e-Science Tools For The Genomic Scale Characterisation Of Bacterial Secreted Proteins

Tracy Craddock[1], Phillip Lord[1], Colin Harwood[2] and Anil Wipat[1]

[1]School of Computing Science and [2]Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne, Tyne and Wear, UK

Abstract

Within bioinformatics, a considerable amount of time is spent dealing with three problems; *heterogeneity*, *distribution* and *autonomy* – concerns that are mirrored in this study and which e-Science technologies should help to address. We describe the design and architecture of a system which makes predictions about the secreted proteins of bacteria, describing both the strengths and some weaknesses of current e-Science technology.

The focus of this study was on the development and application of e-Science workflows and a service oriented approach to the genomic scale detection and characterisation of secreted proteins from *Bacillus* species. Members of the genus *Bacillus* show a diverse range of properties, from the non-pathogenic *B. subtilis* to the causative agent of anthrax, *B. anthracis*. Protein predictions were automatically integrated into a custom relational database with an associated Web portal to facilitate expert curation, results browsing and querying. We describe the use of workflow technology to arrange secreted proteins into families and the subsequent study of the relationships between and within these families. The design of the workflows, the architecture and the reasoning behind our approach are discussed.

1. Introduction

Bioinformatics has become one of the major applications areas for e-Science. The development of high-throughput technologies and the desire to understand organisms as complex systems, rather than the more traditional approach of studying their component parts, is placing new requirements for a computing infrastructure.

Outside of a few specialist centres, the majority of bioinformatics tasks lack the extreme requirements for raw computing power and large-scale storage, typical of physics. Moreover, bioinformatics tools and datasets tend to be highly heterogeneous in content and structure; the datasets are often widely geographically dispersed, and, as they are often maintained and deployed by individual scientists within their own laboratories, autonomously controlled. Dealing with these three problems – heterogeneity, distribution and autonomy – often forms a significant part of the work load of a bioinformatician.

Many bioinformatics experiments can be represented as a series of workflows, integrating a number of programs and data sources to test a hypothesis. These workflows, normally called “pipelines” within the field, form the bedrock of the computational analysis within bioinformatics. Traditionally, these have been implemented in two different ways. Firstly, in those laboratories with the necessary resources or specialist support, they have often been automated using Perl (which is ideally suited to the manipulation of the textual representations that biology has traditionally used to store its data). The majority of biologists, however, have used cut-and-paste between the myriad of Web sites offering access to underlying computational resources; in a sense, biology has predated the Web Service revolution, albeit in a rather *ad hoc* manner.

In previous work [1,2] we have described the ^{my}Grid project which aims to provide an alternative to these two approaches. The use of Web Services, a workflow enactment engine and a convenient, easy-to-use workflow editor, Taverna [3], have enabled lab biologists to access some of the power of automation available previously only to programmers.

In this paper, we describe the application of ^{my}Grid technology to an additional biological problem. We wish to understand and predict the characteristics and behavior of a family of bacteria, through an analysis of their complete genomic sequences. In this work we focus on a family of bacteria, *Bacillus*, whose members show a diverse range of properties. In particular, we wish to identify the proteins that are produced by these bacteria and secreted across the cytoplasmic membrane.

This problem places different requirements on the workflow and the surrounding architecture than previously described in the bioinformatics workflow domain. Here, we describe in detail the background to the problem and the biological analysis that we wish to perform to address this problem. Finally, we describe the architecture that we have developed to support this analysis and discuss some preliminary results of its application.

2. The Secretome

One of the main mechanisms that bacteria use to interact with their environment is to synthesise proteins and export them from the cell into their external surroundings. These secreted proteins are often important in the adaptation of bacteria to a particular environment. The entire complement of secreted proteins is often referred to as the *secretome*. Characterising secreted proteins and the mechanisms of their secretion can reveal a great deal about the capabilities of an organism. For example, soil organisms secrete macromolecular hydrolases to recover nutrients from their immediate surroundings. During infection, many pathogenic bacteria secrete harmful enzymes and toxins into the extracellular environment. These secreted virulence proteins can subvert the host defence systems, for example by limiting the influence of the innate immune system, and facilitating the entry, movement and dissemination of bacteria within infected tissues [4]. Bacteria may also use secreted proteins to communicate with each other, allowing them to form complex communities and to adapt quickly to changing conditions [5].

