ELSEVIER

FEMS Microbiology Letters 251 (2005) 341–346

FEMS MICROBIOLOGY Letters

www.fems-microbiology.org

Bacillus subtilis YkuK protein is distantly related to RNase H

Łukasz Kniżewski, Krzysztof Ginalski *

Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw University, Pawińskiego 5A, 02-106 Warsaw, Poland

Received 19 May 2005; received in revised form 12 August 2005; accepted 16 August 2005

First published online 30 August 2005

Edited by M.Y. Galperin

Abstract

In addition to one hypothetical viral sequence from *Bacteriophage KVP40*, the PfamA family of unknown function DUF458 (Pfam Accession No. PF04308) encompasses several uncharacterized bacterial proteins including *Bacillus subtilis* YkuK protein. Using Meta-BASIC, a highly sensitive method for detection of distant similarity between proteins, we assign DUF458 family members to the ribonuclease H-like (RNase H-like) superfamily. DUF458 sequences maintain all core secondary structure elements of RNase H-like fold and share several conserved, presumably active site residues with RNase HI, including an invariant DDE motif. In addition to providing a model structure for a previously uncharacterized protein family, this finding suggests that DUF458 proteins function as nucleases. The unusual phyletic pattern, together with a presence of DUF458 in several thermophilic organisms, may suggest a potential role of these proteins in DNA repair in stressful conditions such as an extreme heat or other stress that causes spore formation.

Keywords: YkuK; DUF458 family; RNase H-like fold; Structure/function prediction; Active site; Nuclease

1. Introduction

Ribonuclease H-like (RNase H-like) proteins comprise an important class of nucleic acid modifying enzymes with nuclease or polynucleotidyl transferase activities. The SCOP database [1] currently defines seven different families in RNase H-like superfamily including RNase H [2,3], retroviral integrase [4], mu transposase [5], Tn5 transposase [6], DnaQ-like 3′–5′ exonuclease [7], RuvC resolvase [8] and mitochondrial resolvase ydc2 [9]. These enzymes perform critical functions in various biological processes such as transcription (RNase H) or viral infection (retroviral integrase), abasic DNA repair (exonuclease III), transposition (mu transposase), recombination and recombinatorial DNA repair (RuvC resolvase). The ribonuclease H-like superfamily is defined by a common core fold that includes a central five-stranded, mixed β-sheet flanked by α-helices on both sides (with βββαβαβα topology) [10]. Various reactions catalyzed by the superfamily members share a common metal ion dependent catalytic mechanism supported by the presence of an invariant active site DDE motif [11]. Two highly conserved aspartates are located on spatially adjacent first and fourth β-strands of the core RNase H-like fold, while the position of third active site carboxylate varies depending on the family. Here we show that previously uncharacterized PfamA [12] domain of unknown function DUF458 (Pfam Accession No. PF04308, InterPro Accession No. IPR007405) is yet another family of RNase H-like proteins and identify its potential active site residues.

* Corresponding author. Tel.: +48 22 8749100; fax: +48 22 8749130.
E-mail address: kginal@icm.edu.pl (K. Ginalski).

0378-1097/$22.00
doi:10.1016/j.femsle.2005.08.020

2. Materials and methods

2.1. Identification of DUF458 family members

To identify all DUF458 family members an exhaustive, transitive PSI-BLAST [13] search procedure was applied. Initially, PSI-BLAST search against the NCBI non-redundant protein sequence database (filtered nr, posted 8 March 2005; 2,354,365 sequences) with inclusion threshold of 0.01 until profile convergence was carried out using the consensus sequence of PfamA [12] DUF458 as a query. Consequently, collected sequences were subjected to further PSI-BLAST searches until no new sequences were found. DUF458 family members were also subjected to neighborhood analysis by the STRING database [14] (http://string.embl.de) to detect possible functional associations.

2.2. Structural assignment for DUF458 proteins

Initially, the consensus sequence of DUF458 as well as several members of this family were analyzed with CDD [15] (http://www.ncbi.nlm.nih.gov/Struct/cdd/wrpsb.cgi) and SMART [16] (http://www.smart.embl-heidelberg.de) search tools to detect conserved protein domains annotated in SMART, Pfam and COG [17] databases. This analysis also included searches for transmembrane segments (with TMHMM2 [18]), signal peptides (SignalP [19]), low compositional complexity (CEG [20]) and coiled coil (Coils2 [21]) regions, as well as regions containing internal repeats (Prospero [22]).

Further searches were performed with meta-profile alignment method Meta-BASIC [23] available at http://basic.bioinfo.pl. Meta-BASIC combines the use of sequence profiles and secondary structure predictions (meta-profiles) for query sequence and given protein families with various scoring systems and meta-profile alignment algorithms to detect distant similarity between proteins even if the structure of the reference protein is not known. Specifically, the consensus sequence of DUF458 was compared to all 7,418 PfamA families and to 10,128 proteins (representatives at 90% of sequence identity) extracted from Protein Data Bank (PDB) [24]. The same comparison was also conducted using PSI-BLAST and RPS-BLAST [13].

Finally, both the consensus sequence of DUF458 and one of the family members, *Bacillus subtilis* YkuK protein, were submitted to Meta-Server [25] (http://bioinfo.pl/meta) that assembles various secondary structure prediction and top-of-the-line fold recognition methods. Collected predictions were screened with 3D-Jury [26], the consensus method of fold recognition servers.

2.3. Generation of sequence-to-structure alignment

To define the general conservation pattern in the selected template, *Escherichia coli* RNase HI (PDB|2rn2) [2], its close homologues were collected with PSI-BLAST search against the NCBI non-redundant protein sequence database. Multiple sequence alignments for both DUF458 and RNase HI families were prepared using PCMA program [27] followed by final manual adjustments. Sequence-to-structure alignment between DUF458 and RNase HI families was built manually using 3D assessment procedure [28] taking into account predicted secondary structure (consensus of the results of several secondary structure prediction methods, mainly PSI-PRED [29]), hydrophobic profile of the family and conservation of presumably catalytic residues.

2.4. 3D model building

Based on the final sequence-to-structure alignment, a 3D model of *B. subtilis* YkuK protein was built with the MODELLER program [30] using the *E. coli* RNase HI structure (PDB|2rn2) as a template. The relative orientation of the second α-helix of the conserved fold core was taken from the *E. coli* Tn5 transposase structure (PDB|1b7e) [6], which has similar secondary structure pattern in this region without any inserted additional elements. Finally, the overall quality of the modeled structure was checked with Verify3D program [31].

3. Results and discussion

3.1. DUF458 belongs to RNase H-like superfamily

Initial PSI-BLAST [13] search with DUF458 consensus sequence performed against the NCBI non-redundant protein sequence database (*E*-value threshold of 0.01) identified 16 entirely uncharacterized proteins (predominantly from bacterial species), including sequences belonging to an uncharacterized cluster of orthologues [32] COG1978. Further exhaustive PSI-BLAST searches originated from these DUF458 family sequences did not detect any other proteins with significant *E*-value (below 0.01). Application of other standard sequence similarity search methods such as CDD [15] and SMART [16] database search tools using collected DUF458 sequences did not yield any significant hits to known protein domains. In addition, the 3D-Jury [26] coupled to structure prediction Meta-Server [25], also did not provide any consistent and reliable matches (with scores above 50 [33]) to existing structures. However, using meta-profile alignment method Meta-BASIC [23] DUF458 consensus sequence was mapped to both Piwi domain (PF02171 in Pfam database, *Z*-score 13.37), recently shown to possess RNase H-like structure [34], and a holliday junc-