A Systematic Approach to Understanding Bacterial Responses to Oxygen Using Taverna and Webservices

S. Maleki-Dizaji[1], M. Rolfe[3], P. Fisher[2], M. Holcombe[1]

[1] The University of Sheffield, Computer Science, Sheffield, United Kingdom
[2] The University of Manchester, Computer Science, Manchester, United Kingdom
[3] The University of Sheffield, Department of Molecular Biology and Biotechnology, Sheffield, United Kingdom

Abstract — ***Escherichia coli* is a versatile organism that can grow at a wide range of oxygen levels; although heavily studied, no comprehensive knowledge of physiological changes at different oxygen levels is known. Transcriptomic studies have previously examined gene regulation in *E. coli* grown at different oxygen levels, and during transitions such as from an anaerobic to aerobic environment, but have tended to analyse data in a user intensive manner to identify regulons, pathways and relevant literature. This study looks at gene regulation during an aerobic to anaerobic transition, which has not previously been investigated. We propose a data-driven methodology that identifies the known pathways and regulons present in a set of differentially expressed genes from a transcriptomic study; these pathways are subsequently used to obtain a corpus of published abstracts (from the PubMed database) relating to each biological pathway**

Keywords — ***E. coli*, Microarray, Taverna, Workflows, Web Services**

I. Introduction

Escherichia coli has been a model system for understanding metabolic and bio-energetic principles for over 80 years and has generated numerous paradigms in molecular biology, biochemistry and physiology [1]. *E. coli* is also widely used for industrial production of proteins and speciality chemicals of therapeutic and commercial interest. A deeper understanding of oxygen metabolism could improve industrial high cell-density fermentations and process scale-up. Knowledge of oxygen-regulation of gene expression is important in other bacteria during pathogenesis where oxygen acts an important signal during infection [2] and thus this project may underpin better antimicrobial strategies and the search for new therapeutics. However, current approaches have generally been increasingly reductionist, not holistic. Too little is known of how molecular modules are organised in time and space, and how control of respiratory metabolism is achieved in the face of changing environmental pressures. Therefore, a new systems-level approach is needed, which integrates data from all spatial and temporal domains. Many transcriptomic studies using microarrays have analysed data in a user-intensive manner to identify regulons, pathways and relevant literature. Here, a two-colour cDNA microarray dataset comprising a time-course experiment of *Escherichia coli* cells during an aerobic to anaerobic environment is used to demonstrate a data-driven methodology that identifies known pathways from a set of differentially expressed genes. These pathways are subsequently used to obtain a corpus of published abstracts (from the PubMed database) relating to each biological pathway identified. In this research Taverna and Web Services were used to achieve the goal.

II. Taverna and web services

Web services provide programmatic access to data resources in a language-independent manner. This means that they can be successfully connected into data analysis pipelines or workflows (Figure 1). These workflows enable us to process a far greater volume of information in a systematic manner. Unlike that of the manual analysis, which is severely limited by human resources, we are only limited by the processing speed, storage space, and memory of the computer executing these workflows. Still the major problems with current bioinformatics investigations remain; which are the lack of recording experimental methods, including software applications used, the parameters used, and the use of hyperlinks in web pages. The use of workflows limits issues surrounding the manual analysis of data, i.e. the bias introduced by researchers when conducting manual analyses of microarray data. Processing data through workflows also increases the productivity of the researchers involved in the investigations, allowing for more time to be spent on investigating the true nature of the detailed information returned from the workflows. For the purpose of implementing this systematic pathway-driven approach, we have chosen to use the Taverna workbench[3,4]. The Taverna Workbench allows bioinformaticians to construct complex data analysis pipelines from components (or web services) located on both remote and local machines. These pipelines, or workflows, are then able to be executed over a set of unique data values, producing results which can be visualised within the Taverna workbench itself. Advantages of the Taverna workflow work-

Chwee Teck Lim, James C.H. Goh (Eds.): ICBME 2008, Proceedings 23, pp. 77–80, 2009
www.springerlink.com

