

CIDB: Chlamydia Interactive Database for cross-querying genomics, transcriptomics and proteomics data

Yan Chen^a, Peter Timms^c, Yi-Ping Phoebe Chen^{a,b,*}

^a School of Engineering and Information Technology, Deakin University, Australia

^b Australia Research Council Centre of Excellence in Bioinformatics, Australia

^c School of Life Sciences, Queensland University of Technology, Australia

Received 19 February 2007; received in revised form 13 August 2007; accepted 13 August 2007

Abstract

Chlamydiae are important pathogens of humans, birds and a wide range of animals. They are a unique group of bacteria, characterized by their developmental cycle. *Chlamydia* has been difficult to study because of their obligate intracellular growth habit and lack of a genetic transformation system. However, the past 5 years has seen the full genome sequencing of seven strains of *Chlamydia* and a rapid expansion of genomic, transcriptomic (RT-PCR, microarray) and proteomic analysis of these pathogens. The Chlamydia Interactive Database (CIDB) described here is the first database of its type that holds genomic, RT-PCR, microarray and proteomics data sets that can be cross-queried by researchers for patterns in the data. Combining the data of many research groups into a single database and cross-querying from different perspectives should enhance our understanding of the complex cell biology of these pathogens. The database is available at: <http://www3.it.deakin.edu.au:8080/CIDB/>.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Genomics; Transcriptomics; Proteomics Chlamydia data

1. Introduction

Chlamydiae are virulent pathogens that reside in humans, birds and a wide range of animals (Schachter et al., 1986). In humans, Chlamydia cause significant diseases such as trachoma (over 500 million people affected), sexually transmitted infection leading to salpingitis and pelvic inflammatory disease and infertility in women, community acquired respiratory disease, bronchitis, sinusitis and pneumonia in some patients. Most recently, they have been linked to atherosclerosis (Bedson et al., 1980; Gerbase et al., 1998; Saikku et al., 1998).

Phylogenetically, Chlamydia are a unique group of bacteria, characterized by their developmental cycle that involves the inter-conversion between two distinct morphological forms, a small metabolically inactive elementary body (EB) and a large metabolically active and multiplying reticulate body (RB). The complete developmental cycle ranges from 48 to 72 h depending on the infecting strain, host cell and environmental conditions.

In addition to acute infections, Chlamydia often manifest as chronic on-going infection which leads to the eventual pathology and serious disease sequelae. *In vitro* models have been developed to study chronic or persistent phase of the *Chlamydia* developmental cycle (Beatty et al., 1993) and several studies have shown altered gene expression and proteomic profiles in persistent cultures as compared to normal cultures (Molestina et al., 2002; Hogan et al., 2003). The persistent state can be induced *in vitro* via several conditions, including nutrient deprivation, gamma-interferon treatment, bacteriophage infection, long-term continuous culture and antibiotic treatment (Kutlin et al., 1999). The last one is highly relevant *in vivo* and may explain the poor response of chronic *C. pneumoniae* infections to antibiotic treatment.

Chlamydiae have been found to be difficult to study due to their obligate intracellular growth habit and lack of a genetic transformation system. However, the past 5 years has seen the full genome sequencing of seven strains of *Chlamydia* and a rapid expansion of genomic, transcriptomic and proteomic analysis of these pathogens. Many groups have conducted transcriptomics (quantitative RT-PCR on specific genes (Kutlin et al., 2001; Mathews et al., 2001), microarray on the whole transcriptome (Belland et al., 2003; Nicholson et al., 2003) and

* Corresponding author at: School of Engineering and Information Technology, Deakin University, Australia.

E-mail address: phoebe@deakin.edu.au (Y.-P. Chen).

proteomics studies (Chen, 2005; Chen and Chen, 2006; Knapp and Chen, 2007)) and this rapid collection and analysis of data will enhance our understanding of the complex biology of the important pathogens.

However, this wealth of new data is presently not well connected for the purpose of analysis. The Chlamydia Interactive Database (CIDB) was therefore developed as a prototype database, to accommodate the large quantity of experimental data relating to *chlamydial* research and to facilitate researchers cross-query the data sets from multiple perspectives.

