J Biochem Tech (2009) 1(3):85-87
ISSN: 0974-2328

CMP Promoters Database: A systematic study on site-specific transcription factors in CMP genes

Meera A, Lalitha Rangarajan, Savithri Bhat*

Received: 17 March 2008 / Received in revised form: 5 February 2009, Accepted: 31 April 2009 Published online: 14 May 2009

Abstract

CMPP Database (Central Metabolic Pathway Promoter Database) consists of manually annotated regulatory sequences/motif of genes controlling Metabolic Pathways (http://cmpp.biodbs.info) such as Glycolysis' and Krebs' cycle. The CMPP software package written in Microsoft Visual Basics & MS Access was also developed for searching & updating the CMPP database. This database was further used to study the relationship between 10 transcription factors involved in the non-coding upstream region of CMP genes. There is an existence of common motifs between the genes encoding enzymes involved in glycolysis and Krebs cycle revealing an evolutionary relation between the genes of the two pathways. However, some of the common motif such as TATA uncommon motifs as NKX and AP2 making up the structural feature of the non-coding genes are absent from few genes.

Keywords: Non-coding sequence, Phylogeny, TCA, Glycolysis, TRANSFAC, Promoter, Database, Central Metabolic Pathway.

Abbreviations: CMP - Central Metabolic Pathway, TF - Transcription Factor, TFBS -Transcription Factor Binding Site, TCA - Tri Carboxylic Acid cycle, EMBL - European Molecular Biology Laboratory, RNA - RiboNucleic Acid, VB - Visual Basics

Meera A

Department of Electronics and Communication, B.M.S College of Engineering, Bangalore - 560 019, India

Lalitha Rangarajan

Department of Computer Science, University of Mysore, Mysore, India

Savithri Bhat*

Department of Biotechnology, B.M.S College of Engineering, Bangalore -560 019, India

* Tel: 0091-80-26622130; Fax: 0091-80-26614357
Email: savithri.bhat@gmail.com

Introduction

The non-coding upstream sequences including promoter are the important elements for controlling gene expression in both Eukaryotes and Prokaryotes. The information for control of initiation for the synthesis of RNA by the RNA polymerase lies in the promoter region that extends between 200-2000 nucleotides upstream of the Transcription Start Site (TSS) of a gene. The transcription factors (TFs) interact with sequence specific elements or motifs, which are 5-12 nucleotides in length. The motifs appear to be arranged in specific configuration that confers on each gene an individualized spatial or temporal transcription program (Wray et al. 2003). The non-coding upstream regions are not well-conserved but it is assumed that genes exhibiting similar expression patterns share similar configuration of TF in their promoter (Blanco et al. 2006). The TFBS associated to the same TF are known to tolerate sequence substitution without losing functionality, thus promoter regions of genes with similar expression pattern may not show sequence similarity even though they may be regulated by similar configuration of TFs. Although te recent progress in this regard, due to the techniques based on Phylogenetic finger printing (Wasserman and Sandelin 2004) has been made, the lack of nucleotide sequence conservation between functionally related promoter regions may partially explain the limited success of currently available computational methods for promoter characterization (Fickett and Hatzigeorgiou 1997; Tompa et al. 2005). In the approach described here, we developed a database of TF & TFBS of genes encoding enzymes of CMP such as Glycolysis and TCA. This is followed by the identification of the relationship between these TFs of both the pathways.

Method

The promoter sequences for each gene encoding enzymes of Glycolysis and TCA (CMP) were retrieved from EMBL database. The gene coding for individual enzyme were used as input and the search was refined by using 'promoter' as key word. They were further used as an input to extract TFBS of 10 important TFs via TRANSFAC (Transcription Factor search tool) The database was further exploited to study and interpret the relationship existing between the 10 important transcription factors and genes coding for enzymes involved in both the Central Metabolic Pathways.

