Person to contact for this submission: Wladek Minor
(wladek@iwonka.med.virginia.edu; phone: 4349242948; fax: 4349821616)

IUCr

Acta Crystallographica Section F structural communications

The first crystal structure of pyrimidine/thiamin biosynthesis precursor-like domain-containing protein, CAE31940 from protebacterium *Bordetella bronchiseptica* RB50 and evolutionary insight into NMT1/Thi5 family.

Bajor, J.[ac], Tkaczuk, K.L.[ac], Chruszcz, M.[ac], Kagan, O.[bc], Savchenko, A.[bc], Minor, W.[ac]

[a]Department of Molecular Physiology and Biological Physics, University of Virginia, 1340 Jefferson Park Avenue, Charlottesville, VA 22908, USA, [b]Banting and Best Department of Medical Research, 112 College Street, University of Toronto, Toronto, Ontario M5G 1L6, Canada, [c]Midwest Center for Structural Genomics

Correspondence e-mail: wladek@iwonka.med.virginia.edu

Keywords: *Bordetella bronchiseptica* RB50; NMT1/THI5-like domain-containing protein; crystal structure; MCSG;

Abstract

The structure of the CAE31940 protein presented here is the first crystal structure of proteobacterial putative NMT1/Thi5-like domain-containing protein. The CAE31940 structure was solved at 2.0 Å resolution. Apart from the crystal structure we also discuss the sequential 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.

1. 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, then they are covalently linked to yield thiamin phosphate (Zurlinder and Schweingruber 1994; Begley, Chatterjee et al. 2008). This process is well studied in prokaryotes but is still poorly understood in eukaryotes, like i.e. *Saccharomyces cerevisiae*. The *thi5* gene product Thi5 is responsible for the synthesis of 4-amino-5-(hydroxymethyl)-2-methylpyrimidine phosphate in yeast (Maundrell 1990; Wightman and Meacock 2003; Bale, Rajashankar et al. 2010), which appears to be conserved in eukaryotes with thiamin biosynthetic pathways and it was studied in yeast. Thi5 belongs to protein family PF09084 that comprises NMT1 and NMT1/Thi5-like

proteins, structurally exemplified by the ThiY protein from *Bacillus halodurans* c-125 (PDB code: 3IX1). These proteins are proposed to be required for the biosynthesis of the pyrimidine moiety of thiamin (Maundrell 1990; Wightman and Meacock 2003; Bale, Rajashankar et al. 2010).

2. Materials and Methods

2.1. Cloning, Expression and Purification

Selenomethionine (Se-Met) substituted CAE31940 protein was produced using standard MSCG protocols as described by Zhang et al. (Zhang, Skarina et al. 2001). Briefly, gene BB1442 from *Bordetella bronchiseptica* RB50 was cloned into a p15TV LIC plasmid using ligation independent cloning (Aslanidis & Dejong, 1990, Haun et al., 1992, Eschenfeldt et al., 2009). The gene was overexpressed in *E. coli* BL21-CODONPLUS(DE3)-RIPL cells in Se-Met-containing 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 minutes 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 polyhistidine tag (His-Tag) was removed by digestion in recombinant TEV protease and the digested protein was passed again through an affinity column. The flowthrough was dialyzed against a solution containing 300 mM NaCl, 10 mM HEPES pH 7.5 and 1mM TCEP. Purified protein was concentrated to 36 mg/mL and frozen in liquid nitrogen.

2.2. 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-Na citrate at pH 5.6. Crystals were grown at 293 K and formed after one week of incubation. Immediately after harvesting, crystals were transferred into cryoprotectant solution (Paraton-N) without mother liquor, washed twice in the solution and flash cooled in liquid nitrogen.

2.3. Data Collection and Processing

Data were collected at 100 K at the 19-ID beamline (ADSC Q315 detector) of the Structural Biology Center (Rosenbaum, Alkire et al. 2006) at the Advanced Photon Source (Argonne National Laboratory, Argonne, Illinois, USA). The beamline was controlled by HKL-3000 (Minor, Cymborowski et al. 2006). Diffraction data were processed with HKL-2000 (Minor, Cymborowski et al. 2006). Data collection, structure determination, and refinement statistics are summarized in Table 1.

