A putative novel alpha/beta hydrolase ORFan family in *Bacillus*

Naomi Siew[a,b], Harpreet Kaur Saini[c,*], Daniel Fischer[b,c]

[a] *Department of Chemistry, Ben Gurion University, Beer-Sheva 84105, Israel*
[b] *Bioinformatics Group, Department of Computer Science, Ben Gurion University, Beer-Sheva 84105, Israel*
[c] *Buffalo Center of Excellence in Bioinformatics, University at Buffalo, 901 Washington St. suite 300, Buffalo, NY 14203, USA*

Received 30 December 2004; revised 25 March 2005; accepted 11 April 2005

Available online 4 May 2005

Edited by Hans Eklund

Abstract A large number of sequences in each newly sequenced genome correspond to lineage and species-specific proteins, also known as ORFans. Amongst these ORFans, a large number are sequences with unknown structures and functions. We have identified a family of sequences, annotated as hypothetical proteins, which are specific to *Bacillus* and have carried out a computational study aimed at characterizing this family. Fold-recognition methods predict that these sequences belong to the α/β hydrolase fold. We suggest possible catalytic triads for the ORFans and propose a hypothesis regarding the possible families within the α/β hydrolase superfamily to which they may belong.

Keywords: ORFans; Fold recognition; Structure prediction; Alpha/beta hydrolase; *Bacillus*

1. Introduction

The availability of dozens of complete genomes enables studying genomes and evolution from new perspectives. Comparative genomics studies have shown that there are varying levels of protein conservation, from those proteins widely conserved among a large number of organisms, to those found in only a particular lineage or in only one species (see, for example [1–4]). We refer to the lineage-specific proteins as orthologous ORFans, and to the species-specific ones as singleton- or paralogous ORFans [1,5,6].

ORFans are present in almost every completely sequenced genome and their percentage can be as high as 60% [7]. Many of the ORFans are sequences with an unknown function as yet and are usually annotated as hypothetical proteins. Some of these ORFans could be highly divergent sequences that belong to known protein families and have similar structures and functions, while others could be novel proteins, with structures and/or functions that have not been seen before [1,8]. Yet, some of them may correspond to errors, non-expressed, or non-functional proteins [1,9–13] and to sequences in the process of deterioration [14].

One way to infer the function of an uncharacterized protein is through its three-dimensional structure [15–21]. Because of the huge amount of raw sequence data and the slow experimental procedures, the availability of sensitive structure prediction methods could help in identifying ORFans that may be distant relatives of known proteins, and provide clues as to their function, where possible.

As part of a study on ORFans in *Bacilli*, we identified one family of orthologous ORFans which consists of sequences from four *Bacillus* genomes, *B. subtilis*, *B. halodurans*, *B. cereus* and *B. anthracis*. As described below, fold-recognition methods predict with a high level of confidence that these orthologous ORFans belong to the α/β hydrolase fold superfamily.

The α/β hydrolase fold [22] is one of the largest protein superfamilies and one of the largest within the alpha–beta class of folds [23,24]. This is a superfamily of single domain enzymes, which are highly divergent on the sequence level and perform a wide range of reactions, using a large variety of substrates [23–25]. Among them are, e.g., carboxylic acid ester hydrolases, lipid hydrolases, haloperoxidases and dehalogenases. Although large insertions to the basic (canonical) fold are common, thus, making it the most ''plastic'' fold known [24], the 3D structure of the α/β hydrolases is highly conserved, as well as the location of the catalytic triad residues [23–25].

The canonical α/β hydrolase fold is constructed of eight β-strands and of five α-helices [22,23]. Most of the variations include insertions of additional secondary structure elements after strand β6. These insertions are referred to as caps, lids or flaps, and they have proven important in defining the shape of the substrate binding crevices of the enzyme, regulating the accessibility and specificity of the substrate, and perhaps other functions as well [22,25,26]. Other common insertions are found after β3, β4, β7 and β8, and several deletions are also known [23].

The catalytic triad in α/β hydrolases is constructed of a nucleophile, an acid and a histidine residue. The nucleophile can be a serine, cysteine or aspartic acid, and is found at the beginning of a short loop between β5 and α3 in the canonical fold, also referred to as the ''nucleophile elbow''. The conserved histidine is located after β8. The acid can be an aspartic or a glutamic acid [23–25]. The location of the acid in most α/β hydrolases is in a loop between β7 and α5 in the canonical fold [23,25]. However, it could also reside after strand β6 [24,27–30].

The most common catalytic triad in α/β hydrolases known, thus, far is Ser–Asp–His. In the large esterases and several lipases, the triad is Ser–Glu–His and in dienlactone hydrolases the unique Cys–Asp–His triad is found. In dehalogenases and epoxide hydrolases, it is Asp–Asp–His, while in haloalkane dehalogenase it is Asp–Glu–His [23,25].

With the growth in number of fully sequenced genomes, the number of proteins that turn out to be members of the α/β

*Corresponding author. Fax: +1 716 849 6732.
E-mail address: hkaur@bioinformatics.buffalo.edu (H.K. Saini).

0014-5793/$30.00
doi:10.1016/j.febslet.2005.04.030

hydrolase superfamily is constantly growing [23,24]. It is expected that new members of this family will be found as more genomes are sequenced. However, because of the large sequence diversity among the α/β hydrolases, identification of new family members is not always straightforward.

