan/xyloside utilization pathways in *Thermotoga* spp. demonstrated broader effector specificity: in addition to D-xylose (MEC 0.02 mM), the specific DNA-binding of XylR was also suppressed by D-glucose (but with MEC >0.2 mM). Interestingly, it was previously shown that D-glucose is weakly active as an inducer for the xylose-responsive repressor XylR in *B. subtilis* (62). The specific binding of the laminarin/β-glucosides utilization regulator BglR to its cognate DNA operator was affected by both monosaccharide D-glucose and disaccharide cellobiose with MEC 0.2 and 2.0 mM, respectively. The trehalose utilization regulator TreR showed strict preference for this disaccharide (MEC 0.2 mM) and did not respond to D-glucose, even at much higher concentrations. The predicted glucose-responsive regulator GluR, which controls the glucose and trehalose ABC transporters in *T. maritima*, was affected by D-glucose at 2 mM but was not affected by other tested sugars. Finally, the inositol catabolism regulator IolR did not respond to any tested sugars including *myo*-inositol, *myo*-inositol-1-phosphate, *scyllo*-inositol, the *myo*-inositol catabolic pathway intermediates (2-keto-*myo*-inositol, 5-keto-D-gluconate) and D-glucose (Supplementary Figure S3).

Structural determinants of effector specificity in the ROK family of regulators

Based on the combined bioinformatic and experimental analysis, we selected seven presumably isofunctional groups of orthologous ROK regulators in Thermotogae to study the residues that determine specificity of regulators toward their molecular effectors (Figure 1). The high sequence similarity between the ligand-binding domains of the studied regulators and four structurally characterized sugar kinases within the ROK family allows us to assume that the active site in kinases corresponds to the effector-binding site in the regulators from this protein family (Figure 4). Various orthologous groups of ROK regulators in Thermotogae have distinct patterns of residues compared with those that are present in the active site of kinases (Supplementary Table S5). Thus, the catalytically important ATP-binding motif of kinases (DxGxT) is altered in all but two groups. However, another feature important for kinase activity, the zinc-binding motif (CxCGxxGCxE/D), is largely conserved throughout the ROK family. Remarkably, the most significant variations among different groups of ROK regulators were observed for the residues implicated in sugar binding.

Using a multiple sequence alignment of the ligand-binding domains of 47 selected ROK regulators, we computed the SDRs that distinguish regulators from different groups (Supplementary Table S2). Seven SDRs that are presumably involved in effector recognition were selected by mapping into the tertiary structure of TM1224/BglR (Supplementary Figure S1). We also selected five additional sugar-binding residues conserved in the active site of ROK kinases that were not identified as SDRs but which are substituted in some groups of the analyzed ROK regulators (Supplementary Figure S4). The identified two groups of residues in the putative

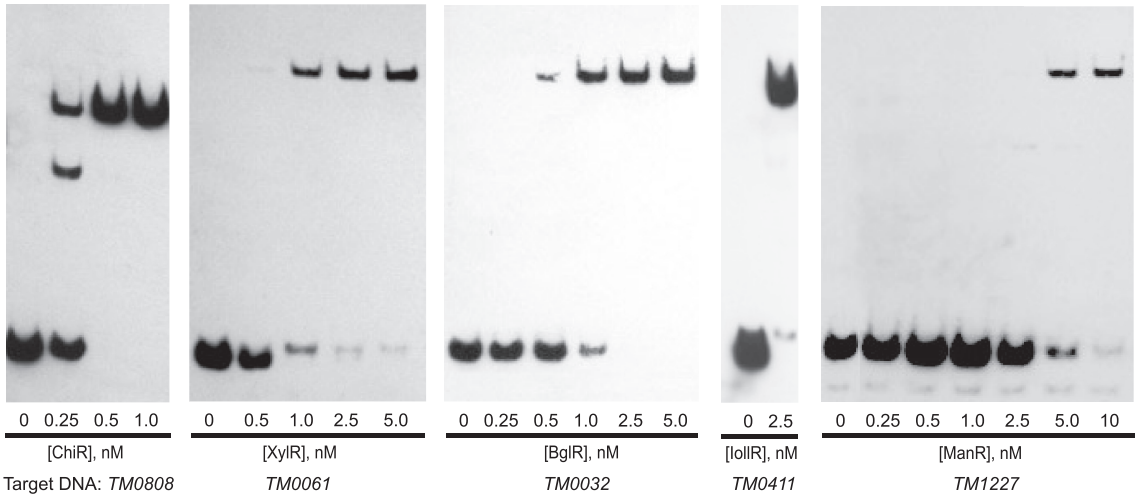

Figure 4. EMSA with *T. maritima* ROK-family regulators and their predicted DNA target fragments. Titration of ROK regulators for binding of their target DNA fragments (0.1 mM). EMSA was performed in the absence (lane 1) and in the presence of increasing protein concentrations.

