COMPUTING SCIENCE

BacillOndex: An Integrated Data Resource for Systems and Synthetic Biology

Goksel Misirli, Jennifer S. Hallinan, Matthew Pocock, Simon J. Cockell, Jochen Weile and Anil Wipat

TECHNICAL REPORT SERIES

No. CS-TR-1237 **February 2011**

TECHNICAL REPORT SERIES

No. CS-TR-1237 February, 2011

BacillOndex: An Integrated Data Resource for Systems and Synthetic Biology

G. Misirli, J.S. Hallinan, M. Pocock, S. J. Cockell, J. Weile and A. Wipat

Abstract

BacillOndex is a semantically-rich, integrated knowledge base for Bacillus subtilis. The system comprises a software plug-in for the Ondex system, allowing a user to mine a variety of Bacillus subtilis data sources, together with the resulting integrated dataset that contains data about genes, gene products and their interactions. The data are presented in a computationally amenable format, and can be visualized using the data integration tool Ondex.

Availability and Implementation: The BacillOndex dataset, the Ondex plug-in and the workflow files are freely available at
http://research.ncl.ac.uk/synthetic_biology/downloads.html

Printed and published by Newcastle University,
Computing Science, Claremont Tower, Claremont Road,
Newcastle upon Tyne, NE1 7RU, England.

Bibliographical details

MISIRLI, G., HALLINAN, J.S., POCOCK, M., COCKELL, S.J., WEILE, J., WIPAT, A.

BacillOndex: An Integrated Data Resource for Systems and Synthetic Biology
[By] G. Misirli, J.S. Hallinan, M. Pocock, S.J. Cockell, J. Weile and A. Wipat

Newcastle upon Tyne: Newcastle University: Computing Science, 2011.

(Newcastle University, Computing Science, Technical Report Series, No. CS-TR-1237)

Added entries

NEWCASTLE UNIVERSITY
Computing Science. Technical Report Series. CS-TR-1237

Abstract

BacillOndex is a semantically-rich, integrated knowledge base for Bacillus subtilis. The system comprises a software plug-in for the Ondex system, allowing a user to mine a variety of Bacillus subtilis data sources, together with the resulting integrated dataset that contains data about genes, gene products and their interactions. The data are presented in a computationally amenable format, and can be visualized using the data integration tool Ondex.
Availability and Implementation: The BacillOndex dataset, the Ondex plug-in and the workflow files are freely available at http://research.ncl.ac.uk/synthetic_biology/downloads.html.

About the authors

Goksel Misirli
Newcastle University, School of Computing Science Newcastle upon Tyne, United Kingdom
Email: goksel.misirli@newcastle.ac.uk

Jennifer Hallinan
Newcastle University, School of Computing Science Newcastle upon Tyne, Tyne and Wear, United Kingdom
Email: j.s.hallinan@newcastle.ac.uk

Matthew Pocock
Newcastle University, School of Computing Science Newcastle upon Tyne, United Kingdom
Email: turingatemyhamster@gmail.com

Simon Cockell
Newcastle University, Bioinformatics Support Unit Newcastle upon Tyne, United Kingdom
Email: s.j.cockell@ncl.ac.uk

Jochen Weile
Newcastle University, School of Computing Science Newcastle upon Tyne, United Kingdom
Email: j.weile@ncl.ac.uk

Anil Wipat
Newcastle University, School of Computing Science Newcastle upon Tyne, Tyne and Wear, United Kingdom
Email: Anil.Wipat@ncl.ac.uk

Suggested keywords

DATA INTEGRATION
DATABASES
SYNTHETIC BIOLOGY
SYSTEMS BIOLOGY
VISUALIZATION
BIOLOGICAL NETWORKS

BacillOndex: An Integrated Data Resource for Systems and Synthetic Biology

Goksel Misirli[1], Jennifer S. Hallinan[1], Matthew Pocock[1], Simon J. Cockell[2], Jochen Weile[1] and Anil Wipat[1,2*]

[1]School of Computing Science, Newcastle University, Newcastle upon Tyne, UK

[2]Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne, UK

ABSTRACT

Summary: BacillOndex is a semantically-rich, integrated knowledge base for *Bacillus subtilis*. The system comprises a software plugin for the Ondex system, allowing a user to mine a variety of *Bacillus subtilis* data sources, together with the resulting integrated dataset that contains data about genes, gene products and their interactions. The data are presented in a computationally amenable format, and can be visualized using the data integration tool Ondex.

Availability and Implementation: The BacillOndex dataset, the Ondex plugin and the workflow files are freely available at http://www.bacillondex.org.