2. Results

In order to identify orthologous ORFan families within the *Bacillus* family, we used the "ORFans searcher" module of the ORFanage [31], the ORFan database, with *B. subtilis* and *B. halodurans* as queries. One of the orthologous ORFan families found consisted of the "hypothetical protein" yjaU from *Bacillus subtilis* (subsp. subtilis str. 168) [32] and the "unknown protein" BH2892 from *B. halodurans* (strain C-125) [33]. When using these two ORFans as queries against NCBI's nr database, two more sequences, from organisms not included in the ORFanage, were found: a sequence from *B. cereus* (strains ATCC 14579 and ATCC 10987) [34] and a sequence from *B. anthracis* (strains A2012 and Ames) [35]. We refer to the last two as Bcereus and Banthracis, respectively. The sequences' gene identification numbers are listed in Table 1. Notice that with the exception of *B. cereus* ATCC 14579, all sequences are annotated as hypothetical proteins.

When used as PSI-Blast [36] queries against the nr NCBI database, each of these sequences finds matches only to the other three sequences. Banthracis, in addition, also finds partial matches in its C-terminus to a number of Eukaryote proteins. These sequences are added to the Psi-Blast search in the second iteration with an initial score of 0.001. Most of Banthracis' sequence aligns only to the other three *Bacillus* sequences and can therefore be referred to as an "ORFan module" (a segment of a non-ORFan sequence which is an ORFan [1]). Thus, in what follows, we regard the four sequences as forming a family of orthologous ORFans.

Fig. 1 shows the multiple sequence alignment of the four ORFans. The ORFans are highly similar to each other, with pairwise sequence identity ranging from 38% to 44% among yjaU, BH2892, and Bcereus or Banthracis, and 97% identity among Bcereus and Banthracis. The secondary structure predicted for Banthracis is shown below the alignment. It can be noted that where the canonical β6 and β8 are expected to be found, a helix is predicted, the first one with a low level of confidence and the latter, perhaps due to being part of an extended 6, with a high level of confidence.

The Conserved Domain Database (CDD) [37], implemented in PSI-Blast, identifies the beginning of the PldB domain in all four sequences, starting from residue 50 (34–41% of the ORFans are aligned to PldB). PldB is a lysophospholipase L2 located in the inner membrane of *Escherichia coli* K-12 [38], and is classified under COG2267, "Lysophospholipase [Lipid metabolism]". It belongs to the α/β hydrolase fold superfamily, with a Ser–Glu–His catalytic triad. In addition, at a higher E-value, CDD identifies the beginning of the MhpC domain in Bcereus and Banthracis starting from residue 16 (39% aligned to MhpC). MhpC is a C–C bond hydrolase from *E. coli* [25], and is classified under COG0596, "Predicted hydrolases or acyltransferases (α/β hydrolase superfamily) [General function prediction only]". Its catalytic triad is Ser–Asp–His.

2.1. Fold-recognition results

To obtain insight regarding the 3D structure of the ORFans, we submitted each one of the four *Bacillus* sequences to the fold-recognition Meta-server at http://BioInfo.PL/Meta [39]. The sequence-structure alignments of Banthracis with two of its best templates, as identified by the Meta-server, were derived directly from the Meta-server's alignment. The alignment of the two templates relative to each other, with respect to their alignment to Banthracis, was adjusted using Dali [40].

Table 2 lists the top 10 predictions obtained for each ORFan by the Meta-server. For each of them, a highly significant 3D-jury score [41,42] above 100 was obtained. Such scores are well above the meta-servers confidence thresholds; large-scale tests [43] have identified that scores greater than 40 correspond to reliable predictions. The predictions point unanimously to the α/β hydrolase fold. However, the templates chosen by the servers belong to several families within this superfamily (e.g., epoxide hydrolase, haloalkane dehalogenase, haloperoxidase, peptidase; see Table 2), and therefore a straightforward classification of the ORFans into one of these families was not possible. It is important to point out that none of our ORFans share any significant sequence similarity to any of the identified templates. Consequently, it is likely that the sequence-structure alignments are not perfect; indeed, the alignments are ambiguous around the location of the catalytic triad, with the catalytic residues in the templates often matching residues in the ORFan sequences that do not demonstrate similar characteristics (e.g., a glycine in the query sequence was often aligned to the catalytic serine in the template). Thus, identification of the putative catalytic residues was not straightforward based on these alignments.

In summary, although fold-recognition predictions clearly indicate that these ORFans belong to the α/β hydrolase fold, it is not straightforward to generate an accurate 3D model or a detailed hypothesis regarding their exact function, or to which of the α/β hydrolase families they belong. This is not surprising

Table 1
Identification numbers and annotations of the *Bacillus* ORFans

Sequence name	Genome	gi number	Annotation
yjaU	*B. subtilis*	16078193|ref|NP 389010.1	Hypothetical protein
BH2892	*B. haloduarns*	15615455|ref|NP 243758.1	Unknown conserved protein in *Bacilli*
Bcereus	*B. cereus* ATCC 14579	30019325|ref|NP 830956.1	Putative hydrolase
	B. cereus ATCC 10987	42780364|ref|NP 977611.1	Conserved hypothetical protein
Banthracis	*B. anthracis* A2012	21399094|ref|NP 655079.1	Hypothetical protein predicted by GeneMark
	B. anthracis str. Ames	30261280|ref|NP 843657.1	Conserved hypothetical protein

