BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 18 2007, pages 2501–2503
doi:10.1093/bioinformatics/btm351

Databases and ontologies

EcoProDB: the Escherichia coli protein database

Hongseok Yun[1,2], Jeong Wook Lee[1], Joonwoo Jeong[2], Jaesung Chung[3], Jong Myoung Park[1,2], Han Na Myoung[1,2] and Sang Yup Lee[1,2,3,*]

[1]Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering (BK21 program), Center for Systems and Synthetic Biotechnology, Institute for the BioCentury, BioProcess Engineering Research Center, [2]Bioinformatics Research Center and [3]Department of BioSystems, Korea Advanced Institute of Science and Technology, 335 Gwahangno, Yuseong-gu, Daejeon 305-701, Korea

Received on March 31, 2007; revised on June 13, 2007; accepted on July 1, 2007

Advance Access publication July 10, 2007

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: EcoProDB is a web-based database for comparative proteomics of *Escherichia coli*. The database contains information on *E. coli* proteins identified on 2D gels along with other resources collected from various databases and published literature, with a special feature of showing the expression levels of *E. coli* proteins under different genetic and environmental conditions. It also provides comparative information of subcellular localization, theoretical 2D map, experimental 2D map and integrated protein information via an interactive web interface and application such as the Map Browser. Users can also upload their own 2D gels, extract core information associated with the proteins and 2D gel results from different experiments and consequently generate new knowledge and hypotheses for further studies.

Availability: EcoProDB database system is accessible at http://eecoli.kaist.ac.kr

Contact: leesy@kaist.ac.kr

1 INTRODUCTION

Recent advances in proteomics along with other omics technologies have facilitated systems-level analysis of organisms toward our understanding of global cellular behaviors (Lee and Lee, 2003). Increasing amounts of proteome data as well as protein interaction and expression profile data are being accumulated. Also, several databases, such as UniProt (http://www.expasy.org) and NCBI (http://www.ncbi.nlm.nih.gov), have been developed. *Escherichia coli* has been a model organism for many years, which is also true for proteomic studies (Tonella *et al.*, 2001). However, even for *E. coli*, most of proteome studies have been performed by 2D gel electrophoresis (Witzmann and Li, 2002), which makes it difficult to retrieve, analyze and compare the data.

In this article, we report the development of a protein database called EcoProDB, which provides comprehensive information on the proteins at the whole proteome level and comparative 2D gel results under different genetic and environmental conditions from literature, using the interactive interface and search algorithm. EcoProDB contains detailed information on protein localization, theoretical and experimental 2D maps, integrated protein information collected from UniProt, the *E. coli* Cell Envelope Protein Data Collection (ECCE) (http://www.cf.ac.uk/biosi/staff/ehrmann/tools/ecce/) and literature, and online tools for comparative analysis and submission for user's data.

2 DATABASE CONTENT

EcoProDB (release 1.2) contains all relevant information on 744 non-redundant proteins; 336 proteins were retrieved from the *E. coli* SWISS 2DPAGE database (http://www.expasy.org/ch2d/) and additional 408 proteins obtained from the published literature (Han and Lee, 2006). Collected data were curated using the UniProt Knowledgebase (release 8.7). The unique feature showing the differential expression levels of proteins under different genetic and/or environmental conditions allows users to compare and predict the increase and/or decrease of the expression levels of particular proteins.

