Biochemical and Biophysical Research Communications 440 (2013) 68–75

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Episodic positive selection during the evolution of naphthalene dioxygenase to nitroarene dioxygenase

Arindam Dutta, Joydeep Chakraborty, Tapan K. Dutta *

Department of Microbiology, Bose Institute, Kolkata 700054, India

ARTICLE INFO

Article history:
Received 29 August 2013
Available online 13 September 2013

Keywords:
Aromatic ring-hydroxylating oxygenase
Naphthalene dioxygenase
Nitroarene dioxygenase
Adaptive evolution
Episodic positive selection

ABSTRACT

Using different maximum-likelihood models of adaptive evolution, signatures of natural selective pressure, operating across the naphthalene family of dioxygenases, were examined. A lineage- and branch-site specific combined analysis revealed that purifying selection pressure dominated the evolutionary history of the enzyme family. Specifically, episodic positive Darwinian selection pressure, affecting only a few sites in a subset of lineages, was found to be responsible for the evolution of nitroarene dioxygenases (NArDO) from naphthalene dioxygenase (NDO). Site-specific analysis confirmed the absence of diversifying selection pressure at any particular site. Different sets of positively selected residues, obtained from branch-site specific analysis, were detected for the evolution of each NArDO. They were mainly located around the active site, the catalytic pocket and their adjacent regions, when mapped onto the 3D structure of the α-subunit of NDO. The present analysis enriches the current understanding of adaptive evolution and also broadens the scope for rational alteration of substrate specificity of enzyme by directed evolution.

1. Introduction

Microorganisms, particularly bacteria, play a crucial role in global bio-geochemical cycles of aromatic compounds using an array of catabolic enzymes. Aromatic ring-hydroxylating oxygenases (RHOs) constitute one such superfamily of enzymes that participate in the oxidative metabolism of a wide variety of aromatic compounds of pharmaceutical, agricultural and environmental significance [1,2]. Though the members of this superfamily work in conjugation with either one or two electron transport proteins (*viz.* ferredoxins and reductases), the substrate recognition and the subsequent catalysis is actually conferred by the large subunit (α) of the $\alpha_n\beta_n$ or α_n type terminal oxygenases [3,4].

The enormous structural variation of aromatics in the environment is responsible for the evolution of various RHO enzyme systems in such a way that they can transform a wide range of such compounds [1]. Several classification schemes have been proposed for RHOs [5,6], the most recent being that of Chakraborty et al. (2012) [7] where the RHOs have been classified into five 'Classes' and eleven 'Types' on the basis of their evolutionary and functional behaviors, in relation to structural configuration of substrates and preferred oxygenation site(s). According to that classification, naphthalene dioxygenases (NDO), together with four types of nitroarene dioxygenases (NArDO) *viz.* nitrobenzene dioxygenase (NBDO), 2-nitrotoluene dioxygenase (2NTDO), 2-chloronitrobenzene dioxygenase (2clNBDO) and 2,4-dinitrotoluene dioxygenase (2,4DNTDO), constitute the naphthalene family of RHOs, which belong to A-IIIαβ type of RHOs.

It may be presumed that NArDOs must have been an outcome of relatively recent evolution since nitroaromatic compounds have been introduced into the environment by human activities. The hypothesis that a direct selective pressure operated during the evolution of NArDOs from NDO was proposed on the basis of the observation that nitroarene degrading bacteria are commonly found only in the nitroarene-contaminated sites [8,9]. A strong evolutionary relationship has indeed been observed among the catabolic operon(s) of nitroarene and naphthalene degradation pathways with frequent occurrence of pseudogenes in nitroarene degradation operons often characterized with deletions and/or frameshift mutations [10,11]. Though the core 3D structure of the catalytic α-subunit of both NDO (PDB: 1NDO) and NBDO (PDB: 2BMO) are quite conserved, nevertheless, the regio- and enantioselectivity of the reactions catalyzed by these enzymes exhibit distinct variations [8] and are anticipated to be the manifestation of variations in amino acid residues within their substrate cleft [8].

A comparison of rate of nonsynonymous substitutions per nonsynonymous site (d_N) to the rate of synonymous substitution per synonymous site (d_S) is an effective approach to detect the nature of selective pressure ($d_N/d_S < 1$, $d_N/d_S = 1$, $d_N/d_S > 1$ represents

* Corresponding author. Address: Department of Microbiology, Bose Institute, P-1/12 C.I.T. Scheme VII M, Kolkata 700054, India. Fax: +91 33 2355 3886.
E-mail address: tapan@jcbose.ac.in (T.K. Dutta).

0006-291X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2013.09.029

negative (purifying), neutral, positive selection, respectively) acting on protein-coding genes. Statistical phylogenetics offers various tools to measure the d_N/d_S ratio (ω) for a set of protein-coding nucleotide sequences along the lineages and/or of individual sites [12,13]. The aim of this study is to unravel the signature of natural selection with a view to test the hypothesis that NArDO evolved from NDO and to identify amino acid residues and regions in NDO/NArDO that have evolved under positive selection.

