BacillOndex: An Integrated Data Resource for Systems and Synthetic Biology

Goksel Misirli[1], Anil Wipat[1], Joseph Mullen[1], Katherine James[1], Matthew Pocock[1,3], Wendy Smith[1], Nick Allenby[2] and Jennifer S. Hallinan[1,*]

[1]School of Computing Science, Newcastle University, Newcastle upon Tyne, UK, NE1 7RU

[2]Demuris Ltd., Newcastle upon Tyne, UK, NE2 4HH

[3]TAMH Ltd., Newcastle upon Tyne

Summary

BacillOndex is an extension of the Ondex data integration system, providing a semantically annotated, integrated knowledge base for the model Gram-positive bacterium *Bacillus subtilis*. This application allows a user to mine a variety of *B. subtilis* data sources, and analyse the resulting integrated dataset, which contains data about genes, gene products and their interactions. The data can be analysed either manually, by browsing using Ondex, or computationally via a Web services interface. We describe the process of creating a BacillOndex instance, and describe the use of the system for the analysis of single nucleotide polymorphisms in *B. subtilis* Marburg. The Marburg strain is the progenitor of the widely-used laboratory strain *B. subtilis* 168. We identified 27 SNPs with predictable phenotypic effects, including genetic traits for known phenotypes. We conclude that BacillOndex is a valuable tool for the systems-level investigation of, and hypothesis generation about, this important biotechnology workhorse. Such understanding contributes to our ability to construct synthetic genetic circuits in this organism.

1 Introduction

1.1 Data Integration and Synthetic Biology

The aim of synthetic biology is to produce organisms with novel, desirable biological functionality, either by re-engineering existing systems, or by designing new organisms from scratch. At the moment, the re-design of existing systems is the most promising approach, although whole-genome design is becoming more feasible [1]. When working with existing organisms it makes sense to integrate and incorporate as much existing information as possible. However, these data are spread amongst a variety of sources, including the scientific literature and numerous online databases [2]. Drawing together these data means that synthetic biologists can use the wealth of data available to inform the design of new genetic circuits.

Integrated data are often represented as networks, 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 binding; genetic, such as synthetic lethal interactions; or more indirect, such as co-citation in a publication [3, 4]. Networks are easy to visualise and browse, and are thus particularly valuable for the exploratory analysis of

* To whom correspondence should be addressed. Email: jennifer.hallinan@ncl.ac.uk

doi:10.2390/biecoll-jib-2013-224

complex datasets. Network analysis is particularly valuable for users who have an interest in a specific biological question, but who lack the time or training to investigate individual data sources exhaustively. Integrated datasets also provide synergistic views of disparate data sources, putting a diverse range of biological interactions into a systems-level context. Integrated networks lend themselves well to manual browsing, and can also be made computationally accessible, using technology such as Web services [5].

Humans bring large amounts of background domain knowledge to the manual analysis of integrated datasets. In order to facilitate computational analysis of such data, including automated reasoning and hypothesis generation [6], the addition of semantic annotations to integrated datasets is particularly valuable. Such annotations are generated from the knowledge captured in the literature, or stored in databases, converted to a standard, unambiguous format. Ideally, this data-representation format should be built upon an ontology: a standardised model of the way in which entities interact. Stevens and colleagues [7] define a biomedical ontology as including, at a minimum, a "vocabulary of terms, and some specification of their meaning". Using an ontology, nodes and edges in a network may carry metadata about the type of entity represented: for example, a node of type `protein` may be related to a node of type `gene` by the relationship `is_product_of`.

1.2 The Ondex Data Integration System

One computational tool for the integration and analysis of semantically-enriched networks is Ondex [8]. Ondex uses purpose-written parsers to extract data from diverse sources and integrate them into a network annotated with semantic metadata. Issues such as reformatting data and matching up the disparate identifiers used by different databases are handled by the system, and users are presented with a network in which nodes may represent any type of *concept* (proteins, genes, publications, protein families, and so on) interacting via *relations*. Both concepts and relations carry semantic *attributes*. All concepts, relations and attributes have 'types', which are organised hierarchically. For example, the concept type `Protein` is a subtype of `Molecule`, which is itself a subtype of `Thing`. This hierarchy means that every `Protein` concept is also a `Molecule` and a `Thing`. Similarly, the relation type `catalyzes` is a subtype of `actively participates in`. Therefore, every statement that `p::Protein catalyzes r::Reaction` means that `p` actively participates in `r` [9].

Ondex networks have been developed for several different organisms, including the yeast *Saccharomyces cerevisiae* [9] and the model plant *Arabidopsis thaliana* [10]. A human Ondex network has been used to identify potential drug repositioning candidates [11]. However, Ondex networks have not yet been constructed for microorganisms, and because microbial data are often stored in different databases from eukaryotic data, existing parsers are not adequate to quickly produce integrated microbial networks. We are interested in engineering novel genetic circuits in the bacterium *Bacillus subtilis* [12, 13]. However, despite being a widely-used model organism, this bacterium is in many ways poorly understood. We developed an Ondex-based integrated dataset, BacillOndex, to facilitate the genome-scale analysis of *B. subtilis*.

