Functional and comparative analysis of novel dehydrogenase genes in

Carboxydothermus hydrogenoformans Z-2901

C P Rajadurai, S Ashok, Nupoor Chowdhary and G Ramesh Kumar*
Bioinformatics Laboratory, AU-KBC Research Centre, Anna University, Chennai 600 044, India

Received 17 January 2012; revised 1 September 2012; accepted 20 October 2012

Carboxydothermus hydrogenoformans Z-2901 is a carboxydotrophic hydrogenogen Gram-positive bacterium species, which utilizes carbon monoxide (CO) as a carbon source and produces hydrogen. During CO-dependent metabolism, dehydrogenase types of enzymes play a significant role in biological hydrogen production. In this organism, CO dehydrogenase (CODH) and other dehydrogenase genes are effectively involved in fixation and degradation of carbon compounds. Several hypothetical open reading frames (ORFs) coding for dehydrogenase genes have to be functionally analyzed in an attempt to understand their biochemical behaviour. In the present paper, authors elaborately describe the functional identification of novel dehydrogenase genes by studying their structural features, such as, motif, conserved signature pattern, folds and evolutionary relationships through functional and comparative studies. Findings from these analyses have strongly suggested the functions of two hypothetical proteins YP_360204 and YP_361460 as succinate dehydrogenase and FAD-dependent dehydrogenase, respectively.

Keywords: Annotation, dehydrogenase, homology, hydrogen, motif, phylogeny

Introduction

Hydrogen is considered as fuel of the future due to its high conversion efficiency, recyclability and non-polluting nature. Besides its application as a fuel, hydrogen can also be used as a potential electron donor for various reactions in biotechnological and chemical industrial processes. It is generally produced from the reformation of methane, combustion of coal and biomass, and most prominently from the metabolism of microorganisms. Biological hydrogen production methods are more convenient, since they are environmentally friendly processes and less energy intensive when compared to the thermochemical and electrochemical processes. Hydrogenase/dehydrogenase types of enzymes play a significant role in bio-hydrogen production. The contribution of these enzymes for hydrogen production in microbes and other organisms has been studied through various in vitro experiments. Several types of hydrogenase enzymes are present in many microorganisms including some algae, trichomonads, anaerobic ciliates, and chytrid fungi. It is noted that bidirectional hydrogenase in microorganisms, such as, cyanobacteria, plays a vital role in hydrogen production as it has Hox proteins containing glycine rich NAD motif (GxGxxGxxxG) and FAD motif (GxGxxxxGx10GxxG). Carbon monoxide-dependent energy metabolism is thought to be strictly depends on the presence of carbon monoxide dehydrogenase or CO dehydrogenase (CODH) enzyme. Other dehydrogenase enzymes, such as, FAD-dependent dehydrogenase, NAD-dependent dehydrogenase, formate-dependent dehydrogenase are also involved in electron transfer and hydrogen production pathways of many hydrogenogens. Several microbial species especially hydrogenogens have a significant research focus because of their ability of biological production of hydrogen gas. Carboxydothermus hydrogenoformans Z-2901 is one among them and it was first isolated from hot spring in Kunashir Island, Russia. This organism generally utilizes CO as a carbon source and produces CO₂ and hydrogen molecule as end products. The peculiar feature of this organism is that it is having five different CODH complexes unlike other carboxydotrophic hydrogenogens. These complexes

*Author for correspondence
Tel: +91-44-22234885; Fax: +91-44-22231034
E-mail: gramesh@au-kbc.org
act as catalytic enzymes for the interconversion of CO to CO$_2$. Hence, *C. hydrogenoformans* has been considered as a model organism for this CO-dependent metabolism and hydrogen production. With the knowledge of annotated genome of *C. hydrogenoformans*, as it has been annotated in the year 2005, various studies on biological hydrogen production have gradually been performed. However, the genome annotation on *C. hydrogenoformans* has resulted in characterization of only 56% gene functions and the rest of (both hypothetical and putative) gene functions are unassigned, whose sequence domains are conserved through various species. Therefore, we hypothesized that several unknown genes function as dehydrogenases are missing in the previous genome information and they can be unveiled through complete functional analysis. In overall, the objective of the present paper work was to model, evaluate and complete functional analysis. In overall, the objective of the present paper work was to model, evaluate and complete functional analysis.