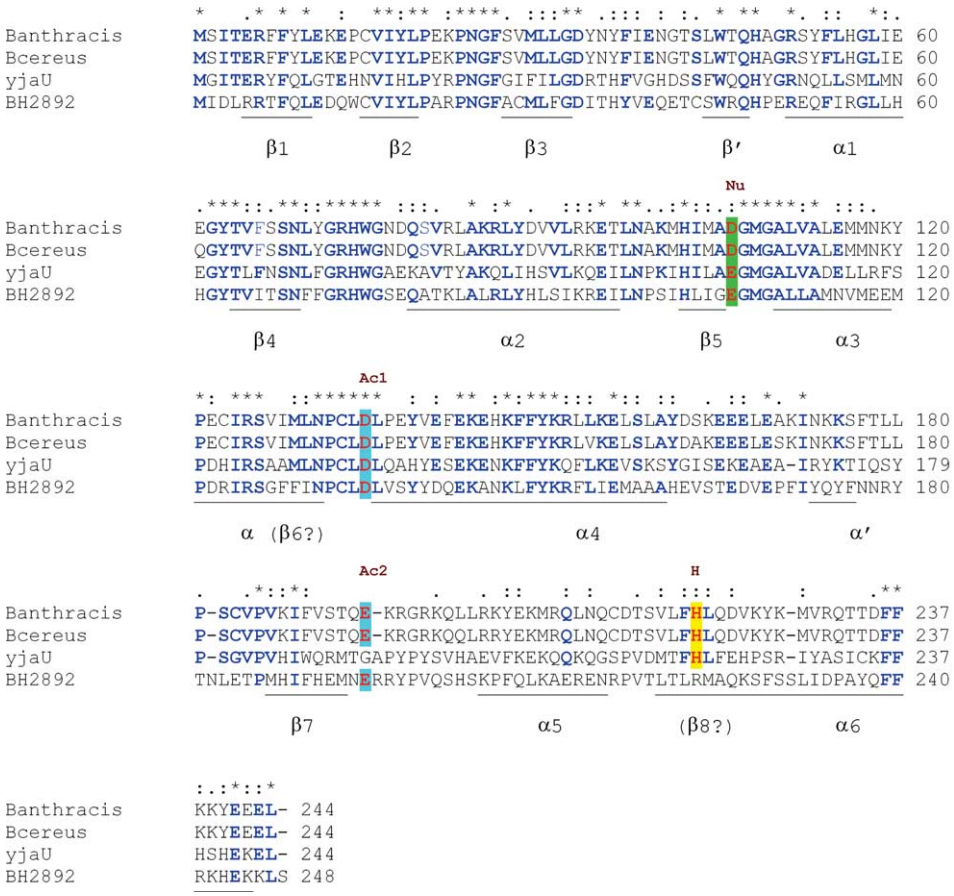

Fig. 1. Multiple sequence alignment of the four ORFans. Residues that are conserved in three or four sequences are colored in blue. The proposed catalytic triad consisting of nucleophile (Nu), acid (Ac) and histidine (H) are highlighted with red letters in color background. The predicted secondary structure elements [48] that correspond to those found in the canonical fold are numbered accordingly. Additional elements are marked with an apostrophe (a β strand after β3 and an α helix after α4). The multiple alignment of the four ORFans was performed using ClustalW [49].

because ORFans are among the hardest prediction targets in bioinformatics due to lack of evolutionary information.

2.2. Identification of the catalytic triad residues

Therefore, we looked for possible candidates for the catalytic triad residues using a combination of the sequence-structure alignments and the secondary structure predictions obtained for the four ORFans. In what follows, we demonstrate this procedure for Banthracis and its sequence-structure alignments with two of the best templates found by the Metaserver, 1bro and 1a8s, both haloperoxidases. 1bro is a bromoperoxidase from *Streptomyces aurefaciens* [44] and 1a8s is a chloroperoxidase from *Pseudomonas fluorescens* [45]. The catalytic triad in both proteins, as in other haloperoxidases, is Ser–Asp–His, corresponding to S98, D228 and H257 in both.

Fig. 2 depicts the sequence-structure alignments of Banthracis to each of 1bro and 1a8s. The secondary structure elements for these templates and predicted secondary structure for Banthracis are shown above the sequences. For clarity, the following segments in the templates' sequences that have no matching sequence in Banthracis were omitted: the beginning of α4, the end of α′ and two additional α-helices following α′. The beginning of 1bro, which contains its β1 strand, was not part of the sequence-structure alignment with Banthracis, and therefore it was omitted from the figure. The areas along the sequence where the two templates are structurally aligned to each other, within a 3-residue shift, consist of the canonical structure and are highlighted by gray boxes.

Fig. 3A shows the structure of 1bro with the catalytic triad residues shown in ball-and-sticks. The catalytic triad consists of Ser–Asp–His, with Ser98 located in nucleophilic elbow between β5 and α3, Asp228 and His257 located after β7 and β8, respectively. Fig. 3B shows our proposed model for Banthracis. This model was generated by the SHGU [46] fold-recognition method using 1bro as a template and was refined by NEST [47]. The possible location of catalytic residues is shown in ball-and-sticks.

2.2.1. The nucleophile. In both 1bro and 1a8s, the nucleophile is a serine. This serine, S98, is part of a conserved motif that is found at the nucleophilic elbow of many α/β hydrolases, G–X–S–X–G, where X is any amino acid [22,25]. S98 is directly aligned to G107 in Banthracis (Fig. 2). The position just before this glycine is occupied by an aspartic acid, D106, in Banthracis and Bcereus and a glutamic acid, E106, in yjaU and BH2892 (marked as "Nu" in Fig. 1). The aspartic or glutamic acid is part of a conserved motif among the four ORFans, {A/G}{D/E}GMG.