1.3 Engineering *Bacillus subtilis*

Bacillus subtilis is a model prokaryote. This bacterium is ubiquitous in soil environments, non-pathogenic, and has a number of characteristics which make it valuable for the biotechnology industry, including the ability take up foreign DNA under certain circumstances, a capacity for spore formation, and the ability to secrete a range of proteins

doi:10.2390/biecoll-jib-2013-224

[14]. *B. subtilis* is widely studied, and much information about its genetics and physiology has been generated.

Even for *B. subtilis*, however, many of the details of interactions between physical elements are not fully understood. Identifying interactions and obtaining biochemical parameters for those interactions and corresponding biochemical reactions has been a challenge [15]. However, qualitative models such as graph-based biological networks can capture the relationships between biological concepts and provide insights into cellular systems [16-19]. In simple, manual genetic circuit designs, interactions between individual parts are specified by a domain expert. In order to automate circuit design, this knowledge must be machine accessible.

The most widely-used strain of *B. subtilis* is 168. This strain is used in almost all academic studies, and for a wide range of industrial processes. Strain 168 was generated from the less malleable parent strain, Marburg, via mutagenesis with X-rays, in the 1940s [20], and is popular largely because it is easily transformed with foreign DNA. Strain 168 is a tryptophan auxotroph, and also has a number of other phenotypic differences from the parent strain, including differences in motility, sporulation and cell wall physiology [21]. However, the genetic bases of these phenotypic differences are still not well understood, and the full extent of the phenotypic effects of the original random mutagenesis is not clear.

B. subtilis 168 has one circular chromosome of 4.2 Megabases (Mb) containing 4354 genes (4176 protein coding genes and 178 RNAs) and is believed to have 192 indispensable, as well as 79 essential genes [22]. The genome of the strain was first sequenced in 1997 [23]. The project had contributions from 25 European laboratories, seven Japanese, two US and one Korean laboratory. Bacterial strains tend to evolve rapidly in laboratory environments and re-sequencing of the genome took place in 2009. This project utilized faster, more accurate sequencing techniques and a recently developed high-level annotation platform, MaGe [24]. In stark contrast to the significant effort required only fifteen years ago, it is now possible to sequence a complete bacterial genome for a few hundred pounds in an afternoon [25]. The potential for large-scale analysis of important organisms such as *B. subtilis* is only just about to be realised.

There is considerable current interest in single nucleotide polymorphism (SNP) analysis. A number of stand-alone tools for SNP analysis are available [26-28], and support for SNP analysis is built into free statistical analysis packages such as R[1] and Excel [29], and commercial tools such as Lasergene[2]. Most synthetic biology focuses on the manipulation of large segments of DNA, whole genes and their attendant control sites, and even entire pathways. The ability to rapidly generate insight into the relationship between small mutations—SNPs—and phenotypic effects could allow future synthetic biologists to carry out fine-grained tuning of bacterial chassis.

Here, we describe the development of an integrated, semantically-annotated Ondex network for *B. subtilis*, and the application of this network to the analysis of SNPs generated when *B. subtilis* 168 was generated from its parent strain, Marburg.

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

[2] http://www.dnastar.com/t-sub-nextgen-genome-solutions-snp-analysis.aspx

doi:10.2390/biecoll-jib-2013-224

2 Methods

2.1 Data Sources

Data about the sequence, functional annotation and interactions of all *B. subtilis* genes and their proteins were obtained from a range of databases (Table 1). In addition, *B. subtilis* microarray data from the KEGG EXPRESSION[3] database were parsed to find the normalised minimum and maximum level of gene expression for each gene over all expression datasets, in order to assess the relative strengths of different promoters. BacillOndex includes information about the expression of genes from 79 microarray experiments. In the network, concepts for these coding sequences (CDSs) were linked to promoter concepts. Therefore, the expression levels of CDS concepts can be used to infer the strengths of upstream promoters.

One of the advantages of the Ondex system is its ability to reuse database parsers in order to rapidly build new datasets. Parsers can be time consuming to write, but once written become a resource for the entire community. Constructing a new knowledge base, such as BacillOndex, using existing parsers takes around an hour, whereas querying the knowledge base is almost instantaneous. We used data sources for which parsers had already been written wherever possible. However, it was necessary to write several new parsers for *B. subtilis*-specific datasets. Although we chose to use KEGG EXPRESSION, other relevant data, such as microarray data from the NCBI Gene Expression Omnibus [30], could be used once the appropriate parsers become available.

Table 1: Data sources used to construct BacillOndex.

Source	Data Type	Reference
BacilluScope	Sequence, annotations	[24]
KEGG	Metabolic pathways	[31]
DBTBS	Transcription factor binding	[32]
STRING	Protein interactions	[33]
KEGG Expression	Microarray	[34]
Gene Ontology	Annotations	[35]
UniProt	Protein sequence features	[36]

New parsers were implemented to convert data from the BacilluScope, DBTBS, STRING and KEGG EXPRESSION databases. For Gene Ontology (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 (FFLs) representing these interactions were added to the knowledge base, along with links to the participating genes.

