Conference Review

From gene regulation to gene function: regulatory networks in Bacillus subtilis

Colin R Harwood1* and Ivan Moszer2

1Department of Microbiology and Immunology, University of Newcastle upon Tyne, Framlington Place, Newcastle upon Tyne, NE2 4HH, UK
2Unité de Génétique des Génomes Bactériens, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

*Correspondence to:
Department of Microbiology and Immunology, The Medical School, University of Newcastle upon Tyne, Framlington Place, Newcastle upon Tyne, NE2 4HH, UK.
E-mail: Colin.Harwood@ncl.ac.uk

Received: 5 December 2001
Accepted: 6 December 2001
Published online: 20 December 2001

Abstract

Bacillus subtilis is a sporulating Gram-positive bacterium that lives primarily in the soil and associated water sources. The publication of the B. subtilis genome sequence and subsequent systematic functional analysis and gene regulation programmes, together with an extensive understanding of its biochemistry and physiology, makes this micro-organism a prime candidate in which to model regulatory networks in silico. In this paper we discuss combined molecular biological and bioinformatical approaches that are being developed to model this organism’s responses to changes in its environment. Copyright © 2001 John Wiley & Sons, Ltd.

Keywords: DNA arrays; functional analysis; gene expression; reporter genes; database conception; database integration

Introduction

Bacillus subtilis has been studied extensively over more than fifty years and is widely regarded as a model for the analysis of the medically, environmentally and industrially important Gram-positive bacteria. Knowledge of its biochemistry and physiology is extensive and B. subtilis strain 168 is highly amenable to genetic manipulation. B. subtilis and other Bacillus species are used in a wide range of industrial processes, for example for the production of extracellular enzymes, vitamins and fine biochemicals [12], and this industrial use has also enhanced our knowledge of the molecular and physiological characteristics of this bacterium.

Post-genomic analyses

The genome sequence of B. subtilis was completed in 1997 [17] by a joint European/Japanese consortium. The genome is 4.2 Mbp in length and encodes approximately 4100 proteins. The completion of the genome sequence was followed by systematic programmes to elucidate the function of genes currently of unknown function and to study genome-wide gene expression. These collaborative programmes are designed to expand knowledge of the molecular biology of strain 168 and, ultimately, to construct a mechanistic model of its behaviour in laboratory-based studies. Data resulting from these programmes have been compiled into databases that are accessible over the Internet: SubtiList (http://genolist.pasteur.fr/SubtiList/), a dedicated DNA sequence database; Micado (http://locus.jouy.inra.fr/cgi-bin/gemnic/madbase/progs/madbase.operl/), which has data on the characterisation of a collection of isogenic mutants; and Sub2D (http://microbio2.biologie.uni-greifswald.de:8880/), that holds data on the analysis of the B. subtilis 168 proteome. A new database, SubScript, for the storage and analysis of transcriptomic data, is currently under construction (see below).

Data from the genome and from the systematic programmes outlined above suggest that a significant proportion of the B. subtilis genome is dedicated to growth and survival in the extremely variable conditions found in the soil. Moreover, it has been proposed that a significant proportion of genes of unknown function (approximately 1800 or 40%) are required for survival. Thus, knowledge of the behaviour of B. subtilis in response to specific
changes in its environment is likely to be of value for elucidating the role of genes of unknown function. To this end, proteomic and transcriptomic methodologies (eg. 2D-PAGE, DNA arrays, yeast two-hybrid analyses) are being directed towards defining regulons, and characterising their cognate regulatory proteins. A particularly valuable resource for the work is the existence of a collection of isogenic mutants (BFA mutants) in which the pMUTin integration vector has been introduced into virtually all genes on the B. subtilis chromosome [23]. pMUTin not only generates target gene mutants, but also links a lacZ transcriptional reporter to the target gene and a controllable promoter to any genes downstream and in the same operon. Interacting European (BACELL Network: http://www.ncl.ac.uk/bacellnet/) and Japanese (JAFAN: http://bacillus.genome.ad.jp/BSORF-DB. html) research consortia are using these mutants, in combination with post-genomic technologies such as DNA array and proteomics, to analyse B. subtilis gene function and regulatory networks in detail.

