Published in final edited form as:
J Struct Funct Genomics. 2014 June ; 15(2): 73–81. doi:10.1007/s10969-014-9180-3.

The crystal structure of pyrimidine/thiamin biosynthesis precursor-like domain-containing protein CAE31940 from proteobacterium *Bordetella bronchiseptica* RB50, and evolutionary insight into the NMT1/THI5 family

Jacek Bajor,
Department of Molecular Physiology and Biological Physics, University of Virginia, 1340 Jefferson Park Avenue, Charlottesville, VA 22908, USA. Midwest Center for Structural Genomics, USA, URL: http://www.mcsg.anl.gov/

Karolina L. Tkaczuk,
Department of Molecular Physiology and Biological Physics, University of Virginia, 1340 Jefferson Park Avenue, Charlottesville, VA 22908, USA. Midwest Center for Structural Genomics, USA, URL: http://www.mcsg.anl.gov/

Maksymilian Chruszcz,
Department of Molecular Physiology and Biological Physics, University of Virginia, 1340 Jefferson Park Avenue, Charlottesville, VA 22908, USA. Midwest Center for Structural Genomics, USA, URL: http://www.mcsg.anl.gov/

Hutton Chapman,
Department of Molecular Physiology and Biological Physics, University of Virginia, 1340 Jefferson Park Avenue, Charlottesville, VA 22908, USA. Midwest Center for Structural Genomics, USA, URL: http://www.mcsg.anl.gov/

Olga Kagan,
Midwest Center for Structural Genomics, USA, URL: http://www.mcsg.anl.gov/. Banting and Best Department of Medical Research, University of Toronto, 112 College Street, Toronto, ON M5G 1L6, Canada

Alexei Savchenko, and
Midwest Center for Structural Genomics, USA, URL: http://www.mcsg.anl.gov/. Banting and Best Department of Medical Research, University of Toronto, 112 College Street, Toronto, ON M5G 1L6, Canada

Wladek Minor
Department of Molecular Physiology and Biological Physics, University of Virginia, 1340 Jefferson Park Avenue, Charlottesville, VA 22908, USA. Midwest Center for Structural Genomics, USA, URL: http://www.mcsg.anl.gov/

Jacek Bajor and Karolina L. Tkaczuk authors contributed equally to the project.

Present Address:
M. Chruszcz, Department of Chemistry and Biochemistry, University of South Carolina, 631 Sumter Street, Columbia, SC 29208, USA

Wladek Minor: wladek@iwonka.med.virginia.edu

Abstract

We report a 2.0 Å structure of the CAE31940 protein, a proteobacterial NMT1/THI5-like domain-containing protein. We also discuss the primary and tertiary structure similarity with its homologs. The highly conserved FGGXMP motif was identified in CAE31940, which corresponds to the GCCCX motif located in the vicinity of the active center characteristic for THi5-like proteins found in yeast. This suggests that the FGGXMP motif may be a unique hallmark of proteobacterial NMT1/THI5-like proteins.

Keywords

Bordetella bronchiseptica RB50; NMT1/THI5-like domain-containing protein; Crystal structure; MCSG

Introduction

Thiamin (vitamin B1) consists of two components: the pyrimidine moiety (4-amino-5-hydroxymethyl-2-methylpyrimidine) and the thiazole moiety (5-(2-hydroxyethyl)-4-methylthiazole). The two moieties are produced by two separate biosynthetic processes, which are then covalently linked to yield thiamin phosphate [1, 2]. This process is well studied in prokaryotes but is still poorly understood in eukaryotes. Thiamin synthesis has been studied to some degree in yeast; in *Saccharomyces cerevisiae* the *thi5* gene product THi5 is responsible for the synthesis of 4-amino-5-(hydroxymethyl)-2-methylpyrimidine phosphate in yeast [3–5]. THi5 appears to be conserved in eukaryotes with thiamin biosynthetic pathways [3–5].

THi5 belongs to a large superfamily known as the NMT1/THI5-like domain proteins (PFam entry PF09084, comprising 7,204 sequences). However, the majority of members of the NMT1/THI5-like superfamily are found in eubacteria, especially *Proteobacteria* (4,295 sequences in 1,354 species). While there is some structural information for the superfamily —for example, a homolog in *Bacillus halodurans*—the family is very diverse in sequence and few structural representatives are known, particularly among proteobacteria. Presented here are the structure and analyses of the two-domain protein CAE31940 from *Bordetella bronchiseptica* RB50 containing pyrimidine/thiamin biosynthesis precursor-like domain, which shed new light on potential proteins taking part in thiamin biosynthesis in this organism.

Materials and methods

Cloning, expression and purification

Selenomethionine (Se-Met) substituted CAE31940 protein was produced using standard MSCG protocols as described by Zhang et al. [6]. Briefly, gene BB1442 from *Bordetella bronchiseptica* RB50 was cloned into a p15TV LIC plasmid using ligation independent cloning [7–9]. The gene was overexpressed in *E. coli* BL21-CodonPlus(DE3)-RIPL cells in