[3] http://www.genome.jp/kegg/expression/

doi:10.2390/biecoll-jib-2013-224

2.2 Data Integration Strategy

Although the Ondex ontology includes a large number of types of biological concepts, the encoding of proteins and RNAs, and transcriptional relationships can be modelled at the gene level using Ondex. However, in synthetic biology it is necessary to work with finer-grained elements of genes, such as promoters and CDSs. Therefore, concept types representing sequence features, including promoters, operators, CDSs, ribosome binding sites (RBSs), shims, terminators and operons, were added to the Ondex ontology in order to construct BacillOndex. In addition, network motifs such as FFLs can be used to construct biological devices, and hence FFL concepts were identified in the network and included in the metadata. GO terms are useful to annotate gene products, the Clusters of Orthologous Groups (COG) numbers can also be used to classify proteins and also find orthologous parts from other species. Therefore, concept types for COG numbers and their categories were also included.

The BacillOndex data model was designed to encompass a wide range of concepts from a variety of source databases (Table 1). Biological concepts such as Protein and CDS are modelled as concept classes as required by the Ondex system. Instances of concept classes inherit the relationships defined in the model. For example, the relationships between proteins and CDSs are represented by `is_encoded_by`, while those between TFs and proteins are annotated as `is_equivalent_to`. Proteins, TFs and enzymes can share properties such as names, but are represented by different concepts. This approach helps to encapsulate role-specific attributes.

The BacillOndex dataset includes the following types of concept:

- CDS, promoter, operator, operon, terminator, shim and RBS concepts derived from sequence-based features;
- Protein, RNA, enzyme, TF and protein complex concepts representing gene products and their aggregates;
- COG class, COG class category, KEGG orthologs enzyme (KOEN), KEGG orthologs gene (KOGE) and KEGG orthologs protein (KOPR) concepts classifying other concepts using orthology terms;
- Cellular component, molecular function, biological process and enzyme classification concepts for location-, function- and biological process-based classifications;
- Reaction, pathway and compound concepts forming the core of the pathways in the network;
- Concepts for FFLs and promoter strength data derived from microarray experiments.

2.3 Systems-level Analysis of Single Nucleotide Polymorphisms

The parent strain, Marburg ATCC 6051, was sequenced by ACGT, Inc.[4] using Illumina[5] HiSeq technology. We obtained 5,092,107 reads, of an average length of 36 bp, providing approximately 40 times coverage. The Marburg sequence was assembled using the Lasergene Core Suite from DNAStar[6]. The assembled sequence was compared with that of *B. subtilis* 168 in order to identify SNPs.

[4] http://www.acgtinc.com/

[5] http://www.illumina.com/index.ilmn

[6] http://www.dnastar.com/

doi:10.2390/biecoll-jib-2013-224

For each strain SNPs were identified using the DNAStar SeqMan Pro[7] tool and a SNP report created, which identified the SNP type of those occurring in CDS: silent SNPs, which do not change the amino acid (sSNPs); non-silent SNPs, which do alter the protein produced (nsSNPs); expression SNPs, occurring in the control regions of genes (eSNPs); or single-base insertions or deletions (INDELs). If the software could not identify the SNP type, it was annotated manually. Manual analysis was also employed to identify the coverage and accuracy of each SNP, and the location of the SNP with respect to the reads; SNPs at read boundaries are less likely to be accurate than those in the middle of a read. BacillOndex was then used to map these SNPs to specific pathways, and infer the possible phenotypic implications of the mutations.

2.3.1 CatSNP

The Cellular Analysis Tool for SNPs (CatSNP) is a SNP filter plug-in for the BacillOndex dataset. It runs in the Ondex Integrator interface, as part of an automated workflow.

CatSNP takes as input the genomic locations, identifies the associated concepts in the BacillOndex dataset, and produces an Ondex subgraph linking genomic location, protein, pathway and other available data.

The CatSNP workflow contains an OXL parser (which takes in the BacillOndex integrated dataset), a SNP parser (which takes in a SNP.txt file containing SNP locations), a filter (which filters SNPs into BacillOndex dataset) and an OXL exporter (Figure 1).

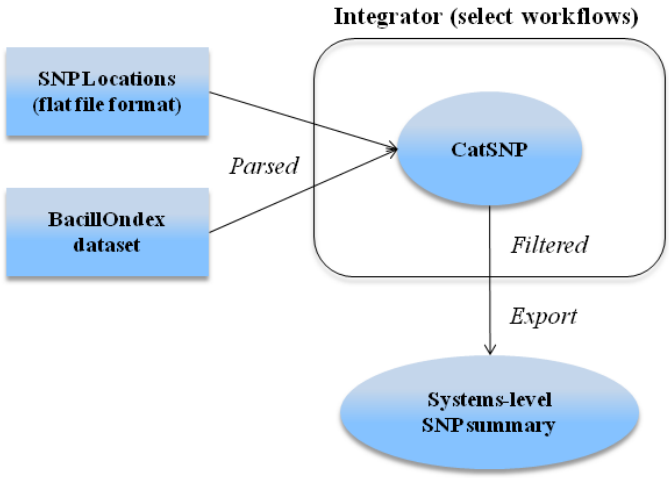

Figure 1. Visualisation of the CatSNP workflow within the Integrator of the Ondex platform.