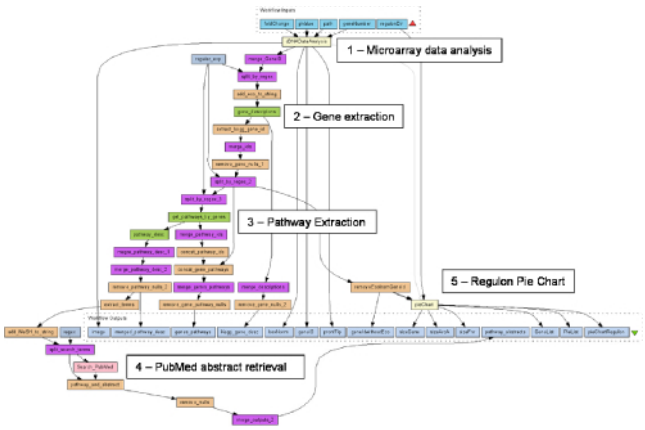


Figure 1 – Workflow Diagram

bench include, repeatable, re-useable the and limiting user bias by removing intermediate manual data analysis. Greater volume of data can be processed in a reduced time period We propose a data-driven methodology that identifies the known pathways from a set of differentially expressed genes from a microarray study (Figure 1). This workflow consists of three parts: microarray data analysis; pathways extraction; and PubMed abstract retrieval. This methodology is implemented systematically through the use of web services and workflows.

A. Microarray Data Analysis

Despite advances in microarray technology that have led to increased reproducibility and substantial reductions in cost, the successful application of this technology is still elusive for many laboratories. The analysis of transcriptome data in particular presents a challenging bottleneck for many biomedical researchers. These researchers may not possess the necessary computational or statistical knowledge to address all aspects of a typical analysis methodology; indeed, this is something which can be time consuming and expensive, even for experienced service providers with many users. Currently available transcriptome analysis tools include both commercial software (GeneSpring [5], ArrayAssist [6]) and non-commercial software (Bioconductor [7]).

The open source Bioconductor package [7] is one of the most widely used suites of tools used by biostatisticians and bioinformaticians in transcriptomics studies. Although both highly powerful and flexible, users of Bioconductor face a steep learning curve, which requires users to learn the *R* statistical scripting language as well as the details of the Bioconductor libraries.

The high overheads in using these tools provide a number of disadvantages for the less experienced user such as, requirement for expensive bioinformatics support, considerable effort in training, less than efficient utilisation of data, difficulty in maintaining consistent standards methodologies, even within the same facility, Difficult integration of additional analysis software and resources and limited reusability of methods and analysis frameworks.

The aim of this work was to limit these issues. Users will, therefore, be able to focus on advanced data analysis and interpretational tasks, rather than common repetitive tasks. We have observed that there is a core of microarray analysis tasks common to many microarray projects. Additionally, we have identified a need for microarray analysis software to support these tasks that has minimal training costs for inexperienced users, and can increase the efficiency of experienced users.

Microarray Data Analysis part provides support to construct a full data analysis workflow, including loading, normalisation, T-test and filtering of microarray data. In addition to returning normalised data, it produces a range of diagnostic plots of array data, including histograms, box plots and principal components analysis plots using R and Bioconductor.

B. Pathway extraction

This part of the workflow searches for genes found to be differentially expressed in the microarray data, selected based on a given p-value from the Microarray Data Analysis part. Gene identifiers from this part were subsequently cross-referenced with KEGG gene identifiers, which allowed KEGG gene descriptions and KEGG pathway descriptions to be returned from the KEGG database.

C. PubMed abstract retrieval

In this part, the workflow takes in a list of KEGG pathway descriptions. The workflow then extracts the biological pathway process from the KEGG formatted pathway descriptions output. A search is then conducted over the PubMed database (using the eFetch web service) to identify up to 500 abstracts related to the chosen biological pathway. At this stage, a MeSH tag is assigned to the search term, in order to reduce the number of false positive results returned from this initial search. All identified PubMed identifiers (PMID) are then passed to the eSearch function and searched for in PubMed. Those abstracts found are then returned to the user along with the initial query string – in this case, the pathway [3].

D. Pie chart

At present, results from transcriptional profiling experiments (lists of significantly regulated genes) have largely been interpreted manually, or using gene analysis software

(i.e. GeneSpring, GenoWiz) that can provide links to databases that define pathways, functional categories and gene ontologies. Many databases, such as EcoCyc [8] and RegulonDB [9], contain information on transcriptional regulators and regulons (genes known to be regulated by a particular transcription factor), and automatic interpretation of a transcriptional profiling dataset using these databases is still in its infancy. When applied to the results of a transcriptional profiling experiment, this may confirm the importance of a regulator that is already known, or suggest a role for a previously unknown regulator, which may be investigated further. The pie chart shown indicates the number of genes in a dataset that are regulated by a known transcriptional regulator, or by combination of regulators, and can suggest previously unknown regulatory interactions. The information for each regulon comes from files that are created manually from the EcoCyc database.

