47. Risco, C. *et al.* Endoplasmic reticulum–Golgi intermediate compartment membranes and vimentin filaments participate in vaccinia virus assembly. *J. Virol.* **76**, 1839–1855 (2002).
48. Sodeik, B., Griffiths, G., Ericsson, M., Moss, B. & Doms, R. W. Assembly of vaccinia virus: effects of rifampin on the intracellular distribution of viral protein p65. *J. Virol.* **68**, 1103–1114 (1994).
49. Szajner, P., Weisberg, A. S., Lebowitz, J., Heuser, J. & Moss, B. External scaffold of spherical immature poxvirus particles is made of protein trimers, forming a honeycomb lattice. *J. Cell Biol.* **170**, 971–981 (2005).
50. Hyun, J. K. *et al.* The structure of a putative scaffolding protein of immature poxvirus particles as determined by electron microscopy suggests similarity with capsid proteins of large icosahedral DNA viruses. *J. Virol.* **81**, 11075–11083 (2007).
51. Federici, B. A. *et al.* in *VIIIth Report of the International Committee on Taxonomy of Viruses* (eds Fauquet, C. M., Mayo, M. A., Maniloff, J., Desselberger, U. & Ball, L. A.) 269–274 (Elsevier–Academic, London, 2005).
52. Stasiak, K., Demattei, M. V., Federici, B. A. & Bigot, Y. Phylogenetic position of the *Diadromus pulchellus* ascovirus DNA polymerase among viruses with large double-stranded DNA genomes. *J. Gen. Virol.* **81**, 3059–3072 (2000).
53. Stasiak, K., Renault, S., Demattei, M. V., Bigot, Y. & Federici, B. A. Evidence for the evolution of ascoviruses from iridoviruses. *J. Gen. Virol.* **84**, 2999–3009 (2003).
54. Feschotte, C. & Pritham, E. J. Non-mammalian c-integrases are encoded by giant transposable elements. *Trends Genet.* **21**, 551–552 (2005).
55. Kapitonov, V. V. & Jurka, J. Self-synthesizing DNA transposons in eukaryotes. *Proc. Natl Acad. Sci. USA* **103**, 4540–4545 (2006).
56. Pritham, E. J., Putliwala, T. & Feschotte, C. *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* **390**, 3–17 (2007).
57. Strömsten, N. J., Bamford, D. H. & Bamford, J. K. *In vitro* DNA packaging of PRD1: a common mechanism for internal-membrane viruses. *J. Mol. Biol.* **348**, 617–629 (2005).
58. Iyer, L. M., Makarova, K. S., Koonin, E. V. & Aravind, L. Comparative genomics of the FtsK–HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. *Nucleic Acids Res.* **32**, 5260–5279 (2004).
59. Burroughs, A. M., Iyer, L. M. & Aravind, L. Comparative genomics and evolutionary trajectories of viral ATP dependent DNA-packaging systems. *Genome Dyn.* **3**, 48–65 (2007).
60. Ginalski, K., Elofsson, A., Fischer, D. & Rychlewski, L. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015–1018 (2003).
61. Krupovič, M. & Bamford, D. H. Putative prophages related to lytic tailless marine dsDNA phage PM2 are widespread in the genomes of aquatic bacteria. *BMC Genomics* **8**, 236 (2007).
62. Chappell, J. D., Prota, A. E., Dermody, T. S. & Stehle, T. Crystal structure of reovirus attachment protein σ1 reveals evolutionary relationship to adenovirus fiber. *EMBO J.* **21**, 1–11 (2002).
63. Graham, S. C. *et al.* Structure of CrmE, a virus-encoded tumour necrosis factor receptor. *J. Mol. Biol.* **372**, 660–671 (2007).
64. Krupovič, M., Cvirkaitė-Krupovič, V. & Bamford, D. H. Identification and functional analysis of the *Rz*/*Rz1*-like accessory lysis genes in the membrane-containing bacteriophage PRD1. *Mol. Microbiol.* **68**, 492–503 (2008).
65. Hendrix, R. W., Lawrence, J. G., Hatfull, G. F. & Casjens, S. The origins and ongoing evolution of viruses. *Trends Microbiol.* **8**, 504–508 (2000).
66. Raoult, D. & Forterre, P. Redefining viruses: lessons from Mimivirus. *Nature Rev. Microbiol.* **6**, 315–319 (2008).
67. Porter, K. *et al.* SH1: a novel, spherical halovirus isolated from an Australian hypersaline lake. *Virology* **335**, 22–33 (2005).
68. Jaatinen, S. T., Happonen, L. J., Laurinmäki, P., Butcher, S. J. & Bamford, D. H. Biochemical and structural characterisation of membrane-containing icosahedral dsDNA bacteriophages infecting thermophilic *Thermus thermophilus*. *Virology* **379**, 10–19 (2008).
69. Jäälinoja, H. T. *et al.* Structure and host-cell interaction of SH1, a membrane-containing, halophilic euryarchaeal virus. *Proc. Natl Acad. Sci. USA* **105**, 8008–8013 (2008).
70. Männistö, R. H., Kivelä, H. M., Paulin, L., Bamford, D. H. & Bamford, J. K. The complete genome sequence of PM2, the first lipid-containing bacterial virus to be isolated. *Virology* **262**, 355–363 (1999).
71. Bamford, D. H. *et al.* Constituents of SH1, a novel lipid-containing virus infecting the halophilic euryarchaeon *Haloarcula hispanica*. *J. Virol.* **79**, 9097–9107 (2005).
72. Epifano, C., Krijnse-Locker, J., Salas, M. L., Salas, J. & Rodríguez, J. M. Generation of filamentous instead of icosahedral particles by repression of African swine fever virus structural protein pB438L. *J. Virol.* **80**, 11456–11466 (2006).
73. Rydman, P. S., Bamford, J. K. & Bamford, D. H. A minor capsid protein P30 is essential for bacteriophage PRD1 capsid assembly. *J. Mol. Biol.* **313**, 785–795 (2001).
74. Raoult, D. *et al.* The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).
75. Moreira, D. & Brochier-Armanet, C. Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC Evol. Biol.* **8**, 12 (2008).
76. Suhre, K. Gene and genome duplication in *Acanthamoeba polyphaga Mimivirus*. *J. Virol.* **79**, 14095–14101 (2005).
77. Filée, J. & Chandler, M. Convergent mechanisms of genome evolution of large and giant DNA viruses. *Res. Microbiol.* **159**, 325–331 (2008).
78. São-José, C., de Frutos, M., Raspaud, E., Santos, M. A. & Tavares, P. Pressure built by DNA packing inside virions: enough to drive DNA ejection *in vitro*, largely insufficient for delivery into the bacterial cytoplasm. *J. Mol. Biol.* **374**, 346–355 (2007).
79. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
80. Yan, X., Chipman, P. R., Castberg, T., Bratbak, G. & Baker, T. S. The marine algal virus PpV01 has an icosahedral capsid with T = 219 quasisymmetry. *J. Virol.* **79**, 9236–9243 (2005).