3 Results

We produced an integrated knowledge base for *B. subtilis* from a range of sources, together with a set of 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

[7] http://www.dnastar.com/t-sub-products-lasergene-seqmanpro.aspx

doi:10.2390/biecoll-jib-2013-224

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. BacillOndex is underpinned by a formal ontology, built as an extension of the standard Ondex ontology (Figure 2).

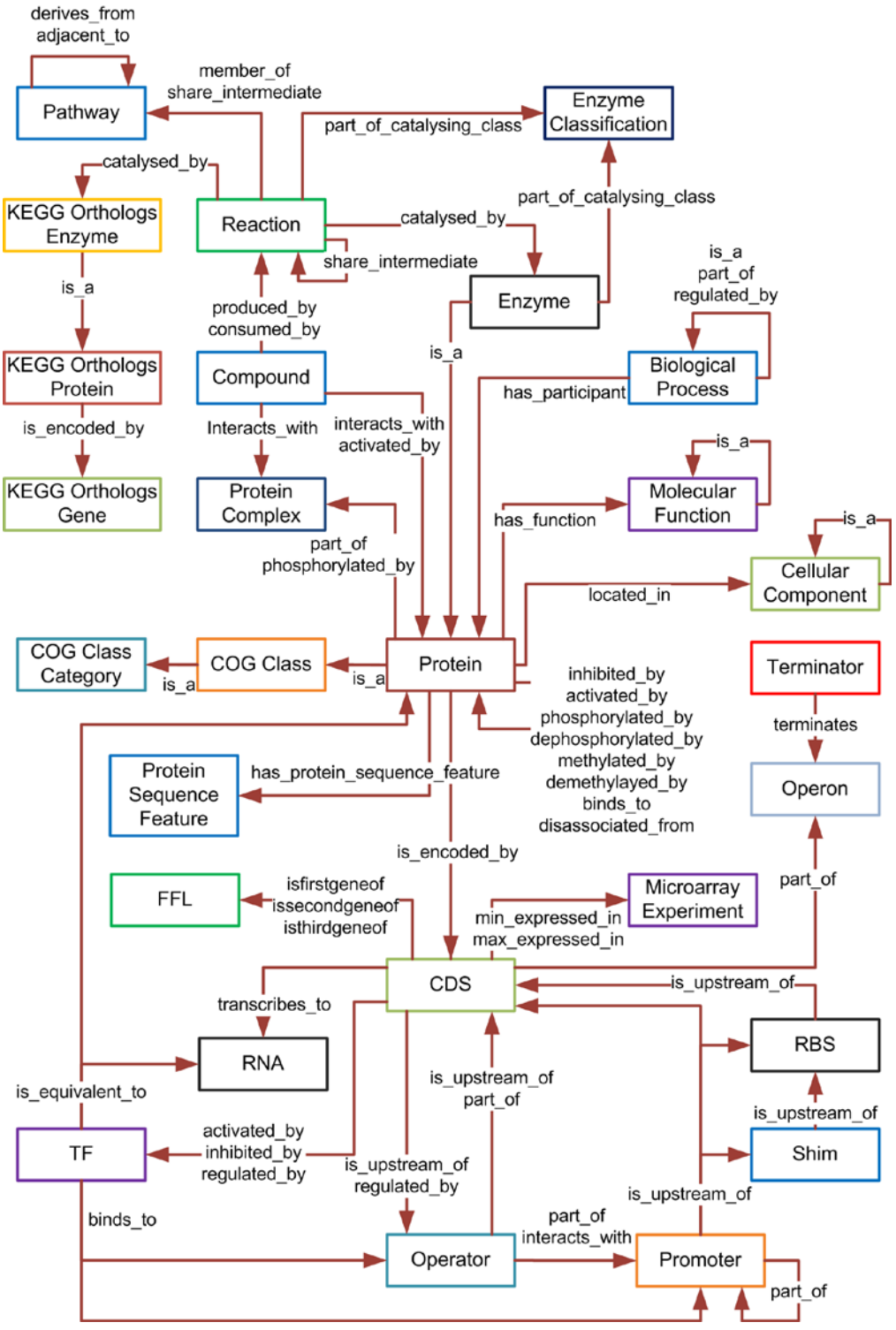

Figure 2. The BacillOndex ontology.

3.1.1 CatSNP Analysis

BacillOndex, in conjunction with a purpose-designed tool, CatSNP, was applied to a genome-wide analysis of the genetic differences between the widely-used laboratory strain of *B. subtilis*, 168, and its parent strain, Marburg. This process produced a set of BacillOndex subgraphs linking each SNP with all available data, at every level of biological organisation (Figure 3). Not all of the concepts shown in Figure 3 were available for every SNP.