The response of B. subtilis to phosphate starvation illustrates the need to develop in silico techniques for modelling regulatory pathways. During phosphate starvation, B. subtilis responds by activating the phosphate stimulus, comprising, at least, the Pho and the σ^B^-dependent general stress regulons [3,13,15,18]. Induction or repression of genes of the Pho regulon enables the cell to use limited phosphate resources more efficiently and to increase accessibility to alternative sources of phosphate. The Pho regulon is controlled by the interaction of at least 3 two-component signal transduction systems [15]. The centre of this regulatory network is the PhoP-PhoR sensor-regulator system [16]. Activation of the response regulator PhoP to PhoP~P during phosphate starvation leads to the induction or repression of genes in the Pho regulon. ResE, involved in the induction of genes during anaerobiosis, is required for the full induction of the Pho regulon, while a third response regulator, Spo0A, terminates the phosphate response, and initiates sporulation if phosphate starvation conditions persist. In addition to the phosphate starvation-specific Pho regulon, phosphate limitation also induces genes of the σ^G^-dependent general stress (σ^B^-GS) regulon [3]. The σ^B^-GS regulon provides a non-specific resistance to stress, protecting DNA, membranes and proteins from the damage [1,11,13]. The activity of σ^B-, an alternative sigma factor modulating promoter choice, is controlled by the anti-sigma factor RsbW. In unstressed cells, when the antagonist protein RsbV is phosphorylated (RsbV~P), RsbW can bind to and inactivate σ^B_ [5]. In response to stress, RsbV is dephosphorylated [5,25,28], allowing it to sequestrate RsbW. As a result, σ^B is free to activate genes in the σ^B-GS regulon [10]. Complex regulatory cascades control the phosphorylation state of RsbV [1,2,24,26,28].

Using a combination of reporter gene, DNA array and 2D-PAGE technologies, we have monitored genome-wide responses to phosphate starvation. phoR- and sigB-null mutants have been used to assign genes to the Pho and the σ^B-GS regulons, and to determine the effects of σ^B_ on the expression of the Pho regulon and vice versa. The data indicated that the Pho and the σ^B-GS regulons are interacting members of the phosphate stimulon (Prágai and Harwood, unpublished), although we do not, as yet, understand the molecular mechanisms underlying these interactions.

To date the analysis of the phosphate stimulon has been carried out by a hypothesis-driven approach. However, the extensive existing knowledge of this system, combined with high throughput data-rich transcriptomic and proteomic methodologies means that data-driven approaches can now be considered. To this end, we are attempting to use the signal transduction (including phosphorylation states and protein-protein interactions) and gene expression data to model the regulatory interactions within the phosphate starvation stimulon. The core of the current model is based around a gene expression weight matrix following the TreMM modelling strategy [27].

**Database integration**

The wealth of information generated by large-scale genomic studies has to be collected and organized in specialised data structures. To this end, a number of dedicated databases have been constructed over the ten past years of B. subtilis genomics. Firstly the SubtiList database holds the core genomic data [20]. Developed in the framework of the genome sequencing project [17], SubtiList is the reference database for the B. subtilis 168 genome. It provides a curated dataset of DNA and protein sequences, combined with the relevant annotations and functional assignments. Information about gene functions and products is continuously updated by linking relevant
From gene regulation to gene function

Bibliographic references. Recently, sequence corrections arising from both systematic verifications and submissions by individual scientists were included in the reference genome sequence [19]. SubtiList is based on a generic relational data schema and World-Wide Web interface developed for the handling of bacterial genomes, called Genolist, which is used for a number of other organisms (http://genolist.pasteur.fr/). A user-friendly WWW interface was designed to allow users to browse easily through genome data and retrieve information according to common biological queries. SubtiList also provides more elaborate tools, such as pattern searching, which are tightly connected to the overall browsing system.

The second database, Micado [8,22], was developed for handling phenotypic data generated in the framework of the B. subtilis functional analysis programme [21]. Specific data structures and displays were developed both for the gel images used for authenticating the BFA mutant constructions, and for the graphs showing the growth and reporter gene activities of the mutants. Transcript and disruption maps are represented through generic physical maps, which are also used for navigating the complete genome. The Micado WWW interface was recently rewritten and its relational data model was ported on a new database management system. A comparable database is managed by JAFAN (http://bacillus.genome.ad.jp/BSORF-DB.html).