Se-Met-containing LB media at 37.0 °C until the optical density at 600 nm reached 1.2. Then the cells were induced by isopropyl-β-D-1-thiogalactopyranoside, incubated at 20.0 °C overnight, and pelleted by centrifugation. Harvested cells were sonicated in lysis buffer (300 mM NaCl, 50 mM HEPES pH 7.5, 5 % glycerol, and 5 mM imidazole), the lysed cells were spun down for 15 min at 16,000 RPM and the supernatant was applied to a nickel chelate affinity resin (Ni–NTA, Qiagen). The resin was washed with wash buffer (300 mM NaCl, 50 mM HEPES pH 7.5, 5 % glycerol, and 30 mM imidazole) and the protein was eluted using elution buffer (300 mM NaCl, 50 mM HEPES pH 7.5, 5 % glycerol, and 250 mM imidazole). The N-terminal polyhistidine tag (His-Tag) was removed by digestion with recombinant TEV protease and the digested protein was passed through a second affinity column. The flow through was dialyzed against a solution containing 300 mM NaCl, 10 mM HEPES pH 7.5 and 1 mMTCEP. Purified protein was concentrated to 36 mg/mL and flash-frozen in liquid nitrogen.

Crystallization

Crystals of CAE31940 used for data collection were grown by the sitting drop vapor diffusion method. The well solution consisted of 0.2 M ammonium acetate, 30 % w/v PEG4000, and 0.1 M tri-sodium citrate at pH 5.6. Crystals were grown at 293 K and formed after 1 week of incubation. Immediately after harvesting, crystals were transferred into cryoprotectant solution (Paratone-N) without mother liquor, washed twice in the solution and flash cooled in liquid nitrogen.

Data collection and processing

Data were collected at 100 K at the 19-ID beamline (ADSC Q315 detector) of the Structural Biology Center [10] at the Advanced Photon Source (Argonne National Laboratory, Argonne, Illinois, USA). The beamline was controlled by HKL-3,000 [11]. Diffraction data were processed with HKL-2,000 [11]. Data collection, structure determination, and refinement statistics are summarized in Table 1.

Structure solution and refinement

The structure of the Se-Met-substituted protein was solved using single-wavelength anomalous diffraction (SAD), and an initial model was built with HKL-3000. HKL-3000 is integrated with SHELXC/D/E [12], MLPHARE, DM, ARP/wARP, CCP4 [13], SOLVE, and RESOLVE [14]. The resulting model was further refined with REFMAC5 [15], and COOT [16]. MOLPROBITY [17] and ADIT [18] were used for structure validation. The coordinates and experimental structure factors were deposited to PDB with accession code 3QSL.

Bioinformatics analyses

Sequence homology searches were performed with PSI-BLAST [19], and structural homology searches were done with HHpred [20, 21] with amino acid sequence of CAE31940 as a seed and FATCAT Database Searches using the crystal structure of CAE31940 with PDB ID 3QSL. Structures were superposed with SSM [22]. The evolutionary history was inferred using the neighbor-joining method [23]. The bootstrap

consensus tree inferred from 500 replicates [24] is taken to represent the evolutionary history of the taxa analyzed [24]. Branches corresponding to partitions reproduced in less than 50 % of bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [24]. The evolutionary distances were computed using the Poisson correction method [25] and are in units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset (complete deletion option). There were a total of 237 positions in the final dataset. Phylogenetic analyses were conducted with MEGA5 [26].

Homology modeling of Thi5 from *Saccharomyces cerevisiae*

The "FRankenstein's Monster" method (in short, best fragment assembly mode) was used to construct a homology model of Thi5 from *Saccharomyces cerevisiae* S288c using the CAE31940 structure as a template. The method comprises cycles of local realignments in uncertain regions, building of alternative models and their evaluation, realignments in poorly scored regions, and merging of the best scoring fragments [27, 28]. PROQ [28] and MetaMQAP were used for the evaluation of models [29], which allowed the prediction of the deviations of individual residues in the homology model from their counterparts in the template structure.

Results and discussion

Overall structure

The two domain structure of the CAE31940 protein from *Bordetella bronchiseptica* RB50 was solved at a resolution of 2.0 Å. The protein crystallized in the orthorhombic $P2_12_12_1$ space group with two polypeptide chains in the asymmetric unit, which had the following unit cell dimensions: a = 60.9 Å, b = 65.0 Å, and c = 160.7 Å (all data collection and structure refinement parameters and statistics are shown in Table 1). The protein consists of 346 residues. Due to the lack of well-defined electron density, 27 amino acids at the N-terminus of chain A and 19 at the N-terminus of chain B could not be modeled. The final C-terminal residue of each chain is also missing in the solved structure.

The tertiary structure is composed of two β-sheets consisting of 5 β-strands each. The first one includes β strands 2↑-1↑-3↑-10↓-4↑ (the strands are numbered by their order in the primary sequence) and is flanked by 13 α-helices. The other β-sheet is composed of β-strands 7↑-6↑-8↑-5↓-9↑ and is surrounded by 6 α-helices. The overall structure of the protein is shown in Fig. 1.

According to the predictions of the PISA server, the CAE31940 protein is monomeric in solution.

Evolutionary relatives and characteristics of CAE31940

Sequential and structural comparisons—Standard sequence searches with PSI-BLAST against a non-redundant protein sequence database found no homolog of CAE31940 with a known structural model. These searches, however, clearly showed that the most

similar relatives (by sequence) of CAE31940 are NMT1/THI5-like domain-containing proteins. Thus subsequent structural homology searches were performed.