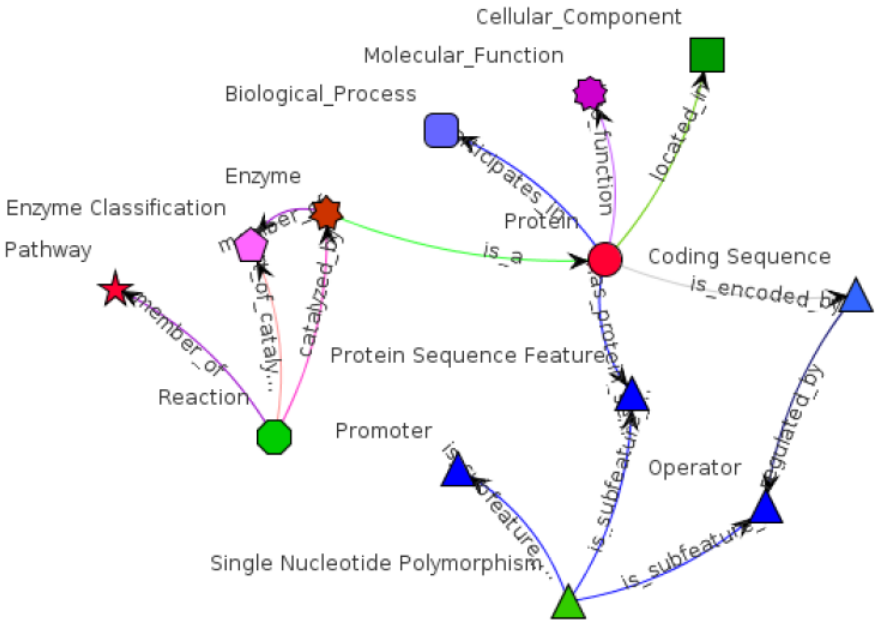

Figure 3. An overview of the concept classes and relation types contained within the graph produced by the BacillOndex dataset showing the relationship between concept classes.

The problem of missing data occurred at all stages of the analysis process. The applicability of any integrated dataset is dependent upon the data available, and even for a well-studied and frequently-used organism such as *B. subtilis* significant amounts of data are unavailable, particularly for genes and proteins of secondary interest to industry. For example, although 65 high-quality SNPs were identified in the alignment of 168 and Marburg, only 57 could be located in BacillOndex. Although we had genomic location information for all SNPs, some genes were simply not represented in the databases upon which we drew. Overall, 44 nsSNPs or INDELS, 5 sSNPs and 8 eSNPs were identified.

The genes containing SNPs had a range of annotations when investigated using GO and KEGG (Table 2). A wide range of concepts was associated with these SNPs (Table 3). There was, however, enough data for some SNPs to allow analysis from the genomic level up to pathways, and even to the inference of biological function (Figure 4).

Table 2: SNPs Identified in *B. subtilis* 168 compared with the parent Marburg strain.

Protein	Annotation	Protein	Annotation
Non-synonymous SNPs		**Expression SNPs**	
TrpC	Phenylalanine, tyrosine and tryptophan biosynthesis	YesS	Transcription, DNA dependent
AroH	Phenylalanine, tyrosine and tryptophan biosynthesis	YisR	Transcription, DNA dependent
SfP	Antibiotic anabolism	PerR	Regulation of transcription
PspC	Antibiotic anabolism	FlgM	Regulation of transcription
GerAA	Spore germination	MtnK	Cysteine and methionine metabolism
SpovG	Cellular spore formation	DegQ	
SigH	Cellular spore formation	**Synonymous SNPs**	
MbL	Cellular morphogenesis	PpsC	Antibiotic anabolism
PiT	Phosphate transport	KipA	Cellular spore formation
AmyD	Transport	PgdS	Hydrolase
PhoD	Folate biosynthesis	IlvC	Valine, leucine and isoleucine biosynthesis
SacA	Starch & sucrose metabolism		
Glt	Alanine, aspartate and glutamate metabolism		
HemA	Porphyrin and chlorophyll metabolism		
RbsR	Regulation of transcription		
TrmD	Methylation		
Rlub	RNA modification		

Table 3: Concepts associated with the high-probability SNPs.

Concept	Number	Concept	Number
SNP	57	Promoter	3
Pathway	18	Shim	1
Reaction	19	Ribosomal Binding Site	1
Enzyme Classification	11	Operator	1
Enzyme	10	GO Cellular Component	8
Protein	40	GO Molecular Function	59
CDS	39	GO Biological Process	48
Protein Sequence Feature	55		