**Materials and Methods**

**Sequence Similarity Search**

To initiate functional predictions of unknown ORFs, the whole genome has to be retrieved from genome databases. The complete genome of *C. hydrogenoformans* Z-2901 in protein fasta format NC_007503 was downloaded from the RefSeq FTP file repository (last updated in 9th February 2007) available in NCBI-Microbial Genome Database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Carboxydotherrus_hydrogenoformans_Z-2901/NC_007503.faa). The downloaded files were manually split into known and unknown protein sequences. Further, the unknown protein sequences were classified into three categories: i) hypothetical, ii) conserved hypothetical and iii) putative. The unknown protein sequences were used as a query for the similarity search.

The functional context of hypothetical ORFs was identified through similarity searches against BLAST, COG, CDD, InterProScan, ProDom, BLOKS and STRING databases. Each of the *C. hydrogenoformans* protein coding sequences (CDSs) previously predicted and annotated was manually re-analyzed based on the diverse approaches, such as, similarity based search approach, neighborhood search, orthologous search, domain based search and proteins family based search. The product descriptions were further manually refined based on the confidence level, *i.e.*, percentage of occurrence of most common positive results from different tools used. In BLAST, the searches were performed using default parameter settings and run against non-redundant database (nr). The results obtained from BLAST with a maximum alignment score and threshold E-value of $1 \times 10^{-6}$ up to $1 \times 10^{-52}$ were chosen as best hits. Next, ortholog sequence search was carried out using COG and domain based search was usually done with ProDom (cut off score 80) and CDD. The protein-pattern search was achieved by using InterProScan. Further, BLOKS and STRING tools were performed using individual global/local search options with expected cutoff E-value of 1.0. The NCBI RefSeq accession numbers were later used as a reference identity for the whole protein sequence.

**Motif and Pattern Search**

Functional prediction based on sequence similarity alone cannot be helpful in assigning the accurate functions of the gene sequence. Henceforth, motif or pattern analysis was implemented to find out significant functions of unknown protein sequences. Clustal X software was used to generate multiple sequence alignment within each predicted protein families. The pattern profiles obtained as a result of multiple sequence alignment were further preceded with motif analysis using ScanProsite tool.

**Fold Recognition**

Fold recognition is another method of predicting the functions of unknown gene by comparing its sequence with structure by a process called threading or folding match. During evolution, the folding patterns of proteins are often preserved and, hence, exploring the secondary structure information of hypothetical proteins through fold recognition may give a clue for studying their biochemical or biophysical functions. Fold recognition for all the predicted dehydrogenase ORFs obtained from sequence similarity search and motif search was done by using PHYRE Protein Fold Recognition Server. It uses a library of known protein structures obtained from the structure databases. The submitted query sequence was searched against a non-redundant
sequence database and a profile was constructed. Further, the close and remote sequence homologs for the submitted query sequence were obtained after iterative PSI-BLAST and thereby the secondary structure alpha-helix (H), beta-strand (E) and coil (C) were predicted based on the best sequence alignment score. The constructed profile and secondary structures were scanned against the fold library using profile-profile algorithm and the best alignments were ranked. From the top ten highest scoring alignments, the first rank model structure (PDB) was then used as a template to generate a 3-D model of the query. The 3-D structure for the predicted dehydrogenase ORFs was generated by using Modeller 9v7 Software. For homology based structure modeling of YP_360204.1 ORF, succinate dehydrogenase protein of *Escherichia coli* (PDB ID: 1NEK) was considered as a template and similarly the ORF YP_361460.1 had FAD utilizing heterotetrameric sarcosine oxidase (PDB ID: 2GAH) protein as a template. This structural modeling work helps to exploit the knowledge of functional regions of the predicted dehydrogenase proteins.

**Whole Genome Pair-wise Comparison**

To obtain an integrated mapping view of homologous regions between *C. hydrogenoformans* and other reference hydrogenogen species, ARTEMIS software was used. Whole genome of all the comparative hydrogenogen species were downloaded as an EMBL file format from EBI database and visualized in order to find the common protein coding sequence distribution pattern across the species. The predicted dehydrogenase genes from similarity search, motif and fold recognition were mapped by marking their locus tag or RefSeq number.

**Phylogenetic Analysis**

Multiple sequence alignment for the newly predicted dehydrogenase proteins was done by using Clustal X software. Protein sequences of hydrogenogens species (to be compared with predicted proteins) were obtained in FASTA format from SWISS-PROT database and visualized in order to find the common protein coding sequence distribution pattern across the species. The predicted dehydrogenase genes from similarity search, motif and fold recognition were mapped by marking their locus tag or RefSeq number.

