Trends in Bioinformatics

ISSN 1994-7941

science alert

ANSInet
an open access publisher
http://ansinet.com

Trends in Bioinformatics 7 (1): 7-12, 2014
ISSN 1994-7941 / DOI: 10.3923/tb.2014.7.12

ECO-MP: *E coli*-Metabolic Pathway-development of Genome-scale Metabolic Pathway Database for *Escherichia coli*

Gopal Ramesh Kumar, Thankaswamy Kosalai Subazini and Ashok Selvaraj
Bioinformatics Lab., AU-KBC Research Centre, Madras Institute of Technology Campus of Anna University, Chromepet, Chennai, Tamil Nadu, 600044, India

Corresponding Author: Gopal Ramesh Kumar, Bioinformatics Lab., AU-KBC Research Centre, Madras Institute of Technology Campus, Anna University, Chennai, 600 044, Tamil Nadu, India Tel: +91 44 2223 2711 Ext: 133 Fax: +91 44 2223 2711

ABSTRACT

Reconstruction of genome-scale metabolic pathway is an essential task to understand cellular processes in *Escherichia coli* K-12 species. ECO-MP, an *E. coli* metabolic pathway database is developed for the investigation of biosynthetic pathways of metabolic network. ECO-MP database is created by BLASTing protein sequences of *E. coli* against KEGG database and the sequences are categorized as known pathways, unknown and constructed pathways. Further, ECO-MP directs the gene or proteins to KEGG pathway in which they involved or otherwise it is linked to generate subnetworks using chemical reactions obtained from KEGG by network expansion method. Totally 2,560 known and 1,730 unknown proteins have been annotated to facilitate metabolic pathway reconstruction in *E. coli*. The subnetworks obtained from ECO-MP can be employed to fill the metabolic holes or to create an alternative pathway by means of network expansion method and it can be accessed at http://ecomp.bioinfo.au-kbc.org.in/.

Key words: *E. coli*, pathway reconstruction, subnetworks

INTRODUCTION

Escherichia coli, a gram negative, genetically engineerable organism can be used as a model organism and being industrially important model organism, understanding metabolic process of *E. coli* is inevitable. Deep comprehensions of metabolic pathway provide the chemistry behind the production of industrial products. The incomplete annotation provides lack in complete perception on information of molecular level interactions which on further analysis can be used to develop drugs by disrupting the reaction networks and also for determining the novel drug targets by analyzing the pathway (Yeh *et al*., 2004). Moreover, to improve the productivity of industrial products like hydrogen and others from *E. coli*, the detailed understanding of metabolic process involved in the production is crucial. Hence, forth, metabolic pathway reconstruction has to be carried out to predict the overall metabolic composition for *E. coli* and those reconstructed pathways can be integrated with functional genomics data. Metabolic pathway reconstruction also provides information about the reactions which has no corresponding enzymes enzymes (Keseler *et al*., 2005). There are many pathway databases such as Cyclone, MedicCyc and PathCase (Le Fevre *et al*., 2007; Urbanczyk-Wochniak and Sumner, 2007; Elliott *et al*., 2008) that has been developed for different organism to integrate functional genomics and metabolic pathway information to obtain more systems biology view of results. The databases such as GenProtEC (Serres *et al*., 2004) and EcoGene (Rudd, 2000) have already been developed with well annotated