2.4. Structure Solution and Refinement

The structure of the selenomethionine 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 (Sheldrick 2008), MLPHARE, DM, ARP/wARP, CCP4 (Bjellqvist, Basse et al. 1994), SOLVE, and RESOLVE (Terwilliger 2004). The resulting model was further refined with REFMAC5 (Murshudov, Skubak et al. 2011), and COOT (Emsley and Cowtan 2004). MOLPROBITY (Chen 2010) and ADIT (Yang, Guranovic et al. 2004) were used for structure validation. The coordinates and experimental structure factors were deposited to PDB with accession code 3QSL.

2.5. Bioinformatics analyses

Sequence homology searches were performed with PSI-BLAST (Altschul, Madden et al. 1997) and structural homology searches were done with HHpred (Soding 2005; Soding, Biegert et al. 2005). Structure superposition was done with SSM (Krissinel and Henrick 2004). The evolutionary history was inferred using the neighbor-joining method (Saitou and Nei 1987). The bootstrap consensus tree inferred from 500 replicates (Felsenstein 1985) is taken to represent the evolutionary history of the taxa analyzed (Felsenstein 1985). 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 (Felsenstein 1985). The evolutionary distances were computed using the Poisson correction method (Zuckerkandl and Pauling 1965) 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 (Tamura, Peterson et al. 2011).

2.6. Homology modeling of Thi5 from *Saccharomyces cerevisiae*

The “FRankenstein’s Monster” method (in short best fragment assembly method) was used to construct a homology of Thi5 from *Saccharomyces cerevisiae* S288c using the CAE31940 structure as a modeling 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 (Kosinski, Cymerman et al. 2003; Wallner and Elofsson 2007). PROQ (Wallner and Elofsson 2007) and a MetaMQAP method were used for the evaluation of models (Pawlowski, Gajda et al. 2008), which allowed the prediction of the deviations of individual residues in the homology model from their counterparts in the template structure.

3. Results and Discussion

3.1. Overall structure

The structure of the CAE3140 protein from *Bordetella bronchiseptica* RB50 was solved at 2.0Å. 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, 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 could not be modeled the N-terminus of the chain A and 19 at the N-terminus of chain B. The last residue at the C-terminus of both chains is 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 Figure 1.

According to the predictions of the PISA server, the CAE3140 protein is monomeric in solution.

3.2. Evolutionary relatives and characteristics of CAE1490

Sequential and structural comparisons

Standard sequence searches with PSI-BLAST found no homologs of CAE31490 with a known crystal structure. These searches, however, clearly showed that the most similar relatives (by sequence) of CAE3140 are NMT1/Thi5-like domain-containing proteins, mainly from proteobacteria. It suggests that the structure of CAE3140 presented here may be the first known structure of a NMT1/Thi5-like domain-containing protein. Thus subsequent structural homology searches were performed.

In order to find the closest structural homologs of CAE3140, HHpred (Soding, Biegert et al. 2005) and FATCAT (Li, Ye et al. 2006

) searches of the most up-to-date PDB database available (pdb70-Dec11) were performed. Their results show that CAE3140 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, 3IX1) belong to the NMT1/THI5-like family according to the PFam (Sonnhammer, Eddy et al. 1997) annotations in the PDB database, and the remaining five are not assigned to any family. All of the identified structural homologs share very little sequence identity with CAE3140, and according to the authors' annotations for each, all act on different substrates than the one predicted for CAE3140.

Of the 5 best-scoring hits in the structural homology searches, the protein with the most similar sequence to CAE3140 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 (Bale, Rajashankar et al. 2010) 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.) CAE3140 shares 24% and 26% sequence identity and 42% and 46% sequence similarity with ThiY and Thi5 respectively. This corresponds to BLAST expectation values of $1x10^{-4}$ and $3x10^{-13}$ respectively. This is rather high sequence similarity and owing that the structure of ThiY is one of the most structurally similar proteins to CAE3140, ThiY superimposes on CAE31490 with an RMSD of 2.5 Å. It is also confirmed by the results of FATCAT structural homolog searches where ThiY ranks as the best hit for CAE3140 with an E-value of 0 and RMSD of 2.7 Å (as presented in Table 2). This may suggest that all three proteins are orthologs and may originate from the same ancestor, which is further supported by the evolutionary analysis presented below.

Superposition 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) (Figure 3B), the

pairwise RMSD of the most structurally similar proteins is 2.55 Å (as shown in Table 2 and Figure 3). However, the motif “FGGXMP” in CAE3140 corresponds to the previously identified “GCCCX” motif in Thi5 from *Saccharomyces cerevisiae* (Bale, Rajashankar et al. 2010): when CAE3140 and the homology model of Thi5 are superimposed the aforementioned motifs are similar in both sequence and 3D structure. Sequence analysis also shows that the “FGGXMP” motif is conserved in all proteobacterial sequences related to CAE3140. Therefore this motif may be a hallmark of proteobacterial members of NMT1/THI5-like domain-containing proteins.

