Journal of Biotechnology 143 (2009) 17–26

Contents lists available at ScienceDirect

Journal of Biotechnology

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

Investigative mining of sequence data for novel enzymes: A case study with nitrilases

Jennifer L. Seffernick[a,b,*], Sudip K. Samanta[c,1], Tai Man Louie[c,d], Lawrence P. Wackett[a,b], Mani Subramanian[c,d]

[a] Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, 140 Gortner Laboratory, 1479 Gortner Ave, St. Paul, MN 55108, USA
[b] BioTechnology Institute, University of Minnesota, 140 Gortner Laboratory, 1479 Gortner Ave, St. Paul, MN 55108, USA
[c] Department of Chemical & Biochemical Engineering, The University of Iowa, Oakdale Research Park, 2501 Crosspark Road, Suite C100, Coralville, IA 52241, USA
[d] Center for Biocatalysis & Bioprocessing, The University of Iowa, Oakdale Research Park, 2501 Crosspark Road, Suite C100, Coralville, IA 52241, USA

ARTICLE INFO

Article history:
Received 23 August 2008
Received in revised form 8 June 2009
Accepted 9 June 2009

Keywords:
Nitrilase
Genome mining
Mandelonitrile

ABSTRACT

Mining sequence data is increasingly important for biocatalysis research. However, when relying on sequence data alone, prediction of the reaction catalyzed by a specific protein sequence is often elusive, and substrate specificity is far from trivial. The present study demonstrated an approach of combining sequence data and structures from distant homologs to target identification of new nitrilases that specifically utilize hindered nitrile substrates like mandelonitrile. A total of 212 non-identical target nitrilases were identified from GenBank. Evolutionary trace and sequence clustering methods were used combinatorily to identify a set of nitrilases with presumably distinct substrate specificities. Selected encoding genes were cloned into *Escherichia coli*. Recombinant *E. coli* expressing NitA (gi91784632) from *Burkholderia xenovorans* LB400 was capable of growth on glutaronitrile or adiponitrile as the sole nitrogen source. Purified NitA exhibited highest activity with mandelonitrile, showing a catalytic efficiency (k_{cat}/K_m) of 3.6×10^4 M^{-1} s^{-1}. A second nitrilase predicted from our studies from *Bradyrhizobium zaponicum* USDA 110 (gi27381513) was likewise shown to prefer mandelonitrile [Zhu, D., Mukherjee, C., Biehl, E.R., Hua, L., 2007. Discovery of a mandelonitrile hydrolase from *Bradyrhizobium japonicum* USDA110 by rational genome mining. J. Biotechnol. 129 (4), 645–650]. Thus, predictions from sequence analysis and distant superfamily structures yielded enzyme activities with high selectivity for mandelonitrile. These data suggest that similar data mining techniques can be used to identify other substrate-specific enzymes from published, unannotated sequences.

1. Introduction

Digital data is in abundance in this post-genomic era. Gene and protein sequences abound while functional elucidations lag far behind. As of February 2008, GenBank holds over 82 million sequence records containing over 85 billion bases that span more than 260,000 named organisms. Of the 570 complete genomes in the database, around 200 were deposited in the last year alone, indicating the rapid influx of new genomic information (Benson et al., 2008). Due in part to the vast biochemical diversity unique to each organism, only approximately 50% of the genes in sequenced organisms have been annotated with a function (Zhou and Miller, 2002; White, 2006). In the case of sequenced eukaryotic organisms like *Aspergillus nidulans*, the extent of known and well-characterized protein annotation is less than 10% (David et al., 2008). In most cases, the base level annotation of gene function results from annotation copied from the closest homolog which itself might be quite distant from any sequence with experimentally determined function. Even when a general functional category is assigned, an indication of substrate (natural and non-natural) specificity of a specific enzyme is most often completely lacking (White, 2006).