G107 in all ORFans is aligned to the nucleophile in several of the templates. In addition, it is immediately followed by a Methionine. It was previously found that the residue just following the nucleophile is related to the substrate binding and is often, but not always, an aromatic residue or a Methionine

Table 2
Fold-recognition results for the *Bacillus* ORFans

Sequence name	3D-Jury score	Server (model number)	PDB	PDB description
yjaU	112.67	SHGU (2)	1yas	Hydroxynitrile lyase
	112.11	SHGU (3)	1ehy	Epoxide hydrolase
	112.00	SHGU (5)	1a88	Chloroperoxidase L
	111.89	SHGU (0)	1mt3	Tricorn interacting factor selenomethionine-F1
	111.78	SHGU (6)	1cv2	Haloalkane dehalogenase
	111.22	SHGU (4)	1ede	Haloalkane dehalogenase
	111.00	SHGU (1)	1j1i	Serine hydrolase
	110.56	SHGU (8)	1azw	Proline iminopeptidase
	109.78	SHGU (7)	1bro	Bromoperoxidase A2
	109.75	3DS3 (0)	1j1iA	Serine hydrolase
BH2892	118.86	BasD (4)	1l7aA	Cephalosporin C deacetylase
	118.36	mBAS (0)	1pfqA	Dipeptidyl peptidase IV Cd26
	118.36	ORFs (0)	1pfqA	Dipeptidyl peptidase IV Cd26
	118.36	mBAS (3)	1l7aA	Cephalosporin C deacetylase
	118.36	ORFs (3)	1l7aA	Cephalosporin C deacetylase
	118.36	ORF2 (1)	1pfqA	Dipeptidyl peptidase IV Cd26
	118.36	ORF2 (3)	1l7aA	Cephalosporin C deacetylase
	118.14	mBAS (1)	1n1mA	Dipeptidyl peptidase IV/Cd26
	118.14	ORF2 (2)	1n1mA	Dipeptidyl peptidase IV/Cd26
	118.14	ORFs (1)	1n1mA	Dipeptidyl peptidase IV/Cd26
Bcereus	114.10	SHGU (2)	1bro	Bromoperoxidase A2
	114.00	SHGU (3)	1jli	Serine hydrolase
	113.78	SHGU (4)	1bro	Bromoperoxidase A2
	113.44	SHGU (6)	1bro	Bromoperoxidase A2
	112.89	SHGU (1)	1jli	Serine hydrolase
	112.56	SHGU (8)	1a88	Chloroperoxidase L
	110.89	SHGU (5)	1ehy	Epoxide hydrolase
	110.00	SHGU (7)	1ehy	Epoxide hydrolase
	109.56	SHGU (0)	1mt3	Tricorn interacting factor selenomethionine-F1
	100.12	3DS3 (0)	1mt3A	Tricorn interacting factor selenomethionine-F1
Banthracis	114.20	ORF2 (1)	1a8s	Chloroperoxidase
	113.27	ORF2 (7)	1a8q	Bromoperoxidase A1
	112.60	SHGU (1)	1bro	Bromoperoxidase A2
	112.20	ORFs (0)	1iunA	Meta-cleavage product hydrolase
	112.20	mBAS (2)	1iunA	Meta-cleavage product hydrolase
	112.20	ORF2 (3)	1iunA	Meta-cleavage product hydrolase
	111.67	mBAS (9)	1a8s	Chloroperoxidase
	111.47	ORFs (1)	1j1iA	Serine hydrolase
	111.47	ORFs (6)	1a8s	Chloroperoxidase
	111.47	mBAS (4)	1j1iA	Serine hydrolase

[29]. While this may imply that the nucleophile in all four ORFans has been deleted or mutated to a glycine and they are therefore non-functional, the existence of a conserved carboxylic acid in proximity to this location in all ORFans, which is part of a conserved motif, seems to have a functional significance. A shift of one residue in the position of the nucleophile may not alter dramatically the shape of the active site and may still enable the protein to retain its catalytic mechanism. Thus, in what follows, we refer to D106 as being the candidate nucleophile in Banthracis and Bcereus.

yjaU and BH2892 have a glutamic acid in this position. No enzyme was reported thus far as having a Glu as the nucleophile. The introduction of such a long residue in this location would probably call for a redesign of the active site defined by the fold [25]. Although, it is possible that yjaU and BH2892 are non-functional due to the nucleophile being mutated from Asp to Glu, the conserved mutation to a carboxylic acid and the conserved motif may hint that E106 in these sequences does have a functional role. It may even correspond to the nucleophile.

2.2.2. The acid. The location of the catalytic acid D228 in 1bro and 1a8s is the more common one, after strand β7. This residue is aligned with a glutamic acid in Banthracis (E194), Bcereus (E194) and BH2892 (E195) and with a glycine in yjaU (G193) (marked as "Ac2" in Fig. 1). Thus, it is possible that the catalytic triad of Banthracis and Bcereus consists of Asp–Glu–His, as seen previously in haloalkane-dehalogenases. However, if indeed E106 is the nucleophile and E195 is the acid in BH2892, then there are two Glu residues in the active site of this protein. The existence of two large residues in this region may cause large shifts in the structure of the oxyanion hole.

Another possibility is that the acid resides after strand β6 (marked as "Ac1" in Fig. 1). This would comply with the existence of G193 in yjaU, since several proteins which are known to be active α/β hydrolases have a non-acidic residue in the catalytic acid position after β7, such as a glycine, proline or asparagine [24,28,29] and their acid was found to reside after β6. Therefore, we looked for possible acid candidates after strand β6. Indeed we found in this location an aspartic acid, D135, that is conserved in all four ORFan sequences (Fig. 1).