Acknowledgements
We thank J. Ravantti for his invaluable comments, and M. Jalasvuori and J.K. Bamford for sharing their unpublished data. This work was supported by the Finnish Center of Excellence Program (2006-2011) of the Academy of Finland (grants 1213467 and 1213992 to D.H.B. and grant 1210253). M.K. is supported by the Viikki Graduate School in Biosciences.

DATABASES
Entrez Genome: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genome
Bam35 | Orf virus | PBCV-1 | PM2 | PRD1 | STIV
Protein Data Bank: http://www.rcsb.org/pdb/home/home.do
adenovirus MCP | PRD1 MCP | PBCV-1 MCP | PM2 MCP | STIV MCP

FURTHER INFORMATION
Dennis H. Bamford's homepage: http://blogs.helsinki.fi/bamford-group/
BioInfoBank Meta Server: http://meta.bioinfo.pl/submit_wizard.pl
Repbase Update: http://www.girinst.org/repbase/update/index.html

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

OPINION

e-Science: relieving bottlenecks in large-scale genome analyses

Tracy Craddock, Colin R. Harwood, Jennifer Hallinan and Anil Wipat

Abstract | The development of affordable, high-throughput sequencing technology has led to a flood of publicly available bacterial genome-sequence data. The availability of multiple genome sequences presents both an opportunity and a challenge for microbiologists, and new computational approaches are needed to extract the knowledge that is required to address specific biological problems and to analyse genomic data. The field of e-Science is maturing, and Grid-based technologies can help address this challenge.

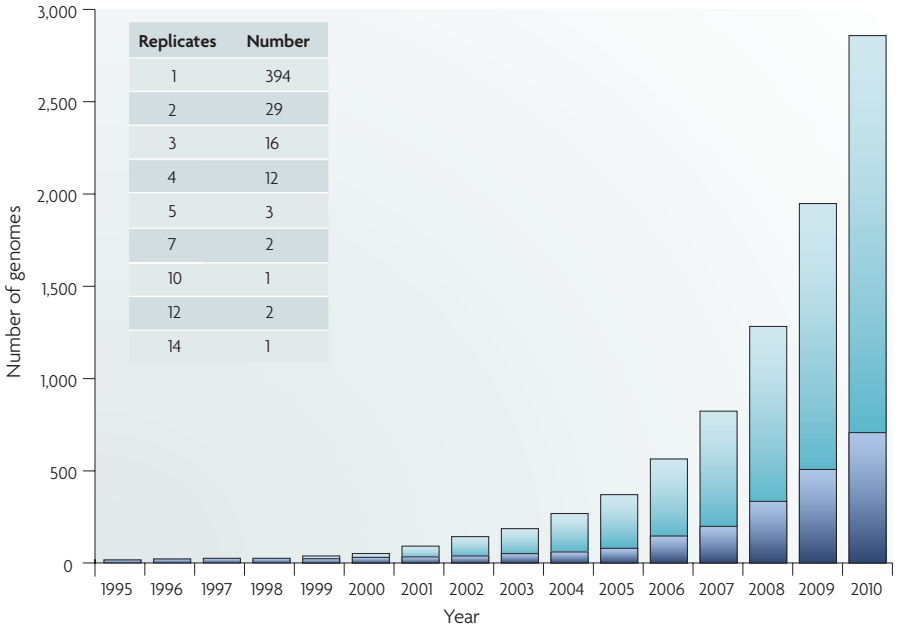
By 30 September 2007, 708 bacterial genome sequences had been submitted to publicly accessible databases, such as those at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL–EBI), the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the National Center for Biotechnological Information (NCBI) (see Further information). Many other genomes are currently being sequenced and several sequences have been completed but have not yet been placed in the public domain. The number of publicly available genomes has doubled approximately every 16 months since 2000. At this rate, more than 2,000 bacterial genomes will be available by the end of 2010 (FIG. 1).

In most cases, the genome of only a single representative of a species has been sequenced. Increasingly, however, replicate genome sequences are appearing in the public domain (FIG. 1). For example, there are now 14 publicly available *Staphylococcus aureus* genomes, with 2 additional strains at the 'finishing' stage. Using replicate genomes, direct comparisons can be made between closely related strains, providing unprecedented insights into the mechanisms involved in genome plasticity and rates of evolution. As

genome sequences are available for representatives of most of the major human and animal pathogens, it is likely that the proportion of replicate sequences will continue to increase.