Along with the core data described above, additional information on comparative proteomics was integrated into EcoProDB, including protein localization, and theoretical and experimental 2D maps. For the comparative subcellular proteomic studies, the data in EcoProDB were categorized into cytosolic, periplasmic and inner or outer membrane proteins by integrating the information on 572 proteins (Lopez-Campistrous *et al.*, 2005) and on 122 proteins from the ECCE database. The theoretical 2D map plots the theoretical isoelectric point (p*I*) versus the theoretical molecular mass (M*w*) of the open reading frames in *E. coli*. These values are calculated using Compute p*I*/M*w* tool (http://kr.expasy.org/tools/pi_tool.html). Each protein spot in the map is linked to the detailed information, such as function, expression, experimental 2D data and cross-references. The experimental 2D map provides a collection of 81 2D gels or autoradiograms which have been published in 23 articles. Total of 2552 spots of proteins were listed with their IDs and properties including protein name, accession number in other databases, functional category, sequence coverage, abundance, fold change under the comparative conditions. When available, the method of protein identification and detailed experimental conditions are also

*To whom correspondence should be addressed.

provided. Thus, EcoProDB can serve as a one-stop resource for retrieving all available information on particular proteins of interest in *E. coli*.

3 SYSTEM ARCHITECHTURE AND IMPLEMENTATION

The EcoProDB is a web-based comparative proteomics system that uses *E. coli* proteins identified on 2D gels as its primary source of data through interactive web applications such as the Map browser. The system employs MySQL5.01 as a relational database management system (RDBMS) and IIS and Apache servers as web server platform. Its web interface is implemented using ASP.NET, PHP, Perl, asynchronous javascript and XML (AJAX) and document object model (DOM). For the flexible integration of information present in EcoProDB and UniProt, the XML-based object model of UniProt was retrieved and mapped onto the core data in EcoProDB. Their corresponding relationship was then stored in the mapping table, which allows efficient integration of the data present in the two resources. The extensible stylesheet language transformation (XSLT) generates the dynamic web page showing the integrated information on proteins (Fig. 1).

4 WEB INTERFACE

For more user-friendly searching, browsing and retrieval of data and information, EcoProDB provides interactive web interfaces including the Map Browser, the Comparable Subcellular Localization and the Keyword Search. The Map Browser was designed to access even the very crowded 2D maps. Some proteins are positioned very closely from one another in the 2D map, which often makes protein identification and analysis difficult. The Map Browser provides the zoom and the pan functions, and can magnify and explore the crowded region. Moreover, combined with search function, the Map Browser can find the location of target proteins. The Comparable Subcellular Localization shows a browsable list of proteins with the subcellular locations annotated in UniProt, published literature and ECCE database. The list can be grouped by their subcellular locations by clicking each column name of their sources. The search function can be used to search for the proteins with special conditions, such as protein name, Swiss-Prot accession number, p*I*, M*w*, description, function and expression, gene name, author, published literature and so forth. In addition to the effective retrieval of information, EcoProDB provides the online tools for users to upload their 2D gel and related information for comparative analysis and/or database expansion. Using the search algorithm