The secretomes of pathogens are therefore of great interest. The comparison of virulent and non-virulent bacteria at the genomic and proteomic levels can aid our understanding of what makes a bacterium pathogenic and how it has evolved [6]. Furthermore, the characterisation of secretory virulence related proteins could ultimately lead to the identification of therapeutic targets.

The interest in protein secretion not only includes the secreted proteins themselves, but also those proteins which form the secretory machinery used to export the proteins across the cytoplasmic membrane, and in the case of Gram-negative bacteria, the outer membrane. Secretory proteins incorporate a short sequence called a signal peptide, which acts as a trafficking signal directing the protein to the secretory machinery. Not all proteins that are translocated from the site of synthesis are secreted into the surrounding medium; *transmembrane proteins* are localised to the cytoplasmic membrane itself; *lipoproteins* attach themselves to the outer surface of the membrane, protruding into the periplasmic space of Gram-negative bacteria or the membrane/wall interface of Gram-positive bacteria. Finally, proteins may also attach themselves to the cell wall by covalent or ionic interactions; the former may be distinguished by an LPXTG motif that is found in the C-terminal domains of mature secreted proteins [7].

Recently, our understanding of the secretome has greatly improved due to the rapidly increasing number of complete bacterial genome sequences, essentially molecular blueprints containing the information that allows the protein repertoire of an organism to be defined. Armed with this sequence information, we can begin to predict which of the proteins an organism is capable of producing are destined to be secreted, as well as the mechanisms of their secretion. As a result, a number of studies have been undertaken, in which bioinformatics programs and resources play a vital role. These studies involve the repeated application of a number of different algorithms to all of the gene and protein sequences encoded on the genome. Many of these algorithms are computationally expensive and, given that an average bacterial genome can encode around 4,000 or more proteins, the process can become computationally bound. In addition, the results of the application of these algorithms needs to be stored and integrated in order to make a prediction about the secretory status of the entire set of proteins encoded by a particular genome. Often, the results of the classification algorithms may be error prone and mechanisms to permit expert human curation and results browsing also need to be established.

Biologists have already begun to apply conventional bioinformatics technology to the prediction and classification of secreted proteins. The ‘first, largely genome-based survey of a secretome’ was carried out using bioinformatics tools on the genome of the industrially important bacterium, *Bacillus subtilis* [8], using legacy tools called from custom scripts in combination with expert curation.

In this study we describe the development and application of e-Science workflows and a service-oriented approach to the genomic scale detection and characterisation of secreted proteins from *Bacillus* species. *Bacillus* species are important not only for their industrial importance in the production of enzymes and pharmaceuticals, but also because of the diversity of characteristics shown by the members of this genus. The *Bacillus* genus includes species that are soil inhabitants, able to promote plant growth and produce antibiotics. The genus also includes harmful bacteria such as *Bacillus anthracis*, the causative agent of anthrax.

We utilised the system to make predictions about the secretomes of 12 *Bacillus* species for which complete genomic sequences are publicly available; this includes *B. cereus* (strain ZK/E33L), *B. thuringiensis* konkukian (strain 97-27), *B. anthracis* (Sterne), *B. anthracis* (Ames ancestor), *B. anthracis* (Ames, isolate Porton), *B. cereus* (ATCC 10987), *B. cereus* (strain ATCC 14579/DSM 31), *B. subtilis* (168), *B. licheniformis* (strain DSM 13/ATCC 14580, sub_strain Novozymes), *B. licheniformis* (DSM 13/ATCC 14580, sub_strain Goettingen), *B. clausii* (KSM-K16) and *B. halodurans* (C-125/JCM 9153). These predictions were automatically integrated into a custom relational database with an associated Web portal to facilitate expert curation, results browsing and querying.

3. Workflow Approach

In this study, the aim was to identify proteins that are likely to be secreted and classify them according to the putative mechanism of their secretion. Such proteins include those exported to the extracellular medium, as well as proteins that attach themselves to the outer surface of the membrane (lipoproteins) and cell wall binding proteins (sortase mediated proteins containing an LPXTG motif). Once secreted proteins had been classified, we then investigated the composition of the predicted secretomes in the twelve species under study.