The exponential increase in whole-genome sequence data and in associated post-genomic studies has generated an urgent requirement for new approaches to facilitate the global, collaborative analyses of the resulting data and the exchange of the results across social, political and technological boundaries. Over the past 5 years there has been a major effort, in the form of e-Science, to develop the technology that is needed to fulfil these requirements. The term e-Science was coined in the United Kingdom to refer to collaborative, global science that is performed *in silico* with a required computational infrastructure[1]. Tasks that require an e-Science approach are typically computationally intensive, and heterogeneous resources must be integrated across a distributed computing network. Conventional Internet-based technologies, such as e-mail and Web browsers, cannot support the current requirements for true e-Science approaches[2]. The infrastructure for e-Science (termed the cyber-infrastructure in the United States) is founded on a novel approach to the distribution of computing, known as the Grid.

e-Science is set to revolutionize computing just as genomics has revolutionized microbiology. This Opinion article reviews the latest developments in the core e-Science tool kit and its application to the analysis of genomes. We illustrate the value of e-Science to microbiologists by discussing how e-Science was used in the genomic analysis of the secretomes of bacteria in the genus *Bacillus*.



Replicates	Number
1	394
2	29
3	16
4	12
5	3
7	2
10	1
12	2
14	1

Figure 1 | **Publicly available bacterial genomes.** Light-blue bars show the accumulated number of sequences; dark-blue bars show the number of sequences published in the indicated year. The table shows the frequency with which replicate genomes of the same species have been sequenced (data from KEGG (Kyoto Encyclopedia of Genes and Genomes) bacterial genomes; see Further information).

e-Science and Grid technology

The term Grid was first coined to describe the use of a large number of distributed computers for computationally intensive problems with a vision that computational power should be as easy to harvest as plugging into a conventional electric power grid[3]. Grids, such as the TeraGrid, in combination with Grid technology, such as the Globus tool kit[4] and Condor[5], have many applications in the analysis of biological data. Grids are ideally suited for bioinformatics tasks, as these tasks often require access to high-performance computational power and large-scale data storage and resources that are only locally available to the most specialist centres.

The requirement for computational power is not the only constraint to effective computational analysis of biological data, however, as the bioinformatics tools and datasets themselves are also highly heterogeneous, both in content and structure. Datasets are often autonomously maintained and deployed by individual laboratories, and are therefore geographically dispersed. The issues of autonomy, distribution and heterogeneity are major problems for the bioinformatician, as they represent barriers to the integration of datasets and can prevent the extraction of meaningful knowledge. More recently, the term Grid has been extended to encompass networks of people and instruments, and it is this type of Grid that provides the ideal infrastructure that is required to make e-Science a reality.

Two Grid technologies are key to the e-Science-based approach to bioinformatics: workflows and Web services.

Workflows. Many bioinformatics tasks operate over distributed data and can be represented as workflows that are analogous to experimental protocols. Each stage in a workflow represents an analytical step that involves the retrieval, storage, processing or representation of data. The benefit of a workflow-based approach is that a series of programs and databases can be integrated to address a particular analytical task, to test a particular hypothesis or to meet a particular scientific objective.

Bioinformatics workflows are typically embedded in custom-written programs, in languages such as Perl, Python or Java, drawing on custom libraries, such as BioPerl[6], BioPython[7] or BioJava[8]. The implementation of bioinformatics applications often requires considerable programming expertise. Applications are not easily reusable, even for slightly different problem domains. A user with limited computer skills often has to use various Web-based tools, cutting and pasting data between applications according to a predetermined series of analyses that seem to meet their needs. e-Science offers an alternative to custom-written scripts and to manual analyses[9]. The data retrieval and analysis requirements of a user are captured in the form of an *in silico* experiment that is analogous to a laboratory-based experiment[10]. A series of tasks that are associated with the running of the experiment are specified in an e-Science workflow that can be created without the need for specialist programming skills (FIG. 2). Workflows are created using a graphical application, such as Taverna[9] (see Further information). Icons that represent tools and services are dragged and dropped onto a workspace and then linked in such a manner that the output from one step becomes the input for the next.

Web services. Grid computing involves the integration of a pool of distributed computing resources, such as databases, instruments and applications, through a network such as the Internet. e-Science workflows can be used

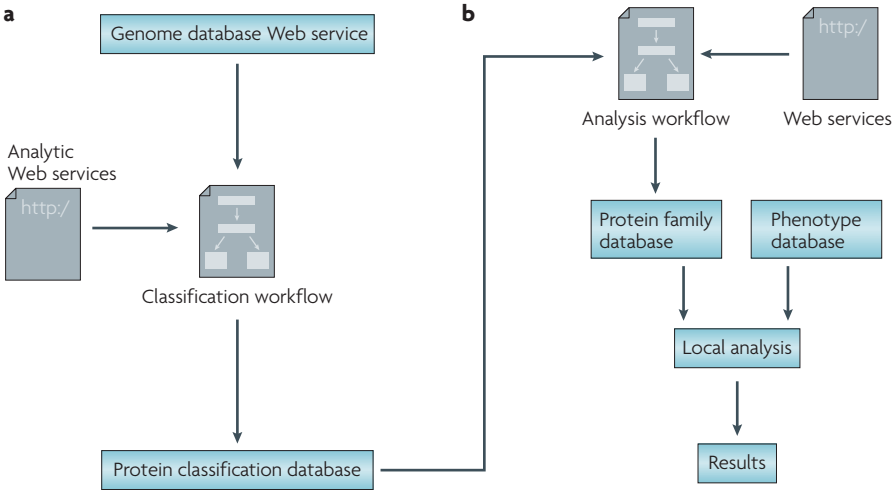


Figure 2 | **Analysis of the *Bacillus subtilis* secretome as an *in silico* experiment. a** | Classification of *B. subtilis* proteins. Data from a genome database that is exposed as a Web service are read into a workflow that incorporates various distributed Web services that implement analytical tools. The output is stored in a local database. **b** | Analysis of protein families. Data from the protein classification database are analysed using a different set of analytical Web services and the results are combined with phenotypic information for analysis and visualization on a local machine.

to draw together these resources to address a particular task in an efficient, automated manner.