In order to find the most similar structural homologs of CAE31940, HHpred [21] and FATCAT [30] searches of the most up-to-date PDB database available (pdb70-Dec11) were performed. Their results show that CAE31940 structurally resembles eight proteins with an HHpred probability of 100 %: the alkanesulfonate-binding protein SsuA from *Xanthomonas axonopodis* (PDB code: 3KSX), periplasmic aliphatic sulphonate binding protein SsuA from *Escherichia coli* (PDB code: 2X26), ThiY periplasmic N-formyl-4-amino-5-aminomethyl-2-methylpyrimidine binding protein from *Bacillus halodurans* (PDB code: 3IX1), DSZB C27S mutant in complex with 2′-hydroxybiphenyl-2-sulfinic acid from *Rhodococcus* sp. (PDB code: 2DE3), periplasmic nitrate-binding protein NrtA from *Synechocystis* PCC 6803 (PDB code: 2G29), protein from *Gloeobacter violaceus* (PDB code: 2XQ7), ABC transporter (BDI_1369) from *Parabacteroides distasonis* (PDB code 3HN0), and bicarbonate transport protein CmpA from *Synechocystis* sp. PCC 6803 (PDB code: 2I49). Three of them (3KSX, 2X26, and 3IX1) belong to the NMT1/THI5-like family according to the PFam [31] annotations in the PDB database, and the remaining five are not assigned to any family. All of the identified structural homologs share comparatively low sequence identity with CAE31940, and according to the authors' annotations for each, all act on different substrates.

Of the 5 best-scoring hits in the structural homology searches, the protein with the most similar sequence to CAE31940 is ThiY, as shown on Fig. 2, and also has the lowest RMSD resulting from structural superposition (Table 2). A recent paper by Bale and co-workers [3] discusses two structural homologs, ThiY (HMP binding protein; PDB code: 3IX1) and Thi5 (HMP-P synthase). (Though the crystal structure of Thi5 is not yet available, the authors created a homology model and discussed the similarities of the two proteins.) CAE31940 shares 24 and 26 % sequence identity and 4 and 46 % sequence similarity with ThiY and Thi5 respectively. This corresponds to BLAST expectation values of 1×10^{-4} and 3×10^{-13} respectively. This is rather high sequence similarity, and considering that the structure of ThiY is one of the most structurally similar proteins to CAE31940, ThiY superimposes on CAE31940 with an RMSD of 2.5 Å. It is also confirmed by the results of FATCAT structural homolog searches (using FATCAT Database Search utility) where ThiY ranks as the best hit for CAE31940 (FATCAT raw score is 603.22) and RMSD of 2.7 Å (as presented in Table 2). This may suggest that all three proteins are orthologs and originate from the same ancestor, which is further supported by the evolutionary analysis presented below.

Structural alignment of the collected structures shows that they superimpose poorly (apart from their central β-sheet): even after removing the most variable regions (β5-β9 and α6-α11) (Fig. 3b), the pairwise RMSD of the most structurally similar proteins is 2.55 Å (as shown in Table 2; Fig. 3). However, the motif FGGXMP in CAE31940 corresponds to the previously identified GCCCX motif in Thi5 from *Saccharomyces cerevisiae* [3] (conserved in the THI5 family of enzymes) when CAE31940 and the homology model of Thi5 are super-imposed the aforementioned motifs are similar in both configuration and in their location in the 3-D structure (Fig. 4). Sequence analysis also shows that the FGGXMP motif

is conserved in all proteobacterial sequences related to CAE31940. Therefore this motif may be a hallmark of proteobacterial members of NMT1/THI5-like domain-containing proteins. The exact role of GCCCX motif in THI5 family of enzymes is not well described. Bale et al. [3] in their work mention that it is conserved in this family and may behave like related cysteine rich sequences (e.g., CX3CX2C) and bind [4Fe-4S] clusters involved in electron transfer or radical reactions.

Evolutionary insight into NMT1/THI5 family—The cladogram presented in Fig. 5 and the multiple sequence alignment (MSA) in Fig. 2 clearly shows three well divided groups (marked in different colors on the MSA). Group A (blue group in Fig. 2) is composed of mainly α/β proteobacterial proteins including the target protein CAE31940, characterized by the FGGX(M/N)P motif. Group B (black group in Fig. 2) contains Thi5-like fungal proteins, with a GCCCX motif instead of the aforementioned FGGXMP motif, as well as ThiY (PDB code: 3IX1). As described previously by Bale and co-workers [3], ThiY is closely related to Thi5. Group C (green group in Fig. 2) contains structural homologs (mainly SsuA) of CAE31940, which act on the sulfonate group in proteins. Group C proteins lack both of the sequential motifs present in groups A and B.

As expected, the Thi5 and ThiY proteins fall into the same group (group B) in the cladogram, together with other fungal homologs of Thi5. The conserved *Candida albicans* CA3427 gene product (PDB code: 2X7Q) groups together with them. CA3427, as described by the authors of its structure [32], represents a new family of proteins exhibiting a generic periplasmic binding protein structural fold. Unlike Thi5 and ThiY, it contains neither the conserved motif nor conserved residues responsible for pyrimidine ring binding. However, it is a fungal protein, like most of the others in group B.