Two workflows were designed and implemented; the *classification* workflow and the *analysis* workflow. The classification workflow is concerned with making predictions about the secretory characteristics of a particular protein from a given set of proteins. The analysis workflow processes the data from the first workflow in order to analyse the function of the secreted proteins that have been found.

A general feature of the workflows is their linear construction. In the classification workflow, for example, at each step, the set is reduced in size, removing those proteins that do not require further classification. A conceptual diagram illustrating the functionality of the classification workflow is shown in Figure 1. The results of the classification process are stored in a remote relational database and the reduced set of proteins passed to the next service in the workflow. In the analysis workflow, data derived from the classification workflow is retrieved from the database and analysed using a further series of service enabled tools. The architectural constraints responsible for this choice of design are discussed further in section 4. The data flow for both workflows combined is summarised in Figure 2.

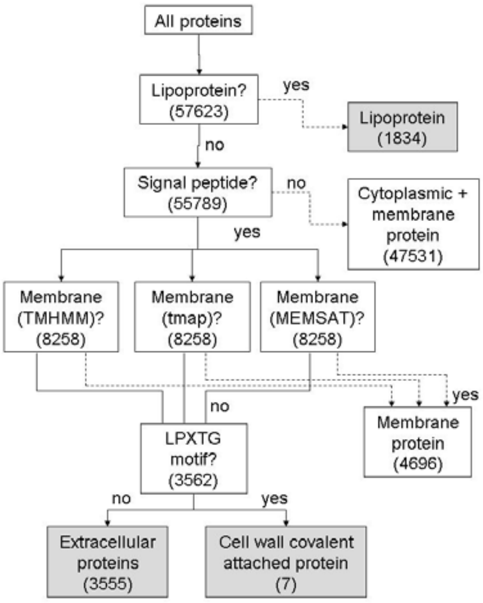


Figure 1. Basic representation of the functionality of the classification workflow. Shaded boxes indicate the set of secreted proteins. The number in brackets represents the total number of proteins to be classified at each level, across the 12 *Bacillus* species.

With respect to the workflow functionality, for the classification workflow, the first objective was the prediction of lipoproteins. This was the responsibility of the first service in the workflow, which takes as input a set of all predicted proteins derived from the EMBL record for the complete genome sequence. This service employed LipoP [9] for the detection of lipoproteins. Following the prediction of lipoproteins, the putative 'non-lipoprotein' set of proteins was analysed for the presence of a signal peptide using a SignalP Web Service [10]; this was performed after the lipoprotein identification because of possible limitations in the efficiency of SignalP at detecting lipoproteins. The use of SignalP at this point also removes most of the proteins from subsequent analysis; this has considerable advantages as the downstream analyses are potentially computationally intensive.

Proteins with a putative N-terminal signal peptide as well as additional transmembrane domains are likely to be retained in the membrane. To identify these putative membrane proteins from among the proteins in the signal peptide dataset, we used a combination of three transmembrane prediction Web Services based on the tools, TMHMM, MEMSAT and tmap, respectively [11]. A subsequent service in the workflow was responsible for integrating the results derived from these three tools, to make a final prediction about the presence of a putative transmembrane protein. Finally, the protein dataset corresponding to proteins with no predicted transmembrane domain was analysed for the cell wall binding amino acid motif LPXTG. The tool, ps_scan (the standalone version of ScanProsite) which scans PROSITE for the LPXTG motif [12] was wrapped as a Web Service and called from the workflow. The classification workflow was validated in its predicted capability by applying it to the proteins of *Bacillus subtilis* whose secretory status has been determined and, to a large extent, experimentally confirmed [8, 13].

For the analysis workflow, the secreted proteins dataset was analysed to provide information about the relationships between the secretomes of the 12 different organisms in the study. Putative secreted proteins were extracted from the database, clustered into families, and the structure, functional composition and relationships between these families were studied. The set of secreted proteins includes those predicted to be lipoproteins, cell wall binding, or extracellular. Transmembrane proteins and cytoplasmic proteins were disregarded. Analysis of the data was initiated by clustering the putative secretory proteins into protein families using the MCL graph clustering algorithm [14]. MCL provides a computationally efficient means of clustering large datasets. BLASTp data was used with the MCL algorithm to identify close relatives of the predicted secreted proteins. This approach follows that of [15], where the BLASTp algorithm provides a score for the putative similarity between proteins. The necessary BLASTp data was retrieved from the Microbase system, a Web Service enabled resource for the comparative genomics of microorganisms [16]. For each predicted secreted protein, similar proteins with a BLASTp expect value less than $1e^{-10}$ were used as input to MCL, (inflation value 1.2).