III. Case study: escherichia coli

Escherichia coli is a model laboratory organism that has been investigated for many years due to its rapid growth rate, simple growth requirements, tractable genetics and metabolic potential [10]. Many aspects of *E. coli* are well characterised, particularly with regards to the most familiar strain K-12, with a sequenced genome[11], widespread knowledge of gene regulation (Regulon DB; [9] and well documented metabolic pathways (EcoCyc; [8]). Indeed, it has been said that more is known about *E. coli* than about any other organism [1] and for these reasons *E. coli* stands out as a desirable organism on which to work (Mori, 2004).

A. Growth conditions

Escherichia coli K-12 strain MG1655 was grown to a steady-state in a Labfors-3 Bioreactor; Infors HT; Bottmingen, Switzerland) under the following conditions (vessel volume 2 L; culture volume 1 L; Evans medium pH 6.9 [12]; stirring 400 rpm; dilution rate 0.2 h^{-1}). To create an aerobic culture 1 L min^{-1} air was sparged through the chemostat, whilst for anaerobic conditions 1 L min^{-1} 5 % CO_2 95 % N_2 ($^{v}/_{v}$) was passed through the chemostat. For steady-state to be reached, continuous flow was allowed for at least 5 vessel volumes (25 hours) before cultures were used. Gas transitions were carried out on steady-state cultures by switching gas supply as required.

B. Isolation of RNA

A steady-state chemostat culture was prepared and samples were removed from the chemostat for RNA extraction just prior to the gas transition and 2, 5, 10, 15 and 20 minutes after the transition. Samples were taken by direct elution of 2 ml culture into 4 ml RNAprotect (Qiagen; Crawley, UK) and using RNeasy RNA extraction kit (Qiagen Crawley, UK) following the manufacturers instructions. RNA was quantified spectrophotometrically at 260 nm.

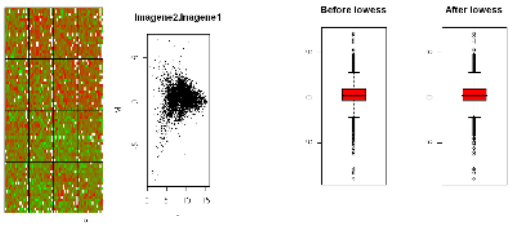
C. Transcriptional Profiling

16 µg RNA for each time point was labelled with Cyanine3-dCTP (Perkin-Elmer; Waltham, USA) using Superscript III reverse transcriptase (Invitrogen; Paisley, UK). Manufacturers instructions were followed for a 30 µl reaction volume, using 5 µg random hexamers (Invitrogen; Paisley, UK) with the exception that 3 nmoles Cyanine3-dCTP and 6 nmoles of unlabelled dCTP was used. Each Cyanine3-labelled cDNA sample was hybridised against 2 µg of Cyanine5-labelled K-12 genomic DNA produced as described by Eriksson[13]. Hybridisation took place to Ocimum OciChip K-12 V2 microarrays (Ocimum; Hyderabad, India) at 42 °C overnight and washed according to the manufacturers instructions. Slides were scanned on an Affymetrix 428 microarray scanner at the highest PMT voltage possible that didn't give excessive saturation of microarray spots. For each time point, two biological replicates and two technical replicates were carried out.

IV. Results

To analyze the transcriptional dataset, the proposed workflow was applied; this workflow can accept raw transcriptional data files and ultimately generates outputs of differentially regulated genes, relevant metabolic pathways and transcriptional regulators, and even potentially relevant published material. This has many advantages compared to standard transcript profiling analyses. From a user aspect, it will be quicker than the time-consuming analysis that currently occurs, and ensures that the same stringency and statistical methods are used in all analyses, and hence should make analyses more user-independent. It can also remove any possibility that users can subconsciously 'manipulate' the data. In order to run the work flow following parameters have been set: NormalizationMethod = rma, Statistical testMethod = limma, p-value = 0.05, foldChange = 1, geneNumber = 100.

The workflow results directly progress from a microarray file to outputs in the form of plots or text files in the case published abstracts from the PubMed database (Figure 2). Tables displaying the processed data can also be visualised From the outputs, the relevance of the transcriptional regulators FNR, ArcA, PdhR were immediately noticeable.