Contact: anil.wipat@ncl.ac.uk

1 INTRODUCTION

Over the past several decades high-throughput technologies have produced large amounts of data in the biological sciences. Much of these data have been made freely available. Such data are an invaluable resource for biologists and bioinformaticists. However, the sheer number of different data sources, together with the fact that each database may use different file formats, data organization and access procedures means that gaining access to all of the known information about any given organism is likely to be a time-consuming, error-prone and frustrating procedure.

Bacillus subtilis is a widely-studied model prokaryote. *B. subtilis* and closely related species play a major role in problems ranging from soil ecology to human disease and industrial use. *Bacillus* species have been exploited for wide variety of industrial applications (Schallmey, Singh & Ward, 2004). This organism is therefore of considerable interest to the research and biotechnology communities. However, data relating to this organism is spread amongst a variety of sources. As well as the scientific literature, there are numerous databases holding *B. subtilis*-specific information.

To overcome the problem of manually interrogating and downloading data from diverse data sources data integration is essential to systems and synthetic biologists. Relevant data can be identified using computational data integration algorithms, downloaded, and combined into a format which can be either manually browsed or computationally analysed. Integrated data is most commonly represented as a network, in which nodes represent either genes or gene products, while edges represent some form of interaction between the nodes. Such interactions may be physical, such as protein-protein bindings; genetic, such as synthetic lethal interactions; or more indirect, such as co-citation in a publication. Biological interaction networks may be enriched with semantic annotations. Both nodes and edges may carry metadata about the type of entity represented: a node of type *protein* may be related to a node of type *gene* by the relationship *is_product of*. The addition of semantic metadata to a network permits automated reasoning over the system, and potentially the discovery of new knowledge.

One computational tool for the integration and analysis of semantically-enriched networks is Ondex[1] (Kohler et al., 2006). Ondex uses dataset-specific parsers to combine a wide range of data into an integrated, semantically annotated network of concepts and relations. Concepts can represent any object of interest in the dataset, not just genes and proteins. Relations between concepts carry annotations based upon an underlying ontology, making an Ondex graph amenable to computational reasoning. Complex workflows may be executed automatically over an Ondex graph, significantly reducing the time and effort needed to re-analyze data or apply an existing analysis to new data.

Ondex networks have been developed for several different organisms including the yeast *Saccharomyces cerevisiae* and the model plant *Arabidopsis thaliana*. A human Ondex network has been used to identify potential drug repositioning candidates (Cockell et al., 2010). However, Ondex networks have not yet been constructed for microorganisms, and because microbial data is often stored in different databases from eukaryotic data, existing parsers are not adequate to quickly produce integrated microbial networks.

Here we describe the development of an Ondex network for the prokaryote *B. subtilis*. The Ondex knowledgebase was constructed using data from a range of sources (Table 1).

BacilluScope was used to obtain sequence and functional annotation data for all *B. subtilis* genes and gene products. DBTBS provided the gene regulatory network, including details of operon membership, promoters and their associated sigma factors. Information on physical protein-protein interactions was taken from STRING, including scores for functional interactions derived from techniques such as protein neighbourhood, protein fusion, co-occurrence and co-expression. The data from KEGG was used to create concepts such as Protein Complex, Enzyme, Compound and Reaction. Start and end positions on the genome were recorded for nucleotide features.

*To whom correspondence should be addressed.

[1] http://www.ondex.org/

Table 1. Data sources used to construct BacillOndex

Source	Data type	Reference
BacilluScope[1]	Sequence, annotations	Barbe et al. (2009)
KEGG[2]	Metabolic pathways	Kanehisa et al. (2010)
DBTBS[3]	Transcription factor binding	Sierro et al. (2008)
STRING[4]	Protein interactions	von Mering et al. (2007)
KEGG Expresssion[5]	Microarray	Goto et al. (2000)
Gene Ontology[6]	Annotations	The Gene Ontology Consortium (2000)

[1] https://www.genoscope.cns.fr/agc/mage/wwwpkgdb/MageHome
[2] http://www.genome.jp/kegg/
[3] http://dbtbs.hgc.jp/
[4] http://string-db.org
[5] http://www.genome.jp/kegg/expression/
[6] http://www.geneontology.org/

B. subtilis microarray data from the KEGG EXPRESSION database was used to find the minimum and maximum strength of gene expression for each gene over all expression datasets, in order to assess the relative strengths of different promoters. Expression values were normalized using an algorithm developed by Dawes & Glassey (2007). *B. subtilis*-specific GO annotations were extracted from unfiltered UniProt GO annotations. GO terms were downloaded in OBO format (Smith et al., 2007) and only the terms with direct relation to proteins were kept in the final dataset.