gene functions. Although, several databases have gene and metabolic pathway information, the pathway, metabolite and enzyme data of complete gene sets are unrevealed (Chou *et al*., 2009). ECO-MP, a reconstructed metabolic pathway database that can assist the purpose of gaining more knowledge on cellular processes of *E. coli* by providing a comprehensive reaction and network data on unknown genes or hypothetical and conserved hypothetical proteins of *E. coli*. The protein functions of unknown genes such as hypothetical and conserved hypothetical proteins must be predicted as they might play a significant role in cellular physiology of microorganisms. Hypothetical proteins are proteins of unknown function with no homology or experimental evidence and conserved hypothetical proteins are unknown proteins with phylogenetic distribution and homology (Rajadurai *et al*., 2011). It encompasses cellular information and biological pathways that characterize the roles of genomic entities in various cellular mechanisms of *E. coli*. ECO-MP database is constructed by exploiting the records from KEGG (Kanehisa *et al*., 2008) which serves as a repository of pathways of around 2770 organisms with 192 eukaryotes and 2578 prokaryotes. Subnetworks are flow of one or more reactions in which the end product of a subnetwork is a reactant of another subnetwork. The pathways are sequence of chemical reactions and those series of chemical reactions are interlinked together for a continuous flow of biochemical process (Schlitt and Brazma, 2007). This flow of process is controlled by certain enzymes or reactants such as metabolite, chemicals which are involved in the accomplishment of inhibiting or reversing the reaction flow. The reactions can be of type single substrate; single product, single substrate; multi-products, multi-substrate; multi-products and multi-substrate; single-product. By altering the amount of reactants in the pathway, formation of certain products can be inhibited or augmented. To shed light on entire biological roles in *E. coli*, the metabolic pathway should be reconstructed with complete information. The metabolic reconstruction using network expansion method involves collecting the stoichiometry information for both known and unknown genes. The stoichiometry information for the genes is collected from KEGG database. Each ORF were searched against KEGG database and enzyme commission numbers were obtained. ORFs which have complete EC No. information are used for subnetwork construction. Graphviz use hierarchical methods were two or more reactions are interconnected using edges generating subnetworks. In this, based on the given reactions and reversibility networks are created by linking the product of one reaction which acts as reactant of another reaction. When there are gaps the network stops reaching dead end. This database is designed to furnish reconstructed pathways from the inference of the biochemical reactions by association of linear reaction and non-linear reaction sets in *E. coli* to form a subnetwork (Schuster *et al*., 2002). Since, the reconstructed network is by joining the reactions together, this approach is known as Bottom-up approach (Natalie *et al*., 2006).

MATERIALS AND METHODS

ECO-MP, an *E. coli* pathway database was designed to offer knowledge on reconstructed pathway by combining two or more chemical compounds by analyzing its activity relationship with each other and also known pathways in KEGG. Metabolic pathway can be considered as a graph, with each nodes being the metabolite, chemicals and enzymes involved in biochemical reactions. The reconstruction is made by network expansion method (Handorf and Ebenhoh, 2007), in which the expansion is made for reactions with product of a reaction being substrate of another reaction. Thus the interdependencies of the reactions are used to reconstruct pathways for *E. coli* by generating single-step pathways which on further analysis can be used to fill metabolic holes or expansion with seed metabolites to create a complete metabolic pathway. Reconstruction of

metabolic networks from biochemical reactions removes blockades in knowledge of manipulating *E. coli* biochemical pathway to enhance industrial products produced by the bacteria. The protein sequences of *E. coli* has been retrieved from EcoCyc (Karp *et al.*, 2002) and blasted against KEGG protein database. The KEGG blast results are categorized based on availability of KEGG-pathway and KEGG-reaction. A set of chemical compounds are listed from the collection of whole reactions. Each reactive compounds and products are analyzed carefully to find out if the chemical compounds occurred in more than one time as product or substrate. In such repetitive process, the particular reactions are linked to form subnetwork that can be further used to fill holes in known pathway in KEGG or to generate sub-pathway.

ECO-MP database is created to facilitate user to reconstruct metabolic pathway by various ways in *E. coli* by utilizing knowledge of subnetworks, alternative pathways and known pathways information represented in it (Fig. 1). ECO-MP database driven source is made using MySql and PHP and it is available publicly. The schema includes three category namely known, unknown and constructed pathways. The pathway table comprises of accession number, sequence, KEGG gene information link, KEGG pathway link and function. The unknown pathway table comprises of accession number, predicted function of sequence which is not having pathway information and only links to KEGG gene information. The constructed table has accession number of sequence which is having both reaction and pathway information and are link to KEGG and corresponding reactions. The database can be queried with sequence function or accession number provided in-house ranging from ecmp 0-4289. The database also can be queried with input sequence in the form of fasta format, ecmp number and functions. For each query refresh button should be used so that options can be chosen easily. The Query-result outputs function, pathway link, if available in KEGG or link to subnetwork constructed internally. In case, if there is no pathway available for a user queried entity, then they are linked to gene information page of KEGG. The FAQ page provides support to the user about the database.