2. Materials and methods

2.1. Sequence retrieval and construction of dataset

The coding sequence of catalytic α-subunit of NBDO (GenBank: AF379638) was used as a query sequence in BLASTn [14] search using the default parameters and 34 full length homologous sequences (showing ⩾99% query coverage and ⩾75% identity) belonging to the naphthalene family of RHOs were retrieved. Corresponding protein sequences of these genes were also downloaded for analysis. Redundant sequences as well as those bearing internal stop codon were excluded from the downloaded set of sequences. Sequences (GenBank: JN655512, GQ184726) showing recombination signals based on recombination analysis using RDP3 (Recombination Detection Program Version3) [15] were discarded from the analysis. Moreover, absence of saturation of synonymous sites in our refined set of sequences was ensured following the methods described by Lynn et al. [16] where pairwise d_N and d_S values were calculated using the maximum-likelihood method implemented in codeml program in PAML4.5 package [13] followed by a correlation analysis using SPSS [17]. Two different datasets, one containing 9 sequences (small dataset) and another containing 21 sequences (large dataset), were finally built (Table S1) to examine the robustness of analysis against dataset size. Both small and large datasets contained 5 NArDOs, apart from 4 and 16 evolutionarily distinct NDOs, respectively.

2.2. Phylogenetic analyses

Nucleotide sequences of both the datasets were aligned and phylogenetic trees were constructed using neighbor-joining algorithm as implemented in MEGA5 [18]. Confidence values for the nodes were obtained by bootstrapping (100 replicates). The nucleotide sequences for each of the datasets were also aligned as codons using the same program. Codon alignments and trees generated for each of the datasets were used in the subsequent analyses.

2.3. Lineage-specific selection analysis

Lineage-specific selection analysis was performed individually for both datasets to detect variation in selection pressure across lineages and to identify lineages with elevated ω values indicating episodes of positive Darwinian selection. Two different approaches were utilized for this purpose *viz*. Genetic Algorithm (GA)-branch method [19] and maximum-likelihood method [12,13].

The GA-branch analysis was performed using the DATAMONKEY web server [20]. This method comes up with optimum branch partition pattern according to ω values but sheds no light on other evolutionary aspects, such as κ value (transition–transversion rate ratio), branch lengths (number of nucleotide substitutions per codon) and number of synonymous/non-synonymous substitutions along the lineages. To overcome these limitations, two different branch models (referred to here as GA-models), one for each dataset, were constructed by codeml program [13] following the identical branch partition pattern, as obtained from the best-fitting models using GA-branch method. During construction of the GA-models, branch labels were set to define various branch-rate classes. The GA-model, one- and free-ratio models were fitted by maximum-likelihood to each dataset followed by comparison of GA-model to one- and free-ratio model by likelihood ratio test (LRT) [21].

2.4. Site-specific selection analysis

In order to infer sites potentially under diversifying selection pressure during the evolution of NDO enzyme family, four site-specific likelihood models, namely M1a, M2a, M7 and M8, were fitted by maximum-likelihood [22,23] and the likelihood values under model M1a and M7 were compared with M2a and M8 respectively by LRT. M1a and M7 are the null models in which positive selection is not allowed while M2a and M8 are the alternative models that allow positive selection. Comparisons of M1a and M7 with M2a and M8 respectively are performed to assess gene evolution under positive selection, which is ensured by the acceptance of alternative models.

2.5. Branch-site specific selection analysis

Both datasets were considered for this analysis to detect positive selection affecting sites along specific lineage(s) (designated as foreground branch) with ω value greater than one as found in the lineage specific analysis. This was performed by fitting modified Branch-site model A and its corresponding null model and subsequent comparison of their likelihood values using LRT. In this study, branch-site analysis was performed in combination with site-stripping method [24] to stringently identify sites, contributing to the significance of the modified branch-site model A over its null model, among the positively selected sites with Bayesian posterior probabilities (*P*-value) >0.95. Before performing this method, the positively selected sites identified by branch-site model A were arranged in descending order of their *P*-values and then the sites were removed sequentially from the alignment starting from the site for which highest *P*-value was obtained. Each resulting stripped alignment was reanalyzed using the same models and parameters. This process was continued until the LRT failed. Sites, whose removal from the alignment resulted in failure of the LRT, were taken to be the sites making the alternate model significant.

3. Results and discussion

3.1. Lineage-specific selection analysis

3.1.1. GA-branch analysis

GA-branch analyses [19] allowed us to identify best-fitting models that led to the classification of lineages within phylogenetic tree into four and six branch-rate classes, for small and large dataset respectively, suggesting the presence of heterogeneity in the ω value among lineages (Fig. 1). Two branch-rate classes, each for small and large datasets, exhibited $\omega > 1$, indicating that some of the lineages have experienced positive selection (Fig. 1). The lineages exhibiting high percentage support in favor of positive selection are summarized in Table S2.

3.1.2. Maximum-likelihood analysis

For both the datasets, ω values displayed by the branch-rate classes were compared with those exhibited by the corresponding GA-models, constructed using the codeml program [13] (Table S3). The values were found to be quite similar indicating that the best-fitting models and their corresponding GA-models are equivalent