2.2.3. The histidine. The catalytic histidine in 1bro and 1a8s is H257, located after β8. H257 is part of a conserved motif in both proteins, GAPHGL. The only histidine in the

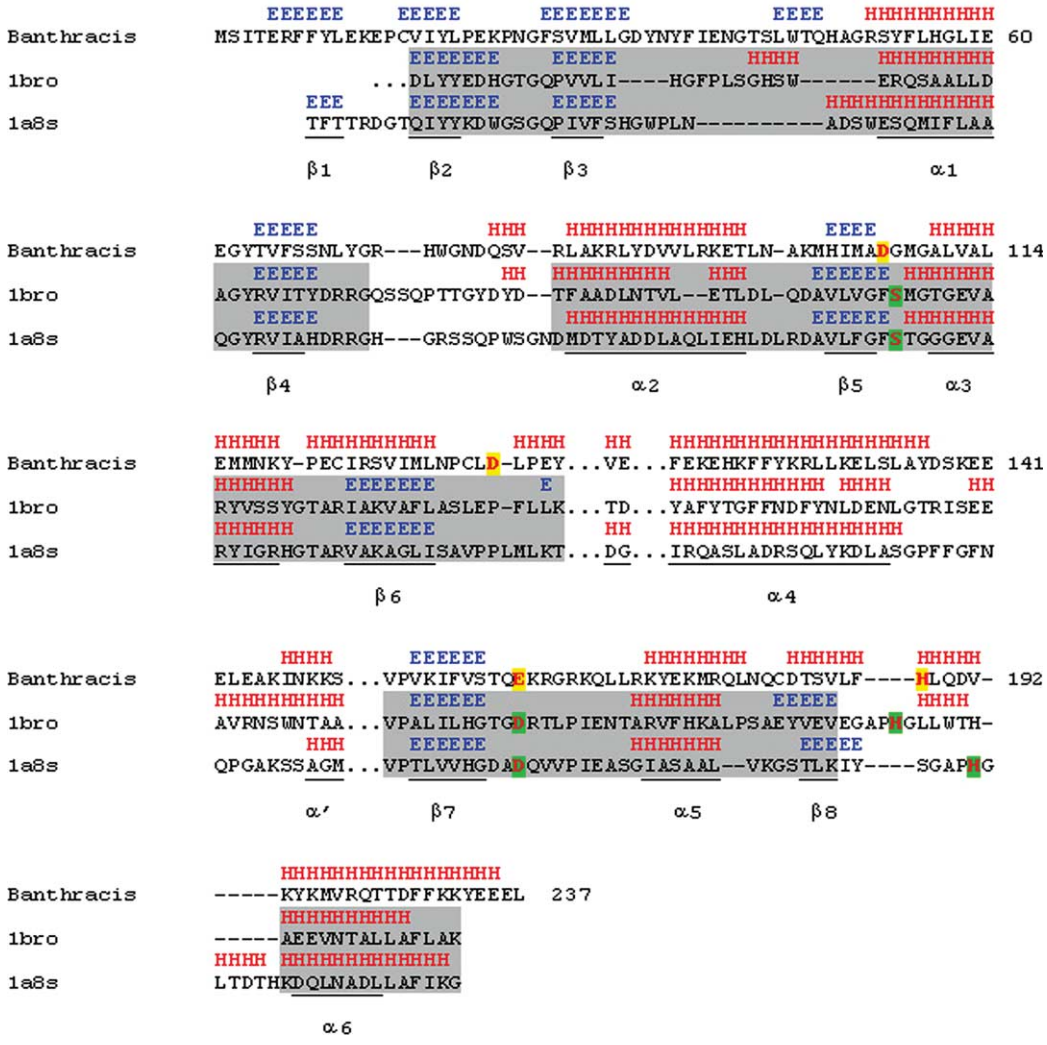

Fig. 2. Sequence-structure alignment of Banthracis with two templates. The sequence-structure alignment of the *B. anthracis* ORFan with each one of two best templates found by the servers, 1bro and 1a8s, both haloperoxidases (see Table 2). Beta strands are shown as blue sequences of the letter E, and alpha helices as red sequences of the letter H. The minimal common secondary structure elements of the three proteins (or the two templates, when Banthracis has a different assignment) are marked below the alignment. Secondary structure elements that correspond to those in the canonical fold are numbered accordingly. The catalytic residues of the templates are marked in red with green background. The suggested residues for Banthracis' catalytic triad are colored in red with yellow background.

Banthracis sequence located in this region is H221 (H221 also in Bcereus and yjaU), and it is part of a conserved motif among these three ORFans, FHL (Fig. 1). Aligned to the histidine residue in these three sequences is an Arginine in BH2892. No histidine is found in BH2892 at this location, suggesting that this sequence may not be functional anymore.

3. Discussion

We have identified a family of orthologous ORFans within the *Bacillus* family and have applied a computational analysis that can help us in the first stage of characterizing it. The classification of the ORFans into a general fold was fairly straightforward. Fold-recognition methods predict with very high levels of confidence that the ORFans belong to the α/β hydrolase fold superfamily. Sequence analyses support this general classification. However, the deeper levels of classification are less evident at the moment. A large number of proteins, belonging to several families within the α/β hydrolase superfamily and having different catalytic triads were identified by the fold-recognition servers as suitable templates for the ORFans. Therefore, detailed sequence analyses were needed in order to gain further insights. Based on these analyses, we suggest the following catalytic triads for the ORFans.