Table 1. Validated binding motifs of ROK-family regulators in *T. maritima*

Regulator	Regulator-binding DNA motif			Regulated target gene[a]	EMSA validation	
	Binding site sequence	Distance to ATG	PWM score		DNA label[b]	MEC[c], nM
TM0110/XylR	GAAATTTCTTTAGAGGAAAAAAT	−45	5.28	*TM0309/xtpN*	biotin	2.5
	AATATTTCCCGAAAGGAAAAAAT	−53	5.67	*TM0071/xloE*	biotin	5.0
	ATAATTGATTGATAGAAAAAATT	−38	4.91	*TM1667/xylA*	biotin	5.0
	ATTATTTCCTGCATATAATTAAT	−61	5.33	*TM0110/ xylR*	biotin	2.5
	ATTTTTTCTTTACAAAAAATAAC	−87	5.61	*TM0061 xynA*	biotin	2.5
TM0808/ChiR	AAGTTGTTTGCGGCATGCAACTA	−27	6.85	*TM0808/chiR*	biotin	0.5
TM0032/BglR	AATTTCTTTCTGAGGAAGATAGA	−45	6.80	*TM0032/bglR*	biotin	2.5
TM0411/IolR	GTTGGTTAGTTAACGATAACAAA	−39	6.04	*TM0411/iolR*	biotin	2.5
TM1224/ManR	AAATAAGTAAAGTTTACTAATTA	−38	7.51	*TM1227/manB*	biotin	10
	TTATTAGTAAGTGTTATTTATTA	−50	5.15	*TM1746/mtpE*	biotin	25
TM0393/TreR	ATTaATTCAagTTACGAATAAAT	−39	5.99	*TM0392/treT*	biotin	10
	tTaTtTTCATTTAACGAAaAAAa	−138	5.80	*Thema_1380/treE*	fluo.	200
TM1847/GluR	ATTTAATTCCtTTGGAAaTTAAT	−121	6.28	*Thema_1377/gluE*	fluo.	100
	ATTTgATTACAAcGTcATTTAAc	−47	5.98	*Thema_1380/treE*	fluo.	100

[a]TM IDs and names of the first genes in candidate-regulated operons are indicated.
[b]Biotin-labeled 49-bp DNA fragments (0.1 nM) and fluorescence-labeled 33-bp DNA fragments (2 nM) were used in EMSA assays.
[c]Minimal effective concentration of regulators that is required to shift at least 80% of target DNA in EMSA experiments. For details, Supplementary Figure S3.

ligand-binding site of ROK regulators are represented as motif logos in Figure 1. The projection of these residues onto available tertiary structures of ROK kinases solved in complex with their cognate substrates, D-glucose or ManNAc, provides insights into the structural basis of their sugar ligand specificity (Supplementary Table S6 and Supplementary Figure S1).

The identified carbohydrate effectors of ROK regulators can be classified by their molecular size (e.g. mono- or disaccharide), structure (e.g. the type of the glycosidic bond in disaccharides) and orientation of hydroxyl groups (Supplementary Figure S5). Some of the observed variations in the ligand-binding site of ROK regulators were correlated with these features of sugar ligands (Figure 1). Thus, Asp226, which is involved in the recognition of the 6′-OH group of the sugar ligand (Supplementary Table S6), is conserved in all analyzed proteins with the exception of the XylR and IolR groups of regulators. In the xylose-responsive regulator XylR, Asp226 is substituted by glutamic acid that has a longer side-chain, thus compensating for the absence of the 6′-OH group in the xylose effector. In the inositol catabolism regulator IolR, the negatively charged Asp226 is substituted by the neutral asparagine. Pro181 in the active site loop is conserved in all analyzed proteins with the exception of TreR, where it is substituted by serine. The effector of TreR is an α-linked glucose disaccharide, trehalose, making it distinct from the effectors of other regulators. In the tertiary structure of TM1224 (2HOE), Pro181 interacts with residues in the sugar-binding β6-β7 loop, likely contributing to coordination of this loop for ligand recognition.