New parsers were implemented to convert data from BacilluScope, DBTBS, STRING and KEGG EXPRESSION. For GO terms and annotations existing parsers were used. The Ondex file for KEGG was downloaded from the Ondex Web site. The new concepts were given the "user friendly" names from BacilluScope as preferred names. Concepts and relations were linked to literature and public databases using the appropriate accession numbers. Following integration the Ondex network was searched for motifs representing positive and negative auto-regulation. Concepts for feed-forward loops representing these interactions were added to the knowledgebase, along with links to the participating genes.

2 RESULTS

We produced an integrated knowledge base for *B. subtilis* from a range of sources and parsers that allow the network to be easily rebuilt and kept up to date. The knowledge base combines genome annotations with data about the genetic regulatory network, biochemical reactions, microarray experiments and protein-protein interactions. The Ondex network contains a number of different concepts: Coding sequence (CDS), Protein, Transcription Factor, Operon, Operator, Promoter, Terminator, RNA, Enzyme, Enzyme Classification, Reaction, Pathway, Compound, COG Class, COG Class Category, Cellular Component, Molecular Function, Biological Process, KEGG Orthologs Enzyme, KEGG Orthologs Gene, Protein Complex, KEGG Orthologs Protein, Feed Forward Loop, Microarray Experiment, Ribosome Binding Site (RBS) and Spacer sequence (Shim). The knowledge base is in the form of an XML file, which can be imported into Ondex. The dataset contains 33,043 concepts and 94,774 relations. We also provide the workflows and relevant parsers to perform the integration, in the form of an Ondex plugin.

BacillOndex will facilitate the accession, visualisation, analysis and exchange of data by the *B. subtilis* research community, and forms the basis for the production of integrated knowledge bases for other microorganisms.

ACKNOWLEDGEMENTS

Funding: Research Councils UK (to J.S.H.); Engineering and Physical Sciences Research Council/National Science Foundation grant number: EP/H019162/1 (to G.M.); Biotechnology and Biological Sciences Research Council Systems Approaches to Biological Research initiative grant number: BB/F006039/1 (to J.W.).

We acknowledge the Ondex development team for their help.

Conflict of Interest: none declared.

REFERENCES

Barbe, V., Cruveiller, S. Kunst, F., Lenoble, P., Meurice, G., Sekowska, A., Vallenet, D., Wang, T. Moszer, I., Médigue, C. & Danchin, A. (2009). From a consortium sequence to a unified sequence: The *Bacillus subtilis* 168 reference genome a decade later. *Microbiology* 155: 1758 - 1775.

Cockell, S. J., Weile, J., Lord, P., Wipat, C., Andriychenko, D., Pocock, M., Wilkinson, D., Young, M. & Wipat, A. (2010). An integrated dataset for in silico drug discovery. *Journal of Integrative Bioinformatics* 7(3): 116.

Dawes, N. L. & Glassey, J. (2007). Normalization of multicondition cDNA microarray data. *Comparative and Functional Genomics*: Article ID 90578 doi:10.1155/2007/90578.

Goto, S., Kawashima, S., Okuji, Y., Kamiya, T., Miyazaki, S., Numata, Y. & Kanehisa, M. (2000). KEGG/EXPRESSION: A database for browsing and analysing microarray expression data. *Genome Informatics* 11: 222 - 223.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research* 38: D355 - D360.

Kohler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Ruegg, A., Rawlings, C., Verrier, P. & Philippi, S. (2006). Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics* 22(11): 1383 - 1390.

Schallmey, M., Singh, A. & Ward, O. P. (2004). Developments in the use of *Bacillus* species for industrial production. *Canadian Journal of Microbiology* 50(1): 1 - 17.

Sierro, N., Makita, Y., de Hoon, M. J. L. & Nakai, K. (2008). DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Research* 36: D93 - D96.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Cuesters, W. Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., The OBI Consortium, Leontis, N., Rocca-Serra, N. P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L. & Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 25: 1251 - 1255.

The Gene Ontology Consortium (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics* 25: 25 - 29.

von Mering, C., Jensen, L. J., Kuhn, M., Chaffron, S., Kruger, B., Snel, B. & Bork, P. (2007). STRING 7 - Recent developments in the integration and prediction of protein interactions. *Nucleic Acids Research* 35(1): D358 - D362.