Web services are formally defined interfaces that allow computational resources to be exposed in a uniform manner. Whereas Web pages are designed for human interaction, Web services are intended to allow computational applications to interact over a distributed computing network. Thus, Web services provide well-defined methods of computational access to remote programs and databases that would not be available from conventional websites. Programs and databases need to be 'exposed' as Web services if they are to be made available in a computationally accessible form that can be used effectively and can be combined with other Web services using Grid technology. Resources that are exposed in this manner can then be incorporated into e-Science workflows, using specialized software called middleware, to connect the software components or applications together. Resources that are not Web-service enabled can be adapted by writing 'wrapper' code to provide a standard interface between the resource and the workflow.

In this Opinion article, the term Web services is used in the generic sense to include a range of technologies that provide computational access to data, including a number of emerging technologies, such as simple object access protocol (SOAP)-based services[11], representational state transfer (REST)-based services[12], BioMoby (an ontology-based interoperability standard between biological data and analytical services; see Further information)[13] and Grid services. Grid services were originally proposed as the service framework for the Grid. However, in recent years the specifications for Web services and Grid services have converged. The specification for Grid services is still evolving, and the use of Grid services is not widespread in the bioinformatics community. Hence, we will only discuss Web services[14]. Currently, over 3,000 bioinformatics tools and databases have been exposed as Web services[15], but many more remain to be adapted. For example, major bioinformatics providers, such as the NCBI and EBI, are committed to exposing their analytical tools[16]. If a required tool or database is not available with a Web-service interface, users have three options: the tools can be deployed locally on the user's machine; tools such as Soaplab[17] can be used to allow a Web-service interface to be developed rapidly; or Web-service interfaces can be manually coded and programs can be installed on either a local or a remote server.

e-Science workflows specify how remote service-based tools and resources can be computationally linked to allow data to be passed between them. The workflows also define the order in which the various Web-service resources are invoked. In this way, complex, customized computing applications can be built up from tools at remote and disparate sites. A number of tools (for example, Taverna[9,15], Triana[18], GPIPE (graphic pipeline generator)[19] and Kepler[20]) are now available to allow both the biologist and the bioinformatician to produce these workflows (BOX 1). The definition, combination and execution of workflows allow experiments to be performed *in silico* using the Grid[21]. Projects such as myGrid[10] (see Further information) and BioMoby[13]

Box 1 | **Resources for Grid-based analysis**

Various resources are available to implement Grid-based analysis, but not all of these can be discussed here owing to space constraints. Resources that are free and open source (OS) can be modified and adapted to the requirements of the user, if necessary.

Web-service providers

Many of the main bioinformatics data providers make some of their resources available as Web services. EMBL–EBI provides a programmatic interface to many of its resources and tools through Web services. The DNA Data Bank of Japan (DDBJ) also makes a number of search and query tools available in Web-services format (Web API for Biology (WABI)), both for itself and for other databases, such as GenBank, RefSeq, Ensembl, OMIM (Online Mendelian Inheritance in Man) and Gene Ontology. Similarly, the NCBI Entrez Utilities Web Service is also available (see Further information).

Workflows and data integration

- Taverna: is part of the myGrid project and allows the design and implementation of workflows using a simple graphical user interface (OS).
- myExperiment is a social networking site for scientists that aims to make it easy to find, use and share scientific workflows, and to build communities.
- BioCatalogue is a project to build and maintain a comprehensive, curated catalogue of Web services for the life-sciences community.
- BioMoby is a system for interoperability between biological data hosts and analytical services (OS).
- BioMart is a query-oriented data integration system that is particularly suited for providing 'data mining'; for example, searches of complex descriptive data (OS).
- Triana is a problem-solving environment that combines an intuitive visual interface with powerful data-analysis tools.
- GPIPE is a prototype graphic pipeline generator that allows the definition of a pipeline, the parameterization of its component methods and the storage of metadata in XML (extensible markup language) format.

provide a Grid infrastructure that is tailored to support *in silico* experiments, and their value to biological research has already been shown[21–23].

The functionality provided by these systems makes it possible for *in silico* experiments to be replicated and shared, and for the knowledge that underlies the purpose of the workflows to be captured. The myExperiment portal[2] (see Further information) has been designed explicitly as a collaborative environment for scientists to deposit workflows and share them with other scientists. The portal includes facilities for online dialogue and the exchange of experiences in a manner that is similar to the popular social collaboration tools of Facebook (see Further information).

A paradigm shift: automated updates. The ability to integrate data and programs is only one of the benefits offered by Grids. As biological datasets increase in size and coverage, it is becoming increasingly necessary to instigate a paradigm shift in the way biologists interact with bioinformatics tools. It is no longer feasible for biologists to manually keep track of all the information that may be relevant to their research. Information must be gathered, sifted and presented to the user in an automated and personalized format. Grid technology offers users the ability to personalize their *in silico* experiment environment, while at the same time facilitating the process of extracting knowledge from the distributed data. This approach is particularly powerful when combined with 'notification' technology. Notification systems allow distributed computing components to respond to events that are relevant to the software in the workflow or its user. It is then possible to develop systems that automatically inform the user about events that could be 'relevant', newly updated items or even hypotheses that have been automatically generated on the basis of previously defined *in silico* experiments. For example, a user can be informed when a dataset of interest is updated and when relevant information is mined and automatically analysed. The user can then be asked to review the results.