Almost all of the homologs of known structure are more remotely related (i.e. in groups B and C) to CAE31940 than the more similar homologs identified in group A (Fig. 5.), which were annotated as NMT1/THI5-like domain-containing proteins (blue group in Fig. 2). That can partially be explained by the source organisms of the proteins grouping with CAE31940—they are all proteobacteria, while almost all of the other homologs of known structure come from other phyla. The only exception is a protein from *Xanthomonas axonopodis* (PDB code: 3KSX) from γ-proteobacteria which falls into a separate branch (group B) in Fig. 5. The *X. axonopodis* protein belongs to the SsuA family and contains neither of the conserved sequential motifs identified. Group C forms a well-separated branch with high statistical support which is far apart from NMT1/THI5 group A. This indicated that the described structure of CAE31940 is the first crystal structure of a NMT1/THI5-like domain-containing protein with the FGGXMP motif conserved in proteobacteria. The function of this domain is still unknown, but due to its structural and sequence similarity to the pyrimidine/thiamin biosynthesis precursor proteins Thi5 and ThiY, it might serve a similar enzymatic function.

Protein data bank accession code

Coordinates and structure factors of CAE31940 have been deposited in the RCSB Protein Data Bank with accession code 3QSL.

Acknowledgments

This research was funded with Federal funds from the National Institute of Health under grants U54-GM074942 and U54-GM094585. The results shown in this report are derived from work performed at Argonne National Laboratory, at the Structural Biology Center of the Advanced Photon Source. Argonne is operated by University of Chicago Argonne, LLC, for the U.S. Department of Energy, Office of Biological and Environmental Research under contract DE-AC02-06CH11357. We thank Dr. Matthew D. Zimmerman for critically reading the manuscript.

References

1. Zurlinden A, Schweingruber ME. Cloning, nucleotide sequence, and regulation of Schizosaccharomyces pombe thi4, a thiamine biosynthetic gene. J Bacteriol. 1994; 176:6631–6635. [PubMed: 7961415]
2. Begley TP, Chatterjee A, Hanes JW, Hazra A, Ealick SE. Cofactor biosynthesis–still yielding fascinating new biological chemistry. Curr Opin Chem Biol. 2008; 12(2):118–125. [PubMed: 18314013]
3. Bale S, Rajashankar KR, Perry K, Begley TP, Ealick SE. HMP binding protein ThiY and HMP-P synthase THI5 are structural homologues. Biochemistry. 2010; 49(41):8929–8936. [PubMed: 20873853]
4. Maundrell K. nmt1 of fission yeast. A highly transcribed gene completely repressed by thiamine. J Biol Chem. 1990; 265(19):10857–10864. [PubMed: 2358444]
5. Wightman R, Meacock PA. The THI5 gene family of Saccharomyces cerevisiae: distribution of homologues among the hemiascomycetes and functional redundancy in the aerobic biosynthesis of thiamin from pyridoxine. Microbiology. 2003; 149(Pt 6):1447–1460. [PubMed: 12777485]
6. Zhang RG, Skarina T, Katz JE, Beasley S, Khachatryan A, Vyas S, Arrowsmith CH, Clarke S, Edwards A, Joachimiak A, et al. Structure of thermo toga maritima stationary phase survival protein SurE: a novel acid phosphatase. Structure. 2001; 9(11):1095–1106. [PubMed: 11709173]
7. Eschenfeldt WH, Lucy S, Millard CS, Joachimiak A, Mark ID. A family of LIC vectors for high-throughput cloning and purification of proteins. Methods Mol Biol. 2009; 498:105–115. [PubMed: 18988021]
8. Aslanidis C, de Jong PJ. Ligation-independent cloning of PCR products (LIC-PCR). Nucleic Acids Res. 1990; 18(20):6069–6074. [PubMed: 2235490]
9. Haun RS, Serventi IM, Moss J. Rapid, reliable ligation-independent cloning of PCR products using modified plasmid vectors. Biotechniques. 1992; 13(4):515–518. [PubMed: 1362067]
10. Rosenbaum G, Alkire RW, Evans G, Rotella FJ, Lazarski K, Zhang RG, Ginell SL, Duke N, Naday I, Lazarz J, et al. The Structural Biology Center 19ID undulator beamline: facility specifications and protein crystallographic results. J Synchrotron Radiat. 2006; 13(Pt 1):30–45. [PubMed: 16371706]
11. Minor W, Cymborowski M, Otwinowski Z, Chruszcz M. HKL-3000: the integration of data reduction and structure solution—from diffraction images to an initial model in minutes. Acta Crystallogr D Biol Crystallogr. 2006; 62:859–866. [PubMed: 16855301]
12. Sheldrick GM. A short history of SHELX. Acta Crystallogr A. 2008; 64:112–122. [PubMed: 18156677]
13. Bjellqvist B, Basse B, Olsen E, Celis JE. Reference points for comparisons of 2-dimensional maps of proteins from different human cell-types defined in a Ph scale where isoelectric points correlate with polypeptide compositions. Electrophoresis. 1994; 15(3–4):529–539. [PubMed: 8055880]
14. Terwilliger T. SOLVE and RESOLVE: automated structure solution, density modification, and model building. J Synchrotron Radiat. 2004; 11:49–52. [PubMed: 14646132]
15. Murshudov GN, Skubak P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA. REFMAC5 for the refinement of macromolecular crystal structures. Acta Crystallogr D Biol Crystallogr. 2011; 67(Pt 4):355–367. [PubMed: 21460454]
16. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr. 2004; 60:2126–2132. [PubMed: 15572765]

17. Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D. 2010; 66:12–21. [PubMed: 20057044]
18. Yang HW, Guranovic V, Dutta S, Feng ZK, Berman HM, Westbrook JD. Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank. Acta Crystallogr D Biol Crystallogr. 2004; 60:1833–1839. [PubMed: 15388930]
19. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25(17):3389–3402. [PubMed: 9254694]
20. Soding J. Protein homology detection by HMM–HMM comparison. Bioinformatics. 2005; 21(7): 951–960. [PubMed: 15531603]
21. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005; 33(Web Server issue):W244–W248. [PubMed: 15980461]
22. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr. 2004; 60(Pt 12 Pt 1):2256–2268. [PubMed: 15572779]
23. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4(4):406–425. [PubMed: 3447015]
24. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 1985; 39:783–791.
25. Zuckerkandl, E.; Pauling, L. Evolutionary divergence and convergence in proteins. Academic Press; New York: 1965.
26. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011; 28:2731–2739. [PubMed: 21546353]
27. Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM. A “FRankenstein’s monster” approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. Proteins. 2003; 53(Suppl 6):369–379. [PubMed: 14579325]
28. Wallner B, Elofsson A. Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins. 2007; 69(Suppl 8):184–193. [PubMed: 17894353]
29. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM. Meta-MQAP: a meta-server for the quality assessment of protein models. BMC Bioinformatics. 2008; 9:403. [PubMed: 18823532]
30. Li Z, Ye Y, Godzik A. Flexible structural neighbourhood: a database of proteins structural similarities and alignments. Nucleic Acids Res. 2006; 34:D277–D280. [PubMed: 16381864]
31. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. 1997; 28(3):405–420. [PubMed: 9223186]
32. Santini S, Claverie JM, Mouz N, Rousselle T, Maza C, Monchois V, Abergel C. The conserved Candida albicans CA3427 gene product defines a new family of proteins exhibiting the generic periplasmic binding protein structural fold. PLoS ONE. 2011; 6(4):e18528. [PubMed: 21494601]

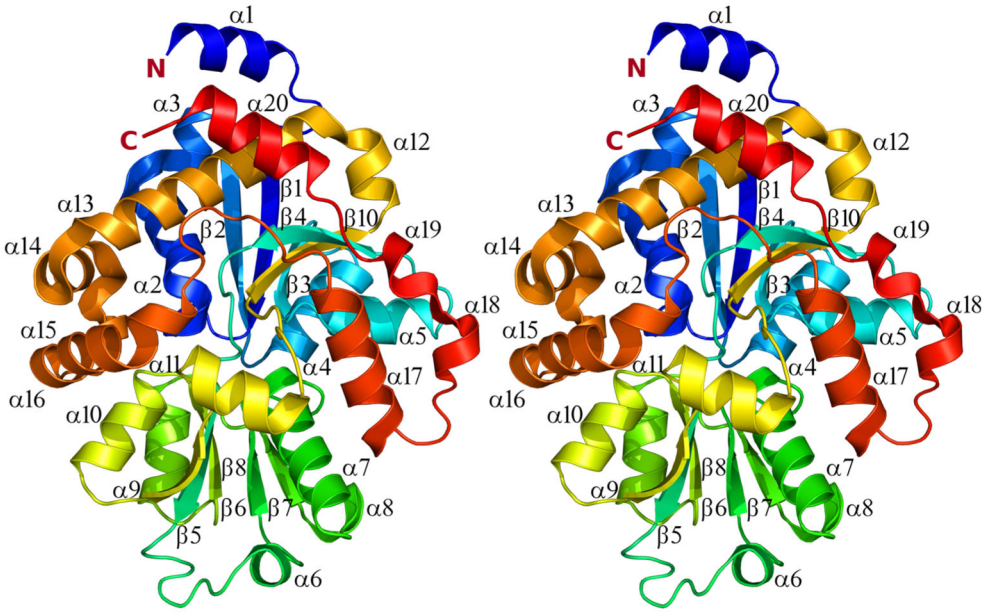

Fig. 1.

Structure of CAE31940 from *Bordetella bronchiseptica* (PDB code: 3QSL) shown in cross-eyed stereo. Structure is colored by primary sequence, from *dark blue* at the N-terminus to red at the C-terminus. The N- and C-termini, as well as the secondary structure elements, are labeled