Table 2. Effectors of ROK-family regulators in *T. maritima* tested by EMSA

Regulator	(Protein) nM	Regulator target gene containing upstream binding site tested	Tested potential ligands		
			Sugar	(Sugar) mM	Effect[a]
TM0808/ChiR	1	*TM0808/ chiR*	Chitobiose	2–20	Partial to full
			Cellobiose	≫20	None
			Gentibiose	≫20	None
			D-Glucose	≫20	None
			GlcNAc	≫20	None
TM0110/XylR	2.5	*TM1667/ xylA*	D-Xylose	0.02–0.2	Partial to full
			D-Glucose	0.2–2.0	Partial to full
			D-Mannose	2–20	Partial to full
TM0032/BglR	2.5	*TM0032/ bglR*	D-Glucose	0.2	Full
			Cellobiose	0.2–2.0	Partial to full
			Chitobiose	≫2	None
			Gentibiose	≫20	Partial
TM0411/IolR	2.5	*TM0411/ iolR*	*myo*-Inositol	≫2	No
			scyllo-Inositol	≫2	No
			MI-Phosphate	≫2	No
			2-Keto-MI	≫2	No
			D-Glucose	≫2	No
TM1224/ManR	50	*TM1227/ manB*	D-Mannose	2.0	Full
			D-Xylose	≫20	No
			D-Glucose	≫20	No
TM0393/TreR	50	*TM0392/ treT*	Trehalose	0.2	Full
			D-Glucose	≫20	No
			Sucrose	≫20	No
TM1847/GluR	50	*Thema_1377/ gluE*	D-Glucose	>2.0	Partial
			D-Mannose	≫20	No
			Trehalose	≫20	No
			D-Xylose	≫20	No

[a]Potential effectors of ROK regulators were tested for their ability to abolish the protein-dependent shift of a target DNA fragment using EMSA experiments. For details, see Supplementary Figure S3.

Two signature residues located in the ligand recognition β6-β7 loop (positions 194 and 196) are presumably responsible for recognition of the 1′- and 2′-OH groups of the sugar ligand (Supplementary Table S6). Large variations in these two residues in the analyzed ROK regulators are correlated with the differences in the 1′- and 2′-OH groups of their corresponding effectors. The three groups of glucose-responsive regulators GluR, XylR and BglR have identical Ala194 and Asn196, whereas other groups of ROK regulators are characterized by other specific patterns of residues in these positions. Hence, the mannose-responsive regulator ManR is characterized by proline in position 194, whereas mannose differs from glucose in the orientation of the 1′- and 2′-OH groups (Supplementary Figure S5). Three signature residues in positions 116, 117 and 254, which are located in the ATP-binding loops in close vicinity to the ligand-binding site, demonstrate the patterns of variation throughout the specificity groups that correlate with particular properties of effectors. Gly116 is conserved in all studied regulators except ChiR and BglR that respond to β-linked disaccharides, whereas Gly254 is substituted by arginine in the GluR and XylR regulators, both of which respond to monosaccharides but not to disaccharides.

Among five sugar-binding residues conserved in the active site of ROK kinases, Gly182, Asn225, Glu277, His280 and Glu301, the first three are absolutely conserved in all studied ROK regulators, whereas the last two residues are substituted in two groups of regulators, ManR and ChiR (Figure 1). The conserved His280 contacts with the 2′-OH group of glucose in ROK kinases (Supplementary Table S6). The His280 substitution correlates with the replacement of the 2′-OH group with the acetylated 2′-NH group in chitobiose (for ChiR), and with an opposite orientation of the 2′-OH group in mannose (for ManR). Another conserved residue, Glu301 that contacts the 1′-OH group, is substituted by aspartic acid in five out of seven ChiR proteins, thus likely contributing to recognition of the chitobiose effector.

DISCUSSION

Sugar catabolic pathways in bacteria are often directly regulated by carbohydrate-responsive TFs that mediate induction of target catabolic operons in the presence of their cognate effectors, usually mono- or disaccharides (27,30,37,54,63,64). Marine bacteria from the Thermotogae phylum provides us with a remarkable case of a lineage-specific expansion and diversification of the carbohydrate catabolic machinery (17), where several families of sugar-related transcriptional regulators, such as ROK and LacI, are represented by multiple paralogs. The ROK family is the largest family of regulators in *Thermotoga* spp., and the model bacterium *T. maritima* alone has seven distinct ROK regulators (Figure 2). Analysis of distant homologs of the ROK regulators did not reveal their potential functional orthologs outside of