Analysing the *Bacillus* secretome

Here, we illustrate the power of e-Science workflows and associated Grid technology by discussing a study aimed at the development and application of workflows for the genomic scale detection and characterization of secreted proteins from *Bacillus* species[24]. The aim of the study was to identify genes that encode proteins that are likely to be secreted across the cytoplasmic membrane and to predict the likely mode of their secretion[25] (BOX 2). Once putative secretory proteins were identified and classified, the composition of the secretome in and between species was investigated. This workflow, if performed manually by querying appropriate databases, reformatting results and running tools individually, would take weeks of effort. This effort would then need to be repeated when analysis was redone: for example, if the data changed, if new tools were developed or if errors in the analysis were discovered. Once designed the workflow takes only a few hours to process the data, and the workflow can be extended or rerun with minimal effort.

Identifying and classifying secretome proteins. Workflows were developed to predict the total complement of secreted proteins from the 12 complete *Bacillus* genomes that were available from the EMBL–EBI genomes database. The outputs of the workflow were stored in a dedicated database to facilitate collaborative annotation and curation of the predicted secretomes. A classification workflow and a secretome analysis workflow were designed to carry out these analyses. For both workflows, the Taverna tool of the myGrid[10] framework was used to specify the remote services to be used and to orchestrate the order in which the analyses were to be performed.

The classification workflow used existing tools, such as LipoP[26] and SignalP[27], to detect the signal peptides of lipoproteins and Sec (secreted) translocase substrates, respectively. A combination of two existing tools, TMHMM (transmembrane hidden Markov model)[28] and MEMSAT (membrane protein structure and topology)[29], was used to detect membrane-bound proteins and to map their putative transmembrane domains. The classification workflow produced a database that was populated with predictions about whether putative gene products were likely to be located in, or secreted across, the cytoplasmic membrane. The genomes to be analysed were specified at the start of the workflow and each was processed in turn.

Analysing secretome proteins. The analysis workflow operates over the populated genomic database to analyse the composition of each predicted secretome, functionally classifying the secretory proteins by arranging them into protein families. This procedure is inherently more complex than that of the classification workflow and therefore requires considerably more computational resources. Additional remote-service-enabled computational resources were invoked using the workflow. These

Box 2 | *Bacillus* protein secretion

Between 5–10% of the proteins that are encoded by a bacterium are secreted, primarily through the ubiquitous Sec (secretion)-dependent pathway[42]. These proteins include enzymes that are involved in cell-wall synthesis and the use of nutrients. For pathogenic bacteria, secreted proteins enable the bacteria to interact with and subvert their target host. In addition to the Sec pathway, a few proteins are secreted by the twin arginine transporter (Tat) pathway. Unlike the Sec pathway, the Tat pathway can secrete predominantly folded proteins[43].

Secretory proteins have trafficking signals that direct them to their cognate transporters and, in some cases, their final locations. The presence of these signals means that secretory proteins are an excellent system with which to illustrate the ability of e-Science to address specific biological questions at a multi-genome level. Automated workflow analyses can be designed to address specific questions that relate to, for example, the identification of proteins of potential industrial importance as novel drug and vaccine targets.

An early event in protein secretion is the identification of substrates that become targeted to the secretion apparatus. In most cases, secretory proteins are recognized by a 20–40 residue amino (N)-terminal extension of the mature protein. This extension, called the signal peptide, is removed by a signal peptidase during the later stages of secretion. The signal peptides of the Sec and Tat pathways have similar physical properties. These include a positively charged N terminus (N region), a hydrophobic core (H region) and a short region that contains the target site for cleavage by signal peptidase (C region). Tat signal peptides are generally longer (~37 amino acids) than those of Sec substrates (~25–30 amino acids), and include a twin arginine motif (SRRXFLK) at the junction between the N and H regions. The signal peptides of Sec and Tat substrates are cleaved by type I signal peptidases (Sips) that recognize the consensus sequence ([AVSE]-X-[AGST]↓).

A subset of secretory proteins is acyl-modified at the N terminus of the mature protein, usually at a cysteine residue. The signal peptides of lipoproteins are generally short (~20 residues) and are cleaved by a type II signal peptidase (LspA) that recognizes the consensus sequence [LITAGMV]-[ASGTIMVF]-[AG]-↓C-[SGENTAQR]. The arrow represents the point of cleavage of the protein sequence by the enzyme.

resources used Grid technology to provide the necessary computational power to carry out the intensive tasks that were required for the analysis. The tasks included 'all-against-all' BLAST (basic local alignment search tool)[30] searches, which were used to establish protein-sequence similarity and execute protein-sequence clustering algorithms, such as TribeMCL[31]. These algorithms were then used to assign proteins to particular families. For the TribeMCL algorithm, the Grid resource Microbase[32] (see Further information) was used to provide access to previously calculated BLASTP (BLAST protein) scores for protein–protein sequence comparisons and to provide a high-performance execution and results management system for implementation of the algorithm. The resulting protein families and their functional compositions were visualized using Web-service-enabled functions that were written in the R statistical language, an open-source environment for statistical computing and graphics.