The third database dedicated to B. subtilis, Sub2D [6,7], was developed to analyse two-dimensional protein gels, with the aim of deciphering new regulons and stimulons. Sub2D was structured according to the guidelines established for the construction of two-dimensional protein databases [4]. It contains information about the methods of protein identification and the environmental conditions that influence the production of a given protein. Beside basic search forms and gene lists, proteins can be selected on the basis on their position on two-dimensional gels. Complex queries allow for the display of all the proteins similarly regulated by a given stimulus or controlled by a single transcriptional regulator.

More recently, two additional databases for B. subtilis genomics have been established in the framework of the current BACCELL Network programme. Firstly, data generated by transcriptome analysis are organized into the SubScript database, which is still under active development. The objectives of SubScript are to provide both a data repository that can be browsed and queried via the Web, and a tool for the analysis of B. subtilis expression data, with the help of standard and original analysis methods. An elaborate data schema, comprising more than fifty classes, has been based on the recommendations published by the Microarray Gene Expression Databases (MGED) consortium (Minimal Information About Microarray Experiments, MIAME) [9], and the corresponding ArrayExpress database model (EBI: http://www.ebi.ac.uk/arrayexpress/). The original version of the schema has been superseded by a significantly improved version that is more appropriate to the precise requirements of the BACCELL Network programme. This newer version was conceived to store experiment parameters and the raw data, as well as their subsequent analyses. A great deal of flexibility in the way experiments can be described has been anticipated in the data model. Specialised interfaces are currently being developed to allow users to enter, retrieve, and analyse expression data.

Finally, the newly developed SPID database contains information about two-hybrid protein interactions [14]. Graphic displays accessible through the WWW facilitate data mining and navigation through the interaction networks and the domains implied in protein pair interactions.

The five major databases on B. subtilis genomics described above generate an obvious requirement for better communication between their respective datasets. Indeed each of them is aimed to become a component of a comprehensive genomic resource for B. subtilis, gathering disparate classes of biological knowledge. This should allow investigators to pose complex questions across the various types of stored information. For example; which genes located close each other (SubtiList), are involved in sporulation (Micado), are co-activated in the presence of glucose in the medium (Sub2D and SubScript), and participate to a single protein complex (SPID)? Such enquiries require a level of data integration that has not currently been achieved, but which is under active investigation.

In the first instance, simple hypertext links through the WWW are being reinforced between the many B. subtilis databases. The use of invariant accession numbers greatly facilitates the definition of cross-references. This does not yet allow questions to be asked across the various data sets, but does reflect progress towards a greater degree of interaction between the databases that does not
require challenging computing techniques: this approach is considered to be the most basic stage of database interoperation. At the opposite end of the data integration ladder is the data warehousing strategy. This approach implies that data schemas can be merged and that common software and technical implementation can be defined. The physical integration of heterogeneous datasets into one single repository, whilst probably the most efficient way of dealing with the complexities of elaborate queries, presents a number of drawbacks, including the need for regular data synchronisation. Universal data description and formalisation languages, such as XML, will help to reach this goal by facilitating exchange of structured data. As for the *B. subtilis* databases, a first level of data warehousing is being implemented naturally as two pairs of databases are being developed in the same locations (SubtiList/SubScript and Micado/SPID). Finally, as an intermediate solution, a truly dynamic database interoperability can be set up, using concepts such as those implemented in the CORBA protocols. This is an attractive solution as it allows the user to perceive one single and unified (virtual) data schema and to ask complex questions on it, while retaining the distributed data management. However, a number of limitations make this technology difficult to apply in practice, including implementation pitfalls and performance bottle-necks not present in the data warehousing approach (notably because of the network traffic). The CORBA architecture solution is being assessed within the small and recent SPID database.

Progress in defining shared knowledge representations, including the use of common nomenclatures, still has to occur in order to break some semantic barriers towards the definition of unified data models, allowing the use of any combination of selection criteria for retrieving information. The development of ontologies for some common and formal descriptions of biological concepts and entities may help to take this route.

Acknowledgement

This work was supported by grants from the European Commission’s Quality of Life programme (contract number QLG2-CT-1999-01455), the UK Biotechnology and Biological Sciences Research Council and the French Centre National de la Recherche Scientifique. Studies on the phosphate stimulus are being carried out by Zoltán Prágai and Nicola O’Conner, and the *in silico* modelling by Nicholas Allenby, Alan C. Ward and Anil Wipat. The development of the SubScript database is being performed by Sandrine Moreira.

References