Hierarchical clustering was performed to identify phylogenetic relations between the *Bacillus* species in the context of their contributions to the protein families. The R package was wrapped as a Web Service for this purpose. R is a package providing a statistical computing and graphics environment (R Project website, http://www.r-project.org/). A distance matrix was constructed using the Euclidean distance, and clustering was carried out using a complete linkage method.

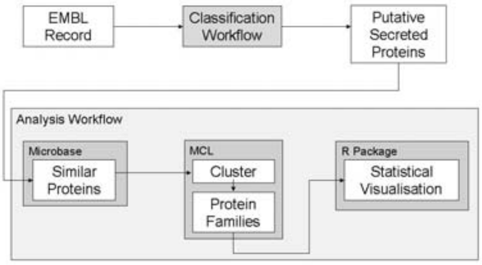


Figure 2. Data flow through the classification and analysis workflows.

4. Architecture

The architectural view of the workflow components involved in the prediction and analysis of the secretomes of the *Bacillus* species is shown in Figure 3.

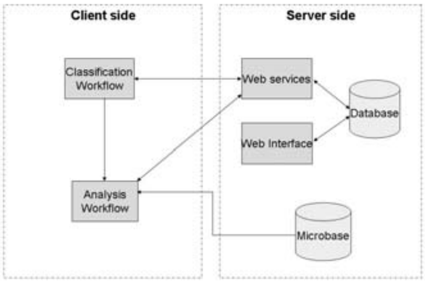


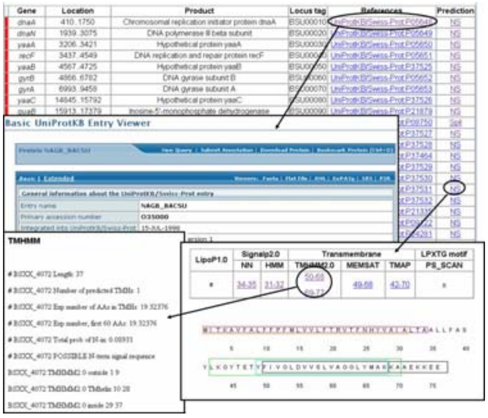
Figure 3. Architectural layout of the classification and analysis workflow components.

We sought to avoid implementing services on the client machine from which the workflows were enacted. This is particularly important for those services interacting with the database. In order to maximise performance, we endeavoured to locate the service as close to the database as possible. However, we also found continual problems with the reliability and availability of services provided by outside, autonomous individuals. Another problem was the restrictions placed on the usage and dissemination of some tools. Running such tools locally required licensing agreements to be signed, limiting the tool usage to the licensees. As a result of these factors, many of the required services were implemented in-house, although still delivered using the Web Services infrastructure. The services were orchestrated using SCUFL workflows and constructed and enacted using the Taverna workbench [3].

The workflows implement the bioinformatics processes outlined in Figure 2. Most of the services both use and return text based information, in standard bioinformatics formats. At all steps of the classification workflow, we have extracted the intermediate results into a custom-designed database. This was achieved using a set of bespoke data storage services which parse the raw textual results, and store them in a structured form. It is these structured data which are used to feed later services in both the classification and analysis workflows. In our case, the custom database is hosted on a machine in close network proximity to the Web Services. This has the significant advantage of reducing the network costs involved in transferring data to and from the database.

After completion of the classification workflow, the custom database contains the data relating to each protein analysed, including the raw data, as well as an integrated summary of the analysis. Tracking the provenance of the data is important in this context, because there are a number of different routes for the classification workflow to designate a protein as 'secretory'. The basic operational provenance provided by Taverna also aids in the identification of service failures [17]. This is particularly important while running the transmembrane domain prediction services, as these run concurrently; a failure in one, therefore, does not impact on the execution of the classification workflow, although may return incomplete conclusions and thus needs to be recorded.

