FEBS Letters 586 (2012) 3908–3913

FEBS Letters

journal homepage: www.FEBSLetters.org

A new family of bacterial DNA repair proteins annotated by the integration of non-homology, distant homology and structural bioinformatic methods

Luciane V. Mello [a,b], Daniel J. Rigden [a,*]

[a] Structural and Chemical Biology, Institute of Integrative Biology, University of Liverpool, UK
[b] School of Life Sciences, Institute of Teaching and Learning, University of Liverpool, UK

ARTICLE INFO

Article history:
Received 3 August 2012
Revised 13 September 2012
Accepted 14 September 2012
Available online 26 September 2012

Edited by Paul Bertone

Keywords:
DNA repair
Function annotation
Domain of unknown function
DUF2086
2-Oxoglutarate
Fe^{2+}-dependent oxygenase
Genome context

ABSTRACT

Different bioinformatics methods illuminate different aspects of protein function, from specific catalytic activities to broad functional categories. Here, a triple-pronged approach to predict function for a domain of unknown function, DUF2086, is applied. Distant homology to characterised enzymes and conservation of key residues suggest an oxygenase function. Modelling indicates that the substrate is most likely a nucleic acid. Finally, genomic context analysis linking DUF2086 to DNA repair, leads to a predicted activity of oxidative demethylation of damaged bases in DNA. The newly assigned activity is sporadically present in phyla not containing near relatives of the similarly active repair protein AlkB.

1. Introduction

Predicting function from sequence and structure is a major endeavour of protein bioinformatics [1]. Protein function may be considered and described at several different levels: the same protein might be most aptly and helpfully considered as a plant defence protein in one context or as a β-1,4-endoglucanase in another [2]. Similarly, bioinformatics methods shed light on different levels of protein function. At the molecular level, a demonstrable evolutionary relationship between a novel protein and a previously characterised enzyme might suggest that the former has the same or a similar activity to the latter [3]. Such an approach is only completely straightforward at high pairwise sequence identities [4,5] but is still often valuable in cases of distant relationships, detected by sensitive sequence comparisons [6–8], provided the results are carefully manually checked and, perhaps, validated by structure modelling. In contrast, non-homology methods such as genome context, gene fusion and phylogenetic profiling [9] provide much broader functional connections between genes or proteins, suggesting that they may participate in the same cellular process without providing specific molecular predictions.

The advent of pyrosequencing combining with metagenomic sampling methods has allowed for a new wave of protein sequence information from previously intractable origins [10]. Although estimates of the rate of increase vary [10,11], this data deluge reveals proteins and protein families bearing no obvious relationship, by routinely applicable sequence comparisons, to currently annotated proteins [11–13]. Even in model organisms, surprising numbers of sequences remain annotated solely as putative or hypothetical proteins. For example, a recent reannotation of an *Escherichia coli* strain concluded with more than a third of predicted proteins lacking predicted function [14]. Fortunately, while the volume of data offer challenges, it also offers opportunities: larger numbers of genomes increase the number of known gene fusion events and improve the statistical power of genome context and phylogenetic profiling methods [15]. Homology methods also benefit, through the discovery of sequences 'bridging' known families and through increased memberships of families, the added information allowing more distant evolutionary relationships to be newly discerned.

A particular challenge to computational annotation are superfamilies spanning functionally divergent groups. Historically these have been particularly susceptible to mis- and over-annotation (e.g. [16]) and are the targets of genome-scale projects, both bioinformatic [17,18] and crystallographic [19]. One such superfamily is the 2-oxoglutarate, Fe^{2+}-dependent oxygenases (here abbreviated

* Corresponding author. Address: Institute of Integrative Biology, University of Liverpool, Crown St., Liverpool L69 7ZB, UK. Fax: +44 151 795 4414.
E-mail address: drigden@liv.ac.uk (D.J. Rigden).

0014-5793/$36.00
http://dx.doi.org/10.1016/j.febslet.2012.09.023

as 2OG oxygenases) [20–22], themselves part of a much larger group called the cupins [23]. The 2OG oxygenases catalyse the incorporation of an oxygen atom, deriving from O_2, into diverse substrates including proteins, nucleic acids and intermediary metabolites [20–22,24]. Here, we apply a triple-pronged bioinformatics approach to annotate DUF2086 (COG3826) based on initial assignment to it of a 2OG superfamily fold. A model of a DUF2086 protein exhibits a strongly electrostatically positive surface near the presumed catalytic site and is strongly predicted to bind DNA. Finally, by genomic context we find that DUF2086 proteins are consistently linked to genes for repair of alkylated DNA. Taken together, the data predict that DUF2086 members act, like their distant AlkB relatives, in the direct repair by oxidative demethylation of damaged DNA. The DUF2086 distribution includes some Firmicutes, a phylum lacking annotated AlkB genes.

2. Materials and methods

HHsearch [7] was used to discern distant homologies between DUFs and other entries in domain databases or proteins of known structure. HHsearch assigns a probability value to a match based on sequence similarity of hidden Markov models representing query and database entry, supplemented by consideration of the match of (predicted) secondary structure between them. In this work only probabilities of greater than 0.8, corresponding to very confident matches, were considered.