In addition to sequence databases, a plethora of informatics software and associated databases have become available, like Pfam (http://pfam.sanger.ac.uk/; Finn et al., 2006) and BLOCKS (Henikoff et al., 1999) that can analyze sequence data for known patterns and motifs to assist in sequence and functional information association. However, even when there is enough information to categorize sequences into superfamilies or groups of related proteins, predicting substrate specificity is not trivial. As a result, automated genome annotation has become an exercise of general assignment of protein relatedness rather than exact functional assignments. It

* Corresponding author at: Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, 140 Gortner Laboratory, 1479 Gortner Ave, St. Paul, MN 55108, USA. Tel.: +1 612 624 4278; fax: +1 612 625 5780.
E-mail address: seffe001@umn.edu (J.L. Seffernick).
[1] Current address: Metabolix, 21 Erie Street, Cambridge, MA 02139, USA.

0168-1656/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2009.06.004

then remains to individual researchers to assess the validity of any assignments.

Attempts to obtain family assignments based on digital information alone is a timely topic and has received a great deal of attention (Mulder et al., 2002; Horan et al., 2005; Finn et al., 2006; Wilson et al., 2007). An important caveat to family assignments, which is just now emerging as a new challenge, is deciphering the substrate specificity of proteins within a given family. This level of analysis will be essential for using genomic data as a platform to select biocatalysts that produce specific, industrially relevant compounds. The broader thrust of this study was therefore to identify specific enzymes from a mixed superfamily and to further select those about to transform specific substrates.

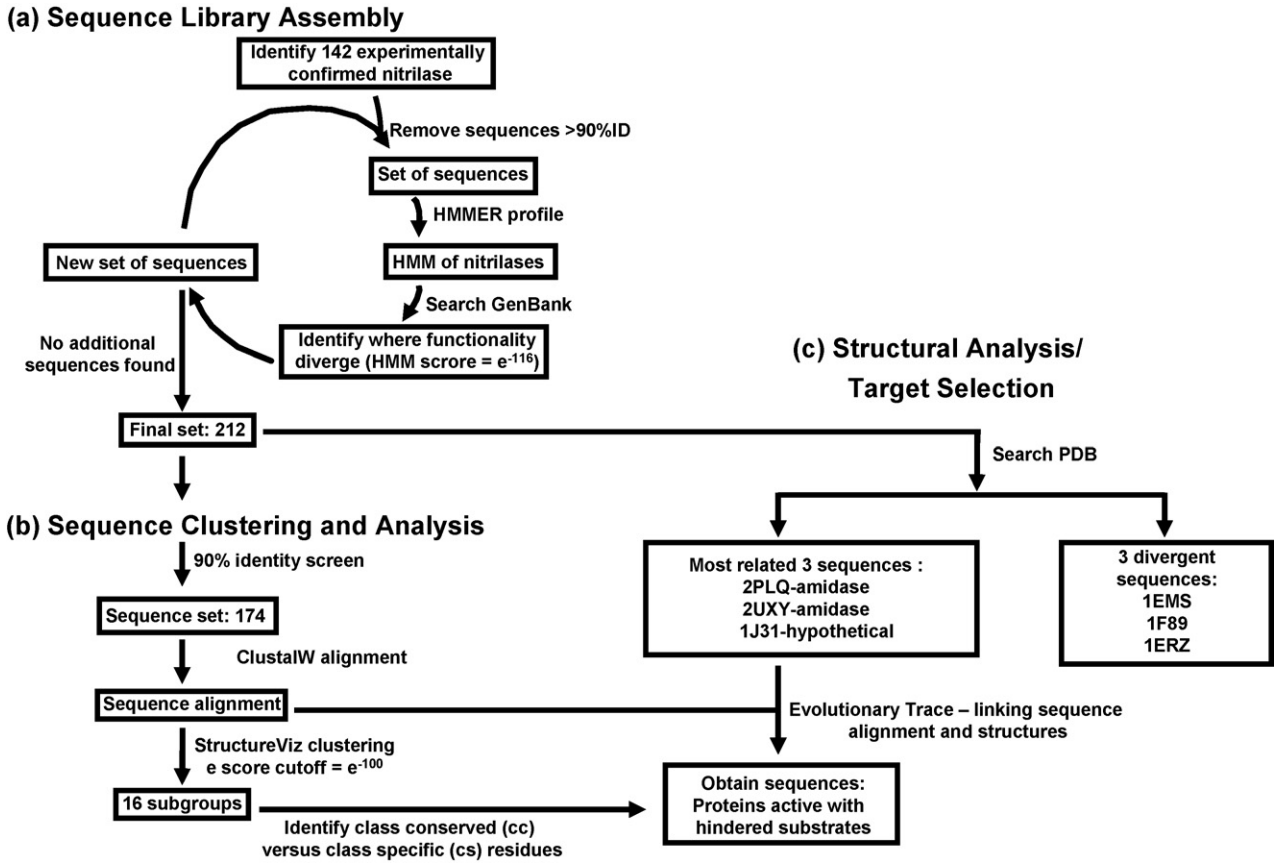
In this study, we chose to search for new nitrilases that would exhibit high specificity towards hindered, bulky substrates such as mandelonitrile. Nitrilases have proven useful in the industrial synthesis of chiral acids and hydroxyacids (Banerjee et al., 2002, 2006). Also, known nitrilases are moderately abundant, indicating that a large, but manageable, data set could be constructed. A typical soil prokaryotic genome of an organism like *Pseudomonas fluorescens* was annotated to contain at least 3 nitrilases (Conboy et al., 2003). Thus, it can be estimated that hundreds of nitrilase gene sequences may exist in public databases. Almost all of those genes encode enzymes with undetermined substrate specificities, kinetic properties, and enantio-selectivity. To our knowledge, no sequence-based predictors of nitrilase function currently exists.