We have developed a Web portal to provide a user-friendly and familiar mechanism for accessing the secretomes data in the database.[1] From this site, users can select the bacterial species in which they are most interested and view the corresponding results. Data is initially displayed as a table from where the users can navigate further to view the details of the classifications. The protein sequences may be viewed along with an overlay of predicted signal peptides and their cleavage sites. Users may also edit and curate the database as appropriate. A screenshot of the database portal is shown in Figure 4.



Gene	Location	Product	Locus tag	References	Prediction
dnaA	410..1750	Chromosomal replication initiator protein dnaA	BSU00010	UniProtKB/Swiss-Prot:P05648	NS
dnaN	1939..3075	DNA polymerase III beta subunit	BSU00020	UniProtKB/Swiss-Prot:P05649	NS
yaaA	3206..3421	Hypothetical protein yaaA	BSU00030	UniProtKB/Swiss-Prot:P05650	NS
recF	3437..4549	DNA replication and repair protein recF	BSU00040	UniProtKB/Swiss-Prot:P05651	NS
yaaB	4567..4725	Hypothetical protein yaaB	BSU00050	UniProtKB/Swiss-Prot:P37525	NS
gyrB	4866..6782	DNA gyrase subunit B	BSU00060	UniProtKB/Swiss-Prot:P05652	NS
gyrA	6993..9458	DNA gyrase subunit A	BSU00070	UniProtKB/Swiss-Prot:P05653	NS
yaaC	14845..15792	Hypothetical protein yaaC	BSU00080	UniProtKB/Swiss-Prot:P37526	NS
guaB	15913..17379	Inosine-5'-monophosphate dehydrogenase	BSU00090	UniProtKB/Swiss-Prot:P21879	NS
				P08750	Spl
				P37527	NS
				P37528	NS
				P37464	NS
				P37529	NS
				P37530	NS
				P37531	NS
				P37532	NS
				P21335	NS
				[illegible]	NS
				[illegible]	NS

Basic UniProtKB Entry Viewer

General information about the UniProtKB/Swiss-Prot entry

Entry name	NAGB_BACSU
Primary accession number	O35000
Integrated into UniProtKB/Swiss-Prot	15-JUL-1998

TMHMM

```
# BSXX_4072 Length: 37
# BSXX_4072 Number of predicted TMHs: 1
# BSXX_4072 Exp number of AAs in TMHs: 19.32376
# BSXX_4072 Exp number, first 60 AAs: 19.32376
# BSXX_4072 Total prob of N-in: 0.00931
# BSXX_4072 POSSIBLE N-term signal sequence
BSXX_4072 TMHMM2.0 outside 1 9
BSXX_4072 TMHMM2.0 TMhelix 10 28
BSXX_4072 TMHMM2.0 inside 29 37
```

LipoP1.0	SignalP2.0		Transmembrane			LPXTG motif
	NN	HMM	TMHMM2.0	MEMSAT	TMAP	PS_SCAN
*	34-35	31-32	50-68 69-7[illegible]	49-68	42-70	*

Figure 4. Screenshot of the Web portal summarising the characteristics of predicted secreted proteins from *B. subtilis* (168).

The analysis process is the most computationally intensive section requiring a large number of BLASTp searches. BLAST is the most commonly performed task in bioinformatics [1], and as such there are many available services which could have been used. However, because of the computational intensive nature of large BLAST searches, we retrieved pre-computed BLAST results from the Microbase database. Microbase is a Grid based system for the automatic comparative analysis of microbial genomes [16]. Amongst other data, Microbase stores all-against-all BLASTp searches for over three hundred microbial genome sequences, and the data are accessible in a Web Service-based fashion.

The final visualisation steps in the analysis workflow were performed using R and MCL. Again, both services were implemented locally, although both were exposed using Web Services. Completion of the entire workflow took approximately 2-3 hours.

This architecture differs from others using ^{my}Grid technology. Whilst the original requirements appeared to favour a highly distributed approach, we have found that the technological and licensing constraints have led to a hybrid approach: a combination of services and workflows, combined with databases for both results storage and pre-caching of analyses. The combination of all of these technologies does result in fairly complex architecture but provides a system that is fast and reliable enough for practical use.