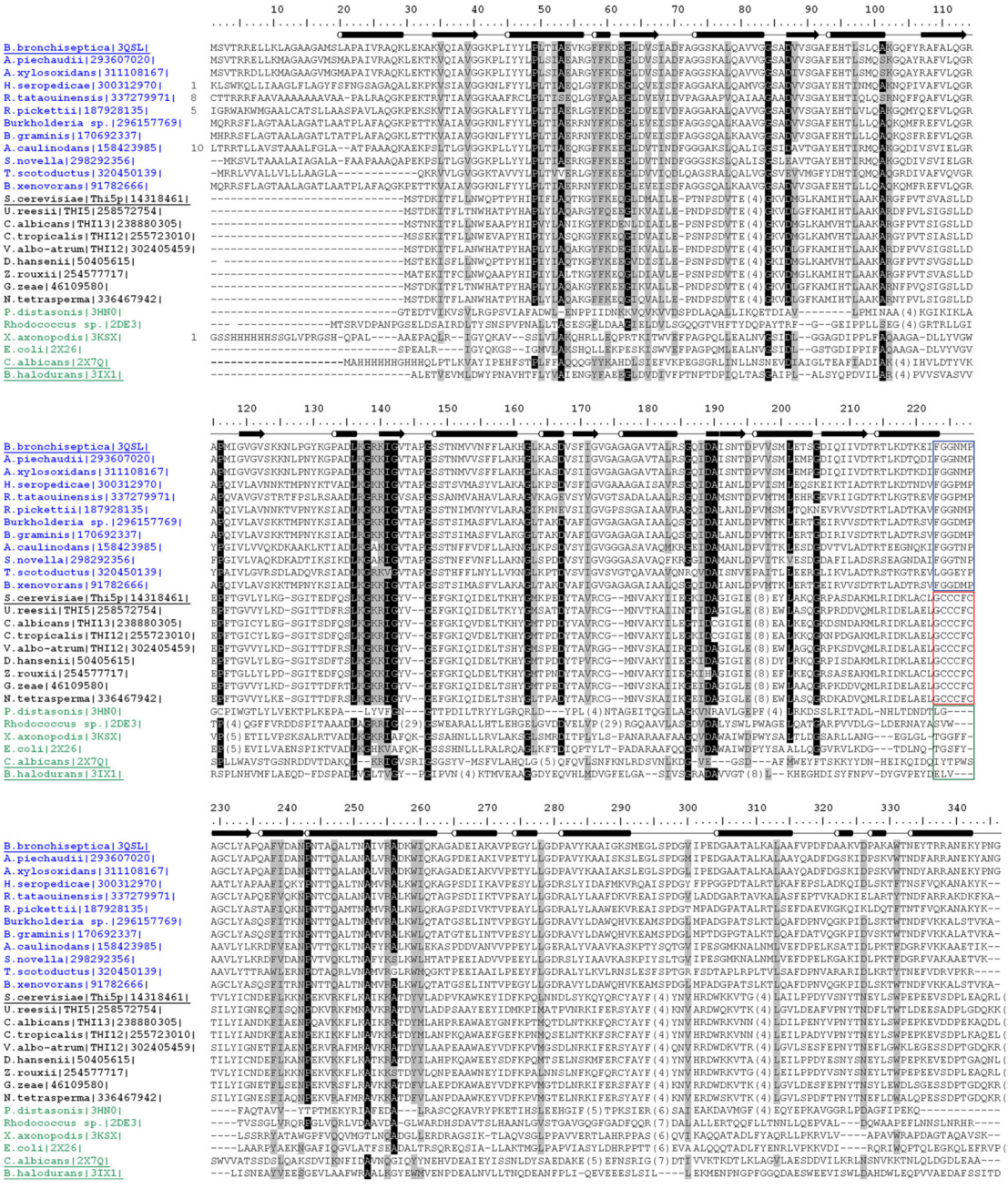

Fig. 2.

Multiple sequence alignment (MSA). The secondary structure of 3QSL is marked above the MSA. Extra residues for some sequences are removed from the alignment at the N- and C-termini for conciseness. The number of residues removed for each sequence is marked at either end of the MSA. Here are the full names of the organisms included in the MSA. The *blue group Bordetella Bronchiseptica* GI:326634536, *Achromobacter piechaudii* GI: 293607020, *Achromobacter xylosoxidans* GI:311108167, *Herbaspirillum seropedicae* GI: 300312970, Ramlibacter *tataouinensis* GI:337279971, *Ralstonia pickettii* GI:187928135, *Burkholderia sp.* GI:296157769, *Burkholderia graminis* GI:170692337, *Azorhizobium*

caulinodans GI:158423985, *Starkeya novella* GI:298292356, *Thermus scotoductus* GI: 320450139, *Burkholderia xenovorans* GI:91782666; the *black group: Saccharomyces cerevisiae*, Thi5, GI:14318461, *Uncinocarpus reesii*, Thi5, GI: 258572754, *Candida albicans*, Thi13, GI:238880305, *Candida tropicalis*, Thi12, GI:255723010, *Verticillium albo-atrum*, Thi12, GI: 302405459, *Debaryomyces hansenii*, GI:50405615, *Zygosaccharomyces rouxii*, GI: 254577717, *Gibberella zeae*, GI: 46109580, *Neurospora tetrasperma*, GI: 336467942; the *green group: Parabacteroides distasonis*, 3HN0, GI: 251837040, *Rhodococcus sp.*, 2DE3, GI: 112490451, *Xanthomonas axonopodis*, 3KSX, GI: 308387823, *Escherichia coli*, 2X26, GI: 294662291, *Candida albicans*, 2X7Q, GI: 326634033, *Bacillus halodurans*, 3IX1, GI: 308387786. Conserved residues are *shaded* when there is >80 % sequence similarity/identity. Similar residues are *shaded in grey* while identical residues are *shaded in black*. The sequence motifs discussed in the text are *boxed* for better visibility