the Thermotogae phylum. A large-scale phylogenetic analysis of ROK-family regulators suggests that all Thermotogae-specific regulators are likely monophyletic (Supplementary Figure S6). These observations allowed us to propose that the expansion of ROK-family regulators in Thermotogae was likely due to massive duplications and subsequent specializations to take control over distinct sugar catabolic pathways. The knowledge of specific functions within a reference set of divergent family members is essential for understanding the molecular mechanisms of functional diversification in large families of regulators. Therefore, in this study, we combined cross-genome context analysis and experimental techniques to assess functions and specificities of the carbohydrate-responsive regulators from the ROK family in Thermotogae. As a result of the combined *in vitro* and *in silico* analyses, the identified and validated novel functional roles and effectors of ROK-family regulators in Thermotogae allowed us to largely expand the reference set of characterized regulators in this family (Figure 3). Interestingly, the phylogentic analysis of the ROK regulators with experimentally characterized specificities or the regulators whose functions were predicted via the comparative genomic analysis (Supplementary Figure S6) suggests a convergent evolution of regulators for certain sugar catabolic pathways. Thus, the phylogenetic tree of ROK regulators suggests that different non-orthologous regulators were independently recruited for transcriptional control of the xylose utilization pathways in Thermotogae, Firmicutes and α-proteobacteria, as well as for the *N*-acetylglucosamine and chitin utilization pathways in Thermotogae, γ-proteobacteria and Actinobacteria.

Based on the expanded reference set of ROK regulators, we performed the structure functional analysis aiming to understand the mechanisms of carbohydrate recognition by these regulators (Figure 1). Using multiple alignment of sugar-binding domains of ROK regulators and homologous sugar kinases, we inferred the signature SDR residues and the conserved residues that presumably play a key role in the recognition of the carbohydrates in the effector-binding site of the ROK family transcriptional regulators. Comparison with the available tertiary structures of ROK kinases bound to their sugar substrates allows us to propose that the sugar effector is surrounded by two groups of conserved residues and two groups of SDRs that are presumably located in the protein loops on the opposite sides of the ligand-binding site (Supplementary Figure S1). Two out of four SDR-containing loops (β4-β5 and β6-β7) are known to move toward the sugar-binding site when protein undergoes the conformational change upon ligand binding (19). Previous mutagenesis studies with the AlsK and NanK kinases have proved the importance of the latter two loops for the substrate specificity of these enzymes (65). We speculate that new specificities (except the XylR group) evolved without disturbing the conserved residues in the buried part of the binding site (positions 225 and 226) and the first side wall of the binding site formed by the conserved residues of the large domain (positions 277 and 280). The substitutions in the moving part of the binding site located in the β4-β5 and β6-β7 loops of the small domain apparently bring the major contribution into the effector specificity. Another group of SDRs located within the β1-β2 and β9-β10 loops in close vicinity to the sugar-binding site could be involved in the recognition of mono- and disaccharides serving as the gates for the sugar-binding site.

We combined the obtained results on the signature residues with phylogenetic analysis of the ROK family proteins and suggested the overall scenario for the evolution of ROK family TFs in the Thermotogae phylum. The largest group of TreR regulators has representatives in all but one Thermotogae genome (Figure 2). Conservation of this regulator suggests that it likely existed in the last common ancestor of Thermotogae. Representatives of the ChiR group of regulators are present in all *Thermotoga* and *Thermosipho* species but not in other Thermotogae, suggesting it has possibly emerged in the common ancestor of these two related lineages. Five other groups of ROK regulators were found only within the *Thermotoga* spp., suggesting their likely emergence by gene duplication in the common ancestor of this genus. Interestingly, three glucose-responsive regulators, GluR, XylR and BglR, form a single cluster on the phylogenetic tree and are characterized by highly similar SDR patterns (Figure 1). The established common response of these regulators to glucose (in addition to more specialized effectors) suggests that an ancestral regulator for this cluster has likely responded to glucose. The glucose-responsive GluR and xylose/glucose-responsive XylR proteins differ only in an SDR in position 226. We speculate that the Asp226–Glu226 substitution in the last common ancestor of XylR has expanded its specificity toward xylose. Similarly, the ability to recognize cellobiose by BglR can be attributed to concurrent substitutions in two SDRs in positions 116 and 254 (Figure 1). Thus, the specialization of duplicated ROK regulators in Thermotogae likely occurred via functional divergence of certain SDRs and conserved residues within the ligand-binding site. These results support the scenario of independent recruitment of non-orthologous regulators for the control of xylose and GlcNAc catabolic pathways in various taxonomic groups of bacteria.