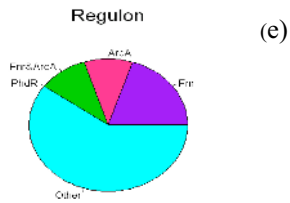
(a) (b) (c)

Gene.ID	logFC	AveExpr	t-value	P.Value	adj.P.Val	B
ybfD	-6.4251	12.53531	-20.9753	8.60E-20	1.59E-16	34.97318
emrY	-6.34166	9.717566	-22.7469	8.21E-21	3.78E-17	37.19892
bglG	-5.74389	11.21273	-20.1239	2.83E-19	1.84E-16	33.83456
b0309	-5.53166	11.29123	-18.5888	2.72E-18	3.62E-16	31.65645
hemA	-5.52479	11.21758	-18.469	3.27E-18	3.62E-16	31.47933
ydcC	-5.48627	12.03441	-19.6757	5.40E-19	1.91E-16	33.21571
ykgH	-5.44609	11.20526	-17.5063	1.48E-17	5.91E-16	30.01513
melB	-5.3852	11.64157	-19.8371	4.27E-19	1.84E-16	33.44017
insB_2	-5.3752	13.0089	-20.8395	1.04E-19	1.59E-16	34.79474

(d)

Query >> Two-component system AND "Metabolic Networks and Pathways"[MeSH Terms] AND "Escherichia coli"
•17616594 2007 AtoC has a dual function as both an antizyme, the posttranslational inhibitor .....
• 16962353 2006 Escherichia coli has several elaborate sensing mechanisms for response to the availability of oxygen and the

(e)



(f)

Figure 2 – Workflow outputs: The workflow produced several at the end of each stage. These show (left-right, top-bottom): (a) raw data image; (b) plotting a MA plot after normalization; (c) box-plotting the summary data pre and post normalisation; (d) filtered and sorted list of differentially expressed genes. (e) tiltle list of relevant paper from; (f) regulon pie chart of Fnr and Arc;

V. Discussion

This workflow has successfully been used to interrogate a transcriptomic dataset and identify regulators and pathways of relevance. This has been demonstrated using experimental conditions in which the major regulators are known; however, the study of less characterised experiments the resulting outputs may have exciting and unanticipated results. From a knowledge point-of-view, an investigators experience is usually over a limited research theme; hence only regulators and pathways already well known to a researcher tend to be examined in detail. This workflow allows easy interrogation of a dataset to identify the role of potentially every known *E. coli* transcriptional regulator or metabolic pathway. This can suggest relevant transcriptional networks and unexpected aspects of physiology that would otherwise have been missed by conventional analysis methods.

Acknowledgment

We thank SUMO team for very useful discussions and SysMO and BBSRC for finical support.

References

1. Neidhardt, F. C. (Ed. in Chief), R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (eds). (1996). ***Escherichia coli* and *Salmonella:* Cellular and Molecular Biology**. American Society for Microbiology. 2 vols. 2898 pages.
2. Rychlik, I. & Barrow, P.A. (2005) *Salmonella* stress management and its relevance to behaviour during intestinal colonisation and infection *FEMS Microbiology reviews* **29** (5) 1021-1040.
3. Fisher, P., Hedeler, C., Wolstencroft, K., Hulme, H., Noyes, H., Kemp, S., Stevens, R., Brass, A. A systematic strategy for large-scale analysis of genotype-phenotype correlations: identification of candidate genes involved in African Trypanosomiasis *Nucleic Acids Research* 2007 35(16): 5625-5633
4. Taverna [http://taverna.sourceforge.net]
5. **GenesSpring** [http://www.chem.agilent.com]
6. **ArrayAssist** [http://www.stratagene.com]
7. **Bioconductor** [http://www.bioconductor.org]
8. Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M. and Karp, P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. Nucleic Acids Research **33** D334-D337.
9. Gama-Castro, S., Jiménez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Peñaloza-Spinola, M.I., Contreras-Moreira, B., Segura-Salazar, J., Muñiz-Rascado, L., Martínez-Flores, I., Salgado, H., Bonavides-Martínez, C., Abreu-Goodger, C., Rodríguez-Penagos, C., Miranda-Ríos, J., Morett, E., Merino, E., Huerta, A.M., Treviño-Quintanilla, L. and Collado-Vides, J. (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Research*. 36 D120-D124
10. Hobman, J.L., Penn, C.W. and Pallen, M.J. (2007) Laboratory strains of *Escherichia coli*: model citizens or deceitful delinquents growing old disgracefully? *Molecular Microbiology* **64** (4) 881-885.
11. Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12 *Science* **277**(5331) 1453-1474.
12. Evans, C.G.T., Herbert, D. and Tempest, D.W. (1970) The continuous culture of microorganisms. 2. Construction of a chemostat. In: Norris JR, Ribbons DW (eds) Methods in microbiology, vol 2. Academic Press, London New York, pp 277–327.
13. Eriksson, S., Lucchini, S., Thompson, A., Rhen, M. and Hinton, J.C. (2003) Unravelling the biology of macrophage infection by gene expression profiling of intracellular Salmonella enterica. *Molecular Microbiology* **47** (1) 103-118.