3.3. Evolutionary insight into NMT1/THI5 family

The cladogram presented in Figure 5 and the multiple sequence alignment (MSA) in Figure 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 CAE3140, characterized by the FGGXMP 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 (Bale, Rajashankar et al. 2010), ThiY is closely related to Thi5. Group C (green group in Fig. 2) contains structural homologs (mainly SsuA) of CAE3140, 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 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: 2Q7X), groups together with them. CA3427, as described by the authors of its structure (Santini, Claverie et al. 2011), 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 CAE3140 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 CAE3140 – 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-protobacteria and 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 also proves that our structure is the first crystal structure of a NMT1/Thi5-like domain-containing protein with the characteristic 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 one can deduct that it can take part in the same process in proteobacteria.

4. Protein data bank accession code

Coordinates and structure factors of CAE3140 have been deposited in the RCSB Protein Data Bank with accession code 3QSL.

Acknowledgements

This research was funded with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200700058C. 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 Matthew Zimmerman for critically reading the manuscript.

Unit cell and space group		
Space group:		P 21 21 21
Unit cell: a=60.91 b=65.03 c=160.69 alpha=90.00 beta=90.00 gamma=90.00		
Data collection		
Wavelength:	0.9792	
Resolution Range:	2.00-50	(2.00-2.05)
Obs reflections:	333947	
Unique reflections:	42851	(2559)
Completness (%):	97.4	(80.8)
Intensity I:	132.1	(9.0)
Sigma I:	4.3	(4.1)
Rmerge I:	0.083	(0.7)
Redundancy:	7.8	(6.4)
Chi squared:	1.523	(0.9)
Percent ref of R-free (%):	5.1	
Refinement		
R-factor (%):	19.0	
R-factor (R-work %):	18.8	(32.3)
R-factor (R-free %):	23.1	(33.1)
Average isotropic B value:	48.5	
Ramachandran outliers (%):	0.0	
Ramachandran favored (%):	98.12	
RMS Deviations		
Bond distance:	0.017	
Bond angle degree:	1.619	
Protein Model		
Number of atoms:	4919	
Number of solvent atoms:	223	
Number of MET and MSE:	20	

* Data for highest resolution shell are in paranthesis

$^{\#}R_{merge} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$.

Table 1. Data collection and structure refinement statistics

Name of the organism	PDB code	protein	HHpred search			
			P-value	E-value	Sequence identity [%]	RMSD [Å]
Xanthomonas 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**			
			score	**P-value**	**chain RMSD [Å]**	**opt-RMSD [Å]**
Bacillus halodurans C-125	3IX1	ThiY	603.22	0.00E+00	2.7	3.0
Xanthomonas 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-15	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

Table 2. Structurally-determined homologs of the CAE3140 protein from from *Bordetella bronchiseptica*. The FATCAT probability cutoff was less than 5.00e-02.