The catalytic triad in Banthracis and Bcereus could be D106–E194–H221, if the acid resides after β7 or D106–D135–H221 if it resides after β6. The first type of catalytic triad has been seen in the haloalkane-dehalogenase family while the second one has been seen in the dehalogenase and the epoxide-hydrolase families. The best templates found for Banthracis by the fold-recognition Meta-server are haloperoxidases having a catalytic triad consisting of Ser–Asp–His. These templates have three helices in the cap region as opposed to only one in Banthracis. Furthermore, analysis of the computational model shows that the side-chains of the proposed catalytic triad residues do not point towards a central cavity. This may be a consequence of the refinement. Thus, it is difficult to say, whether the location of

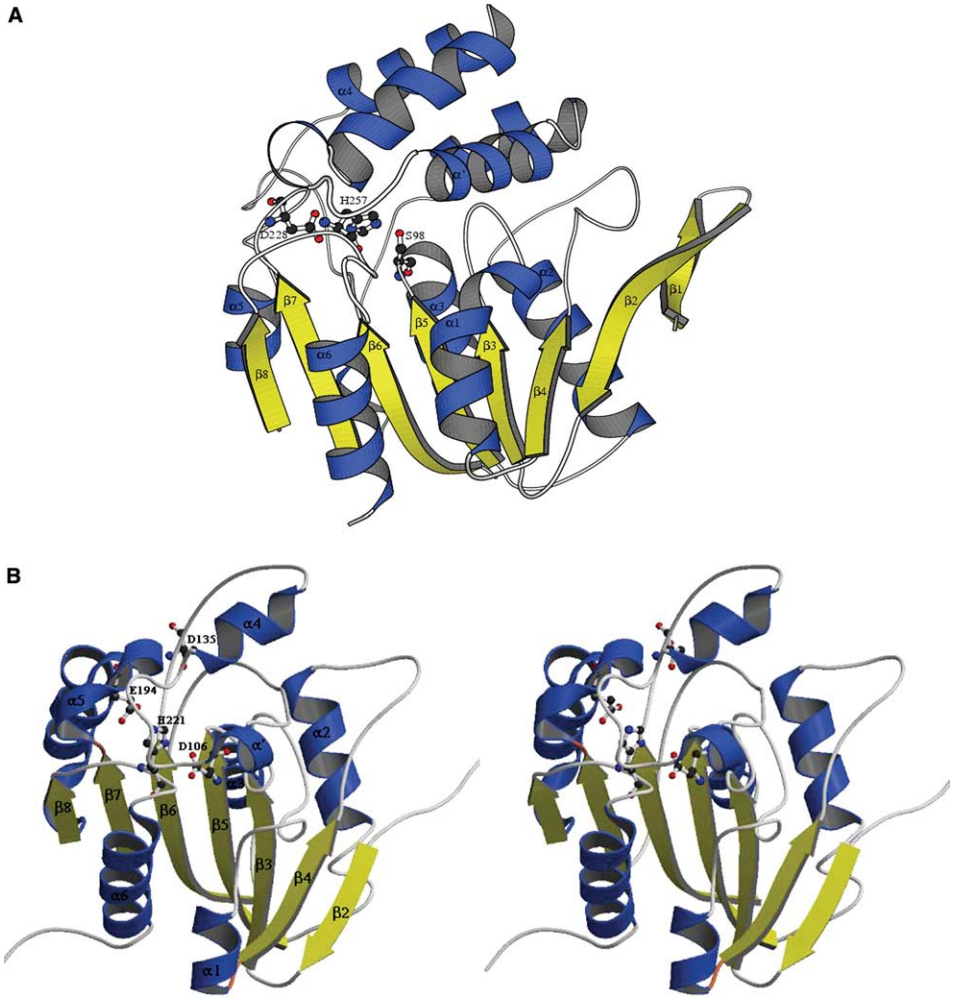

Fig. 3. 1 bro and the predicted model for Banthracis. (A) Ribbon representation of the template 1bro. α-Helices and β-strands are colored in blue and yellow, respectively, and are numbered numerically. The remaining secondary structure is in white color. The catalytic residues are represented as ball-and-sticks. The figure was drawn with MOLSCRIPT program [50]. (B) Our proposed model of the *B. anthracis* ORFan, generated by the fold-recognition program SHGU [46] and refined by NEST [47].

D135 makes it a probable member of the catalytic triad. Our analyses suggest that the ORFans may belong to the haloperoxidase family, but we could not rule out the possibility that they may belong to the dehalogenase or the epoxide-hydrolase families. Alternatively, they could belong to a novel family within the α/β hydrolase superfamily.

yjaU and BH2892 have a glutamic acid aligned to D106 in Banthracis and Bcereus. Since no known α/β hydrolase enzyme is known to have a glutamic acid as the nucleophile, it is possible that these two sequences correspond to non-functional proteins. This hypothesis is further supported in BH2892 since this sequence seems to also lack the essential catalytic histidine. The conserved substitution of an aspartic to a glutamic acid and the conserved motifs these sequences share with Banthracis and Bcereus around the suggested catalytic residues may show that they used to be active. It is possible that these two sequences are in the process of deterioration, the first step being the loss of the active site residues. Alternatively, it could be that yjaU is a functional protein with the novel catalytic triad E106–D135–H221, where the acid is located after β6.

Only experimental studies could verify whether all or some of these ORFans correspond to real, functional proteins, confirm their function and catalytic triad, and find to what family within the α/β hydrolase fold they belong. Nonetheless, our computational analysis lead to proposing a testable hypothesis that can serve as a first step towards characterizing these ORFans.