The carbohydrate recognition by proteins is important to many biological processes including enzymatic reactions, transcriptional regulation and metabolite uptake. The determined specificities of ROK regulators in Thermotogae and the established correlations between the signature residues and the sugar ligands provide a framework for understanding the specificity of yet uncharacterized kinases and regulators in the ROK family. Site-directed mutagenesis of the identified SDRs and structural characterization of the ROK regulators in complexes with their cognate effectors are important directions for the future studies. The results of this analysis will enable the structure functional classification and evolutionary analysis of specificities for the entire ROK family, as it was recently performed for the FGGY family of sugar kinases (66). The integrated approach, which combines comparative structural and evolutionary analyses with other predictive bioinformatic techniques

such as reconstruction of pathways and regulons and focused experiments to test predicted functions, is generally applicable to other families of transcriptional regulators with diverse specificities.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–6 and Supplementary Figures 1–6.

FUNDING

The U.S. Department of Energy, Office of Science (Biological and Environmental Research), as part of Genomic Science Program [contract DE-FG02-08ER64686 with SBMRI and UCSD and contract DE-SC0004999 with SBMRI]; Russian Foundation for Basic Research [10-04-01768 and 12-04-33003 to D.A.R.]; State contract #8135 [application 2012-1.2.2-12-000-1013-079 to M.D.K.]; Russian Academy of Sciences (via the Molecular and Cellular Biology program to M.S.G.). Funding for open access charge: DOE [DE-SC0004999].

Conflict of interest statement. None declared.