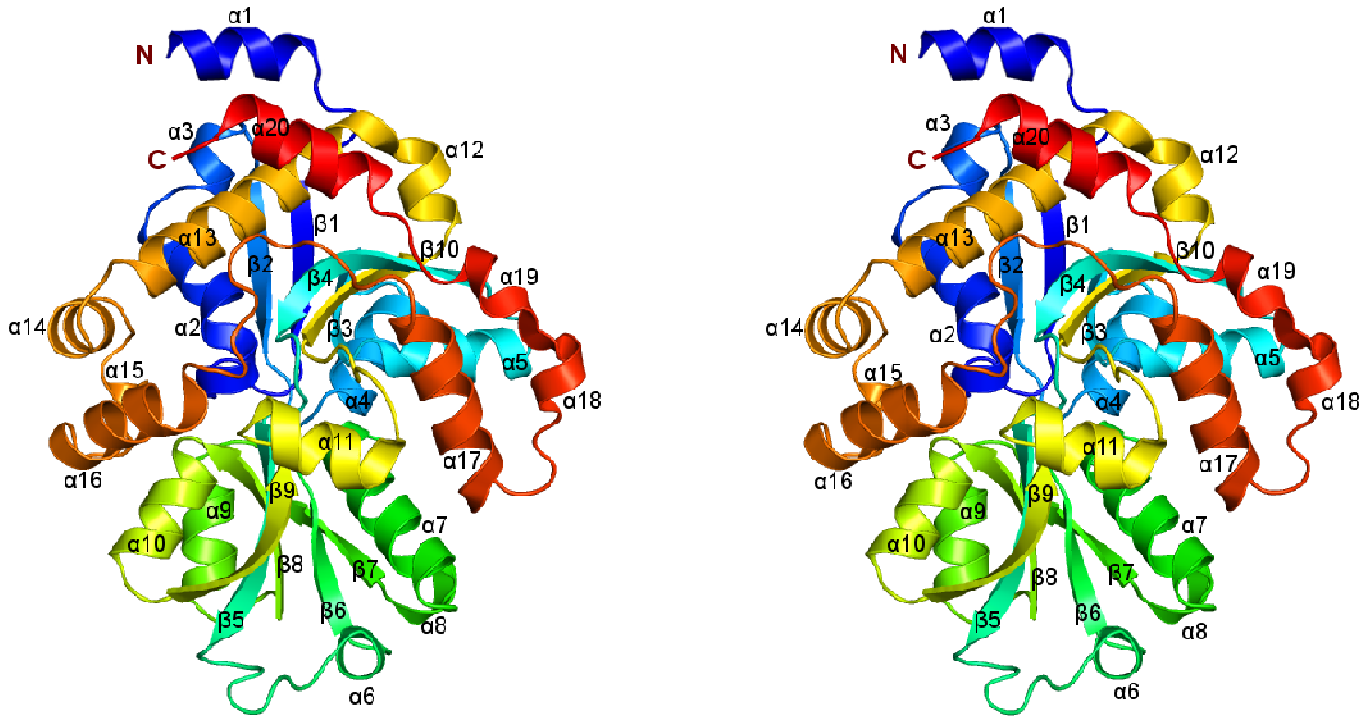

Figure 1. Structure of CAE3140 from *Bordetella bronchiseptica* (PDB code: 3QSL) shown in cross-eyed stereo. Structure is colored by primary sequence, from deep blue at the N-terminus to deep red at the C-terminus. The N and C termini, as well as the secondary structure elements, are labeled.