**Results and Discussion**

**Identification of Potential Dehydrogenase Proteins**

The ultimate aim of sequence similarity based search was to identify sequence features indicative of putative dehydrogenase homologs. At this stage, only five sequences were considered as strong positive dehydrogenase gene. Three ORF sequences YP_358956.1, YP_358959.1 and YP_359691.1 could be considered as positive candidates, although any CODH signature pattern was unidentifiable. Based on the functional context, two ORFs were potentially assigned as succinate dehydrogenase (YP_360204) and FAD-dependent oxidoreductase (YP_361460). The above mentioned five gene functions were assumed to be directly or indirectly involved in hydrogen production as referred from the literature. The average GC content of the five predicted ORFs was 40%, of which two ORFs, YP_360204 and YP_361460 having 100% confidence level were taken for further annotation studies (Table 1).

**Conserved Motif or Pattern Analysis of Predicted ORFs**

The multiple sequence alignment of the predicted dehydrogenase ORFs highlighted the existence of their characteristic signature motif. The conserved motifs for the two predicted dehydrogenase ORFs YP_360204.1 and YP_361460.1 were identified using Clustal X software (Figs 1a & b). Those hypothetical sequences were then compared against their orthologous species (each ORF as a query) obtained from non-redundant blast search.

**Table 1—Description of predicted dehydrogenase functions of hypothetical sequence with confidence level**

<table>
<thead>
<tr>
<th>RefSeq no.</th>
<th>Predicted function</th>
<th>Sequence length (bp)</th>
<th>Confidence level</th>
<th>G+C content</th>
</tr>
</thead>
<tbody>
<tr>
<td>YP_360204</td>
<td>Succinate dehydrogenase</td>
<td>420</td>
<td>100%</td>
<td>38%</td>
</tr>
<tr>
<td>YP_361460</td>
<td>FAD-dependent oxidoreductase</td>
<td>1386</td>
<td>100%</td>
<td>50%</td>
</tr>
</tbody>
</table>
RAJADURAI et al: ANALYSIS OF NOVEL DEHYDROGENASE GENES

For YP_360204.1, the conserved signature pattern was L-Y-R-C-H that indicates the presence of LYR motif and suggested as succinate dehydrogenase, an enzyme involved in Fe-S metabolism (Fig. 2a). Such motifs are generally partially buried regions, structurally clustered along with cytochrome P450, a protein involved in electron transport path. It is well known from the literature that the succinate dehydrogenase plays a crucial role in energy coupling during citrate cycle in several prokaryotes. The presence of G-[AG]-G-P-A-G signature pattern in the ORF, YP_361460.1 revealed that it has GXGXXGXXXL motif specific to FAD-dependent pyridine nucleotide disulphide oxidoreductase function (Fig. 2b). NAD/FAD-dependent oxidoreductase types of enzymes in both aerobic and anaerobic organisms play a significant role in bio-hydrogen production. The contribution of these enzymes for hydrogen production in microbes and other organisms are well studied through various in vitro experiments.

**Fold Recognition and Superimposition of Predicted Dehydrogenase ORFs with Their Templates**

The present fold recognition studies on predicted dehydrogenase ORFs explained their three dimensional fold arrangements with respect to their function. The predicted succinate dehydrogenase ORF YP_360204.1 has Heme-binding four-helical bundle fold. It was found that the SCOP template i.d. 1NEK has the structural similarity of 95% with YP_360204.1, whose predicted backbone RMSD (root mean square deviation) value was obtained as 2.7 Å (optimum RMSD value ranges from 2-4 Å). This transmembrane fold region of succinate dehydrogenase (1NEK) was already reported in E. coli and they were surrounded by anti-parallel left-handed helical pair with beta-sheets in the interior part. The predicted fold contains functionally active
residues ILE 8, LYS 9, GLY 10, ARG 48, LEU 49 and ARG 108, found in the N-terminal regions of 107 to 238 amino acids, and was specific to succinate dehydrogenase enzyme. Similarly, oxidoreductase domain fold region was found to be predicted for YP_361460.1. The SCOP template was 2GAH, which codes for FAD-utilizing heterotetrameric sarcosine oxidase beta-subunit. The fold contains functionally active residues LYS 130, ARG 131, ALA 132, VAL 168, THR 172, GLN 190 and LYS 216, found in the N-terminal region and the predicted RMSD value was 6.10 Å with the template.