The aims of the current research are therefore, (1) to create the actual storage database and (2) to design a bioinformatics data mining tool that would assist researchers to better identify important biological patterns present in the complex datasets. A web-based user query interface was created to facilitate the retrieval, analysis and visualization of the gene profile data. All data is cross-referenced and can be cross-queried from different perspectives, which include transcriptomics (RT-PCR, microarray), proteomics and genomics data, including data from one assay method but analysed under different laboratory conditions (eg. various time points throughout the normal

chlamydial developmental cycle, persistent models, other external factors, such as heat shock).

2. Structure of the database

The CIDB system is a relational database which has been developed using the MySQL (MySQL, 2005) database system. All data (quantitative RT-PCR data, microarray data, proteomics data and promoter prediction data) is stored in a MySQL database on an MS Windows platform. The system architecture is shown schematically in Fig. 1.

The query interfaces of this database are web-based and implemented with JavaServer Pages (JSP, 2005). The middle tier of the website is serviced by the Apache web server integrated with the Tomcat Servlet engine (Tomcat, 2005). JSP pages utilizing MySQL's JDBC driver are created to first query the database, and then tunnel client-side applets to the clients to provide necessary graphical visualization of the query results. Fig. 2 is a screen shot of a client's webpage which currently displays the query result for the gene "cpa". The transcription profiles of this gene under either "normal" or "persistent"