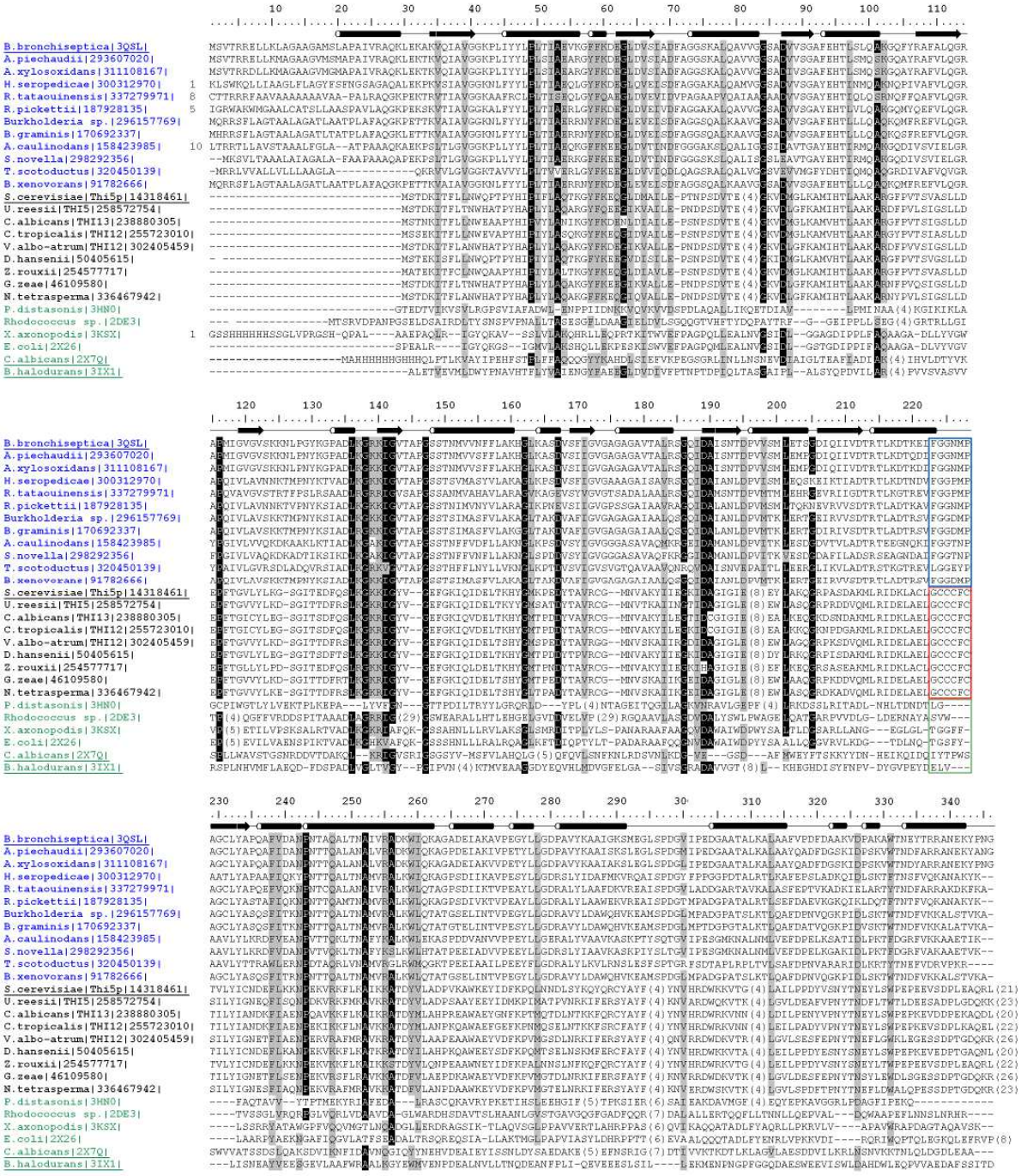

Figure 2. Multiple sequence alignment (MSA). The secondary structure of 3QSL is marked above the MSA. Extra residues for some sequences are removed from the alignment at the N- and C-termini for conciseness. The number of residues removedis marked at either end of of the MSA. Here are the full names 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 sequential motifs discussed in the text are boxed for better visibility.