[1] At http://bioinf.ncl.ac.uk:8081/BaSPP

5. Results

The classification workflow was applied to the predicted proteomes of the 12 *Bacillus* spp. listed in section 2. The resulting predicted secretomes varied in size from 358 proteins in *B. clausii* (strain KSM-K16) (9% of the total proteome) up to 508 proteins in *Bacillus cereus* (strain ZK / E33L) (11% of the total proteome). An investigation into the functional distribution of the proteins comprising the secretome of each strain was carried out by classifying them into functionally related families based on their sequence similarity as defined by BLASTp. The member proteins of the 12 secretomes were arranged into 543 families of 2 or more members. Some 350 proteins showed no similarity to other proteins and hence did not fall into families. Core protein families that contain members from all 12 proteomes are of particular interest since they may represent secreted proteins whose functions are indispensable for growth in all environments. 9% of protein families were found to be 'core' and the functions of these were investigated. The Gene Ontology terms [18] of the genes encoding the proteins in each cluster were examined and then summarised by classifying the terms according to the "SubtiList" classification codes [19]. Figure 5 shows a summary of the different functional classifications of the 'core' secreted protein families. Interestingly, a large number of core families had not been experimentally characterised and remain of unknown function. More predictably, many core proteins were grouped into families concerned with cell wall related functions, transporter proteins and proteins responsible for membrane biogenesis.

In addition to defining core protein families, we were also interested in gaining some insight into the functions of protein families that are specific to pathogens and those that are specific to non-pathogens. Canned queries over the database allowed these results to be easily and repeatedly extracted. Interestingly, only 14 of the 106 protein families that were unique to the secretomes of the potentially pathogenic bacteria (*B. cereus* (ZK / E33L), *B. thuringiensis* konkukian (97-27), *B. anthracis* (Sterne), *B. anthracis* (Ames ancestor), *B. anthracis* (Ames, isolate Porton), *B. cereus* (strain ATCC 10987) and *B. cereus* (ATCC 14579 / DSM 31)), showed similarity to proteins with known functions. Thus, the majority were found to be of unknown function and remain to be characterised.

The secretory protein families that were unique to the non-pathogens showed functions that were indicative of their habitats. Of the 11 unique families, 5 encoded enzymes concerned with the breakdown and transport of plant polysaccharides, 2 were concerned with the structure of flagellae, and the remaining 4 were functionally unclassified.

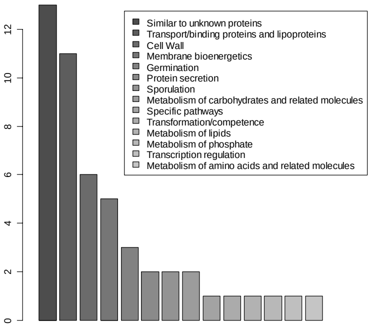


Figure 5. Functional classification of the 'core' secreted protein families. The graph shows the number of 'core' secreted proteins per 'SubtiList' category.

Finally, we were interested to determine whether the secretomes of the pathogenic organisms were closely related to each other in terms of their functional composition than to those of the non-pathogens. The phylogeny of the *Bacillus* strains was investigated in the context of the relationships between their secretomes. This was illustrated using a dendrogram, constructed using the R package, in which the relation between the different *Bacillus* strains is based on their contribution to the predicted secreted protein families (Figure 6). Essentially the dendrogram highlights the level of similarity between the secretomes of the various strains.

Within the *B. cereus* group subcluster (*B. anthracis*, *B. cereus* and *B. thuringiensis*), two sub-clusters were formed by the well-established pathogens (CP000001 *B. cereus*, AE017355 *B. thuringiensis* konkukian, AE017225 *B. anthracis*, AE017334 *B. anthracis*, AE016879 *B. anthracis),* while the two members of questionable pathogenesis (AE017194 *B. cereus*, AE016877 *B. cereus)* formed a separate cluster. The environmental strains (*B. subtilis*, *B. licheniformis*, *B. clausii*, *B. halodurans*) formed a separate cluster from that of the *B. cereus* group organisms.

6. Discussion

From the biologist's perspective: From the perspective of a biologist, construction of a workflow that enables secretory protein prediction over bacterial genomes using multiple prediction tools, integrates the results into a database, and then performs analysis on the families, is a novel development. This approach utilises current bioinformatics programs in order to make predictions, which would otherwise take several days if performed manually. In particular the ease by which a workflow may be re-run as and when new genomes are sequenced is a distinct advantage, especially as the rate of complete genome sequencing continues to increase. The initial time in developing the application should also be considered; though this factor reduces in significance with re-execution of the workflow across a number of genomes.