Analysis of the 12 *Bacillus* secretomes resulted in the categorization of 5,329 putative secreted proteins into 673 families of 2 or more members (618 proteins showed no similarity to other proteins at a BLASTP cut-off e-score of e^{-10}). Using the outputs of the analysis workflow, a dendrogram was generated to reflect the contribution of each of the 12 secretomes to the 673 protein families (FIG. 3). The data reveal a spectrum of protein family members: some proteins appear to be unique to a single *Bacillus* genome sequence, whereas others are shared between several genomes. Of the protein families, 45 had at least 1 representative member in each of the 12 genomes analysed. These were termed the core families. Interestingly, some secreted protein families were encoded exclusively by the genomes of pathogenic strains, whereas other families were only encoded by the genomes of environmental strains.

The functions of the various protein families were analysed using the gene ontology (GO) terms that were assigned in the original EMBL records. Analysis of the core secreted proteins revealed that almost one-quarter were members of families of unknown function. A large proportion of the other families appeared to contain proteins with functional assignments that related to membrane transport (substrate-binding proteins), cell-wall-associated proteins and proteins connected with membrane biogenesis. Of the protein families unique to pathogens, most (13 out of 15) contained proteins that had not been previously characterized. These proteins are currently the focus of laboratory-based functional analysis, as they could represent important virulence determinants and could provide suitable targets for the development of novel antimicrobial agents or vaccines.

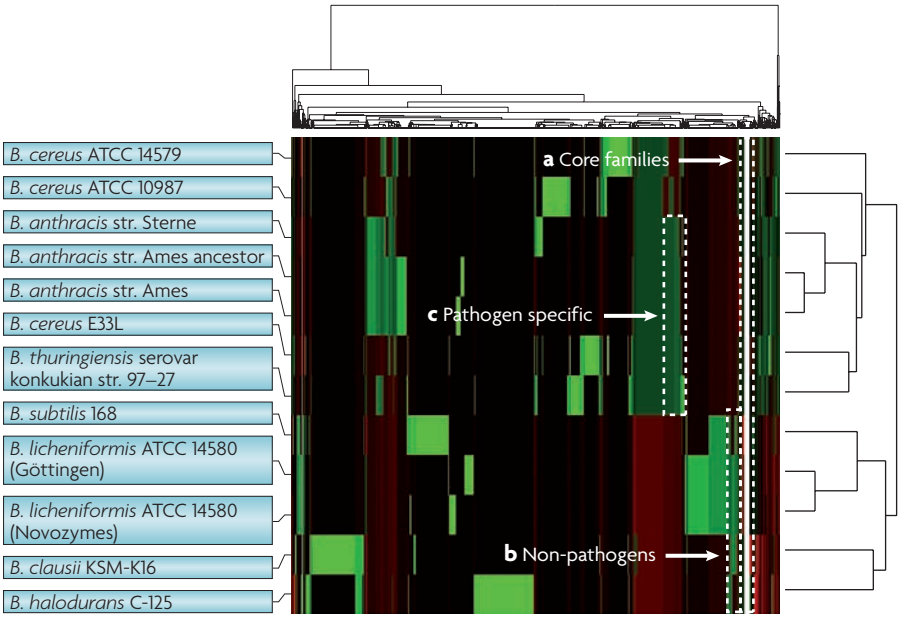


Figure 3 | **A heat map that shows the prevalence of secreted protein families across 12 *Bacillus* species.** The analysis workflow was used to investigate the composition of secreted protein families and the distribution of their members across 12 *Bacillus* genomes. Families that are composed of proteins with a representative from each species (45 of 673; **a**), families that are generally unique to non-pathogens (9 of 673; **b**) and families with members conserved only in pathogen-specific organisms (15 of 673; **c**) are indicated as examples. The intensity of the colour increases along with the number of proteins that contributed to a particular family.

Benefits of the e-Science approach. The computational study of secretory proteins reveals the benefits of using e-Science technology to analyse genomic data. A major advantage of the e-Science approach is the way it simplifies the use of powerful bioinformatics resources and therefore allows the 'ordinary' biologist to create and tailor a multiple-genome-scale analysis. The computational power of the Grid dramatically speeds up the time taken to complete the analysis, thereby facilitating its application to many genomes simultaneously. The biologist also benefits from information that improves the reproducibility of the data. The workflow enactor automatically gathers extensive provenance, detailing the history and resources of the process that was used to generate the data.

Another key element of the workflow approach to e-Science is its flexibility. The two workflows that are involved — one identifies and classifies secretory proteins, whereas the other analyses their sequence similarity and family membership — are composed of multiple processes that can be easily modified. Many workflows are naturally modular in nature, and modular sections of workflows can be substituted for one another. The formulation of workflows still requires an understanding of bioinformatics tools and applications, together with an appreciation of their mode of operation. However, workflows can be designed with little or no programming experience, unlike more conventional approaches, such as Perl scripting. Moreover, many workflow environments, such as Taverna and Triana, offer graphical approaches to workflow composition, further simplifying workflow construction.

One of the principles of e-Science is the emphasis it places on collaborative science. The example discussed above is a typical *in silico* experiment that both uses previously defined resources (such as SignalP, LipoP, TMHMM and MEMSAT) and promotes the sharing of new resources, such as workflows and the resulting databases. Workflows might already exist to perform part or all of a particular analysis, and these can often be found in workflow repositories, such as myExperiment. In the past, existing Web services, such as EMBL file parsers and BLASTP, were often identified by word of mouth or through the Taverna service palette. More recently, Web-service registries have

become available in a Grid environment. These registries, such as BioCatalogue and Feta (the myGrid semantic discovery tool for searching available web services), describe the function and operation of services in a computationally amenable manner (by assigning semantically defined metadata) that supports a search for services with the required specific characteristics. Users can therefore specify, for example, that a BLASTP service must operate on an amino-acid sequence in FASTA format in a named database and return a text report.