Nitrilases are known to be members of the CN hydrolase superfamily (also known as the nitrilase superfamily) of proteins. This superfamily is based on a structural scaffold known as the α–β–β–α sandwich that houses a conserved Glu-Lys-Cys catalytic triad. The superfamily catalyzes nitrilase, amidase, carbamylase, and N-acyltransferase reactions (Pace and Brenner, 2001). The best sequence predictor of nitrilase functionality currently available was found to be the hidden Markov models (HMMs) from Pfam. These models identify nitrilases and their close relatives, the cyanide hydratases and cyanide dihydratases, but identification of the dividing line between these functionalities has not been realized. Furthermore, the ability to predict the substrate specificity of nitrilases is completely absent. Recently, an annotated alkylnitrile nitrilase of *P. fluorescens* was cloned and found to be a novel bifunctional enzyme that catalyzed both nitrilase and nitrile hydratase reactions with a specificity favoring aromatic nitrile compounds like hydroxycinnamonitrile rather than the annotated alkylnitriles (Conboy et al., 2003).

In this study, we were able to identify a set of nitrilases that were subsequently assayed with specific types of substrates. Detailed substrate specificity has only been determined with a handful of nitrilases. Of these enzymes, none are known to exhibit a high level of specificity towards a hindered or bulky substrate such as mandelonitrile. Therefore, the ability to predict such a biocatalyst from genome sequences, with a reasonable level of success, would substantially improve our understanding of nitrilases and obtain a biocatalyst that has a desired catalytic capabilities. This predictive methodology could also be applied towards other enzyme classes.

2. Materials and methods

Scheme 1 depicts the general methodology used to identify novel nitirilase sequences and analyze those sequences for specific traits or features such as substrate specificity. This methodology consists of three main sections described in more detail below: sequence library assembly, sequence clustering and analysis, and



Scheme 1. Methodology for target sequence prediction of nitrilases that utilize large hindered substrates.