doi:10.2390/biecoll-jib-2013-224

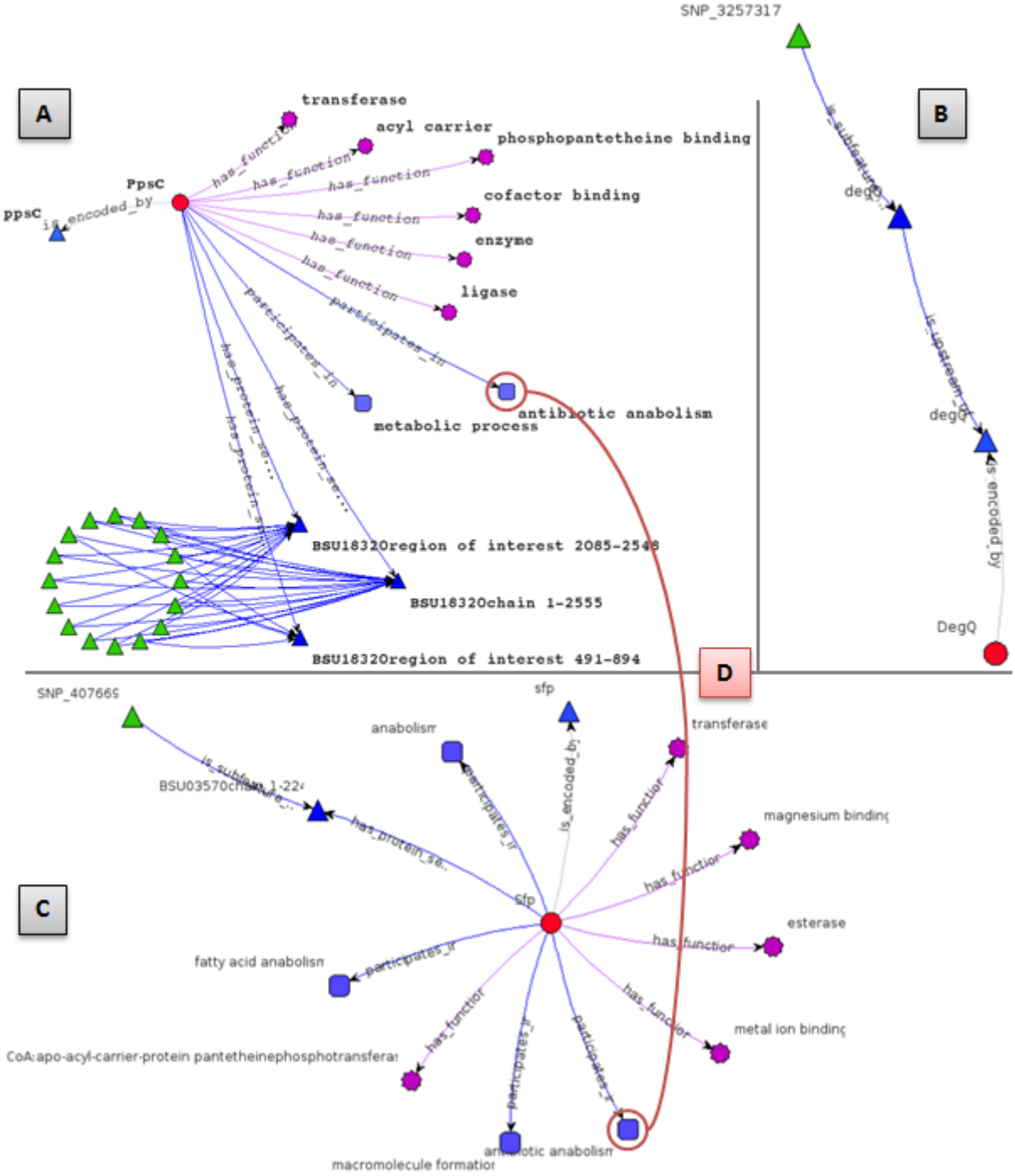

Figure 4. SNPs can be linked to phenotype at several levels of complexity. A: 16 SNPs (1sSNP & 15 nsSNPs- of which 9 are INDELs) associated with the protein chain of the plipistatin synthase subunit C, PpsC. 6 SNPs are found within the adenylation 1 region of the protein chain with 9 found within the epimerization region (please note that for the purpose of this graph the SNP_Labels as well as SNP_relation labels (is_subfeature_of) have been hidden).B: An eSNP within the promoter region of the pleiotropic regulator, DegQ. C: Visualisation of the insertion within the protein chain of the Surfactin synthetase-activating enzyme, SfP. D: It is clear that both SfP and PpsC are involved in the formation of antibiotics.

4 Discussion

The BacillOndex and CatSNP tools together facilitate the analysis of very large amounts of data. In this application we compared two complete bacterial genomes, each of 4.2MB, and reduced this data to 65 SNPs most likely to be relevant to our interests. Although this process still required considerable human input, particularly with respect to establishing the likelihood that the SNPs identified are valid, many of the tasks we did manually could be automated in the future. Even with the manual analysis steps, our workflow allows much more data to be analysed that would be possible without it.

Although the purpose of the workflow is, essentially, to eliminate irrelevant data, it is still possible to produce a useful level of biological detail (Figure 4). As synthetic biologists, we are interested in understanding the genetic basis of differences between the laboratory workhorse *B. subtilis* 168 and its parent strain, Marburg. Our aim is to engineer organisms with novel, predictable behaviour, and in order to do so a systems-level understanding of the organisms in question is valuable [37]. In the case of *B. subtilis* the mutations which created strain 168 were introduced at random, but the organism has had 65 years of laboratory existence in which to evolve both to fit life in a lab, and to adjust the effect of the initial mutations to work together.