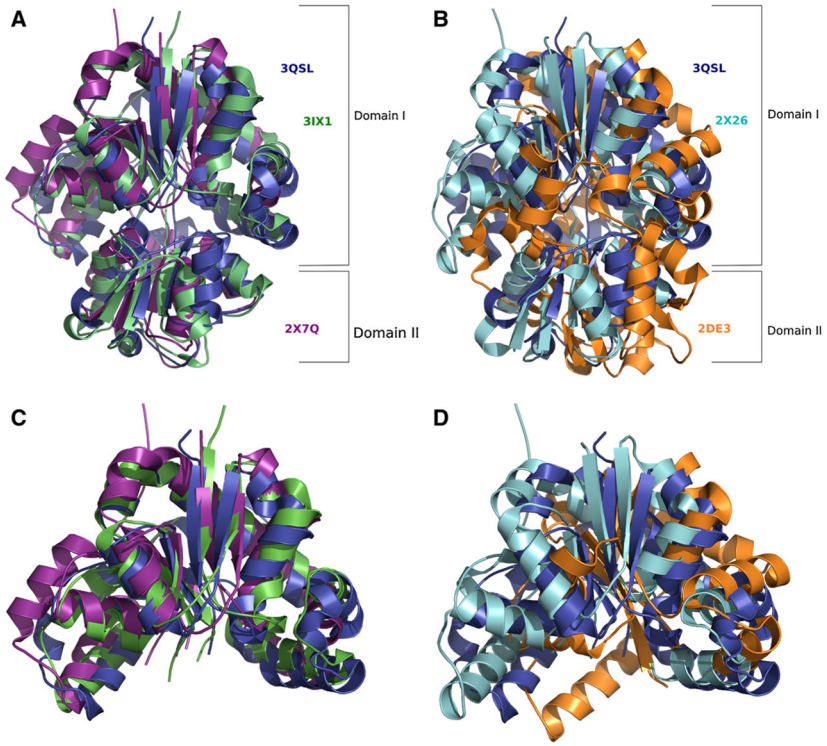

Fig. 3.

Superposition of the most structurally similar homologs. **a** and **b** present C^{α} superposition of CAE31940 with its most structurally similar homologs. In A: CAE31940 PDB ID: 3QSL (*dark blue*), with ThiY (PDB ID 3IX1; *green*), and CA3427 (PDB ID 2X7Q; *violet*); in B: CAE31940 (PDB ID: 3QSL; *dark blue*) with DSZB (PDB ID 2DE3; *orange*) and SsuA (PDB ID 2X26; *cyan*). **c**, **d** shows only the superposition of domain I of respective structures, which gives a better view of the degree of structural similarity in the superposition of single domains from all of the known structures

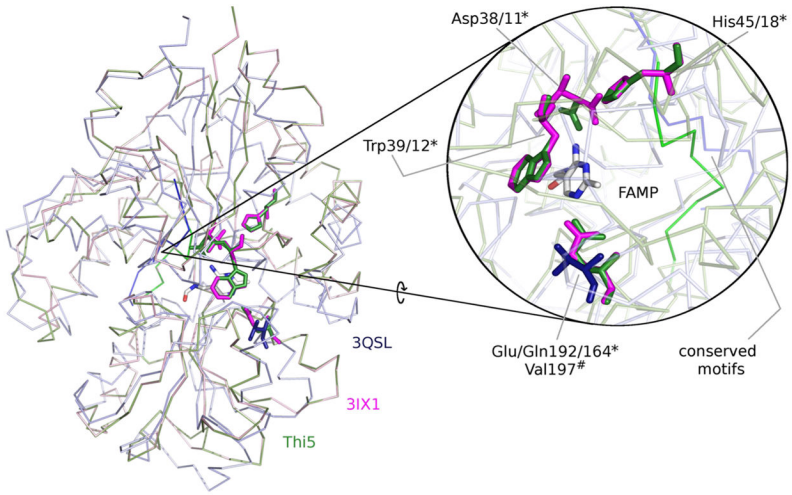

Fig. 4.

Superposition of CAE31940 with ThiY and the homology model of Thi5. The close-up is a rotated (as shown by the *arrow*) image along the *vertical axis* for better presentation of the detailed position of the conserved amino acids. The conserved residues involved in pyrimidine ring binding in ThiY are shown in stick representation and labels are marked with *asterisks* for Thi5 (model, *green*) and ThiY (crystal structure, *magenta*), and residues corresponding to this cluster in CAE31940 (crystal structure, *navy blue*) are marked with *hashes* (#). The conserved motifs GCCCXC and FGGXMP are highlighted and labeled. FAMP stands for N-formyl-4- amino-5-(aminomethyl)-2-methylpyrimidine, the ligand present in ThiY as reported by Bale et al. [3]