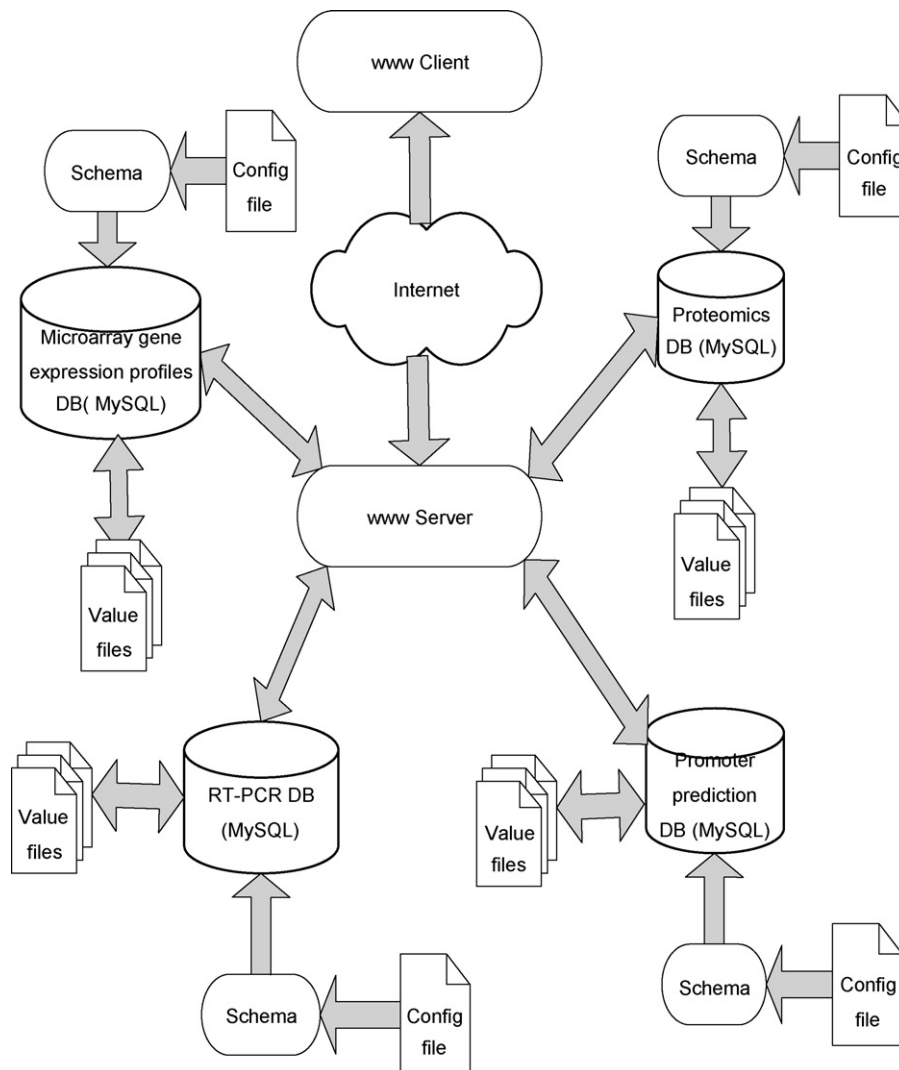


Fig. 1. A schematic illustration of the core database architecture for CIDB.

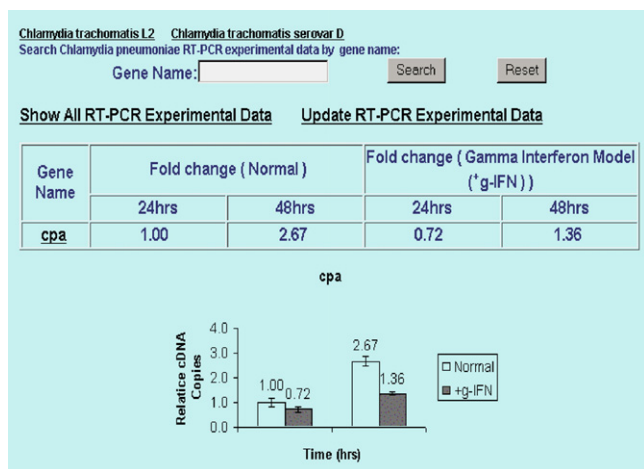


Fig. 2. A screen shot of a client webpage which provides both textual and graphical comparisons of the gene expression profiles.

conditions are displayed. In addition, the bar graph displays direct visual comparison of the gene expression levels under these two biological conditions. JSP is the only reliable means to deliver both textual and graphical comparisons of gene expression profiles across different platforms by tunneling client-side applets.

3. Data availability and query methods

The prototype CIDB database currently contains five types of data sources:

- Quantitative RT-PCR expression data for 66 genes from two developmental time points (24 and 48 h) under both normal growth conditions (Mathews et al., 2001) and also under gamma-interferon induced persistence conditions (Chen, 2005).
- Microarray gene expression profiles for *C. trachomatis* serovar D (Belland et al., 2003) which contains 901 gene expression patterns for six time points (1, 3, 8, 16, 24 and 40 h). Microarray gene expression profiles for *C. trachomatis* L2 (Nicholson et al., 2003), which contain microarray data for 890 genes at two time points (24 and 48 h).
- Promoter data which includes a list of genes for which the promoters have been predicted (Molestina et al., 2002; Chen, 2005). These are arranged into three categories (sigma-66, sigma-54 and sigma-28).
- Proteomic data for 14 genes which are arranged into three categories (up-regulated, down-regulated and unchanged) (Molestina et al., 2002).
- Genomic data is accessed via the Berkeley Genome web site (<http://chlamydia-www.berkeley.edu:4231>) and enables provision of the predicted gene function and the gene arrangement maps.

Currently, there are two main query approaches supported by the web-based user interface, although other types of queries can easily be added to the functions.