HHsearch was used to obtain initial alignments of sequences with possible templates for comparative model building. Metal and 2-oxoglutarate ligands were included in the model and were derived from the structure of *E. coli* AlkB (PDB code 3khc; [25]). The major template used was putative oxygenase from *Shewanella baltica* (PDB code 3dkq; unpublished). Structures were superimposed with MUSTANG [26] and the resulting alignments processed with STACCATO [27]. Models were built with MODELLER [28]. Since the alignments implied the existence of two large insertions in the target with respect to principal template (Fig. 1), the specialised loop modelling protocol of MODELLER was used. For insertion 1, secondary structure restraints deriving from an α-helical prediction by PSIPRED [29] were incorporated. For later models, the structure of an algal prolyl hydroxylase (3gze [30]) was additionally used as template for insertion 2. This was because the 3gze loop is closer in length to that of the target than the much shorter loop in 3dkq and bears sequence similarity to the target. Improved model quality resulted from the use of 3gze as an additional template. The DOPE score [31] of MODELLER, VERIFY_3D [32] and PROCHECK [33] stereochemical analysis were used for model validation e.g. comparison of different possible locations for the insertions in the DUF2086 protein. MODELLER was also used to measure percentage sequence identities. MUSCLE [34] was employed for sequence alignment and Jalview 2 [35] for visualisation and manipulation of the results. An alignment of all DUF2086 full-length sequences, reduced in redundancy to a 95% level in Jalview 2, was used in conjunction with the CONSURF server [36], to map sequence conservation onto the final model. DNA_BIND [37] was used for (model) structure-based prediction of DNA binding ability and APBS to calculate electrostatics [38].

Local mining of STRING [15] with proteins in DUF2086 discovered multiple strong (score > 0.6) connections to families of DNA repair proteins. Such families derived from the eggNOG database [39], which encompasses and extends the COG/KOG database [40]. Genome contexts were visualised in STRING, with BAGET [41] used to discover additional families of DNA repair-related genes in the vicinity of genes coding for DUF2086 proteins. Reciprocal genome BLAST [42] runs were used to confirm relationships within the new families. Phyletic distributions, in particular to determine which species contain a DUF2086 protein but no regular AlkB, were analysed by comparing between Pfam and COG entries and confirmed with BLAST searches.

3. Results and discussion

3.1. DUF2086 contains a 2OG superfamily fold

HHsearch results, based on comparison of hidden Markov models, confidently located a 2OG superfamily fold in DUF2086. For example, with the *Ralstonia solanacearum* protein (UniProt ID: Q8XWA7; locus name RSc2567) searching against Pfam entries, probability scores included 99% for 2OG-FeII_Oxy_3 (PF13640,

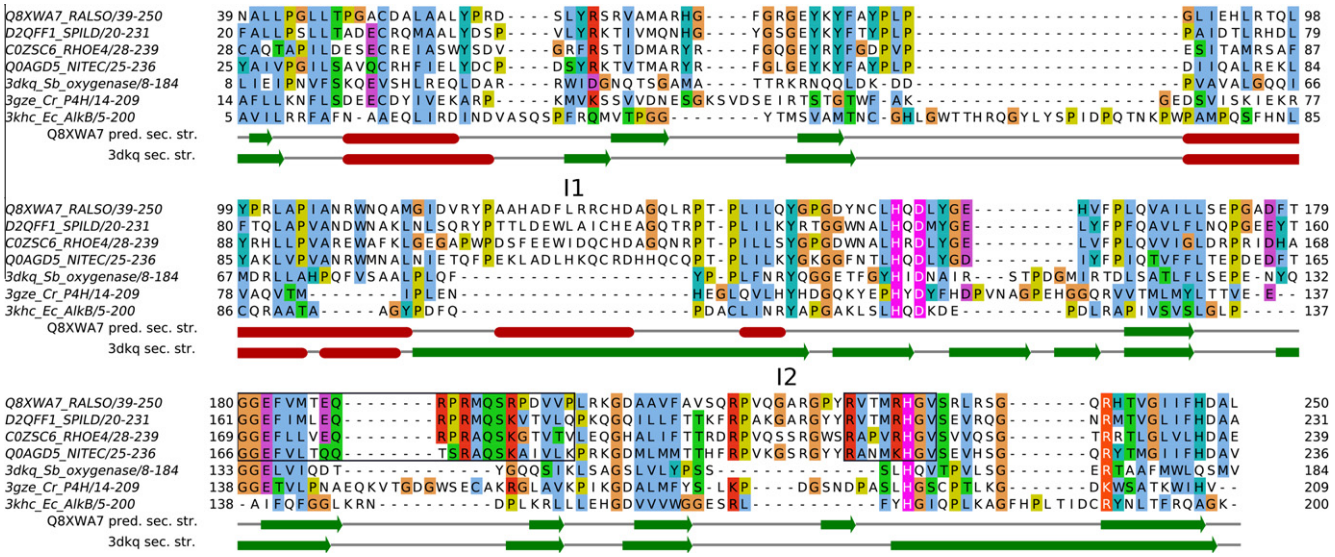

Fig. 1. Sequence alignment of the 2OG oxygenase domains of four representative DUF2086 proteins with three experimentally determined structures. DUF2086 proteins are given their UniProt identifiers, structures are shown as PDB code, followed by abbreviated species (Sb, *Shewanella baltica*, Cr *Chlamydomonas reinhardtii* and Ec, *Escherichia coli*) and activity (P4H, prolyl 4-hydroxylase). Comparison of the predicted secondary structure of the modelled DUF2086 sequence and the actual secondary structure of the principal template, 3dkq, is shown beneath the alignment. Key metal or 2-oxoglutarate binding residues are picked out in pink or orange, respectively. Boxed DUF2086 regions are the principal contributors to a conserved patch near the catalytic site that may be involved in substrate binding (see text).