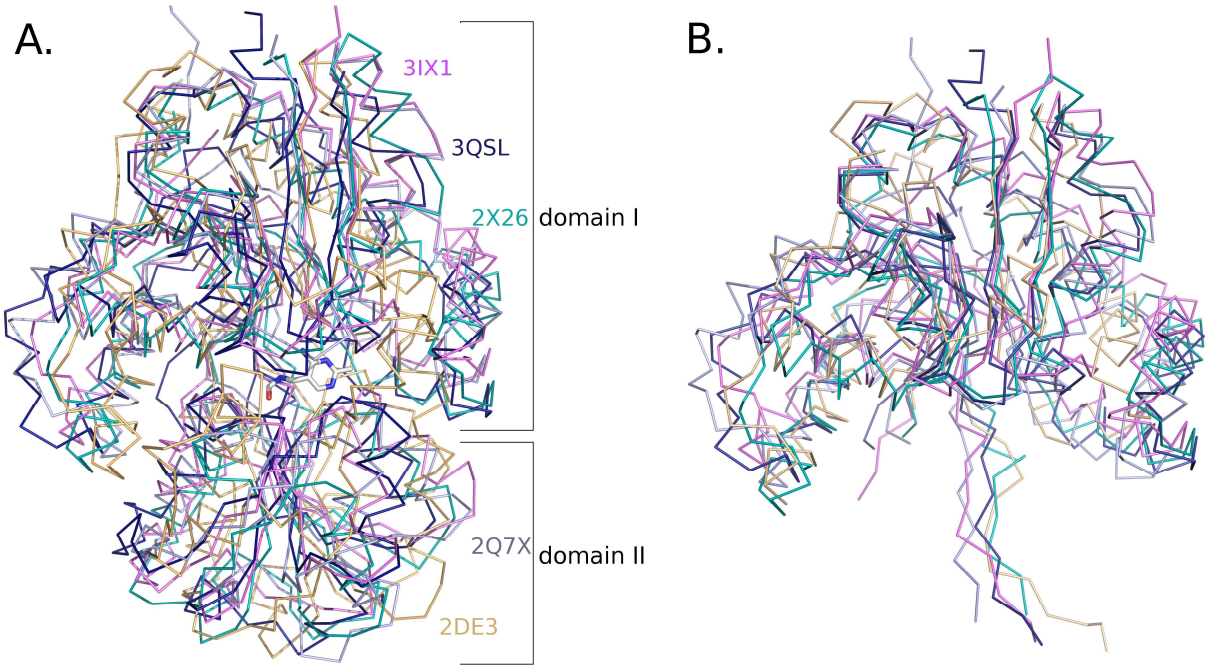

Figure 3. Superposition of the most structurally similar homologs. Panel A presents C^{α} superposition of CAE31940 with its most structurally similar homologs. CAE31940 (dark blue), DSZB (PDB ID 2DE3; yellow), SsuA (PDB ID 2X26; cyan), ThiY (PDB ID 3IX1; pink), and CA3427 (PDB ID 2X7Q; violet). Panel B shows only the superposition of domain I, which gives a better view of the degree of structural similarity in the superposition of single domains from all of the known structures.

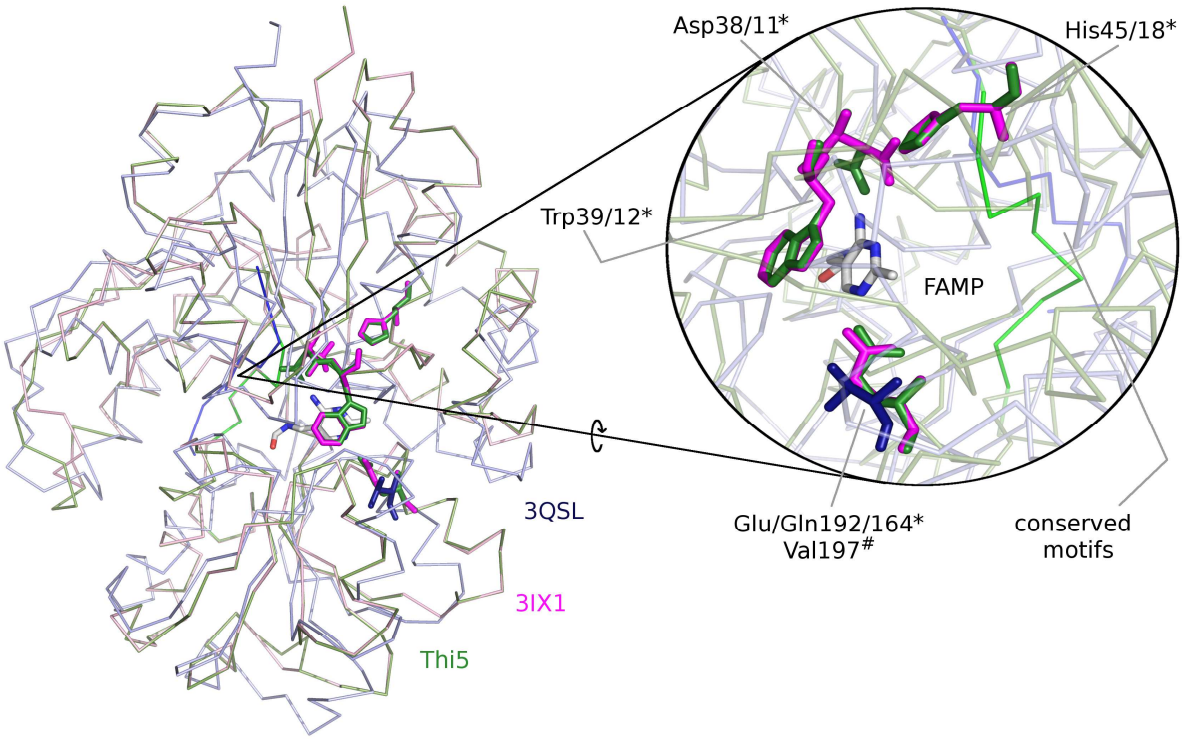

Figure 4. Superposition of CAE31940 with ThiY and the homology model of Thi5. The close-up is a rotated (as shown by the arrow) image along the vertical axis for better presentation of the detailed position of the conserved amino acids. The conserved residues involved in pyrimidine ring binding are shown in stick representation and labels are marked with asterisks for Thi5 (green) and ThiY (magenta), and residues corresponding to this cluster in CAE1940 (navy blue) are marked with hashes (#). The conserved motifs GCCCXC and FGGXMP are highlighted and labeled.