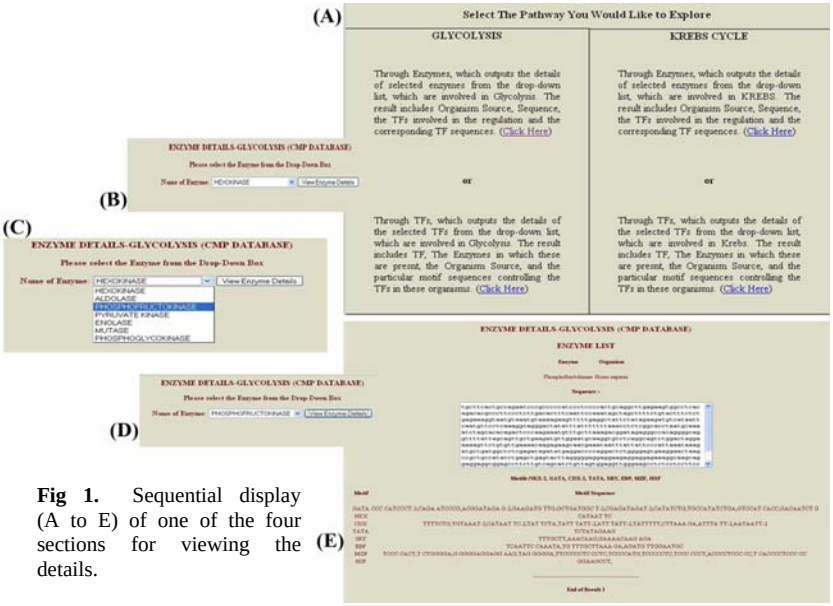

Fig 1. Sequential display (A to E) of one of the four sections for viewing the details.

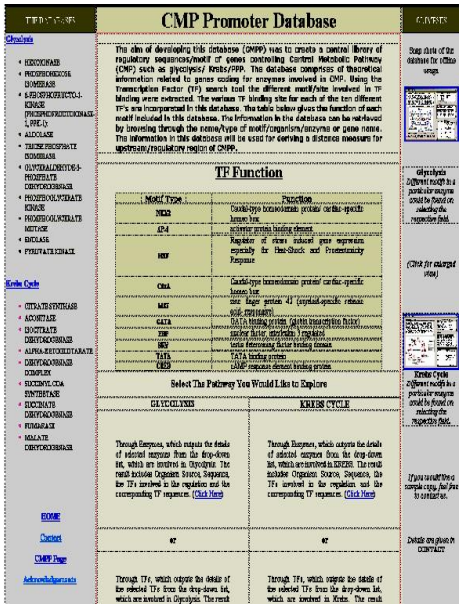

Fig 2. The interface of the CMPP Database showing the information covered in the entire database.

Results

The developed CMPP Software contains the enzyme types , their sources, sequences, the TFs and the corresponding motif sequence on which TFs bind. The dataset could be browsed and extracted through "Enzymes" section or corresponding motifs associated with each (Fig 1 and Fig 2). The software contains information of 16 different sources involved in glycolysis and 8 different sources involved in Krebs cycle. Theoretical information about enzymes was also reserved inside the software.

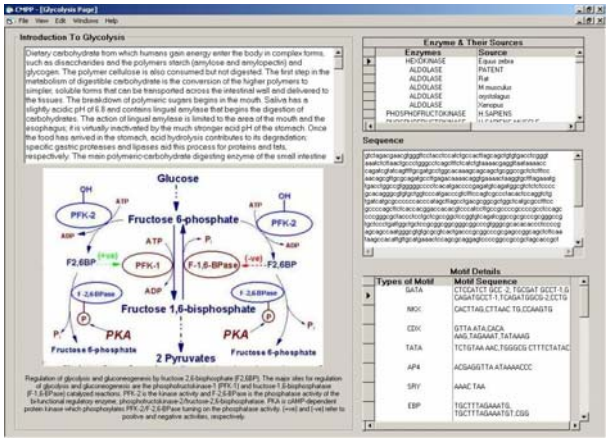

Fig 3. Description of glycolysis, source, motifs and their sequence

The dataset was made available online (http://cmpp.biodbs.info) which has a flexibility to run on various operating systems. The online version could be browsed through either "Enzymes" or "Motifs" present in Glycolysis' or Krebs' pathway as shown in the figure (Fig 3 & Fig 4). Updating the online CMPP database could be done by e-mailing to the webmaster (info@biodbs.info).