RESULTS AND DISCUSSION

An accurate understanding of primary metabolism such as central carbon metabolism, energy metabolism, secondary metabolism and signaling pathways of *E. coli* can be achieved by accessing ECO-MP. It emphasizes the scope of a set of biochemical products which are produced by *E. coli* in its biochemical networks and it can be queried to determine, if a particular protein is found in

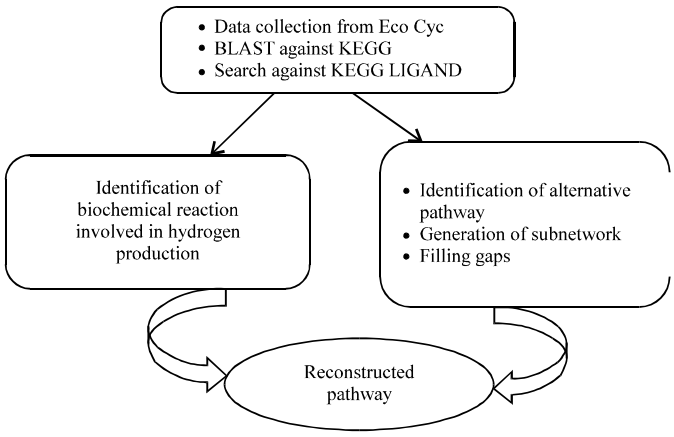

Fig. 1: Work flow diagram of ECO-MP

known pathways of KEGG, or to subnetworks constructed in house. The intact results of ECO-MP database shows that some of the protein sequences have specific functions from KEGG database and for others no hits were found. For some sequences, only biochemical reactions are available in KEGG database. Many other protein sequences are membrane and trans-membrane proteins which do not have any pathway data or biochemical reactions in the KEGG database. Totally 2,560 known and 1,730 unknown proteins have been searched against KEGG pathways to facilitate metabolic pathway reconstruction in *E. coli* and as a result, around 217 protein sequences were linked to subnetworks and 1,504 sequences to known pathways by network expansion method. The homepage of ECO-MP (Fig. 2) has options for accessing the database by sequence, function or accession number in which the accession numbers are also categorized into pathway, unknown pathway and constructed pathway categories are displayed as a tabular format. By a single click on accession number, the user can get further information about the sequence.

Known pathways: The pathways are categorized into known pathways on blasting against KEGG database, if it produced hit function in *E. coli* and the pathway link is available. The information from KEGG are gathered and tabularized and further clicking the KEGG accession links it takes the user to KEGG pathway.

Subnetworks: When there is homology against KEGG database and pathway information is not available then the chemical reactions are used to construct subnetwork by means of network expansion method Handorf and Ebenhoh (2007) and the subnetwork graph is drawn using Graphviz software (Zhao, 2006). This subnetwork information from ECO-MP can be used to reconstruct pathway or in filling metabolic holes. It is confirmed that data retrieved from ECO-MP can be used to identify minimal number of reactions that is to be added for further reconstruction work in *E. coli*. In the subnetwork shown in Fig. 3, 2-3-diaminopropanoate is available as a reactant and produce pyruvate as an immediate product and are not linked to any of the known