In some cases, Web services provide access to large and powerful computational systems that exploit Grid technology behind the scenes. For example, Microbase[32,33], which was used in the secretome analysis, has access to a large all-against-all database of pre-computed BLASTP results for over 300 genomes; this database eliminates the need for these comparisons to be generated repeatedly and speeds up the operation of the workflow. Sometimes, the required services are not available and programming skills are required to develop them *de novo*. However, tool kits, such as myGrid Soaplab, ease the process of exposing conventional bioinformatics tools as Web services. Doing so allows others to use the resources either as a whole or as functional units that can be assembled to form different functional workflows.

e-Science in the future

Grid technology is still evolving and it will be some time before the vision that drives the development of e-Science is fully realized and virtual laboratory environments become commonplace. Many bioinformatics resources still need to be made available in a form that is amenable for use in a workflow environment. Grid technology has yet to develop effective methods to allow the service-based use of software that requires an individual user license. Another challenge is the development of mechanisms which ensure that service providers receive recognition for the use of their tools and resources in *in silico* experiments.

Despite these challenges, systems for the analysis of complete microbial genomes are already beginning to be developed. These systems include e-Fungi[34], Xbase2 (REF. 35), Microbase[32,33], Gnare[36], GADU (genome analysis and database update)[37], CAMERA (community cyberinfrastructure for advanced marine microbial ecology research and analysis)[38] and Puma2 (REF. 39), which annotate and compare new genomes and metagenomes in a semi-automated manner.

Systems that exploit developments in the integration of semantic data (information on relationships between genes, proteins and organisms, using controlled vocabularies and ontologies[40]) are particularly promising. Such systems incorporate knowledge about the structure and meaning of data on the Web (the so-called semantic web) to amalgamate data from disparate sources and allow information that is derived from downstream 'omics'-based studies to be combined automatically and dynamically with the primary sequence data. Systems such as ONDEX[41] (ontology-based index; see Further information) increasingly exploit Grid technology to build integrated warehouses (collections of resources and data) of genomic data, post-genomic experimental results and information gathered automatically from the literature. As a result, genomic data can be integrated with other information in the context of a graph or network that can be explored visually or mined computationally.

Technological advances that are associated with the Grid, and the vision for collaborative research presented by e-Science, are essential for the computational genomics of the future. In addition, the flexibility offered by workflow-based approaches enhances the ability of both biologists and bioinformaticians to analyse multiple genomes effectively. Recent emphasis has been placed on providing systems with enough computing power to perform scalable data analyses, such as comparative genomics. Although this type of research will continue to be worthwhile, future challenges lie in developing systems that generate knowledge that is directly relevant to the biologist, either for system-based or hypothesis-driven research.

Tracy Craddock, Jennifer Hallinan and Anil Wipat are at the School of Computing Science, Claremont Tower, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK.

Colin R. Harwood is at the Institute for Cell and Molecular Biosciences, Medical School, Newcastle University, Newcastle upon Tyne, NE2 4HH, UK.

Correspondence to C.R.H.
e-mail: Colin.Harwood@ncl.ac.uk

doi:10.1038/nrmicro2031

1. Luciano, J. S. & Stevens, R. D. e-Science and biological pathway semantics. *BMC Bioinformatics* **8**, S3 (2007).
2. de Roure, D., Goble, C. & Stevens, R. in *Proc. 2007 IEEE Conf. eScience Grid Comput.* 603–610 (2007).
3. Foster, I., Kesselman, C. & Tuecke, S. The anatomy of the Grid: enabling scalable virtual organizations. *Int. J. High Perform. Comput. Appl.* **15**, 200–222 (2001).
4. Foster, I. & Kesselman, C. Globus: a metacomputing infrastructure toolkit. *Int. J. High Perform. Comput. Appl.* **11**, 115–128 (1997).
5. Thain, D., Tannenbaum, T. & Livny, M. in *Grid Computing* (eds Berman, F., Fox, G. & Hey, T.) 299–335 (2003).
6. Stajich, J. E. *et al.* The Bioperl toolkit: perl modules for the life sciences. *Genome Res.* **12**, 1611–1618 (2002).
7. Chapman, B. & Chang, J. Biopython: Python tools for computational biology. *ACM SIGBIO Newsl.* **20**, 15–19 (2000).
8. Pocock, M., Down, T. & Hubbard, T. BioJava: open source components for bioinformatics. *ACM SIGBIO Newsl.* **20**, 10–12 (2000).
9. Oinn, T. *et al.* Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**, 3045–3054 (2004).
10. Stevens, R. D., Robinson, A. J. & Goble, C. A. myGrid: personalised bioinformatics on the information grid. *Bioinformatics* **19**, I302–I304 (2003).
11. Curbera, F. *et al.* Unraveling the web services web: an introduction to SOAP, WSDL, and UDDI. *IEEE Internet Comput.* **6**, 86–93 (2002).
12. Khare, R. & Taylor, R. N. in *Proc. 26th Int. Conf. Software Eng.* (ed. Taylor, R. N.) 428–437 (2004).
13. Wilkinson, M. D. & Links, M. BioMOBY: an open source biological web services proposal. *Brief. Bioinformatics* **3**, 331–341 (2002).
14. Foster, I., Kesselman, C., Nick, J. M. & Tuecke, S. Grid services for distributed system integration. *Computer* **35**, 37–46 (2002).
15. Hull, D. *et al.* Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.* **34**, W729–W732 (2006).
16. Pillai, S. *et al.* SOAP-based services provided by the European Bioinformatics Institute. *Nucleic Acids Res.* **33**, W25–W28 (2005).
17. Senger, M., Rice, P. & Oinn, T. in *UK e-Science All Hands Meet. 2003* (ed. Cox, S. J.) 509–513 (2003).
18. Majithia, S., Shields, M., Taylor, I. & Wang, I. in *Proc. IEEE Intern. Conf. Web Services* (ed. Shields, M.) 514–521 (2004).
19. Castro, A. G., Thoraval, S., Garcia, L. J. & Ragan, M. A. Workflows in bioinformatics: meta-analysis and prototype implementation of a workflow generator. *BMC Bioinform.* **6**, 87 (2005).
20. Ludäscher, B. *et al.* Scientific workflow management and the kepler system. *Concurr. Comput. Pract. Exper.* **18**, 1039–1065 (2006).
21. Stevens, R. *et al.* myGrid and the drug discovery process. *Drug Discov. Today* **2**, 140–148 (2004).
22. Fisher, P. *et al.* A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis. *Nucleic Acids Res.* **35**, 5625–5633 (2007).
23. Agostini, F. P., Soares-Pinto, D. O., Moret, M. A., Osthoff, C. & Pascutti, P. G. Generalized simulated annealing applied to protein folding studies. *J. Comput. Chem.* **11**, 1142–1152 (2006).
24. Craddock, T., Lord, P., Harwood, C. R. & Wipat, A. in *Proc. 5th UK e-Science All Hands Meet.* 788–795 (2006).
25. Harwood, C. R. & Cranenburgh, R. *Bacillus* protein secretion: an unfolding story. *Trends Microbiol.* **16**, 73–79 (2008).
26. Juncker, A. S. *et al.* Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* **12**, 1652–1662 (2003).
27. Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
28. Sonnhammer, E. L. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182 (1998).
29. Jones, D. T., Taylor, W. R. & Thornton, J. M. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33**, 3038–3049 (1994).
30. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
31. Enright, A. J., Kunin, V. & Ouzounis, C. A. Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.* **31**, 4632–4638 (2003).
32. Sun, Y. *et al.* Exploring microbial genome sequences to identify protein families on the Grid. *IEEE Trans. Inf. Technol. Biomed.*, 11 (2007).
33. Sun, Y. *et al.* in *2005 IEEE International Symposium on Cluster Computing and the Grid* 977–984 (2005).
34. Hedeler, C. *et al.* e-Fungi: a data resource for comparative analysis of fungal genomes. *BMC Genomics* **8**, 426 (2007).