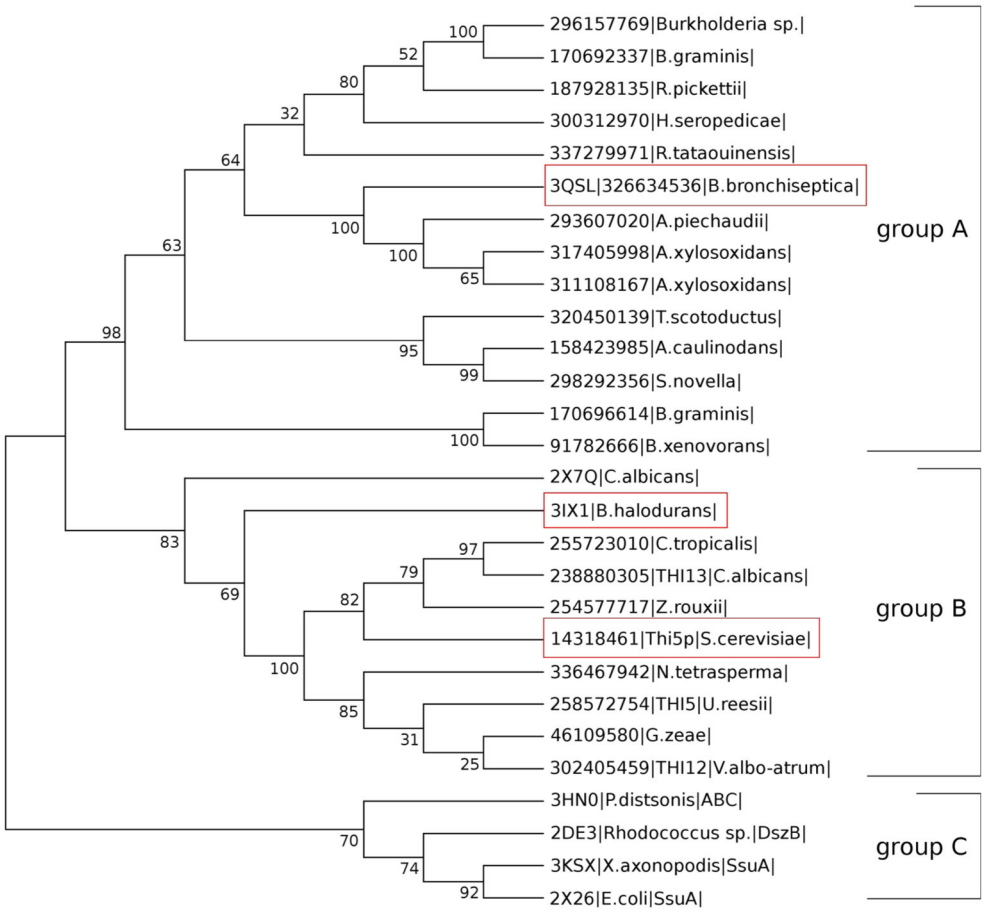

Fig. 5.

Evolutionary relationships of CAE39140 from *Bordetella bronchiseptica* (PDB code: 3QSL) and its most similar homologs as determined by structure and sequence. The *cladogram* was calculated based on the MSA presented in Fig. 2. *Values* at the nodes indicate the statistical support for the particular branches, according to the bootstrap test. The sequences from group A correspond to the blue group in the MSA, those from group B in *black*, and those from group C in *green*

Table 1

Crystallographic parameters, and data collection and refinement statistics

Organism	***Bordetella bronchiseptica*** **PDB code: 3QSL**
Crystal	
Space group	$P2_12_12_1$
Unit cell	a = 60.9 Å b = 65.0 Å, c = 160.7 Å α = β = γ = 90°
Solvent content (%)	44
Data collection	
Diffraction protocol	SAD
Wavelength (Å)	0.9792
Resolution (Å)	50.00–2.00
Highest resolution shell (Å)	2.05–2.00
Unique reflections	42946
Redundancy	7.8 (6.4)
Completeness (%)	97.4 (80.8)
$I/\sigma(I)$	30.7 (2.2)
R_{merge} (%)	8.3 (71.0)
Refinement	
R_{work} (%)	18.8 (32.3)
R_{free} (%)	23.1 (33.1)
RMSD for bond length (Å)	0.017
RMSD for bond angles (°)	1.6
Number of protein atoms	4,652
Number of water molecules	223
Ramachandran plot	
Most favored regions (%)	98.1
Additional allowed regions (%)	1.9

Data for the highest resolution shell are given in parentheses

Ramachandran statistics were calculated with *MOLPROBITY*

R_{merge} was calculated for merged Friedel pairs

Table 2

Structurally-determined homologs of the CAE31940 protein

Name of the organism	PDB code	Protein	HHpred search			
			P value	*E* value	Sequence identity (%)	RMSD (Å)
Xantomonas axonopodis pv.citri	3KSX	SsuA	2.30E−40	5.80E−36	25.2	2.7
Escherichia coli	2X26	SsuA	2.90E−36	1.10E−40	24.7	3.1
Bacillus halodurans C-125	3IX1	ThiY	2.70E−35	1.10E−39	24.2	2.6
Rhodococcus sp.	2DE3	DszB	4.70E−37	1.90E−41	22.9	3.2
Synechocystis sp.	2G29	NrtA	1.30E−30	4.90E−35	25.6	2.6
Candida albicans	2X7Q	New family	1.00E−28	3.90E−33	54.5	2.6
Parabacteroides distasonis	3HN0	Nitrate transport	4.30E−29	1.70E−33	23.8	3.3
Synechocystis sp.	2I49	CmpA	5.10E−27	2.00E−31	26.5	2.8

Name of the organism	PDB code	Protein	FATCAT results			
			Score	*P* value	Chain RMSD (Å)	Opt RMSD (Å)
Bacillus halodurans C-125	3IX1	ThiY	603.22	0.00E+00	2.7	3.0
Xantomonas axonopodis pv.citri	3E4R	SsuA	556.21	0.00E+00	2.9	3.0
Synechocystis sp	2G29	NrtA	545.82	1.11E−16	2.9	3.0
Synechocystis sp	2I48	CmpA	527.35	2.11E−16	3.0	3.1
Archaeoglobus fulgidus	1ZBM	AF1704	377.69	3.80E−12	3.5	3.1
Thermus thermophilus HB8	2CZL	MqnD	381.37	5.50E−12	3.9	3.2
Escherichia coli	2X26	SsuA	523.72	2.84E−11	4.1	3.2
Streptomyces coelicolor	2NXO	SCO4506	344.48	6.67E−11	3.4	3.1

The FATCAT probability cutoff was less than 5.00×10^{-2}

J Struct Funct Genomics. Author manuscript; available in PMC 2015 June 08.