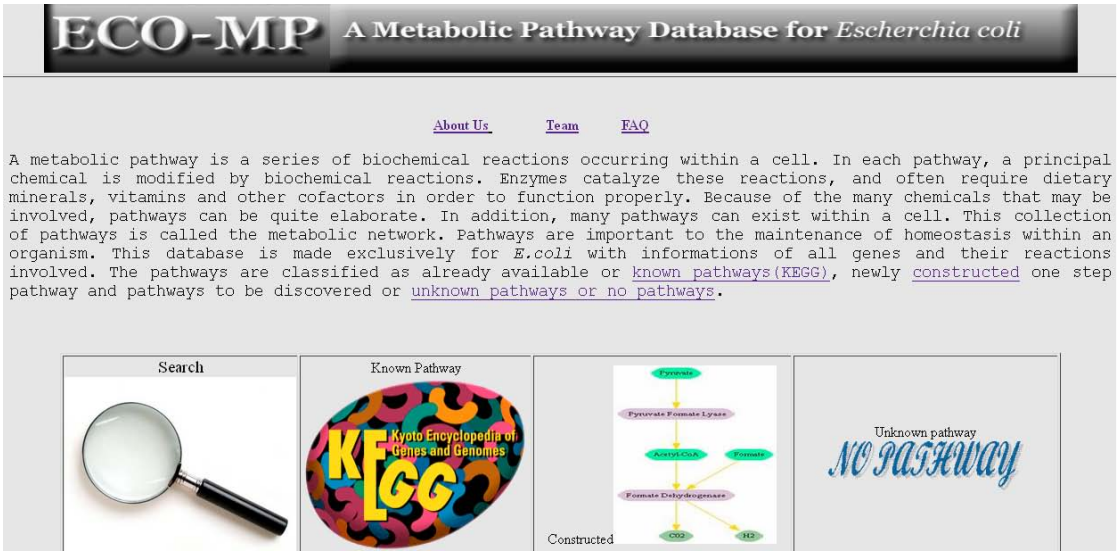

Fig. 2: Snapshot of homepage of ECO-MP. Searches can be done by accession number, function, sequence or by directly clicking the link of ECO-MP accession number

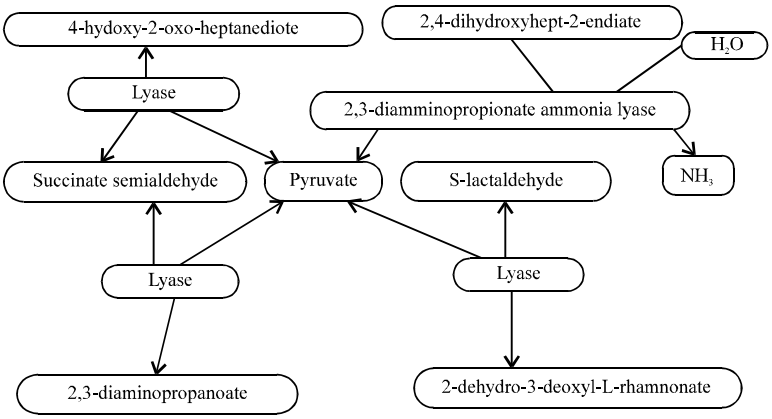

Fig. 3: 2-3-diaminopropanoate linked with pyruvate by network expansion method for hydrogen production

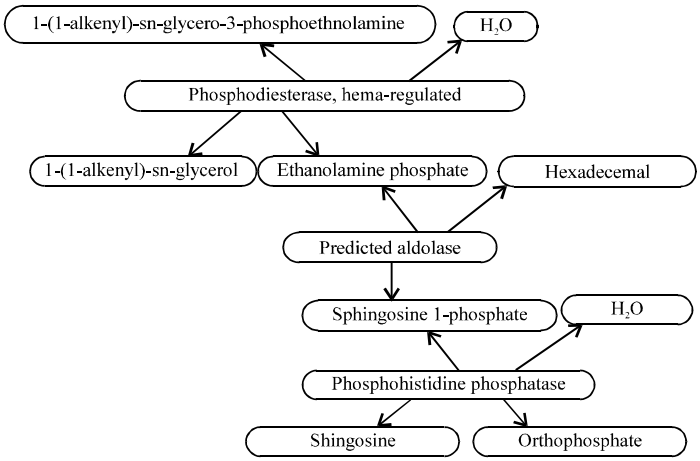

Fig. 4: 1-(1-alkenyl)-sn-glycero-3-phosphoethanolamine linked to sphingolipid metabolism by network expansion method

pathway in KEGG and this knowledge can be applied in filling metabolic holes. Similarly in sphingolipid metabolism, 1-(1-alkenyl)-sn-glycero-3-phosphoethanolamine (Fig. 4) is not specified in the formation of sphingosine which forms a primary part of sphingolipids, with an unsaturated hydrocarbon chain and has potential role in various cellular process. These subnetworks created are the prime step for analyzing the complex biochemical network and however, wet-lab examination has to be performed for further evaluation of those compounds which may not have pathway information in KEGG, since they act as an active participant in chemical reaction.