REFERENCES

1. Galperin,M.Y. (2006) Structural classification of bacterial response regulators: diversity of output domains and domain combinations. *J. Bacteriol.*, **188**, 4169–4182.
2. Minezaki,Y., Homma,K. and Nishikawa,K. (2005) Genome-wide survey of transcription factors in prokaryotes reveals many bacteria-specific families not found in archaea. *DNA Res.*, **12**, 269–280.
3. Rivera-Gomez,N., Segovia,L. and Perez-Rueda,E. (2011) Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology*, **157**, 2308–2318.
4. Whitworth,D.E. and Cock,P.J. (2009) Evolution of prokaryotic two-component systems: insights from comparative genomics. *Amino Acids*, **37**, 459–466.
5. Charoensawan,V., Wilson,D. and Teichmann,S.A. (2010) Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res.*, **38**, 7364–7377.
6. Charoensawan,V., Wilson,D. and Teichmann,S.A. (2010) Lineage-specific expansion of DNA-binding transcription factor families. *Trends Genet.*, **26**, 388–393.
7. Fukami-Kobayashi,K., Tateno,Y. and Nishikawa,K. (2003) Parallel evolution of ligand specificity between LacI/GalR family repressors and periplasmic sugar-binding proteins. *Mol. Biol. Evol.*, **20**, 267–277.
8. Anantharaman,V. and Aravind,L. (2006) Diversification of catalytic activities and ligand interactions in the protein fold shared by the sugar isomerases, eIF2B, DeoR transcription factors, acyl-CoA transferases and methenyltetrahydrofolate synthetase. *J. Mol. Biol.*, **356**, 823–842.
9. Titgemeyer,F., Reizer,J., Reizer,A. and Saier,M.H. Jr (1994) Evolutionary relationships between sugar kinases and transcriptional repressors in bacteria. *Microbiology*, **140**, 2349–2354.
10. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
11. Plumbridge,J.A., Cochet,O., Souza,J.M., Altamirano,M.M., Calcagno,M.L. and Badet,B. (1993) Coordinated regulation of amino sugar-synthesizing and -degrading enzymes in *Escherichia coli* K-12. *J. Bacteriol.*, **175**, 4951–4956.
12. Plumbridge,J.A. (1991) Repression and induction of the nag regulon of *Escherichia coli* K-12: the roles of *nagC* and *nagA* in maintenance of the uninduced state. *Mol. Microbiol.*, **5**, 2053–2062.
13. Dahl,M.K., Degenkolb,J. and Hillen,W. (1994) Transcription of the xyl operon is controlled in *Bacillus subtilis* by tandem overlapping operators spaced by four base-pairs. *J. Mol. Biol.*, **243**, 413–424.
14. Scheler,A. and Hillen,W. (1994) Regulation of xylose utilization in *Bacillus licheniformis*: Xyl repressor-xyl-operator interaction studied by DNA modification protection and interference. *Mol. Microbiol.*, **13**, 505–512.
15. Dubeau,M.P., Poulin-Laprade,D., Ghinet,M.G. and Brzezinski,R. (2011) Properties of CsnR, the transcriptional repressor of the chitosanase gene, *csnA*, of *Streptomyces lividans*. *J. Bacteriol.*, **193**, 2441–2450.
16. Conejo,M.S., Thompson,S.M. and Miller,B.G. (2010) Evolutionary bases of carbohydrate recognition and substrate discrimination in the ROK protein family. *J. Mol. Evol.*, **70**, 545–556.
17. Rodionova,I.A., Yang,C., Li,X., Kurnasov,O.V., Best,A.A., Osterman,A.L. and Rodionov,D.A. (2012) Diversity and versatility of the *Thermotoga maritima* of sugar kinome. *J. Bacteriol.*, **194**, 5552–5563.
18. Nocek,B., Stein,A.J., Jedrzejczak,R., Cuff,M.E., Li,H., Volkart,L. and Joachimiak,A. (2011) Structural studies of ROK fructokinase YdhR from *Bacillus subtilis*: insights into substrate binding and fructose specificity. *J. Mol. Biol.*, **406**, 325–342.
19. Miyazono,K., Tabei,N., Morita,S., Ohnishi,Y., Horinouchi,S. and Tanokura,M. (2012) Substrate recognition mechanism and substrate-dependent conformational changes of an ROK family glucokinase from *Streptomyces griseus*. *J. Bacteriol.*, **194**, 607–616.
20. Mukai,T., Kawai,S., Mori,S., Mikami,B. and Murata,K. (2004) Crystal structure of bacterial inorganic polyphosphate/ATP-glucomannokinase. Insights into kinase evolution. *J. Biol. Chem.*, **279**, 50591–50600.
21. Nakamura,T., Kashima,Y., Mine,S., Oku,T. and Uegaki,K. (2012) Characterization and crystal structure of the thermophilic ROK hexokinase from *Thermus thermophilus*. *J. Biosci. Bioeng.*, **114**, 150–154.
22. Martinez,J., Nguyen,L.D., Hinderlich,S., Zimmer,R., Tauberger,E., Reutter,W., Saenger,W., Fan,H. and Moniot,S. (2012) Crystal structures of N-acetylmannosamine kinase provide insights into enzyme activity and inhibition. *J. Biol. Chem.*, **287**, 13656–13665.
23. Leyn,S.A., Li,X., Zheng,Q., Novichkov,P.S., Reed,S., Romine,M.F., Fredrickson,J.K., Yang,C., Osterman,A.L. and Rodionov,D.A. (2011) Control of proteobacterial central carbon metabolism by the HexR transcriptional regulator: a case study in *Shewanella oneidensis*. *J. Biol. Chem.*, **286**, 35782–35794.
24. Ravcheev,D.A., Li,X., Latif,H., Zengler,K., Leyn,S.A., Korostelev,Y.D., Kazakov,A.E., Novichkov,P.S., Osterman,A.L. and Rodionov,D.A. (2012) Transcriptional regulation of central carbon and energy metabolism in bacteria by redox-responsive repressor Rex. *J. Bacteriol.*, **194**, 1145–1157.
25. Rodionov,D.A. (2007) Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chem. Rev.*, **107**, 3467–3497.
26. Ravcheev,D.A., Best,A.A., Tintle,N., Dejongh,M., Osterman,A.L., Novichkov,P.S. and Rodionov,D.A. (2011) Inference of the transcriptional regulatory network in *Staphylococcus aureus* by integration of experimental and genomics-based evidence. *J. Bacteriol.*, **193**, 3228–3240.
27. Rodionov,D.A., Yang,C., Li,X., Rodionova,I.A., Wang,Y., Obraztsova,A.Y., Zagnitko,O.P., Overbeek,R., Romine,M.F., Reed,S. *et al.* (2010) Genomic encyclopedia of sugar utilization pathways in the *Shewanella* genus. *BMC Genomics*, **11**, 494.
28. Yang,C., Rodionov,D.A., Li,X., Laikova,O.N., Gelfand,M.S., Zagnitko,O.P., Romine,M.F., Obraztsova,A.Y., Nealson,K.H. and