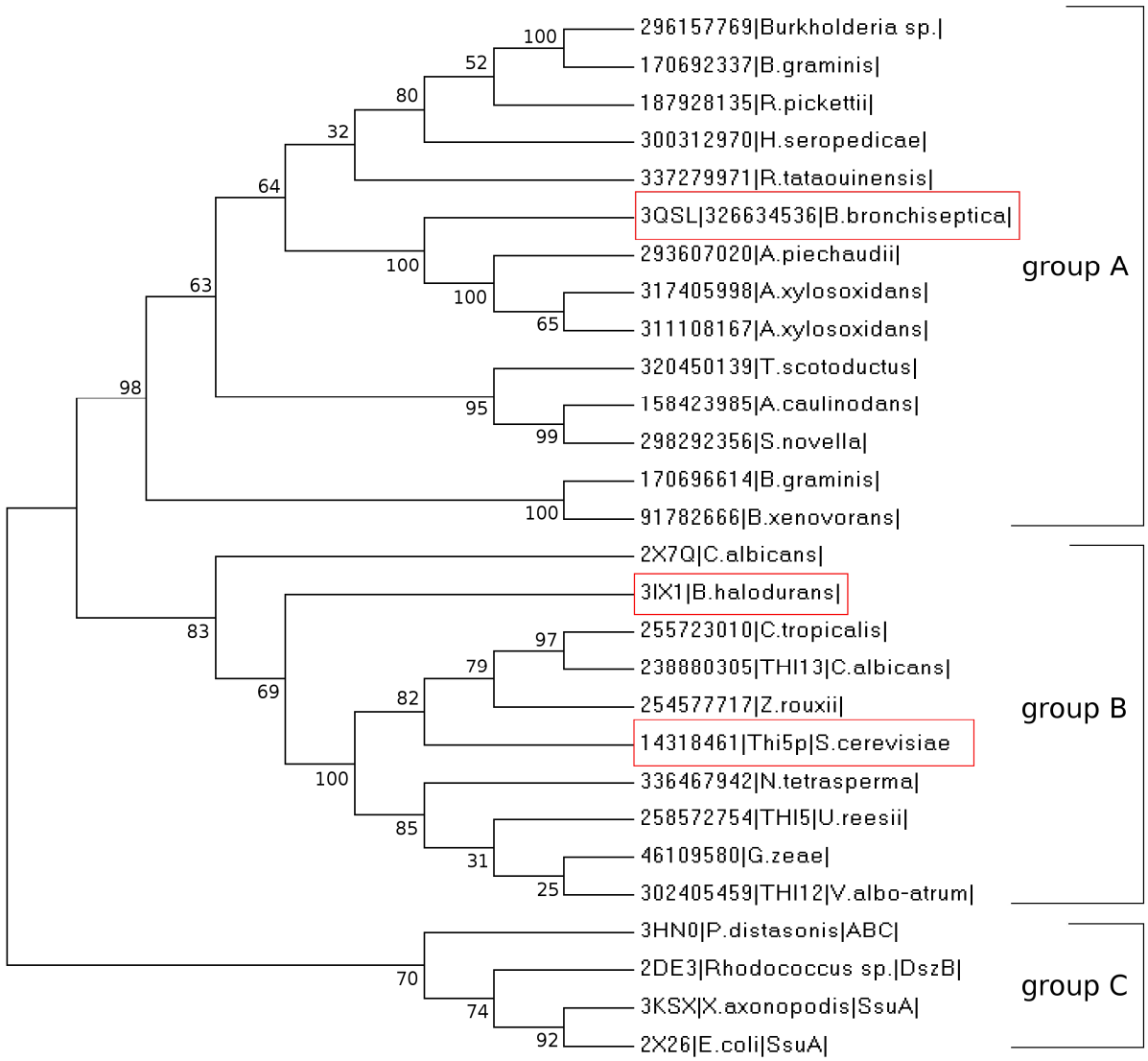

Figure 5. Evolutionary relationships of CAE3140 from *Bordetella bronchiseptica* (PDB code: 3QSL) and its most similar homologs as determined by structure and sequence. The cladogram was calculated based on the MSA presented in Figure 2. Values at the nodes indicate the statistical support for the particular 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.

References

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Bale, S., K. R. Rajashankar, et al. (2010). "HMP binding protein ThiY and HMP-P synthase THI5 are structural homologues." Biochemistry **49**(41): 8929-8936.

Begley, T. P., A. Chatterjee, et al. (2008). "Cofactor biosynthesis--still yielding fascinating new biological chemistry." Curr Opin Chem Biol **12**(2): 118-125.

Bjellqvist, B., B. Basse, et al. (1994). "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 **15**(3-4): 529-539.