We found that many of the SNPs identified were in CDSs involved in biological processes contributing to phenotypic differences known to exist between Marburg and 168. For example, in a laboratory 168 grown on rich medium is generally under very little pressure to form spores. A number of the SNPs we identified occur in proteins related to sporulation. It has previously been speculated that sporulation deficiency in 168 is due to small indels and substitutions [38]. We also identified some previously known mutations, including one in TrpC2, essential for tryptophan metabolism [39], and several that interfere with motility in 168. Some other pathways which are strongly represented in the data, however, such as phenylalanine, tyrosine and tryptophan biosynthesis, and antibiotic anabolism, are less well represented in the literature, and are worth further investigation.

Although we had specific interests underlying the development of our workflow, BacillOndex and CatSNP could be applied to the investigation of many other aspects of *Bacillus* biology. Ondex has already been used to identify candidates for drug repurposing in a human dataset [11], and its application to drug target identification in bacteria would be straightforward.

Many industrial applications involve the production of novel biomolecules. An obvious extension of BacillOndex, with particular value to synthetic biology, would be the incorporation of data from other organisms. By identifying commonalities and differences between different organisms, it should be possible to identify compatible pathways which could easily be moved between organisms, or unexpected interactions, regulatory or otherwise, between chassis and novel circuits.

An obvious limitation of BacillOndex is the lack of data available for some genes, proteins and interactions. Of the 65 SNPs which we identified as potentially interesting using sequence alignment and analysis, only 57 could be identified in BacillOndex. Recent technological advances in both data generation and bioinformatics analysis mean, however, that more data is constantly becoming available. Much of this data, generated by high-throughput experiments, is not published in the peer-reviewed literature, but is deposited in online databases, and is therefore valuable grist to the data miner's mill. We have made BacillOndex and its associated parsers freely available so that new data can be added to the system as it becomes available, by any user. Many of the parsers are generic enough to be used for any organism, making the development of new integrated bacterial datasets relatively straightforward. New parsers can also be developed as data sources come online or change.

doi:10.2390/biecoll-jib-2013-224

Ondex, BacillOndex and the CatSNP plugin are data integration and workflow development tools which we believe will be valuable to systems and synthetic biologists interested in making maximal use of the variety of data available for microbes. We have applied them to an evolutionary analysis of *B. subtilis*, but they are potentially much more widely applicable, and should aid in the investigation of a wide range of biological questions.

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.).

References

[1] D. G. Gibson, J. I. Glass, C. Lartigue, V. N. Noskov, R.-Y. Chuang, M. A. Algire, G. A. Benders, M. G. Montague, L. Ma, M. M. Moodie, *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, 329:52 - 56, 2010.

[2] M. Y. Galperin and X. M. Fernández-Suárez. The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Research*, 40:D1-D8, 2012.

[3] I. Lee, S. Date, A. Adai, and E. Marcotte. A probabilistic functional network of yeast genes. *Science*, 306:1555-1558, 2004.

[4] J.-D. J. Han. Understanding biological functions through molecular networks. *Cell Research*, 18:224-237, 2008.

[5] D. Benslimane, D. Schahram, and S. Amit. Services mashups: The new generation of Web applications. *IEEE Internet Computing*, 12:13 - 15, 2008.

[6] R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, *et al.* The automation of science. *Science*, 324:85-89, 2009.

[7] R. Stevens, C. A. Goble, and S. Bechhofer. Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, 1:398-414, 2000.

[8] J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Ruegg, C. Rawlings, P. Verrier, and S. Philippi. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22:1383-1390, 2006.

[9] J. Weile, M. Pocock, S. J. Cockell, P. Lord, J. M. Dewar, E. Holstein, D. Wilkinson, D. Lydall, J. S. Hallinan, and A. Wipat. Customisable views on semantically integrated networks for systems biology. *Bioinformatics*, 27:1299-1306, 2011.

[10] A. Splendiani, C. Rawlings, S.-C. Kuo, R. Stevens, and P. Lord. Lost in translation: Data integration tools meet the Semantic Web (experiences from the Ondex project). In F. L. Gaol, ed., *Recent Progress in Data Engineering and Internet Technology*. vol. 157 of *Lecture Notes in Electrical Engineering*, Springer Berlin Heidelberg, pp. 87-97, 2012.

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

[12] J. Hallinan, G. Misirli, and A. Wipat. Evolutionary computation for the design of a stochastic switch for synthetic genetic circuits. In *32nd Annual International*

Conference of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2010), Buenos Ares, Argentina, 2010.

[13] J. Hallinan, S. Park, and A. Wipat. Bridging the gap between design and reality: A dual evolutionary strategy for the design of synthetic genetic circuits. presented at the Bioinformatics: The International Conference on Bioinformatics Models, Methods and Algorithms, Algarve, Portugal, 2012.