Osterman,A.L. (2006) Comparative genomics and experimental characterization of N-acetylglucosamine utilization pathway of *Shewanella oneidensis*. *J. Biol. Chem.*, **281**, 29872–29885.
29. Rodionov,D.A., Mironov,A.A. and Gelfand,M.S. (2001) Transcriptional regulation of pentose utilisation systems in the *Bacillus*/*Clostridium* group of bacteria. *FEMS Microbiol. Lett.*, **205**, 305–314.
30. Gu,Y., Ding,Y., Ren,C., Sun,Z., Rodionov,D.A., Zhang,W., Yang,S., Yang,C. and Jiang,W. (2010) Reconstruction of xylose utilization pathway and regulons in Firmicutes. *BMC Genomics*, **11**, 255.
31. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Soding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
32. Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.
33. Huson,D.H., Richter,D.C., Rausch,C., Dezulian,T., Franz,M. and Rupp,R. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, **8**, 460.
34. Mironov,A.A., Vinokurova,N.P. and Gel'fand,M.S. (2000) Software for analyzing bacterial genomes. *Mol. Biol.*, **34**, 253–262.
35. Dehal,P.S., Joachimiak,M.P., Price,M.N., Bates,J.T., Baumohl,J.K., Chivian,D., Friedland,G.D., Huang,K.H., Keller,K., Novichkov,P.S. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.*, **38**, D396–D400.
36. Overbeek,R., Begley,T., Butler,R.M., Choudhuri,J.V., Chuang,H.Y., Cohoon,M., de Crecy-Lagard,V., Diaz,N., Disz,T., Edwards,R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.
37. Rodionov,D.A., Kurnasov,O.V., Stec,B., Wang,Y., Roberts,M.F. and Osterman,A.L. (2007) Genomic identification and in vitro reconstitution of a complete biosynthetic pathway for the osmolyte di-myo-inositol-phosphate. *Proc. Natl Acad. Sci. USA*, **104**, 4279–4284.
38. Novichkov,P.S., Rodionov,D.A., Stavrovskaya,E.D., Novichkova,E.S., Kazakov,A.E., Gelfand,M.S., Arkin,A.P., Mironov,A.A. and Dubchak,I. (2010) RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic Acids Res.*, **38**, W299–W307.
39. Plumbridge,J. (2001) Regulation of PTS gene expression by the homologous transcriptional regulators, Mlc and NagC, in *Escherichia coli* (or how two similar repressors can behave differently). *J. Mol. Microbiol. Biotechnol.*, **3**, 371–380.
40. Rodionov,D.A., Dubchak,I.L., Arkin,A.P., Alm,E.J. and Gelfand,M.S. (2005) Dissimilatory metabolism of nitrogen oxides in bacteria: comparative reconstruction of transcriptional networks. *PLoS Comput. Biol.*, **1**, e55.
41. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
42. Novichkov,P.S., Laikova,O.N., Novichkova,E.S., Gelfand,M.S., Arkin,A.P., Dubchak,I. and Rodionov,D.A. (2010) RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Res.*, **38**, D111–D118.
43. Kalinina,O.V., Mironov,A.A., Gelfand,M.S. and Rakhmaninova,A.B. (2004) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci.*, **13**, 443–456.
44. Mazin,P.V., Gelfand,M.S., Mironov,A.A., Rakhmaninova,A.B., Rubinov,A.R., Russell,R.B. and Kalinina,O.V. (2010) An automated stochastic approach to the identification of the protein specificity determinants and functional subfamilies. *Algorithms Mol. Biol.*, **5**, 29.
45. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
46. Karlin,S. and Brocchieri,L. (1996) Evolutionary conservation of RecA genes in relation to protein structure and function. *J. Bacteriol.*, **178**, 1881–1894.
47. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
48. Lesley,S.A., Kuhn,P., Godzik,A., Deacon,A.M., Mathews,I., Kreusch,A., Spraggon,G., Klock,H.E., McMullan,D., Shin,T. *et al.* (2002) Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. *Proc. Natl Acad. Sci. USA*, **99**, 11664–11669.
49. Osterman,A., Grishin,N.V., Kinch,L.N. and Phillips,M.A. (1994) Formation of functional cross-species heterodimers of ornithine decarboxylase. *Biochemistry*, **33**, 13662–13667.