CONCLUSION

The ECO-MP database is a valuable resource for metabolic pathways and subnetworks which caters research community to analyze the in-depth knowledge of biological processes of *E. coli*. The subnetworks generated in ECO-MP are able to give ideas on creating alternative pathways and filling holes which can help in finding alternate metabolic pathways involved in important biochemical process. The user can interact with ECO-MP through a web interface with flexible searching capabilities by querying gene name or ecmp id and a resulting output with supported

links to pathways and generated subnetwork information. This idea can be extended in altering metabolic pathways to enhance the production of industrial products. The cumulated information of subnetworks and metabolic pathways available in the ECO-MP can be applied in drug discovery process on attaining the knowledge of the disease related pathways. The future release of the database will also include more subnetworks, alternate pathways and comparative analysis of pathways in all *E. coli* strains.

ACKNOWLEDGMENT

The authors thank Mr. R. Sathish Kumar and Dr. Latha Prabakar for perusing the manuscript and providing valuable suggestions.

REFERENCES

Chou, C.H., W.C. Chang, C.M. Chiu, C.C. Huang and H.D. Huang, 2009. FMM: A web server for metabolic pathway reconstruction and comparative analysis. Nucleic Acids Res., 37: 129-134.

Elliott, B., M. Kirac, A. Cakmak, G. Yavas and S. Mayes *et al.*, 2008. PathCase: Pathways database system. Bioinformatics, 24: 2526-2533.

Handorf, T. and O. Ebenhoh, 2007. MetaPath Online: A web server implementation of the network expansion algorithm. Nucleic Acids Res., 35: 613-618.

Kanehisa, M., M. Araki, S. Goto, M. Hattori and M. Hirakawa *et al.*, 2008. KEGG for linking genomes to life and the environment. Nucleic Acids Res., 36: 480-484.

Karp, P.D., M. Riley, M. Saier, I.T. Paulsen and J. Collado-Vides *et al.*, 2002. The ecocyc database. Nucleic Acids Res., 30: 56-58.

Keseler, I.M., J. Collado-Vides, S. Gama-Castro, J. Ingraham and S. Paley *et al.*, 2005. EcoCyc: A comprehensive database resource for *Escherichia coli*. Nucleic Acids Res., 33: 334-337.

Le Fevre, F., S. Smidtas and V. Schachter, 2007. Cyclone: Java-based querying and computing with Pathway/Genome databases. Bioinformatics, 23: 1299-1300.

Natalie, C.D., A.B. Scott, J. Neema, T. Ines and L.M. Monica *et al.*, 2006. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc. Natl. Acad. Sci. USA., 104: 1777-1782.

Rajadurai, C.P., T.K. Subazini and G.R. Kumar, 2011. An Integrated Re-Annotation Approach for Functional Predictions of Hypothetical Proteins in Microbial Genomes. Curr. Bioinform., 6: 450-461.

Rudd, K.E., 2000. EcoGene: A genome sequence database for *Escherichia coli* K-12. Nucleic Acids Res., 28: 60-64.

Schlitt, T. and A. Brazma, 2007. Current approaches to gene regulatory network modelling. BMC Bioinform., Vol. 8. 10.1186/1471-2105-8-S6-S9

Schuster, S., T. Pfeiffer, F. Moldenhauer, I. Koch and T. Dandekar, 2002. Exploring the pathway structure of metabolism: Decomposition into subnetworks and application to *Mycoplasma pneumonia*. Bioinformatics, 18: 351-361.

Serres, M.H., S. Goswami and M. Riley, 2004. GenProtEC: An updated and improved analysis of functions of *Escherichia coli* K-12 proteins. Nucleic Acids Res., 32: D300-D302.

Urbanczyk-Wochniak, E. and L.W. Sumner, 2007. MedicCyc: A biochemical pathway database for *Medicago truncatula*. Bioinformatics, 23: 1418-1423.

Yeh, I., T. Hanekamp, S. Tsoka, P.D. Karp and R.B. Altman, 2004. Computational analysis of *Plasmodium falciparum* metabolism: Organizing genomic information to facilitate drug discovery. Genome Res., 14: 917-924.

Zhao, J.H., 2006. Pedigree-drawing with R and graphviz. Bioinformatics, 22: 1013-1014.