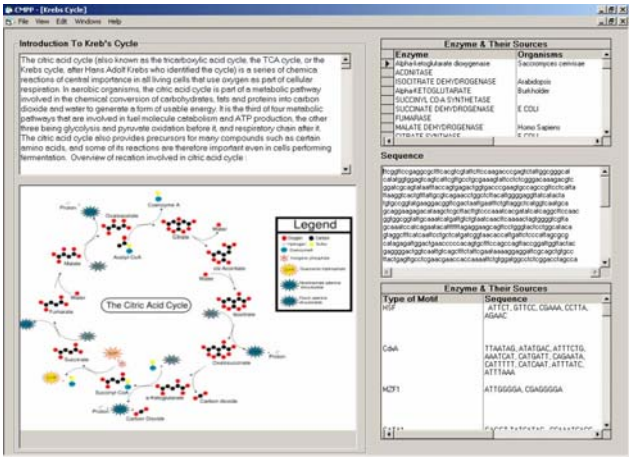

Fig 4. Description of Kreb cycle, source, motifs and their sequence

Discussion

In the present paper a comprehensive analysis of TFs involved in Central metabolic pathway genes has been carried out to draw relationship between the genes in order to trace their origin. The details of the genes encoding enzymes of both glycolysis and TCA and their reactions are depicted in the database. A bioinformatics approach to study the upstream regulatory sequence of Glycolysis and Gluconeogenesis using software tools are reported (Van Helden et al. 2000; Tompa et al. 2005). Under micro array analysis global expression of Krebs' cycle genes were studied and related to various TF sequences present upstream of Aconitase , Isocitrate Dehydrogenase , α-Ketoglutarate Dehydrogenase, Succinate Dehydrogenase, Fumarase , Malate Dehydrogenase and Citrate Synthase (Mark T et.al. 2003**).** Amongst the Transcription factor binding sites for TF screened, motif for HSF was found to be present universally as a common TF in all the enzymes of both the

pathways which reveals the highly conserved nature of HSF or an important TF required for transcription of the CMP genes. The upstream sequence of Pyruvate Kinase & Endolase enzymes of Glycolysis does not have any motif studied except HSF which reveal the evolutionary relationship between these genes, an assumption that they share common ancestry. The upstream sequence of Pyruvate Kinase, Mutase, & Endolase of Glycolysis shows similarity in the types of motifs in more than 50% of the sequences studied.

The non-coding sequences of genes encoding enzymes such as Aldolase of Glycolysis & Alpha-Ketoglutarate Dehydrogenase and Deoxygenase and Succinate Dehydrogenase of TCA were shown to have all the motifs listed revealing the common origin of these enzymes. The TFs such as AP2 & NKX are founded to be the uncommon motifs or the rare motifs for CMP genes. One of the important motifs, TATA was seen to be absent from the genes such as Pyruvate Kinase, Mutase of Glycolysis, Citrate Synthase & Succinate Dehydrogenase of TCA. It is known that TATA may not be always a RNA polymerase binding site in some of the eukaryotic genes (Itay Tiros 2007; Fickett and Hatzigeorgiou 1997). As shown, there was a variation in the sequences making TFBS for almost all the upstream sequences of the genes encoding enzymes of both the pathway. This further proves the non-conserved nature of the motifs present in the upstream region of genes on which a transcription factor binds.

The present database is been used to derive the Transcription factors present upstream of the start site of a gene. The distances between the genes coding for the enzymes of glycolysis and TCA were calculated dependent on presence or absence of motifs (Meera et al. 2009) and TF maps were generated.

This information present in the database could be further used as template to study related metabolic pathways such as lipid or nucleotide or amino acids or carbohydrate, an essential approach for understanding the science underlying the disease development and progression. At present KEGG metabolic pathway is the database which comprises of the information related to the genes coding for enzyme involved in various pathways. The details of TFBs present upstream of the gene are not available in KEGG. The development of this database is intended towards an approach for revealing the importance of upstream non-coding regions of genes in deriving phylogenetic relation. This could also be useful in studying the mutations in these TFs which could lead to diseases.

Further, it could provide a pave to enhance metabolic engineering research. We shall continue to include more scientific information in the database to make it topic specific and application oriented database.

Acknowledgement

We greatly acknowledge Mr. Aby Abraham, Biological Science, Greenville, South Carolina Area, Clemson University, USA, for computational advise on creation of this database tool and Mr. Arun Chandra Shekar, Department of Microbiology, CFTRI, Mysore, India, for his valuable scientific suggestions.

References

Blanco E, Messeguer X, Smith TF, Guigo R (2006) Transcription factor map alignment of promoter regions. PloS Computa Biol. doi:10.1371/journal.pcbi.0020049

Fickett JW, Hatzigeorgiou (1997) Eukaryotic promoter recognition. Genome Res 7:861–878

Itay Tiros, Judith Berman, Naama Barkai (2007) The pattern and evolution of yeast promoter bendability. Trends in Genetics 23:318-321

Kageyama R, Merlina GT, Pastan I (1989) Nuclear factor ETF specifically stimulates transcription from promoters without a TATA box. J Biol Chem 264:15508-15514

Meera A, Lalitha Rangarajan, Savithri Bhat (2009) Computational Approach towards Finding Evolutionary Distance and Gene Order Using Promoter Sequences of Central Metabolic Pathway. Interdiscip Sci Comput Life Sci 1:1–5

Mark T McCammon, Charles B, Epstein et al (2003) Global Transcription analysis of TCA mutants reveals an alternating pattern of gene expression and effects on hypoxia and oxidative genes. Mol Biol Cell 14:958-972

Tompa M, Li N, Bailey TL et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol. 23:137–144

Van Helden J, del Olmo, Perez OJE (2000) Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic acid Research 28(4):1000-1010

Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5:276–286

Wray GA, Hahn MW et al (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20:1377–1419