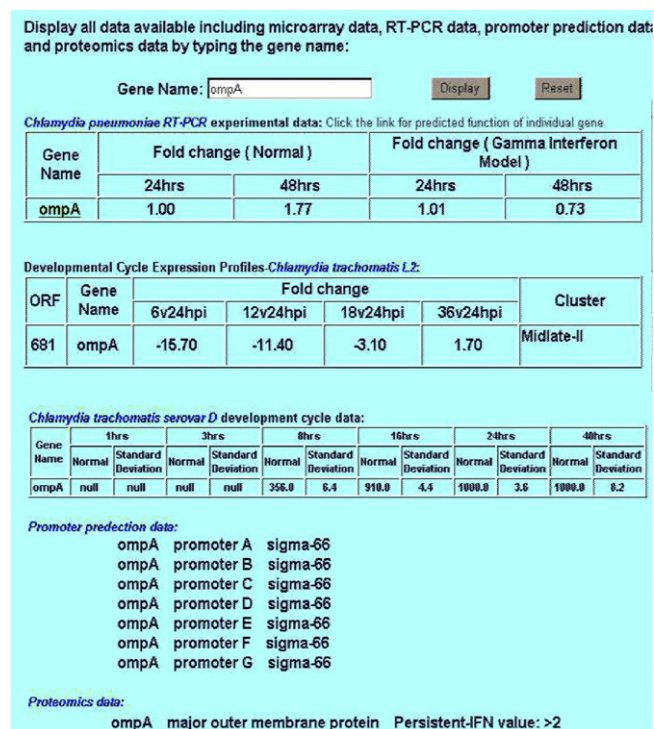


Fig. 3. A screen shot of a client’s webpage with an “all data” query.

- Query “all data” by a gene name and display the information in relation to gene name, predicted function, microarray transcription information (category of gene and value for expression), RT-PCR transcription information (ratios of expression for normal and interferon-induced persistence conditions, if available), proteomics profile (if available), predicted promoter type, and the position of the gene in relation to neighboring genes.
- Analysis of any selected sub-sets of data, by selecting any one of the four variables (microarray data, RT-PCR data, proteomics data and promoter data), choosing “all” or any of the sub-sets of data available for each of these categories. This function is the most powerful use of CIDB.

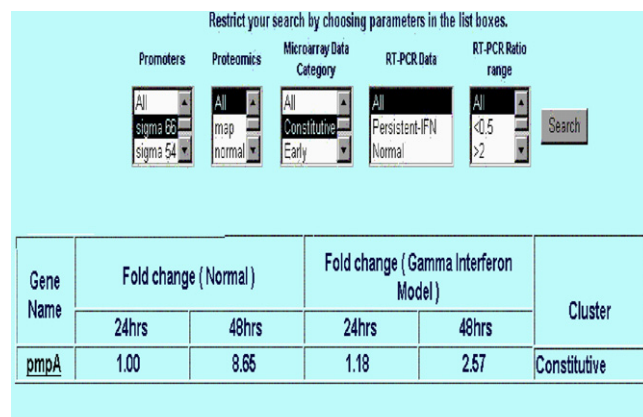


Fig. 4. A screen shot of a type B query.

3.1. Example of query type A: display of “all data” for a particular gene or protein

By using the “All data” link and entering the gene or protein of interest, the CIDB displays all available data, including

microarray data, RT-PCR data (under normal and interferon gamma induced persistence), promoter prediction data and any proteomics data. Direct links to the Berkeley genome web page for information such as “predicted function” and “genome map location” are also provided.



Fig. 5. An illustration of cross-referencing data from external databases for the gene “pgk”. In this case, the gene’s RT-PCR expression profiles are compared with its homology functions, translated amino acid sequence and chromosome gene map position.

[Query example]:

```
SELECT    cpn.*,    ctrl2.*,    ctrd.*,
          ctr_promoter.*, proteomics.*
FROM      cpn, ctrl2, ctrd, ctr_promoter,
          proteomics
WHERE     genename = "ompA";
```

A screen shot of an “all data” query response for the gene “ompA” is shown in Fig. 3.

3.2. Example of query type B: what genes identified as late genes by microarray data have a sigma-66 type promoter defined?

By using the “Cross reference” link, select the following sub-categories; (a) microarray data – “constitutive”, (b) RT-PCR data – “all”, (c) proteomics data – “all”, (d) promoter data – “sigma-66”. Fig. 4 shows the results for this particular query. It shows that of the 27 genes defined as “constitutive” from the microarray data, promoter prediction is only available at present for one of these, gene pmpA and it has a predicted sigma-66 type promoter. Fig. 4 shows a screen shot of this query type.