Model Validation of Predicted Dehydrogenase ORFs

The two predicted dehydrogenase ORF models, YP_360204.1 and YP_361460.1 were further validated with Structural Analysis and Verification Server (SAVES) to check their structural consistency and reliability. To evaluate the backbone confirmation, Ramachandran plot was obtained from PROCHECK analysis. For YP_360204.1 ORF, it was found that 84.5% of amino acids fall in most favoured region, 14.5% in additionally allowed region, 0.8% in generously allowed region and 0.2% in disallowed region (Fig. 3a). Likewise, for YP_361460.1, it was shown that 86.2% of amino acids fall in most favoured region, 13.2% in additionally allowed region, 0.3% in generously allowed region and 0.3% in disallowed region (Fig. 3b). Ramachandran plot for glycine, pre-proline and proline for both the ORFs showed all these three types of amino acids under allowed region. The packing quality of the predicted dehydrogenase ORF models were found in normal range as calculated with the WHATIF program. Using VERIFY_3D and ERRAT program, it was shown that the overall ERRAT quality factor (residues had an average 3D-1D score >0.2) for YP_360204.1 and YP_361460.1 ORFs were 84.65 and 90.15%, respectively.

Functional Linkage through Evolutionary Studies

In this phylogenetic study, predicted dehydrogenase ORFs, YP_360204.1 and YP_361460.1 were searched using BLASTP to obtain known dehydrogenase genes of different carboxydotrophic species with an E value of <10^{-5} as an optimum hit. Phylogenetic analysis for the hypothetical ORF YP_360204.1 of C. hydrogenoformans was performed with the sequences of known succinate dehydrogenase enzyme subunit of seven different carboxydotrophs. Interestingly, the predicted succinate dehydrogenase ORF YP_360204.1 shared the very close paired-evolutionary distance of 2.635920 with S. cerevisiae species (Fig. 4a). This evolutionary distance also suggested that the predicted ORFs YP_360204.1 has a highly coupled functional linkage with succinate dehydrogenase genes. Likewise, YP_361460.1 shared the closest pairwise distance of 0.487080 with C. desulforudis audaxviator (Fig. 4b). The phylogenetic results, therefore explained, those two
predicted ORFs belong to dehydrogenase class as their gene sequences lie in close proximity with less variation with other reference species.

Phylogenetic tree was constructed for the three predicted dehydrogenase ORFs with their orthologous sequences to determine the evolutionary and functional relationship. In *C. hydrogenoformans*, FAD-dependent dehydrogenase/oxidoreductase involved in intermediate electron carrier that functions along with CooF gene. The operon with multi-subunit complex encoding CooS-IV, a CooF homolog, and NAD/FAD-dependent oxidoreductase (CHY0735–8) was suggesting that they reduce ubrerythrin from hydrogen peroxide to water\(^1\). The group of NAD/FAD-dependent oxidoreductase has diversities such as, NADH oxidase, NADH reductase, nitrite reductase etc, which were present in several CO-dependent organisms. Previously, *C. hydrogenoformans* was thought to be an autotroph that strictly depends on only CO. Later, substrate studies with various carbon sources, such as, succinate, revealed that it has heterotrophic capabilities\(^1\). Our phylogenetic results provided a clue for the existence of novel FAD-dependent dehydrogenase/oxidoreductase and succinate dehydrogenase that were missing in the genome table. The two newly predicted ORFs YP_361460.1 (succinate dehydrogenase) and YP_360204.1 (FAD dehydrogenase) shared very close genetic distance of 0.487080 with *S. cerevisiae* and 2.635920 with *C. desulforudis audaxviator*. The results strongly suggested that *C. hydrogenoformans* has heterotrophic nature. Functional analyses suggested the existence of CO dehydrogenase, FAD/NAD dehydrogenase and succinate dehydrogenase, and thus provided an evidence for multiple substrate dependent carbon metabolisms that perhaps facilitate the growth of this species by efficiently oxidizing CO and producing hydrogen.

Acknowledgement

The authors are thankful to the project students Ms Devika Parvathy, Ms Neena and Mr Ashish Daniel for their help in manuscript preparation.

References


16 Ragsdale S W & Wood H G, Acetate biosynthesis by acetogenic bacteria. Evidence that carbon monoxide dehydrogenase is the condensing enzyme that catalyzes the final steps of the synthesis, J Biol Chem, 260 (1985) 3970-3977.


26 Gabriel M H & Kristen L, Choosing BLAST options for better detection of orthologs as reciprocal best hits, Bioinformatics, 24 (2008) 319-324.


39 Mowat C G, Leys D, McLean K J, Rivers S L, Richmond A et al, Crystallization and preliminary crystallographic analysis of a novel cytochrome P450 from Mycobacterium