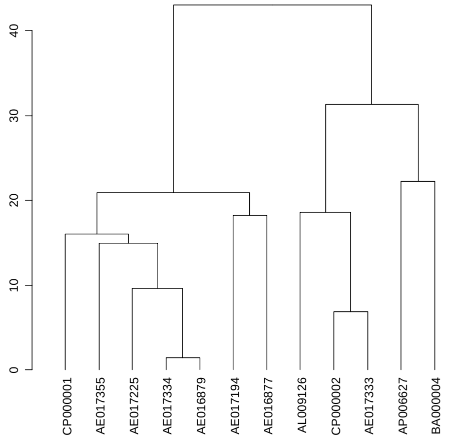


Figure 6. Dendrogram representing the relationship of the *Bacillus* species in terms of their secretome. From left to right: CP000001 *B. cereus* (ZK/E33L), AE017355 *B. thuringiensis* konkukian (97-27), AE017225 *B. anthracis* (Sterne), AE017334 *B. anthracis* (Ames ancestor), AE016879 *B. anthracis* (Ames, isolate Porton), AE017194 *B. cereus* (ATCC 10987), AE016877 *B. cereus* (ATCC 14579/DSM 31), AL009126 *B. subtilis* (168), CP000002 *B. licheniformis* (DSM 13/ATCC 14580, sub_strain Novozymes), AE173333 *B. licheniformis* (DSM 13/ATCC 14580, sub_strain Goettingen), AP006627 *B. clausii* (KSM-K16), BA000004 *B. halodurans* (C-125 / JCM 9153).

Another advantage for the biologist is that the generated data is stored in a custom database making the results available for subsequent analysis. It also provides a way of sharing data and promoting collaboration by providing an accessible and user-friendly Web interface to the database. This approach is different to current workflow methods where users need to take more control of where results are stored and possibly provide a means for the parsing of the data.

In addition to the bioinformatics approach, of course, the data generated by these workflows is also of great interest to microbiologists and this work provides data to prime further, biologically oriented investigations.

***From the eScientist's perspective*:** We have shown here the effectiveness of an e-Science approach, generating biologically meaningful results. These results can be used to generate hypotheses that may be verified by experimental approaches.

This process of data analysis was simplified through the use of Microbase. The comparison of the bacterial proteins had already been done, therefore reducing the computing time required in analysing the results.

Although the details of the workflow are specific to the problem of predicting secretory proteins, the architectural solutions that we have employed represent general issues for e-Science. We have attempted to deal with three key problems – distribution, autonomy and heterogeneity, in an efficient manner. While the e-Science framework has helped, it has been less successful in dealing with some key issues.

Dealing firstly with the problems of **autonomy**; most of the services in the classification workflow are, in fact, provided originally by external parties, autonomous from the workflow authors. As has been mentioned by previous authors [1], the **reliability** of the services deployed by many providers is not high. The problem has been significantly worsened in this case, as the workflows are largely linear; therefore, a failure by any service will cause the entire workflow to fail. We solved this problem by the simple approach of hosting the services locally, which is obviously not ideal. Having developed local services, we would have liked to at least republish them for reuse as a service to the community. There is, however, the second problem of **licensing** agreements: in most cases, we are not allowed to expose services for use by non-licensed individuals.

Perhaps surprisingly, there were few problems introduced by **distribution** in this work. The main recurrent difficulty came from the relatively large datasets over which we were operating. This was one of the motivations for the linear shape of the classification workflow. We can see two, more principled approaches to this problem. Firstly, **improved data transport** facilities, enabling transfer without SOAP packaging, as well as direct transfer between third parties, would reduce many difficulties. The new Taverna2 architecture should enable these functionalities [20]. Secondly, the ability to **migrate workflows** and services closer to the data would provide a significant advantage; in many cases the service executables are smaller than the data they operate over. It should be noted that licensing issues will prevent this in many cases.

Data **heterogeneity** has provided us with fewer problems than expected for a bioinformatics workflow. There are relatively few data types, most often protein lists are passed between the services in the workflows. Data heterogeneity was dealt with through the use of a custom database and parsing code. In the second workflow, the use of the Microbase data warehouse removed many of these problems.