50. Mossessova,E. and Lima,C.D. (2000) Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast. *Mol. Cell*, **5**, 865–876.
51. Yang,C., Rodionov,D.A., Rodionova,I.A., Li,X. and Osterman,A.L. (2008) Glycerate 2-kinase of *Thermotoga maritima* and genomic reconstruction of related metabolic pathways. *J. Bacteriol.*, **190**, 1773–1782.
52. Brune,I., Gotker,S., Schneider,J., Rodionov,D.A. and Tauch,A. (2012) Negative transcriptional control of biotin metabolism genes by the TetR-type regulator BioQ in biotin-auxotrophic *Corynebacterium glutamicum* ATCC 13032. *J. Biotechnol.*, **159**, 225–234.
53. Chen,X., Kohl,T.A., Ruckert,C., Rodionov,D.A., Li,L.H., Ding,J.Y., Kalinowski,J. and Liu,S.J. (2012) Phenylacetic acid catabolism and its transcriptional regulation in *Corynebacterium glutamicum*. *Appl. Environ. Microbiol.*, **78**, 5796–5804.
54. Leyn,S.A., Gao,F., Yang,C. and Rodionov,D.A. (2012) N-acetylgalactosamine utilization pathway and regulon in Proteobacteria. Genomic reconstruction and experimental characterization in Shewanella. *J. Biol. Chem.*, **287**, 28047–28056.
55. Choi,K.H., Seo,J.Y., Park,K.M., Park,C.S. and Cha,J. (2009) Characterization of glycosyl hydrolase family 3 beta-N-acetylglucosaminidases from *Thermotoga maritima* and *Thermotoga neapolitana*. *J. Biosci. Bioeng.*, **108**, 455–459.
56. Sarkar,S., Pramanik,A., Mitra,A. and Mukherjee,J. (2010) Bioprocessing data for the production of marine enzymes. *Marine Drugs*, **8**, 1323–1372.
57. Conners,S.B., Montero,C.I., Comfort,D.A., Shockley,K.R., Johnson,M.R., Chhabra,S.R. and Kelly,R.M. (2005) An expression-driven approach to the prediction of carbohydrate transport and utilization regulons in the hyperthermophilic bacterium *Thermotoga maritima*. *J. Bacteriol.*, **187**, 7267–7282.
58. Nanavati,D.M., Thirangoon,K. and Noll,K.M. (2006) Several archaeal homologs of putative oligopeptide-binding proteins encoded by *Thermotoga maritima* bind sugars. *Appl. Environ. Microbiol.*, **72**, 1336–1345.
59. Boucher,N. and Noll,K.M. (2011) Ligands of thermophilic ABC transporters encoded in a newly sequenced genomic region of *Thermotoga maritima* MSB8 screened by differential scanning fluorimetry. *Appl. Environ. Microbiol.*, **77**, 6395–6399.
60. Qu,Q., Lee,S.J. and Boos,W. (2004) TreT, a novel trehalose glycosyltransferring synthase of the hyperthermophilic archaeon *Thermococcus litoralis*. *J. Biol. Chem.*, **279**, 47890–47897.
61. Frock,A.D., Gray,S.R. and Kelly,R.M. (2012) Hyperthermophilic *Thermotoga* species differ with respect to specific carbohydrate transporters and glycoside hydrolases. *Appl. Environ. Microbiol.*, **78**, 1978–1986.
62. Dahl,M.K., Schmiedel,D. and Hillen,W. (1995) Glucose and glucose-6-phosphate interaction with Xyl repressor proteins from *Bacillus* spp. may contribute to regulation of xylose utilization. *J. Bacteriol.*, **177**, 5467–5472.
63. Zhang,L., Leyn,S.A., Gu,Y., Jiang,W., Rodionov,D.A. and Yang,C. (2012) Ribulokinase and transcriptional regulation of arabinose metabolism in *Clostridium acetobutylicum*. *J. Bacteriol.*, **194**, 1055–1064.

64. Rodionov,D.A., Gelfand,M.S. and Hugouvieux-Cotte-Pattat,N. (2004) Comparative genomics of the KdgR regulon in *Erwinia chrysanthemi* 3937 and other gamma-proteobacteria. *Microbiology*, **150**, 3571–3590.
65. Larion,M., Moore,L.B., Thompson,S.M. and Miller,B.G. (2007) Divergent evolution of function in the ROK sugar kinase superfamily: role of enzyme loops in substrate specificity. *Biochemistry*, **46**, 13564–13572.
66. Zhang,Y., Zagnitko,O., Rodionova,I., Osterman,A. and Godzik,A. (2011) The FGGY carbohydrate kinase family: insights into the evolution of functional specificities. *PLoS Comput. Biol.*, **7**, e1002318.