References

[1] Siew, N. and Fischer, D. (2003) Analysis of singleton ORFans in fully sequenced microbial genomes. Proteins 53, 241–251.
[2] Makarova, K.S. and Koonin, E.V. (2003) Comparative genomics of Archaea: how much have we learned in six years, and what's next? Genome Biol. 4, 115.
[3] Herrero, E., de la Torre, M.A. and Valentin, E. (2003) Comparative genomics of yeast species: new insights into their biology. Int. Microbiol. 6, 183–190.
[4] Boucher, Y. and Doolittle, W.F. (2002) Something new under the sea. Nature 417, 27–28.
[5] Siew, N. and Fischer, D. (2003) Twenty thousand ORFan microbial protein families for the biologist? Structure (Camb) 11, 7–9.
[6] Siew, N. and Fischer, D. (2003) Unraveling the ORFan puzzle. Comp. Funct. Genom. 4, 432–441.
[7] Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman,

S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M. and Barrell, B. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419, 498–511.

[8] Siew, N. and Fischer, D. (2004) Structural biology sheds light on the puzzle of genomic ORFans. J. Mol. Biol. 342, 369–373.

[9] Dujon, B. (1996) The yeast genome project: what did we learn? Trends Genet. 12, 263–270.

[10] Mackiewicz, P., Kowalczuk, M., Gierlik, A., Dudek, M.R. and Cebrat, S. (1999) Origin and properties of non-coding ORFs in the yeast genome. Nucleic Acids Res. 27, 3503–3509.

[11] Schmid, K.J. and Aquadro, C.F. (2001) The evolutionary analysis of orphans from the *Drosophila* genome identifies rapidly diverging and incorrectly annotated genes. Genetics 159, 589–598.

[12] Skovgaard, M., Jensen, L.J., Brunak, S., Ussery, D. and Krogh, A. (2001) On the total number of genes and their length distribution in complete microbial genomes. Trends Genet. 17, 425–428.

[13] Wood, V., Rutherford, K.M., Ivens, A., Rajandream, M.-A. and Barrell, B. (2001) A reannotation of the *Saccharomyces cerevisiae* genome. Comp. Funct. Genom. 2, 143–154.

[14] Amiri, H., Davids, W. and Andersson, S.G. (2003) Birth and death of orphan genes in Rickettsia. Mol. Biol. Evol. 20, 1575–1587.

[15] Eisenstein, E., Gilliland, G.L., Herzberg, O., Moult, J., Orban, J., Poljak, R.J., Banerjei, L., Richardson, D. and Howard, A.J. (2000) Biological function made crystal clear – annotation of hypothetical proteins via structural genomics. Curr. Opin. Biotechnol. 11, 25–30.

[16] Zhang, C. and Kim, S.H. (2003) Overview of structural genomics: from structure to function. Curr. Opin. Chem. Biol. 7, 28–32.

[17] Kim, S.H. (2000) Structural genomics of microbes: an objective. Curr. Opin. Struct. Biol. 10, 380–383.

[18] Brenner, S.E. (2000) Target selection for structural genomics. Nat. Struct. Biol. 7 (Suppl.), 967–969.

[19] Zarembinski, T.I., Hung, L.W., Mueller-Dieckmann, H.J., Kim, K.K., Yokota, H., Kim, R. and Kim, S.H. (1998) Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. Proc. Natl. Acad. Sci. USA 95, 15189–15193.

[20] Goulding, C.W., Parseghian, A., Sawaya, M.R., Cascio, D., Apostol, M.I., Gennaro, M.L. and Eisenberg, D. (2002) Crystal structure of a major secreted protein of *Mycobacterium tuberculosis*-MPT63 at 1.5-A resolution. Protein Sci. 11, 2887–2893.

[21] Teplyakov, A., Obmolova, G., Chu, S.Y., Toedt, J., Eisenstein, E., Howard, A.J. and Gilliland, G.L. (2003) Crystal structure of the YchF protein reveals binding sites for GTP and nucleic acid. J. Bacteriol. 185, 4031–4037.

[22] Ollis, D.L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S.M., Harel, M., Remington, S.J., Silman, I. and Schrag, J., et al. (1992) The alpha/beta hydrolase fold. Protein Eng. 5, 197–211.

[23] Heikinheimo, P., Goldman, A., Jeffries, C. and Ollis, D.L. (1999) Of barn owls and bankers: a lush variety of alpha/beta hydrolases. Struct. Fold Des. 7, R141–R146.

[24] Nardini, M. and Dijkstra, B.W. (1999) Alpha/beta hydrolase fold enzymes: the family keeps growing. Curr. Opin. Struct. Biol. 9, 732–737.

[25] Holmquist, M. (2000) Alpha/beta-hydrolase fold enzymes: structures, functions and mechanisms. Curr. Protein. Pept. Sci. 1, 209–235.

[26] Wei, Y., Contreras, J.A., Sheffield, P., Osterlund, T., Derewenda, U., Kneusel, R.E., Matern, U., Holm, C. and Derewenda, Z.S. (1999) Crystal structure of brefeldin A esterase, a bacterial homolog of the mammalian hormone-sensitive lipase. Nat. Struct. Biol. 6, 340–345.

[27] Nardini, M., Ridder, I.S., Rozeboom, H.J., Kalk, K.H., Rink, R., Janssen, D.B. and Dijkstra, B.W. (1999) The X-ray structure of epoxide hydrolase from *Agrobacterium radiobacter* AD1. An enzyme to detoxify harmful epoxides. J. Biol. Chem. 274, 14579–14586.

[28] Hynkova, K., Nagata, Y., Takagi, M. and Damborsky, J. (1999) Identification of the catalytic triad in the haloalkane dehalogenase from *Sphingomonas paucimobilis* UT26. FEBS Lett. 446, 177–181.