One major future enhancement is to use, notification in conjunction with the workflow. This would provide a way of automatically analysing recently annotated genomes for secretory proteins. The need for users to interact with the workflow would therefore be removed, providing automatic updates of the database with new secretory protein data. We also intend to investigate the migration of workflows from the machine of their conception, to a remote enclosed environment from which they are able to access services that are unable to be exposed directly to third parties.

In conclusion, the investigation of predicted bacterial secretomes through the use of workflows and e-Science technology indicates the potential use of computing resources for maximising the information gained as multiple genomic sequences are generated. The knowledge gained from large scale analysis of secretomes can be used to generate inferences about bacterial evolution and niche adaptation, lead to hypotheses to inform future experimental studies and possibly identify proteins as candidate drug targets.

7. Acknowledgments

We gratefully acknowledge the support of the North East regional e-Science centre and the European Commission (LSHC-CT-2004-503468). We thank the EPSRC and Non-Linear Dynamics for supporting Tracy Craddock.

8. References

1. Stevens RD, Tipney HJ, Wroe CJ, Oinn TM, Senger M, Lord PW, Goble CA, Brass A, Tassabehji M., 2004. *Bioinformatics,* 20 Suppl. 1:I303-I310.
2. Li P, Hayward K, Jennings C, Owen K, Oinn T, Stevens R, Pearce S and Wipat A. *Proceedings of the UK e-Science All Hands Meeting 2004, 31st Aug - 3rd Sept, Nottingham UK*.
3. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P, 2004. *Bioinformatics*, 20(17): 3045-54.
4. Nomura K, He SY, 2005. *PNAS*. 102(10): 3527-3528.
5. Piazza F, Tortosa P, Dubnau D, 1999. *J Bacteriol*., 181(15): 4540-8.
6. Lee VT, Schneewind O, 2001. *Genes Dev.,* 15(14): 1725–1752.
7. Boekhorst J, de Been MW, Kleerebezem M, Siezen RJ, 2005. *J Bacteriol*., 187(14): 4928-34.
8. Tjalsma H, Bolhuis A, Jongbloed JDH, Bron S, van Dijl J.M, 2000. *Microbiol. Mol. Biol. Rev.*, 64(3): 515–547.
9. Juncker AS, Willenbrock H, von Heijne G, Nielsen H, Brunak S, Krogh A, 2003. *Protein Sci*., 12(8): 1652-62.
10. Nielsen, H. & Krogh., A., 1998. *Proc Int Conf Intell Syst Mol Biol. (ISMB 6), 6*: 122-130.
11. Moller S, Croning MDR, Apweiler R, 2001. *Bioinformatics*, 17(7): 646-653.
12. Gattiker A, Gasteiger E, Bairoch A, 2002. *Appl Bioinformatics*, 1(2): 107-108.
13. Tjalsma H, Antelmann H, Jongbloed JDH, Braun PG, Darmon E, Dorenbos R, Dubois JY, Westers H, Zanen G, Quax WJ, Kuipers OP, Bron S, Hecker M, van Dijl, J.M., 2004. *Microbiol. Mol. Biol. Rev.,* 68: 207-233.
14. van Dongen S, 2000. A cluster algorithm for graphs. *Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam*.
15. Enright AJ, Van Dongen S, Ouzounis CA, 2002. *Nucleic Acids Res.,* 30(7):1575-1584.
16. Sun Y, Wipat A, Pocock M, Lee PA, Watson P, Flanagan K, Worthington JT, 2005. *The 5th IEEE International Symposium on Cluster Computing and the Grid* (CCGrid 2005), Cardiff, UK, May 9-12.
17. Zhao J, Stevens R, Wroe C, Greenwood M. Goble C, 2004. In Proceedings of the *UK e-Science All Hands Meeting, Nottingham UK, 31 Aug-3 Sept.*
18. Gene Ontology Consortium, 2006. *Nucleic Acids Res.,* 34 (Database issue):D322-6.
19. Moszer I, Jones LM, Moreira S, Fabry C, Danchin A, 2002. *Nucleic Acids Res.,* 30(1): 62-5.
20. Oinn T & Pocock M, pers. comm