[Query example]:

```
SELECT    cpn.*, ctrl2.category
FROM      cpn, ctrl2, ctr.promoter
WHERE     ctrl2.category= "constitutive"
          and
          ctr_promoter.category =
          "sigma-66"
          and
          cpn.genename =
          ctrl2.genename
          and
          ctrl2.genename =
          ctr_promoter.genename;
```

3.3. Dynamics of the database enquiries

Basic queries and visualization of the differential expression profiles of the selected gene are supported in the query interfaces (see Fig. 1) via web display. Graphical display for comparing fold changes is realized by tunneling client-side applets from server backend to a client’s webpage. These databases (the RT-PCR database, the microarray databases, promoter prediction database and the proteomics database) are loosely federated in this project because separated connections are maintained by the middle tier of the web server as shown in Fig. 1. Data are retrieved separately from different databases before combining them into a unified response to user queries.

Through careful design of the query interface, the system allows a user to make a single query with a combination of several key words. The query is then parsed by JSP database programs into sub-queries for different database sources. The cross-referencing to remote databases is handled automatically with sub-queries. The results are finally combined as a single unified answer, and sent back to the client. Some of the

frequently used data retrieved from the remote databases are cached so that the queries are optimized across multiple databases.

Data integration from multiple heterogeneous data sources is a challenging task. A frequently used approach is to impose a global schema on all related databases. However, such an approach experiences difficulties in resolving inconsistencies among the meanings (?) of entities used in each schema of original databases. Overheads must be paid to preprocess query terms and translate these query terms into the required entity terms used in each individual database. We used an alternative approach to data integration in our database design, based on a work-flow model used for comparing gene expression profiles. Federated databases are connected to maintain efficiency. However, databases are queried on-demand at each stage of the work-flow for gene expression analysis by using the original database schema. This removes the need for a global schema.

To facilitate cross-referencing, sub-queries to comparison databases are embedded in the current query result page as hyperlinks. By following the work-flow of gene expression profile comparison, each click to a hyperlink executes a sub-query to a remote database. The query results are then displayed in a separate window. The embedded sub-queries form a hierarchy which reflects the levels of comparisons to be performed while analyzing gene profiles. Fig. 5 demonstrates a case of cross-referencing of the RT-PCR gene expression data for the gene pgk to its predicted protein function and its relationship with neighboring genes by using the work-flow model. DNA sequence and translated amino acid sequence are cross-referenced in this case for gene profile comparison.

In this case, the gene’s RT-PCR expression profiles are compared with its homology functions, translated amino acid sequence and chromosome gene map position.

The database is available at <http://www3.it.deakin.edu.au:8080/CIDB/>. All the data stored in this database are downloadable from the website.

4. Future research and conclusion

Federating remote and heterogeneous databases of *Chlamydia* gene expression data is a very challenging task. Currently, different approaches such as usage of schemas to virtualize the queries to the remote databases are being actively investigated. This will facilitate researchers from different projects to query and interact with the same interface, and hence reduce the complexity of maintaining multiple database connections and multiple transactions.

A relational database for *Chlamydia* transcriptomics (microarray and individual RT-PCR gene expression profiles), proteomics and genomics data and its associated web-based query interfaces were created successfully. The web-based query interfaces not only allow users to retrieve information about up-regulated and down-regulated expression profiles but also enrich the retrieved data by cross-referencing them with microarray data and other available genomics from loosely federated remote databases. This greatly enhances biologists’

abilities to make informed decision regarding to the quality of the data and accuracy of current experiments.

The other unique feature of the web-based interface is its ability of visualizing and manipulating retrieved data at client-side. Visualization tools are embedded with the query interfaces as client-side applets and are tunneled to client's webpage from the server using JSP. This allows the retrieved gene expressions be visualized and manipulated by different visualization tools appropriate to the types of gene expressions. With these visualization tools, biologists are able to make comparisons between gene expressions obtained during different experiments under different control conditions.

Acknowledgments

Part of this work was supported by an Australian Research Council (ARC) Discovery grant DP0344488 and NIH grant R21 AI5255-01. The authors also wish to acknowledge Richard Hogan, Adam Polkinghorne, David Good and Brian Grech for kindly supplying their biological data.