Chen, V. B., Arendall III, W.B., Headd, J.J., Keedy, D.A., Immormino, R. M., Kapral, G.J., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2010). "MolProbity: all-atom structure validation for macromolecular crystallography." Acta Crystallographica Section D **66**: 12-21.

Emsley, P. and K. Cowtan (2004). "Coot: model-building tools for molecular graphics." Acta Crystallographica Section D-Biological Crystallography **60**: 2126-2132.

Felsenstein, J. (1985). "Confidence limits on phylogenies: An approach using the bootstrap." Evolution **39**: 783-791.

Kosinski, J., I. A. Cymerman, et al. (2003). "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 **53 Suppl 6**: 369-379.

Krissinel, E. and K. Henrick (2004). "Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions." Acta Crystallogr D Biol Crystallogr **60**(Pt 12 Pt 1): 2256-2268.

Li, Z., Y. Ye, et al. (2006

). "Flexible structural neighbourhood - a database of proteins structuctural similarities and alignments." Nucleic Acids Res **34**(database issue): D277-280.

Maundrell, K. (1990). "nmt1 of fission yeast. A highly transcribed gene completely repressed by thiamine." J Biol Chem **265**(19): 10857-10864.

Minor, W., M. Cymborowski, et al. (2006). "HKL-3000: the integration of data reduction and structure solution - from diffraction images to an initial model in minutes." Acta Crystallographica Section D-Biological Crystallography **62**: 859-866.

Murshudov, G. N., P. Skubak, et al. (2011). "REFMAC5 for the refinement of macromolecular crystal structures." Acta Crystallogr D Biol Crystallogr **67**(Pt 4): 355-367.

Pawlowski, M., M. J. Gajda, et al. (2008). "MetaMQAP: a meta-server for the quality assessment of protein models." BMC Bioinformatics **9**: 403.

Rosenbaum, G., R. W. Alkire, et al. (2006). "The Structural Biology Center 19ID undulator beamline: facility specifications and protein crystallographic results." J Synchrotron Radiat **13**(Pt 1): 30-45.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Mol Biol Evol **4**(4): 406-425.

Santini, S., J. M. Claverie, et al. (2011). "The conserved Candida albicans CA3427 gene product defines a new family of proteins exhibiting the generic periplasmic binding protein structural fold." PLoS One **6**(4): e18528.

Sheldrick, G. M. (2008). "A short history of SHELX." Acta Crystallographica Section A **64**: 112-122.

Soding, J. (2005). "Protein homology detection by HMM-HMM comparison." Bioinformatics **21**(7): 951-960.

Soding, J., A. Biegert, et al. (2005). "The HHpred interactive server for protein homology detection and structure prediction." Nucleic acids research **33**(Web Server issue): W244-248.

Soding, J., A. Biegert, et al. (2005). "The HHpred interactive server for protein homology detection and structure prediction." Nucleic Acids Res **33**(Web Server issue): W244-248.

Sonnhammer, E. L., S. R. Eddy, et al. (1997). "Pfam: a comprehensive database of protein domain families based on seed alignments." Proteins **28**(3): 405-420.

Tamura, K., D. Peterson, et al. (2011). "MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. ." Molecular Biology and Evolution (28): 2731-2739.

Terwilliger, T. (2004). "SOLVE and RESOLVE: automated structure solution, density modification, and model building." Journal of Synchrotron Radiation **11**: 49-52.

Wallner, B. and A. Elofsson (2007). "Prediction of global and local model quality in CASP7 using Pcons and ProQ." Proteins **69 Suppl 8**: 184-193.

Wightman, R. and P. A. Meacock (2003). "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 **149**(Pt 6): 1447-1460.

Yang, H. W., V. Guranovic, et al. (2004). "Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank." Acta Crystallographica Section D-Biological Crystallography **60**: 1833-1839.

Zhang, R. G., T. Skarina, et al. (2001). "Structure of Thermotoga maritima stationary phase survival protein SurE: a novel acid phosphatase." Structure **9**(11): 1095-1106.

Zuckerkandl, E. and L. Pauling (1965). Evolutionary divergence and convergence in proteins, Academic Press, New York.

Zurlinder, A. and M. E. Schweingruber (1994). "Cloning, nucleotide sequence, and regulation of Schizosaccharomyces pombe thi4, a thiamine biosynthetic gene. ." Journal of Bacteriology **176**: 6631–6635.