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

[15] H. V. Westerhoff and B. O. Palsson. The evolution of molecular biology into systems biology. *Nature Biotechnology*, 22:1249 - 1252, 2004.

[16] A. Goelzer, F. Bekkal Brikci, I. Martin-Verstraete, P. Noirot, P. Bessieres, S. Aymerich, and V. Fromion. Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. *BMC Systems Biology*, 2:20, 2008.

[17] H. Bolouri. *Computational Modelling Of Gene Regulatory Networks - A Primer*: Imperial College Press, 2008.

[18] C. Henry, J. Zinner, M. Cohoon, and R. Stevens. iBsu1103: a new genome-scale metabolic model of *Bacillus subtilis* based on SEED annotations. *Genome Biology*, 10:R69, 2009.

[19] Y.-K. Oh, B. O. Palsson, S. M. Park, C. H. Schilling, and R. Mahadevan. Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *Journal of Biological Chemistry*, 282:28791 - 28799, 2007.

[20] P. R. Burkholder and N. H. Giles. Induced biochemical mutations in *Bacillus subtilis*. *American Journal of Botany*, 34:345 - 348, 1947.

[21] D. R. Zeigler, Z. Prágai, S. Rodriguez, B. Chevreux, A. Muffler, T. Albert, R. Bai, M. Wyss, and J. B. Perkins. The origins of 168, W23, and other *Bacillus subtilis* legacy strains. *Journal of Bacteriology*, 190:6983-6995, 2008.

[22] K. Kobayashi, S. D. Ehrlich, A. Albertini, G. Amati, K. K. Andersen, M. Arnaud, K. Asai, S. Ashikaga, S. Aymerich, P. Bessieres, *et al.* Essential *Bacillus subtilis* genes. *Proceedings of the National Academy of Sciences*, 100:4678-4683, 2003.

[23] F. Kunst, N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessieres, A. Bolotin, S. Borchert, *et al.* The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, 390:249-256, 1997.

[24] V. Barbe, S. Cruveiller, F. Kunst, P. Lenoble, G. Meurice, A. Sekowska, D. Vallenet, T. Wang, I. Moszer, C. Medigue, *et al.* From a consortium sequence to a unified sequence: the *Bacillus subtilis* 168 reference genome a decade later. *Microbiology*, 155:1758-1775, 2009.

[25] H. Lee and H. Tang. Next-generation sequencing technologies and fragment assembly algorithms. In M. Anisimova, ed., *Evolutionary Genomics: Statistical and Computational Methods*, vol. 855 of *Methods in Molecular Biology*, pp. 155-174, Humana Press, 2012.

[26] H. Mi, A. Muruganujan, and P. D. Thomas. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research*, 41:D377-D386, 2013.

doi:10.2390/biecoll-jib-2013-224

[27] A. Dereeper, S. Nicolas, L. Lecunff, R. Bacilieri, A. Doligez, J. P. Peros, M. Ruiz, and P. This. SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC Bioinformatics,* 12:134, 2011.

[28] M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, and J. P. Mesirov. GenePattern 2.0. *Nature Genetics,* 38:500-501, 2006.

[29] B. Chen, S. Wilkening, M. Drechsel, and K. Hemmincki. SNP_tools: A compact tool package for analysis and conversion of genotype data for MS-Excel. *BMC Research Notes,* 2:214, 2009.

[30] C. L. Barrett, T. Y. Kim, H. U. Kim, B. Ø. Palsson, and S. Y. Lee. Systems biology as a foundation for genome-scale synthetic biology. *Current Opinion in Biotechnology,* 17:488-492, 2006.

[31] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Research,* 36:D480-D484, 2008.

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

[33] C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Krueger, B. Snel, and P. Bork. STRING 7-recent developments in the integration and prediction of protein interactions. *Nucleic Acids Research,* 35:D358-D362, 2006.

[34] S. Goto, S. Kawashima, Y. Okuji, T. Kamiya, S. Miyazaki, Y. Numata, and M. Kanehi-sa. KEGG/EXPRESSION: A database for browsing and analysing microarray expression data. *Genome Informatics,* 11:222 - 223, 2000.

[35] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.* Gene Ontology: Tool for the unification of biology. *Nature Genetics,* 25:25-29, 2000.

[36] M. Magrane and T. UniProt Consortium. UniProt Knowledgebase: a hub of integrated protein data. *Database,* 2011:bar009, 2011.

[37] G. M. Church. From systems biology to synthetic biology. *Molecular Systems Biology,* 1:2005.0032, 2005.

[38] H. Maughan, C. W. Birky, and W. L. Nicholson. Transcriptome divergence and the loss of plasticity in *Bacillus subtilis* after 6,000 generations of evolution under relaxed selection for sporulation. *Journal of Bacteriology,* 191:428-433, 2009.

[39] A. M. Albertini and A. Galizzi. The sequence of the trp operon of *Bacillus subtilis* 168 (trpC2) revisited. *Microbiology,* 145:3319-3320, 1999.