References

- Beatty, W.L., Byrne, G.I., Morrison, R.P., 1993. Morphologic and antigenic characterization of interferon g-mediated persistent *Chlamydia trachomatis* infection in vitro. *Proc. Natl. Acad. Sci.* 90, 3998–4002.
- Bedson, S.P., Western, G.T., Simpson, S.L., 1980. Observations on the aetiology of psittacosis. *Lancet* 1, 235–236.
- Belland, R.J., Zhong, G., Crane, D.D.D.H., Sturdevant, D., Sharma, J., Beatty, W., Caldwell, H.D., 2003. Genomic transcriptional profiling of the developmental cycle of *Chlamydia trachomatis*. *Proc. Natl. Acad. Sci.* 100 (14), 8478–8483.
- Chen, Y.P.P. (Ed.), 2005. *Bioinformatics Technologies*. Springer, p. 396.
- Chen, Q., Chen, Y.P.P., 2006. Mining frequent patterns for AMP-activated protein kinase regulation on skeletal muscle. *BMC Bioinformatics* 7, 394.
- Gerbase, A.C., Rowley, J.T., Mertens, T.E., 1998. Global epidemiology of sexually transmitted diseases. *Lancet* 351 (Suppl. 3), 2–4.
- Hogan, R.J., Mathews, S.A., Kutlin, A., Hammerschlag, M.R., Timms, P., 2003. Differential expression of genes encoding membrane proteins between acute and continuous *Chlamydia pneumoniae* infections. *Microb. Pathog.* 34, 11–16.
- JSP, 2005. *JavaServer Pages™ Technical Information*. <http://java.sun.com/products/jsp/techinfo.html>.
- Knapp, K., Chen, Y.P.P., 2007. An evaluation of contemporary hidden Markov model gene finders with a predicted exon taxonomy. *Nucleic Acids Res.* 35, 317–324.
- Kutlin, A., Roblin, P.M., Hammerschlag, M.R., 1999. In vitro activities of azithromycin and ofloxacin against *Chlamydia pneumoniae* in a continuous-infection model. *Antimicrob. Agents Chemother.* 43, 2268–2272.
- Kutlin, A., Flegg, C., Stenzel, D., Reznik, T., Roblin, P.M., Mathews, S., Timms, P., Hammerschlag, M.R., 2001. Ultrastructural study of *Chlamydia pneumoniae* in a continuous-infection model. *J. Clin. Microbiol.* 39 (10), 3721–3723.
- Mathews, S.A., George, C., Flegg, C., Stenzel, D., Timms, P., 2001. Differential expression of *ompA*, *ompB*, *pyk*, *nlpD* and *Cpn0585* genes between normal and interferon- γ treated cultures of *Chlamydia pneumoniae*. *Microb. Pathog.* 30, 337–345.
- Molestina, R.E., Klein, J.B., Miller, R.D., Pierce, W.H., Ramirez, J.A., Summersgill, J.T., 2002. Proteomic analysis of differentially expressed *Chlamydia pneumoniae* genes during persistent infection of *HEp-2* cells. *Infect. Immun.* 70, 2976–2981.
- MySQL, 2005. *MySQL: The World's Most Popular Open Source Database*. <http://www.mysql.com/>.
- Nicholson, T.L., Olinger, L., Chong, K., Schoolnik, G., Stephens, R.S., 2003. Global stage-specific gene regulation during the developmental cycle of *Chlamydia trachomatis*. *J. Bacteriol.* 185 (10), 3179–3189.
- Saikku, P., Leinonen, M., Mattila, K., Ekman, M.-R., Nieminen, M.S., Makela, P.H., Huttunen, J.K., Valtonen, V., 1998. Serological evidence of an association of a novel *Chlamydia*, TWAR, with chronic coronary heart disease and acute myocardial infarction. *Lancet* 2, 983–986.
- Schachter, J., Grossman, M., Sweet, R.L., Holt, J., Jordan, C., Bishop, E., 1986. Prospective study of perinatal transmission of *Chlamydia trachomatis*. *J. Am. Med. Assoc.* 255, 3374–3377.
- Tomcat, 2005. *The Apache Tomcat Project*. <http://jakarta.apache.org/tomcat/>.