35. Chaudhuri, R. R. *et al.* xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res.* **36**, D543–D546 (2008).
36. Sulakhe, D. *et al.* Gnare: automated system for high-throughput genome analysis with Grid computational backend. *J. Clin. Monit. Comput.* **19**, 361–369 (2005).
37. Sulakhe, D., Rodriguez, A., Wilde, M., Foster, I. A. & Maltsev, N. A. Interoperability of GADU in using heterogeneous grid resources for bioinformatics applications. *IEEE Trans. Inf. Technol. Biomed.* **12**, 241–246 (2008).
38. Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P. & Frazier, M. CAMERA: a community resource for metagenomics. *PLoS Biol.* **5**, e75 (2007).
39. Maltsev, N. A. *et al.* PUMA2-grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res.* **34**, D369–D372 (2006).
40. Schulze-Kremer, S. Ontologies for molecular biology. in *Proc. 3rd Pacific Symp. Biocomput.*, 693–704 (1998).
41. Kohler, J. *et al.* Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics* **22**, 1383–1390 (2006).
42. Papanikou, E., Karamanou, S. & Economou, A. Bacterial protein secretion through the translocase nanomachine. *Nature Rev. Microbiol.* **5**, 839–851 (2007).
43. Berks, B. C., Palmer, T. & Sargent, F. Protein targeting by the bacterial twin-arginine translocation (Tat) pathway. *Curr. Opin. Microbiol.* **8**, 174–181 (2005).

Acknowledgements
The authors acknowledge funding from the UK Engineering and Physical Sciences Research Council and Non-linear Dynamics for a CASE (collaborative awards in science and engineering) studentship to T.C., from Research Councils UK for a fellowship to J.H. and from the European Union (Bacell Health; grant number LSH-2002-1.1.0-1).

DATABASES

Entrez Genome Project: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj
Staphylococcus aureus

FURTHER INFORMATION

Anil Wipat's homepage: http://www.cs.ncl.ac.uk/people/anil.wipat
Colin R. Harwood's homepage: http://www.ncl.ac.uk/camb/staff/profile/colin.harwood
Jennifer Hallinan's homepage: http://www.cs.ncl.ac.uk/people/J.S.Hallinan
BioCatalogue: http://www.biocatalogue.org
BioMart: http://www.biomart.org
BioMoby: http://biomoby.open-bio.org/index.php/what-is-moby
EMBL–EBI (Genome Pages — Bacteria): http://www.ebi.ac.uk/genomes/bacteria.html
Ensembl: http://www.ensembl.org/index.html
Facebook: http://www.facebook.com
GenBank: http://www.ncbi.nlm.nih.gov/Genbank/index.html
Gene Ontology: http://www.geneontology.org
KEGG bacterial genomes: http://www.genome.jp/kegg
Microbase: http://www.microbase.gr
myExperiment: http://www.myexperiment.org
myGrid project: http://www.mygrid.org.uk
NCBI Entrez Genome: http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome
NCBI Entrez Utilities Web Service: http://www.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html
OMIM: http://www.ncbi.nlm.nih.gov/omim
ONDEX: http://ondex.sourceforge.net
RefSeq: http://www.ncbi.nlm.nih.gov/RefSeq
Taverna: http://taverna.sourceforge.net
Triana: http://www.trianacode.org/index.html
WABI: http://www.xml.nig.ac.jp/index.html

ALL LINKS ARE ACTIVE IN THE ONLINE PDF