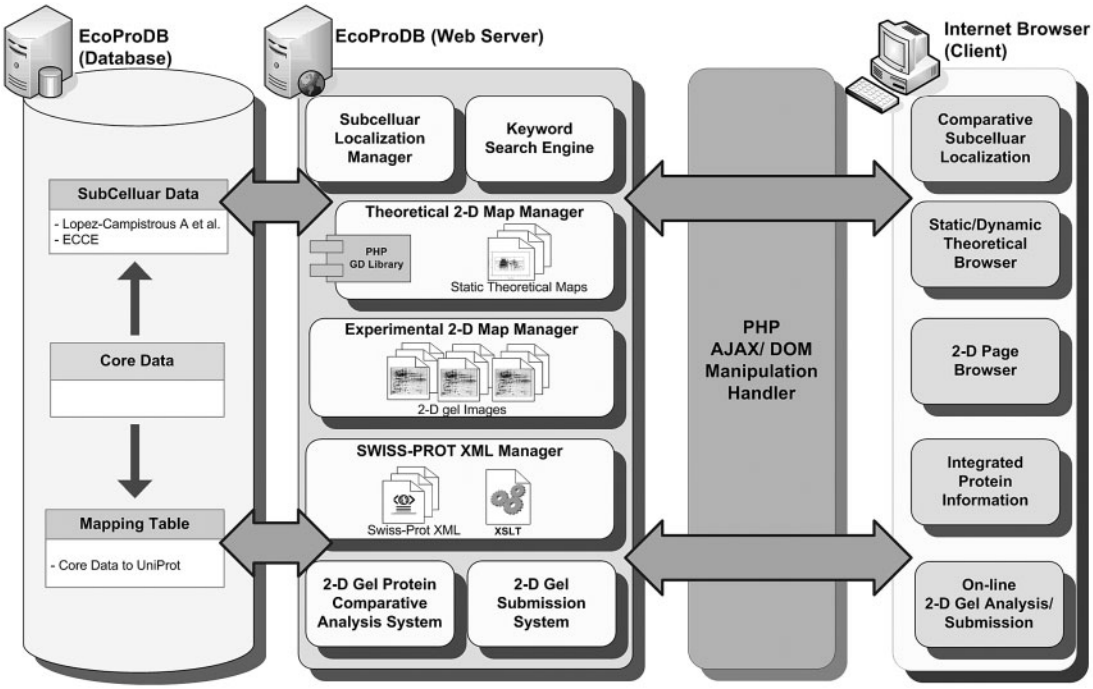

Fig. 1. EcoProDB is a web-based database which allows comprehensive protein analysis at the whole proteome level and comparative analysis of the protein expression levels under different genetic and environmental conditions, using the interactive interface and search algorithm. *E. coli* proteins identified on 2D gels were used as its primary source of data through interactive tools such as Map Browser. The storage backend is composed of MySQL relational database management system. The database records are mapped and integrated onto the XML-based object model of the UniProt, which allows integration of the data present in EcoProDB and UniProt. The dynamic web page showing the integrated protein information is generated through the extensible stylesheet language transformation (XSLT). The data themselves are queried by search engine and map browser using web development technique for interactive web applications such as asynchronous javascript and XML (AJAX) and document object model (DOM).

and comparison tools associated with the 2D Maps in EcoProDB, users can efficiently obtain the relevant information associated with the proteins in their experimental 2D gel before mass spectrometric analysis, and consequently generate new knowledge and hypotheses for further investigation.

5 FUTURE PROSPECT

EcoProDB is the first comprehensive and specialized database for studying *E. coli* proteins in conjunction with proteome analysis. It will be continuously updated and upgraded as new information becomes available. We also plan to integrate the information on metabolic pathways, enzyme activities and regulatory mechanisms into the EcoProDB, which will allow us to better understand the global physiology of *E. coli* under various genetic and environmental conditions.

ACKNOWLEDGEMENTS

This work was supported by the Korean Systems Biology Research Program (M10309020000-03B5002-00000) of the Ministry of Science and Technology through the Korea Science and Engineering Foundation. Further supports by LG Chem Chair Professorship, Microsoft and IBM-SUR program are appreciated.

Conflict of Interest: none declared.

REFERENCES

Han,M.J. and Lee,S.Y. (2006) The *Escherichia coli* proteome: past, present, and future prospects. *Microbiol. Mol. Biol. Rev.*, **70**, 362–439.

Lee,P.S. and Lee,K.H. (2003) *Escherichia coli*-a model system that benefits from and contributes to the evolution of proteomics. *Biotechnol. Bioeng.*, **84**, 801–814.

Lopez-Campistrous,A. *et al.* (2005) Localization, annotation & comparison of the *Escherichia coli* K-12 proteome under two states of growth. *Mol. Cell Proteomics*, **4**, 1205–1209.

Tonella,L. *et al.* (2001) New perspectives in the *Escherichia coli* proteome investigation. *Proteomics*, **1**, 409–423.

Witzmann,F.A. and Li,J. (2002) Cutting-edge technology. II. Proteomics: core technologies and applications in physiology. *Am. J. Physiol. Gastrointest. Liver Physiol.*, **282**, G735–G741.