[29] Fischer, F., Kunne, S. and Fetzner, S. (1999) Bacterial 2,4-dioxygenases: new members of the alpha/beta hydrolase-fold superfamily of enzymes functionally related to serine hydrolases. J. Bacteriol. 181, 5725–5733.

[30] Krooshof, G.H., Kwant, E.M., Damborsky, J., Koca, J. and Janssen, D.B. (1997) Repositioning the catalytic triad aspartic acid of haloalkane dehalogenase: effects on stability, kinetics, and structure. Biochemistry 36, 9571–9580.

[31] Siew, N., Azaria, Y. and Fischer, D. (2004) The ORFanage: an ORFan database. Nucleic Acids Res. 32, D281–D283.

[32] Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.K., Codani, J.J., Connerton, I.F. and Danchin, A., et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390, 249–256.

[33] Takami, H., Nakasone, K., Takaki, Y., Maeno, G., Sasaki, R., Masui, N., Fuji, F., Hirama, C., Nakamura, Y., Ogasawara, N., Kuhara, S. and Horikoshi, K. (2000) Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. Nucleic Acids Res. 28, 4317–4331.

[34] Ivanova, N., Sorokin, A., Anderson, I., Galleron, N., Candelon, B., Kapatral, V., Bhattacharyya, A., Reznik, G., Mikhailova, N., Lapidus, A., Chu, L., Mazur, M., Goltsman, E., Larsen, N., D'Souza, M., Walunas, T., Grechkin, Y., Pusch, G., Haselkorn, R., Fonstein, M., Ehrlich, S.D., Overbeek, R. and Kyrpides, N. (2003) Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. Nature 423, 87–91.

[35] Read, T.D., Peterson, S.N., Tourasse, N., Baillie, L.W., Paulsen, I.T., Nelson, K.E., Tettelin, H., Fouts, D.E., Eisen, J.A., Gill, S.R., Holtzapple, E.K., Okstad, O.A., Helgason, E., Rilstone, J., Wu, M., Kolonay, J.F., Beanan, M.J., Dodson, R.J., Brinkac, L.M., Gwinn, M., DeBoy, R.T., Madpu, R., Daugherty, S.C., Durkin, A.S., Haft, D.H., Nelson, W.C., Peterson, J.D., Pop, M., Khouri, H.M., Radune, D., Benton, J.L., Mahamoud, Y., Jiang, L., Hance, I.R., Weidman, J.F., Berry, K.J., Plaut, R.D., Wolf, A.M., Watkins, K.L., Nierman, W.C., Hazen, A., Cline, R., Redmond, C., Thwaite, J.E., White, O., Salzberg, S.L., Thomason, B., Friedlander, A.M., Koehler, T.M., Hanna, P.C., Kolsto, A.B. and Fraser, C.M. (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. Nature 423, 81–86.

[36] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

[37] Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., Liebert, C.A., Liu, C., Madej, T., Marchler, G.H., Mazumder, R., Nikolskaya, A.N., Panchenko, A.R., Rao, B.S., Shoemaker, B.A., Simonyan, V., Song, J.S., Thiessen, P.A., Vasudevan, S., Wang, Y., Yamashita, R.A., Yin, J.J. and Bryant, S.H. (2003) CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res. 31, 383–387.

[38] Kobayashi, T., Kudo, I., Karasawa, K., Mizushima, H., Inoue, K. and Nojima, S. (1985) Nucleotide sequence of the pldB gene and characteristics of deduced amino acid sequence of lysophospholipase L2 in *Escherichia coli*. J. Biochem. (Tokyo) 98, 1017–1025.

[39] Bujnicki, J.M., Elofsson, A., Fischer, D. and Rychlewski, L. (2001) Structure prediction meta server. Bioinformatics 17, 750–751.

[40] Holm, L. and Sander, C. (1995) Dali: a network tool for protein structure comparison. Trends Biochem. Sci. 20, 478–480.

[41] Ginalski, K., Elofsson, A., Fischer, D. and Rychlewski, L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19, 1015–1018.

[42] Ginalski, K. and Rychlewski, L. (2003) Detection of reliable and unexpected protein fold predictions using 3D-Jury. Nucleic Acids Res. 31, 3291–3292.
[43] Rychlewski, L. and Fischer, D. (2005) LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. Protein Sci. 14, 240–245.
[44] Hecht, H.J., Sobek, H., Haag, T., Pfeifer, O. and van Pee, K.H. (1994) The metal-ion-free oxidoreductase from *Streptomyces aureofaciens* has an alpha/beta hydrolase fold. Nat. Struct. Biol. 1, 532–537.
[45] Hofmann, B., Tolzer, S., Pelletier, I., Altenbuchner, J., van Pee, K.H. and Hecht, H.J. (1998) Structural investigation of the cofactor-free chloroperoxidases. J. Mol. Biol. 279, 889–900.
[46] Fischer, D. (2003) 3DS3 and 3DS5 3D-SHOTGUN meta-predictors in CAFASP3. Proteins 53 (Suppl. 6), 517–523.
[47] Petrey, D., Xiang, Z., Tang, C.L., Xie, L., Gimpelev, M., Mitros, T., Soto, C.S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., Koh, I.Y., Alexov, E. and Honig, B. (2003) Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. Proteins 53 (Suppl. 6), 430–435.
[48] McGuffin, L.J., Bryson, K. and Jones, D.T. (2000) The PSIPRED protein structure prediction server. Bioinformatics 16, 404–405.
[49] Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. and Thompson, J.D. (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31, 3497–3500.
[50] Kraulis, P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24, 946–950.